Life Is CMOS: Why Chase the Life After?

George Sery, Shekhar Borkar, Vivek De
Intel Corporation
RN3-21, 2200 Mission College Boulevard
Santa Clara, CA 95052-8119
001-408-765-9367
george.sery@intel.com

ABSTRACT
This paper discusses potential solutions to the CMOS device technology scaling at gate lengths approaching 10nm. Promising circuit and design techniques to control leakage power are described. Energy-efficient microarchitecture trends for general-purpose microprocessors are elucidated.

Categories and Subject Descriptors
B.7 INTEGRATED CIRCUITS
B.7.1 Types and Design Styles – *Microprocessors and microcomputers, VLSI.*

General Terms
Performance, Design

Keywords
Technology scaling, Leakage control, Microarchitecture

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-100nm technology generation. Some of the key bottlenecks are related to reducing device parasitics such as source/drain resistances and gate overlap capacitances. Excessive subthreshold and gate oxide leakage are also 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. Potential solutions to the device technology scaling challenges at gate lengths approaching 10nm are discussed in Section 2. Section 3 describes some promising circuit and design techniques to control leakage power. Energy-efficient microarchitecture trends are elucidated in Section 4.

2. DEVICE TECHNOLOGY CHALLENGES
2.1 Performance and Energy Scaling
In the sub-200nm technology generation, it is difficult to maintain traditional constant-field supply voltage scaling (Figs. 1 & 2) due to the non-scalability of threshold voltage from excessive leakage current considerations. Essentially, the electric field across the gate dielectric has been increasing by 10% per generation. This has been made possible by the superior long-term reliability offered by physically thinner gate oxides. In order to improve the delay of driving constant capacitance loads such as those posed by interconnects in high performance microprocessor designs, device saturation current per unit width must remain constant or increase from one technology generation to the next. This has been accomplished by reducing the rate of supply voltage scaling, clearly at the expense of escalating switching power density.

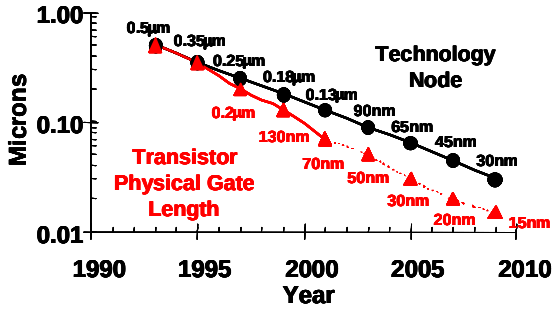

Fig. 1: Physical gate length scaling trend

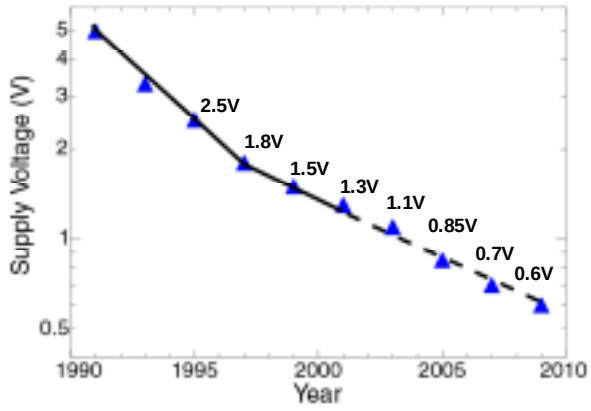

Fig. 2: Supply voltage scaling trend

Extrinsic source/drain resistance, gate overlap capacitance and junction capacitance do not scale in a desired fashion. This limits the circuit delay improvements achievable from large intrinsic device saturation currents. Reducing the depth of the source/drain junction extension or tip improves short-channel effects and thus

DAC 2002, June 10-14, 2002, New Orleans, Louisiana, USA.

allows a shorter gate length. However, tip depths less than 50nm causes the tip resistance to become so large that drive current becomes smaller (Fig. 3). Increasing the tip doping concentration beyond the solid solubility limit can help alleviate this problem to some extent. Furthermore, abruptness of the doping profiles, both in vertical and lateral directions, must be increased.

