GRID DATA FARM

Gfarm Grid File System

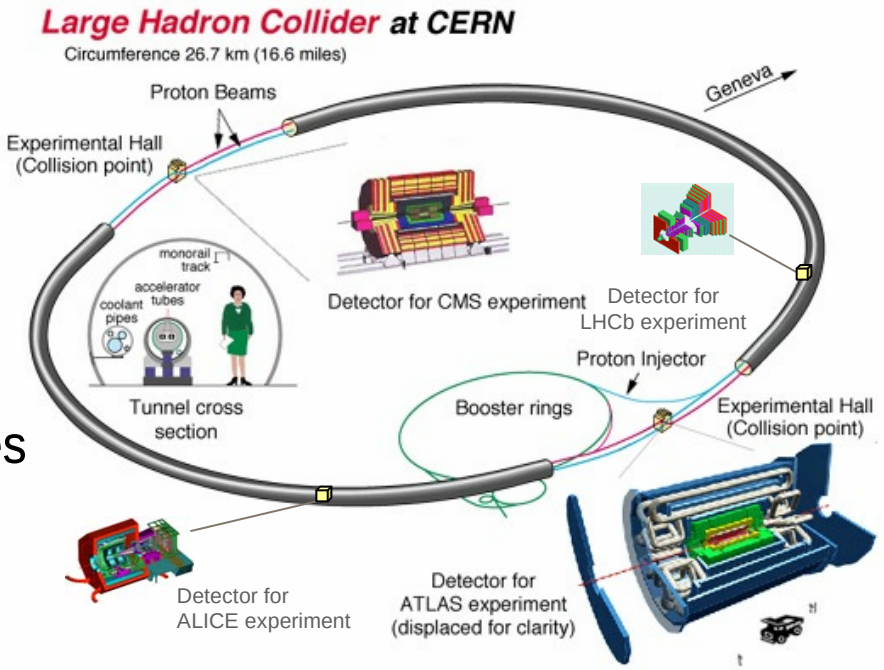
Osamu Tatebe
University of Tsukuba
tatebe@cs.tsukuba.ac.jp

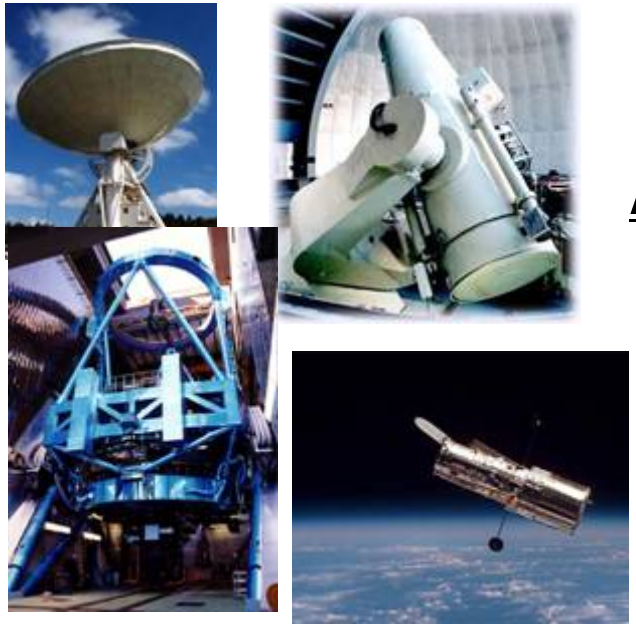
Petascale Data Intensive Computing

GRID DATA FARM

High Energy Physics

- CERN LHC, KEK-B Belle
  - ~MB/collision,
    100 collisions/sec
  - ~PB/year
  - 2000 physicists, 35 countries





Astronomical Data Analysis

- data analysis of the whole data
- TB~PB/year/telescope
- Subaru telescope
  - 10 GB/night, 3 TB/year

