Lattice 2006
July 25
Tucson

Status and physics plan of the PACS-CS Project

Akira Ukawa
Center for Computational Sciences
University of Tsukuba

- *Collaboration members*
- *PACS-CS status*
- *physics plan*
- *Summary*

Related talks:

T. Ishikawa	Spectroscopy session 3(Tue)
K. Ishikawa	Algorithm session 2(Tue)
Y. Kuramashi	Algorithm session 1(Mon)

Collaboration members

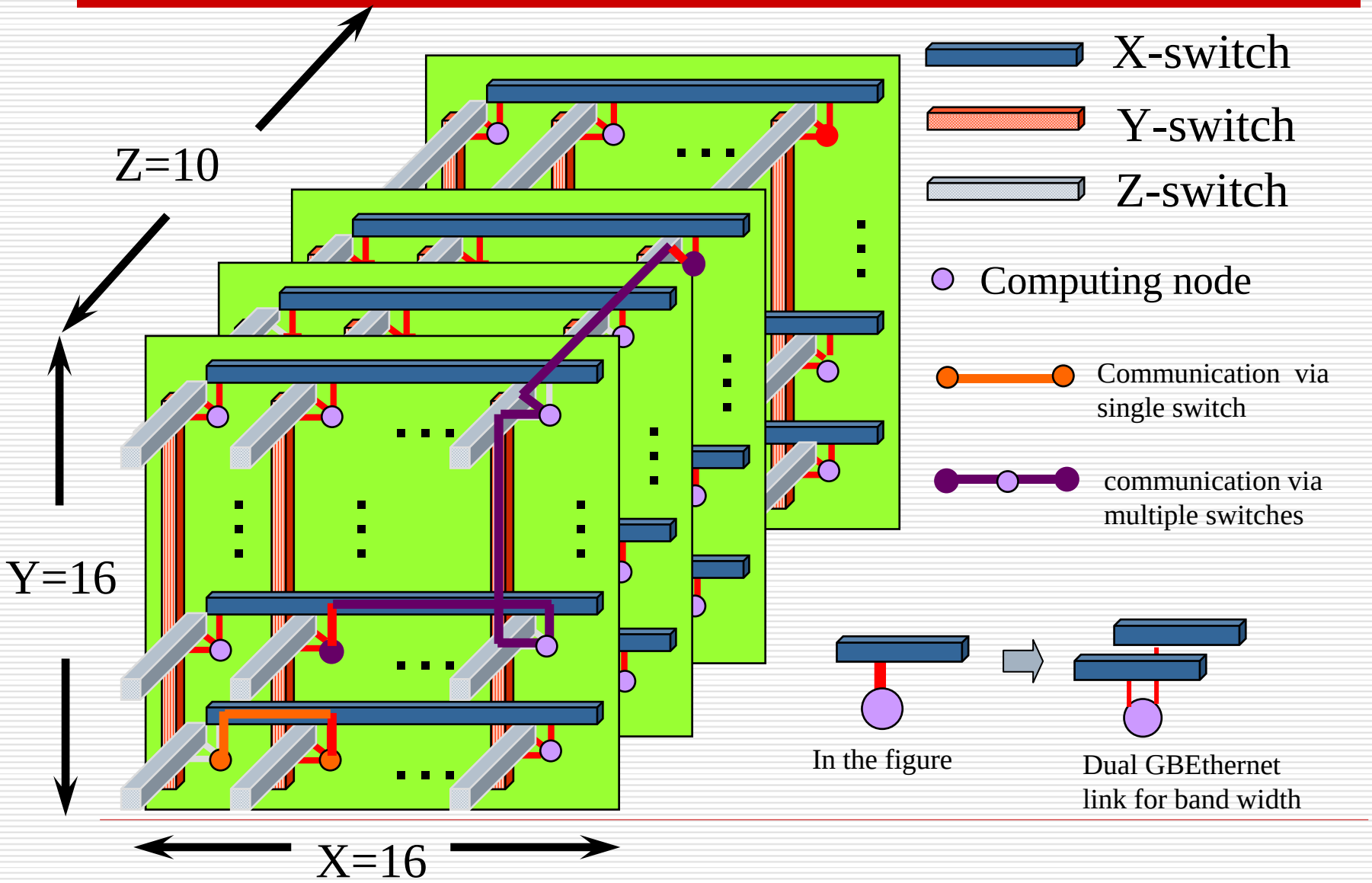
- Physicists

  S. Aoki, T. Ishikawa, N. Ishizuka, K. Kanaya,
  Y. Kuramashi, K. Sasaki, Y. Taniguchi, A. Ukawa,
  T. Yoshie *at Tsukuba*
  K.-I. Ishikawa, M. Okawa *at Hiroshima*
  N. Tsutsui *at KEK*
  T. Izubuchi *at Kanazawa*

