Future Performance Challenges in Nanometer Design

Dennis Sylvester
University of Michigan
Ann Arbor, MI 48109-2122
dennis@eecs.umich.edu

Himanshu Kaul
University of Michigan
Ann Arbor, MI 48109-2122
hkaul@engin.umich.edu

ABSTRACT

We highlight several fundamental challenges to designing high-performance integrated circuits in nanometer-scale technologies (i.e. drawn feature sizes < 100 nm). Dynamic power scaling trends lead to major packaging problems. To alleviate these concerns, thermal monitoring and feedback mechanisms can limit worst-case dissipation and reduce costs. Furthermore, a flexible multi-V_{dd} + multi-V_{th} + re-sizing approach is advocated to leverage the inherent properties of ultra-small MOSFETs and limit both dynamic and static power. Alternative global signaling strategies such as differential and low-swing drivers are recommended in order to curb the power requirements of cross-chip communication. Finally, potential power delivery challenges are addressed with respect to ITRS packaging predictions.

1. INTRODUCTION

Many challenges confront device engineers, circuit designers, system-level architects, and electronic design automation (EDA) tool developers in nanometer (sub-0.1μm) design. They can be broadly categorized as speed, power, reliability, and variability challenges. Specific examples include soft error rates (reliability), increasing V_{th} fluctuations across a large die (variability), full-chip inductance extraction (reliability/signal integrity), rising global interconnect latency (delay), and distributing V_{dd}/GND stably despite large current transients and massive supply currents (power). This paper will center on power-related challenges for high-performance IC design (e.g., for desktop microprocessor (MPU) applications) in the 50nm and 35nm technology nodes at the end of the ITRS. Our discussion will highlight key *challenges* facing designers and EDA developers, existing or proposed *solutions* to these challenges, and new ideas that may help circumvent the biggest challenges.

This paper does not address such important issues as difficulties in synchronization at extremely high clock rates, the impact of growing process variability, signal integrity, etc. We focus on power-related issues because power consumption has more widespread implications than the above issues. For instance, limitations in power management capabilities can fundamentally restrict *performance*.[1] In Section 2, we see that power-related packaging limitations place bounds on die area and integration density. Removing these limits by better packaging/cooling or other methods improves overall performance, not merely power management. Furthermore, static power dissipation and transistor drive current are linked – it is in large part transistor drive current that enables high speed ICs. In general, any roadblocks to the long-standing trends of rising transistor density, die sizes, and clock frequency/throughput can be seen as challenges to performance. Whether they are commonly viewed as reliability problems, signal integrity, power management, etc., the consequence of such challenges is to limit IC performance. To summarize, while the submicron and deep submicron regimes concentrated on maintaining device and circuit speed improvements despite shrinking supply voltages, nanometer design will be most concerned with limiting power consumption while sustaining throughput and reliability.

In the following, Section 2 examines dynamic power, as well as tradeoffs and possible ways to reduce power limitations on performance. Section 3 explores *static* power consumption's increasing importance, even in desktop (non-portable) applications. Section 4 examines difficulties in distributing power to increasingly larger ICs under restrictive performance targets (e.g. IR drop < 5-10% of a shrinking V_{dd}). Throughout, we refer to the 2000 update of the ITRS and highlight important trends and key deviations required to continue relatively unabated along the roadmap. We also use predictive MOS SPICE models [2] as well as a realistic 50nm device model extracted from rigorous process and device simulations [3].

2. DYNAMIC POWER

2.1 Packaging Limitations

With the forecasted increase in MPU power consumption, IC packaging will bear the burden of dissipating even more heat in the future. A package's ability to remove waste heat is defined by the junction-to-ambient thermal resistance (θ_{ja}), expressed as:

$$\theta_{ja} = \left(T_{chip} - T_{ambient}\right) / P_{chip} \quad (1)$$

In (1), T_{chip} is the on-die junction temperature, $T_{ambient}$ is the ambient (outside package) temperature, and P_{chip} is the *maximum* IC power consumption. Given a packaging solution with a fixed θ_{ja} and an MPU design consuming P_{chip}, the resultant on-die temperature can be calculated using (1). Alternatively, when considering packaging solutions for a new MPU, the maximum allowable θ_{ja} can be determined based on constraints for maximum on-die temperature (this is typically limited to ensure correct operation of the MPU). Currently, IC operation frequently pushes the junction temperature beyond 100°C while $T_{ambient}$ is approximately 45°C. With P_{chip} rising, packaging technology must improve (meaning θ_{ja} must decrease) to meet heat dissipation demands. Consistent reduction of thermal junction resistance requires advanced cooling techniques such as larger, more powerful (and louder) fans, liquid cooling, etc. Furthermore, to ensure reliability in nanometer scale MPUs, the ITRS calls for a reduction in junction temperature (from 100°C in 1999 to 85°C in 2002). Due to cost constraints, achieving the corresponding θ_{ja} values for packaging is considered a barrier to scaling – the materials needed are currently unknown. Presently, θ_{ja} values range from 0.6 to 1 °C/W for the workstation/desktop processor markets [4]. ITRS projections call for a θ_{ja} of 0.25 °C/W in 3 years – requiring improvements in the CPU package (ceramics, etc.) as well as heat sinking technology. Allowing T_{chip} to rise allows for less complex and expensive packaging solutions to be used, but this adversely affects circuit performance with respect to leakage current and device reliability. Some packaging experts believe cooled systems are the best alternative for packaging high power density nanometer microprocessor designs. The advantages of cooling the ambient and junction temperatures are well documented: improved voltage scalability due to reduced leakage currents, higher carrier mobilities, lower interconnect resistances, and improved reliability [5]. However, as a reference point, current vapor

[1] That is, any important design metric such as clock speed/throughput, integration density, power dissipation, reliability/yield, etc.

