Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration

Keith A. Bowman, *Member, IEEE*, Steven G. Duvall, *Member, IEEE*, and James D. Meindl, *Life Fellow, IEEE*

***Abstract*—A model describing the maximum clock frequency (FMAX) distribution of a microprocessor is derived and compared with wafer sort data for a recent 0.25-μm microprocessor. The model agrees closely with measured data in mean, variance, and shape. Results demonstrate that within-die fluctuations primarily impact the FMAX mean and die-to-die fluctuations determine the majority of the FMAX variance. Employing rigorously derived device and circuit models, the impact of die-to-die and within-die parameter fluctuations on future FMAX distributions is forecast for the 180, 130, 100, 70, and 50-nm technology generations. Model predictions reveal that systematic within-die fluctuations impose the largest performance degradation resulting from parameter fluctuations. Assuming a 3σ channel length deviation of 20%, projections for the 50-nm technology generation indicate that essentially a generation of performance gain can be lost due to systematic within-die fluctuations. Key insights from this work elucidate the recommendations that manufacturing process controls be targeted specifically toward sources of systematic within-die fluctuations, and the development of new circuit design methodologies be aimed at suppressing the effect of within-die parameter fluctuations.**

***Index Terms*—Critical path delay variations, die-to-die and within-die fluctuations, FMAX distribution, gate delay variations, inter-die and intra-die fluctuations, manufacturing tolerances, maximum clock frequency distribution, parameter variations, technology projections.**

I. Introduction

INTEGRATED circuits have always been vulnerable to inherent die-to-die (inter-die) and within-die (intra-die) parameter fluctuations in the manufacturing process. Die-to-die parameter fluctuations resulting from lot-to-lot, wafer-to-wafer, and a portion of the within-wafer variations affect every element on a chip equally. Conversely, within-die parameter fluctuations consisting of both random and systematic components produce a nonuniformity of electrical characteristics across the chip [1].

Examples of the lot-to-lot and wafer-to-wafer variations include processing temperatures, equipment properties, wafer polishing, and wafer placement. The within-wafer variations have contributions to both die-to-die and within-die fluctuations. An example of the within-wafer variations that impact the die-to-die fluctuations is the resist thickness across the wafer, which is random from wafer to wafer, but deterministic within the wafer. The aberrations in the stepper lens are an example of systematic within-die variations. As an example of random within-die fluctuations, the placement of dopant atoms in the device channel region, which is an intrinsic effect since it cannot be eliminated by external control of conventional manufacturing processes [2], varies randomly and independently from device to device.

Traditionally, die-to-die fluctuations have been the main concern in CMOS digital circuit designs, and the within-die fluctuations have been neglected [1], [3]. As polysilicon gate lengths have decreased below the wavelength of light used in the optical lithography process, however, the systematic and random within-die fluctuations of channel length have exceeded the die-to-die fluctuations [1]. Thus, within-die fluctuations are a growing threat to the performance and functionality of future gigascale integration (GSI).

The importance of accurately estimating the impact of parameter fluctuations on circuit performance is directly related to a company's overall revenue. An overestimation increases the design complexity, possibly leading to an increase in design time, an increase in die size, rejection of otherwise good designs, and even missed market windows [1]. Conversely, an underestimation can compromise the product's performance and overall yield as well as increase the silicon debug time [1]. In summary, overestimating fluctuations impacts the design effort, and underestimating fluctuations impacts the manufacturing effort.

This work demonstrates that the magnitude of both die-to-die and within-die parameter fluctuations significantly influence a processor's maximum clock frequency (FMAX) distribution [4], a measurement performed at wafer sort in which each functional die is tested for its maximum operating clock frequency. In Section II, a model describing the FMAX distribution [4] is presented and compared with wafer sort data for a recent 0.25-μm microprocessor. Employing the FMAX distribution model and a generic critical path model based on physical device and circuit analyses [5]–[7], a new circuit-level methodology, presented in Section III, enables projections, given in Section IV, of the impact of die-to-die and within-die parameter fluctuations on future processor performance [8]. Section V concludes by summarizing the key insights and recommendations. The research objective of this paper is to evaluate the limitations imposed by die-to-die and within-die

Manuscript received March 15, 2001; revised September 24, 2001. This work was supported in part by the Semiconductor Research Corporation, the Defense Advanced Research Projects Agency, and the Intel TCAD Group.

K. A. Bowman was with the Georgia Institute of Technology, Atlanta, GA 30332 USA. He is now with the Intel Corporation, Hillsboro, OR 97124 USA (e-mail: keith.a.bowman@intel.com)

S. G. Duvall was with the Intel Corporation, Santa Clara, CA 95054 USA. He is now with Intel Australia Pty. Ltd., North Sydney, Australia.

J. D. Meindl is with the Georgia Institute of Technology, Atlanta, GA 30332 USA.

Publisher Item Identifier S 0018-9200(02)00663-7.

