The ITRS MPU and SOC System Drivers: Calibration and Implications for Design-Based Equivalent Scaling in the Roadmap

Wei-Ting Jonas Chan[†], Andrew B. Kahng[†‡], Siddhartha Nath[‡] and Ichiro Yamamoto[*]

[†]ECE and [‡]CSE Departments, UC San Diego, La Jolla, CA 92093, USA

[*]Rohm Co., Ltd., Shinyokohama, Japan

{wechan, abk, sinath}@ucsd.edu, Ichiro.Yamamoto@dsn.rohm.co.jp

***Abstract*—The *system driver* models for *microprocessor* (MPU) and *system-on-chip* (SOC) in the *International Technology Roadmap for Semiconductors* [21] (ITRS) determine the roadmap of underlying technology requirements across devices, patterning, interconnect, test, design and other semiconductor supplier industries. In this paper, we describe several fundamental changes in the ITRS MPU and SOC system driver models as of the recently-released 2013 edition of the roadmap. We first present new *A-factor* (i.e., layout density) models for the logic and memory components of the MPU and SOC drivers; these updated density models comprehend the industry's shift to FinFET devices below the foundry 20nm node. We also describe updated architectural, total chip area, and total chip power models for the MPU and SOC drivers. Notably, we model the growing *uncore* portion of MPU products, and the growing presence of *graphic processing units* (GPUs) and other *peripheral cores* (PEs) in SOC architectures. The updated SOC architectural model enables more realistic scenario-based power modeling for the SOC driver.**

The 2013 ITRS update of system driver models embodies extensive calibration with foundry data as well as product structural analysis reports from a leading analysis firm (Chipworks). The model calibration reveals that the industry has contended with a "scaling gap" since 2008, whereby traditional Moore's-Law density scaling of 2× per node has failed due to patterning limitations on layout design, as well as manufacturability and performability challenges of *Metal-1 half-pitch* (M1HP) scaling. Growing design margins due to reliability, yield, variability, etc. have also contributed to the slowdown of density scaling. We describe how this scaling gap can potentially be compensated if the semiconductor industry urgently pursues *design-based equivalent scaling* (DES), which substantially changes the area and power model trajectories of MPUs and SOCs in the ITRS System Drivers Chapter. Finally, we note that as a consequence of the updated A-factor, area and power models in the 2013 ITRS, the industry now faces a 20% more daunting power management challenge than had been predicted in the 2011 roadmap.

I. Introduction

The *International Technology Roadmap for Semiconductors* [21] (ITRS) roadmaps technology requirements for devices, patterning, interconnect, test, design and other semiconductor supplier industries. The requirements are determined using the *system driver* models for *microprocessor* (MPU) and system-on-chip (SOC). The key system drivers have been updated based on different marketing requirements in the past decade. In 2001 [22], there was only one SOC driver without any other derived variations. In 2005 [23], the SOC driver was split into SOC high-performance and SOC power-efficient. In 2007 [24], the SOC driver was categorized into *Networking* (SOC-NW), *Consumer Portable* (SOC-CP), and *Consumer Stationary* (SOC-CS) to address the specialization trends. In Table I, we show that some drivers disappear from the 2013 roadmap because the application contexts have changed. Due to the fast growth of the mobile market, the *MPU Power Connectivity Cost* (MPU-PCC) driver is removed as its application context is subsumed by the SOC-CP driver. Since game consoles and PCs use similar processors, we remove the SOC-CS driver.

TABLE I. Summary of changes between 2011 and 2013 ITRS MPU and SOC drivers.

Year	2011	2013
MPU High Performance (MPU-HP)	Exists	Exists
MPU Cost Performance (MPU-CP)	Exists	Exists
MPU Power Connectivity Cost (MPU-PCC)	Exists	Dropped
SOC Consumer Portable (SOC-CP)	Exists	Exists
SOC Consumer Stationary (SOC-CS)	Exists	Dropped
SOC Networking (SOC-NW)	Exists	Exists

The heartbeat of the roadmap for layout density is the minimum feature size, that is, the *Metal-1 half-pitch* (M1HP), also referred to as F. $(\text{M1HP})^2$, in conjunction with *A-factor* values, yields models for the layout density of logic and memory blocks in the MPU and SOC drivers. The derivations of area and power models based on A-factor are described by Jeong and Kahng in [7]. The 2013 *System Drivers Chapter* uses the model in [7], calibrates the model with state-of-the-art MPU and SOC products, and then projects these models for the next 15 years.

Formally, A-factor is a multiplier of the $(\text{M1HP})^2$ unit of area. Below the foundry 20nm node, the industry shifts to FinFET-based devices for memory (SRAM) and logic cells. In this paper, we present new A-factor models that comprehend the industry's shift to FinFET devices, and describe several fundamental changes in the MPU and SOC system driver models in the 2013 ITRS roadmap. We describe updated architectural templates, as well as total chip area and total chip power models, for the MPU and SOC drivers. For MPU products, we model the growing *uncore* components, such as *graphic processing units* (GPUs), on-chip networking, multiple bus interfaces, etc. For SOC products, we model the growing number of GPU cores, *radio-frequency and analog/mixed signal circuits* (RF/AMS), IO and other *processing engines* (PEs). The updated SOC architectural model enables more realistic usage scenario-based power modeling for the SOC driver, and we update the SOC power models on a per-scenario basis.

The system driver models are extensively calibrated with foundry data as well as structural analysis reports from a leading analysis firm (Chipworks) [16] [17]. The calibration process reveals that a "scaling gap" has existed since 2008, that is, traditional Moore's-Law density scaling of 2× per node has slowed down due to patterning limitations on design, as well as manufacturability and performability challenges of the M1HP scaling. The industry has been contending with this gap with architecture- to device-level enhancements. Growing design margins due to reliability,

978-1-4799-6492-5/14/$31.00

yield, variability, etc. have also contributed to the slowdown of density scaling. Furthermore, the ITRS M1HP scaling trajectory has been reset in 2013, and slowed down by one technology node from 2013 to 2019; this induces a further slowdown of density scaling to prevent a die area explosion. In this work, we describe a quantified requirement for *design-based equivalent scaling* (DES), which can potentially compensate the scaling gap. However, DES, substantially changes the modeled area and power of the MPU and SOC drivers.[1] The updated A-factor, area and power models in the 2013 ITRS reveal a 20% more daunting power management challenge than had been predicted in the 2011 roadmap. The industry needs innovation in low-power design techniques to overcome this challenge. The updates to the System Drivers models between the 2011 and 2013 ITRS editions are summarized in Table II.

