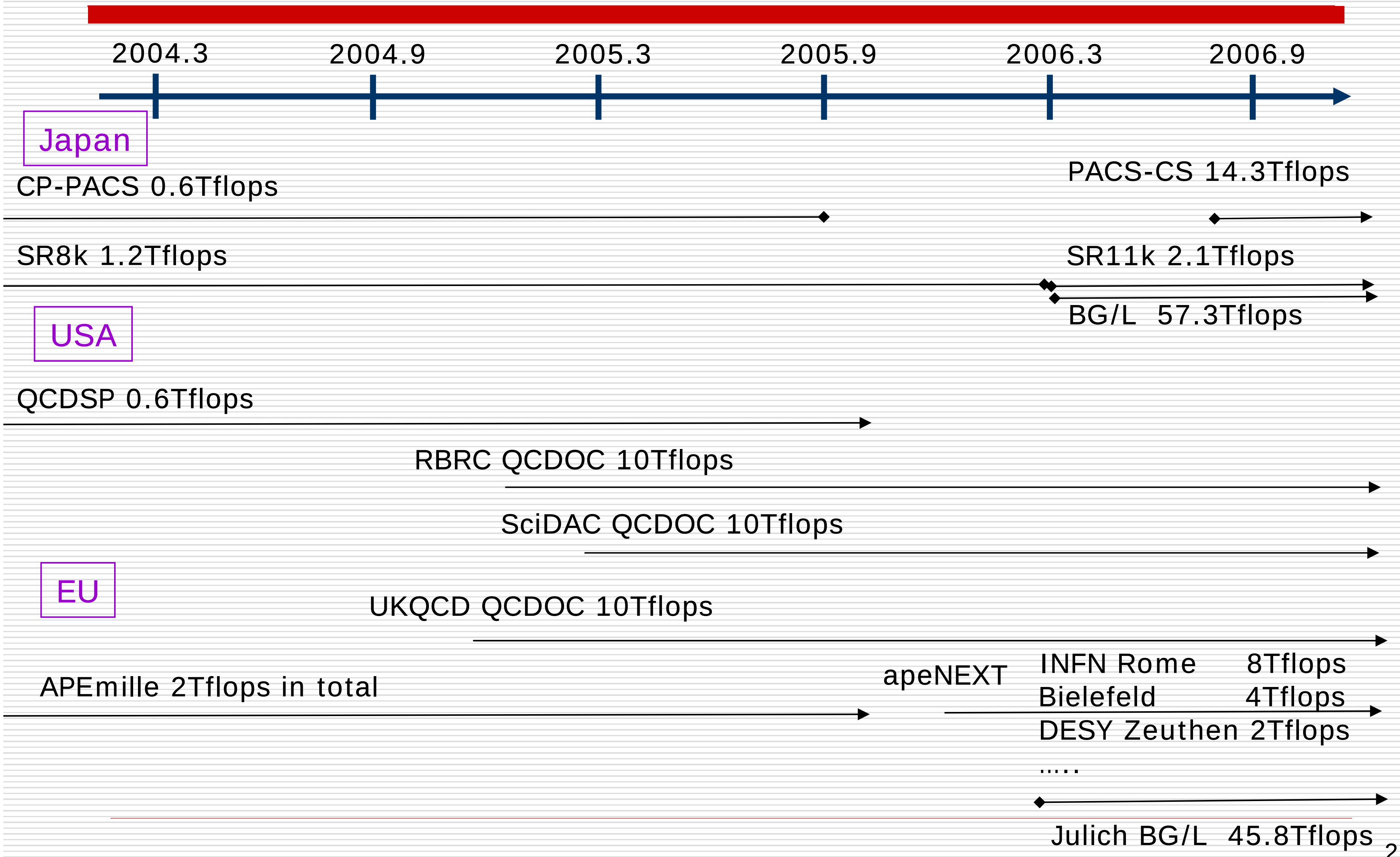
Looking back and looking ahead

- Lattice QCD in an international setting -

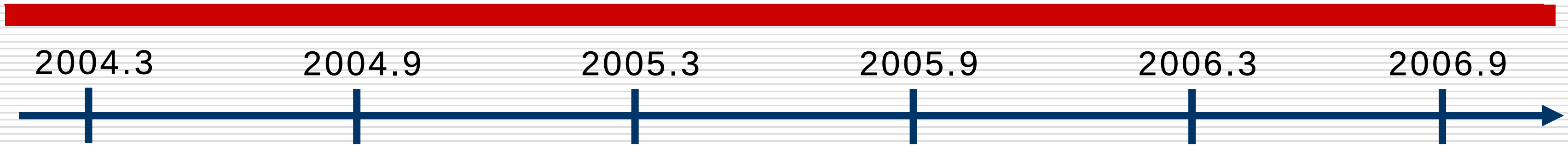
Akira Ukawa
Center for Computational Sciences
University of Tsukuba

- Looking back
- Status of ILDG
- Status of PACS-CS Development
- Looking ahead

Looking back: machines



Looking back: algorithms

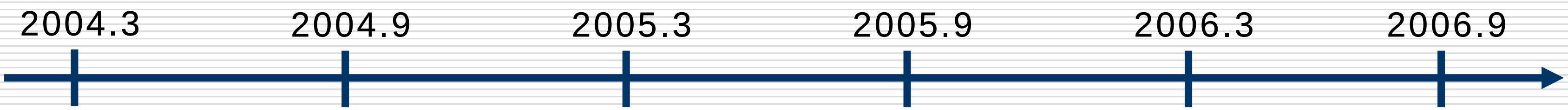


- Odd number of flavors has become standard:
  - Multi-boson methods
  - Polynomial HMC
  - Rational HMC

- Acceleration tricks that seem to work
  - Hasenbusch preconditioning
  - Domain decomposition
  - Multi-time step evolution

promising 5-10 times speedup for Wilson-type quark action

Looking back: shift toward chiral regime



- Lighter quarks with conventional actions:
  - Staggered quarks have been doing it since long time — *MILC Staggered Program since around 2000*
  - Berlin wall for Wilson-type quarks may be down now… — *Luescher et al 2005*
  - Light quarks with twisted mass — *Twisted Mass Collaboration*

- Realistic simulations with chiral actions — *UKQCD/RBC 2+1 domain wall program 24^3x64x16*
  - 2+1 runs with domain wall
  - Serious attempt toward dynamical overlap

- Revolutionary period
  - "once every 10 years" event
  - previous was 1996 with CP-PACS/QCDSP

- Perhaps more exciting since
  - Finally Nf=2+1
  - Finally chiral

    No excuse necessary to experiment/theory colleagues outside lattice QCD
  - Burst of activities is worldwide

Status of ILDG

International Lattice Data Grid

- Proposed by UKQCD in June 2002;
  International Research Infrastructure for
  - Sharing lattice data
  - Physics collaborations
- First workshop in Edinburgh (19-20 Dec. 2002)
  - UK, Germany, Japan, USA
  - Metadata and Middleware Working groups set up
- Since then,
  - 6 bi-annual workshops using Internet
  - Annual report at Lattice conferences
- Official opening of ILDG in June 2006
  - Opening of web sites in UK/USA/Japan/Germany
  - Search and download of public configurations

Technical preparations

- QCDml v1.1
  - Standard for configuration file description
  - Adopted in August 2004/some updates
- Binary file format v1.0
  - Standard for configuration file format
  - Adopted in May 2005
- Middleware architechture
  - ILDG stipulates only the interface; detailed implementation left to each country
  - Adopted in Dec. 2004/refined in May 2005

Tremendous amount of work done and being done by the members of the Metadata and Middleware Working Groups

Organizational aspects

- Particpating countries as of now
  - UK/USA/Germany/Japan
  - France/Italy showing interest
- ILDG Board
  - One representative from each country
  - Board chair with one year term
    - 2003 R. Kenway(UK)
    - 2004 A. Ukawa(Japan)
    - 2005 R. Brower(USA)
    - 2006 K. Jansen(Germany)
- Web page

  http://www.lqcd.org/ildg/

ILDG Data sharing policy 7 July 2004

In addition to the normal practice of sharing data within restricted groups for specific joint projects, collaborations that are generating substantial sets of gauge configurations should

- *mark up their data using the QCDML standard;*
- *adopt a policy to make their data generally available as soon as possible;*
- *announce on the ILDG web pages, at the time of production, their chosen action and parameter values, and when their configurations will be made generally available through ILDG.*

Japan Lattice Data Grid (JLDG)

- objectives
  - construct a QCD Data Grid for Japanese collaborations
  - construct a gateway to/from ILDG
- institutions
  - Tsukuba, KEK, Kyoto, Hiroshima,....
  - collaboration agreed and started in Nov. 2005
- previous activities
  - Lattice QCD Archive (LQA) @ CCS, Tsukuba
  - File Mirroring over Hepnet-J/sc

Lattice QCD Archive (LQA)

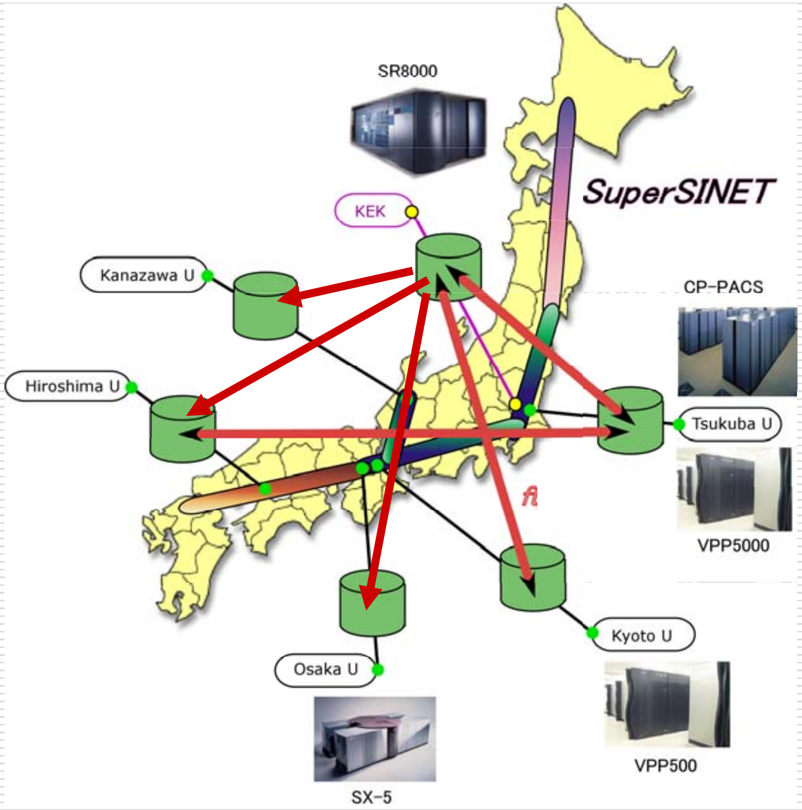
http://www.lqa.ccs.tsukuba.ac.jp/

- stores gauge configurations and makes them available to lattice community world-wide
  - CP-PACS Wilson-clover Nf=2 configs
  - CP-PACS/JLQCD Wilson-clover Nf=2+1 configs (soon)
- set up in Dec.2003 and maintained by CCS
- prototype implementation of ILDG architecture
  - MDC on Xindice, No RC, Interactive search
  - XML files written based on old QCDML draft
  - file format based on previous proposal
- will be dissolved and absorbed into JLDG

