# Science Cloud
# - Migrating from Grid to Cloud -

**Yoshio Tanaka**
**AIST, Japan**

# Contents

- Introduction of two examples of migration from Grid to Cloud
  - From GEO Grid to GEO Cloud???
  - From Grid Computing to HPC Cloud

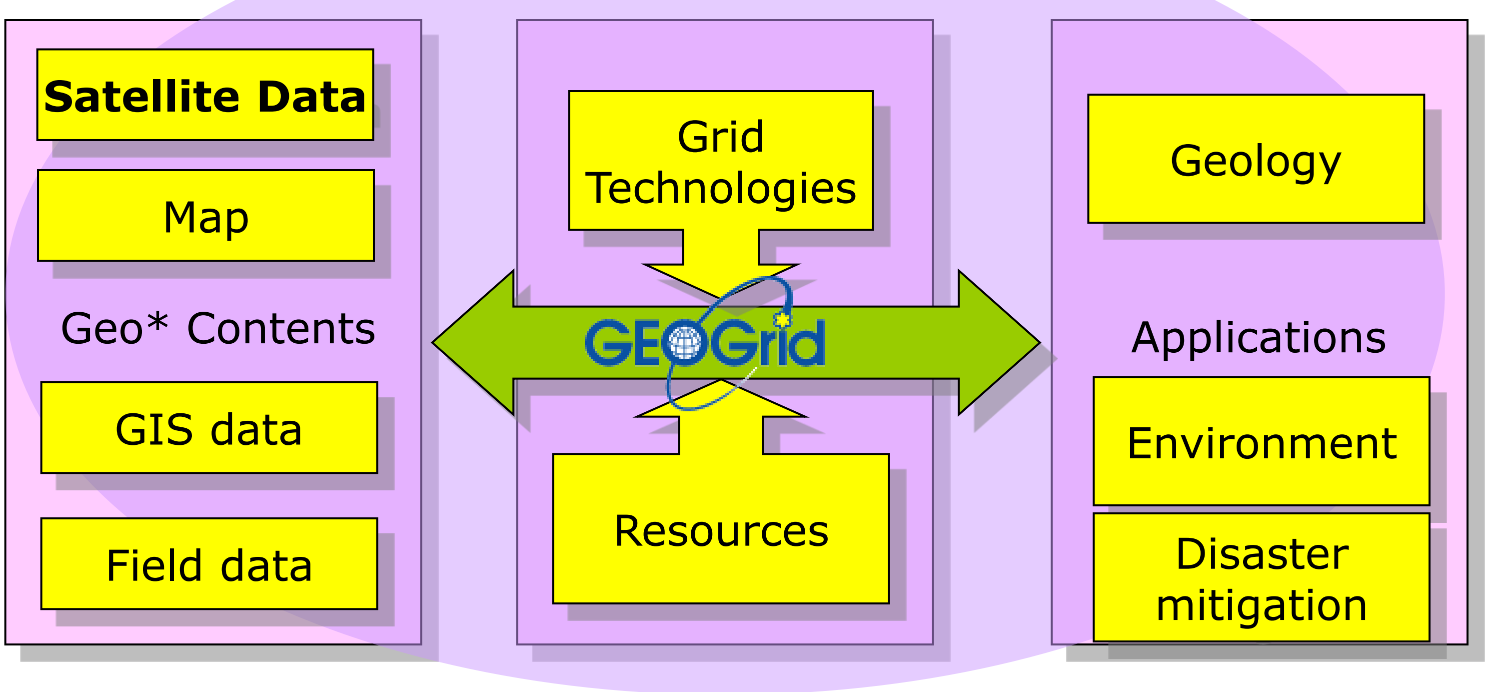
Should not distinguish GEO Cloud and HPC Cloud
Just be Science Cloud…

# What is the GEO Grid ?

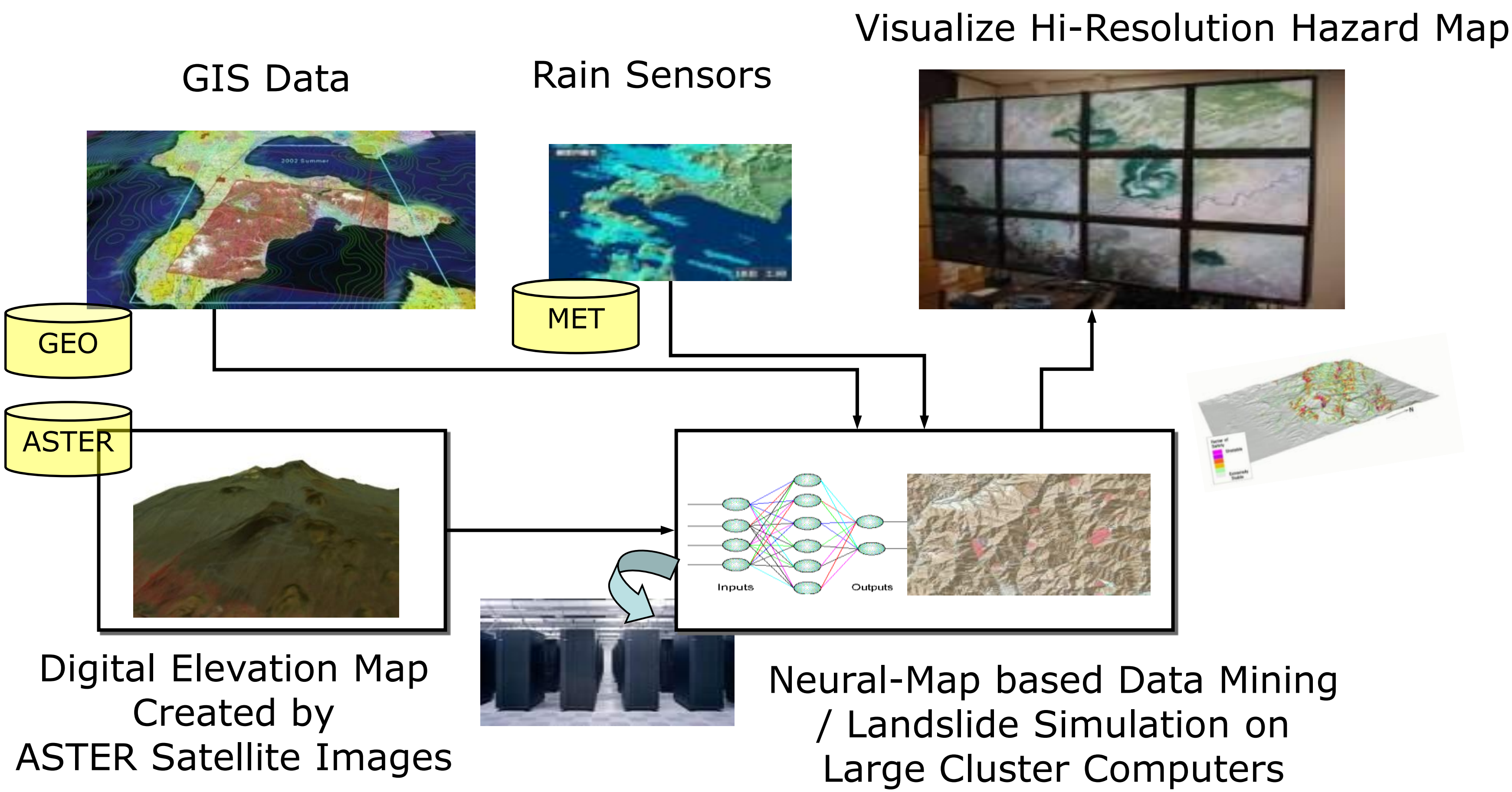
- The GEO (Global Earth Observation) Grid is aiming at providing an *E-Science Infrastructure* for worldwide Earth Sciences communities to accelerate GEO sciences based on the concept that relevant data and computation are *virtually integrated* with a certain *access control* and ease-of-use interface those are enabled by a set of Grid and Web service technologies.

**AIST: OGF Gold sponsor (a founding member)**

**AIST: OGC Associate member (since 2007)**



# Use Case: Creation of Hazard Map

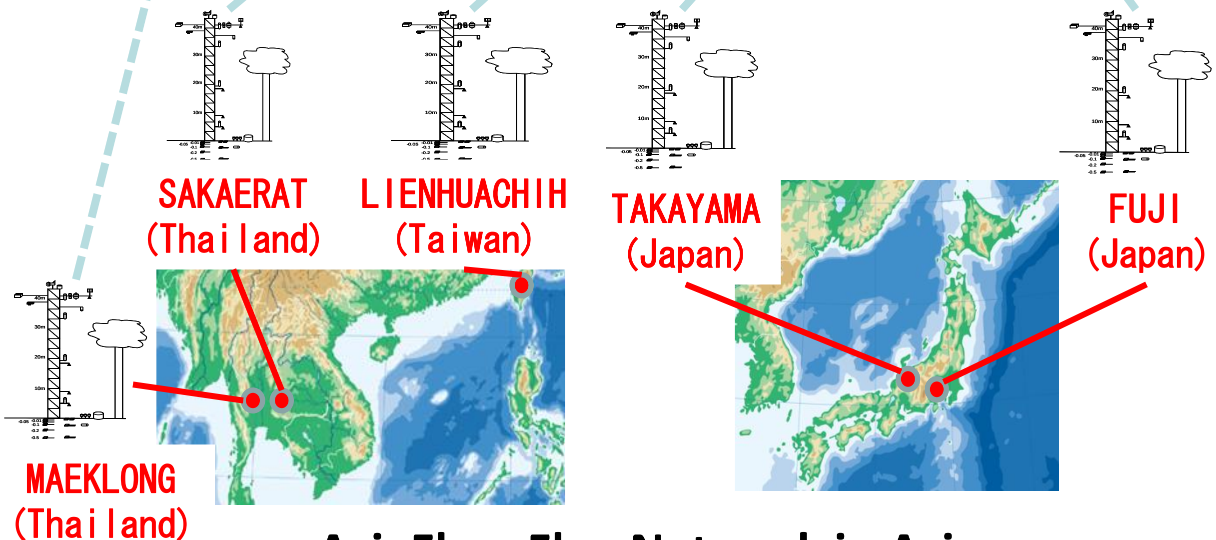


# Asia GEO Grid Initiative

**3 years Project funded by MEXT/Japan.**
**Promoting International Collaboration.**
- **AIST / Japan**
- **NIES / Japan**
- **NARL, NCHC / Taiawn**
- **NECTEC / Thailand**
- **VAST / Vietnam**

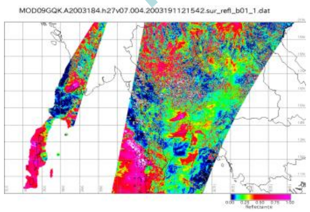
1. Search Catalogue
2. Acquire Data
3. Process Data
4. Visualize

