Pathfinding for 22nm CMOS Designs using Predictive Technology Models

Xia Li, Wei Zhao, Yu Cao*, Zhi Zhu, Jooyoung Song, David Bang, Chi-Chao Wang*,
Seung H. Kang, Joseph Wang, Matt Nowak, Nick Yu

QCT Process Technology, Qualcomm Incorporated, 5775 Morehouse Dr., San Diego, CA 92121, USA
*Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287, USA

Abstract – **Traditional IC scaling is becoming increasingly difficult at the 22nm node and beyond. Dealing with these challenges increase product development cycle time. For continued CMOS scaling, it is essential to start design explorations in new process nodes as early as possible. Such an effort requires having Predictive Technology Models, which bridge technological and design practices, in order to assess the performance impact of future key modules. In this paper we propose a strategy that enables simultaneous investigation of advanced process and design concepts. Based on a customized predictive methodology and silicon data at 90-45nm nodes, compact transistor and interconnect models are developed for the next generation CMOS technology. We capture the heuristic device behavior during the scaling, which helps us to gain key insights that allow us to make tradeoffs of circuit performance metrics for next technology node.**

I. INTRODUCTION

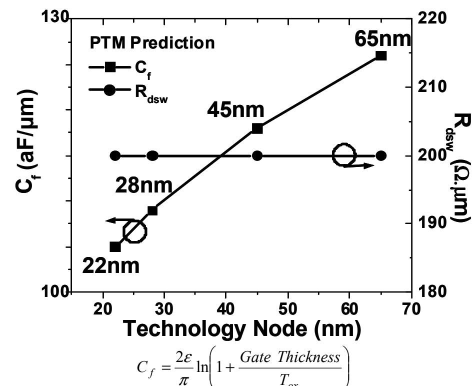
CMOS technology scaling is increasingly challenged by physics and manufacturing limits at 22nm and beyond [1]. These challenges reduce the predictability of circuit performance and increase the development cycle for new IC products. Continued CMOS scaling, it requires comprehending the technology impacts and capabilities as early as possible. This enables identification of potential issues, allows adaptive design decisions up front, and guarantees managing the time to market. Objecting such early stage exploration requires Predictive Technology Models (PTM) to assess performance trends, evaluate key modules before Silicon is ready, and facilitates the development of future CMOS technology [2-3].

This work proposes such a predictive strategy to enable simultaneous exploration of low power CMOS process and design concepts. We incorporate the general PTM methodology [2-4], with customized enhancements of transistor-level and interconnect-level physical effects [1]. These customized PTM models are systematically calibrated to 90-45nm Poly/SiON silicon data and published high-k/metal gate (HK/MG) information. Section II briefly presents the predictive methodology. We leverage these predictive models to quantify projections of various behaviors of CMOS devices, interconnect, and representative design modules, such as ring oscillator (RO), standard cell and SRAM for the 22nm node. The results reveal the trends, potential, and limits of technology scaling, and provide important guidance and insight into process and design choices (Section III).

II. METHODOLOGY OF PREDICTIVE MODELING

The PTM was first introduced in 2000 [1] and was further improved in 2006 [2]. It covers both frontend-of-the-line (FEOL) devices and backend-of-the-line (BEOL) metal interconnects. Predictions of FEOL technology rely on a set of fundamental equations that capture the essential behavior of charge and carrier transport, rather than the full set of BSIM models [2]. The electrostatic models emphasize the dependence of V_{th} on channel length (e.g., DIBL), channel doping, HALO, etc. The transport part of the model adopts the velocity saturation model with overshoot behavior [2]. In addition, the impact of layout dependent stress effects are embedded into mobility and V_{th} predictive models.

By using this set of models, only about ten primary parameters are needed to enable the prediction [2]. This simplification allows easier calibration of critical model parameters with published data [2, 4], and the modeling of scaling trends [2]. The general PTM models are available at http://www.eas.asu.edu/~ptm/. In this work, PTM is further customized for low power applications with multiple V_{th} choices [5]. Benefiting from the process continuity scaling, we handle secondary device effects with better confidence, including the body effect, temperature dependence and parasitic capacitances. For example, the trend of gate fringe capacitance (C_f) is calculated based on a physical equation, with source/drain resistance (R_{dsw}) remaining constant (Fig. 1). Figures 2 and 3 predict I-V characteristics [6][7].



$$C_f = \frac{2\varepsilon}{\pi}\ln\left(1+\frac{Gate\ Thickness}{T_{ox}}\right)$$

Fig. 1 The scaling trends of C_f. R_{dsw} keeps constant.