Hepnet-J/sc

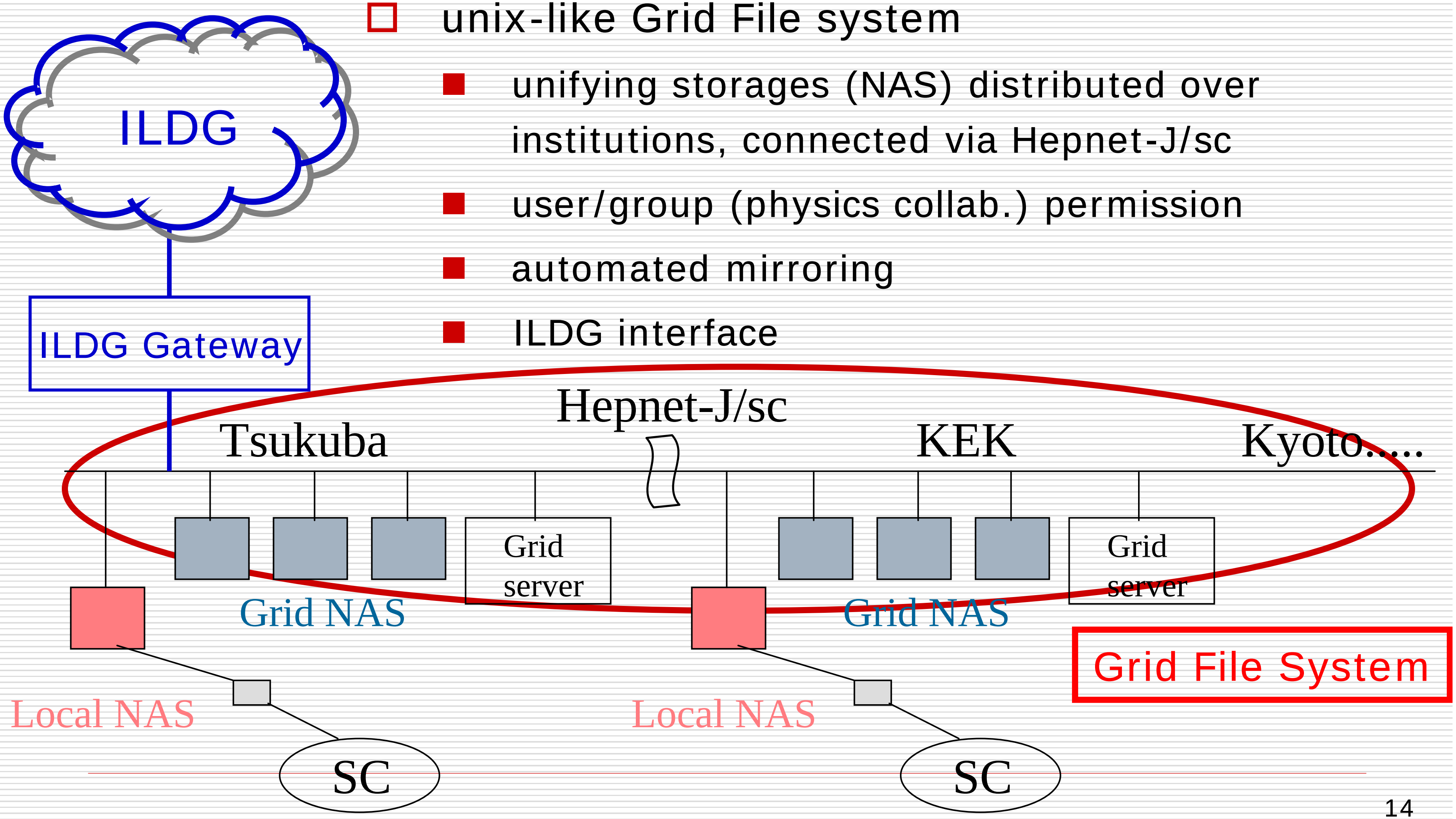
- Japanese domestic network for theoretical high energy physics
  - uses SuperSINET 1Gbps private networks (NII)
  - major LQCD sites in Japan are connected
  - file mirroring for Japanese collab's (60TB, 6 sites)

    maintained by hand
  - data distributed over many disks, because data size exceeds partition size



JLDG design

- unix-like Grid File system
  - unifying storages (NAS) distributed over institutions, connected via Hepnet-J/sc
  - user/group (physics collab.) permission
  - automated mirroring
  - ILDG interface

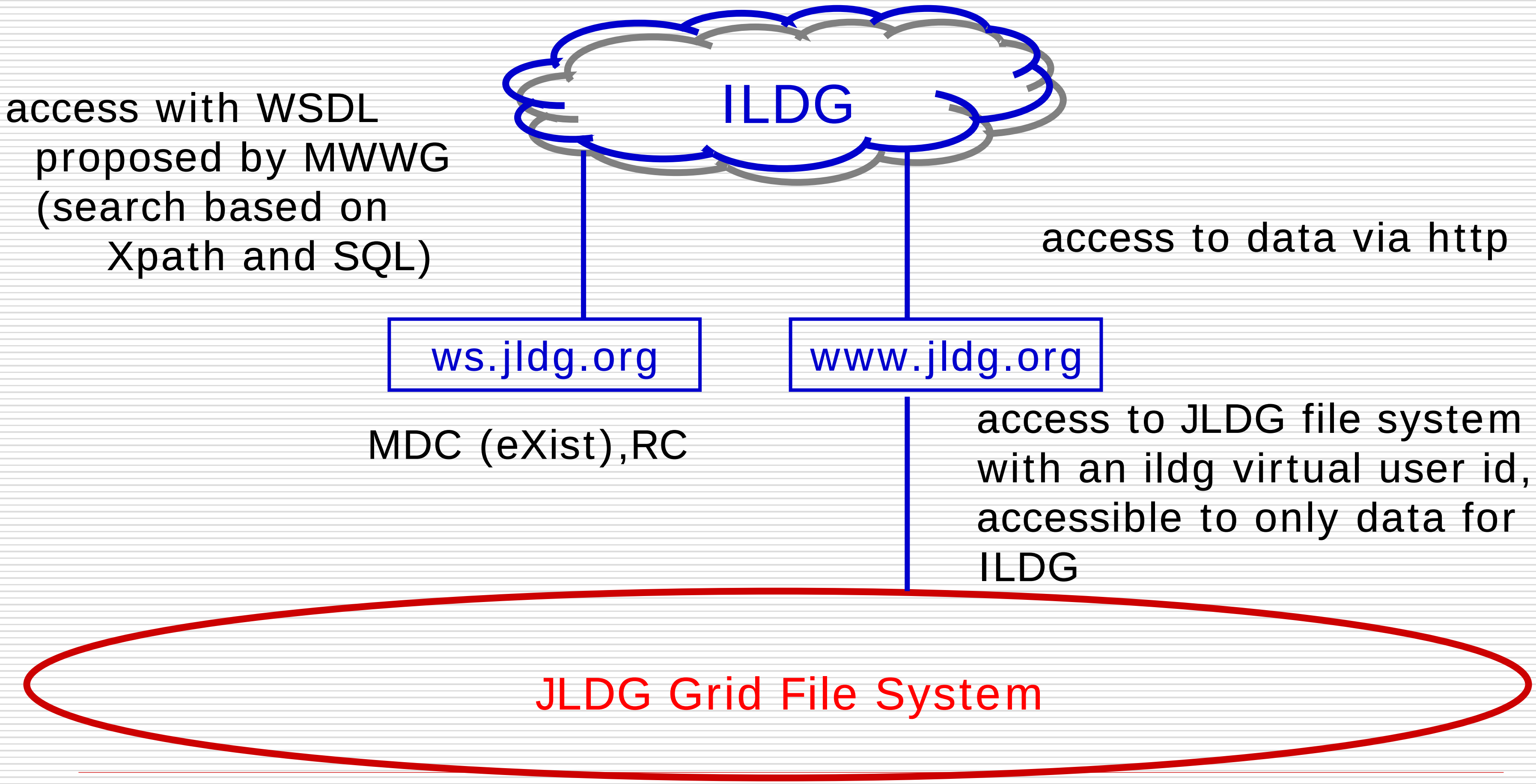


JLDG architecture : File System

- based on Gfarm (a national grid data farm software)
  - grid-wide file system: Gfarm enables us to share any file contents (configuration, propagator, etc...) between sites.
  - authentication: private CA handled by Globus toolkit coordinated with VOMS. File/directory permission enables us to protect unpublished data.
- usual unix software works on the Grid file system without any change.
  - by hooking libc
  - an advantage to develop ILDG gateway

JLDG Architecture: ILDG Gateway

- hardware and middleware



JLDG Status

- File system
  - test implementation @ Tsukuba and KEK
  - 1^{st} stage will be completed in March 2006 (no group permission)
  - 2^{nd} stage before June 2006
- Middleware
  - under development, will be completed in March
  - test and tuning will continue to May
- Data
  - ensemble and configuration markup completed, based on the latest MDWG proposal (QCDml1.3.0)
  - converting LQA data to ILDG format on-going.

