Parameter Variations and Impact on Circuits and Microarchitecture

Shekhar Borkar, Tanay Karnik, Siva Narendra, Jim Tschanz, Ali Keshavarzi, Vivek De
Circuit Research, Intel Labs, JF3-334, 2111 NE 25th Ave, Hillsboro, OR 97124.
shekhar.y.borkar@intel.com

ABSTRACT
Parameter variation in scaled technologies beyond 90nm will pose a major challenge for design of future high performance microprocessors. In this paper, we discuss process, voltage and temperature variations; and their impact on circuit and microarchitecture. Possible solutions to reduce the impact of parameter variations and to achieve higher frequency bins are also presented.

Categories and Subject Descriptors
B.7.1 *Microprocessors and microcomputers, VLSI.*

General Terms: Design, Performance, Reliability.

Keywords: Parameter variation, high performance deisgn, body bias.

1. INTRODUCTION
Systematic and random variations in process, supply voltage and temperature (P, V, T) are posing a major challenge to the future high performance microprocessor design [1,2]. Technology scaling beyond 90nm is causing higher levels of device parameter variations, which are changing the design problem from deterministic to probabilistic [3,4]. The demand for low power causes supply voltage scaling and hence making voltage variations a significant part of the overall challenge. Finally, the quest for growth in operating frequency has manifested in significantly high junction temperature and within die temperature variation. We discuss the impact of P, V, T variations on circuits and microarchitecture. We will also present possible solutions to reduce or tolerate the parameter variations in high frequency microprocessor designs.

The paper is organized as follows. Process, supply voltage and temperature variations are introduced in Section 2. The serious impact of these variations on circuits and microarchitecture is presented in Section 3. Section 4 consists of possible solutions to mitigate parameter variation including V_t modulation by forward body bias (FBB), leakage reduction by reverse body bias (RBB) or power/performance tradeoff by adaptive body bias (ABB). Section 4 also includes approaches to control supply voltage and temperature variations. We conclude this paper in Section 5.

DAC 2003, June 2-6, 2003, Anaheim, California, USA.

2. VARIATIONS – P, V, T
In this section, we will introduce the P, V, T variations based on measurement data in advanced CMOS technologies.

2.1 Process Variations
Figure 1 plots distributions of frequency and standby leakage current (I_{sb}) of microprocessors in a wafer. The spread in frequency and leakage distributions is due to variation in transistor parameters, causing about 20x variation in chip leakage and 30% variation in chip frequency. This variation in frequency has introduced the concept of frequency binning. 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} may have to be discarded, affecting the yield.

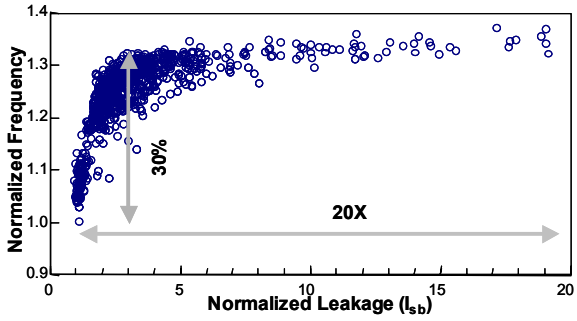


Figure 1: Leakage and frequency variations

The spread in standby current is due to variation in channel length and variations in the threshold voltage. Figure 2 illustrates the die-to-die V_t distribution and its resulting chip I_{sb} variation. V_t variation is normally distributed and its 3σ variation is about 30mV in a 180nm CMOS logic technology. This variation causes a significant variation in circuit performance and leakage. The most critical paths in a chip may be different from chip to chip. Figure 2 also shows the 20x I_{sb} variation distribution in detail.

