# The ITRS Design Technology and System Drivers Roadmap: Process and Status

Andrew B. Kahng
CSE and ECE Depts., Univ. of California at San Diego
abk@ucsd.edu

## ABSTRACT

The Design technology working group (TWG) is one of 16 working groups in the *International Technology Roadmap for Semiconductors* (ITRS) effort. It is responsible for the ITRS' Design Chapter, which roadmaps design technology requirements and potential solutions for elements of the semiconductor supply chain that are produced by the electronic design automation (EDA) industry. The Design TWG is also responsible for the ITRS' System Drivers Chapter, which roadmaps the key product classes that drive the leading-edge requirements for process and design technologies. Through these activities, the Design TWG sets a number of fundamental parameters in the overall ITRS: layout density, die size, maximum on-chip clock frequency, total chip power, SOC and MPU architecture models, etc. This paper reviews the process by which the Design TWG evolves its roadmap content, and some of the key modeling and roadmapping questions that the semiconductor and EDA industries will face in the near term.

## 1. INTRODUCTION

As noted in [13], technology roadmaps seek "precompetitive" specifications of future technical requirements and challenges. Potential solutions are identified, investigated, pruned, productized, standardized, and delivered to the marketplace – in a synchronized, timely, and cost-effective manner – to ensure a continued stream of technology benefits. The *International Technology Roadmap for Semiconductors* (ITRS) [22] is one of the most successful roadmapping efforts ever: well over 1000 scientists and engineers worldwide collaborate to synchronize a wide range of industries and technologies (automated test equipment, assembly and packaging, photomask, electronic design automation (EDA), lithography, interconnect, device, etc.) so that the "Moore's Law" semiconductor value proposition can continue. The broad scope of the ITRS is essential, e.g., the roadmap for design technology must comprehend (i) lithography and restricted design rules; (ii) die stacking and 3D integration; (iii) device and interconnect electrical performance, variability and robustness; (iv) ATE, BIST and BISR overheads and production costs; (v) product-level trajectories for RF blocks, IO bandwidth and processing capability; and many other futures. The ITRS's 15-year horizon reflects the lead times needed to identify and develop production-worthy technologies.

All technology roadmaps struggle with the tension between "roadmapping" and "extrapolation". An uncalibrated roadmap lacks credibility. On the other hand, unthinking extrapolation from historical data risks "driving by the rear-view mirror", and can result in absurd projections at the 15-year horizon. Meaningful roadmapping of technology requirements and potential solutions requires at least the following elements.

- *Metrics.* What cannot be measured cannot be tracked or improved. EDA tools heuristically address large-scale, NP-hard optimizations, and design quality is strongly determined by flow and methodology ("it's the magician, not the wand"). Thus, it is challenging to identify metrics that capture the progress of design technology.
- *Understanding of contexts and needs for technology.* Contexts ranging from process technology to market forces affect the need for technology. For example, the trajectory of mobile consumer SOC products has driven rapid innovation in low-power design techniques spanning embedded memory design, power and clock gating, dynamic voltage scaling, etc. At the same time, these low-power design techniques must acknowledge process and material attributes such as discreteness of FinFET device widths starting at the 16nm foundry node, or increasingly dominant reliability and aging mechanisms.
- *Holistic selection of potential solutions.* Technology roadmapping must holistically model and predict impacts of potential technology solutions, at many levels. For example, solutions to a "power crisis" in IC design may come from manufacturing technologists (e.g., process innovation to reduce Vth variation), device and circuit technologists (introduction of FinFET and resistive RAM), and system designers (heterogeneous multi-core SOC architectures) – as well as design and test technologists (asynchronous design flow, on-chip variability monitoring and adaptivity, etc.). All potential solutions cost money to develop and deploy. Thus, as discussed in [11], a mindset of "shared red bricks" in the semiconductor technology roadmap is critical to achieve proper allocation of R&D resources.[1]

**The ITRS Design Technology Working Group.** The Design technology working group (TWG) is one of 16 TWGs in the ITRS. With over 50 industry and academic contributors from all five regional semiconductor industry associations (USA, EU, Japan, Taiwan, Korea), the Design TWG is responsible for the ITRS Design Chapter, which roadmaps design technology requirements and potential solutions relevant to the EDA industry, and the ITRS System Drivers Chapter, which roadmaps the key product classes that drive leading-edge requirements for process and design technologies.

Figure 1 shows how the Design and System Drivers chapters have consistently evolved over the past decade. First, the Design Chapter gives a *quantified* Design Technology roadmap with metrics, potential solutions, and mappings from requirements to potential solutions. This matches the structure and metrics-oriented "look and feel" of other ITRS chapters. Second, an increasingly comprehensive set of System Drivers has been developed that maintain alignment to key segments of the semiconductor industry. Each update to the System Drivers (e.g., the acknowledgment of a hard platform power limit in the MPU roadmap, starting in 2007) has ripple effects across Overall Roadmap Technology Characteristics (ORTCs) such as layout density, transistor count, die size, chip power and frequency – as well as fundamental technology metrics owned by other technology working groups. These interactions are conceptually depicted in Figure 2.[2] The System Drivers also enable


*DAC'13,* May 29 - June 07 2013, Austin, TX, USA.


[1] In ITRS parlance, a "red brick" is a technology requirement that has no known solution (the term stems from the coloring convention in ITRS technology requirement tables). For example, to solve the problem of poor interconnect RC scaling, are R&D dollars best invested in new dielectric materials, new interconnect and barrier materials, better overlay control, more accurate signal integrity analyses in EDA tools, scalable many-core GALS architectures, or ...? Or, to solve the problem of exploding (and widening) modes and corners in signoff, should variation be reduced in the process itself, or should statistical signoffs be adopted, or should "signoff at typical" be adopted in combination with adaptivity [3], or ...?

