Research, Development and Scientific Application of Gfarm File System

Masahiro Tanaka

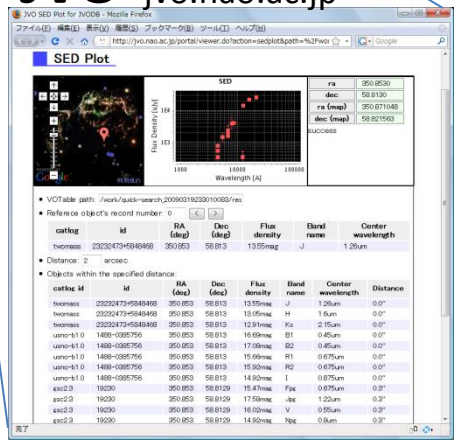
University of Tsukuba

Self Introduction

- I am originally a researcher in Astronomy.
- I was engaged in development of **Japanese Virtual Observatory (JVO)**.
- I am engaged in research on **Gfarm** since last April.
  - Case studies in Astronomy.

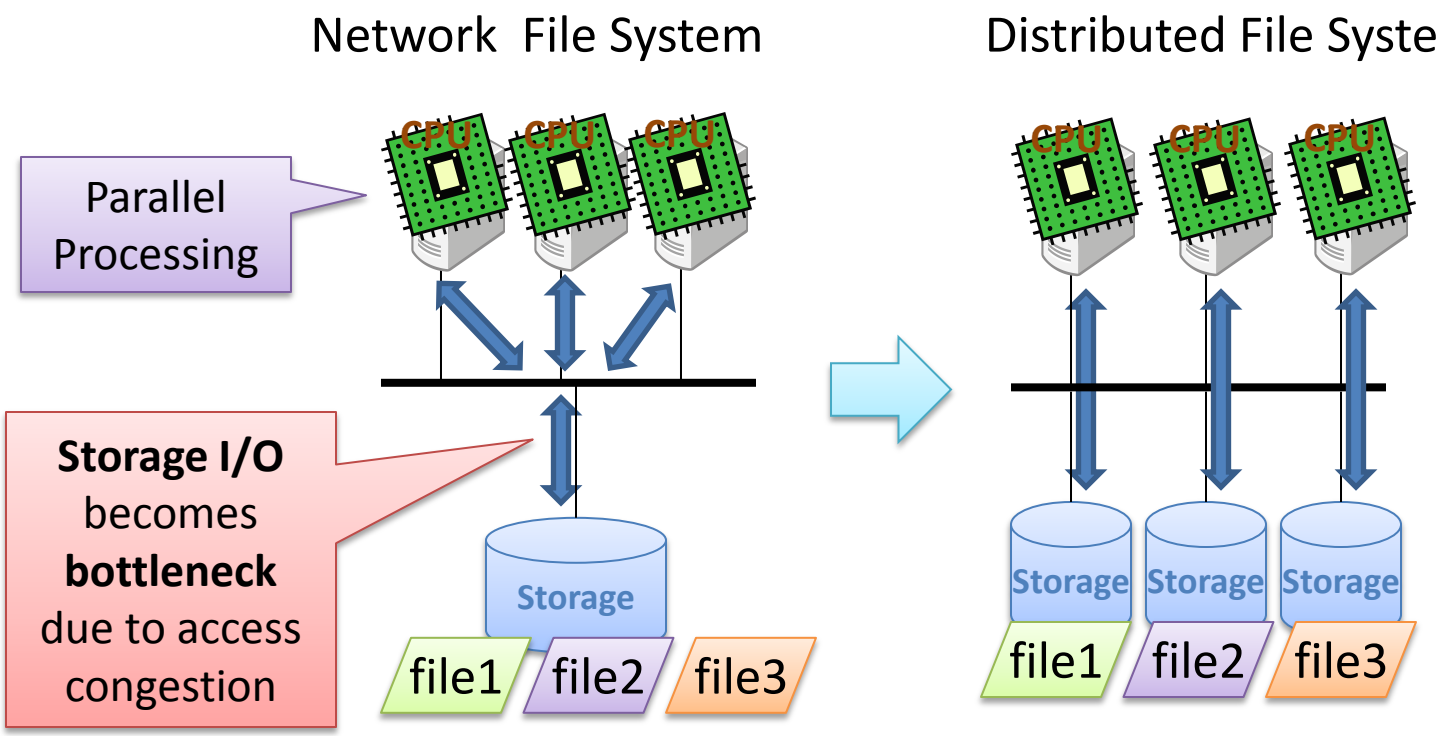
International Virtual Observatory Alliance (IVOA)