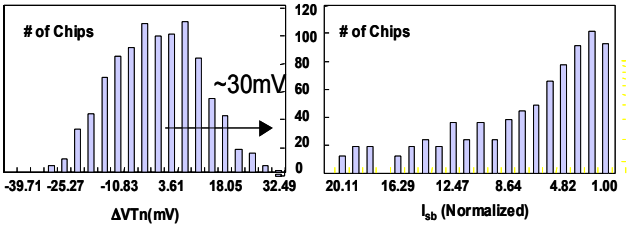


Figure 2: Die-to-die V_t, I_{sb} variation

2.2 Supply Voltage Variations

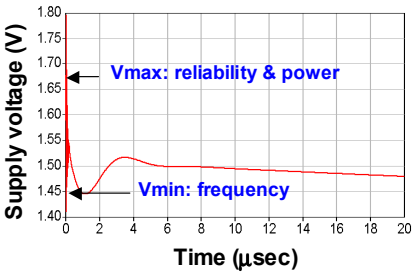


Figure 3: Supply voltage variation

Variations in switching activity across the die and diversity of the type of logic, result in uneven power dissipation across the die. This variation results in uneven supply voltage distribution and temperature hot spots, across a die, causing transistor subthreshold leakage variation across the die.

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 (V_t), and (2) to meet the transistor performance goals. Maximum Vcc is specified as a reliability limit for a process and minimum Vcc is required for the target performance. Vcc variation inside the max-min window is shown in Figure 3. This figure shows a droop in Vcc, which degrades the performance. Packaging and platform technologies do not follow the scaling trends of CMOS process. Therefore, power delivery impedance does not scale with Vcc and ΔVcc has become a significant percentage of Vcc.

2.3 Temperature Variation

Figure 4 shows the thermal image of a leading microprocessor die with as high as 120°C hot spots.

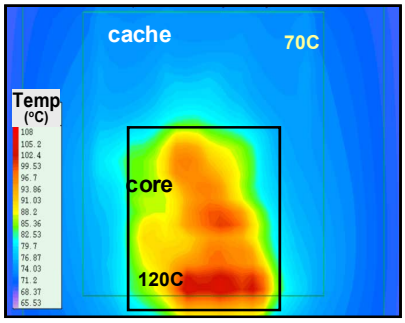


Figure 4: Within die temperature variation

Within die temperature fluctuations have existed as a major performance and packaging challenge for many years. Both the device and interconnect performance have temperature dependence, with higher temperature causing performance degradation. Additionally, temperature variation across communicating blocks on the same chip may cause performance mismatches, which may lead to logic or functional failures.

The net consequence of the P, V, T variation manifests itself on chip frequency variation. Figure 5 rolls up the distribution of microprocessor dies in 180nm technology across a frequency range. Variation in frequency is due to variation of other parameters discussed in Figure 2. This frequency distribution has serious cost implications associated with it: Low performing

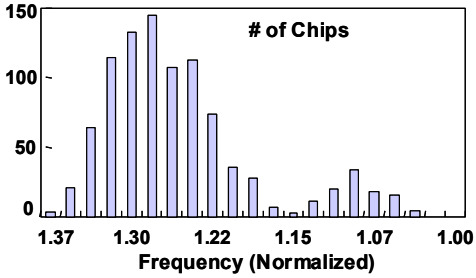


Figure 5: Die-to-die frequency variation

parts need to be discarded which in turn affects the yield and hence the cost.

3. IMPACT ON CIRCUITS AND MICROARCHITECTURE

The P, V, T variations impact all levels of design. In this section we will highlight some of the impact process has on circuit and microarchitecture design choices.

Dual-Vt circuit designs [5,6] can reduce leakage power during active operation, burn-in and standby. Two V_t's are provided by the process technology for each transistor. High-V_t transistors in performance critical paths are either upsized or are made low-Vt to provide the target chip performance. Figure 6 is a conceptual view of this tradeoff. On the left side, larger transistor sizes increase the relative probability of achieving the target frequency at the expense of switching power. The right chart shows that increasing low-V_t usage also boosts the probability, but with a penalty in leakage power. It was shown in [5,6], that by carefully employing low-V_t devices, 24% delay improvement is possible to trade off leakage and switching power components, while maintaining the same total power.

