From Blue Gene to Cell

Power.org Moscow, JSCC Technical Day
November 30, 2005

Dr. Luigi Brochard
IBM Distinguished Engineer
Deep Computing Architect
luigi.brochard@fr.ibm.com

Technology Trends

- As frequency increase is limited due to power limitation
- Dual core is a way to :
  - 2 x Peak Performance per chip (and per cycle)
  - But at the expense of frequency (around 20% down)
- Another way is to increase Flop/cycle

IBM innovations

- POWER :
  - FMA in 1990 with POWER: 2 Flop/cycle/chip
  - Double FMA in 1992 with POWER2 : 4 Flop/cycle/chip
  - Dual core in 2001 with POWER4: 8 Flop/cycle/chip
  - Quadruple core modules in Oct 2005 with POWER5: 16 Flop/cycle/module
- PowerPC:
  - VMX in 2003 with ppc970FX : 8 Flops/cycle/core, 32bit only
  - Dual VMX+ FMA with pp970MP in 1Q06
- Blue Gene:
  - Low frequency , system on a chip, tight integration of thousands of cpus
- Cell :
  - 8 SIMD units and a ppc970 core on a chip : 64 Flop/cycle/chip

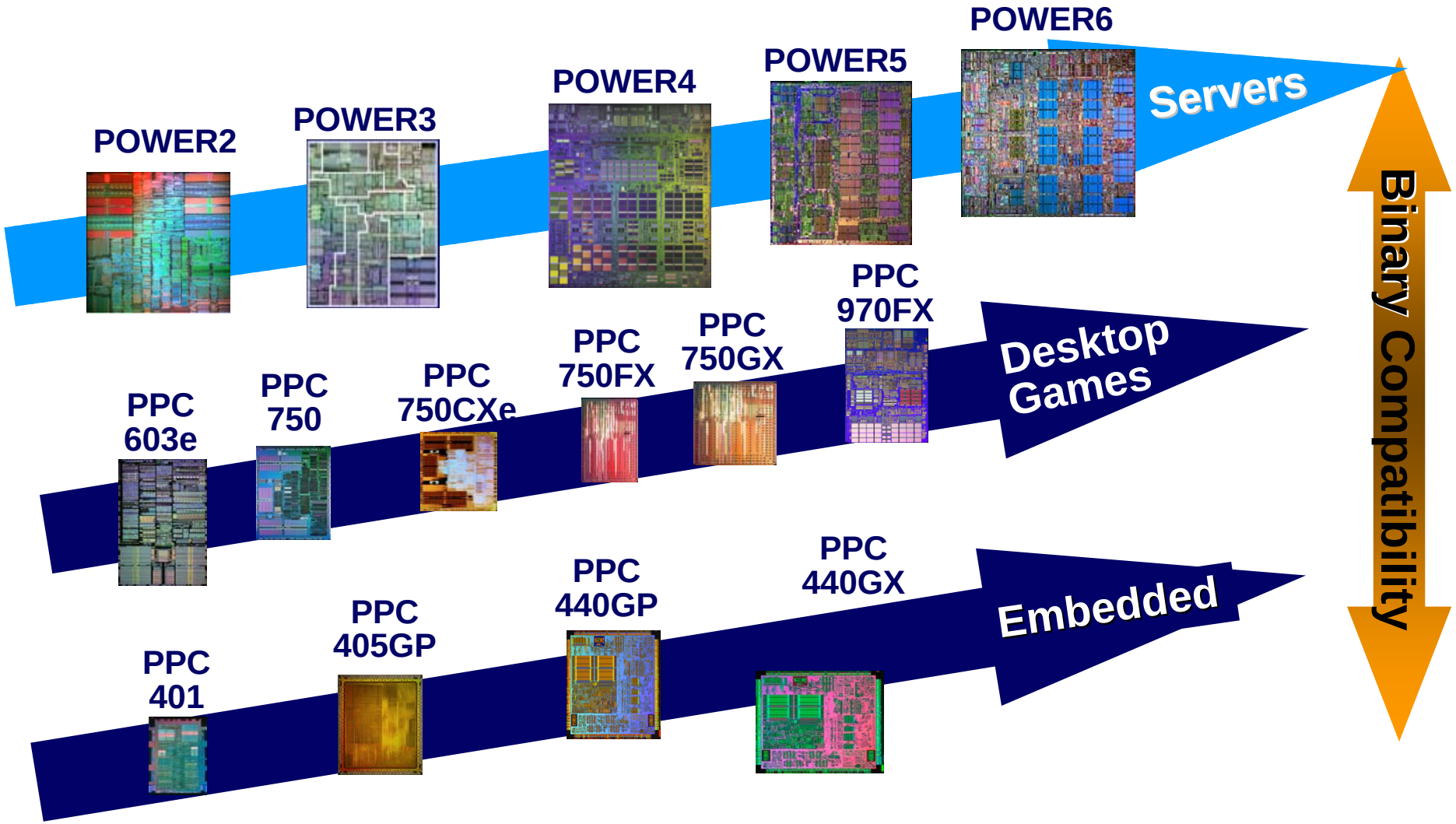
Technology Trends

- As needs diversify, systems are heterogeneous and distributed
  - GRID technologies are an essential part to create cooperative environments based on standards

IBM innovations

- IBM is :
  - a sponsor of Globus Alliances
  - contributing to Globus Tool Kit open souce
  - a founding member of Globus Consortium
- IBM is extending its products
  - Global file systems :
    - Multi platform and multi cluster GPFS
  - Meta schedulers :
    - Multi platform and multi cluster Loadleveler
- IBM is collaborating to major GRID projects
  - DTF in US
  - DEISA in Europe

PowerPC : The Most Scalable Architecture

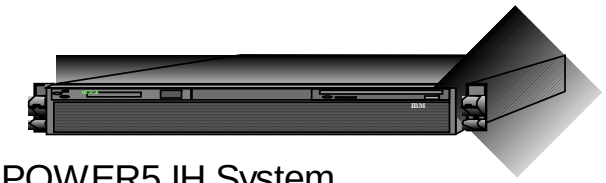


POWER5 Improves High Performance Computing

- Higher “Sustained/Peak FLOPS” ratio compared to POWER4
  - Increased rename resources allow higher instruction level parallelism (from 72 to 120)
- Significant reduction in L3 and memory latency
- Improved memory bandwidth per FLOP
- Simultaneous Multi Threading (SMT)
- Fast barrier synchronization operation
- Advanced data prefetch mechanism
- Stronger SMP coupling and scalability

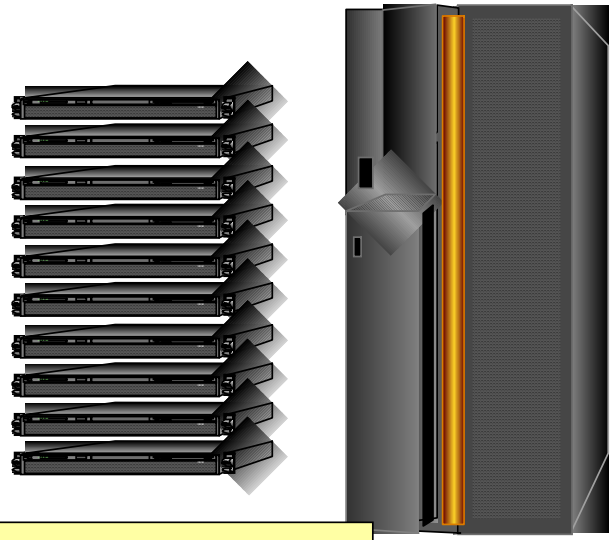
POWER5 p575 Overview

.



