A Portable Multimedia Terminal

Successful personal communications terminals will depend upon the smooth integration of computation and communications facilities in a lightweight unit.

Samuel Sheng, Anantha Chandrakasan, and R.W. Brodersen

The personal communications industry has seen explosive growth in the past several years, especially in the number and types of services and technologies. In voiceband communications, systems such as mobile analog cellular telephony, radio pagers, and cordless telephones have become commonplace, despite their limited nature and sometimes poor quality of transmission. In computing, portable "notebook" computers are boasting capabilities far in excess of the desktop machines of five years ago, and multi-MIPS, RISC-based, portable workstations are available. Despite the myriad technologies to be had, however, little integration of these diverse services — the combination of computation and communications facilities in a portable unit — has occurred. Thus, our vision of a future personal communications system (PCS) centers on such integration of services to provide ubiquitous access to data and communications via a specialized, wireless multimedia terminal.

A schematic view of such a system is shown in Fig. 1. The wireless terminal is in full duplex communication with a networked base station, which serves as a gateway between the wired and wireless mediums. Via a base station, users access services over the high-speed communications backbone, including communicating with another person also linked into the network. This idea can also be extended to a user communicating not only with another person, but with network "servers." Because the data bandwidth of future fiber-optic networks is easily in excess of 10 Gb/s, these centralized servers can provide a wide variety of information services to users. A personal communications system will likely include the four key features that follow.

Access to large commercial databases that contain information such as international and domestic news, financial information, traffic data, transportation schedules, voice mail, telephone numbers, news, bulletin boards, and educational material is necessary. The continuous connectivity afforded by personal communications systems has several advantages. Many sources of information are of a transitory nature, such as stock pricing, local news, and so on, making distribution by other means such as CD-ROM impractical. Furthermore, given sufficiently large database servers, libraries of books, journal archives, and other currently "paper-intensive" media can be placed on-line; these databases would allow for instantaneous recovery of all types of information, without the need to be at a terminal physically attached to the wired network.

Second, a PCS would have access to digital video databases containing both entertainment and educational media, such as animated information sequences, taped lectures, movies, news clips, and other isochronous data. Unlike today's television broadcasts, video databases can be made available on-demand, giving users the freedom to access video information as needed. Video data will be necessarily stored in a compressed format for minimization of both storage space and transmission bandwidth, thus requiring that the wireless terminals at least support video decompression.

Simplified entry mechanisms such as voice-recognition and handwriting-recognition interfacing to access the above functions would also be available. The design of an effective user interface to access such a vast information storehouse is a critical issue. By using speech recognition and pen-based input, supported by large, speaker-independent recognizers placed on the network, such interfacing and information access can be tremendously simplified. Placing the recognition units on the network conserves power in portable units and enables much larger and more complex recognition algorithms to be employed. Recognizer servers can also make use of context-sensitive analysis, which can increase recognition accuracy by determining which words are most likely to be used in a given application [1].

Fourth, the system would provide support for a distributed computing environment, such as MIT's X-Window system. In distributed computing environments, computation need not take place on a local machine; instead, computation is performed by programs executing on one or more remote machines, which may have no computing capability except that required to act as an intelligent display device. Many such inexpensive "X-terminals" already exist. Unlike

ROBERT W. BRODERSEN is a member of the Electrical Engineering and Computer Science faculty of the University of California, Berkeley.

ANANTHA P. CHANDRAKASAN and SAMUEL SHENG are Ph.D. candidates in the Electrical Engineering and Computer Science department at the University of California, Berkeley.

0163-6804/92/$03.00 1992© IEEE

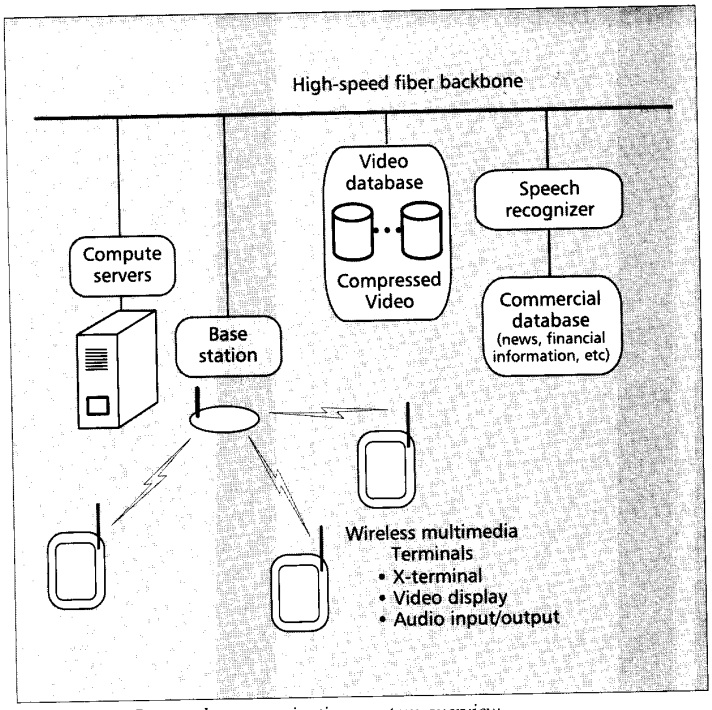
TTY terminals that can only communicate with a single host machine, X-terminals possess all of the necessary networking capability to communicate with as many remote servers as needed. The multimedia terminal will be based on this model: remote computation servers will be used to run applications like spreadsheets, word processors, etc., with the results being transmitted to the terminal. Likewise, supercomputer-class servers will perform intensive tasks requiring simulations, 3-D image rendering, and computer-aided design.

Clearly, the cornerstone of the entire system lies in the ability of multimedia terminals and wireless communications links to support all of the aforementioned services. Correspondingly, it is this desire for portability that translates directly into design constraints on the size and weight of the terminals, the power they consume, and the frequency bandwidth needed in the wireless links.

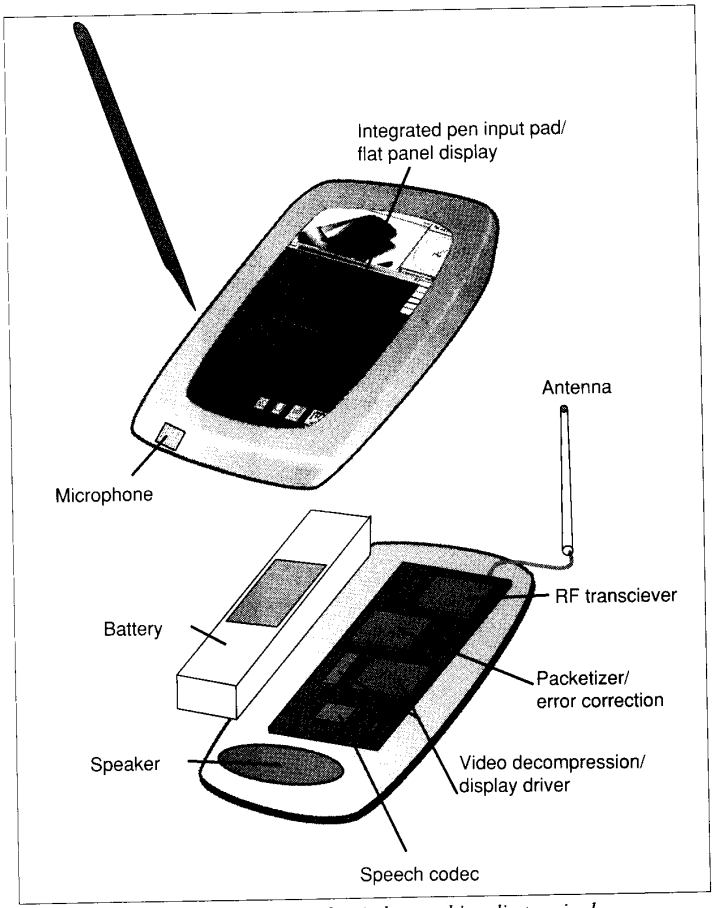
A diagram of a portable terminal is shown in Fig. 2. To minimize power, only those functions that are absolutely necessary are implemented: the analog RF transceiver; baseband processing for communications, such as equalization, coding, and packetization; the image decompression unit and a display driver; and the speech codec. Because size is also an issue, a pen input system is integrated directly onto a compact flat-panel display, eliminating the need for a large keyboard and providing greater visual feedback than possible with mouse- or trackball-based interfaces. More important, the system is asymmetric in nature: High-quality, full-motion video is only supported in the downlink from the base station to the portable. This must be accounted for the in the design, as the bandwidth requirements in the reverse link from the portable unit are thus considerably less than the link from the base station. For video teleconferencing, a low-rate, reduced-quality video uplink might also be supported; however, the asymmetric bandwidth requirements will still remain.

