Design and Analysis of Optical Flow Switched Networks

by

Guy E. Weichenberg

B.A.Sc., University of Toronto (2001)
S.M., Massachusetts Institute of Technology (2003)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

February 2009

Author .....
Department of Electrical Engineering and Computer Science
December 1, 2008

Certified by....
Vincent W. S. Chan
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by.....
Muriel Médard
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by.....
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Design and Analysis of Optical Flow Switched Networks

by

Guy E. Weichenberg

Submitted to the Department of Electrical Engineering and Computer Science on December 1, 2008, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

In the four decades since optical fiber was introduced as a communications medium, optical networking has revolutionized the telecommunications landscape. It has enabled the Internet as we know it today, and is central to the realization of Network-Centric Warfare in the defense world. Sustained exponential growth in communications bandwidth demand, however, is requiring that the nexus of innovation in optical networking continue, in order to ensure cost-effective communications in the future.

In this thesis, we present Optical Flow Switching (OFS) as a key enabler of scalable future optical networks. The general idea behind OFS—agile, end-to-end, all-optical connections—is decades old, if not as old as the field of optical networking itself. However, owing to the absence of an application for it, OFS remained an underdeveloped idea—bereft of how it could be implemented, how well it would perform, and how much it would cost relative to other architectures. The contributions of this thesis are in providing partial answers to these three broad questions. With respect to implementation, we address the physical layer design of OFS in the metro-area and access, and develop sensible scheduling algorithms for OFS communication. Our performance study comprises a comparative capacity analysis for the wide-area, as well as an analytical approximation of the throughput-delay tradeoff offered by OFS for inter-MAN communication. Lastly, with regard to the economics of OFS, we employ an approximate capital expenditure model, which enables a throughput-cost comparison of OFS with other prominent candidate architectures. Our conclusions point to the fact that OFS offers significant advantage over other architectures in economic scalability. In particular, for sufficiently heavy traffic, OFS handles large transactions at far lower cost than other optical network architectures. In light of the increasing importance of large transactions in both commercial and defense networks, we conclude that OFS may be crucial to the future viability of optical networking.

Thesis Supervisor: Vincent W. S. Chan
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Muriel Médard
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I owe my two research supervisors, Vincent W. S. Chan and Muriel Médard, an immeasurable debt of gratitude. While any graduate student would be exceedingly fortunate to have either Vincent or Muriel in their corner, I have had the great fortune of having them both. Their support—intellectual, financial, and emotional in tough times—enabled my doctoral studies, and shaped my development as a researcher. Their mentorship and living examples, moreover, have inspired the person that I have become (the good parts), and the person that I aspire to be. To the extent that I am an engineer, I must credit Vincent greatly. His passion for technical challenges has been contagious, and his uncanny engineering intuition has inspired me to always look beyond the equations and understand things at a gut level. He has, moreover, been a pristine example of patience, generosity, and integrity to follow. Among many things, I thank Muriel for nurturing my analytical maturity, and for instilling in me the value of being bold and aiming high in research. Her broad-minded wisdom and the astonishing breadth and depth of her life, moreover, continue to serve as a shining beacon for me as I move forward. In addition to my research supervisors, I would like to express my appreciation to Eric Swanson, the third member of my thesis committee. His engineering wisdom, imparted through valuable discussions, critiques, and suggestions, has significantly improved this thesis. I also wish to thank Eric for his friendship and general life advice, which have meant much to me.

I have been fortunate to have met Terry McGarty during my latter years at MIT. I thank him for broadening my horizons beyond the rigors of my technical work, for his wellspring of sage counsel, and for the doors that he has opened for me. I have also benefitted from the avuncular wisdom of Bob Gallager at several critical junctures along my journey at MIT. The friendship and administrative support offered by Doris Inslee, Brian Jones, Michael Lewy, Pardis Parsa, Sue Patterson, Liz Reid, Rosangela dos Santos, and Kathy Sullivan are much appreciated and have made my years as a graduate student very comfortable. Though I have had many enjoyable interactions with my peers at MIT over the years, those with the Old Guard—Lillian Dai, Kyle Guan, Danielle Hinton, Julius Kusuma, Etty Lee, Paul Njoroge, Florent Ségonne, Yonggang Wen—and the New Guard—Anurupa Ganguly, James Glettler, MinJi Kim, and Andrew Puryear, Jay Kumar Sundararajan, Lei Zhang—have been among the most memorable. I am especially grateful to Desmond Lun, a perennial

and supportive presence since we joined Muriel's group together several years ago; and to Ted Sargent, whose role in my life has evolved from respected teacher to very dear (and still respected) friend.

Lastly, I have my family to thank. I owe any modest successes in my life to my parents' sacrifices and devotion in providing me with opportunities that they only dreamt of for themselves. I also wish to thank my two little brothers, Oren and Shawn, for their cherished company during my visits home to Toronto. These final words are reserved for Haixia—my ballast, my inspiration, and my very best friend. The two of us have come a long way, and I credit this to her idealism, steadfastness, and love. This thesis is dedicated to her.

Guy Weichenberg
Cambridge, Mass.

The research in this thesis was generously supported by: Defense Advanced Research Projects Agency grants HR0011-08-1-0008 and HR0011-06-1-0011; National Science Foundation grants CNS-0626800 and ANI-0335217; and the Natural Sciences and Engineering Research Council of Canada.

Contents

List of Figures

List of Tables

Acronyms

10GEPON	10 Gigabit Ethernet Passive Optical Network
AON	All-Optical Network
AOTF	acousto-optic tunable filter
APD	avalanche photodiode
APON	Asynchronous Transfer Mode Passive Optical Network
ASE	amplified spontaneous emission
ATM	Asynchronous Transfer Mode
AWG	arrayed waveguide grating
BER	bit error rate
BPON	Broadband Passive Optical Network
CapEx	capital expenditure
CCW	coupled cavity waveguide
CGM	cross-gain modulation
CIOQ	combined input- and output-queued
CPM	cross-phase modulation
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
DBR	distributed Bragg reflector
DCF	dispersion compensating fiber
DCS	digital cross-connect
DFB	distributed-feedback laser
DFG	difference frequency generation
DN	distribution network
DoD	Department of Defense
DPT	Dynamic Packet Transport
DSF	dispersion-shifted fiber
ECC	error-correcting code
EDC	electronic dispersion compensation
EDF	erbium-doped fiber
EDFA	erbium-doped fiber amplifier
EIT	electromagnetically induced transparency
EPON	Ethernet Passive Optical Network
EPS	Electronic Packet Switching

ESCON	Enterprise Serial Connection
FB	feed-back
FBG	fiber Bragg grating
FDDI	Fiber Distributed Data Interface
FDL	fiber delay line
FEC	forward error correction
FF	feed-forward
FIFO	first-in first-out
FP	Fabry-Perot
FWM	four-wave mixing
GEPON	Gigabit Ethernet Passive Optical Network
GIG	Global Information Grid
GMPLS	Generalized Multi-Protocol Label Switching
GPON	Gigabit Passive Optical Network
HFC	hybrid fiber coax
HORNET	Hybrid Opto-electronic Ring Network
HOT	Highly Optimized Tolerance
IP	Internet Protocol
IQ	input-queued
LAN	local-area network
LDPC	Low Density Parity Check Codes
LED	light-emitting diode
MAC	medium access control
MAN	metropolitan-area network
MEMS	micro-electro-mechanical systems
MLM	multiple-longitudinal mode
MONET	Multiwavelength Optical Networking
MPLS	Multi-Protocol Label Switching
MQW	multi-quantum well
MTIT	Multitoken Interarrival Time
MWM	Maximum Weight Matching
MZI	Mach-Zehnder interferometer
NGI ONRAMP	Next Generation Internet Optical Network for Regional Access using Multi-wavelength Protocols
OADM	optical add-drop multiplexer
OBS	Optical Burst Switching
OCDM	optical code-division multiplexing
OCS	Optical Circuit Switching
OEO	optical-electronic-optical
OFS	Optical Flow Switching
OLT	optical line terminal
ONU	optical network unit

OOK	on-off keying
OpEx	operating expenditure
OPS	Optical Packet Switching
ORION	Overspill Routing in Optical Networks
OXC	optical cross-connect
P2P	peer-to-peer
PDL	polarization-dependent loss
P-K	Pollaczek-Khinchin
PMD	polarization-mode dispersion
PON	passive optical network
PPP	Point-to-Point Protocol
PSC	passive star coupler
QoS	quality of service
RAM	random access memory
RF	radio-frequency
ROADM	reconfigurable optical add-drop multiplexer
RPR	Resilient Packet Ring
RWA	routing and wavelength assignment
SAN	storage-area network
SBS	stimulated Brillouin scattering
SDH	Synchronous Digital Hierarchy
SFG	sum frequency generation
SG	sampled grating
SG/DBR	sampled grating distributed Bragg reflection
SHG	second harmonic generation
SLLN	Strong Law of Large Numbers
SLM	single-longitudinal mode
SNR	signal-to-noise ratio
SOA	semiconductor optical amplifier
SOC	Self-Organized Criticality
SONET	Synchronous Optical Network
SPM	self-phase modulation
SRS	stimulated Raman scattering
SSS	Static Service Split
TaG	Tell-and-Go
TCP	Transport Control Protocol
TDM	time-division multiplexing
TFF	thin-film resonant cavity filter
TFMF	thin-film resonant multicavity filter
TOAD	Terahertz Optical Asymmetric Demultiplexer
TPON	Passive Optical Network for Telephony
TWIN	Time-domain Wavelength Interleaved Network

UDP	User Datagram Protocol
UNI	Ultrafast Nonlinear Interferometer
VCSEL	vertical cavity surface-emitting laser
VOD	video-on-demand
VOIP	voice-over-IP
VOQ	virtual output queue
WAN	wide-area network
WC	wavelength converter
WDM	wavelength division multiplexing
WSS	wavelength-selective switch
WWW	World Wide Web

Notation

Roman symbols

A_n	Vector of cumulative number of exogenous arrivals of transaction types by time slot n
$\mathcal{A}$	Admissible rate region
a_p	EDFA cross-section area of pump mode inside fiber
$\tilde{a}_p$	EDFA transition cross-section at pump frequency
a_w	Cost of WAN amplifier and dispersion compensation (per wavelength)
$B(x, y, z)$	Binomial probability of y successes from x trials with trial success probability z
b_i	Number of bursts traversing the i^{th} link along a burst's path
$b_{p,h}$	Number of burst types that traverse, but do not originate on the h^{th} hop of a burst's path
$b_{s,h}$	Number of burst types that originate on the h^{th} hop of a burst's path
$\hat{C}^{\text{EPS}}$	Total network cost under EPS
$\hat{C}^{\text{OCS}}$	Total network cost under OCS/EPS
$\hat{C}^{\text{OFS}}$	Total network cost under OFS
$\hat{C}_a^{\text{OFS}}$	Total cost of an OFS DN
$\hat{C}_a^{\text{PON}}$	Total cost of a PON
$\hat{C}_m^{\text{EPS}}$	Total cost of an EPS MAN
$\hat{C}_m^{\text{OFS}}$	Total cost of an OFS MAN
C_w^{EPS}	Average cost of a WAN end-to-end wavelength-granular EPS connection
C_w^{OBS}	Average cost of a WAN end-to-end wavelength-granular OBS connection
C_w^{OCS}	Average cost of a WAN end-to-end wavelength-granular OCS connection
C_w^{OFS}	Average cost of a WAN end-to-end wavelength-granular OFS connection
$\mathcal{C}$	Clique inequality rate region
c_r	Cost of tunable WDM regenerator
c_t	Cost of tunable WDM transponder
D_n	Vector of number of departures from network queues at time slot n
d	Diameter of a graph
E_n	Vector of number of entrances (exogenous and endogenous) to network queues at time slot n

$\mathbf{E}[\cdot]$	Expectation function
$\text{erfc}(x)$	Gaussian error function $\equiv \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-y^2} dy$
F	Number of flow types in a network
F_n	Noise figure $\equiv \frac{\text{SNR}_{\text{in}}}{\text{SNR}_{\text{out}}}$
F_n^d	Noise figure of down-link segment in inter-MAN OFS communication
F_n^l	Noise figure of intra-MAN OFS communication
F_n^R	Noise figure of pre- and main amplifiers
F_n^u	Noise figure of up-link segment in inter-MAN OFS communication
$F_{n,d}$	Effective noise figure of destination DN stage
$F_{n,d}^e$	Effective noise figure of destination DN stage, assuming external amplification
$F_{n,d}^i$	Effective noise figure of destination DN stage, assuming internal amplification
$F_{n,m}^d$	Effective noise figure of destination MAN
$F_{n,m}^s$	Effective noise figure of source MAN
$F_{n,s}$	Effective noise figure of source DN stage
$F_{n,s}^e$	Effective noise figure of source DN stage, assuming external amplification
$F_{n,s}^i$	Effective noise figure of source DN stage, assuming internal amplification
$F_{n,w}^d$	Effective noise figure of down-link WAN segment
$F_{n,w}^u$	Effective noise figure of up-link WAN segment
f	Number of fibers with inter-MAN OFS traffic connecting a MAN to the WAN
G_d	Optical gain of destination DN stage
G_m^d	Optical gain of destination MAN stage
G_m^s	Optical gain of source MAN stage
G_s	Optical gain of source DN stage
G_w^d	Optical gain of down-link WAN segment
G_w^u	Optical gain of up-link WAN segment
g_d^e	Optical gain of amplifier in destination DN stage with external amplification
g_d^i	Optical gain of amplifier in destination DN stage with external amplification
g_m	Optical gain of MAN amplifier
g_s^e	Optical gain of amplifier in source DN stage with external amplification
g_s^i	Optical gain of amplifier in source DN stage with external amplification
h	Planck constant $\equiv 6.62 \times 10^{-34}$ J $\cdot$ s
h_w	Average number of hops of a WAN connection
$h_{\mathscr{G}}(\cdot,\cdot)$	Average shortest path distance of a Generalized Moore Graph as a function of number of nodes and maximum node degree
$h_{\mathscr{M}}(\cdot,\cdot)$	Average shortest path distance of a Moore Graph as a function of maximum node degree and diameter

$I(\cdot)$	Indicator function
I_0	Photocurrent corresponding to a 0 bit
I_1	Photocurrent corresponding to a 1 bit
I_d	Dark current
k	Number of tributaries per DN
k_B	Boltzmann constant $\equiv 1.38 \times 10^{-23}$ J/K
L	Random variable representing the length of a flow or burst
$L_{\max}$	Maximum length of a flow
$\mathscr{L}(\cdot)$	Lyapunov function
l_a	Optical amplifier spacing on a fiber link
l_m	Average MAN fiber link length
l_u	Average end-user fiber link length in the DN
l_w	Average WAN fiber link length
$\mathscr{M}(\cdot,\cdot)$	Moore Bound as a function of node degree and diameter
m	Cost of OLT chassis (per wavelength)
$\mathbb{N}^c$	Set of c-dimensional vectors with components belonging to the set of natural numbers
n_a	Number of end-users within an access network
$\tilde{n}_a$	Number of access networks per MAN
n_m	Number of nodes within a MAN
n_{sp}	Spontaneous emission factor
n_t	Maximum number of end-users per DN tributary
n_u	Number of end-users per MAN
n_w	Number of nodes within a WAN
P_a	Amplified signal power incident at a photodetector
$P_{\max}$	Maximum launch power per wavelength channel
$\overline{P}_{\text{OFS}}$	Total EDFA pump power, normalized by saturation pump power, required for an OFS DN
P_p	EDFA pump power
$\overline{P}_p$	EDFA pump power normalized by saturation pump power
P_p^{sat}	EDFA saturation pump power
$\overline{P}_{\text{PON}}$	Total EDFA pump power, normalized by saturation pump power, required for a PON
P_s	Input optical signal power
$P_s(i)$	Probability that a burst of type i is successfully received
P_{sens}	Receiver sensitivity
P_{sens}^T	Thermal noise-limited receiver sensitivity
P_{sp}	Spontaneous emission noise power
$\Pr(\cdot)$	Probability of an event
$\mathcal{P}$	OFS capacity region
$\overline{p}_d^i$	EDFA pump power, normalized by saturation pump power, required for a destination DN tributary

$\widetilde{p}_d^i$	EDFA pump power, normalized by saturation pump power, required for the bus of a destination DN
$\overline{p}_s^i$	EDFA pump power, normalized by saturation pump power, required for a source DN tributary
$p_X(\cdot)$	Probability density function or mass function of random variable X
Q	Number of queues in a network
$\overline{Q}_{M_d}^{D_d}(\omega)$	Secondary request queue associated with destination DN D_d in MAN M_d and wavelength channel ω
$\hat{Q}_{M_s}^{D_s}(\omega)$	Secondary request queue associated with source DN D_s in MAN M_s and wavelength channel ω
$Q_{M_s}^{M_d}$	Primary request queue associated with source MAN M_s and destination MAN M_d
$\mathscr{Q}$	Q-factor $\equiv \frac{I_1 - I_0}{\sigma_1 + \sigma_0}$
q	Average idle time between burst attempts
$\hat{q}$	Electronic charge $\equiv 1.602 \times 10^{-19}$ C
$\hat{R}$	Photodetector responsivity
R_L	Resistance of load resistor
$\mathbb{R}_+$	Set of nonnegative real numbers
r_m	Cost of MAN router port
r_w	Cost of WAN router port
S	Throughput of a wavelength channel
s_m	Cost of OXC port with MAN amplification
s_w	Cost of OXC port with WAN amplification and dispersion compensation
T	Absolute temperature
t	Number of transceivers per fiber link
t_l	Cost of tunable long-reach WDM transceiver
t_m	Cost of tunable medium-reach transceiver
U^{EPS}	Utilization of a WAN wavelength channel under EPS
U^{OCS}	Utilization of a WAN wavelength channel under OCS/EPS
U^{OFS}	Utilization of a WAN wavelength channel under OFS
u	Cost of an OFS scheduler
W	Average queueing delay of a transaction
$\hat{W}_{G,k}(\cdot)$	Average queueing delay of $M/G/k$ queueing system as a function of offered load
$\hat{W}_{M,k}(\cdot)$	Average queueing delay of $M/M/k$ queueing system as a function of offered load
$\mathscr{W}^*(\cdot)$	Maximum weight in a MaxWeight scheduling policy as a function network queue lengths
$\mathcal{W}(\cdot)$	Lambert function; inverse of $f(x) = xe^x$
w	Number of wavelength channels per fiber link

w_l	Number of wavelength channels per fiber link for intra-MAN communication via OFS
w_m	Number of OFS WAN wavelength channels provisioned for a source-destination MAN pair
w_t	Number of wavelength channels per fiber link for inter-MAN communication via OFS
w_u	Number of wavelength channels of uniform all-to-all intra-MAN traffic
X	Service time of a primary request
X_n	Vector of number of transactions residing in network queues at time slot n
Y	Queueing delay experienced by a primary request at the head of its queue
Z_d	Time spent by a secondary destination request in its queue prior to reaching the head of the queue
Z_s	Time spent by a secondary source request in its queue prior to reaching the head of the queue

Greek symbols

α_f^a	Cost of deploying fiber (per km·wavelength) in the access environment
$\hat{\alpha}_f^a$	Cost of deploying fiber (per km) in the access environment
α_f^m	Cost of deploying fiber (per km·wavelength) in the MAN
$\hat{\alpha}_f^m$	Cost of deploying fiber (per km) in the MAN
α_f^w	Cost of deploying fiber (per km·wavelength) in the WAN
$\hat{\alpha}_f^w$	Cost of deploying fiber (per km) in the WAN
α_m	$M/G/k$ delay scaling factor $\equiv \frac{\hat{W}_{M,w_m}(\lambda_m)}{\hat{W}_{M,1}(\lambda_m/w_m)}$
α_p	EDFA pump erbium-doping absorption coefficient
α_s	EDFA signal erbium-doping absorption coefficient
α_t	$M/G/k$ delay scaling factor $\equiv \frac{\hat{W}_{M,w_t}(w_t\lambda_c)}{\hat{W}_{M,1}(\lambda_c)}$
$\beta_{c_r}^e$	Average number of non-nodal regenerations of a WAN connection in OCS and OFS
$\beta_{c_r}^o$	Average number of non-nodal regenerations of a WAN connection in EPS
Γ_d^e	Worst-case overall gain from ingress of destination DN to end-user, assuming external amplification
Γ_d^i	Worst-case overall gain from ingress of destination DN to end-user, assuming internal amplification
Γ_s^e	Worst-case overall gain from end-user to egress of source DN, assuming external amplification
Γ_s^i	Worst-case overall gain from end-user to egress of source DN, assuming internal amplification

γ_d^e	Tap coupling coefficient across bus in destination DN, assuming external amplification
γ_d^i	Tap coupling coefficient across bus in destination DN, assuming internal amplification
γ_m	Combined insertion loss of MAN OXC and fiber run immediately following
γ_s^e	Tap coupling coefficient across bus in source DN, assuming external amplification
γ_s^i	Tap coupling coefficient across bus in source DN, assuming internal amplification
γ_t	Tap coupling coefficient across bus in source DN tributary, assuming internal amplification
Δ	Degree of a node in a graph
Δ^*_{EPS}	Optimal node degree of an EPS MAN
Δ^*_{OFS}	Optimal node degree of an OFS MAN
Δf	Effective electrical bandwidth
$\Delta\nu_{\text{opt}}$	Optical bandwidth of spontaneous-emission noise
$\Delta\nu_{\text{sp}}$	Effective bandwidth of spontaneous emission
δ	Excess loss of a passive 2×2 coupler
ζ_d^e	Tributary (excess and splitting) loss in destination DN, assuming external amplification
ζ_d^i	Tributary (excess and splitting) loss in destination DN, assuming internal amplification
ζ_s^e	Tributary (excess and splitting) loss in source DN, assuming external amplification
ζ_s^i	Tributary (excess and splitting) loss in source DN, assuming internal amplification
η	Photodetector quantum efficiency
η_p	EDFA ratio of absorption to emission at the pump frequency
η_s	EDFA ratio of absorption to emission at the signal frequency
κ_a	Cost ratio of 10 Gbps to 40 Gbps amplifier and dispersion compensation equipment
κ_e	Cost ratio of 10 Gbps to 40 Gbps electronics
κ_h	Ratio of average transaction length to average transaction length plus hardware reconfiguration time
κ_l	Ratio of PON line-rate to WAN line-rate
κ_r	Cost ratio of shorter-reach transmission system equipment used in EPS to longer-reach transmission system equipment used in OCS and OFS
κ_w	WAN capacity build ratio of EPS to OCS or OFS
Λ	Vector of arrival rates of a set of stochastic arrival processes
λ_c	Arrival rate of OFS flows for a source-destination MAN pair, normalized by the number of provisioned wavelength channels

λ_m	Arrival rate of OFS flows for a source-destination MAN pair
ν	Optical frequency
ρ_u	End-user duty cycle assuming the WAN line-rate
σ_0^2	Variance of 0 bit noise photocurrent fluctuation
σ_1^2	Variance of 1 bit noise photocurrent fluctuation
σ_s^2	Variance of signal shot noise photocurrent fluctuation
$\sigma_{\text{s-sp}}^2$	Variance of spontaneous emission shot noise photocurrent fluctuation
$\sigma_{\text{sg-sp}}^2$	Variance of signal-spontaneous beat noise photocurrent fluctuation
$\sigma_{\text{sp-sp}}^2$	Variance of spontaneous-spontaneous beat noise photocurrent fluctuation
σ_T^2	Variance of thermal noise photocurrent fluctuation
τ	Time duration of MAN and WAN reconfiguration
τ_{sp}	EDFA spontaneous lifetime of excited state
Φ	Substochastic vector

Other notation

$\overline{X^k}$	k^{th} moment of random variable X
X^T	Transpose of vector X
$\|X\|$	Euclidean norm of vector X
$\|X\|_1$	Sum of the components of vector X

Chapter 1

Introduction

OPTICAL networking is unfolding as a two-generation story. The initial impetus behind optical networking was the prospect of tapping the vast usable bandwidth of optical fiber—roughly 30 THz—to meet increasing telephony traffic demands. First-generation optical networks of the 1970s and 1980s thus employed optical fibers as replacements for copper links, but otherwise maintained traditional architectures[1] that were tailored to the use of electronic networking components (see Figure 1-1(a)). Informational bottlenecks were thereby preempted at transmission links, but still loomed at network nodes whose operations were constrained by the speed of electronics.

Second-generation optical networks, which began emerging in the 1990s, employ optical networking devices—in addition to fiber—in novel network architectures to mitigate electronic bottlenecks arising from steep growth in data application traffic (see Figure 1-1(b)). Data traffic served by these networks, in addition to possessing orders of magnitude more bits per unit time, is characterized by detailed statistics (e.g., those governing traffic burstiness, transaction lengths) and quality of service (QoS) demands that are quite different from those of the telephony traffic served by first-generation optical networks. This changing nature of network traffic, coupled with the novel properties and cost structures of optical networking devices vis-à-vis electronic networking devices, have thus begged a fundamental rethinking of optical network architecture. In response, measured architectural advancements in the wide-area network (WAN)[2] environment occurred, reducing the cost per transmitted bit by exploiting wavelength division multiplexing (WDM) in conjunction with optical amplification and switching. Nevertheless, the end-user's access to the vast

[1]Here, we take network architecture to mean the implementation of a network's functions (as defined by the needs of its users) via the specification of subsystems, components, and interfaces at the physical layer through to the session layer.

[2]Telecommunications networks are conventionally partitioned into three hierarchical tiers: the WAN, metropolitan-area network (MAN), and access network. (The local-area network (LAN) is a subset of the access network.) Although the boundaries between these network tiers are not well-defined, WANs are thought of as networks spanning inter-regional or global distances of 1000 km or more; access networks provide connectivity between end-users and the service provider; and MANs serve as the intermediary between access networks and WANs and are primarily responsible for aggregating/disaggregating data [214].

core network[3] bandwidth has been restrained by the lag in architectural innovation in the MAN and access environments. To a significant extent, then, the future economic viability of optical networking hinges on cost-effective access to core network bandwidth.

In this thesis, we present OFS as an attractive candidate network architecture which will help support future traffic growth by providing this desired cost-effective access [58,59,105]. OFS is an end-to-end transport service, in that it directly connects source and destination end-users through the access network, MAN, and WAN. It is, moreover, well-suited to users with large transactions (i.e., those that can fully utilize a wavelength channel for hundreds of milliseconds or longer), which are expected to contribute significantly to future traffic volume[4]. Furthermore, OFS can be readily implemented with today's device technology [111], since it possesses a simple, all-optical data plane that is separated from its electronic control plane. In addition to improving the QoS for its direct end-users, OFS has the additional important benefit of lowering access costs for *all* users by relieving WAN routers of the onerous burden of serving large transactions.

■ 1.1 Drivers of optical network architecture

In this section, we address the aforementioned two principal drivers of optical network architecture—traffic trends and device technology—in more detail, with a view towards motivating our focus in this thesis on OFS. While this may suggest a chain of causality in which applications and device technology determine network architecture, the picture is, in reality, more complicated: feedback certainly exists in which network architecture also influences the trajectory of device innovation (e.g., widely-tunable lasers), and applications, and by extension the traffic they generate, arise and adapt to constraints imposed by network architecture.

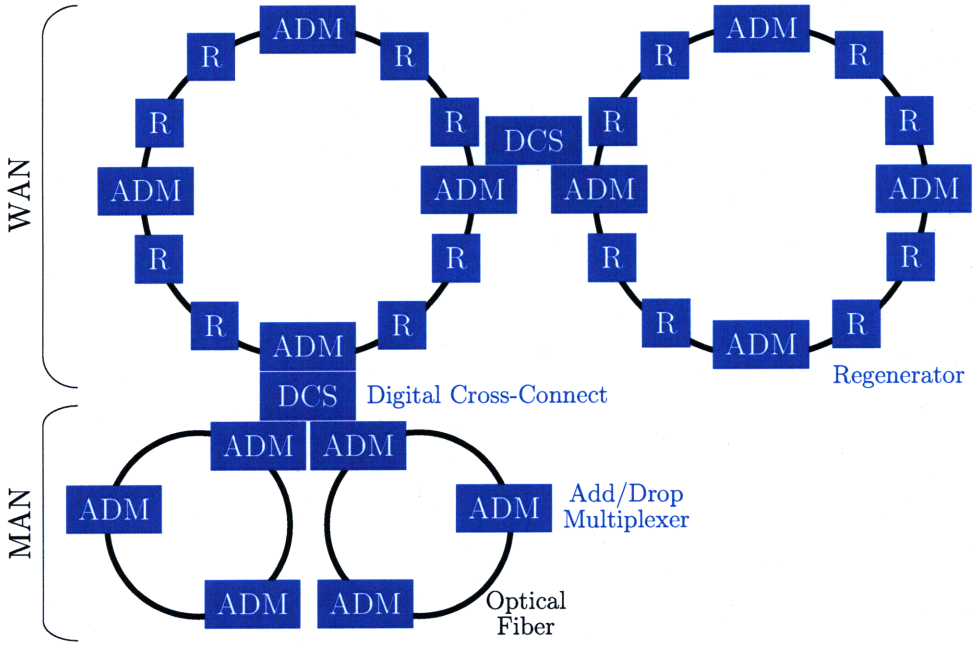
■ 1.1.1 Traffic

A critical, initial step in architecting a network is an assessment of the network users' needs. History, however, through a long trail of foolhardy prophecies on application innovation—"640 kB ought to be enough for anyone"[5]—urges us to exercise caution in this endeavor. With respect to Internet traffic, the prime example of traffic being carried by an optical network, the trends are especially surprising [292]. In the early 1990s, when World Wide Web (WWW) traffic accounted for a couple percent of all traffic on the Internet, who could have expected that by the turn of the millennium WWW traffic would account for about three quarters of Internet traffic? Similarly,

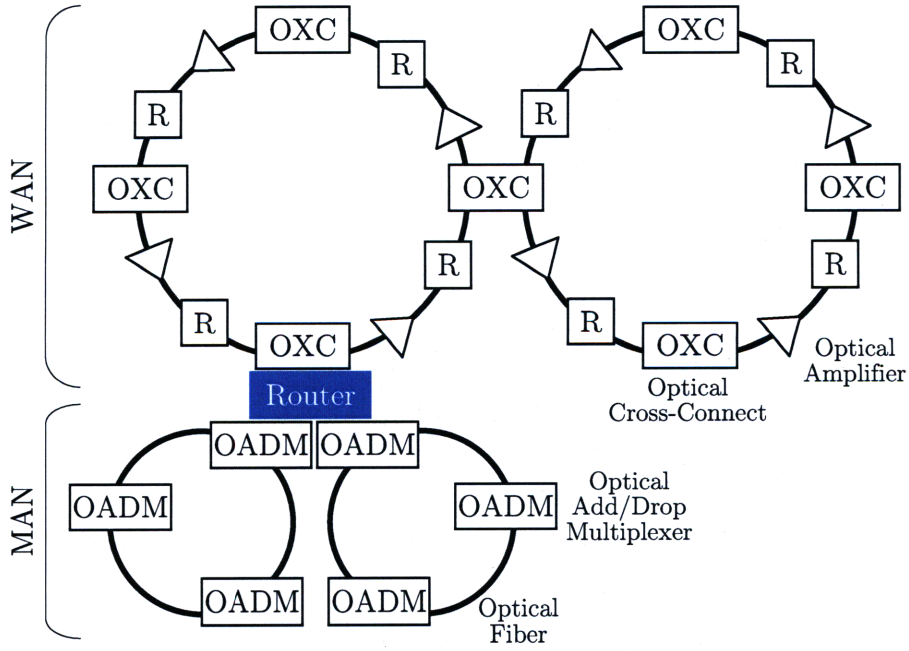
[3]In this thesis, we use the terms "core network" and "WAN" interchangeably.

[4]As will be discussed shortly, the proportion of *transactions* generating this large proportion of traffic volume can be quite small.

[5]Attributed to Bill Gates, co-founder and chairman of Microsoft Corporation, in 1981 in reference to the question of required memory on a personal computer.



(a) An example of a first-generation optical network based upon SONET or SDH rings.



(b) An example of a second-generation optical network which evolved from the first-generation optical network of Figure 1-1(a).

Figure 1-1. Examples of first- and second-generation optical networks. Electronic networking devices are drawn in blue; optical networking devices are drawn in black and white, and are explained in detail in Appendix A. Note that the access environment has not been shown.

in the year 2000 when peer-to-peer (P2P) traffic was virtually nonexistent, who could have expected it to account for two thirds of Internet traffic just four years hence? These trends are indeed humbling and suggest that detailed predictions can run significant dangers—in addition to the ridicule of posterity—if the design of complex, capital-intensive engineering systems, such as optical communication networks, are specifically tailored to them. In this section, we shall therefore review recent traffic trends and suggest, in only broad strokes, what is plausible in the future.

One of the putative invariants of the Internet is the exponential growth in the traffic that it carries. Internet Protocol (IP) traffic, which represents the lion's share of Internet traffic, has doubled roughly every two years—an equivalent growth of 40% per year—since the year 2000, and similar growth is expected in the foreseeable future [6, 82]. Part of this exponential growth is attributable to increased penetration of Internet access in businesses and households. However, with penetration growing at approximately 20% per year [11], the remaining growth of IP traffic has arisen from more bandwidth-consuming applications. Indeed, traffic arising from P2P applications—file sharing, voice-over-IP (VOIP), streaming media, instant messaging, software publication and distribution, media publication and distribution—has carried much of the remaining growth of IP traffic over the last several years [292].

What can be expected of data traffic in the future? With respect to Internet traffic, it is safe to conclude that significant growth will occur from increased penetration of Internet access alone. Asia and Africa, for instance, which collectively represent 70% of the world's population, boast penetrations of only 11% and 4%, respectively, as of 2007 [11], leaving much room for growth. Beyond more connected human users, devices such as televisions, mobile telephones, and sensors, will increasingly be connected to the Internet, further swelling the volume of traffic carried by the Internet. Compounding more prevalent access to the Internet, more sophisticated, bandwidth-consuming applications are envisioned. In particular, a host of video applications on the horizon—video-on-demand (VOD) to the computer and television, and real-time video communications, for example—are expected to sustain much of the exponential growth in Internet traffic over the next several years (e.g., see Figure 1-2) [6]. In addition, grid computing, an emerging means of sharing distributed processing and data resources for academic (e.g., e-science) and commercial applications (e.g., data backup), will require very large data rates of Tbps for durations of seconds, minutes, or even hours. Beyond the Internet, the Global Information Grid (GIG) is responsible for fulfilling all of the communication needs of the US Department of Defense (DoD) in both wartime and peacetime [230]. In order to bring to fruition the military doctrine of Network-Centric Warfare, the GIG will need to support large volumes of traffic in an on-demand fashion so that an integrated view of future battlefields—by means of real-time data fusion, synchronization, and visualization within a "service-oriented architecture"—may be delivered to any operating point in the world [259]. Lastly, and perhaps most importantly, the recurring lesson of the past two decades has been

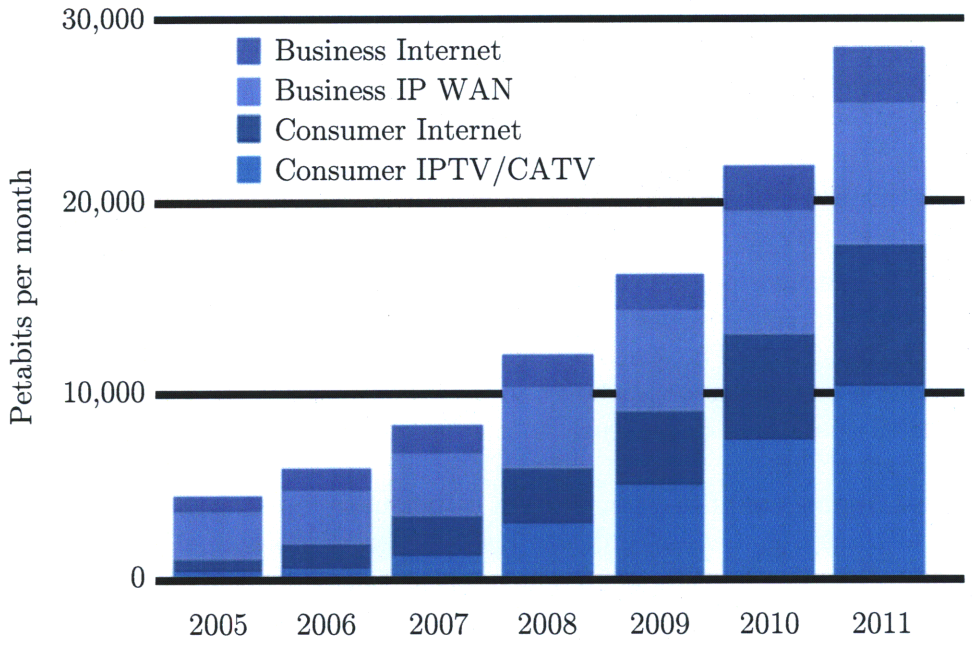


Figure 1-2. Historical trends and future projections of Internet traffic, according to Cisco Systems [6].

that the openness of the network architectures, such as the Internet, serves as fertile ground for the sprouting of unimagined applications.

Our discussion has thus far focused on the growth of the average, or first moment, of traffic. While an important metric for coarse dimensioning of network resources, the first moment of traffic suppresses several important properties of traffic that impact design and performance of networks. In particular, the following (user-imposed) properties of transactions weigh heavily on the design and performance of networks:

- The *size* of the transaction in bits sent. For a fixed volume of aggregate traffic, a larger proportion of smaller transactions is more conducive to smoothing of traffic—temporally and spatially—via statistical multiplexing, and hence better utilization of network resources.

- The maximum *delay* that is tolerable before a transaction's request for service is fulfilled. More stringent delay requirements obviously hamper efficient utilization of network resources.

- The *elasticity* of a transaction, or its tolerance to variations in transmission bit rate. Elasticity in transactions is desirable as it enables better utilization of network resources.

- The *multicast* or *unicast* nature of transactions. Multicast transactions offer potential for savings in network resource consumption compared with unicast transactions to the same number of destinations via: i) less redundancy of data along common links from a source to destinations, and ii) replication of identical sources along the edge of the WAN such that different surrogates are responsible for serving different MANs (e.g., Akamai [8]).

While recognizing the importance of these dimensions of traffic, we shall refrain from detailed forecasting along them because, i) per our previous comments, this is a difficult, if not impossible, undertaking; and ii) it is not strictly necessary, since the focus of this thesis is OFS, which retains its attractiveness as a network architecture under broad assumptions with respect to these traffic dimensions. We shall therefore focus our discussion of traffic such that it is pertinent to motivating our present consideration of OFS.

Several arguments of varying degrees of persuasiveness regarding the nature of future traffic may be made in order to justify a serious consideration of OFS. First, as discussed previously, there is broad consensus that video will contribute to future traffic growth [6]. VOD, as well as other elastic video content with transaction sizes on the order of many gigabytes, would be excellent candidate transactions for OFS. For example, OFS communication across a WAN could be used to update video content residing in content distribution nodes within MAN; and after customization of content (e.g., insertion of tailored advertisements), intra-MAN OFS communication could be used for communication between content distribution nodes and end-users. In addition, inelastic real-time video with required rates on the order of wavelength channel rates—should such applications be developed—would also be well-suited to OFS. Lastly, grid computing applications in the academic, commercial, and defense sectors, entailing massive transfers of data, usually in an elastic fashion, would also be suitable for OFS service. Thus, even though the fine contours of future applications may be difficult to predict, it is quite plausible that these applications will result in large transactions (suitable for service via OFS) being a prominent component of future traffic.

Beyond arguments which rely on visions of future applications, other arguments for large Internet transactions have relied on power law distributions[6], another putative invariant of Internet traffic, and of complex systems more generally. Empirical studies of data network traffic, which began in earnest in the early 1990s, yielded surprises that challenged traditional assumptions of data traffic being Poisson in nature. In particular, traffic in both LAN [185] and WAN [235] environments were shown to exhibit *self-similarity*. Self-similar traffic, unlike Poisson traffic which has a characteristic burst length that is smoothed out by averaging [316], exhibits long-range dependence and high burstiness over a wide range of time scales. The self-similar

[6]A nonnegative random variable X is said to have a power law distribution if $\Pr[X \geq x] \sim cx^{-\alpha}$ for constants $c > 0$ and $\alpha > 0$, where $f(x) \sim g(x)$ denotes that the ratio of $f(x)$ to $g(x)$ approaches unity as x grows large.

nature of traffic has important implications for network design, such as buffer design and admission control. Indeed, it has been shown that ignoring the self-similar nature of traffic in favor of traditional Poisson assumptions results in optimistic performance predictions [39, 227].

It has been widely acknowledged that the most probable origin of the self-similarity of traffic is the *heavy-tailed* nature of data transactions. That is, self-similar traffic arises when the distribution of data transactions exhibit heavy tails which decline like power laws with exponents close to unity [73, 288, 317][7]. Power law distributions, in fact, are known to permeate many complex systems arising in both Nature and human engineering [2], but the deep origin of this phenomenon is being vigorously debated[8]. In spite of the debate over why power law relationships exist in Internet traffic, and more broadly in Nature, there seems to be agreement on their expected future existence [88]—and, by extension, the future existence of a significant proportion of traffic that is amenable to service via OFS.

Note that, in addition to the aforementioned self-similarity of traffic that arises from the application layer and which exists at coarse time-scales (i.e., hundreds of milliseconds or larger), the recent explosion of WWW traffic—and its attendant protocols, such as Transport Control Protocol (TCP)—has created additional "small-time scaling" dynamics (i.e., hundreds of milliseconds or smaller) which are distinctly different [100]. In this thesis, we are more concerned with the coarse time-scale dynamics of traffic because: i) they are more intrinsic since they are not manifestations of network protocols, and ii) they influence network performance and architecture more significantly.

■ 1.1.2 Optical and electronic networking devices

Owing to the different natures of electrons and photons, electronic and optical networking devices can be quite different functionally; and even when functionally similar, may possess very different cost structures. For instance, optics cannot supplant electronics at network nodes—at least not in the foreseeable future—for a variety of reasons succinctly captured by Chan in [54]. First, viable and cost-effective optical technologies do not exist for critical router building blocks, such as random access memory (RAM). Second, when optical equivalents do exist, they are prohibitively

[7] Another mathematically legitimate, but less plausible, origin of self-similar traffic is a self-similar number of users engaging in transactions that are exponentially distributed [298].

[8] The statistical physics concept of Self-Organized Criticality (SOC), for instance, has been suggested as the mechanism by which complexity arises in many natural, and even engineered systems, with power laws as a byproduct [23]. More recently, the theory of Highly Optimized Tolerance (HOT) systems has been put forth as an alternative explanation for the ubiquity of power law relationships. Roughly speaking, power laws are one of the signatures of HOT systems which arise from robust design—either via explicit engineering planning, or mutation and natural selection—for a specific level of tolerance to uncertainty, which is traded off against the cost of compensating resources [2, 50, 51].

expensive[9], owing to the fact that the current cost of an optical logic gate is at least 10 orders of magnitude higher than that of an electronic logic gate. Third, the fundamental limit of minimum switching energy of an optical logic gate is on the order of $\overline{h}\nu$ (i.e., Planck constant multiplied by the electromagnetic frequency), which is significantly larger than the corresponding limit for electronics of k_BT (i.e., Boltzmann constant multiplied by the temperature). Thus, optics will generate more heat than electronics for the same number of logical operations. Given the large number of logical operations required of a router, heat management for optical routers will therefore be a significant issue.

Optical networking technology[10], however, does present several scalability advantages over electronics apart from transmission capacity [54]. Indeed, the impetus behind first-generation optical networks was the vast usable bandwidth of optical fiber in comparison to copper links. Second-generation optical networks, moreover, have exploited additional properties of optics over the last decade to dramatically further reduce the cost per transmitted per bit. For instance, the deployment of WDM technology has enabled high utilization of the fiber medium while only requiring that network devices operate at the peak wavelength channel capacity, which is matched to the maximum speed of electronics; the erbium-doped fiber amplifier (EDFA) has enabled simultaneous optical amplification of multiple wavelength channels; and the optical cross-connect (OXC), in conjunction with Generalized Multi-Protocol Label Switching (GMPLS), has enabled transparent optical bypass of expensive electronic routers. Currently, and in the near future, the attractiveness of optics as a broadcast medium is leading to simpler implementations of multicast, narrowcast, and multiple-access in the access environment [54].

The complementarity of electronics and optics—the ability for electronics to perform complex operations at fine granularity, and the ability of optics to perform simple operations en masse cost-effectively—leads us to conclude that, at least in the foreseeable future, economically viable network architectures will incorporate electronic and optical networking technologies. Indeed, such complementarity, with an increasingly prominent role awarded to optics, has been evident throughout the evolution of optical networks. In order to maintain the economic viability of optical networking in the face of growing bandwidth demands, further optical networking device innovation—coupled with architecture innovation—that exploits other attractive properties of optics will therefore be necessary. The OFS network architecture, which we describe next, is a logical next step in this evolution.

[9]For instance, the current cost of electronic time slot interchangers is at least 6 orders of magnitude lower than that of wavelength converters, their optical analog.

[10]For a detailed survey of optical networking technology, we refer the reader to Appendix A.

■ 1.2 Historical evolution towards OFS

In this section, we begin with a brief historical evolution of optical network architecture, focusing on the WAN, where most significant architecture developments in optical networking have occurred. The section culminates with an overview of OFS, which emerges as a promising next step in this story.

■ 1.2.1 Historical evolution of optical networking

In the days of first-generation optical networks, WAN transmission capacity was the most precious network resource. Optical networks serving bursty data traffic thus adopted the Electronic Packet Switching (EPS) architecture in lieu of the circuit-switched architecture used in telephone networks, since EPS utilizes this resource most efficiently. In these EPS networks, electronics-based routers, interconnected by optical fiber links, made routing and scheduling decisions in a distributed manner. Transactions were typically broken up into small IP packets; then placed into Point-to-Point Protocol (PPP) frames and/or Asynchronous Transfer Mode (ATM)[11] cells; and ultimately placed into SONET/SDH frames which operated directly above the optical fiber physical layer. As depicted in Figure 1-1(a), both the WAN and MAN assumed ring-based topologies, with aggregation of data into higher rate SONET/SDH streams occurring in the MAN rings. Access network and LAN architectures feeding into these rings were largely based upon Ethernet [1, 38, 104], Fiber Distributed Data Interface (FDDI) [254]; or in storage-area network (SAN) settings, Fibre Channel or Enterprise Serial Connection (ESCON) [70].

Traffic trends in the 1990s exposed the Achilles Heel of the EPS architecture to be its scalability: the difficulty, due to the complexity of electronic packet processing at routers, to keep apace with the exponential growth in bandwidth demand. Indeed, even if electronic processing advanced with Moore's Law, which is arguably optimistic owing to the super-linear complexity of switching and routing computation at routers, it would be unable to serve future traffic demands in an economically viable manner. Following the deployment of WDM technology in the WAN, second-generation optical network architecture innovation in the form of Multi-Protocol Label Switching (MPLS)[12], and subsequently GMPLS, arose to mitigate the looming crisis in EPS [289]. MPLS alleviated congestion at routers by enabling electronic circuit-like tunnels in the data link layer which bypassed computationally-intensive routing and switching operations at the network layer via table lookup. GMPLS generalized[13] this notion of tunnels to the optical domain, thus enabling optical bypass of routers

[11] ATM was initially conceived as a layer 2 and 3 replacement for IP offering QoS guarantees, but IP's ubiquity prevented this displacement. Consequently, ATM's (likely short-lived) niche is the layer 2 protocol underneath IP.

[12] MPLS is currently displacing ATM from its layer 2 niche.

[13] In fact, GMPLS also generalized the MPLS notion of tunnels to include layer 2 interfaces (e.g., ATM, Ethernet), and time-division multiplexing (TDM) interfaces (e.g., SONET/SDH) [289].

via OXCs at WAN network nodes. The benefit of MPLS and GMPLS has been the significant reduction of onerous electronic computation in the network core; but this has come at the expense of a somewhat increased computational burden at the ingress and egress of the WAN.

Second-generation optical networking has, meanwhile, occurred at a slower pace outside the WAN. In MANs, WDM technology has only recently been adopted in conjunction with traditional SONET/SDH ring topologies, thus adding the frequency domain to the time domain as a means of multiplexing data [214]. In residential and small/medium sized business access networks, hybrid fiber coax (HFC) architectures are being employed as interim architectures by cable service providers; whereas, the passive optical network (PON), which we shall return to in section 1.3.3, is viewed as a longer-term solution and is now beginning to be deployed in earnest. Ethernet, it should be noted, is still very much the de facto layer 1 and layer 2 standard in LANs, and serious efforts are being made to extend its reach to the MAN and WAN environments [4].

■ 1.2.2 Overview of OFS

In spite of technological and architectural advancements in the WAN which have led to significant reductions in the cost per transmitted bit, the end-user's access to the vast core network bandwidth has been restrained by the lag in advancements in the MAN and access network environments. In this section, we briefly introduce OFS as an end-to-end, all-optical transport service which provides end-users with cost-effective access to the core network bandwidth by means of exploiting the complementary strengths of optics and electronics [58, 59, 105, 111].

In a sense, OFS may be viewed as a generalization of the GMPLS concept of optical bypass of routers in the WAN to end-users. This generalization, however, presents a host of new challenges at multiple layers of network architecture, which are the focus of this thesis. Moreover, as we shall see, the extension of optical bypass renders OFS a sensible architecture only for users with large transactions; for otherwise, the network management burden and the cost of end-user equipment required in setting up an end-to-end, all-optical connection for a small transaction would outweigh the benefits of OFS. Thus, OFS is an architecture that is envisioned to serve end-users with large bandwidth demands exclusively, while end-users with comparatively small bandwidth demands will more efficiently be served by architectures such as EPS, MPLS or GMPLS.

In OFS, end-users request long duration end-to-end lightpaths by communicating, via an EPS control plane, with scheduling nodes assigned to their respective MANs. These scheduling nodes, in turn, coordinate transmission of data across the WAN in the EPS control plane. Owing to the intractable complexity of coordinating transmission among many OFS users, the control plane for OFS is envisioned to have a mixture of centralized processes which occur on coarse time scales, and distributed processes which occur on fine time scales. In particular, centrally computed candidate

routes for WAN reconfiguration will be disseminated to scheduling nodes on the order of seconds or minutes (or even longer). Thus, scheduling of individual transactions may occur in a distributed fashion with exchange of little information for coordination, and with physical layer reconfiguration times in the metro-area on the order of tens of milliseconds. The access environment, in contrast, is envisioned to have a broadcast structure, with access to it granted by the end-to-end scheduling algorithm.

In OFS, it is assumed that the smallest granularity of bandwidth that can be reserved across the core is a wavelength[14]. Motivated by the minimization of network management and switch complexity in the network core, transactions are served as indivisible entities. That is, data cells comprising a transaction traverse the network contiguously in time, along the same wavelength channel (assuming no wavelength conversion), and along the same spatial network path. This is in contrast to EPS, where transactions are broken up into cells, and these cells are switched and routed through the network independently. Note that in OFS networks, unlike packet switched networks, all queuing of data (in addition to admission control) occurs at the end-users, thereby obviating the need for buffering in the network core. A core node is thus equipped with a bufferless OXC. The elimination of buffering at OFS nodes presents a significant scalability advantage, since the queueing subsystem of EPS routers is becoming their major bottleneck as the number of ports and line-rates increase [177]. However, this absence of buffering, coupled with the requirement of serving transactions as indivisible blocks of data, renders the efficient utilization of network resources more difficult. Mixing transactions with different QoS requirements may provide significant advantages in this regard.

Summary from end-user and network operator perspectives

In closing this section, we summarize the salient features of OFS from both an end-user perspective and a network operator perspective.

OFS entails the following from an end-user's perspective:

- Payload transmission as a contiguous block of data with arbitrary format.
- Congestion and flow control via admission control and buffering at the end-user terminal.
- Delay arising from reservation of network resources (at least one round-trip time), and possibly from queueing behind other transactions.
- Substantially lower cost than current architectures (i.e., EPS, GMPLS) for sufficiently heavy network traffic.

and the following from a network operator's perspective:

[14]In the event that several single users have transactions which are not sufficiently large to warrant their own wavelength channels, they may multiplex their data for transmission across the WAN via dynamic broadcast group formation, albeit at the delay of aggregating these transactions.

- Network management and control via an EPS control plane.
- Quasi-static logical topology in the WAN for decoupling of slow centralized network management and fast distributed control.
- Agile optical switching in the MAN.
- Aggregation in the access environment via broadcast and reservation/scheduling.
- Mixing of transactions with different QoS to achieve efficient utilization of network resources.

1.3 Alternative architectures for the future

In this section, we present alternative candidate architectures to OFS for future optical networks in the WAN, MAN, and access environments. In spite of the organization of this section, we emphasize that these three hierarchical network tiers are not independent and, in fact, as is the case with OFS, can be rather tightly coupled. Lastly, we note that the survey presented in this section is not exhaustive, but represents a collection of architectures which are prominently represented in the literature.

1.3.1 WANs

As discussed earlier, the primary function of a WAN is data transport over distances as long as thousands of kilometers. Owing to the appreciable signal degradation inherent in long-haul transport over fiber, expensive apparatus such as amplifiers, regenerators, and dispersion compensation equipment is required along WAN fiber spans. The primary driver behind WAN architectures, such as those presented in this subsection, is therefore achieving high utilization of fiber links as a means of minimizing the high cost of transport.

Optical Burst Switching and Tell-and-Go

Optical Burst Switching (OBS), like OFS, is an architecture which attempts to draw upon the strengths of electronics and optics by employing an electronic control plane that is separated from the optical data plane in both time and space. In OBS, packets are assembled at WAN ingress nodes, as in MPLS and GMPLS, according to destination and QoS requirements, to form longer bursts. Prior to transmission of a burst, a control packet is sent into the control plane to request, at each intermediate node, that an all-optical path be set up for the ensuing burst. Without any acknowledgement regarding the success of the control packet in reserving a path, the burst is then transmitted into the core on a wavelength channel, reaching its destination egress node if the necessary resources at intermediate nodes are available. Failure to reach

the destination egress node arises when a burst contends with another burst for resources at one or more of the intermediate nodes, causing the burst to be discarded[15], since intermediate nodes use OXCs which have no buffering capability.

The above description forms the essence of the OBS architecture, although there is a great deal of variability in how OBS networks are designed[16]. While most of the implementations of OBS aggregate bursts at the edge of the WAN, one variation of OBS, analyzed subsequently in this thesis, entails end-users acting as sources of data bursts [294]. We refer to this variation as Tell-and-Go (TaG) since it is based upon the TaG protocol for ATM networks [313], in which end-users act as sources for bursts of data. Another variation, Wavelength-Routed OBS [173], is essentially traditional Optical Circuit Switching (OCS)—WAN ingress routers assemble packets into bursts, and centralized, advanced scheduling sets up lightpaths for these bursts. We emphasize that the distinction between OFS and these various OBS architectures is that in OFS all-optical connections are set up in a *scheduled* fashion between *end-users* rather than between WAN nodes.

Optical Packet Switching

The Optical Packet Switching (OPS) architecture is similar to the EPS architecture, with the important exception that data is processed in the optical domain rather than in the electronic domain. The benefit of OPS is that, if implemented, it would yield maximum utilization[17] of fiber links without the burden of costly optical-electronic-optical (OEO) conversions. Formidable obstacles, however, exist in the realization of OPS networks, which stem from the infancy of the required optical technology [329]. Appendix A reviews the status of optical buffering and optical logic technology which would be required for the implementation of optical routers.

Hybrid architectures

Hybrid optical network architectures—architectures comprising two or more stand-alone optical network architectures (e.g., OBS and OCS)—for WANs have been proposed as a means of accommodating widely varying service requirements of users [3, 53, 63, 68, 113, 249, 300]. Hybrid architectures may be classified according to how tightly integrated the operations of its constituent subarchitectures are. Earlier proposed hybrid architectures include Multiwavelength Optical Networking (MONET) [307], and the current WAN standard, GMPLS.

At one extreme are hybrid architectures whose component subarchitectures operate in parallel with little interaction, as they are granted their own statically dedicated network resources. In the middle of the taxonomical spectrum are architectures

[15]In some variations of OBS, burst discardment is a last resort should other ways of handling contending bursts, such as segmentation [305] or deflection routing [62], fail.

[16]For a survey of alternatives, we refer the reader to [26,27,245,304,325] and the references therein.

[17]We shall elaborate on the capacity performance of WAN architectures in Chapter 5.

whose network resources are shared among component subarchitectures, but the reallocation of resources occurs on coarse time scales (i.e., hundreds of milliseconds or longer). For instance, the hybrid architectures proposed in [182, 208, 320, 328] combine OBS and OCS and share network resources via a unified control plane. Finally, there are hybrid architectures whose component subarchitectures are intimately integrated in that allocation of resources occurs on a packet by packet basis. Examples are Overspill Routing in Optical Networks (ORION) [63, 68, 113, 300], OpMiGua [34], and SLIP-IN [249], all of which permit packets and wavelength circuits to coexist on wavelength channels. While architectures that more intimately integrate component subarchitectures are capable of achieving better resource utilization, they incur significantly more control plane and hardware complexity relative to architectures exhibiting less integrated subarchitectures [113].

Other architectures

Light-trails is an IP over WDM architecture that treats an end-to-end lightpath similarly to a bus, in that intermediate nodes, each comprising a drop coupler, optical shutter, and add coupler in serial, can: i) copy incoming data while also allowing it to continue on to the next node (i.e., multicast), or ii) drop incoming data and replace it with new data to continue on to the next node [126, 128]. Owing to its ability to gracefully handle subwavelength communication between nodes on linear topologies, the Light-trails concept is particularly well-suited to MAN and LAN environments [129] which are often based upon buses or rings. Moreover, Light-trails does not suffer from the same drawbacks (e.g., fairness issues) as most previous architectures based on linear topologies as a result of more sophisticated medium access control (MAC) protocols [193]. We presented the architecture in this section, however, because the Light-trails concept has been extended to mesh topologies for use within WANs [127, 130].

The Time-domain Wavelength Interleaved Network (TWIN) architecture, proposed in [314], was motivated by the belief that hybrid architectures—and even architectures employing switching in more than one domain (i.e., GMPLS)—are unnecessarily expensive, and that a single, unified transport architecture is preferable. Transport in the TWIN architecture occurs via TDM over WDM, and is implemented with special OXCs in the WAN core working in concert with fast, tunable lasers residing at the WAN edge. In TWIN, each WAN egress node is assigned a wavelength, and data from several WAN ingress nodes wishing to transmit to the same egress node are interleaved onto the egress node's wavelength and routed via a multipoint-to-point tree. Thus, in contrast to typical OXC operation, the OXCs used in TWIN are required to merge incoming signals of the same wavelength onto the same outgoing wavelength. The switching time of these OXCs is not required to be fast, since these devices are reconfigured only when the multipoint-to-point trees are reset. Nevertheless, the TWIN architecture will face substantial challenges in its implementation, owing to the high precision and speed of tunability required of the lasers residing at

the WAN edge, as well as the complications in scheduling arising from propagation delays [255] and reuse of wavelength channels [228] in the core.

■ 1.3.2 MANs

Recall that the MAN serves as the connective tissue between the WAN and end-users residing in access networks. As the interface between access networks, which generate bursty traffic of various bit rates and formats, and the WAN, for which link capacity is the precious resource, the primary function of future MANs is to perform aggregation/disaggregation in a cost-effective manner.

Ring-based architectures

As discussed in section 1.2, traditional MAN architectures—and even ones presently being deployed which employ WDM—are based upon SONET/SDH ring topologies. Recognizing these topologies as the point from which future architectures will migrate, most of the proposed next-generation MAN architectures are based upon ring physical topologies. Moreover, ring topologies have long been recognized as conferring improved network management and control, such as via simpler routing and failure protection, as compared to other topologies [183].

Some near-future proposals for ring-based WDM MANs recognize the continued demand for SONET/SDH gear, and therefore employ a "next-generation" form of SONET/SDH that utilizes generic framing protocols for mapping several different protocols used in LANs [214]. Interconnection among these WDM rings can be achieved with static optical add-drop multiplexers (OADMs), or dynamic reconfigurable optical add-drop multiplexers (ROADMs)/OXCs [214], which lend to different traffic grooming techniques [308,309] and MAC protocols [67].

Some longer-term proposals maintain a SONET/SDH lower layer, but employ packet-based framing above this layer. Resilient Packet Ring (RPR)[18], for instance, is a packet-based standard that employs layer 2 Ethernet technology atop SONET/SDH with the added capability of guaranteeing three classes of QoS [9]. Other packet-based architectures, such as Hybrid Opto-electronic Ring Network (HORNET), eliminate SONET/SDH altogether, and transport IP packets over a WDM physical layer. This architecture employs OADMs at nodes for optical bypass of some transiting wavelengths, but also requires OEO conversions at nodes for packets on dropped wavelengths that are required to continue along the ring to their destination nodes. The Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) MAC protocol is employed in HORNET to arbitrate transmission of packets on wavelength channels. Other IP over WDM approaches employ complex token-based MAC protocols as a means to avoid costly OEO conversions (e.g., Multitoken Interarrival Time (MTIT) [42], and the protocols in [107,147]).

[18]Dynamic Packet Transport (DPT) is a similar competing architecture whose primary difference is the MAC protocol it employs [84, Chapter 7].

The Next Generation Internet Optical Network for Regional Access using Multi-wavelength Protocols (NGI ONRAMP) architecture is another ring-based architecture employing IP over WDM [105, 179], although its design is easily extendable to other topologies. In this architecture, MAN nodes each comprise an IP router and an OXC. The IP router is responsible for the aggregation/disaggregation of data from/to access networks, as well as for routing of data transiting along the MAN ring. The OXC enables bypass of the node's router and allows for wavelength channels to be transparently added/dropped, thus providing support for heterogeneous traffic formats. As OFS was first proposed by the NGI ONRAMP consortium, the OXCs at MAN nodes also provide the optical bypass necessary for OFS operation.

More aggressive architecture proposals include a MAN version of the aforementioned ORION architecture in which OCS and OPS are merged on common wavelength channels [226]. Such architectures are arguably impractical in that they involve transparent insertion and removal of data on wavelength channels on fine time-scales.

Architectures based upon other topologies

In the 1990s, the All-Optical Network (AON) consortium unveiled one of the first expansive all-optical network architectures [18, 160]. The architecture, which encompasses access network, MAN, and WAN environments is based upon a tree topology[19]. Each of the nodes in the MAN tier of the AON tree topology employ either an arrayed waveguide grating (AWG) or OXC for wavelength routing of data between the WAN and LANs, or among LANs. In addition, a passive star coupler (PSC) and a passive splitter/combiner positioned in parallel with each node's AWG or OXC enables multicast service.

In [196, 197], a star topology employing an AWG hub is proposed, in which reconfigurability of lightpaths is achieved by tunable transceivers or arrays of fixed transceivers. Owing to the poor survivability of star topologies—a single failure at the hub entails complete failure of the network—the previous architecture was modified to yield two alternative architectures. In the first alternative, a PSC is used in parallel with the AWG [99]. Normal operation involves both devices operating in concert to enhance throughput; and, in the event that either the AWG or PSC fails, the other device still provides full connectivity among nodes. In the second alternative, the AWG-based star network is embedded[20] within a ring topology [197]. Under normal operation, the AWG and the ring are used to efficiently route lightpaths for a larger number of nodes; and in the event of one or more, node or link failures, the network is able to maintain full connectivity.

Outside of MAN topologies based upon the canonical ring, star, and tree families, the present author has investigated meshed physical topologies where resilience to

[19]Tree topologies are not as resilient to failures as ring topologies, but they are particulary efficient at routing data among nodes, particularly when most data is destined for the root node (i.e., the WAN).

[20]In particular, only a subset of nodes on the ring are attached to the AWG hub.

multiple failures is of primary concern [312]. In this work, Circulants were identified as a rich family of graphs with desirable reliability properties. Lastly, Guan has investigated MAN physical topology design using a cost model comprising realistic switching and fiber plant costs. He has shown that a family of graphs, known as Generalized Moore Graphs, minimizes network cost under uniform intra-MAN traffic [121, 123, 124]. See Appendix B for a discussion of (Generalized) Moore Graphs.

■ 1.3.3 Access networks

An access network is the end-user's means of connecting to the MAN and WAN, and usually supports up to several hundred end-users within a geographic span of up to twenty kilometers. While the boundary between the MAN and access network is not well-defined—and may even be disappearing—access networks are viewed today as the network connecting a service provider's facility and the end-user.

Research on optical access networks has occurred in roughly two waves. The first wave of research occurred in the 1980s and early 1990s, coinciding with other research on WDM technology. Many interesting and far-reaching ideas (for the time) were put forth, but none of the proposed systems were ever implemented outside R&D test-beds, if even there. Following a decade-long lull, PONs began to be deployed in commercial settings, which rekindled research interest in optical access networks. The result has been a second wave of near-term research that has proceeded in rough lock-step with deployed commercial systems.

First wave of optical access network research

An initial wave of research on access network architectures employing WDM and other optical networking technology began nearly two decades ago. Owing to the small number of end-users residing in an access network, proposed architectures generally employed passive optical devices—which have low capital expenditure (CapEx) and operating expenditure (OpEx) costs[21] compared with active components—in simple broadcast topologies [57, 61, 66, 87, 133, 155, 157, 205, 212, 213, 277, 278]. Most architectures are based upon a star topology, though significant variation exists in the number and type (i.e., fixed-tuned, tunable, array) of end-user transceivers, and the MAC protocol employed. Some architectures achieve simple and inexpensive implementations by employing unslotted random-access MAC protocols (e.g., Aloha) and a small number of transceivers at end-users, but suffer from poor throughput and delay performance owing to channel and receiver collisions [61, 205]. Other architectures achieve better performance, but at the expense of complex, often slotted, reservation schemes and require multiple, tunable transceivers at end-users [66, 133, 155, 157]. Yet other architectures achieve good performance with a simple MAC protocol and a minimal number of fixed-tuned transceivers, but exhibit limited scalability in number of

[21] Network costs are generally partitioned into CapEx and OpEx, which represent the capital cost of equipment and the operating cost of running the network, respectively.

supported end-users because wavelength channels are dedicated to end-users [87,277]. Bus topologies have also been proposed as the basis for next-generation optical access networks [24, 139, 215, 310, 311]. While this family of topologies uses less fiber to connect end-users than a star, it is plagued by fairness issues of user bandwidth allocation. Thus, most bus-based architectures require complex MAC protocols to ensure fairness. Lastly, some proposed optical access network architectures incorporate optical amplification with otherwise passive components in order to enhance the reach and number of supported users [57].

Second wave of optical access network research

While PONs have been discussed in the literature for two decades, and have even been in existence in scattered settings[22], insufficient bandwidth demand coupled with prohibitive deployment costs have not rendered them economically viable until recently. The generic architecture of a present-day PON is illustrated in Figure 1-3. All-optical transmission is carried out on optical fiber between an optical line terminal (OLT) residing at a service provider's central office and an optical network unit (ONU) which resides close to or at an end-user. Although Figure 1-3 depicts the OLT connecting to the ONUs via a star topology, arbitrary topologies are in principle possible. In fact, the single key feature of PONs is that *passive* optical networking devices (e.g., PSCs) are employed at remote nodes in order to connect the OLT to the ONUs.

Several variations exist within the family of PONs currently deployed, which are distinguished by the way data is framed in layers 1 and 2 and by the MAC protocol employed for accessing the shared upstream fiber medium. The remote nodes in present-day PONs comprise a PSC. Early PON standards[23] employed downstream broadcast from the OLT on a single wavelength channel, whereas upstream transmission occurred on a different wavelength on the same fiber, or on the same wavelength on a different fiber, and employed some form of TDM to avoid channel collisions among ONUs. PSCs at remote nodes with splitting ratios anywhere from 16 to 128 were prescribed in conjunction with geographic spans of 10-20 km. Aggregate supported bit rates began at 155 Mbps with TPON, and are now as high as 2.5 Gbps for GPON. Lastly, in these PONs, a relatively inexpensive Fabry-Perot laser or light-emitting diode (LED) and an inexpensive uncooled pinFET receiver are employed at OLTs and ONUs.

In the future, these deployed PON architectures are expected to migrate to WDM as a means of dramatically increasing the aggregate bandwidth of earlier PONs. For instance, the next-generation Ethernet PON, 10 Gigabit Ethernet Passive Optical Network (10GEPON), will have WDM capability built in [7]. While some proposed

[22]British Telecom, for instance, deployed early versions of PONs in the 1990s [250].

[23]Examples include: Passive Optical Network for Telephony (TPON) [274], Asynchronous Transfer Mode Passive Optical Network (APON) [195], Broadband Passive Optical Network (BPON) [296], Ethernet Passive Optical Network (EPON) [174], Gigabit Ethernet Passive Optical Network (GEPON) [10], and most recently Gigabit Passive Optical Network (GPON) [52].

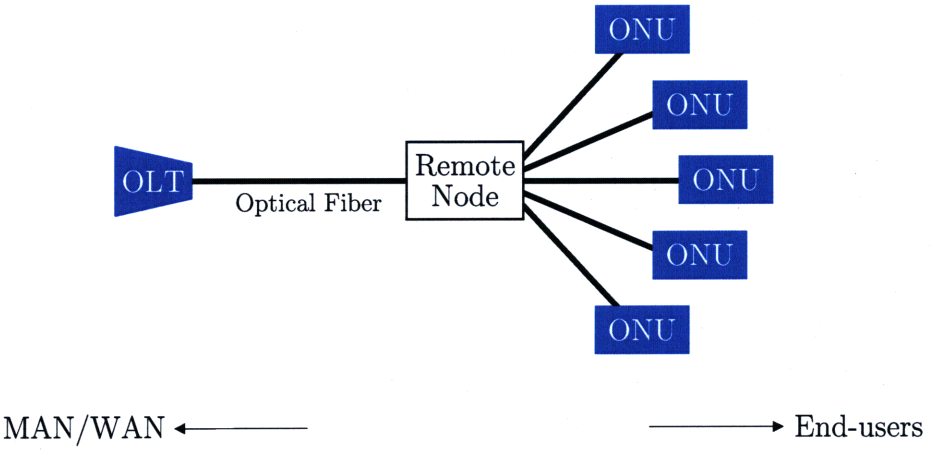


Figure 1-3. Generic PON architecture being implemented today. Only passive optical networking device(s) are deployed in the remote node.

WDM PONs use a PSC at the remote node as in the case of earlier PONs [94,120,297], most proposals employ an AWG at the remote node to carry out wavelength routing [35,83,94,95,120,250,266,327], which enables point-to-point dedicated services to be provided to ONUs [95,250]. WDM PONs are expected to suffer from higher capital costs stemming from the need for multiple, high-quality transceivers at the OLT, and possibly at the ONUs. Low cost WDM sources are therefore critical to the future economic viability of these networks.

Recently, optical code-division multiplexing (OCDM) has been proposed as a multiplexing technique for PONs [15, 132]. While such PONs are attractive by virtue of their extended reach (~100 km) and security [132, 165], front-end optical encoders/decoders are expensive and the number of supported users is limited by interference and noise [165].

Lastly, optical amplification, which represents a break from the passive paradigm of deployed PONs, is being considered as a means of extending the reach and the number of supported end-users. Such architectures, which would employ optical amplification at remote nodes with either semiconductor optical amplifiers (SOAs) or EDFAs, are expected to result in cost savings through a reduction in the number of OLTs, notwithstanding the additional CapEx and OpEx costs of amplifiers [57, 77,78,114,115,252,273,286,291]. Such amplified PONs, which are expected to have geographic spans of 100 km and split ratios of up to 1000, would effectively consolidate the access network and MAN environments [77,188,251].

1.4 Notes on this thesis

OFS as a high-level concept has been in existence for approximately two decades [56], and first appeared in the literature almost a decade ago [58]. Nevertheless, the progression from high-level concept to detailed architecture, analysis, and proof-of-concept had not begun in earnest until very recently. In [110], B. Ganguly reports on the first OFS test-bed, implemented in the Boston area. An investigation into the throughput-delay performance of OFS in the metro environment is also reported in this work, and lends strong support to the use of scheduling in OFS. In [109], A. Ganguly addresses a special form of OFS particularly well-suited for critical defense applications. In this work, flows are assumed to possess such strict delay requirements that scheduling of data and/or waiting for the dissemination of global network state information is unacceptable. Fast optical probing techniques are therefore used to query the availability of, and subsequently reserve, network resources for these OFS flows.

This thesis complements these aforementioned studies by developing and assessing the OFS concept in several new directions. Our contributions, in particular, are in providing partial answers to the following three broad questions:

- How should OFS be implemented?
- How does OFS perform?
- What are the economic properties of OFS?

Our contributions in answering these three questions are captured in Table 1.1; and we organize our contributions according to chapter in the following subsection.

1.4.1 Thesis overview

In accordance with the discussion in section 1.2.2, the design and analysis of OFS networks in this thesis reflect a necessary loose coupling between the WAN environment and the MAN and access environments. Consequently, the organization of this thesis reflects a structure in which the WAN and MAN/access are treated in separate chapters.

In Chapter 2, we address OFS in the context of the wide-area via a comparative analysis of network capacity with other prominent transport architectures. An important byproduct of this analysis is a capacity-achieving scheduling algorithm appropriate for intra-MAN OFS communication. Chapter 2 subsumes the following works:

- G. Weichenberg, V. W. S. Chan, and M. Médard, "On the capacity of optical networks: A framework for comparing different transport architectures," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, Barcelona, Spain, April 2006.

Implementation of OFS
• MAN physical layer design (sections 3.1–3.3, 5.4) *Optimization of Generalized Moore Graphs to minimize network cost under shortest path routing with uniform intra- and inter-MAN traffic* • Access network physical layer design (sections 3.4–3.6, 5.5) *Design of internally amplified distribution networks that respect OFS's physical layer constraints while supporting many end-users for efficient statistical multiplexing* • Intra-MAN scheduling algorithm (section 2.1.2) *Family of online scheduling algorithms with tunable throughput-delay tradeoff* • Inter-MAN scheduling algorithm (section 4.2) *Online low-complexity scheduling algorithm employing sequential reservation with small performance penalty for large networks*
Performance of OFS
• Comparative capacity analysis (Chapter 2) *Characterization of OFS capacity region, and comparison with other architectures; uniform all-to-all throughput comparison for special topologies* • Throughput-delay tradeoff for inter-MAN communication (sections 4.3–4.4) *Analytical approximation of throughput-delay tradeoff with above inter-MAN scheduling algorithm; exploration of OFS design parameter tradeoffs*
Economics of OFS
• Parametric CapEx cost model (sections 5.1–5.5) *Formulation of cost model for all geographic network tiers which captures major sources of CapEx* • Global physical layer design (section 5.6) *End-to-end cost optimization of physical layer for OFS and other architectures* • Throughput-cost comparison of architectures (sections 5.6–5.8) *Throughput-CapEx cost comparison of OFS with other architectures assuming homogeneous and hybrid architectures, and optimized physical layer*

Table 1.1. Thesis contributions.

- G. Weichenberg, V. W. S. Chan, and M. Médard, "On the capacity of optical networks: A framework for comparing different transport architectures," *Optical Communications and Networking Series of the IEEE Journal on Selected Areas in Communications (JSAC-OCN)*, vol. 25, no. 6, pp. 84–101, 2007.

In Chapter 3, we address the physical layer design of the OFS data plane in the metro and access environments. In addition to deriving high-level physical layer constraints, we propose and analyze alternative families of distribution networks (DNs) that differ in where optical amplification of signals is carried out. Chapter 3 subsumes the following works:

- G. Weichenberg, V. W. S. Chan, and M. Médard, "Access network design for Optical Flow Switching," *Proceedings of IEEE Global Telecommunications Conference (Globecom)*, Washington D.C. November 2007.

In Chapter 4, we begin by outlining a simple and sensible scheduling algorithm for OFS networks. We then proceed with an approximate throughput and delay performance analysis of OFS networks. Chapter 4 subsumes the following work:

- G. Weichenberg, V. W. S. Chan, and M. Médard, "Performance Analysis of Optical Flow Switching," submitted to IEEE International Conference on Communications (ICC), Dresden, Germany, June 2009.

In Chapter 5, we formally introduce the notion of cost via an approximate CapEx model. This enables us to carry out a throughput-cost comparison of OFS with other prominent candidate architectures—OPS/EPS and OBS/TaG—in the context of both homogeneous and hybrid architectures. Chapter 5 subsumes the following works:

- G. Weichenberg, V. W. S. Chan, and M. Médard, "Cost-efficient optical network architectures," *Proceedings of European Conference on Optical Communications (ECOC)*, Cannes, France, September 2006.

- G. Weichenberg, V. W. S. Chan, and M. Médard, "On the throughput-cost tradeoff of multi-tiered optical network architectures," *Proceedings of Globecom*, San Francisco, November 2006.

- G. Weichenberg, V. W. S. Chan, and M. Médard, "Access network design for Optical Flow Switching," *Proceedings of Globecom*, Washington D.C. November 2007.

- G. Weichenberg, V. W. S. Chan, E. A. Swanson, and M. Médard, "Throughput-Cost Analysis of Optical Flow Switching," submitted to IEEE/OSA Optical Fiber Communication Conference (OFC), San Diego, March 2009.

We conclude this thesis in Chapter 6. In this chapter, we provide a summary of contributions of this thesis and outline directions for future research within the theme of OFS networks.

In Appendix A, we survey building blocks for future optical networks. Much of the content of this thesis relies upon a high-level understanding of these building blocks. Appendix A subsumes the following works:

- E. Yamazaki, G. Weichenberg, A. Takada, and T. Morioka, "Polarisation-insensitive parametric wavelength conversion without tunable filters for converted light extraction," *IEEE Electronics Letters*, vol. 42, no. 6, pp. 365–367, 2006.

- E. Yamazaki, F. Inuzuka, G. Weichenberg, A. Takada, J. Yamawaku, T. Morioka, and M. Koga, "Waveband path virtual concatenation with contention resolution provided by transparent waveband conversion using QPM-LN waveguides," *IEIC Technical Report*, vol. 106, no. 69, pp. 55–60, 2006.

In Appendix B, we survey a family of graphs, Moore Graphs, and their superset, Generalized Moore Graphs. As will be discussed in this appendix, Moore Graphs and Generalized Moore Graphs serve as attractive candidate physical topologies for MANs.

Chapter 2

OFS in the Wide-Area: A Comparative Capacity Analysis

In this chapter, we consider OFS in the context of the wide-area. As discussed earlier, optical networking technology for WANs has advanced steadily and, in fact, many technologies developed a decade ago have yet to be deployed commercially. The optical networking components and systems required for the realization of OFS are therefore well within the reach of present-day and near-term technology, and we thus do not address such issues in any detail. The physical topology of the WAN is another element of network design beyond the scope of this thesis, since predetermined geographic and right-of-way constraints curtail much of the control of the network designer in this area of planning.

The focus of this chapter is therefore on the higher-layer issue of resource provisioning and access. In particular, we turn our attention in this chapter to the capacity or throughput performance limits of OFS networks as well as the means by which to attain them. We define network capacity as the set of exogenous traffic rates that can be stably supported by a network under its operational constraints. By operational constraints of a network, we mean the set of capabilities and rules followed by the network, such as the presence of core buffering, wavelength conversion capability, and the nature of the scheduling policy employed. Along with our results for OFS networks, we address the capacity of two prominent transport architectures—EPS/OPS and OBS—in order to obtain benchmarks by which to assess the performance of OFS networks. Each transport architecture's physical and operational properties impose constraints on its capability for logical topology reconfiguration, and naturally lead to different performance regimes.

