Los Alamos
NATIONAL LABORATORY

EST.1943

Delivering science and technology
to protect our nation
and promote world stability

NNSA
National Nuclear Security Administration

Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

Recent and Expected Advances in HPC

Future CFD Technologies Workshop

Dr. Josip Loncaric

HPC Technology Futures Lead

Los Alamos National Laboratory

NNSA
National Nuclear Security Administration

Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

LA-UR-17-30925

Outline: Exciting Times! Adapt to Survive.

January 7th, 2018

11:30 – 12:00

Kissimmee, FL

- **Dennard scaling for CMOS and Moore's law: Zenith of CMOS**
- **Architecting future HPC platforms**
  - Technical feasibility
  - Market feasibility
  - Cost
  - Programmability
  - Productive use
- **End of Moore's law**
  - Exploration: quantum, neuromorphic, …
- **What's next?**

Zenith of CMOS? Really???

- **Really really:**
  - Zenith in the sense of CMOS progress stalling
    - CMOS progress stall is already affecting us. CMOS expected to hit an economic wall in 2021
    - ITRS 2015 edition is the last one ever – no more coming
      - https://www.semiconductors.org/main/2015_international_technology_roadmap_for_semiconductors_itrs/
      - SIA statement: “Faced with ever-evolving research needs and technology challenges, industry leaders have decided to conclude the ITRS and transition to new ways to advance semiconductor research and bring about the next generation of semiconductor innovations.”
  - Not in the sense of CMOS going away soon
    - Trillion dollar electronics industry will ride CMOS to the bitter end
    - Semiconductor industry isn’t focusing on replacements yet ($ reason)
  - CMOS won’t end because we ran out of sand
    - “Stone age didn’t end because we ran out of stones.”
  - CMOS will only be displaced when something better shows up
    - Buggy and horse whip industry was just fine until cars showed up
    - New: International Roadmap for Devices and Systems: http://irds.ieee.org/

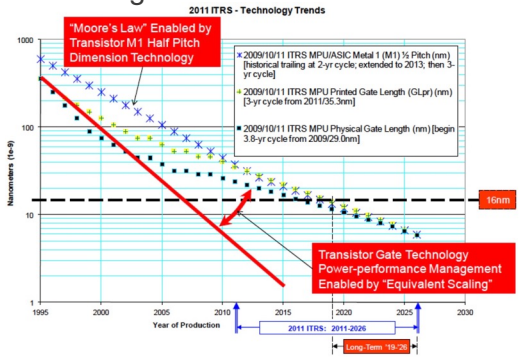
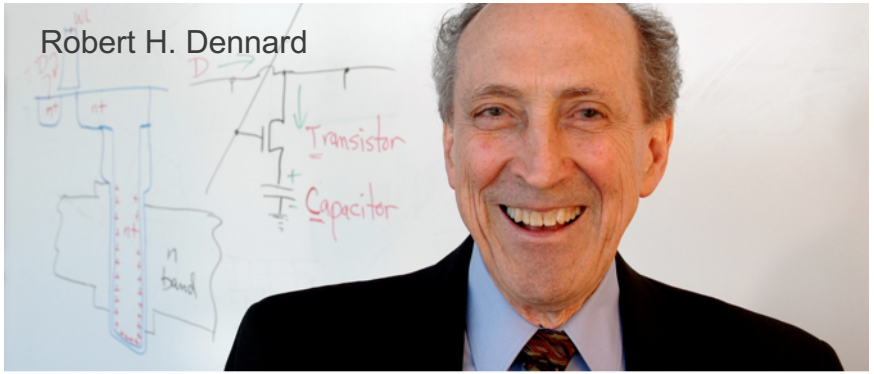
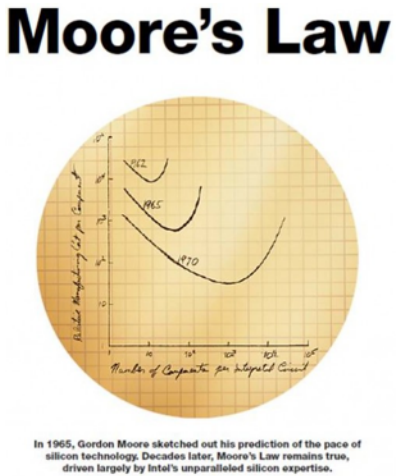


Figure 11 2011 ITRS—MPU/High-performance ASIC Half Pitch and Gate Length Trends

Stalled CMOS progress allows alternatives to catch up: Watch this space!

Dennard Scaling and Moore's Law





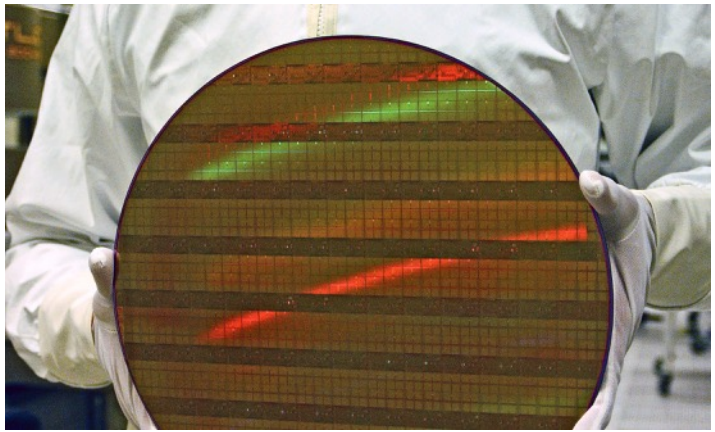
- **Dennard's scaling for Complementary Metal-Oxide-Semiconductor (CMOS)**
  - Voltage and current proportional to the linear dimensions of a transistor
  - Power proportional to the area of a transistor
- **Moore's law is enabled by Dennard's scaling for CMOS**
  - …and economics of finer lithography required to make smaller transistors

Moore's law is about economics. Dennard scaling is physical.

