SAND2008-2662P

Issues in the Future of Computing

Erik P. DeBenedictis
Sandia National Labs

Presented April 24, 2008
New Mexico State University
Las Cruces, NM

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the
United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Introduction

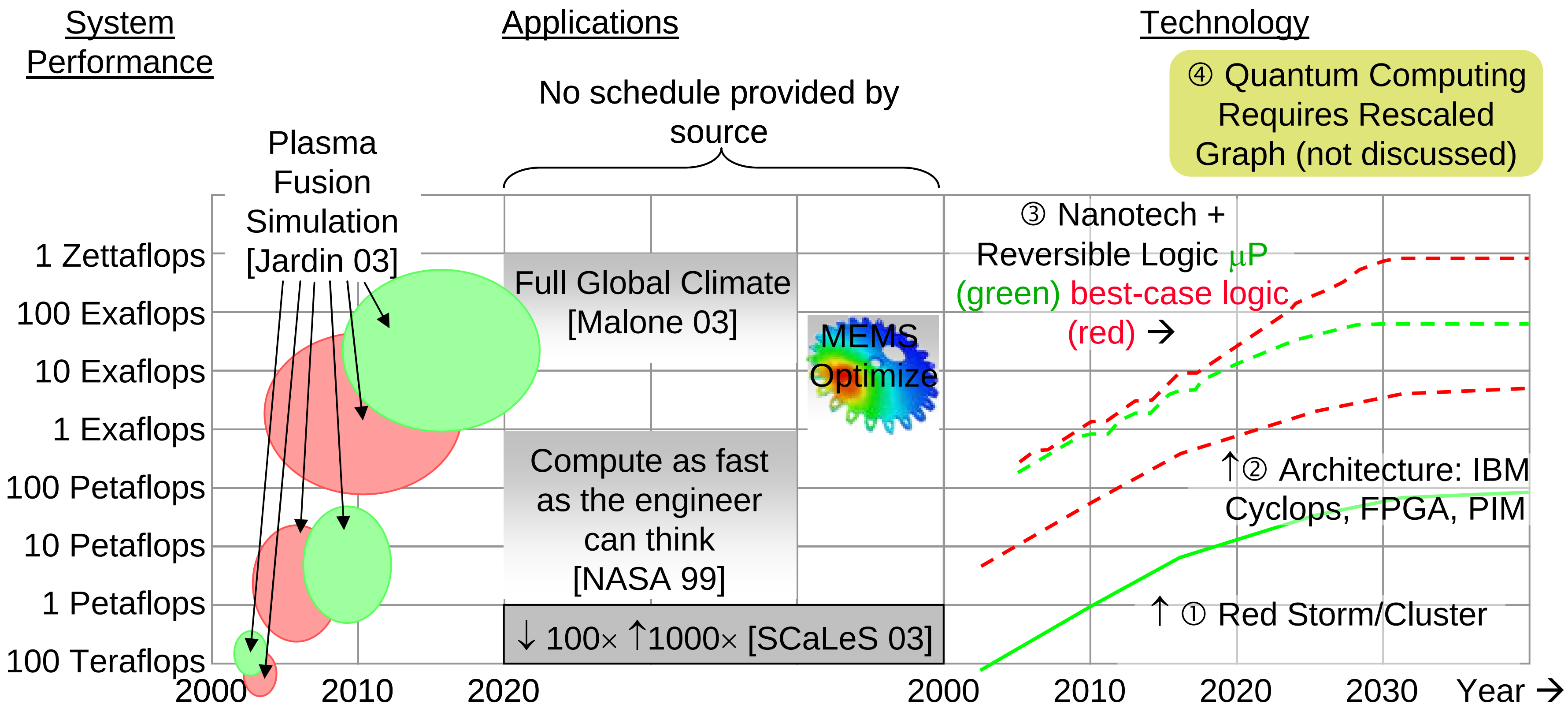
- Since the early 1980s, μPs doubled in performance every 18 months
- The computer industry was predictable, if uninteresting
- The computer industry could replace products every few years with a faster model, to great profit
- A paradigm shift became apparent about 6 months ago
  - The flat lining of clock rate and multi-core architectures preceded the paradigm shift, but it took a couple years for users to realize the impact
- I will try to sort this out in this talk

Outline

- Why Strive for Zettaflops?
  - Global Warming Mission
- The CMOS Roadmap
- Industry’s Responses 1 & 2 to Maturing CMOS
- Limits to Computing and Avoidance
- Industry’s Responses 3 to Maturing CMOS
- Conclusions

Sandia
National
Laboratories

Applications and $100M Supercomputers



System Performance

Applications

Technology

④ Quantum Computing Requires Rescaled Graph (not discussed)

No schedule provided by source

Plasma Fusion Simulation [Jardin 03]

Full Global Climate [Malone 03]

MEMS Optimize

Compute as fast as the engineer can think [NASA 99]

↓ 100× ↑1000× [SCaLeS 03]

③ Nanotech + Reversible Logic µP (green) best-case logic (red) →

↑② Architecture: IBM Cyclops, FPGA, PIM

↑ ① Red Storm/Cluster

[Jardin 03] S.C. Jardin, "Plasma Science Contribution to the SCaLeS Report," Princeton Plasma Physics Laboratory, PPPL-3879 UC-70, available on Internet.
[Malone 03] Robert C. Malone, John B. Drake, Philip W. Jones, Douglas A. Rotman, "High-End Computing in Climate Modeling," contribution to SCaLeS report.
[NASA 99] R. T. Biedron, P. Mehrotra, M. L. Nelson, F. S. Preston, J. J. Rehder, J. L. Rogers, D. H. Rudy, J. Sobieski, and O. O. Storaasli, "Compute as Fast as the Engineers Can Think!" NASA/TM-1999-209715, available on Internet.
[SCaLeS 03] Workshop on the Science Case for Large-scale Simulation, June 24-25, proceedings on Internet a http://www.pnl.gov/scales/.
[DeBenedictis 04], Erik P. DeBenedictis, "Matching Supercomputing to Progress in Science," July 2004. Presentation at Lawrence Berkeley National Laboratory, also published as Sandia National Laboratories SAND report SAND2004-3333P. Sandia technical reports are available by going to http://www.sandia.gov and accessing the technical library.

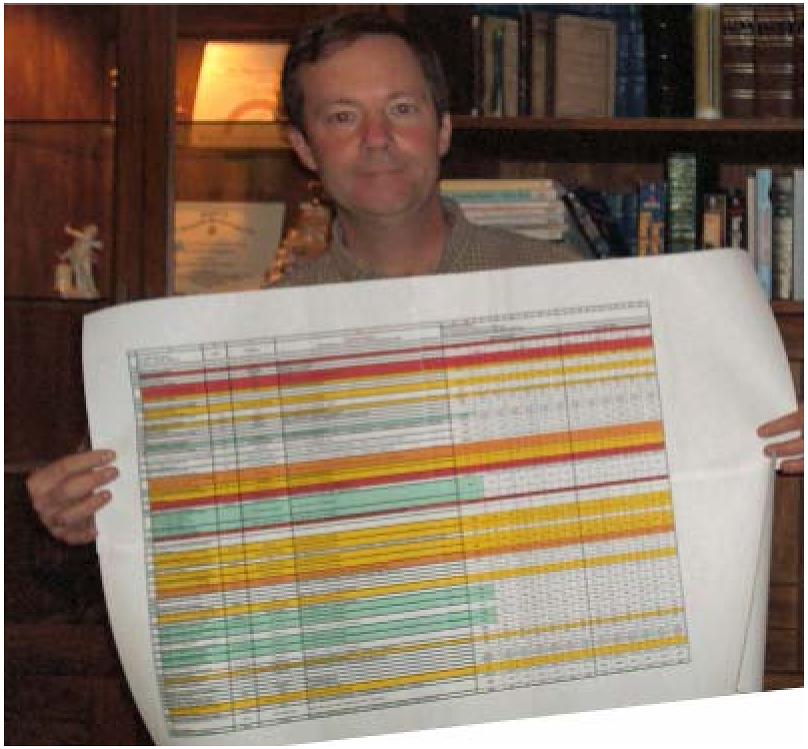
Outline

- Why Strive for Zettaflops?
  - Global Warming Mission
- The CMOS Roadmap
- Industry’s Responses 1 & 2 to Maturing CMOS
- Limits to Computing and Avoidance
- Industry’s Responses 3 to Maturing CMOS
- Conclusions

Sandia
National
Laboratories

ITRS Process Integration Spreadsheet

- Big Spreadsheet
  - Columns are years
  - Rows are 100+ transistor parameters
  - Manual entry of process parameters by year
  - Excel computes operating parameters
  - Extra degrees of freedom go to making Moore’s Law smooth – not the best computers

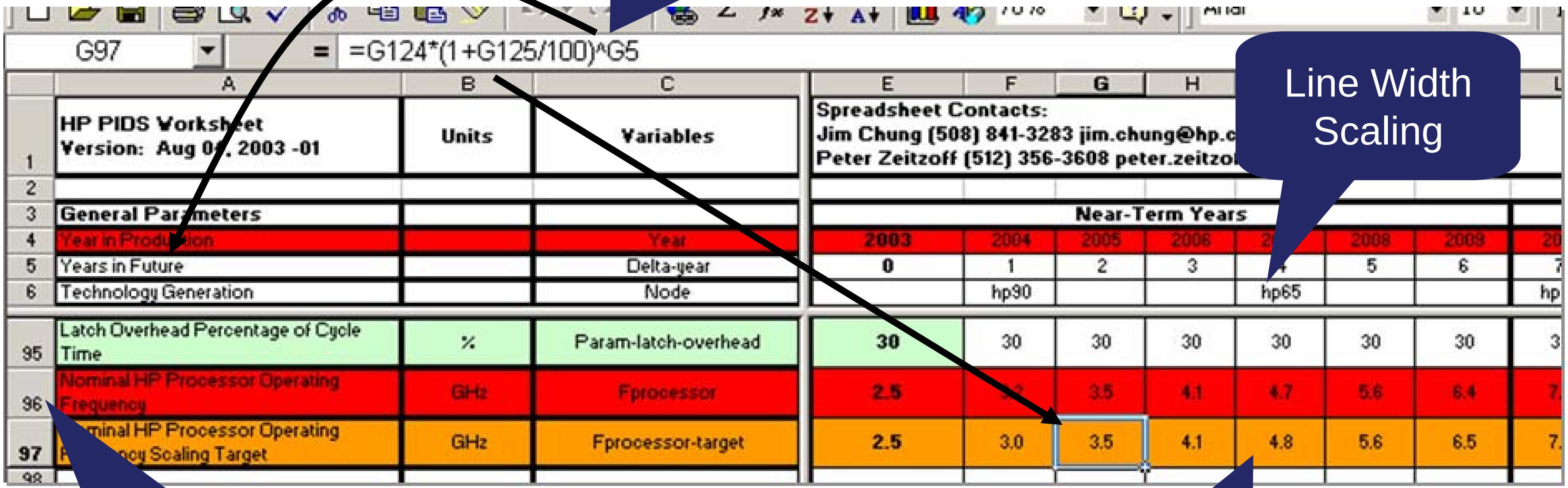


Clock Rate Flat Lined

- Clock rate flat lined a couple years ago, as vendors put excess resources into multiple cores
- This is a historical fact and evident to everybody, so there is little reason to comment on the cause
- However, it has profound architectural consequences (later slide)