Status of PACS-CS

Parallel Array Computer System for Computational Sciences

Our strategy

Essentially, an MPP with commodity components

Build a “semi-dedicated” cluster appropriate for lattice QCD (and a few other applications)

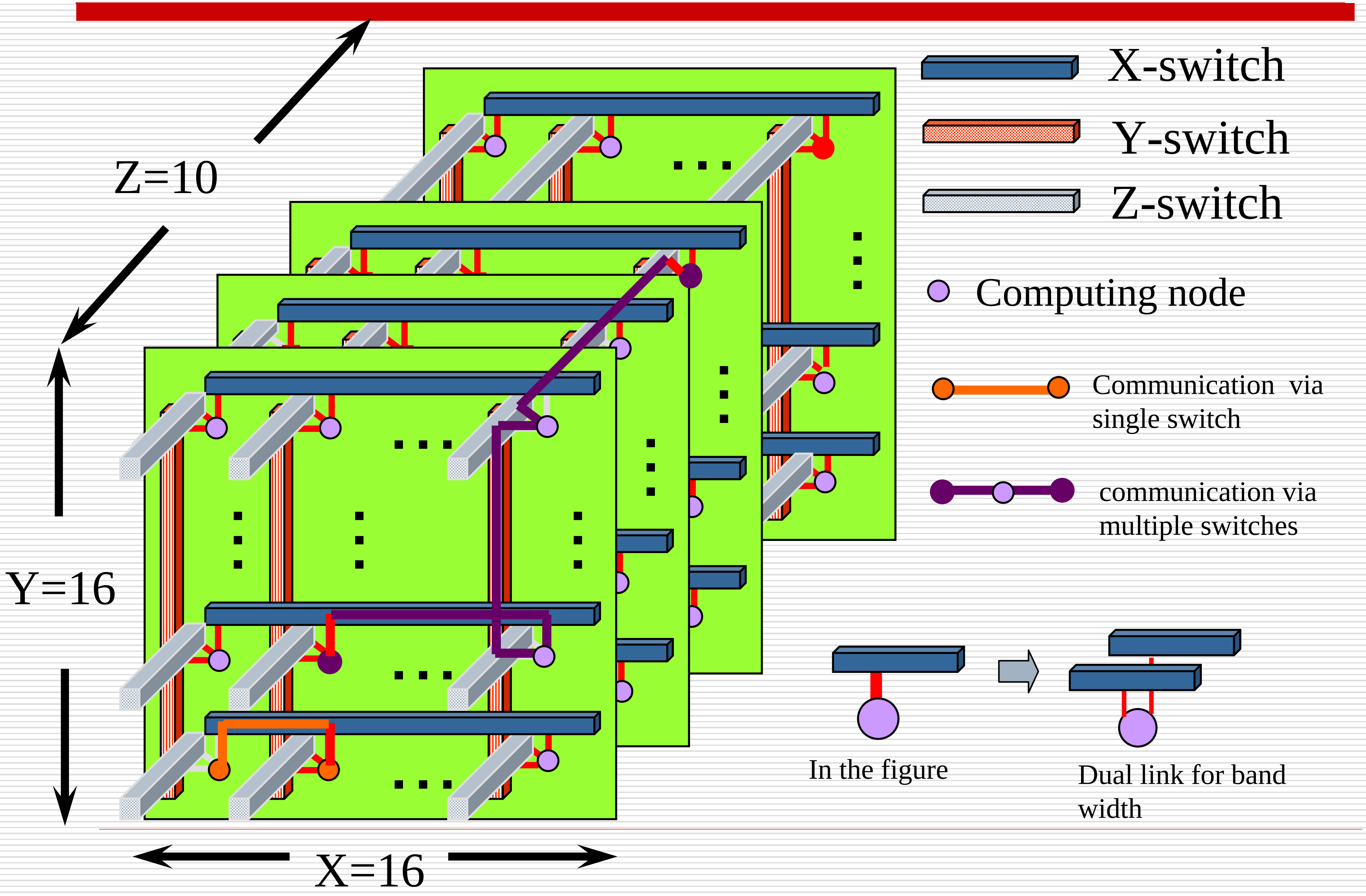
- Single CPU/node with fastest memory bus available
- Judicious choice of network topology

  (3-dimensional hyper-crossbar)
  - Multiple Gigbit Ethernet from each node for high aggregate bandwidth
  - Large number of medium-size switches to cut switch cost
- Mother board design to accommodate these features

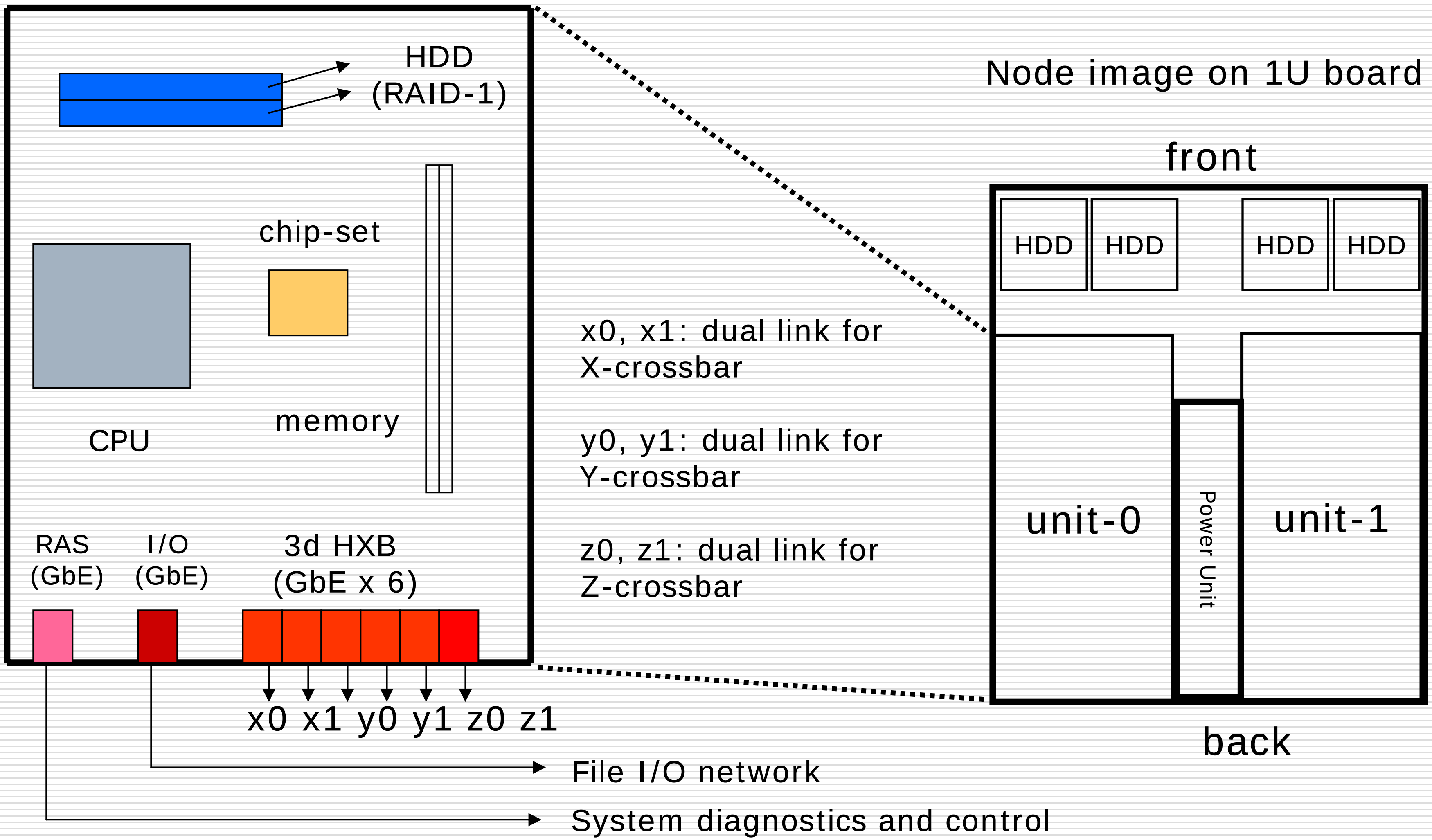
PACS-CS hardware specifications

- Node
  - Single low-voltage Xeon 2.8GHz 5.6Gflops
  - 2GB PC3200 memory with FSB800 6.4GB/s
  - 160GB disk (Raid1 mirror)
- Network
  - 3-dimensional hyper-crossbar topology
  - Dual Gigabit Ethernet for each direction,
    i.e., 0.25GB/s/link and an agregate 0.75GB/s/node
    (better than InfiniBand(x4) shared by dual CPU)
- System size
  - 16x16x10=2560 nodes, 14.3Tflops peak, 5.12TB memory,

