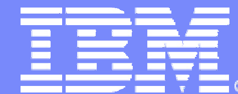
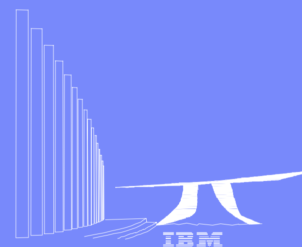


The Limits of CMOS Scaling from a Power-Constrained Technology Optimization Perspective

D. J. Frank
IBM T.J. Watson Research Center, Yorktown Heights, NY

Purdue University seminar Oct. 4, 2006



Presentation is © 2006 IBM
© 2004 IBM Corporation

Acknowledgements

- Wilfried Haensch
- Ghavam Shahidi
- Omer Dokumaci
- Mary Wisniewski
- Mike Scheuermann
- Phillip Restle
- Steve Kosonocky
- Evan Colgan
- Philip Wong
- Yuan Taur
- Paul Solomon
- Bob Dennard

© 2004 IBM Corporation

Outline

1. **General Limitations to Scaling**
2. **Power-Constrained Technology Optimization**
   - Models and Assumptions
   - Optimization Results
3. **Minimum Energy from Optimization**
4. **Open Questions**
5. **Summary**

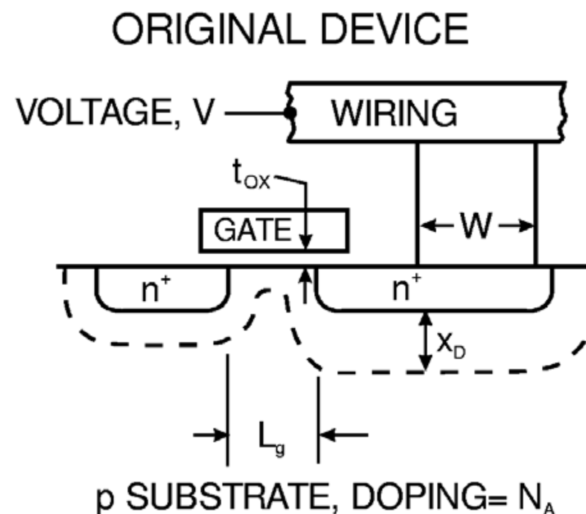
© 2004 IBM Corporation

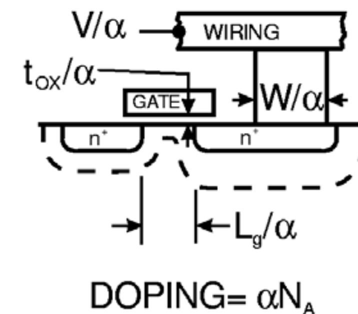
1. Limitations to Scaling

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

Basic idea of Scaling:

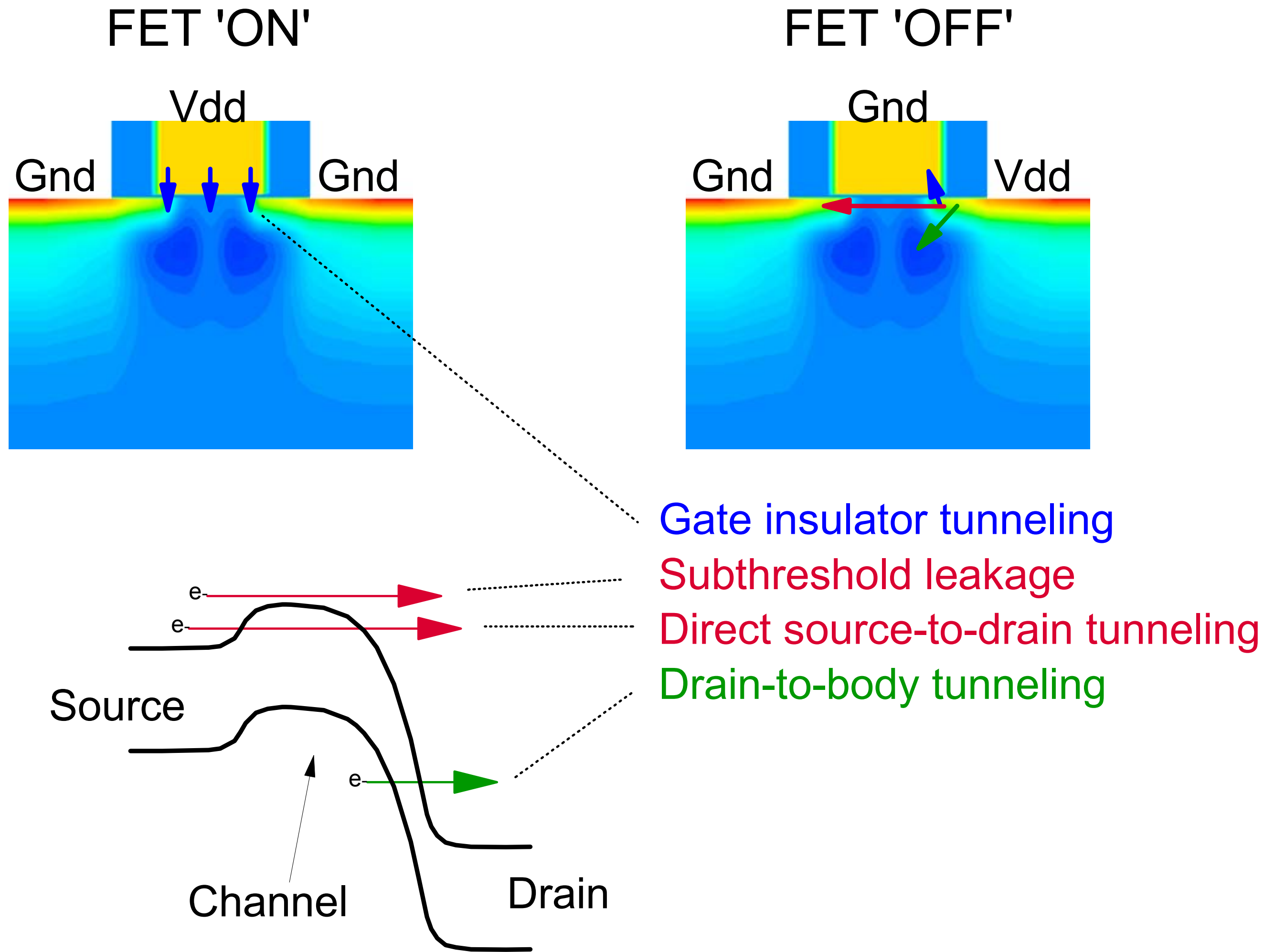
Adjust dimensions, voltages, & doping to achieve smaller FET with same electrostatic behavior.





© 2004 IBM Corporation

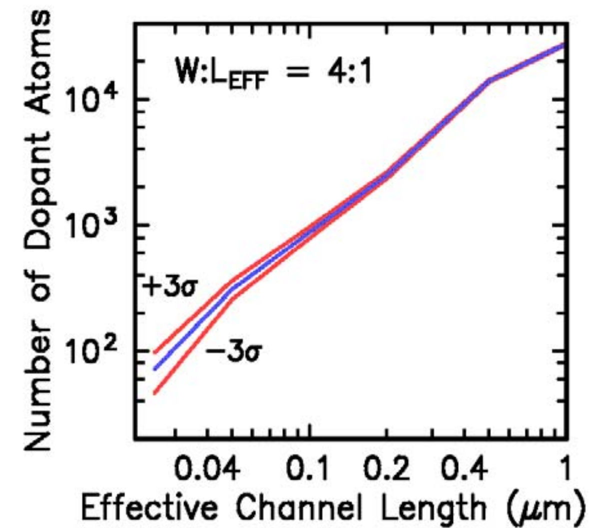
Quantum Mechanical Tunneling Leakage Currents



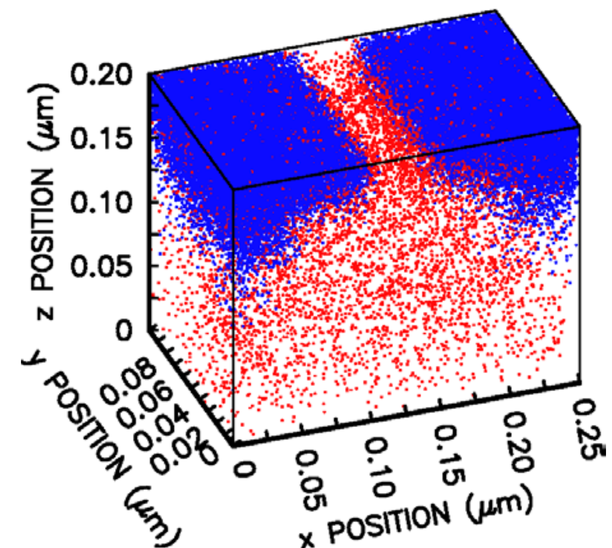
© 2004 IBM Corporation

Discrete dopant fluctuations

- **The number of dopant atoms in the depletion layer of a MOSFET has been scaling roughly as $L_{eff}^{1.5}$.**
- **Statistical variation in the number of dopants, N, varies as $N^{1/2}$, causing increasing V_T uncertainty for small N.**
- **Specific threshold uncertainties depend on the details of the doping profiles.**
- **3D simulations are required to accurately evaluate these dopant fluctuations.**
- **A preprocessor (called MCMESH3D) was written for FIELDAY:**
  - Checks every Si atom site to see if it is a dopant
  - Transfers these dopants to the simulation mesh
- **3D FIELDAY simulations of subthreshold current are run on ~100 different cases to statistically evaluate σ_{VT} for any given design.**
  - Use constant mobility model to avoid unphysical mobility dependence on dopant positions.



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

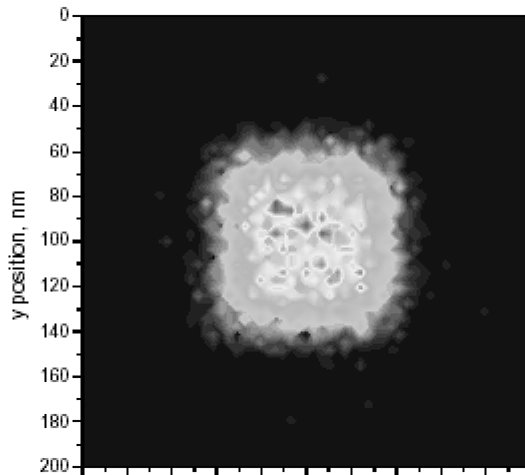


[D. J. Frank, et al., Symp. VLSI Technol., p.169, 1999 and D. J. Frank and H.-S. P. Wong, IWCE, p.2, May 2000]

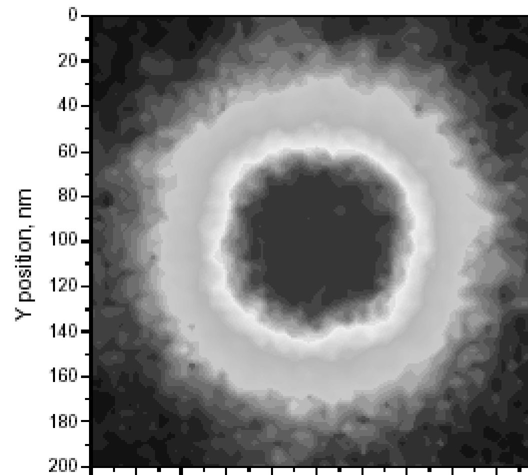
© 2004 IBM Corporation

Simulated Contact Hole Exposure --Discreteness of photons

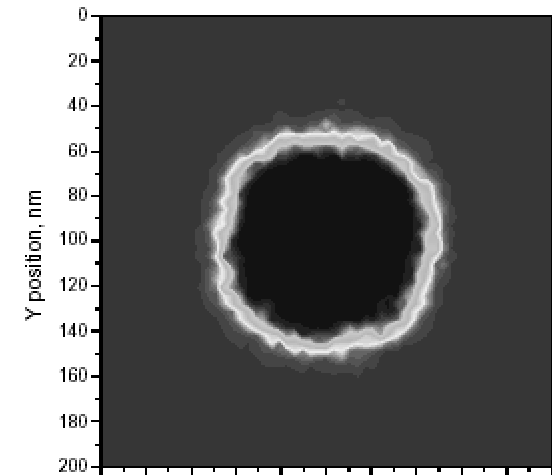
Photons absorbed



Deprotected polymer



Disolved polymer



©2003, SPIE

Monte Carlo simulation of exposure and development of a 80 nm contact hole using EUV lithography.

