Trends and Challenges in High Speed Microprocessor Design

Kerry Bernstein
IBM Microelectronics
Essex Junction, VT USA
Phone: (802) 769-6897 Fax: (802) 769-6744 Internet: kbernste@us.ibm.com

Abstract

Entropy is a worthy adversary! High performance logic design in next-generation CMOS lithography must address an increasing array of challenges in order to deliver superior performance, power consumption, reliability and cost. Technology scaling is reaching fundamental quantum-mechanical boundaries! This paper reviews example mechanisms which threaten deep submicron VLSI circuit design, such as tunneling, radiation-induced logic corruption, and on-chip delay variability. Architectures, circuit topologies, and device technologies under development are explored which extend "evolutionary" concepts and introduce "revolutionary" paradigms. It will be these revolutionary technologies which will bring our industry to the threshold of human compute capability.

Introduction

The overwhelming success of VLSI arises from the convergence of advances in multiple disciples: MOSFET device design, process development, innovative new circuit topologies, and power new state machine architectures. Each has consistently contributed opportunities for improving transaction throughput. So successful has been this progression, that limits in the design space must now be confronted. This progression, known as scaling[1], has provided benchmarks for each discipline, generation over generation. We first examine the scaling experience, look at example mechanisms limiting continued scaling, and then explore how designs have responded to these new capabilities and limitations. Finally, we will muse over the compute power continued scaling may enable.

Scaling Experience

Scaling refers to the practice of simultaneously reducing a collection of key electrical and physical design parameters by a constant value. Figure 1 shows the application of scale factor a to the physical dimensions of a MOSFET. Frank, etal describes how the retention of these relationships preserves device optimization[2]. This relationship has in fact been preserved, more or less, through multiple generations of CMOS Lithography, yield the performance trend shown in Figure 2. Also evident is the requirement more recently for the constant infusion of innovative structures and materials to sustain this improvement. The underlying engine driving this capability is photolithography. Smaller and smaller minimum critical dimensions have given rise to channel length reductions seen. speed, provided the other following boundary conditions are met.

Scaling Limitations

Nonetheless, even with innovation, "Moore's Law" has been observed to be eroding. A roll-off in device performance arises from the inability to scale threshold voltage as quickly as supply. This results in (V_{GS}-V_T) overdrive voltage reduction.

Process tolerance presents a second challenge to scaling. As critical parameter tolerance becomes harder to maintain at smaller lithography, the amount of timing margin consumed by the resulting delay variation impacts yield. Figure 3 shows the offset between the functionality window (defined by the dispersion of delay in paths of varying composition) and the full process tolerance window. It is evident that to maintain yield, performance must be sacrificed in the form of margin.

Aside from process, voltage and temperature variation across die also contribute to delay variability. Typically a design may present up to 3% performance change per 10degC in temperature change, or 5% performance change per 100 mV in supply voltage variation. A fourth challenge to scaling lies in its intrinsic response to radiation events. Alpha particles arising from semiconductor materials or high energy protons or neutron daughters of cosmic ray events both have the opportunity to deliver charge necessary to corrupt the content of a bistable, or to glitch a logic level. Figure 4 shows the steady decrease in QCRIT, the minimum charge necessary to induce an event, against scaling. As feature dimensions are reduced, the capacitance reservoir of charge balancing an event is also reduced.

A fifth challenge is associated with the integrity of gate dielectrics in the MOSFET. As dimensions sizes reduce, it is essential to reduce gate oxide thickness for the gate to retain control over the inversion layer formation. Thinner dielectrics have higher tunneling currents and more frequent breakdowns.

Design Response

To understand how designs have exploited this capability and address its emerging limitations, it is useful to examine the Patterson-Hennessy Formula[3] for performance contributions.

