INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

2003 EDITION

SYSTEM DRIVERS

THE ITRS IS DEVISED AND INTENDED FOR TECHNOLOGY ASSESSMENT ONLY AND IS WITHOUT REGARD TO ANY COMMERCIAL CONSIDERATIONS PERTAINING TO INDIVIDUAL PRODUCTS OR EQUIPMENT.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SYSTEM DRIVERS

SCOPE

Future semiconductor manufacturing and design technology capability is developed in response to economic drivers within the worldwide semiconductor industry. The ITRS must understand how technology requirements arise for product classes whose business and retooling cycles drive the semiconductor sector. Until 2001, the ITRS focused on microprocessor (MPU), dynamic random-access memory (DRAM), and application-specific integrated circuit (ASIC) product classes, with some mention of system-on-chip (SOC) and analog/mixed-signal circuits. The unstated assumption was that technological advances needed only be straight-ahead and "linear", and would be deployed in all semiconductor products. For this reason, specifics of the product classes (e.g., MPU or ASIC) were not required. Today, introduction of new technology solutions is increasingly application-driven, with products for different markets making use of different combinations of technologies at different times. General-purpose digital microprocessors for personal computers have been joined as drivers by mixed-signal systems for wireless communication and embedded applications. Wall-plugged servers are being replaced by battery-powered mobile devices. In-house, single-source chip designs are being supplanted by SOC and system-in-package (SIP) designs that incorporate building blocks from multiple sources.

The purpose of the 2003 ITRS System Drivers Chapter is to update and more clearly define the system drivers as used in previous ITRS editions. Together with the Overall Roadmap Technology Characteristics, the System Drivers Chapter provides consistent framework and motivation for technology requirements across the respective ITRS technology areas and the 15-year span of the ITRS. The main contribution of the Chapter consists of quantified, self-consistent models of the system drivers that support extrapolation into future technologies and adapt more smoothly to future technology developments. We focus on four system drivers: system-on-chip (SOC), microprocessor (MPU), analog/mixed-signal (AMS), and embedded memory. Before describing these drivers, we briefly survey key market drivers for semiconductor products. The reader is also referred to the *NEMI roadmap*, http://www.nemi.org.

MARKET DRIVERS

Table 8 contrasts semiconductor product markets according to such factors as manufacturing volume, die size, integration heterogeneity, system complexity, and time-to-market. Influence on the SOC, AMS and MPU drivers is noted.[1]

Table 8 Major Product Market Segments and Impact on System Drivers

MARKET DRIVERS	*SOC*	*ANALOG/MS*	*MPU*
I. Portable and Wireless			
1. Size/weight ratio: peak in 2004 2. Battery life: peak in 2004 3. Function: 2×/2 years 4. Time-to-market: ASAP	Low power paramount Need SOC integration (DSP, MPU, I/O cores, etc.)	Migrating on-chip for voice processing, A/D sampling, and even for some RF transceiver function	Specialized cores to optimize processing per microwatt.
II. Broadband			
1. Bandwidth: 2× / 9 months 2. Function: 20%/yr increase 3. Deployment/Operation Cost: flat 4. Reliability: asymptotic 99.999% 5. Time-in-market: long 6. Power: W/m^3 of system	Large gate counts. High reliability. Primarily SOC.	Migrating on-chip for signal recovery, A/D sampling, etc.	MPU cores and some specialized functions.

[1] *The market drivers are most clearly segmented according to cost, time-to-market, and production volume. System cost is equal to Manufacturing cost + Design cost. Manufacturing cost breaks down further into non-recurring engineering (NRE) cost (masks, tools, etc.) and silicon cost (raw wafers + processing + test). The total system depends on function, number of I/Os, package cost, power and speed. Different regions of the (Manufacturing Volume, Time To Market, System Complexity) space are best served by FPGA, Structured-ASIC, or SOC implementation fabrics, and by single-die or system-in-package (SIP) integration. This partitioning is continually evolving.*

Table 8 Major Product Market Segments and Impact On System Drivers (continued)

MARKET DRIVERS	*SOC*	*ANALOG/MS*	*MPU*
III. Internet Switching			
1. Bandwidth: 4×/3–4 yrs. 2. Reliability 3. Time-to-market: ASAP 4. Power: W/m^3 of system	Large gate counts. High reliability. More reprogrammability to accommodate custom functions.	Migrating on-chip for MUX/DEMUX circuitry. MEMS for optical switching.	MPU cores, FPGA cores and some specialized functions.
IV. Mass Storage			
1. Density: 60% increase/year 2. Speed: 2× by 2007 3. Form factor: shift toward 2.5"	High-speed front-end for storage systems. Primarily ASSP. Shift toward large FPGA and COT, away from ASIC costs and design flows	Highest-speed A/D sampling on chip. Increases for higher precision in positioning, "inertia knowledgeable" actuation, on-chip power control. MEMS sensing on R/W head as an SIP option.	High-speed hardware for, e.g., "look-ahead" in DB search, MPU instruction pre-fetch, data compression, S/N monitoring, failure prediction.
V. Consumer			
1. Cost: strong downward pressure 2. Time-to-market: <12 mos 3. Function: high novelty 4. Form factor 5. Durability/safety 6. Conservation/ecology	High-end products only. Reprogrammability possible. Mainly ASSP; more SOC for high-end digital with cores for 3D graphics, parallel proc, RTOS kernel, MPU-MMU-DSP, voice synthesis and recognition, etc.	Increased integration for voice, visual, tactile, physical measurement (e.g., sensor networks). CCD or CMOS sensing for cameras.	For "long-life" mature products only. Decrease in long design cycles, and in use of high-cost non-prepackaged functions and design flows.
VI. Computer			
1. Speed: 2×/2 years 2. Memory density: 2×/2 years 3. Power: flat to decreasing, driven by cost and W/m^3 4. Form factor: shrinking size 5. Reliability	Large gate counts. High speed. Drives demand for digital functionality. Primarily SOC integration of custom off-the-shelf MPU and I/O cores.	Minimal on-chip analog. Simple A/D and D/A. Video i/f for automated camera monitoring, video conferencing. Integrated high-speed A/D, D/A for monitoring, instrumentation, and range-speed-pos resolution.	MPU cores and some specialized functions. Increased industry partnerships on common designs to reduce development costs (requires data sharing and reuse across multiple design systems).
VII. Automotive			
1. Functionality 2. Ruggedness (external environment, noise) 3. Reliability and safety 4. Cost	Mainly entertainment systems. Mainly ASSP, but increasing SOC for high end using standard H/W platforms with RTOS kernel, embedded software.	Cost-driven on-chip A/D and D/A for sensor and actuators. Signal processing shifting to DSP for voice, visual. Physical measurement ("communicating sensors" for proximity, motion, positioning). MEMS for sensors.	

SYSTEM ON CHIP DRIVER

SOC is a yet-evolving *product class and design style.* The most important observation is that SOC integrates technology and design elements from other system driver classes (MPU, embedded memory, AMS—as well as reprogrammable logic) into a wide range of high-complexity, high-value semiconductor products. Manufacturing and design technologies for SOC are typically developed originally for high-volume custom drivers. The SOC driver class most closely resembles, and is evolved most directly from, the ASIC category since reduced design costs and higher levels of system integration are its principal goals.[2] The primary difference from ASIC is that in SOC design, the goal is to maximize *reuse* of existing blocks or "cores"—i.e., minimize the amount of the chip that is newly or directly created. Reused blocks in SOC include analog and high-volume custom cores, as well as blocks of *software* technology. A key challenge is to invent, create and maintain reusable blocks or cores so that they are available to SOC designers.[3] Economic viability of SOC also requires that validation of reuse-based SOC designs becomes easier than for equivalent "from-scratch" designs.

SOC represents a confluence of previous product classes in several ways. As noted above, SOCs integrate building blocks from the other system driver classes, and are subsuming the ASIC category. The quality gap between full-custom and ASIC/SOC is diminishing: 1) starting in the 2001 ITRS, overall ASIC and MPU logic densities are modeled as being equal; and 2) "custom quality on an ASIC schedule" is increasingly achieved by on-the-fly ("liquid") or tuning-based standard-cell methodologies. Finally, MPUs have evolved into SOCs: 1) MPUs are increasingly designed as cores to be included in SOCs, and 2) MPUs are themselves designed as SOCs to improve reuse and design productivity (as discussed below, the ITRS MPU model has multiple processing cores and resembles an SOC in organization[4]). The most basic SOC challenge is presented by implementation productivity and manufacturing cost, which require greater reuse as well as platform-based design, silicon implementation regularity, or other novel circuit and system architecture paradigms. Another challenge is the heterogeneous integration of components from multiple implementation *fabrics* (e.g., reprogrammable, memory, analog and RF, MEMS, and software).

The SOC driver class is characterized by heavy reuse of intellectual property (IP) to improve design productivity, and by *system* integration that potentially encompasses heterogeneous technologies. SOCs exist to provide low cost and high integration. *Cost* considerations drive the deployment of low-power process and low-cost packaging solutions, along with fast-turnaround time design methodologies. The latter, in turn, require new standards and methodologies for IP description, IP test (including built-in self-test and self-repair), block interface synthesis, etc. *Integration* considerations drive the demand for heterogeneous technologies (flash, DRAM, analog and RF, MEMS, FRAM, MRAM, chemical sensors, etc.) in which particular system components (memory, sensors, etc.) are implemented, as well as the need for chip-package co-optimization. Thus, SOC is the driver for convergence of multiple technologies not only in the same system package, but also potentially in the same manufacturing process. We discuss the nature and evolution of SOCs with respect to three variants driven respectively by multi-technology integration (MT), high performance (HP), and low power and low cost (LP). This partition is by no means disjointed, but rather reflects separate driving concerns (e.g., low-power design *is* high-performance design, but must also reduce package and system cost).

---

[2] *Most digital designs today are considered to be ASICs. ASIC connotes both a business model (with particular "handoff" from design team to ASIC foundry) and a design methodology (where the chip designer works predominantly at the functional level, coding the design at Verilog/VHDL or higher level description languages and invoking automatic logic synthesis and place-and-route with a standard-cell methodology). For economic reasons, custom functions are rarely created; reducing design cost and design risk is paramount. ASIC design is characterized by relatively conservative design methods and design goals (cf. differences in clock frequency and layout density between MPU and ASIC in previous ITRS editions) but aggressive use of technology, since moving to a scaled technology is a cheap way of achieving a better (smaller, lower power, and faster) part with little design risk (cf. convergence of MPU and ASIC process geometries in previous ITRS editions). Since the latter half of the 1990s, ASICs have been converging with SOCs in terms of content, process technology, and design methodology.*

[3] *For example, reusable cores might require characterization of specific noise or power attributes ("field of use", or "assumed design context") that are not normally specified. Creation of an IC design artifact for reuse by others is substantially more difficult (by factors estimated at between 2× and 5×) than creation for one-time use.*

[4] *The corresponding ASIC and structured-custom MPU design methodologies are also converging to a common "hierarchical ASIC/SOC" methodology. This is accelerated by customer-owned tooling business models on the ASIC side, and by tool limitations faced by both methodologies.*

