On the Economic and Technological Forces Shaping Mobile Transceiver Architecture

by

Christopher A. Aden

Submitted to the System Design and Management Program
in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Engineering and Management

at the

Massachusetts Institute of Technology

February 2009

Signature of Author ____

Christopher A. Aden
System Design and Management Program
February 2002

Certified by ____

John M. Grace
Thesis Supervisor
Industry Co-Director
System Design & Management Program

Certified by ____

Carliss Baldwin
Thesis Advisor
William L. White Professor of Business Administration
Harvard Business School

Certified by ____

Patrick Hale
Director
System Design & Management Program

On the Economic and Technological Forces Shaping Mobile Transceiver Architecture

by
Christopher A. Aden

Submitted to the System Design and Management Program
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Engineering and Management at the Massachusetts Institute of Technology

Abstract

The thesis examines the evolution of mobile transceiver architecture using the management framework pioneered by Carliss Baldwin and Kim Clark. The thesis begins with an introduction and an overview of the wireless communication value network. The author subsequently distills the salient aspects of the Baldwin and Clark management framework predicated on bottleneck analysis, modularity, and return on invested capital. The prominence of bottleneck analysis motivates a technical chapter that summarizes the bottlenecks relevant to all wireless communication systems, namely data rate, error rate, and battery life. A brief chapter discussing the dominant wireless communication network architecture, TDMA and CDMA, corroborates the bottleneck analysis and effectively assigns the error rate and battery life bottlenecks to the handset ODM and supplier layers of the value network. With a clear vision of the competitive bottlenecks, the evolution of transceiver architecture is presented in the context of the aforementioned management framework. Through this analysis, design power is shown to have passed from handset ODMs to integrated circuit suppliers. A noteworthy byproduct of the analysis is the genesis of the bottleneck tree whereby new layers of bottlenecks are emergent upon a firm's selection of a particular design architecture that targets the strategic bottleneck layer. Finally, the thesis is concluded with a summary of the ground covered and the author's opinions of how the architecture may yet evolve and the future nature of the competitive landscape.

Thesis Supervisor: John M. Grace
Industry Co-Director, System Design and Management Program

Table of Contents

1 Introduction

The handset industry has experienced rapid growth and change since the emergence of the Motorola DynaTAC in 1973. By some standards, the wireless communication device, qua cell phone, has become the most pervasive technology since and perhaps eclipsing the automobile. In developed markets, peoples' purchasing requirements for handsets resemble those of automobiles more than conventional telephones. Beyond meeting utilitarian needs of consumers, handsets have become lifestyle enriching, productivity enhancing, and even status symbols – much like automobiles. If polled, most consumers would probably place handsets in the same category occupied by computers as opposed to that of telephones or automobiles; even though the status-quo functionality is inherited from telephone expectations. However, contrary to consumers' intuition, handsets are not personal computers. Deep investigation into the wireless communication bottlenecks illuminate the Pandora's Box consumer electronics firms have opened by marketing new handsets as portable computers. The danger lies in over-hyping computer-like applications that require data rates that exceed the information carrying capacity of wireless channels.

The goal of this thesis is to present a general management framework and industry specific analysis that can be used to establish viable strategies for firms competing in the handset radio market. The goal will be realized by examining specific aspects of the handset industry through the lens developed by Baldwin and Clark in their seminal work Design Rules and subsequent papers on footprint strategies [1,2]. In particular, this thesis will apply the framework consisting of bottleneck analysis, modularity, and return on invested capital (ROIC) to the wireless communication value network in an attempt to explain the evolution of radio architecture and the migration of design power. As a matter of scope, the bulk of the technical collateral will be focused on the analog portion of radio architecture. Although the thesis will attempt to capture the shift in performance bottlenecks owing to the emergence of more capable digital technology, a rigorous treatment of legacy digital applications is conspicuously absent from the thesis.

The paper is organized into a logical sequence of chapters wherein each chapter pulls information from previous chapters in a manner that preserves the natural sequence of the framework presented in this thesis. Chapter 2 establishes the players in the handset value network and presents the business strategies and interdependencies among the four major layers: wireless network operators, handset original design\equipment manufacturers (ODM\OEM), and component suppliers. Chapter 3 develops the management framework by which value in complex systems is identified, through bottleneck analysis; created, through modularity and outsourcing; and realized, by improved ROIC. Chapter 4 provides high level explanations of the bottlenecks unique to wireless communication systems and Chapter 5 discusses the standards that embody the bottlenecks. Chapter 6 explores the contemporary architectures that deliver on the specifications originated by the standards bodies and in doing so exposes new layers of solution specific bottlenecks. Moreover, design structure matrices (DSMs) presented in Chapter 6 connect the modularity-derived strategy from Chapter 3 with contemporary radio architectures. Chapter 7 provides a summary and conclusions.

From Baldwin and Clark, growth by way of maximizing ROIC is the incentive that motivates firms to establish design rules and outsource non-bottleneck components; or components that do not fit within the strategic scope of the firm [2]. The same incentive acts throughout the value network and creates the demand for specialized modules which leads to competition for market share of the legacy and newly defined modules. Over time, the growth of the market can be slowed, if not reversed, as price pressure exerted from up-stream layers forces system integrators to outsource key system designs to lower cost providers.

In the context of the wireless communication industry, price pressure originating from carriers exerted on handset ODMs incentivizes the ODMs to seek lower cost design strategies. As a direct consequence, ownership of the option portfolios comprised of sub-systems are naturally abrogated by the ODM to lower cost suppliers. This trend is explained by a slightly modified version of the net option value of the modular operators derived by Baldwin and Clark [1]. Stated simply, when the cost associated with manufacturing a modular system exceeds the value promised by modular architecture, ownership of the option portfolio is ceded to a firm or group of firms with a more cost effective manufacturing technology. The value of modularity is not necessarily lost to the new firms in possession of the subsystem, provided they foster the antecedents required to maximize modular option value.

Firms such as Motorola and Nokia were market leaders in analog and digital handsets so long as the bottlenecks they chose to address, namely those pertaining to semiconductor manufacturing and waveform engineering, remained at the forefront of enabling wireless communication. As radio technology matured, these firms jettisoned the component business to maximize ROIC and chose to identify and adopt emergent bottlenecks such as operation of higher layers of network architecture, usability, and new modes of data consumption. In addition, new entrants such as Research in Motion and Apple quickly established themselves as leaders along the new bottleneck dimensions. Per the antecedents for modular architecture, component suppliers such as Texas Instruments strengthened their position, albeit in a different market, by extending the relevance of their integrated circuit manufacturing technology and absorbing the functionality of legacy analog components.

Motorola and TI were both behaving rationally by trying to maximize ROIC but to different ends. Motorola successfully reduced its footprint as a handset ODM, in response to mounting price pressure from carriers and growing competition from other ODMs, but in doing so relinquished control of the option portfolio related to the physical layer designs and subsequently, its entitlement in terms of enabling technology. On the other hand, TI improved its ROIC by absorbing functionality that led to higher utilization of capital. Companies like TI increased its footprint in the architecture and actually gained control of the physical layer option portfolio. The remainder of this report will discuss the forces acting on the architecture that created the cost dilemma and catalyzed the shift in “design power”.

2 Handset Value Network & Economics

The Introduction provided an overview of this thesis research. Chapter 2 provides a top down introduction to the entire handset value network spanning wireless service providers, handset ODM's, and component suppliers. The goal of the chapter is to illustrate structure in the value network and highlight the tight coupling of technology and economics that ultimately determines the nature of the competition in the industry. Figure 2-1 illustrates the value network

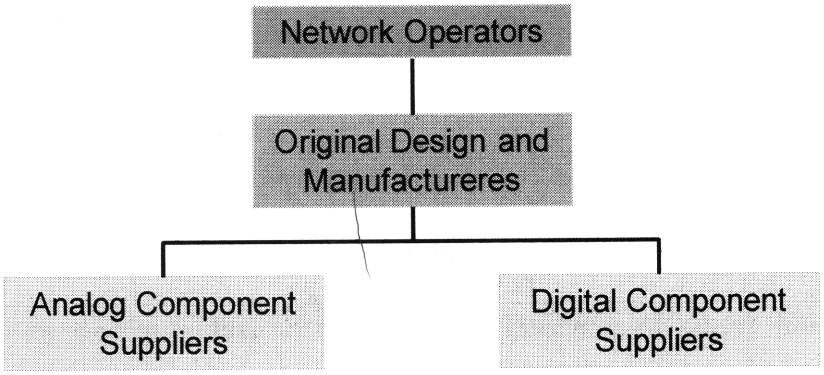

Network Operators
Original Design and Manufactureres
Analog Component Suppliers
Digital Component Suppliers

Figure 2-1: Wireless Handset Value Network.

2.1 Wireless Service Providers

The wireless service providers, otherwise know as network operators or carriers, provide the back-end infrastructure that truly enables wireless communication. The handset, or mobile station, is merely a client device that operates within the constraints of the networks that are designed, deployed, and maintained by the operators. Because the network operators invest the most capital relative to the other layers in the value network, it follows that they also assume the most down-side risk. The risk derives from the operators' uncertainty about whether or not their business models enable them to recoup the investment costs, satisfy debt holders, and meet investors' expectations.

The network operators' ability to make money derives from their control of spectrum. Spectrum in the general sense refers to all frequencies of electromagnetic radiation. In this report, the term spectrum will refer only to the contiguous band of frequencies spanning approximately 500MHz to 10GHz that can effectively support wireless communication. The primacy of spectrum is unique to this industry. Although the physical argument for the primacy of spectrum will be tabled until Chapter 4, for this chapter, the reader need only be concerned with spectrum's scarcity.

Spectrum is a very unique commodity. Its presence is ubiquitous but directly observable only via specially designed test equipment and indirectly via portals that send and receive information carried by the spectrum. What's more, spectrum, like crude oil, is a finite common resource and therefore is subject to the same "tragedies" of other finite resources [3]. Unlike the neo-classical commodities, such as crude oil, over production of spectrum doesn't result in a zero sum gain. Rather, over utilization of spectrum manifests itself as interference that renders all of the effected airwaves un-suitable for communication. To eliminate the possible scenario of over utilization, property rights are sold to network operators by government agencies such as the Federal Communications Commission (FCC) in the United States.

The property rights granted by regional governments apply to specific swaths of spectrum called bands and are strictly enforced so that interference among different bands is avoided. In the United States, prior to 1993 the FCC assigned bands using a lottery system. Speculation and a secondary market for spectrum quickly led to the passage of The Omnibus Budget Reconciliation Act of 1993 which required the government to collect monetary compensation in exchange for spectrum property rights [4]. Since the legislation, sans the lobbying effort of special interest groups, the transactions for spectrum acquisition have taken place via auction format. Owing to its scarcity and value, the auction prices for spectrum often reach into the billions of dollars. For example, in 2008 Verizon Wireless and AT&T spent a combined $16 billion on the United States' 700 MHz block that was recently reclaimed from the legacy television broadcasting application[5].

The huge sums of money necessary to acquire spectrum comprise one of three major capital entry barriers unique to the network operator layer in the value network. The remaining two barriers include network deployment and sales and marketing. Network deployment includes but is not limited to base station installation and maintenance. In this context, base stations refer to the geographically fixed appliances responsible for transmitting and receiving information to and from mobile stations. Each of the base stations is locked to a particular communication standard but is often co-located with other base stations from different networks. Suffice it to say that investments in the infrastructure are non-trivial. For example, between 2005 and 2008, AT&T would have invested over $20 billion in network infrastructure [6]. The sales and marketing cost incurred by network operators are compulsory to stimulate subscription revenue that is necessary to recoup the huge investment in spectrum and infrastructure and satisfy investors and debt holders.

The business models network operators employ to monetize their investments fall into one of two categories, post-paid subscription or pre-paid\pay-as-you go. The subscription based business model generates revenues from usage contracts that generally span several years. The long commitment requires that the network operators maintain credibility and relevance to the available user base. In this industry, credibility derives from performance and reputation. Relevance typically requires that an operator proves that it can meet performance expectations of its target market. Meeting performance expectations requires network operators to work with handset ODMs to optimize the users' experience when consuming information on the network.

The economics of the entire value network is predicated on the service providers' business model. The handset cost to consumers is reduced by subsidies from the service provider to the ODM in return for longer term subscriptions. The subsidy cost incurred by the network operator is then amortized over the life of the subscription. The pre-paid business model is a subtle alternative to generating revenue through long term subscriptions. Network operators that generate pre-paid revenue charge the user a-priori and limit the amount of network usage commensurate with the terms of the prepaid contract. Since the long-term subscription is no longer part of the contract, subsidies to ODMs are lowered if not eliminated altogether, which means the ODMs are no longer sheltered from the economics relating demand with cost.

From the previous paragraph, revenues generated by network operators either derive from a per-unit of service model, such as per kilobyte for data or per minute for voice, or from a subscription based model, or a hybrid of both. In comparison with traditional subscription based business models, there is nothing special about the subscription based model applied to cellular service. Equation (2.1.1) gives the present value of a subscription based revenue model wherein the prospect lifetime value (PLV) is the context dependent proxy for the net present value (NPV) used in financial valuations [7].

$$PLV = CLV - A = \frac{R \cdot P}{1 + d - R} - A \qquad \textbf{(2.1.1)}$$

The PLV is simply the difference between the customer lifetime value (CLV) and the acquisition cost. The CLV is comprised of the profit attributed to each subscriber, *P*, the retention rate, *R*, of the network operator, and a cost of capital, *d*. Network operators typically report churn (*1-R*), as opposed to retention rate, to communicate the health of the business to investors and potential lenders.

The following example illustrates the dramatic effect churn and profitability can have on the subscriber based business model. First assume a network operator generated $7B operating profit from 70 million subscribers. This translates into $100 of profit per subscriber. In turn, suppose the company gains 5 million new subscription based wireless customers. Also assume that the operator spent $600 million in SG&A to acquire the new customers. In this scenario, the acquisition cost incurred by the operator is $120 per new subscriber. Last, assume that a vanilla hurdle rate of %10 is used to discount the cash flows. The PLV for this example is plotted in Figure 2.1-1 as a function of churn.

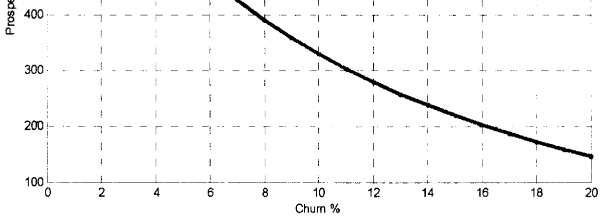


Figure 2.1-1: Operator Subscription Based Model.

Clearly, the network operators' growth and profitability is determined by high profit margin, low churn, and low acquisition costs associated with new subscribers. Owing to the exponential nature of the customer lifetime value (CLV), the prospect lifetime value is dramatically affected by churn. Moreover, a seemingly low level of 6% churn effectively drops the profits of the subscription-only service provider in half. As an aside, carriers typically report profit as the difference between average revenue per user (ARPU) and the incremental operating cost associated with each subscriber.

Carriers have options available to them when it comes to managing the amount of revenue lost to churn. The first option is obvious, an operator can minimize churn by providing differentiated service at a competitive cost, thereby meeting or exceeding consumer expectations. However, differentiation is elusive in an industry when all competitors are equally aware of the performance bottlenecks and of the finite solution space. One alternative is to diversify the business model with the intent to become less sensitive to churn. The illusion of diversity might come in the form of the pre-paid access model. This model is fool's gold because the only major difference is the duration of the contract, and the fact that payment is received ex-ante. By offering more closely spaced decision points to the consumer, the commoditization of differentiated performance is hastened.

One differentiation strategy employed by network operators attempting to minimize churn is to flee the commoditized market for voice services and invest in high performance applications up-market, such as email and web-browsing. In addition to reduced churn, such applications typically pay a double dividend in the form of higher ARPU. The common denominator in the performance of up-market wireless applications is high data rates. Higher data rates require new networks and higher performance handsets which implies some level of collaboration between carriers and ODMs. To mitigate the downside risk of a network launch or major improvement that targets new functionality, the network deployment is accompanied by a fierce marketing campaign complemented by large handset subsidies. For example, in order to monetize its new high speed 3G network in 2007, AT&T invested approximately $16B and $13B billion in cost-of-goods-sold (COGS) and selling, general, and administrative expenses (SG&A), respectively [8]. The former is compared to the $4B of revenue generated from equipment sales. This means that in 2007 AT&T invested approximately $25B in excess of the actual 3G network! Of that, $9B was paid to handset ODMs in the form of subsidies.

To the extent that the market's need for data intensive applications continues to grow, network operators will continue to flee commoditization by providing more capable networks and continue to subsidize higher performance handsets. Markets being over-served by high performance networks and handsets will naturally lead to commodity pricing for network service and the emergence of the low cost handset market. Owing to the price pressure associated with commodity subscriptions and pre-paid voice service, it is unlikely that network operators will offer significant subsidies to handset venders competing in the status-quo handset arena. The primacy of cost owing to the ubiquity of service and lack of subsidies could noticeably alter the handset architecture and the downstream value network.

2.2 Original Design and Manufacturers

Handset manufacturers supply the portals needed to access information residing on the networks erected by the service providers. It is a common practice for handset ODMs to sell a *majority* of handsets to network operators who in turn package the devices with subscription contracts and sell them to users. For example, 95.7% of Research in Motion's revenues in FY2007 came from sales to carriers [9]. By selling products directly to carriers, handset ODM's effectively outsource the sales channel to the wireless service providers. Although relinquishing the distribution channel is a risky endeavor, the subsidies and marketing muscle provided by the carriers make the economics almost impenetrable. Apple once thought that it could sell its popular handset, the iPhone, without network subsidies but reversed its strategy prior to the release of the second generation 3G iPhone.

Foreshadowing Chapter 4, the underlying networks determine the maximum achievable capacity (data or users) of the wireless communication experience. Given that the network operators maintain exclusive ownership of this bottleneck and the handset distribution channels, and that competition in the handset market is fierce; handset ODMs retain very little market power. Like many producers of goods, ODMs' profits are determined by the margins and volume associated with sales. What's more, carriers' demands for device exclusivity coupled with the rise and fall of form factor preference in the market make platform strategies very practical. To the extent that the communication functionality is identical across product families, economies of scale and scope can be realized by device-agnostic physical layer designs within the system level platform architecture[1]. At the system level, exclusion and augmentation is practiced to control of the bill of materials and differentiate, respectively. Again alluding to Chapter 4, it's the physical layer design that operates on the information-bearing spectrum and thus comprises the essential component in a handset design. Stated simply, a handset is not capable of communication without a physical layer design; and therefore all other components are considered secondary in a handset qua communication portal.

The early days of the handset market belonged to the firms that pioneered the technology that enabled portable wireless communication. Firms such as Motorola and Nokia, who pioneered high performance military radios enjoyed early success in the handset market because they already possessed the engineering knowledge to create viable mobile

[1] The physical layer refers to the network protocol that standardizes the transmission and reception of information over the wireless channel.

stations. They were the first movers. Examples of early Motorola and Nokia "handsets" are illustrated in Figure 2.2-1.

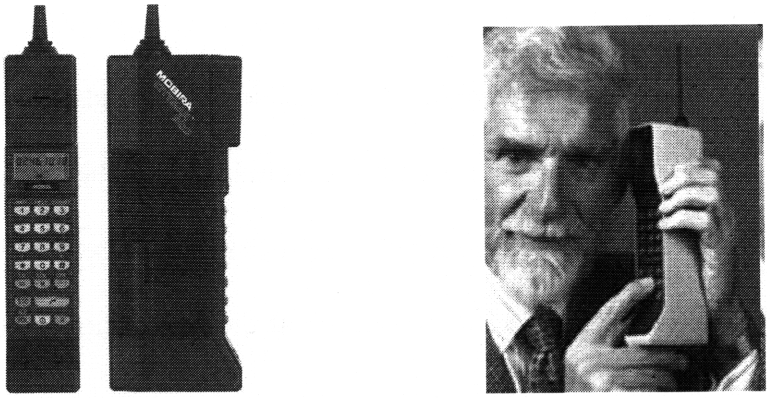


Figure 2.2-1: Nokia Mobira Cityman and Motorola Dynatec. Motorola and Nokia leveraged engineering experience gained from producing military radios to launch early handsets.

Moreover, these firms invested heavily in large design and manufacturing operations to capitalize on their expertise. The following quotation taken from an email dialogue with Dennis Buss, former Vice President of Texas Instruments, illustrates the prevailing strategy followed by legacy handset manufacturers.

> *"Back in the 70's and early 80's, many system companies followed the strategy of "vertical integration". Companies like Motorola felt like they had an advantage because they made communication equipment, and they produced the component ICs. There was very little outsourcing by MOT in these times. As time went on, they found that they were, in fact, at a disadvantage because they had only one supplier of components, and they adopted a strategy where MOT SPS was treated as just one of MOT's many IC suppliers... NOK never had their own IC capability. They did many of their own designs, but they partnered with TI go get them produced."*

What's certain is that as the market continued to grow, the intellectual property (IP) diffused out from the vertically integrated firms such as Motorola and Nokia. Whether by design or attrition, the leaked IP coupled with capital markets and entrepreneurs established the antecedents of a value network capable of supporting modular architectures. The quotation above suggests that handset manufacturers understood the benefits of modularity and restructured their operations accordingly. Over time, a competitive market for components enabled the transition to exclusive outsourcing that allowed firms, such as Motorola and Nokia, to improve the return on assets and subsequently ROIC.

Judging from contemporary architecture, wherein the handset manufacturers' hardware footprint is virtually non-existent; one must conclude that the bottleneck strategy at the handset level has migrated. The emergence of upstarts, such as Research in Motion and Apple, has identified the application specific experience as the new critical bottleneck at the handset layer in the value network. RIM targets email and connectivity in a manner that meets the real-time communication needs of business users while *maximizing network performance*. The uniqueness of the benefits provided by RIM, a handset ODM, to their customers, the service providers, merits deeper investigation. RIM differs from their competitors in that BlackBerry handsets ease the bandwidth constraints that the networks inherit from Shannon's Principle[2]. RIM accomplishes this feat not by cheating the physics, but rather by filtering information and controlling the load that Blackberry users impose on the network. Therefore, RIM helps service providers manage the capacity related bottlenecks while delivering a product that generates high ARPU. Apple meanwhile has targeted the portable internet as the core bottleneck in the user experience paradigm. Apple delivers best-in-class usability by contributing years of software experience to a handset design that is otherwise exclusively outsourced. What remains to be seen is whether or not Apple can continue to deliver high levels of quality in the presence of increased network congestion.

New entrants Apple and RIM cater to a very small percentage of the handset market labeled smart-phones by industry pundits. The lion's share of the market is still owned by

[2] Shannon's Principle will be discussed in detail in Chapter 4.

Nokia, Samsung, and others. Figure 2.2-2 illustrates the breakdown of market share in Western Europe and North America for the dominant handset ODM's. The firms in the figure supply handsets to the network operators competing in the markets for both high performance service as well as commodity voice and data service. Regardless of the end application, cost is the primary determinant of sales in the market for handsets serving also-ran networks. Primacy of cost coupled with the emergence of inexpensive and capable integrated circuit (IC) technology have shifted ownership of the physical layer architecture from ODMs to the component suppliers. If the performance of IC technology continues to improve, then it stands to reason that imminent cost pressure will shift the ownership of all communication enabling designs to the suppliers. ODMs will be forced to adapt and compete for the non-critical subset of the overall handset design.

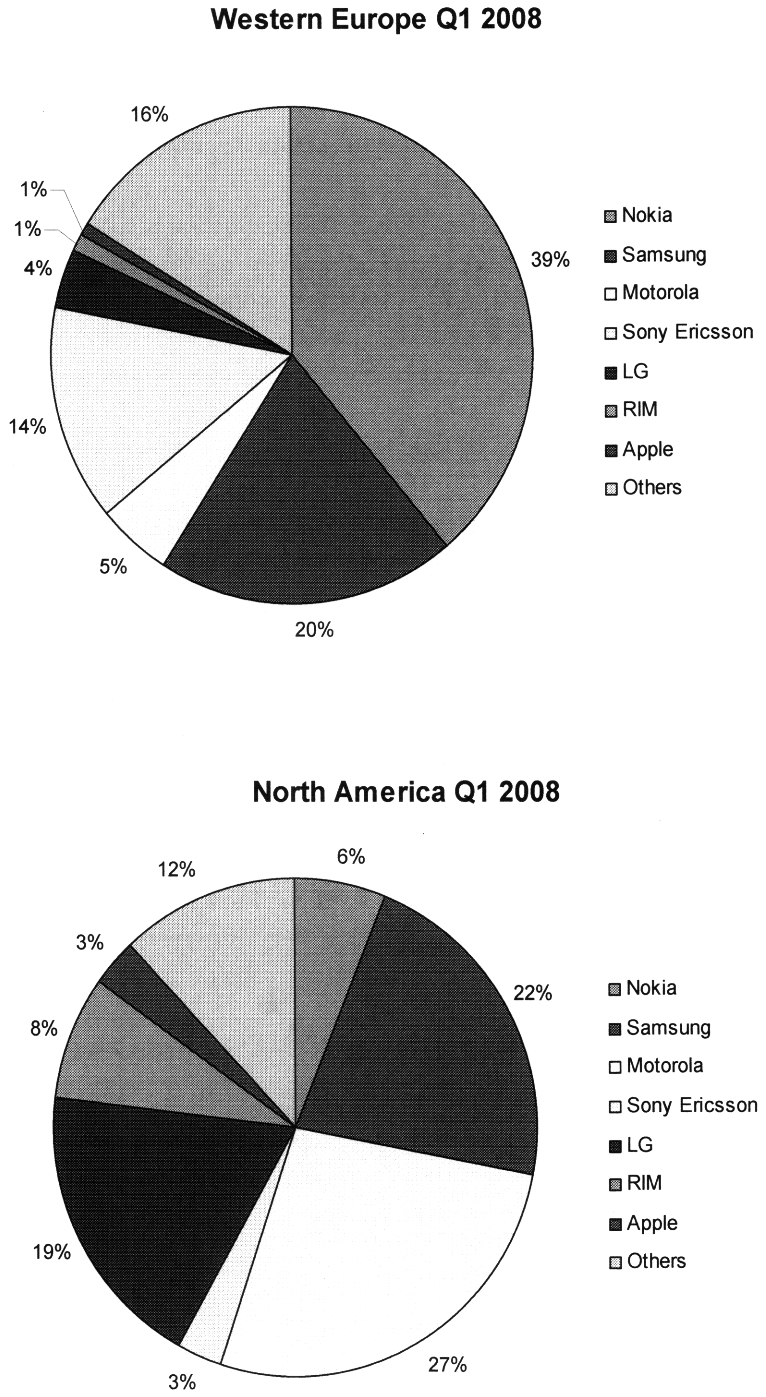


Figure 2.2-2: Market Share Figures [10]

2.3 Component Suppliers

Outsourcing non-bottleneck functions is made possible by the antecedents of modular architecture: well defined design rules and a healthy market for components. To be sure, contemporary handsets deliver functionality that far exceeds the original design for voice communication. As the user expectations and architecture evolved, so too did the need for new component markets and component suppliers. Much of the enhanced functionality, such as location based services and social networking, is the product of augmenting the core design with new applications and design rules that facilitate new modes of information consumption. However, to the extent that the core competencies required to deliver non-perfunctory functionality were never within the purview of handset manufacturers; it is very difficult to argue that handset manufacturers should have invested early in the development of the enabling technologies. What's more relevant to this thesis is how handset manufacturers responded to architectural changes within the core systems they did control. These systems are defined by the physical layer electronics that truly enable wireless communication.

The market for legacy wireless communication components emerged from the option value promised by modularization and the improved return on assets provided by outsourcing. Much like the evolution of computer architecture, handset manufacturers outsourced most if not all of components to suppliers. As standards matured and performance became commoditized, the burden of physical layer ownership began to outstrip its value. In an architecture comprised of many individually packaged integrated circuits, the most obvious path to cost reduction was monolithic integration of multiple functions. However, rigid performance specifications required the implementation of analog functions by specialized transistor technologies that were vastly different than the inexpensive digital IC technologies that existed to meet the digital application needs. For this reason, modular boundaries in the physical layer design emerged between the analog and digital domains. As a consequence, the supplier layer in the value network is split into two complementary groups, analog suppliers and digital suppliers. See Figure 2-1.

The partitioning of the suppliers by domain is determined by physics, economics, and design expertise. The physical limitations of digital transistor technologies justify the existence of analog IC technologies. Firms that serve the analog market identify strategies that rely heavily on technological and knowledge-based competition barriers. Notwithstanding, as their technology becomes more mature, firms that specialize in advanced digital technologies continue to encroach on the analog component suppliers' market with viable low cost alternatives. The firms that specialize in such manufacturing technologies believe that by increasing their footprint in the system architecture they will sell more silicon and improve their return on assets (ROA).

To date, there remain three meta-modules within handsets that enable wireless communication: the digital baseband, the analog RF radio, and the analog RF front end. The markets for these designs are highly competitive and intra-market substitution is the dominant market force within the supplier layer of the value network. Figure 2.3-1 illustrates the modular boundaries in the contemporary handset architecture. Briefly, the digital baseband module provides all of the digital processing functionality required to filter and condition information immediately before\after transmission\reception. This module inherits all of the scale advantages specific to high density CMOS technology; and most importantly the manufacturing cost follows Moore's Law. The analog RF radio is the module that enables wireless communication by mapping low frequency analog waveforms to and from RF waveforms. The reason for the existence of this function will be discussed in detail in Chapter 6. The final enabling module that comprises the architecture is the RF front-end, which is responsible for amplifying RF waveforms and switching among frequency bands. In legacy designs, the technologies employed by each module are mutually exclusive. The physical barriers that define the bottlenecks in the architecture are illustrated in Figure 2.3-2.

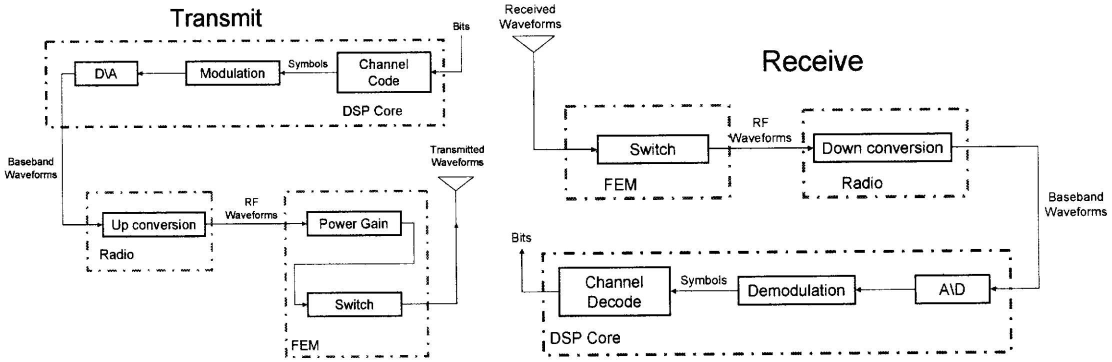


Figure 2.3-1: Modular Boundaries in Contemporary Handset.

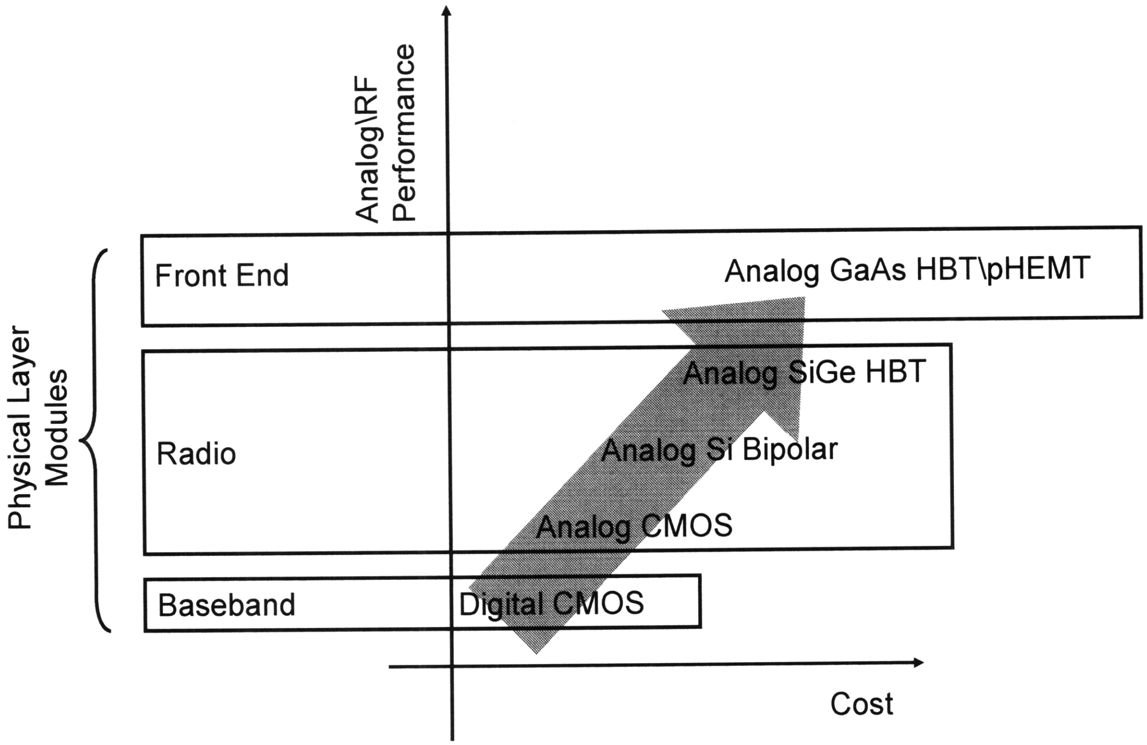


Figure 2.3-2: Technological Composition of Physical Layer Architecture.

3 Bottlenecks, Modularity and ROIC

The previous chapters required the reader to accept at face value the terms bottleneck, return-on-invested-capital, and modularity. This chapter will identify and explore each of these concepts with the intent to establish inter-relationships and relevance to competitive strategy. In as much, all three terms are introduced independently and then collapsed into a strategic framework that can then be applied to the wireless communication industry. The organization of the chapter is as follows. First, the concept of bottleneck analysis is introduced and explored. Next, the six core modular operators are reviewed to establish the dimensions along which firms can choose to modularize their product architecture or identify new market opportunities. Finally, the focus of the discussion is shifted toward implementation of a framework to improve ROIC and realize a successful competitive strategy.

3.1 Bottleneck Analysis

Firms compete for market share by committing to a business strategy that matches the firms' strengths with market opportunities. This thesis builds on the strategies proposed by Baldwin and Clark whereby businesses gain a competitive advantage by investing in architectural knowledge [2]. To be more specific, investment in architectural knowledge provides firms a deep appreciation for absolute and relative bottlenecks that impact a particular design. Awareness of both bottlenecks and a firm's strengths can provide management a guide to maximize growth of their market share. In this context, absolute bottlenecks are defined as those that limit the overall system performance [2]. The mathematical equivalent to the previous statement is presented in Equation (3.1.1). In (3.1.1), the x_i terms refer to generic performance bottlenecks.

$$X = \min(x_1,...x_n) \qquad \textbf{(3.1.1)}$$

Relative bottlenecks identify shortcomings of a product architecture that, in aggregate, affect a particular system performance metric [2]. Equation (3.1.2) describes the effect of relative bottlenecks on performance metric, X.

$$X = \sum x_i \qquad \textbf{(3.1.2)}$$

The perception of knowledge about a system's bottlenecks provides a firm with a context dependent map that matches competencies, or investments, with value-related attributes of a system. Armed with this knowledge, a firm could well choose to focus on a set of bottlenecks while outsourcing those that aren't deemed strategic enough to address internally. Outsourcing enables the firm to shed its commitment to capital and improve its ROIC. The last section in this chapter will show that, in a competitive market, the firms with the highest ROIC will always outperform the competition.

Returning to the bottleneck Equations (3.1.1) and (3.1.2), the implications of each class of bottlenecks are similar. In both cases, firms target utility of systems to differentiate themselves from the competition. In some cases, the impact of relative bottlenecks on a system's performance me be dwarfed by the absolute system bottleneck. In this scenario, firms focused on relative bottlenecks will likely compete exclusively on cost. In the case where the system performance is dominated by the presence of an absolute bottleneck, and the system performance is below market expectations; then firms enter the market focused on the absolute bottleneck hoping to improve utility and charge a premium. In both cases, an efficient market will attract capital and competitive firms. Chapter 4 will explore the bottlenecks pertaining to wireless communication systems. Chapter 6 will investigate the designs that target the bottlenecks that are introduced in Chapter 4 and as a consequence introduce new design specific bottlenecks.

Unless a firm is vertically integrated, it must either act as a supplier to system integrators or act as a system integrator itself. Bottleneck analysis can be used to determine which paradigm is right for a firm. If a firm possesses pertinent architectural knowledge about a system, and by virtue of this knowledge it decides that the system integrator role is most strategic; then it must determine which aspects of the system to produce internally, if any, and which to outsource. Likewise, a firm that is best suited to fill the supplier role can use bottleneck analysis to determine the components to supply and more importantly a competitive strategy that will provide an intra-competition advantage and protect the firm from aggressive system integrators or downstream suppliers attempting to migrate upstream. In this context, firms will divest themselves of capital addressing bottlenecks that are outside their strategic scope.

Outsourcing functions of a larger system is made possible by modular architecture. The term module is defined by Baldwin and Clark in the following excerpt from Design Rules [1]. “A module is a unit whose structural elements are powerfully connected among themselves and relatively weakly connected to elements in other units. Clearly there are degrees of connection, thus there are gradations of modularity.” In particular, a modular architecture embodies the Baldwin and Clark definition by comprising of groups of loosely connected modules that contribute uniquely to system performance. The connections, or interfaces, provide strict specification whereby one module can be substituted by another. An architecture is modular to the extent that these interfaces can be established.

The concept of modularity naturally leads to two important terms that are used to describe modular architecture: information hiding and complexity. Information hiding is embodied by the "black box" model used to describe modular architecture. The principle behind information hiding is independent of perspective. In a hidden module, the designer of the module need not be bothered by the larger system design parameters, and the system designer need not worry about inner workings of the hidden module. On the contrary, visible modules are "seen" by many different modules within the system; for that reason they are coined architectural modules by Baldwin and Clark [1][2]. A term that is extremely important in the discussion of modular architecture is complexity. Complexity refers to the number of tasks attributed to a particular module or system. Modules that accomplish more tasks are said to be more complex.

Modularity provides multiple benefits to an economy and a firm. From an operational point of view, a truly modular architecture enables efficiency through the division of labor and the emergence of skilled specialists; where each specialist focuses on a particular module. In situations where inherent value exists, the specialists can contribute to the growth of an economy by engaging capital markets for the funds necessary to start new ventures. Assessing value is the domain of the venture capitalists. The amount of value deriving from modular architecture is determined by the particular modular operator that is projected on system. The value that modularity creates within firm boundaries pertains to firms' outsourcing strategies. Without the thin interfaces of modular architecture, whereby very little coupling exist among modules, agency and transaction cost can render outsourcing inefficient [1]. On the contrary, well defined modular boundaries entice firms to practice highly developed outsourcing strategies to improve the bottom line.

Per the introduction paragraph, the contributions of modularity to firms and economies will be discussed by way of the six modular operators developed by Baldwin and Clark [1]. The chapter will conclude with the discussion of ROIC as a proxy for growth and enabled by the modular operators. The modular operators covered in the next three sections are listed below:

- Splitting: breaking the design and tasks into modules
- Substitution: replacing one module with another
- Augmentation: adding a new module to the system
- Exclusion: removing a module from the system
- Inversion: tearing functionality away from existing modules and folding it into others
- Porting: reusing a core module for different application\or different system

Note, the discussion of the operators follows closely the treatment by Baldwin and Clark, albeit in lesser detail. The interested reader is strongly encouraged to read the Baldwin and Clark text, Design Rules [1].

3.2 Splitting and Substitution

The inherent value created in a modular architecture is closely related to financial options. In the financial industry, the owner of an option has the right, not the obligation, to exercise [11]. In its simplest form, a financial option provides a floor to the down side risk assumed by the owner of the option by allowing the purchase transaction to occur ex-post performance of the security or portfolio. In simpler terms, the owner of an option elects to exercise (purchase) if the option value is observably positive. The value associated with modular architecture is similar. In the context of modularity, a firm that fills the role of system integrator elects to include a particular module only after it is shown to outperform the incumbent or if the module provides additional value to the system. Accordingly, splitting and substitution are the most natural extensions of modular value creation. What's more, much of the analytical machinery derived from the splitting and substitution operators are portable to other modular operators.

Splitting refers to the act of breaking the dependencies within a system. The antecedent to splitting is the existence of modular boundaries, or thin interfaces. To this end, splitting is merely a cost center and a means to an end. Interestingly enough, splitting qua splitting is the only requirement for outsourcing and cost reduction. However, a firm that possesses the power to split a design may also be able to benefit from the emergence of a market for components that have been "split out" of the firm. These benefits refer to the value created by the substitution operator. The splitting and substitution operators are illustrated in Figure 3.2-1. In the figure, the upper-most integrated architecture is split into independent modules. The interfaces among these modules are specified by the system architect or by a standards body. What's more, because the interfaces are well defined and well known, a market for substitutes emerges, independent of the integrator.

