HMCS-G : Grid-enabled Hybrid Computer System for Computational Astrophysics

Taisuke Boku*, Mitsuhisa Sato*, Kenji Onuma*
Jun'ichiro Makino**, Hajime Susa***, Daisuke Takahashi*
Masayuki Umemura*, Akira Ukawa*
**Center for Computational Physics, University of Tsukuba*
***Graduate School of Science, University of Tokyo*
****Department of Physics, Rikkyo University*

Outline

- Background (HMCS or HMCS-L)
- Concept of HMCS-G
- Design issues & Implementation
- Performance Evaluation
- Demonstration (?)
- Summary & Future Works

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

Basic concept of HMCS

- Computational power required for computational physics
  - In the state-of-the-art computational physics, various physical phenomena have to contribute for detailed and precise simulation
  - Some of them require enormous computational power in large problem size
    - FFT: O(N log$N$)
    - gravity, molecular dynamics: O(N^2)
    - nano-scale material: O(N^3), O(N^4), ...
  - Ordinary general purpose machine is not enough in many cases
  - Requirement of Special Purpose Machines

Basic concept of HMCS (cont'd)

- General Purpose Machines:
  Variety of algorithm and easy programming for multi-purpose utilization
- Special Purpose Machines:
  Absolute computational power with very limited type of calculation
- We need both !

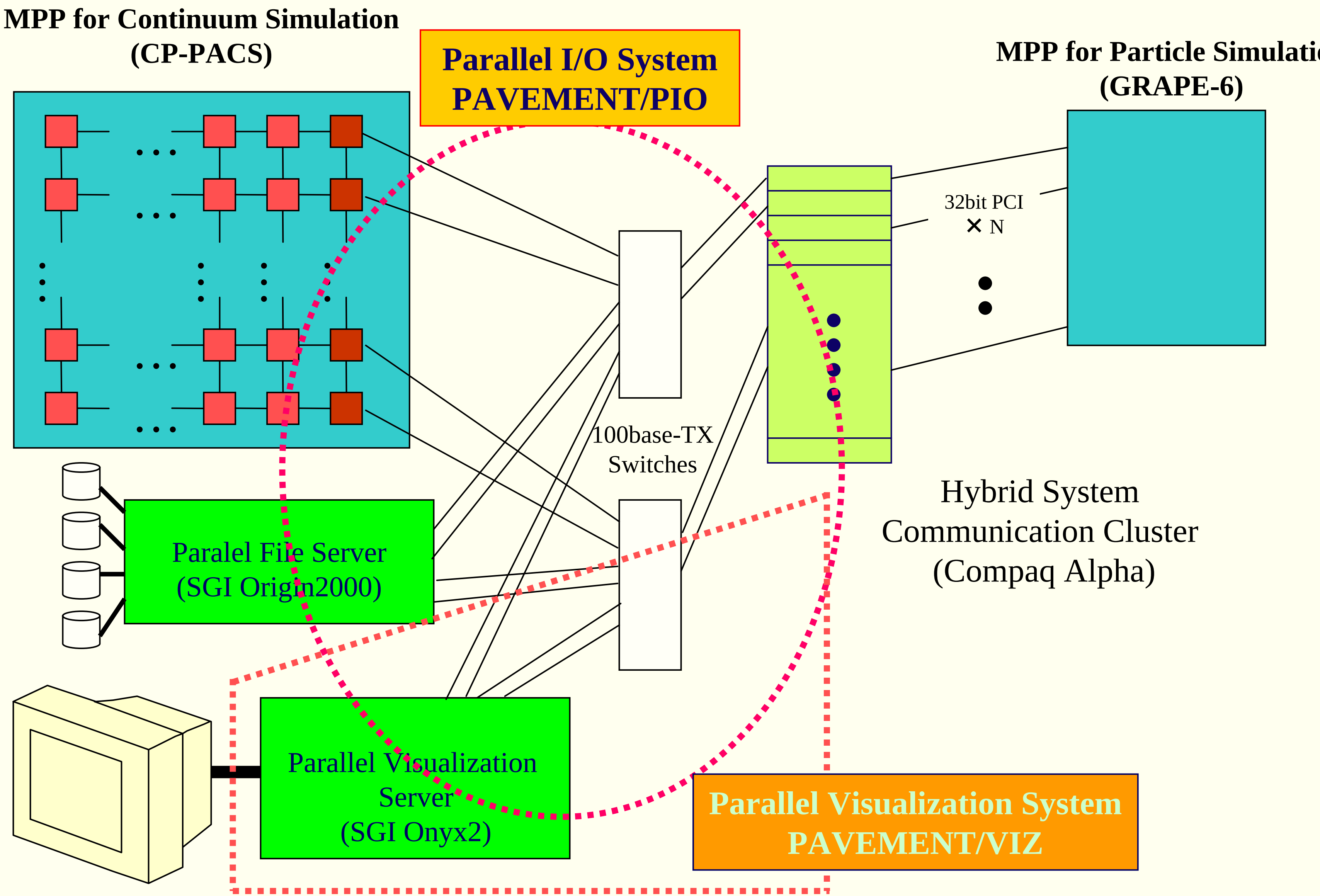
Heterogeneous Multi-Computer System
combining high speed computation power
with high bandwidth network

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 with 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



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 installed in University of Tokyo ⇒ 63 Tflops
- On each board, all particles data are set onto SRAM memory, and each target particle data is injected into the pipeline, then acceleration data is calculated
- Gordon Bell Prize at SC2000, SC2001 (Prof. Makino, U. Tokyo) also nominated at SC2002

CP-PACS

- General purpose MPP with 2048 PU + 128 IOU
- CPU: PVP-SW (pseudo vector processing with sliding window) feature for vector proc. with 300 MFLOPS peak performance
- Network: 3-D HXB (hyper-crossbar) with 300 MB/s/link
- I/O: 8 GB RAID5 disk ✕ 128 = 1TB, 100MB/s HIPPI, 100base-TX Ethernet ✕ 16
- Total peak performance: 614.4 GFLOPS
  ⇒ No. 1 in TOP500 list at Nov. 1996

Clusters in CCP (Center for Computational Physics)

- Alpha based system
  - hyades: 600 MHz single 21264, 16 nodes, 100base-TX Ethernet
  - orion: 833 MHz dual 21264, 29 nodes, dual 100base-TX Ethernet
- IA-32 based system
  - perseus: 2.8 GHz dual Xeon, 36 nodes, Myrinet2000
- Misc.
  - alice: 1800+ dual AthlonMP, 17 nodes
  - cecily: 800 MHz quad IA-64 (Itanium), 4 nodes
  - dennis: 2.4 GHz dual Xeon, 1000base-T Ethernet
  - ...

