INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

2011 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 comprehend how technology requirements arise for product classes whose business and retooling cycles drive the semiconductor sector. The unstated assumption that technological advances are deployed in all semiconductor products, independent of the specifics of key product classes, is no longer valid. Today, introduction of new technology solutions is increasingly application-driven – i.e., *applications drive technology*. Computer microprocessors have been joined as drivers by mixed-signal systems, battery-powered mobile devices, wall-plugged consumer devices, and networking devices. In-house chip designs are replaced by system-on-chip (SOC) and system-in-package (SIP) designs that incorporate building blocks from multiple sources.

The purpose of the 2011 ITRS System Drivers Chapter is to update existing ITRS system drivers, and to continue adding further drivers to capture the increasing breadth of the semiconductor industry. Together with the *Overall Roadmap Technology Characteristics*, the System Drivers Chapter provides a consistent framework and motivation for technology requirements across the respective ITRS technology areas. This chapter consists of quantified, self-consistent models of the system drivers that support extrapolation into future technologies. We focus on four system drivers: system-on-chip (SOC – including increasing mentions of system-in-package technology), microprocessor (MPU), analog/mixed-signal (AMS), and embedded memory. The system-on-chip driver is defined according to key markets: consumer stationary, consumer portable, and networking. We first briefly survey key *market drivers* for semiconductor products. The reader is also referred to the International Electronics Manufacturing Initiative (iNEMI) roadmap, *http://www.inemi.org*.

MARKET DRIVERS

Table SYSD1 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 SYSD1 Major Product Market Segments and Impact on System Drivers

Market Drivers	*SOC*	*Analog/MS*	*MPU*
I. Portable/consumer			
1. Size/weight ratio: peak in 2004 2. Battery life 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. Medical			
1. Cost: slight downward pressure (~1/2 every 5 years) 2. Time-to-market: >12 months 3. Function: new on-chip functions 4. Form factor often not important 5. Durability/safety 6. Conservation/ ecology	High-end products only. Reprogrammability possible. Mainly ASSP, especially for patient data storage and telemedicine; more SOC for high-end digital with cores for imaging, real-time diagnostics, etc.	Absolutely necessary for physical measurement and response, but may not be integrated on chip	Often used for programmability especially when real-time performance is not important Recent advances in multicore processors have made programmability and real-time performance possible
III. Networking and communications			
1. Bandwidth: 4×/3–4 years 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

[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 integration. This partitioning is evolving.*

Table SYSD1 Major Product Market Segments and Impact on System Drivers (continued)

IV. Defense			
1. Cost: not prime concern 2. Time-to-market: >12 months 3. Function: mostly on SW to ride technology curve 4. Form factor may be important 5. High durability/safety	Most cases leverage existing processors, but some requirements may drive towards single-chip designs with programmability	Absolutely necessary for physical measurement and response, but may not be integrated on chip	Often used for programmability, especially when real-time performance is not important Recent advances in multicore processors have made programmability and real-time performance possible
V. Office			
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-position 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)
VI. 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 HW 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	

A/D—analog to digital ASSP—application-specific standard product D/A—digital to analog DEMUX—demultiplexer DSP—digital signal processing FPGA—field programmable gate array i/f—interface I/O—input/output HW—hardware MEMS—microelectromechanical systems MUX—multiplexer RTOS—real-time operating system

SYSTEM-ON-CHIP DRIVER

SOC is a *product class and design style* that 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 has evolved from the ASIC driver discussed in early editions of the ITRS; reduced design costs and higher levels of system integration are its principal goals.[2] 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]

SOC represents a confluence of previous product classes in several ways. As noted above, SOCs integrate building blocks from the other system driver classes. The quality gap between full-custom and ASIC/SOC has steadily diminished: 1) starting in the 2001 ITRS, overall ASIC and MPU logic densities were modeled as being equal; and 2) "custom quality on an ASIC schedule" has been increasingly achieved by improved physical synthesis and tuning-based standard-cell

[2] *The term "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 in Verilog/VHDL (very high speed integrated circuits hardware description language) 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.*

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]). We also note that particular market sectors, notably networking hardware and gaming systems, increasingly feature very demanding performance specifications. In some cases, required performance metrics – e.g., per-die floating point operations per second, or per-die external I/O bandwidth – rise above those of conventional drivers such as the MPU driver. Given these specifications, it is the SOC designs in such sectors that have become the drivers of key design requirements and solutions. Growth in key parameters, such as number of cores per die, maximum frequency per core, and per-pin I/O bandwidth, is increasingly led by these drivers.

As noted above, 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. A second basic challenge is the heterogeneous integration of components from multiple implementation *fabrics* (such as reprogrammable logic, memory, analog and radio frequency (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 of heterogeneous technologies, 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, ferroelectric RAM (FeRAM), magnetic RAM (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. This chapter discusses the nature and evolution of SOCs with respect to several variants driven respectively by integration (multi-technology integration or MT), high performance (HP) with emphasis on (a) networking and (b) consumer stationary segments, and low power/cost (LP) with emphasis on the consumer portable segment.

SOC/SIP MULTI-TECHNOLOGY

The need to build heterogeneous systems on a single chip arises from 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, 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". First integrations of each technology within standard CMOS processes—not necessarily together with other technologies, and not necessarily in volume production—will evolve over time. 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 (such as Flash, DRAM). Physical scale in system applications (ear-mouth = speaker-microphone separation, or distances within a car) also affects the need for single-die integration, particularly of sensors.

SOC NETWORKING DRIVER

Examples of high-performance SOC designs include processors for high-end gaming (cf. the SOC Consumer Stationary (SOC-CS) Driver, below) and networking applications. SOCs for high-speed networking drive requirements for off-chip I/O signaling (which in turn create significant challenges to test, assembly and packaging, and design). Historically, chip I/O speed (per-pin bandwidth) has been scaling more slowly than internal clock frequency. 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, culminating with today's state-of-the-art serial links at 10 Gbit/s, now moving to 40 Gbit/s and 100 Gbit/s. Future networking requires scalable, power-limited, cost-driven SOC solutions that can deliver rich multimedia content and support advanced IP-based applications and services (seamless mobility, entertainment, home networking, etc.). Given the fundamental differences between core speeds and memory and I/O latencies, the

[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.*

trajectory for networking SOCs is toward multicore architectures with heterogeneous on-demand accelerator engines, and with integration of on-board switch fabric and L3 caches. We now motivate and describe a new multicore SOC platform architecture – the SOC Networking Driver – which targets the embedded networking space.

Because networking needs can no longer be met by increasing the operating frequencies on single-core architectures, any networking SOC solution will exploit multicore architecture to achieve added performance. Even so, thermal management challenges and hard power limits in the embedded space (e.g., 20W or 40W) prevent multicore alone from delivering the necessary performance increases. Consequently, integration of accelerator engines, on-chip switch fabric, and more on-board cache hierarchy will also be applied in the quest for incremental performance. These evolutions are examples of *design equivalent scaling* (refer to the *Executive Summary* for definitions). From an SOC platform perspective, the challenge of expanding the achievable performance-power envelope goes beyond silicon, to encompass such issues as contention for bus bandwidth and memories, scalability, and unused processing cycles due to lack of programming visibility. Leveraging the raw hardware capability requires greater investment in software enablement and simulation environment. This motivates the SOC Networking Driver architecture illustrated in Figure SYSD1, which shows the multicore and accelerator engine (MC/AE) aspects necessary to address needs of the embedded networking space. The MC/AE architecture is designed not only to provide superior performance and energy efficiency, but also to ease the industry's transition to multicore processors via explicit investment in the complementary software enablement ecosystem.

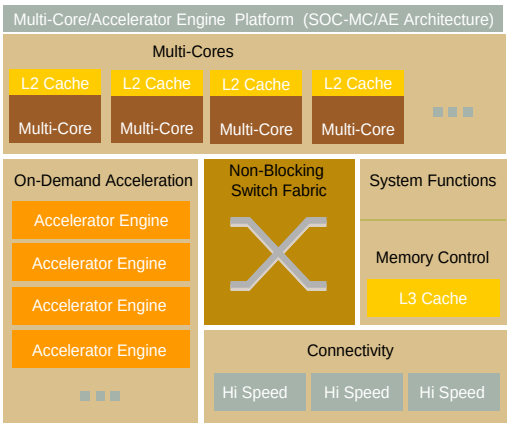

Figure SYSD1 SOC Networking Driver Architecture Template

Figure SYSD2 illustrates the anticipated growth over time in performance and number of cores for the SOC Networking Driver with an assumed 30W power envelope targeting the mid-range switching/routing workload segment of the embedded networking space. Model assumptions include the following.

- Die area is constant.
- Number of cores increases by 1.4× / year.
- Core frequency increases by 1.05× / year.
- On-demand accelerator engine frequency increases by 1.05× / year.
- Underlying fabrics – logic, embedded memory (cache hierarchy), on-chip switching fabric, and system interconnect – will scale consistently with the increase in number of cores.

The figure shows a roughly 1000× increase in the system processing performance metric, which is the product of number of cores, core frequency, and accelerator engine frequency. Per the scenario shown, anticipated 22nm system performance is >20× (with 80+ cores) the system performance of an 8-core implementation at 45nm in 2009.

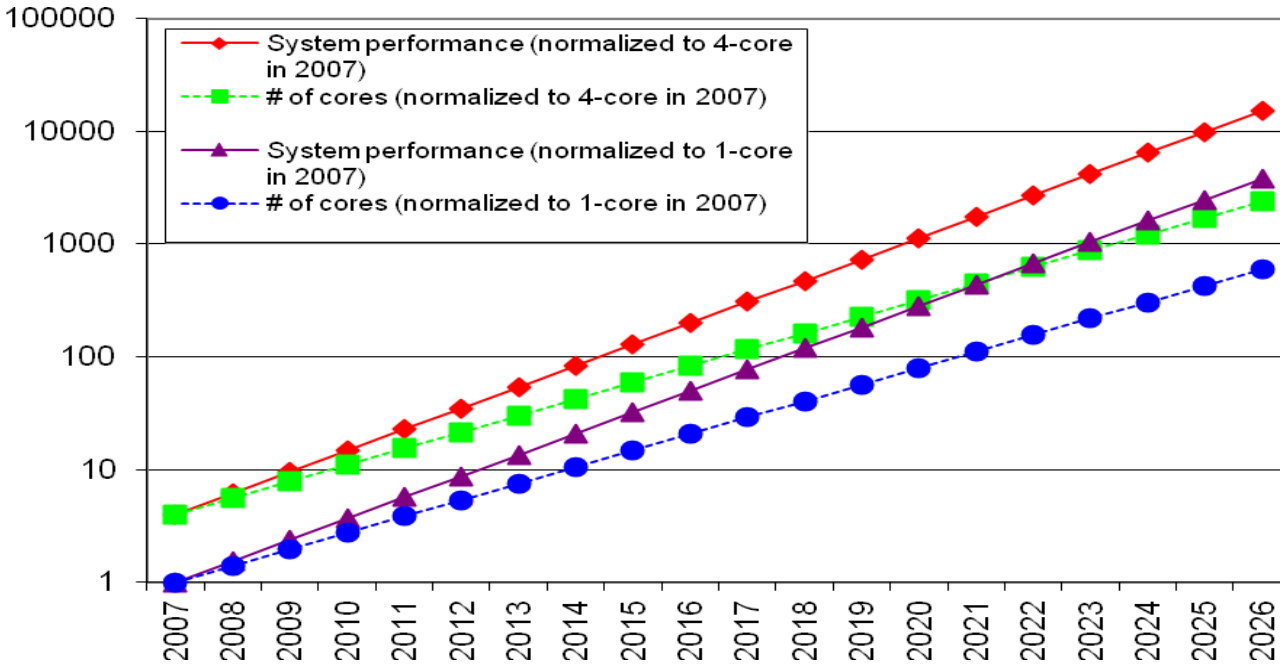

Figure SYSD2 SOC Networking Driver MC/AE Platform Performance

SOC CONSUMER DRIVER

The SOC Consumer Driver captures a typical SOC class that reflects a wide variety of consumer electronics applications. Due to short product life cycles and rapidly growing needs for functionality and performance in consumer products, the key requirements for the SOC Consumer Driver are to achieve high performance and function, and short time-to-market. The SOC Consumer Driver is classified into two categories, Consumer Portable and Consumer Stationary, with typical applications being mobile telephony and high-end gaming, respectively. The two different categories are distinguished mainly by power consumption requirement: the Consumer Portable Driver must minimize power consumption to maintain product battery life, while the Consumer Stationary Driver has high performance as its most important differentiator.

