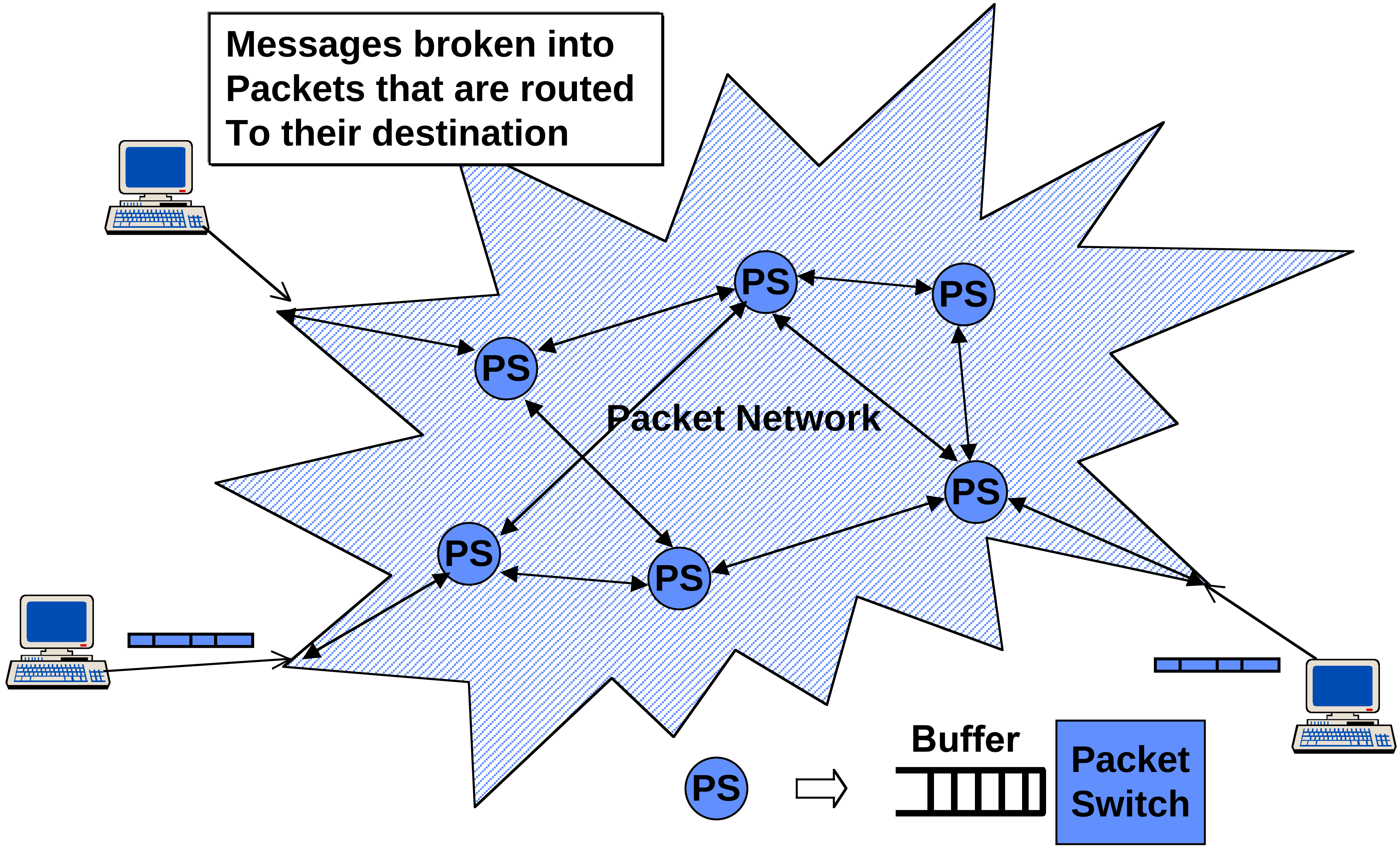
16.36: Communication Systems Engineering

Lecture 17/18: Delay Models for Data Networks

Eytan Modiano

Packet Switched Networks



Queueing Systems

- Used for analyzing network performance

- In packet networks, events are random
  - Random packet arrivals
  - Random packet lengths

- While at the physical layer we were concerned with bit-error-rate, at the network layer we care about delays
  - How long does a packet spend waiting in buffers ?
  - How large are the buffers ?

