# Design Optimizations for Microprocessors at Low Temperature

Arman Vassighi, Ali Keshavarzi, Siva Narendra, Gerhard Schrom,
Yibin Ye, Seri Lee[1], Greg Chrysler[2], Manoj Sachdev[3], Vivek De

Circuits Research, [1]Platform Technology, Intel Labs, Hillsboro, OR,
[2]Advanced Technology, Intel Labs, Chandler, AZ, [3]University of Waterloo, Canada

## ABSTRACT

We investigate trade-offs in microprocessor frequency and system power achievable for low temperature operation in scaled high leakage technologies by combining refrigeration with supply voltage selection, body bias, transistor sizing and shorter channel length. Reducing channel length provides better frequency and power improvement than forward body bias. When, the leakage power is more than 30% of chip power, combining refrigeration with enhancing technology by shorter channel length provides the best trade-off for power and frequency.

**Categories & Subject Descriptors:** Design

**General Terms:** Design, Performance

**KEYWORDS:** Low temperature, Electrothermal modeling, microprocessor, power, frequency, CMOS, cooling, refrigeration

## 1. INTRODUCTION

Potential advantages of using refrigeration for cooling processors have been reported in the past [1]. Using the algorithm shown in Fig. 1, we investigate trade-offs in microprocessor clock frequency, energy efficiency (MIPS/Watt), die area and system power when we use active cooling to reduce the operating temperature of the microprocessors below typical hot temperature of 110°C. However, we are not looking at sub-ambient operating temperatures. We would like to lower microprocessor system temperature below 110°C depending on cooling efficiency of the technique we have selected. We are interested to find out if low-temperature CMOS operation has any merit for scaled technologies where transistor subthreshold leakage is relatively high. And if yes, what kind of device, circuit, and design choices are applicable for high performance microprocessors.

We consider active cooling with and without refrigeration. We investigate several active cooling techniques including: air cooling, liquid cooling and refrigeration. Refrigeration is the most effective cooling solution and is considered for junction temperatures not much below the ambient temperature. We have considered cooling power into our system power trade-offs. Consequently, we study the above mentioned trade-offs achievable by combining active cooling with (1) supply voltage (Vcc) selection, (2) applying body bias, (3) sizing of transistors in critical and non-critical paths on chip, and (4) reduction of channel length (L) as a function of different process technology worst case leakage limit. System power is the total of chip power (switching and leakage) and power consumed by the cooling system. Analytical models are used for frequency, power, die area, etc. in an electrothermal analysis tool (Fig. 1). The tool performs analysis of (1) frequency, limited by logic and interconnect RC paths, (2) system energy efficiency, (3) chip switching and leakage powers, including subthreshold and gate oxide leakage, (4) package and cooling system characteristics, (5) die area, (6) gate oxide reliability-limited maximum Vcc constraints, and (7) maximum temperature in a self-consistent manner. The model parameters and input parameters to the tool are typical values. Some of the parameters are extracted from device measurements, process files, and chip measurements.


*DAC 2004,* June 7-11, 2004, San Diego, California, USA.


## 2. SELF-CONSISTENT ELECTROTHERMAL OPTIMIZATION

To account for the change in temperature, we have developed an electrothermal analysis tool to self-consistently compute temperature, power, and operating frequency of the microprocessor. Starting with an initial assumption of temperature (T), the electrothermal analysis tool first computes frequency and power. Then it computes the new temperature (T) from this power based on package and cooling system characteristics, and computes frequency and power again with this new T. These iterations continue until temperatures computed in consecutive steps are close where the convergence has occurred. If we do not achieve convergence, it indicates thermal runaway [3]. When the iterations converge, that is the final self-consistent junction temperature. It also produces values for frequency, chip switching and leakage power, active cooling system power, and die area that are consistent with the final temperature, given (1) thermal resistance of the packaging and cooling system, (2) coefficient of performance (COP) for active cooling (defined as ratio of cooled power to the power consumed by the cooling system), and (3) chip design and process technology characteristics (Fig. 2).

