32-bit Processor Core at 5-nm Technology: Analysis of Transistor and Interconnect Impact on VLSI System Performance

Chi-Shuen Lee[1], Brian Cline[2], Saurabh Sinha[2], Greg Yeric[2], H.-S. Philip Wong[1]

[1]Deprtment of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. Email: chishuen@stanford.edu

[2]ARM Inc., Austin, TX 78735, USA.

Abstract—A 32-bit commercial processor core is implemented at 5-nm design rules to study transistor and interconnect technology options and the impact of increasing interconnect resistance on system performance. Insights obtained are: 1) The major benefit of downscaling FET gate length is reducing MEOL parasitics instead of the intrinsic gate capacitance. 2) 2D-material-based FETs can achieve ~2× better core-level energy-delay-product in theory compared to the projected Si FinFET; contact resistivity $<6\times10^{-8}$ Ω-μm^2 is required for 2D-FETs to match the core performance using Si FinFET. 3) Signal routing optimization can mitigate the impact of BEOL resistance such it contributes to 15%-35% of the total delay at the cost of using more cells and vias, which is not manifest if a ring oscillator with fixed wire load is used without performing full place-and-route. 4) Thinning Cu diffusion barrier can improve system performance up to 11% and alleviate BEOL variation impact.

Introduction

As CMOS technology scaling continues toward sub-10-nm nodes, system performance gain with scaling diminishes due to undesired effects such as the short-channel effect, increasing middle-end-of-line (MEOL) parasitics and back-end-of-line (BEOL) wire/via resistance (R_{wire}/R_{via}). New transistor and interconnect technologies are developed to sustain the scaling trend, but understanding of the performance gains that these new technologies can bring to systems is lacking. Here we present a design flow that integrates technology modeling and VLSI physical design to implement a commercial 32-bit processor at projected 5-nm design rules to study the contributions of various technologies to system performance, including Si FinFET, nanowire-FET (NWFET), monolayer Black Phosphorous (BP) and MoS_2 FETs, and graphene as BEOL wires. The aim here is to study the impact of different technology characteristics on system performance. While the absolute numbers may vary with assumptions, the conclusions remain valid in general.

Design Flow Overview

Fig. 1 depicts the design flow. Transistor and interconnect are the two main inputs. MEOL parasitic resistance (R_{MEOL}) and capacitance (C_{MEOL}) are extracted at the logic gate-level parasitic extraction step using standard EDA tools to accurately account for the 3D device structures at nanoscale. Physical design steps (place & route) are carried out according to industry standard. For simplicity, the SRAM is removed from the core to focus on the logic; design rules are simplified (e.g. no complications from quadruple patterning) to avoid potential routing congestion to focus on the technology impact on system performance.

Transistor Modeling

Device structures of the FinFET, NWFET, and 2D Planar FETs are illustrated in Fig. 2. The 5-nm design rules are listed in Table I. The gate length (L_G) is chosen such that the subthreshold slope (SS) is maintained at 70 mV/dec for all the FETs [1-4]. The BSIM-CMG model [6] is used to model all the FETs with the carrier mobility (μ) and velocity (v) extracted from experimental or simulation data as shown in Fig. 3. Projected on-state current (I_{ON}) and gate capacitance (C_{gate}) at a nominal V_{DD} = 0.6 V are shown in Fig. 4 and 5. When projecting the experimental-data-based models down to the 5-nm dimension, μ and v are assumed to remain unchanged (an optimistic assumption for bulk-material FETs since μ degrades quickly with shrinking channel thickness [7]), while the electrostatic gate control and capacitances are based on numerical simulations.

Standard-Cell Parasitic Resistance and Capacitance

Fig. 6 shows the components of R_{MEOL} and the transistor source/drain (S/D) resistance (R_{SD}). R_{SD} is calculated analytically [8], using a Gaussian doping profile decaying from N_{SD} at the S/D to N_{Gjun} at the gate-extension junction (Fig. 7). NWFET has higher R_{SD} than FinFET due to the smaller cross-section area of the S/D extensions. For 2D-FETs, R_{SD} = 12.4 kΩ/□·L_{SPA} (experimental data) [9] is used. As shown in Fig. 4, I_{ON} of FinFET is limited by the S/D contact resistance (R_{CON}) due to small L_{CON} constrained by the fixed contacted gate pitch (CGP) and non-scalable L_G, while NWFET current is limited by R_{SD}.