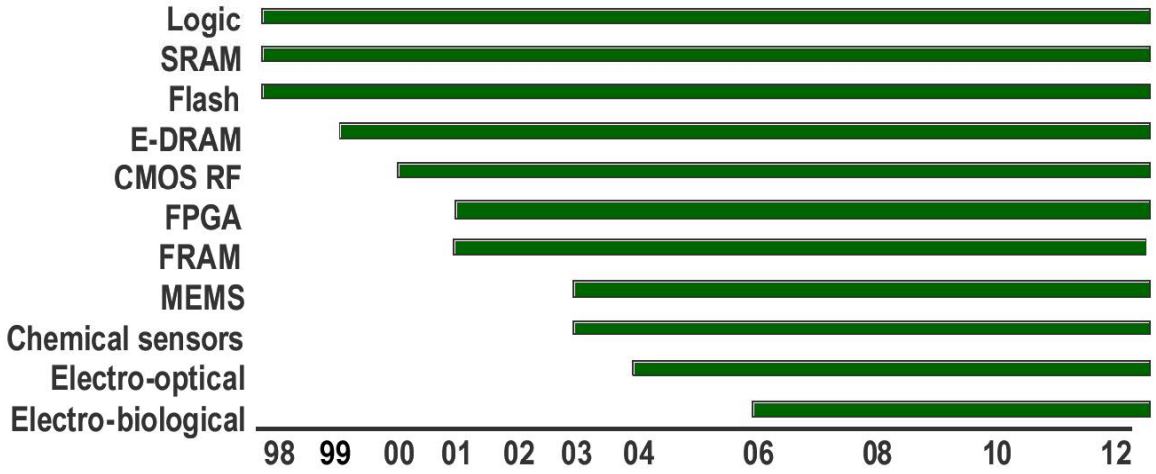

Figure 9 First Integration of Technologies in SOC with Standard CMOS Process

SOC MULTI-TECHNOLOGY

The need to build heterogeneous systems on a single chip is driven by such considerations as cost, form-factor, connection speed/overhead, and reliability. Thus, process technologists seek to meld CMOS with MEMS, and other sensors. Process complexity is a major factor in the cost of SOC-MT applications, since more technologies assembled on a single chip requires more complex processing. The total cost of processing is difficult to predict for future new materials and combinations of processing steps. However, at present cost considerations limit the number of technologies on a given SOC: processes are increasingly modular (e.g., enabling a flash add-on to a standard low-power logic process), but the modules are not generally "stackable". Figure 9 shows how first integrations of each technology within standard CMOS processes—not necessarily together with other technologies, and not necessarily in volume production—might evolve. CMOS integration of the latter technologies (electro-optical, electro-biological) is less certain, since this depends not only on basic technical advances but also on SOC-MT being more cost-effective than multi-die SIP alternatives. Today, a number of technologies (MEMS, GaAs) are more cost-effectively flipped onto or integrated side-by-side with silicon in the same module depending also on the area and pin-count restrictions of the respective product (e.g. Flash, DRAM). Physical scale in system applications (e.g., ear-mouth = speaker-microphone separation, or distances within a car) also affect the need for single-die integration, particularly of sensors.

SOC HIGH-PERFORMANCE

Examples of SOC-HP include network processors and high-end gaming applications. Since it reflects MPU-SOC convergence, SOC-HP follows a similar trend as MPU and is not separately modeled here. However, one aspect of SOC-HP merits discussion, namely, that instances in the high-speed networking domain drive requirements for off-chip I/O signaling (which in turn create significant technology challenges to Test, Assembly and Packaging, and Design). Historically, chip I/O speed (per-pin bandwidth) has been scaling much more slowly than internal clock frequency. This is partly due to compatibility with existing slow I/O standards, but the primary limitation has been that unterminated CMOS signals on printed circuit boards are difficult to run at significantly greater than 100MHz due to slow settling times. During the past decade, high-speed links in technology initially developed for long-haul communication networks have found increasing use in other applications. The high-speed I/O eliminates the slow board settling problems by using point-to-point connections and treating the wire as a transmission line. Today the fastest of these serial links can run at 10Gbit/s per pin.

A high-speed link has four main parts: a transmitter to convert bits to an electrical signal that is injected into the board-level wire, the wire itself, a receiver that converts the signal at the end of the wire back to bits, and a timing recovery circuit that compensates for the delay of the wire and samples the signal on the wire at the right place to get the correct data. Such links are intrinsically mixed-signal designs since receivers, transmitters, and timing recovery all require analog blocks (e.g., the VCO discussed as part of the Mixed-Signal driver is a key component of a timing recovery circuit). Broadly speaking, high-speed links are used in optical systems, chip-to-chip connections, and backplane connections. We now discuss each of these applications in slightly more detail.

Optical links generally push link performance the hardest; since there are generally a small number of optical signals, these links can tolerate relatively complex and power hungry interface circuits. Today, optical links run at 10Gbit/s per pin, and are expected to continue to scale up in frequency as projected in the Test Chapter (high-speed serial links discussion). Initially, electronics for these links were created in non-CMOS technologies, since CMOS was thought incapable of meeting the high-speed requirements. However, over the past five years, many researchers have developed circuits that can run at 10 Gbit/s. While some papers have demonstrated links that run as fast as 1 FO4 delay per bit, most links run at 2–4 FO4 delays per bit, which yields 10 Gbit/s in the 180 nm node. Continuing to scale link speed with technology should be possible from the circuits' standpoint, but will become difficult due to parasitics and packaging. Signals at this speed are highly sensitive to any discontinuities in their signal path. Even if controlled impedance packaging is used, vias in the package or board can cause impedance changes that will degrade the signal. The 1–2 pF parasitic capacitance from the ESD device will also significantly degrade the signal. Thus, continued performance scaling will require significant work in ESD, package and board design.

Chip-to-chip interconnections communicate information between two chips located on the same board, usually close to each other. The main metric driving the design of these links is not Gbit/s since it is generally possible to use a number of links in parallel to connect these chips. For example, if going twice as fast requires 10× the area and 10× the power, it is better to use two links in parallel. Thus, these links are optimized for performance and cost, not just performance. In general, the highest chip-to-chip link speeds are 2–4 times slower than the highest optical link speeds. Bit times for these links vary dramatically, e.g., point-to-point links are available today with bit times ranging from about 2.5 ns (400 Mbit/s) to .4ns (2.5 Gbit/s). This wide range of performance reflects dependencies on the number of IO required (higher IO counts have slower speeds), the degree of risk the designer is willing to take, and sometimes an existing I/O standard. Design of robust high-speed I/O is still a mixed-signal problem that cannot be automated or checked with current tools. Thus, many design teams are still conservative when choosing I/O rates. As technology scales and design tools become more robust, bit times should approach 4-8 FO4 delays, but this will require additional circuitry to compensate for package and other parasitic effects.

The last major application for high-speed links is in networking, where two chips on different boards must communicate. The signal path is still point-to-point, but travels from one chip through its package to the local board, through a connector to another board, through another connector to the destination board, and then through that board and receiver package to the receiver chip. For high bandwidth each chip generally has a large number of links, so that performance per unit cost is critical. The principal difference from chip-to-chip links is that the "wire" between the two chips has worse electrical properties. Wire issues are a serious concern as speeds increase through 10 Gbit/s, which is achieved in the 90 nm node. These I/O considerations also show the trade-off between SOC and SIP solutions in the high-speed area.

SOC LOW-COST, LOW-POWER

Examples of SOC-LP include portable and wireless applications such as PDAs or digital camera chips. Table 9 sets requirements for various attributes of a low-power, consumer-driven, handheld wireless device ("PDA") with multimedia processing capabilities, based in part on the model created by the Japan Semiconductor Technology Roadmap Working Group 1 and originally introduced in the 2000 ITRS update (Design Chapter). Key aspects of the model are as follows.[5]

- The system design consists of embedded blocks of CPU, DSP and other processing engines, and SRAM and embedded DRAM circuits. Processor core logic increases by 4× per node, and memory content increases by 2–4× per node.[6]

[5] *Other aspects of the model, which are not essential to the following analyses, address external communication speed (increasing by 6× per node in the near term, starting from 384 Kbps in 2001) and addressable system memory (increasing by 10× per node, starting from 0.1 Gb in 2001).*

[6] *The PDA contained approximately 20 million transistors in 2001, and will contain approximately 41 million transistors in 2004. The model assumes that increasing parallel computation will be required in each generation of the device, to support video, audio and voice recognition functionality. This is reflected in CPU and DSP content (e.g., number of cores), which increases four-fold (4×) per technology node to match the processing demands of the corresponding applications. (By comparison, MPU logic content is projected to double with each node.) Overhead area (I/O buffer cells, pad ring, white space due to block packing, analog blocks, etc.) is fixed at 28% of the die. The 41M transistor count in 2004 is broken down as follows. A typical CPU/DSP core (e.g., ARM) today is approximately 30–40K gates, or 125K transistors. We assume 16 such cores on chip in 2004, i.e., 2M CPU/DSP core transistors. In 2004, the "peripheral" logic transistor count is 23M transistors, and this count grows at 2×/node thereafter. SRAM transistor count is 16M in 2004, and grows at 2×/node thereafter. The composition of SRAM versus DRAM depends on the ratio of memory to logic. We assume that embedded DRAM (eDRAM) is cost effective when at least 30% of the chip area is memory; this trigger point occurs at 16 Mb in 2004. Once triggered, the eDRAM content quadruples every technology node. (While the SOC-LP PDA is a "single-chip design", we do not imply any judgment as to whether multi-die or single-die implementation will be more cost-effective.)*

- Die size increases on average by 10% per node through 2018 to accommodate increased functionality; this matches historical trends for the application domain.
- Layout densities for memory and logic fabrics are the same as for the MPU driver, with eDRAM density assumed to be 3× SRAM density.
- Maximum on-chip clock frequency is approximately 5–10% of the MPU clock frequency at each node.

Peak power dissipation is limited to 0.1 W at 100^0C, and standby power to 2.1 mW, due to battery life.

Table 9 System Functional Requirements for the PDA SOC-LP Driver

YEAR OF PRODUCTION	2003	2006	2009	2012	2015	2018
Process Technology (nm)	101	90	65	45	32	22
Supply Voltage (V)	1.2	1	0.8	0.6	0.5	0.4
Clock Frequency (MHz)	300	450	600	900	1200	1500
Application (maximum required performance)	Still Image Processing	Real Time Video Codec (MPEG4/CIF)		Real Time Interpretation		
Application (other)	Web Browser Electric Mailer Scheduler	TV Telephone (1:1) Voice Recognition (Input) Authentication (Crypto Engine)		TV Telephone (>3:1) Voice Recognition (Operation)		
Processing Performance (GOPS)	0.3	2	14	77	461	2458
Required Average Power (W)	0.1	0.1	0.1	0.1	0.1	0.1
Required Standby Power (mW)	2	2	2	2	2	2
Battery Capacity (Wh/Kg)	120	200	200	400	400	400

SOC TRENDS

SOC presents Design, Test, PIDS and other areas with a number of technology challenges, such as development of reusable analog IP. The most daunting SOC challenges are:

- *design productivity improvement of > 100% per node*, with needs including platform-based design[7] and integration of programmable logic fabrics (Design),[8]
- *management of power* especially for low-power, wireless, multimedia applications (Design, PIDS),
- *system-level integration of heterogeneous technologies* including MEMS and optoelectronics (PIDS, FEP, Design), and
- *development of SOC test methodology*, with needs including test reusability and analog/digital BIST.

Since SOC is aimed at low-cost and rapid system implementation, and since power is one of the grand challenges in recent ITRS editions, it is appropriate to consider implications of *power management* on the achievable space of SOC designs. The following discussion develops trend analyses for the SOC-LP driver with respect to this issue.