3-dimensional hypercrossbar network



Board layout: 2 nodes /1U board



x0, x1: dual link for X-crossbar

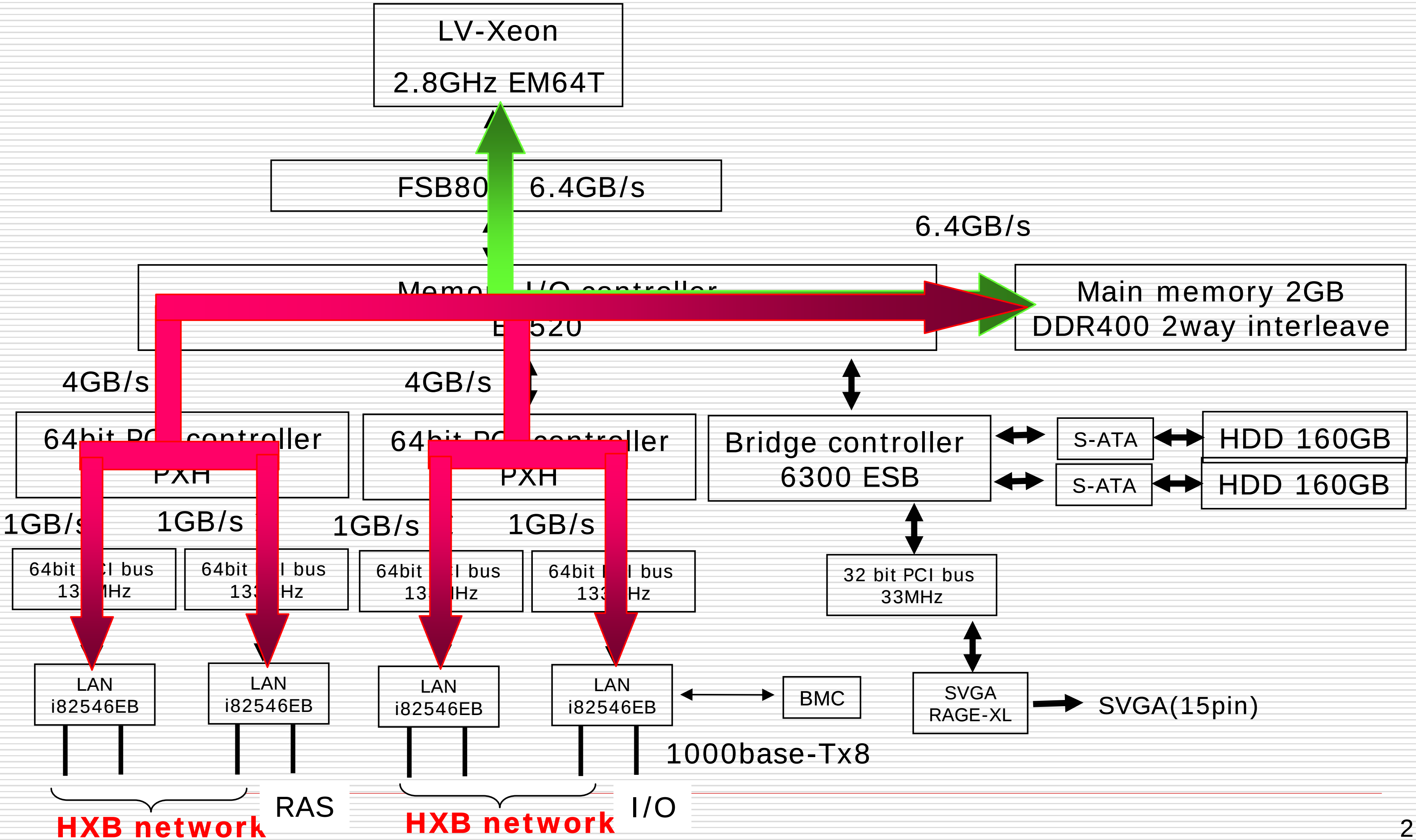
y0, y1: dual link for Y-crossbar

z0, z1: dual link for Z-crossbar

File I/O network

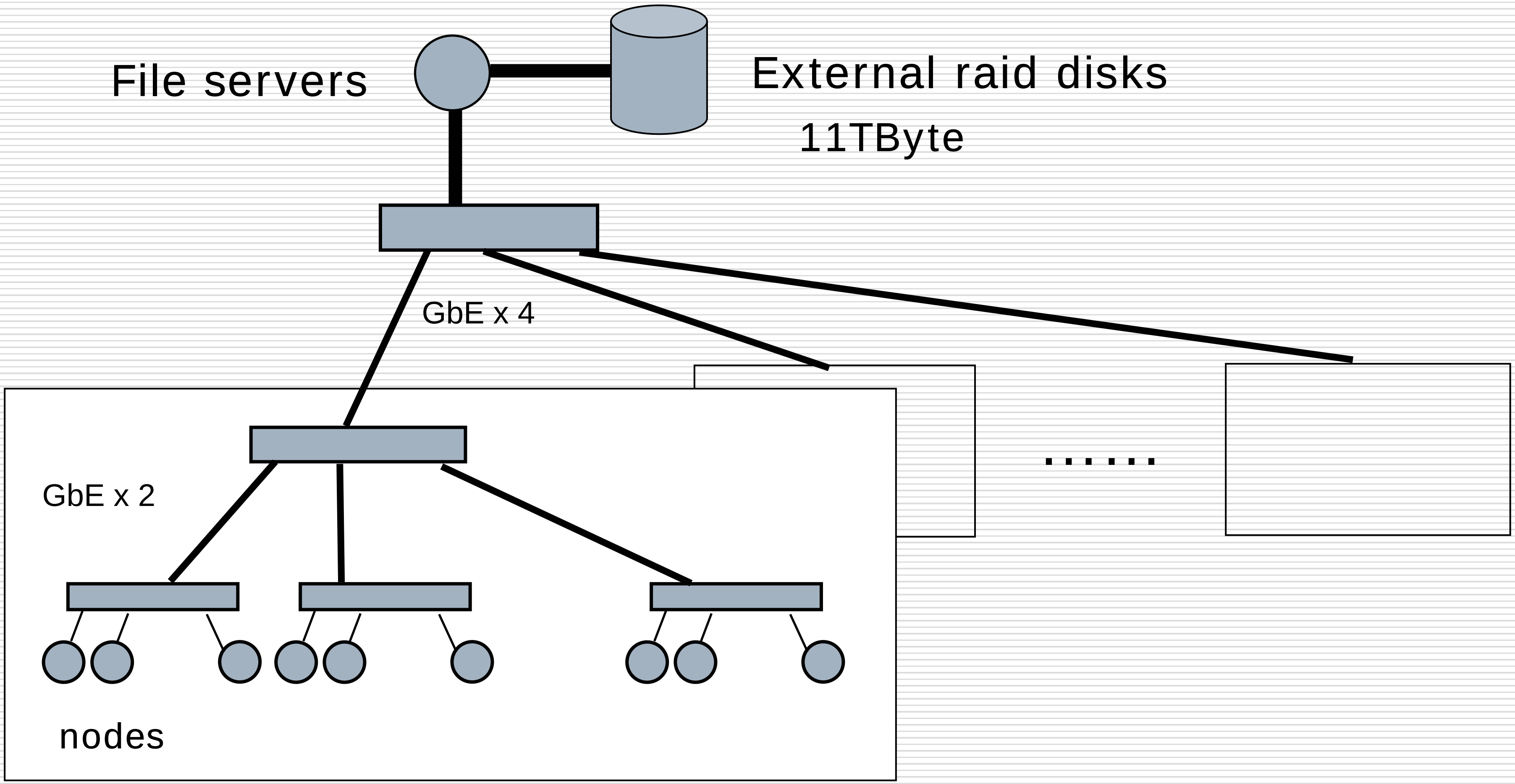
System diagnostics and control

Node block diagram



File server and external I/O

Separate tree network for file I/O



PACS-CS software

- OS
  - Linux
  - SCore (cluster middleware deveoped by PC Cluster Consortium http://www.pccluster.org/index.html.en)
  - 3D HXB driver based on SCore PMv2
- Programming
  - MPI for communication
  - Library for 3D HXB network
  - Fortran, C, C++
- Job execution
  - System partition (256nodes, 512nodes, 1024nodes, ...)
  - Batch queue using PBS
  - Job scripts for file I/O

Node performance

Written and optimized by K. Ishikawa

- Mult benchmark v2.62_sse3_64
  - Measures performance for Wilson-clover hopping term

$$(1+c_{sw}F\cdot\sigma)^{-1}\sum_{\mu}\left((1-\gamma_{\mu})U_{n\mu}+(1+\gamma_{\mu})U^{*}_{n\mu}\right)$$

  - Compiled with

    Intel C Compiler for EM64T, Version 8.1

    Intel Fortran Compiler for EM64T, Version 8.1
  - Same hardware spec as PACS-CS

    LV-Xeon 2.8GHz EM64T/FSB800/DDR2 2GB 2-way interleave

  - 8x8x8x64 result
    - C with SSE3 assembler coding 1.87Gflops (33%)
    - C with Intel intrinsic function 1.91Gflops (34%)
    - Fortran 1.45Gflops (26%)

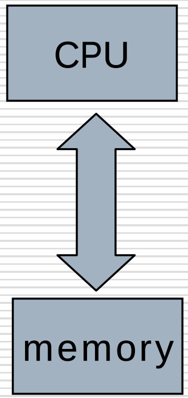
QCDMult benchmark performance analysis

- #floating operations and I/O with Mult routine
  - #flop executed 1896
  - #I/O needed 5088 Byte

  $\Big\}$ **2.68 Byte/flop**
- Since max I/O possible is 6.4GByte/s,

  max floating speed = 6.4/2.68 **2.39 Gflops(37.3%)**

	flop			Load (Byte)		Store (Byte)	Byte/ flop
	mult	add	total	U	p	q	
t	168	120	288	288	384	192	3.00
x	144	192	336	288	576	192	3.14
y	144	192	336	288	576	192	3.14
z	144	192	336	288	576	192	3.14
clover	288	312	600	672	192	192	1.76
total	888	1008	1896	1824	2304	960	2.68



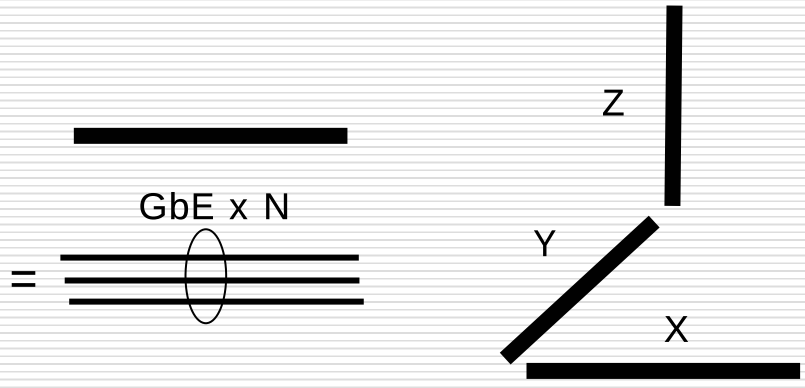
Network driver PM/Ethernet-HXB

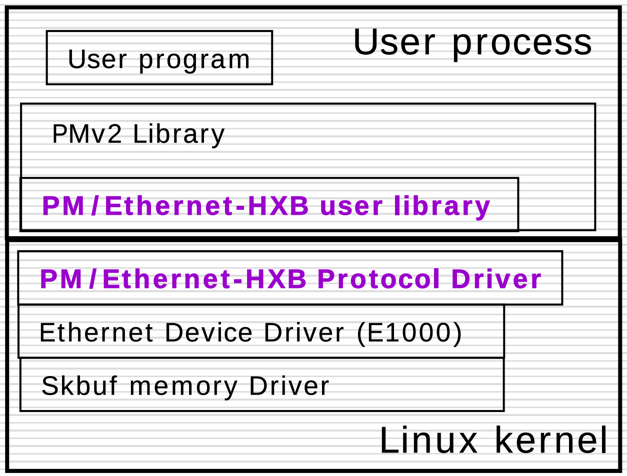
Being developed by S. Sumimoto, K. Kumon, T. Boku, M. Sato