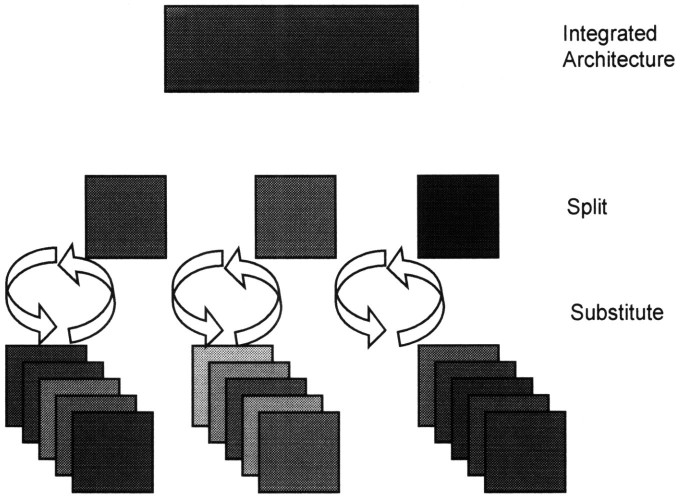


Figure 3.2-1: Splitting and Substitution

The value created from the splitting and substitution operators derive from the value inherent in a portfolio of options. For the purpose of illustration, assume that an integrated architecture is split symmetrically. In other words, the architecture that functions to accomplish N tasks is divided into j modules of equal complexity. Assume also that the technical potential embodied by the variance of each task is identical. Under these conditions, the value created by a modular architecture is equal to the integrated system value, S_0, plus any contribution due to modular splitting and substitution. See Equation (3.2.1). For the time being, ignore the subscript k - it will be made relevant later in the section.

$$S = S_0 + NOV(j,k) \qquad \textbf{(3.2.1)}$$

The net option value (NOV) derives from the right, not obligation to exercise and includes both value created by the option as well as the cost incurred by the system architect to expose the option.

A designer reserves the right to replace a module only after it outperforms the incumbent. To that end, the downside risk from the option holder's point of view is completely eliminated. Figure 3.2-2 illustrates option value for a module described by a standard normal distribution. The expected value of this module is given by: $\int_0^{\infty} v_i p(v_i) dv_i = 0.3989$.

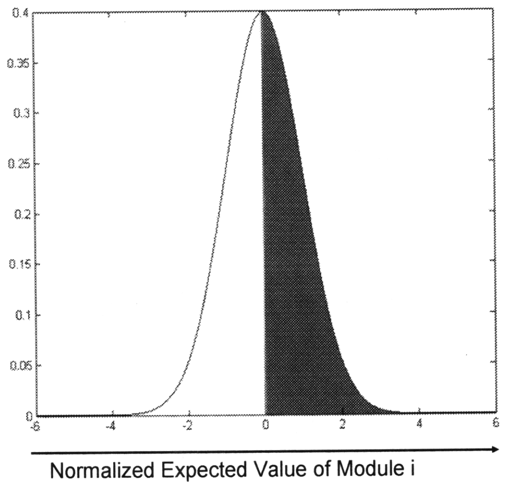


Figure 3.2-2: Modular Option Value [1].

In a system comprised of many well defined modules, the total value due to splitting and substitution is given by Equation (3.2.2).

$$V = \sum_{i=1}^{j} E(v_i) \tag{3.2.2}$$

Each v_i, in (3.2.2) is determined by the technical potential and complexity associated with the corresponding module. From basic statistics, the variance of the sum of identically distributed random variables is the product of the number of variables and the underlying variance [12]. In the context of the symmetric module architecture, each module's variance is given in Equation (3.2.3) [1].

$$\sigma_v{}^2 = \frac{N}{j}\sigma^2 \tag{3.2.3}$$

From the equation, the variance is associated with the technical potential of the symmetric tasks and the module's complexity. The expected value of each module, assuming a normal distribution similar to that presented in Figure 3.1-2, is calculated by first normalizing the random variable: $E(z_i) = \frac{E(v_i)}{\sigma_v}$. Combining terms yields the option value for each module and is presented in Equation (3.2.4).

$$E(v_i) = \sigma_v E(z_i) = 0.3989\sqrt{\frac{N}{j}}\sigma \qquad \textbf{(3.2.4)}$$

Combining Equation (3.2.4) with Equation (3.2.2) yields the net value determined by the option portfolio.

$$V = 0.3989\sqrt{jN}\sigma \qquad \textbf{(3.2.5)}$$

Equation (3.2.5) confirms a well-know property of financial options. Namely, a portfolio of options is more valuable than an option on a portfolio of assets. In this context, the value of the portfolio of options is $\sqrt{j}$ times that of the option on an integrated system. The NOV of the modularized architecture is realized by adding, in ad-hoc fashion, the cost associated with modularization. The expression for the symmetric modular system is given in Equation (3.2.6). The cost terms include those associated with developing the modular interfaces, c_j, the cost associated with testing each module, c_k, and the system level testing, $T(j,k)$.

$$NOV(j,k) = E(z_i)\sigma\sqrt{Nj} - c_j j - c_k k - T(j,k) \qquad \textbf{(3.2.6)}$$

It can be inferred from Equation (3.2.6) that the index k is associated with the testing of modules. In fact, k enumerates the number of different experiments for each module. Various experiments per module not only introduce cost, but also change the normalized distribution function according to Equation (3.2.7) [1].

$$E(z_i) = Q(k) = k\int_0^\infty v_i P(v_i)^{k-1} p(v_i) dv_i \qquad \textbf{(3.2.7)}$$

Figure 3.2-3 illustrates the affect that multiple experiments can have on option value. From the figure, k=1 corresponds to the area of the shaded region in Figure 3.2-2. Notice how the rate by which experiments improve option value of a particular module diminish with each successive experiment. Combining Equation (3.2.6) and (3.2.7) gives the corrected option value complete with the benefits of experimentation. See Equation (3.2.8).

$$NOV(j,k) = Q(k)\sigma\sqrt{Nj} - c_j j - c_k k - T(j,k) \tag{3.2.8}$$

Note, in (3.2.8) the symmetry argument is extended to include experiments per module.

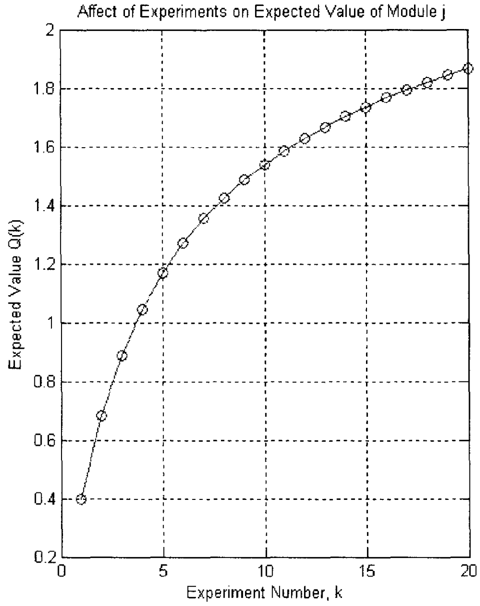


Figure 3.2-3: Affect of experiments on expected option value

Close examination of Equation (3.2.8) reveals a host of relevant insight. Assuming the number of experiments is fixed at one, net option value is positive provided the cost associated with splitting the architecture is not prohibitive. To the extent that the cost of splitting is incurred only during the modularization phase, then it is sunk for each epoch beyond the first when options can be exercised. However, if variable costs such as those associated with manufacturing accompany a modularized architecture, then the cost is carried throughout the life of the design. Indeed targeting the manufacturing cost could very well fit into the bottleneck analysis introduced in the beginning of this chapter. Also interesting is the competing costs associated with experimentation. Assume for a moment that the cost of experimentation corresponds to the cost of capital and resources required to complete a design. In this scenario, cost scales linearly with *k* and will eventually dominate the value added by *Q(k)*. Subsequently, an optimal number of experiments must exist for each combination of cost and system technical potential. Last but not least is the cost associated with test and measurement of systems. These tests can be prohibitive provided blind prototypes are the only means of evaluating system performance. Fortunately, the advancement of simulation software has eliminated all but the final few prototyping iterations [13].

Although concise and illuminating, Equation (3.2.8) has limited usefulness to a firm because of the symmetry constraints. Equation (3.2.9) implicitly drops the symmetry assumption and that is visible by Equation (3.2.10).

$$S = S_0 + \sum_{s=1}^{j}\left(NOV_s - c_{MFG,s}\right) - c_{YIELD}(j) \qquad \textbf{(3.2.9)}$$

$$NOV_s = \max\left(\sigma_s \sqrt{n_s} Q(k_s)\right) - c_{k,s}(n_s)k_s - Z_s \qquad \textbf{(3.2.10)}$$

These equations mark the first departure from the Baldwin and Clark framework wherein a manufacturing cost and yield penalty is incurred by the owner of the option portfolio. The manufacturing cost is determined by the physical and logistic challenges inherent in assembling multi-module architecture. The yield penalty reflects the percentage of defects per batch of manufactured systems. Owing to the random variable nature of defects, like the option value, the cost will be a stochastic function and will depend on defect densities. As will be shown in Chapter 6, the additional manufacturing cost supports the migration of module ownership to the level in the value network with the lowest manufacturing-cost structure.

The levers whereby firms can control modular value are exposed in Equation (3.2.10). Namely, technical potential, σ_s, complexity, n_s, and visibility, Z_s, determine the realizable option value inherent in a particular module. The term c_j does not appear in Equation (3.2.10) because it is immediately sunk after the first iteration and as such does not contribute to the marginal cost of production [14]. One of the most profound implications of (3.2.10), and indeed modularity, is that it encourages firms to place risky bets on modules that exhibit high technical uncertainty $(NOV_S \propto \sigma_S)$. Restated, it is rational to experiment with the difficult tasks that are safeguarded in an integral architecture [1]. The visibility term is unique to also new to the formulation (3.2.10). A simple explanation for Z_s is that cost is incurred for experimenting with modules that force the re-design of others. Baldwin and Clark show that the visibility cost can be prohibitive even when technical uncertainty is high [1].

To wrap up, designs that provide for splitting and substitution via modular architecture were shown to offer distinct performance advantages when compared to otherwise identical integral architectures. The operators create value by eliminating the downside risk of un-fruitful design iterations. Value was shown to grow proportionally with the number of experiments, or options, for each module. Not withstanding, modularity is not free; and under some scenarios, the cost associated modular architecture may be prohibitive. Last but not least, by relaxing the symmetry assumption, firms can determine which modules offer the highest returns and subsequently where to invest capital.

3.3 Augmenting and Excluding

As the term suggests, augmenting simply means adding a new module where one did not previously exist. Exclusion is a complementary operator that provides for easy augmentation. As Baldwin and Clark suggest, these operators are difficult to model because they create, or at least recognize, new peaks in the "value landscape". This is problematic because the strike price for the option is un-known a priori. Nonetheless, this section will show by way of tacit examples and recycled models that augmentation and exclusion add value to a modular system. Both augmentation and exclusion are illustrated in Figure 3.3-1.

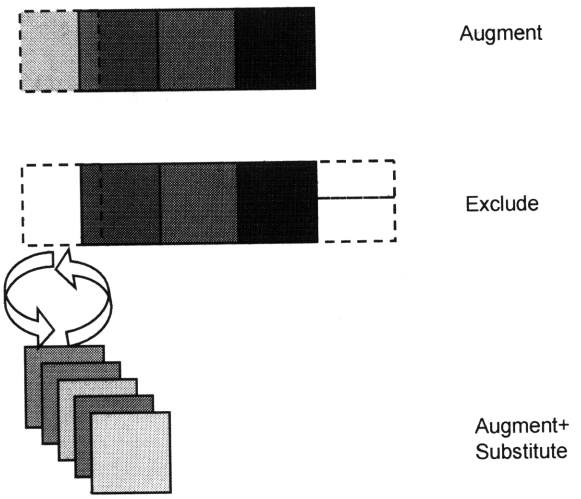


Figure 3.3-1: Augmenting and Excluding

The baseline in Figure 3.3-1 is a modular system. The upper portion of the figure illustrates two important features. First a new green module has been added to the original three. Second, the new module is visible to the pre-existing red module. Under these circumstances, the red module will have to be redesigned which incurs a cost. The corresponding net option value for augmentation is presented in Equation (3.3.1).

$$NOV_a = \max\left(\sigma_a \sqrt{n_a} Q(k_a)\right) - c_{k,a}(n_a)k_a - Z_a \qquad \textbf{(3.3.1)}$$

The NOV for an augmented exactly replicates the NOV from Section 3.1 which presumes the new module is subject to substitution. To reiterate, the challenge in modeling the augmentation operator is associated with the baseline system value, S_0. Typically, the value contributed to S_0 by adding functionality exceeds the marginal value derived from modularity. To the extent that firms can reserve the right to augment a system without redesign accentuates the value of the augmentation operator when paired with the exclusion operator.

The term exclusion, like augmentation, completely describes the operator. Exclusion is illustrated in the second row in Figure 3.3-1 by the dashed placeholders. In fact, it's the exclusion operator that provides a firm the real-option to augment a design. Systems that are designed using the exclusion operator fall into the category of designs called platforms. Platforms enable firms to minimize cost for a family of products designed for different markets by manufacturing the salient features expected by all markets and excluding the features that are market specific. Platform architectures employing exclusion improve a firm's operating profit by adding economies of scope while diversifying the market risk. If a system was designed to serve multiple markets by way of exclusion, the augmentation operator is applied ex-post to customize the platform for each specific market. In comparison, all of the aforementioned benefits are lost to the firm that designs integral systems to serve each market. The downside risk is exacerbated if functionality for an uncertain market is embedded in an integral design.

The previous paragraph suggests that the exclusion and augmentation operators require a different level of abstraction prior to modeling. Whereas the splitting and substitution operators assume a static task count, the augmentation and exclusion operators establish a design with a variable task count. This subtle feature obfuscates the ability to compare on the virtues of modularity alone because the one-to-one comparison between an integral and modular architecture for an established number of tasks collapses into the realm of substitution. Baldwin and Clark sidestep this anomaly by proposing a model that is predicated on real options [1]. To summarize, the value of exclusion is captured by a model wherein designers test the market prior to development and use the cursory experiment to make the go, no-go decision on the system as a whole. The model allows designers to experiment with any number of modules, present and future, in an attempt to gauge market acceptance. If the market tests do not support the proposed design evolution, then the design is abandoned. If on the other hand market tests reveal that a particular design can meet both present and future needs, then designers can architect a system that accommodates the design evolution by way of exclusion and augmentation. The Baldwin and Clark model is captured in Equation (3.3.2).

$$\begin{aligned} NOV_{SYS} &= \max\left(V_{SYSTEM} - C_{Design}, 0\right) \\ &\Rightarrow (\text{Proceed, Abandon}) \\ &= \max\left(S_0 + \sum NOV_s + \sum NOV_a - C_{Design}, 0\right) \\ &\Rightarrow (\text{Proceed, Abandon}) \end{aligned} \qquad \textbf{(3.3.2)}$$

The following example illustrates the value of exclusion and augmentation in the context of handset architecture. Suppose that a carrier has uncertain plans to deploy a high speed network within the next two years. To support this and other carriers, a handset ODM is beginning to design a new product that addresses the emerging market for social networking. Nonetheless, the handset ODM knows that if the new network is deployed then it will have to redesign the physical layer to include a new high performance radio chipset;

otherwise the handset will fall out of favor. In order to prevent massive redesign of their product in the wake of the new network deployment, the handset ODM designs the printed circuit board and chooses a baseband processor to support both the legacy and the proposed network. The designers purposely exclude the higher performance radio in the initial launch to save cost; but reserve the right to augment the design later by adding the new radio chip.

3.4 Inverting and Porting

Inversion and porting offer improved efficiencies to already modular architectures by reducing redundancies. Inversion effectively tears redundant functionality away from modules and creates a separate architectural module possessing the expropriated functionality. The new module is then referenced by the affected hidden modules. The inversion operator is illustrated in Figure 3.4-1.

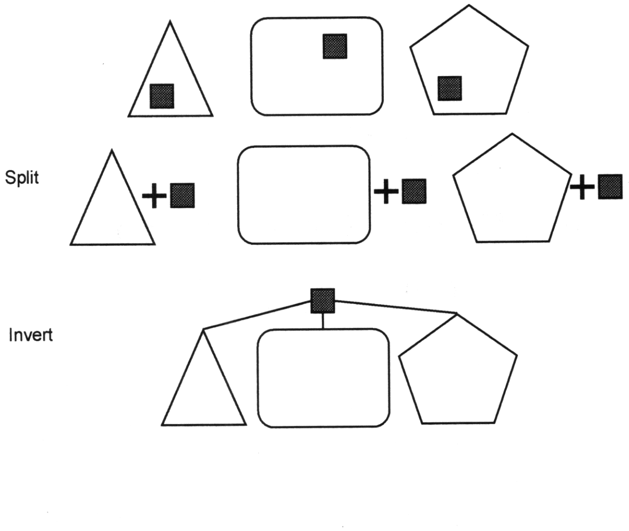


Figure 3.4-1: Inversion

From the figure, the obvious sign that a design would benefit from inversion is the presence of redundancy in hidden modules. The inversion operator simply identifies the redundancy, splits it from the hidden modules, and re-maps the architecture to allow the hidden modules to reference the new architectural model.

The net option value attributed to the inversion operator is given in Equation (3.4.1).

$$NOV_{INV} = V_{INV}(j,k) - C_{INV}(j,k) - \sigma_{INV}\sqrt{n}Q(m) + C(n)m - Z_{INV}(m) \qquad \textbf{(3.4.1)}$$

The indices j and k refer to the symmetric module introduced in Section 3.1; m is the number of hidden modules that once included the inverted functionality; and n is the number of tasks corresponding to the old functionality. The first term captures the option value inherent in the architectural module that is being inverted during the process. If the architectural module is symmetric than its option value is determined by j sub-modules requiring k experiments. In logical fashion, the second term suggests that the cost of designing the architectural model is commensurate with the design parameters j and k. The third term represents the lower bound on the value destroyed by eliminating the tasks from the hidden modules. In exchange for the value destroyed by inversion, the cost associated with experimenting with each hidden module is recouped in the fourth term. The last term captures the cost to redesign each of the hidden modules that now reference the new architectural module.

An example of inversion in handset architecture is the analog to digital conversion function in a multi-mode handset. The digital signal processor, as its name suggests, operates on discrete signals coded by bits of information. On the contrary, analog waveforms carry information across the wireless channel. Subsequently, the analog waveforms must first be converted to their digital equivalent prior to digital signal processing. Rather than implementing an analog to digital converter on the back end of each receiver, it is inverted and integrated with the DSP chip. As a consequence, the receiver need only pass a specification compliant analog waveform to the once-instantiated analog to digital converter.

The porting operator behaves in a manner similar to the inversion operator. The fundamental difference between inversion and porting is that porting allows for similar, not identical, functionality to be reused in a design. Porting comes in handy when users wish to add a layer of abstraction on top of an already sound foundation. The example used by Baldwin and Clark refers to high level programming languages, such as C, that make use of assembly language modules [1] that are operating system (OS) dependant. Because the functionality split from each relevant module is unique, the porting operator must include a translator to make the output of the architectural module compliant. In the programming example, the translator is called a compiler. Figure 3.4-2 illustrates the porting operator.

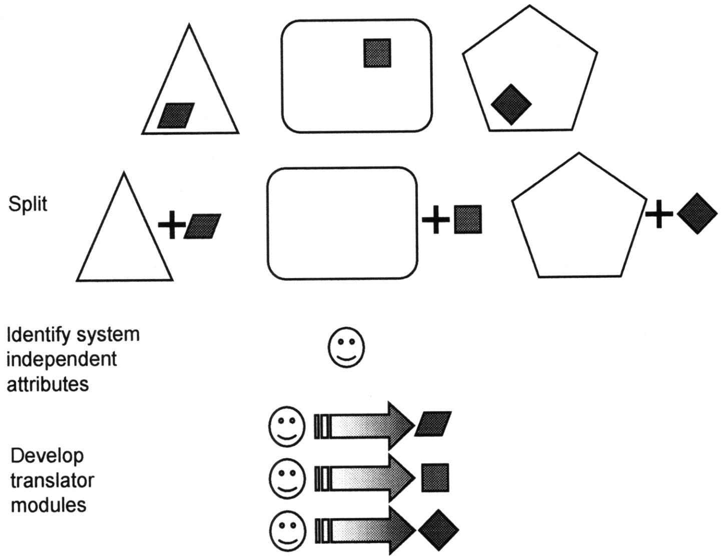

Split
Identify system independent attributes
Develop translator modules

Figure 3.4-2: Porting

Figure 3.4-2 depicts three separate modules containing three separate sub-modules. From the previous discussion, the sub-modules need only belong to the same class of functions in the sense that they serve each module in a similar fashion. The porting operator requires the designer to split the sub-module out of each design, create a common representation, and then develop design specific translators for each design. In the figure, the small shapes represent different modules (different OS's) belonging to the same class (OS), the smiley represents the common representation (C++), and the colored arrows represent the context dependant translators (compiler).

To the extent that a single architectural module resembles an inverted module, its option value is captured by equation (3.4.1). However, if the module is made portable among a multitude (*M*) of different systems, then the module realizes additional option value given by Equation (3.4.2).

$$NOV_{PORT} = V(M,S) + (M-1)C_{SAVE} - MC_{TRANS} - C_{BOOT} \qquad \textbf{(3.4.2)}$$

The composition of Equation (3.4.2) is different from NOV expressions corresponding to the previous five operators. The first term represents the option value of switching among systems M in the presence of switch cost S. The second term gives the cost savings associated with not having to re-invent the ported module for all but the design that is targeted for the initial prototype. The third term represents the cost associated with the design effort required to write separate translators. The final term C_{BOOT} captures the cost associated with finding the agnostic representation that is capable of spanning the class of systems that will employ the ported module.

3.5 Multiplication of Modular Operators

The last three sections provided an introduction to the modular operators defined by Baldwin and Clark in Design Rules [1]. Each of the operators was shown to create value when applied to the design of an artifact. When the collection of operators is applied throughout the hierarchy of a design, the modular operators multiply value. Figure 3.5-1 illustrates the effect modularity can have on the design of an artifact. In the figure, each of the modular operators is applied to the once-integrated system. The compounding nature of modularity is observable in the figure as operators are applied to modules, sub-modules, and disparate designs.

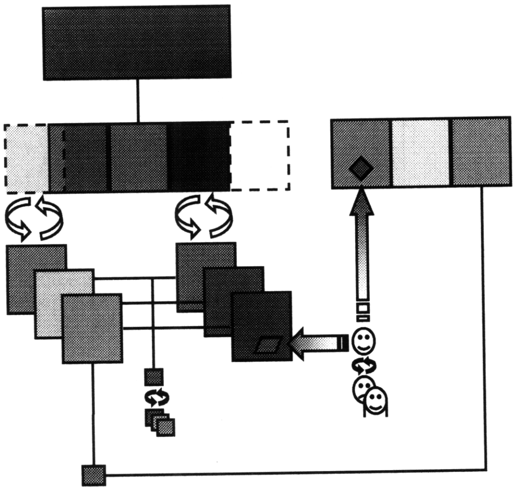


Figure 3.5-1: Multiplicative property of modular operators

3.6 Return on Invested Capital and Growth

The remainder of this chapter will introduce the financial machinery that substantiates the claim that successful firms enjoy higher return on invested capital than their competition. Like the discussion devoted to bottlenecks and modularity, this work is paraphrased from the work of Baldwin and Clark [2]. As an aside, the completeness and rigor of the Baldwin and Clark frameworks summarized in this report separate them from other strategic frameworks wherein mental models seek to substitute deep thought and analysis with simple intuition. It is the author's opinion that the aforementioned intuition based frameworks complement the Baldwin and Clark models - which present a more rigorous standard.

Return on invested capital is a simple accounting ratio that is often subjugated by analysts and investors by return on equity (ROE) or other context dependant ratios [15]. See Equation (3.6.1).

$$ROIC = \frac{\text{Net Income}}{\text{Invested Capital}} \qquad \textbf{(3.6.1)}$$

Although ROE provides investors a useful financial abstraction, it does little to predict the evolution of market share in a competitive market. To the long term investor, a firm's ability to win market share and improve profitability determines the worth of shares he or she holds. To that extent, growth and profit margin reign supreme over abstract ratios. Equation (3.6.2) gives the simple expression for a single firm's economic growth.

$$G = \frac{\Delta Q}{Q} \qquad \textbf{(3.6.2)}$$

Assuming the ratio of invested capital to production is fixed, i.e. that there has been no new infusion of technology, the change in production is determined by the asset turnover ratio and the change in invested capital.

$$\kappa = \frac{Q}{C_I} \qquad \textbf{(3.6.3)}$$

$$\Delta Q = \kappa \cdot \Delta C_I \qquad \textbf{(3.6.4)}$$

The maximum re-invested capital per period can be expressed as the product of the profit, the number of goods sold, and the percentage not owed to the government for taxes.

$$I = (P - C) \cdot Q \cdot (1 - t) = \Delta C_I \qquad \textbf{(3.6.5)}$$

This amount multiplied by the ratio of quantity produced to capital, normalized by quantity shows growth to be equivalent to return on invested capital [2].

$$G = \frac{\kappa \cdot (P - C) \cdot Q \cdot (1 - t)}{Q} = \frac{(P - C) \cdot Q \cdot (1 - t)}{C_I} = ROIC \qquad \textbf{(3.6.6)}$$

Change in quantity for each firm is subsequently given as the product of ROIC and the last period's quantity produced.

$$\Delta Q_{i,n} = ROIC_{i,n-1} \cdot Q_{i,n-1} \qquad \textbf{(3.6.7)}$$

The change in quantity in the industry is the sum of all the new production.

$$\Delta Q_{T,n} = \sum \Delta Q_{i,n-1} \qquad \textbf{(3.6.8)}$$

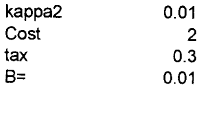
Assuming all firms are price takers, and that the change in price varies linearly with changes in quantity, then the new price will be lowered by added supply.

$$\Delta P_n = -B \cdot \Delta Q_{T,n-1} \qquad \textbf{(3.6.9)}$$

Growth in sales, and subsequently ROIC, can be calculated in a sequential manner.

$$G_{i,n} = \frac{(P_n + \Delta P_n - C) \cdot Q_{n-1} \cdot (1-t)}{C_{I,n}} = \frac{(P_n - B \cdot \Delta Q_{T,n-1} - C) \cdot Q_{i,n-1} \cdot (1-t)}{C_{I,n}} \qquad \textbf{(3.6.10)}$$

A fictional duopolistic competition described in Table 3.6-1 was simulated to illustrate the effects of ROIC on growth and profitability in a competitive market. The table illustrates the state of the market and the firms at a frozen moment in time. The only difference in the firms' business is that Firm 1 possesses an asset turnover ratio that is twice that of Firm 2. The market conditions are set by cost, tax rate, and elasticity of the market; all of which are observable data points. The results are captured in Table 3.6-2 and in Figure 3.6-1.



kappa1	0.02
kappa2	0.01
Cost	2
tax	0.3
B=	0.01

Table 3.6-1: Competitive Example

time	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
P	20	19.811	19.58471	19.31423	18.99179	18.6089	18.15664	17.62612	17.00928	16.29992	15.495	14.59616	13.61114	12.55476	11.44907	10.32236
Delta P		0.9%	1.1%	1.4%	1.7%	2.0%	2.4%	2.9%	3.5%	4.2%	4.9%	5.8%	6.7%	7.8%	8.8%	9.8%
Firm 1 Q	50	62.6	78.20956	97.46366	121.0888	149.894	184.748	226.5367	276.0951	334.111	400.9996	476.7605	560.8354	652.0026	748.3468	847.3433
Firm 1 C	2500	3130	3910.478	4873.183	6054.438	7494.699	9237.401	11326.83	13804.76	16705.55	20049.98	23838.02	28041.77	32600.13	37417.34	42367.17
Firm 1 ROIC	0.252	0.249354	0.246186	0.242399	0.237885	0.232525	0.226193	0.218766	0.21013	0.200199	0.18893	0.176346	0.162556	0.147767	0.132287	0.116513
Delta ROIC		1.1%	1.3%	1.5%	1.9%	2.3%	2.7%	3.3%	3.9%	4.7%	5.6%	6.7%	7.8%	9.1%	10.5%	11.9%
Firm 2 Q	50	56.3	63.31932	71.11348	79.7324	89.21598	99.58843	110.8515	122.9768	135.8973	149.5006	163.6231	178.0503	192.5219	206.746	220.4209
Firm 2 C	5000	5630	6331.932	7111.348	7973.24	8921.598	9958.843	11085.15	12297.68	13589.73	14950.06	16362.31	17805.03	19252.19	20674.6	22042.09
Firm 2 ROIC	0.126	0.124677	0.123093	0.1212	0.118943	0.116262	0.113096	0.109383	0.105065	0.100099	0.094465	0.088173	0.081278	0.073883	0.066144	0.058257
Delta ROIC		1.1%	1.3%	1.5%	1.9%	2.3%	2.7%	3.3%	3.9%	4.7%	5.6%	6.7%	7.8%	9.1%	10.5%	11.9%

Table 3.6-2: Duopoly competition

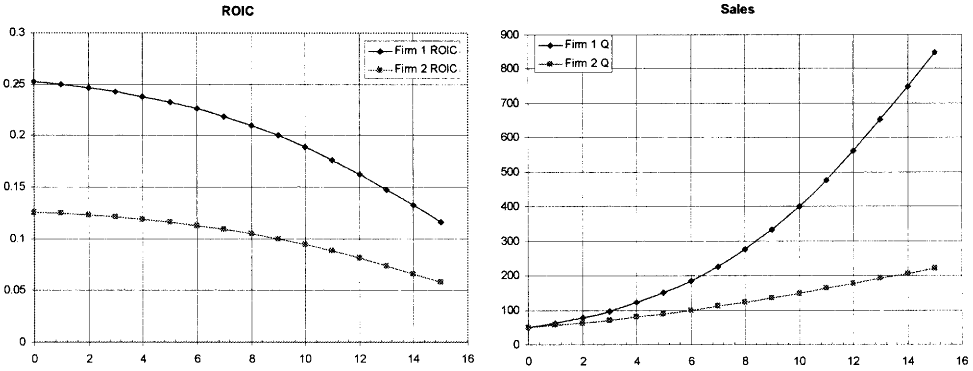


Figure 3.6-1: ROIC and growth in duopolistic competition

Figure 3.6-1 clearly illustrates the outcome of the competition to the extent that investors continue to fund the growth of Firm 2. In a real competition, managers acting as "good" agents for their investors and debt holders will exit the market when the ROIC drops below the weighted average cost of capital (WACC) of the firm [2]. In the example above, the imminent exit of Firm 2 from the market cedes market power to Firm 1 in the form of monopoly. Over time, high profit margins coupled with changes in the market exposes the winning firm to new competition. Emergent firms that recognize the market opportunity are likely to enter and compete at an even higher ROIC due to new technology and evolved business strategies.

3.7 Summing Up

This chapter provides a fundamental business strategy framework for firms engaged in a market for complex systems. Clearly, a firm must provide a benefit to be considered part of the value network. Value is created by improving on performance limitations that occur at bottlenecks in a system's architecture. Firms have the choice to address one or all of the bottlenecks in a system that is valued by the market. The firm that opts to focus on a subset of bottlenecks wisely reaps the benefits of modularity and selective outsourcing. As the holder of the portfolio of options, the firm possessing a modular design creates much more value than a vertically integrated firm that at best offers options on a portfolio to customers in the market. The double dividend comes in the form of improved asset turnover as a result of outsourcing that leads to higher ROIC and growth.

Clearly, the antecedent to successful execution of the strategies coupled to modularity is knowledge of relevant bottlenecks. Since the focus of this thesis is to apply the frameworks introduced in this chapter to the wireless handset industry, an understanding of the domain specific bottlenecks is absolutely necessary. To this end, the reader must explore the science and standards that bound the performance of wireless communication systems. Chapter 4 will introduce the science; and Chapter 5 will discuss prevalent standards that embody the science and enable the substitution operator to thrive. Chapter 6 will investigate the contemporary architectures and align them with the modified operators and emergent trends.

4 Waveform Engineering

Chapter 4 establishes the natural bottlenecks in wireless communication systems. The bottlenecks define the modes of competition at the handset and supplier layers in the value network. Owing to the nature of the science behind the bottlenecks, readers inexperienced in the field of electrical engineering may find this chapter to be dense and confusing. Unfortunately, because wireless communication bottlenecks are predicated on science, the depth of the discussion is unavoidable. That being said, this chapter is NOT a mathematically rigorous treatment of digital communication systems, although much of the machinery that would appear in a communications text book is at least mentioned in the sections that follow. Footnotes, ancillary "boxes", and periodic summaries are employed to make the chapter as readable as possible. Those readers who find themselves getting mired in the abstracted math can choose to skip to the final section in this chapter, albeit at the expense of the intuition gained from the methodical development of relevant concepts.

Waveform engineering refers to the science of transforming information into waveforms capable of propagation via analog channels. The digital waveform engineering process consists largely of modulation and pulse shaping. The term modulation refers to the process of mapping one or more bits to analog waveforms. Pulse shaping is the product of both careful engineering and channel imperfections. The term channel refers simply to the wireless transmission medium. This chapter will prove that performance of all wireless communication systems is embodied by three performance bottlenecks: error rate, data rate, and battery life.

Chapter 4 is split into three parts. Part 1, Sections 4.1-4.5, establishes design variables, system constraints, and performance bottlenecks determined by the bandpass nature of contemporary wireless communication systems. Part 2, Sections 4.6-4.9, exposes additional variables, constraints and performance bottlenecks by considering the antecedents of successful communicate events that occur the context of part 1. Part 3, Section 4.10 and 4.11, pulls the information from the two previous parts together and relates it to the management framework introduced in Chapter 3.

More specifically, section 4.1 introduces a heuristic model for bandpass waveforms. Section 4.2 establishes the importance of waveform power in determining compliance with regulation and battery life conservation. Section 4.3 presents a generalized expression for bandpass waveforms and illustrates the relationship between the baseband information carried by a waveform and the power in the bandpass waveform. Section 4.4 refines the power formulation from Section 4.3 with features specific to pulsed modulating waveforms, the dominant design in digital communications systems. Section 4.5 exposes the constraints, design variables, and bottlenecks that emerge upon consideration of the spectral and power qualities of the pulse modulated bandpass waveforms. Per the portioning stated in the previous paragraph, the focus on the chapter shifts to the theory that determines a successful communication event. In as much, Section 4.6 introduces the machinery that enables error free communication of information imparted on bandpass waveforms. Section 4.7 introduces the systematic schemes used to map information to the baseband pulsed modulated waveforms. Section 4.8 employs the machinery from Section to 4.6 to postulate the error-free receiving task. Finally, Section 4.9 introduces noise into the previously perfect environment to expose the error rate and battery life bottlenecks. Section 4.10 assembles the body of work from the first nine sections into a concise system of inter-related performance bottlenecks. Section 4.11 discusses the strategic implications of Section 4.10.

4.1 Fundamentals

All physical waveforms bearing information are functions of time. However, every time domain waveform can be decomposed into a frequency domain equivalent[3]. The frequency domain content determined by the decomposition determines the bandwidth of the waveform and is of utmost importance to communication system designers owing to the nature of the band regulations. In general, communication waveforms fall into one of two general categories: baseband or bandpass. Baseband waveforms are centered about zero frequency, often referred to as direct current (DC). Bandpass waveforms are centered about a carrier frequency (f_c). Per regulation, these bands are finite and well guarded. Therefore, the spectral qualities of communication waveforms are designed carefully to optimize spectral efficiency and minimize spread in the frequency domain. For example, sharp transitions in the time domain, such as edges, are avoided because they map to large and sometimes infinite sums of sinusoids in the frequency domain.

Both categories of waveforms are symmetric about DC. Because frequency is the time derivative of the phase of a sinusoid, the Fourier transform of a time domain waveform contains both positive and negative frequencies corresponding to forward and reverse phase derivatives. However bandpass waveforms are centered about a carrier frequency and therefore the frequency content on either side of the carrier corresponds to unique positive frequencies. For instance, a *baseband* waveform limited to ± 10 Hz can be filtered using any component specified up to 10 Hz because the component does not discriminate between

[3] The process of mapping time domain waveforms to the frequency domain is dependant on the periodic properties of the waveform. Truly periodic waveforms are decomposed into a finite sum of tones using a mathematical abstraction called a Fourier series. All other waveforms are mapped to the frequency domain using an operation called a Fourier transform. Both operations represent tones present in the time domain waveform with spikes in the frequency domain. The amplitude of the spikes is determined by the energy or power of the tones present in the time domain waveform.

$\pm f$ centered about 0. On the other hand, the same waveform centered about f=100 Hz requires components specified from 90 Hz to 110 Hz. For this reason, bandpass waveforms are often labeled double sideband, while baseband waveforms are labeled single sideband. Figure 4.1-1 illustrates the spectral components of both baseband and bandpass waveforms.

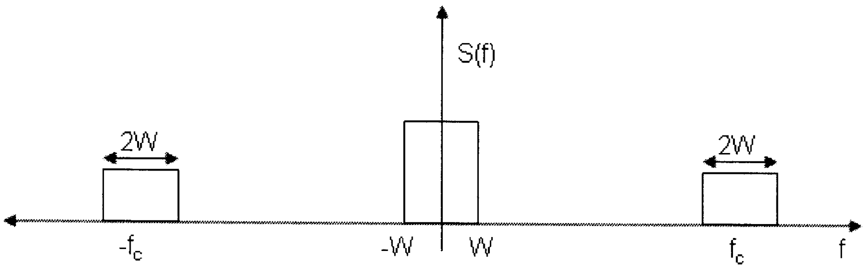


Figure 4.1-1: Frequency domain representation of bandpass waveform. The single sideband baseband waveform is symmetric about f=0 and spans [-W,W]. The bandpass waveforms span the same 2W but are centered about ± f_c.

4.2 Frequency and Power

There are three primary reasons why power in a bandpass waveform, described as a function of frequency, is desirable. First, per the regulatory discussion in Chapter 2, communication events are confined to bands. Encroachment is determined by power spillage from a waveform residing in one band into unlicensed neighboring bands. Second, per energy conservation, a transmitted waveform naturally dissipates the power supply of a transmitter proportionally to the power in the waveform. Therefore, in a portable device waveform power is an end in itself because it directly impacts operational battery life. Lastly, foreshadowing Section 4.9, detection performance of a receiver is largely determined by the energy or power in a waveform. This section presents the mathematical machinery required to describe the power in arbitrary waveforms as a function of frequency.

The *time domain* waveforms for energy and power are given by Equation (4.2.1) and (4.2.2), respectively[4].

$$E = \int_{-\infty}^{\infty} \left\| x(t) \right\|^2 dt \qquad \textbf{(4.2.1)}$$

$$P = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \left\| x(t) \right\|^2 dt \qquad \textbf{(4.2.2)}$$

What's important to communication system designers is the energy or power corresponding to the *frequency* content in the waveforms. Application of Parseval's Theorem to Equations (4.2.1) and (4.2.2) yield the frequency domain equivalent expressions for energy and power [16].

$$E = \int_{-\infty}^{\infty} \left\| X(f) \right\|^2 df \qquad \textbf{(4.2.3)}$$

$$P = \int_{-\infty}^{\infty} \lim_{T \to \infty} \frac{\left\| X(f) \right\|^2}{T} df \qquad \textbf{(4.2.4)}$$

In the equations above, the term $X(f)$ is the Fourier transform of the time domain waveform, $x(t)$. The integrands in (4.2.3) and (4.2.4) are called the energy spectral density (ℰ𝒮𝒟) and the power spectral density (𝒫𝒮𝒟) of the waveform, respectively [17]. In situations when the Fourier transform of *x(t)* cannot be evaluated, owing to mathematical complexity, the 𝒫𝒮𝒟 is determined by applying the Fourier transform to the autocorrelation function of *x(t)*[5].

[4] Close examination of Equation (4.2.1) and (4.2.2) reveals that a waveform with finite energy has zero power, and a waveform with finite power contains infinite energy. In other words, a waveform can either be an energy signal or a power signal, but not both.

[5] A problem arises when the Fourier transform, and subsequently the PSD, of x(t) cannot be calculated directly, as in the case of an infinite pulse train. Fortunately, most waveforms of interest are either deterministic or stochastic,ergodic, and wide sense stationary. In this case, the PSD can be generated using the Fourier transform of the autocorrelation function of x(t) - which is tractable for communication waveforms of interest. Wide

4.3 Bandpass Waveforms

Section 4.1 made the claim that contemporary wireless communication systems employ bandpass waveforms to communicate information. Section 4.2 provided a general equation that is used to describe the power in an arbitrary waveform. This section will connect the concepts in the two previous sections by establishing the power spectral density of a bandpass waveform. Also noteworthy, this section introduces the two information carrying constituents of a bandpass communication waveform, the "I" and "Q", that will be the focus of proceeding sections.

