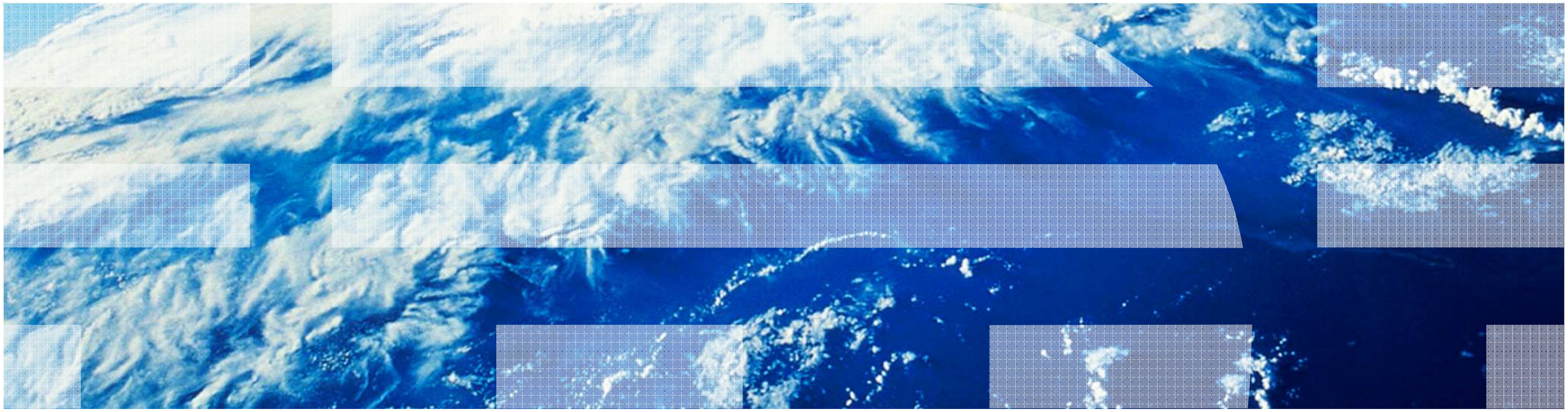
January 25, 2012

Robert W. Wisniewski
Chief Software Architect
Blue Gene Supercomputer Research

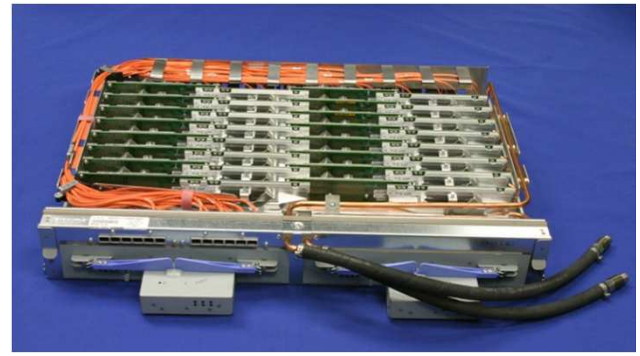
On behalf the Blue Gene Team

# BlueGene/Q: Architecture, CoDesign; Path to Exascale

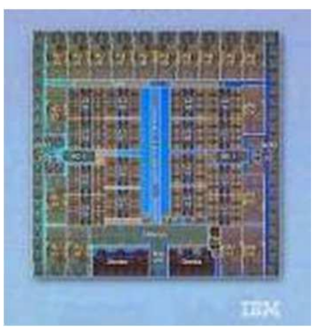



# Blue Gene/Q

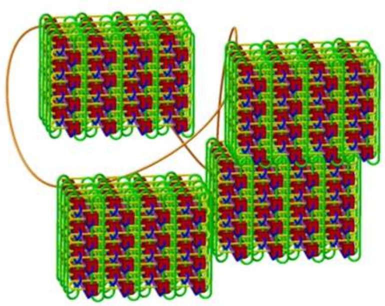
Industrial Design



32 Node Board



BQC DD2.0



5D torus

4-rack system

# Top 10 reasons that you need Blue Gene/Q

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**

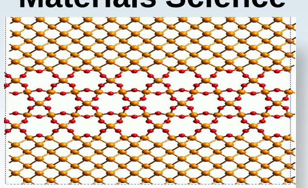
# Examples of Applications Running on Blue Gene

## Developed on L, P; many ported to Q

| Application | Owner | Application | Owner | Application | Owner |
|---|---|---|---|---|---|
| CFD Alya System | Barcelona SC | DFT iGryd | Jülich | BM: SPEC2006, SPEC openmp | SPEC |
| CFD (Flame) AVBP | CERFACS Consortium | DFT KKRnano | Jülich | BM: NAS Parallel Benchmarks | NASA |
| CFD dns3D | Argonne National Lab | DFT ls3df | Argonne National Lab | BM: RZG (AIMS,Gadget,GENE, GROMACS,NEMORB,Octopus, Vertex) | RZG |
| CFD OpenFOAM | SGI | DFT PARATEC | NERSC / LBL | Coulomb Solver - PEPC | Jülich |
| CFD NEK5000, NEKTAR | Argonne, Brown U | DFT CPMD | IBM/Max Planck | MPI PALLAS | UCB |
| CFD OVERFLOW | NASA, Boeing | DFT QBOX | LLNL | Mesh AMR | CCSE, LBL |
| CFD Saturne | EDF | DFT VASP | U Vienna & Duisburg | PETSC | Argonne National Lab |
| CFD LBM | Erlanger-Nuremberg | Q Chem GAMESS | Ames Lab/Iowa State | MpiBlast-pio Biology | VaTech / ANL |
| MD Amber | UCSF | Nuclear Physics GFMC | Argonne National Lab | RTM – Seismic Imaging | ENI |
| MD Dalton | Univ Oslo/Argonne | Neutronics SWEEP3D | LANL | Supernova Ia FLASH | Argonne National Lab |
| MD ddcMD | LLNL | QCD CPS | Columbia U/IBM | Ocean HYCOM | NOPP / Consortium |
| MD LAMMPS | Sandia National Labs | QCD MILC | Indiana University | Ocean POP | LANL/ANL/NCAR |
| MD MP2C | Jülich | Plasma GTC | PPPL | Weather/Climate CAM | NCAR |
| MD NAMD | UIUC/NCSA | Plasma GYRO (Tokamak) | General Atomics | Weather/Climate Held-Suarez Test | GFDL |
| MD Rosetta | U Washington | KAUST Stencil Code Gen | KAUST | Climate HOMME | NCAR |
| DFT GPAW | Argonne National Lab | BM:sppm,raptor,AMG,IRS,sphot | Livermore | Weather/Climate WRF, CM1 | NCAR, NCSA |

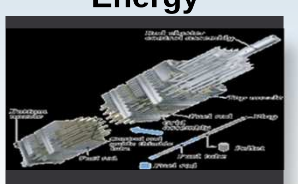
## *Accelerating Discovery and Innovation in:*

Materials Science



Silicon Design

Energy

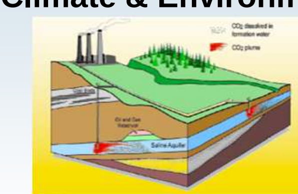


Next Gen Nuclear

Engineering

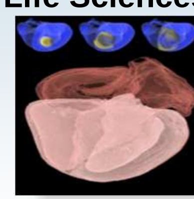
High Efficiency Engines

Climate & Environment



Oil Exploration

Life Sciences

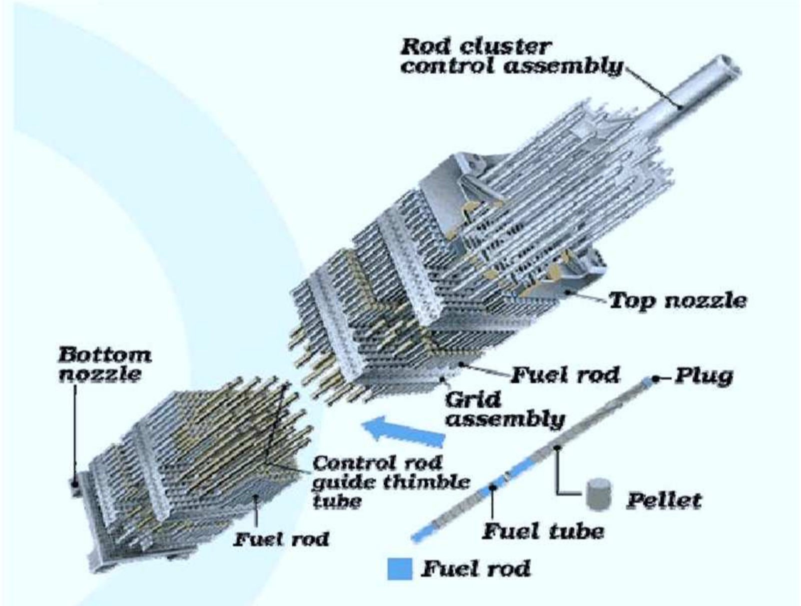


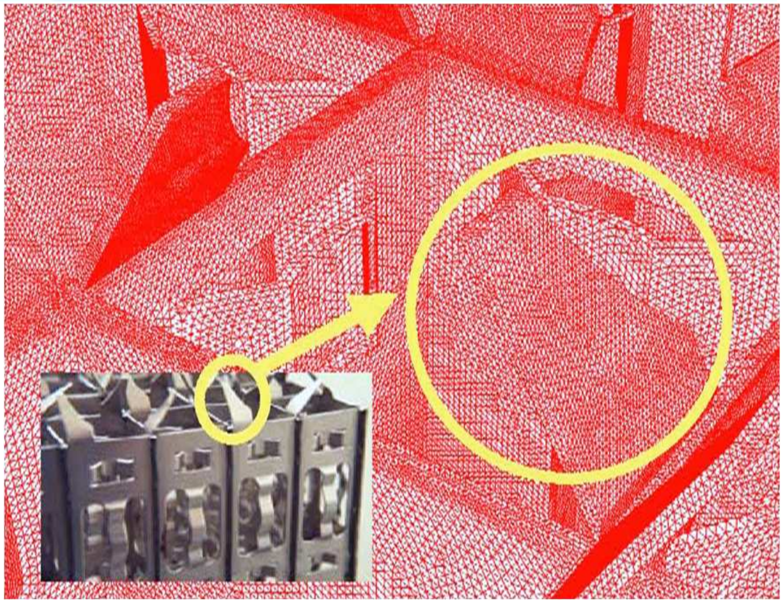
Whole Organ Simulation

ɔn

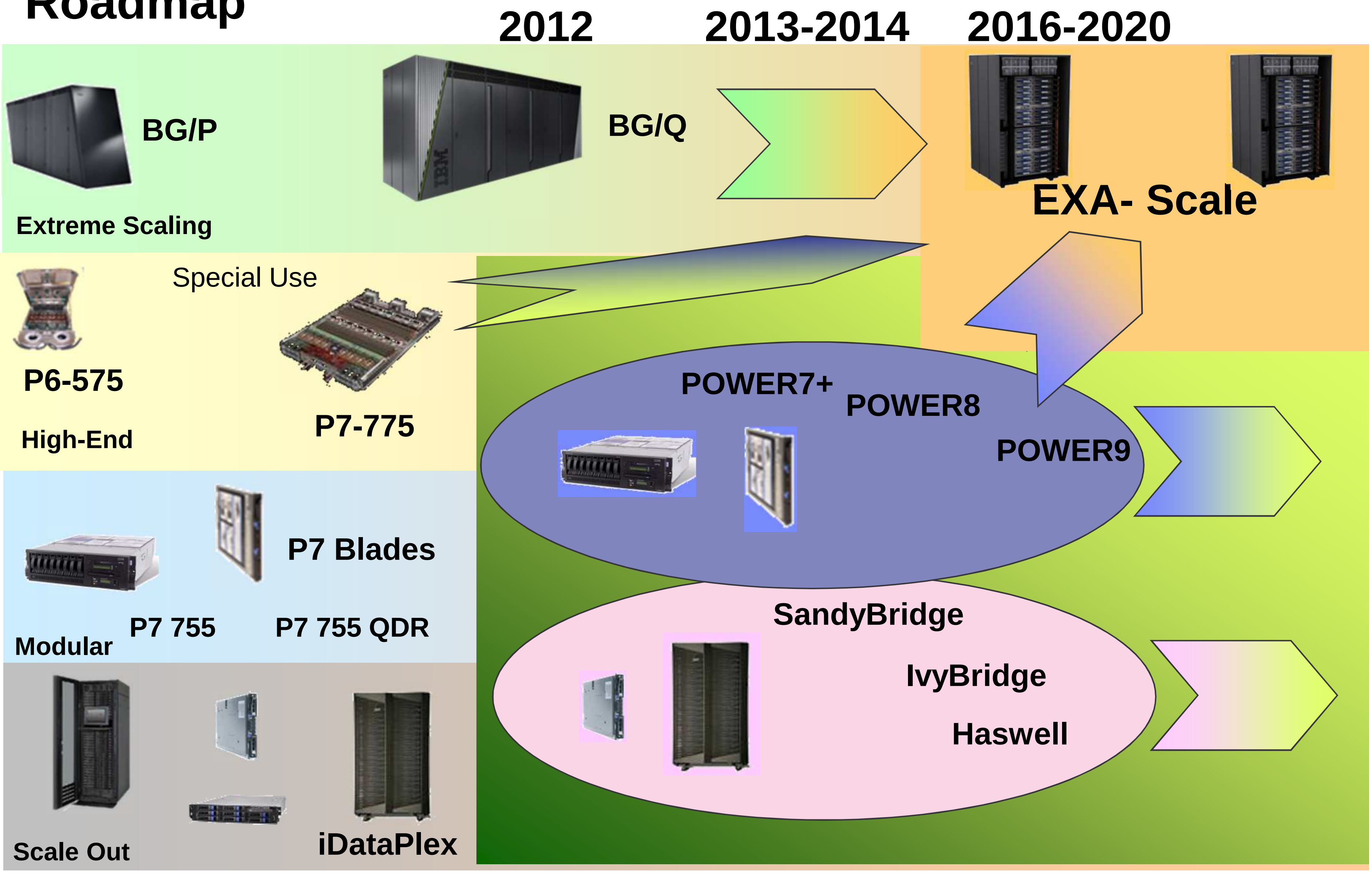
# Blue Gene/Q Expanded Apps Reach

- **Ease of Programming**
  - –More memory/node
  - –Enhanced I/O
  - –Ease of porting
- **BROADER Application Front**
  - –Graph 500
  - –Life Sciences
  - –Uncertainty Quantification
- **Increasing capability- Example**
  - –L: a few fuel rods (5x5)
  - –P: fuel assembly (17x17)
  - –Q: nuclear reactor (~200 assemblies)