- Computer scientists

  T. Boku, M. Sato, D. Takahashi, O. Tatebe *at Tsukuba*

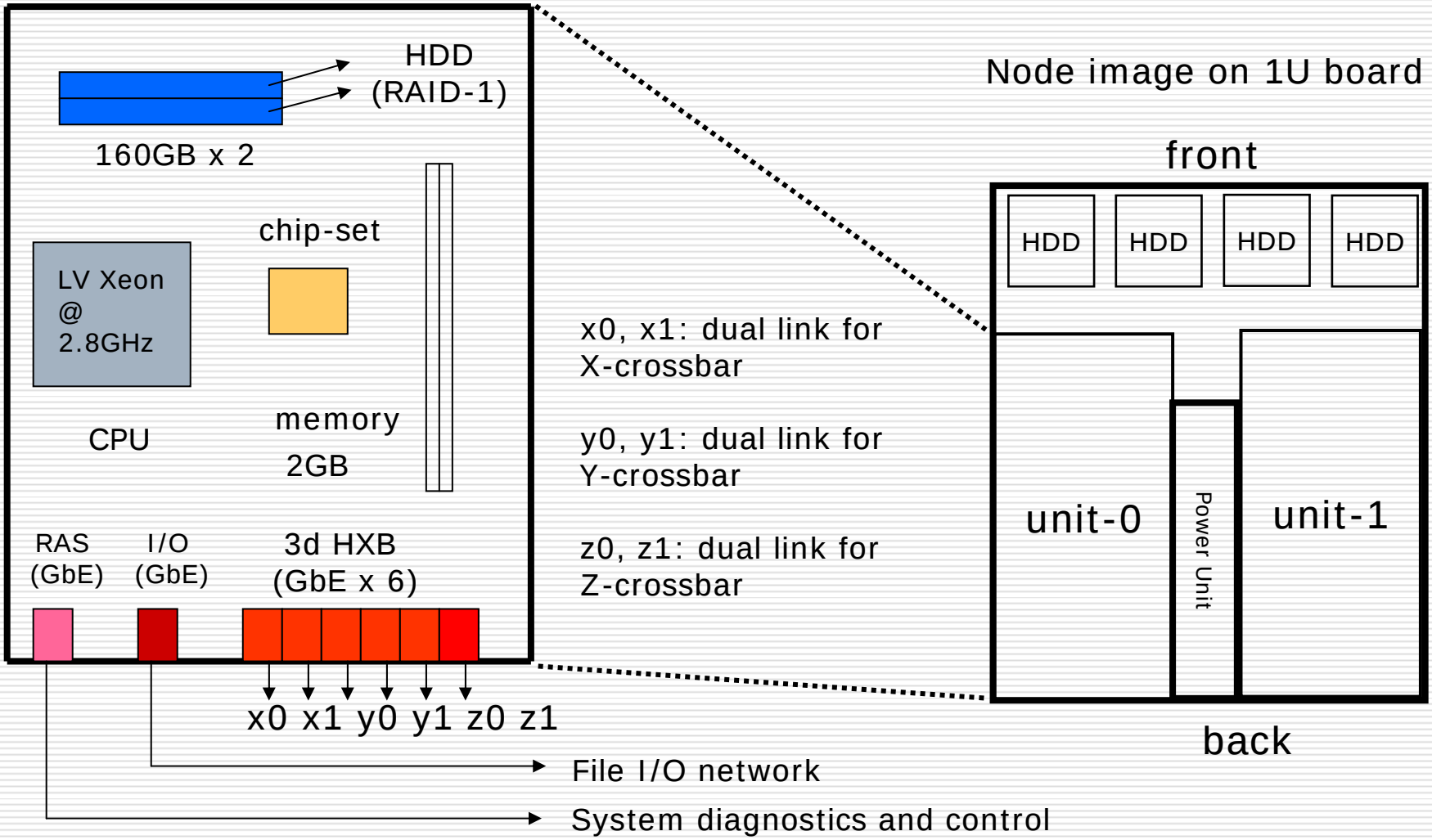
Schematic diagram of PACS-CS/2560



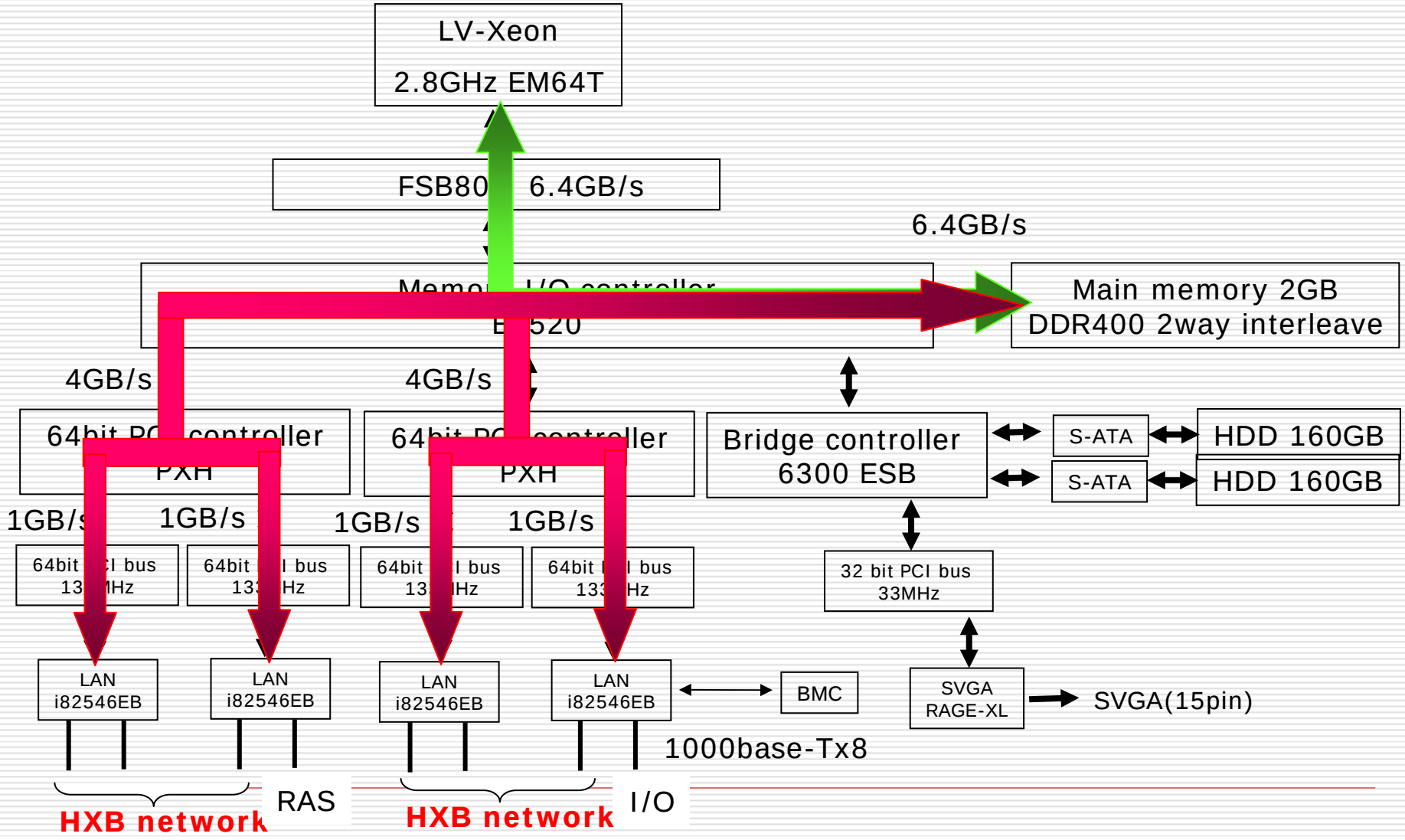
PACS-CS specifications

#nodes	2560 (16x16x10)
peak performance	14.3Tflops
node	single CPU+memory+HDD+8 GBEthernet ports
CPU	Intel LV Xeon EM64T, 2.8GHz, 1MB L2 cache
memory	2GB/node (5.12TB/system)
network	3dimensional hyper-crossbar uses dual GBEthernet/link
network performance	250MB/s/direction 750MB/s/node (3dim. simultaneous send/receive)
local HDD	160GBx2 (RAID-1) (410TBx2/system)
#racks	59 racks
footprint	100㎡
power	545kW

