Supercomputing 2004 - Status und Trends (Conference Report)

Peter Wegner

- SC2004 conference
- Top500 List
- BG/L
- Moors Law, problems of recent architectures
- Solutions
- Interconnects
- Software
- Lattice QCD machines
  - DESY @SC2004
  - QCDOC
- Conclusions

Supercomputing 2004 - Status und Trends

Supercomputing 2004 - Status und Trends

Supercomputing = (Processor Performance) +
Network Speed + Storage Capacity

SCinet

Aggregate WAN transport delivered to the Industry and Research Exhibitors is expected to exceed 150 Gbps

StorCloud

1 PetaByte of randomly accessible storage
1 GigaByte per second backup bandwidth

Supercomputing 2004 - Status und Trends

Top 500 List (Linpack Benchmark, Jack Dongarra, Erich Strohmaier, Hans W. Meuer)

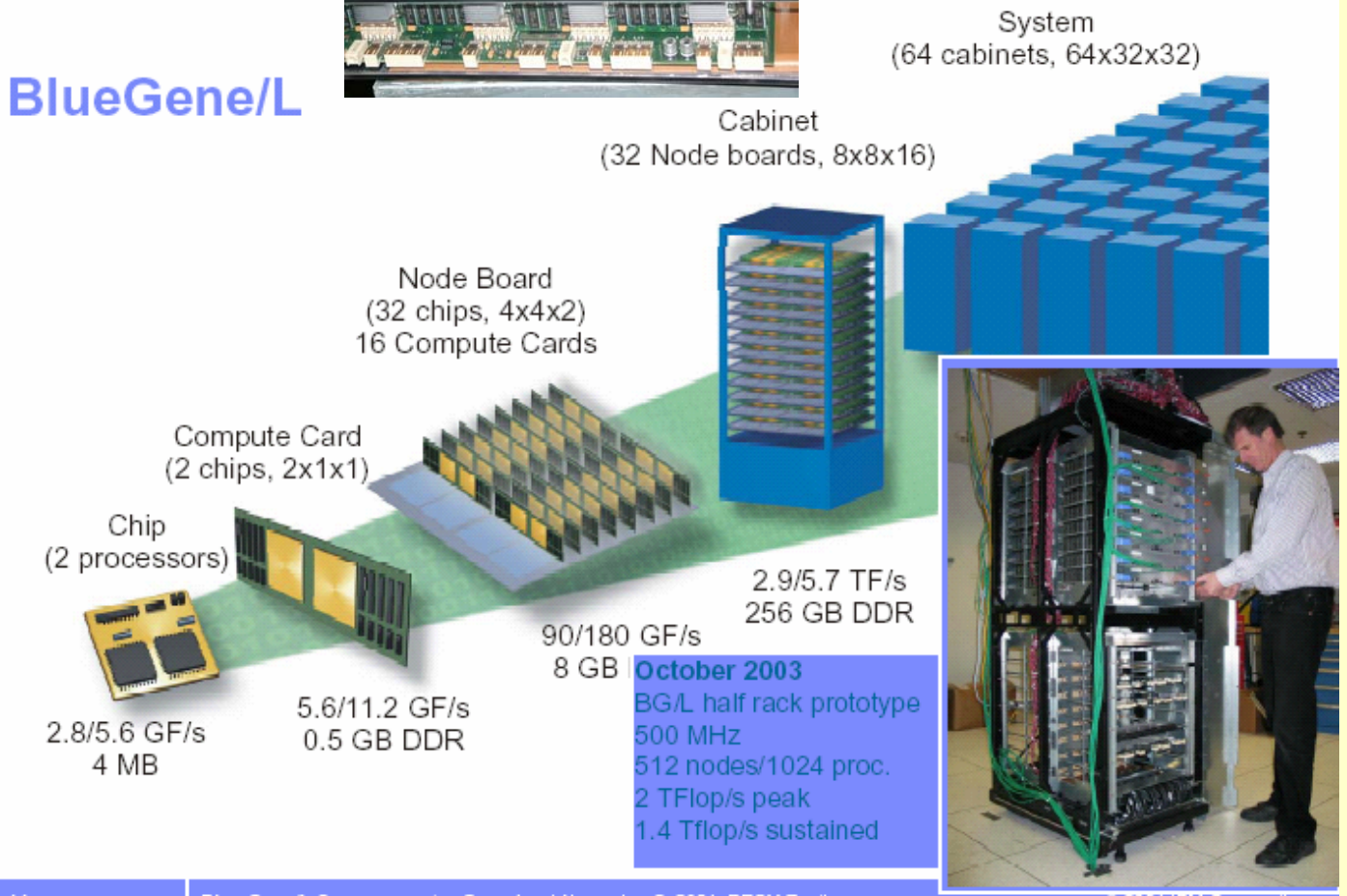
Rank	Computer/Processors/Manufactor	R_{max} R_{peak}	Installation Site Country/Year	Inst. type Installation Area
1	*BlueGene/L beta-System* **BlueGene/L DD2 beta-System (0.7 GHz PowerPC 440)** / 32768 IBM	70720 91750	United States/2004	**Research**
2	*Columbia* **SGI Altix 1.5 GHz, Voltaire Infiniband** / 10160 SGI	51870 60960	**NASA/Ames Research Center/NAS** United States/2004	**Research**
3	**Earth-Simulator** / 5120 NEC	35860 40960	**The Earth Simulator Center** Japan/2002	**Research**
4	*MareNostrum* **eServer BladeCenter JS20 (PowerPC970 2.2 GHz), Myrinet** / 3564 IBM	20530 31363	**Barcelona Supercomputer Center** Spain/2004	**Research**
5	*Thunder* **Intel Itanium2 Tiger4 1.4GHz - Quadrics** / 4096 California Digital Corporation	19940 22938	**Lawrence Livermore National Laboratory** United States/2004	**Research**