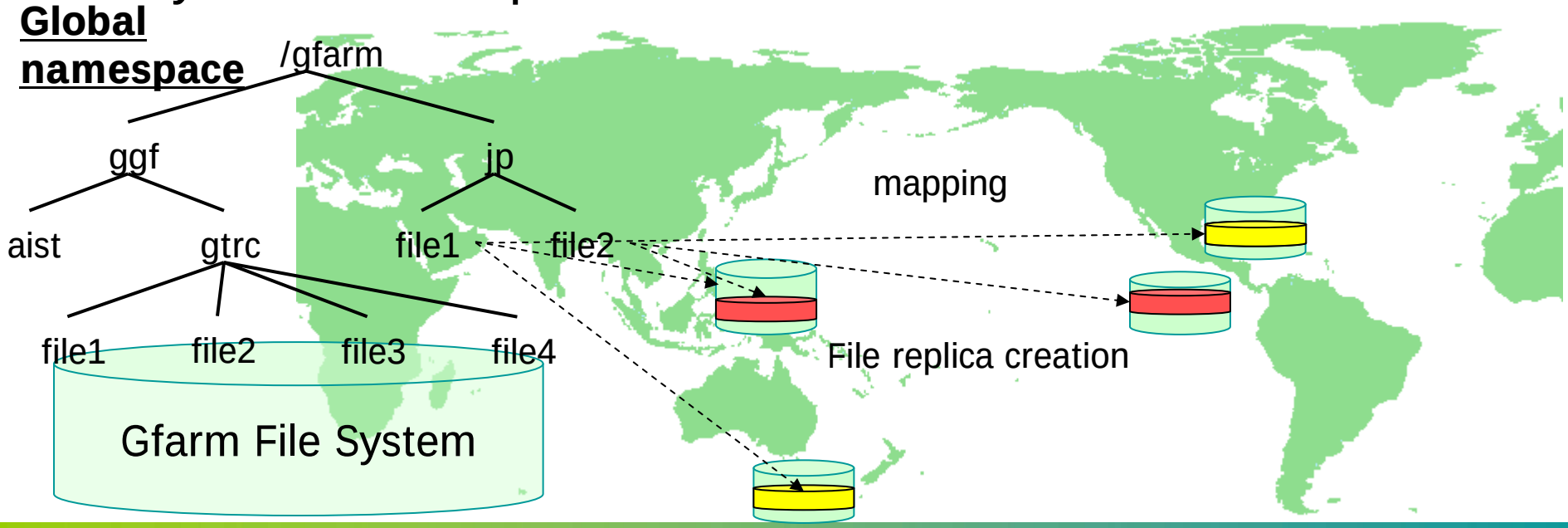
Petascale Data Intensive Computing Requirements

GRID DATA FARM

- **Storage Capacity**
  - Peta/Exabyte scale files, millions of millions of files
- **Computing Power**
  - > 1TFLOPS, hopefully > 10TFLOPS
- **I/O Bandwidth**
  - > 100GB/s, hopefully > 1TB/s within a system and between systems
- **Global Sharing**
  - group-oriented authentication and access control

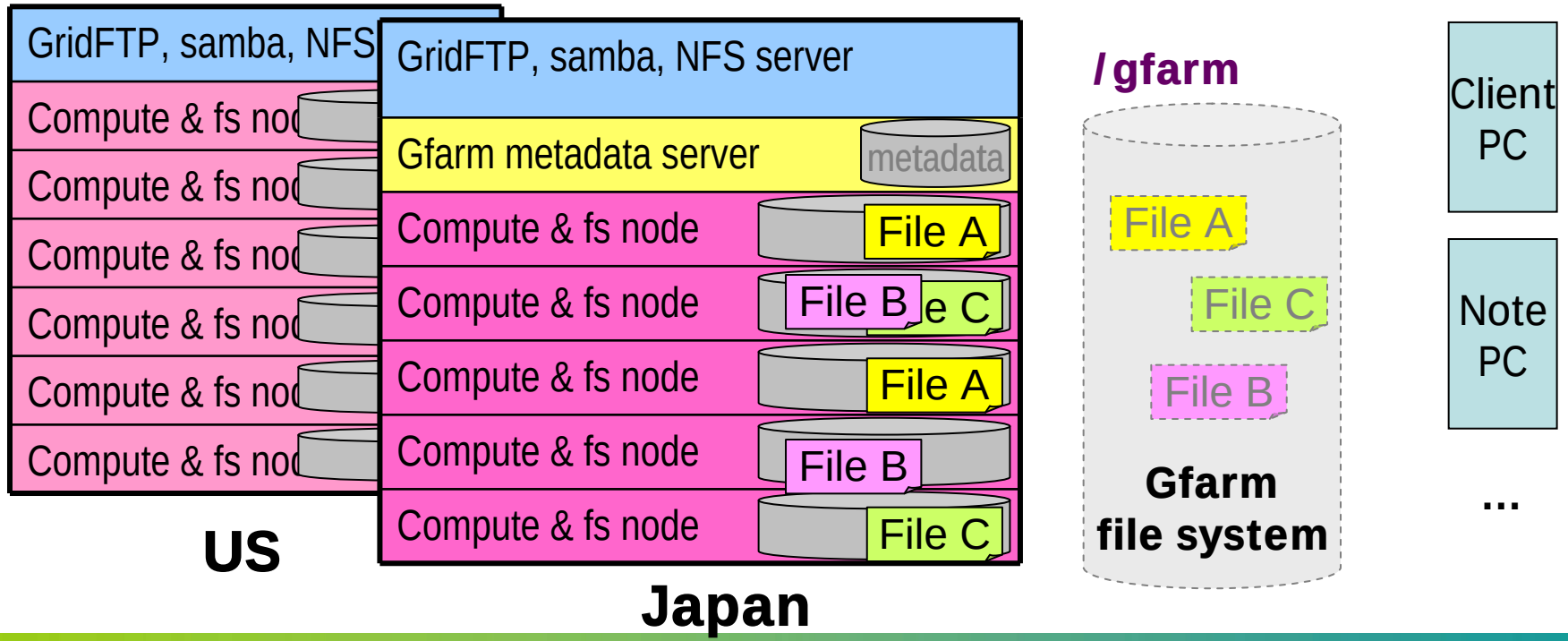
Gfarm Grid File System [CCGrid 2002]