JVO jvo.nao.ac.jp



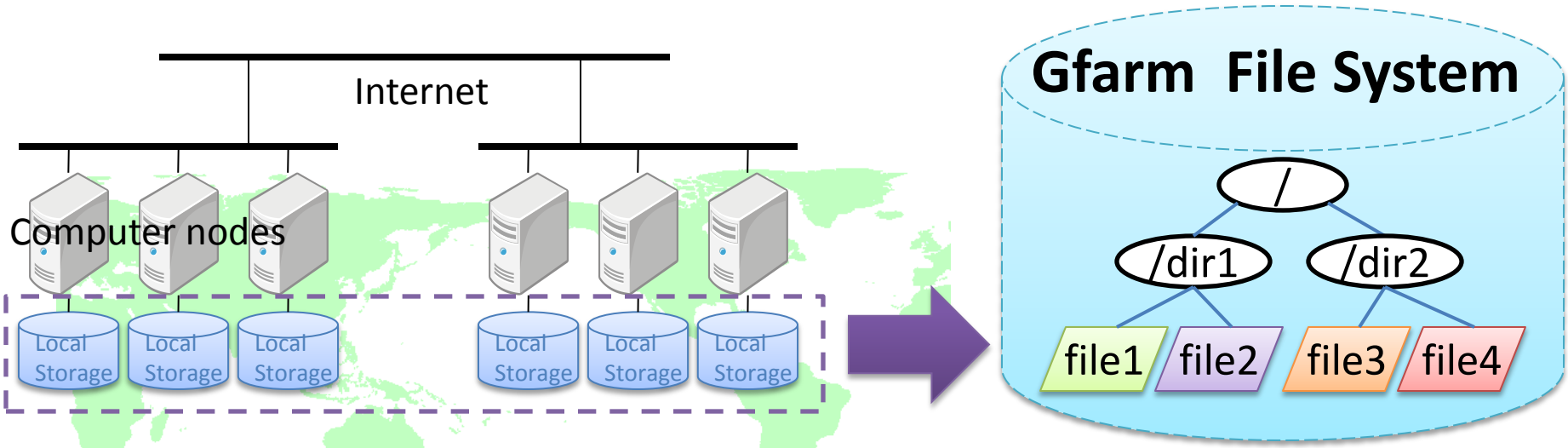
Introduction of Gfarm File System

Needs for Distributed File System

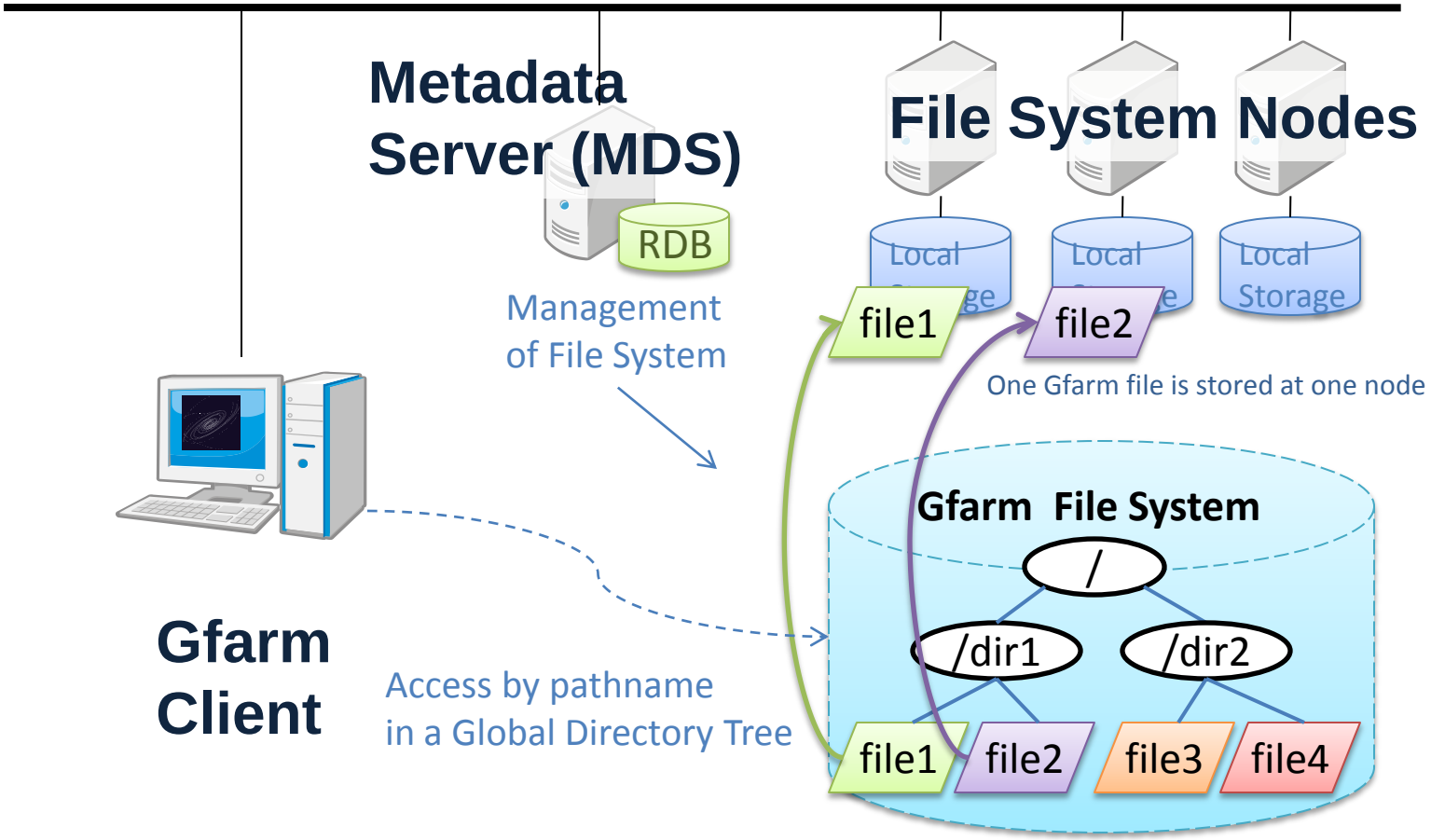


Gfarm

- Wide-area distributed file system
- Global namespace to federate storages
- Main developer : Osamu Tatebe
- Open source development
  - http://datafarm.apgrid.org/



Gfarm Components



Gfarm Programs