**IT Infrastructure for sharing Earth observation data and computing resources.**




AsiaFlux: Flux Network in Asia

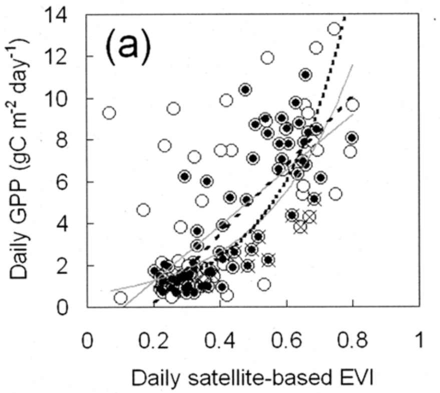
Fish-eye Digital Camera



Satellite Data

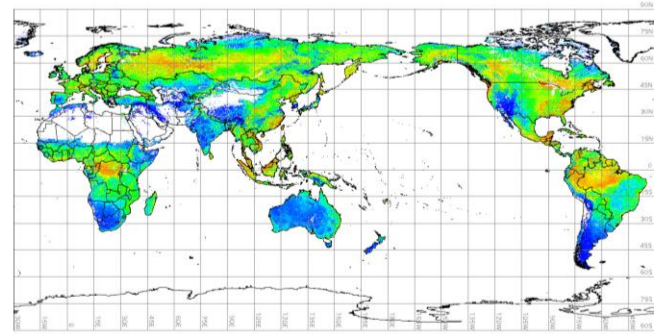
# Use Case and Research Issues
## - Federation of $CO_2$ Flux data and Satellite data -

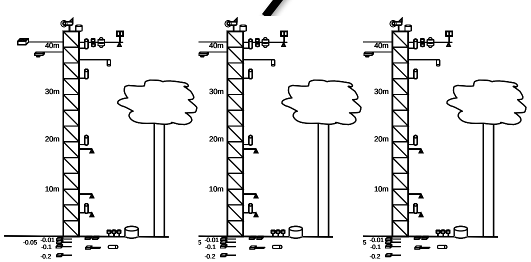
Calibration of Satellite Data using In-situ Observation Data




Apply to the similar vegetation area

**Global $CO_2$ Map generated from In-situ data and Satellite data**

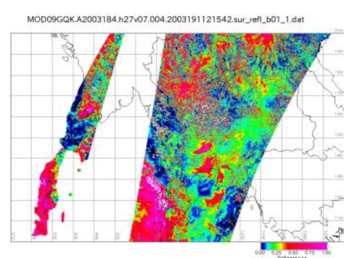




MKL SKR TKY

**Flux Tower**

**Digital Camera**




**Satellite data**

## Research Issues and approach

**(1) Development of IT infrastructure which federates distributed and heterogeneous Earth observation data.**
**Approach: Integration of Grid and OGC standards**

**(2) Establishment of multi-disciplinary and cross boundary scientific community**
**Approach: Linking IT and application networks**

# Overview

1. Catalog search
2. Data query
3. Processing
4. Results & visualization

OGC Open Geospatial Consortium, Inc.

CSW | WMS | WFS | WCS | WPS

OpenGridForum OPEN FORUM | OPEN STANDARDS

**GEO Grid**

- **Integration of heterogeneous and distributed DBs**
- **Flexible and user friendly security**
- **High performance and scalable catalogue services**
- **Development of model cases at Flux observation sites**

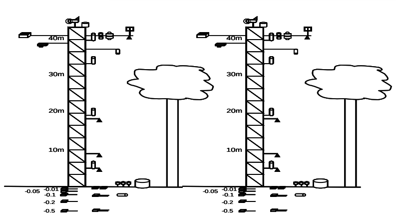
Outcomes of the other research projects

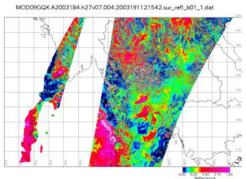
**One of research themes in this project**

**Research Theme 1**
Evaluate and improve GEO Grid middleware for practical use through empirical experiments.

**Research Theme 2**
Implementation of secure and scalable OGC standards using Grid technologies.

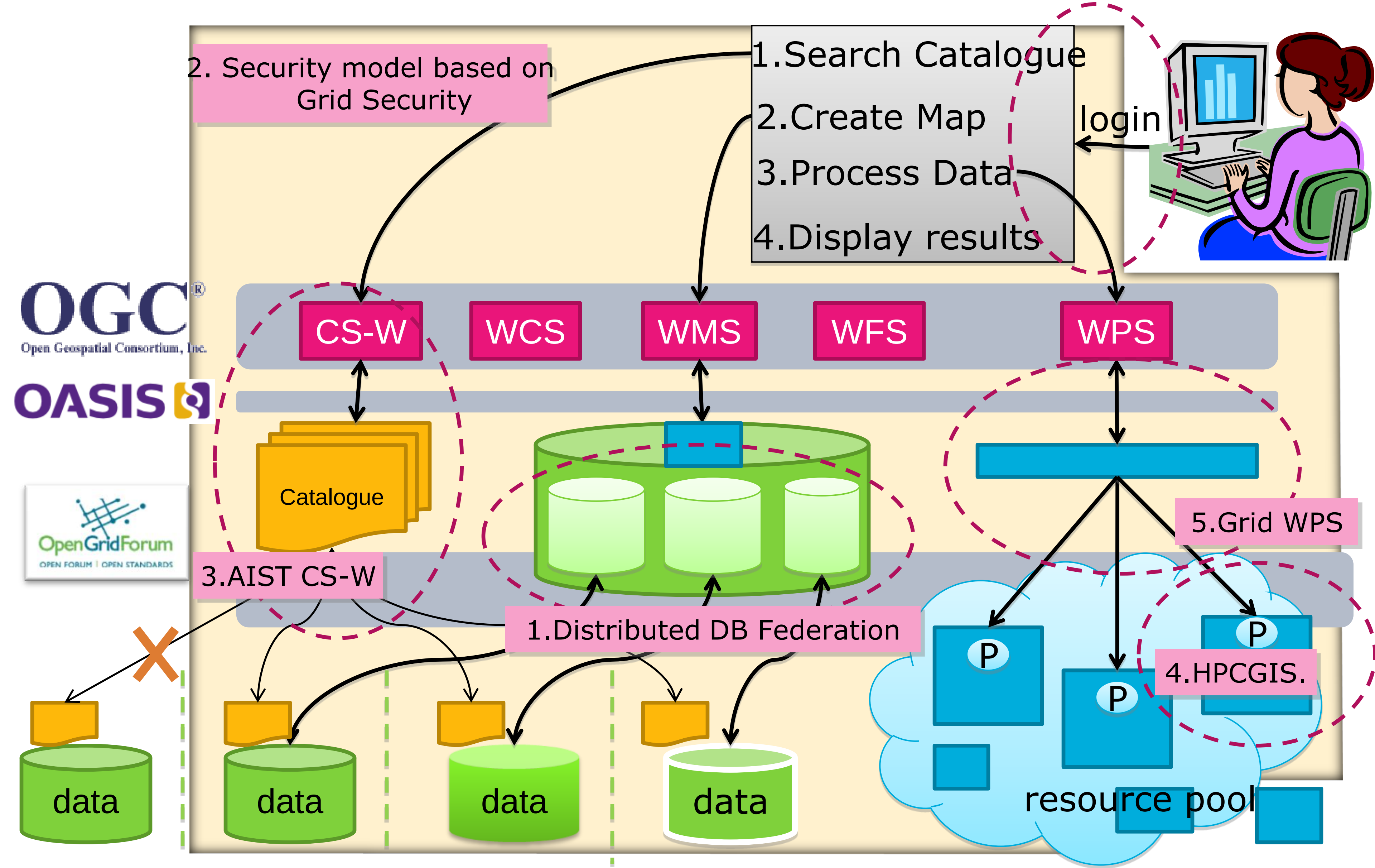
**Research Theme 3**
Development of Asia GEO Grid Infrastructure which federates heterogeneous and distributed Earth observation data.





**Expected outcome**
- ➢Development of sustainable infrastructure and community for Earth sciences
- ➢Contribution to GEOSS and International standards

# GEO Grid Enhanced Architecture

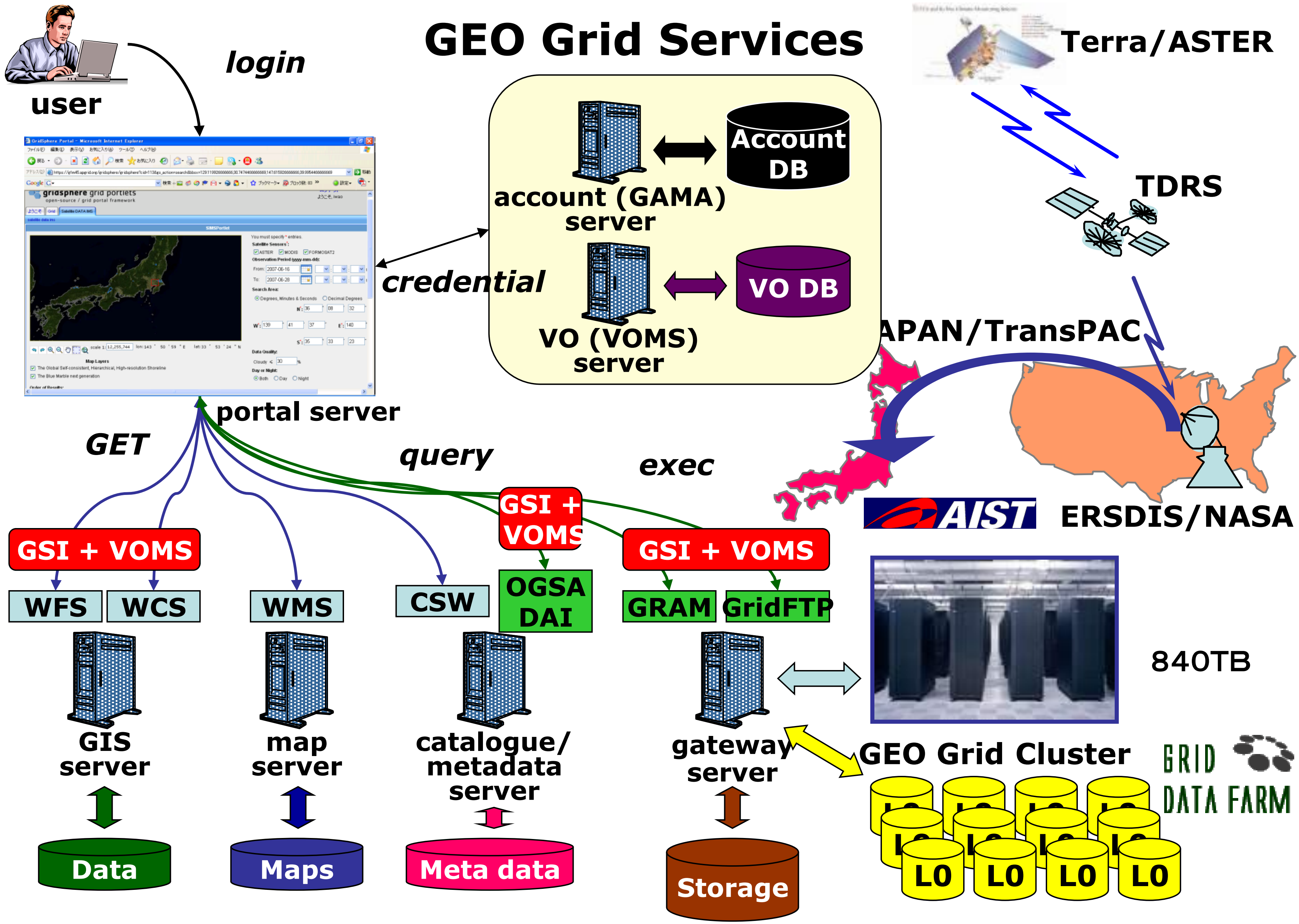


# Basic Concept of Grid

- Sharing resources between organizations.
- Key concept and technology is creation of Virtual Organizations (VOs).
  - Security is the key
- "Open & Standard" is important for sharing connecting different organizations.
- Originated from Academia (HPC, Sci. & Tech.), but…
  - Meta-computing is not practical.
    - Performance, fault tolerance
    - Small user base
    - Data Grid may still work.
  - Failed in business use.
    - Less requirements for resource sharing with different companies.

