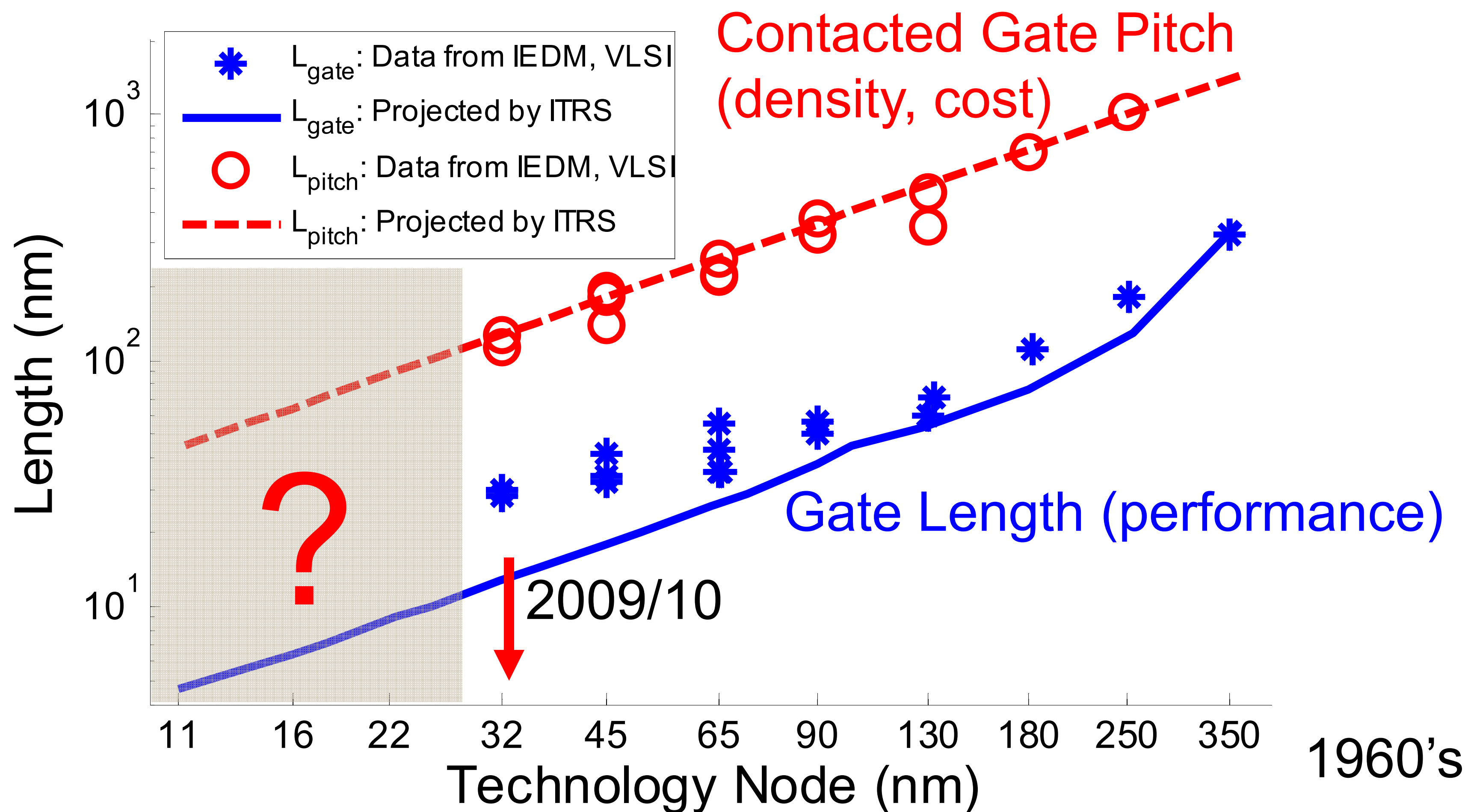
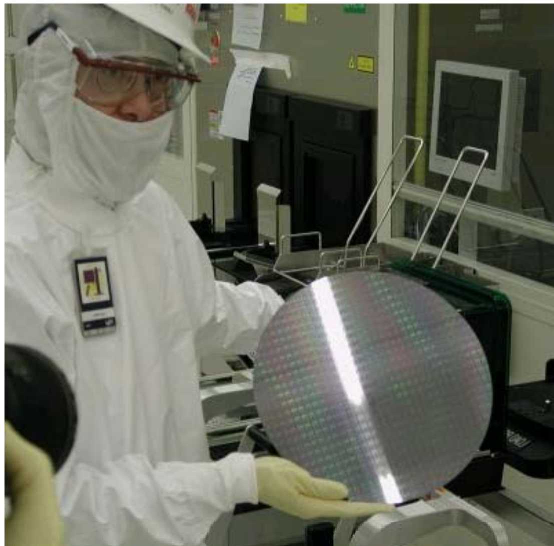
Compact Modeling Aided Technology Design and Projection Considering System-Level Performance

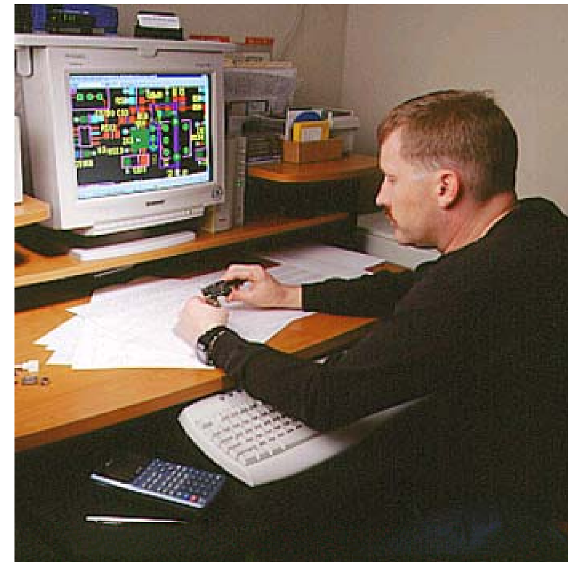
Lan Wei, H.-S. Philip Wong
Stanford University
{lanw, hspwong}@stanford.edu

CMOS Device Pitch Scaling



In the old days…





Device → Library → Circuit

Proportional scaling → Intrinsic on-state performance improvement → Delay, power, area

In the future

Non-scalable factors

- Leakage
- V_{dd} (V_{th})
- On current
- Parasitics – R, C

Device +Circuit

Outline

Near future CMOS Technology

- Gate control vs. current drive

Mid future CMOS Technology

- Parasitic engineering

Far future Post-Si Technology

- Carbon nanotube FET (CNFET)

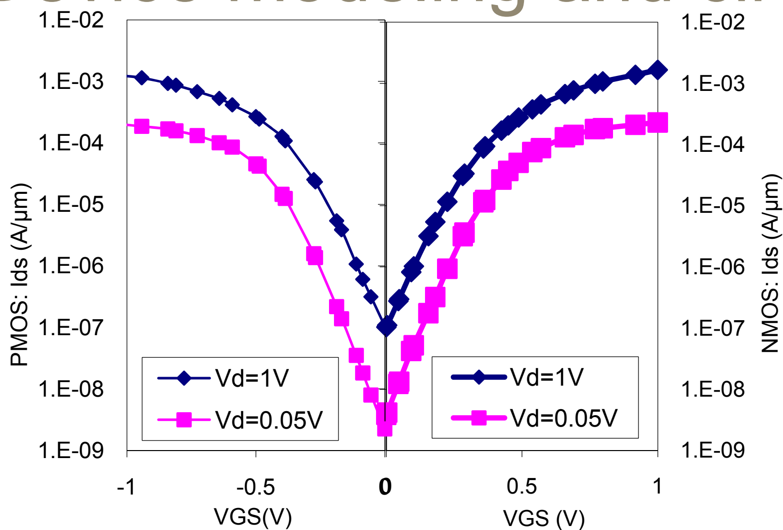
CMOS Scaling Today

- **Present path**
  - Higher velocity (strain, exotic channel materials Ge, III-V)
  - Good electrostatics for gate length scaling
    - Ultra-thin body SOI, FinFET
    - 1D FETs (nanowires, carbon nanotubes)

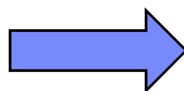
Which way is more efficient?