[2] In over 17 years of NTRS and ITRS roadmap participation, I have witnessed a steady rise in the prominence of "design" within the ITRS. Originally highly process-centric, the roadmap now increasingly relies on "design-based equivalent scaling" [24] and

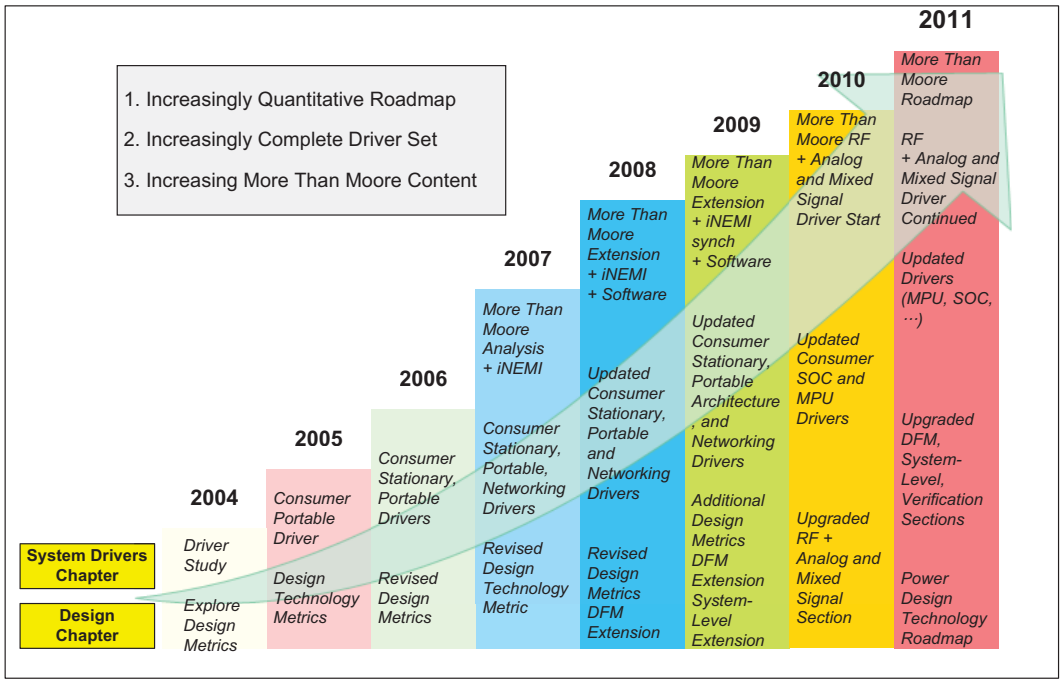


**Figure 1: Roadmap from ITRS System Drivers and Design chapters. [Source: ITRS Design ITWG 2011 Public Conference presentation, December 2011, Songdo, Korea.]**

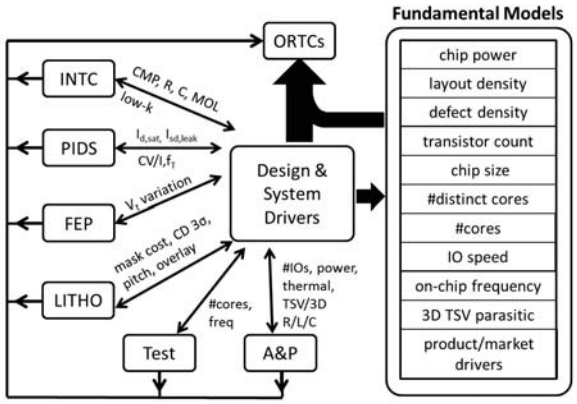


**Figure 2: Increasingly central role of Design TWG in ITRS roadmap definition.**

stronger alignment (cf. "More Than Moore") between the ITRS's chip-level roadmap and system product-level roadmaps such as iNEMI [21].

**Organization of This Paper.** The remainder of this paper is organized as follows. Section 2 outlines the process and overarching objectives that guide the evolution of Design and System Drivers content. Several examples then give the "flavor" of how the roadmap evolves. Two aspects of the System Drivers Chapter are the System Driver model evolution, which is discussed in Section 3, and the "A-factor" approach that underlies projection of density scaling in the ITRS, which is discussed in Section 4. Two aspects of the Design Chapter are the low-power design technology roadmap, which is discussed in Section 5, and the evolution of Design for Manufacturability (Variability, Reliability) content, which is discussed in Section 6. Section 7 concludes with some thoughts on modeling and roadmapping issues that the semiconductor and EDA industries will face in the near term.

## 2. DESIGN TWG GOALS AND PROCESS

Like every other technology working group in the ITRS, the Design TWG places the interests of its industry and R&D community – i.e., EDA and VLSI CAD – first and foremost. In ITRS cross-TWG interactions, the Design TWG must respond to questions such as "How much variability can designers tolerate?" (Lithography TWG) or "What is the $J_{max}$ limit for on-chip global interconnects?" (Interconnect TWG) or "What tradeoff between leakage and drive currents is best for mobile SOCs?" (Process Integration, Devices and Structures (PIDS) TWG). The roadmap for DFT is jointly owned with the Test TWG. The roadmap for off-chip IO bandwidth is jointly owned with the Test TWG and the Assembly and Packaging (A&P) TWG. And the roadmap for 3D/TSV based integration is jointly owned with a number of other TWGs, notably A&P, Test, Interconnect and Front-End Processing (FEP). All of these interactions entail asynchronous, off-line dialogues year-round with designers, EDA technologists and researchers so that perspectives from IC design, and from IC design automation, are correctly represented.

ITRS challenges and technology requirements directly inform the research priorities and funding allocations of a number of government funding agencies and industry consortia worldwide, and the phrase "According to the ITRS, ..." is often given as motivation in academic research papers. Thus, Design TWG activities often include advocacy for the importance of EDA technology and academic research. Furthermore, "key messages" in the Design Chapter can seed future trends in academic research and research funding. Three examples of such advocacy and messaging are as follows.