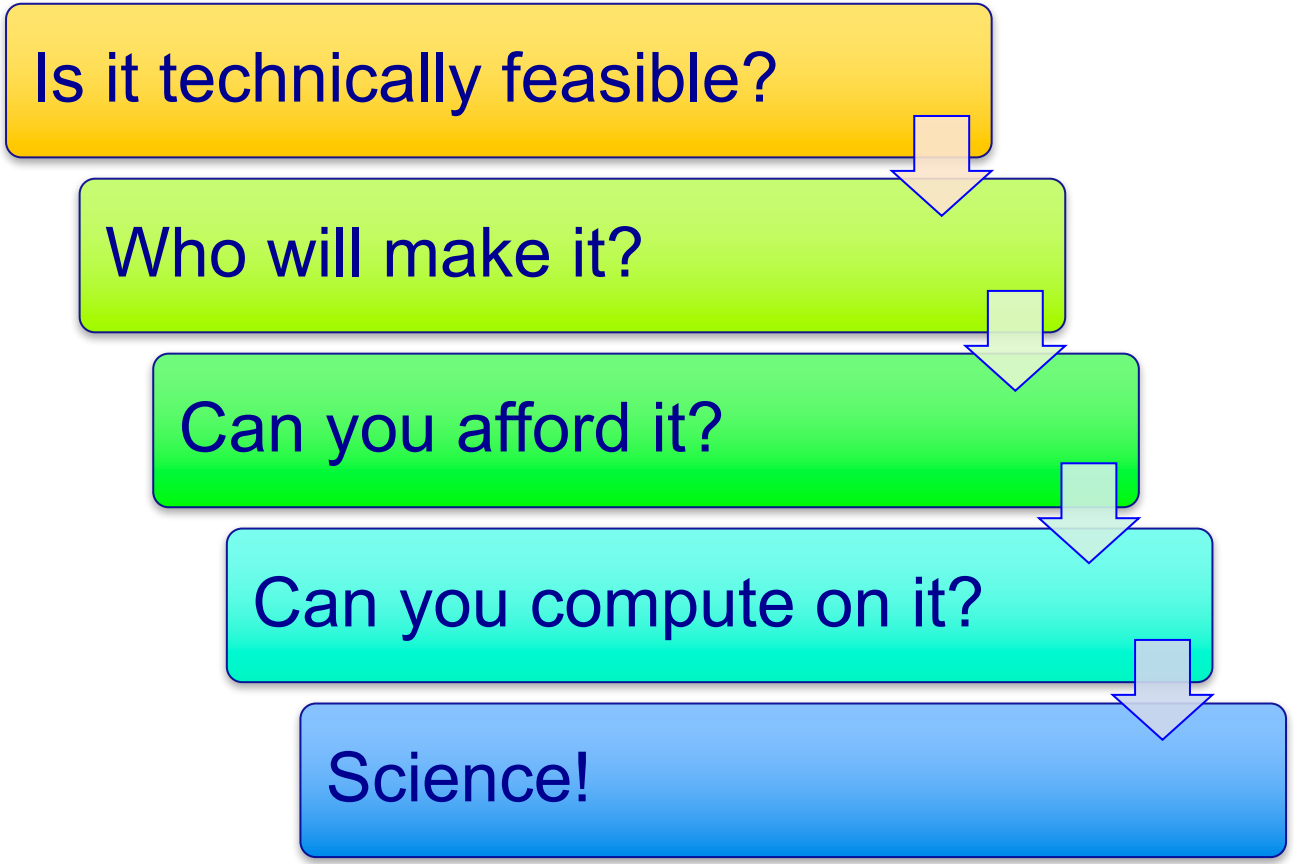
Exponential Growth & Moore's Law

- **Doubling of transistor density every 1.5-2 years**
  - Delivers twice the performance potential
  - Exponential growth since 1960s delivered unprecedented gains
  - However: *"If something can't go on forever, it won't"*

- **Flip side of Moore's law**
  - Exponential growth in *complexity*
  - Today:
    - Designing multi-billion transistor processors
    - Processors cost $1+ billion to develop
    - Fabrication facility costs ~$20 billion
      - Intel invested $4.1 billion into ASML Holdings, to accelerate next (EUV) lithography technology and tools for larger 450-mm silicone wafers; progress still slow; only four EUV machines sold in 2016 at $110 million each, production use planned for 2018-2020 at 7nm technology node (for some layers of the chip), TSMC's 3nm fab projected to cost $20 billion (2017), EUV finally ready (IEEE Spectrum, 1/5/2018)

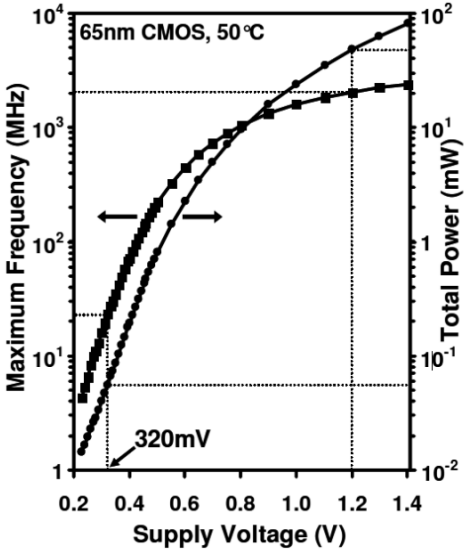


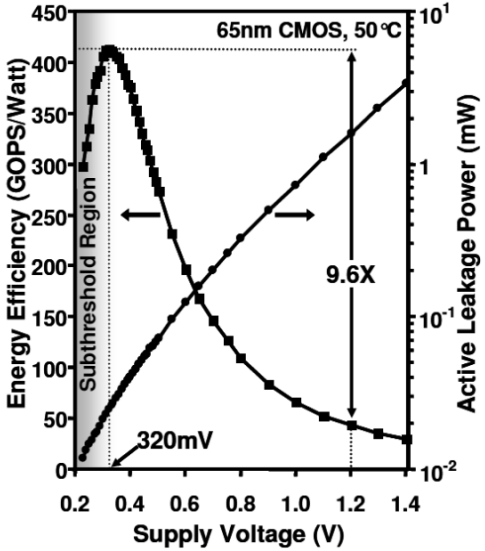
Science Depends on the Available Computer

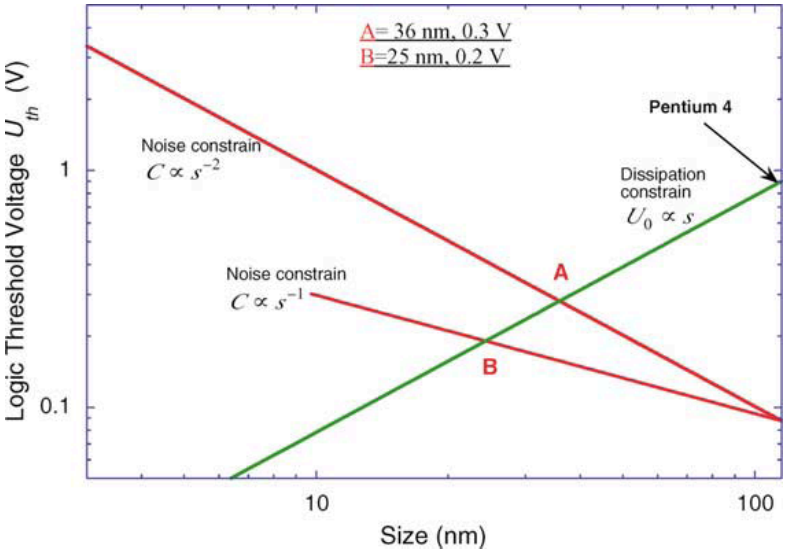


Is it Technically Feasible?

Semiconductor Physics Drives Major Changes







Sources:

(1) ExaScale Computing Study, DARPA, 2008

(2) End of Moore's law: thermal (noise) death of integration in micro and nano electronics, L.B. Kish, 2002

Energy and Power Challenge

Memory and Storage Challenge

Concurrency and Locality Challenge

Resiliency Challenge

Semiconductor Scaling Changed Around 2004

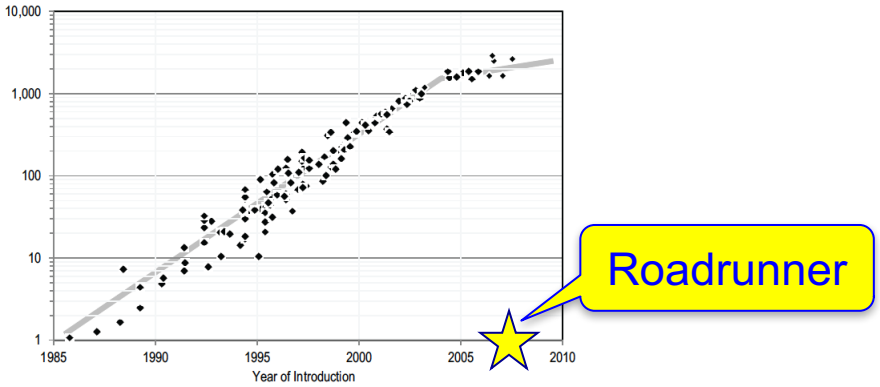
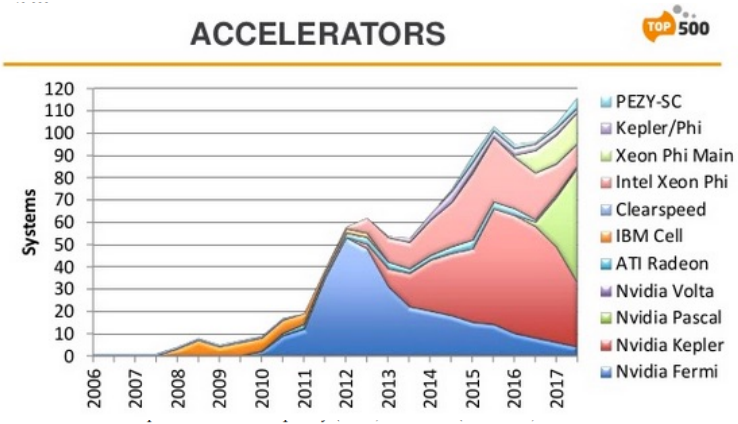


FIGURE A.1 Integer application performance (SPECint2000) over time (1985-2010).