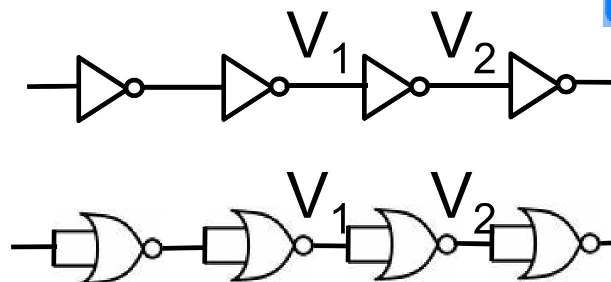
$$\tau = \frac{CV}{I} = \frac{\left(C_{\text{intrinsic}} + C_{par}\right) \cdot V}{I}$$

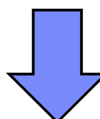
Device modeling and circuit simulation

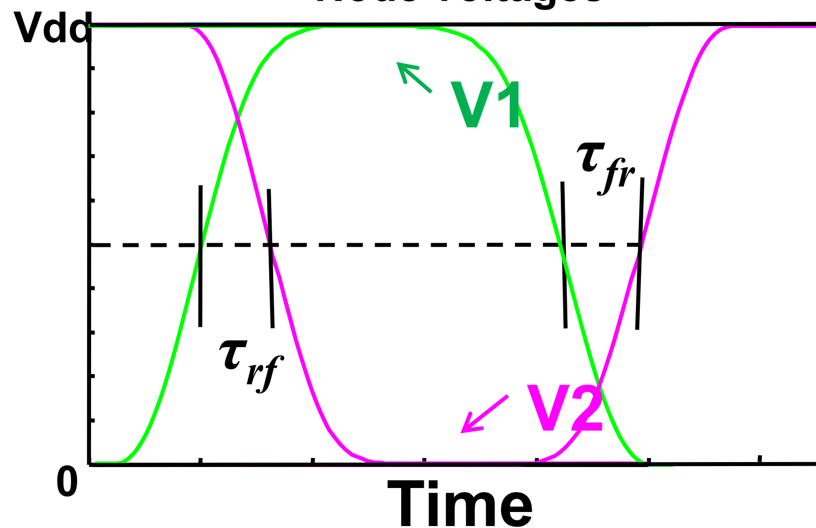


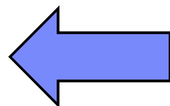
Ref: A. Khakifirooz et al, *T-ED09*
S, Natarajari, et al, *IEDM 08.*

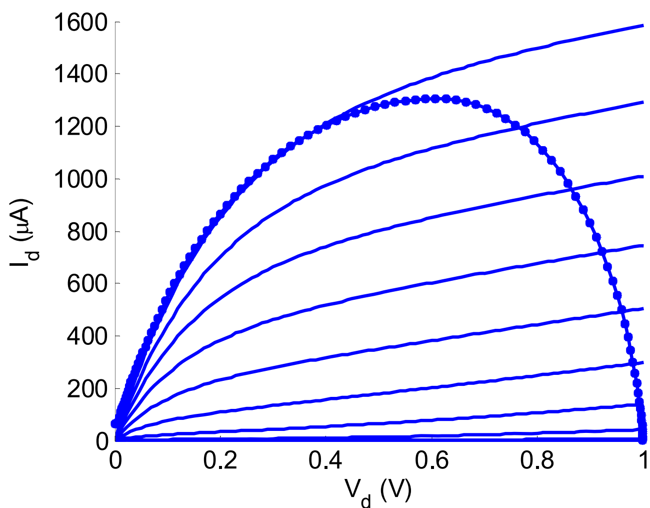






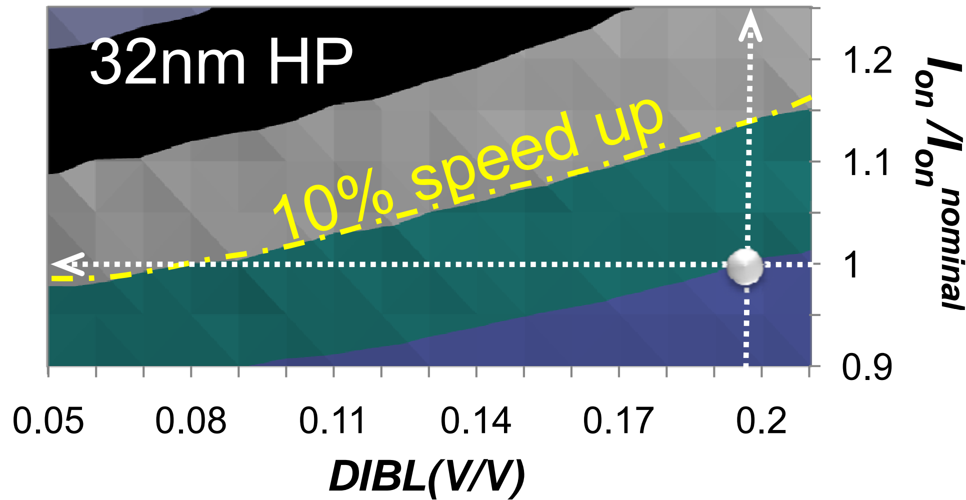






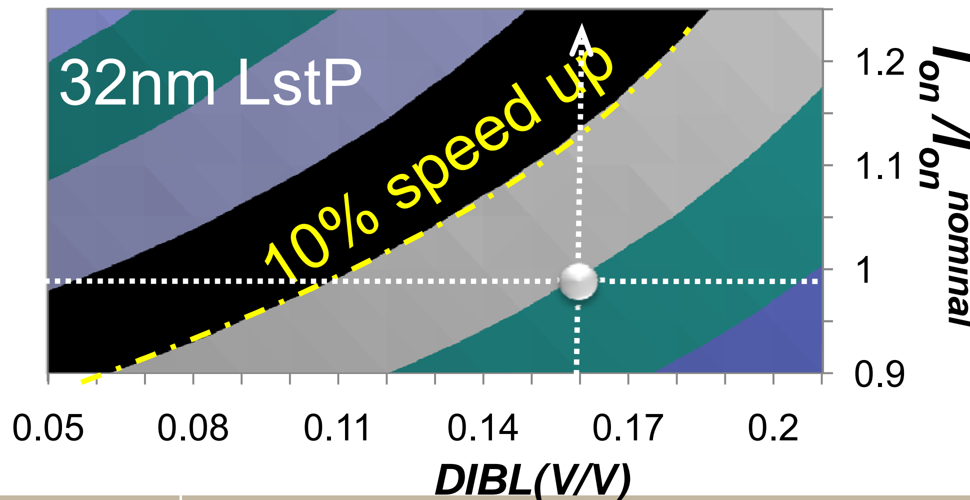
L. Wei et al, SSDM2009
T. Skotnicki , IEDM 2009 Short course