All physical bandpass waveforms can be represented in compact form by Equation (4.3.1).

$$s(t) = a(t)\cos\left(2\pi f_c t + \Theta(t)\right) \qquad \textbf{(4.3.1)}$$

By identity, the waveform defined by Equation (4.3.1) can also be represented by a weighted sum of quadrature shifted sinusoids oscillating at the carrier frequency f_c. See Equation (4.3.2).

$$\begin{aligned} s(t) &= x(t)\cos\left(2\pi f_c t\right) - y(t)\sin\left(2\pi f_c t\right) \\ x(t) &= a(t)\cos\left(\Theta(t)\right) \\ y(t) &= a(t)\sin\left(\Theta(t)\right) \end{aligned} \qquad \textbf{(4.3.2)}$$

sense stationary stochastic signals are unique in that the expectation value is independent of time and the autocorrelation is dependant only on the time difference between epochs. An ergodic process is one in which the time average is equivalent to the ensemble average. The autocorrelation function for an ergodic wide sense stationary waveform is given in Equation (4.2.5).

$$\begin{aligned} R_{xx}(\tau) &= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)dt \\ &= E\left[x(t)x(t+\tau)\right] \end{aligned}$$

An intuitive example of an ergodic wide sense stationary signal is an infinite random binary pulse train.

The functions $\cos(2\pi f_c t)$ and $\sin(2\pi f_c t)$ comprise the two dimensional basis spanning bandpass, or quadrature, signal space[6]. The functions x(t) and y(t), often referred to as the in-phase (I) and quadrature (Q) components of a bandpass waveform, exist as baseband waveforms that are used to modulate the basis functions. What's more, the I and Q waveforms comprise two of the three degrees of freedom that are employed by contemporary modulation schemes to communicate information using bandpass waveforms[7].

From Section 4.2, designers require the frequency domain representation of waveforms. Calculation of the Fourier transform and $\mathcal{PSD}$ is made easy by introducing the equivalent bandpass waveform expressions given below.

$$s(t) = \Re\left\{z(t)e^{j2\pi f_c t}\right\} \qquad \textbf{(4.3.3)}$$

$$z(t) = x(t) + jy(t) \qquad \textbf{(4.3.4)}$$

Equation (4.3.4) is called the complex envelope of the waveform $s(t)$. Note, the complex envelope combines both the I and Q components of the waveform in a single complex expression[8]. Using this equivalent notation, the Fourier transform and $\mathcal{PSD}$ of a quadrature waveform can be derived easily [17]. The results are given in Equation (4.3.5) and (4.3.6), respectively.

[6] What is a basis? The Cartesian coordinate axis spanning two dimensional Euclidean space provides an intuitive analogy that can help explain the concept of a basis. In Euclidean space, any arbitrary vector can be generated by first assigning weights along the $\hat{x}$ and $\hat{y}$ Cartesian axes and then summing the result. Similarly, in a two dimensional signal space, any waveform can be generated by assigning weights to quadrature shifted sinusoids and summing the result. The $\hat{x}$ and $\hat{y}$ vectors spanning two dimensional Euclidean space comprise a basis in the same way that the quadrature shifted sinusoids comprise a basis spanning two dimensional signal space. Digressing for the sake of semantics; in the context of waveform engineering, the term quadrature refers to the 90 degree phase shift between sine and cosine functions. For this reason, bandpass waveforms are often referred to as quadrature waveforms. The synonyms bandpass and quadrature will be used interchangeably in this thesis to describe the special class of waveforms employed by wireless communication systems.

[7] The third degree of freedom is the duration of the pulse stream discussed in Section 4.4.

[8] It is common in data sheets to see the time domain in-phase and quadarature waveforms to be referred simply by their acronyms: I and Q.

$$S(f) = \frac{1}{2}\left[Z(f - f_c) + Z(-f - fc)\right] \qquad \textbf{(4.3.5)}$$

$$\Phi_S(f) = \frac{1}{4}\left[\Phi_Z(f - f_c) + \Phi_Z(-f - fc)\right] \qquad \textbf{(4.3.6)}$$

The key observation from Equations (4.3.5) and (4.3.6) is that the Fourier transform and $\mathcal{PSD}$ of a bandpass waveform are determined by shifted versions of the baseband complex envelope.[9] This means that system designers need only bother with the properties of the I and Q waveforms and the carrier frequency in order to completely define a bandpass communication waveform. The next section will illustrate the systematic approach by which digital information is embedded on the I and Q baseband waveforms.

4.4 Mapping Information to Waveforms

Digital systems communicate waveforms over brief time intervals called pulses. The general form of a *baseband* digital communication pulse train is given in Equation (4.4.1) wherein $\{a_k\}$ is the set of possible symbols that represents discrete random data bits and $g(t - kT)$ is a pulse function occupying period *kT*. Throughout this report the term symbol will be used to refer to members of the set $\{a_k\}$ and can represent single bits or M-ary messages. To reiterate, the information is communicated in the form of symbols carried by pulses.

$$s(t) = \sum_{k=-\infty}^{\infty} a_k p(t - kT_s) \qquad \textbf{(4.4.1)}$$

[9] The expression for $\mathcal{PSD}$ captured in equation (4.3.6) is generic and applies to both deterministic and wide sense stationary stochastic signals [17].

The duration of the pulse T_s is determined by the baud rate $\left(\frac{symbols}{\sec}\right)$ of the system. From the Section 4.3, the $\mathcal{PSD}$ of the baseband pulse train is generated by calculating the Fourier transform of the autocorrelation of $s(t)$ [18]. The result is given in Equation (4.4.2)[10].

$$\Phi_S(f) = \frac{\sigma_a^2 |P(f)|^2}{T_s} + \left(\frac{\mu_a}{T_s}\right)^2 \sum_{k=-\infty}^{\infty} \left| P\left(\frac{k}{T_s}\right)\right|^2 \delta\left(f - \frac{k}{T_s}\right) \qquad \textbf{(4.4.2)}$$

Bandpass waveforms behave similarly, except that they employ the orthogonality property of sinusoidal basis functions to transmit two independent pulse trains simultaneously[11]. The I and Q pulse trains are given in Equation (4.4.3). In this case, the symbols are specified by <u>signal</u> sets $\{x_k\}$ and $\{y_k\}$.

$$\begin{aligned} x(t) &= \sum_{k=-\infty}^{\infty} x_k p(t - kT_s) \\ y(t) &= \sum_{k=-\infty}^{\infty} y_k q(t - kT_s) \end{aligned} \qquad \textbf{(4.4.3)}$$

From the last section, the complex envelope is formed by inserting Equation (4.4.3) into Equation (4.3.4).

Recall from Equation (4.3.6) that the $\mathcal{PSD}$ of a bandpass waveform is determined by the Fourier transform of the autocorrelation function of its complex envelope[12]. See Equation (4.4.4).

[10] Several important observations can be gleaned from Equation (4.4.2). First, the $\mathcal{PSD}$ contains both continuous and discrete components. The continuous components are proportional to the variance of the symbols, σ_a^2. The discrete spectra occur at multiples of the data rate and are proportional to the mean of the symbols, $\mu_a = E[\{a_k\}]$. Subsequently, independent of the pulse shape, $p(t)$, a system designer can minimize the power and bandwidth of the waveform by choosing a symbol set $\{a_k\}$ with zero mean.

[11] This implies that a quadrature waveform transmits twice the information of a baseband waveform. However, the baseband waveform only occupied half of the non-negative spectrum; so the spectral efficiency is exactly the same [16].

[12] The autocorrelation function is merely a product of the waveform and a time shifted version of itself.

$$R_{zz}(\tau) = R_{xx}(\tau) + R_{yy}(\tau) \qquad \textbf{(4.4.4)}$$

By inspection, the autocorrelation function of a digitally modulated bandpass waveform is the sum of the I and Q autocorrelation functions. It follows that the Fourier transform of (4.4.4) yields a 𝒫𝔖𝔇 that is simply a two-fold replication of Equation (4.4.2), one for each autocorrelation function, and shifted to the carrier frequency. The 𝒫𝔖𝔇 for a pulsed quadrature waveform with symmetric signals $\{x_k, y_k\}$ centered about the origin is given in Equation (4.4.5)[13].

$$\Phi_S(f) = \frac{\sigma_x^2 |P(f)|^2}{T_s} + \frac{\sigma_y^2 |Q(f)|^2}{T_s} \qquad \textbf{(4.4.5)}$$

From Equation (4.4.5), the power in a bandpass waveform is determined by the variance in the signals, the baud rate, and the Fourier transform of the pulse functions, p(t) and q(t). All else being equal, a waveform comprised of a diverse signal set with high variance and high baud rate ($1/T_S$) consumes more power than a slow binary waveform. It will be shown in Section 4.7 that the variance in the signal set is dependant on the modulation scheme. Section 4.6 will address the pulse shape design variable and describe the trade-off between robustness and bandwidth.

In review, three system level design parameters have been revealed in this section: baud rate T_s, signal sets $\{x_k, y_k\}$, and the pulse shape for $p(t)$ and $q(t)$. From an architectural point of view, handsets act as clients on the wireless network; and therefore the

$$\begin{aligned} R_{zz}(\tau) &= E[(x(t) + jy(t))(x(t+\tau) + jy(t+\tau))] \\ &= E[x(t)x(t+\tau)] + E[y(t)y(t+\tau)] \\ &= R_{xx}(\tau) + R_{yy}(\tau) \end{aligned}$$

[13] Note, the symmetric symbol set sets the mean μ_a to 0.

aforementioned parameters are determined upstream by the network designers and transferred down to the handset layer in the form of design rules.

4.5 Pulse Shaping and Baud Rate

Recall, Section 4.2 established the primacy of power spectral density. Section 4.3 captured the $\mathscr{PSD}$ for general bandpass waveforms. Section 4.4 refined the analysis to include pulse trains carrying signals $\{x_k, y_k\}$, but made no mention of the pulse shape. A logical question arises: Is there an optimal pulse shape for waveforms employed by digital communication systems? The answer, as one might expect, is that a Pareto optimal solution is the best a designer can hope for. The competing objectives along the Pareto front are bandwidth, as determined by the frequency-localized power, and robustness which captures the receivers ability to extract transmitted information. From Equation (4.4.5), bandwidth is inversely proportional to baud rate. The robustness argument follows from Equation (4.4.3) and is described in this section.

In order to properly detect the signals $\{x_k, y_k\}$ from the received waveform, samples at epoch kT_s should not include remnants of signals sent during any previous or future periods. Failure to comply with this constraint leads to an error phenomena referred to as intersymbol interference (ISI) [16,17,18]. The most intuitive pulse shape that meets the ISI criteria is a rectangle of width T_s scaled to $\{x_k, y_k\}$. However, from Section 4.1, the sharp edges of the square pulse map to infinite spectral occupation in the frequency domain.

Since spectral real estate is of primary importance to system designers, inverting the problem and starting with a bandwidth efficient waveform certainly makes sense. As it turns out, a pulse with rectangular shaped <u>spectrum</u> maximizes spectral efficiency. Its corresponding time domain pulse shape is that of a sinc function. See Equation (4.5.1) for the mathematical representation of rectangular spectrum and its corresponding pulse shape, the sinc function. The forward and reverse arrow indicates a Fourier transform pair.

$$rect\left(\frac{f}{2W}\right) \Leftrightarrow \frac{1}{\mathrm{T_s}}\cdot \mathrm{sinc}\left(\frac{\mathrm{t}}{\mathrm{T_s}}\right) = \frac{\sin(2W\cdot \pi t)}{\pi t} \tag{4.5.1}$$

A graphical illustration of the optimal pulse shape and its spectrum are presented in Figure 4.5-1. The blue traces in Figure 4.5-1 illustrate the optimal frequency domain pulse. The red and green traces illustrate a more robust time domain pulse at the expense of spread in the frequency domain. To put the waveforms in proper context, Figure 4.5-2 illustrates a snapshot of an infinite pulse train. Notice how each pulse peaks when all of the others cross through zero. This is a visual confirmation of the ISI error criteria, otherwise known as the Nyquist criteria.

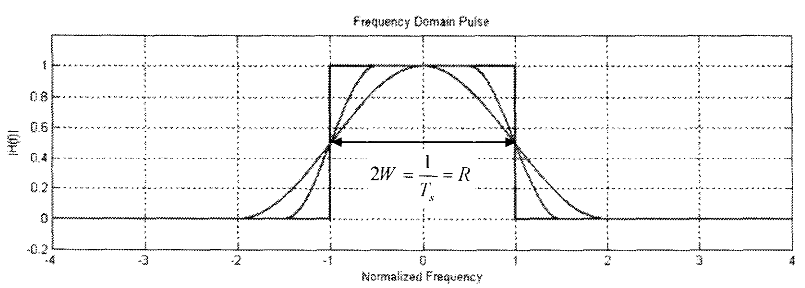


Figure 4.5-1: Band Limited Pulses and Corresponding Spectrum. The figure illustrates the class of pulses with finite and localized frequency content. The undulations away from t=0 have a higher potential to cause interference. The figure shows that pulses with suppressed oscillation tend to spread in the frequency domain. Since the time and frequency domain representation of signals are duals of one

another, the square wave from the lower figure, if perceived in the time domain, would exhibit a frequency signature from the upper figure. This is exactly why square wave pulses are not used in bandpass communication systems.

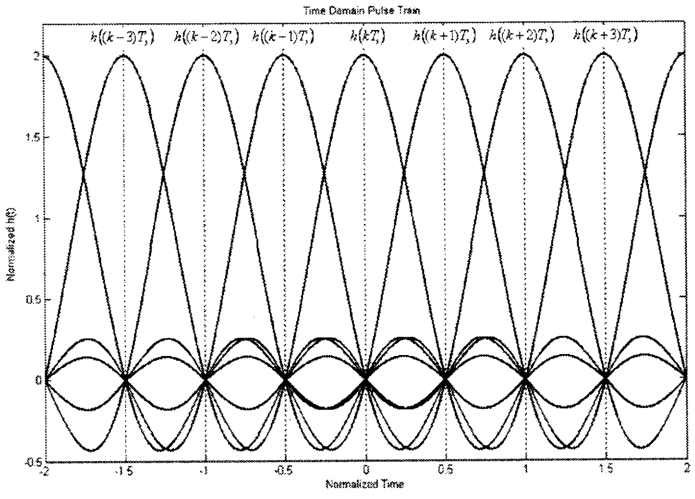


Figure 4.5-2: Sinc Function Pulse Train. Nyquist compliant pulses, when sampled in the center of the period, are not conaminated by adjacent pulses.

Implicit in all filtered pulse shapes consistent with Figure 4.5-1 is compliance with the Nyquist criteria. Simply stated, the Nyquist criterion for a pulse shape dictates that in the absence of timing errors, a waveform sampled at the center of its period will contain only information embedded in the target waveform[14]. When exposed to timing *error*, even Nyquist waveforms experience ISI. Such is the case when sampling occurs at instances left-of-center or right-of-center of the target pulse. The consequence of ISI is that the sampled waveform includes information from past and future pulses

The robustness improvements realized by a filtered pulse are illustrated in Figure 4.5-3. The black trace is the time domain representation of a standard sinc function pulse train corresponding to the blue trace from Figure 4.5-1. The blue trace corresponds to a filtered

[14] The Nyquist criterion is intimately linked to the Sampling Theorem that is also used to describe aliasing of sampled waveforms [16].

sinc function represented by the green or red trace from Figure 4.5-1. From 4.5-3, sampling errors originating from the blue family of pulses are less pronounced than those from the black pulses because undulations from neighboring pulses are suppressed. From figure 4.5-1, the filtering operation does not come without cost. Reflecting back on the figure, notice that the red time domain pulse exhibits the smallest undulations, but that its bandwidth is the largest of the three pulses. The remainder of this section describes the compromise between robustness, bandwidth and data rate.

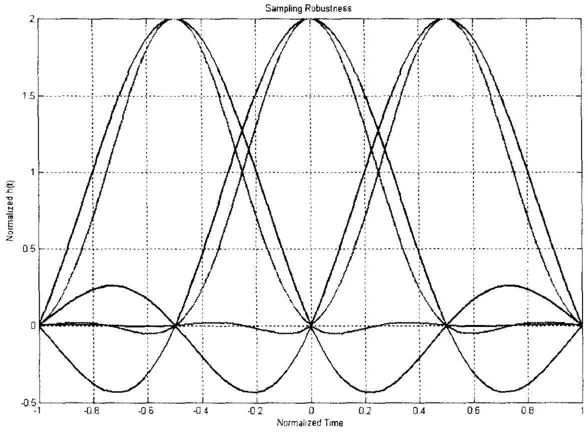


Figure 4.5-3: Sinc Pulse and Filtered Pulse Comparison. The unfiltered sinc pulse (black) occupies the smallest amount of frequency space for a given baud rate, 1/T. Therefore, it also supports the highest data rates for a given bandwidth. However, oscillations are the most pronounced and therefore, unfiltered data are more prone to timing related error. The filtered pulse (blue) exhibits much better robustness to timing errors since the oscillations are smaller, but from 4.5-1, occupies more spectrum.

From Equation (4.4.3), the pulse period determines the baud rate in any pulse-based waveform. See Equation (4.5.2).

$$R = \frac{1}{T_s} \qquad \textbf{(4.5.2)}$$

From Equation (4.5.1), the pulse width of an unfiltered sinc function also determines the double sideband-bandwidth of the waveform in the frequency domain.

$$B = 2W = \frac{1}{T_s} \qquad \textbf{(4.5.3)}$$

Therefore, the data rate achieved by a sinc shaped pulse is exactly the double sideband-bandwidth of the pulse. However, per the previous discussion, if a sinc shaped pulse is not robust enough for practical communication systems then it is filtered. From Figure 4.5-1, the filtering stretches the bandwidth of the waveform beyond the 2W bandwidth of the sinc function, but maintains the same data rate 2W. If a hard limit is placed on the bandwidth of the waveform, as determined by the $\mathcal{PSD}$ of the waveform, then the data rate must be reduced to accommodate the filtering. For example, if the available channel is 5MHz, and the ISI filter requires and additional 200kHz on both sides of 2W, then the baud rate of the system must be reduced to 46 mbps.

For the reasons discussed above, a pulse shaping filter is implemented in the transmitter to conserve bandwidth and minimize ISI. Unfortunately, the wireless channel is lousy and introduces noise and "shapes" the pulse. Since the goal in communicating information is to transmit and receive intelligible waveforms, the effect of the channel imperfections must be managed by the receiver. It can be shown that a filter at the receiver exactly matched to the combination of the pulse shaping filter and the undesirable channel filter minimizes the effect of noise corruption on the received waveform [19]. In other words, the affect of pulse shaping must be reversed so that sampled signals $\{x_k, y_k\}$ can be recovered from the received waveform. Owing to its optimal noise performance, the "matched filter" has become a pillar in all receiver architectures. Additional constraints placed on the pulse shaping filter are introduced in Box 4.5-1. The un-interested reader need only know that the pulse shaping filter and matched filter combination must not violate Equation (4.4.3); otherwise recovery of signals will be impossible.

Box 4.5-1. The filtering operation is described mathematically in Equation (4.5.4) wherein the pulse shape *g(t)* includes the effects of both the pulse shaping filter and the channel.

$$\begin{aligned} r(t) &= s(t) * h(t) \\ &= \int_{-\infty}^{\infty} s(\tau)h(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_k g(\tau - kT_s)h(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_k g(\tau - kT_s)g(-(t-\tau))d\tau \\ &= \int_{-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_k g(\tau - kT_s)g(\tau - t))d\tau \end{aligned} \qquad \textbf{(4.5.4)}$$

The first equality in (4.5.4) is a statement of one of the foundations of linear time invariant systems [ref]. In words, the output a filtered linear time invariant signal is the convolution of the input waveform with the filter's impulse response, *h(t)*. The second line substitutes the mathematical form of the convolution operator for the asterisk. Recall that the ISI condition requires that $r(T_j)=a_j$ so that the modulated information can be extracted by the receiver. Equation (4.5.5) illustrates the ISI condition at the receiver wherein the baseband pulse train was used to simplify the analysis.

$$r(jT) = \int_{-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_k g(\tau - kT_s)g(\tau - jT_s)d\tau = \begin{cases} a_j & \mathrm{j=k} \\ 0 & \mathrm{j \neq k} \end{cases} \qquad \textbf{(4.5.5)}$$

Equation (4.5.5) implies that the pulse train $\sum_{k=-\infty}^{\infty} g(\tau - kT_s)$ forms an orthonormal basis for the received waveform *r(t)* provided the waveform is sampled at instant kT_s. The importance of basis functions and inner products in the context of digital communication will be explored in the next section.

This section concludes with a survey of the ground covered so far in this chapter. Section 4.1 discussed general properties of baseband and bandpass wave forms. Section 4.2 introduced the concepts of energy and power and described the machinery used to analyze communication waveforms. Section 4.3 explained in more detail the dominant design in contemporary communication systems, bandpass waveforms. Notably, the $\mathscr{PSD}$ of bandpass waveforms were shown to derive from the baseband complex envelope. Building on the previous sections, Section 4.4 introduced the general form of digital communication waveforms. Finally Section, 4.5 introduced ISI, the Nyquist criterion and the compromises therein: robustness and bandwidth\data rate. The design parameters exposed through these five sections are: signal sets, symbol period\baud rate\bandwidth, and pulse shape. The performance bottlenecks implicit in these design variables is data rate, i.e. the amount of information that can be transmitted over an analog channel, and battery life owing to the power carried by the transmitted waveforms. The next three sections explore the formal processes used to design signal sets and receive information. These concepts reveal new design variables, add color those already listed, and augment the list of performance bottlenecks.

4.6 Signal Spaces and Inner Products

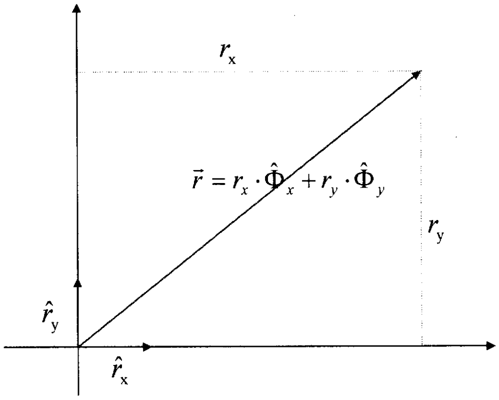
Assume for the sake of argument that the wireless handset is responsible for both transmit and receive functions. Wireless communication is initiated by the transmitter module. In as much, modulation schemes are employed to embed signals $\{x_k, y_k\}$ on the pulsed baseband functions $x(t)$ and $y(t)$ at a rate $\frac{1}{T_U}$ determined by the up-link baud rate of the system. Next the $x(t)$ and $y(t)$ waveforms are shaped by a filter $p(t)$ to address ISI and conserve bandwidth. The shaped pulses are then up-converted to $cos(2\pi f_c t)$ and, $sin(2\pi f_c t)$ respectively, and combined to realize the bandpass waveform of Equation (4.3.2). Finally, the quadrature waveform is amplified by a power amplifier and transmitted over the wireless channel via the antenna.

Conversely, the receiver's role is to collect bandpass waveforms and generate symbol streams. The receive process is initiated when the *noise corrupted* waveform is collected by the antenna. The bandpass waveform is projected onto a phase locked orthogonal quadrature basis and thereby separated into the in-phase and quadrature pulses *x(t)* and *y(t)*. The modulated signals $\{x_k, y_k\}$ are extracted by passing *x(t)* and *y(t)* through a filter matched to the combination of the original pulse shaping filter and channel filter and sampled at a rate $\frac{1}{T_D}$. Finally, the receiver attempts to determine the correct signal set.

The previous paragraph introduced a new concept, projection, which refers to the process of extracting information from a signal by examining its components along the directions of an orthonormal basis. Technically, this concept is employed twice in the receive process. The first projection separates the in-phase and quadrature waveforms comprising the received waveform. The second projection extracts $\{x_k, y_k\}$ from the filtered baseband waveforms. In mathematical terms, the projection process is captured by the inner product between a vector or waveform and its respective basis. The two bases of interest in communication systems are the quadrature basis and the infinite pulse train. Box 4.6-1 builds on the Cartesian analogy introduced in the footnotes of Section 4.3 to illustrate the inner product (projection) operation in both Cartesian space and in signal space. In the final analysis, Box 4.6-1 demonstrates that the I and Q components can be extracted from a bandpass waveform by correlating it with the in-phase and quadrature carrier tone.

Box 4.6-1. Most readers are probably familiar with projections of vectors on a Cartesian axis in 2-D Euclidean space. In the Cartesian example in the figure below, a vector $\vec{r} = r_x \cdot \hat{r}_x + r_y \cdot \hat{r}_y$ is decomposed into its r_x and r_y components along $\hat{r}_x$ and $\hat{r}_y$, respectively. The term projection is immediately apparent in a two dimensional Cartesian space as r_x is simply the component of $\vec{r}$ projected on $\hat{r}_x$ and similarly r_y is the projection of $\vec{r}$ on $\hat{r}_y$. Notice how the projection, or inner product, of $\hat{r}_x$ on $\hat{r}_y$ is zero. What's more, the inner product of each basis vector on itself is exactly one.

In signal space, the orthonormal basis *vectors* are replaced by orthonormal *functions*. Inner products are determined by correlating the signal with the desired basis function over time interval T. The correlation operator is given as $\int_{-T/2}^{T/2} r(t)\Phi_z(t)dt$ where $r(t)$ is the waveform of interest and $\Phi_z(t)$ is one dimension of an orthonormal basis function. The end result is the same as the geometric example. Namely, the inner product of a function with its basis functions generate the components of the function in the *direction* of each basis.



$$\hat{\Phi}_x = \hat{r}_x \qquad \Phi_x = \sqrt{\frac{2}{T}}\cos(2\pi f_c t)$$

$$\hat{\Phi}_y = \hat{r}_y \qquad \Phi_y = \sqrt{\frac{2}{T}}\sin(2\pi f_c t)$$

$$\left\langle \hat{\Phi}_x, \hat{\Phi}_x \right\rangle = 1 \qquad \int_{-T/2}^{T/2} \sqrt{\frac{2}{T}}\cos(2\pi f_c t)\sqrt{\frac{2}{T}}\cos(2\pi f_c t)dt = 1$$

$$\left\langle \hat{\Phi}_y, \hat{\Phi}_y \right\rangle = 1 \qquad \int_{-T/2}^{T/2} \sqrt{\frac{2}{T}}\sin(2\pi f_c t)\sqrt{\frac{2}{T}}\sin(2\pi f_c t)dt = 1$$

$$\left\langle \hat{\Phi}_x, \hat{\Phi}_y \right\rangle = 0 \qquad \int_{-T/2}^{T/2} \sqrt{\frac{2}{T}}\cos(2\pi f_c t)\sqrt{\frac{2}{T}}\sin(2\pi f_c t)dt = 0$$

$$\left\langle \vec{r}, \hat{\Phi}_x \right\rangle = r_x \qquad \int_{-T/2}^{T/2} r(t)\sqrt{\frac{2}{T}}\cos(2\pi f_c t)dt = x(t)$$

$$\left\langle \vec{r}, \hat{\Phi}_y \right\rangle = r_y \qquad \int_{-T/2}^{T/2} r(t)\sqrt{\frac{2}{T}}\sin(2\pi f_c t)dt = y(t)$$

Figure 4.6-1 illustrates a generic receiver architecture. From the figure, $x(t)$ and $y(t)$ are recovered from the received waveform by forming the inner product between the incident waveform $r(t)$ and the basis functions, $\sqrt{\frac{2}{T}}\cos(2\pi f_c t)$ and $\sqrt{\frac{2}{T}}\sin(2\pi f_c t)$, that are generated within the receiver. To extract $\{x_k, y_k\}$ from $x(t)$ and $y(t)$, a second inner product must be formed between the baseband waveforms and basis defined by the non-overlapping pulse functions developed in Section 4.5 [16,18][15]. For analysis purposes only, the bases are combined and the inner products evaluated simultaneously in Equation (4.6.1) [20].

$$\begin{aligned}
\Phi_{x,k} &= \sqrt{\frac{2}{T}}\cos(2\pi f_c t)g(kT-t)\\
\Phi_{y,k} &= \sqrt{\frac{2}{T}}\sin(2\pi f_c t)g(kT-t)\\
\left\langle \Phi_{x,k},\Phi_{x1,k}\right\rangle &= \int_{-T/2}^{T/2}\sqrt{\frac{2}{T}}\cos(2\pi f_c t)g(t-kT)\sqrt{\frac{2}{T}}\cos(2\pi f_c t)g(t-kT)dt = 1\\
\left\langle \Phi_{y,k},\Phi_{y,k}\right\rangle &= \int_{-T/2}^{T/2}\sqrt{\frac{2}{T}}\sin(2\pi f_c t)g(t-kT)\sqrt{\frac{2}{T}}\sin(2\pi f_c t)g(t-kT)dt = 1\\
\left\langle \Phi_{x,k},\Phi_{y,k}\right\rangle &= \int_{-T/2}^{T/2}\sqrt{\frac{2}{T}}\cos(2\pi f_c t)g(t-kT)\sqrt{\frac{2}{T}}\sin(2\pi f_c t)g(t-kT)dt = 0\\
\left\langle \Phi_{x,k},\Phi_{x,j\neq k}\right\rangle &= \int_{-T/2}^{T/2}\sqrt{\frac{2}{T}}\cos(2\pi f_c t)g(t-kT)\sqrt{\frac{2}{T}}\cos(2\pi f_c t)g(t-jT)dt = 0\\
\left\langle \Phi_{y,k},\Phi_{y,j\neq k}\right\rangle &= \int_{-T/2}^{T/2}\sqrt{\frac{2}{T}}\sin(2\pi f_c t)g(t-kT)\sqrt{\frac{2}{T}}\sin(2\pi f_c t)g(t-jT)dt = 0
\end{aligned} \qquad \textbf{(4.6.1)}$$

In words, Equation (4.6.1) says that the information $\{x_k, y_k\}$ carried by a bandpass waveform comprised of a high frequency carrier modulated by a slow moving pulse train, subject to the constraints of Section 4.5, can be recovered by projecting the waveform on a combined orthonormal basis. The basis is given by the first two rows, $\Phi_{x,k}$ and $\Phi_{y,k}$, in Equation (4.6.1). The 3rd through 7th rows in Equation (4.6.1) merely confirm the conditions of an orthonormal basis. Namely that the projection of a basis function on itself is exactly one and the projection of a basis function on another is exactly zero [16].

The onus falls on the transmitter and receiver to generate the basis functions and form the inner products to impart and extract the information on the transmitted and received waveforms, respectively. In practice, the inner product meta-task is split between the analog

[15] The mutual orthogonality between bases alludes to the number of degrees of freedom per burst in a packet based system. Per Nyquist, a packet of duration T has at most $R \cdot T = 2WT$ degrees of freedom.

and digital components in the transmitter and receiver. The partitioning is determined by the architecture and will be addressed in Chapter 6 of this thesis.

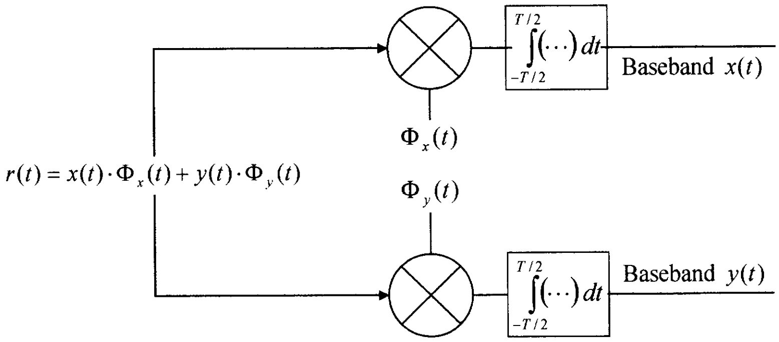

$T/2$
$\int_{-T/2}^{T/2}(\cdots)dt$
$-T/2$
Baseband $x(t)$
$\Phi_x(t)$
$r(t) = x(t)\cdot\Phi_x(t) + y(t)\cdot\Phi_y(t)$
$\Phi_y(t)$
$T/2$
$\int_{-T/2}^{T/2}(\cdots)dt$
$-T/2$
Baseband $y(t)$

Figure 4.6-1: Generic Bandpass Waveform Receiver Architecture. The incoming waveform is multiplied by the basis functions and passed to a matched filter. Thefilters pass two baseband pulse trains carrying information.

4.7 Modulation

Section 4.6 introduced signal spaces and inner products as means and method for imparting and recovering information on bandpass waveforms. The process of mapping $\{x_k, y_k\}$ onto the carrier waveform is known as modulation. Contemporary modulation schemes are divided into three major categories: frequency shift keying (FSK), phase shift keying (PSK), and amplitude shift keying (ASK). ASK is conditionally coined pulse amplitude modulation (PAM) when referring to baseband waveforms or quadrature amplitude modulation (QAM) when referring to quadrature waveforms. Frequency shift keying maps each signal to a different frequency. Phase shift keying maps each signal to a different phase offset. Finally, QAM maps signals to different amplitudes along directions of the in-phase and quadrature basis. The bandpass waveforms and available signals are listed in Table 4.7-1. The index $1 \le i \le M$ determines the symbol coordinate[16].

Modulation	BandPass Waveform	x_k	y_k
FSK	$s_i(t) = Ag(t)\cos(2\pi(f_c + \Delta f_i)t + \phi)$	$A\cos(2\pi\Delta f_i t)$	$A\sin(2\pi\Delta f_i t)$
PSK	$s_i(t) = Ag(t)\cos\left(\frac{2\pi(i-1)}{M} + 2\pi f_c t + \phi\right)$	$A\cos\left(\frac{2\pi(i-1)}{M}\right)$	$A\sin\left(\frac{2\pi(i-1)}{M}\right)$
QAM	$s_i(t) = A_i g(t)\cos(\Theta_i + 2\pi f_c t + \phi)$	$A_i\cos(\Theta_i)$	$A_i\sin(\Theta_i)$

Table 4.7-1: Wireless communication modulation schemes. Frequency shift keying employs different carrier frequencies for each information couple $\{x_k, y_k\}$. Phase shift keying maps coordinates associated with phase to different symbols. QAM also uses phase but also provides a second degree of freedom, amplitude, to the symbols.

Table 4.7-1 suggests that the I and Q constituents of a bandpass waveform provide a mapping of the communicated information to unique coordinates in a 2-D Euclidean space. Subsequently, the x_k and y_k terms in Table 4.7-1 define a particular vector given by Equation 4.7.1.

$$\vec{s}_i = (s_{i,x}, s_{i,y}) \qquad \textbf{(4.7.1)}$$

Each vector in the set of all $\vec{s}_i$ represents a possible symbol. In turn, each symbol maps to a collection of one or more bits. The collection of coordinates consisting of all the possible vectors $\vec{s}_i$ is referred to as the constellation of the modulation scheme. The shape of the

[16] Note, the primary difference between FSK and the other two schemes, is that the FSK $\{x_k, y_k\}$ defines a vector *spinning* around the circle of radius A at a rate Δf_i.

constellation is unique to each scheme. Figure 4.7-1 illustrates the constellations for 8PSK and 16QAM modulation schemes.

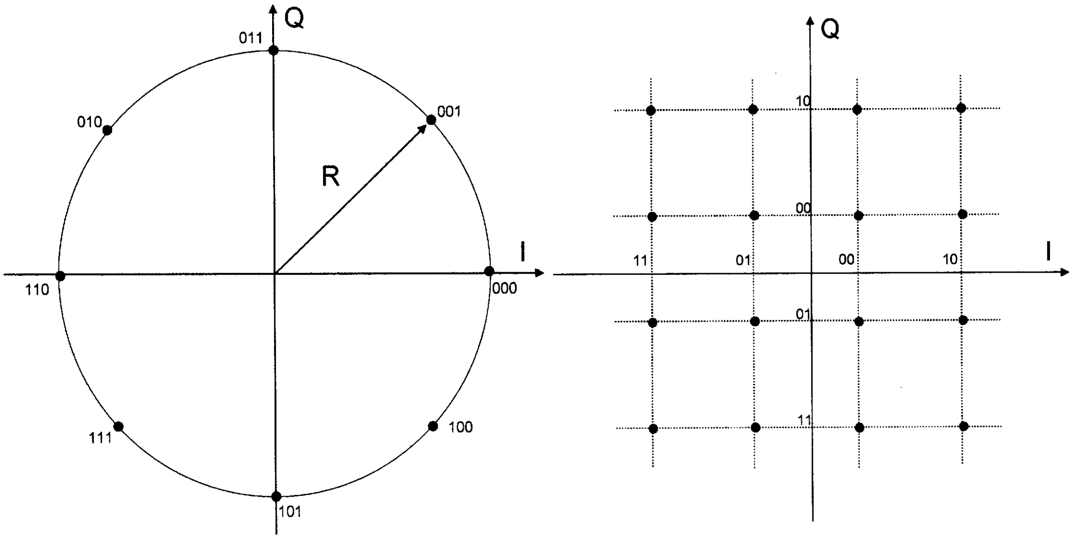


Figure 4.7-1: Sample Constellation Diagrams. Vector representation of 8PSK and 16 QAM. 8 PSK realizes 3 bits of data, 16QAM realizes 4 bits of data. The arrangement of bits in the 8PSK constellation, called Gray coding, matches the bit error rate with the nearest neighbor symbol error rate by requiring a single bit difference between nearest neighbors.

In the figure, and in general, the size of the constellations is proportional to symbol separation which is determined by a waveform's energy. Equation (4.7.2) evaluates the energy expression given in Equation (4.2.1) for the special case of quadrature waveforms. The result shows that the energy in a quadrature waveform is determined by the weight of the signals.

$$\begin{aligned} E &= \int_{-T/2}^{T/2} \|s(t)\|^2 dt = \int_{-T/2}^{T/2} \left[x(t)\Phi_{x,k}(t) + y(t)\Phi_{y,k}(t)\right]^2 dt \\ &= A_k^{\ 2}\left[\cos^2\Theta(t) + \sin^2\Theta(t)\right] \\ &= A_k^{\ 2} \end{aligned} \tag{4.7.2}$$

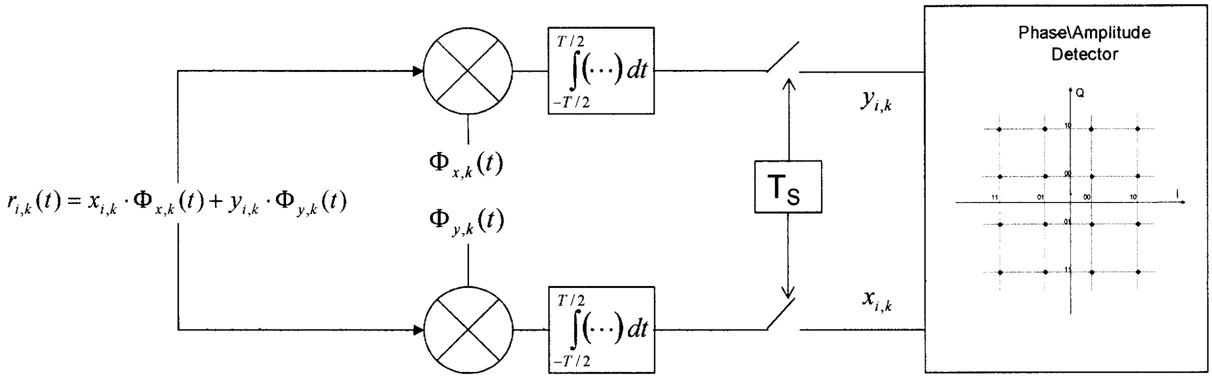
The single most important take-away from Equation (4.7.2) is that the energy in a transmitted bandpass waveform varies proportionally with separation among the coordinates in the constellation. It will be shown in Section 4.9 that error rate is also proportional to symbol separation. Therefore, in order to achieve arbitrarily low error rates, the transmitter must provide increasing levels of power to the transmitted waveform which directly impacts battery life in a portable device.

4.8 Demodulation

The Demodulation process refers to the extraction of signals from received waveforms. The process is made trivial by the material in Section 4.6. Equation (4.8.1) expresses the inner products between the received signals and the meta-basis functions defined in Equation (4.6.1).

$$\begin{aligned} x_{i,k}(t) &= \int_{-T/2}^{T/2} \mathrm{r}_{\mathrm{i,k}} \cdot \Phi_{x,k}(t)dt \\ y_{i,k}(t) &= \int_{-T/2}^{T/2} \mathrm{r}_{\mathrm{i,k}} \cdot \Phi_{y,k}(t)dt \end{aligned} \qquad \textbf{(4.8.1)}$$

To facilitate discussion in the remainder of this chapter without introducing architectural bias, the components responsible for the inner products are referred to as correlators. Figure 4.8-1 illustrates the demodulation process in a high level block diagram. In the figure, the received waveform is multiplied by the two basis functions and passed to the correlators where the inner products are formed and the results *are sampled* and passed to the decision circuitry. To reiterate, the demodulation process involves both analog and digital signal processing and is determined by the design architecture. In the absence of random fluctuations between the transmitter and the receiver, the decision circuitry is 100% capable of detecting the proper symbol sent by the transmitter. Unfortunately, various noise sources exist that introduce uncertainty into the detection process. The next section will focus on the uncertainty in the detection process owing to corruption of waveforms by noise.



