Impact of CMOS Technology Scaling on the Atmospheric Neutron Soft Error Rate

Peter Hazucha and Christer Svensson, *Member, IEEE*

***Abstract*—We investigated scaling of the atmospheric neutron soft error rate (SER) which affects reliability of CMOS circuits at ground level and airplane flight altitudes. We considered CMOS circuits manufactured in a bulk process with a lightly-doped p-type wafer. One method, based on the empirical model, predicts a linear decrease of SER per bit with decreasing feature size L_G. A different method, based on the MBGR model, predicts even faster decrease of SER per bit than linear. If the increasing number of bits is taken into account, then the SER per chip is not expected to increase faster than linearly with decreasing L_G.**

***Index Terms*—Circuit reliability, scaling, single event upset, soft error rate, technology characterization.**

I. Introduction

SCALING of MOS transistor dimensions is a key factor in the improvement of performance of CMOS technologies. For deep submicron technologies, the energy-delay product decreases approximately with the fourth power of the dimension scaling factor. This power efficiency can be attributed to a quadratic reduction of switching charge, linear increase in speed, and linear decrease in supply voltage. At the same time, transistor area decreases quadraticaly. The trend shows a slow increase in die area which means that the number of transistors increases at least quadraticaly [1].

In this paper, we analyze scaling of soft error rate (SER) in CMOS technologies. We consider circuits manufactured in a bulk CMOS process and operating in the natural environment from ground level to airplane flight altitudes (<60 kft). At ground level, there are three major contributors to SER. The first component are alpha particles emitted by decaying radioactive impurities in packaging and interconnect materials. Alpha particles are amenable to theoretical and experimental investigation and several scaling studies have already been published [2]–[5]. The second component are atmospheric neutrons with energies below 1 MeV, which interact with the ^{10}B isotope present in boro-phosphosilicate glasses and p-type regions [6], [7]. Modeling of SER due to thermal neutrons benefits from extensive research in nuclear energy and weapons. Well established nuclear models and experimental nuclear cross section data are available. The analysis is also simplified by a limited number of reaction channels. The third component are atmospheric neutrons with energies >1 MeV. Theoretical modeling of neutron SER in this energy region is very complicated. The number of reaction channels is relatively large, but very few neutron nuclear cross sections have been measured. Historically, the atmospheric neutrons have been believed to be dominating at high altitudes, and for circuits which are not sensitive to alpha particles [8], [9]. The question of relative importance of fast neutrons with respect to thermal neutrons is still open. The sum of atomic numbers of secondary particles from a reaction of thermal neutrons with ^{10}B can not exceed eleven, and these secondary particles tend to have lower LET and energy compared to the secondaries from a ^{28}Si(n, x) reaction. This observation suggests that effects of thermal neutrons may be important for circuits with critical charges close to the alpha particle threshold for a particular technology.

In this paper, we investigate scaling of SER caused by fast neutrons of the atmospheric spectrum, i.e. with energies >1 MeV. Alpha particles and thermal neutrons may be equally important, although they are not considered in this paper. Effects of the three types of radiation are additive and the present work can be extended in future to include alpha particles and thermal neutrons.

It has been argued, that the rapidly decreasing switching charge coupled to the increasing number of transistors per die results in increasing SER per chip. One of the first studies of SER caused by background radiation at ground level was published by Wallmark *et al.*in 1962 [10]. Their conclusion was that scaling will be limited either by excessive heat generation or by SER caused by cosmic rays. Indeed, excessive heat generation has become a severe problem, especially in high performance microprocessors, and has been one of the reasons for reduction of supply voltage. Unfortunately, lower supply voltage results in lower switching charge, which increases SER. If only dimensions were scaled, but not voltage, then the decrease in charge would be linear, instead of quadratic. In 1982, Pickel studied SER in the galactic cosmic environment [11]. For bulk CMOS processes, SER per bit of an SRAM was found to remain constant, or increase at approximately linear pace, depending on the assumptions about funneling. A similar study by Petersen [12] found approximately constant SER per bit. Due to increase in the number of bits per die (see the beginning of this Section), SER per chip can be expected to increase quadraticaly. For alpha particles, both SOI and bulk SRAMs showed approximately constant SER per bit, for feature sizes from 0.4 to 0.19 μm [4]. Another alpha particle study of a 16-bit multiplier found increase by a factor of ten for scaling from 0.6 μm to 0.12 μm [3]. In [5], alpha SER has been shown to increase by three orders of magnitude from 0.25 μm to 0.05 μm. However, it is not clear, if the authors considered

Manuscript received July 24, 2000. This work was suppored by the Swedish Foundation for Strategic Research and the Intel Corporation.

The authors are with the Department of Physics and Measurement Technology, Linköping, University, SE-581 83 Linköping, Sweden (e-mail: {pehta; chs}@ifm.liu.se).

Publisher Item Identifier S 0018-9499(00)11196-7.

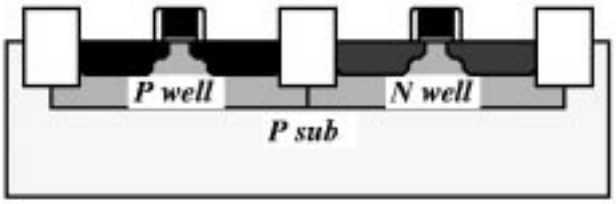

Fig. 1. Simplified cross section of a twin-well CMOS process on a P type substrate.

other effects besides reduction of critical charge, e.g., reduction of collection area, increase in the number of transistors and die size, etc. In [13], atmospheric neutron SER for several generations of DRAMs and bipolar SRAMs was reported. The DRAMs exhibited an improvement of a factor of 2000 per bit, but the critical charge in DRAMs does not change very much with the feature size. SER of the bipolar SRAMs did not show any increase. As pointed out by the authors, this trend does not indicate that the intrinsic circuit susceptibility to soft errors did not change. Rather, it shows a concern of the manufacturer with the improvement of SER reliability. Similarly, the experimental data on SRAMs [9], [14]–[16] indicate that for the past ten years, the neutron SER cross section has been between 10^{-14} and 10^{-12} cm^2/bit, corresponding to sea-level SER from 200 to 2^*10^4 FIT/Mbit (unit FIT = Failure-In-Time gives the number of errors in 10^9 hours).

The goal of this study was to establish a relation between the atmospheric neutron soft error rate and technology feature size. Our approach is different from the previous scaling studies [3]–[5], [11], [12] which modeled SER from first principles. For fast neutrons such an approach is very complicated. We start with a detailed SER characterization of a single technology. For this purpose we will use a verified empirical model [17]. With help of finite-element device simulations the empirical model will be translated to other technologies. It should be noted that this translation is much easier than development of a fully predictive model, because it does not require modeling of nuclear reactions. SER susceptibility of several technology generations will be compared.