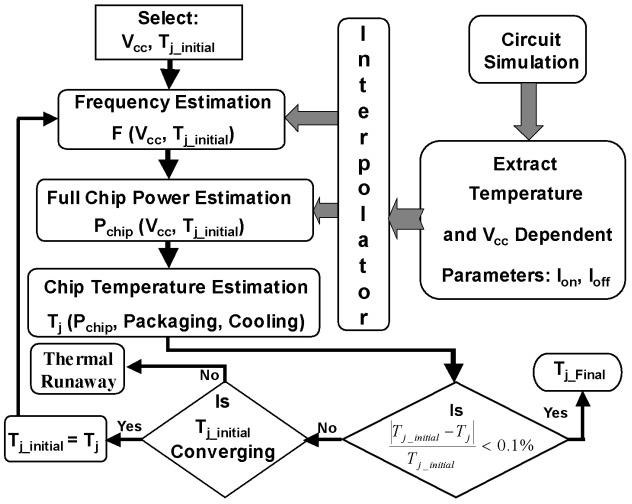


**Figure 1. Algorithm of the implemented Electrothermal analysis tool for this study [2].**

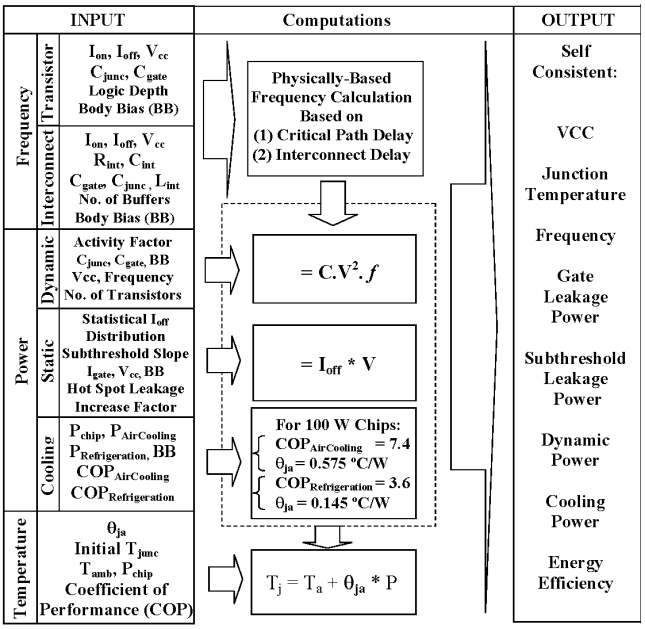


**Figure 2. Algorithm of self-consistent physically-based Electrothermal modeling approach.**

We have attempted to keep our tool physically-based. Our frequency calculations mimic microprocessor's frequency limitation by considering critical path delay and the role of interconnects. We use circuit parameters such as supply voltage, body bias value, number of buffers used in long interconnect lines, and logic depth in critical paths. We use transistor parameters including: Ion, Ioff, Cj, and Cgate and interconnect parameters including: Cint and Rint. The critical path logic depth will be used to transition from transistor to microprocessor frequency calculation.

We calibrated our tool to actual microprocessor measurements. For power calculations, we incorporated switching power, leakage power and cooling power. We ignored the short circuit power. Dynamic switching power is computed according to appropriate microprocessor activity factor, chip supply voltage, chip switching capacitance, body bias, and the area based on the number of transistors. Static leakage power has also considered gate leakage. The chip leakage is derived based on statistical transistor leakage distributions [4]. We have also considered the role of hot spots on the chips for leakage and maximum operating frequency of our chip. Cooling power was computed based on chip power and COP and thermal resistance of different cooling solutions. We have captured these considerations in Fig. 2.

To demonstrate how our modeling works, we show in Fig. 3 the results of optimization for an example microprocessor in a low-leakage 130nm process technology for a typical package and air cooling system. Solutions are obtained for different Vcc values, and an operating point is accepted only if Vcc does not exceed the gate oxide reliability-limited maximum (Vmax) at the final T.