SOC CONSUMER DRIVER DESIGN PRODUCTIVITY TRENDS

Table SYSD2 shows required design productivity trends common to both SOC Consumer Portable and SOC Consumer Stationary Drivers. The underlying model makes the following assumptions. Required design effort is assumed constant. Design effort is assumed to be proportional to the size of the logic circuit portion. Design effort for reused logic is assumed to be half the effort needed for newly designed logic of equal size; this is because reused logic is not free, but requires effort for functionality modifications and design steps up to implementation and final physical verification. Design reuse effort is free for non-logic circuits, such as memory and pure analog. Reuse rate in all years is determined by a linear fit to values of 54% in 2011 and 98% in 2026. With these assumptions, maintaining constant SOC design effort requires a 10× design productivity improvement for newly designed logic over the next ten years to 2019. To solve this productivity challenge, several approaches must be combined. First, design abstraction levels must be raised. Second, the degree of automation, particularly in design verification and design implementation, must be increased. Finally, reuse rate must be increased, with an accompanying reduction in effort overhead for design reuse also being required.

Table SYSD2 SOC Consumer Driver Design Productivity Trends

Years	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026
SoC-CP Total Logic Size	1.00	1.32	1.79	2.32	2.96	3.77	4.70	5.85	7.45	9.70	11.65	15.50	19.56	24.40	31.23	38.10
Required %of reused design	54%	58%	62%	66%	70%	74%	78%	82%	86%	90%	92%	94%	95%	96%	97%	98%
Required Productivity for new designs (Normalized to 2011)	1.00	1.22	1.60	2.02	2.50	3.08	3.72	4.48	5.51	6.93	8.17	10.67	13.34	16.48	20.89	16.48
Required Productivity for reused designs (Normalized to productivity for new designs at 2011)	1.00	1.22	1.60	2.02	2.50	3.08	3.72	4.48	5.51	6.93	8.17	10.67	13.34	16.48	20.89	25.24

SOC CONSUMER PORTABLE (SOC-CP) DRIVER

The SOC Consumer Portable (SOC-CP) Driver (known in prior ITRS editions as SOC-PE, for "power-efficient") increasingly represents SOC designs; it spans portable and wireless applications such as smart media-enabled telephones, tablets and digital cameras, as well as other processing purposes such as high-performance computing and enterprise applications. The SOC-CP driver is based on a model created by the Japan Semiconductor Technology Roadmap Design Group.

Figure SYSD3 shows the resulting required attributes of a power-efficient, consumer-driven, possibly wireless device with multimedia processing capabilities.

- Its typical application area is electronic equipment categorized as "Portable/Mobile Consumer Platforms", as this application area will make rapid progress in the foreseeable future across semiconductor technology generations.
- Typical requirements for this type of SOC ("Portable/Mobile Consumer Platforms") dictate a rapid increase in processing capability, despite an upper bound constraint on power to maintain battery lifetime. Processing power increases by 1000× in the next ten years, even as dynamic power consumption does not change significantly.
- Lifecycles of "Portable/Mobile Consumer Platform" products are and will continue to be short. Hence, design effort cannot be increased, and must remain at current levels for the foreseeable future.

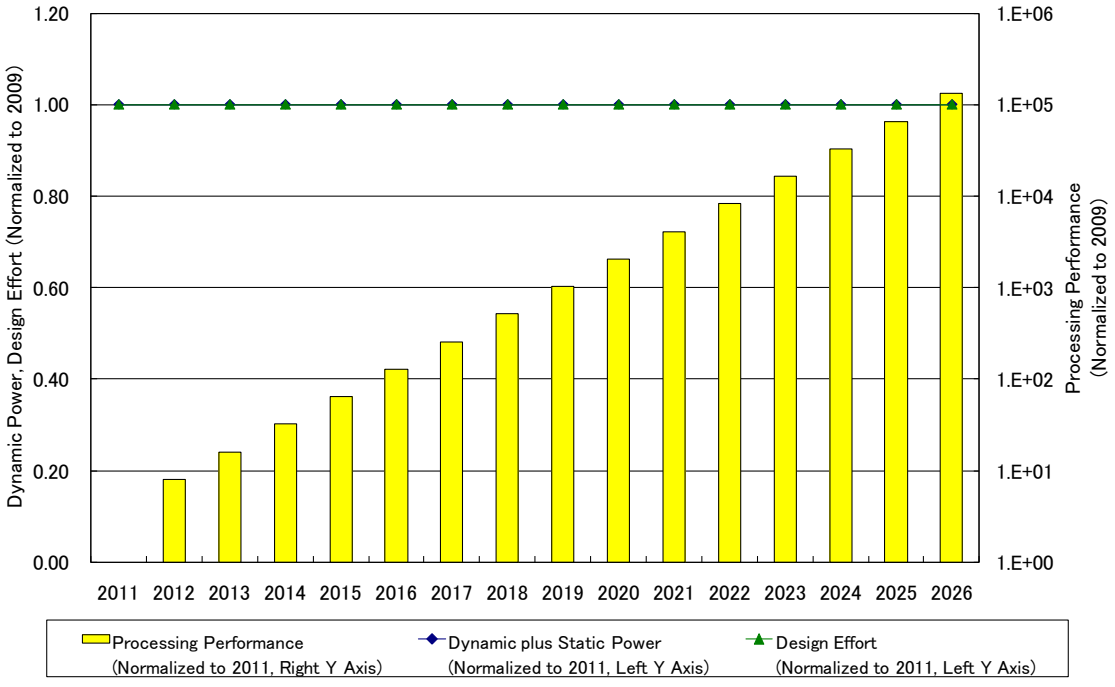

Figure SYSD3 Several Trends for the SOC Consumer Portable Driver

Figure SYSD4 shows an architecture template for the SOC Consumer Portable Driver. The SOC embodies a highly parallel architecture consisting of a number of main processors, a number of PEs (Processing Engines), peripherals, and

memories. Here, a PE is a processor customized for a specific function. A function with a large-scale, highly complicated structure will be implemented as a set of PEs. This architecture template enables both high processing performance and low power consumption by virtue of parallel processing and hardware realization of specific functions. The architecture does not require specific processor array architectures or symmetric processors; its essential feature is the large number of PEs embedded within the SOC to implement a set of required functions.

Based on this architecture template, Figure SYSD5 shows quantified design complexity trends for the SOC Consumer Portable Driver. Underlying model assumptions are as follows. 1) There are 2 to 4 main processors with approximately constant complexity, and the number of main processors will continually grow in the future. 2) Peripherals will also maintain constant complexity. 3) For PEs, average circuit complexity will stay constant, and the number of PEs will continue to grow subject to a die size limit of 49 mm^2 that gradually decreases to around 44 mm^2. Hence, the number of PEs grows rapidly in subsequent years. 4) The amount of main memory is assumed to increase proportionally with the number of PEs.

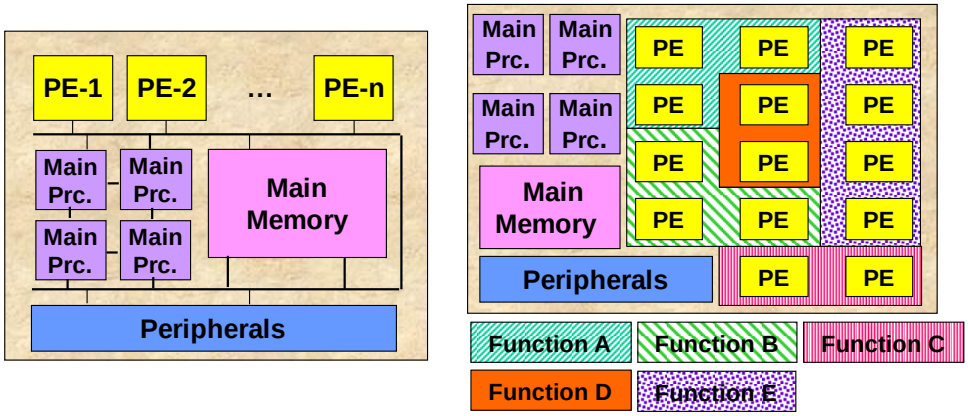

Figure SYSD4 SOC Consumer Portable Driver Architecture Template

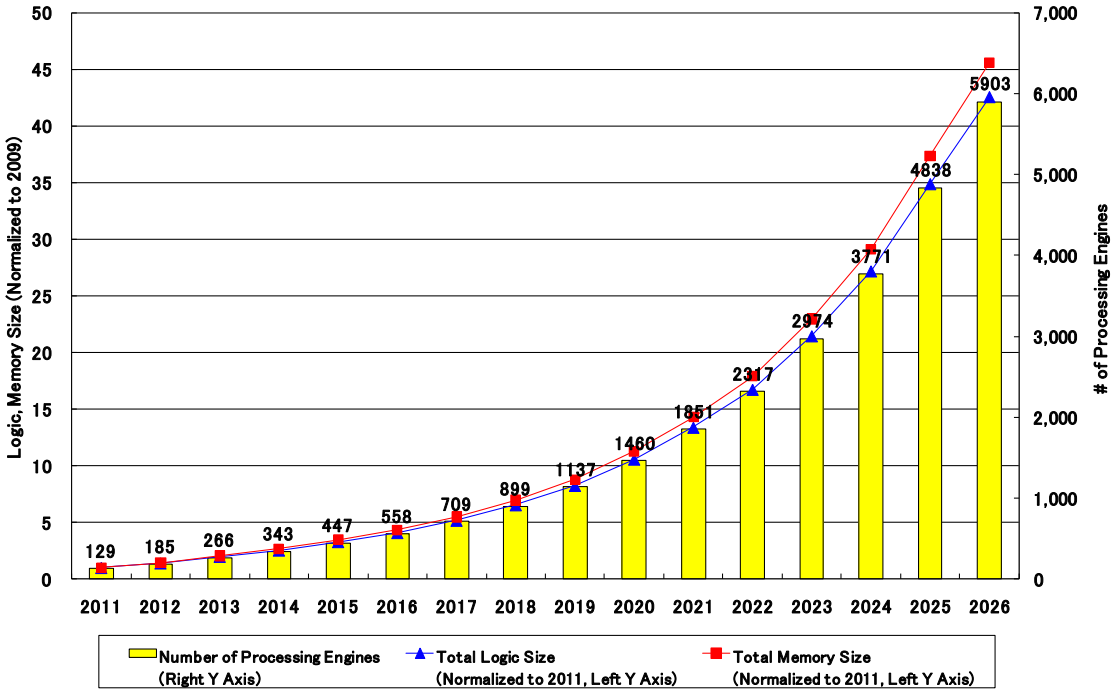

Figure SYSD5 SOC Consumer Portable Design Complexity Trends

SOC CONSUMER PORTABLE POWER CONSUMPTION TRENDS

While design complexity is a key trend, power consumption is also a critical factor for the design of SOC Consumer Portable chips. Figure SYSD6 shows the trend for total chip power, using LOP transistor performance parameters from the *PIDS Chapter*; interconnect performance parameters from the "Interconnect Technology Requirements" in the *Interconnect Chapter*, and circuit complexity parameters from Table SYSD2 above. We note the following.