Efficiency depends on applications



HP: Improve I_{on} is efficient

Transport property: strain/new materials

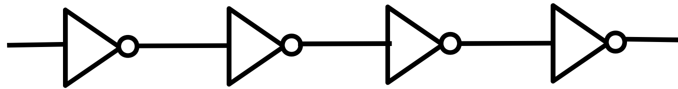


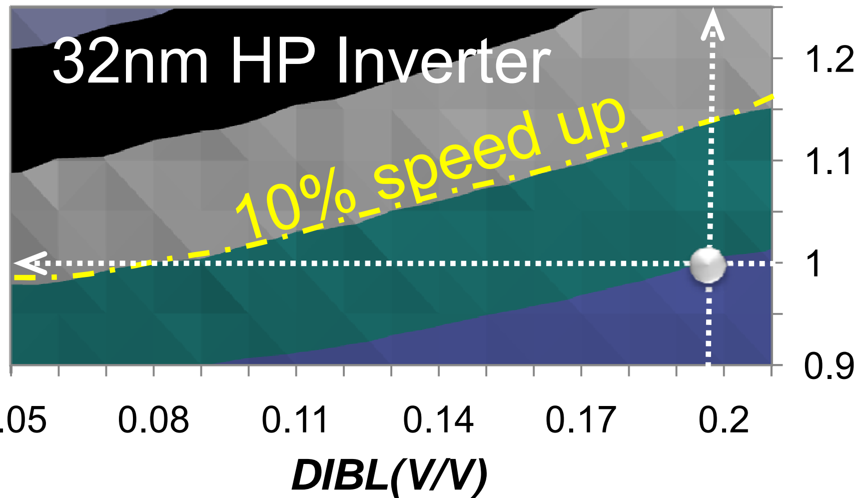
Improve *DIBL* is efficient

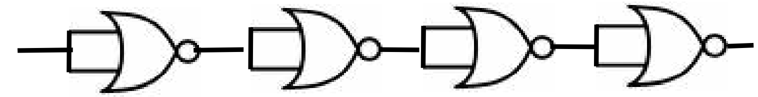
Electrostatic property: FDSOI/FinFET/GAA

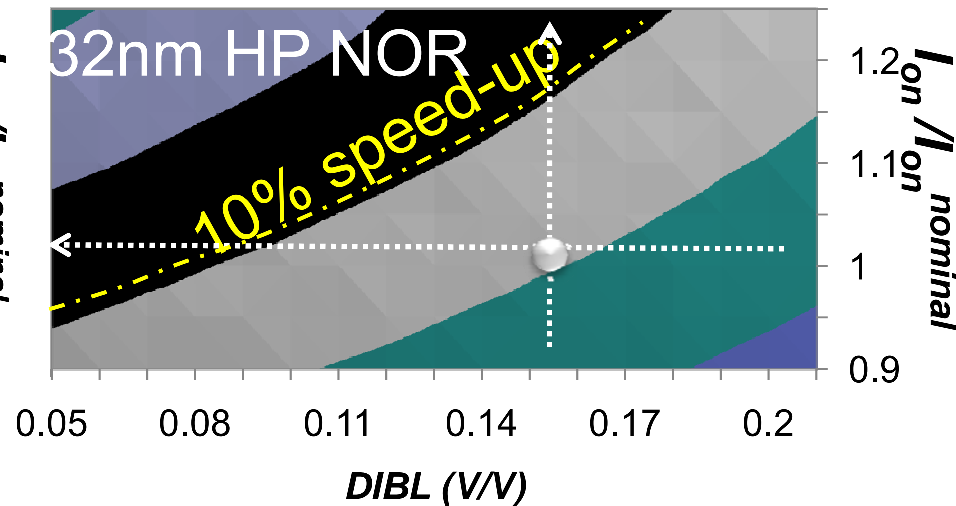
L. Wei et al, SSDM2009
T. Skotnicki , IEDM 2009 Short course

Efficiency also depends on circuit complexity









INV Improve I_{on} is efficient

Transport property: strain/new materials

NAND/NOR: Improve *DIBL* is efficient

Electrostatic property: FDSOI/FinFET/GAA

Take-home message #1

The most efficient knob to improve device performance at circuit-level depends on

- Application
- Circuit design

The Future of CMOS Scaling

- **In the limit of $L_g \rightarrow 0$ and ballistic transport**
  - Speed is determined by parasitic capacitance and series resistance
  - Realistic performance estimation must include parasitics
    - A 3D problem, some parastics do not scale

$$\tau = \frac{CV}{I} = C \cdot R = \left(C_{\text{intrinsic}} + C_{par}\right) \cdot \left(R_{channel} + R_{series}\right)$$

< 50%   > 50%   < 70%   > 30%

Performance Boosting By Parasitics Engineering

$$\tau = \frac{CV}{I} = C \cdot R = \left(C_{\text{intrinsic}} + C_{par}\right) \cdot \left(R_{channel} + R_{series}\right)$$

50% 50% 70% 30%

Performance Boosting By Parasitics Engineering

$$\tau = \frac{CV}{I} = C \cdot R = \left(C_{\text{intrinsic}} + C_{par}\right) \cdot \left(R_{channel} + R_{series}\right)$$

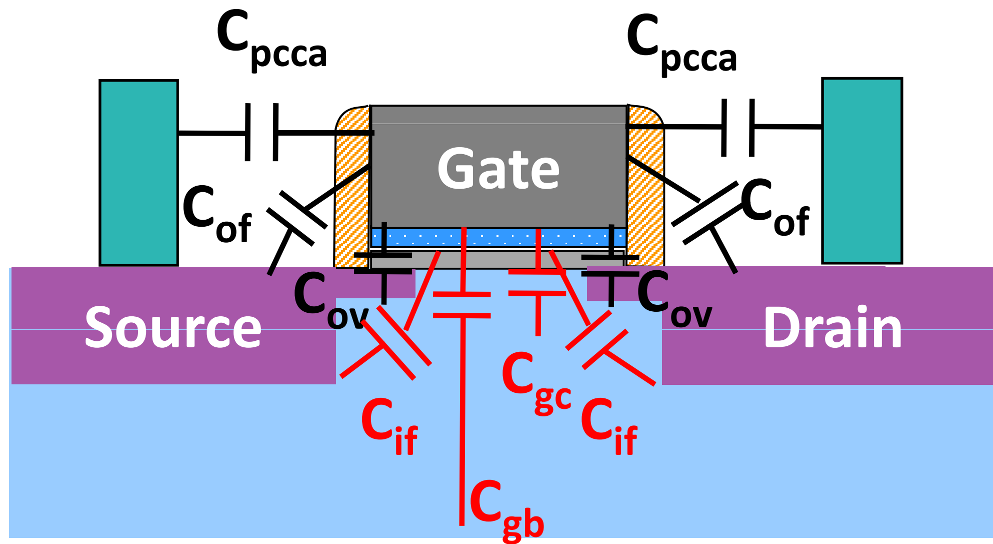
50% 50% 70% 30%

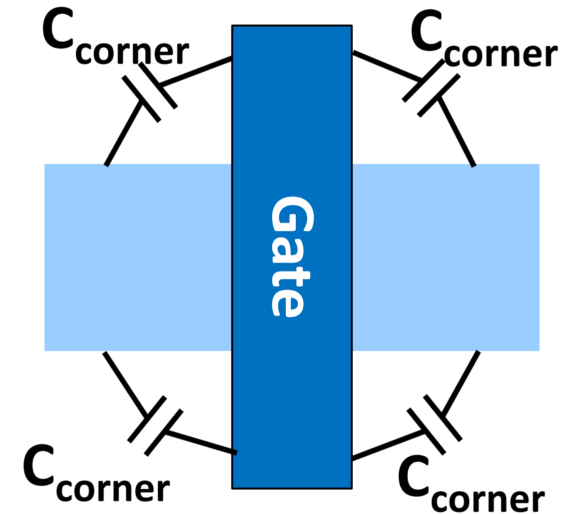
Reduce parasitics when:

- Keep scaling L_{pitch}
- Relax L_{gate} scaling

Gate *Contact, isolation, spacer, etc.* *Pitch*

Capacitance Components





Ref: L. Wei et al, VLSI-TSA 2009
T. Skotnicki, IEDM 2009, short course

Capacitance Model

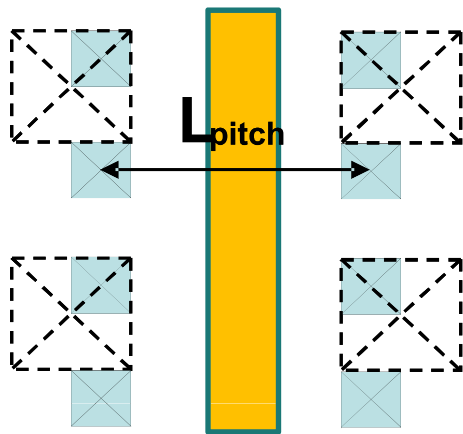
- **Implemented in MASTAR5**
  - Used by ITRS Roadmap 2009 Edition
  - Bulk, FDSOI, Double-gate FET, FinFET
  - http://www.itrs.net/models/htm

gate-to-channel capacitance	$C_{gc_bulk}=C_{gc_SOI}=\frac{\varepsilon_{ox}}{t_{ox_el}}(L_g-\Delta L)W, C_{gc_DG}=2C_{gc_bulk}.$
overlap capacitance	$C_{ov_bulk}=C_{ov_SOI}=0.5\frac{\varepsilon_{ox}}{t_{ox_el}}\Delta L\cdot W, C_{ov_DG}=2C_{ov_bulk}.$
gate-to-substrate off-state capacitance	$C_{gb_bulk}=series\left[\frac{\varepsilon_{ox}}{t_{ox}}(L-\Delta L)W,\frac{\varepsilon_{si}}{t_{dep}}(L-\Delta L)W\right], C_{gb_DG}=0,$ $C_{gb_SOI}=series\left[\frac{\varepsilon_{ox}}{t_{ox}}(L_g-\Delta L)W,\frac{\varepsilon_{si}}{t_{si}}(L_g-\Delta L)W,\frac{\varepsilon_{box}}{t_{box}}(L_g-\Delta L)W\right]$
inner-fringe capacitance	$C_{if_bulk,SOI}=\frac{2\varepsilon_{si}}{\pi}W\ln\left(1+\frac{\min(0.5L_g-t_{ox},t_{eff})}{t_{ox}}\right), C_{if_DG}=\frac{4\varepsilon_{si}}{\pi}W\ln\left(1+\frac{\min(0.5L_g-t_{ox},t_{eff})}{t_{ox}}\right)$
outer-fringe capacitance	$C_{of_bulk}=C_{of_SOI}=0.5C_{of_DG}$ $=\frac{\varepsilon_{cap}W}{\pi}\ln\left(\frac{PCCA+\sqrt{t_{ox}^2+PCCA^2}}{t_{ox}}\right)+0.35\frac{\varepsilon_{cap}W}{\pi}\ln\left(\frac{\pi W}{t_{ox}}\right)$

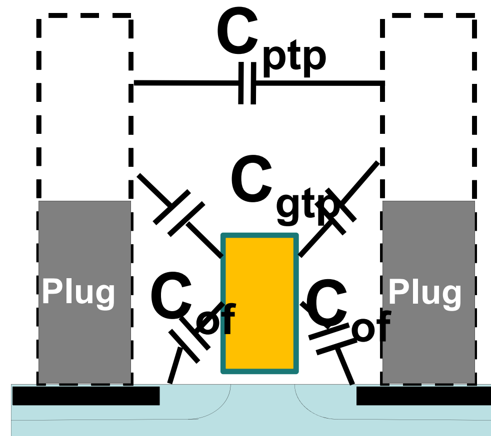
gate-to-contace capacitance	$C_{PCCA}=\frac{0.5\pi\varepsilon_{cap}W}{\ln\left[\frac{2\pi(PCCA+L_g)}{2L_g+\tau_{bk}\cdot t_{poly_eff}}\right]}+\frac{\varepsilon_{cap}Wt_{poly_eff}}{PCCA}+\frac{\varepsilon_{spacer}Wt_{rSD_eff}}{t_{spacer}}$ $\tau_{bk}=\exp\left[2-2\sqrt{1+\frac{2(t_{poly_eff}+L_g)}{PCCA}}\right]$
corner capacitance	$C_{corner_bulk}=C_{corner_SOI}=\frac{2\pi\varepsilon_{cap}(t_{poly_eff}+2W_{ext})}{\cosh^{-1}\left(\frac{t_{eff}+2t_{ox}}{t_{eff}}\right)}, C_{corner_DG}=2C_{corner_bulk}$
junction capacitance	$C_{j_bulk}=W\left[x_j+0.5(L_{pitch}-L_g)\right]\sqrt{\frac{q\varepsilon_{si}}{2V_{bi}}\frac{N_aN_d}{N_a+N_d}}\Big/\left(1+\frac{V}{V_{bi}}\right)^{0.33}$ $C_{j_SOI}=\frac{\varepsilon_{box}}{t_{box}}W\cdot 0.5(L_{pitch}-L_g),\ C_{j_DG}=0.$
	$t_{eff}=x_j(bulk), t_{si}(SOI)$ or $0.5t_{si}(DG)$ $t_{rSD_eff}=0(bulk,DG)$ or $t_{rSD}(SOI)$ $t_{poly_eff}=t_{poly}(bulk,DG)$ or $(t_{poly}-t_{rSD})(SOI)$

L. Wei et al, VLSI-TSA 2009

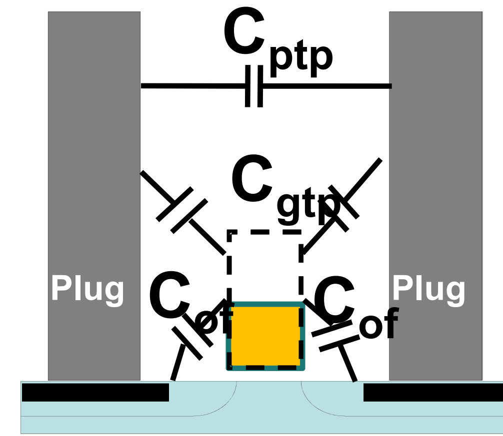
Selective Device Structure Scaling: Basic Idea