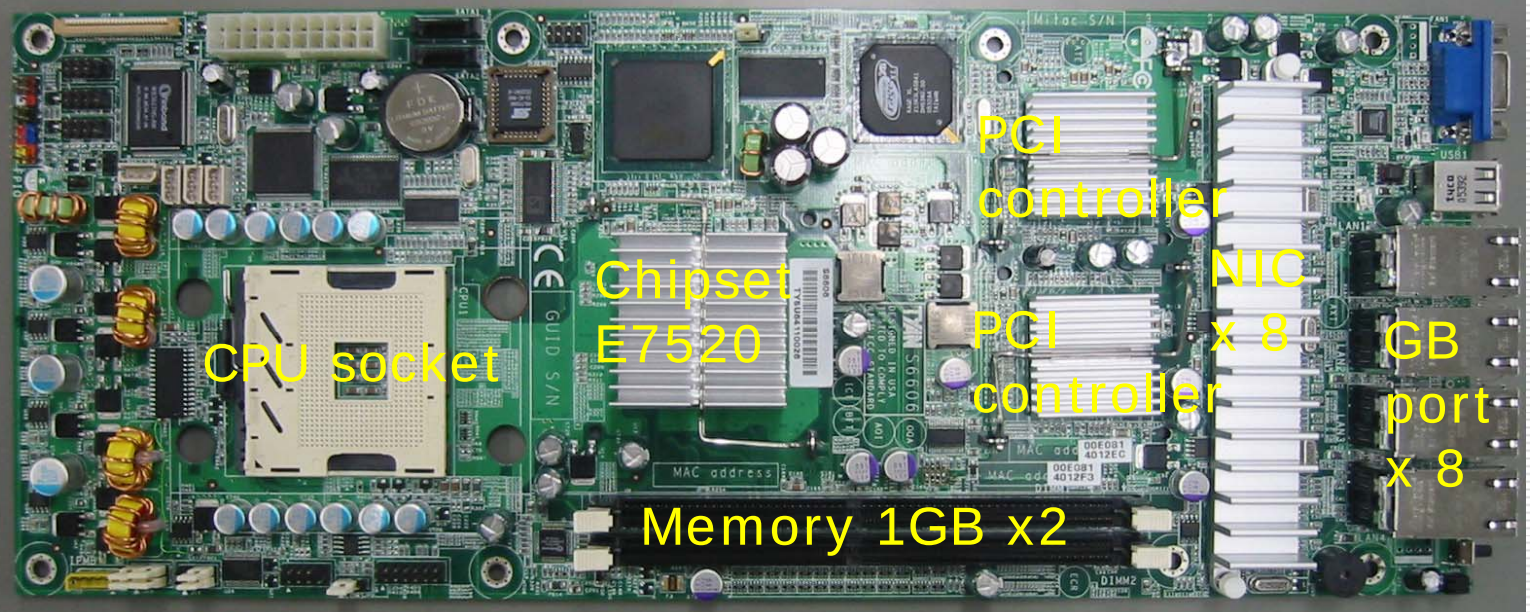
Board layout: 2 nodes /1U board



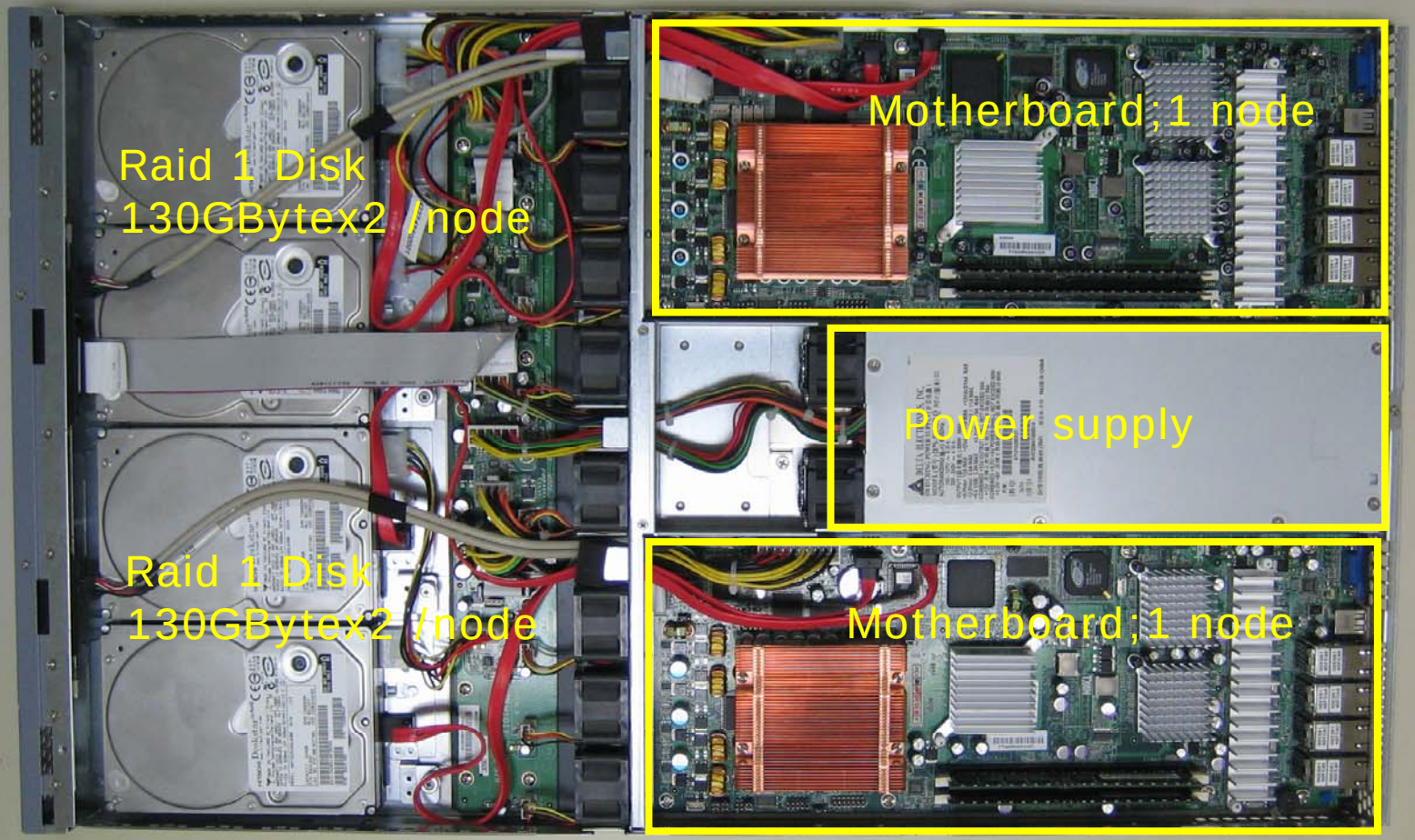
Node block diagram