Recently, Tosaka *et al.* published an easy-to-use Modified Burst Generation Rate (MBGR) model for calculation of atmospheric neutron SER [18]. We will perform an independent scaling study using the MBGR model. Scaling trends derived from the empirical model and from the MBGR model will be compared.

II. Technology Assumptions

Fig. 1 shows a cross section of a twin-well CMOS process on p-type substrate. Technologies for high-performance logic applications make use of lightly-doped substrates or substrates with a lightly-doped epitaxial layer grown on top of a highly-doped substrate (SOI technologies are not discussed in this paper). A twin-well process allows an independent optimization of doping profiles for NMOS and PMOS transistors. SER depends mainly on doping profiles underneath the drain and source of a MOSFET [19] which is sometimes explained by the sensitive (collection) depth dependence on the substrate doping concentration [18]. Drain extensions close to the channel have to be shallow in order to suppress short-channel effects. However, this increases parasitic drain resistance and junction capacitance. Low junction capacitance implies a wider depletion region and lower doping density, which is not compatible with another requirement on high punch-through voltage between the source and drain. Below 0.35 μm, contacts have been salicided. Too shallow extensions make salicidation difficult and may lead to increased junction leakage. For best performance, shallow extensions together with additional implants (e.g., a halo) are fabricated close to the channel, and deep extensions with a wider depletion region are used under the contacts. The shallow extensions occupy typically 20–30% of drain area.

TABLE I
Supply Voltage V_{CC} for Logic CMOS Circuits

L_G [μm]	0.8	0.6	0.35	0.1
V_{CC} [V]	5/3.3	5/3.3	3.3/2.5	1.2/0.9

In this study, we investigate four different technology generations. We refer to the technologies by their drawn gaten length L_G, which is 0.8, 0.6, 0.35, or 0.1 μm. All technologies use a lightly-doped p-type wafer as a starting material (Fig. 1). For L_G equal 0.8 and 0.6 μm, the doping profiles were obtained from Austria Mikro Systeme (AMS). The profiles used for drain and well were approximately gaussian with peak concentrations at the top of the silicon substrate. Buried layers, retrograde doping, or thin epitaxial layers were not employed in these technologies. For more information on these technologies as well as the exact doping profiles the interested reader is advised to contact the AMS [20]. Doping profiles for L_G equal 0.35 and 0.1 μm were derived from the AMS profiles. With decreasing L_G, all other distances, e.g., well thickness and depletion region width, have to scale proportionately. This was accomplished by rising the dopant concentrations and decreasing the decay lengths of the profile distributions.

Layout dimensions for L_G from 0.8 to 0.35 μm followed the AMS design rules. For $L_G = 0.1$ μm, the dimensions were scaled from the AMS 0.35 μm process. For L_G from 0.8 to 0.35 μm, BSIM3 transistor models [21] and diode models were provided by the AMS. For $L_G = 0.1$ μm, the models were scaled from a 0.13 μm technology [21]. These models were used in circuit simulations with HSPICE.

For each technology, supply voltage V_{CC} for high-performance and low-power applications is given in Table I. Low-power circuits operate at reduced V_{CC}. Identical V_{CC} was used for the 0.8 and 0.6 μm technologies.

III. Scaling Theory for the Atmospheric Neutron SER

A. Introduction

Neutron-induced soft errors are generated by secondary charged particles which are created in collisions of neutrons with silicon atoms (collisions with ^{10}B are not considered here). The angular and energy distribution of the secondary

particles is therefore *independent* of technology generation. A soft error occurs if collected charge Q exceeds critical charge Q_{CRIT} of a circuit node. Magnitude of Q for a given particle strike depends on doping profiles of drain and substrate, and voltages applied on the drain and well junctions. For a given technology and circuit node, Q_{CRIT} depends on V_{CC}, and whether the strike occurs on a p- or n-type drain. For many circuits, e.g., SRAMs, Q_{CRIT} depends on the shape of the charge collection waveform. For a given particle strike, the waveform shape depends on doping, and voltage on the drain and well junctions. We consider all these effects separately and find their impact on neutron SER.

We start with the empirical model calibrated and verified for $L_G = 0.6$ μm. For this technology, the dependence of atmospheric neutron SER on the drain type, V_{CC}, and Q_{CRIT} was measured. Also, effective values for time constants of the collection waveforms were extracted from measurements. From device simulations, we will find the dependence of collected charge Q and the waveform time constant on doping, drain type, and voltage. This will allow us to translate the empirical model to technologies with L_G equal 0.1, 0.35, and 0.8 μm. We will determine Q_{CRIT} from circuit simulations, and calculate atmospheric neutron SER of a six-transistor SRAM cell for each technology. Finally, we will repeat the calculation using the MBGR model and compare both scaling trends.

B. Empirical Model for the 0.6 μm Technology

The empirical model was presented in [17]. For a given circuit node, the atmospheric neutron cross section CS is calculated as

$$CS(V_{CC}, A, Q_{CRIT}) = A \times p_{ENV}(V_{CC}, Q_{CRIT}(V_{CC}, \overline{s}_{\mathrm{EFF}})) \quad (1)$$

where

V_{CC} is the supply voltage,
A is the drain area,
Q_{CRIT} is the critical charge, and
p_{ENV} is a dimension-less function.

For $L_G = 0.6$ μm, p_{ENV} was measured in [17] (Fig. 2). The measurements were carried out with the WNR neutron beam at Los Alamos National Laboratory, NM. The energy spectrum of the WNR beam is very similar to the atmospheric spectrum and is often used for accelerated measurements of the atmospheric neutron SER [14]–[18]. In this way, the atmospheric neutron SER cross section can be measured directly.

For a given V_{CC} and drain type, an exponential dependence was fitted to p_{ENV}.

$$p_{ENV} = K \times \exp\left(-\frac{Q_{CRIT}}{Q_S}\right) \quad (2)$$

Extrapolation to $Q_{CRIT} = 0$ gives $p_{ENV} = K$. Cross section at this point is determined by the rate at which the secondary particles intercept the drain, rather than by the amount of collected charge. Therefore, K should not depend on V_{CC} or doping profiles. Fig. 2 shows that the exponentials with a common point of interception provide a good fit to p_{ENV} at all conditions. Collection slope Q_S depends strongly on doping and V_{CC}.

