Heterogeneous Multi-Computer System
A New Platform for Multi-Paradigm Scientific Simulation

Taisuke Boku, Hajime Susa,
Masayuki Umemura, Akira Ukawa
Center for Computational Physics, University of Tsukuba

Junichiro Makino, Toshiyuki Fukushige
Department of Astronomy, University of Tokyo
Outline

• Background
• Concept and Design of HMCS prototype
• Implementation of prototype
• Performance evaluation
• Computational physics result
• Variation of HMCS
• Conclusions
Background

• Requirements to Platforms for Next Generation Large Scale Scientific Simulation
  – More powerful computation power
  – Large capacity of Memory, Wide bandwidth of Network
  – High speed & Wide bandwidth of I/O
  – High speed Networking Interface (outside)
  – …

• Is it enough? How about the quality?
• Multi-Scale or Multi-Paradigm Simulation
Multi-Scale Physics Simulation

- Various level of interaction
  - Newton Dynamics, Electro-Magnetic interaction, Quantum Dynamics, ...
- Microscopic and Macroscopic Interactions
- Difference in Computation Order
  - $O(N^2)$: ex. N-body
  - $O(N \log N)$: ex. FFT
  - $O(N)$: ex. straight-CFD
- Combining these simulation, Multi-Scale or Multi-Paradigm Computational Physics is realized
HMCS – Heterogeneous Multi-Computer System

- Combining Particle Simulation (ex: Gravity interaction) and Continuum Simulation (ex: SPH) in a Platform
- Combining General Purpose Processor (flexibility) and Special Purpose Processor (high-speed)
- Connecting General Purpose MPP and Special Purpose MPP via high-throughput network
- Exchanging particle data at every time-step

Prototype System: CP-PACS + GRAPE-6
(JSPS Research for the Future Project “Computational Science and Engineering”)
Block Diagram of HMCS

MPP for Continuum Simulation (CP-PACS)

Parallel I/O System PAVEMENT/PIO

MPP for Particle Simulation (GRAPE-6)

Parallel File Server (SGI Origin2000)

Paralel File Server (SGI Onyx2)

100base-TX Switches

32bit PCI \times N

Hybrid System Communication Cluster (Compaq Alpha)

Parallel Visualization System PAVEMENT/VIZ

Parallel Visualization Server (SGI Onyx2)
PACS

- Pseudo Vector Processor with 300 Mflops of Peak Performance \times 2048 \Rightarrow 614.4 \text{ Gflops}
- I/O node with the same performance \times 128
- Interconnection Network: 3-D Hyper Crossbar (300MB/s / link)
- Platform for General Purpose Scientific Calculation
- 100base-TX NIC on 16 IOUs for outside comm.
- Partitioning is available (Any partition can access any IOU)
- Manufactured by Hitachi Co.
- Operation from 1996.4 with 1024 PUs, from 1996.10 with 2048 PUs
CP-PACS
(Center for Computational Physics)
GRAPE-6

- The 6th generation of GRAPE (Gravity Pipe) Project
- Gravity calculation for many particles with 31 Gflops/chip
- 32 chips / board ⇒ 0.99 Tflops/board
- 64 boards of full system is under implementation ⇒ 63 Tflops
- On each board, all particles (j-particles) data are set onto SRAM memory, and each target particle (i-particle) data is injected into the pipeline and acceleration data is calculated
- Gordon Bell Prize at SC01 Denver
GRAPE-6
(University of Tokyo)

8 board \times 4 \text{ system}

GRAPE-6 board (32 chips)
GRAPE-6 (cont’d)

(Top View)   (Bottom View)

Daughter Card Module (4 chip / module)
Host Computer for GRAPE-6

- GRAPE-6 is not a stand-alone system
  ⇒ Host computer is required
- Alpha CPU base PC (Intel x86, AMD Ahtlon are also available)
- Connected via 32bit PCI Interface Card to GRAPE-6 board
- A host computer can handle several GRAPE-6 boards
- It is impossible to handle an enormous number of particles with a single host computer for complicated calculation
Hyades (Alpha CPU base Cluster)

- Cluster with Alpha 21264A (600MHz) × 16 node
- Samsung UP1100 (single CPU) board
- 768 MB memory / node
- Dual 100base-TX NIC
- 8 nodes are equipped with GRAPE-6 PCI card
  ⇒ Cooperative work with 8 GRAPE-6 boards under MPI programming
- One of 100base-TX NICs is connected with CP-PACS via PIO (Parallel I/O System)
- Linux RedHat 6.2 (kernel 2.2.16)
- Operated as a data exchanging and controlling system to connect CP-PACS and GRAPE-6
PAVEMENT/PIO

- Parallel I/O and Visualization Environment
- Connecting multiple parallel processing platforms with commodity-based parallel network
- Automatic and dynamic load balancing feature to utilize spatial parallelism for applications
- Utilizing multiple I/O processors of MPP not to make bottleneck in communication
- Providing easy-to-program API with various operation modes (user-oriented, static or dynamic load balancing)
MPP – DSM system example

CP-PACS

I/O processor (PIO server)
Calculation processor (user process)

SMP or Cluster

Switch

PIO server
User process (thread)
HMCS Prototype

Massively Parallel Processor
CP-PACS
(2048 PUs, 128 IOUs)

Parallel Visualization Server
SGI Onyx2 (4 Processors)