Random events

- **Arrival process**
  - **Packets arrive according to a random process**
  - **Typically the arrival process is modeled as Poisson**

- **The Poisson process**
  - **Arrival rate of λ packets per second**

  - **Over a small interval δ,**

    P(exactly one arrival) = $\lambda\delta$
    P(0 arrivals) = 1 - $\lambda\delta$
    P(more than one arrival) = 0

  - **It can be shown that:**

$$\text{P}(n \text{ arrivals in interval } T) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}$$

The Poisson Process

$$P(n \text{ arrivals in interval } T) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}$$

n = number of arrivals in T

It can be shown that,

$E[n] = \lambda T$

$E[n^2] = \lambda T + (\lambda T)^2$

$\sigma^2 = E[(n - E[n])^2] = E[n^2] - E[n]^2 = \lambda T$

Inter-arrival times

- Time that elapses between arrivals (IA)

$P(IA <= t) = 1 - P(IA > t)$

$= 1 - P(0 \text{ arrivals in time } t)$

$= 1 - e^{-\lambda t}$

- This is known as the exponential distribution
  - Inter-arrival CDF = $F_{IA}(t) = 1 - e^{-\lambda t}$
  - Inter-arrival PDF = $d/dt\, F_{IA}(t) = \lambda e^{-\lambda t}$

- The exponential distribution is often used to model the service times (I.e., the packet length distribution)

Markov property (Memoryless)

$$P(T \le t_0 + t \mid T > t_0) = P(T \le t)$$

$$\Pr oof:$$

$$P(T \le t_0 + t \mid T > t_0) = \frac{P(t_0 < T \le t_0 + t)}{P(T > t_0)}$$

$$= \frac{\int_{t_0}^{t_0+t} \lambda e^{-\lambda t} dt}{\int_{t_0}^{\infty} \lambda e^{-\lambda t} dt} = \frac{-e^{-\lambda t} \Big|_{t_0}^{t_0+t}}{-e^{-\lambda t} \Big|_{t_0}^{\infty}} = \frac{-e^{-\lambda(t+t_0)} + e^{-\lambda(t_0)}}{e^{-\lambda(t_0)}}$$

$$= 1 - e^{-\lambda t} \;\; = P(T \le t)$$

- **Previous history does not help in predicting the future!**
- **Distribution of the time until the next arrival is independent of when the last arrival occurred!**

Example

- Suppose a train arrives at a station according to a Poisson process with average inter-arrival time of 20 minutes

- When a customer arrives at the station the average amount of time until the next arrival is 20 minutes
  - Regardless of when the previous train arrived

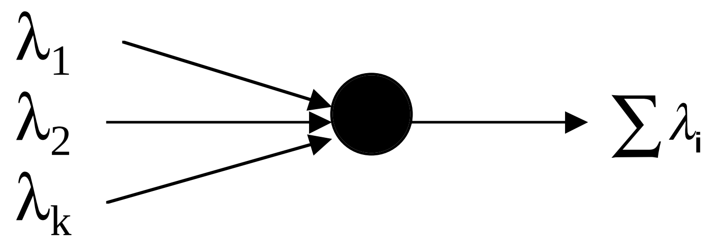
- The average amount of time since the last departure is 20 minutes!

- Paradox: If an average of 20 minutes passed since the last train arrived and an average of 20 minutes until the next train, then an average of 40 minutes will elapse between trains
  - But we assumed an average inter-arrival time of 20 minutes!
  - What happened?