- Key features
  - Trunking of multiple Ethernet links

    up to N=8;

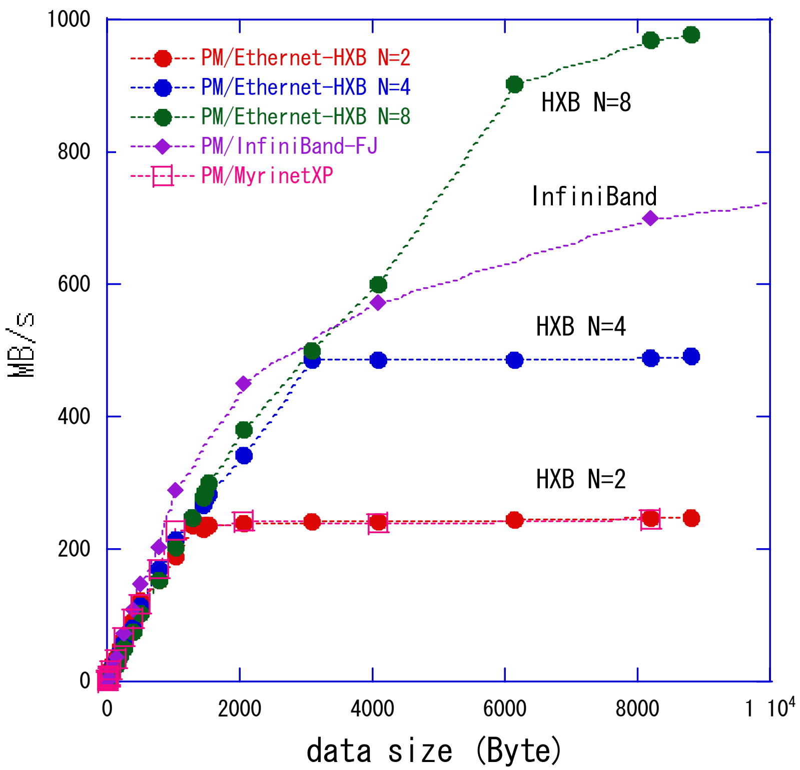
    use N=2 (250MB/s max) for PACS-CS
  - Routing on a multi-dimensional crossbar network (X->Y->Z)
  - Zero-copy communication





PM/Ethernet-HXB : PM-level performance

- 1 dimension
  - throughput



- latency

latency	
PM/MyrinetXP	4.2 microsec
PM/InfiniBand-FJ	7.8 microsec
PM/Ethernet-HXB	15.1 microsec

Including 3 microsec for switch latency

- Multi-dimensions with N=2

	max bandwidth	latency
1-dimension	247MB/s (99%)	15.1 microsec
2-dimension	241MB/s (96%)	29.1 microsec
3-dimension	237MB/s (95%)	43.2 microsec

Including 3 microsec for switch latency for each dimension

Network performance estimate (I)

- 32x32x32x64 lattice on 8x8x8=512 nodes
  - 4x4x4x64/node
- BiCGStabL2 solver
  - QCDMult called 8 times
  - Calculation 2208*Nsite/2=4.522Mflop in addition
  - Communication Global sum called 6 times
- QCDMult
  - Calculation (1296+600)*(Nsite/2) =3.883 Mflop
  - Communication Ny*Nz*Nt/2*192 =96kByte
    for +xyz simultaneously
    Same for –xyz

Network performance estimate (II)

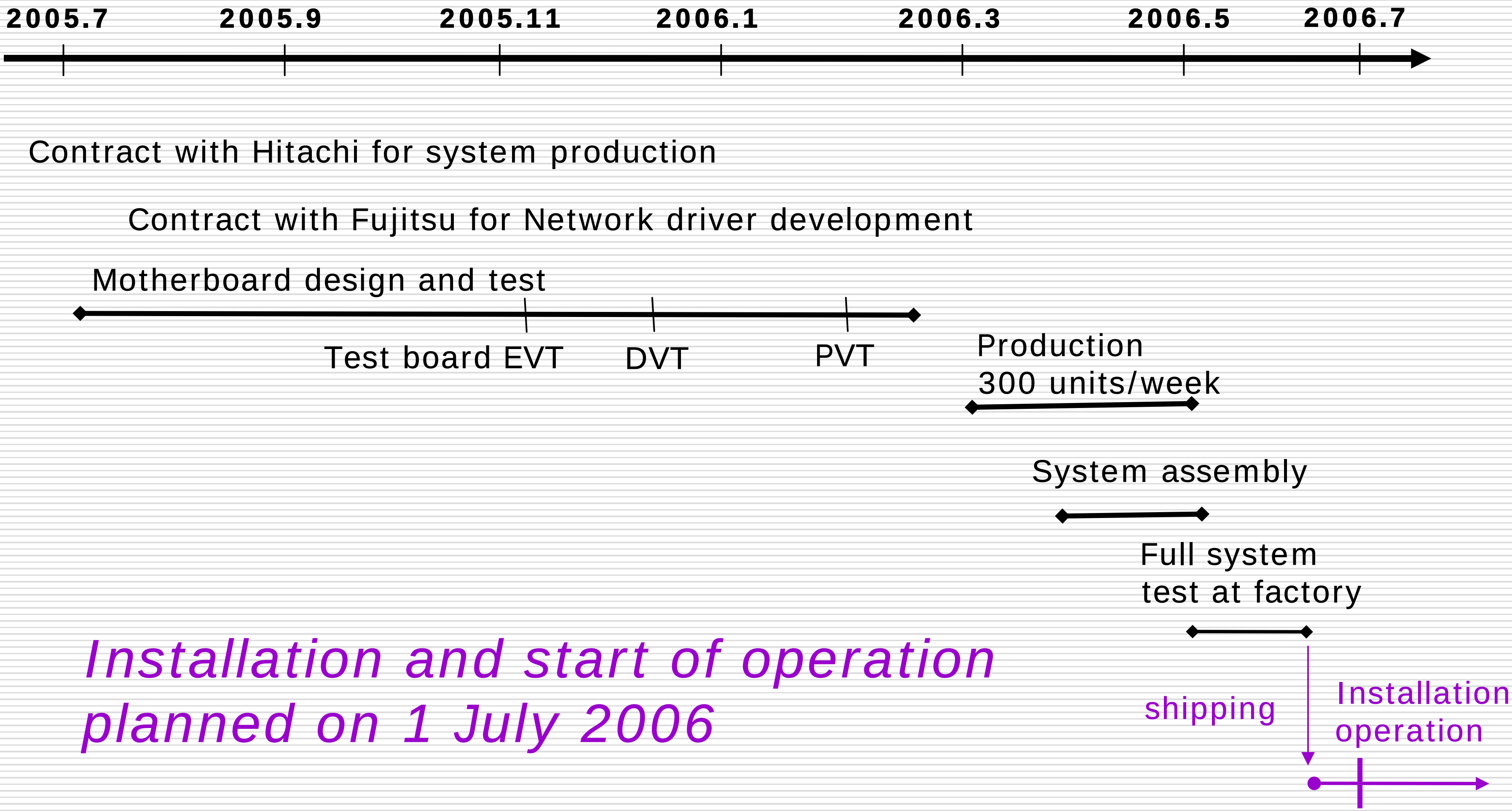
- Assumptions
  - 2Gflops/node
  - 250MB/s/link, 15microsec latency
    (N1/2=3.75kByte)

- Breakdown (BiCGStabL2)

calculation	17.79msec	71.6%
Neighbor communication	6.24msec	25.1%
Global sum (9 step cascade)	0.81msec	3.3%

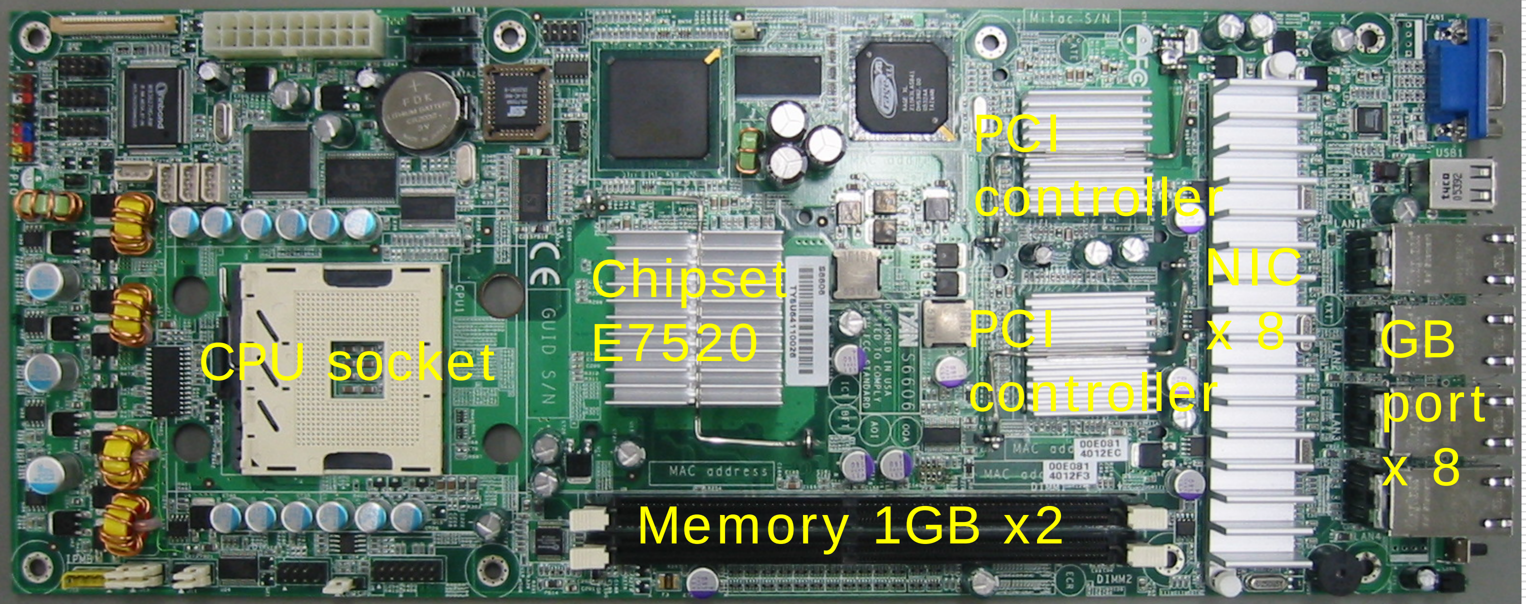
Network performance balanced;
In particular, latency OK for global sum

