Modeling the Effect of Technology Trends on Soft Error Rate of Combinational Logic

Premkishore Shivakumar Michael Kistler *

Stephen W. Keckler Doug Burger Lorenzo Alvisi

Department of Computer Sciences
University of Texas at Austin
Austin, TX USA
{*pkishore,kistler,skeckler,dburger,lorenzo*}*@cs.utexas.edu*

IBM Technical Contacts: John Keaty, Rob Bell, and Ram Rajamony

Abstract

This paper examines the effect of technology scaling and microarchitectural trends on the rate of soft errors in CMOS memory and logic circuits. We describe and validate an end-to-end model that enables us to compute the soft error rates (SER) for existing and future microprocessor-style designs. The model captures the effects of two important masking phenomena, electrical masking and latching-window masking, which inhibit soft errors in combinational logic. We quantify the SER in combinational logic and latches for feature sizes from 600nm to 50nm and clock rates from 16 to 6 fan-out-of-4 delays. Our model predicts that the SER per chip of logic circuits will increase eight orders of magnitude by the year 2011 and at that point will be comparable to the SER per chip of unprotected memory elements. Our result emphasizes the need for computer system designers to address the risks of SER in logic circuits in future designs.

1 Introduction

Two important trends driving microprocessor performance are scaling of device feature sizes and increasing pipeline depths. In this paper we explore how these trends affect the susceptibility of microprocessors to soft errors. Device scaling is reduction in feature size and voltage levels of the basic devices on the microprocessor. The basic motivation for device scaling is to improve processor performance, since smaller devices require less current to turn on or off, and thus can be operated at higher frequencies. Pipelining is a microarchitectural technique for improving performance by increasing instruction level parallelism (ILP). Pipelining is a well accepted and almost universally adopted technique in microprocessor design. Five stage pipelines are quite common, and even processors with six to eight pipeline stages are considered to be relatively simple designs. Recent processors have aggressively applied the techniques of pipelining, with some current designs using upwards of twenty stages [9]. Such designs are commonly referred to as *superpipelined* designs.

Our study focuses on *soft errors*, which are also called transient faults or single-event upsets (SEUs). These are errors in processor execution that are not due to design or manufacturing defects, but instead due to electrical noise or external radiation. In particular, we are interested in soft errors caused by cosmic rays. The existence of cosmic ray radiation has been known for over 50 years, and the capacity for this radiation to create transient faults in semiconductor circuits has been studied since the early 1980s. As a result, most modern microprocessors already incorporate mechanisms for detecting soft errors. These mechanisms are typically focused on protecting memory elements, particularly caches, using error-correcting codes (ECC), parity, and other techniques. Two key reasons for this focus on memory elements are: 1) the techniques for protecting memory elements are well understood and relatively inexpensive in terms of the extra circuitry required, and 2) caches take up a large part, and in some cases a majority, of the chip area in modern microprocessors.

Past research has shown that combinational logic is much less susceptible to soft errors than memory elements. This is because three phenomena provide combinational

*This author is also employed at the IBM Austin Research Laboratory.

logic a form of natural resistance to soft errors: 1) logical masking, 2) electrical masking, and 3) latching-window masking. We develop models for electrical masking and latching-window masking to determine how these are affected by device scaling and superpipelining. Then based on a composite model we estimate the effects of these technology trends on the soft error rate (SER) of combinational logic. Finally using an overall chip area model we compare the SER/chip of combinational logic with the expected trends in SER of memory elements.

The primary contribution of our work is an analysis of the trends in SER for SRAM cells, latches, and combinational logic. Our models predict that by 2011 the soft error rate in combinational logic will be comparable to that of unprotected memory elements. This is extremely significant because current methods for protecting combinational logic have significant costs in terms of chip area, performance, and/or power consumption in comparison to protection mechanisms for memory elements. Technology trends will lead to a significant reduction in both electrical and latching-window masking, which accounts for a major portion of the increase in SER of combinational logic.

The rest of this paper is organized as follows. Section 2 provides background on the nature of soft errors, and a method for estimating the soft error rate of memory circuits. Section 3 introduces our definition of soft errors in combinational logic, and examines the phenomena that can mask soft errors in combinational logic. Section 4 describes in detail our methodology for estimating the soft error rate in combinational logic. We present our results in Section 5. Section 6 discusses the implications of our analysis and simulations. Section 7 summarizes the related work, and Section 8 concludes the paper.

2 Background

2.1 Particles that cause soft errors

In the early 1980s, IBM conducted a series of experiments to measure the particle flux [27]. Flux is generally a measure of rate of flow; in this paper the flux of cosmic ray particles is expressed as the number of particles of a particular energy per square centimeter per second. For our work, the most important aspect of these results is that particles of lower energy occur far more frequently than particles of higher energy. In particular, a one order of magnitude difference in energy can correspond to a two orders of magnitude larger flux for the lower energy particles. As CMOS device sizes decrease, they are more easily affected by these lower energy particles, potentially leading to a much higher rate of soft errors.

This paper investigates the soft error rate of combinational logic caused by atmospheric neutrons with energies greater than 1 mega-electron-volt (MeV). This form of radiation, the result of cosmic rays colliding with particles in the atmosphere, is known to be a significant source of soft errors in memory elements. We do not consider atmospheric neutrons with energy less than 1 MeV since we believe their much lower energies are less likely to result in soft errors in combinational logic. We also do not consider alpha particles, since this form of radiation comes almost entirely from impurities in packaging material, and thus can vary widely for processors at a particular technology. The contribution to the overall soft error rate from each of these radiation sources is additive, and thus each component can be studied independently.

2.2 Soft Errors in Memory Circuits

In most modern microprocessors, combinational logic and memory elements are constructed from the same basic devices – NMOS and PMOS transistors. Therefore, we can utilize techniques for estimating the SER in memory elements to assess soft errors in combinational logic. We will also use these techniques directly to compute the SER in memory elements for a range of device sizes, and compare the results to our estimates of SER for combinational logic.

High-energy neutrons lose energy in materials mainly through collisions with silicon nuclei that lead to a chain of secondary reactions. These reactions deposit a dense track of electron-hole pairs as they pass through a p-n junction. Some of the deposited charge will recombine, and some will be collected at the junction contacts. When a particle strikes a sensitive region of an SRAM cell, the charge that accumulates could exceed the minimum charge that is needed to flip the value stored in the cell, resulting in a soft error. The smallest charge that results in a soft error is called the *critical charge* (Q_{CRIT}) of the SRAM cell [6]. The rate at which soft errors occur is typically expressed in terms of *Failures In Time (FIT)*, which measures the number of failures per 10^9 hours of operation. A number of studies on soft errors in SRAMs have concluded that the SER for constant area SRAM arrays will increase as device sizes decrease [21, 20, 12], though researchers differ on the rate of this increase.

