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

X-switch
Y-switch
Z-switch
Computing node

Communication via single switch
communication via multiple switches

In the figure
Dual GBEthernet link for band width
## 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   | 100m² |
| power       | 545kW |
Board layout: 2 nodes /1U board

160GB x 2

LV Xeon @ 2.8GHz

CPU

chip-set

memory 2GB

RAS (GbE) I/O (GbE)

3d HXB (GbE x 6)

x0, x1: dual link for X-crossbar

y0, y1: dual link for Y-crossbar

z0, z1: dual link for Z-crossbar

HDD (RAID-1)

Node image on 1U board

front

unit-0

HDD

Power Unit

unit-1

HDD

back

x0 x1 y0 y1 z0 z1

File I/O network

System diagnostics and control
Node block diagram

LV-Xeon
2.8GHz EM64T

FSB800 6.4GB/s

Main memory 2GB DDR400 2way interleave

64bit PCI controller
PXH

64bit PCI controller
PXH

Bridge controller
6300 ESB

S-ATA
HDD 160GB
S-ATA
HDD 160GB

LAN
i82546EB
LAN
i82546EB
LAN
i82546EB
LAN
i82546EB

BMC

SVGA
RAGE-XL

SVGA(15pin)

1000base-Tx8

I/O

RAS

HXB network

HXB network
Mother board

- CPU socket
- Chipset E7520
- Memory 1GB x2
- PCI controller
- PCI controller
- NIC x 8
- GB port x 8
2 mother boards on a single 1U board

Raid 1 Disk
130GByte \times 2 \text{ /node}

Motherboard; 1 node

Raid 1 Disk
130GByte \times 2 \text{ /node}

Power supply

Motherboard; 1 node
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
Alltogether 10 units arranged in 5 rows
PACS-CS software

- OS
  - Linux
  - 3D HXB driver based on SCore PMv2

- Programming
  - MPI for communication
  - Library for 3D HXB network
  - Fortran, C, C++

- Job execution
  - System partition (256 nodes, 512 nodes, 1024 nodes, …)
  - Batch queue using PBS
  - Job scripts for file I/O
some partitions

PU512S[1–5] 16x16x2

PU256C[1–8] 8x8x4

PU256S[1–10] 16x16x1

PU512C[1–4] 8x8x8
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

\[
\text{flop} \cdot \text{I/O} = 1896 \cdot 5088 = 2.68 \text{ Byte/flop}
\]

- Since max I/O possible is 6.4GByte/s, max possible flops = 6.4/2.68 = 2.39 Gflops (37.3%)

<table>
<thead>
<tr>
<th>flop</th>
<th>Load (Byte)</th>
<th>Store (Byte)</th>
<th>Byte/flop</th>
</tr>
</thead>
<tbody>
<tr>
<td>mult</td>
<td>U</td>
<td>p</td>
<td>q</td>
</tr>
<tr>
<td>t</td>
<td>168</td>
<td>288</td>
<td>288</td>
</tr>
<tr>
<td>x</td>
<td>144</td>
<td>288</td>
<td>336</td>
</tr>
<tr>
<td>y</td>
<td>144</td>
<td>288</td>
<td>336</td>
</tr>
<tr>
<td>z</td>
<td>144</td>
<td>288</td>
<td>336</td>
</tr>
<tr>
<td>clover</td>
<td>288</td>
<td>600</td>
<td>672</td>
</tr>
<tr>
<td>total</td>
<td>888</td>
<td>1896</td>
<td>1824</td>
</tr>
</tbody>
</table>
Network performance/expectations

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

- expected (BiCGStabL2)
  - Calculation: 17.79 msec, 71.6%
  - Neighbor communication: 6.24 msec, 25.1%
  - Global sum: 0.81 msec, 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: Ave. 586.8 (78.2%), Min. 559.2 (74.6%), Max. 619.3 (82.6%)
  - 512node: Ave. 582.0 (77.6%), Min. 434.0 (57.9%), Max. 629.6 (84.0%)

- Global sum (MPI_Allreduce) (msec)
  - 8B: Ave. 0.420, Min. 0.344, Max. 0.727
  - 800kB: Ave. 257, Min. 491, Max. 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.,
    \[ \frac{m_\pi}{m_\rho} \approx 0.3 \quad i.e., \quad \frac{m_{ud}}{m_s} \approx 0.15 \quad \text{or so} \]
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**

<table>
<thead>
<tr>
<th>beta</th>
<th>a(fm)</th>
<th>lattice size</th>
<th>kappa_ud</th>
<th>mpi/mrhol</th>
<th>kappa_s</th>
<th>lattice size</th>
<th>kappa_ud</th>
<th>mpi/mrhol</th>
<th>kappa_s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.83</td>
<td>0.122</td>
<td>16^3x32</td>
<td>0.13655–0.13825</td>
<td>0.61–0.78</td>
<td>0.13710, 0.13760</td>
<td>24^3x48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.90</td>
<td>0.100</td>
<td>20^3x40</td>
<td>0.13580–0.13700</td>
<td>0.63–0.77</td>
<td>0.13580, 0.13640</td>
<td>32^3x64</td>
<td>0.13700–0.13770</td>
<td>0.63–0.30</td>
<td>0.1364</td>
</tr>
<tr>
<td>2.05</td>
<td>0.070</td>
<td>28^3x56</td>
<td>0.13470–0.13560</td>
<td>0.63–0.78</td>
<td>0.13510, 0.13540</td>
<td>40^3x80</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Algorithm tests

Y. kuramashi, algorithm session 1

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

<table>
<thead>
<tr>
<th>kappa_ud</th>
<th>mpi/mrho</th>
<th>(N0,N1,N2)</th>
<th>Npoly</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.13700</td>
<td>0.62(1)</td>
<td>(4,5,6)</td>
<td>130</td>
</tr>
<tr>
<td>0.13741</td>
<td>0.5</td>
<td>(4,5,8)</td>
<td>140</td>
</tr>
<tr>
<td>0.13759</td>
<td>0.4</td>
<td>(4,5,12)</td>
<td>140</td>
</tr>
<tr>
<td>0.13770</td>
<td>0.3</td>
<td>(4,5,14)</td>
<td>140</td>
</tr>
</tbody>
</table>

- 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