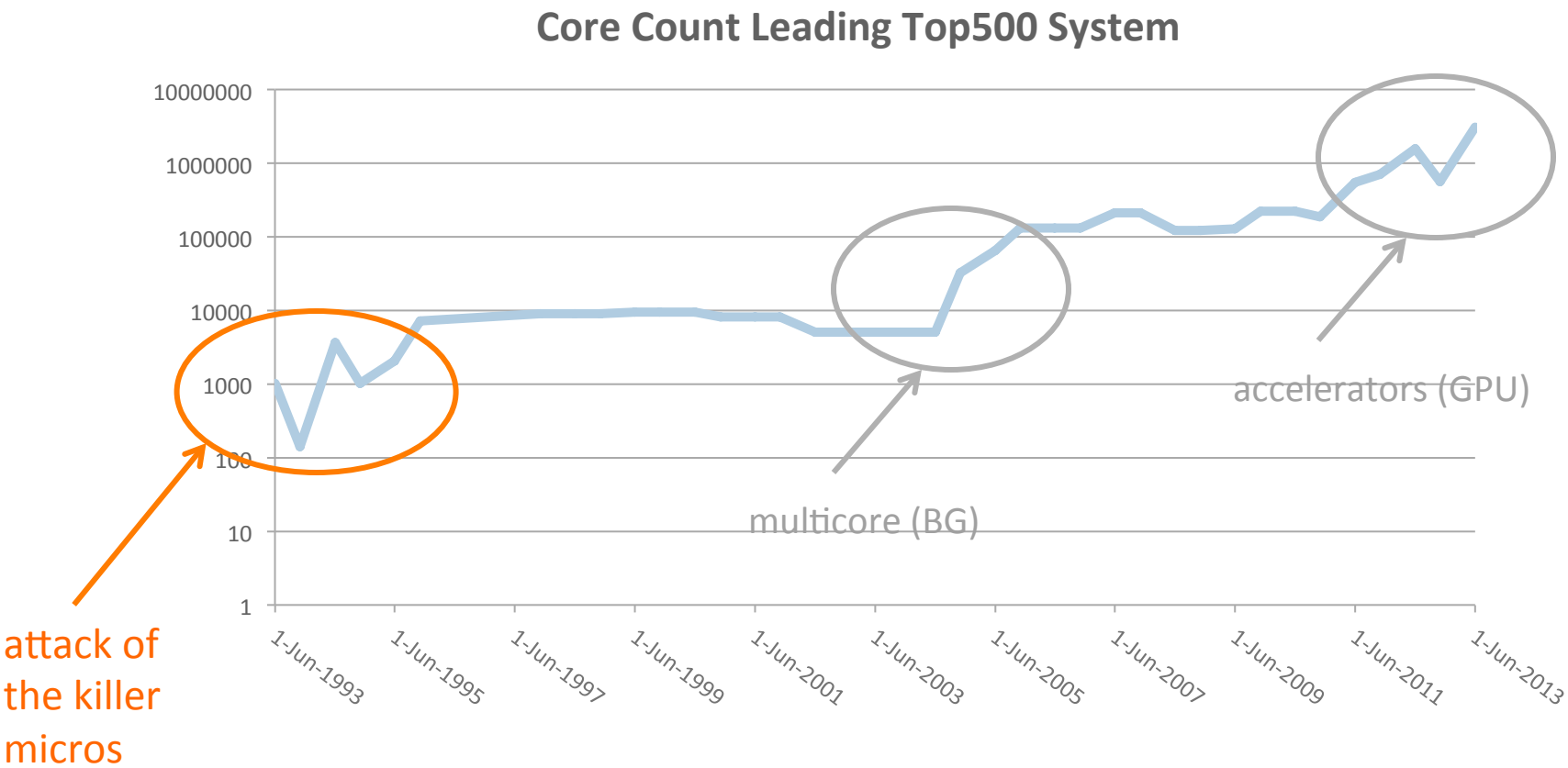
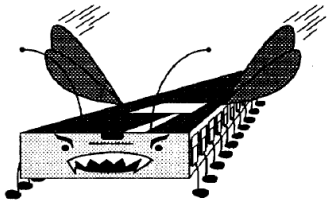
Supercomputing: The Coming Decade