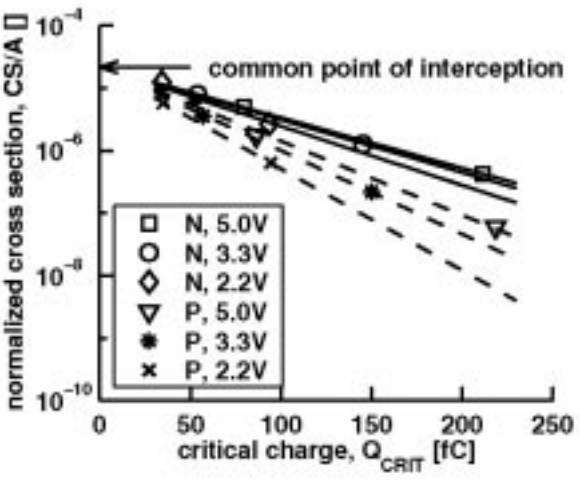

Fig. 2. Normalized atmospheric neutron cross section for the 0.6 μm technology plotted for different V_{CC} and drain types (N or P) [17]. The cross section data were measured at WNR, Los Alamos, NM.

TABLE II
MODEL PARAMTERS FOR $L_G = 0.6$ μm AND $V_{CC} = 5$ V

N type drain			P type drain		
K []	Q_S [fC]	T [ps]	K []	Q_S [fC]	T [ps]
$2.2*10^{-5}$	54	175	$2.2*10^{-5}$	37	125

Q_{CRIT} depends on V_{CC} and parameter $\overline{s}_{\mathrm{EFF}}$ describing the time-dependence of the collection waveform. A one-parameter waveform

$$I(t) = \frac{2}{T\sqrt{\pi}}\sqrt{\frac{t}{T}}\exp\left(\frac{-t}{T}\right) \quad (3)$$

from [22] was used for approximating the shape of collection waveforms. Effective parameter $\overline{s}_{EFF} = T$ was measured separately for p- and n-type drains. Table II gives the empirical model parameters for $L_G = 0.6$ μm and $V_{CC} = 5$ V.

C. Empirical Model for Other Technologies

Translation of the empirical model to other technology requires finding the collection slope Q_S and time constant T for n- and p-type drains. Parameter K has same value for all technologies. If Q_{CRIT} approaches zero then virtually any secondary particle intercepting a sensitive drain causes a soft error. Therefore, the normalized cross section (which equals K for $Q_{CRIT} = 0$) should not depend on the substrate doping profiles (see also Section III-B. and Fig. 2). Fig. 3 shows the transformation of slope Q_S.

The baseline technology is the 0.6 μm technology at zero bias. When the doping profiles change from $L_G = 0.6$ μm to some other technology, then Q_S is multiplied by factor f_D for a given technology. Next, when voltage is increased for a given technology from 0 V to V_{CC}, then Q_S is multiplied by factor

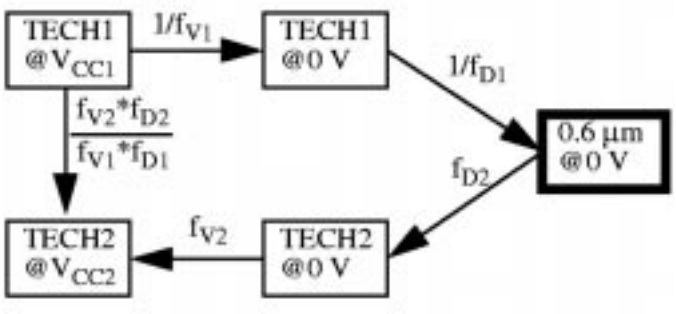

Fig. 3. Transformation of charge collection slope Q_S from technology $TECH1$ to technology $TECH2$.

f_V, which depends on L_G, drain type, and V_{CC}. This gives the overall slope transformation as

$$Q_{S2} = \frac{f_{V2}}{f_{V1}} \times \frac{f_{D2}}{f_{D1}} \times Q_{S1}, \tag{4}$$

where Q_{S1} and Q_{S2} give slope Q_S for technologies $TECH1$ and $TECH2$, respectively. It should be noted, that f_{V2}/f_{V1} and f_{D2}/f_{D1} give impacts of scaling of V_{CC} and doping, respectively. These simple transformations can work only if there is a strong correlation between collected charges for $TECH1$ and $TECH2$. Additionally, the relation should be linear in order to preserve the exponential dependence (2). We performed device simulations of identical particle strikes on a drain diode for each technology and drain type. The transistor drain was modeled as a cylindrical diode. We used DESSIS simulator from the ISETCAD simulation package [23]. The rotational symmetry of the structure and particle track allowed running DESSIS with a two-dimensional cylindrical coordinate system [24]. Despite great reduction in the simulation time and improvement in mesh resolution this approach did not introduce a significant error because the transformation of slope Q_S makes use of a ratio of collected charges under various conditions rather than the absolute value of collected charge. We activated Shockley–Reed–Hall and Auger recombination, doping-dependent mobility model, high-field velocity saturation and carrier–carrier scattering. A reaction of a neutron with silicon can result only in secondary particles ranging from hydrogen to silicon. As a representative set, we selected ^{4}He, ^{9}Be, ^{12}C, ^{16}O, ^{20}Ne, ^{25}Mg, and ^{28}Si. Each particle was simulated twice, with energy corresponding to ranges of 5 and 10 μm in silicon. The particles were incident vertically in the middle of a drain, and propagated downward. The track LET profiles were obtained from TRIM [25]. The radial dependence was modeled by a gaussian with a track radius of 0.04 μm. The time dependence was modeled by a gaussian with a standard deviation of 1 ps. The peak of generation rate occurred at 10 ps. For each drain type, technology, and bias conditions, 14 different tracks were simulated (7 particles by 2 ranges).

Collected charge for the NMOS and PMOS transistors is shown in Figs. 4 and 5, respectively. Collected charge for $L_G = 0.8$ μm is larger than for $L_G = 0.6$ μm because of lower doping, which enhances funneling and diffusion. For $L_G = 0.35$ and 0.1 μm, collected charge is much smaller than for $L_G = 0.6$ μm. These simulations were carried out at zero bias.

