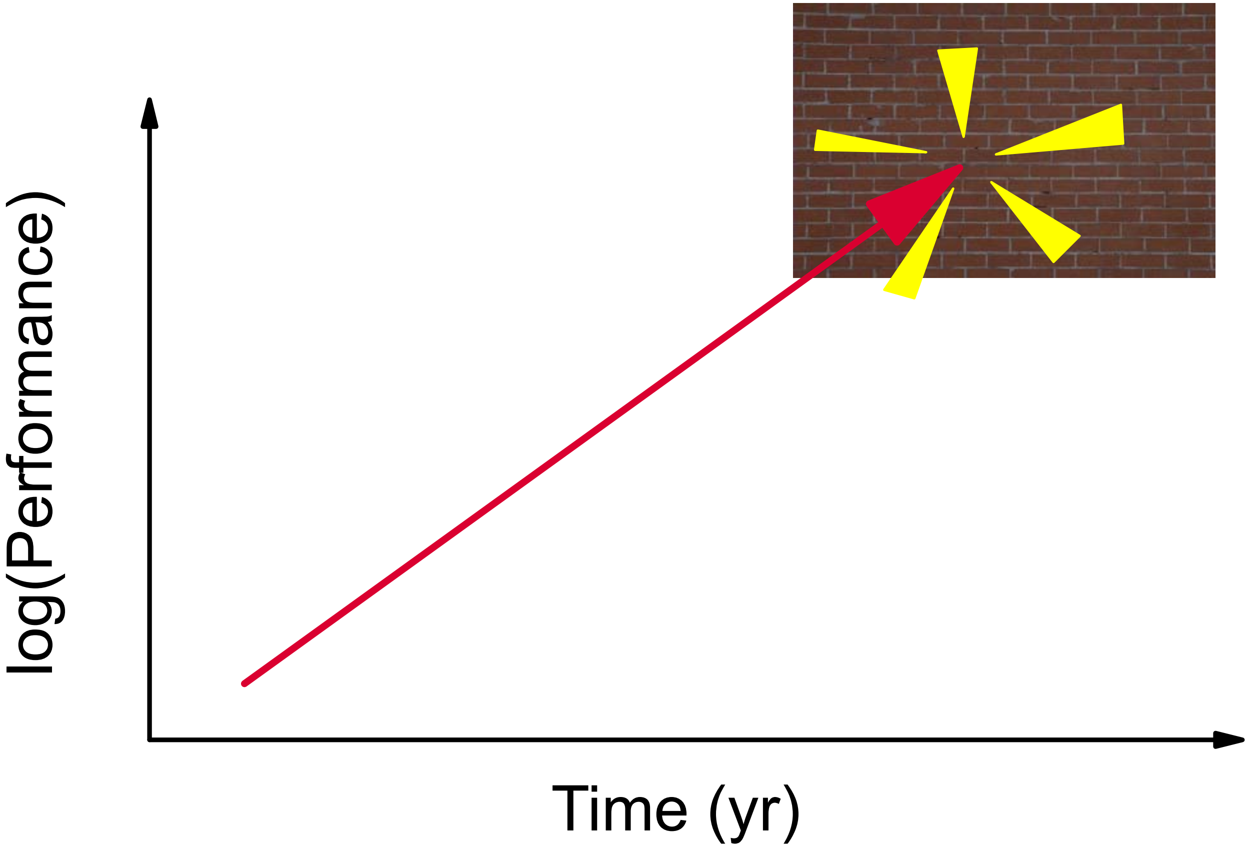
Technology Optimization and Performance Projections

David J. Frank

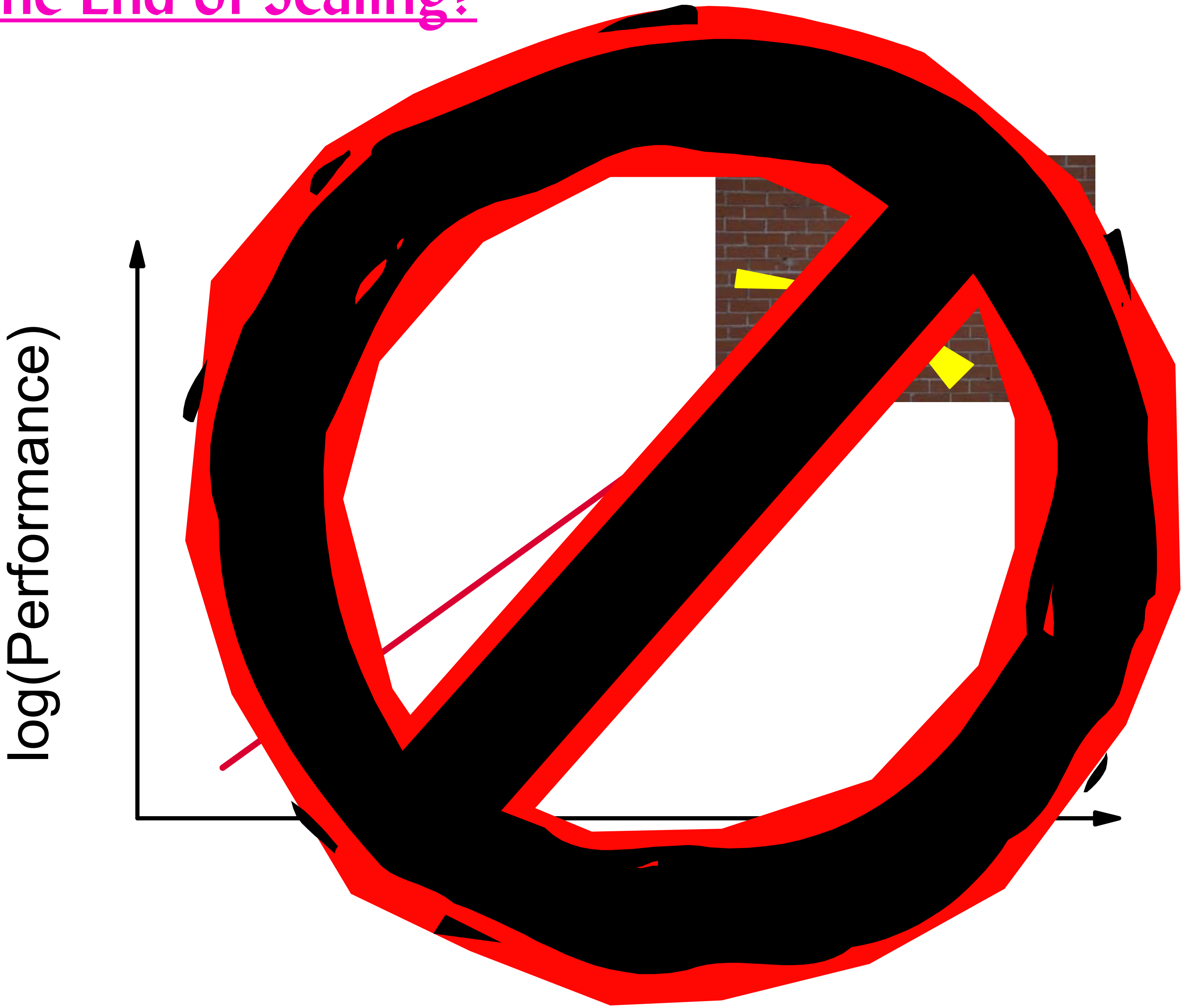
12/4/09
International Winter School for Graduate Students
IIT, Bombay, India

1. Introduction

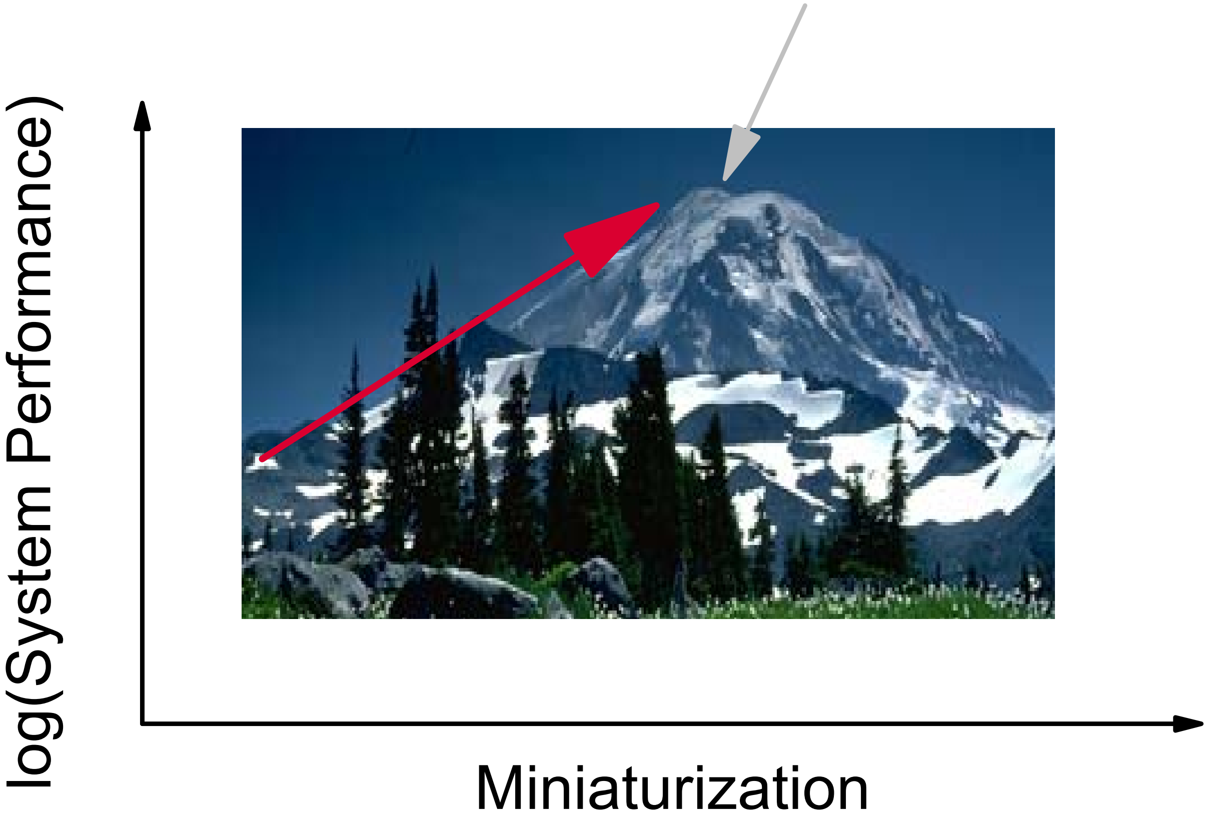
The End of Scaling?



The End of Scaling?



The End of Scaling is Optimization



Stop when you get to the top.

Outline

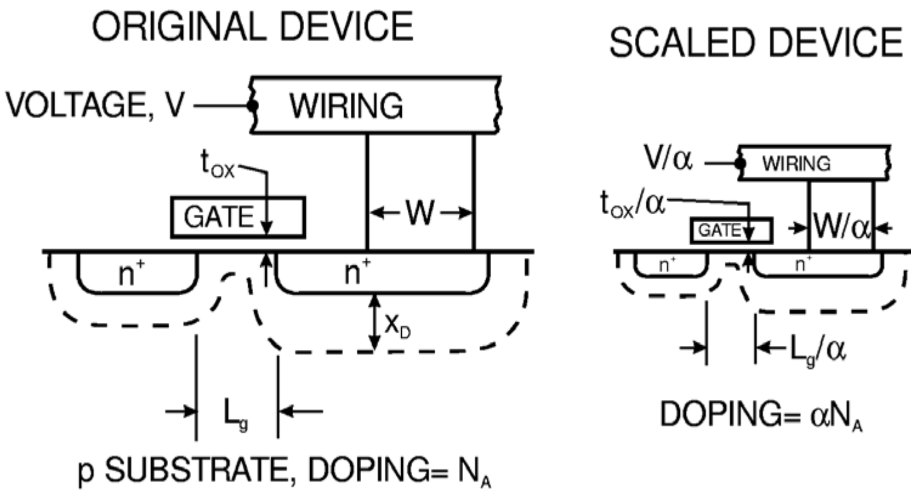
1. Review limitations to Scaling
2. Optimizing technology within constraints
3. Optimization results
4. Technology projections

1. Limitations to Scaling

1. Electrostatic constraints
2. Quantum mechanical leakage currents
3. Discreteness of matter and energy
4. Thermodynamic limitations
5. Practical and environmental constraints on power

Basic idea of Scaling:

Adjust dimensions, voltages, & doping to achieve smaller FET with same electrostatic behavior.



Electrostatic constraints on FET design

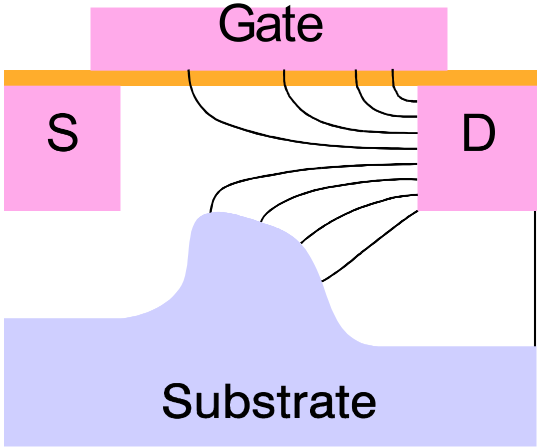
1. A good design should have
    - A. High output resistance
    - B. High gain
    - C. Low sensitivity to variations
    - D. High transconductance
    - E. High drain current
    - F. High speed

} Long channel behavior (A–C)

} Short channel behavior (D–F)

2. Must choose a compromise:

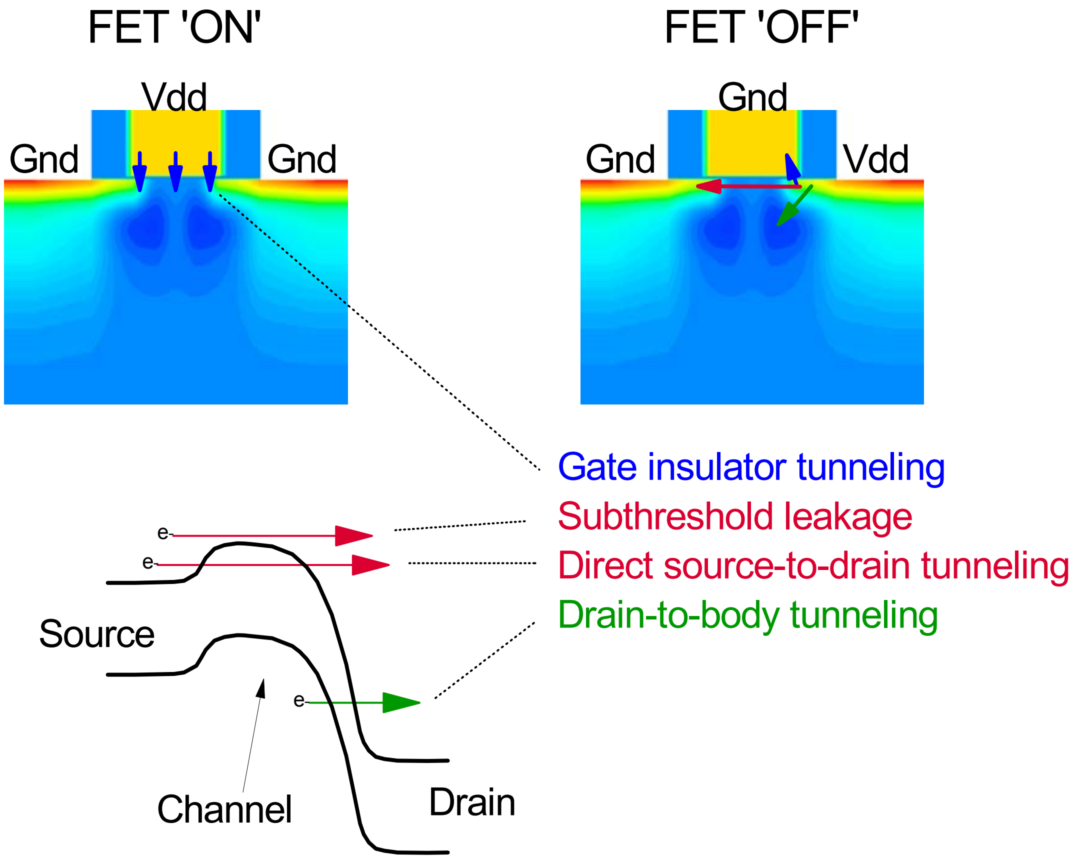
Short, but not so short that 2D effects kill A, B, C.



Quantum Mechanical Tunneling Leakage Currents

Currents increase exponentially as the barriers become thinner.

Everything becomes leaky.



Gate insulator tunneling

Subthreshold leakage

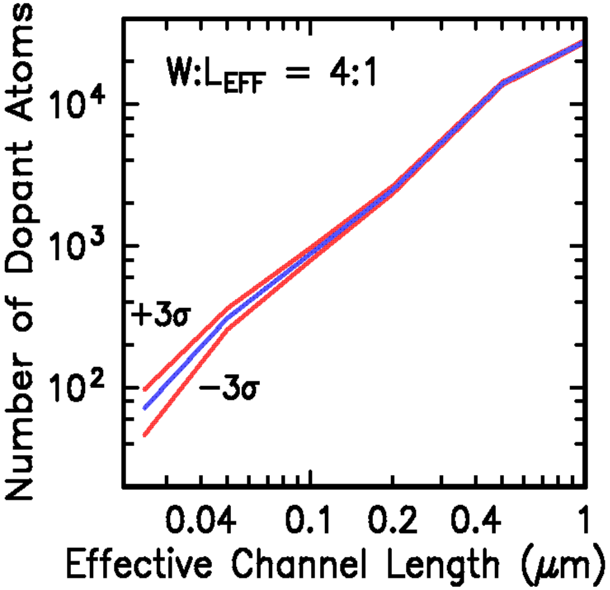
Direct source-to-drain tunneling

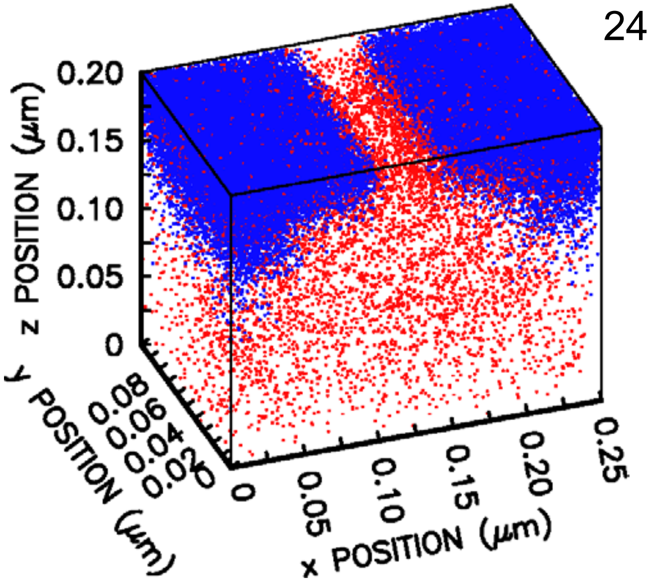
Drain-to-body tunneling

Atomistic effects

The number of dopant atoms in the depletion layer of a MOSFET has been scaling roughly as Leff$^{1.5}$.

Statistical variation in the number of dopants, N, varies as $N^{1/2}$, causing increasing V_T uncertainty for small N.





249,403,263 Si atoms
68,743 donors
13,042 acceptors

[D. J. Frank, et al., 1999 Symp. VLSI Tech.]

Thermodynamic limitations

- The Boltzmann distribution determines the subthreshold slope and leakage current, V_T, and diode leakage currents, too.

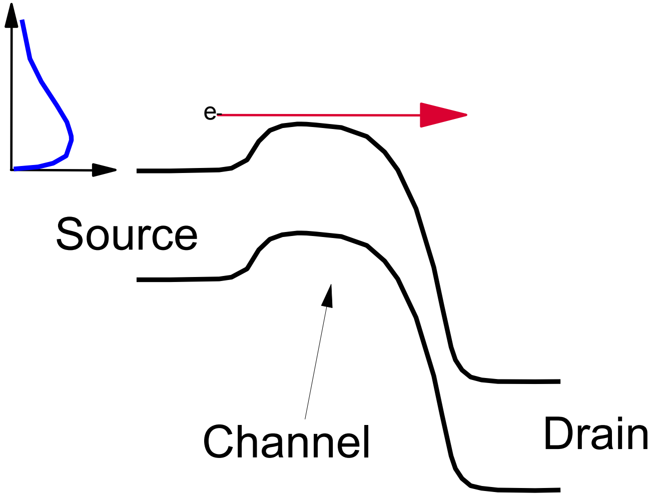
$$I_{subVT} = I_0\, e^{e(V_G - V_T)/\eta kT}$$

- V_T can only be scaled by reducing the temperature, which is not acceptable for many applications.

