Sub-90nm Technologies--Challenges and Opportunities for CAD

Tanay Karnik, Shekhar Borkar, Vivek De
Circuit Research, Intel Labs, Hillsboro, OR 97124.
tanay.karnik@intel.com

ABSTRACT

Future high performance microprocessor design with technology scaling beyond 90nm will pose two major challenges: (1) energy and power, and (2) parameter variations. Design practice will have to change from deterministic design to probabilistic and statistical design. This paper discusses circuit techniques and design automation opportunities to overcome the challenges.

1. INTRODUCTION

We are encountering several challenges in maintaining historical rates of performance improvement and energy reduction with CMOS technology scaling as we enter the sub-90nm technology generation. Excessive subthreshold and gate oxide leakage are emerging as serious problems. In addition, energy efficiency of the microarchitecture of general-purpose microprocessors is starting to play a more critical role in the performance vs. power and area trade-offs. The scaling issues are discussed in Section 2, followed by Section 3 describing required design automation solution from EDA tools, and the conclusion in Section 4.

2. SCALING TRENDS & DESIGN ISSUES

As technology scales beyond 90nm, transistor density will continue to double, allowing higher integration. Transistor delay will also continue to improve, at least modestly to 30% reduction per generation. However, power dissipation, delivery, density, and parameter variations will prohibit to take advantage of the performance and integration capacity.

2.1 Technology Scaling

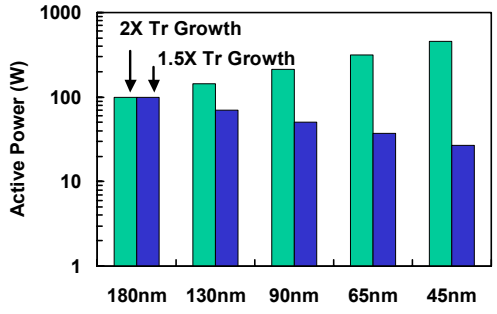

Fig. 1: Active Power Trend

Supply voltage (Vcc) will continue to scale modestly by 15%, not by the historic 30% per generation, due to (1) difficulties in scaling threshold voltage (Vt), and (2) to meet transistor performance goals.

Fig. 1 shows growth in active power of a microprocessor assuming historical 2X growth in number of transistors and with hypothetical 1.5X growth. Clearly, following historic trend would push the active power way over the power envelope, limiting transistor growth, reducing integration capacity, and the die size over technology generations.

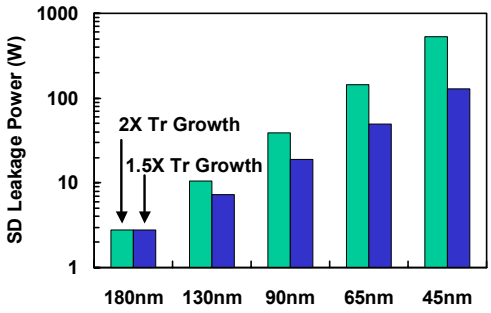

Fig. 2: SD Leakage Power Trend

Vt will continue to reduce modestly to meet the transistor performance demand, increasing source-drain subthreshold (SD) leakage. Fig. 2 projects SD leakage power of the microprocessor with 2X and 1.5X transistor growth. Notice that even with modest reduction in Vt, the SD leakage power will increase substantially, questioning viability of even 1.5X increase in transistors each generation, and suggesting even less integration of transistors, and smaller die size.

Design parameter variations will play even an important role in the chips designed beyond 90nm. Fig. 3 plots frequency and standby leakage current (I_{sb}) of microprocessors in a wafer. The spread in standby current is due to variation in channel lengths causing variations in the threshold voltage. Notice that the highest frequency chips have a wide distribution of leakage, and for a given leakage, there is a wide distribution in the frequency of the chips. The highest frequency chips with large I_{sb}, and low frequency chips with reasonably high I_{sb} will have to be discarded, affecting the yield.