DAC 2001, June 18-22, 2001, Las Vegas, Nevada, USA.

compression based refrigeration techniques are expensive, on the order of $1 per watt cooled. Such measures for desktop applications in the next decade will likely not be needed, due to improved heat sinking technology and evolving low power design techniques applied to high-end processors.

The above packaging-constrained system perspective leads into the concept of dynamic thermal management [6]. Thermal management techniques can take a number of forms. Transmeta's approach dynamically varies the supply voltage when the CPU is not heavily loaded. Simpler techniques can be used with only minor changes to a straightforward processor implementation. An example is the thermal monitor in Intel's Pentium 4 design [7], which has an on-chip temperature sensor (a diode with a fixed voltage across it) along with a reference current source and current comparator to determine when the on-die temperature exceeds a given value. This temperature corresponds to a power dissipation level for the microprocessor (determined by (1)). When the temperature (and power consumption) is exceeded, the internal clock frequency is reduced, limiting power and performance / throughput.

The importance of dynamic thermal management techniques lies in their ability to reduce P_{chip} in (1) to the *effective worst-case* power dissipation rather than the theoretical worst-case [6]. The effective worst-case power consumption, as found by running power-hungry applications, is about 75% of the theoretical worst-case, which is determined using synthetic input code sequences that are not realized in practice [7,8]. This difference has major implications for packaging costs and design flexibility. Small increases in the maximum power can lead to significantly more exotic, expensive cooling techniques. For example, Intel engineers found that a rise in power consumption from 65 to 75 W would triple cooling costs due to the need for additional heat pipe technology to achieve the required θ_{ja} [7]. With an effective 25% reduction in P_{chip}, the allowable θ_{ja} is 33% higher, translating to less expensive heat sinking, quieter and smaller fans, and avoidance of refrigerated or liquid-cooled solutions.

2.2 Global Signaling

Propagation of global signals across a large die in a shrinking clock period is one of the foremost challenges in nanometer design [1,9,10]. It appears likely that global signaling will use a slower clock than localized logic such as datapaths (despite the fact that multi-cycle nets can be broken up using latches). For example, a recent Intel microprocessor clocks the integer ALUs at a higher rate than other sections of the design. Even with relaxed timing constraints on global communication, substantial power is consumed to achieve the desired global clock speeds. [9] demonstrates that using unscaled top level wiring, ITRS projected global clock frequencies can be met. Based on the current signaling paradigm of inserting large CMOS buffers along an RC line, this requires over 50 W of power in the nanometer regime. The proliferation of repeaters (nearly 10^6 required at 50-nm compared to about 10^4 in a large 180nm microprocessor [11]) heightens difficulties in power distribution and floorplanning[2].

An alternative is to use advanced signaling strategies such as differential and/or low-swing drivers and receivers for global communication [12]. In many cases, these approaches can lead to power *and* delay savings due to smaller voltage transitions as well as major reductions in the magnitude of power grid current transients. For instance, the Alpha 21264 uses differential low-swing buses to communicate between functional units [8]. Worst-case power for these buses was reduced significantly by limiting the voltage swing to 10% of V_{dd}. Differential signaling increases routing area, but the increase may be less than the expected factor of 2 due to the use of shield wires in global signaling to limit coupling from neighboring signals on long lines. Furthermore, shielding may be insufficient to limit inductively coupled noise, whereas low-swing differential signaling creates less noise and is more noise immune than single-ended full-swing CMOS [13]. While further study is necessary to determine worst-case noise behavior and tolerable voltage swings, the Alpha design demonstrates that the approach is already viable today. With trends indicating rising power consumption for global communication, the use of alternative signaling strategies will likely increase.

[2] Repeater clusters constrain repeater placement to ease floorplanning and simplify insertion of repeaters late in the design. Resulting power densities can exceed 100 W/cm^2, complicating power distribution.

2.3 Library Optimization

While most high performance microprocessors rely heavily on custom design, library optimization can still enhance performance in these applications. System complexity and the resulting design productivity needs mean that some components of nearly every IC design will draw from a cell library. Advances in library generation, and synthesis tools that take advantage of improved libraries, can together yield more automated, less expensive design flows. Recent work claims libraries are one important reason that custom designs are significantly faster (6-8X) than counterpart ASIC designs [14,15]. For instance, [15] asserts that the lowest performance level (smallest) gates in modern libraries are nearly 10X larger than minimum-sized gates, leading to major power increases due to overdriving small loads. However, most current libraries contain a large number of drive strengths, including some very near minimum size. As evidence, we cite the same 180 nm library as [15]: the smallest standard cell inverter has an input capacitance of just 1.5fF (smaller than the custom gate in [15]) and the smallest inverter with balanced rise/fall delays has an input capacitance of 6.6fF [16]. Other leading-edge libraries contain a rich set of drive strengths (e.g. 11 2-input NANDs, 16 inverter sizes), dual output polarities, and single pin inverted inputs on NAND/NOR's.