(a): Reduce L_{cont}

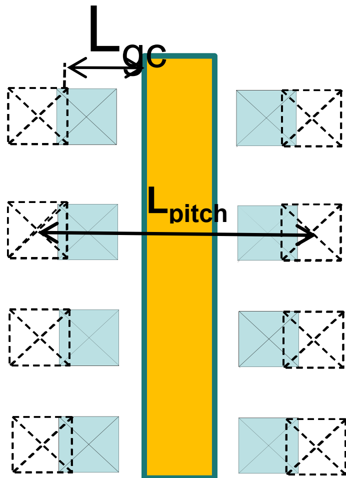


(b): Reduce H_{plug}



(c): Reduce H_{gate}

Parasitic capacitance ↓
Series resistance ↑



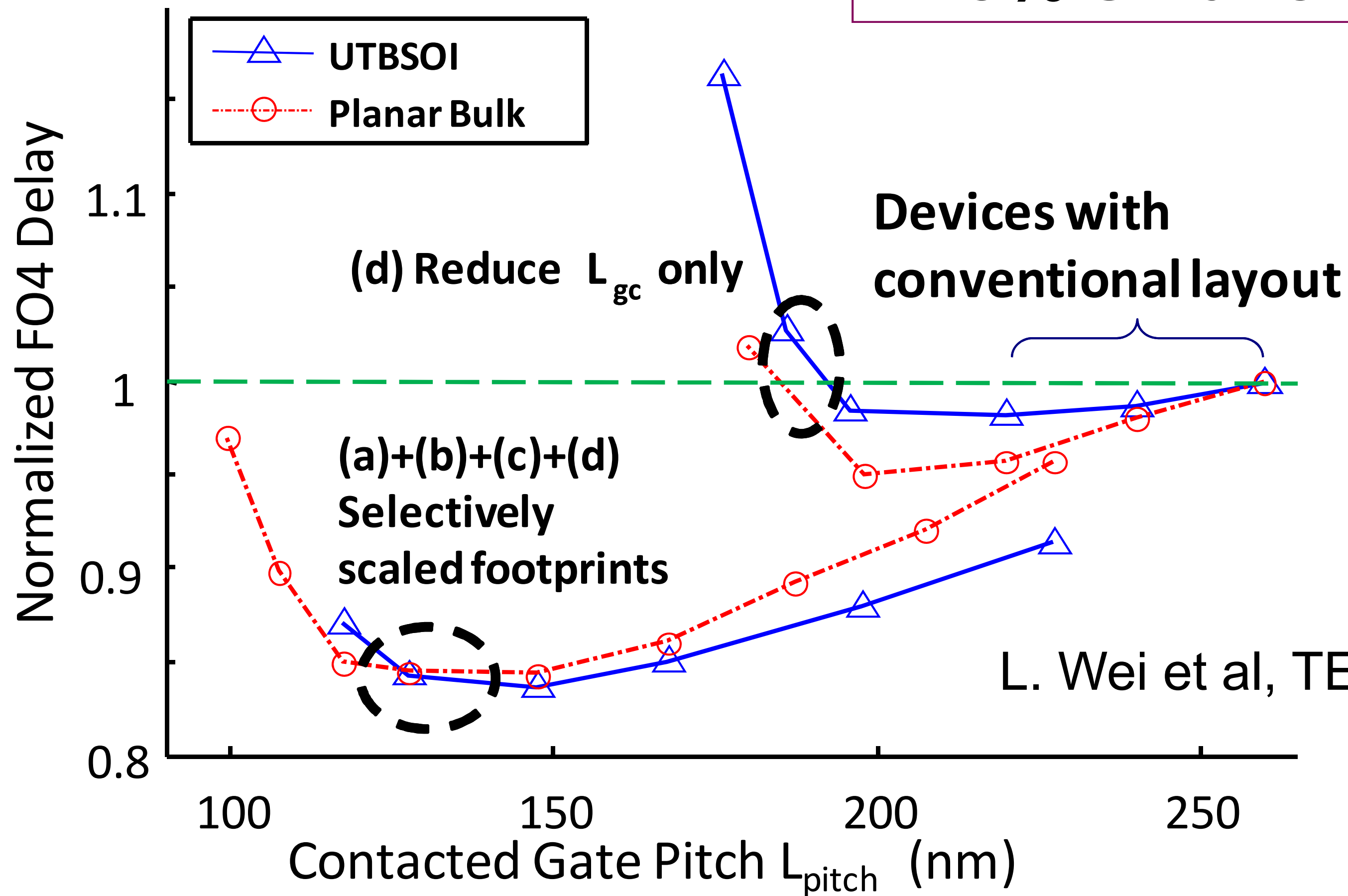
(d): Reduce L_{gc}

Parasitic capacitance ↑
Series resistance ↓

$$\tau = R \cdot C \downarrow$$

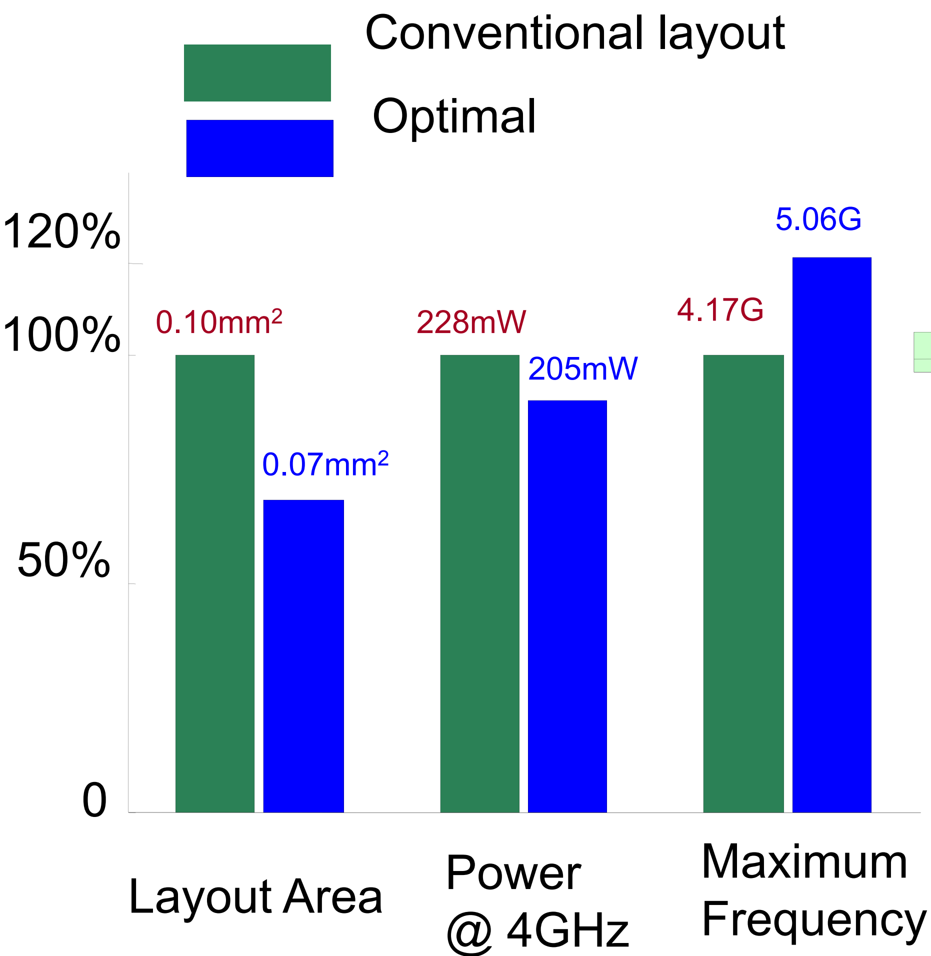
Inverter Delay Improvement – (bulk CMOS and UTBSOI)