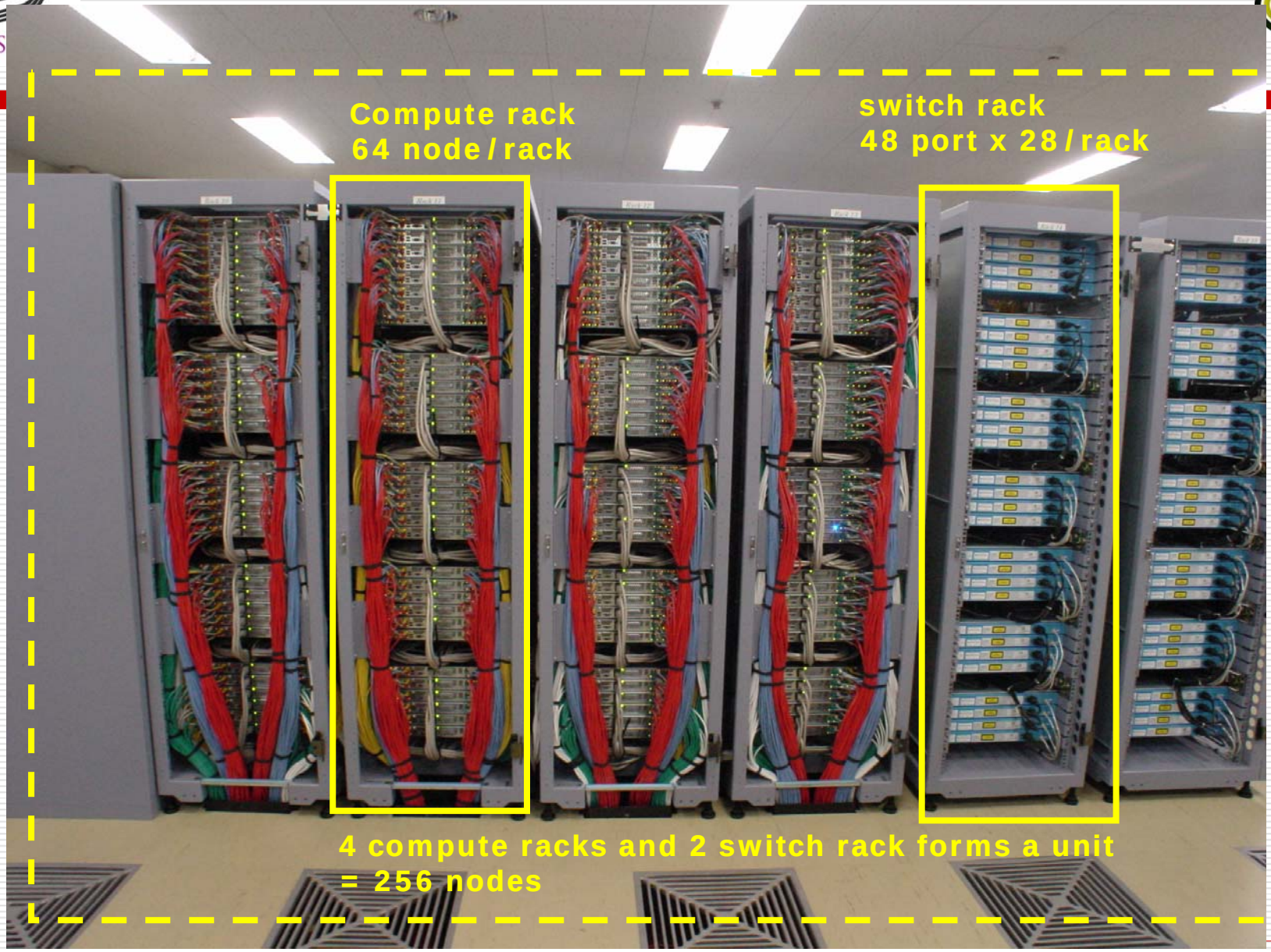


Mother board



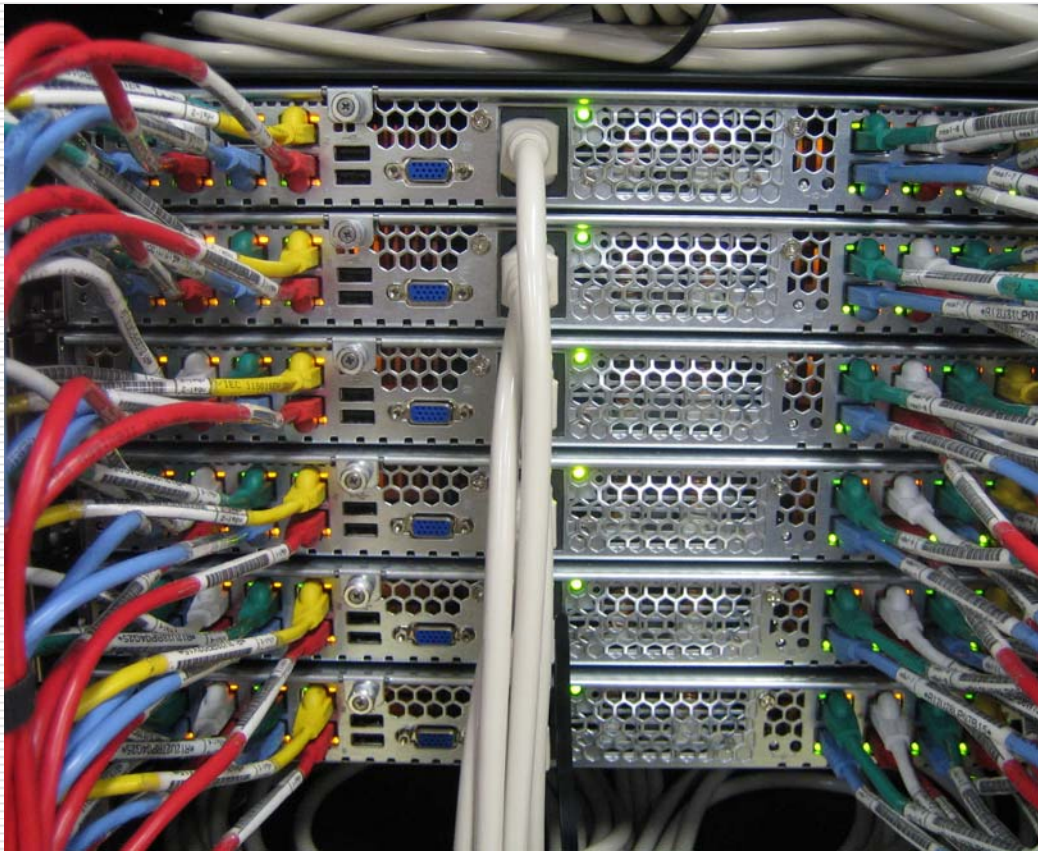
2 mother boards on a single 1U board




Compute rack
64 node / rack
switch rack
48 port x 28 / rack
4 compute racks and 2 switch rack forms a unit
= 256 nodes