# Basic Concept of Cloud

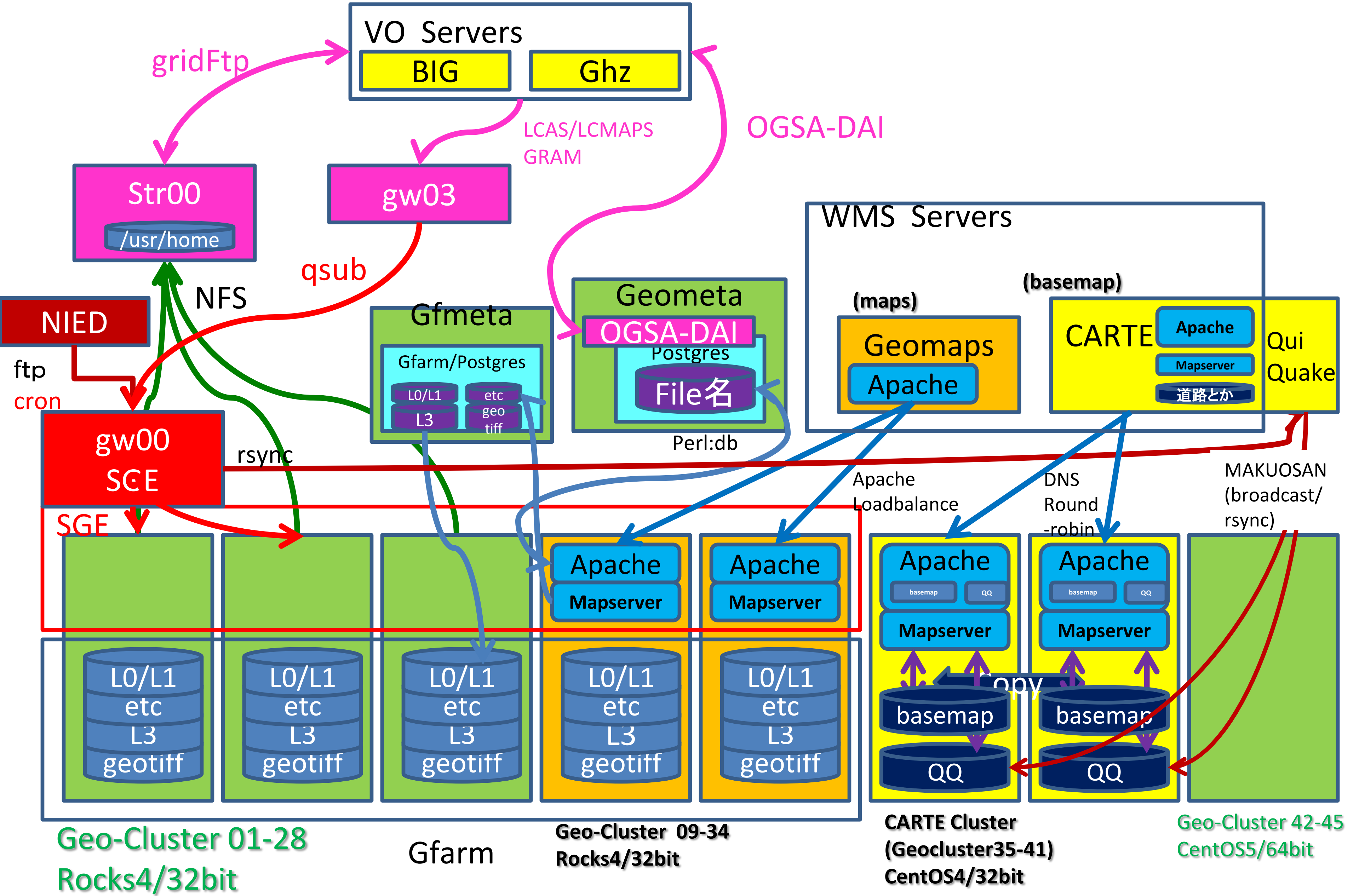
- Provides services (Infrastructure, Platform, Software) using resources in datacenter(s).
  - Increase system utilization.
  - High availability and elasticity.
  - High cost performance.
- (at this moment) Less interests for resource sharing between different companies, i.e. less interests for "Open & Standard".
  - Make it proprietary, Go forward for vendor lock-in.
  - Not sure for future.
- Originated from Business area.
  - Focus on specific capabilities
  - Make it compact, and easy to use.
  - Going to be extended to HPC (e.g. Amazon CCI, CGI).


GEO Grid Services
user
login
credential
portal server
account (GAMA) server
Account DB
VO (VOMS) server
VO DB
Terra/ASTER
TDRS
APAN/TransPAC
ERSDIS/NASA
GET
query
exec
GSI + VOMS
GSI + VOMS
GSI + VOMS
WFS
WCS
WMS
CSW
OGSA DAI
GRAM
GridFTP
GIS server
map server
catalogue/ metadata server
gateway server
Data
Maps
Meta data
Storage
840TB
GEO Grid Cluster
L0
L0
L0
L0
GRID DATA FARM

# Motivation for migrating to Cloud

- Do we need Grid protocol (e.g. GRAM, GridFTP)?
  - Application developers use not Grid middleware/protocol but OGC standards.
- Do we need Grid Security?
  - Delegation is necessary for third-party file transfer.
  - But key management is burden for end users.
  - Installation/configuration of VOMS is not easy.
- GEO Grid system is stably in operation, but not extendable.
  - Data server and computing server are tightly coupled.
  - It’s hard to use resources outside organization.
- Is GEO Grid Design appropriate for use by business partners?

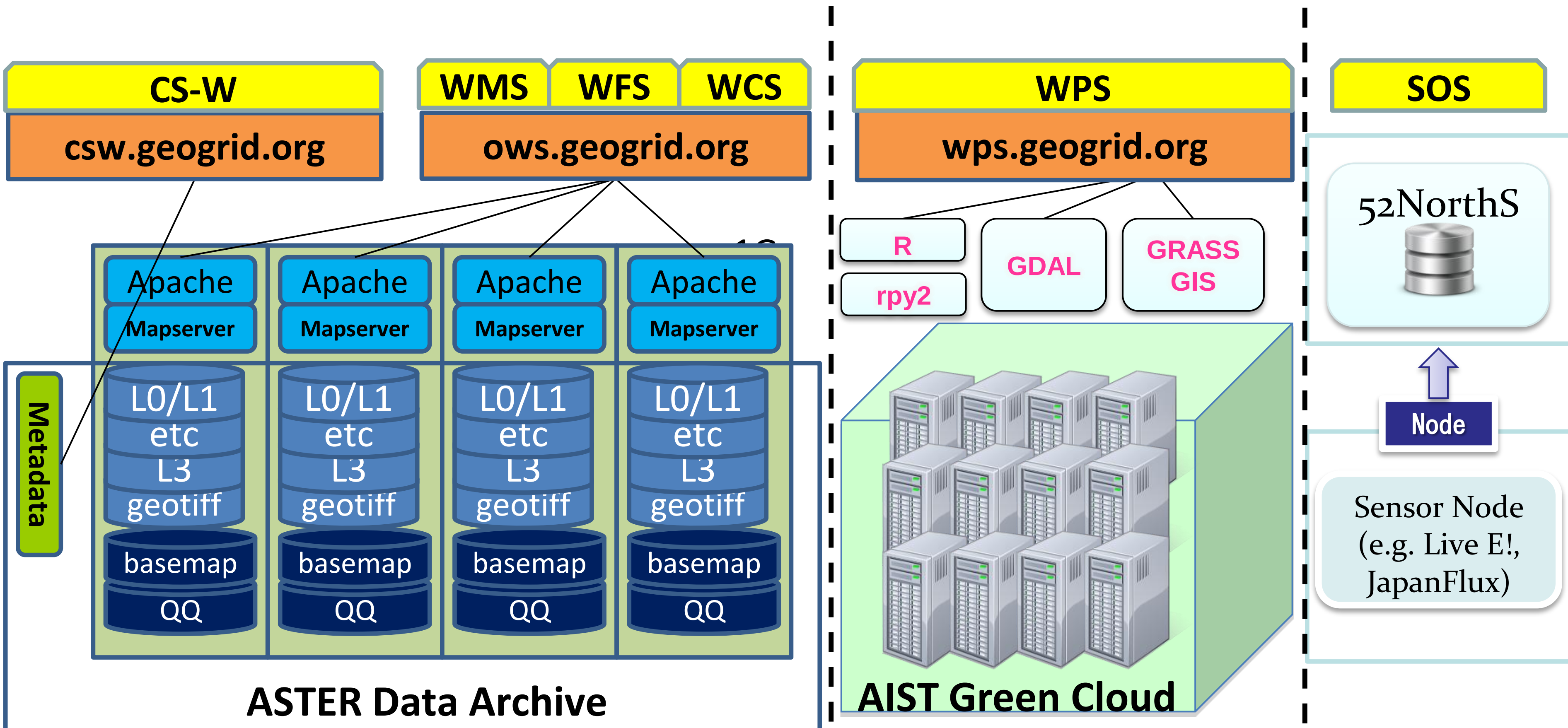
# GEO Grid System Configuration as of May 2010