ITRS Spreadsheet Structure

Target is exponential in “Years in Future”



G97 = =G124*(1+G125/100)^G5

	A	B	C	E	F	G	H				L
1	HP PIDS Worksheet Version: Aug 04, 2003 -01	Units	Variables	Spreadsheet Contacts: Jim Chung (508) 841-3283 jim.chung@hp.c Peter Zeitzoff (512) 356-3608 peter.zeitzo							
2											
3	General Parameters						Near-Term Years				
4	Year in Production		Year	2003	2004	2005	2006	20	2008	2009	20
5	Years in Future		Delta-year	0	1	2	3	4	5	6	7
6	Technology Generation		Node		hp90			hp65			hp
95	Latch Overhead Percentage of Cycle Time	%	Param-latch-overhead	30	30	30	30	30	30	30	3
96	Nominal HP Processor Operating Frequency	GHz	Fprocessor	2.5	3.2	3.5	4.1	4.7	5.6	6.4	7.
97	Nominal HP Processor Operating Frequency Scaling Target	GHz	Fprocessor-target	2.5	3.0	3.5	4.1	4.8	5.6	6.5	7.
98											

Line Width Scaling

Fprocessor is result of 96 rows of targets, inputs, and iterative calculation

Result usually matches to one decimal place!

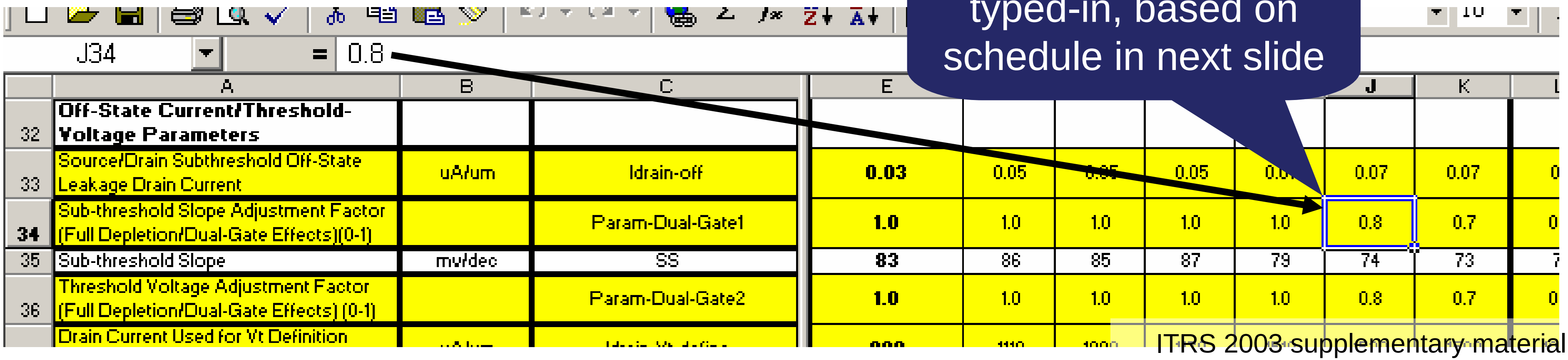
ITRS 2003 supplementary material

Do Demo

- Do a demo here of the actual ITRS spreadsheet
  - Illustrate how some parameters are Excel-generated exponentials
  - Other parameters are input by panels of experts based on schedules for technology innovations (high K dielectrics)
  - Other parameters are computed
  - Parameters are hand tweaked to make the curve look smooth
- Performance model is 10 gate delays with 30% latch overhead (no wire)

User Inputs

- Some factors will scale exponentially by definition, yet others will scale based on projections of engineers
- Supply voltage, doping levels, layer thicknesses, <u>leakage</u>, <u>geometry</u>, mobility, parasitic capacitance



These values are typed-in, based on schedule in next slide

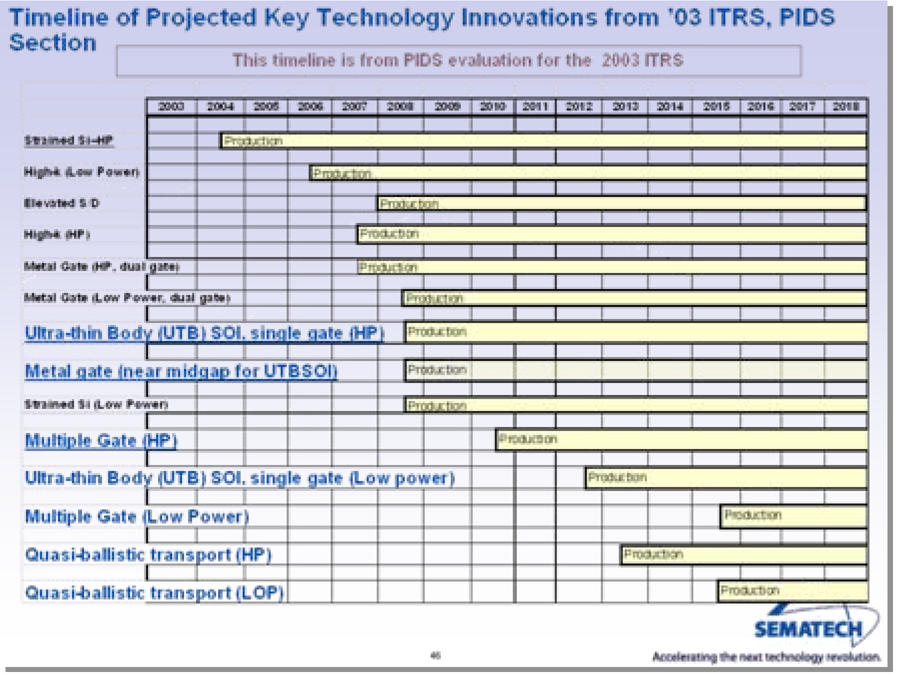
J34 = 0.8

	A	B	C	E					J	K	L
32	**Off-State Current/Threshold-Voltage Parameters**										
33	Source/Drain Subthreshold Off-State Leakage Drain Current	uA/um	Idrain-off	**0.03**	0.05	0.05	0.05	0.0[illegible]	0.07	0.07	0
34	Sub-threshold Slope Adjustment Factor (Full Depletion/Dual-Gate Effects)(0-1)		Param-Dual-Gate1	**1.0**	1.0	1.0	1.0	1.0	0.8	0.7	0
35	Sub-threshold Slope	mv/dec	SS	**83**	86	85	87	79	74	73	7
36	Threshold Voltage Adjustment Factor (Full Depletion/Dual-Gate Effects) (0-1)		Param-Dual-Gate2	**1.0**	1.0	1.0	1.0	1.0	0.8	0.7	0
	Drain Current Used for Vt Definition	uA/um	Idrain-Vt-define	[illegible]	1110	1000					

ITRS 2003 supplementary material

Schedule of Innovations

- **To make the calculations fit the projection of a smooth “Moore’s Law,” certain variables must be adjustable**
- **The independent variables are a “schedule of innovations,” or technology advances that must enter production on certain years**



Timeline of Projected Key Technology Innovations from ’03 ITRS, PIDS Section

This timeline is from PIDS evaluation for the 2003 ITRS

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Strained Si-HP		Production														
High-k (Low Power)				Production												
Elevated S/D						Production										
High-k (HP)					Production											
Metal Gate (HP, dual gate)					Production											
Metal Gate (Low Power, dual gate)						Production										
Ultra-thin Body (UTB) SOI, single gate (HP)						Production										
Metal gate (near midgap for UTBSOI)						Production										
Strained Si (Low Power)						Production										
Multiple Gate (HP)								Production								
Ultra-thin Body (UTB) SOI, single gate (Low power)										Production						
Multiple Gate (Low Power)													Production			
Quasi-ballistic transport (HP)											Production					
Quasi-ballistic transport (LOP)													Production			

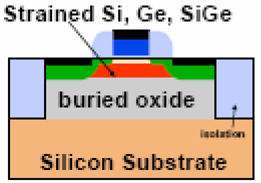
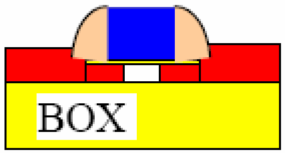
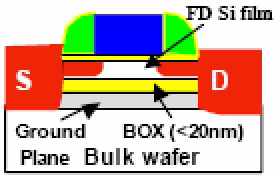
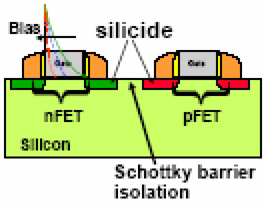
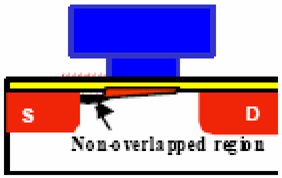
SEMATECH

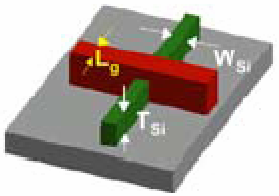
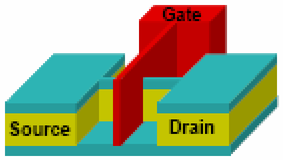
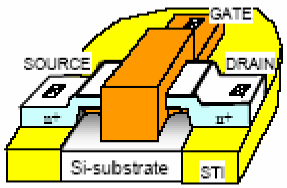
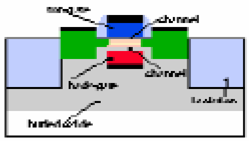
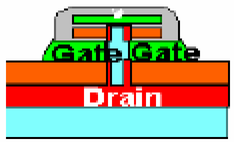
46

Accelerating the next technology revolution.