- Speed is very sensitive to V_T/V_{DD} ratio.

- Irreversible computation => All switching energy is converted to heat.

- All leakage currents and IR drops are irreversible => More heat.



Practical and Environmental Issues

- **Power consumption and heat removal are limited by practical considerations.**
- **Low power applications are often battery powered**
  - Many must be lightweight => power < ~few watts.
  - Disposable batteries can cost >> $500/watt over life of device.
  - Rechargeables can cost > $50/watt over life of device.
- **Home electronics is limited to <~1000W by heating of the room and cost of electricity.**
- **High performance is limited by difficulty of heat removal from chip (~100 W/chip). (Cost of electricity is ~$5/watt over life.)**

2. Optimizing Technology within Constraints

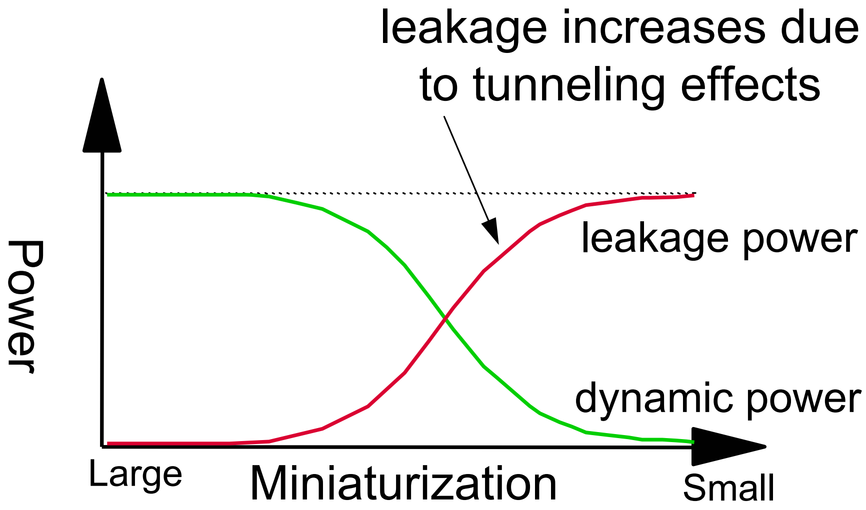
Fixed architectural complexity

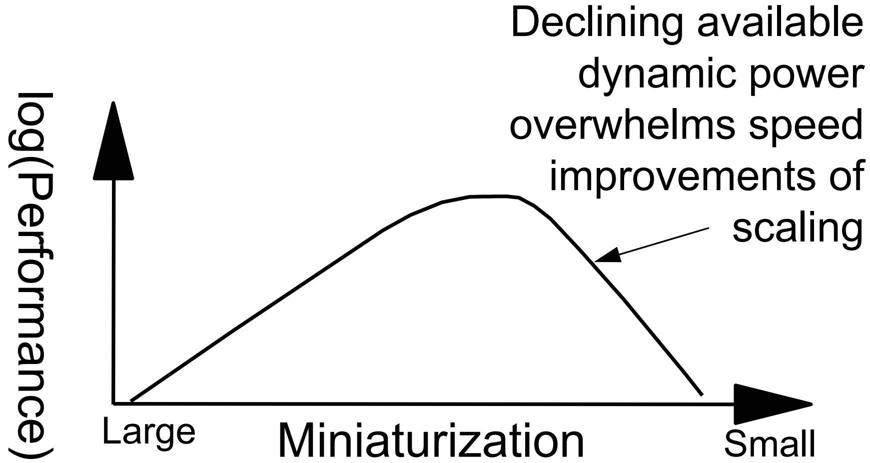
+ Fixed power constraints

+ Device physics

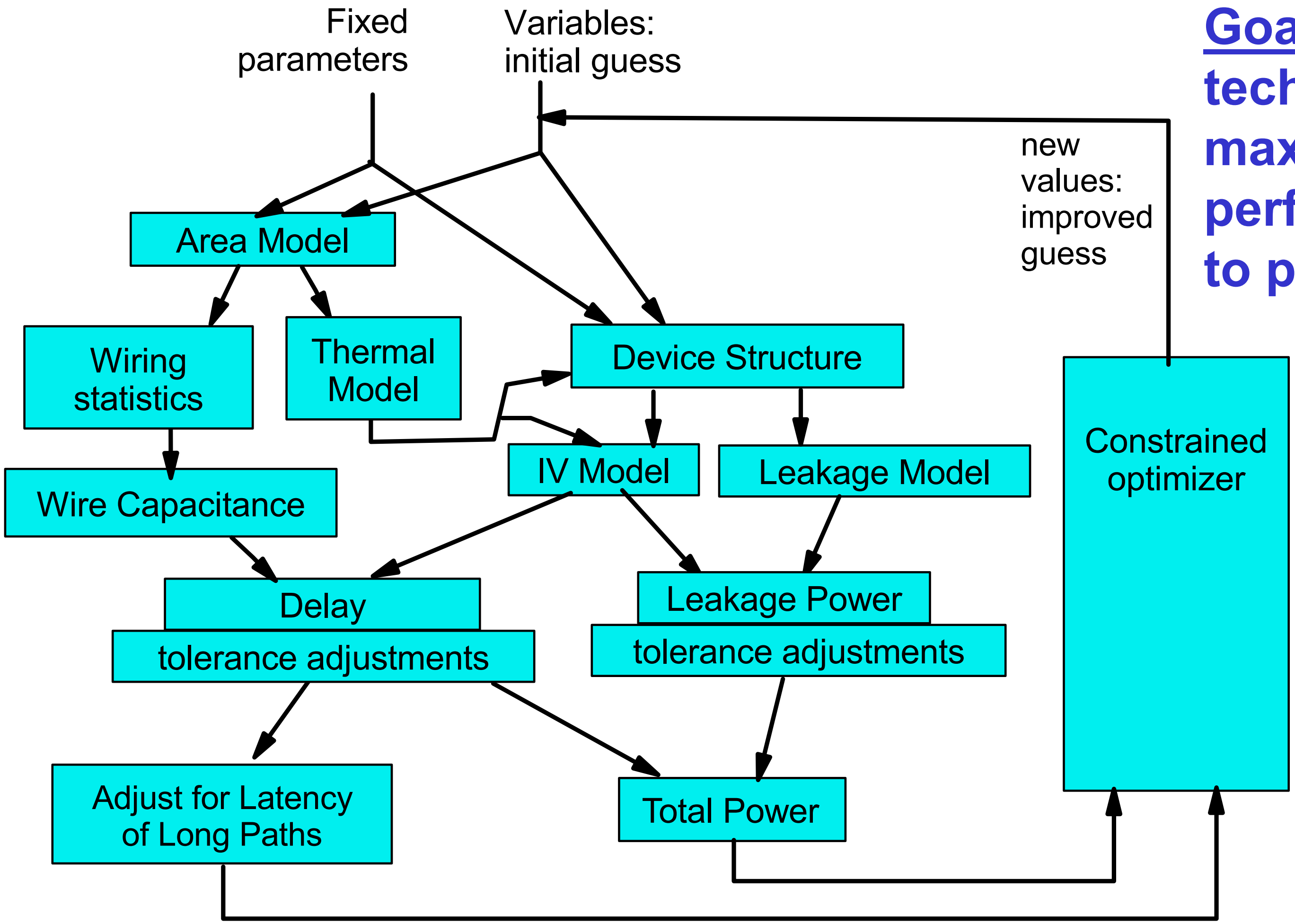
= Existence of an optimal technology with maximal performance.

- Practicality imposes power constraints.
- Electrostatics imposes geometric constraints
- Thermodynamics imposes voltage constraints.
- Quantum mechanics imposes miniaturization constraints due to tunneling.





Background: Schematic organization of an optimization program

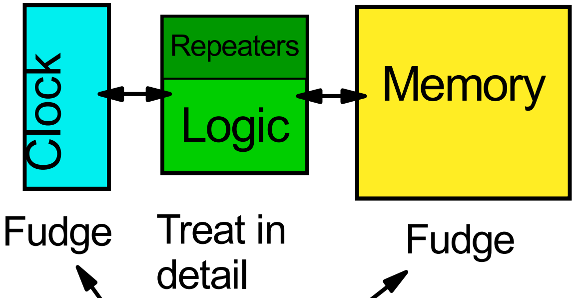


Goal: optimize device technology to maximize chip-level performance, subject to power constraints.

Models and Approximations

System Assumptions

- Processor chip is assumed to have a fixed number of cores, each with a specified number of logic gates.
- Only the logic within the cores is considered within the optimizations.
- The clock and memory aspects of the chip are assumed to scale in the same way as the logic (delay, power, and area).
- Core-to-core and core-to-memory communication is not dealt with.

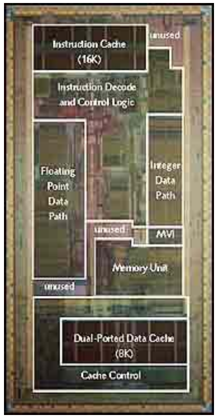


Treat these by simple scaling from the logic part.

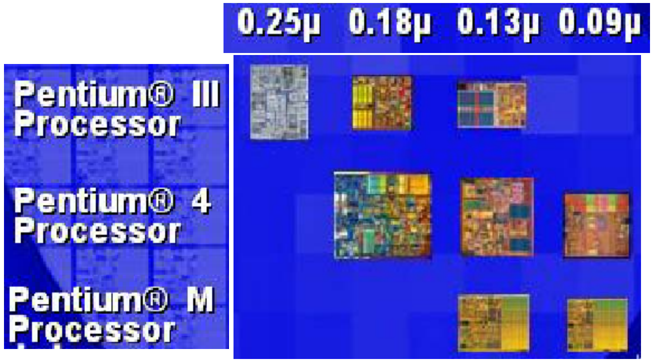
How much area do the processor cores take?

100% to 25%, generally decreasing with generation:

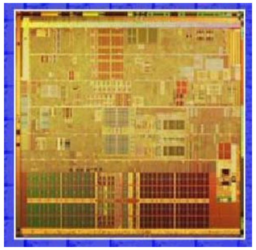
100%



Alpha 21264 ('96)
15M FETs, L1 cache only

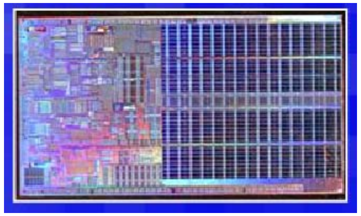


70%

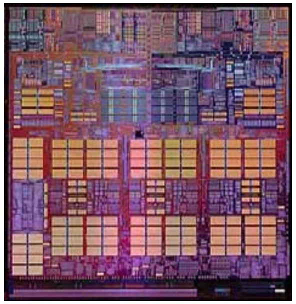


Prescott. 125M FETs

40%

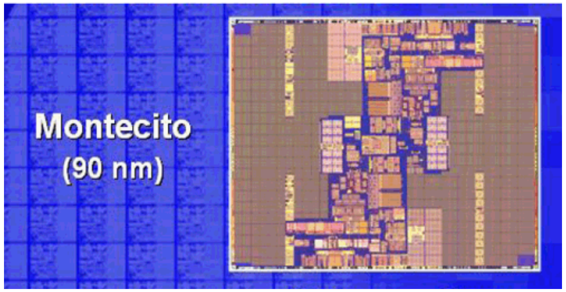


Dothan, 140M FETs



40%

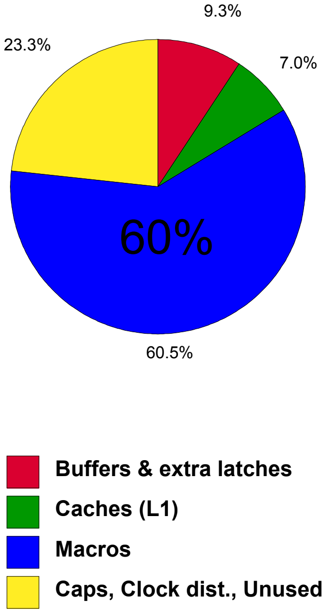
Power4, 174M FETs

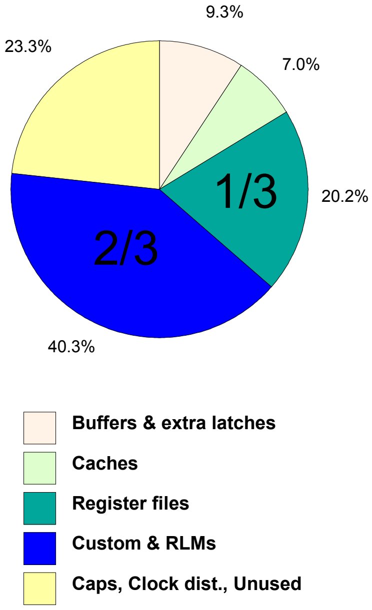


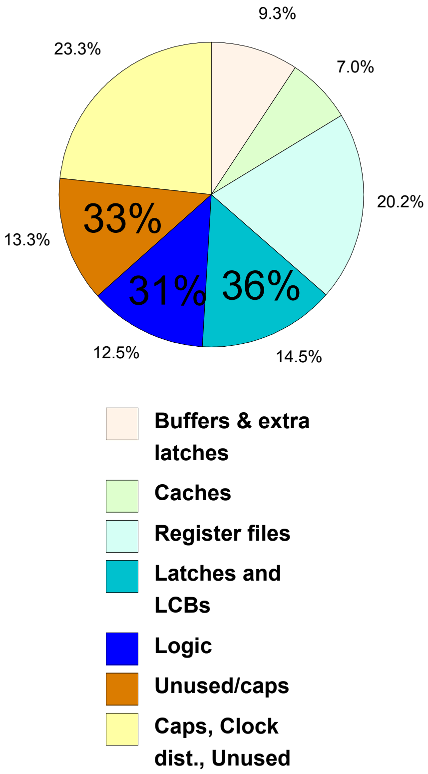
~25% 2 cores, 1.72B FETs

Area usage within a processor core

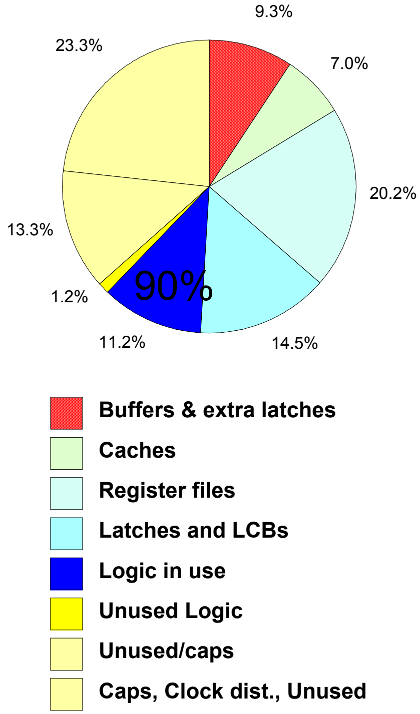
Approximate area fractions for a high-performance microprocessor core in leading-edge technology







data from:
M. Scheuermann and M. Wisniewski



- ► Processors built with nanotechnology are likely to have similar area usage statistics.
- ► Nanotechnology may require additional area allocations for defective circuitry.
- ► Estimates of power and computational densities should take into account realistic area efficiencies.

Optimization Approaches

1. Engineering approach:

Maximize system performance, at fixed power.

Use total logic transition rate (LTR),

LTR = N_{gates} x activity factor/logic depth x 1/Delay

Relatively little dependence on architectural details.

✗ 2. Business approach:

Maximize Return on Investment (ROI).

$$ROI = \frac{LTR \cdot t_{Life}}{Area \cdot C_A + Power \cdot C_P} = \frac{t_{Life}}{C_A} \cdot \frac{LTR}{Area(1 + P/P_{econ})}$$

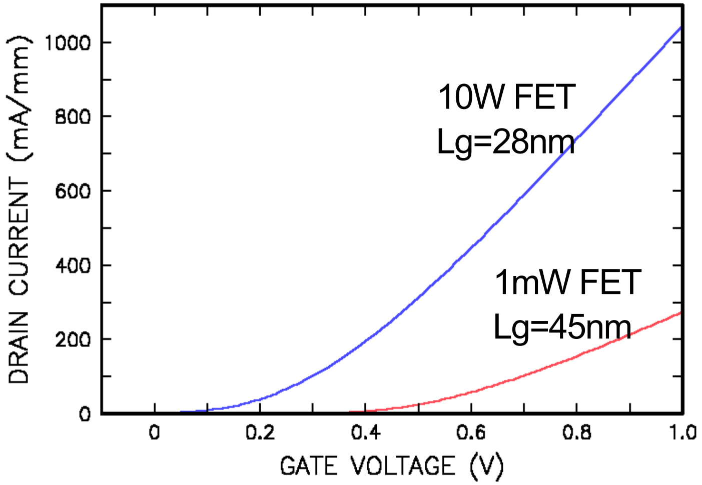
FET Model

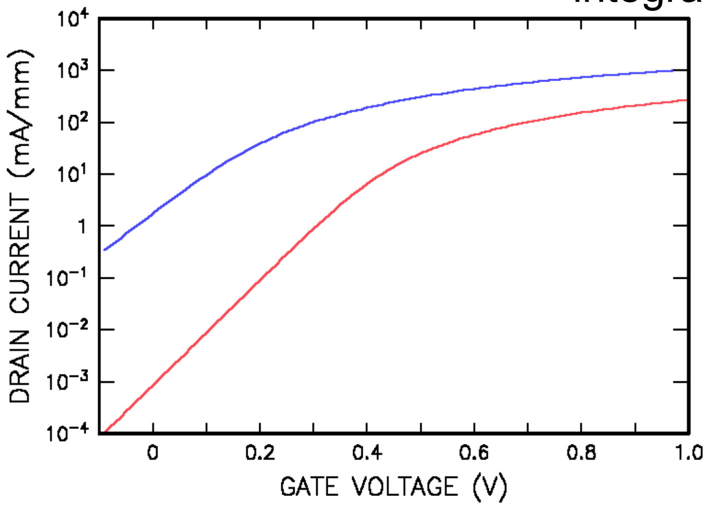
Using a general temperature-dependent short-channel FET model in which V_T, t_D, and t_{ox} are coupled, halo doping effects are included, and V_T is set by the doping.

Modified alpha power model:

$$I_D(V_{GS}) = \frac{W\varepsilon_I}{t_{ox}^{eff}} \frac{\eta kT}{e} \left(\frac{\eta kT/e}{FI\, E_C L_{CH}} \right)^{\gamma} \mu_0 \left(\frac{\mu(E_\perp)}{\mu_0} \right)^{s} E_C\, \mathrm{F}_\alpha \left(\frac{V_{GS}-V_T}{\eta kT/e} \right)$$

Fermi-Dirac integral of order α





Circuit Delay Estimation

Basic circuit elements are:

FI=2, FO=1.65 wire-loaded NAND gates for logic

inverters for repeaters, FO ~ 1.2

Delay calculations:

$$\tau_1 = \frac{V_{DD}\,(C_{parasitic} + C_{wire} + C_{gateload})}{2I^{*}_{Deff}}$$

Current is adjusted to account for noise and variations.

$$\tau_2 = R_{wire}(C_{wire} + C_{gateload})$$

$$\tau_3 = L_{wire}/(c/2)$$

Propagation delay

$$\tau = \frac{\tau_1 + \left(\tau_2^{4/3} + \tau_3^{4/3}\right)^{3/4}}{0.5 + \left(1 - V_T/V_{DD}\right)/(1+\alpha)}$$

Final delay empirically merges the separate components. [Eble's thesis]

Correction for V_T/V_{DD}.

Power Calculation

$$P_{TOT} = P_{DYN} + P_{subVT} + P_{OX} + P_{B2B}$$

$$P_{DYN} = \frac{\alpha}{\ell_D} N_{CKT} \frac{1}{2} \langle C \rangle (V_H - V_L) V_{DD} / \tau$$

$$P_{subVT} = 1.7 N_{CKT} V_{DD} I_{off} (V_T, V_{DD}, t_{ox}, \eta, L_G, \tfrac{W}{L})$$

$$P_{ox} = A_{core} D_{ox} (\tfrac{W}{L}) V_{DD} J_{ox} (V_T, V_{DD}, t_{ox}, \eta)$$

$$P_{B2B} = \frac{1}{3} A_{core} D_{ox} (\tfrac{W}{L}) V_{DD} J_{B2B} (F_{Max}, V_{DD})$$

$$LTR = \frac{\alpha}{\ell_D} N_{CKT} \frac{1}{\tau}$$

Note that cross-through power is not included.