TABLE II. SUMMARY OF CHANGES BETWEEN 2011 AND 2013 ITRS MPU AND SOC ROADMAPS.

	Year	**2011**	**2013**
	Calibration source	**Public domain data**	**Public domain data + structure analysis reports**
A-factor	Logic	$60F^2$ (bulk)	$60F^2$ (FinFET)
	SRAM	$175F^2$ (bulk)	$155F^2$ (FinFET)
	Uncore	N/A	Added
MPU Driver	Overhead	Exists	Updated
	DES	N/A	Added
	Uncore	N/A	Added
SOC Driver	Overhead	Exists	Updated
	DES	N/A	Added
	Power	Single scenario	Multiple scenarios

The contributions of this paper are summarized as follows.

(1) We present new A-factor models for the SRAM bitcell and NAND2 logic cell for both FinFET and bulk devices. Our models are developed with inputs from experts at various consortia, and calibrated with foundry data at 20*nm* as well as1 product structural analysis reports from a leading chip analysis firm [17].

(2) We derive area and power models for MPU and SOC drivers using the new A-factor models alone with component block dimensions and floorplans of latest products from product structural analysis reports. Our area models comprehend recent trends in MPU and SOC product die such as growth in the number of GPUs. We describe new overheads that are used in the area model for uncore components in the MPU. Furthermore, we add scenario-based power modeling for the SOC-CP driver.

(3) We study scaling trends of MPU and SOC products and we report a "scaling gap" that has existed since 2008. In addition, the reset and slowdown of M1HP requires the semiconductor industry to adopt "design-based equivalent scaling". DES substantially changes the area and power models.

The remainder of this paper is organized as follows. In Section II, we present new A-factor models for FinFET-based SRAM and logic cells. Section III presents new architectural templates for MPU and SOC drivers and Section IV describes our calibration process. We introduce DES, and present MPU and SOC area and overhead models using DES, in Section V. Finally, we present the power models in Section VI and conclude the paper in Section VII.

[1]The ITRS System Drivers Chapter area and power models give "centerline" projections; it is mentioned that the models cannot fit every actual product design.

II. A-FACTOR UPDATES

As noted above, A-factors of SRAM bitcells and standard-cell logic gates are multipliers of $(\text{M1HP})^2$. The MPU and SOC area models rely on A-factors of SRAM and logic cells (A_{SRAM} and A_{logic} in Equations (1) and (2) respectively). As the dominant device architecture is now shifting from bulk to FinFET, in the 2013 ITRS we introduce new A-factor models based on both (i) new layout guidelines from our industrial collaborators and (ii) analysis reports from Chipworks. We describe the main differences of the layout models and calibration flow in this section. To compare the 2011 and 2013 layout models, we show the layouts of the 2011 model in Figure 1.

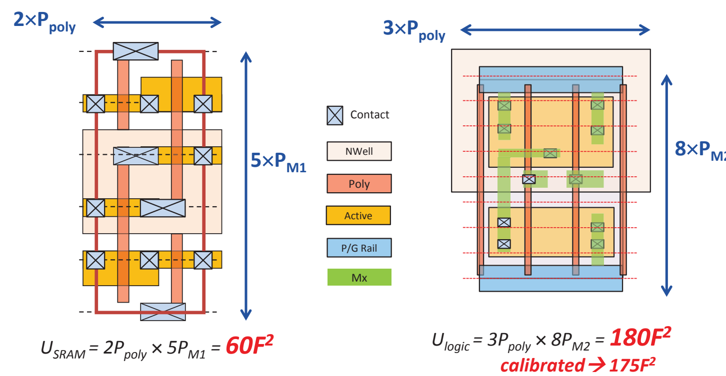

Fig. 1. SRAM bitcell and NAND2 layout for bulk (2011 ITRS).

To reflect the latest progress in circuit designs, we consider the following in the new layouts:

(1) FinFET-based logic cells and SRAM cells,
(2) discreteness of FinFET sizes in cell design,
(3) FinFET pitch in the SRAM cell layout, and
(4) impacts of changing pitches of Mx layers (e.g., the Metal-2 pitch (P_{M2}) changes from 1.25× to 1× of the Metal-1 pitch (P_{M1})).

Table III shows the pitch conversions used to derive 2013 ITRS A-factors for SRAM and logic cells from 2011 values.

TABLE III. COMPARISON OF PITCH CONVERSION USED IN 2011 AND 2013 A-FACTOR MODELS.

Layer	**Normalization to P_{M1}**	
	2011	**2013**
F	0.50	0.50
M1 (P_{M1})	1.00	1.00
M2 (P_{M2})	1.25	1.00
Polysilicon (P_{poly})	1.50	1.50
Fin Pitch (P_{fin})	–	0.75
P/G Track Width	–	1.50

One of the key features in the 2013 model is the pitch of fins P_{fin}. The main limitation of P_{fin} is the lithography used to produce the spacer in (*sidewall image transfer* (SIT) or *self-aligned double patterning* (SADP) based patterning) during the *front-end-of-line* (FEOL) process [4] [15]. In practice, we set P_{fin} to be $0.75\times$ of the Metal-1 pitch P_{M1}. We derive A-factors of SRAM 6T bitcells and logic NAND2 standard cells based on inputs from roadmapping consortia and integrated device manufacturers (IDMs). For example, in the FinFET regime, we use a discrete device sizing ratio in the 6T SRAM bitcell of 1:2:1 for the *pull-up* (PU), *pull-down* (PD) and *pass-gate* (PG) transistors. While high-density bitcells could use 1:1:1 ratioing at the cost of design margin, we apply the 1:2:1 ratioing as a “middle ground” that takes into account the increased reliability challenge in advanced nodes. Figure 2 shows our updated layout for the 6T SRAM bitcell, which we refer to as FinFET SRAM. Interconnect widths are $2\times 0.75\times P_{fin}$ for bitlines on the two sides. For the active devices, we use the following spacings:

- $1\times P_{fin}$ for each of the pull-down N-channel transistor,
- $1\times P_{fin}$ for each of P/N channel isolation, and
- $1\times P_{fin}$ for P-channel transistors.