Marc Snir
Argonne National Laboratory &
University of Illinois at Urbana-Champaign

Punctuated Equilibrium and Extinctions in HPC



The 1990 Big Extinction: The Attack of the Killer Micros (Eugene Brooks, 1990)

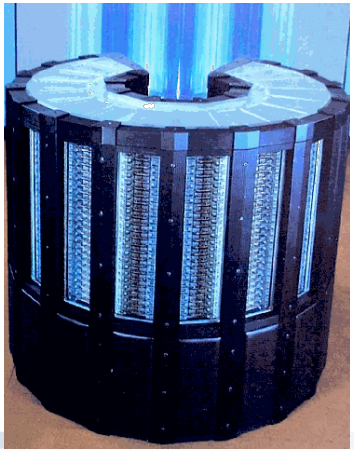


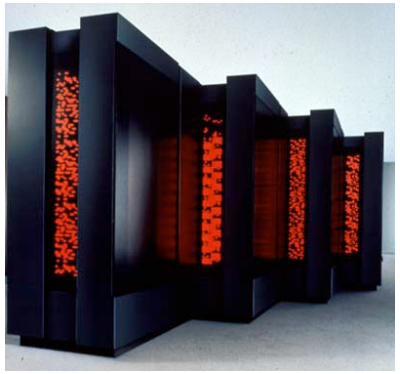
Shift from bipolar vector machines & to clusters of MOS micros

- *Roadblock*: bipolar circuits leaked too much current – it became too hard to cool them (even with liquid nitrogen)
- MOS was leaking very little – did not require aggressive cooling
- MOS was used in fast growing markets: controllers, workstations, PCs
- MOS had a 20 year history and clear evolution path ("Moore's Law")
- MOS was slower
  - Cray C90 vs. CM5 in 1991: 244 MHz vs. 32 MHz

- Perfect example of "good enough" technology

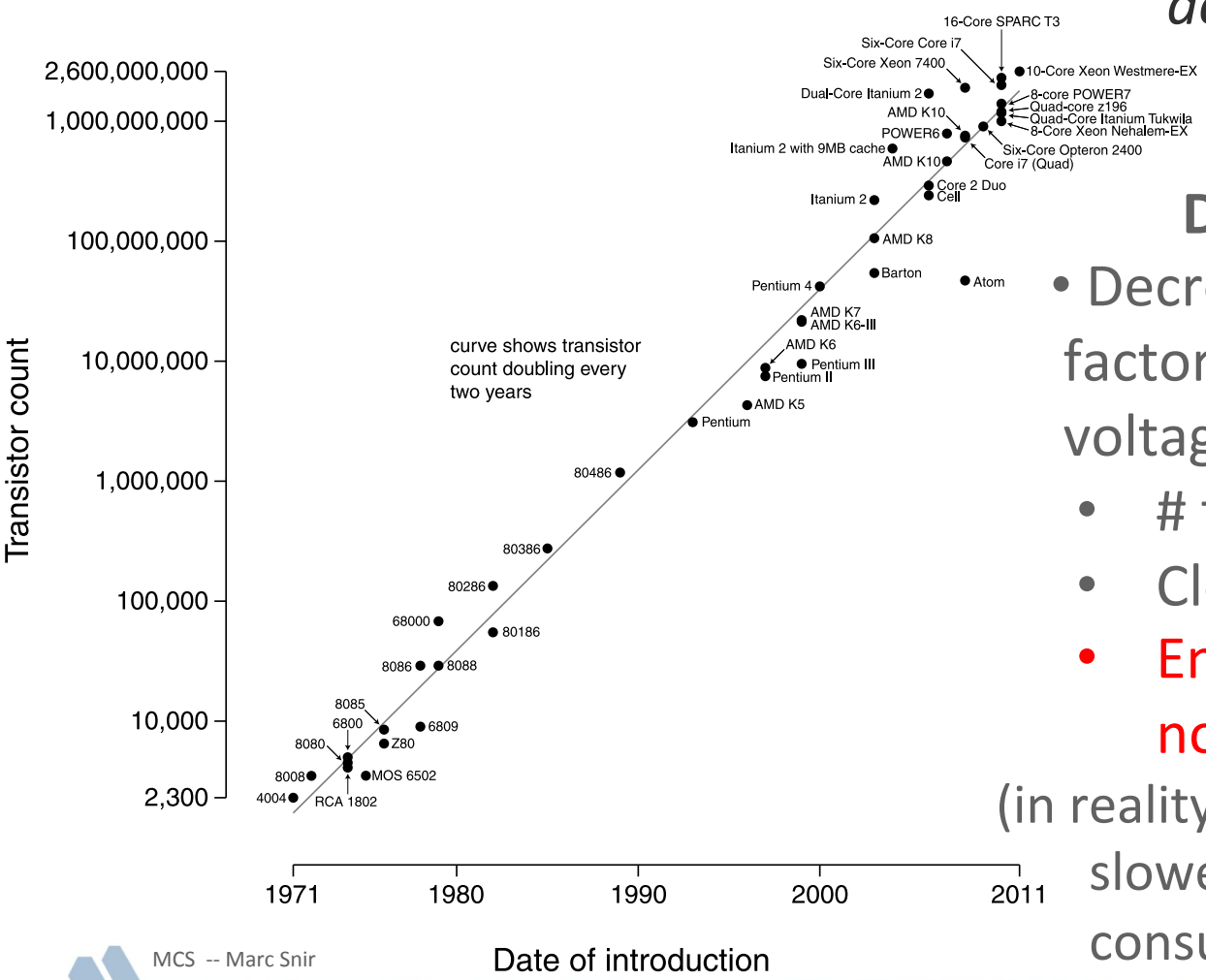
(Christensen, *The Innovator's Dilemma*)