[J. Cobb, et al., Proc SPIE]

© 2004 IBM Corporation

Material Properties

1. Bandgap.

The Si bandgap does not scale, but

(a) this is not a major problem, and

(b) it can be overcome by forward biasing the body.

2. Dielectric constants.

ITRS roadmap requires high-k gate insulators, but there are few materials that come close to satisfying all the demands:

High k

High barrier (for both electrons and holes, preferably)

Stable on Si at anneal temperatures

No traps or interface states

High reliability

Hafnium silicate-based dielectrics are presently the most promising.

© 2004 IBM Corporation

Thermodynamic limitations

- 1. The Boltzmann distribution. This causes subthreshold leakage current.
- 2. Irreversible computation => All switching energy is converted to heat.
- 3. All leakage currents and IR drops are irreversible => More heat.
- 4. Subthreshold slope sets fundamental limits on logic swing, but this limit is not usually important.

© 2004 IBM Corporation

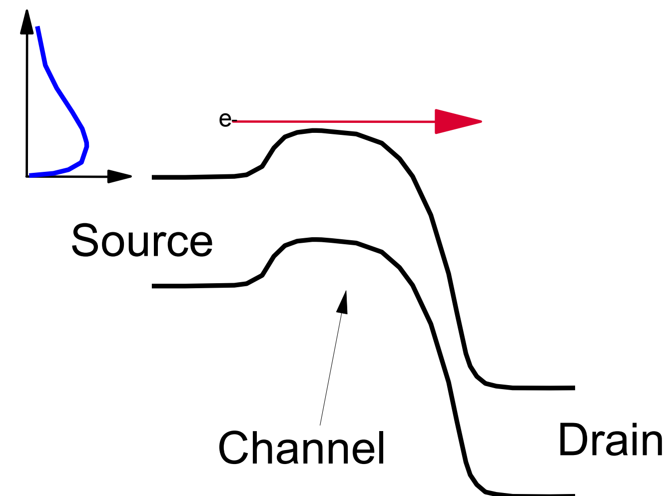
Thermal Current (subthreshold leakage)

- 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.



© 2004 IBM Corporation

Practical and Environmental Issues

- Power consumption and heat removal are limited by practical considerations.
- Low power applications must be 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.)

© 2004 IBM Corporation

2. Technology Scaling Limits from System Performance Optimization

- Since the end of scaling is dominated by practical considerations, it is application-dependent, requiring optimization across device, circuit, and architecture.
- In the past, device, circuit and architecture design have proceeded in parallel, to increase product throughput and reduce complexity, but in an era of diminishing returns, greater performance can be achieved by optimizing across the boundaries.
- As an initial exploration of this regime, a tool has been designed to capture the essential features of each complexity level in order to evaluate the impact of technology options on the performance of future systems.

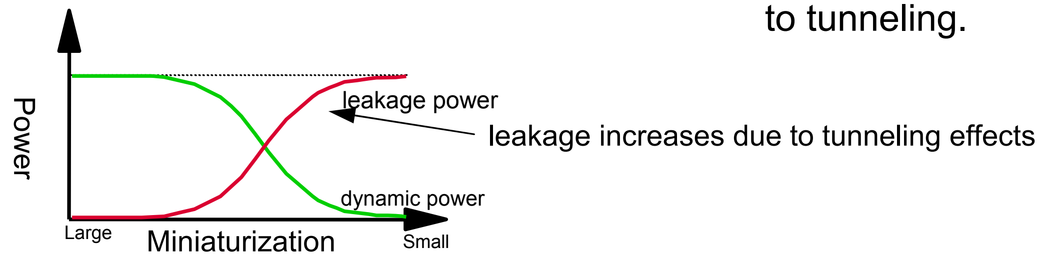
© 2004 IBM Corporation

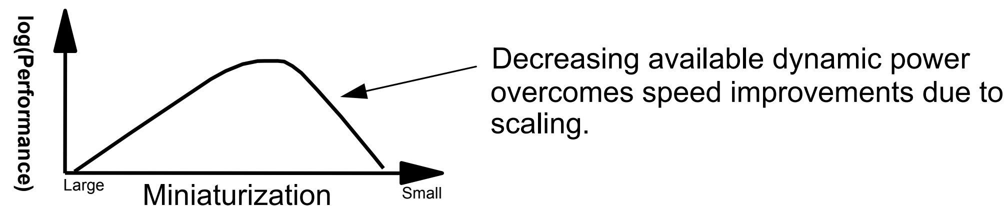
Concept

Existence of an Optimal Technology

Fixed architectural complexity
+ Fixed power constraints
+ Device physics
= Existence of an optimal tech-
nology with maximal performance.

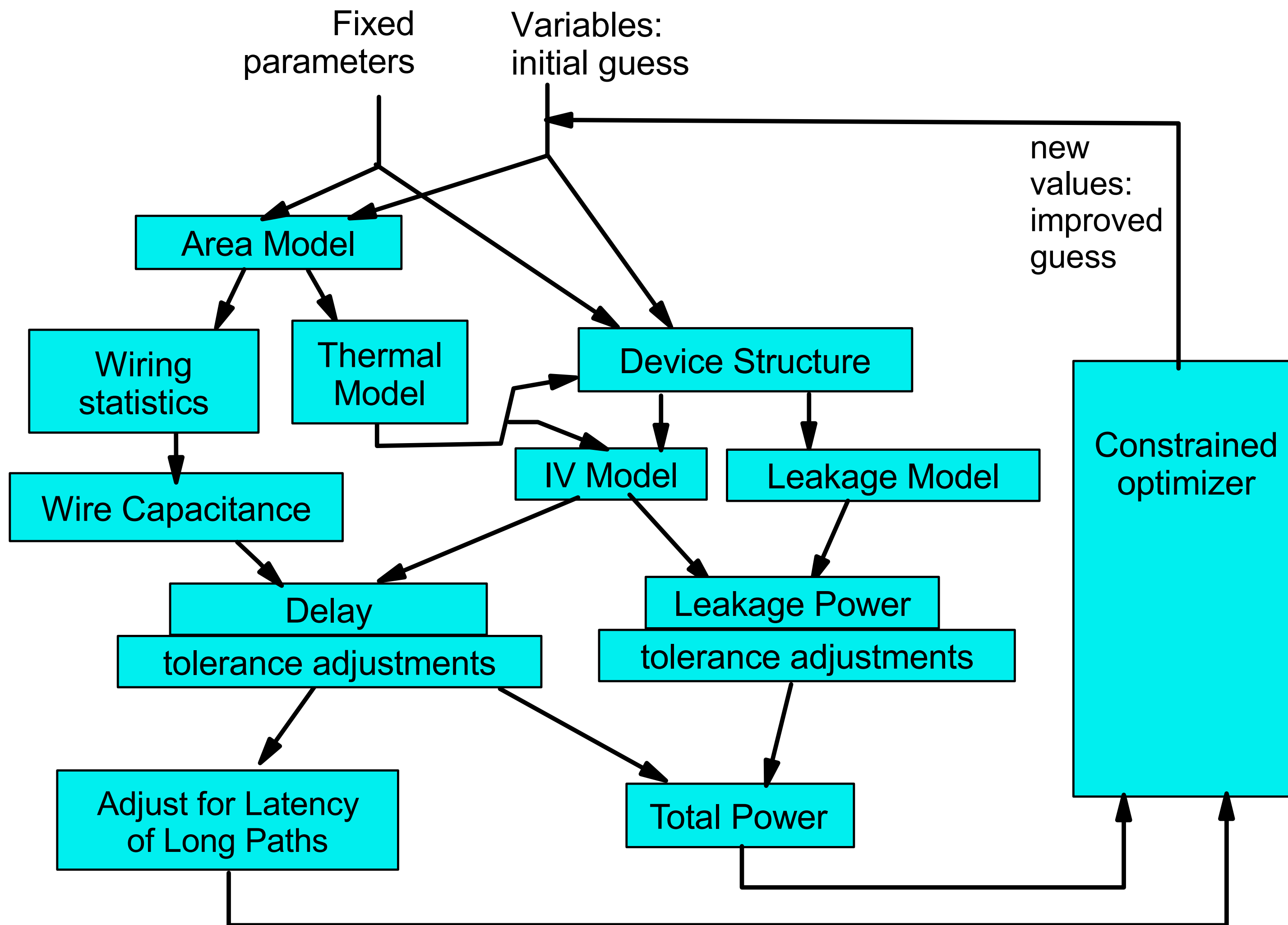
- Practicality imposes power constraints.
- Electrostatics imposes geometric constraints
- Thermodynamics imposes voltage constraints.
- Quantum mechanics imposes miniaturization constraints due to tunneling.





© 2004 IBM Corporation

Schematic organization of optimization program



© 2004 IBM Corporation

Assumptions and Model Details

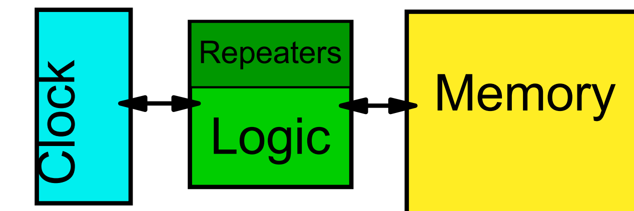
- **Chip-level assumptions**
- **Optimization metric**
- **Device IV curves**
- **Circuit delay**
- **Power dissipation**
- **Thermal model**
- **Wire models**
- **Accounting for variations**

© 2004 IBM Corporation

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.

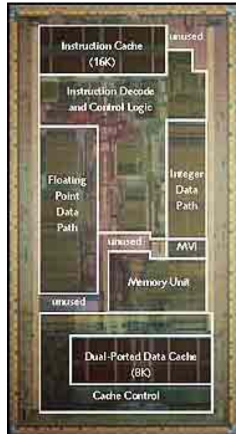


© 2004 IBM Corporation

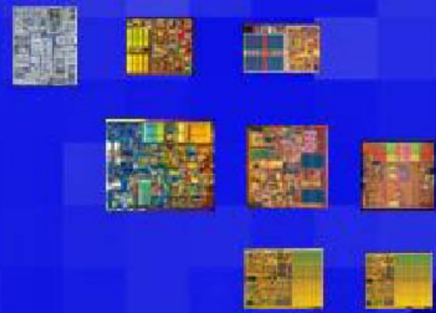
How much area do the processor cores take?

100% to 25%, generally decreasing with generation:

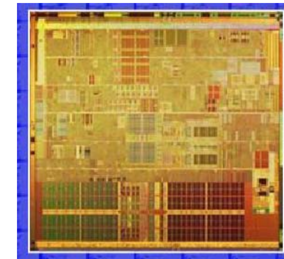
100%



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

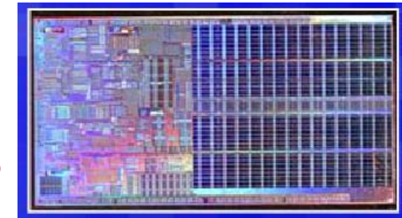


70%

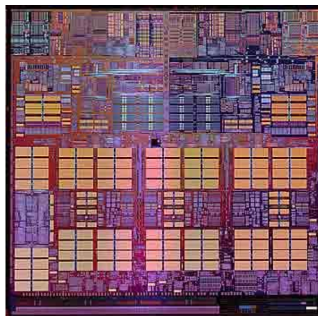


Prescott. 125M FETs

40%

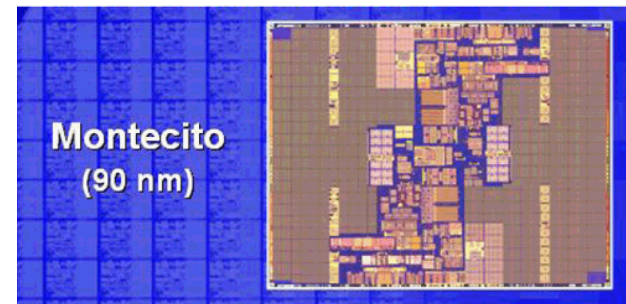


Dothan, 140M FETs



40%

Power4, 174M FETs



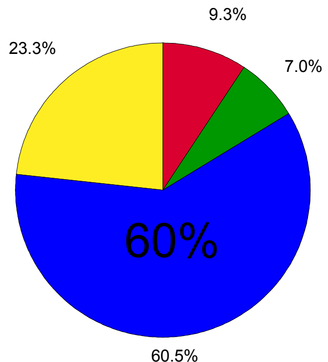
~25% 2 cores, 1.72B FETs

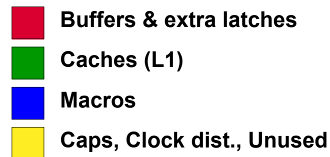
© 2004 IBM Corporation

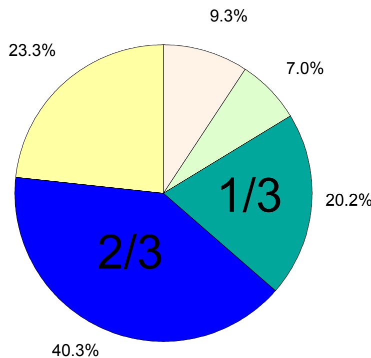
Area usage within a processor core

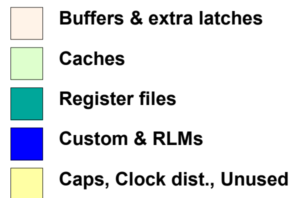
Approximate area fractions for a high-performance microprocessor core in leading-edge technology