- Commodity-based distributed file system that federates storage of each site
- It can be mounted from all cluster nodes and clients
- It provides scalable I/O performance wrt the number of parallel processes and users
- It supports fault tolerance and avoids access concentration by automatic replica selection



Gfarm Grid File System (2)

- Files can be shared among all nodes and clients
- Physically, it may be replicated and stored on any file system node
- Applications can access it regardless of its location
- File system nodes can be distributed



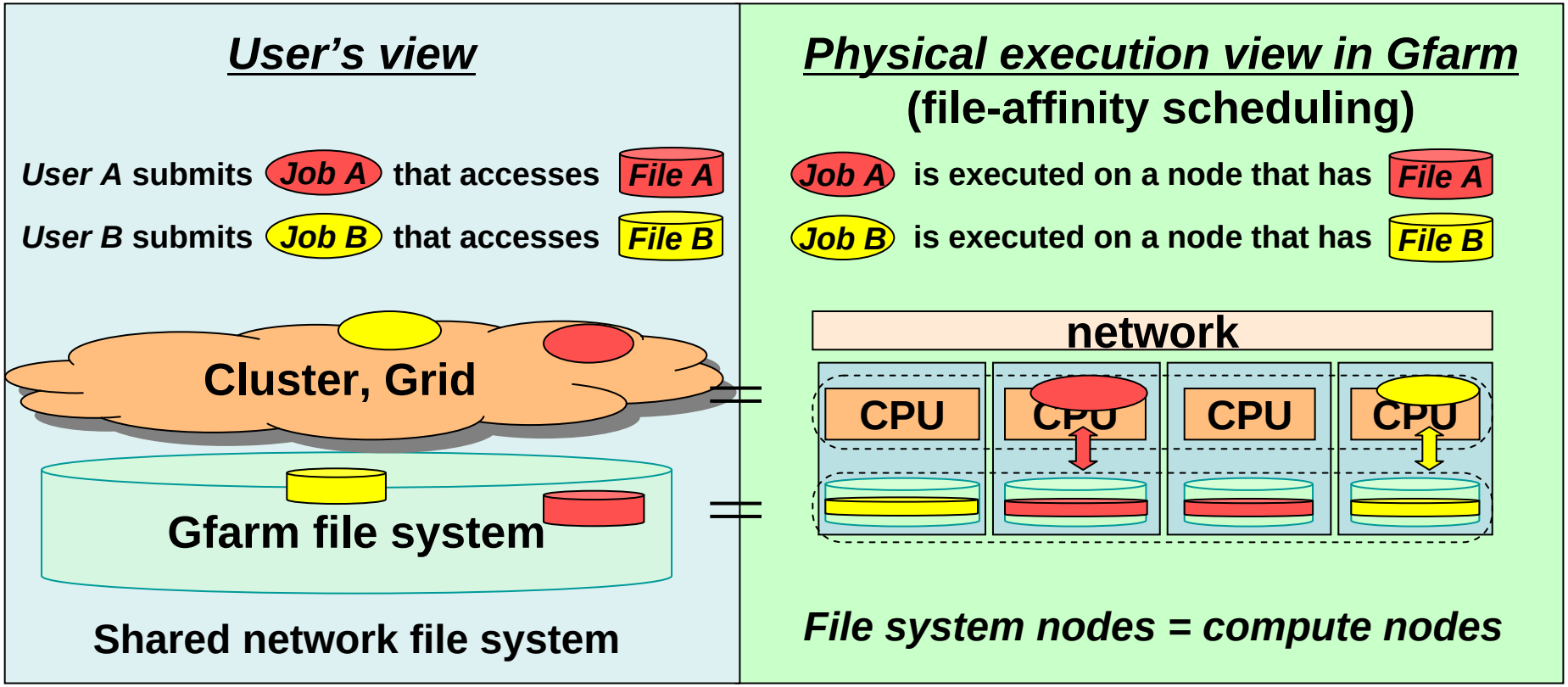
Scalable I/O Performance

GRID DATA FARM

- **Decentralization of disk access putting priority to local disk**
  - When a new file is created,
    - Local disk is selected when there is enough space
    - Otherwise, near and the least busy node is selected
  - When a file is accessed,
    - Local disk is selected if it has one of the file replicas
    - Otherwise, near and the least busy node having one of file replicas is selected
- **File affinity scheduling**
  - Schedule a process on a node having the specified file
    - Improve the opportunity to access local disk