$r_{i,k}(t) = x_{i,k} \cdot \Phi_{x,k}(t) + y_{i,k} \cdot \Phi_{y,k}(t)$
$\Phi_{x,k}(t)$
$\Phi_{y,k}(t)$
$\int_{-T/2}^{T/2} (\cdots)\,dt$
$\int_{-T/2}^{T/2} (\cdots)\,dt$
Ts
$y_{i,k}$
$x_{i,k}$
Phase\Amplitude Detector
Q
I
11
01
00
10

Figure 4.8-1: Demodulation Block Diagram. Demodulation differs from Figure 4.6-2 by the addition of a detection stage. This stage makes a decision about which signal was received based on the maximum likelihood statistic introduced in Section 4.9.

4.9 Detection

Over a perfect noiseless channel, infinite amounts of data can be packed into a PSK or QAM constellation simply by assigning more and more bits to each symbol and packing the symbols infinitesimally close to one another. In reality, wireless channels are imperfect and communication systems suffer from the effects of noise. In this context, noise is defined as random fluctuations in the channel that degrade the receiver's ability to correctly identify the transmitted symbols. The lowest achievable noise level, called the noise floor, is modeled by additive white Gaussian noise (AWGN) [16,17,18,20]. By addressing the issues associated with AWGN, this section will identify the final performance bottleneck in wireless communication systems, error rate.

The effect of the noise induced uncertainty is minimized by choosing the received signal that maximizes the a posteriori probability that $\vec{s}_i$ was sent given $\vec{r}_i$ was observed,

$p(\vec{s}_i|\vec{r}_i)$. Although this probability lends itself to bit error rate estimation, the "after the fact" nature of the conditional probability renders it useless in determining symbol selection criteria. Using Bayes Theorem, the condition on the a posteriori probability can be mapped to the a priori probability $p(\vec{r}_i|\vec{s}_i)$ [16]. Digressing a moment for the sake of semantics, the expression $p(\vec{r}_i|\vec{s}_i)$ is the joint probability density function (PDF) for receiving each of the symbols present in a constellation conditional on $\vec{s}_i$ having been sent.

By definition, a signal subjected to AWGN is modeled as the sum of the deterministic symbol and a Gaussian random variable [18]. From Section 4.3, a bandpass waveform is defined in two dimensional signal space. Therefore, the probability of receiving the symbol $\vec{r}_{i,k}$ given $\vec{s}_{i,k}$ was transmitted can be described by the PDF of a two dimensional Gaussian process, parameterized by period *k*. See Equation (4.9.1).

$$p(\vec{r}_{i,k}|\vec{s}_{i,k}) = p(r_{i,k,x}, r_{i,k,y}|s_{i,k,x}, s_{i,k,y}) = \frac{1}{(2\pi)^{\frac{N}{2}} \det[C_k]^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{r}_{i,k} - E(\vec{r}_{i,k}))^T [C_k]^{-1} (\vec{r}_{i,k} - E(\vec{r}_{i,k}))} \qquad \textbf{(4.9.1)}$$

In (4.9.1), $\vec{s}_{i,k}$ and $\vec{r}_{i,k}$ represent the transmitted and received signals, respectively, and $[C_k]$is the $(\vec{r}_{i,k} - \vec{s}_{i,k})$ correlation matrix.

From Section 4.8, the received waveform is processed by the correlators and sampled prior to being passed to the detection circuitry. Per the definition of AWGN, the input to each correlator in is the sum of a deterministic waveform and a stochastic noise process. Therefore, the demodulator output can be described by Equation (4.9.2). The index *z* is introduced to simplify the analysis and refers to either component of the two dimensional quadrature basis. In words, the equation says that the input of the detector is a pair of signals each corrupted by a random variable. Since $n_{z,k}$ is a random variable, then so too is $r_{iz,k}$.

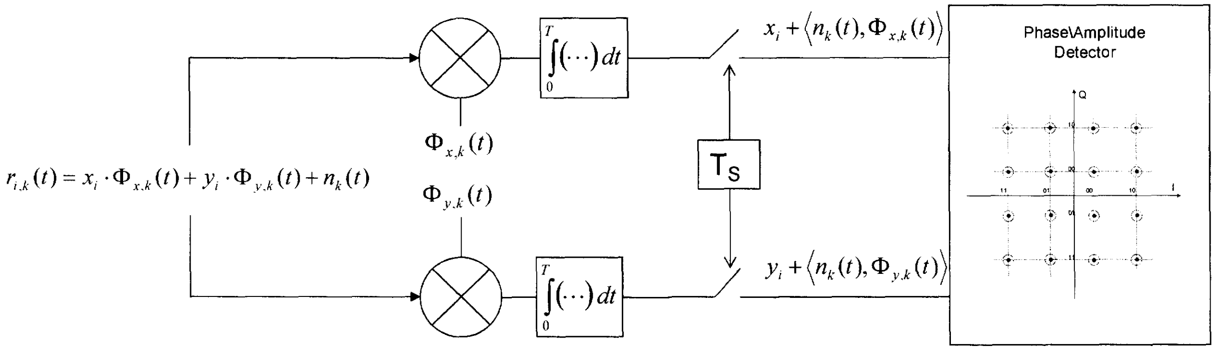
$$
\begin{aligned}
& r_{iz,k} = s_{iz,k} + n_{z,k} \\
& \text{deterministic:} \quad s_{iz,k} = \left\langle s_{iz,k}(t), \Phi_{z,k}(t) \right\rangle \\
& \text{stochastic:} \quad n_{z,k} = \left\langle n_k(t), \Phi_{z,k}(t) \right\rangle
\end{aligned}
\tag{4.9.2}
$$

For illustrative purposes, the impact of noise on the block diagram of Figure 4.8-1 is presented in Figure 4.9-1. The uncertainty represented by the red circles in Figure 4.9-1 illustrates the correlators' inability to remove noise parallel to their respective basis functions. Clearly, each correlator allows some noise to pass straight through to the decision circuitry. That being said, from Equation (4.9.2), *only* noise components parallel to the basis functions are passed by the correlators; all other noise directions are deemed irrelevant[17]. Box 4.9-1 illustrates the irrelevance principle and determines the variance that is used to populate the correlation matrix from Equation (4.9.1). Equation (4.9.3) re-states the contribution from Box 4.9-1.

$$
\left[C_k\right] = \begin{bmatrix} \sigma_{xx,k} & \sigma_{xy,k} \\ \sigma_{yx,k} & \sigma_{yy,k} \end{bmatrix} = \frac{N_0}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
\tag{4.9.3}
$$

Because the correlation matrix from Equation (4.9.3) is diagonal, the probability distribution in (4.9.1) can be maximized by maximizing the probability of the detected signals along each dimension independently.

[17] This concept is often referred to by the irrelevance principle.



$r_{i,k}(t) = x_i \cdot \Phi_{x,k}(t) + y_i \cdot \Phi_{y,k}(t) + n_k(t)$
$\Phi_{x,k}(t)$
$\Phi_{y,k}(t)$
$\int_0^T (\cdots)\,dt$
$\int_0^T (\cdots)\,dt$
T_S
$x_i + \langle n_k(t), \Phi_{x,k}(t) \rangle$
$y_i + \langle n_k(t), \Phi_{y,k}(t) \rangle$
Phase\Amplitude Detector
Q
I
11
01
00
10
10
00
01
11

Figure 4.9-1: Effects of Noise on 16QAM Receiver. This figure illustrates the process by which decisions are made by the detection circuit. The red circles centered about the symbols in the constellation illustrate the uncertainty injected into the system by noise. The correlators pass only noise parallel to the respective basis functions. The x and y coordinates define a circle in 2-D Euclidean space. The size of the circles grows inversely with signal to noise ratio.

Box 4.9-1. Referring back to the joint PDF of Equation (4.9.1); clearly, if $n_k(t)$ is a zero-mean process, the term $E\left(\vec{r}_{i,k}\right)$ is simply $\vec{s}_{i,k}$. If this were not the case, an offset is introduced immediately following the correlators to correct for the noise bias. Nonetheless, the expected value of $\vec{r}_{i,k}$ is exactly $\vec{s}_{i,k}$. With this analysis in mind, the correlation matrix can be populated with the help of Equation (4.9.4).

$$\begin{aligned}\sigma_{zz',k} &= E\left([r_{iz,k}-\mu_{iz,k}][r_{iz',k}-\mu_{iz',k}]\right)\\ &= E\left([s_{iz,k}+n_{z,k}-s_{iz,k}][s_{iz',k}+n_{z',k}-s_{iz',k}]\right)\\ &= E\left(n_{z,k}n_{z',k}\right)\end{aligned} \qquad \mathbf{4.9.4}$$

The variance of the correlator outputs, $E\left(n_{z,k}n_{z',k}\right)$, is derived in Equation (4.9.5).

$$\begin{aligned}E\left(n_{z,k}(t)n_{z',k}(\tau)\right) &= E\left(\int_{-T/2}^{T/2} n_k(t)\cdot\Phi_{z,k}(t)dt\int_{-T/2}^{T/2} n_k(\tau)\cdot\Phi_{z',k}(\tau)d\tau\right)\\ &= \int_{-T/2}^{T/2}\int_{-T/2}^{T/2} E\left(n_k(t)n_k(\tau)\right)\cdot\Phi_{z,k}(t)\Phi_{z',k}(\tau)\,dt\,d\tau\\ &= \int_{-T/2}^{T/2}\int_{-T/2}^{T/2}\frac{N_0}{2}\delta(t-\tau)\cdot\Phi_{z,k}(t)\Phi_{z',k}(\tau)\,dt\,d\tau\\ &= \begin{cases}\dfrac{N_0}{2} & \mathrm{z}=\mathrm{z'}\\ 0 & \mathrm{z}\neq\mathrm{z'}\end{cases}\end{aligned} \qquad \mathbf{4.9.5}$$

The emergence of the term $\frac{N_0}{2}\delta(t-\tau)$ in Equation (4.9.5) is a consequence of the wide sense stationary and ergodic properties of white noise and derives from its autocorrelation function. The N_0 term is the power in the noise signal and the factor of two derives from the model used to represent noise in a bandpass system [18]. The contaminated received signals were shown to be uncorrelated and time independent. The former is established by the orthogonal bafunctions that are used to waveforms. The later is explained with the help from the following figure.

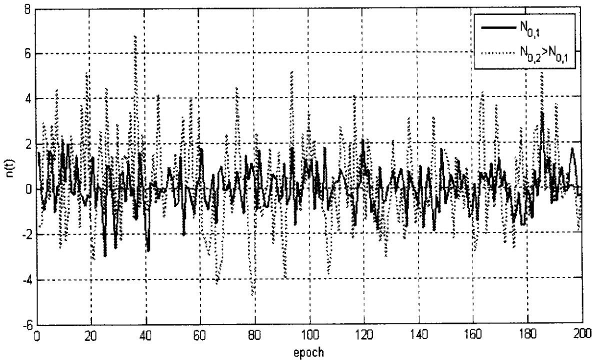


Figure 4.9-2: White Gaussian Noise. The noise floor in a wireless communication system is modeled by white Gaussian noise. The figure shows two noise processes. Both have zero mean. The red noise process clearly exhibits more variation than the blue noise process. The randomness of a given noise process suggests that its autocorrelation is exactly zero for all non-zero time shif.

The figure illustrates a random process that is completely uncorrelated from one epoch to the next. The correlation is non-zero only when the two correlated waveforms are measured concurrently as is evident by the emergence of the delta function in Equation (4.9.5).

Before proceeding, an interesting thought experiment will help bring the theory back to the realm of practical application. Assume for a moment that an ASK bandpass waveform is received. The ASK modulation scheme allows one of two possible states *per dimension,* ±A[18]. Equation (4.9.3) says that to maximize the joint distribution function, the probability function along each basis direction can be maximized independently. The observed signal along the in-phase direction is generated by a sample of the inner product between the received waveform and the in-phase correlator. If the in-phase correlator output is closer to A than –A, then the probability $p\left(r_{ix,k}\middle|A\right) \geq p\left(r_{ix,k}\middle|-A\right)$. Therefore, assigning $s_{ix,k}$=A maximizes the a priori conditional probability function. Again, because the effect of noise on either component of the quadrature waveform is uncorrelated with the other, the same logic can be used to determine $s_{iy,k}$. Moreover, the independent decisions maximize the overall PDF in Equation (4.9.1).

The thought experiment suggests that the probability distribution $p\left(\vec{r}_{i,k}\middle|\vec{s}_{i,k}\right)$ is maximized by choosing the most likely signal from the set of allowed signals per each direction in the basis function. To test the thought experiment, the correlation matrix in Equation (4.9.5) is inserted into (4.9.1).

$$\begin{aligned} p(\vec{r}_{i,k}|\vec{s}_{i,k}) &= \left(\frac{N_0}{2}2\pi\right)^{-\frac{N}{2}} e^{-\frac{1}{N_0}(\vec{r}_{i,k}-\vec{s}_{i,k})^T(\vec{r}_{i,k}-\vec{s}_{i,k})} \\ \ln\left(p(\vec{r}_{i,k}|\vec{s}_{i,k})\right) &= -\frac{N}{2}\ln(\pi N_0) - \frac{1}{N_0}\left(\vec{r}_{i,k}-\vec{s}_{i,k}\right)^T\left(\vec{r}_{i,k}-\vec{s}_{i,k}\right) \end{aligned} \qquad \textbf{(4.9.6)}$$

[18] This scheme exactly describes QPSK or 4QAM modulation.

Clearly (4.9.6) is maximized by minimizing the noise or, as illustrated by Equation (4.9.7), by minimizing the vector magnitude separation between the received signal and the allowable signals.

$$\left(\vec{r}_{i,k}-\vec{s}_{i,k}\right)^T\left(\vec{r}_{i,k}-\vec{s}_{i,k}\right)=\left\|\vec{r}_{i,k}-\vec{s}_{i,k}\right\|^2=\left(r_{ix,k}-s_{ix,k}\right)^2+\left(r_{iy,k}-s_{iy,k}\right)^2 \qquad \textbf{(4.9.7)}$$

The practice of assigning symbols based on minimizing the distance between observed and allowable signals in the context of maximizing the conditional distribution is called the maximum likelihood principle [16]. The last term in the equality of (4.9.7) explicitly confirms the previous thought experiment. That is to say, choosing the symbol that is geometrically closest to an available symbol yields the highest probability of properly detecting the transmitted information.

Of particular interest in the design of communication systems is the error rate performance of different modulation schemes. Errors occur when the maximum likelihood principle is used to assign the *wrong* symbol to a received waveform. The error phenomena are captured in Equation (4.9.8).

$$\begin{aligned} p\left(\left\|\vec{r}_{i,k}-\vec{s}_{i,k}\right\|^2>\left\|\vec{r}_{i,k}-\vec{s}_{i',k}\right\|^2\middle|\vec{s}_{i,k}\right)&=p\left(\left\|\vec{s}_{i,k}+\vec{n}_k-\vec{s}_{i,k}\right\|^2>\left\|\vec{s}_{i,k}+\vec{n}_k-\vec{s}_{i',k}\right\|^2\middle|\vec{s}_{i,k}\right)\\ &=p\left(\left\|\vec{n}_k\right\|^2>\left\|\vec{s}_{i,k}-\vec{s}_{i',k}+\vec{n}_k\right\|^2\middle|\vec{s}_{i,k}\right) \end{aligned} \qquad \textbf{(4.9.8)}$$

In words, (4.9.8) says that an error occurs when the absolute value of the noise exceeds the absolute value of the sum of the noise and the difference between signals. This observation is illustrated in Figure 4.9-3 for a QPSK\4QAM system. From Figure 4.9-3, an error occurs when the noise component anti-parallel to the difference vector is exactly half the separation between the two symbols being examined. Consequently, the error threshold between any two points in the constellation is $\frac{d_{ii'}}{2}$ where $d_{ii'}$ is the separation between symbols i and i'.

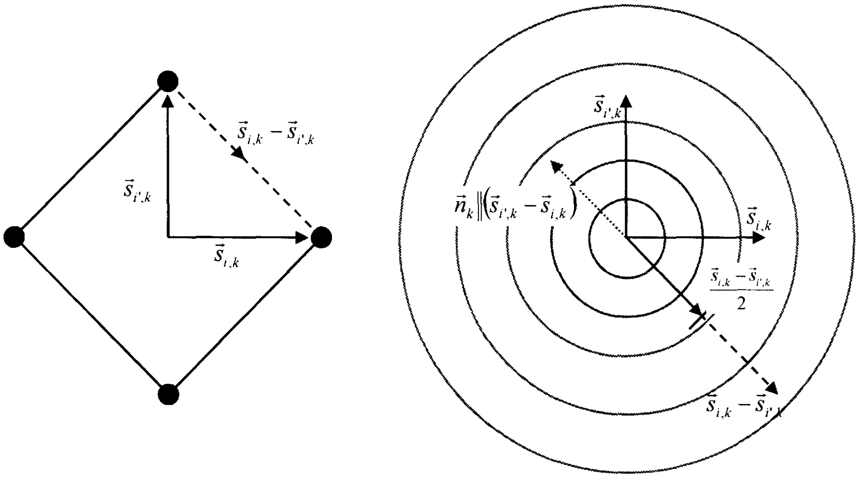

$\vec{s}_{i,k} - \vec{s}_{i',k}$
$\vec{s}_{i',k}$
$\vec{s}_{i,k}$
$\vec{s}_{i',k}$
$\vec{n}_k \| (\vec{s}_{i',k} - \vec{s}_{i,k})$
$\vec{s}_{i,k}$
$\frac{\vec{s}_{i,k} - \vec{s}_{i',k}}{2}$
$\vec{s}_{i,k} - \vec{s}_{i',k}$

Figure 4.9-3: Symbol Errors. The figure on the left illustrates a 4QAM\QPSK constellation diagram. The difference vector from Equation (4.9.8) is illustrated on the left by the dotted line and redrawn from the origin on the right. Only when the noise parallel and opposite to the difference vector exceeds half of the difference vector does an error occur. This is shown in the figure on the right by the dotted blue line exactly anti-parallel to the difference vector and half its length. To generalize, an error occurs when a detected symbol crosses the perpendicular bisector between the correct symbol and any others in the constellation.

The error criteria described above can be used to upper bound the probability of error for a given constellation by assuming the minimum distance between any two symbols is replicated throughout the constellation. The minimum distance $d_{\min}$ is determined by nearest neighbor separation. Figure 4.9-5 illustrates the process used to assign an upper bound on the error probability for an 8PSK modulation scheme. The nearest neighbors are illustrated by the dotted blue lines labeled $d_{\min}$.

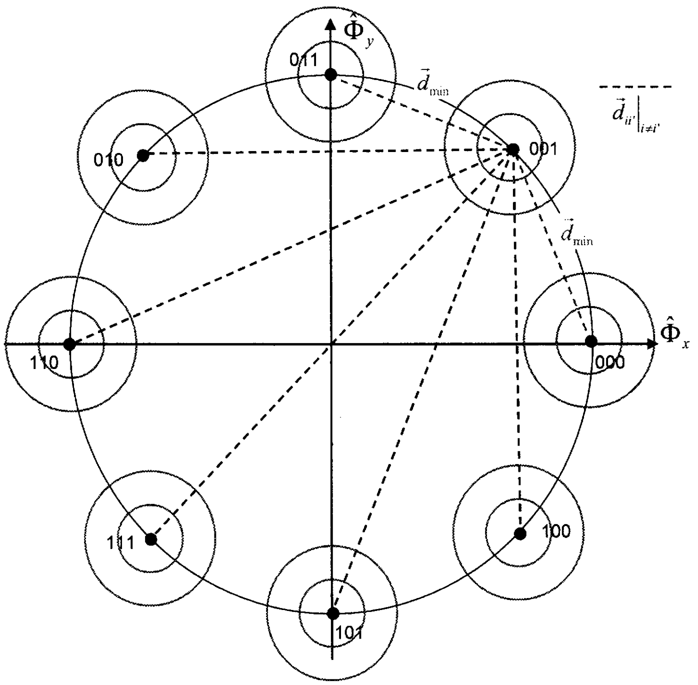


Figure 4.9-4: Conceptual Error Diagram for 8PSK Modulation Scheme. Splitting the perpendicular bisectors between each symbol determines the error domain in a constellation. A sum over each of the coordinates in the constellation will determine the domain in an error test. The errors can be upper bounded by assuming the minimum separation is repeated between all coordinates; or can be reasonably approximated by only counting nearest neighbors.

Since the effect of noise is characterized by a jointly Gaussian distribution, the probability of error is determined by normalizing the error threshold by the standard deviation of the corrupted waveform given by Equation (4.9.3) then simply referencing error-function tables. The normalized threshold is presented in Equation (4.9.9).

$$z = \left(\frac{d_{ii'}}{2} \frac{1}{\sqrt{\sigma_{xy}}} \right) = \frac{d_{ii'}}{2} \sqrt{\frac{2}{N_0}} = \frac{d_{ii'}}{\sqrt{2N_0}} \qquad \textbf{(4.9.9)}$$

Figure 4.9-5 illustrates the affect of a noisy environment on the probability of error. From the figure, the noise level in the red environment is higher than that of the blue environment. Subsequently, a user in both environments receiving identical waveforms of equal power will observe more errors in the red environment than in the blue[19]. Combining Equation (4.9.9) with the nearest neighbor approximation provides the upper bound for the probability of error for a given constellation [20]. The shortcoming of this approximation is that it is overly pessimistic. A more objective approximation is made by considering only the nearest neighbors. This approximation is referred to as the nearest neighbor approximation and is given in Equation (4.9.10).

$$Pe \approx M_{d_{MIN}} Q\left(\frac{d_{\min}}{\sqrt{2N_0}}\right) = M_{d_{MIN}} Q\left(\sqrt{\frac{d_{\min}{}^2}{2N_0}}\right) \qquad \textbf{(4.9.10)}$$

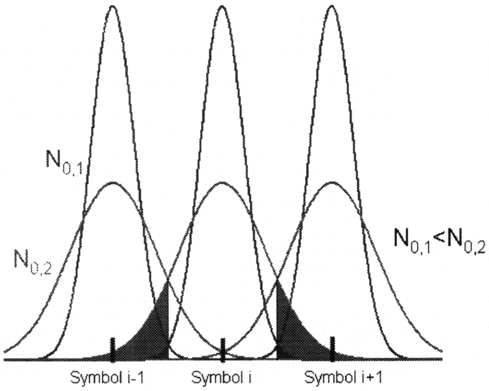


Figure 4.9-5: Effect of N_0 on probability of error. The variance in the red distribution is much larger than that of the blue distribution. The shaded areas in the tails start at $d_{min}/2$ and capture the probability of detecting the wrong signal.

The presence of d_{min} in the numerator of the argument in (4.9.10) indicates that the probability of error is proportional to symbol separation. See Figure 4.9-4. From Section 4.7, separation in signal space is determined by the amplitude of the received waveform. From Equation (4.7.2), the amplitude of the received waveform is the square root of the

[19] The red and blue distributions match the noise figure from Box 4.9-1.

waveform energy, $E=A_i^2$. Taken together, the error probability can be reduced by simply pumping more power in to the transmitted signal.

In the final analysis, this section illustrates a prominent architectural bottleneck; namely- the error rate of a bandpass waveform in the presence of noise. Even with access to an infinite power supply, saturation of solid state amplifiers provides an upper limit to signal separation and subsequently achievable error rates[20]. Despite the upper bound, "power level" officially joins the list of design parameters that was presented in Section 4.5 and in doing so re-enforces the battery life bottleneck. To the extent that power level impacts both error rate and battery life, the two are related. Moreover, Section 4.10 will formally present the interrelationships among the bottlenecks and design variables exposed in the first nine sections of this chapter. Nonetheless, to be consistent with Section 4.7, the design parameter "signal set" is replaced with "modulation scheme". The revamped list includes modulation scheme, baud rate, pulse shape, and power level; and the list of bottlenecks now include: data rate, error rate, and battery life. The constraint inherited by this section is noise level[21].

4.10 Spectral Efficiency and Power

The previous section closed the set of system level performance bottlenecks present in wireless communication systems. The final set consists of: data rate, error rate, and battery life. A comparison of the performance of each modulation scheme discussed in Section 4.7 along the bottleneck dimensions is presented in Figure 4.10-1. In the figure, the x-axis identifies the independent variable, SNR. In general, the noise is a constraint determined by

[20] It is worth mentioning that the discussion in this section was focused on symbol error rates; that are different from bit error rates for all but binary symbols [20].

[21] Wireless communication is affected by other more dominant noise sources such as multi-path and fading. Unfortunately, this level of engineering detail is beyond the scope of this thesis. For more information, see the text by Goldsmith [20].

the environment, so the x-axis captures the design variable: power level which maps directly to the battery life bottleneck. The y-axis quantifies the spectral efficiency of each modulation scheme at a given error rate. Spectral efficiency is simply the available data rate per unit of bandwidth. From Section 4.4 and 4.5, bandwidth is determined by the design variables baud rate, pulse shape, and symbol variance which is, in turn, determined by the modulation scheme. Therefore, spectral efficiency is a general metric that can be used to compare data rates for various communication systems as a function of the aforementioned design variables and waveform power subject to a particular error rate. In as much, each of the design variables and bottlenecks has been captured by the curves in the figure. The error rates presented in the figure, 1% and 0.001%, were determined by first assuming the nearest neighbor approximation described in Section 4.9 for each of the modulation schemes [20].

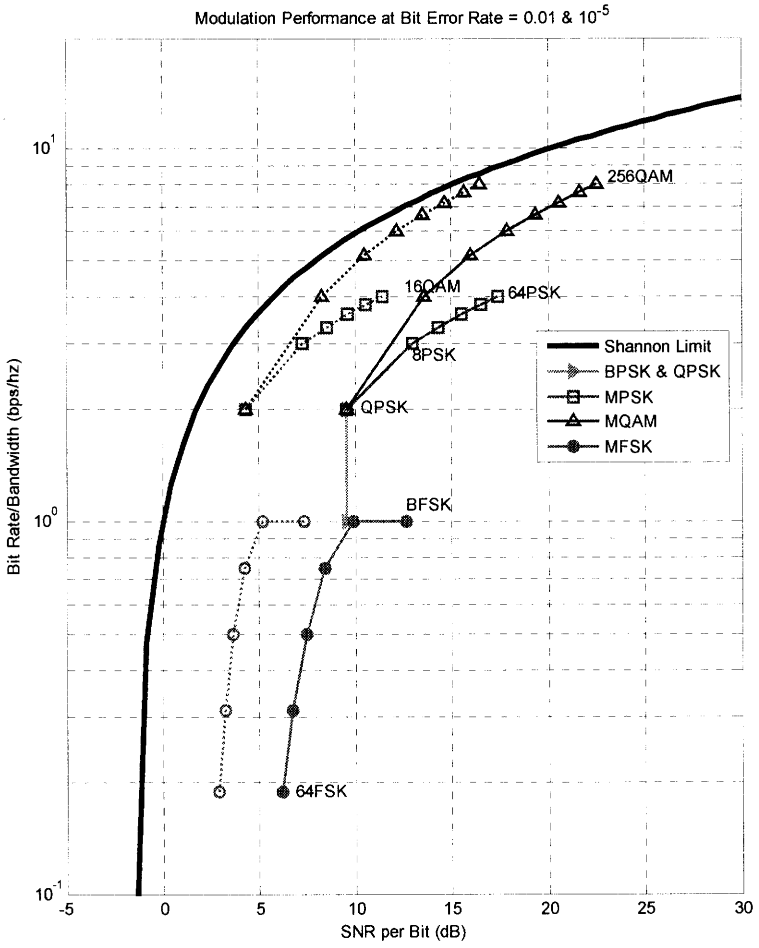


Figure 4.10-1: Comparison of Modulation Schemes. The solid line curves illustrate optimal performance levels of systems designed for 10^{-5} error rate. In comparison, the dotted lines illustrate performance of the same modulation schemes characteristic of a 1% error rate. Clearly, the only difference is the SNR required to meet error rate expectations. Assuming noise is inherited by the environment, the power level variable is isolated by the different families of curves.

The black curve in Figure 4.10-1 invokes the famous Shannon Theorem which bounds the available capacity of a white Gaussian noise channel. Shannon's Theorem is given in Equation (4.10.1) and can be found in most communication and information theory text books [16,18].

$$C = B\log\left(1+\frac{S}{N}\right) \qquad \textbf{(4.10.1)}$$

The expression simply states that channel capacity, in units of bits per second (bps), is proportional to the bandwidth of the bandpass waveform and the base 2 logarithm of the signal to noise ratio. From the figure, the family of curves approaches the Shannon Limit as the expected error rate is allowed to grow. This observation is exploited by system designers who include redundancy in the transmitted information, called channel coding\decoding, in order to achieve lower error rates than the figure suggests; however at the expense of data rate. Unfortunately, channel coding is outside the scope of this thesis, but indeed represents an additional design parameter.

The curves in the figure demonstrate the compromises inherent in capacity, power, and bandwidth. In the limit of maximum capacity and fixed modulation scheme, data rates existing on the Shannon frontier can only grow with additional spectrum allocation-regardless of power. Notwithstanding, the migration from QPSK to 8PSK provides a noticeable bump in available data rate, but at the expense of power consumption.

4.11 Bottlenecks in Wireless Communication Systems

The remainder of this chapter distills the engineering results into a form that is consistent with the goal of this thesis. Table 4.11-1 aggregates all of the information from this chapter in a table that maps the design parameters to the performance bottlenecks. The first four columns list the design variables that were identified in this chapter. The remaining two columns refer to the design variables introduced by the physical artifacts that embody the physical layer design (formally introduced in Chapter 6). The last row in the table is merely a placeholder that will be employed later to explain the migration of bottlenecks within the value network. The table certainly illustrates the high level of interconnectedness that binds the performance bottlenecks to the design variables. Unfortunately, the table does little to illuminate the competitive strategies pertaining to the layers of the wireless value network.

	Modulation Scheme	Baud Rate	Pulse Shape	Signal Power	Noise Added	Electronic Power Consumption
Data Rate	X	X	X	-	-	-
Battery Life	X	X	X	X	-	X
Error Rate	X	-	X	X	X	-
Experience	-	-	-	-	-	-

Table 4.11-1: Matching Design Parameters to Bottlenecks.

Recall that Chapter 3 presented a framework wherein firms determine internal development and outsourcing strategies to maximize ROIC by identifying and claiming bottlenecks. Table 4.11-2 assigns the dominant strategies to the three layers in the value network. This table includes an additional column for standards bodies. In an ideal world, the standards bodies exist to serve the entire value network by creating a set of design rules that spans the interconnected design variables presented in the previous table. In the real world, a standards body may very well be serving a single layer or firm in the value network. Nonetheless, bottlenecks are attributed to competitive firms and therefore are not relevant to the standards bodies. In the table, the letter L indicates legacy identification with a particular bottleneck whereas the C indicates contemporary ownership of a particular bottleneck.

	Network Operators	Handset ODMs	Component Suppliers	Standards Bodies
Data Rate	L	-	-	-
Battery Life	-	L	C	-
Error Rate	-	L	C	-
Experience	L	C	-	-

Table 4.11-2: Matching Value Network to Bottlenecks.

Table 4.11-2 illustrates the dominant characteristics and historical trends in the industry. The first column illustrates the network operators' stakes in the game. They possess the spectrum which, from Section 4.10, dominates the data rate performance bottleneck by imposing an active constraint on the amount of transmissible data. What's more, the network operators, until recently, have maintained possession of the user experience bottleneck in an attempt to control the quality of service across handset platforms. Slowly, ODMs such as Apple are wrestling the user experience bottleneck away from the Network Operators. Downstream, the Handset ODMs have traditionally focused on offering peak performance within the constraints of the data rate bottleneck, by focusing on battery life, error rate, and most recently - user experience. The component suppliers have also staked claim on the battery life and error rate bottlenecks, but were beholden to the handset ODM's. Chapter 6 will show that by effectively splitting functionality ascribed to the transmitter and receiver, the legacy (L) handset ODMs enjoyed the option value associated with the substitution operator as well as the benefits of a reduced manufacturing and design footprint. However, changes in the architecture have shifted ownership of the error rate and battery life bottlenecks to the contemporary (C) component suppliers.

The final table in this chapter, Table 4.10-3, illustrates physical layer design hierarchy by assigning possession of design variables to layers in the value network. The table illustrates the primacy of the standards bodies in that they create the design rules that pertain to all members of the value network. The Network Operators merely choose from available standards and pass the design rules downstream to the handset ODMs. Per the discussion in Section 3.4, value is proportional to the number of levels of hierarchy in a particular design.

The fact that ODMs and component suppliers share the focus on battery life and error rate is immaterial from a modular value point-of-view, provided one is not cannibalizing the other. That being said, value is most likely destroyed when levels of hierarchy are squeezed out of a design if the conditions for modularity are not nurtured by the firms assuming control of the design. For example, if component suppliers gain sole possession of the battery life and error rate bottlenecks, then level 3 in Table 4.10-3 is effectively eliminated to the dismay of all upstream members of the value network.

	Modulation Scheme	Baud Rate	Pulse Shape	Signal Power	Noise Added
Network Operators	2				-
Handset ODMs	3	3	3	3	1
Component Suppliers	4	4	4	4	2
Standards Bodies	1				-

Table 4.11-3: Matching Design Parameters to Value Network.

5 Wireless Communication Standards

The importance of standards was introduced in Section 4.11. To reiterate, standards drive the uncertainty out of the wireless system architecture and provide a set of design rules that enable a modular value network to emerge. All contemporary wireless communication systems are embodied by standards. At the highest level of abstraction, the standards can be divided into two types, time domain multiple access (TDMA) and code division multiple access (CDMA). Each standard is determined by the format used to support simultaneous users and various data rates. TDMA systems enable multiple concurrent users by granting exclusive access to specific communication channels for a well defined time periods, called slots. TDMA standards include the dominant Global System for Mobile Communications (GSM) and its data derivatives: General Packet Radio Service (GPRS) and Enhanced Data Rates for GSM Evolution (EDGE) [21]. CDMA systems enable multiple users by allowing each to share the entire communication band simultaneously. CDMA standards include those pioneered by Qualcomm such as cdmaONE and cdma2000 and its data derivatives, Evolution-Data Optimized (EV-DO) as well as new wideband CDMA systems being deployed with traditional GSM networks such as the Universal Mobile Telecommunications Service (UMTS).

5.1 Time Division Multiple Access

Owing to the ubiquity of the GSM standard, TDMA systems are considered the dominant design in the wireless network industry. TDMA networks assign users to non-overlapping frequency channels and time slots. Multiple slots are aggregated into frames. To prevent interference, groups of channels are *spatially* separated among base station cells [21].

The design of an idealized hexagonal GSM network is illustrated in Figure 5.1-1. Each of the colors indicates a different allocation of available channels for the transmit and receive bands, respectively. For example, the channel set corresponding to the blue base stations is not reused in any of the adjacent cells.

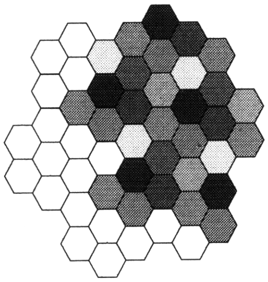


Figure 5.1-1: TDMA Cellular Network [27].

From the figure and the previous discussion, a cell's subscriber capacity is determined by the number of available channels, the number of time slots per frame, and an interleaving constant. Interleaving refers to the practice of assigning users to alternate frames. For example, half rate communication is realized by granting users access to every other frame. The capacity of an ideal hexagonal GSM network is given in Equation (5.1.1)[22].

$$\text{Network Capacity} = \frac{Ch}{7} \cdot N_{Slot} \cdot K_{Interleave} \qquad \textbf{(5.1.1)}$$

In particular, the GSM standard specifies a transmit and receive band each comprised of 124 200kHz channels spanning 25 MHz. From equation (5.1.1), the number of simultaneous users per cell for full rate voice communication is limited to 142. Equation (5.1.1) also identifies a key relationship between network capacity and bandwidth. Stated simply,

[22] Note, frequency hopping enables much higher level of frequency re-use.

network capacity in a TDMA system is directly proportional to the number of spectrally defined channels.

In GSM systems, the data rate performance bottleneck was established by legacy digital voice communication. Moreover, the GSM standard ensures full rate voice communication of 13kilo-bits per second (kbps) or half rate voice communication of 5.6 kbps by interleaving two users. The demand for higher data rates led to the emergence of GPRS, EDGE, and subsequent standards. The evolution of wireless communication standards predicated on the GSM base standard is presented in Table 5.1-1. Judging from the table, GPRS and EDGE are evolutionary data variants of GSM that appear to be backward compatible with GSM TDMA architecture. It can be shown that only minor upgrades on the part of the network operators are required to migrate legacy GSM networks to GPRS and EDGE [22]. In particular, GPRS introduced packet switching and variable data rates into the GSM standard[23]. EDGE followed with the 8PSK modulation scheme that enabled higher data rates. From Chapter 4, higher modulation schemes require higher SNR or more redundancy in the form of channel coding to realize acceptable error rates. Options in the form of variable modulation schemes and redundancy were included in the GPRS and EDGE standards to provide users with the flexibility to operate under different noise conditions.

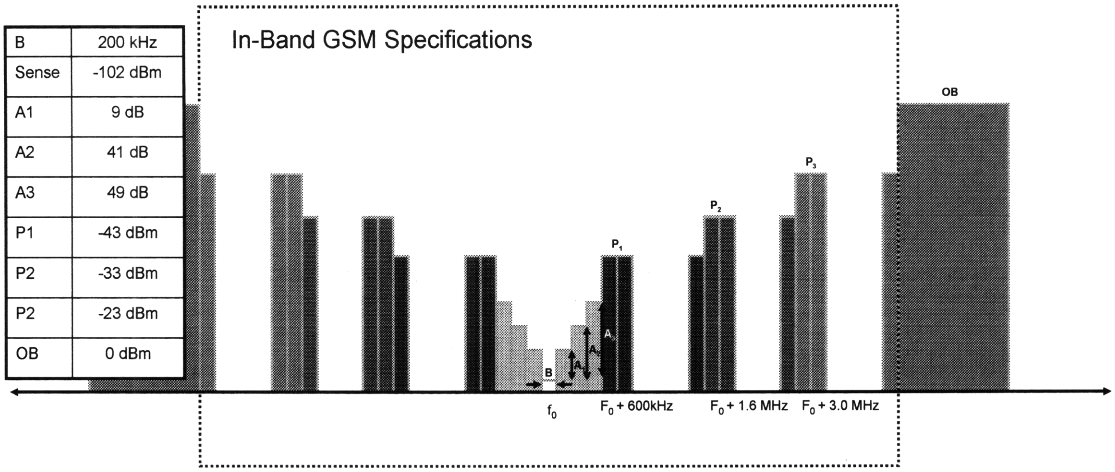
[23] Packet switching refers to asynchronous data transmissions wherein information is communicated via one or many allocated channel in bursts. Channels are released on completion of the packet communication event. The alternative to packet switching is called circuit switching wherein channels are dedicated for the duration of the entire connection. Packet switching enables higher spectrum utilization in a TDMA network.

	GSM	GPRS	EDGE
TX Band	25 MHz	25 MHz	25 MHz
RX Band	25 MHz	25 MHz	25 MHz
Channel Bandwidth	200 kHz	200 kHz	200 kHz
Modulation Scheme	GMSK (FSK)	GMSK (FSK)	GMSK & 8-PSK
Baud Rate	270 ksps*	270 ksps*	270 ksps*
Bit Rate	270 kbps	270 kbps	270 kbps
Frame Duration	4.615 ms	4.615 ms	4.615 ms
Slot Number	8	8	8
Bit Rate/Slot	33.75 kbps	33.75 kbps	101.25 kbps
Data Rate/Slot (payload)	13 kbps	22.8 kbps	69.2 kbps
Available User Data Rate/Slot		8 kbps-20 kbps	8.8 kbps-59.2 kbps
Available User Data Rate/Frame		64 kbps-160 kbps	70.4 kbps-473.6 kbps

Table 5.1-1: GSM System Performance. The GSM standard divides the 25MHz band into 124 channels of 200 kHz each. Each channel supports a 270 kbps waveform that actually violates the Nyquist criteria. Control of channels is divided up into time slots that are aggregated into frames. GPRS enables packet switched data and works with the existing GSM standard. The EDGE standard improves data rates by providing higher order modulation schemes. Higher effective data rates are realized by granting users control over multiple slots

To ensure minimum data rates are delivered to subscribers, the receiver sensitivity, neighboring channel power levels, and SNR is specified by the GSM standard. Figure 5.1-2 illustrates the European GSM 900MHz specification for receiver designs [23]. The importance of out-of-channel interference is discussed in Chapter 6. For the time being, suffice it to say that non-idealities in components map power in adjacent channels to the target channel, thereby raising the effective noise level[24]. Likewise, the transmitter is well specified to ensure that transmitted waveforms are confined to the proper channel. Figure 5.1-3 illustrates the shape of a transmitted waveform that specified by the GSM standard [23].