What is HMCS-G ?

Importance of Grid for Computational Physics

- Efficient utilization of world-wide generic HPC resources (MPP, cluster, storage, network, etc.)
  = Quantitative Contribution
- Sharing special purpose machines installed to small number of institutes from all over the world
  = Qualitative Contribution

↓

HMCS-G is a concept of hybrid computational system to combine General purpose and Special purpose machines based on Grid-RPC

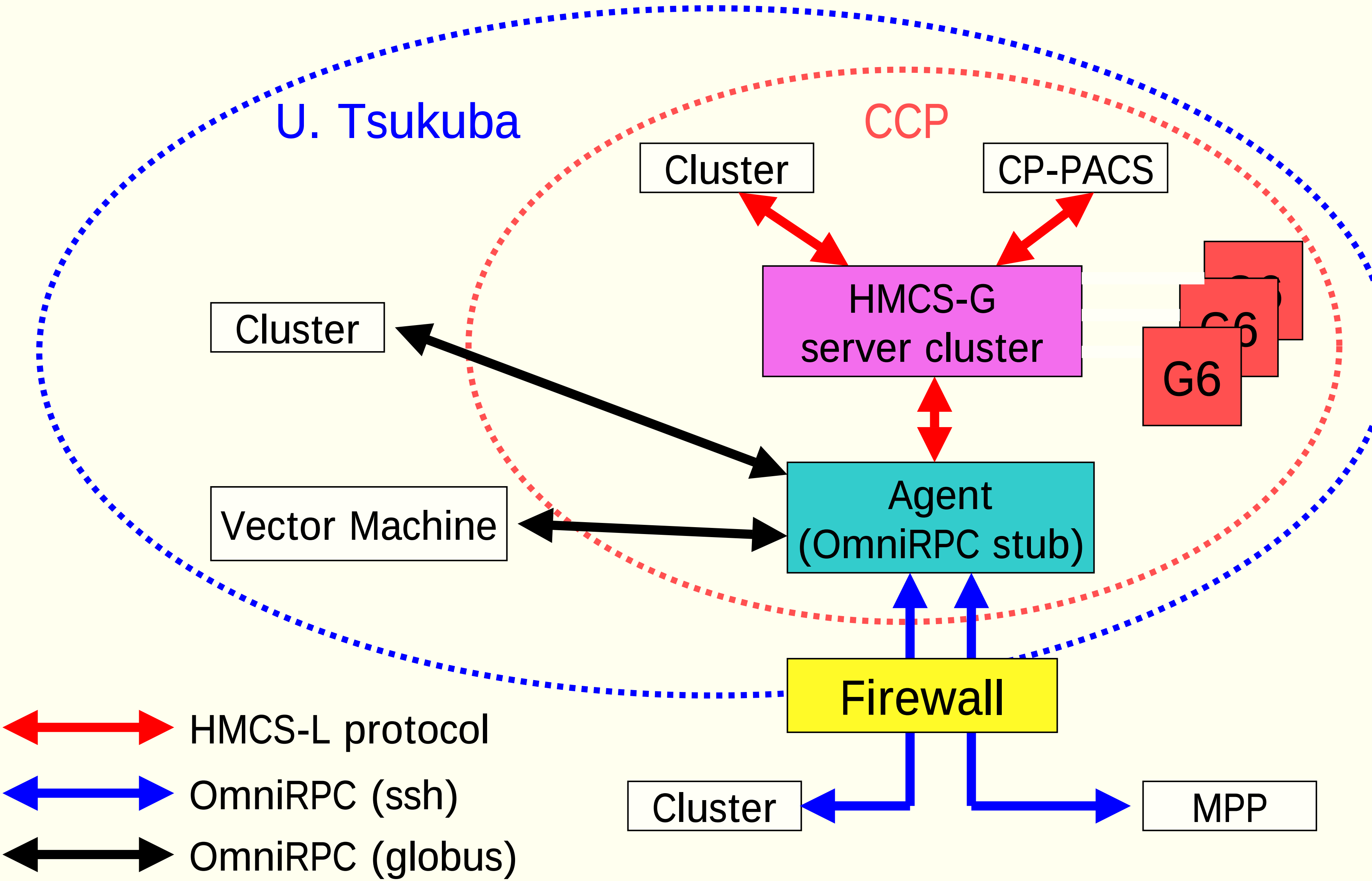
HMCS-G: Grid-enabled HMCS (for Gravity

- Purpose
  - Sharing special purpose machine GRAPE-6 among users over the world who needs gravity calculation
  - Providing local and remote services of GRAPE-6 facility access simultaneously
  - Utilizing GRAPE-6 resource efficiently
  - Hiding long network access latency through communication buffers between end-points

HMCS-G: Grid-enabled HMCS (cont'd)

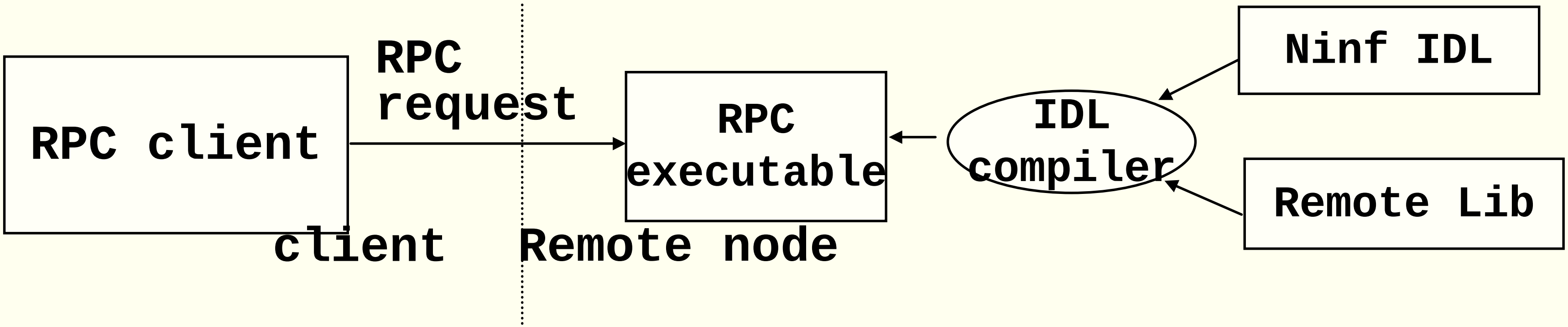
- Implementation
  - Multiple clients support for GRAPE-6 server cluster
  - OmniRPC: Grid-enabled RPC
  - Authentication: globus & ssh