This recent increase in library complexity seems to be closing the gap slightly between custom designed cells and those from libraries. However, more work needs to be done: a recent study [17] demonstrates the potential of on-the-fly cell generation layered on top of a pre-existing rich library. Results show 15-22% power reductions with fixed timing, and one design achieved 13.5% speed gains and 18% power reduction. In these cases overnight optimizations created hundreds of new cells, adding flexibility to the original library and more closely approximating a *custom* design approach. The new cells serve to exactly match load conditions (limiting overdrive of small capacitances) and allow for imbalanced P/N sizing if advantageous.

2.4 Multiple-V_{dd}

Multiple supply voltages on a chip will be one of the most valuable tools for designers to fight the rise of dynamic power in nanometer design. Only a few designs based on this concept, all with relatively low clock speeds, have been reported [18,19]. However, results are promising, and the slow acceptance in high-performance MPUs seems primarily due to a lack of urgency in dynamic power reduction.

The general idea most often applied is that of clustered voltage scaling (CVS) [20]. With two V_{dd} levels ($V_{dd,h}$ and $V_{dd,l}$), the circuit is partitioned so that non-critical gates run at $V_{dd,l}$ and only critical gates use $V_{dd,h}$. Level conversions, performed when gates running at $V_{dd,l}$ fan-out to gates at $V_{dd,h}$, are reduced by clustering $V_{dd,l}$ and $V_{dd,h}$ gates together to minimize the number of such interactions.

Analysis indicates that $V_{dd,l}$ should be around 0.6 to 0.7 times $V_{dd,h}$ to maximize power savings. The dynamic power reduction by using two V_{dd} levels is readily calculated if one can estimate the fraction of cells that can be assigned to $V_{dd,l}$. Existing media processor designs that use CVS report that ~75% of all gates can tolerate $V_{dd,l}$ without altering the critical path delay. Similarly, path slack distributions for high-end MPUs show that over half of all timing paths commonly use less than half the clock cycle [21,22]. Using $V_{dd,l} = 0.65 * V_{dd,h}$, this yields a 45-50% dynamic power reduction, considering 8-10% additional

level conversion power. In [18], area overhead due to constrained cell placement, level converters, and added power grid routing was found to be 15%. The impact of post-synthesis transistor re-sizing on multi-V_{dd} processes is discussed in Section 3.3.

The key challenges to the use of multiple supplies on a chip lie in minimizing area overhead and providing EDA tool support for V_{dd} cell selection, placement given new clustering constraints, dual power grid routing, and enhanced library generation capabilities. In Section 3.3 we describe the major improvements that can be achieved in the delay vs. V_{dd} design space by use of multiple threshold voltages. With this new concept, the idea of using multiple supplies on a chip becomes much more powerful.

3. STATIC POWER

3.1 ITRS Projections & Analysis

The ITRS predicts an increase in MOSFET off current (I_{off}) by a factor of 2 per generation. The author of [23] projects a 5X rise in I_{off}/generation. Figure 1 shows the relative importance of static and dynamic power for an inverter driving a fan-out of 4 with an average interconnect load. 70 nm and 50 nm technologies are explored; results indicate that for logic with switching activities on the order of 0.01 to 0.1, static power can approach and exceed 10% of dynamic power. The ITRS calculates the expected increase in static power consumption due to I_{off} and sets constraints to limit static power to 10% of the maximum power dissipation of the MPU. Hence at 35 nm, an MPU can draw 30A of current *in standby*. Even with this mild restriction on static power consumption, the reduction needed by circuit/architecture innovations reaches 98% at the end of the roadmap [1]. Unchecked, static power would reach kilowatt levels, dwarfing dynamic power. Circuit and architectural techniques have been proposed to reduce standby power. These approaches will become standard in low-power applications and experience with these designs will ease integration into high performance ICs. Some of these techniques are described in the following section.

To give further perspective on I_{off} scaling, we examined recent literature on advanced CMOS processes, noting the I_{on}, I_{off}, V_{dd}, and T_{ox} (oxide thickness) values. Results are summarized in Table 1. The key point of this table is that, while very good I_{on}/I_{off} characteristics are achieved, there are no examples of sub-1 V technologies that come close to meeting ITRS expectations. For instance, the 70nm technologies described in [26,28] offer leakage currents below that projected by the ITRS with I_{on} values slightly lower than forecast. However, the V_{dd} value required to achieve this performance is 1.2 V – not 0.9 V as expected for 70nm. This V_{dd} increase gives a 78% rise in dynamic power.

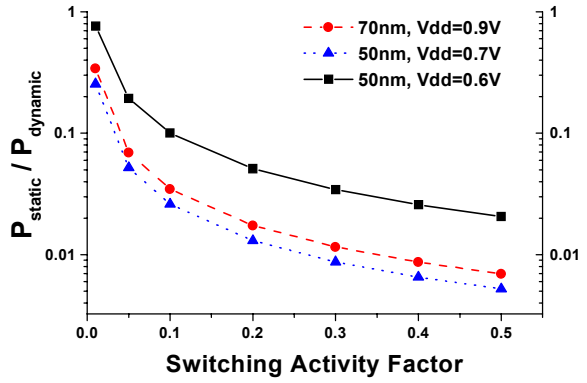

Figure 1. The ratio of static power consumption to dynamic power for an inverter with fan-out of 4 and average wiring load. Temperature is 85°C.

Table 1. Recent NMOS device results, compared with ITRS projections.