0018–9200/02$17.00 © 2002 IEEE

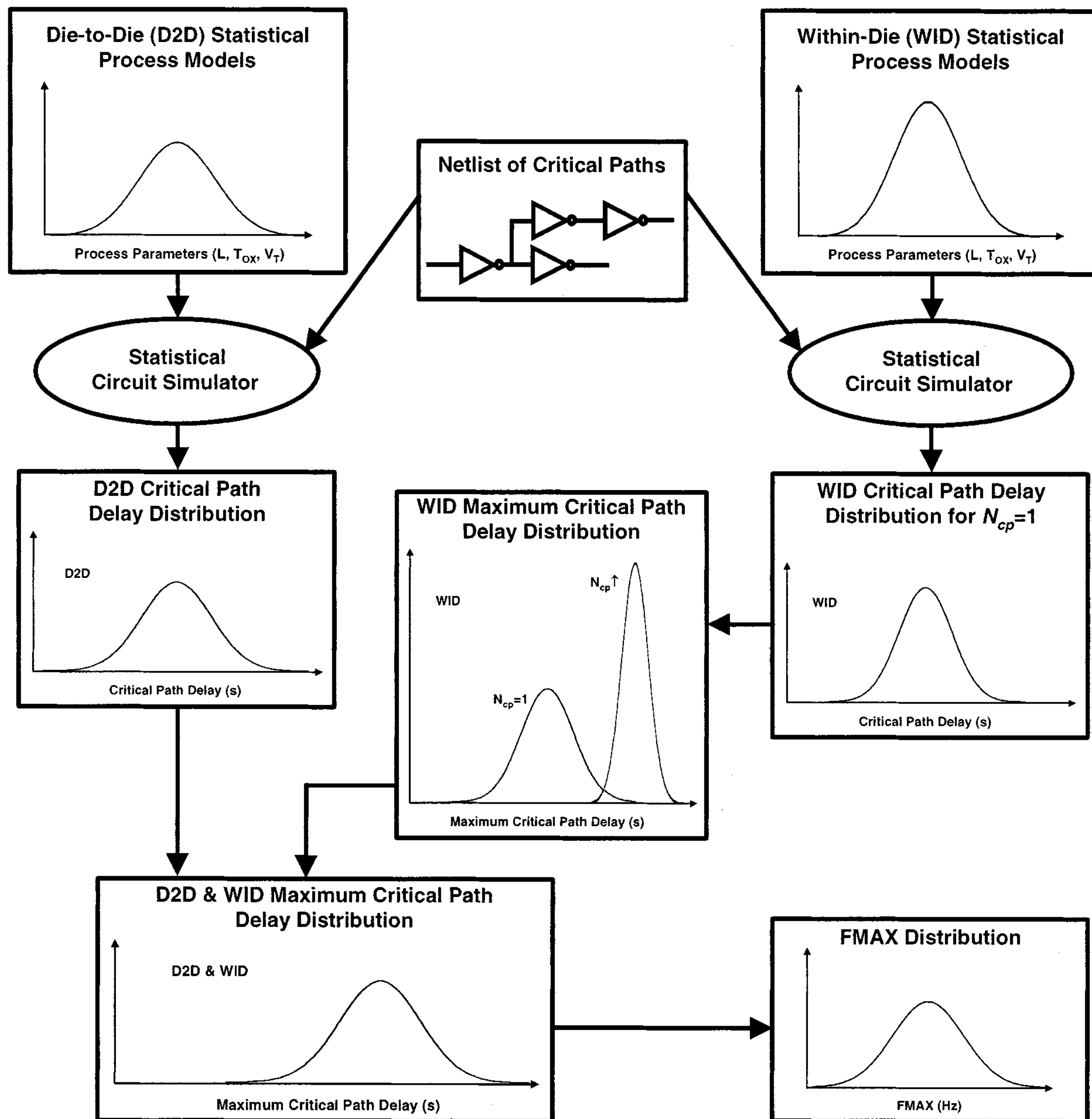

Fig. 1. Flowchart for describing the FMAX distribution. $N_{\rm cp}$ is the number of independent critical paths on a chip.

parameter fluctuations on a product's performance to facilitate opportunities for further advancement of GSI systems.

II. FMAX Distribution Model

An overview of the FMAX distribution model is presented in Fig. 1 and described in detail in the following subsections. First, in Section II-A, the individual contributions of die-to-die (D2D) and within-die (WID) fluctuations on the nominal critical path delay distribution are determined by simulating representative speed-limiting paths for a specific microprocessor using D2D and WID process models based on measured data. The simulated critical path delay distribution resulting from WID fluctuations is the distribution of one specific critical path. In Section II-B, a number $N_{\rm cp}$ of independent critical paths for the chip is estimated to calculate the within-die maximum critical path delay distribution for the entire chip. Since D2D fluctuations affect each critical path on a chip equally, the D2D maximum critical path delay distribution is represented by the D2D nominal critical path delay distribution. Next, the two maximum critical path delay distributions resulting from D2D and WID fluctuations are statistically combined in Section II-C, and then mapped to a frequency distribution in Section II-D. The FMAX distribution model is compared to measured data, and key insights are offered in Section II-E.

A. Impact of Die-to-Die and Within-Die Fluctuations on the Critical Path Delay Distribution

The FMAX distribution model is based upon statistical simulations of critical paths for a 0.25-μm microprocessor [4]. Die-to-die fluctuations are simulated using statistical process files, which are generated by mapping the electrical-test data

	Path 1	Path 2	Path 3	Nominal Path
Normalized μ_{Tcp}	1.00	0.77	0.51	**1.00**
D2D: σ_{Tcp}/μ_{Tcp} (%)	8.63	8.59	9.74	**8.99**
WID: σ_{Tcp}/μ_{Tcp} (%)	2.65	3.19	3.32	**3.05**

Fig. 2. Statistical summary of the die-to-die (D2D) and within-die (WID) fluctuations on three critical paths for a 0.25-μm microprocessor as well as the nominal path.

to model parameters. Within-die fluctuations are simulated through models calibrated with data obtained from a WID-variation test chip. Fig. 2 summarizes the statistical simulations of three critical paths by providing the mean delay $(\mu_{T_{\rm cp}})$ and the ratio of the standard deviation to the mean delay $(\sigma_{T_{\rm cp}}/\mu_{T_{\rm cp}})$ corresponding to D2D and WID fluctuations. The mean critical path delays are different for the three simulated paths, since some circuits require execution in less than one clock cycle. The nominal mean critical path delay $T_{\rm cp,nom}$ is assumed equal to the longest path delay. The D2D and WID nominal critical path standard deviations, $\sigma_{{\rm D2D}-T_{\rm cp,nom}}$ and $\sigma_{{\rm WID}-T_{\rm cp,nom}}$, respectively, are calculated individually by averaging the ratio of the standard deviation to mean path delay for all three simulated paths. Using the nominal mean and standard deviations provided in Fig. 2, the critical path delay density functions resulting from D2D and WID parameter fluctuations are modeled as normal distributions

$$f_{{\rm D2D}-T_{\rm cp,nom}} = N\left(T_{\rm cp,nom}, \sigma^2_{{\rm D2D}-T_{\rm cp,nom}}\right) \quad (1)$$

and

$$f_{{\rm WID}-T_{\rm cp,nom}} = N\left(T_{\rm cp,nom}, \sigma^2_{{\rm WID}-T_{\rm cp,nom}}\right) \quad (2)$$

respectively.