Our metric of network capacity is particularly relevant to core networks because, owing to the high cost of supporting transport traffic, capacity is a precious commodity in the core (but not necessarily in the access network). We recognize, however, that while the question of the capacity limits of a network is important, it is not the only performance criterion by which the network should be assessed. Delay, of course, is a performance metric that is just as important as capacity. However, owing to the fact that any queuing delay in OFS networks occurs at the transmitting end-user, the issue of delay cannot be addressed until we consider MAN and access network

design for OFS networks in Chapter 3. Ultimately, a network should not be judged on performance alone, but rather on the performance-cost trade-off it presents to the end-user. Thus, the most useful comparison would include a detailed complexity/cost model for each of the candidate networks—the subject of Chapter 5.

Our approach to characterizing the capacity of optical networks differs from those of preceding works in various respects. Other works have applied the matrix decomposition results of Birkhoff [33] and von Neumann [306], traditionally used to analyze switches, to networks [41, 60, 314]. Inherent in such approaches are two limiting assumptions. First, the network, when viewed as a switch, is nonblocking. Although the underlying physical topology of a network is rarely a complete graph (i.e., a graph in which each node shares an edge with every other node), these approaches achieve nonblocking logical topologies by assuming sufficiently many active wavelengths in a fiber. Second, switching of data in the network is cell-based, even though data transactions are naturally variable in length. These cell-based schemes thus either employ framing or segmentation and reassembly of transactions which result in additional overhead, and possibly larger transaction delays. Our work is general in that arbitrary numbers of wavelengths are considered. Furthermore, data transactions are treated as indivisible entities in OFS and OBS, which is motivated from the fact that these two transport mechanisms consume resources for transmission set-up. Furthermore, we investigate the relationship among the capacity regions of the optical network architectures as a function of the number of switch ports per fiber in core network nodes. Finally, our approach to characterizing the capacity region is constructive in that online, capacity-achieving scheduling policies are outlined. It should be noted that most of the network capacity results in the literature—including some of the results in the present work—fall within the framework introduced by Tassiulas and Ephremides in their seminal work [290].

This work invokes several results from switching and networking theory, the background for which will be presented as required in the following sections. In section 2.1, we formally define network capacity and characterize capacity regions of OFS, as well as EPS/OPS and OBS. In section 2.2, we apply these results to two important network topologies: bidirectional rings and Moore Graphs. In section 2.3, we investigate the relationship among the capacity regions of the architectures as a function of the number of switch ports per fiber in core nodes. We conclude this chapter in section 2.4.

■ 2.1 Capacity of optical transport architectures

We model networks as directed graphs, where graph arcs and vertices represent directed fiber links and network nodes, respectively. Each fiber can support a maximum of t unit capacity active wavelength channels, and we assume that each node is equipped with t transceivers per fiber, one for each wavelength channel. We assume that each of these active wavelength channels carries data which is aggregated from

the end-users associated with a particular WAN ingress node. For example, a WAN ingress node that has f_{out} outgoing fibers can support a maximum of $f_{\text{out}}t$ wavelengths of traffic. Thus, $f_{\text{out}}t \geq N\rho_u$, where N is the number of end-users associated with the WAN ingress node, and ρ_u is each end-user's duty cycle (i.e., the fraction of time that an end-user has enough data to occupy a wavelength channel). We further assume that at each node there exist, in addition to dedicated transceivers, any other processing equipment (depending upon the network architecture) that may be required to support each active wavelength channel. We assume no wavelength conversion capability in our networks, which allows for the active wavelength channels to be decoupled and considered independently in the forthcoming capacity analyses. Thus, it suffices to examine only one of these channels in isolation[1].

In the remainder of this work, we assume that time is slotted and we neglect propagation delay. Furthermore, we only consider the case of unicast transactions. We associate exactly one transaction type with each source-destination node pair[2], and we denote the number of such transaction types by F. Let $A_n(i)$ denote the cumulative number of exogenous cell arrivals to the network of transaction type i by time slot n. As in [200], we assume that the arrival process $\{A_n(i)\}_{n=1}^{\infty}$ for each i satisfies the Strong Law of Large Numbers (SLLN) and is stationary. Let vector Λ denote the set of exogenous rates:

$$\lambda_i = \lim_{n\to\infty} \frac{A_n(i)}{n}.$$

Note that we use the terms *rate*, *traffic rate*, *flow rate*, and *burst rate* interchangeably in this work, depending upon the context.

The following definition relates to how transaction types traverse the network.

Definition 2.1 (Routing, Simple routing) *For each transaction type, we list all paths in the network from the corresponding source to the corresponding destination. A routing is this list of transaction types and possible paths, along with a set of probabilities that transactions follow these paths in the network. If each transaction type is restricted to follow exactly one path from source to destination, then the routing is a simple routing.*

In the above definition, we assume that transactions are routed independently of one another.

We associate with each of the F transaction types a queue where transactions of that type enqueue upon entry to the network. Each such queue resides at a network node—the source node corresponding to the transaction type. Note that these queues may be virtual queues in the sense that they may not physically exist—as in the case

[1]Clearly, this is not true for a delay analysis, as several servers working together can achieve lower expected delay than several servers working independently under the same traffic intensity.

[2]We assume this for simplicity. It is straightforward to generalize the results in this chapter to an arbitrary number of transaction types associated with each source-destination pair.

of OFS in which transactions enqueue at end-users rather than WAN ingress nodes—but are convenient mathematical constructs. Also, at each node along a path that a transaction type may follow toward its destination, there may exist a queue dedicated to that transaction type and path. For example, let us assume that in an EPS/OPS network a transaction of a certain type may follow one of k paths, each of length h_i where $i = 1, \ldots, k$, toward its destination. Then, regardless of whether the paths overlap in the network, there exist $1 + \sum_{i=1}^{k}(h_i - 1)$ queues in the network dedicated to the transaction type. We shall denote the total number of queues in the network by Q, and each queue is identified by a unique index from $1, \ldots, Q$. Note that in an OFS network, since all transactions traverse the network in a single hop, there exist exactly $Q = F$ queues in the network, one for each transaction type at its source node.

Let X_n be a Q-dimensional vector whose i^{th} element represents the number of data transactions in the i^{th} queue at time slot n. Likewise, let the i^{th} element of the Q-dimensional vectors D_n and E_n represent the number of departures from and entrances (exogenous or endogenous) to the i^{th} queue at time slot n, respectively. Hence, $X_{n+1} = X_n + E_n - D_n$.

Definition 2.2 (Rate-stability) *A system of queues is rate-stable if:*

$$\lim_{n\to\infty} \frac{X_n}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} (E_i - D_i) = 0, \text{ with probability 1.}$$

Remark: This is a relatively weak form of stability. For a discussion of other, stronger forms of stability, see appendix 2.A.1.

A scheduling policy, roughly speaking, is a rule of determining which queues in the network to serve in any time slot, based upon the queue occupancy information. More precisely:

Definition 2.3 (Scheduling policy) *A scheduling policy is a fixed mapping of X_n into a probability distribution on D_n which respects the connectivity and operational constraints of the underlying network.*

Remarks:

1. A *distributed* scheduling policy determines the probability that a queue at some node is served based only upon on the queue occupancies at that node, and thus can be implemented in a distributed fashion at each node. This is in contrast to a *centralized* scheduling policy which requires that every node have knowledge of all queue occupancy information.

2. A scheduling policy that is dependent on queue occupancy information is an *online* scheduling policy. A degenerate scheduling policy whose output probability distribution is independent of queue occupancy information is an *offline*

scheduling policy. For example, the Static Service Split (SSS) policy described in section 2.1.2 is an offline scheduling policy.

3. In a *work-conserving* scheduling policy, a queue is never left unserved if it has a transaction which can be served in conjunction with the queues already selected for service by the scheduling policy. A scheduling policy that is not work-conserving is said to be *nonwork-conserving.*

We are now ready to introduce our notion of network capacity:

Definition 2.4 (Capacity region) *The capacity region of a network is the closure of the set of exogenous traffic rate vectors for which the system of queues in the network is rate-stable for some routing and for some scheduling policy.*

Remarks:

1. We emphasize that the capacity region of a network is not tied to a particular routing. Rather, it is the closure of the collection of achievable traffic rates taken over the set of all routings.

2. The above definition determines the set of achievable exogenous traffic rate vectors to within the "boundary" of the network's capacity region. In other words, the capacity region of a network and its set of achievable exogenous traffic rate vectors differ at most by the boundary of the capacity region. Practically speaking, this ambiguity is not important since traffic rate vectors arbitrarily close to the boundary can be implemented.

Definition 2.5 (Admissibility) *A set of exogenous traffic rates is admissible if a routing exists for which every channel in the network is offered a rate of traffic which is less than or equal to its channel capacity.*

Remark: It is clear that the capacity region of a network must lie within the set of admissible rates, for otherwise, rate-stability would obviously be violated for at least one of the network's queues.

We conclude with the following relational definition:

Definition 2.6 (Dominance) *Set A is dominated by set B if $A \subseteq B$.*

■ 2.1.1 EPS/OPS networks

Recall that a packet-switched network is an interconnection of routers, which we model as an interconnection of cell-based, input-queued (IQ) switches[3] which make

[3]While the IQ switch design is less general than that of the combined input- and output-queued (CIOQ) switch, these switches are optimal with respect to capacity performance assuming unicast flows.

scheduling decisions in a distributed fashion. Transport along links is carried out in optical fiber using WDM, and switching/routing functions at network nodes are carried out in the electronic domain (i.e., EPS) or optical domain (i.e., OPS). Although at present, the capabilities of electronic logic greatly exceed those of optical logic, we do not draw a distinction between networks employing electronics versus those employing optics for logic because our intent is to characterize the fundamental capacity limits of packet-switched networks should the full gamut of logical operations become practically feasible (in any domain) in the future. For brevity, we shall therefore employ the convention of "OPS networks" in lieu of "EPS/OPS networks" or "packet-switched networks" in the remainder of this chapter, with the understanding that whatever is said applies to both OPS and EPS networks.

Each IQ switch employs virtual output queues (VOQs) with infinite buffering capability at its input ports; that is, each input port keeps a separate queue for each output port to which it may be connected. In our model, input and output ports correspond to wavelength channels. In an OPS network without wavelength conversion capability—which is the case of principal interest in section 2.1.1[4]—each IQ switch is effectively a collection of parallel nonblocking IQ switches, one for each wavelength channel.

To obtain a view of the operational constraints of OPS network, we propose a taxonomy of networks of IQ switches. Our taxonomy is based upon two axes, as illustrated in Figure 2-2 (and Figure 2-1). The first axis characterizes networks according to the online/offline nature of the scheduling policy used. Recall that online scheduling policies make use of queue occupancy information while offline policies do not. Thus, the capacity region for online policies must dominate that of offline policies. The second axis relates to the nature of the information available to individual switches when making scheduling decisions. In centralized scheduling each network switch is privy to network-wide information, and in distributed scheduling each switch only has access to local information at that node. The capacity region for centralized policies therefore dominates that of distributed policies. This leads us to conclude that scheduling policies in quadrant 2 of Figure 2-2 have the largest capacity region. We note that OPS falls into quadrants 3 and 4 of Figure 2-2.

Until recently, the literature on switch scheduling mostly examined the performance of a switch in isolation. An important result by McKeown *et al.* [204] shows, through the use of two different scheduling policies based upon Maximum Weight Matching (MWM), that stability of an IQ switch can be attained for any admissible traffic vector with independent arrival processes. MWM is a class of scheduling policies that employs some weighting function to assign each VOQ a weight, and then matches the switch's input ports to its output ports according to the matching which

[4]Wavelength conversion will be addressed at the end of section 2.1.1, in that it will be shown that wavelength conversion capability does not provide any benefit with respect to *capacity performance* in OPS networks.

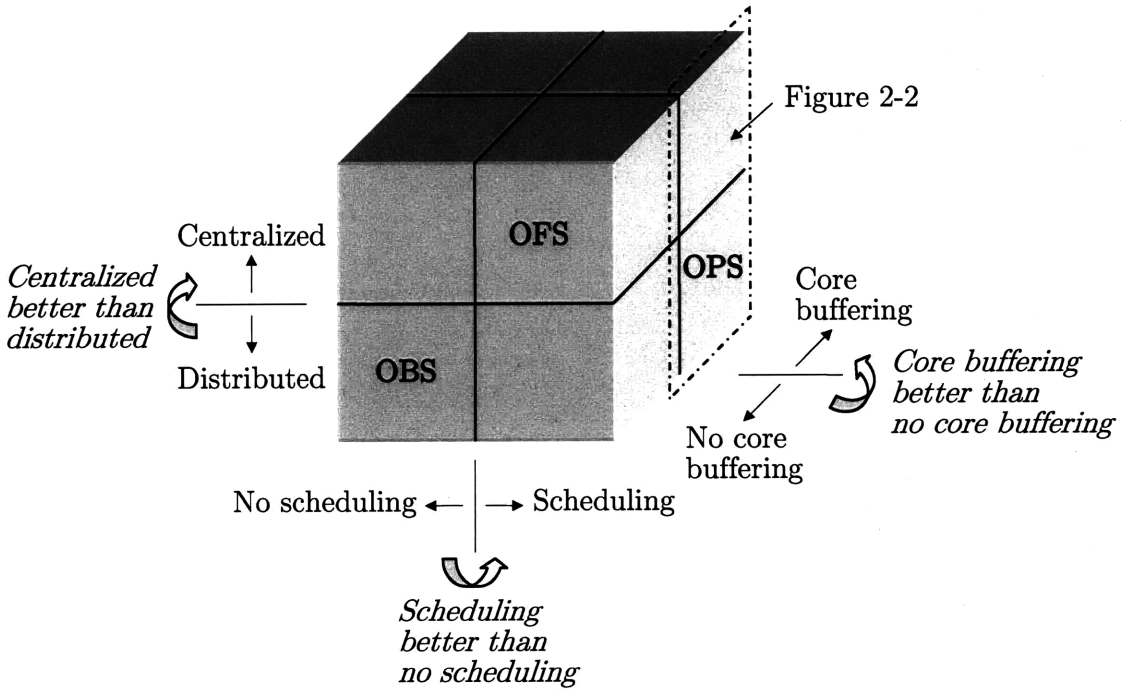


Figure 2-1. Taxonomy of optical network architectures. Relative merit is based upon capacity performance. Figure 2-2 further categorizes the indicated region according to the nature of the scheduling used.

achieves the maximum weight. Examples of weighting functions are the number of cells residing in the VOQ, or the age of the oldest cell in the VOQ.

As shown by Andrews and Zhang in [21], McKeown *et al.*'s result does not extend to networks of IQ switches. In [200], Marsan *et al.* show, however, that there exist scheduling policies in quadrants 3 and 4 of Figure 2-2 which are rate-stable as long as the offered load is stationary, satisfies the SLLN, and lies in the interior of the admissible rate region of the network. This result assumes a simple routing, but permits multiple transaction types to be associated with each source-destination node pair. A minor adaptation of this result leads to the following theorem:

Theorem 2.1 *The capacity region of an OPS network is the admissible rate region for the network. That is, OPS networks achieve the maximum possible capacity region.*

Proof. To show that the capacity region of an OPS network is the admissible rate region for the network, it suffices to show that any traffic vector in the interior of the admissible rate region has an associated routing such that the system of queues is rate-stable. Consider such a vector of traffic rates Λ^c and an associated routing which renders it admissible. We show that Λ^c and this routing can be mapped to a vector of traffic rates and an associated routing consistent with the formulation in [200], which is admissible and hence achievable. This allows us to conclude that Λ^c is in the capacity region of the network.

Owing to the independence of the transaction types, it is straightforward to see that Λ^c may be expressed as a decomposition $\Lambda^c = \sum_{i=1}^{k} \Lambda_i^s$ for some positive integer k, where each Λ_i^s is an admissible vector of traffic rates associated with a simple routing i. Note that, in general, different simple routings may require a given transaction type to follow different paths in the network from source to destination. Let us now temporarily redefine a transaction type as a path in the network, rather than a source-destination pair. Transactions of the same type, according to our previous definition, that follow different paths are now considered to belong to different transaction types. With this new definition in place, we take a union of transaction types over all simple routings i. We now have a collection of transaction types, differentiated by their associated paths in the network, along with corresponding traffic rates derived from the Λ_i^s vectors. Since we have not created, destroyed, or rerouted any transactions in the network—but merely renamed them—the resulting load on the network remains admissible, and hence achievable by [200]. This implies that Λ^c is achievable too. □

Remark: For an alternative proof of Theorem 2.1, first, note that any multihop transaction may be decomposed into a sequence of single-hop transactions. Thus, any admissible vector of traffic rates and associated routing may be alternately expressed as an admissible vector with positive rate components for single-hop transaction types only. Because this latter vector is clearly associable with a simple routing, this vector along with a simple routing falls under the network routing formulation of [200]. The mathematical interchangeability of these two vectors and routings, and hence the

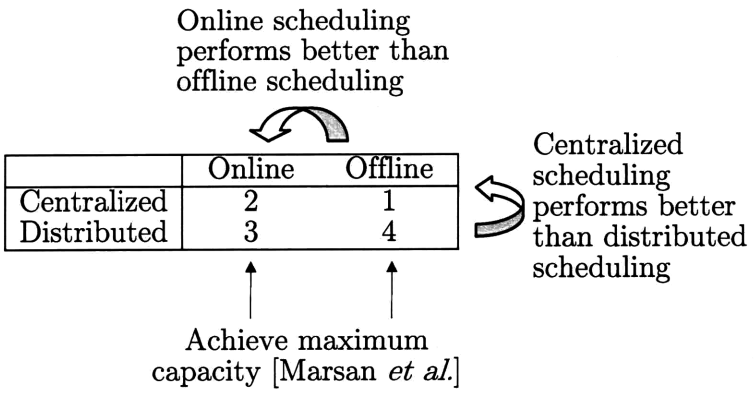


Figure 2-2. Taxonomy of networks of IQ switches. This matrix is a further categorization of the core buffering/scheduling region indicated in Figure 2-1. The quadrants labeled 1, 2, 3, 4 represent the four possible network types with the two criteria.

achievability of the former admissible vector and routing, is evident in the proofs of [200].

Theorem 2.1 implies that, as far as network capacity is concerned, wavelength conversion capability does not provide any advantage in OPS networks. To see this, first note that the capacity region of an OPS network with w unit capacity wavelength channels, with or without wavelength conversion capability, must respect the admissibility constraints which state that no link may be subscribed beyond rate w. Theorem 2.1, however, implies that any traffic rate vector in this region is achievable without wavelength conversion. (To see this, multiply any such vector Λ by $1/w$ and assign the resulting scaled vector to each of the w wavelength channels. By Theorem 2.1, the scaled vector lies in the capacity region of a single wavelength channel OPS network. Hence, vector Λ lies in the capacity region of the OPS network with w wavelength channels and no wavelength conversion.) Since the capacity region of an OPS network with wavelength conversion must dominate that of the same OPS network without wavelength conversion, and since the latter achieves the maximum capacity region, we conclude that wavelength conversion indeed does not provide any capacity performance advantage. This result is summarized in Corollary 2.2 in section 2.3.1.

■ 2.1.2 OFS networks

We now address the capacity region of OFS networks, and more generally, networks with buffering at source nodes but no buffering capability in the network core (e.g., OCS). In such networks, data is scheduled to traverse the network from source to destination without being buffered at intermediate nodes. To characterize the capacity region of this family of networks it is helpful to view a network as a large,

generalized switch. Viewed this way, input and output ports correspond to the nodes in the network, and a connection between an input and output port represents a transaction type, or flow, being serviced between the two nodes corresponding to the ports.

Before proceeding, we need the following definitions:

Definition 2.7 (Feasible network state) *A feasible network state is a set of transaction types, or flows, that can be simultaneously serviced, while respecting the connectivity and operational constraints of the underlying network.*

Definition 2.8 (Stable set) *A stable set of an undirected graph is a (possibly empty) set of vertices in which no two vertices have an edge connecting them.*

Definition 2.9 (Incidence vector of a stable set) *The incidence vector of a stable set of an undirected graph with F nodes is the binary vector $(e_1, e_2, \ldots, e_F)$, where $e_i = 1$ if node i is in the stable set, and $e_i = 0$ otherwise.*

Definition 2.10 (Stable set polytope) *The stable set polytope of an undirected graph is the convex hull of the incidence vectors of all stable sets.*

Definition 2.11 (Conflict graph) *The conflict graph associated with a network and a simple routing is an undirected graph in which vertices represent the set of transaction types or flows to be served in the network. An edge exists between vertices i and j if the flows corresponding to nodes i and j cannot simultaneously exist in the network (i.e., the flows share at least one link).*

Note that a stable set of a conflict graph of a network represents a feasible network state. Building on the work in [49, 280], the capacity region of a network without core buffering is related to the stable set polytopes of the conflict graphs of the network. By means of time-sharing among all possible stable sets, we may conclude that the rate region defined by the convex hull of the stable set incidence vectors lies within the capacity region of the network. That this convex hull is exactly the capacity region of the network follows from the next lemma, which is an adaptation of [275, Proposition 1] to our present context. It essentially says that the capacity region of a network without core buffering is characterized by the set of all offline, static, random scheduling schemes.

Consider a SSS scheduling policy for a network without core buffering, which chooses in each time slot the feasible network state (i.e., stable set) k for service with probability ϕ_k. Define Φ as the vector of probabilities ϕ_k, and $v(\Phi)$ as the function which maps an SSS policy's Φ into the vector of flow rates $v = (v_1, v_2, \ldots, v_F)$. Specifically:

$$v_j = \sum_k \phi_k I_j(k)$$

where $I_j(k)$ is an indicator function that has the value of unity if the feasible network state k services flow j.

Lemma 2.1 *Consider a network lacking core buffering with flow rate vector Λ. For there to exist a scheduling policy under which the network is stable, the condition:*

$$\Lambda \leq v(\Phi), \text{ for at least one SSS policy } \Phi$$

is necessary, and the condition:

$$\Lambda < v(\Phi), \text{ for at least one SSS policy } \Phi$$

is sufficient, where the inequalities hold component-wise.

This lemma assists us in proving the following theorem:

Theorem 2.2 *The capacity region $\mathcal{P}$ of a network lacking core buffering is the convex hull of the union (over all simple routings) of the stable set incidence vectors of the conflict graphs.*

Proof. We first show that every traffic rate vector Λ that lies within the capacity region of the network also lies within the convex hull of the union (over all simple routings) of the stable set incidence vectors of the conflict graphs. Toward this end, note that the necessity condition in Lemma 2.1 provides, for each such vector Λ, an SSS policy with associated probability vector Φ such that $\Lambda \leq v(\Phi)$. It can be shown that Φ can always be transformed into another probability vector Φ' such that $\Lambda = v(\Phi')$ (see appendix 2.A.2). Recall that, by definition, the probability vector Φ' represents a collection of probabilities of the network being in its feasible states. Thus, Λ can be expressed as a convex combination of incidence vectors of feasible network states with weights equal to Φ'. Since each feasible network state corresponds to a stable set of some conflict graph, it follows that Λ lies within the convex hull of the union (over all simple routings) of the stable set incidence vectors of the conflict graphs.

We now show that a vector in the convex hull of the union (over all simple routings) of the stable set incidence vectors of the conflict graphs also lies within the capacity region of the network. Consider such a vector Λ^* formed by a convex combination of stable set incidence vectors with weights Φ^*. Note that Λ^* can be trivially mapped to an SSS policy with associated probability vector Φ^*. Under this SSS rule, it is straightforward that ϕ_k^* represents the average fraction of time that the network is in feasible state k. Thus, the vector $v(\Phi^*)$ represents the flow rates achieved under this SSS rule. Recalling that Λ^* is a convex combination of stable set incidence vectors with weights Φ^*, it must be true that $\Lambda^* = v(\Phi^*)$. This proves that Λ^* is achievable and therefore lies within the capacity region of the network. □

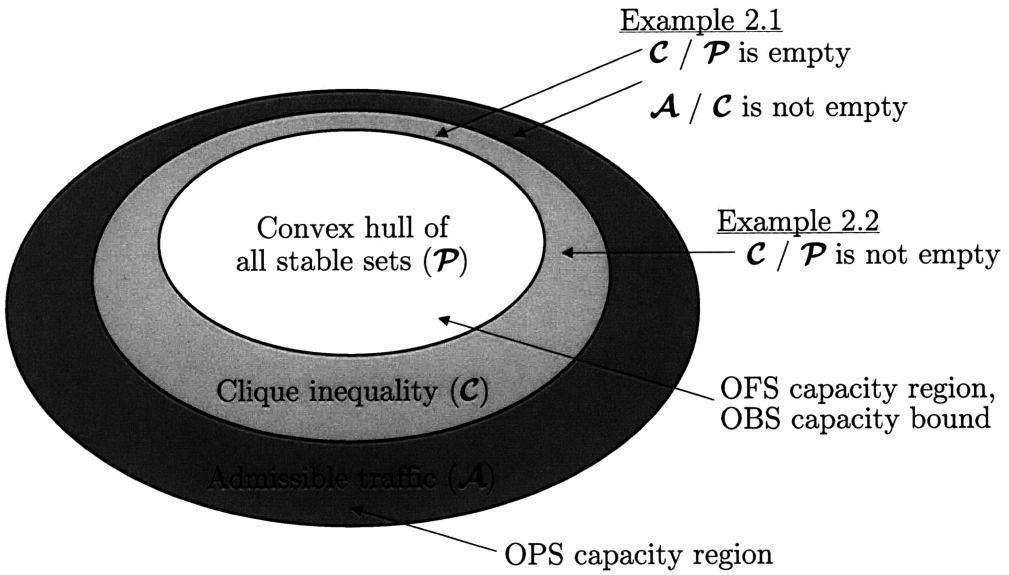


Figure 2-3. Relationship among different rate regions when $w = t$.

Remark: We emphasize that Theorem 2.2 does not characterize the capacity region of OFS networks, but provides an outer bound for it. This is because the only constraint that we imposed in the theorem's derivation was the absence of buffering in the network core. In particular, we allowed for flows to be broken up into arbitrary granularities and serviced piecemeal in a cell-based fashion, as in OCS. As discussed earlier, the assumption of being able to break up flows and service them piecemeal is contrary to the spirit of OFS. We, therefore, naturally wonder if there is an inherent sacrifice in the capacity region of OFS relative to the region $\mathcal{P}$. Our main result of this section (Theorem 2.3) shows that the answer is no.

We now turn our attention to Figure 2-3 which illustrates the relationship among $\mathcal{P}$, the admissible rate region $\mathcal{A}$, and a region that, for a reason which will soon become clear, we call the clique inequality region $\mathcal{C}$. We introduce the region $\mathcal{C}$ into our discussion because, among other reasons, it provides intuition as to why the regions $\mathcal{A}$ and $\mathcal{P}$ differ. The concept of a clique—a fully connected subgraph—is important for the following discussion.

Let $Z = (z_1, z_2, \ldots, z_F)$ be a point in F-dimensional Euclidean space representing a set of flow rates. The following two sets of inequalities are satisfied if Z lies within the stable set polytope of the conflict graph for a particular simple routing [263]:

- Trivial constraints: $0 \leq z_i \leq 1$, for all i.
- Clique inequalities: $\sum_{i \in K} z_i \leq 1$, for every clique K of the conflict graph.

This leads us to the following definition:

Definition 2.12 (Clique inequality region) *The clique inequality region $\mathcal{C}$ is the convex hull of the union (over all simple routings) of the rate regions defined by the trivial constraints and the clique inequalities.*

It can be shown that the stable set polytope of the conflict graph for a particular simple routing is the integer hull of the polytope defined by the trivial constraints and clique inequalities. Hence, $\mathcal{P} \subseteq \mathcal{C}$. Since the problems of finding maximum-size stable sets and cliques in a graph are NP-complete, a simple inequality characterization of these regions generally does not exist [263].

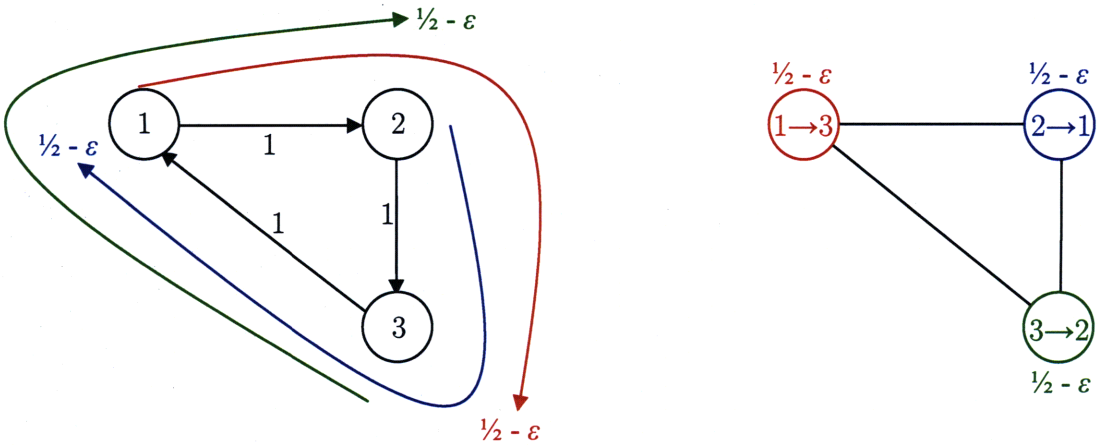
In general, $\mathcal{C} \subseteq \mathcal{A}$ because the clique inequalities are stricter than the admissibility constraints, which state that no link may be oversubscribed. The clique inequalities require that any flows which form a clique in the conflict graph have an aggregate capacity less than unity. If for each clique all the flows in the clique share at least one common link, then the clique inequalities are equivalent to the admissibility constraints. However, as shown in Example 2.1 below, it is possible for flows to form a clique without all merging at a particular link. Therefore, we conclude that the clique inequalities are stricter than the admissibility constraints, and consequently define a smaller rate region. Thus, $\mathcal{P} \subseteq \mathcal{C} \subseteq \mathcal{A}$.

It is worth noting that, while the clique inequalities are always satisfied within the stable set polytope of a conflict graph, they exactly characterize the stable set polytope for a particular family of conflict graphs known as perfect graphs [263]. In a perfect graph, the chromatic number[5] equals the size of the largest clique for each of its induced subgraphs. For this family of graphs, the task of finding the region $\mathcal{C} = \mathcal{P}$ is solvable in polynomial time [263]. While it is true that for perfect graphs $\mathcal{P} = \mathcal{C}$, it is not necessarily true that $\mathcal{A} = \mathcal{C} = \mathcal{P}$. Therefore, even for network topologies which maximize $\mathcal{P}$ relative to the clique inequality region $\mathcal{C}$, an OPS architecture atop the same network topology will generally have a larger capacity region. This is illustrated in the following example:

Example 2.1 (Three node ring) *Consider the network and associated transaction types drawn in Figure 2-4(a), where each transaction type has rate $1/2 - \varepsilon$ and the capacity of each link is unity. The corresponding conflict graph, which is perfect, is drawn in Figure 2-4(b). Since the clique inequalities are not satisfied, a schedule that can accommodate this traffic demand does not exist for networks which lack core buffering. However, the traffic demand is clearly admissible, which implies that an OPS network can accommodate the demand.*

Consider the same network, but under the all-to-all traffic demand illustrated in Figure 2-5(a). The corresponding conflict graph, which is also perfect, is drawn in Figure 2-5(b). We assign the single-link transaction types rate x and the two-link transaction types rate y. Figure 2-5(c) then illustrates the different rate regions $\mathcal{A}$,

[5]The chromatic number of a graph is the least number of colors required to color its vertices such that adjacent vertices have different colors.



(a) Network lacking core buffering (b) Conflict graph

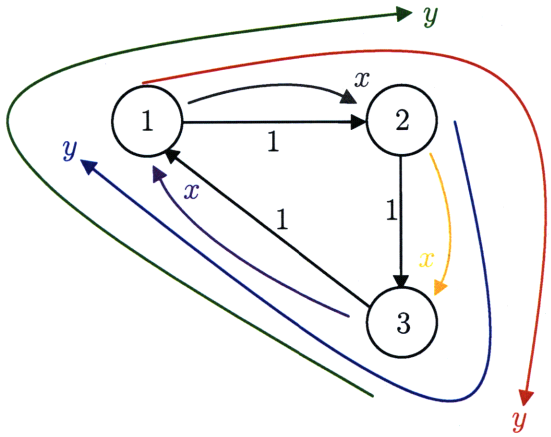
Figure 2-4. Illustration of the network considered in Example 2.1 and its associated conflict graph when only two-link transaction types are assumed.

$\mathcal{C}$, and $\mathcal{P}$ as functions of x and y. The region $\mathcal{A}$ is the triangular region with vertices $(0,0)$, $(1,0)$, and $(0,1/2)$; $\mathcal{C}$ and $\mathcal{P}$ are the trapezoidal region within $\mathcal{A}$ with vertices $(0,0)$, $(1,0)$, $(1/3,1/3)$, and $(0,1/3)$. We note that these capacity regions, which assume uniform traffic, are actually a two-dimensional cross-sections of the unconstrained six-dimensional capacity regions. We also note that, assuming uniform all-to-all traffic (i.e., all possible transaction types have equal rates: $x = y$), then a network without core buffering can achieve the same set of rates as an OPS architecture. As evident in Figure 2-5(c), the maximum common operating point of these two architectures is $(1/3,1/3)$. We see in section 2.2.1 the capacity equivalence of these architectures is a general property of networks having ring topologies under uniform all-to-all traffic.

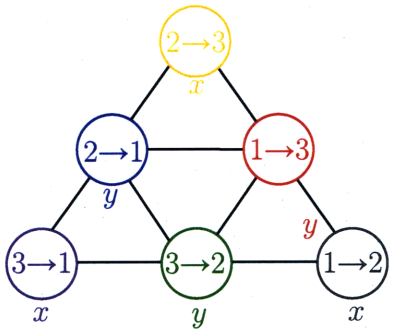
Finally, owing to the perfectness of the conflict graph, the capacity region for networks which lack core buffering is completely characterized by the trivial and clique inequalities on the flow rates $z_{i\to j}$:

$$
\begin{aligned}
0 \le z_{i\to j} \le 1, \quad & \text{for } i,j = 1,2,3 \text{ and } i \neq j \\
z_{1\to 2} + z_{1\to 3} + z_{3\to 2} &\le 1 \\
z_{2\to 1} + z_{1\to 3} + z_{3\to 2} &\le 1 \\
z_{2\to 1} + z_{3\to 1} + z_{3\to 2} &\le 1 \\
z_{2\to 1} + z_{1\to 3} + z_{2\to 3} &\le 1
\end{aligned}
$$

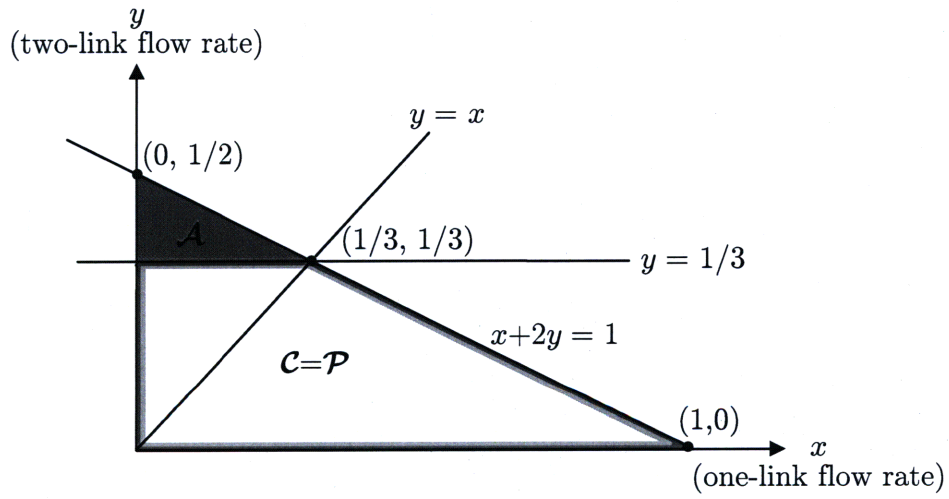
We now investigate online, cell-based algorithms that achieve rate-stability in the region $\mathcal{P}$. In [20, 275], the authors propose a family of scheduling policies known as MaxWeight scheduling (with MWM as a special case) in the context of a generalized,



(a) Network lacking core buffering



(b) Conflict graph



(c) Rate regions

Figure 2-5. Illustration of the network considered in Example 2.1, its associated conflict graph when an all-to-all traffic demand is assumed, and the different rate regions when one-link transaction types have rate x and two-link transaction types have rate y.

cell-based switch. The switch model employed assumes that switch states follow a finite state, discrete time Markov chain, where, in each state, the switch has an associated finite set of scheduling choices. We may view a network lacking core buffering as such a generalized switch with a single state Markov chain in which the finite set of scheduling choices correspond to the feasible network states that can be used to service flows. In [275, Lemma 5], the author proves, using fluid model techniques, that MaxWeight scheduling policies achieve the maximum capacity region. This leads to the following:

Lemma 2.2 *For networks lacking core buffering, the capacity region $\mathcal{P}$ can be achieved using an online, cell-based MaxWeight scheduling algorithm (with MWM as a special case).*

Remarks:

1. The definition of stability employed in [275] involves the existence of a set of positive recurrent states in the Markov chain underlying the state of the generalized switch. It is straightforward to see that this notion of stability implies rate-stability.

2. Lemma 2.2 generalizes the optimality of MWM scheduling in two ways. First, it broadens the class of optimal scheduling policies to the MaxWeight family, which have other attractive properties (e.g., low delay). Second, and more importantly for our discussion, it demonstrates the optimality of MaxWeight scheduling for generalized switches which differ from traditional nonblocking switches in that they may have additional constraints. In the context of this chapter, these additional constraints correspond to topology and resource constraints in the network. Unfortunately, applying a MaxWeight schedule to a generalized switch is equivalent to finding the maximum-weight stable set of a (conflict) graph, which is known to be NP-complete [263]. Polynomial-time algorithms for finding the maximum-weight stable set of a graph do exist, however, for perfect graphs and their complements, t-perfect graphs, and claw-free graphs [263].

The next lemma is an application of [112, Lemma 1] to our constrained switch model of networks lacking core buffering. The fluid model techniques employed in the proof in [112] are immediately applicable to our model. Specifically, in the fluid limit of a generalized switch process, the capacity region achieved by a scheduling algorithm that is "suboptimally bounded" is indistinguishable from that of an optimal scheduling algorithm.

Lemma 2.3 *The capacity region of a network lacking core buffering can be achieved by an online, cell-based scheduling algorithm if the value of the weight of the matching it uses at each time slot is less than the maximum weight by at most a bounded constant.*

In [112], the authors investigate the performance of scheduling algorithms for non-blocking IQ switches which are flow-based—they switch flows of variable number of cells as indivisible entities, rather than segment them into cells and switch them with cell-based schemes. The authors show that any flow-based algorithm which is work-conserving cannot be stable for all admissible traffic rates. In particular, for a scheme to be rate-stable, it is necessary to periodically ensure that all of the switches ports are free and to then "resynchronize" the switch state with the state of all the queues. This requires the switch to wait for periods of time for some of the ports to become free. The authors thus propose a family of nonwork-conserving scheduling algorithms based upon MWM that switch variable length flows as indivisible entities, which can be adapted to our OFS model. The previous lemma is instrumental because it implies that if we wait for bounded periods of time for the purpose of this resynchronization, then the weight of our matching at any instant in time is less than the optimal weight by a bounded constant. By waiting for an arbitrarily long (but bounded) period of time between resynchronizations of the switch, we can ensure that the bandwidth waste due to waiting is arbitrarily small. Indeed, it is the long-term *fraction* of time spent waiting for the ports to become free that governs the achievable throughput, and the size of the flows is therefore irrelevant (as long as they are bounded).

Specifically, in the algorithm proposed in [112], multiple variable-length flows are concatenated, up to a fixed aggregate number of cells, to form a large block which is transmitted through the switch during a single configuration. In particular, if the maximum flow length is $L_{\max}$ data cells, then flows of the same class of traffic are aggregated into blocks up to $kL_{\max}$ time slots long (where $k \geq 1$ is an integer), which allows at least k flows to be aggregated in each such block (see Figure 2-6). The duration of each network configuration is $kL_{\max}$ (core) time slots, which allows all of the aggregated flows in a block to be transmitted monolithically from an input port to an output port. It is straightforward to see that the maximum number of cells that are "wasted" at the end of each block of $kL_{\max}$ time slots as a result of an integer number of variable-length flows being unable to use all $kL_{\max}$ slots is $L_{\max} - 1$, yielding the fraction of bandwidth wasted as:

$$\frac{L_{\max} - 1}{kL_{\max}} \approx \frac{1}{k}$$

for $L_{\max}$ large. Thus, the achievable rate region is $\mathcal{P}$ scaled by $\frac{k-1}{k}$ (i.e., $1/k$ of the switch's capacity is wasted). The achievable data rates are therefore governed by k and not by $L_{\max}$.

Using the previous lemma, we have the following result which is an adaptation of [112, Theorem 3]:

Lemma 2.4 *There exist online, nonwork-conserving, flow-based scheduling algorithms which are rate-stable for the interior of the capacity region of a network lacking core buffering, provided that the maximum flow length is bounded.*

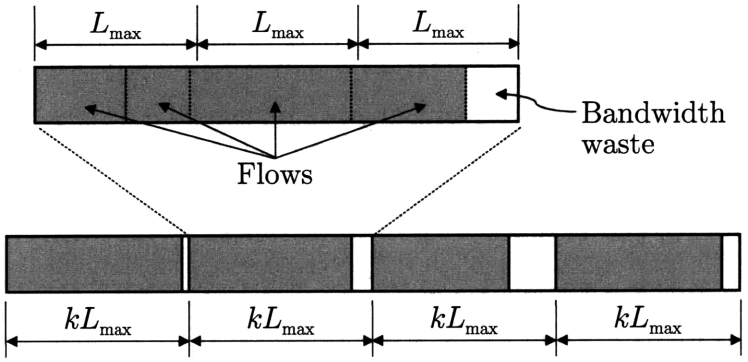


Figure 2-6. The aggregation mechanism employed to accommodate variable-length flows with $k = 3$.

Remark: In the discussion following Definition 2.3, we proposed a strict definition for a distributed scheduling policy, which allows only local queue occupancy information to be used in scheduling decisions at each node. More relaxed definitions, however, allow for varying degrees of information to be exchanged among nodes and for this information to be used in scheduling decisions at nodes. Lemmas 2.3 and 2.4 imply that, if a distributed scheduling policy executed at each node uses stale network state information from other nodes, it can still achieve optimal capacity performance provided that the weights of the matchings it uses are suboptimally bounded. More generally, the capacity regions of other distributed OFS scheduling policies can be deduced from the relationships of the matching weights that they achieve relative to the maximum weight matching[6].

This immediately leads us to our main OFS result:

Theorem 2.3 *The capacity region of an OFS network is the convex hull of the union (over all simple routings) of the stable set incidence vectors of the conflict graphs.*

Remark: A stronger form of stability than rate-stability can, in fact, be proven for the interior of the OFS capacity region. See 2.A.1 for details.

Extensions

In this section, we invoked certain assumptions regarding the operational properties of the OFS architecture which rendered the analysis simpler. Some of these assumptions, such as the absence of wavelength conversion, may be relaxed in a straightforward manner. Specifically, in Theorem 2.3, stable sets of conflict graphs, which correspond to *feasible network states*, served as the building blocks for the construction of the OFS capacity region. In a more general setting, the stable set incidence vectors need to be substituted with flow incidence vectors which represent feasible network states,

[6]See, for example, [221].

and the other results of this section still hold. For example, the approach used to derive the capacity region of OFS may be generalized to accommodate wavelength conversion in the core, although separability of the wavelength channels may no longer be invoked since wavelength channels are coupled. In this case, instead of using stable set incidence vectors of conflict graphs as the building blocks of the capacity region, we employ flow incidence vectors which abide by the constraint that no physical link is subscribed by more than w flows, where w is the number of wavelengths per fiber link and $w = t$. The capacity region is then given by the convex hull of the union, taken over all simple routings, of these flow incidence vectors. Limited wavelength conversion—for example, in wavelength range and/or node location—can similarly be handled by employing the flow incidence vectors representing appropriate feasible network states.

The scheduling algorithm used to arrive at Theorem 2.3 may be modified to account for latency costs of reconfiguring the WAN logical topology[7] (e.g., transceiver tuning time, OXC switch fabric reconfiguration time, and control information dissemination time). Toward this end, let us assume that every reconfiguration of the WAN requires τ time slots during which data cannot be sent. If we append τ slots for potential reconfiguration to the end of each block of $kL_{\max}$ slots, then, by the same reasoning preceding Lemma 2.3, the achievable rate region is further scaled by:

$$\frac{kL_{\max}}{kL_{\max} + \tau}.$$

In order to mitigate the burden of reconfiguration, a positive constant may be added to the present configuration's weight as a means of expressing the preference to not reconfigure the network. Since the weight associated with any configuration is less than the maximum weight by this constant at most, Lemma 2.3 implies that the rate region of $\mathcal{P}$ scaled by:

$$\frac{L_{\max}(k-1)}{kL_{\max} + \tau}$$

is achievable. The achievable rate region is, in fact, generally larger, but the extent to which this is the case depends upon the traffic statistics.

As mentioned earlier, the *online* scheduling algorithm upon which Theorem 2.3 is based is NP-complete as it entails the computation of the maximum-weight stable set of the conflict graph. However, for certain families of graphs—perfect graphs and their complements, t-perfect graphs, and claw-free graphs—this computation can be performed in polynomial-time [263]. An interesting extension is the investigation of techniques for modifying a conflict graph into one of these aforementioned special graphs (to enable polynomial-time computation of the maximum-weight stable set), and to assess the imputed penalty in the capacity region. Perfect graphs, for example, are characterized by the absence of a hole or anti-hole. A conflict graph can therefore

[7]Related extensions are discussed in [40].

be converted to a perfect graph by augmenting any hole or anti-hole with a link. The addition of such a link to the conflict graph imposes an additional restriction on the coexistence of two transaction types in a feasible configuration of the original network, which entails a capacity performance penalty. Alternatively, if the conflict graph is such that each of its vertices can be covered q times by a family of p induced perfect subgraphs, then an *offline*, polynomial-time scheduling algorithm can be constructed with a capacity penalty of at most p/q. This follows from the fact that the stable set polytope of each induced subgraph can be characterized in polynomial-time [263]; and from [116, Corollary 2.3.5], which upper bounds the imperfection ratio of the original imperfect graph by p/q, leading to a scaling-down of the capacity region by p/q [169, Theorem 9].

In closing this subsection, we note that Corollary 2.3 in section 2.3.2, where the case of $w \geq t$ is considered for OFS, may be viewed as a generalization of the results in this section.

■ 2.1.3 OBS networks

As discussed in section 1.3.1, there is a great deal of variability in how OBS networks are designed. The particular OBS model that we consider here, though simple, captures the spirit of the OBS transport philosophy. Our OBS model is based upon the common implementation of OBS in which packets are assembled at WAN ingress nodes, according to destination and QoS, to form bursts. Core nodes, as in OFS, are equipped with OXCs which have no buffering capability. We also assume that, as in OFS, wavelength conversion is not used in the network core. Recall that in OBS, contention for resources among bursts may occur at one or more of the intermediate nodes. In our model, such contention for resources is resolved via burst discardment.

Based upon the above model, OBS networks can be viewed as incarnations of OFS networks in that they lack buffering capability in the core, and that they require bursts to be serviced as indivisible entities. However, owing to the fact that they employ random-access instead of scheduling[8], OBS networks are generally characterized by nonzero burst blocking probabilities. Specifically, the fact that bursts may require retransmission can lead to instability on an individual link, even if the offered traffic is admissible [19]. Furthermore, the lack of coordination among core links implies that resources are wasted if they are consumed by bursts that are eventually discarded. This is illustrated in Example 2.2. For these two reasons, OBS networks are generally incapable of achieving rate-stability within the OFS capacity region. This leads to the following result:

[8]This is true of all variations of OBS, except for Wavelength-Routed OBS [173]. As discussed earlier, this version of OBS is essentially OCS, in that advanced scheduling sets up lightpaths for bursts.

Corollary 2.1 *The capacity region of an OBS network is dominated by the convex hull of the union (over all simple routings) of the stable set incidence vectors of the conflict graphs.*

Remarks:

1. The degree to which the capacity of an OBS network differs from the capacity region of the analogous OFS network depends upon the traffic statistics, and on network architecture parameters such as burst aggregation and retransmission policies.

2. Obtaining analytic expressions for OBS network capacity regions is related to, and in fact more difficult than, characterizing the stability regions of retrial queues, for which analytic solutions are available only under special circumstances [92,98]. As a result, with the exception of works such as [253], analytic studies of OBS network performance have usually only considered a single edge or core OBS node in isolation. These analyses, however, neglect the key property of OBS networks that resources are wasted if they are consumed by bursts that are eventually discarded.

We now analyze the performance of OBS networks under a further simplified model in order to gain a rough quantitative sense as to how these networks compare to their OPS and OFS counterparts. We model a wavelength channel in an OBS network as a multiple-access system with a finite number of users which represent the burst types on that link. Because access to a wavelength channel is mediated by switch ports or tunable lasers, we assume that channel capture occurs by bursts. That is, if the channel is free when a burst transmission is attempted on it, then the channel is reserved for the duration of the burst. For analytical tractability, a burst traversing an OBS network is transmitted along a series of wavelength channels on links that is treated as a cascade of independent multiple-access systems. Although the actual load is reduced as the network egress is approached, the load is unreduced when requests are made by the control packet preceding a burst. This is because the control packet requests resources from all nodes along the burst's intended path regardless of whether its requests at previous nodes were successful.

We assume that each burst type on a link produces, independent of other burst types, a renewal process comprising a sequence of independent transmit and idle states representing the aggregate of fresh arrivals and retransmissions. Each transmit state, which corresponds to a burst being transmitted, has length drawn from a distribution $p_L(n)$ with mean $\overline{L}$; and each idle state is geometrically distributed with mean q. While this assumption may not be realistic because retransmissions corresponding to the same burst should have the same length, the derived throughput will not be affected owing to the independence of link capture and burst length. A shortcoming of our model, however, is that the distribution of idle states is independent of whether or not there are backlogged packets attempting retransmission.

We assume that wavelength channel capture requires that all burst types not transmit on that wavelength channel. The probability that a burst type successfully transmits is assumed to be independent of previous attempts to transmit. Finally, links are assumed to be independent. Hence, the probability that a given burst of type j is successfully received at its destination h_j hops away is:

$$P_s(j) = \prod_{i=1}^{h_j} \Pr(\text{success on hop } i).$$

Depending on how a wavelength is chosen at the source node, we have the following two alternative OBS models from which to derive the product terms in the above equation. From this probability of success, both an approximate average delay for a burst type and the throughput of a link may be found. The average queueing delay W_j experienced by a burst of type j from the time it arrives at its source node to the time it begins its successful transmission, neglecting propagation delay, is approximately:

$$\begin{aligned} W_j &= (\text{avg. no. retransmissions}) \left(\overline{L} + q\right) \\ &= \left(\frac{1}{P_s(j)} - 1\right) \left(\overline{L} + q\right) \end{aligned}$$

where, inherent in this expression is the assumption that retransmission attempts are independent. This assumption is reasonable when there are many burst types traversing each link and each burst type offers a light traffic load to links [101]. The throughput of link i is:

$$S(i) = \sum_{j=1}^{F} P_s(j) \frac{\overline{L}}{\overline{L} + q} I_i(j),$$

where F is the number of types of bursts in the network, and $I_i(j)$ is the indicator function that has the value of unity if a burst of type j traverses link i. The fractional term in the above summation represents the probability that a burst of type j is transmitting on link i and its form follows from the ergodicity of the transmit-idle renewal process.

OBS Model 1

In this model, it is assumed that one of the t wavelength channels is selected with uniform probability at a source node wishing to transmit a burst. Thus, the probability that a given burst of type j is successfully received at its destination h_j hops away is:

$$P_s(j) = \prod_{i=1}^{h_j} \sum_{u=0}^{b_i - 1} B\left(b_i - 1, u, \frac{\overline{L}}{\overline{L} + q}\right) \left(\frac{t-1}{t}\right)^u,$$

where b_i is the number of bursts traversing the i^{th} link along the burst's path, and $B(x, y, z)$ denotes the binomial probability[9] of y successes from x trials with individual trial success probability z. In this expression, the summation index u represents the number of burst types which are attempting transmission on a link at that instant in time.

OBS Model 2

This OBS model is a refinement of the previous model in that a source node wishing to transmit a burst does not choose a wavelength randomly, but instead chooses a vacant wavelength if one exists. To conduct this analysis correctly, the order in which bursts attempt transmission on each link must be taken into account. Since such an exact analysis is intractable, we carry out an approximate analysis in which we assume that:

1. It is known on which wavelengths bursts originating at upstream nodes attempt transmission on the present link, even if their transmission has yet to begin.

2. These upstream bursts select their wavelength channels independently and uniformly. This is a reasonable assumption provided that these upstream bursts originate at different links.

Considering the impact of these two simplifying assumptions on the analysis, we expect that their effects offset each other to some extent.

For brevity, in the following we shall refer to burst types whose paths include the present link as the first hop as *originating* on the present link; and to burst types whose paths include the present link but not as the first hop as *not originating* on the present link.

The probability that a burst is successfully carried on its first link is given by:

$$\Pr(\text{success on hop 1}) = \sum_{x=0}^{b_{p,1}} \sum_{y=0}^{\min(t-1,x)} \sum_{z=y}^{\min(t,b_{s,1}+y)-1} U(t,y,x) \\ B\left(b_{p,1}, x, \frac{\overline{L}}{\overline{L}+q}\right) B\left(b_{s,1}-1, z-y, \frac{\overline{L}}{\overline{L}+q}\right), \quad (2.1)$$

In this expression, the index x represents the number of burst types that do not originate on the burst's first link but that are attempting transmission on the link; the index y represents the number of burst types that do not originate on the burst's first link that have successfully captured wavelength channels on the link; the index z represents the total number of wavelength channels that have been captured on the burst's first link; and the index difference $z - y$ is the number of burst types originating on the burst's first link that are (successfully) transmitting on the link.

[9]For future reference, we define $B(x, y, z)$=0 if $0 \leq y \leq x$ is violated.

The parameters $b_{s,h}$ and $b_{p,h}$ denote the number of burst types that originate and do not originate, respectively, on the h^{th} hop of the present burst's path. Finally, the term $U(t, y, x)$ is the probability that exactly y wavelength channels on a link have had transmission attempted on them by x transmitting burst types that do not originate on the link. The computation of $U(t, y, x)$ is discussed in appendix 2.A.3.

The probability that the burst is successfully carried on hop h $(2 \leq h \leq h_j)$ is similarly given by:

$$\Pr(\text{success on hop } h) = \sum_{x=0}^{b_{p,h}-1} \sum_{y=0}^{\min(t-1,x)} \sum_{z=y}^{\min(t,b_{s,h}+y)-1} \left(\frac{t-z}{t}\right) U(t,y,x) \\ B\left(b_{p,h}-1, x, \frac{\overline{L}}{\overline{L}+q}\right) B\left(b_{s,h}, z-y, \frac{\overline{L}}{\overline{L}+q}\right), \quad (2.2)$$

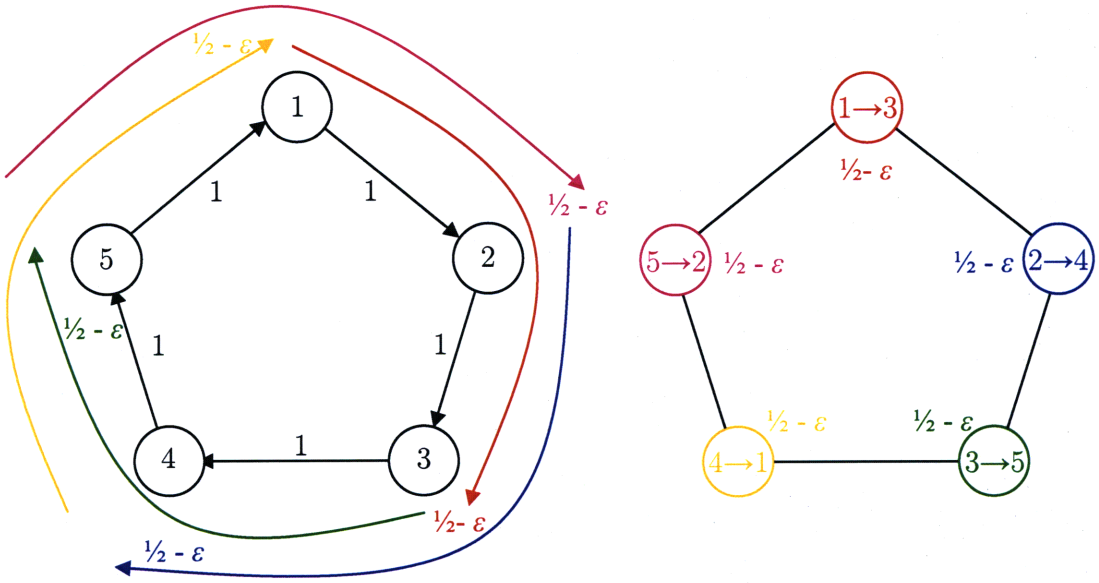
where the expression $\frac{t-z}{t}$ is the probability that none of the z busy wavelengths correspond to the present burst's wavelength.

In the following example, we illustrate through a simple network topology and routing that the capacity region of OFS is smaller than that of OPS, and that the capacity region of OBS is, in turn, significantly worse than that of OFS.

Example 2.2 (Five node ring) *Consider the five node ring depicted in Figure 2-7(a), and assume that* $t = 1$. *Drawn in the figure are also the offered transaction types, each of which is of rate* $1/2 - \varepsilon$. *The capacity of each link is unity, implying that the traffic pattern is admissible and can therefore be serviced by an OPS architecture. In fact, if we constrain ourselves to uniform traffic, the capacity region of this OPS network comprises the set of rates bounded above by* $1/2$.

By examining the conflict graph drawn in Figure 2-7(b), it is apparent that the clique inequalities are satisfied and that this traffic demand is therefore contained in the clique inequality region $\mathcal{C}$. *To determine whether the traffic pattern is contained within the OFS stability region* $\mathcal{P}$, *we first observe that, at any instant in time, at most two of the flows may be serviced using an OFS architecture. Thus, at any instant in time, at least one link in the network is unutilized. This under-utilization of at least* $1/5$ *of link resources implies that the traffic demand cannot be accommodated, as nearly full link utilization is required (when* ε *is very small). In fact, it can be shown that for uniform traffic, a rate bounded above by* $2/5$ *can be offered by each flow, which would achieve the* $4/5$ *utilization bound just mentioned. Thus, assuming uniform traffic, then the capacity region of this OFS network contains the set of flow rates bounded above by* $2/5$.

In an OBS implementation of the network, each link is shared by two contending bursts and each burst traverses two links. Thus, using the above approximate OBS



(a) Network lacking core buffering (b) Conflict graph

Figure 2-7. Illustration of the network considered in Example 2.2, and its associated conflict graph.

analysis, the throughput of each link (and by symmetry, the network) is:

$$S = \frac{2\overline{L}q^2}{(\overline{L}+q)^3}.$$

Assuming that $\overline{L} = q$*, which is equivalent to letting each burst type's aggregate (fresh arrivals plus retransmissions) rate be* $1/2$*, then the throughput of the system is* $1/4$*. If we assume that* $3\overline{L} = 2q$*, which is equivalent to letting each burst type's aggregate rate be* $2/5$*, then the throughput of the system is* $36/125 = 0.288$*. It can be shown that throughput is maximized when* $2\overline{L} = q$*, or when bursts are offered at aggregate rates of* $1/3$*. This yields a maximum throughput of* $8/27 \approx 0.296$*. Thus, under the above traffic assumptions, the capacity of this OBS network is limited to burst rates of less than approximately* $4/27$*.*

Remark: The above example illustrates that OPS networks, owing to their ability to buffer in the core, have a larger capacity region than OFS networks which lack core buffering. A significant performance disparity is also observed between OFS and OBS, which are physically similar architectures. This performance difference can be attributed to the benefit of scheduling over random-access, analogous to the benefit of TDM over Aloha in a multiple-access channel (with capture). Example 2.2, and its generalization to larger rings, thus illustrate that the capacity inequivalence of OPS, OFS and OBS exists not just for contrived networks, but also for realistic network topologies such as rings, and for realistic routings such as shortest-path routing.

■ 2.2 Topology case studies

In this section, we investigate the OPS, OFS and OBS capacity properties of two important classes of network topologies: bidirectional rings and Moore Graphs. We assume that the traffic rates in each network are uniform all-to-all of magnitude r. That is, each node sends each other node in the network traffic at rate r. We further assume that there are N nodes in each network, and, as before, that there are t active wavelength channels (each of unit capacity) per link.

■ 2.2.1 Bidirectional rings

In the following, we assume that shortest path routing is employed, as it is clear that throughput is maximized in bidirectional rings under this routing.

OPS

Under shortest path routing, the average (directed) link load in a bidirectional ring is:

$$\overline{L} = \begin{cases} r\frac{N^2-1}{8}, & \text{if } N \text{ is odd,} \\ r\frac{N^2}{8}, & \text{if } N \text{ is even.} \end{cases}$$

In the case of N odd, there is a unique shortest path connecting each node pair, and each link's load is exactly equal to $\overline{L}$. In the case of N even, owing to the existence of two shortest paths for nodes which are diametrically spaced, we ensure that each link's load is exactly $\overline{L}$ by routing $r/2$ units of traffic along each of the two paths connecting each diametrically spaced node pair.

Since the maximum link load in the network must be less than t (which is each link's unidirectional capacity) and $\overline{L}$ is the exact load on each link, Theorem 2.1 allows us to conclude that any r, such that:

$$r < \begin{cases} \frac{8t}{N^2-1}, & \text{if } N \text{ is odd} \\ \frac{8t}{N^2}, & \text{if } N \text{ is even,} \end{cases}$$

is achievable.

OFS

We now show that the maximum achievable rate r by an OFS architecture is the same as in OPS. We do this by proposing a set of feasible network states for one of the t wavelength channels over which we shall share our time equally, and show that doing so yields the same maximum r as in OPS. Let us number the N nodes in the ring in a clockwise fashion from $1, \ldots, N$, and let us label the bidirectional link between nodes i and $[(i \bmod N) + 1]$ link i.

For the case of N even, we shall associate with each $k = 1, \ldots, N/2$ a set of $N/2$ network configurations, for a total of $N^2/4$ configurations over which we shall time-share equally. For a fixed $k \neq N/2$, let us configure the network as follows:

- Node 1 and node $k+1$ transmit to each other over links $1, 2, \ldots, k$.
- Node $k+1$ and node $N/2+1$ transmit to each other over links $k+1, k+2, \ldots, N/2$.
- Node $N/2+1$ and node $N/2+1+k$ transmit to each other over links $N/2+1, N/2+2, \ldots, N/2+k$.
- Node $N/2+1+k$ and node 1 transmit to each other over links $N/2+1+k, N/2+k+2, \ldots, N$.

By rotating this configuration clockwise by one node $N/2-1$ times, we obtain $N/2$ distinct configurations. For the case of $k = N/2$, each of the $N/2$ configurations involves a different pair of diametrically spaced nodes transmitting to each other twice—once in each of the clockwise and counter-clockwise directions. Now, considering the ensemble of $N^2/4$ configurations, we see that each node pair transmits to each other exactly twice. Thus, by time-sharing equally over these $N^2/4$ configurations, the fraction of time that each node pair transmits to each other is $8/N^2$. For a network with t wavelength channels, this implies that any r less than $8t/N^2$ is achievable, as in OPS.

For the case of N odd, a very similar analysis to the above implies that any r less than $8t/(N^2-1)$ is achievable, as in OPS. We omit the analysis for the sake of brevity.

OBS

In order to determine the maximum achievable r in an OBS network, we employ the two OBS models of the previous section and numerically maximize their expressions.

Figure 2-8 illustrates the maximum session rate performance as a function of network size for the different switching architectures. The most immediate observation from the figure is that the OPS and OFS architectures significantly outperform the OBS architecture (under both OBS models). Another observation is that, owing to the sparse connectivity of the ring architecture, the capacity performance falls sharply as the number of nodes increases, especially for OBS. Finally, our second OBS model outperforms the first, as expected. This is a result of source nodes choosing wavelength channels more intelligently in the second model. By doing so, there is a traffic "smoothing" effect when the number of wavelength channels is increased[10].

[10]An analogous result is evident in the finite buffer $M/M/m/m$ queueing system: when the number of servers m is increased, while increasing the traffic commensurately, a decreasing blocking probability is achieved.

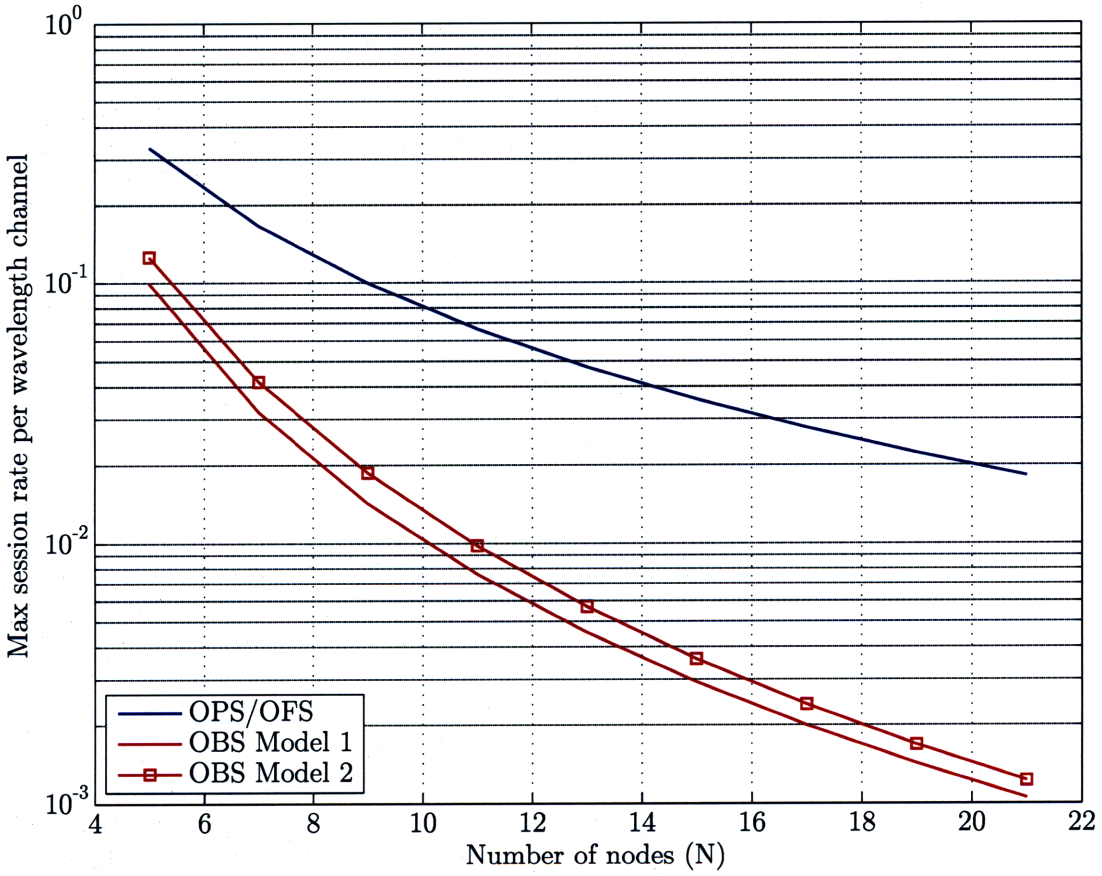


Figure 2-8. Maximum session rate (r) normalized by the number of wavelength channels versus number of network nodes for the bidirectional ring. Uniform all-to-all traffic and $t = \lfloor N/2 \rfloor$ wavelength channels are assumed.

■ 2.2.2 Moore Graphs

As discussed in Appendix B in more detail, Moore Graphs are a family of graphs which achieve the Moore bound—an upper bound on the number of nodes in a graph given a diameter and node degree. These graphs have been previously shown to minimize switching and fiber costs in optical networks under uniform all-to-all traffic conditions [123]. These attractive properties of Moore Graphs motivate our present consideration of this family of graphs. It should be noted, however, that aside from degenerate instances (e.g., complete graphs, rings), Moore Graphs seldom exist.

OPS

For a Moore Graph with diameter d and degree Δ, it can be shown that shortest path routing achieves an average link load of:

$$\overline{L} = r \sum_{i=1}^{d} i(\Delta - 1)^{i-1}.$$

Owing to the symmetry of Moore Graphs and the uniqueness of the shortest path, the network load is also perfectly distributed among all of the links. Thus, by Theorem 1, any r less than $tr/\overline{L}$ is achievable.

OFS

In [124], it was shown that in order to achieve uniform all-to-all traffic of rate $r = 1$ without wavelength conversion, the number of wavelength channels t required is bounded as follows:

$$\sum_{i=1}^{d} i(\Delta - 1)^{i-1} \leq t_{\min} \leq 1 + \sum_{i=1}^{d} i(\Delta - 1)^{i-1}$$

or equivalently:

$$\frac{h_{\mathscr{M}}(n_m - 1)}{\Delta} \leq t_{\min} \leq 1 + \frac{h_{\mathscr{M}}(n_m - 1)}{\Delta}, \tag{2.3}$$

where $h_{\mathscr{M}}(\cdot, \cdot)$ is the average shortest path distance of the Moore Graph with degree Δ and diameter d. Let us now consider one wavelength channel in isolation. By time-sharing equally over the configurations of individual channels used in the scheme proposed in [124], a per session rate arbitrarily close to $(t_{\min})^{-1}$ is achievable on each wavelength channel. This implies that for a network with t wavelength channels a traffic rate $t/t_{\min}$ is achievable, and is thus a lower bound for the maximum achievable rate r.

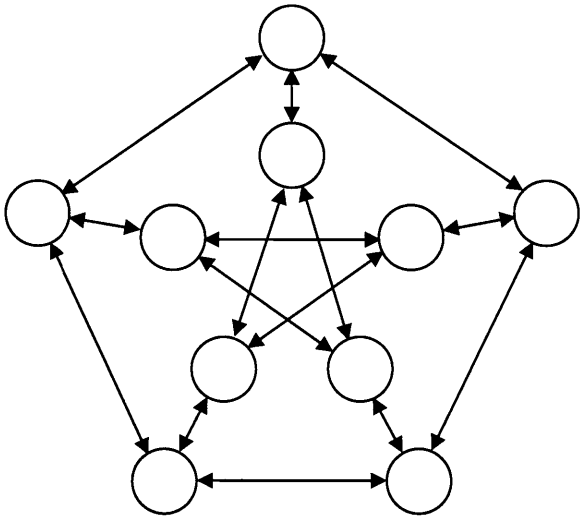


Figure 2-9. Illustration of the Petersen graph, a 10 node Moore Graph with degree 3 and diameter 2.

OBS

As in the case of bidirectional rings, the two OBS models in the previous section provide a basis for a numerical maximization of r.

Example 2.3 (Petersen graph) *Consider the Petersen graph drawn in Figure 2-9, which is a 10 node Moore Graph with degree 3 and diameter 2. Assuming uniform all-to-all traffic of magnitude r, the previous discussion implies that, under an OPS architecture, any $r < t/5$ can be achieved, and that under an OFS architecture r is lower bounded by $t/6$. In fact, for the special case of the Petersen graph, the work in [124] implies that any $r < t/5$ can be achieved. These results, along with numerical results for OBS, are summarized in Table 2.1 and Figure 2-10 for a range of wavelength channels t. Note that r, normalized by the number of channels, and depicted in Figure 2-10, is constant except in the case of OBS Model 2 for the same reasons discussed for bidirectional rings.*

We now generalize the above example to Moore Graphs of degree $\Delta = 3$. In Figure 2-11, we plot the maximum r normalized by the number of wavelength channels versus number of network nodes for Moore Graphs of degree $\Delta = 3$ and assuming $t = 5$ wavelength channels. Note that this plot represents actual capacity performance *only if* the Moore Graphs exist. As in the case of the bidirectional ring, the most immediate observation is that the OPS and OFS architectures significantly outperform the OBS architectures. Also, since the node degree remains constant while the number of network nodes increases, the performance falls sharply, especially for OBS.

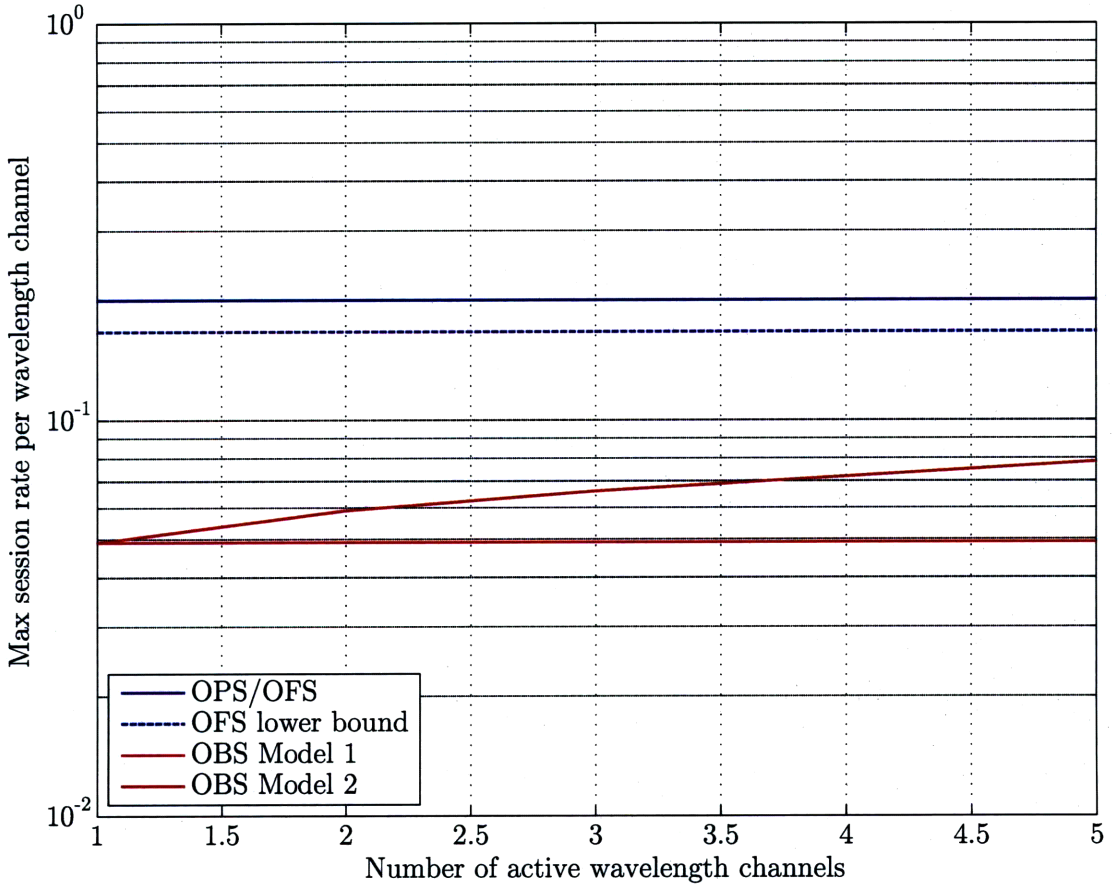


Figure 2-10. Maximum session rate (r) normalized by the number of wavelength channels versus number of wavelength channels for the Petersen graph in Example 2.3.

	$\mathbf{t=1}$	$\mathbf{t=2}$	$\mathbf{t=3}$	$\mathbf{t=4}$	$\mathbf{t=5}$
OPS	0.200	0.400	0.600	0.800	1.000
OFS exact	0.200	0.400	0.600	0.800	1.000
OFS lower bound	0.167	0.333	0.500	0.667	0.833
OBS Model 1	0.049	0.098	0.147	0.196	0.245
OBS Model 2	0.049	0.118	0.198	0.287	0.391

Table 2.1. Maximum session rate r versus number of transceivers t for the Petersen graph considered in Example 2.3.

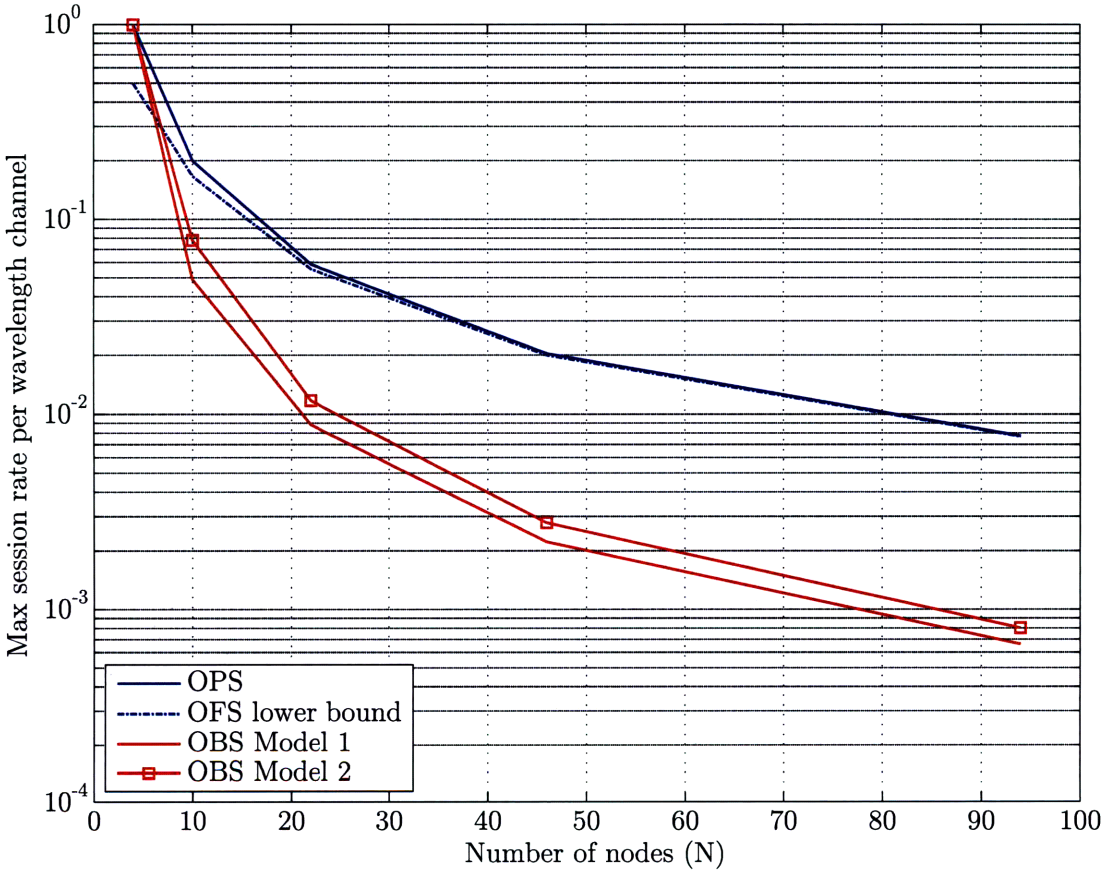


Figure 2-11. Maximum session rate (r) normalized by the number of wavelength channels versus number of network nodes for Moore Graphs. Node degree $\Delta = 3$, and $t = 5$ wavelength channels are assumed.