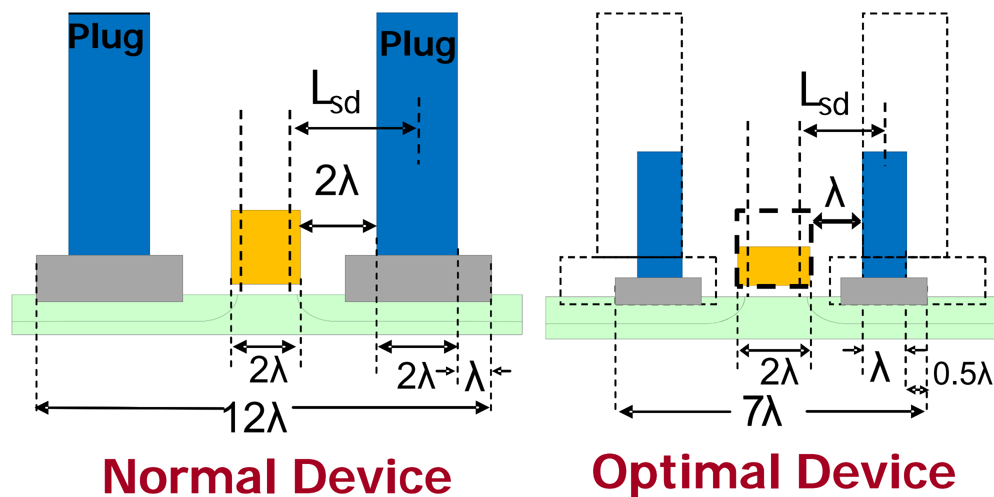
- **15% faster**
- **45% smaller area**



L. Wei et al, TED 2009.

Circuit-Level Improvement



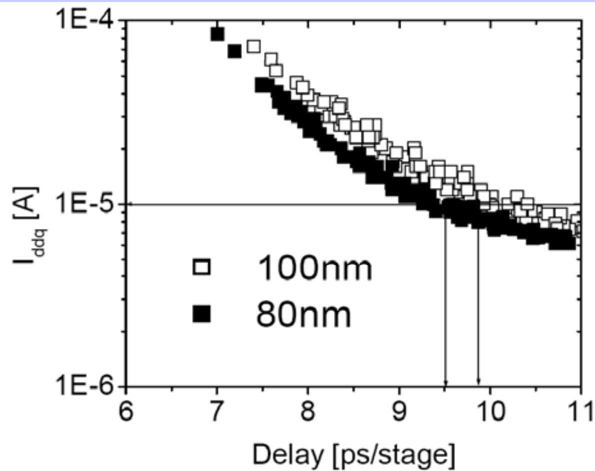


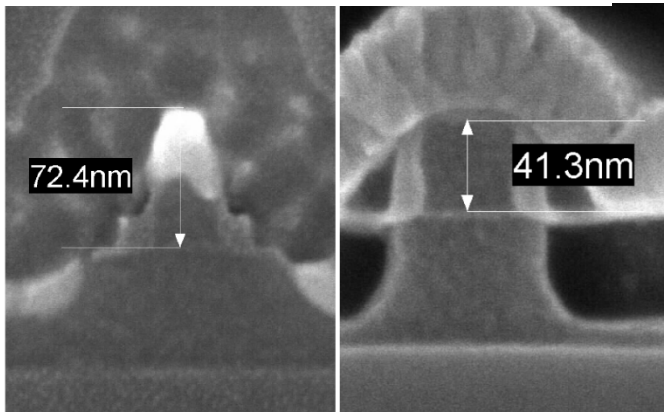
- Layout area: **30% smaller**
- Speed: **25% faster** (Device 15%+Interconnect 10%)
- Power: **10% smaller**

L. Wei, et al, TED 2009.

Device Structure Scaling – Experimental Work

Reduced Gate Height

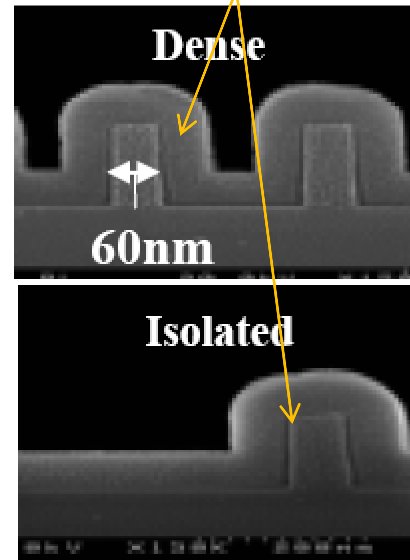


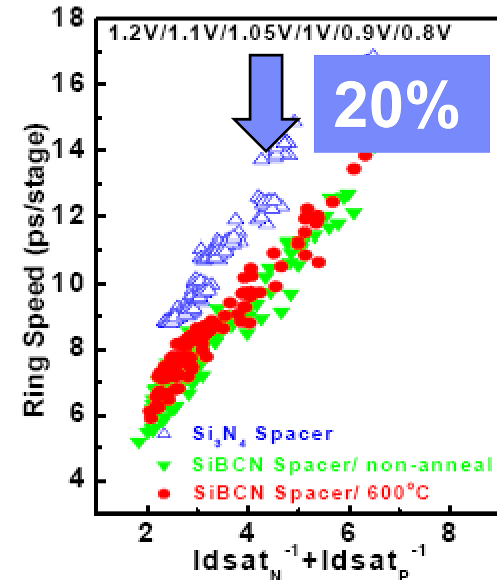


Z. Ren et al. ICSICT, paper A1.6 (2008) [IBM]

SiBCN (k=5.2)

Low-k spacer





Low-k spacer lowers the gate-fringing capacitance, thus speeds up the ring oscillator

Ref: Ko, et al, VLSI 2008, pp. 108-109 [TSMC]

Take-home message #2

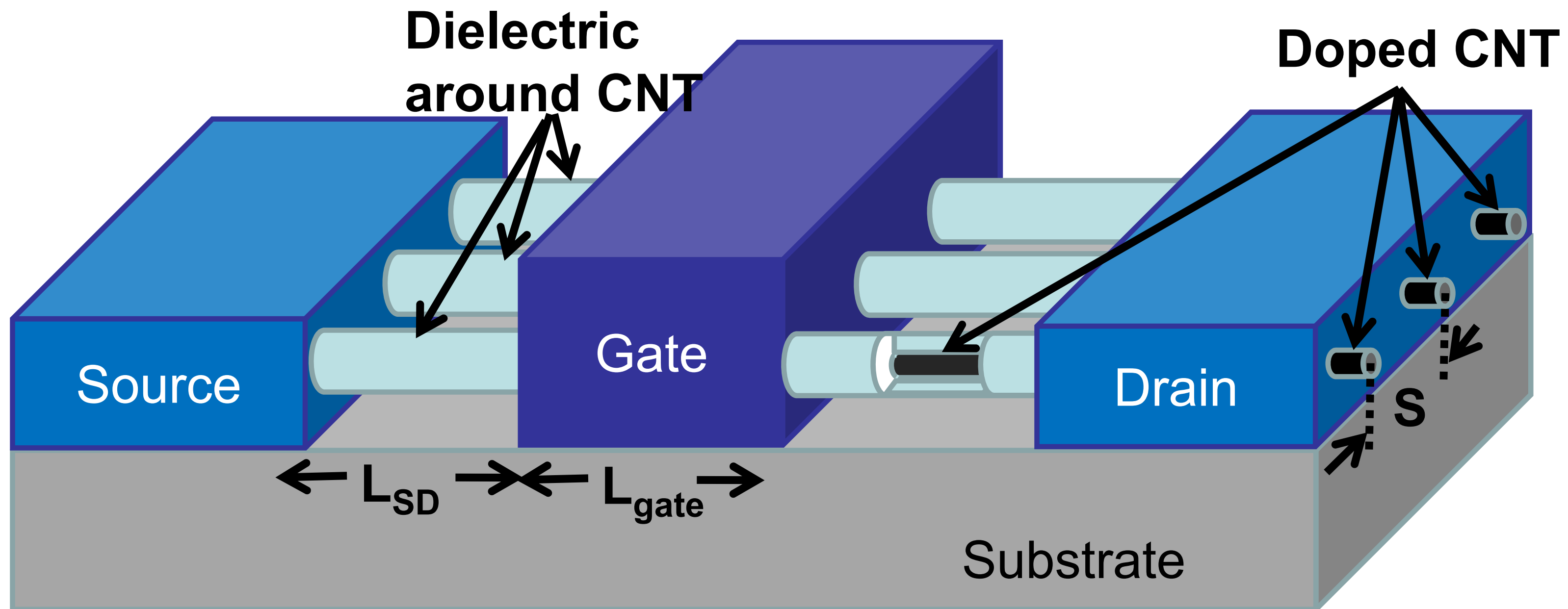
Parasitic engineering is vital for circuit-level performance

- **"Selective device structure scaling"** can effectively improve device performance at the circuit-level and extend the technology roadmap for a few generations beyond the currently perceived limits.