The CMOS Age: Moore's Law & Dennard Scaling

Moore's Law: *The number of transistors per chip doubles every 2+ years*



Dennard Scaling:

- Decrease feature size by a factor of λ and decrease voltage by a factor of λ; then
  - # transistors increase by λ^2
  - Clock speed increases by λ
  - Energy consumption does not change

(in reality, voltage decrease was slower; clock speed and energy consumption increased faster)

Stein's Law: *If something cannot go forever, it will stop*

- Dennard scaling ended at around 130 nm in 2001-2004
  - Leakage (static energy) does not scale – it increases as device size shrinks
- Growth in density continues (multicore), but clock speed is (slowly) decreasing

While power consumption is an urgent challenge, its leakage or static component will become a major industry crisis in the long term, threatening the survival of CMOS technology itself, just as bipolar technology was threatened and eventually disposed of decades ago

International Technology Roadmap for Semiconductors (ITRS) 2011

- The ITRS "long term" is the 2017-2024 timeframe.

On Our Way to the Next Extinction?

- **History repeats itself:**
  - CMOS technology has hit a power wall
    - Clock speed is not raising
  - Alternative materials are not yet (?) ready (gallium arsenide and other III-V materials; nanowires, nanotubes)
- **History does not repeat itself:**
  - ✔ There is a much larger industrial base investing in continued improvements in current technologies
  - ✗ An alternative "good enough" technology IS NOT ready
  - ✗ There is much more code that needs to be rewritten if a new model is needed (>200MLOCs)

The Physical & Engineering Limits

- Transistor size cannot shrink forever
  - Need a few hundred atoms per gate
  - 5 nm is the limit for 2D (5 nm = 20 atoms)– might get denser with 3D
- Decreased return on feature size: Performance improvement is not proportional to size reduction
  - Additional spacing and larger safety margins needed to reduce interference, handle manufacturing variances, etc.
- Reduced leakage requires technology innovation
  - New materials (III/V, nanotubes…), 3D devices
- Need new light sources
  - Current 192 nm
- …

Technological challenges of high-performance logic scaling

Argonne

Year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Gate Length	**18**	**16.7**	**15.2**	**13.9**	**12.7**	**11.6**	**10.6**	**9.7**	**8.8**	**8.0**	**7.3**
Equivalent Oxide Thickness	🟢	🟢	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🔴
Source-Drain Leakage*	🟢	🟢	🟡	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🔴
Threshold Voltage*	🟢	🟡	🟡	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🔴
CV/I Intrinsic Delay	🟢	🟡	🟡	🟡	🔴	🔴	🔴	🔴	🔴	🔴	🔴
Total Gate Capacitance	🟢	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🔴
Drive Current	🟢	🟡	🟡	🟡	🔴	🔴	🔴	🔴	🔴	🔴	🔴

- Time line shown for best performing multi-gate transistor technology.
- Based on ITRS reports (2011*, 2012* and 2013 eds.)