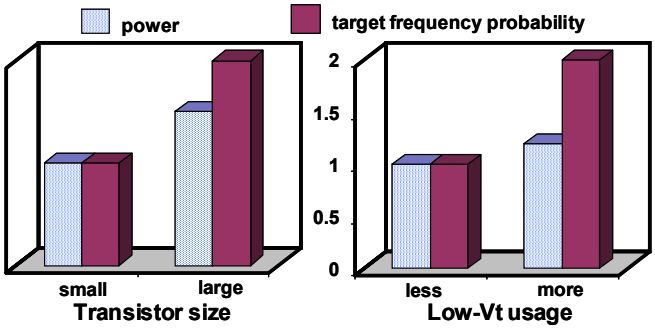


Figure 6: Circuit design tradeoffs

The number of critical paths that determine the target frequency vary depending on the microarchitecture design choice. Microarchitecture designs that demand increased parallelism and/or functionality require increase in the number of critical paths. Designs that require deeper pipelining, to support higher frequency of operation, require increase in the number of critical paths and decrease in the logic depth. The impact process variation has on these choices are described next.

Testchip measurements in Figure 7 show that as the number of critical paths on a die increases, within-die delay variations among critical paths cause both mean (µ) and standard deviation (σ) of the die frequency distribution to become smaller. This is consistent with statistical simulation results [1] indicating that

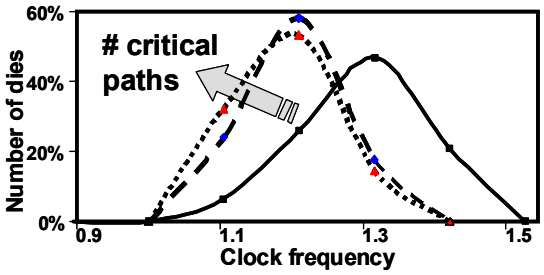


Figure 7: Die-to-die critical path distribution

the impact of within-die parameter variations on die frequency distribution is significant. As the number of critical paths exceeds 14, there is no noticeable change in the frequency distribution. So, microarchitecture designs that increase the number of critical paths will result in reduced mean frequency, since the probability that at least one of the paths is slower will increase.

Historically, the logic depth of microarchitecture critical paths has been decreasing to accommodate a 2x growth in the operating frequency every generation, faster than the 42% supported by technology scaling. As the number of logic gates that determine the frequency of operation reduces, the impact of variation in device parameter increases, as illustrated in Figure 8. Measurement on 49-stage ring oscillators showed that σ of the within-die frequency distribution was 4x smaller than σ of the device saturation current distribution [1]. However, measurements on a testchip containing 16-stage critical paths show that σ's of within die (WID) critical path delay distributions and NMOS/PMOS drive current distributions are comparable.

Logic depth: 16		
NMOS Ion σ/μ	PMOS Ion σ/μ	Delay σ/μ
5.6%	3.0%	4.2%

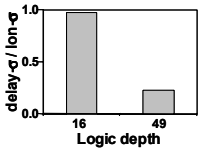


Figure 8: Impact of logic depth

Figure 9 summarizes the impact of process variation on the microarchitecture design choices. The bar charts conceptually show that with either smaller logic depth or with increasing number of microarchitecture critical paths, performance improvement is possible. However, the probability of achieving the target frequency that translates to performance, drops due to the impact of within-die process variation.

4. VARIATION TOLERANCE AND REDUCTION

In this section, we describe some of the research and design work to enhance the variation tolerance of circuits and microarchitecture and to reduce the variations by clever circuit and microarchitectural techniques. We first describe the body

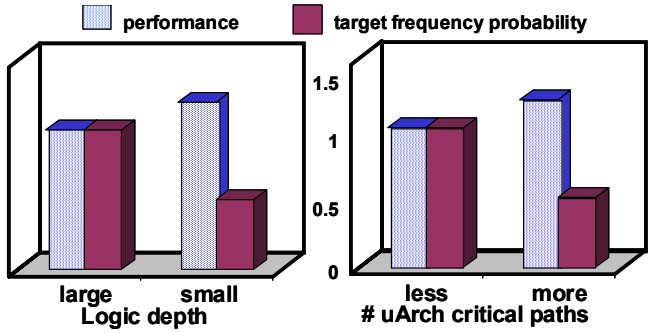


Figure 9: Microarchitecture tradeoffs

(substrate) biasing techniques, followed by supply voltage and temperature variation tolerance methods.