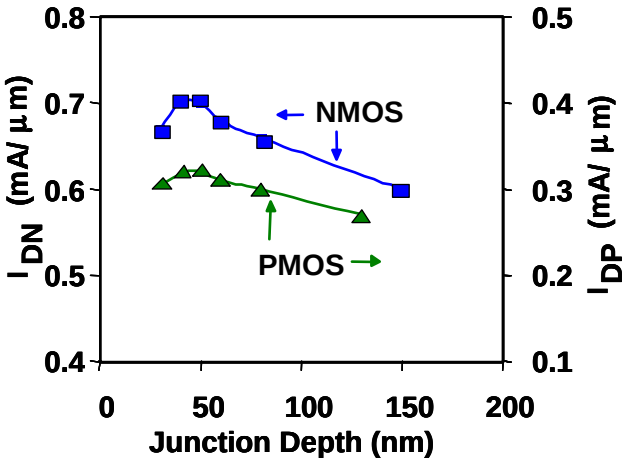

Fig. 3: Optimal source/drain extension depth

Spreading resistance from the inversion layer to the source/drain extension region also limits drive current. This makes it difficult to scale the gate overlap length below 10nm. In addition, increasing channel doping, demanded by sub-surface punch through control and short-channel effect reduction, causes capacitance of the gate-edge junction sidewall to increase with technology scaling and degrades delays of wide-OR circuits such as bitlines in the cache. In spite of all these limitations, device CV/I delays well beyond a terahertz is achievable at sub-1V supply voltages using a traditional planar bulk CMOS device structure in the 15-30nm gate length regime (Figs. 4 & 5).

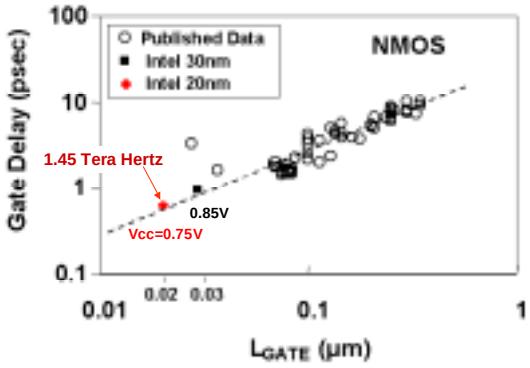

Fig. 4: CV/I delay scaling trends for bulk CMOS

A fully-depleted SOI device structure, referred to as the Depleted Substrate Transistor (DST), is promising for alleviating many of the challenges discussed before (Fig. 6). The subthreshold swing is much steeper compared to either bulk or partially depleted SOI devices when the silicon film thickness is below 30nm, resulting in a fully depleted channel. This allows Vt to be reduced for a specific leakage target and boosts drive current. Furthermore, the oxide layer below the silicon channel completely eliminates sub-surface punch through and junction leakage currents. Therefore, channel doping can be reduced. This reduces the gate-edge junction sidewall capacitance dramatically.