Scalable I/O performance in distributed environment

GRID DATA FARM



Do not separate storage and CPU (SAN not necessary)

Move and execute program instead of moving large-scale data

exploiting local I/O is a key for scalable I/O performance

FIRST – CCS Astrophysics Simulator

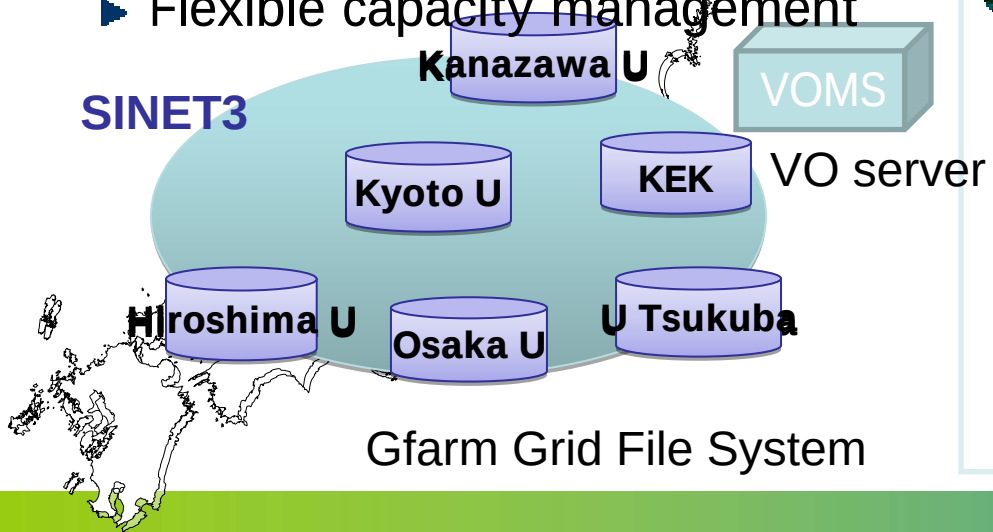
GRID DATA FARM

- **256 Compute Node**
  - Dual Xeon for 256 nodes
  - Blade-GRAPE (240 nodes)
  - 3.1 TFlops + 33 TFlops
- **Gfarm file system**
  - 12.8 TB (36 GB x 240 + 250 GB x 16 + 480 GB)

```
% df /gfs/home/tatebe
Filesystem           1K-blocks      Used Available Use% Mounted on
gfarmfs           13292988192 3062931612 9554800896  25% /gfs/home/tatebe
```

Japan Lattice Data Grid – Advanced Nationwide Data Sharing

- **Nationwide distributed file system to share QCD data**
  - Transparent data access regardless of the data location
  - Efficient data access with fault tolerance thanks to incorporated file replicas management
  - Flexible capacity management



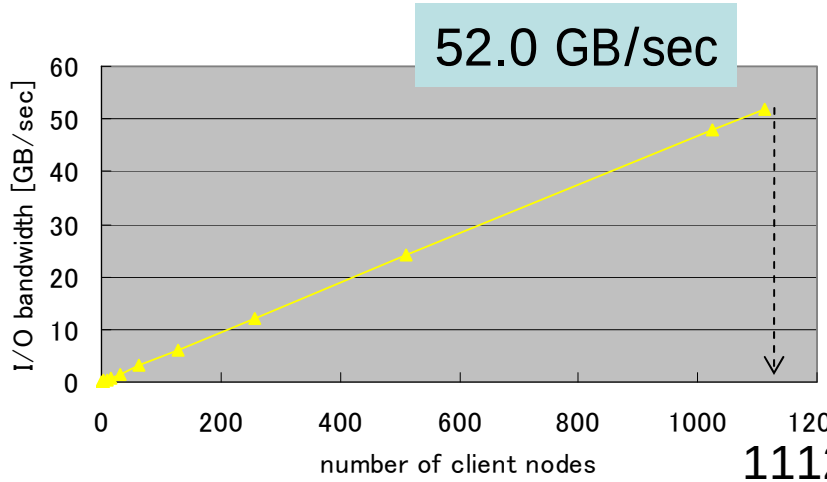
- **Virtual Organization (VO) membership management**
  - Project base, independent from real organizations
  - VO based (project based) Access control
  - Easy access with single sign on