Two approaches can be used to derive the power dissipation for the SOC-LP model. The first approach is to accept the system specifications (0.1 W peak power, and 2 mW standby power) in a "top-down" fashion. The second approach is to derive the power requirements "bottom-up" from the implied logic and memory content, as well as process and circuit

[7] *Platform-based design is focused on a specific application domain. The platform embodies the hardware architecture, embedded software architecture, design methodologies for IP authoring and integration, design guidelines and modeling standards, IP characterization and support, and hardware/software verification and prototyping. Derivative designs may be rapidly implemented from a single platform that has a fixed portion and a variable portion that permits proprietary or differentiated designs. (See: H. Chang et al., Surviving the SOC Revolution: A Guide to Platform-based Design, Boston:Kluwer Academic, 1999.)*

[8] *A programmable logic core is a flexible logic fabric that can be customized to implement any digital logic function after fabrication. The structure of a programmable logic fabric may be similar to an FPGA capability within specific blocks of the SOC. They allow reprogrammability, adaptability and reconfigurability, which greatly improve chip productivity. Applications include blocks that implement standards and protocols that continue to evolve, changing design specifications, and customization of logic for different, but related, applications and customers.*

parameters. Logic power consumption is estimated based on $\alpha CV_{dd}^2 f + I_{off} V_{dd}$ model for dynamic plus static power, using area-based calculations similar to those in the MPU power analysis. The memory power consumption model also uses $\alpha CV_{dd}^2 f + I_{off} V_{dd}$ with a different factor for α.[9] For these calculations, we refer to the low-power device roadmap described in the PIDS Chapter. It is almost certain that future low-power SOCs will integrate multiple (LOP, LSTP, HP) technologies simultaneously within the same core, to afford greater control of dynamic power, standby power, and performance.

Figure 10 shows the "bottom-up" *lower bound* for total chip power at an operating temperature of 100°C, assuming that all logic is implemented with LOP or LSTP devices and operates as described in Footnote 25. We say that this is a lower bound since in practice some logic would need to be implemented with faster, higher-current devices. The figure suggests that SOC-LP power levels will exceed the low-power requirements of the PDA application, and further provides a breakdown of power contributions for each case. As expected, LOP power is primarily due to standby power dissipation while LSTP power is primarily due to dynamic power dissipation[10]. Total chip power using only LOP devices reaches 1.39 W in 2018, mostly due to a sharp rise in static power after 2012. Total chip power using only LSTP devices reaches 1.27 W in 2018; almost all of this is dynamic power.

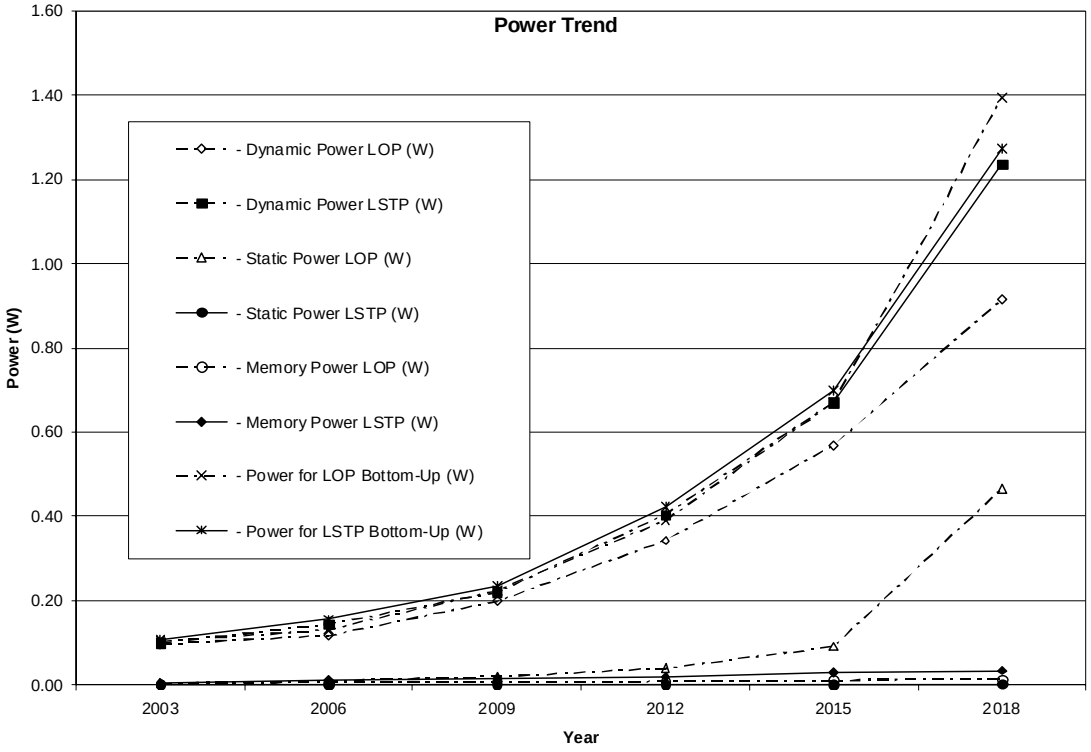

Figure 10 Total Chip Power Trend for SOC-LP PDA Application

[9] *I_{off} denotes the NMOSFET drain current at room temperature, and is the sum of the NMOS sub-threshold, gate, and junction leakage current components, as described in the PIDS chapter. Details of active capacitance density calculations, dependences on temperature and threshold, etc. may be found in the PIDS Chapter documentation and in the following supplemental file. The activity of logic blocks is fixed at 10%. The activity of memory blocks is estimated to be 0.4% based on the following analysis of large memory designs. We first assume that a memory cell contributes 2 gate capacitances of minimum size transistors for switching purposes, accounting for source/drain capacitances, contact capacitances and wiring capacitance along the bit lines. A write access requires power in the row/column decoders, word line and M bit lines, sense amplifiers and output buffers. We consider memory to be addressed with 2N bits and assume that memory power is due primarily to the column capacitances, and that $Mx2^N$ bits are accessed simultaneously out of $2^N x 2^N$ possible bits. Then $\alpha = M/2^N$ which is the ratio of accessed bit to total bits in the memory. For example, for a 16 Mbit memory, M=16 and N=12; hence α=0.4%.*

[10] *At 25°C, dynamic power dissipation dominates the total power in both the LOP and LSTP cases.*

Table 10 Power Management Gap for SOC LP-PDA

	2003	*2006*	*2009*	*2012*	*2015*	*2018*
Total LOP Dynamic Power Gap (X)	**0.0**	**0.2**	**1**	**2.4**	**4.7**	**8.1**
Total LSTP Dynamic Power Gap (X)	**0.0**	**0.4**	**1.2**	**3.00**	**5.7**	**11.4**
Total LOP Standby Power Gap (X)	**0.37**	**3.44**	**8.73**	**18.79**	**44.38**	**231.9**
Total LSTP Standby Power Gap (X)	**-0.98**	**-0.96**	**-0.90**	**-0.78**	**-0.53**	**0.10**

Table 10 shows the implied *power management gap*, i.e., the factor improvement in power management that must be achieved jointly at the levels of application, operating system, architecture, and IC design. Required power reduction factors exceed 8× for dynamic power, and can be large with respect to standby power unless the design is dominated by LSTP devices. Here, the Total Power Gap is defined as (Total Power – 0.1 W)/0.1 W (the PDA total power requirement). Similarly, the Total Standby Power Gap is defined as (Total Standby Power – 2 mW)/2m W (the PDA total standby power requirement). Negative values indicate the lack of any power management gap (i.e., existing techniques suffice).

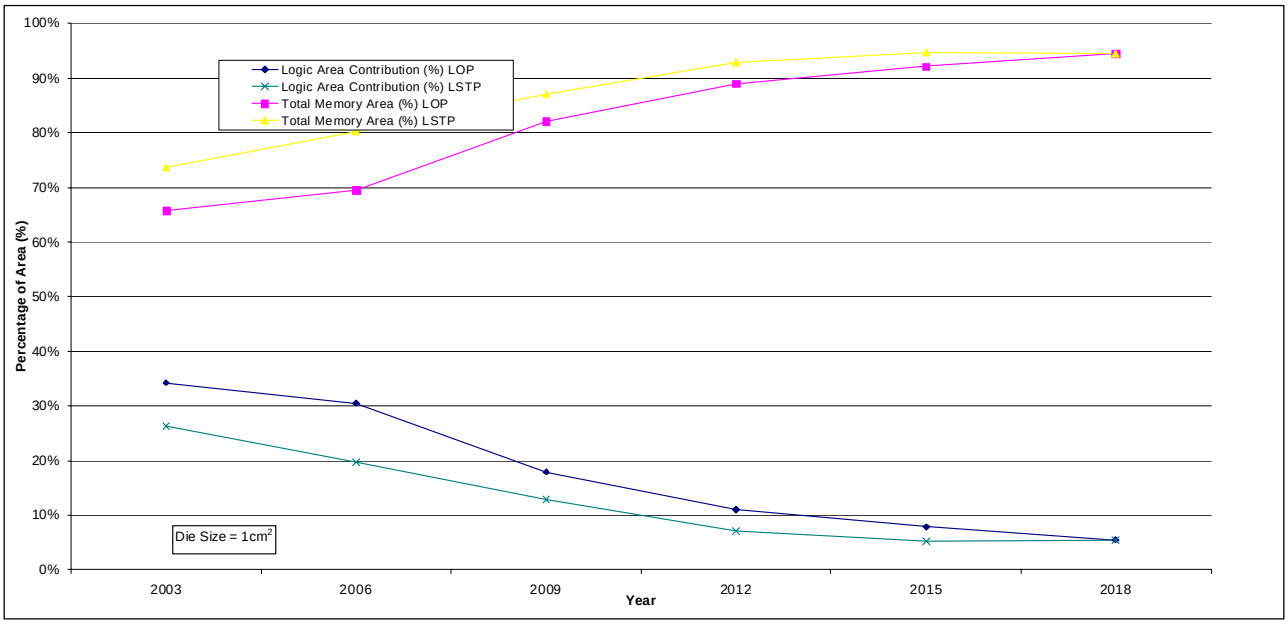

Figure 11 Power Gap Effect on Chip Composition

Figure 11 projects logic/memory composition of SOC-LP designs, assuming that chip power is constrained to 0.1W and that chip size is constrained to 100 mm^2. Memory content outstrips logic content faster with LSTP devices since their operating power is much higher than that of LOP devices. Both models indicate that chips will be asymptotically dominated by memory by 2018 without substantial improvements in power management capability. The SOC-LP PDA driver chip size is projected to grow at approximately 10% per node even though power remains flat at 0.1 W; this would lead to even more extreme memory-logic imbalances in the long-term years.

A final trend, mentioned earlier but worth revisiting, is the blurring of SOC and SIP integration choices. SIP allows exploitation of individually optimized chip technologies, and has rapidly emerged as an enabler of cost-effective mixed-technology system implementation. For example, 1) chip-laminate-chip SIP approaches lead to cost-effective DRAM-logic integration, via a thin-film laminate within the BGA package that provides top-level power, ground and clock distribution along with built-in decoupling capacitors (DRAM bare dice are flip-chip mounted on one side, and ASIC bare dice on the other); and 2) silicon-on-silicon SIP approaches enable integration of one or more RF ICs, flip-chip assembled onto a thin-film, highly resistive substrate with high-quality embedded passives. The choice between SOC and SIP is ultimately a function of system value and cost, with respect to such metrics as power, speed, area, reliability, testability, and yield. SIP-driven challenges are noted throughout the *Design*, *Test*, and *Assembly and Packaging* Chapters of this ITRS. The remainder of this chapter discusses detailed models and roadmaps for the MPU, AMS, and embedded Memory drivers.

MICROPROCESSOR (MPU) DRIVER

In high-volume custom designs, performance and manufacturing cost issues outweigh design or other non-recurring engineering (NRE) cost issues, primarily because of the large profits that these chips can potentially produce. These large profits result from very large sales volumes. Large volumes alone are neither necessary nor sufficient to warrant the custom design style, special process engineering and equipment, etc. often associated with such parts; the key is that the expected return on the combined NRE and manufacturing investment must be positive. Within the high-volume custom arena, three dominant classes today are MPUs, memory[11] and reprogrammable (e.g., field-programmable gate array (FPGA)). In this section, we focus on the MPU as one of the key system drivers for semiconductor products. MPUs use the most aggressive design styles and manufacturing technologies to achieve their goals. It is for these high-volume parts that changes to the manufacturing flow are made, new design styles and supporting tools are created (the large revenue streams can pay for new tool creation), and subtle circuits issues are uncovered (not all risks taken by designers work out). Indeed, MPUs drive the semiconductor industry with respect to integration density and design complexity, power-speed performance envelope, large-team design process efficiency, test and verification, power management, and packaged system cost. While MPUs (and high-volume custom designs in general) are extremely labor-intensive, they create new technology and automation methods (in both design and fabrication) that are leveraged by the entire industry.