http://www.top500.org

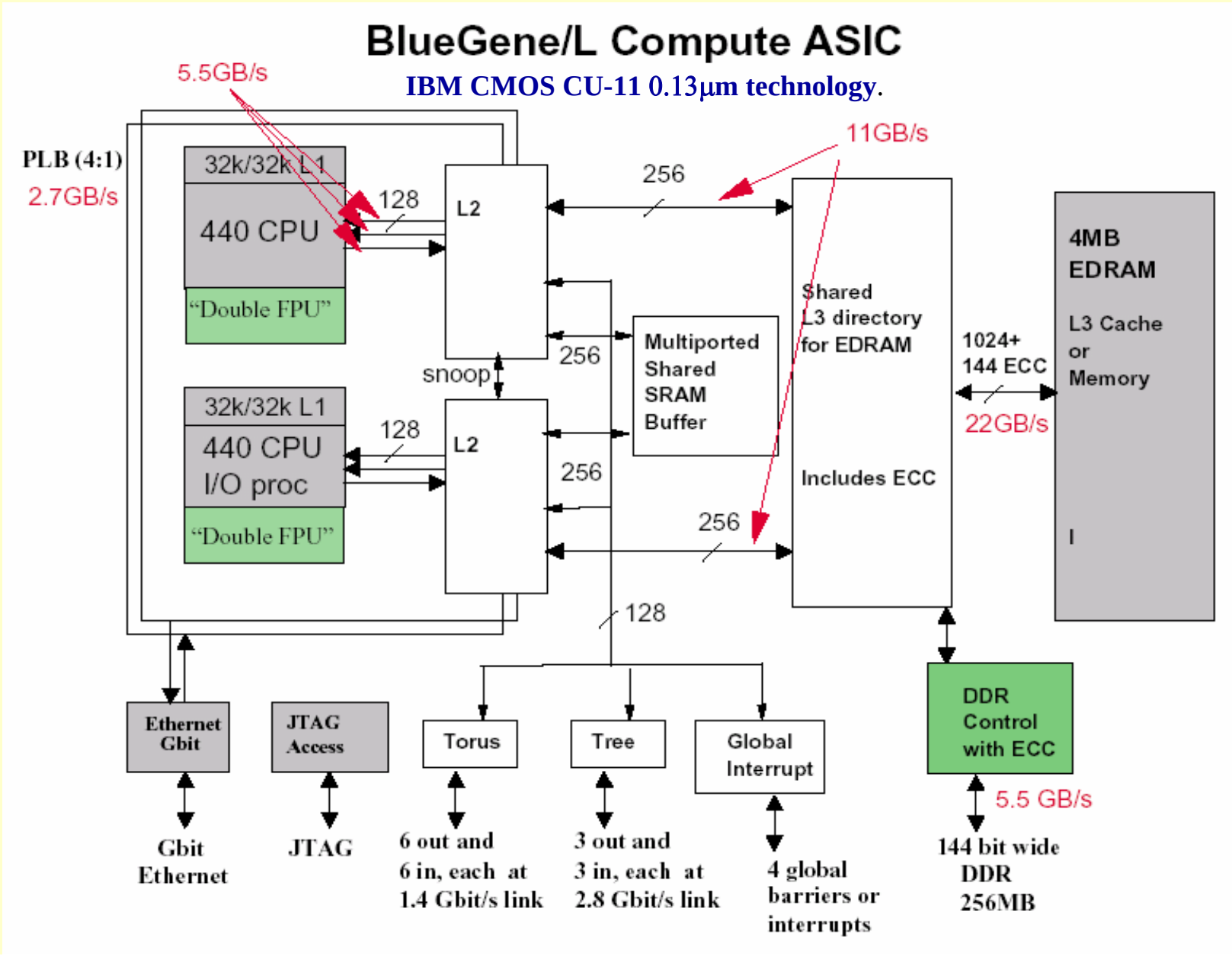
Supercomputing 2004 - Status und Trends

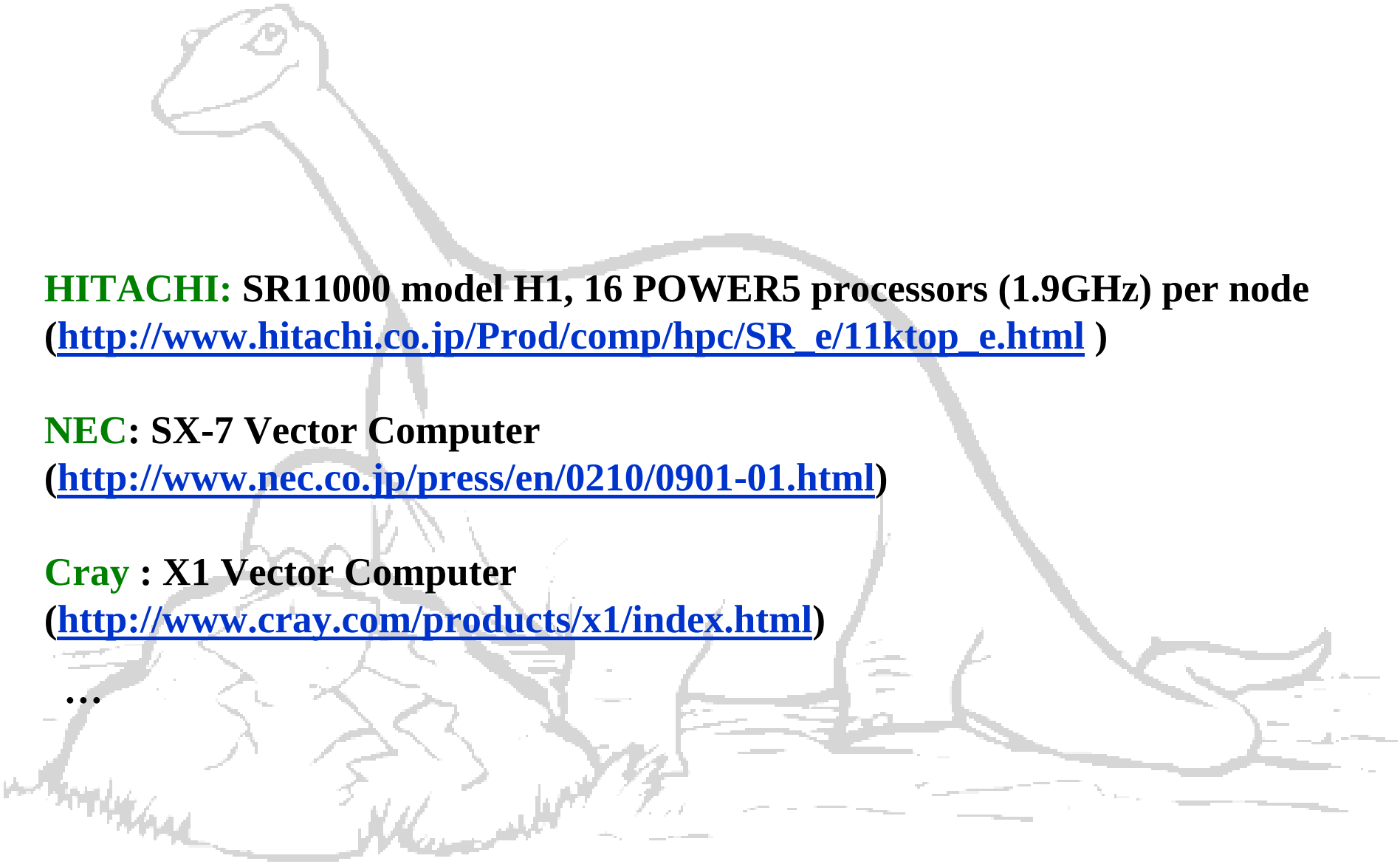
Blue Gene/L Architecture

Burkhard Steinmacher-Burow
IBM Watson / Böblingen



Supercomputing 2004 - Status und Trends





HITACHI: SR11000 model H1, 16 POWER5 processors (1.9GHz) per node (http://www.hitachi.co.jp/Prod/comp/hpc/SR_e/11ktop_e.html)

NEC: SX-7 Vector Computer (http://www.nec.co.jp/press/en/0210/0901-01.html)

Cray : X1 Vector Computer (http://www.cray.com/products/x1/index.html)

…