🟢 technology available
🟡 solutions known
🔴 no known solutions

(courtesy Denis Mamaluy)

U.S. DEPARTMENT OF ENERGY

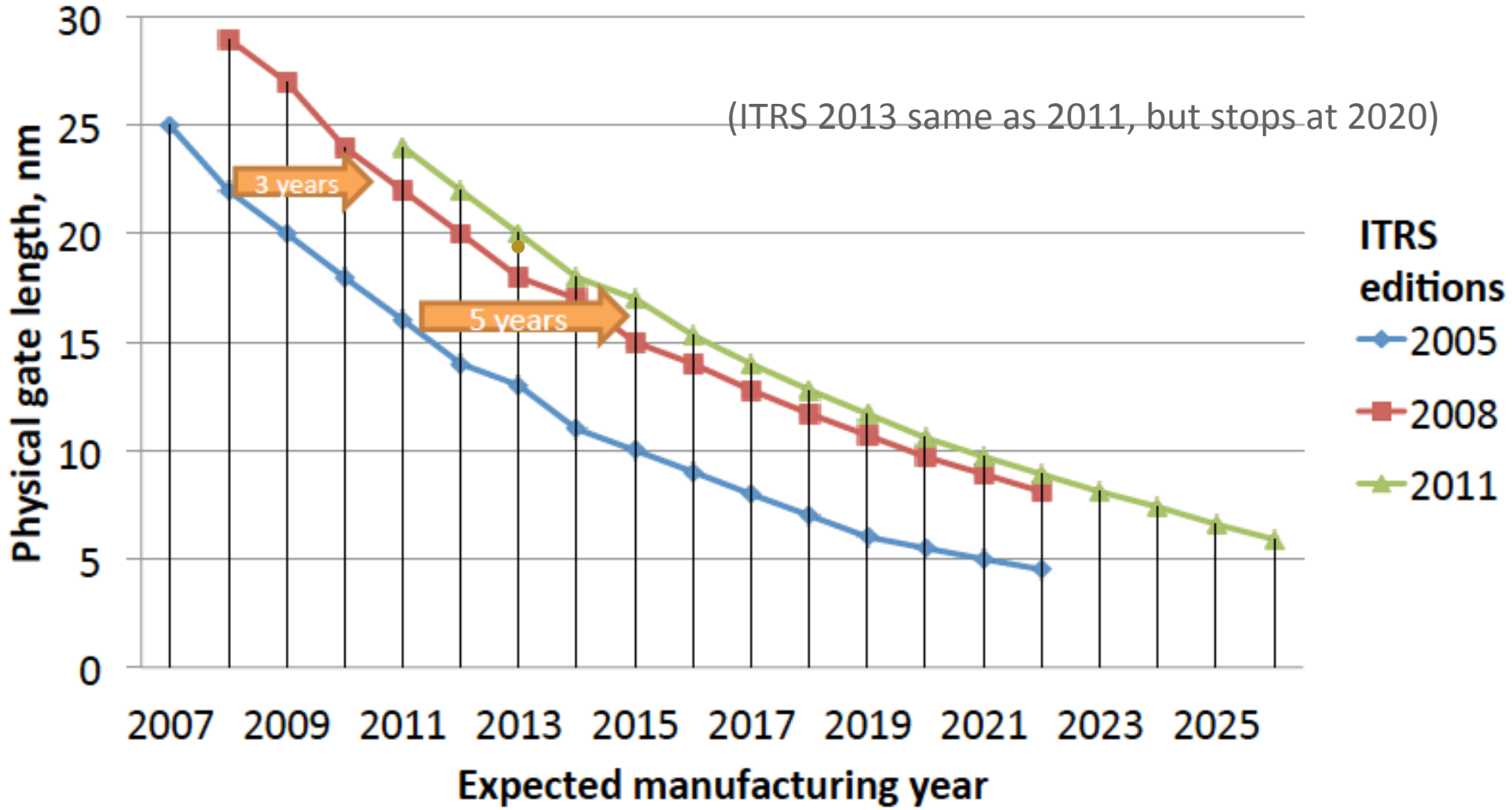
The Economical Limits

- Cost per transistor has not decreased last year
  - Market for increased performance at increased cost is very small
- Investments for new fabs keep growing, resulting in increased consolidation
  - Some predict only two vendors will be left below 22nm
- Cost of manufacturing chips keep increasing
  - More materials, more masks, more passes
- IC market cannot grow forever faster than GDP
  - Fast growth is necessary to amortize the large investments in new fabs