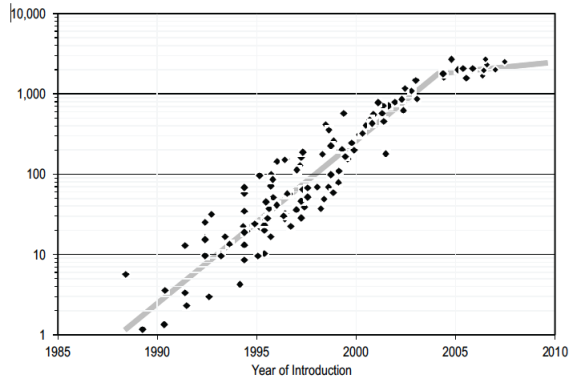


FIGURE A.2 Floating-point application performance (SPECfp2000) over time (1985-2010).

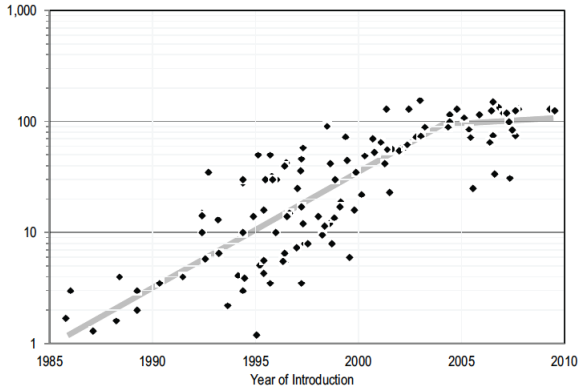
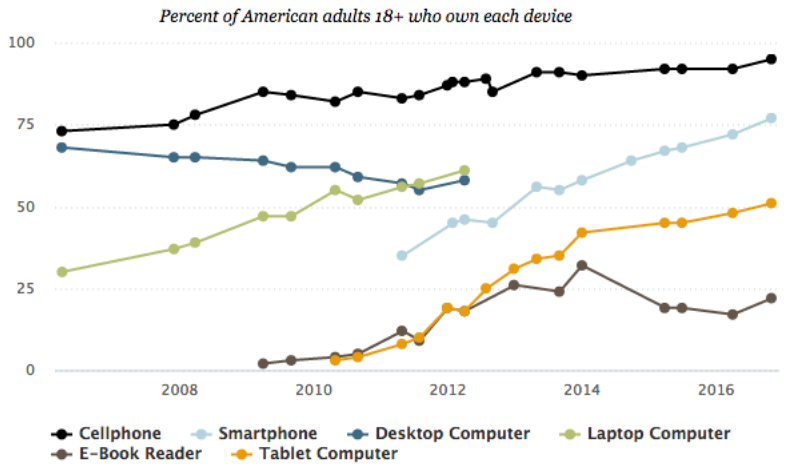


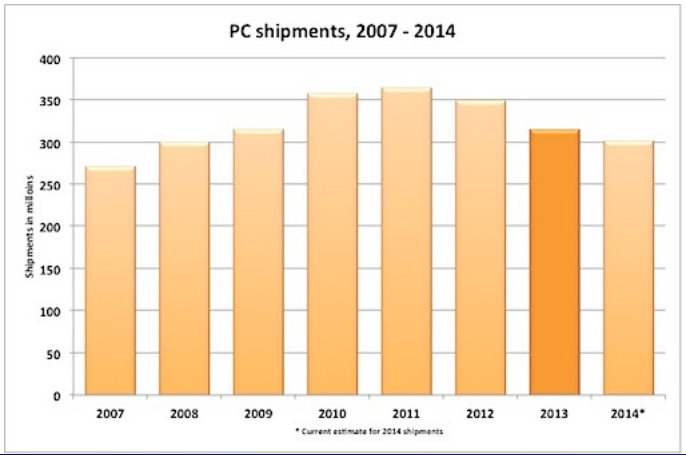
FIGURE A.4 Microprocessor power dissipation (watts) over time (1985-2010).

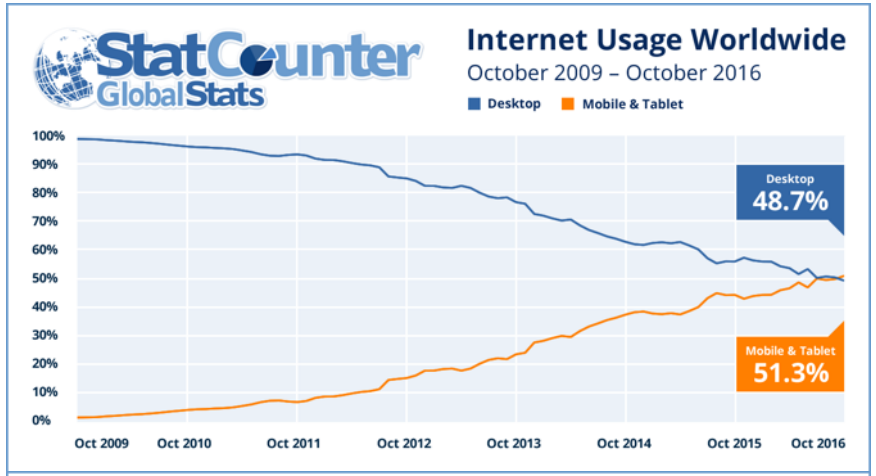
Source: “The Future of Computing Performance: Game Over or Next Level?”, NRC, 2011

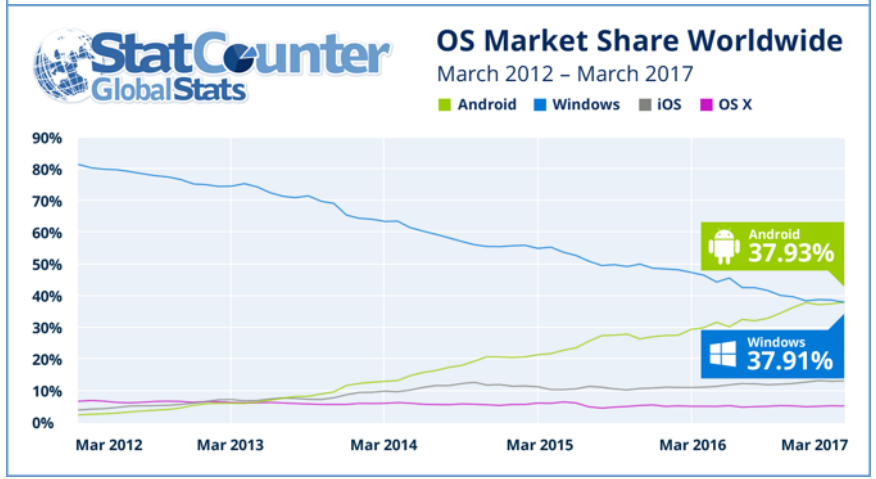
Who Will Make It?