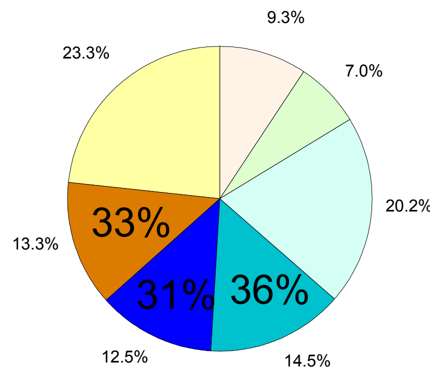
data from:
M. Scheuermann and M. Wisniewski

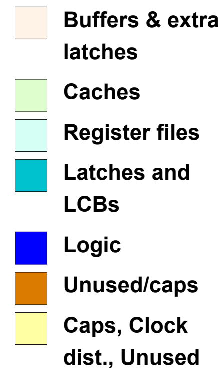


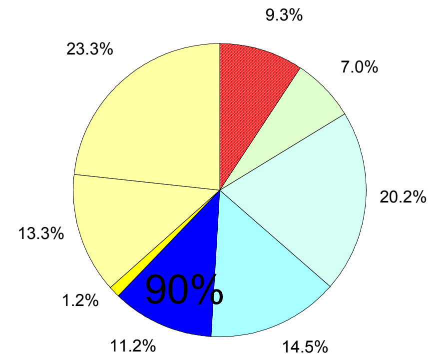


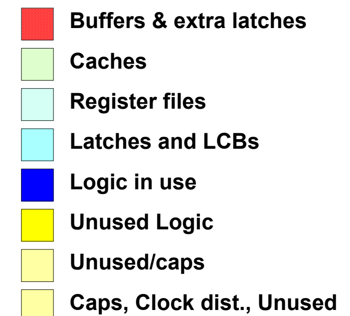












- Processors built with nanotechnology are likely to have similar area usage statistics.
- Nanotechnology may require additional area allocations for defective circuitry.
- Estimates of power and computational densities should take into account realistic area efficiencies.

© 2004 IBM Corporation

Optimization Approaches

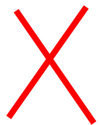
1. Engineering approach:

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).

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

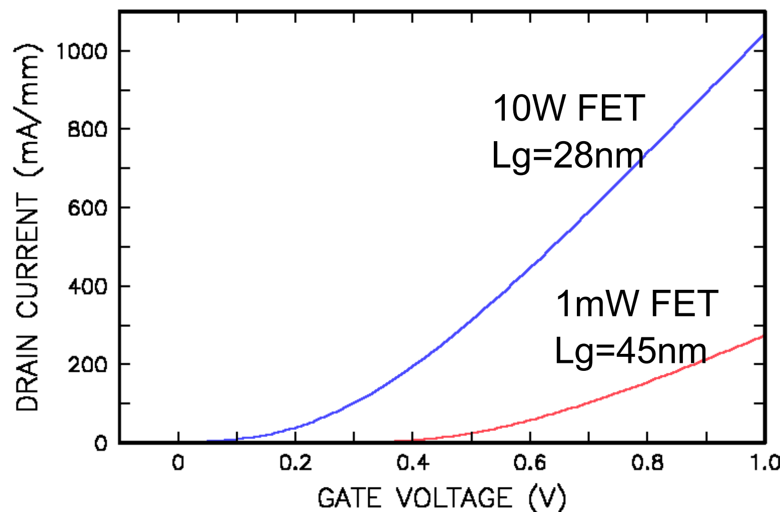
© 2004 IBM Corporation

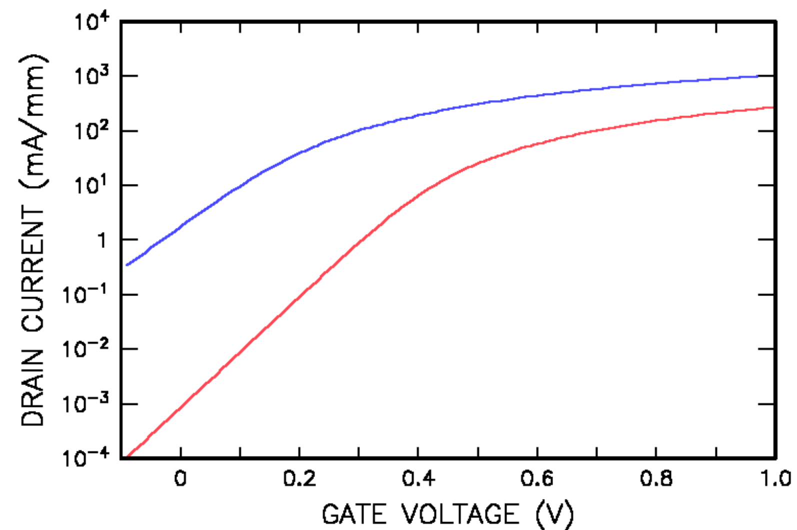
FET Model

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:

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





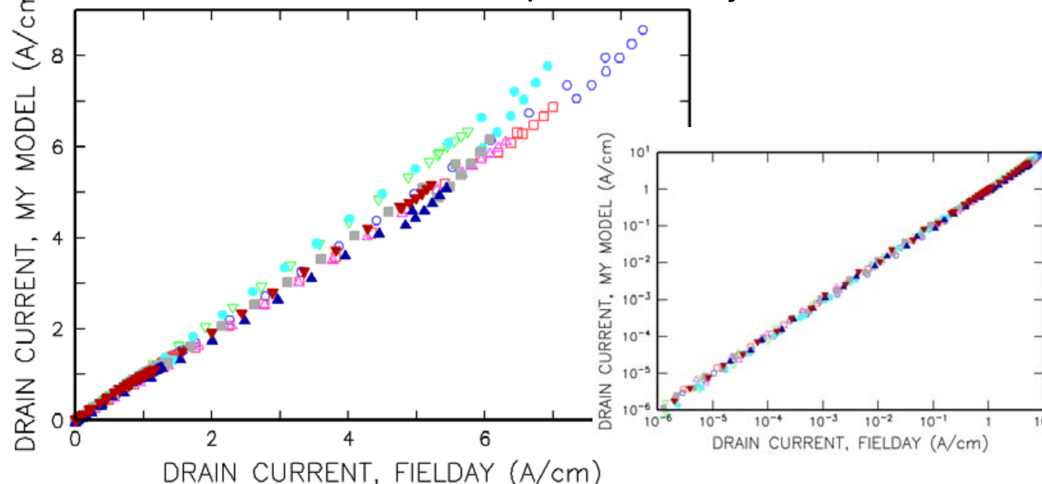
© 2004 IBM Corporation

Calibrating FET IV Model

- Optimizer FET model is fit to IV curves from 2D device simulator (FIELDAY).
- An empirical relation for the effective body doping vs channel length is used, which allows excellent fitting to the FIELDAY data.
- Best fits occur when x_{ovlp} is an optimization variable, allowing overlap to decrease with generation.
- Evaluated 10 parameter fits to 90nm and 45nm technology node IV's separately, and 14 parameter fit to the data jointly.

	Separate Fits		Joint Fit	
	90nm	45nm	90nm	45nm
ND	3.26E18	6.00E18	2.29E18	4.35E18
xe	50.7	36.9	60.6	45.8
xoverlap	5.69	0.29	9.66	-0.55
beta	1.45	1.49	1.465	=
gamma	0.321	0.132	0.243	=
n1	-0.399	-0.643	-0.755	=
n2	2.33	1.84	2.14	=
tan theta	1.375	1.20	1.275	=
mult factor	2.04	1.97	1.933	=
Rc	0.0126	0.0162	0.0127	0.0161
Chi2	2.54	3.09	9.63	-

Correlation plots for the joint fit:



$$L_{CH} = L_G - x_{ovlp}$$

$$N_{eff} = N_D \frac{\left(L_{CH}/x_e\right)^{n_1}}{1+\left(L_{CH}/x_e\right)^{n_2}}$$

$$I_D \sim \mu_{mult} (L_{CH})^{-\gamma} (V_G - V_T)^{\beta}$$

© 2004 IBM Corporation

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 corrections are based on Eble's thesis.

Correction for V_T/V_{DD}.

© 2004 IBM Corporation

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.

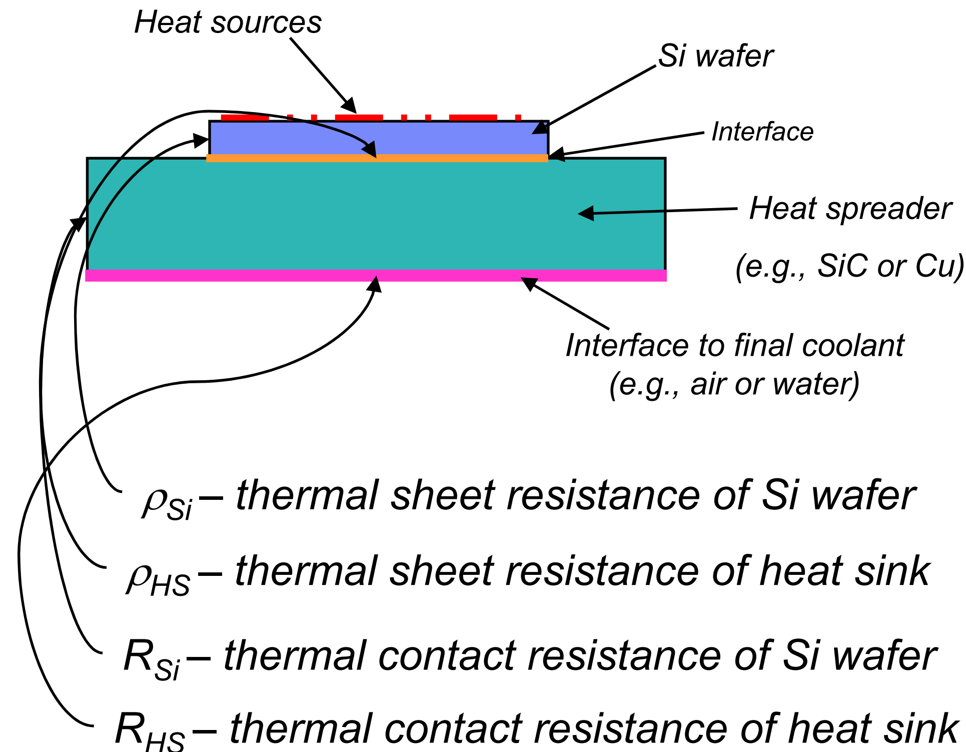
τ = mean delay for a single loaded logic gate

$\frac{\alpha}{\ell_D}$ is activity factor divided by logic depth. Usually ~0.015 in recent optimizations.