In addition to the transistor, parasitic BEOL resistance and capacitance play an increasingly important role in determining circuit performance. A predictive model for simple BEOL structures was presented in [4]. As shown in Table 1 [6, 8-9], more complicated BEOL structures and physical effects, such as high-k cap layer, etch damage layer, and metal grain scattering effects are existed in advanced process nodes [1]. We have developed a new capacitance model that incorporates these advanced features [10]. The new model decomposes the electrical field into various regions and solves each basic

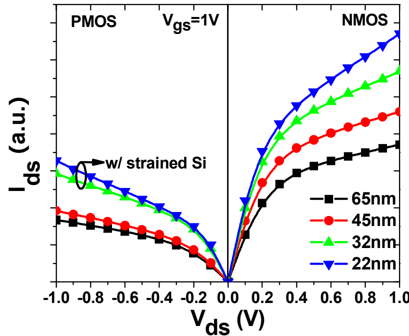


Fig. 2 The nominal I_{ds}-V_{ds} curves of scaled CMOS devices.

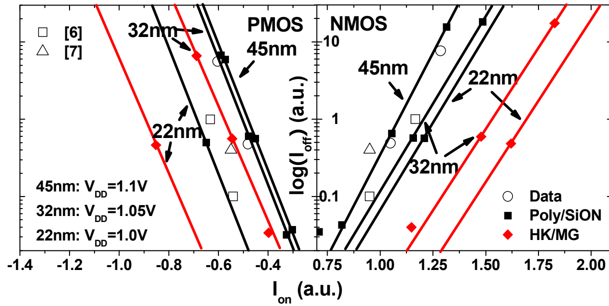


Fig. 3 I_{on}/I_{off} are predicted for Poly/SiON CMOS and HK/MG devices from 45nm to 22nm nodes. Part of the data is from silicon.

TABLE 1. INTERCONNECT SCALING TREND (*: PREDICTION).

Technology (nm)	65 [6]	45 [6]	32 [8]	22
Gate Pitch w/ contact (nm)	260	162	130	90 [9]
Contact Pitch (nm)	200	126	110	80*
M1 Pitch (nm)	180	126	100	70*
Intermediate Metal Pitch (nm)	200	126	100	70*
IMD k value	2.9	2.5	2.4	2.2*

TABLE 2. PTM MODELS OF WIRE AND CONTACT/VIA RESISTANCE.

Metal Resistivity	$\rho_m = \rho_{bulk}(1 + d/w)$
Metal Resistance	$R_m = \left(\frac{\rho_m \cdot L}{A_{metal}}\right) \parallel \left(\frac{\rho_l \cdot L}{A_{liner}}\right)$
Contact/Via Resistance	$R_c = \frac{\rho_m \cdot H/\pi}{r^2 + (2r+T_t)T_t(\rho_m/\rho_l)} + \frac{\rho_l \cdot T_t/\pi}{(r+T_t)^2} \propto (\rho_m \cdot H + \rho_l \cdot T_t)/W^2$

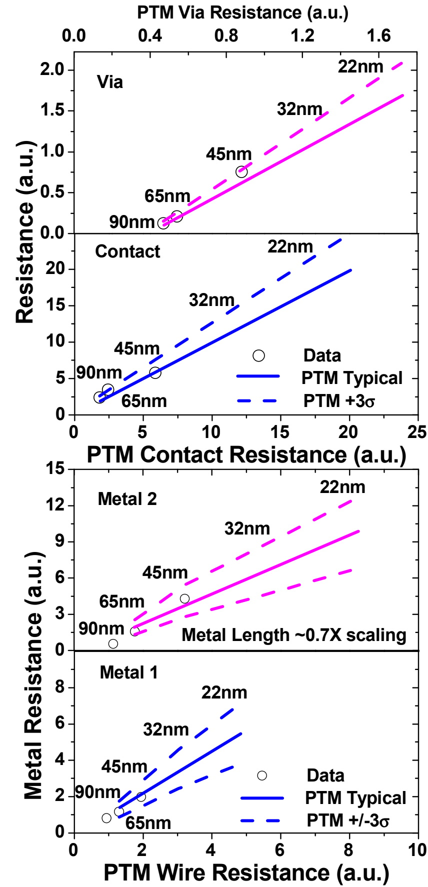


Fig. 4 Scaling trends of contact/via and metal wire resistance, for typical and +3σ values. Open circles are silicon data. Dashlines are 3σ trends.

component into a closed-form solution. Such a physical approach is convenient to incorporate new structures and materials, minimizing the complexity and the error in the model fitting process. Metal wire and contact/via resistance models are also developed, as shown in Table 2.