Ref	ITRS node (nm)	T_{ox} (Å) (electrical)	V_{dd}	I_{on} (µA/µm)	I_{off} (nA/µm)
[24]	50-70	18	0.85	514	100
[25]	100	21	1.2	860	10
[26]	70	25	1.2	697	10
[27]	100	27	1.2	800	10
[28]	70	32	1.2	650	3
[29]	100	13 - physical	1.0	723	16
ITRS	100	12-15 physical	1.2	750	13
ITRS	70	8-12 physical	0.9	750	40
ITRS	50	6-8 physical	0.6	750	80

While the current literature is not scalable to sub-1 V supplies, there will be improvements when these processes come online (2-5 years). Looking at historical references, reports of pre-production technologies tend to underestimate I_{on} by ~20% compared to actual performance several years later [30,31]. Unfortunately, most of the gains in I_{on} from R&D to production have been obtained from aggressive oxide scaling. This performance "lever" may be approaching the end of its usefulness; even with high-dielectric materials, maintaining current scaling trends for the effective oxide thickness faces a number of barriers in nanometer design.

This point is further described using a set of compact MOSFET I-V expressions to project the scaling of I_{on} and I_{off} in nanometer scale processes [32]. I_{on} is expressed as:

$$I_{on} = I_{dsat0} \Big/ \left(1 + \frac{2I_{dsat0}R_s}{V_{dd} - V_{th}} - \frac{I_{dsat0}R_s}{V_{dd} - V_{th} + E_{sat}L_{eff}}\right) \quad (2)$$

R_s is the parasitic source resistance (set according to [1]), E_{sat} is the lateral electric field required to saturate the carrier velocity, and L_{eff} is the effective gate length (final, as-etched dimension in [1]). I_{dsat0} is:

$$I_{dsat0} = \frac{W\mu_{eff}C_{oxe}}{2L_{eff}} \frac{(V_{dd} - V_{th})^2}{1 + (V_{dd} - V_{th})/E_{sat}L_{eff}} \quad (3)$$

Here μ_{eff} is the effective mobility, which is a function of gate voltage and T_{ox}. C_{oxe} is the electrical oxide capacitance, described later.

Off current (per unit width) is estimated as [33]:

$$I_{off} = 10 \times \left(10^{\frac{-V_{th}}{85mV}}\right) \mu A/\mu m \quad (4)$$

85 mV is the assumed subthreshold swing parameter throughout scaling (taken at room temperature to match [1])[3]. An analytical analysis of the ITRS on/off current projections is summarized in Table 2. The V_{th} for each technology is set to meet 750µA/µm for I_{on}. We make the following observations:

1. Including *electrical* oxide thickness is important and should be considered in the ITRS. Electrical oxide thickness reflects the finite inversion layer thickness (i.e. the inversion layer is not a sheet of charge located at the Si/SiO_2 interface) and gate depletion effects (GDE) [32]. The net effect is that the oxide appears ~0.7 nm thicker than the physical oxide layer. Advanced gate materials may limit the contribution of GDE, however the quantization of the inversion layer will be unaffected. An analysis ignoring GDE but incorporating inversion layer thickness (denoted "metal gate" in Table 2) shows I_{off} decreases by 78% at 35 nm. Enhanced current resulting from a thinner effective gate oxide allows a 55 mV increase in V_{th}, significantly reducing I_{off}.

[3] Technologies such as fully-depleted SOI may reduce this value considerably (i.e. by 20%), making lower thresholds feasible given fixed I_{off} constraints.

Table 2. Analytical model results for I_{off} scaling. Values in ( ) for 50nm are results for V_{dd}=0.7V.

ITRS node (nm) ⇒	180	130	100	70	50	35
C_{oxe} (normalized)	1	1.23	1.45	1.68	2.13	2.46
C_{ox} (physical)	1	1.32	1.67	2.08	3.13	4.17
V_{th} required to meet I_{on}	0.3	0.29	0.22	0.14	0.04 (0.12)	0.11
I_{off} (nA/μm)	3	4	26	210	3205 (432)	456
I_{off} (metal gate)	1	1.4	8.7	55	666 (100)	103
ITRS I_{off} projections	7	10	16	40	80	160

2. A 0.6 V supply voltage for 50 nm high-performance parts will make it difficult to achieve the desired I_{on}/I_{off} targets. A V_{dd} of 0.7 V is more realistic (given that V_{dd} for 35 nm is projected as 0.6V), reducing off current by nearly 7X but increasing dynamic power by 36%. Extracted 50 nm device parameters support this – simulations demonstrate a marked increase in I_{off} at V_{dd}=0.6 V to meet ITRS I_{on} (I_{off} = 2.6 μA/μm at 0.6 V, 430 nA/μm for 0.7 V).

3. The projected I_{off} from the models is 3 nA/μm for 180 nm, rising to 456 nA/μm for 35 nm. The increase of 152X is markedly higher than the ITRS value of 23X[4]. Furthermore, the leakage current at 35 nm here is 2.9X larger than ITRS projections. This translates to additional static power reduction required by circuit design techniques. In general, the 2X increase in I_{off}/generation listed in [1] allows just a 25mV drop in V_{th} in each technology. Following this constraint, the models show a 16% loss in I_{on} by the end of the roadmap[5]. We note, however, that the 152X increase in I_{off} across the roadmap is much less than predicted by [23] which anticipates a 3125X rise by 35nm.

3.2 Multiple-V_{th} Approaches

Several approaches have been developed to reduce CMOS static power consumption. This section briefly highlights several of these techniques that use multiple thresholds on a single chip to limit I_{off}.

