IBM's Blue Gene/Q System and Implications for Simulation and Data Analysis

Kirk E. Jordan
Emerging Solutions Exec. & Assoc. Prog. Director
Computational Science Center
IBM T.J. Watson Research

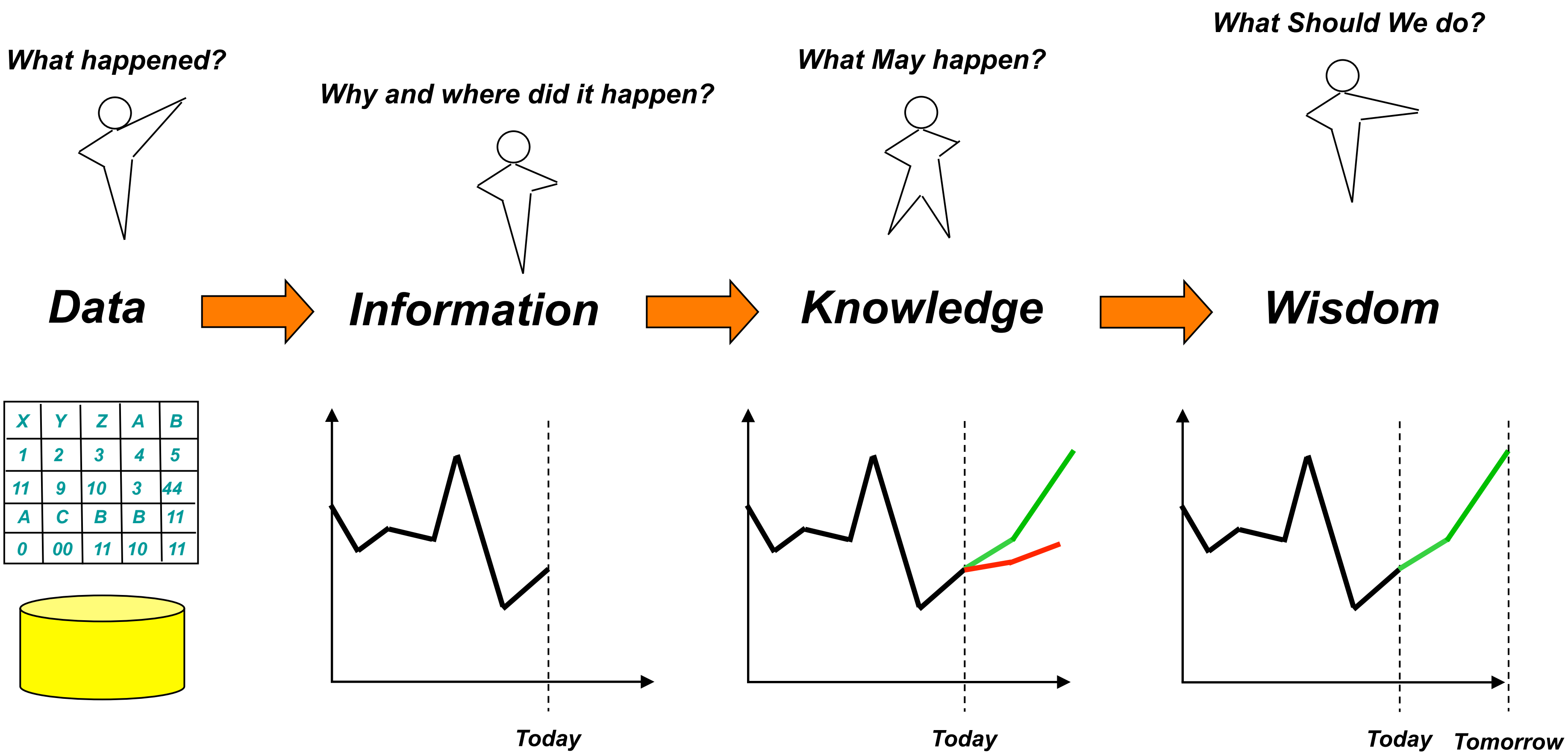
kjordan@us.ibm.com

Outline

- Background toward Exascale
- Overview of Blue Gene/Q Hardware & Software
- Smart Planet – SmartData/Analytics
- Comments Simulation and Analytics
- Closing Remarks