- *The Design Cost Model.* Although tremendous product differentiation comes from design and design technology, EDA industry revenues, and levels of R&D investment and academic research funding, have been stagnant. With this in mind, quantifying the *value of design technology* has been one of the high-level goals for the Design TWG within the ITRS effort. Since 2001, the Design Chapter has included a highly influential Design Cost model [14] [12] that now encompasses both hardware and software development costs (salary and overhead of engineers, EDA tool cost per seat, interoperability costs, etc.). The cost model quantifies the impact of design technology innovation and resulting productivity improvements. For example, the hardware design costs for a consumer portable SOC design in 2011 are estimated at $25.7M, versus $7708M had design technology innovations between 1993 and 2009 not occurred.
- *Key Messages.* Over the years, the Design TWG has formulated specific key messages within the ITRS. Since 2001, an overarching message has been that "cost of design is the greatest obstacle to continuation of semiconductor roadmap". In the 1998-2001 time frame, the Design TWG also advocated a "Living ITRS" mindset wherein all technology roadmap projections and models could be implemented on a common platform, to enable interoperability and cross-checking for consistency.[3] More specific messages have also been given over the years. For example, in 2009 the Design Chapter's

"More Than Moore" to deliver scaling of semiconductor product value in the face of non-ideal performance, power, density and variability scaling.

[3]The GTX (MARCO GSRC Technology Extrapolation) package [4] for some years provided a realization of this goal, but is no longer maintained.

key messages were that (i) software and system-level design productivity are critical to the roadmap of semiconductor value; (ii) design reliability roadmapping was a necessary addition to the roadmap; (iii) system-level design techniques would ultimately be crucial to managing power; and (iv) design technology innovations must keep on schedule through the end of the roadmap in order to contain design costs. New messages in 2011 and 2012 included (i) roadmapping focus at the design-manufacturing interface has evolved from "manufacturability" to a more general "variability", which now entails an even broader question of how systems will maintain reliability and be resilient; (ii) design technology innovations must keep on schedule through the end of the roadmap in order to contain power; and (iii) the importance of cross-TWG interactions is continually growing, whether for More Than Moore, 3D, Beyond CMOS, or even the basic device and lithography roadmaps.

- *Grand Challenges.* The ITRS Executive Summary calls out a subset of each working group's "difficult challenges", and categorizes these as either "Enhancing Performance" or "Cost-Effective Manufacturing", and as either near-term (within the next seven years) or long-term (between eight and 15 years out). In the 2005-2011 ITRS editions, power management, design productivity, and DFM were consistently listed as near-term grand challenges for design. The roadmap noted that power management challenges would need to be addressed across multiple levels, especially system, design, and process technology. Moreover, to maintain design quality in advanced process nodes, design implementation productivity must improve to the same degree that design complexity is scaled – with improvement of design productivity and IP reuse being key considerations. Long term challenges have evolved from management of leakage power consumption in the 2005-2009 roadmaps to design of concurrent software and design for reliability and resilience in the 2011 roadmap.

The Design TWG operates in a distributed manner, with each major Design Chapter section or System Driver model maintained by a distinct subteam. Different geographies tend to assume natural responsibilities for content, e.g., European contributors have responsibility for the AMS/RF content, and Japanese contributors have responsibility for the SOC system driver models. New content is constantly developed according to identified gaps in roadmap coverage, e.g., Design Chapter updates in 2009 and 2011 include (i) a 3D/TSV design technology section, (ii) a hardware-related software development cost component for Design Cost model, and (iii) a low-power design technology roadmap. Following ITRS convention, the U.S. TWG co-chairs coordinate worldwide efforts and serve as the editors for all published content.[4]

# 3. KEY SYSTEM DRIVER MODELS

As noted above, the System Drivers Chapter models and projects key semiconductor product classes that create the need for continued semiconductor innovation [5–7]. The 2011 System Drivers Chapter identifies three microprocessor (MPU) drivers (high-performance (HP), cost-performance (CP) and power-connectivity-cost (PCC)) and three System-On-Chip (SOC) drivers (consumer portable (CP), consumer stationary (CS) and networking (NW)).[5] Each driver should provide impetus for specific technology objectives, e.g., the SOC-CP driver drives lower leakage (or standby) power consumption, given the severe battery life requirement of mobile devices. For each MPU and SOC system driver, the ITRS roadmaps scaling of parameters such as number of cores, number of SRAM and logic transistors, layout density, frequency and power.

[4]Resources and dedicated bandwidth in support of the ITRS have not yet recovered from the 2008-2009 economic downturn. All suggestions, participation in ITRS meetings, and other contributions are always welcome; interested individuals should contact the Design TWG co-chairs, Dr. Andrew B. Kahng (abk@ucsd.edu) and Dr. Juan-Antonio Carballo (jantonio@ieee.org).

[5]MPU-HP are server products, e.g., Intel Xeon and AMD Opteron. MPU-CP are desktop products, e.g., Intel Core i7 and AMD Phenom. MPU-PCC are handheld and micro-server products, e.g., Intel Atom and Marvell Armada. SOC-CP are handheld products, e.g., Qualcomm Snapdragon and Samsung Exynos. SOC-CS are products for game consoles, e.g., IBM Cell BE and WonderMedia (Via) WM series. SOC-NW are multi-core network processors, e.g., Broadcom XLP864 and Calxeda ECX-1000.

### MPU Driver Modeling

The ITRS MPU driver model has for many years scaled the number of logic transistors and the number of SRAM transistors by $2\times$ per technology node. Since dimensions shrink by $0.7\times$ per node, and nominal layout density therefore doubles, this simple scaling model allows die size to remain constant across technology nodes.