■ 2.3 Dependence on number of switch ports

In section 2.1, we investigated the capacity performance of OPS, OFS and OBS assuming equal switch port counts in optical packet switches and OXCs at core nodes. Specifically, we assumed that each fiber could support a maximum of w unit capacity active wavelength channels, and that each node is equipped with $t = w$ transceivers per fiber, one for each wavelength channel. This assumption led us to the result that the capacity region of OPS dominates that of OFS, and that the capacity region of OFS dominates that of OBS.

In this section, we investigate the capacity regions of the optical network architectures as a function of the number of switch ports available at core node switching devices. This generalization is motivated by:

1. The incommensurate complexity and cost of comparable OPS, OFS and OBS architectures. For example, the present cost of an all-optical logic gate, a building block of OPS networks, is several orders of magnitude more than the cost of an electronic logic gate, the analogous building block for OFS and OBS networks. We more fully address the issue of cost in Chapter 5.

2. The relevance of this scenario to MANs—which are admittedly beyond the scope of this chapter—where only a fraction of the possible wavelength channels may be lit with data.

As before, we assume that each node generates (terminates) a maximum of t unit capacity wavelengths of traffic per outgoing (incoming) fiber, and hence has t tunable transceivers per fiber. However, depending upon the architecture, we may permit a larger number of wavelengths w to be carried on a fiber and switched at nodes.

■ 2.3.1 OPS networks

As in section 2.1.1, we assume that optical packet switches may switch up to t wavelengths of traffic per fiber. Thus, each fiber may carry a maximum of t unit capacity active wavelength channels. The capacity region of the network is, as discussed at the end of section 2.1.1, given by the following corollary of Theorem 2.1:

Corollary 2.2 *The capacity region of an OPS network with t unit capacity active wavelength channels is given by the admissibility constraints which state that no link may be subscribed beyond rate t.*

Remark: As discussed at the end of section 2.1.1, this result holds for OPS networks with and without wavelength conversion capability.

■ 2.3.2 OFS networks

In OFS, we allow the OXCs at core nodes to switch $w \geq t$ wavelength channels per fiber. We allow a larger number of switch ports in the OFS architecture because

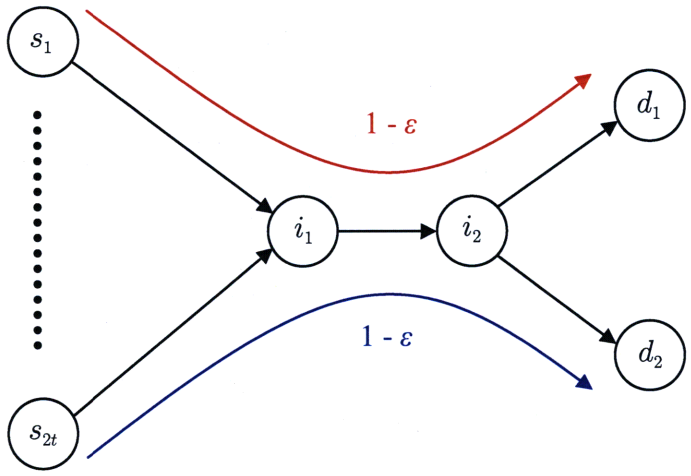


Figure 2-12. Illustration of the network considered in Example 2.4.

OFS core nodes are much simpler, and thus cheaper, to build than OPS core nodes. As we see in the following example, this relaxation in the number of OXC switch ports allows for certain traffic rates to be achievable in OFS networks but not in OPS networks.

Example 2.4 (Bottleneck) *Consider the network drawn in Figure 2-12, where nodes* $s_1,\ldots,s_t$ *each send data at rate* $1-\varepsilon$ *to node* d_1 *and nodes* $s_{t+1},\ldots,s_{2t}$ *each send data at rate* $1-\varepsilon$ *to node* d_2 *via the intermediate nodes* i_1 *and* i_2*. Since the offered load to the link between* i_1 *and* i_2 *is close to* $2t$ *when* ε *is very small, this traffic pattern cannot be serviced with an OPS architecture. However, an OFS architecture can accommodate this traffic provided that* $w \geq 2t$*.*

In characterizing the capacity region of OFS under this relaxed assumption, we first note that a feasible network state corresponds to an ensemble of w stable sets—one for each wavelength—subject to the constraint that no more than t flows per fiber may originate or terminate at a network node. We point out that these w stable sets may be identical, and that they, furthermore, may belong to different conflict graphs altogether as it is certainly possible to route transactions differently over different wavelength channels. Let us now define the incidence vector of a w-stable set of a network as the sum of the incidence vectors of w stable sets of (possibly different) conflict graphs of the network. Then, by similar reasoning as in section 2.1.2, we have the following:

Corollary 2.3 *The capacity region of an OFS network with* $w \geq t$ *is the convex hull of all possible* w*-stable set incidence vectors of the network, subject to the constraint that the* w*-stable sets result in no more than* t *flows per fiber originating or terminating at a network node.*

Remark: The above corollary assumes that the network does not have wavelength conversion capability. However, wavelength conversion may be handled as discussed at the end of section 2.1.2.

■ 2.3.3 OBS networks

By similar reasoning as in section 2.1.3, the capacity region of OBS is dominated by that of OFS.

The development of an approximate throughput analysis of OBS under the assumption that OBS core nodes have $w \geq t$ ports per fiber resembles that of section 2.1.3. In particular we generalize the two OBS models presented in that section. As before, the probability that a given burst of type j is successfully received at its destination h_j hops away is:

$$P_s(j) = \prod_{i=1}^{h_j} \Pr(\text{success on hop } i).$$

The average queueing delay of a burst type, and the throughput of the link i are also the same as in section 2.1.3:

$$W_j = \left(\frac{1}{P_s(j)} - 1\right)(\overline{L} + q)$$

$$S(i) = \sum_{j=1}^{F} P_s(j) \frac{\overline{L}}{\overline{L} + q} I_i(j),$$

where F is the number of types of bursts in the network, and $I_i(j)$ is the indicator function that has a value of unity if a burst of type j traverses link i.

Generalization of OBS Model 1

Employing previous notation, the probability that a particular burst is successfully carried on its first hop is:

$$\Pr(\text{success on hop 1}) = \sum_{l=0}^{\min(t, b_{s,1})-1} \sum_{u=l}^{l+b_{p,1}} \left(\frac{w-1}{w}\right)^u B\left(b_{s,1} - 1, l, \frac{\overline{L}}{q + \overline{L}}\right) B\left(b_{p,1}, u - l, \frac{\overline{L}}{q + \overline{L}}\right),$$

where $\frac{\overline{L}}{q+\overline{L}}$ denotes the probability that a burst of a particular type is attempting transmission. The probability that the burst is successfully carried on hop h ($2 \leq h \leq h_j$) is similarly given by:

$$\Pr(\text{success on hop } h) = \sum_{l=0}^{\min(t,b_{s,h})} \sum_{u=l}^{l+b_{p,h}-1} \left(\frac{w-1}{w}\right)^u B\left(b_{s,h}, l, \frac{\overline{L}}{q+\overline{L}}\right) B\left(b_{p,h}-1, u-l, \frac{\overline{L}}{q+\overline{L}}\right).$$

In the above two expressions, the summation index u represents the total number of burst types which are attempting transmission on the link at that instant in time, and the summation index l represents the number of those burst types that originate at the link.

Generalization of OBS Model 2

In generalizing the second analytical OBS model, $t = w$ is substituted into equations (2.1) and (2.2). The probability that the burst is successfully carried on its first hop is given by:

$$\Pr(\text{success on hop } 1) = \sum_{x=0}^{b_{p,1}} \sum_{y=0}^{\min(w-1,x)} \sum_{z=y}^{\min(w,b_{s,1}+y,y+t)-1} U(w,y,x) B\left(b_{p,1}, x, \frac{\overline{L}}{\overline{L}+q}\right) B\left(b_{s,1}-1, z-y, \frac{\overline{L}}{\overline{L}+q}\right),$$

and the probability that the burst is successfully carried on hop h ($2 \leq h \leq h_j$) is given by:

$$\Pr(\text{success on hop } h) = \sum_{x=0}^{b_{p,h}-1} \sum_{y=0}^{\min(w-1,x)} \sum_{z=y}^{\min(w,b_{s,h}+y,y+t+1)-1} \left(\frac{w-z}{w}\right) U(w,y,x) B\left(b_{p,h}-1, x, \frac{\overline{L}}{\overline{L}+q}\right) B\left(b_{s,h}, z-y, \frac{\overline{L}}{\overline{L}+q}\right),$$

where the upper limits on the innermost summations now account for the fact the number of wavelength channels and transceivers can be different. The interpretations of these equations are otherwise identical to those of equations (2.1) and (2.2).

■ 2.3.4 On the relationship between the number of wavelength channels and the number of transceivers

Our last result relates the capacity regions of OPS, OFS, and OBS according to the relationship between w and t:

Theorem 2.4 *1. If $w = t$, then the capacity region of OPS dominates that of OFS, and the capacity region of OFS dominates that of OBS.*

2. *If $w > t$, then it is possible for the capacity regions of OFS and OBS to contain traffic vectors not in the capacity region of OPS. Likewise, it is possible for the capacity region of OPS to contain traffic vectors not in the capacity regions of OFS and OBS.*

3. *If $w - t$ is sufficiently large, then the capacity region of OFS dominates the capacity region of OPS. If there are dedicated wavelength receivers in each fiber (to avoid receiver collisions), then the capacity region of OBS is identical to that of OFS.*

Proof. (1) follows from Theorem 2.1, Theorem 2.3, and Corollary 2.1.

To show (2), consider the five node ring illustrated in Figure 2-7 with $t = 10$ and $w = 11$. For an example of a traffic vector that can be accommodated by OPS but not by OFS or OBS, let the five flows illustrated in Figure 2-7 each have rate $5 - \varepsilon$, where ε is very small. In order for this traffic vector to be accommodated, links must carry an average of $10 - 2\varepsilon$ wavelengths of traffic. However, even with 11 wavelengths available on each fiber, this is impossible to support with an OFS architecture, and hence, an OBS architecture. On the other hand, the traffic vector is clearly admissible, and thus serviceable with an OPS architecture. For an example of a traffic vector that can be accommodated by OFS and OBS but not OPS, define flows of rate $5 - \varepsilon$ and $6 - \varepsilon$ between nodes 1 and 5, and nodes 2 and 4, respectively. Since $11 - 2\varepsilon$ units of traffic pass through node 3, this traffic vector can be supported by an OFS and an OBS architecture but not by an OPS architecture (when ε is very small).

To prove (3), we show that any admissible OPS traffic vector can be accommodated by an OFS network with a large enough number of wavelengths. When there are at least as many wavelengths as there are flow types, then the OFS network may be viewed as a large nonblocking switch, in which the set of all tunable transmitters in the network correspond to the switch's input ports and the set of all tunable receivers represent the switch's output ports. By the results of [112], there exist flow-based scheduling algorithms that are rate-stable for the set of admissible traffic rates. If we additionally assume that there are sufficiently many receivers per fiber at destination nodes to avoid receiver collisions, then the capacity region of OBS is identical to that of OFS. This is because, provided that we can assign each burst type its own wavelength, once a burst enters the network it is guaranteed to reach its destination without collision. If, however, we assume that receiver collisions are possible, then it is no longer true that the OBS capacity region is equivalent to the OFS capacity region. □

Remark: Theorem 2.4 provides an indication of the relative performance of the OPS and OFS architectures if the costs of the architectures are made comparable while holding the number of transceivers per fiber t constant. When switch ports have commensurate costs in OPS and OFS core nodes, then the capacity performance of OPS dominates that of OFS. However, when switch ports in OPS core nodes are far more

expensive than in OFS core nodes, then the converse is true: OFS outperforms OPS. Finally, when switch port cost in OPS core nodes is only moderately more expensive than in OPS core nodes, then it is possible that neither architecture dominates.

■ 2.4 Conclusion

In this chapter, we employed a framework based upon network capacity—the set of exogenous traffic rates that can be stably supported under operational constraints—to analyze the performance of OFS in the WAN. We emphasize that the work presented in this chapter constitutes a "best-case" throughput comparison among optical network architectures, in that: i) a tolerance to unbounded delay is implied throughout, and ii) any capacity inefficiencies arising from coupling with MAN architectures are neglected. Our analysis was constructive in that algorithms were outlined which achieve the OFS capacity limits. More practical algorithms which account for reconfiguration latency of hardware were also discussed. Unfortunately, these algorithms are generally NP-complete, and will therefore be difficult to implement for inter-MAN OFS communication, as we discuss further in section 4.2. These algorithms, however, may be appropriate for intra-MAN OFS communication, since this scheduling problem may be sensibly decoupled from that of inter-MAN OFS communication on fine time-scales.

Our study was comparative in that we compared the capacity performance of OFS to that of OPS and OBS. We showed that, under the assumption of an equal number of switch ports per fiber at core nodes, the capacity region of OPS dominates that of OFS, and that the capacity region of OFS dominates that of OBS. These differences in capacity performance arose because of the benefits of core buffering and scheduling. We also applied these results to two important families of graphs—bidirectional rings and Moore Graphs—under uniform all-to-all traffic and observed that the performances under OPS and OFS were the same or almost the same, while the performance under OBS was significantly worse.

Motivated in part by the incommensurate cost of comparable transport architectures, we investigated the dependence of relative capacity performance on the number of switch ports per fiber at core nodes. When this number is significantly larger in OFS than in OPS, we found that OFS outperforms OPS. This can be attributed to the fact that OFS exploits its higher capacity network core in spite of its lack of core buffering. This is a useful result because core routers are more expensive than OXCs with the same number of ports operating at the same line-rates. Finally, we showed that when the number of switch ports per fiber in core nodes is only moderately larger in OFS than in OPS, then it is possible that neither OPS nor OFS dominates.

Finally, as mentioned at the outset of this thesis, the goal of the network designer should be to determine which network architecture meets end-user requirements with the minimum cost. While we were motivated in section 2.3 by the importance of cost in assessing a network, a detailed consideration of the optical transport net-

work architectures in this respect is deferred to Chapter 5. As we shall see there, a performance-cost tradeoff casts OFS in a more positive light than a strict performance comparison.

A notable avenue for extending the work in this thesis is a generalization to include multicasting. The capacity regions of individual OXCs with multicast capability were recently investigated in [281]. However, the characterization of the capacity regions of networks of such switches, along with efficient algorithms with which to operate these networks, is an important open problem. Indeed, results in this area could be very useful for video content distribution, an increasingly important network application.

2.A Appendix

2.A.1 Further notes on the stability of OFS

In this appendix, we discuss stronger forms of stability than rate-stability, which was the definition of stability employed in this chapter. We also show that OFS networks are strongly stable in the interior of their capacity regions.

Stability definitions

In the following stability definitions, any norm definition can be used.

Definition 2.13 *A system of queues is rate-stable if:*

$$\lim_{n\to\infty} \frac{X_n}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} (E_i - D_i) = 0, \text{ with probability 1.}$$

Note that rate-stability allows queue lengths to indefinitely grow with sublinear rate.

Definition 2.14 *A system of queues is weakly stable if, for every $\epsilon > 0$, there exists a $B > 0$ such that:*

$$\lim_{n\to\infty} Pr(\|X_n\| > B) < \epsilon.$$

Note that weak stability implies that the servers at the queues are able to serve all of the offered transactions. To see this, note that for a transaction to never be served with nonzero probability, say $\hat{\epsilon}$, it is necessary (under a FIFO policy for simplicity) for the transaction to arrive at a queue which is infinite in size. However, weak stability guarantees that for sufficiently large n, the probability that the queue size exceeds a finite size B (which depends on $\hat{\epsilon}$) is less than $\hat{\epsilon}$, which is a contradiction. Hence, weak stability implies that every transaction must be served. However, weak stability is not sufficiently strong to ensure that the delay experienced by every customer is bounded.

Definition 2.15 *A system of queues is strongly stable if:*

$$\limsup_{n\to\infty} \mathbf{E}[\|X_n\|] < \infty.$$

Note that, by virtue of Little's Theorem, strong stability implies boundedness of average customer delays.

Strong stability implies weak stability, and weak stability implies rate-stability [201]. Some definitions of stability, such as the one used in [20, 275], imply that the discrete time Markov chain underlying the state of the queues is positive recurrent. Note that this definition of stability implies weak stability, and hence rate-stability [187].

Proving stability: Fluid limits and Lyapunov drift

In the work done in [74] which uses fluid limits to prove stability, the only form of stability of the original network that can be ascertained is that of rate-stability. However, as shown in [199, 275], by relaxing the initial condition of the derived fluid system, a stronger form of stability—weak stability—can be proven.

Lyapunov drift arguments are often used to prove strong stability of systems of queues. Specifically, the following result is often invoked to prove strong stability:

Theorem 2.5 ([176, 206]) *Let X_n be a Q-dimensional Markov chain, whose elements are nonnegative integers. If there exists a nonnegative valued function $\{\mathscr{L}(\cdot) : \mathbb{N}^Q \to \mathbb{R}_+\}$, which we call a Lyapunov function, such that:*

$$\mathbf{E}\left[\mathscr{L}(X_{n+1}) - \mathscr{L}(X_n)|X_n\right] < \infty$$
$$\limsup_{\|X_n\|\to\infty} \frac{\mathbf{E}\left[\mathscr{L}(X_{n+1}) - \mathscr{L}(X_n)|X_n\right]}{\|X_n\|} < -\epsilon$$

for some $\epsilon > 0$, where $\|X\|$ is the Euclidean norm of vector X, then the Markov chain $\{X_n\}_{n=0}^{\infty}$ is positive recurrent and the system of queues is strongly stable.

OFS and strong stability

We now show that strong stability can be achieved within the OFS capacity region. For simplicity, we shall consider nonwork-conserving adaptations of a particular family of MaxWeight scheduling policies whose maximum weight at time slot n is given by:

$$\mathscr{W}^*(X_n) = \max_{D\in\Pi}\{{X_n}^{\beta} D^T\}, \text{ with } \beta \geq 1$$

where D is any departure vector belonging to the set of feasible network states Π, and the exponentiating operation applies component-wise. Recall that in an OFS network, we do not use the maximum weight at each time slot, although we ensure

that the weight of the configuration used is within a bounded constant K of the maximum weight.

The Lyapunov function that we employ to prove strong stability of OFS networks is:

$$\mathscr{L}(X_n) \equiv \|{X_n}^{\beta+1}\|_1 = \sum_{i=1}^{Q} (x_n^i)^{\beta+1},$$

where, $\|X\|_1$ is the sum of the components of vector X. Now:

$$\begin{aligned}
&\limsup_{\|X_n\|\to\infty} \frac{\mathbf{E}\left[\mathscr{L}(X_{n+1}) - \mathscr{L}(X_n)|X_n\right]}{\|X_n\|} \\
&= \limsup_{\|X_n\|\to\infty} \frac{\mathbf{E}\left[\|{X_{n+1}}^{\beta+1}\|_1 - \|{X_n}^{\beta+1}\|_1|X_n\right]}{\|X_n\|} \\
&= \limsup_{\|X_n\|\to\infty} \frac{\mathbf{E}\left[\|(X_n + A_n - D_n)^{\beta+1}\|_1 - \|{X_n}^{\beta+1}\|_1|X_n\right]}{\|X_n\|} \\
&= \limsup_{\|X_n\|\to\infty} \frac{\mathbf{E}\left[[(X_n + A_n - D_n)^{\beta} - {X_n}^{\beta}]{X_n}^T + (X_n + A_n - D_n)^{\beta}(A_n - D_n)^T|X_n\right]}{\|X_n\|} \\
&= \limsup_{\|X_n\|\to\infty} \frac{\mathbf{E}\left[{X_n}^{\beta}(A_n - D_n)^T|X_n\right]}{\|X_n\|}, \text{ for } \|X_n\| \text{ sufficiently large} \\
&= \limsup_{\|X_n\|\to\infty} \frac{{X_n}^{\beta}\left(\mathbf{E}[A_n] - \mathbf{E}[D_n|X_n]\right)^T}{\|X_n\|},
\end{aligned}$$

where the penultimate equality follows from the fact that A_n and D_n are binary vectors and thus contribute negligibly as $\|X_n\| \to \infty$. Now, since $\mathbf{E}[A_n]$ lies strictly within the OFS capacity region, there exists an $\alpha > 1$, such that $\alpha\mathbf{E}[A_n]$ lies on the boundary of the capacity region. Note that it must be true that:

$$\begin{aligned}
{X_n}^{\beta}\mathbf{E}[A_n]^T &< \alpha {X_n}^{\beta}\mathbf{E}[A_n]^T \\
&\leq {X_n}^{\beta} {D_n^*}^T \\
&\leq {X_n}^{\beta}\mathbf{E}[D_n|X_n]^T + K,
\end{aligned}$$

where:

$$D_n^* = \arg\max_{D\in\Pi}\{{X_n}^{\beta} D^T\}$$

is the departure vector yielding the maximum weight. The penultimate inequality follows from the fact that, roughly speaking, $\alpha\mathbf{E}[A_n]$ is a weighted average of all the feasible D_n and must therefore result in a suboptimal weight. The last inequality follows from the fact that the weight used by the OFS scheduling policy is at most a

bounded constant K away from the optimal weight. The above inequalities yield:

$$\begin{aligned}\limsup_{\|X_n\|\to\infty} \frac{\mathbf{E}\left[\mathscr{L}(X_{n+1}) - \mathscr{L}(X_n)|X_n\right]}{\|X_n\|} &\leq \limsup_{\|X_n\|\to\infty} \frac{(1-\alpha){X_n}^{\beta}\mathbf{E}[A_n]^T + K}{\|X_n\|} \\ &\leq (1-\alpha)a_{\min}\frac{\|{X_n}^{\beta}\|_1}{\|X_n\|} \\ &\leq (1-\alpha)a_{\min} \equiv 2\epsilon \\ &< 0,\end{aligned}$$

where $a_{\min}$ is the minimum non-null component of the vector $\mathbf{E}[A_n]$. This fulfills the second condition of Theorem 2.5. The fact that $\mathbf{E}[A_n {A_n}^T] < \infty$ trivially guarantees the first condition of the theorem. Thus, the system of queues is strongly stable.

■ 2.A.2 Algorithm in proof of Theorem 2.2

We now detail the algorithm in the proof of Theorem 2.2 which transforms probability vector Φ, where $v(\Phi) \geq \Lambda$, to probability vector Φ', such that $v(\Phi') = \Lambda$, and where Λ is some achievable traffic rate vector. The algorithm works by sequentially transforming the elements of the rate vector $v(\Phi)$ to the corresponding elements in Λ (assuming that these elements differ), while ensuring that the evolving vector is the result of a convex combination of stable set incidence vectors which correspond to feasible network states.

We now detail one iteration of the algorithm. Assume that the j^{th} element of $v(\Phi)$, denoted v_j and corresponding to the rate of flow j, is strictly greater than the corresponding element λ_j of Λ. Recall that the SSS policy Φ corresponds to a convex combination of stable set incidence vectors which sum to $v(\Phi)$. We now restrict our attention to the incidence vectors in this convex combination which correspond to serving flow j (i.e., incidence vectors in this convex combination whose j^{th} element is '1'). We denote the number of such vectors by k. Let us associate with each of these k incidence vectors an index i, and let us denote their corresponding convex coefficients α_i, $i = 1, \ldots, k$. These k incidence vectors, along with their corresponding convex coefficients, form a subconvex (i.e., $\sum_{i=1}^{k} \alpha_i \leq 1$) sum Λ_j whose j^{th} element is v_j. We wish to transform Λ_j into another vector Λ'_j whose j^{th} element is λ_j but is otherwise identical to Λ_j. We define i_t as the minimum index such that $\sum_{i=1}^{i_t} \alpha_i > \lambda_j$. We now modify the incidence vectors $i_t, \ldots, k$ and their convex coefficients in the following way, such that the rate of flow j equals λ_j:

1. We change the j^{th} element of incidence vectors $i_t + 1, \ldots, k$ from '1' to '0'. The convex coefficients of these modified incidence vectors remain the same.

2. We change the convex coefficient of incidence vector i_t to $\alpha^a_{i_t} = \lambda_j - \sum_{i=1}^{i_t-1} \alpha_i$.

3. We introduce a new incidence vector identical to incidence vector i_t, except that the j^{th} element is '0' instead of '1'. The coefficient associated with this new incidence vector is $\alpha^b_{it} = \alpha_{i_t} - \alpha^a_{it}$.

It is easy to see that the modified vectors resulting from steps 1) and 3) are still stable set incidence vectors since they correspond to subsets of feasible network states. Furthermore, the subconvex combination resulting of the above modified vectors and coefficients exactly equals Λ'_j, as desired. Finally, the sum of the new set coefficients remains $\sum_{i=1}^{k} \alpha_i$. This implies that, after combining these modified incidence vectors with the original incidence vectors which did not contribute to flow j's rate, the ensemble of incidence vectors and their coefficients form a convex combination whose j^{th} element is λ_j. Therefore, by repeating the above procedure for each element of $v(\Phi)$ that is strictly greater than the corresponding element in Λ, we form a new probability vector Φ' satisfying $v(\Phi') = \Lambda$.

■ 2.A.3 Computation of U(t,y,x) in section 2.1.3

Consider an urn containing t balls, each colored distinctly. Now draw a single ball from the urn, note its color, and replace it in the urn. Repeat this procedure $x - 1$ times for a total of x draws. $U(t, y, x)$ is the probability that exactly y distinct colors were observed over the course of the x trials. We define $U(t, y, x) = 0$ if the conditions $1 \leq y \leq t$ and $y \leq x$ are violated. Otherwise, we compute $U(t, y, x)$ using the following recurrence relation:

$$U(t,y,x) = \begin{cases} \left(\frac{1}{t}\right)^{x-1}, & \text{if } y = 1, \\ \binom{t}{y}\left(\frac{y}{t}\right)^{x}\left[1 - \sum_{i=1}^{y-1} U(y,i,x)\right], & \text{otherwise.} \end{cases}$$

The case of $y = 1$ is straightforward. For the other case, the intuition behind the above expression is as follows. We first assume that the y distinct colors out of the t possible choices are fixed. Then, with these fixed y distinct colors, we find the probability that, over the x trials, we only select from this set of y colors. This probability is $(y/t)^x$. Note that, in addition to accounting for the probability that *exactly* y colors are observed, this expression also includes the sequences of x trials where *fewer* than y distinct colors are observed. We therefore we need to subtract out these cases with fewer than y distinct colors from the previous expression. The aggregate probability of these undesired events is $\left(\frac{y}{t}\right)^x \sum_{i=1}^{y-1} U(y,i,x)$. Finally, the number of ways of selecting the y distinct colors at the start is $\binom{t}{y}$.

Chapter 3

OFS in the Metro-Area and Access: Physical Layer Design

In this chapter, we address the physical layer design of the OFS data plane in the metro-area and access. Owing to the all-optical nature of data transmission in OFS, we confine our attention to all-optical networking components exclusively[1]. Our objective is to employ a subset of these components in conjunction with an appropriate scheduling algorithm, to be discussed in the next chapter, to meet the operational requirements of OFS in an economically attractive fashion. Indeed, owing to the absence of buffering and OEO conversions in the interior of OFS networks, the economic viability of OFS hinges, in large part, on cost-effective deployment of all-optical networking components in the metro-area and access to carry out optical aggregation of data—while, of course, respecting the stringent physical layer constraints imposed by the architecture. In the metro-area, reconfigurability via (expensive) OXCs is economically justifiable, owing to the large number of end-users supported; whereas (less expensive) broadcast architectures, coupled with reservation/scheduling, are appropriate for the access, where the number of supported end-users is significantly smaller.

All-optical data transmission in OFS, in addition to constraining the data plane building blocks employed to be all-optical, has implications for the logical partitioning of the network. As discussed in section 1.3.3, the boundary between the access and metro environments may be blurred under some optical network architectures. This is indeed the case with OFS, as all-optical data transmission without buffering promotes consolidation of the metro and access environments. The PON architectures discussed in section 1.3.3 involve OEO conversions at the OLT for data aggregation/disaggregation, and have therefore been designed in the absence of coupling with the MAN and WAN, rendering them inappropriate for OFS. In contrast, the network designs proposed in this chapter are intended for the collective region outside of the WAN and exhibit coupling with the WAN.

[1]In long-haul transport, regeneration of optical signals is often required. Presently (2008), electronic regeneration is a more practical alternative to all-optical regeneration. Nevertheless, we omit a consideration of electronic (and optical) regeneration from our present discussion, as we assume that any signal regeneration occurs in the WAN.

A final remark on this chapter's scope: OFS, as explained earlier, is an architecture that is envisioned to exclusively serve high-end end-users with very large bandwidth demands. As detailed in Chapter 5, users with comparatively small bandwidth demands will be more prudently served by architectures such as EPS. Thus, the network designs discussed in this chapter are proposed as partial solutions in the broader context of hybrid networks.

This chapter is organized as follows. In the next section, we outline the modeling assumptions employed throughout this chapter. In section 3.2, we derive high-level physical layer constraints for both inter- and intra-MAN OFS communication. In section 3.3, we begin a transition into a greater level of detail by laying out further physical layer design assumptions for the metro-area and access. In section 3.4, we characterize the physical layer performance for passive DNs with optical amplification at the interface with the MAN. In section 3.5, we propose and analyze a family of DNs which employ optical amplification *within* the DN in order to improve upon the performance of DNs with external amplification. In section 3.6, we address the issue of the pump power required to implement the aforementioned DN designs. Finally, in appendix 3.A.1, we provide background theory on the detection of optically amplified signals which forms the basis of this chapter.

■ 3.1 Modeling assumptions

In its most general form, the task undertaken in this chapter—physical layer design of an OFS MAN and access network with minimal cost—is intractable for a multitude of reasons. The identification of future user service requirements, for instance, is a pertinent but difficult, if not impossible task, per our discussion in section 1.1.1. The metro-area and access, furthermore, generally cannot be optimized in isolation for they exhibit coupling with the wide-area—and this coupling is especially strong in OFS networks which do not possess buffering or wavelength conversion at the MAN-WAN interface. Moreover, the future costs of devices—existing, and yet to be developed—cannot be faithfully forecasted. In this chapter, we therefore consider OFS MAN and access network design under simplifying, though still realistic, assumptions. The intention is to elucidate the qualitative implications of OFS on metro-area and access physical architecture, with a view towards ultimately integrating OFS within a hybrid optical network architecture.

■ 3.1.1 Devices

Many optical networking devices for MANs and access networks are beginning to be deployed commercially and are thus exhibiting somewhat stable cost structures. We shall consequently employ such commercially available technology in our network architectures rather than functionally similar devices that are still in development. In the case of optical amplifiers—discussed in detail in section A.7 of Appendix A—we employ the EDFA, a relatively mature technology, rather than the Raman amplifier

or SOA, which are limitedly deployed or still under development. Much of the design and analysis carried out in this chapter, however, is parametric in nature, and thus amenable to straightforward modification to reflect improvements in device performance and cost. We shall also avoid the use of technologies which are presently economically prohibitive or still in the process of being developed (e.g., wavelength converters, OXCs capable of multicast and tunable waveband switching).

Wavelength-selective switches—discussed in detail in sections A.4 and A.8 of Appendix A—are key components in our OFS MAN designs. Such devices may be broadly categorized as reconfigurable, and thus active in that they require external power for operation (e.g., MEMS-based OXC); or passive in that they require little[2] or no power, and hence statically configured (e.g., AWG). While passive switches are attractive in that they are far less expensive and more reliable than active switches, the combination of their wavelength-selective and static natures results in a static partitioning of wavelength resources among the different outputs. This is an unattractive property for highly dynamic traffic or for traffic that cannot be forecasted accurately at the time of node deployment. Owing to these significant shortcomings of statically configured devices, we omit their consideration as candidate building blocks for the network architectures in this chapter.

■ 3.1.2 Network topology

In our network model, a single WAN connects many MANs employing OFS. An OFS MAN node comprises an OXC with direct connections to adjacent MAN nodes as well as one or more access networks based upon optical DN architectures. The bidirectional links forming these connections are actually implemented with two contradirectional fiber links, as is done in practice. Moreover, we employ separate passive networking components (e.g., taps, PSCs) for up-link and down-link communication in the access environment. The architectural implications of this are: i) design parameters (e.g., tap coupling ratios) could be optimized separately for up-link and down-link communication, and ii) intra-DN communication requires that a signal be routed through the parent MAN node's OXC[3].

Within the physical topology of each MAN, we assume the existence of an embedded regular tree topology with root node located at the WAN edge (see Figure 3-1(a)). Under normal operating conditions, inter-MAN traffic is assumed to be carried solely on the fibers of the links in the embedded tree[4]; whereas the fibers on the links out-

[2]An example of a passive device requiring external power is an AWG, which requires power for temperature control.

[3]As discussed further in section 3.5.4, this allows intra-tributary communication to benefit from optical amplification, thereby potentially allowing more users to be supported in each tributary.

[4]Tree topologies, as discussed earlier, are known to be efficient at routing data among nodes when most data is destined for the root node (i.e., the WAN), which ultimately lends to cost-effective architectures for inter-MAN communication. It is also worth noting that other common topologies, such as stars and buses, are special cases of trees.

side the embedded tree are assumed to carry only intra-MAN traffic. However, in the event of network element failures in the MAN, or significant deviations from expected traffic—considerations beyond the scope of this work—it may be necessary to reroute inter-MAN traffic outside the embedded tree.

While, in principle, the mesh topologies within which these tree topologies are embedded may be arbitrary, we shall assume that they are based upon Moore Graphs[5] (e.g., Figure 3-1(b)), because, as discussed earlier, topologies based upon this family of graphs lend themselves to cost-effective MAN architectures [121]. In Chapter 5, we shall generalize some of the work in [121] to justify our focus on Moore Graphs as the basis for MAN topologies. In reference to Figure 3-1, we denote by Δ the number of bidirectional links connecting the root node to other MAN nodes. Our Moore Graph assumption implies that MAN nodes which are internally located in the embedded tree topology also have Δ links: one link to the parent node, and $\Delta - 1$ links to child nodes. A leaf node in the embedded tree possesses a single link to its parent node, but additionally possesses $\Delta - 1$ links, outside of the embedded tree, to other MAN nodes (e.g., see Figure 3-1(b)). The number of nodes in a MAN, n_m, is thus given by:

$$\begin{aligned} n_m &= 1 + \Delta + \Delta(\Delta - 1) + \cdots + \Delta(\Delta - 1)^{d-1} \\ &= 1 + \Delta \left[\frac{1 - (\Delta - 1)^d}{2 - \Delta} \right], \end{aligned} \tag{3.1}$$

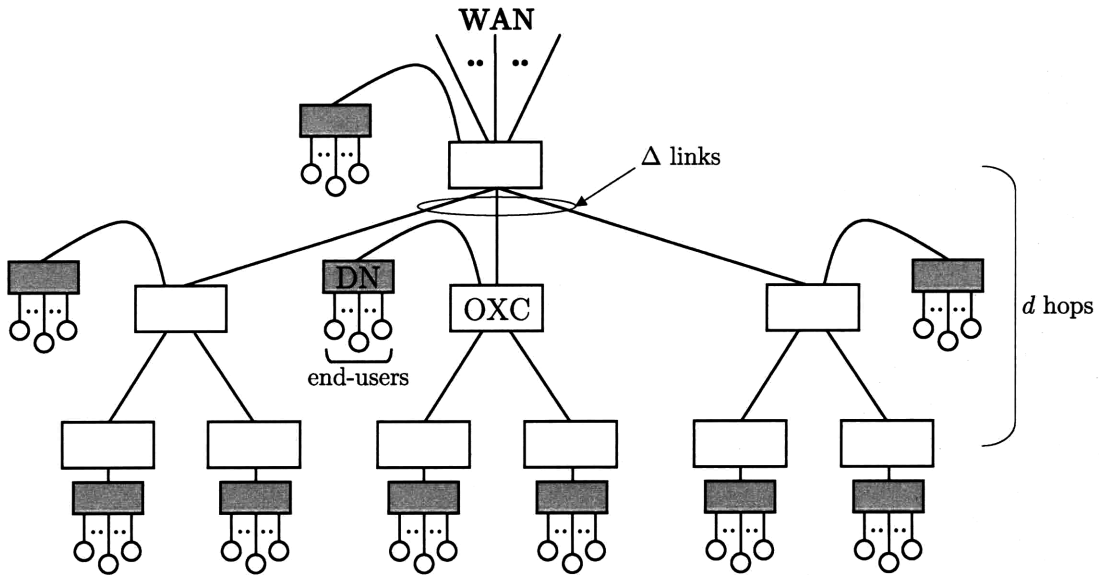
where d denotes the diameter of the embedded tree, and also equals the diameter of the MAN Moore Graph topology.

■ 3.1.3 Amplifier placement and configuration

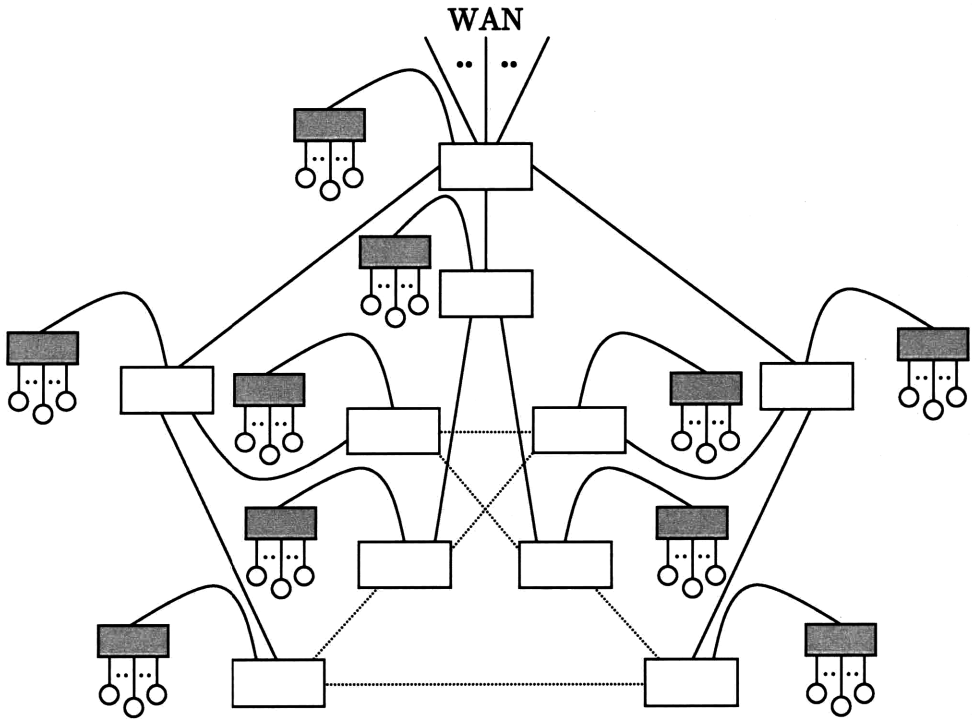
The issue of amplifier placement in optical networks has been addressed extensively in the literature via various problem formulations, all of which are known to be difficult under general conditions. For point-to-point links or linear network topologies—which are good models for WAN links in first-generation optical networks—the merits and optimizations of different amplifier configurations are well understood for a single optical amplifier technology (e.g., see [81, 151]). In [287], dynamic programming and heuristic approaches are applied to linear topologies to determine the combination of optical amplifier technologies which minimizes amplifier cost. For arbitrary meshed topologies, even with a single amplifier technology, amplifier minimization entails a mixed integer nonlinear programming formulation [246], which is difficult to solve efficiently.

For the metro-area, a single amplifier is usually sufficient to compensate for the losses between nodes. However, motivated by cost and reliability concerns, some studies of MAN amplifier placement attempt to further decrease the number of amplifiers. For instance, the minimum number of amplifiers for WDM ring topologies is

[5](Generalized) Moore Graphs are discussed at length in Appendix B.



(a) Embedded tree portion of the OFS MAN.



(b) Mesh OFS MAN based upon the Petersen graph. Note that the tree topology in Figure 3-1(a) is embedded within this topology. Fiber links *not* in the embedded tree are drawn with dotted lines.

Figure 3-1. An example of an OFS MAN based upon a Moore Graph (Petersen graph) with $\Delta = 3$ and $d = 2$. A MAN node (drawn as a white box) comprises an OXC with one or more access DNs (drawn as grey boxes) connected. Optical amplifiers are not drawn.

investigated in [293] via approximate linear and nonlinear programming techniques. With the development of low-cost, reliable, uncooled EDFAs for the metro-area, the importance of marginally reducing the number of EDFAs per network has diminished. In this chapter, we therefore assume that each MAN OXC is compensated by its own EDFA, as is the practice today (see Figure 3-2).

In the access environment, approaches to minimizing the number of optical amplifiers in arbitrary topologies have employed mixed integer linear/nonlinear programming techniques [247, 248], or simulated annealing heuristics [106]. In [57, 210], the assumption of a regular access network topology (e.g., star, bus, tree) permitted analytic tractability. In these studies, however, instead of minimizing amplifiers or their cost, the number of supportable end-users is maximized by appropriately tailoring the passive and active components in the access network. Moreover, remotely-pumped amplifiers are employed in lieu of individually pumped amplifiers as a means of lowering cost. While the general theory of remotely-pumped amplifiers is well-known (e.g., see [81, 279]), the work in [210] specializes it to access networks employing remotely-pumped *distributed* amplification with end-users regularly tapping power from the network. Our approach to access network design is similar: we tailor passive components and remotely-pumped (lumped) amplifiers in our access networks in order maximize the number of supportable end-users, subject to the physical layer constraints imposed by end-to-end OFS communication.

■ 3.2 High-level physical layer design

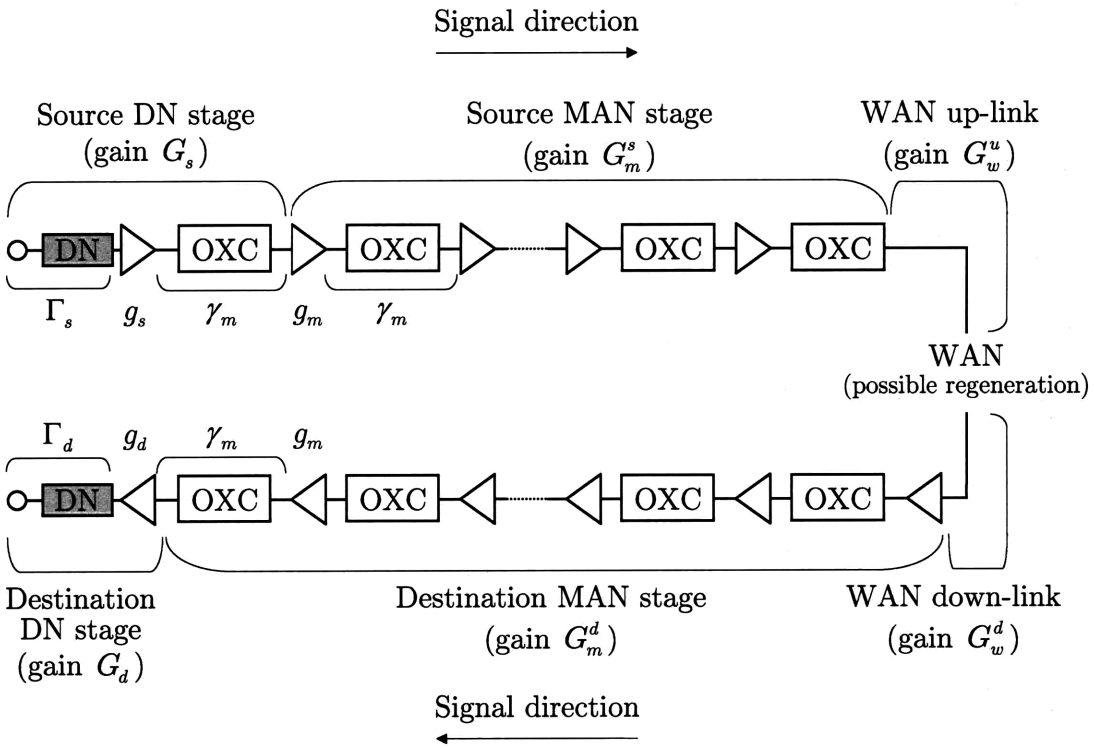
Based upon the above assumptions, and as illustrated in Figure 3-2, the maximum number OXCs in a MAN that an inter- or intra-MAN optical signal passes through is $d + 1$. Intra-MAN signals, furthermore, pass through two DNs in the same MAN, whereas inter-MAN signals pass through one DN in each MAN. Depending on whether and where signal regeneration occurs in the WAN, either inter- or intra-MAN communication may impose the more stringent physical layer requirements. We shall therefore consider both scenarios in the following.

■ 3.2.1 Inter-MAN communication

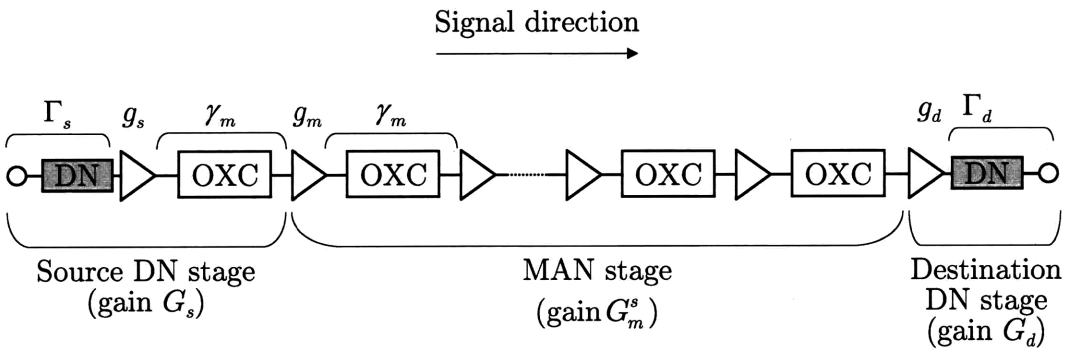
To address the impact of inter-MAN OFS communication on the physical layer design of the MAN and DN, we separately analyze the following segments of inter-MAN communication:

1. The up-link segment, comprising the portion of the network beginning from the source end-user's transmitter, through the source MAN, and ending at the first point of regeneration (if any) in the WAN.

2. The down-link segment, comprising the portion of the network beginning from the last point of regeneration (if any) in the WAN, through the destination MAN, and ending at the destination end-user's receiver.



(a) Inter-MAN OFS communication.



(b) Intra-MAN OFS communication.

Figure 3-2. Illustration of the OFS physical layer for inter- and intra-MAN communication.

These two physical layer segments are drawn in Figure 3-2(a).

Equation (3.24) in the appendix, which represents the formula for the overall noise figure of a cascade of optical gain elements, yields the following for the noise figure of the up-link segment:

$$F_n^u = F_{n,s} + \frac{F_{n,m}^s - 1}{G_s} + \frac{F_{n,w}^u - 1}{G_s G_m^s},$$

where $F_{n,s}$, $F_{n,m}^s$, and $F_{n,w}^u$ are the noise figures of the source DN stage, source MAN stage, and up-link portion of the WAN before the first instance of signal regeneration, respectively; and G_s and G_m^s are the overall optical amplifications of the source DN stage, and source MAN stage, respectively. We point out that this equation does not account for thermal noise. Equation (3.27) from the appendix provides us with an upper bound on this noise figure in terms of lower level system parameters:

$$F_n^u \leq \frac{\eta P_{\max}}{\mathcal{Q}^2 \overline{h} \nu \Delta f},$$

where we refer the reader to the chapter appendix for an explanation of the parameters in this formula. We may now combine the previous two relationships to obtain the following bound:

$$F_{n,s} + \frac{F_{n,m}^s - 1}{G_s^m} + \frac{F_{n,w}^u - 1}{G_s G_m^s} \leq \frac{\eta P_{\max}}{\mathcal{Q}^2 \overline{h} \nu \Delta f}. \tag{3.2}$$

To obtain a numerical feel for the above constraint in a typical system, let us assume the following: perfect quantum photodetector efficiency ($\eta = 1$); a maximum launch power per wavelength of $P_{\max} = 5$ dBm ≈ 3 mW, for which both intra- and inter-channel fiber nonlinearities are insignificant; $\mathcal{Q} = 6$, corresponding[6] to a bit error rate (BER) of 10^{-9}; $\nu = 193.5$ THz, corresponding to a wavelength of 1550 nm; and $\Delta f = 20$ GHz, corresponding to a bit rate of 40 Gbps. Let us, furthermore, assume gain "transparency" in the source DN and MAN stages; that is, the optical power amplifications compensate exactly for the optical power losses in each of the stages (i.e., $G_s = G_m^s = 1$). These typical assumptions yield:

$$F_{n,s} + F_{n,m}^s + F_{n,w}^u \lesssim 3.43 \times 10^4 \approx 45 \text{ dB},$$

provided that the received signal power is well above the thermal noise-limited receiver sensitivity P_{sens}^T.

[6]Note that, by employing forward error correction (FEC), the required value of $\mathcal{Q}$ may be reduced, albeit at the expense of channel throughput.

A similar analysis for the down-link segment of inter-MAN communication yields the following bound:

$$F_n^d \equiv F_{n,w}^d + \frac{F_{n,m}^d - 1}{G_w^d} + \frac{F_{n,d} - 1}{G_w^d G_m^d} \leq \frac{\eta P_{\max}}{\mathcal{Q}^2 \overline{h} \nu \Delta f}, \tag{3.3}$$

where $F_{n,d}$, $F_{n,m}^d$, and $F_{n,w}^d$ are the noise figures of the destination DN stage, destination MAN stage, and down-link portion of the WAN from the final instance of signal regeneration, respectively; and G_w^d and G_m^d are the overall optical gains of the WAN down-link segment, and destination MAN stage, respectively. Assuming, as we did above, typical system parameters as well as gain transparency, we have:

$$F_{n,d} + F_{n,m}^d + F_{n,w}^d \lesssim 45 \text{ dB},$$

provided, again, that the received signal power is well above the thermal noise-limited receiver sensitivity.

■ 3.2.2 Intra-MAN communication

To address the impact of intra-MAN OFS communication on the physical layer design of the MAN and DN, we analyze the network segment drawn in Figure 3-2(b). We point out that the analysis in this subsection subsumes the case of intra-MAN communication within the same DN. This is because intra-DN communication requires that a signal be routed to the parent MAN node OXC since, as discussed in section 3.1.2, up-link and down-link communication occur on separate networking components.

Similar to the previous subsection's analysis, we have the following expression for the overall noise figure for intra-MAN communication:

$$F_n^l = F_{n,s} + \frac{F_{n,m}^s - 1}{G_s} + \frac{F_{n,d} - 1}{G_s G_m^s},$$

where $F_{n,s}$, $F_{n,m}^s$, $F_{n,d}$, G_s, and G_m^s are as defined previously. As above, we obtain the following bound:

$$F_{n,s} + \frac{F_{n,m}^s - 1}{G_s} + \frac{F_{n,d} - 1}{G_s G_m^s} \lesssim \frac{\eta P_{\max}}{\mathcal{Q}^2 \overline{h} \nu \Delta f}, \tag{3.4}$$

which, for gain transparency and typical system parameters, reduces to:

$$F_{n,s} + F_{n,m}^s + F_{n,d} \lesssim 45 \text{ dB},$$

provided, again, that the received signal power is well above the thermal noise-limited receiver sensitivity.

■ 3.3 Detailed physical layer design modeling assumptions

In the following two sections, we consider alternative physical layer designs for DNs. For concreteness, our discussion will take place in the context of bus-based physical topologies, along which k tributaries exist (e.g., see Figure 3-3). Each of these tributaries supports the same number of end-users via a topology which is itself a simple DN (e.g., star, tree, bus). Note that an additional group of end-users—greater in number than a tributary—may be supported beyond the tributary furthest from the DN head-end. This family of topologies that we choose to focus on is, of course, just one of numerous candidates, and similar analyses as the following could be carried out for these other candidate topologies. The insights gained from our analysis of this family of topologies, however, are valuable in that they carry over to arbitrary topologies.

In the following two sections, we treat the physical layer design of source and destination DNs separately. Recall from section 3.1.2 that since up-link and down-link DN communication occurs on separate fibers, a group of end-users may be configured differently for up-link and down-link communication.

In the remainder of this chapter, we assume the MAN physical layer to be as illustrated in Figure 3-2; namely, a lumped EDFA is positioned just prior to the input ports of each OXC. The gain g_m furnished by each such EDFA compensates exactly for the insertion loss of the OXC as well as for the fiber run immediately following it. Thus, gain transparency is achieved in the MAN (i.e., $G_m^s = G_m^d = 1$). Since a signal traversing a MAN travels, in the worst case, d hops, equations (3.23) and (3.24) yield:

$$F_{n,m}^s = \frac{2dn_{\text{sp}}^m (g_m - 1)}{g_m} + 1 \approx 2dn_{\text{sp}}^m + 1 \tag{3.5}$$

$$F_{n,m}^d \approx 2(d+1)n_{\text{sp}}^m + 1, \tag{3.6}$$

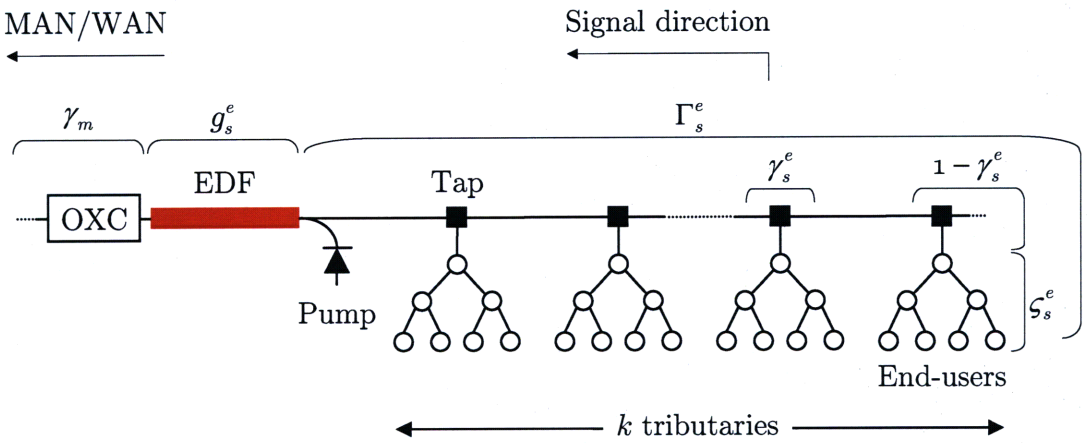
where n_{sp}^m is the spontaneous emission factor of each MAN stage EDFA, and we have assumed that $g_m \gg 1$.

■ 3.4 DNs with external amplification

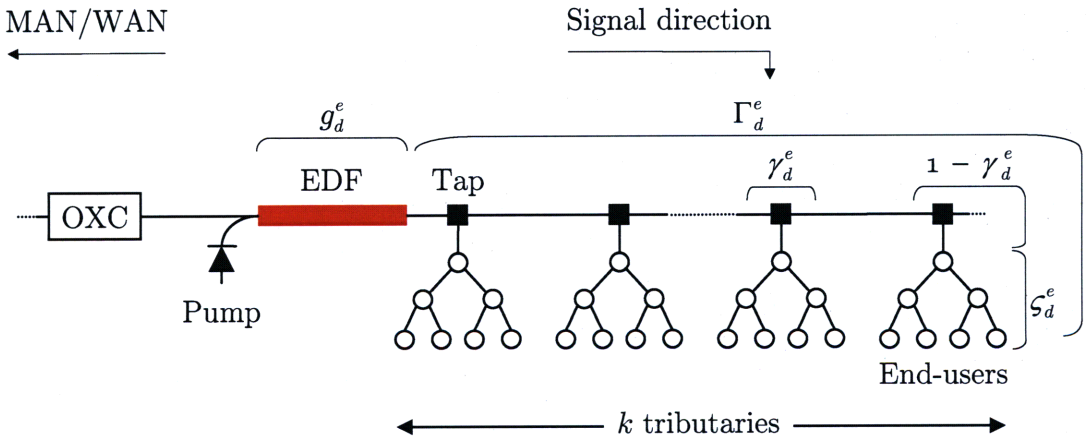
In this section, we consider the simple case of DNs with external amplification at the DN-MAN interface, as drawn in Figure 3-3.

■ 3.4.1 Source DN

In a source DN with external amplification, as drawn in Figure 3-3(a), a signal sent by an end-user incurs the entire power loss of passing through the DN before being amplified by an EDFA. This design has the advantage of simplicity, in that it requires the deployment of a single, lumped amplifier at the egress of the source DN. However,



(a) Externally amplified source DN.



(b) Externally amplified destination DN.

Figure 3-3. Externally amplified DNs with a bus-based physical topology. An EDFA is placed at the egress and ingress of the source and destination DN, respectively.

because the loss of optical power from the DN *precedes* optical amplification, the amplified spontaneous emission (ASE) noise injected by the EDFA could be significant relative to the attenuated signal power emerging from the DN. Indeed, as can be seen from equation (3.24), the noise figure of the first of a cascade of optical gain elements has the most significant impact on the overall noise figure. More concretely, as illustrated in Figure 3-3(a), the source DN stage comprises the following sequence: i) source DN with worst-case (splitting and excess) loss $\Gamma_s^e \ll 1$; ii) EDFA with gain g_s^e; and iii) OXC and fiber run with combined loss $\gamma_m = (g_m)^{-1}$. Equations (3.23) and (3.24) from the appendix may be invoked to obtain the following expression for the source DN stage noise figure:

$$F_{n,s}^e = \frac{2n_{\text{sp}}^s\left(g_s^e - 1\right) + g_m}{g_s^e \Gamma_s^e},$$

where n_{sp}^s is the spontaneous emission factor of the EDFA in the source DN stage. If we assume transparency in the source DN stage—that is:

$$G_s^e \equiv \frac{\Gamma_s^e g_s^e}{g_m} = 1,$$

then the noise figure reduces to:

$$F_{n,s}^e \approx \frac{2n_{\text{sp}}^s}{\Gamma_s^e} + 1 \approx \frac{2n_{\text{sp}}^s}{\Gamma_s^e},$$

where we have also assumed that $g_s^e, g_m \gg 1$. Thus, the noise figure is inversely proportional to the total attenuation Γ_s^e experienced in the DN. Minimizing the noise figure $F_{n,s}^e$ therefore requires maximizing Γ_s^e. In the example bus-based DN depicted in Figure 3-3(a) with k tributaries, Γ_s^e is given by:

$$\Gamma_s^e = \zeta_s^e\left(1 - \gamma_s^e\right)\left(\gamma_s^e\right)^{k-1}, \tag{3.7}$$

where ζ_s^e is the worst-case loss across a tributary, and γ_s^e is the coupling ratio across a tap from a bus input to a bus output and is constant for all taps along the bus[7]. It can be shown that:

$$\gamma_s^e = 1 - \frac{1}{k}$$

yields a minimum noise figure of:

$$F_{n,s}^e \approx \frac{2kn_{\text{sp}}^s}{\zeta_s^e\left(1 - \frac{1}{k}\right)^{k-1}} \rightarrow \frac{2ekn_{\text{sp}}^s}{\zeta_s^e}, \text{ as } k \rightarrow \infty. \tag{3.8}$$

[7]We acknowledge that a customized coupling ratio γ_s^e at each tap would yield a higher value of Γ_s^e (i.e., less loss), but such an implementation is arguably impractical.

■ 3.4.2 Destination DN

In a destination DN with external amplification, as depicted in Figure 3-3(b), optical amplification with gain:

$$g_d^e = \frac{1}{\Gamma_d^e} = \frac{1}{\zeta_d^e \left(1 - \gamma_d^e\right) \left(\gamma_d^e\right)^{k-1}} \tag{3.9}$$

precedes all of the worst-case attenuation Γ_d^e entailed by passing through the DN. In the absence of fiber nonlinearities, equation (3.24) in the chapter appendix indicates that positioning the EDFA in this manner is optimal with respect to the noise figure. In this case, the noise figure is given by:

$$F_{n,d}^e = \frac{2n_{\text{sp}}^d \left(g_d^e - 1\right) + \left(\Gamma_d^e\right)^{-1}}{g_d^e},$$

where n_{sp}^d is the spontaneous emission factor of the destination DN stage EDFA. Again, assuming transparency and that $g_d^e \gg 1$, we have:

$$F_{n,d}^e \approx 2n_{\text{sp}}^d + 1. \tag{3.10}$$

■ 3.4.3 Total noise figure

Combining equations (3.2), (3.3), and (3.4) under gain transparency with our above expressions for the source DN, MAN, and destination DN stage noise figures in equations (3.5), (3.6), (3.8), and (3.10) for the bus-based DN, we have:

$$2n_{\text{sp}} \left(\frac{k}{\zeta_s^e \left(1 - \frac{1}{k}\right)^{k-1}} + d \right) + F_{n,w}^u \lesssim \frac{\eta P_{\max}}{\mathscr{Q}^2 \overline{h}\nu \Delta f} \approx 45 \text{ dB} \qquad \text{(up-link)}$$

$$2n_{\text{sp}} \left(d + 2\right) + F_{n,w}^d \lesssim \frac{\eta P_{\max}}{\mathscr{Q}^2 \overline{h}\nu \Delta f} \qquad \text{(down-link)}$$

for inter-MAN communication, and:

$$2n_{\text{sp}} \left(\frac{k}{\zeta_s^e \left(1 - \frac{1}{k}\right)^{k-1}} + d + 1 \right) \lesssim \frac{\eta P_{\max}}{\mathscr{Q}^2 \overline{h}\nu \Delta f}$$

for intra-MAN communication. Note that in the above equations, we have assumed that:

$$n_{\text{sp}} \equiv n_{\text{sp}}^s = n_{\text{sp}}^m = n_{\text{sp}}^d.$$

■ 3.5 DNs with internal amplification

In the event that the previous noise figure constraints (and/or fiber nonlinearity constraints) cannot be respected[8] with externally amplified DNs, we may employ alternative physical layer designs for DNs which remedy this by improving the noise figures of the DN stages. In particular, the designs we consider in this subsection employ optical amplification *within* the DN via multiple segments of erbium-doped fiber (EDF) which are remotely pumped by a common laser. Remote pumping is attractive in that it: i) eliminates the overhead costs associated with deploying more than one pump, and ii) reduces the total pump power expended in the DN to overcome absorption in the EDFs[9]. However, the latter efficiency in pump power can be negated if fiber loss is significant in the DN—a serious concern for 980 nm pumps, which operate outside the low loss regime of fiber. Moreover, a single pump design is less resilient to component failures than a multiple pump design. Lastly, remote pumping is more vulnerable to stimulated Brillouin scattering (SBS), but this vulnerability can be mitigated using the techniques discussed in section 3.5.2.

The designs proposed in this section are shown to improve performance since a more distributed form of optical amplification ensures that: i) a transmitted signal from a source DN is not so severely attenuated before being amplified for the first time; and, ii) a signal amplified at the ingress of a destination DN is not so greatly amplified that the received signal is affected by nonlinear fiber impairments.

■ 3.5.1 Source DN

As discussed previously, the key to reducing the noise figure of the source DN stage is preventing excessive attenuation of the signal prior to optical amplification. Strategic placement of remotely-pumped EDF segments—the precise location generally depending upon the physical topology of the DN—is therefore a cost-effective design.

To illustrate this idea, we return to the bus-based DN in Figure 3-3. In Figure 3-4(a), we depict a modified version of this DN in which an EDF segment is inserted in the bus after each of the k tributaries, all of which are remotely-pumped by a common laser. The rightmost $k-1$ EDF segments are designed[10] to each provide a gain of g_s^i which compensates exactly for the tap coupling loss of γ_s^i (which we, again, assume to be constant for all taps along the bus); whereas, the final EDF segment is designed to provide a gain of $\tilde{g}_s^i$ which compensates exactly for: i) the tributary and coupling loss of $\zeta_s^i(1-\gamma_s^i)$; and ii) the OXC and fiber run loss of γ_m at the ingress to the MAN. We define the source DN stage noise figure for this configuration as the

[8]Violation of these constraints would occur as a result of high splitting loss, in an effort to multiplex many users in a DN to efficiently utilize network resources.

[9]The required pump power to overcome absorption in an EDF is given by the ordinate intercept in Figure 3-10.

[10]As discussed in further detail in section 3.6, the gain furnished by an EDF is a strong function of the length of the EDF and the pump power injected into it.

noise figure seen by the worst-case end-user in the furthest tributary from the MAN. It can be shown that this noise figure is given by:

$$F_{n,s}^{i} \approx \frac{2kn_{\mathrm{sp}}}{\zeta_s^i} + \frac{2n_{\mathrm{sp}}}{\zeta_s^i\,(g_s^i - 1)} + 1,$$

provided that the spontaneous emission factors of all EDF segments are equal to n_{sp}, and that $\tilde{g}_s^i \gg 1$. For a fixed ζ_s^i, the noise figure decreases with increasing g_s^i (i.e., decreasing γ_s^i). Thus, we have:

$$F_{n,s}^{i} \to \frac{2kn_{\mathrm{sp}}}{\zeta_s^i} + 1, \quad \text{as } \gamma_s^i \to 0. \tag{3.11}$$

This optimized noise figure, however, comes at the expense of a pump with large output power since a large portion of the pump power is tapped to every tributary en route to the MAN. This can be avoided, however, by using a tap made from a WDM coupler which employs a different coupling ratio for the pump wavelength than for the signal wavelength band. Comparing this expression with equation (3.8), we observe that the benefit of internal amplification with this configuration amounts to a decrease in the noise figure of at most $10\log e \approx 4.34$ dB, which occurs in the limit of a large number of tributaries k. The modesty of this performance improvement is the result of the bulk of the amplification in this configuration, $\tilde{g}_s^i$, occurring just before the MAN OXC—significantly downstream from the worst-case tributary.

To remedy this, we consider the more complex (and costly) source DN configuration drawn in Figure 3-4(b). In this configuration, a portion of the gain of $\tilde{g}_s^i$ in the previous configuration is moved into the individual tributaries. Specifically, an EDF providing a gain of:

$$\overline{g}_s^i = \frac{1}{\zeta_s^i\,(1-\gamma_s^i)} \tag{3.12}$$

is employed just before the tap in each tributary to compensate for the preceding tributary loss and the coupling loss which immediately follows; and the gain of the EDF segment just before the OXC is reduced to:

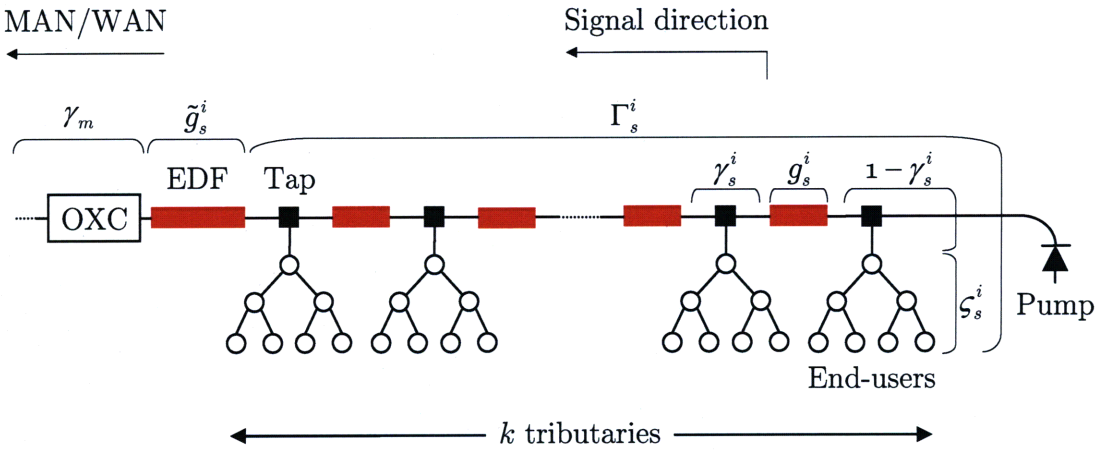
$$\tilde{g}_s^i = g_m.$$

The EDF segments on the bus each still provide a gain of:

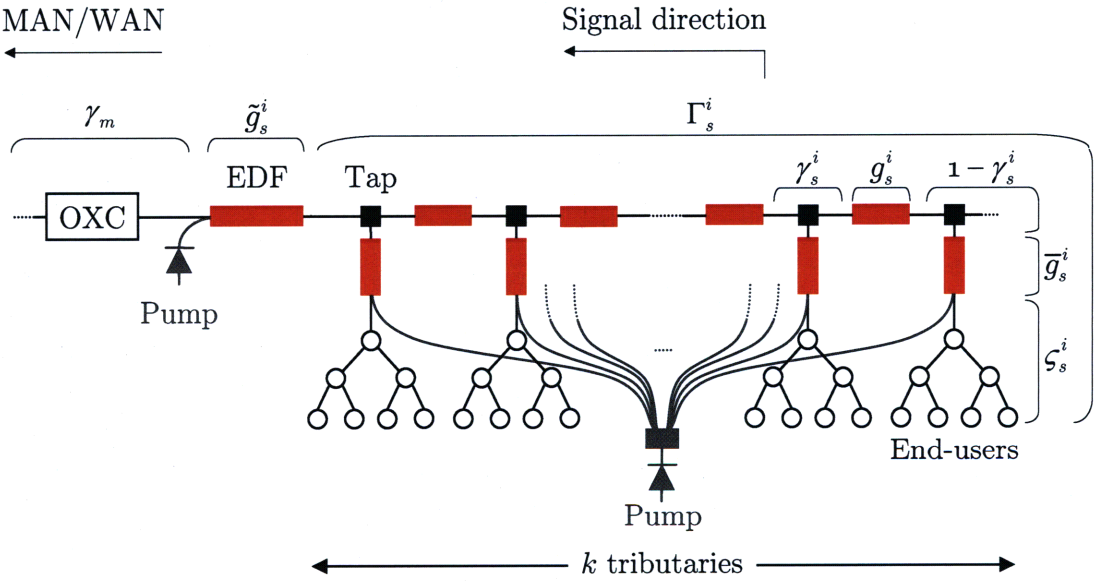
$$g_s^i = \frac{1}{\gamma_s^i},$$

to compensate exactly for the coupling loss that immediately follows.

In Figure 3-4(b), two possible pumping configurations are illustrated. In one alternative, the pump is located at the head-end of the bus, and pump power is channeled along the bus and into the tributary EDFs in a counter-propagating direction.



(a) Internally amplified source DN with EDFs in the bus only.



(b) Internally amplified source DN with EDFs in the bus and tributaries, and with two alternative pumping configurations (drawn in gray).

Figure 3-4. Internally amplified source DNs with a bus-based physical topology. EDF segments are placed throughout the DNs in order to improve the noise figure by moving optical amplification upstream.

The taps, responsible for connecting the tributaries to the bus, are WDM couplers[11] which have a different coupling ratio for the pump wavelength than for the signal wavelengths to ensure that equal pump power reaches each tributary EDF. This pumping configuration is attractive, provided that the fiber loss is not significant. In the event that the fiber loss is significant, it is desirable for the pump to be situated closer to the tributary EDFs and to pump these EDFs more directly—as they require more pump power than the bus EDFs—and allowing the residual pump power to couple into the bus. This can be achieved by immediately splitting the output power of a centrally located laser pump, and then coupling the power into each tributary EDF; or employing a dedicated pump for each tributary EDF.

The noise figure of this configuration can be shown to be:

$$\begin{aligned} F_{n,s}^{i} &\approx \frac{2n_{\mathrm{sp}}}{\zeta_s^i} + \frac{1}{g_s^i} + \frac{2n_{\mathrm{sp}}(k-1)\left(g_s^i - 1\right)}{g_s^i} + 2n_{\mathrm{sp}} \\ &\leq 2n_{\mathrm{sp}}\left(\frac{1}{\zeta_s^i} + k\right) + 1, \end{aligned}$$

where the approximation arises from setting the spontaneous emission factors of all EDF segments to n_{sp}, and noting that $\overline{g}_s^i, g_m \gg 1$. The upper bound on $F_{n,s}^i$ indicates that when the tributary loss ζ_s^i is quite significant, the noise figure is weakly dependent upon the number of tributaries k for any coupling ratio γ_s^i. The dependence of $F_{n,s}^i$ on k, in fact, completely vanishes as the coupling ratio γ_s^i tends to unity:

$$F_{n,s}^{i} \to 2n_{\mathrm{sp}}\left(\frac{1}{\zeta_s^i} + 1\right) + 1, \quad \text{as } \gamma_s^i \to 1. \tag{3.13}$$

The absence of a (strong) dependence on k is in contrast to the noise figure expressions in equations (3.8) and (3.11), which are proportional to k. The design implication of this is that, for a fixed number of supported users, a low source DN stage noise figure can be achieved with the internally amplified configuration in Figure 3-4(b) by employing a large number of tributaries k, each supporting a small number of users with low loss ζ_s^i. However, we point out that thermal noise—which is not accounted for in these calculations—imposes a limit on the number of tributaries k supportable. A low noise figure, however, comes at the expense of the large amount of power required to simultaneously pump all k tributaries. This tradeoff seems to be inherent, regardless of the detailed topology: an improved noise figure comes at the expense of amplification closer to the user, which requires replicated EDF segments and pump power.

[11]Such couplers are sometimes referred to as dichroic couplers.

■ 3.5.2 Destination DN

In section 3.4, we mentioned that the externally amplified DN configuration depicted in Figure 3-3(b) minimizes the destination DN stage noise figure in the absence of nonlinear fiber impairments. However, if the attenuation Γ_d^e of the destination DN is significant (e.g., because of a large number of supported users), the amplification required at the DN ingress, g_d^e, may be quite large. If the attenuation in the DN, furthermore, occurs gradually over distances of a kilometer or more, the effects of fiber nonlinearities on signal quality cannot be neglected. The particular nonlinear effects expected to most severely impede detection in DNs are the intra-channel effects of SBS and self-phase modulation (SPM). In Figure 3-5, the threshold power for the onset of SBS is plotted as a function of fiber distance traveled by the signal. (We conservatively assume in this plot that all of the signal power lies within the 20 MHz gain bandwidth of SBS. The threshold power can be significantly greater if the signal has broad spectral width, which can be achieved by [250]: i) directly modulating the laser source, ii) dithering the laser slightly in frequency, or iii) using phase modulation rather than amplitude modulation schemes.) In Figure 3-6, the factor by which an optical pulse broadens as a result of SPM is depicted as a function of distance traveled by the pulse for different transmit powers.

Since these nonlinear effects can be significant for signal powers above 10 dBm per wavelength, we consider destination DN configurations in which the total gain in the DN is distributed across multiple EDF segments remotely-pumped by a common laser. Again, our discussion will take place in the context of the bus-based topologies considered thus far. The first configuration we consider, depicted in Figure 3-7(a), employs EDF segments only in the bus portion of the DN, similar to Figure 3-4(a) for the source DN. In this configuration, the first EDF segment provides a gain of $\tilde{g}_d^i$ which compensates exactly for the loss $(1-\gamma_d^i)\,\zeta_d^i$ that arises when a signal is coupled into a tributary and passes through the tributary to the end-user. The gain g_d^i provided by each of the following $k-1$ EDF segments compensates exactly for the tap coupling loss γ_d^i on the bus that immediately precedes it. Defining the noise figure for this configuration $F_{n,d}^i$ as the noise figure seen by the worst-case end-user in the furthest tributary from the MAN, we have:

$$F_{n,d}^i \approx 2n_{\text{sp}}\left[1+\zeta_d^i(k-1)\left(\gamma_d^i+\frac{1}{\gamma_d^i}-2\right)\right]+1,$$

where we have assumed that the spontaneous emission factors of all EDF segments are equal to n_{sp}, and the approximation arises from $\tilde{g}_d^i \gg 1$. Comparing this expression with equation (3.10), we observe that introducing EDF segments along the bus increases the noise figure by a term that is proportional to k but that is small nevertheless.

If, in the above configuration, the signal power emerging from the initial EDF segment is still sufficiently high that the effects of fiber nonlinearities are significant,

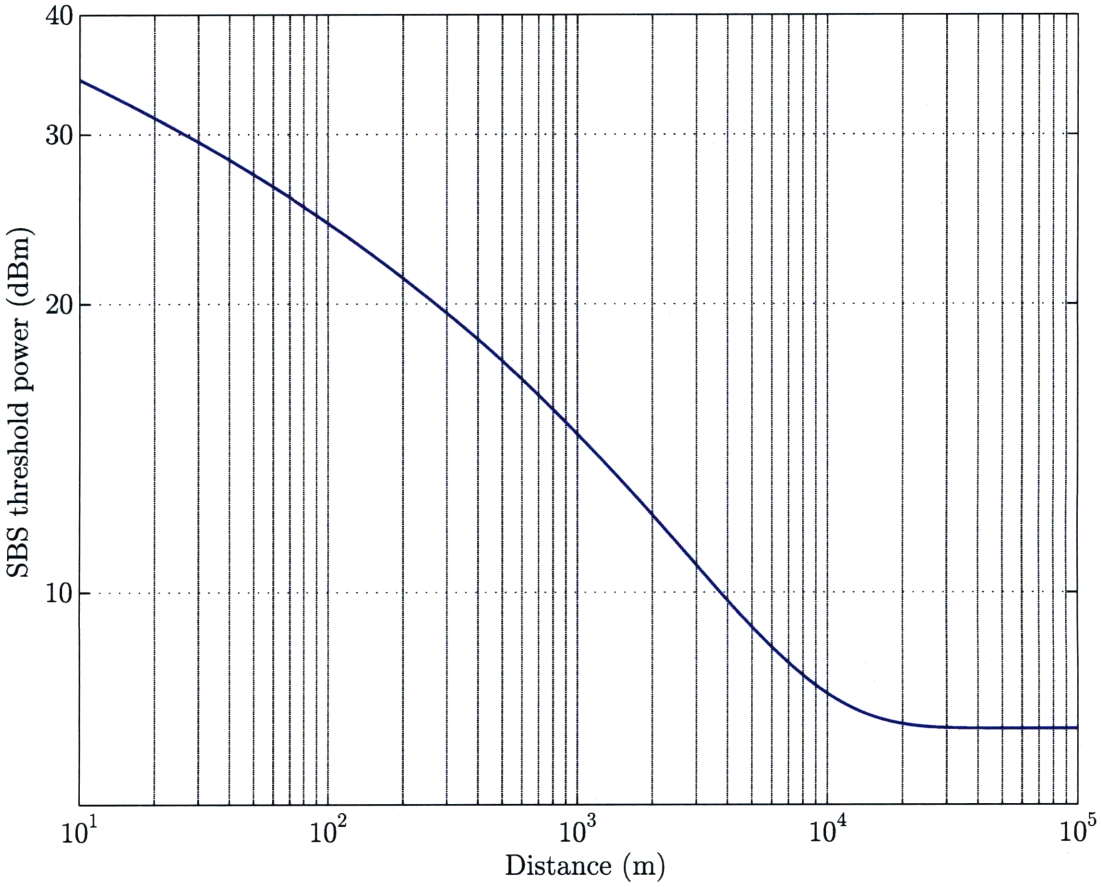


Figure 3-5. SBS threshold power (P_{th}) versus fiber distance traveled by optical signal. The threshold power is given approximately by [269]: $P_{\text{th}} \approx \frac{21ba_e}{g_B l_e}$, where a_e and l_e are the effective area and length of the fiber, respectively; $g_B \approx 4 \times 10^{-11}$ m/W is the Brillouin gain coefficient; and the value of b lies between 1 and 2 depending on the relative polarizations of the pump and Stokes waves. In the above plot, we have assumed that $b = 1$ and $a_e = 50\ \mu\text{m}^2$.