Future Foresight and Right Action

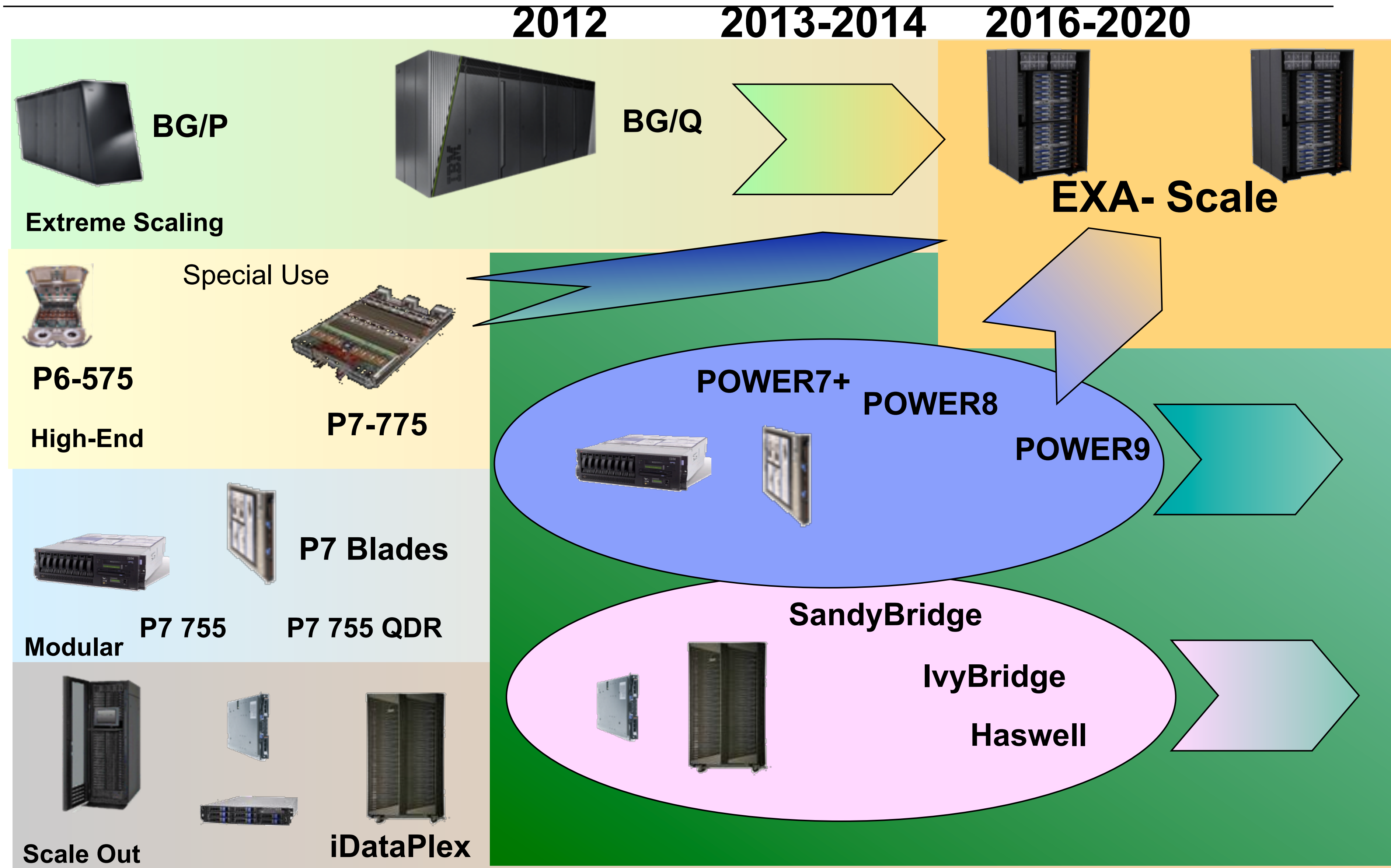
IBM Research



Simulation & Modeling

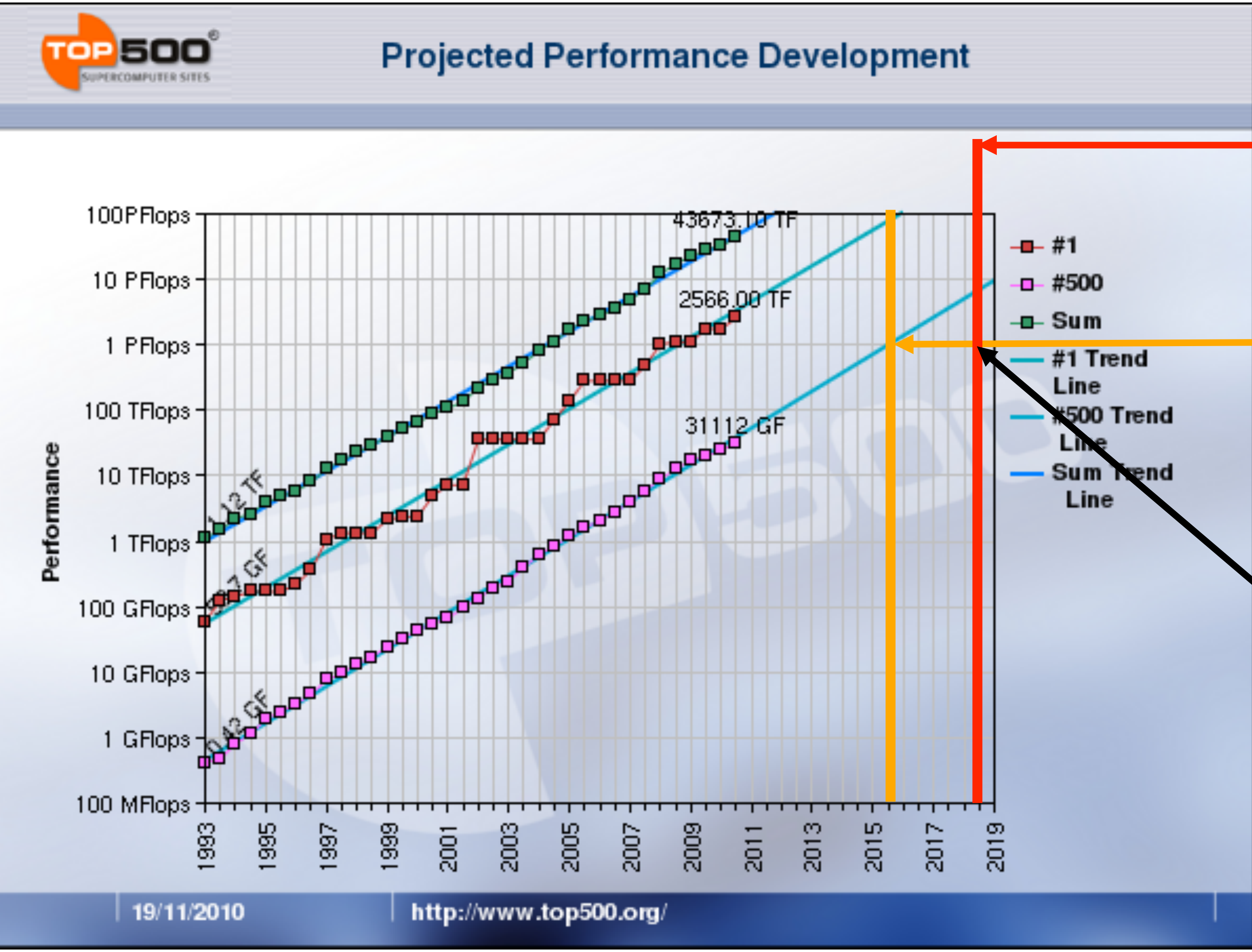
Roadmap

IBM Research



Trends in Computing Performance

IBM Research



1EFlop ~2nd half of 2018 – est cost $200M

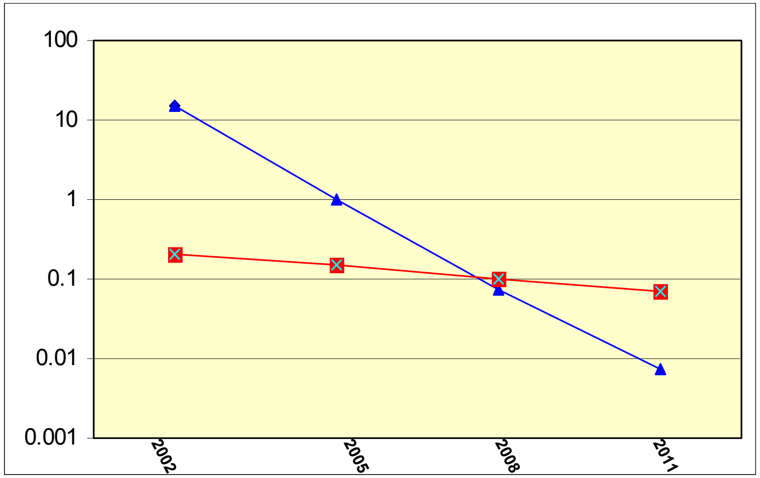
1PFlop ~2nd half of 2015 is bottom of Top500

1PFlop ~2nd half of 2018 – est. cost < $200K

Computer Design Challenges

- Core Frequencies ~
  - 2-4 GHz, will not change significantly as we go forward
  - 100,000,000 Cores to deliver an Exaflop
- Power
  - At today's MegaFlops / Watt: 2 GW needed (~$2B/yr)
  - Power reduction will force simpler chips, longer latencies, more caches, nearest neighbor network
- Memory and Memory Bandwidth
  - Much less memory / core (price)
  - Much less bandwidth / core (power / technology)
- Network Bandwidth
  - Much less network bandwidth per core (price / core) (Full fat tree ~$1B to $4B)
  - Local network connectivity
- Reliability
  - Expect algorithms / applications will have to permit / survive hardware fails.
- I/O Bandwidth
  - At 1 Byte / Flop, an EXAFLOP system will have 1 EXABYTE of Memory.
  - No disk system can read / write this amount of data in reasonable time. (BG/P 4TB ~1min but disk array ingest at ~15min)

GFLOPs vs DRAM Price Reductions



- Exascale Computing
  - O(100 M) compute engines working together
- Capability delivered has the potential to be truly revolutionary
- However
  - Systems will be complex
  - Software will be complex
  - Applications will be complex
  - Data Centers will be complex
  - Maintenance / Management will be complex

A suggested design point from Pete Beckman

Exascale Systems Targets

Systems	2009	2018	Difference Today & 2018
System peak	**2 Pflop/s**	**1 Eflop/s**	**O(1000)**
Power	**6 MW**	**~20 MW (goal)**	
System memory	0.3 PB	32 - 64 PB	O(100)
Node performance	125 GF	1.2 or 15TF	O(10) – O(100)
Node memory BW	25 GB/s	2 - 4TB/s	O(100)
Node concurrency	12	O(1k) or O(10k)	O(100) – O(1000)
Total Node Interconnect BW	3.5 GB/s	200-400GB/s (1:4 or 1:8 from memory BW)	O(100)
System size (nodes)	18,700	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	225,000	O(billion) + [O(10) to O(100) for latency hiding]	O(10,000)
Storage Capacity	15 PB	500-1000 PB (>10x system memory is min)	O(10) – O(100)
IO Rates	0.2 TB	60 TB/s	O(100)
MTTI	days	O(1 day)	- O(10)

From Rick Stevens: http://www.exascale.org/mediawiki/images/d/db/PlanningForExascaleApps-Steven.pdf

Top 10 Highlights of Blue Gene/Q (overview)

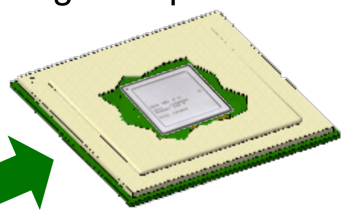
1. **Ultra-scalability for breakthrough science**
   - System can scale to 256 racks and beyond (>262,144 nodes)
   - Cluster: typically a few racks (512-1024 nodes) or less.
2. **Highest capability machine in the world (20-100PF+ peak)**
3. **Superior reliability: Run an application across the whole machine, low maintenance**
4. **Highest power efficiency, smallest footprint, lowest TCO (Total Cost of Ownership)**
5. **Low latency, high bandwidth inter-processor communication system**
6. **Low latency, high bandwidth memory system**
7. **Open source and standards-based programming environment**
   - Red Hat Linux distribution on service, front end, and I/O nodes
   - Lightweight Compute Node Kernel (CNK) on compute nodes ensures scaling with no OS jitter, enables reproducible runtime results
   - Automatic SIMD (Single-Instruction Multiple-Data) FPU exploitation enabled by IBM XL (Fortran, C, C++) compilers
   - PAMI (Parallel Active Messaging Interface) runtime layer. Runs across IBM HPC platforms
8. **Software architecture extends application reach**
   - Generalized communication runtime layer allows flexibility of programming model
   - Familiar Linux execution environment with support for most POSIX system calls.
   - Familiar programming models: MPI, OpenMP, POSIX I/O
9. **Broad range of scientific applicability at superior cost/performance**
10. **Key foundation for exascale exploration**

Blue Gene/Q

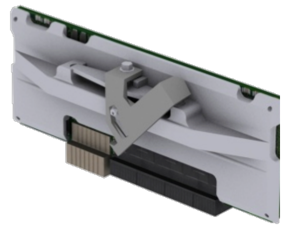
IBM Research

1. Chip:
16+2 µP
cores

2. Single Chip Module



3. Compute card:
One chip module,
16 GB DDR3 Memory,
Heat Spreader for H_2O Cooling



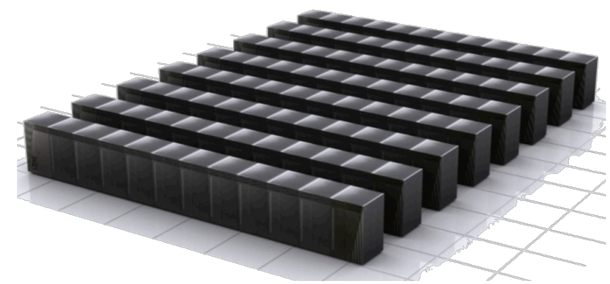
4. Node Card:
32 Compute Cards,
Optical Modules, Link Chips; 5D Torus