Therefore, reliability considerations set our maximum allowable supply voltage. The highest optimal frequency and corresponding T are set by Vcc = Vmax at that T. The x-axis of Fig. 3 represents chip operating frequency. The y-axis has captured multiple parameters including power, temperature, supply voltage and reliability maximum allowed supply voltage. As supply voltage increases, chip frequency increases, and the junction temperature rises. However, maximum reliability limited supply voltage reduces at higher temperatures due to degraded gate oxide reliability performance. Consequently the maximum supply voltage is determined at the cross section of the Vcc and reliability limited max Vcc curves. This sets our junction temperature, frequency and power of the chip accordingly. The optimal operating frequency is 2.7 GHz at Vcc of 1.5V and Tj of 81°C where the system power is 82W. Interconnect RC delays with repeaters can also limit the maximum frequency (Fig. 2) since RC delays change with T in a different way from transistor performance and circuit delay. Also, RC delay is relatively insensitive to Vcc change, whereas circuit delays in logic paths change significantly with Vcc. An optimum design should be such that interconnects do not limit chip frequency and power at the optimum frequency. This allows transistors to provide their highest potential performance. This is shown in Fig. 3 where interconnect dashed line increases power without improving the chip frequency after optimal operating point.

## 3. LOW LEAKAGE VS. HIGH LEAKAGE TECHNOLOGY TRADE-OFFS

Fig. 4 shows the frequency and power trade-offs for iso-reliability high performance operation and iso-power operation conditions when we incorporate refrigeration active cooling. The relative contributions of cooling power, dynamic power, and leakage power demonstrate how we can trade-off leakage power for

cooling power. This is best shown in iso-power case. For a constant power limit of 80W, frequency increases by 4.5% going from air cooling to refrigeration in a low-leakage technology, and by 7.5% for high-leakage technology (Fig. 4). This happens because when leakage is a large percentage of total power (31% in this case), leakage power reduction due to lower T, translates to more savings in total chip power. Then, power overhead of the cooling system will have less impact on total system power.

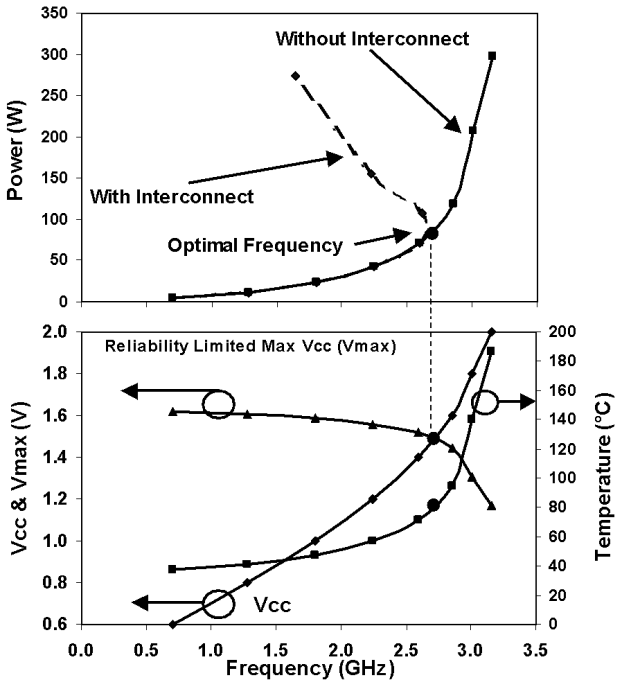


**Figure 3. Optimization of microprocessor operating frequency subject to reliability constraints.**

To achieve highest operating frequency in line with microprocessor applications, we should study the iso-reliability case as shown in Fig. 4. For a low-leakage technology, the reliability-limited frequency improves by 12% and the system power increases by 35% going from air cooling to refrigeration. When leakage is higher, frequency increases by 17% for 62% increase in system power. Therefore, frequency vs. power trade-off is worse when leakage is higher. Frequency improvements in both cases come from operation at reduced temperature and higher Vmax allowed at lower T due to utilizing refrigeration.