The diagram shows that no direct user computation is supported within the portable itself; instead, it is wholly dependent on the network servers to provide desired functionality. Although this has immediate benefits in terms of reducing power consumption, it provides another advantage: Data that is highly sensitive to corruption will not be transmitted over the wireless network.

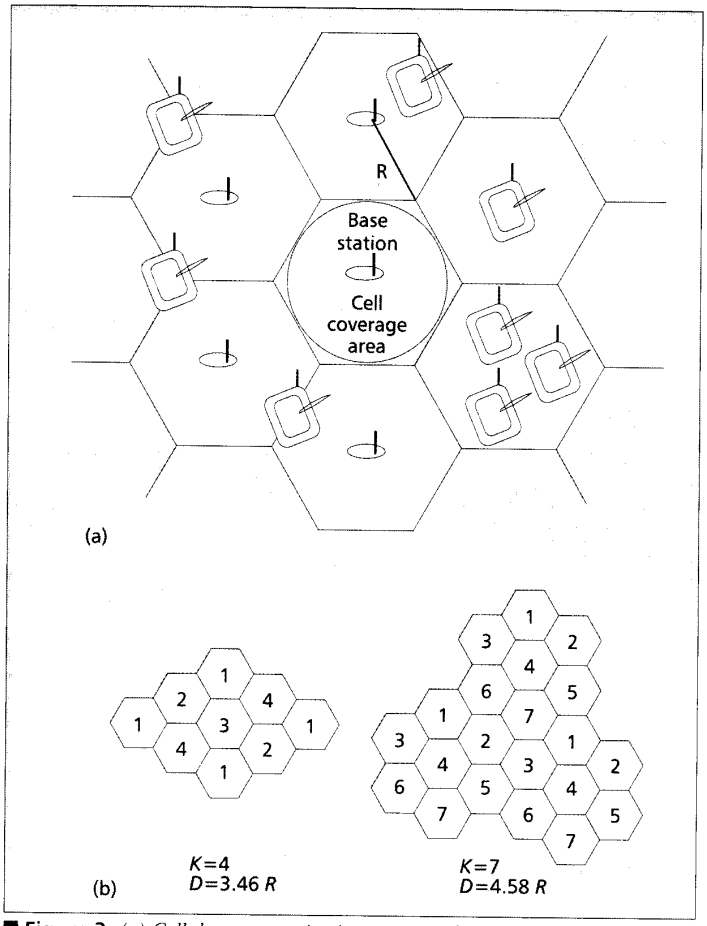
Existing distributed computation environments are dependent on the fact that data transmitted over the network has high integrity — bit-error rates on wired Ethernet are typically on the order of 1 in 10^{12} bits, and further protection is gained by packet retransmission after errors. On wireless networks, however, this is not true. Even after extensive application of error-correction coding, it is still difficult to attain error rates even remotely as low as this. User "computation" data, such as spreadsheets or simulation results, simply cannot be allowed to sustain any corruption. For wireless systems, this translates into an inordinate amount of transmission overhead in terms of coding and data retransmission. On the other hand, user "multimedia" information, such as voice and image data, is relatively tolerant of bit errors — an error in a single video frame or an audio sample will not significantly change the meaning or usefulness of the data. The ability to coexist with an error-prone transmission environment has tremendous impact on the overall system design.



■ **Figure 1.** *Personal communications system overview.*



■ **Figure 2.** *Diagram of the portable wireless multimedia terminal.*



■ **Figure 3.** *(a) Cellular communications system; (b) typical frequency reuse patterns.*

Thus, the portable unit described above is truly a terminal dedicated to multimedia personal communications, and not simply a notebook computer with a wireless LAN/modem attached to it.

The remainder of this article will focus on several of the major design issues behind portable multimedia terminals: spectrally efficient picocellular networking, low-power digital design, video data compression, and integrated wireless RF transceivers. Optimizing performance in each of these areas is crucial in meeting the performance requirements of the overall system and providing a small, lightweight terminal for personal communications.

Picocell Networking and Spectrum Usage

Because full-motion digital video is to be supported, spectrum usage is of great concern — per-user data rates can easily exceed 2 Mb/s even with the best compression schemes reported to date. This data rate is not needed on a continuous basis; regular computation tasks such as word processing or use of a spreadsheet require only slight screen changes on a frame-by-frame basis over a small region, usually on the order of a single character or a few pixels. Hence, it is probable that the peak data rate required by users will be much larger than the overall time-average data rate. Minimizing overall system bandwidth consumption while supporting a large number of users accessing data simultaneously is paramount.

One method to achieve this goal is to physically reduce peak user data rates via data compression techniques (we will discuss this later). Another technique, applied at the system level, is to utilize cellular networking techniques to achieve spatial frequency reuse. Because such a personal communications system will first be used as a step beyond conventional wireless LANs, an indoor picocellular transmission environment will be of primary concern. (The techniques described in this article are applicable to both indoor and outdoor environments.)

The advantages in improved spectral efficiency afforded by cellular systems have been employed extensively in present-day analog mobile radiotelephony, where large-scale cells exploit these advantages to a limited extent. By scaling down cell sizes, tremendous increases in spectral efficiency can be achieved. A simple cellular scheme, as shown in Fig. 3a, consists of dividing the entire service area for the personal communication system into "cells" of radius R, with a single base station serving all mobile users within that cell. Each cell uses its own distinct set of frequencies. As users move from cell to cell, their transactions with the network are "handed off" from base station to base station, reconfiguring the network dynamically as the need arises.

The key benefit of cellular systems is that they allow the network to achieve spatial multiple access. If two cells are separated by sufficient distance, each can use the same frequency bands at the same time without resulting in disastrous cochannel interference. Thus, extensive frequency reuse becomes possible, as opposed to an umbrella scheme where every user must be assigned a different frequency slot. Fig. 3b shows several classical reuse patterns [2]; such patterns are typically characterized by a frequency reuse factor K, which represents the number of distinct frequency sets that need to be used to cover the entire service area. Instead of one user per frequency band, the network can now support N users per band, where N is the number of cells in the service area using that band. From the point of view of spectrum usage, each user effectively consumes only B/N Hz of bandwidth, where B is the physical bandwidth needed to support transmission, thus drastically increasing overall spectral efficiency.

Clearly, minimizing the physical distance D between cells using the same frequency, by reducing the cell size R, yields the greatest frequency reuse, and hence the greatest gains in efficiency.[1] Therefore, it is clear that the number of users supportable within the same overall system bandwidth increases quadratically as R decreases, because of the increased number of cellular subdivisions within the service area. Minimizing R (and hence D) is critical in achieving high levels of spectral efficiency.