4.1 Body Bias Control Techniques

Device performance can be improved by lowering V_t, however, that leads to higher leakage current I_{sb}. One possible method to trade off performance with leakage power is to apply a separate bias to critical devices.

4.1.1 Vt Modulation by Forward Body Bias (FBB)

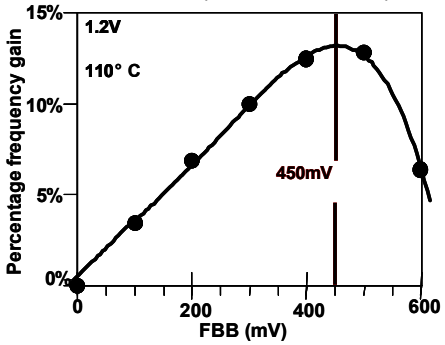
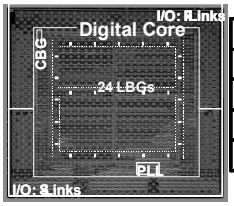


Figure 10: Optimal FBB for sub-90nm generations

Device V_t is a function of body to source potential (V_{BS}). V_t can be modulated for higher performance by forward biasing the body. This method also reduces the impact of short channel effects, hence reducing V_t variations. Figure 10 plots the percentage frequency gain as a function of FBB. It was shown empirically that 450mV is the optimal FBB for sub-90nm generations at high temperature [7].



Die size	10.1 X 10.1 mm2
Technology	150nm CMOS
Transistors	6.6 million
Area overhead	2%
Power overhead	1%

Figure 11: Forward body bias evaluation testchip

A 6.6M transistors communications router chip [8], with on-chip circuitry to provide forward body bias (FBB) to PMOS devices during active operation and zero body bias (ZBB) during standby mode, was implemented in a 150nm CMOS technology (Figure 11). FBB is withdrawn during standby mode to reduce leakage power.

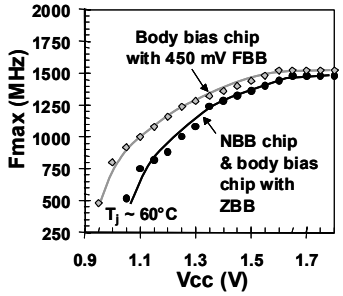


Figure 12: Forward body bias results

Performance of the chip is compared with the original design that has no body bias (NBB) in Figure 12. F_{max} of the NBB and FBB router chips are compared from 0.9V to 1.8V Vcc at 60°C (Figure 13). The FBB chip with forward body bias achieves 1GHz operation at 1.1V, compared to 1.25V required for the NBB chip. The switching power is 23% smaller at 1GHz. The frequency of FBB is 33% higher than NBB at 1.1V.

4.1.2 Leakage Reduction by Reverse Body Bias (RBB)

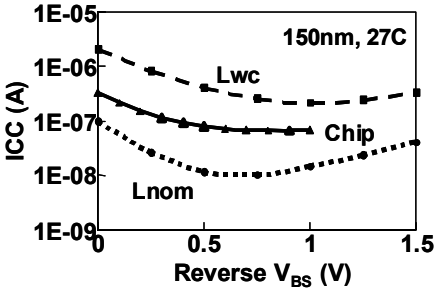


Figure 13: Leakage reduction by reverse body bias

An alternate method for reducing I_{sb} is to apply reverse V_{BS}. Figure 13 plots the leakage current for the worst-case channel length (L_{wc} dashed) and the nominal channel length (L_{nom} dotted) as a function of RBB. The measured full-chip leakage current is within these upper and lower leakage current bounds over a range of RBB values. The optimum RBB value derived from the measured chip for minimum leakage is 500mV [9]. Using RBB values larger than this value causes the junction leakage current to increase and overall leakage power to go up. However, effectiveness of RBB reduces as channel lengths become smaller or V_t is lowered. Essentially, the V_t-modulation capability by RBB weakens as short-channel effects become worse or body effect diminishes due to lower channel doping.