HMCS-G block diagram



OmniRPC

- A thread-safe RPC based on the Ninf grid RPC
  - RPC executable (library programs)
    - **Libraries wrapped with stub programs for RPC.**
    - **Generated by Ninf IDL.**
    - **Invoked by RPC request.**
  - Programming interface:
    - **A simple language-independent interface: `OmniRpcCall`**
    - **Ease-to-use and familiar-looking for existing programming languages such as C, Fortran, …**



OmniRPC Basic API

- **A simple language-independent interface:** OmniRpcCall
  - OmniRpcCall(**FUNC_NAME**, ...);
- **Ease-to-use and familiar-looking for existing programming languages such as C, Fortran, …**

```
double A[n][n],B[n][n],C[n][n]; /* Data Decl.*/

dmmul(n,A,B,C);                /* Call local function*/
```

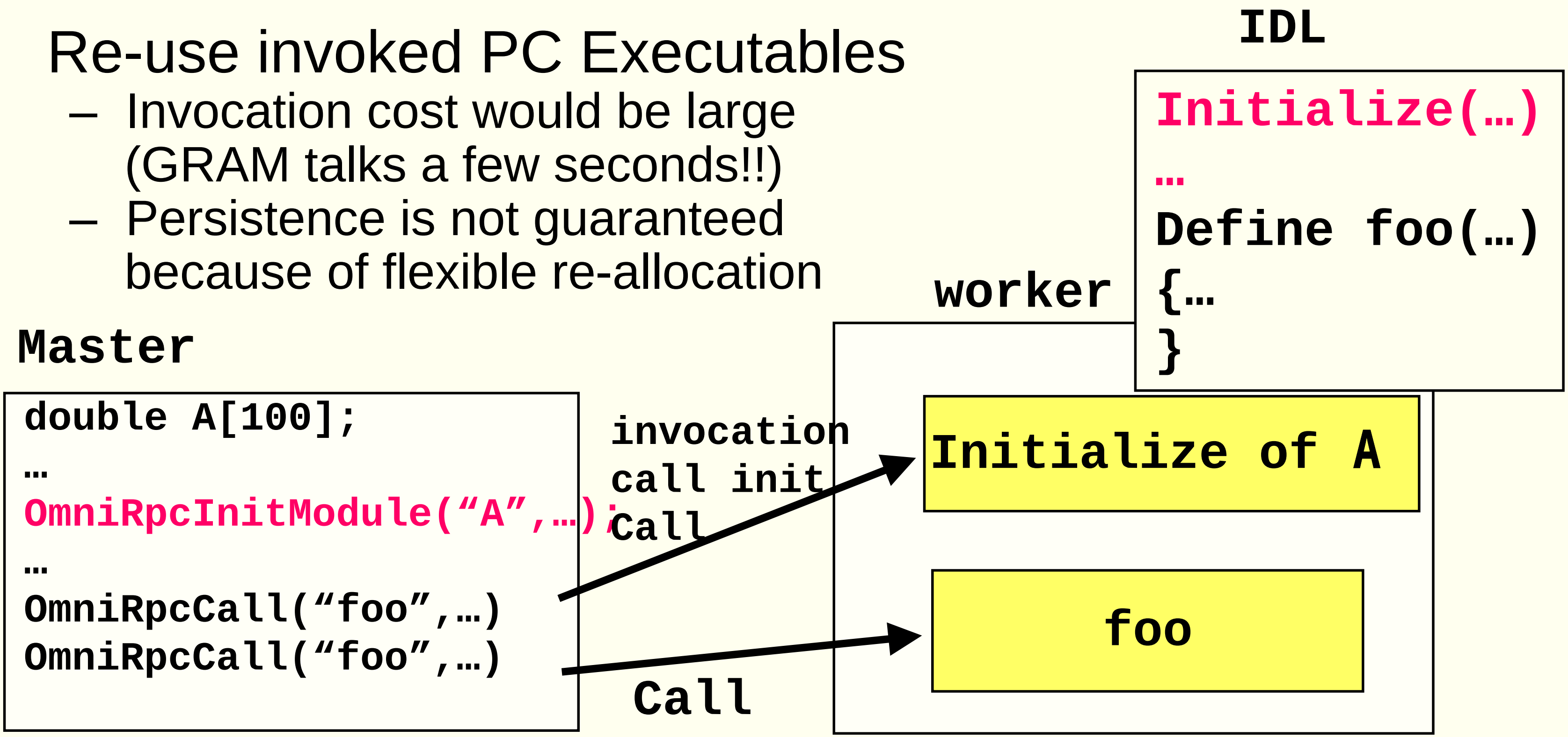
"Gridify"

```
OmniRpcCall("dmmul",n,A,B,C); /*  Call Ninf Func */
```