- **Software packaging for advanced data sharing**
  - Commodity hardware and open source software
  - Globus, VOMS, Naregi-CA, Gfarm, Uberftp, . . .
  - Easy deployment

Particle Physics Data Analysis

- **O. Tatebe et al, "High Performance Data Analysis for Particle Physics using the Gfarm File System", SC06 HPC Storage Challenge, Winner – Large Systems, 2006**

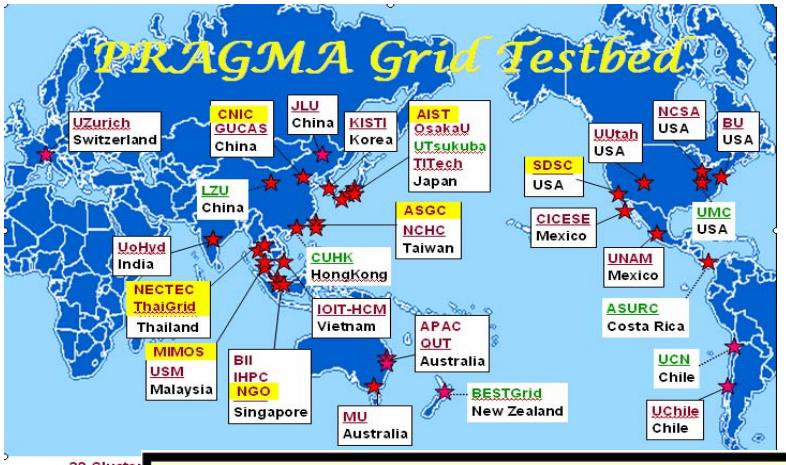
• Construct 26 TB of Gfarm FS using **1112** nodes
• Store all 24.6 TB of Belle experiment data
• 52.0GB/s in parallel read
→ **3,024** times speedup
• 24.0GB/s in skimming process for b → s γ decays using 704 nodes
→ **3 weeks to 30 minutes**

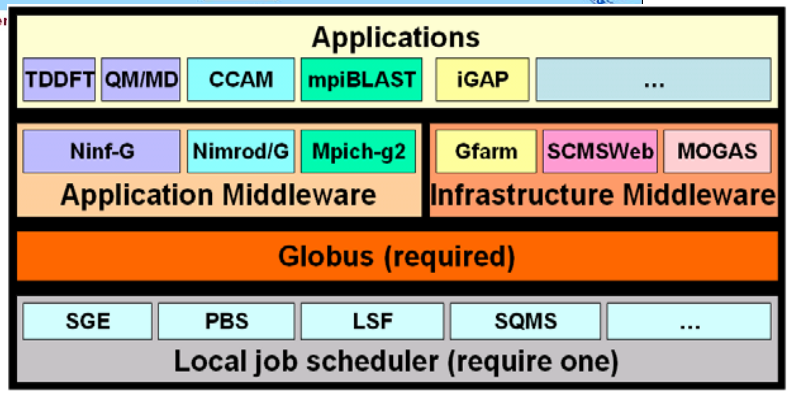


PRAGMA Grid

- **C. Zheng, O. Tatebe et al, "Lessons Learned Through Driving Science Applications in the PRAGMA Grid", Int. J. Web and Grid Services, Inderscience Enterprise Ltd., 2007**

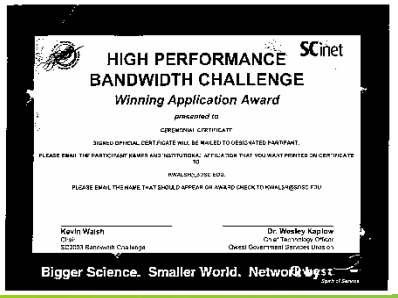
• Worldwide Grid testbed consisting of 14 countries, 29 institutes
• Gfarm file system is used for file sharing infrastructure
→ **executable, input/output data sharing possible in Grid**
→ no explicit staging to a local cluster needed