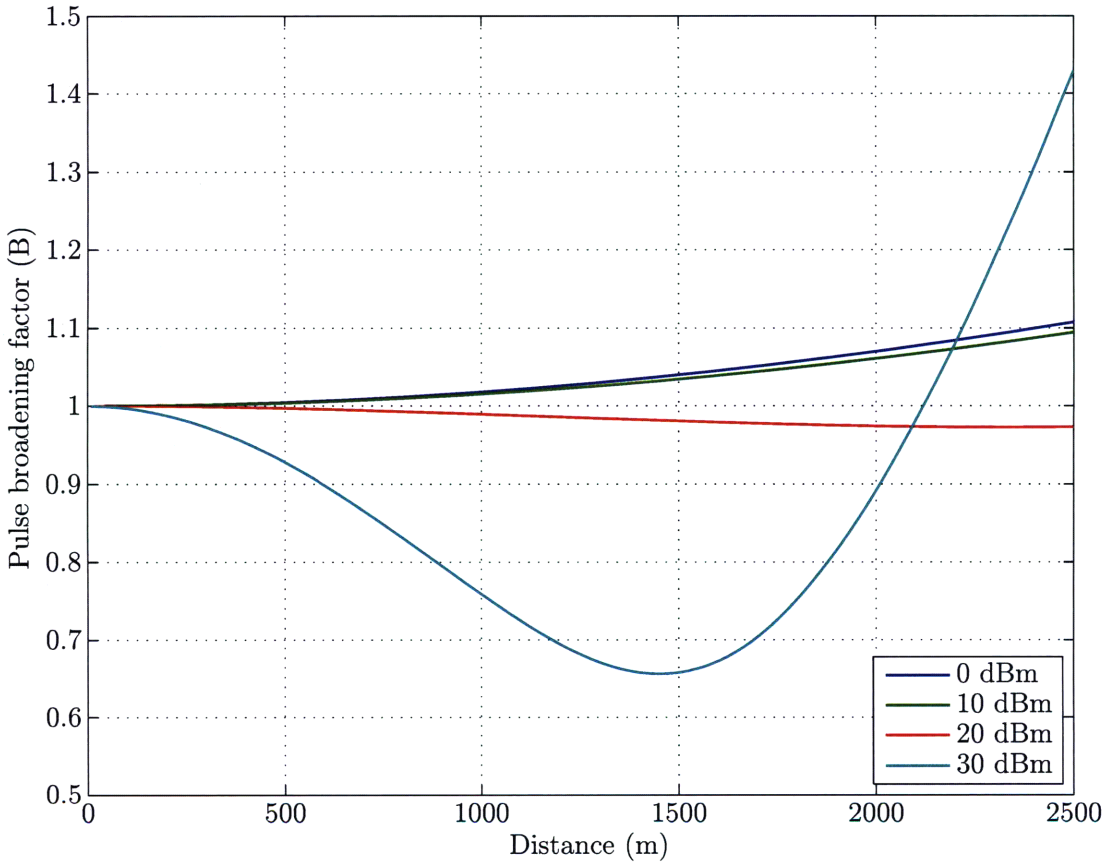
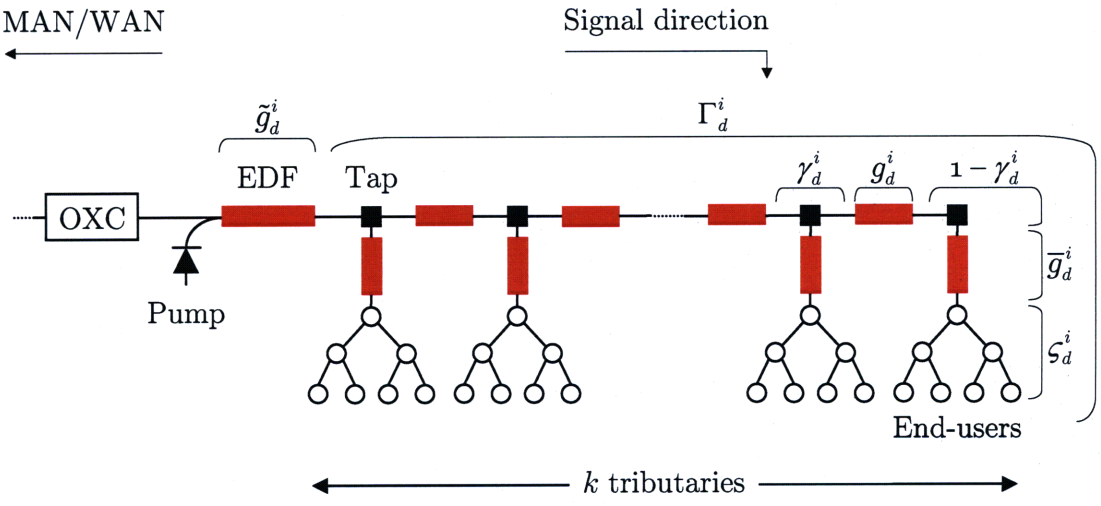
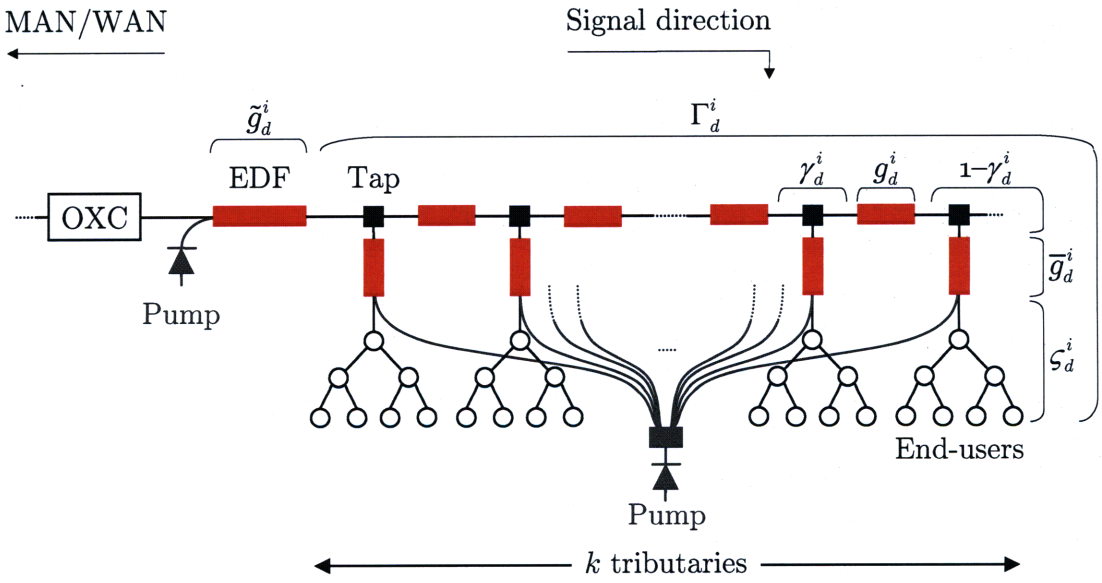


Figure 3-6. SPM pulse broadening factor (B) versus distance (l) traveled for a non-chirped Gaussian pulse for different transmit powers. The pulse broadening factor is given approximately by [237]: $B \approx \sqrt{1 + \sqrt{2}\frac{l_e l}{l_{NL} l_D} + \left(1 + \frac{4l_e^2}{3\sqrt{3}l_{NL}^2}\right)\frac{l^2}{l_D^2}}$, where l_e is the effective length of the fiber; l_D is the dispersion length; and l_{NL} is the nonlinear length. In the above plot, transmission at 40 Gbps in single-mode fiber at a wavelength of 1550 nm is assumed, leading to anomalous dispersion.



(a) Internally amplified destination DN with EDFs in the bus only.



(b) Internally amplified destination DN with EDFs in the bus and tributaries, and with two alternative pumping configurations (drawn in gray).

Figure 3-7. Internally amplified destination DNs with a bus-based physical topology. EDFA segments are placed throughout the DNs in order to mitigate the effects of fiber nonlinearities arising from a large initial gain.

we can resort to a destination DN design that moves some of the initial gain $\tilde{g}_i^d$ into the tributaries, albeit at the expense of a higher noise figure. The design depicted in 3-7(b) is analogous to the source DN design in Figure 3-4(b). In particular, the gain of the first EDF segment is reduced to:

$$\tilde{g}_d^i = \frac{1}{1-\gamma_d^i}, \tag{3.14}$$

to compensate exactly for the eventual tap coupling loss into the tributary; the gain provided by the EDF in each tributary is:

$$\overline{g}_d^i = \frac{1}{\zeta_d^i}, \tag{3.15}$$

to compensate exactly for the tributary loss that follows it; and the EDF segments on the bus each still provide a gain of:

$$g_d^i = \frac{1}{\gamma_d^i},$$

to compensate exactly for the immediately preceding tap coupling loss. The noise figure of this configuration can be shown to be:

$$F_{n,d}^i \approx 2n_{\text{sp}} \left[\gamma_d^i + 1 + (k-1)\left(\gamma_d^i + \frac{1}{\gamma_d^i} - 2\right)\right] + 1,$$

where we have assumed that the spontaneous emission factors of all EDF segments are equal to n_{sp}, and the approximation arises from $\overline{g}_d^i \gg 1$. It can be shown that the choice of coupling ratio that minimizes the noise figure is:

$$\gamma_d^i = \sqrt{1-\frac{1}{k}}. \tag{3.16}$$

While the corresponding noise figure achieved is cumbersome, it is easy to see that it satisfies the following inequality:

$$\min_{\gamma_d^i} F_{n,d}^i < 4n_{\text{sp}} + 1. \tag{3.17}$$

The design implication of this is that the effects of nonlinear fiber impairments can be avoided altogether—provided the geographic span of a tributary is not large—with a penalty of at most $2n_{\text{sp}}$ in the destination DN stage noise figure. Again, the price paid for this low noise figure is the larger amount of power required to simultaneously pump all k tributaries.

■ 3.5.3 Total noise figure

Combining equations (3.2), (3.3), and (3.4) under gain transparency with our above expressions for the source DN, MAN, and destination DN stage noise figures in equations (3.5), (3.6), (3.13), and (3.17) for the bus-based DN, we have:

$$2n_{\text{sp}}\left(\frac{1}{\zeta_s^i}+d+1\right)+F_{n,w}^u \lesssim \frac{\eta P_{\max}}{\mathcal{Q}^2\overline{h}\nu\Delta f} \approx 45 \text{ dB} \qquad \text{(up-link)}$$

$$2n_{\text{sp}}\,(d+3)+F_{n,w}^d \lesssim \frac{\eta P_{\max}}{\mathcal{Q}^2\overline{h}\nu\Delta f} \qquad \text{(down-link)}$$

for inter-MAN communication, and:

$$2n_{\text{sp}}\left(\frac{1}{\zeta_s^i}+d+3\right) \lesssim \frac{\eta P_{\max}}{\mathcal{Q}^2\overline{h}\nu\Delta f}$$

for intra-MAN communication. Note that in the above equations, we have assumed that:

$$n_{\text{sp}} \equiv n_{\text{sp}}^s = n_{\text{sp}}^m = n_{\text{sp}}^d.$$

■ 3.5.4 Number of supportable users per DN tributary

As discussed in section 3.1.2, we employ separate passive components for up-link and down-link communication. This design requires intra-tributary signals to pass through the parent MAN node OXC before returning to the tributary for reception at the destination end-user. Though circuitous, this routing allows intra-tributary signals to be optically amplified en route to the destination end-user; whereas, given our DN designs in this chapter, if an intra-tributary signal remained within the tributary, it could not be amplified en route. This added optical amplification allows more end-users to be supported within each tributary[12].

Using the equations in section 3.5.3, which capture the physical layer constraints for reliable OFS communication, we obtain the following approximate bound on the loss that can be tolerated within a source DN tributary:

$$\begin{aligned}\zeta_s^i &\gtrsim \left[\frac{1}{2n_{\text{sp}}}\left(\frac{\eta P_{\max}}{\mathcal{Q}^2\overline{h}\nu\Delta f}-F_{n,w}\right)-d-3\right]^{-1} \\ &\approx 2n_{\text{sp}}\left(\frac{\eta P_{\max}}{\mathcal{Q}^2\overline{h}\nu\Delta f}-F_{n,w}\right)^{-1}, \end{aligned} \qquad (3.18)$$

[12]The only exceptions are tributaries based upon star topologies. This is because, in these topologies, intra-DN signals incur the same loss when departing the tributary for the parent MAN node OXC as when reaching the destination end-user directly. Nevertheless, the performance penalty in employing the circuitous routing through the parent MAN node is very minor.

where we have assumed that $F_{n,w} = F^u_{n,w} = F^d_{n,w}$ by symmetry. In translating this into a bound on the number of supportable users in a tributary, we consider both star and bus tributary topologies, which represent best- and worst-case scenarios, respectively. Other topologies, such as trees, should support an intermediate number of users.

In both of the following analyses, we denote the excess loss of a tap based upon a passive 2×2 coupler by δ. Owing to the short distances traveled within a tributary, fiber loss is neglected.

Star topology

In a tributary based upon a star topology that supports n_t end-users, a $1 \times n_t$ passive combiner is used for up-link communication in the source DN. In addition to the inherent splitting loss of $1/n_t$, an excess loss associated with the combiner must be accounted for. We assume that such a device is fabricated by cascading $\log_2 n_t$ stages of $\frac{n_t}{2}$ 3 dB couplers. The excess loss of such a device may thus be approximated by:

$$1 - (1 - \delta)^{\log_2 n_t},$$

since a signal passes through exactly $\log_2 n_t$ couplers in traveling from any input port to the output port. We acknowledge that excess loss may scale differently with n_t using an integrated optics implementation for the combiner.

The tributary loss constraint for the tributary may therefore be expressed as:

$$\zeta^i_s \leq \frac{(1 - \delta)^{\log_2 n_t}}{n_t},$$

which translates to the following upper bound on the number of supportable users:

$$n_t \leq \left(\zeta^i_s\right)^{1/[\log_2(1-\delta)-1]}. \tag{3.19}$$

Figure 3-8 depicts the number of supportable users for different values of ζ^i_s.

Bus topology

For the bus topology, a signal generated at an end-user must pass through every end-user's tap en route to the head-end of the tributary in the worst case. Denoting the tap coupling coefficient across the tributary bus by γ_t and the number of end-users supported per tributary by n_t, we have the following loss constraint:

$$\zeta^i_s \leq (1 - \gamma_t)\gamma_t^{n_t - 1}(1 - \delta)^{n_t},$$

which translates to the following upper bound on the number of supportable users:

$$n_t \leq \frac{\log\left(\frac{\zeta_s^i \gamma_t}{1-\gamma_t}\right)}{\log\left[\gamma_t(1-\delta)\right]}. \tag{3.20}$$

In maximizing n_t over all possible values of γ_t, we obtain the following optimality condition:

$$(1-\delta)\gamma_t^* = \left(\frac{\zeta_s^i \gamma_t^*}{1-\gamma_t^*}\right)^{1-\gamma_t^*},$$

which does not have a closed form solution for γ_t^*. Figure 3-8 depicts a numerical optimization of n_t over γ_t^* for different values of ζ_s^i. As evident in this figure, for reasonable values of δ and ζ_s^i, we have $\gamma_t^* \approx 1$. We may thus invoke the following approximation of the optimality condition:

$$(1-\delta) \approx \left(\frac{\zeta_s^i}{1-\gamma_t^*}\right)^{1-\gamma_t^*},$$

which has the following solution:

$$\gamma_t^* \approx \frac{\mathcal{W}\left[\frac{1}{\zeta_s^i}\ln\left(\frac{1}{1-\delta}\right)\right] - \ln\left(\frac{1}{1-\delta}\right)}{\mathcal{W}\left[\frac{1}{\zeta_s^i}\ln\left(\frac{1}{1-\delta}\right)\right]}, \tag{3.21}$$

where $\mathcal{W}(\cdot)$ is the Lambert Function, defined as the inverse of the function:

$$f(x) = xe^x,$$

and is well approximated by:

$$\mathcal{W}(x) \approx \ln x - \ln(\ln x) + \frac{\ln(\ln x)}{\ln x},$$

when $x \gtrapprox 3$. The above approximation for γ_t^* may be substituted into equation (3.20) to obtain a good approximation of the number of supportable users n_t.

■ 3.6 Required pump power

In this section, we address the (approximate) required pump power needed to implement the EDFAs in sections 3.4 and 3.5. We devote special attention to this issue because the pump cost dominates the cost of an EDFA, and also constitutes a significant fraction of the overall cost of a DN.

The gain of an EDFA is a strong function of the injected pump power and the EDF length, in addition to a host of other parameters. In particular, in the unsaturated

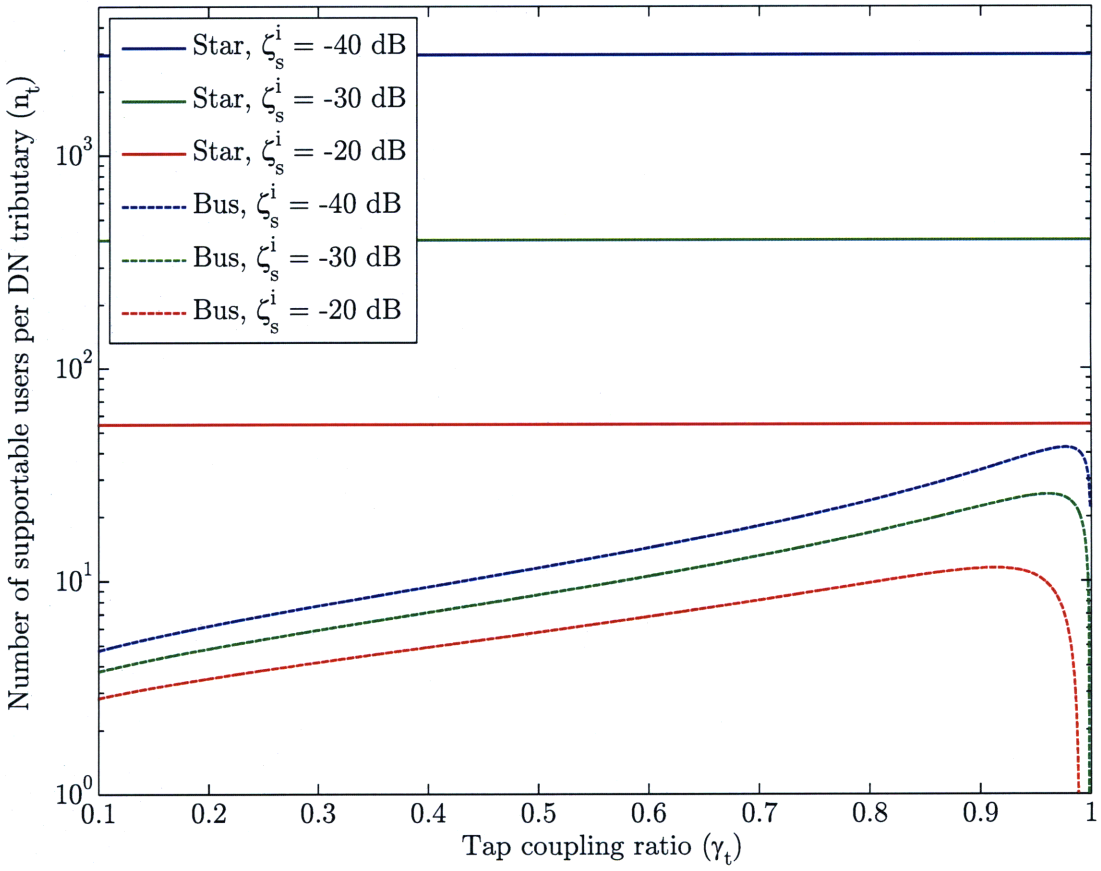


Figure 3-8. Number of supportable end-users per DN tributary (n_t) versus tap coupling ratio (γ_t) for star and bus tributary designs, and for different tributary losses (ζ_s^i). It is assumed that $\delta = 0.1$.

gain regime of an EDFA's operation, the relationship between the normalized pump power $\overline{P}_p$, amplifier gain g, and length of the EDF l is given by the following equation [81]:

$$\overline{P}_p \equiv \frac{P_p}{P_p^{\text{sat}}} = \alpha_p l \frac{V_s}{1 - \exp\left[\alpha_p l \left(V_s - 1\right)\right]}, \tag{3.22}$$

where:

$$P_p^{\text{sat}} = \frac{\overline{h}\nu_p a_p}{\tau_{\text{sp}} \tilde{a}_p}$$

$$V_s = \frac{1 + \eta_p}{1 + \eta_s} \left(1 + \frac{\ln g}{\alpha_s l}\right),$$

where ν_p is the pump frequency; a_p is the cross-section area of the pump mode inside the fiber; τ_{sp} is the spontaneous lifetime of the excited state; $\tilde{a}_p$ is the transition cross-section at the pump frequency ν_p; α_p and α_s are the pump and signal erbium-doping absorption coefficients, respectively; and η_p and η_s are the ratios of absorption to emission at the pump and signal frequencies, respectively. Figure 3-9 depicts the required normalized input pump power as a function of EDF length for various values of g, as given by equation (3.22). As illustrated in the figure, for each value of g, there exists a minimum normalized input pump power (with corresponding EDF length). The relationship between this minimum normalized input pump power and the gain g is depicted in Figure 3-10. While the precise mathematical relationship is mathematically cumbersome, Figure 3-10 illustrates that the relationship is approximately linear when g is expressed in dB. In reality, however, this approximate linear relationship cannot persist indefinitely: amplifier self-saturation by ASE and laser oscillation prevent the EDFA gain from being increased indefinitely [81].

In the configurations proposed in section 3.4, the required pump powers and EDF lengths can be obtained directly from equation (3.22). However, in the configurations proposed in section 3.5, the pump power can pass through several optical elements (e.g., fiber, taps) before amplifying the signal. Computation of the required pump power from the laser as well as the EDF lengths must therefore generally account for the attenuation of the pump power through these elements. Since the configurations of Figures 3-4(b) and 3-7(b) were shown to possess a significant noise figure advantage relative to their counterparts in Figures 3-4(a) and 3-7(a), we shall confine our attention to the former two configurations. In these two configurations, the gains furnished—and pump power required—by the EDFs in the tributaries are expected to be greater than those of the EDFs in the bus. Thus, computation of the approximate required pump power is straightforward in that it depends almost entirely on the tributary design. In particular, a good approximation of the required pump power from the laser can be obtained by first employing $g = \overline{g}_s^i$ or $\overline{g}_d^i$ in equation (3.22) to determine the normalized pump power required for a single tributary; and then

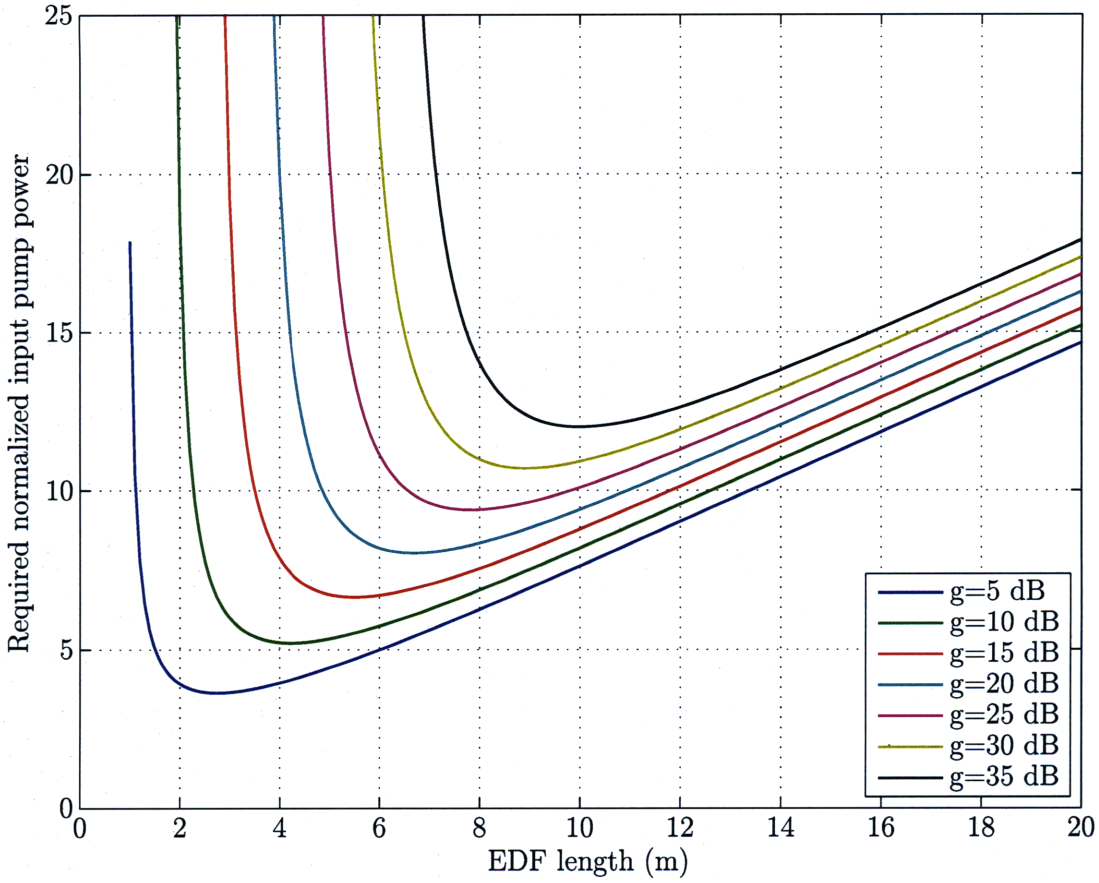


Figure 3-9. Required normalized input pump power $(\overline{P}_p)$ versus EDF length (l). The following typical parameters are assumed for a 1480 nm pump: $\alpha_p = 1.32\ \mathrm{m}^{-1}$, $\alpha_s = 1.50\ \mathrm{m}^{-1}$, $\eta_p = 0.23$, and $\eta_s = 1.3$. These parameters correspond to a saturation power P_p^{sat} of approximately 1 mW. Adapted from [210, Figure 3-2].

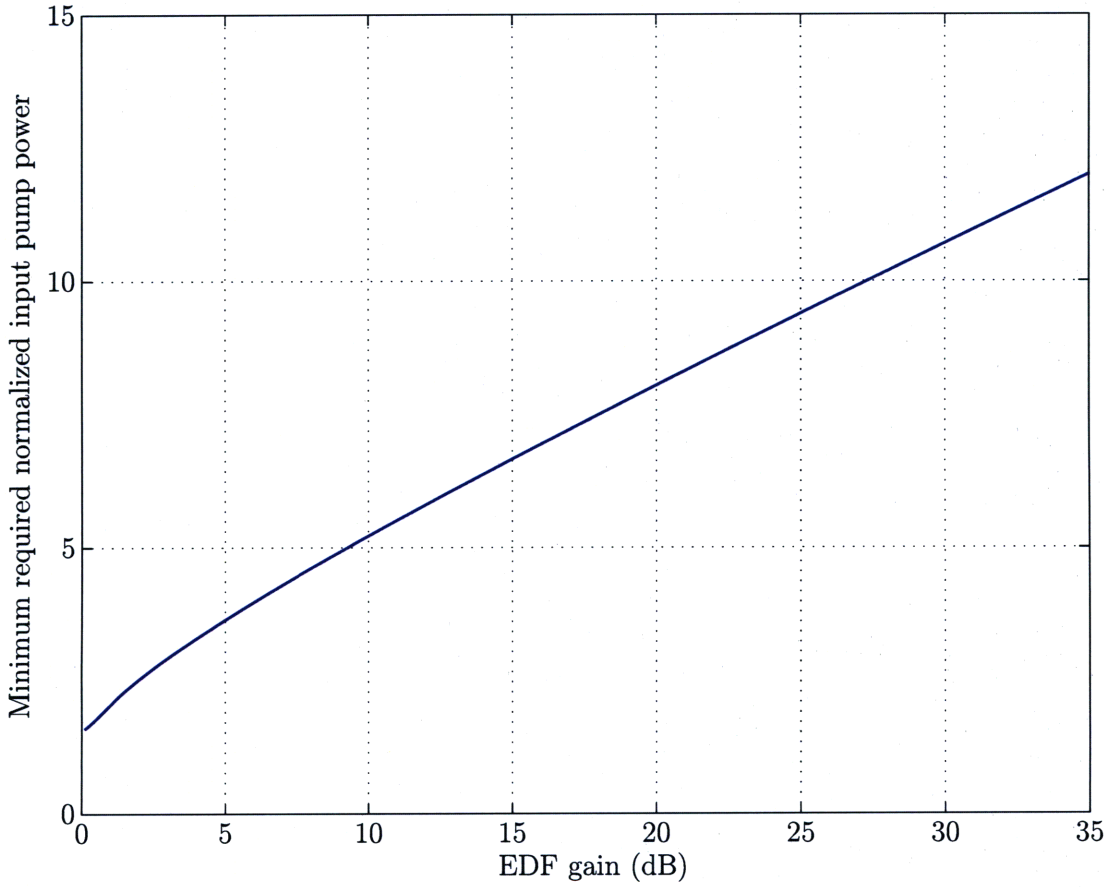


Figure 3-10. Minimum required normalized input pump power $(\overline{P}_p)$ versus EDF gain (g). The following typical parameters are assumed for a 1480 nm pump: $\alpha_p = 1.32$ m^{-1}, $\alpha_s = 1.50$ m^{-1}, $\eta_p = 0.23$, and $\eta_s = 1.3$. These parameters correspond to a saturation power P_p^{sat} of approximately 1 mW. Adapted from [210, Figure 3-6].

multiplying this by k to obtain the required normalized power to pump all of the tributaries. This is illustrated in the following example.

Example 3.1 *Consider a bus-based DN with $k = 10$ tributaries.*

For the externally amplified DN configurations of Figure 3-3, assume tributary losses of:

$$\zeta_s^e = -5 \ dB$$
$$\zeta_d^e = -15 \ dB$$

and optimal bus coupling ratios of:

$$\gamma_s^e = \gamma_d^e = 1 - \frac{1}{k} = -0.46 \ dB.$$

Furthermore, assume a loss of:

$$\gamma_m = -10 \ dB$$

for a MAN OXC and the fiber run following it. The required EDFA gains are thus:

$$g_s^e = \frac{1}{\Gamma_s^e \gamma_m} \approx 29.2 \ dB$$
$$g_d^e = \frac{1}{\Gamma_d^e} \approx 29.2 \ dB,$$

where we have invoked equations (3.7) and (3.9). According to equation (3.22) (or Figure 3-10), the implied required normalized pump powers for both source and destination EDFAs is at least 10.5.

Let us now consider the internally amplified DN configurations of Figures 3-4(b) and 3-7(b). Assume tributary losses of:

$$\zeta_s^i = -5 \ dB$$
$$\zeta_d^i = -15 \ dB$$

and optimal bus coupling ratios of:

$$\gamma_s^i = \gamma_d^i = \sqrt{1 - \frac{1}{k}} = -0.23 \ dB.$$

Then, using equation (3.12), the required EDFA gain per tributary for the source configuration is:

$$\overline{g}_s^i = \frac{1}{\zeta_s^i (1 - \gamma_s^i)} \approx 17.9 \ dB.$$

Similarly, using equations (3.14) and (3.15), the required EDFA gains for the destination configuration are:

$$\tilde{g}_d^i = \frac{1}{1-\gamma_d^i} \approx 12.9 \ dB$$
$$\overline{g}_d^i = \frac{1}{\zeta_d^i} = 15 \ dB.$$

According to equation (3.22) (or Figure 3-10), the implied required normalized pump powers per tributary are at least 7.5 and 6.7 for the source and destination configurations, respectively; and the most upstream EDF segment in the destination configuration requires a normalized pump power of approximately 6.1. Accounting for the total number of tributaries that need to be simultaneously pumped, we arrive at aggregate normalized pump powers of at least 75 and 73 for the source and destination configurations, respectively.

As expected, the pump powers required for internally amplified DN configurations are higher than those of externally amplified DN configurations for gain transparency. But, as discussed earlier, the noise figures of internally amplified DNs may be considerably lower.

■ 3.7 Conclusion

In this chapter, we addressed the physical layer design of OFS networks in the metro and access environments. We derived high-level noise figure constraints, and proposed alternative designs to respect these constraints. For the access environment, in particular, the choice of DN design will depend upon a host of factors, such as the number and geographic distribution of end-users to be supported, the physical layer design of the WAN prior to (respectively, after) the first (last) instance of regeneration, and, of course, the cost structures of the alternatives. We address the latter issue of cost in more detail in Chapter 5.

The contributions made in this chapter can be extended in various ways. For example, detailed design and analysis of DN topologies other than the bus-based ones examined here should be addressed. However, as noted earlier, the insights gained from analyzing this family of topologies should carry over to arbitrary DN topologies. In addition, we chose to focus on amplification with EDFAs, as they represent, by far, the most mature, and hence cost-effective, optical amplification technology. Raman amplifiers, however, should they come down in cost, could be combined with EDFAs to provide significant noise figure improvements—a technique already employed in WANs, particularly for sub-marine links. Another potentially attractive future technology is the remotely-powered and controlled OXC for the DN [55], which could be used in lieu of passive devices to improve noise figure performance, and/or support a larger base of users if DN resources are under-utilized.

■ 3.A Appendix

■ 3.A.1 Direct detection of optically amplified signals

When a signal is passed through a single optical amplifier, the amplified signal power incident at the photodetector can be written as:

$$P_a = GP_s + P_{\text{sp}},$$

where G is the amplifier gain, P_s is the input optical signal power, and:

$$P_{\text{sp}} = (G-1)n_{\text{sp}}\overline{h}v\Delta\nu_{\text{sp}}$$

is the spontaneous emission noise power added to the signal, in which n_{sp} is the spontaneous emission factor, $\overline{h}$ is the Planck constant, ν is the optical frequency, and $\Delta\nu_{\text{sp}}$ is the effective bandwidth of spontaneous emission which can be approximated by the amplifier bandwidth.

The variance of the photocurrent fluctuations at the detector can be written as:

$$\sigma^2 = \sigma_T^2 + \sigma_s^2 + \sigma_{\text{sp-sp}}^2 + \sigma_{\text{sg-sp}}^2 + \sigma_{\text{s-sp}}^2,$$

corresponding to thermal noise, signal shot noise, spontaneous-spontaneous beat noise, signal-spontaneous beat noise, and spontaneous emission shot noise, respectively, where (e.g., see [14, section 8.6]):

$$\begin{aligned}
\sigma_T^2 &= \frac{4k_BTF_n^R\Delta f}{R_L} \\
\sigma_s^2 &= 2\hat{q}[\hat{R}(GP_s + P_{\text{sp}}) + I_d]\Delta f \\
\sigma_{\text{sp-sp}}^2 &= 4\hat{R}^2[(G-1)n_{\text{sp}}\overline{h}\nu]^2\Delta\nu_{\text{opt}}\Delta f \\
\sigma_{\text{sg-sp}}^2 &= 4\hat{R}^2GP_s(G-1)n_{\text{sp}}\overline{h}\nu\Delta f \\
\sigma_{\text{s-sp}}^2 &= 4\hat{q}\hat{R}(G-1)n_{\text{sp}}\overline{h}\nu\Delta\nu_{\text{opt}}\Delta f,
\end{aligned}$$

and k_B is the Boltzmann constant, F_n^R is the noise figure of the pre- and main amplifiers, Δf is the effective electrical bandwidth (approximately half of the bit rate), T is the absolute temperature, R_L is the resistance of the load resistor, $\hat{q}$ is the electronic charge, $\hat{R}$ is the responsivity of the detector, I_d is the dark current, and $\Delta\nu_{\text{opt}}$ is the optical bandwidth of the spontaneous-emission noise.

For an optically amplified signal which is shot noise limited at the amplifier's input and signal-spontaneous beating noise limited at the output, the noise figure is given by:

$$F_n \equiv \frac{\text{SNR}_{\text{in}}}{\text{SNR}_{\text{out}}} = \frac{2n_{\text{sp}}(G-1)}{G}.$$

Neglecting $\hat{R}P_{\text{sp}}$ and I_d in the previous shot noise equation and substituting in F_n and $\hat{R} = \eta\hat{q}/\overline{h}v$, where η is the quantum efficiency of the photodetector, the previous noise power component equations can be written as:

$$\begin{aligned}
\sigma_T^2 &= \frac{4k_BTF_n^R\Delta f}{R_L} \\
\sigma_s^2 &= \frac{2\hat{q}^2\eta GP_s\Delta f}{\overline{h}\nu} \\
\sigma_{\text{sp-sp}}^2 &= (\hat{q}\eta GF_n)^2\Delta\nu_{\text{opt}}\Delta f \\
\sigma_{\text{sg-sp}}^2 &= \frac{2(\hat{q}\eta G)^2F_nP_s\Delta f}{\overline{h}\nu} \\
\sigma_{\text{s-sp}}^2 &= 2\hat{q}^2\eta GF_n\Delta\nu_{\text{opt}}\Delta f.
\end{aligned}$$

Let us consider the case of a cascade of amplifiers. By redefining the noise figure as:

$$F_n \equiv \frac{1 + 2n_{\text{sp}}(G-1)}{G}, \tag{3.23}$$

the noise figure now accounts for both signal shot noise and signal-spontaneous beat noise. This is known as the standard optical noise figure definition [80], and represents a truncation of a more general definition which includes contributions from spontaneous-spontaneous beat noise and spontaneous emission shot noise (but still neglects thermal noise). We choose to employ the truncated noise figure—which is accurate for large signals—as it allows for tractable cascading analysis, in that it omits the aforementioned spontaneous-spontaneous beat noise and spontaneous emission shot noise whose noise figure contributions are power-dependent. In particular, for a cascade of optical gain elements, including power attenuation elements, the following is an expression for the effective noise figure:

$$F_n^{\text{eff}} = F_{n,1} + \frac{F_{n,2}-1}{G_1} + \frac{F_{n,3}-1}{G_1G_2} + \cdots + \frac{F_{n,k}-1}{G_1G_2\cdots G_{k-1}}, \tag{3.24}$$

where $F_{n,j}$ and G_j are the noise figure and the gain of the j^{th} element. Thus, we can readily obtain the output signal-to-noise ratio (SNR) of a cascade of optical elements.

Output BER and receiver sensitivity

Our performance criterion is BER. For direct detection, it is common to assume that the noise contributions render the photodetector's current—I_0 and I_1 in the case of a 0 bit and 1 bit, respectively—Gaussian, and that the BER is thus given by:

$$\text{BER} = \frac{1}{2}\text{erfc}\left(\frac{\mathcal{Q}}{\sqrt{2}}\right), \tag{3.25}$$

where:

$$\mathcal{Q} = \frac{I_1 - I_0}{\sigma_1 + \sigma_0},$$

in which σ_0 and σ_1 are the noise standard deviations of a 0 bit and 1 bit, respectively, and:

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-y^2} dy.$$

To compute the BER of the detected signal for on-off keying (OOK), we have $I_0 = 0$, and:

$$I_1 = \hat{R} G_{\text{total}} P_1 = 2\hat{R} G_{\text{total}} P_{\text{in}},$$

where G_{total} is the overall gain of the cascade of optical elements, and P_{in} is the average input power assuming equally likely 0 and 1 bits. Now, considering only signal shot noise, signal-spontaneous beat noise, and thermal noise, the output noise power for the 0 and 1 bits are[13]:

$$\sigma_0 = \sigma_T$$

and:

$$\sigma_1 = \left(G_{\text{total}}^2 F_n^{\text{eff}} \sigma_{1,\text{in}}^2 + \sigma_T^2 \right)^{1/2},$$

respectively, where $\sigma_{1,\text{in}}^2$ is the (corresponding electrical) noise power of a 1 bit at the input of the cascade. Substituting the above into equation (3.25), we have:

$$\text{BER} = \frac{1}{2} \text{erfc} \left(\frac{\sqrt{2} \hat{R} G_{\text{total}} P_{\text{in}}}{\left(G_{\text{total}}^2 F_n^{\text{eff}} \sigma_{1,\text{in}}^2 + \sigma_T^2 \right)^{1/2} + \sigma_T} \right). \tag{3.26}$$

The inverse problem—namely, determining the required power to achieve a specified BER—is known as the receiver sensitivity problem[14]. Thus, by comparing equations (3.25) and (3.26), we have the following equation for the receiver sensitivity P_{sens}:

$$P_{\text{sens}} \equiv P_{\text{in}}|_{\mathcal{Q}} = \frac{\mathcal{Q} \left[\left(G_{\text{total}}^2 F_n^{\text{eff}} \sigma_{1,\text{in}}^2 + \sigma_T^2 \right)^{1/2} + \sigma_T \right]}{2\hat{R} G_{\text{total}}}.$$

In the limit of large gain G_{total}, the above can be simplified by substituting in the shot-noise limited:

$$\sigma_{1,\text{in}}^2 = \frac{4\hat{q}^2 \eta \Delta f P_{\text{sens}}}{\bar{h}\nu}$$

[13]This may not be a good approximation for the 0 bit, since spontaneous-spontaneous beat noise may be significant. However, owing to the larger magnitude of σ_1, this approximation is not critical to equation (3.26), which is where σ_0 is ultimately used.

[14]To achieve a BER of 10^{-9}, which is a common standard, it can be shown that $\mathcal{Q} \approx 6$ in equation (3.25).

to obtain:

$$P_{\text{sens}} = \mathcal{Q}^2 \frac{\overline{h}\nu \Delta f F_n^{\text{eff}}}{\eta}. \tag{3.27}$$

Chapter 4

OFS Performance Evaluation

In this chapter, we focus our attention on the algorithmic implementation and performance evaluation of inter-MAN OFS communication. We focus on this mode of communication because the algorithms and analyses in Chapter 2, as well as the work of [110], are directly applicable to intra-MAN OFS communication. We begin by addressing the scheduling algorithm which arbitrates access to resources in OFS networks. The critical underpinnings of our proposed algorithm are that WAN wavelength channels are a precious resource and should therefore be efficiently utilized; and that traffic in the WAN will be sufficiently heavy and smooth such that a quasi-static logical topology is reasonable. Using our proposed algorithm, we then conduct an approximate throughput-delay analysis of OFS networks under inter-MAN communication. We remind the reader again that the physical layer and algorithmic implementations proposed in this chapter are envisioned as partial solutions in the broader context of hybrid networks.

This chapter is organized as follows. In the next section, we outline the physical layer and traffic assumptions employed throughout this chapter. In section 4.2, we begin with a motivating discussion on the properties of a sensible scheduling algorithm for inter-MAN OFS communication, and then proceed to propose our own simple algorithm which meets these criteria. In section 4.3, we derive an analytic approximation of the throughput-delay tradeoff for OFS under our scheduling algorithm. In section 4.4, we carry out a numerical study of our results in section 4.3 to further our understanding of the tradeoffs in the OFS architecture design space. We conclude this chapter in section 4.5.

■ 4.1 Modeling assumptions

In this section, we build upon the modeling assumptions in section 3.1 by elaborating our physical layer model in the next subsection, followed by a discussion of our traffic model in section 4.1.2.

■ 4.1.1 Network topology and other physical layer issues

In our network model, a single WAN connects n_w MANs, all of which employ OFS. As drawn in Figure 4-1, a MAN is connected to the WAN via a MAN node residing

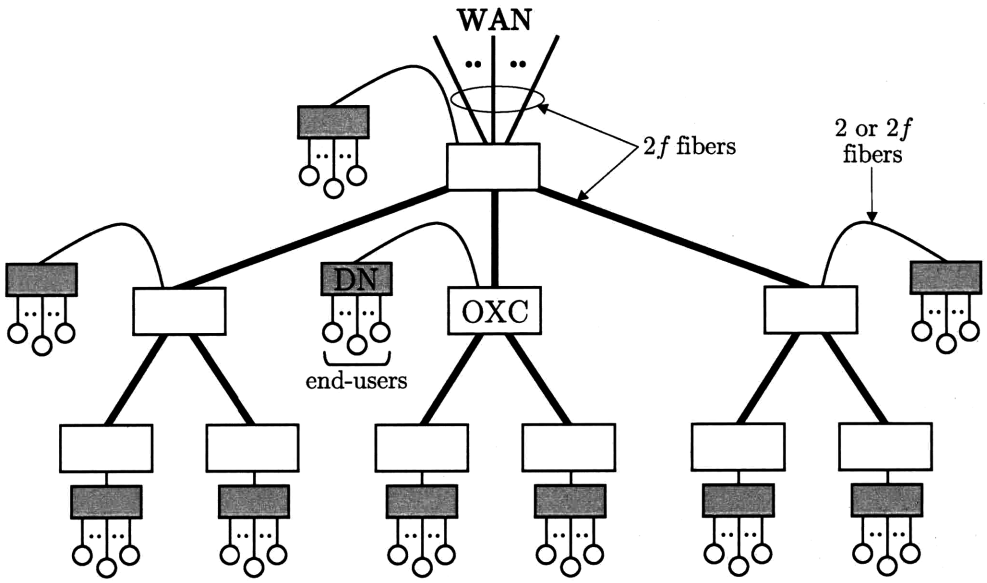
at the MAN-WAN interface. The wavelength channels provisioned for inter-MAN OFS communication reside within f fibers in each direction connecting this node to the rest of the WAN. Note that these fibers may carry traffic corresponding to other transport mechanisms besides OFS (e.g., EPS).

As discussed in section 3.1.2, our MAN physical topologies are based upon Moore Graphs[1]. An OFS MAN node comprises an OXC with direct connections to adjacent MAN nodes as well as one or more access networks based upon the optical DN architectures discussed in Chapter 3. We denote the total number of such DNs per MAN by $\tilde{n}_a$.

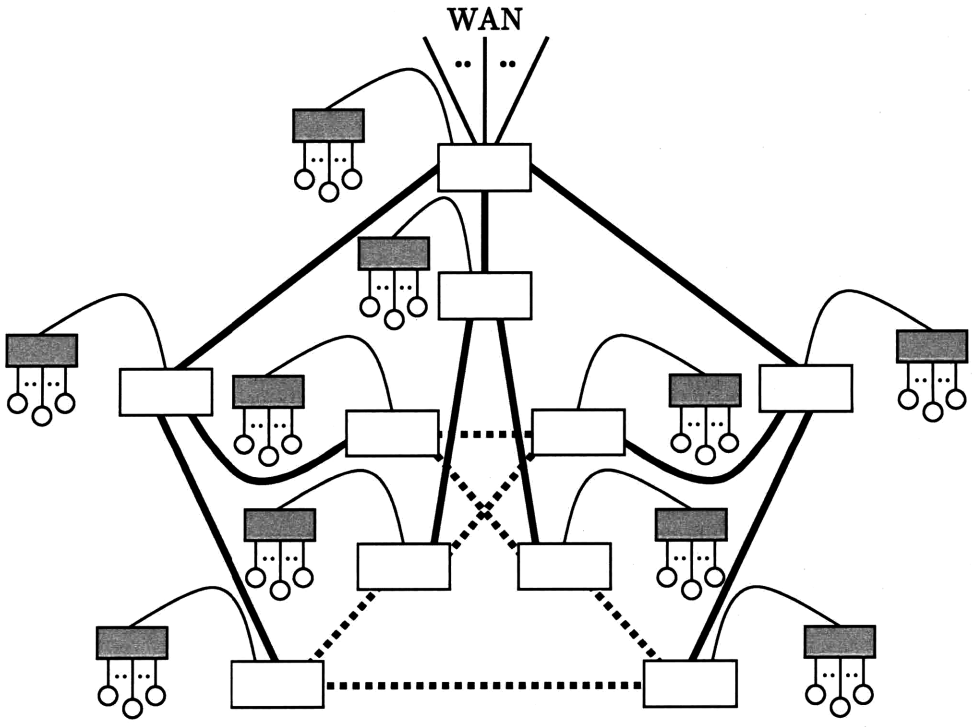
Recall that, under normal operating conditions, inter-MAN OFS traffic is assumed to be carried along the embedded tree of the MAN physical topology, whereas the portion of the topology outside the embedded tree carries OFS traffic that is only intra-MAN in nature. Since intra- and inter-MAN OFS traffic could coexist on the same fiber in the embedded tree, we shall assume that the wavelength channels employed for OFS communication are partitioned into w_l channels for local intra-MAN communication, and w_t channels for inter-MAN communication. This separation between wavelength channels for inter- and intra-MAN communication is made for analytical tractability, but may also prove to be sensible in implementing real networks, since it enables simpler resource scheduling decisions, albeit with some potential performance penalty. The decision as to how many wavelength channels to allocate to inter- and intra-MAN traffic will occur on coarse time-scales and will depend upon traffic statistics and requirements. This is an important issue, but is beyond the scope of this thesis.

Owing to fact that WAN wavelength channels are precious network resources, the scheduling algorithm that we employ should ensure that they are efficiently utilized. In order for our scheduling algorithm to do this in a manner that is not computationally prohibitive, we shall require that, for each WAN wavelength channel provisioned for inter-MAN OFS communication, there exists a dedicated wavelength channel in each link of the embedded tree in both source and destination MANs. For example, if the Boston MAN collectively has 50 wavelength channels provisioned for OFS communication to all other MANs in the WAN, then each link in the Boston MAN embedded tree has 50 dedicated wavelength channels, one corresponding to each of the 50 aforementioned provisioned WAN wavelength channels. If wavelength conversion exists at the MAN-WAN interface, then the colors of the wavelength channels in the MAN links need not be the same as their WAN counterparts. This flexibility enables the dedicated wavelength channels to be packed into a minimal number of fibers in an embedded tree link since wavelength colors for the MAN and WAN may be chosen independently. In the absence of wavelength conversion capability at the MAN-WAN interface, wavelength continuity between WAN and MAN wavelength channels must be respected. Thus, for a given routing and wavelength assignment (RWA) in the WAN, a larger than minimal number of fibers per embedded tree link is generally

[1](Generalized) Moore Graphs are discussed at length in Appendix B.



(a) Embedded tree portion of the OFS MAN.



(b) Mesh OFS MAN based upon the Petersen graph. Note that the tree topology in Figure 4-1(a) is embedded within this topology. Fiber links *not* in the embedded tree are drawn with dotted lines.

Figure 4-1. An example of an OFS MAN based upon a Moore Graph (Petersen graph) with $\Delta = 3$ and $d = 2$. A MAN node (drawn as a white box) comprises an OXC with one or more access DNs (drawn as grey boxes) connected. Optical amplifiers are not drawn.

required to support the dedicated wavelength channels. The exact number of fibers per link is very much situation dependent, so we shall assume for the remainder of this chapter that the links in the embedded tree topology include f fibers in each direction to ensure the one-to-one correspondence to the fibers connecting the MAN to the WAN[2]. In the absence of wavelength conversion at the MAN-WAN interface, the size of each OXC within the embedded tree can be minimized by not terminating each wavelength channel on each fiber with an OXC port. For example, a wavelength-selective switch (WSS)[3] could additionally be deployed before each OXC's input in order to block inactive wavelength channels while steering the active channels into the OXC input ports. This design comes, of course, at the price of the additional WSS.

With respect to the number of fibers connecting each DN to its parent MAN node, we consider two extreme cases: i) two fibers, one in each direction, and ii) $2f$ fibers, f in each direction such that there is one-to-one correspondence to the $2f$ fibers connecting the MAN to the WAN. Clearly, the performance corresponding to the second case will be better; but, as we shall see, the performance margin is not great under expected future network dimensions (i.e., number of nodes, volume of traffic, etc.). In addition to providing similar performance, the case of two fibers per DN is attractive in that it is a simpler and more scalable design: less hardware at end-users is required, and no modifications to this hardware are required as the number of fibers f in the MAN increases. Indeed, under this scenario, any upgrades or changes to the network occur strictly within the MAN. We shall also consider the role of wavelength conversion between each DN and its parent MAN node. It will be shown that wavelength conversion capability within each DN provides some (though not very significant) performance benefit under expected future network dimensions.

Let us now provide rough estimates for some of the parameters defined above. In the near- to medium-term future, we anticipate the number of MANs (n_w) to be on the order of several hundred[4], and the number of fibers f to be on the order of ten. Since, under the assumption of two fibers per DN, each DN can generate or sink at most one fiber worth of traffic, we have $\tilde{n}_a \geq f$. In fact, potential geographic constraints limiting the number of end-users per DN, along with the fact that the

[2]While we shall assume $2f$ fibers for each of these links, a smaller number of fibers per link may suffice for identical performance. Assume that the number of DNs in the subtree rooted at node i is $\tilde{n}_a^i$, and that each of these DNs is connected to node i with α fibers in each direction. If $2f_i$ fibers connect MAN node i to its parent node in the embedded tree, then:

$$f_i = \min(f, \alpha \tilde{n}_a^i)$$

ensures no contention for wavelength channels in the MAN among DNs.

[3]WSSs are briefly discussed in section A.8 of Appendix A.

[4]The US Census Bureau estimates that in 2006 the number of "incorporated places" (i.e., cities, towns, villages, boroughs, municipalities) in the US with populations over 100,000—good candidates for distinct MANs—to be approximately 260 [5].

bandwidth in a fiber will likely be shared with other transport mechanisms, suggests that $\tilde{n}_a \gg f$, with $\tilde{n}_a$ being on the order of a hundred or more.

■ 4.1.2 Traffic

As discussed earlier, we shall confine our attention in this chapter to serving the inter-MAN traffic demand. We shall assume uniformity in inter-MAN traffic demands, as well as uniformity in DN traffic demands. The former assumption implies that each MAN communicates with every other MAN on at most:

$$w_m = \frac{fw_t}{n_w - 1}$$

wavelength channels in each direction.

OFS traffic is assumed to be generated at end-users such that the aggregate traffic generated for a particular destination MAN arrives according to a Poisson process with rate λ_m. This Poisson assumption is reasonable for commercial networks, since the superposition of many stationary, identical, and independent point processes—which are reasonable models themselves for individual flow sources—is well-known to converge to a Poisson point process [75]. In the event that there do not exist sufficiently many flow sources to multiplex, we argue that OFS may not be an appropriate architecture. Indeed, this conclusion is supported by our work in the following chapter. (In contrast, defense networks may be lightly loaded, but may still require OFS for certain applications. In these networks, it is conceivable that WAN resources may be significantly over-provisioned for critical applications with very stringent delay deadlines—such as the transport of large amounts of sensor data relating to an imminent enemy threat—that may be best served by OFS. The capacity performance of such defense networks may be assessed using our work in Chapter 2. The issue of delay for such lightly loaded networks, however, is largely an open question which we do not pursue in this thesis, but which has been partially addressed in B. Ganguly's work in [110].)

The duration of flows are modeled as identical and independently distributed random variables with probability density function $p_L(\cdot)$ and k^{th} moment $\overline{L^k}$. Lastly, we consider only unicast transactions, although we recognize the increasing importance of multicast transactions, particularly for video content distribution.

■ 4.2 Scheduling algorithm for inter-MAN communication

Before outlining the scheduling algorithm used to serve inter-MAN traffic, we review some features of OFS networks that provide important guidelines for our scheduling algorithm design.

As discussed in the previous subsection, we focus on networks for which there exists significant multiplexing of flows in each MAN. Following the work of Cao *et al.* in [47, 48], appreciable statistical smoothing arises from the multiplexing of

many flows on wavelength channels. We emphasize that this applies to heavy-tailed flow length distributions exhibiting long-range temporal dependence despite erroneous claims to the contrary[5]. Such smoothing of aggregated traffic renders a quasi-static WAN logical topology sensible for serving this traffic. That is, changes to the WAN logical topology will be required on time-scales that are on the order of many flows[6]. Consequently, in the scheduling algorithm design for the scheduling of individual flows, it may be assumed that the wavelength channels provisioned for inter-MAN communication are static.

Owing to the high cost of provisioning wavelength channels in the wide-area, a scheduled approach should be employed which ensures high utilization of these resources. Ideally, this scheduled approach would be enable reservation of end-to-end optical paths without decoupling of the three geographic network tiers. However, such approaches are implementationally infeasible owing to the computational effort given the scale of the problem. In Chapter 2, for instance, we employ a scheduling algorithm that uses the maximum-weight stable set of the conflict graph at each network reconfiguration. Determining the maximum-weight stable set of a graph is known to be NP-complete [263]. Even polynomial-time heuristics, such as linear-programming relaxations (e.g., see [262,263]), require prohibitive computational effort given the size of an end-to-end OFS scheduling problem. Indeed, in OFS, the number of resources to be scheduled is on the order of the product of the number of wavelength channels per fiber and the number of DNs sharing a WAN, which could easily be on the order of 10^5. Running such algorithms to schedule individual flows would thus be infeasible[7].

Our approach in this chapter is to exploit the unique characteristics of the different geographic network tiers in OFS to devise an algorithm which, in addition to having constant complexity, lends to an analytic delay study. Our simple scheduling algorithm, in particular, reserves end-to-end optical paths for flow transmission via a sequential reservation process in which wavelength channels in the MAN and WAN are reserved first, followed by simultaneous but separate reservations of the source and destination DN wavelength channels. Owing to the fact that wavelength chan-

[5]In [47,48], it is shown that statistical multiplexing of flows does not change the *temporal* features of long-range dependence; but it does dampen the *magnitude of the excursions* from the mean of the aggregated traffic.

[6]We do not address the algorithms responsible for carrying out this slow reconfiguration of the WAN logical topology in this thesis.

[7]Distributed matching algorithms—tailored, in particular, to wireless settings—with constant complexities have recently been proposed [131, 191, 261]. These algorithms achieve complexities independent of network size by operating only on local information, and, in so doing, sacrifice a fraction of capacity. Analogous distributed, constant complexity algorithms for weighted stable set optimization would be attractive for OFS scheduling; but such algorithms only exist under special circumstances (i.e., fork-free graphs), in which case they may be mapped to distributed matching algorithms [194, 209].

nels in DNs are not heavily loaded, the latter reservation step should entail very little contention, thereby ensuring efficient end-to-end scheduling.

■ 4.2.1 Scheduling algorithm description

In the following description of the scheduling algorithm, we assume that each DN is connected to its MAN by two fibers. The case of $2f$ fibers is simpler and will be addressed at the end. In addition, we neglect the possibility of transmitter and receiver collisions which arise when two or more flows simultaneously require an end-user's transmitter or receiver, respectively. This is a reasonable assumption when each end-user transmits flows only occasionally, which we assume to be the case. We now illustrate the scheduling algorithm by stepping through the process by which an end-to-end all-optical path is established for the transmission of a particular flow. The algorithm is summarized in Figure 4-2.

Consider a flow that is generated at an end-user residing in DN D_s within MAN M_s and that is destined for an end-user residing in DN D_d within MAN M_d. As soon as this flow is ready for transmission, the source end-user sends a "primary request" r_w to the scheduling node associated with M_s requesting an end-to-end all-optical path for its flow transmission.

At a MAN's scheduling node, there exist $n_w - 1$ first-in first-out (FIFO) queues, one queue corresponding to every possible MAN destination. Each queue can be thought as the queue for an $M/G/w_m$ queueing system, in that the w_m wavelength channels dedicated to transmission from M_s to M_d eventually serve the primary requests waiting in it. After it arrives at M_s's scheduling node, r_w is placed at the end of the queue associated with M_d which we denote $Q_{M_s}^{M_d}$. Once r_w reaches the head of $Q_{M_s}^{M_d}$, r_w's flow is assigned to wavelength channel ω, the first of the w_m wavelength channels dedicated to transmission from M_s to M_d to have the primary request it is serving depart. After this wavelength channel assignment is made, an all-optical path is established on wavelength channel ω from the edge of D_s to the edge of D_d, passing through M_s, the WAN, and M_d. (Such a path is guaranteed to exist since there are $2f$ fibers within each link on this path, one of which with a dedicated ω wavelength channel for communication from M_s to M_d.) Now, in order to reserve the single outgoing ω wavelength channel in D_s and the single incoming ω wavelength channel in D_d, two additional secondary requests, r_s and r_d, respectively, are sent. During this process, r_w remains at the head of $Q_{M_s}^{M_d}$.

Secondary request r_s joins the end of the "source" secondary queue associated with D_s's ω wavelength channel, denoted by $\hat{Q}_{M_s}^{D_s}(\omega)$, which is physically located in M_s's scheduling node; and secondary request r_d joins the end of the "destination" secondary queue associated with D_d's ω wavelength channel, denoted by $\overline{Q}_{M_d}^{D_d}(\omega)$, that is physically located in M_d's scheduling node. These queues contain secondary requests to use the ω wavelength channel on D_s's outgoing fiber and D_d's incoming fiber, respectively. Note that at any instant in time there can be at most one secondary

1. A source node residing in DN D_s within MAN M_s indicates its desire to transmit a flow to an end-user residing in DN D_d within MAN M_d by sending a primary request r_w to the scheduling node associated with M_s, where it is enqueued at FIFO queue $Q_{M_s}^{M_d}$.

2. Once r_w reaches the head of $Q_{M_s}^{M_d}$, r_w's flow is assigned to wavelength channel ω, the first of the w_m wavelength channels dedicated to transmission from M_s to M_d to have a primary request depart. An all-optical path is established on wavelength channel ω from the edge of D_s to the edge of D_d, passing through M_s, the WAN, and M_d.

3. To reserve the ω wavelength channel on D_s's outgoing fiber and D_d's incoming fiber:

   (a) Secondary request r_s is sent to FIFO source secondary queue $\hat{Q}_{M_s}^{D_s}(\omega)$ located within M_s's scheduling node.

   (b) Secondary request r_d is sent to FIFO secondary queue $\overline{Q}_{M_d}^{D_d}(\omega)$ located within M_d's scheduling node.

4. When r_s and r_d reach the heads of their respective secondary queues, they each notify both M_s's and M_d's 's scheduling nodes. As soon as M_s's and M_d's scheduling nodes have received *both* notifications, flow transmission may begin.

5. After the flow transmission is complete, r_w, r_s, and r_d depart their queues, thereby freeing up their reserved network resources.

Figure 4-2. Summary of the scheduling algorithm for inter-MAN OFS communication.

request associated with flows destined for M_d. When r_s and r_d each reach the heads of their respective secondary queues (which we assume to be FIFO), they each notify both M_s's and M_d's 's scheduling nodes. As soon as M_s's and M_d's scheduling nodes have received *both* notifications, they instruct the source and destination end-users, respectively, to begin flow transmission immediately. After the flow transmission is complete, r_w, r_s, and r_d depart their queues.

The scheduling algorithm for the case in which a DN is connected to its MAN with $2f$ fibers is simpler because the source and destination DN queues $\hat{Q}_{M_s}^{D_s}(\omega)$ and $\overline{Q}_{M_d}^{D_d}(\omega)$ are no longer necessary. This is because the correspondence between each of these fibers and the $2f$ fibers in the MAN and WAN preclude contention for DN wavelength channel resources among flows with different destination MANs. Thus, for the case of $2f$ fibers, once primary request r_w reaches the head of $Q_{M_s}^{M_d}$, instruction is sent to the scheduling node associated with M_d is to set up an all-optical path in

M_d from the WAN to D_d on channel ω. The two scheduling nodes then instruct the source and destination end-users to begin flow transmission immediately.

■ 4.3 Performance analysis

In the following performance analysis, we neglect propagation delay of requests and transactions.

In our description of the scheduling algorithm for inter-MAN communication in section 4.2.1, we mentioned that primary requests for a source-destination MAN pair may be modeled as customers of an $M/G/w_m$ queueing system with arrival rate λ_m. The service time in this model is the time spent by a primary request r_w at the head of its primary queue, which comprises flow transmission time in addition to the time spent reserving wavelength channels in the source and destination DNs (in the case of two fibers per DNs). Unfortunately, there is no exact solution for the $M/G/w_m$, except in the special cases of exponential service times, $w_m = 1$, or $w_m = \infty$. We must therefore resort to an approximate performance analysis of an OFS network.

At high offered traffic loads, the $M/G/w_m$ queueing system is well approximated by the $M/G/1$ queueing system operating under the same normalized traffic load (i.e., $1/w_m$ of the load of the former system). The approximation, however, tends to be somewhat loose at medium offered loads and especially so at light loads, where the $M/G/\infty$ system is a better approximation. While light loading may be a highly unlikely operating point owing to the high cost of provisioning MAN and WAN wavelength channels, a medium load is realistic since it may provide a good balance between wavelength channel utilization and queueing delay experienced by flows. A more accurate approximation of the performance of the $M/G/w_m$ queueing system is therefore needed. Because most of such approximations—including the most common one, which we focus on in this chapter—require an analysis of the analogous single server system [37], we address the performance of the single server queueing system model of our OFS network.

To this end, we randomly split the Poisson process of intensity λ_m representing the flow arrivals of each source-destination MAN pair into w_m "child" Poisson processes of intensity:

$$\lambda_c = \frac{\lambda_m}{w_m},$$

one child Poisson arrival process for each wavelength channel dedicated to the source-destination MAN pair. We also replace queue $Q_{M_s}^{M_d}$ in MAN M_s by w_m independent, parallel queues:

$$Q_{M_s}^{M_d}(\omega_1), Q_{M_s}^{M_d}(\omega_2), \ldots, Q_{M_s}^{M_d}(\omega_{w_m}),$$

each corresponding to a wavelength channel dedicated to the source-destination MAN pair and accepting primary requests from the corresponding child Poisson process.

To compute the queueing delay of this system we shall eventually apply the Pollaczek-Khinchin (P-K) formula to the single server queue $Q_{M_s}^{M_d}(\omega)$ accepting pri-

mary requests from flows generated in M_s, destined for M_d, and employing wavelength channel ω. The arrival rate to this queue is λ_c, and we shall denote the service time of a primary request in this queue by X with k^{th} moment $\overline{X^k}$. In the case of $2f$ fibers per DN, we simply have:

$$X = L + \tau,$$

where L is the flow transmission time and τ is the hardware reconfiguration time, since no additional time is required to reserve resources at the source and destination DNs. However, in the case of two fibers per DN, X comprises not only the sum of the flow transmission time L and hardware configuration time τ, but also any time spent at the head of queue $Q_{M_s}^{M_d}(\omega)$ reserving source and destination DN resources. We therefore generally express the average service time $\overline{X}$ as:

$$\begin{aligned} \overline{X} &= \overline{L} + \overline{Y} + \tau \\ &\approx \overline{L} + \overline{Y}, \end{aligned} \tag{4.1}$$

where $\overline{Y}$ is the average time spent at the head of the primary queue reserving resources in both the source and destination DNs, and is equal to zero in the case of $2f$ fibers per DN. The approximation for the average service time neglects the hardware reconfiguration time, because in OFS's most sensible regime of operation, as characterized in the next chapter, $\overline{L} + \overline{Y} \gg \tau$. More concretely, $\overline{L} + \overline{Y}$ is, at the very least, on the order of hundreds of milliseconds when OFS is economically viable, whereas τ is likely to be on the order of a few tens of milliseconds.

Recall that, in the case of two fibers per DN, after a primary request reaches the head of $Q_{M_s}^{M_d}(\omega)$, secondary requests are sent to each of $\hat{Q}_{M_s}^{D_s}(\omega)$ and $\overline{Q}_{M_d}^{D_d}(\omega)$, where D_s and D_d are the source and destination DNs of the flow, respectively. At each of these two secondary queues there could be up to $f-1$ other secondary requests associated with other destination and source MANs, respectively[8]. However, since flows are equally likely to be generated at, or destined for each of the DNs in a MAN, the probability that a primary request at $Q_{M_s}^{M_d}(\omega)$ generates a secondary request to a particular $\hat{Q}_{M_s}^{i}(\omega)$ or $\overline{Q}_{M_d}^{j}(\omega)$ is $1/\tilde{n}_a$. The arrival rate of secondary requests to a secondary queue contributed from each of the f contending primary queues is $\lambda_c/\tilde{n}_a$, for an aggregate arrival rate at each secondary queue of $f\lambda_c/\tilde{n}_a \ll 1$. Now, for a fixed aggregate arrival rate of $f\lambda_c/\tilde{n}_a$, as f and $\tilde{n}_a$ get proportionately large—corresponding to a large WAN and large MANs, respectively—the arrival process of secondary requests to each secondary queue is known to converge to a Poisson process [75]. We thus model the arrival process to each secondary queue as a Poisson process of rate $f\lambda_c/\tilde{n}_a$, with the approximation becoming increasingly accurate as f and $\tilde{n}_a$ become large.

[8]The upper bound on the number of potential contending secondary requests is equal to the number of times a wavelength channel is reused over the f fibers connecting the MAN to the WAN. We shall assume it to be f for simplicity and to maintain a conservative performance analysis.

We now address the calculation of the moments of Y. Recall that Y is the time spent at the head of the primary queue reserving resources in both the source and destination DNs, and is thus the maximum of the two queueing delays experienced by the two peer secondary requests. Since the service time of each secondary request already enqueued at a secondary queue is itself coupled to the state of the queue in which its peer secondary resides, the characterization of Y is quite difficult. Thus, given the Poisson assumption of secondary request arrivals to secondary queues, we derive upper and lower bounds for $\overline{Y}$ and $\overline{Y^2}$ which will be used in conjunction with the P-K formula to obtain optimistic and pessimistic approximations, respectively, for the queueing delay experienced by a flow. These approximations do not serve as strict bounds since the Poisson assumption is an approximation.

■ 4.3.1 Optimistic approximation for primary request service time

We may compute simple lower bounds for $\overline{Y}$ and $\overline{Y^2}$ by viewing the time spent to reserve the source and destination DNs as equal to the time spent by a single secondary request in a single queue with service time drawn from $p_L(\cdot)$. This bound can be thought as arising if only one of the two DNs needs to be reserved rather than both. In this case, the P-K formula yields the following bound for the first moment of Y:

$$\overline{Y} > \overline{Y_l} = \frac{f\lambda_c \overline{L^2}}{2\left(\tilde{n}_a - f\lambda_c \overline{L}\right)},$$

and the following bound for the second moment of Y:

$$\begin{aligned}\overline{Y^2} > \overline{Y_l^2} &= 2\overline{Y_l}^2 + \frac{f\lambda_c \overline{L^3}}{3\left(\tilde{n}_a - f\lambda_c \overline{L}\right)} \\ &= \frac{\left(f\lambda_c \overline{L^2}\right)^2}{2\left(\tilde{n}_a - f\lambda_c \overline{L}\right)^2} + \frac{f\lambda_c \overline{L^3}}{3\left(\tilde{n}_a - f\lambda_c \overline{L}\right)},\end{aligned}$$

where the first equality directly follows from the Takács recurrence formula for the moments of the waiting time in an $M/G/1$ queueing system (e.g., see [171]). These bounds can now be employed to yield the following approximations for the first two moments of the service time in the primary request queue:

$$\begin{aligned}\overline{X} &\approx \overline{L} + \overline{Y_l} \\ &= \overline{L} + \frac{f\lambda_c \overline{L^2}}{2\left(\tilde{n}_a - f\lambda_c \overline{L}\right)} \\ \overline{X^2} &\approx \overline{(L + Y_l)^2} \\ &= \overline{L^2} + 2\overline{L} \cdot \overline{Y_l} + \overline{Y_l^2}\end{aligned} \tag{4.2}$$

$$= \frac{\tilde{n}_a \overline{L^2}}{\tilde{n}_a - f\lambda_c \overline{L}} + \frac{\left(f\lambda_c \overline{L^2}\right)^2}{2\left(\tilde{n}_a - f\lambda_c \overline{L}\right)^2} + \frac{f\lambda_c \overline{L^3}}{3\left(\tilde{n}_a - f\lambda_c \overline{L}\right)}. \tag{4.3}$$

We note that the above expressions for $\overline{X}$ and $\overline{X^2}$ include contributions from $\overline{L^2}$ and $\overline{L^3}$, respectively. This (undesirable) proportionality to higher-order moments of L arises from the service time X containing a queueing delay term, itself proportional to the second moment of L. However, the contributions from these higher-order moments of L can be made arbitrarily small by designing the network such that $\tilde{n}_a \gg f\lambda_c \overline{L}$. Since $f\lambda_c \overline{L}$ is interpreted as the load offered to each wavelength channel *color* in a MAN, designing the network in this way ensures that there is little contention for each wavelength channel in each DN.

■ 4.3.2 Pessimistic approximation for primary request service time

In order to compute upper bounds for $\overline{Y}$ and $\overline{Y^2}$, we consider secondary requests, instead of being sent simultaneously after a primary request reaches the head of its queue, are sent sequentially. Specifically, we assume that the secondary request reserving the destination DN is sent only *after* the secondary request responsible for reserving the source DN reaches the head of its queue in effect reserving the source DN. Thus we have:

$$\overline{Y} < \overline{Y_u} = \overline{Z_s} + \overline{Z_d},$$

where Z_s is the time spent by a secondary source request in its queue prior to reaching the head of the queue; and Z_d is the time spent by a secondary destination request in its queue prior to reaching the head of its queue.

By invoking the Poisson approximation for the arrival process to the secondary destination queue, we may thus treat the queue as an $M/G/1$ queueing system. We therefore have the following for the first three moments of Z_d:

$$\overline{Z_d} = \frac{f\lambda_c \overline{L^2}}{2\left(\tilde{n}_a - f\lambda_c \overline{L}\right)}$$

$$\overline{Z_d^2} = \frac{\left(f\lambda_c \overline{L^2}\right)^2}{2\left(\tilde{n}_a - f\lambda_c \overline{L}\right)^2} + \frac{f\lambda_c \overline{L^3}}{3\left(\tilde{n}_a - f\lambda_c \overline{L}\right)}$$

$$\overline{Z_d^3} = 3\overline{Z_d} \cdot \overline{Z_d^2} + \frac{f\lambda_c \overline{L^3} \cdot \overline{Z_d}}{\tilde{n}_a - f\lambda_c \overline{L}} + \frac{f\lambda_c \overline{L^4}}{4\left(\tilde{n}_a - f\lambda_c \overline{L}\right)}$$

$$= \frac{f\lambda_c \overline{L^2}}{\left(\tilde{n}_a - f\lambda_c \overline{L}\right)^2} \left\{ \frac{3\left(f\lambda_c \overline{L^2}\right)^2}{4\left(\tilde{n}_a - f\lambda_c \overline{L}\right)} + f\lambda_c \overline{L^3} \right\} + \frac{f\lambda_c \overline{L^4}}{4\left(\tilde{n}_a - f\lambda_c \overline{L}\right)},$$

where the Takács recurrence formula has been used in all three cases.

To compute the first two moments of Z_s, we invoke the Poisson approximation for the arrival process of secondary requests to the secondary source queue, and thus treat the queue as an $M/G/1$ queueing system. Note that in this queueing system, the service time of a customer is $L + Z_d$. Thus, we have the following for the first moment of Z_s:

$$\begin{aligned}\overline{Z_s} &= \frac{f\lambda_c\overline{(L+Z_d)^2}}{2\left(\tilde{n}_a - f\lambda_c\left(\overline{L}+\overline{Z_d}\right)\right)} \\ &= \frac{f\lambda_c\left(\overline{L^2}+2\overline{L}\cdot\overline{Z_d}+\overline{Z_d^2}\right)}{2\left(\tilde{n}_a - f\lambda_c\overline{L} - f\lambda_c\overline{Z_d}\right)},\end{aligned}$$

and the following for the second moment of Z_s:

$$\begin{aligned}\overline{Z_s^2} &= 2\left\{\frac{f\lambda_c\overline{(L+Z_d)^2}}{2\left(\tilde{n}_a - f\lambda_c\left(\overline{L}+\overline{Z_d}\right)\right)}\right\}^2 + \frac{f\lambda_c\overline{(L+Z_d)^3}}{3\left(\tilde{n}_a - f\lambda_c\overline{L} - f\lambda_c\overline{Z_d}\right)} \\ &= \frac{1}{2}\left\{\frac{f\lambda_c\left(\overline{L^2}+2\overline{L}\cdot\overline{Z_d}+\overline{Z_d^2}\right)}{\tilde{n}_a - f\lambda_c\overline{L} - f\lambda_c\overline{Z_d}}\right\}^2 + \frac{f\lambda_c\left(\overline{L^3}+3\overline{L^2}\cdot\overline{Z_d}+3\overline{L}\cdot\overline{Z_d^2}+\overline{Z_d^3}\right)}{3\left(\tilde{n}_a - f\lambda_c\overline{L} - f\lambda_c\overline{Z_d}\right)},\end{aligned}$$

where the previous expressions for the moments of Z_d should be substituted in.

We are now able to form the following pessimistic approximations for the first two moments of the service time in the primary request queue:

$$\begin{aligned}\overline{X} &\approx \overline{L} + \overline{Y_u} \\ &= \overline{L} + \overline{Z_s} + \overline{Z_d} \qquad (4.4) \\ \overline{X^2} &\approx \overline{(L+Y_u)^2} \\ &= \overline{L^2} + \overline{Z_s^2} + \overline{Z_d^2} + 2\overline{L}\cdot\overline{Z_s} + 2\overline{L}\cdot\overline{Z_d} + 2\overline{Z_s}\cdot\overline{Z_d}. \qquad (4.5)\end{aligned}$$

As in the optimistic approximation for primary request service time, these expressions for $\overline{X}$ and $\overline{X^2}$ include contributions from higher-order moments of L. In fact, the contributions are from even higher moments: $\overline{L^3}$ and $\overline{L^4}$ corresponding to $\overline{X}$ and $\overline{X^2}$, respectively. The reason that yet higher-order moments of L arise is because the service time X contains a queueing delay term nested within another queueing delay term, which itself is proportional to the second moment of L. However, as before, the contributions from these higher-order moments of L can be made arbitrarily small by designing the network such that $\tilde{n}_a \gg f\lambda_c\overline{L}$, which results in significantly reduced contention for resources in DNs.

■ 4.3.3 Wavelength conversion in DNs

In the presence of wavelength conversion capability at the head-end of each DN, there is only one secondary source queue and one secondary destination queue associated with each DN. Each of these queues is served by w_t wavelength channels akin to an $M/G/w_t$ queueing system. In this case, the maximum number of contending secondary requests at a queue would be $(n_w - 1)w_m$ rather than f as in the case of no wavelength conversion. However, since there are w_t candidate wavelength channels serving this queue instead of just one, the normalized arrival rate of secondary requests to a secondary queue is again $f\lambda_c/\tilde{n}_a \ll 1$. The convergence of this arrival process to Poisson is faster in this case than in the case of no wavelength conversion owing to the fact that $(n_w - 1)w_m > f$. Furthermore, the delay experienced by secondary requests at these secondary queues will be less than in the case of no wavelength conversion owing to the statistical smoothing gain.

To evaluate quantitatively the approximate performance gain provided by wavelength conversion capability in DNs, we invoke the following well-known, simple approximation of the delay experienced by a customer in the $M/G/k$ queueing system *prior* to reaching the head of the queue [37]:

$$\hat{W}_{G,k}(k\rho, p_L) \approx \frac{\hat{W}_{G,1}(\rho, p_L)\hat{W}_{M,k}(k\rho, p_L)}{\hat{W}_{M,1}(\rho, p_L)}, \tag{4.6}$$

where $\hat{W}_{G,k}(\cdot,\cdot)$ and $\hat{W}_{M,k}(\cdot,\cdot)$ are the queueing delays of the $M/G/k$ and $M/M/k$ queueing systems, respectively, as functions of offered traffic load and service distribution. We point out that $\hat{W}_{G,1}(\rho, p_L)$ is given simply by the P-K formula and thus depends only on the first two moments of the distribution p_L; whereas $W_{M,k}(k\rho, p_L)$ is given by (e.g., see [171]):

$$W_{M,k}(k\rho, p_L) = \frac{\overline{L}P_Q}{k(1-\rho)},$$

with P_Q given by the Erlang C formula:

$$P_Q = \frac{(k\rho)^k}{k!(1-\rho)}\left[\sum_{i=0}^{k-1}\frac{(k\rho)^i}{i!} + \frac{(k\rho)^k}{k!(1-\rho)}\right]^{-1},$$

and thus depends only on the first moment of the distribution p_L.