Supercomputing 2004 - Status und Trends

Moors Law

Gordon Moore from INTEL predicted at the beginning of IC mass production in the 1970 that within every period of 18 month the complexity (number of transistors on one chip) will be doubled.

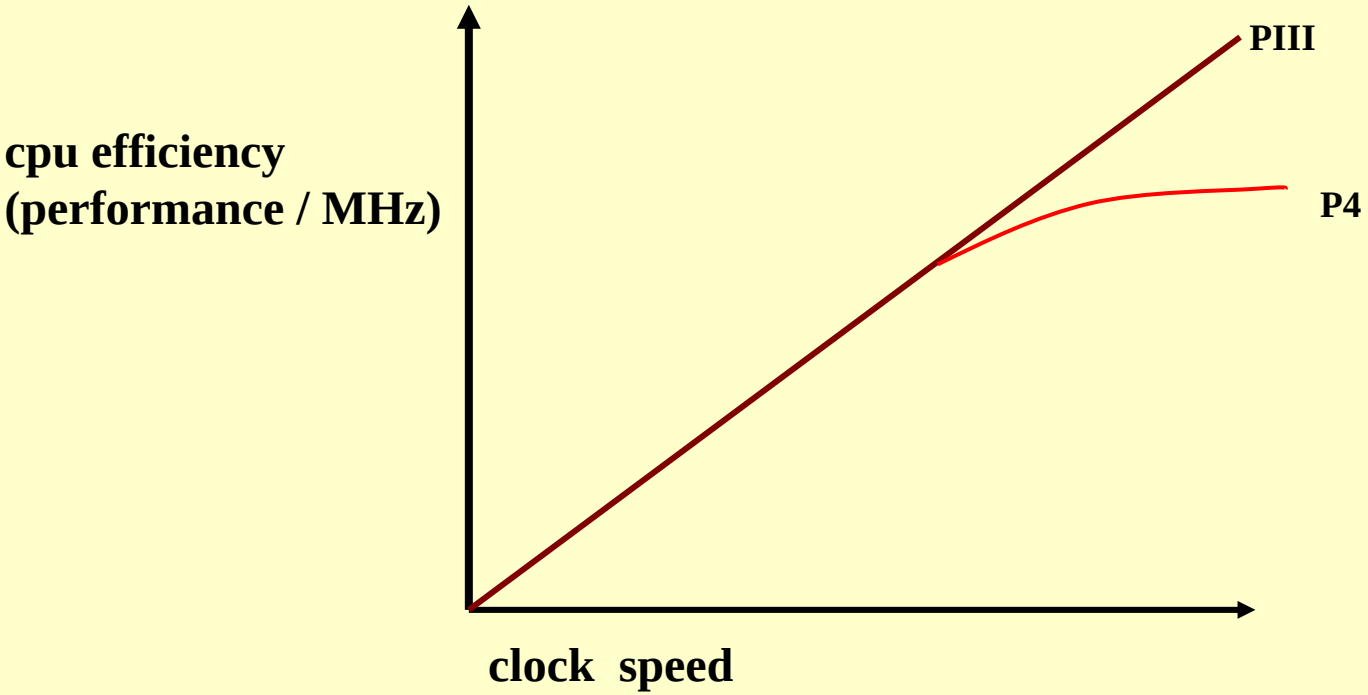
No prediction about power consumption, about performance, about efficiency …

Problems:
memory bandwidth, performance loss due to the increase in pipeline size, cooling …

Solutions:
multi-core CPUs, multithreaded architectures (or both), software pre-fetch (HITACHI), …