With an indoor environment, it is no longer feasible to have only a single network transceiver station serving all of the terminals in the building. Due to the 5 to 15 dB attenuation through walls, the total microwave output power from all of the transmitters would have to be inordinately (and dangerously) high [3]. However, this attenuation can be taken advantage of by a cellular network — each room naturally becomes its own cell. Likewise, the cellular

[1] *The frequency reuse distance is geometrically related to K and R by $D = R\sqrt{3K}$*

scheme now moves into three dimensions because the floors also provide RF isolation. Even if walls are not present, the use of electrically-adjustable directional antennas (such as a small phased-array device) can provide the same effect. The cells are now extremely small, on the order of about five meters; R is usually dictated by the size of the room, and K can be as low as 3 to 4, depending on how much attenuation is provided by the walls. If K is increased to 6 or 7, the assumption that cochannel interference is negligible becomes reasonable for most indoor office environments.

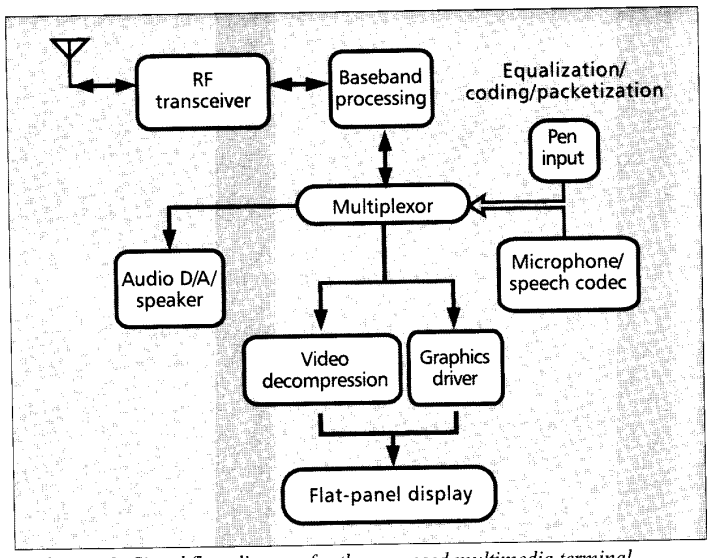
In light of the above considerations, the total amount of spectrum that will be consumed to provide the outlined services can now be addressed. After examination of the user density in a typical office environment,[2] such as those found in modern buildings with open-area soft-partition cubicles, cells with a radius of five meters typically contain 12 to 16 active users. In the worst-case, a 2 Mb/s data rate for full-motion video using a linear DQPSK (differential quadrature phase-shift keying) modulation scheme and design parameters from existing systems [4-5], would require a transmission bandwidth of approximately 1.5 MHz per user using a 20 percent excess bandwidth raised-cosine pulse shape and a 25 percent loss for packetization, equalizer training, and other overhead. Assuming that half of the 16 users in the cell demand the complete 2 Mb/s data rate for full-motion digital video, and the remainder need 256 kb/s each[3] for lower data rate applications, a picocellular system with $K = 7$ would use approximately 100 MHz of bandwidth. Although 100 MHz is a considerable amount of spectrum, this is amortized over large numbers of people using this spectrum simultaneously within multiple buildings. Because the bandwidth of 100 MHz is designed to support full motion video and other multimedia network services for all users, this allocation of spectrum is not unreasonable, given the level of service provided by the system, especially when compared to the spectrum allocated for existing systems such as NTSC television.

There is another significant advantage to picocellular wireless systems: because transmit power is scaled down as cells move closer together to reduce interference, the power consumed in the portable's transmitter to drive the antenna is correspondingly reduced. Whereas existing cellular systems use 1 watt of transmit power for voiceband RF links in 5 mile cells, a picocellular system with 5 meter cells requires only milliwatts to maintain the link [6].

Implementing Portable Terminals

Picocellular networking ameliorates several important issues in providing portable multimedia-based communications systems. Many challenges remain, however, in building the required functionality into terminal hardware.

A signal flow diagram for a terminal is shown in Fig. 4. Incoming data can be one of three types: digital video, screen graphics, or sampled audio. Outgoing data can be either voice or pen input. The asymmetry in the data rates of the uplink and the downlink is clear; no high-rate signals are intended for transmission from the mobile to the base station. The hardware design must also reflect this asymmetry:



■ **Figure 4.** *Signal flow diagram for the proposed multimedia terminal.*

within the analog RF block, the receiver design becomes critical, because it must demodulate a high-rate signal corrupted by noise and distortion, whereas the transmitter is relatively simple, with low data rate and output power requirements. Likewise, the algorithm chosen for the video decompression should be designed, if possible, to make decompression as simple as possible and with little consideration for compression complexity, because compression can be performed by one of the network servers.

One key consideration is how long the portable can function between battery rechargings. Ideally, it should be able to operate for one work day, or eight to ten hours of battery life. Given that conventional batteries typically possess 20 watt-hours for each pound of battery weight, and a limit of one pound of batteries in the portable, the entire portable can consume no more than 2 watts of power. Furthermore, projections of progress in battery technologies show that only a 20 percent improvement in battery capacity will occur over the next ten years. Thus, power minimization becomes a serious concern.

The largest power consumer in current portables is the backlighting of flat-panel displays. As display technologies improve screen contrast, however, this requirement will be significantly relaxed, implying that low-power techniques for implementing the analog and digital core circuitry are needed.

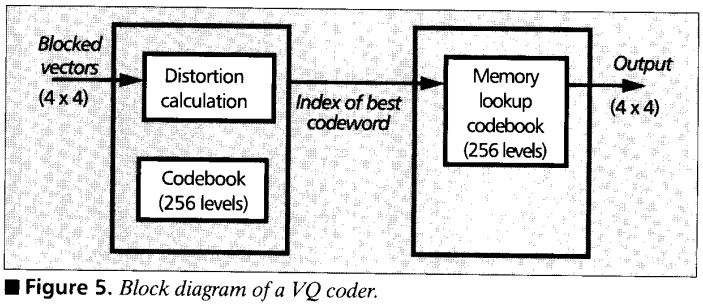
With present-day technology — single-chip packaging using printed circuit boards for interconnection — one-third or more of the total power is consumed by the chip's input/output (I/O) because the capacitances at the chip boundaries are usually much larger than the capacitances inside the chip. Typical values range from a few 10s of femtofarads at internal chip nodes, up to 10s of picofarads at the chip interface attributed to pad capacitances and board traces. To reduce the power consumed in the I/O, low-capacitance, high-density interconnect methods will be employed, such as the emerging multichip module (MCM) technologies. MCM integrates many individual chip die into a single structure, reducing the size of interchip capacitances to the same order-of-mag-

[2] *For example, the EECS graduate student research facility at the University of California, Berkeley.*

[3] *256 kb/s is typical of the peak data rate afforded by most wireless LAN systems, and is reflective of data rates used in existing X-terminals; this value may be considerably lower depending on the level of user activity.*

DCT Algorithm	Multiplies	Additions
Brute force	4096	4096
Row-col DCT	1024	1024
Chen's algorithm	256	416
Lee's algorithm	192	464
Feig's algorithm	54	462

■ **Table 1.** *Complexity of various DCT algorithms (8 x 8 blocks).*



■ **Figure 5.** *Block diagram of a VQ coder.*

nitude as on-chip capacitances and minimizing the power consumed in the I/O drivers. Thus, with MCM, the majority of the power is consumed within the functional core of the chip itself, as opposed to the interface. Also, because the packing density has increased, and with the ever-decreasing size of CMOS circuitry (down to 0.2 μm line widths), over 10^{10} transistors can be placed within a single eight-by-eleven-inch MCM substrate. Area constraints imposed by available silicon are no longer of great issue, allowing for greater possibilities in power optimization, as we will discuss later.

For analog RF transceivers, however, there are other design considerations beyond low-power implementation. Due to size considerations, traditional discrete element design is not feasible for a small, portable unit such as the proposed multimedia terminal; single-chip integration techniques that exploit advances in silicon CMOS (as opposed to gallium arsenide, GaAs) must be explored, to address cost and manufacturability concerns. Likewise, the fact that digital circuitry is readily available on-chip also opens up new possibilities: analog performance requirements can be reduced at the expense of increased digital signal processing.