The height of the FinFET SRAM bitcell is the same as that of the bulk SRAM bitcell, that is, $2\times P_{poly}$, where P_{poly} is the pitch of polysilicide. We illustrate the bulk and FinFET layouts of SRAM bitcells in Figures 1 and 2 respectively. With $P_{fin} = 0.75\times P_{M1}$ and $P_{poly} = 1.5\times P_{M1}$, we derive the unit area (U_{SRAM}) of a FinFET SRAM bitcell as shown in Equation (1). After calibration with structural analysis reports of state-of-the-art SRAM products [17], the A-factor of FinFET SRAM is adjusted to 60, that is, the same as the A-factor of bulk SRAM in the 2011 ITRS roadmap. Going forward, discreteness of fin counts will bring challenges in SRAM design in that required read/write margins will be more difficult to attain with fewer device size choices. Circuit techniques such as read- and write- boosting will be required to compensate decreased performance margins of FinFET SRAM.

$$\begin{aligned} U_{SRAM} &= 2p_{poly}\times 6.5p_{fin} = 14.625(P_{M1})^2 \\ &= 58.5F^2 = A_{SRAM}F^2 \end{aligned} \tag{1}$$

$$\begin{aligned} U_{logic} &= 3P_{poly}\times 9P_{M2} = 40.5(P_{M1})^2 \\ &= 162F^2 = A_{logic}F^2 \end{aligned} \tag{2}$$

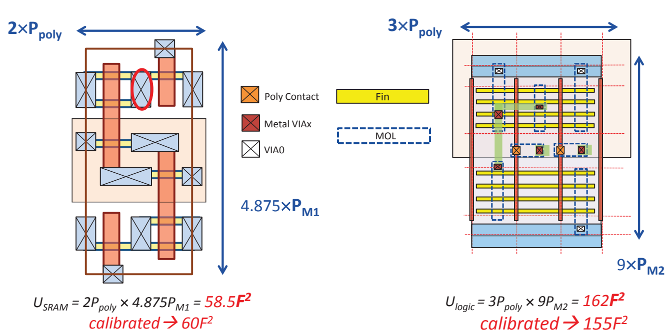

Fig. 2. SRAM bitcell and NAND2 layout for FinFET (2013 ITRS).

The A-factor of a FinFET-based logic cell using the layout of a NAND2 cell is shown on the right side of Figure 2, and is calculated based on the following assumptions.

- The height of power and ground (P/G) tracks is $1.5\times P_{M2}$.
- The pull-up fin count is same as the pull-down fin count, and is four according our collaborative researchers.
- Each cell needs to be at least $3\times P_{poly}$ pitches wide to layout four transistors. So, the width of the NAND2 cell is $3\times P_{poly} = 9F$, after applying the conversions in Table III.
- The height of the NAND2 cell is set to 9 tracks to accommodate eight fins, P/G tracks, and routing space to route and access the *middle of line* (MOL) layers. Thus, the height of the NAND2 cell is $9\times P_{M2}$.

The unit area (U_{logic}) of a logic standard cell (NAND2) is derived using Equation (2). We calibrate the A-factor with Chipworks [17] structural analysis reports of NAND2 cells in state-of-the-art products, as well as feedback from our collaborators in design houses. After calibration, the A-factor is adjusted to be $155F^2$. The change of A-factors from the 2011 ITRS roadmap is shown in Table IV.

TABLE IV. SUMMARY OF A-FACTOR FOR BOTH NAND2 AND SRAM BITCELL IN 2011 AND 2013 ITRS MODELS.

	SRAM		NAND2	
Year	2011	2013	2011	2013
BULK	$60F^2$	$60F^2$	$175F^2$	$175F^2$
FinFET	N/A	$60F^2$	N/A	$155F^2$

III. ARCHITECTURAL MODELS OF MPU AND SOC

The SOC Consumer Portable (SOC-CP) Driver 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 processing power of SOC-CP grows much slower than the previous SOC-CP roadmap ($1000\times$ improvement was predicted from 2011 to 2021 in the 2011 ITRS roadmap). This is because the number of processing engines scaling has slowed down, as a result of increased scaling of Graphic Processing Engine (GPU) cores. At the same time, total SOC-CP power is also predicted to increase because of the strong demand for compute-heavy functionality and ever-improving user experience. Advanced low-power design technologies will be crucial for this segment.

Figure 3 shows an architecture template for the SOC-CP driver. The SOC embodies a highly parallel architecture consisting of a number of main processors, a number of GPUs, peripherals, and memories. Due to rising demand for high-definition graphics and video playback, the number of GPUs is expected to rapidly increase. A processing engine (PE) is a processor customized for a specific function, and has a large-scale, highly complex structure. The architecture template of SOC-CP enables both high processing performance and low power consumption by virtue of parallel processing and hardware realization of specific functions. The computation-intensive graphics and media functions are executed by GPUs, and the remaining sets of functions are implemented by

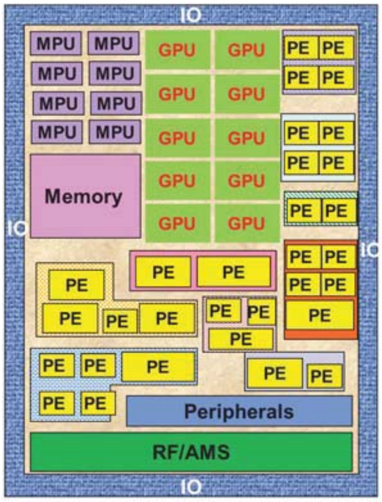

Fig. 3. SOC architecture template.