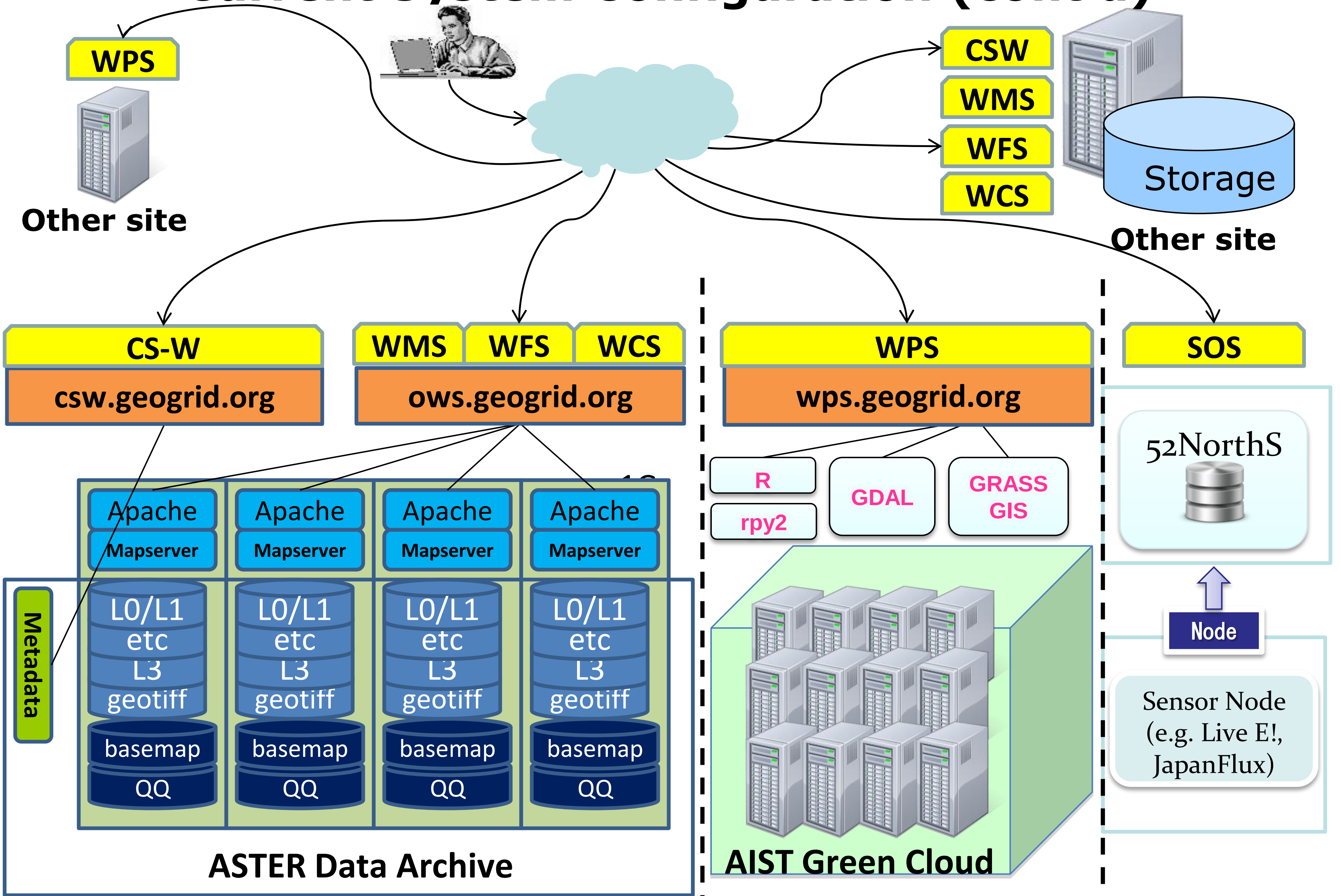
# What we did

- Clearly separate data service and computation services.
- Computation services are provided on virtual clusters.
- Create VM images for computation services so that the services will be available from the other Cloud vendors including private sectors.

# Current System Configuration



# Current System Configuration (cont'd)



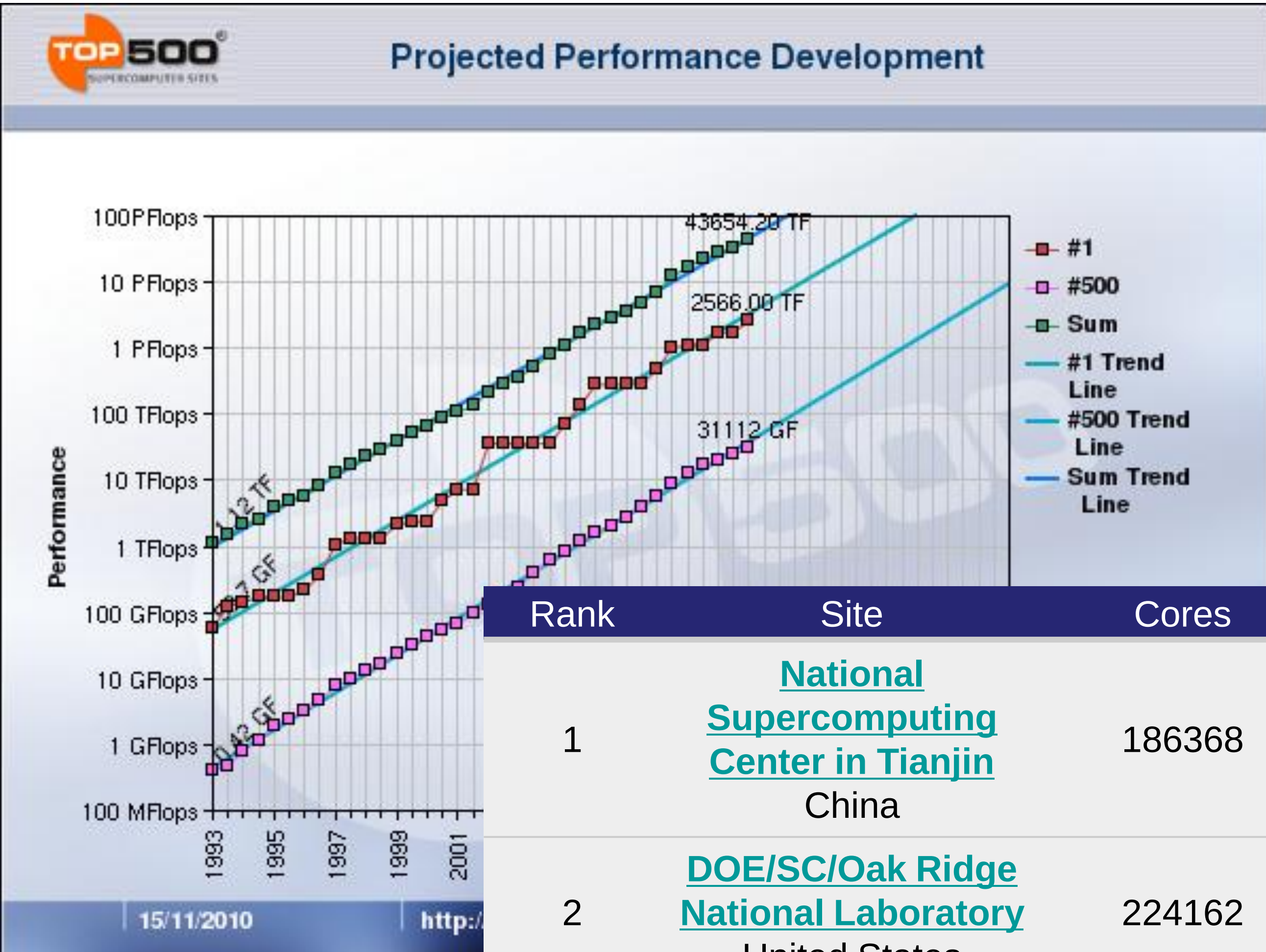
# Summary of GEO Grid -> GEO Cloud

- Re-configure GEO Grid system to make it simple and extendable.
- Good to collaborate with the other organizations, especially with private sectors.
- Need to consider appropriate security.

# Contents

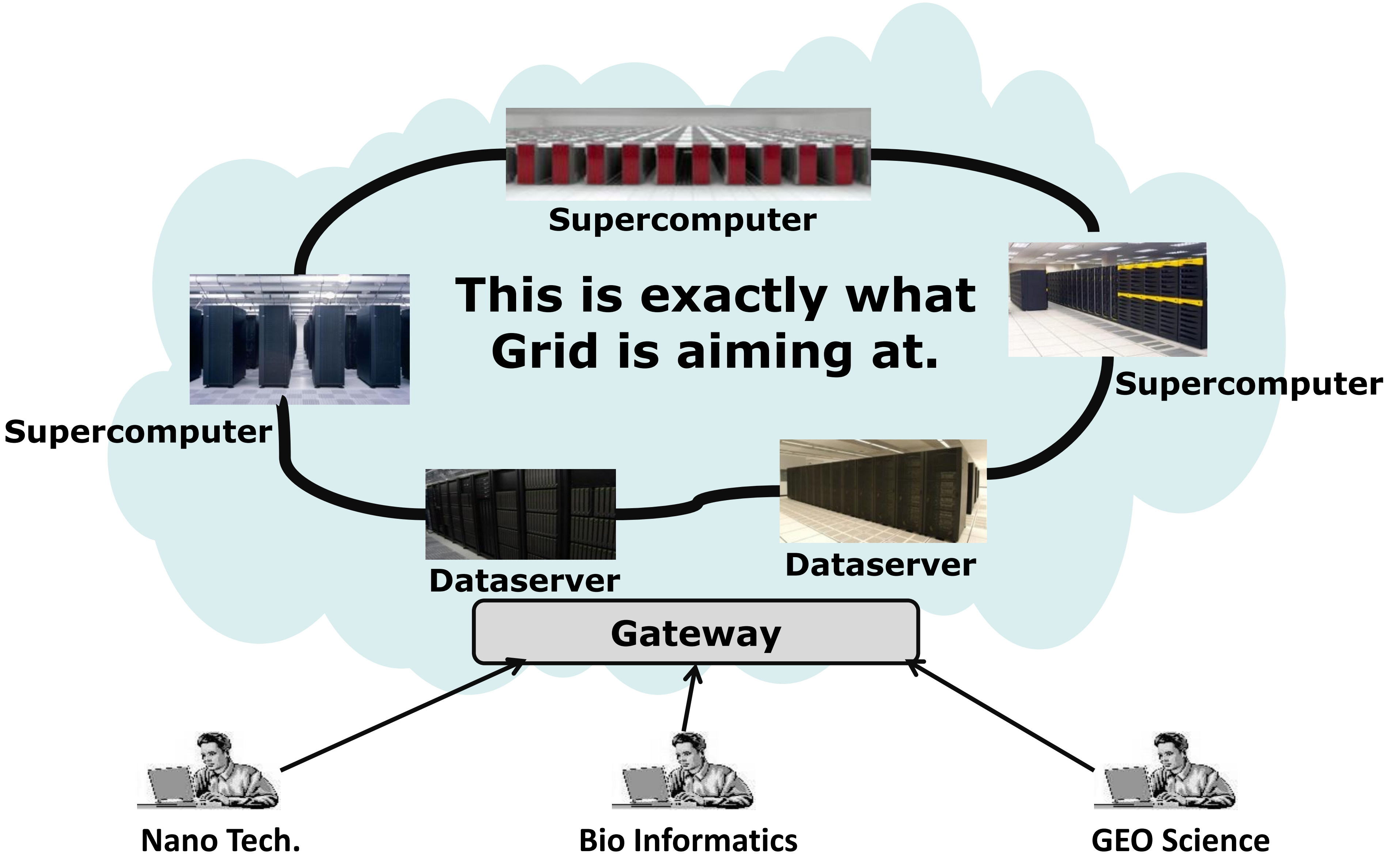
- Introduction of two examples of migration from Grid to Cloud
  - From GEO Grid to GEO Cloud???
  - From Grid Computing to HPC Cloud