Optimistic approximation for primary request service time

To modify the previous optimistic approximation to account for wavelength conversion, we simply scale Y_l by a constant factor:

$$\alpha_t \equiv \frac{\hat{W}_{M,w_t}\left(w_t\lambda_c\overline{L}, p_L\right)}{\hat{W}_{M,1}\left(\lambda_c\overline{L}, p_L\right)},$$

where the previous formula may be substituted in. The optimistic approximations for the first two moments of X then become:

$$\overline{X} \approx \overline{L} + \frac{\alpha_t f\lambda_c\overline{L^2}}{2\left(\tilde{n}_a - f\lambda_c\overline{L}\right)} \tag{4.7}$$

$$\overline{X^2} \approx \frac{\tilde{n}_a\overline{L^2} + f\lambda_c\overline{L}\cdot\overline{L^2}(\alpha_t - 1)}{\tilde{n}_a - f\lambda_c\overline{L}} + \alpha_t^2\left[\frac{\left(f\lambda_c\overline{L^2}\right)^2}{2\left(\tilde{n}_a - f\lambda_c\overline{L}\right)^2} + \frac{f\lambda_c\overline{L^3}}{3\left(\tilde{n}_a - f\lambda_c\overline{L}\right)}\right]. \tag{4.8}$$

The interpretation of these expressions is virtually identical to that of equations (4.2) and (4.3).

Pessimistic approximation for primary request service time

To modify the previous pessimistic approximation to account for wavelength conversion, we must similarly scale $\overline{Z_s}$ and $\overline{Z_d}$. As above, $\overline{Z_d}$ is scaled by the constant factor α_t. However, since:

$$Z_s \approx L + \alpha_t Z_d,$$

we must scale $\overline{Z_s}$ by the following factor:

$$\alpha_t' \equiv \frac{\hat{W}_{M,w_t}\left[w_t\lambda_c\left(\overline{L} + \alpha_t\overline{Z_d}\right), p_{L+\alpha_t Z_d}\right]}{\hat{W}_{M,1}\left[\lambda_c\left(\overline{L} + \alpha_t\overline{Z_d}\right), p_{L+\alpha_t Z_d}\right]}.$$

The pessimistic approximations for the first two moments of X then become:

$$\overline{X} \approx \overline{L} + \alpha_t'\overline{Z_s} + \alpha_t\overline{Z_d}$$

$$\overline{X^2} \approx \overline{L^2} + \left(\alpha_t'\right)^2\overline{Z_s^2} + \alpha_t^2\overline{Z_d^2} + 2\alpha_t'\overline{L}\cdot\overline{Z_s} + 2\alpha_t\overline{L}\cdot\overline{Z_d} + 2\alpha_t\alpha_t'\overline{Z_s}\cdot\overline{Z_d}.$$

Again, the interpretation of these expressions is virtually identical to that of equations (4.4) and (4.5).

■ 4.3.4 Total queueing delay

Given the above optimistic and pessimistic approximations for the first two moments of X, we now turn to computing the total queueing delay seen by a transaction. To

do this, we invoke the P-K formula with respect to the primary request queue with an additional term reflecting the queueing delay experienced by a primary request while at the head of its queue:

$$\hat{W}_{G,1} = \overline{Y} + \frac{\lambda_c \overline{X^2}}{2\left(1 - \lambda_c \overline{X}\right)}, \tag{4.9}$$

with the following stability condition:

$$\lambda_c \overline{X} < 1.$$

We now invoke the $M/G/k$ approximation in equation (4.6) in conjunction with equation (4.9) to obtain the following approximation of the total queueing delay experienced by a flow, including the time spent reserving DN wavelength channels just prior to flow transmission:

$$W \approx \left[\overline{Y} + \frac{\lambda_c \overline{X^2}}{2\left(1 - \lambda_c \overline{X}\right)}\right] \left[\frac{\hat{W}_{M,w_m}\left(\lambda_m \overline{X}, p_X\right)}{\hat{W}_{M,1}\left(\lambda_c \overline{X}, p_X\right)}\right]. \tag{4.10}$$

We note that there are alternative approximations of the $M/G/w_m$ queue, but these require knowledge of the analogous $M/G/1$ system or the first two moments of the service time X [37], all of which have been addressed above. Equations (4.2) and (4.3), or equations (4.4) and (4.5), or their wavelength conversion counterparts may be substituted into equation (4.10) ultimately yielding optimistic and pessimistic approximations of W, respectively.

The throughput of a WAN wavelength channel is given simply by:

$$S = \lambda_c \overline{L} \tag{4.11}$$

$$1 > \lambda_c \overline{X}. \qquad \text{(Stability condition)}$$

For a given flow length distribution $p_L(\cdot)$, a bound on the maximum achievable throughput may be obtained by substituting either equation (4.2) or (4.7) into equation (4.11), and then maximizing over all values of λ_c, subject to the above stability condition. By carrying out this optimization for equation (4.2), the throughput of a WAN wavelength channel is bounded as follows:

$$S_{\max} < \begin{cases} \dfrac{\overline{L}^2\left(1+\frac{\tilde{n}_a}{f}\right) - \overline{L}\sqrt{\overline{L}^2\left(1+\frac{\tilde{n}_a}{f}\right)^2 + 2\frac{\tilde{n}_a}{f}\left(\alpha_{wc}\overline{L^2} - 2\overline{L}^2\right)}}{2\overline{L}^2 - \alpha_{wc}\overline{L^2}}, & \text{if } 2\overline{L}^2 \neq \alpha_{wc}\overline{L^2}, \\ \dfrac{\tilde{n}_a}{\tilde{n}_a + f}, & \text{if } 2\overline{L}^2 = \alpha_{wc}\overline{L^2}, \end{cases} \tag{4.12}$$

where $\alpha_{wc} = \alpha_t$ if wavelength conversion exists in DNs, and $\alpha_{wc} = 1$ otherwise. Note that exponential flows correspond to the case of $2\overline{L}^2 = \alpha_{wc}\overline{L^2}$, whereas constant and

heavy-tailed flows correspond to the first case. The above throughput maximization, recall, neglects any hardware configuration time τ, a reasonable assumption provided that transaction lengths are on the order of hundreds of milliseconds or larger. Moreover, the above throughput maximization of a WAN wavelength channel implies a tolerance to unbounded delay. We invoke this condition delay-insensitivity in our comparison of OFS to other architectures in Chapter 5.

■ 4.4 Numerical results

In Figures 4-3, 4-4, and 4-5, the approximations of the queueing delay derived from equation (4.10) are plotted versus WAN wavelength channel utilization (equation (4.11)) for three flow length distributions: constant, exponential, and truncated heavy-tailed, respectively. Figure 4-6 superimposes the optimistic approximations of the aforementioned three figures on the same axes.

For each set of flow length distributions, the approximations, which differ only in the way DN reservations are carried out, yield similar performances at low loads since there is very little contention for DN resources. The performances diverge with increasing traffic load—especially between the case of no wavelength conversion (i.e., red curves) and the other cases (i.e., blue and black curves)—owing to the increasing role of DN reservation time. The performance difference among the approximations is greatest for the truncated heavy-tailed distribution, followed by the exponential distribution, and then constant length flows. Indeed, it is in this order of flow distributions that wavelength conversion becomes less useful in ensuring that flows do not get delayed by extremely large flows ahead of them. In comparing performance across distributions, we observe that constant length flows offer the best delay-throughput tradeoff, followed by exponentially distributed flows, and then truncated heavy-tailed distributed flows. These results again indicate very large flows impeding subsequent flows to the greatest extent in the truncated heavy-tailed distribution and to the least extent in the constant distribution.

In Figure 4-7, we illustrate the impact of the number of DNs per MAN on the delay-throughput tradeoff presented by equations (4.10) and (4.11). As expected, the gap between the optimistic and pessimistic approximations for a the same $\tilde{n}_a$ narrows as the number of DNs per MAN increases. In addition, the performance of the two fibers per DN case converges to that of the $2f$ fibers per DN case as the number of DNs per MAN increases. This, of course, is because as $\tilde{n}_a$ increases, the amount of traffic per DN decreases, ultimately resulting in less contention for DN resources. Figure 4-8 illustrates the relationship between the maximum throughput possible—that is, assuming a tolerance to unbounded delay—and the number of DNs per MAN, as captured by equation (4.12). For low values of $\tilde{n}_a$, the marginal throughput gain from increasing $\tilde{n}_a$ is large, but diminishes when $\tilde{n}_a$ grows beyond 100 or so. Moreover, the throughput gain with increasing $\tilde{n}_a$ is fastest with constant length transactions and slowest with heavy-tailed transactions. The explanation for this trend is, again,

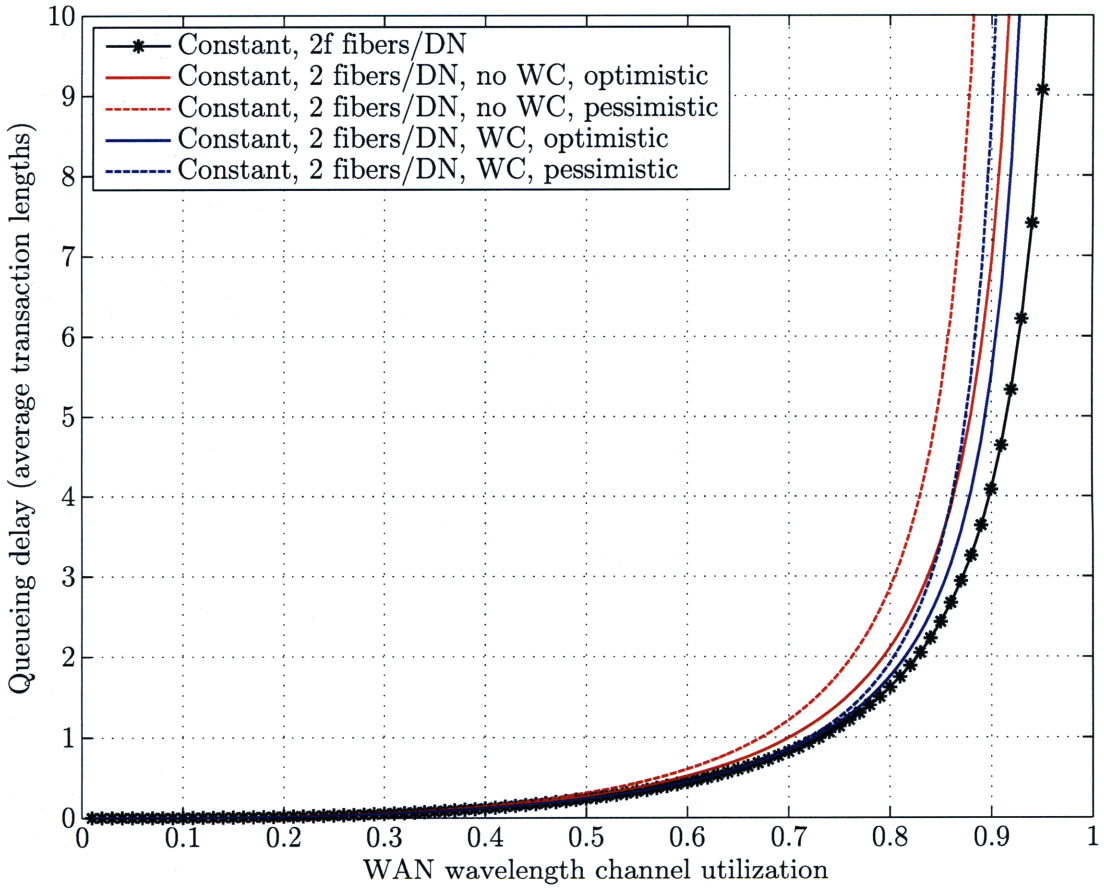


Figure 4-3. Different approximations of queueing delay versus throughput for DNs with two fibers and $2f$ fibers and constant length flows of duration 10s. It is assumed that $\tilde{n}_a = 200$, $f = 15$, $w_t = 20$, and $w_m = 3$.

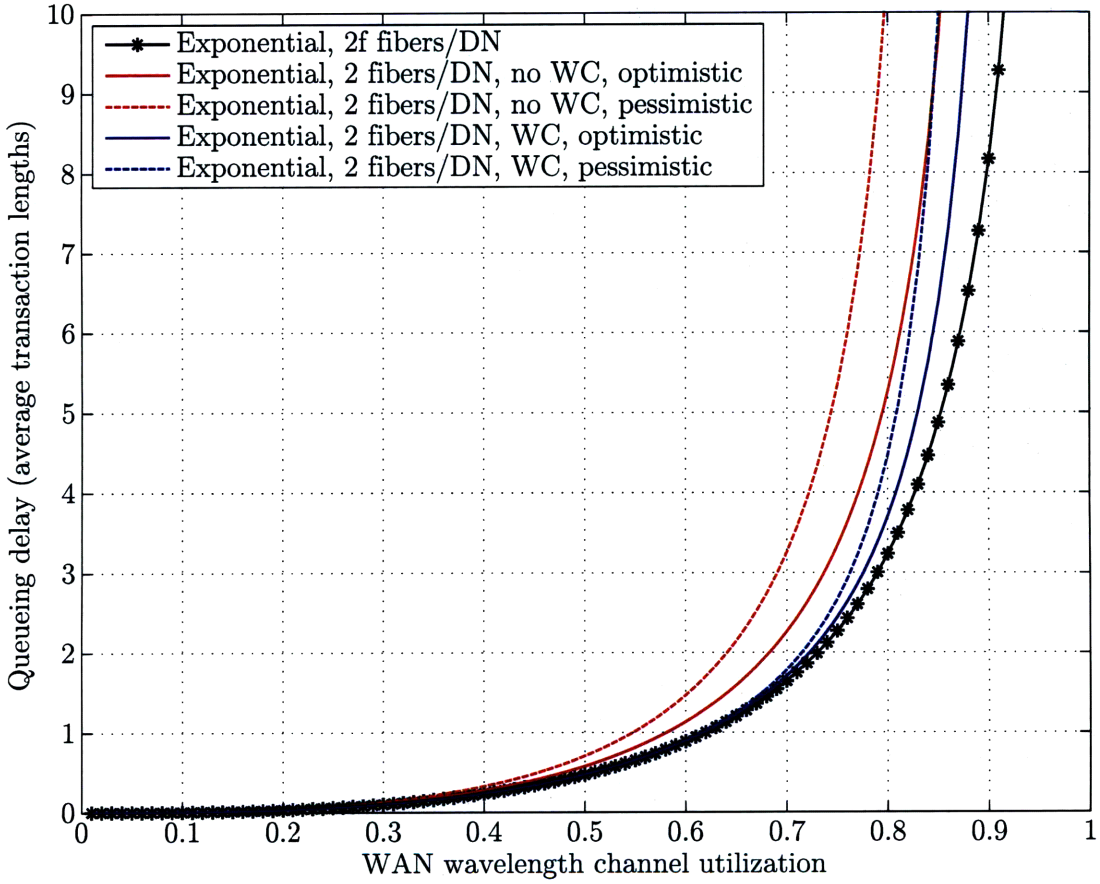


Figure 4-4. Different approximations of queueing delay versus throughput for DNs with two fibers and $2f$ fibers. In all cases, flow lengths are exponentially distributed with average duration 10s. It is assumed that $\tilde{n}_a = 200$, $f = 15$, $w_t = 20$, and $w_m = 3$.

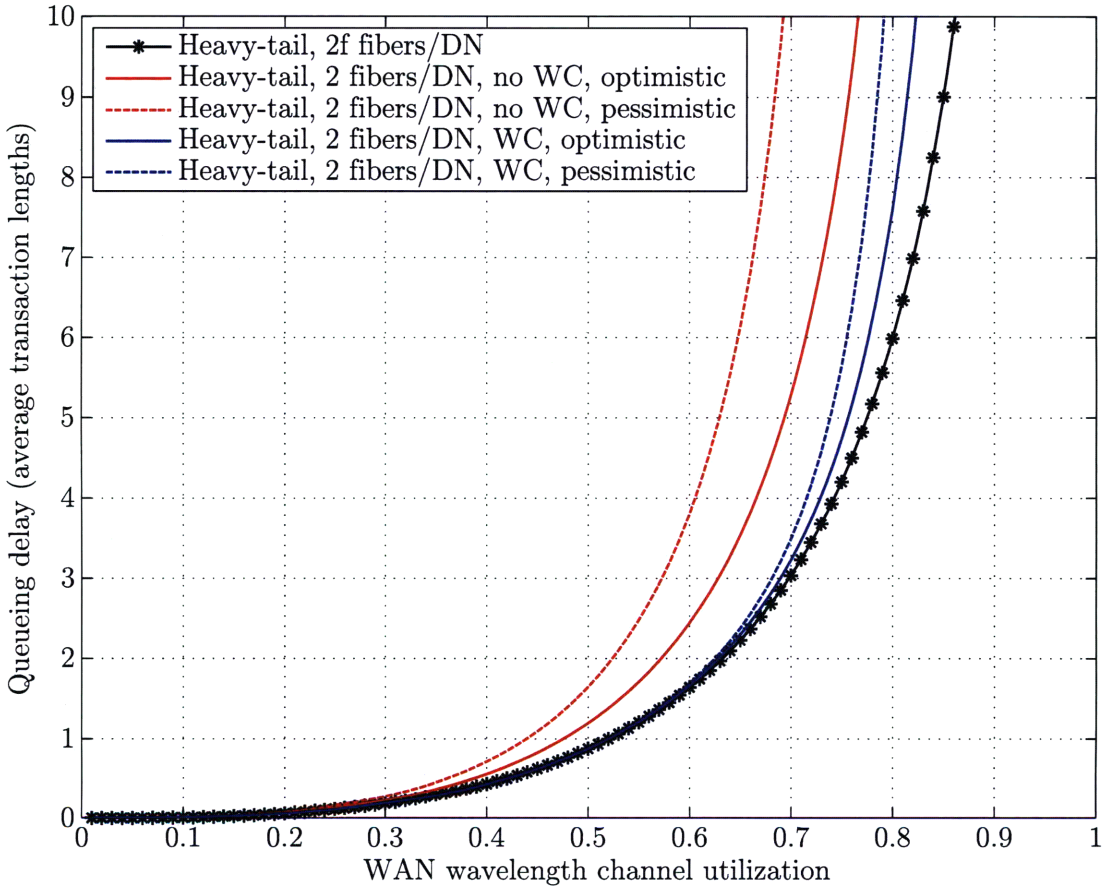


Figure 4-5. Different approximations of queueing delay versus throughput for DNs with two fibers and $2f$ fibers. In all cases, flows lengths are drawn from a truncated heavy-tailed distribution with: a tail that decays with exponent -1.5, an expected flow duration of 10s, and lower and upper flow length bounds of 1s and 100s, respectively. It is assumed that $\tilde{n}_a = 200$, $f = 15$, $w_t = 20$, and $w_m = 3$.

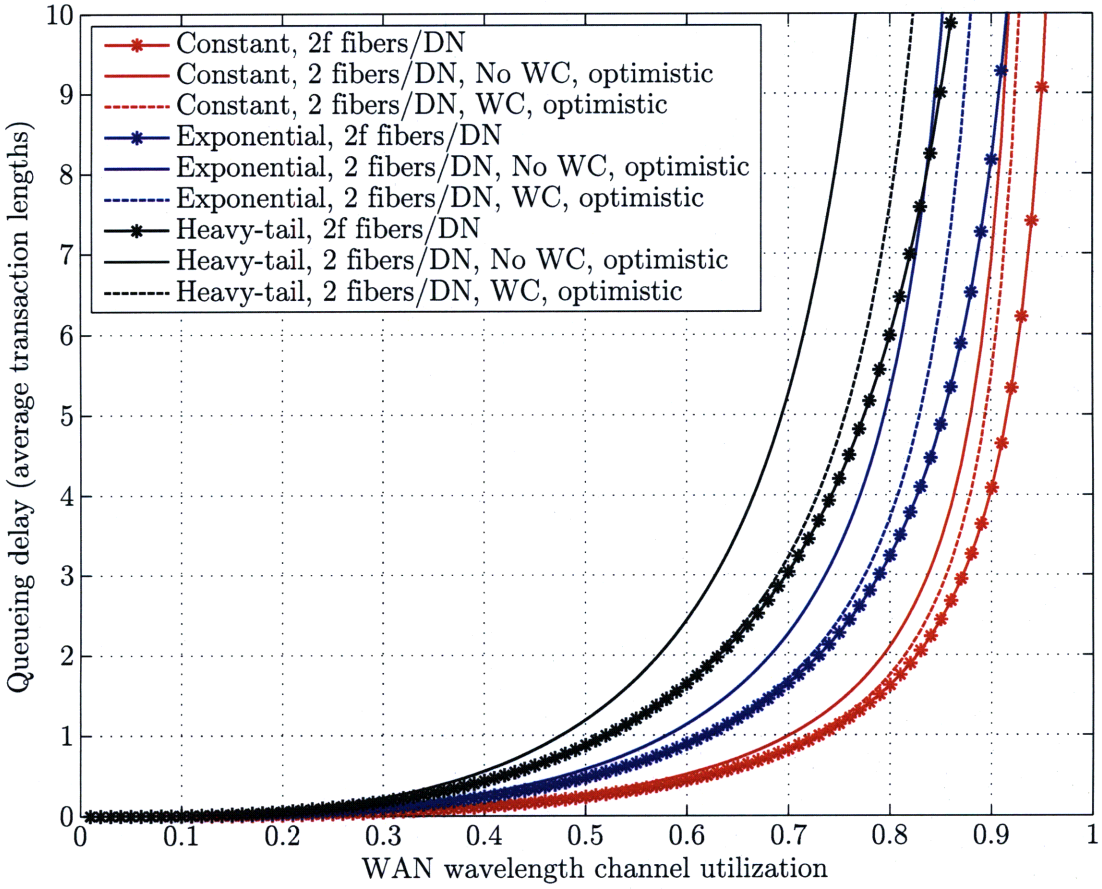


Figure 4-6. Queueing delay versus throughput for DN with two fibers and $2f$ fibers under three flow length distributions. In all cases, flows have an average duration of 10s. In the truncated heavy-tailed case, the distribution tail decays with exponent -1.5, and the lower and upper flow length bounds are 1s and 100s, respectively. It is assumed that $\tilde{n}_a = 200$, $f = 15$, $w_t = 20$, and $w_m = 3$.

related to the frequency with which very large transactions create a contention for resources, thereby diminishing throughput.

Since $\tilde{n}_a$ is a design parameter that is under the control of the network architect, it is interesting to investigate its relationship—under the assumption of two fibers per DN—with another notable design parameter: w_m, the number of wavelength channels provisioned for each MAN pair. The approximations derived for the first two moments of the service time X, with and without wavelength conversion, may be expressed as functions of the ratio:

$$\frac{\lambda_c}{\tilde{n}_a} = \frac{\lambda_m}{w_m \tilde{n}_a},$$

rather than of the individual parameters w_m and $\tilde{n}_a$. In other words, the first two moments of X are determined by the per channel arrival rate of flows to the source DNs. Since λ_m should be assumed to be an imposed parameter beyond the control of the network architect, there exists a simple tradeoff between $\tilde{n}_a$ and w_m for fixed first two moments of X.

The tradeoff between $\tilde{n}_a$ and w_m, for a fixed *total queueing delay* W, however, is not as simple because equation (4.10) exhibits a dependence on w_m individually via the parameters:

$$\lambda_c = \frac{\lambda_m}{w_m} \quad \text{and} \quad \alpha_m \equiv \frac{\hat{W}_{M,w_m}\left(\lambda_m \overline{X}, p_X\right)}{\hat{W}_{M,1}\left(\lambda_c \overline{X}, p_X\right)} = \frac{\hat{W}_{M,w_m}\left(\lambda_m \overline{X}, p_X\right)}{\hat{W}_{M,1}\left(\lambda_m \overline{X}/w_m, p_X\right)}.$$

While this dependence results in a tradeoff between $\tilde{n}_a$ and w_m that cannot be captured in closed form, we shall comment on this tradeoff qualitatively. Figure 4-9 depicts typical tradeoff curves for these two parameters juxtaposed with the above simple inverse proportionality relationship for fixed first two moments of X. As illustrated in this figure, a scaling of the number of DNs $\tilde{n}_a$ by a factor of, say k, results in a scaling down of the number of provisioned wavelength channels w_m by a factor of less than k. Mathematically, this occurs because a decrease in w_m arising from an increase in $\tilde{n}_a$ serves to increase both λ_c and α_m in equation (4.10). On an intuitive level, a decrease in w_m arising from an increase in $\tilde{n}_a$ is dampened, in part, because decreasing w_m increases the load per wavelength channel λ_c, resulting in larger total queueing delay. The exact manner in which λ_c impacts delay is given by the P-K formula. Decreasing w_m, furthermore, increases the ratio α_m because there is less statistical smoothing of flows arising from multiple wavelength channels working in concert to serve flows. The dependence of ratio α_m on $\lambda_c \overline{X}$ is illustrated in Figure 4-10. The ratio is observed to increase monotonically as an "S" curve, exhibiting small rates of increase outside intermediate values of $\lambda_c \overline{X}$. In reference to Figure 4-9, the colored curves most distant from the black inverse proportionality curve are impacted most by changes in statistical smoothing (i.e., the points at which $\tilde{n}_a = 50$ and $w_m = 10$ correspond to operating points near the middle of the "S"

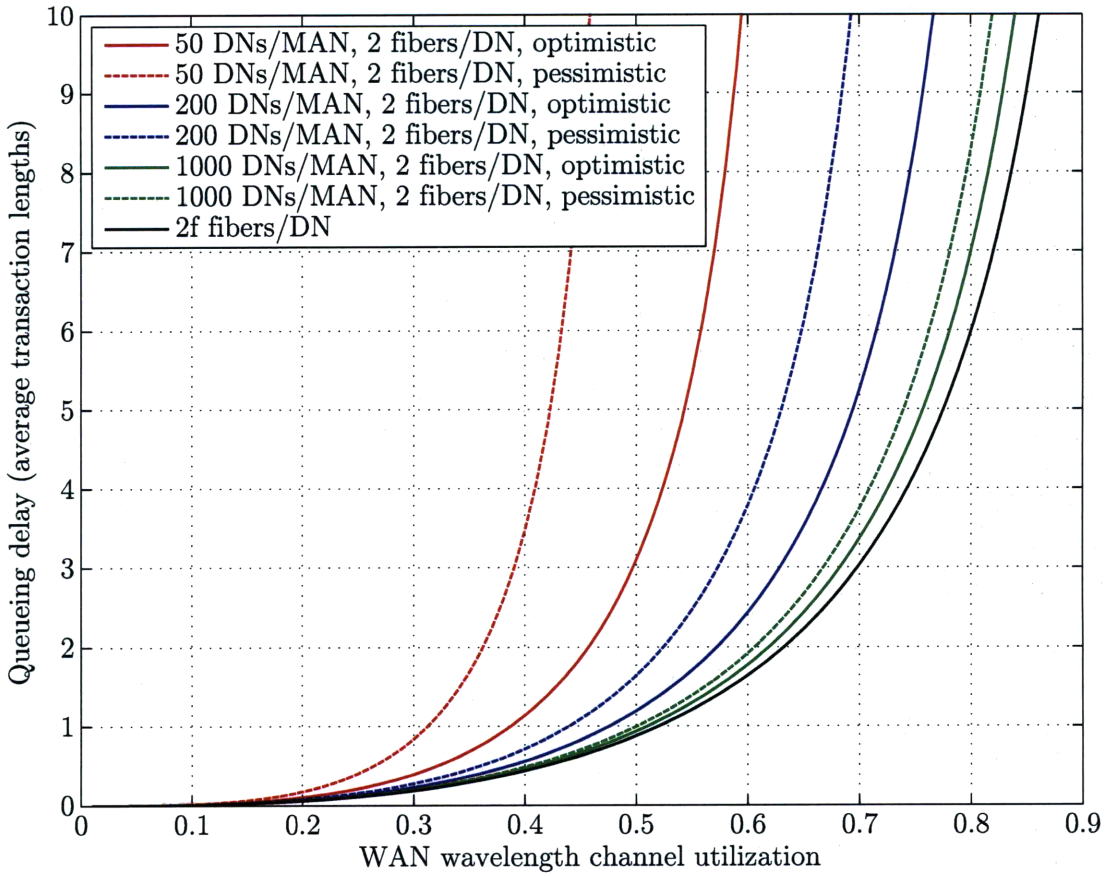


Figure 4-7. Queueing delay versus throughput for a truncated heavy-tailed flow distribution with different numbers of DNs ($\tilde{n}_a$) per MAN with two fibers per DN and no wavelength conversion. The case of $2f$ fibers per DN is also shown as a performance benchmark. In all cases, flows lengths are drawn from a truncated heavy-tailed distribution with: a tail that decays with exponent -1.5, an expected flow duration of 10s, and lower and upper flow length bounds of 1s and 100s, respectively. It is assumed that $w_m = 3$ and $f = 15$.

curve in Figure 4-10). Lastly, we note that, as $\tilde{n}_a$ grows large, each of the colored curves asymptotically approaches the value of w_m required for an analogous network in which each DN is connected to the MAN with $2f$ fibers (i.e., $X = L$) to achieve the same performance.

The previous numerical results and discussion provide us with justification for narrowing the space of design alternatives in the remainder of this thesis. First, equipping each DN with $2f$ fibers for bidirectional communication provides little performance benefit for large MANs in which there are hundreds of DNs per MAN. This observation, coupled with the aforementioned practical concerns regarding network upgrades, render this design alternative less attractive than the case of two fibers per

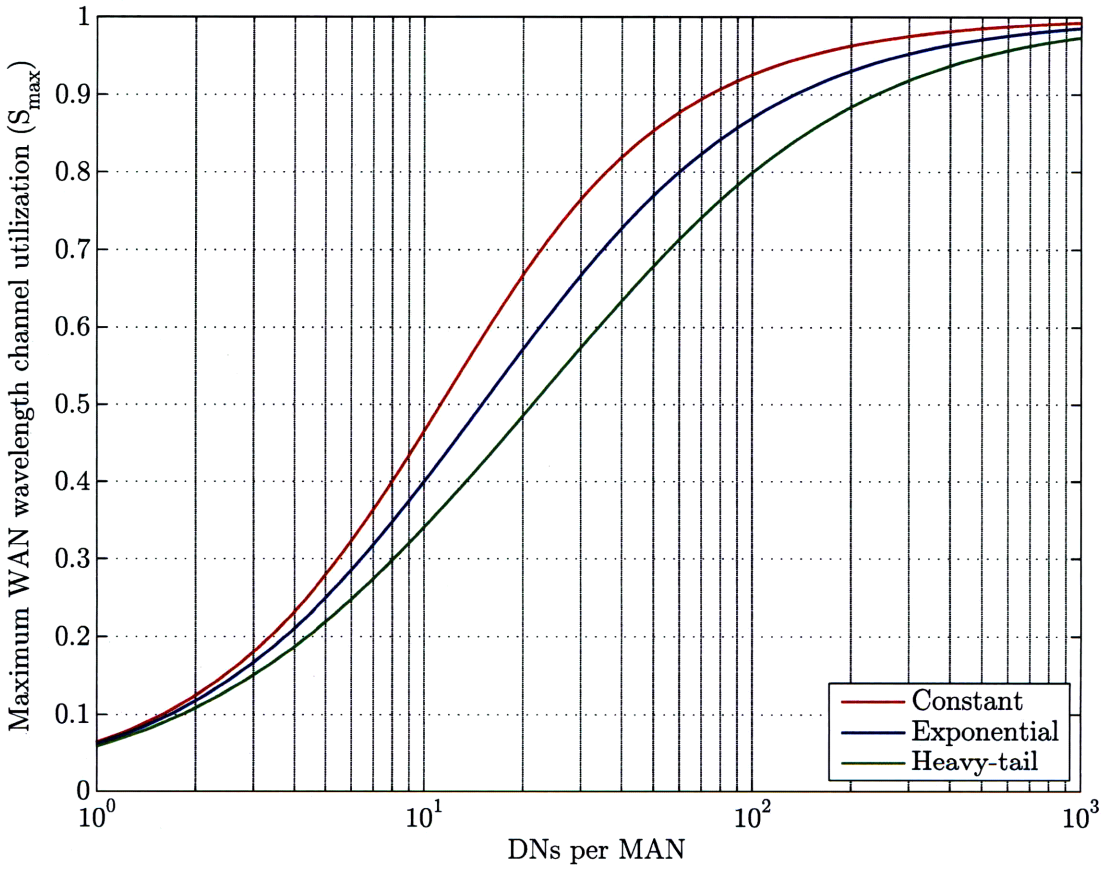


Figure 4-8. Maximum throughput versus number of DNs per MAN for different flow distributions. In the truncated heavy-tailed case, the distribution tail decays with exponent -1.5, and the lower and upper flow length bounds are 1s and 100s, respectively. Two fibers per DN, no wavelength conversion, and $f = 15$ is assumed.

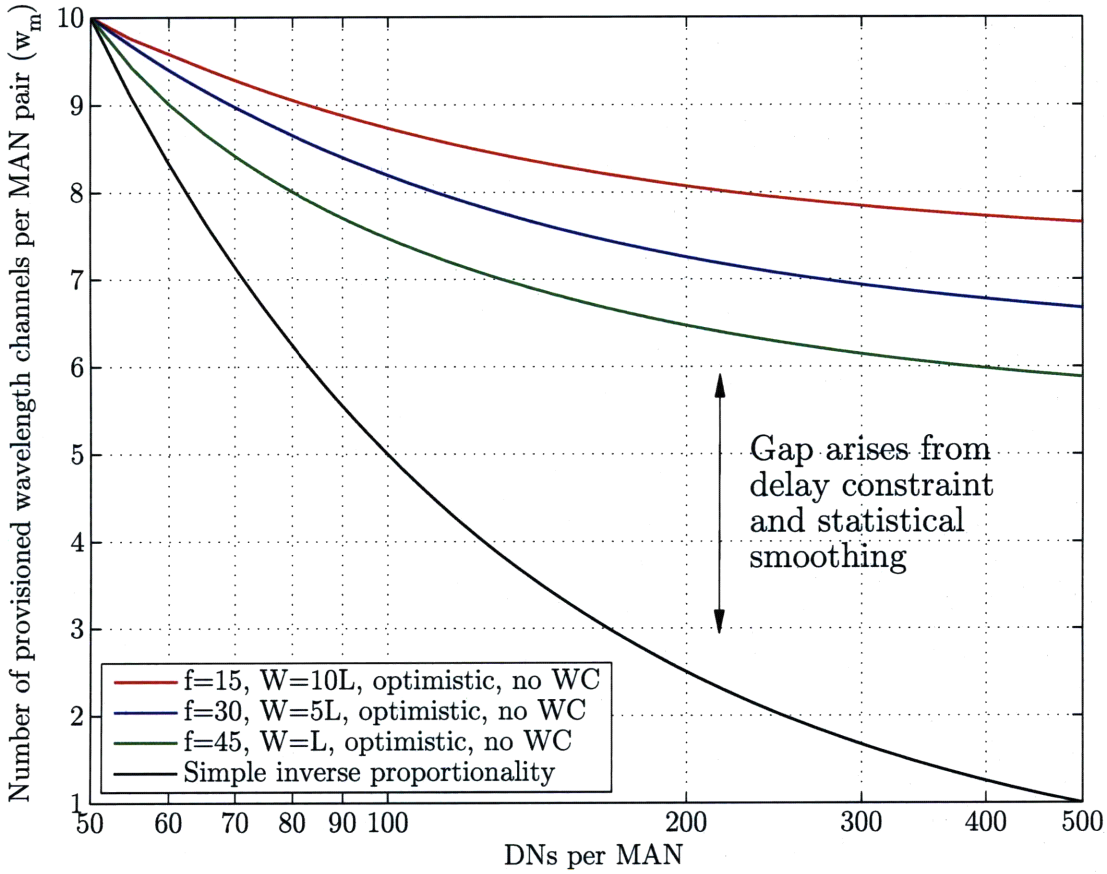


Figure 4-9. Tradeoff between the number of DNs per MAN ($\tilde{n}_a$) and the number of provisioned wavelength channels per MAN pair (w_m). Each curve represents the tradeoff for *different* fixed offered traffic and average delay. An optimistic analysis is employed with no wavelength conversion. In addition, a simple inverse proportionality is drawn in black for comparison.

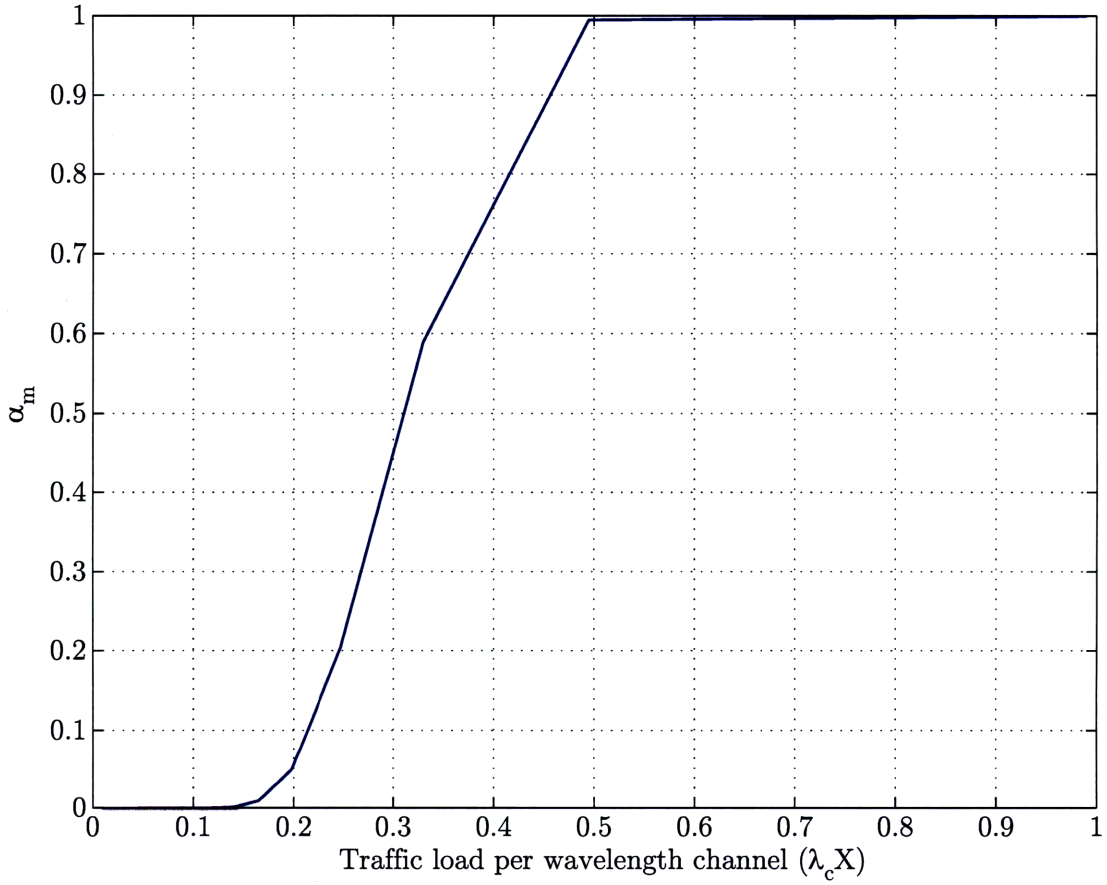


Figure 4-10. Ratio $\alpha_m \equiv \frac{\hat{W}_{M,w_m}(\lambda_m)}{\hat{W}_{M,1}(\lambda_c)}$ versus traffic load per wavelength channel $\left(\lambda_c \overline{X}\right)$. The jaggedness of the curve is a consequence of the integrality requirement on w_m.

DN. We therefore confine our attention in the following chapter to the case where each DN is equipped with two fibers for bidirectional communication.

Wavelength conversion was found to provide a moderate performance, with the benefit decreasing for large MANs. In spite of this performance benefit, we do not consider wavelength conversion in the remainder of this thesis, as the present-day costs of the relevant technologies are prohibitive.

■ 4.5 Conclusion

In this chapter, we began by proposing a simple scheduling algorithm to arbitrate access to network resources for inter-MAN OFS communication. Using this scheduling algorithm, we then conducted an approximate throughput-delay analysis of OFS networks, and explored the tradeoffs in the OFS architecture design space.

The work in this chapter may be extended in various directions. First, generalizing the number of wavelength channels in MAN links to be less than that required for a dedicated channel to exist for every inter-MAN OFS channel may be of interest in order to lower the cost of the MAN. If the loading on these MAN links is still light, only a small performance penalty should be expected. However, the scheduling algorithm would likely need to be more complex, possibly requiring more than the two sequential reservation stages in our scheduling algorithm.

The traffic model employed in this chapter involved a single class of service. In future networks, traffic with different QoS requirements will likely exist. Incorporating a class of best-effort traffic with a long-term throughput requirement, but no delay requirement, would be a straightforward extension in that this traffic would appear "invisible" to higher-priority traffic. A more challenging endeavor is generalizing the work in this chapter to allow for multiple classes of traffic with different delay constraints—a very plausible requirement for future networks.

Whereas this chapter focused on a performance analysis of inter-MAN OFS communication, future work should address the case of intra-MAN communication. As mentioned at the outset of this chapter, the work in Chapter 2 of this thesis is applicable to a metro-setting. In particular, the capacity-achieving scheduling algorithms are suitable for scheduling intra-MAN flows. Furthermore, since capacity is not as precious a resource in the MAN as in the WAN, a frame size equal to a small multiple of the maximum flow length may be employed, thereby trading wavelength channel utilization for lower flow delay. Intra-MAN OFS communication has also been addressed by B. Ganguly in [110], in which a valuable numerical and simulation study was carried out to determine the throughput-delay tradeoff for a particular scheduling algorithm with finite horizon under Poisson traffic. These contributions notwithstanding, there is a need for (approximate) analytical results governing throughput-delay tradeoffs under alternative scheduling algorithms and traffic models. Lastly, this chapter also began with the assumption that there is a quasi-static separation of MAN resources for inter- and intra-MAN communication. While this is a reasonable

assumption that enabled the analysis in the chapter, the algorithms that dynamically allocate resources (on coarse time-scales) for inter- and intra-MAN communication need to be addressed.

Chapter 5

Performance-Cost Comparison of OFS with Other Architectures

We complement the previous three performance-focused chapters by introducing the notion of cost in the present chapter. While the previous chapters were central to our understanding of the functioning and performance of OFS, it is only by viewing the architecture through the lens of network economics that we are able to evaluate its real-world viability. In this chapter, we therefore carry out a simple performance-cost study to substantiate our claim at the outset of this thesis that OFS is an economically attractive candidate architecture for end-users wishing to send large transactions. Our performance-cost study compares OFS to the following prominent optical network architectures: EPS, OCS, GMPLS, and OBS/TaG. Since some architectures are applicable only in the wide-area (e.g., OCS), we must employ them in conjunction with other architectures in the metro-area and access (i.e., EPS) to enable an *end-to-end* comparison with OFS.

Ideally, our study would compare the total network cost imputed by the different architectures in achieving a certain level of performance, as captured, for example, by constraints on throughput and delay metrics. Our cost model, however, is incomplete in that it focuses on CapEx components. Our consideration of performance is also admittedly incomplete in that we address only throughput. The reason we neglect delay is that end-to-end delay analyses, akin to that of Chapter 4 for OFS, have yet to be achieved for the other architectures we consider. We therefore resort in this chapter to a comparative throughput-CapEx cost analysis, wherein a tolerance to unbounded transaction delay is implied. Unbounded delay is, of course, not acceptable in real optical networks. Nevertheless, provided that the costs of the architectures considered in this chapter scale (roughly) commensurately in the presence of finite delay constraints, the insights gleaned and conclusions reached from our simple study should be applicable to the design of real-world optical networks.

This chapter is organized as follows. In the next section, we outline, at a high level, our physical topology and traffic assumptions for all three geographic network tiers. In section 5.2, we discuss our general approach in modeling network cost, as well as some detailed assumptions about the manner in which we model device and system costs. Sections 5.3–5.5 contain our actual cost models for the three geographic

network tiers; and in section 5.6, we combine these network cost models to obtain *total* network cost metrics. In section 5.7, we carry out a throughput-cost comparison of OFS with other prominent optical network architectures. In section 5.8, we consider OFS's role within economically attractive hybrid optical network architecture. We conclude this chapter in section 5.9.

■ 5.1 Topology and traffic assumptions

In our cost study, we shall assume that all of the architectures considered operate on the same WAN fiber plant topology drawn in Figure 5-1. This 60 node network was introduced in [267] as a representative US carrier backbone network. Relevant attributes of this fiber plant topology are listed in Table 5.1. The assumption of a pre-existing fiber plant is a good assumption for countries, such as the US, which have established telecommunication infrastructures. The assumption of a pre-existing fiber plant, however, is less appropriate for countries, such as China, which are presently developing their telecommunication infrastructures; and is entirely inapplicable to many of the countries on the African continent which have virtually no installed fiber. Nevertheless, assuming the same fiber plant under all architectures is reasonable since the layout of the fiber plant is governed, to a large extent, by right-of-way and geographic considerations. The sets of WAN source-destination average traffic demands that we consider are uniformly scaled versions of the set employed in [267, Chapter 8]. This latter traffic set reflects actual US backbone network traffic, and is therefore not uniform all-to-all in nature.

In the metro-area, unlike in the wide-area, we do not assume a fixed fiber plant topology over which all architectures operate. Instead, we analytically optimize the fiber topology in accordance with the switching and fiber deployment costs particular to each architecture. Our rationale for optimizing the MAN topology is that the metro-area is undergoing significant development with MANs requiring significant expenditures to augment existing fiber plant topologies. In particular, in existing MANs, node degrees are low [267], so much fiber needs to be deployed as MANs migrate to more meshed topologies [31]. We restrict our consideration of fiber plant topologies to those that are based upon regular graphs[1] with nodal symmetry[2], since such topologies are reasonable models of real MANs and are more analytically tractable. Similarly to [121], we find that the family of Generalized Moore Graphs minimizes MAN cost, albeit with different dimensions for different architectures. With respect to traffic in the MAN, we assume that intra-MAN traffic is uniform all-to-all, whereas inter-MAN traffic is uniform all-to-one (and one-to-all) to (from) the root node of the

[1] A graph is said to be *regular* of degree Δ if there are Δ outgoing/incoming edges to each of its nodes.

[2] Two nodes u and v in a graph are similar if there is an automorphism which maps u onto v. A graph in which all nodes are similar is *node-symmetric*.

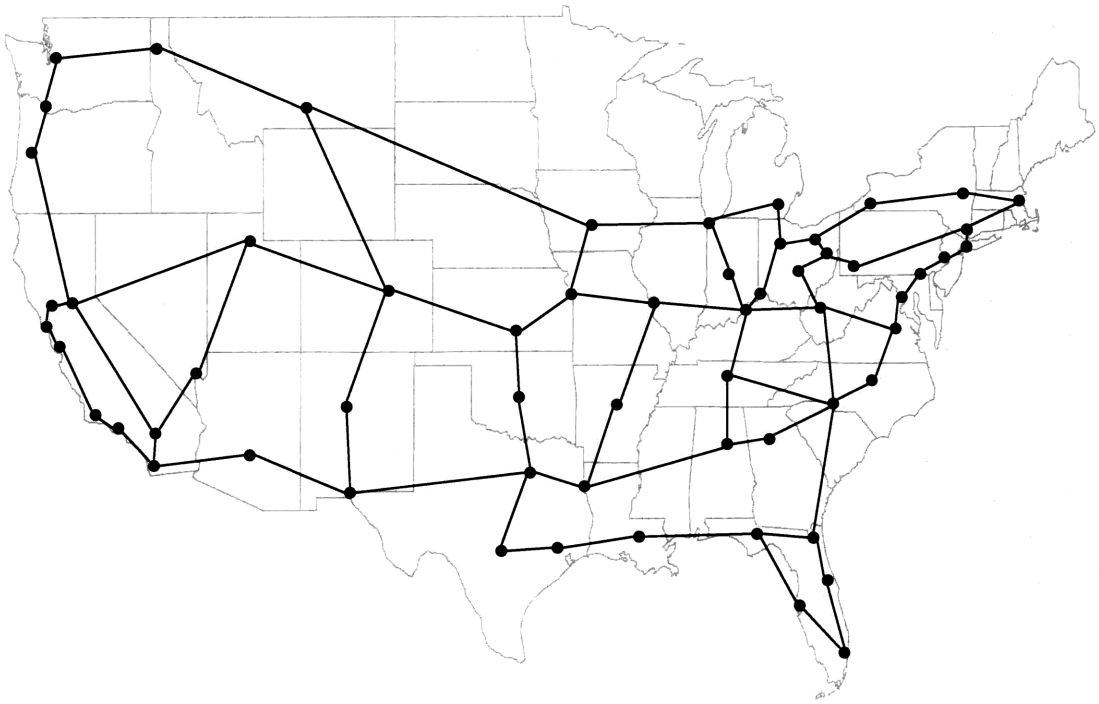


Figure 5-1. WAN topology of the US considered throughout this chapter. Reproduced from [267, Fig. 8.1].

Parameter	Symbol	Value
Number of nodes	n_w	60
Number of links	–	77
Average node degree	–	2.6
Largest node degree	–	5
Average link length	l_w	450 km
Longest link length	–	1200 km
Optical amplifier spacing	l_a	80 km
Number of wavelength channels per fiber link	w	200
Average length of an end-to-end connection	–	1950 km
Average number of hops of an end-to-end connection	h_w	4
Average number of non-nodal regenerations along an end-to-end connection in EPS	$\beta^e_{c_r}$	0.1
Average number of regenerations along an end-to-end connection in OCS or OFS	$\beta^o_{c_r}$	0.3
Capacity efficiency scaling factor	κ_w	0.95
Line-rate	–	40 Gbps

Table 5.1. Important WAN parameters, and their values for the numerical studies in sections 5.7 and 5.8. These numerical values correspond to the reference WAN in Figure 5-1, and the traffic carried on it (adapted from [267, Tbls. 8.1 and 8.2]).

Parameter	Symbol	Value
Number of nodes	n_m	30
Average link length	l_m	10 km
OFS hardware reconfiguration time	τ	10 ms
Line-rate	–	40 Gbps

Table 5.2. Important MAN parameters, and their values for the numerical studies in sections 5.7 and 5.8.

Parameter	Symbol	Value
Average end-user link length	l_u	35 m
End-user duty cycle	ρ_u	0.001
Ratio of PON line-rate to WAN line-rate	κ_l	0.25
OFS DN Line-rate	–	40 Gbps
PON Line-rate	–	10 Gbps

Table 5.3. Important access network parameters, and their values for the numerical studies in sections 5.7 and 5.8.

embedded MAN tree. Typical values for relevant MAN parameters, which we invoke later in this chapter, are listed in Table 5.2.

In the access environment, we employ the DN designs introduced in Chapter 3 as the basis for both our non-OFS PONs and OFS DNs. As discussed in Chapter 3 and further in section 5.5, DNs employing remotely-pumped EDFs within tributaries can be tailored to support end-users economically. With regard to traffic generated or sunk at access networks, we assume uniformity. However, with respect to individual end-user data rate requirements, we consider both homogeneous and heterogeneous requirements.

■ 5.2 Cost modeling approach and assumptions

In this section, we begin by outlining, at a high level, the cost modeling approach employed throughout this chapter. This is followed by detailed assumptions on the manner in which we model device and system costs.

■ 5.2.1 High-level approach

Our cost modeling approach in this chapter is to carefully account for the cost of network resources required to establish communication from a transmitting end-user to a receiving end-user. In the access networks, MANs, and the WAN, network resources are less than 100% utilized owing to a combination of inherent operational constraints

of the architectures and implementation tradeoffs made. Thus, when computing the cost of achieving communication, the cost of sacrificed resources, in addition to the cost of resources actually consumed in carrying data, must be accounted for.

In order to compute this cost of end-to-end communication, it is helpful to imagine that a traffic demand, rather than being a set of long-term average rates of communication, is actually a set of constant instantaneous rates of communication which are served with statically provisioned routes, albeit with reconfigurable devices (e.g., OXCs). To be clear, we can safely make this assumption (owing to our insensitivity to delay) for analytic simplicity, but this is *not* how we envision future optical networks to operate. Future optical networks will, of course, serve traffic that is bursty and delay-sensitive in nature, and these networks will therefore require reconfiguration of lightpaths in all three geographic network tiers. Moreover, per our work in Chapter 2, it is the efficiency with which different architectures carry out this reconfiguration—especially in the WAN because it is most heavily loaded—that determines how much traffic can be carried between a source-destination node pair. While we parameterize this efficiency factor in our cost model, we point out that any meaningful insight into the design of future optical networks should be robust to the nature of offered traffic as well as the underlying fiber plant topology. Said another way, given the uncertainty in the nature of future traffic and in the fiber plant topology, the cost structures of the architectures considered must be significantly different for any definitive conclusions to be reached in this chapter. Indeed, as we shall see later in this chapter, the cost structures of the architectures are sufficiently different that any realistic advantage conferred from efficiency in lightpath topology reconfiguration is immaterial to our general conclusions.

To simplify the mathematics involved, we neglect the integrality constraints on all discrete network parameters (e.g., number of switch ports, amplifiers on a link).

■ 5.2.2 Detailed modeling assumptions

The relative costs employed in this chapter, collected from a variety of sources, are organized in Tables 5.4 and 5.5. In this subsection, we address the underlying assumptions in these tables.

CapEx cost components

Our model focuses on CapEx costs and neglects ongoing OpEx costs, which constitute a significant portion of a network's cost, and, moreover, differ across architectures. Indeed, OpEx to CapEx ratios ranging from 1.3 to 4.0 have been reported in [303], indicating both the importance and high variability of OpEx costs. The significance of OpEx notwithstanding, carriers tend to evaluate network design alternatives based upon CapEx, owing to the difficulty in forecasting (and even tracking) the wide range of OpEx costs—electricity, office space rental, and labor for equipment maintenance and services sales, to name a few. Interestingly, the number of OEO conversions in

Network element	Symbol	Relative cost	
		10 Gbps	40 Gbps
Tunable medium-reach transceiver	t_m	$0.3x$	$0.75x$
Tunable long-reach WDM transceiver	t_l	$0.4x$	x
Tunable WDM transponder	c_t	$40x$	$100x$
Tunable WDM regenerator	c_r	$56x$	$140x$
Optical terminal chassis (per wavelength)	m	$2.5x$	$2.5x$
WAN amplifier and dispersion compensation (per wavelength)	a_w	$2x$	$3.1x$
WAN OXC port with amplification and dispersion compensation	s_w	$8x$	$9.1x$
WAN router port	r_w	$120x$	$300x$
MAN OXC port with amplification	s_m	$5x$	$5.1x$
MAN grooming/router port	r_m	$60x$	$150x$
OFS scheduler	u	$2x$	$2x$
Access fiber deployment (per km·wavelength)	α_f^a	$0.2x$	$0.2x$
MAN fiber deployment (per km·wavelength)	α_f^m	$0.2x$	$0.2x$
WAN fiber deployment (per km·wavelength)	α_f^w	$0.1x$	$0.1x$
Access network EDFA pump power (per 100 mW)	a_a	$2x$	$2x$

Table 5.4. Relative costs of network elements for both 10 Gbps and 40 Gbps line-rates [232, 265, 267, 282]. Optical reach of 2500 km, 200 wavelengths per fiber, and bidirectional element function are assumed. As of 2008, $x \approx \$1000$.

Parameter	Symbol	Value
Cost ratio of 10 Gbps to 40 Gbps amplifier and dispersion compensation equipment	κ_a	0.64
Cost ratio of 10 Gbps to 40 Gbps electronics	κ_e	0.4
Cost ratio of 600 km to 2500 km optical reach equipment	κ_r	0.63

Table 5.5. Cost scaling parameters, and their values for the numerical studies in sections 5.7 and 5.8.

a network, a CapEx component captured in our cost model, has been shown to be a rough indicator for a network's OpEx, since the electronics in these conversions consume a major portion of the costs related to power, required office space, and maintenance [25, 267].

Our cost model addresses CapEx insofar that *major cost components which differ from architecture to architecture are captured.* A potential shortcoming of this cost model—in addition to the omission of OpEx—is the possible neglect of significant sources of cost which are roughly constant across architectures, resulting in an overemphasis of the cost differences among architectures.

Fiber deployment

The cost of deploying fiber depends largely on whether the fiber plant of the network pre-exists. For "green-field" networks, where the fiber plant does not pre-exist, cables, each containing tens of fibers, need to be installed underground. The cost of a fiber link would reflect the material cost of the fiber strand and the aforementioned installation cost—including right-of-way cost—amortized over the number of fiber strands installed. For networks in which the fiber plant pre-exists, however, much of the fiber installation cost is a sunk cost.

In either case, the cost of deploying a link c_f is well modeled as a linear function of its length l:

$$\hat{\alpha}_f l \equiv (\alpha_f w)\, l,$$

where the proportionality constant $\hat{\alpha}_f$ reflects the various factors discussed above. The range of values of $\hat{\alpha}_f$ is quite large, but generally falls within $\$2000 - \$50,000$/km [12, 121], with the exception of highly urbanized regions (e.g., New York City) where $\hat{\alpha}_f$ can be over an order of magnitude larger [202]. In Table 5.4, we specialize $\hat{\alpha}_f$ for the three geographic tiers.

Line-rate

In our cost study, we have chosen to focus on networks with WAN line-rates (i.e., the bit rate of a wavelength channel) of 40 Gbps for two reasons: i) most future systems and architecture proposals assume either 40 Gbps or 100 Gbps line-rates as a means of economically accommodating steeply increasing traffic demands; and, ii) CapEx costs associated with 40 Gbps transmission can be more accurately estimated today than those of 100 Gbps transmission. While we assume that MANs and OFS DNs operate at the same line-rate as the WAN, our PON model (i.e., for non-OFS architectures) employs 10 Gbps equipment as a means of lowering access network cost. For the sake of generality, we shall employ the parameter κ_l to capture the ratio of PON to WAN line-rates.

In Table 5.4, we include relative costs of both 10 Gbps and 40 Gbps equipment. Consistent with empirical trends, it is expected that the cost of transceivers, transponders, and regenerators for 40 Gbps equipment will stabilize to approximately 2.5 times

the cost of the corresponding 10 Gbps equipment [117, 267]. Again, for the sake of generality, we shall employ the parameter κ_e to capture the cost ratio of such slower line-rate equipment to faster line-rate equipment.

Also, for fixed optical reach[3], the costs of amplifiers and dispersion compensation equipment scale with maximum fiber capacity (i.e., number of wavelengths per fiber times the line-rate). Specifically, for every doubling in fiber capacity, the costs of amplifiers and dispersion compensation equipment have empirically increased by approximately 25% [267]. Again, for the sake of generality, we shall employ the parameter κ_a to capture the cost ratio of such slower line-rate equipment to faster line-rate equipment.

Optical reach

All of the WAN architectures considered require amplification, chromatic dispersion and polarization-mode dispersion (PMD) compensation, and regeneration. However, the cost of this equipment is significantly dependent upon the optical reach of the underlying transmission systems. In the EPS architecture, signals undergo OEO conversions at nodes in order to be processed electronically at switches or routers. Regeneration of optical signals therefore comes for "free", and there is little economic incentive to build expensive transmission systems with optical reaches well beyond a typical link length. On the other hand, in WAN architectures that employ optical bypass at nodes, regeneration of signals happens only deliberately at expensive regeneration sites (i.e., no "free" regeneration at nodes), so there is more economic incentive to build transmission systems with optical reaches that are much longer than a typical link length. Note that, because longer reach systems cost significantly more than shorter reach systems, there exists an intermediate reach that is cost-optimal for a network (which is very much dependent on the network's parameters). Based upon anecdotal experience with vendors and carriers, the author of [267] has observed that, all other factors held constant, the costs of amplifiers, dispersion compensation equipment, transponders, and regenerators increase by approximately 25% for every doubling in optical reach.

In our present study, rather than carrying out an onerous optimization of optical reach for each architecture, we invoke the results of the network economics studies of [267] to make reasonable assumptions for the optical reaches of the architectures considered. For architectures employing optical bypass at nodes, we shall assume an optical reach of 2500 km, which was shown to be close to optimal for the network parameters assumed thus far. For the EPS architecture, which entails OEO conversions at each node, we shall assume an optical reach of 600 km, which is equivalent to one third longer than the average link length in Figure 5-1. Given the aforementioned

[3]Optical reach is defined as the maximum distance that an optical signal can travel before it degrades to the point of requiring regeneration. Regeneration of optical signals is discussed in section A.9.1.

scaling of equipment cost with optical reach, relevant EPS equipment costs:

$$\kappa_r \equiv 1 - 1.25^{\log_2(600/2500)} \approx 63\%$$

of the corresponding equipment for 2500 km reach systems for the same line-rate. For the sake of generality, we employ the parameter κ_r to represent this cost ratio.

Linearity of device costs

Lastly, implicit in our cost model are the assumptions that: i) all of the wavelength channels in fibers are utilized by traffic, and ii) the cost of a device with multiple wavelength channel ports scales linearly with the number of such ports. These assumptions allow us to express network element costs as per wavelength channel costs by simply amortizing the total cost of the element over the number of wavelength channels it supports. The first assumption allows us to express network element costs that are most naturally expressed as per fiber costs (e.g., fiber, optical amplifiers) as per wavelength channel costs by simply amortizing the per fiber cost over the maximum number of wavelength channels per fiber. The second assumption allows us to employ per wavelength channel costs for a switch or a router by fixing a reasonable dimension for the device (e.g., 256×256 ports for an OXC), and then assuming that the device cost is a linear function of port count for modest perturbations from this operating point. This is a realistic assumption for OXCs based upon 3D micro-electro-mechanical systems (MEMS) technology, since the number of costly parts scales linearly with the number of ports (see appendix A.8.2). OXCs based upon 3D MEMS technology have, moreover, been shown to the most economically attractive switches for networks with approximately ten or more nodes [121, Chapter 6].

In addition, we shall assume that the cost of a laser pump scales linearly with its output power [267]. For laser pump powers in excess of hundreds of milliwatts, the cost scaling for a single device may be superlinear. However, an equivalent output power with linear cost may be engineered by cascading several smaller sources.

■ 5.3 WAN cost model

The major CapEx components in the WAN are: fiber (i.e., material, trenching, and right-of-way costs), optical amplification, chromatic dispersion compensation, PMD compensation, regeneration, as well as switching, routing, and grooming at nodes. For all architectures, we assume that the fiber plant depicted in Figure 5-1 pre-exists. Therefore, while we account for fiber deployment cost, we do not optimize the WAN topology in our cost model. In accounting for the costs, we consider a "typical" bidirectional end-to-end connection in the WAN of wavelength granularity under each architecture (see Figure 5-2).

Note that, a fair comparison among the architectures must account for their different WAN capacities, as discussed in Chapter 2. In other words, we must be mindful

that, for a given WAN fiber topology, different architectures (loosely speaking) may permit different amounts of offered traffic to be carried[4]. Per our work in Chapter 2, the difference in network capacity among architectures is sensitive to the underlying fiber plant topology. With respect to the WAN topology that we consider in Figure 5-1, it was observed empirically in [267] that architectures employing optical bypass without wavelength conversion (e.g., OCS, OFS) require approximately 5% more capacity build than architectures with wavelength conversion at each WAN node (e.g., EPS) to carry a static set of realistic (i.e., nonuniform) source-destination average traffic values[5]. We capture the capacity efficiency of EPS by scaling the total cost of EPS by the (topology dependent) factor $\kappa_w < 1$. However, as mentioned earlier in this chapter, the cost structures of the architectures considered are sufficiently different that our conclusions in this chapter are relatively insensitive to κ_w.

Beyond fundamental differences in capacity efficiency captured by κ_w, implementation issues can impact the throughput achieved under each architecture. Indeed, in Chapter 4, we proposed a *practical* scheduling algorithm for inter-MAN communication that, even in the absence of delay constraints, results in a throughput penalty relative to an optimal, but infeasible, scheduling algorithm. Similarly, there are implementation issues at routers in EPS and OCS that, in reality, result in a throughput penalty. Algorithms achieving maximum throughput are too computationally intensive to be executed at routers, and therefore less complex algorithms that perform suboptimally are employed (e.g., [69]). While the throughput penalty incurred from this algorithm suboptimality is simple to model parametrically, estimating this penalty numerically is difficult owing to the proprietary nature of this information. Our model does not account for this penalty, so the results of our numerical studies in sections 5.7 and 5.8 are conservative with respect to OFS.

■ 5.3.1 EPS

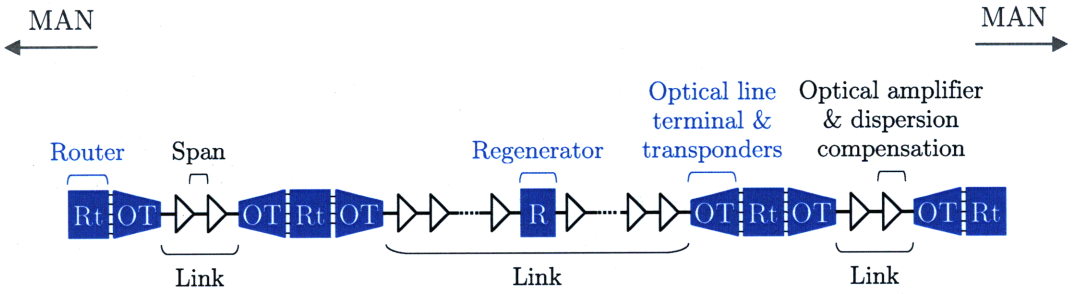
In our EPS cost model, we account for the router, transponder, and OLT costs at nodes. Links joining nodes comprise multiple fiber spans, each of which is terminated with an amplifier and dispersion compensation equipment, or a regenerator, when not terminated at a WAN node (see Figure 5-2(a)). The average cost of a bidirectional end-to-end wavelength-granular WAN connection is therefore given by:

$$C_w^{\text{EPS}} = 2\left\{h_w\left[(r_w + \kappa_r c_t + m) + \left(\frac{l_w}{l_a} - 1\right)\kappa_r a_w + l_w \alpha_f^w\right] + \beta_{c_r}^e \kappa_r (c_r - a_w)\right\}, \tag{5.1}$$

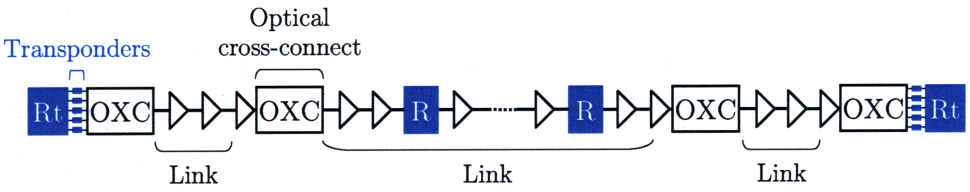
where the first term within the brackets captures the router, transponder, and OLT costs in terminating each of the h_w links along an average WAN wavelength-granular

[4]Recall from Chapter 2 that packet-switched architectures (i.e., EPS, OPS) maximize network capacity, whereas architectures employing optical bypass (i.e., OCS, OFS) do not.

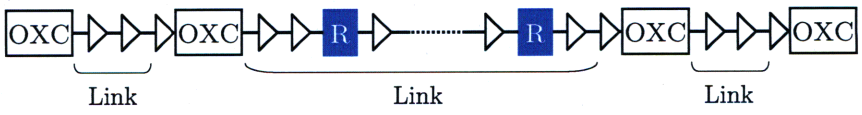
[5]Note that accounting for buffering at WAN nodes, as well as dynamic traffic, would change this disparity in capacity build.



(a) Typical end-to-end WAN connection under EPS. Note that the long length of the middle link requires an additional non-nodal regeneration.



(b) Typical end-to-end WAN connection under OCS and OBS.



(c) Typical end-to-end WAN connection under OFS and TaG.

Figure 5-2. Sample end-to-end WAN connection under EPS, OCS/OBS, and OFS/TaG. Electronic networking devices are drawn in blue; optical networking devices are drawn in black and white. In all three scenarios, the connection comprises three hops.

path; the second term within the brackets captures the amplification and dispersion compensation costs along the h_w links, each of which comprises l_w/l_a fiber spans; the third term within the brackets captures the fiber deployment cost of these links; and the last term in the braces captures the cost of the average number ($\beta^e_{c_r}$) of non-nodal regenerations required. Note that the cost scaling factor of $\kappa_r < 1$ has been applied to account for the fact that EPS transmission system equipment is less expensive than the longer-reach equipment in Table 5.4.

■ 5.3.2 OCS and OBS

In our OCS and model, drawn in Figure 5-2(b), an end-to-end wavelength-granular WAN connection begins and ends at WAN edge routers. These routers are connected to the optical layer via OXCs (which are also responsible for switching transiting traffic). Intermediate nodes along a connection keep data in the optical domain by performing optical bypass with OXCs. As in the case of EPS, links joining nodes comprise fiber spans which are terminated with amplifiers and dispersion compensation equipment, or regenerators. The average cost of a bidirectional end-to-end wavelength-granular WAN connection is therefore given by:

$$C_w^{\text{OCS}} = \frac{2}{\kappa_w}\left\{(r_w + c_t) + h_w\left[s_w + \left(\frac{l_w}{l_a} - 1\right)a_w + l_w\alpha_f^w\right] + \beta^o_{c_r}(c_r - a_w)\right\}, \quad (5.2)$$

where the first term within the braces captures the router and transponder costs at the WAN edge; the second term captures the OXC costs in terminating either end of the h_w links, the amplification and dispersion compensation costs along these links, and the fiber deployment cost of these links; and the last term captures the cost of the average number ($\beta^o_{c_r}$) of regenerations required. Note that the reach of the transmission system used in OCS is different from that of EPS, resulting in different costs for many components (i.e., no κ_r factor). A difference in optical reach, coupled with the fact that optical bypass at intermediate nodes does not furnish "free" regeneration, results in a different number of required regenerations from EPS. Lastly, the capacity scaling factor of $\kappa_w < 1$ has been applied to reflect the inferior capacity efficiency compared to EPS.

Since OBS and OCS differ only in the functioning of their control planes, the cost model for OBS is identical to that of OCS derived above.

■ 5.3.3 OFS and TaG

Our OFS WAN model, drawn in Figure 5-2(c), is very similar to our OCS model with the notable exception that the edge routers are no longer needed at the WAN edge. Recall that, in OFS, all-optical connections are set up between *end-users* rather than between WAN nodes, so all WAN nodes simply provide optical bypass. The average cost of the WAN portion of a wavelength-granular OFS connection is therefore given

by:

$$C_w^{\text{OFS}} = \frac{2}{\kappa_w}\left\{h_w\left[s_w + \left(\frac{l_w}{l_a} - 1\right)a_w + l_w\alpha_f^w\right] + \beta_{c_r}^o\left(c_r - a_w\right)\right\}, \tag{5.3}$$

when there is no regeneration at the WAN-MAN interface; and:

$$C_w^{\text{OFS}} = \frac{2}{\kappa_w}\left\{c_t + h_w\left[s_w + \left(\frac{l_w}{l_a} - 1\right)a_w + l_w\alpha_f^w\right] + \beta_{c_r}^o\left(c_r - a_w\right)\right\}, \tag{5.4}$$

when there is regeneration at the WAN-MAN interface. While similar to the above OCS cost expression, the absence of the router port cost (and transponder cost) can be quite significant. In fact, as we shall we in section 5.7, a factor of two in normalized cost can arise assuming present day cost structures.

As in the case of OCS and OBS, OFS and TaG differ only in the functioning of their control planes. Thus, the cost model for TaG is identical to that of OFS derived above.

■ 5.4 MAN cost model

In contrast to the WAN for which we assume a given fiber plant topology, we undertake an optimization of the MAN topology in this section. In optimizing the MAN topology, we assume that the number of MAN nodes n_m is a fixed parameter constrained by factors beyond the scope of this thesis. We restrict our consideration of fiber plant topologies to those that are based upon regular graphs with nodal symmetry, as in [121]. However, rather than dimension all links and nodal switches identically as in [121], we augment the links and nodal switches in the embedded tree portion of the MAN with more fibers[6] and ports, respectively, to account for the inter-MAN traffic that these network elements additionally support.

In designing our MAN topologies, we shall see that Generalized Moore Graphs minimize network cost over all families of regular graphs—albeit with different dimensions for different architectures—and these graphs therefore form the basis for our MAN topologies. In [121, Theorem 8], Guan shows that this family of graphs minimizes CapEx network cost under uniform all-to-all traffic among MAN nodes. In our present setting, we employ a similar network cost model as in [121], but we consider a more general traffic demand that accounts for inter-MAN communication. In particular, we assume, as in [121], that intra-MAN OFS traffic is uniform all-to-all; but, in addition, there exists uniform all-to-one (and one-to-all) traffic to (from) the root node on the augmented tree portion of the MAN topology representing inter-MAN OFS traffic. Under this more general traffic scenario, Generalized Moore

[6]We assume that the marginal material cost—excluding amplifiers—of augmenting links with more fiber is negligible, as compared with trenching and right-of-way costs. Our fiber deployment cost models in this chapter, therefore, do not explicitly differentiate between augmented and non-augmented links. In other words, all links are assumed to have identical fiber deployment cost regardless of whether the are part of the embedded tree or not.

Graphs still minimize network cost. As we shall see in the following subsection, this is the case because the total MAN cost is a monotonically increasing function of the average shortest path distance of the underlying topology. Since, as discussed in Appendix B, Generalized Moore Graphs minimize this quantity among all graphs with the same number of nodes n_m and node degree Δ, we conclude that this family of graphs minimizes network cost in our present setting. In spite of the optimality of the broad family of Generalized Moore Graphs, we shall focus our attention on the special subset of Moore Graphs when formulating our cost models. This will simplify the mathematics involved, but the insights gained should be applicable to Generalized Moore Graphs.

The major CapEx components in the MAN that we consider are: fiber, optical amplification (at optical nodes), as well as switching, routing, and grooming at nodes. For simplicity, we shall assume an average MAN fiber link length of l_m, resulting in uniform MAN fiber link costs. Node infrastructure (e.g., huts) are considered to be a significant expense, should they need to be built, but we again assume this cost to be roughly the same across the different architectures, and do not include this in our cost model.

■ 5.4.1 Optimality of Generalized Moore Graphs

In this subsection, we substantiate our claim that Generalized Moore Graphs minimize total MAN cost under our present assumptions. Moreover, we analytically solve for the optimal node degree which minimizes total MAN cost, as a function of other network and cost parameters, when the topology is based upon a Moore Graph.

Our assumption of uniform fiber link costs results in the following simple expression for the total fiber cost of a MAN with degree Δ:

$$n_m \Delta \hat{\alpha}_f^m l_m.$$

To compute the total nodal cost, we dimension the switching resources (i.e., router or OXC ports) required at each of the n_m nodes, and then multiply by the per port cost. OXCs are dimensioned differently depending on whether they are internal to the embedded spanning tree portion of the topology or not. Switching equipment internal to the embedded tree support both intra- and inter-MAN pass-through traffic and therefore require more ports than the switching equipment outside of the embedded tree which only supports intra-MAN pass-through traffic.

Similarly to [121, Chapter 5], the total cost of switch ports required to support w_u wavelength channels of uniform all-to-all intra-MAN traffic is:

$$k w_u h_u n_m \left(n_m - 1\right),$$

where h_u is the average shortest path distance, and k is the cost per port.

To compute the total cost of the switch ports required to support inter-MAN traffic, we consider two cases. In the first case, which is applicable to electronic (dis)aggregation (i.e., EPS), we provide just enough resources to support all-to-one (and one-to-all) traffic via the embedded tree. Since each wavelength channel of traffic is terminated by a switch port at either end of each link it occupies, the total number of switch ports required is equal to the aggregate traffic load. The total cost of the switch ports required to support w_a wavelength channels of all-to-one (and one-to-all) traffic is therefore:

$$2kw_a h_a \left(n_m - 1\right),$$

where h_a is the average distance from the root node of the embedded tree to each of the remaining $(n_m - 1)$ MAN nodes. In the second case that we consider, which is applicable to optical (dis)aggregation, we employ the design discussed in section 4.1.1 for OFS. In particular, we dimension all of the switching equipment within the embedded tree identically such that, for each provisioned WAN wavelength channel there exists a dedicated wavelength channel in each link of the embedded tree. Thus, the total cost of the required switch ports is:

$$2kw_m(n_w - 1)(n_m - 1),$$

which, we point out, is independent of the MAN topology aside from the number of MAN nodes n_m. We also point out that this expression corresponds to provisioning w_m wavelength channels, *not* supporting w_m wavelengths worth of traffic. Recall from Chapter 4 that the price paid for the simplicity of our scheduling algorithm is some waste in WAN wavelength channel utilization. This is in contrast to MANs with electronic (dis)aggregation, which we assume to be capable of fully utilizing wavelength channels. As we remarked in section 5.3 in the context of the wide-area, full utilization of wavelength channels cannot be achieved in the metro-area because of the prohibitive computational burden of optimal router algorithms. This idealization of MANs with electronic (dis)aggregation results in a conservative analysis with respect to OFS.

Lastly, we must account for the switch ports required for adding/dropping traffic from/to access networks. Assuming that $\hat{w}$ wavelength channels on each fiber connecting an access network to its parent MAN node are terminated with a switch port, the total cost of required switch ports is:

$$k\tilde{n}_a\hat{w}.$$

Collecting the above cost terms, we have the following expression for the total MAN cost for case 1—electronic (dis)aggregation:

$$\hat{C}_m^{\text{EPS}} = n_m\Delta\hat{\alpha}_f^m l_m + k\left(n_m - 1\right)\left[w_u h_u n_m + 2w_a h_a\right] + k\tilde{n}_a\hat{w}, \tag{5.5}$$

and the following expression for case 2—optical (dis)aggregation:

$$\hat{C}_m^{\text{OFS}} = n_m \Delta \hat{\alpha}_f^m l_m + k(n_m - 1)\left[w_u h_u n_m + 2w_m(n_w - 1)\right] + k\tilde{n}_a \hat{w}. \tag{5.6}$$

These equations indicate that the graph underlying the MAN physical topology influences the total cost of the MAN via the parameters n_m, Δ, h_u and h_a. From Appendix B, we know that Generalized Moore Graphs minimize h_u and h_a for given n_m and Δ. In fact, for Generalized Moore Graphs:

$$h_u = h_a = h_{\mathscr{G}},$$

where $h_{\mathscr{G}}$ is the average shortest path distance for a Generalized Moore Graph and is discussed in Appendix B. Thus, as claimed, Generalized Moore Graphs minimize network cost under our present assumptions, and we therefore restrict our attention to this family of graphs for the remainder of this section.

Optimal node degree for Generalized Moore Graphs

We now solve for the node degree Δ^* that minimizes total MAN cost in equations (5.5) and (5.6). In doing so, we invoke the approximation for Generalized Moore Graphs:

$$h_u = h_a = h_{\mathscr{G}} \approx \log_\Delta n_m$$

in order to simplify the mathematics involved.