Carbon Nanotube FET (CNFET)

TECHNOLOGY
Structure innovation
Material innovation

DESIGN
Optimization
System-level impact

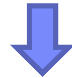


Avouris, et al, Nature Nanotechnology, 2007
L. Wei, et al, IEDM 2009

L. Wei et al, IEDM 2009
L. Wei et al, ESSDERC 2008

Transport Model

Ballistic transport

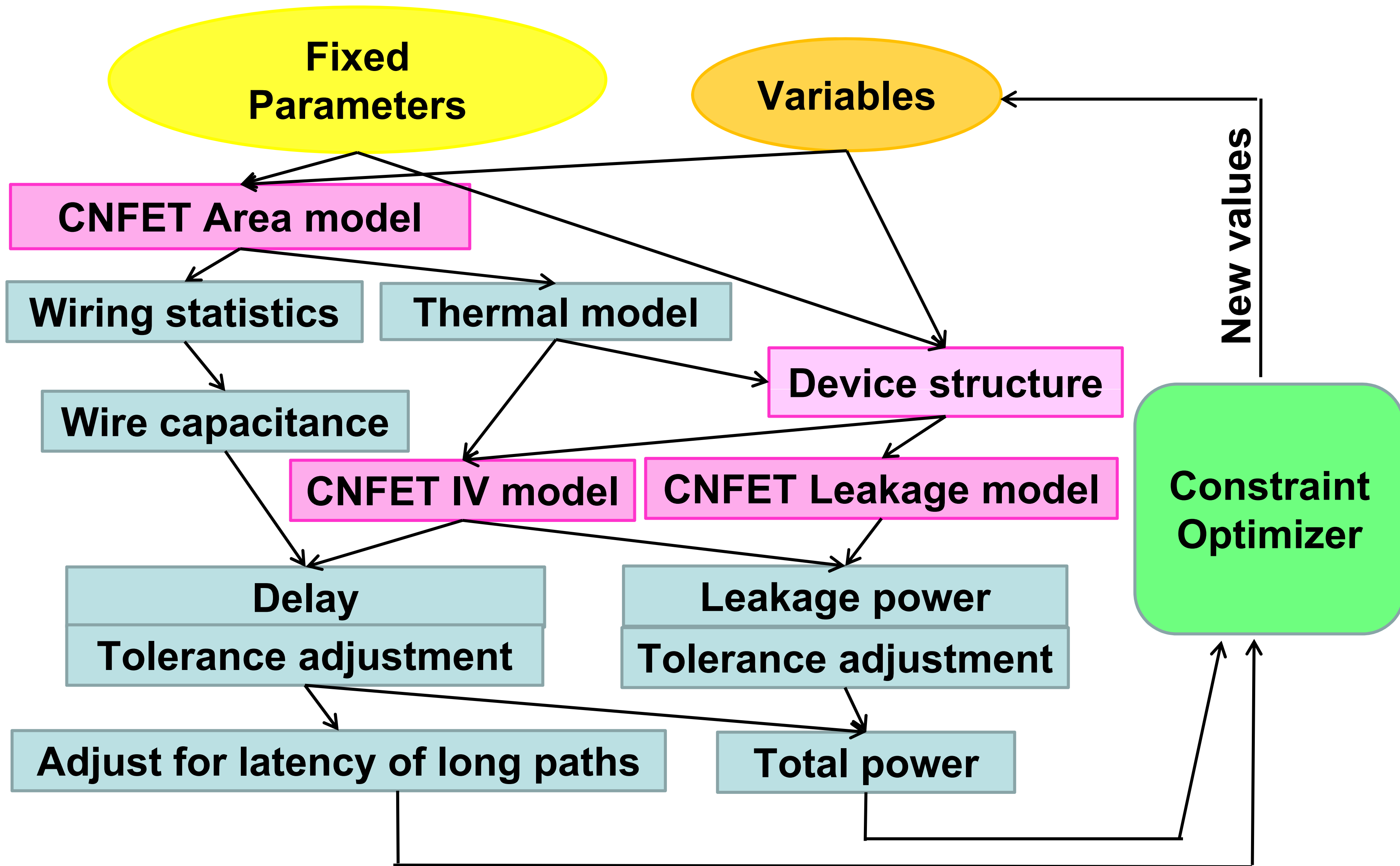


Piecewise constant quantum capacitances

Non-iterative model for φ_{ch}

Highest barrier=max (φ_{ch}, φ_S, φ_D)

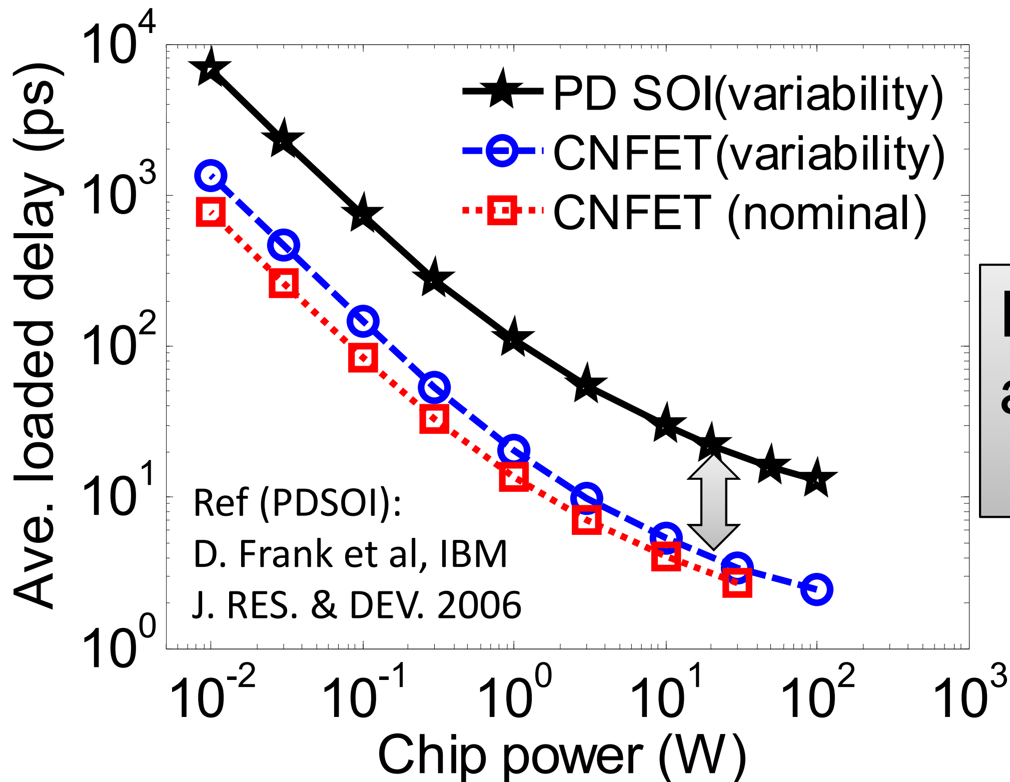
Non-iterative model for intrinsic device current