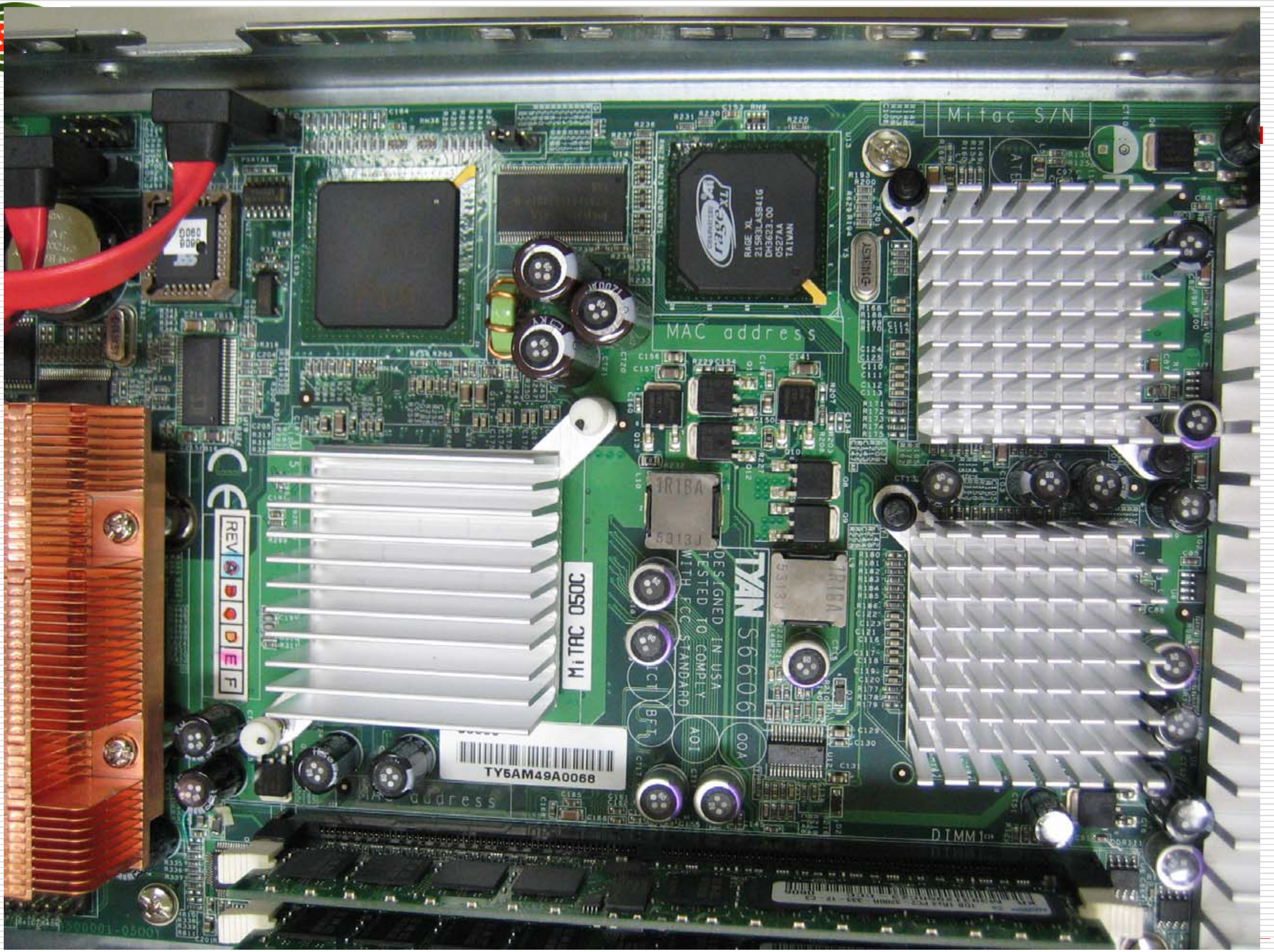
Design/testing/production

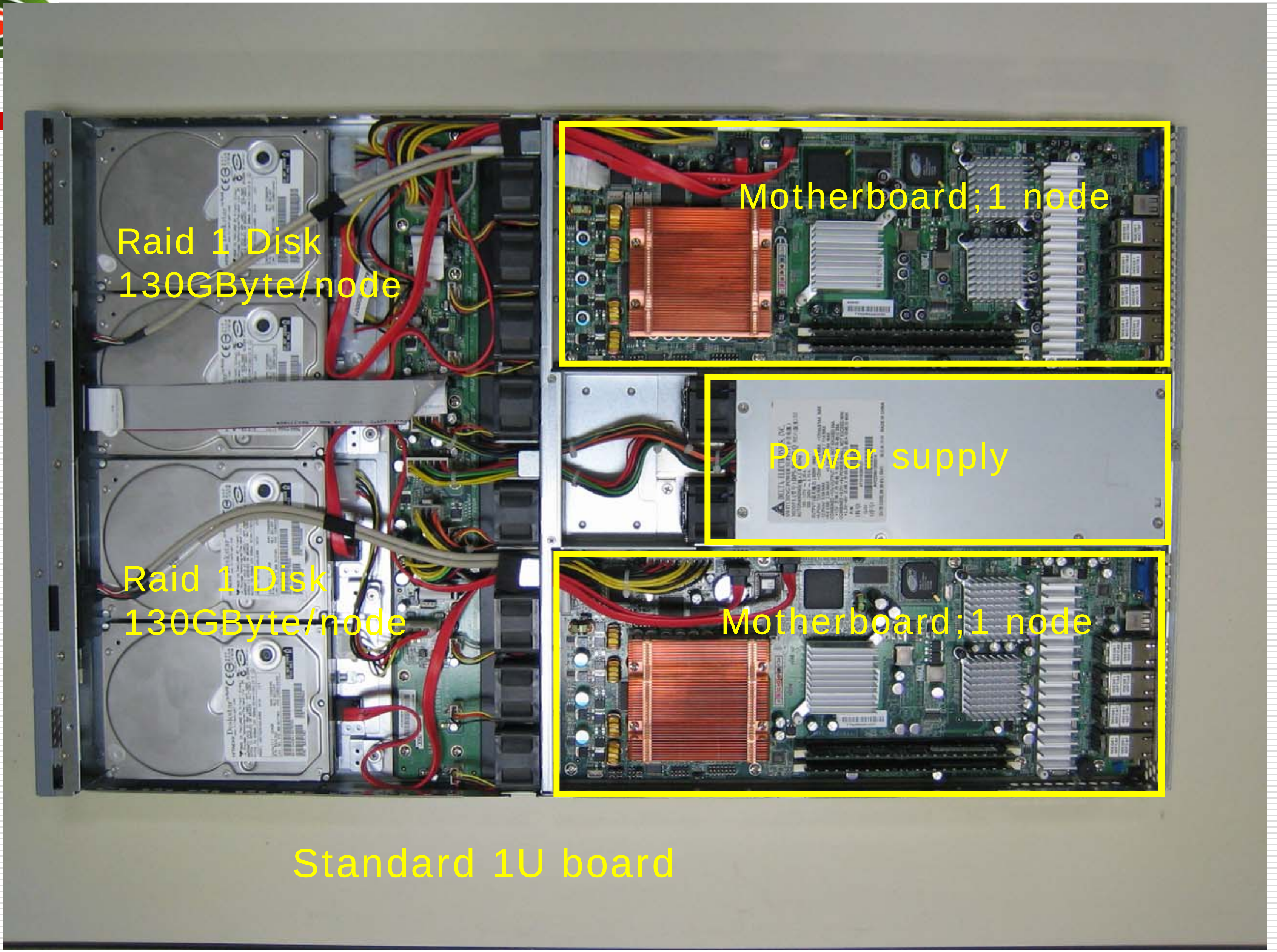


Installation and start of operation planned on 1 July 2006

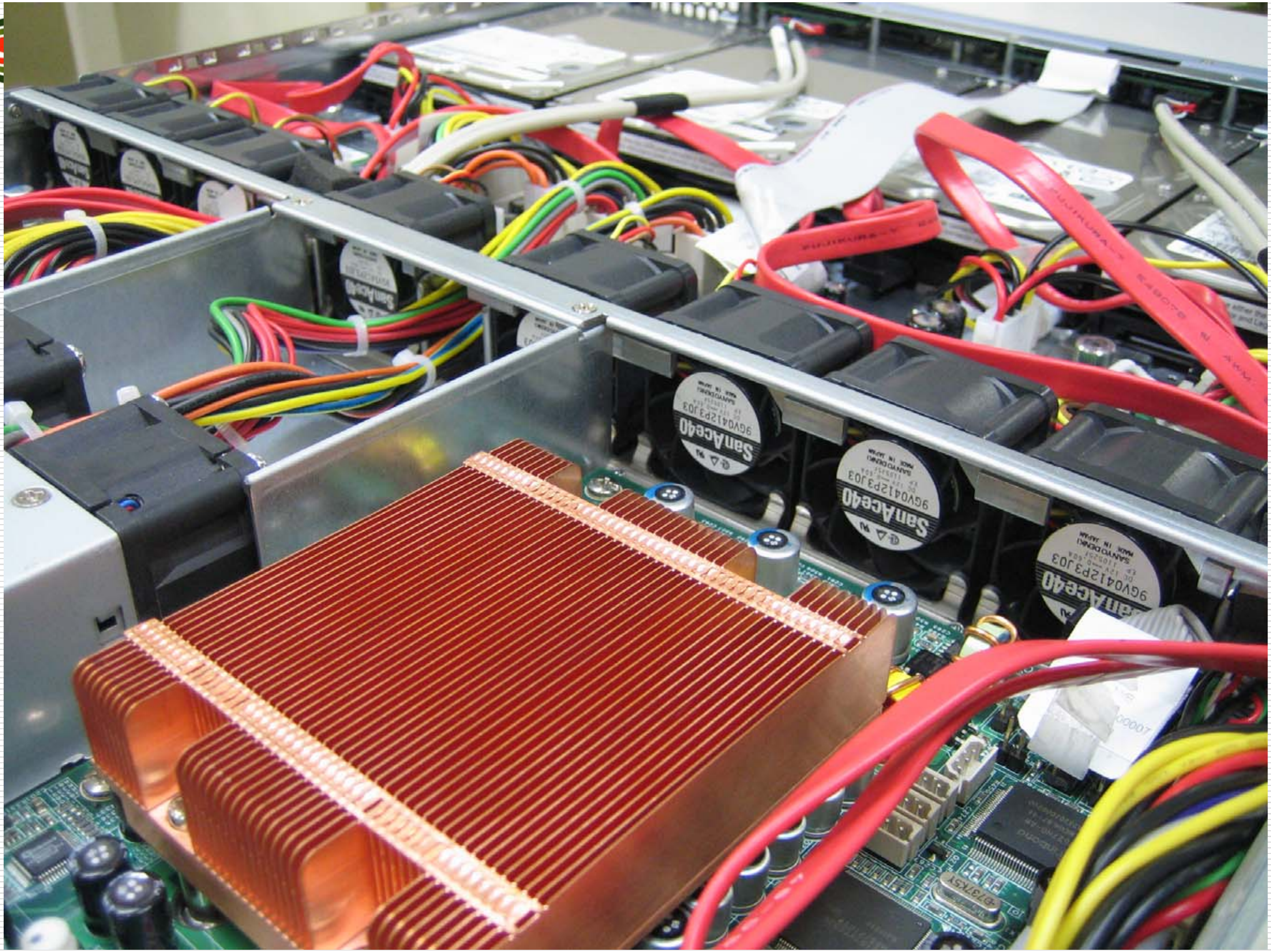
Mother board




Mitac S/N
MAC address
MiTAC 050C
DESIGNED IN USA
TESTED TO COMPLY
WITH FCC STANDARD
TY5AM49A0068
DIMM1

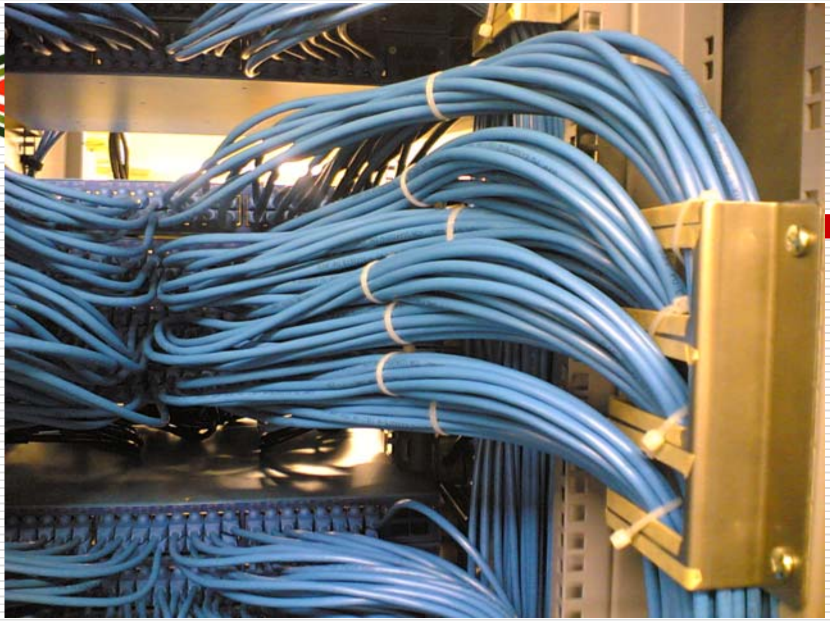

Raid 1 Disk
130GByte/node
Motherboard;1 node
Power supply
Raid 1 Disk
130GByte/node
Motherboard;1 node
Standard 1U board

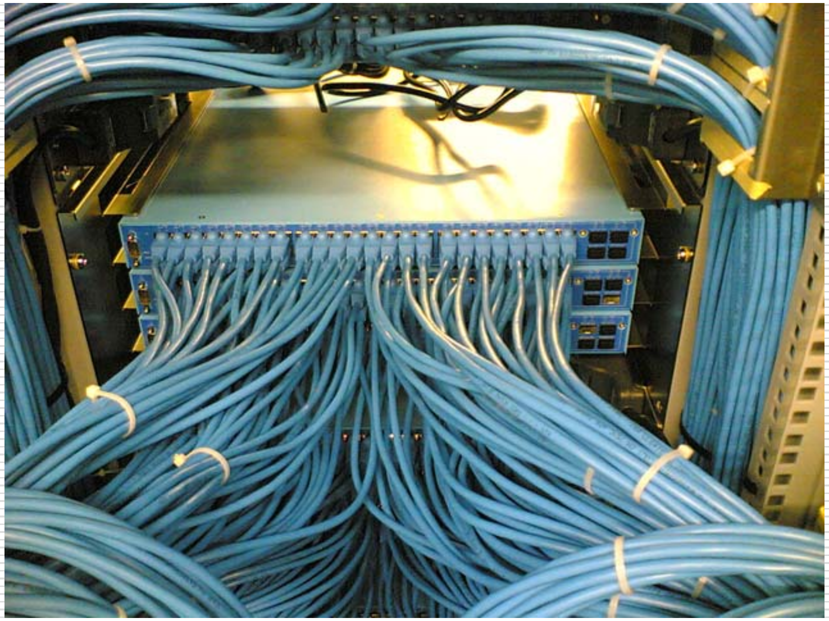