The optimality condition for the case of electronic (dis)aggregation in equation (5.5) is then given by:

$$\Delta_{\text{EPS}}^* \left(\ln \Delta_{\text{EPS}}^*\right)^2 = \frac{k \ln n_m \left[w_u(n_m - 1) + 2w_a \frac{n_m - 1}{n_m}\right]}{\hat{\alpha}_f^m l_m}.$$

Solving for Δ_{EPS}^* yields:

$$\Delta_{\text{EPS}}^* \approx \frac{k \ln n_m \left[w_u n_m + 2w_a\right]}{4\hat{\alpha}_f^m l_m} \left[\mathcal{W}\left(\sqrt{\frac{k \ln n_m \left[w_u n_m + 2w_a\right]}{4\hat{\alpha}_f^m l_m}}\right)\right]^{-2} \tag{5.7}$$

provided that:

$$2 \leq \Delta_{\text{EPS}}^* \leq n_m - 1,$$

where, recall, $\mathcal{W}(\cdot)$ is the Lambert Function, defined as the inverse of the function $f(x) = xe^x$.

Similarly, the optimality condition corresponding to the case of optical (dis)aggregation in equation (5.6) is given by:

$$\Delta^*_{\text{OFS}} \left(\ln \Delta^*_{\text{OFS}}\right)^2 = \frac{k w_u (n_m - 1) \ln n_m}{\hat{\alpha}^m_f l_m},$$

which yields:

$$\Delta^*_{\text{OFS}} \approx \frac{k w_u n_m \ln n_m}{4\hat{\alpha}^m_f l_m} \left[\mathcal{W}\left(\sqrt{\frac{k w_u n_w \ln n_m}{4\hat{\alpha}^m_f l_m}}\right)\right]^{-2}, \tag{5.8}$$

provided that

$$2 \leq \Delta^*_{\text{EPS}} \leq n_m - 1.$$

Note that both expressions for Δ^*_{EPS} and Δ^*_{OFS} are of the form:

$$\Delta^*(x) = \frac{x}{\mathcal{W}\left(\sqrt{x}\right)^2},$$

which we plot in Figure 5-3. For both electronic and optical (dis)aggregation, the argument of this function is proportional to the ratio of switching to fiber deployment cost. From Figure 5-3, as this ratio increases, the optimal node degree increases. This trend conforms with our intuition: as switching becomes more expensive relative to fiber deployment, it is economically advantageous to route traffic along fewer hops by increasing the node degree versus routing traffic along more hops (corresponding to lower node degree) which requires more switching resources.

■ 5.4.2 MANs with electronic (dis)aggregation (EPS)

In this subsection, we address the design of a MAN with electronic (dis)aggregation at nodes, as drawn in Figure 5-4(a). Such a MAN architecture is essentially EPS applied in the metro-area, and is compatible with EPS, OCS, or OBS employed in the wide-area. In the metro-area, electronic switching equipment consists of some combination of metro IP routers, SONET/SDH boxes, and Ethernet switches. To simplify our cost model, we shall employ a single, roughly averaged cost corresponding to a "MAN grooming/router port" in Table 5.4 with line-rate 40 Gbps. Owing to OEO conversions at each MAN node, and to fiber loss and nonlinearities being negligible over the short distances traveled in the metro-area, neither optical amplification nor dispersion compensation are required in these MANs.

In contrast to our approach for the WAN in which we considered a wavelength-granular connection, we consider the total cost of constructing a MAN with n_m nodes of degree Δ. To begin, let us adapt equation (5.5) to our present setting by substituting in the following parameters:

$$k = r_m + c_t + m$$

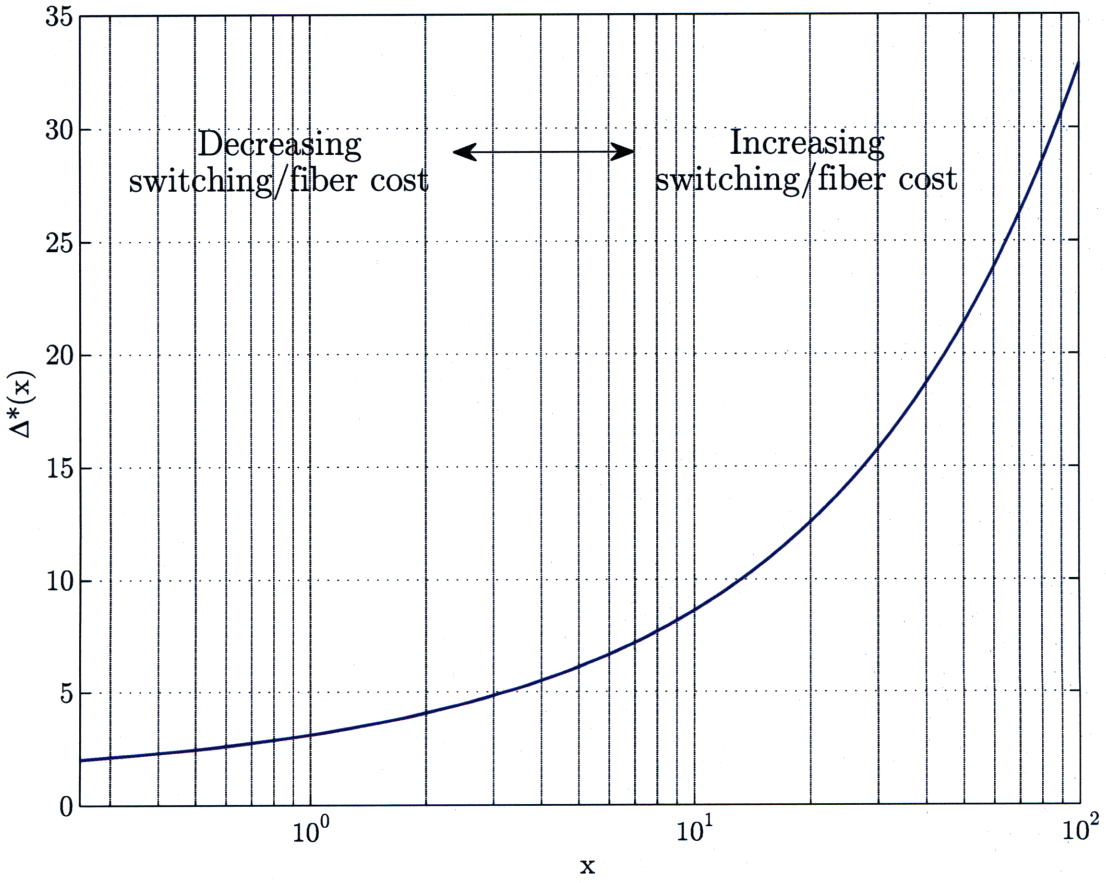


Figure 5-3. Generic optimal node degree for Generalized Moore Graphs: $\Delta^*(x) = \frac{x}{\mathcal{W}(\sqrt{x})^2}$.

$$h_u = h_a = h_{\mathscr{M}}$$
$$w_a = \frac{w_m(n_w - 1)}{n_m}$$
$$\hat{w} = \kappa_l w$$

where $h_{\mathscr{M}}$ is the average shortest path distance of a Moore Graph and is discussed in Appendix B; the second last substitution arises from the fact that $w_m(n_w - 1)$ provisioned WAN wavelength channels of inter-MAN traffic corresponds to uniform all-to-one (or one-to-all) traffic of magnitude $w_m(n_w - 1)/n_m$; and the last substitution arises from our assumption (later in section 5.5.1) that the line-rate in an electronically aggregated access network (e.g., a PON) is slower than that of the MAN. The total MAN cost is then given by:

$$\hat{C}_m^{\text{EPS}} = n_m \Delta \hat{\alpha}_f^m l_m + (r_m + c_t + m)(n_m - 1)\left[w_u h_{\mathscr{M}} n_m + 2h_{\mathscr{M}} \frac{w_m(n_w - 1)}{n_m}\right] + \kappa_l (r_m + c_t + m)\tilde{n}_a w, \quad (5.9)$$

where equation (5.7) may be substituted in for Δ.

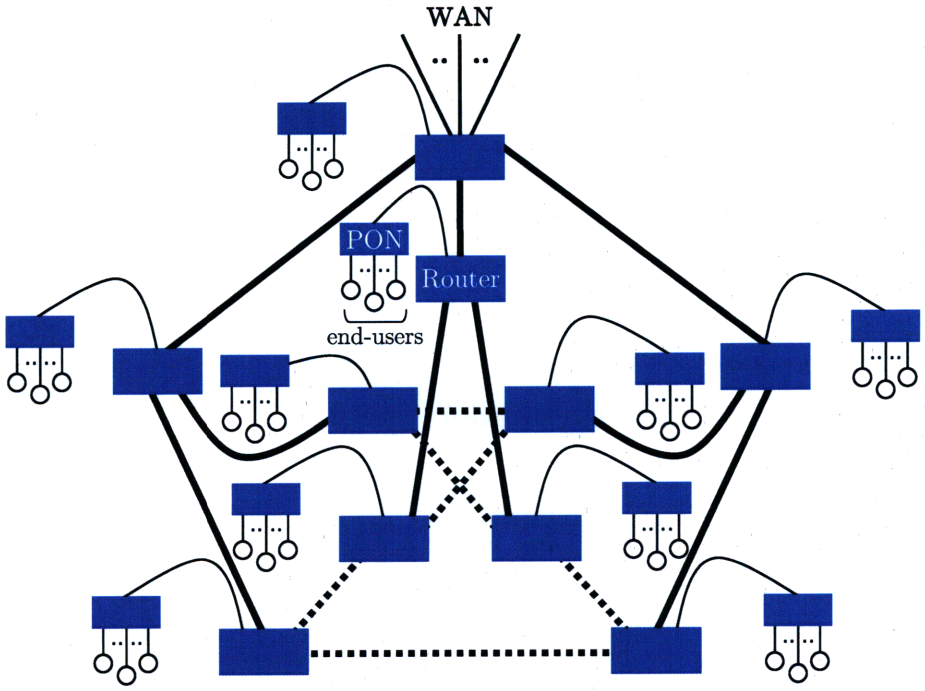
■ 5.4.3 MANs with optical (dis)aggregation

In this subsection, we address the design of a MAN with an all-optical data plane—drawn in Figure 5-4(b))—which is an appropriate MAN model for the OFS and TaG architectures. In an analogous manner to the previous subsection, we begin with equation (5.6) as the basis for our cost model. As in MANs employing electronic aggregation, fiber loss and nonlinearities are assumed to be negligible owing to the short distances traveled in the metro-area. However, MANs that employ all-optical switching generally require optical amplification to compensate for the insertion losses at OXCs, since the signal is not regenerated for "free" at each hop, as in electronic aggregation. Moreover, for OFS, there is the additional cost u of the MAN scheduler.

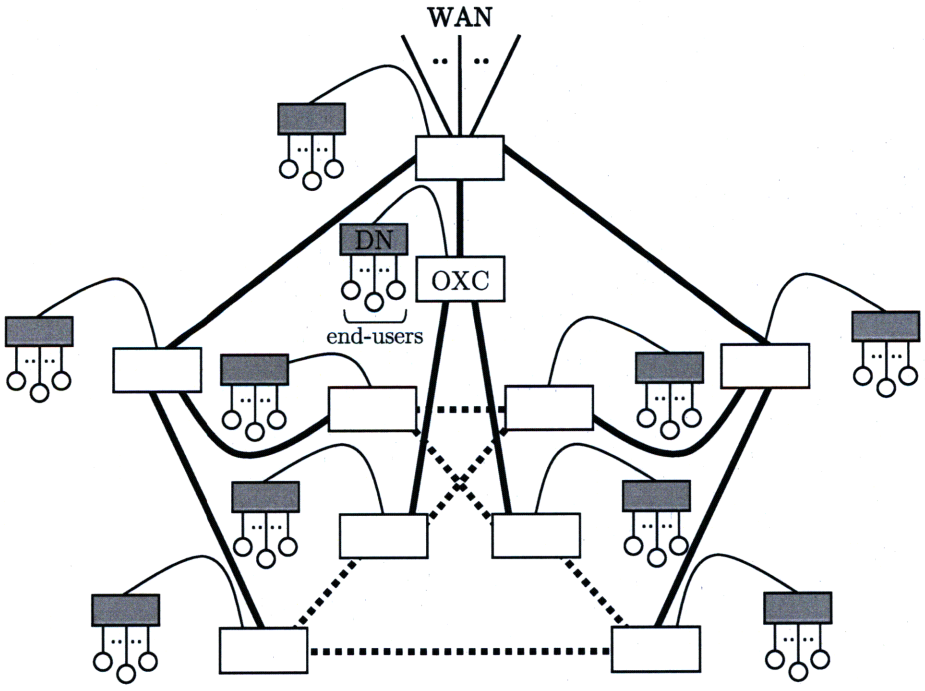
OFS

To specialize equation (5.6) for an OFS MAN based upon a Moore Graph, we substitute in the following parameters:

$$k = s_m$$
$$h_u = h_{\mathscr{M}}$$
$$\hat{w} = w$$



(a) Electronically (dis)aggregated MAN.



(b) Optically (dis)aggregated MAN.

Figure 5-4. MANs based upon a Moore Graph (Petersen graph) with $\Delta = 3$ and $d = 2$. Fiber links *not* in the embedded tree are drawn with dotted lines. Electronic networking devices are drawn in blue; optical networking devices are drawn in black and white.

and add the cost u of the MAN scheduler. The total OFS MAN cost is then given by:

$$\hat{C}_m^{\text{OFS}} = n_m \Delta \hat{\alpha}_f^m l_m + s_m(n_m - 1)\left[w_u h_{\mathscr{M}} n_m + 2w_m(n_w - 1)\right] + s_m \tilde{n}_a w + u, \quad (5.10)$$

where equation (5.8) may be substituted in for Δ.

TaG

The physical design of a MAN under the TaG architecture is identical to that of OFS, except for the absence of a scheduler in TaG (i.e., $u = 0$).

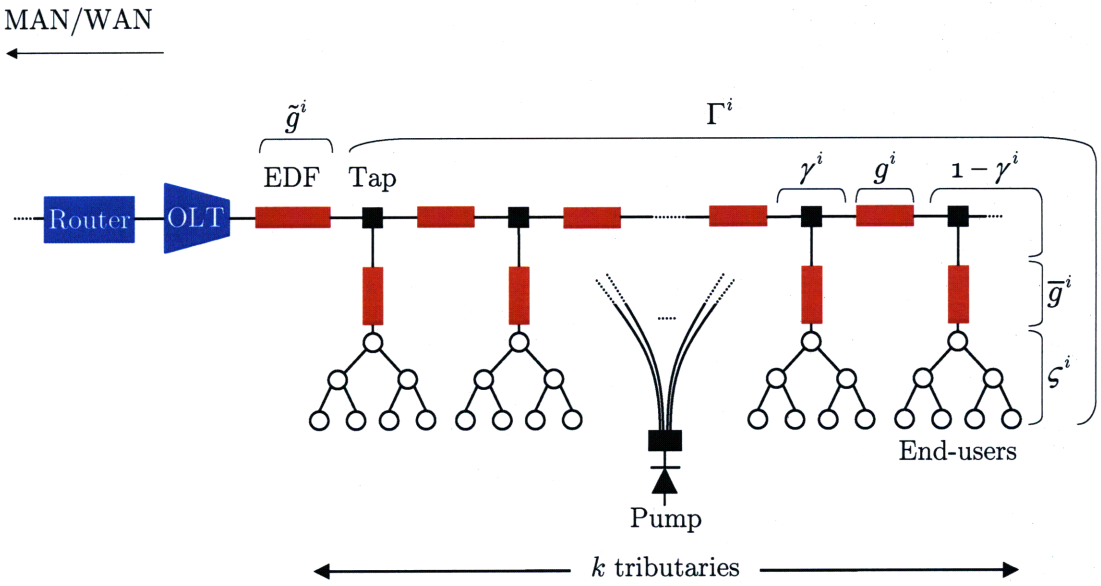
■ 5.5 Access network cost model

In the access environment, the major cost components that we consider are: optical amplifier pumps, OLTs, transceivers[7], and fiber. We expect that the cost of shared passive components, including EDF segments, to be insignificant compared with the above cost components; and allow the costs of passive components required for each end-user (e.g., coupler-based tap) may be lumped into the transceiver cost. In Chapter 3, we presented several candidate DN architectures for OFS. We found remotely-pumped EDFs within tributaries—see Figure 5-5(b)—to be an economical way of supporting many end-users, while adhering to the physical layer constraints posed by reliable OFS communication. In PONs, the physical layer constraints are less stringent owing to signal regeneration capability at head-end OLTs. Nevertheless, the internally amplified DN designs of Chapter 3—augmented with a head-end OLT—are attractive candidate designs for these non-OFS PON settings, and we therefore confine our attention to these designs in this section.

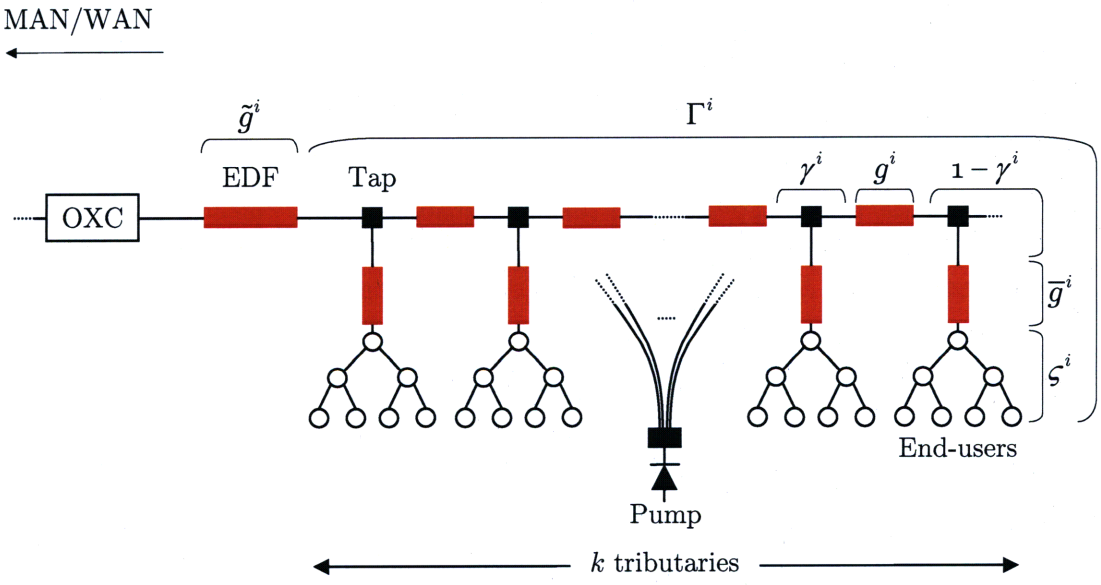
In Chapter 3, we saw that in these internally amplified DN architectures, the required amplifier pump power scales approximately linearly with the number of tributaries k. The amount of pump power required per tributary will depend upon the loss incurred, which, in turn, depends upon the number of supported end-users. Since each tributary of end-users requires its own *parallel* pump power—an expensive resource—it is economically advantageous for each tributary to support the maximum number of end-users. Owing to the aforementioned linearity of pump power cost, the total pump power cost for a MAN is roughly independent of how the collection of all tributaries in the MAN are partitioned into DNs[8]. For example, a MAN with ten DNs each supporting ten tributaries would require approximately the same amount of pump power as a MAN with twenty DNs each supporting five tributaries, provided that the tributaries in both scenarios support the same number of end-users. However, since the fiber pair joining each DN to the MAN needs to be terminated by switch

[7]Transceivers are discussed in sections A.5 and A.6.

[8]This neglects the secondary effects discussed at the beginning of section 3.5 (e.g., fiber loss, excess pump power to overcome absorption).



(a) PON



(b) OFS DN

Figure 5-5. Internally amplified access networks based upon the designs in section 3.5. Up-link and down-link portions are shown in a single diagram.

ports at its parent MAN switch, one would be inclined to design each DN to support the maximum number of users as a means of minimizing the total number of these terminating OXC ports. This maximum number of supportable end-users for a DN would be determined from performance considerations. In the present chapter, where we do not consider delay constraints, the maximum number of end-users that can be supported in a DN is given simply by the data rate capacity of a fiber divided by the average data rate requirement of a user. In the presence of delay constraints—which is beyond the scope of this thesis—this number of supportable users will necessarily be smaller.

The above approach of maximizing the number of end-users per DN indeed results in the most economical design for PONs. We can optimize PON design in this isolated manner because a PON is only loosely coupled with its adjoining MAN as result of OEO conversion and buffering at the head-end OLT. On the other hand, the optimization of OFS DNs is more complicated because of the tight coupling of the access and metro environments in OFS. One of our important observations in Chapter 4 was that, for a fixed aggregate amount of inter-MAN OFS traffic, a larger number of DNs over which to distribute this traffic results in better performance. In particular, a larger number of DNs per MAN results in less contention for resources in each DN, thereby resulting in higher WAN wavelength channel utilization, as illustrated in Figures 4-7 and 4-8. This ultimately requires fewer provisioned WAN wavelength channels than an otherwise identical scenario with fewer DNs. We therefore conclude that in OFS, it can be economically advantageous to employ more than the minimum number of DNs per MAN, each supporting less than the maximum number of users. We shall address this idea analytically in section 5.5.2.

The challenges in accurately estimating fiber cost are particularly acute in the access environment owing to the importance of how residential and business communities are geographically laid out. We shall nevertheless attempt a rough approximation of this cost. To do so, we shall estimate the geographic density of broadband connections in the US, accounting for both anticipated growth in broadband access penetration, and the fact that the vast majority of connections occur in urban environments. A recent study indicated that, as of the end of the first quarter of 2008, there existed approximately 64 million broadband connections in the US [184]. Since the geographic area of the US is approximately 10 million squared kilometers, this translates to an average broadband connection density of 6.4 per squared kilometer. In metropolitan areas—which is were the vast majority of broadband connections are located—the density of such connections can be expected to be about 100 times greater[9] [315]. Additionally accounting for anticipated growth in broadband access penetration, we shall assume 1000 broadband connections per squared kilometer. Moreover, assuming that these broadband connection points are uniformly distributed spatially within urban regions, the distance between neighboring connections is approximately $l_u \approx 35$ m. To compute the fiber cost that this entails, we assume that

[9]We assume that broadband connection density scales commensurately with population density.

end-users are connected to the DN by a single (bidirectional) fiber run to a nearest neighboring connection (i.e., in a tree- or bus-like fashion). Thus, associated with each end-user is a fiber cost of $\hat{\alpha}_f^a l_u$.

With the above discussion as our guide, we shall now formulate cost models for non-OFS PONs as well as OFS DNs.

■ 5.5.1 PONs

As mentioned earlier, and as drawn in Figure 5-5(a), we shall assume a PON architecture based upon the internally amplified DN design in Chapter 3 augmented with an OLT at the head-end.

Since the OLT at the head-end of the PON converts signals into a form appropriate for the MAN, the technology employed in the PON need not be compatible with the MAN. Noting that access networks carry far less traffic than MANs, we allow lower line-rate equipment in the PON as a means of lowering cost. In particular, we assume that end-users are equipped with medium-range transceivers with peak rates of 10 Gbps; and we model an OLT as a "MAN grooming/router port" with the same slower line-rate.

Let us now compute the maximum number of supportable users per PON. First, let us assume that fibers within the PON employ the same number of wavelength channels per fiber w as the MAN and WAN, but with the lower line-rate of 10 Gbps. This under-utilization of fiber capacity in PONs is a sacrifice that is made for lower cost equipment in the access environment[10]. Letting ρ_u denote each end-user's duty cycle at the WAN line-rate—that is, the fraction of time that the end-user would be transmitting data at the WAN line-rate—the maximum number of supportable users in the PON is given by:

$$n_a = \frac{\kappa_l w}{\rho_u},$$

where, recall, a tolerance to infinite delay is implicit.

To determine the maximum number of supportable users per PON tributary n_t, we invoke equation (3.18), setting $F_{n,w} = 1$ and accounting for the slower line-rate with the factor κ_l:

$$\zeta_s^i \approx \frac{2\kappa_l n_{\text{sp}} \mathcal{Q}^2 \overline{h}\nu\Delta f}{\eta P_{\max}}.$$

This expression for ζ_s^i may then be substituted into equation (3.19) to obtain the number of supportable users for a star tributary; or, along with equation (3.21), into equation (3.20) to obtain the number of supportable users for a bus tributary. This number of supportable users per tributary may then be used to compute the destination tributary loss parameter ζ_d^i.

[10]More than w wavelength channels at this slower line-rate entails higher quality transceivers with narrower linewidths.

As discussed earlier, optimal PON design requires the maximum number of tributaries to be supported. This maximum number of tributaries k that can be supported per PON is then given by:

$$k = \frac{n_a}{n_t} = \frac{\kappa_l w}{\rho_u n_t}.$$

To compute the pump power required for a PON with k tributaries and loss parameters ζ_s^i and ζ_d^i, we carry out a calculation similar to that of Example 3.1 in section 3.6. Recall from equation (3.16) that for a PON with k tributaries, the optimal bus coupling ratio[11] is:

$$\gamma_d^i = \sqrt{1 - \frac{1}{k}}.$$

Assuming the same coupling ratio for γ_s^i, equations (3.12), (3.14), and (3.15) yield the following required EDFA gains for the PON:

$$\begin{aligned}
\overline{g}_s^i &= \frac{1}{\zeta_s^i (1 - \gamma_s^i)} = \left[\zeta_s^i \left(1 - \sqrt{1 - \frac{1}{k}}\right)\right]^{-1} \\
\tilde{g}_d^i &= \frac{1}{1 - \gamma_d^i} = \left(1 - \sqrt{1 - \frac{1}{k}}\right)^{-1} \\
\overline{g}_d^i &= \frac{1}{\zeta_d^i}.
\end{aligned}$$

The implied normalized pump power for each of these EDFA gains may be computed with equation (3.22). Denoting these implied normalized pump powers by $\overline{p}_s^i$, $\tilde{p}_d^i$, and $\overline{p}_d^i$, the total required normalized pump power for the PON is given approximately by:

$$\overline{P}_{\text{PON}} \approx \tilde{p}_d^i + k\left(\overline{p}_s^i + \overline{p}_d^i\right).$$

Using the above, the total cost of a PON may be expressed as:

$$\hat{C}_a^{\text{PON}} = a_a \overline{P}_{\text{PON}} + \kappa_e w(r_m + c_t + m) + n_a\left(\kappa_e t_m + \hat{\alpha}_f^a l_u\right), \tag{5.11}$$

where the first term captures the cost of the amplifier pump power; the second term captures the cost of the OLT at the head-end; and the third term captures the transceiver and fiber cost associated with each end-user.

■ 5.5.2 OFS DNs

Our treatment of OFS DNs in this subsection is very similar to that of PONs. However, we shall defter the step of determining the number of DNs per MAN to the next

[11] As discussed in Chapter 3, we assume constant tap values on the bus. Customized tap values would yield better performance (i.e., less required pump power, or larger number of supportable end-users), but custom taps are arguably impractical to implement.

subsection, as it entails an optimization encompassing all three geographic network tiers.

As drawn in Figure 5-5, the critical difference between an OFS DN and a PON is the absence of an OLT in the OFS DN. The implications of this are tighter physical layer constraints, and the requirement that end-user transceivers operate at the WAN line-rate. The physical layer constraint from equation (3.18) remains:

$$\zeta_s^i \approx 2n_{\text{sp}} \left(\frac{\eta P_{\max}}{\mathscr{Q}^2 \overline{h} \nu \Delta f} - F_{n,w} \right)^{-1},$$

which may be substituted into equation (3.19) to obtain the number of supportable users n_t for a star tributary; or, along with equation (3.21), into equation (3.20) to obtain the number of supportable users for a bus tributary. After determining the optimal number of DNs per MAN $\tilde{n}_a$ in section 5.6.3, the number of end-users per DN follows simply as:

$$n_a = \frac{n_u}{\tilde{n}_a},$$

and the number of tributaries k per DN is:

$$k = \frac{n_a}{n_t},$$

where n_u is the total number of end-users per MAN. The optimal bus coupling ratio is computed in the same manner as in the PON case, yielding similar expressions for the required EDFA gains and normalized pump powers.

The total cost of an OFS DN may then be expressed as:

$$\hat{C}_a^{\text{OFS}} = a_a \overline{P}_{\text{OFS}} + n_a \left(t_l + \hat{\alpha}_f^a l_u \right), \tag{5.12}$$

where the first term captures the cost of the amplifier pump power $\overline{P}_{\text{OFS}}$, and the second term captures the transceiver and fiber cost associated with each end-user.

■ 5.6 Total network cost

In this section, we assemble our previous cost models for individual geographic network tiers into a model for total network cost to be used in the performance-cost study that follows. Our approach is to focus on a particular MAN and sum all of the costs in supporting intra- and inter-MAN communication associated with it. For intra-MAN communication, we assume w_u wavelengths of uniform all-to-all traffic; that is, associated with each MAN node pair is w_u wavelengths of traffic in each direction. For inter-MAN communication, we assume that associated with each MAN

pair are w_m wavelengths of traffic in each direction[12]. These w_m wavelengths of traffic are assumed to be generated/sunk uniformly among the n_m nodes in each of the source/destination MANs.

We consider the following three end-to-end architectures to serve the above traffic demand:

- "EPS". In this architecture, EPS is used in the wide-area (section 5.3.1), electronic (dis)aggregation is used in metro-area (section 5.4.2), and PONs are used in the access (section 5.5.1).
- "OCS/EPS". In this architecture, OCS is used in the wide-area (section 5.3.2), electronic (dis)aggregation is used in metro-area (section 5.4.2), and PONs are used in the access (section 5.5.1).
- "OFS". The OFS architecture model is captured in sections 5.3.3, 5.4.3, and 5.5.2. We employ the scheduling algorithm proposed in Chapter 4, and we assume no regeneration at the MAN-WAN interface.

These three architectures represent an evolution from electronic to optical switching, from the network core towards the end-users at the network edge. To be sure, these three architectures are not exhaustive of the space of architecture alternatives. For instance, in the previous three sections we also described cost models for OBS and TaG, which we do not consider any further in this chapter. The purpose of formulating CapEx cost models for these architectures was to highlight the fact that they are identical to those of OCS and OFS, respectively. The critical differences between OBS and TaG on the one hand, and OCS and OFS on the other, lie in how network resources are allocated. The former architectures employ random-access approaches for resource reservation, which are known to result in inferior throughput performance relative to the latter scheduled approaches. (Indeed, this was discussed at length in Chapter 2.) As a result, OBS and TaG are guaranteed to appear inferior to OCS and OFS in a throughput-CapEx study, and we therefore do not consider these architectures any further.

■ 5.6.1 EPS

Under the EPS architecture, the cost of supporting w_u wavelengths of uniform all-to-all intra-MAN traffic, and w_m wavelengths of traffic with every other MAN is given by:

$$\hat{C}^{\text{EPS}} = \frac{1}{2} w_m (n_w - 1) C_w^{\text{EPS}} + \hat{C}_m^{\text{EPS}} + \tilde{n}_a \hat{C}_a^{\text{PON}}, \tag{5.13}$$

where equations (5.1), (5.9), and (5.11) should be substituted in. Note that a factor of 1/2 has been applied to the WAN portion because the cost of the bidirectional

[12]The assumption of uniform all-to-all traffic among WAN nodes is made for simplicity. Per our discussion in section 5.1, traffic among WAN nodes is known to be nonuniform.

WAN wavelength channels should be shared equally by the two MANs for which they are provisioned.

As discussed in Chapter 2, the EPS architecture is maximizes the WAN capacity region. Denoting the maximum EPS WAN wavelength channel utilization by U^{EPS}, we therefore have:

$$U^{\text{EPS}} = 1.$$

Note that the optimal utilization of EPS is implicit in equation (5.13), in that exactly w_m wavelength channels are necessary to support w_m wavelength channels of traffic. We reiterate that this involves the idealization that electronic switches and routers are capable of executing algorithms resulting in maximum throughput, which is not the case in reality.

■ 5.6.2 OCS/EPS

Under the OCS/EPS architecture, the cost of supporting the same intra- and inter-MAN traffic is given by:

$$\hat{C}^{\text{OCS}} = \frac{1}{2} w_m (n_w - 1) C_w^{\text{OCS}} + \hat{C}_m^{\text{EPS}} + \tilde{n}_a \hat{C}_a^{\text{PON}}, \tag{5.14}$$

where equations (5.2), (5.9), and (5.11) should be substituted in.

Note that, under the OCS architecture, once a wavelength channel is provisioned, full utilization of its informational capacity can be achieved. However, as discussed in section 5.2.1—and in more depth in Chapter 2—the OCS architecture is less efficient than the EPS architecture in reconfiguring lightpaths in the WAN. Recall that this inefficiency, which we parameterized by κ_w, arises from the absence of buffering and wavelength conversion in the OCS WAN. Consequently, the OCS WAN capacity must be overbuilt by a factor of κ_w^{-1} relative to the EPS WAN capacity to provide the same throughput performance. The maximum utilization of an OCS WAN wavelength channel is therefore:

$$U^{\text{OCS}} = \kappa_w.$$

Again, we point out the implied idealization here that electronic switches and routers are capable of executing optimal algorithms.

■ 5.6.3 OFS

Under the OFS architecture, the cost of supporting w_u wavelengths of uniform all-to-all intra-MAN traffic, and *provisioning* w_m wavelength channels for communication with every other MAN is given by:

$$\frac{1}{2} w_m (n_w - 1) C_w^{\text{OFS}} + \hat{C}_m^{\text{OFS}} + \tilde{n}_a \hat{C}_a^{\text{OFS}},$$

where equations (5.3) [or (5.4)], (5.10), and (5.12) should be substituted in. Recall that, even after being provisioned, each WAN wavelength channel is unable to be fully utilized because of the inefficiency of our simple scheduling algorithm. Thus, if we wish to support w_m wavelengths of traffic between each MAN pair, we must scale up the amount of provisioned inter-MAN network resources by the reciprocal of the channel throughput, as given by equation (4.12). Moreover, let us reintroduce the idea of capacity waste due to MAN hardware reconfiguration between transactions. We neglected this hardware reconfiguration time τ in equation (4.1) because we assumed in Chapter 4 that the average transaction length $\overline{L}$ was much larger than τ. However, in the cost study to follow, we also consider using OFS for small transactions for which τ may no longer be negligible compared with $\overline{L}$. The fraction of bandwidth wasted due to hardware reconfiguration is given simply by:

$$\frac{\tau}{\overline{L}+\tau} \equiv 1 - \kappa_h.$$

Accounting for all of the aforementioned capacity inefficiencies, and substituting in equation (5.10) for the cost of the MAN portion, the total cost under OFS is given by:

$$\hat{C}^{\text{OFS}} = \kappa_h^{-1} \left\{ S_{\max}^{-1} \left[\frac{1}{2} w_m (n_w - 1) C_w^{\text{OFS}} + 2 w_m s_m (n_m - 1)(n_w - 1) \right] \right.$$
$$\left. + \left[n_m \Delta \hat{\alpha}_f^m l_m + s_m (n_m - 1) w_u h_{\mathcal{M}} n_m + s_m \tilde{n}_a w + u \right] + \tilde{n}_a \hat{C}_a^{\text{OFS}} \right\}, \quad (5.15)$$

where the first bracketed term captures the scaled up inter-MAN network resources; the second bracketed term captures the remaining MAN resources, which are used solely for intra-MAN communication; and the last term captures the cost of the $\tilde{n}_a$ DNs.

Recall that our OFS DN cost model in section 5.5.2 did not involve an optimization of the number of DNs per MAN $\tilde{n}_a$. Let us now carry out the minimization of $\hat{C}^{\text{OFS}}$ over all values of $\tilde{n}_a$. To simplify the mathematics involved, we invoke an approximation for $S_{\max}$. As illustrated in Figure 4-8, $S_{\max}$ is approximately linear with respect to $\log(\tilde{n}_a)$ for the intermediate range of $\tilde{n}_a$—which happens to also be our range of practical interest. In Figure 5-6, we re-plot this figure with the normalized quantity $\log(\tilde{n}_a/f)$ on the abscissa. A linear approximation to one of the curves is also included, highlighting the accuracy of the approximation for the range of practical interest. For a given flow length distribution, we can therefore make the following approximation:

$$S_{\max} \approx \mu_1 \ln\left(\frac{\tilde{n}_a}{f}\right) + \mu_2,$$

where the constants μ_1 and μ_2 are specific to the flow length distribution. After invoking this approximation for $S_{\max}$, the optimality condition for minimizing $\hat{C}^{\text{OFS}}$

is given by:

$$\tilde{n}_a \left[\mu_1 \ln\left(\frac{\tilde{n}_a}{f}\right) + \mu_2\right]^2 = \frac{\mu_1 \left[\frac{1}{2} w_m (n_w - 1) C_w^{\text{OFS}} + 2 w_m s_m (n_m - 1)(n_w - 1)\right]}{s_m w}.$$

Defining ξ as the right-hand side of this equation, it can be shown that the optimal number of DNs is given by:

$$\tilde{n}_a^* \approx f \exp\left\{2\mathcal{W}\left[\frac{\exp\left(\frac{\mu_2}{2\mu_1}\right)}{2\mu_1\sqrt{\frac{f}{\xi}}}\right] - \frac{\mu_2}{\mu_1}, \right\},$$

which can be substituted into equation (5.15) to obtain the minimum cost of an OFS network.

With respect to WAN wavelength channel utilization, our discussion of OCS/EPS in the previous subsection applies to OFS. Namely, OFS suffers the same WAN reconfiguration inefficiency as OCS in relation to EPS. Moreover, the maximum throughput of an OFS WAN wavelength channel is diminished by factors κ_h and $S_{\max}$ owing to hardware reconfiguration overhead and contention for resources in DNs, respectively. Consequently, we have:

$$U^{\text{OFS}} = \kappa_w \kappa_h S_{\max}.$$

We note that, under the idealization that electronic switches and routers are capable of executing optimal algorithms, OFS has a lower maximum WAN wavelength channel utilization than both EPS and OCS/EPS. OFS must therefore possess a more attractive cost structure than these two architectures to be a viable alternative. We show that this is indeed the case in the next two sections.

■ 5.7 Throughput-cost comparison of architectures

In this section, we carry out a throughput-cost comparison of the three end-to-end network architectures discussed in the previous section. By numerically varying the critical network parameters in the previous section's models for total network cost, we shall deepen our understanding of each architecture's cost drivers as well as how the architectures compare.

■ 5.7.1 Modeling assumptions

In this subsection, we discuss our modeling assumptions beyond those in the previous four sections. We remind the reader that the network and cost parameters used, which reflect the state of present-day networks, are summarized in Tables 5.1, 5.2, 5.3, 5.4, and 5.5.

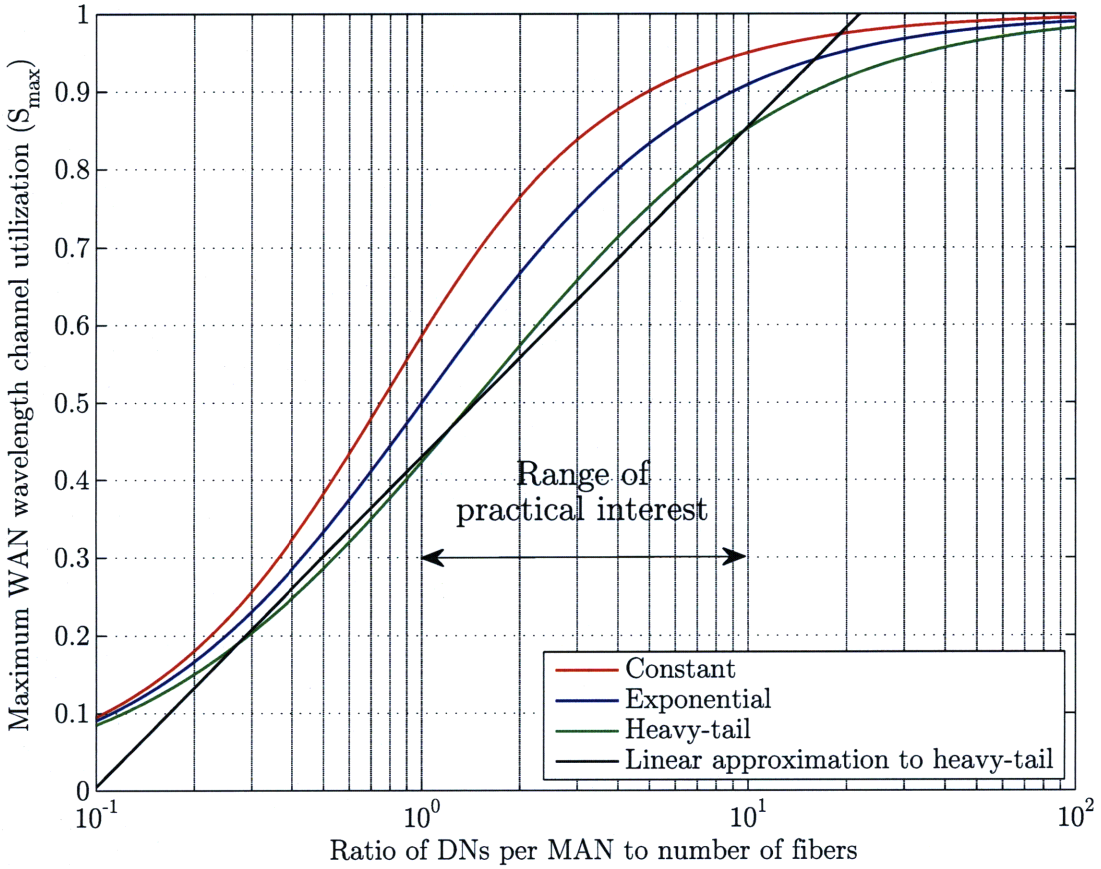


Figure 5-6. Maximum throughput versus ratio of number of DNs per MAN to number of fibers ($\tilde{n}_a/f$) for different flow distributions. In the truncated heavy-tailed case, the distribution tail decays with exponent -1.5, and the lower and upper flow length bounds are 1s and 100s, respectively. A linear approximation to the heavy-tail distribution is also drawn, for which $\mu_1 \approx 0.18$ and $\mu_2 \approx 0.43$. Two fibers per DN, and no wavelength conversion is assumed.

In section 5.6, we defined the first moments of intra- and inter-MAN traffic in terms of w_u and w_m, respectively. In a performance study where only throughput is considered, knowledge of just the first moment of traffic suffices for an analysis under the EPS and OCS/EPS architectures. However, as we saw at the end of Chapter 4—via equation (4.12) specifically—even with a tolerance to infinite delay, the first and second moments of the transaction length distribution affect the maximum achievable throughput on an OFS WAN wavelength channel. Moreover, this maximum achievable throughput is further discounted by a factor κ_h which reflects the relationship between the first moment of the transaction length distribution and the hardware reconfiguration time τ. Therefore, the transaction length distribution of data impacts the performance of OFS. Consistently with our discussion in section 1.1.1, we shall assume that transaction lengths are drawn from a (truncated) heavy-tailed distribution. Moreover, we shall assume that the average transaction length $\overline{L}$ scales linearly with an end-user's average data rate. This is tantamount to the number of transactions sent by an end-user being constant over a fixed period of time. For the purpose of our cost study, we shall assume that the number of transactions sent in a 24 hour period is on the order of 100. Another traffic parameter that must be specified in our study is the proportion of traffic generated in MAN being intra- or inter-MAN in nature. Per our discussion in section 1.1.1, this is a difficult parameter to predict; but since the results of our study are not very sensitive to this parameter, we shall assume an equal proportion of intra- to inter-MAN traffic throughout.

When comparing the costs of supporting a fixed traffic demand under each architecture, we assume that the underlying networks have been optimized in the manner outlined in the last several sections of this chapter. The WAN, as we discussed in section 5.1, is assumed to be fixed, and based upon the realistic fiber plant topology drawn in Figure 5-1. The MANs, on the other hand, are optimized in that they are based upon Moore Graph topologies with optimal node degrees, as derived in section 5.4.1. Similarly, the access networks are based upon optimized versions of the internally amplified DNs in section 5.5. Per section 5.5.1, PONs are assumed to support the maximum number tributaries, each supporting the maximum number of end-users. OFS DNs, on the other hand, are optimized via the global optimization outlined in section 5.6.3.

Lastly, we reintroduce into our study the integrality constraints on various network parameters that have we neglected in our mathematical analyses thus far. In the absence of these integrality constraints, our foregoing analyses are accurate, as long as the networks considered are "large' in the sense that quantities such as wavelength channels provisioned, node degree, number of DNs per MAN, and so forth are much larger than unity. In "small" networks, however, where expensive network resources with large capacity (e.g., WAN wavelength channels) can be very lightly utilized, the omission of integrality constraints can yield misleading results. Since we are considering a range of network sizes in our throughput-cost study, we shall invoke integrality constraints when applicable.

■ 5.7.2 Numerical results

In Figure 5-7, we indicate the minimum-cost architecture as a function of number of end-users per MAN and average end-user data rate. When aggregate MAN bandwidth demand—given by the product of abscissa and ordinate values—is relatively low, EPS is seen to be the most sensible architecture. Electronic switches and routers, to be sure, are less economically scalable technologies than OXCs, but they operate at finer data granularities than OXCs. Thus, when aggregate traffic is low, it is wasteful to provision entire wavelength-granular OXC ports that are poorly utilized—which is why EPS is the minimum-cost architecture in this regime of operation. However, when bandwidth demand between each MAN pair is on the order of multiple wavelengths, optical switching in the WAN is sensible, rendering OCS/EPS the minimum-cost architecture. In fact, under heavy aggregate traffic, the cost difference between EPS and OCS/EPS scales approximately linearly with the product of this aggregate traffic and the difference in cost between a router and OXC port. As aggregate traffic grows even larger, such that the traffic carried on MAN links is on the order of wavelengths, optical switching in the MAN and at the access boundary is most economical, rendering OFS the minimum-cost architecture. Before moving on, a comment on the very large abscissa values in Figure 5-7 (and in the other figures in this chapter) is warranted. Owing to our tolerance to unbounded delay, the abscissa values in our figures are very large. In the presence of realistic delay constraints, the abscissa values at which transitions between optimal architectures occur could be an order of magnitude or more lower.

In Figure 5-8, we depict a horizontal cross-section of Figure 5-7 at a MAN population of 10^6 end-users. On the ordinate, we plot total network cost normalized by the number of end-users and by the average end-user data rate. The figure indicates that when end-users have average data rates below 5×10^7 bps, the EPS and OCS/EPS architectures have the lowest normalized cost. Intuitively, this, again, is because relatively little expensive electronic equipment is necessary to support the aggregate traffic in these architectures; whereas in OFS, wavelength-granular optical equipment is wastefully over-provisioned in the MAN—and to a lesser extent in the WAN—along with expensive long-haul transceivers at end-users operating at the WAN line-rate. Beyond data rates of 5×10^7 bps, however, we see that OFS is the most cost-efficient architecture because: i) the aforementioned optical equipment in the WAN, MAN, and DN equipment is better utilized, and ii) this equipment is more economically scalable than analogous electronic equipment.

In Figures 5-9, 5-10, and 5-11, we deconstruct the normalized total cost in Figure 5-8 according to geographic tier for the different network architectures. We point out that the occasional jaggedness of the curves arises from our enforcement of the integrality constraints on applicable network parameters. Sharp bumps in the curves arise, for example, from additional wavelength channels, or new DNs being provisioned. Figure 5-9 illustrates that, asymptotically, the WAN constitutes the majority of EPS network cost, followed by the MAN, and then the access network. Since all

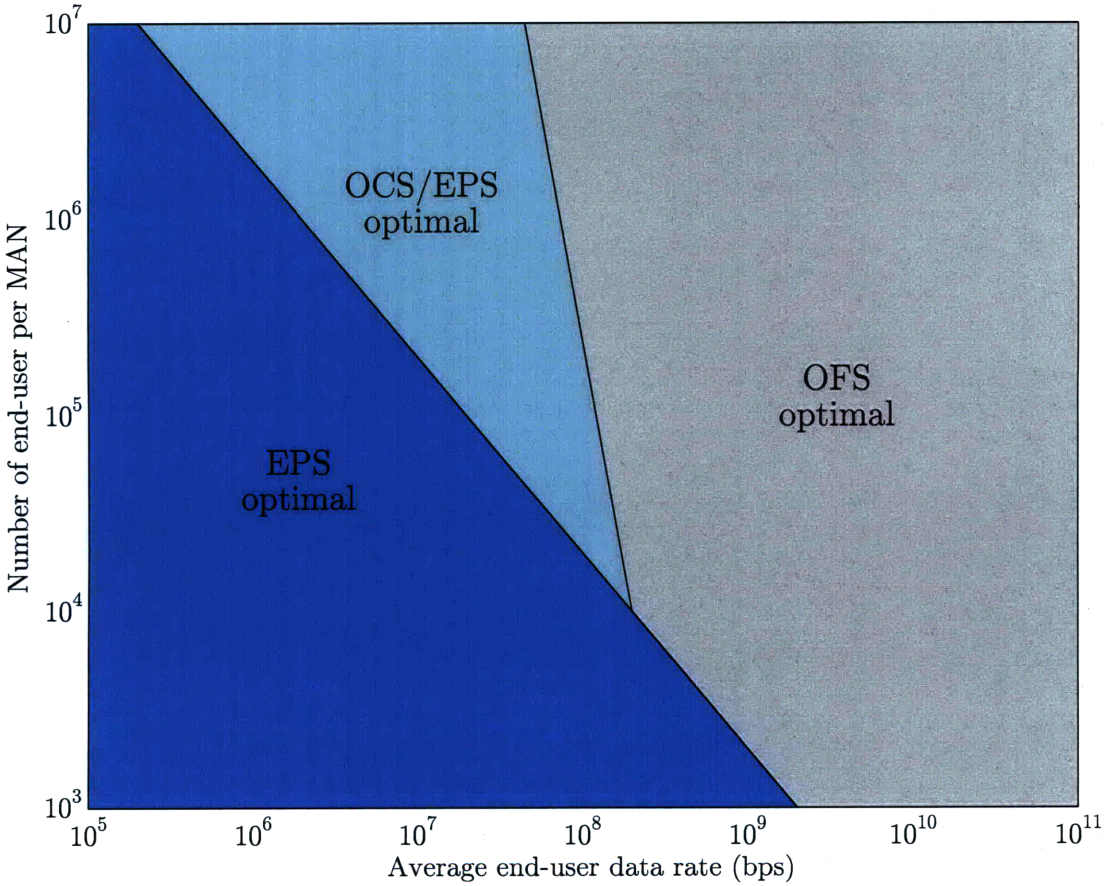


Figure 5-7. Minimum-cost architecture as a function of MAN size and average end-user data rate. It is assumed that: transactions have a truncated heavy-tailed distribution, and that DNs have two fibers and no wavelength conversion.

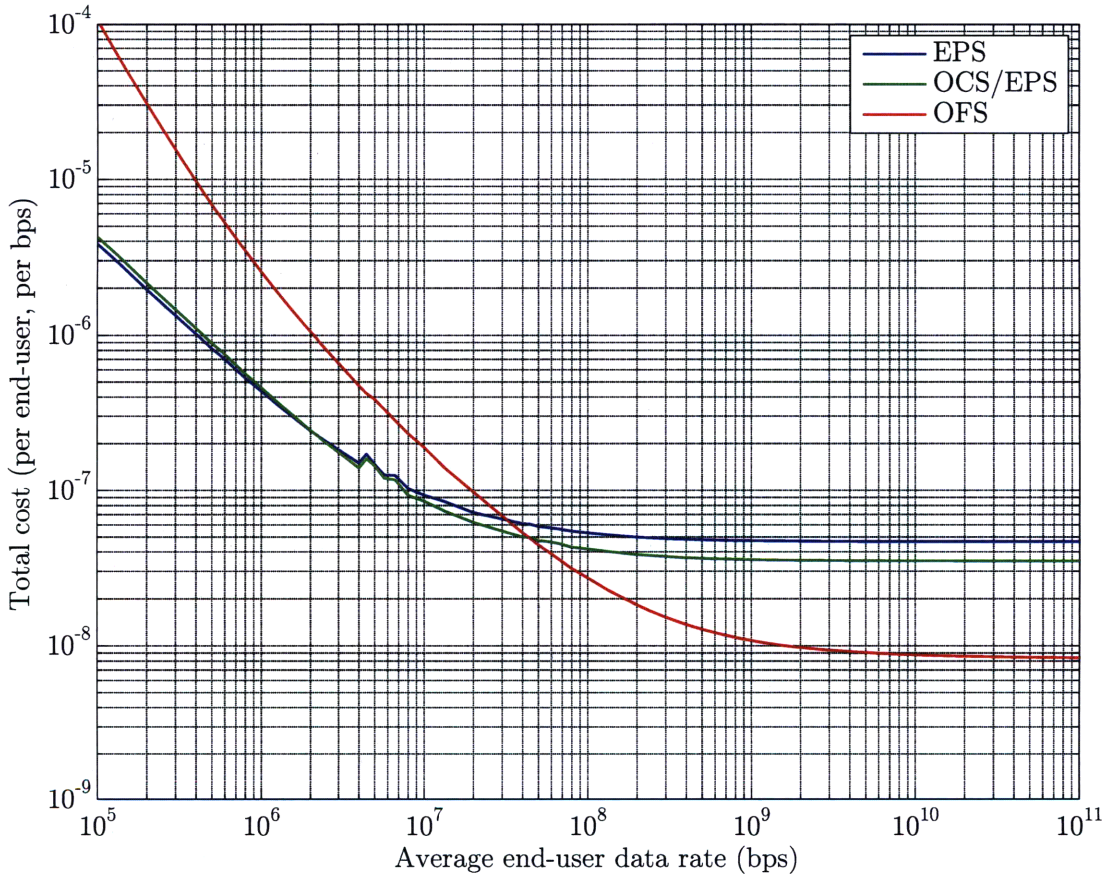


Figure 5-8. Normalized total network cost (in units of "x" used in Table 5.4) versus average end-user data rate. It is assumed that: each MAN has an end-user population of 10^6, transactions have a truncated heavy-tailed distribution, and DNs have two fibers and no wavelength conversion.

three geographic tiers embrace electronic switching in EPS, it is natural that their costs follow their geographic scales. That is, sending information through the WAN is more expensive than sending it through the MAN because of dispersion compensation, optical amplification costs, and larger fiber deployment costs. This same explanation may be used in reference to the MAN and access network cost trends for OCS/EPS drawn in Figure 5-10. The WAN cost component, however, is lowest in OCS/EPS because of the economic scalability of optical switching vis-à-vis electronic switching. Lastly, let us address the component cost trends for OFS in Figure 5-11. Analogous to EPS, OFS embraces a homogeneous paradigm of switching in all geographic tiers—optical switching. We might, therefore, similarly expect the wide-area to constitute the majority of the network cost, followed by the metro-area, and then the access. Figure 5-11 illustrates that is the case, except that the MAN cost is approximately the same as the WAN cost in OFS. The reason for this is that the metro-area is a difficult setting for low-cost network design owing to the dual role that a MAN plays: (dis)aggregation of data, and short-range transport. This is contrast to networks in the wide-area and access which may be more easily optimized for their single roles of transport and aggregation, respectively. MAN design, moreover, is especially difficult with optical components, owing to the absence of economically viable wavelength conversion and buffering.

In Figure 5-12, the last of our plots in this section, we depict average WAN wavelength channel utilization as a function of average end-user data rate. Per Chapter 2 (and our idealized assumption about electronic routers operating optimally), the EPS WAN wavelength channel utilization is unity. WAN wavelength channel utilization under OCS/EPS and OFS exhibits an oscillatory convergence—but to κ_w, since these architectures are less efficient than EPS at using WAN capacity. The damped oscillations during convergence, again, arise from the integrality constraints of certain network parameters. We also observe that the rate of convergence for OFS is slower than that of OCS/EPS. This is a result of the fact that OFS WAN wavelength channel utilization involves the convergence to unity of parameters κ_h and $S_{\max}$, which, recall, capture inefficiencies due to hardware reconfiguration time and our simple scheduling algorithm, respectively.

In summary, we conclude that *OFS is the most cost-scalable architecture of all*, in that its asymptotic normalized cost is several times lower—approximately a factor of four in Figure 5-8—than that of competing architectures. Nevertheless, there is potential for improvement in the throughput-cost tradeoff offered by OFS. In contrast to the wide-area and access environments, which have been significantly optimized, the metro-area retains opportunities for further optimization, albeit with greater difficulty. The scheduling algorithm and MAN physical layer design that we proposed in Chapter 4, though sensible, may not be the most cost-efficient of all pragmatic approaches. Recall that our design required that, for each WAN wavelength channel provisioned for inter-MAN OFS communication, there exist a dedicated wavelength channel in each link of the embedded tree in both source and destination MANs. If an

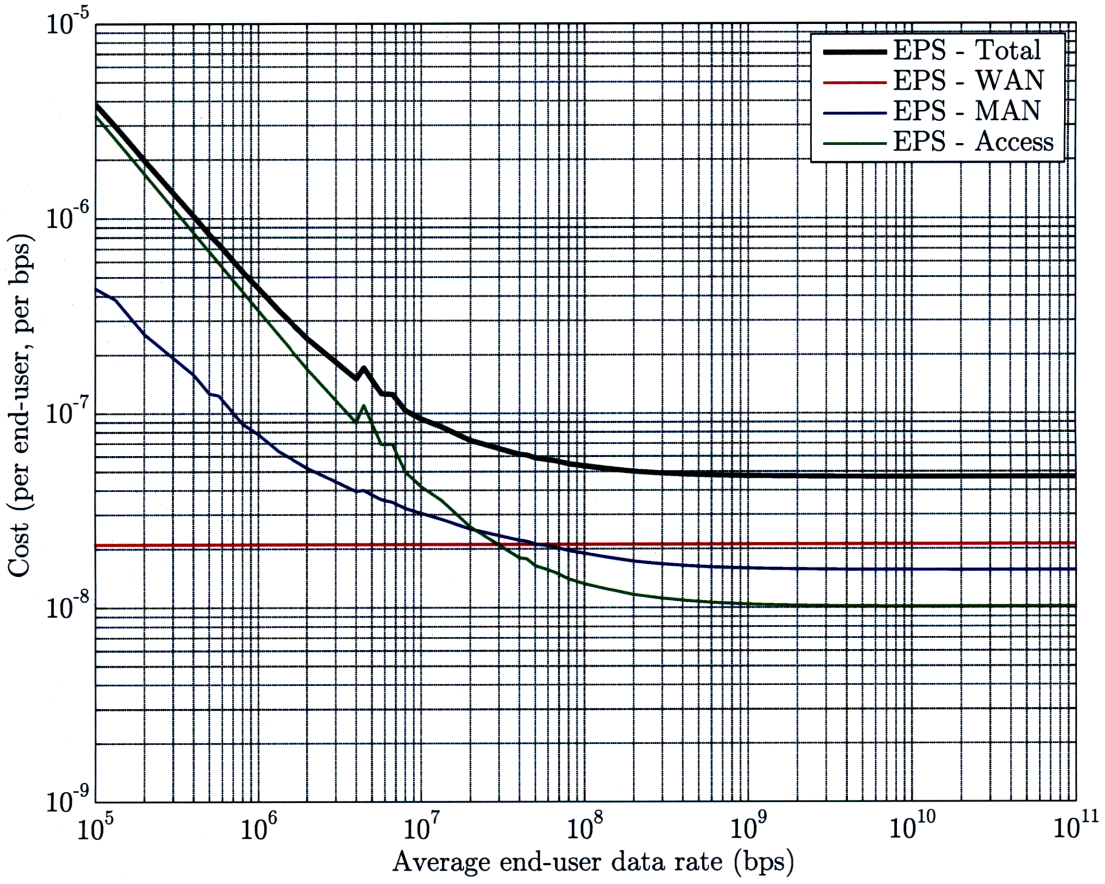


Figure 5-9. Normalized EPS cost components (in units of "x" used in Table 5.4) versus average end-user data rate. It is assumed that: each MAN has an end-user population of 10^6, transactions have a truncated heavy-tailed distribution, and DNs have two fibers and no wavelength conversion.

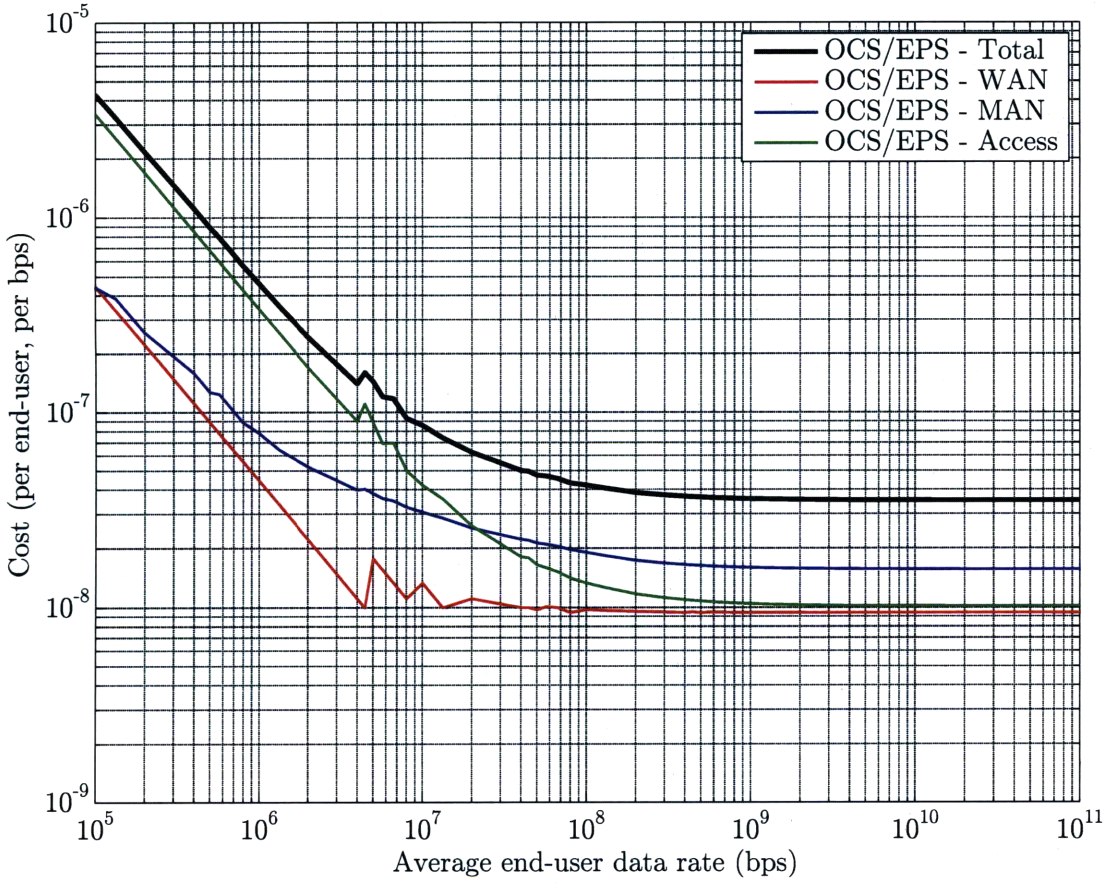


Figure 5-10. Normalized OCS/EPS cost components (in units of "x" used in Table 5.4) versus average end-user data rate. It is assumed that: each MAN has an end-user population of 10^6, transactions have a truncated heavy-tailed distribution, and DNs have two fibers and no wavelength conversion.

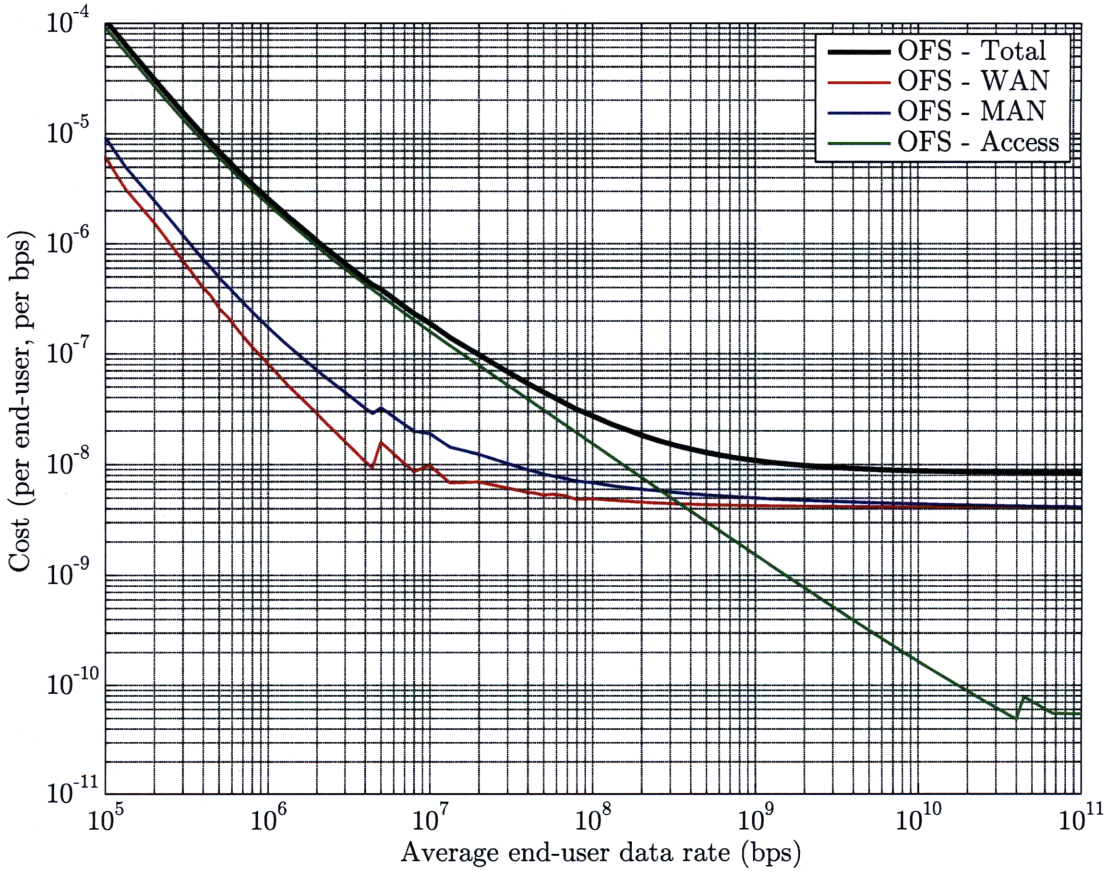


Figure 5-11. Normalized OFS cost components (in units of "x" used in Table 5.4) versus average end-user data rate. It is assumed that: each MAN has an end-user population of 10^6, transactions have a truncated heavy-tailed distribution, and DNs have two fibers and no wavelength conversion.

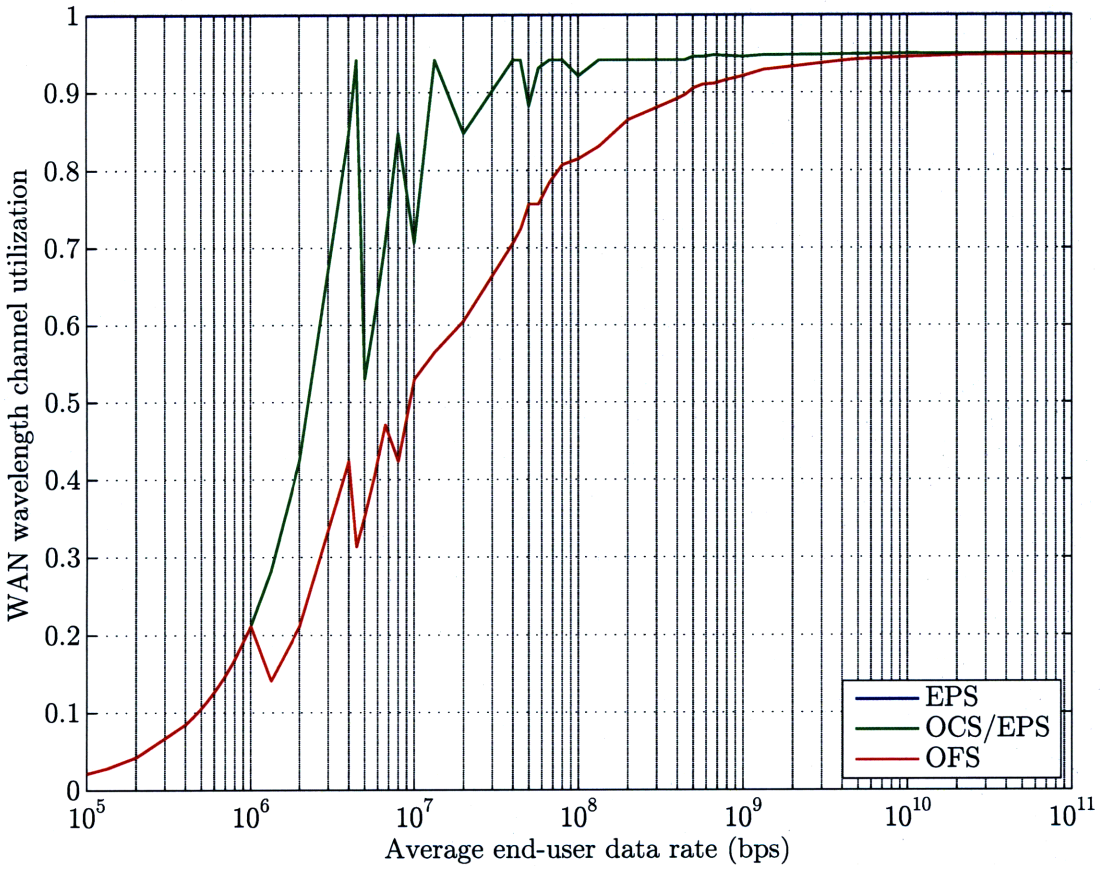


Figure 5-12. WAN wavelength channel utilization versus average end-user data rate. It is assumed that: each MAN has an end-user population of 10^6, transactions have a truncated heavy-tailed distribution, and DNs have two fibers and no wavelength conversion.

implementationally feasible scheduling algorithm that relaxes this dedicated channel requirement could be designed, the cost of the MAN could be reduced. Moreover, since the MAN constitutes a significant portion of total network cost under the OFS architecture, any such improvements to the MAN would noticeably impact total network cost.

As a final remark, we point out that, while the precise values at which transitions in the optimal architecture occur are sensitive to the exact parameter values assumed, the general trends observed are manifestations of present-day cost structures of architectures and their building blocks. Thus, in the absence of disruptive technologies with radically different cost structures, we expect that the trends observed in these figures to hold for a reasonable range of parameter values.

■ 5.8 Hybrid network architectures

In the previous section, we investigated the throughput-cost tradeoffs offered by different *homogeneous* network architectures. In our study, we assumed that a MAN supports a uniform base of end-users, each sending transactions with average length $\overline{L}$ proportional to the long-term average end-user data rate. We found that when end-users have small transactions to send (i.e, $\overline{L}$ is small), EPS is the most sensible architecture; when end-users have large transactions to send (i.e, $\overline{L}$ is large), OFS is most sensible; and for intermediate-sized transactions, OCS/EPS may be most prudent. Given that the size of a transaction greatly impacts the efficiency with which it is served by an architecture, hybrid architectures—architectures comprising two or more of the aforementioned homogeneous network architectures—may be economically advantageous to homogeneous. We devote our attention in this section to investigating this hypothesis.

■ 5.8.1 Modeling assumptions

As discussed in section 1.3.1, great variation exists in the design of hybrid network architectures, as there is a wide range in how tightly the component subarchitectures may be integrated. In this section, we shall focus on one form of hybrid architectures in which component subarchitectures operate in parallel with little interaction. Since the metro-area and access designs are identical for EPS and OCS/EPS, we shall allow end-users belonging to these two architectures to share resources (i.e., wavelength channels, switches/router ports) in these environments, but not in the wide-area where their transport mechanisms differ. Owing to the significant differences between these two architectures and OFS in all three geographic network tiers, OFS end-users are allocated their own network resources from end-to-end. The design of more integrated hybrid architectures may provide better performance-cost tradeoffs than those considered here—but not before their many outstanding physical layer and protocol issues are resolved.

Consistently with our discussion of network traffic in section 1.1.1, we shall assume that the lengths of transactions generated or sunk in a MAN are drawn from a (truncated) heavy-tailed distribution. For simplicity, we shall assume that this aggregate distribution arises from an end-user population with average data rates also drawn from a heavy-tail distribution. In particular, for an end-user with average rate r generating, or sinking, transactions of average length l, we shall assume the following relationship for the truncated heavy-tailed probability distributions of these parameters:

$$p_L(l) = p_R(r).$$

Note that this equality holds when each end-user's transactions are all of constant length. This, of course, is an idealization, as end-users in real networks will likely generate transactions of various sizes.

Motivated by our results in section 5.7, we confine our attention to hybrid architectures in which transactions—or equivalently end-users—are partitioned for service as shown in Figure 5-13. Transactions (equivalently, end-users) are partitioned into three contiguous regions such that the transactions (end-users) in each partition are served exclusively by the indicated architecture. The optimal hybrid architecture is defined as the architecture which minimizes total network cost by judicious positioning of the inner two dotted boundary lines in Figure 5-13. Note that a hybrid architecture reduces to a homogeneous architecture when the dotted boundary lines overlap appropriately.

As part of our study, we shall investigate how the composition of the optimal hybrid architecture changes as average end-user data rates increase. In increasing the average end-user data rate, we shall assume both a rightward shift and expansion of the truncated heavy-tailed distribution drawn in Figure 5-13. Specifically, we assume that increasing average end-user data rates arise from both: i) increases in the lower limit of transaction size (equivalently, average end-user rate), and ii) commensurate scaling in the ratio of the upper limit to lower limit of transaction size (equivalently, average end-user rate). We acknowledge that other means of increasing average end-user data rate exist, and we shall comment on they may change our results in the next section.

■ 5.8.2 Numerical results

In Figure 5-14, we indicate the minimum-cost hybrid architecture as a function of the number of end-users per MAN and average end-user data rate. As in Figure 5-7—and for the same reasons discussed in section 5.7.2—when aggregate MAN traffic is relatively low the homogeneous EPS architecture is optimal, and when aggregate MAN traffic is relatively high the homogeneous OFS architecture is optimal. However, for intermediate aggregate traffic, we observe that hybrid architectures become preferable to homogeneous architectures. When the number of end-users per MAN is small to moderate and the average end-user data rate is moderate, aggregate MAN

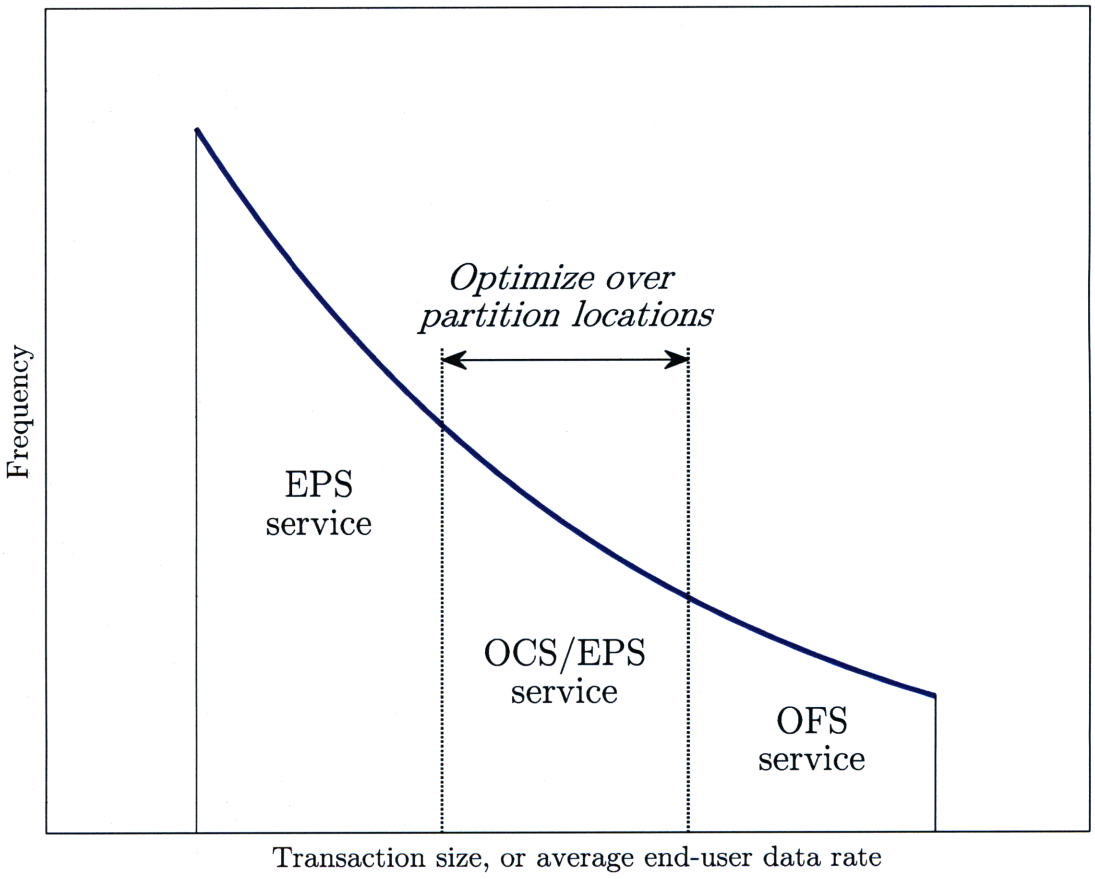


Figure 5-13. Partitioning of the truncated heavy-tail distribution into architecture service regions.

traffic is in the lower range of intermediate values, and EPS is a component of the optimal hybrid architecture—the other component being OFS—since it serves intermediate and low-end end-users most economically. Similarly, when the number of end-users per MAN is moderate to large and the average end-user data rate is low to moderate, aggregate MAN traffic is in the lower range of intermediate values, and EPS is part of the optimal hybrid architecture. However, owing to the large number of end-users per MAN in this latter scenario, the composition of the (remainder of the) optimal hybrid architecture is further refined: at lower average end-user data rates, OCS/EPS serves the intermediate and (few) high-end end-users; whereas at intermediate average end-user data rates, OCS/EPS serves the intermediate end-users and OFS serves the high-end end-users. When the number of end-users per MAN is large and average end-user data rates are moderate, aggregate MAN traffic shifts to the higher range of intermediate values. In this regime, there is sufficient traffic generated from intermediate and low-end end-users (i.e., multiples of wavelengths) such that electronic switching in the WAN is no longer most economically viable, and OCS/EPS and OFS therefore constitute the minimum-cost hybrid architecture.

In Figure 5-15, we depict a horizontal cross-section of the minimum-cost hybrid architecture in Figure 5-14 at a MAN population of 10^6 end-users. On the ordinate, we plot (sub)architecture cost normalized by the number of end-users served by the (sub)architecture and the average data rate of end-users served by the (sub)architecture. Consistent with our results in section 5.7, the asymptotic normalized costs are lowest for OFS, followed by OCS/EPS, and then EPS. The black curve, which represents the normalized cost of the entire hybrid architecture, is essentially a weighted average of the three colored subarchitecture curves. At low average end-user data rates the (black) hybrid architecture curve follows the (blue) EPS curve, and for high average end-user data rates the (black) hybrid architecture curve follows the (red) OFS curve, suggesting the dominance of these architectures at these two extremes. Figure 5-16, which depicts the proportion of traffic served by each subarchitecture as a function of the average data rate of end-users, corroborates this expectation. We observe that at low average end-user data rates, EPS serves the vast majority of traffic; while at high average end-user data rates, OFS serves the vast majority of traffic. OCS/EPS acts as a transitional bridge between these two extremes, serving a significant portion of traffic at intermediate average end-user data rates. Lastly, in Figure 5-17, we depict the proportion of end-users served by each subarchitecture as a function of the average end-user data rate. As expected, first EPS, then OCS/EPS, and lastly OFS serve the vast majority of end-users as average end-user data rate increases. Interestingly, at moderate average end-user data rates, OFS serves the majority of traffic in spite of serving only a very small minority of end-users. This arises from the heavy-tailed nature of traffic—namely, that a major proportion of traffic originates from a minor proportion of transactions.