In Table 2, the effect of electron scattering is also considered, where d is electron scattering coefficient; ρ_m and ρ_{bulk} are metal resistivity and metal bulk resistivity, respectively; A_{metal} and A_{liner} are the area of metal and the barrier metal liner, respectively; H is the height of contact/via; T_t and ρ_l are barrier metal liner thickness and resistivity; r is the radius of contact/via; and W is the structure width. Based on Tables 1 and 2, the PTM BEOL model projects the scaling trend of contact/via and metal resistances, and their variations. Figure 4 shows comparison between model and silicon data.

III. PATHFINDING 22NM DESIGN ANALYSIS

Through SPICE simulations, PTM offers an insightful pathway to evaluate the trends and tradeoffs of circuit performance metrics, for given low power design constraints. As an example, the SRAM cell static noise margin (SNM) trend in Fig. 5 illustrates that a 22nm HK/MG device may still have adequate SNM [1, 6, 8-9, 11-12]. PTM HK/MG SNM is below the average of published data at 22nm node. This may

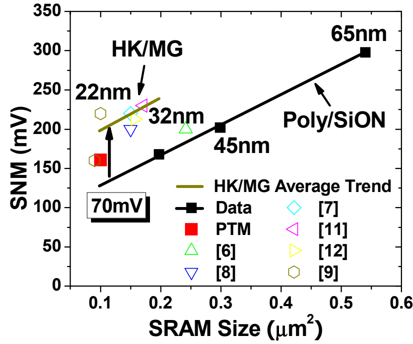


Fig. 5 Trends of SRAM static noise margin. Red square is PTM HK/MG SNM. HK/MG average trend line guides eye view.

imply that the PTM model and the layout are not optimized for a scaled SRAM cell.

Figure 6 shows a smooth reduction in the delay of an inverter-based FO=4 ring oscillator (RO). During the V_{DD} scaling, the device delay and dynamic energy trends are similar across generations (Figs. 6 and 7). However, it is observed that the RO delay at 22nm increases rapidly at lower voltage due to the strain effect.

Figure 8 shows total power trend of RO at 10% duty cycle. If we keep the same I_{off} target, HK/MG may not save total energy (i.e., standby and active energy) at the same power voltage at 22nm. However, a design with HK/MG devices allows further V_{DD} reduction at the same I_{on} target, since HK/MG effectively boosts drive current compared with Poly/SiON. Thus, it helps to reduce I_{off} and total energy with lower V_{DD}. If the duty cycle is below 10%, then total energy

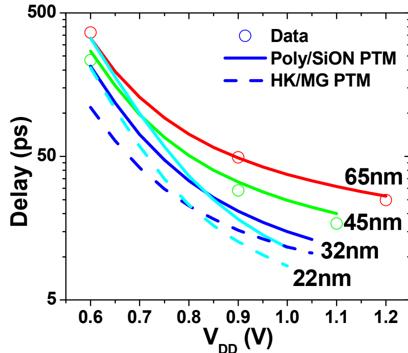


Fig. 6 Device delay trends of RO (inverter based, FO=4) during V_{DD} scaling.

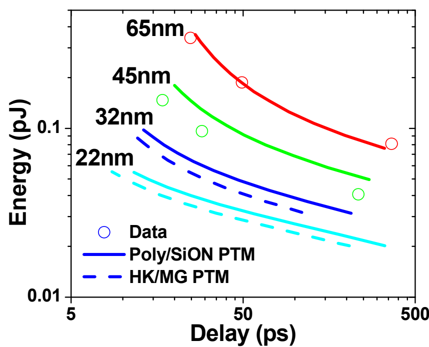


Fig. 7 Device dynamic energy vs. delay trends of RO FO=4 during V_{DD} scaling.