Supercomputing 2004 - Status und Trends

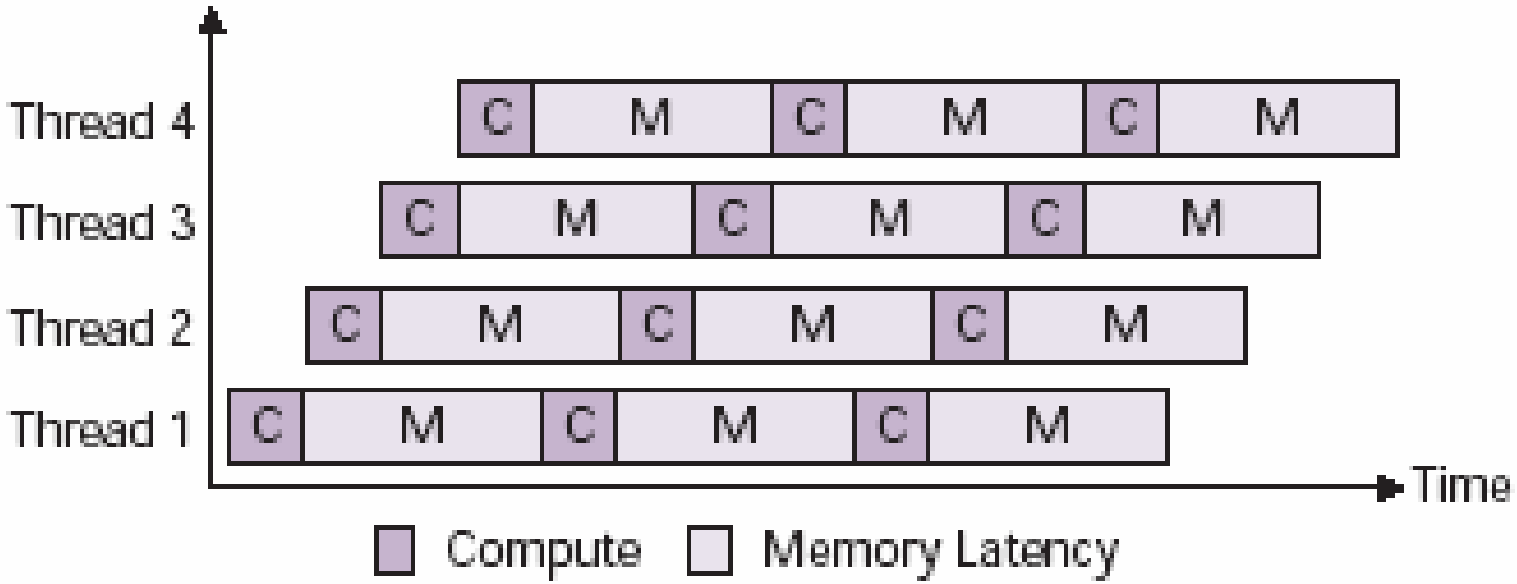
CPU efficiency at increasing clock speed



Marc Tremblay, SUN

Supercomputing 2004 - Status und Trends

Multi-core multithreaded Architectures

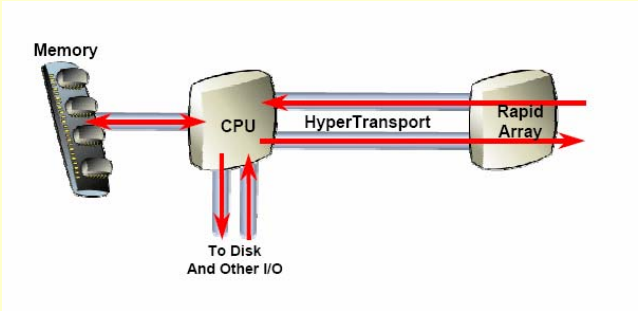


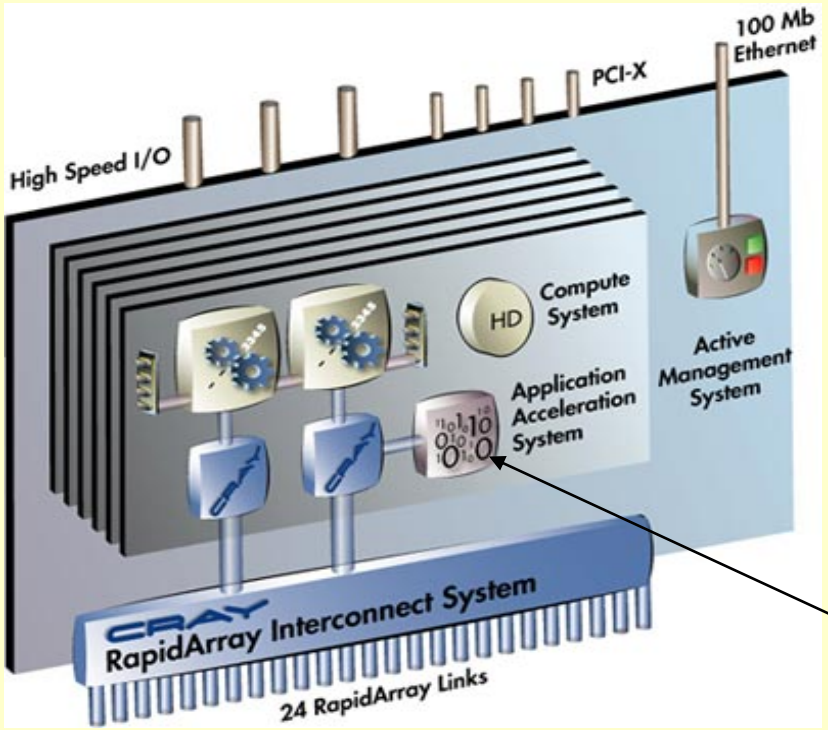
Marc Tremblay, SUN

Supercomputing 2004 - Status und Trends

Reconfigurable co-processors

Cray XD1:





Compute System

- 12 AMD Opteron 32/64 bit, x86 processors per "shelf"
- Packaged as six 2-way SMP's
- Linux
- Very densely packaged

RapidArray Interconnect System

- 12 communications processors
- 1 terabit per second switch fabric
- 30X Gigabit ethernet performance