In closing this section, a few additional comments are in order. As in the case of homogeneous architectures, the curves in our plots of this section exhibit sharp bumps.

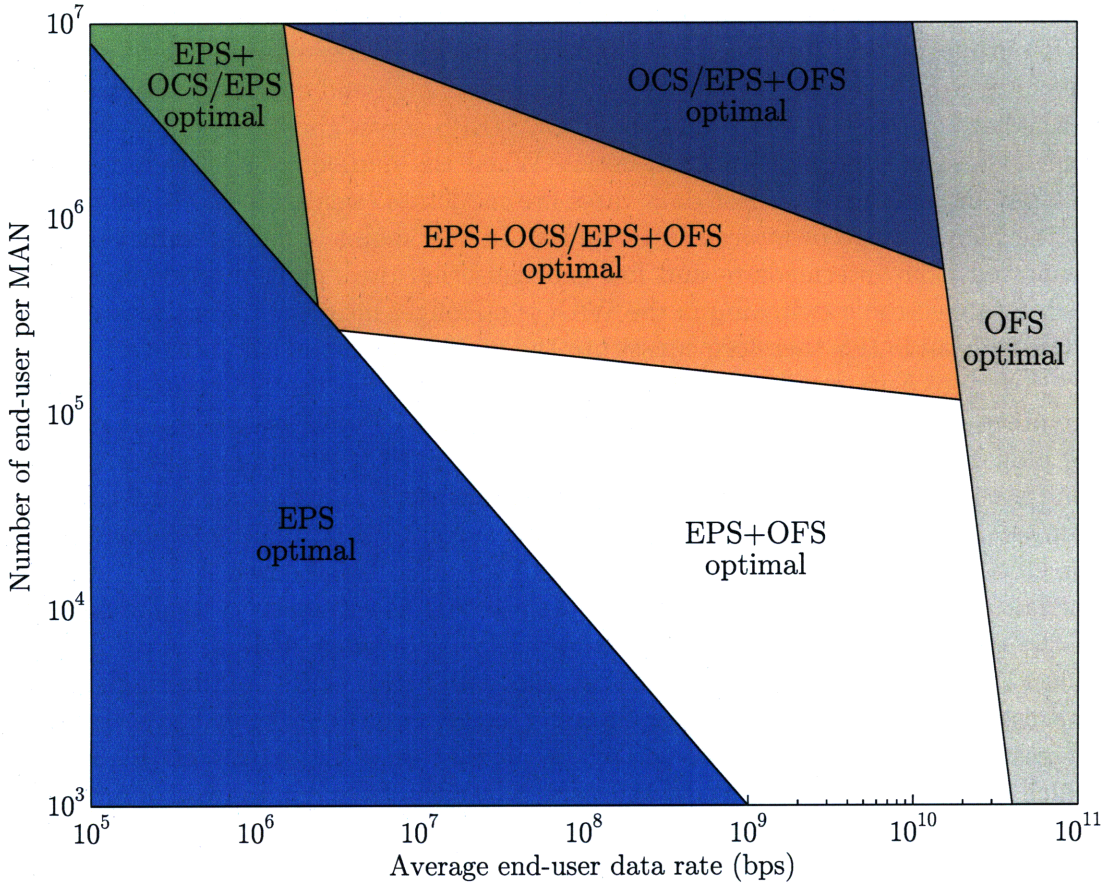


Figure 5-14. Minimum-cost hybrid architecture as a function of MAN size and average end-user data rate. It is assumed that average end-user data rates are drawn from a truncated heavy-tailed distribution with initial lower limit 10^3 bps and width 10^4 bps; and that DNs have two fibers and no wavelength conversion.

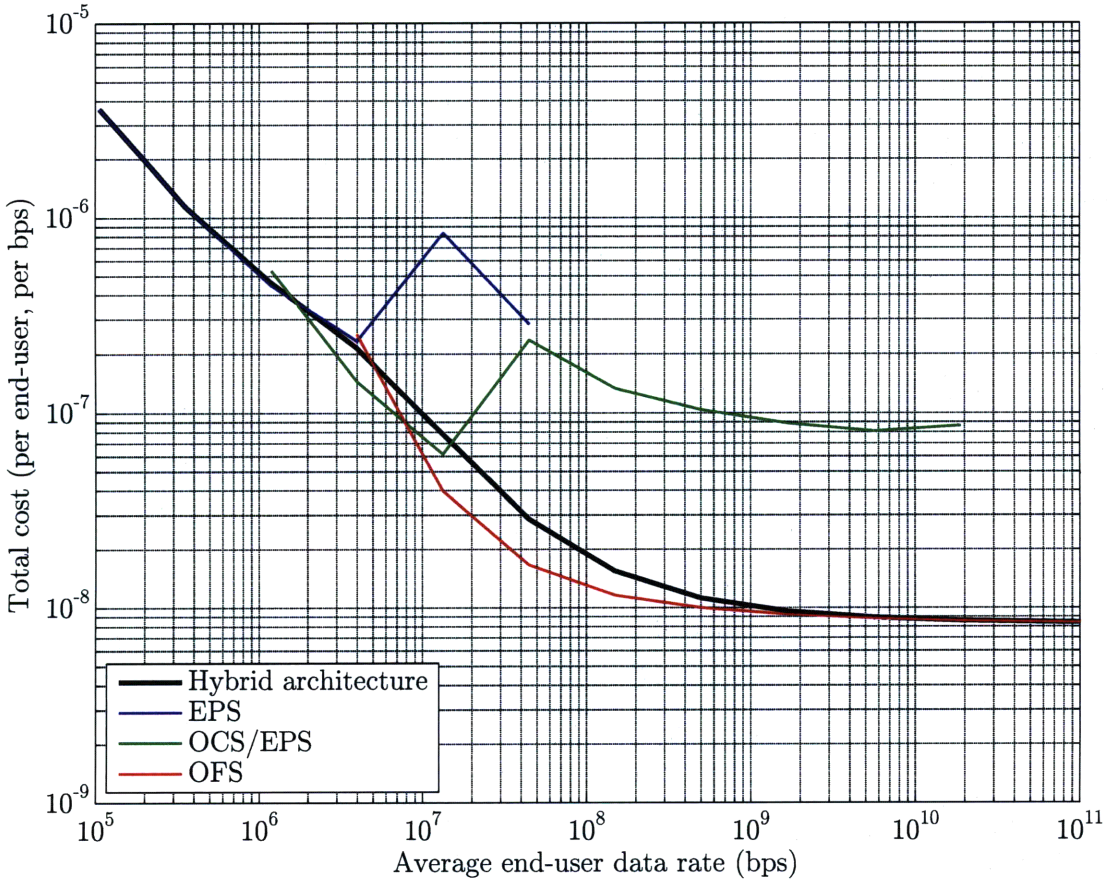


Figure 5-15. Normalized cost components of the minimum-cost hybrid architecture (in units of "x" used in Table 5.4) versus average end-user data rate. It is assumed that average end-user data rates are drawn from a truncated heavy-tailed distribution with initial lower limit 10^3 bps and width 10^4 bps; each MAN has an end-user population of 10^6; and that DNs have two fibers and no wavelength conversion.

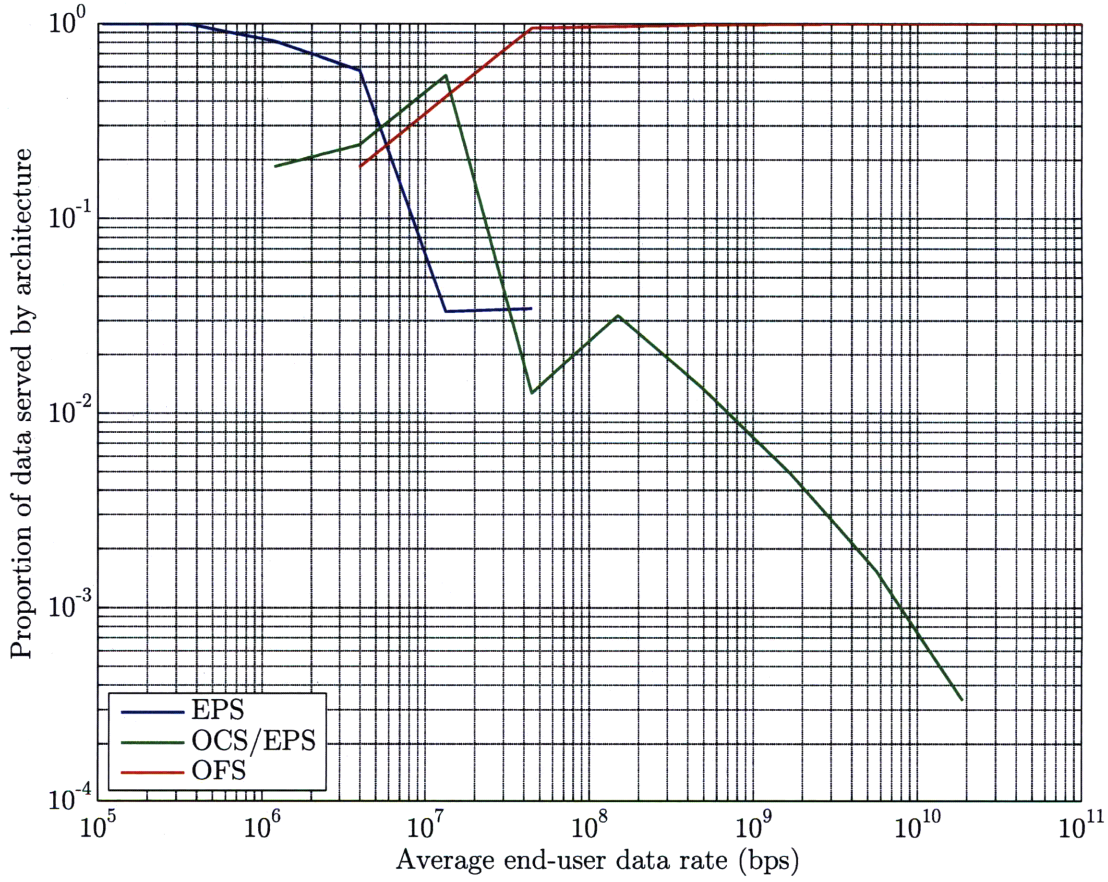


Figure 5-16. Fraction of data served by subarchitectures versus average end-user data rate. It is assumed that average end-user data rates are drawn from a truncated heavy-tailed distribution with initial lower limit 10^3 bps and width 10^4 bps; each MAN has an end-user population of 10^6; and that DNs have two fibers and no wavelength conversion.

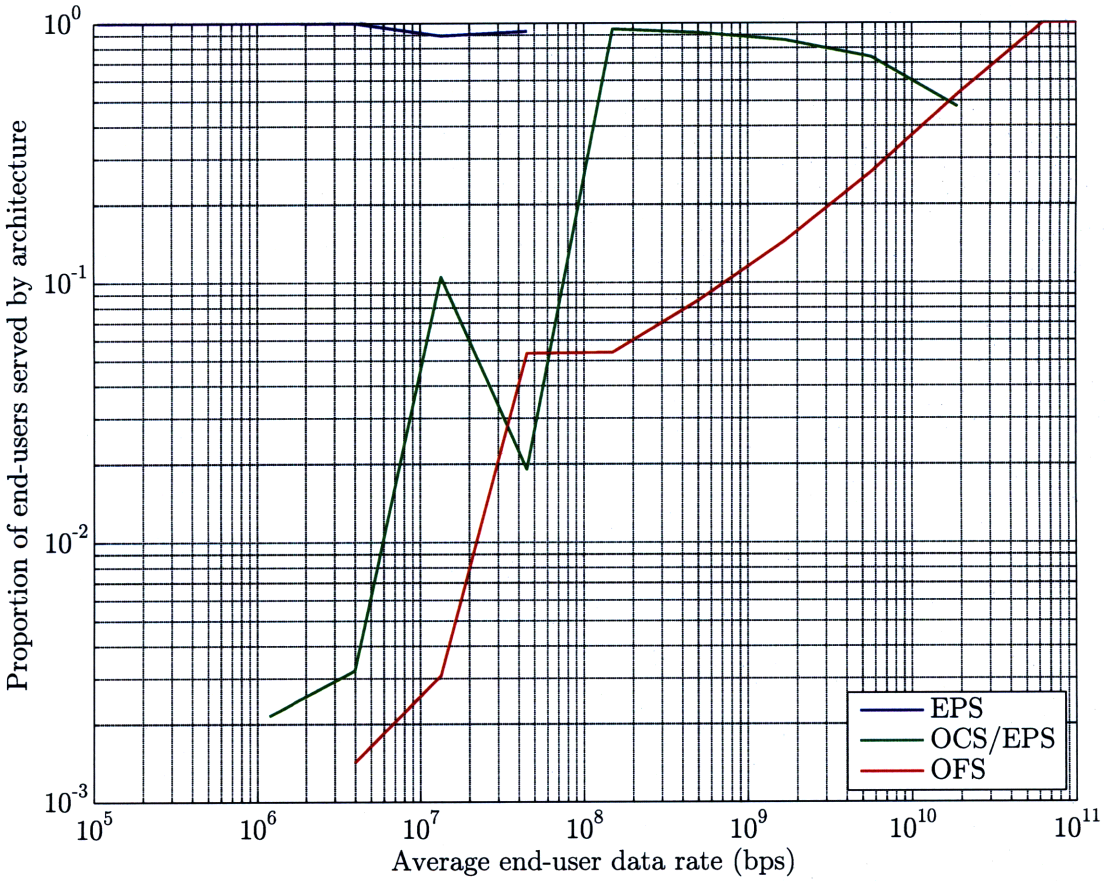


Figure 5-17. Fraction of end-users served by subarchitectures versus average end-user data rate. It is assumed that average end-user data rates are drawn from a truncated heavy-tailed distribution with initial lower limit 10^3 bps and width 10^4 bps; each MAN has an end-user population of 10^6; and that DNs have two fibers and no wavelength conversion.

Again, this arises from the integrality constraints on certain network parameters: occasionally, for small perturbations in parameter values, the integrality constraints induce sudden changes in the composition of the optimal hybrid architecture. We have mitigated these discontinuities by averaging over nearby points, but the jaggedness of the curves remains as a vestige of these discontinuities. Lastly, we mentioned in the previous subsection that there exist alternatives to our method of increasing average end-user data rate by right-shifting and expanding the heavy-tailed distribution. If, for instance, we kept the lower limit of the distribution constant and only increased the upper limit, we would have observed that EPS and OCS/EPS would retained roles in the optimal hybrid architecture at larger average end-user data rates. Conversely, if we kept the width of the distribution constant and only shifted it rightwards, we would have observed EPS and OCS diminishing in importance in the optimal hybrid architecture at smaller average end-user data rates. The conclusion, however, that OFS increases in importance with increasing average end-user data rate is impervious to these modeling assumptions.

■ 5.9 Conclusion

Our investigation of OFS in this thesis culminated in this chapter with a coupling of performance, as derived in previous chapters, with economic attributes vis-à-vis other prominent architectures. Network economics were addressed via approximate CapEx cost models for each of the three geographic network tiers. In an effort to provide as meaningful a comparison as possible, we optimized the architectures under consideration to minimize their respective costs, drawing on our work in previous chapters. Our main conclusion is that *OFS is the most cost-scalable architecture of all*, in that its asymptotic normalized cost is several times lower than that of competing architectures. As such, for sufficiently large average end-user data rates, OFS was observed to be the most economically attractive homogeneous architecture and a critical component of hybrid architectures. Owing to our tolerance to unbounded transaction delay, the average end-user data rates at which OFS becomes economically viable are indeed very large. However, under realistic tolerances to transaction delay, it is expected that OFS would become viable at significantly lower average end-user data rates. Our analysis, moreover, was conservative with respect to OFS in that both the throughput suboptimality of electronic routers, and hardware reconfiguration times were neglected for the other architectures.

The work in this chapter may be extended along various directions. With respect to cost, the most salient limitation of our work was the omission of OpEx. Since OpEx is usually comparable or larger than CapEx, any quantitative—if not qualitative—adjudication among architectures must account for sources of OpEx. With respect to performance, a natural extension would be the inclusion of delay constraints as intimated above. While accounting for delay may not change the qualitative nature of our conclusions, it would allow for more realistic estimates of network costs, as well as

a better sense of the thresholds demarcating the optimality of different architectures. Lastly, the work in this chapter could be broadened to include additional candidate architectures. Even among the architectures that we considered in this chapter, several different variations exist—OFS with wavelength conversion at the MAN-WAN interface, for instance—let alone altogether different architectures, including more tightly integrated hybrid architectures.

Chapter 6

Conclusion

In the four decades since optical fiber was introduced as a communications medium, optical networking has revolutionized the telecommunications landscape. It has enabled the Internet as we know it today, and is central to the realization of Network-Centric Warfare in the defense world. Sustained exponential growth in communications bandwidth demand, however, is requiring that the nexus of innovation in optical networking continue, in order to ensure cost-effective communications in the future.

In this thesis, we presented OFS as a key enabler of scalable future optical networks. The OFS architecture, in particular, was shown to be a cost-effective way of serving large transactions, which are increasing in importance with every passing day. Moreover, only modest technological hurdles exist before OFS can be implemented: the required device technology is presently available, and the remaining algorithmic and protocol challenges should be manageable.

■ 6.1 Summary of contributions

The general idea behind OFS—agile, end-to-end, all-optical lighpaths—is decades old, if not as old as the field of optical networking itself. However, owing (most likely) to the absence of an application for it, OFS remained an under-developed idea—bereft of how it could be implemented, how well it would perform, and how much it would cost relative to other architectures. The contributions of this thesis are in providing partial answers to these three broad questions. Our major conclusions are captured in Table 6.1.

In the introductory chapter of this thesis, we motivated our consideration of OFS, as well as our approach to resolving some of its key implementation issues. In particular, our review of traffic trends and the properties of electronic and optical networking devices led us to a picture of OFS as an architecture that could serve large transactions in a cost-effective manner by means of an all-optical data plane employing end-to-end lighpaths.

In Chapter 2, we addressed OFS in the context of the wide-area via a comparative analysis of network capacity. Employing a constructive approach—in that capacity-achieving algorithms (most appropriate for the metro-area) were outlined—the capacity region of packet-switched architectures (i.e., EPS, OPS) was shown to subsume

that of OFS and OCS, while OFS and OCS, in turn, outperformed OBS. These performance differences arose from the benefits of buffering and scheduling, respectively. For two important families of graphs—bidirectional rings and Moore Graphs—under uniform all-to-all traffic, we observed that the performances under EPS/OPS and OFS/OCS were the same, or almost the same, while the performance under OBS was significantly worse. These results suggest that OBS should be dismissed as a cost-efficient candidate architecture, and that consideration of other attributes (e.g., transaction size, aggregate traffic volume) is required to adjudicate between EPS/OPS and OFS.

In Chapter 3, we addressed the physical layer design of the OFS data plane in the metro and access environments. We began by deriving high-level physical layer constraints for both inter- and intra-MAN OFS communication. For the metro-area, we specialized our results to topologies based upon Generalized Moore Graphs, which were shown (in Chapter 5) to minimize network cost. For the access environment, we first specialized our results to DNs with optical amplification close to the interface with the MAN as a means of supporting a significant number of end-users for efficient statistical multiplexing. In order to support more end-users by improving upon the noise figures of these DNs, we then proposed and analyzed a family of DNs employing optical amplification *within* the DN. Lastly, we addressed the number of supportable end-users, and the approximate pump power required in the aforementioned DN designs.

We began Chapter 4 by outlining the key elements of a scheduling algorithm for inter-MAN OFS communication. This discussion then led us to propose a simple and sensible algorithm embodying these key properties. The remainder of the chapter was devoted to a performance analysis of OFS under our proposed scheduling algorithm. We derived an approximation to the throughput-delay tradeoff offered for inter-MAN OFS communication, followed by an exploration into some of the key tradeoffs in the architecture design space. We found that, for large networks, our simple inter-MAN scheduling algorithm in conjunction with a quasi-static WAN logical topology results in a small penalty in the throughput-delay tradeoff.

Lastly, in Chapter 5, we addressed the economic aspects of OFS. In particular, we introduced the notion of cost via an approximate CapEx model, which enabled a throughput-cost comparison of OFS with other prominent candidate architectures—OPS/EPS, OCS, and OBS/TaG. We also explored hybrid optical network architectures which combine the aforementioned architectures as a means of further lowering overall network cost. Our conclusions from these studies affirmed our expectation that OFS offers a significant advantage over other architectures in economic scalability. In particular, for sufficiently heavy traffic, OFS was observed to handle large transactions at far lower cost than other optical network architectures. In light of the increasing importance of large transactions in both the commercial and defense networks, OFS may therefore be crucial to the future economic viability of optical networks.

Implementation of OFS
• MAN physical layer design (sections 3.1–3.3, 5.4) *Generalized Moore Graphs minimize network cost under shortest path routing with uniform intra- and inter-MAN traffic; node degree optimization balances switching and fiber costs* • Access network physical layer design (sections 3.4–3.6, 5.5) *Internally amplified distribution networks respect OFS's physical layer constraints while supporting many end-users with efficient statistical multiplexing*
Performance of OFS
• Comparative capacity analysis (Chapter 2) *Capacity region of EPS/OPS subsumes that of OFS/OCS due to buffering (and wavelength conversion); OFS/OCS outperform OBS due to scheduling; for bidirectional rings and Moore Graphs under uniform all-to-all traffic, EPS/OPS and OFS/OCS throughputs are (almost) the same, while OBS throughput is significantly worse* *→ OBS should be dismissed as a cost-efficient candidate architecture, and consideration of other attributes is required to adjudicate between EPS/OPS and OFS* • Throughput-delay tradeoff for inter-MAN communication (Chapter 4) *Simple inter-MAN scheduling with a quasi-static WAN logical topology results in a small penalty in the throughput-delay tradeoff for large networks*
Economics of OFS
• Throughput-cost comparison of architectures (Chapter 5) *OFS handles large transactions at far lower cost than other optical network architectures under heavy traffic, and is therefore a critical component of future optical network architectures*

Table 6.1. Major thesis conclusions.

■ 6.2 Future work and challenges

To be sure, the work presented in this thesis does not constitute an exhaustive examination of the OFS architecture. The concluding sections of each chapter, for instance, outline natural extensions to the work carried out in this thesis. Moreover, as mentioned in the introductory chapter, concurrent studies exist on other aspects of OFS: in [110], B. Ganguly reports on the first OFS test-bed, and investigates throughput-delay performance of OFS in the metro-area; and in [109], A. Ganguly addresses a special form of OFS in which flows with strict delay requirements are scheduled via fast optical probing techniques.

Beyond natural extensions to this thesis and the concurrent work carried out in [109,110], there exist avenues of research in further developing and analyzing OFS. In terms of further fleshing out the OFS architecture, mechanism(s) ensuring reliable end-to-end communication for OFS are needed. Presently, the de facto transport layer protocol in communications networks is TCP, which was designed to provide reliable communication of data and flow control in packet-switched networks that are prone to unreliable transmission and congestion. Since OFS is a scheduled flow-based transport architecture, there is no need for TCP's congestion and flow control mechanisms, which are sure to impede the efficiency with which data is transmitted through the network. Instead, a lightweight transport layer protocol, akin to User Datagram Protocol (UDP), should be used in conjunction with an error-checking mechanism, such as Low Density Parity Check Codes (LDPC). Should any errors be detected, a request for retransmission of part, or even the whole, transaction may be made via feedback to the transmitter. With the low error rates of current optical transmission technology, this would not be inefficient in terms of throughput, but would entail additional delay. For transactions with stringent delay requirements, an error-correcting code (ECC) with greater built-in redundancy could be used, albeit at the expense of channel throughput and additional computational resources for decoding [55,59].

Another interesting avenue of future research is less related to how OFS itself may be implemented than to how it may efficiently coexist with other subarchitectures in a hybrid network setting. As discussed earlier, there are varying degrees of subarchitecture integration as far as network resource sharing is concerned. Naturally, it is preferable to integrate component subarchitectures as tightly as possible, as this enables better utilization of network resources. In practice, however, the challenge will be to design of algorithms and protocols that are capable of computing and disseminating network resource allocation information within the time-frames required for agile resource reconfiguration at the architecture level.

On a less technical note, the real-world implementation of OFS may face barriers related to standardization and regulation. In many countries, such as the US, several competing companies may exist at any geographic network tier. While networks operated and owned by different companies (in the US) are mandated by law

to interconnect with one another, proprietary technology and nonstandardized data formats are often used within each network, and, moreover, a minimal amount of control plane information is exchanged among networks. These latter realities present an encumbrance to implementing OFS. Data format conversions at network interfaces add additional cost to be sure, but may not prove to be very detrimental to the viability of OFS. (In fact, regeneration of signals at the MAN-WAN interface can actually confer performance benefits.) On the other hand, since OFS entails rapid setup of end-to-end lightpaths, frequent updates of network resource availability are required, even if WANs possess quasi-static logical topologies. Thus, to implement OFS, regulatory action mandating cooperation among networks may prove necessary. In defense settings, OFS will likely face less severe challenges in standardization and regulation. The US DoD, for instance, being the sole user of the GIG, either owns or leases all of its network resources. It therefore may be able to exert its leverage on suppliers and/or lessors to enable the realization of OFS on the GIG.

To be sure, the obstacles facing the deployment of OFS—technical, regulatory, or otherwise—are significant. Nevertheless, there is good reason to be optimistic about OFS's role in future networks: OFS is presently poised as the only optical network architecture with both the scalability and near-term feasibility to neutralize the economic threat posed by the growth of communications demands. It is this apparent indispensability of OFS that should provide the impetus to surmount these obstacles, and render OFS a major milestone in the evolution of optical network architecture.

Appendix A

Optical Networking Components

In this appendix, we survey a (non-exhaustive) selection of devices which will be the building blocks for future optical networks. As we shall see, optical network devices are not functionally equivalent to electronic network devices, owing to fundamental differences between photons and electrons.

Optical network components may be classified as either passive or active. Passive devices are components that require no external energy for their operation. Active components, on the other hand, require some type of external energy either to perform their function, or to be used over a wider operating range than a passive device, thereby offering greater application flexibility.

■ A.1 Fiber

Optical fiber is the key enabling technology in optical networks, as it serves as the physical communication channel in these networks. An optical fiber is a cylindrical dielectric waveguide that confines and guides light waves along its axis. A fiber comprises a cylindrical silica core surrounded by a silica cladding. The difference in the core and cladding indices determines how light signals travel along the fiber.

Fibers introduced early on, which had core diameters of 50 to 85 μm, are known as multimode fibers. Because the diameters of these fibers are large compared to the operating wavelength of the light signals they guide, light propagation in these fibers may be understood with the geometric optics model, where light is treated as a ray bouncing back and forth in the core, being reflected at the core-cladding interface. A light signal comprises multiple light rays, each of which potentially takes a different path through the fiber. Each of these different paths corresponds to a propagation mode, and the length of the different paths is different. At the end of the fiber, the different modes therefore arrive at slightly different times, resulting in a smearing of the pulse. The smearing of a pulse is known as dispersion, and this specific form of smearing is known as intermodal dispersion [162].

Fibers with core diameters of approximately 8 to 10 μm are known as single mode fibers. In these fibers, the core is a small multiple of the operating wavelength range of the light signal. In order to understand light propagation in single mode fibers, we must abandon the geometric optics model in favor of the wave theory of

light. This forces all the energy in a light signal to travel in the form of a single mode [156, 225]. Single mode fiber eliminates intermodal dispersion, thus enabling a significant increase in the bit rate and distance possible in optical networks.

Light traveling in fiber loses power over distance mainly because of absorption and scattering, the interaction of light waves with phonons (molecular vibrations), in the fiber. Attenuation is an important property of fiber because, together with the distortion properties of fiber, it determines the maximum transmission distance possible. The degree of attenuation depends on the wavelength of the light and on the fiber material. Silica fiber has three low-loss windows in the 0.8, 1.3, and 1.55 μm wavelength bands.

When optical communication systems take advantage of the low loss window at the 1.55 μm wavelength, chromatic dispersion becomes a limiting impairment[1]. Chromatic dispersion is a form of dispersion in optical fiber, where, even in single mode fiber, the fundamental physical properties of silica cause different frequency components of a pulse propagate at different speeds [162]. This effect again causes smearing of the pulse at the output, as in intermodal dispersion. The wider the spectrum of the pulse, the more the smearing due to chromatic dispersion. The high chromatic dispersion at 1.55 μm motivated the development of dispersion-shifted fiber (DSF). DSF is carefully designed to have zero dispersion at the 1.55 μm wavelength. However, by this time there was already a large installed base of standard single-mode fiber deployed for which this solution could not be applied [250]. Thus, dispersion compensating fiber (DCF), which is designed to provide significant negative chromatic dispersion, is often employed between fiber spans, within the huts that also house optical amplifier equipment[2]. Significant drawbacks of DCF, besides its high cost and high loss, are its static nature of compensation in time and the difficulty in matching its dispersion profile to that of the transmission band (which is not constant). As a result, electronic dispersion compensation (EDC), presently an active area of research, actively employs feedback from links to dynamically adjust dispersion levels on a per wavelength basis in response to network churn [267]. PMD is another form of dispersion that arises when the two orthogonally polarized modes of light carrying a signal waveform have slightly different propagation constants leading to smearing of the pulse. This phenomenon arises from the birefringence of optical fiber, resulting from the fact that fiber is not perfectly circular symmetric in practice. PMD has traditionally been compensated for in the electronic domain by means of equalization. At bit rates of 40 Gbps and higher where PMD is especially severe, electronic equalization is very difficult to carry out. Optical PMD compensation is

[1]It turns out that standard silica optical fiber has virtually no chromatic dispersion in the 1.3 μm band, but has significant dispersion in the 1.55 μm band. Thus, chromatic dispersion is not an issue in systems using 1.3 μm wavelength light.

[2]Chirped fiber Bragg gratings (FBGs), discussed in more detail in section A.4.1, are a less commonly employed alternative to compensating for chromatic dispersion.

thus employed by separating the two polarizations of light and delaying the faster of the two polarizations, all in the optical domain.

In addition to being limited by the various forms of dispersion, the information-carrying capacity of fiber is limited by various nonlinear effects [161]. Nonlinear effects may be categorized as resulting either from scattering in the silica medium, or from the dependence of the refractive index on the intensity of the light signal [13]. In stimulated Raman scattering (SRS), energy is transferred from a shorter wavelength signal to a longer wavelength signal when two or more signals at different wavelengths are injected into a fiber. SRS is a broadband effect, in that the transfer of energy occurs over large wavelength spacings of signals. In the case of SBS, the power lost in the scattering process is transferred to an acoustic wave. As opposed to SRS, SBS is a narrowband process, in that the downshift in frequency for systems operating in the 1.55 μm band is approximately 11 GHz. The most important of the second category nonlinear effects are SPM, cross-phase modulation (CPM), and four-wave mixing (FWM). In SPM, the dependence of the refractive index on intensity induces a phase shift that is proportional to the intensity of the pulse. Different parts of the pulse undergo different phase shifts, which gives rise to broadening of the pulse. When more than one signal is present, the nonlinear effects of SPM are enhanced because of the intensities of the other signals in other channels. This nonlinear effect is CPM. FWM is a fiber nonlinearity in which three optical frequencies f_1, f_2 and f_3 mix to produce a fourth frequency, given by $f_4 = f_1 + f_2 - f_3$. When this new frequency falls in the transmission window of the original frequencies, it can cause severe crosstalk[3].

In addition to the aforementioned fibers which are designed to transmit light over long distances with minimal change in the signal, there exist specialty fibers which are used to manipulate a light signal in different ways. As we shall see, specialty fibers can be used for signal amplification, wavelength selection, and wavelength conversion.

■ A.2 Couplers

Couplers can be used for a variety of purposes in optical networks. For example, couplers are used to split light into multiple streams, combine multiple light streams, or tap off a portion of optical power. Couplers are also used as building blocks in more complex devices, such as the Mach-Zehnder interferometer (MZI), which transfers a selective range of optical power from one fiber to another. Couplers are characterized by the number of input and output ports they have, in addition to whether they are fabricated from fused optical fibers or by means of planar optical waveguides using materials such as lithium niobate ($LiNbO_3$).

The directional coupler, or the basic 2×2 coupler, consists of two input ports and two output ports and is used to combine and split signals in an optical network. The

[3]Crosstalk is the gain dependence of an optical signal in one channel on the presence or absence of signals in other channels.

2×2 coupler takes a fraction α, known as the coupling ratio, of the power from input 1 and places it on output 1 and the remaining fraction $1 - \alpha$ on output 2. Likewise, a fraction $1 - \alpha$ of the power from input 2 is distributed to output 1 and the remaining power to output 2. The 2×2 coupler can be designed to make the coupling ratio either dependent or independent of wavelength. When the coupling ratio is dependent upon the wavelength, the device is called a WDM or dichroic coupler.

Directional couplers used to tap off a small portion of the power from a light stream are called taps. The fraction of power tapped off—usually less than 10%—is used to monitor the signal level or quality of a link.

■ A.2.1 Passive star coupler

The PSC is a passive broadcasting device which generalizes the symmetric directional coupler. Optical signals entering from its multiple input ports are combined and divided equally among its output ports. PSCs do not contain any wavelength-selective elements as they do not attempt to separate individual wavelength channels. The number of input and output ports in a PSC need not be the same. The theory and construction of the PSC are detailed in [91, 211, 283].

The broadcast property of the PSC makes it ideal for distributing information to all nodes in WDM networks. Star topology networks based upon the PSC as the central broadcast device require a lower power budget compared to networks with a linear bus topology or a tree topology. These advantages have led to numerous proposals for PSC-based broadcast-and-select networks, as we shall see in the coming chapters.

■ A.3 Isolator and circulator

An optical isolators is a reciprocal[4] passive device that allows light to pass through it in only one direction. Such a device is important in preventing scattered or reflected light from traveling in the reverse direction. Isolators often find use at the outputs of optical amplifiers and lasers primarily to prevent reflections from entering these devices, which would degrade their performance.

An optical circulator is a nonreciprocal multiport passive device that directs light sequentially from port to port in only one direction. In a three-port circulator, an input signal on port 1 is sent out on port 2, an input on port 2 is sent out on port 3, and an input signal on port 3 is sent out on port 1. Circulators are often used to construct optical add/drop elements.

[4]A reciprocal device works identically if the inputs and outputs are reversed. That is, if a wave propagates from an input of the device to an output, a wave injected at the output will propagate backward to the input in exactly the same way.

A.4 Wavelength-selective devices

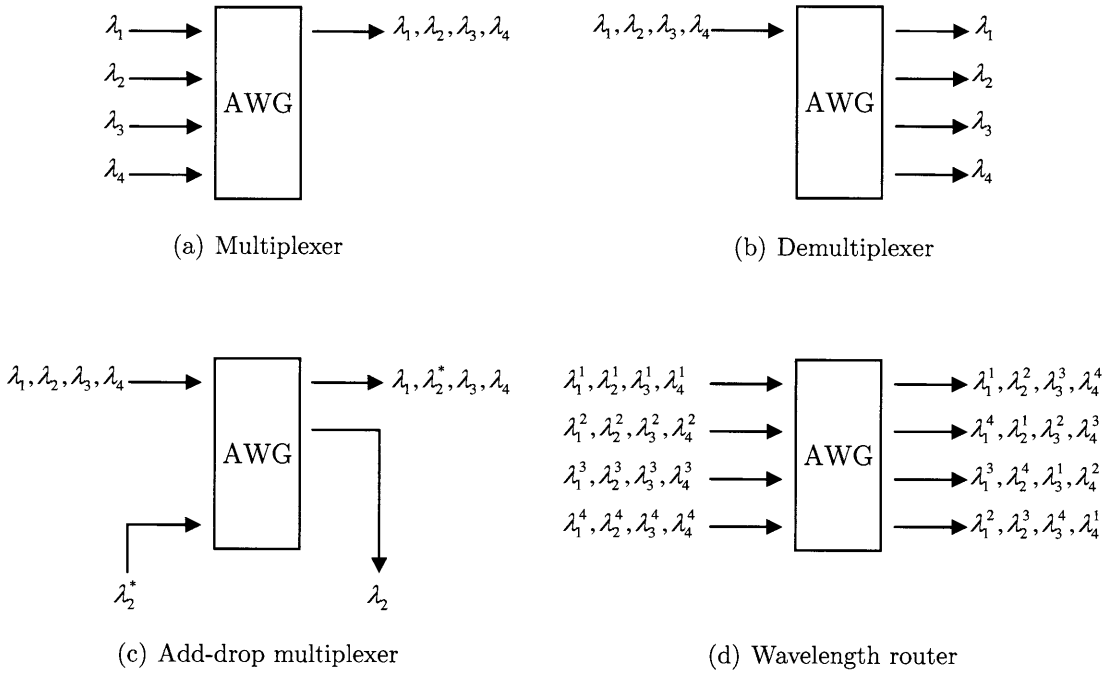
In this subsection, we examine filters which are basic wavelength-selective devices that can be used in WDM networks to achieve functions such as multiplexing and demultiplexing of wavelength signals, or for more complex tasks such as wavelength switching. The performance of a filter is characterized by its: (a) insertion loss, which is the input-to-output loss, (b) polarization-dependent loss (PDL), which is the dependence of loss on the state of polarization of the input light, (c) temperature sensitivity, (d) passband flatness, and the sharpness of its passband skirts to avoid crosstalk.

A.4.1 Gratings

A grating is a device used to combine or separate individual wavelengths by means of a periodic structure or perturbation in a material. The variation in the material has the property of reflecting or transmitting light in a certain direction depending on the wavelength. Gratings exploit interference among multiple optical signals originating from the same source but with different relative phase shifts [36].

Gratings can be categorized as either transmitting or reflecting. In a transmitting grating, multiple narrow slits are equally spaced on a grating plane. Light incident from a source on one side of the grating is transmitted through each slit, spreading out in all directions by the phenomenon of diffraction. On an imaging plane parallel to the grating plane, an interference pattern is observed. Since different wavelengths interfere constructively at different points on the imaging plane, the grating effectively separates a WDM signal into its constituent wavelengths. If the transmission slits are replaced by narrow reflecting surfaces, with the rest of the surface being non-reflecting, then we have a reflection grating. The principle of operation of this device is similar to that of the transmission grating. The majority of gratings used in practice are reflection gratings since they are easier to fabricate. Gratings can also be fabricated in concave geometry, where the slits or reflecting surfaces are located on the arc of a circle.

Bragg gratings are devices where the periodic perturbation occurs in the propagating medium. This perturbation is usually a periodic variation in the refractive index of the medium. One incarnation of a Bragg grating is the FBG, in which two ultraviolet light beams set up a periodic interference pattern in a section of the core of a germania-doped silica fiber [30, 46, 164]. Since this material is sensitive to ultraviolet light, the interference pattern induces a permanent periodic variation in the core refractive index along the direction of light propagation. The grating reflects light with a wavelength corresponding to twice the grating period. Long-period fiber gratings, which have periods much greater than a wavelength, are fabricated in the same manner as FBGs and are used primarily as filters inside EDFAs to compensate for a non-flat gain spectrum [301, 302].



(a) Multiplexer (b) Demultiplexer

(c) Add-drop multiplexer (d) Wavelength router

Figure A-1. Different AWG functions.

■ A.4.2 Arrayed waveguide grating

The AWG is a passive wavelength-selective device consisting of multiport couplers interconnected by an array of waveguides. Dragone *et al.* [89, 90] discuss the construction and physical properties of the AWG.

The AWG has several uses, as illustrated in Figure A-1 [118, 142, 203, 284]. The AWG can be used as a wavelength multiplexer, a wavelength demultiplexer, or it can combine multiplexing and demultiplexing functions as an OADM. The most powerful application of the AWG is a strictly nonblocking wavelength permutation router.

■ A.4.3 Fabry-Perot filter

A Fabry-Perot (FP) filter or interferometer consists of the cavity formed by two highly reflective mirrors placed parallel to each other. The input light beam to the filter enters the first mirror at right angles to its surface. After one pass through the cavity, a part of the light leaves the cavity through the second facet and a part is reflected. A part of the reflected wave is again reflected by the initial facet to the second facet. For those wavelengths for which the cavity length is an integral multiple of half the wavelength in the cavity — so that a round trip through the cavity is an integral multiple of the wavelength – all the light waves transmitted through the second facet add in phase. This is known as the Bragg condition, and such wavelengths are called the resonant wavelengths of the cavity.

An FP filter can be tuned to select different wavelengths in one of several ways. The simplest approach is to change the cavity length. The same effect can be achieved by varying the refractive index within the cavity. Mechanical tuning of the filter can be effected by moving one of the mirrors so that the cavity length changes. Another approach to tuning is to use a piezoelectric material within the cavity. A piezoelectric filter undergoes compression on the application of a voltage. Thus, the length of the cavity filled with such a material can be changed by the application of a voltage.

■ A.4.4 Multilayer dielectric thin-film filter

A thin-film resonant cavity filter (TFF) is an FP interferometer, where the mirrors surrounding the cavity are implemented with multiple reflective dielectric thin-film layers. This device acts as a bandpass filter, passing through a particular wavelength and reflecting all the other wavelengths. The wavelength that is passed through is determined by the cavity length. A thin-film resonant multicavity filter (TFMF) consists of two or more cavities separated by reflective dielectric thin-film layers [264]. As more cavities are added, the top of the passband becomes flatter and the skirts become sharper.

In order to obtain a multiplexer or a demultiplexer, a number of these filters can be cascaded. Each filter passes a different wavelength and reflects all the others. When used as a demultiplexer, the first filter in the cascade passes one wavelength and reflects all the others onto the second filter. The second filter passes another wavelength and reflects the remaining ones, and so on. This device has many features that make it attractive for system applications. As mentioned, it is possible to have a very flat top on the passband and very sharp skirts. Furthermore, the device is extremely stable with regard to temperature variations, has low loss, and is insensitive to the polarization of the signal.

■ A.4.5 Mach-Zehnder interferometer

An MZI is an interferometric device that makes use of two interfering paths of different lengths to resolve different wavelengths. MZIs are typically constructed with integrated optics and consist of two 3 dB directional couplers interconnected through two paths of differing lengths. The substrate is usually silicon, and the waveguide and cladding regions are silica.

MZIs are useful as both filters and (de)multiplexers. Even though there are better technologies for making narrow band filters, such as TFMFs, MZIs are still useful in realizing wideband filters. For example, MZIs can be used to separate the wavelengths in the 1.3 and 1.55 μm bands. Narrowband MZI filters can be fabricated by cascading a number of stages, but this leads to large loss. Furthermore, the passband of narrowband MZIs is not flat, whereas TFMFs can have flat passbands and good stopbands. Finally, crosstalk performance of MZIs is far from ideal. MZIs are useful as two-input, two-output multiplexers and demultiplexers. They can also be used as

tunable filters, where the tuning is achieved by varying the temperature of one of the arms of the device. This causes the refractive index of that arm to change, which in turn affects the phase relationship between the two arms and causes a different wavelength to be coupled out. The tuning time is of the order of several milliseconds.

■ A.4.6 Acousto-optic tunable filter

The acousto-optic tunable filter (AOTF) is presently the only device that is capable of selecting several wavelengths simultaneously. The AOTF is an example of an optical device whose construction is based upon the interaction of sound and light. In an AOTF, an acoustic wave is used to create a Bragg grating in a waveguide, which is then used to perform wavelength selection. In an AOTF, launching multiple acoustic waves simultaneously allows the Bragg condition to be satisfied for multiple optical wavelengths simultaneously. Thus, an AOTF may be used as a 2×2 dynamic wavelength cross-connect. Such an AOTF is an unconstrained two-input, two-output dynamic cross-connect, since the set of wavelengths to be exchanged can be changed by varying the frequencies of the acoustic waves launched into the device. In principle, larger dimensional dynamic cross-connects can be built by cascading 2×2 cross-connects.

The AOTF has not yet lived up to its promise as either a versatile tunable filter or a wavelength cross-connect. One reason for this is the high level of crosstalk that is present in the device. Another reason is that the passband width is fairly large (at least 100 GHz), rendering it unsuitable for dense WDM systems where channel spacings are on the order of tens of GHz. However, recent theoretical work indicates that some of these problems, particularly crosstalk, may be mitigated [271] . The crosstalk issues that arise in AOTFs when used as wavelength cross-connects are discussed in detail in [152].

■ A.5 Transmitters

Before discussing optical transmitters, we touch upon the processes by which photons can be absorbed and emitted. In a simple two-level atomic system, an electron may move from energy level E_1 to a higher energy level E_2 by either being pumped externally to that level, or by absorbing the energy from an incident photon. An excited electron may then return to the ground state either spontaneously or by stimulation, emitting a photon in the process. The corresponding photon generation processes are called spontaneous emission and stimulated emission, respectively. Spontaneously generated photons have random phases and frequencies, as electrons may return to the ground state from any other energy level. This type of light thus has broad spectral width and is incoherent. Stimulated emission occurs when an external simulant, such as an incident photon, causes an excited electron to drop to the ground state. The photon emitted in this process has the same energy as the incident photon and is in phase with it. This type of light is thus coherent.

Light sources for optical networks have evolved from the LED to the laser, of which there are several types. The four main lasers genres are the Fabry-Perot (FP) laser, distributed-feedback laser (DFB), tunable laser, and vertical cavity surface-emitting laser (VCSEL). A laser is a device which converts electrical energy into monochromatic light by the enclosure of an optical amplifier in a reflective cavity that causes the amplifier to oscillate via positive feedback.

When assessing the merits of a laser as a light source for a WDM application, the following properties are desired [250]: (a) high output powers in the 0-10 dBm range (pump lasers for amplifiers are required to produce much higher power levels), (b) narrow spectral width so that the signal can pass through intermediate filters, and multiple channels can be closely spaced, and (c) small wavelength drift relative to the wavelength spacing between adjacent channels.

■ A.5.1 Light-emitting diode

The LED is a good, inexpensive alternative to a laser for low-data rate, short-distance optical network applications. An LED is a forward-biased *pn*-junction which relies on spontaneous emission from the recombination of injected minority carriers to produce light. Owing to the broad spectrum of light produced by spontaneous emission, LEDs are broad spectrum light sources. Furthermore, LEDs are low-power optical sources with typical output powers on the order of -20 dBm, and cannot be directly modulated at data rates higher than a few hundred Mbps.

For applications which require low data rates but narrow spectral widths, LEDs can be used in conjunction with a technique known as spectral slicing as an inexpensive alternative to lasers. Spectral slicing involves placing a narrow passband optical filter in front of an LED, so that the optical filter selects a portion of the LED's output. Because different filters can be used to select non-overlapping spectral slices of the LED output, one LED can be shared by a number of users.

■ A.5.2 Fabry-Perot laser

In the FP laser, the gain medium is contained in an optical cavity of rectangular geometry with partially reflecting mirrors, or facets, on two sides. The facets provide partial internal reflection, as in the FP filter discussed previously. In an FP cavity, light is reflected back and forth between the mirrors with wavelengths for which the cavity length is an integral multiple of half the wavelength being reinforced by constructive interference. If the combination of the amplifier gain and the facet reflectivity is sufficiently large, the amplifier will start to produce light output in the absence of an input signal. At this point, the device no longer acts as an amplifier but as an oscillator. This occurs because the stray spontaneous emission, which is always present at all wavelengths within the bandwidth of the amplifier, gets amplified even without an input signal and appears as the light output. Since the amplification process is due to stimulated emission, the light output of a laser is coherent.

Recall that in order for the FP laser to oscillate at a particular wavelength, the wavelength must be within the bandwidth of the gain medium that is used, and the length of the cavity must be an integral multiple of half the wavelength. For a given laser, the wavelengths that satisfy this second condition are called the longitudinal modes of that laser. The FP laser will usually oscillate simultaneously in several longitudinal modes, and is thus known as a multiple-longitudinal mode (MLM) laser. MLM lasers have large spectral widths, typically around 10 nm.

■ A.5.3 Distributed-feedback laser

DFB lasers achieve single-longitudinal mode (SLM) operation and are thus appropriate for high-speed optical transport systems. In contrast to the FP laser where feedback of light is isolated at the facets of the cavity, in a DFB laser a series of closely spaced reflectors provide feedback in a distributed fashion throughout the cavity. By tailoring the design of these reflectors, which are normally some kind of grating, the device can be made to oscillate in an SLM with narrow linewidth. In this section of the cavity, the incident wave undergoes a series of reflections which add in phase if the Bragg condition is satisfied. The Bragg condition will be satisfied for a number of wavelengths, but the strongest transmitted wave occurs at the wavelength for which the grating period is equal to half the wavelength, rather than some other integer multiple of it.

When the grating used by a laser to achieve feedback is no longer within the gain cavity but outside the gain region, the laser is called a distributed Bragg reflector (DBR) laser. The main advantage of the DBR laser is that the gain region is decoupled from the wavelength selection region. Thus, it is possible to control both regions independently, thereby enabling tunability of the laser. Specifically, by changing the refractive index of the wavelength selection region, the laser can be tuned to a different wavelength without affecting its other operating parameters.

For WDM systems, it is very useful to package multiple DFB lasers at different wavelengths inside a single package. This device can then serve as a multiwavelength light source, or as a tunable laser if only one of the lasers in the array is turned on at a time. These lasers can all be grown on a single substrate in the form of an array. The primary reason that these laser arrays are not manufactured in large volumes is the relatively low yield of the array as a whole.

■ A.5.4 Vertical cavity surface-emitting laser

A VCSEL is another type of laser that achieves SLM operation. In our discussion of the FP laser, we noted that MLMs arise because of the large cavity length of the cavity relative to the wavelength of the light used. VCSELs are lasers with cavities sufficiently small that only one longitudinal mode occurs within the gain bandwidth of the laser. For ease of fabrication, this thin active layer is deposited on a semiconductor substrate. This leads to stacks of up to 30 thin mirroring layers placed on both sides

of a semiconductor substrate. The layers of mirrors selectively reflect a narrow range of wavelengths, with the top stack having slightly less reflectivity which allows light to exit from the top [198].

The advantages of VCSELs compared to edge-emitting lasers include simpler and more efficient fiber coupling, easier packaging and testing, and the ability to be integrated into multiwavelength arrays. In a WDM system, many wavelengths are transmitted simultaneously over each link. Usually, this requires a separate laser for each wavelength. The cost of the transmitters can be significantly reduced if all the lasers can be integrated on a single substrate [140]. Moreover, an arrayed laser can be used as a tunable laser by simply turning on only the required laser in the array. The use of surface-emitting lasers enables the fabrication of two-dimensional arrays of lasers. Much higher array packing densities can be achieved using surface-emitting rather than edge-emitting lasers because of this added dimension. However, it is harder to couple light from the lasers in this array into optical fiber. These arrayed lasers have the same yield problem as other arrayed laser structures.

■ A.5.5 Tunable laser

Tunable lasers are a key technology for WDM networks, as they allow reconfigurability of optical networks, are essential to optical packet switched networks (if deployed), and alleviate inventory difficulties associated with fixed lasers [140,172]. Tunability of lasers can be achieved by three different mechanisms [250]: (a) injecting current into a semiconductor laser (to change the refractive index of the material, thereby changing the lasing wavelength), (b) temperature tuning (since the wavelength sensitivity of a semiconductor laser to temperature is approximately 0.1 nm/°C), and (c) mechanical tuning.

Table A.1 lists examples of tunable laser technologies and typical performance values [166]. Owing to their existence for some time now, DFB lasers are quite reliable but their tuning times are slow. The conventional DBR laser may be modified by injecting current into the Bragg region that is decoupled from the gain region. This allows the wavelength to be controlled independently of the output power in a simple structure, with high output power and modest tuning range. The sampled grating distributed Bragg reflection (SG/DBR) laser employs two gratings which are interrupted, or sampled, at different periods [154]. This tunable laser has a fairly complex structure and a large tuning range of 40 nm. The VCSEL/MEMS laser achieves tunability by having the upper mirror be a movable MEMS membrane [299]. The cavity spacing can be adjusted by moving the upper mirror by applying a voltage across the upper and lower mirrors. This implementation offers simple tuning over a large range, but the hybrid packaging structure is complex. Finally, as discussed earlier, another method of achieving laser tunability is the use of an array of wavelength-differentiated lasers, where only one is turned on at any time.

Type	Tuning	Range	Speed	Output	Advantages	Disadvantages
DFB	T	3-9 nm	10 s	2 mW	Reliable	Small range; slow
DBR	C	8-10 nm	10 ms	5 mW	Simple device; high output	Intermediate range
SG/DBR	C	40 nm	10 ms	2 mW	Wide range	Complex control
VCSEL/ MEMS	M	32 nm	10 ms	20 mW	Wide range; high output	Hybrid packaging

Table A.1. Tunable laser technologies and performance. (C: current, T: temperature, V: voltage.) Adapted from [166].

■ A.5.6 Other lasers

In addition to DFB, DBR, and VCSEL lasers, the external cavity laser can achieve SLM operation. The external cavity laser suppresses all but one longitudinal mode by using another cavity, called an external cavity, following the primary cavity where gain occurs. Like the primary cavity, the external cavity has resonant wavelengths. The laser is thus capable of oscillating only at wavelengths that resonate with both the primary and external cavities. By suitable design of the two cavities, it can be ensured that only one wavelength in the gain bandwidth of the primary cavity satisfies this condition. Furthermore, if the grating or wavelength-selective mirror in the external cavity can be changed, then the external cavity laser may be used as a tunable laser.

Mode-locked lasers are MLM lasers in which the relative amplitudes and phases associated with the different modes are 'locked' to values representing Fourier coefficients for a periodic pulse train [13,323]. Mode-locked lasers are thus used to generate narrow optical pulses for high speed TDM systems.

In addition to being used as light sources for data transmission, lasers are required to supply external energy to optical amplifiers. Pump lasers for EDFAs are required to produce power levels of 100-200 mW, and pump lasers for Raman amplifiers may reach a few watts. To reach these power levels, multiple semiconductor lasers can be combined using a combination of wavelength and polarization multiplexing techniques [250].

■ A.6 Detectors

An optical receiver consists of a photodetector and electronics for amplifying and processing the signal. In this subsection, we discuss the photodetector, which converts incident optical power to a proportional electrical current, a photocurrent, using direct detection. The photodetectors used in optical transmission systems are semiconduc-

tor photodiodes, which in the simplest case, consists of a reverse-biased *pn*-junction. A photon incident on the semiconductor can be absorbed by an electron in the valence band, elevating the electron to the conduction band. Each electron-hole pair produced this way in the depletion region contributes to the photocurrent. For absorption to occur, the photon energy must be at least as great as the semiconductor's bandgap energy. Thus, photon absorption only occurs for wavelengths smaller than a characteristic wavelength known as the cutoff wavelength. Examples of materials with sufficiently large cutoff wavelengths to allow absorption and conversion over the entire spectrum of interest in optical fiber communication systems are InGaAs and InGaAsP.

The fraction of energy of an optical signal that is absorbed and gives rise to a photocurrent is called the efficiency of the photodetector. For transmission at high bit rates over long distances, it is especially desirable that the photodetector possess an efficiency close to 1 since optical power levels are low. The absorption coefficient of semiconductor material currently in use is high enough that a slab of material on the order of 10 μm thick is sufficient for an efficiency of nearly 1.

■ A.6.1 pin photodiodes

The *pin* photodiode is a modification of the aforementioned photodiode in which a lightly doped intrinsic semiconductor is introduced between the *p*-type and *n*-type semiconductors. The purpose of this construction is to extend the depletion region so that a larger fraction of power falls on the depletion region, thereby enabling higher efficiencies. In the previous *pn* photodiode, minority carriers in the *p*-type and *n*-type regions diffuse slowly into the depletion region, thereby distorting the photodetector response.

■ A.6.2 Avalanche photodiode

The photocurrent generated by a *pn* or a *pin* photodiode is limited by the fact that each incident photon produces at most one electron of photocurrent. Such a limitation is a serious issue when the incident optical power is very low, as thermal noise from the circuitry following the photodiode may swamp the signal current. The avalanche photodiode (APD) overcomes this limitation by subjecting generated electrons to a high electric field, which endows them with sufficient energy to induce a chain reaction in which additional electrons are produced through collisions of previous electrons with the semiconductor material. This process is called avalanche multiplication. The APD, while effectively overcoming thermal noise, introduces its own noise due to the inherent randomness of the electron multiplication process. Thus, there is a trade-off between the multiplicative gain of the APD and the noise introduced by this multiplicative gain.

■ A.7 Amplifiers

Optical amplifiers play a pivotal role in rendering WDM a practical networking technology, as they enable the simultaneous amplification of many wavelengths in a transparent fashion. We next discuss the different types of optical amplifiers [14, 250].

■ A.7.1 Erbium-doped fiber amplifiers

An EDFA consists of a length of silica fiber whose core is doped with ionized erbium atoms [28, 79]. The fiber is pumped using a pump signal from a laser, typically at a wavelength of 980 nm or 1480 nm. The optical pumping process requires the use of three energy levels. The top energy level to which the electron is elevated must lie above the desired lasing level. After reaching its excited state, the electron must release some of its energy and drop to the desired lasing level. From this level a signal photon triggers the excited electron into stimulated emission, resulting in the release of a new photon with an identical wavelength as the signal photon.

The following properties render the EDFA the most popular amplifier choice for current optical communication systems: (a) the availability of compact and reliable high-power semiconductor pump lasers, (b) it is an all-fiber device, making it polarization independent and easy to couple light in and out of it, (c) the simplicity of the device, and (d) it introduces no crosstalk when amplifying WDM signals.

■ A.7.2 Raman amplifiers

Raman amplifiers exploit SRS. Raman amplification relies on simply pumping the same silica fiber used for transmitting the data signals [161, 219, 220]. This can be achieved in two ways. In a lumped or discrete amplifier, the amplifier comprises a sufficiently long spool of fiber along with the appropriate pump lasers in a package. In a distributed amplifier, the fiber can simply be the fiber span of interest, with the pump attached to one end of the span.

Unlike EDFAs, Raman amplifiers can be used to provide gain at any wavelength. Also, multiple pumps at different wavelengths and different powers can be used to tailor the overall gain spectrum. The biggest challenge in realizing Raman amplifiers is building the required high-power pump sources at the correct wavelengths. Unlike EDFAs, Raman amplifiers respond almost instantaneously to pump power. Therefore, it is important to keep the pump power constant. Otherwise, the gain varies in time and the fluctuations in pump power appear as crosstalk. Another major concern with Raman amplifiers is the crosstalk among the WDM signals. A modulated signal at a particular wavelength depletes the pump power, effectively imposing the same modulation on the pump signal, which in turn affects the gain seen by the next wavelength. These last two sources of noise in Raman amplifiers can be mitigated significantly by employing a counter-propagating pump geometry which effectively averages out the gain variations.

■ A.7.3 Semiconductor optical amplifiers

A SOA is based upon the same technology as a semiconductor laser. An SOA is essentially a forward biased *pn*-junction. Light is amplified through stimulated emission when it propagates through the active region of the junction [229]. The two ends of the active region are also given an antireflection coating to eliminate ripples in the amplifier gain as a function of wavelength.

In practice, bandwidths on the order of 100 nm can be achieved with SOAs, which is much larger than what is achievable with EDFAs. In fact, signals in the 1.3 and 1.5 μm bands can be simultaneously amplified in SOAs. In spite of this, EDFAs are preferred to SOAs mainly because SOAs introduce severe crosstalk when used in WDM systems. Furthermore, the gains and output powers achievable with EDFAs are higher, and the coupling losses and PDL are lower with EDFAs since the amplifier is a fiber. Finally, SOAs require very high quality anti-reflective coatings on their facets, which can be difficult to fabricate.

■ A.8 Switches

Optical switches may be used in optical networks for a variety of applications, such as provisioning of lightpaths, protection switching, and switching of packets in packet switched networks. Cost-efficient optical switching is expected to usher in the next generation of optical networking by enabling the deployment of the OXC and WSS in the wide- and metro-area environments. An OXC provides the same functionality for optical networks that an electronic digital cross-connect (DCS) provides for telephone networks. Specifically, OXCs enable dynamic wavelength routing for reconfiguration of optical networks while maintaining network transparency. A WSS is a special case of an OXC in that it has one input (output) port and multiple output (input) ports.

An attractive candidate OXC architecture is illustrated in Figure A-2 [250]. This design uses wavelength switching planes in order to simplify the switch fabric from an unnecessarily complex fabric that allows every wavelength input to be connected to every wavelength output. In this design, a collection of signals entering over a fiber is first demultiplexed by an OLT. All signals at a given wavelength are sent to a switch dedicated to that wavelength, and the signals from the outputs of the switches are multiplexed back together by the OLTs. In addition, the OXC allows a maximum number of wavelengths to be dropped, which requires an additional optical switch to be connected to the wavelength plane switches.

For optical switches with large port counts, a good switch architecture will minimize the number of switch elements used, possess uniform loss across input-output port combinations, minimize the number of crossovers[5], and be nonblocking. Note that there are different degrees to which switches can be nonblocking. Wide-sense

[5]In integrated optics, unlike integrated electronic circuits, connections should be made in a single layer because if the paths of two waveguides cross, power loss and crosstalk occur. Note that this is not an issue in free-space switches, such as MEMS.

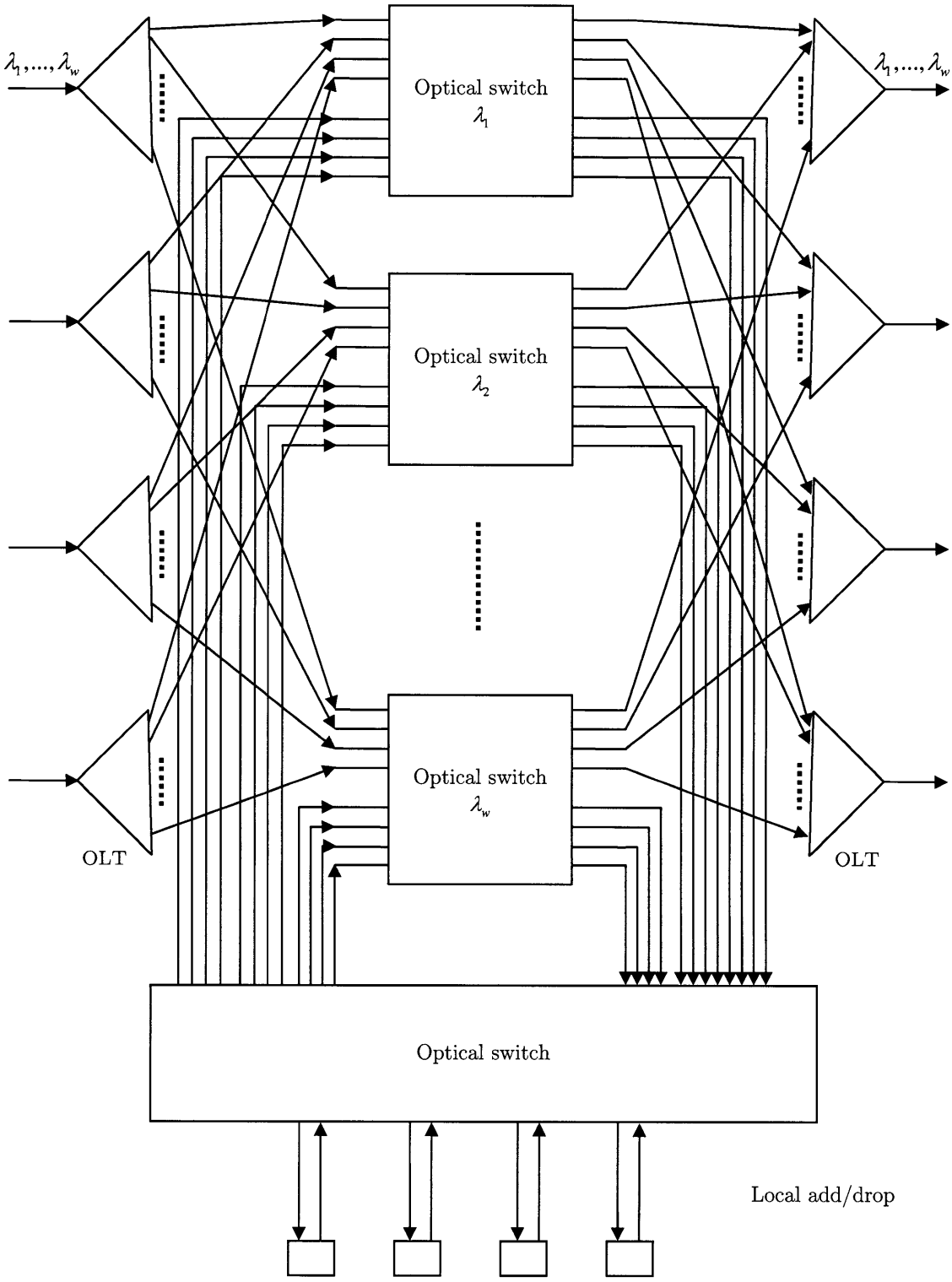


Figure A-2. Candidate OXC architecture. This switch has four fiber inputs and four fiber outputs, and each fiber has w wavelengths. In addition, any four wavelengths may be dropped/added. Adapted from [250].

nonblocking switches can connect any unused input to any unused output without requiring any existing connection to be rerouted. They achieve this property by means of predefined routings through the switch. Strict-sense nonblocking switches allow any unused input to be connected to any unused output regardless of how existing connections were routed through the switch. Rearrangeably nonblocking switches may require rerouting of connections to achieve the nonblocking property.

■ A.8.1 Switch architectures

Large optical switches may be architected in several ways, as discussed in the following [207, 272].

Crossbar

The crossbar architecture is wide-sense nonblocking and employs a matrix array of 2×2 switches. To connect input i to output j, the path taken traverses the 2×2 switches in row i until column j is reached, and then traverses the switches in column j until output j is reached. In general, an $n \times n$ crossbar requires n^2 switches. The shortest path length is 1 and the longest path length is $2n-1$, resulting in undesirable nonuniform loss.

Clos

The Clos architecture is strict-sense nonblocking. To construct an $n \times n$ switch, we employ three parameters m, k, and p, and set $n = mk$. The first and third stages consist of k $m \times p$ switches. The middle stage consists of p $k \times k$ switches. Each of the k switches in the first stage is connected to all the switches in the middle stage. Likewise, each of the k switches in the third stage is connected to all the switches in the middle stage. If $p \geq 2m - 1$, then the switch is strict-sense nonblocking. Assuming that the individual switches in each stage are crossbars, the minimum number of switch elements required is approximately $4\sqrt{2}n^{3/2} - 4n$, which is significantly lower than the n^2 switch elements required for a crossbar. Furthermore, the Clos architecture has more uniform loss than the crossbar.

Spanke

The Spanke architecture is strict-sense nonblocking. An $n \times n$ switch is constructed by combining n $1 \times n$ switches along with n $n \times 1$ switches. The Spanke architecture is attractive when a $1 \times n$ optical switch can be built using a single switch element that does not need to be built out of 1×2 or 2×2 switch elements. This is the case for 3D MEMS, as we shall see shortly, where $2n$ such switch elements are needed to build an $n \times n$ switch. Thus, the switch cost scales linearly with n, which is significantly better than other architectures. In addition, each connection passes through two switch elements, which provides low and uniform loss regardless of input-output combination.

Type	Size	Loss	Crosstalk	PDL	Switching time
Bulk mechanical	8×8	3 dB	55 dB	0.2 dB	10 ms
2D MEMS	32×32	5 dB	55 dB	0.2 dB	10 ms
3D MEMS	$10^3 \times 10^3$	5 dB	55 dB	0.5 dB	10 ms
Thermo-optic silica	8×8	8 dB	40 dB	Low	3 ms
Bubble-based	32×32	7.5 dB	50 dB	0.3 dB	10 ms
Liquid crystal	2×2	1 dB	35 dB	0.1 dB	4 ms
Polymer	8×8	10 dB	30 dB	Low	2 ms
Electro-optic $LiNbO_3$	4×4	8 dB	35 dB	1 dB	10 ps
SOA	4×4	0 dB	40 dB	Low	1 ns

Table A.2. Comparison of optical switching technologies. Adapted from [250].

Beneš

The Beneš architecture is rearrangeably nonblocking, and is the most efficient switch architecture in terms of the number of 2×2 switches it uses. In general, an $n \times n$ Beneš switch requires $(n/2)(2\log_2 n - 1)$ 2×2 switches, where n is a power of 2. The loss is the same through every path in the switch, as each path passes through $2\log_2 n - 1$ 2×2 switches. Its two main drawbacks are that it is not wide-sense nonblocking, and that a number of waveguide crossovers are required, making it difficult to fabricate with integrated optics.

■ A.8.2 Switch technologies

Several different technologies exist to build optical switches, which are compared in Table A.2 and discussed in greater detail below [207, 250]. With the exception of the 3D MEMS switch, the switches employ the crossbar architecture.

Bulk mechanical switches

Bulk mechanical switches carry out switching by mechanical means, such as moving a mirror in and out of the optical path, or bending or stretching fiber in the interaction region of a coupler to change the coupling ratio [163, chapter 13]. This family of switches has low insertion losses, low PDL, low crosstalk, and is inexpensive to build. Owing to millisecond switching times and low port counts, these switches are suited to small wavelength cross-connects or protection switching. As with most mechanical components, long-term reliability is an issue.

Micro-electro-mechanical systems

MEMS switches are mechanical devices employing miniature movable mirrors with dimensions ranging from a few hundred micrometers to a few millimeters, and are typically fabricated using silicon substrates. These mirrors are deflected from one position to another using a variety of electronic actuation techniques [181, 190, 222, 224, 256].

The simplest MEMS designs correspond to 2D MEMS, where the mirrors, arranged in a crossbar, are each in one of two states. In one state, the mirror is flat in line with the substrate and the light beam is not deflected. In the other state, the mirror is in a vertical position and the light beam, if present, is deflected. While 2D MEMS devices are reliable, they do not scale well with the number of ports.

In more complex MEMS implementations, known as 3D MEMS, each mirror may rotate freely on two distinct axes. Each mirror is controlled in an analog fashion in order to achieve a continuous range of angular deflections. A mirror of this type can be used to implement a $1 \times n$ switch (i.e., a WSS), and an $n \times n$ switch may thus be realized by employing $2n$ mirrors in a Spanke architecture. While very scalable with the number of ports, 3D MEMS suffer from poor reliability, owing to rapid ageing and sensitivity to mechanical disturbances.

In [122], Guan and Chan show that the crossover point for cost-efficiency of 3D versus 2D MEMS occurs at a port count of 32 or 64.

Bubble-based waveguide switches

Bubble-based waveguide switches employ planar waveguides where the switch actuation is based upon technology similar to that of ink-jet printers [103]. Light travels along one of a series of intersecting waveguides. Under normal conditions, light continues along the same waveguide at the crossover points. In order to deflect light into an intersecting waveguide, fluid at a crossover point is heated creating an air bubble. This air bubble then reflects the light at the crossover point into an intersecting waveguide. This technology may realize relatively low-cost, easily manufacturable small switch arrays with switching times on the order of tens of milliseconds.

Liquid crystal switches

Liquid crystal cells provide another way for building small optical switches, such as WSSs [146, 231, 318]. These switches employ polarization effects to perform switching. By applying a voltage to a suitably designed liquid crystal cell, the polarization of the light passing through the cell can be rotated or not. This can then be combined with passive polarization beam splitters and combiners to yield a polarization-independent switch. Switching times are on the order of a few milliseconds. Like bubble-based waveguide switches, liquid crystal switches are solid-state devices and can potentially be manufactured in volume at low cost.

Electro-optic switches

Electro-optic switches make use of the electro-optic effect, where an applied voltage induces a change in the refractive index of a material. A 2×2 electro-optic switch can be realized by employing lithium niobate in the coupling region of an MZI. When a voltage is applied to this region, thereby changing its refractive index, the proportion of power coupled from the input waveguides to the output waveguides is controlled. An electro-optic switch is capable of quickly changing its state in less than 1 ns. This switching time is determined by the capacitance of the electrode configuration. In addition to fast switching times, lithium niobate switches allow modest levels of integration. However, they tend to have high loss and PDL, and are more expensive to build than mechanical switches.

Thermo-optic switches

Thermo-optic switches are 2×2 integrated-optic MZIs, constructed on waveguide material whose refractive index is a function of temperature. These devices have been made on silica as well as polymer substrates, and have poor crosstalk. Also, the thermo-optic effect is quite slow, resulting in switching speeds on the order of a few milliseconds.

Semiconductor optical amplifier switches

SOA switches employ SOAs which can be used as on-off switches by varying the voltage to the device. If the bias voltage is reduced, no population inversion is achieved, and the device absorbs input signals. If the bias voltage is present, the input signal is amplified. Switching speed is on the order of 1 ns for such devices.

■ A.9 Wavelength converters

A wavelength converter (WC) converts data from an incoming wavelength to a different outgoing wavelength. WCs are important components in WDM networks for two reasons: (a) data may enter the network at a wavelength that is not suitable for use within that network, and (b) WCs may be needed within the network to improve utilization of the available wavelengths on the network links.

WCs may be classified on the range of wavelengths that they can handle at their inputs and outputs. The input and output of a WC may be either fixed or variable wavelength, leading to four types of WCs. Critical properties of WCs beyond the nature of the inputs and outputs are the range of input optical powers that the WC can handle, whether the WC is transparent to the bit rate and modulation format of the input signals, and whether it introduces additional noise or phase jitter to the signal.

There are five fundamental ways of achieving wavelength conversion [22, 64, 76, 93, 97, 150, 223, 276, 321, 322, 326]: (a) optoelectronic, (b) optical gating, (c) interfero-

metric, (d) wave mixing, and (e) nonlinear parametric amplification. The latter four approaches are all-optical but are not yet mature for commercial use.

■ A.9.1 Optoelectronic

Currently, commercial wavelength conversion is achieved via optoelectronics, where an optical receiver first converts the optical signal at the input wavelength into an electric current, and this current is then used to modulate an optical transmitter at the desired output wavelength. Optoelectronic wavelength conversion has limited transparency to bit rate and data format, is limited in speed by electronics, and is relatively expensive to implement. The optoelectronic approach is usually variable input, since, as discussed previously, a single photodetector can be used for a wide range of input wavelengths. The output is either fixed or variable, depending on whether a fixed or tunable laser is used, respectively.

The quality and transparency of the optoelectronic conversion depend on which of the three possible forms of regeneration is used. In 1R regeneration, which is completely transparent to modulation format, the receiver converts incoming photons to electrons, which are then amplified by an analog radio-frequency (RF) amplifier before driving the laser. 1R regeneration is compatible with analog data, but noise is added to the signal and nonlinear effects and dispersion are not mitigated. In 2R regeneration, the signal is regenerated as in 1R regeneration, and is then passed through a logic gate for reshaping. This form of regeneration suffers from additional phase jitter. Finally, in 3R regeneration, the data is regenerated, reshaped and re-timed, which completely eliminates the deleterious effects of nonlinearities, noise, and dispersion [64]. Since re-timing is a bit-rate specific operation, transparency, however, is lost.

■ A.9.2 Optical gating

Optical gating is a technique in which changes in intensity of an input optical signal are transferred to an unmodulated probe signal through an intermediate optical device. In the context of wavelength conversion, the original signal powers an SOA, whose gain, owing to the nonlinear cross-gain modulation (CGM) phenomenon, is dependent upon the input signal power [93,223]. Specifically, as the input signal power increases, carriers in the gain region of the SOA are depleted, resulting in smaller amplifier gain. Thus, if a low-power probe signal at a different wavelength is sent into the SOA, it experiences low gain when there is a 1 bit in the input signal, and a higher gain when there is a 0 bit [250]. Because the carrier dynamics within the SOA occur on a picosecond time scale, the gain is able to respond to bitwise fluctuations of the input signal provided the bit rate is below 10 Gb/s. A shortcoming of this approach to wavelength conversion is that the achievable extinction ratio is small since the gain does not drop to zero when there is an input 1 bit. Furthermore, variations

in the carrier density of the SOA change the refractive index, which in turn changes the phase of the probe, creating pulse distortion.

■ A.9.3 Interferometric techniques

One approach to wavelength conversion employing interferometric techniques uses an MZI in which each arm is fitted with an SOA, and with an asymmetric coupling ratio. The signal is sent in one arm and the probe in the other. If no signal is present, then the probe signal comes out unmodulated. When the signal is present, it induces a phase change in each amplifier by CPM. The phase change induced by each amplifier on the probe is different because different amounts of signal power are present in the two amplifiers, owing to the asymmetric coupling ratio. The MZI translates this relative phase difference between the two arms on the probe into an intensity-modulated signal at the output.

This approach requires much less signal power to achieve a large phase shift compared to a large gain shift in optical gating. This method also produces a better extinction ratio because the phase change can be converted into a digital amplitude-modulated output signal by the interferometer. Like optical gating, the bit rate that can be handled is at most 10 Gb/s. However, unlike optical gating, precise control of the SOA bias current is required because the phase of signals passing through the device are sensitive to changes in this current [250].

Other interferometric techniques for wavelength conversion can be found in [64, 125].

■ A.9.4 Wave mixing

The wave mixing approach to wavelength conversion exploits the FWM phenomenon. If a signal at frequency f_s and a probe at frequency f_p are mixed, then signals at frequencies $2f_p - f_s$ and $2f_s - f_p$ are produced by FWM. The four-wave mixing power can be enhanced by using an SOA, provided that the mixed signals fall within the amplifier bandwidth [223]. The main advantage of wave mixing is that it is transparent to modulation format and bit rate, since both amplitude and phase are preserved during the mixing process [250]. The disadvantages of this approach are that the unwanted waves must be filtered out of the SOA output, and the conversion efficiency's dependence on the wavelength separation between the signal and probe [93].

■ A.9.5 Nonlinear parametric amplification

In nonlinear parametric amplification, second harmonic generation (SHG), sum frequency generation (SFG), and difference frequency generation (DFG) are used to exchange power among electromagnetic waves of different frequencies [324]. Efficient conversion of power among waves, however, only occurs if the stringent so-called phase matching conditions are met among the waves and the medium in which they

interact. The device often employed to satisfy the phase matching conditions, and hence enable wavelength and waveband conversion [321,322], is a periodically-poled lithium niobate crystal. In this ferroelectric device, regions of periodically reversed spontaneous polarization are created which reset the relative phase between interacting waves, such that, on average, the proper phase relationship among waves of certain frequencies is maintained.

■ A.10 Buffers

In the context of switched optical networks, there are at least three potential reasons for storing or buffering a packet at a switch before it is forwarded on one of its outgoing links [250]. First, the incoming packet must be buffered while its header is processed to determine how the data must be routed. This is usually a fixed delay that can be implemented simply. Second, the required switch input and/or output port may not be free, requiring the packet to be queued at an input buffer. The switch input may not be free because other packets that arrived on the same link have to be served earlier. The switch output port may not be free because data from other input ports are being switched to it. Third, after the packet has been switched to the required output port, the outgoing link from this port may be busy transmitting other data, thus requiring that the packet wait for its turn. The latter delays are variable and are implemented differently from the fixed delay required for header processing.