The ITRS MPU driver reflects general-purpose instruction-set architectures (ISAs) that are found standalone in desktop and server systems, and embedded as cores in SOC applications. The MPU system driver is subject to market forces that have historically led to 1) emergence of standard architecture platforms and multiple generations of derivatives, 2) strong price sensitivities in the marketplace, and 3) extremely high production volumes and manufacturing cost awareness. Key elements of the MPU driver model are as follows (studies in this Chapter can be run in the *GTX tool;* MPU content is provided *GTX study*.

1. Three types of MPU—Historically, there have been three types of MPU: cost-performance (CP), reflecting "desktop", and high-performance (HP), reflecting "server" and power-connectivity-cost (PCC). As predicted in the 2001 ITRS, the increasing market acceptance of battery-limited mobile designs (often with wireless connectivity) lead to the creation of a new power-connectivity-cost (PCC) category for MPUs. At the same time, the CP segment that traditionally referred to "desktops" is now expanding to span a much larger portion of the price-performance tradeoff curve, ranging from low-end, low-cost traditional "servers" to "mobile desktops" (i.e., laptops used primarily in AC mode) and "blade" servers. As a consequence, the performance gap between the CP and HP categories is shrinking. However, there will remain a market for truly high-end servers, driving design effort disproportionate to product volume because of large margins involved. As predicted previously, the new PCC category will start taking on characteristics of high-performance, low power SOC design, with an emphasis on convenience through battery life extension and wireless connectivity. However, the larger margins and volumes of a PCC design will justify much greater design effort as compared to a traditional SOC.

2. Constant die area—Die areas are constant (140 mm^2 for CP, 310 mm^2 for HP, 70–100 mm^2 for PCC) over the course of the roadmap, and are broken down into logic, memory, and integration overhead. Integration overhead reflects the presence of white space for interblock channels, floor plan packing losses, and potentially growing tradeoff of layout density for design turnaround time. The core message, in contrast to previous ITRS models, is that power, cost and interconnect cycle latency are strong limiters of die size. To first order, additional logic content would not be efficiently usable due to package power limits, and additional memory content (e.g., larger caches, more levels of memory hierarchy integrated on-chip) would not be cost-effective beyond a certain point.[12] Furthermore, the difficulty of accurate architectural performance simulations with increasingly deeper interconnect pipelining (caused due to process scaling) will also limit die growth size.

3. Multi-core organization—MPU logic content reflects multiple processing units on-chip starting from the 130 nm node, primarily in the HP and high-end CP categories. This integrates several factors: 1) organization of recent and planned commercial MPU products (both server and desktop); 2) increasing need to reuse verification and logic design, as well as standard ISAs; 3) ISA "augmentations" in successive generations (e.g., x86, MMX and EPIC) with likely continuations for encryption, graphics and multimedia, etc.; 4) the need to enable flexible management of power at the architecture, OS

[11] *Memory is a special class of high-volume custom design because of the very high replication rate of the basic memory cells and supporting circuits. Since these cells are repeated millions of time on a chip, and millions of chips are sold, the amount of custom design for these parts is extraordinary. This has led to separate fabrication lines for DRAM devices, with some of the most careful circuit engineering needed to ensure correct operation.*

[12] *Multi-core organization and associated power efficiencies may permit slight growth in die size, but the message is still that die areas are flattening out.*

and application levels via SOC-like integration of less efficient, general-purpose processor cores with more efficient, special-purpose "helper engines"[13]; 5) the limited size of processor cores (the estimate of a *constant* 20–25 million transistors per core[14] is a conservative upper bound with respect to recent trends); and 6) the convergence of SOC and MPU design methodologies due to design productivity needs. While increasingly complex single core designs will continue for a few more years, they will compete with equivalent multi-core designs especially in the HP and high-end CP categories. During this period, the number of cores in multi-core designs is projected to double with each successive technology node.

4. Memory content—The MPU memory content is initially 512 KBytes (512 × 1024 × 9 bits) of SRAM for CP and 2 MBytes for HP in the 180 nm node. Memory content, like logic content, is projected to double with each successive technology node, not with respect to absolute time intervals (e.g., every 18 months).[15,16]

5. Layout density—Due to their high levels of system complexity and production volume, MPUs are the driver for improved layout density.[17] Thus, MPU driver sets the layout densities, and hence the transistor counts and chip sizes, stated in the Overall Roadmap Technology Characteristics. The logic and SRAM layout densities in the 2001 ITRS ORTC tables are analogous to the DRAM "A-factor," and have been calibrated to recent MPU products. Logic layout densities reflect average standard-cell gate layouts of approximately $320F^2$, where F is the minimum feature size of the technology node.[18] At 65 nm and below layout density scaling will slow due to the increased complexity of spacing constraints imposed by sub-resolution lithography techniques such as optical proximity correction (OPC) and phase shift mask (PSM). Projections show this impact to be as much as 20% at the 65 nm node. As a result, the scale factor of 0.7 will yield a density improvement of only 0.55–0.6. As noted above, the logic layout density may improve significantly with the advent of novel devices. SRAM layout densities reflect use of a 6-transistor bit cell (the fitted expression for area per bit cell in units of F^2 = 223.19F (μm) + 97.74) in MPUs, with 60% area overhead for peripheral circuitry.

6. Maximum on-chip (global) clock frequency—MPUs also drive maximum on-chip clock frequencies in the *Overall Roadmap Technology Characteristics*; these in turn drive various aspects of the Interconnect, *PIDS*, *FEP* and *Test* roadmaps. The MPU maximum on-chip clock frequency has historically increased by a factor of 2 per generation. Of this, approximately 1.4× has been from device scaling (which runs into t_{ox} and other limits); the other 1.4× has been from reduction in number of logic stages in a pipeline stage (e.g., equivalent of 32 fanout-of-4 inverter (FO4 INV) delays[19] at 180 nm, and 24–26 FO4 INV delays at 130 nm). There are several reasons why this historical trend will not continue: 1) well-formed clock pulses cannot be generated with period below 6–8 FO4 INV delays; 2) there is increased overhead (diminishing returns) in pipelining (2–3 FO4 INV delays per flip-flop, 1–1.5 FO4 INV delays per pulse-mode latch); 3) thermal envelopes imposed by affordable packaging discourage very deep pipelining, and 4) architectural and circuit innovations will increasingly counter the impact of worsening interconnect RCs (relative to devices) rather than

---

[13] A "helper engine" is a form of "processing core" for graphics, encryption, signal processing, etc. The trend is toward architectures that contain more special-purpose, and less general-purpose, logic.

[14] The CP core has 20 million transistors, and the HP core has 25 million transistors. The difference allows for more aggressive microarchitectural enhancements (trace caching, various prediction mechanisms, etc.) and other performance support.

[15] The doubling of logic and memory content with each technology node, rather than with each 18- or 24-month time interval, is due to essentially constant layout densities for logic and SRAM, as well as conformance with other parts of the ITRS. Specifically, the ITRS remains planar CMOS-centric, there is evidence that non-planar "emerging research devices" are moving into development, possibly as early as for the 45 nm node (VLSI Symp'03). Adoption of such novel device architectures would allow improvements of layout densities beyond what is afforded by scaling alone.

[16] Deviation from the given model will likely occur around the 90 nm node with adoption of denser embedded memories (eDRAM). Adoption of eDRAM, and integrated on-chip L3 cache, will respectively increase the on-chip memory density and memory transistor count by factors of approximately 3 from the given values. While this will significantly boost transistor counts, it is not projected to significantly affect the chip size or total chip power roadmap. Adoption of eDRAM will also depend strongly on compatibility with logic processes (notably the limited process window that arises from scaling of oxide thickness), the size and partitioning of memory within the individual product architecture, and density-performance-cost sensitivities.

[17] ASIC/SOC and MPU system driver products have access to similar processes, as forecast since the 1999 ITRS. This reflects emergence of pure-play foundry models, and means that fabric layout densities (SRAM, logic) are the same for SOC and MPU. However, MPUs drive high density and high performance, while SOCs drive high integration, low cost, and low power.

[18] A 2-input NAND gate is assumed to lay out in an 8×4 standard cell, where the dimensions are in units of contacted local metal pitch (MP = 3.16 x F). In other words, the average gate occupies 32 x $(3.16)^2$ = $320F^2$. For both semi-custom (ASIC/SOC) and full-custom (MPU) design methodologies, an overhead of 100% is assumed.

[19] A FO4 INV delay is defined to be the delay of an inverter driving a load equal to 4 times its own input capacitance (with no local interconnect). This is equivalent to roughly 14 times the CV/I device delay metric that is used in the PIDS Chapter to track device performance. An explanation of the FO4 INV delay model used in the 2003 ITRS is provided in supplemental material.

contribute directly to frequency improvements. The 2003 ITRS MPU model continues the historic rate of advance for maximum on-chip global clock frequencies, but flattens the clock period at 12 FO4 INV delays during the 90 nm node (*a plot of historical MPU clock period data is provided*). This is a change from the projection of 16 FO4 INV delays made in 2001; projections based on circuit and architecture advances made since the 2001 ITRS indicate that the minimum achievable logic depth is closer to 10–12 F04. The message remains that after the 90 nm node, clock frequencies will advance only with device performance in the absence of novel circuit and architectural approaches.[20]

MPU EVOLUTION

An emerging "centralized processing" context integrates 1) centralized computing servers that provide high-performance computing via traditional MPUs (this driver), and 2) *interface remedial processors* that provide power-efficient basic computing via, e.g., SOC integration of RF, analog/mixed-signal, and digital functions within a wireless handheld multimedia platform (refer to the low-power SOC PDA model, above). Key contexts for the future evolution of the traditional MPU are with respect to design productivity, power management, multi-core organization, I/O bandwidth, and circuit and process technology.

Design productivity—The complexity and cost of design and verification of MPU products has rapidly increased to the point where thousands of engineer-years (and a design team of hundreds) are devoted to a single design, yet processors reach market with hundreds of bugs. This is leading to a decreasing emphasis on the use of heavy customization and fancy circuit families resulting in an increasing use of design automation such as logic synthesis and automatic circuit tuning. The resulting productivity increases have allowed processor development schedules and team sizes to flatten out. Improvements in design tools for analysis for timing, noise, power and electrical rules checking have also contributed to a steady increase in design quality.

Power management—Power dissipation limits of packaging (despite being estimated to reach 200 W/cm^2 by the end of the ITRS) cannot continue to support high supply voltages (historically scaling at 0.85× per generation instead of 0.7× ideal scaling) and frequencies (historically scaling by 2× per generation instead of 1.4× ideal scaling).[21] Past clock frequency trends in the MPU system driver have been interpreted as future CMOS device performance (switching speed) requirements that lead to large off-currents and extremely thin gate oxides, as specified in the PIDS Chapter. Given such devices, MPUs that simply continue existing circuit and architecture techniques would exceed package power limits by factors of nearly 4× by the end of the ITRS; alternatively, MPU logic content and/or logic activity would need to decrease to match package constraints. Portable and low-power embedded contexts have more stringent power limits, and will encounter such obstacles earlier. Last, power efficiencies (e.g., GOPS/mW) are up to four orders of magnitude greater for direct-mapped hardware than for general-purpose MPUs; this gap is increasing. As a result, traditional processing cores will face competition from application-specific or reconfigurable processing engines for space on future SOC-like MPUs.