# Roadmap

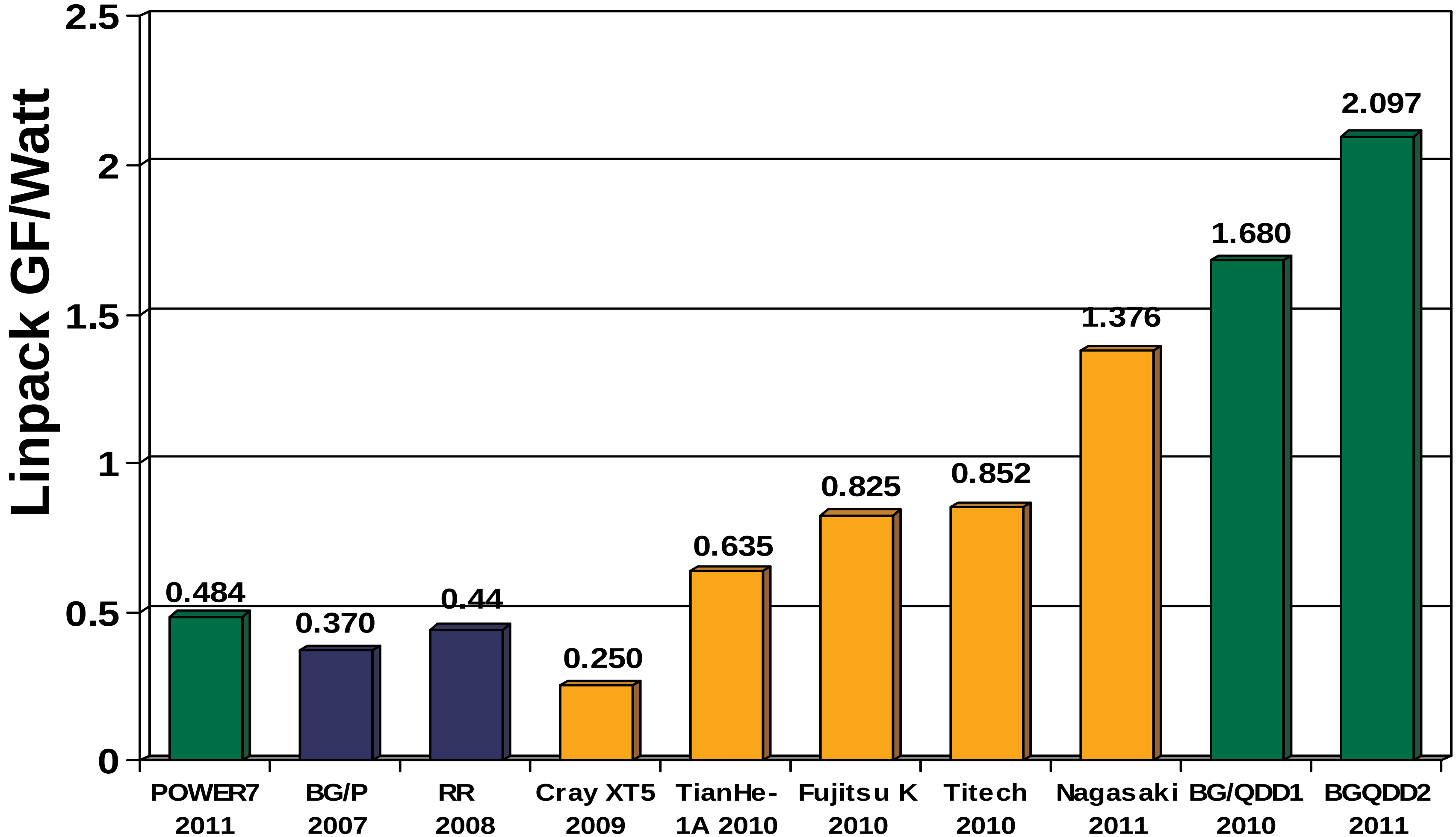

2012
2013-2014
2016-2020
BG/P
BG/Q
EXA- Scale
Extreme Scaling
Special Use
P6-575
P7-775
High-End
POWER7+
POWER8
POWER9
P7 Blades
Modular
P7 755
P7 755 QDR
SandyBridge
IvyBridge
Haswell
Scale Out
iDataPlex

# *October 7, 2009: President Obama presented the 2008 National Medal of Technology and Innovation to IBM, the only company so honored, for the Blue Gene family of supercomputers…*


*The US Government and IBM represent world leadership in high performance computing.*

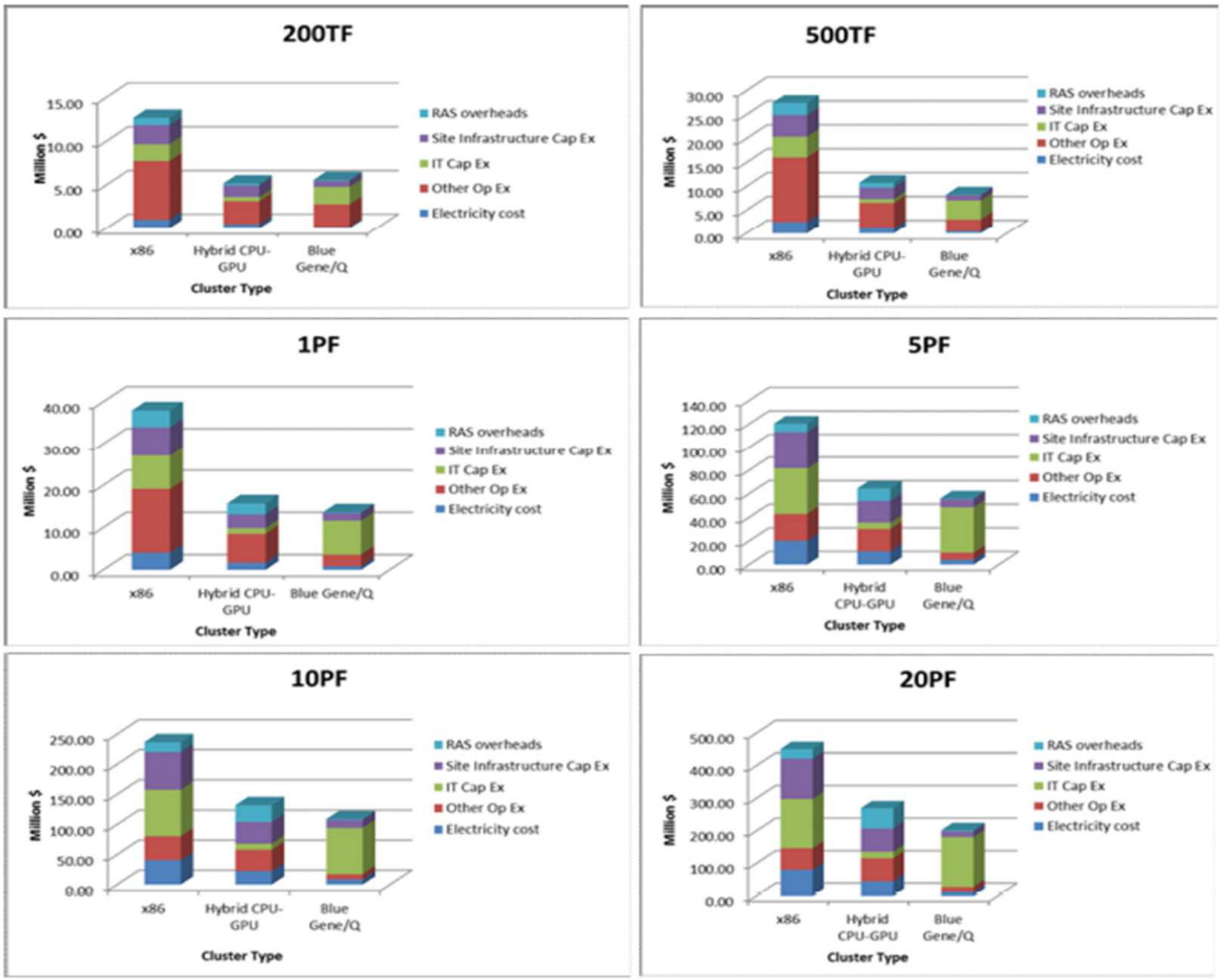
# System Power Efficiency (Green500 06/2011)

**At $.10/kWh => 1MW savings in power saves $1M/year. TCO saving is much more.**
**Low power is key to scaling to large systems**



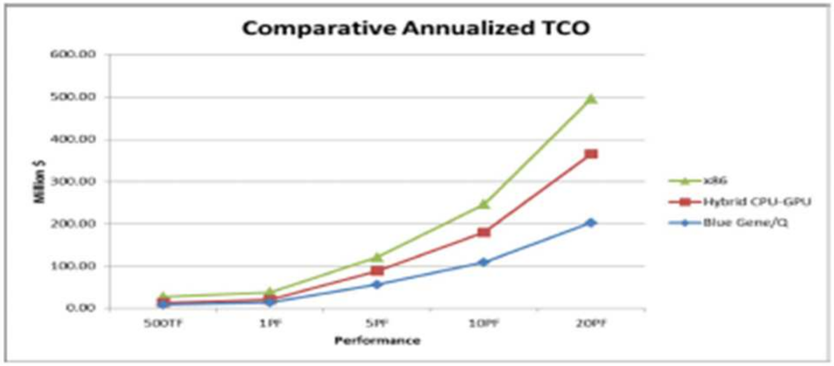

*Source: www.green500.org*

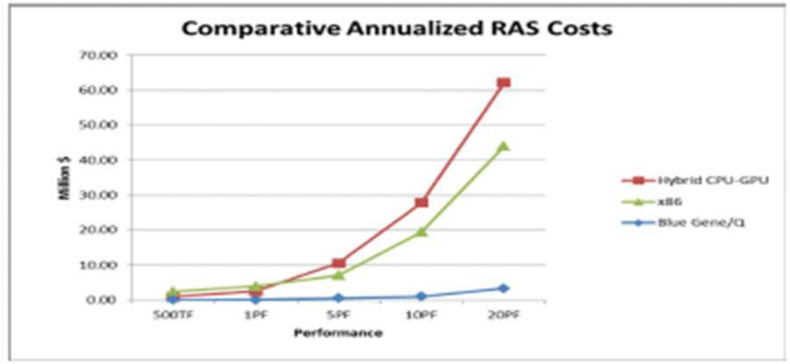
# Annualized TCO of HPC Systems (Cabot Partners)

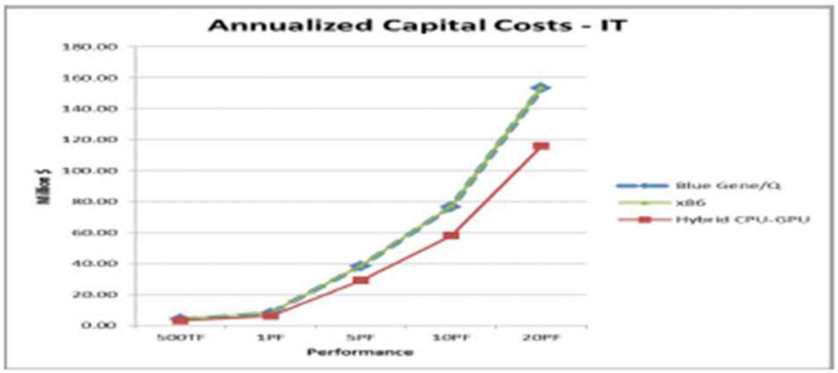


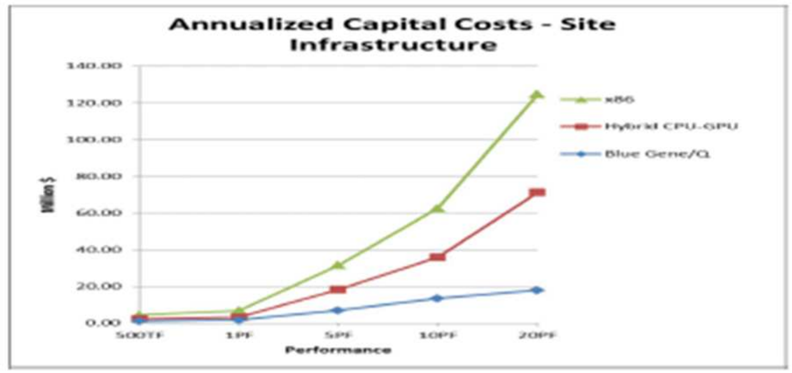

BG/Q saves ~$300M/yr!