The Market Constraints

- Leading market for IC is mobile. The drivers in the market have little overlap with HPC.
  - ✔ low power
  - ✔ system on chip
  - ✗ small form factor
  - ✗ integration of analog and MEMS
  - ✗ limited interest in low error rate
  - ✗ no interest in 64 bit floating point and higher

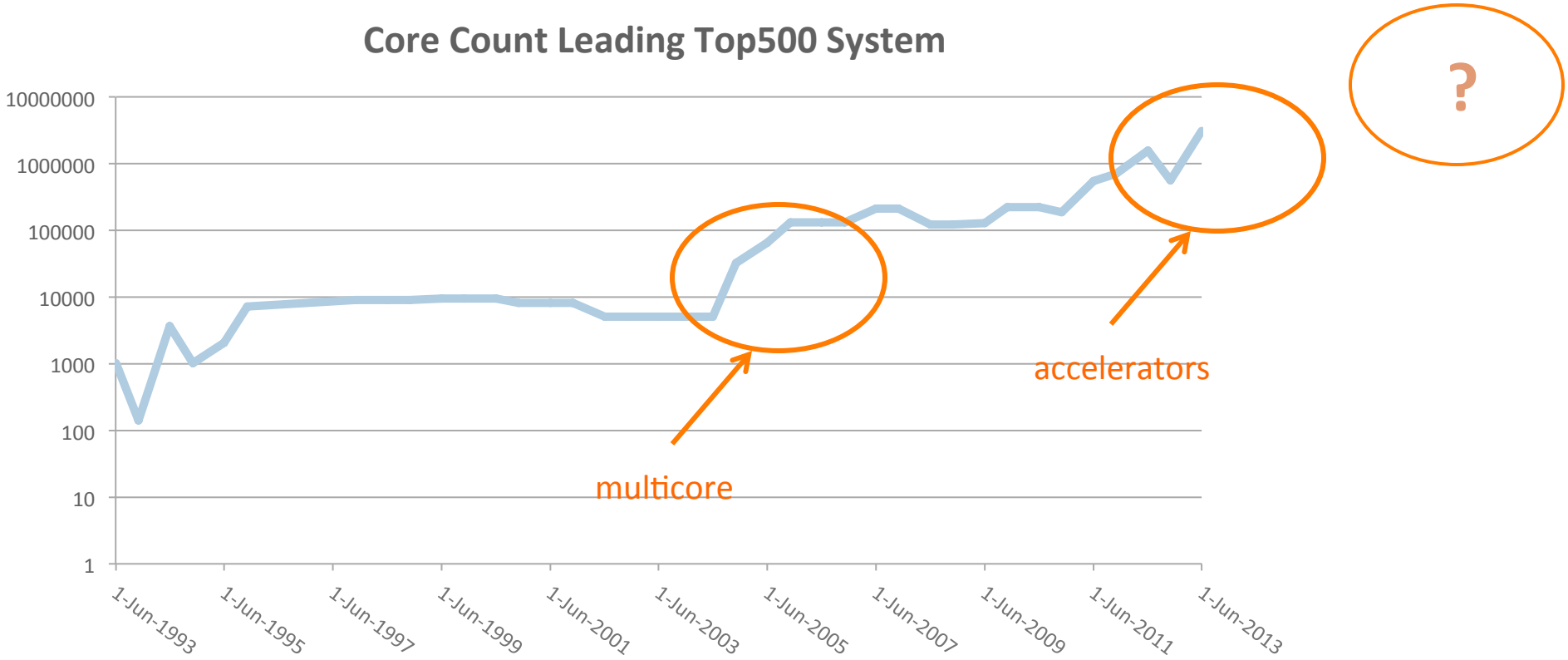
The Future Is Not What It Was



The sky is falling, but not immediately

(courtesy J. Aidun)

The Impact of the Energy Wall on HPC
What Next?