Major Market Shifts Are In Progress

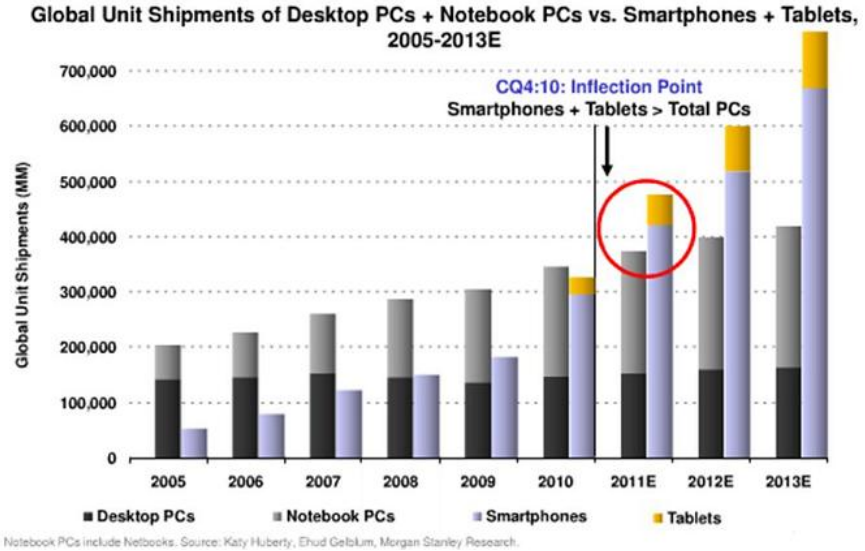




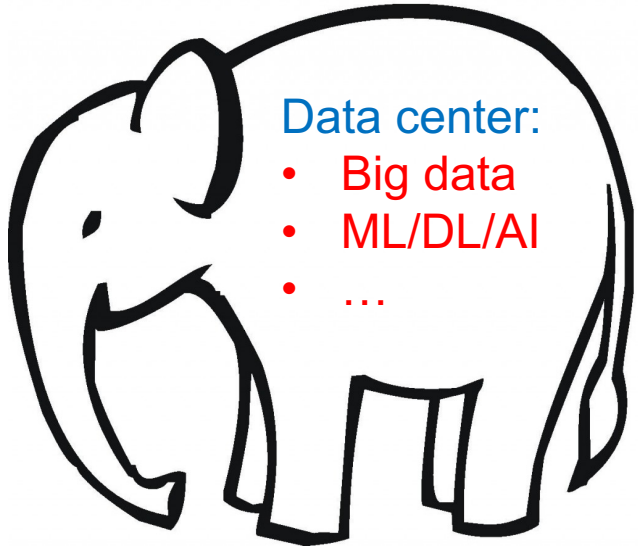




Past: Leverage Desktop PC Growth to Build HPC
Present: Desktop PC Sales Are Stagnating & Declining



Net



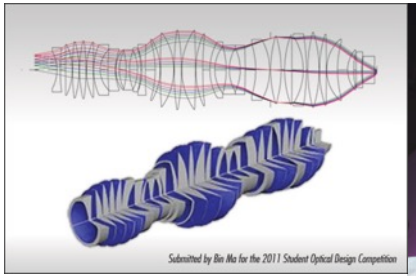
For first time in a decade, PC sales slip below 63 million

Jon Swartz , USA TODAY Published 12:51 p.m. ET April 12, 2017 | Updated 7:49 p.m. ET April 12, 2017

PC in a decline, mobile device market and networked services growing.

Who Is Driving This Market?

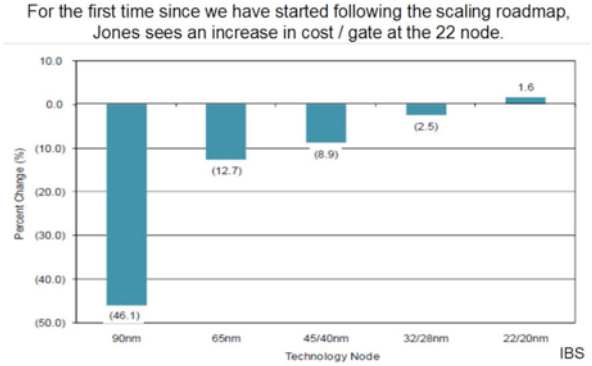
- ***Not us!***
  - 2014 electronics industry: >$1,000 Billion
  - 2016 semiconductor industry: $339 Billion
  - 2016 Intel total revenue: $59.4 Billion
  - 2016 manufacturing equip. market: $39.1 Billion
  - 2016 Intel PC Client Group revenue: $32.9 Billion
  - 2017 cost of a fabrication facility: $20 Billion
  - 2016 Intel Data Center Group revenue: $17.2 Billion
  - 2016 global HPC market: $11.2 Billion
  - 2011 cost of processor development: >$1 Billion
  - NNSA budget per supercomputer: $0.2 Billion (~1% of Intel DCG)
  - LANL IC budget for supercomputers: $0.0 Billion (roundoff error)
- **Inevitable conclusion: Widely applicable progress needed**
  - Semiconductor industry needs much larger markets than a single HPC platform or even the global HPC market
  - Leads to trade-offs in power-performance-memory space
  - Economics 101: People need a reason to part with their money
  - Not enough money = no processors
    - New option: Mild processor customization for $10-100 million (depending on details)



No choice: We <u>must</u> leverage broader industry trends

End of Moore's Law: Follow the Money!

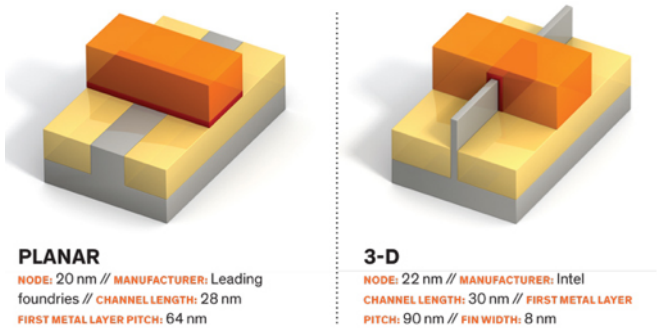
- **Price/gate no longer decreasing**
  - Price/gate actually *increased* at 22/20nm
    - Intel still claims economic gains, future TBD
  - Lithography: Extreme UV challenges ($, time)
  - Wafer sizes: 450mm fabs expensive ($)
    - When deployed, 450mm may enable some cost reductions



- **Functionality gains**
  - Gaining ~1.6x instead of 2x per technology generation
- **Losing "Separation of Concerns"**
  - Power efficiency, resilience, scaling, programmability challenges
  - Co-design needed at many levels, complex optimization tradeoffs
- **If designers can't deliver >>10% benefit per generation, will people buy?**
  - Bob Colwell's answer: CMOS progress will stall around 2020-2022
    - http://archive.hpcwire.com/hpcwire/2013-08-29/moore_s_law_we_miss_you_already.html
    - Current expectation: stall, then slow CMOS progress to about 2028

World's Total Electric Power to IT?

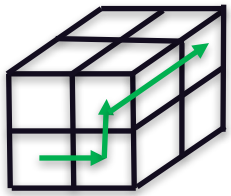
- **Bernd Hoefflinger, editor of *Chips 2020:***
  - "They expect 1000 times more computations per second within a decade. If we were to try to accomplish this with today's technology, we would eat up the world's total electric power within five years. *Total electric power!"*
  - Data centers consume about 8% of worldwide electricity today
  - Example: Worldwide bitcoin power use 3,441 MW https://digiconomist.net/bitcoin-energy-consumption
- **Power efficiency increases through:**
  - Shorter wires
  - 3-D circuit designs
  - FinFET transistors
  - 3-D merged transistors
  - Redesign computations, e.g. multiplication unit
  - Redesign chip circuits using communication circuitry
  - Advanced power management & power efficient architectures
    - Many implications for applications



Why 3D?
Energy Use Grows With Distance * Bits Moved

- **Smaller transistors are closer together and therefore more efficient**
  - This was good while we could reduce feature sizes, but…
  - Economics says there will soon be some minimum size --- call that unit size
- **3D integration to the (temporary) rescue**
  - Energy proportional to Manhattan distance to N units: diameter = $D*N^{1/D}$
    - In 2D: 2*sqrt(N)
    - In 3D: 3*cuberoot(N)
    - Energy efficiency advantage of 3D vs. 2D = $2/3 * N^{1/6}$