# Millions of Cores will come soon.



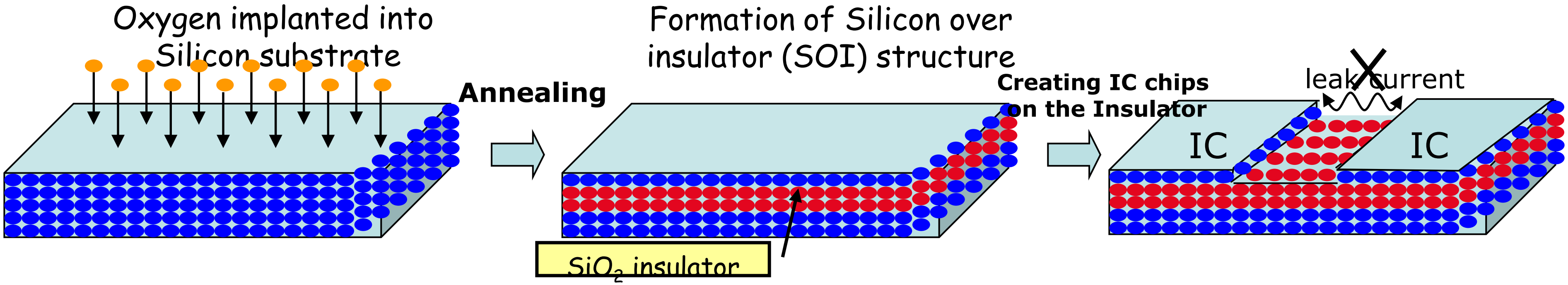

| Rank | Site | Cores | $R_{max}$ | $R_{peak}$ |
|---|---|---|---|---|
| 1 | National Supercomputing Center in Tianjin China | 186368 | 2566.00 | 4701.00 |
| 2 | DOE/SC/Oak Ridge National Laboratory United States | 224162 | 1759.00 | 2331.00 |
| 3 | National Supercomputing Centre in Shenzhen (NSCS) China | 120640 | 1271.00 | 2984.30 |

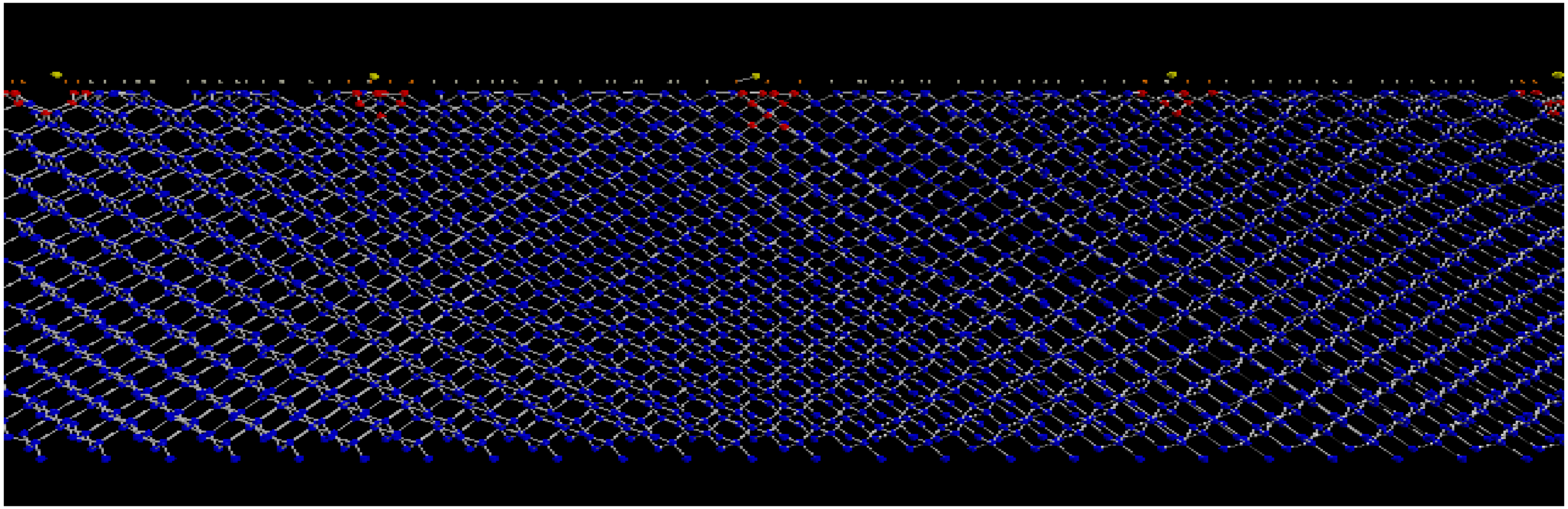
# Image of HPC Cloud



# Grid-enabled SIMOX Simulation on Japan-US Grid Testbed at SC2005

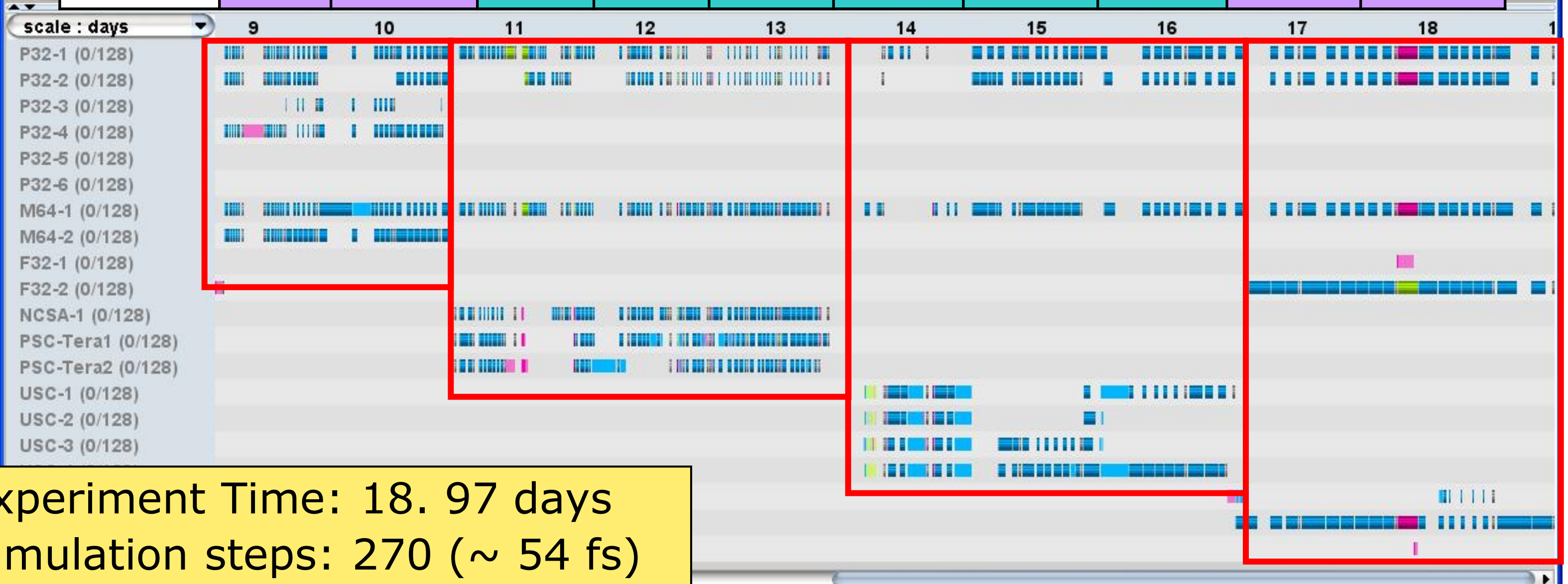
- A technique to fabricate a micro structure consisting of Si surface on the thin $SiO_2$ insulator
- Allows to create higher speed with lower power consumption device



# We could learn from Grid experiments


| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| QM1 | P32 | P32 | P32 | P32 | P32 | USC | USC | USC | ISTBS | ISTBS |
| QM2 | P32 | P32 | NCSA | NCSA | NCSA | USC | USC | USC | Presto | Presto |
| QM3 | M64 | M64 | M64 | M64 | M64 | M64 | M64 | M64 | M64 | M64 |
| QM4 | P32 | P32 | TCS | TCS | TCS | USC | USC | USC | P32 | P32 |
| QM5 | P32 | P32 | TCS | TCS | TCS | USC | USC | USC | P32 | P32 |
| Reserve | F32 | F32 | P32 | P32 | P32 | P32 | P32 | P32 | F32 | F32 |




- Experiment Time: 18. 97 days
- Simulation steps: 270 (~ 54 fs)

Phase 1 Phase 2 Phase 3 Phase 4

# Heterogeneity was PAIN!!

- Heterogeneity did exist in various places
  - OS, Library version
  - in more details of the system configuration
    - Configuration of queuing systems
      - e.g. max wall clock time
    - disk quota limit
    - Firewall configuration
- We need to install and test applications on each supercomputer one by one.
  - We should not expect end users to do the same!

# Resource Management was PAIN!!

- We made a cross-site reservation for TeraGrid resources, but…
  - In some cases, our jobs were not activated when reserved time had been reached due to miss-configuration of the system.
  - We need to contact to help@teragrid.org to fix the problem.
  - In some cases, we need to ask special (manual) operation for experiments.
    - give us a special (dedicated) queue
    - need help for unexpected errors (jobs were not activated)
    - more easy operation for cross-site reservation is expected
  - We have no good solution to solve time-difference problem!

# What has happened in PRAGMA

- Migrating from Globus-based Grid to VM-based infrastructure.
- PRAGMA Grid is not a collection of homogeneous resources.
  - A user has to install and test applications on each resource one-by-one.
  - PRAGMA Grid is a collection of a large number of small clusters, this does not motivate end users to use PRAGMA Grid…
- In PRAGMA 17th @ Daejeon, we agreed to migrate from traditional Globus-based Grid to VM-based infrastructure.
- We have tested
  - sharing VM images between two virtual clusters,
  - Extending to Amazon EC2.

# Grid Computing to HPC Cloud

- Supposed Scenario
  - Application developers creates a VM image which includes application as well as running environment.
  - Submit the VM image to HPC Cloud.
  - The running environment and application will be deployed on a selected HPC resource.
  - Start the application run.