A method for estimating SER in CMOS SRAM circuits was recently developed by Hazucha & Svensson [8]. This model estimates SER due to atmospheric neutrons (neutrons with energies $>$ 1MeV) for a range of submicron feature sizes. It is based on a verified empirical model for the 600nm technology, which is then scaled to other technologies. The basic form of this model is:

$$SER \propto F \times A \times \exp\left(-\frac{Q_{CRIT}}{Q_S}\right) \quad (1)$$

where

F	is the neutron flux with energy > 1 MeV (in n*cm^{-2}s^{-1}),
A	is the area of the circuit sensitive to particle strikes (in cm^2),
Q_{CRIT}	is the critical charge (in fC), and
Q_S	is the charge collection efficiency of the device (in fC)

Two key parameters in this model are the critical charge (Q_{CRIT}) of the SRAM cell and the charge collection efficiency (Q_S) of the circuit. Q_{CRIT} depends on characteristics of the circuit, particularly the supply voltage and the effective capacitance of drain nodes. Q_S is a measure of the magnitude of charge generated by a particle strike. These two parameters are essentially independent, but both decrease with decreasing feature size. From Equation 1 we see that SER will increase exponentially as Q_{CRIT} becomes comparable to Q_S. SER is also proportional to the area of the sensitive region of the device, and therefore decreases proportional to the square of the device size. Hazucha & Svensson used this model to evaluate the effect of device scaling on the SER of memory circuits. They concluded that SER-per-chip of SRAM circuits should increase at most linearly with decreasing feature size.

3 Soft Errors in Combinational Logic

A particle that strikes a p-n junction within a combinational logic circuit can alter the value generated by the circuit. However, a transient change in the value of a logic circuit will not affect the results of a computation unless it is captured in a memory circuit. Therefore, we define a soft error in combinational logic as a transient error in the result of a logic circuit that is subsequently stored in a memory circuit of the processor.

A transient error in a logic circuit might not be captured in a memory circuit because it might be *masked* by one of the following three phenomena:

Logical masking occurs when a particle strikes a portion of the combinational logic that is blocked from affecting the output due to a subsequent gate whose result is completely determined by its other input values.

Electrical Masking occurs when the pulse resulting from a particle strike is attenuated by subsequent logic gates due to the electrical properties of the gates to the point that it does not affect the result of the circuit.

Latching-Window Masking occurs when the pulse resulting from a particle strike reaches a latch, but not at the clock transition where the latch captures its input value.

These masking effects have been found to result in a significantly lower rate of soft errors in combinational logic in comparison to storage circuits in equivalent device technology [16]. However, these effects will diminish significantly as feature sizes decrease and the number of stages in the processor pipeline increases. For example, electrical masking will be reduced by device scaling because smaller transistors are also faster and therefore will have less attenuation effect on the pulse. Also, deeper processor pipelines result in higher clock rates, which means the latches in the processor will cycle more frequently, which reduces the opportunity for latching-window masking.

3.1 Combinational Logic Model

The datapath of modern processors can be extremely complicated in nature, typically composed of 64 parallel bit lines and divided into 20 or more pipeline stages. We have chosen to use a much simpler model for the purposes of estimating the SER of combinational logic. Our model is just a one-wide chain of homogeneous gates terminating in a latch. Figure 1 illustrates this pipeline model. The gates we use in our study are all static combinational logic gates. Many modern microprocesors also employ dynamic logic because it occupies less area and offers greater flexibility for techniques such as time borrowing. These devices are commonly designed for high performance, and as a result have lower noise margins and may be more susceptible to soft errors. We believe our model can be extended to estimate the SER for dynamic logic and other circuit styles. The number of gates in the chain is dependent on the degree of pipelining in the microarchitecture, which we characterize by the number of fan-out-of-4 inverter (FO4) gates that can be placed between two latches in a single pipeline stage. The FO4 metric is technology independent and 1 FO4 roughly corresponds to 360 pico-seconds times the transistor's drawn gate length in microns [10]. During the last twelve years technology has scaled from 1000nm to 130nm and the amount of logic per pipeline stage has decreased from 84 to 12 FO4 contributing to a total of 60-fold increase in clock frequency in the Intel family of processors. Aggressive pipelining could reduce this to as few as 6 in five to seven years from now. For a given degree of pipelining, the number of gates in the pipestage is largest number that does not exceed the total delay of the corresponding FO4 chain.

In our model, a latch consists of a passgate, a forward inverter and a feedback inverter, where the forward inverter is

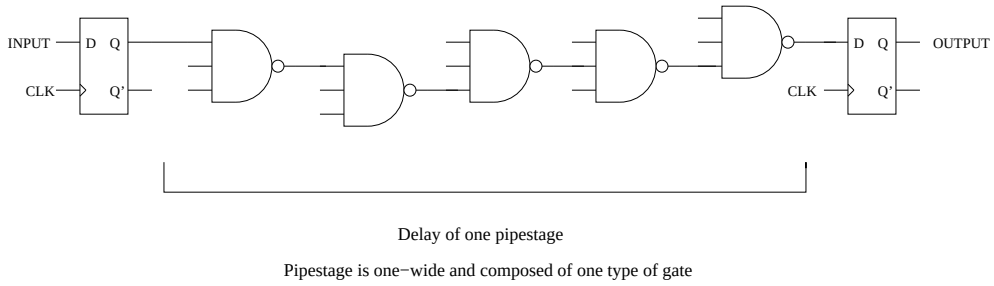

Figure 1. Simple model for a processor pipeline.

about 6 times larger than the feedback inverter and the transistors are all of minimum length. We use level sensitive latches in our pipeline model because they occupy less area than edge triggered flip-flops and so are more suitable for superpipelining. They also allow for time borrowing techniques and offer less load to the clock distribution network thus reducing the clock skew in the chip.

4 Methodology

Our methodology for estimating the soft error rate in combinational logic considers the impact of CMOS device scaling and the microarchitectural trend toward increasing depth of processor pipelines. We determine the soft error rate using analytical models for each stage of the pulse from its creation to the time it reaches the latch. Figure 2 shows the various stages the pulse passes through and the corresponding model used to determine the effect on the pulse at that stage. In the first stage an error current pulse is produced from the charge and the corresponding voltage pulse is also generated. The electrical masking model simulates the effect of the electrical properties of the gates on the pulse. Finally a model for the latching window determines the probability that the pulse is successfully latched. The following sections describe each model in detail.

4.1 Device Scaling model

In practice, technology parameters are scaled to achieve certain physical objectives such as constant power density or constant electric field strength. Our method for constructing technology parameters uses values from the Semiconductor Industry Association (SIA) technology roadmap [25], with minor adjustments to ensure that the delay of a fan-out-of-4 (FO4) inverter satisfies Equation 2. These technology parameters are used in the circuit simulations and analytical models for estimating SER of combinational logic.

$$t_{\text{FO4 delay}}(\text{in ps}) = 360 * \text{feature size (in } \mu\text{m}) \qquad (2)$$

4.2 Charge to Voltage Pulse model

When a particle strikes a device it produces a current pulse with a very rapid rise time, but a more gradual fall time. The shape of the pulse can be approximated by a one-parameter function [6] shown in Equation 3.

$$I(t) \propto \frac{Q}{T} \times \sqrt{\frac{t}{T}} \times \exp\left(-\frac{t}{T}\right) \qquad (3)$$

Q refers to the amount of charge collected due to the particle strike. The parameter T is the time constant of the transistor for the charge collection process. If T is large it takes more time for the charge to recombine and if T is small the current pulse dies faster. The rapid rise of the current pulse is captured in the square root function and the gradual fall of the current pulse is produced by the negative exponential dependence.

The current pulse produced by a particle strike results in a voltage pulse at the output node of the device. We use hspice to determine the characteristics of this voltage pulse. The voltage pulse is described in a 3 parameter form - rise time, fall time and effective width (width of the pulse at half the supply voltage) and is used as input to the electrical masking analytical model.

4.3 Electrical Masking Model

Electrical masking is the composition of two electrical effects that reduce the strength of a pulse as it passes through a logic gate. Circuit delays caused by the switching time of the transistors cause the rise and fall time of the pulse to increase. Also, the amplitude of a pulse with short duration may decrease since the gate may start to turn off before the output reaches its full amplitude. The combination of these two effects reduce the width of a pulse, making it less likely to cause a soft error. These effects are illustrated in Figure 3. The effect cascades from one gate to the next because at each gate the slope decreases and hence the amplitude also decreases.

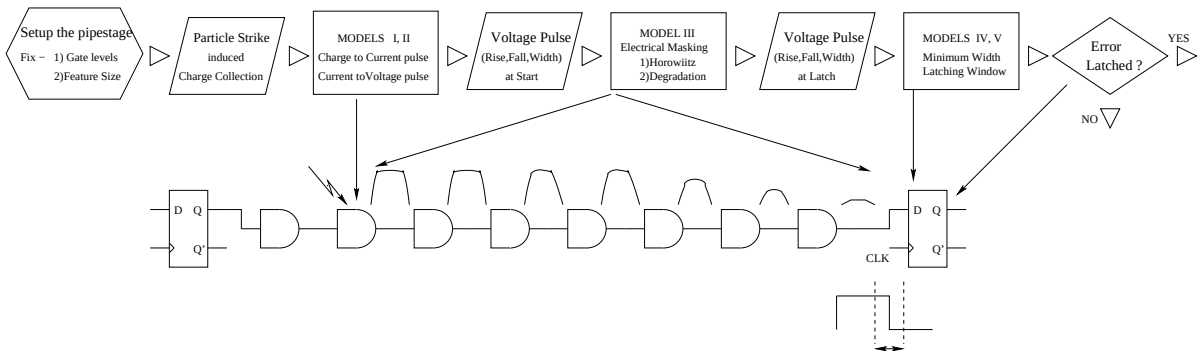

Figure 2. Overview of process to determine if a charge leads to a soft error

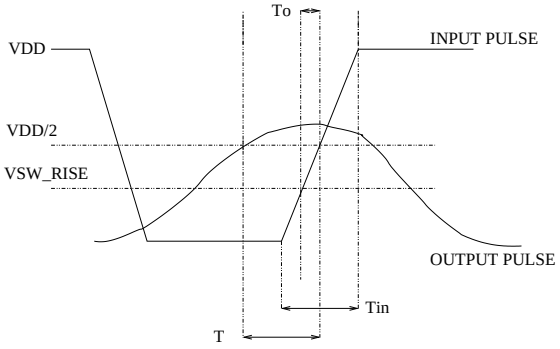

Figure 3. The Electrical Masking Effect. Degradation of a pulse passing through a transistor.

We constructed a model for electrical masking by integrating two existing models. We use the Horowitz rise and fall time model [11] to determine the rise and fall time of the output pulse, and the Logical Delay Degradation Effect Model [3] to determine the amplitude, and hence the duration, of the output pulse.

Horowitz rise and fall time model: The Horowitz model calculates the rise and fall time of the output pulse based on the gate switching voltages, CMOS model parameters and the input rise/fall time. The model is very sensitive to the values for the rise and fall switching voltages of the gates. We used an iterative bisection method to determine values for the switching voltages. This procedure adjusted the switching voltages until the rise and fall times predicted by the model were within 15% of values obtained from hspice simulations.

Delay degradation model: Delay degradation occurs when an input transition occurs before the gate has completely switched from its previous transition. The new input transition causes the gate to switch in the opposite direction leading to a degradation in the amplitude of the output pulse. We use the "Delay Degradation Model" proposed and validated by Bellidio-Diaz *et al.* [3] to determine how a voltage pulse degrades as it passes through a logic gate. This model determines the amplitude of the output pulse based on the time between the output transition and the next input transition, $(T - T_0)$, and the time needed for the gate to switch fully, which is proportional to $\frac{T_{in}}{3}$. These parameters are illustrated in Figure 3.

4.4 Pulse latching model

Recall that our definition of a soft error in combinational logic requires that an error pulse is captured in a memory circuit. In our model, this means that the pulse is stored into the level-sensitive latch at the end of a pipeline stage. We only consider a value to be stored in the latch if it is present (and stable) when the latch closes, since it is this value that is then passed to the next pipeline stage.

When a voltage pulse reaches the input of a latch, we use an hspice simulation to determine if it has sufficient amplitude and duration to be captured by the latch. The simulation is done in two steps. First we determine the pulse start time, the shortest time between the rising edge of the pulse and clock edge for which the pulse could be latched. This is similar to a setup time analysis for the latch, except that the input data waveform has the slope of the pulse at the latch input. The second step is to determine the minimum duration (measured at the threshold voltage) pulse that could be latched. For this step, we position the rising edge of the pulse at the point determined in the first step, and then vary the duration until the minimum value is determined. We studied the nature of the pulse start time and minimum duration using separate experiments and found that the pulse start time is a linear function of the rise time of the pulse, and the minimum duration is a linear function of the rise

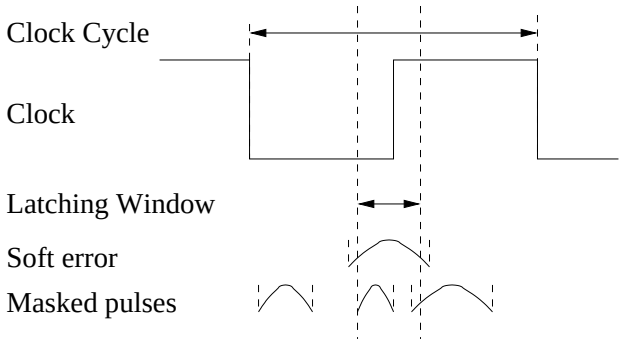

Figure 4. Latching Window Masking

time and fall time. For example, the pulse start time (in ps) of our pipeline latch in our 600nm technology can be computed as follows:

$$\text{start} = 54.02 + 0.42 \times t_{\text{rise}}$$

and the minimum duration (in ps) is given by

$$\text{duration} = 88.74 + 0.4 \times t_{\text{rise}} + 0.42 \times t_{\text{fall}}$$

In our method for computing SER for combinational circuits, the test to determine if a pulse can be latched is performed very frequently. Therefore, it is important that this test be done efficiently so that run times for the model are reasonable. The pulse start time and minimum duration given by these models correlate very highly with the pulse start time and minimum duration determined from hspice simulations, and therefore allow us to replace an expensive simulation run with a very inexpensive calculation without significant loss in accuracy.

4.5 Latching-Window Masking Model

A latch is only vulnerable to a soft error during a small window around its closing clock edge. The size of this *latching window* is simply the minimum duration pulse that can be latched, which depends on the pulse rise and fall time. A pulse that is present at the latch input throughout the entire latching window will be latched and causes a soft error. If a pulse partially overlaps the latching window, there is the possibility that it may also cause a soft error, since it could prevent the data from satisfying the latch setup and hold time requirements. We believe this is a secondary effect and therefore we have ignored it in our model. This simplification results in a more conservative estimate of SER. Figure 4 illustrates our model of latching window masking. Only a pulse that completely overlaps the latching window results in a soft error. If the pulse arrives after the latching window has opened, dies before the latching window closes, or does not have sufficient duration to cover the whole window, the pulse will be masked.

Let d represent the duration of the pulse on arrival at the latch input. We assume that the pulse can arrive at any point during the clock cycle with equal probability. Let w represent the size of the latching window corresponding to this pulse, and let c represent the clock cycle time. If a latching window for the latch starts after time t and ends before time $t + d$, the pulse is present at the latch input throughout the entire latching window and results in a soft error. In any other case the pulse is masked and no soft error occurs.

We can determine the probability that the pulse causes a soft error by computing the probability that a randomly placed interval of length d overlaps a fixed interval of length w within an overall interval of length c. This probability is given in by the following equation:

$$\Pr\{\text{soft error}\} = \begin{cases} 0 & \text{if } d < w \\ \frac{d-w}{c} & \text{if } w \leq d \leq c + w \\ 1 & \text{if } d > c + w \end{cases}$$

Note that when $d < w$, the probability of a soft error is zero, but this is not an effect of latching window masking, since the pulse does not have sufficient duration to be latched.

4.6 Estimating SER for Combinational Logic

Now we combine our models for electrical and latching-window masking with Hazucha & Svensson's method for estimating SER to obtain a method of estimating SER in combinational logic circuits. Given the feature size and degree of pipelining, the basic steps to compute SER for a combinational logic circuit are:

1. Compute the contribution to SER for each gate in the pipe stage, and
2. The total SER for the circuit is then the sum of the SER contributions computed in the previous step.

To compute the SER contribution for a given gate in the pipestage, we simulate a particle strike to the drain of the gate using our charge to current pulse model. We use our current pulse to voltage pulse model to determine the voltage pulse that is produced when the current pulse reaches the next transistor of the pipestage. Following this, our electrical masking model is used to determine the characteristics of voltage pulse when it reaches the latch input. We use the pulse-latching model to determine if the pulse that reaches the latch input has sufficient amplitude and duration to cause a soft error. If so, we compute SER for

this charge value using the model of Hazucha & Svensson, and the probability that soft error occurs using the latching-window masking model.

Recall that a key parameter to the Hazucha & Svensson model for SER is Q_{CRIT}, which is the smallest charge required to cause a soft error. In memory circuits, soft errors are essentially deterministic, in that no charge less that Q_{CRIT} can cause a soft error, and every charge of Q_{CRIT} or larger results in a soft error with probability 1.0. In combinational logic, we need to consider the probability of latching-window masking when computing SER for combinational logic. This is done by considering a range of charge values. The lower bound of this range is Q_{CRIT}, and the upper bound of the range is Q_{CMAX}, the smallest charge that has probability of 1.0 of being latched according to our latching-window masking model, or which has a probability within epsilon of all greater charge values. We then calculate the SER (using the model of Hazucha & Svensson) for m equally spaced charge values between Q_{CRIT} and Q_{CMAX} (we used $m = 20$ for the results presented in this paper). The values of the Q_S and T parameters for 600nm, 350nm and 100nm are taken directly from [8]. The Q_S and T values scale approximately linearly with technology in a log-log scale [7], so we determined the parameters for the remaining technologies from the curve obtained by fitting the existing points to a straight line in a log-log scale. The curve fitting was done using Matlab and the correlation coefficients were high enough for the errors to be insignificant. All our experiments use a value for the neutron flux of $F = 0.00565$, corresponding to sea level in New York City.

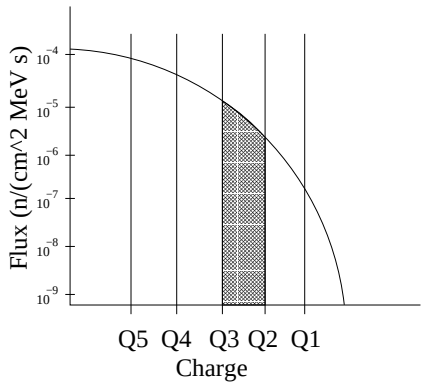

Figure 5. Computing SER using a range of charges with varying probability of latching. The contribution of the shaded region to overall SER is the SER for charges greater than Q_3 minus the SER for charges larger than Q_2, multiplied by the soft error probability associated with charge Q_3.

To compute the overall SER for the gate from these m SER values, we must determine the SER for each range of charges and then weight this SER by the probability a soft error occurs (e.g. is not masked by latching window masking). The SER value given by the Hazucha & Svensson model accounts for all charge values larger than the specified charge. Therefore, the SER for the range of charges from Q_{low} to Q_{high} is just $(SER(Q_{low}) - SER(Q_{high}))$. Thus, we compute the overall SER with the following formula:

$$SER = p_1 \times SER(Q_1) + \sum_{i=2}^{n} p_i \left(SER(Q_i) - SER(Q_{i-1})\right)$$

where Q_i are the m charge values arranged in decreasing order ($Q_1 = Q_{CMAX}$, and $Q_m = Q_{CRIT}$) and p_i is the probability that charge Q_i causes a soft error (is not latching-window masked). Note that since Q is monotonically decreasing, we have $p_i >= p_{i+1}$ for $1 \leq i < n$. This computation is illustrated in Figure 5.

5 Results

5.1 Memory Circuits

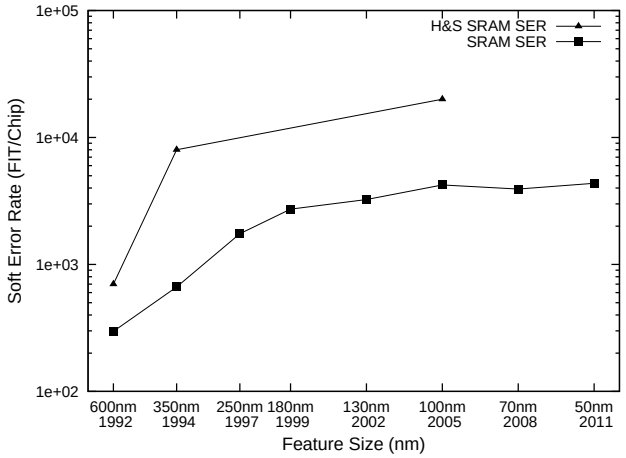

Figure 6. SER of a constant area SRAM array at different technologies

We estimated the SER of a constant area SRAM array using Hazucha & Svensson's model and our CMOS technology parameters. We used hspice simulations to determine Q_{CRIT} values for each technology. We simulated a current pulse at the drain of one node of the SRAM cell and sampled the cell later to see if the value had changed. Figure 6 presents our results, along with the results of a

similar experiment reported by Hazucha and Svensson [8]. Our results show good correlation with those of Hazucha and Svensson; both results show the same basic trend, and the absolute error is less than one order of magnitude for all technologies, which can be attributed to differences in CMOS parameters. The graph shows that the SER per chip is slightly increasing with decreasing feature size. There are four basic factors that combine to produce this trend. The drain area of each transistor, which is the region sensitive to particle strikes, decreases quadratically as feature size decreases, but since the SRAM array occupies a constant area, the number of bits increases quadratically and offsets this effect. Critical charge also decreases significantly with decreasing feature size, primarily due to lower supply voltage levels, but charge accumulation in the transistor also decreases and effectively offsets the reduction in critical charge.

5.2 Individual Elements

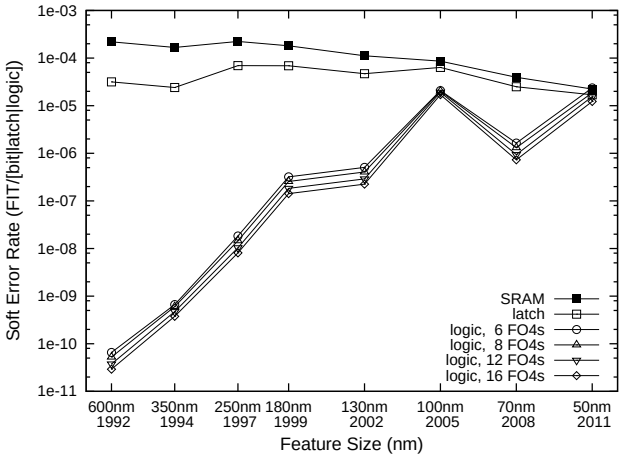

Figure 7. SER of a single SRAM cell/latch/logic chain for different feature sizes and pipeline depths

The circuits of a modern microprocessor fall into three basic classes: SRAM cells, latches, and combinational logic. We estimated the SER for an individual SRAM cell, latch, and logic chain using methodology described in Section 4. Figure 7 shows the predicted SER for a variety of feature sizes and pipeline depths. The x-axis plots the feature size of the CMOS technology, arranged by actual or expected date of adoption, and the y-axis plots the SER for each element on a log scale. The SER of a single SRAM cell declines gradually with decreasing device size, while the SER of a latch stays relatively constant. The SER for a single logic chain shows the most significant change – increasing over five orders of magnitude from 600nm to 50nm.

Reductions in feature size affect both Q_S and Q_{CRIT}, while changes in clock rate only affect Q_{CRIT}. The left graph of Figure 8 plots these parameters for the feature sizes and pipeline depths we consider in our study. The values shown are for NMOS devices. For combinational logic, the graph shows Q_{CRIT} values for a particle strike 0, 4, and 16 FO4 gate-delays from the latch.

Recall that the ratio Q_{CRIT}/Q_S is an exponent in the denominator of Equation 1. When this ratio is large, this factor dominates the SER model and produces a low SER value. When Q_{CRIT}/Q_S is small, this factor has much less influence on SER, and other factors such as the area of the sensitive region dominate. The right graph of Figure 8 shows the trend of this ratio with decreasing feature size. Note that all the curves appear to be asymptotically approaching 1.0.

SRAMs and Latches: The right graph of Figure 8 shows that Q_{CRIT}/Q_S of SRAMs is relatively small for all feature sizes, and decreases monotonically with feature size until 100nm, where it levels off at just over 1.0. As a result, the primary effect of device scaling on SER of a single SRAM cell is the reduction in sensitive area, leading to gradual downward trend shown in Figure 7. The Q_{CRIT}/Q_S ratio for latches is larger than for SRAMs at large feature sizes, but decreases more rapidly than SRAMs with decreasing feature size, and by 100nm has converged to almost the same value as SRAMs. This explains the relatively flat SER for a single latch shown in Figure 7. Device scaling in memory elements affects the critical charge and charge collection efficiency almost equally because smaller transistors are more sensitive to a particle strike but have very little sensitive volume for charge collection.

Combinational Logic: Device scaling has a significant effect on the SER of logic circuits. The transistors in the logic gates are typically wider and thus have greater capacitance than those used in SRAMs and latches. This greater capacitance reduces the size of the pulse generated by a strike at the node and so we expect the Q_{CRIT}/Q_S values to be much larger for logic. The 0 FO4 curve for logic plots Q_{CRIT}/Q_S for a particle strike just before the latch, and thus includes no electrical masking effect. This curve shows the same basic trend as SRAMs and latches, but is much larger at large feature sizes. From 600nm to 50nm, the Q_{CRIT}/Q_S ratio decreases by almost a factor of 10 for 0 FO4s of logic, compared to a factor of 5 reduction for latches and a factor of 3.5 reduction for SRAMs. This steep reduction in Q_{CRIT}/Q_S is primarily due to two factors. First, the output node capacitances associated with the gate

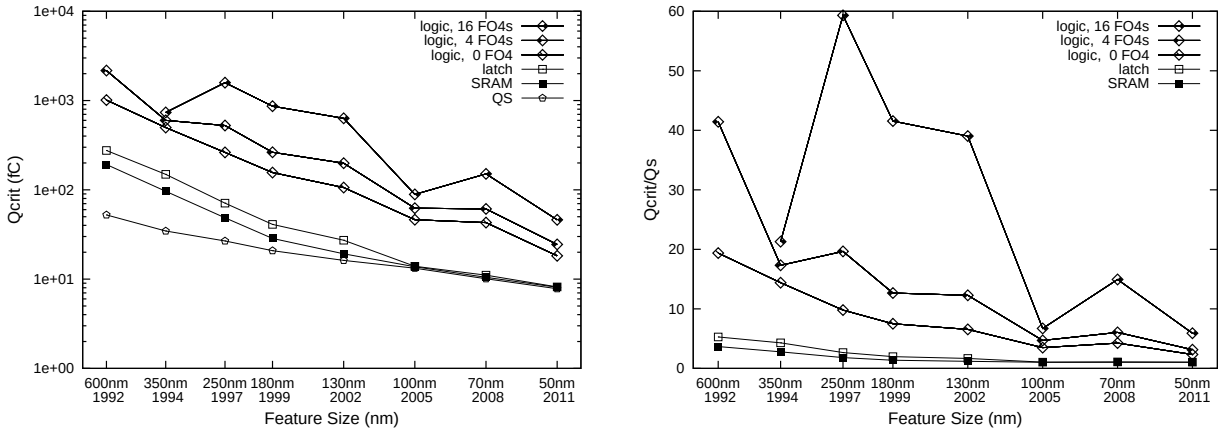

Figure 8. Critical charge for an SRAM cell/latch/logic chain by feature size and pipeline depth

decrease quadratically with feature size and so the gate becomes more sensitive. Second, the electrical masking effect is reduced significantly at smaller feature sizes. The difference in Q_{CRIT} for different number of FO4 gates within a single technology indicates the effect of superpipelining on electrical masking. From the left graph of Figure 8 the ratio of Q_{CRIT} values between 16 FO4 and 0 FO4 is about 6 at 250nm whereas it is only 3 at 50nm. These two factors contribute mainly to the increase in SER for combinational logic shown in Figure 7.

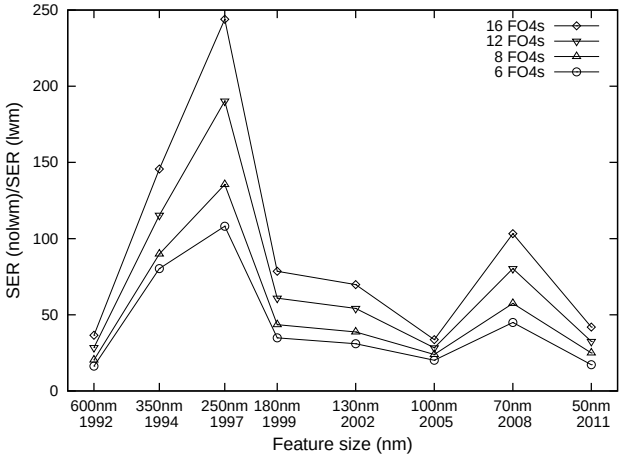

Figure 9. Effect of latching-window masking by feature size and pipeline depth.

To evaluate the effect of technology trends on latching window masking, we recomputed the SER of combinational logic with the assumption that any charge larger than Q_{CRIT} will result in a soft error. Then we divided by the original SER value to obtain a ratio that indicates the effect of latching window masking for a given feature size and pipeline depth. Figure 9 presents the results of this analysis. From the graph we can see that for each feature size the latching-window masking effect decreases with decreasing number of gates between latches. This is because at lower clock rates the latching window occupies a smaller fraction of the clock period. For a given pipeline depth latching window masking is only a function of the relative widths of the latching window and the error pulse at the latch. As we go to lower feature sizes latches have much shorter response times and so have smaller latching windows. The error pulse then has more avenues of overlap with the latching window and hence the masking probability is lower. For each pipeline depth the graph drops between 250nm and 50nm confirming the expected trend. The graph does show considerable variation at the other technologies and we think this is due to some inconsistencies in the underlying transistor models.

5.3 Processor SER

Our primary interest is not only in the reliability of individual gates and latches, but also in the processor as a whole. As feature sizes decrease, the number of transistors that can be placed on a fixed size die increases quadratically, creating significantly greater opportunity for soft errors. Since the rate of soft errors is different in SRAM cells, latches and logic, the SER of the processor will depend on the chip area devoted to each type of device. To estimate the SER rate of the total chip we have developed a chip model that describes the transistor decomposition into logic, SRAMs and latches. The chip model has mechanisms to redistribute the transistors as we move to deeper pipelines or lower feature sizes. From the chip model we determine the total number of SRAM bits, latches and pipeline stages and

then scale the per unit SER of each circuit by their number on the chip to obtain the SER/chip.

We used the Alpha 21264 microprocessor as the basis for constructing our chip model. The Alpha 21264 was designed for a 350nm process and has 15.2 million transistors on the die [15]. We performed detailed area analysis on die photos of the Alpha 21264 [14] to determine the number of transistors devoted to different structures on the chip. From our analysis we concluded that approximately 20% of transistors are in logic circuits and the remaining 80% are in storage elements in the form of latches, caches, branch predictors, and other memory structures. Our chip model applies this basic allocation to all feature sizes. The total number of transistors on the processor is scaled quadratically from the baseline Alpha 21264 based on feature size. Table 1 presents the total number of transistors per chip, and the transistors devoted to each circuit class for each technology based on this assumption.

The allocation of memory element transistors to SRAM cells and latches depends on the number of latches required by the processor pipeline, which depends on pipeline depth. We allocate one latch for each pipestage, where the number of pipestages is given by Equation 4.

$$\text{pipestages} = \frac{\text{logic\_transistors}}{\text{gates\_per\_pipestage} \times \text{transistors\_per\_gate}} \quad (4)$$

The remaining memory element transistors are allocated to SRAM cells. A typical SRAM bit requires 6 transistors, the level sensitive latch we use in our model consists of 6 transistors, and we assume each logic gate also uses 6 transistors. These assumptions are quite realistic and using slightly different values for these numbers will not affect the overall trend noticeably. Table 2 illustrates how our model allocates transistors to SRAM bits, latches, and logic gates for the 350nm feature size. The table shows the number of SRAM bits, latches, and logic gates, and the corresponding percentage of chip area, for the four pipeline depths we examine in this paper.

Using the SER of the individual elements presented in the previous section and our chip model, we computed the SER for each class of components for each feature size and pipeline depth of our study. The results are presented in Figure 10. As discussed above, SER/chip of SRAM shows little increase as feature size decreases. Furthermore, pipeline depth has no noticeable effect, since the percentage of chip area allocated to SRAM changes very little. SER/chip in latches increases only slightly for all pipeline depths, a combined effect of the relatively constant SER/latch and the increasing number of latches at smaller feature sizes. SER/chip of latches increases for deeper pipelines, due solely to the greater number of latches required for deeper pipeline microarchitectures.

SER/chip in combinational logic increases dramatically from 600nm to 50nm, from 10^{-8} to approximately 1.0, or eight orders of magnitude. This is simply the composition of a 10^5 increase in SER per individual logic chain and more than 100 increase in logic chains per chip. At 50nm and a 6 FO4 pipeline, the SER per chip of logic exceeds that of latches, and is within two orders of magnitude of the SER per chip of unprotected memory elements. For processors that use ECC to protect a large portion of the memory elements on the chip, logic will quickly become the dominant source of soft errors.

6 Discussion

The primary focus of our study has been to establish the basic trend in SER of combinational logic and the major influences on this trend. A number of other factors may have some influence on this trend but we excluded from our work to simply the modeling and analysis. This section discusses the most important of these factors and how they might affect the SER of combinational logic.

Circuit Implementations We restricted our analysis to static combinational logic circuits and level sensitive latches. Modern microprocessors frequently employ a much more diverse set of circuit styles, including dynamic logic, latched domino logic, edge-triggered flip flops, and a variety of latches designed for particular performance/power/area/noise margin tradeoffs. We believe our model could be extended to include any of these additional circuit styles. Our methodology can also be applied to heterogeneous gates in the logic chain if each gate type can be appropriately characterized.

The effect of dynamic logic may be particularly significant, since these circuits include maintain state within the gates themselves to prevent the output degrading due to leakage effects. This feature will probably reduce electrical masking significantly since each gate will reinforce the error pulse. In addition, dynamic logic typically uses smaller transistors, which have lower node capacitance making them more susceptible to soft errors.

Logical Masking Logical masking is another masking effect that inhibits soft errors in combinational logic and could have a significant effect on SER. Since we model a pipestage using a simple linear string of gates, we are actually modeling a minimal active path to the latch, which is the most conservative approximation of the logical masking effect. However, our model also places every logic gate on an active path to a latch, which understates the effect of logical masking. Thus, it is unclear how our results

Device size	Total	SRAM	Latches	Logic gates
600nm	5.17 M	4.07 M (78.8%)	0.06 M (1.2%)	1.03 M (20.0%)
350nm	15.2 M	11.9 M (78.8%)	0.19 M (1.2%)	3.04 M (20.0%)
250nm	29.7 M	23.4 M (78.8%)	0.37 M (1.3%)	5.95 M (20.0%)
180nm	57.4 M	45.2 M (78.8%)	0.71 M (1.3%)	11.4 M (20.0%)
130nm	110 M	86.7 M (78.8%)	1.37 M (1.2%)	22.0 M (20.0%)
100nm	186 M	146 M (78.8%)	2.32 M (1.2%)	37.2 M (20.0%)
70nm	380 M	299 M (78.8%)	4.75 M (1.2%)	76.0 M (20.0%)
50nm	744 M	586 M (78.8%)	9.31 M (1.2%)	148 M (20.0%)

Table 1. Transistors per chip for 16 FO4 pipeline using quadratic scaling assumption

Pipeline depth	SRAM bits	Latches	Logic gates
16 FO4s	1995 K (78.8%)	32 K (1.2%)	507 K (20.0%)
12 FO4s	1984 K (78.3%)	42 K (1.7%)	507 K (20.0%)
8 FO4s	1963 K (77.5%)	63 K (2.5%)	507 K (20.0%)
6 FO4s	1942 K (76.7%)	84 K (3.3%)	507 K (20.0%)

Table 2. Chip Model for 350nm device size

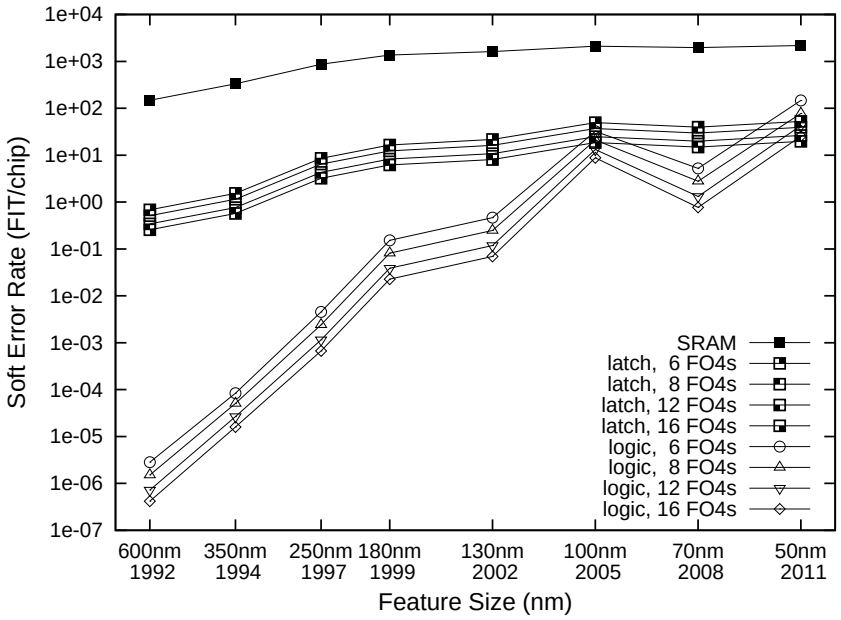

Figure 10. Total chip SER for logic, latches and SRAMs across different gate lengths and pipeline depths

might change if we extended our model to incorporate logical masking. Incorporating logical masking would likely increase the complexity of the model dramatically, since the model would need to consider actual circuits and associated inputs. Massengill et. al. developed a specialized VHDL simulator that could analyze soft faults in an actual circuit [17] and model the effects of logical masking. They found that effect of logical masking on SER depends heavily on circuit inputs.

Effects similar to logical masking can also occur in memory elements. For example, if a soft error occurs in a memory element that holds dead data – data that will not be used again – it is in some sense logically masked. Another example is a soft error in a memory structure such as

a branch predictor, which may lead to reduced performance but not produce incorrect results. Our methodology is consistent in that we do not consider logical masking effects for either memory elements or logic.

Finally, it seems unlikely that logical masking will be significantly affected by the technology trends we consider in this study. Device scaling provides more transistors on the processor die which may encourage more speculative processing, which could increase the potential for logical masking. Deeper pipelines will entail some increase in complexity of control path in the processor, leading to a slightly higher potential for logical masking. However, such effects are unlikely to have a significant effect on overall SER.

7 Related Work

Although this is the first paper to model the effect of both technology scaling and superpipelining on the soft error rate of combinational logic, previous experimental work has been done to estimate the soft error rate of storage and combinational logic in existing technologies [21, 5, 13, 16, 20].

IBM developed the SEMM (Soft-Error Monte Carlo Modeling) program [19] to determine whether chip designs meet SER specifications. The program calculates the SER of semiconductor chips due to ionizing radiation based on detailed layout, process information and circuit (Q_{CRIT}) values. The modified Burst generation rate (BGR) model [26] uses nuclear theory to calculate the reaction products of particle strikes and hence arrives at the collected charge as opposed to Hazucha's empirical model that we use in this paper which uses measured SER cross sections.

Some work has also been done to estimate the SER in combinational logic. Liden *et al.* compared the soft error rate due to direct particle strikes in latches with the soft error rate from error pulses propagating through the logic gates [16]. They considered a circuit implemented in 1000nm technology clocked at 5MHz. They conclude that the errors are predominantly due to direct strikes to latches and only 2% of the total observed errors are from the logic chain. The results in this paper show the same behavior at high feature sizes and low clock rates but the behavior is dramatically different at lower feature sizes because the masking effects are greatly reduced there. Baze *et al.* studied electrical masking in a chain of inverters and concluded that for pulses that successfully get latched electrical masking does not have any significant effect on SER [2]. They also allude to various parameters such as the chip model and the clock rate as factors that might affect the impact of this effect on overall SER. Buchner *et al.* investigated latching window masking in combinational and sequential logic [4]. They concluded that while the SER of sequential logic is independent of frequency, combinational logic SER increases linearly with clock rate. In this paper we model in detail the effect of electrical and latching window masking and illustrate the impact of technology and frequency scaling, and the chip model on the two effects.

Seifert *et al.* used experiments and simulation to determine the trend of soft error rate in the family of Alpha processors [24]. They conclude that the α particle susceptibility of both logic and memory circuits has decreased over the last few process generations. Our study shows an increasing susceptibility to neutron-induced soft errors, particularly in logic circuits, due to device scaling and greater neutron flux at lower energies [27]. They also found that the errors in combinational logic are predominantly due to direct strikes to pipeline latches, rather than error propagation in logic. Our simulations agree with this result at current feature sizes, but predict that SER of logic will approach SER of latches as feature sizes decrease. They also concluded that for a given feature size, clock rate has little influence on SER. This is also consistent with our results, as shown in Figure 10.

8 Conclusion

In this paper, we have presented an analysis of how two key trends in microprocessor technology, namely device scaling and superpipeling, will affect the susceptibility of microprocessor circuits to soft errors. The primary impact of device scaling is that the on-currents of devices decrease and circuit delay decreases. As a result, particles of lower energy can generate sufficient charge to trigger the device and cause a soft error. This result was demonstrated using a series of simulations of basic circuits in a range of device sizes. Since the flux of particles is substantially greater for particles of lower energy, it follows that all circuits will experience higher soft error rates due to device scaling.