5b. IO drawer:
8 IO cards w/16 GB
8 PCIe Gen2 x8 slots
3D I/O torus

5a. Midplane:
16 Node Cards

6. Rack: 2 Midplanes

7. System:
96 racks, 20PF/s



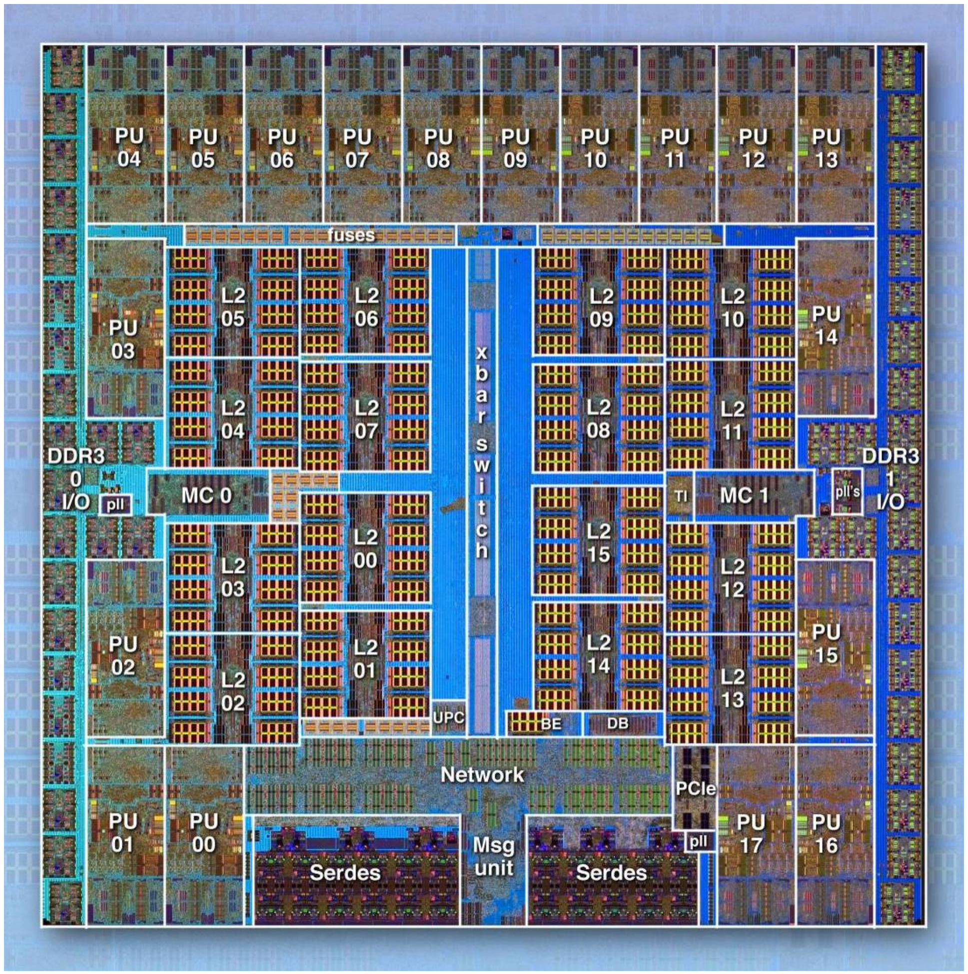
- **Sustained single node perf: 10x P, 20x L**
- **MF/Watt: (6x) P, (10x) L (~2GF/W, Green 500 criteria)**
- **Software and hardware support for programming models for exploitation of node hardware concurrency**

- **BG/Q (209 TF/rack, 2000 MF/W) – 45nm ASIC**
- **Scales >256 racks, 53.6 PF/s, 16 core/64 thread SOC 16 GB/ Node**
- **Speculative execution, sophisticated L1 prefetch, transactional memory, fast thread handoff, compute + IO systems**

BlueGene/Q Compute chip

IBM Research

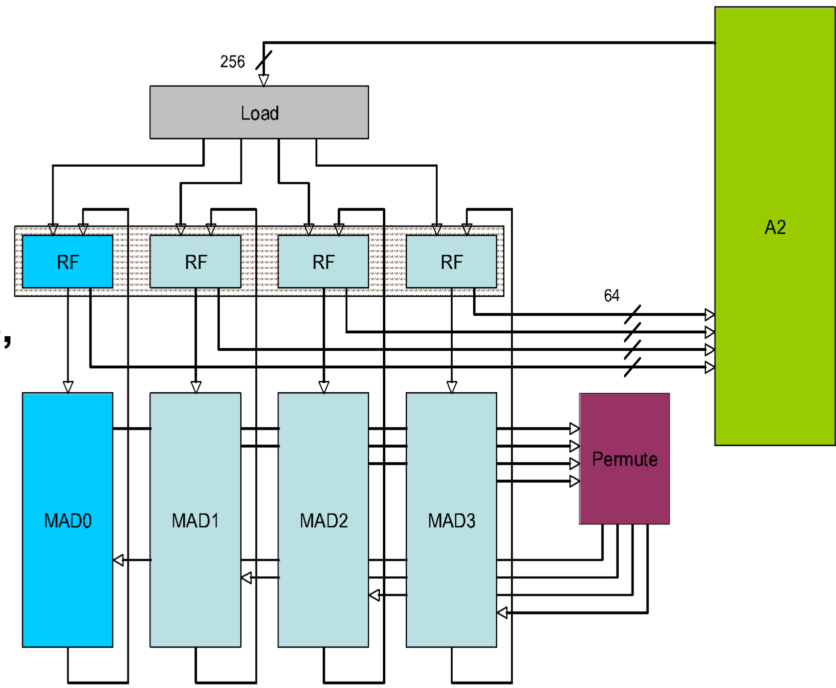
System-on-a-Chip design : integrates processors, memory and networking logic into a single chip



- **360 mm² Cu-45 technology (SOI)**
- **16 user + 1 service PPC processors**
  - plus 1 redundant processor
  - all processors are symmetric
  - 11 metal layer
  - each 4-way multi-threaded
  - 64 bits
  - 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 @ 55 W
- **Central shared L2 cache: 32 MB**
  - eDRAM
  - multiversioned cache – supports transactional memory, speculative execution.
  - supports scalable atomic operations
- **Dual memory controller**
  - 16 GB external DDR3 memory
  - 42.6 GB/s DDR3 bandwidth (1.333 GHz DDR3) (2 channels each with chip kill protection)
- **Chip-to-chip networking**
  - 5D Torus topology + external link
    → 5 x 2 + 1 high speed serial links
  - each 2 GB/s send + 2 GB/s receive
  - DMA, remote put/get, collective operations
- **External (file) IO -- when used as IO chip.**
  - PCIe Gen2 x8 interface (4 GB/s Tx + 4 GB/s Rx)
  - re-uses 2 serial links
  - interface to Ethernet or Infiniband cards

QPX Overview

- Instruction Extensions to PowerISA
- 4-wide double precision FPU SIMD (BG/L,P are 2-wide)
- Also usable as 2-way complex SIMD (BG/L had 1 complex arithmetic)
- Alignment: new module that support multitude of alignments (before only 16, now simultaneous 8,16, 32…)
- Attached to AXU port of A2 core – A2 issues one instruction/cycle to AXU
- 4R/2W register file
  - 32x32 bytes per thread
- 32B (256 bits) datapath to/from L1 cache, 8 concurrent floating point operations (FMA) + load +store

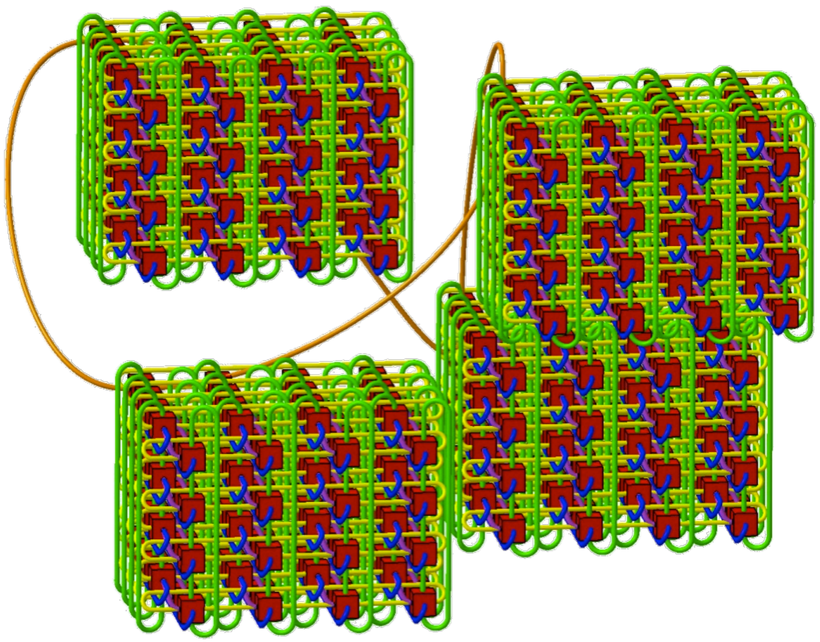


NNSA/SC/IBM Blue Gene /Q

- # of cores: 65,536
- # of nodes: 4096 (4 racks)
- R_{max}: 677 TF
- R_{peak}: 838.9 TF
- N_{max}: 2719743
- Power: 85 kW (network excluded)
- Sustained perf: 80.72%
- GF/W: 1.99