# 4. POWER, FREQUENCY AND ENERGY COMPARISONS FOR OPTIMAL DESIGNS AT LOW TEMPERATURE

Now that we have studied the active cooling for different amount of worst case process technology leakage constraints, we can investigate the optimum design for low temperature CMOS operation. We consider the following design techniques for optimal low temperature operation: changing Vcc, changing transistor channel length and enhancing the process technology, changing transistor sizing, and applying body bias. Fig. 5 shows

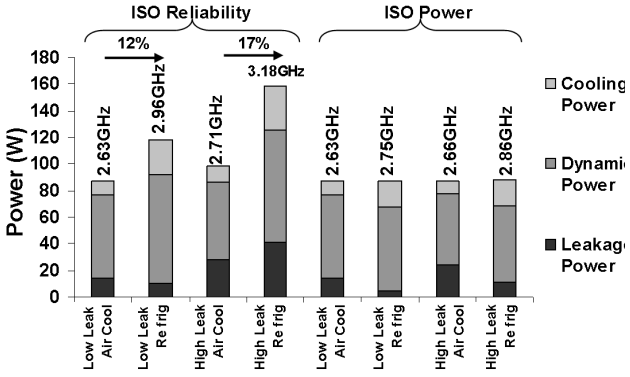


**Figure 4. Reliability and power limited maximum frequency achievable for low and high leakage technologies with refrigeration.**

trade-offs in system power, energy efficiency and die area vs. frequency offered by forward body bias (FBB), shortening L, changing Vcc and transistor sizing, with and without refrigeration.

System power and system energy efficiency as a function of chip frequency is plotted in Fig. 5. These graphs are normalized to air cooling power, energy efficiency and frequency. When we utilize refrigeration combined with a design technique, we are interested in smallest slope in system power versus chip frequency. This corresponds to maximum chip frequency increase for lowest increase in system power. For system energy efficiency, we are interested in maximum change in frequency and highest possible energy efficiency. Fig. 5 shows applying forward body bias in addition to refrigeration increases frequency but the rate of system power increase is rather steep. Applying 0.4V FBB increases frequency by an additional 2.7% and increases power by 27%. The best FBB trade-off is when its value is limited to 100mV. FBB also degrades energy efficiency by 16%. Decreasing Vcc from 1.56V to 1.4V lowers frequency and system power. However, at lower Vcc values, the rate of chip slow-down is much higher than the achieved power saving. Reducing sizing by lowering transistor width has similar trade-offs as supply voltage.

Table 1 summarizes integration of different design solutions and explores the design space for iso-power and iso-frequency conditions. Combined refrigeration with shorter L (enhancing technology), appropriate Vcc selection and transistor sizing provides the highest frequency for any system power limit and the highest energy efficiency for any target frequency. Highest frequency increase of 11% is achieved for iso-power at Vcc of 1.41V, temperature of 31°C and 11% smaller area for enhancing the technology in our design space. If we perform iso-frequency analysis, enhancing technology (shorter L), provides 38% total system power saving at Vcc of 1.36V, temperature of 15°C and 33% smaller area. In both cases we improve the energy efficiency by 11% and 62% respectively.