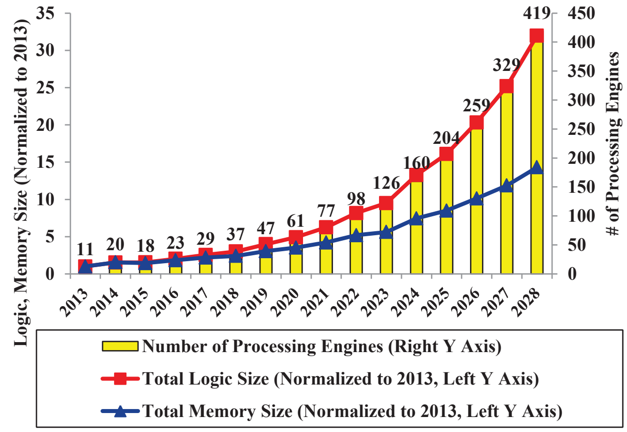

Fig. 4. Design complexity trends of the SOC-CP driver in terms of number of processing engines, logic size, and memory size.

corresponding PEs. Based on this architecture template, Figure 4 shows the design complexity trends for the SOC-CP driver. We make the following assumptions in the SOC model.

(1) There are four processing cores with identical complexity, and the number of these processors will continually grow in the future.

(2) GPU cores are modeled separately from existing peripherals and PEs. This is due to the increasing demand of high-definition display interfaces and 3D graphic processing requirements by gaming, as well as graphical applications on smartphone and tablets. Area of GPU cores increases the die area of SOC-CP and consumes a significant part of the chip power.

(3) The peripheral contains I/O circuits such as multi-channel high-speed memory interfaces, USB, HDMI, etc., as well as integrated RF/AMS circuits for wireless communications.

(4) The area of PEs decreases because several functions are implemented by the programmable components, such as CPUs and GPUs.

(5) Our survey of die areas of recent mobile products suggests that the die area of SOC-CP must be increased to $140mm^2$ from $100mm^2$. The die area is expected to remain constant through the roadmap in light of yield and form-factor considerations.

A major change in the MPU model is the addition of "uncore" overhead ($O_{uncore-logic}$) to model GPU, bus interfaces, display ports and networking components. These components occupy around 30% of the MPU die based on publicly available MPU floorplan data as well as structural analysis reports from Chipworks [17]. The uncore consists mainly of logic cells, which consume power. Due to lack of detailed information, we could not model the transistor count of uncore, but instead use a logic area overhead in our MPU model. Table V shows the initial value used in 2013 and the growth rate of the uncore overhead.

The new SOC-CP architectural model points out that the required processing performance of SOC-CP presents a severe challenge to designers, as seen in Figure 5. The growth of required processing performance, which is defined as the product of core frequency and number of cores (both main processors and GPUs) (Equation (3)), remains similar to the previous projection in the 2011 ITRS (around 10× every three years). However, the growth rates of core frequency and number of cores do not meet the requirement as shown in Figure 5. Potential solutions for this gap range from better hardware/software partitioning to high-level synthesis for interface components.

$$\{Processing\ performance\} = \{\#main\ processors\} \times \{main\ processor\ frequency\} + \{\#GPUs\} \times \{GPU\ frequency\} \quad (3)$$

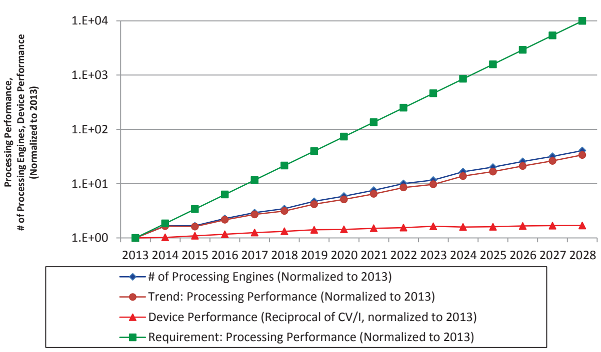

Fig. 5. Processing performance requirement of the SOC-CP driver.

IV. Model Calibration and the Density Scaling Gap

The 2013 ITRS driver model revisions have seen a significantly improved process of model calibration. Previous derivations of A-factor models and die area models (e.g., [7]) have been based on chip size and capacity information from published data sheets, white papers and die photos for state-of-the-art MPU and SOC products. However, these data lack details of implementation, such as physical locations and sizes of function blocks, actual device dimensions, interconnect pitches, etc. During the 2013 ITRS revision cycle, access was granted by a leading chip analysis firm, Chipworks [17], to a large collection of structural analysis reports for recent MPU and SOC products. Analyses of actual pitches, patterning styles, and basic cell layouts in recent technology nodes have greatly enhanced the 2013 ITRS calibration of A-factor models for both logic and SRAM. The calibrated A-factor models are then used in the MPU and SOC chip area models.

From Chipworks reports [17], we obtain the SRAM and NAND2 cell layouts, and use cell length and width from these reports to calibrate A-factors for SRAM and logic cells. We study reports over a wide range of technology nodes from foundry 65nm to 20nm. The results are shown in Figure 6. The calibration process reveals that the A-factor of SRAM bitcell remains at $60F^2$ since 65nm. The A-factor of NAND2 cells gradually decreases from 65nm to a value of $155F^2$ at the foundry 20nm node.

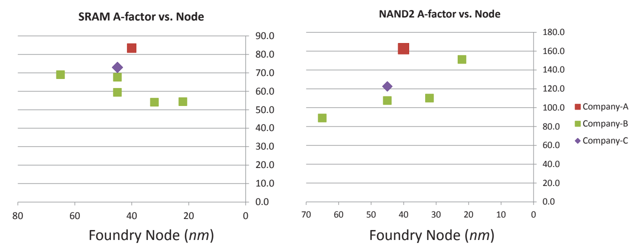

Fig. 6. SRAM and NAND2 A-factors derived from Chipworks [17] analysis reports for different nodes.