C_{MEOL}, including all metal-to-metal and fringe capacitances calculated based on a 3D field solver, dominates over C_{gate} as shown in Fig. 5. C_{MEOL} vs. gate spacer dielectric constant for FinFET is shown in Fig. 8, indicating that 68% of the total C_{MEOL} is attributed to the gate-to-S/D capacitance.

BEOL Interconnect Resistance and Capacitance

The wire capacitance (C_{wire}) is calculated by the extraction tool with wire dimensions and interlayer dielectric constant as the inputs (see Table I.). Fig. 9 illustrates the model for R_{wire} and R_{via}. Cross-section-area-dependent Cu resistivity (ρ_{Cu}) is calculated based on the Steinhögl model calibrated to recent experimental data [10]. As shown in Fig. 10, while C_{wire} remains roughly constant as the critical dimension (CD) scales down, R_{wire} and R_{via} rise drastically due to increasing electron scattering and the non-scaled Cu diffusion barrier thickness (t_B).

System-Level Performance Analysis: FinFET

Fig. 11 shows total energy consumption vs. clock frequency (E-F) of the processor core implemented using Si FinFET. Each symbol represents a target clock frequency for a V_{DD}. Connecting the optimal E-F points generates a Pareto optimal curve as the performance metric of a technology. Optimal E-F for different FET off-state current (I_{off}) and a slow-corner analysis accounting for device variations are presented in Fig. 12. For simplicity, we will only discuss the nominal-case device with I_{off}= 1nA/μm, but the conclusions will not be affected. Fig. 13 shows the effect of intrinsic drive current (I_{Dint}) on cell-level I_{ON} ($I_{\mathrm{ON,cell}}$) and the core iso-energy clock frequency (f_{CLK}) for different S/D contact resistivity (ρ_{con}). As I_{Dint} increases by 50%, $I_{\mathrm{ON,cell}}$ and f_{CLK} only increase by 23% and 15%, respectively, due to R_{MEOL}, R_{wire}, and R_{via}, even for a low ρ_{con} of 2×10^{-9} Ω-cm^2.

978-1-5090-3902-9/16/$31.00 ©2016 IEEE

NWFET and Gate Length Scaling

One advantage of NWFETs is the good electrostatic gate control, allowing L_G to shrink without increasing subthreshold leakage current, which lowers intrinsic gate capacitance (C_{gint}) and spares more space for L_{CON} and L_{SPA} to reduce MEOL RC. As shown in Fig. 14, reducing MEOL RC is more effective than lowering C_{gint}: for a hypothetical FinFET with L_G that can scale from 18 nm to 11 nm without affecting SS and DIBL, the reduced parasitics will improve core energy-delay-product (EDP) by 30%, while lowering C_{gint} only improves EDP by 13.7%. Optimization of L_{SPA} with a fixed CGP for FinFET and NWFET is shown in Fig. 15. Larger L_{SPA} helps NWFET achieve lower C_{MEOL} and thus lower EDP, despite lower $I_{ON,cell}$. However, Fig. 16 shows that if ρ_{con} can be lowered to 2×10^{-9} Ω-cm^2, FinFET is not limited by R_{CON} anymore and will outperform NWFET, which still suffers from high R_{SD}. To lower the R_{SD} of NWFET to the level of R_{MEOL}, the extensions need to be doped uniformly as high as 2×10^{20} cm^{-3} (rather than a Gaussian profile), which may cause more serious short-channel effect.