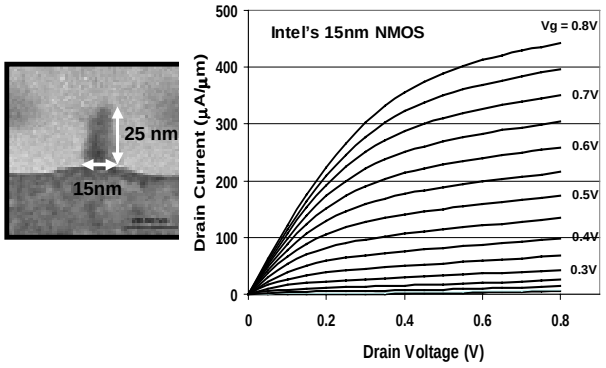

Fig. 5: Intel's 15nm bulk NMOS transistor

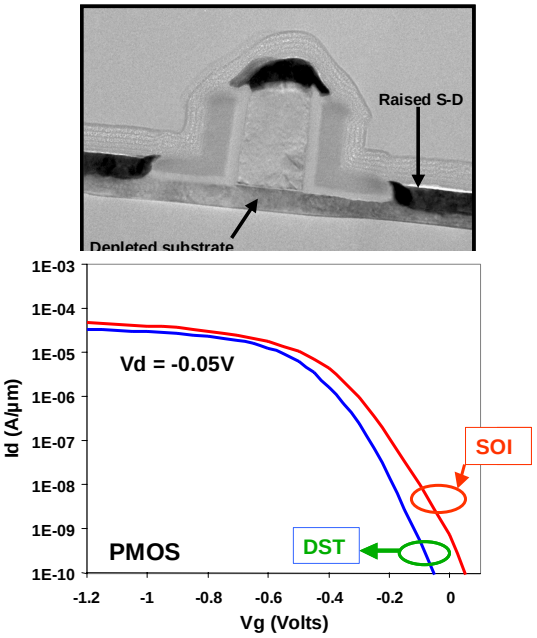

Fig. 6: Depleted substrate transistor (DST) with raised S/D

The source/drain extension depth in DST can be scaled by simply scaling the silicon thickness to improve short-channel effects. The buried oxide layer also serves as a diffusion stopper and creates more abrupt vertical doping profiles in the source/drain region. When combined with a raised source/drain structure, the drive current improvement due to lower parasitic resistance is as much as 30%. The main challenge associated with further development of DST with conventional polysilicon gates is achieving a sufficiently tight control of the silicon film thickness since the threshold voltage is quite sensitive to the film thickness. This problem can be alleviated to some extent by migration to a metal gate electrode whose work function is chosen appropriately to provide the appropriate Vt. Two different metals may be needed for NMOS and PMOS. When the gate length is pushed to the DIBL limit, threshold voltage sensitivity to variations in silicon film thickness will still need to be dealt with. In any case, DST provides a promising scaling path to sub-20nm technology generation.

2.2 Subthreshold and Gate Oxide Leakages

Subthreshold leakage current of a transistor is increasing by ~5X per generation. At high temperature, it exceeds 1000nA/um in sub-100nm technology nodes (Fig. 7). As the physical gate oxide thickness approaches sub-10Å regime, gate oxide leakage becomes larger than 100A/cm^2 (Fig. 8) due to direct band-to-band tunneling. Although gate oxide leakage increases weakly with temperature, it accelerates exponentially with increase in supply voltage at a rate of 2X larger leakage for every 100mV increase in voltage. Junction leakage is an additional component of concern since it is increasingly dictated by tunneling as channel doping concentrations approach 5X10^{18} cm^{-3} in the channel.

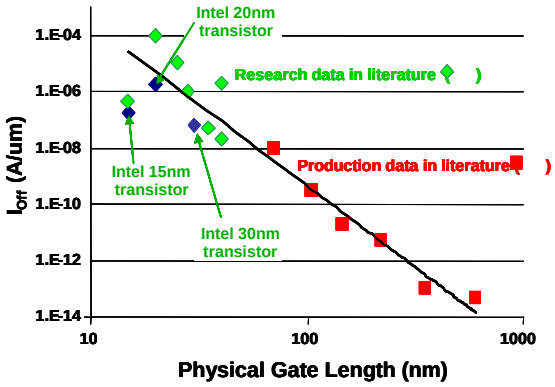

Fig. 7: Subthreshold leakage scaling trend

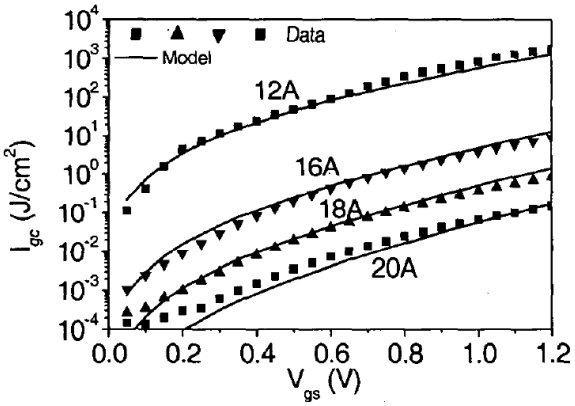

Fig. 8: Gate oxide leakage scaling with thickness and voltage

There is a compelling need to use high-K dielectrics as the gate dielectric material to replace silicon dioxide or oxynitride. High-K dielectric materials would allow electrical thickness of the gate dielectric to be scaled to provide large capacitance per unit area, while keeping the tunneling leakage per unit area within acceptable limits due to larger physical thickness. Scaling of electrical oxide thickness is essential to provide sharp subthreshold swing, large drive current and control short-channel effects. Characteristics of several candidate high-K dielectrics are compared in Fig. 9. HfO_2 and ZrO_2 provide the smallest gate leakage, 2-3 orders of magnitude smaller than oxide, for a target electrical thickness in the sub-10Å regime. Since the bandgap reduces with increasing permittivitty, gate leakage due to thermal emission dominates for materials with very high K values. Thus, Ta_2O_5 is not as attractive for this application. Of course, many process integration challenges need to be resolved and silicon-dielectric interface quality needs to be improved for these new dielectric materials to provide improvements to CMOS circuit delay.