V. Design-Based Equivalent Scaling and Updated Area, Overhead Models

Geometric scaling of transistors, popularly known as "Moore's Law", has been a long-term thrust for the semiconductor industry. Historically, Moore's Law had been continuously scaling at 2× per node by huge investments and research efforts from the industry to overcome plenty of barriers. However, the 2× per node scaling by Moore's Law is slowing down in the recent technology nodes as we mentioned above due to the scaling gap [19] [20] [26]. From our study of recent semiconductor products, only 1.6× improvement is realized by designs as shown in Figures 7 and 8. More specifically, the scaling of M1HP is expected to slow down by one technology node from 2013 to 2019 due to process and other challenges of copper BEOL interconnects, and has been reset to 40nm in 2013 (compared to 27nm in 2013 from the 2011 ITRS roadmap) as shown in Figure 9. We observe that the scaling of Moore's Law must be supported by both geometric scaling and design techniques to maintain the equivalent 2× per node scaling trend. We refer these design techniques, such as design for variability, low-power design, heterogeneous multicore architectures, etc., as "design-based equivalent scaling" (DES). DES, added as a new parameter in 2013 ITRS MPU/SOC models to compensate the slowness of physical scaling, is defined as a shrink (i.e., reduction) factor for chip area due to non-geometric scaling. Figure 10 illustrates die area explosion for the high-performance MPU driver without DES (IS (w/o DES) in the figure) and with DES (IS (w/ 6y-DES) in the figure). The "WAS" line in the figure illustrates the 2011 ITRS die area roadmap.

A. MPU Area and Overhead Models

The ITRS MPU driver reflects high- and cost-performance desktop and server systems that are general-purpose instruction-set architectures. The scaling of usable transistors has slowed down from 2× for each technology generation to 1.6× due to the

TABLE V. Area overheads and DES.

Overhead/DES	Initial Value	Growth Rate
DES	1.00	0.93× per year until 2019; fixed at 0.63× from 2020
O_{SRAM}	1.30	1.30 until 2019; 1.26× per node from 2020
$O_{Afactor-logic}$	1.00	1.12× per node
$O_{Afactor-SRAM}$	1.66	1.15× per node
$O_{uncore-logic}$	1.00	1.12× per node
O_{logic}	1.40	Fixed
$O_{eq-logic}$ (post DES)	1.40	1.26× per node
$O_{integration}$	1.24	Fixed

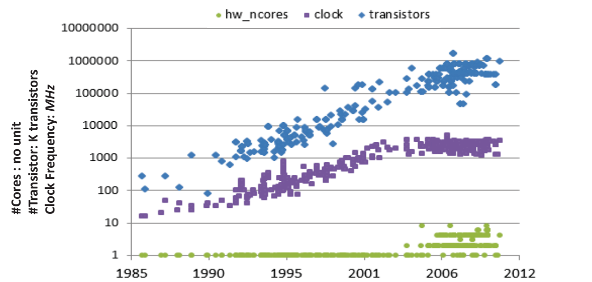

Fig. 7. Number of cores, transistors, and clock frequency scaling trend from CPUDB [18].

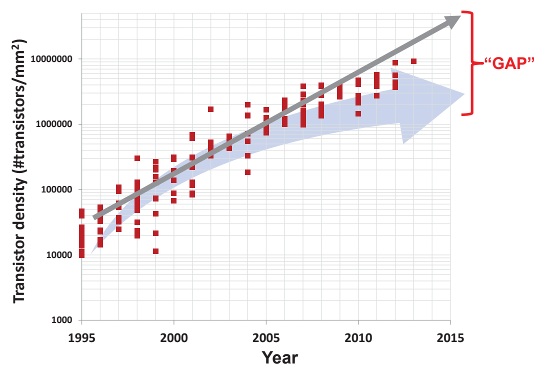

Fig. 8. Scaling gap in transistor density.

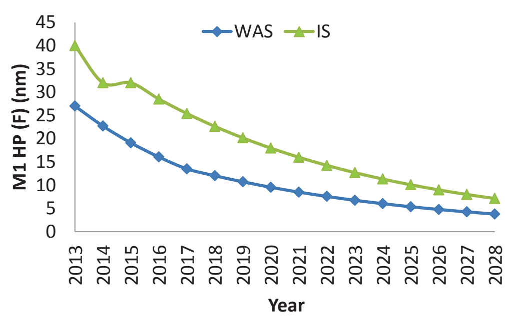

Fig. 9. Slowdown of Metal-1 half-pitch scaling from the 2011 ITRS ("WAS") to the 2013 ITRS ("IS").

difficulty of applying all the available transistors to enable more functions. This implies that the scaling of number of processing cores and the number of transistors slows down from 1.6× per each technology node to only 1.26× per node. This "utilization barrier" [21] introduces new overheads in the MPU SRAM and logic density models, and these overheads grow with each node, as compared to the 2011 ITRS MPU area models in which the overheads are constant throughout the roadmap.

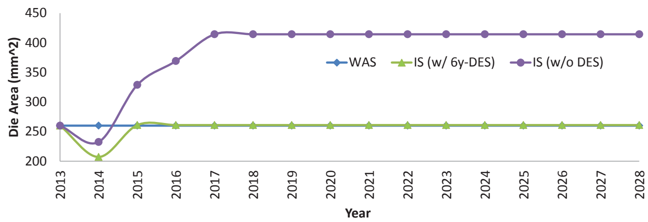

Fig. 10. The 2013 ITRS's proposed recovery of the historical scaling of area density, during the years 2013-2019, with design-based equivalent scaling (DES).