More Feature of Gfarm Grid File System

- **Commodity PC based scalable architecture**
  - Add commodity PCs to **increase storage capacity** in operation much more than **petabyte scale**
    - Even PCs at distant locations via internet
- **Adaptive replica creation and consistent management**
  - Create multiple file replicas to **increase performance and reliability**
  - Create file replicas at distant locations for **disaster recovery**
- **Open Source Software**
  - Linux binary packages, ports for *BSD, . . .
    - It is included in Naregi, Knoppix HTC edition, and Rocks distribution
  - **Existing applications** can access w/o any modification



SC03 Bandwidth SC|05 StorCloud SC06 HPC Storage Challenge Winner

GRID DATA FARM

Design and implementation of Gfarm v2

Design policy of Gfarm v2

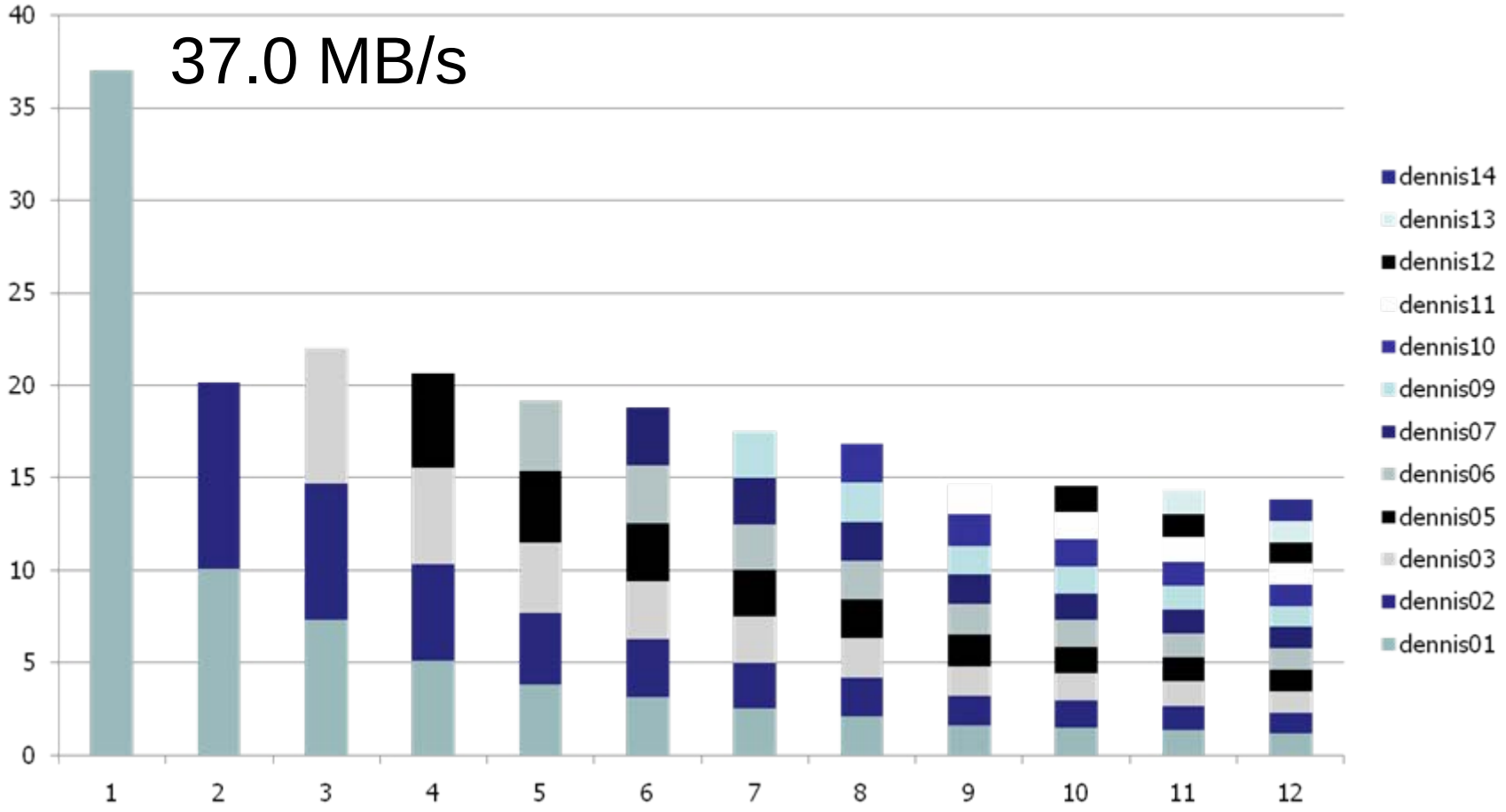
- **Inherit architectural benefit of scalable I/O performance using commodity platform**
- **Design as a POSIX compliant file system**
  - Solve security problems in Gfarm v1
- **Improve performance for small files**
  - Reduce metadata access overhead
- **Grid file system -> Distributed file system**
  - Still benefit from improvement of local file system
- **Compete with NFS, AFS, and Lustre**