B. Impact of Within-Die Fluctuations on the Maximum Critical Path Delay Distribution

The impact of WID fluctuations on *one* critical path is described in (2). The probability of one critical path satisfying a specified maximum delay $t_{\rm max}$ is calculated as

$$P_{{\rm WID}-T_{\rm cp,nom}}(t \le t_{\rm max}) = F_{{\rm WID}-T_{\rm cp,nom}}(t_{\rm max}) = \int_0^{t_{\rm max}} f_{{\rm WID}-T_{\rm cp,nom}}(t)dt \quad (3)$$

where t is the variable critical path delay. $F_{{\rm WID}-T_{\rm cp,nom}}$ is the WID cumulative distribution for one critical path. A chip, however, contains *many* critical paths, all of which must satisfy the worst-case delay constraint [2]–[4], [9]. The paths may be completely dependent (correlation = 1), independent (correlation = 0), or some correlation between 0 and 1. If two paths are completely dependent, only one distribution is required to model the worst-case delay for both paths. If two paths, however, are not completely dependent, both paths must be statistically combined to obtain the worst-case delay. Assuming a number $N_{\rm cp}$ of independent critical paths for the entire chip [2], the probability of satisfying $t_{\rm max}$ is

$$P_{\rm WID}(t \le t_{\rm max}) = F_{\rm WID}(t_{\rm max}) = \left(F_{{\rm WID}-T_{\rm cp,nom}}(t_{\rm max})\right)^{N_{\rm cp}} \quad (4)$$

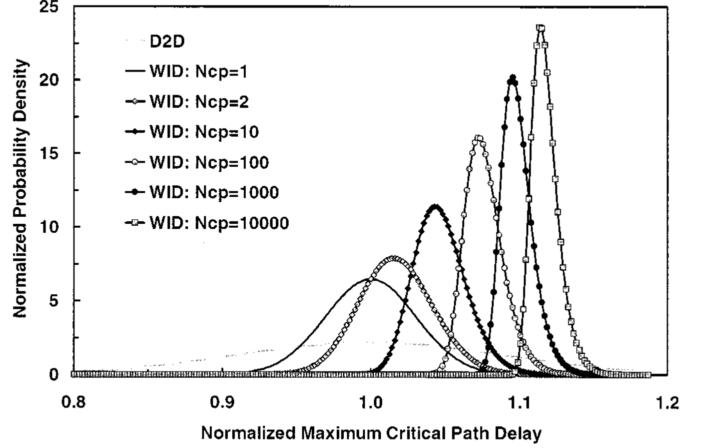

Fig. 3. Within-die (WID) maximum critical path delay distribution for different values of $N_{\rm cp}$ and the die-to-die (D2D) critical path delay distribution.

where $F_{\rm WID}$ is the chip's WID cumulative distribution. The chip's WID maximum critical path delay density function is then calculated by taking the derivative of (4) with respect to $t_{\rm max}$ as

$$\begin{aligned} f_{\rm WID}(t_{\rm max}) &= \frac{dF_{\rm WID}(t_{\rm max})}{dt_{\rm max}} \\ &= N_{\rm cp}\frac{dF_{{\rm WID}-T_{\rm cp,nom}}(t_{\rm max})}{dt_{\rm max}} \times \left(F_{{\rm WID}-T_{\rm cp,nom}}(t_{\rm max})\right)^{N_{\rm cp}-1} \\ &= N_{\rm cp} f_{{\rm WID}-T_{\rm cp,nom}}(t_{\rm max}) \times \left(F_{{\rm WID}-T_{\rm cp,nom}}(t_{\rm max})\right)^{N_{\rm cp}-1}. \end{aligned} \quad (5)$$

Fig. 3 illustrates the dependency of the WID maximum critical path delay density function (5) on $N_{\rm cp}$. As $N_{\rm cp}$ increases, the mean delay increases and the standard deviation decreases. Since the *slowest* critical path limits the chip's overall performance, the probability of a longer cycle time increases as $N_{\rm cp}$ increases. For example, when only one path is considered, the probability of a delay less than $T_{\rm cp,nom}$ is equal to 0.5. When two independent critical paths are considered, the probability that the delay is less than $T_{\rm cp,nom}$ is $(0.5)^2 = 0.25$. Notice, however, that increasing $N_{\rm cp}$ from 1 to 10 has a greater impact on the mean and variance of the WID distribution than increasing $N_{\rm cp}$ from 10^3 to 10^4, thus elucidating the decreasing dependency of the WID distribution on $N_{\rm cp}$ as $N_{\rm cp}$ increases to relatively large values. As the number of transistors per chip increases and the number of average gate delays per critical path is reduced [10], $N_{\rm cp}$ is expected to increase for each technology generation, therefore diminishing the relative sensitivity of the FMAX predictions to $N_{\rm cp}$.

For further insight, Fig. 4 plots the WID maximum critical path delay density function (5) on a logarithmic scale for $N_{\rm cp} = 1, 10, 10^2$, and 10^3 to illustrate the nonnormal shape of the WID distribution. Fig. 4 also plots the WID cumulative distribution for one critical path (3) on the right axis. The dependency of the WID density function (5) on $N_{\rm cp}$ has two competing factors, $N_{\rm cp}$ and $\left(F_{{\rm WID}-T_{\rm cp,nom}}\right)^{N_{\rm cp}-1}$. As $N_{\rm cp}$ increases, $f_{\rm WID}$ decreases dramatically for values of $F_{{\rm WID}-T_{\rm cp,nom}} \ll 1$ and increases linearly for values of $F_{{\rm WID}-T_{\rm cp,nom}}$ approaching 1. With

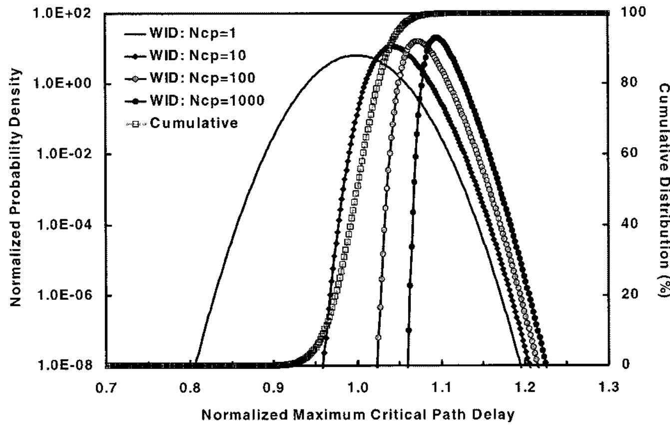

Fig. 4. Within-die (WID) maximum critical path delay distribution for different values of $N_{\rm cp}$ plotted on a logarithmic scale and the cumulative distribution for $N_{\rm cp} = 1$.

an increase in $N_{\rm cp}$, the resulting $f_{\rm WID}$ exhibits both a larger mean and a smaller variance. Thus, as the distribution shifts into the region where $F_{{\rm WID}-T_{\rm cp,nom}}$ approaches 1, $f_{\rm WID}$ becomes less sensitive to further increases in $N_{\rm cp}$.