The powers are computed separately for logic and for repeaters.

τ = mean delay for a single loaded logic gate

$\frac{\alpha}{\ell_D}$ is activity factor divided by logic depth. Usually ~0.012 in recent optimizations.

Communication and Wiring Models

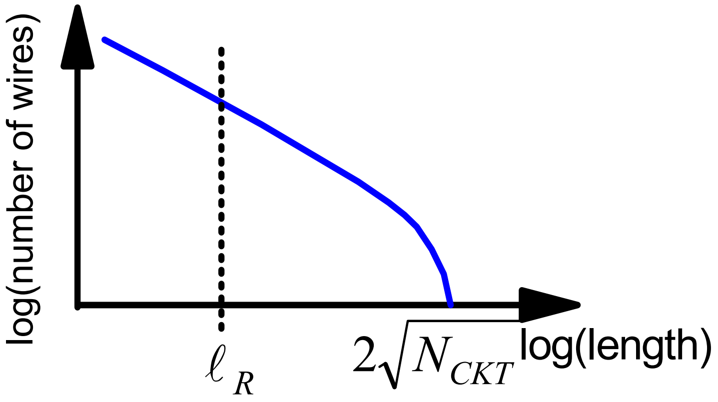
Assume wire lengths distributed according to Rent's rule.

$$i_{net}(\ell) = \frac{4FO}{3+FO}\left(\ell^{2r-3} - \left(2\sqrt{N_{CKT}}\right)^{2r-3}\right)$$

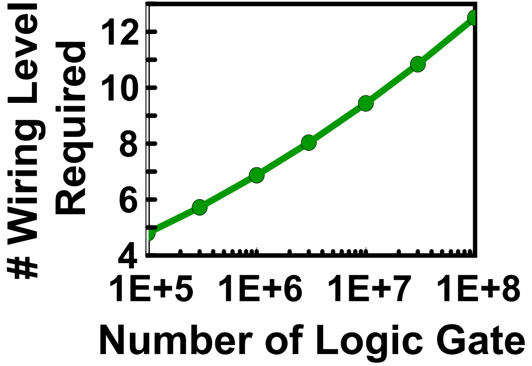
$$\langle L_{noRptr} \rangle = \frac{\int_{1}^{\ell_R} \ell i_{net}(\ell)\,d\ell}{\int_{1}^{\ell_R} i_{net}(\ell)\,d\ell}$$

$$N_{Rptr} = \int_{\ell_R}^{\ell_{Max}} \ell i_{net}(\ell)\,d\ell \Big/ \ell_R$$

Units are gate pitches.
r = Rent exponent, 0.6, here.



From optimizations:



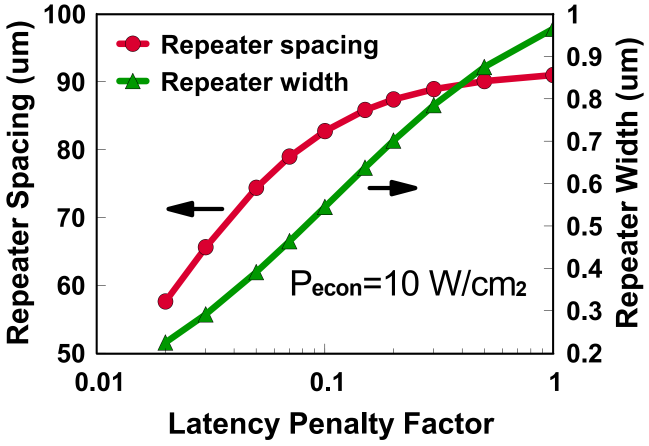
Repeater Model

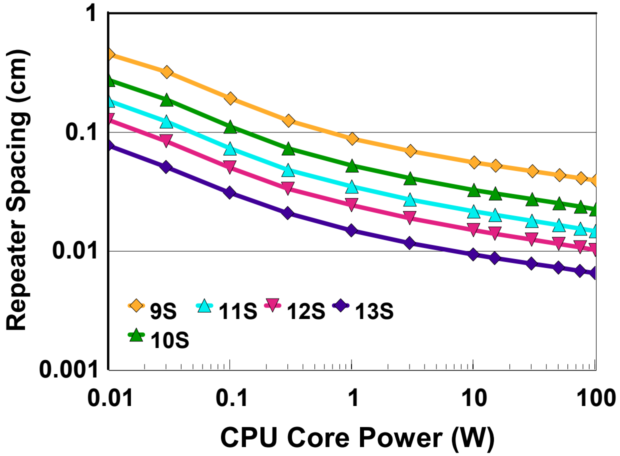
Long wires receive repeaters with a spacing that is optimized.

Long wire delay can be absorbed into pipeline depth, but the latency causes inefficiency, so we use a latency penalty factor: γ.

$$CPI_{eff} = (1-\gamma) + \gamma\tau_{instr}/\tau_{cycle}$$

$$LTR_{eff} = LTR/CPI_{eff}$$





Local Variation Modeling

- Variation sources:
  - **Signal Coupling noise**
  - **Supply noise**
  - **Statistical doping variations**
  - **LER gate length variations**
- Consequences modeled:
  - **Increased static power**
    - **combine 1 sigma of doping, length, and noise**
  - **Critical path delay distribution**
    - **yield-based, using estimated critical path distribution,**
    - **and 1 sigma of doping and length, and worst case noise.**
  - **Single stage functionality**
    - **use worst case (~6 sigma) of doping and length, no noise.**

Accounting for variations

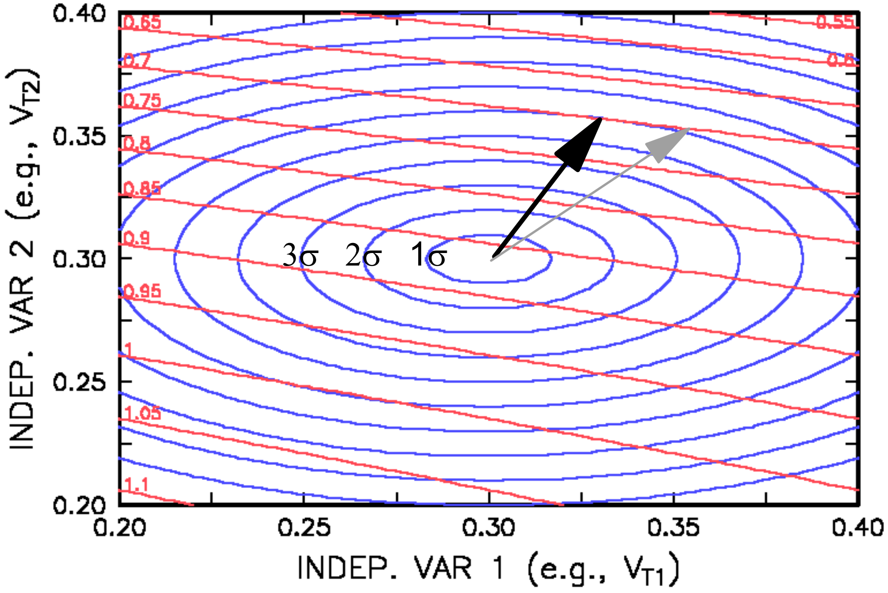
- A complete most-probable worst-case-vector methodology is used to handle both local and global variability.

Worst-case vectors:

Blue curves are contours of constant probability.

Red curves are contours of function to be minimized.

Two variable example:



MPWC vector

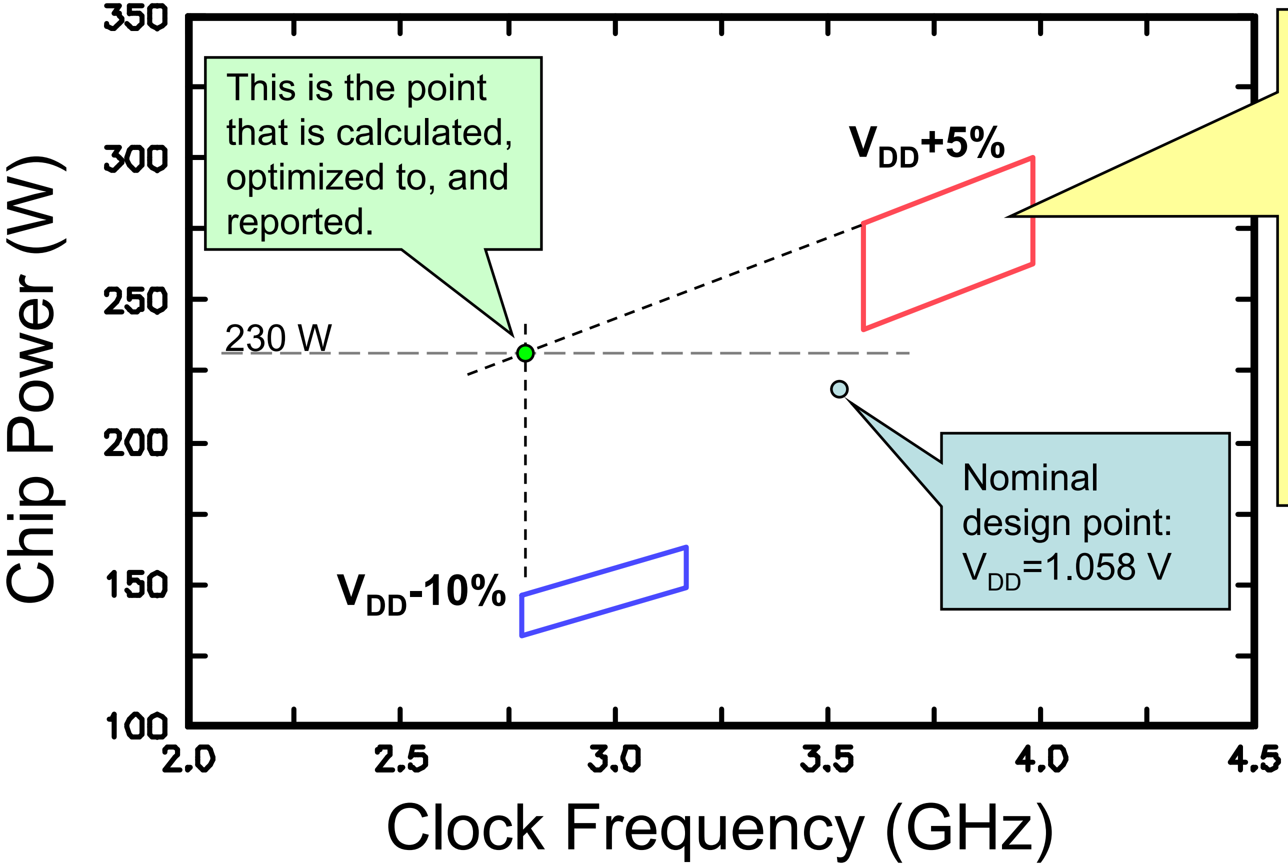
Murphy vector

Power vs Frequency Variation Windows

Optimizations for high-perf processor, 45 nm node technology, PDSOI

Power-constrained optimization parameters: V_{DD}, V_{Tn}, V_{Tp}, L_G, repeater size and spacing.

Area constrained is also constrained, to 5.6 cm^2, by adjusting the widths.



This is the point that is calculated, optimized to, and reported.

Nominal design point: V_{DD}=1.058 V

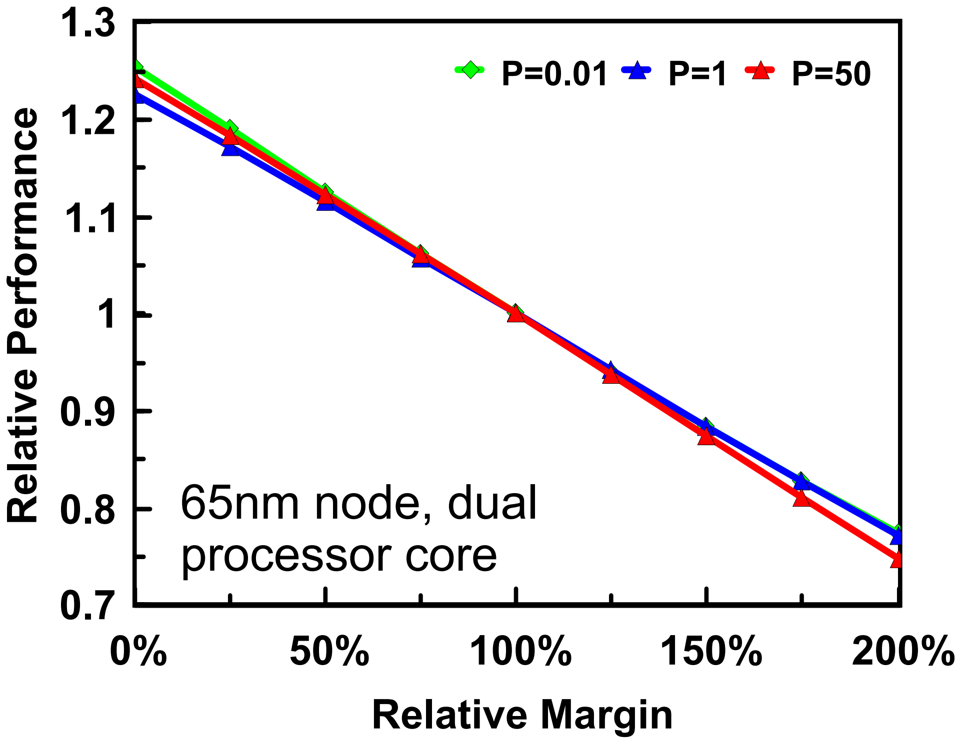
These boxes are for ±0.675 sigma, for 50% yield.

Power and delay are treated as independent: 25% of yield is lost for each.

The data shows 10% frequency variation per sigma.

Impact of variability on performance

- Atomistic effects are leading to greater device variability.
- Increasing variability requires larger design margins.
- Designing for larger margins decreases performance.



Increased variability requires:

- ☐ Higher supply voltages
- ☐ Less scaled FETs

Summary: Models, Assumptions and Approximations

- Power modeling
  - Dynamic switching energy plus static power mechanisms including sub-threshold current, gate oxide tunneling, and body-to-drain band-to-band tunneling.
- Device modeling
  - Bulk MOSFETs: V_T, and depletion depth determined by the halo doping, 2D effects are taken into account.
  - Gate length is fully optimized, not set by the technology node.
- Circuit modeling
  - Delay is for FI=FO=2 or 3 NAND gates, based on model from J.C. Eble's thesis [Ga.Tech. '98].
  - Capacitance includes gate, parasitic, and wire parts (Rent's rule).
  - Wire resistance includes temperature dependence and surface scattering in small wires.

Summary: Models, Assumptions and Approximations

- Chip-level modeling
  - Allocate fixed fraction of chip power and area to logic, and assume fixed number of logic gates. Logic part is optimized, and the rest is assumed to scale similarly.
  - Assume multiple processor cores are interconnected in a way that does not greatly add to the wiring burden.
  - Long wires are fatter, and receive repeaters with a spacing that is optimized.
  - Long wire delay is accounted for using a latency penalty factor.
  - On-chip tolerance/variability and noise is accounted for.

3. Optimization Results

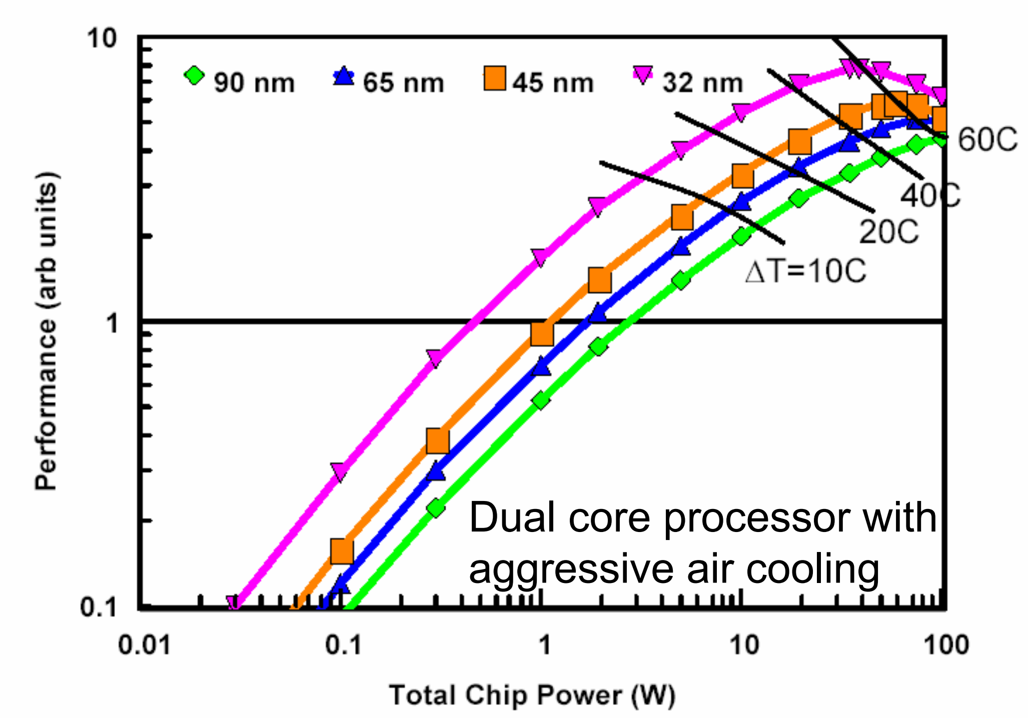
- General results
- Evaluating specific possible device directions
  - Increasing mobility
  - High-k gate dielectric and metal gates
  - 3D stacking
  - Better heat sinks
  - Sub-ambient cooling
  - Multi-processor tradeoffs

Optimize by technology node

For each node, pre-specify the following parameters:

- Wire half-pitch,
- gate overlap,
- halo scalelength,
- contact resistance,
- LER sigma,
- ACLV,
- mobility,
- gate depletion,
- k_wire,
- k_gate

Optimizations over 7 variables:
t_{ox}, L_g, N_D, <w>, V_{dd}, S_{rpt}, <w_{rpt}>

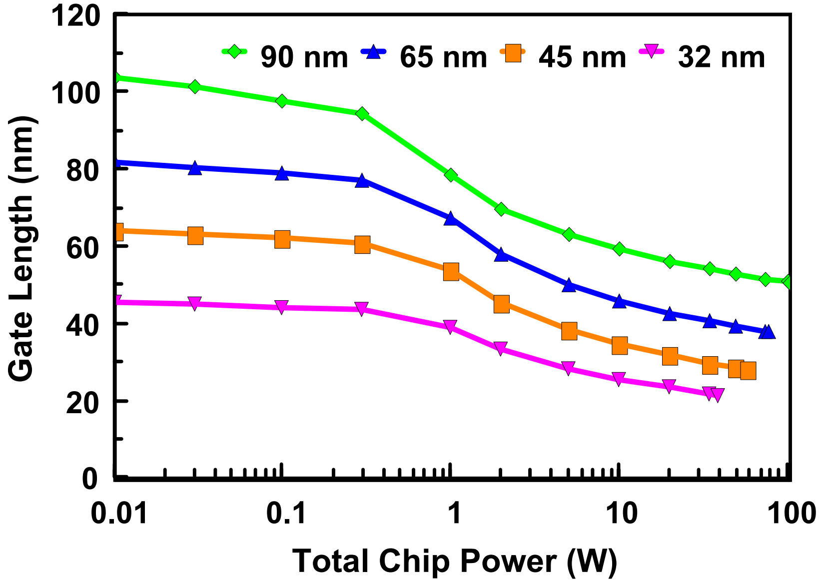


Dual core processor with aggressive air cooling

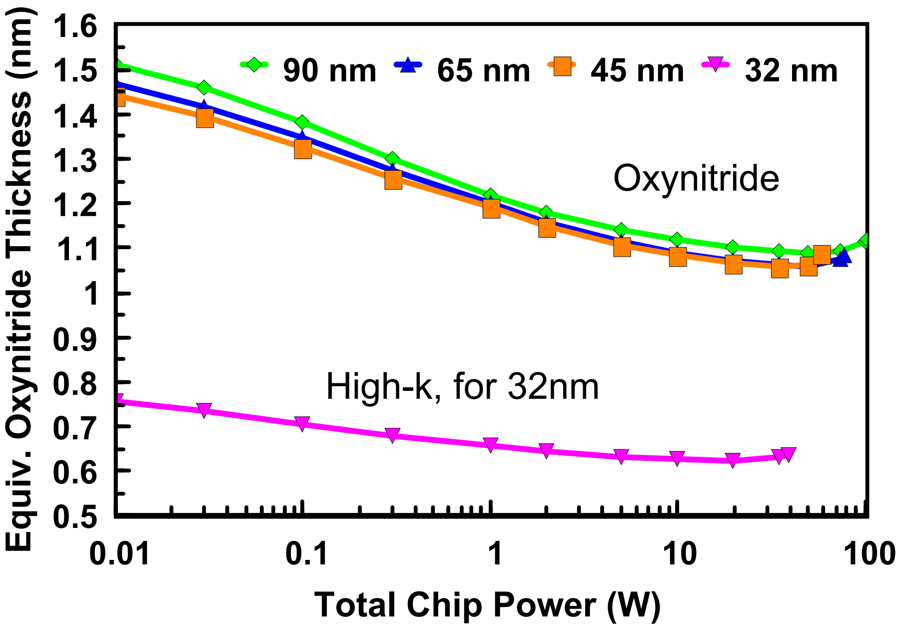
Note that the L_G, t_{ox}, V_{DD}, V_T, width, etc. are **NOT** preselected. They are solved for by the optimizations.

