Exascale?

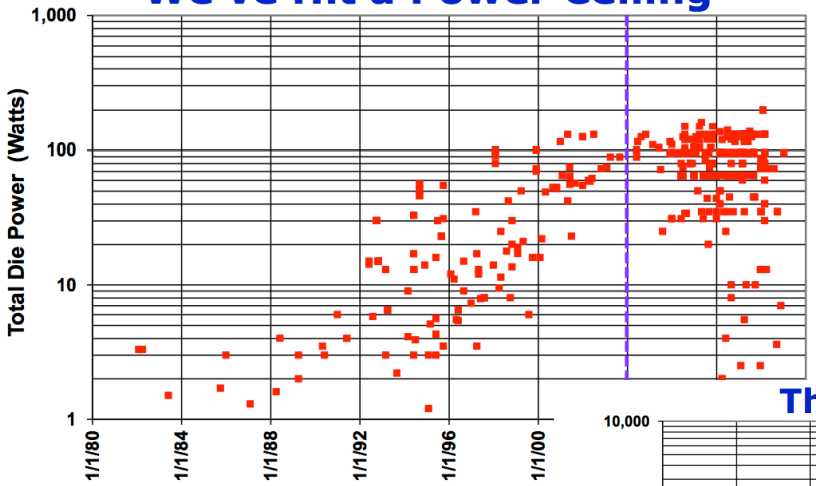
Pete Beckman, Argonne National Laboratory

Director, Exascale Technology and Computing Institute

Co-Director, Northwestern University – Argonne Institute of Science and Engineering

Data from Peter Kogge, Notre Dame

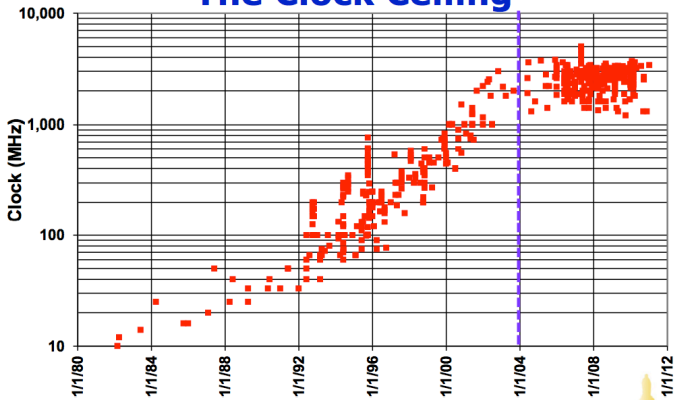
We've Hit a Power Ceiling



data from www.cpudb.stanford.edu

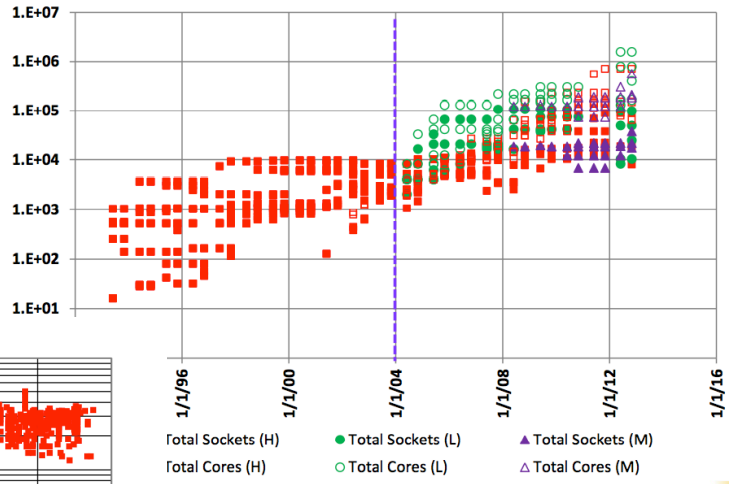
UNIVERSITY OF NOTRE DAME Argonne 30 Years: May 14, 2013

The Clock Ceiling



data from www.cpudb.stanford.edu

Sockets and Cores Growing



Argonne 30 Years: May 14, 2013 ENABLING INNOVATION

What's the Power Problem?

- *Mira*: Blue Gene/Q System
  - 20 times faster than BG/P *Intrepid* (10 PF)
  - ~4 times more power (~4 MW)
  - ~5X more power efficient than BG/P
- Repeat twice to reach Exascale?
  - 400 times faster than BG/Q *Mira* (4 EF)
  - ~16 times more power (**~64 MW**)
  - ~25X more power efficient than BG/Q

Japan:

Japan´s Policy toward Exascale Computing

Yoshio KAWAGUCHI
Office for Promotion of Computing Science / MEXT
27 February, 2014

MEXT
MINISTRY OF EDUCATION

Project Arrangement between US and Japan on R&D Collaboration for HPC System Software Development

Agreement between US and Japan on Cooperation in R&D in Sci. and Tech. (1988)

The Implementing Arrangement Concerning Cooperation in R&D in Energy and Related Fields
MEXT, Japan ⇔ DOE, US
<April 30, 2013>

※Cooperative area described in this arrangement:
(Nuclear Fusion Science, High Energy Physics, Nuclear Physics, Computer Science, etc)


※At Joint High Level Committee Meeting on Science and Technology cooperation (April 30, 2013)

TBC (As One of cooperative area)

Project Arrangement to the Implementing Arrangement between US and Japan Concerning Computer Science and Software Related to HPC for Open Scientific Research

- Participating Organizations:
  National Lab. (DOE, US), RIKEN (Japan)
- Cooperative Area:
  System Software
- Major Contents
  Description for Forms of Cooperation and Management of Project etc

Japan Exascale System Development

Outline:

- Double-digits (higher)performance by 2020
- Push state of the art in power efficiency, scalability & reliability
- Enable unprecedented application capability
- AICS RIKEN in charge of exascale systems development
- Total project cost ca. JPY140 billion with about JPY 110 billion from the government's budget (JPY 1.2 billion for 2014)

Schedule:

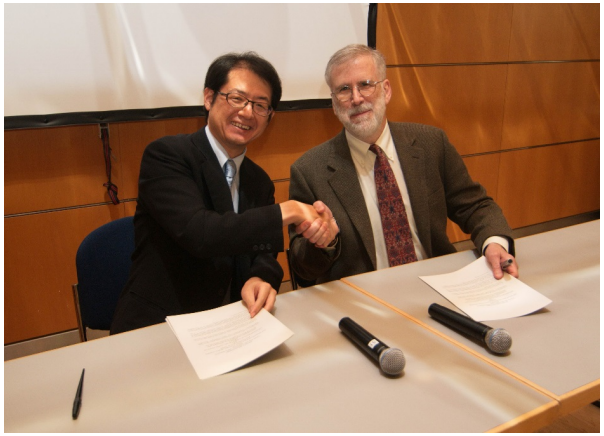
Basic design for GP & Accelerator components and thorough review the system specification

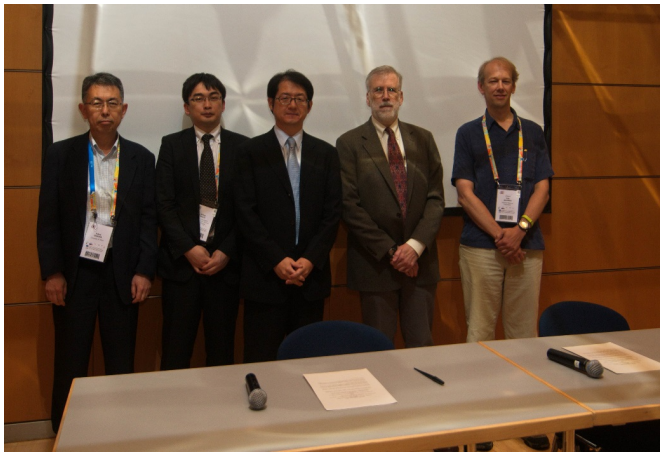
- Architecture: general-purpose plus accelerator components
- Target performance of 1 Exaflops, 100 times as powerful as K
- Power consumption of 30-40MW (cf. K computer: 12.7 MW)

Year of 2013	2014	2015	2016	2017	2018	2019	2020
System	Basic Design		Trial Production / Detailed Design		Manufacture	Installation / Coordination	Operation
Application	Development and utilization of Application for Exascale Computing						

1	China	Tianhe-2
2	USA	Titan
3	USA	Sequoia
4	Japan	K Computer
5	USA	Mira
6	Switzerland	Piz Daint
7	USA	Stampede
8	Germany	JUQUEEN
9	USA	Vulcan
10	USA	???

US-Japan Agreement





Yoshio Kawaguchi (MEXT, Japan) and William Harrod(DOE, USA)

Yoshio Kawaguchi
Director
Office for the Promotion of Computing Science,
Research Promotion Bureau
Ministry of Education, Culture, Sports, Science and Technology (MEXT)

Shinya TAHATA
Director for Information Science and Technology, Information Division,
Research Promotion Bureau
Ministry of Education, Culture, Sports, Science and Technology (MEXT)

Yutaka Ishikawa, Shinya Tahata, Yoshio Kawaguchi, William Harrod, Peter Beckman

Europe:

High Performance Computing in Horizon 2020

Big Data and Extreme Scale Computing Workshop
February 26-28, 2014 – Fukuoka Japan

Excellence in Science
DG CONNECT
European Commission

Jean-Yves Berthou ANR
on behalf of the European Commission Leonardo Flores, Panagiotis Tsarchopoulos, Aniyan Varghese

An integrated HPC approach | **Excellent Science**

- HPC strategy combining three elements:

(a) Computer Science: towards **exascale** HPC; *A special FET initiative focussing on the next generations of exascale computing as a key horizontal enabler for advanced modelling, simulation and big-data applications* [HPC in Future and Emerging Technologies (FET)]

(b) providing **access** to the best supercomputing facilities and services for both industry and academia; *PRACE - world-class HPC infrastructure for the best research* [HPC in e-infrastructures]

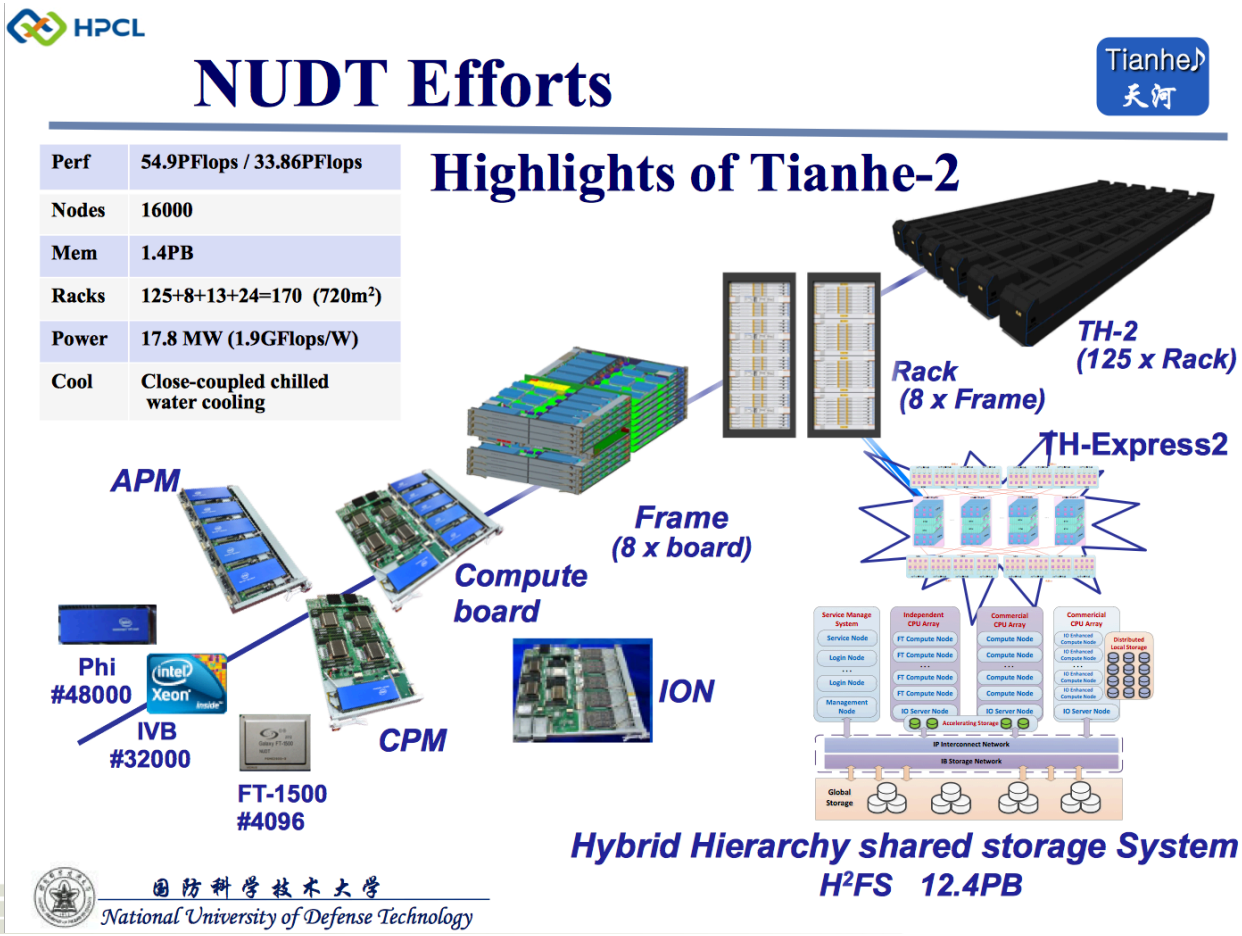
(c) achieving excellence in HPC **applications**; *Centres of Excellence for scientific/industrial HPC applications in (new) domains that are most important for Europe* [HPC in e-infrastructures]

- complemented with training, education and skills development in HPC

(a) and (c) will be implemented in the context of the HPC Public-Private Partnership

- ***HPC PPP** starting 1st January 2014: **700 m€ for the period 2014-2020** (€143,4 million in Calls in 2014-2015)*

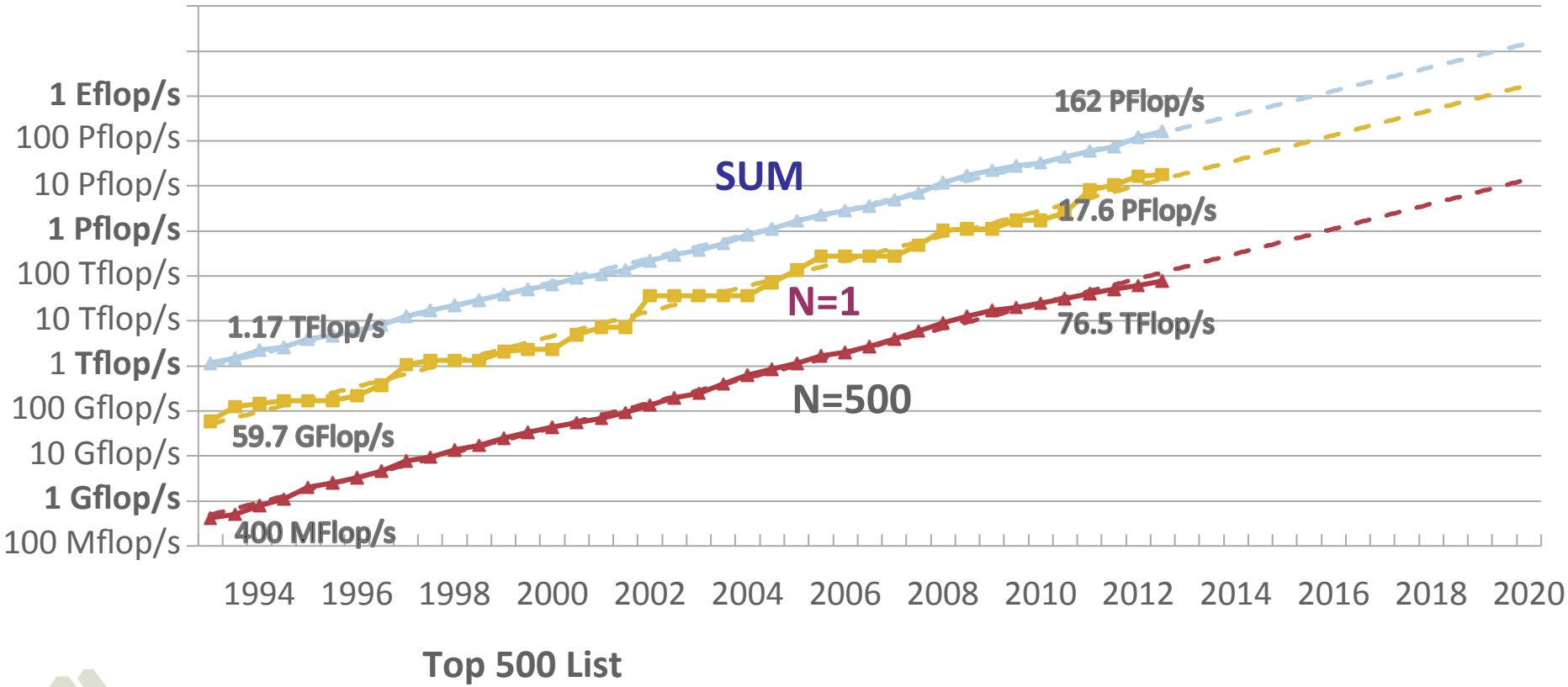
China: (currently with fastest machine)



Chinese Tianhe-2

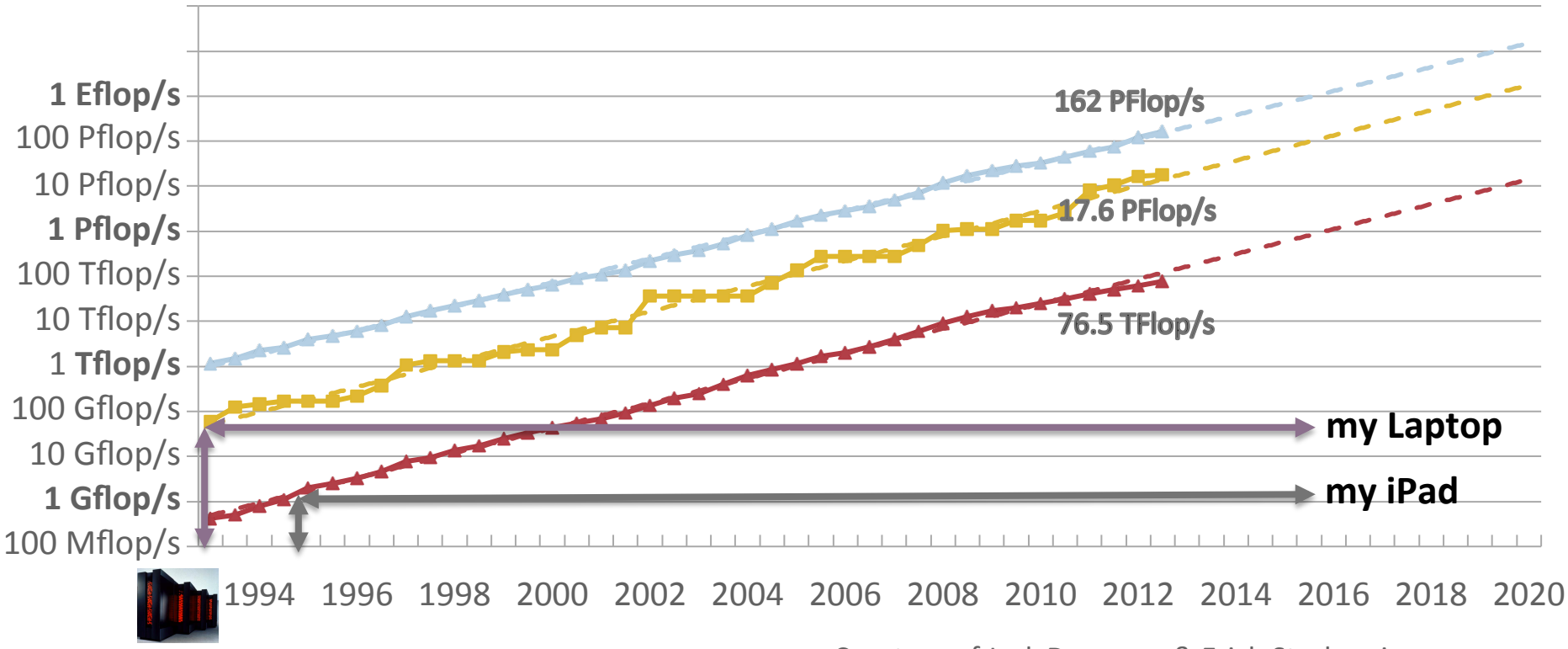
32K Intel Ivy Bridge Xeon Sockets
48K Intel Phi Sockets
~3M cores
~55 PF/sec peak
~30 PF/sec Linpack
~24 MW power with cooling

History of Supercomputer Performance



Courtesy of Jack Dongarra & Erich Strohmaier

History of Supercomputer Performance



Top 500 List

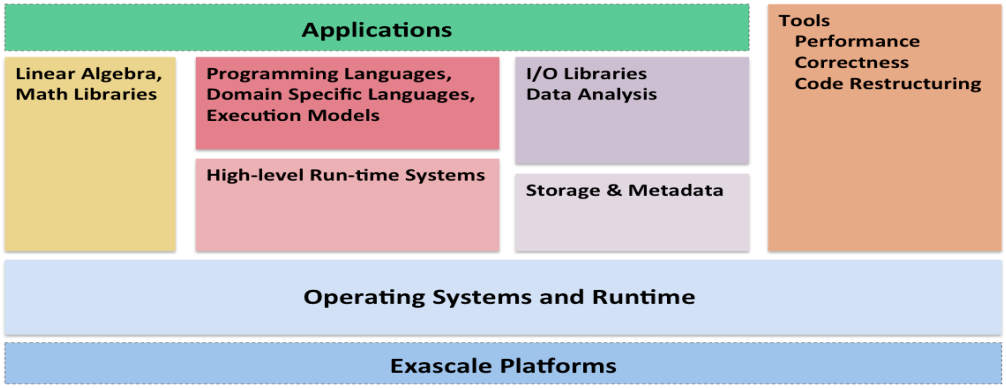
Courtesy of Jack Dongarra & Erich Strohmaier

To Reach Exascale: Will There be a Revolution?

The Software in the Middle: The Forces of Change

- Memory
- Threads
- Messaging
- Resilience
- Power

Co-Design Applications



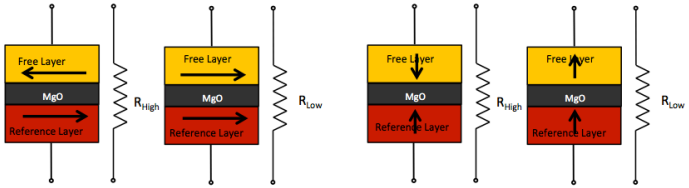
FastForward / DesignForward

Memory: Technology Summary from Rob Schreiber

New memory on the horizon

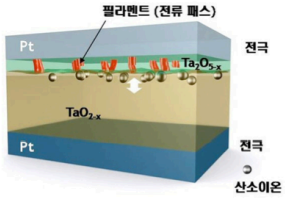
- Spin-Torque-Transfer RAM (STTRAM)
  - Grandis (54nm, acquired by Samsung)
- Phase-Change RAM (PCRAM)
  - Samsung (20nm, diode, up to 8Gb)
  - Micron and Nokia – In phones now
- Resistive RAM (ReRAM)
  - Panasonic (180nm process, 4-layer xpoint)
  - Unity Semi (64MB, acquired by Rambus)

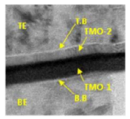
Spin transfer torque (STTRAM)



In-plane MTJ Perpendicular MTJ

ReRAM





Samsung, HP-Hynix, Sandisk, Toshiba

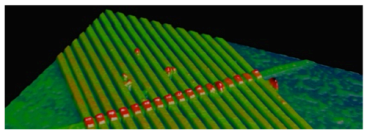
32Gb test chip (Sandisk/Toshiba. 24 nm. ISSCC 2013)

Fast (tens of nsecs) for both read and write

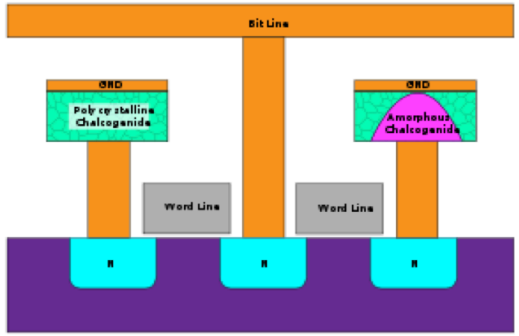
Good data retention and reliability

3D -- 2 to 4 layers

MLC possible

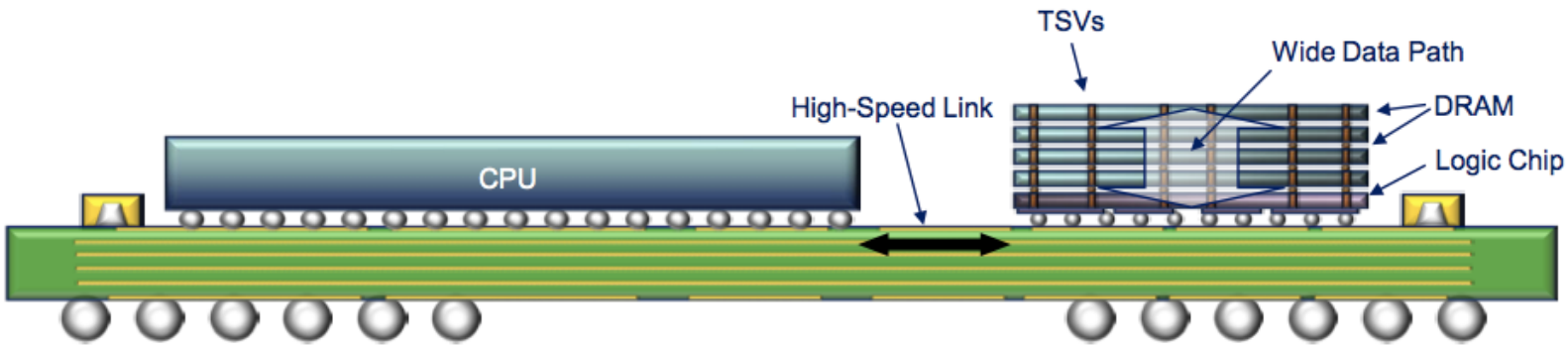


PCRAM



- Shipping today
- MLC (limited by resistance drift)
- Slow, expensive writes
- Wearout issue

Memory will be local, and include NVRAM



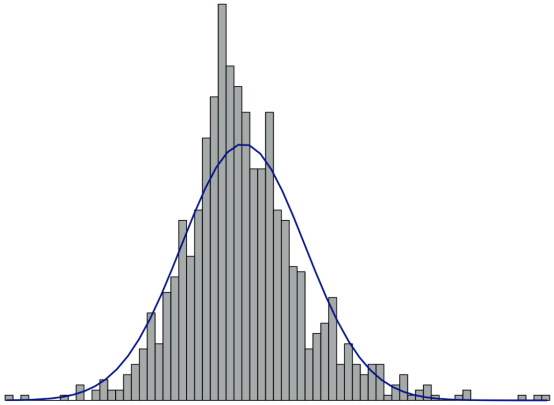
- Helps reduce power
- Helps with resilience
- Helps with cost
- Helps with **Big Data** (capacity)

Great! Where is the Revolution?
Programming Model and OS Interface:

- Need dynamic run-time to move ‘pages’ between RAM and NVRAM
- Need programming model that can hint (or manage) write-once and read only & data movement
- Coherence islands
- Add persistence to programming model
- Programming model for PIM

Threads/Tasks: Managing Exploding Parallelism

- Dynamic parallelism and decomposition
  - Programmer cannot hand-pick granularity / resource mapping
    - (equal work != equal time)

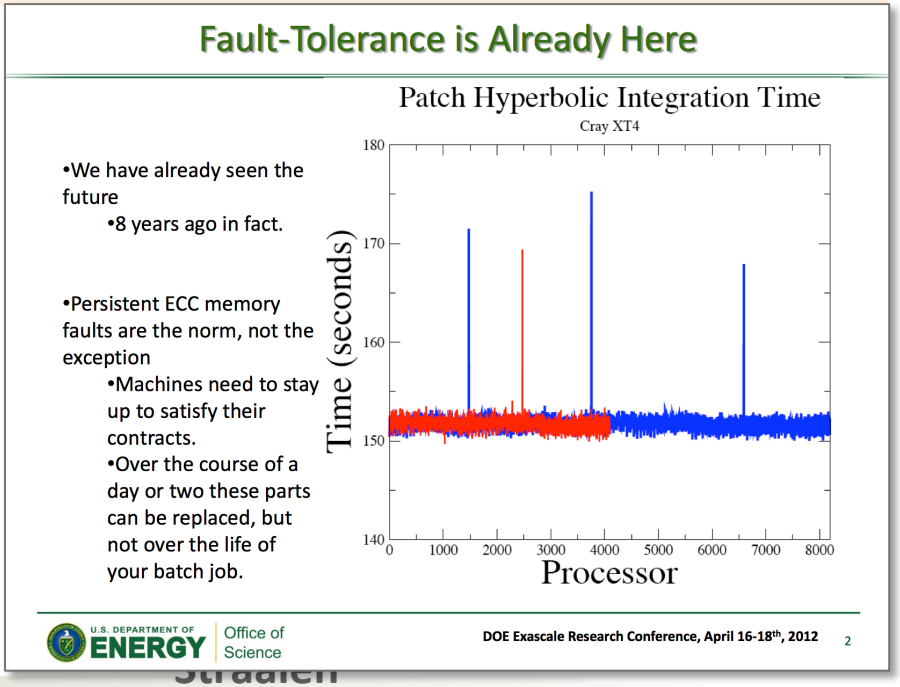


Variability is the new norm:

Power

Resilience

Intranode Contention



Google (re-discovers) OS Noise & Contention

Software techniques that tolerate latency variability are vital to building responsive large-scale Web services.

BY JEFFREY DEAN AND LUIZ ANDRÉ BARROSO

The Tail at Scale

Component-Level Variability Amplified By Scale

A common technique for reducing latency in large-scale online services is to parallelize sub-operations across many different machines, where each sub-operation is co-located with its portion of a large dataset. Parallelization happens by fanning out a request from a root to a large number of leaf servers and merging responses via a request-distribution tree. These sub-operations must all complete within a strict deadline for the

Living with Latency Variability

The careful engineering techniques in the preceding section are essential for building high-performance interactive services, but the scale and complexity of modern Web services make it infeasible to eliminate all latency variability. Even if such perfect behavior could

Reducing Component Variability

Interactive response-time variability can be reduced by ensuring interactive requests are serviced in a timely manner

-second service-level response time as the system scales and frequency gh-latency outliers varies.

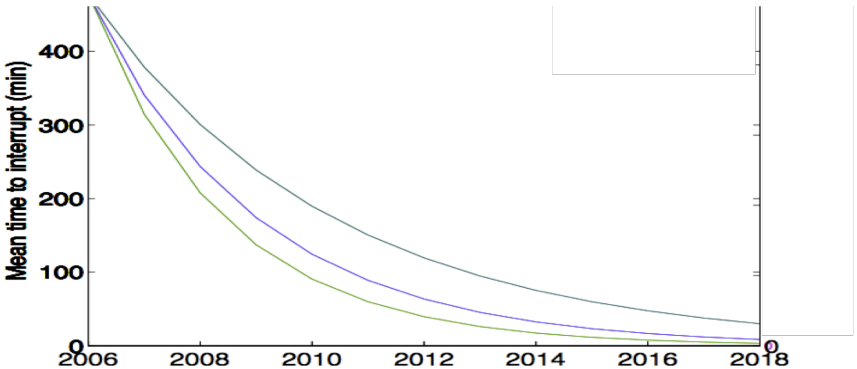
Messaging (send/recv & put/get): Where is the Revolution?

- We have done inter-node messaging for decades… what is new?
- Millions of outstanding threads (messages) per socket to hide latency
- Hardware Wake-On Threads (no polling)
  - BG/Q
  - X86 Mwait
- Network -> Cache/NVRAM Injection
- New OS/R Interfaces

Fault Predictions are Hard

What will be the Revolution?

Example Prediction from 2007

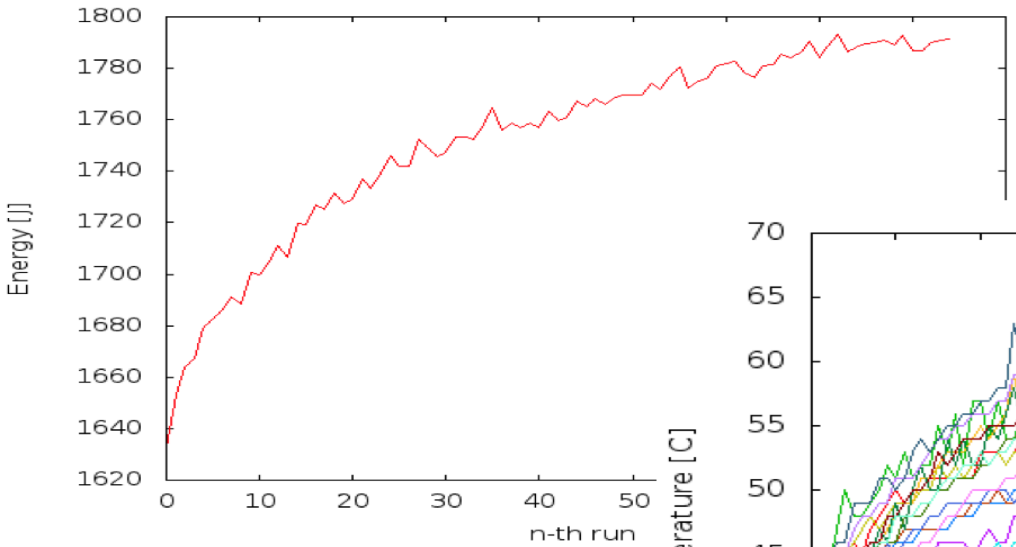


"Over the past thirty years there have been several predictions of the eminent cessation of the rate of improvement in computer performance. Every such prediction was wrong. They were wrong because they hinged on unstated assumptions that were overturned by subsequent events."

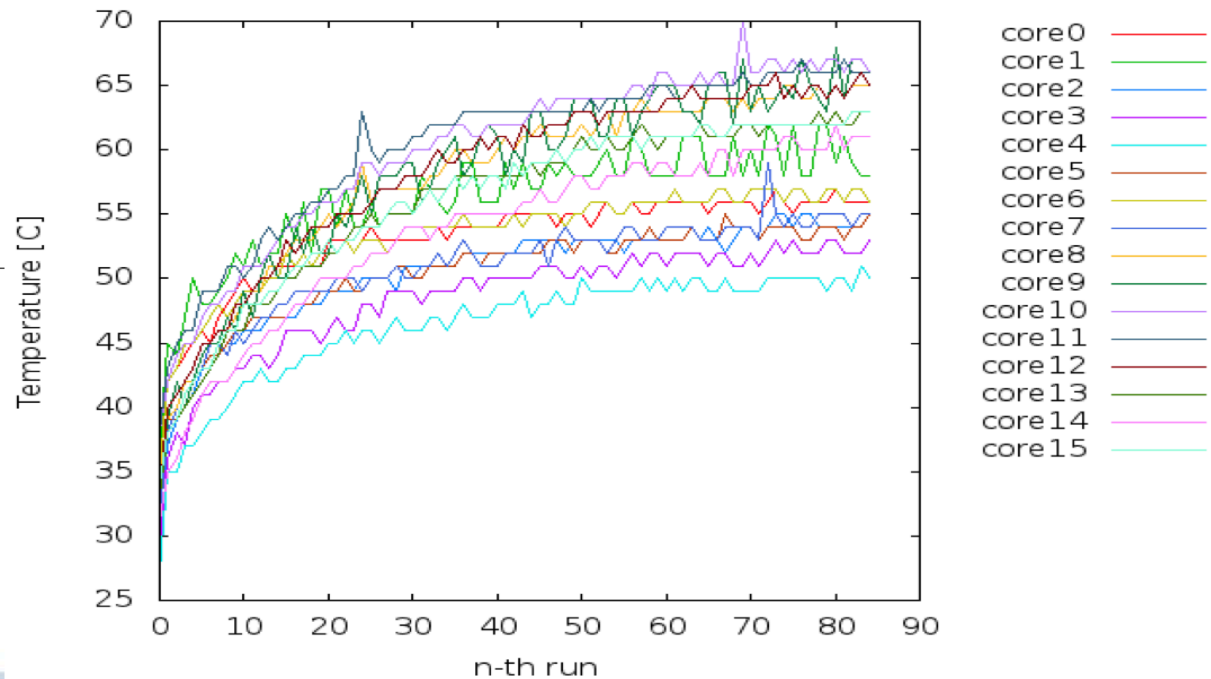
What we do know:

- Driving down power increases faults
- Vendors have great market incentive to redesign for reliable hardware
- Our current HPC software is very fragile
- We should improve resilience
- Build solutions at multiple layers

Exploring Power and Temp



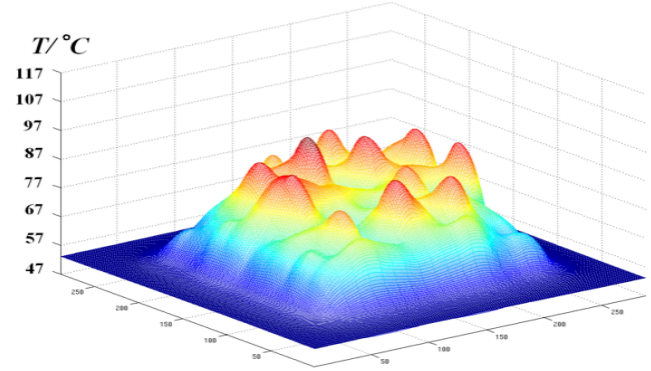
Sandy Bridge
8 cores x 2 sockets

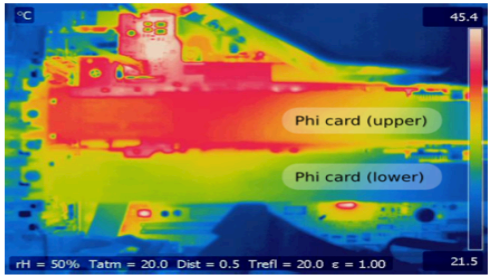


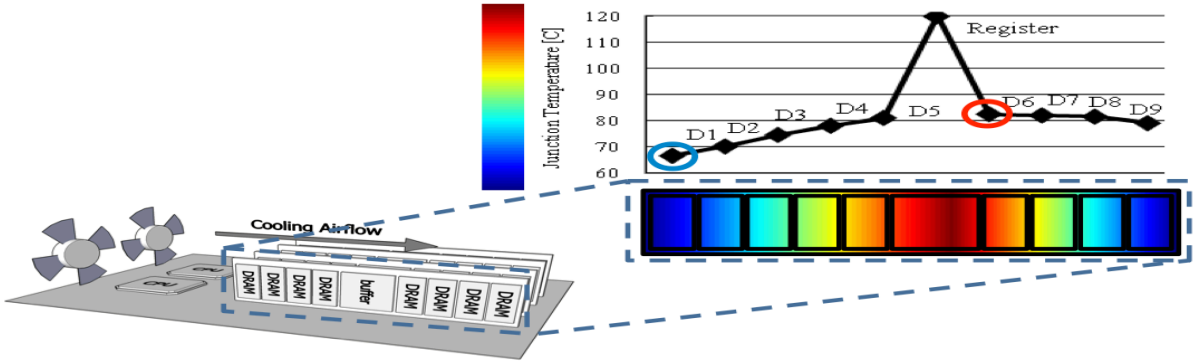
CMOS-based thermal sensors available via MSR (cpuonline)

Thermal Effects in High Performance Systems

- Temperature variation
  - Across cores, chips, memory versus cores, across nodes
    - input data dependent
    - spatial and temporal variation
    - hard to predict at design time
    - dynamic architecture highly desired







NORTHWESTERN UNIVERSITY

Courtesy: Seda Öğrenci Memik

Revolution Areas:

- Architecture
- Memory
- Threads
- Messaging
- Resilience
- Power

Questions? (then time for some more fun topics)