- The model applied here simply extrapolates from current state-of-the-art technology and component technology roadmaps within the ITRS. The resulting power consumption substantially exceeds power efficiency requirements, despite the 2011 *PIDS Chapter* reverting the 2009 changes to LOP device supply voltage in long-term years. We expect that the required power-efficiency of competitive consumer portable products, as well as the global quest for more "green" and energy-efficient electronics products, will lead to a design power-centric device roadmap in future ITRS editions.
- Potential solutions are discussed in the *Design Chapter*. Specific solutions for SOC Consumer Portable include architecture optimization in high-level design stages based upon power consumption analysis, and customized PE realization.
- The total power consumption requirement of 0.5W was established in the 2009 edition of the ITRS. We note the existence of exceptions; for example, in newer tablet products portable consumer processors dissipating 2W may be acceptable, given the physical product dimensions and advanced power management techniques.
- Due to the discontinuous trajectory of supply voltage in the future, logic switching (i.e., dynamic) power, and/or memory static power, can have non-monotone behavior, e.g., from 2013 to 2014, from 2018 to 2019, and from 2020 to 2021.

SOC CONSUMER PORTABLE PROCESSING PERFORMANCE TRENDS

The SOC Consumer Portable driver's processing performance can be assumed proportional to the product of device performance and the number of PEs on the SOC. Figure SYSD7 shows that there remains a superlinearly growing gap between the processing requirement and the available processing performance. This gap can potentially be solved by increasing the number of PEs, subject to power and design effort constraints. Potential solutions are discussed in the *Design Chapter*, and include appropriate hardware/software (HW/SW) partitioning in high-level design stages, as well as automated interface technology from high-level design stages to implementation design stages (e.g., high-level synthesis).

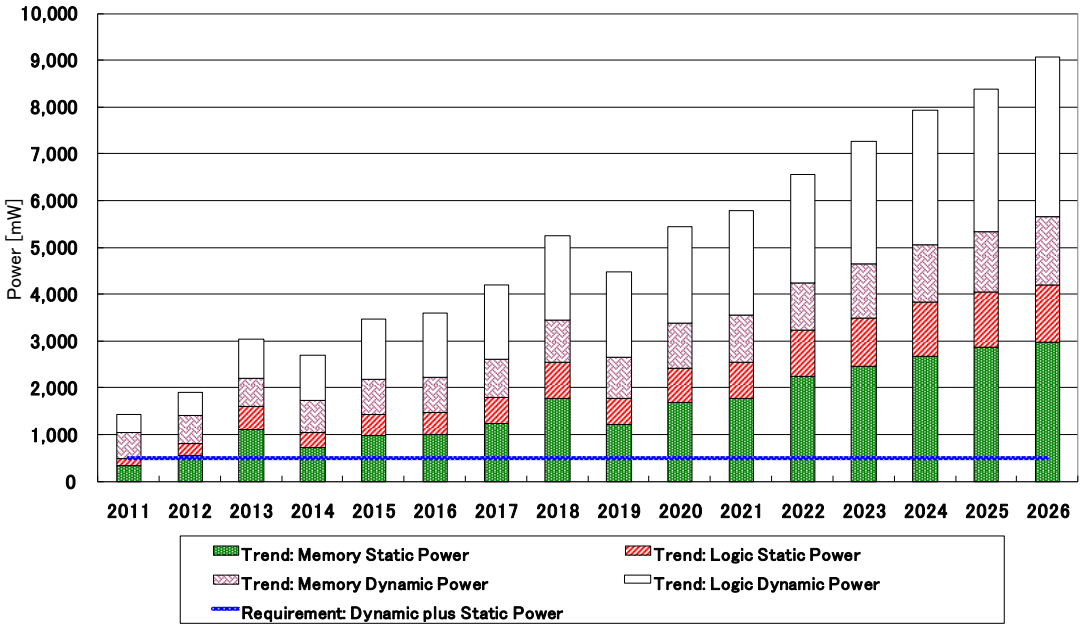

Figure SYSD6 SOC Consumer Portable Power Consumption Trends

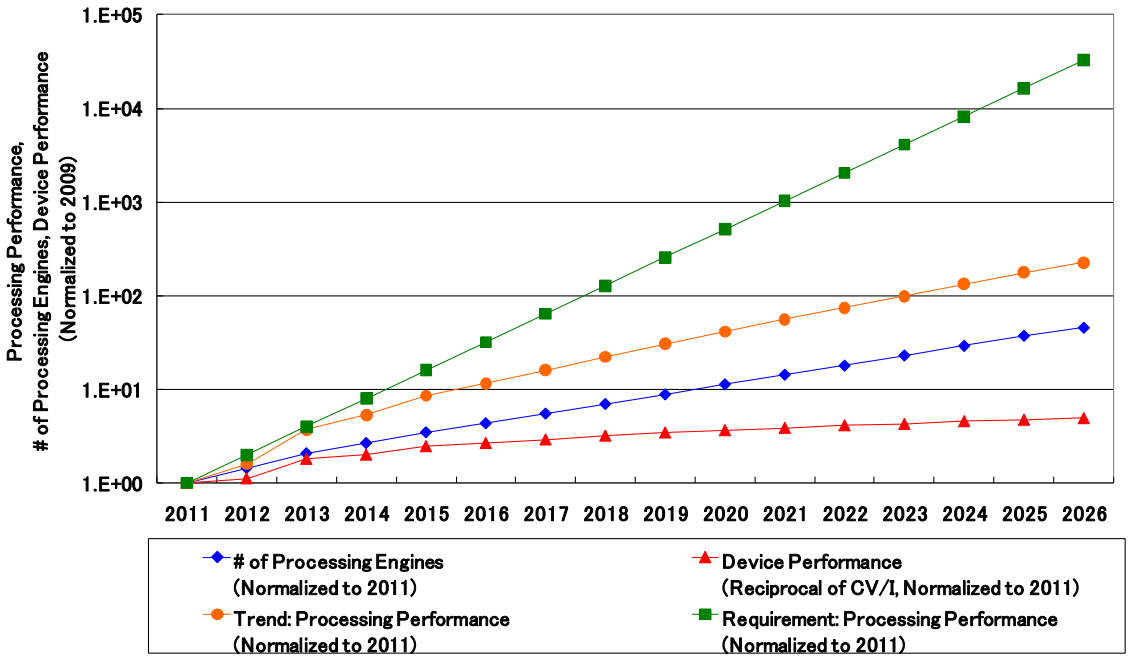

Figure SYSD7 SOC Consumer Portable Processing Performance Trends

SOC CONSUMER STATIONARY (SOC-CS) DRIVER

The SOC Consumer Stationary (SOC-CS) Driver represents SOC designs over a wide variety of applications in digital consumer electronic equipment, such as high-end game machines; these are assumed to be typically used in a tethered (non-mobile) environment. Key aspects of the model are as follows.

- Processing performance is the most important differentiator. As shown in Figure SYSD9, required processing performance in the year 2022 will be more than 70 TFlops.
- Functions will be implemented and realized mainly by software. Thus, high processing power is required, and this SOC will have many Data Processing Engines (DPEs).
- In comparison with the SOC Consumer Portable driver, this driver has worse performance–to-power ratio, but superior functional flexibility to support adding or modifying functions.
- Because it is easy to add or modify functions, the lifecycle of SOC Consumer Stationary designs is relatively long, and as a result the application area is wide.

Figure SYSD8 shows a typical architecture template for the SOC Consumer Stationary driver. The SOC features a highly parallel architecture consisting of a number of main processors, a number of DPEs, and I/O for memory and chip-to-chip interfaces. Here, a DPE is a processor dedicated to data processing which achieves high throughput by eliminating general-purpose features. A main processor is a general-purpose processor which allocates and schedules jobs to DPEs. A main processor, along with a number of DPEs, constitutes the basic architecture. The number of DPEs will be determined by required performance and chip size. Of all types of SOC that are modeled in this chapter, this SOC-CS driver will potentially have the largest number of DPEs in order to achieve required performance objectives.

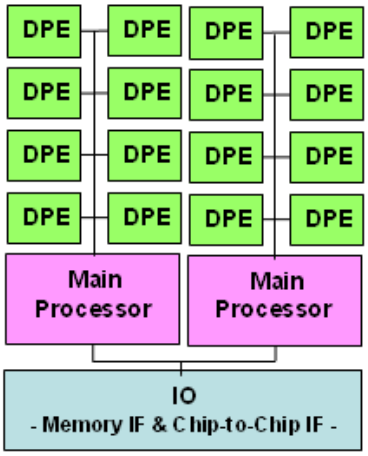

Figure SYSD8 SOC Consumer Stationary Driver Architecture Template

SOC CONSUMER STATIONARY DESIGN COMPLEXITY TRENDS

Based on the SOC-CS architecture template, quantified design complexity trends are shown in Figure SYSD9. The most interesting aspect is the rapid growth in number of DPEs. Underlying model assumptions are as follows.

- The SOC die size is constant at 220mm^2 based on published data for recent gaming processor products.
- Both the main processor and the DPE have constant circuit complexity, so that their respective layout areas decrease in proportion to the square of M1 pitch.
- A main processor is assumed to be able to control up to 8 DPEs.

SOC CONSUMER STATIONARY PERFORMANCE TRENDS

The SOC Consumer Stationary driver's processing performance can be assumed proportional to the product of device performance and the number of DPEs on the SOC. Figure SYSD10 shows SOC Consumer Stationary processing performance trends. Required processing performance grows rapidly, by approximately 250× over the next fifteen years. Key potential solutions to achieve the required performance include various design technologies (particularly in the logical, circuit and physical design stages) to maximize circuit performance. Automated design methodologies such as high-level synthesis are of course important as well.