3.2.1 MTCMOS and variants

Multi-Threshold CMOS (MTCMOS) gates a high-V_{th} transistor with a sleep mode signal to virtually eliminate leakage current in idle states [34]. The sleep transistor is placed between ground and fast low-V_{th} CMOS logic. As it is in series, it adds delay, which can be reduced by increasing its area. Disadvantages include no leakage reduction in active mode, increased device area, and additional overhead for routing sleep signals.

Other related techniques include dual-V_{th} domino logic [35], substrate biasing to modify V_{th} in standby [36], and using negative NMOS gate voltages to bias the devices further into cut-off [37]. A single-threshold leakage reduction technique combines the concepts of sleep transistors and state dependent leakage [38]. All these techniques trade off area to limit static power and most only reduce leakage in standby mode. In practice, they are currently limited to portable applications such as notebook processors. Also, some of the proposed methods do not scale well – the use of domino logic for example, and substrate bias controlled V_{th} (body bias is less effective at controlling V_{th} in scaled devices). Dual V_{th} insertion, described next, is the only technique used in current high-end MPUs.

3.2.2 Dual-V_{th}

Recently, circuit designers gained access to multiple threshold voltages on a single IC to select between gates that use high or low thresholds. The impact of V_{th} on the delay and power of gates such as inverters and NANDs is profound. As seen in (4), a reduction in V_{th} (with constant V_{dd}) exponentially increases off current and roughly linearly reduces propagation delay. An additional threshold adjust ion implantation step allows designers to choose from a wider range within the power-performance design envelope. Gates located on critical paths can be assigned fast low V_{th}, while gates that are not timing critical can tolerate high V_{th} and slower response times. Algorithms have been developed to optimally assign gates to either high or low threshold voltages [22,39]. Typical results show leakage power reductions of 40-80% with minimal penalty in critical path delay compared to all low-V_{th} implementations.

It is instructive to examine the scaling properties of a dual-V_{th} approach to limiting I_{off}. Based on (2)-(4), we consider two NMOS devices in the same technology with thresholds offset by 100 mV. The high-V_{th} device has its V_{th} set so that I_{on} is 750 μA/μm. Figure 2 shows the increase in I_{on} for the low-V_{th} device. The relative difference in I_{off} between the two devices will remain constant throughout the roadmap (at about a 15X increase in I_{off} for 100 mV reduction in V_{th}). Given that the off current change is constant, the steady improvement in I_{on} with scaling demonstrates that the dual-V_{th} (or multi-V_{th}) approach to leakage reduction is inherently scalable. Figure 2 also shows the resulting I_{off} increase for I_{on} to rise 20% beyond the high-V_{th} case. At 35 nm, just a 7X rise in I_{off} is required to yield 20% drive current improvement, compared with a factor of 54X today. Published data from [21,40] validate the models, as seen in Figure 2.

3.3 Scalable Dynamic/Static Power Approach

The combination of multiple V_{dd}'s, multiple V_{th}'s, and intra-cell size and V_{th} assignments points to a highly flexible, scalable, cost-effective design approach to dynamic *and* static power minimization. With two voltage supply values available, different V_{th}'s will allow designers or EDA tools to choose to emphasize speed, standby power, or dynamic power. Figure 3 demonstrates the potential of the multi-V_{dd} + multi-V_{th} approach. In 35nm technology, a reduction in V_{dd} from nominal (0.6V) to 0.2V incurs a severe delay penalty (normalized delay is 3.7X that at 0.6V). However, by reducing V_{th} in the gates using 0.2V supplies, the delay increase is *less than 30% while dynamic power is 89% lower and static power is constant*. These compelling results are the product of two powerful ideas: 1) MOSFET drive current when using sub-1V V_{dd} are very sensitive to V_{th}, so small reductions in V_{th} achieve major current gains. 2) Static power decays roughly quadratically with V_{dd} reductions (given a fixed V_{th}) due to shrinking I_{off} and a smaller V_{dd} value. Figure 3 harnesses these two concepts by slowly reducing V_{th} as V_{dd} is dropped so that I_{off} rises at the same rate V_{dd} is shrinking, keeping P_{static} constant. Figure 4 shows that the vast improvements in dynamic power and a constant P_{static} push the ratio of $P_{dynamic}/P_{static}$ towards 1 for low switching activ-

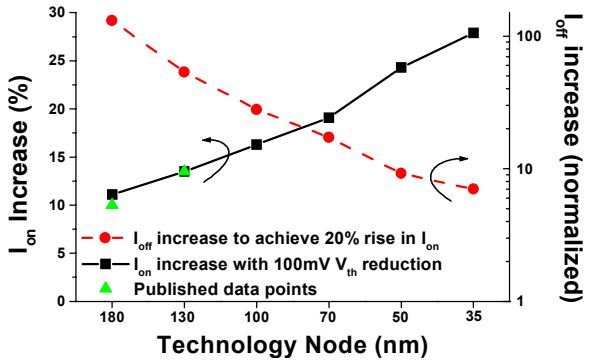

Figure 2. I_{on} increases more rapidly with a 100mV change in V_{th} for scaled technologies. I_{off} penalty for 20% I_{on} gain reduces with scaling.

[4] The slope of I_{off} vs. technology is larger for the models as well, meaning a fast rise in leakage may be ahead.