C. Combining the Die-to-Die and Within-Die Maximum Critical Path Delay Distributions

The impact of both D2D and WID fluctuations on the chip's maximum critical path delay distribution is analyzed by combining the individual D2D and WID distributions. Shifting the D2D and WID distributions, (1) and (5), respectively, by $-T_{\rm cp,nom}$, the resulting distributions are expressed as

$$f_{\Delta T_{\rm D2D}} = N\left(0, \sigma^2_{{\rm D2D}-T_{\rm cp,nom}}\right) \tag{6}$$

and

$$f_{\Delta T_{\rm WID}}(t) = N_{\rm cp} f_{{\rm WID}-T_{\rm cp,nom}}(t - T_{\rm cp,nom}) \left(F_{{\rm WID}-T_{\rm cp,nom}}(t - T_{\rm cp,nom})\right)^{N_{\rm cp}-1}. \tag{7}$$

The density functions $f_{\Delta T_{\rm D2D}}$ and $f_{\Delta T_{\rm WID}}$ represent the deviations in delay from $T_{\rm cp,nom}$. Assuming $f_{\Delta T_{\rm D2D}}$ and $f_{\Delta T_{\rm WID}}$ are independent, the maximum critical path delay is calculated as

$$T_{\rm cp,max} = T_{\rm cp,nom} + \Delta T_{\rm D2D} + \Delta T_{\rm WID} \tag{8}$$

where $\Delta T_{\rm D2D}$ and $\Delta T_{\rm WID}$ are the deviations in the nominal critical path delay resulting from D2D and WID fluctuations, respectively. The maximum critical path delay density function resulting from both D2D and WID fluctuations is then calculated from a convolution

$$f_{T_{\rm cp,max}} = f_{T_{\rm cp,nom}} * f_{\Delta T_{\rm D2D}} * f_{\Delta T_{\rm WID}} \tag{9}$$

where $f_{T_{\rm cp,nom}}$ is an impulse at $T_{\rm cp,nom}$

$$f_{T_{\rm cp,nom}}(t) = \delta(t - T_{\rm cp,nom}). \tag{10}$$

Fig. 3 also plots the D2D critical path delay distribution (1). Although the D2D distribution is independent of $N_{\rm cp}$ since D2D fluctuations have an equal effect on each critical path on a chip, the WID distribution approaches an impulse function with an increasing mean delay as $N_{\rm cp}$ increases. As the D2D and WID distributions are statistically combined through (9), the resulting distribution has a mean equal to that of the WID distribution and a variance resulting predominantly from the D2D distribution. Thus, WID fluctuations determine the mean of the maximum critical path delay distribution, and D2D fluctuations determine the variance.

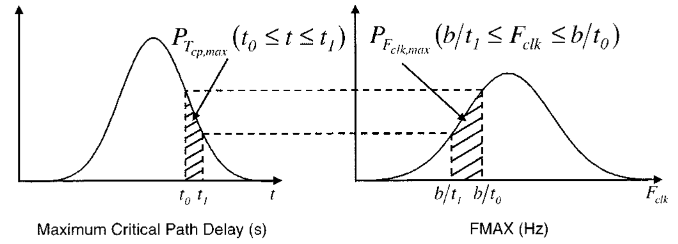

Fig. 5. Mapping the maximum critical path delay distribution to the maximum clock frequency distribution.

D. Mapping the Maximum Critical Path Delay Distribution to the Maximum Clock Frequency Distribution

The combined delay distribution in (9) is now mapped to a frequency distribution. The maximum clock frequency is calculated as

$$F_{\rm clk,max} = \frac{b}{T_{\rm cp,max}} \tag{11}$$

where b is the clock skew factor ($b = 0.9$, assumes 10% clock skew). Fig. 5 illustrates the mapping of $f_{T_{\rm cp,max}}$ to the maximum clock frequency density function $f_{F_{\rm clk,max}}$. The probability that the maximum critical path delay is within some interval $t_0 \le t \le t_1$ is equal to the probability that the maximum clock frequency is within the interval $b/t_1 \le F_{\rm clk} \le b/t_0$

$$\begin{aligned} P_{T_{\rm cp,max}}(t_0 \le t \le t_1) &= \int_{t_0}^{t_1} f_{T_{\rm cp,max}}(t)dt \\ &= P_{F_{\rm clk,max}}\left(\frac{b}{t_1} \le F_{\rm clk} \le \frac{b}{t_0}\right) \\ &= \int_{b/t_1}^{b/t_0} f_{F_{\rm clk,max}}(F_{\rm clk})\, dF_{\rm clk}. \end{aligned} \tag{12}$$

Define Δt as the difference between t_1 and t_0, as

$$\Delta t = t_1 - t_0 \tag{13}$$

and $\Delta F_{\rm clk}$ as the difference between b/t_0 and b/t_1, as

$$\Delta F_{\rm clk} = \frac{b}{t_0} - \frac{b}{t_1} = b\frac{t_1 - t_0}{t_0 t_1} = b\frac{\Delta t}{t_0 t_1}. \tag{14}$$

As Δt approaches zero, the integrals in (12) may be approximated as

$$\int_{t_0}^{t_1} f_{T_{\rm cp,max}}(t)dt \cong f_{T_{\rm cp,max}}(t_0)\Delta t \tag{15}$$

and

$$\int_{b/t_1}^{b/t_0} f_{F_{\rm clk,max}}(F_{\rm clk})\, dF_{\rm clk} \cong f_{F_{\rm clk,max}}\left(\frac{b}{t_0}\right)\Delta F_{\rm clk} \tag{16}$$