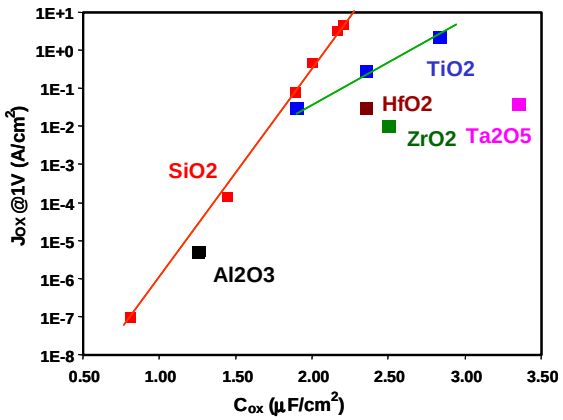

Fig. 9: Capacitance and leakage of high-K dielectric materials

The scaling of electrical gate oxide thickness is limited by poly depletion and separation of inversion layer charge from the oxide-silicon interface at high vertical fields due to quantum-mechanical (QM) effects (Fig. 10). Each of these effects adds 5Å to the effective electrical oxide thickness at the highest gate voltage. These effects become more significant as gate voltage increases. Thus, when averaged over the entire gate voltage range from Vt to the maximum supply voltage, their impacts on drive current are less severe. Nevertheless, in order to maximize the benefit of migrating to a high-K gate dielectric, poly depletion should be reduced or eliminated. Increasing poly doping beyond the solid solubility limit is desirable. Transition to a metal gate fully eliminates poly depletion. But metal gates with appropriate work functions for NMOS and PMOS must be identified and process integration issues must be resolved. Combining metal gates with DST to set the threshold voltage by work function engineering is a promising approach that also addresses the poly depletion problem. However, additional process complexities due to two different gate electrode metals, one for NMOS and one for PMOS, will be incurred.

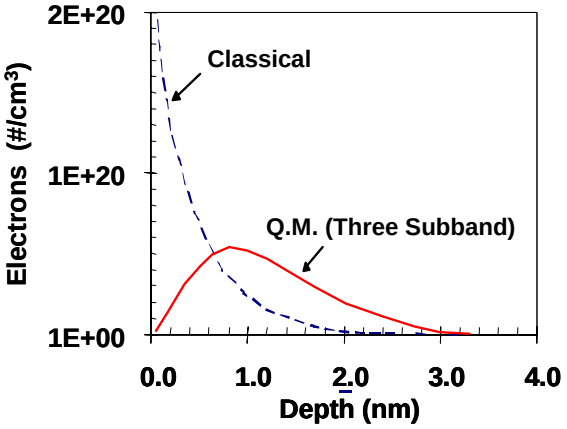

Fig. 10: Increase in electrical oxide thickness by QM effects

3. LEAKAGE CONTROL TECHNIQUES

Leakage power is becoming a larger fraction of the total active power of microprocessors (Fig. 11). This poses serious challenges for heat removal and power delivery in high performance processors. Excessive leakage power can also cause thermal runaway during burn-in, and impact burn-in cost. Subthreshold leakage dominates at high temperature and gate oxide leakage is a significant contributor to the burn-in leakage power due to the higher voltage used. Of course, standby leakage power at room temperature also needs to be kept sufficiently small for battery-operated systems.