MOSFET Scaling Trends, Challenges, and Key Technology Innovations through the End of the Roadmap, Peter M. Zeitzoff

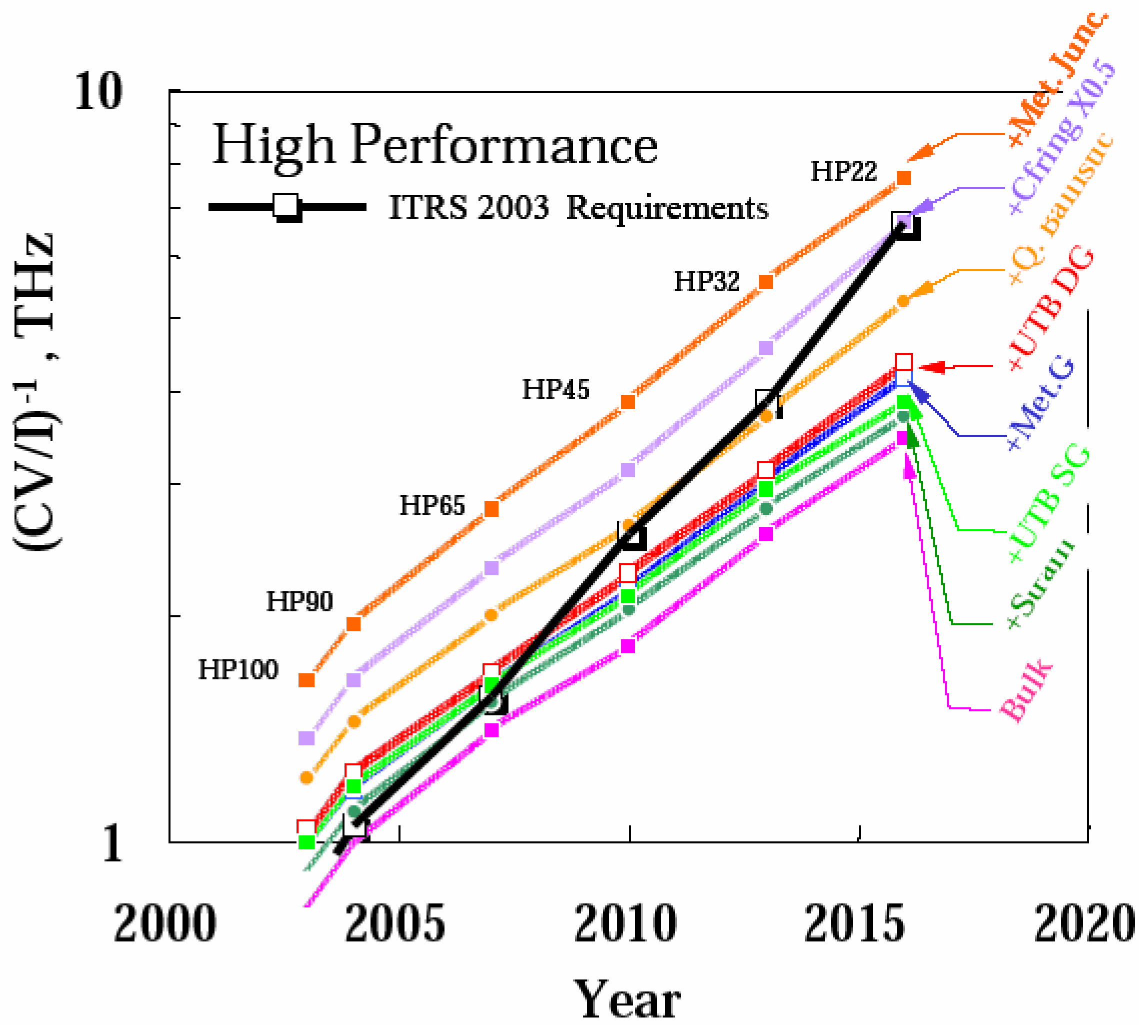
ITRS Transistor Geometries

Transport-enhanced FETs	*Ultra-thin Body SOI FETs*		*Source/Drain Engineered FETs*	
 Strained Si, Ge, SiGe buried oxide isolation Silicon Substrate		 FD Si film S D Ground Plane BOX (<20nm) Bulk wafer	 Bias silicide nFET pFET Silicon Schottky barrier isolation	
Strained Si, Ge, SiGe, SiGeC or other semiconductor; on bulk or SOI	Fully depleted SOI with body thinner than 10 nm	Ultra-thin channel and localized ultra-thin BOX	Schottky source/drain	Non-overlapped S/D extensions on bulk, SOI, or DG devices

N-Gate (N>2) FETs	*Double-gate FETs*			
				
Tied gates (number of channels >2)	Tied gates, side-wall conduction	Tied gates planar conduction	Independently switched gates, planar conduction	Vertical conduction

ITRS 2003 Emerging Devices Section Pages 4 and 5

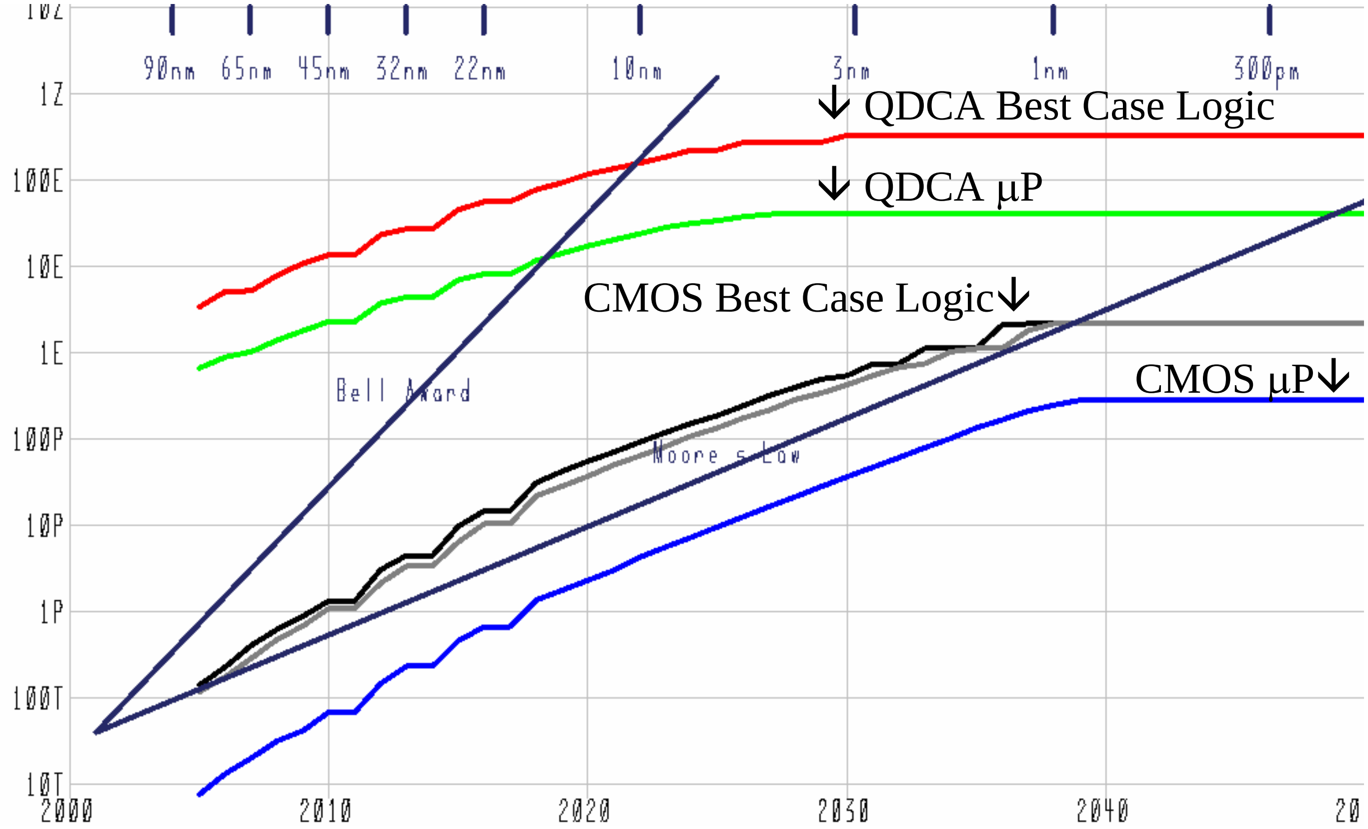
ITRS Technology Progression



ITRS 2003 Emerging Devices Section Page 12

Workup for Climate Modeling

- **Conclusion: CMOS to 200 Petaflops; QDCA to .5 Zettaflops**

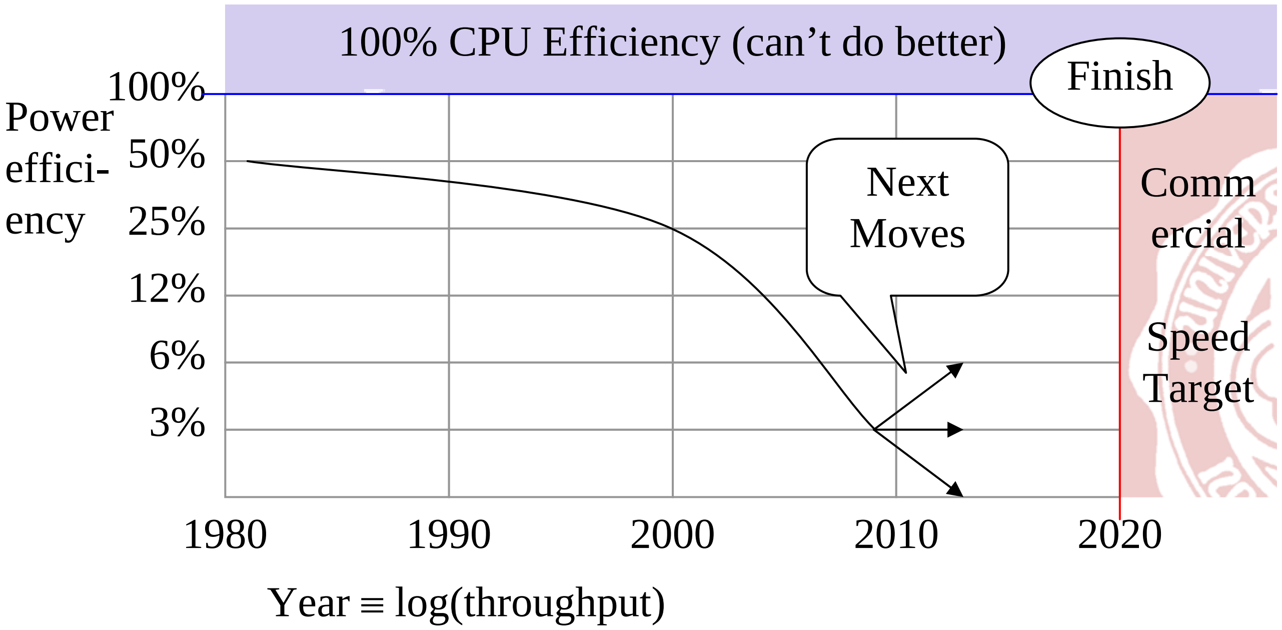


Outline

- Why Strive for Zettaflops?
  - Global Warming Mission
- The CMOS Roadmap
- Industry’s Responses 1 & 2 to Maturing CMOS
- Limits to Computing and Avoidance
- Industry’s Responses 3 to Maturing CMOS
- Conclusions

The Architecture Game

- **This is my diagram from a paper to illustrate CMOS architecture in light of CMOS scaling limits**
- **[Discuss]**



Next Moves:

↗ Switch to Vector Arch.
↗ Switch to SIMD Arch.
↗ Add Coprocessor
↗ Scale Linewidth
→Increase Parallelism
↘ Increase Cache
↘ More Superscalar
↘ Raise Vdd and Clk

“More Moore” and “More than Moore”

To Be Continued…

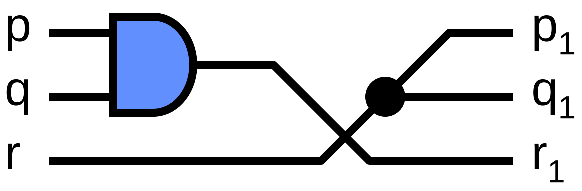
Outline

- Why Strive for Zettaflops?
  - Global Warming Mission
- The CMOS Roadmap
- Industry’s Responses 1 & 2 to Maturing CMOS
- Limits to Computing and Avoidance
- Industry’s Responses 3 to Maturing CMOS
- Conclusions

Landauer’s Limit and How to Avoid It

- **The original exposition of the connection between classical computing and heat generation**
  - **R. Landauer, “Irreversibility and Heat Generation in the Computing Process,” *IBM Journal of Research and Development*, vol. 5, Jul. 1961, pp. 183-191.**

Landauer's Paper (I)



Test Machine

Keys to argument:
Energy = TS
T = temperature
S = Entropy
$S = k_B \ln W$
k_B = Boltzmann's constant, 1.38×10^{-23}
W = number of states

For a fixed set of Boolean values for p, q, and, let W= the number of thermodynamic states of the physical apparatus. The W's are about the same for all sets of Boolean values.

From R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM Journal of Research and Development*, vol. 5, Jul. 1961, pp. 183-

Landauer’s Paper (II)

On input, we assert each of 8 states has equal probability

$$S_{inital} = -k_B \sum_{i = 1}^{8} 1/8 \ln(1/8) = 2.0794k_B$$

On output, states α and β have p=1/8 and γ and δ have p=3/8

$$S_{final} = -k_B(1/8 \ln(1/8) + 1/8 \ln(1/8) + 3/8 \ln(3/8) + 3/8 \ln(3/8)) = 1.2555k_B$$

“$S_{final} \geq S_{initial}$“ by second law of thermodynamics (for whole system – oops), or

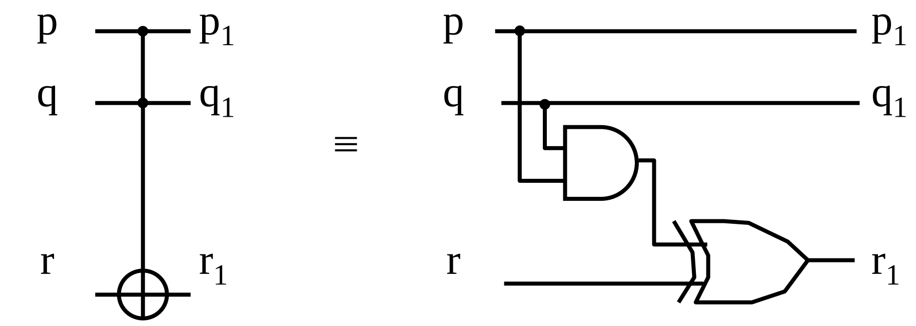
$S_{final} = S_{initial}$ + heat, heat > .8239k_BT

So basically, the output state has less information than the input, so some of the information appears as heat.

In today’s devices, heat is much greater than .8239k_BT; Landuaer’s analysis says .8239k_BT is a lower bound for an AND gate with balanced inputs

How to Avoid Landauer’s Heat Generation

- **Answer: Use gates that avoid reducing states**
  - **I. e. use gates that don’t destroy information**
  - **Use gates that are logically reversible**



Toffoli Gate Functional Equivalent

  - **If p and q are true, flip r**
  - **Function is its own inverse**

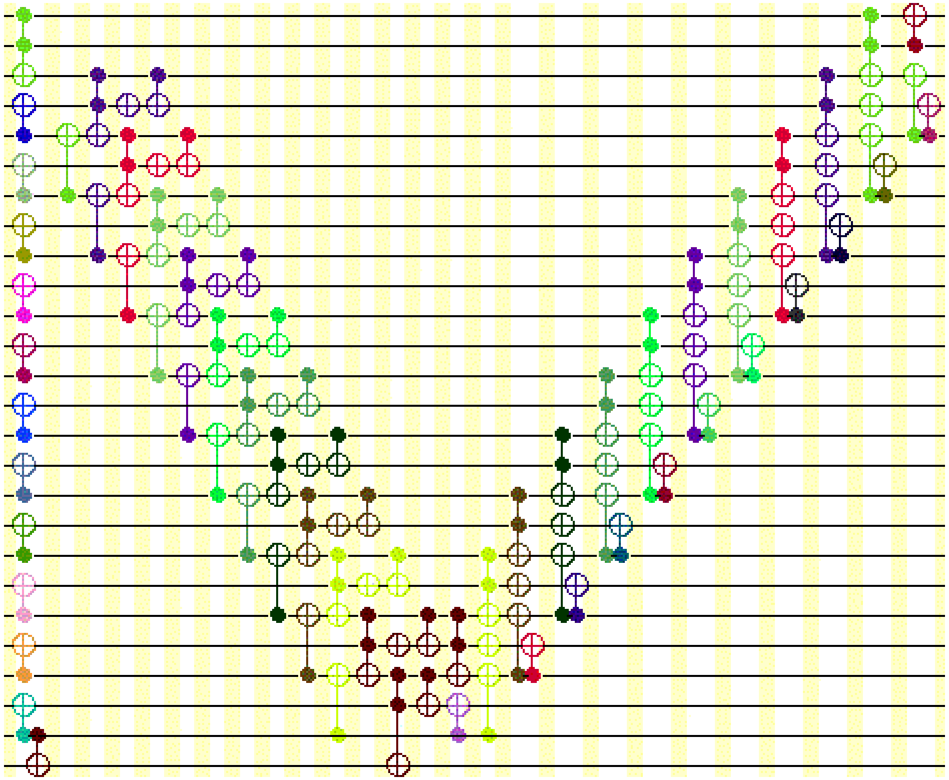
How to Avoid Landauer’s Heat Generation

- **The Toffoli gate just rearranges the 8 states**
- **By Landauer’s argument, minimum entropy generation is zero**

BEFORE CYCLE				AFTER CYCLE		
p	q	r		p_1	q_1	r_1
1	1	1	→	1	1	0
1	1	0	→	1	1	1
1	0	1	→	1	0	1
1	0	0	→	1	0	0
0	1	1	→	0	1	1
0	1	0	→	0	1	0
0	0	1	→	0	0	1
0	0	0	→	0	0	0

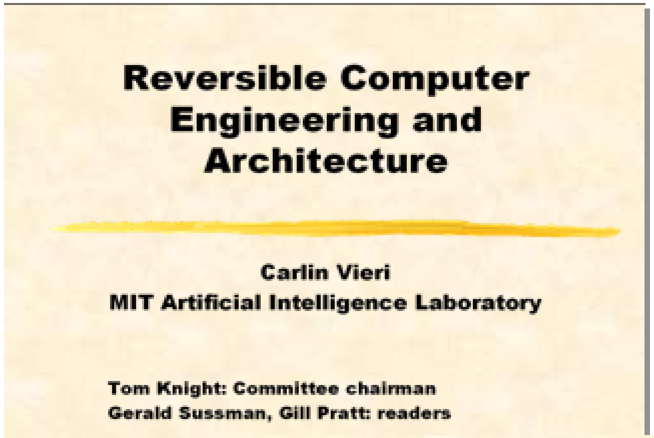
But Can You Compute?

- Yes, Toffoli is universal
  - Typically used with CNOT, invert, and there are “garbage disposal” issues
- Furthermore, there are other gates that are universal and reversible, like Fredkin
- Adder →
  - From top a_0, b_0, a_1, b_1…
  - From top a_0, $(a+b)_0$, …



Reversible Microprocessor Status

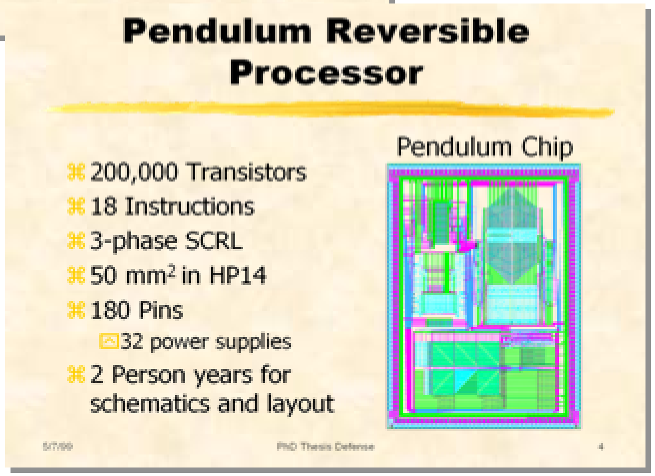
- Status
  - Subject of Ph. D. thesis
  - Chip laid out (no floating point)
  - RISC instruction set
  - C-like language
  - Compiler
  - Demonstrated on a PDE
  - However: really weird and not general to program with +=, -=, etc. rather than =



Reversible Computer Engineering and Architecture

Carlin Vieri
MIT Artificial Intelligence Laboratory

Tom Knight: Committee chairman
Gerald Sussman, Gill Pratt: readers



Pendulum Reversible Processor

- 200,000 Transistors
- 18 Instructions
- 3-phase SCRL
- 50 mm^2 in HP14
- 180 Pins
  - 32 power supplies
- 2 Person years for schematics and layout

Pendulum Chip

5/7/99 PhD Thesis Defense 4

Logic Gates & Computer Heat, Conclusions

- George Boole introduced the world to universal AND-OR-NOT logic, and we stuck with it
  - AND & OR are not information-preserving and must generate heat
- Other universal gate sets need not generate heat (Toffoli, Fredkin), but they are less known

George Boole
(1815-1864)

(April 17, 2008)

Reversible Logic Parameters

- We need some data point on performance
- Graph to right from a published paper by Lent et. al. Notre Dame on quantum dot cellular automata
- However, architectural considerations say their operating points are not ideal

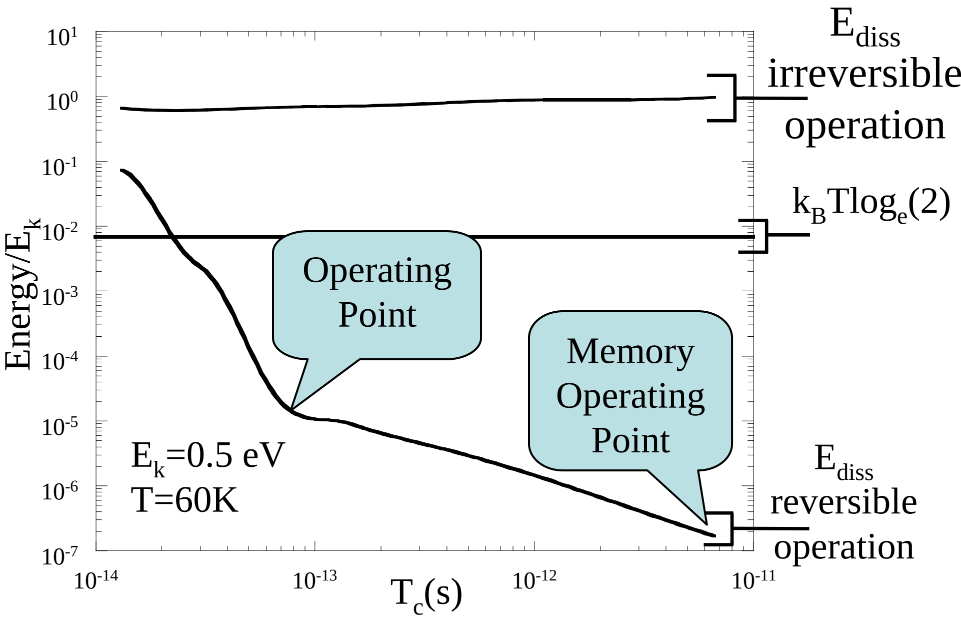
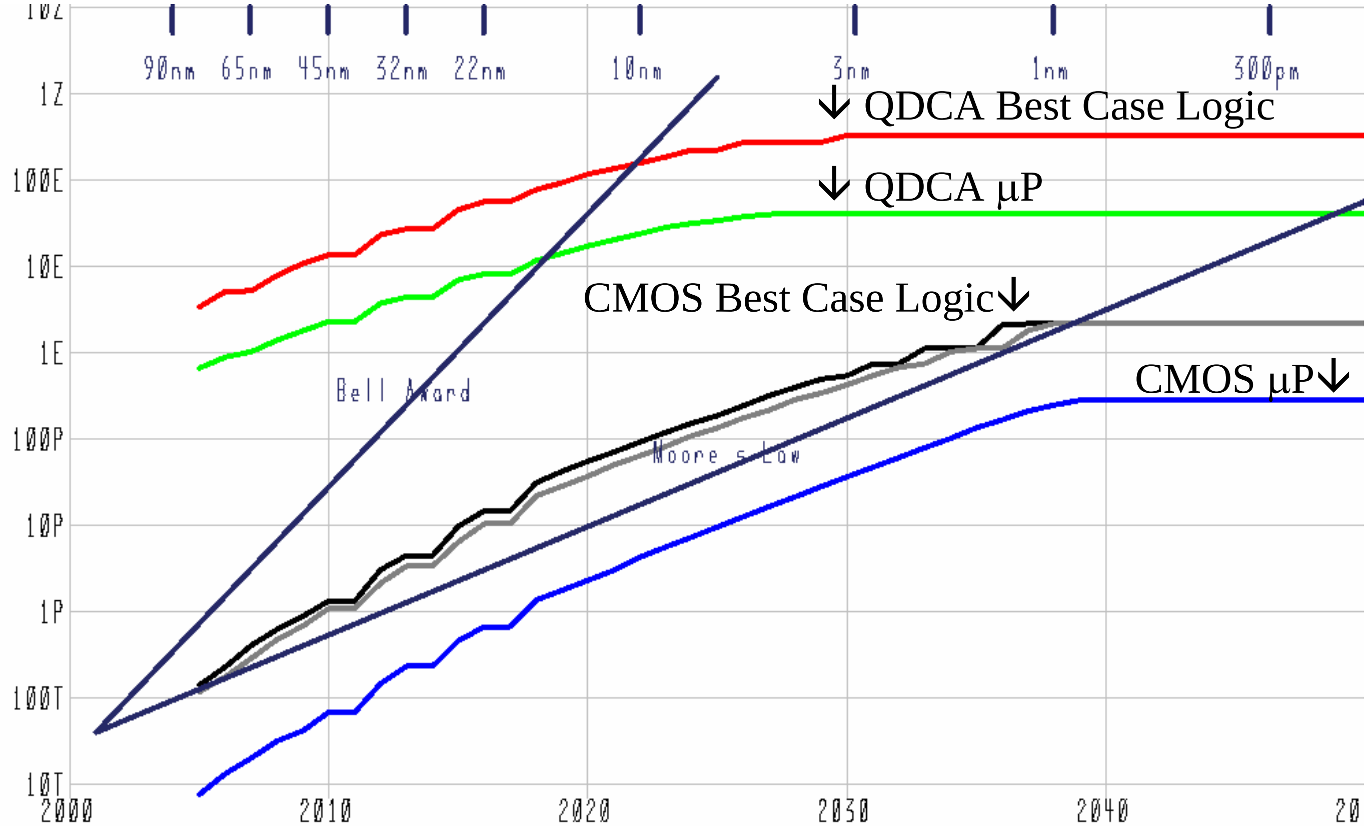


Figure 8: Molecular Quantum Dot Cellular Automata speed-energy curve for irreversible and reversible operation (courtesy of C. Lent) with operating points used in this paper labeled.

Workup for Climate Modeling

- **Conclusion: CMOS to 200 Petaflops; QDCA to .5 Zettaflops**



CMOS and Beyond CMOS Limits

- CMOS per ITRS roadmap
  - With operating points adjusted for climate modeling machines instead of matching Moore’s Law
  - 200 Petaflops @ 2 MW
- DARPA Exascale study
  - 1 Exaflops @ >2 MW

- A New Computing Device
  - Notre Dame QDCA
  - Reversible Logic
  - .5 Zettaflops

Sandia National Laboratories

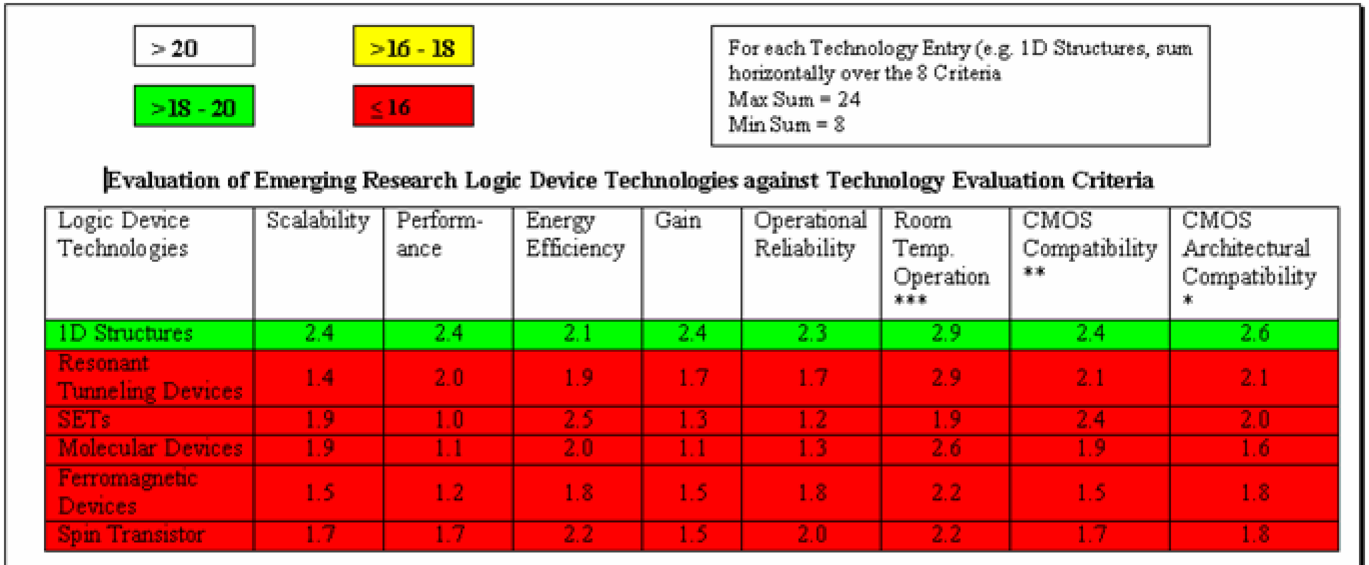
Outline

- Why Strive for Zettaflops?
  - Global Warming Mission
- The CMOS Roadmap
- Industry’s Responses 1 & 2 to Maturing CMOS
- Limits to Computing and Avoidance
- Industry’s Responses 3 to Maturing CMOS
- Conclusions

Transistor Replacement Alternatives 2006

- ITRS ERD [see below]
  - Influential over industrial and government funding

- International Technology Roadmap for Semiconductors (ITRS) Emerging Research Devices (ERD) architecture panel. All new devices are inadequate except CNFET



□ > 20

🟨 >16 - 18

🟩 >18 - 20

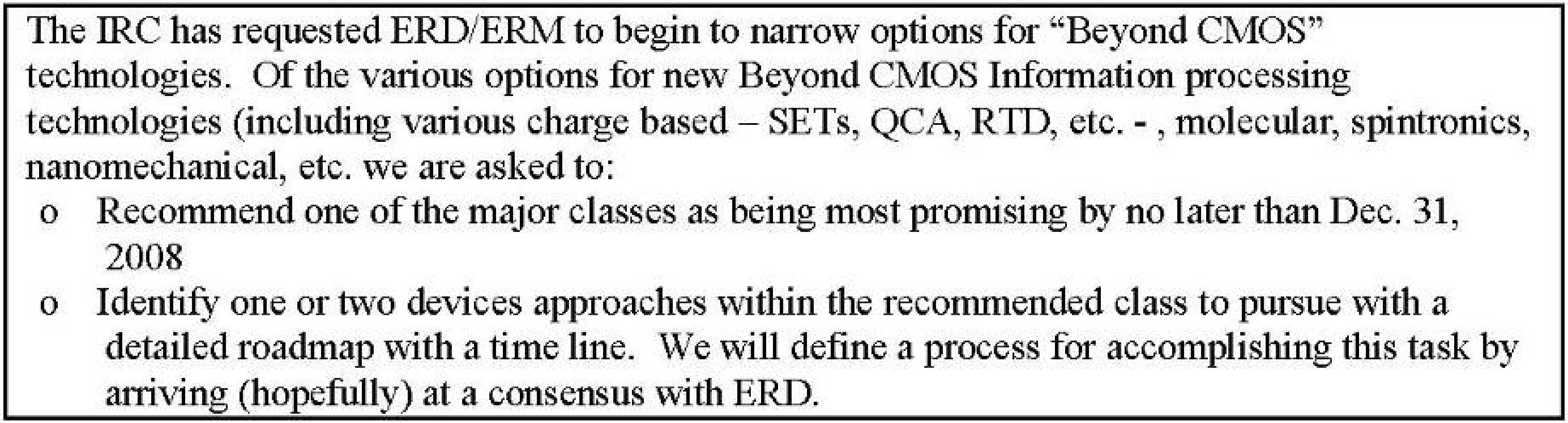
🟥 ≤ 16

For each Technology Entry (e.g. 1D Structures, sum horizontally over the 8 Criteria
Max Sum = 24
Min Sum = 8

Evaluation of Emerging Research Logic Device Technologies against Technology Evaluation Criteria

Logic Device Technologies	Scalability	Perform-ance	Energy Efficiency	Gain	Operational Reliability	Room Temp. Operation ***	CMOS Compatibility **	CMOS Architectural Compatibility *
1D Structures	2.4	2.4	2.1	2.4	2.3	2.9	2.4	2.6
Resonant Tunneling Devices	1.4	2.0	1.9	1.7	1.7	2.9	2.1	2.1
SETs	1.9	1.0	2.5	1.3	1.2	1.9	2.4	2.0
Molecular Devices	1.9	1.1	2.0	1.1	1.3	2.6	1.9	1.6
Ferromagnetic Devices	1.5	1.2	1.8	1.5	1.8	2.2	1.5	1.8
Spin Transistor	1.7	1.7	2.2	1.5	2.0	2.2	1.7	1.8

Selecting Successor to CMOS by 12/31/2008



The IRC has requested ERD/ERM to begin to narrow options for "Beyond CMOS" technologies. Of the various options for new Beyond CMOS Information processing technologies (including various charge based – SETs, QCA, RTD, etc. - , molecular, spintronics, nanomechanical, etc. we are asked to:

- o Recommend one of the major classes as being most promising by no later than Dec. 31, 2008
- o Identify one or two devices approaches within the recommended class to pursue with a detailed roadmap with a time line. We will define a process for accomplishing this task by arriving (hopefully) at a consensus with ERD.

- **From meeting and e-mail to committee ↑**
- **The semiconductor industry is waking up**
- **Downselect "beyond CMOS" options through a advocate/skeptic competition**

Downselect Criteria

Basic description	This section comprises a description of the proposed device family. The section may include textual and graphical descriptions but should be independent of (or parameterized by) feature size F	
Principle of Operation	Control mechanism	*Thermal injection over gate barrirer*
	Operating temperature	*Usually 25C - 125C*
Materials and Geometry	Base	*Si*
	Device Architecture	*FET*
	Patterning	*Lithography*
	Design	*2D layout*
	Circuit element	*Transistor, 3 or 4 terminal*
	Device density as a function of feature size F	*~ 1/F^2*
	Size in units of feature size F of a gate equivalent to a 2-input NAND gate, including contacts and isolation and necessary peripheral circuitry	*>~65 F^2*
State variables and control	State variable	*Voltage*
	Number of logic states	*2 (high and low)*
Logic Family	Information processing basis	*Universal set comprising NAND, NOR, NOT logic gates, also pass gates*
	Interconnects	*Wire*
	Compatible memory	*SRAM (fast) , DRAM (dense)*
	Clock	*CMOS based clock circuits*
	CMOS compatible	*N/A*

Downselect Criteria

Limitations	This section comprises a list of known limiting factors for performance and manufacturing	
Materials and Geometry	Sources of variability	LER, Doping fluctuations ~ 1/SQRT (LW)
	External parasitics	Access resistance, fringe capacitance
State variables and control	Noise margin	(Vdd-Vth)/ KT/q > 5
	QM limit	Tuneling: Band to Band, Source-to-Drain

Performance Potential	This section comprises an extrapolation of the technology to about the year 2020, stipulating F=14 nm. Provide best estimate numerical values.	
Switching speed and energy	Intrinsic speed of single element	Lchan/v ~ 0.1ps
	Self Gain	gm/gd ~ Vdd/DIBL
	Proposed clock rate	xxx
	Switching Energy per gate or gate equivalent @ proposed clock rate	0.5*Cload*Vdd^2
	Static Power Dissipation per gate or gate equivalent	Vdd*Ioff*(2/5)
Interconnect	Interconnect delay per micron	RC
	Interconnect energy as a function of distance at proposed clock rate	CV^2

Outline

- Why Strive for Zettaflops?
  - Global Warming Mission
- The CMOS Roadmap
- Industry's Responses 1 & 2 to Maturing CMOS
- Limits to Computing and Avoidance
- Industry's Responses 3 to Maturing CMOS
- Conclusions

Conclusions (I/III)

- End-user applications: Understanding and mitigating global warming to save the Earth
  - The climate modeling community can supply representatives that say 1 Zettaflops is needed
  - “Faster computers are better,” but there are few other specific examples
- New computer required >2 Exaflops
  - DARPA IPTO is preparing a plan for 1 Exaflops but that looks like a stretch goal for mature CMOS
  - Reference Zettaflops workshop that there is no CMOS solution beyond 1 Exaflops

Conclusions (II/III)

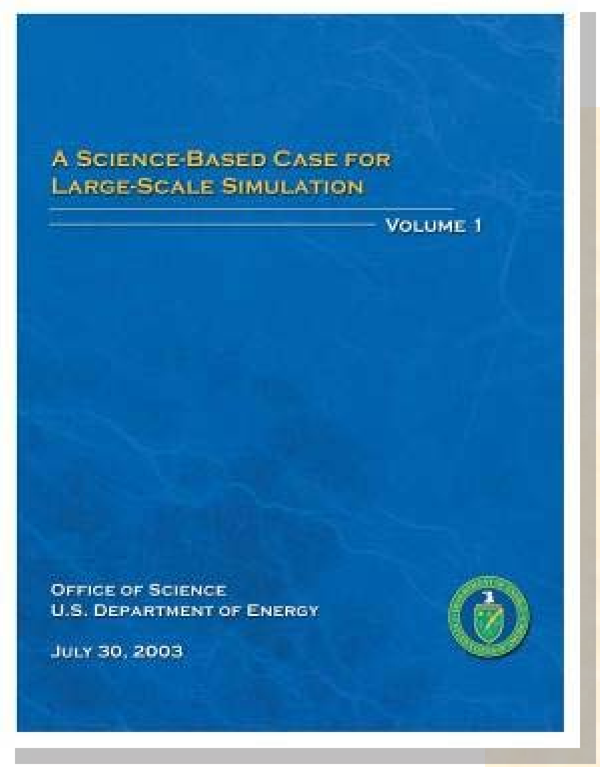
- Physical science research is seeking to discover a new computing device
  - ITRS calls this the “new switch”
    - we can guarantee it won’t be a switch
  - NRI, NSF, maybe national labs have infrastructure in place and can distribute research funds
  - Downselect competition complete12/31/2008

Conclusions (III/III)

- Determiners of Progress
  - – Industry has a CMOS roadmap for a dozen years
  - – CMOS architecture progress and parallel processing will be required to use advances
  - – Industry is searching for a new device, starting now
  - – Key point: Device won’t work with AND-OR-NOT logic (user retraining?)

Climate Modeling as an Application

- **SCaLeS study included section on climate**
- **Understanding and mitigating global warming analyzed and requires 1 Zettaflops**

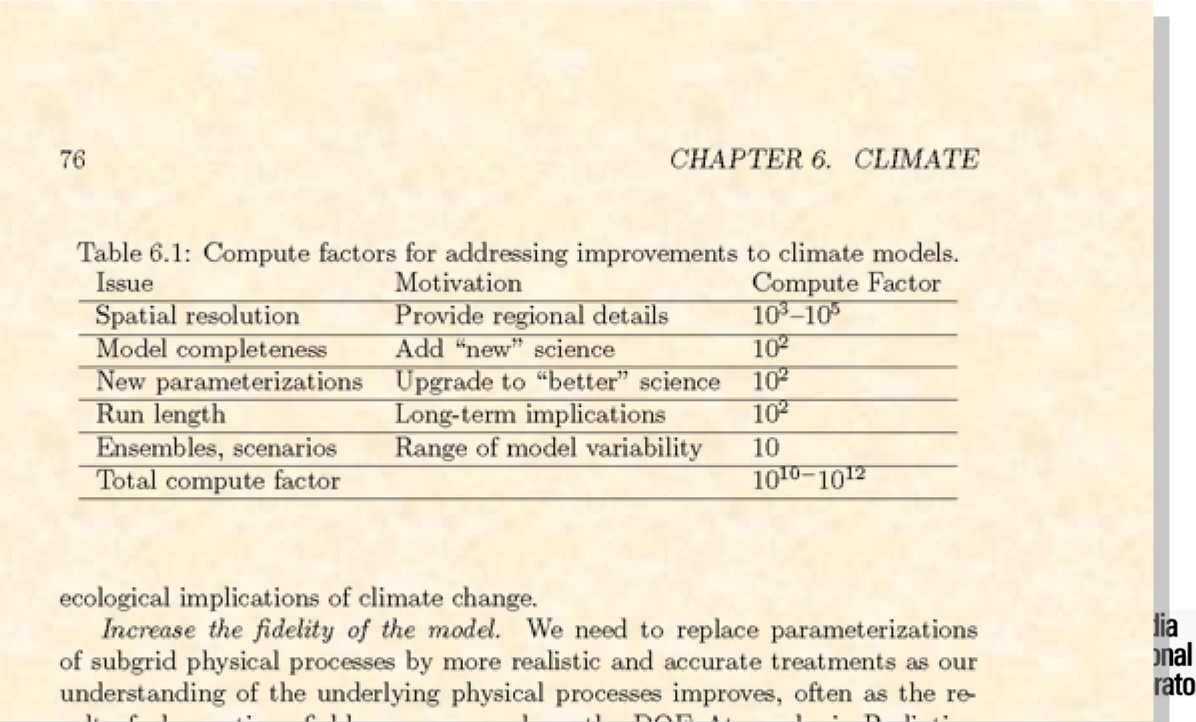


A SCIENCE-BASED CASE FOR LARGE-SCALE SIMULATION

VOLUME 1

OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

JULY 30, 2003



76 CHAPTER 6. CLIMATE

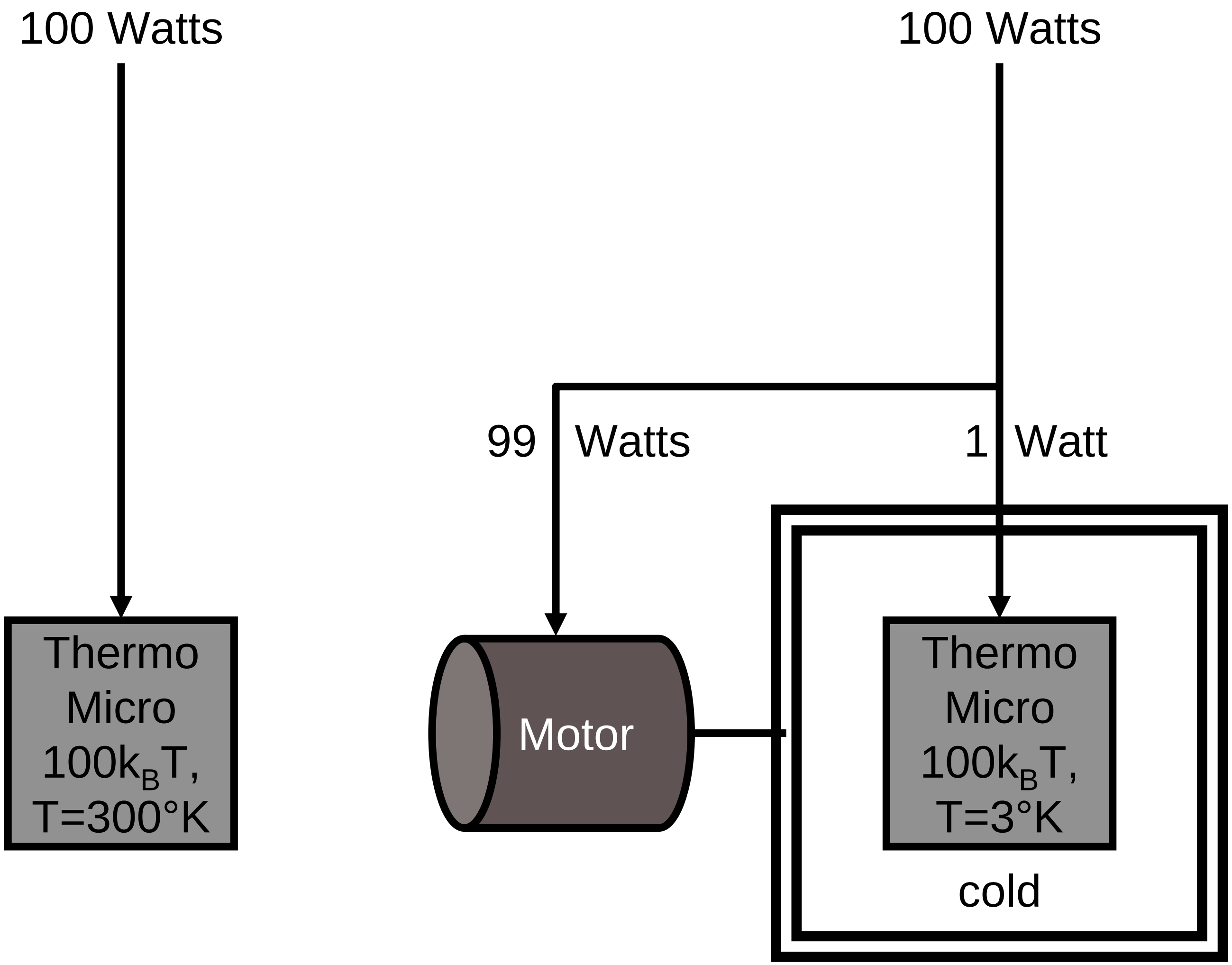
Table 6.1: Compute factors for addressing improvements to climate models.

Issue	Motivation	Compute Factor
Spatial resolution	Provide regional details	10^3–10^5
Model completeness	Add “new” science	10^2
New parameterizations	Upgrade to “better” science	10^2
Run length	Long-term implications	10^2
Ensembles, scenarios	Range of model variability	10
Total compute factor		10^{10}–10^{12}

ecological implications of climate change.

Increase the fidelity of the model. We need to replace parameterizations of subgrid physical processes by more realistic and accurate treatments as our understanding of the underlying physical processes improves, often as the re-

Cutting Temperature



Cutting Temperature

Carnot Efficiency $\eta_c = \frac{T_c}{T_h\text{-}T_c}$

Specific Power $1/\eta_c = \frac{T_h\text{-}T_c}{T_c}$

Specific power is watts input power required to remove one watt at the cooling temperature

Idea:
To cut computer power, let's cool the active devices to 3° K. This will cut minimum power per reliable operation from $100k_B \times 300$ to $100k_B \times 3$, cutting device power by 100 fold!

Specific Power $1/\eta_c = \frac{T_h\text{-}T_c}{T_c}$

$$= \frac{300 - 3}{3}$$

$$= 99$$

Thus, we cut device power to 1% of original power at the price of a refrigerator consuming 99% of the original power, for resulting total power consumption of 100% of original power.

However, refrigerators are typically <20% efficient, so we're actually in the hole by 5× …
but it is cheaper to dissipate power in a big motor than an expensive chip.

How to Project Uniprocessor Performance

- **Let’s assume industry makes the innovations called for by the ITRS on schedule**
- **However, companies will not be constrained to do everything like the ITRS**
  - **Engineers can choose any power supply voltage they like**
  - **Doping levels can be changed**

- **Evaluate**

max(SpecFP)

engineering
← choices,
architecture

and report performance and architecture as a function of years into the future

UT Austin Study (2000)

- The Study
  - Clock Rate versus IPC: The End of the Road for Conventional Microarchitectures, Vikas Agarwal, M.S. Hrishikesh, Stephen W. Keckler, Doug Burger. 27th Annual International Symposium on Computer Architecture

- Conclusions (to be Explained)
  - Modified ITRS roadmap predictions to be more friendly to architectures
  - Concluded there would be a 12%/year growth…
  - However, recent growth has been ~30%, with industry’s maneuver to cheat the analysis instructive

Wire Delay Coverage in ITRS

- **Wire delay added to ITRS 2002 edition**

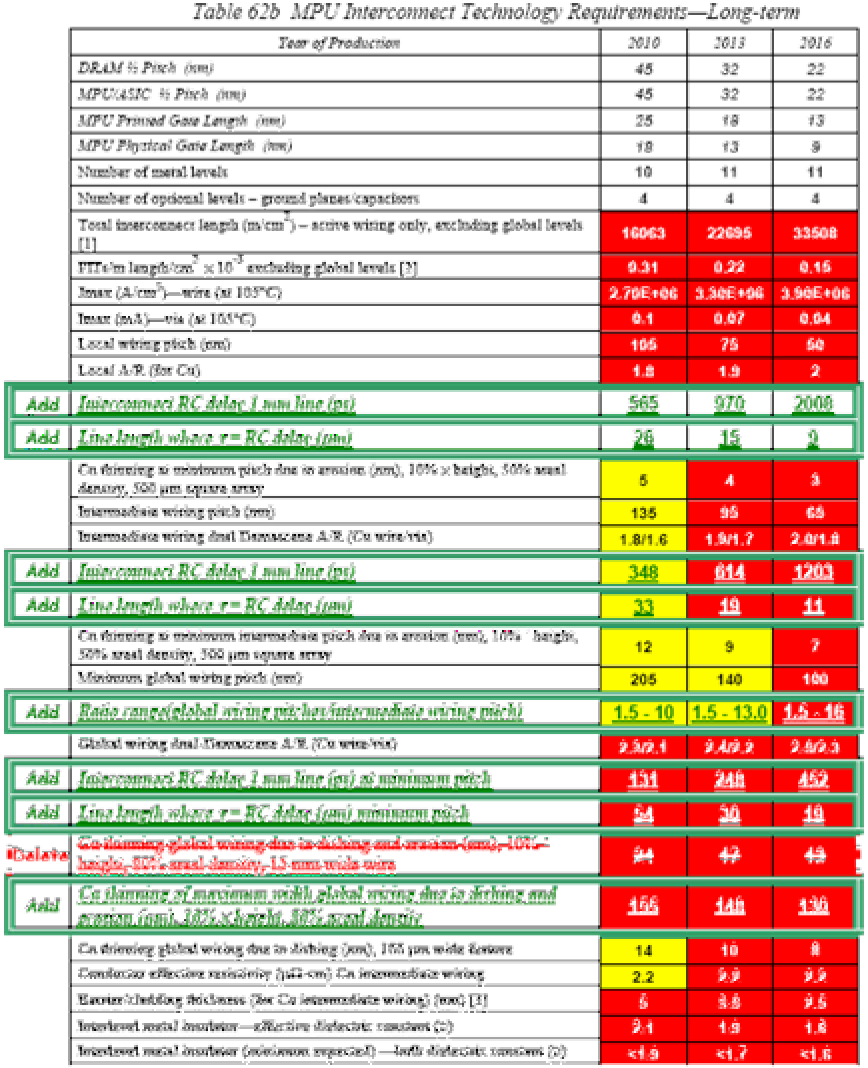


Table 62b MPU Interconnect Technology Requirements—Long-term

	Year of Production	*2010*	*2013*	*2016*
	DRAM ½ Pitch (nm)	45	32	22
	MPU/ASIC ½ Pitch (nm)	45	32	22
	MPU Printed Gate Length (nm)	25	18	13
	MPU Physical Gate Length (nm)	18	13	9
	Number of metal levels	10	11	11
	Number of optional levels – ground planes/capacitors	4	4	4
	Total interconnect length (m/cm^2) – active wiring only, excluding global levels [1]	16063	22695	33508
	FITs/m length/cm^2 × 10^{-3} excluding global levels [2]	0.31	0.22	0.15
	Jmax (A/cm^2)—wire (at 105°C)	2.70E+06	3.30E+06	3.90E+06
	Imax (mA)—via (at 105°C)	0.1	0.07	0.04
	Local wiring pitch (nm)	105	75	50
	Local A/R (for Cu)	1.8	1.9	2
Add	*Interconnect RC delay 1 mm line (ps)*	565	970	2008
Add	*Line length where τ = RC delay (µm)*	26	15	9
	Cu thinning at minimum pitch due to erosion (nm), 10% × height, 50% areal density, 500 µm square array	5	4	3
	Intermediate wiring pitch (nm)	135	95	65
	Intermediate wiring dual Damascene A/R (Cu wire/via)	1.8/1.6	1.9/1.7	2.0/1.8
Add	*Interconnect RC delay 1 mm line (ps)*	348	614	1203
Add	*Line length where τ = RC delay (µm)*	33	19	11
	Cu thinning at minimum intermediate pitch due to erosion (nm), 10% × height, 50% areal density, 500 µm square array	12	9	7
	Minimum global wiring pitch (nm)	205	140	100
Add	*Ratio range(global wiring pitches/intermediate wiring pitch)*	1.5 - 10	1.5 - 13.0	1.5 - 16
	Global wiring dual Damascene A/R (Cu wire/via)	2.3/2.1	2.4/2.2	2.5/2.3
Add	*Interconnect RC delay 1 mm line (ps) at minimum pitch*	131	248	452
Add	*Line length where τ = RC delay (µm) minimum pitch*	54	30	19
Delete	~~Cu thinning global wiring due to dishing and erosion (nm), 10% × height, 80% areal density, 15 mm wide wire~~	24	17	13
Add	*Cu thinning of maximum width global wiring due to dishing and erosion (nm), 10% x height, 80% areal density*	156	140	130
	Cu thinning global wiring due to dishing (nm), 100 µm wide feature	14	10	8
	Conductor effective resistivity (µΩ-cm) Cu intermediate wiring	2.2	2.2	2.2
	Barrier/cladding thickness (for Cu intermediate wiring) (nm) [3]	6	3.8	2.5
	Interlevel metal insulator—effective dielectric constant (κ)	2.1	1.9	1.8
	Interlevel metal insulator (minimum expected) —bulk dielectric constant (κ)	<1.9	<1.7	<1.6

Modeling Wire Delay

- For some year in the future
  - ITRS and other models project a clock rate
  - ITRS and other models project a signal propagation velocity
  - Divide the two figures to get d=distance traveled in one clock cycle
  - Chip area/d^2 is plotted at right →

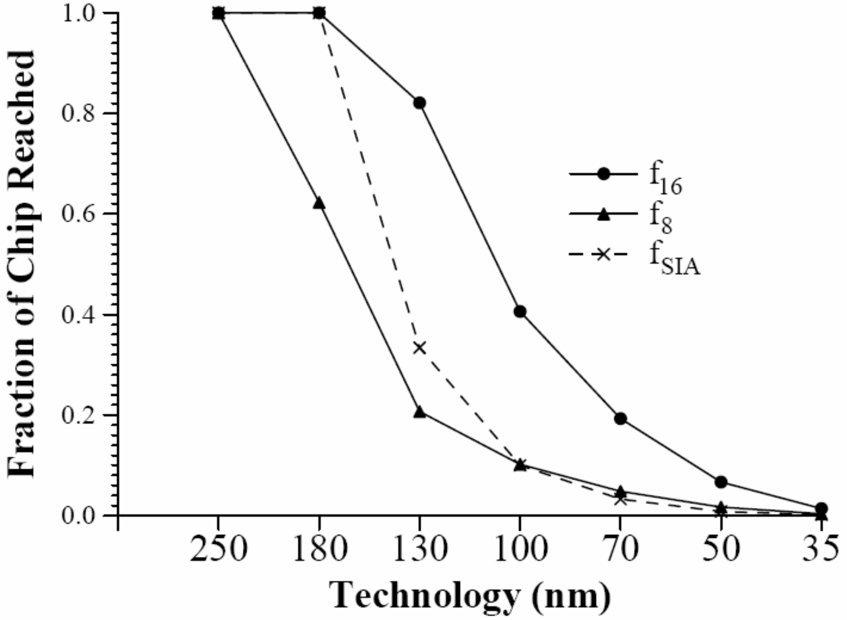


Figure 4: Fraction of total chip area reachable in one cycle.

- **Figure 4 from “Clock Rate versus IPC: The End of the Road for Conventional Microarchitectures,” Vikas Agarwal, M.S. Hrishikesh, Stephen W. Keckler, and Doug Burger**

Cache Performance

- **Authors used ECacti cache modeling tool**
- **ECacti lays out caches in terms of banks, associatively, etc.**
- **As technology progresses, size of cache accessible in 3 cycles decreases**
- **Remedy is obvious, but has consequences: increase depth of pipelining**

This graph for a
3 cycle cache access

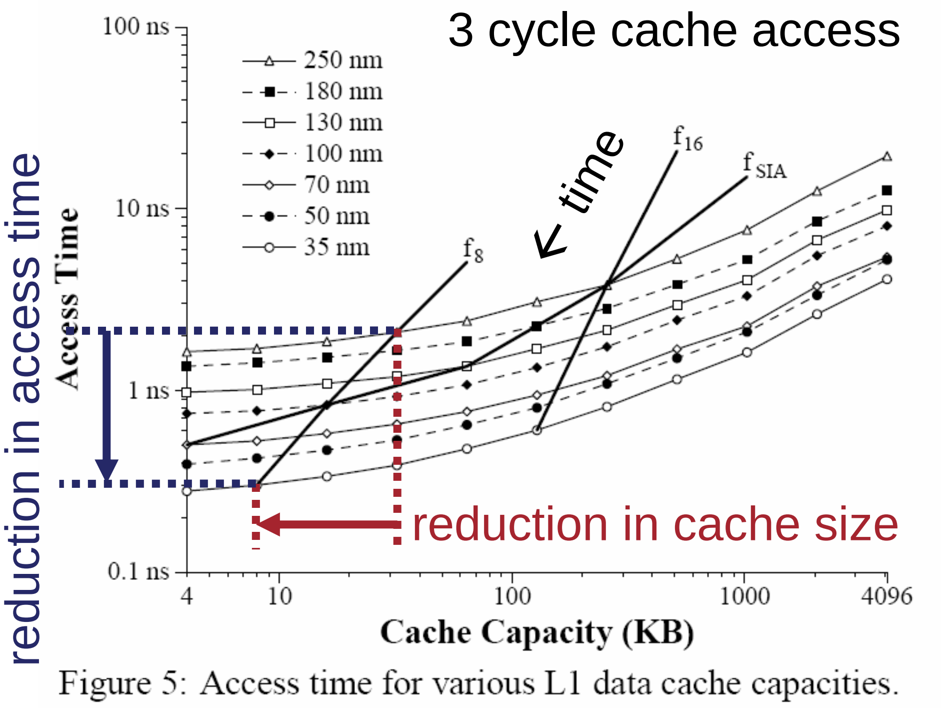


Figure 5: Access time for various L1 data cache capacities.

- **Figure 5 from "Clock Rate versus IPC: The End of the Road for Conventional Microarchitectures Vikas Agarwal, M.S. Hrishikesh, Stephen W. Keckler, and Doug Burger**

Modeling Pipelined μP

- Authors used SimpleScalar, cycle accurate simulator of a DEC Alpha 21264
- However, actually models hypothetical future μPs with parameterized
  - Cache parameters
  - Pipeline depth
  - Branch prediction
  - Technology (clock speed)
- Authors used SimpleScalar to model the 18 SPEC95 benchmarks for 500 million instructions each
  - Adjustments to avoid initialization
- Question to answer: What is the best architecture, and how well does it work?

Simulation Results

- **Results shown at right → are noted by author to be “remarkably consistent”**
- **If fact, the results are almost the same as the clock rate increase**
- **Conclusion: To first order, SPEC ratings will increase with speed of clock**
  - **Noting that this analysis is per μP core, and SPEC is for one core**

Pipeline = caches same size but more pipelining to keep access rate same
Capacity = cut cache size so access is possible without cutting clock rate

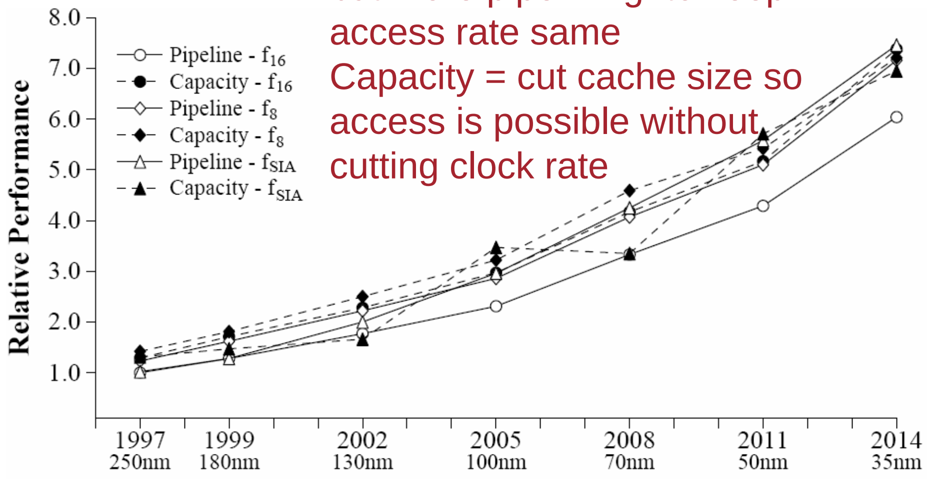


Figure 7: Performance increases for different scaling strategies.

- **Figure 7 from “Clock Rate versus IPC: The End of the Road for Conventional Microarchitectures Vikas Agarwal, M.S. Hrishikesh, Stephen W. Keckler, and Doug Burger**