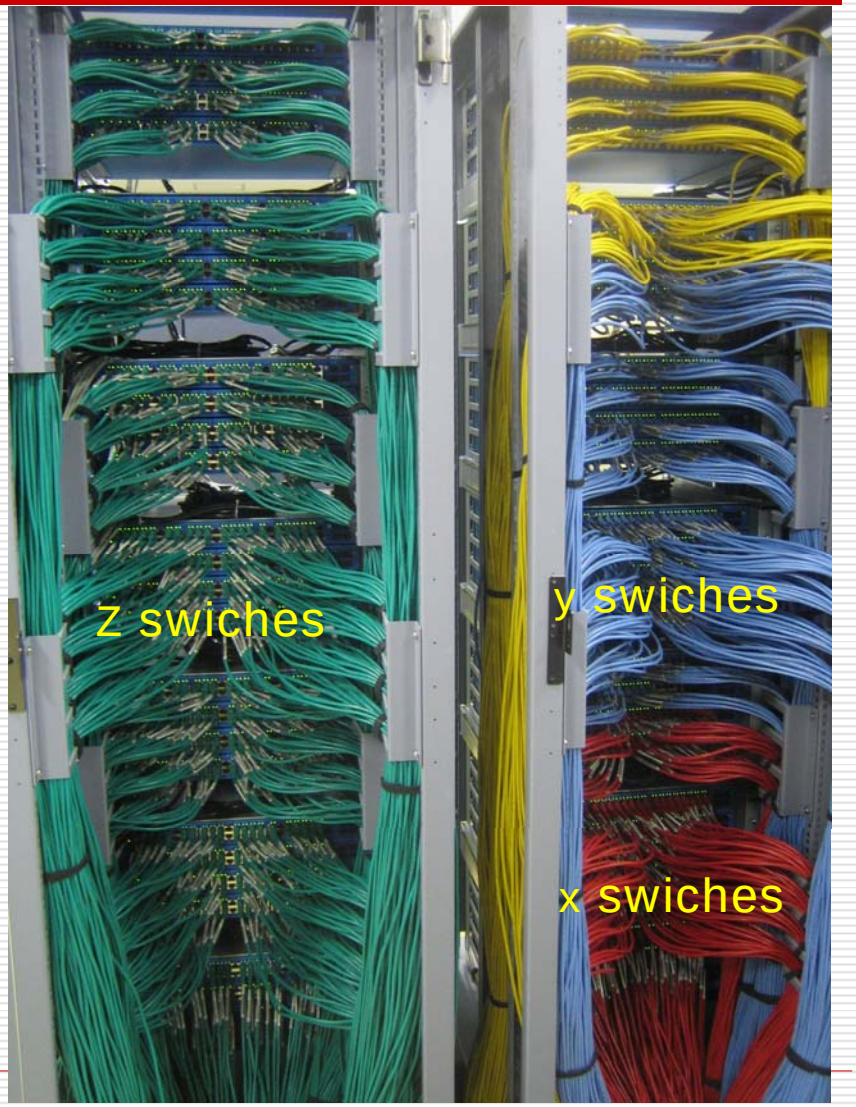
GBitEther cables and switches (I)

- 8 cables from each node
  - 6 for hypercrossbar network (Red,Blue,Green)
  - 1 for external file I/O (white)
  - 1 for RAS (yellow)
- 8 x 2560=20480 cables over 400km in length



GBitEther cables and switches (II)