POWER5 IH System

- 2U rack chassis
  - Rack: 24" X 43 " Deep, Full Drawer



12 Servers / Rack
96/192 Processors / Rack

	POWER5 IH Node
Architecture	8 or16 POWER5 1.9 or 1.5 GHz
L3 Cache	144MB / 288MB
Memory	2GB - 256GB DDR1
Packaging	2U (24" rack) 12 Nodes / Rack
DASD / Bays	2 DASD (Hot Plug)
I/O Expansion	6 slots (PCI-X)
Integrated SCSI	Ultra 320
Integrated Ethernet	4 Ports 10/100/1000
RIO Drawers	Yes (1/2 or 1)
LPAR	Yes
Switch	HPS and Myrinet
OS	AIX & Linux

p575 Sustained Performance

p5-575 8 way Benchmark publications

Sq IH/ p5-575	1w	8w
SPECint_2000	1,456	
SPECfp2000	2,600	
SPECint_rate2000		167
SPECfp_rate2000		282
Linpack DP	1.776	
Linpack TPP (n=1000)	5.872	34.5
Linpack HPC	7,120	56,6
STREAM standard		41,5
STREAM tuned		55,7

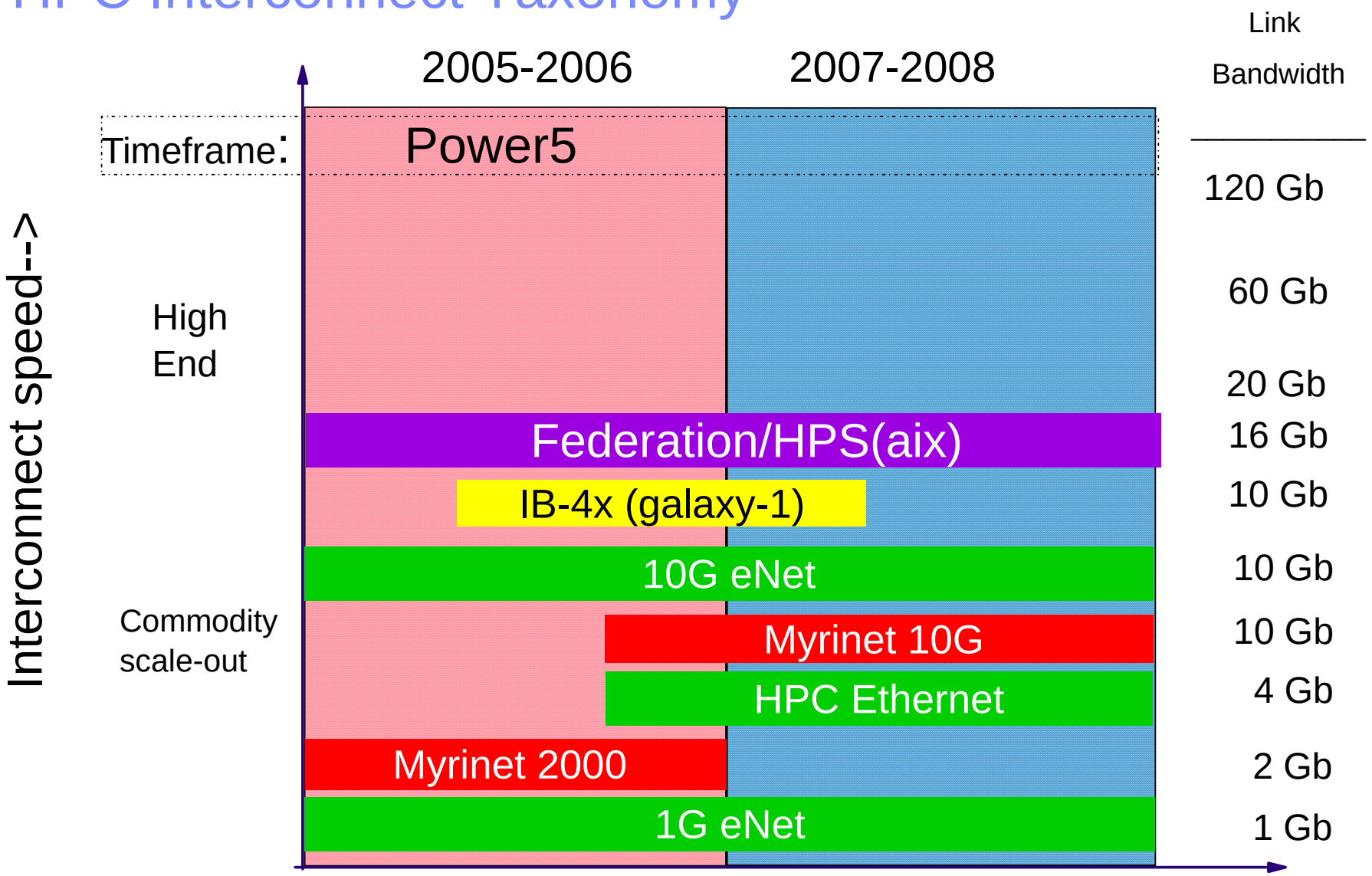
p5-57516 way Benchmark publications

Sq IH/ p5-575	1w	16w
SPECint_2000	1,143	
SPECfp2000	2,185	
SPECint_rate2000		238
SPECfp_rate2000		385
Linpack DP		
Linpack TPP (n=1000)		
Linpack HPC		87,3
STREAM standard		42,6
STREAM tuned		55,8

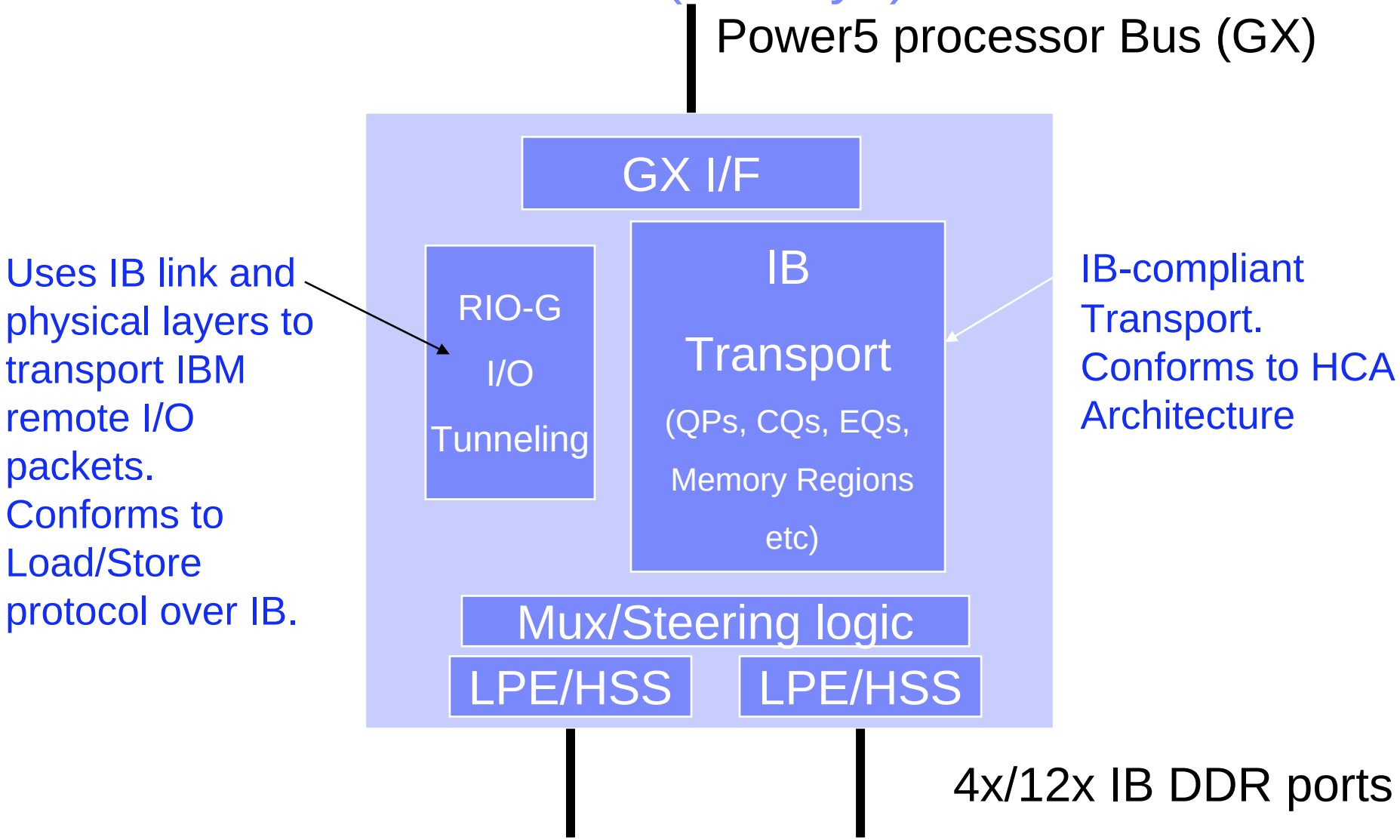
p575 Peak Performance and Memory Bandwidth