To examine these implementation details more fully, we will discuss three distinct design issues. The first is image compression, both as a means of spectrum reduction and as an example of how the choice of algorithms can take advantage of the aforementioned asymmetry to reduce power consumption. Second, we will show how low-power digital system design leads to large reductions in power consumption. Third, we will analyze the design of analog RF transceiver to exploit monolithic integration techniques and the underlying digital nature of the transmitted signal.

Image Compression

As stated above, frequency reuse is only one means of reducing spectrum consumption. Concomitant with frequency reuse is the idea of reducing the amount of physical bandwidth needed by each user, and hence reducing the bandwidth consumed by the overall system. Because high-resolution, full-motion video is to be supported, transmission of 640-by-480 pixel images, digitized at 24 bits/pixel, would require a bandwidth of 220 Mb/s in an uncompressed format at 30 frames per second. Thus, it is clear that video compression techniques are crucial in making wireless video transmission feasible.

The video module performs the decoding and display interface functions and converts a compressed data stream to an uncompressed video stream that is displayed on the LCD display. The decompression module can be implemented using a variety of algorithms, such as transform-based schemes, vector quantization and subband coding. The selection of the algorithm for the portable terminal depends not only on the traditional criteria of achievable compression ratio and the quality of reconstructed images, but also on computational complexity (and hence power) and robustness to higher bit error rates. The choice of an algorithm to implement the decompression function is the most important in meeting the power constraints. The basic complexity of the computation must be optimized and, as shown in the next section, the ability to parallelize an algorithm will be critical.

Most current compression standards (for example, JPEG and MPEG) are based upon the Discrete Cosine Transform (DCT). The basic idea in intraframe schemes such as JPEG is to apply a two-dimensional DCT on a blocked image (typically eight pixels by eight pixels) followed by quantization to remove correlations within a given frame. In the transform domain, most of the image energy is packed into only a few of the resulting coefficients, and compression is achieved by transmitting only a carefully chosen subset of the coefficients. While these standards specify the use of DCT as the transformation to be applied, they do not specify the algorithm to be used. Table 1 shows a comparison of a few algorithms that can be used to implement the DCT [7-8]. Minimizing the operation count is important in minimizing the switching events and the hence the power consumption.

A primary characteristic of the DCT is the symmetric nature of the computation; that is, the coder and decoder have equal computational complexity. However, an alternative compression scheme is that of vector quantization (VQ) coding, which is asymmetrical in nature and has been unpopular due to its complex coder requirements. The basic idea behind VQ coding is to group the image data into a vector and quantize it. Fig. 5 shows a block diagram of a VQ coder. On the coder side, the input is first blocked into a vector (for example, a four-by-four block of video is 16 bytes). This vector is compared to the entries in the codebook with the goal of minimizing the expected error or distortion between the input vector and its reproduction for a given bit rate. The codeword corresponding to the closest match (or the match that minimizes the distortion) is transmitted (in this case, the index is one byte long and 16:1 compression is achieved). On the decoder side, a simple memory look-up is used to reproduce the image data. The design of the codebook and distortion measure have been extensively discussed in literature. Clearly, the distortion computation is much more computationally intensive than the decoder, as the entire decoder

is nothing more than a lookup table; however, algorithms have been proposed that reduce the computational complexity in the coder with little loss in reconstructed image quality [9]. Table 2 shows the requirements for two strategies.

The simple memory-lookup decoder of VQ algorithms makes them well-suited for single-encoder, multiple-decoder systems (in which video is compressed and stored once, but is accessed multiple times by multiple users). If one-way video communication is desired, the VQ solution provides a means of implementing real-time decompression requiring little computation and power. For portable terminals, supporting one-way communication in which a user accesses various databases on the wired network is a very important feature.

The algorithms described above are based on removing redundancy inside a frame (intraframe), and are quite robust against higher bit-error rates. For a typical channel BER of 10^{-6}, this translates to an average of only a couple of local regions per second having corrupted image data. This magnitude of error is acceptable for video as human perception is not sensitive to such small, local errors. Likewise, the errors do not accumulate, because the next frame will be fully transmitted with its own local error independent of the previous frame. On the other hand, interframe schemes remove temporal redundancy (i.e., correlation between frames) and are not as robust against bit errors. The basic idea in these schemes is to apply DCT or VQ on the difference of the current image and the previous image. Since only the difference information is transmitted, a bit error can cause many regions to be corrupted. Also, errors accumulate in the case of interframe coding, making it less desirable; by "resetting" the errors every so often by transmitting a complete frame, however, such differential error propagation can be alleviated. However, it is clear that the "best" algorithm for a wireless environment will likely be quite different than one chosen for the low-BER situation assumed by the present compression standards.

Low Power Systems Design

A massive amount of computing resources will be required to achieve the performance desired of portable multimedia terminals—each unit needs to support both complex modulation schemes and sophisticated data decompression algorithms. Likewise, the power needed to drive these functions can easily be prohibitive for portable operation. Total power consumption must be minimized and the required throughput of the overall system must be maintained. However, because processing is bounded by real-time constraints, as in image decompression and channel equalization, once the throughput performance is met there is no advantage is making computation any faster, opening up a major degree of freedom to the designer. This section comprises an overview of techniques for minimizing the power consumption in digital CMOS circuits.

Sources of Power Dissipation

The main contribution to power consumption in CMOS circuits is attributed to the charging and discharging of parasitic capacitors during logical transitions. The average switching energy of a CMOS gate (or the power-delay product) is equal to $C_{avg} \cdot V_{dd}^2$, where C_{avg} is the average capacitance being switched per clock and V_{dd} is the supply voltage. Average power consumption is given by the product of the energy per transition and the frequency of operation, f_{clk}. Power is also consumed by short circuit currents (which arise from a direct current path from supply to ground during switching) and by subthreshold leakage currents (parasitic currents that flow when the transistors are supposed to be turned off). The total power dissipation in CMOS circuits is summarized in the following equation:

$$P_{total} = C_{avg} \cdot V_{dd}^2 \cdot f_{clk} + I_{sc} \cdot V_{dd} + I_{leakage} \cdot V_{dd} \quad \text{Eq. 1}$$

By careful circuit and technology design, the short circuit and leakage components can be minimized and the goal of low power design becomes the minimizing of C_{avg}, V_{dd}, and f_{clk}, while retaining the required functionality.

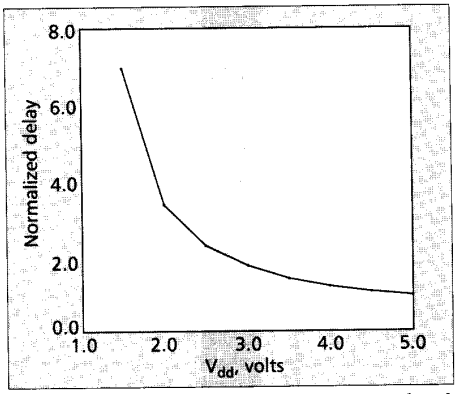
The quadratic dependence of energy on voltage has been experimentally verified for various circuits, and it is clear that operating at the lowest possible voltage is most desirable for optimizing the energy per computation. Unfortunately, reducing the supply voltage comes at the cost of a reduction in the computational throughput. This is seen from Fig. 6, a plot of normalized delay vs. V_{dd} for a typical 2μm CMOS gate, with the gate delays increasing as the supply voltage is dropped. Even though the exact analysis of the delay is quite complex, it is found that a simple first-order derivation adequately predicts the experimentally determined dependence quite well and is given by:

$$T_d = \frac{C_L \cdot V_{dd}}{I} = \frac{C_L \cdot V_{dd}}{K_p (W/L)(V_{dd} - V_t)^2} \quad \text{Eq. 2}$$

where C_L is the load capacitance, I is the current drive, K_p is a process dependent parameter, W/L

Algorithm	Coder per pixel	Decoder per 16 pixels
Full-search	256 subtract, multiply, add 256 memory access	1 memory access
Tree search with a differential codebook	8 multiply, add 8 memory access	1 memory access

■ **Table 2.** *Computational complexity of VQ algorithms for a vector size of 16 (4 x 4).*



■ **Figure 6.** *Plot of normalized delay vs. supply voltage at* V_{dd} *(delay at* V_{dd} *= 5V is normalized to 1).*

Architecture optimization will play a critical role in optimizing power dissipation by enabling low-voltage operation.

is the size or the strength of a transistor, and V_t is the threshold voltage (the gate-to-source voltage at which the transistor "starts" to conduct current) [16]. It is clear from the above equation that the delays increase drastically when the supply voltage approaches the threshold voltage. Given that it is desirable to run at low supply voltages for energy optimization and the requirement that the real-time throughput constraints be met, techniques for compensating increased delays at low voltage must be considered.