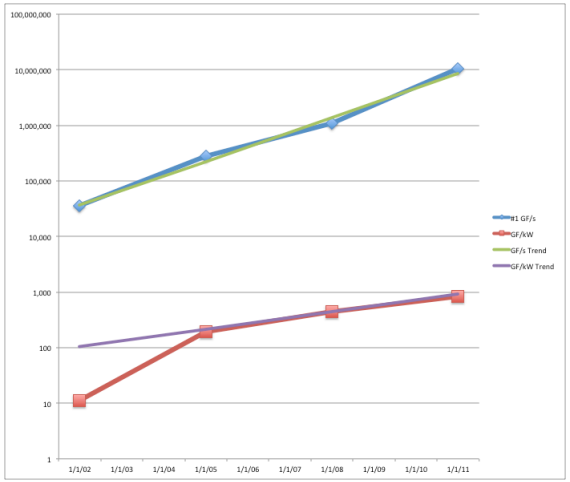
- **Examples:**
  - N=729 has 3D energy efficiency advantage 2x
  - N=8,304 advantage is ~3x
  - N=46,656 advantage is 4x

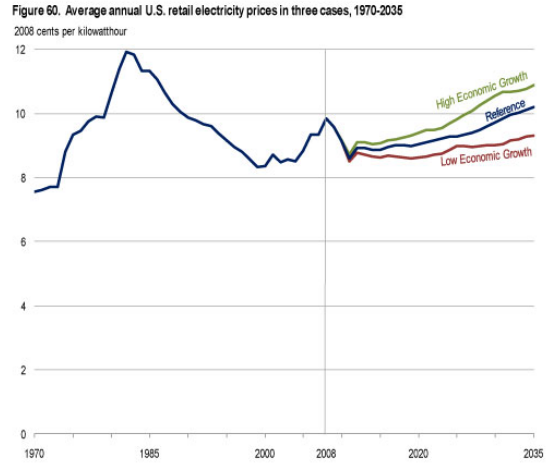


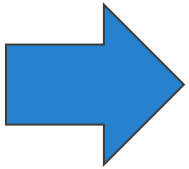
- **We are not quite there yet, but some 3D stacking is already here**
  - KNL with MCDRAM: about 5x bandwidth advantage (finer wire pitch, shorter wires)
  - Node as a package: Tighter integration than motherboard

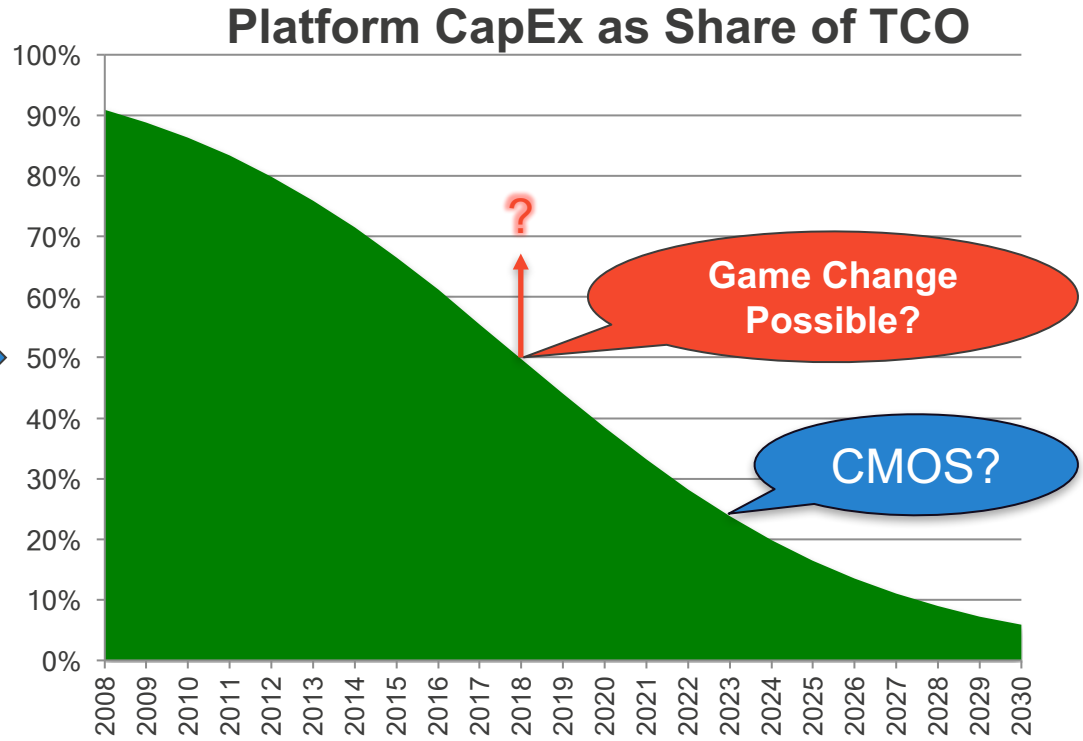
Can You Afford It?

Power Costs Are Growing, May Limit HPC Progress









Sources:
(1) Top500 Nov. 2011 list
(2) U.S. Energy Information Administration, Annual Energy Outlook 2010

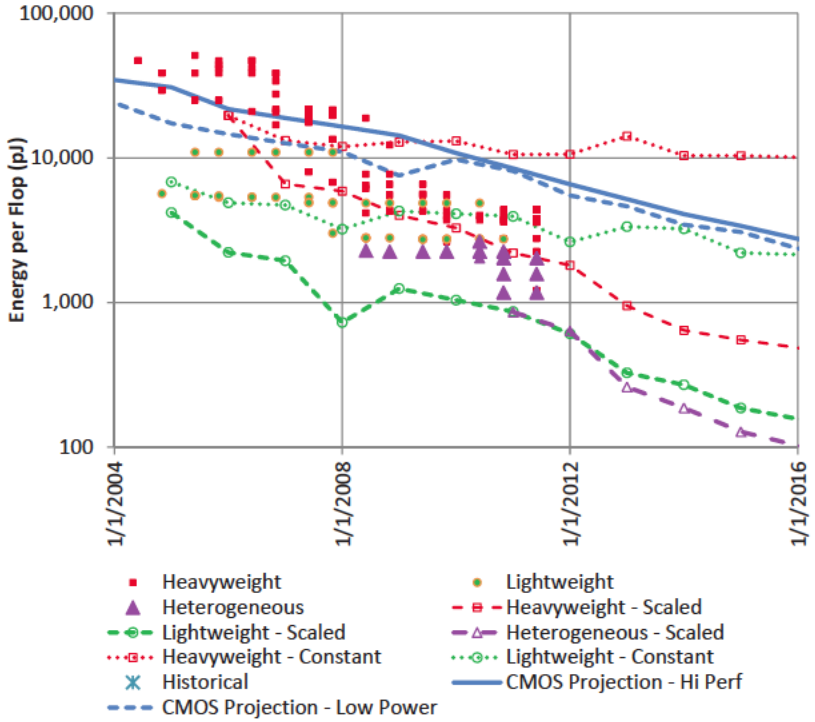
Heterogeneous and Lightweight Nodes Promising, Heavyweight Nodes Require Too Much Power

"Looking forward, the same power dissipation limits that caused the inflection will continue to have an effect. Using the same amount of silicon area to increase the intrinsic computation power of a chip will likely require the basic clock rate of such chips to not just flatten, but actually decline, meaning that even more brute parallelism will be needed to make advances in high end performance. Matching this performance increase will require increased memory bandwidth, which in turn may not scale as much as logic. In fact, […] the differences between the scaled and constant models is almost totally driven by the memory access and I/O power differences, […].