Processor Type	Frequency	Peak Perf by node/rack GFlops	Memory BW Byte/Flop
p575 single core DDR1	1.9 GHz	60 / 720	1,6
p575 dual core DDR1	1.5 GHz	90 / 1150	1.1

HPC Interconnect Taxonomy



IBM HCA for Power5 (Galaxy1)



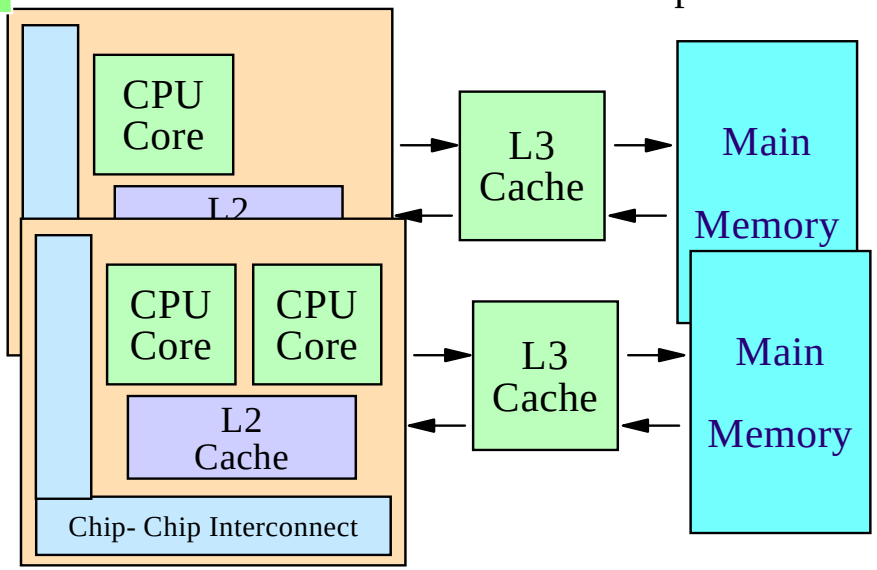
- Power PC 970

PowerPC 970 and POWER4 Comparaison

Memory Bandwidth 12.8 GB/s

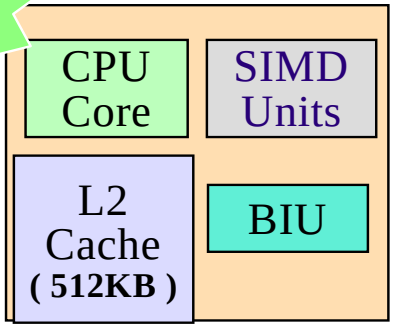
POWER4+

- SMP Optimized
- Balanced system/bus design
- One or two processor cores



Memory Bandwidth 6.4 GB/s

PowerPC 970



- Single processor core
  - ✓2 Load/store units
  - ✓2 Fixed point units
  - ✓2 IEE Floating point units
  - ✓2 SIMD sub-units (VMX 32 bit)
  - ✓Branch unit
  - ✓Condition register unit

- 8 Execution pipeline
  - 2 load / store units
  - 2 fixed point units
  - 2 DP multiply-add execution units
  - 1 branch resolution unit / 1 CR execution unit

- 8 prefetching streams

BladeCenter Server Overview

- **Enterprise-class shared infrastructure**
  - **Shared power, cooling, cabling, switches means reduced cost and improved availability**
  - **Enables consolidation of many servers for improved utilization curve**
- **High performance and density: 16 blades in a 7U rack => 192 processors in a 42U rack**
- **Available in 2 way or 4 way Xeon(HS20, HS40), 2 way Opteron (JS20) or 2 way PowerPC (JS20)**

Scale out platform for growth

BladeCenter JS20 Roadmap

GA1-JS20 1.6GHz
- 2x (std.) 1.6 GHz PPC 970
- 512 MB (srd.) to 4GB Memory
- 0-2 40GB IDE drives
- SLES 8, RHEL AS

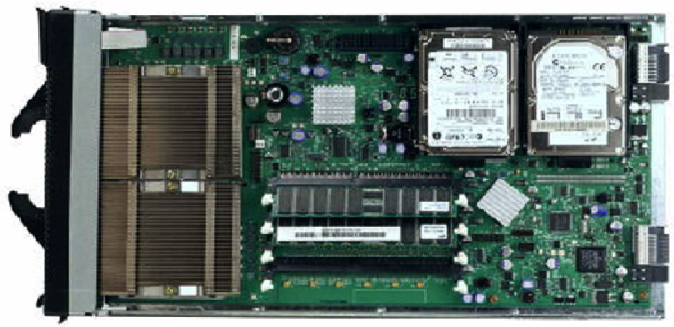
JS20 2.2GHz
- 60GB IDE Drives
- 11/12/04

GA2-JS20 2.2GHz
- 2x (std.) 2.2 GHz PPC 970+
- 512MB (std.) to 4GB Memory
- 0-2 40, 60GB IDE drives
- SLES 8/9, RHEL AS 3, AIX 5L V5.2/5.3 Myrinet, CSM, USB DVD-ROM

GA2 2Q05
- InfiniBand
- iSCSI
- 2GB DIMMs
- BCT & NEBSupport

GA3 : 1Q06

GA4/5



* All statements regarding IBM future directions and intent are subject to change or withdrawal without notice.