Technology and Circuit Considerations

The reduction in speed due to lowering supply voltage can be compensated to some extent by scaling the feature size, the smallest geometry that characterizes the technology, into the submicron regime [6]. As the feature size scales into the submicron region (< 1μm), the velocity of electrons in the transistor saturates at high supply voltages (or at high electric fields), with the gate delays becoming relatively independent of the supply voltage. As a result, the supply voltage can be dropped to some extent without a significant reduction in speed. This fact has been exploited to reduce the supply voltage to 3.3V in several emerging low-voltage applications. For example, this approach was found to achieve a 60 percent reduction in power when compared to a 5V operation [10]. While this approach allows a reduction of voltage to some extent, it is clearly much more advantageous to run at much lower supply voltages to minimize the energy per computation.

The fabrication technology itself can also be optimized for low-power design, in particular by adjusting the threshold voltage V_t of the MOS devices. Because the gate delays increase as V_{dd} approaches the threshold voltage and the fact that the power consumed is only weakly dependent on V_t, the threshold voltage should be made as small as possible. To date, scaled MOS circuits have not exploited reduced supply voltages mainly because designers have been focused on fabricating the fastest possible circuits. However, the threshold voltage cannot be lowered to zero; subthreshold leakage currents become increasingly significant as V_t is reduced. First-order MOS theory assumes that the drain current goes to zero for a gate to source voltage V_{gs} less than or equal to V_t. In reality, at $V_{gs} = V_t$, the drain current is not zero, and tapers off exponentially as V_{gs} goes further below V_t, yielding subthreshold leakage conduction. This leakage can represent a significant component of the overall power consumption is V_t if made too low. The question then becomes what threshold is optimal for low power design. By considering the trade-offs between the switching component of energy and the leakage component of energy, a threshold voltage in the range of 0.3V to 0.5V was found to be optimal. [15]

Several options are available in choosing and optimizing the basic circuit approach and topology for implementing various logic and arithmetic functions. For example, a pass-transistor logic family was found to be more efficient in terms of the number of transistors (and hence capacitance) required to implement various functions when compared to a conventional CMOS logic family [11]. Reducing the average activity (i.e., the average number of transitions per clock cycle) can also result in a dramatic reduction of the switching energy. For example, it is important to minimize switching activity by powering down execution units when they are not performing important operations. While the design of synchronous circuits requires special design effort and power-down circuitry to detect and shut down unused units, self-timed asynchronous logic has an inherent capability to power-down unused modules, because transitions occur only when requested. Optimizing transistor sizing is another method for optimizing the energy per computation. To minimize the parasitic capacitances, it is desirable to use the smallest devices (the smallest possible transistor size) as much as possible, which also assists in minimizing the interconnect routing lengths and their associated parasitic capacitances.

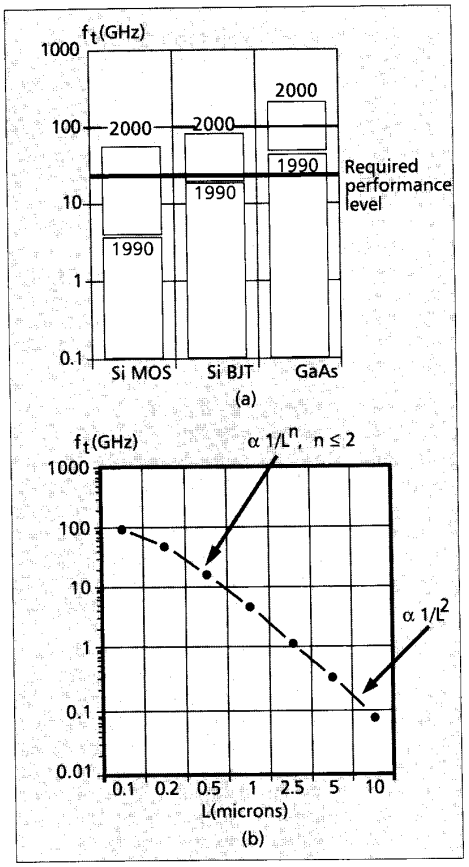
Architectural Optimization

Architecture optimization will play a critical role in optimizing power dissipation by enabling low-voltage operation. One major consideration in optimizing for dedicated signal-processing functions such as video decompression and channel equalization, as opposed to general purpose computing, is that once real-time operation is achieved, there is no advantage to making the computation any faster. Combined with the increasing density of VLSI systems because of feature size scaling and high-density packaging, an architectural strategy has been developed that can be used to trade off area and power for a fixed throughput.

One way to compensate for increased delays at low supply voltages is to use architectures that reduce the speed requirements of operations while keeping throughput constant. Parallel hardware duplication can be used to lower the supply voltage for a fixed throughput. By using parallel and identical units, the speed requirements on each unit are reduced, allowing for a reduction in voltage. For example, consider a datapath module running at a frequency of f while switching an average capacitance C at a supply voltage of 5V. Parallelizing by a factor of two will result in two units working at $f/2$ while maintaining the original throughput. Because the speed requirements on each module have lowered by factor of two, the voltage can be lowered until the gate delays increase by a factor of two (this corresponds to a voltage of 2.9V, as shown in Fig. 6). Power then becomes equal to $(2C) \cdot (2.9)^2 \cdot f/2$, which is approximately 3 times lower than the original implementation with one unit. Likewise, pipelining is a similar technique that can be exploited to recoup speed loss at lowered supply voltages, thus enabling a significant reduction in power.

As stated in previous section, the VQ algorithm is a good choice for minimizing the computational complexity in the video decoder. However, it is clear that the ability of the algorithm to be parallelized is also very important because it is desirable to operate at very low supply voltages (where the gate delays are larger) for optimizing the energy per computation. It turns out that by organizing the lookup-table memory in a parallel fashion and applying pipelining, the memory access requirements can be reduced, allowing for operation at low supply voltages (1.5V).

Both pipelined and parallel architectures represent elegant examples of the direct trade-off between silicon area (number of transistors) and power consumption. With advanced packaging, such as



■ **Figure 7.** *(a) Analog performance comparison for MOS, BJT, and GaAs; (b) f_t vs. channel length for silicon MOS devices.*

multichip modules, area is no longer of great concern, but power is.

RF Transceiver Design

This section focuses on designing an analog modulator/demodulator in hardware capable of supporting the required baseband data rates (in excess of 2 Mb/s). In addition to meeting the system performance requirements, another major design goal is that the analog hardware be simplified as much as possible. Because of spectrum congestion, personal communications systems must operate at carrier frequencies well above 1 GHz, and the resulting complexity and difficulties in implementation imply that simplifying the circuitry or relaxing the required analog performance should be paramount.