It is also interesting to note that with these power constraints, only the heterogeneous model has a chance at a peak (not sustained) exaflop/s by 2020, but at a power of in excess of 100MW." ← based on 2011 model assumptions

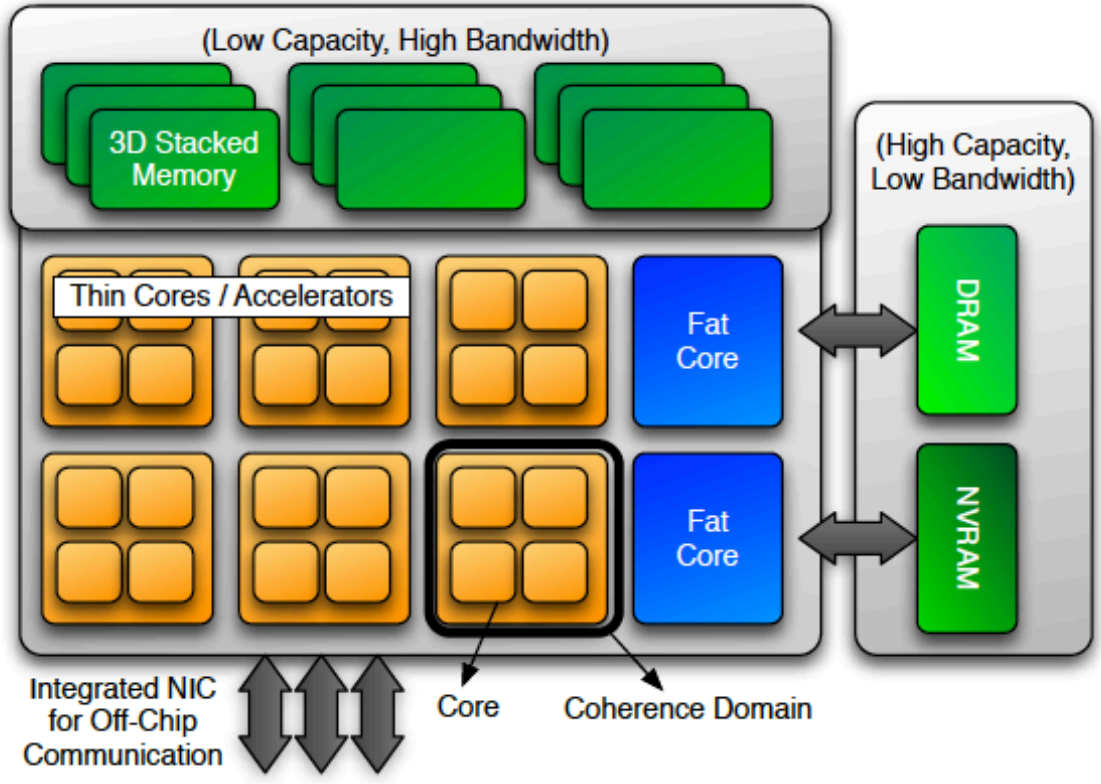
"**Using the TOP500 to Trace and Project Technology and Architecture Trends," P. M. Kogge & T. J. Dysart, SC11, 2011**

See also: "Yearly Update: Exascale Projections for 2013," by Kogge & Resnick, SAND2013-9229, Oct. 2013



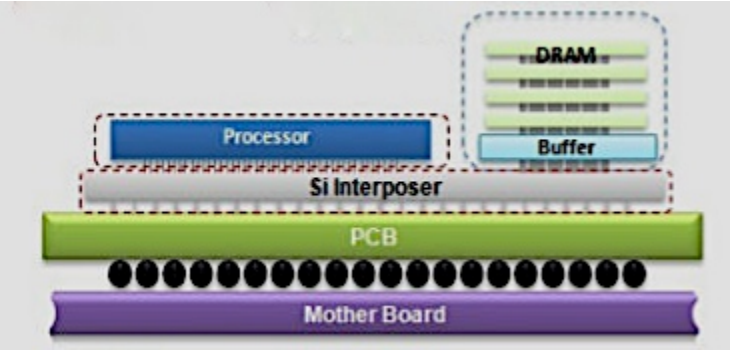
Heavyweight: Jaguar class machines
Lightweight: Blue Gene class machines
Heterogeneous: Roadrunner class machines
Scaled: Power follows ITRS projections
Constant: Constant RAM+I/O power per use

DOE-funded Vendor Research ⇒ Abstract Machine Model for an Exascale Node



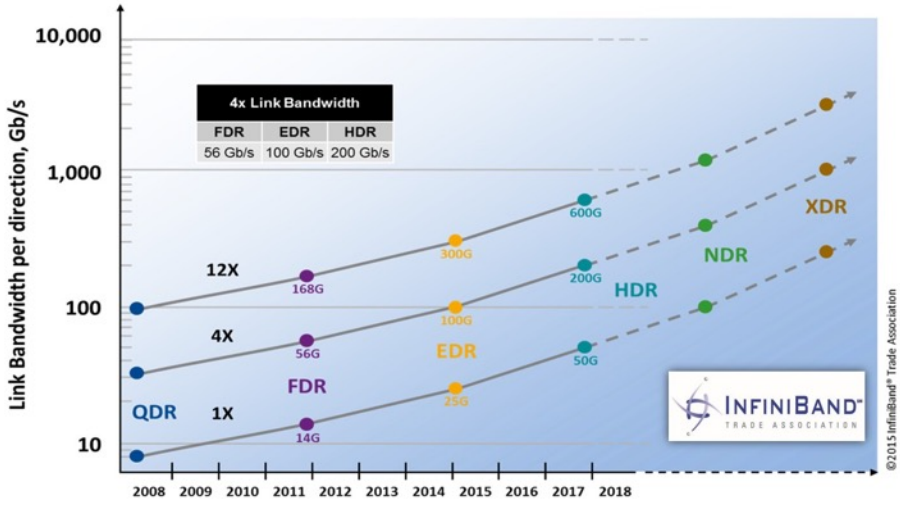
J.A. Ang et al: "Abstract Machine Models and Proxy Architectures for Exascale Computing," Rev 2.0, July 2016

More Building Blocks: Memory, Networks



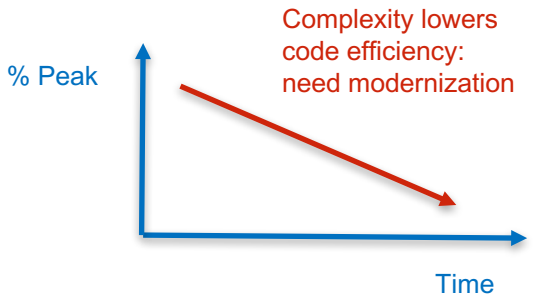
- **Memory:**
  - On-package memory has more pins, shorter traces, lower f, ~5x lower data transfer energy/bit
    - Trinity MCDRAM latency similar to DDR4
    - HBM (one-time gain): ~2-3x lower latency
    - BW*delay product growing: Little's law
  - Capacity/BW needs similar to cell phones

- **Network:**
  - All technologies improving similar to IB
    - Other protocols: OPA, RoCE, etc.
  - BW = f*width, latency not improving
    - BW*delay product also growing: Little's law
  - Short haul electrical then optical
  - SiPh delayed, coming any year now
  - Increasing energy cost of larger networks: tapering