- Blue Gene

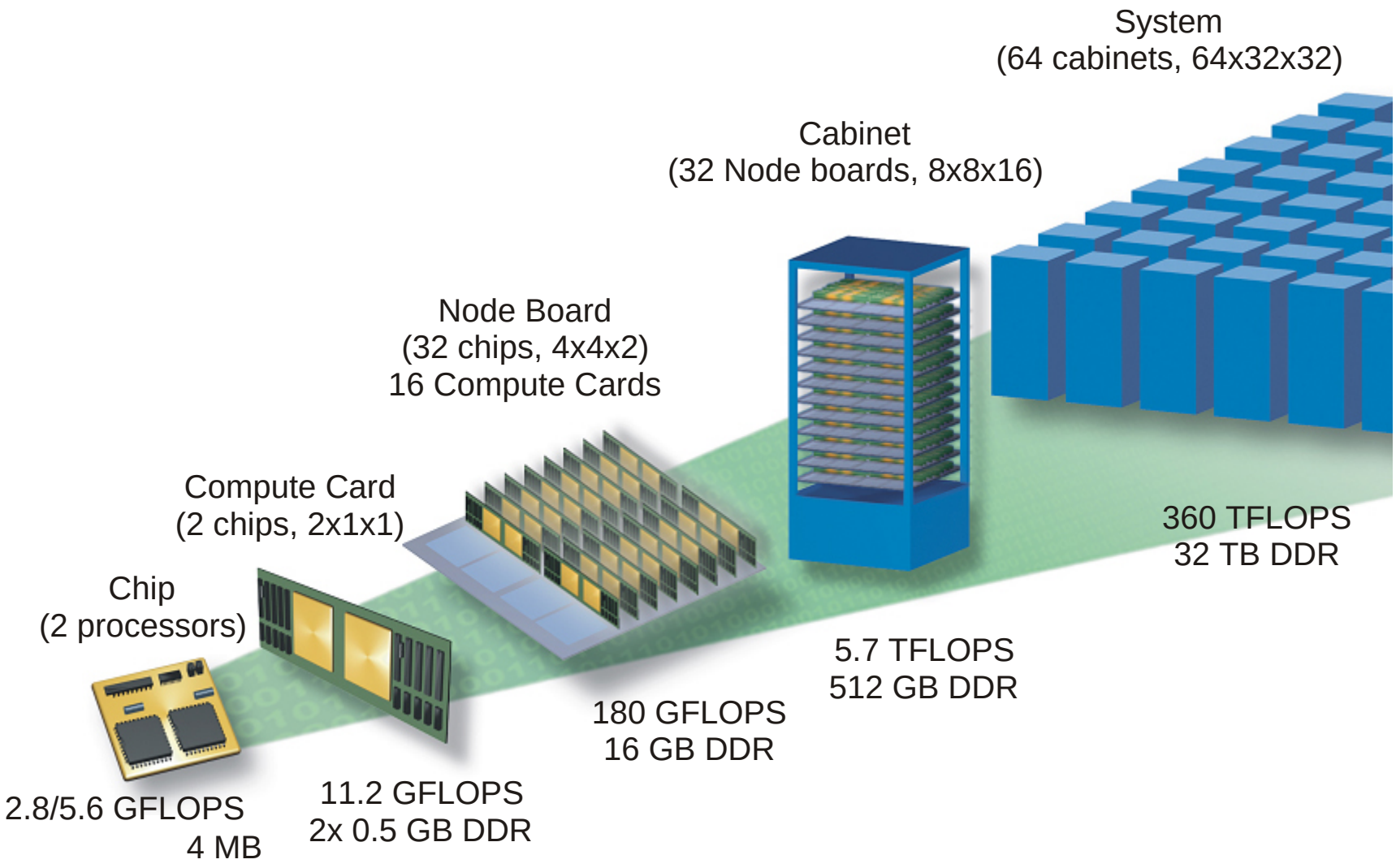
BlueGene/L Project Motivations

- **Traditional supercomputer-processor design is hitting power/cost limits**
- **Complexity and power are major driver for cost and reliability**
- **Integration, power, and technology directions are driving toward multiple modest cores on a single chip rather than one high-performance processor**
  - Optimal design point is very different from standard approach based on high-end superscalar nodes
  - Watts/FLOP will not improve much from future technologies.
- **Applications for supercomputers do scale fairly well**
  - Growing volume of parallel applications
  - Physics is mostly local

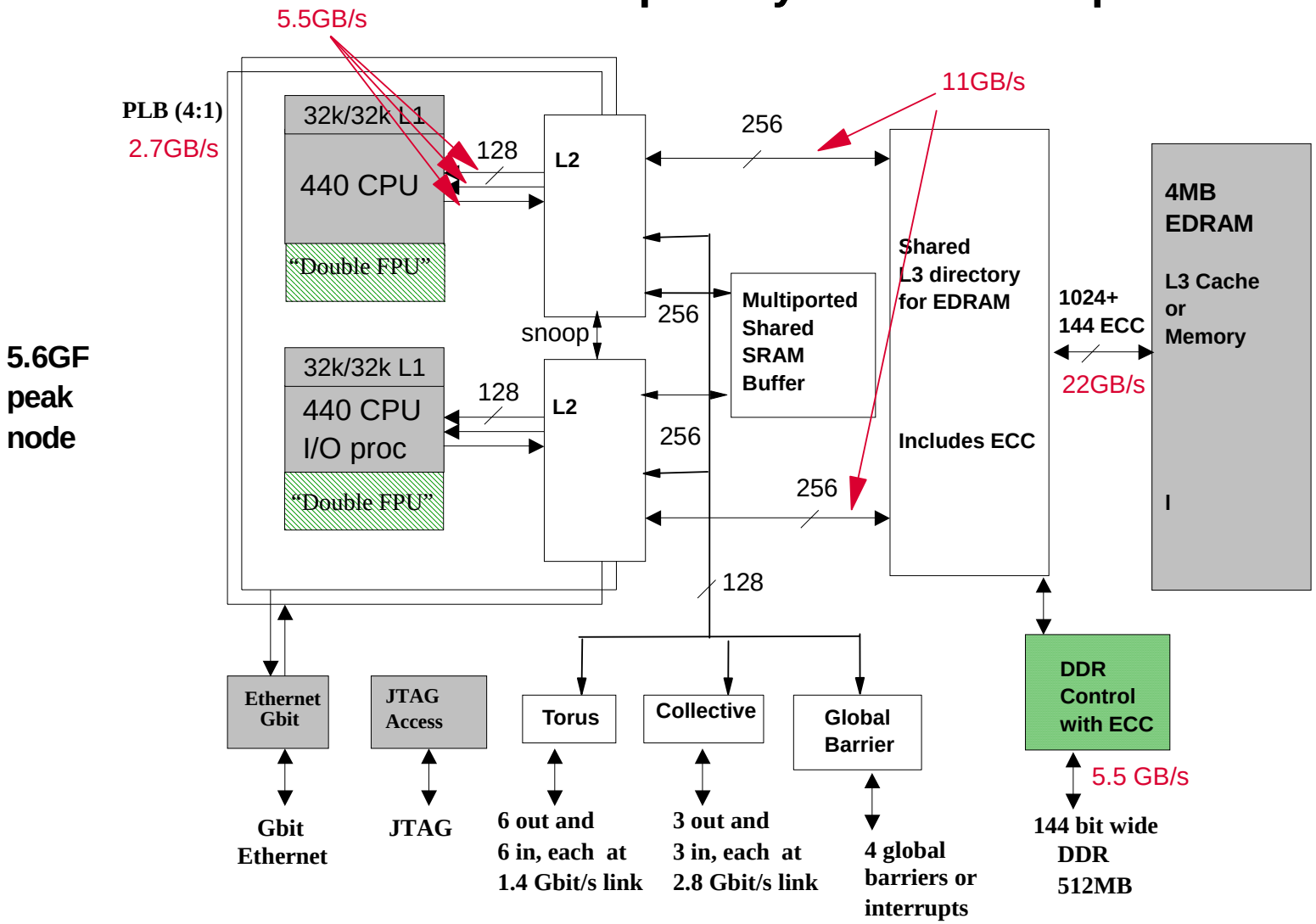
IBM approach with Blue Gene

- **Use embedded system-on-a-chip (SOC) design**
  - **Significant reduction of complexity**
    - **Simplicity is critical, enough complexity already due to scale**
  - **Significant reduction of power**
    - **Critical to achieving a dense and inexpensive packaging solution.**
  - **Significant reduction in time to market, lower development cost and lower risk**
    - **Much of the technology is qualified.**
- **Utilize PowerPC architecture and standard messaging interface (MPI).**
  - **Standard, familiar programming model and mature compiler support.**
- **Integrated and tighly coupled networks**
  - **To reduce wiring complexity and sustain performance of applications on large number of nodes**
- **Close attention to RAS (reliability, availability, and serviceability) at all system levels.**

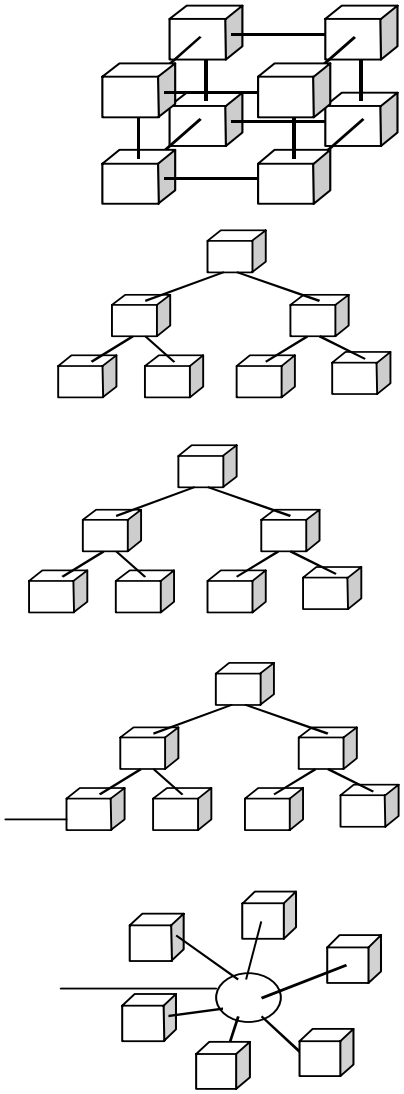
BlueGene/L



BlueGene/L Compute System-on-a-Chip ASIC



Blue Gene Networks



3 Dimensional Torus

- 32x32x64 connectivity
- Backbone for one-to-one and one-to-some communications
- 1.4 Gb/s bi-directional bandwidth in all 6 directions (Total 2.1 GB
- ~100 ns hardware node latency