2D-Material-Based Transistors: MoS$_2$ and BP

Ultrathin (~1 nm) channel materials such as 1D carbon nanotubes and 2D layered materials provide excellent electrostatic control while maintaining high mobility. Earlier works [11] showed carbon nanotube FETs offer 10× EDP benefit over Si CMOS at constant power density. However, analysis of system-level performance using 2D-FETs is still lacking. Here MoS$_2$ and BP are selected to represent the 2D-FET family because MoS$_2$ is the most "mature" among the family and BP possesses the highest mobility in theory. However, their high R_{CON} may be a problem. To understand the theoretical performance of 2D-FETs and provide a target for $R_{CON,}$ two models are created for each of BP and MoS$_2$: an experimental-data-based model (Exp-model) and a simulation-based model (Sim-model). BP and MoS$_2$ FETs are compared against Si FinFET in Fig. 17-18. Theoretically, BP (MoS$_2$) FET can provide 2.2× (1.7×) better EDP at 0.6 V compared to Si FinFET assuming $\rho_{con}=10^{-8}$. The superior energy efficiency is due to the lower C_{MEOL} (~0.5× of FinFET) thanks to the 2D planar structure and short L_G enabled by the ultrathin channel. However, the lowest reported ρ_{con} to date is 3×10^{-7} Ω-cm^2 for few-layer MoS$_2$ [7], while Fig. 18 shows that $\rho_{con} < 6\times10^{-8}$ Ω-cm^2 is required to compete with Si FinFET. Note that the required ρ_{con} for BP/MoS$_2$ FETs does not need to be as low as the ρ_{con} for Si FinFETs thanks to the shorter L_G (and larger L_{CON}, L_{SPA}). Comparison of the Exp-models and Sim-models shows a gap of 5-7× between today's experiments and theoretical prediction in terms of the core EDP performance.

Graphene as BEOL Interconnect Wires

The growing R_{wire} and R_{via} have raised concerns about the dominance of wires in VLSI systems [12]. Multilayer graphene (MLG) is an alternative to Cu since wire resistivity (ρ_W) of 3.2 μΩ-μm has been demonstrated experimentally [13]. At the 5-nm node, MLG's R_{wire} is ~0.25× of Cu's R_{wire}. However, obtaining low R_{via} for MLG may be a challenge. To explore the potentials of MLG interconnect, three scenarios are considered in Fig. 19: 1) *Cu*: Cu wires and vias as the baseline; 2) *MLG-CuVia*: MLG as wires while R_{via} is assumed to be the same as the *Cu* case; 3) *MLG-EdgeVia*: MLG as wires and R_{via} is calculated based on Fig. 19a. As shown in Fig. 19b-c, MLG's low R_{wire} improves core EDP by 11%, but for *MLG-EdgeVia* whose R_{via} is ~3.6× of the *Cu*, EDP is degraded by 11%. While *MLG-CuVia* gives 73% lower R_{wire} compared to *Cu*, the EDP only improves by 11%, because the place-and-route (P&R) tool can break long wires into short pieces and/or route signals through high-level metal (HLM) to minimize the impact of R_{wire}. As the statistics show in Fig. 20a-b, most wires in the critical paths (CP) use HLM, and the average net length of *MLG-CuVia* is 7% longer than *Cu*. As a result, $R_{wire}+R_{via}$ contributes to 15%-35% of the total CP delay for >90% of the final designs (Fig. 20c). However, the cost is using more vias and cell area as shown in Fig. 20d-e: the *Cu* case needs 12% more vias and cell area than *MLG-CuVia* because more buffers are needed to drive the resistive wires. For this reason, the impact of R_{via} will increase as R_{wire} grows, as shown in Fig. 21 for the sensitivity study of the core EDP to R_{via}: The 2×-R_{wire} case is more sensitive to R_{via} compared to the 0.5×-R_{wire} case. This impact of R_{via} on system performance is not manifest without performing full P&R process. Analysis based on a ring oscillator with fixed wire load will conclude that the higher the R_{wire}, the less sensitive of the system performance to R_{via}, because higher R_{wire} will overshadow the effect of R_{via}.

BEOL Interconnect Resistance Variability

Dimensional scaling not only increases R_{wire} and R_{via}, but also leads to higher variability. Fig. 22 shows the distribution of R_{wire} and R_{via} for different levels of CD variation (σ_W, see W in Fig. 9), assuming σ_W follows a normal distribution. Since R_{wire} and R_{via} rise rapidly with decreasing W, σ_W is the major source for the long tail in the resistance distribution. Monte-Carlo SPICE simulations are performed for the top 10 CPs including actual routing information. Cumulative distribution of the CP delay and 95-percentile delay penalty over 1,000 samples are shown in Fig. 23 for different σ_W and t_B. For the default 2-nm t_B, delay penalty is 5% for σ_W = 1 nm and the penalty grows to 24% for σ_W = 2 nm; whereas for 0-nm t_B (e.g. Co as vias [14] or graphene as Cu diffusion barrier [15]), the penalty is <5% even for σ_W = 2.5 nm. Removing the 2-nm t_B can not only improve the nominal core EDP by 11% but also greatly increase the tolerance of dimensional variations.