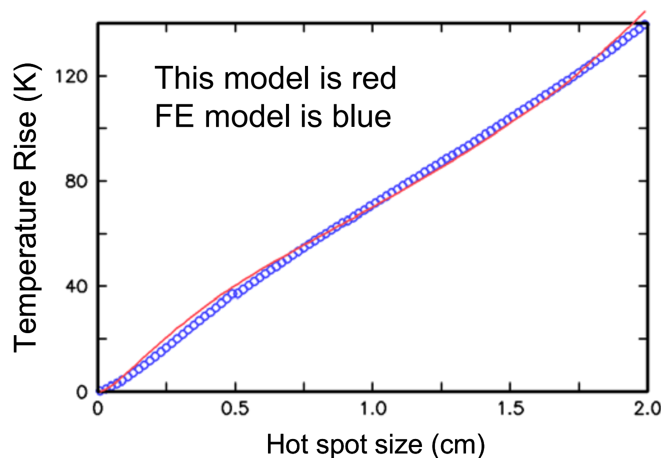
© 2004 IBM Corporation

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.**



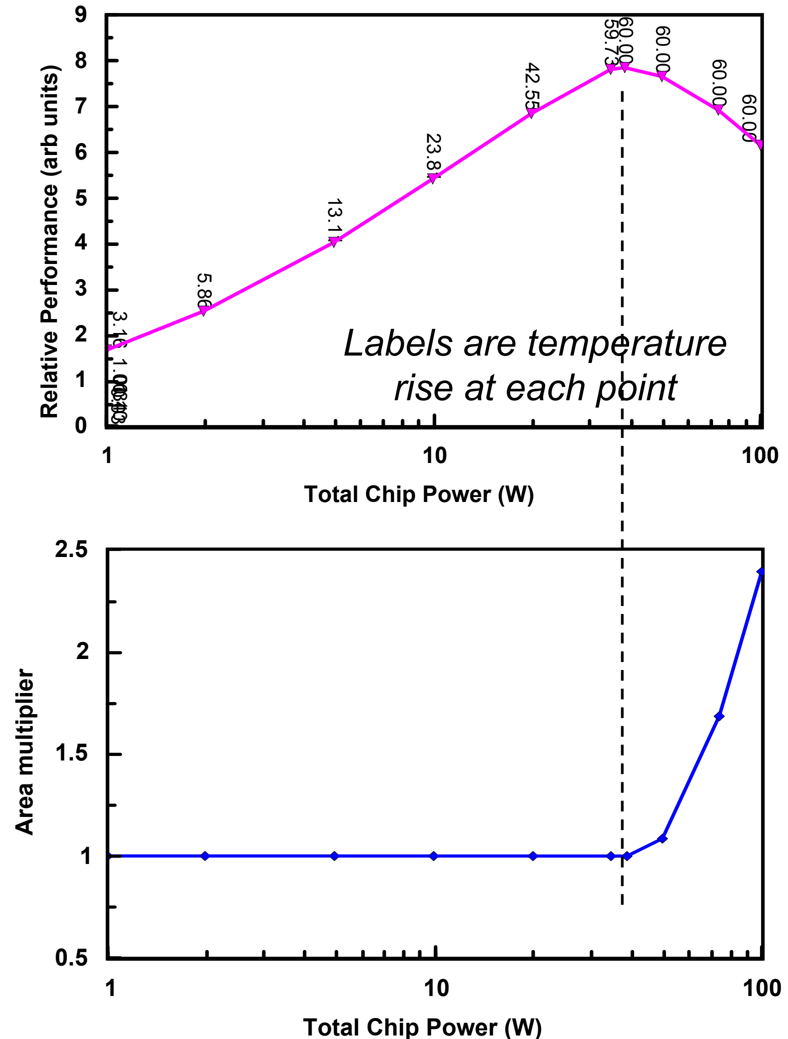
Comparison of simplified analytic model with detailed numerical model.



© 2004 IBM Corporation

Temperature rise constraint

- To prevent excessive heating, a constraint is introduced:
- If the power level would cause the maximum chip temperature to exceed the constraint value, the core area is increased above its nominal estimated size (e.g., by diluting the core with cache) until the temperature rise is just equal to the constraint.
- This makes longer average wire length, but prevents excess temperatures.



© 2004 IBM Corporation

Communication and Wiring Models

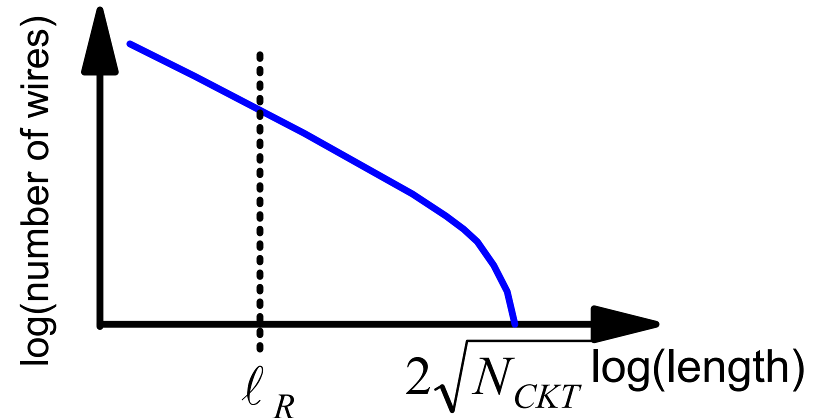
Assume wire lengths distributed according to Rent's rule.

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

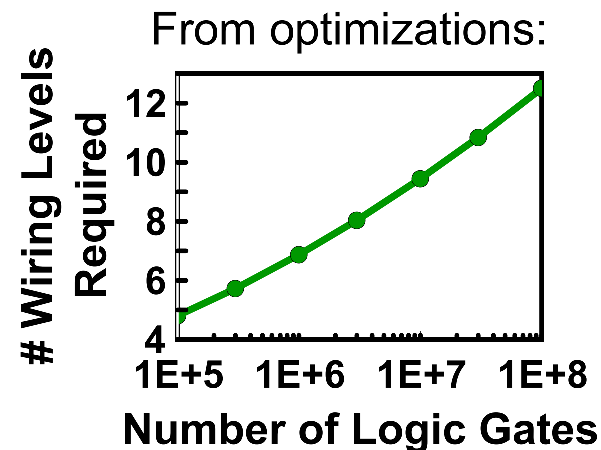
$$\langle L_{noRptr} \rangle = \frac{\int_1^{\ell_R} \ell i_{net}(\ell) d\ell}{\int_1^{\ell_R} i_{net}(\ell) d\ell}$$

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

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



From optimizations:



© 2004 IBM Corporation

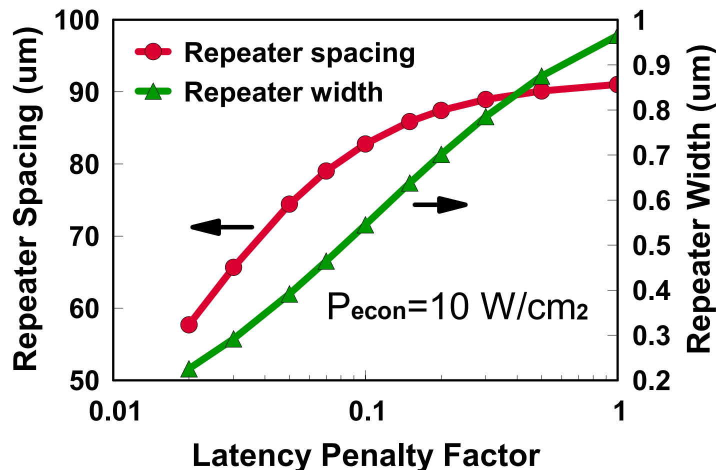
Repeater Model

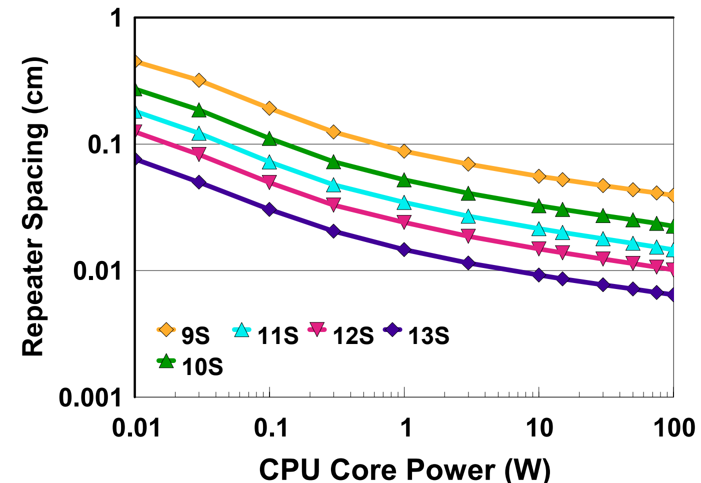
Long wires receive repeaters with a spacing that is optimized.

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

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

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





© 2004 IBM Corporation

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.**

© 2004 IBM Corporation

Optimization Results

- **General results**
- **Evaluating specific possible device directions**
  - Increasing mobility
  - High-k gate dielectric and metal gates
  - BEOL improvement only
  - Better heat sinks
  - Sub-ambient cooling
  - Multi-processor tradeoffs

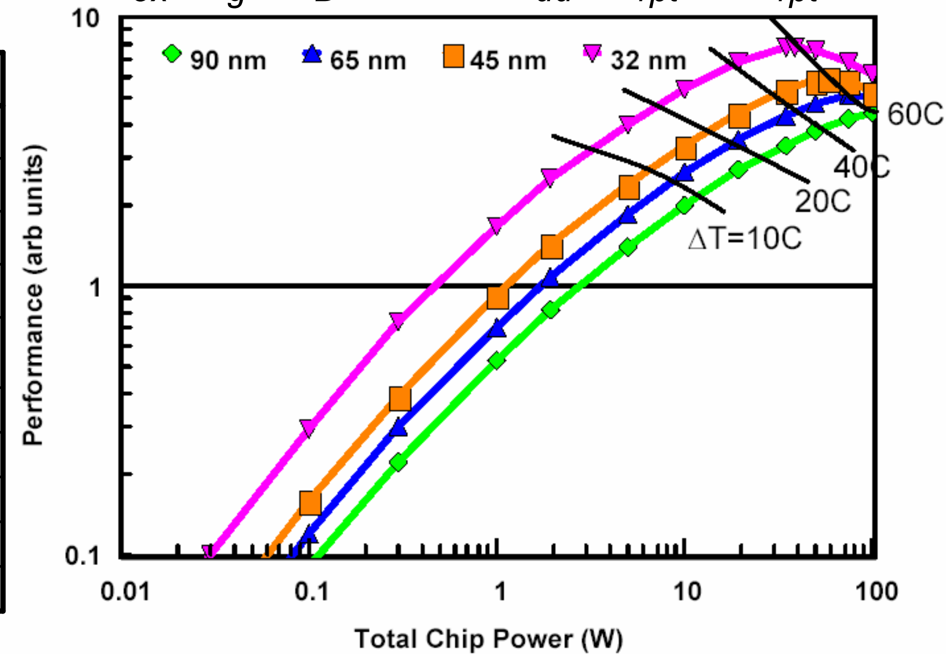
© 2004 IBM Corporation

Optimize by generation

Dual core processor with aggressive air cooling

Technology node (nm)	130	90	65	45	32
Wire 1/2 pitch (nm)	175	120	90	70	50
gate overlap (nm)	19	8.74	3	-1.03	-1.5
halo scale len (nm)	61.5	53.4	46.3	40.2	34.9
Rcs (Ohm cm)	0.0129	0.0136	0.0144	0.0152	0.0161
LER sigma Lg @W=1um (nm)	0.28	0.28	0.28	0.28	0.28
ACLV (nm)	3.9	2.7	1.8	1.3	1
mob. enhancement	1	1.4	1.7	2	2
gate depletion (nm)	0.4	0.4	0.3	0.3	0.01
k_BEOL	3.9	3.5	3.2	2.8	2.5
Gate insulator	oxynitride	oxynitride	oxynitride	oxynitride	HfO_2