[5] This includes a reduction of 37% at 50 nm (19.6% if V_{dd} = 0.7 V).

ity gates at V_{dd}=0.2V[6]. If a constraint is set that $P_{dynamic}$ must be 10X larger than P_{static} (as in the ITRS), a V_{dd} of about 0.44V is attainable, providing 46% dynamic power reduction. More options are available; Figure 3 shows that if threshold voltage is scaled less aggressively than required to maintain constant P_{static}, delay increases more quickly but remains reasonable at 1/3 the nominal V_{dd} value. In this scenario, the static power is being reduced linearly with V_{dd} so that P_{static} is 1/3 that of a gate using V_{dd}=0.6V.

Now, consider post-synthesis transistor re-sizing, which reduces power by down sizing transistors off critical paths [21]. As a result, more paths approach criticality; this makes the application of multi-V_{dd} approaches less advantageous since fewer cells than assumed above (75%) can move to $V_{dd,l}$. This point highlights the sub-optimal nature of today's low power design techniques. If, before transistor re-sizing, slack distributions demonstrate a large number of paths with significant slack, the current approach is to down size the corresponding cells, slowing down that path. This approach provides a sublinear reduction in power with respect to the size reduction (sublinear since interconnect capacitance will not scale down and represents a constant factor in the total capacitance). Instead of such re-sizing efforts, a lower supply voltage could be used, providing a *quadratic* drop in power. Leakage power will be significantly reduced in this case due to the V_{dd} reduction as well as I_{off} which also decreases. The combination of multiple V_{dd}'s, multiple V_{th}'s, and transistor re-sizing needs to be harnessed in future EDA tools to achieve excellent power/performance results.

Combining the above multi-V_{dd} + multi-V_{th} optimization strategy with the on-the-fly cell generation approach of Section 2.3, designers and EDA tools can fully explore the design space of dynamic power, static power, and timing slack. One example of unique gate layouts that could help face the power challenges of nanometer design is the use of different V_{th}'s inside a cell. Particularly, the use of different threshold transistors in a stacked arrangement can give fairly substantial leakage savings with minimal delay penalties. Furthermore, the state dependence of leakage can be leveraged in cases with stacked multi-V_{th}'s without additional sleep transistors that sacrifice area and dynamic power.

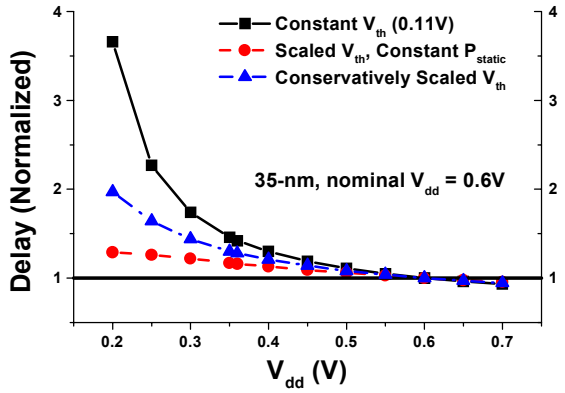

Figure 3. Delay increase due to V_{dd} reduction can be effectively offset by reducing V_{th}.

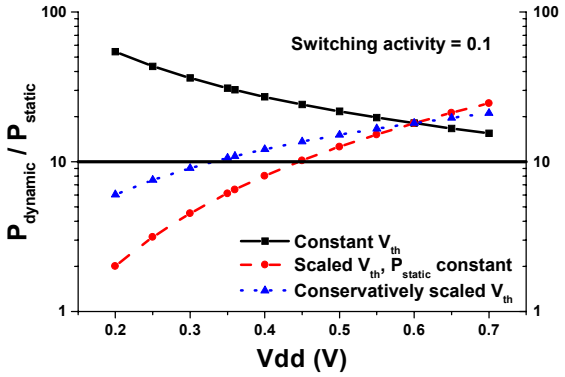

Figure 4. The ratio of dynamic to static power drops when using low V_{th}'s to reduce delay penalty in low V_{dd} gates. 35-nm technology is modeled.

[6] $P_{dynamic}$ is calculated using a fan-out of 4 and an average wiring load. Gates are inverters with $W_n/L=4, W_p/L=8$.

4. POWER DISTRIBUTION

Flip-chip and grid array packaging allows distribution of V_{dd}/GND and signals throughout a die, rather than just at the periphery. This increased flexibility makes power grid IR drops substantially more manageable, to meet 10% IR drop constraints, etc. However, in this section we show that current ITRS projections for power/grid pad connectivity in nanometer designs do not fully take advantage of grid array capabilities and lead to power distribution problems.

Based on BACPAC models [41], we examine the scalability of typical power grid distribution in the face of quickly rising chip current supplies. Hot-spots are considered since uniform power density assumptions are overly optimistic. A hot-spot is defined to have a localized power density four times larger than a uniform power density approximation (given by P_{chip} / A_{chip})[7].

Figure 5 shows the required power rail width (normalized to minimum top-level metal width) to ensure <10% IR drop in "hot-spots" of a design in scaled technologies using the minimum allowable bump pitch. This figure focuses on top-level routing only, assuming that the remainder of the power grid is under the designers control whereas the top-level granularity is technology-limited[8]. 35 nm is less restricted than 50 nm due to a reduction in power density at 35 nm[9]. In general, while the trend seems alarming (roughly quadratic increase in power rail linewidth, normalized to minimum allowable linewidth), even 35 nm results are manageable, in that V_{dd} and GND rails that are 16X minimum width will consume less than 4% of top-level routing resources (based on 80 µm bump and power-grid pitch). The total routing resources consumed due to power routing is around 17-20% as a constant factor of 16% is used to reflect the need for large metal "landing pads" for the bumps. The continued reductions in bump pitch allow V_{dd}/GND to be supplied at finer granularities where it is most needed.