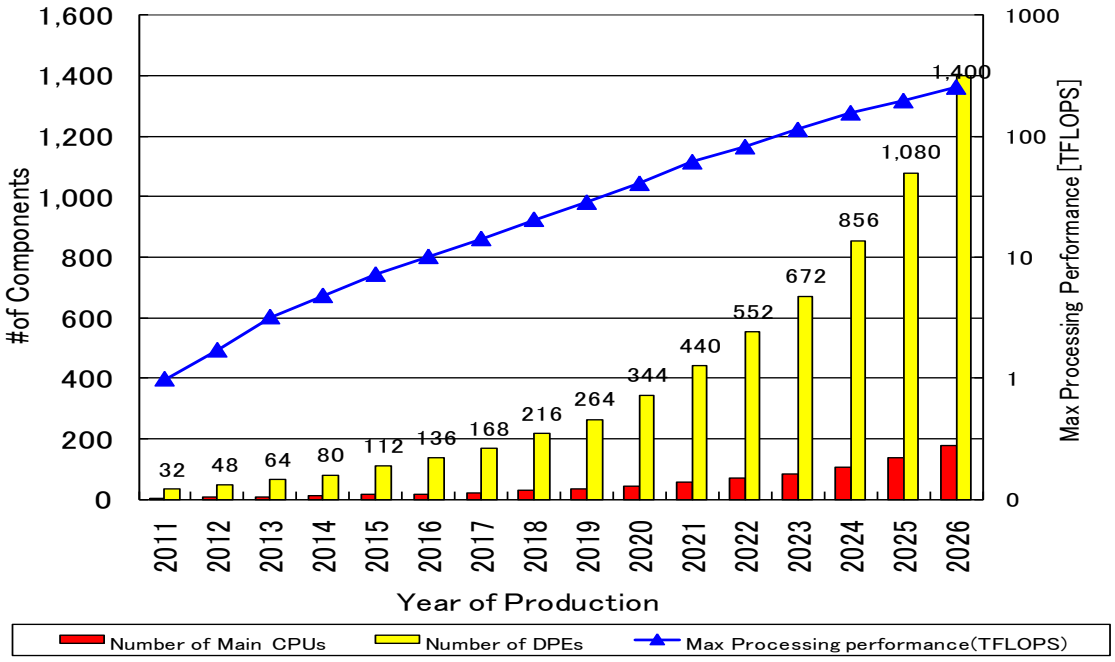

Figure SYSD9 SOC Consumer Stationary Design Complexity Trends

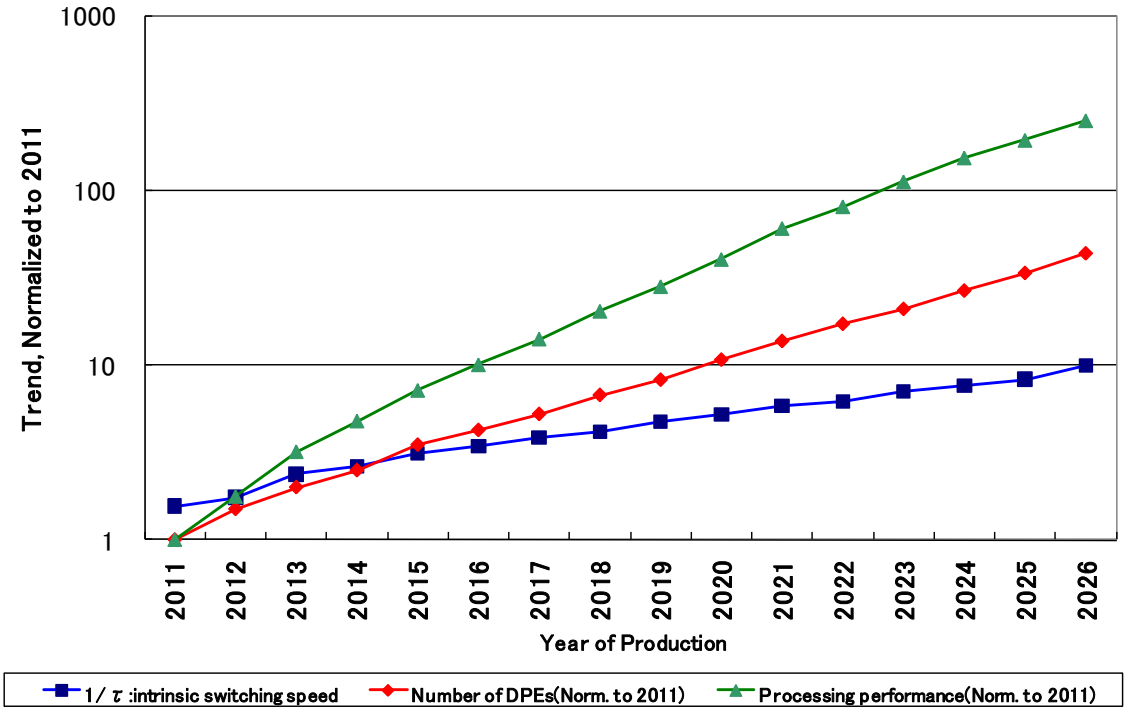

Figure SYSD10 SOC Consumer Stationary Performance Trends

SOC CONSUMER STATIONARY POWER CONSUMPTION TRENDS

An explosion in power consumption will be a critical consideration for the design of future SOC Consumer Stationary chips. Figure SYSD11 shows the trend for total chip power, decomposed into switching and leakage power, across logic and memory. The analysis is based on transistor performance parameters from the *PIDS Chapter*; interconnect performance parameters from the *Interconnect Chapter*, and the design complexity trends presented above. Power consumption as of 2011 is obtained from published data for recent gaming processor products. We note the following.

- Unlike the SOC Consumer Portable Driver, the SOC Consumer Stationary Driver is generally free from battery life issues; however, the rapid increase in power consumption will result in critical chip packaging and cooling issues.
- Leakage power will be much greater than the calculated value shown in Figure SYSD11, due to variability and temperature effects.
- Power consumption per DPE will decrease according to trends for supply voltage and insulator dielectric constant. However, this will be outweighed by the increase in number of DPEs per chip.

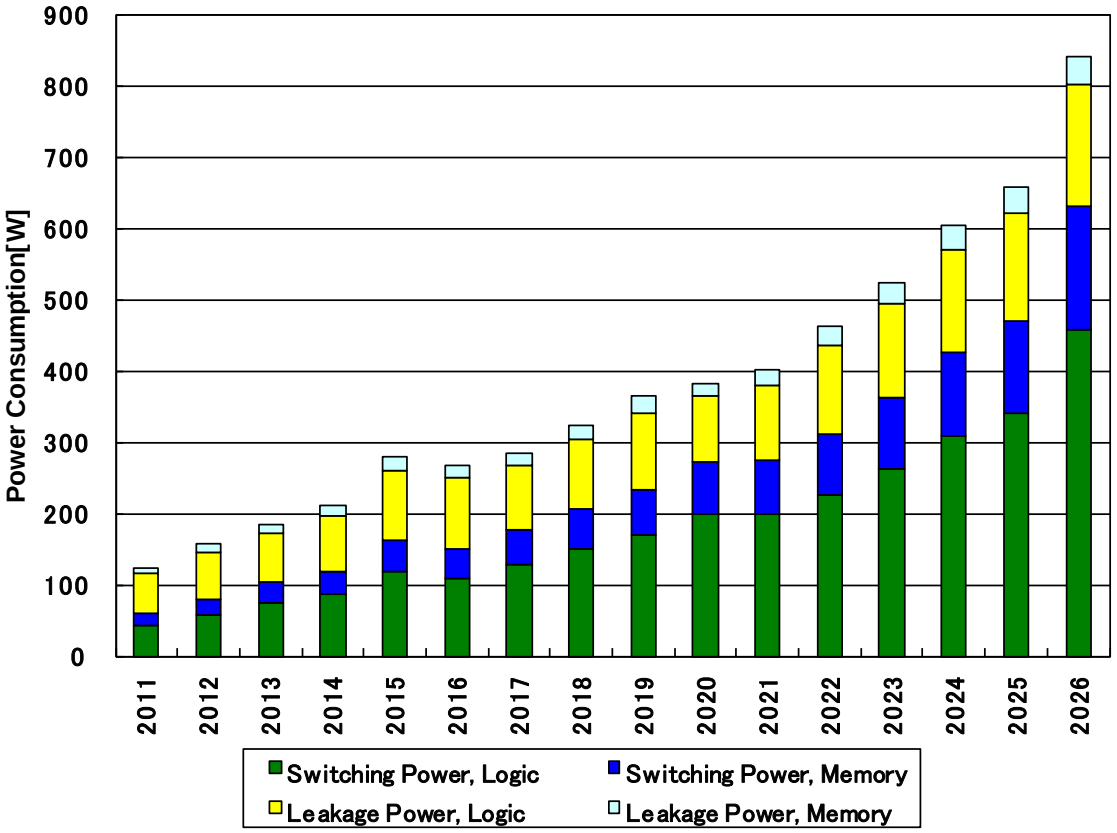

Figure SYSD11 SOC Consumer Stationary Power Consumption Trends

Clearly, the trend in Figure SYSD11 highlights a pressing need to develop new solutions, beyond those already embedded in the technology roadmaps of the *PIDS* and *Interconnect* chapters, so that actual power consumption remains within acceptable limits.

MICROPROCESSOR (MPU) DRIVER

In high-volume custom designs, performance and manufacturing cost issues outweigh design or other non-recurring engineering cost issues, primarily because of the large profits that these chips can potentially produce as a result of sales volumes. Large volumes 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 combined NRE

and manufacturing investment must be positive. Within the high-volume custom arena, three dominant classes today are MPUs, memory[5] and reprogrammable logic (e.g., FPGA). The MPU product class is a key system driver for semiconductor products because it uses the most aggressive design styles and manufacturing technologies. 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 in the linked study in the electronic chapter version).

1. *Three types of MPUs*—Historically, there have been three types of MPUs: 1) cost-performance (CP), reflecting "desktop," 2) high-performance (HP), reflecting "server," and 3) power-connectivity-cost (PCC), reflecting "mobile handheld device". As predicted in the 2001 ITRS, the increasing market acceptance of battery-limited mobile designs (often with wireless connectivity) leads to the creation of a new 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 that is disproportionate to product volume because of the large margins involved. As has already been predicted, 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, along with such added components as graphics, non-volatile memory and high speed I/Os. However, the larger margins and volumes of a PCC design will justify much greater design effort compared to a traditional SOC.
2. *Constant die area*—Die areas are constant (140 mm^2 for CP, 260 mm^2 for HP, 70–100 mm^2 for PCC) over the course of the roadmap, and are currently broken down into logic, memory, and integration overhead. Integration overhead reflects the presence of white space for interblock channels, floorplan packing losses, and potentially growing tradeoff of layout density for design turnaround time. The message here, in contrast to that of previous ITRS models, is that power, cost and interconnect cycle latency are strong limiters of die size. 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 and *uncore* components integrated on-chip) would not be cost-effective beyond a certain point.[6] Furthermore, the difficulty of accurate architectural performance simulations with increasingly deeper interconnect pipelining (a consequence of non-ideal process and device scaling) will also limit die growth size.
3. *Multi-core organization*—MPU logic content reflects multiple processing units on chip starting at the 130 nm generation, 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 (for example, x86, multimedia instructions (MMX), and explicitly parallel instruction computing (EPIC) with continuations for encryption, graphics, and multimedia, etc.); 4) the need to enable flexible management of power at the architecture, operating system (OS) and application levels via SOC-like integration of less efficient, general-purpose processor cores with more efficient, special-purpose "helper engines"[7]; 5) the increase in transistor complexity of processor cores [8] (the

[5] *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 times on a chip, and millions of chips are sold, the amount of custom design for these parts is extraordinary. This aspect has led to separate fabrication lines for DRAM devices, with some of the most careful circuit engineering needed to ensure correct operation.*

[6] *Multi-core organization and associated power efficiencies may permit slight growth in die size, but the message is still that die areas are not increasing.*

[7] *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.*

number of logic transistors per processor core is projected to increase by a factor of 1.4× or less with each technology generation); 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 multicore designs especially in the HP and high-end CP categories. The number of logic cores in the MPU model was reset to 4 in 2007 and projected to increase by a factor of 1.4× with each technology generation. The "power wall" and the currently limited capability to exploit available parallelism together limit the scaling of cores; recent trends suggest a factor of 2× every 4 years. In combination with scaling of the number of transistors per core, the number of logic transistors in the ITRS MPU model doubles with each successive technology generation.