The 2013 ITRS roadmap uses two types of MPU drivers, *high-performance* (HP) to represent server systems and *cost-performance* (CP) to represent desktop systems. Our surveys using public domain data indicate that the MPU-CP products span a much larger portion of the price-performance tradeoff curve in the 2011 ITRS. The products range from low-end, low-cost traditional desktops to laptops used primarily in AC mode (also known as *mobile desktops*) and low-cost blade servers. We use a constant die area over the course of the roadmap, $140mm^2$ for CP and $260mm^2$ for HP. In the 2011 ITRS, die area is broken down into logic, memory, and integration overhead. Integration overhead models (i) the white space for routing channels between blocks; (ii) area loss due to thermal-aware floorplans; and (iii) lower utilization to reduce congestion and enable signoff of designs with smaller design turnaround time. In the 2013 ITRS, we add new area overheads to model logic transistors in the uncore components of the MPU die, such as GPUs, network, bus interfaces, and other peripherals. Guardbands for reliability, variability, etc., increase the area of logic transistors, so we add A-factor overhead ($O_{Afactor-logic}$) to model this increase per node.

We derive the MPU area model in a bottom-up manner using logic and SRAM cells and other chip-level information. The unit logic cell area S_{logic} is shown in Equation (4). The parameters N_{core} and N_{gate} are the number of cores and the number of logic gates per core, respectively; O_{logic} is the logic overhead.

$$S_{logic} = O_{logic} \cdot U_{logic} \cdot N_{core} \cdot N_{gate} \tag{4}$$

To derive the logic transistor density $D_{tr,logic}$, we use the number of transistors in a NAND2 gate ($N_{tr,nand2} = 4$). In the 2013 area model, $O_{Afactor-logic}$ and $O_{uncore-logic}$, mentioned above, are used in Equation (5) to model the scaling gap, and DES to compensate the scaling gap.

$$D_{tr,logic} = \frac{N_{tr,nand2}}{O_{logic} \cdot O_{Afactor-logic} \cdot O_{uncore-logic} \cdot U_{logic}} \cdot \frac{1}{DES} \tag{5}$$

Similarly, the area occupied by SRAM, S_{SRAM}, is calculated using Equation (6). With parameters N_{bits}, the number of bits per core, and O_{SRAM}, the SRAM peripheral overhead.

$$S_{SRAM} = O_{SRAM} \times U_{SRAM} \times N_{core} \times N_{bits} \tag{6}$$

To derive the SRAM transistor density of logic $D_{tr,SRAM}$, we use the number of transistors in a SRAM cell ($N_{tr,bitcell} = 6$) and Equation (7). In the 2013 area model, $O_{Afactor-SRAM}$, mentioned above, is used in Equation (5) to model the scaling gap, and DES to compensate the scaling gap.

$$D_{tr,SRAM} = \frac{N_{tr,bitcell}}{O_{SRAM} \cdot O_{Afactor-SRAM} \cdot U_{SRAM}} \cdot \frac{1}{DES} \tag{7}$$

The total die area S_{die} is calculated using Equation (8), where $O_{integration}$ is the integration overhead.

$$S_{die} = O_{integration} \times (S_{logic} + S_{SRAM}) \tag{8}$$

The above area and density models, along with the supply voltage and capacitance parameters from the ITRS *Process Integration, Devices, and Structures Chapter* (PIDS) [27] and *Interconnect Chapter* (INTC) [25], are used to develop the power models for MPU and SOC. We explain the 2013 changes in the power model in Section VI.

The areas, transistor counts, and transistor density scaling trends of both MPU and SOC drivers are summarized in Table VI.[2] In the 2013 ITRS roadmap, the transistor count of MPU-HP has been reset to 7.14B transistors, and MPU-CP has been reset to 2.54B transistors.[3] The difference in transistor count allows for more aggressive microarchitectural improvements such as prefetching and other prediction mechanisms, trace caching, and the introduction of accelerators such as encryption and graphics/media. The number of logic cores in the MPU model has been reset to eight for MPU-HP and four for MPU-CP, and projected to increase by a factor of 1.26× with each technology node. The "power wall" and the limited exploitation of available parallelism together limit the scaling of cores. Recent trends suggest a factor of 1.26× scaling of logic transistors per core with every node. The MPU memory content has been reset to 12MBytes (12 × 1,048,576 × 9 bits) of SRAM for CP and 58MBytes for HP in 2013. Memory content is also projected to scale at 1.6× with each successive technology node. Based on public domain data and structural analysis reports of leading-edge products from Chipworks [17], we calibrate the overheads. $O_{uncore-logic}$ and $O_{Afactor-logic}$ are 1.0 in 2013 and grow at 1.12× per node. $O_{integration}$ is reset to 1.24 in 2013 and remains constant throughout the roadmap. The overheads are summarized in Table V. Separately, to compensate the reset and slowdown of M1HP scaling from 2013 to 2019, DES is set to 1.0× in 2013 and scales at 0.93× per year from 2013 to 2019, and 0.63× per year from 2020 onward.

TABLE VI. SUMMARY OF CHANGES BETWEEN 2011 AND 2013 ITRS MPU AND SOC AREA MODELS.

	Area Model Channges	
Year	**2011**	**2013**
Dominant Devices	Bulk	FinFET
Transistor Scaling	2× per node	1.6× per node
MPU-HP Area	$260mm^2$	$260mm^2$
SOC-CP Area	$100mm^2$	$140mm^2$
MPU-HP #Transistor (Logic + SRAM)	8.85B@2013	7.14B@2013
MPU-HP #Transistor (Logic)	1.60B@2013	3.68B@2013
MPU-HP #Transistor (SRAM)	7.25B@2013	3.46B@2013
SOC-CP #Transistor (Logic + SRAM)	2.02B@2013	2.40B@2013
SOC-CP #Transistor (Logic)	0.28B@2013	1.57B@2013
SOC-CP #Transistor (SRAM)	1.74B@2013	0.83B@2013

[2] A new node [21] occurs every two years from 2013 up to 2019, and then every three years beyond 2019.