*MPU Die Size.* The 2009 MPU model update [10] set a constant die area of $260mm^2$ for MPU-HP and $140mm^2$ for MPU-CP. The model for logic density ($D_{tr,logic}$) is

$$D_{tr,logic} = \frac{N_{tr,nand2}}{O_{logic} \cdot U_{logic}} \quad (1)$$

where $N_{tr,nand2}$ (number of transistors in a NAND2 gate) is four, $O_{logic}$ (logic overhead due to design integration) is 2.0 (i.e., 100% area overhead for whitespace), and $U_{logic}$ (the area of a unit NAND2 gate) is calculated using the "A-factor" described below in Section 4. The model for SRAM density ($D_{tr,SRAM}$) [10] is

$$D_{tr,SRAM} = \frac{N_{tr,bitcell}}{O_{SRAM} \cdot U_{SRAM}} \quad (2)$$

where $N_{tr,bitcell}$ is the number of transistors in a SRAM bitcell, $O_{SRAM}$ (overhead due to peripheral circuits) is assumed to be 1.6 (i.e., 60% area overhead), and $U_{SRAM}$ (the area of a unit SRAM bitcell) is calculated using another A-factor, also described in Section 4. While the 2009 MPU model remains accurate with respect to number of cores, or total number of transistors, die areas of recent server MPU products have grown rapidly, reaching $\sim 530mm^2$ in the 2012-2013 time frame. Moreover, the simple model of cores + SRAM does not acknowledge the growth of "uncore" elements (memory controllers, IO controllers, GPU cores, on-chip networking, etc.) in MPU products. These considerations make it likely that the 2013 ITRS edition will see substantial revision of the MPU-HP model with respect to both A-factors and architecture.

*MPU Frequency.* Figure 3 overlays historical changes in the ITRS maximum on-chip frequency roadmap with product data from the Stanford CPUDB [18]. The 2001 System Drivers Chapter observed that rapid MPU frequency increases up to that time had been enabled by reduction in the number of fanout-of-four (FO4) delays per clock period. That is, microarchitecture (aggressive pipelining, with fewer stages of logic per pipeline stage) had been used to increase frequency at a faster rate than the intrinsic growth of device switching speed. At that time (2001), a basic limit of 12 FO4 delays (in which useful computation could be performed during a clock cycle) was being reached, and so the roadmap was modified to improve frequency only as device speeds improved (17%/year improvement in CV/I metric, in the PIDS roadmap).

In 2007, a market-driven platform power limit of 130W per die was acknowledged, and the MPU frequency roadmap was revised to increase by just 8% per year to meet this power limit.[6] The slowing of frequency enabled the PIDS device roadmap to also slow the CV/I improvement to 13%/year, which eased the challenge of managing leakage currents. Subsequently, during the 2009-2011 roadmapping cycle, device technologists found that even the 13%/year CV/I improvement was incompatible with leakage current requirements; hence, the likely scenario for 2013 and beyond is for 4%/year frequency increase in MPU products (still with design-based equivalent scaling in the form of switching factor reductions), along with some limited "headroom" of 8%/year improvement in the device CV/I metric.

### System Driver Futures

During the 12 years since the System Drivers Chapter was introduced, many structural changes have occurred in the marketplace. As these shifts occur, the set of system drivers, and their intrinsic models, are subject to change.

- The SOC-CS driver was introduced at a time when the IBM Cell BE was highly visible in the game console market. Today, game consoles are primarily driven by high-end CPU-GPU fusion products such as AMD A10-5800K, which is essentially an instance of the MPU-CP driver. Thus, the need for an SOC-CS driver may be obsolete.

[6]With this 8%/year frequency growth, a "magic, design-based" 5% reduction per year in the chip's switching activity factor had to be added into the MPU model, to keep MPU power flat. Although actual product frequencies were already visibly flattening, it was felt that a model with 0% frequency increase would stall device and circuit innovation needed by other semiconductor products.

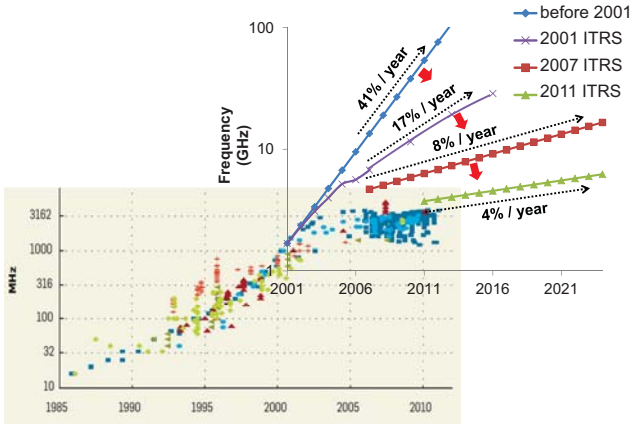


**Figure 3: Frequency scaling roadmap.**

- While the MPU-CP driver originally reflected the desktop PC market and the "shrink" version of the MPU-HP "lead" processor, today we see that desktop processors simply use the same architecture as either server- or handheld-class products. Accordingly, MPU-CP may also be considered for removal as a separate driver.
- If SOC-CS and MPU-CP drivers are both removed from the roadmap, the key remaining drivers will be MPU-HP and SOC-CP. SOC-CP, which reflects the handheld market, is rapidly becoming more general-purpose and integrates GPU IPs such as Mali, PowerVR, etc. The architecture and scaling models for SOC-CP may be considered for change.
- It may also be noted that the MPU-PCC driver is evolving toward the micro-server market and away from the handheld market. Thus, the roadmap for MPU-PCC may need to change as well.
- Even as the MPU-CP and SOC-CS drivers become less important to the technology roadmap, new system drivers may arise from automotive, defense, medical and energy management applications, aligning with recent More Than Moore foci.