Multi-core organization—In an MPU with multiple cores per die, the cores can be 1) smaller and faster to counter global interconnect scaling, and 2) optimized for reuse across multiple applications and configurations. Multi-core architectures allow power savings as well as the use of redundancy to improve manufacturing yield.[22] Organization of the MPU model also permits increasing amounts of the memory hierarchy on chip (consistent with processor-in-memory, or large on-chip eDRAM L3 starting in the 90 nm generation). Higher memory content can, if only in a relatively trivial way, afford better "control" of leakage and total chip power.

---

[20] Unlike the ITRS clock frequency models used through 2000 (refer to Fisher/Nesbitt 1999), the 2003 model does not have any local or global interconnect component in its prototypical "critical path". This is because local interconnect delays are negligible, and scale with device performance. Furthermore, buffered global interconnect does not contribute to the minimum clock period since long global interconnects are pipelined– i.e., the clock frequency is determined primarily by the time needed to complete local computation loops, not by the time needed for global communication. Pipelining of global interconnects will become standard as the number of clock cycles required to signal cross-chip continues to increase beyond 1. "Marketing" emphases for MPUs necessarily shift from "frequency" to "throughput" or "utility".

[21] To maintain reasonable packaging cost, package pin counts and bump pitches for flip-chip are required to advance at a slower rate than integration densities (refer to the Assembly and Packaging Chapter). This increases pressure on design technology to manage larger wakeup and operational currents and larger supply voltage IR drops; power management problems are also passed to the architecture, OS and application levels of the system design.

[22] Replication enables power savings through lowering of frequency and V_{dd} while maintaining throughput (e.g., two cores running at half the frequency and half the supply voltage will save a factor of 4 in CV^2f dynamic capacitive power, versus the "equivalent" single core). (Possibly, this could allow future increases in die size.) More generally, overheads of time-multiplexing of resources can be avoided, and the architecture and design focus can shift to better use of area than memory. Redundancy-based yield improvement occurs if, e.g., a die with k-1 instead of k functional cores is still useful.

Evolutionary microarchitecture changes (super-pipelining, super-scalar, predictive methods) appear to be running out of steam. ("*Pollack's Rule*" observes that in a given process technology, a new microarchitecture occupies 2–3× the area of the old (previous-generation) microarchitecture, while providing only 1.4–1.6× the performance.) Thus, more multithreading support will emerge for parallel processing, as well as more complex "hardwired" functions and/or specialized engines for networking, graphics, security, etc. Flexibility-efficiency tradeoff points shift away from general-purpose processing.

Input/output bandwidth—In MPU systems, I/O pins are mainly used to connect to memory, both high-level cache memory and main system memory. Increased processor performance has been pushing I/O bandwidth requirements. The highest-bandwidth port has traditionally been used for L2 or L3 cache, but recent designs are starting to integrate the memory controller on the processor die to reduce memory latency. These direct memory interfaces require more I/O bandwidth than the cache interface. In addition to the memory interface, many designs are replacing the system bus with high-speed point-to-point interfaces. These interfaces require much faster I/O design, exceeding Gbit/s rates. While serial links have achieved these rates for a while, integrating a large number of these I/O on a single chip is still challenging for design (each circuit must be very low power), test (need to have a tester that can run this fast) and packaging (packages must act as balanced transmission lines, including the connection to the chip and the board).

Circuit and process technology—Parametric yield ($/wafer after bin-sorting) is severely threatened by the growing process variability implicit in feature size and device architecture roadmaps (*Lithography* and *PIDS*), including thinner and less reliable gate oxides, subwavelength optical lithography requiring aggressive reticle enhancement, and increased vulnerability to atomic-scale process variability (e.g., implant). This will require more intervention at the circuit and architecture design levels. Circuit design use of dynamic circuits, while attractive for performance in lower-frequency or clock-gated regimes, may be limited by noise margin and power dissipation concerns; less pass gate logic will be used due to body effect. Error-correction for single-event upset (SEU) in logic will increase, as will the use of redundancy and reconfigurability to compensate for yield loss. Design technology will also evolve to enable consideration of process variation during design and analysis and its impact on parametric yield (bin-splits). The need for power management will require a combination of techniques from several component technologies:

1. application-, OS- and architecture-level optimizations including parallelism and adaptive voltage and frequency scaling
2. process innovations including increased use of SOI
3. circuit design techniques including the *simultaneous* use of multi-V_{th}, multi-V_{dd}, minimum-energy sizing under throughput constraints, and multi-domain clock gating and scheduling
4. novel devices that decrease leakage

MPU CHALLENGES

The MPU driver strongly affects design and test technologies (distributed/collaborative design process, verification, at-speed test, tool capacity, power management), as well as device (off-current), lithography/FEP/interconnect (variability) and packaging (power dissipation and current delivery). The most daunting challenges are:

- *design and verification productivity* (e.g., total design cost, number of bug escapes) (Design)
- *power management and delivery* (e.g., GOPS per mW) (Design, PIDS, Assembly and Packaging)
- *parametric yield at volume production* (Lithography, PIDS, FEP, Design)

MIXED-SIGNAL DRIVER

AMS chips are those that at least partially deal with input signals whose precise values matter. This broad class includes RF, analog, analog to digital and digital to analog conversion, and, more recently, a large number of mixed-signal chips where at least part of the chip design needs to measure signals with high precision. These chips have very different design and process technology demands than digital circuits. While technology scaling is always desirable for digital circuits due to reduced power, area and delay, it is not necessarily helpful for analog circuits since dealing with precision requirements or signals from a fixed voltage range is more difficult with scaled voltage supplies. Thus, scaling of analog circuits into new technologies is a difficult challenge. In general, AMS circuits (e.g., RF and analog design styles) and process technologies (e.g., silicon-germanium, embedded passives) present severe challenges to cost-effective CMOS integration.

The need for precision also affects tool requirements for analog design. Digital circuit design creates a set of rules that allow logic gates to function correctly: as long as these rules are followed, precise calculation of exact signal values is not needed. Analog designers, on the other hand, must be concerned with a large number of "second-order effects" to obtain the required precision. Relevant issues include coupling (capacitance, inductance, resistance and substrate affecting the integrity of signals and supply voltages) and asymmetries (local variation of implantation, alignment, etching, and other fabrication steps all affect the predictability of the electrical performance). Analysis tools for these issues are mostly in place but require expert users; synthesis tools are preliminary. Manufacturing test for AMS circuits remains essentially unsolved.

Most analog and RF circuitry in today's high-volume applications is part of SOCs. The economic regime of a mainstream product is usually highly competitive: it has a high production volume, and hence a high level of R&D investment by which its technology requirements can drive mixed-signal technology as a whole. Mobile communication platforms are the highest volume circuits driving the needs of mixed signal circuits. Therefore we restrict here to the needs of those circuits. Even here, when formulating an analog and mixed-signal (AMS) roadmap, simplification is necessary because there are too many different circuits and architectures. We restrict our discussion to four basic analog circuits:

1. Low-noise amplifier (LNA)
2. Voltage-controlled oscillator (VCO)
3. Power amplifier (PA)
4. Analog to digital converter (ADC)

The design and process technology used to build these basic analog circuits also determines the performance of many other mixed-signal circuits. Thus, the performance of these four circuits, as described by figures of merit (FoMs), is a good basis for a mixed-signal roadmap.

The following discussion develops these FoMs in detail. Unless otherwise noted, all parameters (e.g., gain *G*) are given as absolute values and not on a decibel scale. We also avoid preferences for specific solutions to given design problems; indeed, we have sought to be as open as possible to different types of solutions since unexpected solutions have often helped to overcome barriers. (Competition, e.g., between alternative solutions, is a good driving force for all types of advances related to technology roadmapping.) Any given type of circuit will have different requirements depending on its purposes. Therefore, certain performance indicators can be contradictory in different applications.[23] To avoid such situations, we adjust the figures of merit to the analog and RF needs of a mobile communication platform. Last, we evaluate the dependence of the FoMs on device parameters, so that circuit design requirements can lead to specific device and process technology specifications. Extrapolations are proposed that lead on the one hand to a significant advance of analog circuit performance and on the other hand to realistic and feasible technology advances. These parameters are given in the AMS, RF-Transceiver, and Power Amplifier Tables of the *RF and Analog/Mixed-signal Technologies for Wireless Communications* section of the PIDS Chapter.

LOW-NOISE AMPLIFIER (LNA)

Digital processing systems require interfaces to the analog world. Prominent examples of these interfaces are transmission media in wired or wireless communication. The LNA amplifies the input signal to a level that makes further signal processing insensitive to noise. The key performance issue for an LNA is to deliver the undistorted but amplified signal to downstream signal processing units without adding further noise.

LNA applications (GSM, CDMA, W-LAN, GPS, Bluetooth, etc.) operate in many frequency bands. The operating frequency and, in some cases, the operating bandwidth of the LNA will impact the maximum achievable performance; nonlinearity must also be considered to meet the specifications of many applications. These parameters must be included in the FoM. On the other hand, different systems may not be directly comparable, and in fact have diverging requirements. For example, very wide bandwidth is needed for high-performance wired applications, but this increases power consumption. Low power consumption is an important design attribute for low-bandwidth wireless applications. For wide-bandwidth systems, bandwidth may be more important than linearity to describe the performance of an LNA. To avoid contradictory design constraints, we focus on the *wireless* communication context.

[23] *Certain cases of application are omitted for the sake of simplicity, and arguments are given for the cases selected. In many cases, we have limited our considerations to CMOS since it is the prime technological driving force and in most cases the most important technology. Alternative solutions (especially other device families) and their relevance will be discussed for some cases, as well as at the end of this section.*

The linearity of a low noise amplifier can be described by the output referenced third order intercept point ($OIP3 = G \times IIP3$ where G is the gain and $IIP3$ is the input referenced third order intercept point). A parameter determining the minimum signal that is correctly amplified by a LNA is directly given by the noise figure of the amplifier, NF. However, (NF-1) is a better measure of the contribution of the amplifier to the total noise, since it allows the ratio between the noise of the amplifier $N_{amplifier}$ and the noise already present at the input N_{input} to be directly evaluated. These two performance figures can be combined with the total power consumption P. The resulting figure of merit captures the dynamic range of an amplifier versus the necessary DC power. For roadmapping purposes it is preferable to have a performance measure that is independent of frequency and thus independent of the specific application. This can be achieved by assuming that the LNA is formed by a single amplification stage, so that the FoM scales linearly with operating frequency f. With these approximations and assumptions, a figure of merit (FoM_{LNA}) for LNAs is defined:

$$FoM_{LNA} = \frac{G \cdot IIP3 \cdot f}{(NF - 1) \cdot P} \quad (1)$$

Making further simplifying assumptions, and neglecting "design intelligence", the evolution of the FoM with technology scaling can be extrapolated.[24] Future trends of relevant device parameters for LNA design, including maximum oscillation frequency f_{max}, quality of inductors, inner gain of the MOSFETs ($g_m/g_{ds}|_{L\_min}$), and RF supply voltages are shown in the AMS Table and RF-Transceiver Table in the *RF and Analog/Mixed-signal Technologies for Wireless Communications* section of the PIDS Chapter. The evolution of the FoM from recent best-in-class published LNAs shows a clear trend towards better performance for smaller device dimensions; this is in good agreement with the increase in the quality of the devices needed for LNA design. Extrapolating these data into the future, an estimate of future progress in LNA design is obtained as shown in Table 11.

VOLTAGE CONTROL OSCILLATOR (VCO)

Another key component of RF signal processing systems is the VCO. The VCO is the key part of a phase-locked loop (PLL), which synchronizes communication between an integrated circuit and the outside world in high-bandwidth and/or high-frequency applications. The key design objectives for VCOs are to minimize the timing jitter of the generated waveform (or, equivalently, the phase noise) and to minimize the power consumption. From these parameters a figure of merit (FoM_{VCO}) is defined:

$$FoM_{VCO} = \left(\frac{f_0}{\Delta f}\right)^2 \frac{1}{L\{\Delta f\} \cdot P} \quad (2)$$

Here, f_0 is the oscillation frequency, $L\{\Delta f\}$ is the phase noise power spectral density measured at a frequency offset Δf from f_0 and taken relative to the carrier power, and P is the total power consumption.