4. *Memory content*—The MPU memory content was reset to 4 MBytes (4 × 1,048,576 × 9 bits) of SRAM for CP and 16 MBytes for HP in 2007. Memory content, like logic content, is projected to double with each successive technology generation, not with respect to absolute time intervals (e.g., every 18 months).[9, 10]

5. *Layout density*—Due to their high levels of system complexity and production volume, MPUs are the driver for improved layout density.[11] 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 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 $175F^2$, and SRAM layout densities reflect use of a 6-transistor bitcell of approximately $60F^2$ (N.B.: an 8-transistor bitcell would require approximately 84 F^2), where F is the logic M1 half-pitch of a given technology.[12] Layout density is projected to double with each technology generation, according to the scale factor of 0.7 for contacted M1 pitch. SRAM layout densities reflect use of a 6-transistor bit cell (via a fitted expression for area per bit cell in units of F^2) 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*, *Process Integration, Devices, and Structures (PIDS)*, *Front End Processes (FEP)* and *Test* roadmaps. Through the 2000 ITRS, the MPU maximum on-chip clock frequency was modeled to increase by a factor of 2 per generation. Of this, approximately 1.4× was historically realized by device scaling (17%/year improvement in CV/I metric); the other 1.4× was obtained by reduction in number of logic stages in a pipeline stage (e.g., equivalent of 32 fanout-of-4 inverter (FO4 INV) delays[13] at 180 nm, going to 24–26 FO4 INV delays at 130 nm). As noted in the 2001 ITRS, there are several reasons why this historical trend could 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 increasingly defer the impact of worsening interconnect RCs (relative to devices) rather than contribute directly to frequency improvements. Recent

---

[8] *Initially, the CP core was reset to 40 million transistors, and the HP core to 50 million transistors, in 2007. The difference allows for more aggressive microarchitectural enhancements (trace caching, various prediction mechanisms, etc.) and introduction of auxiliary engines (encryption, graphics/media, etc.).*

[9] *The doubling of logic and memory content with each technology generation, 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. While the ITRS remains planar CMOS-centric, evolution to UTB FD SOI and multi-gate FETs is under way. Adoption of such novel device architectures would allow improvements of layout densities beyond what is afforded by scaling alone.*

[10] *Adoption of eDRAM, and integration of on-chip L3 cache, can 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 does not significantly affect the chip size or total chip power roadmap.*

[11] *ASIC/SOC and MPU system driver products have access to similar processes, as forecast since the 1999 ITRS. This reflects the 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.*

[12] *Through the end of the roadmap, a 2-input NAND gate is assumed to lay out in 8 × 3 grids, where the vertical dimension is in units of contacted local metal (Metal 2) pitch ($P_{M2} = 2.5 \times F$), and the horizontal dimension is in unit of contacted poly pitch ($P_{poly} = 3 \times F$). A 6-transistor SRAM bit cell is assumed to lay out in 2×5 grids, where the vertical dimension is in units of contacted poly pitch ($P_{poly} = 3 \times F$), and the horizontal dimension is in units of contacted Metal1 pitch ($P_{M1}=2 \times F$). In other words, the average gate occupies ($8 \times 2.5F$) × ($3 \times 3F$) = $180F^2$ and the average SRAM bitcell occupies ($2 \times 3F$) × ($5 \times 2F$) = $60F^2$. After fitting with data from production libraries, 175 is chosen for the logic A-factor. For both semi-custom (ASIC/SOC) and full-custom (MPU) design methodologies, an overhead of 100% is assumed for logic, and 60% is assumed for SRAM. The A-factor of 175 (logic) is a significant reduction from the value of 320 used in previous ITRS editions; likewise, the value of 60 (SRAM) is also a significant reduction from previous editions. This was a fundamental cause of the MPU Driver die size reduction in the 2009 ITRS.*

[13] *A FO4 INV delay is defined to be the delay of an inverter driving a load equal to 4× its own input capacitance (with no local interconnect). This is equivalent to roughly 14× 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 as updated in the 2007 ITRS is linked in the 2007 ITRS online edition.*

editions of the ITRS flattened the MPU clock period at 12 FO4 INV delays at 90 nm, so that clock frequencies advanced only with device performance in the absence of novel circuit and architectural approaches. Modern MPU platforms have stabilized maximum power dissipation at approximately 120W due to package cost, reliability, and cooling cost issues. With a flat power requirement, the updated MPU clock frequency model was reset to 4.7 GHz in 2007 and is projected to increase by a factor of *at most* 1.25× per technology generation, despite aggressive development and deployment of low-power design techniques.[14]

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 SOC Consumer Portable model in Figure SYSD4). Key contexts for the future evolution of the traditional MPU are with respect to design productivity, power management, multicore organization, I/O bandwidth, and circuit and process technology.

Design productivity—The complexity and cost of design and verification of MPU products have 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 aspect is leading to decreasing emphasis on heavy customization and exotic circuit topologies, and increasing use of design automation tools 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 tools for analysis of timing, noise and power, and for verification of physical and electrical design rules, 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 timeframe) cannot continue to *cost-effectively* 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 1.25× ideal scaling).[15] 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 not be commercially viable; 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 (for example, 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.[16] The MPU model also permits increasing amounts of the memory hierarchy on chip (consistent with processor-in-memory, or large on-chip eDRAM L3 cache). Higher memory content (with lower power densities than logic) is an 'easy' path to controlling leakage and total chip power. In general, 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

---

[14] The new "constant" power MPU model depends on evolution of a "Design Factor", such that dynamic and leakage power reduction techniques together compensate the 1.25× increase of clock frequency with each technology generation. The Design Factor for dynamic power corresponds to a 15% reduction in switching activity factor per unit area with each technology generation; this will be achieved by improved design and partitioning of architectures/functions, and by extreme use of existing low-power techniques such as pin swapping, gate sizing, hierarchical clock gating, etc. Dynamic voltage and frequency scaling can contribute to the Design Factor of both dynamic power and leakage power. We believe that these Design Factors are actually conservative, in that the 'slack' between maximum achievable clock frequencies and projected clock frequencies allows superlinear reductions in chip power due to added flexibility of logic and physical design optimizations. Faster progress by the industry in achieving Design Factor-based power reduction can enable lower power budgets and/or higher clock frequencies in future.

[15] 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.

[16] 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 replication could allow future increases in chip 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, for example, a die with k-1 instead of k functional cores is still useful.

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 will shift away from general-purpose processing, and heterogeneity will generally be in the form of more low-power small cores coexisting with traditional large cores. Indeed, recent publications suggest that such small cores may dominate MPU scaling. Large cores will scale at a rate less than 2× per 4 years beyond 2013 while the small cores will scale by 2× per technology node. MPU core logic along with L1 and L2 caches are anticipated to occupy around 30% of die area, while the last-level L3 caches are anticipated to occupy another 30% of die area.

Input/output bandwidth—In MPU systems, I/O pins are mainly used to connect to high-level cache 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 (a tester is required that can run this fast) and packaging (packages must act as balanced transmission lines, including the connection to the chip and the board). Even I/O interfaces such as PCIe are integrated, leading to more I/O pins and increased packaging costs, along with improvements in throughput and latency. Memory controller, interconnect fabric, I/O controllers and other I/O blocks are anticipated to occupy approximately 30% of die area.

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 consider process variation during design and analysis, along with the impact of variation on parametric (binned) yield. The need for power management will require a combination of techniques from several component technologies, including the following.

- application-, OS-, compiler- and architecture-level optimizations including parallelism and heterogeneous-core aware scheduling of tasks, as well as process-, operating condition-, and aging-adaptive voltage and frequency scaling;
- process innovations including increased use of silicon-on-insulator (SOI);
- 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; and
- 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., giga operations per second (GOPS) per mW) (*Design*, *PIDS*, *Assembly and Packaging*); and
- *Parametric yield at volume production* (*Lithography*, *PIDS*, *FEP*, *Design*).
- *MPU scaling. Adequate return of investment must be obtained on scaling of cores, as application-, operating system-, compiler- and architecture-level improvements all seek maximal use of resources for highest possible throughput and lowest possible idle power.*

MIXED-SIGNAL DRIVER

RF/mm-wave/analog/mixed-signal chips are those that at least partially deal with input signals whose precise values matter. This broad class of functions encompasses RF, mm-wave, 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 (such as mm-wave, RF and analog design styles) and process technologies (e.g., III-V, silicon-germanium, embedded passives) present severe challenges to cost-effective CMOS integration. However, clever system combinations of RF or mm-wave or analog and digital circuitry, like digital calibration of precision analog/RF/mm-wave circuits or analog/RF/mm-wave self test features, offer potential for functionality and cost scaling at almost the same rate as digital circuits. Reduced accuracy in analog circuits designed in nanoscale CMOS can often be traded off against speed which comes almost for free in those process nodes. Nevertheless, it should be emphasized that a thick-oxide 180-nm MOSFET in a 32-nm CMOS process has significantly improved analog performance when compared to a 180-nm MOSFET in the 180-nm node.

The need for precision clearly 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. RF, mm-wave and 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, and their accuracy is still insufficient for many problems in low-power analog, high speed mixed-signal, RF and mm-wave design. Synthesis tools must concentrate on analog- and RF/mm-wave specific layout synthesis and do not currently take into account all precision matching and electromagnetic-field needs for such designs. Manufacturing test for AMS circuits still needs to be improved but the trend towards SOC also gives opportunities for analog, RF and mm-wave built-in self test (BIST). Finally, practical reuse of building blocks in a scaled version of a device is complicated due to missing verification automation and overwhelming device metallization parasitics in nanoscale CMOS nodes. For analog designs, regression techniques must be enhanced by methods for verifying parametric compliance ranges.

Most analog and RF circuitry in today's high-volume applications is part of SOCs. This trend is now extending into the mm-wave domain. 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. When formulating an analog and mixed-signal (AMS) roadmap, simplification is necessary because there are many different circuit types and topologies. Since the 2001 ITRS edition, this section has discussed four basic analog and RF circuits. Those are not only critical components, but their performance requirements are also representative and most important for RF, mm-wave and analog parts of the SOC. Additionally, in the 2011 edition we have introduced a fifth building block and have expanded the requirements for the amplifiers to include broadband applications.

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

The design and process technology used to build these basic RF, analog/mixed-signal and high-speed digital circuits also determines the performance of many other mixed-signal circuits. Thus, the performance of these five circuits, as described by figures of merit (FoMs), is a good basis for a mixed-signal roadmap. (Future roadmap editions will likely add power management requirements, which are now also a dominant design consideration.)

The following discussion develops these FoMs in detail. Unless otherwise noted, all parameters (e.g., gain, *G, noise figure, NF*) are given as absolute values and not on a decibel scale. Preferences for specific solutions to given design problems are avoided. Instead, different types of solutions are encouraged since unexpected solutions have often helped to overcome barriers. (Competition, such as 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.[17] To avoid such situations, the figures of merit correlate to the analog and RF needs of a mobile communication platform. Lastly, this section evaluates 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 to a significant advance of analog, RF and mm-wave circuit performance as well as to realistic and feasible technology advances. These parameters are given in the *RF and Analog/Mixed-signal Technologies Chapter*.