Active Management System

- Dedicated processor
- High availability
- Single system management and control

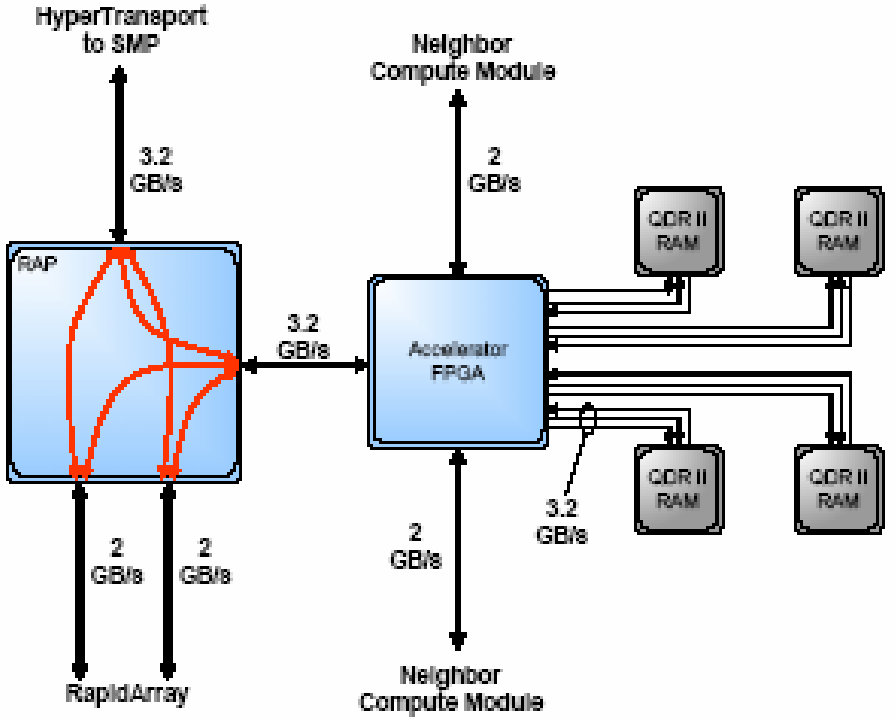
Application Acceleration System

- 6 FPGA co-processors

Supercomputing 2004 - Status und Trends

Reconfigurable co-processors

Cray XD1:

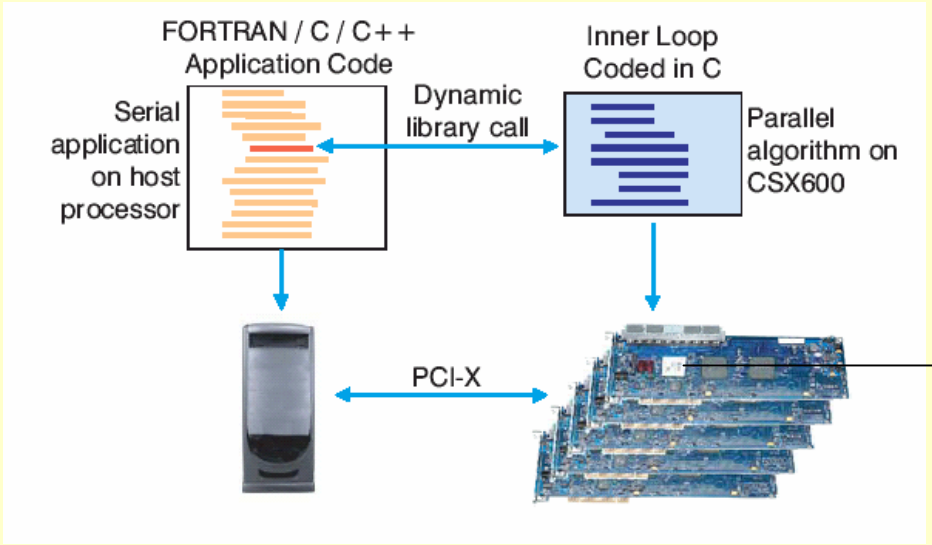


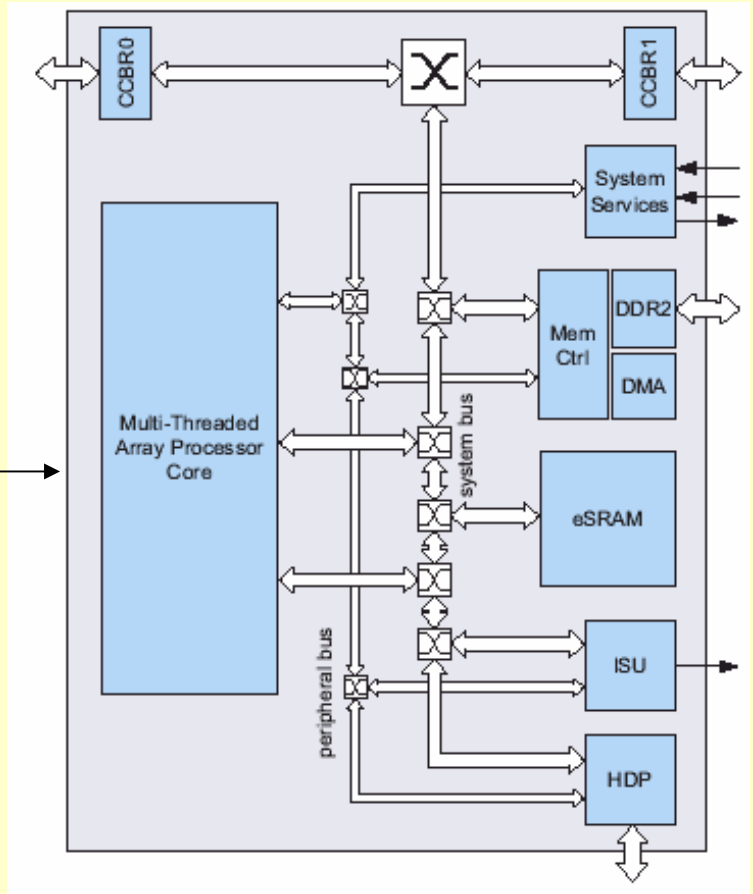
The RapidArray processor provides the interface for the FPGA to connect to the local Opteron processors as well as the RapidArray fabric. The interface's speed and performance match that of the HyperTransport links, but the link protocol is simplified to reduce logic requirements in the FPGA.

Supercomputing 2004 - Status und Trends

Reconfigurable co-processors

Clearspeed:

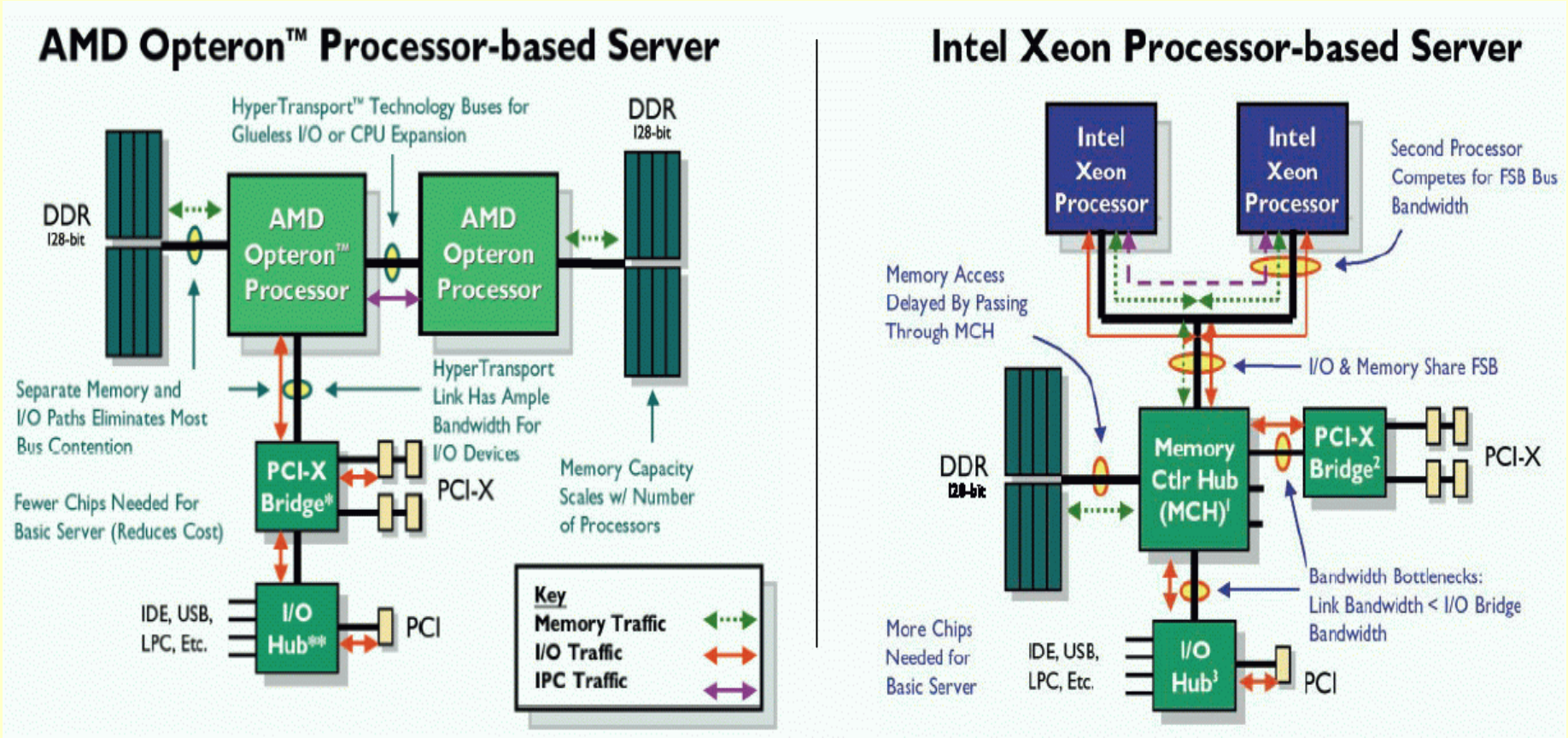




Linux Networks ?

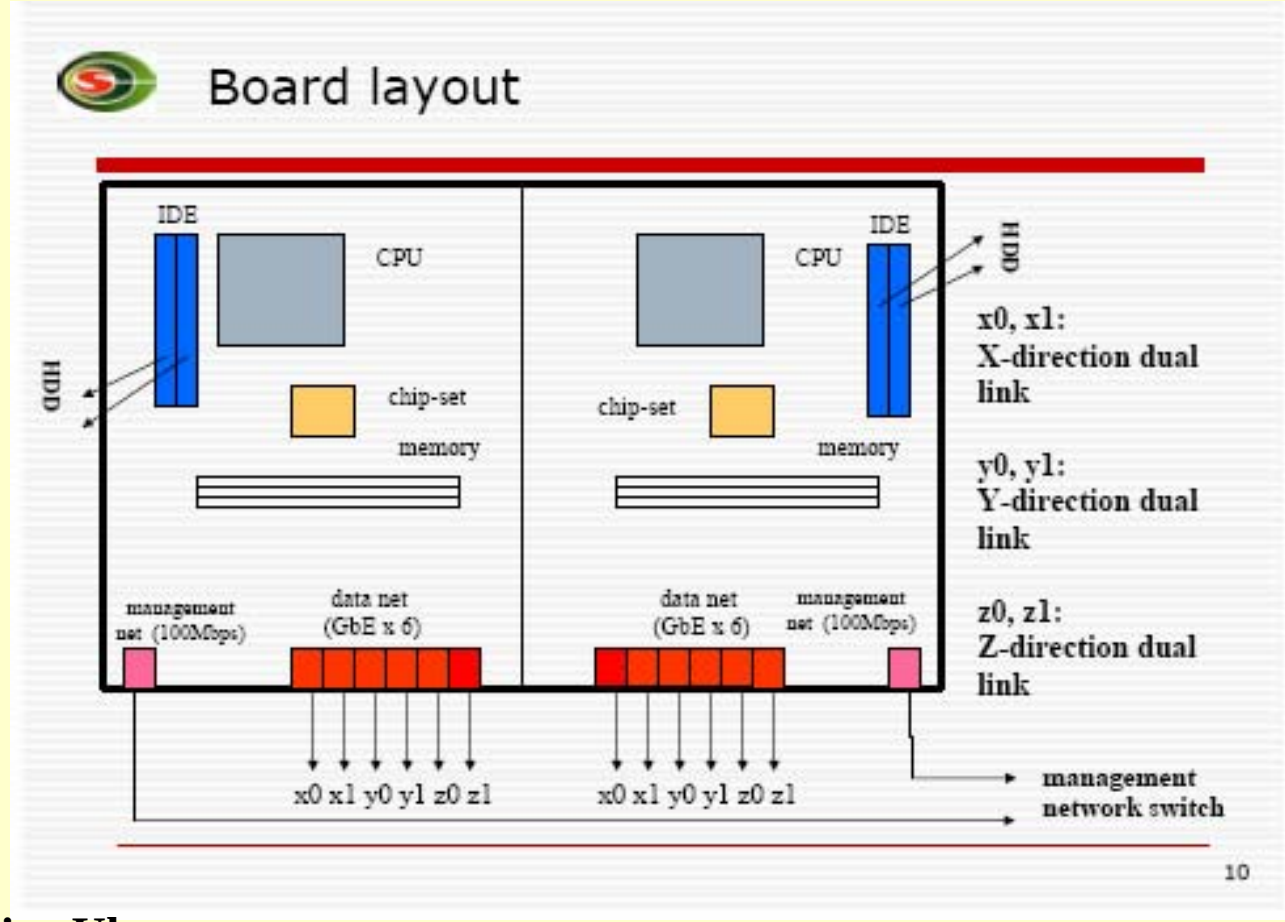
Supercomputing 2004 - Status und Trends

X86 architectures, bottlenecks



Supercomputing 2004 - Status und Trends

X86 architectures, dual chipset XEON Motherboards



Akira Ukawa, ***University of Tsukuba***

German-Japanese Workshop
27 November, 2004

A Similar Architecture using Infiniband interconnect was presented on SC2004 by James Ballew from Raytheon