# 4. LAYOUT DENSITY A-FACTORS

In the ITRS System Drivers Chapter and Overall Roadmap Technology Characteristics, *A-factors* enable the modeling of unit cell areas of SRAM and standard-cell logic circuit fabrics, in terms of the M1 half-pitch, $F$. SRAM layout density is mainly determined by Mx pitches and poly pitch in a bulk technology. With FinFET devices, the fin pitch ($P_{fin}$) becomes the dominant factor for SRAM layout. On the other hand, the density of *standard cells* is mainly decided by the cell height (in M2 tracks) and the poly pitch. Since the 2009 ITRS, the A-factor for a 6T SRAM bitcell has been $60F^2$, and the A-factor for a 2-input NAND gate has been $175F^2$ [10]. These values are based on various ratios between, e.g., poly, M1, and M2 layer pitches (design rules) as summarized in the left half of Table 1, as well as on the canonical layouts shown in Figures 4(b) and 5(b) [10].

As the industry moves to double-patterning, FinFETs with discrete gate widths, and "middle of line" (MOL) layers to enable local access to transistors, the fundamental A-factor scaling models will likely require significant revisions. For example, in future NAND2 cell layouts, M1 may no longer be the most congested metal layer, so M2 pitch ($P_{M2}$) may shrink to be the same as M1 pitch ($P_{M1}$). Furthermore, with emerging FinFET (multi-gate) devices, fin pitch ($P_{fin}$) cannot be arbitrarily small, and gate width is in quanta of fins. Based on these considerations, the A-factor of the bulk NAND2 layout may evolve to $144F^2$ (Figure 4(b)), i.e., $W_{cell} = 3P_{poly}$, $H_{cell} = 8P_{M2}$, and hence $A_{Bulk,NAND2} = W_{cell} \times H_{cell} = 144F^2$. The area of the FinFET NAND2 layout may be set to $162F^2$ (Figure 4(a)), i.e., $W_{cell} = 3P_{poly}$, $H_{cell} = 9P_{M2}$, and hence $A_{FinFET,NAND2} = W_{cell} \times H_{cell} = 162F^2$.[7] Industry colleagues have observed that contacted poly pitch (CPP) appears more difficult to scale than Mx (local metal) pitch. In other words, Mx pitch seems to be scaling at a rate faster than $0.7\times$ per node, while CPP scales at a rate slower than $0.7\times$ per node, even as the product of the two pitches achieves $0.5\times$ area scaling. Such a trend, if continued, may eventually change the A-factor modeling and A-factor values.

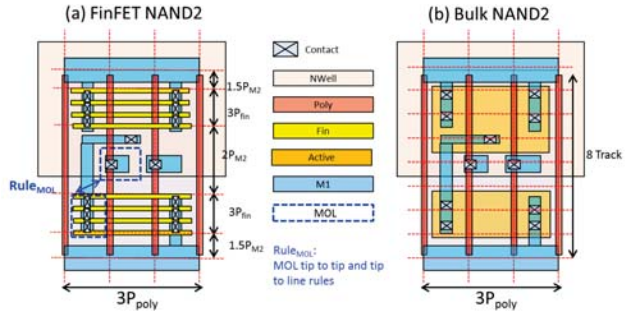


**Figure 4: Layout of NAND2 cells for (a) FinFET and (b) bulk.**

**Table 1: Pitch conversions used in the A-factor models.**

| Layer | Pitch/M1 (2009) | Pitch/M1 (revised) |
|---|---|---|
| $F$ | 0.50 | 0.50 |
| $P_{M1}$ | 1.00 | 1.00 |
| $P_{poly}$ | 1.50 | 1.50 |
| $P_{M2}$ | 1.25 | 1.00 |
| $P_{fin}$ | N/A | 0.75 |
| P/G track width | N/A | 1.50 |

Figures 5(a) and (b) respectively give canonical layouts for FinFET and bulk 6T SRAM bitcells. Each layout uses two poly channels, so the bitcell height is $2P_{poly}$. The width of the bitcell depends on (i) the distances between bitline to wordline on each end; (ii) transistor separations (P to N); and (iii) distance between n-active (N to N); these parameters differ for FinFET and bulk. The width of the 6T bulk SRAM is $5P_{M1}$ from Figure 5(b), and the A-factor of the bulk 6T SRAM is therefore $60F^2$ [10].

Derivation of an A-factor for a FinFET-based 6T SRAM bitcell must consider two main issues. First, a pitch conversion between $P_{fin}$, $P_{M1}$ and $P_{poly}$ must be determined; industry experts suggest $P_{fin} = 0.75 \times P_{M1}$. Second, the β ratio (i.e., the ratio of fin counts between the PU and PD transistors in FinFET SRAM bitcell) is critical for read stability [9, 16], and affects fin counts and layout. For example, using a β ratio of 2 along with the pitch ratios in Table 1 would set the width of routing regions of bitlines to $2\times 0.75P_{fin}$, PD NMOS to $2 \times P_{fin}$, P/N channel separation to $2 \times 1.5P_{fin}$, and PU PMOS to $1 \times P_{fin}$. The A-factor of the FinFET 6T SRAM would then be calculated as $67F^2$. (Note that the area overhead can be less in 8T FinFET bitcells compared to 6T FinFET bitcells, since additional read transistors in 8T bitcells provide read margin protection. By assuming $\beta = 1.0$ and the layout in Figure 6, the A-factor of 8T FinFET SRAM would be calculated as $72F^2$.)

# 5. LOW-POWER DESIGN

In response to power and energy being identified as *the* grand challenge for the semiconductor roadmap, the Design TWG in 2011 added a Low-Power Design technology roadmap to the Design Chapter. The low-power design roadmap contains a mix of future solutions spanning electrical, functional and software realms [13]. Projected low-power design innovations include (i) frequency islands and near-threshold computing at the circuit level; (ii) heterogeneous parallel processing, many core software development tools, and hardware/software co-partitioning at the architecture level; and (iii) power-aware software and software virtual prototyping at the software level. Figure 8 shows that with low-power innovations the SOC-CP driver dissipates 3.5W (with 48.8M logic gates) in 2011. Low-power design innovations will help limit the power to 8.22W when the number of logic gates grows by more than 40x to 1995.5M in 2026.