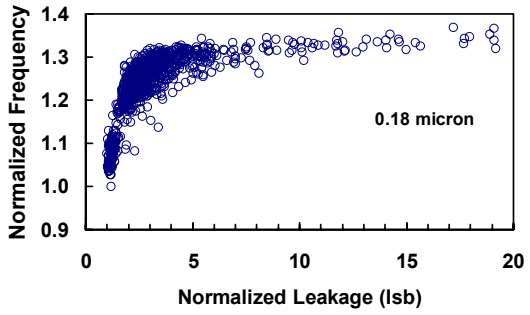

Fig. 3: Frequency & Standby Leakage Distribution

Variations in switching activity across the die, and diversity of the type of logic, results in uneven power dissipation across the die, as shown in Fig. 4 [1]. This variation results in uneven supply voltage distribution, temperature hot spots, and thus variation in subthreshold leakage across the die. Therefore, it

0-7803-7607-2/02/$17.00 ©2002 IEEE

will be important to design with parameter variations in mind, changing the design style from today's deterministic design to probabilistic and statistical design.

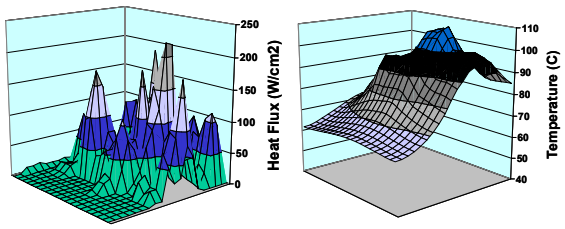

Fig. 4: Power Density & Temperature Variation

2.2 Switching Power Reduction

Active power density of memory, such as on-chip SRAM cache, is an order of magnitude smaller than logic (Fig. 5). Therefore, the overall processor performance can be improved in a more energy-efficient manner by using more memory than logic [1]. This effectively reduces the overall activity factor of the chip. Future microarchitectures will use even bigger on-die caches to deliver higher performance within a tight power budget.

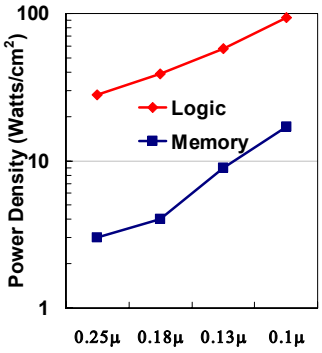

Fig. 5: Power density of memory vs. logic

Improving single-thread performance of general-purpose processors has historically required lots of logic transistors to exploit instruction level parallelism. For improving performance by 40%, the number of logic transistors has to be doubled (Fig. 6). This is quite inefficient use of transistor resources and has been considered appropriate for the amount of flexibility obtained. However, with the power constraints emerging as paramount, future microarchitectures will incorporate special-purpose functionality to improve the benchmark performance in a more area and energy-efficient manner (Fig. 7) [2].

Dual-Vcc designs [8,9] will be needed in the future to reduce switching power as well as leakage power (both subthreshold and gate oxide leakages) without impacting overall performance. Latency-critical units will use high-Vcc and non-critical units will use low-Vcc. Furthermore, throughput-oriented special purpose functional blocks can use low-Vcc to reduce power and recover performance loss by hardware replication to maintain

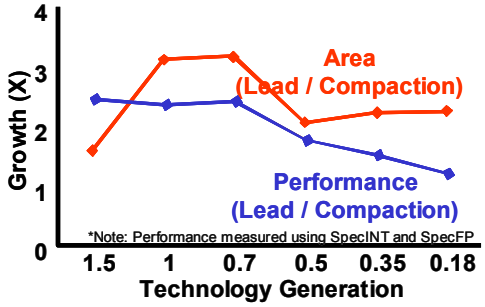

Fig. 6: Microarchitecture efficiency – Pollack's Rule

	Die Area	Power	Performance
General Purpose	**2X**	**2X**	**~1.4X**
Special Purpose	**<10%**	**<10%**	**1.5-4X**

Fig. 7: Special-purpose hardware efficiency

throughput. Transistors will be available in plenty in future technology generations. Thus hardware replication offers an attractive means of reducing power consumption [2]. Deeper pipelining also causes the clock power to become a more dominant component of processor power. Low swing clock distribution to reduce clock power will be useful, provided that jitter and skew issues at low Vcc can be managed [9]. Efficient level converters and proper management of noise injection from high-Vcc to low-Vcc domains will be critical.