Collective Network

- Global Operations
- 2.8Gb/s per link , 68TB/s aggregate bandwidth
- Arithmetic operations implemented in tree
  - Integer/ Floating Point Maximum/Minimum
  - Integer addition/subtract, bitwise logical operations
- Latency of tree less than 2.5usec to top, additional 2.5usec t
- Global sum over 64k in less than 2.5 usec (to top of tree)

Global Barriers and Interrupts

- Low Latency Barriers and Interrupts

Gbit Ethernet

- File I/O and Host Interface
- Funnel via Global Tree network

Control Network

- Boot, Monitoring and Diagnostics

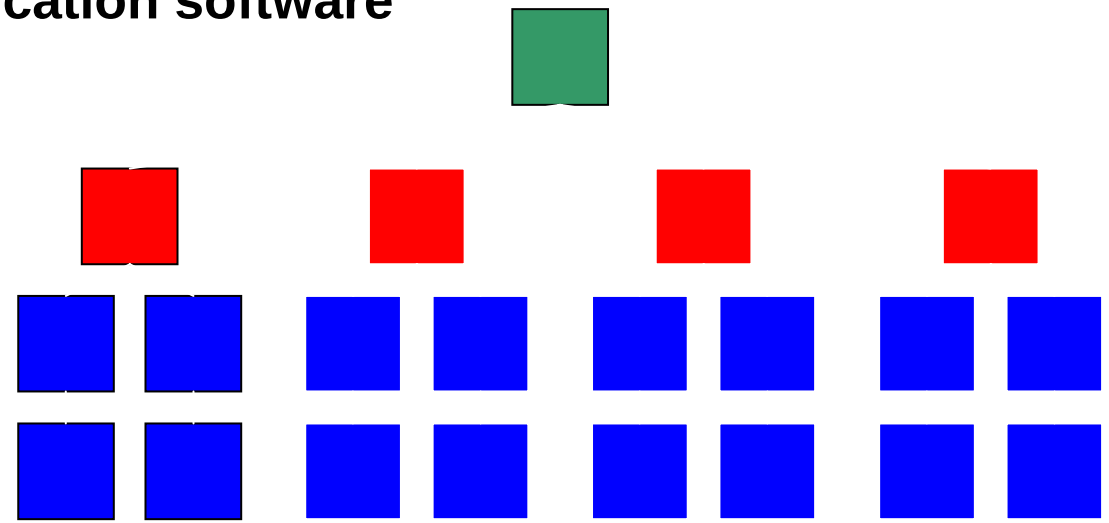
BG/L is a well balanced system

System	Memory Bandwidth GB/s Byte/Flop	Memory Latency ns cycles	Network Latency us cycles	Network Barrier 128 cpu us cycles
BG/L	**2.2** **0,39**	**110** **77**	**3.35** **2345**	**6,75** 4725
Opteron Infiniband	**6** **0,34**	**120** **264**	**4.98** **17900**	**39** 85800
POWER5 Federation	**41** **0,67**	**120** **228**	**4.8** **8160**	**29,5** 50150

(*) POWER5 barrier is based on 8 way node , Opteron is based on 4 core blades

BlueGene/L Software Hierarchical Organization

- Compute nodes **dedicated to running user application, and almost nothing else - simple compute node kernel (CNK)**
- I/O nodes **run Linux and provide a more complete range of OS services – files, sockets, process launch, signaling, debugging, and termination**
- Service node **performs system management services (e.g., heart beating, monitoring errors) - transparent to application software**

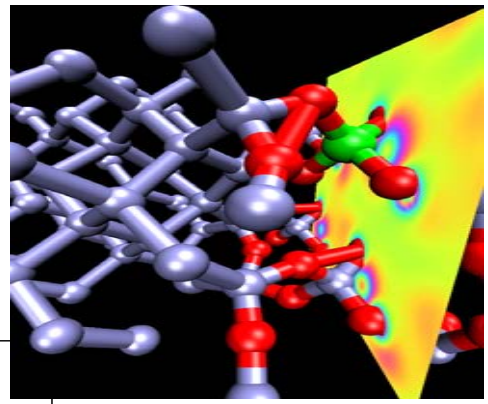


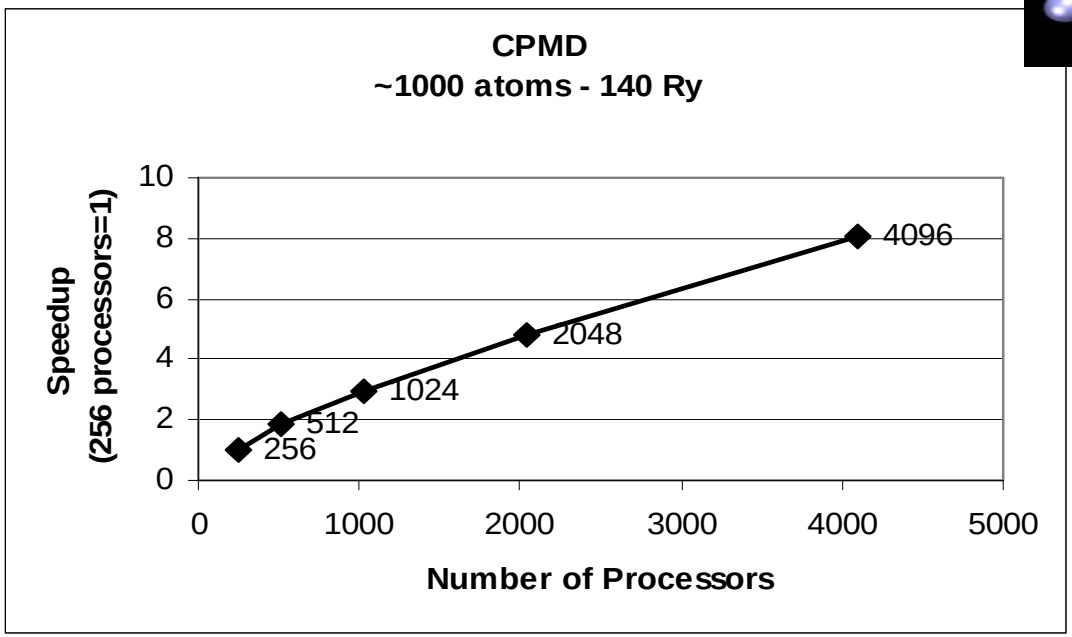
Programming Models and Development Environment

- **Familiar Aspects**
  - SPMD model - Fortran, C, C++ with MPI (MPI1 + subset of MPI2)
    - Full language support
    - Automatic SIMD FPU exploitation
  - Linux development environment
    - User interacts with system through FE nodes running Linux – compilation, job submission, debugging
    - Compute Node Kernel provides look and feel of a Linux environment – POSIX system calls (with restrictions)
  - Tools – support for debuggers (Etnus TotalView), MPI tracer, profiler, hardware performance monitors, visualizer (HPC Toolkit, Paraver, Kojak)
- **Restrictions (lead to significant scalability benefits)**
  - Strictly space sharing - one parallel job (user) per partition of machine, one process per processor of compute node
  - Virtual memory constrained to physical memory size
- **Same environment as other IBM POWER systems**