Concomitant with the goal of simplification is the desire for monolithic integration of as much of the analog circuitry as possible. Traditionally, the realm of gigahertz-band RF front-end circuitry has been dominated by designs using discrete GaAs transistors and stripline filters, which take up a significant area on a circuit board and consume excessive power in driving high-speed analog signals across board interconnects, especially when matching to standard impedance levels (typically 50Ω). For portable applications, low power and small size are critical, thus requiring highly integrated analog technologies.

Silicon for High-Speed Analog

Given the advances in silicon processing and technology of the past decade, digital technologies have seen a breakthrough in both performance and size through the use of device scaling, especially in the arena of silicon complementary MOSFETs. For micro- and millimeter-wave IC technologies, GaAs has been used almost exclusively, despite the extra cost and processing difficulties, simply because silicon devices have not possessed the necessary performance. However, the same benefits derived from MOS scaling for digital circuits are reflected in analog applications as well.

The achievable unity-gain bandwidth f_t for the three major IC technologies — silicon CMOS, silicon bipolar, and GaAs — is shown in Fig. 7a. The f_t of a device is a measure of the maximum frequency that a device can operate and still provide active gain; above f_t, the intrinsic device parasitics dominate and the transistor appears to be a passive network. As the carrier frequencies must be placed in the low microwave bands around 1.5 GHz, an f_t of approximately 20 GHz will be needed to meet the performance requirements. Clearly, conventional silicon CMOS is below this point, silicon bipolar is just above this point, and GaAs is well ahead of both.

Silicon CMOS technologies, however, will continue to scale rapidly owing to the demand for denser and higher-performance memory chips and digital microprocessors. Several vendors to date have reported experimental memories using 0.2 to 0.3 μm technologies. For comparison, the projected gains in f_t over the next ten years is shown in Fig. 7a. Clearly, all technologies will be well above the required performance threshold. To examine the behavior of a MOSFET under scaling, a plot of f_t under scaling of the channel length L is shown in Fig. 7b. For long-channel devices, with $L > 1$ μm, the f_t increases quadratically as L decreases. However, as velocity saturation and other short-channel effects begin to become significant, the f_t tapers off to vary inverse linearly with L [12]. Even so, projecting from the reported ft of 5 GHz for today, in only a few years this will have jumped above 50 GHz as deep-submicron technologies become readily available. Thus, complete implementation of a 1.5-GHz transceiver in silicon CMOS is both possible and desirable, given the benefits of reduced parasitics, provision for digital processing functions on the same chip, and lower manufacturing costs over bipolar, GaAs, or hybrid BiCMOS.

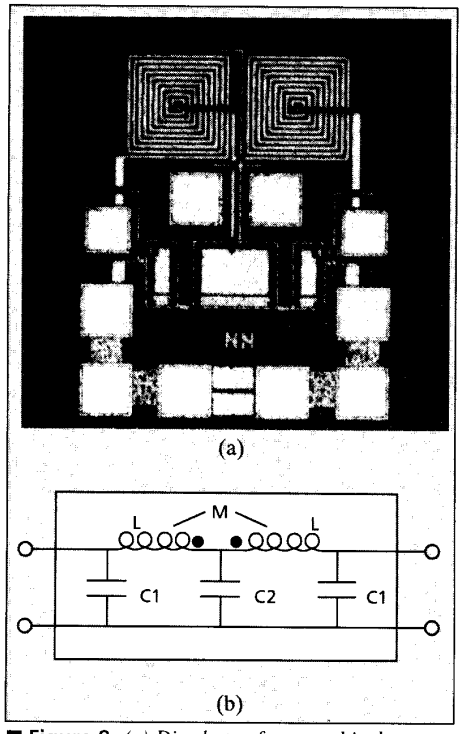
As an example of the impact of scaled technologies, the performance of an amplifier is directly related to the achievable f_t of the underlying technology. Assuming a single dominant-pole frequency response in the amplifier, the relationship between gain and bandwidth is thus given by:

$$(\text{Gain per stage}) \cdot (-3\text{ dB bandwidth}) \approx f_t \qquad \text{Eq. 3}$$

Thus, for an f_t of 20 GHz, and a required -3 dB bandwidth of 2 GHz, a maximum gain of approximately 20 dB can be reasonably expected from a single-stage MOS amplifier, which is easily comparable to the performance achieved by existing bipolar and GaAs amplifiers.

Digital technologies have seen a breakthrough in both performance and size through the use of device scaling, especially in the arena of silicon complementary MOSFETs.

One technique to simplify the receiver architecture is to use passive mixing, which takes advantage of the fact that the underlying signal is discrete-time in nature.



■ **Figure 8.** *(a) Die photo of an on-chip, low-pass filter using integrated stripline inductors; (b) Equivalent lumped circuit schematic (after Meyer and Nguyen [13]).*

Filters

For the transmitter, very little analog filtering is required, because the baseband transmit filter is more easily implemented in the digital domain; the signal is already bandlimited at the baseband before modulation into the passband. The only analog filter that is desirable is at the output of the power amplifier, where it is used to weakly bandlimit the signal to eliminate spurious frequencies from the oscillator. For the receiver, however, an anti-imaging filter is needed to eliminate all the undesirable frequency bands before demodulation.

Several options exist for implementing this passband anti-imaging filter; the conventional method is to utilize a high-Q off-chip filter to perform the necessary filtering. Generally, electromechanical filters such as ceramic resonators or surface-acoustic wave (SAW) filters are popular for use at frequencies below 1 GHz, and can achieve good performance at low cost. However, they are not tunable, and because they are dependent on material vibrational modes, they are not easy to fabricate for operation above 1 GHz. On the other hand, the electromagnetic resonators, such as L-C tank networks, suspended strip-line filters, and combline resonators, can easily operate well above 1 GHz, but tend to be physically large. For example, one commercially available combline filter achieves an extremely accurate fifth-order response over a passband of 2.2 to 2.3 GHz; however, its dimensions are 6 cm by 1.5 cm by 1.5 cm — far too large to be used in a small portable terminal [17].

Because operation frequencies are in the gigahertz band, another filtering option — the use of on-chip L-C filters — is available to hardware designers, because long aluminum traces on the chip die begin to exhibit transmission line effects and stripline behavior. By literally drawing a microstrip L-C circuit out of a spiral aluminum trace, a usable inductor is thus created. In the bandpass filter circuit shown in Fig. 8 (as developed by Meyer and Nguyen [13]), an effective inductance of 9 nH was achieved at 4.0 GHz, resulting in a filter Q of 10 with the entire circuit fabricated completely on-chip. At these frequencies, the Q was limited primarily by the 100Ω parasitic series resistance of the trace, and is certainly comparable to the performance found in off-chip L-C filters implemented with discrete components or striplines.

Frequency Conversion and Sampling Demodulation

A simplified block diagram for a conventional transmitter/2-stage heterodyne receiver is shown in Fig. 9. This structure will work as a transceiver for the personal communications system; however, simplifying this structure to achieve high integrability using silicon MOS, as well as minimizing power consumption, are important considerations.