- Answer: You tend to arrive during long inter-arrival times
  - If you don't believe me you have not taken the T

Properties of the Poisson process

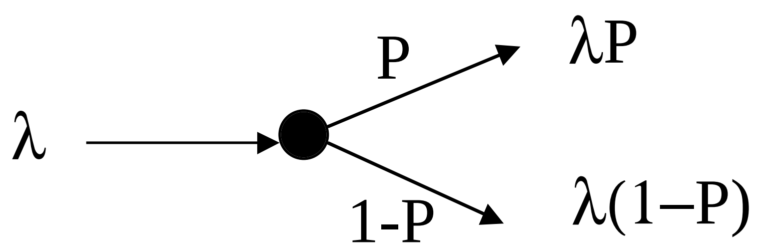
- **Merging Property**



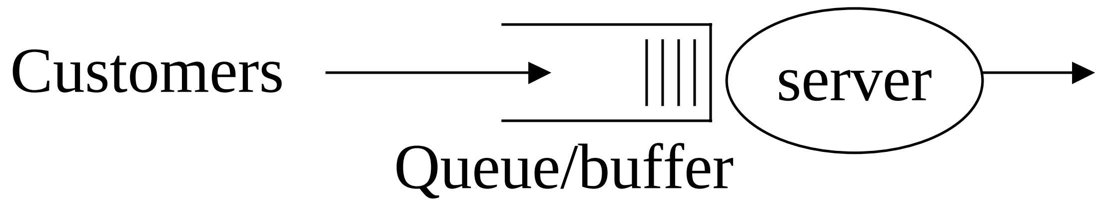
Let A1, A2, … Ak be independent Poisson Processes of rate λ1, λ2, …λk

$$A = \sum A_i \text{ is also Poisson of rate } = \sum \lambda_i$$

- **Splitting property**
  - **Suppose that every arrival is randomly routed with probability P to stream 1 and (1-P) to stream 2**
  - **Streams 1 and 2 are Poisson of rates Pλ and (1-P)λ respectively**



Queueing Models



- **Model for**
  - **Customers waiting in line**
  - **Assembly line**
  - **Packets in a network (transmission line)**

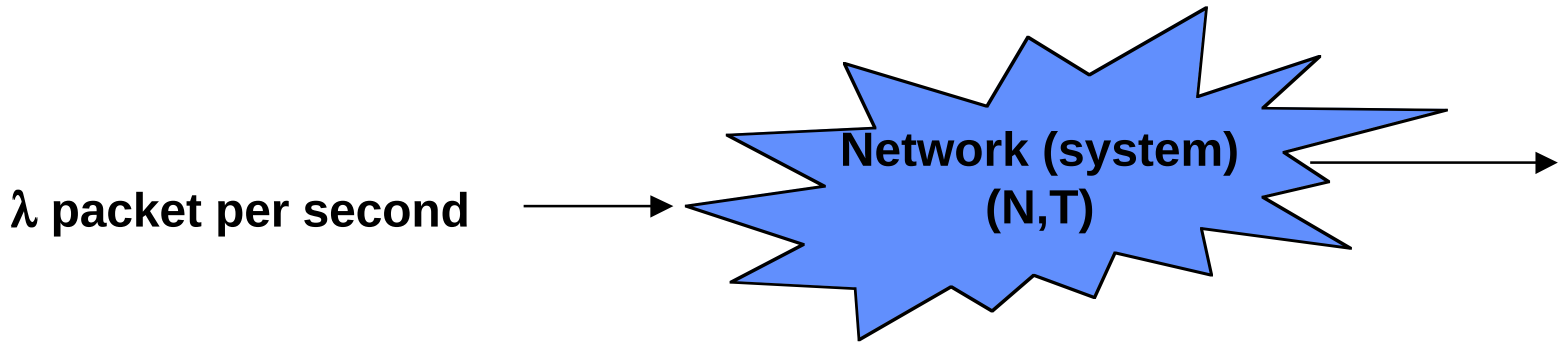
- **Want to know**
  - **Average number of customers in the system**
  - **Average delay experienced by a customer**

- **Quantities obtained in terms of**
  - **Arrival rate of customers (average number of customers per unit time)**
  - **Service rate (average number of customers that the server can serve per unit time)**

Analyzing delay in networks (queueing theory)

- **Little's theorem**
  - **Relates delay to number of users in the system**
  - **Can be applied to any system**
- **Simple queueing systems (single server)**
  - **M/M/1, M/G/1, M/D/1**
  - **M/M/m/m**
- **Poisson Arrivals =>** $P(\text{n arrivals in interval T}) = \frac{(\lambda \mathrm{T})^{\mathrm{n}} e^{-\lambda \mathrm{T}}}{\mathrm{n}!}$
  - **λ = arrival rate in packets/second**
- **Exponential service time =>** $P(\text{service time} < \mathrm{T}) = 1 - \mathrm{e}^{-\mu \mathrm{T}}$
  - **μ = service rate in packets/second**