LOW-NOISE AMPLIFIER

Digital processing systems require interfaces to the analog world. Prominent examples of these interfaces are transmission media in wired or wireless communication. The low-noise amplifier (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 (Long Term Evolution (LTE), wideband code division multiple access (W-CDMA), wireless local area network (WLAN), global positioning system (GPS), Bluetooth, automotive radar, passive imaging, 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 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.

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/factor of the amplifier, F. However, (F-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 $N_{amplifier}$ (or its equivalent noise temperature) and the noise already present at the input N_{input} (or the reference temperature) 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 a tuned narrow-band 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 and knowing that $(F-1)$ increases linearly with the operating frequency f. In a well-designed LNA, $OIP3 = G \times IIP3$ is limited by the supply voltage and should not be a function of frequency. With these approximations and assumptions, a figure of merit (FoM_{LNA}) for tuned LNAs is defined:

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

Making further simplifying assumptions in which the output 3rd compression point is limited by $V_{DD} \times I_{DC} = P$, and neglecting "design intelligence", the evolution of the FoM_{LNA} with technology scaling can be extrapolated from the transistor F_{MIN} at frequency f to obtain an upper physical limit of the LNA FoM:

$$FoM_{LNA-UL} = \frac{f}{(F_{MIN}-1)} \quad [2]$$

Future trends of relevant device parameters for LNA design, including transistor minimum noise figure/factor, F_{MIN}, maximum oscillation frequency, f_{max}, quality factor of inductors, and RF supply voltages are shown in the *RF and Analog/Mixed-signal Technologies Chapter*. In the long term, linearity issues in particular may increasingly be solved by digital calibration techniques.

As mentioned, in certain applications, the amplifier bandwidth is more important than its linearity. In such cases, a modified FoM can be derived which augments [1] with the relative bandwidth of the amplifier, BW_R:

[17] *Certain cases of application are omitted for the sake of simplicity, and arguments are given for the cases selected. Considerations focus on 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.*

$$FoM_{LNA-pi} = \frac{G \cdot IIP3 \cdot f \cdot BW_R}{(F-1) \cdot P} \quad [3]$$

Here, the worst-case noise figure measured over the amplifier bandwidth must be included. In passive imaging LNAs, the linearity figure of merit is irrelevant and can be removed altogether.

Finally, in optical communication systems and in would-be future optical chip-to-chip interconnect links, a low noise broadband transimpedance amplifier, or TIA, plays the same role as the tuned LNA does in wireless systems. The performance metrics of the TIA are the transimpedance gain, Z [Ohm], power consumption, P [mW], the maximum bit rate, R_B [Gb/s], which replaces f in [1], the maximum input current for nominal error-rate operation, I_{MAX} [mA], and the rms input referred noise current, $i^{rms}{}_n$ [uA]:

$$FoM_{TIA} = \frac{Z \cdot I_{MAX} \cdot R_B}{i^{rms}{}_n \cdot P} \quad [4]$$

As in the case of tuned LNAs, for roadmapping purposes, the input equivalent noise current can be linked to the minimum noise figure of the transistor, I_{MAX} is limited by the supply voltage and by the transistor breakdown voltage, while the product between transimpedance gain and data rate scales with the cutoff frequency, f_T.

VOLTAGE-CONTROLLED OSCILLATOR

The voltage-controlled oscillator (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 first to minimize the phase noise (or, equivalently, timing jitter of the generated waveform) and second to minimize the power consumption that achieves the minimum phase noise requirement. From these parameters a figure of merit (FoM_{VCO}), measured in W^{-1}, is defined:

$$FoM_{VCO} = \left(\frac{f_0}{\Delta f}\right)^2 \frac{1}{L\{\Delta f\} \cdot P} \quad [5]$$

Here, f_0 is the oscillation frequency, $L\{\Delta f\}$ is the phase noise power spectral density in a 1Hz band measured at a frequency offset Δf from f_0 and taken relative to the carrier power, and P represents the power consumption of the VCO core. FoM_{VCO} is also often expressed in dB relative to 1mW.

All the measurements must be carried out at room temperature (25 °C) for an accurate VCO performance comparison. Because FoM_{VCO} peaks usually at much lower current density than that at which the minimum phase noise is reached, the FoM must be calculated at the current density where the best VCO phase noise is measured.

This definition recognizes that, in a given transistor technology and at a specified frequency offset, Δf, the phase noise and the output power of the VCO decrease with the square of the frequency. 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 the amplitude of the oscillation, the quality factor of the LC tank, the noise of the active and passive components in the VCO, 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 maximum allowed voltage swing of the active elements in the VCO, inversely proportional to V_{dd}, and proportional to the square of the quality factor of the LC tank. The value of the chosen bias current density is a compromise between minimizing the contribution of *1/f* noise, noise figure, and keeping the amplitude of the oscillation sufficiently high. In this way, FoM_{VCO} is linked to technology development, primarily through f_{MAX}, breakdown voltage and minimum noise figure.

As frequency is increased, layout parasitics are more predominant, and achieving wide VCO tuning range becomes more challenging. The $FoM_{VCO\text{-}T}$ [W^{-1}] is defined to normalize the VCO performance to carrier frequency, carrier offset, DC power, and frequency tuning range:

$$FoM_{VCO-T} = \left(\frac{f_0}{\Delta f}\right)^2 \left(\frac{FTR}{10}\right)^2 \frac{1}{L\{\Delta f\} \cdot P} \quad [6]$$

Here, *FTR* [in percentage points] is the normalized VCO frequency tuning range, and is equal to $200(F_{max}-F_{min})/(F_{max}+F_{min})$. F_{max} and F_{min} are the maximum and minimum VCO oscillation frequencies. As with the previous FoM,

the measurements must be carried out at room temperature (25 °C), and the FoM must be calculated at the current density where the best VCO phase noise is measured. The FoM_{VCO-T} is derived from the expanded Leeson's model and was introduced in [1]. It takes into account the phase noise degradation due to the VCO gain, and it is normalized to a 10% frequency tuning range which is usually required for RF systems. FoM_{VCO-T} is also often expressed in dB relative to 1mW, and as the VCO performance is increased the FoM_{VCO-T}, which is a positive number, increases.

It should be noted that, in mm-wave VCOs, the output power, P_{out}, is often added to the numerator to reflect the difficulty of generating power at frequencies approaching the f_{MAX} of the transistor technology. This is particularly important for automotive radar applications. A third VCO FoM is employed in these situations:

$$FoM_{VCO-PA} = \left(\frac{f_0}{\Delta f}\right)^2 \frac{P_{OUT}}{L\{\Delta f\} \cdot P} \quad [7]$$

As for the previous FoM, the measurements must be carried out at room temperature (25 °C), and the FoM must be calculated at the current density where the best VCO phase noise is measured. CMOS cross-coupled VCOs tend to score better with FoM_{VCO} while Colpitts, bipolar VCOs rank highest with FoM_{VCO-PA}.

In addition to technology scaling-related trends, a further design-related trend towards digitally-controlled oscillators or digital PLLs is observed. Such VCOs have better flexibility and cost structure; hence, they are preferred especially in regimes where different radios are integrated onto a single chip.

Figure SYSD12 shows the evolution of FoM_{VCO-T} (in dB) for mm-wave VCOs integrated in CMOS technologies ranging from the 130 to the 45-nm node. The overall trend versus frequency is negative, with a slope of slightly higher than 20 dB per octave. This reflects one of the challenges with mm-wave VCO design which is to achieve good tuning range. The negative trend may also reflect the lack of data in 45-nm CMOS. There is no clear technological advantage between 45-nm and 130-nm CMOS for the FoM_{VCO-T}.

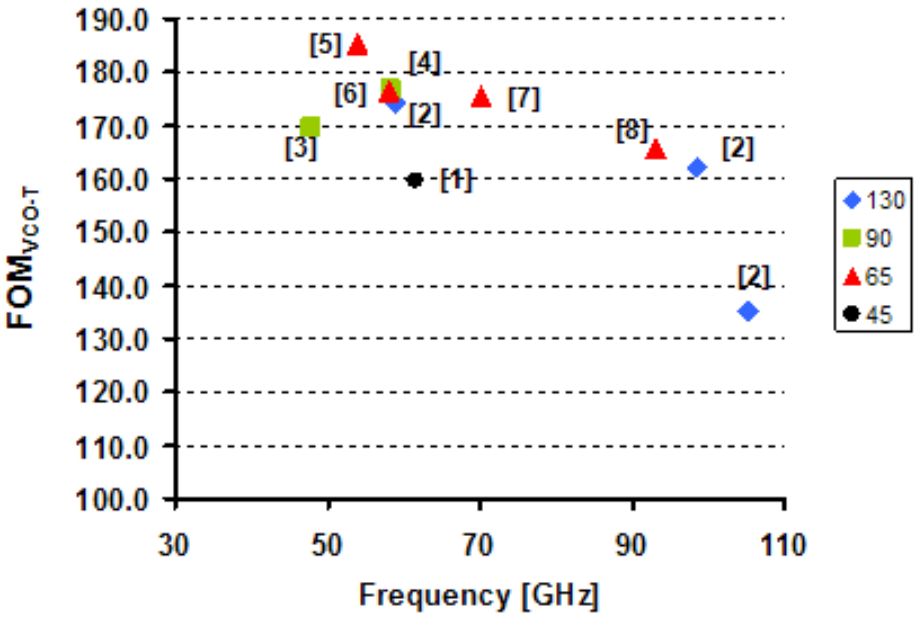

Figure SYSD12 VCO Performance for Mm-Wave Circuits

POWER AMPLIFIER

Power amplifiers (PAs) 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, due to technological issues such as low breakdown voltage and low-Q passives, are restricted to applications where relatively small transmit power is needed. For discrete PAs with higher transmit power (maybe integrated within a

SIP), other technologies like bipolar or compound semiconductor technologies have advantages (*RF and Analog/Mixed-Signal Technologies Chapter*).

To establish a performance figure of merit, several key parameters must be taken into account. These include saturated output power, P_{out}, power gain, *G*, carrier frequency, *f*, linearity (in terms of the 1-dB output compression point, OP1dB), 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. In addition, linearity issues in the future may increasingly be solved by digital calibration techniques and fully digital PAs. To remain independent of the design approach and the specifications of different applications, this parameter is omitted in the figure of merit. To compensate for the 20 dB/decade roll-off of the PA's transistor RF-gain, a factor of f^2 is included into the figure of merit. This results in:

$$FoM_{PA} = P_{OUT} \cdot G \cdot PAE \cdot f^2 \quad [8]$$

Finally, restricting to the simplest saturated PA architecture with highest efficiency (class B or F) and making further simplifications where a 50% PAE is assumed for CMOS and SiGe PAs, enables correlation between the FoM and device parameters such as the peak f_T current density or $I_{ON}/2$, where the maximum saturated power is typically obtained for a SiGe HBT or MOSFETs, respectively, the device breakdown voltage, and f_{max}.

$$FoM_{PA\_HBT} = \frac{J_{pfT}}{2} \min\{V_{DD}, \frac{BV_{CEO}}{2}\} \cdot \min\{MSG, \frac{f_{MAX}^2}{f^2}\} \cdot 0.5 \cdot f^2 \quad [9]$$

$$FoM_{PA\_MOS} = \frac{I_{ON}}{4} \min\{V_{DD}, \frac{BV_{GD}}{2}\} \cdot \min\{MSG, \frac{f_{MAX}^2}{f^2}\} \cdot 0.5 \cdot f^2 \quad [10]$$

This figure of merit is in [GHz^{2}W/mm]. Another important technology parameter is the quality factor of the passive matching elements, assumed infinite in [9] and [10] above. From the scaled device parameters for future technology generations (see the *Power Amplifier Tables in RF and Analog/Mixed-signal Technologies Chapter*), we can predict future PA FoM values. FoMs of best-in-class CMOS PAs have increased in recent years, strongly correlated with progress in active and passive device parameters and with the introduction of digitally-enhanced CMOS PAs (digital PAs) operated in saturated power mode. However, most commercial PAs in cellular phones and WLAN applications, and practically all mm-wave PAs, continue to be fabricated in SiGe HBT and III-V technologies.