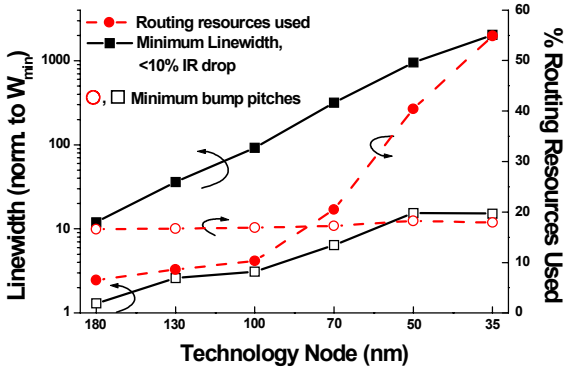

Figure 5. IR drop scaling trends based on minimum allowable bump pitch (open symbols) and ITRS bump/pad count projections (solid symbols).

[7] The factor of four stems from estimating that half the chip area is consumed by memory (having about 1/10th the power density of logic) and that certain logic areas may have twice the power density of others.

[8] Meaning that the chip's access to V_{dd}/GND is limited by how often connections can be made to the off-chip supplies.

[9] Total power at 50 nm increases only slightly while the area jumps 15%.

However, ITRS projections for microprocessor pad counts do not correspond to the minimum achievable bump pitch. For instance, a bump pitch of 80 µm is estimated to be attainable at 35 nm, but the number of bumps actually used is 4416, translating to an effective bump pitch of 356 µm. Since IR drop is strongly dependent on the periodicity of power connections, this large bump pitch results in a staggering increase in wiring resources needed to maintain adequate IR drops. Figure 5 also shows the required power rail widths under the ITRS assumptions of bump/pad count. At 35 nm, the required line width is over 2000X the minimum allowable; this is the result of a roughly constant bump pitch of around 350 µm throughout the roadmap. More V_{dd} and GND connections will be required and advances in technology should be leveraged rather than consuming extra routing resources. In addition, with just 1500 V_{dd} bumps at 35 nm, ITRS bump current capability projections are incompatible with the worst-case current draw of 300A in such a design. This also points to the need for more V_{dd}/GND connections at the chip-to-package level.

Finally, rising supply currents and the use of sleep or standby modes to reduce power have potential consequences in power distribution. Awakening from standby results in large current transients, placing an extreme burden on the power distribution network to limit inductive noise. Using the minimum bump pitch will help here as well, providing a low inductance path to each gate on the chip. Alternate logic styles may minimize current transients and provide superior power-delay characteristics. One option is MOS current mode logic (MCML), which burns static power but yields much smaller current transients while providing comparable performance and lower total power in high activity circuitry such as datapaths [42]. If a point is reached where static CMOS leakage currents are intractable, current steering logic families such as MCML may provide solutions.

5. CONCLUSIONS

The main points of this paper are:

1. Power management techniques such as on-chip temperature monitors and multiple voltage supplies will reduce dynamic power, enabling cheaper packaging and higher integration densities.

2. Alternative techniques to CMOS repeaters for global signaling need to be investigated and mated with EDA tools (similar to buffer insertion tools today but using different primitive components) to minimize power consumed in global communications.

3. A multi-layered approach to power reduction (both dynamic and static) is described, combining multiple threshold and supply voltages with flexible gate layouts using different thresholds and device sizes within a gate. Non-critical gates are first assigned to a reduced V_{dd}, followed by sizing and V_{th} selection to reduce power most efficiently.

4. Power distribution will be manageable from the standpoint of IR drop – given changes in the ITRS to take advantage of technological advancements in flip-chip packaging. However, large current transients may be exacerbated by the use of sleep/standby modes.

6. ACKNOWLEDGMENTS

The authors thank Kurt Keutzer, Andrew Kahng, and Dave Chinnery for valuable comments, Pin Su, Charles Kuo, and Min She for device models, Richard Hamilton for packaging discussions, and Philippe Hurat and Martin Lefebvre at Cadabra Design.