Optimization results

Gate Length vs Power



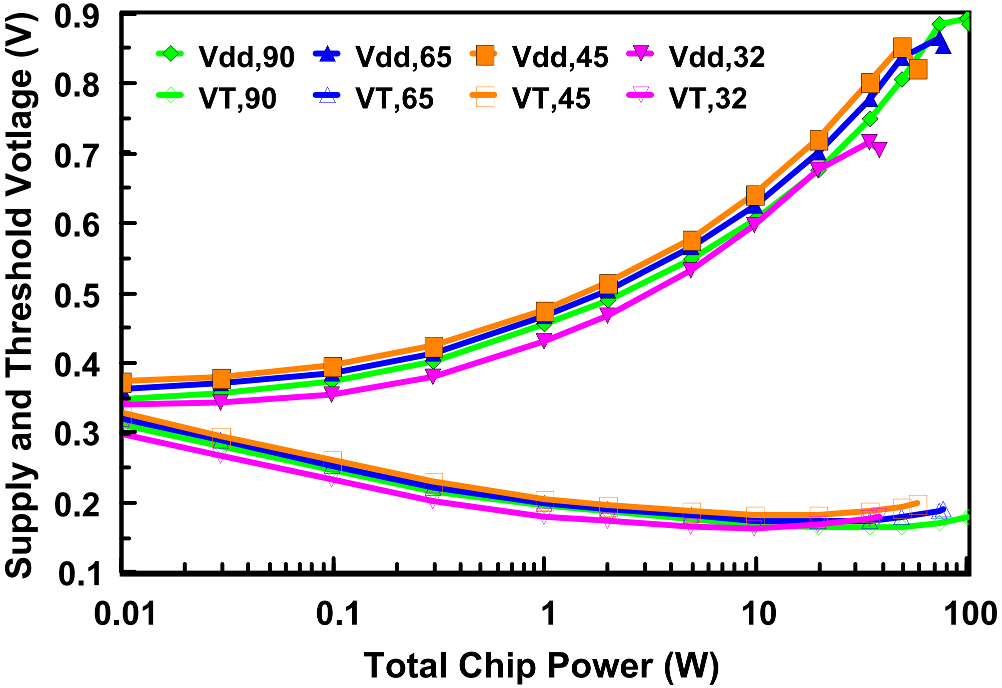
Oxide Thickness vs Power



Dual core processor with aggressive air cooling

(High-k case assumes 0.3nm barrier layer, bandedge metal gate, HfO_2-like insulator characteristics.)

Optimization results

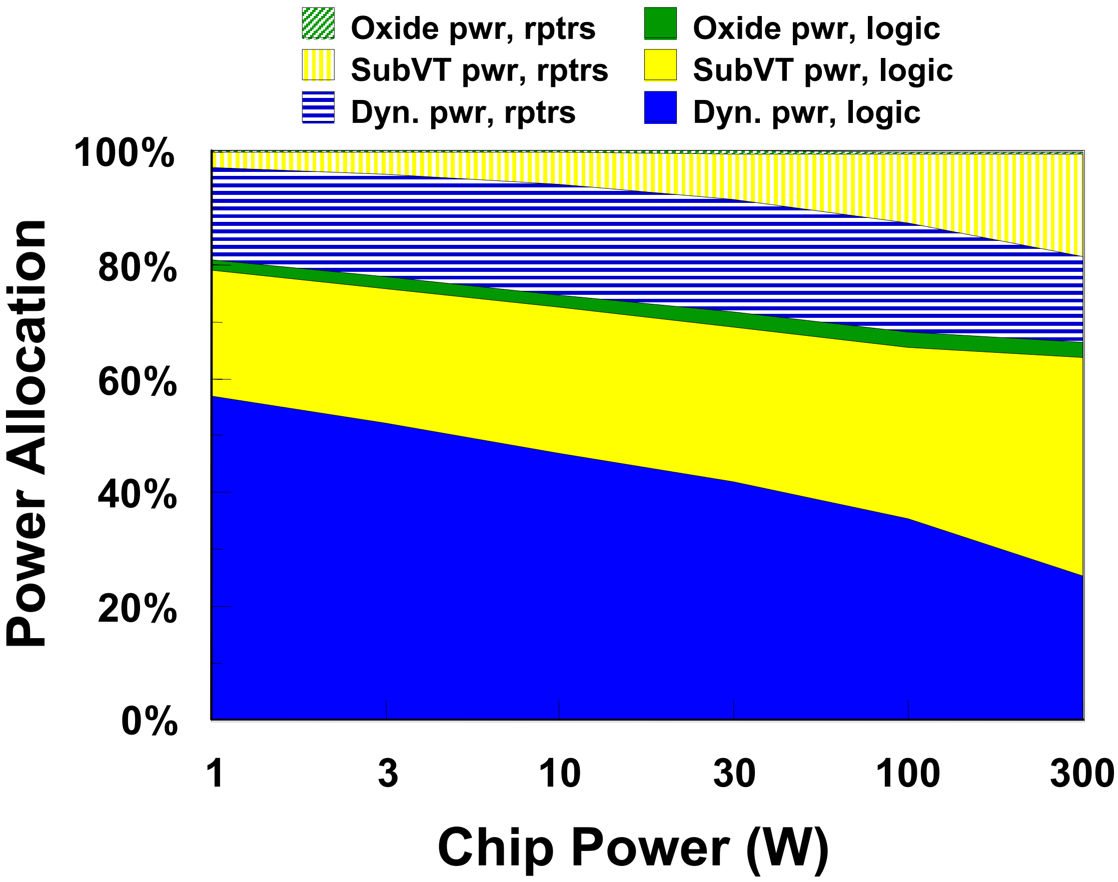


Voltages vs Power

Dual core processor with aggressive air cooling

- Supply voltages are lower for low power applications.
- High-k lowers V_{DD} ~ 15% at the 45nm generation.

Optimal Power Allocation Fractions



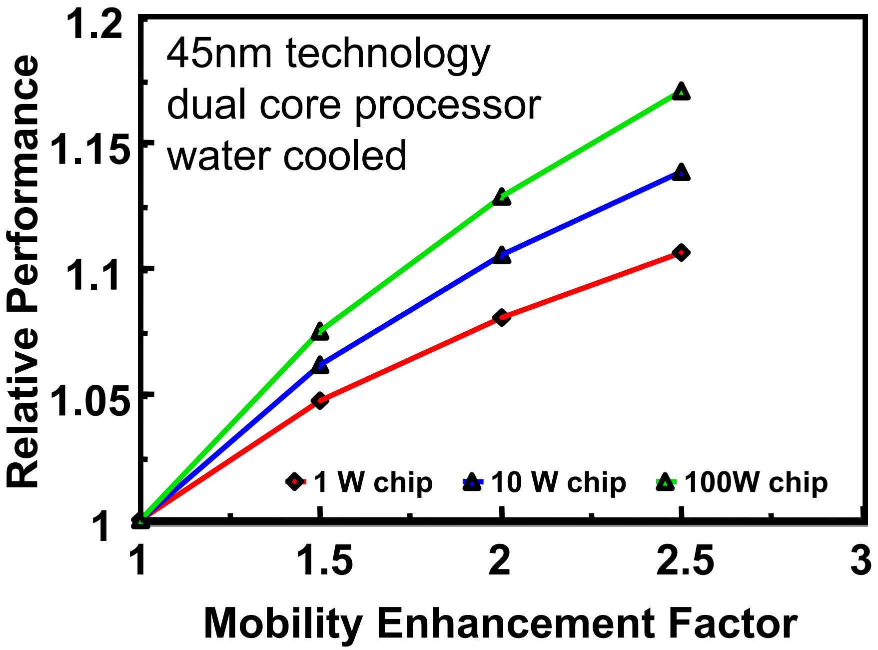
Active power fraction:
70% at low power to
40% at high power.

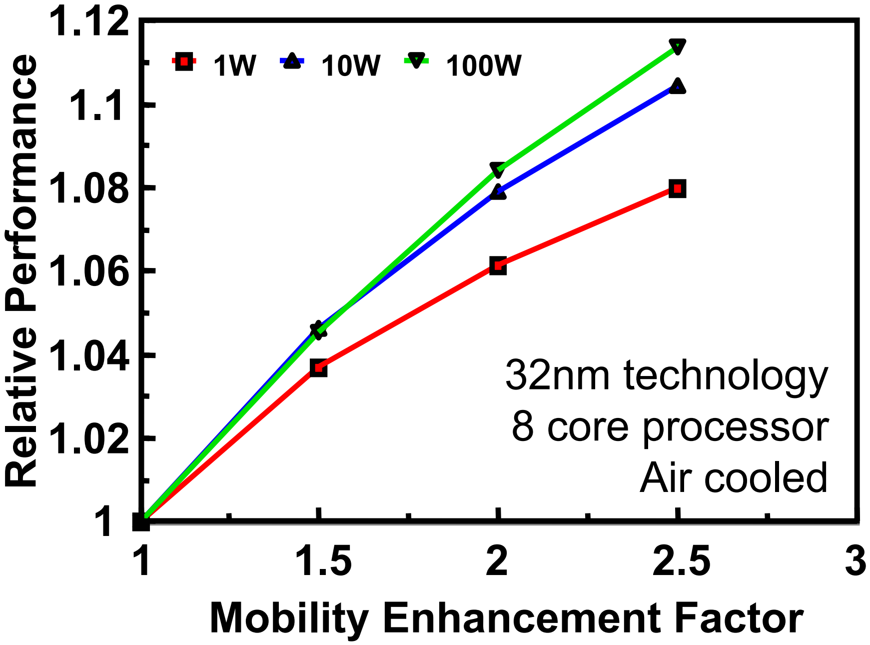
45nm technology with microchannel
heat sink and water cooling.
4 core chip.

Mobility dependence

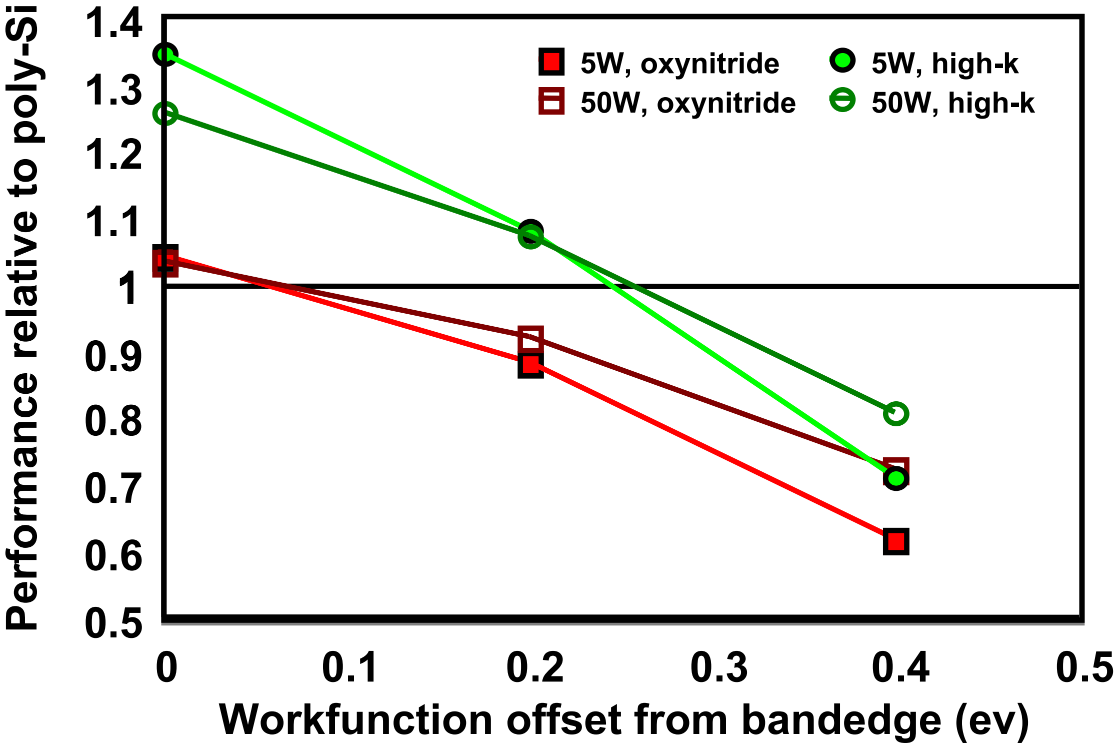
Enhanced mobility has greatest benefit at high power.

Even for large mobility enhancements, performance boost is modest: 10-15%.



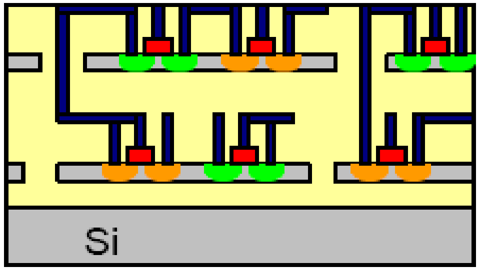


Metal-gate workfunction for high-k and oxynitride



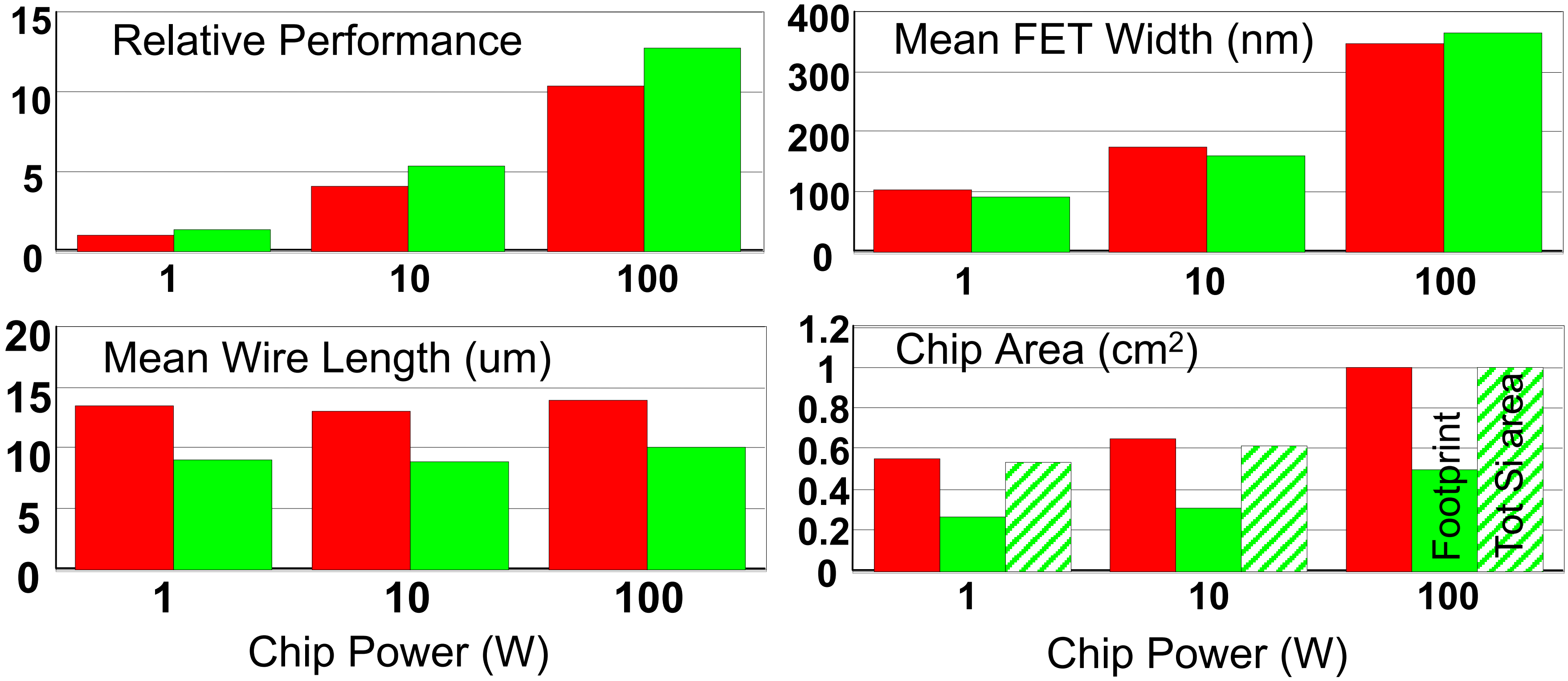
45nm node, dual core processor
with aggressive air cooling

3D stacking



Multiple layers offer higher performance due to shorter wires.

RED = 1 Layer, **GREEN = 2 Layers**

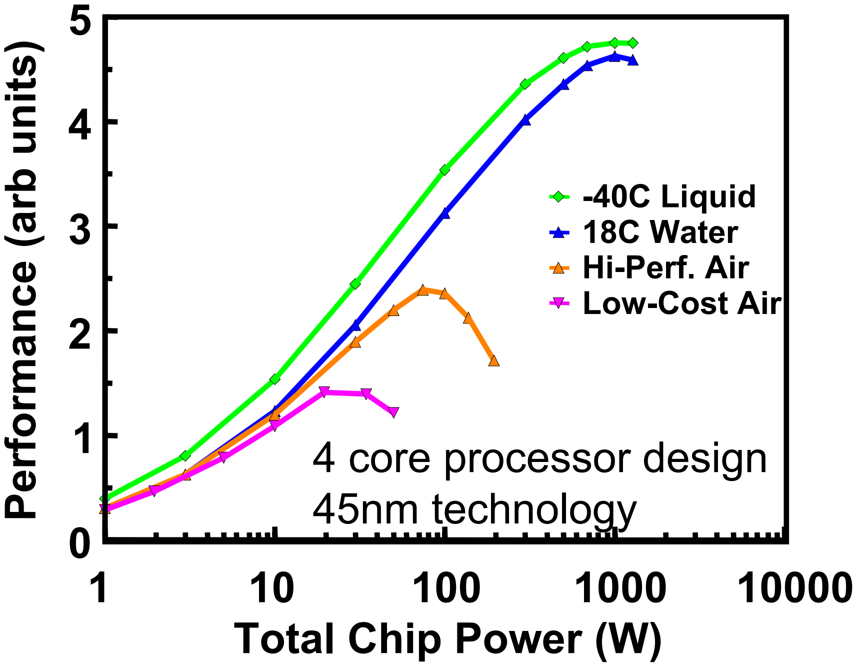


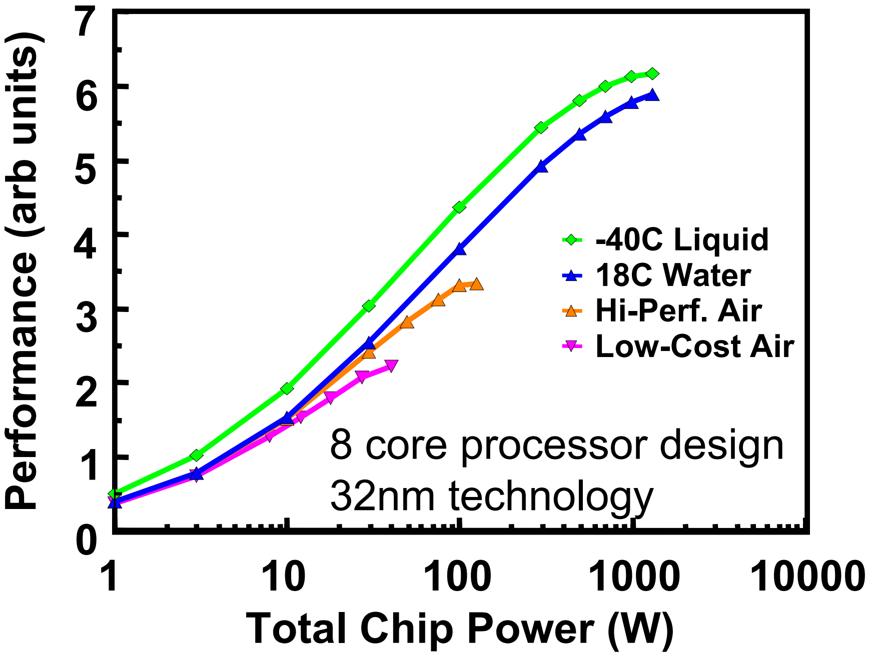
(4 core, 45nm node, water cooling.)