2.3 Leakage Power Control

Dual-Vt designs [3,4] can reduce leakage power during active operation, burn-in and standby. Two Vt's are provided by the process technology for each transistor. Performance-critical transistors are made low-Vt to provide the target chip performance. Since the full-chip frequency is dictated by only a fraction of transistors in the critical paths, the selective Vt assignment is possible without degrading overall chip performance achievable by using a single low-Vt transistor everywhere. Fig 8 shows an example circuit block, where all low Vt design provides 24% delay improvement over all high Vt design. Notice that as you start inserting low Vt devices (Y axis), the delay improves (X axis). Only 34% of the total transistor width needs to be low-Vt. Typically, low-Vt device leakage is 10X higher than high-Vt. Thus, by carefully employing low-Vt up to 34% of the total width, 24% delay improvement is possible with only ~3X increase in leakage.

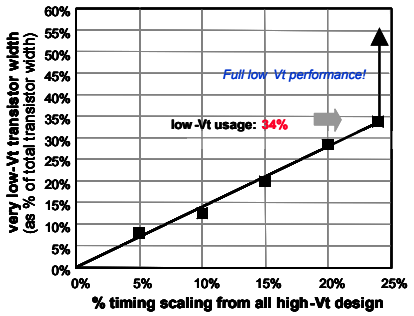

Fig. 8: Performance vs. leakage in dual-Vt designs

Another technique to reduce leakage power during burn-in and standby is to apply reverse body bias (RBB) to the transistors to increase [2] Vt since high performance is not required during these modes. There is an optimal RBB value that minimizes leakage power. Using RBB values larger than this value causes the junction leakage current to increase and overall leakage power to go up. In sub-90nm technology generation, approximately 500mV RBB is optimal [2]. 2-3X reduction in leakage current is achievable. However, effectiveness of RBB reduces as channel lengths become smaller or Vt values are lowered (Fig. 9). Essentially, the Vt-modulation capability by RBB weakens as short-channel effects become worse or body effect diminishes due to lower channel doping. Therefore, RBB becomes less effective with technology scaling and as leakage currents are pushed higher by shorter L or lower Vt.

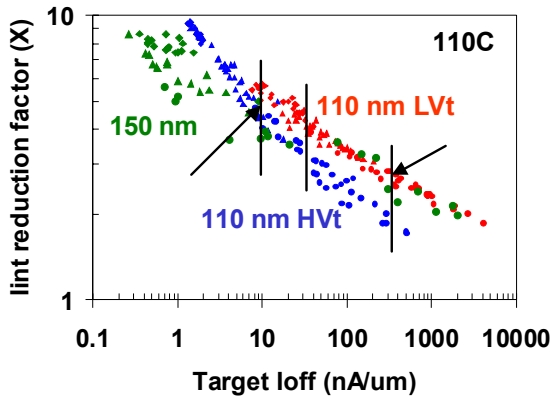

Fig 9: Subthreshold leakage reduction by reverse body bias

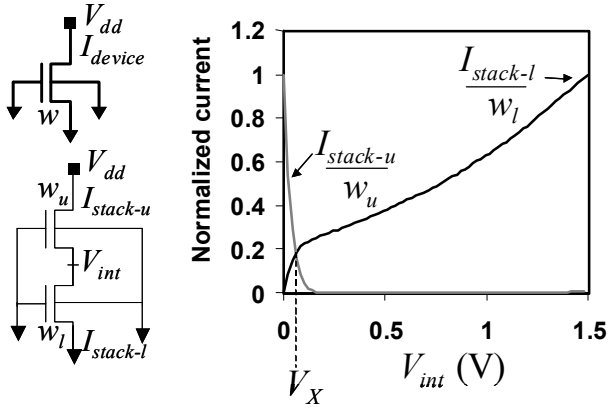

Fig. 10: Leakage current of transistor stacks – stack effect

Leakage current through series-connected transistors or transistor "stacks", with more than one device "off", is at least an order of magnitude smaller than that through a single device (Fig. 10). This so-called "stack effect" can be exploited for leakage reduction in circuits. The stack effect factor, defined as the ratio of single device leakage to stack leakage, increases as the DIBL factor becomes larger and supply voltage increases. As the rate of supply voltage scaling diminishes and DIBL effects become stronger with technology scaling, the effectiveness of leakage reduction by stacks becomes higher.