4.1.3 Power/Performance and Bin Improvement by Adaptive Body Bias (ABB)

The previous two subsections presented the advantages of both FBB and RBB. It is possible to utilize both of these approaches as shown in Figure 14. Due to the frequency spread in fabricated parts caused by process variations, the low frequency parts may be discarded for lower performance and the high frequency parts may be discarded for higher leakage power. As shown on the right side, devices can be adaptively biased to increase the performance of the slow parts by FBB and to decrease leakage power of the fast parts by RBB.

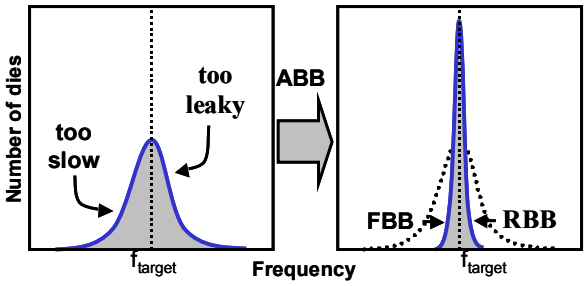
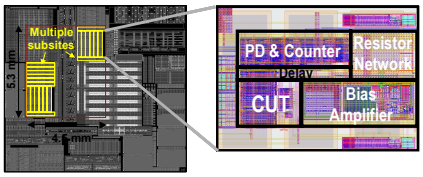


Figure 14: Target frequency binning by adaptive body bias

A testchip (Fig. 15) has been implemented in a 150nm CMOS technology to evaluate effectiveness of the adaptive body bias (ABB) technique [10] for minimizing impacts of both die-to-die and within-die parameter variations on processor frequency and active leakage power. The bias is based on a 5-bit digital code, which provides one of 32 different body bias values with 32mV resolution to PMOS transistors. NMOS body is biased externally across the chip.



Technology	150nm CMOS
Number of subsites per die	21
Body bias range	0.45V FBB to 0.5V RBB
Bias resolution	32 mV

Figure 15: Adaptive body bias evaluation testchip

Bidirectional ABB is used for both NMOS and PMOS devices to increase the percentage of dies that meet both frequency requirement and leakage constraint. As a result, die-to-die frequency variations (σ/μ) reduce by an order of magnitude, and 100% of the dies become acceptable (Figure 16). Bin 2 is the highest frequency bin while Bin 1 is the lowest acceptable frequency bin – any dies that are slower than Bin 1 are discarded. Almost 50% of dies with NBB fell below Bin 1 but are recovered using ABB. In addition, 30% of the dies are now in the highest frequency bin allowed by the power density limit. WID-ABB (applying multiple bias values per die to compensate for within-die as well as die-to-die variation) reduces σ of the die frequency distribution by 50%, compared to ABB. In addition, almost all of the dies are accepted in the highest possible frequency bin, compared to 30% for ABB.

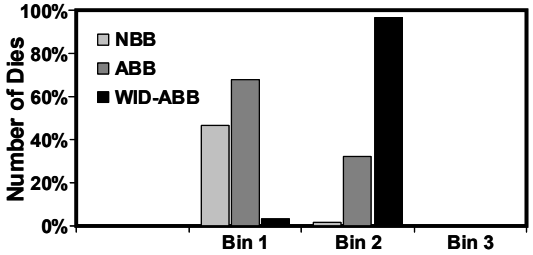


Figure 16: Adaptive body bias results

4.2 Supply Voltage and Temperature Control Techniques

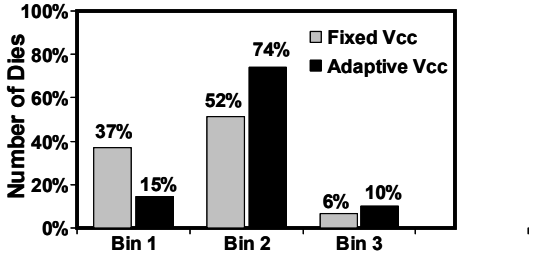


Figure 17: Bin improvement by adaptive Vcc

As introduced in Section 3, variations in switching activity across the die, and diversity of the type of logic, results in uneven power dissipation across the die. This variation results in uneven supply voltage distribution and temperature hot spots.