Swich rack viewed from back



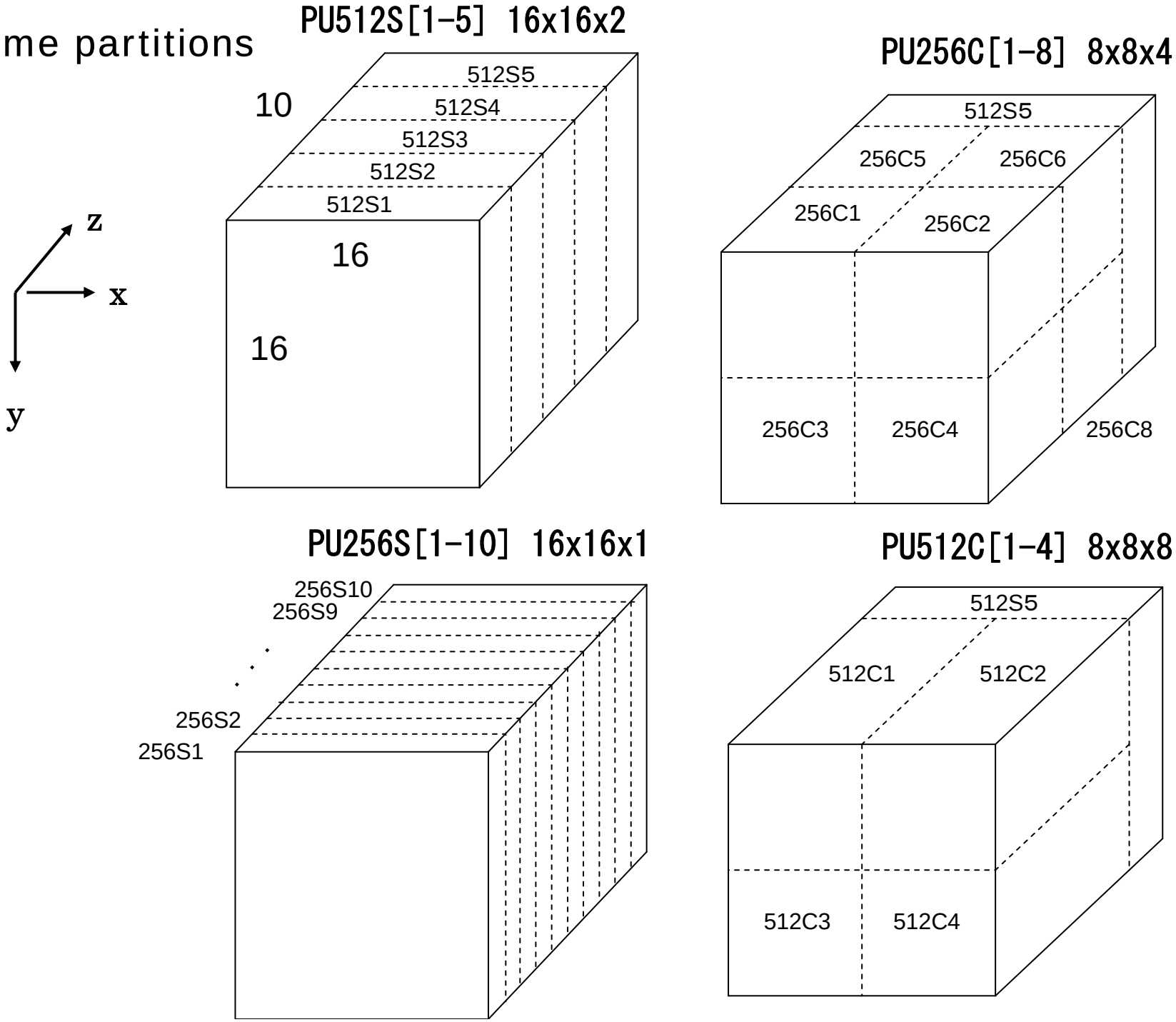
cables under the floor

PACS-CS/2560

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

some partitions



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
  - 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%)

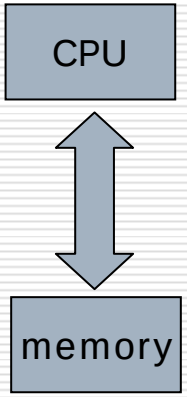
QCDMult performance expectations

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

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

  max possible flops = 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 performance/expectations

- Assumptions
  - 2Gflops/node
  - 750MB/s/node for simultaneous send to xyz directions, 15microsec latency

- expected (BiCGStabL2)

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

If realized, network performance balanced for Wilson-clover simulations

Network performance/measurements

- 3 dim. simultaneous send (MB/s and % to peak) using current PMv2 network driver

	256node	512node
Ave.	586.8 (78.2%)	582.0 (77.6%)
Min.	559.2 (74.6%)	434.0 (57.9%)
Max.	619.3 (82.6%)	629.6 (84.0%)

- Global sum (MPI_Allreduce) (msec)

	8B	800kB
Ave.	0.420	257
Min.	0.344	491
Max.	0.727	52

Preliminary values

Expect improvement with PMvX driver under development (reduced buffer copy)

Initial physics plan for PACS-CS

- Complete the Wilson-clover Nf=2+1 program
  - current status: → T. Ishikawa, Spectroscopy session 3
    - Three lattice spacings and continuum extrapolation

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

  - Encouraging results on the meson spectrum and light quark masses
  - But, Light quark masses only down to

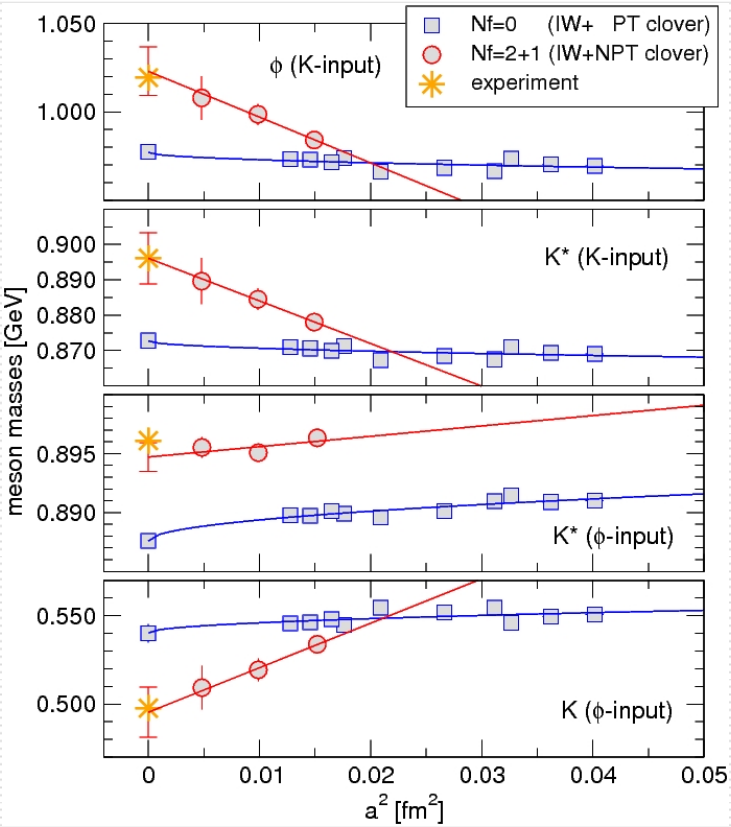
$$\frac{m_\pi}{m_\rho} \approx 0.6 \quad i.e., \quad \frac{m_{ud}}{m_s} \approx 0.5$$