Conclusion

Full place-and-route of a 32-bit processor core at 5-nm design rules shows that: 1) NWFET outperforms FinFET because of better L_G scalability and lower C_{MEOL}, while FinFET will be able to maintain its advantage if ρ_{con} can be lowered to 2×10^{-9} Ω-μm^2. 2) 2D-FETs can reduce C_{MEOL} by 50% and theoretically provides ~2× better core EDP compared to Si FinFETs for the same ρ_{con}; $\rho_{con} < 6\times10^{-8}$ Ω-μm^2 is required for 2D-FETs to match the core EDP using Si FinFET with ρ_{con} = 10^{-8} Ω-μm^2 (Note that ρ_{con} of 2D-FETs does not need to be lower than Si FinFETs); 3) Optimized signal routing can mitigate the impact of wire resistance at the cost of cell area and vias. As R_{wire} increases, impact of R_{via} also increases. 4) Thinning Cu diffusion barrier can improve core EDP up to 11% and tolerate 2× more variation in the critical dimension.

Acknowledgement

Supported in part by Stanford SystemX Alliance, Stanford INMP, NSF NCN-NEEDS, STARnet SONIC. Work by CSL performed in part at ARM Inc.

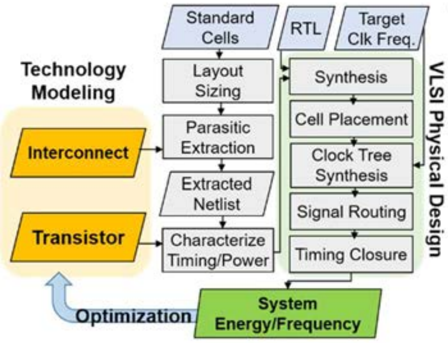


Fig. 1. Integrated design flow for technology modeling and VLSI system implementation.

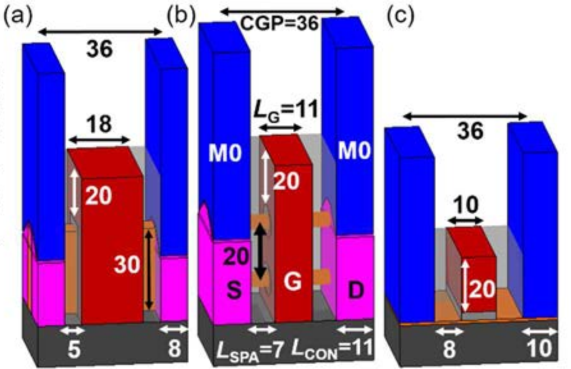


Fig. 2. FET device structures of (a) FinFET, (b) NWFET, and (c) 2D-FETs. All labeled numbers are in the unit of nm. The Fin and NW widths are both 5 nm.

Table I. Projected 5-nm Design Rules. Aspect ratio of 2 is assumed for all metal wires and vias.

layer	Width	Pitch
Contacted gate pitch	-	36 nm
Fin/NW stack	5 nm	21 nm
Metal 1/2/3	12 nm	24 nm
Metal 4/5	18 nm	36 nm
Metal 6/7	24 nm	48 nm
Via 0/1/2/3	12 nm	-
Via 4/5	18 nm	-
Via 6/7	24 nm	-