Nov 2011 TOP500 - Rank 17

Inter-Processor Communication



Network Performance

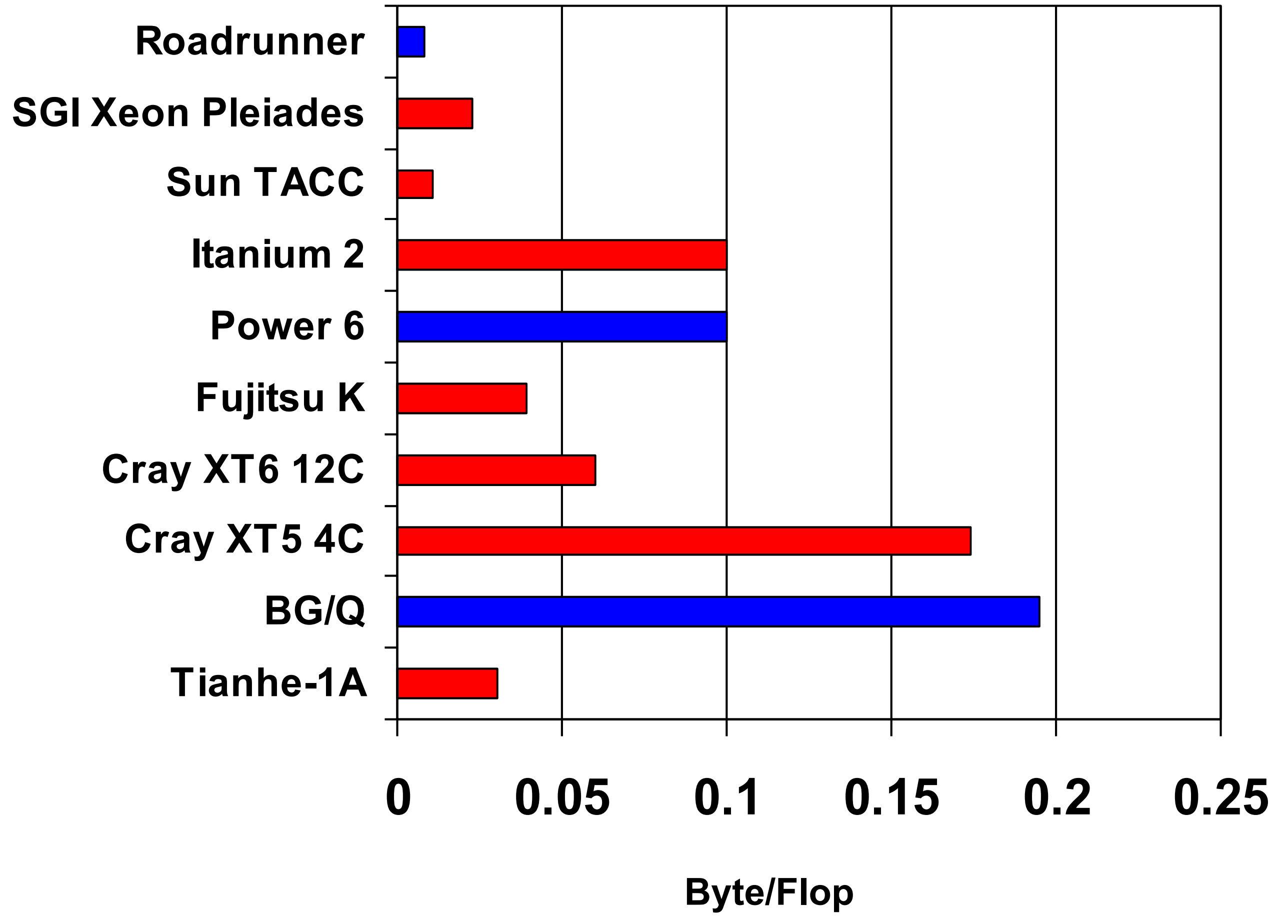
- **All-to-all: 97% of peak**
- **Bisection: > 93% of peak**
- **Nearest-neighbor: 98% of peak**
- **Collective: FP reductions at 94.6% of peak**

- **Integrated 5D torus**
  - Virtual Cut-Through routing
  - Hardware assists for collective & barrier functions
  - FP addition support in network
  - RDMA
    - Integrated on-chip Message Unit
- **2 GB/s raw bandwidth on all 10 links**
  - each direction -- i.e. 4 GB/s bidi
  - 1.8 GB/s user bandwidth
    - protocol overhead
- **5D nearest neighbor exchange measured at 1.76 GB/s per link (98% efficiency)**
- **Hardware latency**
  - Nearest: 80ns
  - Farthest: 3us

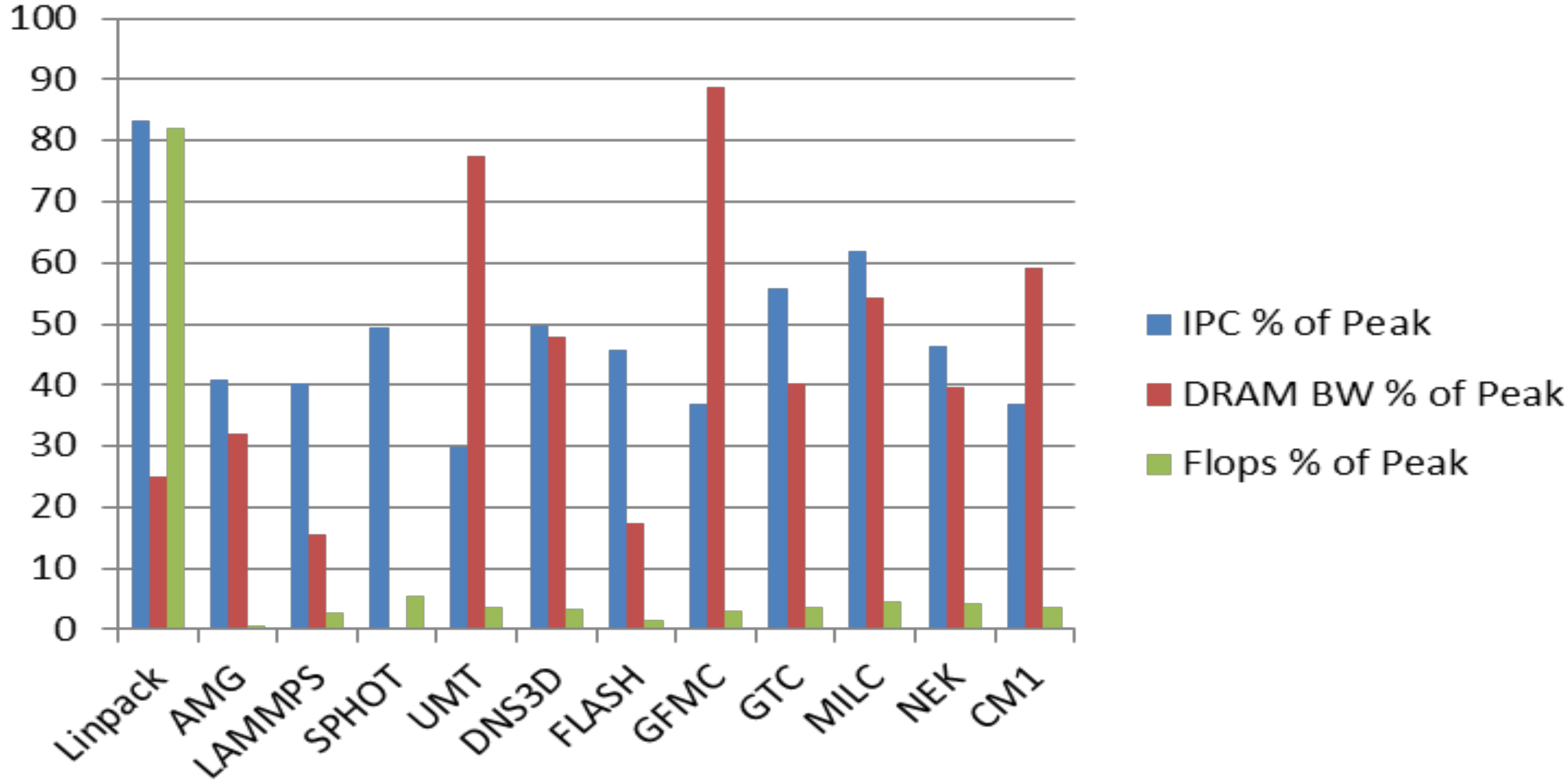
    (96-rack 20PF system, 31 hops)
- **Additional 11th link for communication to IO nodes**
  - BQC chips in separate enclosure
  - IO nodes run Linux, mount file system
  - IO nodes drive PCIe Gen2 x8 (4+4 GB/s)