Cooling scenario optimizations

Forced liquid cooling through microchannel fins may permit <u>very</u> high power densities.

Optimized (maximum) performance increases as the ~log of the power.



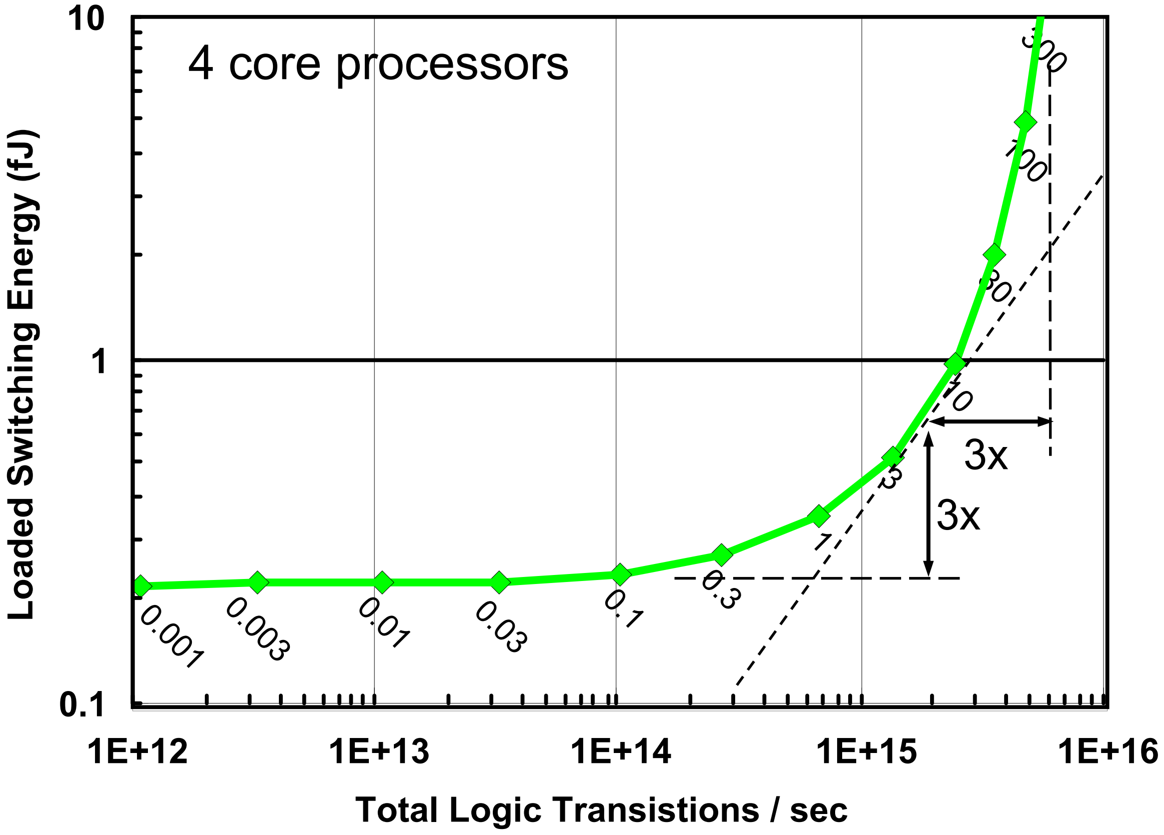


Optimized over 7 variables: L_g, t_{ox}, N_d, <w>, D_{rptr}, $<w_{rptr}>$, V_{dd}.

Low temperature case does not include refrigerator power.

Multiprocessor motivation

The energy / performance tradeoff is very steep at the high end. Lower power, more parallel processors potentially offer more computation for the same total power level.

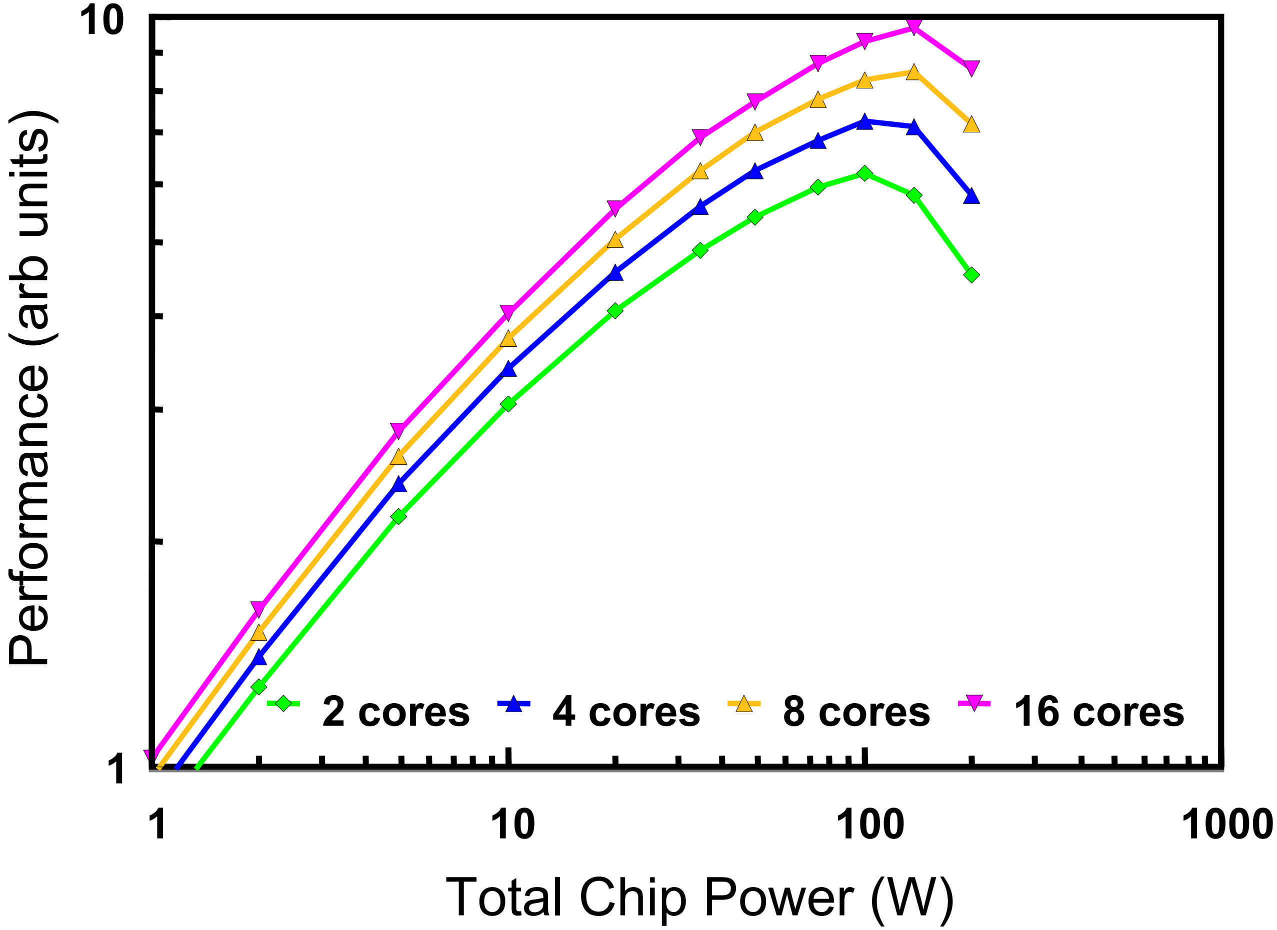


4-processor chips with micro-channel water cooling, optimizing ‘everything’.

9 variables: t_{ox}, L_g, N_D, $\langle w \rangle$, V_{dd}, w_{HP}, S_{rpt}, $\langle w_{rpt} \rangle$, x_{halo}

Dependence on number of cores

Constant total number of transistors, divided equally among n cores:

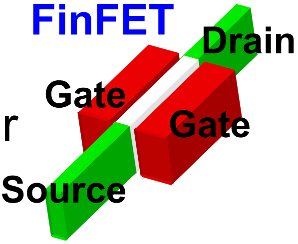


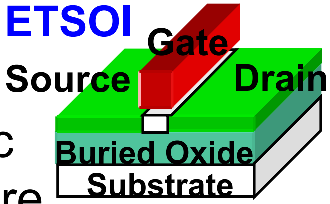
4. Future Projections (22 → 11nm)

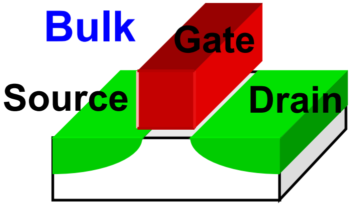
- Device options
- General results
- Technology projections
- Beyond 11nm?

Device Options

- PDSOI
  - IBM's best understood technology.
- FinFET
  - Improved electrostatic control of channel offers shorter gates, lower voltages, higher speed. Workfunction V_T control. Not entirely planar.
- ETSOI
  - Somewhat improved electrostatic control compared to PDSOI. More compatible with conventional planar processing. Workfunction V_T control.
- Shallow Bulk
  - Shallow junctions, raised S/D.





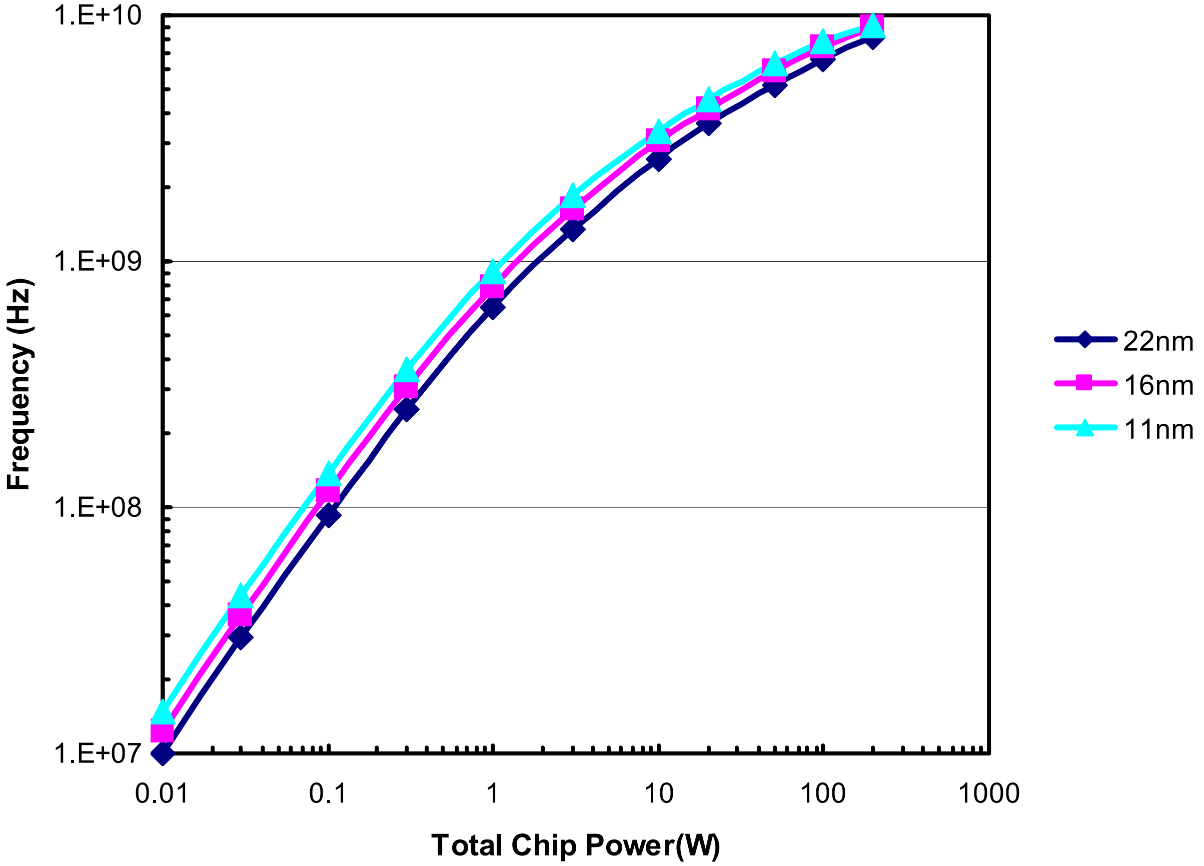


The ETSOI and FinFET devices simulated using an in-house scaling model.

Comparable source/drain resistance and parasitic capacitance models were implemented for ETSOI, FinFET, and for shallow bulk MOSFETs.

General optimization results – performance vs power

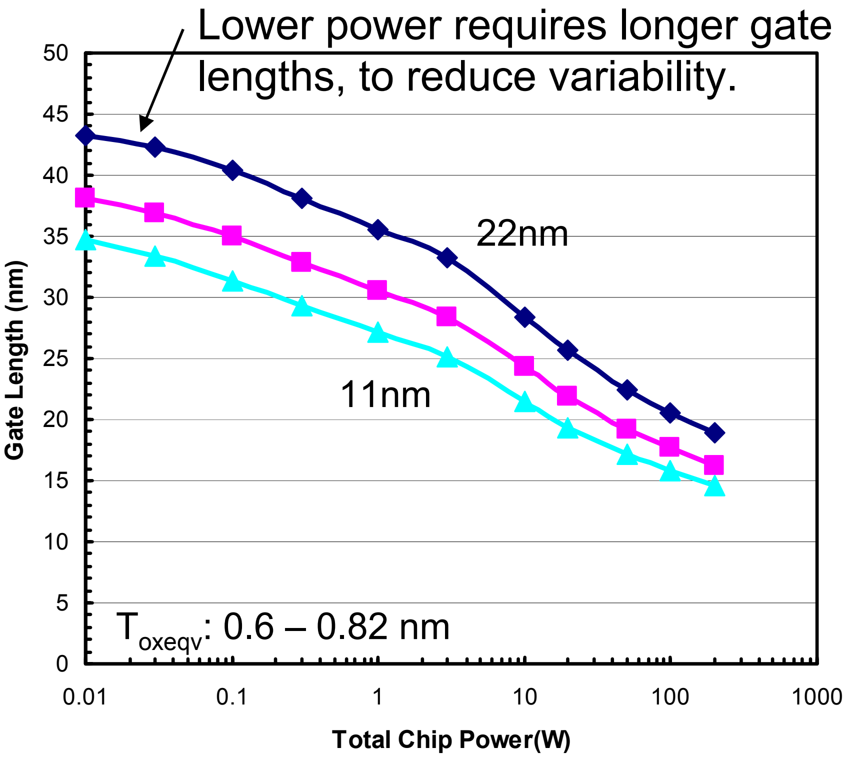
As always, the easiest way to increase performance is to increase the power.



Conditions: PDSOI, 4 core processor chip, constraining total chip power

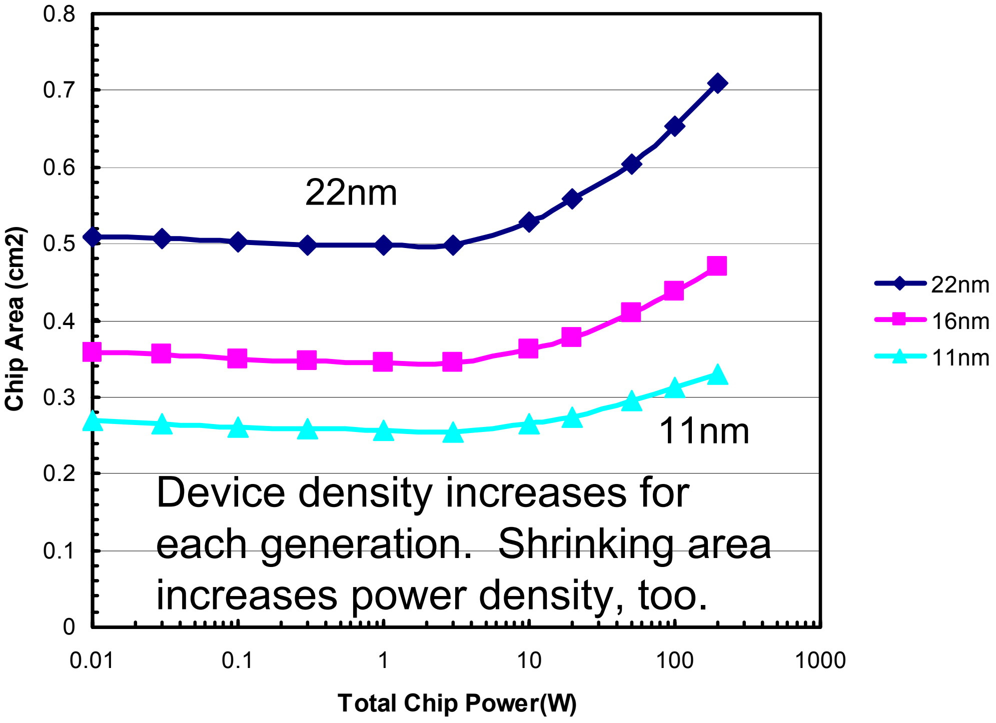
Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing.

Gate Length and Chip Area vs Power



Lower power requires longer gate lengths, to reduce variability.

T_{oxeqv}: 0.6 – 0.82 nm



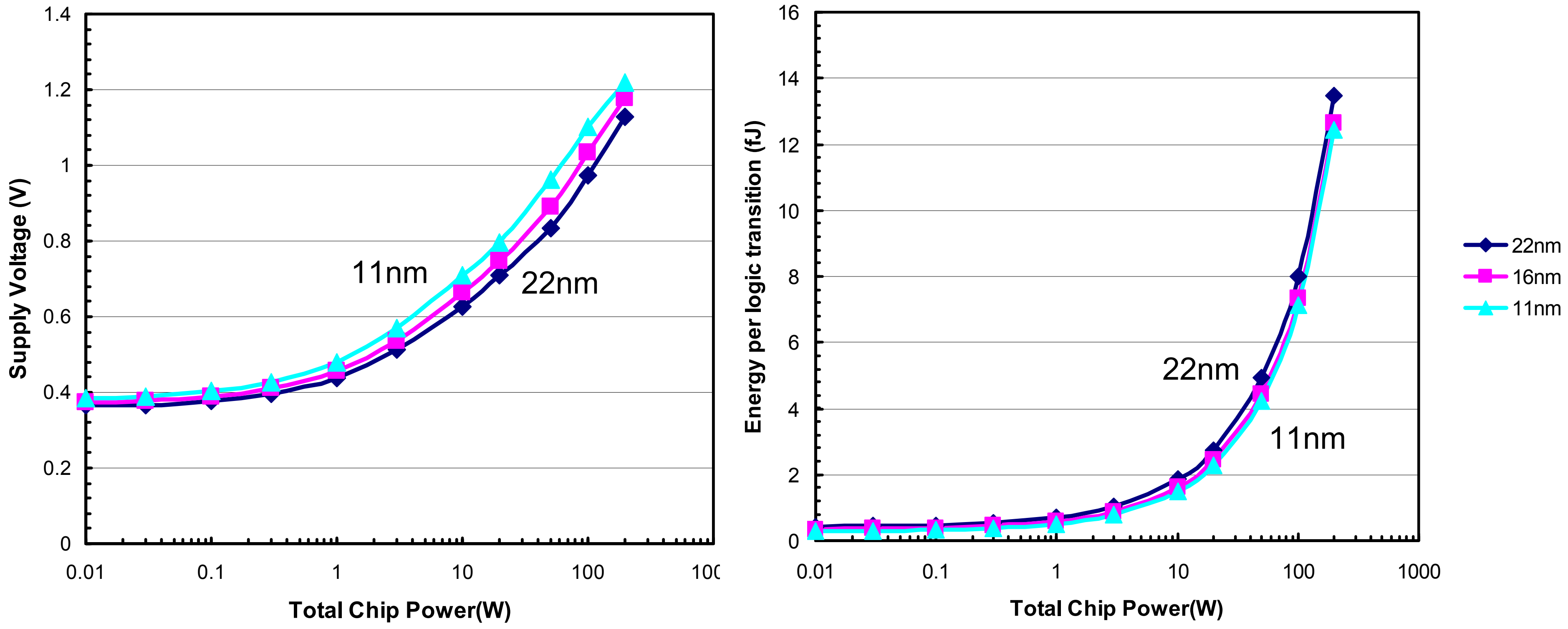
Device density increases for each generation. Shrinking area increases power density, too.