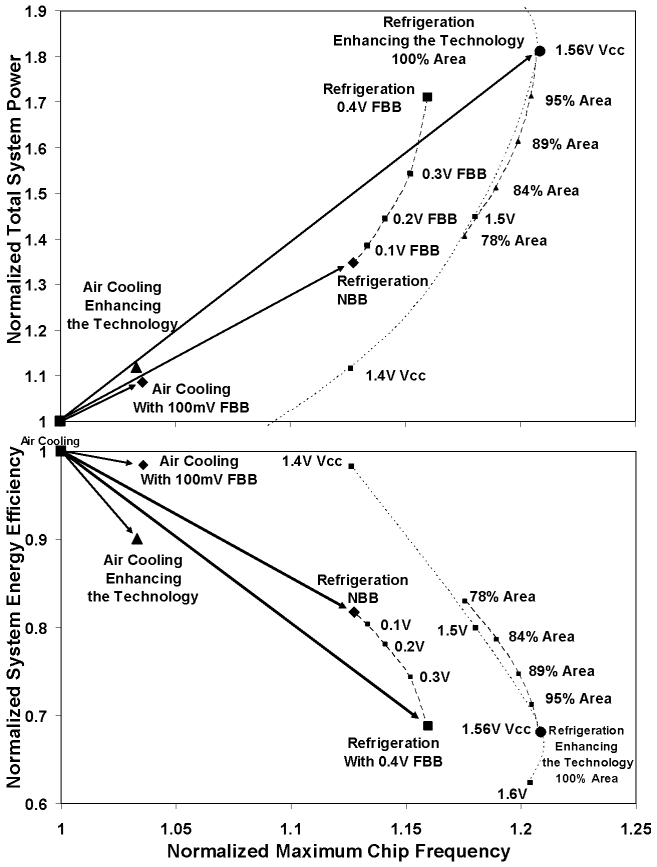


**Figure 5. Trade-offs in system power, energy efficiency, die area, and frequency by different circuit and design techniques.**

In summary, Table 1 shows improvements in frequency for equal power and reduction in power for a specific target frequency for air cooling and refrigeration when transistor sizing and supply voltage are optimized for optimal forward body bias, and shorter L. Die area changes are also compared. Reducing L provides better frequency and power improvement than FBB in all cases. Also, combining refrigeration with shorter L is the best for power and frequency. Furthermore It provides lowest die area when comparing power at equal frequency, and second best when comparing frequency at equal power.

## 5. CONCLUSIONS

We investigated trade-offs in microprocessor frequency and system power achievable by combining refrigeration with supply voltage selection, body bias, transistor sizing and shorter channel length. Reducing channel length provided better frequency and power improvement than forward body bias. Also, combining refrigeration with shorter length is the best for power and frequency. Frequency is 11% higher at equal power and power is 38% lower at equal frequency, compared to air cooling, in a case study for an example microprocessor in a 130nm process technology.

## 6. ACKNOWLEDGEMENT

We would like to acknowledge collaborations of different electrical and mechanical disciplines by providing physical parameters used in this study. We also acknowledge valuable discussion with Prof. Banerjee of UCSB and Prof. Segura of UIB.

## 7. REFERENCES

[1] I. Aller, et. al., 2000 ISSCC, pp. 214-215.

[2] K. Banerjee, L. Sheng-Chih, A. Keshavarzi, S. Narendra, V. De, 2003 IEDM, pp. 36.7.1 - 36.7.4.

[3] A. Vassighi, O. Semenov, M. Sachdev, "Thermal Runaway Avoidance", 2004 IRPS.

[4] S. Narendra., V. De, S. Borkar, D. A. Antoniadis, A.P. Chandrakasan, IEEE Journal of Solid-State Circuits , March 2004, pp. 501 – 510.

**Table 1. Optimum design space for active cooling at iso-power and iso-frequency conditions.**

| Iso Power | Vcc | Temp | %Area | %Leakage | %Cooling | %Energy | %Frequency |
|---|---|---|---|---|---|---|---|
| Air Cooling Reference | 1.49 | 80 | 100% | 19 | 14 | 100% | 100% |
| Air Cooling with 100 mV FBB | 1.47 | 81 | 95% | 21 | 14 | 102% | 102% |
| Air Cooling with Enhamced Technology | 1.4 | 82 | 89% | 31 | 14 | 107% | 107% |
| Refrigeration with 200 mV FBB | 1.6 | 30 | 73% | 13 | 29 | 108% | 108% |
| Refrigeration with Enhanced Technology | 1.41 | 31 | 89% | 18 | 28 | 111% | 111% |

| Iso Frequency | Vcc | Temp | %Area | %Leakage | %Cooling | %Energy | %Power |
|---|---|---|---|---|---|---|---|
| Air Cooling Reference | 1.49 | 80 | 100% | 19 | 14 | 100% | 100% |
| Air Cooling with 100 mV FBB | 1.45 | 76 | 89% | 19 | 14 | 115% | 87% |
| Refrigeration with 200 mV FBB | 1.46 | 20 | 73% | 8 | 29 | 134% | 75% |
| Air Cooling with Enhamced Technology | 1.34 | 72 | 78% | 27 | 14 | 137% | 73% |
| Refrigeration with Enhanced Technology | 1.36 | 15 | 67% | 12 | 30 | 162% | 62% |