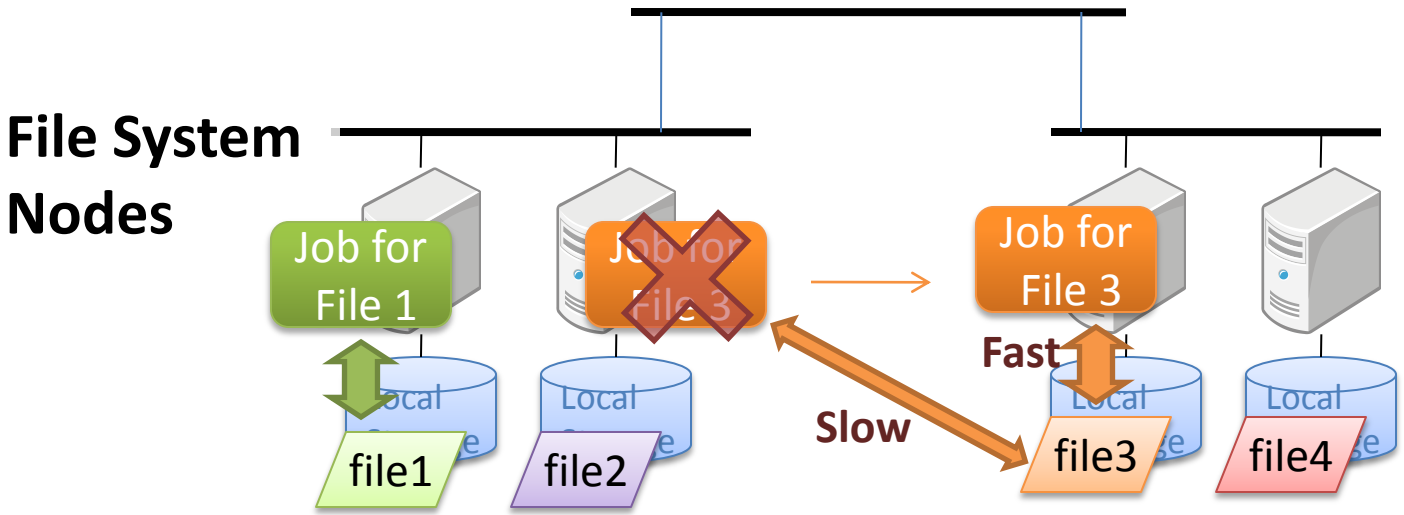
- Metadata Server daemon : **gfmd**
- FileSystem node daemon : **gfsd**
- Mount Gfarm using FUSE : **gfarm2fs**
- Unix-command-like user programs :
  - **gfls, gfcp, gfmv, gfrm, gfmkdir, gfrmdir, gfln, gfdf**
  - **gfuser, gfgroup, gfchmod, gfchown, gfchgrp**
- Manipulate Gfarm files
  - **gfreg, gfexport, gfrep, gfwhere**
- Extended attribute
  - **gfxattr, gffindxmlattr**