This definition does not contain the absolute value of the operating frequency since there is no clear correlation between the operating frequency and the figure of merit. The definition also neglects the tuning range of the VCO since the necessary tuning range strongly depends on the application. In this tuning range, FoM_{VCO} should be evaluated at the frequency where phase noise is maximal.

Phase noise is mainly determined by thermal noise of the active and passive components in the VCO, the quality factor of the LC tank, the amplitude of the oscillation, and—close to the carrier frequency—by the 1/f noise of the active components of the VCO. FoM_{VCO} is roughly proportional to the overdrive voltage of the active elements in the VCO, inversely proportional to VDD, and proportional to the square of the quality factor of the LC tank. The value of the chosen overdrive voltage is a compromise between minimization of the contribution of 1/f noise and keeping the amplitude of the oscillation sufficiently high. In this way, FoM_{VCO} is linked to technology development. Based on a prediction of the relevant device parameters for future technology nodes (see the RF-Transceiver Table *RF and Analog/Mixed-signal Technologies for Wireless Communications* section of the PIDS Chapter.), an extrapolation of the VCO FoM for future technology nodes is given in Table 11. The increasing trend of FOM_{VCO} corresponds to the evolution of FoMs from recent best in class published accounts for VCOs. The FoMs are in good agreement with the data of the best available devices needed for VCO design in these technologies.

[24] *R. Brederlow, S. Donnay, J. Sauerer, M. Vertregt, P. Wambacq, and W. Weber, "A Mixed-signal Design Roadmap for the International Technology Roadmap for Semiconductors (ITRS)," IEEE Design and Test, December 2001.*

POWER AMPLIFIER (PA)

Power amplifiers are key components in the transmission path of wired or wireless communication systems. They deliver the transmission power required for transmitting information off-chip with high linearity to minimize adjacent channel power. For battery-operated applications in particular, minimum DC power at a given output power is required.

CMOS PAs are in a relatively infant stage but will have advantages in system-on-chip applications where relatively small transmit power is needed. For discrete PAs with higher transmit power, e.g. for cellular basestation applications, other technologies like bipolar or compound semiconductor technologies have advantages (*RF and Analog/Mixed-signal Technologies for Wireless Communications* section of the PIDS Chapter). To remain under the SOC umbrella, only CMOS power amplifiers will be discussed here. System-in-Package options may use alternative technologies depending on system performance and cost.

To establish a performance figure of merit, several key parameters must be taken into account. These include output power P_{out}, power gain G, carrier frequency f, linearity (in terms of IIP3), and power-added-efficiency (PAE). Unfortunately, linearity strongly depends on the operating class of the amplifiers, making it difficult to compare amplifiers of different classes. To remain independent of the design approach and the specifications of different applications, we omit this parameter in our figure of merit. To compensate for the 20 dB/decade roll-off[25] of the PA's RF-gain, a factor of f^2 is included into the figure of merit. This results in:

$$FoM_{PA} = P_{out} \cdot G \cdot PAE \cdot f^2 \qquad (3)$$

Finally, restricting to the simplest PA architecture (class A operation)[26] and making further simplifications enables correlation between the FoM and device parameters.[27] The key device parameters are seen to be the quality factor of the available inductors and f_{max}; values for these parameters are mapped in the PA Table of the PIDS Chapter. FoMs of best-in-class CMOS PAs have increased by approximately a factor of two per technology node in recent years, strongly correlated with progress in active and passive device parameters. From required device parameters for future technology nodes (see the Power Amplifier Tables in *RF and Analog/Mixed-signal Technologies for Wireless Communications* section of the PIDS Chapter), we can deduce requirements for future PA FoM values, as shown in Table 11.

ANALOG-TO-DIGITAL CONVERTER (ADC)

Digital processing systems have interfaces to the analog world: audio and video interfaces, interfaces to magnetic and optical storage media, and interfaces to wired or wireless transmission media. The analog world meets digital processing at the analog-to-digital converter (ADC), where continuous-time and continuous-amplitude analog signals are converted to discrete-time (sampled) and discrete-amplitude (quantized). The ADC is therefore a useful vehicle for identifying advantages and limitations of future technologies with respect to system integration; it is also the most prominent and widely used mixed-signal circuit in today's integrated mixed-signal circuit design.

The main specification parameters of an ADC relate to sampling and quantization. The resolution of the converter, i.e., the number of quantization levels, is 2^n where n is the "number of bits" of the converter. This parameter also defines the maximum signal to noise level $SNR = n \cdot 6.02 + 1.76 \quad [dB]$. The sampling rate of the converter, i.e., the number of n-wide samples quantized per unit time, is related to the bandwidth that needs to be converted and to the power consumption required for reaching these performance points. The Shannon/Nyquist criterion states that a signal can be reconstructed whenever the sample rate exceeds twice the converted bandwidth: $f_{sample} > 2 \times BW$.

To yield insight into the potential of future technology nodes, the ADC FoM should combine dynamic range, sample rate f_{sample} and power consumption P. However, these nominal parameters do not give accurate insight into the effective performance of the converter; a better basis is the effective performance extracted from measured data. Dynamic range is extracted from low frequency signal-to-noise-and-distortion ($SINAD_0$) measurement minus quantization error (both values in dB). From $SINAD_0$ an "effective number of bits" can be derived as $ENOB_0 = (SINAD_0 - 1.76)/6.02$. Then, the

[25] *Most CMOS PAs are currently operated in this regime. Using DC-gain for applications far below* f_t *would result in a slightly increased slope.*
[26] *R. Brederlow, S. Donnay, J. Sauerer, M. Vertregt, P. Wambacq, and W. Weber, "A Mixed-signal Design Roadmap for the International Technology Roadmap for Semiconductors (ITRS)," IEEE Design and Test, December 2001.*
[27] *R. Brederlow, S. Donnay, J. Sauerer, M. Vertregt, P. Wambacq, and W. Weber, "A Mixed-signal Design Roadmap for the International Technology Roadmap for Semiconductors (ITRS)," IEEE Design and Test, December 2001.*

sample rate may be replaced by twice the effective resolution bandwidth (2×ERBW) if it has a lower value, to establish a link with the Nyquist criterion:

$$FoM_{ADC} = \frac{\left(2^{ENOB_0}\right) \times \min(\{f_{sample}\}, \{2 \times ERBW\})}{P} \quad (4)$$

For ADCs, the relationship between FoM and technology parameters is strongly dependent on the particular converter architecture and circuits used. The complexity and diversity of ADC designs makes it nearly impossible to come up with a direct relationship, as was possible for the basic RF circuits. Nevertheless, some general considerations regarding the parameters in the FoM are proposed,[28] in some cases, it is possible to determine performance requirements of the design from the performance requirements of a critical subcircuit. The device parameters relevant for the different ADC designs are stated in the AMS Table of the PIDS Chapter. The trend in recent years shows that the ADC FoM improves by approximately a factor of 2 every three years. Taking increasing design intelligence into account, these past improvements are in good agreement with improvements in analog device parameters. Current best-in-class is approximately 1600 [giga-conversion-steps per second and watt] for stand-alone CMOS/BiCMOS, and approximately 800 [giga-conversion-steps per second and watt] for embedded CMOS. Expected future values for the ADC FoM are shown in Table 11. Major advances in design are needed to maintain performance increases for ADCs in the face of decreased voltage signal swings and supplies. In the long run, fundamental physical limitations (thermal noise) may block further improvement of the ADC FoM.

Table 11 Projected Mixed-Signal Figures of Merit for Four Circuit Types.

Year of Production	*2003*	*2006*	*2009*	*2012*	*2015*	*2018*	*Driver*
RF-CMOS ½ Pitch	*130*	*90*	*65*	*45*	*32*	*22*	
FoM_{LNA} [GHz]	*40*	*80*	*160*	*200-400*	*250–500*	*300–600*	*RF-Transceiver PIDS Table*
FoM_{VCO} [1/J] 10^{22}	*0.7*	*0.9*	*1.1*	*1.9*	*2*	*2.4*	*RF-Transceiver PIDS Table*
FoM_{PA} [W•GHz^2] 10^4	*6*	*12*	*24*	*40–50*	*80–90*	*100–130*	*PA PIDS Table*
FoM_{ADC} [GHz/W]10^3	*0.8*	*1.2*	*1.6–2.5*	*2.5–5*	*4–10*	*6–20*	*AMS PIDS Table*

MIXED-SIGNAL EVOLUTION

Evolution of the mixed-signal driver, including its scope of application, is completely determined by the interplay between cost and performance. The above figures of merit measure mixed-signal *performance*. However, *cost* of production is also a critical issue for practical deployment of AMS circuits. Together, cost and performance determine the sufficiency of given technology trends relative to existing applications, as well as the potential of given technologies to enable and address entirely new applications.

Cost estimation—Unlike high-volume digital products where cost is mostly determined by chip area, in mixed-signal designs area is only one of several cost factors. The area of analog circuits in an SOC is typically in the range of 5–30%; economic forces to reduce mixed-signal area are therefore not as strong as for logic or memory. Related considerations include:

- analog area can sometimes be reduced by shifting the partitioning of a system between analog and digital parts (e.g. auto-calibration of A-to-D converters)
- process complexity is increased by introducing high-performance analog devices, so that solutions can have less area but greater total cost
- technology choices can impact design cost by introducing greater risk of multiple hardware passes (tapeout iterations)
- manufacturing cost can also be impacted via parametric yield sensitivities
- a SIP solution with multiple die (e.g., large, low-cost digital and small, high-performance analog) can be cheaper than a single SOC solution

Such considerations make cost estimation very difficult for mixed-signal designs. We may attempt to quantify mixed-signal cost by first restricting our attention to high-performance applications, since these also drive technology demands.

[28] *R. Brederlow, S. Donnay, J. Sauerer, M. Vertregt, P. Wambacq, and W. Weber, "A Mixed-signal Design Roadmap for the International Technology Roadmap for Semiconductors (ITRS)," IEEE Design and Test, December 2001.*

Next, we note that analog features are embodied as high-performance passives or analog transistors, and that area can be taken as a proxy for cost.[29] Since scaling of transistors is driven by the need to improve density of the digital parts of a system, analog transistors can simply follow, thus rendering it unnecessary to specifically address their layout density. At the same time, total area in most current AMS designs is determined by embedded passives; their area consumption dominates the cost of the mixed-signal part of a system. Therefore, the AMS Table, FR-Transceiver Table and PA Table of the PIDS Chapter set a roadmap of layout density for on-chip passive devices that is necessary to improve the cost/performance ratio of high-performance mixed-signal designs.

Estimation of technology sufficiency—Figure 12 shows ADC requirements for recent applications in terms of a power/performance relationship. Under conditions of constant performance (resolution × bandwidth), a constant power consumption is represented by a straight line with slope –1. Increasing performance—which is achievable with better technology or circuit design—is equivalent to a shift of the power consumption lines toward the upper right. The data show a very slowly moving technological "barrier-line" (Table 11) for ADCs for a power consumption of 1W (Figure 12). Most of today's ADC technologies (silicon, SiGe, and III-V compound semiconductor technologies and their hybrids) lie below the 1W barrier-line, and near-term solutions for moving the barrier-line more rapidly are unknown.