Build Once, Run Everywhere

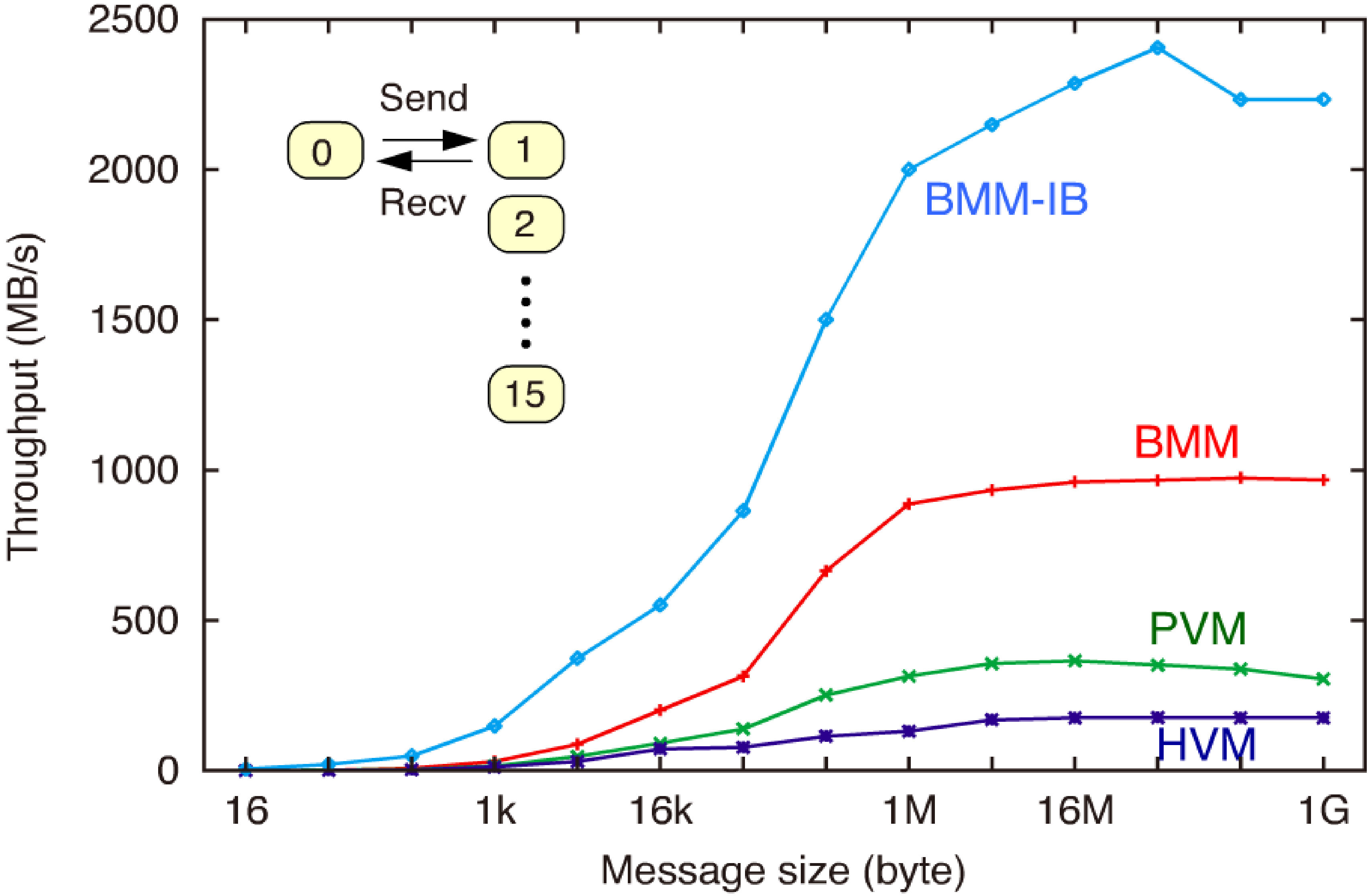
# Research Issues

- How do users create VM images?
- How do we select appropriate HPC resources and monitor our application run?
  - Meta-scheduler / broker
  - Monitoring system
- How do we hide heterogeneity of networks?
  - IB, Myrinet, 10G, …
- What is the appropriate security?
- How do we share data/file between multiple HPC resources?
- <u>Is performance acceptable</u>?

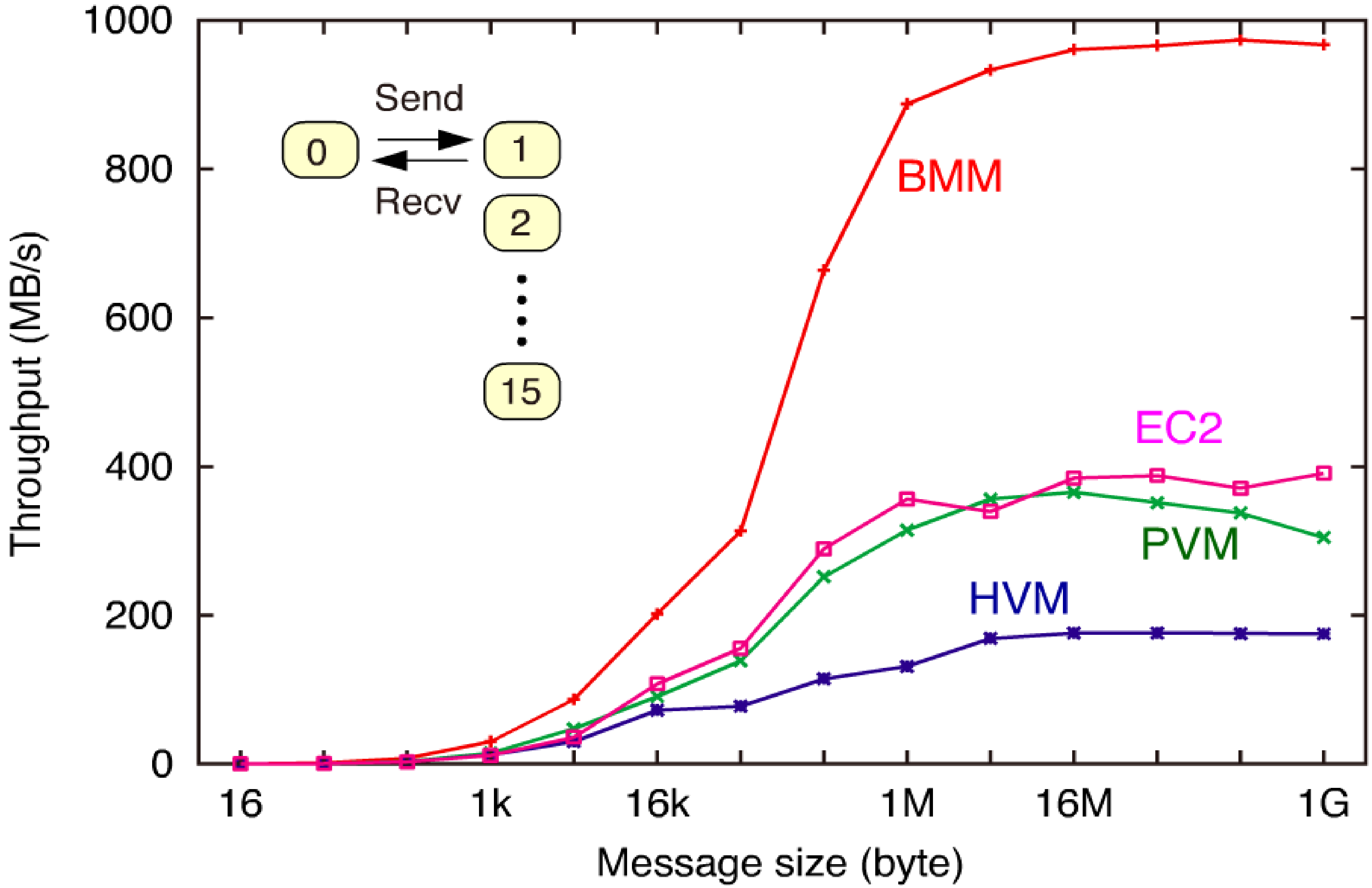
# VM Performance

| | AIST Green Cloud (AGC) | EC2 Cluster Instance (EC2) |
|---|---|---|
| CPU | E5540 (2.53GHz)<br>2 Sockets 8 Cores | X5570 (2.93GHz)<br>2 Sockets 8 Cores |
| Mem. | 48 GB | 23 GB |
| HD | SAS 300GB x 2<br>(HW RAID1) | root 20GB (EBS)<br>Ephemeral 850GB x 2 |
| Comm. | InfiniBand + 10GEther | 10GEther |
| OS | CentOS 5.5<br>BMM, PVM, HVM | CentOS 5.4<br>HVM |