Persistence Model of OmniRPC

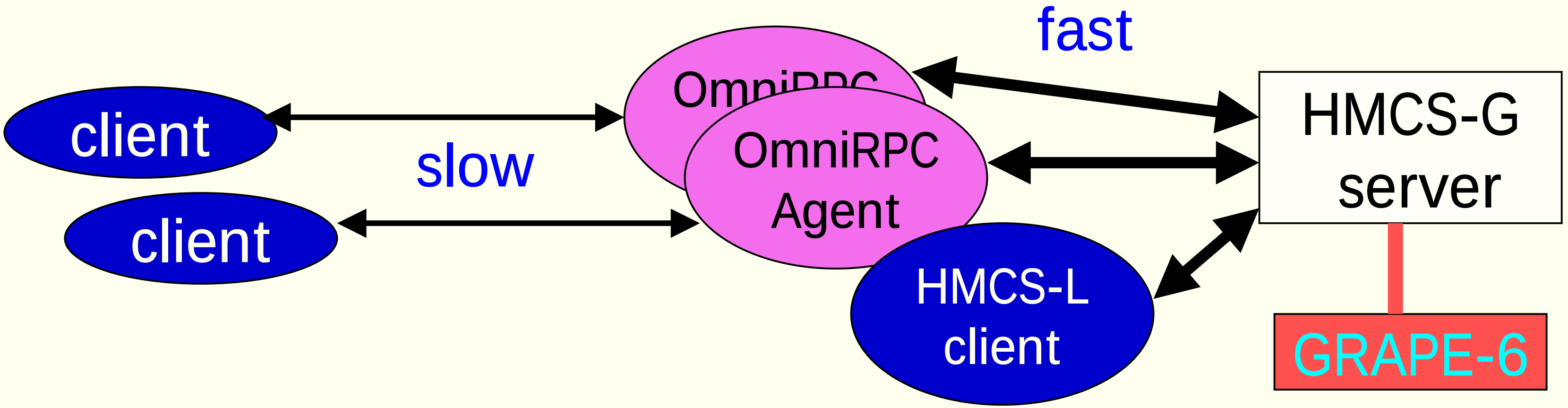
- Automatic initializable remote module
  - Limited persistence model between initialization and each RpcCall
  - Useful for master-workers models, sharing the same data between RpcCall
  - API: OmniRpcInitModule

- Re-use invoked PC Executables
  - Invocation cost would be large (GRAM talks a few seconds!!)
  - Persistence is not guaranteed because of flexible re-allocation



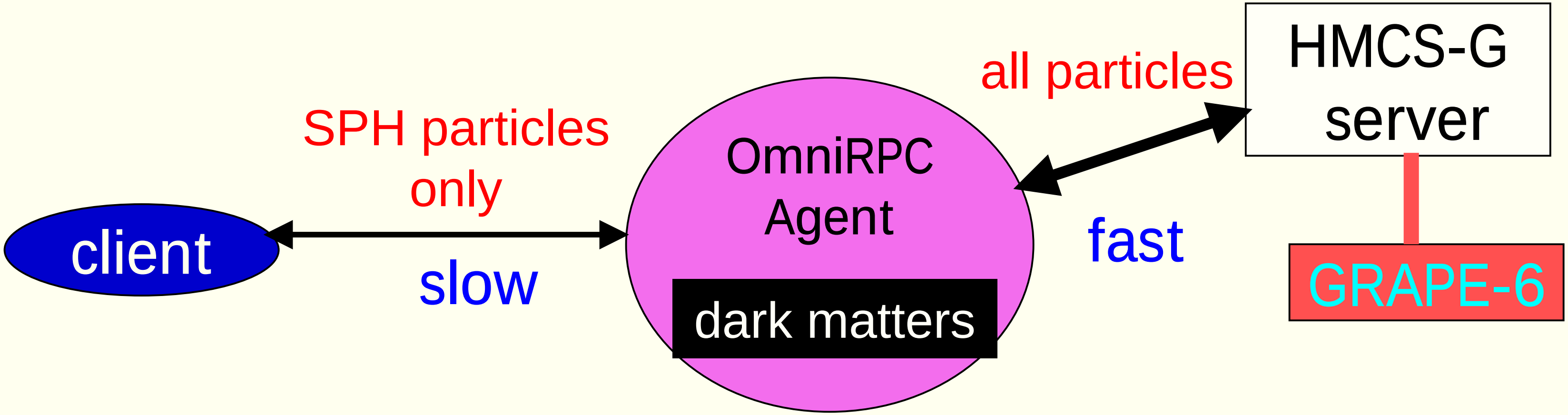
Agent (OmniRPC stub)

- Authentication and Accounting
- Various benefits for application users
  - easy API
  - globus option for de facto standard authentication
  - ssh option for easy system installation
- Data buffering and communication speed-gap absorbing
- Co-working with original HMCS-L clients



Dark matter localization

- In our RT-SPH (Radiative Transfer with Smoothed Particle Hydrodynamics) simulation model, a half of particles are dark-matters
- They do not need to be calculated with SPH particles on client side
- Agent keeps dark-matters and does not exchange them with MPP (cluster) by persistency feature of OmniRPC
  ⇒ Data transfer amount is reduced



Misc. features for Grid-enabling

- Previous version (HMCS-L) connects just a pair of client and server
  ⇒ No authentication, communication phase control, robustness
- HMCS-G provides
  - Authentication
    - ***globus & ssh***
    - ***globus = de facto standard of Grid***
    - ***ssh = easy implementation, no firewall problem***
  - Resource scheduling
    - ***Do not lock GRAPE-6 server for long time***
    - ***Fine grained phase control***
  - Robustness
    - ***Time-out mechanism to detect network or client failure***

Programming

- User Level:
  call gg6_calc(n, unit_t, unit_x, eps2)