[24] Although it is common practice to differentiate noise from interference by using SNR or SIR (signal to interference ratio), the effects on bit error rate are the same. Rather than introducing another acronym, noise and SNR will refer equally to physical noise as well as interference.



B	200 kHz
Sense	-102 dBm
A1	9 dB
A2	41 dB
A3	49 dB
P1	-43 dBm
P2	-33 dBm
P2	-23 dBm
OB	0 dBm

Figure 5.1-2: GSM 900 Receiver Specifications. Owing to non-idealities in the receiver, the receiver design specifications must include power levels of both out-of-band waveforms as well as adjacent channels. The onus falls on the designer to meet or exceed specifications.

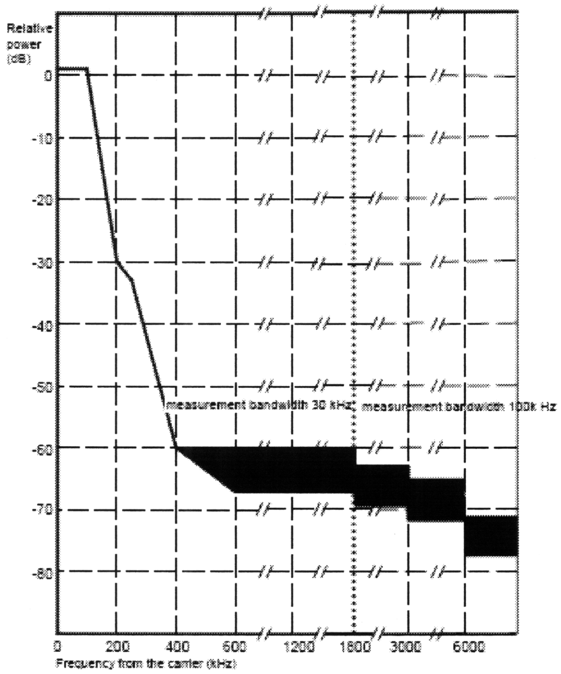


Figure 5.1-3: Transmitter Frequency Mask. As was the case with the receiver, non-ideal component technologies leave open the possibility that the transmitter will inject power into adjacent channels. This shortcoming is addressed by introducing strict transmitter specifications.

5.2 Code Division Multiple Access

Rather than parsing up spectral real estate, CDMA systems allow users to *share* available spectrum. CDMA is achieved by encoding each user's waveform in such a way that it appears as noise to all but the intended receiver [24]. Disguising various waveforms as noise requires spreading the original data waveform about a larger frequency band than is required to support the target data rate. As a consequence, CDMA systems are commonly referred to as spread spectrum systems. To enable spectrum sharing, CDMA assigns unique orthogonal codes to transmissions that when correlated with non-intended receivers appear as noise. The transmitted waveforms are generated by multiplexing the user data with a much higher rate "chip" stream prior to pulse shaping. Since chips are transmitted over the wireless channel, it is the chip rate that must conform to the Nyquist criteria introduced in Chapter 4. See Figure 5.2-1 for a simple illustration of the encoding process [25].

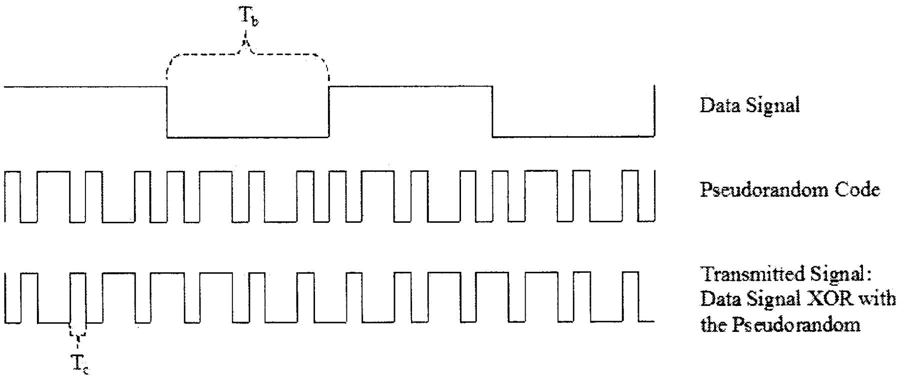


Figure 5.2-1: CDMA Data Encoding. The data defined by period T_b is multiplexed with a pseudorandom code to create a chip stream of period T_c. Note, the chip rate is many times faster than the actual data rate. Since the chips are transmitted over a wireless channel in a CDMA system, then the chip waveform must comply with the Nyquist criteria [26].

The *network* architecture enabled by the CDMA communication standard is illustrated in Figure 5.2-2. Similar to the TDMA network, the base stations are arranged in a grid wherein each cell supports a localized group of subscribers. However, as the figure clearly depicts, CDMA networks differs from TDMA networks by allowing network operators to *reuse* spectrum in each cell. Intra-cell waveform collisions are avoided by multiplexing information with distinct orthogonal codes [24]. Unfortunately, CDMA network architecture does not come without its share of challenges. Recall that error rates are determined by SNR. The noise like property of overlapping waveforms can render the entire band unusable by raising the noise floor. The onus lies on the network designers to minimize interference by regulating the amount of power that mobile stations transmit [27].

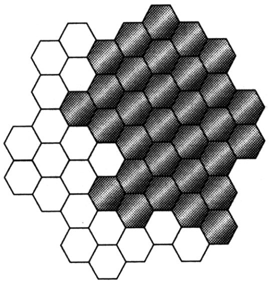


Figure 5.2-2: CDMA Cellular Network. The full band is used in each cell [27].

In the absence of redundancy, the maximum data rate in a CDMA system is determined by the static chip rate[25]. Per the discussion in Chapter 4, the error rate bottleneck, is determined by the modulation scheme, pulse shape, and transmit power design variables. Assuming the first two variables are fixed by the CDMA standard, then acceptable error rates are determined exclusively by SNR. Equation (5.2.1) provides a simplified expression for the signal to noise ratio in terms of: energy per bit, noise contributed by the receiver ($F_N K_0 TB$), and noise generated by other users in the cell.

$$\frac{E_b}{N_0} = \frac{\dfrac{P_j}{R_S}}{\dfrac{F_N K_0 TB}{B} + \dfrac{\sum_{i \neq j}^{N} P_i}{B}} \qquad \textbf{(5.2.1)}$$

[25] As will be demonstrated, this upper limit is accompanied by significant channel coding.

The numerator is the $\mathfrak{PSD}$ of the CDMA waveform divided by the symbol rate of the waveform. The denominator is the sum of the noise power contributed by the receiver and all overlapping channels. Both terms are normalized by the channel bandwidth to convert power to $\mathfrak{PSD}$. The constants F_N and K_0T are the receiver noise figure and thermal effects, respectively. Both of these terms fall into the "Noise Added" category in Figure 4.11-1. Good designs possess the quality that receiver noise is dominated by overlapping channel noise – not noise generated by the receiver modules. That being said, the user generated noise is directly proportional to the number of subscribers on the system and inversely proportional to the bandwidth occupied by the waveform. Therefore, as was the case for TDMA systems, network capacity in CDMA systems is proportional to bandwidth by way of acceptable error rates.

Equation (5.2.1) offers a great deal of insight into the context dependant bottlenecks unique to CDMA systems. The numerator is proportional to the *data* rate; therefore slower *data* rates are less sensitive to noise. The noise floor is determined by the noise added by the receiver design which comprises one of the modes of competition among handset ODMs and suppliers. The second term in the denominator is the most interesting from a network performance point of view. From Equation (4.4.5) and (4.5.1), slow waveforms contain less power than waveforms derived from faster symbol rates. Therefore, the contributions to noise due to overlapping waveforms are much lower if the overlapping waveforms derive from slower pulse rates. What's more, because the interference noise in the denominator is inversely proportional to the real bandwidth of the transmitted signal, i.e. the chip rate, higher chip rates facilitate more user capacity within the cell. Taken together, low speed control channels will contribute less noise than high speed data channels.

The CDMA IS-95 (also called CDMAone) sets the *chip rate* at 1.2288 Msps (mega symbols per second). Similar to the GSM standard, the *bit rate* for IS-95 was inherited by voice communication. The CDMAone standard allows voice transmission to assume one of four possible rates depending on conversation activity: [1200, 2400, 4800, 9600] bps [24]. The encoded voice data is subjected to a 1:3 convolutional channel coder prior to modulation. Subsequently, the maximum bit rate passed to the modulator is 28.8 kbps [28]. Under better noise conditions, the 1:3 convolutional coder is replaced by a 1:2 coder enabling effective data rate of 14.4kbps (up from 9.6 kbps). To make use of this flexibility and to optimize waveform power, each channel is divided into temporal frames called power control groups that last 20ms [24]. The data rate and power of each power group is variable and chosen to optimize overall system performance.

The information from the previous paragraphs can be used to estimate CDMA network capacity, *N*. Equation (5.2.2) provides a simplification of the previous SNR expression wherein the power of each signal is assumed to be equivalent. This assumption can be shown to maximize performance [27].

$$\frac{E_b}{N_0} = \frac{\dfrac{P}{R_S}}{\dfrac{F_N K_0 T B}{B} + \dfrac{(N-1)P}{B}} \approx \frac{\dfrac{B}{R}}{(N-1)} \qquad \textbf{(5.2.2)}$$

The expression on the far right assumes that the first term in the denominator is dominated by the second. The log-scale equivalent of Equation (5.2.2) is presented in Equation (5.2.3) [27].

$$\left(\frac{E_b}{N_0}\right)_{dB} = 10\log\left(\frac{B}{R}\right) - 10\log(N-1) \qquad \textbf{(5.2.3)}$$

By convention, a bit-wise SNR of 8 dB achieves reasonable error rates. From the discussion in the last paragraph, under favorable operating conditions the data rate R assumes one of four possible rates: [3.6, 7.2, 14.4, 28.8] kbps[26]. Assuming a uniform distribution, the average gross symbol rate per channel is 13.5 kbps. The QPSK modulation scheme is employed by CDMA systems thereby reducing the gross data rate to a symbol rate of 6.75 ksps. Given the aforementioned assumptions, evaluation of Equation (5.2.3) yields a channel capacity of approximately 30 channels per 1.25 MHz band. Three of those channels are used for network overhead [24]. Therefore, the net serviceable capacity per cell is 27 per 1.25 MHz bandwidth. For comparison with the GSM standard, the capacity is scaled up to 25 MHz yielding 540 simultaneous users. Based on this simple analysis, CDMA systems offer a factor of 4 capacity advantage over GSM systems occupying the same bandwidth[27]!

As was the case for the GSM based networks, the demand for data intensive applications greatly impacted the evolution trajectory of CDMA networks. Table 5.2-1 captures the evolution of Qualcomm's dominant CDMA standards for forward link communication[28] [29]. From the table, like GSM, the CDMA standards remained true to the installed bandwidth: 1.25 MHz. To address the primacy of data in the presence of simultaneous user traffic and strict bandwidth constraints, packet switching in the form of time division multiplexing was introduced into the CDMA 2000 network architecture. Also evident from the table, the quantum leaps in data rate were enabled by new modulation

[26] The data rates are determined by the legacy voice coding options scaled by ½ to realize 1:2 convolutional coding: $(1200, 2400, 4800, 9600) \cdot 3 \cdot \frac{1}{2}$

[27] Frequency hopping applied to the GSM base standard renders network capacity comparable.

[28] Forward link communication refers to the data origination from base stations and termination at the mobile station. Market demand for data rates in the forward direction is deemed to be more relevant than reverse link communication. Clearly this will change as the market for Web 2.0 applications for handsets matures.

schemes and bandwidth expansion. In the jargon of this thesis, improvement on the data rate bottleneck was provided by focusing on the modulation scheme design variable and relaxing the bandwidth constraint. Also noteworthy is the change in channel coding algorithms that enabled the migration to smaller $\frac{B}{R}$[29].

	Pure CDMA		TDM CDMA		
	CDMA ONE	CDMA2000-1x	CDMA2000-EVDO-R0	CDMA2000-EVDO-RA	CDMA2000-EVDO-RB
TX Band	1.25 MHz	1.25 MHz	1.25 MHz	1.25 MHz	1.25 MHz
RX Band	1.25 MHz	1.25 MHz	1.25 MHz	1.25 MHz	1.25 MHz
Channel Bandwidth	1.25 MHz	1.25 MHz	1.25 MHz	1.25 MHz	3x1.25 MHz
Chip Rate	1.228 MChps	1.228 MChps	1.228 MChps	1.228 MChps	3x1.228 MChps
Modulation Scheme	QPSK	QPSK	QPSK,8PSK,16 QAM	QPSK,8PSK,16 QAM	QPSK,8PSK,16 QAM,64 QAM
Coding	1:2	1:2	1:5, 1:3, 2:3	1:5, 1:3, 2:3, 5:6, 5:12	1:5, 1:3, 2:3, 5:6, 5:12
Baud Rate (ksps)	3.6-28.8	307	96-921.6	921.6	3x921.6
Bit Rate/Ch (kbps)	1.2-14.4	307	2458	3072	14745
Slot Time (ms)	N\A	N\A	1.67	1.67	1.67
Frame Time (ms)	20	20	26.7	26.7	26.7

Table 5.2-1: CDMA Evolution (Forward Link Only). The original CDMA ONE standard provided voice service that was well served by 14.4 kbps. CDMA 2000 introduced higher data rates and is backward compatible with CDMA ONE. CDMA2000 EVDO introduced packet switching (TDMA) and higher data rates enabled by higher order modulation schemes and more advanced coding algorithms. Revision A and B provided more of the same.

5.3 Network Impact on Handset Architecture

Section 4.11 briefly introduced standards bodies as agents acting on behalf of the competitive members of the value network. The last two sections demonstrated the value the added by standards bodies. In short, the standards bodies provide the design rules that enable the emergence of new markets including consumer markets, handset markets, and component markets. Standards also package otherwise incommunicable bottlenecks into something comprehensive that can, in turn, be used to establish credibility in their target market: the end user. For example, network operators promote standards, such as "3G", and peak data rates

[29] Note, from (6.2.3), the number of subscribers is proportional to B/R. Therefore a symbol rate equal to the chip rate would require significant gain from the channel coding algorithm.

to avoid confusing the lay user. With regard to bottlenecks, the standards bodies provide a blueprint, addressing both data rate and error rate, which specifies all of the network level design parameters: modulation scheme, baud rate, pulse shape, signal power, and channel coding algorithm.

In particular, Section 5.1 and 5.2 presented the salient features that define the physical layer performance of the dominant contemporary wireless networks. From the discussions therein, the fundamental difference in network architecture was identified as the methods by which the available spectrum is assigned to users. The TDMA networks, such as those complying with the GSM standard, assign exclusive channels to each user whereas CDMA networks assign unique codes to each user thereby enabling users to share the entire band. That withstanding, the design variables and constraints pertaining to both networks were shown to be exactly the same and corroborate the bottleneck analysis presented in Chapter 4. Namely, improved data rate and error rate bottlenecks were shown to be functions of the design variables: pulse shape, modulation scheme, baud rate, power level, and channel coding algorithm[30]. What's more, analysis of network capacity for each of the dominant standards illuminated reinforced the bandwidth constraint. Recall, this constraint was first identified in Figure 4.10-1.

In general, if standards like those introduced in the previous sections are endorsed and deployed by the network operators, then firms competing in the market for designs that operate on the waveforms specified by the standards bodies are left with two design variables: noise added and power consumption, that impact the bottlenecks associated with battery life and error rate. What's more, the error rate bottleneck is one that is lower

[30] Pulse shaping was omitted from the list as it is compulsory and does not differentiate wireless communication performance.

bounded by the standards. The next chapter focuses on the physical layer designs targeting battery life and error rate.

6 Physical Layer Architecture

Chapter 4 established the system level bottlenecks and design variables associated with wireless communication. Chapter 5 introduced the dominant wireless communication standards and in doing so reduced the set of performance bottlenecks and design variables available to the handset and supplier layers of the value network to error rate and battery life. This chapter presents the design architectures employed by both layers to address these bottlenecks. As one might expect, given the discussion in Section 3.4, *new sets of performance bottlenecks emerge as artifacts are designed to address the system level bottlenecks*. The intent of this chapter is to identify and classify these new bottlenecks and to examine the trends in the evolution of handset architecture through the lens of the Baldwin and Clark management framework introduced in Chapter 3. Design structure matrices of the architectures presented in this chapter are included to aid the understanding of the interdependencies, degrees of modularity, and the size of the footprint of each layer in the designs.

Since an ancillary goal of the thesis is to better understand the migration of the digital domain boundary towards the antenna, the focus is placed on the analog sub-systems. A thorough analysis of complete physical layer designs would require focus on both the digital and analog functions, but is beyond the scope of this thesis. Instead, the thesis will develop the emergence of a bottleneck tree by considering various analog architectures that are core to the physical layer designs. Indeed, Section 6.4 will show that digital substitutes are only viable when they exceed the analog performance metrics developed in Section 6.1. Otherwise a change in architecture is required.

6.1 Common Challenges

The performance of physical components is frequency limited. For example, the gain of solid state amplifiers is inversely proportional to frequency. Since all analog and digital designs require significant gain from constituent transistor amplifiers to function, the designs typically fail well below the unity gain frequency determined by the transistor technology employed by each design. To address this and other frequency limitations, designs are chosen that convert the high frequency bandpass waveforms, used to communicate information over the wireless channel, to lower frequencies that are easily processed by contemporary micro-electronic components[31]. The target frequencies are determined by the frequency plan which is shaped by the wireless standard, component technologies, and noise sources.

The science of converting waveforms to different frequencies is called heterodyning. The first heterodyning radio architecture is attributed to Edwin Armstrong in 1918 [30]. The age of the design coupled with the pre-existence of integrated circuit technology suggests that much of the integration technology and engineering knowledge was available prior to the emergence of the wireless handset market. Handsets designed for digital communication systems also benefited from the pre-existence of legacy analog handset designs. Figure 6.1-1 illustrates one of the very first analog handset designs, the Motorola DynaTAC. The figure shows that even in the early stages of analog handset design, functionality is *split* and distributed to corresponding modules relevant to a heterodyning communication system. From left to right, the modulator performs the analog equivalent of the functions described in

[31] Recall from Chapter 4 that information is only dependant on bandwidth and is independent of carrier frequency.

Section 4.7. The frequency synthesizer generates a tone called the local oscillator (LO) that is employed by the mixer to map waveforms to new frequencies. The low noise amplifier (LNA) is used to boost the received waveform and the power amplifier is used to boost the transmitted waveform. The duplex filter separates the transmit and receive bands in a full duplex communication system. Lastly, the demodulator performs the analog equivalent functions of Section 4.8.

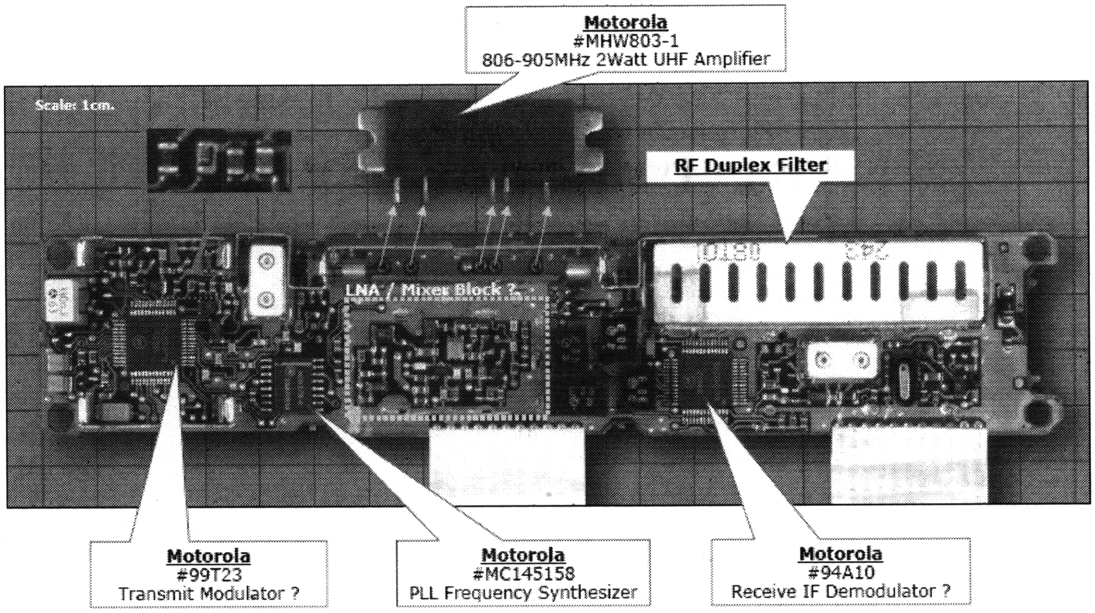


Figure 6.1-1: Motorola DynaTAC Radio Board [31].

Close examination of Figure 6.1-1 leads to one very important question that relates to modular frameworks. How does one define a module in the context of circuit designs? To the extent that a module refers to an integrated and packaged component that accomplishes a task or group of tasks, then the DynaTAC is already modularized and should enjoy all of the benefits captured by the equations from Chapter 3. Equations (3.2.9) and (3.2.10) are repeated here for convenience.

$$S = S_0 + \sum_{s=1}^{j}\left(NOV_s - c_{MFG,s}\right) - c_{YIELD}(j) \qquad (3.2.9)$$

$$NOV_s = \max\left(\sigma_s \sqrt{n_s} Q(k_s)\right) - c_{k,s}(n_s)k_s - Z_s \qquad (3.2.10)$$

Clearly, the DynaTAC design includes scores of small modules that support the larger tasks and modules described in the last paragraph. The green box magnifies an area on the printed circuit board containing four rigidly defined modules. What's more, the mixer and LNA designs are comprised exclusively of these smaller discrete modules. As a consequence the n_s term in (3.1.10) is large. The problem is that these support modules serve a role that is many layers below the bottlenecks targeted by the designer of the system; and they typically possess very little, if any, technical uncertainty[32]. Subsequently, the S_0 is not affected by the presence of these modules and the standard deviation, σ_s, is effectively zero. Then, due to the cost terms in (3.2.9) and (3.2.10), the overall system value owing to the presence of these modules is grossly negative.

The result of the economic pressure deriving from high cost of ownership of the super-modular design stimulated the evolution of integrated circuit technology enabling the integration of the supporting modules into the meta-modules that they support. Figure 6.1-2 illustrates the modular architecture that will serve as the baseline for the remainder of this section. In addition to identifying the value-related tasks and modules, the figure also delineates the divide between the digital and analog implementation of contemporary modules. Today, the migration of the line from left to right is the dominant trend in the evolution of physical layer designs. Indeed this migration is reshaping the supplier layer as

[32] On the contrary, the physical uncertainty of the parameters is often times very high – contributing to high yield costs.

analog suppliers are being pushed farther towards the antenna while the digital IC manufacturers fill the void.

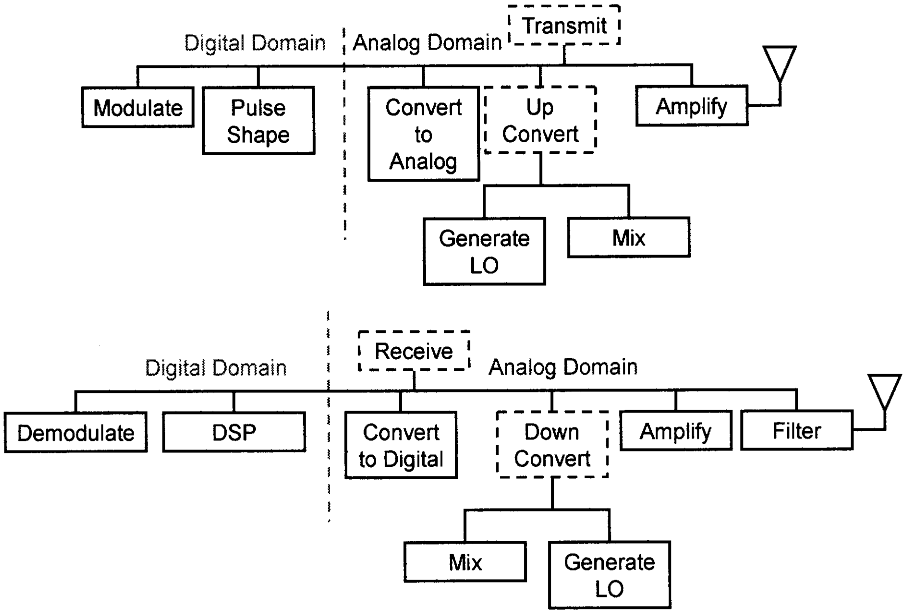


Figure 6.1-2: Baseline Transmitter (top) and Receiver (bottom) Architectures. The split functionality and the assignment of tasks to analog and digital modules is a key take-away from this figure. The primary role of the analog portion of the design is to translate the received waveform to a lower frequency suitable for digital conversion and subsequent digital signal processing. On the transmit side, the analog sub-system up converts the quadrature waveform to the proper carrier, provides power gain, and transmits the waveform.

Chapter 2 asserted the existence of hard modular boundaries at the integrated circuit level that partition the modules into those that perform either analog or digital tasks. Very briefly, high gain and output impedance made certain transistor technologies more suitable for analog applications, whereas lower performing transistors, but superior economics, dominated the digital applications in the design of the physical layer. Notice from Figure 6.1-2 that all of essential communication functions introduced in Chapter 4 are executed in the digital domain. The reason is simple, digital systems enjoy the benefits of Moore's law and they outperform and outlast functionally-equivalent analog systems. Note, this statement makes no mention of the architectural differences between analog and digital designs –

which are significant. Notwithstanding, the function of the analog portion of the receiver is purely to convert the communicable waveforms to frequencies suitable for digital conversion and subsequent digital signal processing. The analog portion of the transmitter is reserved for up-conversion and amplification of the transmitted waveform.

The fundamental task associated accomplished by heterodyning designs is frequency translation of the complex envelope from an RF carrier to an intermediate frequency. An intermediate frequency (IF) is defined as a frequency that serves the design of a radio by relaxing the component level frequency constraints[33]. All frequencies between the carrier and DC are potential intermediate frequencies. The IF is generated by multiplying, or “mixing”, a tone called the local oscillator (LO) with a target waveform. The multiplication re-centers the target waveform about *both* the sum *and* difference frequencies of the RF and the LO [32]. The inability of the mixer to distinguish the sum from the difference limits the usefulness of the mixer in a crowded frequency band. In particular, if a tone or band is present IF away from the LO and on the opposite side of the LO as the intended RF, be it sum or difference, that tone or band is converted to the IF along with the targeted RF. This undesirable tone or band is referred to as the image (IM) of the target RF. Figure 6.1-3 illustrates the mixing process for band conversions.

[33] The term radio refers to the transmitter and receiver designs the are responsible for frequency translation.

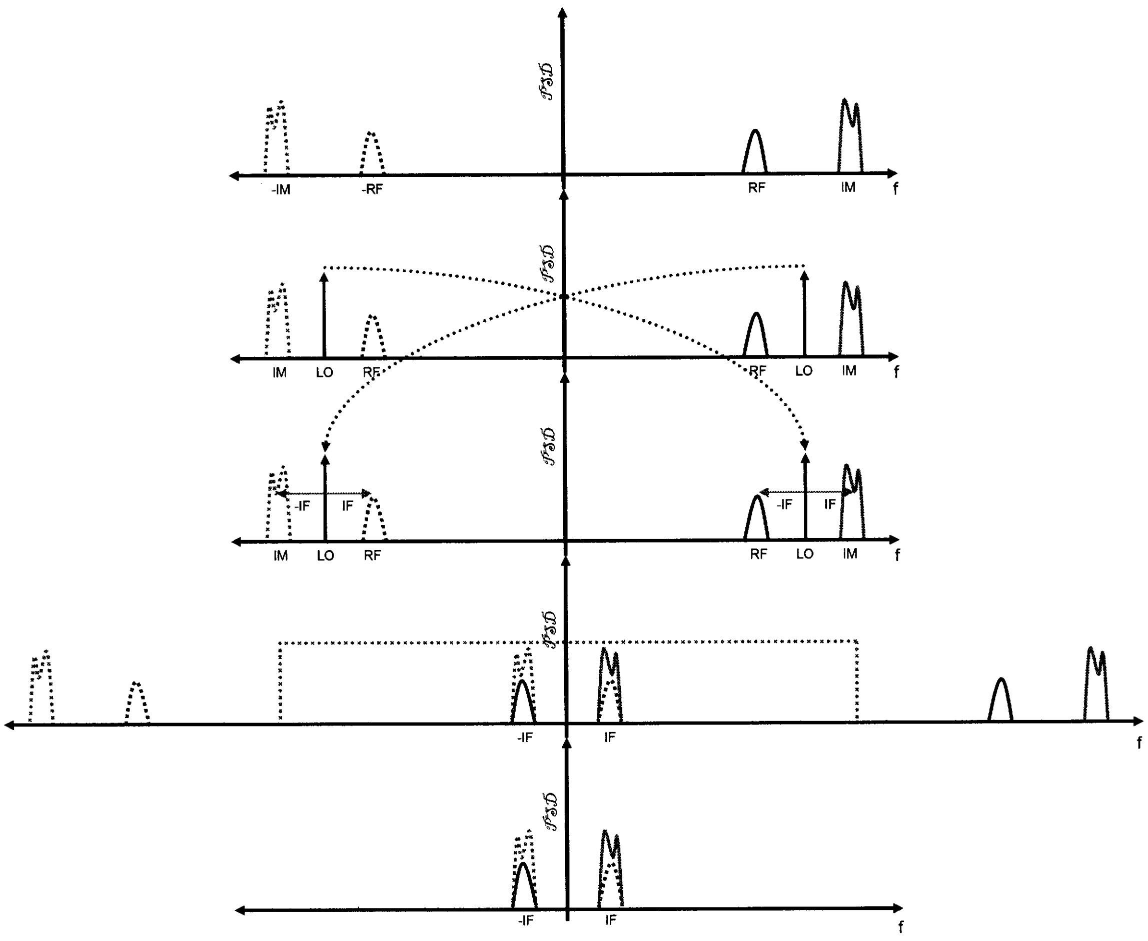


Figure 6.1-3: Mixing Process. From top to bottom: bandpass waveform and adjacent blocker, introduction of a LO, mirroring the LO tone through the y-axis (consequence of convolution operator), application of the convolution operator in frequency domain, low pass filtering. From the figure, the multiplication in the time domain is a convolution in the frequency domain. The convolution re-centers the RF quadrature waveform about the IF but does not discriminate between the targeted waveform and the image on the negative side of the y-axis. Both waveforms are superimposed in the IF band.

Clearly, if the image is eliminated, then only the RF is converted to the IF and the mixing operation will not contribute noise to the SNR. If the image is not completely eliminated, then it will certainly degrade the SNR. The simple expression for mixer induced SNR is given in Equation (6.1.1).

$$\frac{S}{N}(dB) = RF(dBm) - IM(dBm) \qquad \textbf{(6.1.1)}$$

Because all designs employ some form of down conversion, the standards bodies specify power levels of out-of-band and in-band potential images, called blockers. Knowledge of the blocker frequencies and power levels enables designers to manage the adverse effects of mixing that are illustrated in Figure 6.1-3 and captured by (6.1.1). Refer back to Figure 5.1-2 for the GSM blocker profile. This profile will be used to demonstrate performance through out the remainder of this chapter. Note, many authors and engineers distinguish noise from interference and subsequently introduce the term SIR. However, because interference and random noise lead to the same SNR impairments, the distinction is NOT made in this thesis.

The target IF and the method by which image frequencies are managed determines the architecture of both the receiver and transmitter. Legacy architectures employ filters to remove IM frequencies prior to mixing. The presence of these filters is the hallmark of the super-heterodyne architecture. On the contrary, contemporary image reject architectures employ cancellation techniques to eliminate filters from the architecture and subsequently reduce cost. These two architectures are compared by evaluating the IF purity when subjected to identical blocker conditions. Both architectures are introduced in Section 6.2.

Figure 6.1-3 illustrated the fundamental challenge of mixer based frequency conversion, given a spectrally pure local oscillator. Unfortunately, practical local oscillators include parasitic frequency "skirts" on either side of the desired tone that also degrade the SNR. Figure 6.1-4 illustrates the mechanism by which the SNR is degraded by imperfect oscillators through the process called reciprocal mixing [35]. From the figure, power in the frequency skirts mix with close-in blockers to corrupt the desired signal. The frequency content about the desired LO is called phase noise and is typically specified in dB below the desired signal (dBc) at a particular Δf from the desired LO. Equation (6.1.2) captures the SNR determined by reciprocal mixing.

$$\frac{S}{N}(dB) = RF(dBm) - (PN + BL + BW)(dBm) \qquad \textbf{(6.1.2)}$$

From the figure, reciprocal mixing degrades the SNR in a manner that is similar and indistinguishable from the degradation due to image conversion. In a modular designs, frequency synthesizers responsible for generating the LO are typically compared in terms of phase noise.

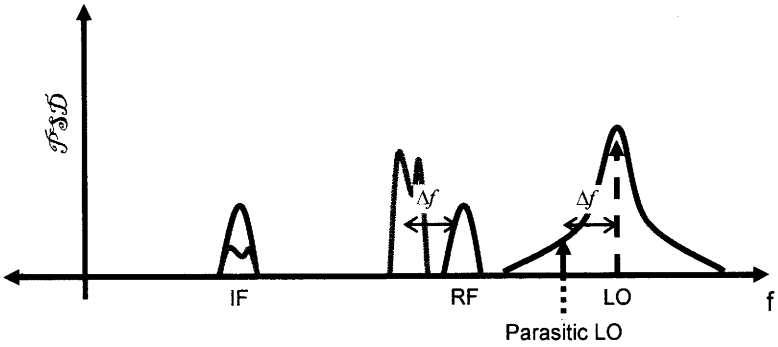


Figure 6.1-4: Reciprocal Mixing Due to Phase Noise. The imperfect LO tone mixes with blocker waveforms leading to IF contamination.

Image conversion to the IF is not the only interference-based parasitic effect that must be managed by designers. The other form of systematic SNR degradation derives from the non-linear qualities of the gain components, such as the amplifiers and mixers, in a design. The process by which nonlinearity degrades SNR is called intermodulation distortion (IMD). IMD is a result of parasitic self-mixing of the various frequency components present in a given band [32]. The degree to which the IMD products contribute to the SNR of a waveform is determined by both the power of the relevant waveforms and the strength of the non-linearities in each of the components in a design. Unfortunately, unlike image related noise where filtering or cancellation can mitigate the effects, there is no easy fix for IMD because the source of the noise is inherent in the components themselves.

The lion's share of IMD is generated by transistors employed by amplifier and mixer designs[34]. Furthermore, not all IMD products generate parasitic effects in the frequency bands of interest. In general, IMD effects are described relative to the terms in the Taylor Series model of a particular component's relationship between input and output waveforms. Stated simply, relevant IMD products, or the relevant Taylor Series terms, are determined by the frequency plan of a particular design. For example, in a non-zero IF based architecture the odd order IMD product is critical because it generates interference in the IF band of the received waveform and therefore contributes directly to SNR degradation [35]. Architectures that do away with the IF and convert directly to baseband are affected by the even order IMD product [35]. Box 6.1-1 describes an abstraction called an intercept point *(IPx)* that is used to approximate amplifier IMD[35]. The box focuses on the SNR due to the 3rd order IMD product. The take-away from the box is captured by equation (6.1.3)[36].

$$\left.\frac{S}{N}\right|_{OUT}(dB) = RF_{IN}(dBm) - 3BL_{IN}(dBm) + 2IP3(dBm) \qquad \textbf{(6.1.3)}$$

From the equation and the box, the SNR due to 3rd order IMD is proportional to the blocker *(BL)* power level and inversely proportional to IP3 and the targeted RF waveform *(RF)*. Note, owing to the non-zero bandwidth of real waveforms, they too behave as blockers unto themselves and degrade SNR even in the absence of external blockers. This property establishes an upper-bound on the power that can be extracted from a component without

[34] Electro-mechanical filters are also known to generate low levels of IMD.

[35] The notation IPx refers to the intermodulation product targeted by the analysis. Figure 7.1-3 illustrates the effect of 3rd order IMD and subsequently establishes the third order intercept point (IP3).

[36] The expression is derived from: $\left.\frac{S}{N}\right|_{OUT} = RF_{OUT} - (OP3 - 3(IP3 - BL_{IN}))$ which is derived by inspection from Figure 7.1-3.

destroying the SNR. Similar figures and expressions exist for other order IMD products (2nd, 4th, 5th, etc).

Box 6.1-1. Figure 7.1-6 illustrates a graphical tool that is often used to specify IMD performance. The figure can be unraveled as follows. The horizontal and vertical axes are log scale representations of the input power and output power, respectively. The black curve illustrates the measured output power of an amplifier as a function of input power. Notice, the amplifier is linear at low power levels and saturates at higher levels generating the IMD products. The figure also depicts a model of the amplifier performance consisting of the linear power performance as well as the third order IMD product as a function of input power. The intersection of the two lines is captured on the x axis by the input referred 3rd order intercept point (IP3) and on the y-axis as the output referred third order intercept point (OP3). In words, the IP3 is the input power level by which the third order IMD is exactly equal to the fundamental output power level. Similarly, the OP3 is determined by the output levels wherein this intersection occurs. Per the Taylor series expansion assumed by the model, the slope of the 3rd order IMD line is exactly three times that of fundamental power level.

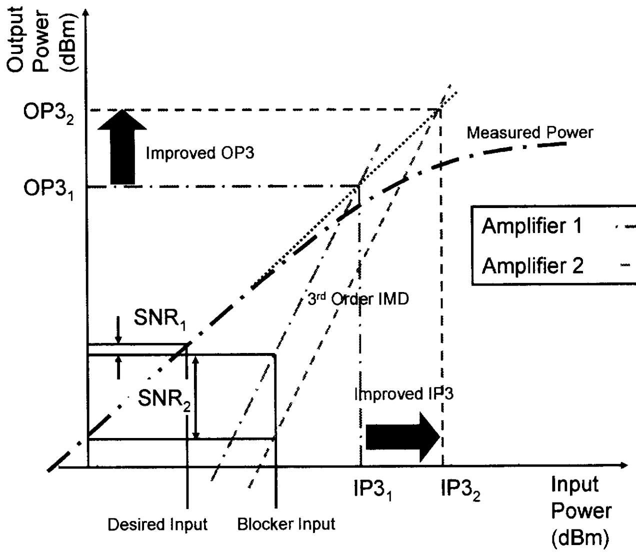


Figure 6.1-5: 3rd Order IMD and 3rd Order Intercept Point (IP3).

The previous discussion described IMD products originating from single components and provided a closed form approximation to SNR predicated on IP3 data. However, from Figure 6-1.1 and 6.1-2, a design is comprised of many non-linear components, each capable of generating IMD. Figure 6.1-6 provides a simplified generic representation of a cascaded system. Each shape represents a different component (amplifier, mixer, filter, etc).

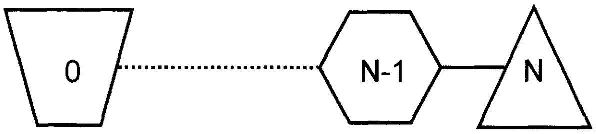


Figure 6.1-6: Generic Cascaded Design. Each stage possesses a gain and IP3\OP3. IMD of the cascaded design is determined recursively from Equations (7.1.2) and (7.1.3).

Intuitively, one would expect that the effects of earlier stages would be compounded by the power gain in later stages. Equations (6.1.4) and (6.1.5) provide a set of recursive expressions for IP3 and OP3 that validates intuition [33] [37].

$$IP3(dBm) = \frac{IP3_N IP3_{N-1}}{IP3_N + G_{N-1} IP3_{N-1}} \qquad \textbf{(6.1.4)}$$

$$OP3(dBm) = G_N G_{N-1} IP3_{CAS} = \frac{G_N OP3_N OP3_{N-1}}{OP3_N + G_N OP3_{N-1}} \qquad \textbf{(6.1.5)}$$

From the expressions, the net effect of cascaded systems lowers the overall IP3 of a design. For example, consider a design comprised of an amplifier followed by a mixer. The conversion gain of the mixer is assumed to be unity and the amplifier's power gain is set to 100. In this case, assuming the IP3's of the two components are comparable, the IP3 of the mixer will dominate the performance of the system. These tradeoffs permeate all the way through the design.

The previous noise sources: image noise, phase noise, and IMD derived from parasitic mixing products of one form or another. The SNR is also degraded by noise intrinsic to the components themselves. The intrinsic noise sources present in integrated

[37] Note, each of the terms in the expression is given in linear power units (watts). Application of (7.1.1) first requires translation to log scale (dB\dBm).

circuit technologies include frequency dependant flicker noise, white thermal noise, and white shot noise [32,34]. Flicker noise is attributed to electron traps found at the interface of semiconductor materials. Flicker noise is unique in that it follows a $1/f$ trajectory from DC to the noise floor. Thermal noise is attributed to thermal fluctuations in resistors. Finally, shot noise is attributed to random movement of electrons across semiconductor junctions. For the purpose of this report, it is important to note that each of these noise sources, save thermal noise, is technology dependant. Firms seeking to improve noise performance might well choose one technology over another exclusively for its low noise qualities.

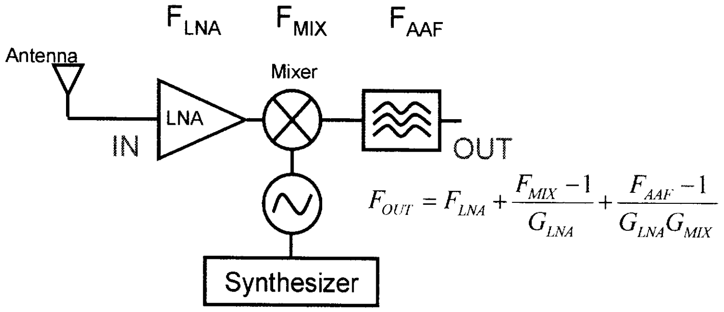
Intrinsic noise is generated by each analog component present in the receive and transmit chains, respectively. Subsequently, each component degrades the SNR of the waveform as it propagates through the system. Equation (6.1.6) introduces the noise figure, *NF*, which defines the SNR transfer function due to noise corruption.

$$\left.\frac{S}{N}\right|_{OUT}(dB) = \left.\frac{S}{N}\right|_{IN}(dB) - NF(dB) \qquad \textbf{(6.1.6)}$$

Unfortunately the noise figure is lower bounded by zero which simply means that the affect of upstream noise cannot be eliminated by downstream waveform processing [32]. Strictly modular designs, wherein the transmit and receive chains consist of well defined cascaded components, employ the Friis Noise Equation to determine noise performance of the entire chain [35]. Note, the noise figure is simply the dB equivalent of the noise factor, F.

$$F_n = F_1 + \frac{F_2 - 1}{G_1} + \ldots + \frac{F_n - 1}{\prod_{i=1}^{n} G_i} \qquad \textbf{(6.1.7)}$$

Figure 6.1-7 illustrates the usefulness of the Friis Equation by modeling the noise factor of a primitive receiver comprised of a LNA, a mixer, a synthesizer, and a filter. It is assumed that the synthesizer does not contribute any intrinsic noise to the waveform. Moreover, close examination of Equation (6.1.7) and Figure 6.1-7 reveals the relative importance of noise contributions in the early stages of waveform processing chain. In particular, the noise performance of later stages is discounted proportionally to the gain of the previous stages. As a consequence, the noise performance of the first few stages immediately following the antenna in a receiver design contribute the most to the NF.



$$F_{OUT} = F_{LNA} + \frac{F_{MIX} - 1}{G_{LNA}} + \frac{F_{AAF} - 1}{G_{LNA} G_{MIX}}$$

Figure 6.1-7: Demonstration of Friis Equation. The noise contribution in later stages is discounted by the gain in previous stages extending primacy to the noise figure of the first few stages in a design.

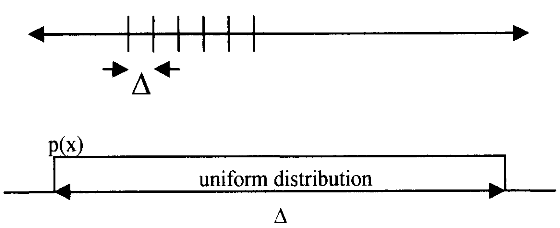
The final noise source impacting the generic design illustrated in Figure 6.1-2 derives from the quantization noise introduced by the analog-digital\digital-analog converters (ADC\DAC). The SNR of a conventional ADC is given in equation (6.1.8) and derived in Box 6.1-1 [36].

$$\frac{S}{N}(dB) = R(dB) + 1.76(dB) + 6.02b(dB) \qquad \textbf{(6.1.8)}$$

The term R is the ratio of the input signal to full scale range of the ADC, the 1.76dB term originates from a uniformly distributed noise probability, and the term b refers to the number of bits in the ADC\DAC. From Equation (6.1.8), fractional R can only degrade the SNR; and for this reason the full scale range of the ADC is carefully matched to the analog waveform being sampled. For example if the full scale range of the ADC is 1V RMS and the incoming waveform is 1uV RMS, then for a 9dB SNR, the ADC would require 21 bits! This is precisely the reason that gain is needed in a receiver design. In general, the term dynamic range of the ADC is defined as the power levels of the sampled waveform by which the SNR is comfortably above the system noise requirements. Subsequently, the dynamic range of the ADC is designed with a-priori knowledge of the possible waveform power levels and various gain stages in a receiver.

Box 6.1-1. The signal to noise ratio of an ADC is determined by the precision penalty of not being able to exactly represent analog waveforms by digital words. Clearly, as the number of bits in the representation grows, so does the precision. However, precision via more bits comes at the expense of added power consumption, or in other words, as a battery life penalty. Ideally, ADCs are designed to operate below the noise floor of the system determined by the other analog components in the receive chain.

The error is ultimately determined by quantization error calculated as the mean square error (MSE). Dividing the signal level, S, by the noise yields the often encountered SNR expression derived in this box.



b=number of quantization bits
V=full range of possible analog input signal
s=percentage of full scale range of input signal

$$E_{RMS} = \sqrt{\int_{-\Delta/2}^{\Delta/2} prob\left(|x| \le \frac{\Delta}{2}\right) \cdot x^2 dx} = \frac{\Delta}{\sqrt{12}}$$

$$V = 2^{b-1}\Delta$$

$$S_{RMS} = \frac{sV}{\sqrt{2}}$$

$$\frac{S}{N} = \frac{S_{RMS}}{E_{RMS}} = \frac{s2^{b-1}\Delta}{\sqrt{2}} \frac{\sqrt{12}}{\Delta}$$

$$\left(\frac{S}{N}\right)_{dB} = 20\log_{10} s + 20\log_{10}\frac{\sqrt{6}}{2} + 20\log_{10} 2^b$$
$$= R + 1.76 + 6.02b$$

Section 5.3 hypothesized that two performance bottlenecks are available to handset and component designers: error rate and battery life. This section presented the ubiquitous noise sources that originate in the analog portion of the physical layer design and that impact error rate via the SNR. Although tacit knowledge of the noise contributions is beneficial to component designers; observed in isolation, they don't provide the information required by firms to formulate strategies. To that end, Box 6.1-2 aggregates all of the noise sources introduced in this section and generates a figure of merit that can be used to classify them using the bottleneck framework introduced in Chapter 3. Equation (6.1.9) restates the figure of merit derived in Box 6.1-2. In words, Equation (6.1.9) collapses the SNR of the entire analog portion of the generic communication system in Figure 6.1-2 into two functions, one determined by the radio performance and the other by the ADC.

$$\frac{s}{n}=\frac{1}{\max\left(\frac{im}{rf}+\frac{pn\cdot bl\cdot bw}{rf}+\frac{bl^3+rf^3}{rf\cdot ip3^2}+\frac{n_{FLOOR}}{rf}F,\frac{\sqrt{6}}{2}\cdot s\cdot 2^b\right)} \tag{6.1.9}$$

From the equation, the SNR and subsequently the data rates are proportional to the exogenous variable: RF power – but only to an extent. Once the self induced IMD begins to dominate as a result of too much RF power, then the data rate degrades with the square of the input power. Furthermore, Equation (6.1.9) shows the performance of designs of this nature to be proportional to IP3 and inversely proportional to image power, oscillator phase noise, blocker power level, and noise factor. That being said, each of the noise terms contributes to system performance in a relative sense. On the contrary, the term to the right of the comma in (6.1.9) represents the noise introduced by the ADC\DAC quantization errors. Although quantization noise could in theory be added to the sum on the left of the comma, it only makes sense to do so when the ADC\DAC contributes noise to the sum – which is deemed unacceptable in practice. Finally, comparison of Equations (3.1.1) and (3.1.2) with (6.1.9), establishes the absolute bottlenecks in the analog portion of the heterodyne communication systems, radio noise and quantization error. What's more, the radio noise is comprised of four relative bottlenecks pertaining to image rejection, phase noise, intermodulation distortion, and intrinsic noise figure.

Box 6.1-2. Earlier in this chapter, the equations for SNR were presented in log scale (dB) units to conform to industry convention. To combine the expressions, they must first be converted to SI units. The original expressions and the SI equivalents are presented in this box. Note, lower case letters were used to denote different units.

To evaluate system level performance, the SI expressions are factored for the term N$_{OUT}$ then combined in the snr performance ratio. Note, the ratio includes the self induced IMD.

$$RF - IM = \frac{S}{N_{OUT}} \Rightarrow \frac{rf}{im}$$

$$RF - \left(PN + BL + BW\right) = \frac{S}{N_{OUT}} \Rightarrow \frac{rf}{pn \cdot bl \cdot bw}$$

$$RF - 3BL + 2IP3 = \frac{S}{N_{OUT}} \Rightarrow \frac{rf \cdot ip3^2}{bl^3}$$

$$\frac{RF}{N_{IN}} - NF = \frac{S}{N_{OUT}} \Rightarrow \frac{rf}{n_{IN}}\frac{1}{F}$$

$$R + 1.76 + 6.02b = \frac{S}{N_{OUT}} \Rightarrow s \cdot \frac{\sqrt{6}}{2} \cdot 2^b$$

$$\frac{s}{n} = \frac{1}{\max\left(\frac{im}{rf} + \frac{pn \cdot bl \cdot bw}{rf} + \frac{bl^3 + rf^3}{rf \cdot ip3^2} + \frac{n_{FLOOR}}{rf}F, \frac{\sqrt{6}}{2} \cdot s \cdot 2^b\right)}$$

The remainder of this section will target the battery life bottleneck and then summarize the results. Contemporary handsets support diverse functionality that place a heavy burden on battery technology. For example, in addition to legacy cellular connectivity, the Nokia N95 supports music, FM radio, video, Bluetooth and WiFi connectivity, GPS, photography, and gaming [37]. Operation of each ancillary feature draws significant power from the battery and subsequently comprises a system of relative bottlenecks. Notwithstanding, this thesis is only concerned with the performance bottlenecks owing to legacy wireless communication.

Figure 6.1-8 illustrates power consumption of the Nokia N95 during GSM voice communication. The mean power consumed in region 3, where communication takes place, is 710 mW. To isolate the power consumed by the communication process, the device dependant background power must be de-embedded from the power observed in the figure.

Figure 6.1-9 illustrates standby power consumption in the offline mode. From the teardown report, the data in Figure 6.1-8 corresponds to region 5 in Figure 6.1-9 [37]. Moreover, the ambient power consumption in this region is approximately 250 mW. Therefore, the power being consumed by the communication function is approximately the difference, 460 mW.

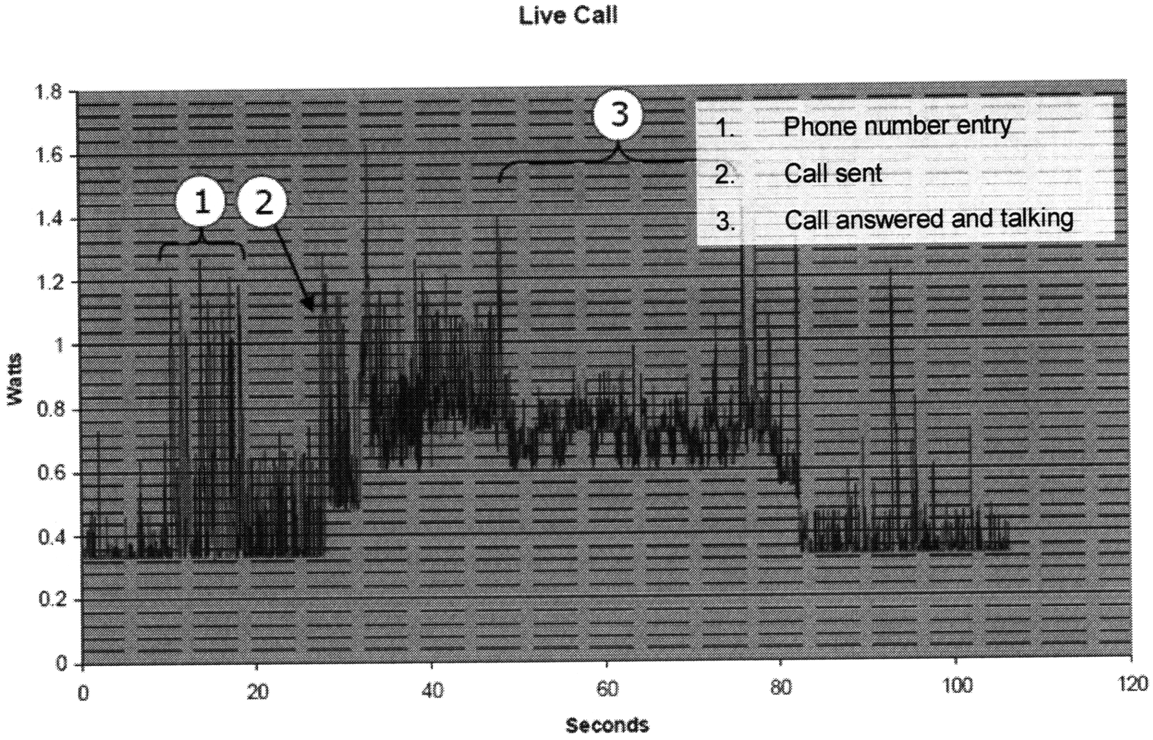


Figure 6.1-8: Nokia N95 Power GSM Power Consumption (maximum backlight intensity). [37]

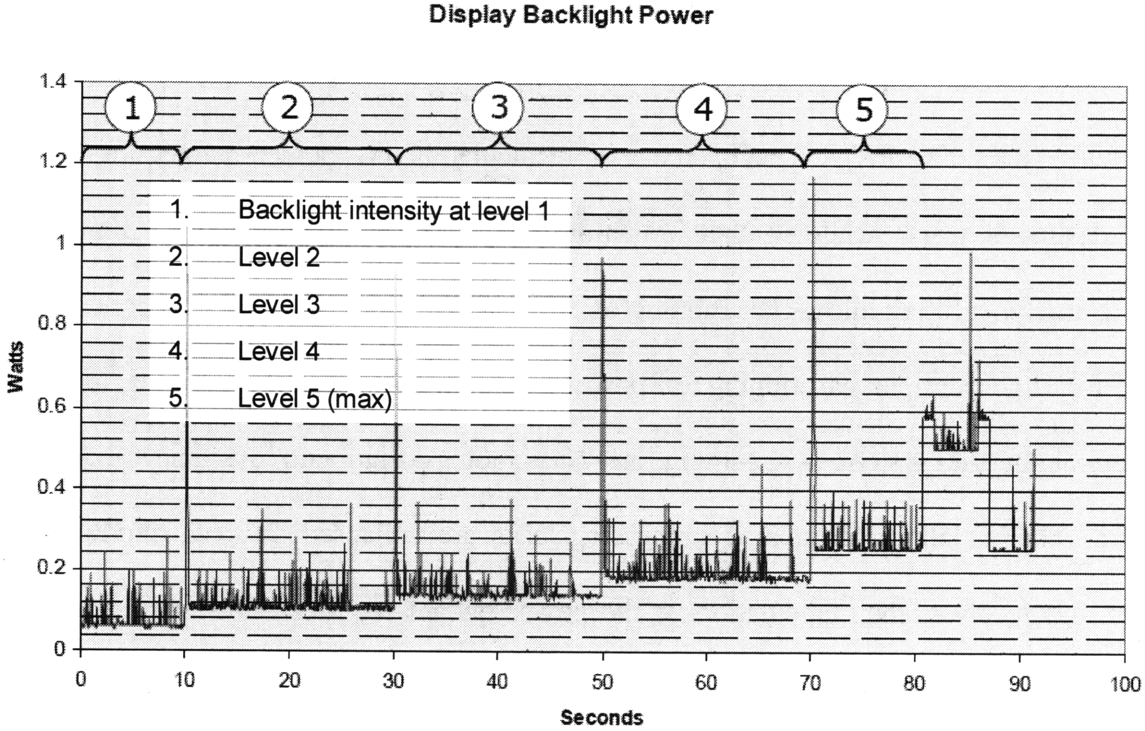


Figure 6.1-9: Nokia N95 Power Offline Power Consumption. [37]

Dividing the 460mW between transmit and receive functions requires knowledge of the communication protocol. In GSM systems, the transmission and reception are mutually exclusive events. What's more, owing to the TDMA qualities of the network, each user is granted access to the frequency channel for a predetermined fraction of the communication frame. From Table 5.1-1, the user has access to a particular channel during one of the eight network specified time slots in a frame. Therefore, the receiver is on for 1/8th of the time and the transmitter is on for 1/8th of the time. If the GSM standard is used as a benchmark for RF output power, then approximately 2 watts of output power is required of the transmitter. If transmitter efficiency is approximated by 60%, per power amplifier supplier specifications, then the transmitter sinks 420 mW of power leaving only 40mW of power for the receive function. Subsequently, the power budget assumes the form: $\{RX, TX\} = \{9\%, 91\%\}$.

Per the baseline architecture of Figure 6.1-2, the power amplifier is *split* from the remainder of the architecture. Therefore, the power budget can be divided among at least three modules (RX, TX, PA). The exercise of recalculating the power budget for the expanded design is worth while in that it demonstrates the importance of the *splitting* operator. Namely, splitting allows partitioning of bottlenecks among different components. However, for this example partitioning of power among three variables requires more information about the system. Assuming the N95 radio design employs image-reject architecture, wherein all of the analog processing tasks are collapsed into two modules, transmitter and receiver, then calculation of the power budget requires knowledge of the power consumption of these two modules. From professional literature, a typical image-reject radio design draws 250mW and 210 mW from the receiver and transmitter, respectively [38]. Equation (6.1.10) expresses GSM power consumption of the 3-module design.

$$\frac{(P_{TX}+P_{PA})}{8}+\frac{P_{RX}}{8}=P_{COM} \qquad \textbf{(6.1.10)}$$

Solving (6.1.10) for the power consumed by the power amplifier yields $P_{PA}=3.2\ W$. Is this reasonable? PA suppliers advertise power added efficiencies of 54% for 34dBm of output power [39][38]. Applying this specification to the previous result, 54% of 3.2 W of DC power translates to 1.72 W of RF power. Unfortunately, it's impossible to determine the exact PA power contribution to the RF output power during the experiment because the power varies with distance to the base station. However, given that the maximum nominal power from the

[38] Power added efficiency is defined as the efficiency for power added to the transmitter output. However, although obfuscated by the log scale, the bulk of the power consumption is due to high absolute output power (~>33dBm) and is largely independent of transmitter output power (~<10dBm).

GSM standards is 2W, this approximation seems very reasonable [23]. In the final analysis, the power budget is divided as follows: $\{RX, TX, PA\} = \left\{\frac{0.25}{3.66}, \frac{0.21}{3.66}, \frac{3.2}{3.66}\right\} = \{7\%, 6\%, 87\%\}$.

Clearly, the PA contributes overwhelmingly to the battery life performance bottleneck. In the context of the bottleneck framework from Chapter 3, every function that draws power must be considered a relative bottleneck. However, given the relative disparity, the PA could reasonably be considered an absolute bottleneck in the system performance for voice communication. That withstanding, handsets operating predominantly in receive mode would drift towards the RX bottleneck or perhaps equal partitioning. Applications such as web browsing in which users consume information from the network but don't necessarily contribute information would realize a receive-only scenario.

Assuming that handsets will always require 2-way voice communication to meet status-quo expectations of consumers, then maximizing battery life via the power amplifier module best addresses the quasi-absolute bottleneck in the system. From Equation (6.1.9), RF power is absolutely necessary to transmit information wirelessly; therefore, primacy must be placed on efficiency. The theoretical maximum efficiency of an amplifier is given as 78.5 % [32]. From the previous example, 54% PAE is a far cry from the maximum efficiency. The reason for the relative inefficiency is not lack of engineering knowledge but rather the error rate performance penalty realized by efficient PA operation. Figure 6.1-10 illustrates the performance tradeoffs of a published hetero-junction bipolar transistor (HBT) power amplifier [40]. Note, at PIN=24 dBm the PAE is very close to the theoretical maximum. However, the amplifier is also pushed 5dB into compression which means it is highly non-linear and subsequently generates prohibitively high levels of IMD. To see this, the trends from Figure 6.1-10 are plotted along with the salient features of Figure 6.1-6 in Figure 7.1-

11. In the latest figure, the efficiency and SNR are compared for two similar input power levels. From the figure, input power level B achieves better efficiency than input power A, but suffers a much lower SNR.

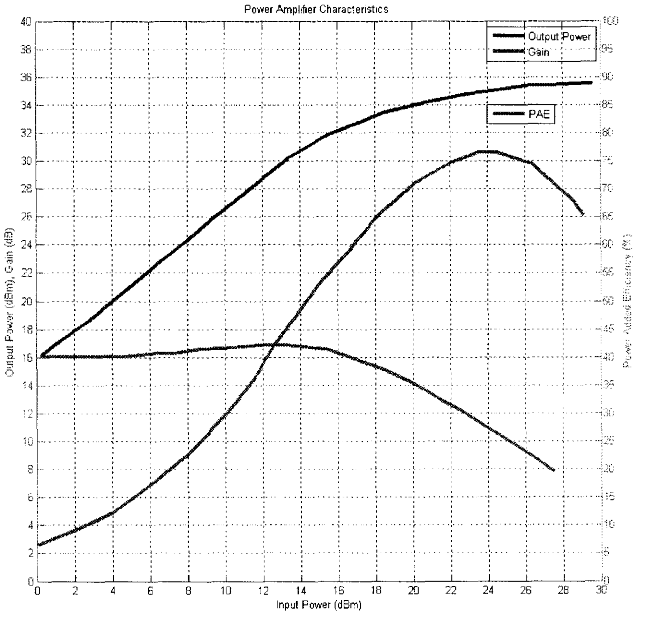


Figure 6.1-10: Power Amplifier Performance. The efficiency (PAE) peaks at a level where IMD is high. System designers must realize a compromise in order to address the error rate and battery life bottlenecks in handset designs.

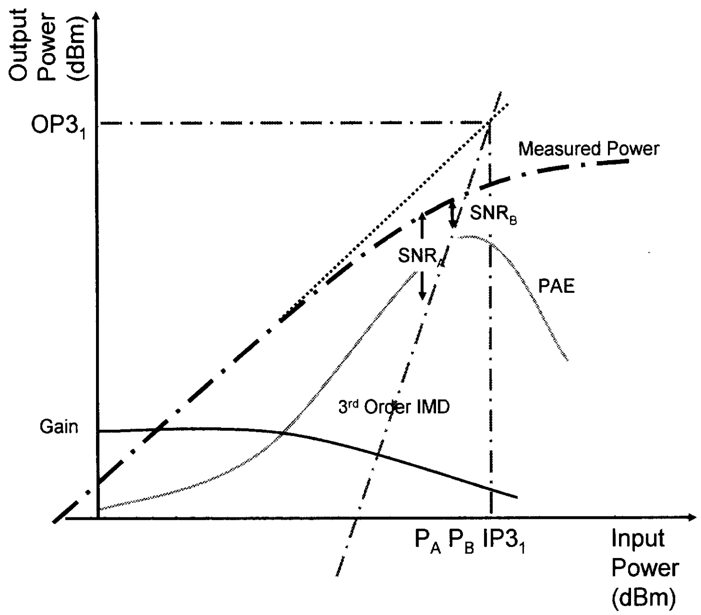


Figure 6.1-11: Efficiency vs. IMD Compromise. Notice, the SNR for the PA operating in backed-off operational modes exhibits a much higher SNR but comes at the expense of efficiency.

The previous discussion illustrates the significance of the splitting operator. In the last example, splitting the power amplifier task from the other communication tasks removed the battery life bottleneck from consideration, or at least relaxed its importance. In general, the net effect of splitting is that it frees firms and designers to identify particular tasks\modules that are linked with bottlenecks in the system. The remaining tasks\modules are subject to outsourcing if doing so enables a firm to reach a higher level of ROIC. In situations where the bottlenecks are not completely isolated in modules, the compromises are left to the discretion of the system designer and are communicated in the form of detailed specifications outlined in the design rules [1]. .

To conclude this section on common challenges impacting all analog radio designs, the newly exposed bottlenecks of a general heterodyning communication system will be assigned to modules. The new bottlenecks emergent upon addressing the error rate system-level bottleneck with a heterodyning design include image noise, phase noise, IMD, intrinsic noise, and quantization noise. Of those five, all but quantization noise fit into the relative bottleneck category. Likewise, examination of the electronic power consumption design variable exposed all transmit and receive components as contributors to the relative bottlenecks impacting the battery life system level absolute bottleneck. To that end, the power consumed from the transmission task, and in particular the PA component, was shown to dominate that of the receive task. The next section examines the dominant design architectures that target these bottlenecks.

	Transceiver Design Variables\Modules					
	Frequency Plan	DAC	Synthesizer	Mixer	Amplifier	Filter
Image Noise (IM)	X			X		X
Phase Noise (PN)			X			
IMD	X			X	X	X
Intrinsic Noise (F)				X	X	X
Quantization (Q)		X				
Electronic Power Consumption (P)		X	X	X	X	X

Table 6.1-1: Mapping of Bottlenecks to Modules.

6.2 Modular Superheterodyne Design

The legacy superheterodyne receiver architecture is illustrated by the block diagram in Figure 6.2-1. The corresponding signal flow is illustrated in Figure 6.2-2. Simultaneous examination of both figures provides a mapping of form to function. In as much, the waveform is first received by the antenna. The band select filter rejects most of the out-of-band blockers before passing the waveform to the low noise amplifier (LNA) for amplification. Upon leaving the LNA, the waveform is down-converted to a *high* intermediate frequency by a mixer which multiplies the received waveform with the first LO tone. The high IF provides the band-select filter with enough spectral separation to suppress the red image band that invariably gets down-converted to the IF. Next, the IF amplifier adds power to the down-converted waveform to compensate for filter losses. The channel select filter removes in-band blockers just before the second mixer and tunable LO down-convert the waveform to a low IF or baseband, depending on the frequency plan and the DSP. The filtered channel or channels are then digitized by the ADC which communicates the digital equivalent of the down-converted waveform to the digital signal processor for separation of the I and Q waveforms, filtering, error-reduction, signal processing, and finally demodulation.

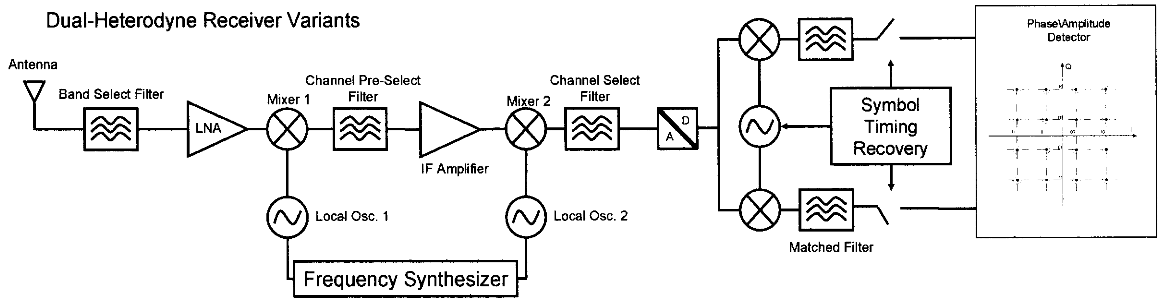


Figure 6.2-1: Legacy Superheterodyne Receiver Architecture.

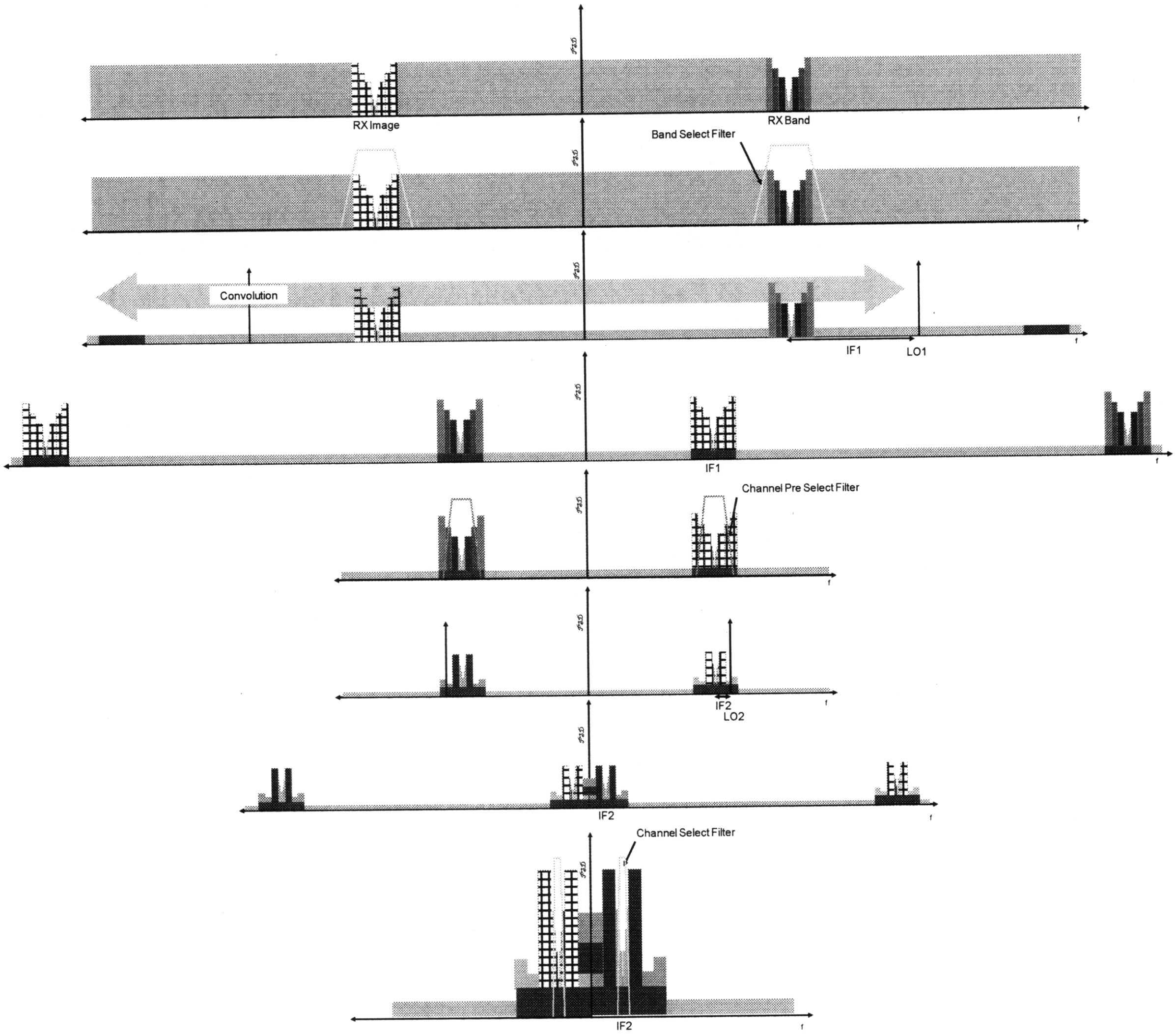
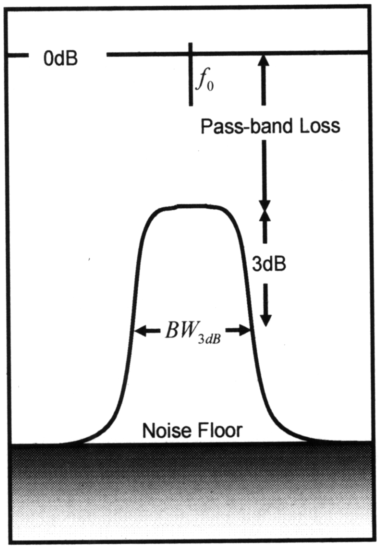


Figure 6.2-2: Superheterodyne Receive Signal Flow. The top row illustrates the frequency domain representation of a waveform embedded in a swath of frequency space. Note, per the discussion in Section 4.1, the image of the positive frequency band is mirrored about the DC axis. To eliminate the out-of-band interference, the received waveform is passed through a high Q filter. The third row shows the operands of the mixing operation; which is effectively a multiplication of the band of interest with a sinusoid. Note, a multiplication in the time domain maps over to a convolution in the frequency domain which requires the frequency domain representation of both operands. The frequency domain equivalent of a real sinusoid is simply two impulse functions centered at the positive and negative frequency of the tone. The convolution in the time domain was introduced in Equation (4.5.5) and functions no differently in the frequency domain. Basically, the multiplying tone is mirrored about the y-axis then swept across all frequency space with the waveform fixed. The product is evaluated at each point in continuous space. The result of the convolution (or mixing operation) is captured in the 4th row. The process is repeated in the last 4 rows of the figure. Note the final down-conversion is not shown in this figure.

Figure 6.2-2 identified a new bottleneck in the heterodyning architecture. Namely, limitations on filter selectivity determine the choice of intermediate frequencies in the double superheterodyne architecture. In plain words, the selectivity of a particular filter captures how well it removes unwanted waveforms outside the band of interest. A bandpass filter's selectivity performance is specified by the center frequency and width of the filter's 3dB pass band[39]. Figure 6.2-3 illustrates the salient features of a bandpass filter. What's not shown in the figure is that the 3dB pass band is proportional to the filter's center frequency and inversely proportional to a physical parameter called the quality factor (Q). See Equation (6.2.1).

$$BW_{3dB} \propto \frac{f_0}{Q} \tag{6.2.1}$$

Equation (6.2.1) means that in order to remove the effects of neighboring interferers from a waveform centered at an arbitrary frequency, filters must achieve arbitrarily high Q values or the targeted waveform must be translated to arbitrarily low frequencies. For example, a waveform spanning 200 kHz at 900 MHz would require a filter possessing Q=4500. Unfortunately, physical limitations of contemporary filters constrain Q to less than 1000.



[39] 3 dB frequency is simply the frequency wherein the

Figure 6.2-3: Filter Transfer Function. A filter is a component that removes spectral content outside the band of interest. It is defined by the 3dB bandwidth (BW_{3dB}) and the center frequency. From Equation (7.2.1), arbitrarily high selectivity filters at high frequencies are impossible to produce.

The aforementioned practical Q limitation imposes constraints on the superheterodyne design. Since the filter rejection properties improve proportionally with distance from the passband, maximum image suppression is achieved by spreading the RF and LO as far apart as possible. By identity, the IF is the difference between the LO and the RF. Therefore, maximum image suppression is realized by high IF designs. On the other hand, maximum selectivity is achieved by very narrow band filters. Therefore, assuming a reasonable Q value, narrow band filtering is made possible by converting the RF to a very low IF. This apparent paradox is relaxed by implementing the dual IF architecture of Figure (6.2-1) and (6.2-2) wherein two IF's are employed, the first for image rejection and the second for channel selection.

The degrees of modularity in this and subsequent design architectures can be compared using the design structure matrix (DSM). Figure 6.2-4 illustrates the DSM for the double superheterodyne design in Figure 6.2-1. A DSM is a square matrix of components or tasks that defines a particular architecture. A binary DSM lists all components along the axes of the matrix and places a binary true mark in each column where a dependency exists. For example, from Figure 7.2-4, the antenna design is dependant on the carrier frequency, and the receive band bandwidth, the system impedance, and the gain. A thorough DSM illustrates dependencies, sequences, and complexity of a particular architecture. Iterative dependencies appear as ones above the main diagonal and sequential design structure appears as ones below the main diagonal.

The DSM in Figure 6.2-4 illustrates the highly ordered architecture of a superheterodyne receiver. What's more, the component design block exhibits exceptional levels of modularity as can be seen by the absence of dependencies among the components. This does not happen by chance. Rather, specifications generated early by system designers are outlined to communicate critical parameters that are relevant to system performance, such as the bottlenecks determined in Section 6.1. It is exactly these specifications that enable the market for components as well as base line performance expectations set by the system designer.

The assumptions behind this and subsequent DSMs are as follows. Per Section 2.1, regulation sets the carrier frequency and bandwidth of communication waveforms. From Chapter 5, standards bodies determine all of the design variables associated with a particular waveform, save those related to noise added and power consumed by the designs themselves. As such, the entries in the red box comprise the system level design rules that determine minimum levels of acceptable performance and provide the fundamental constraints placed on ODMs and suppliers. The entries contained in the green box represent design variables unique to each design as determined by the signal flow graphs of Figure 6.2-2. These variables include: frequency plan, value related features of the modules in the block diagram, and limits on the performance bottlenecks exposed in this section. For completeness, the ADC and DSP entries in the DSM represent place-holders for the meta-modules that fill out the physical layer design. The entries captured by the yellow block identify with each component in the block diagram of Figure 6.2-1. Provided the system design parameters are complete, these entries in the DSM identify the market for substitute modules that derive from the modular architecture. Finally, the last four entries in the DSM illustrate the measured performance along the bottleneck dimensions and are the purview of the system integrators.

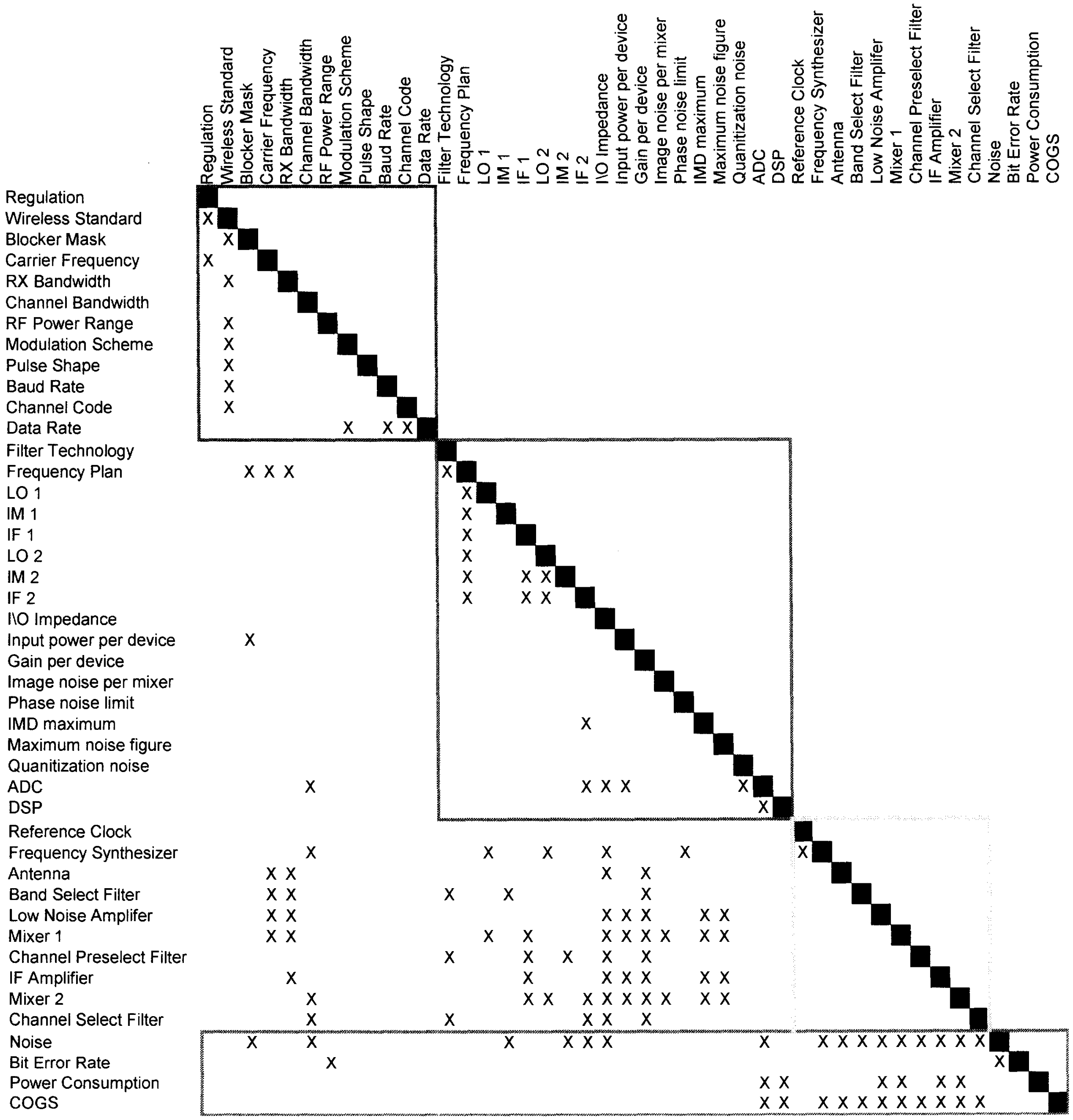


Figure 6.2-4: Double Superheterodyne DSM. The DSM for the superheterodyne design illustrates an architecture that is highly ordered and modular. The layers in the value network are captured by the colored blocks in the DSM. The network operators adopt standards that determine the design rules in the pink box in the upper left corner. The ODMs inherit the system specifications and determine the frequencies of operation and place constraints on the implementation level bottlenecks described in Section 6.1. Downstream, the component suppliers inherit the specs from the ODMs and compete for

"design wins" in the architecture. The last four rows in the DSM capture the option value that the ODMs reserve by filling the role of system integrator.

Further examination of the DSM reveals some very interesting qualities with regard to the legacy wireless communication value network. To start, the colored boxes map design decisions to different layers is the value network. The red box includes those design variables controlled by the network operators and standards bodies. The green box establishes the design parameters owned by the system designers. The yellow box belongs to the component suppliers. Finally, the blue box belongs to firms that fill the system integrator role and who are accountable for the performance of the physical layer designs along the bottleneck dimensions. Note, the bit error rate is explicitly identified in the DSM whereas the battery life bottleneck is implied by power consumption. Judging from the interdependencies, the system designers control most of the downstream design activity by specifying design rules for component modules. Meanwhile, the system integrators reap the benefits of the design rules specified by the system designers by realizing the *option value* of the complete design. In the legacy value network, the handset ODMs filled the roles of system designer and system integrator. Notice the absence of coupling in the yellow box. This simply means that component suppliers were not beholden to one another for their contribution to the system; thereby minimizing the risk due to schedule slippage and agency cost born by the system integrators. By minimizing the dependencies in a design, such as has been done in the modular superheterodyne design, clever system designers and integrators create a very competitive market for modules.

6.3 *Integrated Modular Image Reject Designs*

The demand for lower manufacturing cost by way of component integration led to the gradual improvement of specialized semiconductor processes enabling complete radio designs without image filters [41]. From Equations (3.2.9) and (3.2.10), the economics driving the demand for integrated modular designs derive from the manufacturing and yield cost associated with ownership of traditional modular designs as well as the low technical potential of the modules employed by legacy superheterodyne designs. The simple economics led to the revisiting of image-reject architectures invented by Hartley and Weaver in the 1920's and 1950's, respectively [42,43]. The primary benefit of these architectures, with regard to manufacturing cost, is that the image is rejected rather than filtered thus eliminating the need for IF filters. The intent of the image reject architecture is realized by complex circuit designs wherein the quality of the image rejection is largely determined by the precision of the IC manufacturing technology.

In addition to providing cost savings, migration from the superheterodyne architecture to image reject architecture increased complexity and technical potential. The manufacturing cost advantage, high technical potential, and well published design rules led to the emergence of a market for image reject radio designs and eventual obsolescence of modular superheterodyne architecture. From a performance bottleneck point of view, the image reject architecture did away with some of the filter selectivity bottlenecks but introduced two new bottlenecks into the design, phase control and transistor manufacturing uniformity.

Contemporary image reject architecture is predicated on complex signal analysis, which assumes only that a waveform is separable into real and imaginary components

[44,45]. Equations (4.3.3) and (4.3.4) guarantee the complex-separable quality of bandpass waveforms. From the last section, frequency translation was shown to be the product of waveform multiplication. Figure 6.3-1 illustrates the multiplication of two complex waveforms *x(t)* and *y(t)* yielding *z(t).* In the figure, per the condition for complex signal analysis, all three waveforms have been separated into real and imaginary components. If *x(t)* represents a physical waveform, such as one received by the antenna, then $x_i(t)$ is exactly zero[40]. In this case, only the top two mixers in Figure 6.3-1 are active. On the other hand, if a complex waveform is presented to the multiplier, such as in a transmitter design, then generation of a real bandpass waveform can be directly implemented using the 1st and 3rd multipliers.

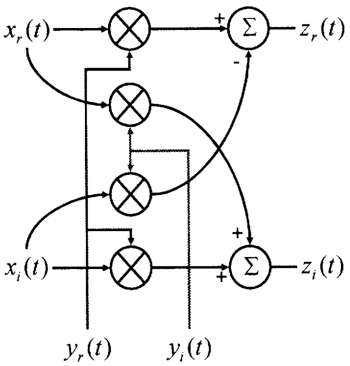


Figure 6.3-1: Complex Signal Multiplication.

In this section, the virtues of complex signal analysis are captured by three contemporary integrated modular receiver designs. The differences between each of the designs flow from the different configurations of the complex multiplier in Figure 6.3-1 as well as the final IF presented to the ADC. Figure 6.3-2 illustrates the first of two Low-IF architectures. This particular architecture is coined Low-IF owing to a target IF that is low

[40] A physical waveform must be real.

enough to enable integrated filtering yet high enough to avoid the pitfalls associated with zero IF designs[41]. Recall from Figure 6.1-3 that a sinusoid contains both positive and negative frequency components and that the existence of these components is responsible for IM distortion in a receiver. If the complex multiplier from Figure 6.3-1 is employed, then the sinusoid modulating tone can be replaced by the Fourier transform pair given in Equation (6.3.1). The complex exponential function in Equation (6.3.1) is referred to as a complexoid [44].

$$e^{\pm j2\pi f_{LO}t} = \cos 2\pi f_{LO}t \pm j\sin 2\pi f_{LO}t \Leftrightarrow \partial(f \pm f_{LO}) \tag{6.3.1}$$

Notice from the delta function on the right of (6.3.1) that the complexoid contains only a positive *or* negative frequency component. Recall that the Fourier transform of a sinusoid contains both. In comparison, the convolution of a bandpass waveform with a complexoid eliminates the need for filtering because the tone image that would mix the interferer to the desired IF doesn't exist. The product of the received waveform with the complexoid is captured in the second and third row of Figure 6.3-3 [42]. From the 4th row of the figure, a low-Q low pass *integrated* filter is employed to suppress close-in blockers and eliminate the up-converted images. Note, the waveform immediately following the low pass filter includes both the target waveform and the image, mirrored about the y-axis. A specially designed filter called a polyphase filter is used to eliminate the image [45,46]. To the extent that the image has been completely removed from the spectrum of the waveform by the polyphase filter, a real sinusoid (as opposed to a complexoid) generated in the digital domain is used to shift the waveform down to baseband for channel selection and demodulation. In the context

[41] Flicker noise, 2nd order IMD, and offsets are three of the most prominent.

[42] Note, a multiplication operation in the time domain maps to a convolution operation in the frequency domain. What's more, the convolution operator mirrors the modulating waveform about the y-axis.

of Figure 6.3-1, only the upper-most and lowest multipliers are needed for the final down-conversion.

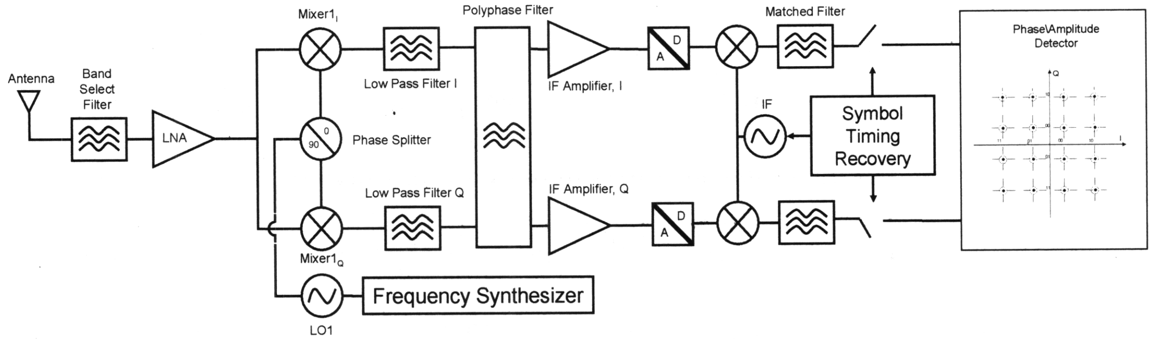


Figure 6.3-2: Low IF Design 1.

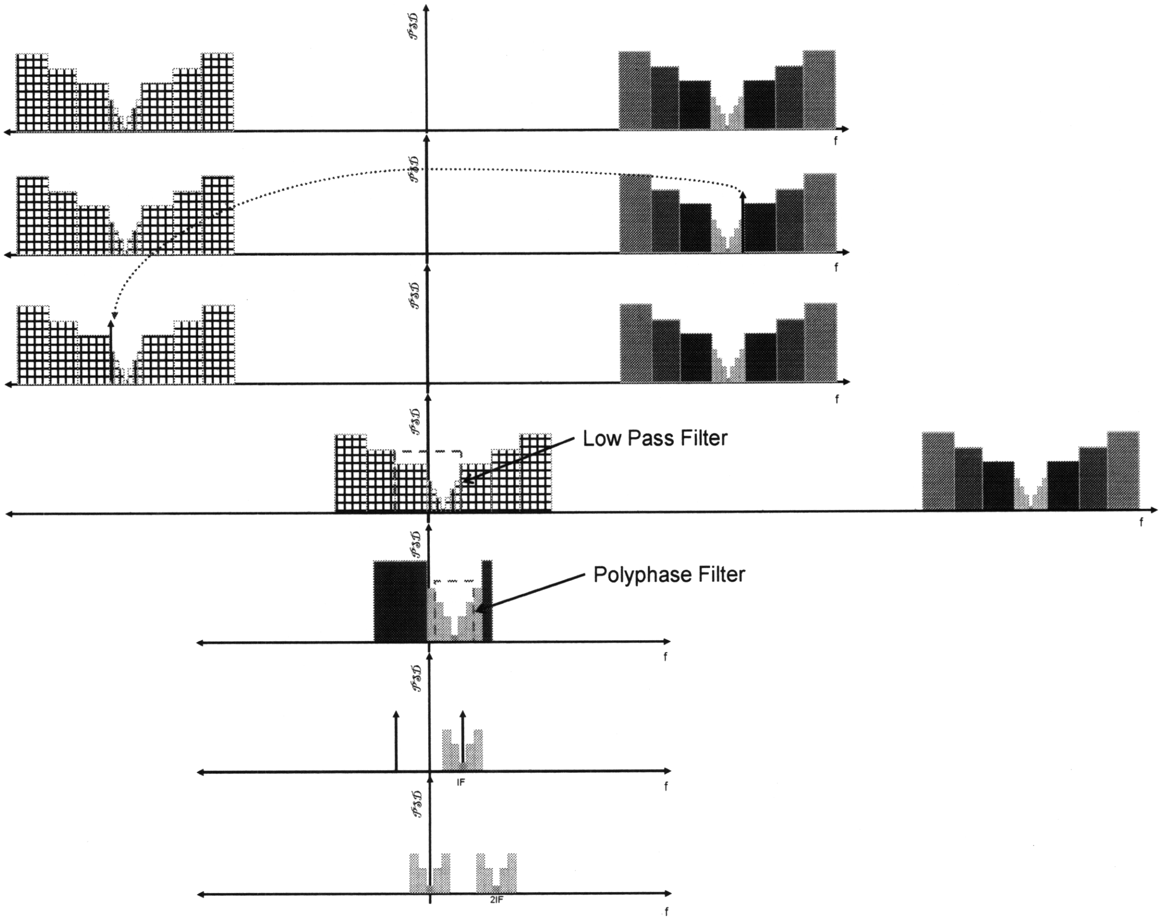


Figure 6.3-3: Low-IF Down-conversion Via Polyphase Filters. The high Q receive band filter is employed for this and all receivers discussed in this section. Therefore, the top two rows in Figure 7.2-3 are relevant and precede the top row shown in this figure. The second row illustrates the fundamental difference between the operation of designs of this architecture versus the superheterodyne. In this

design, a single delta function corresponding to the positive or negative complexoid of a cosine term is used to multiply the waveform. Per the discussion in the caption of Figure 7.2-3, the delta function is mirrored to the opposite side of the DC axis prior to the multiplication operation. The product is a low IF centered waveform that is low-pass filtered to eliminate the adjacent channels. The half multiplication is employed to down-convert the lo-IF waveform to baseband.

The second architecture that employs complex analog signal processing techniques is also a Low-IF architecture but uses standard filters in place of the polyphase filter introduced in the last design. See Figure 6.3-4 for the block diagram. Notice that the front end of this design is the same as the design in Figure 6.3-2. Therefore, the waveform processing up until the omission of the polyphase filter is exactly the same. Figure 6.3-5 depicts the function corresponding to the form of Figure 6.3-4[43]. With regard to the baseband mixing function, since the polyphase filter is omitted from this design architecture, mixing the Low-IF waveform with a plain sinusoid will degrade the SNR. Therefore, a complex exponential is employed in the baseband down-conversion in the digital domain. Because the target waveform is complex, the full machinery of Figure 6.3-1 is employed in Figure 6.3-4 for the baseband conversion.

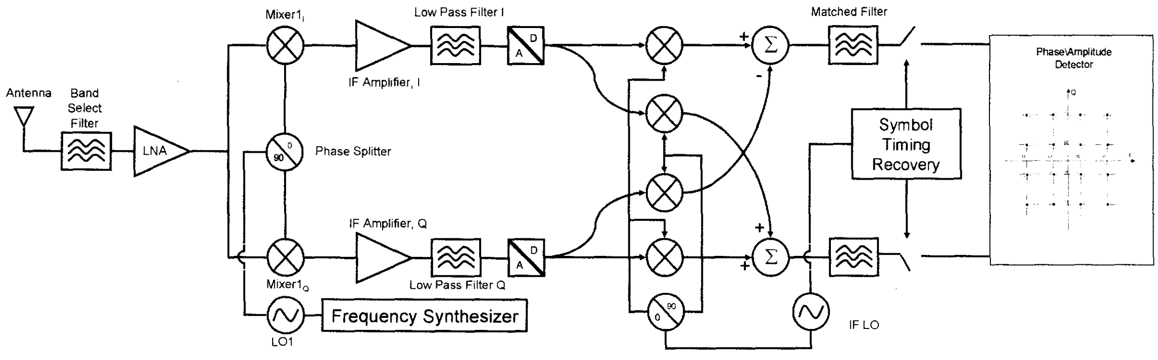


Figure 6.3-4: Low-IF Receiver 2.

[43] Owing to the similarities, this figure assumes the first four rows of Figure 7.2-8.

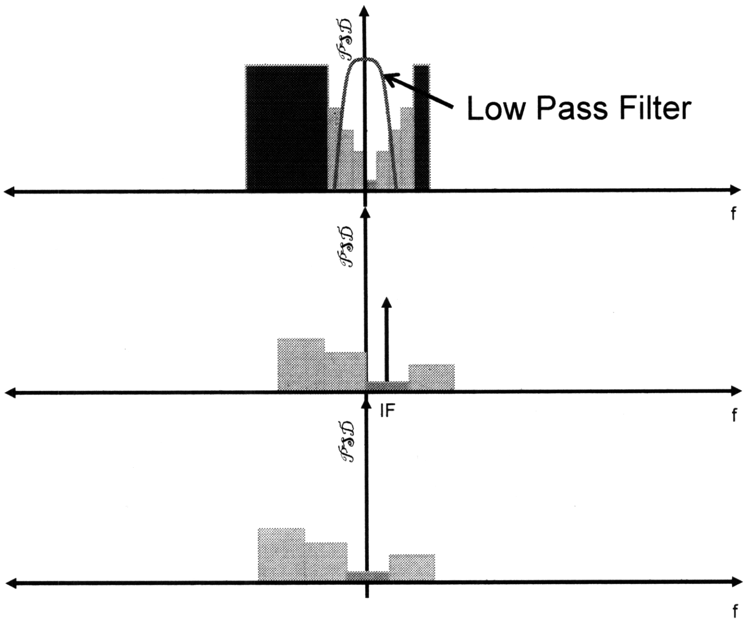


Figure 6.3-5: Low-IF Down-conversion Via Quadrature Mixing. The polyphase filter is absent from this design, and instead a low pass filter is used to suppress adjacent channel blockers. The presence of positive and negative frequency components in the low IF waveform places constraints on the digitally implemented baseband down-conversion design.

The final image reject receiver architecture discussed in this section is called the zero IF architecture. Figure 6.3-6 and 6.3-7 illustrate the block diagram and signal flow, respectively. As the name suggests, zero IF designs do away with the IF and convert the received waveform directly to baseband. The down-converted baseband waveform is then low-pass filtered and digitized prior to demodulation. The seemingly simple architecture is complicated by the need for DC offset correction that is used to negate the effects of LO leakage. LO leakage is the term used to describe coupling of the strong LO signal to the received signal path prior to mixing. The result of self mixing between the LO leakage and the LO injected into the mixer by the synthesizer is a DC offset that often overwhelms the target waveform. Assuming the components in the receiver were designed with the received RF power distribution in mind, a non-corrected baseband waveform corrupted by LO self-mixing can saturate the ADC. In the jargon of this report, the absolute bottleneck associated with ADC dynamic range would become active, and severely limit the performance of the design.

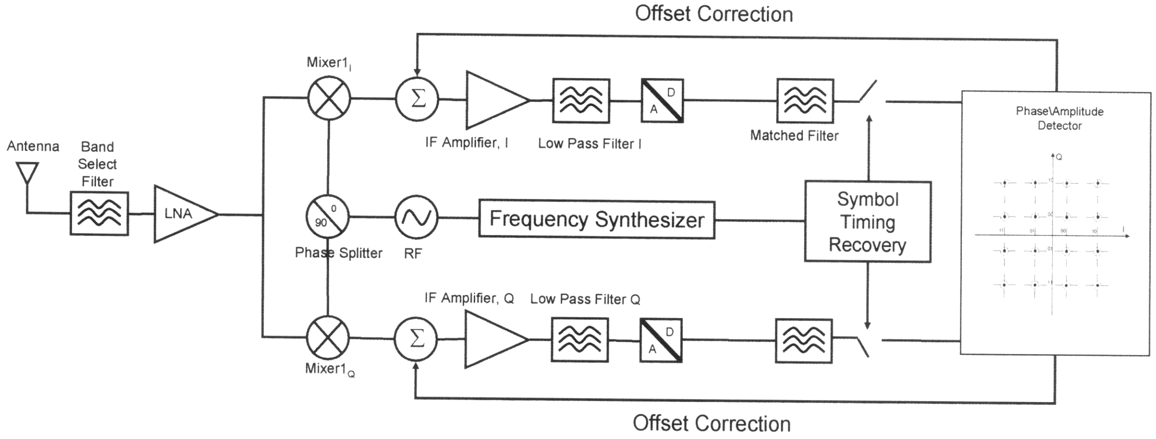


Figure 6.3-6: 0 IF Architecture

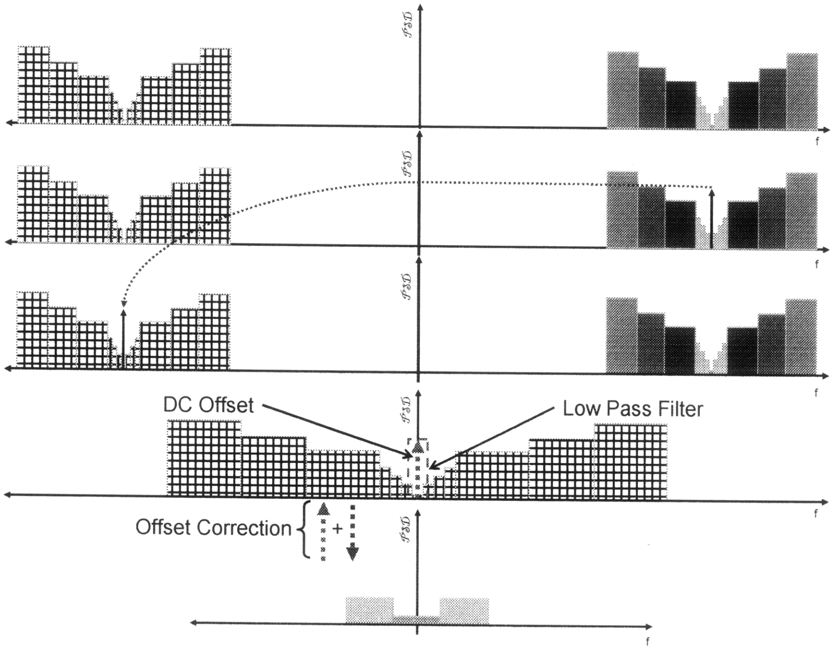


Figure 6.3-7: Direct Conversion Signal Flow. The complex exponential is used to down-convert the entire receive band. The effects of the DC offset are mitigated by the feedback loop between the DSP and the radio.

In comparison, the 2nd Low-IF and zero IF receiver architectures exhibit striking similarity. Of all of the commonalities, the most important is the placement of the ADCs in the receive chain. The ADC is the interface between the analog and digital domain. From a design point of view, the analog portion of a particular design is fixed for the life of the product. Notwithstanding, *all* functionality to the right of the ADC is software configurable; or in other words can be changed ex-post design. Because the placement of the ADC's is the same, it is reasonable to believe that a single design could support both low-IF and zero IF receivers. This assumes that the *exclusion* and *augmentation* operators are employed to handle the offset issue inherent in direct conversion and provided the ADC is designed with *inversion* in mind.

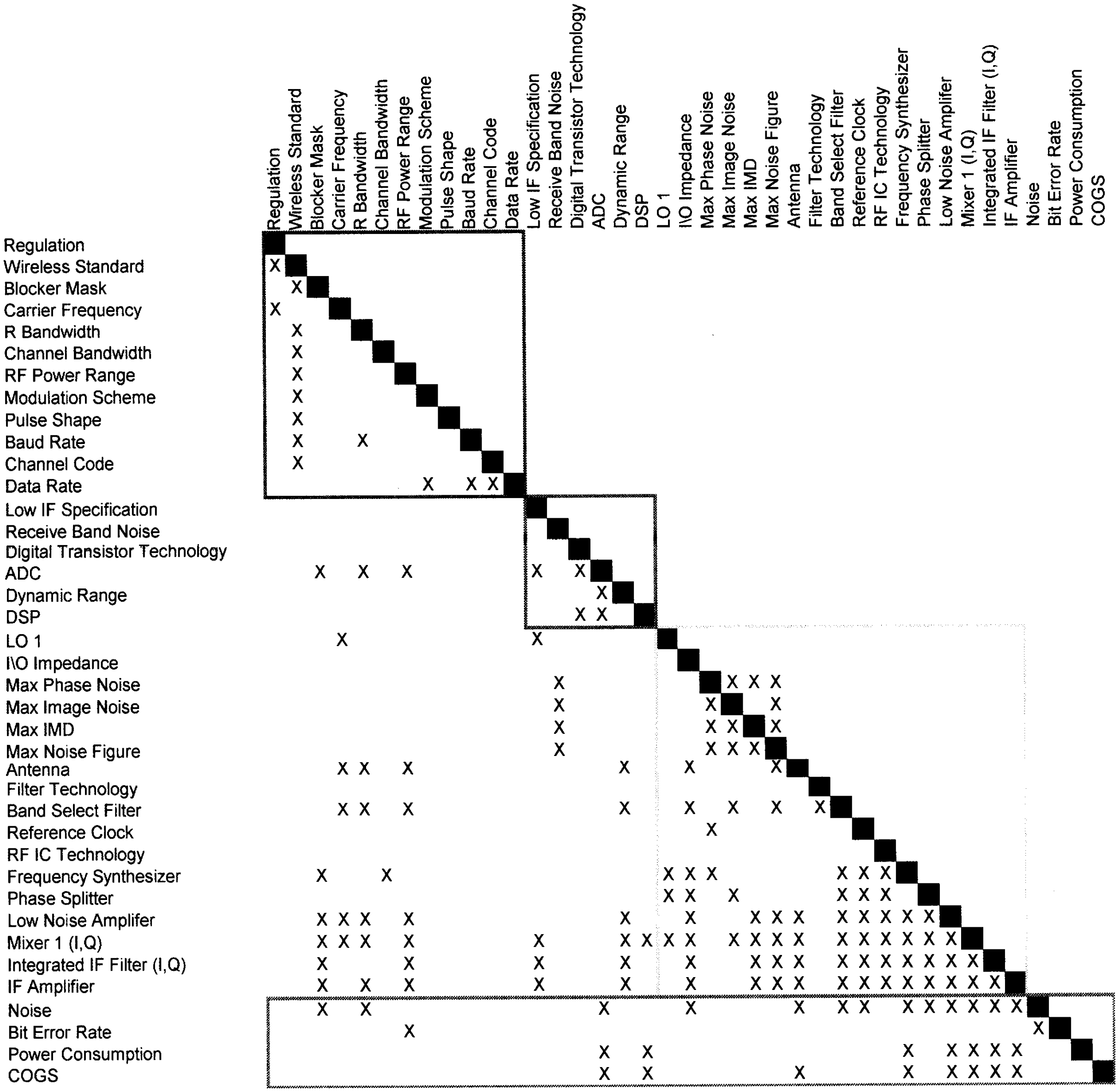
The viability of the three radio architectures described in this sub-section depend on the quality of the complexoid applied to the multipliers. If the sinusoids applied to the multipliers from Figure 6.3-1 are not exactly 90 degrees out of phase, or in other words if error terms are introduced into Equation (6.3.1), then image conversion will contaminate the converted IF band. Equation (6.3.2) through (6.3.3) captures the effect of phase and amplitude induced errors. In the limit of the error terms approaching zero $(\phi,\varepsilon \to 0)$, P goes to one and N goes to zero.

$$y(t) = \cos 2\pi f_{LO} t \pm j(1+\varepsilon)\sin(2\pi f_{LO} t + \phi) = Pe^{j2\pi f_{LO} t} + Ne^{-j2\pi f_{LO} t} \qquad \textbf{(6.3.2)}$$

$$P = \frac{1+(1+\varepsilon)e^{j\phi}}{2}$$
$$N = \frac{1-(1+\varepsilon)e^{-j\phi}}{2} \qquad \textbf{(6.3.3)}$$

Equation (6.3.2) says that amplitude or phase errors are manifested in the form of a mirror image of the desired complexoid. Clearly from Figure 6.1-3, any parasitic complexoid will mix with the image of the target RF band thereby nullifying the intent of the design. It can be shown that a phase difference of only one degree contributes -35dB of image noise to the IF. Similarly, an amplitude error of 1% contributes -40dB of image noise to the IF. Likewise, post conversion, since the downstream demodulation decisions are based on the I and Q waveforms from the original down-conversion, amplitude and phase matching in the signal paths are critical. This short discussion rationalizes an additional abstraction of the performance bottlenecks described in Section 6.1. Namely, phase and amplitude precision replaces the more direct image generated noise bottleneck that is relevant to the modular superheterodyne design.

For comparison, the DSM of the 2nd low IF receiver discussed in this section is presented in Figure 6.3-8. Notice that in the case of the low-IF architecture, the box defining the design variables associated with the component suppliers is now much larger than that belonging to the system designers. This is a consequence of the system designers' abrogation of component specifications to the firms designing the radio solutions. Whereas before the emergence of the integrated design architecture the system designers held a powerful up-stream position in the design flow, post integrated-design-architecture the system designers identify merely two modules, the ADC and DSP, and specify only the combined noise performance. Additionally, the new design architecture introduces cyclical interdependencies in what was once a sequential and independent design flow. Interestingly, the block containing cyclical dependencies belong to the owner of the receiver module who reserves the right to trade off the noise contributions determined by Equations (6.1.7), (6.1.9), and (6.1.11). The increase in technical potential relative to the modular superheterodyne design derives from this flexibility but, due to manufacturing capital barriers, comes at the expense of the number of firms capable of supplying component modules. To the extent that handset ODMs continue to fill both the smaller system design role as well as the system integrator roles, they still possess the keys to the error rate bottleneck and reap the modular architecture option value. However, their importance in the value network, at least from a design point of view, is very much diminished.



	Regulation	Wireless Standard	Blocker Mask	Carrier Frequency	R Bandwidth	Channel Bandwidth	RF Power Range	Modulation Scheme	Pulse Shape	Baud Rate	Channel Code	Data Rate	Low IF Specification	Receive Band Noise	Digital Transistor Technology	ADC	Dynamic Range	DSP	LO 1	I\O Impedance	Max Phase Noise	Max Image Noise	Max IMD	Max Noise Figure	Antenna	Filter Technology	Band Select Filter	Reference Clock	RF IC Technology	Frequency Synthesizer	Phase Splitter	Low Noise Amplifer	Mixer 1 (I,Q)	Integrated IF Filter (I,Q)	IF Amplifier	Noise	Bit Error Rate	Power Consumption	COGS
Regulation	■																																						
Wireless Standard	X	■																																					
Blocker Mask		X	■																																				
Carrier Frequency	X			■																																			
R Bandwidth		X			■																																		
Channel Bandwidth		X				■																																	
RF Power Range		X					■																																
Modulation Scheme		X						■																															
Pulse Shape		X							■																														
Baud Rate		X			X					■																													
Channel Code		X									■																												
Data Rate								X		X	X	■																											
Low IF Specification													■																										
Receive Band Noise														■																									
Digital Transistor Technology															■																								
ADC			X		X		X						X		X	■																							
Dynamic Range																X	■																						
DSP															X	X		■																					
LO 1				X									X						■																				
I\O Impedance																				■																			
Max Phase Noise														X							■	X	X	X															
Max Image Noise														X							X	■		X															
Max IMD														X							X	X	■	X															
Max Noise Figure														X							X	X	X	■															
Antenna				X	X		X										X			X				X	■														
Filter Technology																										■													
Band Select Filter				X	X		X										X			X		X		X		X	■												
Reference Clock																					X							■											
RF IC Technology																													■										
Frequency Synthesizer			X			X													X	X	X						X	X	X	■									
Phase Splitter																			X	X		X					X	X	X		■								
Low Noise Amplifer			X	X	X		X										X			X			X	X	X		X	X	X	X	X	■							
Mixer 1 (I,Q)			X	X	X		X						X				X	X	X	X		X	X	X	X		X	X	X	X	X	X	■						
Integrated IF Filter (I,Q)			X				X						X				X			X			X	X	X		X	X	X	X	X	X	X	■					
IF Amplifier			X		X		X						X				X			X			X	X	X		X	X	X	X	X	X	X	X	■				
Noise			X		X											X				X					X		X	X		X	X	X	X	X	X	■			
Bit Error Rate							X																													X	■		
Power Consumption																X		X												X		X	X	X	X			■	
COGS																X		X							X					X		X	X	X	X				■

Figure 6.3-8: Lo IF Receiver DSM. In comparison with Figure 6.2-4, this DSM shows that the design power of the system designers is greatly reduced. Otherwise, the design flow is very similar with one exception. In a integrated modular design, the system designers, now the component designers, reserve the right to trade off noise sources in order to meet the target error performance. The DSM suggests that a design sequence to the integrated modules defined in the block diagram. In reality, the integrated modules are design concurrently using advanced CAD tools. The cost incurred by rework owing to corruption of the natural sequence is low in the CAD tool design domain.

The remainder of this section will be devoted to the integrated modular transmitter module. Recall from Equation (4.3.2) that a bandpass waveform can be synthesized by modulating the analytic term $e^{-j2\pi f_c t}$ with the complex envelope of the information bearing baseband waveform. Stated this way, the implementation of Figure 6.3-1 that is widely used in receiver designs becomes immediately relevant to the transmitter design. What's more, since the desired output from the transmitter is a real waveform, the multipliers associated with the imaginary part of the product are excluded from the design. Figure 6.3-9 illustrates a typical digital bandpass transmitter. The signal flow associated with the transmitter is illustrated in Figure 6.3-10. In short, the DSP provides the real and imaginary components of the baseband waveform to the quadrature multipliers. The positive frequency complexoid translates the baseband waveform to the RF carrier frequency. The real RF waveform is then amplified and filtered on its way to the antenna. Note, before the power amplifier, the phase noise bottleneck is of primary concern. Clearly from Figure 6.3-9, only phase noise can corrupt the waveform prior to the power amplifier. From the latter part of section 6.1, battery life and IMD bottlenecks dominate once the waveform encounters the PA.

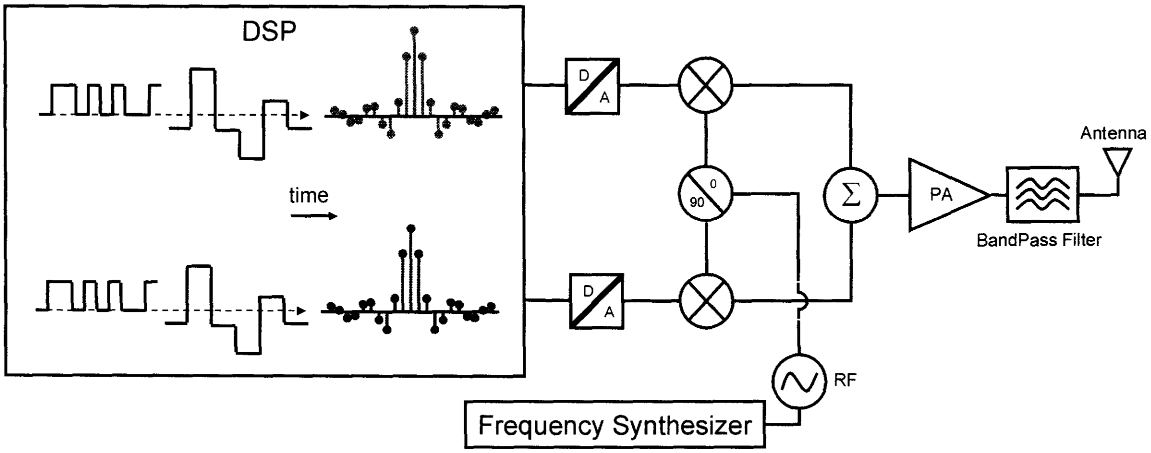


Figure 6.3-9: Direct Conversion Transmitter.

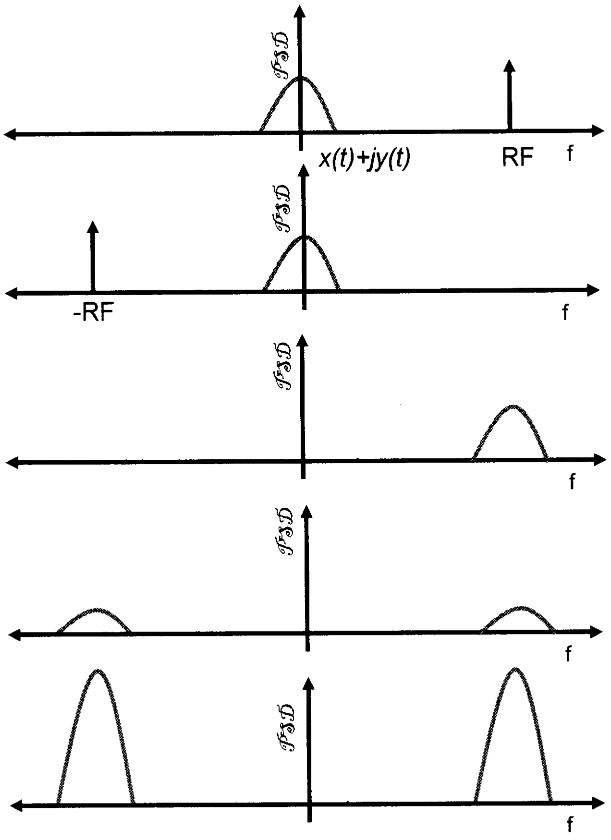


Figure 6.3-10: Transmit Signal Flow. The first two rows illustrate the mechanics of the convolution operator. Row 3 illustrates the waveform translation. The omission of the imaginary component of the complex product effectively splits the waveform into mirror images of itself. The real waveform is then boosted by the power amplifier.

Both the superheterodyne design architecture and the integrated modular architecture employed the splitting and substitution operators to create value. In the case of the superheterodyne architecture, the modular operators embodied by the design rules were defined by system designers at the handset layer of the value network and disseminated the component layer to foster component markets and create competition. In the integrated modular designs, the operators were employed within the walls of the manufacturing firm producing the receivers and transmitters. Notwithstanding, contemporary designs also demonstrate the effects of the inversion and porting operators. Comparison of the transmit and receive meta-tasks in Figure 6.1-2 in the context of the integrated modular designs suggests redundancy that could be eliminated by careful application of the *inversion* and *porting* operators. Figure 6.3-11 illustrates a consolidated solution neutral design that delivers the functionality for both the transmit and receive meta-tasks.

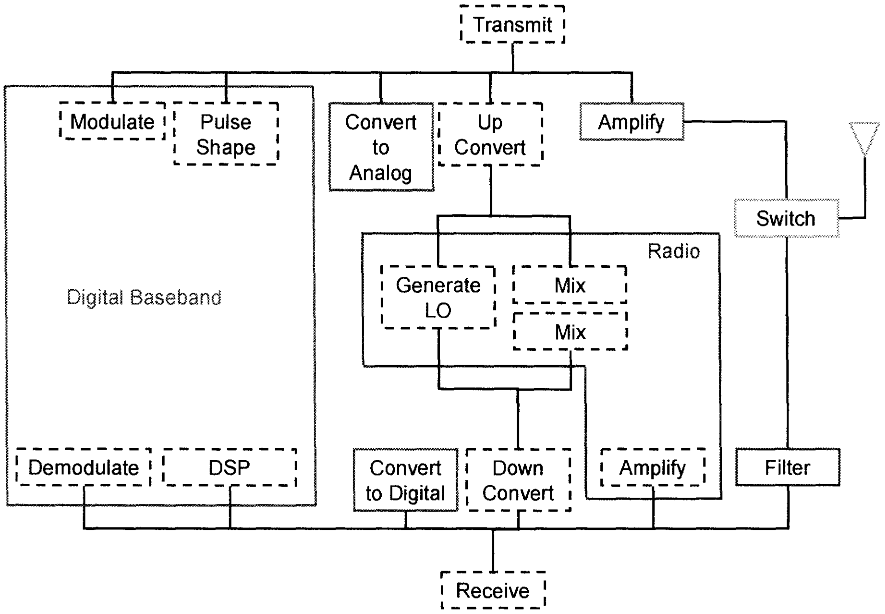


Figure 6.3-11: Contemporary Transceiver Architecture. The tasks in dashed boxes are split and assigned to integrated meta-modules in handset designs. Those tasks not aggregated into meta-modules exist as standalone modules themselves. That being said, the analog to digital and digital to analog converters are often co-located on the same chip as the digital baseband.

The combined design is referred to as a transceiver. From the figure, the tasks have been divided into eight modules: digital baseband, radio, ADC, DAC, receive filter, power amplifier, switch and antenna. In most systems, the ADC and DAC modules are implemented on the digital baseband chip adjacent to the DSP core [36]. What's more, owing to the similarity in the amplifier and switch transistor technology, these two modules have been combined (not integrated!) in a meta-module referred to as the front end module.

Figure 6.3-12 illustrates an integrated modular design employing the low-IF receiver architecture and direct conversion transmitter.

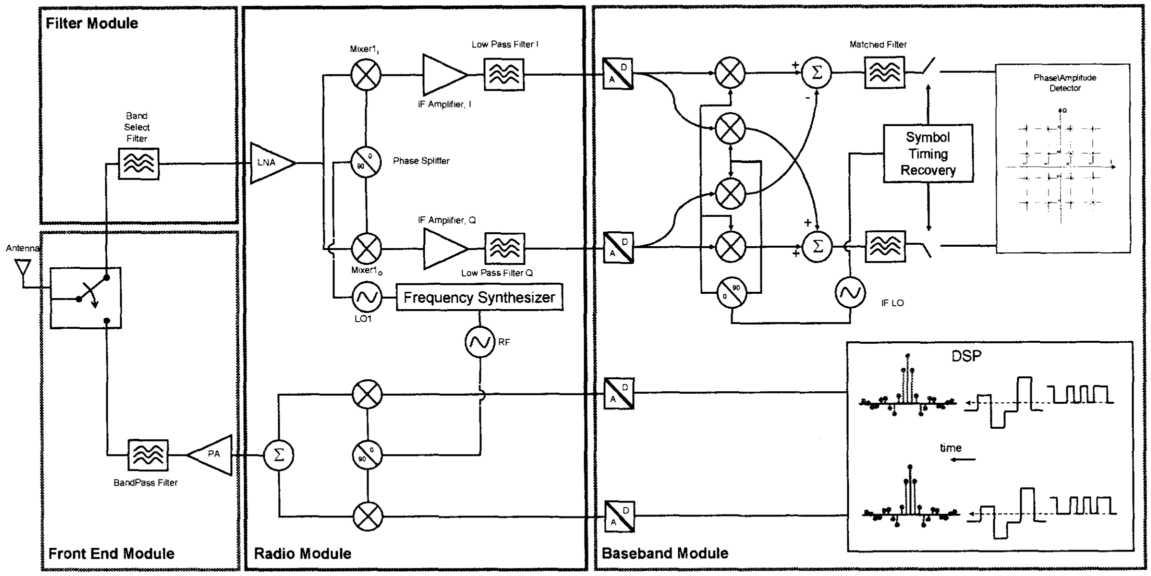
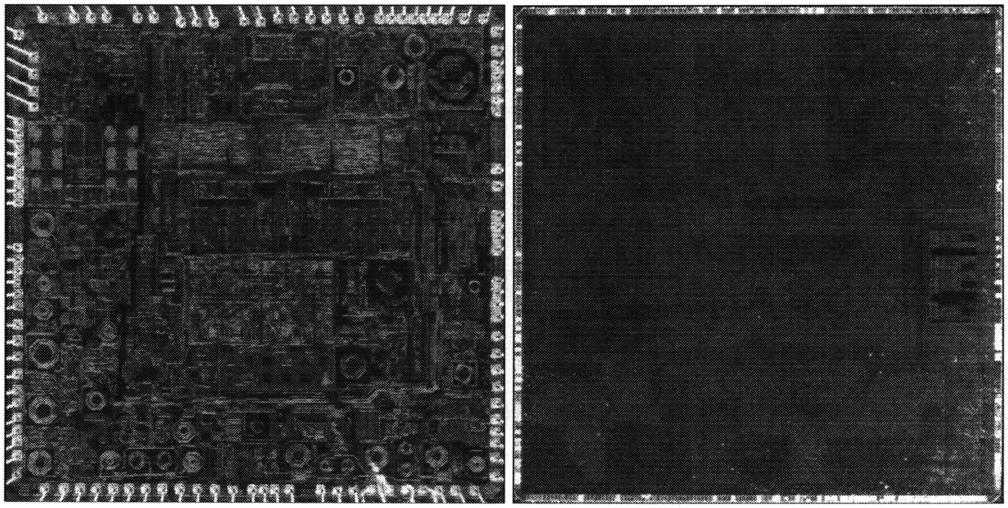


Figure 6.3-12: Contemporary Transceiver Design. The transceiver design illustrates the 4 meta-modules present in contemporary transceivers. True to the figure, most designs of this nature include 4 disparate integrated circuits connected on the handset circuit board.