A technique to increase yield in the high frequency bins, is to apply adaptive Vcc. Figure 17 shows the advantage of adaptive Vcc over fixed Vcc. Bin 3 is the highest frequency bin, while Bin 1 is the lowest acceptable frequency bin. The dark bars indicate that adaptive Vcc has pushed more than 20% dies from Bin 1 to Bin 2 and even Bin 3, as well as recovered those dies that fell below Bin 1.

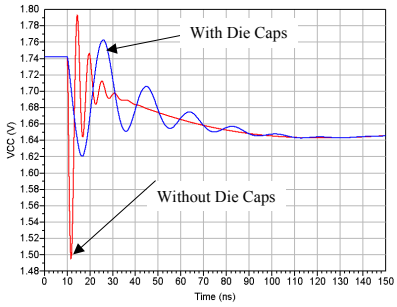


Figure 18: Effectiveness of on-die decoupling capacitors

Adaptive Vcc does not solve the ΔVcc problem. Figure 18 shows a well-known technique to mitigate voltage variations, namely, adding on-die decoupling capacitors. ΔVcc reduces from 15% to <10% by carefully placing appropriate amount of decaps on microprocessor dies [11]. However, it has to be noted that decoupling capacitors cost silicon area. In sub-90nm technologies designers are facing a challenge of gate oxide leakage. Decap layouts are designed for high capacitance to area ratio, however, that tends to increase the gate oxide area. The oxide leakage and area penalty must be traded off with ΔVcc.

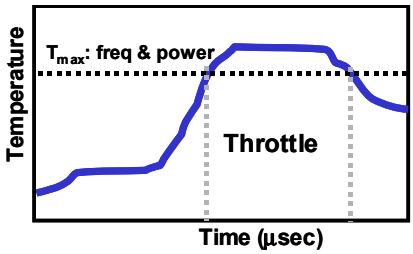


Figure 19: Temperature based Vcc/frequency throttling

Maximum temperature and within die temperature variations have to be controlled. Throttling is used to control the die temperature. Figure 19 explains the method. When the temperature rises above the maximum limit set by frequency and power, the operating frequency is lowered, followed by the die Vcc. Subsequently, the power consumption drops followed by a drop in on-die temperature. When the die comes out of power saving mode, Vcc is raised followed by frequency. Many commercial processors incorporate throttling.

5. CONCLUSIONS

We presented a major challenge for future designs at circuit and microarchitecture levels, namely, parameter variations. We discussed process, voltage and temperature variations; and their impact on circuit and microarchitecture. Possible solutions to reduce the impact of parameter variations and to achieve higher frequency bins were also presented. Variations will pose significant challenge, necessitating a shift in the design paradigm, from today's deterministic design to statistical or probabilistic design.

6. REFERENCES

[1] Bowman, K., et. al, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration", IEEE Journal of Solid-State Circuits, Volume 37, Feb 2002, pp.183-190.

[2] Borkar, S., "Parameter Variations and Impact on Circuits & Microarchitecture", C2S2 MARCO review, March 2003.

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

[4] Karnik, T., et al., "Sub-90nm Technologies—Challenges and Opportunities for CAD", ICCAD 2002, pp. 203-206.

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

[6] 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.

[7] Tschanz, J., et al., "Dynamic-Sleep Transistor and Body Bias for Active Leakage Power Control of Micro-processors", ISSCC 2003, pp.102-103.

[8] Narendra, S., et al., "1.1 V 1 GHz communications router with on-chip body bias in 150 nm CMOS", ISSCC 2002, pp. 270-271.

[9] Keshavarzi, A., et al., "Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual Vt CMOS Ics", ISLPED 2001, pp.207-210.

[10] Tschanz, J., et al., "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage", ISSCC 2002, pp.422-423.

[11] Rahal-Arabi, T., et al., "Design and validation of the Pentium III and Pentium 4 processors power delivery", VLSI Symposium 2002, pp220-223.