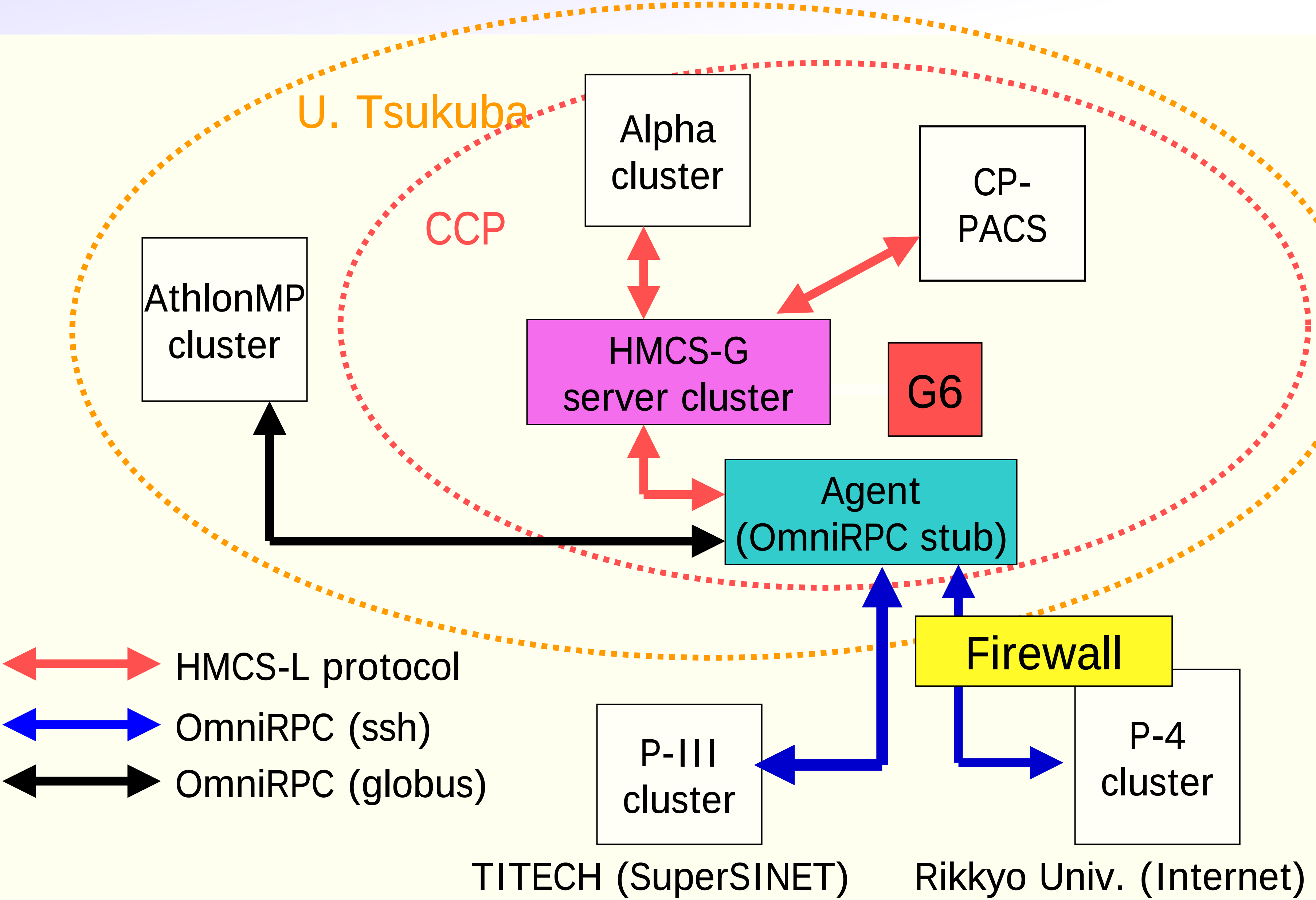
- OmniRPC Client Level:
  OmniRpcExecCall(handle, "gg6_unit", *n, *newunit_t, *newunit_x, *eps2, error);

- Network stub (Agent) Level (= HMCS-L):
  rg6_calc(n, unit_t, unit_x, error);

API (current version)

- **gg6_init(char *agent, int key)**
  initialize & specify agent
- **gg6_start(int nio, int mode)**
  specify # of nodes, utilization mode (currently, only mode 1)
- **gg6_unit(int np, int unit_t, int unit_x)**
  specify # of particles and magnitude of calc.
- **gg6_calc1(double mass[], double x[][3], double f_old[], double phiold[])**
  request actual calculation
- **gg6_wait1(double acc[][3], double f[])**
  retrieve calculation result
- **gg6_end()**
  end of calculation

Current Environment



Network condition

- CCP local
  - 1000base-SX on backbone, 100base-TX for leaves
- Inside the university
  - 1000base-LX on backbone, 100base-TX for leaves
- Between U. Tsukuba and TITECH (Tokyo Inst. of Tech.)
  - SuperSINET (1Gbps dedicated link)
- Between U. Tsukuba and Rikkyo U.
  - Commodity Internet (b/w ??)