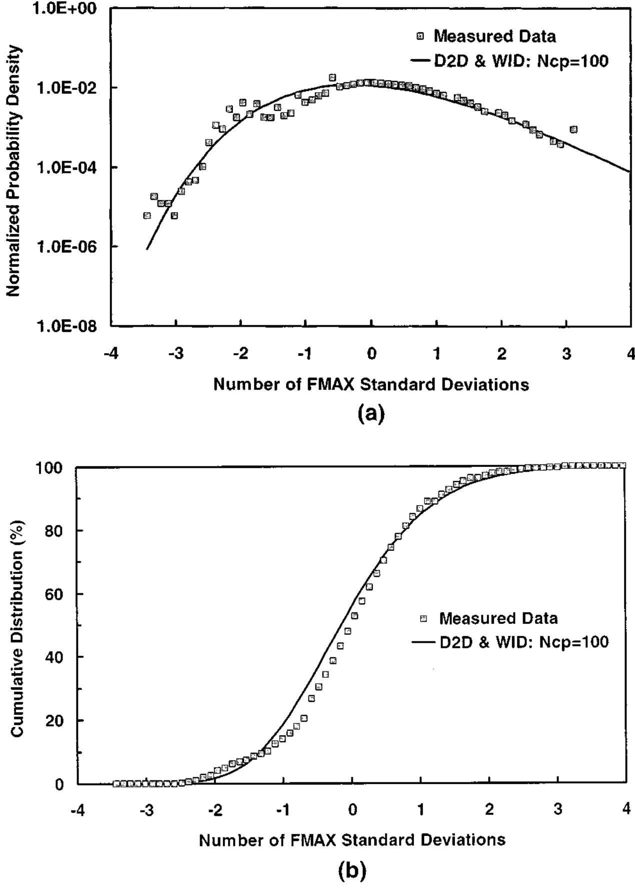

Fig. 6. Comparison of the model projections with measured data for the maximum clock frequency (a) probability density and (b) cumulative distribution.

where (14) is simplified as

$$\Delta F_{\text{clk}} = b\frac{\Delta t}{t_0 t_1} \cong b\frac{\Delta t}{t_0^2}. \tag{17}$$

Substituting (15)–(17) into (12) and replacing t_0 with t, the maximum clock frequency density function is derived as

$$f_{F_{\text{clk,max}}}\left(\frac{b}{t}\right) = f_{T_{\text{cp,max}}}(t)\frac{t^2}{b}. \tag{18}$$

E. FMAX Model Verification

Fig. 6(a) compares the FMAX distribution model, described in (2), (3), (6), (7), (9), (10), and (18), for both D2D and WID parameter fluctuations with $N_{\text{cp}} = 100$ against the FMAX measured data obtained at wafer sort for a recent 0.25-μm microprocessor. The wafer sort data represents measurements taken for approximately 50 000 dies. The predicted FMAX distribution agrees closely with the distribution of measured data in *mean*, *variance*, and *shape*. Fig. 6(b) validates the model with measured data for the cumulative FMAX distribution. Fig. 7 plots the distributions resulting from only D2D and only WID parameter fluctuations to illustrate their individual effects on the FMAX distribution. These results clearly reveal that *within-die fluctuations directly impact the FMAX mean and die-to-die fluctuations impact the FMAX variance.*

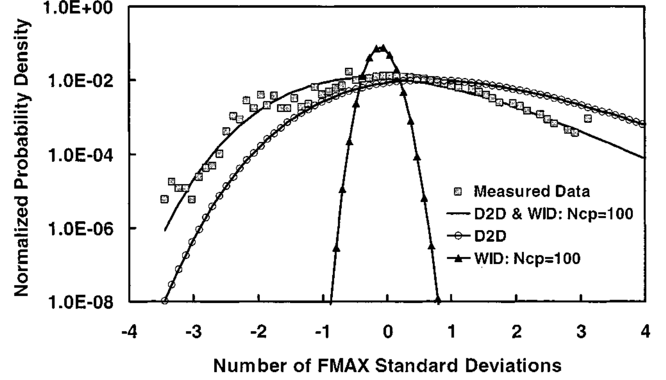

Fig. 7. Individual contributions of die-to-die (D2D) and within-die (WID) parameter fluctuations to the FMAX distribution.

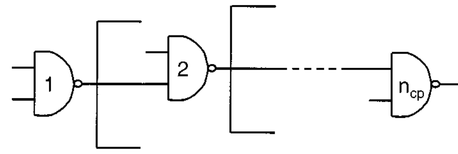

Fig. 8. Generic critical path (GCP) model, where n_{cp} is the number of average gate delays.

III. Generic Critical Path (GCP) Model

As discussed in Section II-A, the critical path delay distributions resulting from die-to-die and within-die fluctuations are calculated from D2D and WID statistical simulators, which both use a SPICE-equivalent circuit simulator [11] with a 0.25-μm process file and a netlist of speed-limiting paths from a specific microprocessor. Many of the device parameters provided in the process file are empirically calculated to fit measured I–V data. In projecting the impact of parameter fluctuations on future circuit performance, it is unclear how these empirical parameters might scale with technology. Therefore, a generic speed-limiting path model is developed through physically based device and circuit analyses [5]–[7] to evaluate the critical path delay distributions resulting from D2D and WID fluctuations.

As illustrated in Fig. 8, the generic critical path (GCP) is modeled by a number n_{cp} of identical two-input static CMOS NAND gates with a fan-out of three, where each gate drives an average wiring capacitance. The static CMOS logic gate is chosen for its low standby power drain, large operating margins, scalability, and flexibility of logic functions [12]. The average propagation delay through a two-input NAND gate is modeled by averaging the delay through two series-connected nFETs and the delay through one pFET, given as

$$T_{\text{nand}} = \frac{f_{\text{ineff}} T_{pd,n} + T_{pd,p}}{2} \tag{19}$$

where f_{ineff} is the effective fan-in factor [13], [14] for series-connected MOSFETs, and $T_{pd,n}$ and $T_{pd,p}$ are the nFET and pFET CMOS propagation delays [7], respectively, as derived from the physical alpha-power law model [5]. The critical path delay is then calculated as

$$T_{\text{cp}} = n_{\text{cp}} T_{\text{nand}}. \tag{20}$$

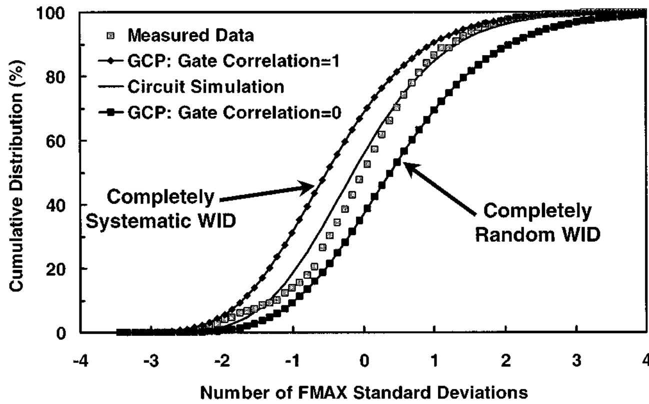

Fig. 9. Comparison of the FMAX projections using both the GCP model and the circuit simulation with measured data.