[3] We use publicly announced transistor counts of Intel Xeon E5420 (released in 2007) and Intel Core i7-920 (released in-related 2008) to reset transistor count in 2013 for MPU-HP and MPU-CP respectively. We scale transistor counts from 2007/2008 to 2013 using the 1.6× per node scaling model to obtain 7.14B for MPU-HP and 2.54B for MPU-CP.

B. SOC Area and Overhead Models

In the 2013 ITRS roadmap, the die area of SOC-CP is reset to 140mm^2 after calibrating with recent product surveys. The transistor count has been reset to 2.4B transistors[4], very close to that of MPU-CP. Processing cores in SOC-CP are increasingly complex and implement similar features such as out-of-order execution, pipelining of up to 15 stages, large L2 caches, etc. The number of cores has been reset to four to limit power of the die. The number of logic transistors per core and the number of cores scale at the same rate as in the MPU-CP driver, that is, at 1.26× per node. Recent trends suggest a factor of 1.26× scaling of logic transistors per core with every node. The memory content of the SOC-CP is reset to 3.87MBytes and the number of SRAM transistors scales at 1.6× per node. The A-factor overhead $O_{Afactor-SRAM}$, that accounts for margin for reliability, write assist, etc., is 1.66 in 2013 and scales at 1.15× per node.

VI. Power Modeling and Power Management Gap

The methodology to model power in the 2013 ITRS roadmap is similar to that of the 2011 ITRS roadmap, which is explained in detail in [7]. We update device parameters, such as gate capacitance, average gate width value, transistor capacitance, capacitance of intermediate layers, and supply voltage, from the latest ITRS PIDS [27] and INTC [25] chapters. We retain the model of switching ratio from the 2011 power model described in [7]. The transistor density $D_{tr,logic}$ is derived from the 2013 area model described in Section V-A for MPU and Section V-B for SOC. We reset the maximum operating frequency in 2013 for MPU-HP to 5.5GHz and SOC-CP to 2.0GHz based on public domain data of leading-edge products in these driver classes. Table VII summarizes the main changes in the power model from 2011 to the 2013 roadmap. With the updated device parameters, we model MPU power using Equation (9).

TABLE VII. Summary of changes between 2011 and 2013 in the ITRS MPU and SOC power model.

Year	2011	2013
Power Scenario of MPU	Single Scenario	Single Scenario
Power Scenario of SOC	Single Scenario	Multiple Scenarios
Frequency scaling of MPU and SOC	1.04× / year	1.04× / year
SOC-CP Max Frequency	1.37GHz@2013	2.0GHz@2013
MPU-HP Max Frequency	7.34GHz@2013	5.5GHz@2013

$$P_{total} = \left(D_{dynamic,logic} + D_{static,logic}\right) \cdot S_{logic} + \left(D_{dynamic,SRAM} + D_{static,SRAM}\right) \cdot S_{SRAM} \quad (9)$$

In the equation, $D_{dynamic,logic}$ and $D_{dynamic,SRAM}$ are the dynamic power per unit area, and $D_{static,logic}$ and $D_{static,SRAM}$ are the static power per unit area, dissipated by the logic and SRAM transistors respectively. The modeling details for both dynamic and static power per unit area are explained in [7].

Since mobile devices have very strict power limits, idle functional blocks in the SOC-CP are aggressively power-gated. For example, GPUs are gated when only voice communications are required. Due to this scenario-dependent nature of the SOC-CP driver, it is not practical to maintain the single-scenario assumption of the 2011 ITRS roadmap. Moreover, SOC-CP drivers are required to support more and more application scenarios in the future. To address this evolution, the 2013 roadmap introduces power modeling based on weighted activity factors in each major block of the SOC-CP driver, across multiple usage scenarios. Table VIII shows a matrix that relates the five function blocks (CPU, GPU, PE, IO and RF) to four usage scenarios (voice, gaming, multimedia and maintenance). We calculate the power of each function block P_i, where $i \in$ {CPU, GPU, PE, IO, RF}, using Equation (9), and then calculate the total power from the power of each function block using Equation (10), where W_i is the weighted activity factor of each block and varies with different usage scenarios.

$$P_{SOC} = \sum_i P_i \cdot W_i \quad (10)$$

Due to the rapid growth predicted for the number of GPUs, we observe that the maximum power consumption of SOC-CP corresponds to the gaming scenario in 2028 in Table IX; this is consistent with our observation about the rapid growth of gaming-related requirements in the mobile device market. The power trend of the gaming scenario from 2013 to 2028 is shown in Figure 11. We expect that the newly-added GPUs become the most power-hungry component in the SOC-CP driver. Since the power consumption is predicted to increase to support higher gaming requirements, power in the gaming scenario will exceed 9W, an increase of 20% as compared to the 2011 ITRS SOC-CP power roadmap, at the 15-year horizon. As the typical peak power limit for mobile SOC is less than 4W, there remains a significant "power management gap" in the roadmap. SOC designers will require more aggressive low-power design techniques, ranging from architecture- to device-level, to address this power management gap.

TABLE VIII. Weighted activity factor of different functional categories among the scenarios.

	CPU	GPU	PE	IO	RF
Voice	0.50	0.00	1.00	1.00	1.00
Gaming	0.15	1.00	0.30	0.30	0.30
Multimedia	0.50	0.75	1.00	1.00	1.00
Maintenance	0.30	0.00	0.25	0.80	0.50

TABLE IX. Total SOC-CP power of different scenarios in 2013 and 2028.

	Power@2013 (W)	Power@2028 (W)
Voice	3.5	2.2
Gaming	3.2	9.1
Multimedia	5.1	8.5
Maintenance	1.9	1.2

VII. Conclusions

In this paper, we have described key aspects of the MPU and SOC system driver models in the 2013 ITRS roadmap. Important changes from previous driver models include new A-factors for SRAM and logic cells based on FinFET devices, and new area and power models for the MPU and SOC drivers. Updated architectural templates include uncore components for MPU drivers, as well as GPUs, PEs and other peripherals for the SOC driver. These

[4]We use publicly announced transistor counts of Qualcomm MSM 8974 (released in 2013) to reset transistor count in 2013 for SOC-CP.