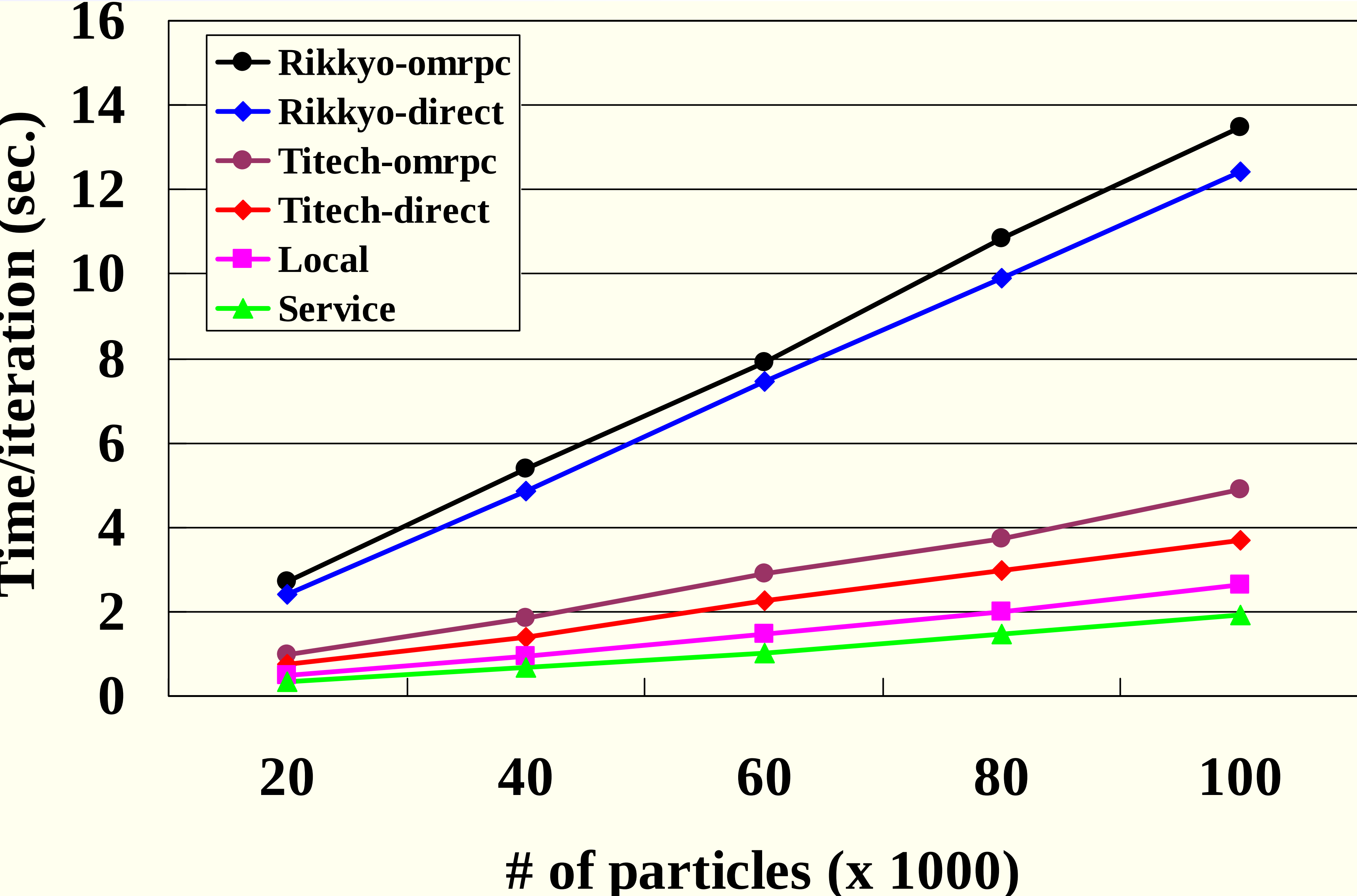
Performance results

- GRAPE-6 pure computation time with 1 node
  - 0.8 sec for 64K particles
  - 2.1 sec for 128K particles

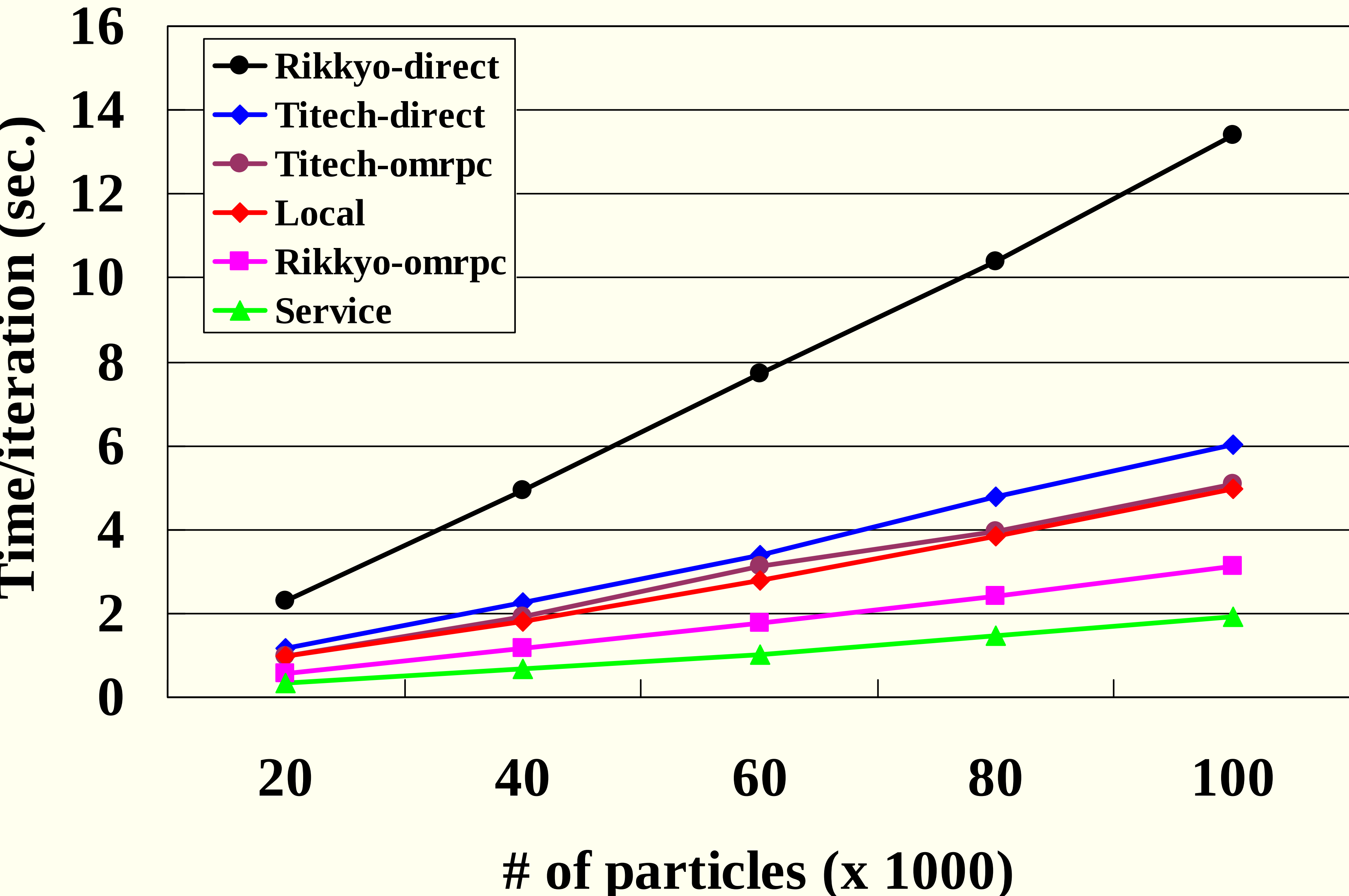
Computation load for 128K particles: 1.3 TFLOP

- Turn around time for 1 time step with 64K particles [total] (communication)
  - [1.2 sec] (0.4 sec) (local, direct)
  - [1.7 sec] (0.9 sec) (university, OmniRPC-globus)
  - [2.1 sec] (1.3 sec) (university, OmniRPC-ssh)
  - [2.8 sec] (2.0 sec) (TITECH, OmniRPC-ssh) *SuperSINET*
  - [9.1 sec] (8.3 sec) (Rikkyo, OmniRPC-ssh) *Internet*