The front end module is the remaining pure modular design wherein the constituent sub-modules are not dependant on the same manufacturing technology. In some scenarios, FEMs include the electro-mechanical receive filters which effectively collapse the pink and green boxes in Figure 6.3-12 into a single box. The preservation of modularity in the FEM owes to the fact that each component technology that is tasked with meeting performance requirements is vastly different from the others. For example, power amplifier designs are comprised of GaAs heterojunction bipolar transistors; filters are implemented using surface or bulk acoustic wave resonators, switches are made using GaAs pHEMT transistors, and adhoc filters such as those used in the transmit filter and inter-component matching employ surface mount passive components[44].

Figure 6.3-13 illustrates the integrated modular architecture for the transceiver and the baseband chip that includes the ADC\DAC and signal processing circuitry. Figure 6.3-14 illustrates the front end module found in contemporary handsets [37]. From the figures, each component appears very different from the others. More will be said about the architectural differences in Figure 6.3-13 in the subsequent section. Meanwhile, the differences between Figure 6.3-13 and Figure 6.3-14 are painstakingly obvious. In the FEM, each sub-component is attached to a printed circuit board and connected to one another in a fashion that realizes the power and switching intent of the FEM. Because each component is fabricated using disparate technologies, there exists no appreciable integration threat. The same cannot be said for the separate modules in 6.3-13. Nonetheless, the front-end module is judged by the system designers and integrators to be similar to packaged single-chip solutions and therefore inherit the cost expectations deriving from monolithic integration. These expectations make ROIC immediately relevant to firms competing in the FEM space.



[44] The term matching refers to setting impedances particular to the frequency of operation.

Figure 6.3-13: RF Transceiver Design (left) and Digital Baseband Design (right) in Nokia N95.[37]. In comparison, the relative elegance of the digital baseband chip occurs as a result of auto-generated artwork. To the contrary, owing to tacit knowledge that is difficult to map to automated computer programs, RF designs are generated by engineers and technicians. [37]

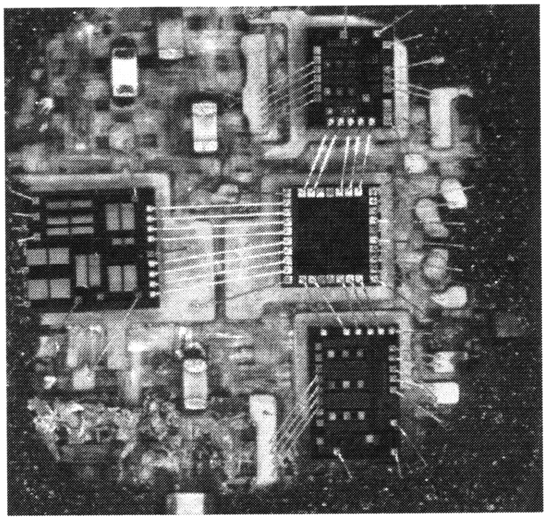


Figure 6.3-14: GSM\GPRS Front End Module Sans Receive Filter. Notice the rather "uncivilized" assembly of this module relative to the integrated designs from Figure 6.3-12. Firms competing in the FEM space focus on design of the commoditized analog ICs and manufacturing excellence to achieve high ROIC. [37]

6.4 Digital Radio

Figure 6.1-2 illustrated the fundamental boundary between analog and digital modules present in heterodyne receiver architecture. Figure 6.3-11 illustrated the evolution of architecture from strictly modular to integrated modular architecture. The causes of the migration of architecture were hypothesized to derive from the value promised by the modular operators influenced by increased technical potential coupled with lower manufacturing cost. A natural question to ask is whether the architecture has reached its steady state or if stakeholders can expect revolutionary new changes in the architecture. Assuming the marginal cost of manufacturing is not grossly affected by integration of the larger chips, such as the digital baseband and transceiver, one could certainly make the case that the levels of modularity present in the contemporary physical layer designs are very close to optimal; especially given the technology challenges associated with designing analog components with digital IC technologies. What is needed to justify higher levels of modular integration is either an economic paradigm shift or the emergence of a killer application wherein an integrated architecture provides enabling functionality.

In fact both exist. On the purely economic side of the argument, the presence of Moore's Law incentivizes digital design firms to integrate more functions provided the ensuing designs scale with transistor feature size. With regard to killer applications, the notion of an integrated platform transceiver architecture capable of spanning several frequency bands and supporting many different network protocols offers a compelling case for additional integration; especially if the alternative is an aggregation of channel specific transmitters and receivers. Provided the physical design of such a system adds no direct value, per se, but rather is a conduit for the value added by software customization, then the entire value proposition of application specific hardware design is weakened by the software derived platform solution. In the context of Chapter 3, the ROIC of a software definable solution is higher than one comprised of specific hardware because the return on assets is higher. The statement assumes that the economies of scale can be realized by combining both the analog and digital designs via a single manufacturing process.

Implicit in both arguments for additional modular integration is a shift in architecture. The purely economic case requires strictly scalable designs in the Moore's Law sense. This means that reduction of the transistor feature size must reduce chip area without impacting the design [47]. Notwithstanding, the platform design scenario requires the radio design to be ex-post customizable to transmit and receive on different channels using different network protocols. The digital nature of the baseband module enables the later which shifts the focus to software definable transmit and receive functionality. Withstanding, the legacy architectural features inhibiting both Moore's Law scalability and ex-post customization can be traced to traditional RF artwork constraints and the presence of rigidly designed analog filters, respectively.

The poor scalability of RF designs flows from both transistor physics and impedance matching requirements between stages. The physics of deep sub-micron CMOS technology is such that the performance metrics important to analog\RF design don't scale linearly with transistor channel length [34]. This means that the analog portion of an integrated design must be redesigned upon each new generation of CMOS technology. What's more, on-chip impedance matching requires components such as inter-layer capacitors and spiral inductors that have properties determined by metal geometry and dielectric properties and therefore are un-affected by transistor scaling. Figure 6.3-13 subtly illustrates the scalability difference between a contemporary RF design and a scalable digital baseband design. The figure shows the artistic qualities of the RF design owing to the integrated matching components as compared to the dense nature of the digital baseband design. From the figure and the previous discussion, an integrated modular design containing both transceiver and baseband designs will only scale proportionally to the percentage of the digital footprint to the overall design. Assuming the designs on the left and right in Figure 6.3-13 are drawn to scale and that the two designs are monolithically integrated, then reduction in feature size will exhibit diminishing returns.

On the application side, the analog filter problem inhibits software definable radio in two ways. First, the presence of the channel select filter in the receiver fixes it to a particular frequency band and eliminates any possibility of serving multiple bands. If a receiver design is to be made capable of serving multiple bands, then the channel select filter must be abandoned. The second issue is a consequence of the anti-aliasing filters that necessarily precede the ADCs. The primacy of analog to digital conversion in software defined radios was introduced by Mitola in 1995 [48]. Per Mitola's work, *"The placement of the A/D/A converters as close to the antenna as possible and the definition of radio functions in software are the hallmarks of the software radio."* However, per the aliasing constraints placed on sampled waveforms, some form of filtering must take place prior to analog to digital conversion [16]. The challenge is enabling flexible filters that can support sampling waveforms of various bandwidths.

Reflecting on the aforementioned challenges, a viable modular integrated architecture combining the digital baseband and RF transmitters and receivers should exhibit the following qualities: minimal traditional RF circuitry, flexible anti-aliasing filters, and absence of channel-select filters. Whether by chance, or owing to the logic of flexible integrated circuits, the last two features are achieved by designs exhibiting a reduced traditional RF footprint. Recent monolithic designs pioneered in academia and industry have proven to be simultaneously specification compliant for Bluetooth, GSM, and 802.11 standards [49][45]. Figure 6.4-1 illustrates the receiver architecture common to both designs. The operation of the receiver is illustrated in Figure 6.4-2 and follows the direct conversion receiver design presented by Bagheri [49].

[45] GSM channels spanning 200kHz and 802.11 channels spanning 20MHz illustrate high degrees of software scalability.

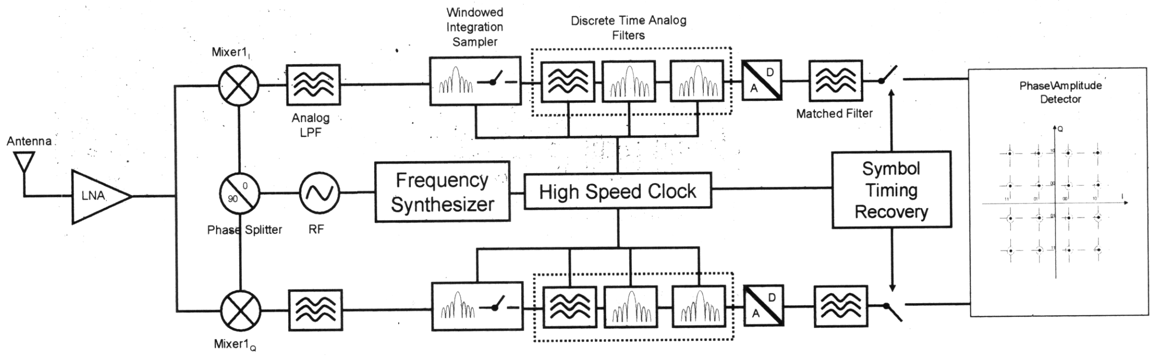


Figure 6.4-1: Software Defined Radio Architecture. The mixer and LNA comprise the lone members of traditional RF modules. The remainder of the design is implemented using contemporary digital integrated circuit technologies. [48]

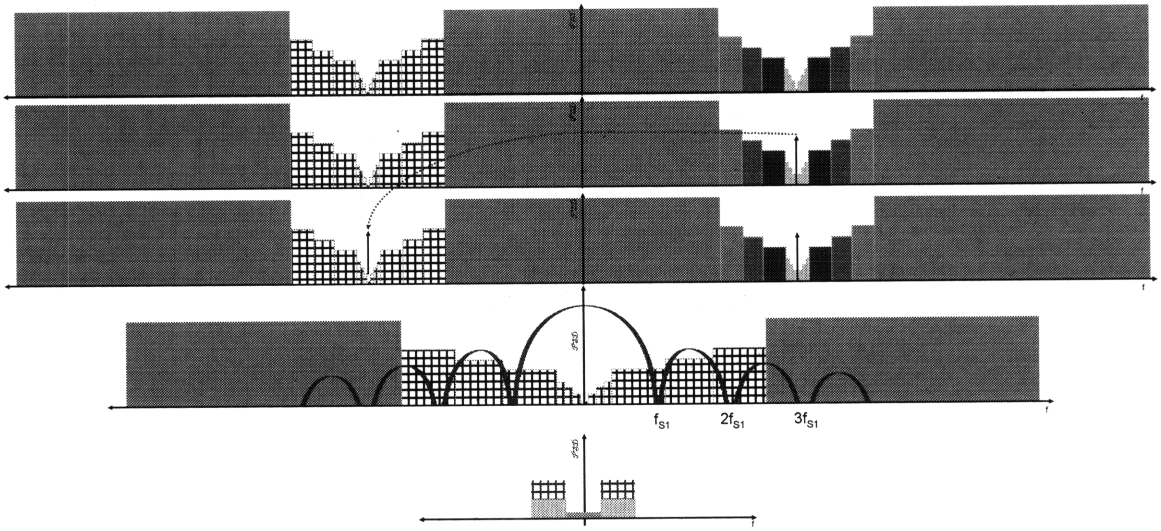


Figure 6.4-2: Software Defined Radio Signal Flow. The direct conversion mixer translates the I and Q waveforms to baseband. Aliasing effects at baseband are minimized within the nulls of the windowed integration samplers and discrete time filters. Subsequently, the level is determined by the depth of the nulls in the sampler and filters.

The software defined radio architecture is enabled by the introduction of the windowed integration samplers. From figure 6.4-2, the intent of the windowed integrated sampler is to filter the baseband converted waveform in a manner that is flexible *but* does not introduce aliasing in the downstream analog to digital conversion function. The nature of the sampler is such that the transfer function assumes a sinc function thereby introducing nulls in the passband at multiples of the sampling frequency [49]. See Equation (6.4.1)

$$H_{WI}(f) = \frac{T_W}{\tau}\frac{\sin(\pi f T_W)}{\pi f T_W} \qquad (6.4.1)$$

The sinc function characteristic is optimal because the nulls in the filter response occur at multiples of the sampling frequency which eliminates aliasing in the 0 IF waveform of interest [49]. Due to the rigid sensitivity requirements of wireless communication systems, imperfect implementation of the samplers, and the need to decimate the sampled waveform, additional filtering stages are required[46]. Discrete time analog filters are employed to both decimate the sampled waveform and provide additional filtering [49].

Figures 6.4-1 and 6.4-2 illustrate a design architecture that operates similar to the previous zero IF architecture. The fundamental difference between the two designs is the substitution of the analog baseband processing chain with the combination of the windowed integration samplers and the discrete time filters. From Equation (6.4.1), the transfer function of the windowed integration samplers is determined by the sampling\switching frequency responsible for moving charge on and off capacitors. Note, higher sampling frequencies map to closer spaced nulls in the transfer function that leads to better rejection of interferers. Anecdotally, decimation and additional filtering provided by the discrete time filters are also controlled by a slowed down version of the clock signal. Decimation is beyond the scope of this work, but suffice it to say that in order to ease the ADC requirements, the sampling rate is decimated to the Nyquist rate of the desired channel [49].

The dependence on the clock frequency means that the filter nulls can be changed ex-post design thereby realizing the antecedent of platform architecture and of software defined radio. The bottlenecks in this architecture include all of those introduced in Sections 6.1 and

[46] Decimation refers to the process of reducing the sampling rate.

6.3 and add clock speed and precision. Those familiar with microprocessors will identify clock speed as the primary performance bottleneck in standard digital designs. What's more, the nature of the windowed integration samplers and the discrete time filters is that they are comprised of switched capacitor networks that are indeed consistent with Moore's Law scalability. With regard to the IMD bottleneck, since the band select filter is eliminated from the design, each of the components in the receive chain must exhibit high even order intermodulation performance. Switched capacitors are highly linear circuits which mean that the burden of IMD is mainly placed on the LNA and the mixer [49].

The DSM of the digital radio is presented in Figure 6.4-3. In comparison with the Low IF integrated modular DSM, this DSM shows that the traditional system designers have been completely removed from the design process. Also, the separate dependencies on the RF and digital transistor technologies have been collapsed into a single high precision digital technology. From the system integrators' point of view, the value received by choosing a portfolio of options has been replaced by the option on a portfolio that is the digital radio design. However, given the platform qualities of this design it is reasonable to believe that a handset firm that sources digital radio designs to serve all of its wireless communication needs would achieve a higher ROIC than one that sources integrated modular designs for each frequency band and standard. What's more, the modular value *can be* conserved, at least in comparison with the integrated modular design, provided the component manufacturer possesses the resources and internal market to promote the modular operators internally. From this perspective, the component suppliers act as system designers and promote modular design internally.

	Regulation	Wireless Standard	Blocker Mask	Carrier Frequency	R Bandwidth	Channel Bandwidth	RF Power Range	Modulation Scheme	Pulse Shape	Baud Rate	Channel Code	Data Rate	Digital IC Technology	DSP Core	Receive Band Noise	ADC	Dynamic Range	I\O Impedance	Max Phase Noise	Max Image Noise	Max IMD	Max Noise Figure	Antenna	Reference Clock	High Speed Clock	Frequency Synthesizer	Phase Splitter	Low Noise Amplifer	Mixer 1 (I,Q)	Passive Filter (I,Q)	Windowed Integration Sampler	Discrete Time Filters	Noise	Bit Error Rate	Power Consumption	COGS
Regulation	■																																			
Wireless Standard	X	■																																		
Blocker Mask		X	■																																	
Carrier Frequency	X			■																																
R Bandwidth		X			■																															
Channel Bandwidth		X				■																														
RF Power Range		X					■																													
Modulation Scheme		X						■																												
Pulse Shape		X							■																											
Baud Rate		X			X					■																										
Channel Code		X									■																									
Data Rate								X		X	X	■																								
Digital IC Technology													■																							
DSP Core													X	■																						
Receive Band Noise															■																					
ADC					X		X						X	X	X	■																				
Dynamic Range																X	■																			
I\O Impedance																		■																		
Max Phase Noise															X				■	X	X	X														
Max Image Noise															X				X	■	X	X														
Max IMD															X				X	X	■	X														
Max Noise Figure															X				X	X	X	■														
Antenna				X	X		X										X	X				X	■													
Reference Clock																			X					■												
High Speed Clock																									■											
Frequency Synthesizer			X			X							X					X	X				X	X	X	■										
Phase Splitter													X					X		X			X	X			■									
Low Noise Amplifer			X	X	X		X						X				X	X			X	X	X					■								
Mixer 1 (I,Q)			X	X	X		X						X	X			X	X		X	X	X	X			X	X	X	■							
Passive Filter (I,Q)			X		X		X						X				X				X	X	X					X	X	■						
Windowed Integration Sampler			X		X	X	X						X				X					X	X		X			X	X	X	■					
Discrete Time Filters																									X			X	X	X	X	■				
Noise			X			X								X		X							X			X	X	X	X	X	X	X	■			
Bit Error Rate							X																										X	■		
Power Consumption														X		X										X		X	X	X					■	
COGS														X		X							X			X		X	X	X	X	X				■

Figure 6.4-2: Digital Receiver DSM. In comparison to the previous two DSMs, this one shows that the ODM's presence in the design flow has been completely eliminated. Otherwise, the DSM exhibits the same properties as those of the integrated modular design. Implicit in this DSM is the elimination of the analog component group in the supplier layer of the value network. Today, system integrator role is still filled by handset ODMs. As a matter of semantics, owing to their absence in the design flow, the system integrators in this DSM resemble more original equipment manufacturers than they do ODMs. This observation implies a new strategy on behalf of handset firms: achieve highest ROIC in manufacturing operations or find new ways to add value.

6.5 Trends and Observations

As a general observation, the previous four sections suggest that new bottlenecks emerge at each new layer of a design hierarchy. In choosing a design architecture to address a particular bottleneck or a collection of bottlenecks, new designs introduce new layers of performance bottlenecks or at least color the bottlenecks with solution specific attributes. The emergence of new bottlenecks corroborates the multiplication quality of the modular operators acting on a design. Figure 6.5-1 illustrates the concept in the form of a bottleneck tree.

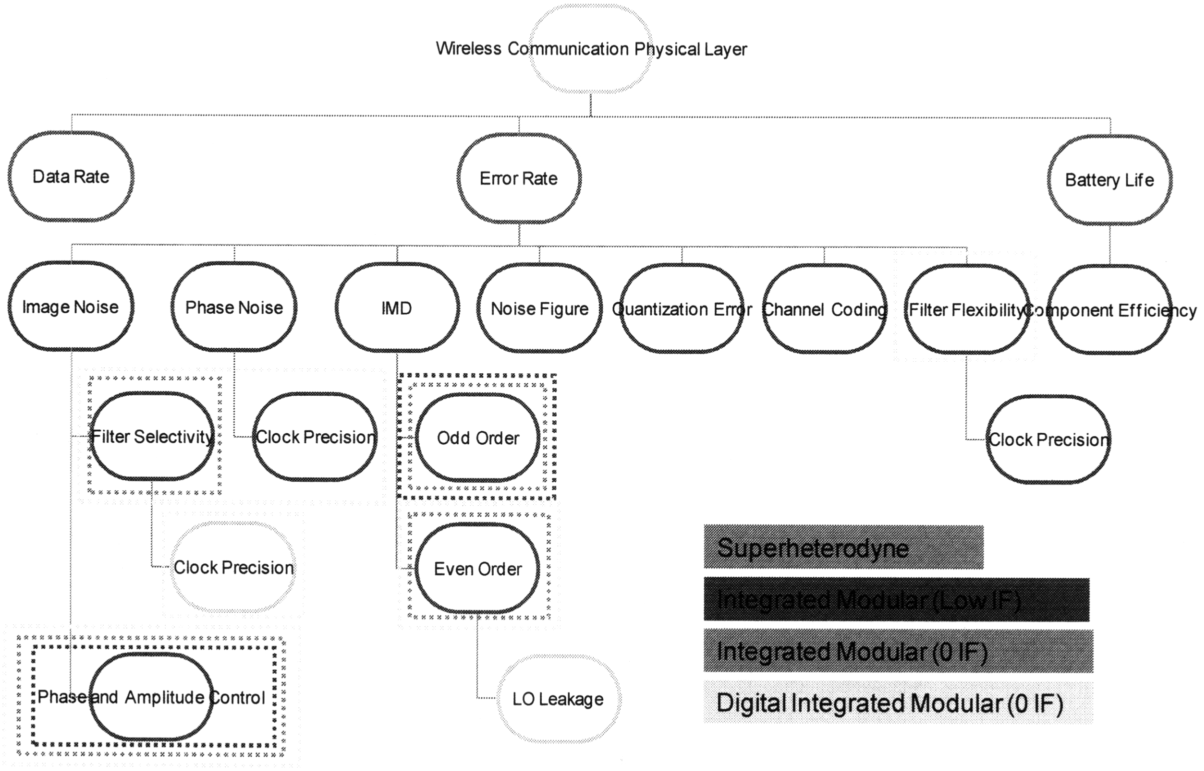


Figure 6.5-1: Bottleneck Tree Associated with Physical Layer of Heterodyning Radio Design.

In the figure, each row is a collection of bottlenecks that is directly related to the design targeting a bottleneck at the next highest level of abstraction in the overall architecture. Bandpass communication architecture introduced the first layer of bottlenecks derived in Chapter 4: data rate, error rate, and battery life. This chapter focused on the bottlenecks visible to the handset and supplier layers in the value network: data rate and battery life. Per section 6.1, conventional electronics and battery technology introduced component efficiency in to the battery life branch. Heterodyning design architecture introduced phase noise, IMD, noise figure, and quantization noise, into the error rate branch. Focusing on the IMD branch, the legacy superheterodyne design introduced filter selectivity and odd-order intermodulation to the image noise and IMD bottlenecks, respectively. Integrated modular architecture replaced the filter selectivity with phase and amplitude precision. The transition from low-IF to 0-IF replaced odd-order intermodulation with even-order intermodulation and introduced feedback. The contemporary emergence of digital radio architecture fostered a new peak in the value network associated with flexibility. Implementation using windowed integration samplers and discrete time signal processing led to additional refinement of the flexibility bottleneck, clock precision.

It remains to be seen whether this notion of bottleneck hierarchy is relevant to all complex designs. However, to the extent that it is, one can hypothesize that firms that make the decision to optimize communication networks and information filters around addressing certain bottlenecks could very well initiate the failure mechanisms described by Henderson and Clark [50]. If true, to mitigate the risk of failure due to architectural lock-in, a firm would need to invest in solving unpopular bottlenecks in order to make the seamless transition to emergent architectures. To preserve ROIC, the same firm may well choose to acquire relevant knowledge and technologies *after* the architecture shifts and new bottlenecks emerge.

7 Conclusions

Chapter 1 identified a shift in the competitive market wherein legacy handset firms were beginning to be challenged by new firms that identify with application specific bottlenecks. Chapter 2 provided an overview of the value network comprised of network operators, handset manufacturers, and component suppliers. The market power was shown to reside with the network operators owing to ownership of spectrum. The network operators offer subscriptions to consumers and subsidize handsets that enable them to differentiate from competing networks. The dominant subscription model was divided into two categories. The first category is predicated on long term subscriptions that enable network operators to subsidize otherwise expensive handsets. The second category, pre-paid\pay-as-you-go, lacks the lengthy subscription contracts and therefore requires lower subsidies or elimination of subsidies that in turn creates the demand for low cost handsets. In either case, profits based on the subscription model were shown to be inversely proportional to churn, which was naturally lower for long term subscription contracts. To target the long term contract market, network operators have identified with the user experience bottleneck that leads to investment in spectrum complemented by subsidized handsets. In response to the downward price pressure applied by network operators, handset firms have abrogated the physical layer designs to suppliers to focus on higher abstractions of functionality. In parallel, firms or product lines focused on the perfunctory performance market characterized by eroding ASPs and smaller subsidies from the network operators have also relinquished control of the physical layer design to minimize cost.

The observation of the trends in the industry motivated the adoption of a suitable management framework that could be used to explain the cause and effect of the migration of design control. The work of Baldwin and Clark was deemed suitable owing to its relevance in other high tech industries such as the personal computer [1]. The Baldwin and Clark framework established the concept of bottlenecks being absolute or relative and hypothesizes that firms develop products to improve on existing bottlenecks and align development strategies to be consistent with high ROIC [2]. In as much, competitive firms attempt to maximize ROIC by directly addressing some bottlenecks, while outsourcing others. Strategies associated with internal and outsourced design partitioning naturally identifies with the Baldwin and Clark framework of modularity.

The existence of a well established value network and dynamic bottlenecks necessitate a strong understanding of the nature of the bottlenecks so that each can be ascribed to the proper layer in the value network. Also of considerable import, the economics coupled with the nature of the bottlenecks may well provide an indication as to which layer in the value network that the bottlenecks will settle. Through a dense overview of relevant communication theory, the system level bottlenecks were shown to consist of data rate, error rate, and battery life. Of the three, the data rate remains comfortably in the hands of the network operators and an adhoc member of the value network, the standards bodies. The remaining bottlenecks were outsourced by the network operators to the handset ODMs and have eventually passed directly to the component suppliers. To identify the design rules that enable the handset and component markets to exist, the two dominant standards for cellular communication were examined. It was shown that both TDMA and CDMA systems are constrained by the same bottlenecks derived in Chapter 4.

The rigid set of specifications developed by the standards bodies coupled with physical limitations in solid state transistor technology led to the natural emergence of the heterodyning transceiver architecture. In the context of this design architecture, five new bottlenecks deriving from the error rate bottleneck were revealed. Knowledge of these bottlenecks coupled with well defined design rules fostered the emergence of a truly modular design, the superheterodyne radio. However, the handset ODMs that benefited from modular option value also suffered from the high cost of manufacturing. To address the manufacturing cost, integrated circuit technology improved to the point where a filter-less integrated modular architecture emerged. This new architecture allowed ODMs to shed manufacturing capital and grow at a higher rate. The integrated modular designs were shown to have changed the bottlenecks at the implementation level of the designs but conserved all of the system level bottlenecks deriving from the heterodyning macro-level design architecture. Per the DSM's in Chapter 6, the adoption of the new design came at the expense of the ODM's footprint in the design process which naturally led to reformulation of their development strategies. Firms once considered innovators of the physical layer components began to outsource the designs. For example, Motorola possessed a physical layer components business unit until 2004 when it was spun out as Freescale Semiconductor. Similarly, Nokia designed transceivers internally until 2007 when the company entered into a licensing and outsourcing agreement with ST Microelectronics for future radio designs.

New economics coupled with the promise of platform radio solutions emerged from the integrated modular designs which all but eliminated the handset ODM from the design of the physical layer. The gradual shift of the physical layer design to the integrated architectures was stated to conserve modular option value provided the firms now in possession of the design architecture enable an internal structure to promote the modular operators. Otherwise, the upstream members of the value network, who have knowingly and willingly abrogated the design to the component layer, are forced to accept designs that are effectively options on portfolios and worth less than if they resembled portfolios of options.

The question remains, how will the value network look in the future? As before, one must first examine the top of the value network. The network operators' entitlement is currently a matter of policy. With the emergence of digital radio, the once fantastic dream of cognitive radio is becoming feasible [48]. Cognitive radio is a variant of bandpass communication wherein the communication device identifies unused spectrum to communicate information independent of ownership of the spectrum being consumed. In a cognitive radio environment, the spectral allocation is handled at the time of the communication event. Sans a secondary market for spectrum, an efficient system would necessarily violate the property rights held by the network operators and therefore would require a change in regulation policy. Interestingly enough, to the extent that unoccupied airwaves are being horded, one could make the argument that the current system of assigning property rights to network operators to address the tragedy of the commons is in fact socially inefficient. Nonetheless, the technology has arrived and it remains to be seen if or when the policy makers will react.

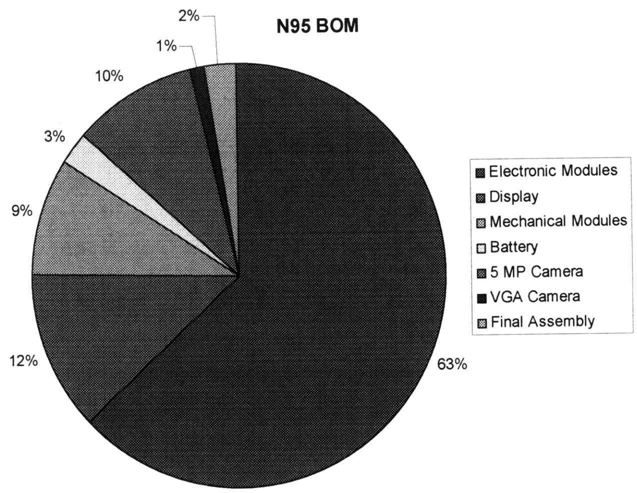
Provided cognitive radio becomes a reality, the largest jolt to the value network will occur at the top. In a scenario where all airwaves must be made available for consumption, it is reasonable to believe the airwaves will be reclaimed by government, else a secondary market system for unused spectrum would have to emerge. Either scenario would likely require regulation of spectrum allocation algorithms, which is the domain of standards bodies. In the new value landscape, the handset ODMs aren't likely to digress. In other words, the revival of a truly modular architecture wherein the ODM is in sole possession of the physical layer design architecture is highly unlikely. Notwithstanding, the contemporary bottlenecks associated with application specific experience is likely to endure. Therefore, the handset ODMs would be wise to grow their competence along the experience bottleneck.

Per the trends in Chapter 6, the solutions to the application specific experience bottlenecks will foster the emergence of implementation layer bottlenecks. In scenarios that require access to the information cloud and on-board computing power, the quality of the experience will likely be dominated by management of spectrum and processing related resources. The module responsible for resource management is the operating system. Unlike the modules discussed in this thesis, the OS module is exclusively software based.

Operating system battles are being waged today between the likes of Nokia's Symbian OS, Windows Mobile, Apple OS, Research in Motion, and now Google's Android OS. The battle lines being drawn resemble those of the closed versus open operating systems for computer operating systems whereby firms have differentiated their OS modules by promoting different levels of customization. In comparison with desktop computer bottlenecks, handset technology is unique in that the data rate and battery life bottlenecks that were subjugated in computer designs are thrust into the spotlight. It stands to reason that the importance of these bottlenecks will cause a significant divergence between computer and handset OSs. Assuming a software defined radio chip is available, the handset ODMs could achieve high ROIC by way of differentiated performance and low electronic bill-of-materials provided the OS enables the chip to efficiently cull information from various frequency bands and communication standards. All the while, the OS should be intelligent enough to minimize power consumption. The later is a feature that is all too familiar to handset ODMs and therefore shouldn't enable differentiated performance. Even prior to the emergence of software defined radio as the status quo, firms have been embracing single chip solutions to the physical layer bottlenecks manufactured by Texas Instruments, Qualcomm, ST Microelectronics, and Infineon. Per the analysis in this report, the intent sans truly cognitive radio availability is to realize high ROIC by eliminating manufacturing cost.

The dominant strategy followed by component suppliers has been well documented in this thesis. Whether by aggressive forward integration on the part of the suppliers or willing abrogation on the part of the handset ODMs, then end result is the same. Possession of physical layer designs has passed from the handset ODMs to the component suppliers. Handset ODM's maintain some power in the integrated modular designs that are often associated with new standards and high performance systems; but for the most part the bottlenecks at the physical layer are being addressed by the component suppliers. Two questions remain. Is additional integration possible? Can the firms that manufacture these system-on-a-chip solutions realize modularity related option value, or is it forever lost?

From the analysis in this thesis, the effort required to change the architecture must be a response to an economic value proposition. In the migration of the physical layer design, a modified version of the Baldwin and Clark modular operators was identified as a key economic driver in the emergence of the integrated modular architecture. To the extent that handset designs employ many integrated circuits, it is reasonable to believe that the economic stimulus responsible for integrated modular transceiver designs and digital radio designs will also apply to full integrated modular systems. Figure 7.3 shows an approximation to the cost contributions of components to the overall BOM for the N95 [37]. Clearly from the chart, the electronic module cost still dominates the overall BOM. Therefore, it stands to reason that if the handset ODMs continue to migrate towards the user experience bottleneck, as opposed to the physical component level bottlenecks, then ROIC will be maximized by lower cost solutions to the electronic components. In other words, the suppliers will continue to be rewarded for higher levels of integration of the electronic modules.



7-3: Nokia N95 Bill of Materials.

The resources required to design and manufacture large integrated systems are enormous. From a strictly manufacturing point of view, the capital required for fabrication equipment coupled with a competitive market will necessarily trim the market to the most operationally fit firms. Design on the other hand could lend itself to a distributed model wherein firms cognizant of the design rules inherited by the digital manufacturing technology could compete in the market for implementation specific bottlenecks within the domains of their expertise. What's more, in a distributed design environment where manufacturing is outsourced, a new financial metric may supplant ROIC as the determinant of growth of the design firms; namely return on human capital.

8 References

[1] Baldwin, Carliss. Clark, Kim. Design Rules-The Power of Modularity, Volume 1. MIT Press. 2000.

[2] Baldwin, Clark. "Architectural Innovation and Dynamic Competition: The Smaller Footprint Strategy. Working Paper." Ver 1.0. August 2006.

[3] Micro Economics for Business Decisions. Course Notes. Fall 2007.

[4] Snider, J.H. "THE ART OF SPECTRUM LOBBYING, America's $480 Billion Spectrum Giveaway, How it Happened, and How to Prevent it From Recurring" New America Foundation. 08/2007. http://www.newamerica.net/files/art_of_spectrum_lobbying.pdf.

[5] Lasar, Matthew. "Verizon, AT&T rule 700MHz auction; Block D fate unsettled" 03/20/2008. http://arstechnica.com/news.ars/post/20080320-verizon-att-rule-700mhz-auction-block-d-fate-unsettled.html. 09/05/2008.

[6] American Telephone and Telegraph. "AT&T Offers Nation's Fastest 3G Network." 07/10/2008. http://www.att.com/gen/press-room?pid=4800&cdvn=news&newsarticleid=25921. 09/05/2008.

[7] Marketing Management Class Notes. MIT Sloan School of Management. Fall 2006.

[8] American Telephone and Telegraph. 2007 Annual Report.

[9] Research in Motion. 2007 Annual Report.

[10] Hoffman, Baxter, and Prohm. "Wireless Equipment Industry Outlook" Cowan and Company. July 1, 2008.

[11] Brealey, Myers, Allen. Principles of Corporate Finance. McGraw Hill-Irwin. 2006.

[12] http://en.wikipedia.org/wiki/Variance. 09/08/2008.

[13] Thomke, Stefan. "Managing Experimentation in the Design of New Products" Management Science. Vol 44. No 6. June 1998.

[14] Pindyck and Rubinfeld. Microeconomics, 6th Edition. Pearson Prentice Hall. 2005.

[15] Higgins, Robert C. Analysis for Financial Management McGraw Hill Irwin. 2007.

[16] Gallager, Robert. Principles of Digital Communication. Cambridge University Press. 2008.

[17] Couch, Leon. Digital and Analog Communication Systems 4th Ed. Prentice Hall. 1997.

[18] Proakis, John. Digital Communications 4th Ed. McGraw Hill 2001.

[19] Turin, George. "An Introduction to Matched Filters." IRE Transactions on Information Theory. 1960.

[20] Goldsmith, Andrea. Wireless Communications. Cambridge University Press. 2005.

[21] Harte, Lawrence. Introduction to Global System for Network Communications (CDMA)-Physical Channels, Logical Channels, Network, and Operation. Althos Publishing. 2005.

[22] Ericsson. EDGE: Introduction to high-speed data in GSM/GPRS networks. 2005.

[23] European Telecommunications Standards Institute (ETSI). Digital cellular telecommunications system (Phase 2+);Digital Radio transmission and reception (GSM 05.05 version 8.5.1 Release 1999).

[24] Harte, Lawrence. Introduction to Code Division Multiple Access (CDMA)-Network, Services, Technologies and Operation. Althos Publishing. 2004.

[25] http://en.wikipedia.org/wiki/Cdma. 09/15/2008.

[26] http://en.wikipedia.org/wiki/Code_division_multiple_access. 09/16/2008.

[27] http://www.cdg.org/technology/cdma_technology/a_ross/cdmarevolution.asp. 09/16/2008.

[28] http://www.cdg.org/technology/cdma_technology/a_ross/Reverse.asp. 09/16/2008.

[29] http://en.wikipedia.org/wiki/Evolution-Data_Optimized. 09/16/2008.

[30] http://en.wikipedia.org/wiki/Superheterodyne_receiver. 09/24/2008

[31] Portelligent. "Motorola DynaTAC "Brick Phone"Q uick Turn Retro-Teardown" May 2006.

[32] Lee, Thomas. The Design of CMOS Radio-Frequency Integrated Circuits. Cambridge University Press. 2001

[33] http://www.rfcafe.com/references/electrical/ip3.htm. 09/22/2008

[34] Sze, S.M. Physics of Semiconductor Devices. John Wiley and Sons. 1981.

[35] Rudell, Weldon, Ou, Lin, Gray. "An Integrated GSM/DECT Receiver: Design Specifications" UCB Electronics Research Lab Memo. Memo #: UCB/ERL M97/82. April 28, 1998.

[36] Lee, Edward. "Programmable DSP Architectures: Part 1." IEEE ASSP Magazine. October 1988.

[37] Portelligent. "Nokia N95 GSM/EDGE 850/900/1800/1900MHz + WCDMA/HSDPA 2100MHz UMTS Cellular Phone" Report #11807-070416-TMf. 2007

[38] Gotz, et al. "A Quad-Band Low Power Single Chip Direct Conversion CMOS Transceiver with CA-Modulation Loop for GSM." European Solid State Circuits Conference. 2003.

[39] Skyworks Solutions. SKY77318 Power Amplifier Module Data Sheet.

[40] Niwa, Ishigaki, Shimawaki, Nashimoto. "A Composite-Collector InGaP/ GaAs HBT with High Ruggedness for GSM Power Amplifiers" IEEE International Microwave Symposium. 2003.

[41] Discussion with Erik Org of Bitwave Semiconductor

[42] Hartley, R. "Modulation System" U.S. Patent 1 666 206, Apr. 1928.

[43] Weaver. "Design of RC Wide-Band 90-Degree Phase-Difference Network." Proceedings of the IRE. Vol 42. Issue 4. 1954.

[44] Martin, Kenneth. "Complex Signal Analysis is Not Complex" IEEE Trans. Circuits and Systems. Vol. 51. No 9. Sept. 2004.

[45] Crols and Steyaert. "Low-IF Topologies for High-Performance Analog Front Ends of Fully Integrated Receivers" IEEE Trans. Circuits and Systems. Vol. 43. No 3. Sept. 1998.

[46] Razavi. "RF CMOS Transceivers for Cellular Telephony" IEEE Communications Magazine. August 2003.

[47] Buss, Dennis (Texas Instruments, MIT) "Technology Dependant Roadmap for Ubiquitous Computing." Presentation given to University of Michigan. Date unknown.

[48] Mitola, Joe. "The Software Radio Architecture." IEEE Comm. Mag. May 1995.

[49] Bagheri, et al. "An 800-MHz-GHz Software-Defined Wireless Receiver in 90-nm CMOS." IEEE JSSC. Vol. 41, No. 12, December 2006.

[50] Henderson, R., & Clark, K. B. (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. Administrative Science Quarterly, 35, 9-30.