[7] The A-factor calculation for FinFET NAND2 may be based on the following assumptions: (i) the heights of P/G rails, which scale poorly in recent nodes due to current delivery and electromigration reliability reasons, are assumed to be $1.5P_{M2}$; (ii) pullup fin count is the same as pulldown fin count; and (iii) cell width of $3P_{poly}$ is still valid for the FinFET-based design. Based on these assumptions and the pitch conversions given in the second column of Table 1, width of the cell is $3P_{Poly}$, and height of the cell is $1.5P_{M2} + 3P_{fin} + 2P_{M2} + 3P_{fin} + 1.5P_{M2} = 9.5P_{M2}$, if $P_{fin} = 0.75P_{M1}$. These calculations suggest that the track number should be more than eight. Recently, GlobalFoundries and ARM have implemented a 14nm FinFET library with 9-track cells [20]. In light of this, the preceding discussion has assumed a track height of nine.

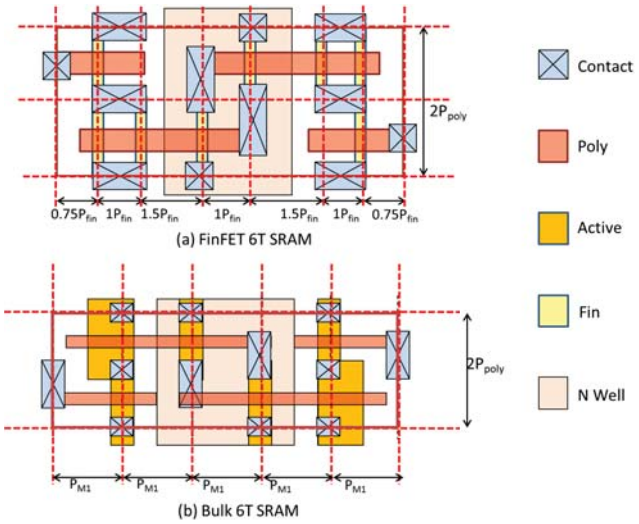


**Figure 5: Layout of 6T SRAMs for (a) FinFET and (b) bulk.**

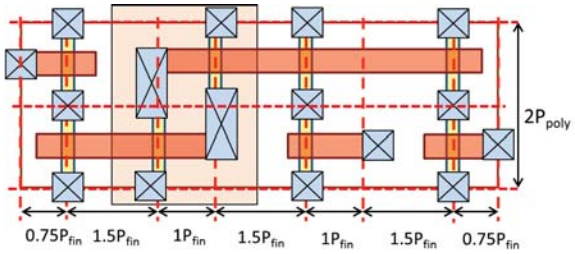


**Figure 6: Layout of 8T FinFET SRAM.**

Figure 8 shows that even if future low-power innovations are developed and deployed according to the low-power design roadmap, power of mobile SOC-CP designs will keep increasing. This is unacceptable in the mobile context; indeed, the SOC-CP driver has a flat power consumption requirement of ∼2W through the end of the roadmap. This is not a new story: Figure 7 from the 2001 Design Chapter predicts that percentage of logic that can be turned on reduces steadily to 2%-6% around 2012, i.e., what researchers have recently termed "dark silicon" [8], [17]. The inability to manage power limits the amount of (switched) logic content in an SOC, which in turn limits product value.

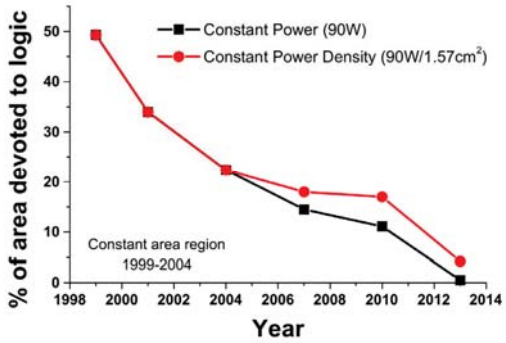


**Figure 7: Dark Silicon projection. [Source: 2001 Design Chapter.]**

In 2012, new additions to the low-power design roadmap include (i) approximate computing (variable-accuracy computing, e.g., flexibly from 64b to 16b); (ii) 4D computing (reconfiguration of circuits on the fly); and (iii) adaptivity (recapturing overdesign due to wearout and variation margins, etc.). To manage power to extreme limits, future low-power innovations must also improve the accuracy of power modeling and estimation. Chips are becoming heterogeneous systems (complex entities with multi-processor software environments) with unpredictable behavior and performance (more of a chip is turned off at any given moment, i.e., dark silicon). In this context, accurate estimation of chip power becomes very difficult.

## 6. DFM, VARIABILITY, RESILIENCE

Increasing process variability, mask cost, data size and lithography hardware limitations pose significant design challenges across different abstraction levels. The ITRS Design Chapter first introduced the design for manufacturing (DFM) section in 2005 to discuss DFM requirements and the corresponding solutions. DFM requirements can be broadly classified as (1) *fundamental economic limitations*, and (2) *variability and lithography limitations*. Requirements due to economic limitations focus on mask cost, which is a key limiter for SOC innovations coming from small companies and emerging-market entities. Requirements due to variability and lithography limitations include quantified bounds on the variability of supply voltage, threshold voltage, critical dimension, circuit performance and circuit power consumption.