GB ports x 8
GB ports x 8


SanAce40
9GV0412P3J03


8
6
4
2
ID
7
5
3
1

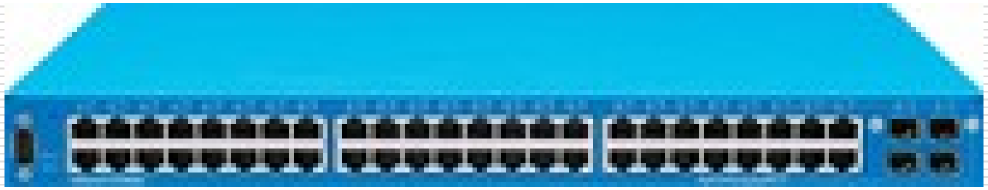
Switch rack






switch

- 48 port GB switch
- Hitachi Wires Apresia 4348GT
- Latency 3～5 microsec
- Throughput test OK

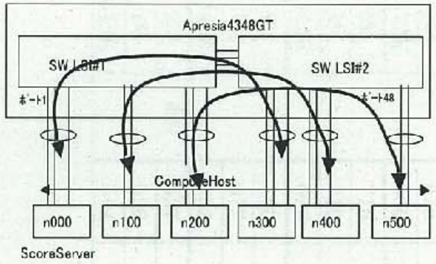


Switch test

sink_burst(SW-LSI跨り)

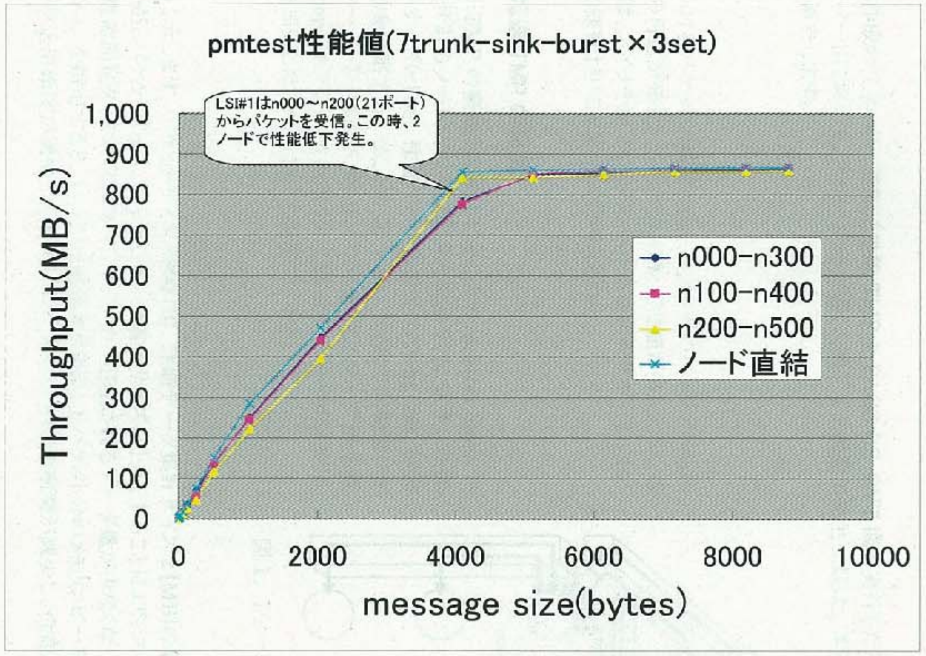
pmtest性能テスト

・測定日 2006/1/27
・測定オプション -sink/-burst -iter 1,000,000 ackskip=12
6ノード
42ポート(7trunk)を使用した、全通信がSW LSIを跨るテスト。



・測定結果(7trunk) (MBytes/sec)