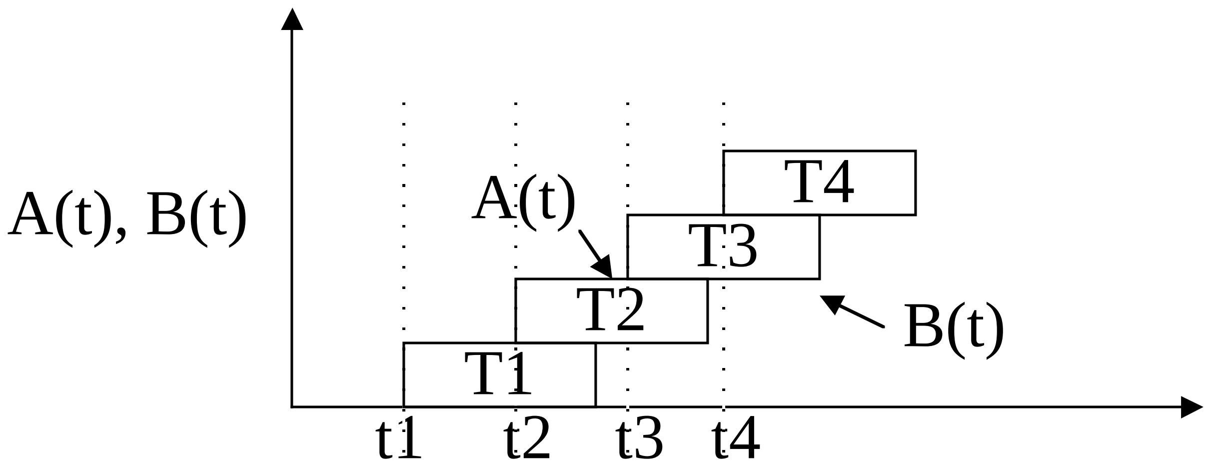
Little's theorem



- N = average number of packets in system
- T = average amount of time a packet spends in the system
- λ = arrival rate of packets into the system
(not necessarily Poisson)

- Little's theorem: $N = \lambda T$
  - Can be applied to entire system or any part of it
  - Crowded system √ long delays
    On a rainy day people drive slowly and roads are more congested!

Proof of Little's Theorem



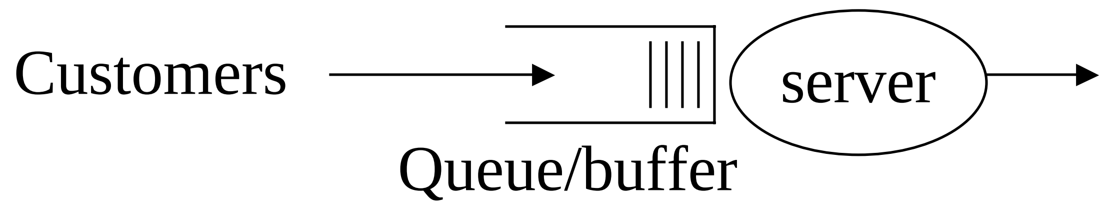
- A(t) = number of arrivals by time t
- B(t) = number of departures by time t
- t_i = arrival time of i^{th} customer
- T_i = amount of time i^{th} customer spends in the system
- N(t) = number of customers in system at time t = A(t) - B(t)

$$N = \lim_{t\to\infty} \frac{\sum_{i=1}^{A(t)} T_i}{t}, \quad T = \lim_{t\to\infty} \frac{\sum_{i=1}^{A(t)} T_i}{A(t)} \Rightarrow \sum_{i=1}^{A(t)} T_i = A(t)T$$

$$N = \frac{\sum_{i=1}^{A(t)} T_i}{t} = \left(\frac{A(t)}{t}\right)\frac{\sum_{i=1}^{A(t)} T_i}{A(t)} = \lambda T$$

Application of little's Theorem

- **Little's Theorem can be applied to almost any system or part of it**
- **Example:**