    ↔ IB/10G Ethernet ↔ file system & world

Inter-Processor Peak Bandwidth per Node

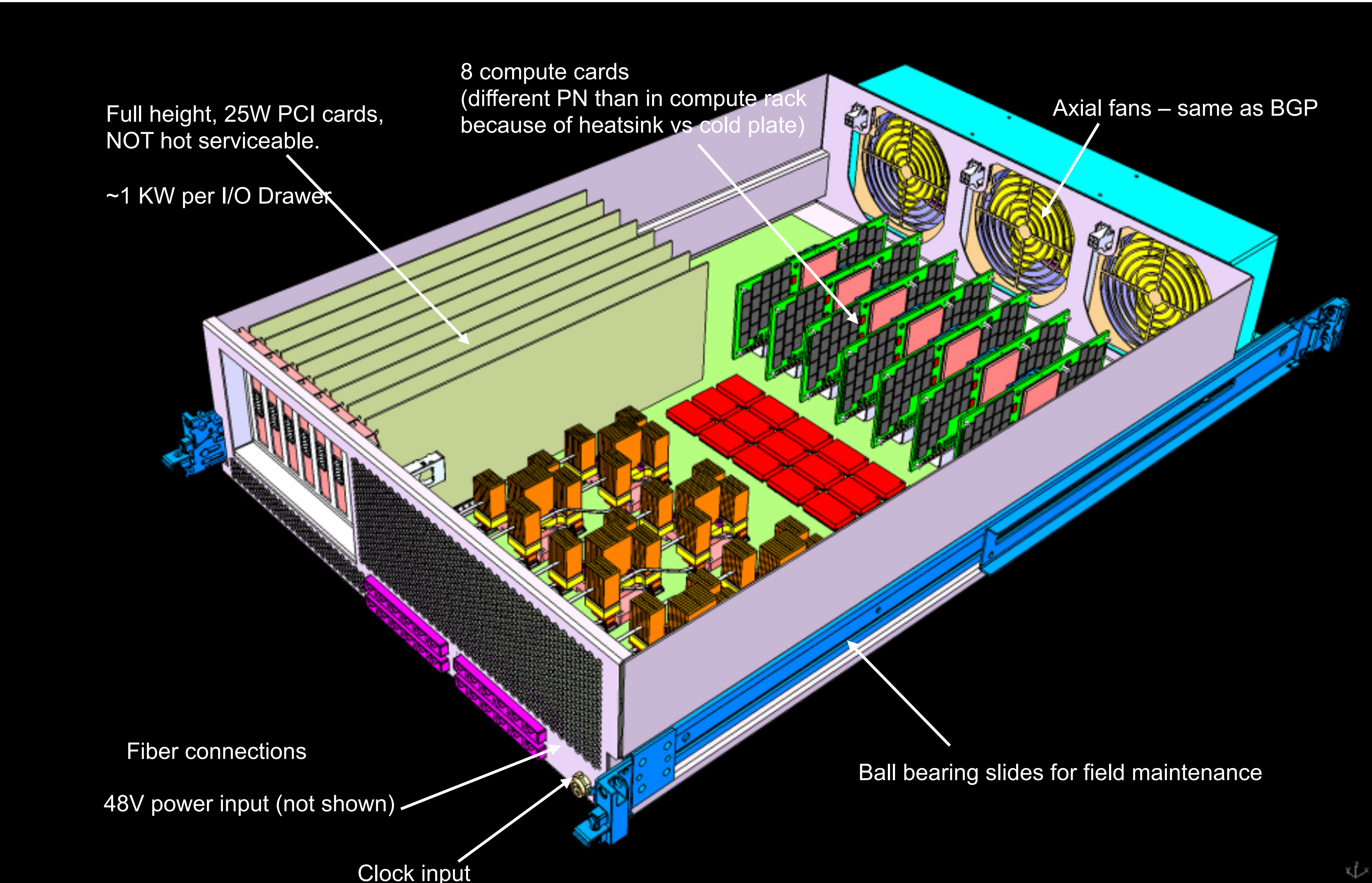


Application Characterization Snapshot from Blue Gene / Q

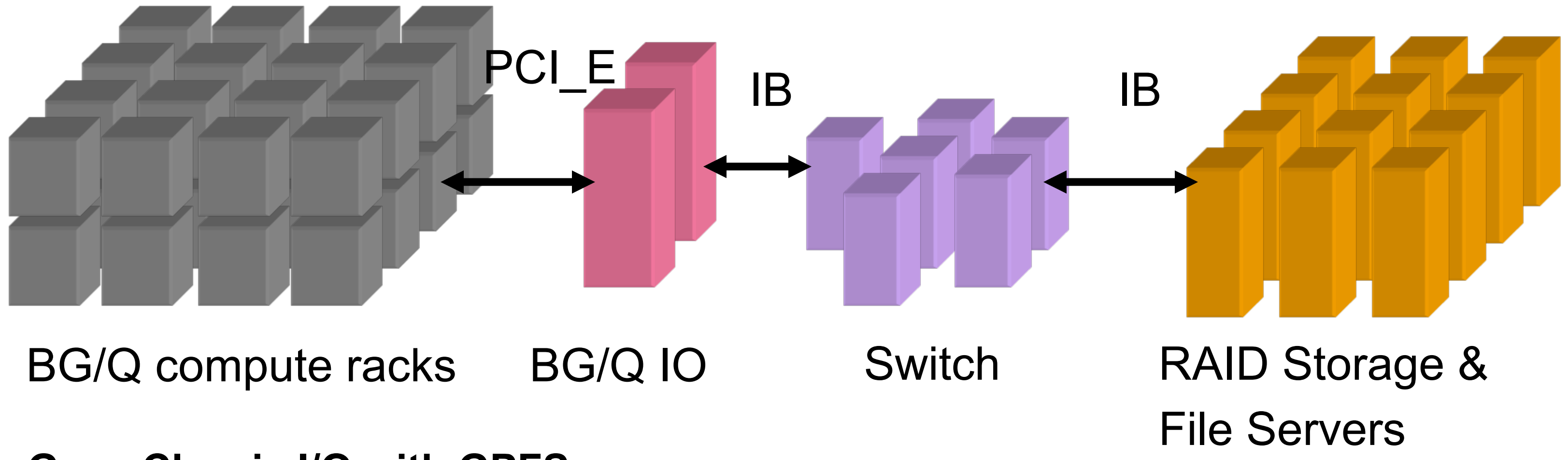


- This data was obtained on a prototype Blue Gene / Q rack.
  - AMG, LAMMPS, SPHOT, UMT are NNSA (Sequoia) benchmarks
  - DNS3D, FLASH, GFMC, GTC, MILC and NEK are Office of Science (ANL) applications.
  - CM1 is a weather / climate app from NCAR
- Even within these three simple metrics, balances are significantly different for different applications.
  - Linpack is a clear outlier
  - Apps except Linpack have low fraction of floating point peak
  - Apps except Linpack have many integer instruction for each floating point operation
  - Main memory bandwidth requirements differ significantly between apps.

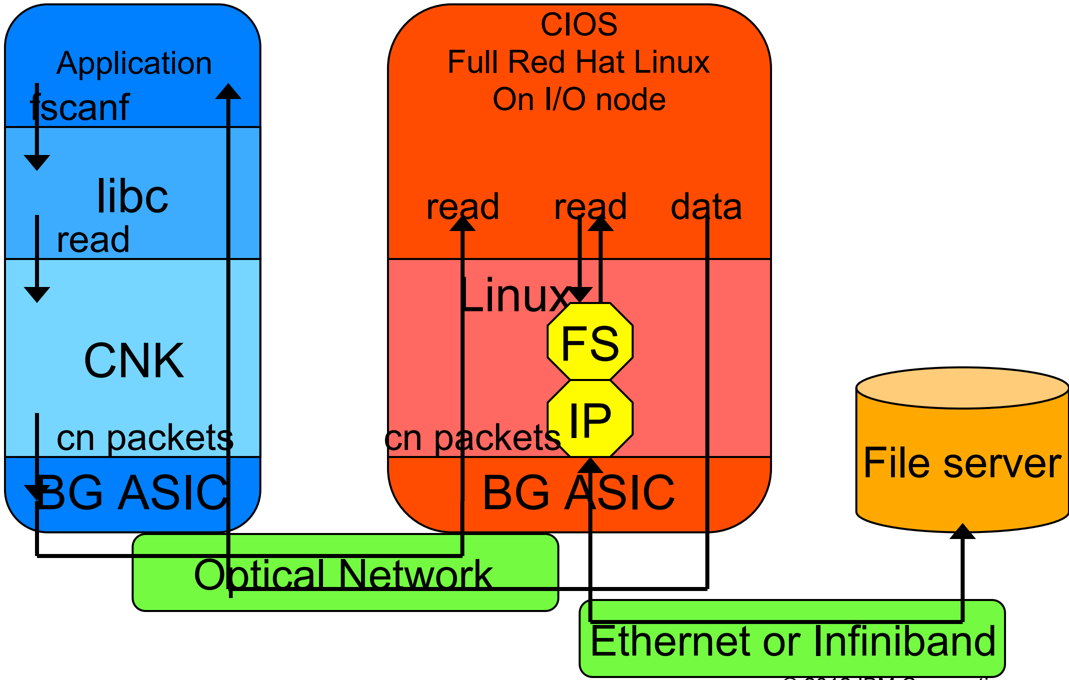
BG/Q I/O Drawer



Classical I/O



- **BlueGene Classic I/O with GPFS clients on the logical I/O nodes**
- **Similar to BG/L and BG/P**
- **Uses InfiniBand switch**
- **Uses DDN RAID controllers and File Servers**
- **BG/Q I/O Nodes are not shared between compute partitions**
  - **IO Nodes are bridge data from function-shipped I/O calls to parallel file system client**
- **Components balanced to allow a specified minimum compute partition size to saturate entire storage array I/O bandwidth**



BG/Q innovations will help programmers cope with an exploding number of hardware threads

- **Exploiting a large number of threads is a challenge for all future architectures. This is a key component of the BGQ research.**
- **Novel hardware and software is utilized in BGQ to**

  a) Reduce the overhead to hand off work to high numbers of threads used in OpenMP and messaging through hardware support for atomic operations and fast wake up of cores.

  b) Multiversioning cache to help in a number of dimensions such as performance, ease of use, and RAS.