ANALOG-TO-DIGITAL CONVERTER

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

$$FoM_{ADC} = \frac{2^{ENOB_0} \cdot \min\{f_{sample}, (2 \cdot ERBW)\}}{P} \quad [11]$$

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;[18] in some cases, it is possible to determine performance requirements of the design from the performance requirements of a critical subcircuit. The device parameters are relevant for the different ADC designs (refer to the data in the *RF and Analog/Mixed-Signal Technologies 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. 2011 best-in-class ADCs were approximately 2100 [giga-conversion-steps per second and watt] for stand-alone CMOS/bipolar CMOS (BiCMOS), and approximately 800 [giga-conversion-steps per second and watt] for embedded CMOS. 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.

SERIALIZER-DESERIALIZER

The serializer-deserializer (SerDes) is a key component in wire line and fiber optic communication systems which are increasingly implemented in CMOS. It includes an N:P multiplexer, a P:N demultiplexer (where N > P and P is typically 1 or 4), a (clock and data recovery circuit) CDR and a (clock multiplication unit) CMU. The highest serial data rates in single-chip serializers or deserializers have reached 56 Gbs while multi-chip or parallel-lane solutions exceed 100 Gbs. A trend towards integrating multi-bit high-speed DACs in the output section of serializers, to allow for higher-order modulation formats such as 16 QAM, has become apparent in recent years. A figure of merit FoM_{SerDes} [Gbs/mW] can be defined for this family of circuits by recognizing that the key performance goals are to maximize the data rate, R_B, while reducing power consumption for a given Mux-DeMux ratio.

$$FoM_{SerDes} = \frac{R_B \cdot MuxDeMux_{ratio}}{P} \quad [12]$$

The key device parameters are seen to be f_T (typically a device technology with $f_T \geq 3 \times R_B$ is required), the peak f_T current density (lower is better), intrinsic slew rate S_{Li} and the CML delay. From the predicted evolution of device parameters for future technology generations (see the SerDes Tables in the *RF and Analog/Mixed-signal Technologies Chapter*), we can deduce future FoM_{SerDes} values.

MIXED-SIGNAL EVOLUTION

Cost estimation—Evolution of the mixed-signal driver, including its scope of application, is completely determined by the interplay between cost and performance. 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. 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. However, this is likely to change during the next 5-10 years with multi-radio and phased array integration on a single die. Related considerations include the following.

- Analog area can sometimes be reduced by shifting the partitioning of a system between analog and digital parts (for example, auto-calibration of ADCs, linearity tuning of PAs, digital PAs).
- 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.

[18] *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.*

- Test cost can dominate the production cost at mm-wave frequencies. Implementing self-test in such applications can have the most dramatic impact on reducing the overall production cost.
- 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. It is possible to quantify mixed-signal cost by first restricting our attention to high-performance applications, since these also drive technology demands. Next, note that RF and analog features are embodied as high-performance passives or analog transistors, and that area can be taken as a proxy for cost.[19] 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 tables in the *Wireless 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. In parallel, serious efforts are on-going to alter system architectures in a direction where passives are replaced by digital circuitry.

Estimation of technology sufficiency—Figure SYSD13 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 towards the upper right. The data show a technological "barrier-line" moving with an order of magnitude per ten years for ADCs for a power consumption of 1W. Most of today's ADC technologies (silicon, SiGe, and III-V compound semiconductor technologies and their hybrids) lie below the 1W barrier-line, and though near-term solutions for moving the barrier-line more rapidly are unknown, the 2010 position (2.13 GHz/mW) of the barrier enables emerging high data-rate communication fields with acceptable dissipation in the conversion function.

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 80 dB 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 at a power consumption of less than 1W. For applications that need high-performance PAs, often a SIP solution with SiGe heterojunction bipolar transistors (HBTs) and III-V devices for the PA and CMOS for the other parts of the analog front-end are the best choice.

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 SYSD13 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 and at increasingly higher frequency and/or data rates will continuously drive the semiconductor industry into such new products and markets.

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

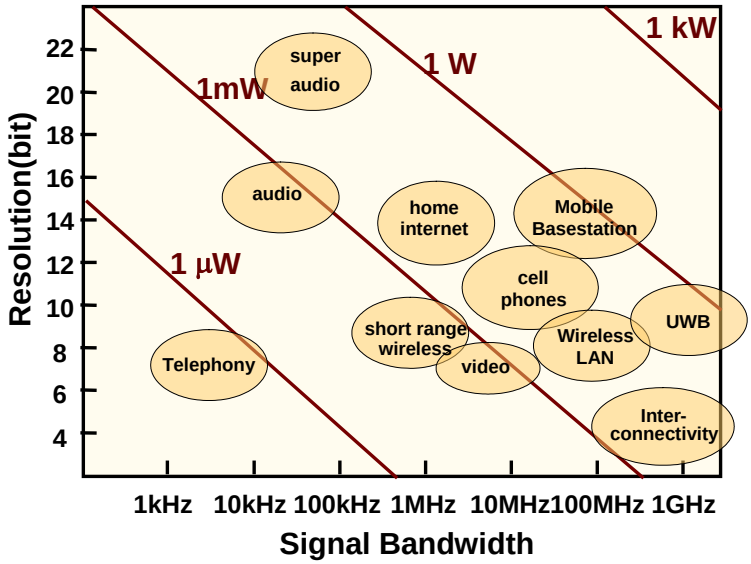

Figure SYSD13 Recent ADC Performance Needs for Important Product Classes

MIXED-SIGNAL CHALLENGES

It is important to separate mm-wave and, to a less extent, RF design from analog design. Millimeter-wave systems have greatly benefited from CMOS, BiCMOS and III-V device scaling. By designing in the current rather than in the voltage domain, and by biasing MOSFETs at relatively large current densities of 0.2-0.4 mA/µm or higher, RF and mm-wave circuits and systems in nanoscale CMOS can be made relatively robust to threshold voltage variation and leakage while taking advantage of higher transistor f_T, f_{MAX} and lower noise figure.

For most of today's mixed-signal designs operating below 3 GHz, for all PAs—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, imply a decrease of the maximum achievable signal level. This has a strong impact on mixed-signal product development for SOC solutions. Improvements in simulation accuracy for parametric behavior do not scale with the rate at which lithographic complexity increases in the most advanced CMOS process nodes. Therefore, fixed layout rules are needed, which limit flexibility and potential reuse. Even though first-pass success of analog, RF and mm-wave designs is increasing due to better simulation tools and models, the typical development time for new mixed-signal parts is still longer than for digital and memory parts. If not area-constrained, 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; however, such reuse sometimes depends on a second type of MOSFET that does not scale its maximum operating voltage and has a similar or better electrical behavior. This challenge has led to the specification in the *PIDS Chapter* of a mixed-signal CMOS transistor that uses a higher analog supply voltage and improves across multiple digital technology generations. Even with such a device, it is time-consuming to scale an analog block to a new technology node due to the lack of any efficient methodology for parametric verification. Furthermore, such a scaling approach minimizes cost benefits from analog scaling, and is therefore undesirable for designs with larger amounts of analog content. In such cases, complete redesigns even including architectural changes are still common. In summary, voltage reduction and development time of analog circuit blocks are still major obstacles to low-cost and efficient scaling of mixed-signal functions. 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*);
- *Increasingly overwhelming layout parasitics in nanoscale CMOS* which render schematic level design useless before layout extraction (*Modeling and Simulation*, *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, more accurate compact models, 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, inductive effects in the immediate metalization on top of the device, 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 and mm-wave 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*); and
- *Lack of reuse-oriented parametric regression suites for analog verification*.

EMBEDDED MEMORY DRIVER

SOC designs contain an increasing number and variety of embedded RAM, read-only memory (ROM), and register file memories. Interconnect and I/O bandwidths, design productivity, and system power limits all point to a continuing trend toward higher levels of memory integration in microelectronic systems. Driving applications for embedded memory technology include code storage in reconfigurable applications (such as 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 I/O 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 integration alternatives. Levels of logic-memory integration will also reflect tradeoffs in hardware-software partitioning (for example, software is more flexible, but must be booted and consumes more area) as well as code-data balance (software must be available to fill code memory, and both non-volatility and applications must be present for data memory). I/O pin count and signaling speeds determine how system organization trades off bandwidth versus storage, e.g., (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 versus 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 such as 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, and 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 static random-access memory (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.

Memory cell size and performance, due to high multiplicities of instantiation, have very direct impact on cost and performance. Thus, the amount of engineering work spent for optimization is much higher compared to all other basic circuits discussed here. *Tables SYSD3a and SYSD3b* give technology requirements for the three currently dominant types of embedded memory: CMOS embedded SRAM, embedded non-volatile memory (NVM), and embedded 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 generation. Voltage scaling involves multiple considerations, such as the relationship between retention time and read operating voltage, or the impact of supply and threshold voltage scaling on PMOS device requirements starting at 45 nm. More nascent ferroelectric RAM, magnetoresistive RAM, and phase-change memory technologies are discussed in the *Emerging Research Devices Chapter*.