Conditions: PDSOI, 4 core processor chip, constraining total chip power

Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing.

Voltage and Energy vs Power

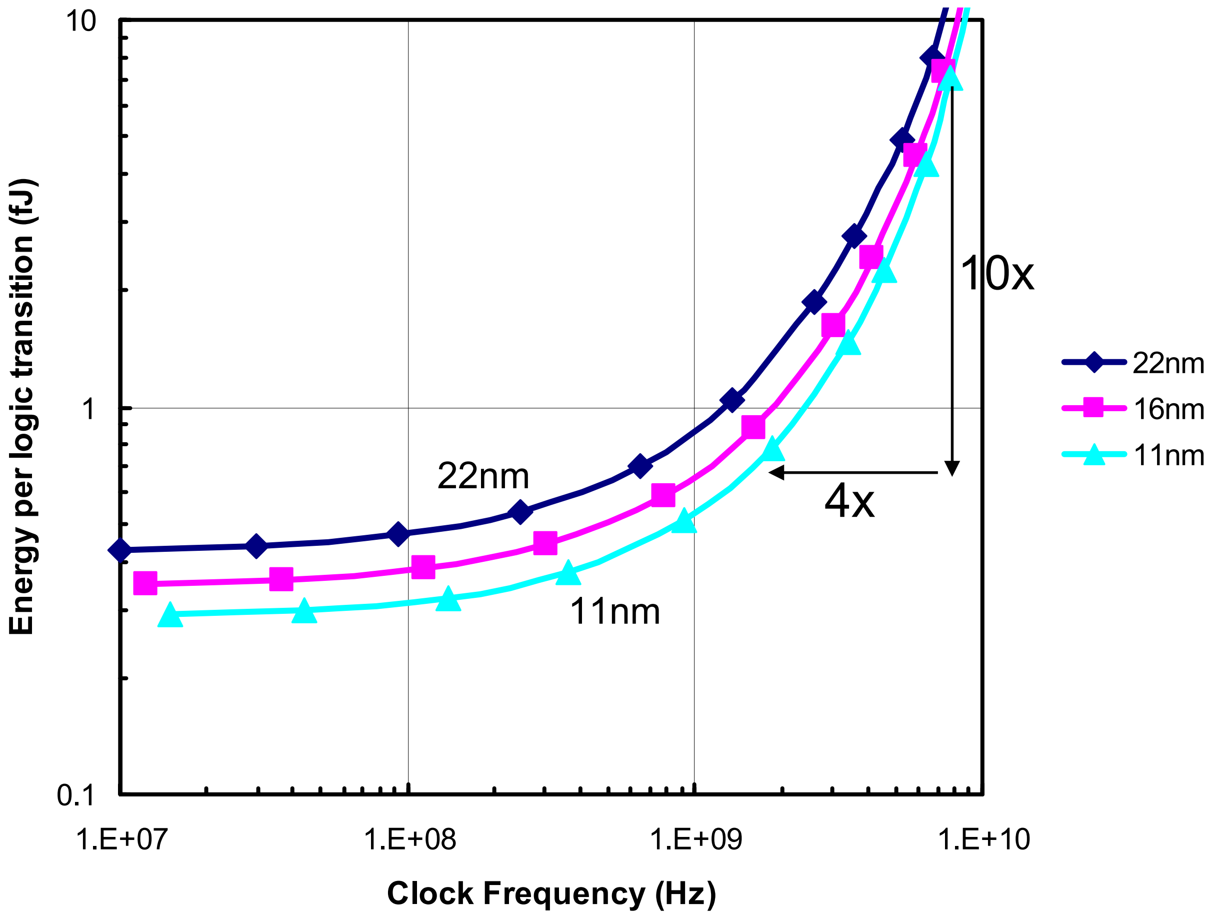
Optimal supply voltage can become quite low for low power constraints, leading to very low energy use per logic transition.



Conditions: PDSOI, 4 core processor chip, constraining total chip power

Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing.

Energy vs performance trade-off

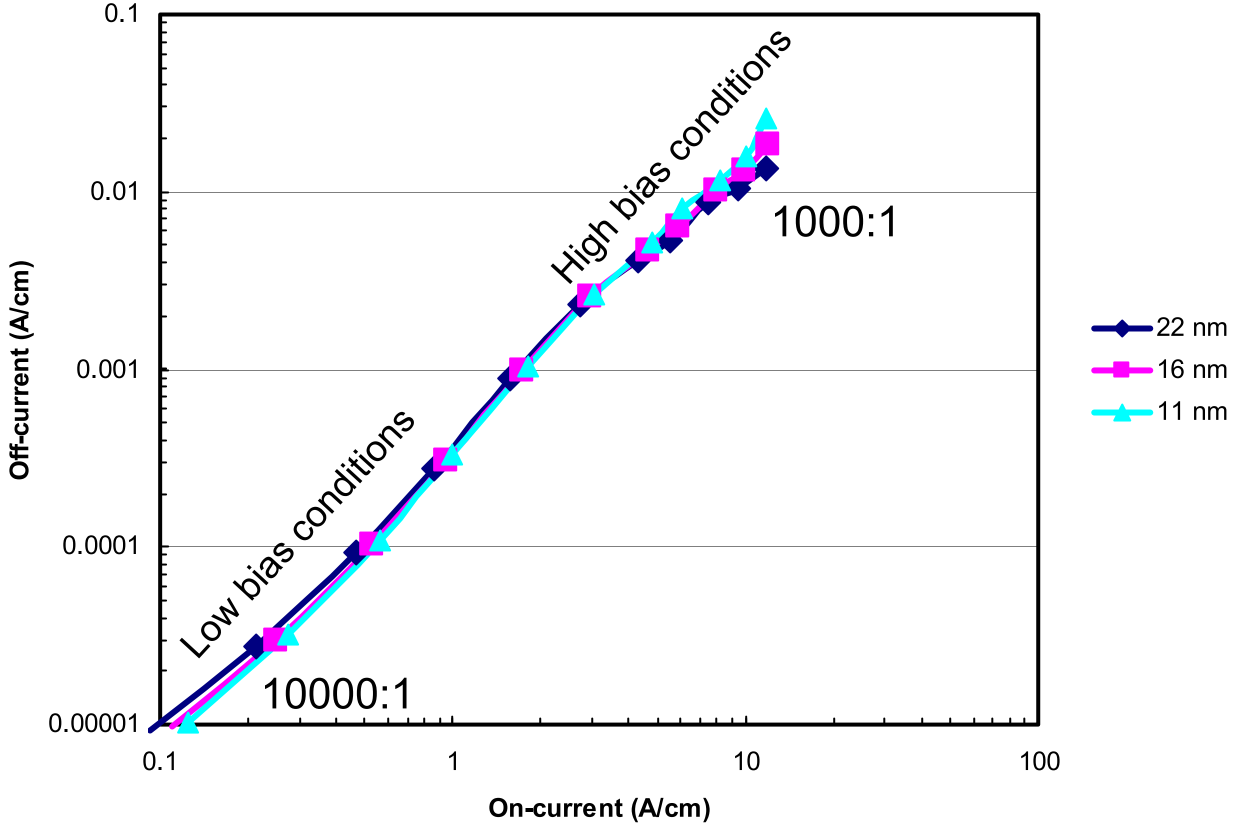


Conditions: PDSOI, 4 core processor chip, constraining total chip power

Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing.

Optimal On/Off Ratio

PDSOI nFETs, currents measured at nominal process and bias conditions

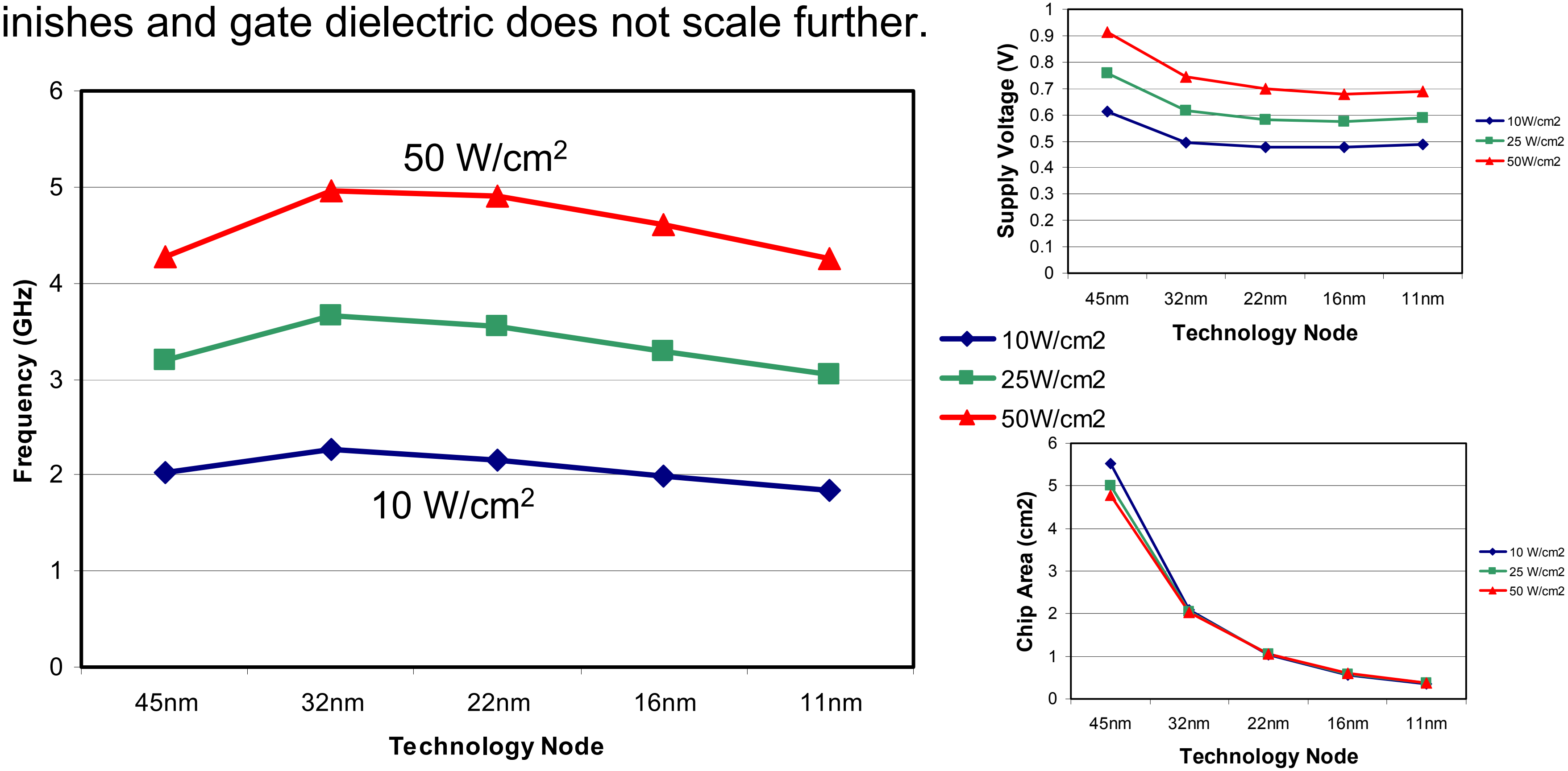


Conditions: PDSOI, 4 core processor chip, constraining total chip power

Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing.

Performance at constant power density (PDSOI)

Performance increases at 32nm due to hi-k introduction, but then falls as strain diminishes and gate dielectric does not scale further.

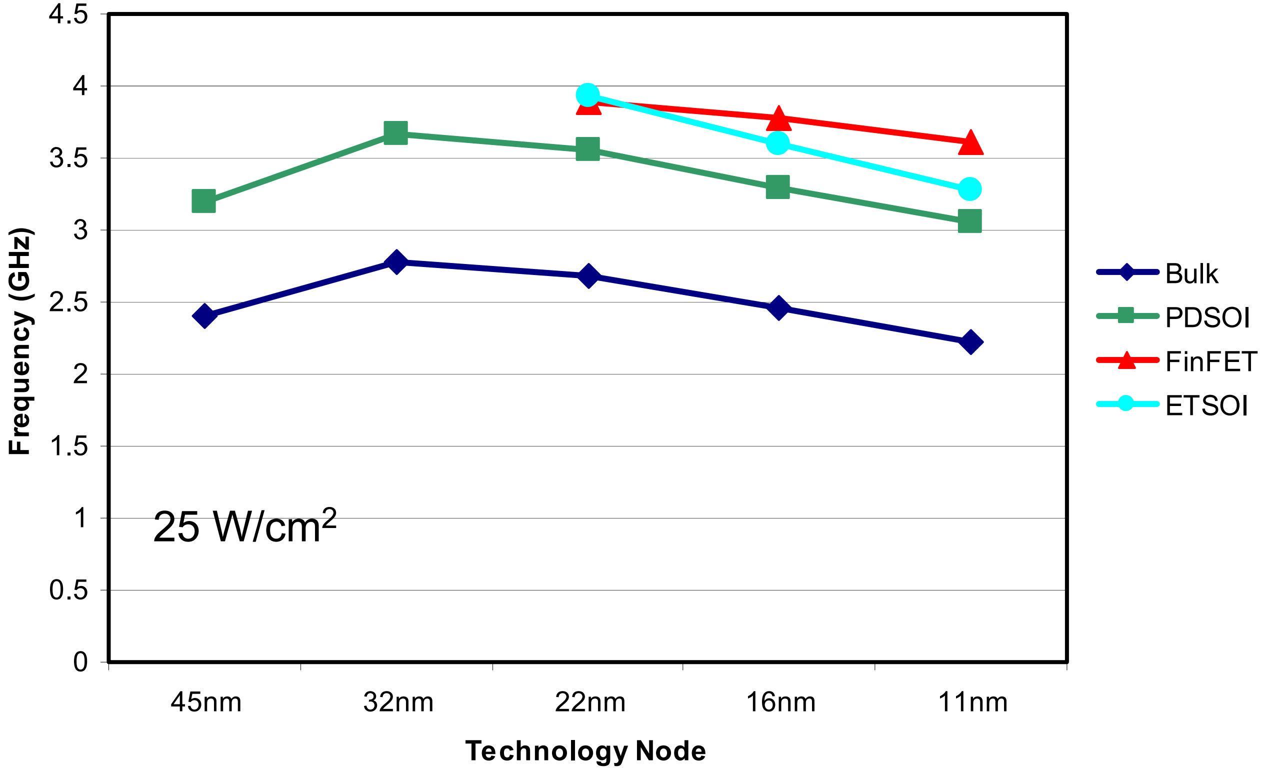


Conditions: PDSOI, 4 core processor chip, constraining total chip power density

Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing.

Performance at constant power density – comparing technologies

ETSOI and FinFET offer moderate performance advantage over PDSOI for 22nm node and beyond. The industry should transition to FinFET at 16nm to avoid performance loss.



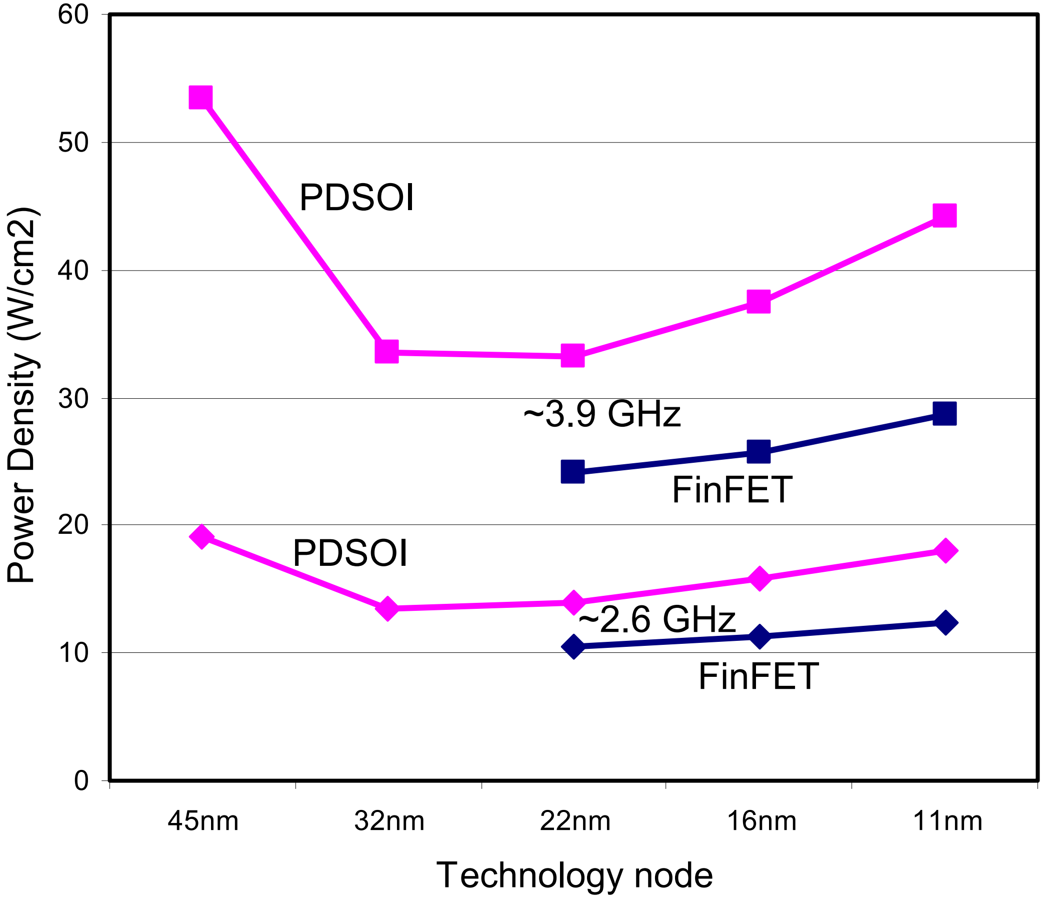
Conditions: 4 core processor chip, constraining total chip power at 25 W/cm^2

Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing, fin height, sidewall thickness (Fin), Si thickness (ET).

Power density at constant performance

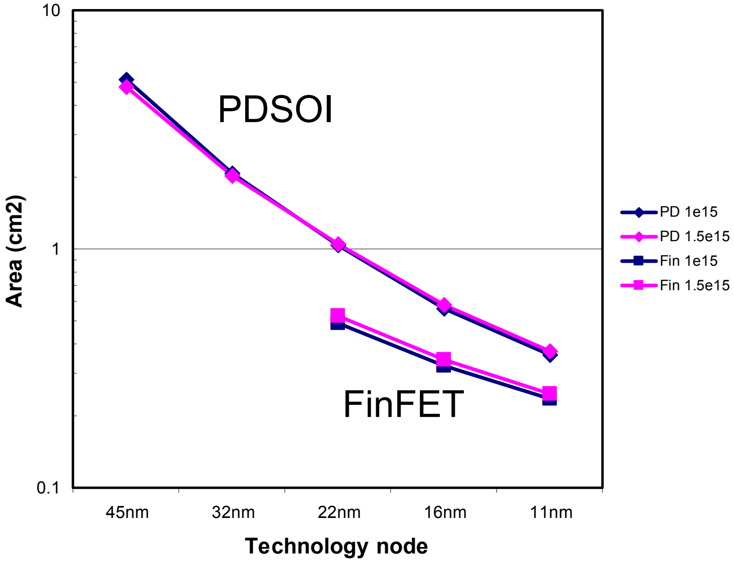
Required power density drops at 32nm due to transition to hi-k, then rises due to decreasing strain and lack of scaling of gate insulator.

Transition to FinFET at 16nm to obtain continued improvement.