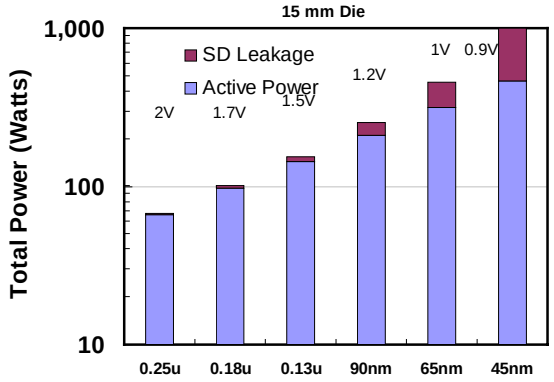

Fig. 11: Switching and leakage power scaling trends

3.1 Dual-Vt and Body Bias

Dual-Vt designs 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. Rest of the transistors are made high-Vt to minimize leakage power. Since the full-chip frequency is dictated by only a fraction of transistors in the critical paths, this selective Vt assignment is possible without degrading overall chip performance achievable by using a single low-Vt transistor everywhere. Fig 12 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 in this example, to get the same frequency as using low-Vt everywhere. 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 ~3X increase in leakage, compared to all high-Vt design.

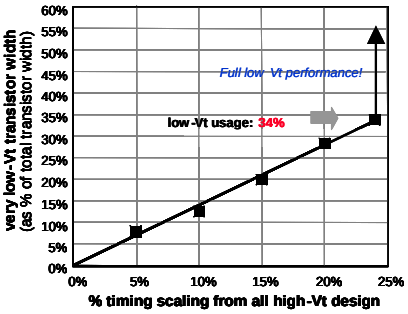

Fig. 12: Performance vs. leakage in dual-Vt designs

Another technique to reduce leakage power during burn-in and standby is to apply reverse body bias (RBB) to the transistors to increase Vt since high performance is not required during these modes. There is an optimal reverse body bias value that minimizes leakage power. Using reverse body bias values larger than this value causes the junction leakage current to increase and overall leakage power to go up. In sub-100nm technology generation, approximately 500mV RBB is optimal. 2-3X reduction in leakage current is achievable. However, effectiveness of RBB reduces as channel lengths become smaller or Vt values are lowered (Fig. 13). 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.

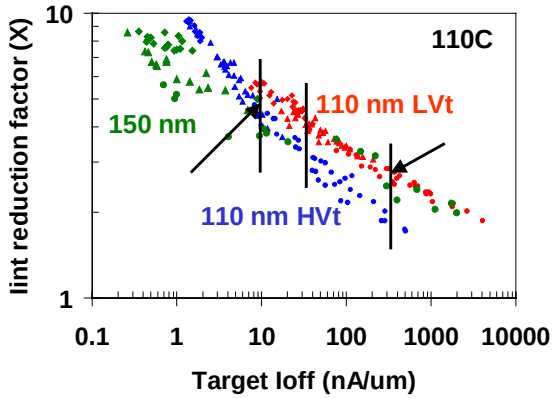

Fig. 13: Subthreshold leakage reduction by reverse body bias

3.2 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. 14). 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. 15).