In calculating the critical path delay distributions that are used by the FMAX distribution model for the analysis in Section II, a rigorously developed WID fluctuation model is employed. This model is empirically derived through an analysis of manufacturing data specific to the 0.25-μm technology generation. The WID fluctuation model represents systematic within-die parameter variations by expressing the device-to-device correlation as a function of the distance between the devices. This correlation function, however, is significantly influenced by specific manufacturing capabilities. Currently, there is little insight into understanding how this distance correlation might scale for future technology nodes. Therefore, the GCP model analyzes two separate WID fluctuation cases: 1) completely dependent gates (gate correlation $= 1$) and 2) completely independent gates (gate correlation $= 0$), which may be viewed as extreme conditions of systematic and random fluctuations, respectively.

Using only the D2D and WID device parameter standard deviations for the 0.25-μm technology, critical path delay distributions are calculated through the GCP model for a gate correlation of one and zero. The results of these critical path delay distributions are inputs into the FMAX distribution model described in (2), (3), (6), (7), (9), (10), and (18). Fig. 9 compares the cumulative distributions of FMAX projected by the GCP model with a gate correlation of 1 and 0 to measured and simulated distributions. The circuit simulation clearly provides much better agreement with measured data than either of the GCP projections due to the accuracy of the systematic WID correlation model. *The GCP model, however, enables a key insight into the projections by establishing boundaries of the actual FMAX distribution with the two extreme cases of completely systemic and completely random within-die fluctuations. Moreover, Fig. 9 illustrates that systematic within-die fluctuations (gate correlation = 1) decrease the FMAX mean more severely than random within-die fluctuations (gate correlation = 0).*

This result can be explained physically as follows. In the completely systematic case, the variations have the same impact on every element in a critical path so that

$$\frac{\sigma_{T_{\rm cp}}}{T_{\rm cp}} = \frac{n_{\rm cp}\sigma_{T_{\rm nand}}}{n_{\rm cp}T_{\rm nand}} = \frac{\sigma_{T_{\rm nand}}}{T_{\rm nand}} \tag{21}$$

where $\sigma_{T_{\rm cp}}$ and $\sigma_{T_{\rm nand}}$ are the standard deviations of the critical path delay distribution and the NAND gate delay distribution, respectively. For completely systematic WID fluctuations, the ratio of the standard deviation to mean is equal for the critical path delay distribution and the gate delay distribution. In the case of completely random fluctuations, however, the fluctuations in the critical path delay are expected to have an averaging effect over the number of gates in the path [3] such that

$$\frac{\sigma_{T_{\rm cp}}}{T_{\rm cp}} = \frac{\sqrt{n_{\rm cp}}\sigma_{T_{\rm nand}}}{n_{\rm cp}T_{\rm nand}} = \frac{1}{\sqrt{n_{\rm cp}}}\frac{\sigma_{T_{\rm nand}}}{T_{\rm nand}}. \tag{22}$$

For completely random WID fluctuations, the ratio of standard deviation to mean for the critical path delay distribution is inversely proportional to the square root of $n_{\rm cp}$ [3]. Thus, for $n_{\rm cp}$ greater than one, systematic WID fluctuations induce worse performance degradation than random WID fluctuations.

Technology Generation (nm)	180	130	100	70	50
L_{Gate} (nm)	140	85	65	45	32
t_{OX} (nm)	2.5	1.9	1.5	1.5	1.5
V_{DD} (V)	1.52	1.40	1.04	1.00	0.70
V_{TL} (V)	0.30	0.30	0.28	0.38	0.36
N_A ($\times 10^{18}$ cm^{-3})	0.85	1.37	1.88	3.00	2.74
n_{cp} (#gates/CP)	10	9	8	7	6
N_{cp} (#CP/chip)	10^2	10^3	10^3	10^4	10^4
I_{OFF} (nA/μm)	5	10	20	40	80
$F_{CLK,nom}$ (GHz)	1.25	2.10	3.50	6.00	10.00

Fig. 10. Nominal values used in the projection analysis.

IV. Impact of Parameter Fluctuations on Future FMAX Distributions

In projecting the impact of parameter fluctuations on future circuit performance, the nominal values in Fig. 10 are selected judiciously by using the International Technology Roadmap for Semiconductors (ITRS) [15] as a guideline. The nominal values of gate channel length $L_{\rm Gate}$, maximum source-to-drain leakage current $I_{\rm OFF}$, and on-chip local clock frequency $F_{\rm CLK,nom}$ are all provided by the ITRS [15]. The value of gate oxide thickness $t_{\rm OX}$ is chosen from the ITRS [15] for the 180, 130, and 100-nm technology generations, however, the value of $t_{\rm OX}$ is not reduced below the 100-nm technology generation as forecast by the ITRS. The continued scaling of $t_{\rm OX}$ as projected by the ITRS assumes the development of a high-κ gate dielectric material to replace the native oxide as the gate insulator. The ITRS, however, emphasizes that there are currently "no known solutions" [15] for this prediction. Recent studies have estimated the minimum value of $t_{\rm OX}$ necessary for retaining the bulk properties of SiO_2 to be approximately 1.5 nm [16], [17], the value to which $t_{\rm OX}$ is limited in this analysis. The long channel threshold voltage $V_{\rm TL}$ and average doping concentration $N_{\rm A}$ are calculated using the physical alpha-power law's subthreshold drain current model [5] and the ITRS projections for $I_{\rm OFF}$ [15]. The supply voltage $V_{\rm DD}$ and $n_{\rm cp}$ are calculated by equating $F_{\rm CLK,nom}$ to the product of b (clock skew factor) and the reciprocal of $T_{\rm cp}$ (20) while maintaining relative agreement with the nominal saturation drain current $I_{\rm ON}$ and the range of $V_{\rm DD}$ values provided in the ITRS [15]. $N_{\rm cp}$ is estimated by assuming the ratio of independent critical paths to the number of transistors per chip remains relatively constant. As

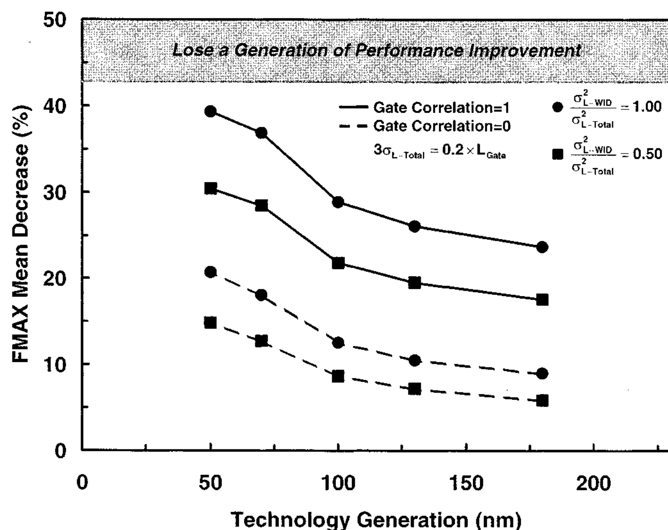

Fig. 11. Reduction in FMAX mean resulting from within-die parameter fluctuations versus technology generation.

discussed in Section II-B, the sensitivity of the FMAX distribution to N_{cp} is essentially negligible for sufficiently large values ($\sim 10^2$–10^4) of N_{cp}.