Table SYSD3a Embedded Memory Requirements—Near-term

Year of Production	*2007*	*2008*	*2009*	*2010*	*2013*
DRAM ½ Pitch (nm)	*65*	*55*	*50*	*45*	*35*
CMOS SRAM High-performance, low standby power (HP/LSTP) DRAM ½ pitch (nm), Feature Size – F	*65*	*65*	*65*	*45*	*35*
6T bit cell size (F^2) [1]	**140F²**	**140F²**	**140F²**	**140F²**	**140F²**
Array efficiency [2]	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**
Process overhead versus standard CMOS – #added mask layers	**2**	**2**	**2**	**2**	**2**
Operating voltage – V_{dd} (V) [4]	**1.1**	**1/1.1**	**1/1.1**	**1**	**0.9/1**
Static power dissipation (mW/Cell) [5]	**3E-4/1E-6**	**3E-4/1E-6**	**3E-4/1E-6**	**5E-4 / 1.2E-6**	**1E-3/1.5E-6**
Dynamic power consumption per cell (mW/MHz) [6]	**4.5E-7/7E-7**	**4E-7/6.5E-7**	**4E-7/6E-7**	**3E-7/5E-7**	**2.5E-7/4.5E-7**
Read cycle time (ns) [7]	**0.3/1.5**	**0.3/1.5**	**0.3/1.5**	**0.2/1.2**	**0.15/0.8**
Write cycle time (ns) [7]	**0.3/1.5**	**0.3/1.5**	**0.3/1.5**	**0.2/1.2**	**0.15/0.8**
Percentage of MBU on total SER	**16%**	**16%**	**16%**	**32%**	**64%**
Soft error rate (FIT/Mb) [8]	**1150**	**1150**	**1150**	**1200**	**1250**
Embedded Non-Volatile Memory (code/data), DRAM ½ pitch (nm)	*90*	*90*	*90*	*90*	*65*
Cell size (F^2) – NOR FLOTOX [9]	**10F²**	**10F²**	**10F²**	**10F²**	**10F²**
Array efficiency – NOR FLOTOX [10]	**0.6**	**0.6**	**0.6**	**0.6**	**0.6**
Process overhead versus standard CMOS – #added mask layers [3]	**6–8**	**6–8**	**6–8**	**6–8**	**6–8**
Read operating voltage (V)	**2V**	**2V**	**2V**	**1.8V**	**1.5V**
Write (program/erase) on chip maximum voltage (V) – NOR [11]	**10V**	**10V**	**10V**	**10V**	**10V**
Static power dissipation (mW/cell) [5]	**1.00E-06**	**1.00E-06**	**1.00E-06**	**1.00E-06**	**1.00E-06**
Dynamic power consumption per cell (mW/MHz) [6]	**6.00E-09**	**6.00E-09**	**6.00E-09**	**6.00E-09**	**4.00E-09**
Read cycle time (ns) – NOR FLOTOX [7]	**15**	**15**	**15**	**15**	**15**
Program time per cell (µs) – NOR FLOTOX [12]	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Erase time per cell (ms) – NOR FLOTOX [12]	**10.0**	**10.0**	**10.0**	**10.0**	**10.0**
Data retention requirement (years) [12]	**10**	**10**	**10**	**10**	**10**
Endurance requirement [12]	**100000**	**100000**	**100000**	**100000**	**100000**
Embedded DRAM, ½ pitch (nm)	*90*	*90*	*65*	*65*	*45*
1T1C bit cell size (F^2) [13]	**12–30**	**12–30**	**12–30**	**12–30**	**12–30**
Array efficiency [2]	**0.6**	**0.6**	**0.6**	**0.6**	**0.6**
Process overhead versus standard CMOS – #added mask layers [3]	**3–5**	**3–5**	**3–5**	**3–5**	**3–6**
Read operating voltage (V)	**2**	**2**	**1.8**	**1.7**	**1.6**
Static power dissipation (mW/Cell) [5]	**1.00E-11**	**1.00E-11**	**1.00E-11**	**1.00E-11**	**1.00E-11**
Dynamic power consumption per cell (mW/MHz) [6]	**1.00E-07**	**1.00E-07**	**1.00E-07**	**1.50E-07**	**1.60E-07**
DRAM retention time (ms) [12]	**10**	**10**	**4**	**4**	**2**
Read/Write cycle time (ns) [7]	**3**	**3**	**2**	**2**	**1.6**
Soft error rate (FIT/Mb) [8]	**60**	**60**	**60**	**60**	**60**

FIT—failures in time FLOTOX—floating gate tunnel oxide MBU—multiple bit upsets NAND—"not AND" logic operation NOR—"not OR" logic operation

Table SYSD3b Embedded Memory Requirements—Long-term

Year of Production	*2016*	*2019*	*2022*	*2024*	*2026*
DRAM ½ Pitch (nm)	*25*	*18*	*13*	*10*	***Not currently projected***
CMOS SRAM High-performance, low standby power (HP/LSTP) DRAM ½ pitch (nm), Feature Size – F	*25*	*18*	*13*	*10*	
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 – #added mask layers [3]	**2**	**2**	**2**	**2**	
Operating voltage – V_{dd} (V) [4]	**0.8/0.9**	**0.7/0.8**	**0.7/0.8**	**0.7/0.8**	
Static power dissipation (mW/cell) [5]	**2E-3/2E-6**	**3E-3/2.5E-6**	**5E-3/3E-6**	**6E-3/3.5E-6**	
Dynamic power consumption per cell (mW/MHz) [6]	**2E-7/4E-7**	**1.5E-7/3E-7**	**1E-7/2E-7**	**0.5E-7/1E-7**	
Read cycle time (ns) [7]	**0.1/0.5**	**0.07/0.3**	**0.07/0.3**	**0.07/0.3**	
Write cycle time (ns) [7]	**0.1/0.5**	**0.07/0.3**	**0.07/0.3**	**0.07/0.3**	
Percentage of MBU on total SER	**100%**	**100%**	**100%**	**100%**	
Soft error rate (FIT/Mb) [8]	**1300**	**1350**	**1400**	**1450**	
Embedded Non-Volatile Memory (code/data), DRAM ½ pitch (nm)	*40*	*32*	*32*	*25*	*20*
Cell size (F^2) – NOR FLOTOX [9]	**$10F^2$**	**$10F^2$**	**$10F^2$**	**$10F^2$**	**10F2**
Array efficiency – NOR FLOTOX [10]	**0.6**	**0.6**	**0.6**	**0.6**	**0.6**
Process overhead versus standard CMOS – #added mask layers [3]	**6–8**	**6–8**	**6–8**	**6–8**	**6-8**
Read operating voltage (V)	**1.3V**	**1.2V**	**1.1V**	**1.0V**	**1.0V**
Write (program/erase) on chip maximum voltage (V) – NOR [11]	**9V**	**9V**	**9V**	**9V**	**9**
Static power dissipation (mW/cell) [5]	**1.00E-06**	**1.00E-06**	**1.00E-06**	**1.00E-06**	**1.00E-06**
Dynamic power consumption per cell (mW/MHz) [6]	**3.50E-09**	**3.00E-09**	**3.00E-09**	**3.00E-09**	**2.50E-09**
Read cycle time (ns) – NOR FLOTOX [7]	**12**	**10**	**10**	**8**	**7**
Program time per cell (µs) – NOR FLOTOX [12]	**1**	**1**	**1**	**1**	**1**
Erase time per cell (ms) – NOR FLOTOX [12]	**10**	**10**	**10**	**10**	**10**
Data retention requirement (years) [12]	**10**	**10**	**10**	**10**	**10**
Endurance requirement [12]	**100000**	**100000**	**100000**	**100000**	**10000**
Embedded DRAM, ½ pitch (nm)	*35*	*25*	*25*	*20*	***Not currently projected***
1T1C bit cell size (F^2) [13]	**12–50**	**12–50**	**12–50**	**12–50**	
Array efficiency [2]	**0.6**	**0.6**	**0.6**	**0.6**	
Process overhead versus standard CMOS – #added mask layers [3]	**3–6**	**3–6**	**3–6**	**3–6**	
Read operating voltage (V)	**1.5**	**1.5**	**1.5**	**1.5**	
Static power dissipation (mW/cell) [5]	**1.00E-11**	**1.00E-11**	**1.00E-11**	**1.00E-11**	
Dynamic power consumption per cell (mW/MHz) [6]	**1.70E-07**	**1.70E-07**	**1.70E-07**	**1.70E-07**	
DRAM retention time (ms) [12]	**1**	**1**	**1**	**1**	
Read/Write cycle time (ns) [7]	**1.3**	**1**	**1**	**1**	
Soft error rate (FIT/Mb) [8]	**60**	**60**	**60**	**60**	

FIT—failures in time FLOTOX—floating gate tunnel oxide MBU—multiple bit upsets NAND—"not AND" logic operation NOR—"not OR" logic operation

Definitions of Terms for Tables SYSD3a and SYSD3b:
[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 needed over standard CMOS logic process in 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 as I_standby × V_{dd} (off-current and V_{dd} are taken for 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 0.8 V_{dd} swing on the Bitline (BL). Determine the WL capacitance per cell (CWL) and BL capacitance per cell (CBL). Then: dynamic power consumption per MHz per cell = V_{dd} × CWL (per cell) × (V_{dd}) + V_{dd} × CBL (per cell) × (V_{dd}) ×10^6.
[7] Read cycle time is the typical time it takes to complete a READ operation from an address. (This number refers to a random access and not to a sequential one.) Write cycle time is the typical time it takes to complete a WRITE operation to an address. Both cycle times depend 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. 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, and standard NAND architecture. Data refer to the NVM device requirements table in the PIDS Chapter.
[11] Maximum voltage required for operation, typically used in WRITE operation. Data refer to the NVM device requirements table in the PIDS Chapter.
[12] 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.
[13] Size of the standard cell for embedded trench DRAM cell. Data refers to the DRAM requirements table in the PIDS Chapter.

CONNECTION TO SYSTEM-LEVEL ROADMAP: SOC-CP POWER CONSUMPTION PILOT

The iNEMI roadmap is to systems and boards as the ITRS roadmap is to chips. The ITRS Design ITWG is continuing its efforts to selectively align aspects of the iNEMI and ITRS roadmaps, so as to ensure that system drivers have realistic input parameter targets. A 2007 pilot study regarding this alignment focused on (a) the SOC Consumer Portable driver and (b) the power and energy aspects of the design. The parameters selected for alignment include voltage supply, energy consumption, standby power, runtime before recharge, operating temperature range, thermal design power (hottest chip), maximum current per chip, thermal design flux, cooling method and passives usage. Several parameters did not exist either in the ITRS or iNEMI models, showing the need for deeper future collaboration. Figure SYSD13 shows the two key comparisons that were possible during this exercise.

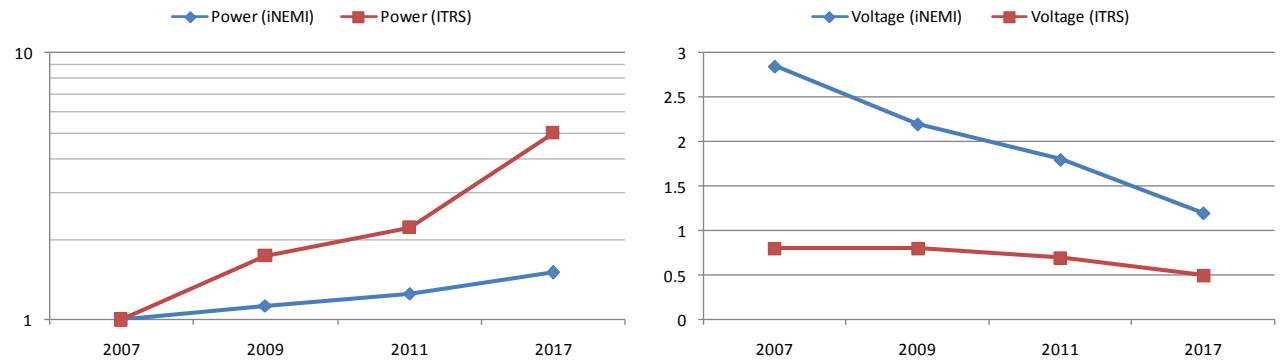

Figure SYSD14 ITRS-iNEMI System-to-Chip Power Comparison Trends

As the left side of the figure shows, the gap between system-level and chip level power specifications is growing in relative terms, which highlights a potential crisis as chips become too hot for the board to handle. The right side of the figure shows another potentially critical issue: voltage supplies seem to be falling at the system level much faster than at the chip level, which may indicate pressure to reduce voltage supplies at the chip level.