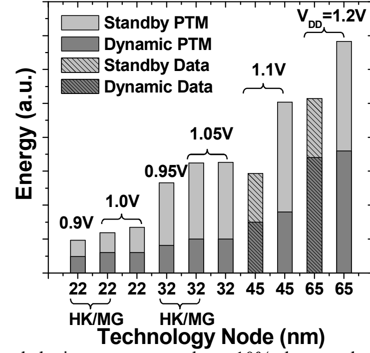


Fig. 8 Total device energy trends at 10% duty cycle. PTM has a relative trend of silicon data (dash line filling). Results that are not labeled with HK/MG are for Poly/SiON.

reduction is marginal at lower frequency because of leakage power dominated.Therefore, multiple V_{DD} domains are more effective than frequency tuning for low power design, especially with increased device count per product.

As transistor delay is reduced, parasitic RC delay becomes a larger portion in total path delay. In addition, V_{DD} drop due to local wiring increases with wire resistance with miniaturization of feature size. Because of these reasons, BEOL is increasingly important at 22nm and beyond. For example, Fig. 9 shows the IR drop at various technology nodes. Assuming M1 length is 50% of cumulative standard cell width of each generation, local V_{DD} drop from M6 to active region dramatically increases as metal resistance and, more importantly, contact resistance become larger during technology scaling. As contact resistance is more dominant,

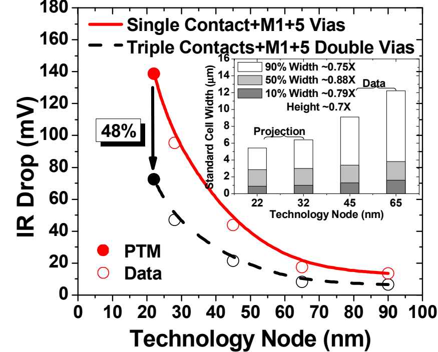


Fig. 9 V_{DD} drop at 500μA (M6 to M1 by stack via). The insert is the trends of 10/50/90% of cumulative standard cell width of technology nodes.

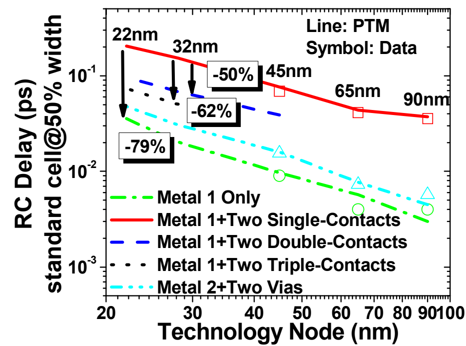


Fig. 10 Interconnect RC delay at 50% of cumulative standard cell width. Lines and dash lines are PTM. Open marks are silicon data.

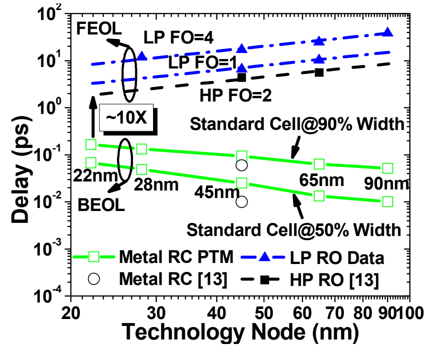


Fig. 11 FEOL only delay trend. Gap of FEOL and BEOL delay becomes smaller (HP ~10X, LP ~20X) while technology is scaled to 22nm.

use of triple-contacts reduces the IR drop by 48% in a 22nm design (Fig. 9). With two contacts, the RC delay of M1 wire increases by 79% at 22nm. Adoption of double- and triple-contact reduces M1 RC by 50% and 62%, respectively (Fig. 10). As shown in Fig. 11, the delay gap between FEOL and BEOL reduces by 10X and 20X at 22nm for high performance (HP) and low power (LP) applications, respectively [13].

The above studies compare the relative importances of FEOL and BEOL on circuit performance. We conclude that BEOL significantly increases the impact on both path delay and local IR drop, as shown in Figs. 9-11.

Finally, we have also studied the impact of technology scaling on circuit performance. By decomposing into various components, we can identify key factors that limit the performance and adaptively find technological or design solutions. Figures 12 and 13 show the decomposition of RO delay and dynamic energy, respectively. Assuming that the M1 length in a stage of RO is 90% of cumulative standard cell width, we observe a larger impact of gate fringe capacitance C_f and BEOL parasitics on delay and energy in the 22nm node. By identifying these critical components, we are able to optimize process and circuit design at process development.