  - Wish to go down to lighter quark masses, e,g,

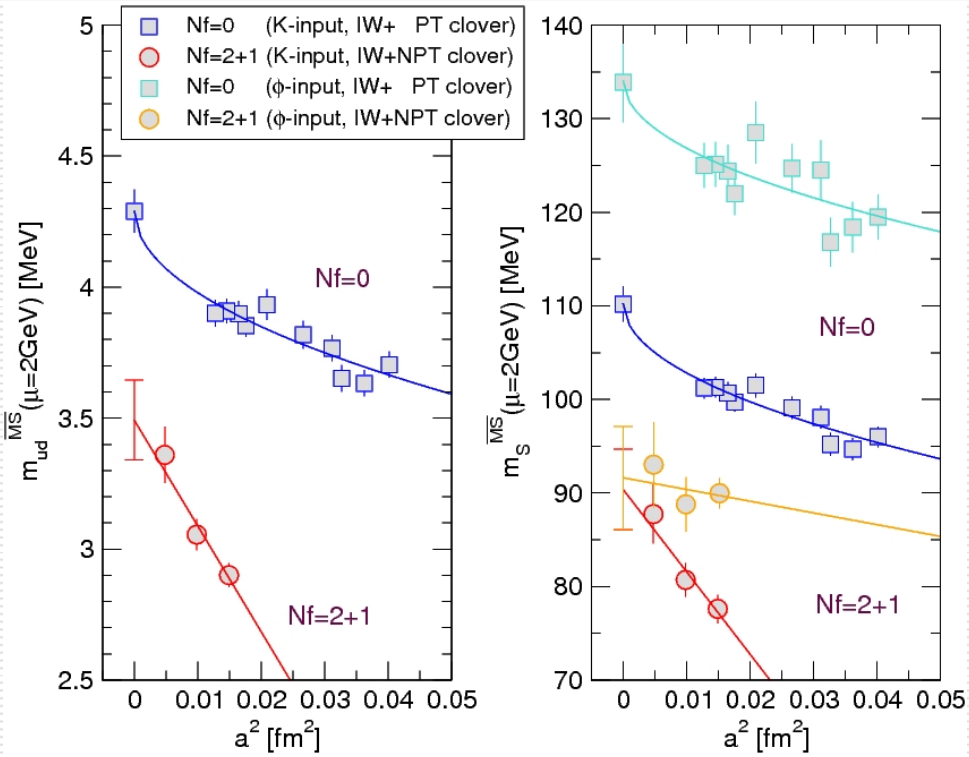
$$\frac{m_\pi}{m_\rho} \approx 0.3 \quad i.e., \quad \frac{m_{ud}}{m_s} \approx 0.15 \qquad \text{or so}$$

CP-PACS/JLQCD results

Meson masses



Light quark masses



Strategy toward lighter quark masses

- Algorithm
  - (degenerate) up-down quarks:
    Luescher's domain-decomposed HMC algorithm
  - strange quark: ⇨ K. Ishikawa, algorithm session 2
    polynomial HMC with UV filtering (factor 2 speed up)
- Lattice parameters planned

		CP-PACS/JLQCD run				planned PACS-CS run			
beta	a(fm)	lattice size	kappa_ud	mpi/ mrho	kappa_s	lattice size	kappa_ud	mpi/ mrho	kappa_s
1.83	0.122	16^3x32	0.13655-0.13825	0.61-0.78	0.13710, 0.13760	24^3x48			
1.90	0.100	20^3x40	0.13580-0,13700	0.63-0.77	0.13580, 0.13640	32^3x64	0.13700-0.13770	0.63-0.30	0.1364
2.05	0.070	28^3x56	0.13470-0.13560	0.63-0.78	0.13510, 0.13540	40^3x80			

Algorithm tests

Y. kuramashi, algorithm session 1

- Beta=1.90, 16^3x32, 1000-2000 trajectories

kappa_ud	mpi/mrho	(N0,N1,N2)	Npoly
0.13700	0.62(1)	(4,5,6)	130
0.13741	0.5	(4,5,8)	140
0.13759	0.4	(4,5,12)	140
0.13770	0.3	(4,5,14)	140

- Examined
  - Magnitude of force
  - Auto-correlations
  - Acceptance, etc

→ If 4 Tflops sustained, perhaps feasible to plan
a half year for 10^4 trajectories at beta=1.90?
2 years for three beta values and the continuum limit?

summary

- *PACS-CS, a 14.3 Tflops large-scale cluster, in place at University of Tsukuba*
- *Plan to complete the improved Wilson-clover program toward light quarks exploiting the domain decomposition acceleration idea*
- *Hope to start production when we go back to Tsukuba*