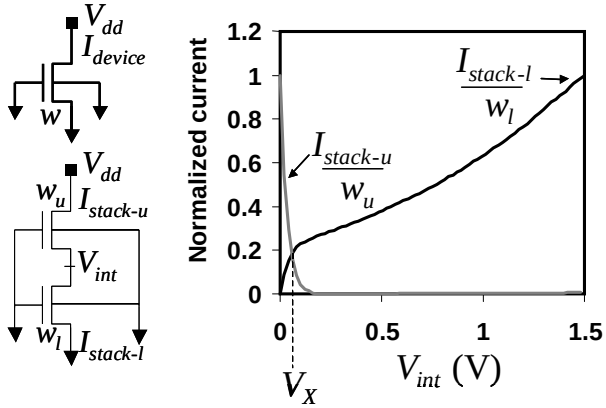

Fig. 14: Leakage current of transistor stacks – stack effect

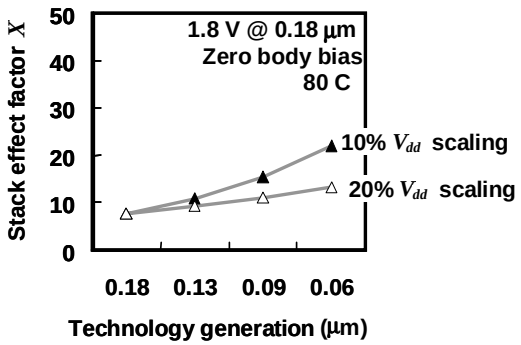

Fig. 15: Scaling of stack effect factor

Leakage reduction by stack effect can be exploited by converting a single transistor to a two-transistor stack in a logic circuit. The widths of these transistors can be half of the original size or other combinations can be chosen to preserve same input capacitance load as the original single device. Leakage vs. delay trade-off provided by this "stack forcing" technique applied to both high-Vt and low-Vt devices is illustrated in Fig. 16. Clearly, stack forcing can be used to emulate additional higher Vt devices without increasing process complexity. Stack forcing can be applied to transistors in non-critical paths in single-Vt or dual-Vt designs to reduce overall chip leakage power without impacting chip performance. Also, robustness of leakage-sensitive circuits can be improved by this technique.

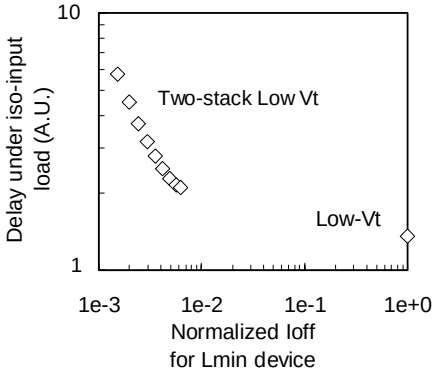

Fig. 16: Leakage vs. delay trade-offs by stack forcing

Leakage vs. delay trade-offs offered by stack forcing are compared with similar trade-offs achievable by increasing transistor channel lengths (Fig. 17). Increasing transistor length reduces leakage because of threshold roll-off and width reduction mandated by preserving the original input capacitance. In sub-100nm technology, where halo doping is used, reverse Vt roll-off is typically observed for channel lengths higher than nominal. Furthermore, two-dimensional 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 confirm this behavior and 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 if the channel length needs to be more than doubled to achieve the target low leakage.

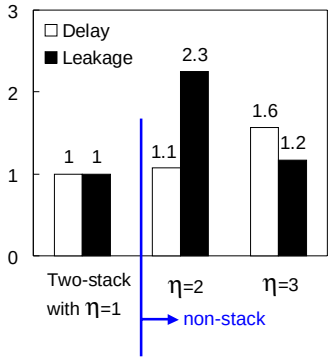

Fig. 17: Stack forcing vs. longer channel length

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. 18). 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 uS. This time reduces further with technology scaling as leakage levels increase, making this technique more attractive. Of course, EDA tools will be needed to identify this "lowest leakage" input vector efficiently during design phase for each circuit block.

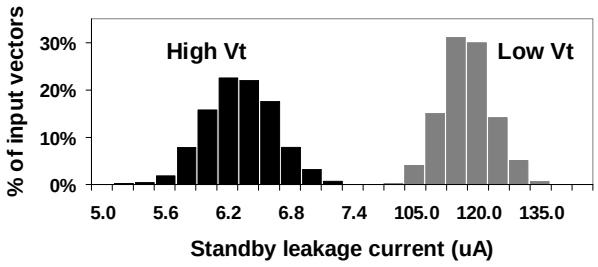

Fig. 18: Leakage control by natural stacks

4. MICROARCHITECTURE TRENDS

To evaluate effectiveness of microarchitecture in delivering higher performance, consider Pollack's rule [4]. Fig 19 plots growth in performance of a new and an old microarchitecture in the same process technology, and growth in the area to implement them. Notice that on an average a 2X growth in area provides only 1.4X increase in the performance—a square law. This shows that traditional microarchitectures, exploiting instruction level parallelism, have not been power efficient in delivering performance.

This is further elaborated in Fig 20, which shows estimated increase in die area, performance, and power due to microarchitecture advances such as super-scalar, dynamic, and netburst. The growth in area and power reflects growth in the number of transistors, and power hungry circuit styles employed for implementation. Notice that each advance has consumed about 2X power delivering 40% more performance. Therefore, we must find alternate energy efficient microarchitectures to continue to deliver higher performance.

Applications will have to lend themselves to incorporate thread-level parallelism, followed by multi-processing to deliver near-linear performance with power. Furthermore, certain application tasks could be easily served by special purpose hardware on the die tailored for the applications, and thus power efficient.