CPMD - Alessandro Curioni, Salomon Billeter, Wanda Andreoni

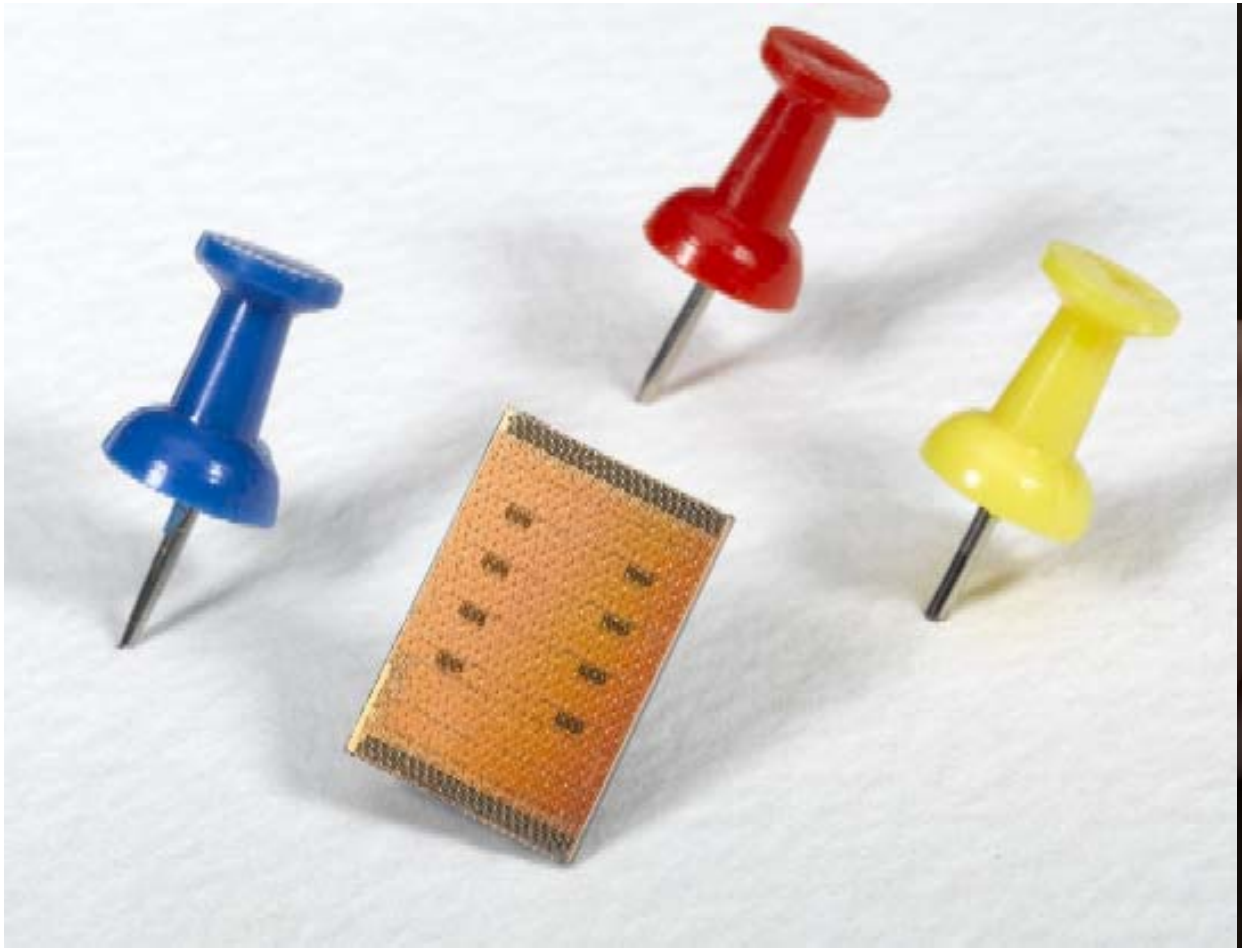
Developed at IBM Zurich and other Universities
from Car Parinello method for Molecular Dynamics
Uses Plane Wave Basis functions, FFT, MPI_Collectives
Demo – Si/SiO2 Interface **% peak ~ 60 % VNM**
Ongoing project : IBM/LLNL PdH Hydrogen Storage

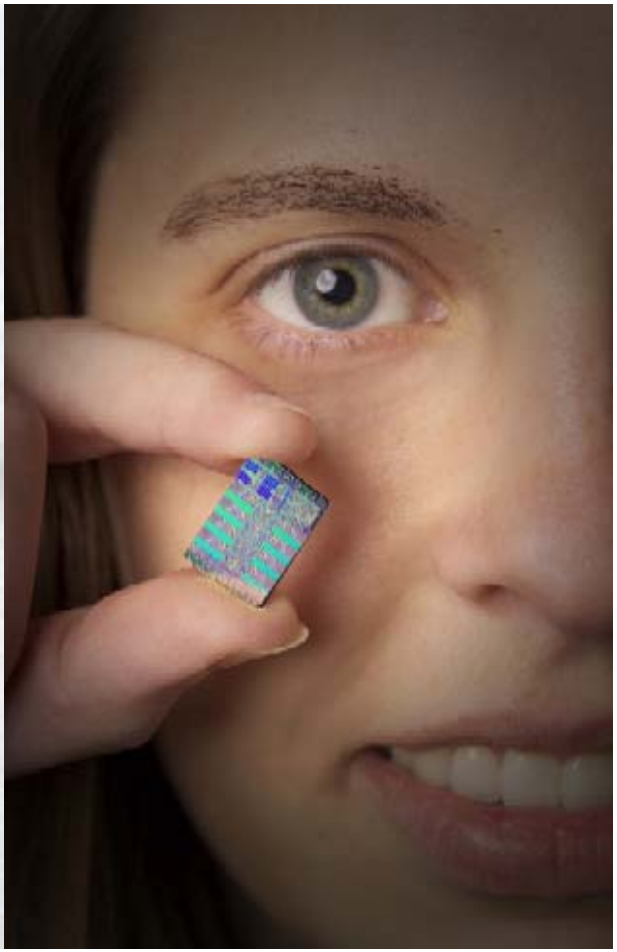




10 sec/step on 2048 BG/L
25 sec/step on 1400 Xeon Cluster at LLNL.

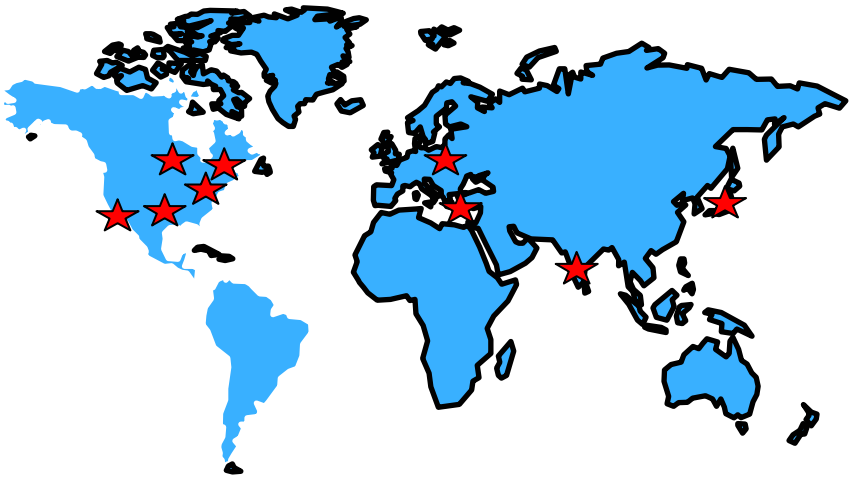
Cell





Cell History

- IBM, SCEI/Sony, Toshiba Alliance formed in 2000
- Design Center opened in March 2001
  - Based in Austin, Texas
- Single CellBE operational Spring 2004
- 2-way SMP operational Summer 2004
- February 7, 2005: First technical disclosures
- November 9, 2005: Open Source SDK Published