As discussed earlier, the WID parameter fluctuations directly impact the FMAX mean. Using the GCP and the FMAX distribution models, Fig. 11 projects the impact of within-die parameter fluctuations on the FMAX mean for the 180, 130, 100, 70, and 50-nm technology generations. The GCP model assumes the 3σ effective channel length L deviation is 20% of the nominal gate length [15]. Since L is among the most difficult device parameters to control in the manufacturing process as well as one of the most influential on circuit performance, only L and the corresponding parameters that are dependent on L (e.g., effective threshold voltage, drain current, etc.) are varied in this projection analysis. All other device and circuit parameters such as V_{DD}, t_{OX}, N_A, interconnect capacitance, etc., are assumed to remain unchanged, thus resulting in a more optimistic projection. Fig. 11 provides a range of percentages (50% and 100%) for the ratio of the WID channel length variance to the total channel length variance (WID and D2D). These ranges are plotted for the GCP model using a gate correlation of one (completely systematic WID fluctuations) and a gate correlation of zero (completely random WID fluctuations).

Fig. 11 includes a shaded region to indicate the limit at which within-die parameter fluctuations degrade the FMAX mean such that the performance gained from a generation of transistor scaling is completely lost. Typical technology scaling decreases the gate delay T_{Gate} by 30% [18]. The improvement in clock frequency resulting strictly from technology scaling is

$$\%F_{CLK} \text{ increase } = \frac{\frac{1}{0.7T_{Gate}} - \frac{1}{T_{Gate}}}{\frac{1}{T_{Gate}}} \approx 43\%. \quad (23)$$

The clock frequency improvement from one technology generation to another is also aided by architecture advances, such as reducing the number of gate delays in a critical path. The limit at 43% provides a criterion for evaluating the impact of the WID parameter fluctuations on circuit performance.

In analyzing Fig. 11 for the 50-nm technology generation, the GCP model using the gate correlation of one projects a degradation in the FMAX mean of 30% and 39% corresponding to ratios of WID channel length variance to total channel length variance of 50% and 100%, respectively. For the same technology node and ratios of channel length variance, the GCP model using a gate correlation of zero predicts a decrease in performance of 15% and 21%. Fig. 11 indicates that the performance degradation resulting from systematic WID fluctuations is much worse than the performance loss resulting from the random WID fluctuations. Since WID fluctuations directly impact the FMAX mean, *the systematic within-die fluctuations are the most significant performance limiter resulting from parameter fluctuations*. This result is of concern since characterizations of a 0.18-μm manufacturing process indicate a more systematic than random WID fluctuation [19]. Fig. 11 projects that essentially *a generation of performance gain can be lost due to systematic within-die fluctuations* at the 50-nm technology node. The key recommendation from this analysis is for manufacturing process controls to focus primarily on the sources of systematic within-die fluctuations such as stepper lens aberrations. Moreover, new circuit design methodologies that suppress the impact of within-die parameter fluctuations should be investigated.

V. Conclusion

A model for the maximum clock frequency (FMAX) distribution is presented and compared with wafer sort data for a recent 0.25-μm microprocessor. Model predictions agree closely with measured data in mean, variance, and shape, and reveal that within-die fluctuations primarily impact the FMAX mean, and die-to-die fluctuations the FMAX variance. The impact of parameter fluctuations on future circuit performance is then analyzed by using a physically based generic critical path model to determine the critical path delay distributions resulting from die-to-die and within-die fluctuations. Utilizing the results of these distributions with the FMAX distribution model, projections are made for the 180, 130, 100, 70, and 50-nm technology generations. Results indicate that systematic within-die fluctuations are the most significant performance limiter resulting from parameter fluctuations. Assuming a 3σ channel length deviation of 20%, projections for the 50-nm technology generation indicate that approximately a generation of performance improvement can be lost due to systematic within-die fluctuations. As device fluctuations increase with decreasing dimensions and the probability of longer critical path delay deviations increases with growing transistor density, within-die parameter fluctuations may become the key obstacle to improving the performance of GSI. To overcome this barrier, results of this work suggest a two-pronged plan of attack: 1) develop manufacturing process controls to reduce the sources of systematic within-die fluctuations such as stepper lens aberrations, and 2) explore new circuit design methodologies aimed at suppressing the impact of within-die parameter fluctuations on future circuit performance.

Acknowledgment

The authors would like to express their sincere appreciation to G. Sery, A. Brand, N. Hakim, R. Gee, S. Hu, M. Wesela, E. Cohn

of Intel, and J. Joyner of Georgia Tech for their helpful advice and kind support.