While the rate of improvement in ADC performance has been adequate for handset applications, this is clearly not the case for applications such as digital linearization of GSM base-stations, or handheld/mobile high-data rate digital video applications. For example, a multi-carrier GSM base-station with a typical setup of 32 carriers requires over 80dB of dynamic range. Implementing digital linearization in such a base-station with a 25 MHz transmitter band requires ADCs that have sampling rates of 300 MHz and 14 bits of resolution. According to Table 11 and assuming progress at recent rates, it will be perhaps until after 2010 before ADCs with such performance are manufactured in volume. While system designers would like to have such ADCs now, silicon and SiGe technologies have the necessary bit resolution (large numbers of devices per unit area) but not the speed; on the other hand, III-V compound semiconductor technologies have the speed but not the bit resolution. This motivates consideration of solutions that potentially increase the rate of ADC improvement at reasonable costs—e.g., use of compound semiconductors for their speed (perhaps combinations of HBTs, HEMTs, and resonant tunneling diodes), and hybrids of both CMOS and compound semiconductor technologies. The challenge for compound semiconductors is to increase the number of devices per unit area and to be co-integrated with CMOS processing. This is discussed in greater detail in the new *RF and Analog/Mixed-signal Technologies for Wireless Communications* section of the PIDS Chapter.

Enabling new applications—For a given product, the usual strategy to increase unit shipments is to reduce cost while increasing product performance. However, this is not the only driver for the semiconductor business, especially for products that include mixed-signal parts. Rather, improving technology and design performance enables *new* applications (comparable to the realization of the mobile handset in recent years), thus pushing the semiconductor industry into new markets. Analysis of mixed-signal designs as in Figure 12 can also be used to estimate design needs and design feasibility for future applications and new markets. We see that increasing performance is equivalent to the ability to develop new products that need higher performance or lower power consumption than is available in today's technologies. Alternatively, when specifications of a new product are known, one can estimate the technology needed to fulfill these specifications, and/or the timeframe in which the semiconductor industry will be able to build that product with acceptable cost and performance. In this way, the FoM concept can be used to evaluate the feasibility and the market of potential new mixed-signal products. The ability to build high performance mixed-signal circuitry at low cost will continuously drive the semiconductor industry into such new products and markets.

[29] *In analog designs, power consumption is often proportional to area—and since power is included in all four figures of merit, we have already implicitly considered area and cost criteria. Nonetheless, area requirements should be stated explicitly in a roadmap.*

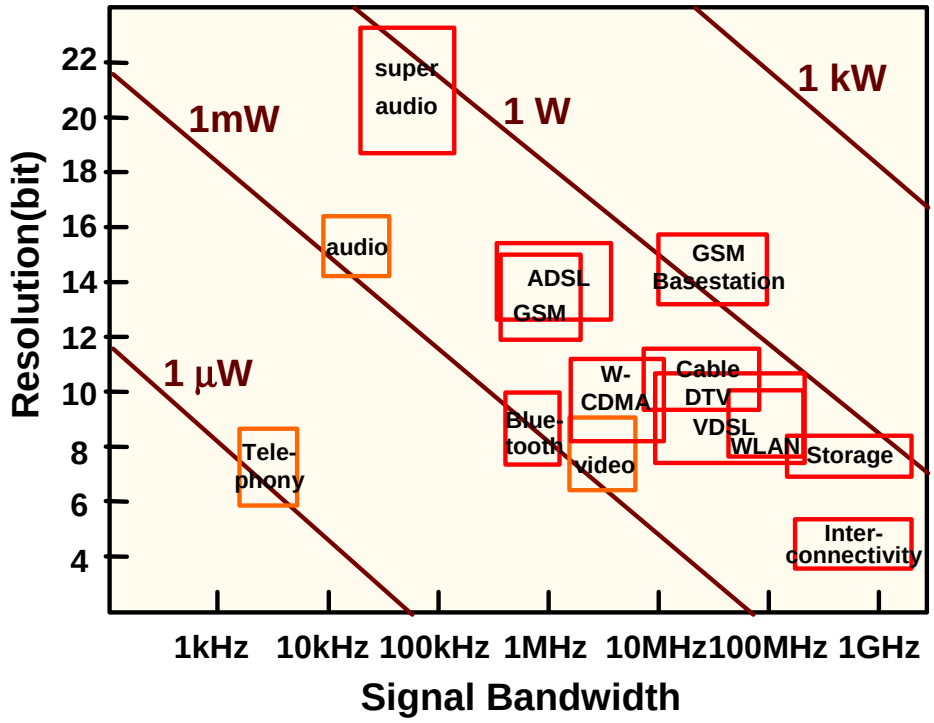

Figure 12 Recent ADC Performance Needs for Important Product Classes

MIXED-SIGNAL CHALLENGES

For most of today's mixed-signal designs—and particularly in classical analog design—the processed signal is represented by a voltage difference, so that the supply voltage determines the maximum signal. Decreasing supplies—a consequence of constant-field scaling—means decreasing the maximum achievable signal level. This has a strong impact on mixed-signal product development for SOC solutions. Typical development time for new mixed-signal parts is much longer than for digital and memory parts; sheer lack of design resources thus becomes another key challenge. An ideal design process would reuse existing mixed-signal designs and adjust parameters to meet interface specifications between a given SOC and the outside world, but such reuse depends on a second type of MOSFET that does not scale its maximum operating voltage. This has led to the specification in the PIDS Chapter of a mixed-signal CMOS transistor that uses a higher analog supply voltage and stays unchanged across multiple digital technology generations. Even with such a device, voltage reduction and development time of analog circuit blocks are major obstacles to low-cost and efficient scaling of mixed-signal functions. In summary, the most daunting mixed-signal challenges are:

- *decreasing supply voltage*, with needs including current-mode circuits, charge pumps for voltage enhancement, and thorough optimization of voltage levels in standard-cell circuits (PIDS, Design)
- *increasing relative parametric variations*, with needs including active mismatch compensation, and tradeoffs of speed versus resolution in product definition (PIDS, FEP, Lithography, Design)
- *increasing numbers of analog transistors per chip*, with needs including faster processing speed and improved convergence of mixed-signal simulation tools (Modeling and Simulation, Design)
- *increasing processing speed (carrier or clock frequencies)*, with needs including more accurate modeling of devices and interconnects, as well as test capability and package- and system-level integration (Test, Assembly and Packaging, Modeling and Simulation)
- *increasing crosstalk* arising from SOC integration, with needs including more accurate modeling of parasitics, fully differential design for RF circuits, as well as technology measures outlined in the PIDS Chapter (PIDS, Modeling and Simulation, Design)
- *shortage of design skills and productivity* arising from lack of training and poor automation, with needs including education and basic design tools research (Design)

EMBEDDED MEMORY DRIVER

SOC designs contain an increasing number and variety of embedded RAM, ROM, and register file memories. Interconnect and IO bandwidths, design productivity, and system power limits all point to a continuing trend of high levels of memory integration in microelectronic systems. Driving applications for embedded memory technology include code storage in reconfigurable applications (e.g., automotive), data storage in smart or memory cards, and the high memory content and high performance logic found in gaming or mass storage systems.

The balance between logic and memory content reflects overall system cost, power and IO constraints, hardware-software organization, and overall system and memory hierarchy. With respect to cost, the device performance and added mask levels of monolithic logic-memory integration must be balanced against chip-laminate-chip or other system-in-package (SIP) integration alternatives. Levels of logic-memory integration will also reflect tradeoffs in hardware-software partitioning (e.g., software is more flexible, but must be booted and consumes more area) as well as code-data balance (e.g., software must be available to fill code memory, and both non-volatility and applications must be present for data memory). IO pin count and signaling speeds determine how system organization trades off bandwidth versus storage: 1) memory access can be made faster at the cost of peripheral overhead by organizing memory in higher or lower bank groups; and 2) access speed also depends on how pin count and circuit complexity are balanced between high speed low pin count connections or higher pin count lower speed connections.

Memory hierarchy is crucial in matching processor speed requirements to memory access capabilities. This fact is well known in the traditional processor architecture domain and has led to the introduction of several layers of hardware-controlled caches between “main” memory and foreground memory (e.g. register files) in the processor core. At each layer, typically one physical cache memory is present. However, the choice of hierarchy also has strong implications for power. Conventional architectures increase performance largely at the cost of energy-inefficient control overheads, e.g., prediction/history mechanisms and extra buffers that are included around highly associative caches. From the system point of view, the embedded multimedia and communication applications that are dominant on portable devices can profit more from software-controlled and distributed memory hierarchies. Different layers of the memory hierarchy also require highly different access modes and internal partitionings. The use of page/burst/interleaving modes and the physical partitioning in banks, subarrays, divided word/bitlines must in general be optimized per layer. Increasingly dominant leakage power constraints also lead to more heterogeneous memory hierarchies.

Scaling presents a number of challenges to embedded memory fabrics. At the circuit level, amplifier sense margins for SRAM, and decreased I_{on} drive currents for DRAM, are two clear challenges. Smaller feature sizes imply greater impact of variability, e.g., with fewer dopants per device. With larger numbers of devices integrated into a single product, variability leads to greater parametric yield loss with respect to both noise margins and leakage power (there is an exponential dependence of leakage current on V_{th}). Future circuit topologies and design methodologies will need to address these issues. Error-tolerance is another challenge that becomes severe with process scaling and aggressive layout densities. Embedded memory soft-error rate (SER) increases with diminishing feature sizes, and affects both embedded SRAM and embedded DRAM, as discussed in the Design Chapter. Moving bits in non-volatile memory may also suffer upsets. Particularly for highly reliable applications such as in the automotive sector, error correction is a requirement going forward, and will entail tradeoffs of yield and reliability against access time, power, and process integration. Finally, cost-effective manufacturing test and built-in self-test, for both large and heterogeneous memory arrays, is a critical requirement in the SOC context.

Since memory cell size and performance due to its high multiplication rate has very direct impact on cost and performance the amount of engineering work spend for optimization is much higher compared to all other basic circuits discussed here. Tables 12a and 12b give technology requirements for the three currently dominant types of embedded memory: CMOS embedded static random-access memory (SRAM), embedded non-volatile memory (NVM), and embedded dynamic random-access memory (DRAM). Those parameters arise from the balance of circuit design consideration and technology boundary conditions given by the logic requirements tables in the PIDS chapter. Aggressive scaling of CMOS SRAM continues due to high-performance and low-power drivers, which require scaling of read cycle time by 0.7× per node. Voltage scaling involves multiple considerations, e.g., the relationship between retention time and read operating voltage, or the impact of supply and threshold voltage scaling on pMOS device requirements starting at the 45 nm node. More nascent ferroelectric RAM, magnetoresistive RAM, and phase-change memory technologies are discussed in the *Emerging Research Devices* section of PIDS Chapter.