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

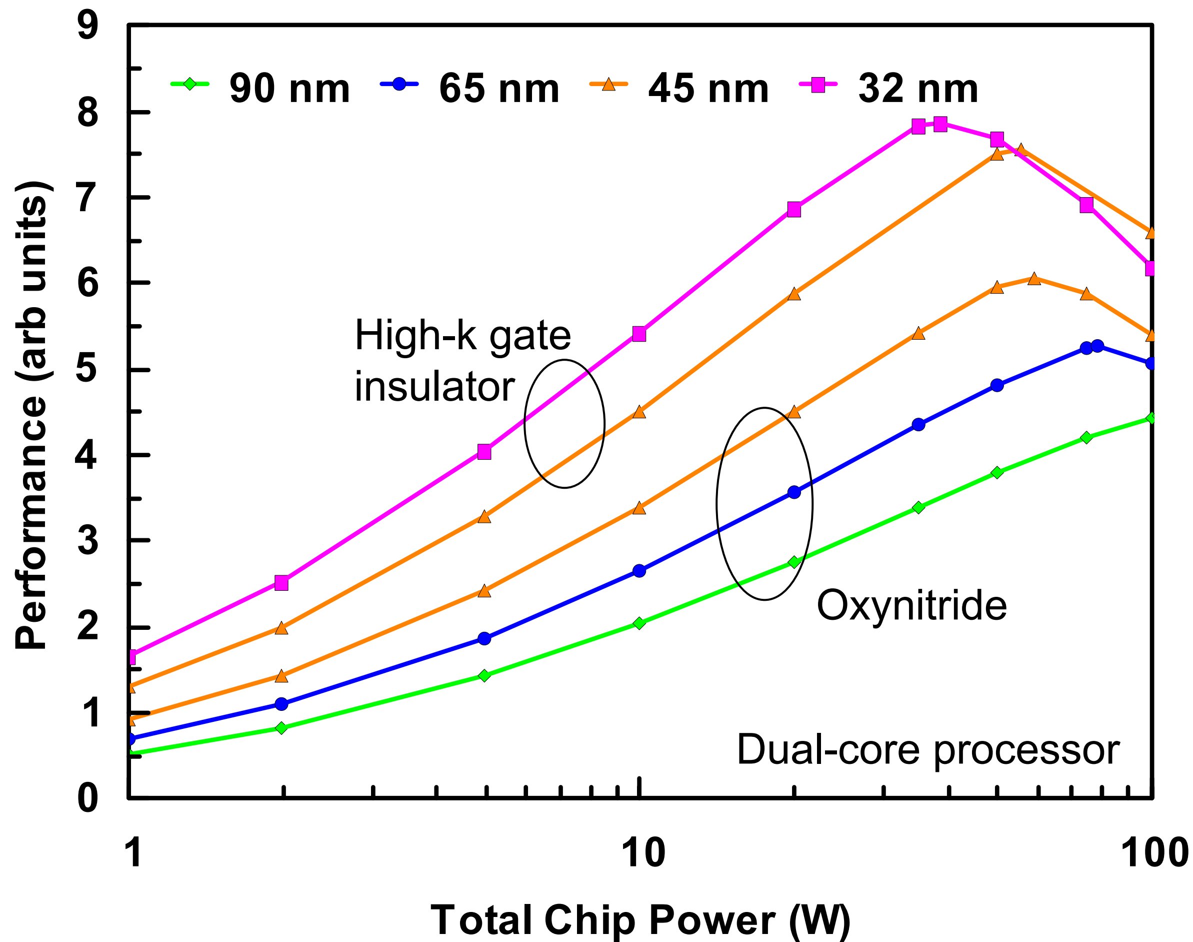


Note that the L_G, t_{ox}, V_{DD}, V_T, etc. are **NOT** preselected. They are solved for by the optimizations.

© 2004 IBM Corporation

Optimize by generation

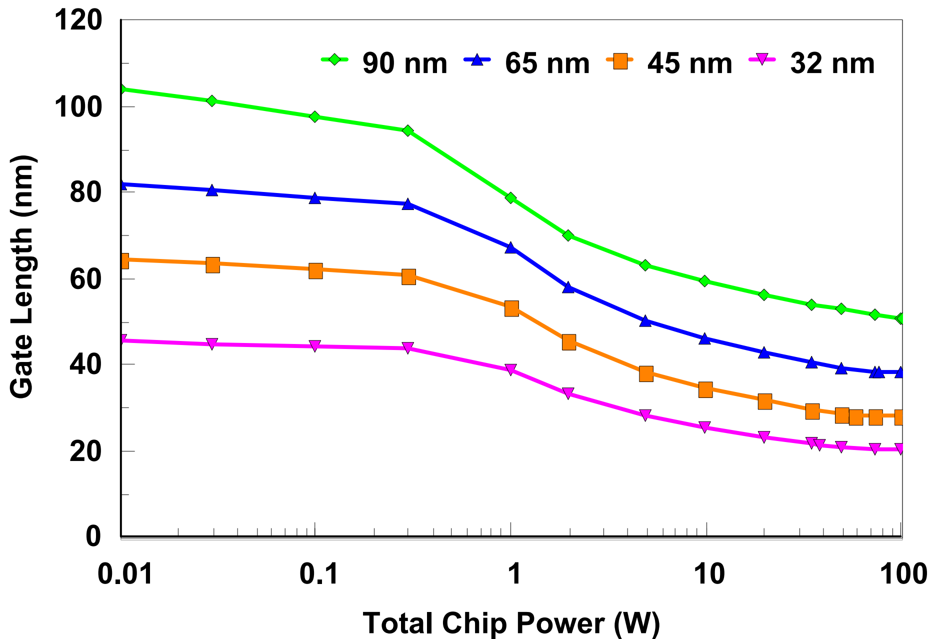
Dual core processor with aggressive air cooling

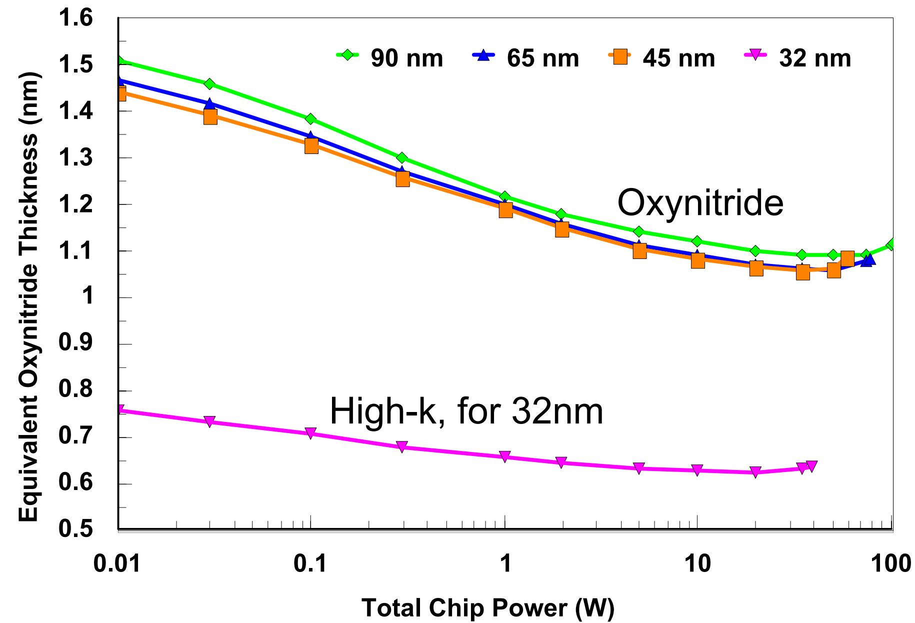


© 2004 IBM Corporation

Optimize by generation

Dual core processor with aggressive air cooling



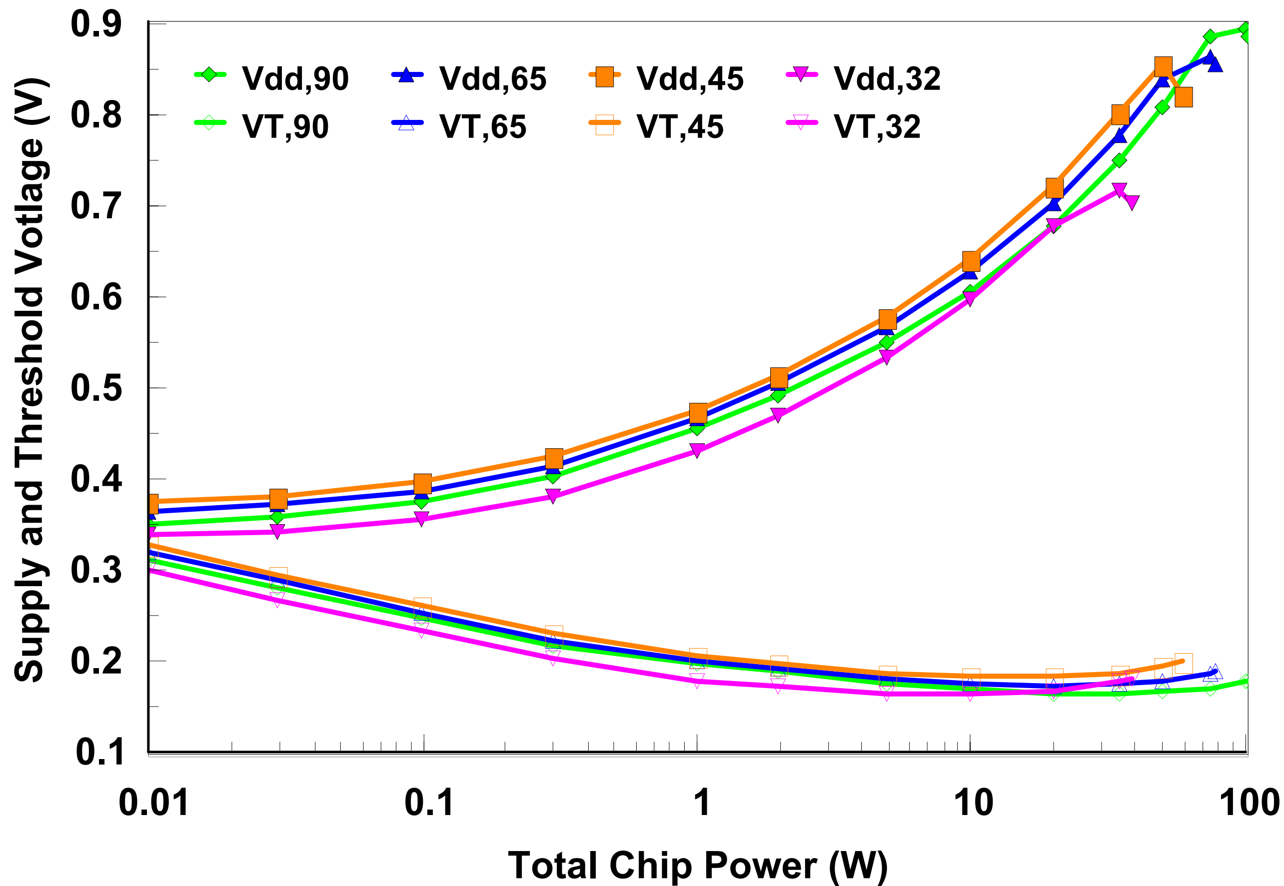


(high-k case assumes 0.3nm barrier layer, bandedge metal gate, HfO_2-like insulator characteristics.)