7. REFERENCES

[1] http://public.itrs.net, ITRS, 2000 update.
[2] Y. Cao, *et al*., "New paradigm of predictive MOSFET and interconnect modeling for early circuit simulation," *Proc. CICC*, pp. 201-204, 2000.
[3] Personal communication, Pin Su & Charles Kuo.
[4] R. Viswanath, *et al*., "Thermal performance challenges from silicon to systems," *Intel Technology Journal*, 3rd quarter, 2000.
[5] I. Aller, *et al*., "CMOS circuit technology for sub-ambient temperature operation," *Proc. ISSCC*, pp. 214-215, 2000.
[6] D. Brooks and M. Martonosi, "Dynamic thermal management for high-performance microprocessors," *Proc. High-Performance Comp. Arch.*, 2001.
[7] S.H. Gunther, *et al*., "Managing the impact of increasing microprocessor power consumption," *Intel Technology Journal*, 1st quarter, 2001.
[8] M.K. Gowan, *et al*., "Power considerations in the design of the Alpha 21264 microprocessor," *Proc. DAC*, pp. 726-731, 1998.
[9] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron II: A global wiring paradigm," *Proc. ISPD*, pp. 193-201, 1999.
[10] R. Ho, K. Mai, H. Kapadia, and M. Horowitz, "Interconnect scaling implications for CAD," *Proc. ICCAD*, pp. 425-429, 1999.
[11] R. McInerney, *et al*, "Methodology for repeater insertion in the Itanium microprocessor," *Proc. ISPD*, pp. 99-104, 2000.
[12] H. Zhang, *et al*., "Low-swing on-chip signaling techniques: effectiveness and robustness," *IEEE Trans. VLSI Systems*, pp. 264-272, Jun. 2000.
[13] Y. Massoud, *et al*., "Differential signaling in crosstalk avoidance strategies for physical synthesis," *Proc. TAU*, 2000.
[14] D.G. Chinnery and K. Keutzer, "Closing the gap between ASIC and custom: an ASIC perspective," *Proc. DAC*, pp. 637-641, 2000.
[15] W.J. Dally and A. Chang, "The role of custom design in ASIC chips," *Proc. DAC*, pp. 643-647, 2000.
[16] IBM SA-27E ASIC standard cell datasheet.
[17] P. Hurat, "Beyond physical synthesis," SNUG Europe 2001.
[18] K. Usami, *et al*., "Automated low-power technique exploiting multiple supply voltages applied to a media processor," *IEEE J. Solid-State Circ.*, pp. 463-472, Mar. 1998.
[19] M. Takahashi, *et al*., "A 60-mW MPEG4 video codec using clustered voltage scaling with variable supply-voltage scheme," *IEEE J. Solid-State Circ.*, pp. 1772-1780, Nov. 1998.
[20] K. Usami and M. Horowitz, "Cluster voltage scaling technique for low power design," *Proc. ISLPED*, pp. 3-8, 1995.
[21] C. Akrout, *et al*, "A 480MHz RISC microprocessor in a 0.12µm L_{eff} CMOS technology with copper interconnects," *IEEE J. Solid-State Circ.*, pp. 1609-1616, Nov. 1998.
[22] S. Sirichotiyakul, *et al*., "Standby power minimization through simultaneous threshold voltage and circuit sizing," *Proc. DAC*, pp. 436-441, 1999.
[23] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, pp. 23-29, Jul-Aug 1999.
[24] R. Chau, *et al*., "30nm physical gate length CMOS transistors with 1.0ps NMOS and 1.7ps PMOS gate delays," *Proc. IEDM*, pp. 45-48, 2000.
[25] S. Song, *et al*., "CMOS device scaling beyond 100nm," *Proc. IEDM*, pp. 235-238, 2000.
[26] H. Wakabayashi, *et al*., "45-nm gate length CMOS technology and beyond using steep halo," *Proc. IEDM*, pp. 49-52, 2000.
[27] M.Mehrotra, *et al*., "A 1.2V, sub-0.09µm gate length CMOS technology," *Proc. IEDM*, pp. 419-422, 1999.
[28] I.Y. Yang, *et al*., "Sub-60nm physical gate length SOI CMOS," *Proc. IEDM*, pp. 431-434, 1999.
[29] A. Ono, *et al*., "A 70nm gate length CMOS technology with 1.0V operation," *VLSI Symp. Tech*., pp. 14-15, 2000.
[30] M. Rodder, *et al*., "A scaled 1.8V, 0.18µm gate length CMOS technology: device design and reliability considerations," *Proc IEDM*, pp. 415-418, 1995.
[31] L. Su, *et al*, "A high-performance 0.08µm CMOS," *VLSI Symp. Tech*., pp. 12-13, 1996.
[32] K. Chen and C. Hu, "Performance and V_{dd} scaling in deep submicrometer CMOS," *IEEE J. Solid-State Circ.*, pp. 1586-1589, Oct. 1998.
[33] C. Hu, "Device and technology impact on low power electronics," in *Low Power Design Methodologies*, ed. Jan Rabaey, Kluwer, pp. 21-35, 1996.
[34] S. Mutoh, *et al*., "1V Multi-Threshold CMOS DSP with an efficient power management technique for mobile phone application", *Proc. ISSCC*, pp. 168-169, 1996.
[35] J.T. Kao and A.P. Chandrakasan, "Dual-threshold voltage techniques for low-power digital circuits," *IEEE J. Solid-State Circ.*, pp. 1009-1018, Jul. 2000.
[36] T. Kuroda, *et al*., "A 0.9V, 150MHz, 10mW, 4mm^2, 2-DCT core processor with variable V_T scheme," *IEEE J. Solid-State Circ.*, pp. 1770-1778, Nov. 1996.
[37] H. Kawaguchi, *et al*., "A CMOS scheme for 0.5V supply voltage with pico-ampere standby current," *Proc. ISSCC*, pp. 192-193, 1998.
[38] M.C. Johnson, *et al*., "Leakage control with efficient use of transistor stacks in single threshold CMOS," *Proc. DAC*, pp. 442-445, 1999.
[39] L. Wei, *et al*., "Design and optimization of dual-threshold circuits for low-voltage low-power applications," *IEEE T. VLSI Sys*, pp. 16-24, Mar. 1999.
[40] S. Tyagi, *et al*., "A 130nm generation logic technology featuring 70nm transistors, dual-V_t transistors and 6 layers of Cu interconnects," *Proc. IEDM*, pp. 567-570, 2000.
[41] http://www-device.eecs.berkeley.edu/~dennis/BACPAC,
see also: http://vlsicad.cs.ucla.edu/GSRC/GTX
[42] J.M. Musicer and J. Rabaey, "MOS current mode logic for low power, low noise CORDIC computation in mixed-signal environments," *Proc. ISLPED*, pp. 102-107, 2000.