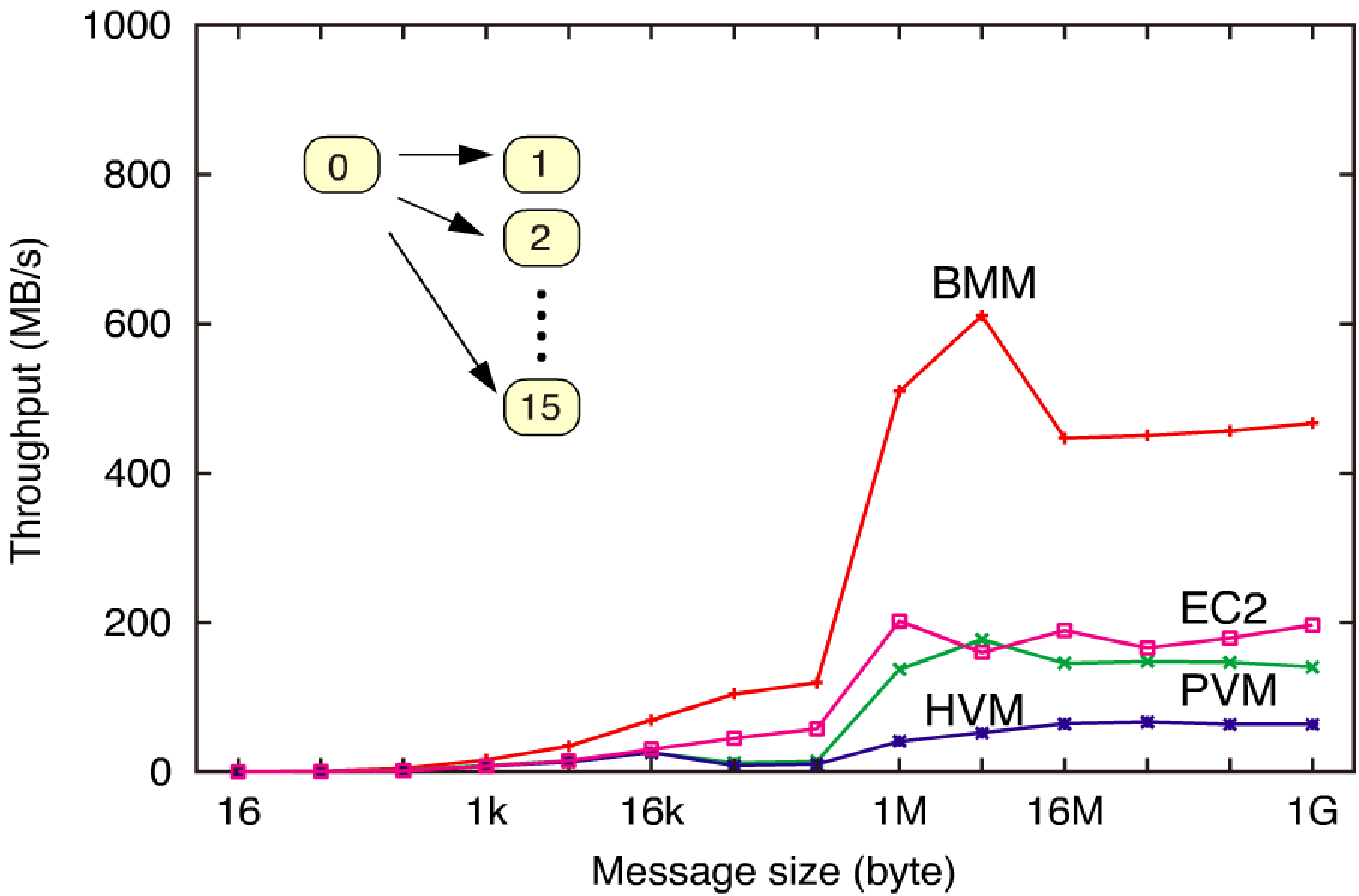
# MPI: PingPong



# MPI: PingPong

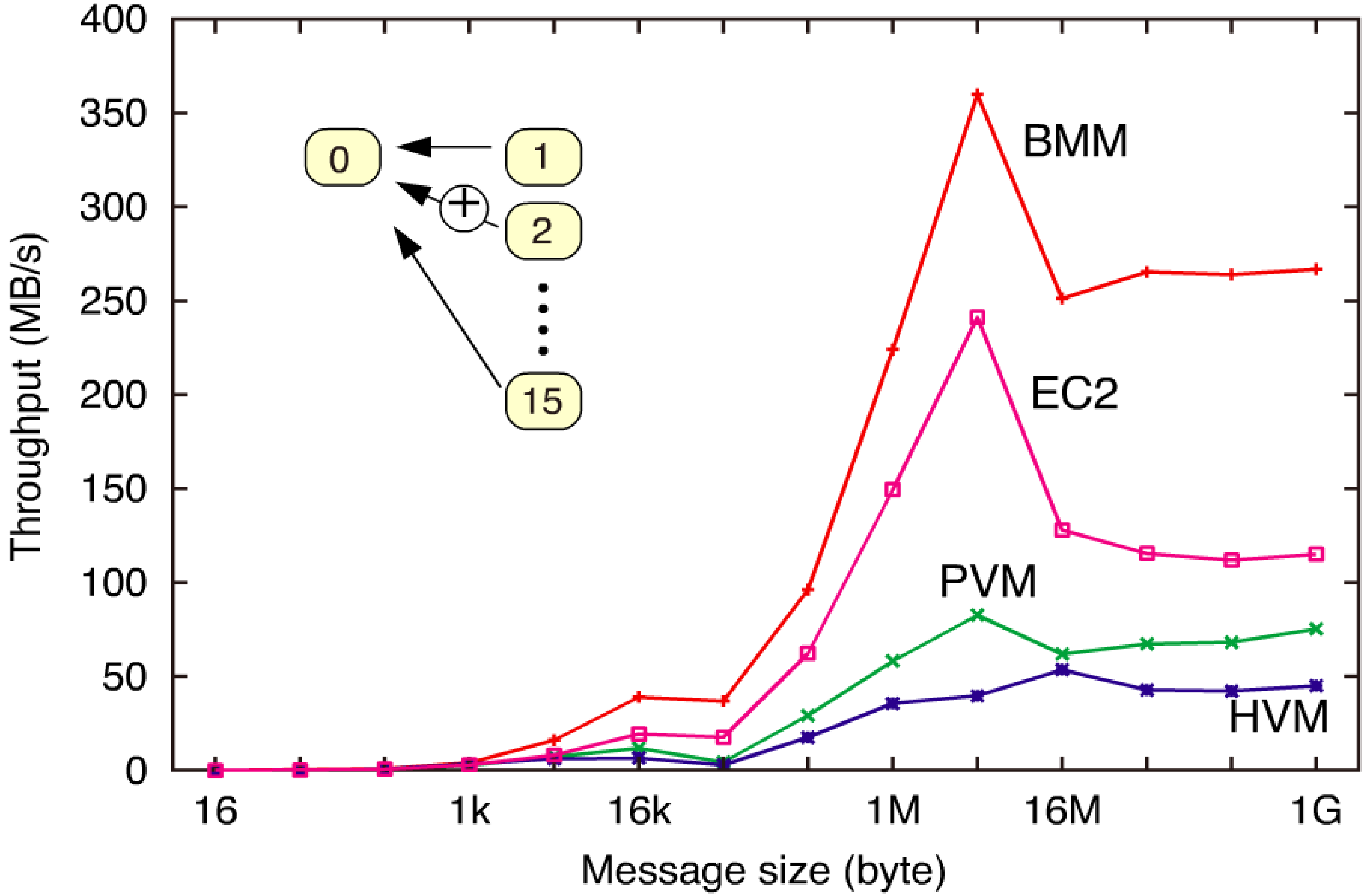


# MPI: Bcast



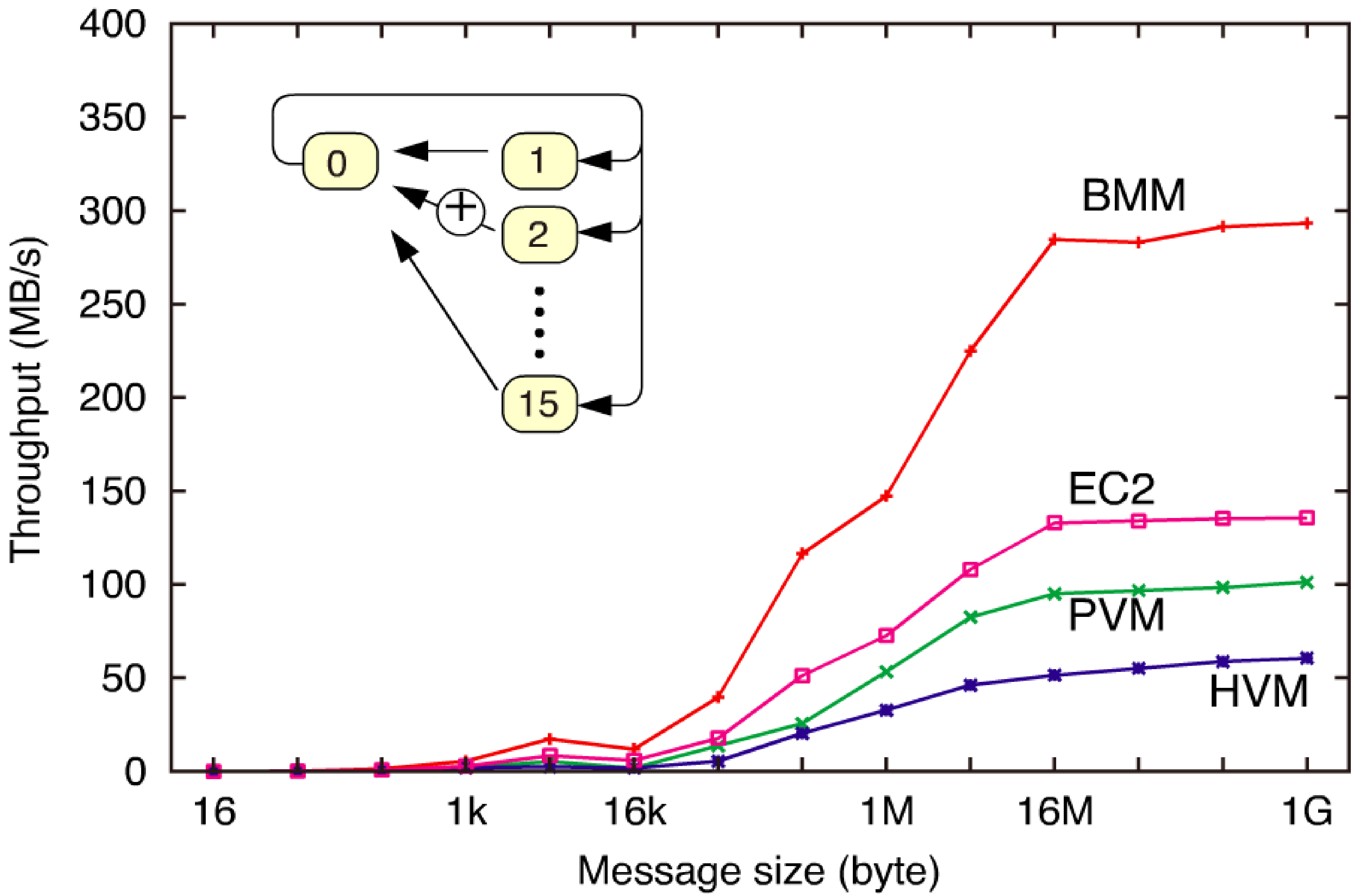
1000
800
600
400
200
0
Throughput (MB/s)
16
1k
16k
1M
16M
1G
Message size (byte)
0
1
2
15
BMM
EC2
PVM
HVM

# MPI: Reduce



0
1
2
15
BMM
EC2
PVM
HVM
Throughput (MB/s)
Message size (byte)
400
350
300
250
200
150
100
50
0
16
1k
16k
1M
16M
1G