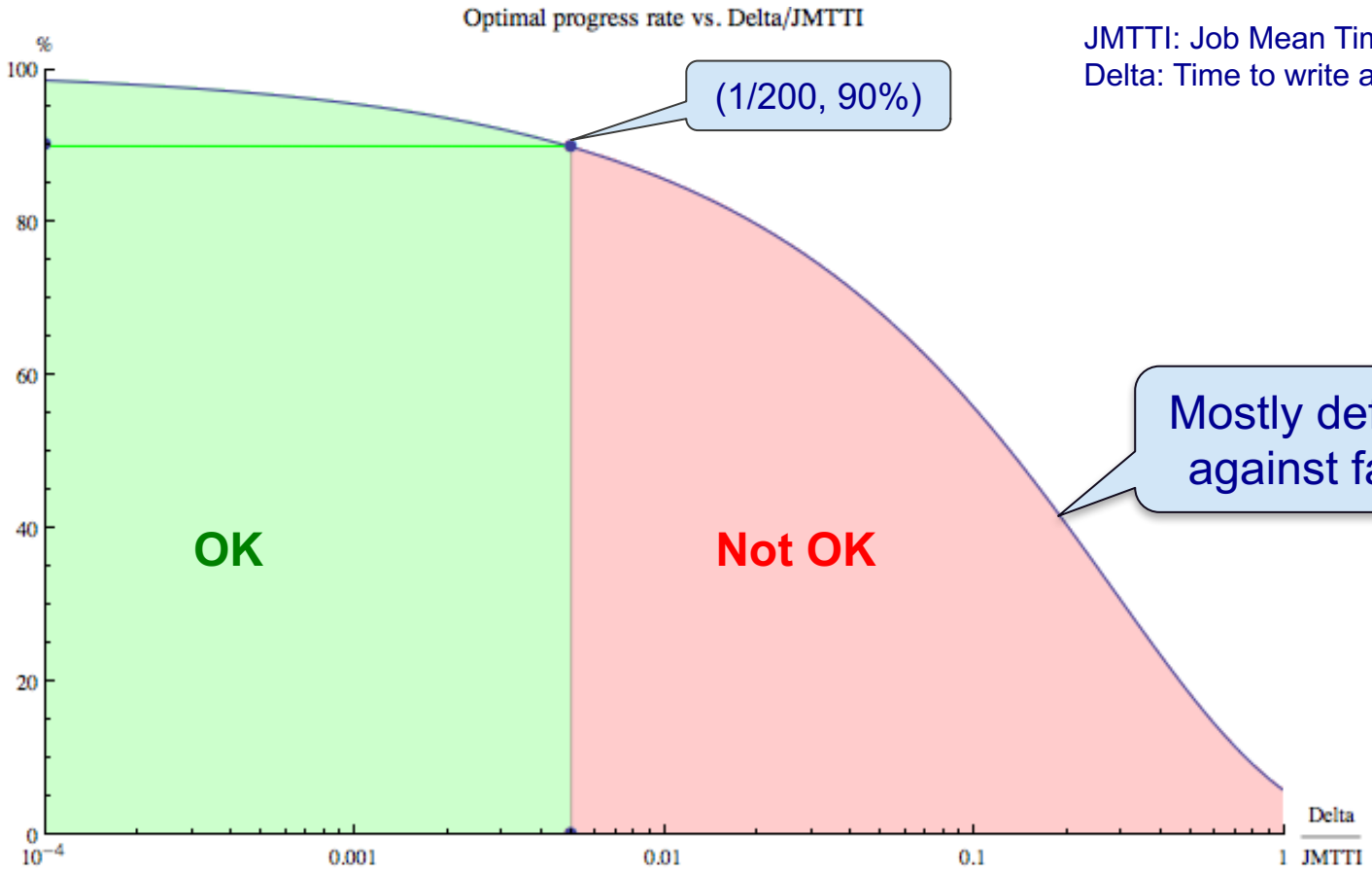
Can You Compute On It?
Requires Changing Codes and Algorithms

- **Changing architectures and programming models:**
  Need for abstraction and new methods
  Will need to adopt newer languages and language constructs
  Need for new tools that analyze codes and identify bottlenecks
  Debugging will be painful
  - Advanced architectures are many-core or heterogeneous
  - No matter what architecture it will likely require an MPI + X approach
  - Efficient use of wider vector widths required
  - Memory per core may be smaller at the performance tier, perhaps with more tiers of memory hierarchy
  - Moving data will be the single biggest cost in time and energy
  - Resiliency is essential for computational progress at full scale
- **Disk performance is not likely to increase in proportion to compute:**
  Need for in-situ analysis
  Need for different I/O paradigms
  - Disk is too slow: Burst buffer needed at multi-PF/s level ⇒ storage tiers evolving
  - Computational progress is a nonlinear function of JMTTI/Delta where Delta is the time to write a checkpoint



Changes mostly due to power limits and concurrency at the node level

It's Not Just a Budget Limit: Computational Progress Rate Decreases With Scale



JMTTI: Job Mean Time To Interrupt
Delta: Time to write a checkpoint

Science:
What Will It Take?

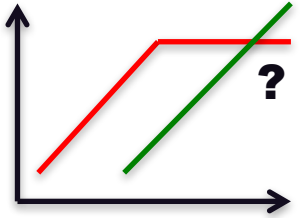
- We will have to manage data movement and locality, hide latency.
- We will have to write code that vectorizes, taking full advantage of accelerators or main processor vector units.

- The compilers are not going to solve this problem for us in the next 3-4 years. Not even close. Quit dreaming.

- This requires expert labor to rewrite our existing codes, and adjust our workflow to new HPC architecture balance

- NNSA is pursuing a two-prong strategy to modernize existing codes and develop new codes

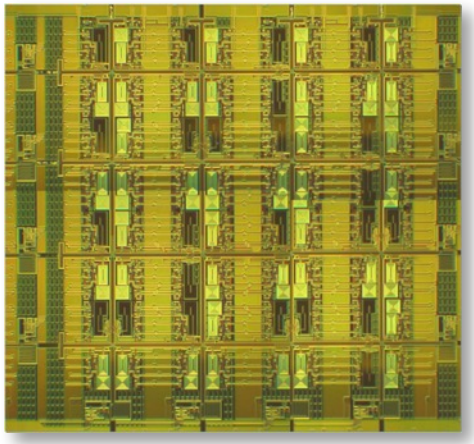
Beyond Moore's Law:
How Will We Compute in 2020, 2030?

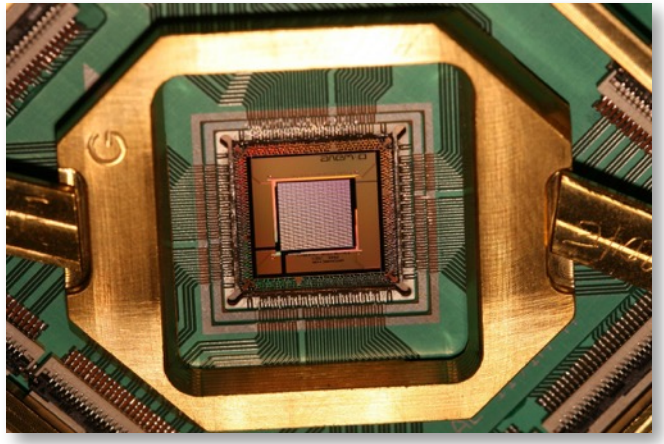


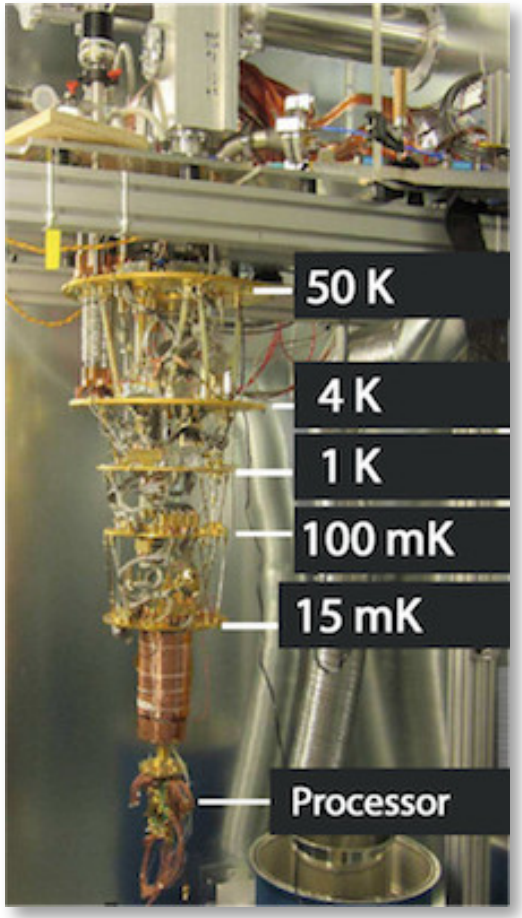
- **There are NO compelling alternatives to CMOS yet**
  - CMOS progress will stall, but continue to dominate for a while
  - Expect product diversification, a la "Internet of things"
  - Expect competing technologies to complement CMOS, then grow
    - DARPA: Investigating ~30 technologies, out of which 2-3 are modestly promising
- **Lessons of Roadrunner continue to be refined**
  - Technology moves on:
    - We can't build a 2010 computer in 2020, and 2020 technology will be complex
    - More of our computing budget will go to pay for electricity in 2020
      - Particularly acute problem for heavyweight architectures
      - Heterogeneous architectures most promising for energy efficiency, lightweight very close
    - Everyone needs a plan to stay in business 5 years from now
      - Semiconductor industry is fickle, rapidly changes direction in response to market forces
      - Survival of the most adaptable
- **Outlook for 2020 to 2030: Exciting times ahead!**
  - 3D circuits, improved packaging, cooling, architectures, memory, software, system features, special purpose designs, new applications, marketing, etc.
  - Communications, biology, physics & materials, control theory, molecular computing & storage, spintronics, photonics, quantum, neuromorphic, etc.
  - Some algorithms won't survive! Science must adapt.