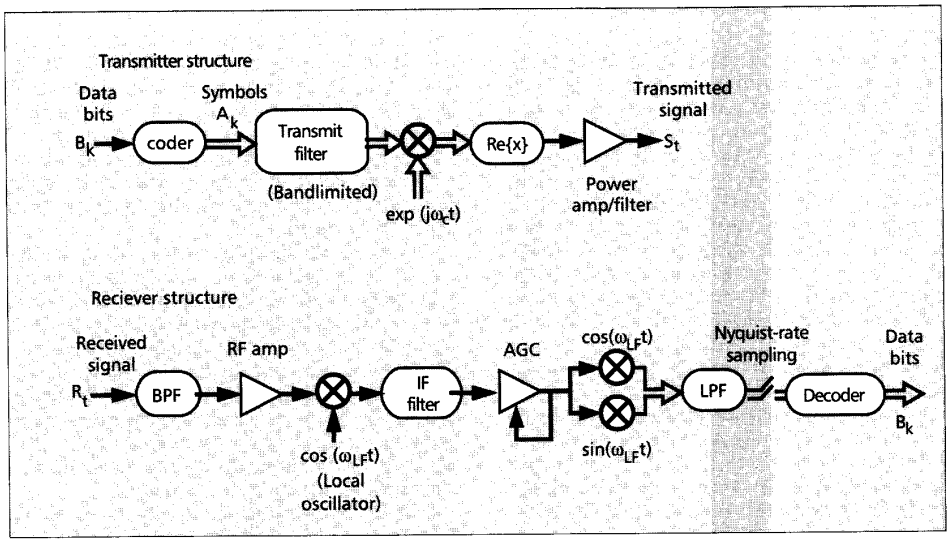
One way to achieve immediate simplification is to utilize a homodyne demodulation strategy in the receiver — that is, elimination of the IF stage entirely. To date, the primary problems with utilizing a homodyne scheme have been accuracy in the local oscillator (LO), gain imbalances between the I and Q paths, and reradiation of the receiver LO signal coupling into the antenna through the front-end amplifier. However, solutions to these problems exist, and in seeking the benefits in complexity and power that a homodyne scheme allows, at least one homodyne chip set for 900 MHz cellular communications has already been developed [14].

At this point, the differences between the transmitter and receiver must be considered. Referring to the block diagrams of Fig. 9, the transmitter can easily use a mixer-oscillator configuration to "upconvert" the signal in a homodyne fashion; methods for doing this using simple mixer-oscillator configurations are well-known. On the other hand, the receiver is required to track frequency and phase changes, and needs to recover the data from a signal corrupted by noise and channel distortion. One technique to simplify the receiver architecture is to use passive mixing, which takes advantage of the fact that the underlying signal is discrete-time in nature.

When considering demodulation, one very useful method of analysis is to examine the frequency domain representation of the signals. If $F(\omega)$ is the original signal spectrum, then the spectrum after ideal demodulation and sampling at a rate equal to ω_{samp} is:

$$\sum_{i=-\infty}^{\infty} F(\omega - i\omega_{samp}) \qquad \text{Eq. 4}$$

Suppose that $F(\omega)$ is modulated at a frequency of $\omega_{carrier}$, and the received signal bandpass sampled at the Nyquist rate of the baseband signal. This rate is much lower than the Nyquist rate for the incoming RF signal. After sampling, the spectrum of the signal has a transform equal to:



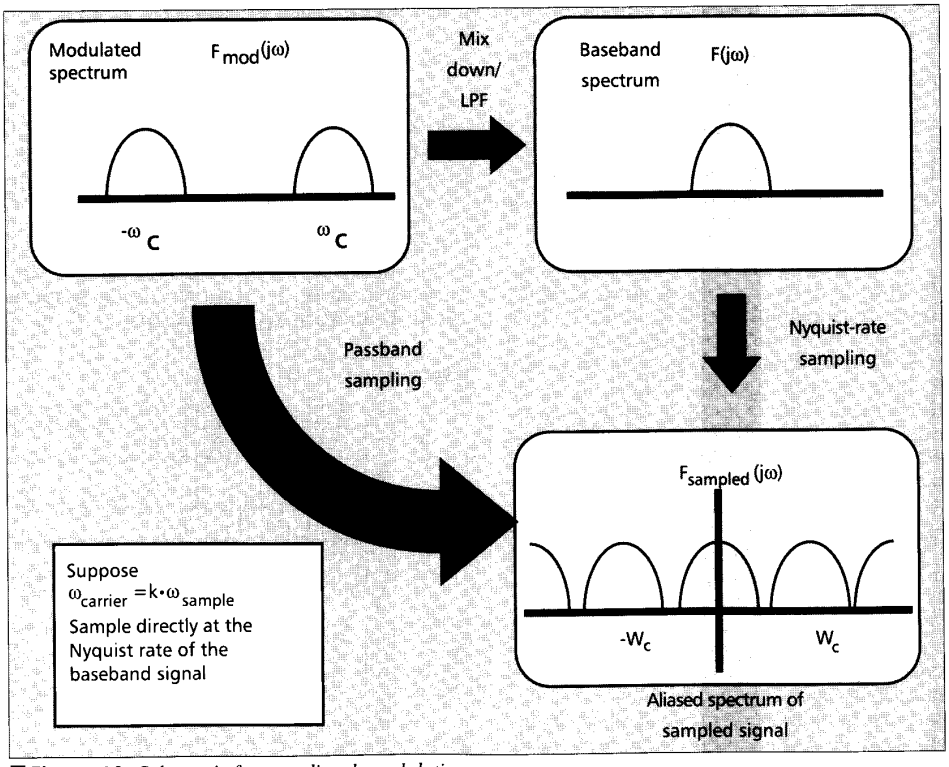
■ **Figure 9.** *Block diagram for a conventional heterodyne transceiver.*

$$\frac{1}{2}\sum_{i=-\infty}^{\infty}(F(\omega-\omega_{carrier}-i\omega_{samp}) + F(\omega+\omega_{carrier}-i\omega_{samp})) \quad \text{Eq. 5}$$

Making the important stipulation that $\omega_{carrier} = k \cdot \omega_{samp}$, where k is an integer, this sum can be simplified to:

$$\sum_{i=-\infty}^{\infty} F(\omega-i\omega_{samp}) \quad \text{Eq. 6}$$

which is precisely the same result achieved by the original demodulation structure. The entire demodulation step has been reduced to a single sampling operation, sampled at the Nyquist rate for the baseband signal. A graphical depiction of this process is shown in Fig. 10. The path indicated by the solid black arrows shows the result of each step of a conventional mixer-sampler structure, and the single path indicated by the large gray arrow is the result of bandpass sampling the modulated signal. Clearly, both will yield the same aliased spec-



■ **Figure 10.** *Schematic for sampling demodulation.*

Because portability places severe constraints on the size and weight of the terminal itself, power is at a premium.

It is evident that an all-MOS RF system operating in the 1- to 2-GHz range will be possible in the near future as deep-submicron devices become commercially available.

trum, provided $\omega_{carrier} = k \cdot \omega_{samp}$.

Furthermore, the sampling rate does not change frequencies, even as the carrier changes. The k in the above stipulation was not specified; this method works so long as the carrier frequency is an integer multiple of the baseband sampling rate. Hence, a voltage-controlled oscillator is not needed and a high accuracy crystal oscillator can be used directly. The hardware and power costs are minimal: an accurate switch, implementable using standard MOS technologies, and a fixed-frequency crystal oscillator. Theoretically, the sampling demodulator is an excellent implementation for the demodulator in terms of hardware and power, utilizing an active device as a simple passive switch. However, the key limitation is the accuracy in the sample-and-hold amplifier: the switch must close quickly enough to capture the 1.5 GHz modulated signal accurately, even though it samples at a much slower rate. From the theory of sample-and-hold design, the performance of such a switch is proportional to the f_t of the device, since the turn-off speed is related to how quickly charge can move through the device. Thus, with scaled technologies, it has been shown that for MOSFETs with 0.2 μm or smaller, minimum feature size will possess the necessary performance [6], thus allowing a receiver which has an extremely simple topology and potentially very low power consumption.

Conclusions

The ultimate goal is to provide a wireless personal communications system capable of supporting a wide variety of services by interfacing, via a fixed base station, into a network of compute servers interlinked by a high-speed fiber-optic link. Beyond providing simple person-to-person communications, such a system will also provide person-to-data communications, such as access to information databases, full-motion video sequences, advanced computation servers. The centerpiece of such a system lies in the development of a small, portable terminal unit, capable of supporting all of these functions while minimizing both bandwidth and power consumption; furthermore, to simplify interfacing, speech and pen-based input will be supported in an integrated fashion.