Gfarm v2 testbed

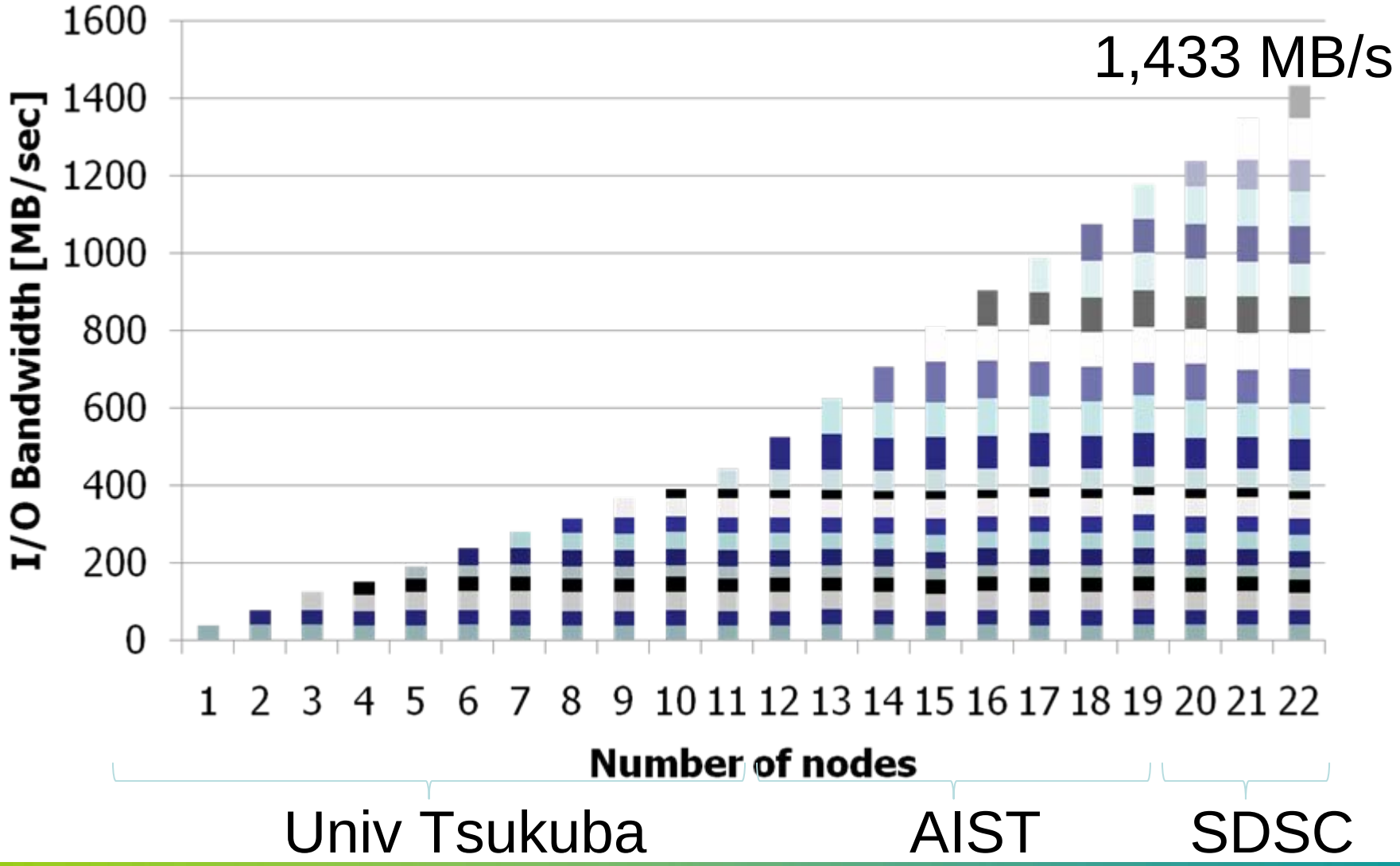
- **Metadata server**
  - Univ of Tsukuba
- **File system nodes**