Gfarm Security

- Authentication
  - Shared key (for non-Grid Cluster)
  - GSI (for Grid)
- Access restriction to Gfarm files
  - Users
  - Groups

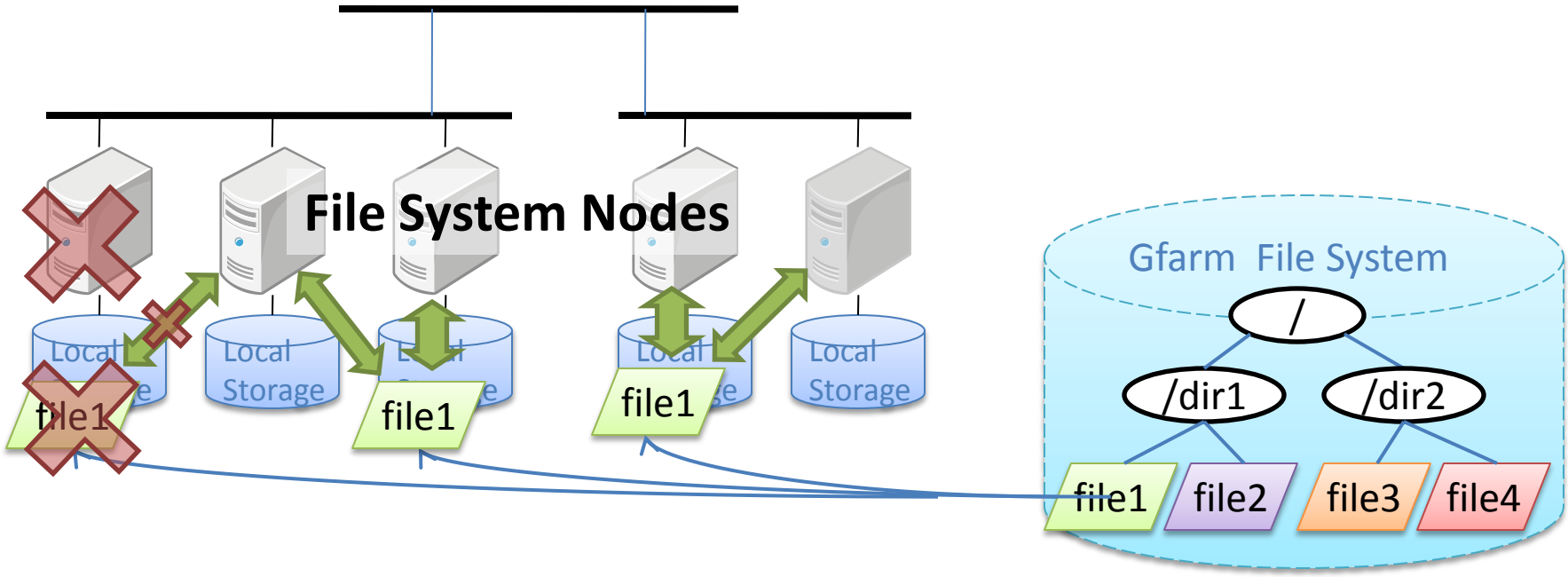
File Affinity Job Scheduling for scalable I/O performance

- Exploit local I/O for scalable I/O performance
- File Affinity Job Scheduling is a key
  - Move and execute program instead of moving large-scale data