Supercomputing 2004 - Status und Trends

Interconnects

Infiniband – absolute winner, but in the a good chance future for low latency
10GBit Ethernet
(10Gig will get to copper in a year or two)

PathScale – HTX-Adaptor - Hypertransport connector
reduced infiniband protocol, accepted by Infiniband switches,
1.5 µs MPI latency,
FPGA ready, ASIC – Q2/2005

Myrinet, QsNET, Dolphin Interconnect (SCALI)

Supercomputing 2004 - Status und Trends

Software

PathScale EKOPATH Compiler Suite for AMD64 and EM64T

PGI Compiler

Intel Compiler

Absoft

TotalView (Etnus)

Allinea DDT (Distributed Debugging Tool)

PBS Pro

SUN Gridengine (SUN, RAYTHEON)

LSF

Supercomputing 2004 - Status und Trends

NIC – FZJ, DESY

NIC/FZJ JUMP Supercomputer
(IBM Regatta, 8.9 TFlops, Rank 30 of Top500 list)
Graphical User Interface for Job Monitoring

NIC/DESY APE massive parallel computers
(550 GFlops/3 TFlops)

Supercomputing 2004 - Status und Trends

Lattice QCD, NIC/DESY Zeuthen poster : APEmille

APE Teraflop Computers for Simulations of Elementary Particle Physics

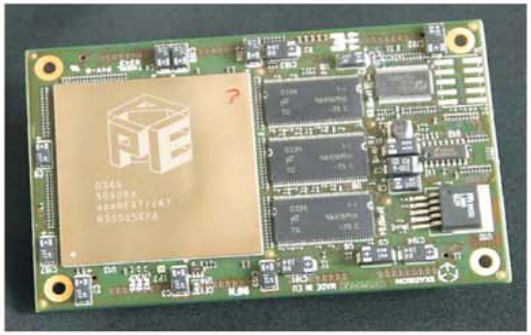
apeNEXT : Development for the Future

apeNEXT Design

New features:

- asynchronous → SPMD architecture
- prefetch queues for local and remote data
- 64bit double precision arithmetics

Processor	
clock	200 MHz
peak performance	1.6 Gflops
sustained performance	30-75% (typical applications)
arithmetics	(a*b+c) complex 64 bit
technology	1 custom chip, 0.18 µ
Network	
topology	3 dimensional, nearest neigthbour
technology	LVDS
bandwidth	200 Mbytes / sec
Price	0.5 Euro / Mflops (peak)



possible apeNEXT Installation at NIC / DESY Zeuthen

Parameters of a possible installation of 20 apeNEXT racks:	
Performance	16 Tflops (peak)
Memory	2.5 - 10 TByte
Power consumption	100 kW
Foodprint	20 m^2

by APE Collaboration

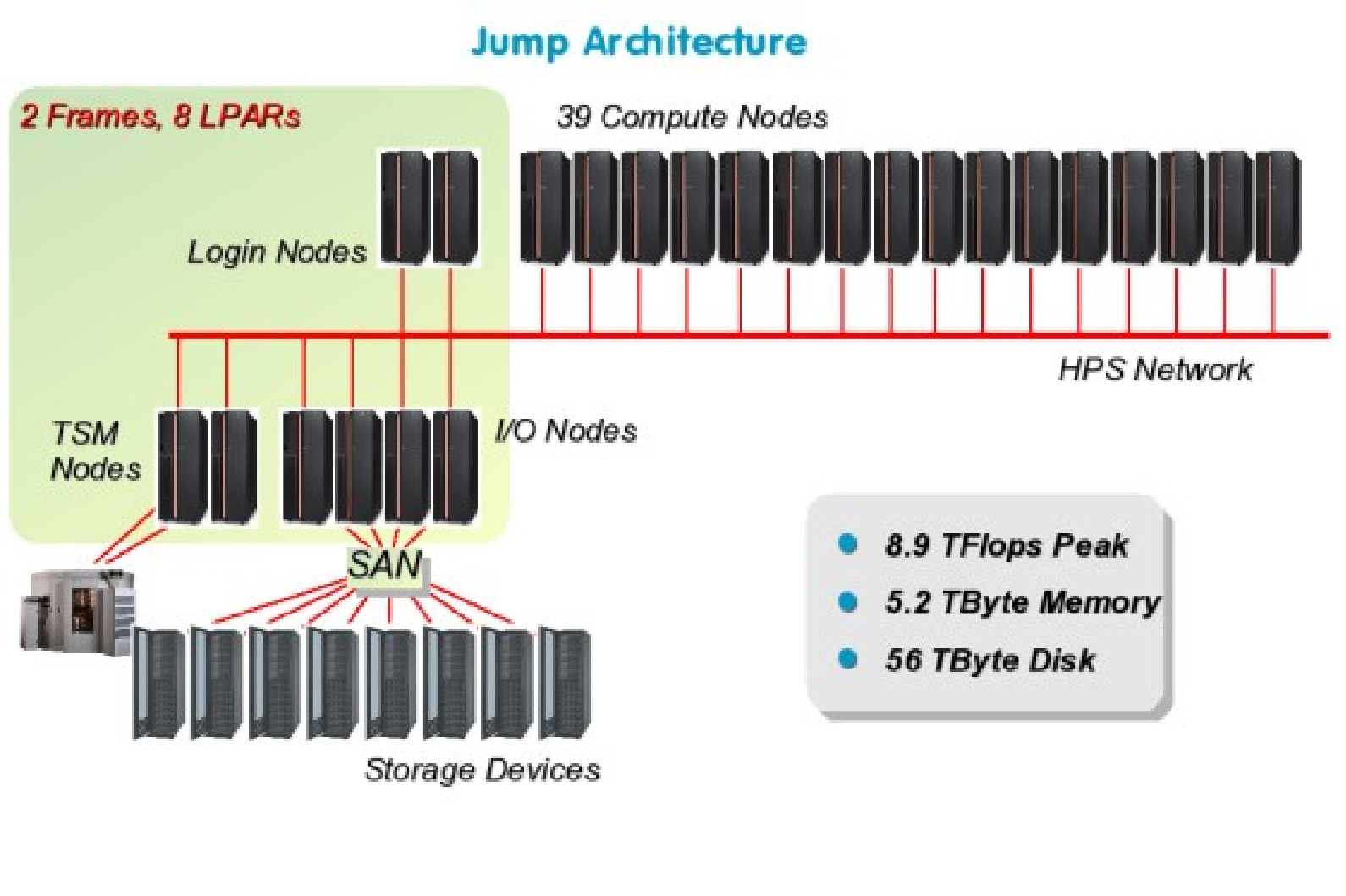
INFN Istituto Nazionale di Fisica Nucleare