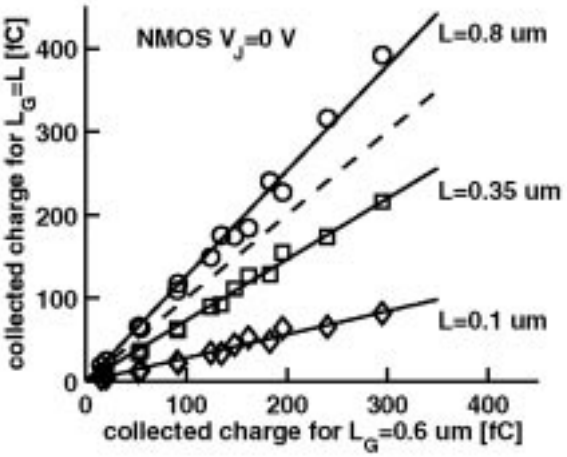

Fig. 4. Collected charge on the outside-the-well NMOS drain. Technologies with smaller feature size collect significantly less charge from the same particle strike. Relation between collected charges for two different technologies is linear. Slope of the lines equals to $f_{D,N}$.

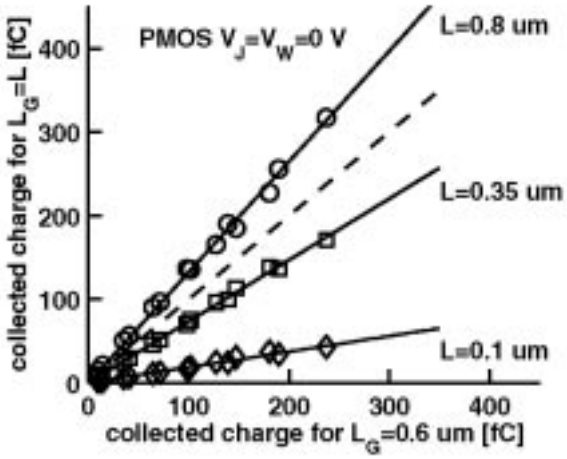

Fig. 5. Collected charge for the inside-the-well PMOS transistor. Both drain and well junction were biased at zero volts. Slope of the lines equals to $f_{D,P}$.

According to Fig. 3, slope of the straight lines equals to f_D for each technology. For $L_G = 0.6$ and 0.8 μm, we ran also particles with 2 μm range. Additionally, all strikes were simulated with a particle starting in the substrate and propagating vertically upward in the direction of a drain. Although the number of simulations increased from the initial 14 to 42, little new information was gained. For $L_G = 0.35$ and 0.1 μm, these additional simulations were omitted to save time. One simulation takes typically 10 hours on a 450 MHz PC with Linux.

In order to find f_V, all simulations were run at increased voltage. For the outside-the-well NMOS transistor, the particles were run with the maximum technology supply voltage $V_J = V_{\mathrm{MAX}}$ across the drain junction. Because the boundary conditions change during a strike as more charge is collected on a circuit node, we interpolated collected charge to $V_J = V_{CC}/2$. Voltage V_{MAX} equals to V_{CC} for high-performance technologies.

The situation is more complicated for the inside-the-well PMOS transistor. Here, collected charge is a function of drain

TABLE III
SLOPE TRANSFORMATION FOR THE HIGH-PERF. TECHNOLOGIES. SUBSCRIPTS N AND P REFER TO N AND P TYPE TRANSISTORS, RESPECTIVELY

L_G [μm]	0.8	0.6	0.35	0.1
V_{CC} [V]	5	5	3.3	1.2
$f_{D,N}$ []	1.26	1.00	0.73	0.28
$f_{V,N}$ []	1.13	1.25	1.15	1.13
$Q_{S,N}$ [fC]	62	54	36	14
$f_{D,P}$ []	1.32	1.00	0.73	0.19
$f_{V,P}$ []	0.92	0.90	0.86	0.76
$Q_{S,P}$ [fC]	50	37	26	5.8

TABLE IV
SLOPE TRANSFORMATION FOR THE LOW-POWER TECHNOLOGIES. SUBSCRIPTS N AND P REFER TO N AND P TYPE TRANSISTORS, RESPECTIVELY

L_G [μm]	0.8	0.6	0.35	0.1
V_{CC} [V]	3.3	3.3	2.5	0.9
$f_{D,N}$ []	1.26	1.00	0.73	0.28
$f_{V,N}$ []	1.08	1.16	1.12	1.10
$Q_{S,N}$ [fC]	59	50	35	13
$f_{D,P}$ []	1.32	1.00	0.73	0.19
$f_{V,P}$ []	0.95	0.93	0.90	0.82
$Q_{S,P}$ [fC]	51	38	27	6.2

junction voltage V_J as well as the well junction voltage V_W. We performed additional simulations with $V_J = 0$ V and $V_W = V_{\rm MAX}$, and again with $V_J = V_{\rm MAX}$ and $V_W = V_{\rm MAX}$. Together with the simulations for $V_J = 0$ V and $V_W = 0$ V, we found collected charge for $V_W = V_{CC}$ and $V_J = V_{CC}/2$ using interpolation in two dimensions. The impact of change in voltages V_J and V_W from 0 V to their final values is given by f_V. According to Fig. 3, factor f_V accounts for the change of circuit bias from 0 V to full supply voltage V_{CC} within the same technology. f_V was found by correlating the collected charges for $V_W = 0$ V and $V_J = 0$ V, and $V_W = V_{CC}$ and $V_J = V_{CC}/2$ in much the same way as the collected charges from different technologies at zero bias were correlated in Figs. 4 and 5.

Tables III and IV give the values of f_D, f_V, and Q_S, for high-performance and low-power technologies, respectively. As expected, f_D for $L_G = 0.6$ μm equals 1. Recall from the previous discussion that K is technology-independent. Its value is given in Table II. Subscripts P and N are used to distinguish between values for p- and n-type drains, respectively.

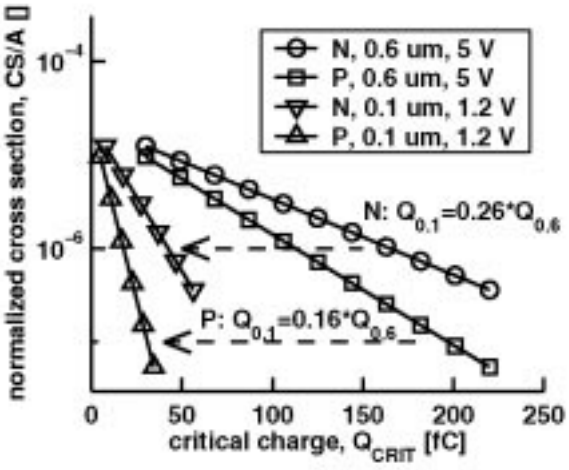

Fig. 6. Transformation of cross sections from the 0.6 μm to the 0.1 μm technology. For $L_G = 0.1$ μm, $Q_S = Q_{0.1}$. For $L_G = 0.6$ μm, $Q_S = Q_{0.6}$.