Time = Instr/pgm x Cycles/Instr x Seconds/Cycle
(1) (2) (3)

The first term is the responsibility of the compiler; improvement in the second term comes from architectural enhancements, which have been responsible for perhaps half of recent performance gains. Out of order execution, speculative branching, multi-threading, and superscalar functional units are examples. These features, while improving through-put, also add extra circuits and devices, increasing power consumption. Figure 5 shows this trend and its implausible trend. The third term falls squarely in the lap of the process and circuit designers. The lies, however, an even more insidious subtle trend (A"red-hat topic"!). Because more and more circuitry has been added to boost architecture performance, wire lengths have not reduced with scaling as die sizes have not shrunk. Worse, these additions create deeper pipelines with less intrinsic delay per pipeline. In short a signal has to go farther than before, and has even less time to get there than

before, both after scaling. The result is that less and less of the chip may be accessed in a given cycle, as shown in Figure 6. The design response as is "logical islands" are not defined during placement, considering which functions must be less than 1 cycle latency away.

Capabilities of the Extended Paradigm

To combat this trend the, high speed microprocessors require constant innovation. A denser device with lower parasitic capacitance and which puts out more current is one such innovation. The Strained Silicon MOSFET[4] is an evolutionary MOSFET improvement (Figure 7); it derives it's performance advantage from strain induced in the layer in which the inversion channel is formed. In the cited reference, a thin SiGe layer is deposited on a Si substrate. With different lattice constants, the resulting strain induces improved mobility in 2 of Silicon's 6 degenerate states. An architectural direction likely to improve throughput is increased parallelism. Figure 8 shows the results of an analysis of various means of achieving equivalent performance. It supports the conclusion that an array of smaller simpler processors run at lower voltages can meet the equivalent performance of fewer processors at high voltage, saving, power and design resource. Finally, new circuit topologies promise to help reduce cycle time. Figure 9 shows Clock-Delayed Domino, an emerging circuit family used in semi-synchronous and "locally-asynchronous-globally-synchronous" microprocessors. At its heart, the circuit is a simple dynamic domino, which has traveling along with it its own clock. The clock can serve one circuit or one time-sliced column of circuits. It's delay is tuned via the passgate beta ratio.

Conclusions

Technology scaling is a paradigm that has indeed served our industry well. It is directly as well as indirectly responsible for the historic performance, density, and power trend known as "Moore's Law." Most recently, quantum-mechanical limitations to scaling have become evident and have required compensation by the designer at the architecture as well as circuit topology level. Innovation in novel MOSFET design, new circuit families, and logic architectures will provide a path for evolution of existing approaches, and buy time to develop revolutionary concepts. Just as scaling up to now has enabled more function to be brought onboard chip with less latency, continued scaling will before long allow our industry to deliver transaction throughput rivaling human compute capability. It is incumbent, then, to wisely invest our physical as well as intellectual resources, to most fruitfully enjoy this future capability.

References

[1]B. Davari, etal., "CMOS Scaling for High Performance and Low Power - The Next 10 Years," *Proceedings of the IEEE*, Vol. 83, No.4, April, 1995, pp. 595-606.

[2] D. Frank, etal., "Generalized Scale Length for Two-Dimensional Effects in MOSFETs", *IEEE Electron Device Letters*, Vol. 19, No. 10, October 1998, pp 385-387.

[3]D. Patterson, etal., "Computer Architecture: A Quantitative Approach", Morgan Kaufmann Publishers, 1995, ISBN 1558603298

[4] K. Rim, etal., "Strained Si NMOSFETs for High Performance CMOS Technology", *Proceedings of 2001 VLSI Technology Symposium*, pp. 59-60.

[5] F. Pollack, "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies", *Micro32*.