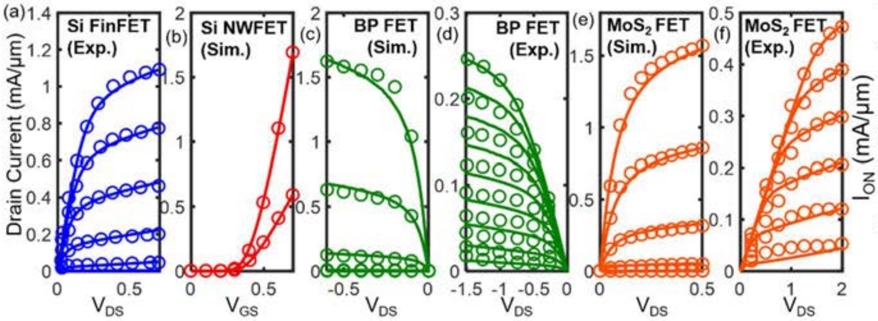


Fig. 3. FET model fitting to experimental (Exp.) or simulation (Sim.) data to extract carrier mobility and velocity. (a) 14-nm Si FinFET [5]. (b) Si NWFET [2]. (c) Sim. BPFET [3]. (d) Exp. BPFET [16]. (e) Sim. MoS$_2$FET [4]. (f) Exp. MoS$_2$FET [17]. Symbol: data; line: model.

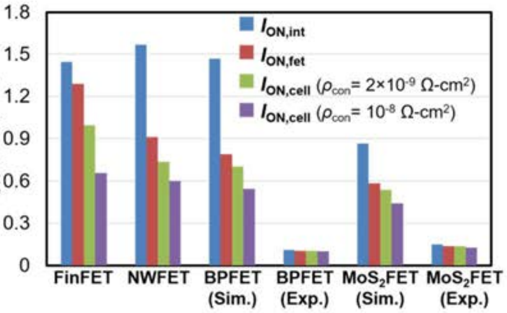


Fig. 4. Projected on-state current (I_{ON}) of all the FETs at nominal V_{DD} of 0.6 V and fixed I_{off} of 1 nA/μm. $I_{ON,int}$: intrinsic I_{ON}; $I_{ON,fet}$: I_{ON} including R_{SD}; $I_{ON,cell}$: I_{ON} including R_{SD}+R_{MEOL}.

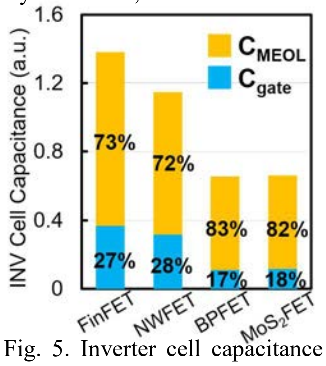


Fig. 5. Inverter cell capacitance breakdown. C_{gate}: transistor-level gate capacitance. C_{MEOL}: MEOL parasitic capacitance.

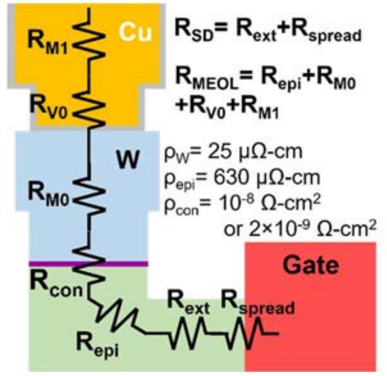

R_{M1}
Cu
R_{V0}
W
R_{M0}
R_{con}
R_{ext}
R_{spread}
R_{epi}
Gate
R_{SD}= R_{ext}+R_{spread}
R_{MEOL}= R_{epi}+R_{M0} +R_{V0}+R_{M1}
ρ_W= 25 μΩ-cm
ρ_{epi}= 630 μΩ-cm
ρ_{con}= 10^{-8} Ω-cm^2 or 2×10^{-9} Ω-cm^2

Fig. 6. Illustration of the parasitic resistance components. Resistivity used to calculate R_{MEOL} during parasitic extraction is shown.

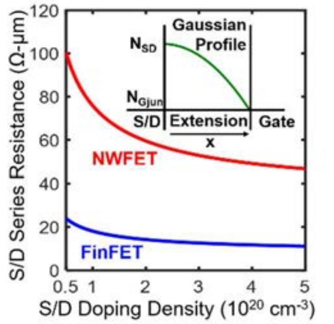


Fig. 7. Source/Drain series resistance (per footprint) vs. doping density. Inset: Gaussian doping profile along the extension. Default N_{SD} = 2×10^{20} cm^{-3}, N_{Gjun} = 10^{19} cm^{-3}.