There are also significant effects of device scaling and superpipeling that only apply to combinational logic circuits. These effects all work to reduce the masking phenomena that currently provide combinational logic with a form of natural protection against soft errors. Logical masking may be reduced due to the decreasing number of latches between gates. Latching-window masking will be reduced due to clock rate increases that result from both device scaling and superpipelining. Electrical masking will be reduced due to the decreasing number of latches between gates and because smaller circuits will result in less degradation of the pulse when it travels through a device.

Both storage elements and combinational logic will become more susceptible to soft errors due to the primary effect of device scaling, but only combinational logic is impacted by the reduction in masking phenomena. Thus, we

conclude that current technology trends will lead to a substantially more rapid increase in the soft error rate in combinational logic than in storage elements. The implication of this result is that further research is required into methods for protecting combinational logic from soft errors.

More recently, a number of schemes have been proposed to detect or recover from transient errors in processor computations. All these techniques are either based on space redundancy or time redundancy. DIVA [1] employs a simple "checker" to verify the results of instructions ready to be committed by the high performance core. Since the recomputations have both a spatial and temporal gap they will not be affected by the temporal or spatial locality of the particles. Both AR-SMT [23] and SRT [22] rely on a hardware approach called "simultaneous multithreading", in which a processor can execute multiple threads at the same time. The basic idea in both approaches is to execute instructions redundantly and then check that the results match before committing the result to architected state. The Out-Of-Order Reliable Superscalar (O3RS) approach [18] focuses on the issue of soft errors in instruction execution. Each instruction is executed twice, the first execution writes its result into the reorder buffer, and the second execution verifies its result against the first. We believe that techniques such as these combined with circuit and process innovations will be required to enable future construction of reliable high performance systems.

References

[1] Todd Austin. DIVA: A Reliable Substrate for Deep Submicron Microarchitecture Design. *International Symposium on Microarchitecture*, November 1999.

[2] M. Base and S. Buchner. Attenuation of Single Event Induced Pulses in CMOS Combinational Logic. *IEEE Trans. on Nuclear Science*, 44(6), December 1997.

[3] M. J. Bellido-Diaz, J. Juan-Chico, A. J. Acosta, M. Valencia, and J.L.Huertas. Logical modelling of delay degradation effect in static CMOS gates. *IEEE Proc-Circuits Devices Syst.*, 147(2), April 2000.

[4] S. Buchner, M. Baze, D. Brown, D. McMorrow, and J. Melinger. Comparison of Error Rates in Combinational and Sequential Logic. *IEEE Transactions on Nuclear Science*, 44(6), December 1997.

[5] H. Cha and J. H. Patel. A Logic-Level Model for α-Particle Hits in CMOS Circuits. In *1993 IEEE International Conference on Computer Design: VLSI in Computers and Processors (ICCD '93)*, pages 538–542, 1993.

[6] L. B. Freeman. Critical charge calculations for a bipolar SRAM array. *IBM Journal of Research and Development, Vol 40, No 1*, January 1996.

[7] Peter Hazucha. Background Radiation and Soft Errors in CMOS Circuits. *Linkping Studies in Science and Technology. Dissertations; 638*, 2000.

[8] Peter Hazucha and Christer Svensson. Impact of CMOS Technology Scaling on the Atmospheric Neutron Soft Error Rate. *IEEE Transactions on Nuclear Science, Vol. 47, No. 6*, Dec. 2000.

[9] Glenn Hinton, Dave Sager, Mike Upton, Darrell Boggs, Doug Carmean, Alan Kyker, and Patrice Roussel. The microarchitecture of the pentium 4 processor. *Intel Technology Journal*, 1, February 2001.

[10] Ron Ho, Kenneth W. Mai, and Mark A. Horowitz. The Future of Wires. In *Proceedings of the IEEE*, volume 89, pages 490–504, April 2001.

[11] Mark A. Horowitz. Timing Models For MOS Circuits. Technical Report SEL83-003, Integrated Circuits Labaratory, Stanford University, 1983.

[12] K. Johansson, P. Dyreklev, B. Granbom, M.C. Calvet, S. Fourtine, and O. Feuillatre. In-flight and ground testing of single event upset sensitivity in static RAM's. *IEEE Transactions on Nuclear Science*, 45:1628–1632, June 1998.

[13] T. Juhnke and H. Klar. Calculation of the soft error rate of submicron CMOS logic circuits. *IEEE Journal of Solid State Circuits*, 30:830–834, July 1995.

[14] Jim Keller. The 21264: A Superscalar Alpha Processor with Out-of-Order Execution. Microprocessor Forum, October 1996.

[15] R. E. Kessler. The Alpha 21264 Microprocessor. *IEEE Micro*, 19(2):24–36, March-April 1999.

[16] Peter Liden, Peter Dahlgren, Rolf Johansson, and Johan Karlsson. On Latching Probability of Particle Induced Transients in Combinational Networks. In *Proceedings of the 24th Symposium on Fault-Tolerant Computing (FTCS-24)*, pages 340–349, 1994.

[17] L. W. Massengill, A. E. Baranski, D. O. Van Nort, J. Meng, and B. L. Bhuva. Analysis of Single-Event Effects in Combinational Logic – Simulation of the AM2901 Bitslice Processor. *IEEE Trans. on Nuclear Science*, 47(6), December 2000.

[18] Avi Mendelson and Neeraj Suri. Designing high-performance and reliable superscalar architectures: The out of order reliable superscalar (o3rs) approach. *International Conference on Dependable Systems and Networks*, June 2000.

[19] P. C. Murley and G. R. Srinivasan. Soft-error Monte Carlo modeling program, SEMM. *IBM Journal of Research and Development, Volume 40, Number 1, 1996*, 1996.

[20] E.L. Peterson, P. Shapiro, J.H. Adams, and E.A. Burke. Calculation of cosmic-ray induced soft upsets and scaling in VLSI devices. *IEEE Transactions on Nuclear Science, Volume: 29 pp. 2055-2063*, December 1982.

[21] J.C. Pickel. Effect of CMOS miniaturization on cosmic-ray-induced error rate. *IEEE Transactions on Nuclear Science*, 29:2049–2054, December 1982.

[22] Steven K Reinhardt and Shubhendu Mukherjee. Transient Fault Detection via Simultaneous Multithreading. *International Symposium on Computer Architecture*, July 2000.

[23] Eric Rotenberg. AR/SMT: A Microarchitectural Approach to Fault Tolerance in Microprocessors. *International Symposium on Fault Tolerant Computing*, 1998.

[24] Normal Seifert, David Moyer, Norman Leland, and Ray Hokinson. Historical Trend in Alpha-Particle induced Soft Error Rates of the Alpha(TM) Microprocessor. In *Proceedings of the IEEE 39th Annual International Reliability Physics Symposium*, 2001.

[25] The International Technology Roadmap for Semiconductors. Semiconductor Industry Association, 1999.

[26] S.Satoh Y.Tosaka, H.Kanata and T.Itakura. Simple method for estimating neutron-induced soft error rates based on modified BGR method. *IEEE Elec. Dev. Lett., Vol. 20, pp. 89-91*, Feb 1999.

[27] J. Ziegler. Terrestrial cosmic ray intensities. *IBM Journal of Research and Development, Vol 42, No 1*, January 1998.