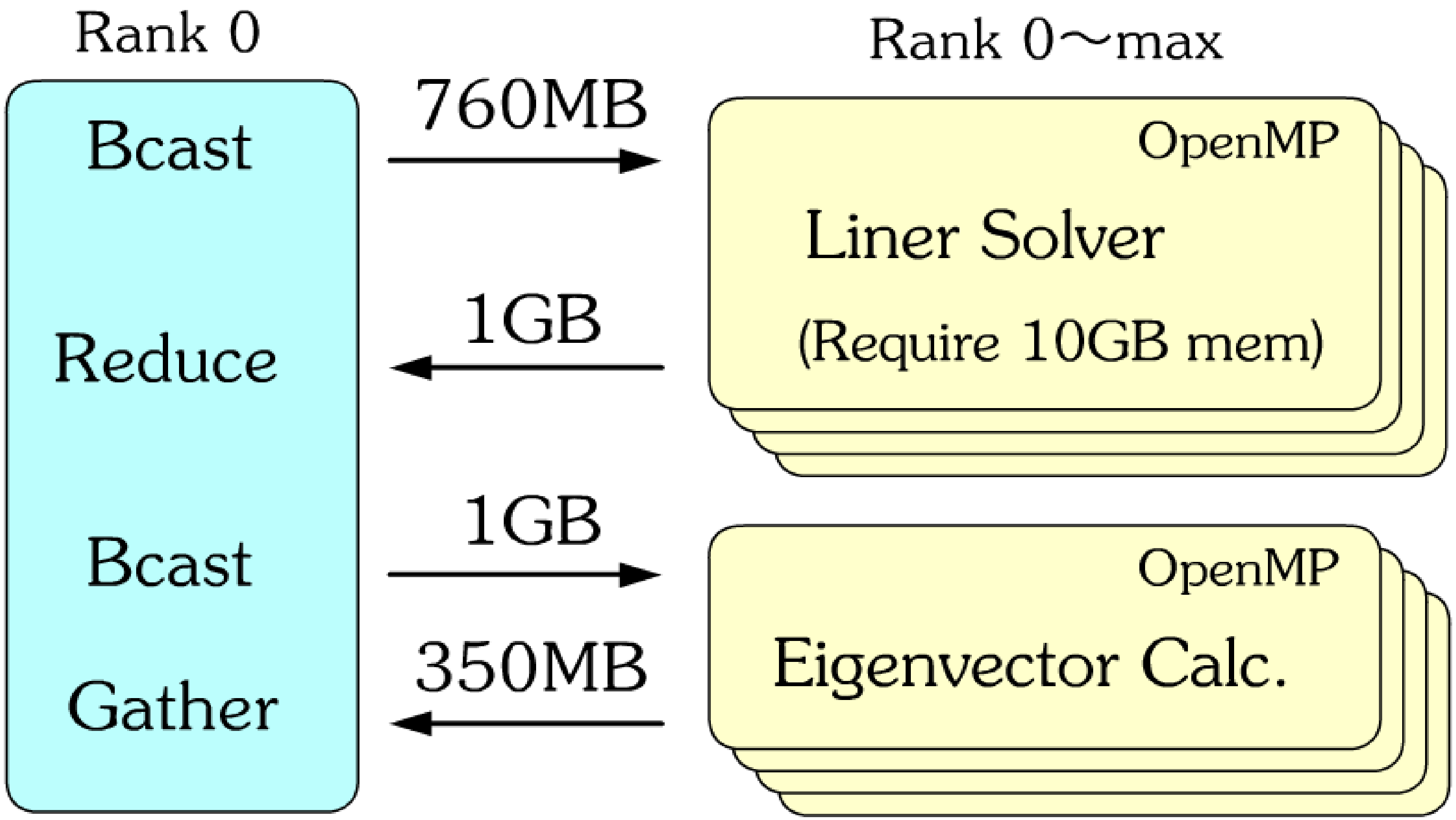
# MPI: Allreduce



# Real HPC Application

Bloss: Non-linear internal eigensolver (~1M dim)

Hierarchical parallel program by MPI + OpenMP



# Single node performance

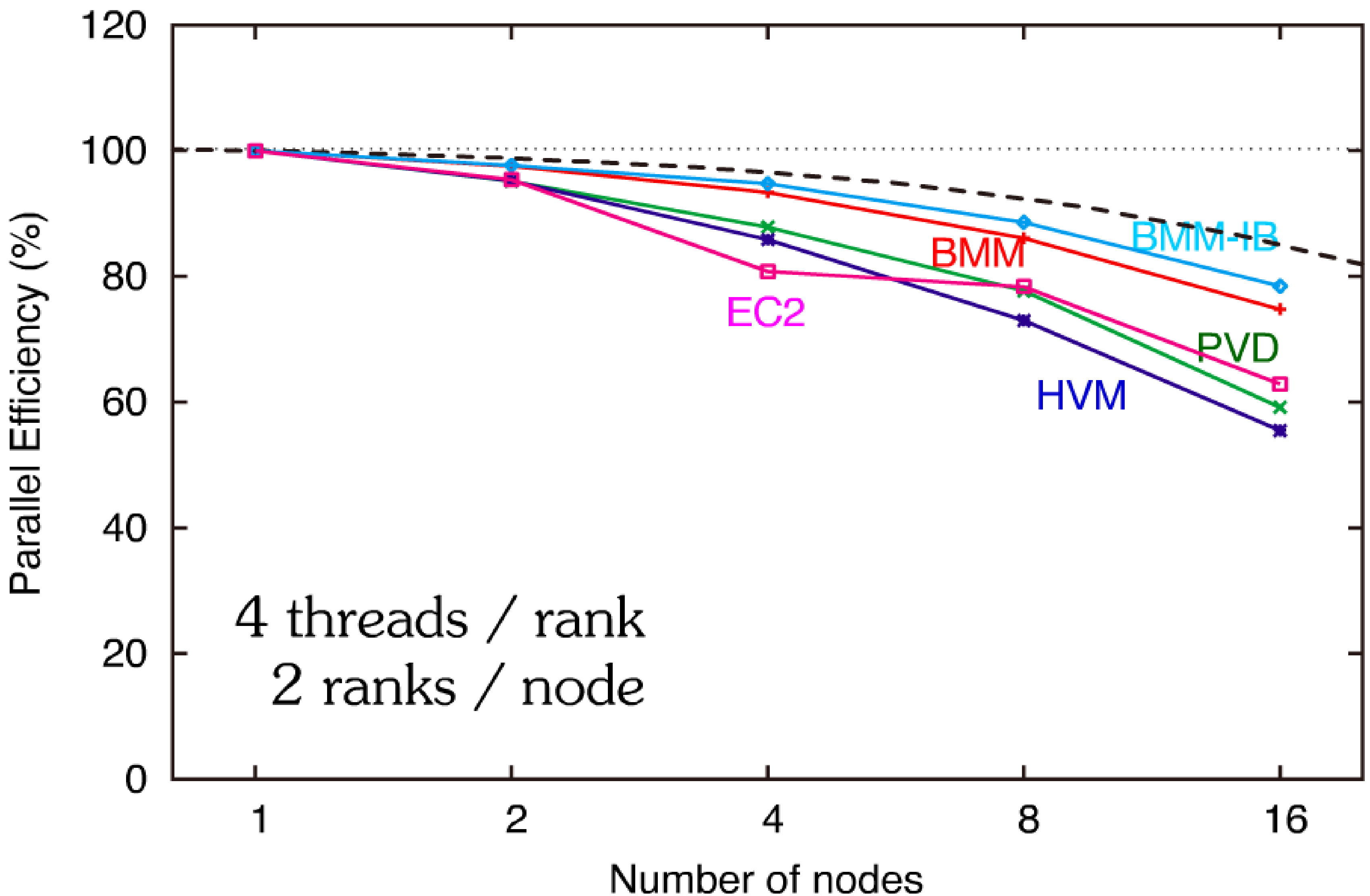
Launch 2 ranks on a single node
4 OpenMP threads per rank

| Machine type | Comp. time (min) |
|---|---|
| BMM | 21.06 |
| PVM | 22.33 |
| HVM | 22.66 |
| EC2 | 20.00 |

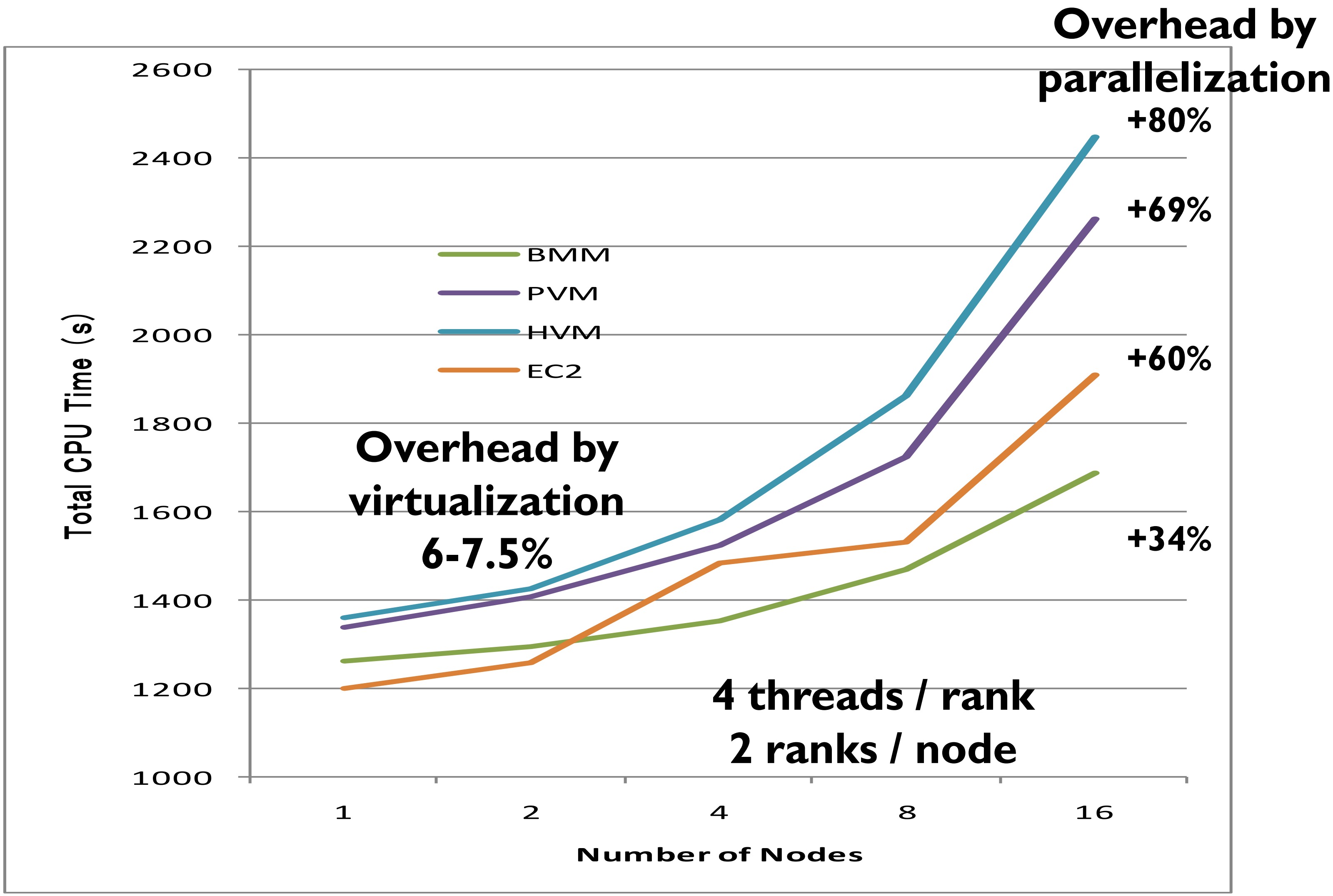
In Bloss, virtualization overhead is ~5 %.

AIST

# Parallel performance



# Parallel performance



# Summary of HPC Cloud

- Provider side
  - No need to investigate HPC Applications.
  - Just concentrate on VM tuning.
- User side
  - Utilize large amount of cloud resources.
  - Build once, run everywhere.

# What should Science Cloud be aiming at?

- Providing large scale and flexible (elastic) high performance computing services.
  - Pay-as-you-go model.
  - Frees end users from burden of system administration and management.
  - Use up-to-date technologies to achieve performance as high as possible.
  - High performance computing for everybody.
- Federating of large-scale data sets
  - Scale up to peta-, exa-, to yota-scale data.
  - Federation of widely distributed data
- Enabling service harmonization of data and computing services.
  - Build applications and higher level services by combining distributed services for solving scientific problems.

# Summary

- Cloud has been rapidly growing.
  - Focus on specific capabilities.
  - Easy to use.
- Science Cloud may encounter with difficult problems (e.g. federation, security) which Grid has been tackling with.
- We should not follow the same way with Grid but our insights gained from Grid experiences must be valuable for solving these problems by light-weight solutions.