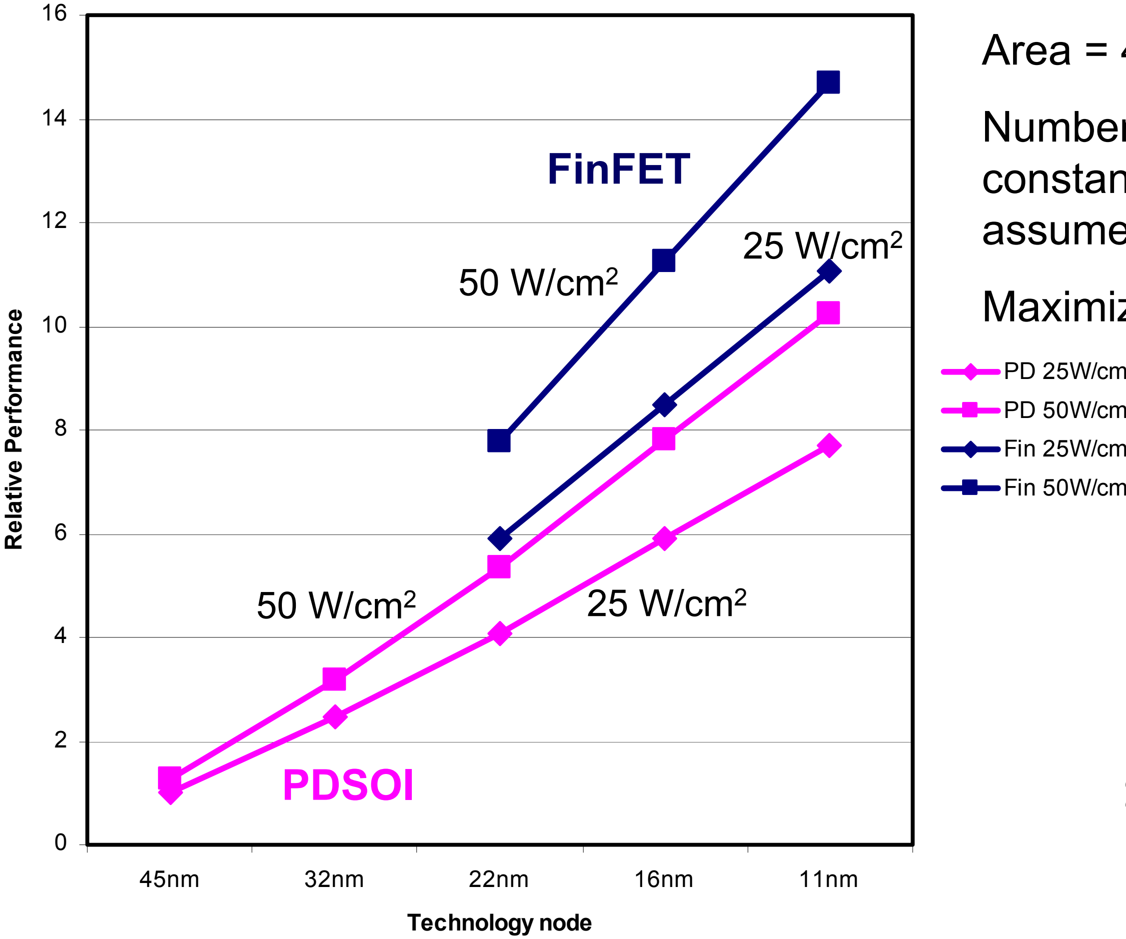


Conditions: PDSOI and FinFET, 4 core processor chip, constraining total chip <u>performance</u>

Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing, fin height, sidewall thickness



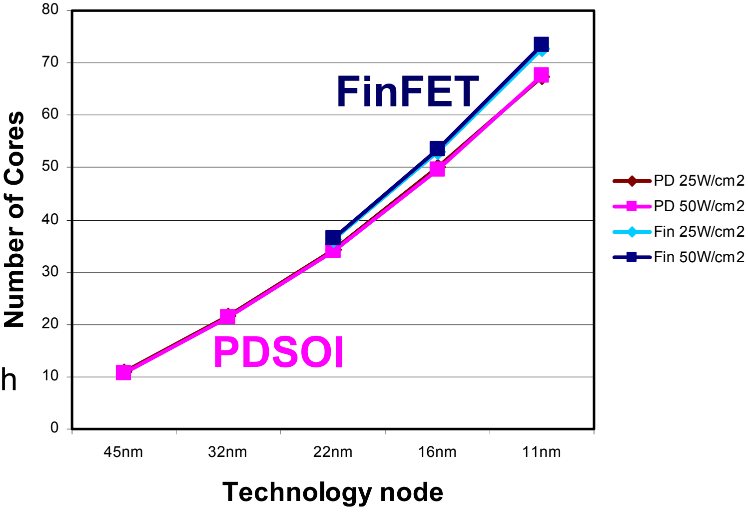
Performance at constant area and power



Area = 4 cm^2, power = 100 W, fixed.

Number of cores is adjusted to maintain constant area. Chip performance is assumed linear with the number of cores.

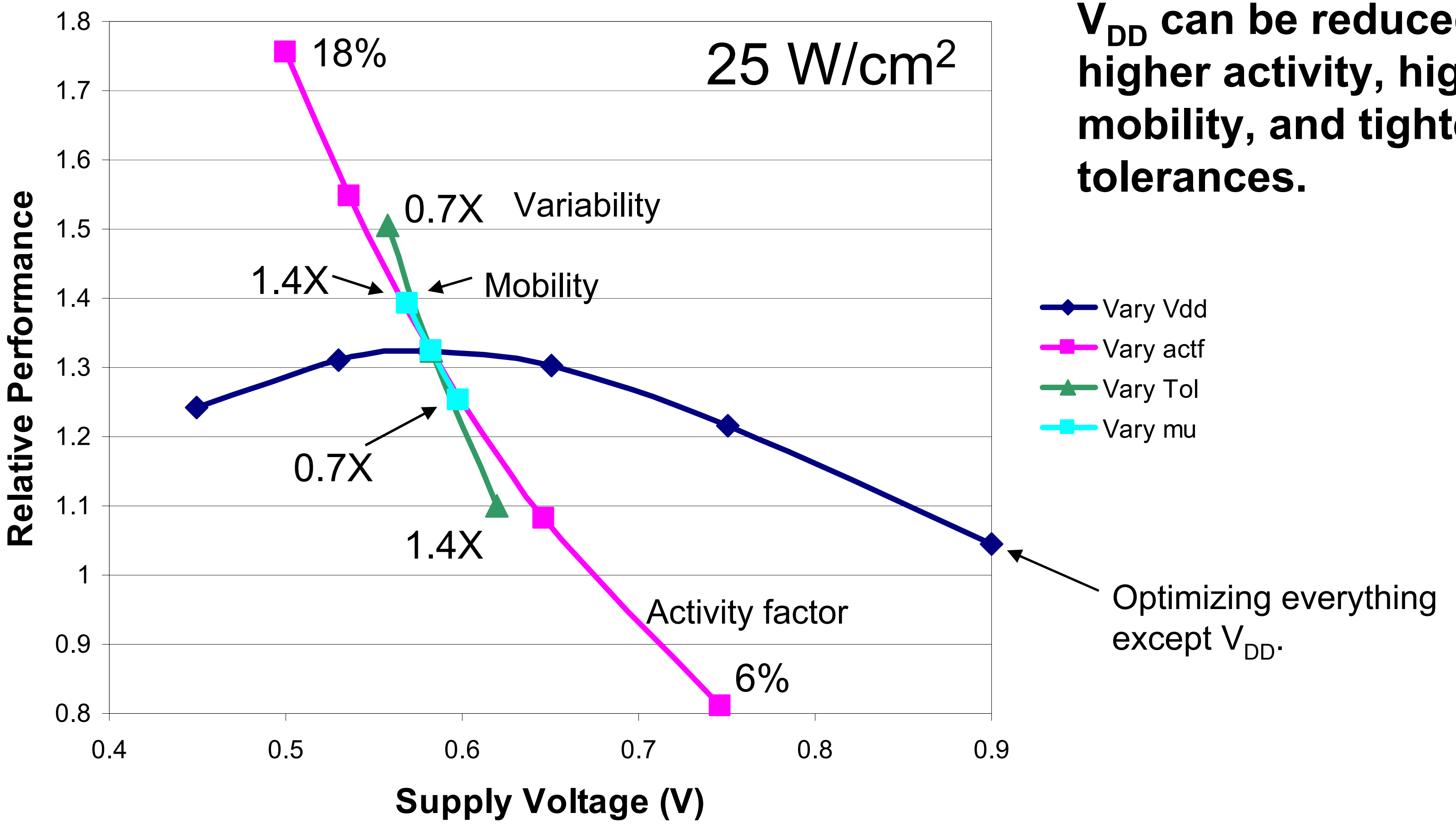
Maximizing total chip performance.



Conditions: PDSOI and FinFET, variable # core processor chip, constraining both chip power and chip area (4 cm^2).

Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing, fin height, sidewall thickness, and number of cores.

Supply voltage considerations



V_{DD} can be reduced by higher activity, higher mobility, and tighter tolerances.

Conditions: PDSOI, 4 core processor chip, constraining total chip power density

Optimizing: V_{DD}, t_{ox}, dopings (for V_Ts), L_G, p:n width ratio, mean widths, repeater size and spacing.

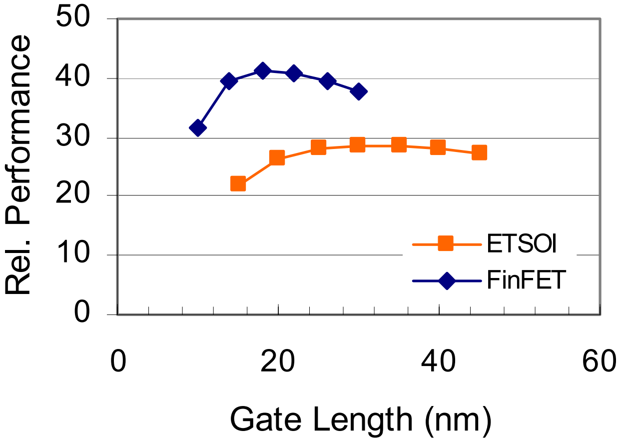
Optimizations vs gate length for FinFETs and ETSOI

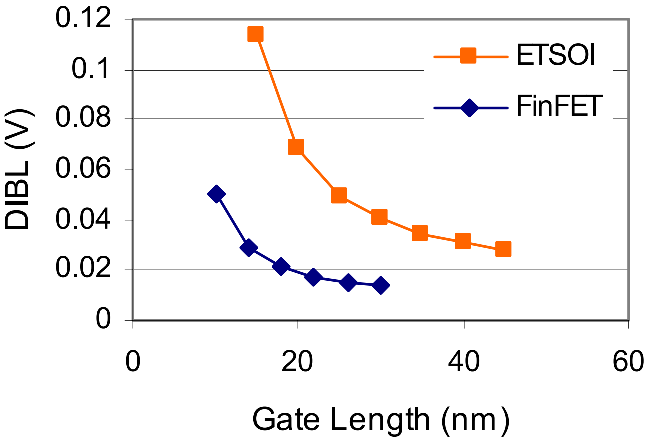
Everything <u>except</u> gate length is optimized. The gate length is scanned.

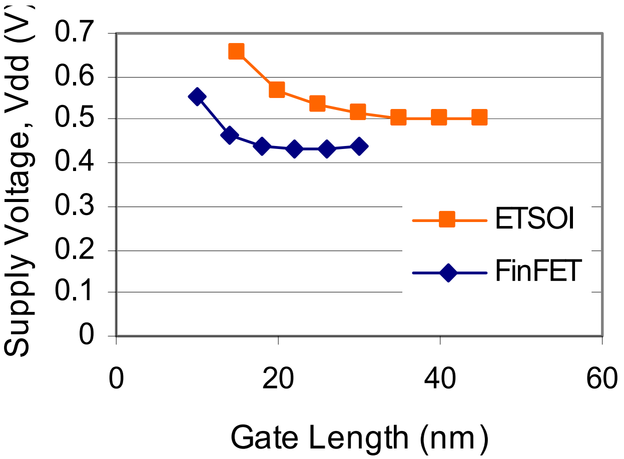
25 W/cm^2 power density constraints.

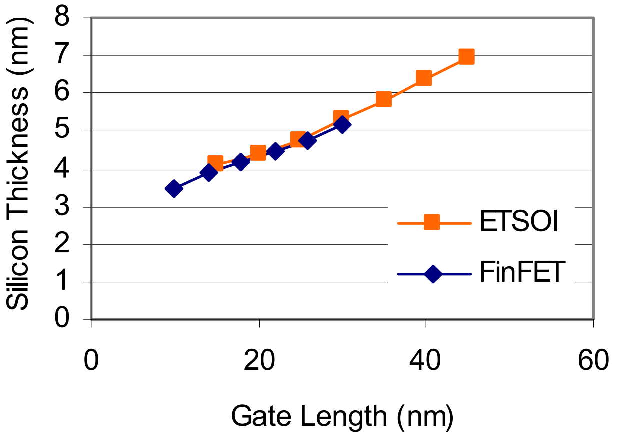
Shorter gate lengths necessitate:

- Higher DIBL
- Higher V_{DD}
- Thinner t_{Si}
- Ultimately, lower performance





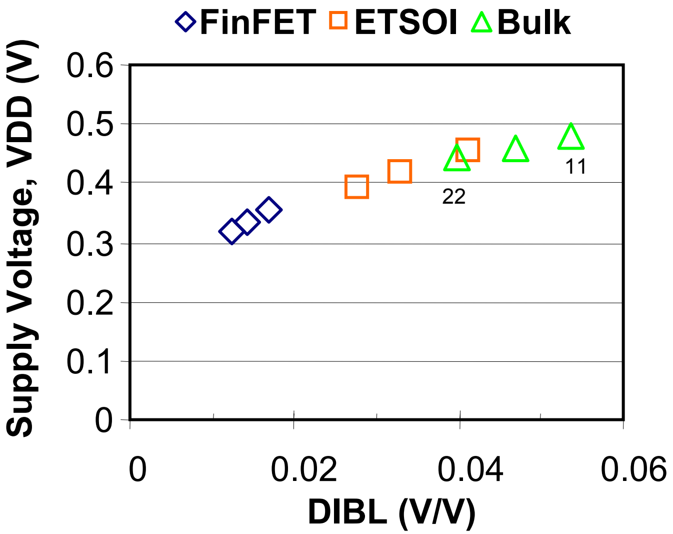


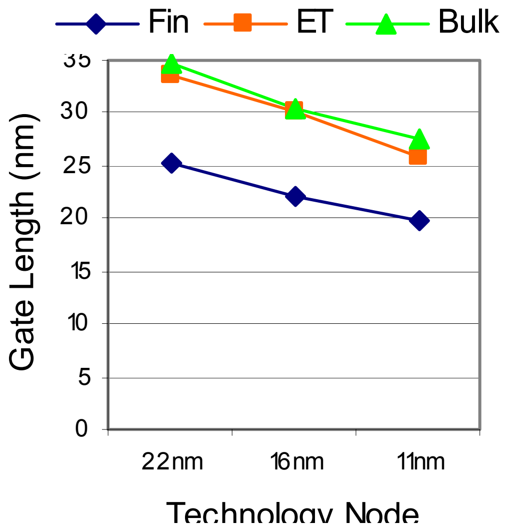


11 nm node

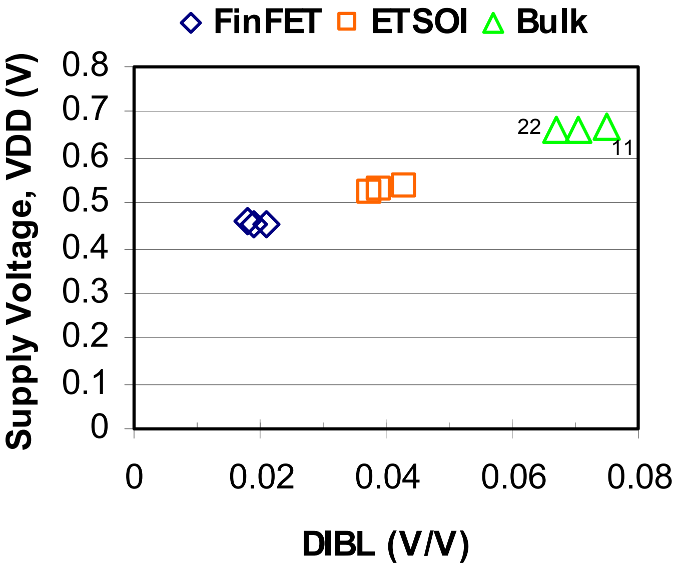
DIBL Scaling enables VDD Scaling

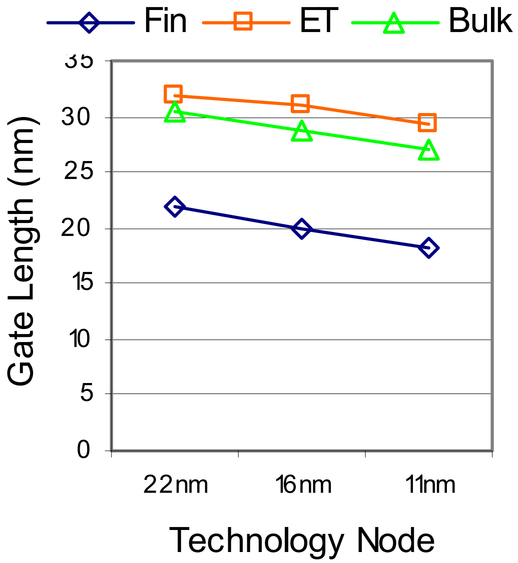
1W





25W/cm^2

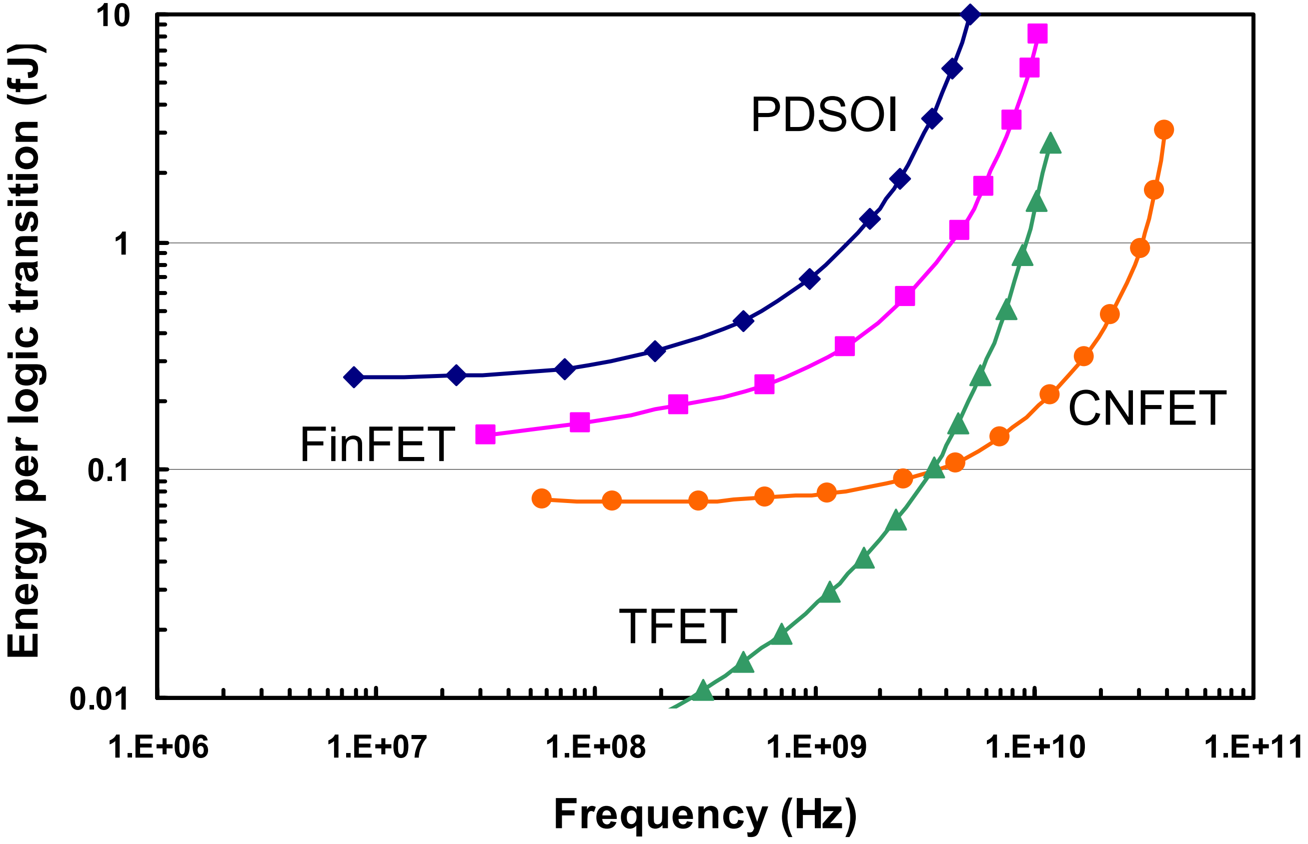




Novel devices for 11nm and beyond

- III-V FinFETs
  - Higher mobility improves drive current
- Tunnel FETs
  - Improved subthreshold slope enables low V_{DD} and low energy operation.
  - To properly model this device, have to be able to calculate the tunneling barrier shapes and band-edge alignments.
  - We are in the process of developing a compact model for TFETs for the optimizer, but results are not yet available. As an interim measure, we can alter the Boltzmann constant in the conventional FET model, to see the impact of steeper subthreshold slope.
- Carbon Nanotube Transistors
  - Ballistic current flow in the channel should enable very high switching speeds for these devices.
  - A compact model for CNTs suitable for the optimizer is being presented at IEDM this year, but results are not available yet.