SONY COMPUTER ENTERTAINMENT

SONY

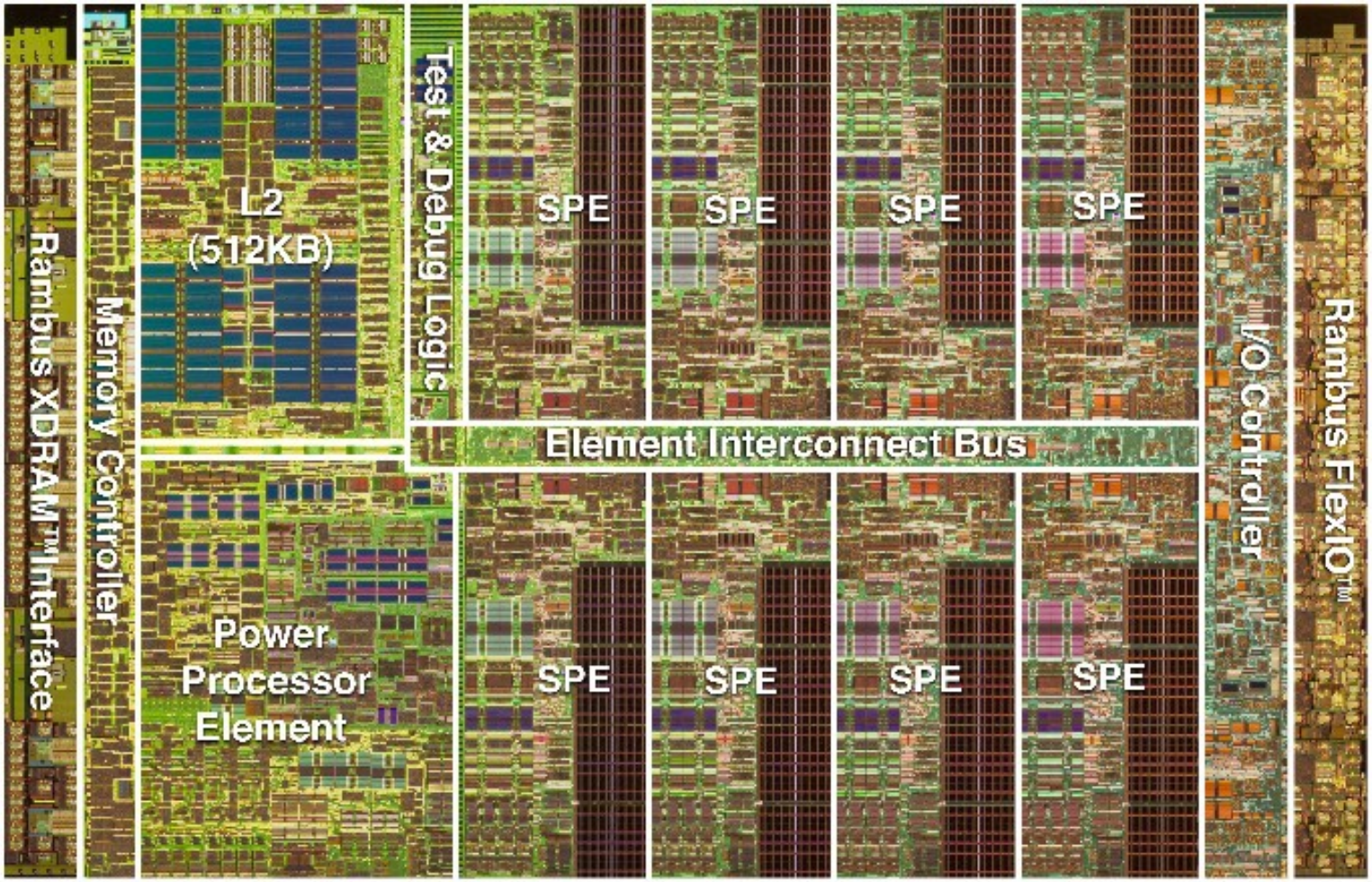
TOSHIBA

IBM

Cell Highlights

- Supercomputer on a chip
- Multi-core microprocessor (9 cores)
- 3.2 GHz clock frequency
- 10x performance for many applications
- Initial Application – Sony Group PS3

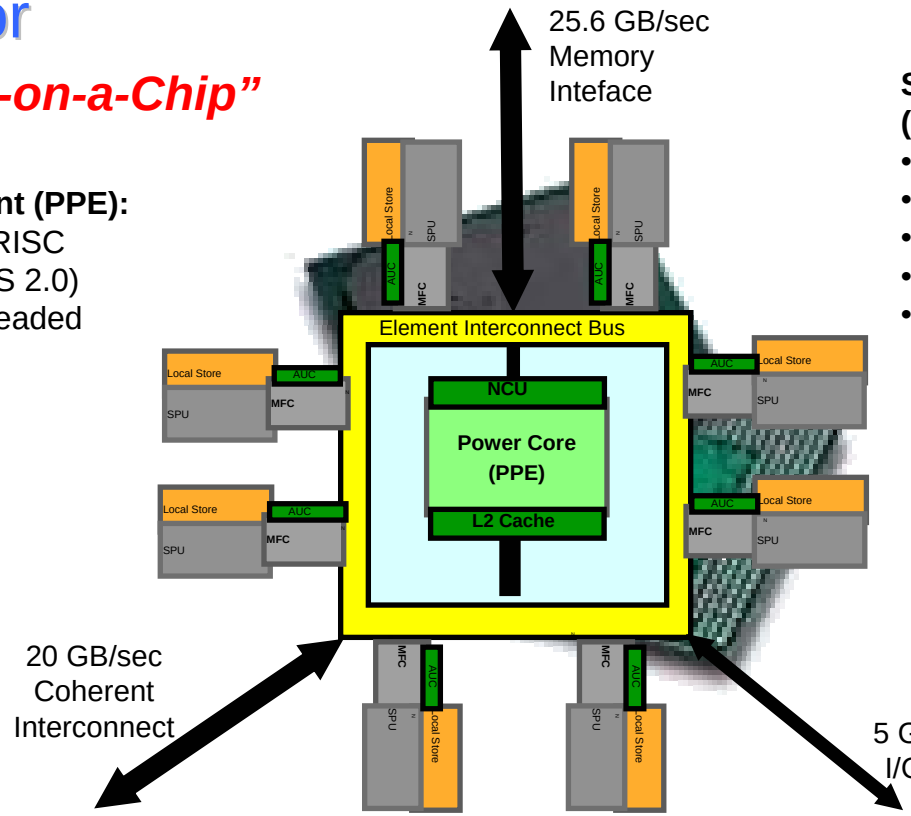
Cell Broadband Engine – 235mm²



Cell Processor

"Supercomputer-on-a-Chip"

Power Processor Element (PPE):
- General Purpose, 64-bit RISC Processor (PowerPC AS 2.0)
- 2-Way Hardware Multithreaded
- L1 : 32KB I ; 32KB D
- L2 : 512KB
- Coherent load/store
- VMX
- 3.2 GHz



Synergistic Processor Elements (SPE):
- 8 per chip
- 128-bit wide SIMD Units
- Integer *and* Floating Point capable
- 256KB Local Store
- Up to 25.6 GF/s per SPE --- 200GF/s total *

Internal Interconnect:
- Coherent ring structure 96B/cycle
- 300+ GB/s total internal interconnect bandwidth
- DMA control to/from SPEs supports >100 outstanding memory requests

External Interconnects:
- 25.6 GB/sec BW memory interface
- 2 Configurable I/O Interfaces
  - Coherent interface (SMP)
  - Normal I/O interface (I/O & Graphics)
  - Total BW configurable between interfaces
  - Up to 35 GB/s out
  - Up to 25 GB/s in

Memory Management & Mapping
- SPE Local Store aliased into PPE system memory
- MFC/MMU controls SPE DMA accesses
  - Compatible with PowerPC Virtual Memory architecture
  - S/W controllable from PPE MMIO
- Hardware or Software TLB management
- SPE DMA access protected by MFC/MMU