Exploration: Quantum Computing

- **Example: D-Wave 2X at LANL**
  - Extreme cold needed for quantum effects
    - At low temperatures, niobium loops become superconducting
    - Electron flux direction can be indeterminate
  - How cold? The chip has to be kept very, very, very cold
    - LANL's D-Wave 2X chip is kept at **10.45 mK**
    - That's 0.01°C above absolute zero
    - For comparison, interstellar space is far warmer: 2700 mK

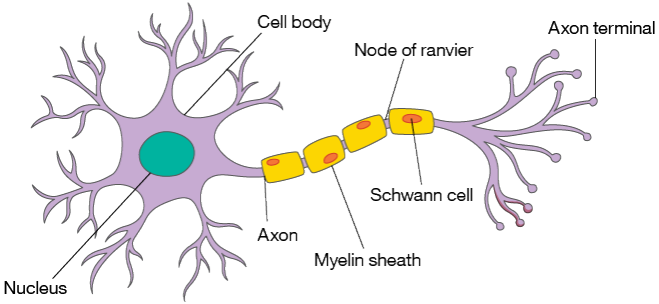


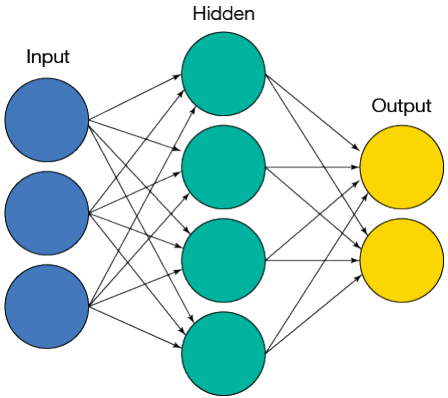


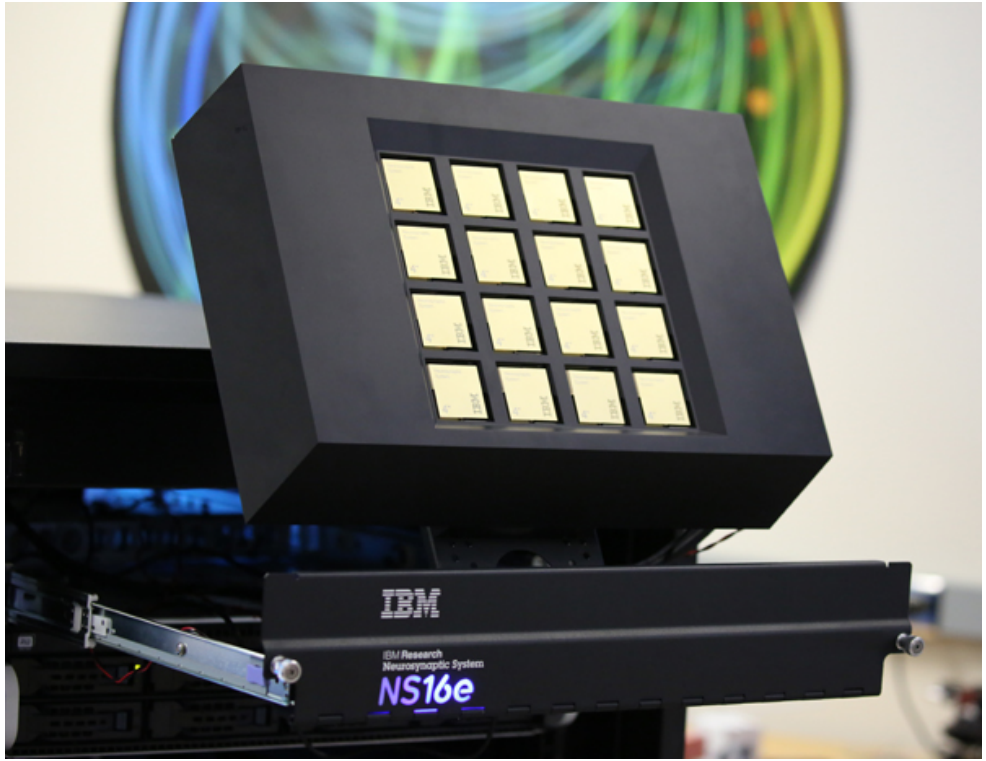


Exploration: Neuromorphic Computing

- **Example: IBM TrueNorth at LLNL**
  - Testbed: 16 million neurons, 4 billion synapses, at about 2-3 Watts, good for ML etc.
  - Also used by LANL and SNL

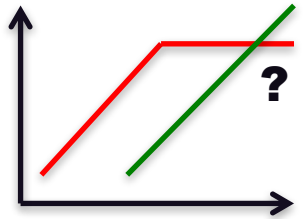






Outlook: Exciting Times!

- **Science = ETC (Experiments, Theory, Computation)**
- **Computation must adapt to physical characteristics of computing technology of the day**
- **Lessons of history**
  - Major epochs of computing technology last ~16 years
  - Capability growth per epoch ~1000x
  - Current epoch estimate: 2004 to 2020 (components), 2008-2024 (systems)
- **Massive concurrency: O(10^6) today, O(10^9–10^{10}) by exascale**
- **Period of diversity and increasing uncertainty after 2020**
  - CMOS progress stalls, alternative technologies eventually catch up
  - "Beyond CMOS" technologies likely by about 2030



- ***Adapt to survive!***

Conclusions

- **Industry, HPC, DOE/NNSA on the path to maximize usefulness of CMOS**
  - CMOS is just too good to be replaced, but we already miss Dennard scaling
- **Physical and budget constraints leave only a narrow path forward**
  - Abstract machine models describe what that looks like
- **Massive scale challenges power, cooling, and system resiliency**
  - Even much larger budgets don't fix machine failures
- **Vendors & ITRS expect slow progress beyond 2020**
  - Slow improvements in energy efficiency
  - Slow improvements in performance per dollar
  - Growing challenges of resiliency for full scale applications
- **Alternative technologies have a chance to catch up by 2030**
  - Evaluating promise of quantum and neuromorphic computing
- **Near term HPC systems chosen based on best value to our mission**
  - Revolutionary additions (quantum, neuromorphic, …) possible in the long run

Abstract

Dennard scaling and Moore's law delivered decades of high performance computing gains, but CMOS progress is expected to stall in early 2020s for both physical and economic reasons. DOE and NNSA investments in exploring future technologies reveal that physics of computation is forcing HPC into a design corner involving heterogeneity in both processing and memory, as described by abstract machine models. Continued exponential growth in delivered functionality requires comparable improvements in energy efficiency, or else information technology is poised to consume the entire global electricity supply. Capability growth also requires budgets and scale, forcing defensive measures to deliver reliable results despite the constant stream of component failures. While there are no compelling alternatives to CMOS today, alternative technologies have a chance to enter the HPC ecosystem by 2030. Prudent planning is required to prepare aerospace engineering applications in time to intercept HPC technology evolution, starting with the machines available today.