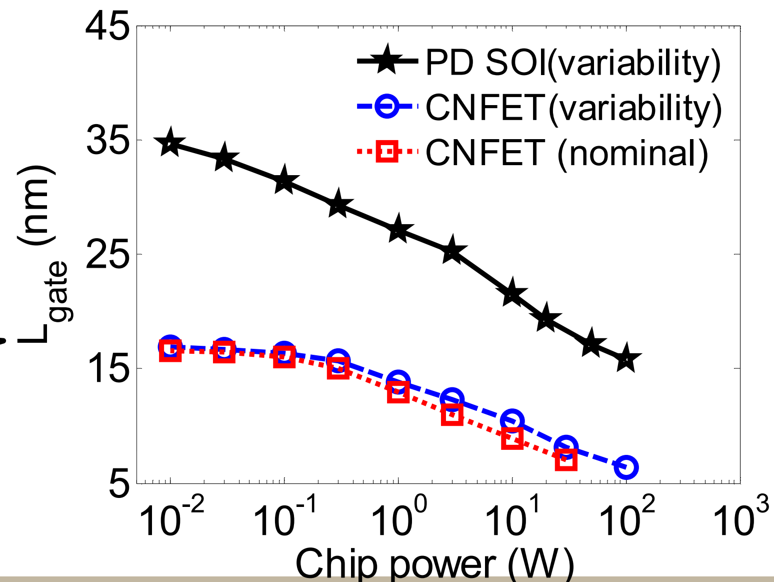
D. Frank et al, IBM J. RES. & DEV. 2006

L. Wei et al, IEDM 2009

L. Wei, et al, IEDM 2009



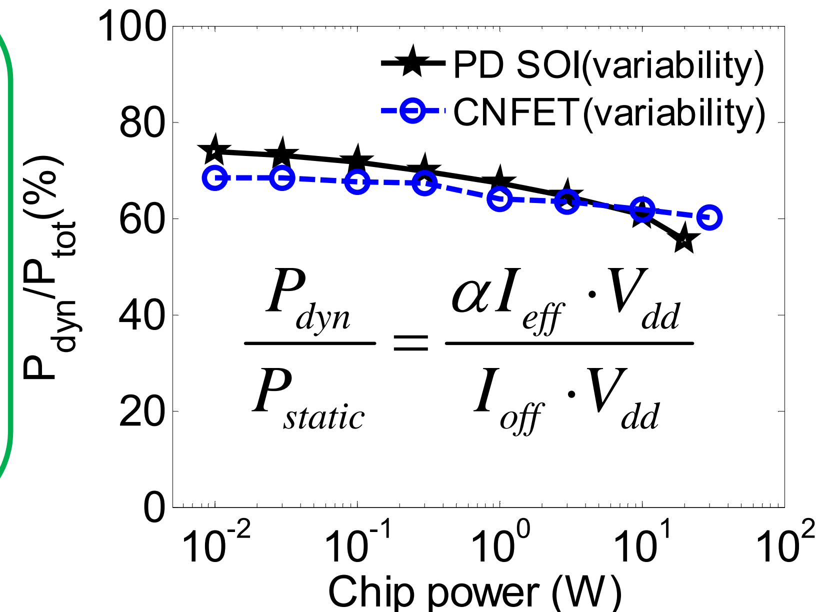
For both high performance and low power applications: CNFET is 5x faster



Device parameters (including L_{gate}) are optimized.

At the optimum performance,

$$\left(\frac{I_{eff}}{I_{off}}\right)_{PDSOI} \approx \left(\frac{I_{eff}}{I_{off}}\right)_{CNFET}$$

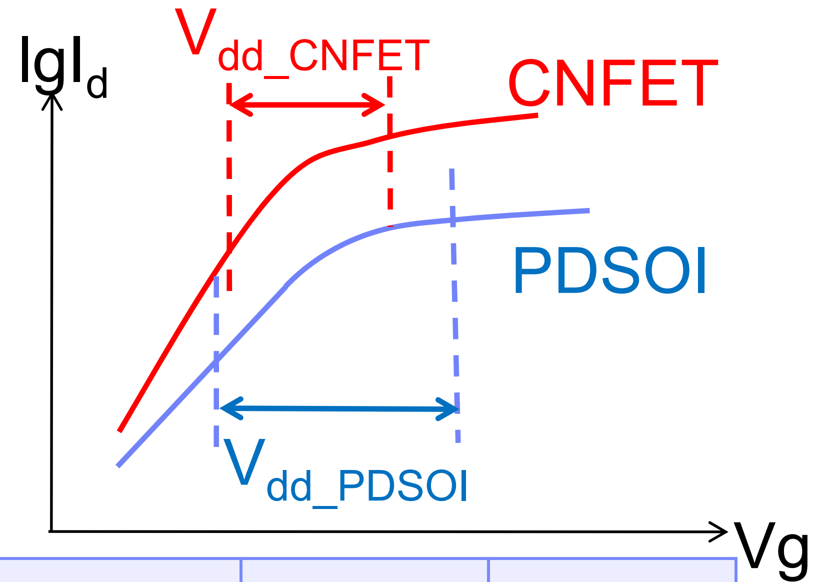


- Conventional way of benchmarking
  - maximizing I_{on} at fixed I_{off} and V_{dd} (not always optimized)
- Optimized design
  - Design criteria depends on application constraints
  - Targeting at the system-level performance

Markovic, et al JSSC 2004 L. Wei et al, IEDM 2009

At the optimum performance,

$$\left(\frac{I_{eff}}{I_{off}}\right)_{PDSOI} \approx \left(\frac{I_{eff}}{I_{off}}\right)_{CNFET}$$



Chip power=10W, incl. variability	L_{gate}	V_{dd}	I_{eff}	I_{eff}/I_{off}	C_{load_ave}
	nm	V	mA/µm		fF
PDSOI (optimal)	21nm	0.7V	0.234	456	1.05
CNFET (optimal)	10nm	0.4V	0.62	334	0.68

Good short channel performance

Ballistic transport

Shorter/narrower devices and shorter wires

Take-home message #3

System-level optimization is essential for device design and benchmarking.

- Design criteria depends on application constraints
- Targets the system-level performance

Conclusions

- **A holistic view is required for device design and engineering**
  – Circuit-level performance must be considered
  – Device design is application-orientated
  – Compact modeling facilitates high-speed circuit and system simulation

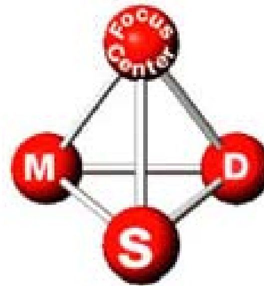
Conclusions

- **A holistic view is required for device design and engineering**
  - Circuit-level performance must be considered
  - Device design is application-orientated
  - Compact modeling facilitates high-speed circuit and system simulation

Device design is not a stand-alone task!

Acknowledgement

- **Sponsors and Collaborators**
  - Dr. David. J. Frank, Dr. Leland Chang, Dr. Jie Deng (IBM)
  - Dr. Frédéric Boeuf, Dr. Thomas Skotnicki (STMicroelectronics)
  - Prof. Dimitri Antoniadis (MIT)
  - Dr. Kwok Ng (SRC)



Stanford INMP

(Toshiba, Intel, TI, IBM, AMD, TEL, AMAT, COSAR)

Conclusions

- **A holistic view is required for device design and engineering**
  - Circuit-level performance must be considered
  - Device design is application-orientated
  - Compact modeling facilitates high-speed circuit and system simulation

References:
L. Wei et al, IEDM07 & 09 (Stanford/IBM)
L. Wei et al, VLSI-TSA 09 (Stanford/STMicroelectronics)
L. Wei et at, TED 09 (Stanford/IBM)
L. Wei et al, SSDM 09 (Stanford/STMicroelectronics/MIT)