DESY Zeuthen

UNIVERSITE PARIS SUD IX

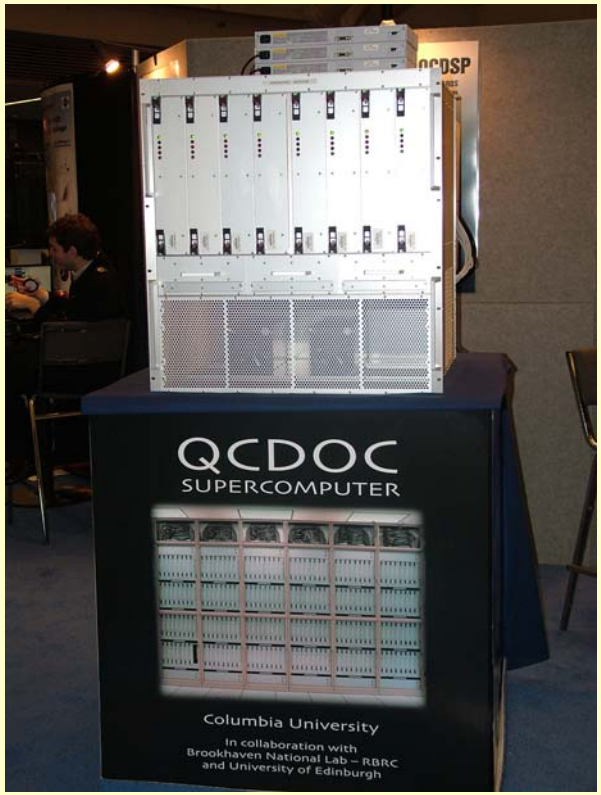
Supercomputing 2004 - Status und Trends

NIC/DESY FZJ



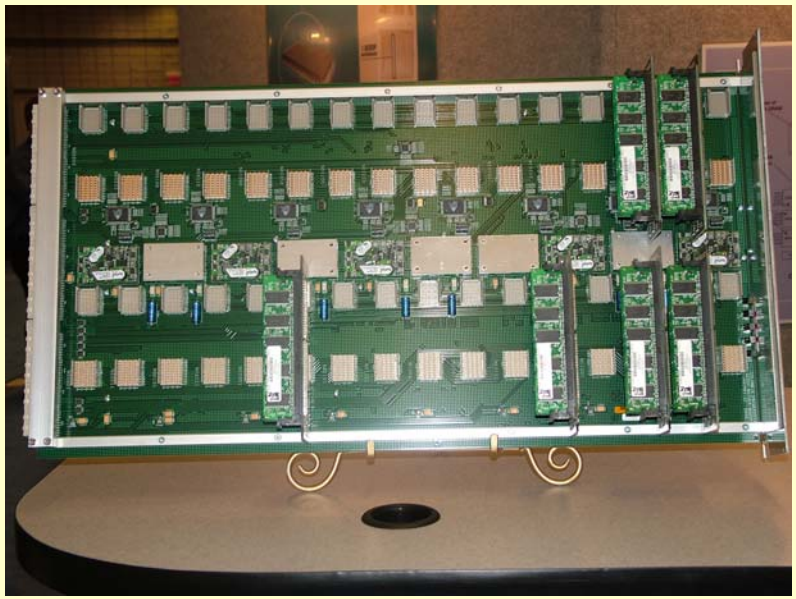
Supercomputing 2004 - Status und Trends

Lattice QCD



QCDOC

Columbia University
RIKEN/Brookhaven National Laboratory (BNL)
UKQCD,
IBM



Supercomputing 2004 - Status und Trends

Columbia University
RIKEN/Brookhaven National Laboratory (BNL)
UKQCD, IBM

MIMD machine with distributed memory system-on-a-chip design

(QCDOC = QCD on a chip)Technology:

IBM SA27E = CMOS 7SF = 0.18 mm lithography process

ASIC combines existing IBM components and QCD-specific, custom-designed logic:

500 MHz PowerPC 440 processor core with 64-bit, 1 GFlops FPU

4 MB on-chip memory (embedded DRAM)

Nearest-neighbor serial communications unit (SCU)

6-dimensional communication network (4-D Physics, 2-D partitioning)

Silicon chip of about 11 mm square, consuming 2 W

Conclusions

•A wide range of supercomputing architectures

- high end supercomputers – Hitachi SR1100, Cray X1, NEC SX7/8 …

-commodity clusters as HPC solution – Linux Networks, IBM, DELL, SGI, SUN, HP, Rack Server …

- more than 20 systems with more than 1500 dual CPU nodes in Top500 list - Blade servers

- SMP systems – IBM (Power5), HP(Itanium), SUN(Sparc, Opteron), SGI (Itanium)

- special designs with integrated commodity CPUs – Cray XD1, Linux Networks ?

-special purpose systems – BG/L, QCDOC, apeNEXT, GRAPE6 (Japan)

•No significant systems performance increase using "classical" technologies, therefore

Direct CPU-Memory connection (AMD)

Multi-core CPUs (IBM, Intel, AMD)

Multi-core – multithreaded CPUs (SUN, MTA)

•Increasing fraction of Opteron HPC systems

•Infiniband as a special High Performance Link beats products like Myrinet and QsNet, in future Gbit10 Ethernet ?

•All commercial compiler vendors are offering optimized compilers for 64-Bit x86 systems (Opteron, EM64T)