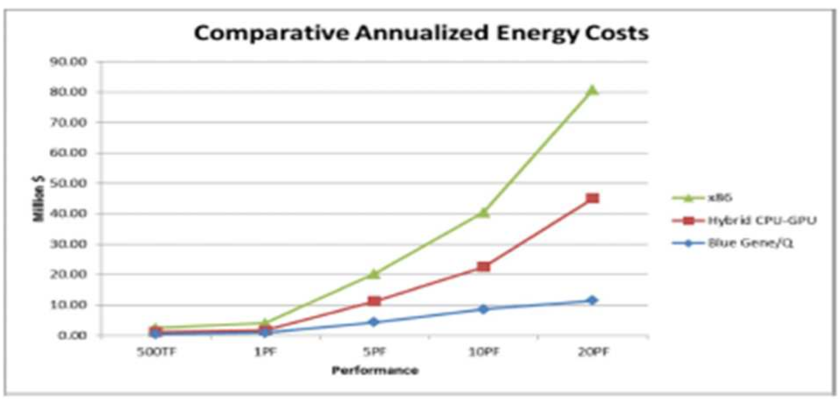
# Annualized TCO & Component Costs vs Peak Performance

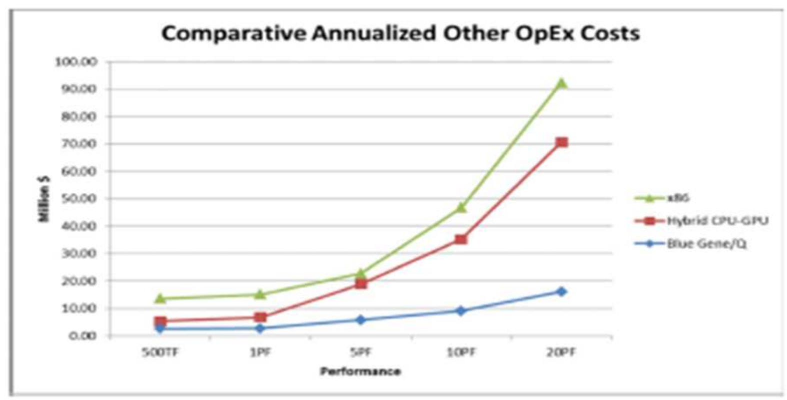


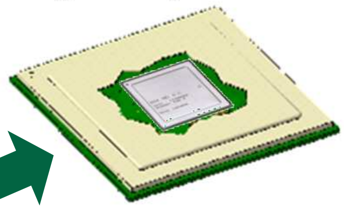
# Blue Gene Evolution

- **BG/L (5.7 TF/rack, 210 MF/W) – 130nm ASIC (2004 GA)**
  - Scales >128 racks, 0.734 PF/s, dual-core system-on-chip,
  - 0.5/1 GB / Node
- **BG/P (13.9 TF/rack, 357 MF/W) – 90nm ASIC (2007 GA)**
  - Scales >256 racks, 3.5 PF/s, quad core SOC, DMA
  - 2/4 GB / Node
  - SMP support, OpenMP, MPI
- **BG/Q (209 TF/rack, 2000 MF/W) – 45nm ASIC (Early 2012 GA)**
  - 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

# Blue Gene/Q

1. Chip:
16+2 $\mu$P
cores

2. Single Chip Module



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

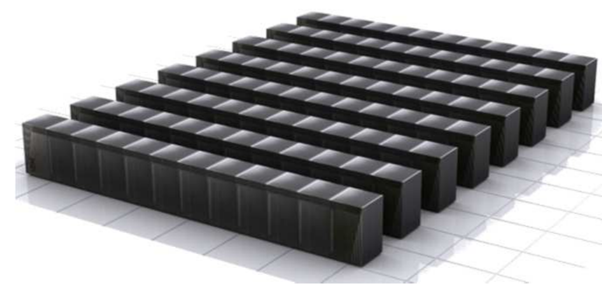
4. Node Card:
32 Compute Cards,
Optical Modules, Link Chips; 5D Torus

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

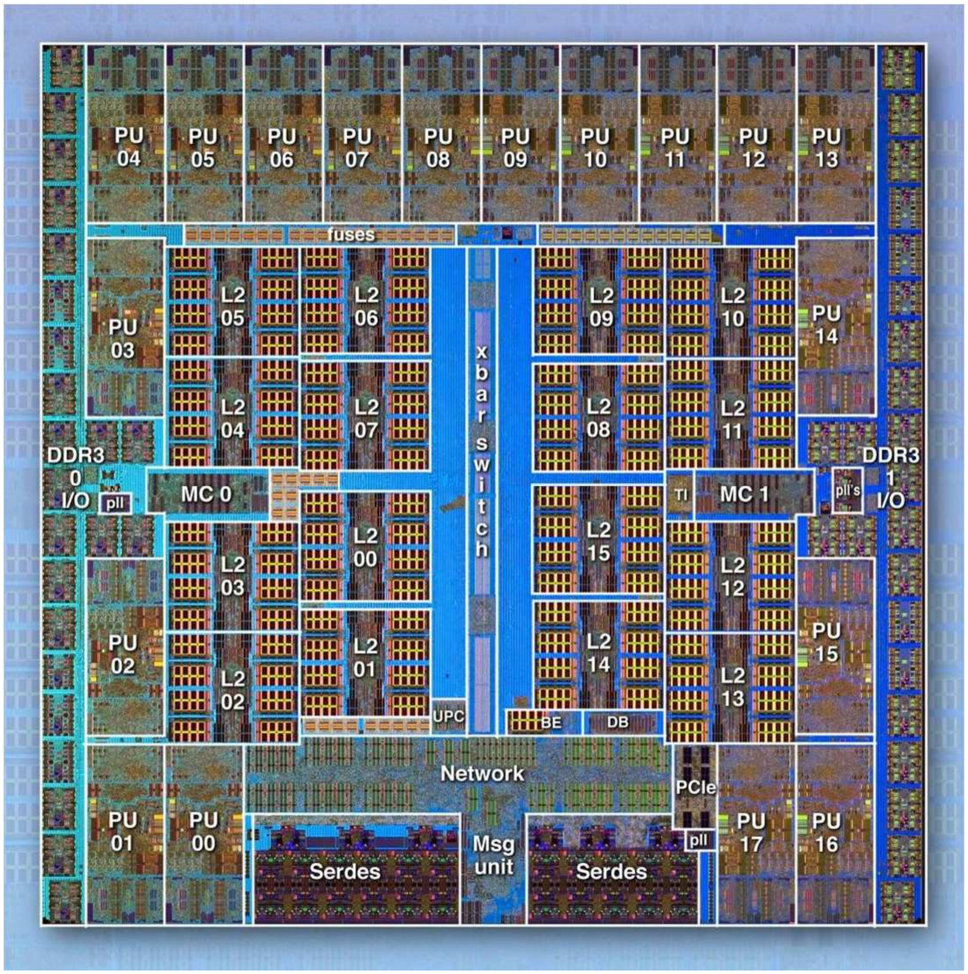
7. System:
96 racks, 20PF/s



- **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**

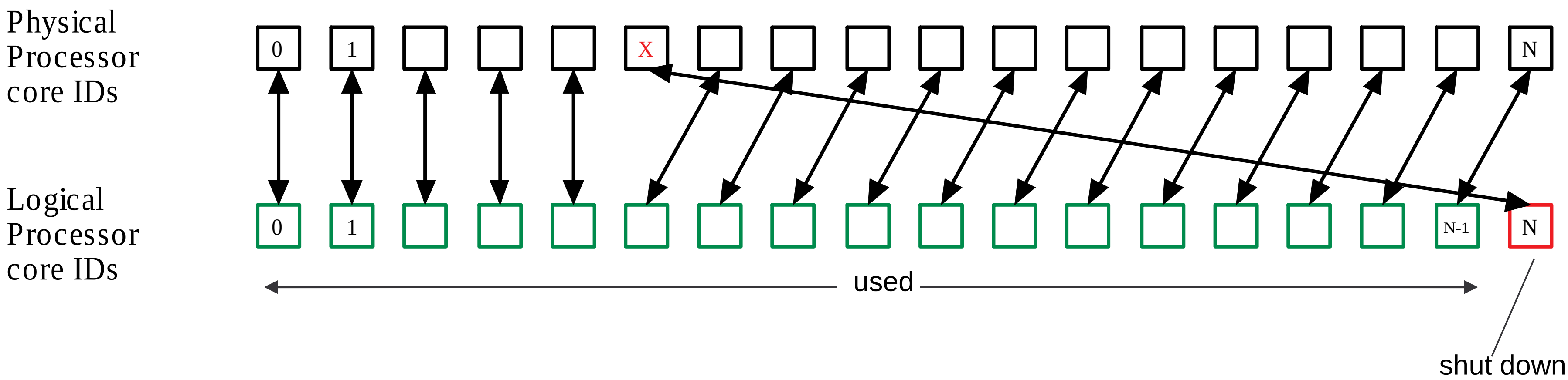
# BlueGene/Q Compute chip

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



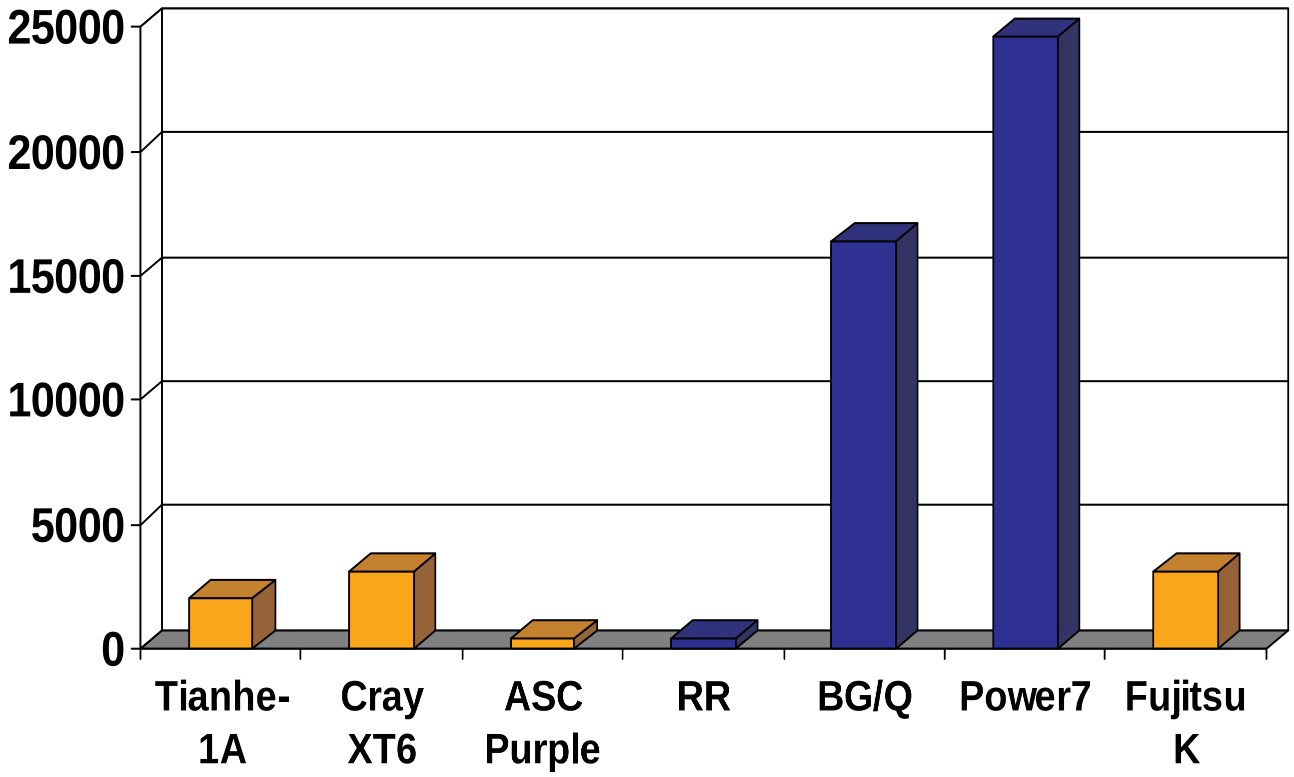

- 360 mm² Cu-45 technology (SOI)
- 16 user + 1 service PPC processors
  - plus 1 redundant processor
  - all processors are symmetric
  - 11 metal layer
  - each 4-way multi-threaded
  - 64 bits
  - 1.6 GHz
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
  - each processor has Quad FPU (4-wide double 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

# Physical-to-Logical mapping of PUnits in presence of a fail

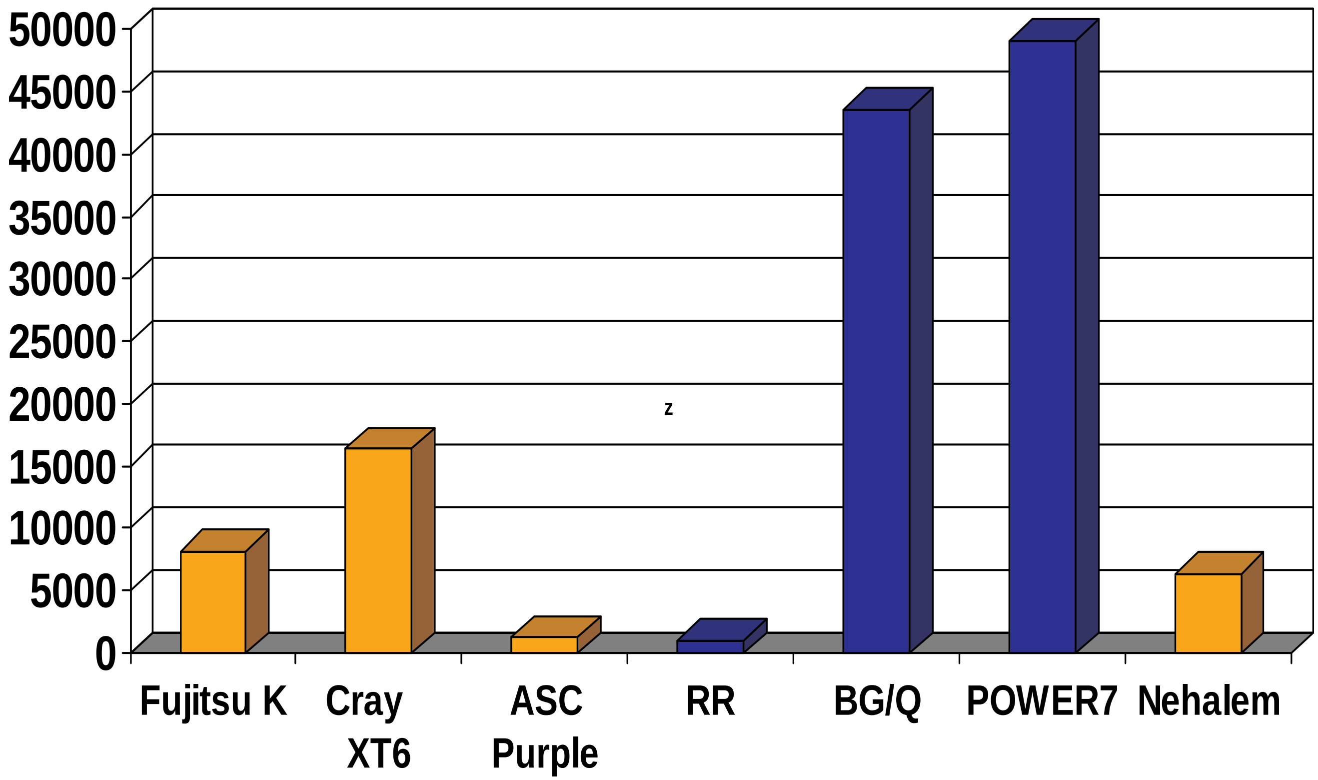



- Inspired by array redundancy
- PUnit N+1 redundancy scheme substantially increases yield of large chip
- Redundancy can be invoked at any manufacturing test stage
  - wafer, module, card, system
- Redundancy info travels with physical part -- stored on chip (eFuse) / on card (EEPROM)
  - at power-on, info transmitted to PUnits, memory system, etc.
- Single part number flow
- Transparent to user software: user sees N consecutive good processor cores.