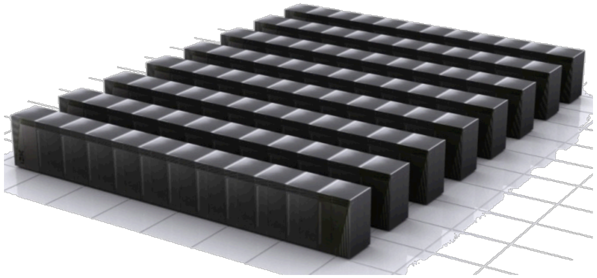
  c) Aggressive FPU to allow for higher single thread performance for some applications. Most will get modest bump (10-25%), some big bump (approaching 300%)

  d) List-Based prefetching for repeated memory reference patterns in arbitrarily long code segments. Also helps achieve higher single thread for some applications.

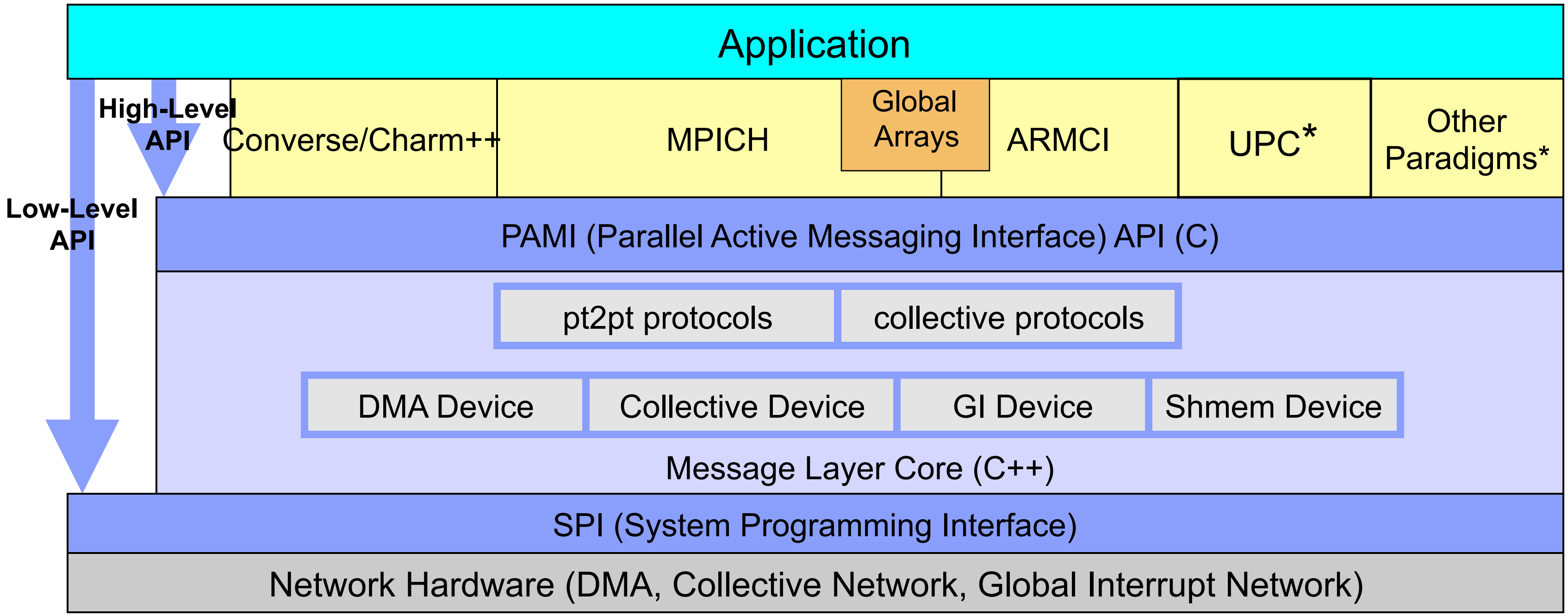
Blue Gene Q Software Innovations

- Standards-based programming environment
  - Linux™ development environment
    - Familiar GNU toolchain with glibc, pthreads, gdb
  - Red Hat on I/O node
  - XL Compilers C, C++, Fortran with OpenMP 3.1
  - Debuggers: Totalview
  - Tools: HPC Toolkit, PAPI, Dyinst, Valgrind, Open Speedshop
- Message Passing
  - Scalable MPICH2 providing MPI 2.2 with extreme message rate
  - Efficient intermediate (PAMI) and low-level (SPI) message libraries, documented, and open source
  - PAMI layer allows easy porting of runtimes like GA/ARMCI, Berkeley UPC, etc,
- Compute Node Kernel (CNK) eliminates OS noise
  - File I/O offloaded to I/O nodes running full Linux
  - GLIBC environment with a few restrictions for scaling
- Flexible and fast job control – with high availability
  - Integrated HPC, HTC, MPMD, and sub-block jobs
  - Noise-free partitioned networks as in previous BG

- **New for Q**
  - Scalability Enhancements: the 17th Core
    - RAS Event handling and interrupt off-load
    - Event CIO Client Interface
    - Event Application Agents: privileged application processing
  - Wide variety of threading choices
  - Efficient support for mixed-mode programs
  - Support for shared memory programming paradigms
  - Scalable atomic instructions
  - Transactional Memory (TM)
  - Speculative Execution (SE)
  - Sub-blocks
  - Integrated HTC, HPC, MPMD, Sub-blocks
  - Integrated persistent memory
  - High availability for service nodes with job continuation
  - I/O nodes running Red Hat



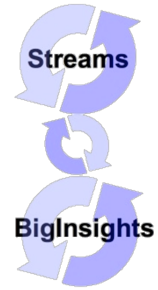
Parallel Active Message Interface



- **Message Layer Core has C++ message classes and other utilities to program the different network devices**
- **Support many programming paradigms**
- **PAMI runtime layer allows uniformity across IBM HPC platforms**

Data Centric Computing

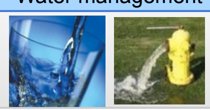
+



Reactive + Deep Analytics Platform

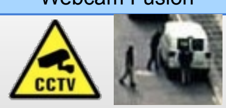
Systems, Services and Solutions Ecosystem

DeepWater
Water management

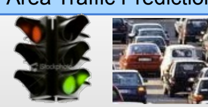


DeepCurrent
Power Delivery

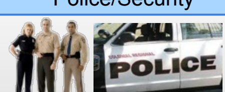
DeepEyes
Webcam Fusion



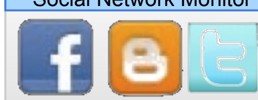
DeepTraffic
Area Traffic Prediction



DeepSafety
Police/Security



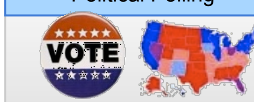
DeepFriends
Social Network Monitor



Let's build a smarter planet

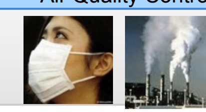
DeepSoil
Farm Prediction

DeepPulse
Political Polling



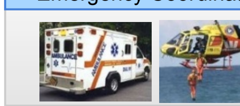
DeepBasket
Food Market Prediction

DeepBreath
Air Quality Control



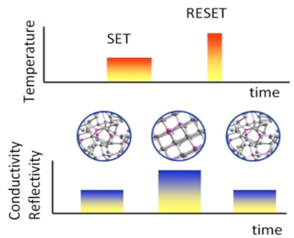
DeepThunder
Local Weather Prediction

DeepResponse
Emergency Coordination



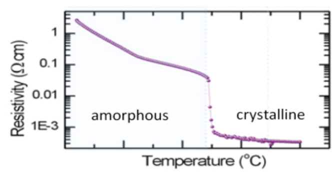
Some Examples of IBM uses of Modeling Simulation and Analytics

Phase Change Materials

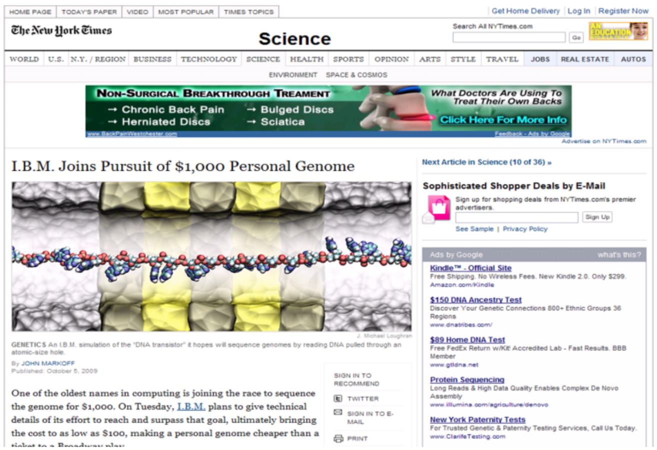


- An alloy contains Ge, Sb and Te
- Normally, PCM has two phases, crystalline and amorphous solid, that are interconverted by heat.

Resistivity changes more than three orders of magnitude between two states.



