SAND2010-7422C
SAND 2010-xxxx

Moore's Law

Erik P. DeBenedictis

New Mexico State University

October 25, 2010

Sandia National Laboratories

Key Points

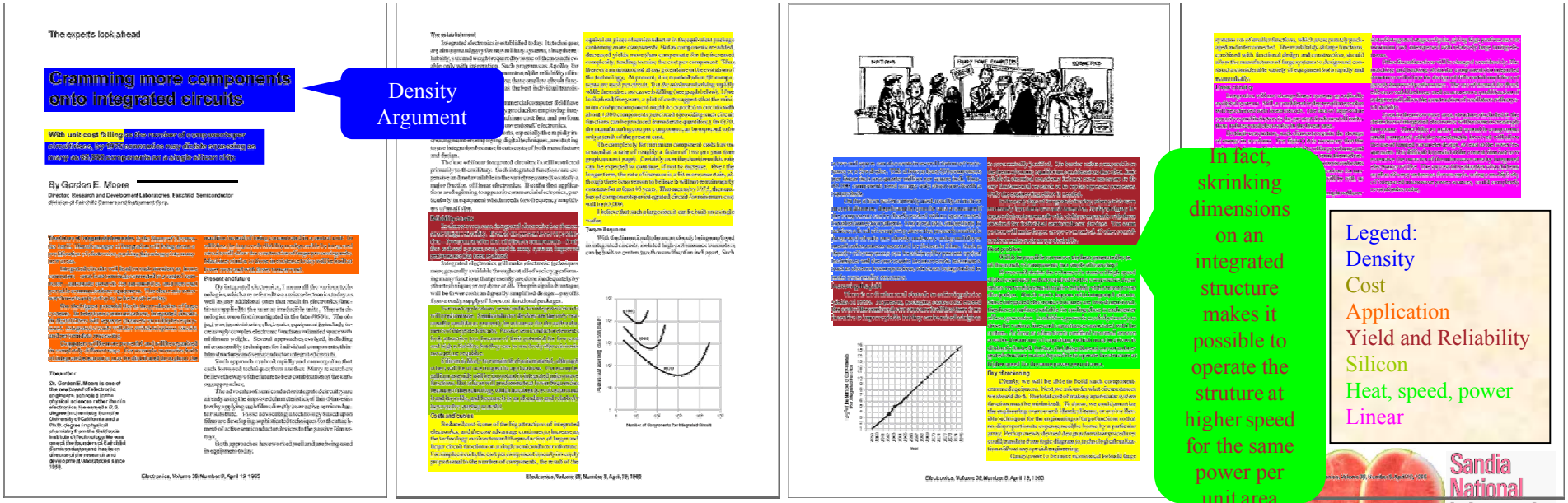
- Gordon Moore wrote a paper in 1965 that became the basis of:
  - Moore's law, formalized technically as semiconductor scaling
  - Growth in Semiconductor industry, formalized by product marketing
- Moore's Law has shifted and is headed towards greater density at decreasing duty cycle per device
- There are other technologies that could supplant CMOS and continue traditional scaling

Agenda

- Gordon Moore's 1965 Paper
- International Technology Roadmap for Semiconductors (ITRS)
  - Innovations
  - Power and Clock Rate
  - System Performance
  - Interconnect
- Faster computing with the power turned off
- Exotic: Reversible Logic
- Conclusions

The New Moore's Law

- **The following text appears in the original article and highlights the role of marketing in redefining Moore's Law:**
  - **"In fact, shrinking dimensions on an integrated structure makes it possible to operate the structure at higher speed for the same power per unit area"**
- **How many scientists in the room could get better research results if they were able to redefine e. g. the speed of light?**



Moore's Paper from Electronics magazine 1965 with topics highlighted

Early Formalization of Moore's Law

- Moore's paper from 1965 implied a basic knowledge of CMOS scaling
  - Later to be →
- Industry expectations were set by Moore's Law
- However, VDD not scaling as expected
- Capacitance model no longer holds
- Software has bursty behavior

- Dennard Scaling
  - Area $1/\kappa^2$
  - Capacitance $1/\kappa$
  - Resistance κ
  - Threshold (V_{th}) $1/\kappa$
  - Current (I_d) $1/\kappa$
  - Gate Delay (τ_{gd}) $1/\kappa$
  - Wire Delay (τ_{wire}) 1
  - Power $1/\kappa^2$ → $1/\kappa^3$

Agenda

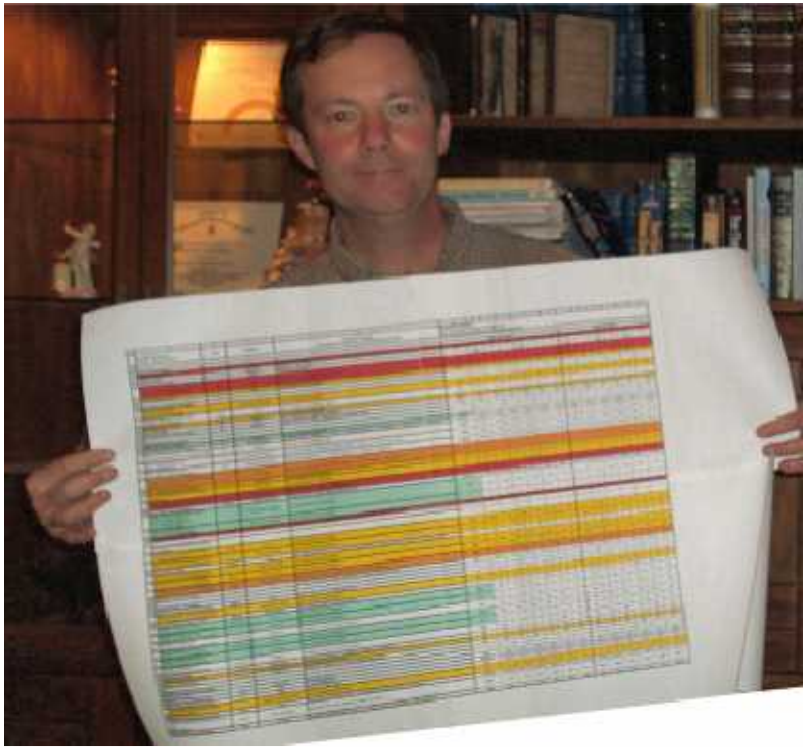
- Gordon Moore's 1965 Paper
- International Technology Roadmap for Semiconductors (ITRS)
  - Innovations
  - Power and Clock Rate
  - System Performance
  - Interconnect
- Faster computing with the power turned off
- Exotic: Reversible Logic
- Conclusions

ITRS Spreadsheet

- Review spreadsheet interactively in Excel
- Points to make
  - Illustrate role and implementation of "targets"
    - Line width
    - Clock rate
  - Illustrate user inputs
    - Sub threshold adjustment factors rows 34 & 36
  - Illustrate rows derived by calculation
  - Illustrate iteration to target
  - Illustrate HP LOP LSTP
- Draw conclusions
  - Industry defines targets
  - Table preparer adds value by scheduling innovations to meet targets
  - Validity depends on innovations occurring on schedule
- Limited example next slide

ITRS Process Integration Spreadsheet

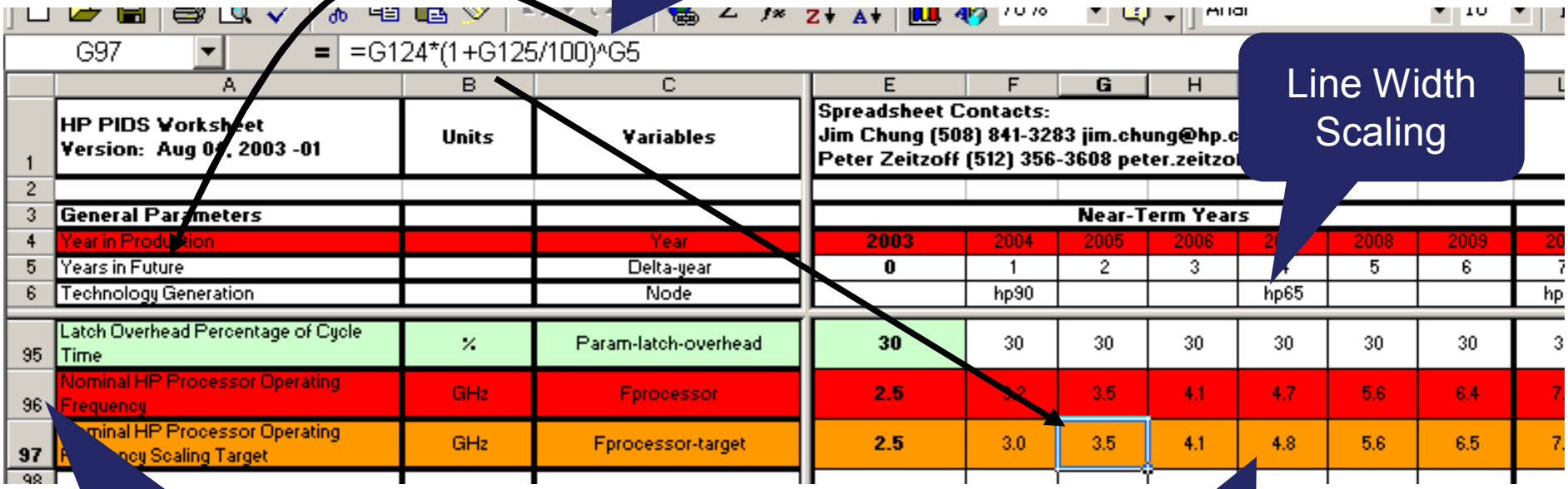
- Big Spreadsheet
  - Columns are years
  - Rows are 100+ transistor parameters
  - Manual entry of process parameters by year
  - Excel computes operating parameters
  - Extra degrees of freedom go to making Moore's Law smooth – not the best computers



Sandia National Laboratories

ITRS Spreadsheet Structure

Target is exponential in "Years in Future"



Line Width Scaling

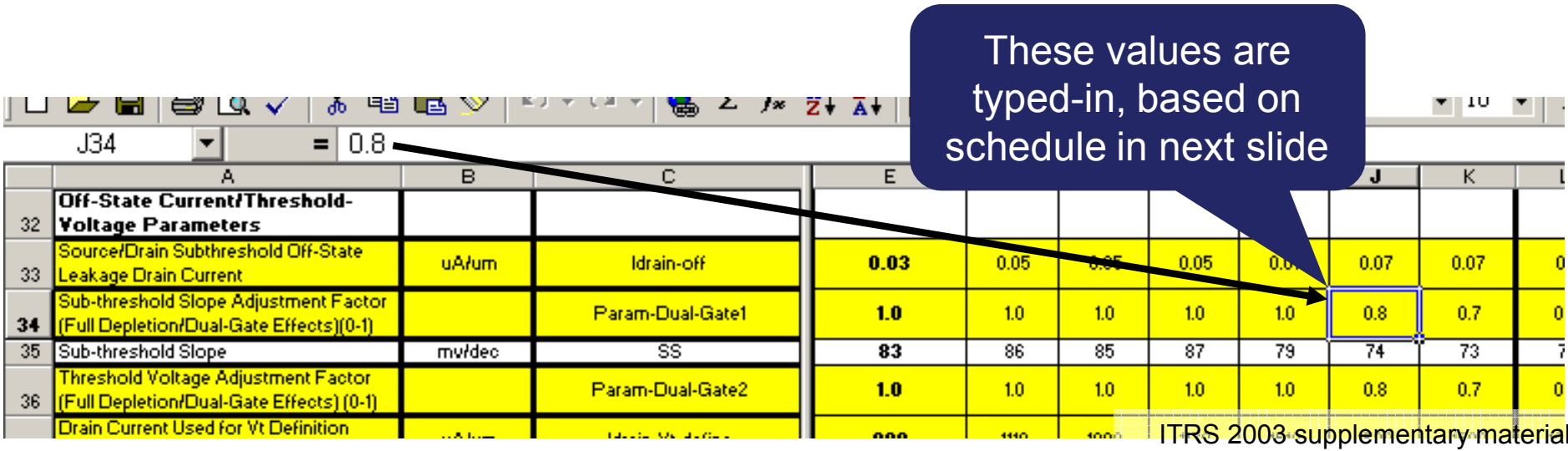
Fprocessor is result of 96 rows of targets, inputs, and iterative calculation

Result usually matches to one decimal place!

ITRS 2003 supplementary material

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**



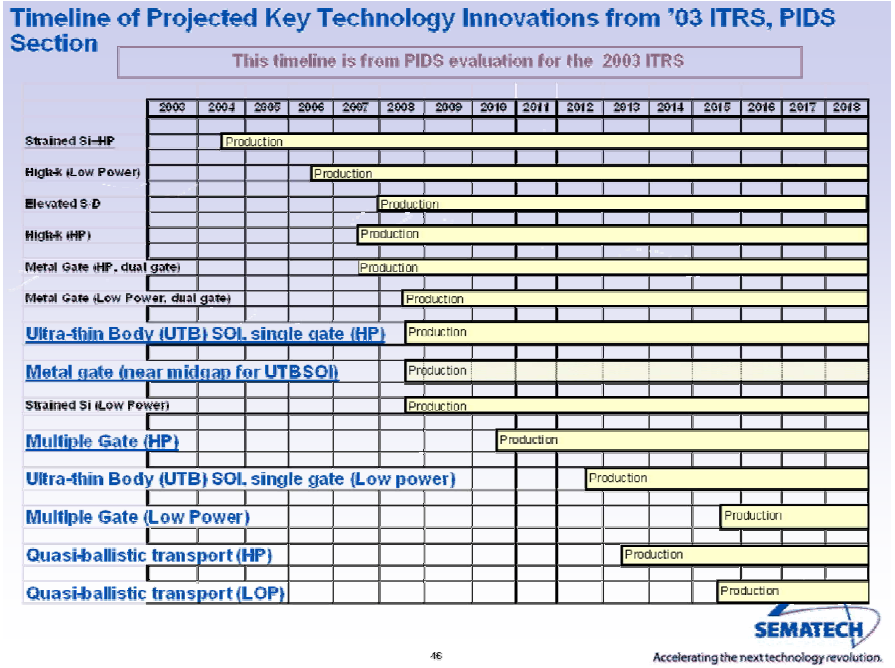


Agenda

- Gordon Moore's 1965 Paper
- International Technology Roadmap for Semiconductors (ITRS)
  - Innovations
  - Power and Clock Rate
  - System Performance
  - Interconnect
- Faster computing with the power turned off
- Exotic: Reversible Logic
- Conclusions

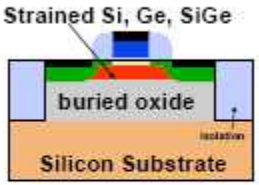
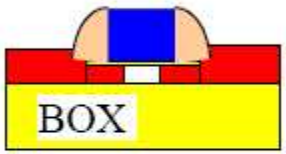
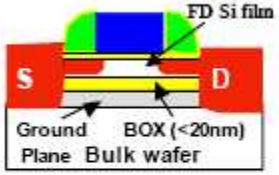
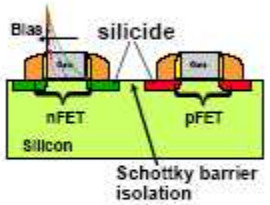
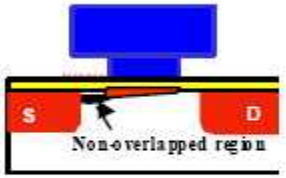
Schedule of Innovations

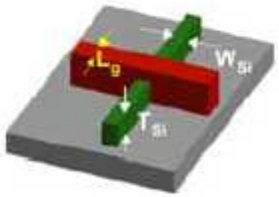
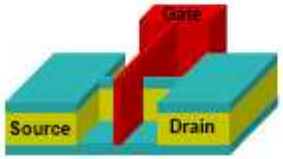
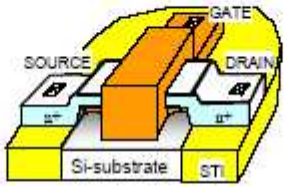
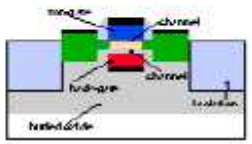
- To make the calculations fit the projection of a smooth "Moore's Law," certain variables must be adjustable
- The independent variables are a "schedule of innovations," or technology advances that must enter production on certain years



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

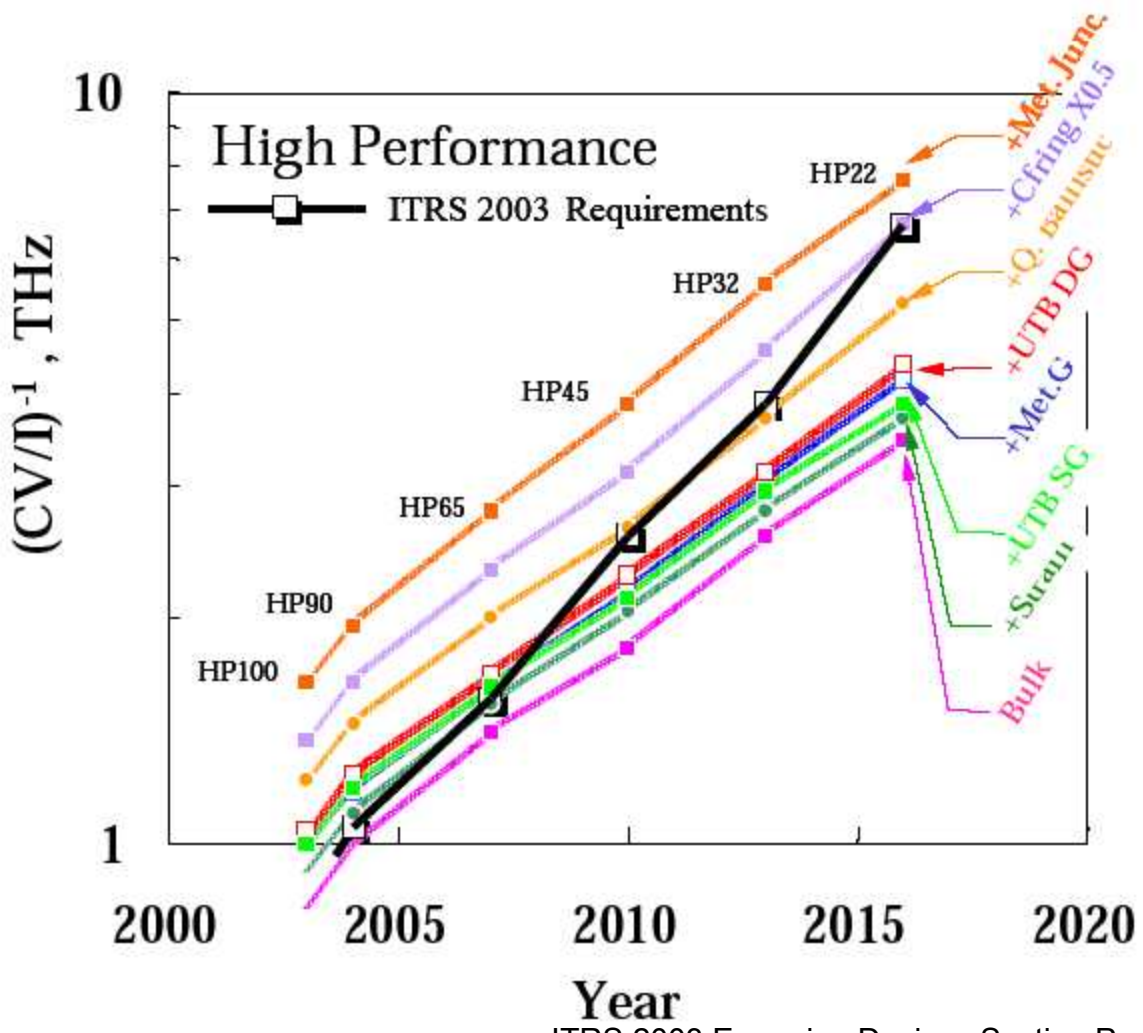
ITRS Transistor Geometries

Transport-enhanced FETs	*Ultra-thin Body SOI FETs*		*Source/Drain Engineered FETs*	
				
Strained Si, Ge, SiGe, SiGeC or other semiconductor; on bulk or SOI	Fully depleted SOI with body thinner than 10 nm		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



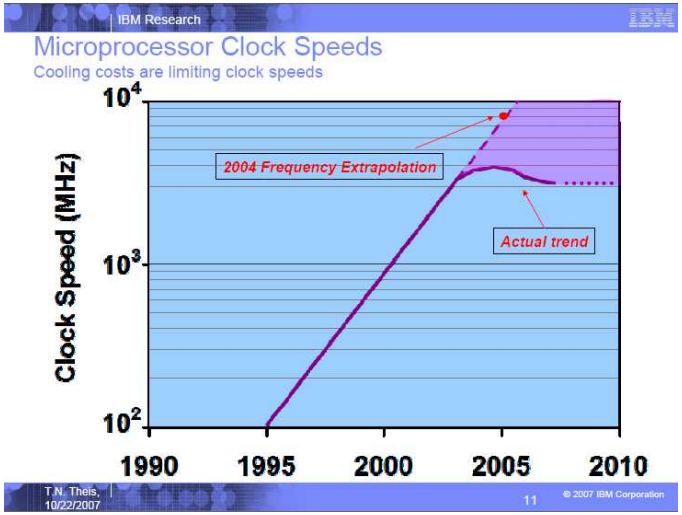


Agenda

- Gordon Moore's 1965 Paper
- International Technology Roadmap for Semiconductors (ITRS)
  - Innovations
  - Power and Clock Rate
  - System Performance
  - Interconnect
- Faster computing with the power turned off
- Exotic: Reversible Logic
- Conclusions

Clock Rate Flat Lined

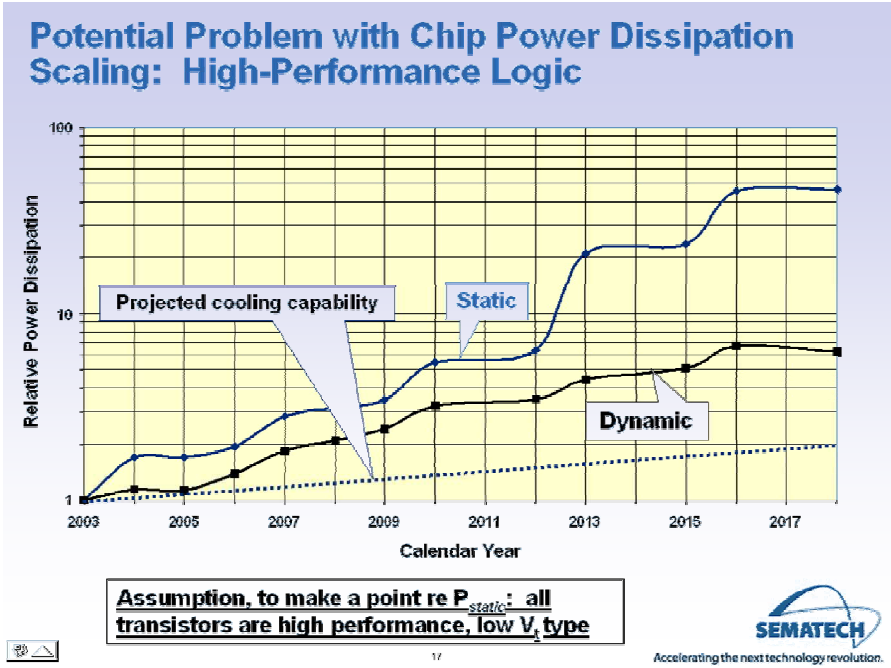
- Clock rate flat lined a couple years ago, as vendors put excess resources into multiple cores
- This is a historical fact and evident to everybody, so there is little reason to comment on the cause
- However, it has profound architectural consequences (later slide)



Sandia National Laboratories

Power Dissipation

- By targeting a smooth exponential increase in performance over time, power dissipation becomes a dependent variable
- Power dissipation per μP chip is not a reported parameter
- Chart shows result



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

Processor Clock Rate

- Processor operating frequency 10 gate delays with 30% latch overhead
- Gate delay assumes FO3, 2× parasitic capacitance
- Gate delay assumes CV^2 charging, hence supply voltage dependence
- However, these are gate level, not system level

ITRS 2003 supplementary material

SUM = =(1/(E94*E92*(1+E95/100)))*1000

	A	B	C	D	E
1	HP PIDS Worksheet Version: Aug 04, 2003 -01	Units	Variables	Parameter Equations (All variables assumed to be of similar dimensional units)	Spreadsheet Con Jim Chung (508) Peter Zeitzoff (5
2					
3	General Parameters				
4	Year in Production		Year	Parameter from ORTC	2003
5	Years in Future		Delta-year	Delta-year = Year - 2003	0
6	Technology Generation		Node	Parameter from ORTC	
92	Nominal Gate Delay (NAND Gate)	ps	Tau-NAND	Tau-NAND = Tau-inverter * Param-NAND-log-eff * Param-NAND-ele-ef	30
93	Nominal Gate Delay Scaling Target	ps	Tau-NAND-target	Tau-NAND-target = Base-NAND / (1 + Yearly-rate) ^ Delta-ye	30
94	Nominal Processor Gate Delays per Cycle		Param-gate-cycle	User-Specified Input Parameter	10
95	Latch Overhead Percentage of Cycle Time	%	Param-latch-overhead	User-Specified Input Parameter	30
96	Nominal HP Processor Operating Frequency	GHz	Fprocessor	Fprocessor = 1 / (Param-gate-cycle * Tau-NAND * (1 + Param-latch-ov	=(1/(E94*E92*(
	Nominal HP Processor Operating	GHz	Fprocessor-target	Fprocessor-target = Base-freq * (1 + Yearly-rate) ^ Delta-year	2.5

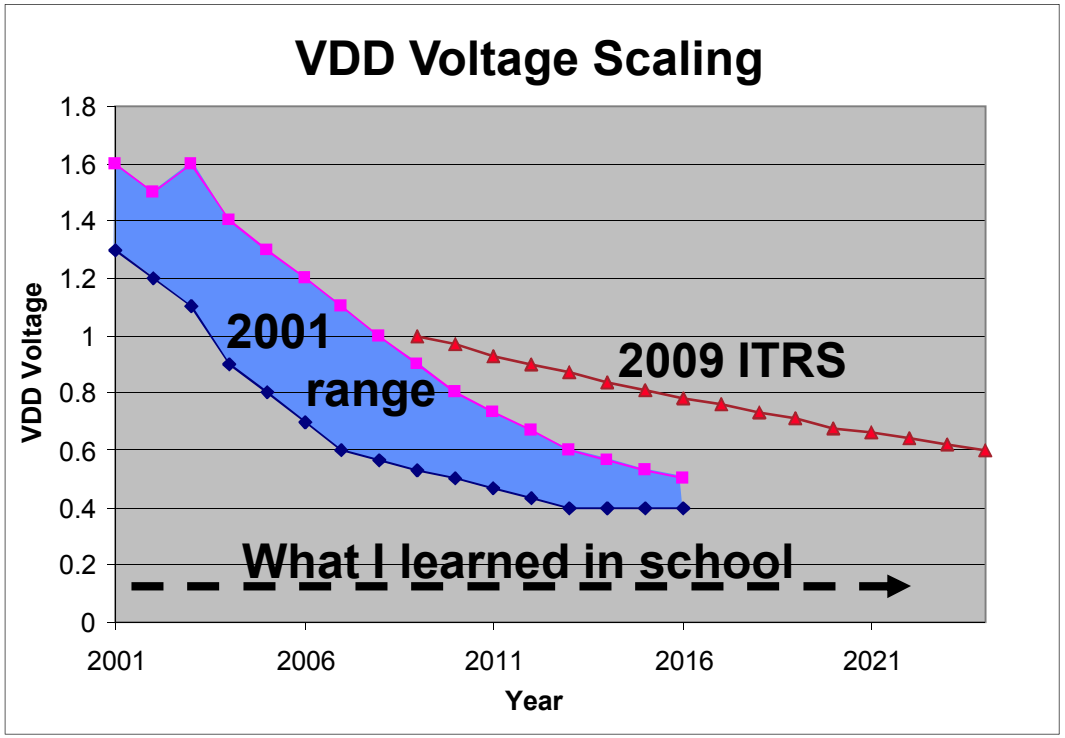


ITRS Scaling Conclusions

- Optimism
  - Density doubles every three years
    - 26% per year
  - Clock rate rises 17% per year
  - Sum is 43%/year!
    - Reasonably close to the 41%/year of ideal scaling!
- Limits of Applicability
  - Power dissipation partially covered
    - However, power dissipation per chip rises
    - Leakage power not covered
  - Timing based on gates, not architecture
    - Wiring delay calculated, but not part of timing model

ITRS Voltage Scaling

- When I went to school in the 1970s, VDD was project to expected to drop to ~130mv
- ITRS was predicting .4-.5 volts as of 2001
- Current predictions are much higher
- Energy is $\frac{1}{2}CV^2$



Data (Backup)

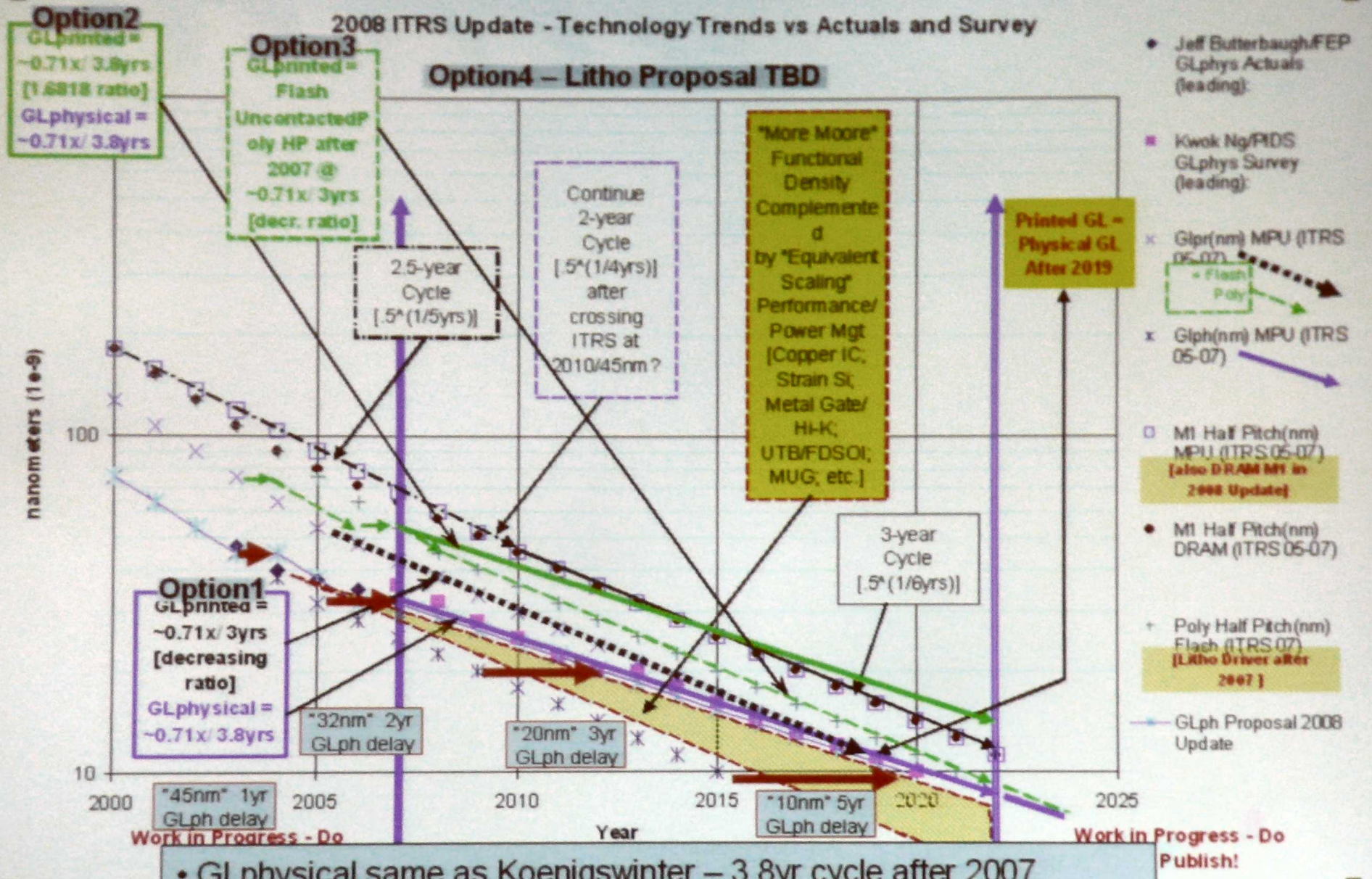
ITRS CMOS scaling is the result of device simulations using a program called MASTAR coupled with computer system models. Everything evolves.

YEAR OF PRODUCTION	2001	2002	2003	2004	2005	2006	2007	2010	2013	2016
DRAM ½ PITCH (nm)	130	115	100	90	80	70	65	45	32	22
MPU / ASIC ½ PITCH (nm)	150	130	107	90	80	70	65	50	35	25
MPU PRINTED GATE LENGTH (nm)	90	75	65	53	45	40	35	25	18	13
MPU PHYSICAL GATE LENGTH (nm)	65	53	45	37	32	28	25	18	13	9
Physical gate length high-performance (HP) (nm) [1]	65	53	45	37	32	28	25	18	13	9
Equivalent physical oxide thickness for high-performance T_{ox} (EOT)(nm) [2]	1.3–1.6	1.2–1.5	1.1–1.6	0.9–1.4	0.8–1.3	0.7–1.2	0.6–1.1	0.5-0.8	0.4-0.6	0.4-0.5
Gate depletion and quantum effects electrical thickness adjustment factor (nm) [3]	0.8	0.8	0.8	0.8	0.8	0.8	0.5	0.5	0.5	0.5
T_{ox} electrical equivalent (nm) [4]	2.3	2.1	2.0	2.0	1.9	1.9	1.4	1.2	1.0	0.9
Nominal power supply voltage (V_{dd}) (V) [5]	1.2	1.1	1.0	1.0	0.9	0.9	0.7	0.6	0.5	0.4

Table ORTC-6 Power Supply and Power Dissipation

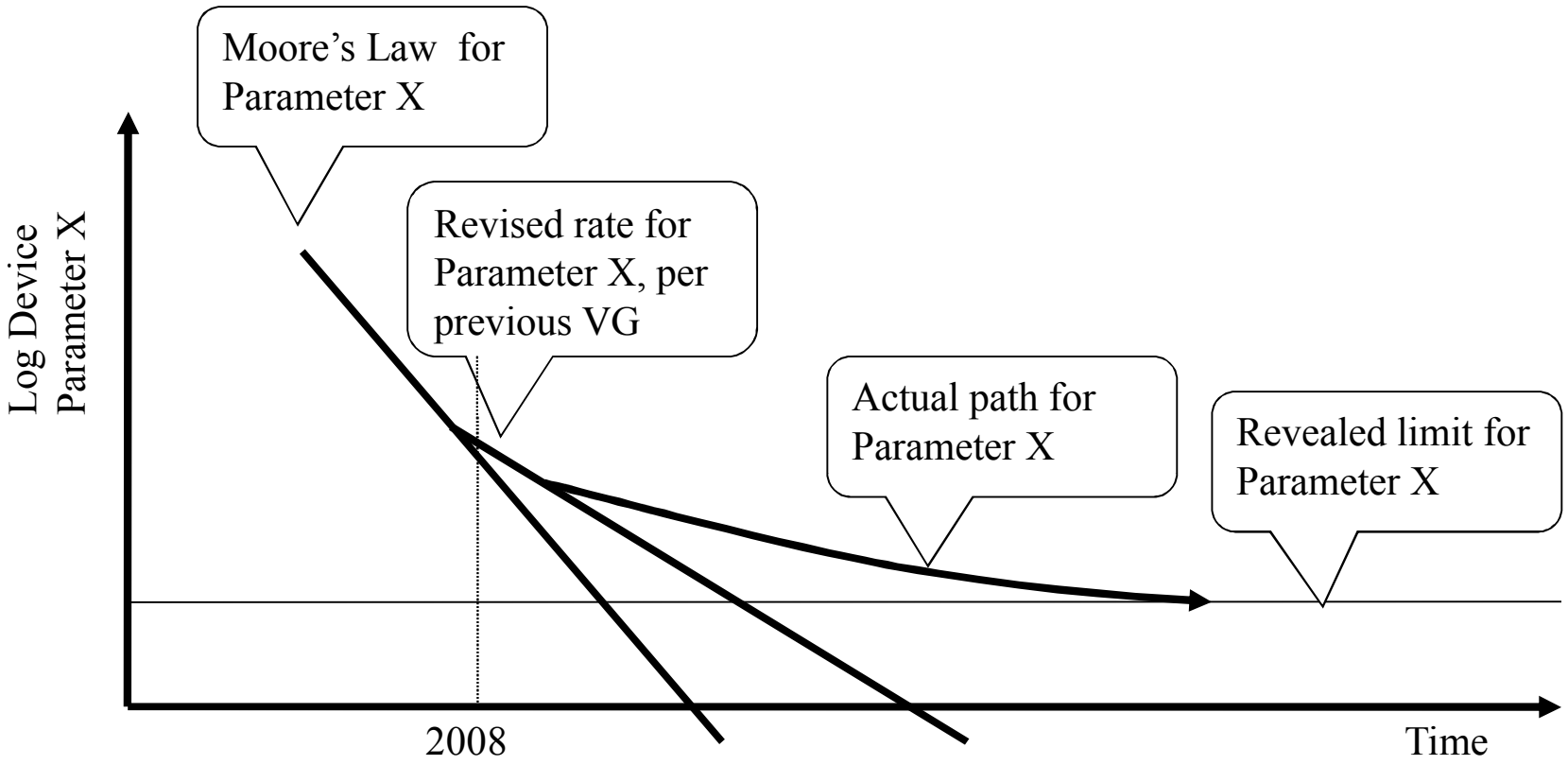
Year of Production	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Flash ½ Pitch (nm) (un-contacted Poly) (f)	38	32	28	25	23	20	18	15.9	14.2	12.6	11.3	10.0	8.9	8.0	7.1	6.3
DRAM ½ Pitch (nm) (contacted)	52	45	40	36	32	28	25	22.5	20.0	17.9	15.9	14.2	12.6	11.3	10.0	8.9
MPU/ASIC Metal 1 (M1) ½ Pitch (nm)	54	45	38	32	27	24	21	18.9	16.9	15.0	13.4	11.9	10.6	9.5	8.4	7.5
MPU Printed Gate Length (GLpr) (nm) ††	47	41	35	31	28	25	22	19.8	17.7	15.7	14.0	12.5	11.1	9.9	8.8	7.9
MPU Physical Gate Length (GLph) (nm)	29	27	24	22	20	18	17	15.3	14.0	12.8	11.7	10.7	9.7	8.9	8.1	7.4
Power Supply Voltage (V)																
Vdd (high-performance)	1.0	0.97	0.93	0.9	0.87	0.84	0.81	0.78	0.76	0.73	0.71	0.68	0.66	0.64	0.62	0.6
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Allowable Maximum Power [1]																
High-performance with heatsink (W)	143	146	161	158	149	152	143	130	130	136	133	130	130	130	Intentionally Blank	Intentionally Blank
Maximum Affordable Chip Size Target for High-performance MPU Maximum Power Calculation [2]	260	260	260	260	260	260	260	260	260	260	260	260	260	260	260	260
Maximum High-performance MPU Maximum Power Density for Maximum Power Calculation	0.46	0.47	0.52	0.51	0.48	0.49	0.46	0.42	0.42	0.44	0.43	0.42	0.42	0.42	0.42	0.42

Sandia National Laboratories



- GLphysical same as Koenigswinter – 3.8yr cycle after 2007
- FEP and PIDS have proposed shifted/interpolated tables
- 3 GLprinted proposal options, plus Litho proposal TBD

Interpretation of Graph





Agenda

- Gordon Moore's 1965 Paper
- International Technology Roadmap for Semiconductors (ITRS)
  - Innovations
  - Power and Clock Rate
  - System Performance
  - Interconnect
- Faster computing with the power turned off
- Exotic: Reversible Logic
- Conclusions



Scaling of Microprocessor Performance

- For a given design, performance proportional to clock rate
- However, designs change with technology
  - More transistors lead to architectures with more "instructions per clock"
  - Signal propagation (wire) delays lead to more pipelining
  - More pipelining leads to larger cache miss penalty
  - Cache miss penalty and desire to run larger programs (a. k. a. "code bloat") leads to larger caches
- Question: What is the roadmap for microprocessor performance?

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**

$$\max_{\text{engineering} \leftarrow \text{choices, architecture}}(\text{SpecFP})$$

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)	19	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	35508
	FITs/m length/cm^2 $\times 10^{-9}$ 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)*	505	970	2008
Add	*Line length where τ=RC delay (mm)*	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	66
	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 (mm)*	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 (mm) 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*	155	148	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]	5	3.5	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 →

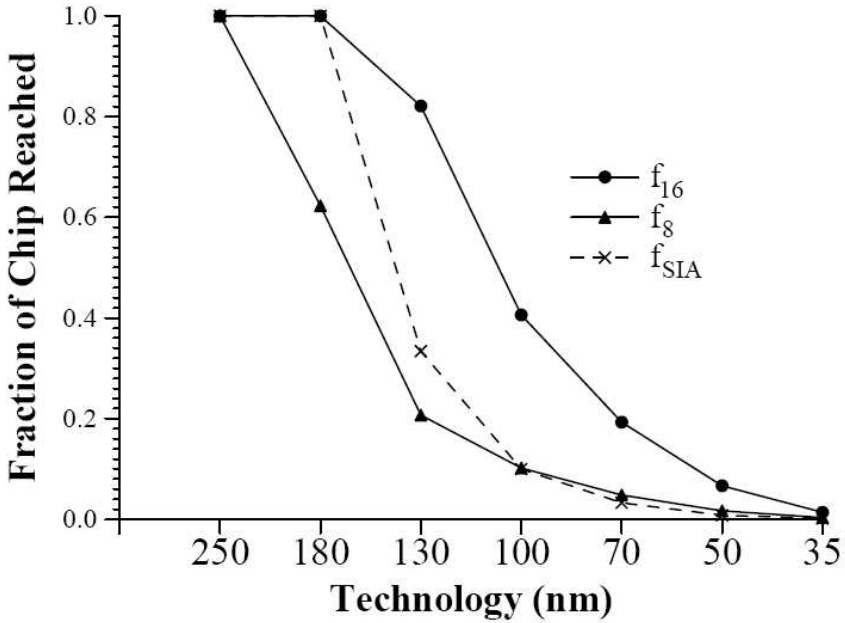


Figure 4: Fraction of total chip area reachable in one cycle.

- 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

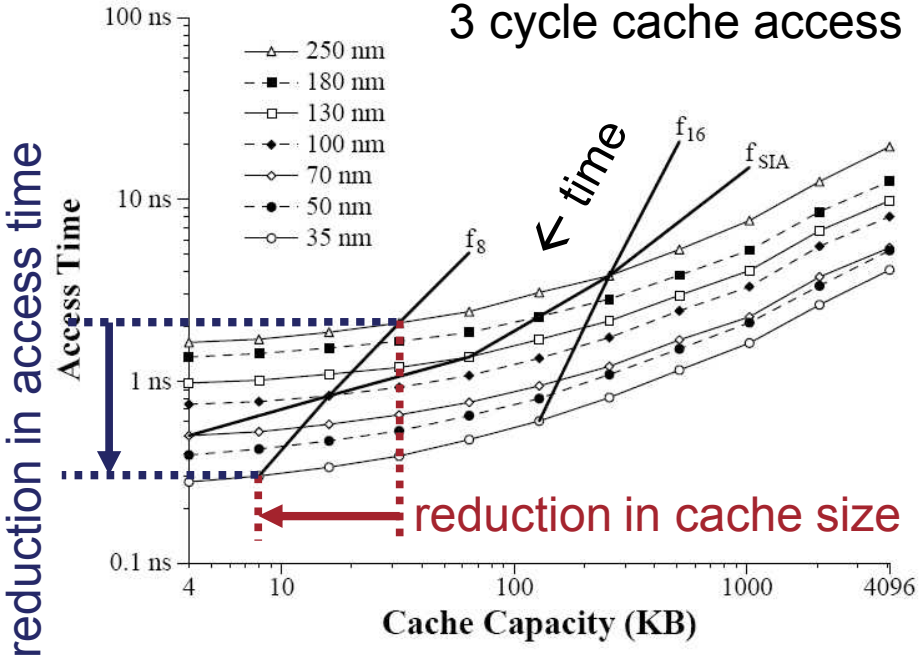


Figure 5: Access time for various L1 data cache capacities.

- **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

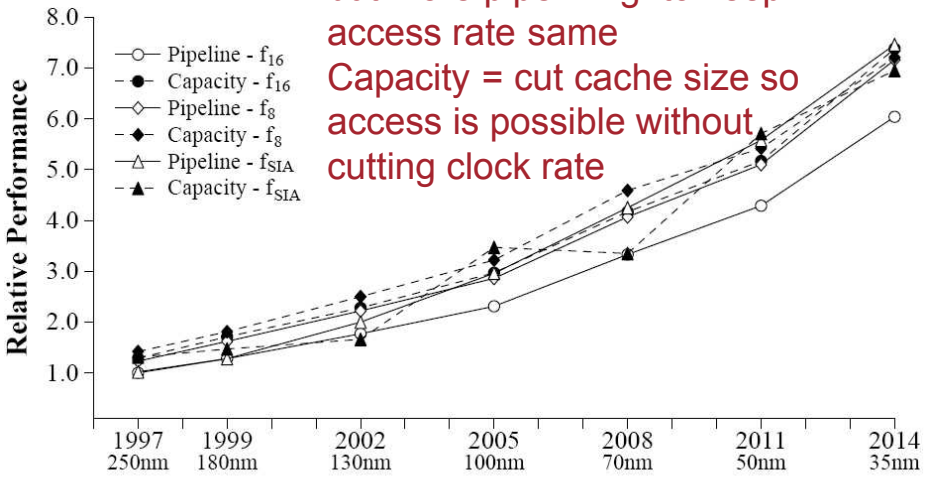


Figure 7: Performance increases for different scaling strategies.

- **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**



Study Conclusions and Discussion

- **UT Austin study concluded that μP performance should increase at about 12%/year**
- **However, it actually increased at 30%/year**
- **What is the discrepancy?**
  - **It is difficult to predict future**
  - **Vendors broke study assumptions by increasing power**
  - **Study was before its time (vendors went multicore this year)**

See Figure 8 from
"Clock Rate versus IPC: The End of the Road for Conventional Microarchitectures",
Vikas Agarwal, M.S. Hrishikesh, Stephen W. Keckler, and Doug Burger

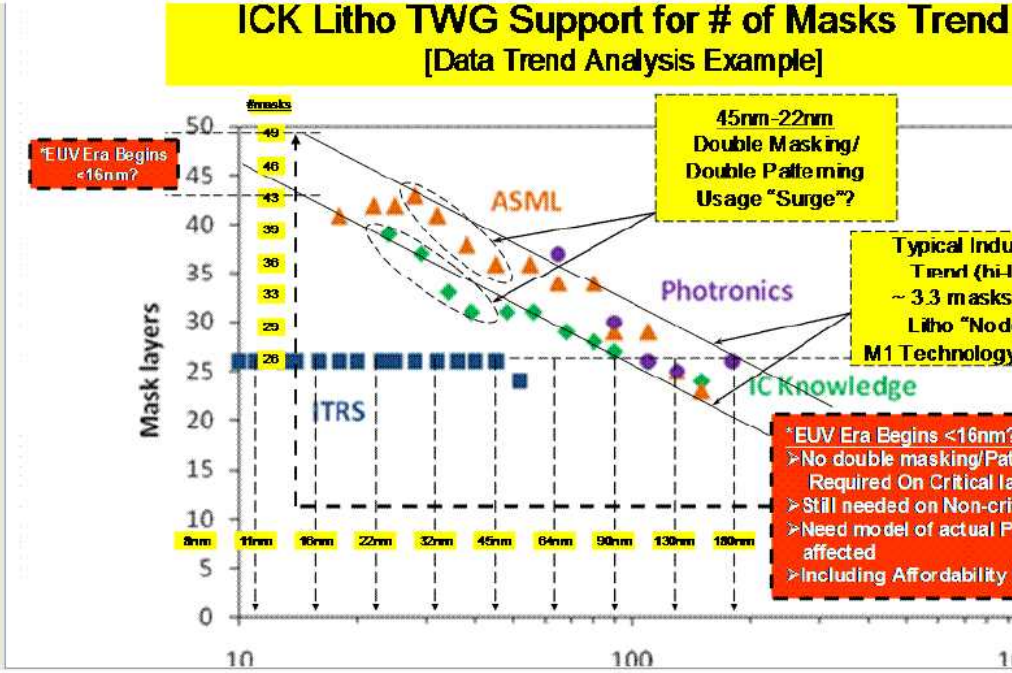


Agenda

- Gordon Moore's 1965 Paper
- International Technology Roadmap for Semiconductors (ITRS)
  - Innovations
  - Power and Clock Rate
  - System Performance
  - Interconnect
- Faster computing with the power turned off
- Exotic: Reversible Logic
- Conclusions

Evolving Moore's Law: More Layers

- **Number of masks is now expected to increase, including number of interconnect layers**
- **Per Moore's Law, power per unit area is constant – but that applies to one metal layer**
- **Another pressure to raise power**



• Preliminary feedback:

•No differentiation among different products

•No mask cost increase by node

•Write time increase

•No breakdown on "non-critical"

• Need to check pu TSMC and Intel

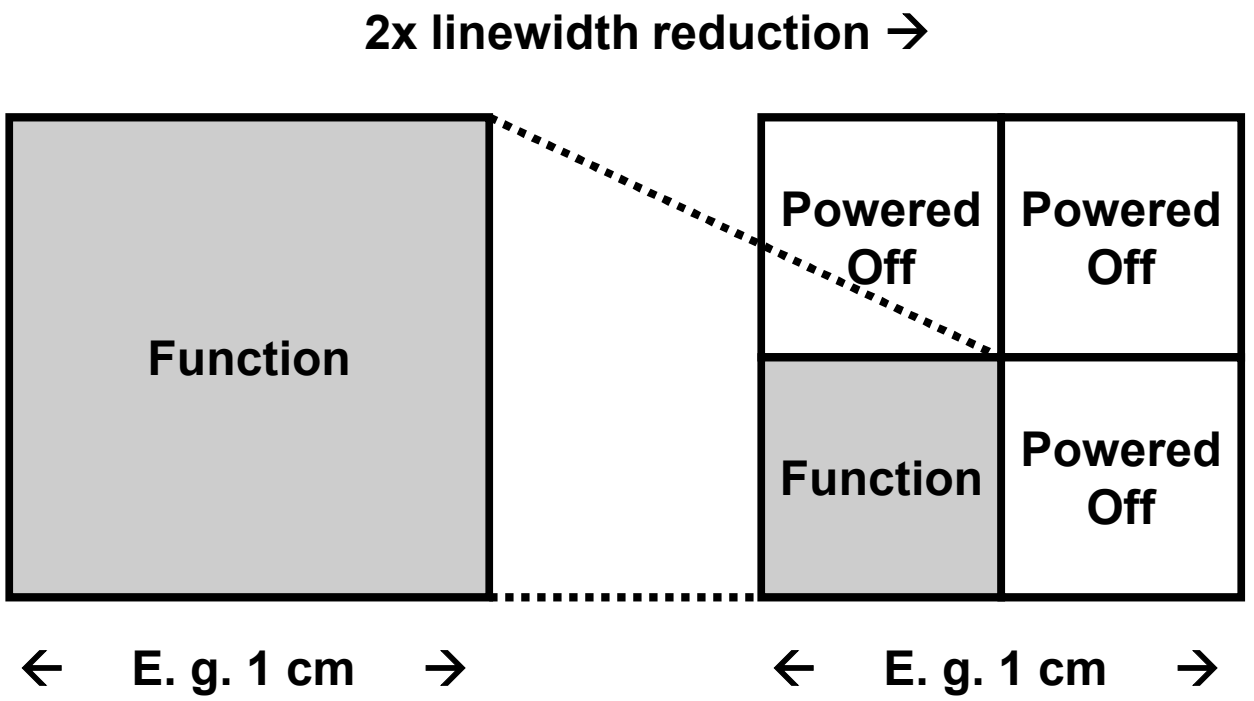
ITRS International Technology Roadmap for Semico

National Laboratories

Agenda

- Gordon Moore's 1965 Paper
- International Technology Roadmap for Semiconductors (ITRS)
  - Innovations
  - Power and Clock Rate
  - System Performance
  - Interconnect
- Faster computing with the power turned off
- Exotic: Reversible Logic
- Conclusions

Main Idea: Think About Power-Off State

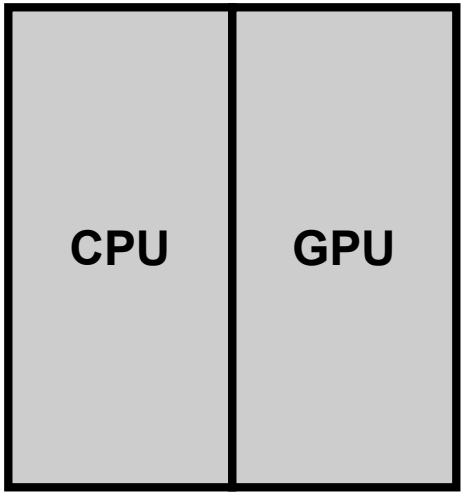


So Moore's Law gives you an additional 3x additional transistors, the complicating factor is that they must be powered off!

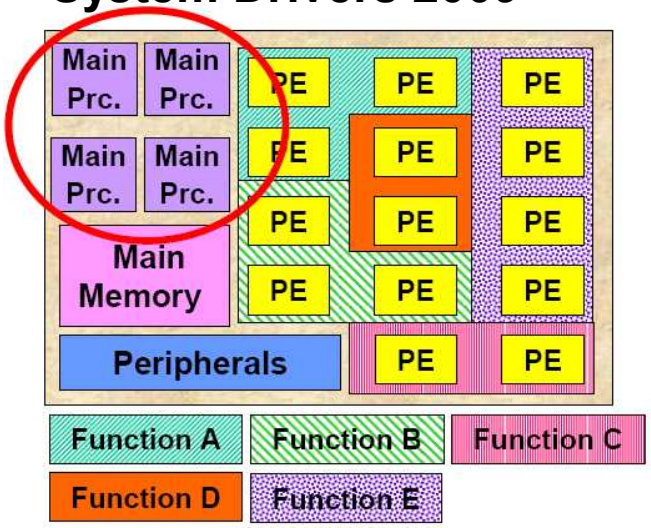
Sandia National Laboratories

How Do You Compute Faster with Powered-off Devices?

Example: CPU/GPU



From ITRS Design and System Drivers 2009



The architectures illustrated tend to be specialized to compute a limited set of functions with high power efficiency.
The ability to dynamically power up and use the most efficient architecture for a task yields a boost in overall power efficiency

Agenda

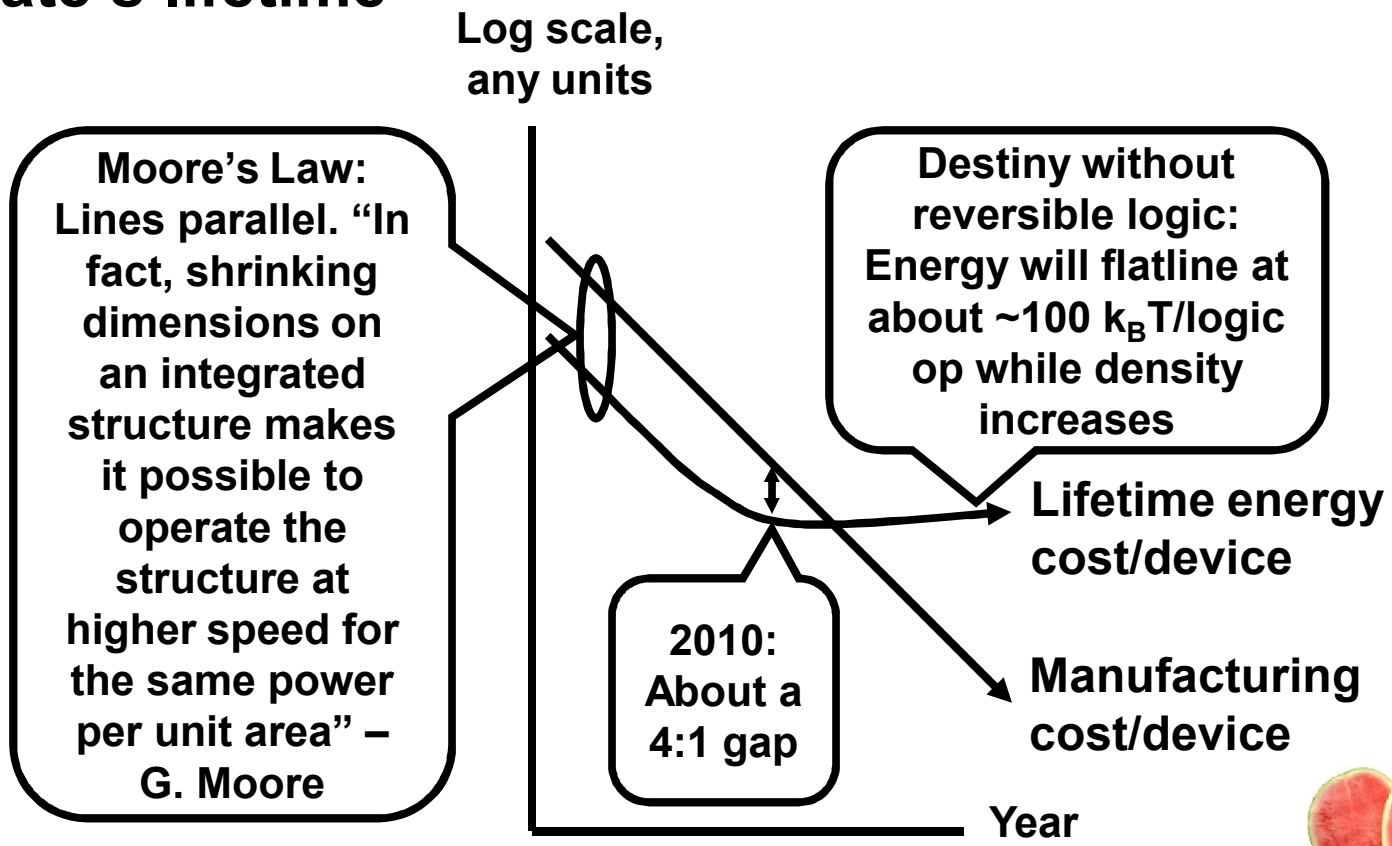
- Gordon Moore's 1965 Paper
- International Technology Roadmap for Semiconductors (ITRS)
  - Innovations
  - Power and Clock Rate
  - System Performance
  - Interconnect
- Faster computing with the power turned off
- Exotic: Reversible Logic
- Conclusions

Exotic: Reversible Logic

- Is energy scaling of logic done?
  - No: Reversible logic 10^6 (22 nm)→4 energy units/gate
  - No: Quantum computing (not discussed)
- Issues
  - There are unstated assumptions about the structure of logic and computation that make some of the exotic computing approaches too weird to understand
  - Hence, an appropriate topic for a university ??
- Technical issues
  - Low-level logic structure needs to change
  - Physical technologies different and low temperatures
  - Implications to computing, such as “uncomputing”

Roadmap for the Cost of Power vs. Device

- Cost of a gate versus the cost of energy over the gate's lifetime

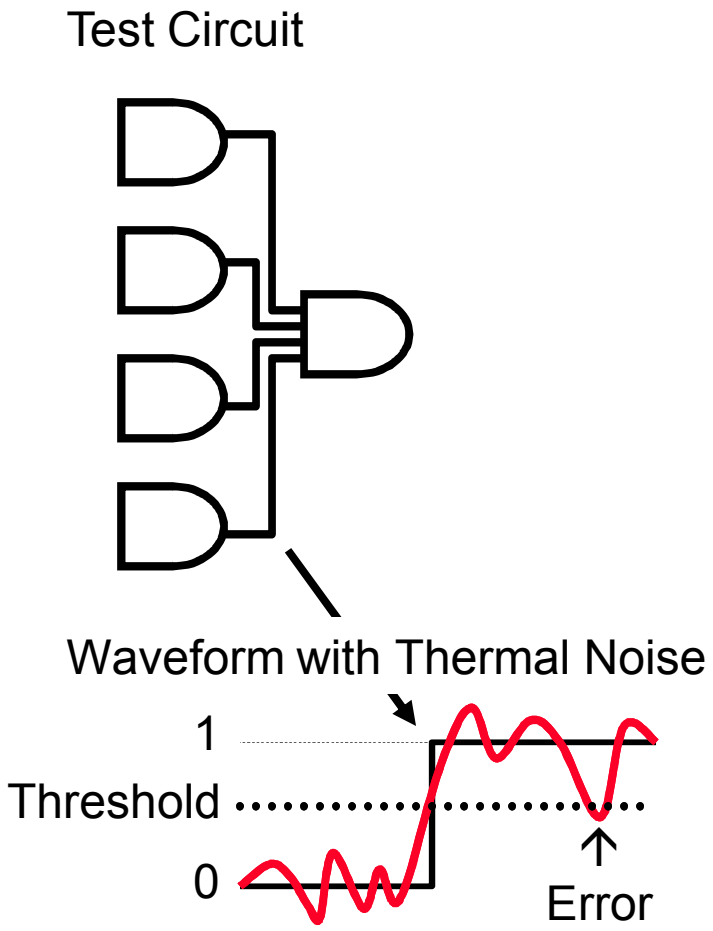


Sandia National Laboratories

$100k_BT$ Signal Energy

- The energy defining a 0 or 1 "competes" with the energy of electrons, atoms, etc. in logic gates
- The thermal energy is about k_BT, sort of
  - k_B is Boltzmann's constant $1.3806503 \times 10^{-23}$ J/K
  - T is temperature in Kelvins
  - (Actually thermal energy is $k_BT\omega$ for bandwidth ω and a typical computer system clocks at rate ω)
- The probability of thermal noise exceeding Nk_BT is about e^{-N}
- For a typical supercomputer $e^{-100} \sim 10^{-44}$ is an OK reliability

Thermal Limit Details



- **Described electrically, but applies generally**
- **Details**
  - **Noise power = $k_B T\omega$**
    - **k_B Boltzmann's constant**
    - **T temperature**
    - **ω bandwidth**
  - **An efficient design clocks every $1/\omega$ seconds**
  - **Noise power per clock period is $k_B T$**
  - **If signal energy is $Nk_B T$, probability that noise exceeds clock is about e^{-N}**

Low-Level Architecture Issues

- Today's logic circuits turn signals into heat between each logic level
  - Thermal limit requires signals to be >100 k_BT energy, otherwise they get confused with thermal noise excursions too often
  - You can't win by reducing T (more later)
- Other logic circuits reuse signal energy across multiple logic levels
  - For example, a $100k_BT$ logic signal that can traverse 50 gates before being lost is equivalent to $2k_BT$/gate
- These are called reversible logic, with subclasses
  - Ballistic logic
  - Partially reversible logic

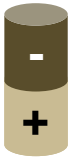
Diagram on 100k_BT Issue

- Signal energy must be more than about 100 k_BT
- The question is whether signal energy can be recycled

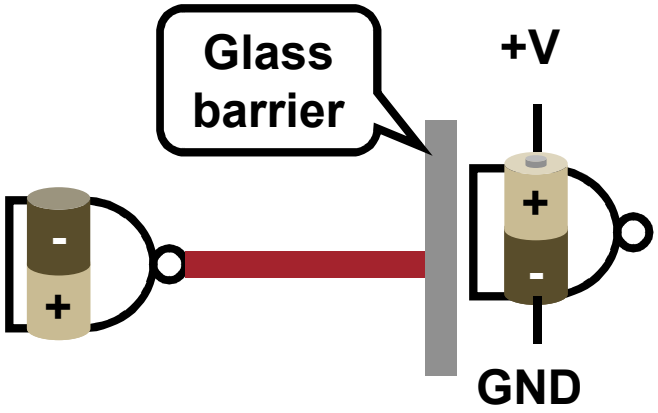
Legend:

Logic 0

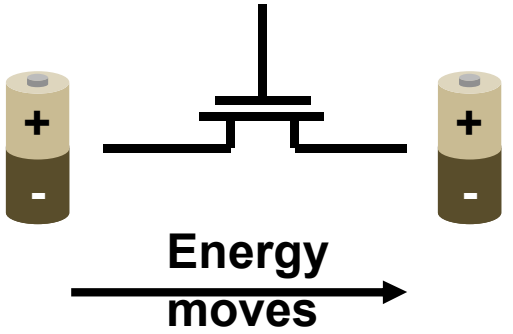
Logic 1



Logic principle in CMOS and all logic in production today: Signal energy regenerated at every logic level



Ballistic logic principle in some memories and logic in laboratory: Signal energy passed between logic levels



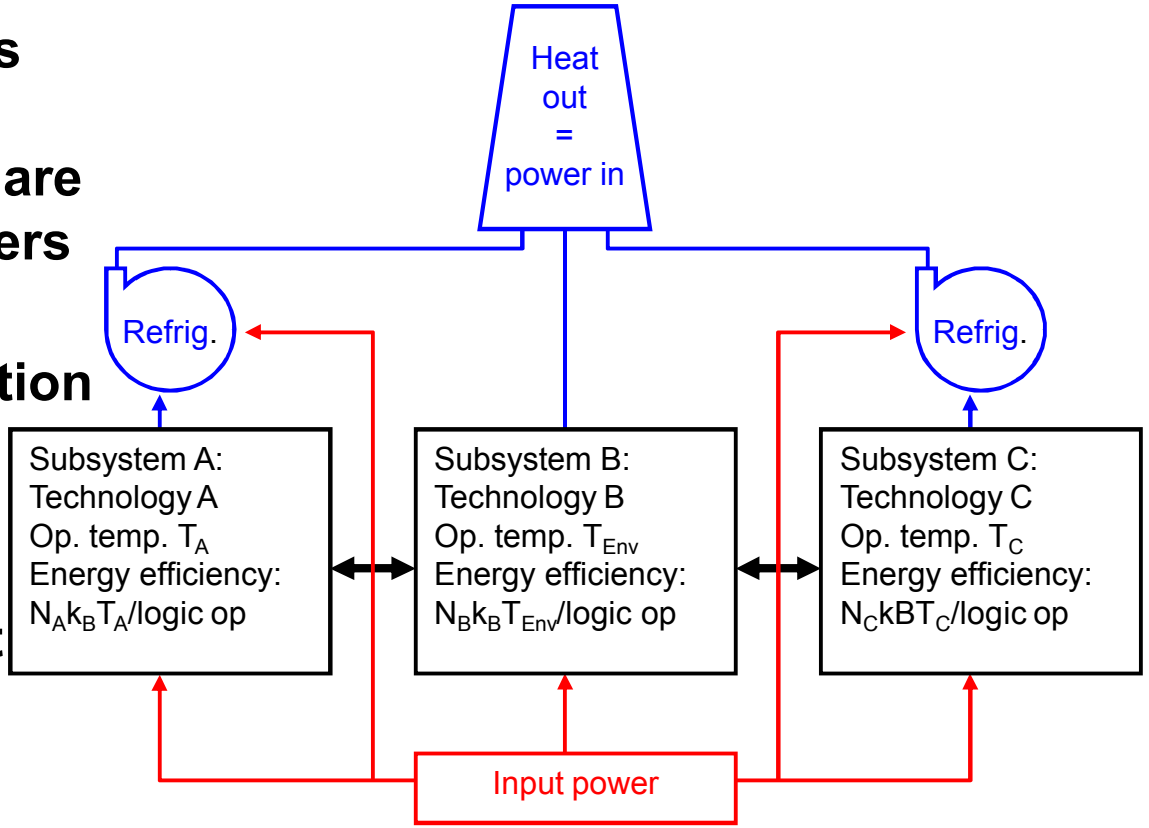
Sandia National Laboratories

Ballistic Logic Issue

- Ballistic logic leads to reversible logic
  - Reversible logic requires logical reversibility
    - NAND and NOR are NOT logically reversible
      - So, learn about Toffoli and Fredkin gates…

Theory: Temperature

- **"Carnot efficiency" is applies directly to refrigeration, but we are dealing with computers**
- **We seek maximum computing as a function of wall power, which is the same as total energy dissipated into the environment at 300K**



Sandia National Laboratories

Theory: Use Coefficients of k_BT

- Assume availability of perfect Carnot-efficient refrigerators
- A technology dissipating Nk_BT_{low} per logic operation will dissipate $300/T_{low}$ Nk_BT_{low} heat into the environment. The T_{low}s cancel and the result is independent of T_{low}.
- Thus, if you want to compare two technologies
  - $N_1k_BT_1$ and
  - $N_2k_BT_2$,
  - just compare N_1 and N_2
- Rhetorical value of this theory
  - Superconducting technologies have been advertised as having low power, but they require a refrigerator that consumes a lot of power. It is necessary to be able to quantify the net benefit/cost in PowerPoint.
  - Compare the coefficients of k_BT.

Bottom Line on Refrigerated Computers

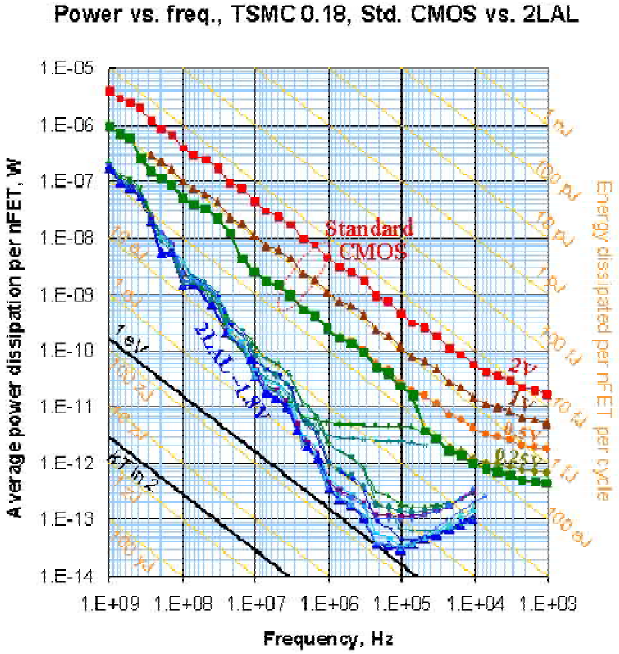
- Which logic is more power efficient?
  - Per-gate energy $N_1 k_B T_1$ at temperature T_1, or
  - per-gate energy $N_2 k_B T_2$ at temperature T_2?
  - Answer: The one with the lower N.
  - Bottom line: Use units of $k_B T$, for T=operating temperature

- Exercise readers to show this. Requires about one napkin, but not really taught in textbooks.

Current Status of CMOS

- Simulation from Mike Frank showing power advantage over conventional CMOS
  - Ignores resonators
- Shows advantage (good)
- Shows best advantage at 200 kHz (too slow)
- This is an example of a general theory involving an "entropy coefficient"

Simulation Results from Cadence



Assumptions & caveats:

- Assumes ideal trapezoidal power/clock waveform.
- Minimum-sized devices, $2\lambda \times 3\lambda$
  - * .18 μm×.24 μm
- nFET data is shown
  - * pFETs data is very similar
- Various body biases tried
  - * Higher V_{th} suppresses leakage
- Room temperature operation.
- Interconnect parasities have not yet been included.
- Activity factor (transitions per device-cycle) is 1 for CMOS, 0.5 for 2LAL in this graph.
- Hardware overhead from fully-adiabatic design style is not yet reflected
  - * >4× transistor-tick hardware overhead in known reversible CMOS design styles

Sandia National Laboratories

STONY BROOK
STATE UNIVERSITY OF NEW YORK

Reversible Computing with nSQUID Arrays

Vasili K. Semenov,

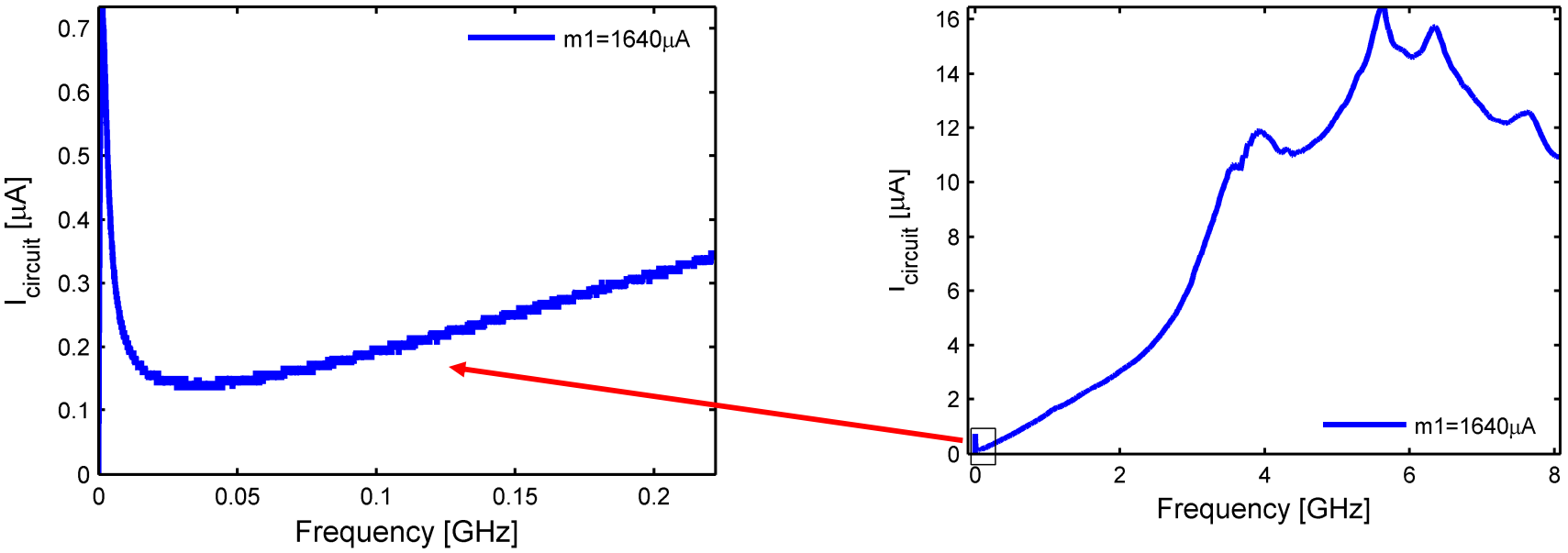
Jie Ren, Yuri Polyakov, Dmitri V. Averin

Department of Physics and Astronomy

Stony Brook University (SUNY)

This work was supported in part by the National Security Agency (NSA) under Army Research Office (ARO) contract number W911NF-06-1-217 and by JST/CREST.

Measurement of Bias Current II.



Discussion of Results

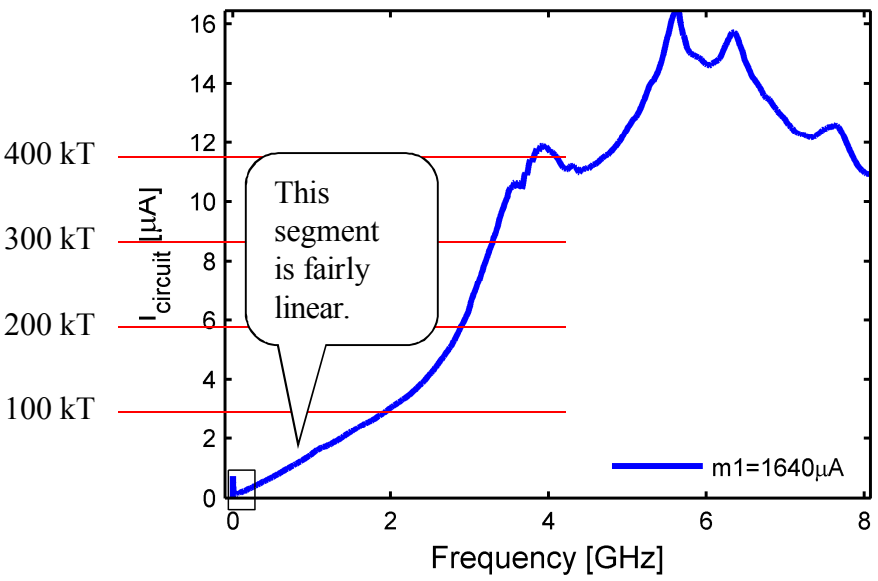
- **We measured that two shift registers operate at bias current as low as ~0.14 μA. It means one shift register consumes ~0.07 μA or 3.5 x k_BTln2!**

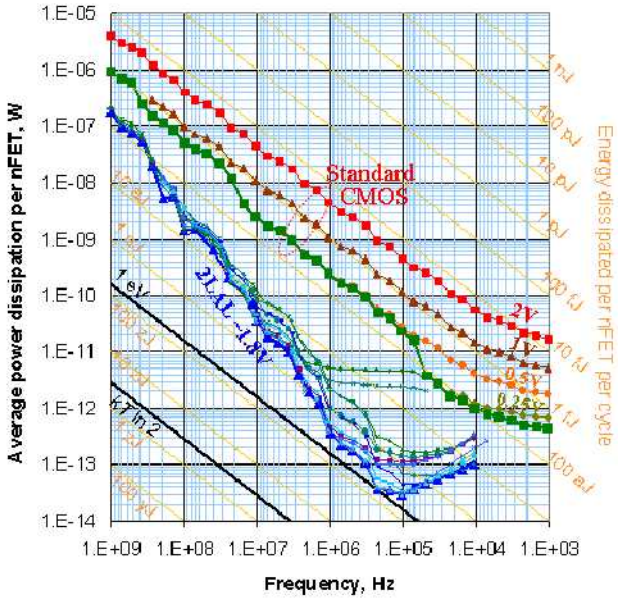
- **This figure is still above 0.02 μA thermodynamic threshold. But scaling of power dissipated in best CMOS gates (about $1.7 \cdot 10^6$ k_BTln2) converts, say, 1 mega Watt of energy to less than 3 Watts at room temperature or less than 0.1 Watt at helium temperature!**

- **It is possible to say that this remarkable result is achieved simply by removing the quantum mechanics from our prospective quantum computer architecture suggested in the framework of this project. In other words, we *experimentally illustrated* potential advantages of quantum computing.**

Speed Graphs

- **Superconducting (low R)**
- **CMOS**



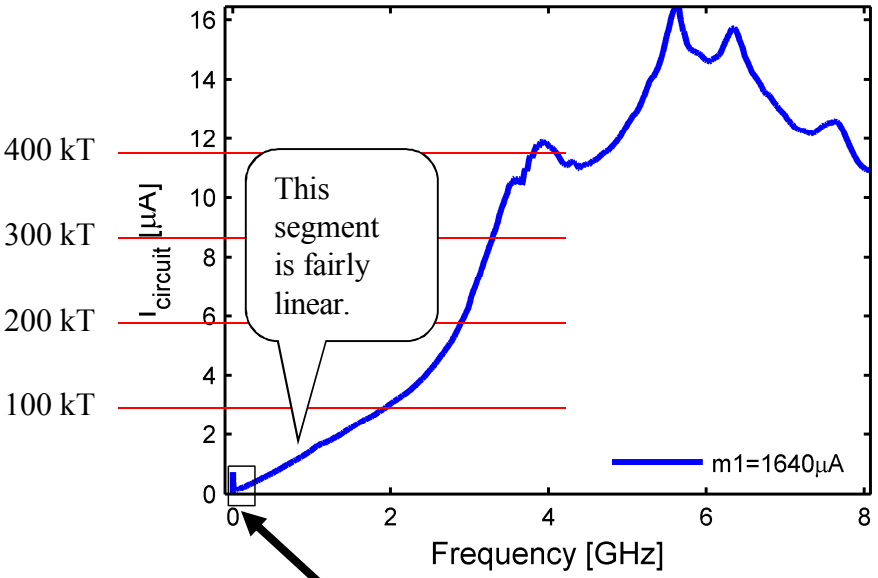


These two authors are plotting similar behaviors, but on different axes. Can we make them comparable, or more so?

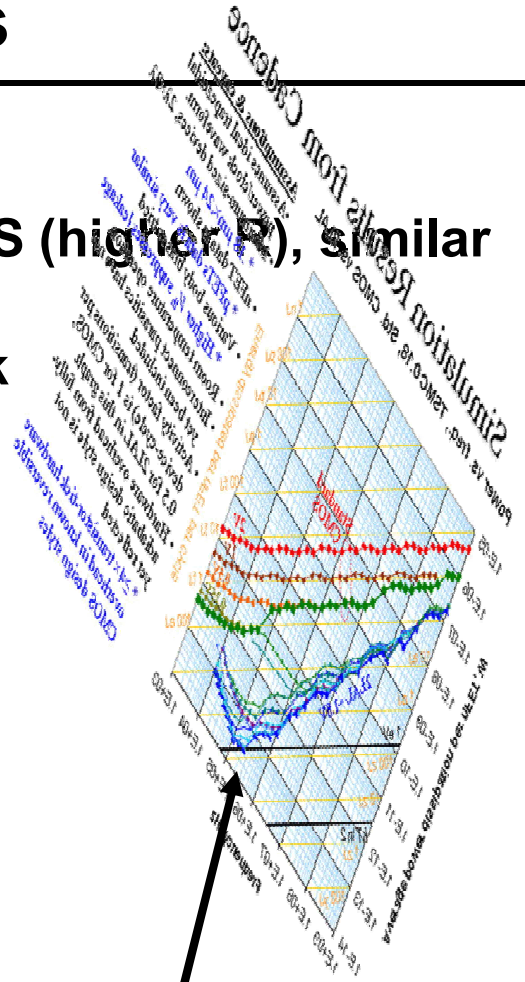
Sandia National Laboratories

Speed Graphs

- **Superconducting (low R) Semenov**



- **CMOS (higher R), similar scale Frank**



Best efficiency is the clock rate where the energy per operation is lowest

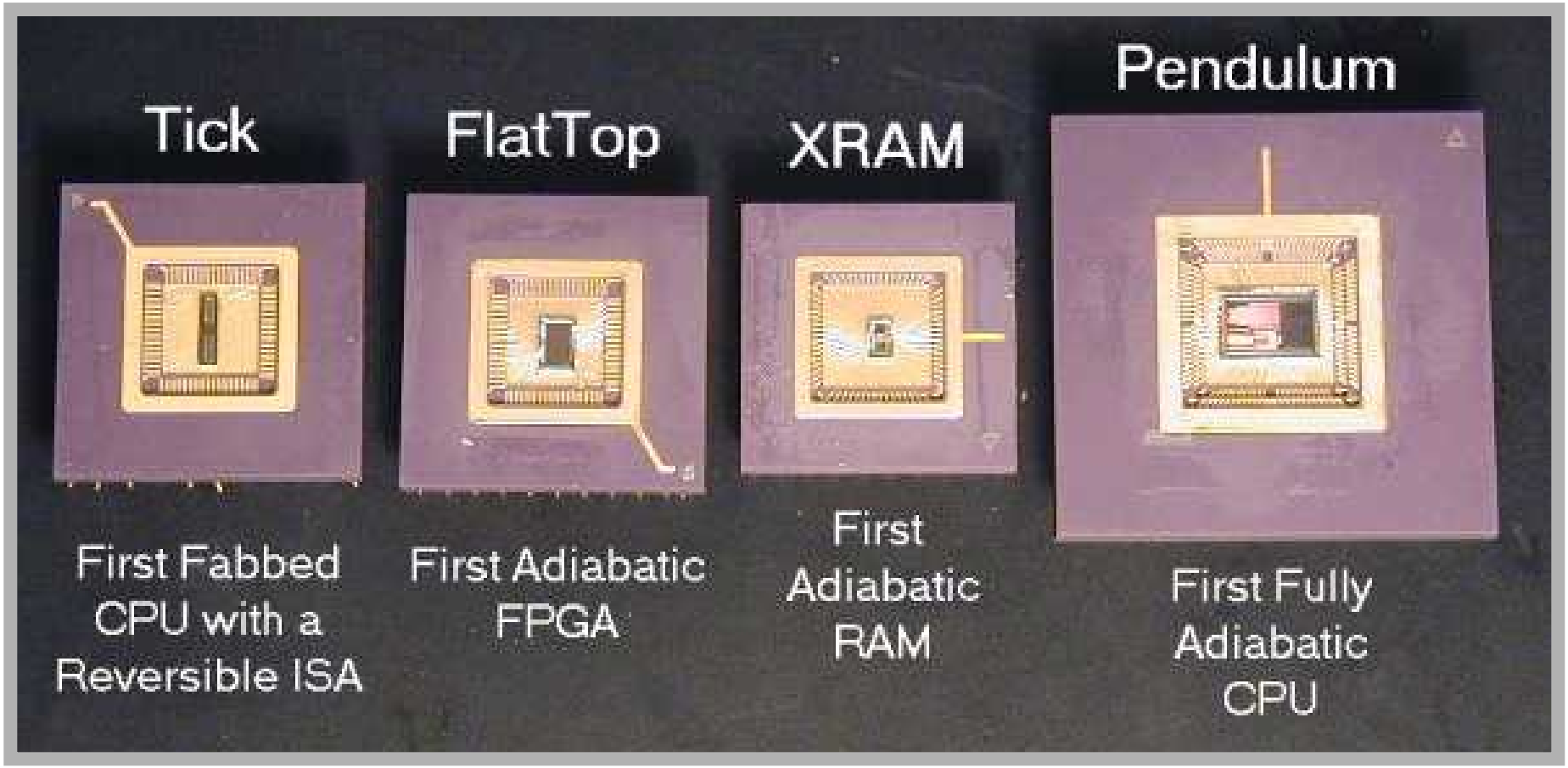
Sandia National Laboratories

Reversible Logic System Status

- In the late 1990s, there was work at MIT in developing reversible logic computers
  - They used reversible logic principles, but did not achieve power advantage
  - They were fully reversible, which overdoes it
  - MIT demonstrated architecture for CPU and memory, compilers
- Also work at Notre Dame now

Reversible / Adiabatic Chips Designed @ MIT, 1996-1999

By the author and other then-students in the MIT Reversible Computing group, under AI/LCS lab members Tom Knight and Norm Margolus.



Reversible Logic Discussion

- Optimists claim CMOS energy will scale to about 100 $k_B T$/gate operation
- Industry would like to keep scaling, but has no story for how to beat the thermal barrier
- It would be timely for universities to become interested
- Google terms: reversible logic, ballistic logic, physics of computation, Landauer's limit

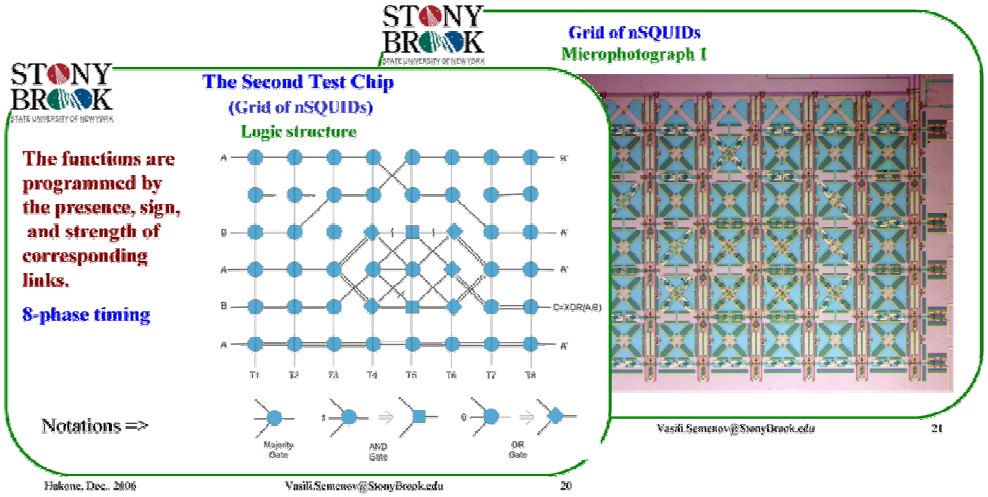
Agenda

- Gordon Moore's 1965 Paper
- International Technology Roadmap for Semiconductors (ITRS)
  - Innovations
  - Power and Clock Rate
  - System Performance
  - Interconnect
- Faster computing with the power turned off
- Exotic: Reversible Logic
- Conclusions

Conclusions

- Preferred solution not available, sorry
  - Single-threaded x86 with 1 Petahertz clock
    - Next milestone 1 Exahertz clock
  - Power ought to not be a problem
- Industry activity now power limited and optimists say it will scale to 100 k_BT/logic operation
  - Except industry will redefine Moore's Law when a better technology comes along
  - Universities are a prime place to find the new technology (how about you?)
- There are possibilities for lower power computing
  - Experimental demonstrations $10^6 \rightarrow 4\ k_BT$
    - These numbers are quantitatively imprecise

Backup



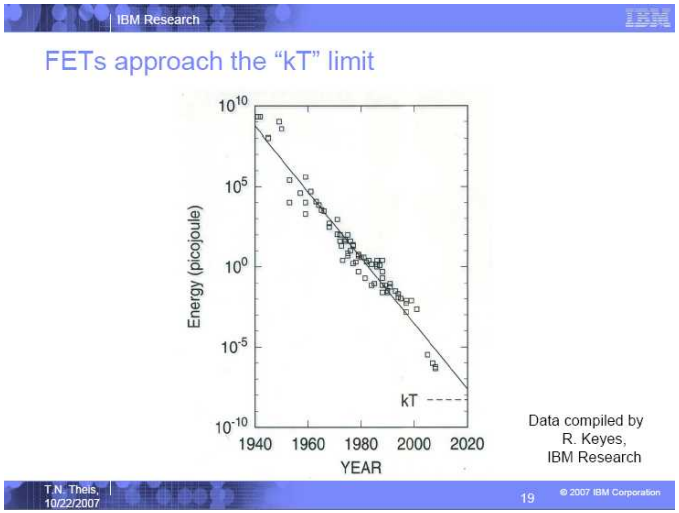
State of the art (Stony Brook): Thermodynamically reversible cells have been fabricated and organized into logic. Could they become the computational components of a supercomputer?

Logic Family:	Device energy:	Gate energy (FO4):
CMOS 22 nm	$2500k_BT$ (transistor)	$120{,}000k_BT$
CMOS 7.5 nm?	$300k_BT$ (transistor)	$14{,}000k_BT$
eSFQ	$1{,}000k_BT$ (JJ)	$9{,}000k_BT$
Reversible	$3.5k_BT$ (1 register stage)	Report objective: Can this work?

Sandia National Laboratories

$k_B T$ Limit Moderates Optimism for Perpetual Exponential Growth

- In past workshops, many participants considered Moore's Law a fact of physics that can overrule finite atom size, Landauer's limit, etc.
- Zettaflops 2007 concludes that Exaflops and Zettaflops are qualitatively different



Sandia National Laboratories