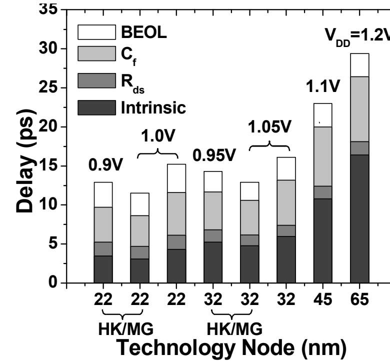


Fig. 12 The decomposition of total RO delay (FO=4), with two contacts/stage and M1 length at 90% of cumulative standard cell width. Results that are not labeled with HK/MG are for Poly/SiON.

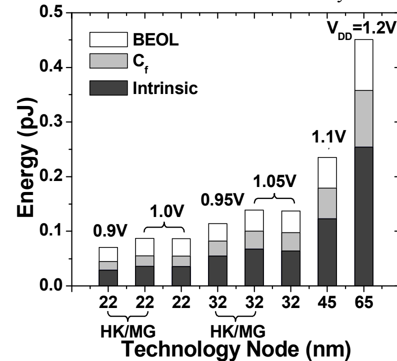


Fig. 13 The decomposition of total dynamic energy RO (FO=4) with the same interconnect loading of Fig. 12. Results that are not labeled with HK/MG are for Poly/SiON.

IV. Conclusion

We have demonstrated a customized PTM methodology for 22nm pathfinding applicable to low power designs. We have evaluated critical performance metrics, e.g., speed and power, with various technological components and design choices with scaled CMOS. Our customized PTM enables exploration of trends and limits at future nodes. It is also applicable to variability studies using statistical PTM FEOL and BEOL models. The proposed methodology allows designers to start competitive design research using advanced CMOS technologies before silicon data is available.

References

[1] "International Technology Roadmap for Semiconductors," 2008.

[2] W. Zhao and Y. Cao, "New generation of predictive technology modeling for sub-45nm early design exploration," *IEEE Trans. Electron Devices*, vol. 53, pp. 2816-2823, Nov. 2006.

[3] W. Zhao, X. Li, M. Nowak, Y. Cao, "Predictive technology modeling for 32nm low power design," *International Semiconductor Device Research Symposium*, TA4-03, 2007.

[4] Y. Cao, T. Sato, M. Orshansky, D. Sylvester, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit simulation," *Custom Integrated Circuits Conf.*, pp. 201-204, 2000.

[5] W. Zhao, F. Liu, K. Agarwal, D. Acharyya, S. Nassif, K. J. Nowka and Y. Cao, "Rigorous extraction of process variations for 65-nm CMOS Design," *IEEE Trans. Semiconductor Manufacturing*, vol. 22, pp. 196-203, Feb. 2009.

[6] K.-L. Cheng, *et al*., "A Highly scaled, high performance 45nm bulk logic CMOS technology with 0.242μm² SRAM Cell," *IEEE International Electron Devices Meeting*, pp. 243-246, 2007.

[7] S.-Y. Wu, *et al*., "A 32nm CMOS low power SoC platform technology for foundry applications with functional high density SRAM," *IEEE International Electron Devices Meeting*, pp. 263-266, 2007.

[8] C. H. Diaz, *et al*., "32nm gate-first high-k/metal-gate technology for high performance low power applications," *IEEE International Electron Devices Meeting, pp. 629-632, 2008.*

[9] B. S. Haran, *et al*., "22 nm technology compatible fully functional 0.1μm² 6T-SRAM cell," *IEEE International Electron Devices Meeting*, pp. *625-628, 2008.*

[10] W. Zhao, X. Li, S. Gu, S. H. Kang, M. Nowak and Y. Cao, "Field-based capacitance modeling for sub-65nm on-chip interconnect," submitted to *IEEE Trans. Electron Devices*, 2009.

[11] S. Natarajan, *et al*., "A 32nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and 0.171μm² SRAM cell size in a 291Mb array," *IEEE International Electron Devices Meeting*, pp. 941-943, 2008.

[12] F. Arnaud, *et al*., "32nm general purpose bulk CMOS technology for high performance applications at low voltage," *IEEE International Electron Devices Meeting*, pp. 633-636, 2008.

[13] K. Mistry, *et al.*," A 45nm logic technology with high-k + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging," *IEEE International Electron Devices Meeting*, pp. 247-250, 2007.