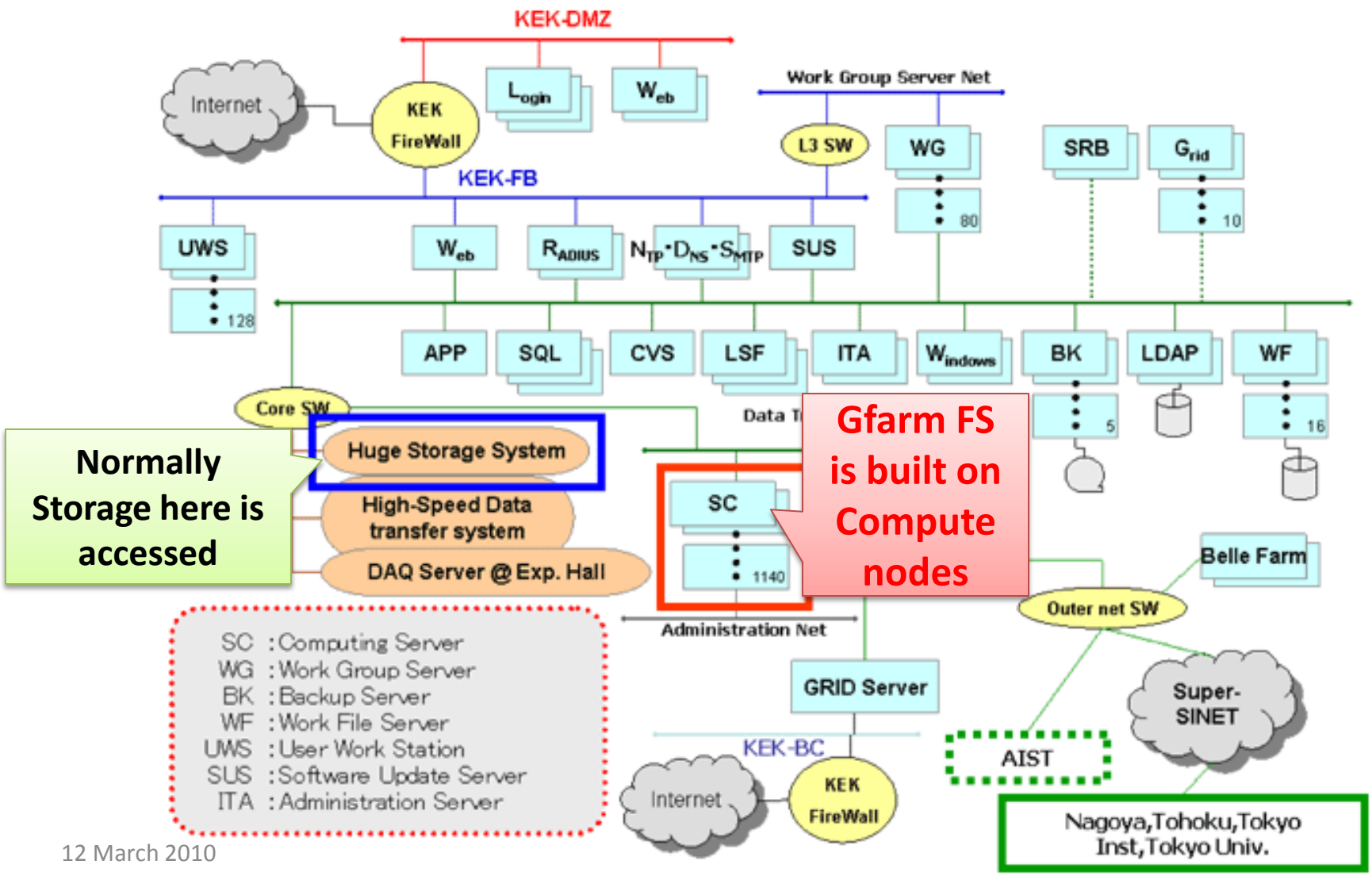
Automatic File Replica Selection

- Files may be replicated and stored in any file system node
  - Fault tolerance
  - Avoids access concentration