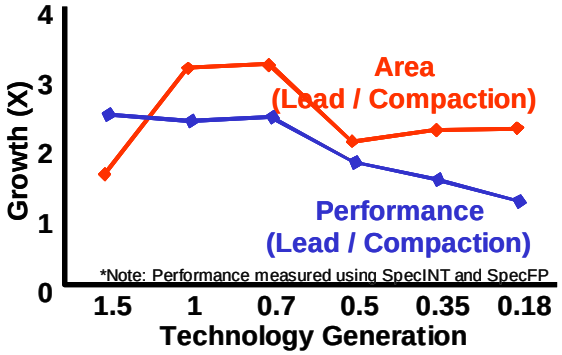

Fig. 19: Microarchitecture efficiency trends

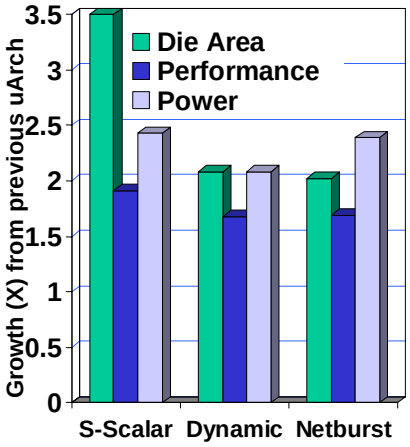

Fig. 20: Performance, power and area trade-offs of general purpose microarchitectures

Fig 21 compares estimated active power density of logic and static memory in a given process technology. Memory power density tends to be an order of magnitude lower than that of logic. This is because only a part of the memory is accessed at any given time. Also, memory transistors can withstand relatively higher threshold voltages, reducing the leakage power compared to logic. To make up for the loss of transistor performance, memory operations can be pipelined, with modest increase in latency.

Therefore future microarchitectures could exploit lower power density of memory to stay on the performance trend, and yet lower active and leakage power. The trend is already evident as shown in Fig 22, which plots cache memory transistors in microprocessors in several technology generations. Future microarchitectures will use even bigger caches to continue to deliver higher performance.

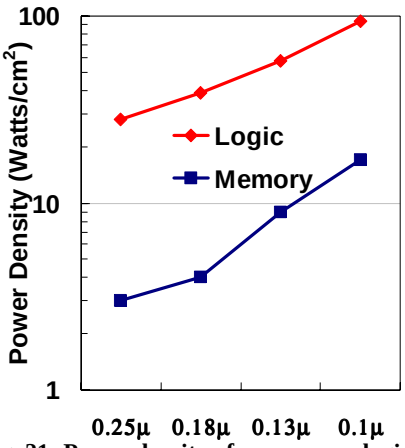

Fig. 21: Power density of memory vs. logic

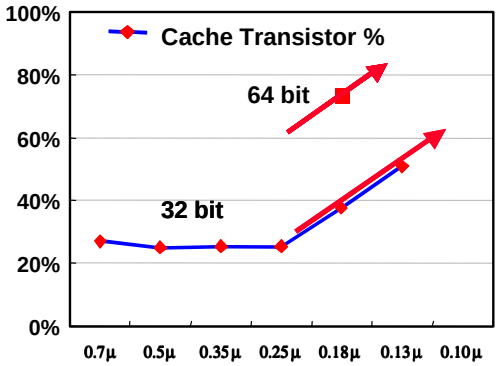

Fig. 22: On-chip memory integration trends

5. CONCLUSIONS

This paper has described CMOS scaling challenges for gate lengths approaching 10nm, and potential solutions in circuits and microarchitecture. These solutions may appear difficult, but are more mature and less risky than other proposed alternatives for CMOS. That is why CMOS is it, for now, and for the foreseeable future.

6. REFERENCES

[1] R. Chau et al., "30 nm physical gate length CMOS transistors with 1.0 ps n-MOS and 1.7 ps p-MOS gate delays", *IEDM 2000*, pp. 45-48.

[2] R. Chau, "30nm and 20nm Physical Gate Length CMOS Transistors", *Silicon Nanotechnology Workshop*, 2001.

[3] R. Chau et al., "A 50nm depleted-substrate CMOS transistor (DST)", *IEDM 2001*, pp. 621-624.

[4] Fred Pollack; New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies; *Micro32*, 1999.