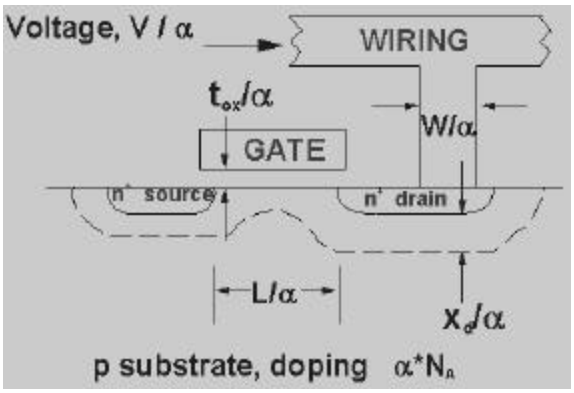

Figure 1: MOSFET Device ideal scaling relationships

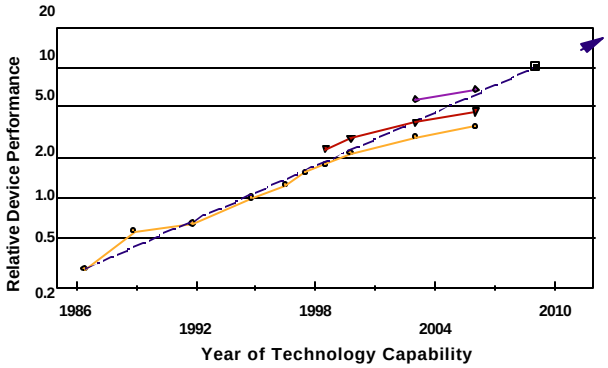

Figure 2: Innovation and Scaling

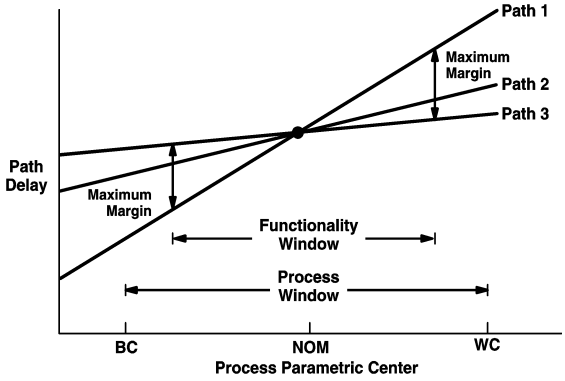

Figure 3: Timing Margin Consumption by Process Variation

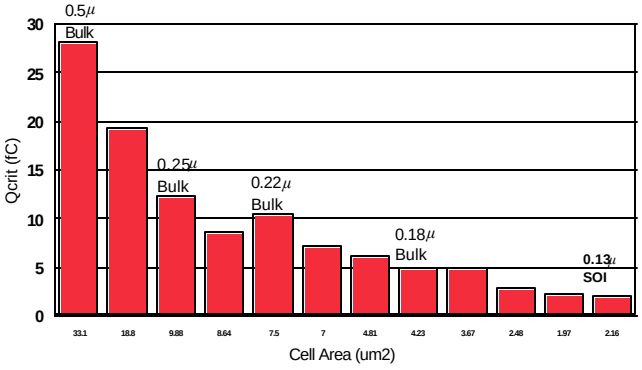

Figure 4: Channel hot electron degradation dependence on V_T.

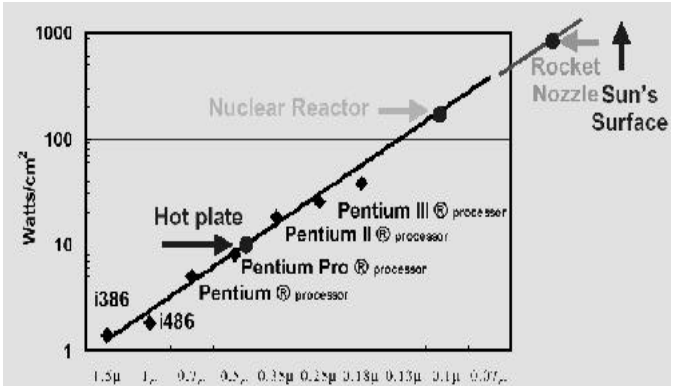

Figure 5: Trends in Power Consumption[5]