REFERENCES

[1] S. G. Duvall, "Statistical circuit modeling and optimization," in *5th Intl. Workshop Statistical Metrology*, June 2000, pp. 56–63.
[2] K. A. Bowman, X. Tang, J. C. Eble, and J. D. Meindl, "Impact of extrinsic and intrinsic parameter fluctuations on CMOS circuit performance," *IEEE J. Solid-State Circuits*, vol. 35, pp. 1186–1193, Aug. 2000.
[3] M. Eisele, J. Berthold, D. Schmitt-Landsiedel, and R. Mahnkopf, "The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits," in *Proc. ISLPED*, Aug. 1996, pp. 237–242.
[4] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2001, pp. 278–279.
[5] K. A. Bowman, B. L. Austin, J. C. Eble, X. Tang, and J. D. Meindl, "A physical alpha-power law MOSFET model," *IEEE J. Solid-State Circuits*, vol. 34, pp. 1410–1414, Oct. 1999.
[6] B. Agrawal, V. K. De, and J. D. Meindl, "Opportunities for scaling FET's for gigascale integration (GSI)," in *Proc. 23rd ESSDERC*, Sept. 1993, pp. 919–926.
[7] K. A. Bowman, L. Wang, X. Tang, and J. D. Meindl, "A circuit-level perspective of the optimum gate oxide thickness," *IEEE Trans. Electron Devices*, vol. 48, pp. 1800–1810, Aug. 2001.
[8] K. A. Bowman and J. D. Meindl, "Impact of *within-die* parameter fluctuations on future maximum clock frequency distributions," in *IEEE CICC*, May 2001, pp. 229–232.
[9] D. J. Frank, P. Solomon, S. Reynolds, and J. Shin, "Supply and threshold voltage optimization for low power design," in *Proc. ISLPED*, Aug. 1997, pp. 317–322.
[10] P. E. Gronowski, W. J. Bowhill, R. P. Preston, M. K. Gowan, and R. L. Allmon, "High-performance microprocessor design," *IEEE J. Solid-State Circuits*, vol. 33, pp. 676–686, May 1998.
[11] *HSPICE User's Manual*, Meta-Software Inc., 1995.
[12] J. D. Meindl, "Low power microelectronics: Retrospect and prospect," *Proc. IEEE*, pp. 619–635, Apr. 1995.
[13] T. Sakurai and R. Newton, "Delay analysis of series-connected MOSFET circuits," *IEEE J. Solid-State Circuits*, vol. 26, pp. 122–131, Feb. 1991.
[14] A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, and J. D. Meindl, "A minimum total power methodology for projection limits on CMOS GSI," *IEEE Trans. VLSI Systems*, pp. 235–251, June 2000.
[15] *International Technology Roadmap for Semiconductors (ITRS)*, Semiconductor Industry Association, 1999.
[16] D. A. Muller *et al.*, "The electronic structure at the atomic scale of ultrathin gate oxides," *Nature*, pp. 758–761, June 1999.
[17] S. Thompson, P. Packan, and M. Bohr, "MOS scaling: Transistor challenges for the 21st century," *Intel Tech. J.*, 3rd qtr. 1998.
[18] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proc. ISLPED*, Aug. 1999, pp. 163–168.
[19] M. Orshansky *et al.*, "Intra-field gate CD variability and its impact on circuit performance," in *IEEE IEDM Tech. Dig.*, Dec. 1999, pp. 479–482.

Keith A. Bowman (S'97–M'02) received the B.S. degree in electrical engineering from North Carolina State University, Raleigh, in 1994 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1995 and 2001, respectively. His doctoral research focused on the impact of power consumption and parameter fluctuations on future circuit performance to enable opportunities for further advancement of gigascale integration.

He is currently a Senior Computer-Aided Design (CAD) Engineer in the Technology CAD Division, Intel Corporation, Hillsboro, OR. In summer of 2000, he performed research in modeling the impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for a 0.25-μm microprocessor while interning with the Technology CAD Division, Intel Corporation, Santa Clara, CA.

Steven G. Duvall (M'94) was born in Cincinnati, OH, in 1956, received the B.S. degree in engineering from Humboldt State University, Arcata, CA, in 1978 and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, in 1980 and 1983, respectively.

He is Director of Strategic Investment for Intel Capital in Australia and New Zealand, based in Sydney, Australia. He joined Intel in 1983 as a Senior Process Engineer and subsequently worked as a Computer-Aided Design Engineer. For ten years, he led the investigation and development of advanced computer-aided design tools for optimizing manufacturing technologies and integrated circuits. He has published extensively on computer-aided design, statistical design, and optimization

Dr. Duvall was appointed an Intel Fellow in 1999. He has twice been awarded Intel's highest individual recognition award.

James D. Meindl (M'56–SM'66–F'68–LF'97) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Carnegie-Mellon University, Pittsburgh, PA, in 1955, 1956, and 1958, respectively.

He is currently the Director of the Joseph M. Pettit Microelectronics Research Center and the Pettit Chair Professor of Microelectronics at the Georgia Institute of Technology, Atlanta, GA. He served from 1986 to 1993 as Senior Vice President for Academic Affairs and Provost of Rensselaer Polytechnic Institute, Troy, NY. From 1967 through 1986, he was with Stanford University, Stanford, CA, where he was the John M. Fluke Professor of Electrical Engineering, Associate Dean for Research in the School of Engineering, Director of the Center for Integrated Systems, Director of the Electronics Laboratories, and founding Director of the Integrated Circuits Laboratory. He is a co-founder of Telesensory Systems, Inc., the principal manufacturer of electronic reading aids for the blind, and served as a member of the Board from 1971 through 1984. From 1965 through 1967, he was founding Director of the Integrated Electronics Division at the Fort Monmouth, NJ, U.S. Army Electronics Laboratories. He is the author of the book *Micropower Circuits* and over 500 technical papers on ultralarge-scale integration, integrated electronics, and medical electronics, and editor of the book *Brief Lessons in High Technology*, which elucidates the most important economic event of our lives, the emergence of the information society.

Dr. Meindl is a Life Fellow of the American Association for the Advancement of Science, and a member of the American Academy of Arts and Sciences and the National Academy of Engineering and its Academic Advisory Board. He most recently was awarded the Georgia Institute of Technology Distinguished Professor Award. He received the IEEE Third Millennium Medal, the 1999 SIA University Research Award, the 1997 Hamerschlag Distinguished Alumnus Award from Carnegie Mellon University, and the 1991 Benjamin Garver Lamme Medal from ASEE. He was the recipient of the 1990 IEEE Education Medal "for establishment of a pioneering academic program for the fabrication and application of integrated circuits" and the recipient of the 1989 IEEE Solid-State Circuits Medal for contributions to solid-state circuits and solid-state circuit technology. At the 1988 IEEE International Solid-State Circuits Conference, he received the Beatrice K. Winner Award. In 1980, he was the recipient of the IEEE Electron Devices Society's J.J. Ebers Award for his contributions to the field of medical electronics and for his research and teaching in solid-state electronics. From 1970 through 1978, he and his students received five outstanding paper awards at IEEE International Solid-State Circuits Conferences, along with one received at the 1985 IEEE VLSI Multilevel Interconnections Conference. His major contributions have been new medical instruments enabled by custom integrated electronics, projections and codification of the hierarchy of physical limits on integrated electronics, and leadership in creation of academic environments promoting high-quality teaching and research.