Comparing devices – Energy vs Performance



[General trends, not exact results.]

Summary

- CMOS scaling is limited by electrostatic, quantum mechanical, discreteness, thermodynamic and practical effects.
- Optimization can and should be used to find the best design points in the midst of these various constraints.
  - Example: low power needs somewhat less scaled devices.
- Technology performance projections based on optimization for PDSOI, FinFETs, ETSOI, and Bulk MOSFETs show:
  - Density improvements should continue, as long as wiring density continues to improve
  - Performance improvements are likely to be rather modest, even with a switch to FinFETs for 16 and/or 11nm nodes.
- Exploratory devices (CNTs and/or TFETs) may offer substantial performance advantages, someday….

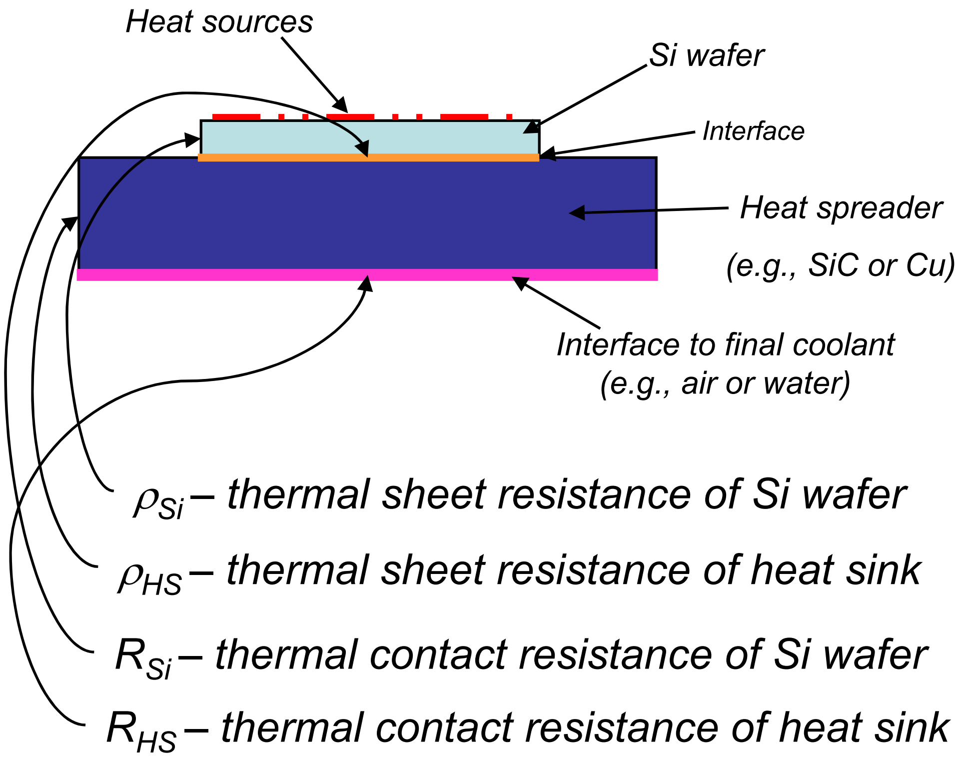
Acknowledgements

- Wilfried Haensch
- Leland Chang
- Paul Solomon
- Steve Koester
- Lan Wei
- Philip Wong
- Ghavam Shahidi
- Mike Scheuermann
- Phillip Restle
- Omer Dokumaci
- Mary Wisniewski
- Steve Kosonocky
- Yuan Taur
- Bob Dennard

Extra slides

Generalized heat sink model

- Two level heat flow model:
  - Flow in the silicon wafer
  - Flow in the heat sink material
- In each layer, the flow can be:
  - 3D (spherical) for spots smaller than thickness
  - 2D (cylindrical) at distances larger than the thickness
- In silicon layer, inhomogeneous power dissipation is accounted for, to estimate maximum junction temperature at hottest point.



Heat sources

Si wafer

Interface

Heat spreader

(e.g., SiC or Cu)

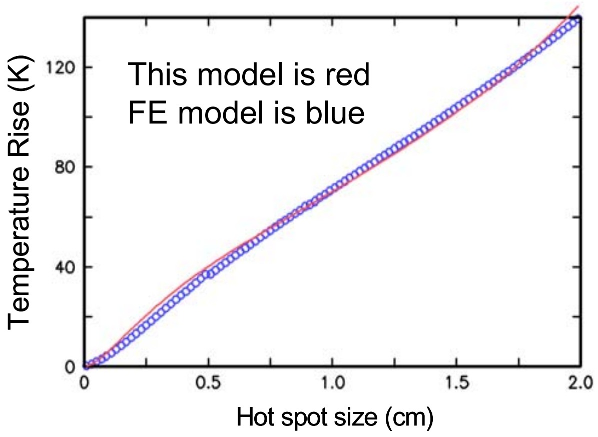
Interface to final coolant
(e.g., air or water)

ρ_{Si} – *thermal sheet resistance of Si wafer*

ρ_{HS} – *thermal sheet resistance of heat sink*

R_{Si} – *thermal contact resistance of Si wafer*

R_{HS} – *thermal contact resistance of heat sink*

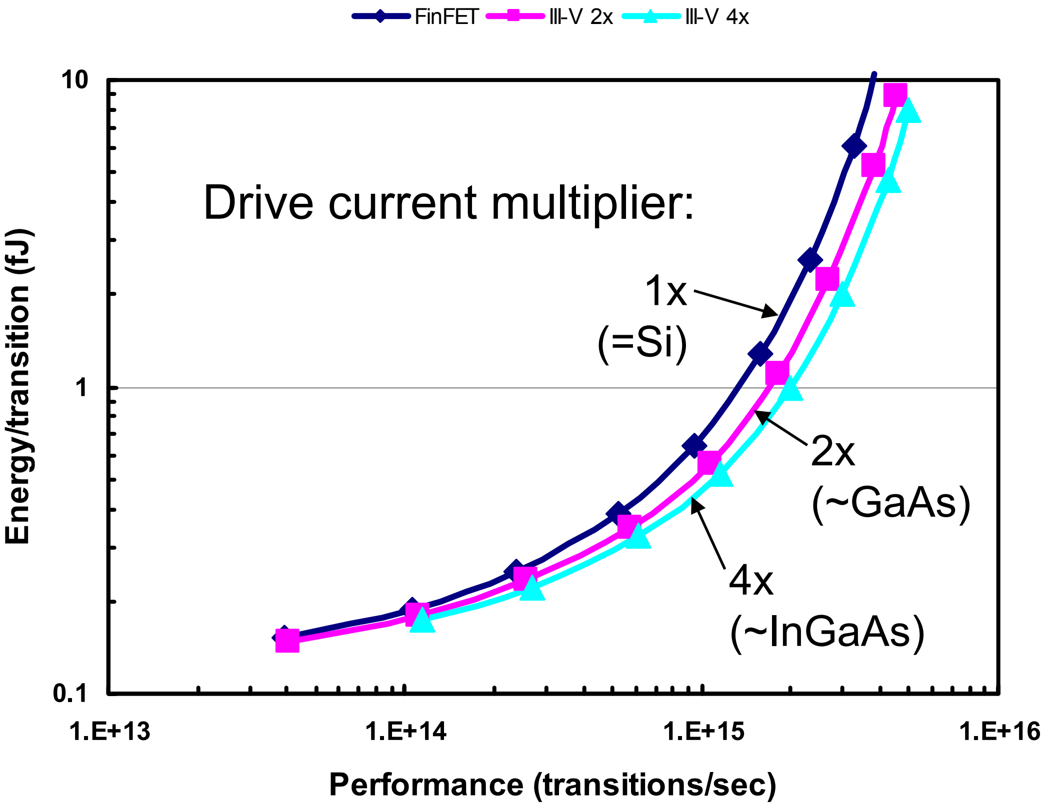


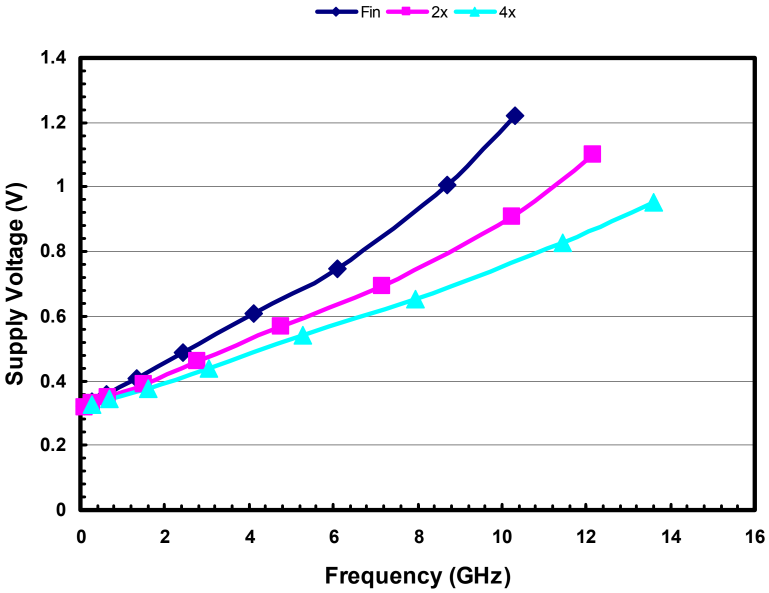
Comparison of simplified analytic model with detailed numerical model.

III-V FinFETs

III-V FinFETs are modeled by increasing the mobility in the conventional model.

Increased mobility enables an improved energy/performance tradeoff by reducing the voltage needed for high performance designs.



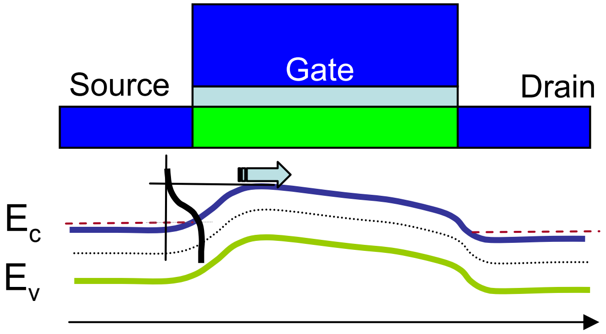


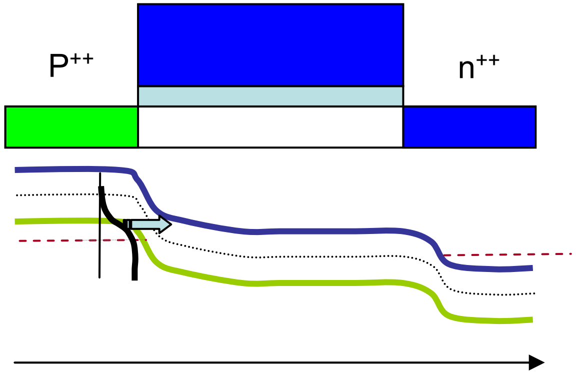
Beating the sub-threshold slope limit

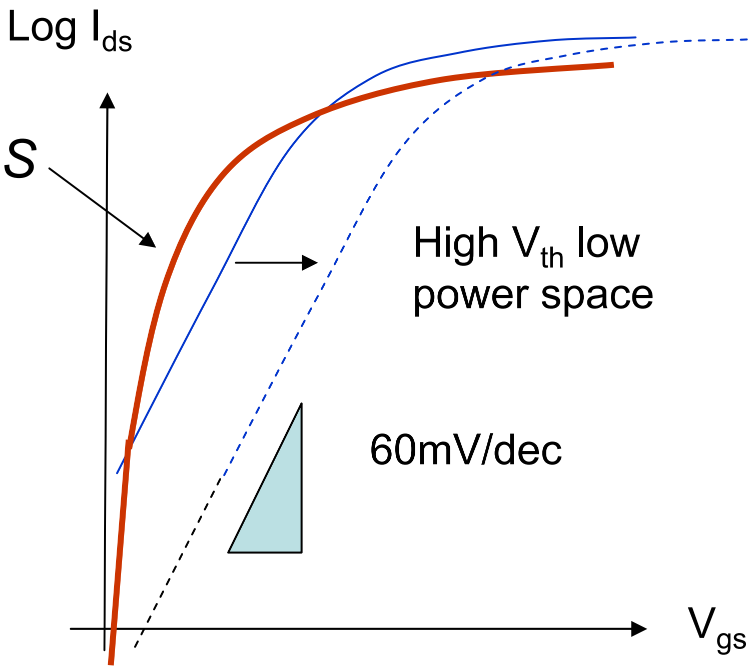
$$T \sim \frac{E^2}{\sqrt{E_G}} e^{-B\frac{\sqrt{E_G^3}}{E}} \sim \frac{\sqrt{E_G^3}}{l^2} e^{-B\sqrt{E_G}l}$$

$$S \sim \frac{V_{gs}^2}{2V_{gs} + D(V_{gs}, V_{ds})}$$

Krishna K. Bhwalka et al. 2005





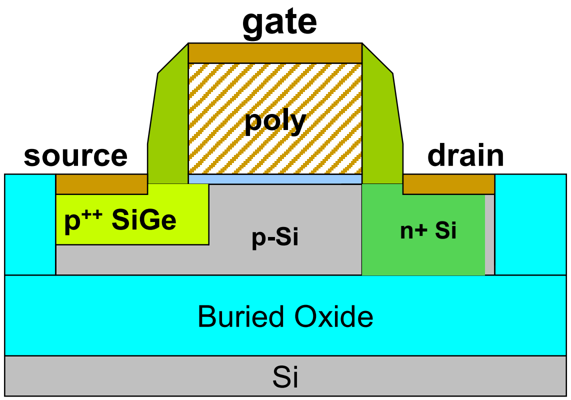


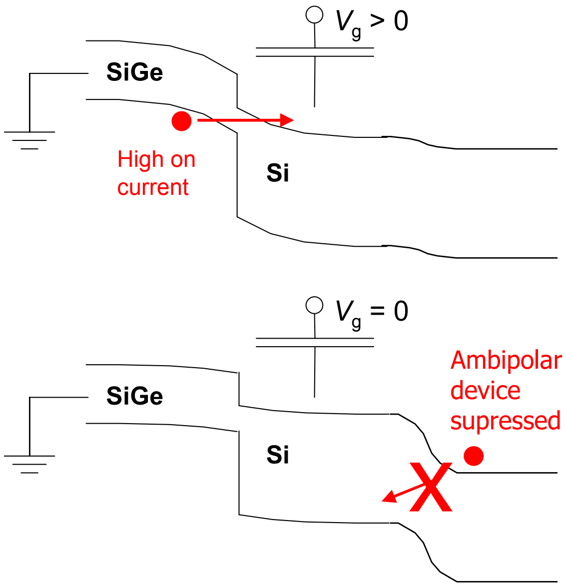
- Tunnel FETs show strong voltage dependence of sub-threshold slope
- On-current not yet on par with conventional high performance FETs at comparable voltages

[Haensch]

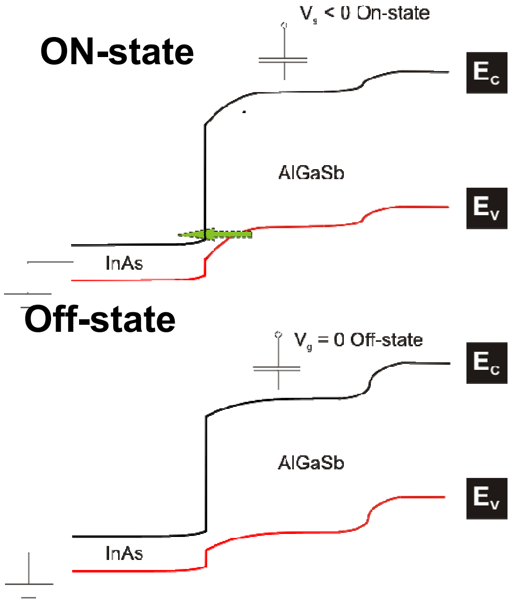
TFET Heterostuctures

Planar SiGe H-TFET





III-V H-TFET



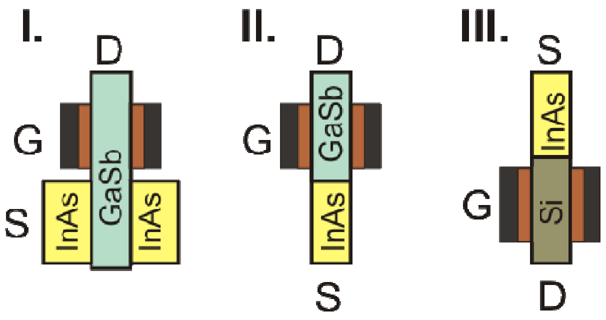
<u>SiGe Heterostuctures</u> offer low effective bandgap, which improves tunneling.

<u>III-V materials</u> offer lower effective masses and more heterojunctions, to further improve tunneling.

<u>Nanowire Geometry</u> Offers:

- Optimum electrostatics with gate all around
- New material combinations
- Integration onto silicon possible
- Scaling to quantum capacitance limit
- Improved *I*on

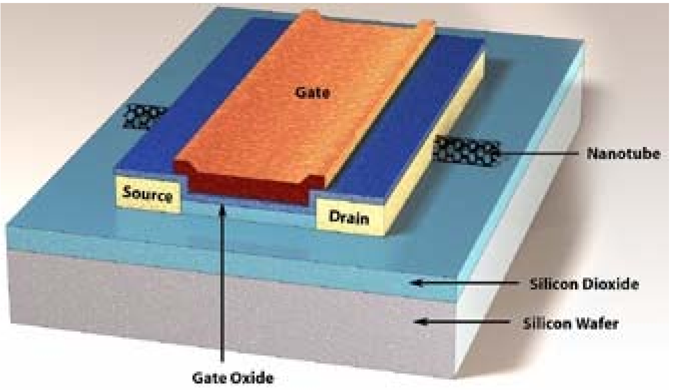
Nanowire Geometry: Device Cross Sections



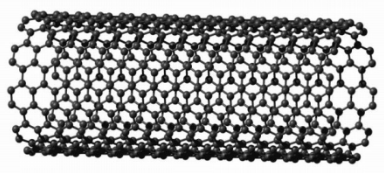
[Koester, Riel, Koswatta]

CNFET: Good or Bad?

- Carbon Nanotube Field Effect Transistor (CNFET)
  - 1D devices
  - Better transport characteristics ☺
  - Worse parasitics ☹
  - Leakage, Variations, etc…



- CNFET: Good or Bad?
  - The judgment highly depends on the application
  - Our approach is to build a optimizer for a full technology, with proper consideration of device properties and system needs.
    - Device modeling
    - Circuit performance benchmarking
    - System application consideration

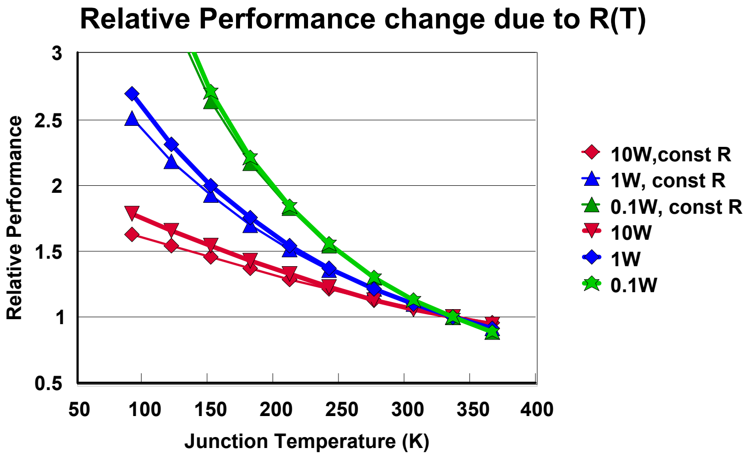


[Lan Wei, Stanford]

Wire Model

Assumed constant 2:1 height to width wiring with equal lines and spaces. (0.062 k_{BEOL} fF/um)

$$\rho(W,T) = \rho_{300\mathrm{K}} \cdot \left(\frac{70\mathrm{nm}}{W} + \frac{\ln\left(1+e^{(T-40)/10}\right)}{26} \right)$$



Consequences of wire resistance model: