Argonne NATIONAL LABORATORY

Reinventing System Software for Exascale Supercomputers

Pete Beckman

Director, Exascale Technology and Computing Institute (ETCi)

Argonne National Laboratory

U.S. DEPARTMENT OF ENERGY

First: Quick Survey
What's Happening in Exascale?

INTERNATIONAL EXASCALE SOFTWARE PROJECT

Published in the International Journal of High Performance Computing Applications

Jack Dongarra, Pete Beckman, Terry Moore, Jean-Claude Andre, Jean-Yves Berthou, Taisuke Boku, Franck Cappello, Barbara Chapman, Xuebin Chi, Alok Choudhary, Sudip Dosanjh, Al Geist, Bill Gropp, Robert Harrison, Mark Hereld, Michael Heroux, Adolfy Hoisie, Koh Hotta, Yutaka Ishikawa, Fred Johnson, Sanjay Kale, Richard Kenway, Bill Kramer, Jesus Labarta, Bob Lucas, Barney Maccabe, Satoshi Matsuoka, Paul Messina, Bernd Mohr, Matthias Mueller, Wolfgang Nagel, Hiroshi Nakashima, Michael E. Papka, Dan Reed, Mitsuhisa Sato, Ed Seidel

Build an international plan for coordinating research for the next generation open source software for scientific high-performance computing

SPONSORS

Where We Are Today:

- Ken Kennedy – Petascale Software Project (2006)
- SC08 (Austin TX) meeting to generate interest
- Funding from DOE's Office of Science & NSF Office of Cyberinfratructure and sponsorship by Europeans and Asians

2008

- US meeting (Santa Fe, NM) April 6-8, 2009
  - 65 people

2009

- European meeting (Paris, France) June 28-29, 2009
  - Outline Report
- Asian meeting (Tsukuba Japan) October 18-20, 2009
  - Draft roadmap and refine report
- SC09 (Portland OR) BOF to inform others
  - Public Comment; Draft Report presented
- European meeting (Oxford, UK) April 13-14, 2010
  - Refine and prioritize roadmap; look at management models

2010

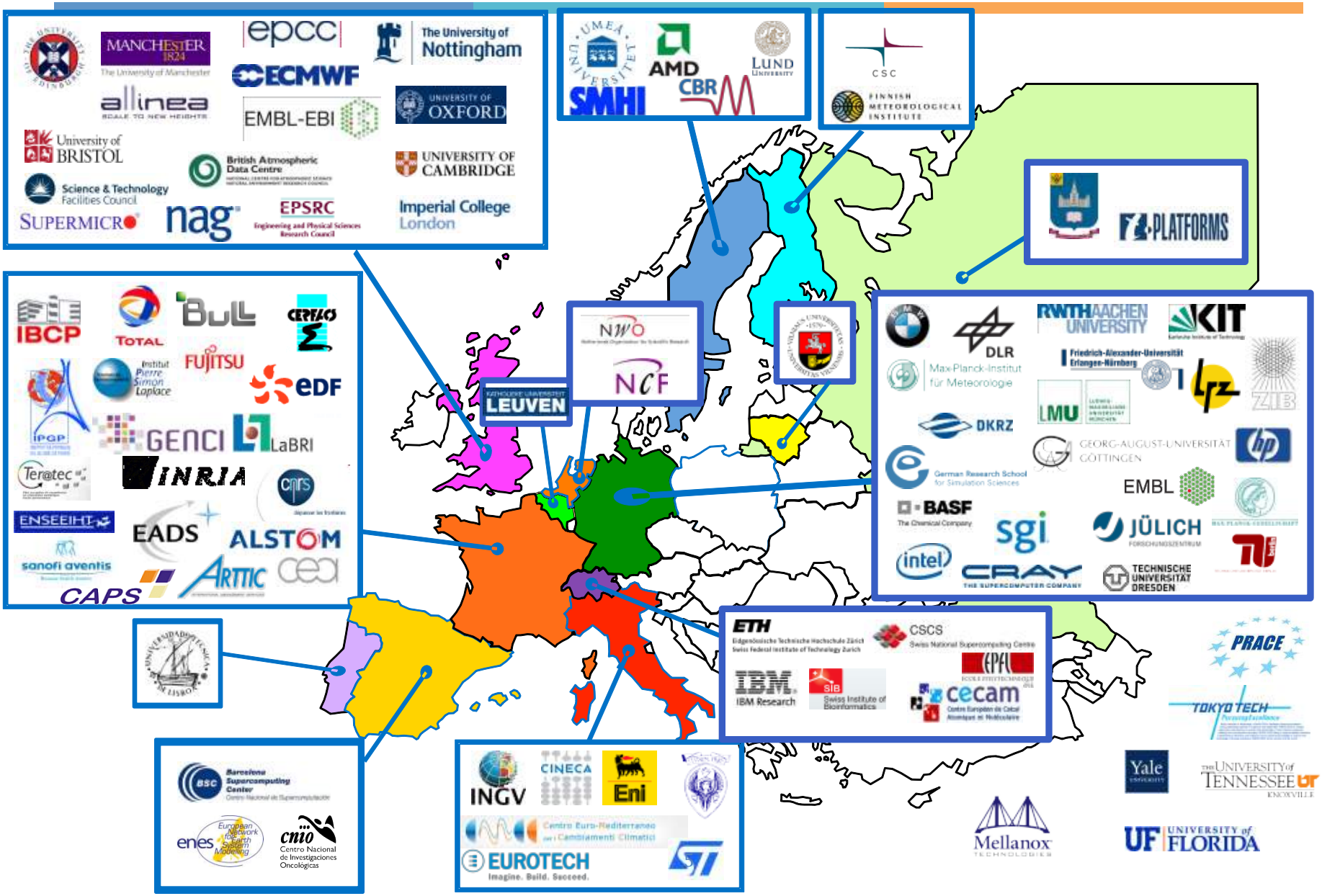
- Maui Meeting October 18-19, 2010
- SC10 (New Orleans) BOF to inform others
- SanFran (Kyoto) Meeting – April 6-7, 2011

2011

- Cologne, Germany – October 6-7, 2011
- Kobe, Japan – April 12-13, 2012

2012

EESI: 150 participants, 100 entities



EESI recommendations – funding targets

European Exascale hardware platforms

Fund 2 to 3 R&D projects **each** aiming at delivering one Exascale hardware platform in 2020, because:

- **need for different kind of architectures for different kind of needs**: weak and strong scaling, capacity/capability, big data oriented computers, field dedicated architectures-*Anton like machine for Molecular Dynamics*(?)
- **Exascale computing is a technological breakthrough compared to Tera and Petascale**, several technological options will have to be addressed
- **need competition between different providers** in order to have (at least) a world class vendor in Europe

Proposal:

- One platform built in Europe, integrated by a European vendor, with mainly European technology.

Should not only target a prototype but a product

- One or two platforms developed in collaboration with non European partners (US, Japan, China, Russia, …) and possibly assembled in Europe

EESI recommendations – budget

Global estimated budget : 2,5 to 3,5 billion euros over the next 10 years

- **Research and Development projects:** 1000 Meuros during 10 years
- **Co-design program**: between 500 to 1000 Meuros for 5 to 10 co-design centers, during 10 years
- **Exascale hardware platforms:** , between 500 to 1000 Meuros for 2 to 3 platforms, during 10 years
- **Technological transfer** through a European Exascale Software Center: 500 Meuros during 10 years

Three Exascale Platform Projects Started in Oct-2011 to Explore European Prototype Architectures

- Goal: jumpstart exascale platforms for Europe
- Joint funding: EC + (some) member states
- Immediate investment modest; $63M total across 3 years ($21M/year)
  - **Mont-Blanc** Project (14.5M€ total)
    - European: ARM (UK), STMicro (France/Italy), BULL (France)
    - + research teams from labs / universities
  - **DEEP** Project (18.5M€ total)
    - EU / US: EXTOLL(German), Intel (US)
    - + research teams from labs / universities
  - **CRESTA** Project (12M€ total)
    - Vampir (German), Cray (UK), Allinea (UK)
    - + research teams from labs / universities

- EESI Plan requests significant, sustained investments in 2 or 3 tracks for 2012
  - 500M€ - 1000M€ over 10 years

EU Announced Funding…

EU to double supercomputing funding to €1.2bn

By Jack Clark, ZDNet UK, 16 February, 2012 16:11 Follow @mappingbabel

Daily Newsletters

Sign up to ZDNet UK's daily newsletter.

Topics

HPC, Supercomputing, Neelie Kroes, European Commission, High-performance computing, Exascale, Exaflop, Petaflop, Curie, Top500, Investment, Funding, PRACE

Sponsored Links

SPSS Business Analytics
Get IBM SPSS Analytic Case Study. See How Top Companies Use SPSS.
www.ibm.com

Foreigner in Japan?
Are Japanese Banks Increasing Your Wealth? Put Your YEN to Work!
www.ObjectiveTrading.co

NEWS Supercomputing in Europe is set to get a boost after the European Commission announced plans to double its funding of high-performance computing.

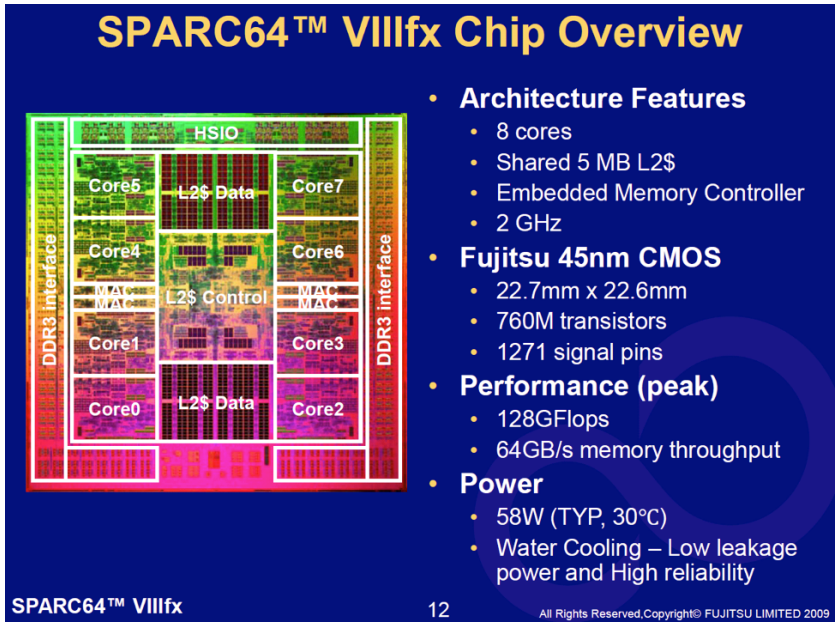
Annual investment in supercomputing equipment, training and research will go from €630m (£522m) to €1.2bn to help Europe "reverse its relative decline in HPC use and capabilities", the Commission said in a statement on Wednesday.

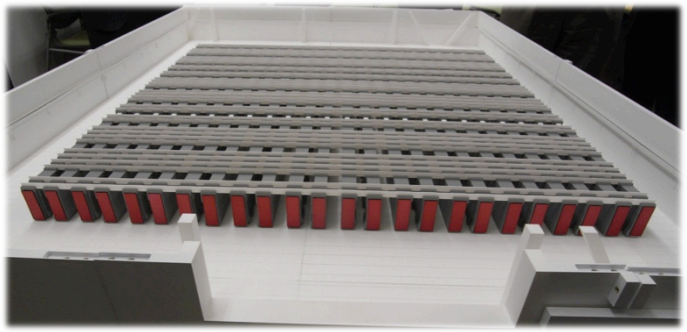
The EU has doubled its funding for supercomputing projects to €1.2bn. Pictured: the MareNostrum computer at the Barcelona Supercomputing Center. *Image credit: Barcelona Supercomputing Center*

Congratulations!

Japan: Current #1: The "K" Computer

The heart of the K computer consists of 80,000 Fujitsu's SPARC64 VIIIfx CPUs





864 Cabinets
10PFlops
1PB

24 Boards / Cabinet

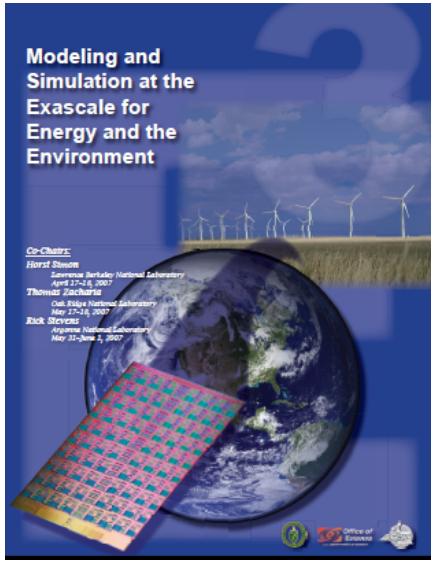
Fujitsu SPARC64™IXfx

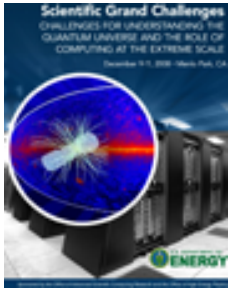
Sept 2011: New chip announced

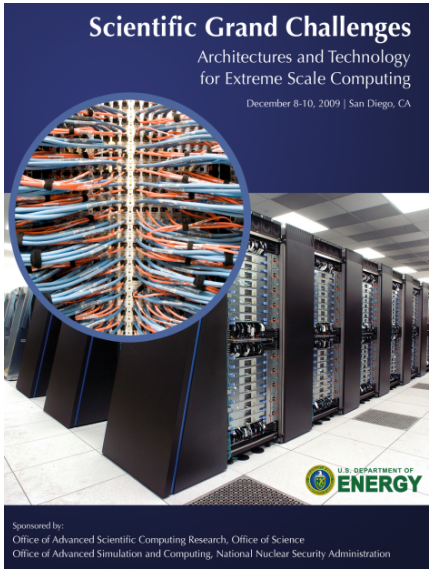
An amazing accomplishment, with unique and advanced system software

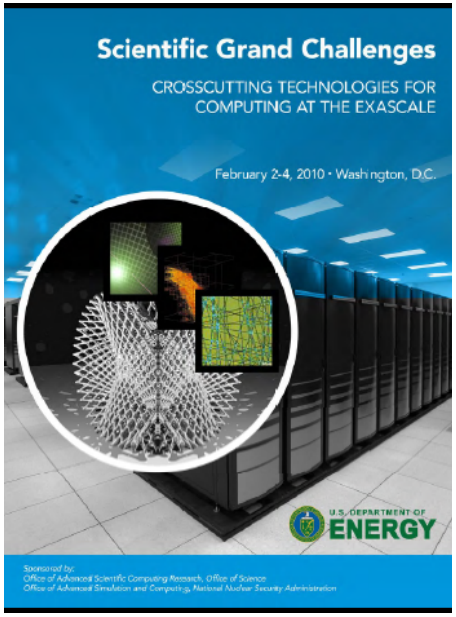
DOE Workshops and Reports

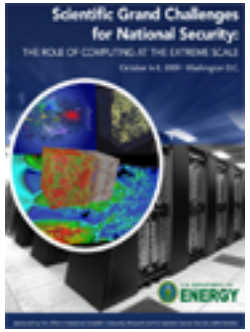
http://science.energy.gov/ascr/news-and-resources/program-documents/











USA:

X-Stack Research Software

- Programming Models
- Storage & File systems,
- Etc.

Three co-design centers established

- Advanced Nuclear Reactors
- Combustion
- Materials in Extreme Environments

What is Co-Design? A buzz-word? Does it just mean “applications”? Does it have a process?

Understanding Co-design (examples from BG/Q)

Architecture ▪ Added MMU ▪ # of cores ▪ # of chips	**General** ▪ Efficiency and safety improvements in the power module design ▪ Open source model ▪ Many MPI improvements
Operating System ▪ # of SW threads ▪ Smart scheduler ▪ Reduced memory footprint ▪ SW Env: (python, shared libs) ▪ Speculative multi-threading system programming interface	**System Management** ▪ Distributed control system ▪ New pervasive security model ▪ Open source, plug-in dynamic allocator ▪ Many RAS usability and performance improvements
Code Development & Tools ▪ Improved performance counters ▪ Improved behavior for TLS and TM to better match application needs ▪ Flexible programming model - MPI everywhere; flexible task/thread ratio ▪ Increased user level APIs	**I/O** ▪ Full size, standard PCI-e cards ▪ Debugger interfaces for IO nodes ▪ Persistent memory uses ▪ Page sizes and scalability improvements

Exascale "RFI" conceptual roadmap
22 responses from companies

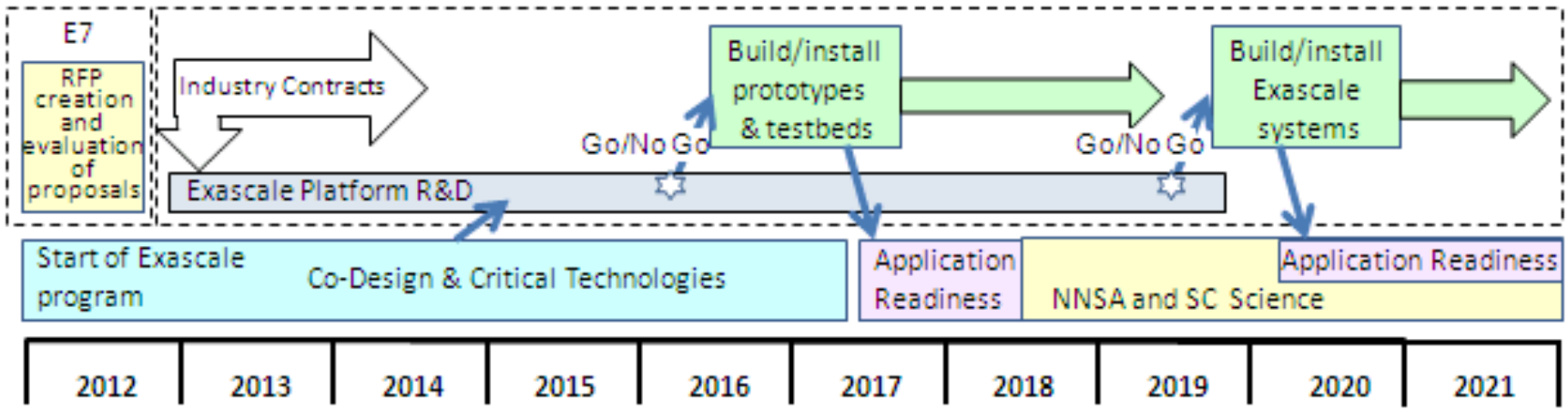


Table 1. Exascale System Goals

Exascale System	Goal
Delivery Date	2019
Performance	1000 PF LINPACK and 300 PF on to-be-specified applications
Power Consumption*	20 MW
MTBAI**	6 days
Memory including NVRAM	128 PB
Node Memory Bandwidth	4 TB/s
Node Interconnect Bandwidth	400 GB/s

*Power consumption includes only power to the compute system, not associated storage or cooling systems.

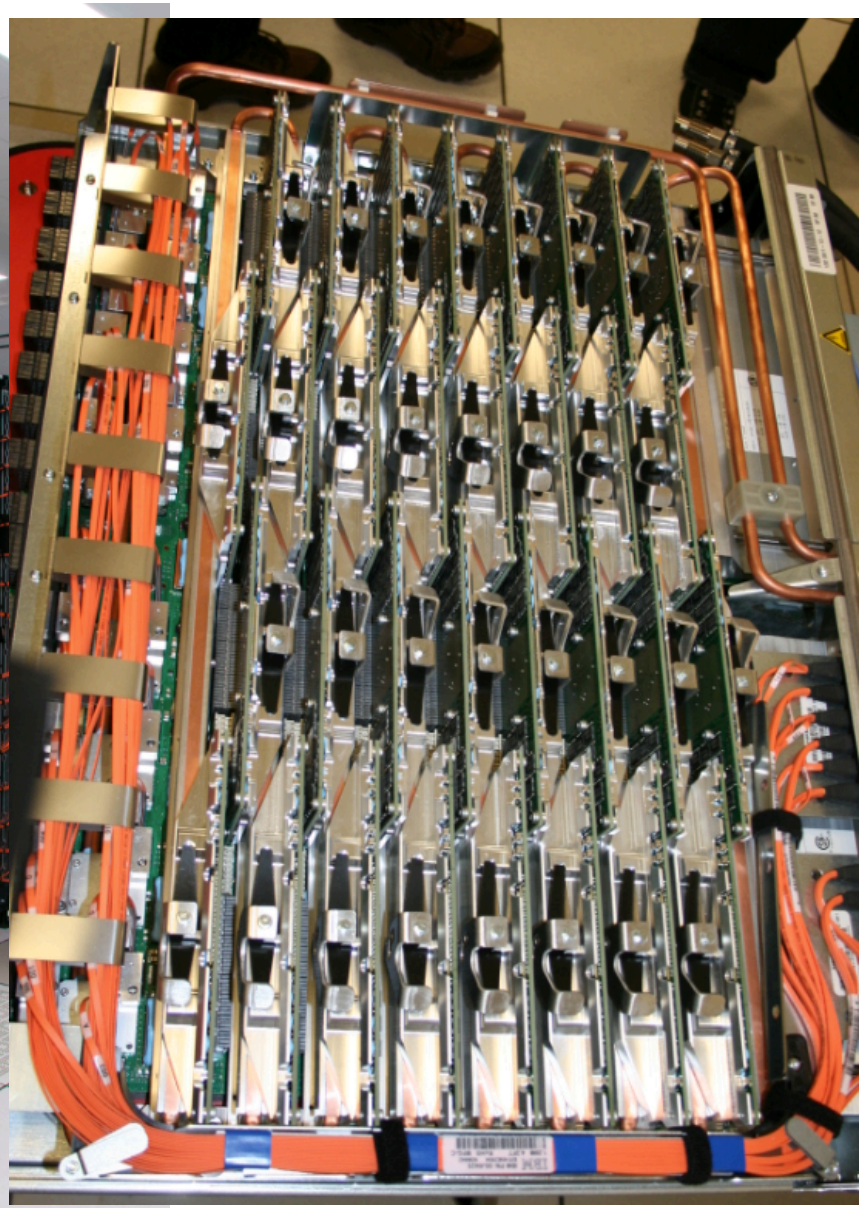
**The mean time to application failure requiring any user or administrator action must be greater than 24 hours, and the asymptotic target is improvement to 6 days over time. The system overhead to handle automatic fault recovery must not reduce application efficiency by more than half.

PF = petaflop/s, MW = megawatts, PB = petabytes, TB/s = terabytes per second, GB/s = gigabytes per second, NVRAM = non-volatile memory.

New at Argonne: BLUE GENE/Q

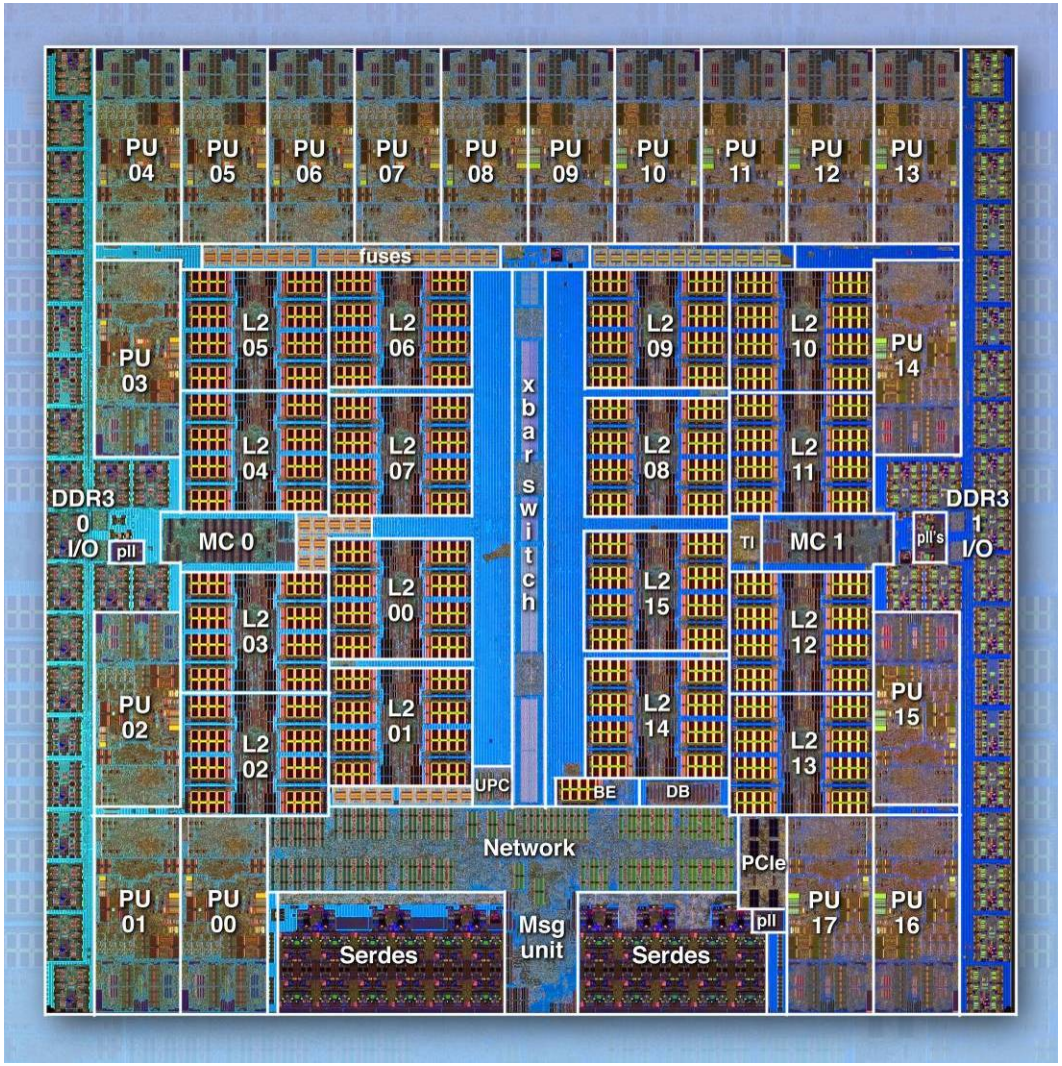
- *Mira* - Blue Gene/Q System
  - 48 racks
  - 48K 1.6 GHz nodes
  - 768K cores & 786TB RAM
  - 384 I/O nodes
  - Peak: 10PF
- Storage
  - ~35 PB capacity, 240GB/s bandwidth (GPFS)
  - Disk storage upgrade planned in 2015
    - Double capacity and bandwidth
- New Visualization Systems
  - Initial system in 2012
  - Advanced visualization system in 2014
    - State-of-the-art server cluster with latest GPU accelerators
    - Provisioned with the best available parallel analysis and visualization software

Two racks of BG/Q are installed and running!



BlueGene/Q Compute chip

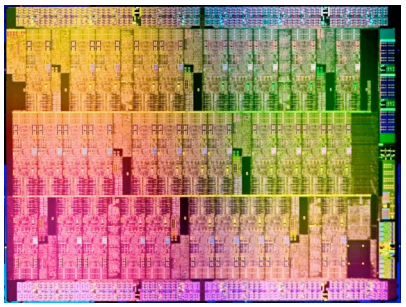
System-on-a-Chip design : integrates processors, memory and networking logic into a single chip



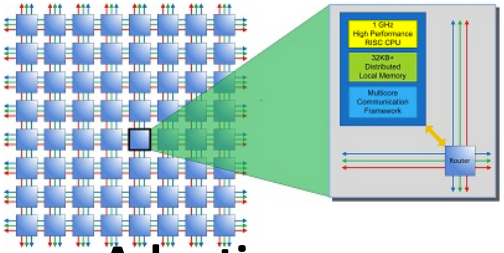
- **360 mm² Cu-45 technology (SOI)**
  - ~ 1.47 B transistors
- **16 user + 1 service processors**
  - plus 1 redundant processor
  - all processors are symmetric
  - each 4-way multi-threaded
  - 64 bits PowerISA™
  - 1.6 GHz
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
  - each processor has Quad FPU (4-wide double precision, SIMD)
  - peak performance 204.8 GFLOPS@55W
- **Central shared L2 cache: 32 MB**
  - eDRAM
  - multiversioned cache will support transactional memory, speculative execution.
  - supports atomic ops
- **Dual memory controller**
  - 16 GB external DDR3 memory
  - 42.6 GB/s
  - 2 * 16 byte-wide interface (+ECC)
- **Chip-to-chip networking**
  - Router logic integrated into BQC chip.
- **External IO**
  - PCIe Gen2 interface

Node Architecture is Rapidly Changing!

How Will System Software Manage CPUs?
How Will They Be Programmed?

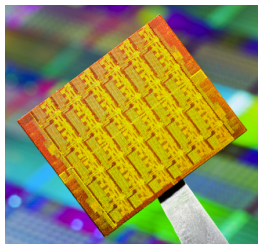


Intel: Knight's Ferry

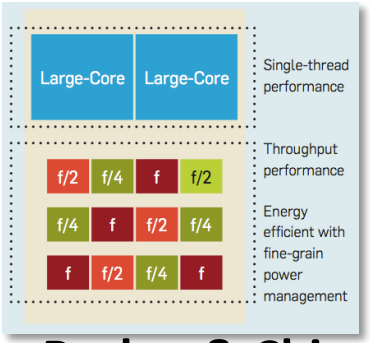


Adaptiva:

Power Constrained Memory Consistency



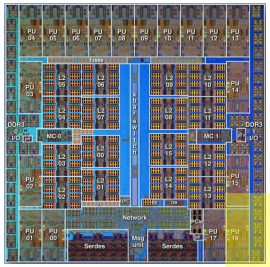
Intel: SCC



Borkar & Chien

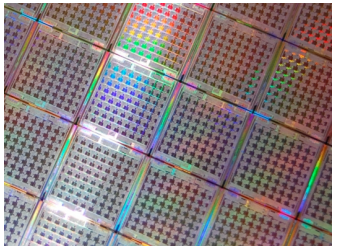
?

Exascale++

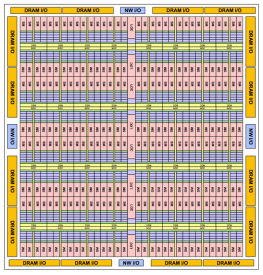


#18

IBM: BG/Q



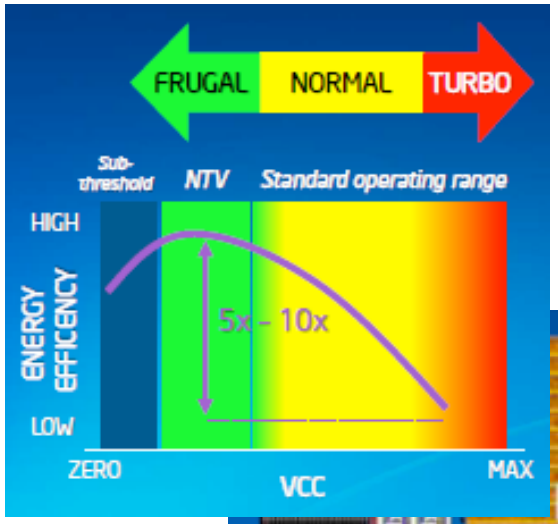
Tilera: GX

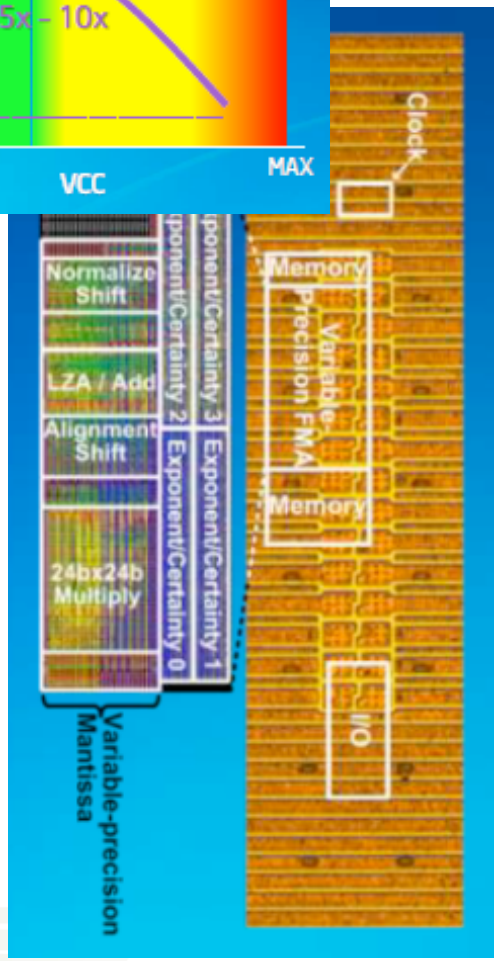


Dally: Echelon

Intel: NTV (Near Threshold Voltage) circuits & variable precision floating point

- In NTV range, 5 to 10 times more efficient
- Demonstrated chip that can go from 3Ghz to 915Mhz
- Prototype variable precision floating point system

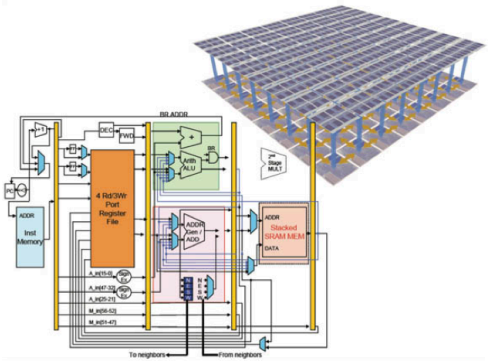




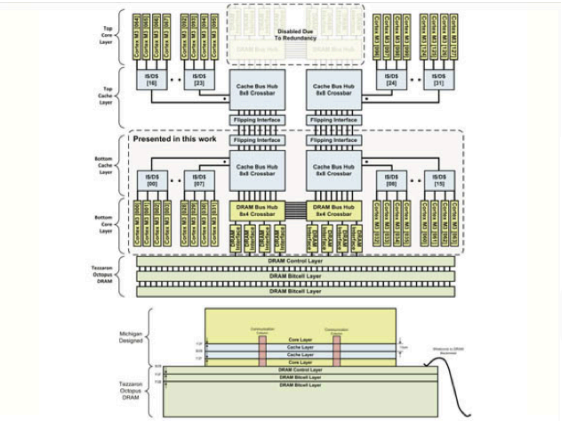
- **Where is the system software?**
  - **Control power based on computation load?**
  - **Control precision based on required error bounds?**

Really Fast, Really Close, Really Small 3D memory…

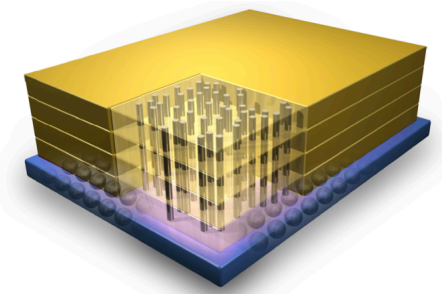
- On-chip RAM small in comparison to today's systems
- Bandwidth will be excellent
- Advanced HPC memory operations possible
- Integrated NIC is the next step
- New data movement methods
- **Where is the system software?**
  - **Memory management for scratchpad, cache, memory power, etc?**
  - **OS that controls threads, tasks, and power**



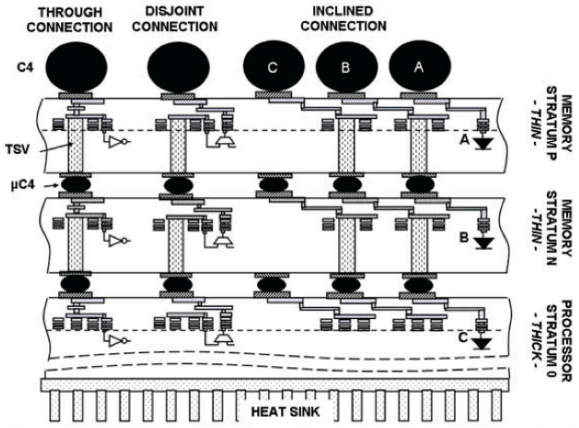
Georgia Tech



Univ of Michigan



Micron HMC



IBM

"Early benchmarks show a memory cube blasting data 12 times faster than DDR3-1333 SDRAM while using only about 10 percent of the power."

We Are Far From Solving In-Socket Parallel Programming:

```
#pragma omp parallel for        \
  default(shared) private(i)    \
  schedule(static,chunk)        \
  reduction(+:result)

  for (i=0; i < n; i++)
    result = result + (a[i] * b[i]);

printf("Final result= %f\n",result);
```

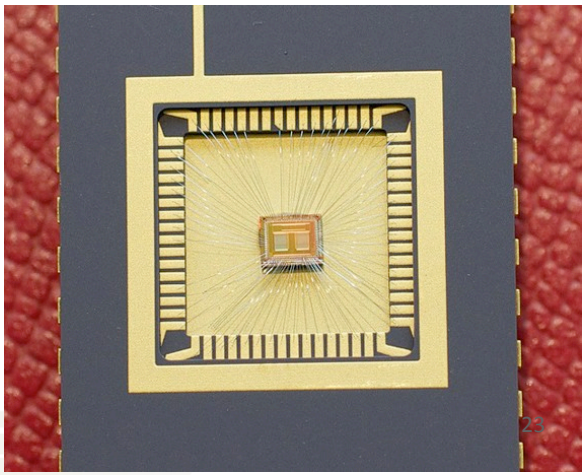
```
float function FTNReductionOMP(data, size)
float data(*)
integer size
ret = 0.0

!dir$ omp offload target(mic) in(size) in(data:length(size))
!$omp parallel do reduction(+:ret)
do i=1,size
        ret = ret + data(i)
enddo
!$omp end parallel do

FTNReductionOMP = ret
```

Clause	Directive					
	PARALLEL	DO/for	SECTIONS	SINGLE	PARALLEL DO/for	PARALLEL SECTIONS
IF	●				●	●
PRIVATE	●	●	●	●	●	●
SHARED	●	●			●	●
DEFAULT	●				●	●
FIRSTPRIVATE	●	●	●	●	●	●
LASTPRIVATE		●	●		●	●
REDUCTION	●	●	●		●	●
COPYIN	●				●	●
COPYPRIVATE				●		
SCHEDULE		●			●	
ORDERED		●			●	
NOWAIT		●	●	●		

- We do not yet have a good in-socket parallel programming model
- OpenMP is a mess
- **Where is the system software?**
  - **Memory management for scratchpad, cache, memory power, etc?**
  - **OS that controls threads, tasks, and power**

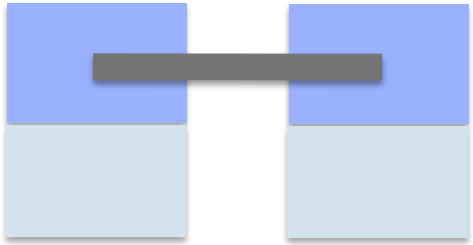
Power, Parallelism, Coherence, Fault, Storage

- Power must be a managed resource
  - More functional units than can run at full speed
  - Variable speed subcomponents
  - NEW… Most performance for a fixed watt
- Restructured node architecture
  - Memory and computing together on package
  - Massive levels of in-package parallelism
  - Variable coherence domains and intrasocket messaging
  - Heterogeneous multi-core
- Complex fault behavior
- NVRAM near computing



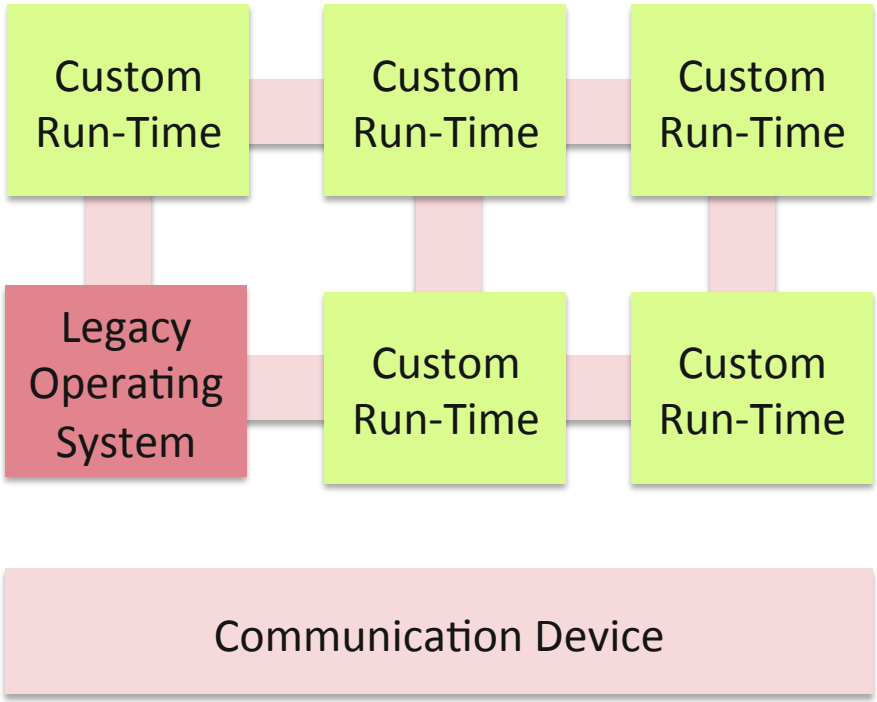
The Future: We must reinvent parallel programming Ignore mem/core -- Focus on computational intensity

- Many programmers are having difficulty with the “abstract machine”. What does the system look like to the programmer.
  - They have focused on “core” as the key architectural component
  - Very complex multi-level mapping: Socket, Core, Hardware Thread
- We can’t program to an exponentially changing component... (num cores)
- We must return to programming per coherence domain
  - (socket counts are changing very slowly)
- Programming model cannot be based on parallelism after the fact (openMP)



Future Designs?

asymmetric designs



Key Features

- User-managed custom run-times
- Power management
  - Dark silicon
  - Intel TurboBoost-like power control
  - Consistency control
- Resilience
  - Fault containment
- Messaging & Active Msgs
  - Inner and Outer Space
  - PGAS & Wakeon
- Distributed memory mgmt?

Questions?