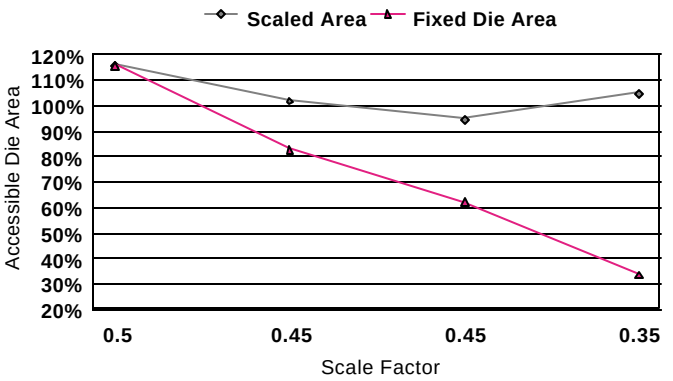

Figure 6: Area of Control [3]

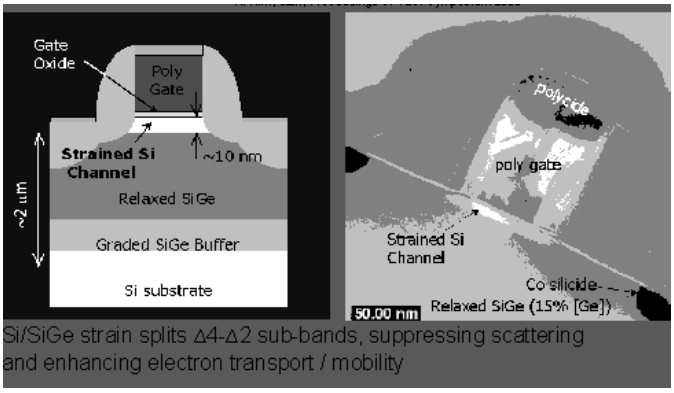

Figure 7: Strained Silicon MOSFET

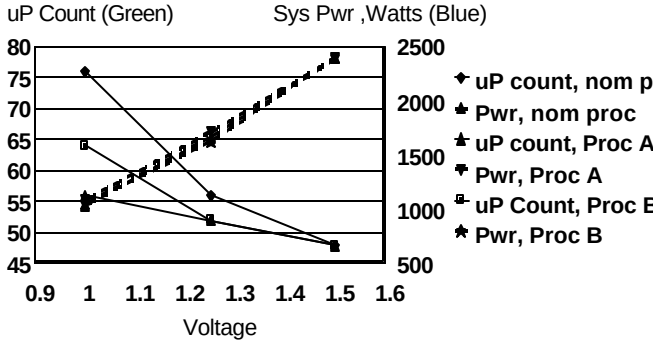

Figure 8: Distributed Processing's Power x Delay advantages.

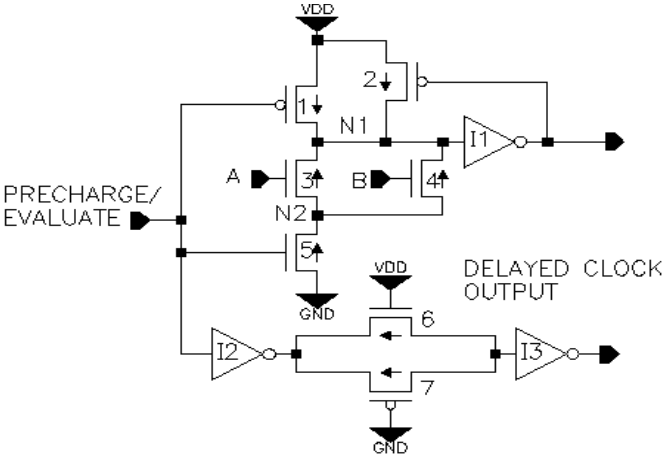

Figure 9: Clock-Delayed Dynamic Domino Logic