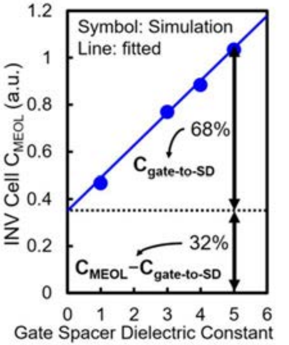


Fig. 8. FinFET MEOL parasitic capacitance of an inverter cell vs. the gate spacer dielectric constant.

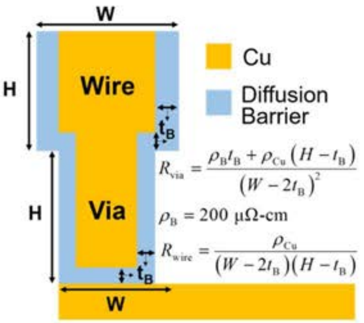

W
H
Wire
H
Via
t_B
t_B
W
Cu
Diffusion Barrier
$R_{via} = \frac{\rho_B t_B + \rho_{Cu}(H - t_B)}{(W - 2t_B)^2}$
ρ_B = 200 μΩ-cm
$R_{wire} = \frac{\rho_{Cu}}{(W - 2t_B)(H - t_B)}$

Fig. 9. Illustration of BEOL wire and via resistance model based on dual damascene process. Default Cu diffusion barrier thickness is 2 nm.

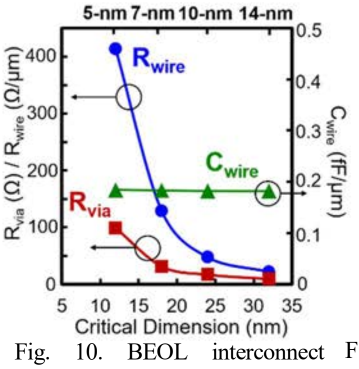


Fig. 10. BEOL interconnect resistance and capacitance vs. the critical dimension (W in Fig. 9). Aspect ratio of 2 is assumed.

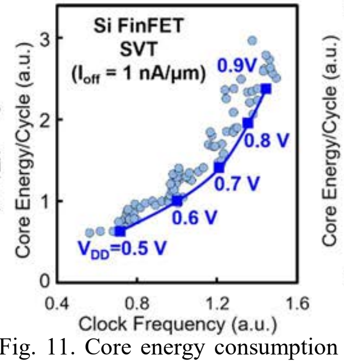


Fig. 11. Core energy consumption per cycle vs. clock frequency on Si FinFET. Each symbol represents a target clock frequency for a V_{DD}. Line: optimal energy curve.

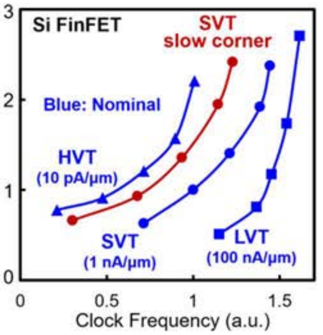


Fig. 12. Core energy vs. frequency for different V_T's. Slow corner: 10% variation in V_{DD}, ρ_{con}, and carrier mobility and velocity. σ_{VT} = 30 mV.

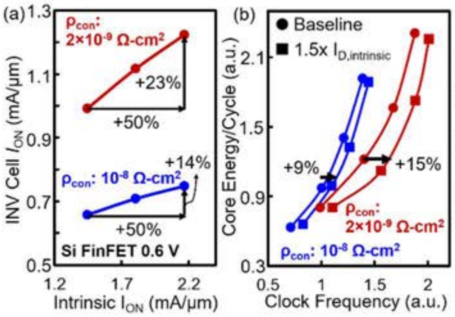


Fig. 13. Effect of 50% increase in the Si FinFET intrinsic current on: (a) cell-level on-current and (b) core iso-energy clock frequency for different contact resistivity.

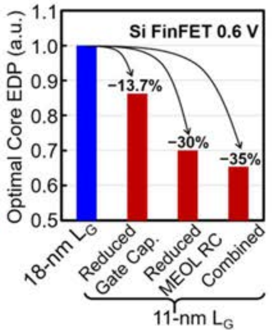


Fig. 14. Performance benefit breakdown of downscaling gate length from 18 nm to 11 nm, assuming SS unchanged.