© 2004 IBM Corporation

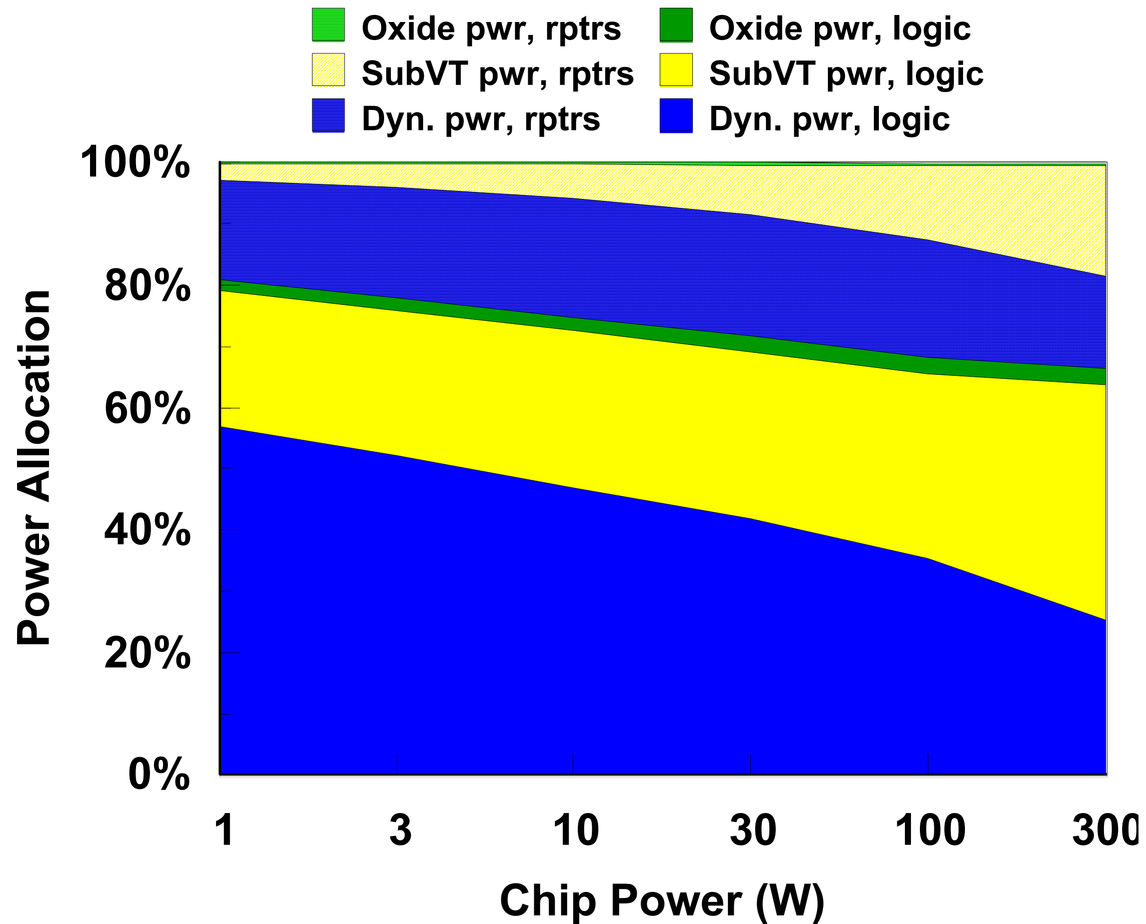
Optimize by generation

Dual core processor with aggressive air cooling



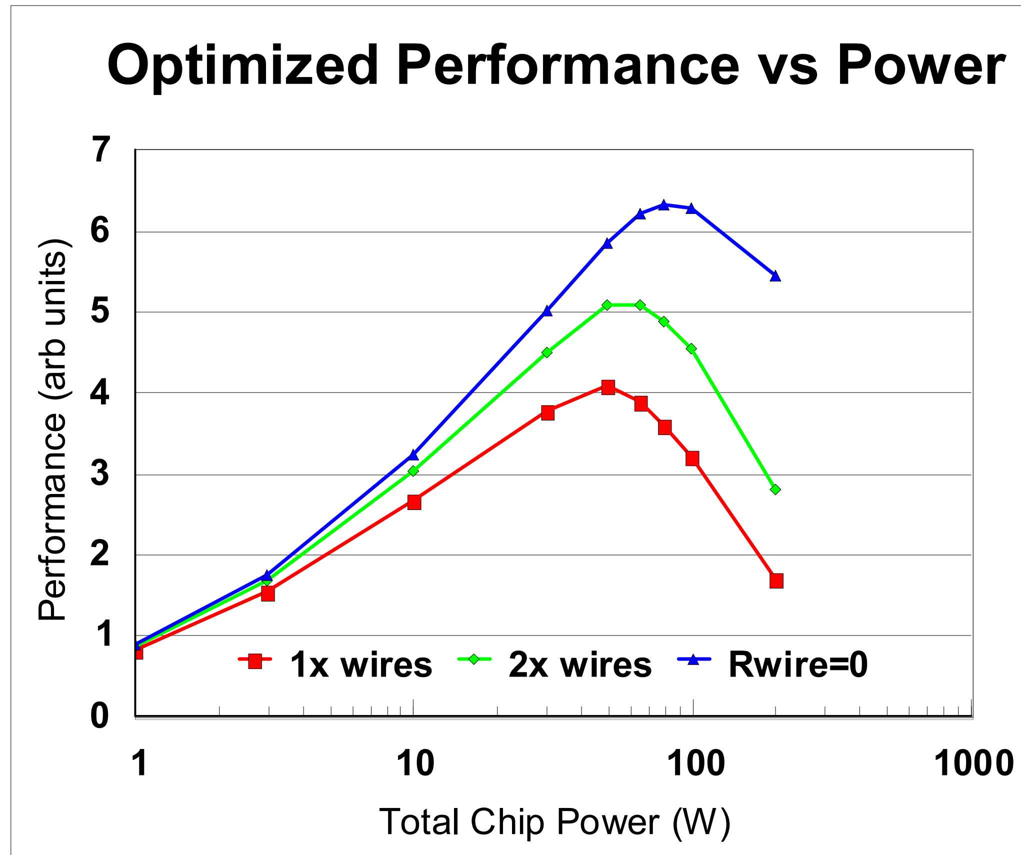
© 2004 IBM Corporation

Optimal Power Allocation Fractions



© 2004 IBM Corporation

Impact of long-wire assumptions



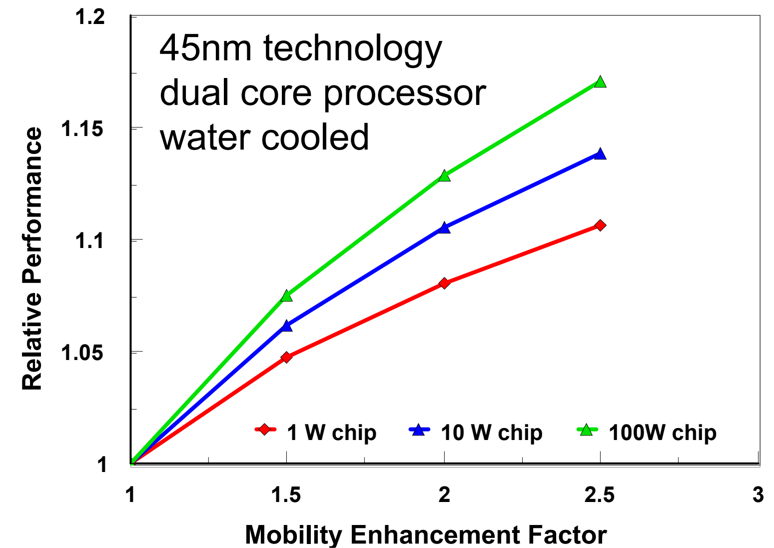
Green case: wires with repeaters are 2x the regular wire.
Red case: all wire is the same size (63.6 nm, here, for 45nm node)
Blue case: zero wire resistance case is for comparison.
(All wires are 2:1, height to width ratio)

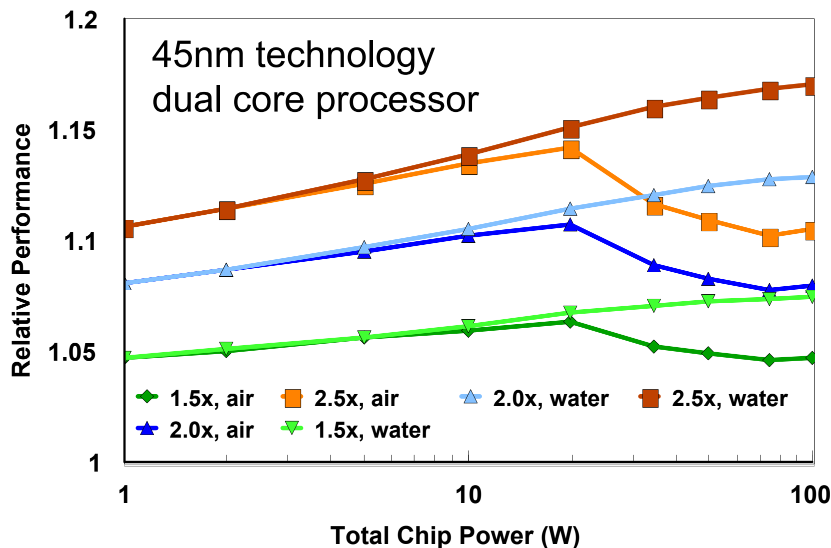
© 2004 IBM Corporation

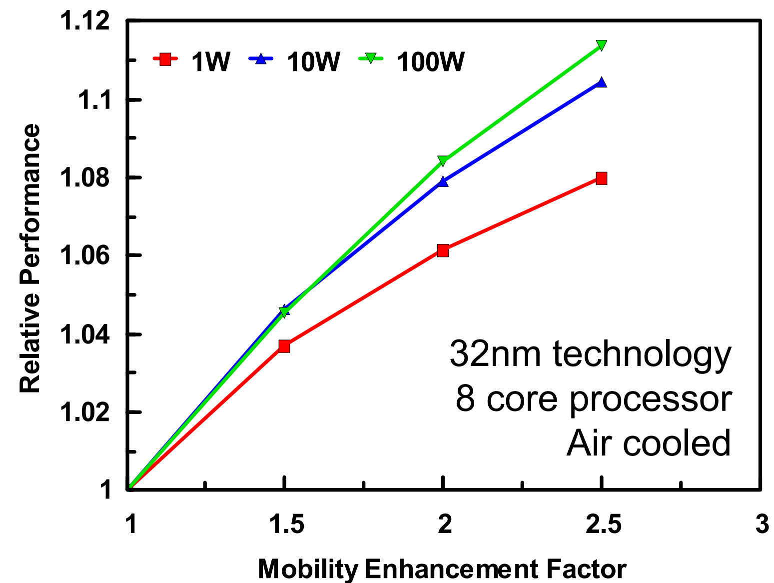
Mobility dependence

Enhanced mobility has greatest benefit at high power.