Orthogonal Scaling

- Scale at multichip package level – get move function in same volume
  - *Reduce communication cost*

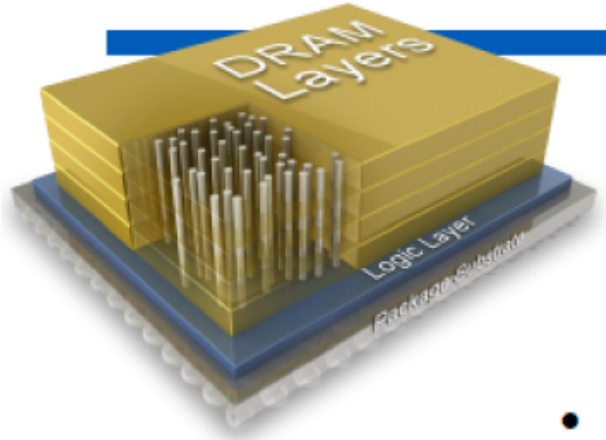
Linpack Energy Budget

- pJoules spent for one flop in Linpack ~2015 technology
  - Floating point unit consumes 10 pJ per flop
  - CPU chip consumes 475 pJ per flop
  - *Note: 1 Exaflop x 475 pj = 475 MWatt*

Step	Target	pJ	#Occurrances	Total pJ	% of Total
Read Alphas	Remote	13,819	4	55,276	16.5%
Read pivot row	Remote	13,819	4	55,276	16.5%
Read 1st Y[i]	Local	1,380	88	121,400	36.3%
Read Other Y[i]s	L1	39	264	10,425	3.1%
Write Y's	L1	39	352	13,900	4.2%
Flush Y's	Local	891	88	78,380	23.4%
Total				334,656	
Ave per Flop				475	

(courtesy P Kogge)

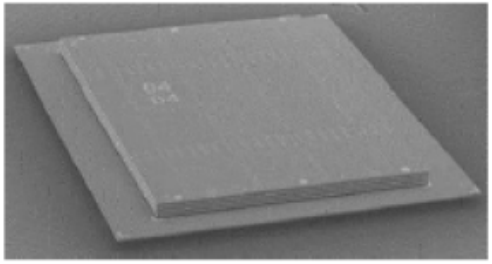
An example of orthogonal scaling: Stacked memory using HMC

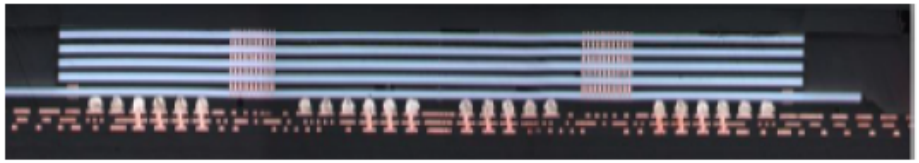


All higher power logic functions including I/Os are localized to a single logic chip on bottom

For 1.28 TB/s performance

- 85% less active signals compared to DDR3
- 90% less board space than DDR4
- 72% less power than DDR4





Source: Micron-IBM HMC development

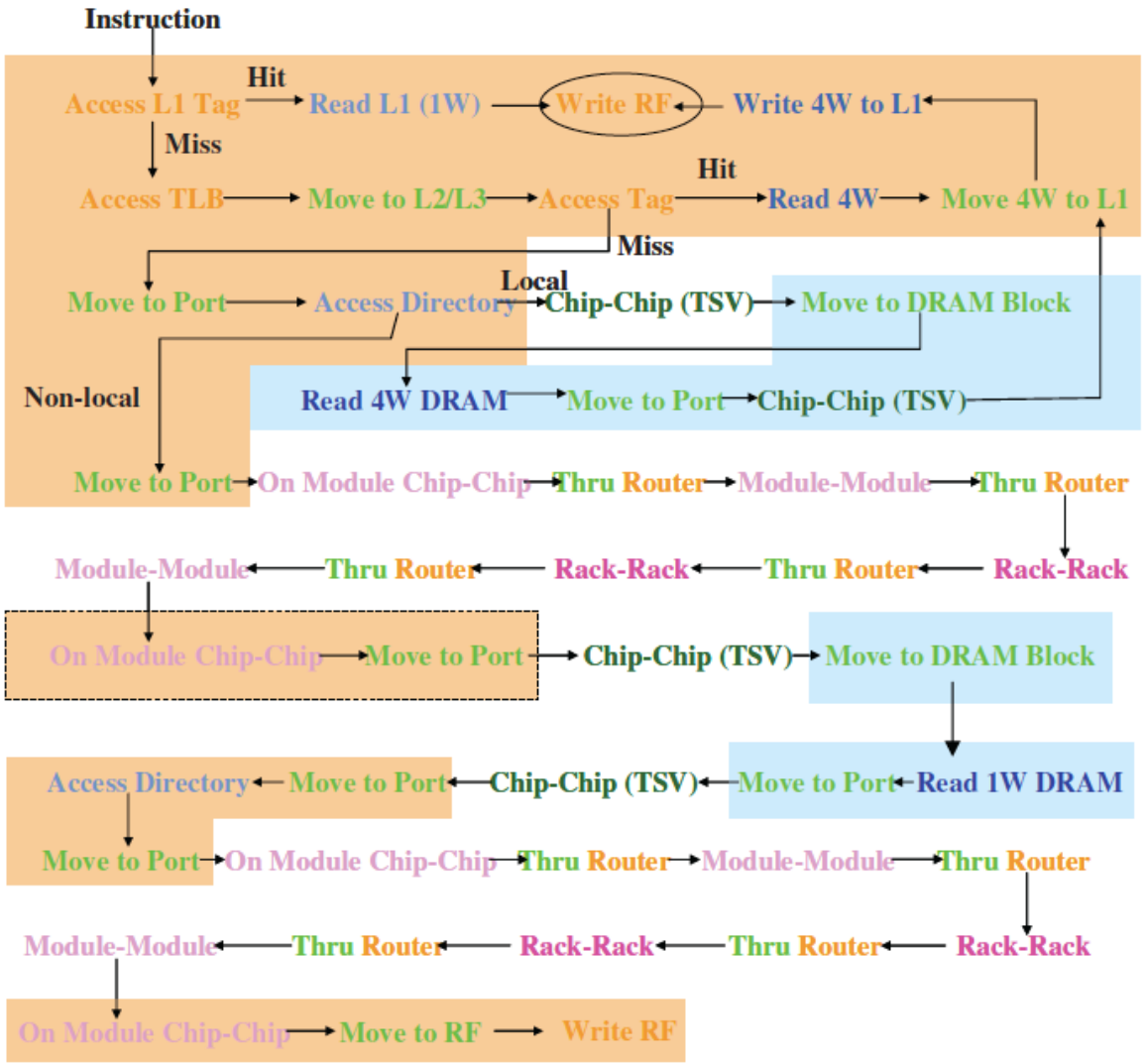
Orthogonal Scaling

- Scale at multichip package level – get move function in same volume

- Issues:
  - Cooling
  - Serviceability

Doing it Better: Frictionless Architecture

- Most energy is "wasted"
  - E.g., 10's of SRAM accesses in order to bring data from memory
  - Load of one memory word is x17 more expensive than it needs be!



Possible Architectural Directions

- Compute accelerators (GPUs)
  - Reduces overhead of instruction decoding, resource management…
- "Memory Accelerators"
  - Reduces overhead of coherent caching
- Approximate computing
  - Reduces storage & computation cost
- Application-specific systems
  - All of the above
  - More reasonable if underlying technology changes slowly – can afford longer design cycle

New Device Technology

- Adiabatic/reversible computing
  - Theoretical limit on switching energy: *ln(2) kT*
  - Current CMOS > *100,000 kT*

  Best demonstrated ~ *3 kT* (nSquid at 4 K) – *Single device*

- Optical (hybrid) computing [Optalysis]
  - analog optical computation (Matrix product, Fourier transform)

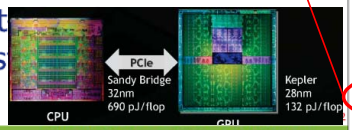
- Cryogenic Computing – Rapid Single Flux Quantum Logic (IARPA)

NVIDIA Projections for CMOS vs. Superconductor RQL Circuits

Power consumption within a GPU

Manufacturing process (and year):	40 nm. ('10)	10 nm. (estim. 2017)	
User platform:	Desktop	Desktop	Laptop
Vdd (nominal)	0.9 V.	0.75 V.	0.65 V.
Target frequency	1.6 GHz.	2.5 GHz.	2 GHz.
Energy for a madd in double-precision	50 pJ.	8.7 pJ.	6.5 pJ.
Energy for a add with integer data	0.5 pJ.	0.07 pJ.	0.05 pJ.
64-bit read from 8 KB. SRAM	14 pJ.	2.4 pJ.	1.8 pJ.
Wire energy (per transition)	240 fJ/bit/mm	150 fJ/bit/mm	115 fJ/bit/mm
Wire energy (256 bits, distance of 10 mm.)	310 pJ.	200 pJ.	150 pJ.

- Communications take the bulk of power consumption.
- And instruction scheduling in an out worse, spending 2000 pJ. for each ins o floating-point).



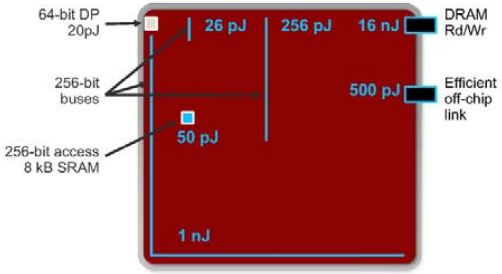
Manuel Ujaldon - Nvidia CUDA Fellow

MIT LL fabrication process technology	248 nm (~2017)	193 nm (~2019)
JJ critical current density J_c	10 kA/cm^2	10 kA/cm^2
Min. JJ critical current I_c	38	20
Frequency	**8.5 GHz**	**10 GHz**
Energy for a DP FP Multiply-Add	**4.2 pJ**	**2.2 pJ**
Energy for a 64-bit integer add	**0.21 pJ**	**0.11 pJ**
64-bit read from a 64x64-bit register file (dynamic)	**0.15 pJ**	**0.08 pJ**
Wire (PTL) energy per non-zero bit (incl. drivers & receivers) 1-20 mm	**0.25 fJ/bit**	**0.13 fJ/bit**
Wire (PTL) energy (256 bits, 1-20 mm) (random data)	**0.032 pJ**	**0.017 pJ**

The High Cost of Data Movement

Fetching operands costs more than computing on them

B. Dally, DOE Exaflops WS, 2011



- Projected energy consumption for RQL processors:
  - **Communication is Cheap, FLOPS are Expensive (cmp. to Communication), Instruction Scheduling & Main memory access costs are TBD**
  - On-chip data transfer takes negligible energy
    - **~5,000-10,000X LESS on-chip than in 10 nm CMOS**
  - Off-chip communication has ~ <u>same</u> negligible energy costs as on-chip one @ the same rates
  - Floating-point operations take most energy
    - **~2-3X LESS energy/op than in 10 nm CMOS** *with the cryocooling efficiency of **0.1%** (**1000 W/W**)*

New Device Technology

- Cryogenic Computing – Rapid Single Flux Quantum Logic
  - Signals propagate from one gate to the next as millivolt picosecond SFQ pulses

- Very different balance:
  - Communication on chip and off chip is free
  - Logic (especially floating point) is expensive
  - Feature size of 200 nm in 2019 (vs. 10 nm or less for CMOS)

- Very different ecosystem
  - No cryogenic cell-phones; current market is for small, special-purpose devices sold by small companies

Something Totally Different

- Quantum Computing
- Neuromorphic computing
- Biocomputing

None seem to apply to scientific simulations

Summary

- Exascale will be there by 2022 or so
- "Business as usual" (riding on Moore's Law and commodity technology) is becoming increasingly harder
- Supercomputers are becoming more "special purpose"
  - Expect most/all supercomputers to use floating point accelerators in a few years; more specialized accelerators to follow
- Can continue to push performance to zetascale
  - Will need to think of supercomputers as unique facilities, such as particle accelerators – not clusters of PCs
- *Supercomputing will become much more interesting*


Questions?