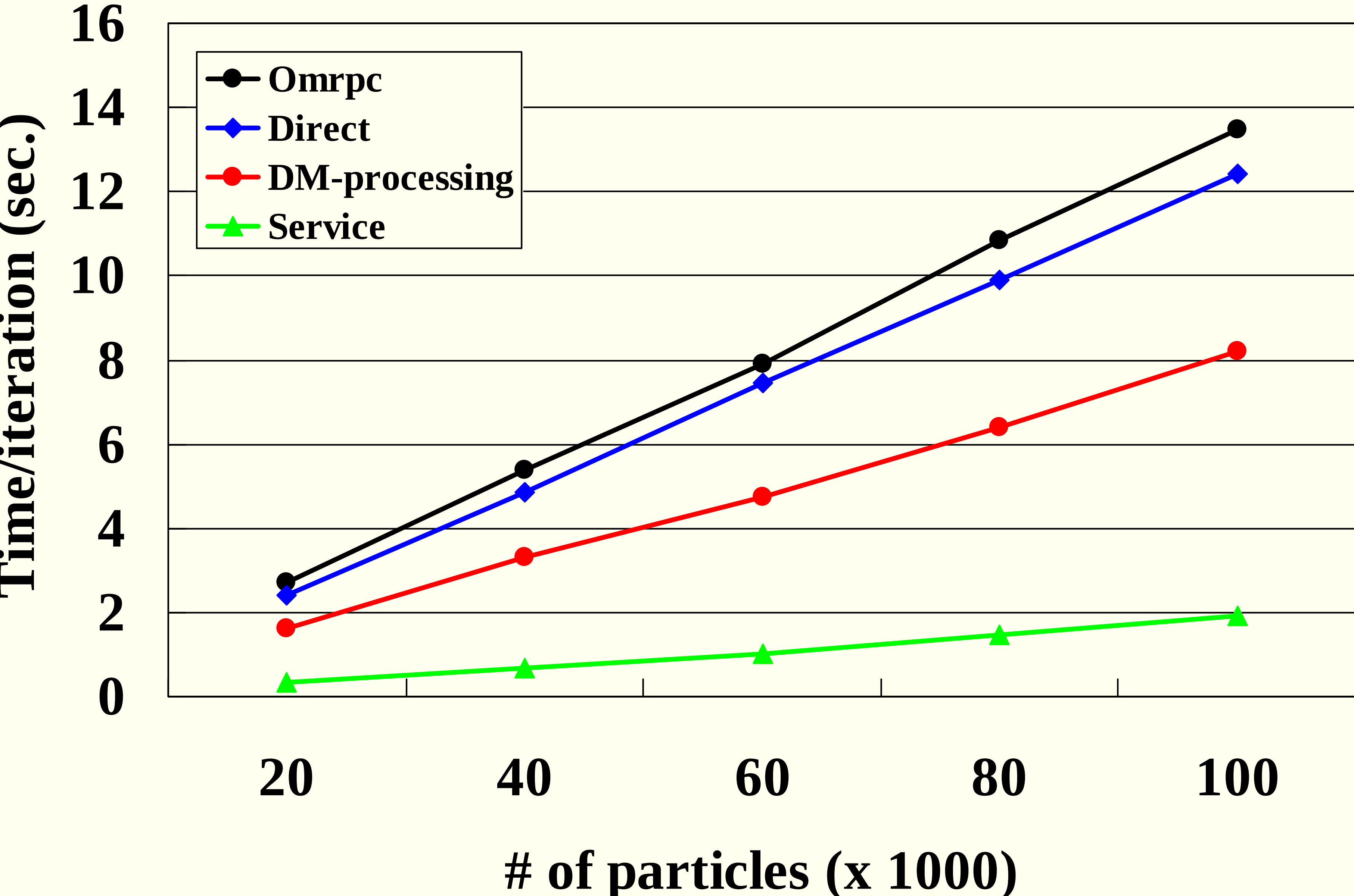
Single client execution time



Local client interleaved by another



Dark-Matter localization (Rikkyo)



On-line Demonstration

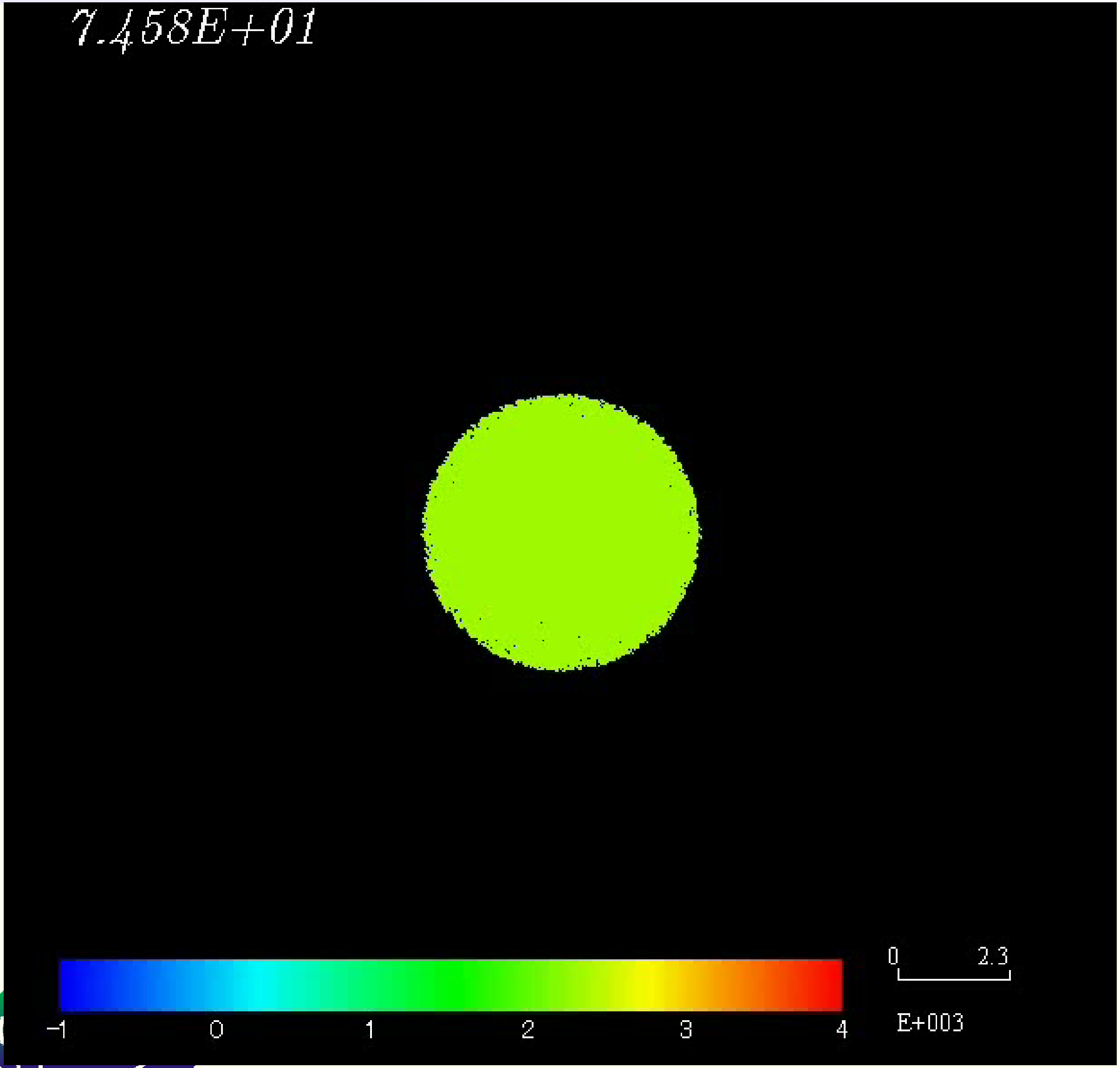
- Run simple client process to compute 10,000 particles with same mass for gravity calculation only on U. Tsukuba and TITECH
- Transfer gravity calculation request to Agent of HMCS-G running in Center for Computational Physics, U. Tsukuba
- Show particles movement on 2-D mapping (actual calculation is performed in 3-D)
- Multiple processes can be served simultaneously