The portable terminal itself will carry no computation capability beyond that needed to support its interfacing functions: the RF transceiver, baseband communications processing, image decompression, and the speech codec. While providing significant power savings by minimizing the amount of circuitry needed in the mobile unit, the key concept behind such a partitioning of functionality between the mobile and the fixed network servers is that no "user" data will be carried over the high bit-error-rate wireless link — data that absolutely cannot tolerate any errors in transmission. Instead, only "multimedia" data such as speech and video will be transmitted — data that is by nature relatively corruption-tolerant.

To minimize the physical bandwidth occupied by the system, a picocellular network architecture of base stations will be utilized to achieve as much spatial frequency reuse as possible. In an indoor environment, the cells are naturally defined by walls and other physical partitions within the building. By examining the user distribution and existing system usage patterns in a typical office environment, a ($K = 7$) cellular system consuming 100 MHz of bandwidth with 5 meter cells will effectively support all users within the service area. Although 100 MHz is a sizable block of spectrum, when compared to existing spectrum consumption by systems such as NTSC television, the system described here is quite attractive, especially given the breadth of services that will be offered.

Because portability places severe constraints on the size and weight of the terminal itself, power is at a premium. Batteries cannot provide much power on a continuous basis without weighing more than is practical or having an extremely short usable life between rechargings. Thus, low-power digital systems design is important. Through optimized circuit design strategies, reduced supply-voltage operation, and architectural techniques such as parallelism and pipelining, the power consumed in the terminal can be reduced dramatically.

The feasibility of a high-performance, monolithic RF transceiver has been addressed. Given the continued advances in scaling of silicon MOS technologies, it is evident that an all-MOS RF system operating in the 1 to 2 GHz range will be possible in the near future as deep-submicron devices become commercially available. In addition, though the use of integrated on-chip filters as well, it is conceivable that the transceiver in a personal communications system can be found entirely on one silicon multichip module, with only a single external timing crystal. The benefits of integration are enormous: reduced parasitic effects, greater manufacturability, and minimized power requirements to drive off-chip loads. Likewise, by examining the basic architecture used in the transceiver, the underlying digital nature of the signal can be used to simplify the resulting circuit considerably: homodyne demodulation using passive sampling techniques is one important example of how this can be achieved.

Acknowledgments

We would like to thank the members of our research group who have been working toward implementing the multimedia terminal described in this paper: Prof. Jan Rabaey, Bill Baringer, Trevor Blumenau, Andy Burstein, Kathy Lu, Roger Doering, Brian Richards, and Tom Truman. Also, Professors Paul Gray and Robert Meyer must be thanked for their invaluable advice, as well as The Fannie and John Hertz Foundation for its kind support.

References

[1] A. Burstein, A. Stoelze, and R.W. Brodersen, "Using Speech Recognition in a Personal Communications System," in *Proc. 1992 International Communications Conf.*, Chicago, IL, June 1992.

[2] W. C-Y. Lee, *Mobile Cellular Telecommunications*, (New York: McGraw- Hill Book Co. 1989).

[3] S.Y. Seidel and T.S. Rappaport, "914 MHz Path-Loss Prediction Models for Indoor Wireless Communications," submitted to *IEEE Trans. Antennas and Propagation*, May 17, 1991.

[4] K. Feher, *Advanced Digital Communications*, (New Jersey: Prentice-Hall Inc. 1987).

[5] K. Raith and J. Uddenfelt, "Capacity of Digital Cellular TDMA Systems," *IEEE Trans. Vehicular Technology*, Vol. VT-40, No.2, pp.323-332, May 1991.

[6] S. Sheng, *Wideband Digital Portable Communications: A System Design*, M.S.Thesis, Univ.of California, Berkeley, 1991.

[7] K.R. Rao and P. Yip, *Discrete Cosine Transform*, (New York: Academic Press 1990).

[8] E. Feig, "On the Multiplicative Complexity of the Discrete Cosine Transform," in *Proc. 1990 SPIE/SPSE Symposium of Electronic Imaging Science and Technology*, Santa Clara, CA, Feb. 1990.

[9] W. C. Fang, C.Y. Chang, and B.J. Sheu, "A Systolic Tree-Searched Vector Quantizer for Real-Time Image Compression," *VLSI Signal Processing IV*, (New York: IEEE Press, 1990).

[10] D. Dahle, "Designing High-Performance Systems to run from 3.3V or lower supplies," in *Proc. 1991 Silicon Valley Computer Conference*, Santa Clara, CA, 1991.
[11] K. Yano *et al.*, "A 3.8-ns CMOS 16x16 Multiplier Using Complementary Pass Transistor Logic," *IEEE Journal of Solid-State Circuits*, Vol. 25, No. 4, pp. 388-395, April 1990.
[12] R.K. Watts, *Submicron Integrated Circuits*, (New York: John Wiley and Sons, 1989).
[13] N.M. Nguyen and R.G. Meyer, "Si-Compatible Inductors and LC Passive Filters," *IEEE Journal of Solid-State Circuits*, Vol. 25, No. 4, pp. 1028-1030, Aug. 1990.
[14] J. Sevenhans, *et al.*, "An Integrated Si-Bipolar RF Transceiver for a Zero-IF 900 MHz GSM Digital Mobile Radio Front-end of a Hand Portable Phone," in *Proc. IEEE 1991 Custom Integrated Circuits Conference*, San Diego, CA, May 12-15, 1991.
[15] A. Chandrakasan, S. Sheng and R.W. Brodersen, "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, Vol. 27, No. 4, pp. 473-484, Apr. 1992.
[16] D.A. Hodges and H.G. Jackson, *Analysis and Design of Digital Integrated Circuits*, (New York: McGraw-Hill Book Co. 1983).
[17] Daden Associates, Inc. Combline Bandpass Filter, #CS2250-100-5SS.

Biographies

SAMUEL SHENG received the B.S. degree in electrical engineering and the B.A. degree in applied mathematics in 1989 from the University of California, Berkeley, where he has just completed his M.S.E.E. thesis work on a design of a wideband portable digital microwave transceiver. He will continue on with the Ph.D. program in the area of integrated RF transceivers, with emphasis on high-speed MOS analog design and low power. During the summer of 1991, he worked in the area of digital communications with the Advanced Technologies Group at Apple Computer in Sunnyvale, CA. Mr. Sheng is a member of Phi Beta Kappa, Eta Kappa Nu, and Tau Beta Pi, and has won the Certificate of Academic Distinction and the Departmental Citation in the Department of EECS from the University of California, Berkeley. He was also a National Merit Scholar and since 1989 has been a fellow of The Fannie and John Hertz Foundation.

ANANTHA P. CHANDRAKASAN received the B.S. (with highest honors) and M.S. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1989 and 1990 respectively. He is currently working toward a Ph.D. degree in electrical engineering at Berkeley. His research interests include low-power techniques for portable real-time applications, video compression, computer-aided design tools for VLSI, and system-level integration. He is a member of Eta Kappa Nu and Tau Beta Pi.

ROBERT W. BRODERSEN received B.S. degrees in electrical engineering and in mathematics at California State Polytechnic University, Pomona, CA in 1966. He received Engineering and M.S. degrees in 1968, and a Ph.D. degree in 1972, from MIT, Cambridge, MA. From 1972 to 1976, he was with the Central Research Laboratory at Texas Instruments, Inc., Dallas, TX, where he was engaged in the study of operation and applications of charge coupled devices. In 1976, he joined the Electrical Engineering and Computer Science faculty of the University of California, Berkeley, where he is now a professor. In addition to teaching, he is involved in research involving new applications of integrated circuits, which is focused in the area of signal processing; and the CAD tools necessary to support this activity. He has won best paper awards at a number of conferences, and in 1979 he received the W.G. Baker award. In 1982 he became a fellow of the IEEE and in 1983, he was corecipient of the IEEE Morris Liebmann award. In 1986 and 1991, he received the Technical Achievement awards in the IEEE Circuits and Systems Society and the Signal Processing Society. In 1988 he was elected to be member of the National Academy of Engineering.