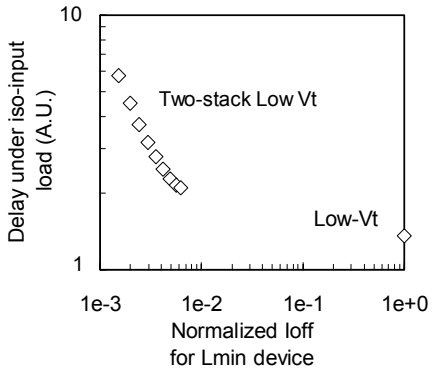

Fig. 11: Leakage vs. delay trade-offs by stack forcing

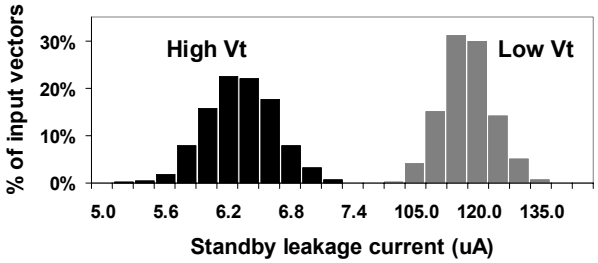

Fig. 12: Leakage control by natural stacks

Leakage vs. delay trade-offs offered by "stack forcing" are compared with similar trade-offs achievable by increasing transistor channel lengths. Increasing transistor length reduces leakage because of Vt roll-off and width reduction mandated by preserving the original input capacitance. In sub-90nm technology, where halo doping is used, reverse Vt roll-off is typically observed for channel lengths higher than nominal. Furthermore, 2D potential distribution effects dictate that doubling the channel length is less effective for leakage reduction than stacking two transistors, especially when DIBL is high. Simulation results show that channel length has to be made 3 times as large to get the same leakage as a stack of two transistors, resulting in 60% worse delay. Clearly, then "stack forcing" for leakage control is preferred.

Typically large circuit blocks contain some series-connected devices in complex logic gates. These so-called "natural stacks" can be exploited to reduce standby leakage. Leakage power of a large circuit block, such as a 32-bit static CMOS Kogge-Stone adder, depends strongly on the primary input vector (Fig. 12). The total "off" device width and the number of transistor stacks with two or more "off" devices change as primary input vectors change. This causes the leakage power to vary with input vector. When a circuit block is "idle", one can store the input vector that provides least amount of leakage at the primary input flops. This can reduce standby leakage power by 2X. There is no performance overhead since this pre-determined input vector can be encoded in the feedback path of the input flip-flop. The minimum time required in standby mode, so that the energy overhead for entry and exit into this mode is less than 10% of the leakage energy saved, is 10's of μS. This time reduces further with technology scaling as leakage levels increase, making this technique more attractive.

3. CHALLENGES FOR CAD

We must find alternate energy efficient microarchitectures to continue to deliver higher performance. The current simulators support trace-driven or execution-based approaches, however, they lack the support for a full system view, including platform, multiple cores/processors, OS, interrupts, etc. Applications will have to lend themselves to incorporate thread-level parallelism, followed by multi-processing to deliver near-linear performance with power. New memory design aids are also required to enable large on-die caches to continue to deliver higher performance.

Active power and performance are proven to be conflicting objectives. Total power, including leakage power, should be included in power-performance tradeoff tools. EDA tools exist for near-optimal device sizing for large circuits. Very few attempts have been reported for a truly integrated dual Vt allocation and device sizing tradeoff [3-6]]. One should also not miss the active power contributed by low Vt devices. Leakage reduction by stack effect can be exploited by automatically converting a single transistor to a two-transistor stack in a logic circuit, subject to performance constraints. EDA tools will need to identify the "lowest leakage" input vector efficiently during standby for each circuit block.