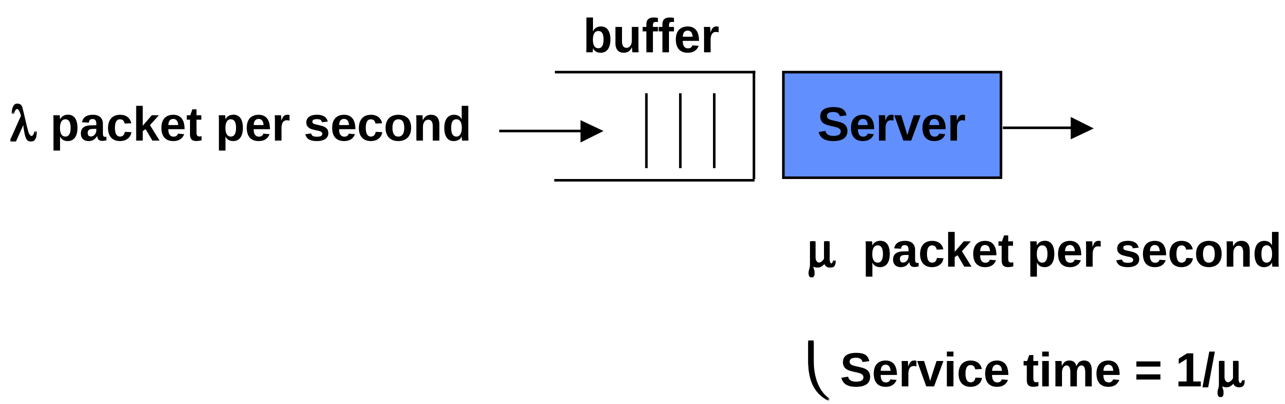
1) The transmitter: D_{TP} = packet transmission time
- **Average number of packets at transmitter = $\lambda D_{TP} = \rho$ = link utilization**

2) The transmission line: D_p = propagation delay
- **Average number of packets in flight = λD_p**

3) The buffer: D_q = average queueing delay
- **Average number of packets in buffer = $N_q = \lambda D_q$**

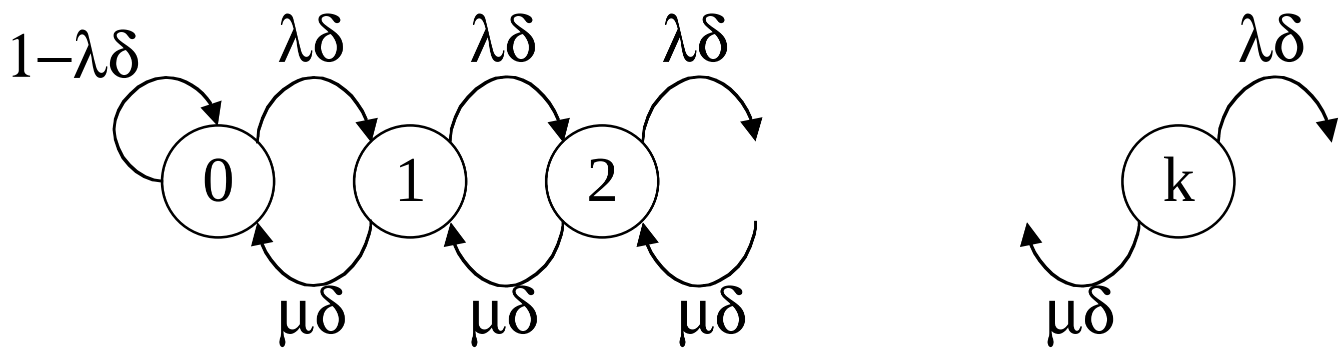
4) Transmitter + buffer
- **Average number of packets = $\rho + N_q$**

Single server queues



- **M/M/1**
  - Poisson arrivals, exponential service times
- **M/G/1**
  - Poisson arrivals, general service times
- **M/D/1**
  - Poisson arrivals, deterministic service times (fixed)

Markov Chain for M/M/1 system



- **State k => k customers in the system**

- **P(I,j) = probability of transition from state I to state j**
  - **As $\delta \Rightarrow 0$, we get:**

    $P(0,0) = 1 - \lambda\delta$, **$P(j,j+1) = \lambda\delta$**

    $P(j,j) = 1 - \lambda\delta - \mu\delta$ **$P(j,j-1) = \mu\delta$**

    P(I,j) = 0 for all other values of I,j.