The DNA Transistor



Desalination Membranes

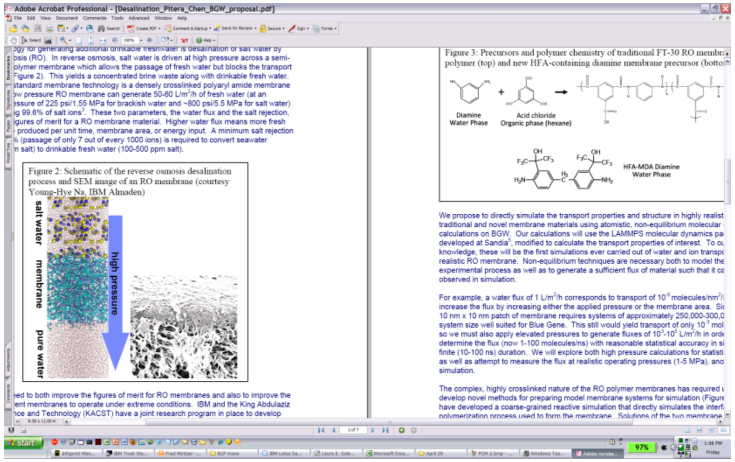
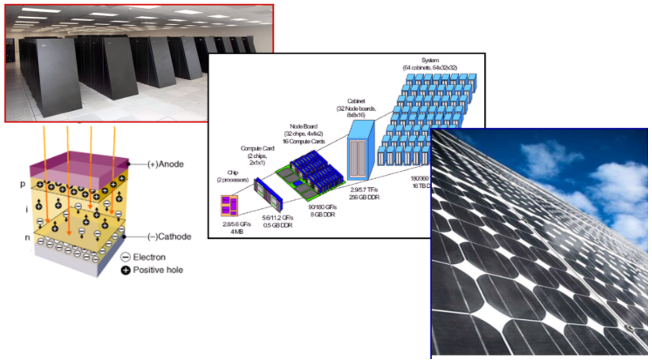


Figure 2: Schematic of the reverse osmosis desalination process and SEM image of an RO membrane (courtesy Young-Hye Na, IBM Almaden)

Figure 3: Precursors and polymer chemistry of traditional FT-30 RO membr polymer (top) and new HFA-containing diamine membrane precursor (botto

Photovoltaic Materials



Fundamental Challenges for 100 PF – 1 EF

- Break out from research in HPC use to large scale production deployment for Research, Industry and Business
  - Essential to develop useable cost-effective solutions for non HPC research users, and for industrial and commercial domains
    - IBM Smart Planet Solutions, Engineering, Finance, Geophysics, Materials, Energy, Climate …

- Workflows will be extremely complex, and will require general purpose systems, software and services solutions
  - at the processor level, system level, data center level
  - Software will have many elements, modeling, simulation, pre/post processing and analysis, uncertainty quantification, sensitivity analysis, visualization, interpretation and presentation

- Data Scales (Exabytes) will require fundamentally new approaches
  - Recognition that data can't move from data center
  - Integrated system solution which spans from Desktop to Exascale solution.

- A full solution will be an extraordinarily complex assembly of systems, software, applications, workflows, data and services.

Workflow Taxonomy

IBM Research

	Description	Examples	Application Set	Team
Capability				
	Calculations not possible on smaller machines Typically a single application Disparate scales define time to solution	Protein Folding Ab-Initio Materials Modeling 1km grid global air circulation	Single Core Application Pre/Post Processing Steps	Small Core Team Team has expert HPC knowledge Team will have significant code knowledge
Complexity				
	Multiple applications cooperating on single workload Coupling between applications	Combined CFD + Structural Cell to Organ Models Environmental Water Management	Multiple Core Applications Complex Linkages Between Apps Data Prep and Analysis	Multiple core teams Mix of HPC, Science, Domain groups Development activities to establish code linkages
Understanding				
	Multiple executions of complex workflow Optimization, Sensitivity Analysis	Integrated Global Climate Structural, CFD, Combustion for Engine Design Aircraft Airflow + Structural	Robust Individual Codes Significant Test and Verification Frameworks Complex Workflows Significant Database Dependencies	Production Quality Codes Primarily non-HPC customers Commercial Grade Service Delivery

Traditional Laboratory Research

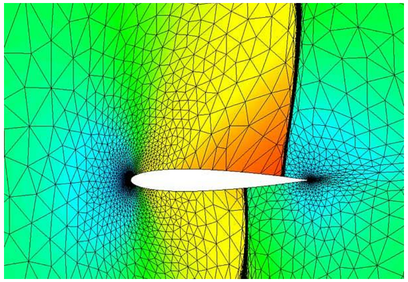
Prototype use only

No commercial impact

Commercial Opportunity requires sophisticated software management, solutions and services

Capability: Computational Needs in Some Domains

IBM Research



CFD Wing Simulation
- 8.3×10^6 Mesh Points
- 5000 flops / mesh point
- 5000 time steps
- **2.5×10^{14} Flops**

CFD Full Plane Simulation
- 3.5×10^{17} Mesh Points
- **8.7×10^{24} Flops**

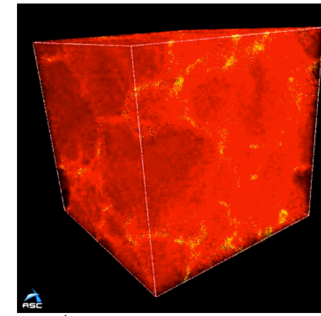
Source:
A. Jameson et al.

Materials Science

Magnetic Materials:
- Current: 2000 Atoms, 2.64 TF/s
- Future: 20000 Atoms, 30 TF/s

Electronic Structure:
- Current 1000 Atoms, 0.5 TF/s
- Future: 10000 Atoms, 100 TF/s

Source: D. Bailey, NERSC



atoms

Digital Movies and Special Effects

- 10^{14} Flops per frame
- 50 Frames / sec
- 90 Min movie
- 2.7×10^{19} Flops

150 Days on 2000 CPUS

Source: Pixar

Spare Parts Inventory Planning

Modeling the optimized deployment of 10,000 part numbers across 100 depots requires
- 2×10^{14} Flops

Industry trend is for rapid frequent modeling for timely business decision support drives higher performance

Source: B. Dietrich, IBM

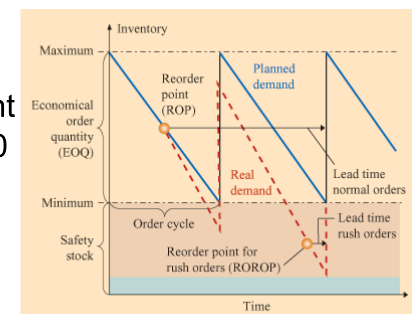
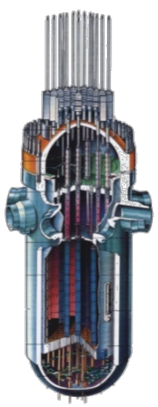


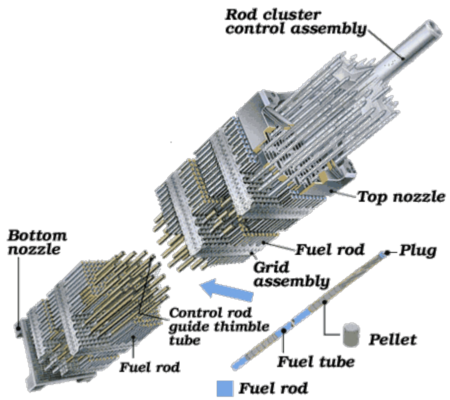
Figure 1

Interdependency among EOQ, safety stock, lead times, and reorder points. Minimum and maximum stock levels are not hard limits, but describe the planning targets without capacity constraints.

Examples: Nuclear Energy, Combustion Applications

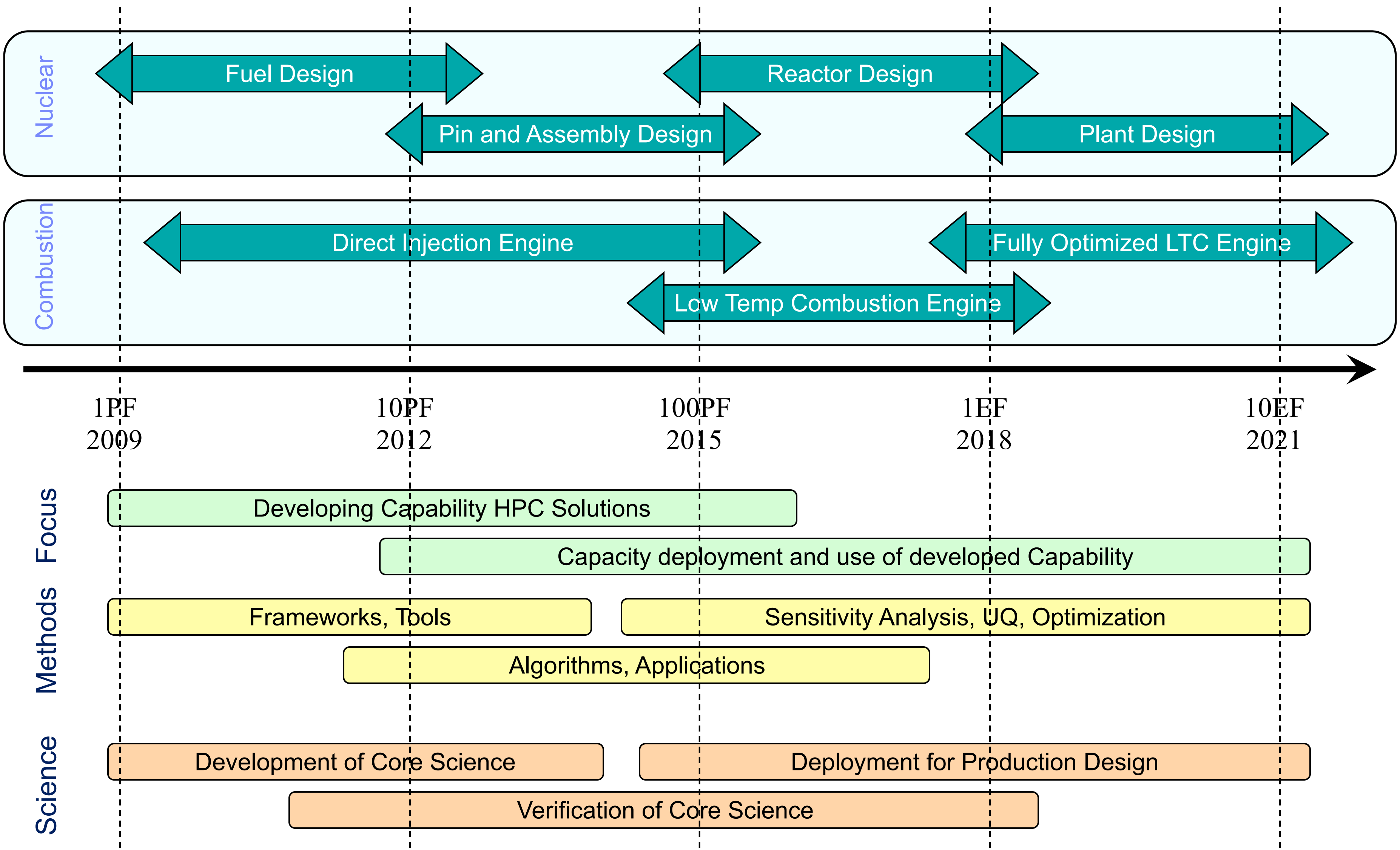
- Nuclear:
  - Next generation reactor design and optimization
  - Develop technologies to improve reliability, safety, increase reactor usable life
  - Develop a sustainable fuel cycle
  - Improve operational management capability
  - Reduce development costs
- Combustion:
  - Gas Turbines, Gasoline and Diesel Engines
  - Increase efficiency,
  - Reduce emissions,
  - Broaden usable fuels
  - Reduce development costs.