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



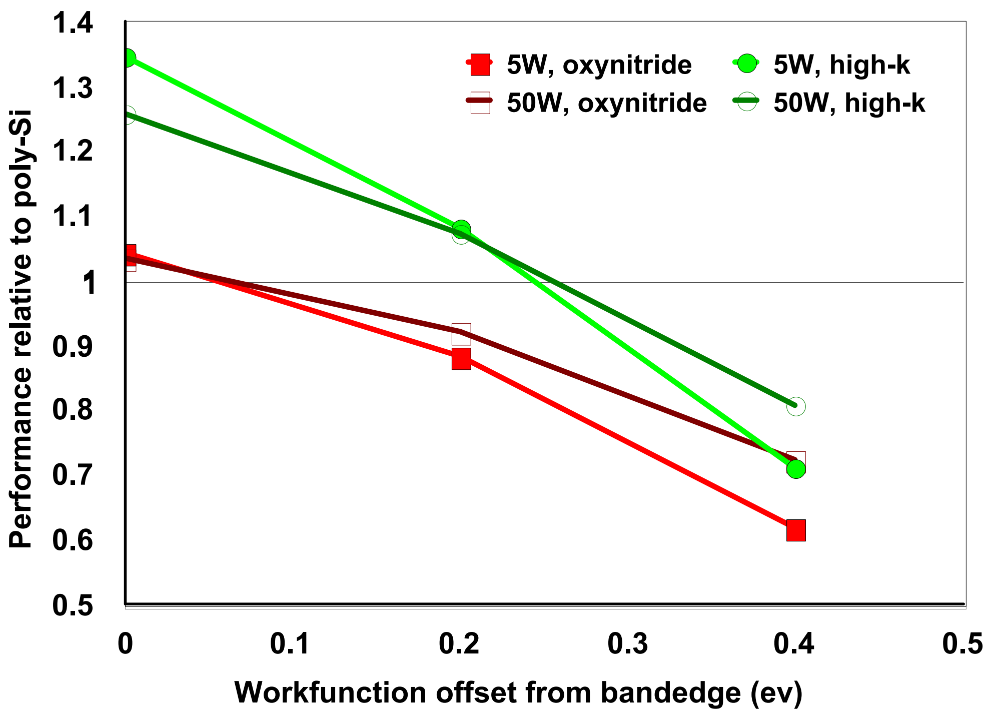


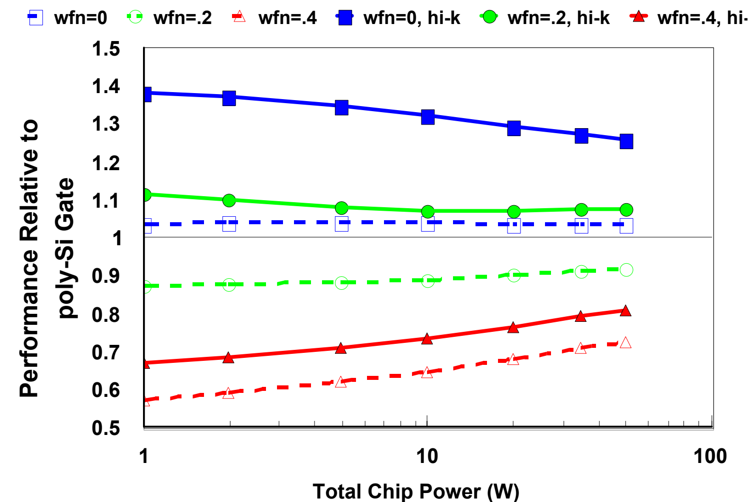


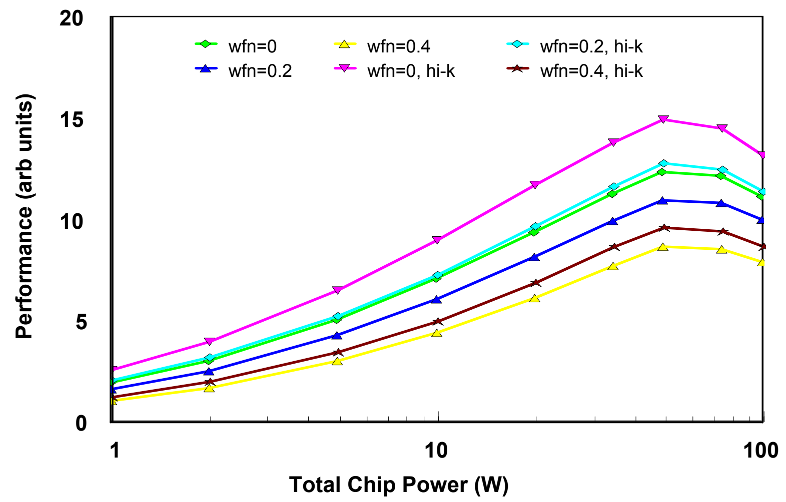
© 2004 IBM Corporation

Metal-gate workfunction for high-k and oxynitride

45nm node, dual core processor with aggressive air cooling



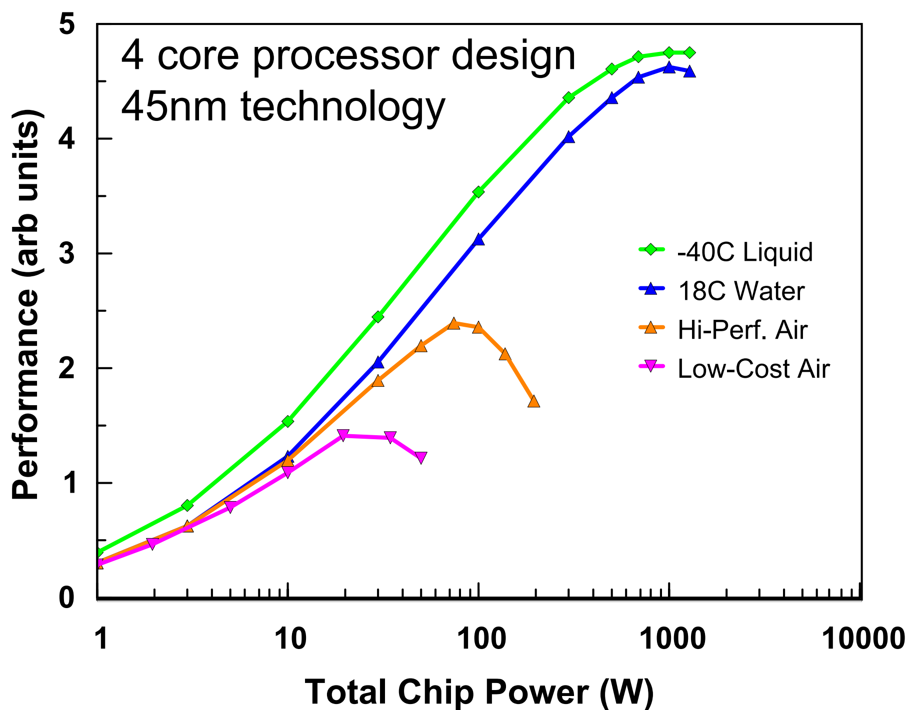


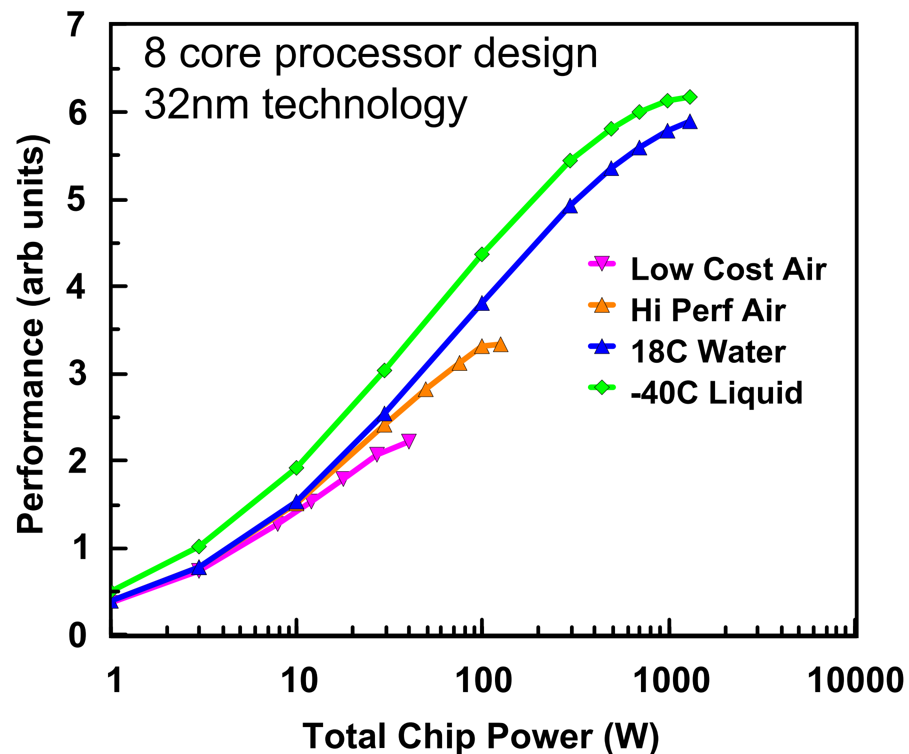


© 2004 IBM Corporation

Cooling scenario optimizations

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



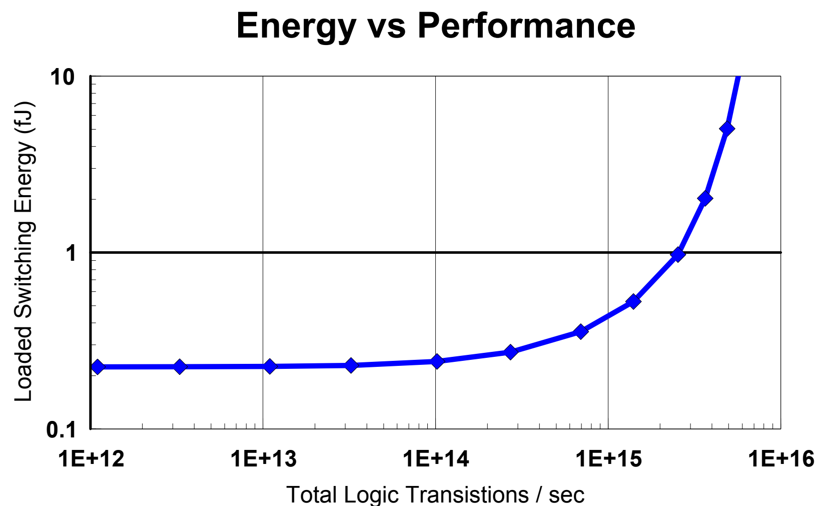


© 2004 IBM Corporation

Multi-processor trade-offs

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.



These results are for 4-processor chips with micro-channel water cooling, pulling out all the stops.

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

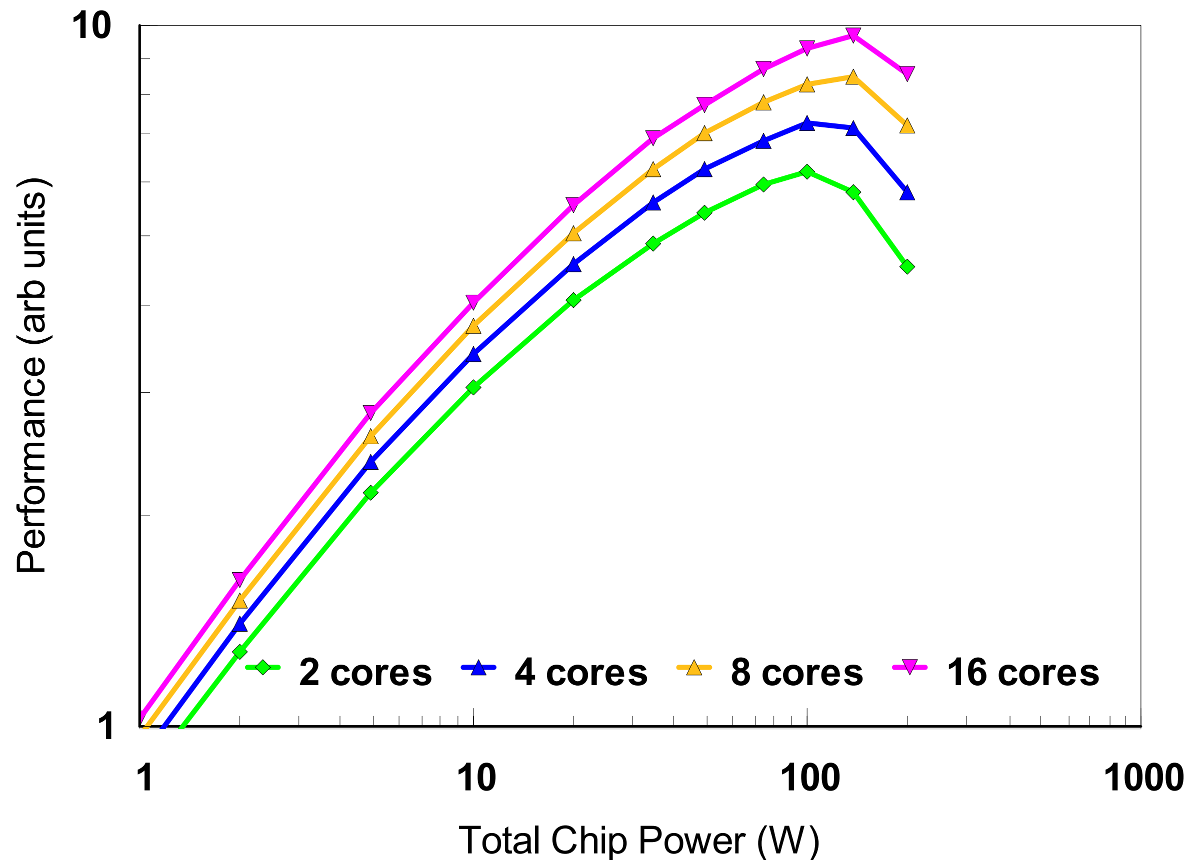
© 2004 IBM Corporation

Optimizations for varying number of processor cores

32nm node optimizations

Aggressive air cooling

Assume: fixed total number of FETs, divided into varying # of cores.



Optimized over 7 variables:

L_g, t_{ox}, N_d, <w>, D_{rptr}, $<w_{rptr}>$, V_{dd}.

© 2004 IBM Corporation

3. What is the best possible?

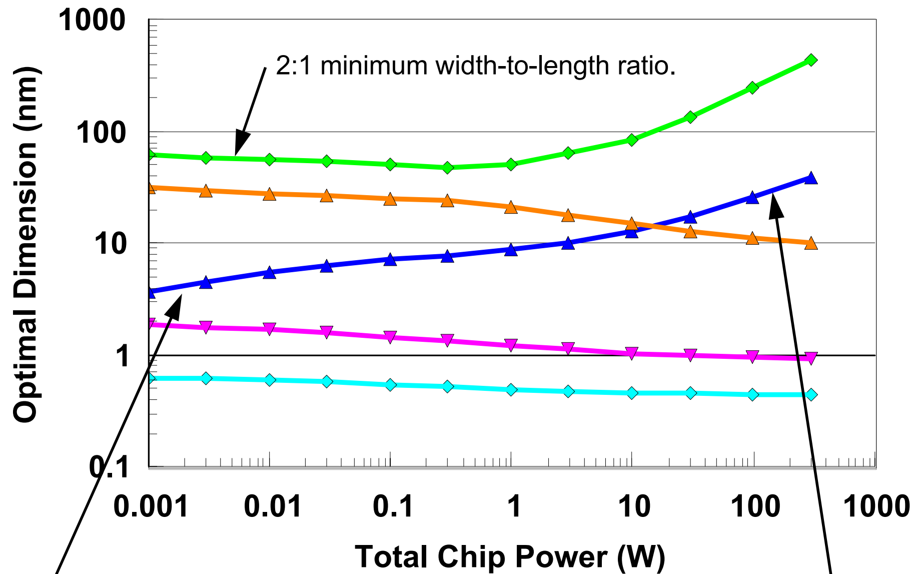
Optimizations across all generations

These optimizations are for hypothetical 4-processor chips with micro-channel water cooling, pulling out all the stops.