- **Birth-death chain: Transitions exist only between adjacent states**
  - **$\lambda\delta$, $\mu\delta$ are flow rates between states**

Equilibrium analysis

- **We want to obtain P(n) = the probability of being in state n**
- **At equilibrium $\lambda P(n) = \mu P(n+1)$ for all n**
  - **$P(n+1) = (\lambda/\mu)P(n) = \rho P(n)$, $\rho = \lambda/\mu$**
- **It follows: $P(n) = \rho^n P(0)$**
- **Now by axiom of probability:**

$$\sum_{i=0}^{\infty} P(n) = 1$$

$$\Rightarrow \sum_{i=0}^{\infty} \rho^n P(0) = \frac{P(0)}{1-\rho} = 1$$

$$\Rightarrow P(0) = 1-\rho$$

$$P(n) = \rho^n (1-\rho)$$

Average queue size

$$N = \sum_{n=0}^{\infty} nP(n) = \sum_{n=0}^{\infty} n\rho^n (1-\rho) = \frac{\rho}{1-\rho}$$

$$N = \frac{\rho}{1-\rho} = \frac{\lambda/\mu}{1-\lambda/\mu} = \frac{\lambda}{\mu-\lambda}$$

- **N = Average number of customers in the system**
- **The average amount of time that a customer spends in the system can be obtained from Little's formula ($N=\lambda T \Rightarrow T = N/\lambda$)**

$$T = \frac{1}{\mu-\lambda}$$

- **T includes the queueing delay plus the service time (Service time = D_{TP} = $1/\mu$)**
  - **W = amount of time spent in queue = T - $1/\mu$ =>** $W = \frac{1}{\mu-\lambda} - \frac{1}{\mu}$
- **Finally, the average number of customers in the buffer can be obtained from little's formula**

$$N_Q = \lambda W = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = N - \rho$$

Example (fast food restaurant)

- Customers arrive at a fast food restaurant at a rate of 100 per hour and take 30 seconds to be served.
- How much time do they spend in the restaurant?

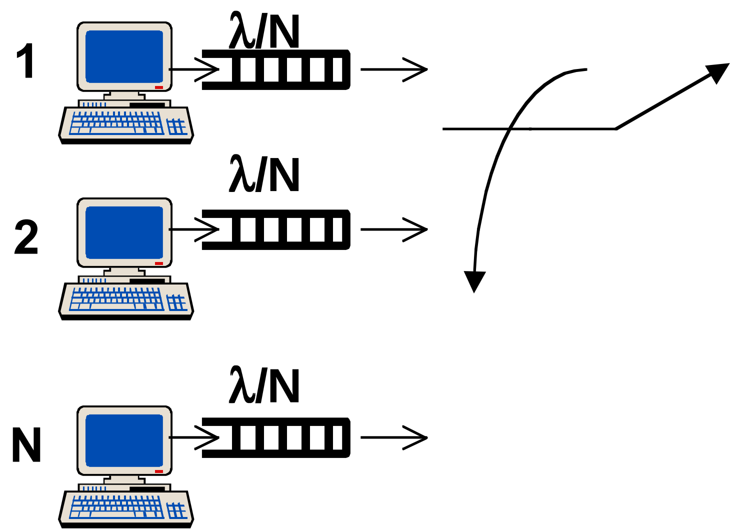
  – Service rate = μ = 60/0.5=120 customers per hour
  – T = $1/\mu - \lambda$ = 1/(120-100) = 1/20 hrs = 3 minutes

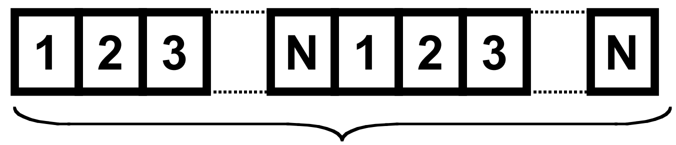
- How much time waiting in line?
  – W = T - $1/\mu$ = 2.5 minutes