Table 12a Embedded Memory Requirements—Near-term

Year of Production	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*
Technology Node		*hp90*			*hp65*		
DRAM ½ Pitch (nm)	*100*	*90*	*80*	*70*	*65*	*55*	*50*
MPU/ASIC ½ Pitch (nm)	*120*	*107*	*95*	*85*	*76*	*67*	*60*
CMOS Static Random Access Memory (HP/LSTP), Technology Node (nm), Feature Size – F	*130*	*90*	*90*	*90*	*65*	*65*	*65*
6T bit cell size (F^2) [1]	**140F²**	**140F²**	**140F²**	**140F²**	**140F²**	**140F²**	**140F²**
Array efficiency [2]	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**
Process overhead versus standard CMOS – number of added mask layers [3]	**1**	**1**	**1**	**2**	**2**	**2**	**2**
Operating voltage – V_{dd} (V) HP/LSTP [4]	**1.2**	**1.2**	**1.1/1.2**	**1.1/1.2**	**1.1**	**1/1.1**	**1/1.1**
Static power dissipation (mW/Cell) HP/LSTP [5]	**1E-4/4E-7**	**1.5E-4/6E-7**	**1.5E-4/6E-7**	**1.5E-4/6E-7**	**3E-4/1E-6**	**3E-4/1E-6**	**3E-4/1E-6**
Dynamic power consumption per cell – (mW/MHz) HP/LSTP [6]	**9E-7/1E-6**	**8E-7/9E-7**	**7E-7/8.5E-7**	**6E-7/8E-7**	**4.5E-7/7E-7**	**4E-7/6.5E-7**	**4E-7/6E-7**
Read cycle time (ns) HP/LSTP [7]	**0.5/2**	**0.4/2**	**0.4/2**	**0.4/2**	**0.3/1.5**	**0.3/1.5**	**0.3/1.5**
Write cycle time (ns) HP/LSTP [7]	**0.5/2**	**0.4/2**	**0.4/2**	**0.4/2**	**0.3/1.5**	**0.3/1.5**	**0.3/1.5**
Soft error rate (FIT/Mb) [8]	**1000**	**1000**	**1000**	**1000**	**1000**	**1000**	**1000**
Embedded Non-Volatile Memory (code/data), Technology Node (nm)	*180*	*130*	*130*	*130*	*90*	*90*	*90*
Cell size (F^2) – NOR FLOTOX /NAND FLOTOX [9]	**$10F^2$/$5F^2$**	**$10F^2$/$5F^2$**	**$10F^2$/$5F^2$**	**$10F^2$/$5F^2$**	**$10F^2$/$5F^2$**	**$10F^2$/$5F^2$**	**$10F^2$/$5F^2$**
Array efficiency – NOR FLOTOX/NAND FLOTOX [10]	**0.6/0.8**	**0.6/0.8**	**0.6/0.8**	**0.6/0.8**	**0.6/0.8**	**0.6/0.8**	**0.6/0.8**
Process overhead versus standard CMOS – number of added mask layers [11]	**6-8**	**6-8**	**6-8**	**6-8**	**6-8**	**6-8**	**6-8**
Read operating voltage (V)	**3.0V**	**2.5V**	**2.5V**	**2.5V**	**2V**	**2V**	**2V**
Write (program/erase) on chip maximum voltage (V) – NOR/NAND [12]	**12V/15V**	**12V/15V**	**12V/15V**	**12V/15V**	**12V/15V**	**12V/15V**	**12V/15V**
Static power dissipation (mW/Cell) [5]	**1.E-06**	**1.E-06**	**1.E-06**	**1.E-06**	**1.E-06**	**1.E-06**	**1.E-06**
Dynamic power consumption per cell – (mW/MHz) [6]	**1.E-07**	**0.8E-07**	**0.8E-07**	**0.8E-07**	**0.6E-07**	**0.6E-07**	**0.6E-07**
Read cycle time (ns) NOR FLOTOX /NAND FLOTOX [7]	**20/1000**	**14/70**	**14/70**	**14/70**	**10/50**	**10/50**	**10/50**
Program time per cell (µs) NOR FLOTOX /NAND FLOTOX [13]	**1.0/1000.0**	**1.0/1000.0**	**1.0/1000.0**	**1.0/1000.0**	**1.0/1000.0**	**1.0/1000.0**	**1.0/1000.0**
Erase time per cell (ms) NOR FLOTOX /NAND FLOTOX [13]	**10.0/0.1**	**10.0/0.1**	**10.0/0.1**	**10.0/0.1**	**10.0/0.1**	**10.0/0.1**	**10.0/0.1**
Data retention requirement (years) [13]	**10**	**10**	**10**	**10**	**10**	**10**	**10**
Endurance requirement [13]	**100,000**	**100000**	**100000**	**100000**	**100000**	**100000**	**100000**
Embedded DRAM, Technology Node (nm)	*130*	*130*	*130*	*90*	*90*	*90*	*65*
1T1C bit cell size (F^2) [14]	**$12F^2$**	**$12F^2$**	**$12F^2$**	**$12F^2$**	**$12F^2$**	**$12F^2$**	**$12F^2$**
Array efficiency [2]	**0.6**	**0.6**	**0.6**	**0.6**	**0.6**	**0.6**	**0.6**
Process overhead versus standard CMOS – number of added mask layers [3]	**4-6**	**4-6**	**4-6**	**4-6**	**4-6**	**4-6**	**4-6**
Read operating voltage (V)	**2.5**	**2.5**	**2.5**	**2**	**2**	**2**	**1.7**
Static power dissipation (mW/Cell) [5]	**1E-10**	**1E-10**	**1E-10**	**1E-10**	**1E-10**	**1E-10**	**1E-10**
Dynamic power consumption per cell – (mW/MHz) [6]	**1.E-07**	**1.E-07**	**1.E-07**	**1.E-07**	**1.E-07**	**1.E-07**	**1.E-07**
DRAM retention time (ms) [13]	**64**	**64**	**64**	**64**	**64**	**64**	**64**
Read/Write cycle time (ns) [7]	**1**	**1**	**1**	**0.7**	**0.7**	**0.7**	**0.5**
Soft error rate (FIT/Mb) [8]	**10**	**10**	**10**	**10**	**10**	**10**	**10**

Table 12b Embedded Memory Requirements—Long-term

Year of Production	*2010*	*2012*	*2015*	*2018*
Technology Node	*hp45*			
DRAM ½ Pitch (nm)	*45*	*35*	*25*	*18*
MPU/ASIC ½ Pitch (nm)	*54*	*42*	*30*	*21*
CMOS Static Random Access Memory (HP/LSTP), Technology Node (nm), Feature Size – F	*45*	*35*	*25*	*18*
6T bit cell size (F^2) [1]	**$140F^2$**	**$140F^2$**	**$140F^2$**	**$140F^2$**
Array efficiency [2]	**0.7**	**0.7**	**0.7**	**0.7**
Process overhead versus standard CMOS – number of mask adders [3]	**2**	**2**	**2**	**2**
Operating voltage – V_{dd} (V)	**1**	**0.9/1**	**0.8/0.9**	**0.8/0.7**
Static power dissipation (mW/Cell) [5]	**5E-4/1.2E-6**	**1E-3/1.5E-6**	**2E-3/2E-6**	**3E-3/2.5E-6**
Dynamic power consumption per cell – (mW/MHz) [6]	**3E-7/5E-7**	**2.5E-7/4.5E-7**	**2E-7/4E-7**	**1.5E-7/3E-7**
Read cycle time (ns) [7]	**0.2/1.2**	**0.15/0.8**	**0.1/0.5**	**0.07/0.3**
Write cycle time (ns) [7]	**0.2/1.2**	**0.15/0.8**	**0.1/0.5**	**0.07/0.3**
Soft error rate (FIT/Mb) [8]	**1000**	**1000**	**1000**	**1000**
Embedded Non-Volatile Memory (code/data), Technology Node (nm)	*65*	*45*	*35*	*25*
Cell size (F^2) – NOR FLOTOX/NAND FLOTOX [9]	**$10F^2/5F^2$**	**$10F^2/5F^2$**	**$10F^2/5F^2$**	**$10F^2/5F^2$**
Array efficiency – NOR FLOTOX/NAND FLOTOX [10]	**0.6/0.8**	**0.6/0.8**	**0.6/0.8**	**0.6/0.8**
Process overhead versus standard CMOS – number of mask adders [3]	**6–8**	**6–8**	**6–8**	**6–8**
Read operating voltage (V) [4]	**1.8V**	**1.5V**	**1.3V**	**1.2V**
WRITE (program/erase) on chip maximum voltage (V) – NOR/NAND [4]	**12V/15V**	**12V/15V**	**12V/15V**	**12V/15V**
Static power dissipation (mW/Cell) [5]	**1.E-06**	**1.E-06**	**1.E-06**	**1.E-06**
Dynamic power consumption per cell – (mW/MHz) [6]	**0.5E-8**	**0.4E-8**	**0.35E-8**	**0.3E-8**
Read cycle time (ns)	**7/35**	**5/25**	**3.5/18**	**2.5/12**
Program time per cell (µs) [13]	**1.0/1000.0**	**1.0/1000.0**	**1.0/1000.0**	**1.0/1000.0**
Erase time per cell (ms) [13]	**10.0/0.1**	**10.0/0.1**	**10.0/0.1**	**10.0/0.1**
Data retention requirement (years) [13]	**10**	**10**	**10**	**10**
Endurance requirement [13]	**100000**	**100000**	**100000**	**100000**
Embedded DRAM, Technology Node (nm)	*65*	*45*	*35*	*25*
1T1C bit cell size (F^2) [14]	**$12F^2$**	**$12F^2$**	**$12F^2$**	**$12F^2$**
Array efficiency [2]	**0.6**	**0.6**	**0.6**	**0.6**
Process overhead versus standard CMOS – number of mask adders [3]	**4–6**	**4–6**	**4–6**	**4–6**
Read operating voltage (V)	**1.7**	**1.6**	**1.5**	**1.5**
Static power dissipation (mW/Cell) [5]	**1E-10**	**1E-10**	**1E-10**	**1E-10**
Dynamic power consumption per cell – (mW/MHz) [6]	**1.5E-07**	**1.6E-07**	**1.7E-07**	**1.7E-07**
DRAM retention time (ms) [13]	**64**	**64**	**64**	**64**
Read/Write cycle time (ns) [7]	**0.4**	**0.3**	**0.25**	**0.2**
Soft error rate (FIT/Mb) [8]	**10**	**10**	**10**	**10**

Definitions of Terms for Tables 12a and 12b:
[1] Size of the standard 6T CMOS SRAM cell as a function of minimum feature size.
[2] Typical array efficiency defined as (core area/memory instance area).
[3] Typical number of extra masks is needed over standard CMOS logic process of equivalent technology. This is typically zero, however for some high-performance or highly reliable (noise immune) SRAMs special process options are sometimes applied like additional high—V_{th} pMOS cell transistors and using higher V_{dd} for better noise margin or zero-V_{th} access transistors for fast read-out.
[4] Nominal operating voltage refers to the HP and LSTP devices in the logic device requirements table in the PIDS chapter.
[5] Static power dissipation per cell in standby mode. This is measured at I_standby x V_{dd}. (off-current and V_{dd} are taken from the HP and LSTP devices in the logic device requirements table in the PIDS Chapter.

[6] This parameter is a strong function of array architecture. However, a parameter for technology can be determined per cell level. Assume full V_{dd} swing on the Wordline (WL) and $0.8V_{dd}$ swing on the Bitline (BL). Determine the WL capacitance per cell (CWL) and BL capacitance per cell (CBL). Then: dyn. power cons. per MHz per cell = V_{dd} × CWL (per cell) × (V_{dd}) + V_{dd} x CBL (per cell) x (V_{dd}) ×10^6.
[7] Read cycle time is the typical time it takes to complete a READ operation from an ADDR. Depends on memory size and architecture. Write cycle time is the typical time it takes to complete a WRITE operation to an ADDR. Depends on memory size and architecture.
[8] A FIT is a failure in 1 billion hours. This data is presented as FIT per megabit.
[9] Size of the standard 1T FLOTOX cell/size of the standard 2T SG cell/size of the standard NAND cell. Cell size is somewhat enhanced compared to stand-alone NVM due to integration issues.
[10] Array efficiency of the standard stacked gate NOR architecture/standard split gate NOR architecture/standard NAND architecture. Data refer to PIDS table the NVM device requirements table in the PIDS chapter.
[11] Extra process steps needed to realize the technology as compared to standard CMOS process.
[12] Maximum voltage required for operation, typically used in WRITE operation. Data refer to the NVM device requirements table in the PIDS chapter.
[13] Program time per cell is typically the time needed to program data to a cell. Erase time per cell is typically the time needed to erase a cell. Data retention requirement is the duration for which the data must remain non-volatile even under worst-case conditions. Endurance requirement specifies the number of times the cell can be programmed and erased.
[14] Size of the standard cell for embedded trench DRAM cell. Data refer to PIDS table the DRAM requirements table in the PIDS chapter.