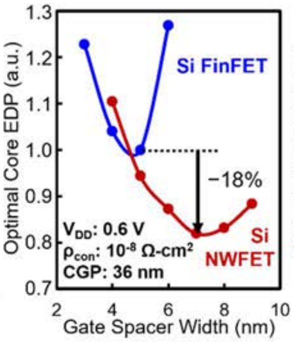


Fig. 15. Optimization of the gate spacer width of FinFET and NWFET to minimize core EDP with a fixed CGP.

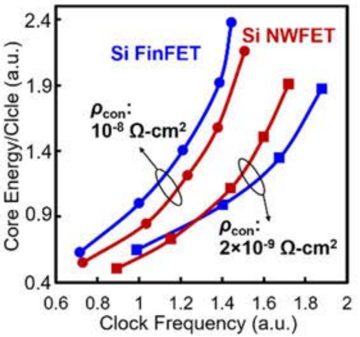


Fig. 16. Core energy consumption vs. clock frequency of FinFET and NWFET for various source/drain contact resistivities.

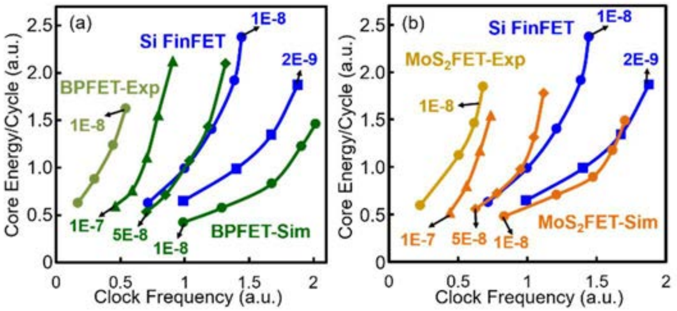
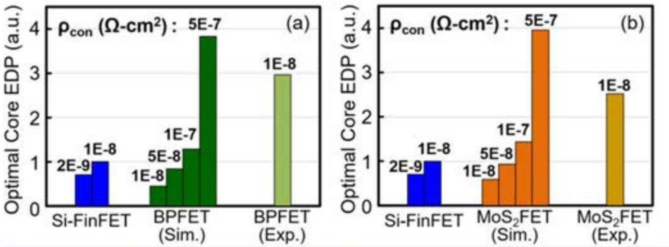
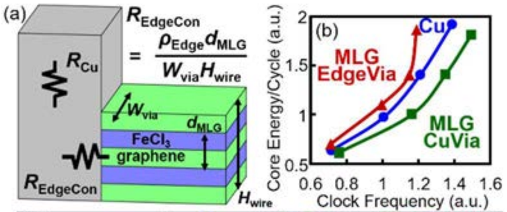


Fig. 17. Comparison of core energy-frequency between (a) BPFET and (b) MoS₂FET against Si FinFET for various contact resistivities ρ_{con} as labeled next to the curves (unit: Ω-cm²). Channel thickness: $t_{BP} = t_{MoS2} = 0.7$ nm and $t_{Si\text{-}Fin} = 5$ nm.



(c) Match	Required BP ρ_{con}	Required MoS$_2$ ρ_{con}
Si FinFET (ρ_{con} 2E-9 Ω-cm²)	3.7E-8 Ω-cm²	2.4E-8 Ω-cm²
Si FinFET (ρ_{con} 1E-8 Ω-cm²)	6.8E-8 Ω-cm²	5.5E-8 Ω-cm²

Fig. 18. Comparison of core EDP at 0.6 V between (a) BPFET and (b) MoS₂FET against Si FinFET for different contact resistivity (labeled on top of the bars) and (c) the required ρ_{con} for BP and MoS₂FETs to match Si-FinFET's EDP performance.



(c)	Cu	MLG-CuVia	MLG-EdgeVia
R_{wire} (Ω/μm)	413	111	111
R_{via} (Ω)	98	98	367
Min. EDP (a.u.)	1	0.89	1.11

Fig. 19. (a) MLG edge-contact scenario. Experimental result of ρ_{Edge} = 150 Ω-μm [18] and d_{MLG} = 0.68 nm are used. FeCl₃ serves to dope the graphene [13]. (b) Core energy vs. frequency. (c) Features of the three interconnect scenarios, whose wire/via dimensions are identical as specified in Table I.