- How many customers in the restaurant?
  – N = λT = 5

- What is the server utilization?
  – $\rho = \lambda/\mu$ = 5/6

Packet switching vs. Circuit switching



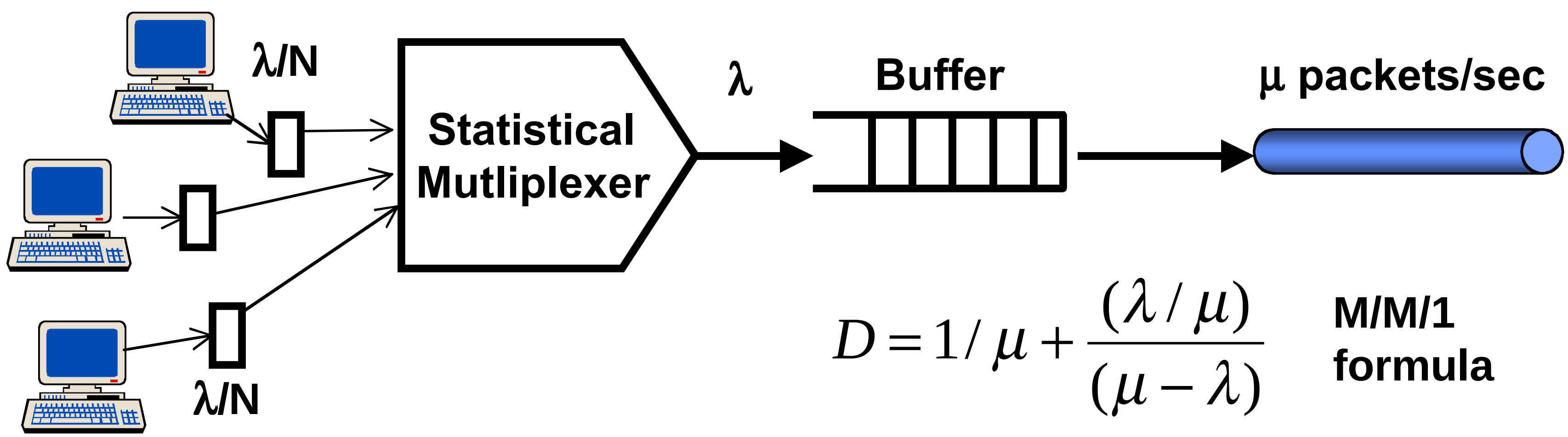


TDM, Time Division Multiplexing
Each user can send μ/N packets/sec and has packet arriving at rate λ/N packets/sec