- In both cases, multiscale multiphase physics problem.
- Includes Computational Fluid Dynamics, Thermohydralics, Structural Mechanics.
- Coupling of different physical domains and simulation approaches a significant issue.
- Nuclear codes need to include also neutronics, materials aging under neutron bombardment.
- Combustion codes include fuel injection, combustion analyses.

Roadmap for Research and Development in HPC Community



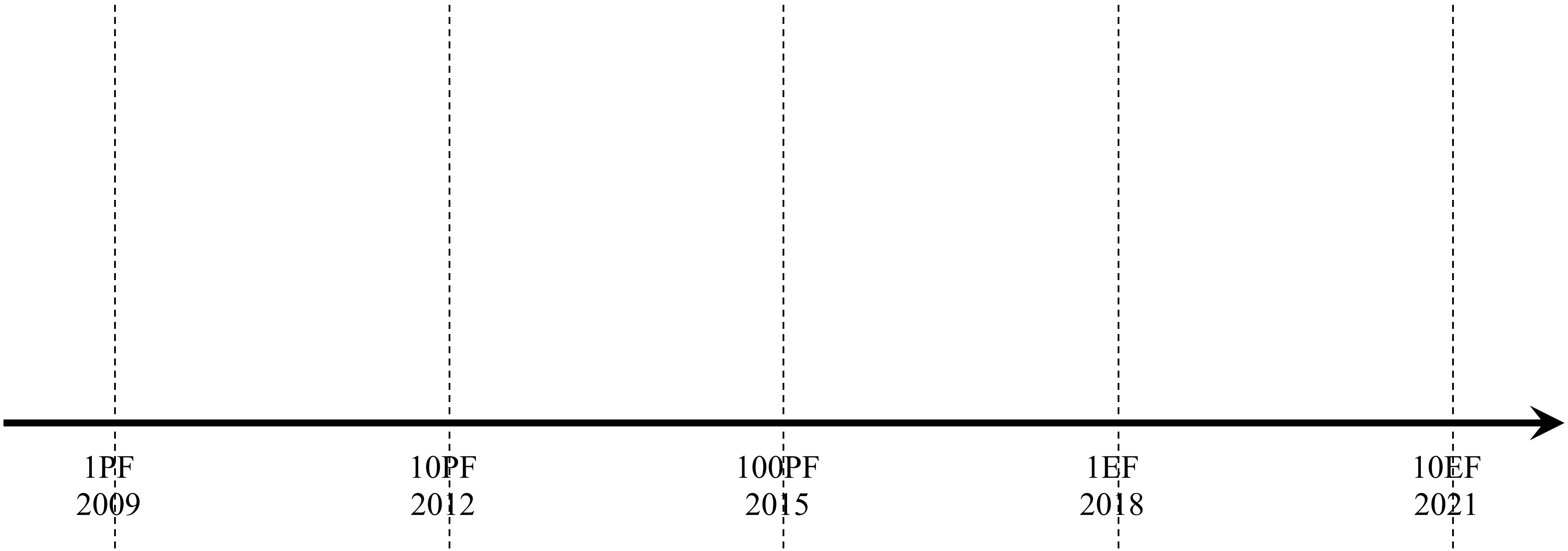
Barriers for Entry for Industrial User (and indeed for any non-HPC research user)

- Computers
  - Currently large companies have small clusters, small companies have nothing
  - Cloud in place, but not HPC Centric
  - Security and Confidentiality a significant concern
- Software
  - Open Source code is often fragile, difficult to use,
  - Usual ISV Code Licensing “per node” or “per core”, actively discourages scaling
  - Customization is complex, expensive
- Expertise
  - Need expertise at all levels to “Play”
  - Systems, Software, Applications, Domain – extremely hard to acquire even for largest organizations
- Enterprise Critical work is the key business opportunity
  - Toy problems, proof of concepts not sufficient.
  - Need security, confidentiality, IP guaranteed
  - Need quality of service and quality of HPC product guarantees to commit.
  - Need qualified cost benefit analysis to convince company executives
- Barriers to Effective Deployment are enormous!

Building the Ecosystem – The Parts

- Participants
  - Industrial / Commercial Users
  - Systems, Software, Services suppliers
  - National Laboratories
  - Academic Researchers
  - Government
- Challenges
  - Collaboration required since no one participant has all the pieces
  - How do both collaborate and compete?
  - Secure systems access suitable for Enterprise Critical applications
  - Expertise available to develop / maintain applications
  - Expertise available to validate / verify computational results
  - Training Services
  - Technology Transfer Services

Roadmap for Industrial / Commercial Deployment?



- Many Activities,
- Many Proof of Concepts
- All the players are engaging

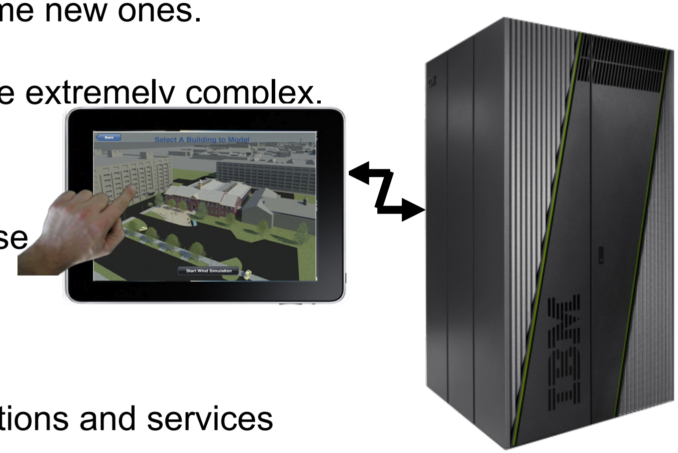
- But
  - Full ecosystem isn't in place, and use remains limited outside HPC research community challenge

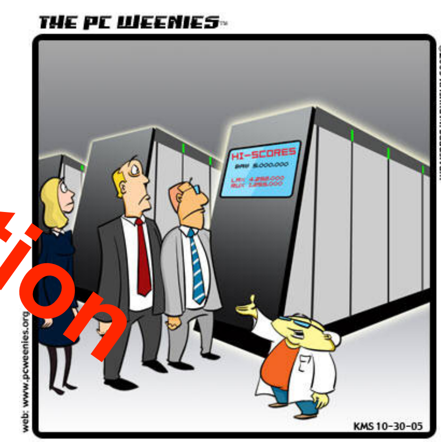
- A Major Challenge for our Community

Summary Remarks

- Blue Gene/Q - more flexible and handle a larger variety of applications
  - Hardware supports several programming models perhaps in even some new ones.
  - Software stack design to help user take advantage of the system.
  - Precursor to Exascale, the Emerging HPC landscape continues to be extremely complex.
- Next 10 years:
  - HPC Capability evolving
    - Fidelity and time to solution relevant for industrial / commercial use
    - Hardware costs continue to fall
  - Focus shifting from Hardware to Services and Solutions
    - Expertise now critical
    - Economic opportunity is development and delivery of robust solutions and services
- We will have succeeded when
  - Stop talking about architecture
  - Focus on real impact: Research, Industry, Business
- Opportunity
  - Focus shifts from single applications to solutions and services
  - Significant opportunities for entry of new players
  - Potential for Ecosystem development to deliver revolutionary Economic impact
- Challenge
  - Ecosystem development to support and enable broad effective adoption.

Thank you for your attention





"OUR SUPERCOMPUTER IS CAPABLE OF PERFORMING TRILLIONS OF COMPUTATIONS PER SECOND, BUT WE PRIMARILY USE IT FOR BIT TORRENT AND GAMING."