Simulation of Galaxy Formation

Simulation of Galaxy Formation based on RT-SPH

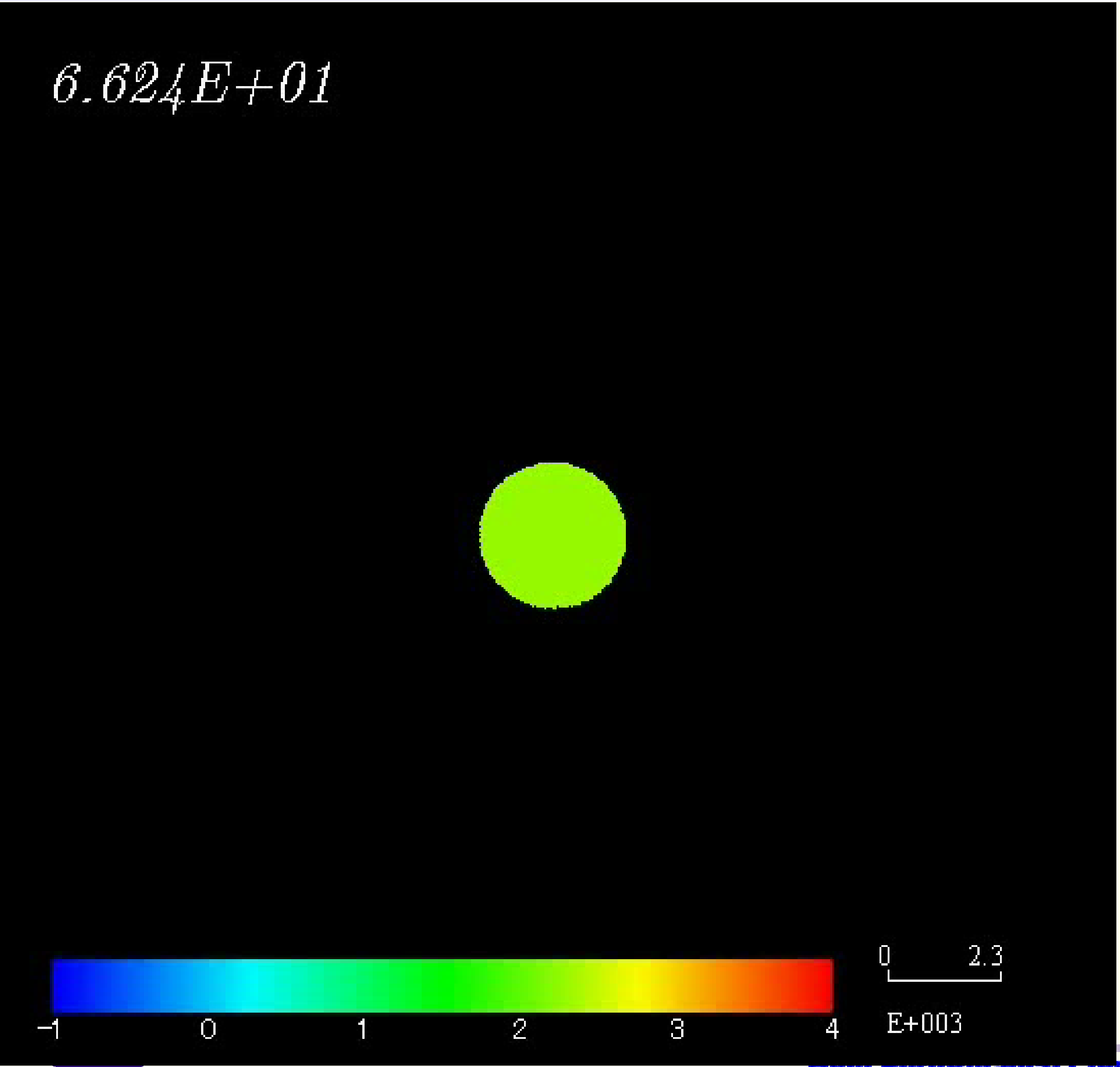
- Smoothed Particle Hydrodynamics with Radiative Transfer (RT-SPH) under Gravity
- Combination of hydro-dynamics computation and gravity calculation
  - Cluster and MPP for RT-SPH
  - GRAPE-6 for gravity
- RT-SPH with 128K particles for 40,000 time steps takes approximately 60 hours with 32 CPUs of Alpha 21264 cluster (DS20L base, 833 MHz)
- Gravity calculation with GRAPE-6 including communication takes 11 hours

Simulation Result (High-z, High-mass)



Blue: "Stars"
Other Color: Temperature

Simulation Result (Low-z, Low-mass)



Blue: “Stars”
Other Color: Temperature

Summary

- HMCS-G enables world-wide utilization of special purpose machine (GRAPE-6) based on OmniRPC
- Multi-physics simulation is very important in next generation computational physics
- This platform concept is expandable to various special purpose systems
- OmniRPC/ssh is very easy to implement for pure application users as well as OmniRPC/globus

Future works

- More efficient GRAPE-6 resource allocation for various size of problems
- Aggregation of multiple GRAPE-6 clusters (in remote site)
- Portal access to the complex of GRAPE-6 clusters (automatic resource allocation)
- Automatic client resource distribution based on generic Grid technology