	Univ Tsukuba	AIST	SDSC
#nodes	14	8	3
RTT [msec]	0.202	0.787	119

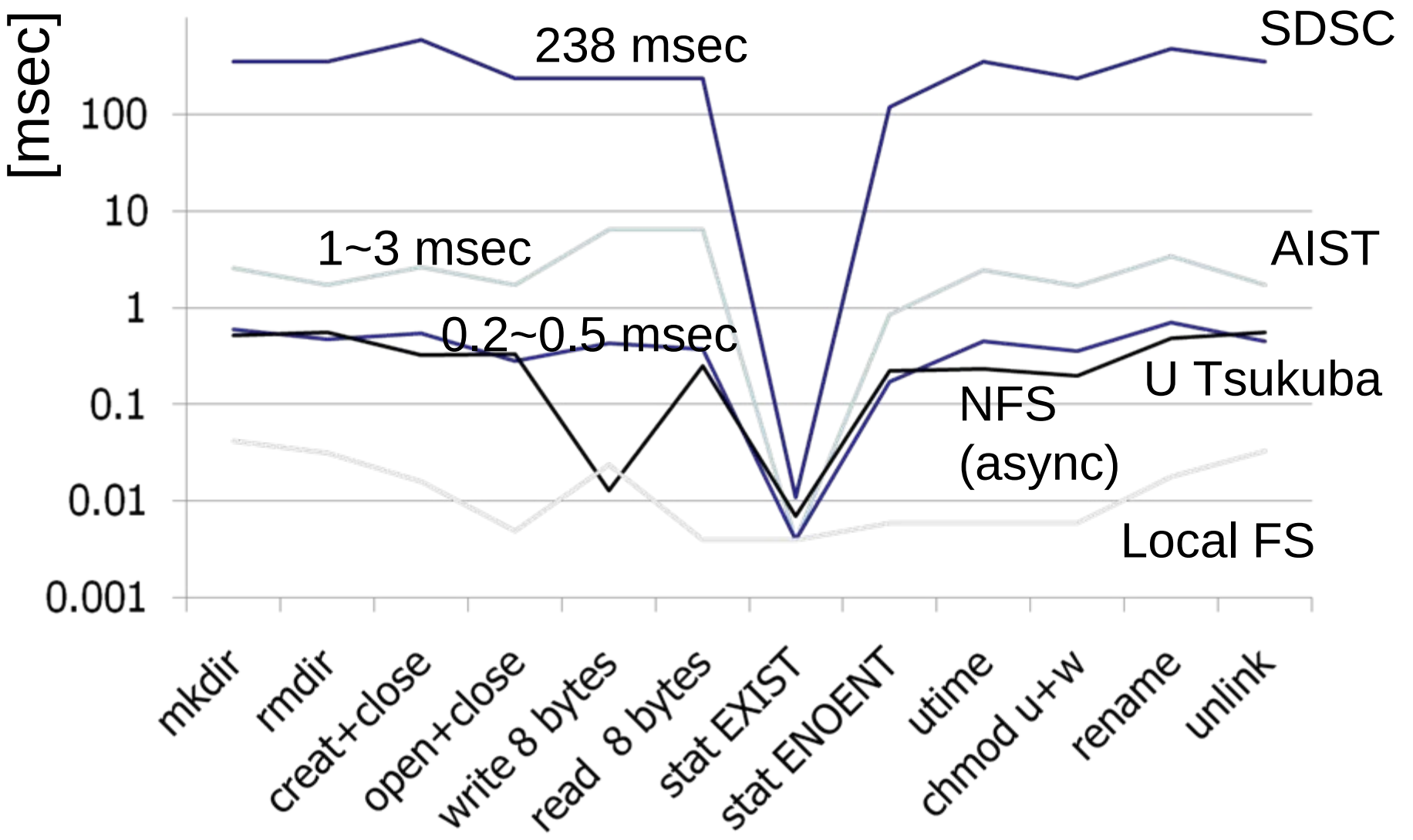
NFS bandwidth (read 1G sep. data)



Gfarm Scalable Bandwidth



Operation latency (2~3RTT)



Summary

- **Gfarm file system**
  - Scalable commodity-based architecture
  - File replicas for fault tolerance and hot spot avoidance
  - Capacity increase/decrease in operation
- **Gfarm v1**
  - Used for several production systems
  - 1000+ clients and file system nodes scalability
- **Gfarm v2**
  - Plug up security hole in Gfarm v1, and improve metadata access performance
  - Comparable performance with NFS for small files in LAN
    - 0.2 ~ 0.5 milliseconds
  - Scalable file IO performance even in distributed environment
    - 1433 MB/sec parallel read IO performance from 22 clients in Japan and US
- **Open Source Development**
  - http://sourceforge.net/projects/gfarm