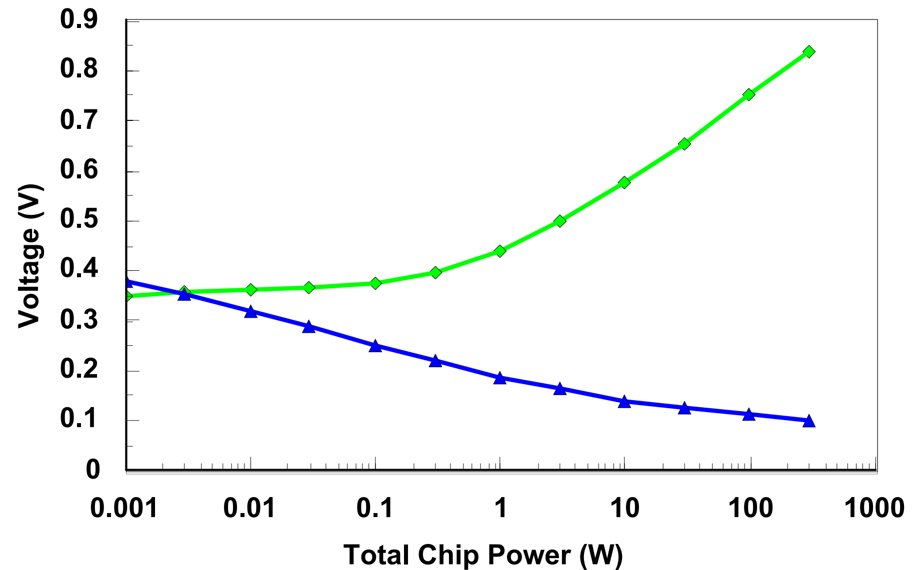
9 variables: t_{ox}, L_g, N_D, $\langle w \rangle$, V_{dd}, S_{rpt}, $\langle w_{rpt} \rangle$, $\mathbf{w_{HP}}$, $\mathbf{x_{halo}}$

Caveats: conventional MOSFET structure, high-performance design practices

Mean FET width · Gate Length · Equiv. oxide thickness · Wire Half-pitch · Hi-k thickness



Supply Voltage · Threshold Voltage (VTsat)



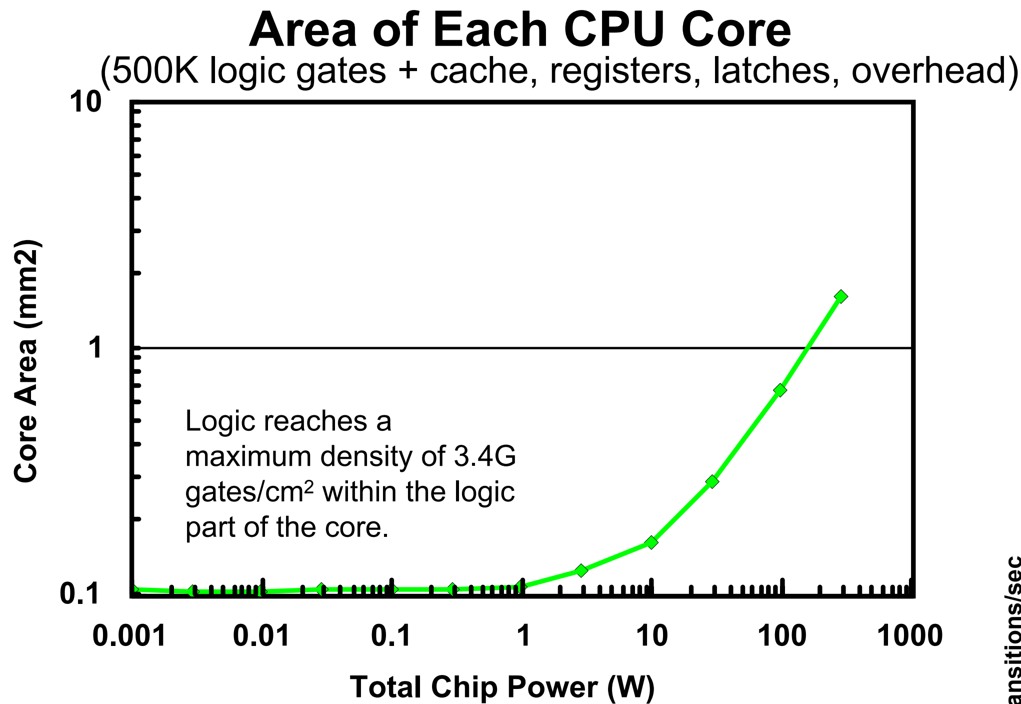
Wire becomes VERY small at lowest power because wire resistance has little impact on the slow speeds.

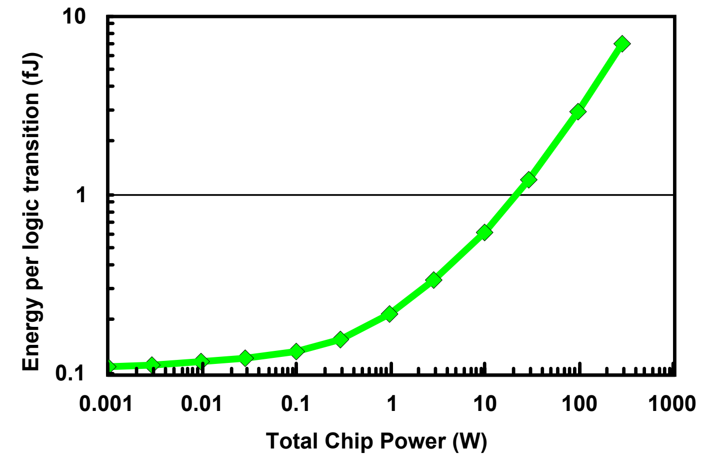
Optimal wire pitch grows for highest performance cases. Gives lower resistance, and the FETs are spreading out because of their width, anyway.

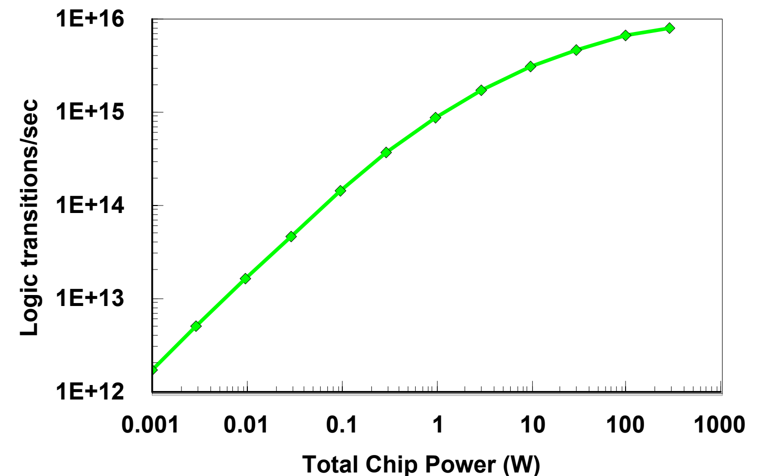
© 2004 IBM Corporation

Continued Optimizations across all generations

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





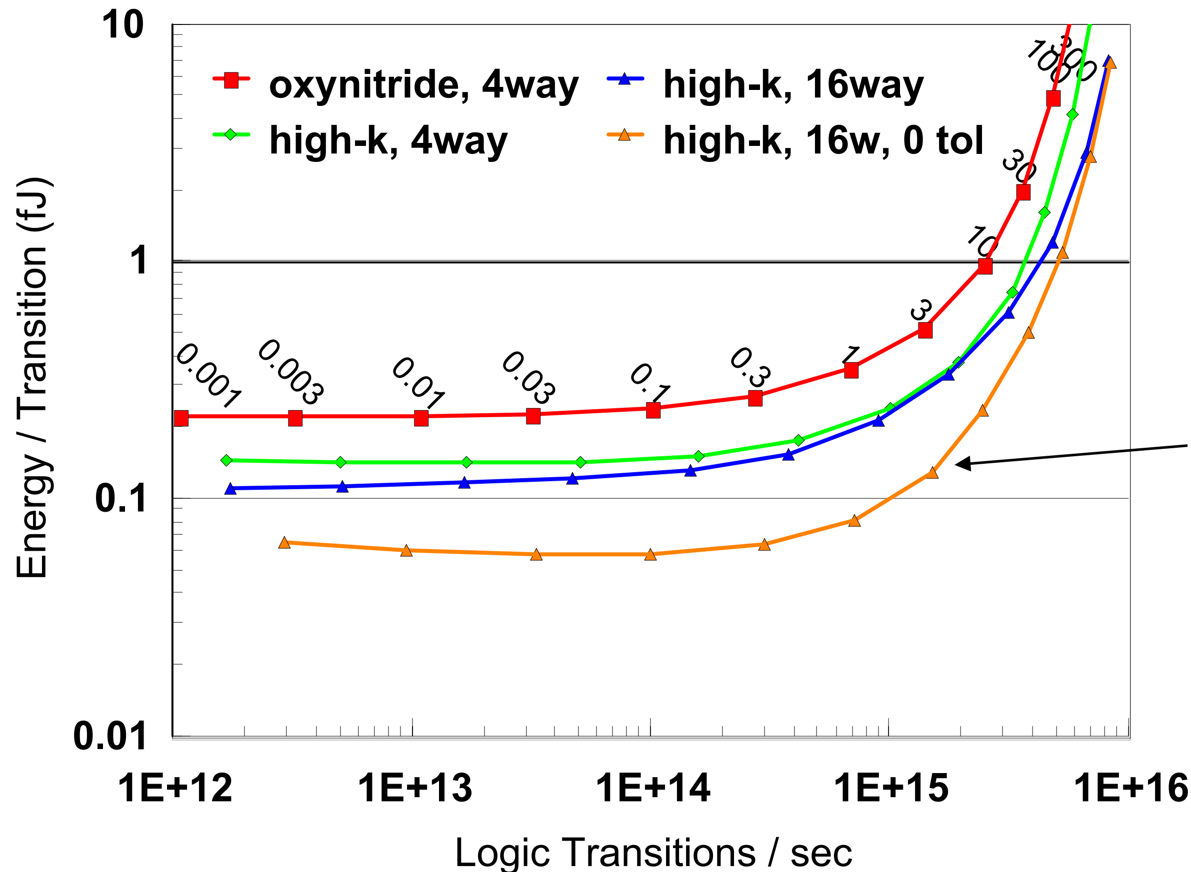


Power constraints limit conventional CMOS scaling to ~3.4G logic gates/cm^2.

The challenge for nanotechnology is to find a way to do significantly better.

© 2004 IBM Corporation

Energy vs Performance



Numbers on each point are the total chip power for that point.

Zero process tolerances is unrealistic, but serves as a lower bound. Gate lengths can be 30% smaller, yielding higher density and shorter wires.

Average loaded switching energy versus performance for cross-generational optimization (9 parameters).

© 2004 IBM Corporation

Minimum Energy breakdown

- **Average logic load capacitance: 0.17fF gate cap, 0.06fF parasitic, 0.40fF wire (3.2um average length)**
- **Minimum supply voltage: ~ 360mV.**
- **Raw logic switching energy: $\frac{1}{2}CV^2$ = 0.04fJ**
- **Ratio of total logic power to active power: ~1.4**
- **Ratio of logic + repeater power to logic power: ~ 1.2**
- **Long wire latency penalty factor: ~1.6**

- **All together: effective energy per 'useful' logic transition: ~0.1fJ**

© 2004 IBM Corporation

4. Open questions

- **Many aspects of VLSI design are tied to the importance or 'criticality' of a net. How can the distribution of 'importance' be modeled? Can this be tied to a distribution of electrical and/or logical effort?**
  - More accurate FET width distribution
  - Multiple VT optimization
  - Wiring hierarchy optimization
- **How should SRAM optimization be tied to logic, if at all?**
- **How to optimize further up the design hierarchy into architecture?**

© 2004 IBM Corporation

5. Summary

- General limitations to scaling have been summarized.
- Power and temperature rise are dominant limiters.
- A set of simplified models have been developed to enable fast turnaround comparative technology optimizations in the presence of power and temperature constraints.
- The dependence of optimal technology parameters on application power requirements has been investigated.
- The dependence of chip performance on potential technology enhancements has also been investigated.
- Performance gains can still be obtained from improved cooling and/or from lower power, slower, more parallel processors.
- Minimum loaded switching energy for conventional CMOS is ~0.1 fJ.
- Open questions are still under investigation.

© 2004 IBM Corporation

The End of Scaling is Optimization

Stop when you get to the top.

Then, try to switch to a different mountain, e.g., some form of nanotechnology.

But, each technology has its own summit, and we need to try to make sure the new peak is actually higher than the one we are already on.

© 2004 IBM Corporation