Supply noise includes a broad frequency spectrum. There is a need for exact power supply noise models and block-level time domain current signature estimation tools. The synthesis tools can reduce global instantaneous current demand by staggering the activity on a die. Intentionally skewed clock domains on the die will also suppress inductive noise. High current low voltage circuits need supply noise characterization tools. Package inductance does not scale enough to compensate for the increase in the instantaneous current. Effective placement of decoupling capacitors and switched capacitor circuits can provide a local source for the large instantaneous current and reduce the inductive noise on power lines. Placement tools need to automatically insert high frequency decoupling capacitors around high current consumption circuits, such as clock drivers, floating point units, etc. Multi-Vcc operation proposed in Section 2 is not a fine grain process-enabled solution like multi-Vt. Interleaved multi-Vcc routing incurs 20% power routing area penalty [9]. Physical separation of Vcc domains is a necessity for multi-Vcc circuits. Logical separation in the netlists manifests in physical separation on the die. Synthesis tools can enable multi-Vcc logical separation for effective floorplanning.

To achieve high performance without short channel effects, Vt can be lowered by forward body bias (FBB) [10]. A technique to reduce leakage power during burn-in and standby is to apply reverse body bias (RBB). Physical design can enable body bias by routing body signal as one more low current power signal. Additionally, floorplanning tools can reserve die areas for global and local body bias signal generators. Standby leakage power can be minimized by inserting a high Vt device, called sleep transistor, in series to normal low Vt circuitry (Fig 13). The sleep transistor is controlled by a special signal to specify active/standby mode. Proper sizing and sharing of sleep transistors across large active circuit blocks needs to be planned during the entire physical design process [7].

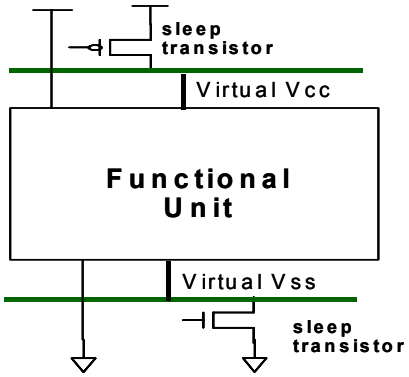

Fig. 13: Sleep transistor technique

Chip manufacturers typically interact with multiple CAD vendors. EDA community should provide unified ASCII data models, industry-wide object-oriented unified framework and interoperability of tools to minimize the burden on DA teams.

4. CONCLUSIONS

This paper has described CMOS scaling challenges for sub-90nm designs and CAD challenges to support energy, parameter variation, and microarchitectural challenges. CMOS is it, for now, and for the foreseeable future. Therefore, design practice will have to change from today's deterministic design to probabilistic and statistical design. We have discussed several circuit techniques and design automation and CAD opportunities to overcome these challenges.

5. REFERENCES

[1] Pollack F., New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies; Micro32, 1999.

[2] Sery G., et al., Life is CMOS: why chase the life after? DAC 2002, 78-83.

[3] Karnik, T., et al., "Total power optimization by simultaneous dual-Vt allocation and device sizing in high performance microprocessors", DAC 2002, pp. 486-491.

[4] Tschanz, J., et al., "Design optimizations of a high performance microprocessor using combinations of dual-Vt allocation and transistor sizing", VLSI Circuits Symposium 2001, pp. 218-219.

[5] Pant, P., et al., Dual-Threshold Voltage Assignment with Transistor Sizing for Low Power CMOS Circuits. TVLSI, Vol. 9, No. 2, 4, 2001, pp. 390-394.

[6] Wei, L., et al., Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications. IEEE TVLSI, Vol. 7, No. 1, March 1999, pp. 16-23.

[7] Anis, M., et al., Dynamic and leakage power reduction in MTCMOS circuits using an automated efficient gate clustering technique. DAC 2002, pp. 480-485.

[8] Kuroda, T., et al., Low-power CMOS digital design with dual embedded adaptive power supplies. JSSC, Vol. 35, Issue 4, April 2000, pp. 652-655.

[9] Krishnamurthy, R., et al., High-performance and low-power challenges for sub-70nm microprocessor circuits. CICC 2002, pp. 125-128.

[10] Keshavarzi, A., et al., Forward body bias for uPs in 130nm technology generation and beyond. VLSI Circuits Symp. 2002, pp. 125-128.