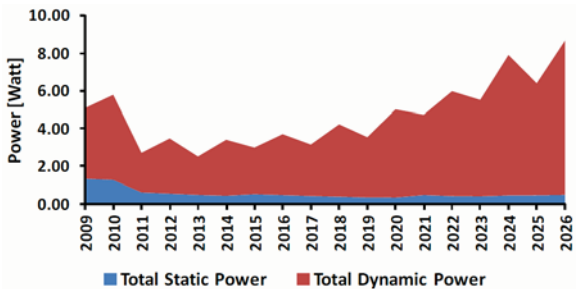


**Figure 8: Impact of low-power design innovations on SOC-CP power consumption. [Source: 2011 System Drivers Chapter.]**

Since variability can cause circuits to exhibit faulty behavior, the DFM section in the 2009 Design Chapter adds projections for circuit-level impacts of variability, focusing on three canonical CMOS logic circuits which are the key components of a digital CMOS design, i.e., (i) SRAM bitcell for storage[8]; (ii) latch for circuit synchronization[9]; and (iii) inverter for logic functions. Failure probabilities for the three canonical circuits in future high-performance technology nodes are obtained by simulating their behavior under the influence of manufacturing process variability. The simulations use *Predictive Technology Model* (PTM) [23] with variability estimates down to 12nm node.

Revised DFM discussion in the 2011 ITRS observes that SRAM failure rate has already become a significant problem in the current technology node. Furthermore, although the latch has a lower failure rate compared to the SRAM, this circuit, too, is predicted to be problematic by the 20nm foundry node. The 2011 analysis also shows that enlarging circuits (i.e., reverse scaling) can be moderately effective in controlling the impact of variability. Other analyses show that failure rate can be reduced by more than an order of magnitude when supply voltage is increased from 90% to 120% of its nominal value, i.e., there is a clear engineering tradeoff between power and robustness.

Over the eight-year history of the Design Chapter's DFM section, potential DFM solutions have been divided into three categories, (i) solutions that address fundamental economic limitations; (ii) solutions that address the impact of variability; and (iii) solutions that address the impact of lithography limitations. Among these, early solutions that directly handle variability (e.g., in timing analysis) have emerged as predicted. The embedding of statistical methods throughout the design flow has been slower than initially forecast, but is still viewed as inevitable. DFM techniques that directly model and simulate lithographic non-idealities are becoming more popular, but will take longer to become qualified in production flows as a consequence of their tighter link to manufacturing models.

## 7. CONCLUDING THOUGHTS

The Design Chapter in the ITRS has for well over a decade defined technology requirements and design challenges for the EDA industry and the VLSI CAD research community. Design technology roadmaps for DFM, low-power design, 3D/TSV integration, More Than Moore, etc. are continually added to maintain relevance of the roadmap. Recent Design Cost and Low-Power Design models highlight the challenges of design productivity, software design cost, and power management in future SOC and MPU designs. At the same time, the System Drivers Chapter has provided models for key market drivers as well as basic chip parameters (layout density,

[8] An SRAM bitcell is considered to be faulty when the SRAM is unable to store the correct logic value during a write operation or when the it fails to preserve the stored logic value during a read operation.

[9] A latch or an inverter is considered to be faulty when its signal delay (e.g., clock-to-output delay for latch) is 10 times the nominal value.

clock frequency, power dissipation, etc.) that bind the ITRS together via the Overall Roadmap Technology Characteristics. The MPU driver model has evolved frequency and power attributes in response to disappearing microarchitectural knobs, emergence of power limits, and challenges of device leakage; further changes (adding uncore elements, evolution of MPU-PCC for micro-server, updated die area modeling) are likely in the near future. The past decade has also seen increased reliance on "design-based equivalent scaling" (e.g., methods for activity factor reduction without compromising throughput or performance) to continue the semiconductor value proposition, and rapidly growing involvement in cross-TWG issues ranging from variability limits to device requirements.

The future of design technology roadmapping, and of the Design TWG's work in the ITRS, will be affected by a variety of technical, business and cultural factors.

- Past foundations of the ITRS seem increasingly shaky. For example, A-factors may no longer be constant across multiple technology nodes. Mx and poly pitches (i.e., horizontal vs. vertical densities) may scale at different rates. The fundamental assumption of 2× density scaling per node may be already long past; whether the industry can flourish with, e.g., 1.4× density scaling per node is an open question.

- Tremendous uncertainty with respect to patterning technology (e.g., timing of EUV, directed self-assembly), cost models (e.g., triple- and quadruple-patterning), device and interconnect structures and properties (tunnel FETs, resistive RAMs, drive vs. leakage currents), and high-value applications all present challenges to the roadmapping of design technology requirements.

- Fewer resources are available for ITRS activity even as the scope of the roadmap widens (MEMS, More Than Moore, new storage and switch elements, 3D integration) and the difficulty of the roadmapping task increases. Greater automation is needed to check consistency and impacts of proposed roadmap changes, a la the "Living ITRS" efforts of a decade ago [4].

- An oligopolistic EDA industry, along with continued consolidation and disaggregation in the semiconductor industry, as well as unwillingness to share competitive (as opposed to pre-competitive) data,[10] means that leading companies more frequently "opt out" of roadmap participation. There is a risk of a "vicious cycle" of decreased roadmap participation and decreased roadmap value.

- Communication across supplier industries, across the design-manufacturing interface, and across academia-industry boundaries is increasingly needed to optimize technology investments and maximize the returns from the roadmapping process. As the industry faces an explosion of post-CMOS, post-optical technology options, it seems appropriate to at least revisit the concept of "shared red bricks".

Against this backdrop, there is some good news: Members of the design, EDA and research communities are willing to find common cause in the design technology roadmap. At the 2009 and 2010 EDA Roadmap Workshops [19], representatives from leading EDA companies, semiconductor companies, and research consortia commenced a dialogue to analyze needs and status of EDA roadmapping.[11] Other discussions sought new mechanisms by which more of the community could contribute to the design technology roadmap. And the really good news for EDA and VLSI CAD: If anything remains essential to the future of Moore's Law scaling, it will be design technology, and design-based equivalent scaling.

[10] It is suboptimal for students at UCSD to "predict" designs and cell libraries that industry has already developed, or for students at Purdue to develop ab initio models for device structures that again have already been developed. Yet, these are the mechanisms by which core material and data is generated in the ITRS today.