$$D = N/\mu + \frac{N(\lambda/\mu)}{(\mu - \lambda)}$$

M/M/1 formula

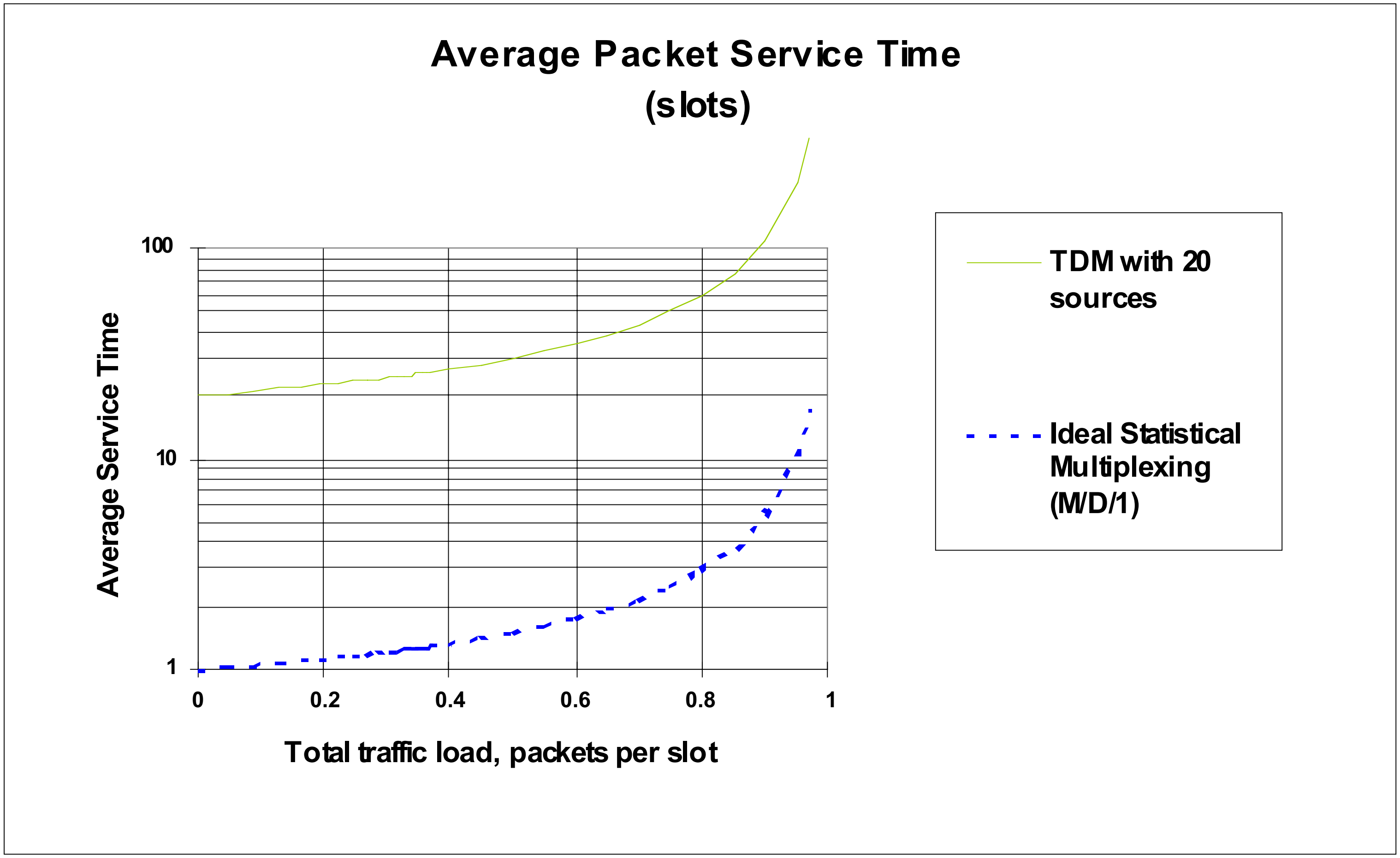
Packets generated at random times



$$D = 1/\mu + \frac{(\lambda/\mu)}{(\mu - \lambda)}$$

M/M/1 formula

Circuit (tdm/fdm) vs. Packet switching



Delay formulas

- **M/G/1**

$$D = \overline{X} + \frac{\lambda \overline{X^2}}{2(1-\lambda/\mu)}$$

- **M/M/1**

$$D = \overline{X} + \frac{\lambda/\mu}{\mu-\lambda}$$

- **M/D/1**

$$D = \overline{X} + \frac{\lambda/\mu}{2(\mu-\lambda)}$$

<u>Delay components:</u>

Service (transmission) time (LHS)

Queueing delay (RHS)

Use Little's Theorem to compute N, the average number of customers in the system

Blocking Probability

- A circuit switched network can be viewed as a Multi-server queueing system
  – Calls are blocked when no servers available - “busy signal”
  – For circuit switched network we are interested in the call blocking probability

- M/G/m/m system
  – Poisson call arrivals and General call duration distribution
  – m servers => m circuits
  – Last m indicated that the system can hold no more than m users

- Erlang B formula
  – Gives the probability that a caller finds all circuits busy

$$P_B = \frac{(\lambda/\mu)^m / m!}{\sum_{n=0}^{m} (\lambda/\mu)^n / n!}$$

Erlang B formula

- Used for sizing transmission line
  - How many circuits does the satellite need to support?
  - The number of circuits is a function of the blocking probability that we can tolerate
    - Systems are designed for a given load predictions and blocking probabilities (typically small)

- Example
  - Arrival rate = 4 calls per minute, average 3 minutes per call
  - How many circuits do we need to provision?
    - Depends on the blocking probability that we can tolerate

Circuits	P_B
20	1%
15	8%
7	30%