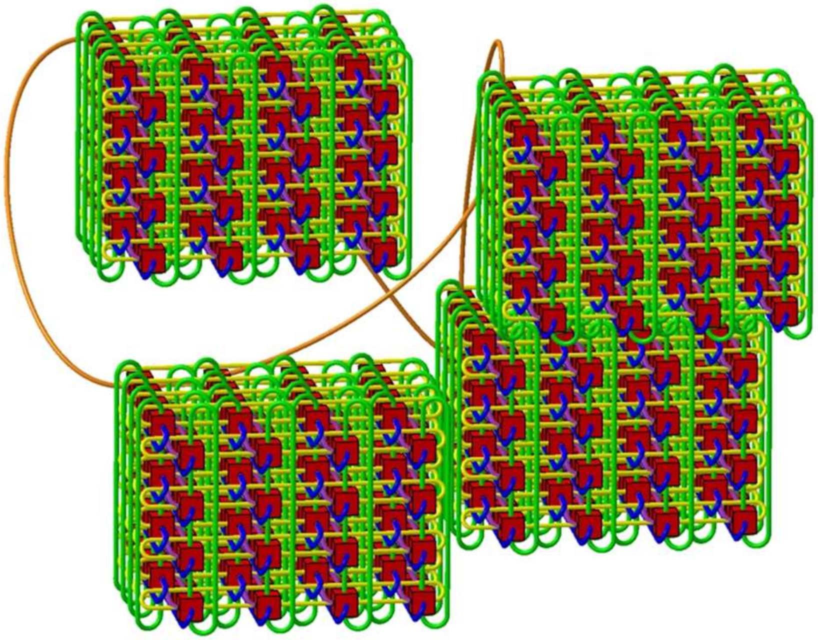
# Main Memory Capacity per Rack



# Main Memory Bandwidth per Rack



# Inter-Processor Communication

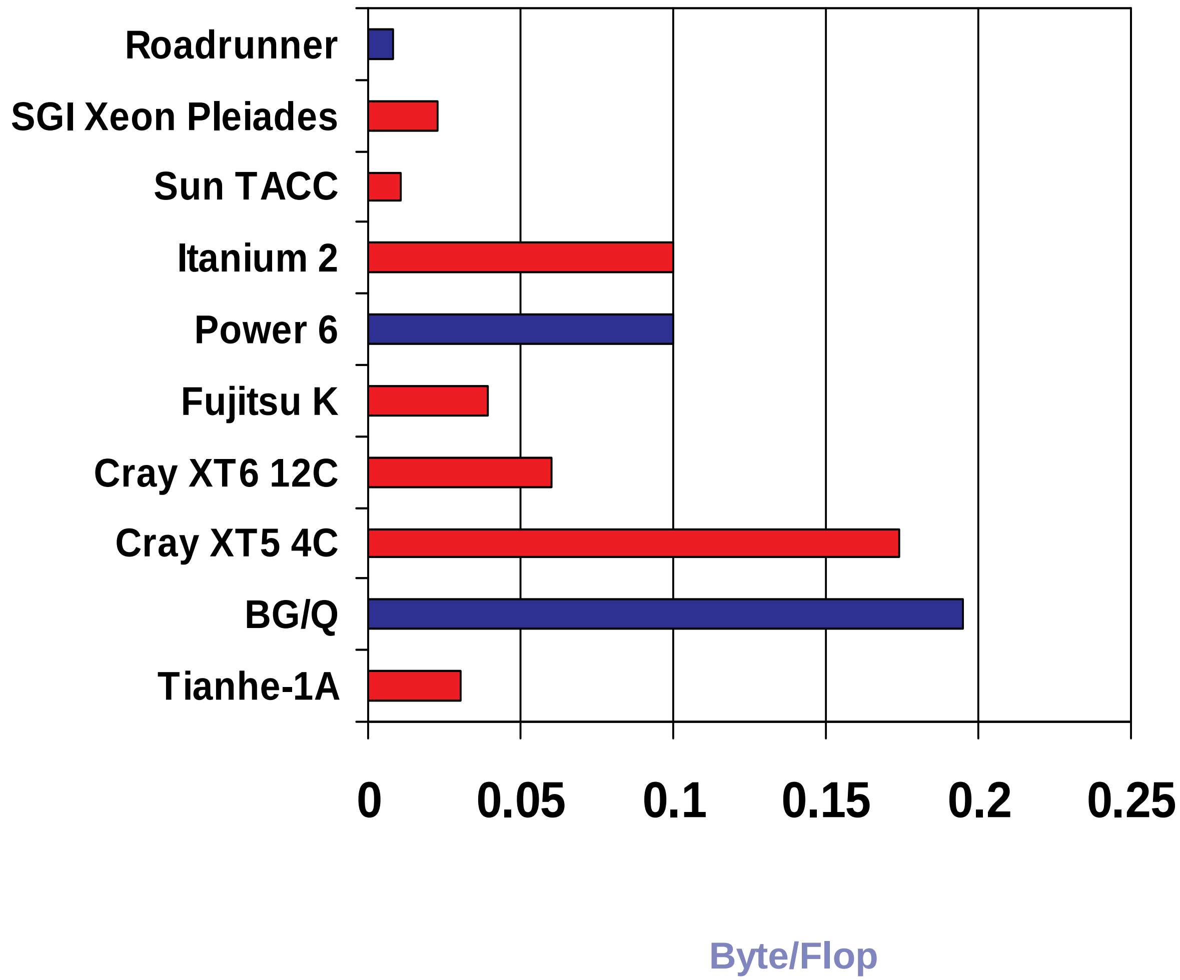


## Network Performance

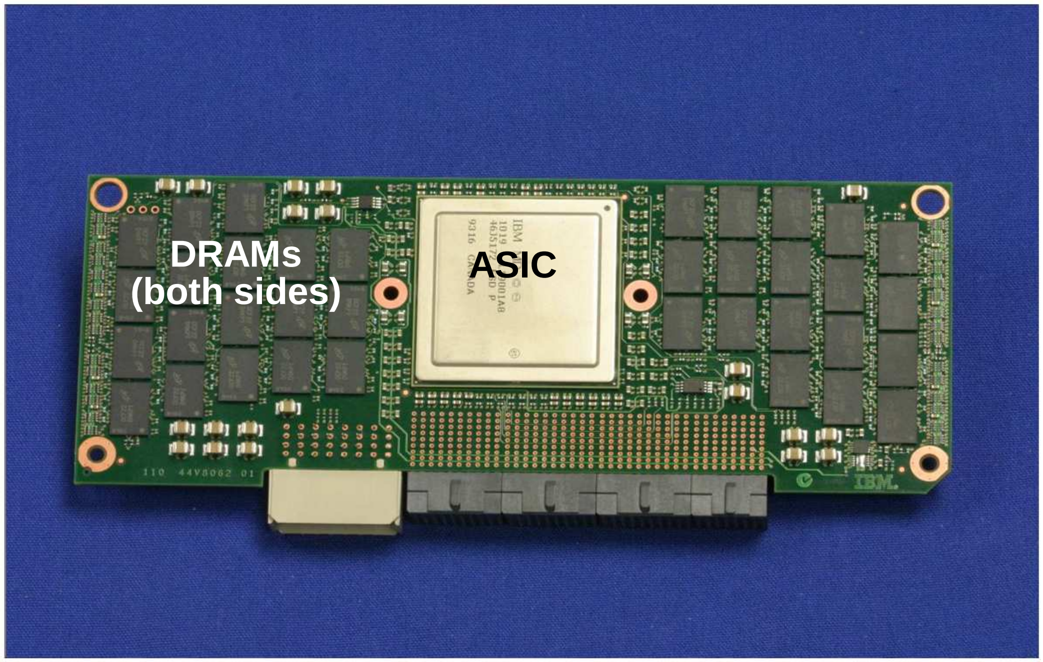
- **All-to-all: 97% of peak**
- **Bisection: > 93% of peak**
- **Nearest-neighbor: 98% of peak**
- **Collective: FP reductions at 94.6% of peak**

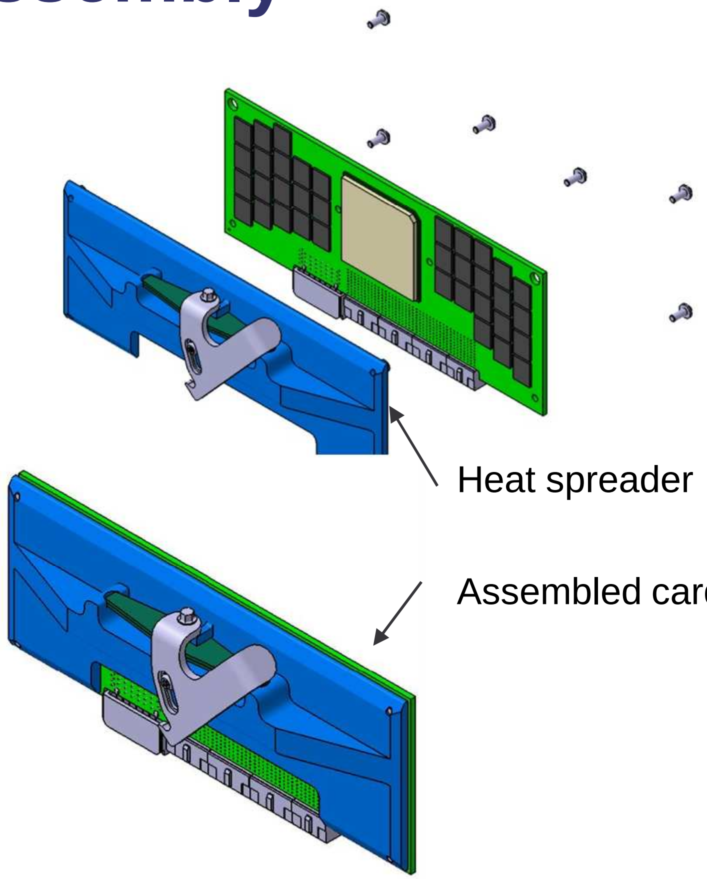
- **Integrated 5D torus**
  - Virtual Cut-Through routing
  - Hardware assists for collective & barrier functions
  - FP addition support in network
  - RDMA
    - Integrated on-chip Message Unit
- **2 GB/s raw bandwidth on all 10 links**
  - each direction -- i.e. 4 GB/s bidi
  - 1.8 GB/s user bandwidth
    - 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 11$^{th}$ 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