[11] The 2009 workshop addressed such questions as "What would make an EDA roadmap more useful?", "Which EDA areas lack most in roadmap efforts?", and "Which EDA areas are behind what the roadmaps say?" The 2010 workshop then identified gaps in the EDA roadmap (system-level executable specification, design-space exploration and pathfinding, EDA scaling requirements in light of evolving computing platforms, power-driven design, and design for resilience), reached agreement on the nature of EDA, and identified challenges in filling in the EDA roadmap gaps (incremental design flows, new design for cost methodologies, and an expanded scope of EDA moving to system-level design).

## Acknowledgments

Dr. Juan-Antonio Carballo has co-chaired the U.S. and International Design TWGs with me for the past decade, and has been particularly influential in the conception of the System Drivers Chapter as well as iNEMI and More Than Moore interactions. Dr. Kwangok Jeong developed and maintained the MPU, power, frequency and A-factor models during the critical years of 2007-2011, which saw many Design-PIDS interactions regarding roadmap for device power vs. performance. This paper would not exist without the help of UCSD Ph.D. students Tuck-Boon Chan, Siddhartha Nath, Wei-Ting Jonas Chan, and Ilgweon Kang. Many participants in the ITRS Design and System Drivers efforts, and in the overall ITRS effort, have contributed valuable insights and perspectives over the years. I also thank Dr. Sani Nassif (who has for years driven the DFM section of the Design Chapter) for organizing the special session which led to the writing of this paper.

## 8. REFERENCES

[1] C. Auth, C. Allen, A. Blattner, D. Bergstrom et al., "A 22nm High Performance and Low-Power CMOS Technology Featuring Fully-Depleted Tri-Gate Transistors, Self-Aligned Contacts and High Density MIM Capacitors", *Proc. Symposium on VLSI Technology*, 2012, pp. 131-132.

[2] V. S. Basker, T. Standaert, H. Kawasaki, C.-C. Yeh et al., "A 0.063 $\mu m^2$ FinFET SRAM Cell Demonstration with Conventional Lithography Using a Novel Integration Scheme with Aggressively Scaled Fin and Gate Pitch", *Proc. IEDM*, 2010, pp. 19-20.

[3] T.-B. Chan and A. B. Kahng, "Tunable Sensors for Process-Aware Voltage Scaling", *Proc. ICCAD*, 2012, pp. 7-14.

[4] A. E. Caldwell, Y. Cao, A. B. Kahng, F. Koushanfar, H. Lu, I. L. Markov, M. R. Oliver, D. Stroobandt and D. Sylvester, "GTX: The MARCO GSRC Technology Exploration System", *Proc. DAC*, 2000, pp. 693-698.

[5] J.-A. Carballo and A. B. Kahng, "ITRS Chapters: Design and System Drivers", *Future Fab International* (36) (2011), pp. 45-48.

[6] J.-A. Carballo and A. B. Kahng, "ITRS Chapters: Design and System Drivers", *Future Fab International* (40) (2012), pp. 54-59.

[7] J.-A. Carballo and A. B. Kahng, "ITRS Chapters: Design and System Drivers", *Future Fab International* (44) (2013), pp. 52-56.

[8] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam and D. Burger, "Dark Silicon and The End of Multicore Scaling", *Proc. ISCA*, 2011, pp. 365-376.

[9] Z. Guo, S. Balasubramanian, R. Zlatanovici, T.-J. King and B. Nikolic, "FinFET-Based SRAM Design", *Proc. ISLPED*, 2005, pp. 2-7.

[10] K. Jeong and A. B. Kahng, "A Power-Constrained MPU Roadmap for the International Technology Roadmap for Semiconductors (ITRS)", *Proc. ISOCC*, 2009, pp. 49-52.

[11] A. B. Kahng, "The Road Ahead: Shared Red Bricks", *IEEE Design and Test of Computers*, 19(2) (2002), pp. 70-71.

[12] A. B. Kahng, "The Road Ahead: The cost of design", *IEEE Design and Test*, 19(4) (2002), pp. 136-137.

[13] A. B. Kahng, "The Road Ahead: Roadmapping Power", *IEEE Design and Test of Computers*, 28(5) (2011), pp. 104-106.

[14] A. B. Kahng and G. Smith, "A New Design Cost Model for the 2001 ITRS", *Proc. ISQED*, 2002, pp. 190-193.

[15] H. Kawasaki, M. Khater, M. Guillorn, N. Fuller et al., "Demonstration of Highly Scaled FinFET SRAM Cells with High-k Metal Gate and Investigation of Characteristic Variability for the 32 nm Node and Beyond", *Proc. IEDM*, 2008, pp. 1-4.

[16] D. Lekshmanan, A. Bansal and K. Roy, "FinFET SRAM: Optimizing Silicon Fin Thickness and Fin Ratio to Improve Stability at Iso Area", *Proc. CICC*, 2007, pp. 623-626.

[17] G. Venkatesh, J. Sampson, N. Goulding, S. Garcia, V. Bryksin, J. Lugo-Martinez, S. Swanson and M. B. Taylor, "Conservation Cores: Reducing the Energy of Mature Computations", *Proc. ASPLOS*, 2010, pp. 205-218.

[18] CPUDB. *http://cpudb.stanford.edu/*

[19] EDA Roadmap Workshop at DAC 2010. *http://vlsicad.ucsd.edu/EDARoadmapWorkshop/*

[20] "GlobalFoundries Details 14nm-XM FinFET Technology Performance, Power and Area Efficiency with a Dual-Core Cortex-A9 Processor Implementation". *http://www.globalfoundries.com/newsroom/2013/20130205-ARM.aspx*

[21] iNEMI. *http://www.inemi.org*

[22] ITRS Edition Reports. *http://public.itrs.net/reports.html*

[23] Predictive Technology Model. *http://ptm.asu.edu*

[24] Design-Based "Equivalent Scaling" to the Rescue of Moore's Law. *http://vlsicad.ucsd.edu/Presentations/talk/UCI-Colloquium-121031-v7-distributed.pdf*