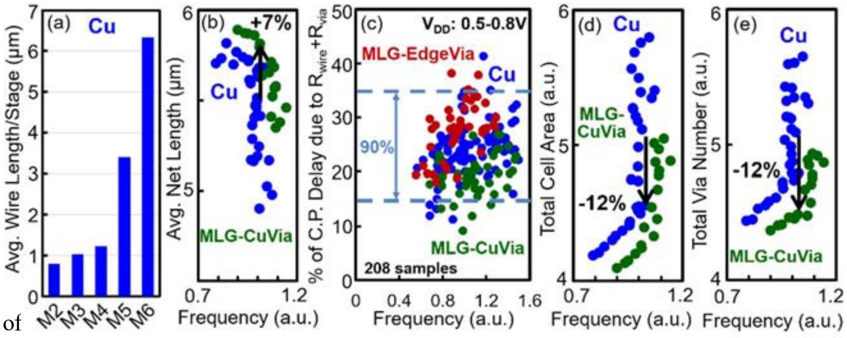


Fig. 20. Core design statistics for the *Cu* and *MLG-CuVia* interconnect models. (a) Average wire length at each layer between each stage in the top 10 critical paths. (b) Average net length. (c) Contribution of $R_{wire}+R_{via}$ to the total critical path delay. (d) Total standard cell area. (e) Total via number.

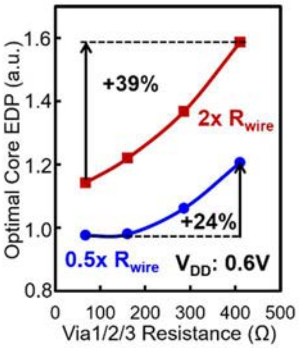


Fig. 21. Sensitivity of optimal core EDP to via resistance for two different wire resistances: 0.5× R_{wire} and 2× R_{wire}.

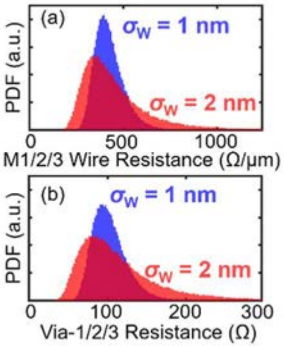


Fig. 22. Probability density function of (a) wire and (b) via resistances for different levels of critical dimension variation (σ_W).

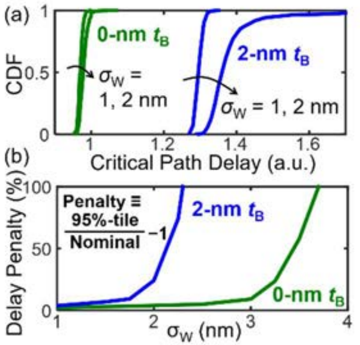


Fig. 23. (a) Cumulative distribution function and (b) penalty of top 10 critical path delay vs. critical dimension variation for 0-nm and 2-nm Cu diffusion barrier thickness (t_B).

Reference

[1] A. Thean, *VLSI Tech. Symp.*, pp. 3.3, 2014.
[2] M. Choi, *SISPAD*, pp. 242, 2015.
[3] X. Cao, *T-ED*, pp. 659, 2015.
[4] L. Liu, *T-ED*, pp. 4133, 2013.
[5] S. Natarajan, *IEDM*, pp. 3.7, 2014.
[6] N. Paydavos, *IEEE Access*, pp. 201, 2013.
[7] C. English, *Nano Lett*., pp. 3824, 2016.
[8] C.-W. Sohn, *T-ED*, pp. 1302, 2013.
[9] A. Rai, *Nano Lett*., pp. 4329, 2015.
[10] A. Pyzyna, *VLSI Tech. Symp*., pp. 120, 2015.
[11] D. Frank, *IEDM Short Course*, 2012.
[12] G. Yeric, *IEDM Short Course*, 2014.
[13] D. Kondo, *IITC*, pp. 189, 2014.
[14] M. H. Veen, *IITC*, pp. 25, 2015.
[15] L. Li, *VLSI Tech. Symp*., pp. 122, 2015.
[16] N. Haratipour, *EDL*, pp. 411, 2015.
[17] L. Yang, *DRC*, pp. 237, 2015.
[18] L. Wang, *Science*, pp. 614, 2014.