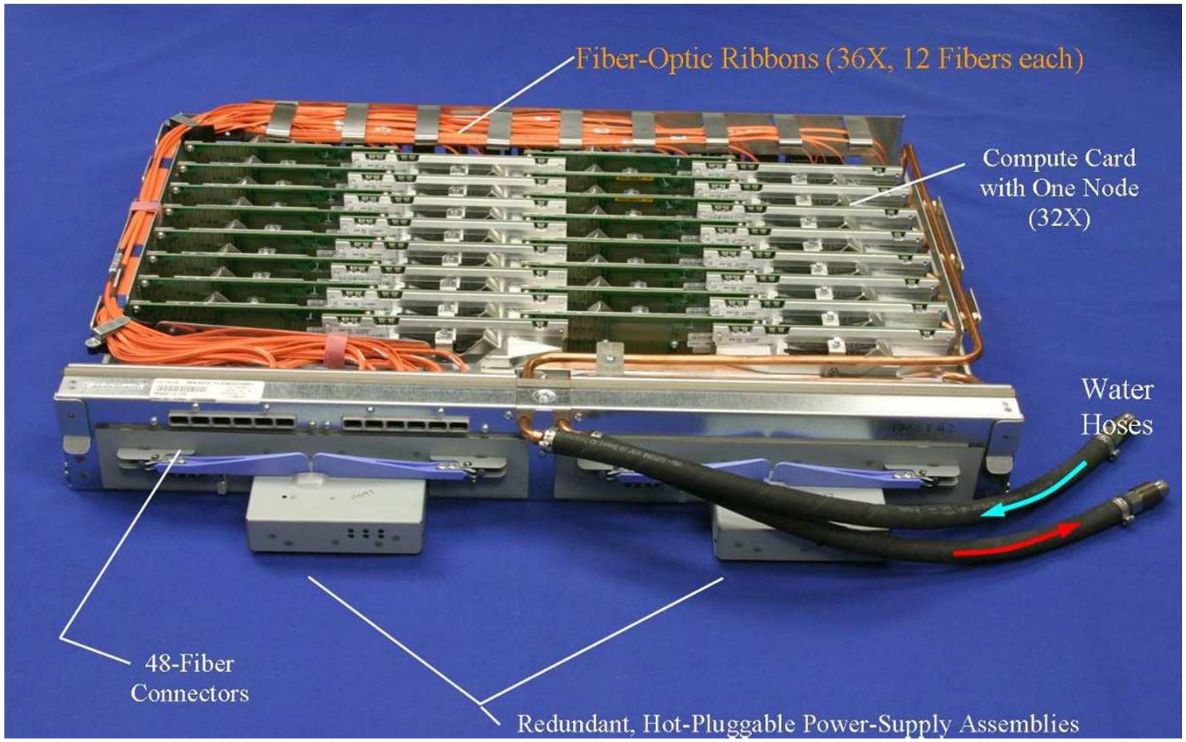
# Blue Gene/Q Compute Card Assembly



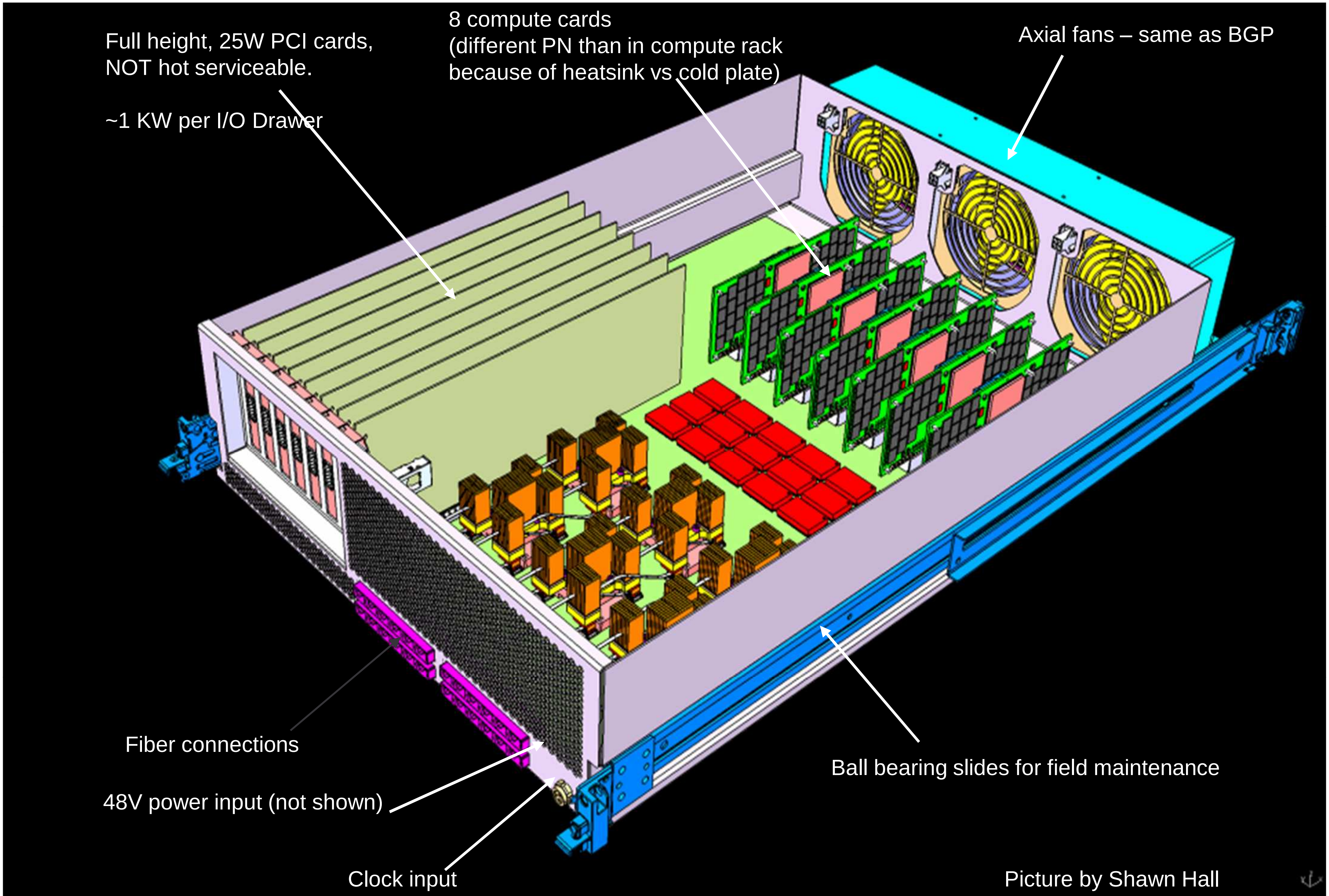




- Basic field replaceable unit of a Blue Gene/Q system
- Compute Card has 1 BQC chip + 72 SDRAMs (16GB DDR3)
- Two heat sink options: Water-cooled → "Compute Node" / air-cooled → "IO Node"
- Connectors carry power supplies, JTAG etc, and 176 Torus signals (4 and 5 Gbps)

# Blue Gene/Q Node Card Assembly

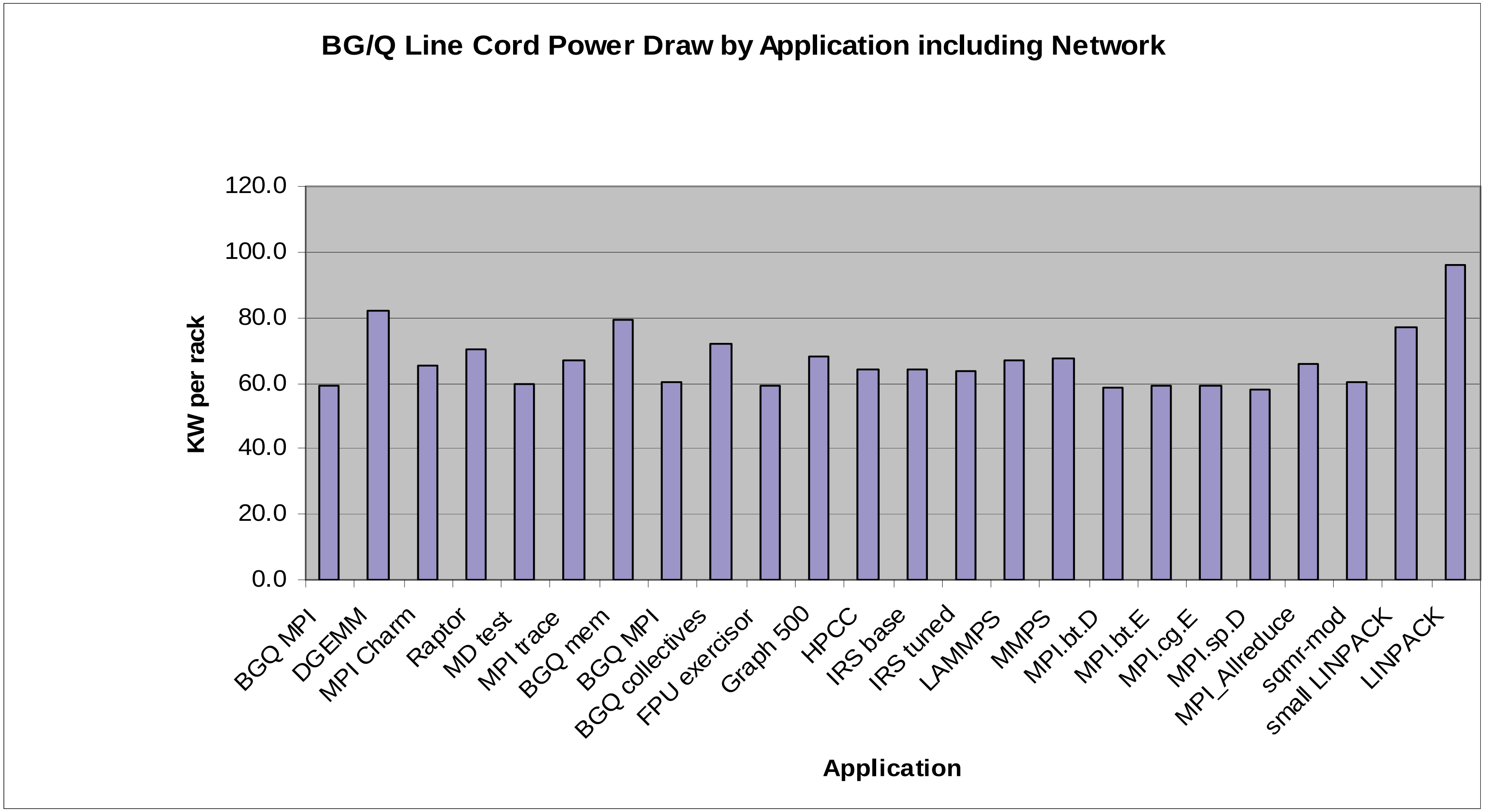



- **Power efficient processor chips allow dense packaging**
- **High bandwidth / low latency electrical interconnect on-board**
- **18+18 (Tx+Rx) 12-channel optical fibers @10Gb/s**
  - Recombined into 8*48-channel fibers for rack-to-rack (Torus) and 4*12 for Compute-to-IO interconnect
- **Compute Node Card assembly is water-cooled (18-25ºC – above dew point)**
- **Redundant power supplies with distributed back-end ~ 2.5 kW**


Full height, 25W PCI cards,
NOT hot serviceable.
~1 KW per I/O Drawer
8 compute cards
(different PN than in compute rack
because of heatsink vs cold plate)
Axial fans – same as BGP
Fiber connections
48V power input (not shown)
Clock input
Ball bearing slides for field maintenance
Picture by Shawn Hall

# BQC Power Measurements

## From 4 rack Prototype System



# Failures per Month per TF

From: http://acts.nersc.gov/events/Workshop2006/slides/Simon.pdf

| | Scale Demonstrated Factor to PF | Failures per month per TF | Power Consumption @PF | Estimated System Cost |
|---|---|---|---|---|
| Cray XT3/XT4 | 10880 CPUs<br>10X to PF<br>~100,000 CPUs | ~.1 - ~1 | ~8MW XT4 | >$150M XT4 |
| Clusters X86/AMD64 | 8000 CPUs<br>12X to PF<br>~100,000 CPUs | 2.6 - 8.0 | ~6MW | >$150M x86 |
| Blue Gene L/P | 131,720 CPUs<br>2.2x to PF<br>294,912 | .01-0.03 | ~2.3MW BG/P | <$100M |

Example: A 100 hr job => BG/Q architecture has 2x advantage in TCO
-MTBF 70 hrs 150 hrs to complete (96 rack BG/Q MTBF target)
-MTBF 7 hrs 309 hrs to complete

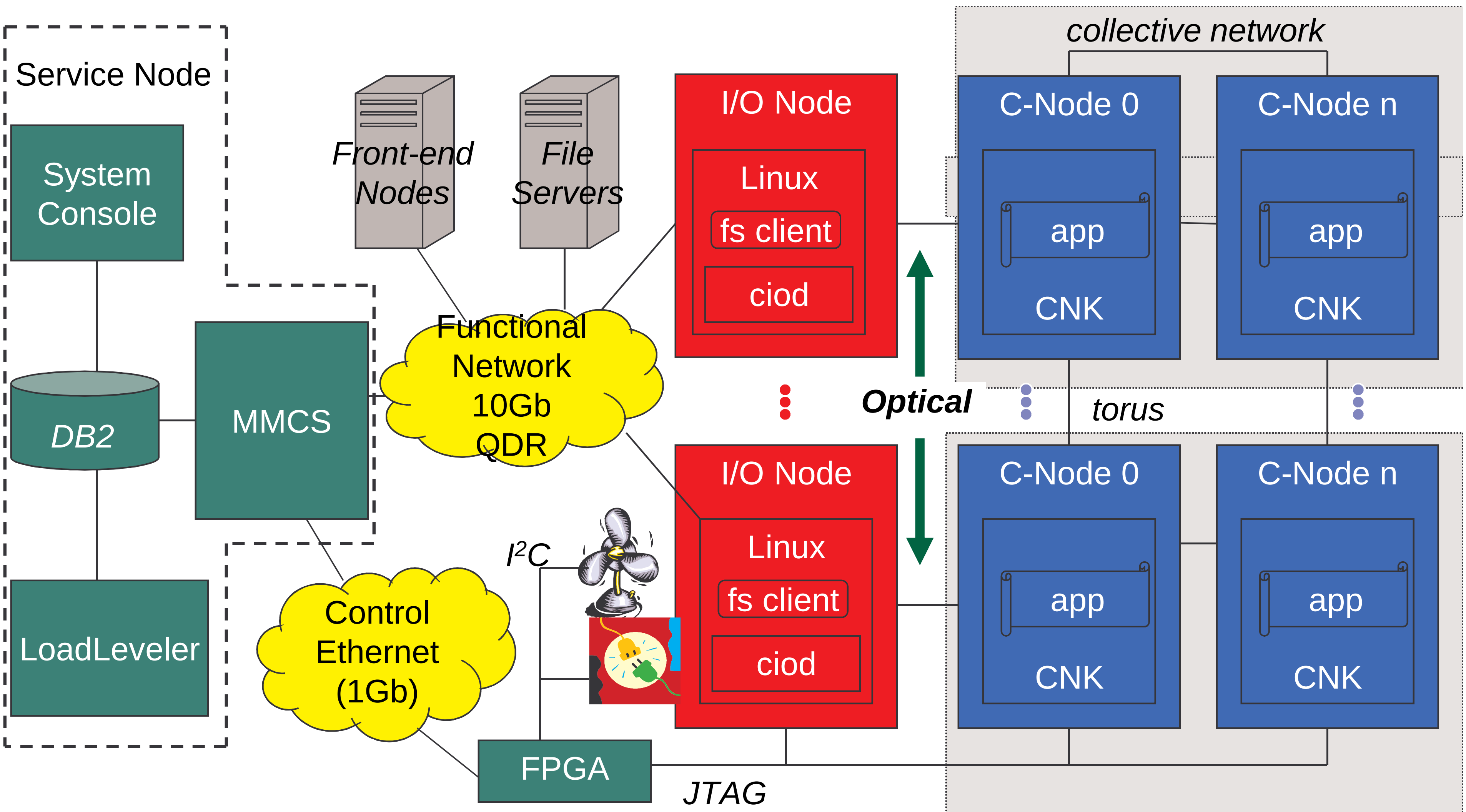
# Blue Gene/Q Software High-Level Goals & Philosophy

- Facilitate extreme scalability
  - Extremely low noise on compute nodes
- High reliability: a corollary of scalability
- Standards-based when possible, leverage other IBM HPC
- Open source where possible
- Facilitate high performance for unique hardware:
  - Quad FPU, DMA unit, List-based prefetcher
  - TM (Transactional Memory), SE (Speculative Execution)
  - Wakeup-Unit, Scalable Atomic Operations
- Optimize MPI and native messaging performance
- Optimize libraries
- Facilitate new programming models