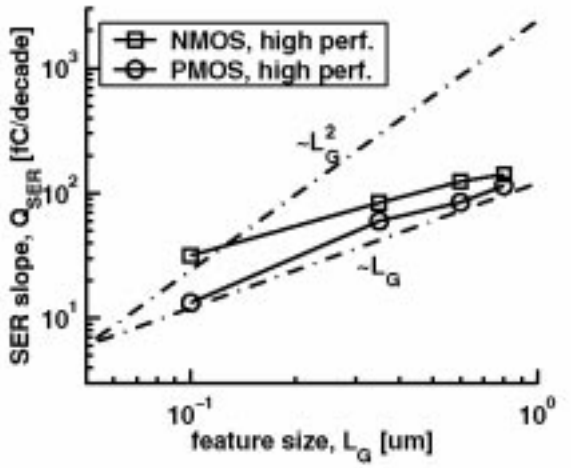

Fig. 7. Scaling of charge collection slope $Q_{\rm SER}$ with feature size L_G.

To clarify the impact of the slope transformation, Fig. 6 gives the transformed cross sections for $L_G = 0.1$ μm and $V_{CC} = 1.2$ V calculated from (4) and values given in Table III. There is significant decrease in the cross section for $L_G = 0.1$ μm. In other models, this decrease is accounted for by smaller sensitive depth and lower value of a BGR function for smaller sensitive depth [18]. It is interesting to compare scaling of charge collection slope Q_S with scaling of L_G. Rather than plotting Q_S directly, we calculated charge $Q_{\rm SER}$ needed to reduce SER by one order of magnitude.

$$Q_{\rm SER} = Q_S \times \ln(10) \tag{5}$$

Fig. 7 shows that $Q_{\rm SER}$ scales approximately linearly with L_G. Identical result was obtained also for the low-power technologies.

The next task is to scale the effective time constant T of the collection waveforms. From the same set of simulations, we found a strong correlation between the time constants for the same particle strike in different technologies. Instead of fitting

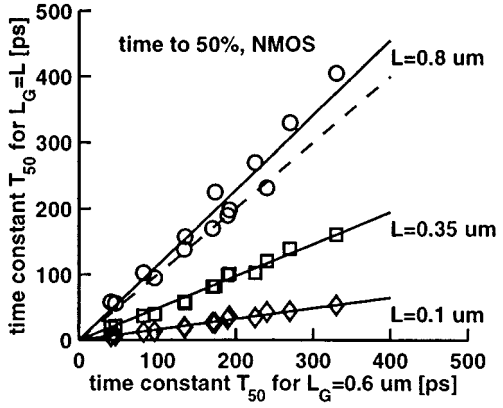

Fig. 8. Time to 50% of collected charge for the NMOS transistor. For smaller feature size L_G, collection of charge is faster. Slope of the lines equals to $f_{T,\,N}$.

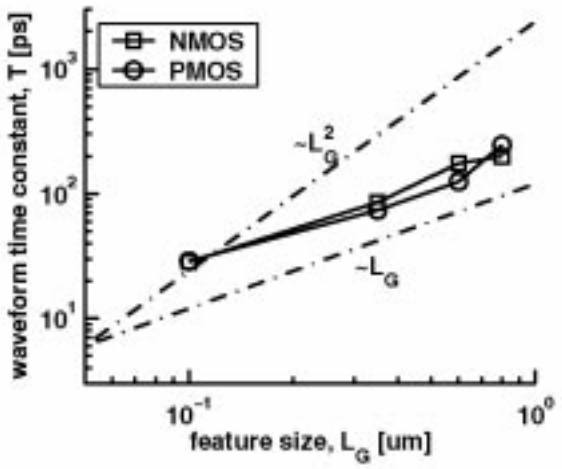

Fig. 9. Scaling of collection waveform time constant T with feature size L_G.

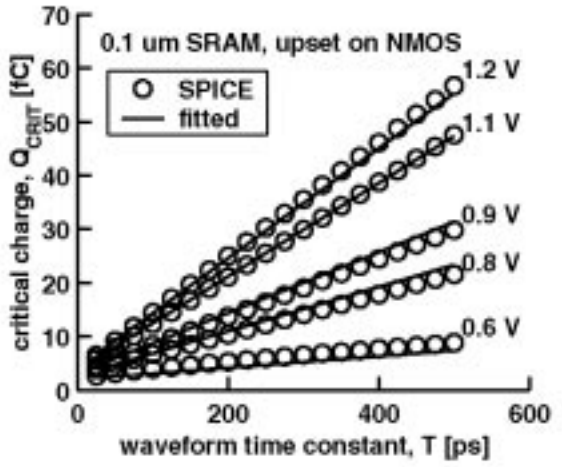

Fig. 10. Critical charge Q_{CRIT} of a SRAM cell. A three-parameter expression can fit the dependence well for all voltages and time constants.

the waveform in (3) to the simulated transients, we determined time T_{50} for collection of 50% of charge. As long as the shape of a collection waveform resembles the shape described by (3) for some parameter T, any extraction method will result in approximately the same value of T. As the shape starts to deviate from (3) the definition of T becomes meaningless. This can occur, e.g., for highly-doped substrates with thin epitaxial layers. Doping profiles in our case (Section II) were such that (3) provided a reasonably good fit to the waveforms obtained from the simulations and changing the criterion from 50% to either 30% or 70% had little effect on the results. Then the time constant T in (3) can be found as

$$T = 0.84 \times T_{50}. \tag{6}$$

Unfortunately, time dependence of current transients depends strongly on the strike location and activated mobility models. Therefore, we used the device simulations to determine the relative change in T when changing the doping profiles and V_{CC}. In analogy with (4), T was transformed as

$$T_2 = \frac{f_{T2}}{f_{T1}} \times T_1. \tag{7}$$

Factor f_T gives the relative change in T_{50} and T compared to the 0.6 μm technology. We neglected the dependence of T_{50} on V_{CC} and give a single value for each L_G corresponding to $V_J = V_{\mathrm{MAX}}/2$ and $V_W = V_{\mathrm{MAX}}$. This approximation does not introduce a significant error because of a second-order impact of T on SER.

Fig. 8 shows a correlation of T_{50} for identical particle strikes on the NMOS drain in different technologies. As L_G scales down, the charge collection process becomes faster. For each technology and transistor type, the time constants calculated from (7) are given in Fig. 9. Together with the waveform in (3), they will be used for finding critical charge Q_{CRIT}. It is interesting that the time constants decrease approximately linearly with L_G. The same trend follow also the propagation delays of logic circuits. Therefore, we do not expect any fundamental change in the circuit response to the collection waveform because of scaling.

It is important to realize that the device simulations were used to find the relative change of T after changing doping and bias. Since the value of T was determined experimentally for $L_G = 0.6$ μm (Section III-B), we were able to determine the value of T for other technologies as well. The value of T obtained in this way has been defined in [17]. Even from Fig. 8 it is obvious that using a single value for a time constant to describe waveforms with varying particle species, locations, and bias would be impossible without some means of averaging. A mathematical basis for such procedure was also given in [17].

D. Scaling of the Critical Charge

Critical charge Q_{CRIT} of a six transistor SRAM cell is a function of V_{CC} and time constant T. Q_{CRIT} was determined from circuit simulations with HSPICE. For $L_G = 0.1$ μm, this dependence is shown in Fig. 10. In order to simplify interpolation, we found it convenient to fit the dependence to a three-parameter expression

$$Q_{CRIT}(V_{CC}, T) = C_0 \left(V_{CC} + (V_{CC} - V_{CC0}) \frac{T}{T_0} \right). \tag{8}$$

For $T = 0$ ps, Q_{CRIT} equals $V_{CC}^* C_0$. Therefore, C_0 can be understood as the effective value of node capacitance. The dependence on time constant T is given by T_0. The increase

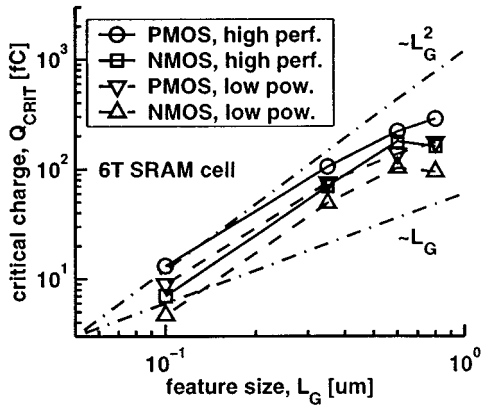

Fig. 11. Scaling of critical charge Q_{CRIT} of a SRAM cell.

TABLE V
FITTING PARAMETERS FOR Q_{CRIT} OF A SRAM CELL

L_G [μm]	0.8	0.6	0.35	0.1
$C_{0,N}$ [fF]	16.6	16.5	12.9	3.42
$V_{CC0,N}$ [V]	1.54	1.60	1.20	0.53
$T_{0,N}$ [ps]	148	102	85	22
$C_{0,P}$ [fF]	18.6	18.3	15.1	4.73
$V_{CC0,P}$ [V]	0.87	0.76	0.74	0.38
$T_{0,P}$ [ps]	98	76	52	15

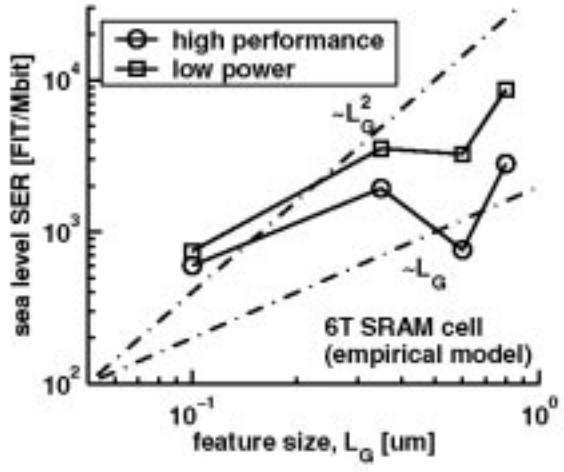

Fig. 12. Scaling of the atmospheric neutron SER at sea level. SER is calculated for a constant number of bits, which corresponds to a shrink of older design into a new process.

in Q_{CRIT} with increasing T is due to the feedback transistor which attempts recovery of the upset node. The longer the transients lasts, the longer is the recovery transistor open and pulls the node voltage back to its original value. Because drain current is larger at higher V_{CC}, there is a factor $V_{CC} - V_{CC0}$. This results in larger increase in Q_{CRIT} with T at higher V_{CC}. Value of V_{CC0} is related to the threshold voltage for a given technology. Fig. 10 shows a fit for $LG = 0.1$ μm. For each technology and transistor type, the fitting parameters are given in Table V. As could be expected, C_0, V_{CC0}, and T_0 decrease with decreasing L_G. For the NMOS transistor, V_{CC0} and T_0 are larger than for the PMOS transistor, because of larger drain current. Sizes of the NMOS and PMOS transistors were equal. Fig. 11 shows the values of Q_{CRIT} for the time constants T in Fig. 9 and corresponding V_{CC}. From 0.8 to 0.6 μm, there is little difference in Q_{CRIT} because V_{CC} was not scaled, and these technologies had similar node capacitances (Table V). Also, lower threshold voltage for $L_G = 0.6$ μm gives larger Q_{CRIT}. From 0.6 to 0.1 μm, Q_{CRIT} decreased almost quadraticaly with decreasing L_G.

E. Scaling of the Collection Area

For the 0.8–0.35 μm technologies, collection area A in (1) was scaled according to the AMS design rules. This trend was extrapolated to $LG = 0.1$ μm. Total reduction in A from 0.8 to 0.1 μm was a factor of 40, i.e., somewhat less than a factor of 64 that would be expected according to linear scaling of all dimensions. Apparently, contact size does not scale as L_G^2.

F. Scaling of the Soft Error Rate

Equations (1) and (2) can be used for calculation of the atmospheric neutron cross sections. The cross sections in Fig. 2 were obtained experimentally from measurements with an atmospheric-like neutron beam and were defined for the neutron flux F with energy >10 MeV. At sea level $F = 0.00565$ n*cm^{-2}s^{-1} (New York City, see [26]). Then, SER at sea level can be calculated as

$$\mathrm{SER} = F \times CS. \tag{9}$$

Fig. 12 shows scaling of SER calculated from the empirical model. The SER values decrease proportionately to L_G. If V_{CC} for $L_G = 0.8$ μm followed the trend from $L_G = 0.1$ to 0.6 μm, then SER per Mbit would be rather constant between 0.35 and 0.8 μm. Compare the low-power 0.35 and 0.6 μm technologies and the high-performance 0.8 μm technology. The rapid decrease of SER per Mbit for $L_G = 0.1$ is a result of the reduction in the collection slope Q_{SER} (Fig. 7), which results in much lower normalized cross section (Fig. 6). Furthermore, the normalized cross section has to be multiplied by sensitive area which decreases almost quadraticaly. These two factors together have larger influence on SER than the reduction in Q_{CRIT} (Fig. 11). Therefore, SER per Mbit decreases for smaller L_G.

The trend in Fig. 12 is typical for a design which was shrunk from an older technology. A circuit, e.g., a memory, which was re-designed to exploit the full potential of a new technology has usually a larger number of bits. If we instead consider an SRAM chip with constant die size of ~1.5 cm^2, then the number of bits increases with decreasing L_G. Fig. 13 shows, that SER for a scaling scenario assuming constant chip size increases approximately linearly with decreasing L_G. Predicted SER at sea level for $L_G = 0.1$ μm is about 2*10^4 FIT, which is equivalent to mean time-to-failure (MTTF) of 5.7 years. At flight altitudes ~40 kft, MTTF decreases ~300 times, i.e., we can expect

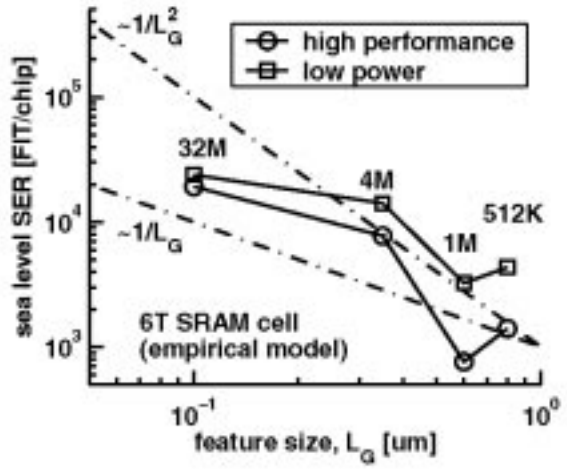

Fig. 13. Scaling of the atmospheric SER at sea level for a constant die size ($\sim$1.5 cm^2). Memory capacity increases to 32 Mbit for $L_G = 0.1$ μm.

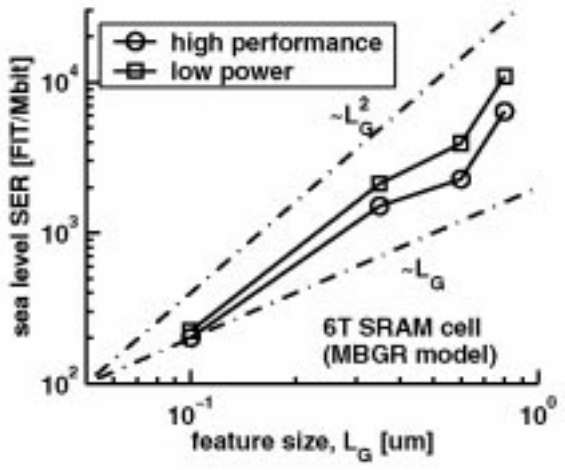

Fig. 14. Scaling of the atmospheric neutron SER at sea level according to the MBGR model. SER is calculated for a constant number of bits. Compare with Fig. 12.

one error per chip in 170 flight hours. Johansson *et al.* [9], observed in-flight SER of 4.5*10^{-3} upset per hour for a 4 Mbit SRAM. From the given neutron flux of 1.4 to 2 n*cm^{-2}s^{-1} [9] the SER at New York City would be approximately 12 700 to 18 200 FIT/chip which is in excellent agreement with our 4 Mbit predictions in Fig. 13. Unfortunately, we have no further information about the measured device except that it was operated at 5 V.

IV. Comparison with the Modified BGR Model

It would be interesting to compare the predicted trends from the empirical model with some other model. The MBGR model [18] is different from our model in several aspects. It uses nuclear theory for calculation of secondary particle spectra while our model uses measured SER cross sections. Collected charge in the MBGR model is calculated by a Monte Carlo technique using a compact charge collection model and the concept of sensitive depth. Our model uses finite-element device simulations to translate the measured charge collection spectra, and therefore, does not need to involve the sensitive depth in the calculations. The MBGR model accounts only for charge collected by funneling, the diffusion term is not included [27]. Our model, derived from measurements, makes use of total charge, including charge collected by diffusion. Finally, Q_{CRIT} in the MBGR model is defined as C*V_{CC}, where C is node capacitance. In our model, we account for the waveform dependence of Q_{CRIT} which is defined as total charge in the collection waveform. Both methods of Q_{CRIT} calculation are correct, at least for SRAMs. The MBGR model neglects diffusion and at the same time neglects the increase of Q_{CRIT} due to the diffusion tail of the collection waveform. Our model includes the diffusion charge as well as a waveform with a diffusion tail, and Q_{CRIT} is defined as total charge in the waveform. For other circuits, e.g., DRAMs, we expect that the MBGR model underestimates SER, because it neglects the diffusion charge.

The MBGR model requires several input parameters [18]. Collection area was taken same as drain area. The sensitive depth was calculated according to [18] for given V_{CC}, and the value for the doping concentration was taken at the metallurgical

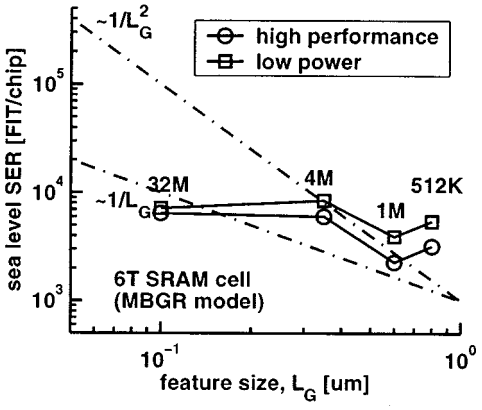

Fig. 15. Scaling of the atmospheric neutron SER at sea level according to the MBGR model assuming constant die size ($\sim$1.5 cm^2). Compare with Fig. 13.

drain junction. Q_{CRIT} was calculated from (8), but instead of using the values from Fig. 9, time constant T was set to zero (see the discussion in the previous paragraph). SER for a constant number of bits is shown in Fig. 14.

The absolute values agree well for L_G from 0.35 to 0.8 μm. For $L_G = 0.1$ μm, the difference is about a factor of 3. A more important observation is that the MBGR model predicts faster decrease of SER per bit than the empirical model, approximately proportional to the square of L_G. If we account for the increasing number of bits, we find that SER per chip is approximately constant, or slowly increasing (Fig. 15). Calculations for $L_G = 0.1$ μm involved several extrapolations. Therefore, the uncertainty can be on the order of a factor of 3 or even larger (in both models). Good agreement between the two models suggests that the results can be trusted within a factor of 3.

V. Conclusion

We investigated future development of the atmospheric neutron SER at sea level. Two different approaches were used. The first one was based on the empirical model derived from measurements and finite-element device simulations. The second approach was based on the theoretical MBGR model.

Soft error rate on per bit basis decreases at least linearly with decreasing feature size L_G. If we account for the increasing number of bits per die then SER per chip does not increase faster than linearly with decreasing L_G. This is caused by rapid decrease of collected charge in highly-doped substrates, smaller reverse voltage on parasitic drain diodes, and reduced collection area. Same scaling rules apply to neutron SER at airplane flight altitudes.

It should be noted that there exists a multitude of silicon CMOS technologies with identical L_G (e.g., bulk based on a n- or p-type wafer, twin/triple/retrograde well). This proliferation most probably results in different SER for different manufacturing processes. The scaling trends derived in this paper apply to the technologies described in Section II which follow predictable scaling trends. Unpredictable trends, e.g., a sudden change from a twin well to a triple well are not considered.

In the future work, it would be interesting to perform a scaling study on alpha particle and thermal neutron SER. To avoid repetition of our work by other researchers, our data was not normalized. Additionally, we included expressions for critical charge Q_{CRIT}, collection slope Q_S, and time constant T for the considered technologies.

ACKNOWLEDGMENT

The authors would like to thank L. Snith and R. Minixhofer from AMS for their support and technology information.

REFERENCES

[1] "International technology roadmap for semiconductors, 1999 edition,", http://www.sematech.org/public/index.htm.
[2] T. C. May and M. H. Woods, "Alpha-particle-induced soft error in dynamic memories," *IEEE Trans. Elec. Dev.*, vol. 26, pp. 2–9, Jan. 1979.
[3] T. Juhnke and H. Klar, "Calculation of the soft error rate of submicron CMOS logic circuits," *IEEE J. Solid-State Circuits*, vol. 30, pp. 830–834, July 1995.
[4] Y. Tosaka, K. Suzuki, and T. Sugii, "Alpha particle induced soft errors in submicron SOI SRAM," in *1995 Symp. VLSI Technol., Dig. Tech. Pap.*, pp. 39–40.
[5] N. Cohen, T. S. Sriram, N. Leland, D. Moyer, S. Bulter, and R. Flatley, "Soft error considerations for deep-submicron CMOS circuit applications," in *Technical Digest of Intern. Elec. Dev. Meet. (IEDM)* , 1999, pp. 315–318.
[6] P. J. Griffin, T. F. Luera, F. W. Sexton, P. J. Cooper, S. G. Karr, G. L. Hash, and E. Fuller, "The role of thermal and fission neutrons in reactor neutron-induced upsets in commercial SRAMs," *IEEE Trans. Nucl. Sci.*, vol. 44, pp. 2079–2086, Dec. 1997.
[7] X. W. Zhu, L. W. Massengill, C. R. Cirba, and H. J. Barnaby, "Charge collection modeling of thermal neutron products in fast submicron MOS devices," *IEEE Trans. Nucl. Sci.*, vol. 46, pp. 1378–1384, Dec. 1999.
[8] T. J. O'Gorman, J. M. Ross, A. H. Taber, J. F. Ziegler, H. P. Muhlfeld, C. J. Montrose, H. W. Curtis, and J. L. Walsh, "Field testing for cosmic ray soft errors in semiconductor memories," *IBM J. Res. Dev.*, vol. 40, pp. 41–49, Jan. 1996.
[9] 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 Trans. Nucl. Sci.*, vol. 45, pp. 1628–1632, June 1998.
[10] J. T. Wallmark and M. Marcus, "Minimum size and maximum packing density of nonredundant semiconductor devices," in *Proc. IRE*, March 1962, pp. 286–298.
[11] J. C. Pickel, "Effect of CMOS miniaturization on cosmic-ray-induced error rate," *IEEE Trans. Nucl. Sci.*, vol. 29, pp. 2049–2054, Dec. 1982.
[12] E. L. Petersen, P. Shapiro, J. H. Adams, and E. A. Burke, "Calculation of cosmic-ray induced soft upsets and scaling in VLSI devices," *IEEE Trans. Nucl. Sci.*, vol. 29, pp. 2055–2063, Dec. 1982.
[13] J. F. Ziegler *et al.*, "IBM experiments in soft fails in computer electronics (1978–1994)," *IBM J. Res. Dev.*, vol. 40, pp. 3–16, Jan. 1996.
[14] E. Normand, D. L. Oberg, J. L. Wert, J. D. Ness, P. P. Majewski, S. A. Wender, and A. Gavron, "Single event upset and charge collection measurements using high energy protons and neutrons," *IEEE Trans. Nucl. Sci.*, vol. 41, pp. 2203–2209, Dec. 1994.
[15] E. Normand, "Extensions of the burst generation rate method for wider application to proton/neutron-induced single event effects," *IEEE Trans. Nucl. Sci.*, vol. 45, pp. 2904–2914, Dec. 1998.
[16] C. A. Gossett, B. W. Hughlock, M. Katoozi, G. S. LaRue, and S. A. Wender, "Single event phenomena in atmospheric neutron environments," *IEEE Trans. Nucl. Sci.*, vol. 40, pp. 1845–1852, Dec. 1993.
[17] P. Hazucha, C. Svensson, and S. A. Wender, "Cosmic ray soft error rate characterization of a standard 0.6 μm CMOS process," *IEEE J. Solid-State Circuits*, Oct. 2000.
[18] Y. Tosaka, H. Kanata, S. Satoh, 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.
[19] D. Burnett, C. Lage, and A. Bormann, "Soft-error-rate improvement in advanced BiCMOS SRAMs," in *1993 IEEE Int. Rel. Phys. Sym.*, 1993, pp. 156–160.
[20] "AMS homepage,", http://www.amsint.com/index.html.
[21] , http://www-device.EECS.Berkeley.EDU/research.html.
[22] L. B. Freeman, "Critical charge calculations for a bipolar SRAM array," *IBM J. Res. Dev.*, vol. 40, pp. 119–129, Jan. 1996.
[23] "ISE homepage,", http://www.ise.com.
[24] P. E. Dodd, "Device simulation of charge collection and single-event upset," *IEEE Trans. Nucl. Sci.*, vol. 43, pp. 561–575, Apr. 1996.
[25] J. F. Ziegler, , http://www.research.ibm.com/ionbeams/.
[26] ——, "Terrestrial cosmic rays," *IBM J. Res. Dev.*, vol. 40, pp. 19–39, Jan. 1996.
[27] L. D. Edmonds, "A simple estimate of funneling-assisted charge collection," *IEEE Trans. Nucl. Sci.*, vol. 38, pp. 828–833, Apr. 1991.