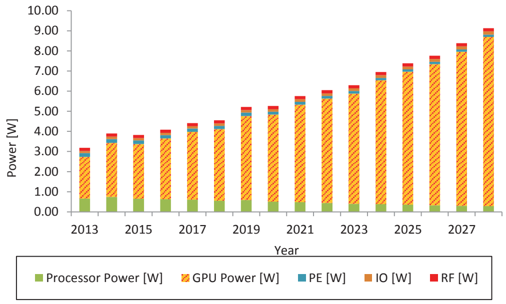

Fig. 11. Per-block type power roadmap of the SOC-CP driver.

added components, along with increased margins due to reliability, patterning, and other considerations, require introduction of new overheads in the models. The 2013 area models for MPU and SOC drivers reflect these new overheads.

A model calibration process using public domain data and structural analysis reports from Chipworks [16] reveals that a "scaling gap" – i.e., a gap between achievable layout density scaling, and the density scaling actually realized in commercial products – has existed since 2008. Transistor density scales at 1.6× per node, not at the 2× per node as roadmapped in the 2011 ITRS. Moreover, the 2013 ITRS resets the Metal-1 half-pitch roadmap, and then slows its scaling by an entire node from 2013 to 2019. According to die area models in the 2011 ITRS, this would cause an explosion in die area. To mitigate this challenge, and to preserve the historical improvement in layout density scaling through the end of this decade, an explicit design-based equivalent scaling (DES) mechanism has been introduced into the 2013 ITRS. We describe how DES potentially compensates the scaling gap and M1HP scaling slowdown, and how DES necessitates significant changes to ITRS chip area and power models. Finally, we describe an updated power model for the SOC-CP system driver, with multiple usage scenarios and calibration from structural analyses of leading-edge SOC products. The new A-factor, area and power models imply a 20% increase in the roadmap's already-identified power management gap, and an even more urgent requirement for low-power design innovation.

Acknowledgments

We gratefully acknowledge many discussions with members of the *ITRS International Technology Working Group* (ITWG) for Design Technology. Dr. Juan-Antonio Carballo and Dr. Mustafa Badaroglu provided extensive guidance and feedback throughout the course of model development and data collection. The updated SOC-CP architecture model and its calibration are due to collective efforts of the *Semiconductor Technology Roadmap Japan* (STRJ) Working Group 1, wherein one of us (I. Y.) led the model development effort. We thank Mr. Masaru Kakimoto and Mr. Toshitada Saito for additional discussions and feedback. We give special thanks to Chipworks for access to numerous recent product structural analysis reports which enabled development of the 2013 A-factor and driver models. Dr. Paolo Gargini coordinated and arranged the access to the product structural analysis reports. We include in this draft several excerpts from recent accounts published by one of us (A. B. K.), and we thank the copyright-holding entities for their indulgence.

References

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

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

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

[4] Y.-K. Choi, T.-J. King and C. Hu, "Spacer FinFET : Nano-Scale CMOS Technology for the Terabit Era", *Proc. International Semiconductor Device Research Symposium*, 2001, pp. 543-546.

[5] A. B. Kahng, "Design Capability Gap", *UCSD CSE Department Technical Report*, CS2013-1002, 2013.

[6] Z. Guo, S. Balasubramanian, R. Zlatanovici, T.-J. King, and B. Nikolic, "FinFET-Based SRAM Design", *Proc. International Symposium on Low Power Electronics and Design*, 2005, pp. 2-7.

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

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

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

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

[11] A. B. Kahng, "The ITRS Design Technology and System Drivers Roadmap: Process and Status", *Proc. Design Automation Conference*, 2013, pp. 1-6.

[12] A. B. Kahng, "Lithography-Induced Limits to Scaling of Design Quality", *Proc. Design-Process-Technology Co-optimization for Manufacturability VIII (SPIE Microlithography Symposium)*, 2014, pp. 905302-1-905302-14.

[13] A. B. Kahng and G. Smith, "A New Design Cost Model for the 2001 ITRS", *Proc. International Symposium on Quality Electronic Design*, 2002, pp. 190-193.

[14] D. Lekshmanan, A. Bansal and K. Roy, "FinFET SRAM: Optimizing Silicon Fin Thickness and Fin Ratio to Improve Stability at Iso Area", *Proc. Custom Integrated Circuits Conference*, 2007, pp. 623-628.

[15] T.-J. K. Liu, "FinFET History, Fundamentals and Future", http://www.eecs.berkeley.edu/~tking/presentations/KingLiu_2012VLSI-Tshortcourse.

[16] Chipworks, http://www.chipworks.com

[17] *Personal communication*, Chipworks, August-September 2013.

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

[19] "Moores Law Has Stopped At 28nm". http://electroiq.com/blog/2014/03/moores-law-has-stopped-at-28nm.

[20] "This is What the Death of Moores Law Looks Like: EUV Rollout Slowed, 450mm Wafers Halted, and an Uncertain Path Beyond 14nm". http://www.extremetech.com/computing/178529-this-is-what-the-death-of-moores-law-looks-like- euv-paused-indefinitely-450mm-wafers-halted-and-no-path-beyond-14nm.

[21] International Technology Roadmap for Semiconductors, http://public.itrs.net/.

[22] System Drivers Chapter in ITRS 2001 Update, http://www.itrs.net/Links/2001ITRS/Home.htm

[23] System Drivers Chapter in ITRS 2005 Update, http://www.itrs.net/Links/2005ITRS/Home.htm

[24] System Drivers Chapter in ITRS 2007 Update, http://www.itrs.net/Links/2007ITRS/Home.htm

[25] Interconnect Chapter in ITRS 2011 Updates, http://www.itrs.net/Links/2011ITRS/Home2011.htm.

[26] "Moore's Law Dead by 2022, Expert Says". http://www.eetimes.com/document.asp?doc_id=1319330.

[27] Process Integration, Devices and Structures (PIDS) Chapter in ITRS 2011 Updates, http://www.itrs.net/Links/2011ITRS/Home2011.htm.