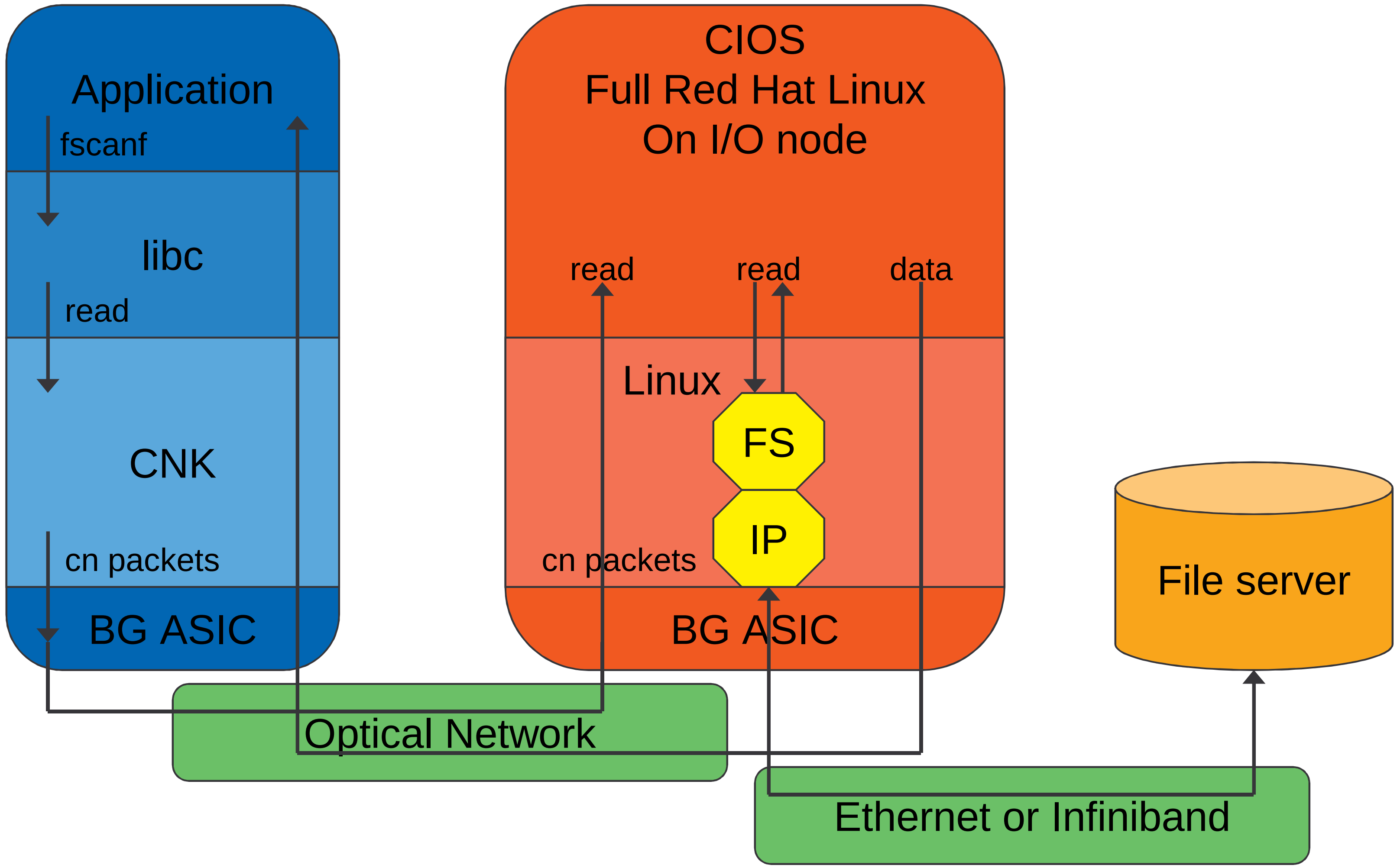
# Software comparison: BG/Q is more general purpose

| Property | | BG/L | BG/Q |
|---|---|---|---|
| Overall Philosophy | Scalability | Scale infinitely, minimal functionality | Scale infinitely, added more functionality |
| | Openness | closed | almost all open |
| Programming Model | Shared Memory | No | Yes |
| | Hybrid | 2 processes 1 thread (software managed) | 1-64 processes 64-1 threads |
| | Low-Level General Messaging | No | PAMI, generic parallel program runtimes, wake-up unit |
| | Programming Models | MPI, ARMCI, global arrays | MPI, OpenMP, UPC, ARMCI, global arrays, Charm++ |
| Kernel | System call interface | proprietary | Linux/POSIX system calls |
| | Library/threading | glibc/proprietary | glibc/pthreads |
| | Linking | static only | static or dynamic |
| | Compute Node OS | CNK | CNK, Linux, Red Hat |
| | I/O Node OS | Linux | SMP Linux with SMT, Red Hat |
| Control | Scheduling | generic API | generic and real-time API |
| | Run Mode | HPC, prototype HTC | Integrated HPC, HTC, MPMD, and sub-blocks, HA with job cont |
| Tools | Tools | HPC Toolkit | HPC Toolkit, Dyninst, Valgrind, PAPI |
| Research Initiatives | OS | Scaling Linux | ZeptoOS, Plan 9 |
| | Big Data | N/A | BGAS (Blue Gene Active Storage), Large memory nodes |
| | Commercial | N/A | Kittyhawk, Cloud, SLAcc |

# Blue Gene System Architecture



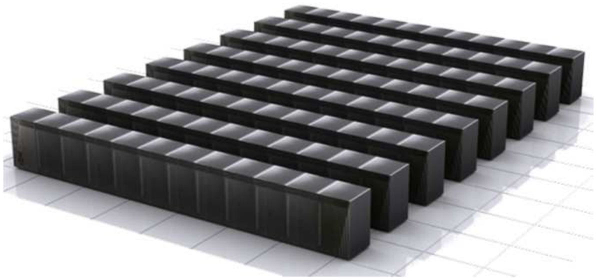
# I/O on Blue Gene/Q



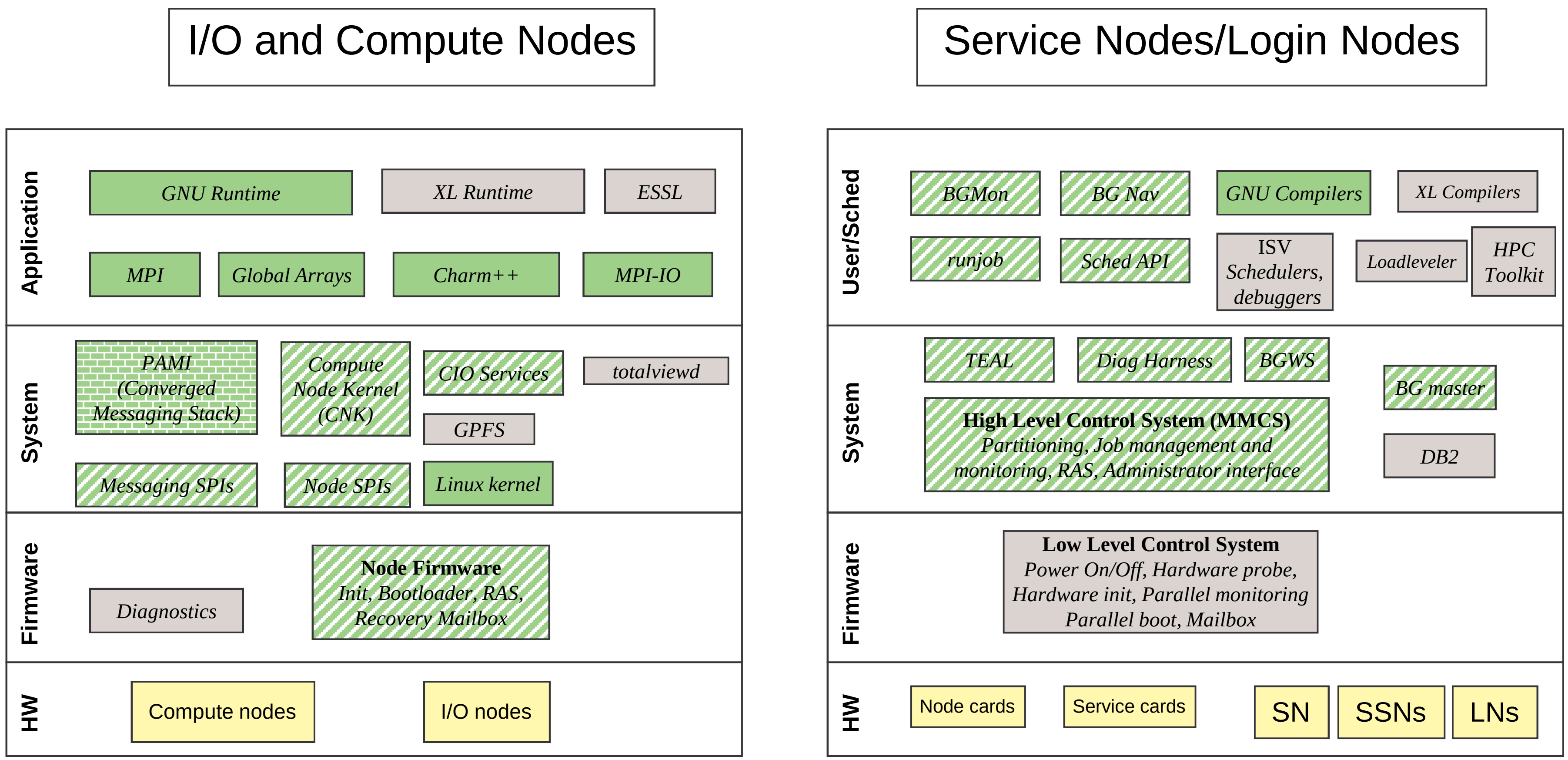
# Blue Gene Q Software Innovations

- 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

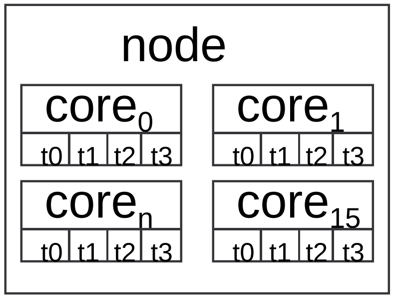


# BG/Q Software Stack Openness



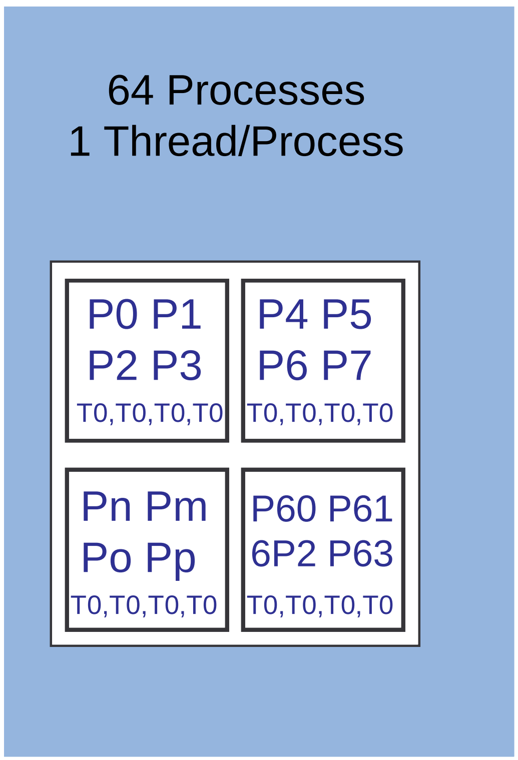

- New open source reference implementation licensed under CPL.
- New open source community under CPL license. Active IBM participation.
- Existing open source communities under various licenses. BG code will be contributed and/or new sub-community started..
- Closed. No source provided. Not buildable.