Parallel File Server
SGI Origin-2000 (8 Processors)

Switching HUB \times 2

Parallel
100Base-TX Ethernet

GRAPE-6 & Hyades
(16 node, 8 board)

8 links
SPH (Smoothed Particle Hydrodynamics)

Representing the material as a collection of particles

\[ \rho(r_i) = \sum_j \rho_j W(|r_i - r_j|) \]

\( W \) : kernel function
Accurate calculation of optical depth along light paths required.
Use the method by Kessel-Deynet & Burkert (2000).

\[ \tau_{TS} = \sum_i \frac{\sigma}{2} \left( n_{E_i} + n_{E_{i+1}} \right) \left( s_{E_{i+1}} - s_{E_i} \right) \]
SPH Algorithm with Self-Gravity Interaction

- Gravity
- GRAPE-6 calculation \(O(N^2)\)
- SPH (Density)
- Radiation Trans.
- Chemistry
- Temperature
- Pressure Calculation
- Newton Dynamics
- CP-PACS \(O(N)\)
- Comm. \(O(N)\)

Iterative process: SPH (Density) → Radiation Trans. → Chemistry → Temperature → Pressure Calculation → Newton Dynamics → Gravity

Computation complexity:
- GRAPE-6: \(O(N^2)\)
- CP-PACS: \(O(N)\)
g6cpplib – CP-PACS API

- `g6cpp_start(myid, nio, mode, error)`
- `g6cpp_unit(n, t_unit, x_unit, eps2, error)`
- `g6cpp_calc(mass, r, f_old, phi_old, error)`
- `g6cpp_wait(acc, pot, error)`
- `g6cpp_end(error)`
Performance (raw – G6 cluster)

- GRAPE-6 cluster performance with dummy data (without real RT-SPH)
- GRAPE-6 board × 4 with 128K particles

<table>
<thead>
<tr>
<th>process</th>
<th>particle data trans.</th>
<th>all-to-all data circulation</th>
<th>set-up data in SRAM</th>
<th>N-body comp.</th>
<th>result return</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>1.177</td>
<td>0.746</td>
<td>0.510</td>
<td>0.435</td>
<td>0.085</td>
</tr>
</tbody>
</table>

Processing time for 1 iteration = 3.24 sec (total)
Scalability with problem size

<table>
<thead>
<tr>
<th>process</th>
<th>n=15</th>
<th>n=16</th>
<th>n=17</th>
</tr>
</thead>
<tbody>
<tr>
<td>data trans.</td>
<td>5.613</td>
<td>10.090</td>
<td>17.998</td>
</tr>
<tr>
<td>all-to-all circulation</td>
<td>0.309</td>
<td>0.476</td>
<td>0.681</td>
</tr>
<tr>
<td>set data to SRAM</td>
<td>0.231</td>
<td>0.362</td>
<td>0.628</td>
</tr>
<tr>
<td>calculation</td>
<td>0.064</td>
<td>0.169</td>
<td>0.504</td>
</tr>
<tr>
<td>TOTAL</td>
<td>6.217</td>
<td>11.097</td>
<td>19.811</td>
</tr>
</tbody>
</table>

# of particles \( N = 2^n \)  (#P=512)
Scalability with # of PUs

<table>
<thead>
<tr>
<th>process</th>
<th>#P=512</th>
<th>#P=1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>data trans.</td>
<td>17.998</td>
<td>10.594</td>
</tr>
<tr>
<td>all-to-all circulation</td>
<td>0.681</td>
<td>0.639</td>
</tr>
<tr>
<td>set data to SRAM</td>
<td>0.628</td>
<td>0.609</td>
</tr>
<tr>
<td>calculation</td>
<td>0.504</td>
<td>0.503</td>
</tr>
<tr>
<td>TOTAL</td>
<td>19.811</td>
<td>12.345</td>
</tr>
</tbody>
</table>

# of particles (N) = 2^{17}

RT-SPH calculation is included
Example of Physics Results
(64K SPH particles + 64K dark matters)
Various implementation methods of HMCS

• HMCS-L (Local)
  – Same as current prototype
  – Simple, but the system is closed

• HMCS-R (Remote)
  – Remote access to GRAPE-6 server through Network (LAN or WAN = Grid)
  – Utilizing GRAPE-6 cluster in time-sharing manner as Gravity Server

• HMCS-E (Embedded)
  – Enhanced HMCS-L: Each node of MPP (or large scale cluster) is equipped with GRAPE chip
  – Combining wide network bandwidth of MPP (or cluster) and powerful node processing power
HMCS-R on Grid

Remote access to GRAPE-6 server via g6cpp API
no persistency on particle data – suitable for Grid
$O(N^2)$ of calculation with $O(N)$ of data amount
High Speed Network Switch

- Local comm. between general purpose and special purpose processors
- Utilizing wide bandwidth of large scale network
- Ideal fusion of flexibility and high performance
Conclusions

- HMCS – Platform for Multi-Scale Scientific Simulation
- Combining General Purpose MPP (CP-PACS) and Special Purpose MPP (GRAPE-6) with parallel network under PAVEMENT/PIO middleware
- SPH + Radiation Transfer with Gravity Interaction ⇒ Detailed simulation for Galaxy formation
- 128K particle real simulation with 1024PU CP-PACS makes new epoch of simulation
- Next Step: HMCS-R and HMCS-E