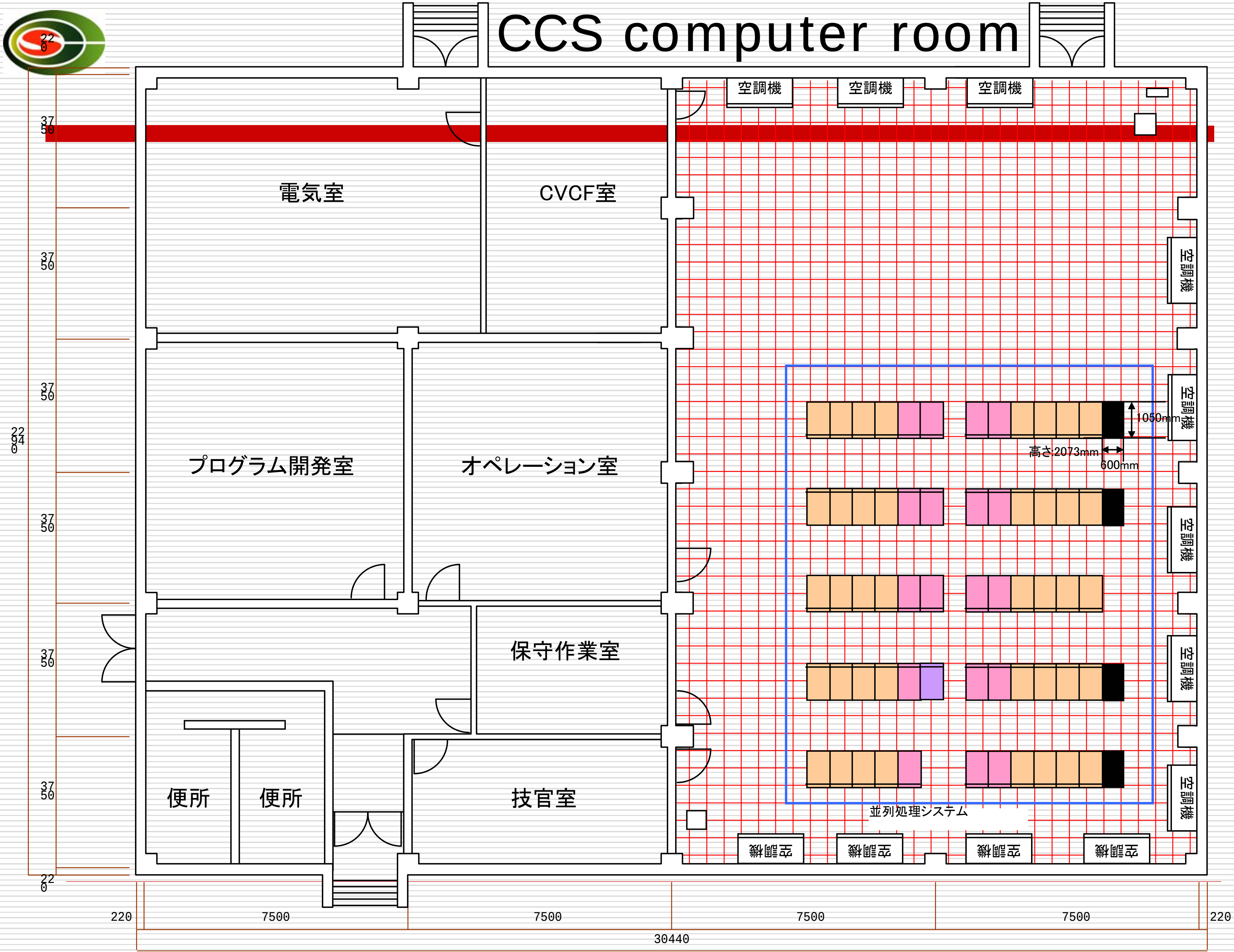
転送長(byte	ノード直結	n000-n300	n100-n400	n200-n500
4	1.1	1.08	0.77	0.85
8	2.2	2.10	1.53	1.71
16	4.4	4.34	3.11	3.53
32	8.8	8.66	6.14	7.20
64	17.6	17.11	12.16	13.78
128	35.7	33.95	24.39	26.15
256	74.0	69.61	57.98	47.56
512	151.2	136.28	139.03	117.07
1024	284.2	249.04	245.34	221.67
2048	469.9	446.11	440.48	395.17
4096	854.6	782.22	776.50	841.82
5120	860.0	848.82	851.90	842.54
6144	862.2	852.04	859.95	850.17
7168	864.1	858.04	863.18	857.39
8192	865.6	860.88	864.85	857.72
8800	866.3	862.16	865.70	858.93



Racking/Power/Footprint/

- Racking
  - Compute rack
    - 2 node/board, 32 board/rack, 40 racks in total
  - Switch rack
    - 48-port GbE switch/board, 19 racks in total
- Power 450kW
- Footprint $100 m^2$

CCS computer room



Looking ahead

- PACS-CS Collaboration
  - Formed in August 2005
  - Kanaya, Aoki, Yoshie, Ishizuka, Kuramashi, Taniguchi, T. Ishikawa, Sasaki, Ukawa/Tsukuba
  - Tsutsui/KEK
  - Okawa, K. Ishikawa/Hiroshima

Initial physics plan for PACS-CS

- Complete the Wilson-clover Nf=2+1 program
  - current run: talk by Tomomi Ishikawa
    - Three lattice spacings

$$a^2 \approx 0.015\,fm^2\,,\,0.01\,fm^2\,,\,0.005\,fm^2$$

    - But, in light quark masses, only down to

$$\frac{m_\pi}{m_\rho} \approx 0.6 \quad i.e., \quad \frac{m_{ud}}{m_s} \approx 0.5$$

  - *Wish to go down to light quark masses, e,g,*

$$\frac{m_\pi}{m_\rho} \approx 0.4 \quad i.e., \quad \frac{m_{ud}}{m_s} \approx 0.2$$ *or less...*

using Luescher's domain-decomposed HMC

DD HMC coding and test (I)

- Nf=2 code for plaquette gauge + naïve Wilson
  - Written and tested on 16x32 at beta=5.6, K=0.15750 (one of Luescher runs)
- Nf=2 code for Iwasaki RG glue + Wilson-clover
  - Written and tested on 16x32 at beta=1.8, Kud=0.1409 (one of CP-PACS runs)
- Nf=2+1 code for Iwasaki RG glue + Wilson-clover (PHMC for strange quark)
  - Written
  - Being tested on 16x32 at beta=1.83, Kud=0.13655, Ks=0.13710 (one of CP-PACS/JLQCD runs)

A paper estimate one year ago

1/a	lattice size		pi/rho	standard HMC 10000traj	domain-decomposed HMC #steps			time/traj(hr)			10000traj	acceleration
(GeV)	Ns	Nt		(days)	N0	N1	N2	calc	comm	total	(days)	
2	24x48		**0.6**	**26**	4	5	5	0.031	0.005	0.037	**4**	**7**
			0.5	**65**	4	5	6	0.058	0.010	0.068	**7**	**9**
			0.4	**180**	4	5	7	0.110	0.019	0.129	**13**	**13**
			0.3	**629**	4	5	8	0.230	0.041	0.271	**28**	**22**
			0.2	**5372**	4	5	9	0.747	0.133	0.880	**92**	**59**
2.83	32x64		**0.6**	**118**	5	6	6	0.181	0.018	0.199	**21**	**6**
			0.5	**303**	5	6	7	0.333	0.033	0.366	**38**	**8**
			0.4	**860**	5	6	9	0.713	0.071	0.784	**82**	**11**
			0.3	**3036**	5	6	10	1.475	0.147	1.622	**169**	**18**
			0.2	**26238**	5	6	11	4.739	0.473	5.213	**543**	**48**

- Only a paper estimate, but more than encouraging .....
- Implementation in progress

DD HMC coding and test (II)

- Scaling test started in February
- Nf=2+1; beta=1.90 1/a=2GeV 16^3x32
- Pi/rho= 0.8 Kud heaviest
- 0.6 Kud lightest
- 0.5 tune from hadron mass data
- 0.4 ditto

We'll soon know how light we can go down with PACS-CS

KEK and JLQCD

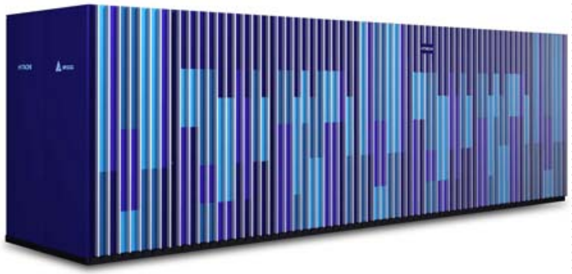
KEK supercomputer facility

- 1985 Hitachi S810/10 350 MFlops
- 1989 Hitachi S820/80 3 GFlops
- 1995 Fujitsu VPP500 128 GFlops
- 2000 Hitachi SR8000 F1 1.2 TFlops

- 2006

  Hitachi SR11000 K1 2.1Tflops

  IBM BlueGene/L 57.3Tflops(10 racks)



Supported by a regular funding for computing
Upgrade every 5-6 years (so far)

JLQCD and physics program

- Members
  - Hashimoto, Kaneko, Yamada, Okamoto, Matsufuru/KEK
  - Kanaya,Aoki,Yoshie,Ishizuka,Kuramashi, Taniguchi, Ukawa/Tsukuba
  - Onogi,Ukita/kyoto
  - Okawa, Ishikawa/hiroshima
- Dynamical overlap program
  - Coding and optimization
  - Choice of gauge action
  - Choice of run parameters

Once more…

2004.3 2004.9 2005.3 2005.9 2006.3 2006.9

Japan

CP-PACS 0.6T

SR8k 1.2Tflop

USA

QCDSP 0.6Tflops

An exciting period ahead

- Finally Nf=2+1
- Finally chiral
- Burst of activities is worldwide

PACS-CS 14.3Tflops

SR11k 2.1Tflops

BG/L 57.3Tflops

RBRC QCDOC 10Tflops

SciDAC QCDOC 10Tflops

EU

UKQCD QCDOC 10Tflops

Ideal timing for ILDG to serve its role

apeNEXT **INFN Rome 8Tflops**
Bielefeld 4Tflops
DESY Zeuthen 2Tflops
.....

Julich BG/L 45.8Tflops

summary

- *We've come a long way since the time of Izu Workshop*
- *An exciting period ahead*
- *Hope ILFT Network has served its purpose in the building up of ILDG and promoting international exchange within our community*
- *But, perhaps time to think about new ideas and new format on how we organize and run the ILFT Network*