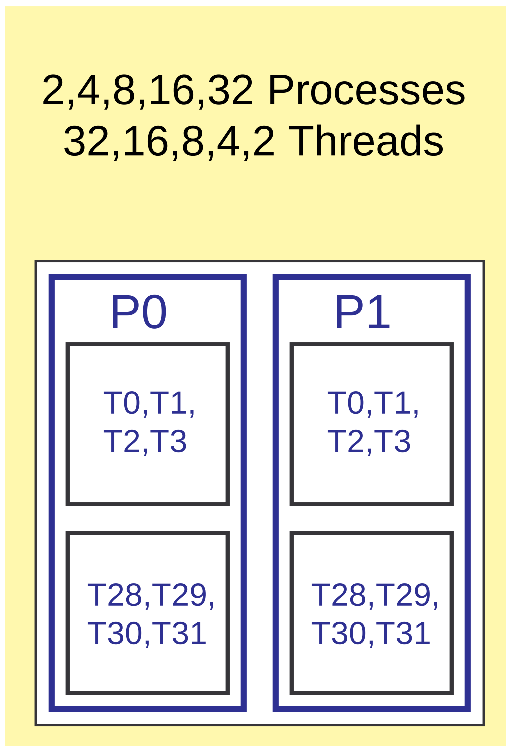
# Execution Modes in BG/Q per Node

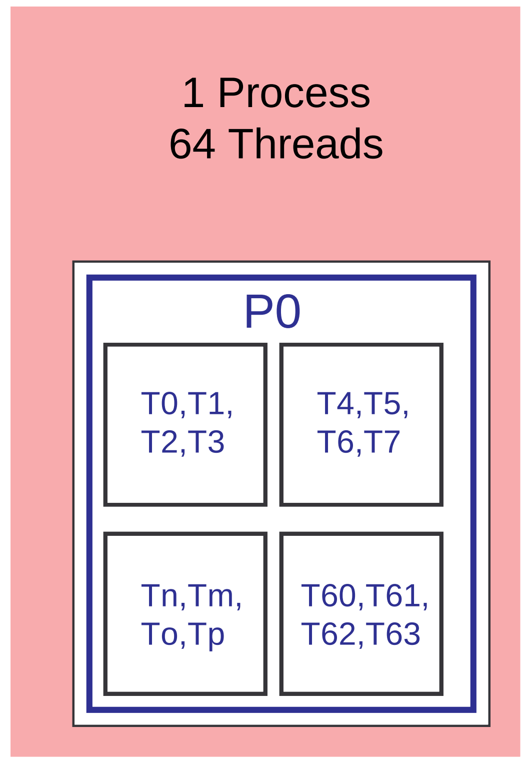



Hardware Abstractions Black
Software Abstractions Blue

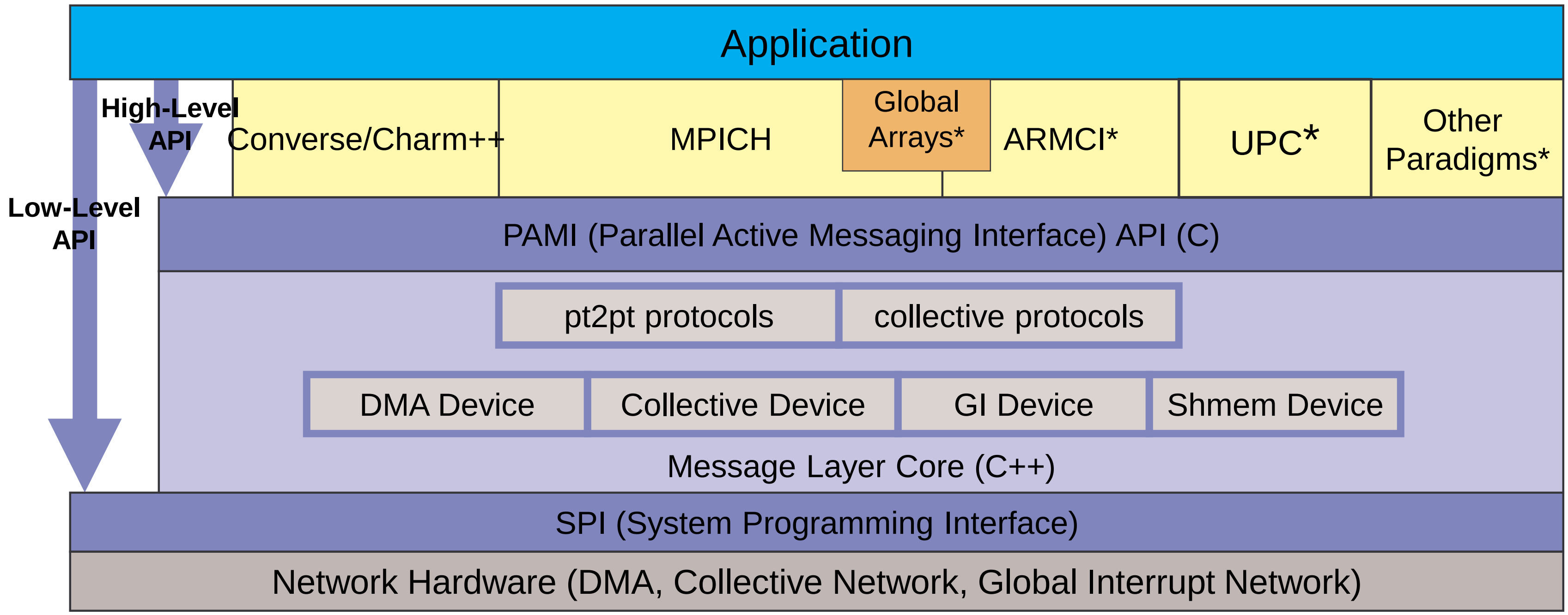
- **Next Generation HPC**
  - **Many Core**
  - **Expensive Memory**
  - **Two-Tiered Programming Model**



# Parallel Active Message Interface

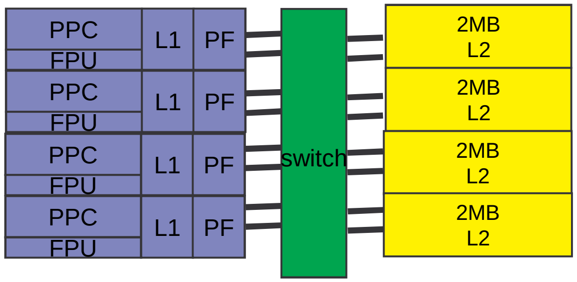



- **Message Layer Core has C++ message classes and other utilities to program the different network devices**
- **Support many programming paradigms**
- **PAMI runtime layer allows uniformity across IBM HPC platforms**

***describes capability not necessarily product support**

# Advantages of Software/Hardware Co-Design on BG/Q
## (helping take advantage of multi-core environment)

- Scalable atomic instructions
  – Enables development of lock-less producer consumer queues with N producers and 1 or more consumers
- Hardware wake-up mechanism
  – Support for OpenMP/MPI and other hybrid programming models
- List-based prefetching
  – Allows efficient use of cache for broader applications
- Multi-valued L2 cache
  – TM and TLS

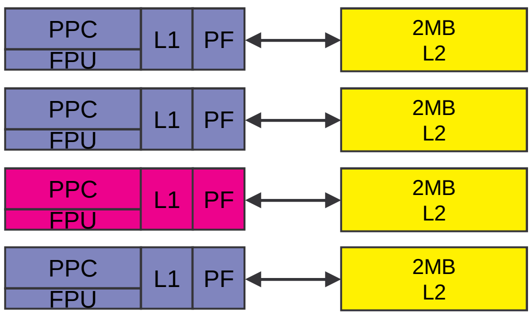



# Standard Atomic Operation

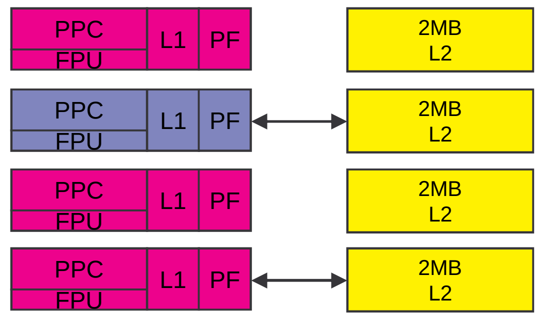
## (lwarx stx on PowerPC)

- N round trips
  - Where N is the number of threads
  - For N=64 and L2 74 cycles →~9500 cycles

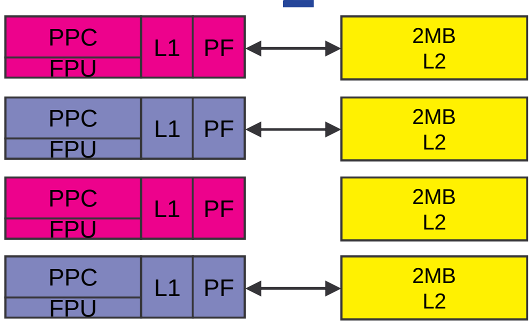
1



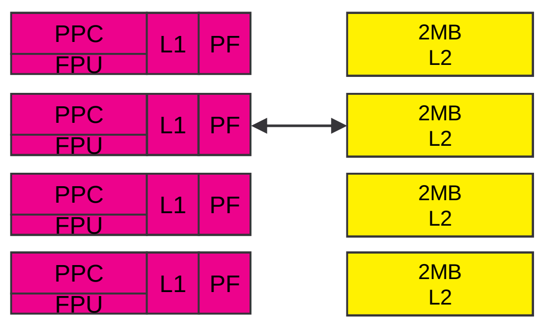

3

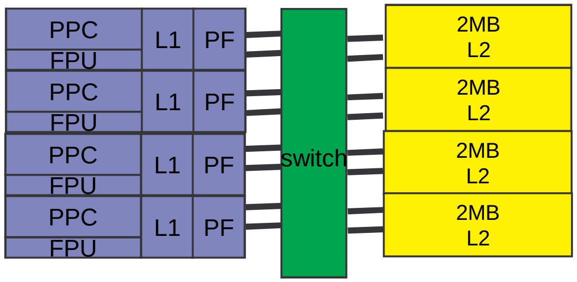



2




4



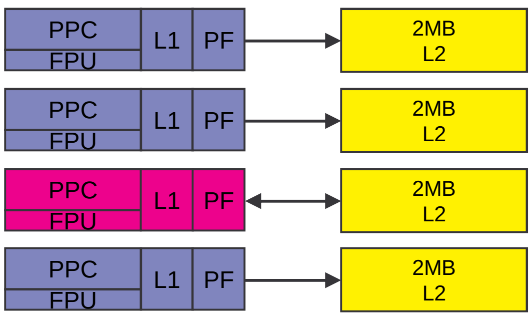



# Scalable Atomic Operation

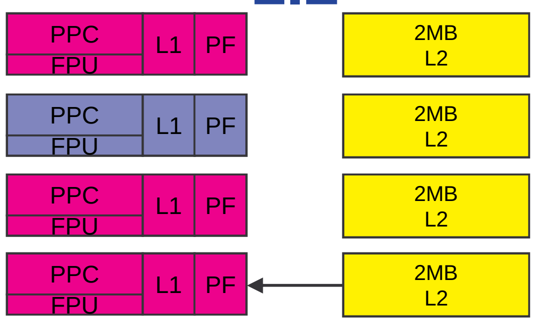
## (fetch_and_inc for example)

- 1 round trips + N L2 cycles
  - Where N is the number of threads
  - For N=64 and L2 74 cycles →~800 cycles
    - Compared to ~9500 cycles for standard

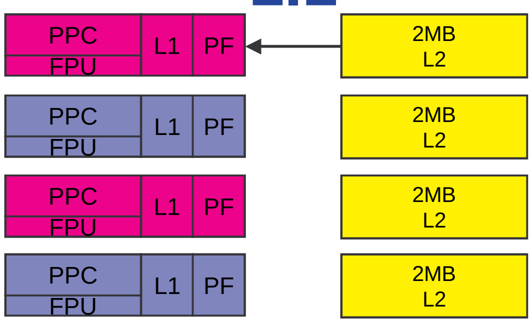
## 1



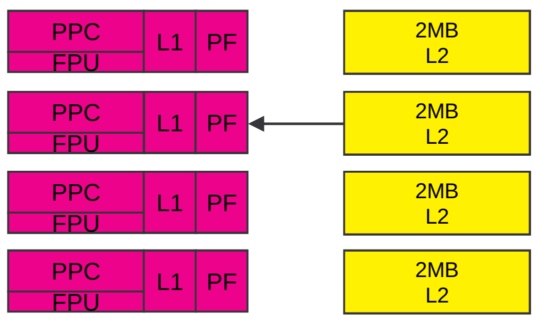

## 1.2



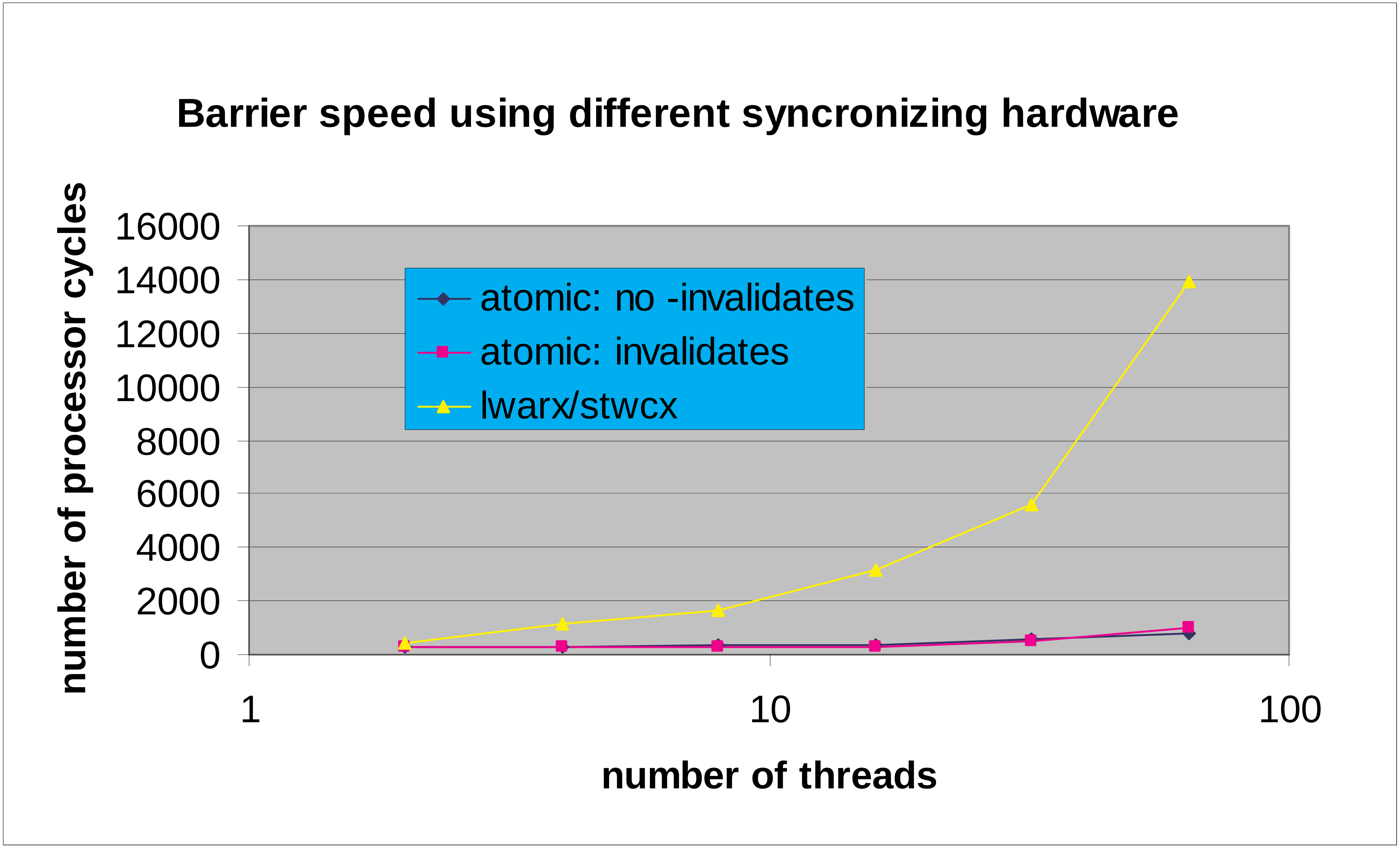

## 1.1



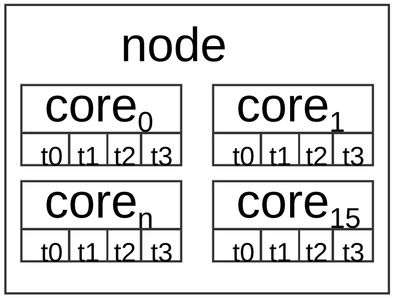

## 1.3



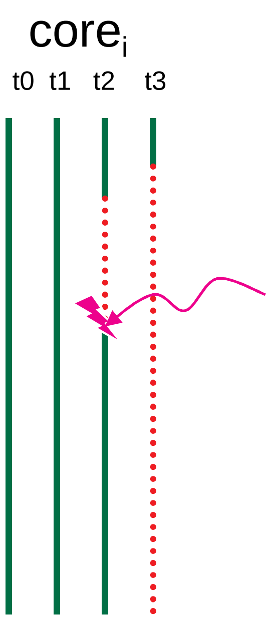
# Use of Scalable Atomic ops



# Wakeup Unit




- Allow hardware threads to stop executing instructions
  - Only two threads needed to keep A2 utilized
- Avoids software polling
- Waiting thread configured to wake up on choice of interrupt