Cell BE Processor Initial Application Areas

- **Cell excels at processing of rich media content in the context of broad connectivity**
  - Digital content creation (games and movies)
  - Game playing and game serving
  - Distribution of dynamic, media rich content
  - Imaging and image processing
  - Image analysis (e.g. video surveillance)
  - Next-generation physics-based visualization
  - Video conferencing
  - Streaming applications (codecs etc.)
  - Physical simulation & science
- **Cell is an excellent match for any applications that require:**
  - Parallel processing
  - Real time processing
  - Graphics content creation or rendering
  - Pattern matching
  - High-performance SIMD capabilities

Cell Alpha Software Development Environment

Distributed on IBM alphaworks & Barcelona Super Computer sites Nov 9th

Development Environment

Documentation

Code Dev Tools

Debug Tools

Performance Tools

Miscellaneous Tools

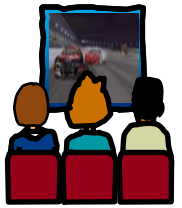
Samples
Workloads
Demos

SPE Optimized Function LIbraries

SPE Management Lib

Linux PPC64 with CellBE Extensions

System Level Simulator

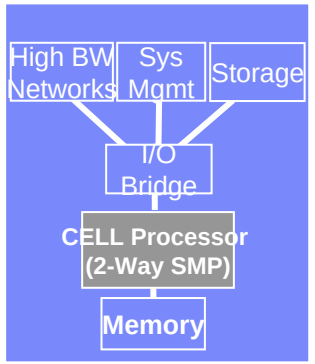


Execution Environment

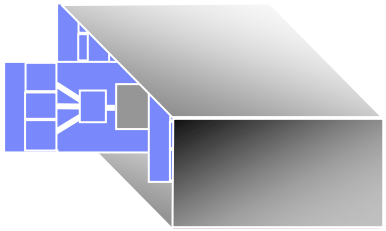
Standards: Language extensions
SPU ABI, Linux Reference ABI

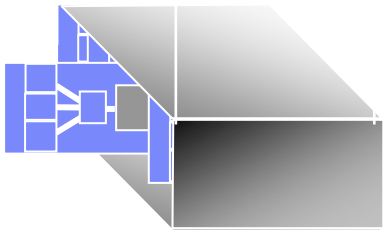
Cell Based Blade

- First Prototype “Powered On”
- 16 Tera-flops in a rack (est.)
  - (equals 1 Peta-flop in 64 racks)
- Optimized for Digital Content Creation, including
  - Computer entertainment
  - Movies
  - Real-time rendering
  - Physics simulation



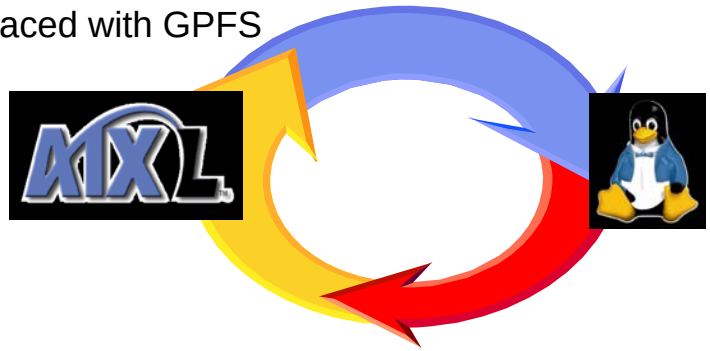
16 TFlop rack



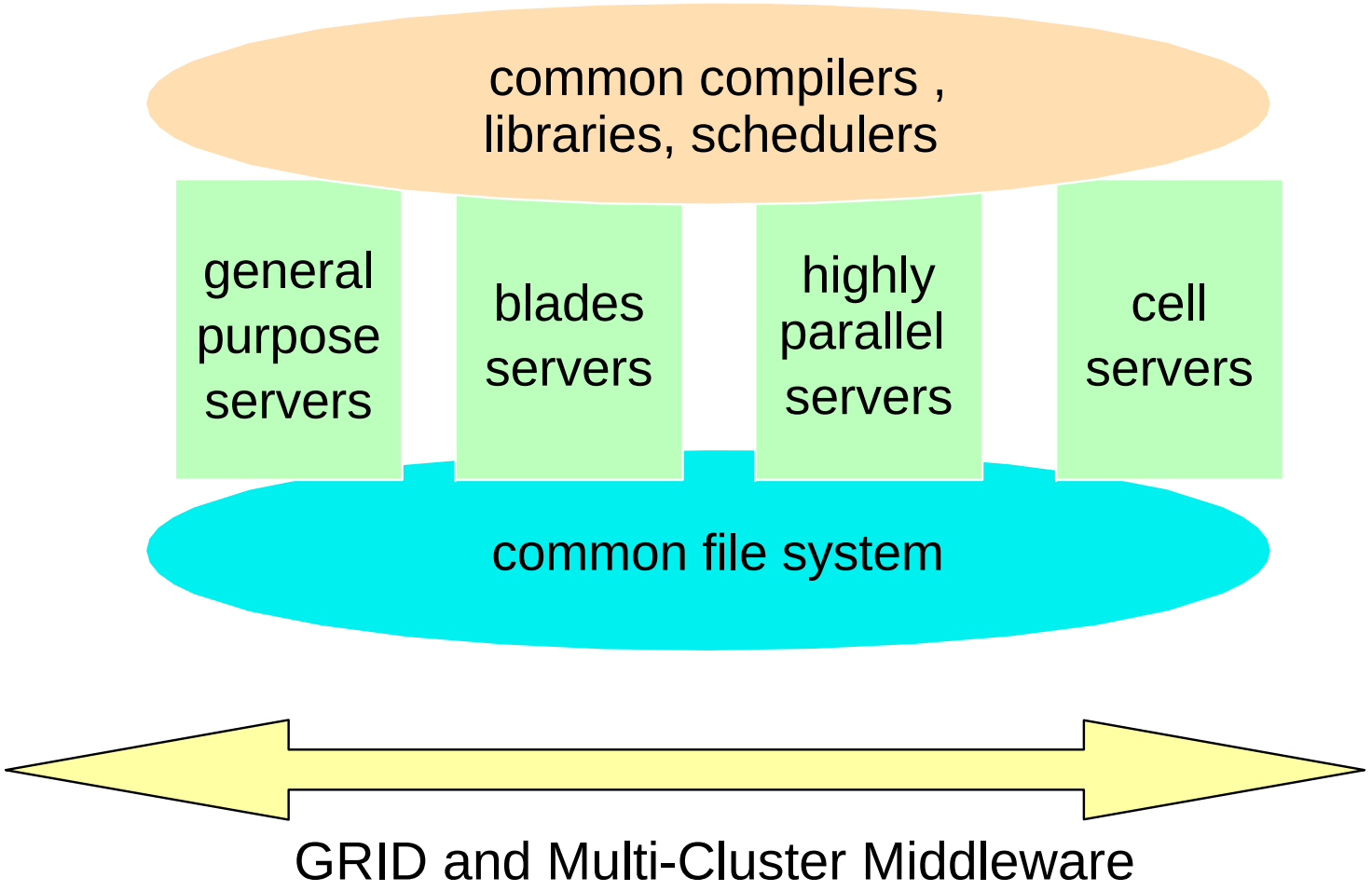


IBM common Software solutions for POWER

- Same Front end tools for POWER/POWERPC:
  - XLF/C/C++, ESSL/pESSL, libmass : compilers, libraries
  - *>Parallel Environment: MPI library and environment*
  - LL: job scheduler within and across AIX/Linux clusters
  - *> introduction of new features in checkpoint/restart and job migration*
- Same system tools for all platforms:
  - CSM :
    - single management for AIX/Linux clusters
  - GPFS :
    - interoperable parallel file system within and across AIX/Linux clusters
    - *> possibilty to licence GPFS to any vendor or customer*
    - *> for Linux x86/IA/ppc architecture*
  - TSM/HSM or HPSS :
    - backup/archive/migrate solutions interfaced with GPFS



IBM integrated heterogeneous solutions



- Questions ?