The lack of good buffering methods in the optical domain is a major impediment. Unlike in the electronic domain, RAM in the optical domain does not exist. Until recently, the fiber delay line (FDL) was the only proposed means of achieving optical buffering. In the last few years, however, new approaches have employed 'slow' light, where the group velocity of a light signal is substantially reduced. These technologies include waveguide resonators, ring resonators, photonic crystals, electromagnetically induced transparency, and coherent population oscillation.

■ A.10.1 Fiber delay lines

The primary method of realizing optical buffers is to use FDLs, which consist of relatively long lengths of fiber. For example, about 200 m of fiber is required for a 1 ms delay, which would be sufficient to store 10 packets, each with 1000 bits at 10 Gb/s. Very small buffers are thus usually used in optical switched networks. Note that unlike in electronic buffers, stored data in FDLs cannot be accessed at an arbitrary point of time. The data can exit the buffer only after a fixed time interval after entering it. This is the time taken for the packet to traverse the fiber length. Of course, by repeated traversals of the same piece of fiber, delays that are multiples of this basic delay can be obtained.

Early FDL buffer architectures were categorized as being either feed-forward (FF) or feed-back (FB), as well as being either single-stage or multi-stage [148]. In FF buffers data is delayed while forwarded toward the output of the node [108, 149],

whereas in FB buffers data is delayed while being fed back to an earlier stage of the node [29]. In single-stage buffers the delay is realized by a set of fixed-length FDLs [29, 108], while in multi-stage buffers the delay is determined by a cascade of FDL and switch pairs [65, 149]. The capacity of FDL buffers can be increased by using WDM in the FDLs. In [45], the problem of dimensioning FDLs in asynchronous networks with variable length packets is addressed.

Another taxonomy of FDL buffers considers whether they are recirculating or traveling. In recirculating buffers, the delay is determined by the loop length and the circulating number. Therefore, the delay time can be varied easily. However, optical amplifiers have to be installed in the loop to compensate for the circulating losses. Eventually, ASE noise accumulates and causes signal degradation. Moreover, this type of buffer cannot handle packets longer than the loop length. By contrast, in traveling buffers, the delay time is determined by the length of the delay line. The drawback of this is that the use of many delay lines is required to avoid losing packets when the traffic load is heavy. Consequently, this type of buffer is bulky.

Recently, recirculating fiber loops in conjunction with wavelength converters were used to create optical buffers capable of handling variable length packets [192, 257, 258]. One such buffering system comprises a wavelength converter, a fiber loop, an optical coupler with two arms connected to form the optical loop, a wavelength shifter, and an optical drop module [258]. In the wavelength converter, the wavelength of each input packet is converted to an initial wavelength that controls the delay time. The wavelength-converted optical packet is fed into the optical loop via the coupler. The wavelength shifter located in the optical loop gives optical packets a uniform wavelength shift per circulation. Therefore, the initial wavelength of the optical signal shifts sequentially. When the wavelength reaches the output wavelength, the optical signal exits from the optical drop module. Consequently, the circulating number, and therefore the delay time, is controlled directly by the initial wavelength.

In addition to the aforementioned structural parameters of FDL buffers, the scheduling algorithm used to arbitrate access to the buffers in a network setting is an important consideration that influences performance [158, 189].

Beyond their most significant shortcoming of not permitting random access to optical data, FDLs are often plagued by bulkiness, manufacturing inaccuracies, and sensitivity to pressure, vibration and temperature.

■ A.10.2 Waveguide resonator

In the waveguide resonator approach [102, 186], buffering of light is achieved by a structure consisting of cascaded chirped grating and phase-shifter sections, providing two optical cavities which generate a wide range of potential filter responses. The active waveguide nature of the filter allows current-controlled tunability of both the spectral response and the corresponding group delay. Unfortunately, using this approach maximum delays on the order of tens of picoseconds are achievable.

■ A.10.3 Ring resonator

The ring resonator approach is based upon inducing large dispersive effects in optical waveguides by coupling the waveguide to an array of optical ring resonators [141]. Since the light field effectively circulates many times in each resonator before passing to the next, the group velocity of propagation of a pulse of light through such a structure is significantly reduced. Like the waveguide resonator approach, maximum delays on the order of tens of picoseconds are achievable. The ring resonator approach is further plagued by fabrication challenges.

■ A.10.4 Photonic crystal

The photonic crystal approach employs a new type of waveguide, known as a coupled cavity waveguide (CCW) [180]. CCWs are impurity band-based waveguides in which the guiding of electromagnetic waves is provided by the coupling of localized defect modes. It is demonstrated that this kind of impurity band can perfectly transmit ultrashort pulses with small group velocity. Unfortunately, only sub-picosecond delays have been achieved for CCWs tens of micrometers long. Furthermore, the fabrication of these devices presents a challenge.

■ A.10.5 Electromagnetically induced transparency

The electromagnetically induced transparency (EIT) phenomenon is a quantum interference effect which acts to reduce the usual absorption a light signal experiences when its frequency is tuned to the resonance frequency of the sample through which it is propagating. The transparency is created by a second light source tuned to another resonance of the sample. The coupling of the second light source induces a large slope in the dielectric function spectrum experienced by the signal, resulting in a slower group velocity for the signal.

Devices which employ EIT have been shown to slow light down by factors as high as 10^7 in atomic vapor cells and Pr-doped Y_2SiO_5 crystals [236, 295]. However, in these devices, the transmission bandwidth of the group velocity is tens of kHz, which is not suitable for optical communication applications. Recently, a semiconductor optical buffer using EIT in a quantum dot medium for application to 40 Gbps optical communication systems was demonstrated. Unfortunately, the factor by which it was able to slow light down was approximately 10^3 [168, 175]. Drawbacks of the aforementioned EIT approaches are that they do not offer tunability via current-control, and that very low temperatures are required for large slowing factors[6].

[6]At room temperature, light can be slowed down by a factor of 70 in a uniform InAs/GaAs quantum dot array.

■ A.10.6 Coherent population oscillation

In [32], slow light propagation at a group velocity of 57.5 m/s was observed at room temperature in a ruby crystal. A quantum coherence effect, coherent population oscillation, produced a very narrow spectral hole in the homogeneously broadened absorption profile of ruby. The resulting rapid spectral variation of the refractive index led to a slower group velocity for the signal. Like EIT, coherent population oscillation does not offer tunability via current-control.

■ A.11 Logic circuits

Optical logic circuits are a key building block for optical packet switches, as current trends indicate that electronic logic circuits will be unable to keep up with optical line-rates. Despite great efforts to develop optical logic circuits, none exist in a reliable, compact and stable form.

All-optical logic gates based upon optical fibers have been studied for more than a decade. However, they are plagued with challenges such as complexity, large volume, and integration difficulty with other devices. On the contrary, all-optical logic gates using semiconductor devices are small, and easy to integrate with other devices. All-optical logic gates using SOAs show particular promise, as they provide high gain, exhibit strong refractive-index change required for nonlinearity, and allow photonic integration [72].

One approach to realizing optical logic gates makes use of photo-generated changes in the carrier density of SOAs. Optical switching using SOAs in the picosecond regime usually involves placing an SOA in an interferometer. This has been demonstrated in a variety of configurations employing the symmetric MZI, Ultrafast Nonlinear Interferometer (UNI), and Terahertz Optical Asymmetric Demultiplexer (TOAD) [96, 153, 233, 234, 270, 285]. It has further been shown that by combining switches, more sophisticated all-optical logic functions may be realized. Proof of concept has been demonstrated for optical address recognizers [43, 71, 119], optical bit-level synchronizers [44, 319], optical clock recovery circuits [159], optical shift registers with inverters [134, 244], optical memories with read-write abilities [86, 143, 242, 243], binary half- and full-adders [238, 239], pseudorandom number generators [241], optical parity checkers [240], and simple optical packet switches [144]. The processing speed of digital optical logic based upon resonant optical transitions, as described above, is restricted to approximately 200 GHz [72]. The reason for this is that SOA recovery depends on the carrier lifetime, which is approximately 1 ns.

Research has subsequently been carried out to investigate SOA gain and index dynamics at sub-picosecond timescales. Early experiments indicated that strong nonlinearities on sub-picosecond timescales in bulk and multi-quantum well (MQW) SOAs can take place [135–138, 167]. Switching experiments employing femtosecond pulses in a high-speed nonlinear interferometer were reported in [216–218]. Wavelength switching experiments employing ultrashort pulses were published in [16, 17, 178]. In [85],

nonlinear carrier dynamics in an MQW SOA in the context of high-speed all-optical logic was investigated, and an AND gate was demonstrated. In [170], an MQW SOA was used to create an optical flip-flop.

Recently, optical logic was demonstrated in SOAs by another technique which employs FWM [53].

At the moment, the largest challenge toward the implementation of individual high-speed optical logic devices based upon SOA nonlinearities is related to the high power consumption of these devices. A related challenge arises when considering the use of high-speed SOA nonlinearities in more sophisticated high-speed all-optical logic circuits. High input powers are projected, owing to power loss in individual gates and the need to cascade several gates in logic circuits.

Appendix B

(Generalized) Moore Graphs

In this appendix, we discuss a family of graphs, Moore Graphs, and their superset, Generalized Moore Graphs. These graphs are known to be attractive candidate network topologies for a variety of communications applications. The study of Moore Graphs and Generalized Moore Graphs in the context of optical networks has been carried out by Guan [121]. In this appendix, we shall restrict our discussion to undirected Moore Graphs, although an analogous discussion exists for directed Moore Graphs [121, Section 4.2].

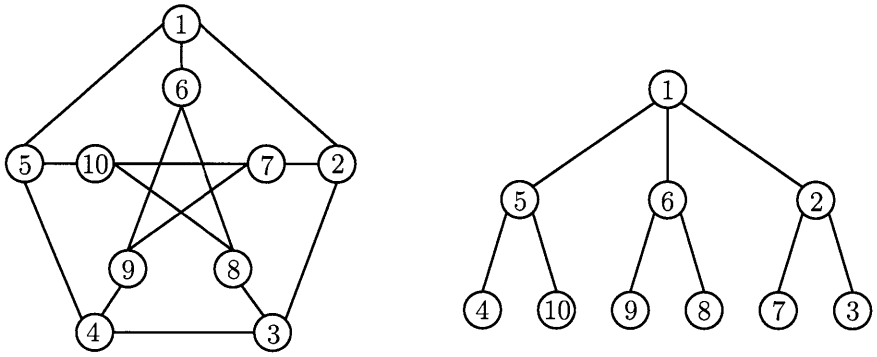
■ B.1 Moore Graphs

A Moore Graph is defined as any undirected graph which achieves the *Moore Bound*. The Moore Bound is the following upper bound on the on the number of nodes $N(\Delta, d)$ that can be supported by an undirected graph with maximum node degree Δ and diameter d [145,268]:

$$N(\Delta, d) \leq \mathscr{M}(\Delta, d) \equiv 1 + \Delta \sum_{i=0}^{d-1} (\Delta - 1)^i$$
$$= 1 + \Delta \frac{(\Delta - 1)^d - 1}{\Delta - 2}.$$

The Moore Bound may be derived by counting the number of nodes in a fully populated, regular tree of degree Δ (i.e., a tree in which all nodes, except leaf nodes, have degree Δ). For a graph to achieve the Moore Bound, a fully-populated regular spanning tree must exist when any of the graph's nodes serve as the root node (see Figure B-1). Thus, there may not exist a realizable Moore Graph for every $\mathscr{M}(\Delta, d)$. Indeed, very few Moore Graphs actually exist. The following is an exhaustive list of known Moore Graphs:

- Complete graphs of any degree.
- Ring graphs with odd number of nodes.
- Petersen Graph: the ten node, $\Delta = 3$, and $d = 2$ graph drawn in Figure B-1.



(a) Petersen Graph.

(b) Spanning tree of the Petersen Graph with root node 1.

Figure B-1. The Petersen Graph, a Moore graph with ten nodes, $\Delta = 3$, and $d = 2$.

- Hoffman-Singleton Graph: the 50 node, $\Delta = 7$, and $d = 2$ graph.

Furthermore, it has been shown that Moore Graphs with $d \geq 3$ do not exist.

The average shortest path distance for the $\mathscr{M}(\Delta, d)$ Moore Graph is given by:

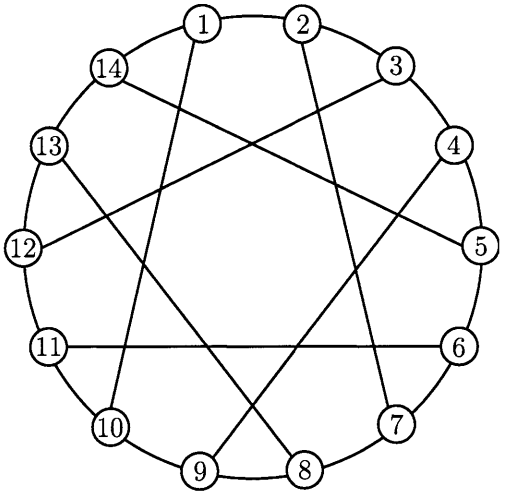
$$h_{\mathscr{M}}(\Delta, d) = \frac{d(\Delta - 1)^d}{(\Delta - 1)^d - 1} - \frac{1}{\Delta - 2}.$$

Furthermore, each node of a Moore Graph has a unique shortest path routing spanning tree, implying that each every source-destination pair has a unique, shortest path [121, Theorem 1]. Lastly, for the $\mathscr{M}(\Delta, d)$ Moore Graph, a balanced traffic load distribution can be achieved for static, uniform all-to-all traffic, with each edge having load $\sum_{i=0}^{d} (\Delta)^{i-1}$ [121, Theorem 2].

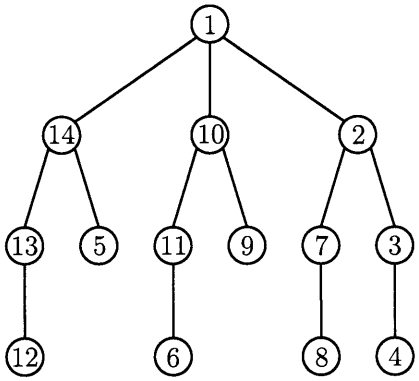
■ B.2 Generalized Moore Graphs

Generalized Moore Graphs are similar to Moore Graphs in that their spanning trees are regular and fully populated *except* possibly the last level of the spanning tree. When the last level of the graph's tree is not fully populated then the Generalized Moore Graph cannot achieve the Moore Bound. Figure B-2 depicts an example of a Generalized Moore Graph, along with one of its spanning trees.

An alternative definition provided in [260] is that a Generalized Moore Graph minimizes the average shortest path distance over all graphs with same number of nodes and maximum node degree. Moreover, since Generalized Moore Graphs are node-symmetric, each shortest path spanning tree individually attains the minimum average shortest path distance. For a Generalized Moore Graph with N nodes and



(a) Heawood Graph.



(b) Spanning tree of the Heawood Graph with root node 1.

Figure B-2. The Heawood Graph, a Generalized Moore graph with 14 nodes, $\Delta = 3$, and $d = 3$.

maximum degree Δ, the average shortest path distance is given by:

$$h_{\mathcal{G}}(N, \Delta) = \frac{\Delta\left[1 - (\Delta - 1)^d\right] + dN(\Delta - 2)^2 + 2d(\Delta - 2)}{(N - 1)(\Delta - 2)^2},$$

where d is the graph diameter, and is given by:

$$d = \left\lceil \log_{\Delta - 1} \frac{N(\Delta - 2) + 2}{\Delta} \right\rceil - 1.$$

In contrast to Moore Graphs, Generalized Moore Graphs do not necessarily have unique shortest path spanning trees, and therefore may not necessarily balance traffic load on graph edges via shortest path routing. Generalized Moore Graphs, however, are more attractive than Moore Graphs in the sense they represent a richer, more densely populated family of graphs.

Bibliography

[1] "Gigabit Ethernet: Accelerating the standard for speed," Internet: http://www.gigabit-ethernet.org, Gigabit Ethernet Alliance, Tech. Rep., 1998.

[2] "Heavy-tailed distributions, generalized source coding and optimal web layout design," California Institute of Technology, Control & Dynamical Systems 107-81, Pasadena, CA 91125, Tech. Rep. CIT CDS 00-001, 2000.

[3] "Optical switching networks: from circuits to packets," *IEEE Communcations Magazine*, vol. 39, no. 3, 2001.

[4] "Special issue on Ethernet transport over wide area networks," *IEEE Communcations Magazine*, vol. 42, no. 2, 2004.

[5] "Annual estimates of the population for incorporated places over 100,000, ranked by July 1, 2006 population: April 1, 2000 to July 1, 2006," Jun. 2007. [Online]. Available: http://www.census.gov/popest/cities/tables/SUB-EST2006-01.csv

[6] "The exabyte era," 2007. [Online]. Available: http://www.cisco.com/

[7] "10Gb/s Ethernet Passive Optical Network (10GEPON)," Jan. 2008. [Online]. Available: http://www.ieee802.org/3/av/index.html

[8] "Akamai's technology overview," May 2008. [Online]. Available: http://www.akamai.com/html/technology/index.html

[9] "IEEE 802.17 Resilient Packet Ring Working Group," Jan. 2008. [Online]. Available: http://www.ieee802.org/17/

[10] "IEEE P802.3ah Ethernet in the First Mile Task Force," Jan. 2008. [Online]. Available: http://www.ieee802.org/3/ah/

[11] "Internet world statistics," Jan. 2008. [Online]. Available: http://www.internetworldstats.com/

[12] "The Telmarc Group," Jun. 2008. [Online]. Available: http://www.telmarc.com

[13] G. P. Agrawal, *Nonlinear Fiber Optics*, 2nd ed. San Diego, USA: Academic Press, 1995.

[14] ——, *Fiber-Optic Communication Systems*, 2nd ed. Wiley-Interscience, 1997.

[15] B. Ahn and Y. Park, "A symmetric-structure CDMA-PON system and its implementation," *IEEE Photonics Technology Letters*, vol. 14, no. 8, pp. 1139–1141, 2002.

[16] T. Akiyama *et al.*, "Nonlinear processes responsible for nondegenerate four-wave mixing in quantum-dot optical amplifier," *AIP Applied Physics Letters*, vol. 74, no. 12, pp. 1753–1755, 2000.

[17] T. Akiyama, H. Kuwatsuka, N. Hatori, Y. Nakata, H. Ebe, and M. Sugawara, "Symmetric highly efficient (0 dB) wavelength conversion based on four-wave-mixing in quantum dot optical amplifiers," *IEEE Photonics Technology Letters*, vol. 14, no. 8, pp. 1139–1141, 2002.

[18] S. B. Alexander *et al.*, "A precompetitive consortium on wide-band all-optical networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 11, no. 5, pp. 714–735, 1993.

[19] E. Altman and A. A. Borovkov, "On the stability of retrial queues," *Queueing Systems*, vol. 26, no. 3, pp. 343–363, 1997.

[20] M. Andrews *et al.*, "Scheduling in a queueing system with asynchronously varying service rates," *Probability in the Engineering and Informational Sciences*, vol. 18, no. 2, pp. 191–217, 2004.

[21] M. Andrews and L. Zhang, "Achieving stability in networks of input-queued switches," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 848–857, 2003.

[22] M. Asobe, O. Tadanaga, H. Miyazawa, and H. Suzuki, "Polarization independent wavelength converter in a periodically poled Zn:$LiNbO_3$ ridge waveguide," *Proceedings of IEEE Conference on Lasers and Electro-Optics (CLEO)*, pp. 484–484, 2002.

[23] P. Bak, *How Nature Works: The Science of Self-Organized Criticality.* New York, USA: Copernicus, 1996.

[24] R. A. Barry *et al.*, "All-Optical Network consortium – ultrafast TDM networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 5, pp. 1014–1029, 1996.

[25] R. Batchellor and O. Gerstel, "Cost effective architectures for core transport networks," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, 2006.

[26] T. Battestilli and H. Perros, "An introduction to Optical Burst Switching," *IEEE Optical Communications*, vol. 41, no. 8, pp. S10–S15, 2003.

[27] P. Bayvel and M. Dueser, "Optical Burst Switching: Research and applications," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, 2004.

[28] P. C. Becker, N. A. Olsson, and J. R. Simpson, *Erbium-Doped Fiber Amplifiers: Fundamentals and Technology.* San Diego, USA: Academic Press, 1999.

[29] G. Bendelli, M. Burzio, P. Gambini, and M. Puleo, "Performance assessment of a photonic atm switch based on a wavelength-controlled fiber loop buffer," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, pp. 106–107, 1996.

[30] I. Bennion *et al.*, "UV-written in-fibre Bragg gratings," *Optical and Quantum Electronics*, vol. 28, pp. 93–135, 1996.

[31] J. Berthold, A. A. M. Saleh, L. Blair, and J. M. Simmons, "Optical networking: Past, present, and future," *IEEE/OSA Journal of Lightwave Technology*, vol. 26, no. 9, pp. 1104–1118, 2008.

[32] M. S. Bigelow, N. N. Lepeshkin, and R. W. Boyd, "Observation of ultraslow light propagation in a ruby crystal at room temperature," *APS Physical Review Letters*, vol. 90, no. 11, p. 113903, 2003.

[33] G. Birkhoff, "Tres observaciones sobre el algebra lineal," *Universidad Nacional de Tucuman Revista, Serie A*, vol. 5, pp. 147–151, 1946.

[34] S. Bjornstad, M. Nord, and D. R. Hjelme, "QoS differentiation and header/payload separation in optical packet swtiching using polarization multiplexing," *Proceedings of European Conference on Optical Communication (ECOC)*, pp. 28–29, 2003.

[35] C. Bock, J. Prat, and S. D. Walker, "Hybrid WDM/TDM PON using the AWG FSR and featuring centralized light generation and dynamic bandwidth allocation," *IEEE/OSA Journal of Lightwave Technology*, vol. 23, no. 12, pp. 3981–3988, 2005.

[36] M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Diffraction and Interference of Light.* Cambridge University Press, 1999.

[37] O. J. Boxama, J. W. Cohen, and N. Huffels, "Approximations of the mean waiting time in an $M/G/s$ queueing system," *Operations Research*, vol. 27, no. 6, pp. 1115–1127, 1979.

[38] R. Breyer and S. Riley, *Switched, Fast, and Gigabit Ethernet*, 3rd ed. New Riders Publishing, 1998.

[39] F. Brichet, J. Roberts, A. Simonian, and D. Veitch, "Heavy traffic analysis of a storage model with long range dependent on/off sources," *Queueing Systems*, vol. 23, pp. 197–215, 1996.

[40] A. Brzezinski, "Scheduling algorithms for throughput maximization in data networks," Ph.D. Dissertation, Massachusetts Institute of Technology, 2007.

[41] A. Brzezinski and E. Modiano, "Dynamic reconfiguration and routing algorithms for IP-over-WDM networks with stochastic traffic," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, vol. 1, pp. 432–443, 2005.

[42] J. Cai, A. Fumagalli, and I. Chlamtac, "The Multitoken Interarrival Time protocol for supporting variable size packets over WDM ring network," vol. 18, no. 10, pp. 2094–2104, 2000.

[43] N. Calabretta, H. de Waardt, G. D. Khoe, and H. J. S. Dorren, "Ultrafast asynchronous multioutput all-optical header processor," *IEEE Photonics Technology Letters*, vol. 16, no. 4, pp. 1182–1184, 2004.

[44] N. Calabretta, Y. Liu, F. M. Huijskens, M. T. Hill, H. de Waardt, G. D. Khoe, and H. J. S. Dorren, "Optical signal processing based on self-induced polarization rotation in a semiconductor optical amplifier," *IEEE/OSA Journal of Lightwave Technology*, vol. 22, no. 2, pp. 372–381, 2004.

[45] F. Callegati, "Optical buffers for variable length packets," *IEEE Communications Letters*, vol. 4, no. 9, pp. 292–294, 2000.

[46] R. J. Campbell and R. Kashyap, "The properties and applications of photosensitive germanosilicate fibre," *International Journal of Optoelectronics*, vol. 9, no. 1, pp. 33–57, 1994.

[47] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "On the nonstationarity of Internet traffic," in *SIGMETRICS/Performance*, 2001, pp. 102–112.

[48] ——, "The effect of statistical multiplexing on the long-range dependence of Internet packet traffic," Bell Laboratories, Tech. Rep., 2002. [Online]. Available: http://cm.bell-labs.com/stat/doc/multiplex.pdf

[49] C. Caramanis, M. Rosenblum, M. X. Goemans, and V. Tarokh, "Scheduling algorithms for providing flexible, rate-based, quality of service guarantees for packet-switching in Banyan networks," *Proceedings of Conference on Information Sciences and Systems (CISS)*, pp. 160–166, 2004.

[50] J. M. Carlson and J. C. Doyle, "Highly optimized tolerance: A mechanism for power laws in designed systems," *Physics Review E*, vol. 60, pp. 1412–1428, 1999.

[51] ——, "Highly optimized tolerance: Robustness and design in complex systems," *Physics Review Letters*, vol. 84, no. 11, pp. 2529–2532, 2000.

[52] A. Cauvin, J. Brannan, and K. Saito, "Common technical specification of the G-PON system among major worldwide access carriers," *IEEE Communications Magazine*, vol. 44, no. 10, pp. 34–40, 2006.

[53] K. Chan, C.-K. Chan, L. K. Chen, and F. Tong, "Demonstration of 20-Gb/s all-optical XOR gate by four-wave mixing in semiconductor optical amplifier with RZ-DPSK modulated inputs," *IEEE Photonics Technology Letters*, vol. 16, no. 3, pp. 897–899, 2004.

[54] V. W. S. Chan, "Editorial," *IEEE Journal on Selected Areas in Communications: Optical Communications and Networking Series*, vol. 23, no. 8.

[55] ——, lecture notes, 6.442 Optical Networks, MIT, Dec. 2004.

[56] ——, personal communication, 2008.

[57] V. W. S. Chan, S. Chan, and S. Mookherjea, "Optical distribution networks," *Optical Networks Magazine*, vol. 3, pp. 25–33, 2002.

[58] V. W. S. Chan, K. L. Hall, E. Modiano, and K. A. Rauschenbach, "Architectures and technologies for high-speed optical data networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 16, no. 12, pp. 2146–2168, 1998.

[59] V. W. S. Chan, G. Weichenberg, and M. Médard, "Optical Flow Switching," *Proceedings of the Workshop on Optical Burst Switching (WOBS)*, 2006.

[60] C. S. Chang, W. J. Chen, and H. Y. Huang, "On service guarantees for input buffered crossbar switches: A capacity decomposition approach by Birkhoff and von Neumann," *Proceedings of IEEE International Workshop on Quality of Service (IWQoS)*, pp. 79–86, 1999.

[61] M. S. Chen, N. R. Dono, and R. Ramaswami, "A new media access protocol for packet switched wavelength division multiaccess metropolitan network," *IEEE/OSA Journal of Lightwave Technology*, vol. 8, no. 8, pp. 1048–1057, 1990.

[62] Y. Chen, H. Wu, D. Xu, and C. Qiao, "Performance analysis of Optical Burst Switched node with deflection routing," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, vol. 2, pp. 1355–1359, 2004.

[63] J. Cheyns, E. V. Breusegem, D. Colle, M. Pickavet, and P. Demeester, "ORION: A novel hybrid network concept: Overspill Routing in Optical Networks," *Proceedings of IEEE International Conference on Transparent Optical Networks (ICTON)*, pp. 144–147, 2003.

[64] D. Chiaroni, B. Lavigne, A. Jourdan, L. Hamon, C. Janz, and M. Renaud, "New 10 Gb/s 3R NRZ optical regenerative interface based on semiconductor optical amplifiers for all-optical networks," *Proceedings of European Conference on Optical Communication (ECOC)*, vol. 5, pp. 41–43, 1997.

[65] I. Chlamtac *et al.*, "CORD: contention resolution by delay lines," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 5, pp. 1014–1029, 1996.

[66] I. Chlamtac and A. Ganz, "Channel allocation protocols in frequency-time controlled high-speed networks," *IEEE Transactions on Communications*, vol. 36, no. 4, pp. 430–440, 1996.

[67] W. Cho and B. Mukherjee, "Architecture and protocols for packet communication in optical WDM metropolitan-area ring networks using tunable wavelength add-drop multiplexers," *Optical Networks Magazine*, vol. 4, no. 5, pp. 71–85, 2003.

[68] K. Christodoulopoulos *et al.*, "Performance evaluation of Overspill Routing in Optical Networks," *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 6, 2006.

[69] S. T. Chuang, S. Iyer, and N. McKeown, "Practical algorithms for performance guarantees in buffered crossbars," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, vol. 2, pp. 981–991, 2005.

[70] T. Clark, *IP SANs.* Addison-Wesley, 2002.

[71] D. Cotter *et al.*, "Self routing of 100 Gbit/sec packets using 6 bit 'keyword' address recognition," *IEEE Electronics Letters*, vol. 31, no. 17, pp. 1475–1476, 1995.

[72] ——, "Nonlinear optics for high-speed digital information processing," *Science*, vol. 286, no. 5444, pp. 1523–1528, 1999.

[73] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, 1997.

[74] J. G. Dai and B. Prabhakar, "The throughput of data switches with and without speedup," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, vol. 2, pp. 556–564, 2000.

[75] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes.* New York, USA: Springer-Verlag, 1988.

[76] S. Danielsen, P. B. Hansen, K. E. Stubjaer, M. Schilling, K. Wunstel, P. Idler, P. Doussiere, and F. Pommerau, "All optical wavelength conversion schemes for increased input power range," *IEEE Photonics Technology Letters*, vol. 10, no. 1, pp. 60–62, 1998.

[77] R. Davey and J. K. F. B. K. McCammon, "Options for future optical access networks," *IEEE Communications Magazine*, vol. 44, no. 10, pp. 50–56, 2006.

[78] I. V. de Voorde, C. M. Martin, I. Vandewege, and X. Z. Oiu, "The super-PON demonstrator: An exploration of possible evolution paths for optical access networks," *IEEE Communications Magazine*, vol. 38, no. 2, pp. 74–82, 2000.

[79] E. B. Desurvire, *Erbium-Doped Fiber Amplifiers: Principles and Applications.* New York, USA: John Wiley, 1994.

[80] ——, *Erbium-Doped Fiber Amplifiers: Device and System Developments.* New York, USA: John Wiley, 2002.

[81] ——, *Erbium-Doped Fiber Amplifiers: Principles and Applications.* New Jersey, USA: John Wiley, 2002.

[82] ——, "Capacity demand and technology challenges for lightwave systems in the next two decades," *IEEE/OSA Journal of Lightwave Technology*, vol. 24, no. 12, pp. 4697–4710, 2006.

[83] A. R. Dhaini, C. M. Assi, M. Maier, and A. Shami, "Dynamic wavelength and bandwidth allocation in hybrid TDM/WDM EPON networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 25, no. 1, pp. 277–286, 2007.

[84] S. S. Dixit, Ed., *IP over WDM: Building the Next Generation Optical Internet.* Wiley-Interscience, 2003.

[85] H. J. S. Dorren *et al.*, "All-optical logic based on ultrafast gain and index dynamics in a semiconductor optical amplifier," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 10, no. 5, pp. 1079–1092, 2004.

[86] H. J. S. Dorren, D. Lenstra, Y. Liu, M. T. Hill, and G. D. Khoe, "Nonlinear polarization rotation in semiconductor optical amplifiers: Theory and application to all-optical flip-flop memories," *IEEE Journal of Quantum Electronics*, vol. 39, no. 1, pp. 141–148, 2003.

[87] P. Dowd, "Random access protocols for high speed interprocess communications based on a passive optical star topology," *IEEE/OSA Journal of Lightwave Technology*, vol. 9, no. 6, pp. 799–808, 1991.

[88] J. C. Doyle, S. Low, J. M. Carlson, F. Paganini, G. Vinnicombe, W. Willinger, and P. Parillo, "Robustness and the internet: Theoretical foundations," in *Robust Design: A Repertoire of Biological, Ecological, and Engineering Case Studies*, E. Jen, Ed. Oxford University Press, 2005.

[89] C. Dragone, "Optimum design of a planar array of tapered waveguides," *Journal of the Optical Society of America*, vol. 7, pp. 2081–2093, 1990.

[90] C. Dragone, C. Edwards, and R. Kistler, "Integrated optics $n \times n$ multiplexer on silicon," *IEEE Photonics Technology Letters*, vol. 3, no. 10, pp. 896–899, 1991.

[91] C. Dragone, C. H. Henry, I. P. Kaminow, and R. C. Kistler, "Efficient multichannel integrated optics star coupler on silicon," *IEEE Photonics Technology Letters*, vol. 1, no. 8, pp. 241–243, 1989.

[92] J. H. Dshalalow, Ed., *Frontiers in Queueing.* Boca Raton, USA: CRC Press, 1996.

[93] T. Durhuus, B. Mikkelsen, C. Joergensen, S. L. Danielsen, and K. E. Stubkjaer, "All optical wavelength conversion by semiconductor optical amplifiers," *IEEE/OSA Journal of Lightwave Technology / Journal on Selected in Communications Special Issue on Multiwavelength Optical Technology and Networks*, vol. 14, no. 6, pp. 942–954, 1996.

[94] A. K. Dutta, N. K. Dutta, and M. Fujiwara, Eds., *WDM Technologies.* Burlington, MA: Elsevier Academic Press, 2004.

[95] F. Effenberger, D. Cleary, O. Haran, G. Kramer, R. D. Li, M. Oron, and T. Pfeiffer, "An introduction to PON technologies," *IEEE Communications Magazine*, vol. 45, no. 3, pp. S17–S24, 2008.

[96] M. Eiselt, W. Pieper, and H. G. Weber, "SLALOM: semiconductor laser amplifier in a loop mirror," *IEEE/OSA Journal of Lightwave Technology*, vol. 13, no. 10, pp. 2099–2112, 1995.

[97] J. M. H. Elmirghani and H. T. Mouftah, "All-optical wavelength conversion technologies and applications in DWDM networks," *IEEE Communications Magazine*, vol. 38, no. 3, pp. 86–92, 2000.

[98] G. I. Falin and J. G. C. Templeton, *Retrial Queues.* Great Britain: Chapman & Hall, 1997.

[99] C. Fan, M. Maier, and M. Reisslein, "The AWG||PSC network: A performance-enhanced single-hop WDM network with heterogeneous protection," *IEEE/OSA Journal of Lightwave Technology*, vol. 22, no. 5, pp. 1242–1262, 2004.

[100] A. Feldmann, A. Gilbert, W. Willinger, and T. Kurtz, "The changing nature of network traffic: Scaling phenomena," *Computer Communication Review*, vol. 28, no. 2, 1998.

[101] M. J. Ferguson, "A study of unslotted Aloha with arbitrary message lengths," *Proceedings of IEEE Data Communications Symposium (DataComm)*, pp. 5.20–5.25, 1975.

[102] M. R. Fisher, S. Minin, and S.-L. Chuang, "Tunable optical group delay in an active waveguide semiconductor resonator," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 11, no. 1, pp. 197–203, 2005.

[103] J. E. Fouquet, "Compact optical cross-connect switch based on total internal reflection in a fluid-containing planar lightwave circuit," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, vol. 1, pp. 204–206, 2000.

[104] H. Frazier and H. Johnson, "Gigabit Ethernet: From 100 to 1000 Mbps," *IEEE Internet Computing*, vol. 3, no. 1, pp. 24–31, 1999.

[105] N. M. Froberg, S. R. Henion, H. G. Rao, B. K. Hazzard, S. Parikh, B. R. Romkey, and M. Kuznetsov, "The NGI ONRAMP test bed: Reconfigurable WDM technology for next generation regional access networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 18, no. 12, pp. 1697–1708, 2000.

[106] A. Fumagalli, G. Balestra, and L. Valcarenghi, "Optimal amplifier placement in multi-wavelength optical networks based on simulated annealing," *Proceedings of SPIE Optical Networking and Communications Conference (OptiComm)*, vol. 3531, pp. 268–279, 1998.

[107] A. Fumagalli and P. Krishnamoorthy, "A low-latency and bandwidth-efficient distributed Optical Burst Switching architecture for metro ring," *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 2, pp. 1340–1344, 2003.

[108] J. M. Gabriagues and J. B. Jacob, "OASIS: A high-speed photonic ATM switch – results and perspectives," *Proceedings of IEEE International Switching Symposium*, pp. 457–461, 1995.

[109] A. R. Ganguly, S.M. Dissertation, Massachusetts Institute of Technology, in preparation.

[110] B. Ganguly, "Implementation and modeling of a scheduled Optical Flow Switching (OFS) network," Ph.D. Dissertation, Massachusetts Institute of Technology, 2008.

[111] B. Ganguly and V. W. S. Chan, "A scheduled approach to optical flow switching in the ONRAMP optical access network testbed," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, pp. 215–216, 2002.

[112] Y. Ganjali, A. Keshavarzian, and D. Shah, "Cell switching versus packet switching in input-queued switches," *IEEE/ACM Transactions on Networking*, vol. 13, no. 4, pp. 782–789, 2005.

[113] C. M. Gauger, P. J. Kuhn, E. V. Breusegem, M. Pickavet, and P. Demeester, "Hybrid optical network architectures: bringing packets and circuits together," *IEEE Communications Magazine*, vol. 44, no. 8, pp. 36–42, 2006.

[114] N. Genay, P. C. andF. Saliou, Q. Liu, T. Soret, and L. Guillo, "Solutions for budget increase for the next generation optical access network," *Proceedings of IEEE International Conference on Transparent Optical Networks (ICTON)*, 2007.

[115] N. Genay, T. Soret, P. Chanclou, B. Landusies, L. Guillo, and F. Saliou, "Evaluation of the budget extension of a GPON by EDFA amplification," *Proceedings of IEEE International Conference on Transparent Optical Networks (ICTON)*, 2007.

[116] S. Gerke, "Weighted colouring and channel assignment," Ph.D. Dissertation, University of Oxford, 2000.

[117] O. Gerstel, personal communication, May 2008.

[118] B. Glance, I. Kaminow, and R. W. Wilson, "Applications of the integrated waveguide grating router," *IEEE/OSA Journal of Lightwave Technology*, vol. 12, no. 6, pp. 957–962, 1994.

[119] I. Glesk, K. I. Kang, and P. R. Prucnal, "All-optical address recognition and self-routing in a 250 Gbit/s packet switched," *IEEE Electronics Letters*, vol. 30, no. 16, pp. 1322–1323, 1994.

[120] K. Grobe and J.-P. Elbers, "PON in adolescence: From TDMA to WDM-PON," *IEEE Communications Magazine*, vol. 46, no. 1, pp. 26–34, 2008.

[121] C. Guan, "Cost-effective optical network architecture – a joint optimization of topology, switching, routing and wavelength assignment," Ph.D. Dissertation, Massachusetts Institute of Technology, 2007.

[122] C. Guan and V. W. S. Chan, "Connectivity architectures of regular optical mesh networks," *Proceedings of IEEE Global Telecommunications Conference (Globecom)*, vol. 3, pp. 2669–2675, 2002.

[123] ——, "Topology design of OXC-switched WDM networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 8, pp. 1670–1686, 2005.

[124] ——, "Cost-efficient physical architecture for OXC-switched WDM mesh networks—(generalized) Moore graphs and their close relatives," *Proceedings of IEEE Global Telecommunications Conference (Globecom)*, 2006.

[125] C. Guillemot *et al.*, "Transparent optical packet switching: the European ACTS KEOPS project approach," *IEEE/OSA Journal of Lightwave Technology*, vol. 16, no. 12, pp. 2117–2134, 1998.

[126] A. Gumaste and I. Chlamtac, "Light-trails: A novel conceptual framework for conducting optical communications," *Proceedings of IEEE Workshop on High Performance Switching and Routing (HPSR)*, 2003.

[127] ——, "Mesh implementation of light-trails: a solution to IP centric communication," *Proceedings of the International Conference on Computer Communications and Networks (ICCCN)*, pp. 178–183, 2003.

[128] A. Gumaste and N. Ghani, "LiTPiC – Light-trails and photonic integrated circuits: Issues of network design and performance," *IEEE Conference on Local Computer Networks*, pp. 37–44, 2007.

[129] A. Gumaste and S. Q. Q. Zheng, "Next-generation optical storage area networks: the light-trails approach," *IEEE Communications Magazine*, vol. 43, no. 3, pp. 72–79, 2005.

[130] S. Q. Z. A. Gumaste, "SMART: An optical infrastructure for future Internet," *Proceedings of IEEE International Conference on Broadband Communications, Networks and Systems (BROADNETS)*, pp. 1–12, 2006.

[131] A. Gupta, X. Lin, and R. Srikant, "Low-complexity distributed scheduling algorithms for wireless networks," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, 2007.

[132] G. C. Gupta, M. Kashima, H. Iwamura, H. Tamai, T. Ushikubo, and T. Kamijoh, "Over 100 km bidirectional, multi-channels COF-PON without optical amplifier," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, pp. 1–3, 2006.

[133] I. M. I. Habib, M. Kavehrad, and C.-E. W. Sundberg, "Protocols for very high-speed optical fiber local area networks using a passive star topology," *IEEE/OSA Journal of Lightwave Technology*, vol. 5, no. 12, pp. 1782–1794, 1987.

[134] K. L. Hall, J. P. Donnely, S. H. Groves, C. I. Fennely, R. J. Bailey, and A. Napoleone, "40 Gbit/sec all-optical circulating shift register with an inverter," *OSA Optics Letters*, vol. 22, no. 19, pp. 1479–1481, 1997.

[135] K. L. Hall, E. P. Ippen, and G. Eisenstein, "Bias-lead monitoring of ultrafast nonlinearities in InGaAsP diode laser amplifiers," *AIP Applied Physics Letters*, vol. 57, no. 2, pp. 129–131, 1990.

[136] K. L. Hall, G. Lenz, A. M. Darwish, and E. P. Ippen, "Subpicosecond gain and index nonlinearities in InGaAsP diode lasers," *Optics Communications*, vol. 111, no. 5, pp. 589–612, 1994.

[137] K. L. Hall, G. Lenz, P. Ippen, and G. Raybon, "Heterodyne pump-probe technique for time-domain studies of optical nonlinearities in waveguides," *OSA Optics Letters*, vol. 17, no. 12, pp. 874–876, 1992.

[138] K. L. Hall, J. Mark, E. P. Ippen, and G. Eisenstein, "Femtosecond gain dynamics in InGaAsP optical amplifiers," *AIP Applied Physics Letters*, vol. 56, no. 18, pp. 1740–1742, 1990.

[139] M. Hamdi, "ORMA: a high-performance MAC protocol for fiber-optic LANs/MANs," *IEEE Communications Magazine*, vol. 35, no. 3, pp. 110–119, 1997.

[140] J. S. Harris, "Tunable long-wavelength vertical-cavity lasers: The engine of next generation optical networks?" *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 6, no. 6, pp. 1145–1160, 2000.

[141] J. E. Heebner and R. W. Boyd, "slow and fast light in resonator-coupled waveguides," *Journal of Modern Optics*, vol. 49, no. 14, pp. 2629–2636, 2002.

[142] Y. Hibino, "An array of photonic filtering advantages: arrayed-waveguide-grating multi/demultiplexers for photonic networks," *IEEE Circuits and Devices Magazine*, vol. 16, no. 6, pp. 21–27, 2000.

[143] M. T. Hill, H. de Waardt, G. D. Khoe, and H. J. S. Dorren, "Fast optical flip-flop by use of MachZehnder interferometers," *Microwave and Optical Technology Letters*, vol. 31, no. 6, pp. 411–415, 2001.

[144] M. T. Hill, A. Srivatsa, N. Calabretta, Y. Liu, H. de Waardt, G. D. Khoe, and H. J. S. Dorren, "1×2 all-optical packet switch using all-optical header processing," *IEEE Electronics Letters*, vol. 37, no. 12, pp. 774–775, 2001.

[145] A. J. Hoffman and R. R. Singleton, "On moore graphs with diameters 2 and 3," *IBM Journal of Research and Development*, vol. 4, pp. 497–504, 1960.

[146] J. Homa and K. Bala, "ROADM architectures and their enabling WSS technology," *IEEE Communications Magazine*, vol. 46, no. 7, pp. 150–153, 2008.

[147] Y.-L. Hsueh *et al.*, "Traffic grooming on WDM rings using optical burst transport," *IEEE/OSA Journal of Lightwave Technology*, vol. 24, no. 1, pp. 44–53, 2006.

[148] D. K. Hunter, M. C. Chia, and I. Andonovic, "Buffering in optical packet switches," *IEEE/OSA Journal of Lightwave Technology*, vol. 16, no. 12, pp. 2081–2094, 1998.

[149] D. K. Hunter, W. D. Cornwell, T. H. Gilfedder, A. Franzen, and I. Andonovic, "SLOB: A switch with large optical buffers for packet switching," *IEEE/OSA Journal of Lightwave Technology*, vol. 16, no. 10, pp. 1725–1736, 1998.

[150] E. Iannone, R. Sabella, L. D. Stefano, and F. Valeri, "All optical wavelength conversion in optical multicarrier networks," *IEEE Transactions on Communications*, vol. 44, no. 6, pp. 716–724, 1996.

[151] M. N. Islam, Ed., *Raman Amplifiers for Telecommunications 2: Sub-Systems and Systems*, 1st ed. Springer, 2003.

[152] J. L. Jackel *et al.*, "Acousto-optic tunable filters (AOTFs) for multiwavelength optical cross-connects," *IEEE/OSA Journal of Lightwave Technology / Journal on Selected in Communications Special Issue on Multiwavelength Optical Technology and Networks*, vol. 14, no. 6, pp. 1056–1066, 1996.

[153] E. Jahn *et al.*, "40 Gbit/s all-optical demultiplexing using a monolithically integrated MachZehnder interferometer with semiconductor laser amplifier," *IEEE Electronics Letters*, vol. 31, no. 12, pp. 1857–1858, 1995.

[154] V. Jayaraman, Z.-M. Chuang, and L. A. Coldren, "Theory, design and performance of extended tuning range semiconductor lasers with sampled gratings," *IEEE Journal of Quantum Electronics*, vol. 29, no. 6, pp. 1824–1834, 1993.

[155] H. B. Jeon and C. K. Un, "Contention based reservation protocols in multiwavelength protocols with passive star topology," *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 3, pp. 1473–1477, 1992.

[156] L. B. Jeunhomme, *Single-Mode Fiber Optics*. New York, USA: Marcel Dekker, 1990.

[157] F. Jia and B. Mukherejee, "The receiver collision avoidance (RCA) protocol for single hop WDM lightwave networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 11, no. 5, pp. 1053–1065, 1993.

[158] S. Jiang, G. Hu, S. Y. Liew, and H. J. Chao, "Scheduling algorithms for shared fiber-delay-line optical packet switches – part ii: The three-stage clos-network case," *IEEE/OSA Journal of Lightwave Technology*, vol. 23, no. 4, pp. 1601–1609, 2005.

[159] O. Kamatani and S. Kawanishi, "Ultrahigh-speed clock recovery with phase lock loop based on four wave mixing in a travelling wave laser diode amplifier," *IEEE/OSA Journal of Lightwave Technology*, vol. 14, no. 8, pp. 1757–1767, 1996.

[160] I. P. Kaminow *et al.*, "A wideband all-optical WDM network," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 5, pp. 780–799, 1996.

[161] I. P. Kaminow and T. L. Koch, Eds., *Optical Fiber Telecommunications IIIA*. San Diego, USA: Academic Press, 1997.

[162] I. P. Kaminow and S. D. Miller, Eds., *Optical Fiber Telecommunications II*. San Diego, USA: Academic Press, 1988.

[163] N. Kashima, *Passive Optical Components for Optical Fiber Transmission*. Boston, USA: Artech House, 1995.

[164] R. Kashyap, *Fibre Bragg Gratings*. San Diego, USA: Academic Press, 1999.

[165] L. G. Kazovsky, W.-T. Shaw, D. Gutierrez, N. Cheng, and S.-W. Wong, "Next-generation optical access networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 25, no. 11, pp. 3428–3442, 2007.

[166] G. Keiser, *Optical Communications Essentials*. New York, USA: McGraw-Hill, 2003.

[167] M. P. Kessler and E. P. Ippen, "Subpicosecond gain dynamics in GaAlAs laser diodes," *AIP Applied Physics Letters*, vol. 51, pp. 1765–1767, 1987.

[168] J. Kim, S. L. Chuang, P. C. Ku, and C. J. Chang-Hasnain, "Slow light using semiconductor quantum dots," *Journal of Physics: Condensed Matter*, vol. 16, no. 35, pp. 3727–3735, 2004.

[169] M. Kim, J. K. Sundararajan, M. Médard, A. Eryilmaz, and R. Koetter, "Network coding in a multicast switch," *IEEE Transactions on Information Theory*, submitted.

[170] Y.-I. Kim, J. H. Kim, S. Lee, D. H. Woo, S. H. Kim, and T.-H. Yoon, "Broadband all-optical flipflop based on optical bistability in an integrated SOA/DFB-SOA," *IEEE Photonics Technology Letters*, vol. 16, pp. 398–400, Feb. 2004.

[171] L. Kleinrock, *Queueing Systems, Volume I: Theory.* Wiley-Interscience, 1975.

[172] K. Kobayashi and I. Mito, "Single frequency and tunable laser diodes," *IEEE/OSA Journal of Lightwave Technology*, vol. 6, no. 11, pp. 1623–1633, 1988.

[173] E. Kozlovski, M. Dueser, A. Zapata, and P. Bayvel, "Wavelength-routed Optical Burst Switched networks," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, pp. 774–776, 2002.

[174] G. Kramer, B. Mukherjee, and G. Pessavento, "Ethernet PON (EPON): Design and analysis of an optical access newtork," *Photonic Network Communications*, vol. 3, no. 3, pp. 307–319, 2001.

[175] P. Ku and C. Chang-Hasnain, "Semiconductor all-optical buffers using quantum dots in resonator structures," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, vol. 1, pp. 76–78, 2002.

[176] P. R. Kumar and S. P. Meyn, "Stability of queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, vol. 40, no. 2, pp. 251–260, 2004.

[177] S. Kumar, J. Turner, and P. Crowley, "Addressing queuing bottlenecks at high speeds," *Proceedings of Symposium on High Performance Interconnects*, pp. 209–224, 2005.

[178] H. Kuwatsuka, T. Akiyama, B. E. Little, T. Simoyama, and H. Ishikawa, "Wavelength conversion of picosecond optical pulses using four wave mixing in a DFB laser," *Proceedings of European Conference on Optical Communication (ECOC)*, vol. 3, pp. 65–66, 2000.

[179] M. Kuznetsov *et al.*, "A next-generation optical regional access networks," *IEEE Communications Magazine*, vol. 38, no. 1, pp. 66–72, 2000.

[180] S. Lan, S. Nishikawa, H. Ishikawa, and O. Wada, "Design of impurity band-based photonic crystal waveguides and delay lines for ultrashort optical pulses," *Journal of Applied Physics*, vol. 90, no. 9, pp. 4321–4327, 2001.

[181] H. Laor, "Construction and performance of a 576×576 single-stage OXC," *Proceedings of IEEE Lasers and Electro-Optics Society Annual Meeting (LEOS)*, vol. 2, pp. 481–482, 1999.

[182] G. M. Lee, B. Wydrowski, M. Zukerman, J. K. Choi, and C. H. Foh, "Performance evaluation of an optical hybrid switching system," *Proceedings of IEEE Global Telecommunications Conference (Globecom)*, vol. 5, pp. 2508–2512, 2003.

[183] T. Lee, K. Lee, and S. Park, "Optimal routing and wavelength assignment in WDM ring networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, pp. 2146–2154, 2000.

[184] Leichtman Research Group, "2.2 million add high-speed Internet in the first quarter of 2008," May 2008. [Online]. Available: http://www.leichtmanresearch.com/press/051508release.html

[185] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[186] G. Lenz, B. J. Eggleton, C. K. Madsen, and R. E. Slusher, "Optical delay lines based on optical filters," *IEEE Journal of Quantum Electronics*, vol. 37, no. 4, pp. 525–532, 2001.

[187] E. Leonardi, M. Mellia, F. Neri, and M. A. Marsan, "Bounds on average delays and queue size averages and variances in input-queued cell-based switches," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, vol. 2, pp. 1095–1103, 2001.

[188] F. Leonberger, "CIPS PON white paper summary and key issues," CIPS OBWG Advance PON Workshop, 2008.

[189] S. Y. Liew, G. Hu, and H. J. Chao, "Scheduling algorithms for shared fiber-delay-line optical packet switches – part i: The single-stage case," *IEEE/OSA Journal of Lightwave Technology*, vol. 23, no. 4, pp. 1586–1600, 2005.

[190] L. Y. Lin, E. L. Goldstein, and R. W. Tkach, "Free-space micromachined optical switches with submillisecond switching time for large-scale crossconnects," *IEEE Photonics Technology Letters*, vol. 10, no. 4, pp. 525–527, 1998.

[191] X. Lin and S. Rasool, "Constant-time distributed scheduling policies for ad hoc wireless networks," *Proceedings of IEEE Conference on Decision and Control*, 2006.

[192] Y. Liu, M. T. Hill, R. Geldenhuys, N. Calabretta, H. de Waardt, G.-D. Khoe, and H. J. S. Dorren, "Demonstration of a variable optical delay for a recirculating buffer by using all-optical signal processing," *IEEE Photonics Technology Letters*, vol. 16, no. 7, pp. 1748–1750, 2004.

[193] A. Lodha, A. Gumaste, P. Bafiia, and N. Ghani, "Stochastic optimization of Light-trail WDM ring networks using Benderapos's Decomposition," *IEEE Workshop on High Performance Switching and Routing (HPSR)*, pp. 1–7, 2007.

[194] V. V. Lozin and M. Milanič, "A polynomial algorithm to find an independent set of maximum weight in a fork-free graph," *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 26–30, 2006.

[195] Y. Maeda, "ATM-PON FTTH access networks and services," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, pp. 66–68, 1999.

[196] M. Maier, *Metropolitan Area WDM Networks.* Kluwer, 2003.

[197] M. Maier and M. Reisslein, "AWG-based metro WDM networking," *IEEE Communications Magazine*, vol. 42, no. 11, pp. S19–S26, 2004.

[198] N. M. Margalit, S. Z. Zhang, and J. E. Bowers, "Vertical cavity lasers for telecom applications," *IEEE Communications Magazine*, vol. 35, no. 5, pp. 164–170, 1997.

[199] M. Marsan, M. Franceschinis, E. Leonardi, F. Neri, and A. Tarello, "Instability phenomena in underloaded packet networks with QoS schedulers," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, vol. 2, pp. 959–969, 2003.

[200] M. Marsan, P. Giaccone, E. Leonardi, and F. Neri, "On the stability of local scheduling policies in networks of packet switches with input queues," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 4, pp. 642–655, 2003.

[201] M. Marsan, E. Leonardi, M. Mellia, and F. Neri, "On the stability of isolated and interconnected input-queueing switches under multiclass traffic," *Technical Report No.16-02-2004*, pp. 1–19, 2004.

[202] T. P. McGarty, personal communication, 2008.

[203] K. McGreer, "Arrayed waveguide gratings for wavelength routing," *IEEE Communications Magazine*, vol. 36, no. 12, pp. 62–68, 1998.

[204] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch (extended version)," *IEEE Transactions on Communications*, vol. 47, no. 8, pp. 1260–1267, 1999.

[205] N. Mehravari, "Performance and protocol improvements for very high speed optical fiber local area networks using a passive star topology," *IEEE/OSA Journal of Lightwave Technology*, vol. 8, no. 4, pp. 520–530, 1990.

[206] S. P. Meyn and R. Tweedie, *Markov chain and stochastic stability.* Berlin, Germany: Springer-Verlag, 1993.

[207] J. E. Midwinter and P. W. Smith, Eds., *IEEE Journal on Selected Areas in Communications: Special Issue on Photonic Switching*, 1988, vol. 6.

[208] I. Miguel *et al.*, "Polymorphic architectures for optical networks and their seamless evolution towards next generation networks," *Springer Photonic Network Communications*, vol. 8, no. 2, pp. 177–189, 2004.

[209] G. J. Minty, "On maximal independent sets of vertices in claw-free graphs," *Journal of Combinatorial Theory, Series B*, vol. 28, pp. 280–304, 1980.

[210] S. Mookherjea, "Optical distribution networks: Signal-to-noise ratio optimization and distributed erbium-doped fiber amplifiers," S.M. Dissertation, Massachusetts Institute of Technology, 2000.

[211] A. A. M.Saleh and H. Kogelnik, "Reflective single-mode fiber-optic passive star couplers," *IEEE/OSA Journal of Lightwave Technology*, vol. 6, no. 3, pp. 392–398, 1988.

[212] B. Mukherjee, "WDM-based local lightwave networks Part I: Single-hop systems," *IEEE Network*, vol. 6, no. 3, pp. 12–27, 1992.

[213] ——, "WDM-based local lightwave networks Part II: Multi-hop systems," *IEEE Network*, vol. 6, no. 4, pp. 20–31, 1992.

[214] ——, *Optical WDM Networks*, New York, 2006.

[215] H. R. Muller *et al.*, "DQMA and CRMA: New access schemes for Gbit/s LANs and MANs," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, pp. 185–191, 1990.

[216] S. Nakamura *et al.*, "Demultiplexing of 168-Gb/s data pulses with a hybrid-integrated symmetric Mach-Zehnder all-optical switch," *IEEE Photonics Technology Letters*, vol. 12, no. 4, pp. 425–427, 2000.

[217] S. Nakamura, Y. Ueno, and K. Tajima, "Ultrafast (200 fs, 1.5 Tb/s demultiplexing) and high repetition (10 GHz) operations of a polarization discriminating symmetric Mach-Zehnder all-optical switch," *IEEE Photonics Technology Letters*, vol. 10, no. 11, pp. 1575–1577, 1998.

[218] ——, "Femtosecond switching with semiconductor optical amplifier based symmetric Mach-Zehnder type optical switch," *AIP Applied Physics Letters*, vol. 78, no. 25, pp. 3929–3931, 2001.

[219] S. Namiki and Y. Emori, "Recent advances in ultra-wideband Raman amplifiers," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, vol. 4, pp. 98–99, 2000.

[220] ——, "Ultra-broadband Raman amplifiers pumped and gain-equalized by wavelength-division multiplexed high-power laser diodes," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 7, no. 1, pp. 3–16, 2001.

[221] M. J. Neely, "Dynamic power allocation and routing for satellite and wireless networks with time varying channels," Ph.D. Dissertation, Massachusetts Institute of Technology, 2003.

[222] D. T. Neilson *et al.*, "Fully provisioned 112 × 112 micro-mechanical optical crossconnect with 35.8 Tb/s demonstrated capacity," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, vol. 4, pp. 202–204, 2000.

[223] D. Nesset, T. Kelly, and D. Marcenac, "All-optical wavelength conversion using SOA nonlinearities," *IEEE Communications Magazine*, vol. 36, no. 12, pp. 56–61, 1998.

[224] A. Neukermans and R. Ramaswami, "MEMS technology for optical networking applications," *IEEE Communications Magazine*, vol. 39, no. 1, pp. 62–69, 2001.

[225] E.-G. Neumann, *Single-Mode Fibers.* Berlin, Germany: Springer-Verlag, 1988.

[226] M. Nord, "Hybrid wavelength routed and optical packet switched ring networks for the metropolitan area network," *Proceedings of IEEE International Conference on Transparent Optical Networks (ICTON)*, 2005.

[227] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, pp. 387–396, 1994.

[228] C. Nuzman and I. Widjaja, "Design and performance evaluation of scalable TWIN networks," *Proceedings of IEEE Workshop on High Performance Switching and Routing (HPSR)*, pp. 152–156, 2005.

[229] M. M. O'Mahony, "Semiconductor laser optical amplifiers for future fiber systems," *IEEE/OSA Journal of Lightwave Technology*, vol. 6, no. 4, pp. 531–544, 1988.

[230] J. Osterholz, "Providing power to the edge: The Global Information Grid Bandwidth Expansion," Global Information Grid Bandwidth Expansion (GIG BE) Program Industry Conference, 2002.

[231] J. S. Patel and Y. Silberberg, "Liquid crystal and grating-based multiple-wavelength cross-connect switch," *IEEE Photonics Technology Letters*, vol. 7, no. 5, pp. 514–516, 1995.

[232] N. S. Patel, "Optical networking: Historical perspectives and future trends," lecture notes, 6.442 Optical Networks, MIT, 2008.

[233] N. S. Patel, K. L. Hall, and K. A. Rauschenbach, "40 Gbit/sec cascadable all-optical logic with an ultrafast nonlinear interferometer," *OSA Optics Letters*, vol. 21, no. 18, pp. 1466–1468, 1996.

[234] D. M. Patrick, A. D. Ellis, D. A. O. Davies, M. C. Tatham, and G. Sherlock, "Demultiplexing using polarization rotation in a semiconductor laser amplifier," *IEEE Electronics Letters*, vol. 30, no. 4, pp. 341–342, 1994.

[235] V. Paxson and S. Floyd, "Wide-area traffic: the failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[236] D. F. Phillips, A. Fleischhauer, A. Mair, R. L. W. R. L, and M. D. Lukin, "Storage of light in atomic vapor," *APS Physical Review Letters*, vol. 86, no. 5, pp. 783–786, 2001.

[237] M. J. Potasek, G. P. Agrawal, and S. C. Pinault, "Analytic and numerical study of pulse broadening in nonlinear dispersive optical fibers," *Journal of Optical Society of America B*, vol. 3, no. 2, pp. 205–211, 1986.

[238] A. J. Poustie, K. J. Blow, A. E. Kelly, and R. J. Manning, "All-optical binary half-adder," *Optics Communications*, vol. 156, no. 1, pp. 22–26, 1998.

[239] ——, "All-optical full-adder with bit differential delay," *Optics Communications*, vol. 168, no. 1, pp. 89–93, 1999.

[240] ——, "All-optical parity checker with bit-differential delay," *Optics Communications*, vol. 146, no. 1, pp. 37–43, 1999.

[241] ——, "All-optical pseudorandom number generator," *Optics Communications*, vol. 159, no. 4, pp. 208–214, 1999.

[242] A. J. Poustie, K. J. Blow, and R. J. Manning, "All-optical regenerative memory for long term data storage threshold and amplitude," *Optics Communications*, vol. 140, no. 4, pp. 184–186, 1997.

[243] ——, "Storage threshold and amplitude restoration in optical regenerative memory," *Optics Communications*, vol. 146, no. 1, pp. 262–267, 1998.

[244] A. J. Poustie, R. J. Manning, and K. J. Blow, "All-optical circulating shift register using a semiconductor optical amplifier in a fiber loop mirror," *IEEE Electronics Letters*, vol. 32, no. 13, pp. 1215–1216, 1996.

[245] C. Qiao and M. Yoo, "Optical Burst Switching OBS – a new paradigm for an optical internet," *Journal of High Speed Networks*, vol. 8, no. 1, pp. 69–84, 1999.

[246] B. Ramamurthy, J. Iness, and B. Mukherjee, "Minimizing the number of optical amplifiers needed to support a multi-wavelength optical LAN/MAN," *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, vol. 1, pp. 261–268, 1997.

[247] ——, "Optimizing amplifier placements in a multiwavelength optical LAN/MAN: the equally powered-wavelengths case," *IEEE/OSA Lightwave Technology, Journal of*, vol. 16, no. 9, pp. 1560–1569, 1998.

[248] ——, "Optimizing amplifier placements in a multiwavelength optical LAN/MAN: the unequally powered wavelengths case," *IEEE/ACM Transactions on Networking*, vol. 6, no. 6, pp. 755–767, 1998.

[249] K. Ramantas, K. Christodoulopoulos, K. Vlachos, E. V. Breusegem, and M. Pickavet, "SLIP-IN architecture: A new hybrid optical switching scheme," *Proceedings of IEEE International Conference on Broadband Communications, Networks and Systems (BROADNETS)*, pp. 1–7, 2006.

[250] R. Ramaswami and K. Sivarajan, *Optical Networks: A Practical Perspective*, 2nd ed. San Francisco, USA: Morgan Kaufmann, 2002.

[251] T. Rand-Nash, "PON OpEx and CapEx modeling," CIPS OBWG Advance PON Workshop, 2008.

[252] K. C. Reichmann and P. P. Iannone, "Wavelength-enhanced passive optical networks with extended reach," *Proceedings of IEEE Worshop on Local and Metropolitan Area Networks*, pp. 60–64, 2007.

[253] Z. Rosberg, H. L. Vu, M. Zukerman, and J. White, "Performance analyses of Optical Burst-Switching networks," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 7, pp. 1187–1197, 2003.

[254] F. E. Ross, "An overview of FDDI: The Fiber Distributed Data Interface," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 7, pp. 1043–1051, 1989.

[255] K. Ross, N. Bambos, K. Kumaran, I. Saniee, and I. Widjaja, "Scheduling bursts in time-domain wavelength interleaved networks," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 9, pp. 1441–1451, 2003.

[256] R. Ryf *et al.*, "1296-port MEMS transparent optical crossconnect with 2.07 Petabit/s switch capacity," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, vol. 4, pp. PD28–1–PD28–3, 2001.

[257] T. Sakamoto, K. Noguchi, R. Sato, A. Okada, Y. Sakai, and M. Matsuoka, "Variable optical delay circuit using wavelength converters," *IEEE Electronics Letters*, vol. 37, no. 7, pp. 454–455, 2001.

[258] T. Sakamoto, A. Okada, O. Moriwaki, M. Matsuoka, and K. Kikuchi, "Performance analysis of variable optical delay circuit using highly nonlinear fiber parametric wavelength converters," *IEEE/OSA Journal of Lightwave Technology*, vol. 22, no. 3, pp. 874–881, 2004.

[259] A. A. M. Saleh and H. Kogelnik, "Evolution toward the next-generation core optical network," *IEEE/OSA Journal of Lightwave Technology*, vol. 24, no. 9, pp. 3303–3321, 2006.

[260] M. Sampels, "Vertex-symmetric generalized Moore graphs," *Discrete Applied Mathematics*, vol. 138, pp. 195–202, 2004.

[261] S. Sanghavi, L. Bui, and R. Srikant, "Distributed link scheduling with constant overhead," *Proceedings of ACM Sigmetrics*, 2007.

[262] S. Sanghavi and D. Shah, "Tightness of LP via Max-Product belief propagation," Massachusetts Institute of Technology, Tech. Rep., 2008. [Online]. Available: http://arxiv.org/abs/cs/0508097

[263] A. Schrijver, *Combinatorial Optimization.* Germany: Springer-Verlag, 2003.

[264] M. A. Scobey and D. E. Spock, "Passive DWDM components using microplasma optical interference filters," *Proceedings of IEEE/OSA Optical Fiber Communication Conference (OFC)*, vol. 242–243, 1996.

[265] S. Sengupta, V. Kumar, and D. Saha, "Switched optical backbone for cost-effective scalable core IP networks," *IEEE Communications Magazine*, vol. 41, no. 6, pp. 60–70, 2003.

[266] J. M. Simmons, "Survivable passive optical networks based on arrayed-waveguide-grating architectures," *IEEE/OSA Journal of Lightwave Technology*, vol. 25, no. 12, pp. 3658–3668, 2007.

[267] ——, *Optical Network Design and Planning.* Springer, 2008.

[268] R. R. Singleton, "On minimal graphs of maximum even girth," *Journal of Combinatorial Theory*, vol. 1, pp. 306–322, 1966.

[269] R. G. Smith, "Optical power handling capacity of low loss optical fibers as determined by stimulated Raman and Brillouin scattering," *Applied Optics*, vol. 11, no. 11, pp. 2489–2494, 1972.

[270] J. P. Sokoloff, P. R. Prucnal, I. Glesk, and M. Kane, "A terahertz optical asymmetric demultiplexer (TOAD)," *IEEE Photonics Technology Letters*, vol. 5, no. 7, pp. 787–790, 1993.

[271] G. H. Song, "Toward the ideal codirectional Bragg filter with an acousto-optic filter design," *IEEE/OSA Journal of Lightwave Technology*, vol. 13, no. 3, pp. 470–480, 1995.

[272] R. A. Spanke, "Architectures for guided-wave optical space switching systems," *IEEE Communications Magazine*, vol. 25, no. 5, pp. 42–48, 1987.

[273] L. Spiekman, D. Piehler, P. Iannone, K. Reichmann, and H.-H. Lee, "Semiconductor optical amplifiers for FTTx," *Proceedings of IEEE International Conference on Transparent Optical Networks (ICTON)*, 2007.

[274] J. Stern *et al.*, "Passive optical local networks for telephony applications," *IEEE Electronics Letters*, vol. 23, pp. 1255–1257, 1987.

[275] A. Stolyar, "Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic," *Annals of Applied Probability*, vol. 14, no. 1, pp. 1–53, 2004.

[276] K. E. Stubkjaer, "Semiconductor optical amplifier-based all-optical gates for high speed optical processing," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 6, no. 6, pp. 1428–1435, 2000.

[277] G. N. M. Sudhakar, N. D. Georganas, and M. Kavehrad, "A multichannel optical star lan and its application as a broadband switch," *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 2, pp. 843–847, 1992.

[278] G. N. M. Sudhakar, M. Kavehrad, and N. D. Georganas, "Access protocols for passive optical star networks," *Computer Networks and ISDN Systems*, pp. 913–930, 1994.

[279] Y. Sun, J. L. Zyskind, and A. K. Srivastava, "Average inversion level, modeling and physics of erbium-doped fiber amplifiers," *IEEE Journal of Quantum Electronics*, vol. 3, no. 4, pp. 991–1007, 1997.

[280] J. K. Sundararajan, S. Deb, and M. Médard, "Extending the Birkhoff–von Neumann switching strategy to multicast switches," *Proceedings of IFIP Networking Conference*, 2005.

[281] ——, "Extending the Birkhoff–von Neumann switching strategy for multicast — On the use of optical splitting in switches," *IEEE Journal on Selected Areas in Communications: Optical Communications and Networking Series*, vol. 25, no. 6, pp. 36–50, 2007.

[282] E. A. Swanson, personal communication, May 2008.

[283] M. Tabiani and M. Kavehrad, "Theory of an efficient $n \times n$ passive optical star-coupler," *IEEE/OSA Journal of Lightwave Technology*, vol. 9, no. 4, pp. 448–455, 1991.

[284] Y. Tachikawa, Y. Inoue, M. Ishii, and T. Nozawa, "Arrayed waveguide grating multiplexer with loop-back optical paths and its applications," *IEEE/OSA Journal of Lightwave Technology*, vol. 14, no. 6, pp. 977–984, 1996.

[285] K. Tajima, S. Nakamura, and Y. Sugimoto, "Ultrafast polarization-discriminating MachZehnder all-optical switch," *AIP Applied Physics Letters*, vol. 67, no. 25, pp. 3709–3711, 1995.

[286] G. Talli and P. D. Townsend, "Hybrid DWDM-TDM long-reach PON for next-generation optical access," *IEEE/OSA Journal of Lightwave Technology*, vol. 24, no. 7, pp. 2827–2834, 2006.

[287] Y. Tao and J. Q. Hu, "Optical amplifier placement for optical networks," Boston University, Tech. Rep., 2003. [Online]. Available: http://people.bu.edu/hqiang/papers/amp1.pdf

[288] M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *Computer Communication Review*, vol. 27, pp. 5–23, 1997.

[289] C. Taskin, H. Evirgen, H. Ekiz, and A. Y. Cakir, "Evolution of the current telecommunications networks and the next generation optical networks," *Proceedings of International Conference on Internet Surveillance and Protection (ICISP)*, 2006.

[290] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

[291] P. D. Townsend *et al.*, "Long reach passive optical networks," *Proceedings of IEEE Lasers and Electro-Optics Society Annual Meeting (LEOS)*, vol. 1, pp. 868–869, 2007.

[292] D. Towsley, "The Internet is flat: A brief history of networking over the next ten years," INFOCOM 2007 Keynote Address, 2007.

[293] A. Tran, R. Tucker, and N. Boland, "Amplifier placement methods for metropolitan WDM ring networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 22, no. 11, pp. 2509–2522, 2004.

[294] J. Turner, "Terabit burst switching," *Journal of High Speed Networks*, vol. 8, no. 1, pp. 3–16, 1999.

[295] A. V. Turukhin, V. S. Sudarshanam, M. S. Shahriar, J. A. Musser, B. S. Ham, and P. R. Hemmer, "Observation of ultraslow and stored light pulses in a solid," *APS Physical Review Letters*, vol. 88, no. 2, pp. 023 602–1–023 602–4, 2002.

[296] H. Ueda *et al.*, "Deployment status and common technical specifications for a B-PON system," *IEEE Communications Magazine*, vol. 39, no. 12, pp. 134–141, 2001.

[297] H. Ueda and E. Maekawa, "Development of an ATM subscriber system (model C)," *NTT Review*, vol. 11, pp. 27–32, 1999.

[298] S. Uhlig and O. Bonaventure, "Understanding the long-term self-similarity of Internet traffic," *Lecture Notes in Computer Science*, vol. 2156, pp. 286–298, 2001.

[299] D. Vakhshoori *et al.*, "2 mW CW singlemode operation of tunable 1550 nm vertical cavity surface emitting laser," *IEEE Electronics Letters*, vol. 35, no. 11, pp. 900–901, 1999.

[300] E. van Breusegem, J. Cheyns, D. Colle, M. Pickavet, and P. Demeester, "Overspill routing in optical networks: A new architecture for future proof IP over WDM networks," *Proceedings of SPIE Optical Networking and Communications Conference (OptiComm)*, 2003.

[301] A. M. Vengsarkar *et al.*, "Long-period fiber-grating-based gain equalizers," *OSA Optics Letters*, vol. 21, no. 5, pp. 336–338, 1996.

[302] ——, "Long-period gratings as band-rejection filters," *IEEE/OSA Journal on Lightwave Technology*, vol. 14, no. 1, pp. 58–65, 1996.

[303] S. Verbrugge *et al.*, "Modeling operating expenditures for telecom operators," *Proceedings of Conference on Optical Network Design and Modeling (ONDM)*, pp. 455–466, 2005.

[304] S. Verma, H. Chaskar, and R. Ravikanth, "Optical Burst Switching: A viable solution for terabit IP backbone," *IEEE Network*, vol. 14, no. 6, pp. 48–53, 2000.

[305] V. M. Vokkarane, J. P. Jue, and S. Sitaraman, "Burst segmentation: An approach for reducing packet loss in Optical Burst Switched networks," *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 5, pp. 2673–2677, 2002.

[306] J. von Neumann, "A certain zero-sum two-person game equivalent to the optimal assignment problem," *Contributions to the Theory of Games*, vol. 2, pp. 5–12, 1953.

[307] R. E. Wagner, R. C. Alferness, A. M. Saleh, and M. S. Goodman, "MONET: Multiwavelength Optical Networking," *IEEE/OSA Journal on Lightwave Technology*, vol. 14, no. 6, pp. 1349–1355, 1996.

[308] J. Wang, W. Cho, V. R. Vemuri, and B. Mukherjee, "Improved approaches for cost-effective traffic grooming in WDM ring networks: ILP formulations, single-hop and multihop connections," *IEEE/OSA Journal of Lightwave Technology*, vol. 19, no. 11, pp. 1645–1653, 2001.

[309] J. Wang and B. Mukherjee, "Interconnected WDM ring networks: Strategies for interconnection and traffic grooming," *Optical Networks Magazine*, vol. 3, no. 5, pp. 10–20, 2002.

[310] G. Watson, S. Ooi, D. Skellern, and D. Cunningham, "HANGMAN Gbit/s network," *IEEE Network*, vol. 4, no. 6, pp. 10–18, 1992.

[311] G. Watson and S. Tohme, "S++ – a new MAC protocol for Gb/s local area networks," *IEEE Journal on Selected Areas in Communication*, vol. 11, no. 4, pp. 531–539, 1993.

[312] G. Weichenberg, "High-reliability architectures for networks under stress," S.M. Dissertation, Massachusetts Institute of Technology, 2003.

[313] I. Widjaja, "Performance analysis of burst admission-control protocols," *IEE Proceedings Communications*, vol. 142, no. 1, pp. 7–14, 1995.

[314] I. Widjaja, I. Saniee, R. Giles, and D. Mitra, "Light core and intelligent edge for a flexible, thin-layered and cost-effective optical transport network," *IEEE Communications Magazine*, vol. 41, no. 5, pp. S30–S36, 2003.

[315] Wikipedia, "List of United States cities by population," Sep. 2008. [Online]. Available: http://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

[316] W. Willinger and V. Paxson, "When mathematics meets the Internet," *Notices of the AMS*, vol. 45, no. 8, pp. 961–970, 1998.

[317] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, 1997.

[318] K.-Y. Wu and J.-Y. Liu, "Liquid-crystal space and wavelength routing switches," *Proceedings of IEEE Lasers and Electro-Optics Society Annual Meeting (LEOS)*, vol. 1, pp. 28–29, 1996.

[319] T. J. Xia *et al.*, "Novel self-synchronization scheme for high-speed packet TDM networks," *IEEE Photonics Technology Letters*, vol. 11, no. 2, pp. 269–271, 1999.

[320] C. Xin, C. Qiao, Y. Ye, and S. Dixit, "A hybrid optical switching approach," *Proceedings of IEEE Global Telecommunications Conference (Globecom)*, vol. 7, pp. 3808–3812, 2003.

[321] E. Yamazaki, F. Inuzuka, G. Weichenberg, A. Takada, J. Yamawaku, T. Morioka, and M. Koga, "Waveband path virtual concatenation with contention resolution provided by transparent waveband conversion using QPM-LN waveguides," *IEIC Technical Report*, vol. 106, no. 69, pp. 55–60, 2006.

[322] E. Yamazaki, G. Weichenberg, A. Takada, and T. Morioka, "Polarisation-insensitive parametric wavelength conversion without tunable filters for converted light extraction," *Electronics Letters*, vol. 42, no. 6, pp. 365–367, 2006.

[323] A. Yariv, *Quantum Electronics*, 3rd ed. New York, USA: John Wiley, 1989.

[324] ——, *Optical Electronics in Modern Communications*, 5th ed. New York, USA: Oxford University Press, 1997.

[325] M. Yoo, M. Jeong, and C. Qiao, "A high speed protocol for bursty traffic in optical networks," *SPIE All-Optical Communication Systems*, vol. 3230, pp. 79–90, 1997.

[326] S. J. B. Yoo, "Wavelength conversion techniques for WDM network applications," *IEEE/OSA Journal of Lightwave Technology / Journal on Selected in Communications Special Issue on Multiwavelength Optical Technology and Networks*, vol. 14, no. 6, pp. 955–966, 1996.

[327] Q. Zhao and C.-K. Chan, "A wavelength-division-multiplexed passive optical network with flexible optical network unit internetworking capability," *IEEE/OSA Journal of Lightwave Technology*, vol. 25, no. 8, pp. 1970–1977, 2007.

[328] X. Zheng, M. Veeraraghavan, N. S. V. Rao, Q. Wu, and M. Zhu, "CHEETAH: circuit-switched high-speed end-to-end transport architecture testbed," *IEEE Communications Magazine*, vol. 43, no. 8, pp. S11–S17, 2005.

[329] P. Zhou and O. Yang, "How practical is optical packet switching in core networks?" *Proceedings of IEEE Global Telecommunications Conference (Globecom)*, vol. 5, pp. 2709–2713, 2003.