IPI (Inter Processor Interrupt)
MU (Messaging Unit Interrupt)
L2

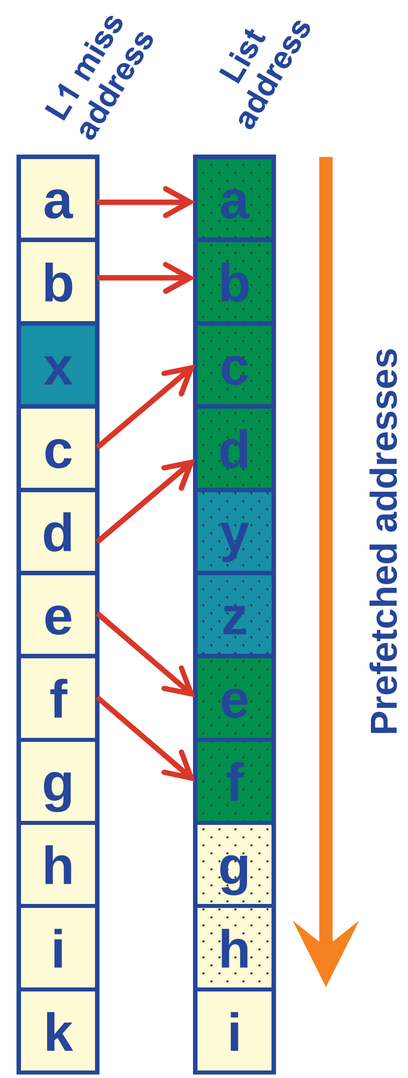
# List-Based prefetching for LLNL IRS Sequoia kernel

- Benchmark with distributed pattern:

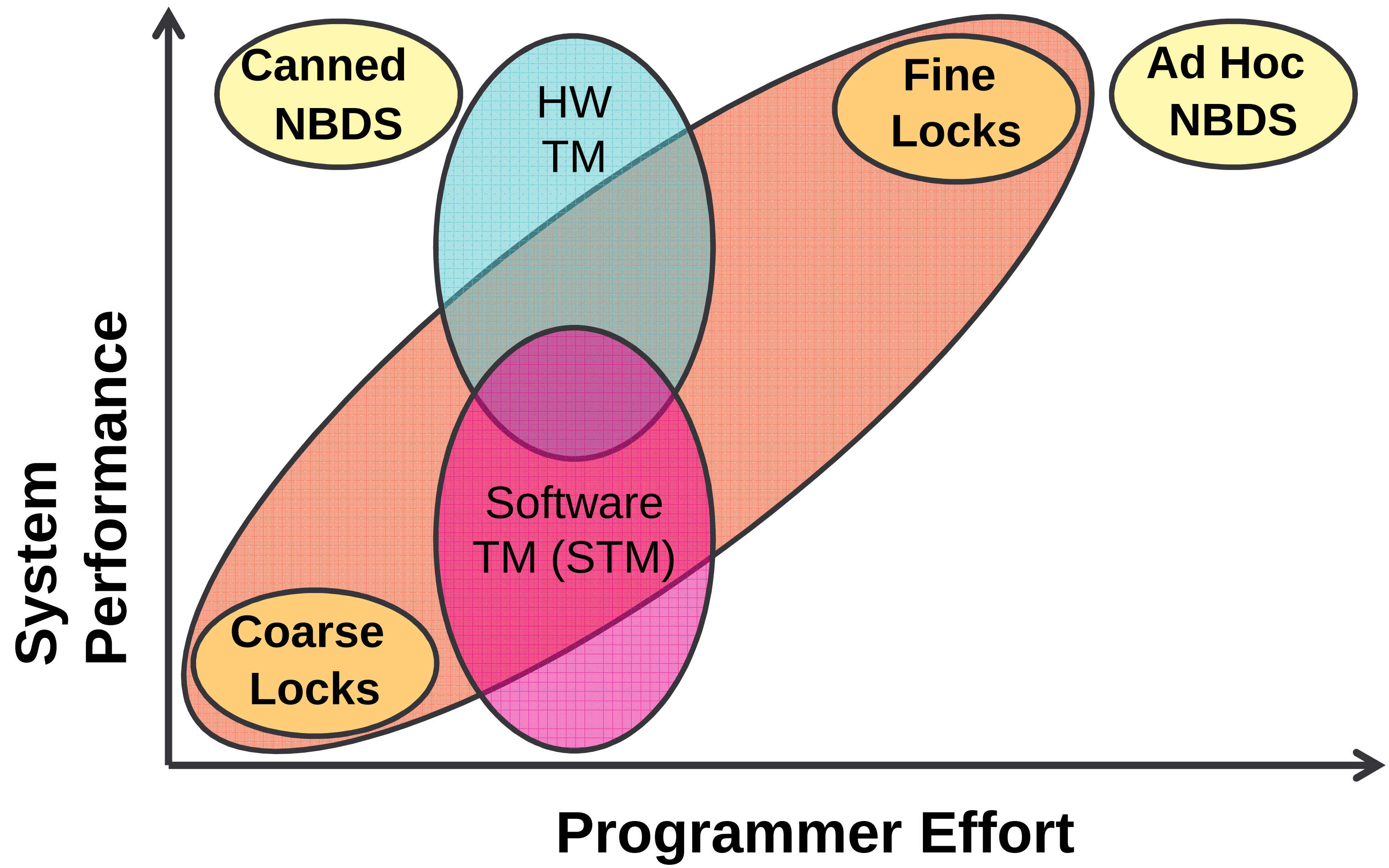
```
for ( kk = kmin ; kk < kmax ; kk++ ) {            <- Start list
    for ( jj = jmin ; jj < jmax ; jj++ ) {
      for ( ii = imin ; ii < imax ; ii++ ) {
         i = ii + jj * jp + kk * kp ;
         b[i] = dbl[i] * xdbl[i] + dbc[i] * xdbc[i] + dbr[i] * xdbr[i] + …
      } } }                                        <- Stop list
```

# “Perfect” Prefetching

- Tolerance to missing or extra cache misses
  - **Possible asynchronous A2 behavior may cause out-of-sync addresses to be issued or initially recorded.**
  - **An L1 miss not matching the next list address will be compared to the next *N* addresses in the list.  A match will cause list prefetching to continue from the point of match.**
  - **An L1 miss not matching these *N*+1 addresses will be discarded and the next miss addressed compared.  *M* sequential such failures will cause the list to be abandoned.**
  - **When a list is abandoned:**
    - **Stream prefetching is activated.**
    - **List recording continues.**
  - **In all cases list recording of each L1 miss address continues until *stop list* is asserted.  The new list then overwrites the original one.**
  - **Such self-healing and adaptation is likely needed since the address pattern will change as the list repeats and prefetching becomes more accurate.**



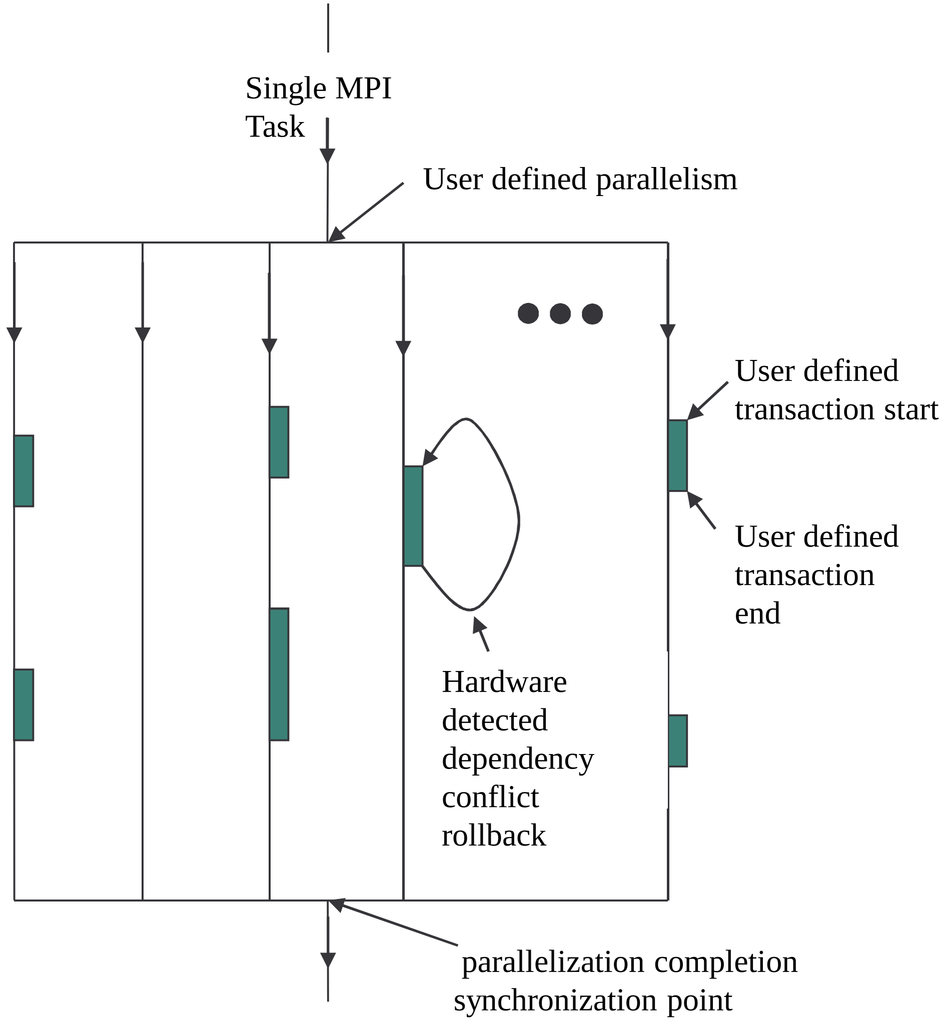
# Concurrency Design Space
# Putting TM in Perspective




Credit: Bill Scherer

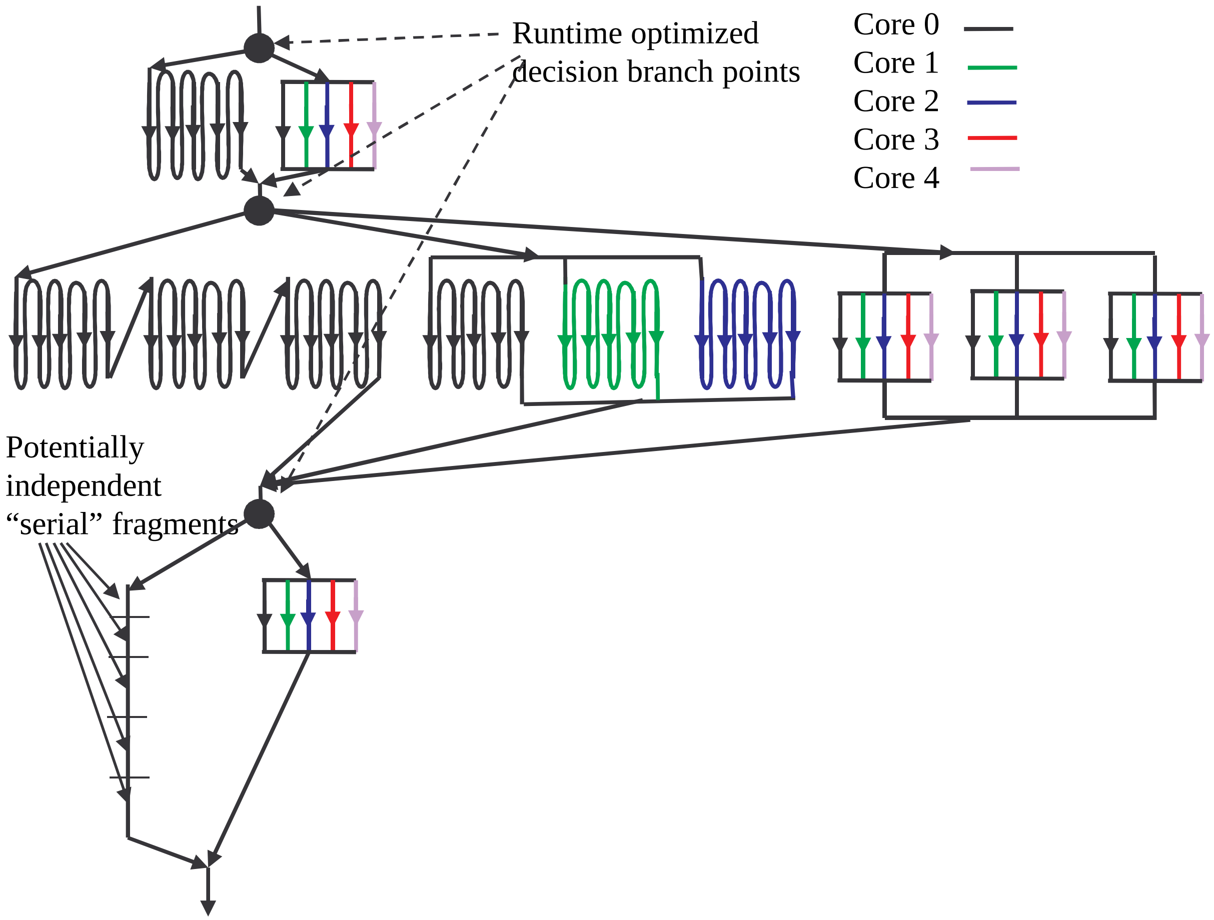
# BlueGene/Q transactional memory mode

- **User program model:**
  - –User defines parallel work to be done
  - –User explicitly defines start and end of transactions within parallel work that are to be treated as atomic
- **Compiler implications:**
  - –Interpret user program annotations to spawn multiple threads
  - –Interpret user program annotation for start of transaction and save state to memory on entry to transaction to enable rollback
  - –At end of transaction program annotation test for successful completion and optionally branch back to rollback pointer.
- **Hardware implications:**
  - –Transactional memory support required to detect transaction failure and rollback
  - –L1 cache visibility for L1 hits as well as misses allowing for ultra low overhead to enter a transaction



# BlueGene/Q 0'th compiler support for TLS

**Program Execution Flow with TLS**



# Summary Blue Gene/Q

## 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. Lowest Total Cost of Ownership

- Highest total power efficiency, smallest footprint
- Typically 2 orders of magnitude better reliability

## 3. Broad range of applications reach

- Familiar programming models
- Easy porting from other environments

## 4. Foundation for Exascale exploration