Gfarm applications to HEP

Computers for Belle experiment



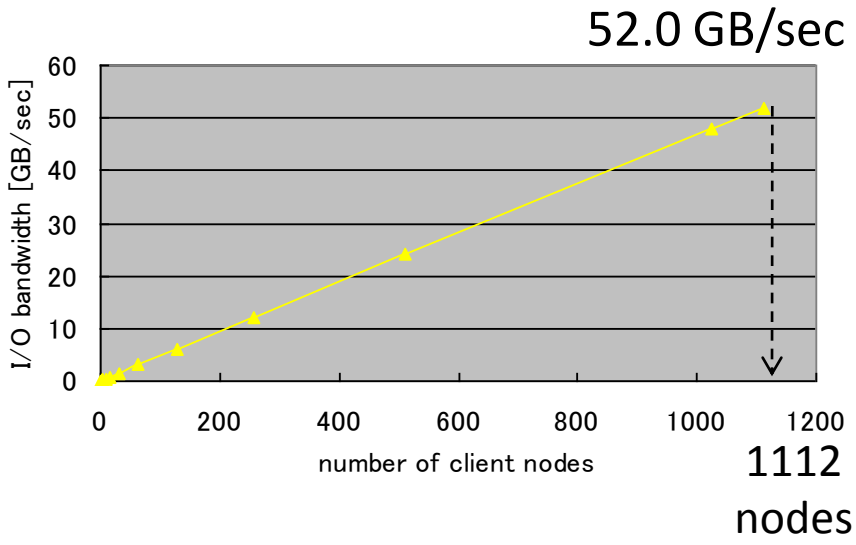
Belle data analysis with Gfarm

- *S. Nishida, N. Katayama, I. Adachi, O. Tatebe, M. Sato, T. Boku, A. Ukawa, "High Performance Data Analysis for Particle Physics using the Gfarm file system", Journal of Physics: Conference Series, 119, 062039, 2008*

- 26 TB of Gfarm FS is constructed with **1112** nodes
- 24.6 TB of Belle experiment data are stored.

- Read performance: **52.0 GB/s**
- Performance of skimming process for b → s γ decays (704 nodes used) : **24.0 GB/s**

→ **3 weeks to 30 minutes**

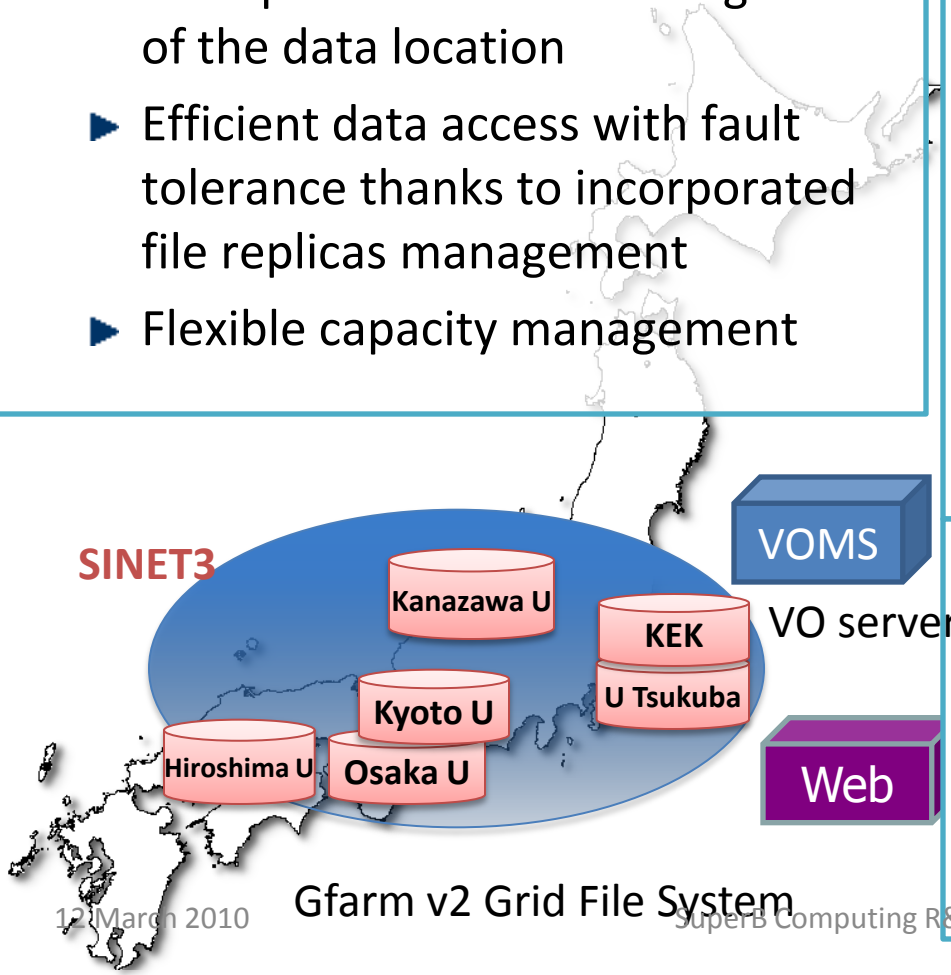


Japan Lattice Data Grid (JLDG)

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

- **Nationwide File Sharing within a research group (VO)**
  - Single sign on
  - Efficient file sharing from distant locations
    - Fast file replication
    - Replica consistency management
  - User and group (VO) based Access control

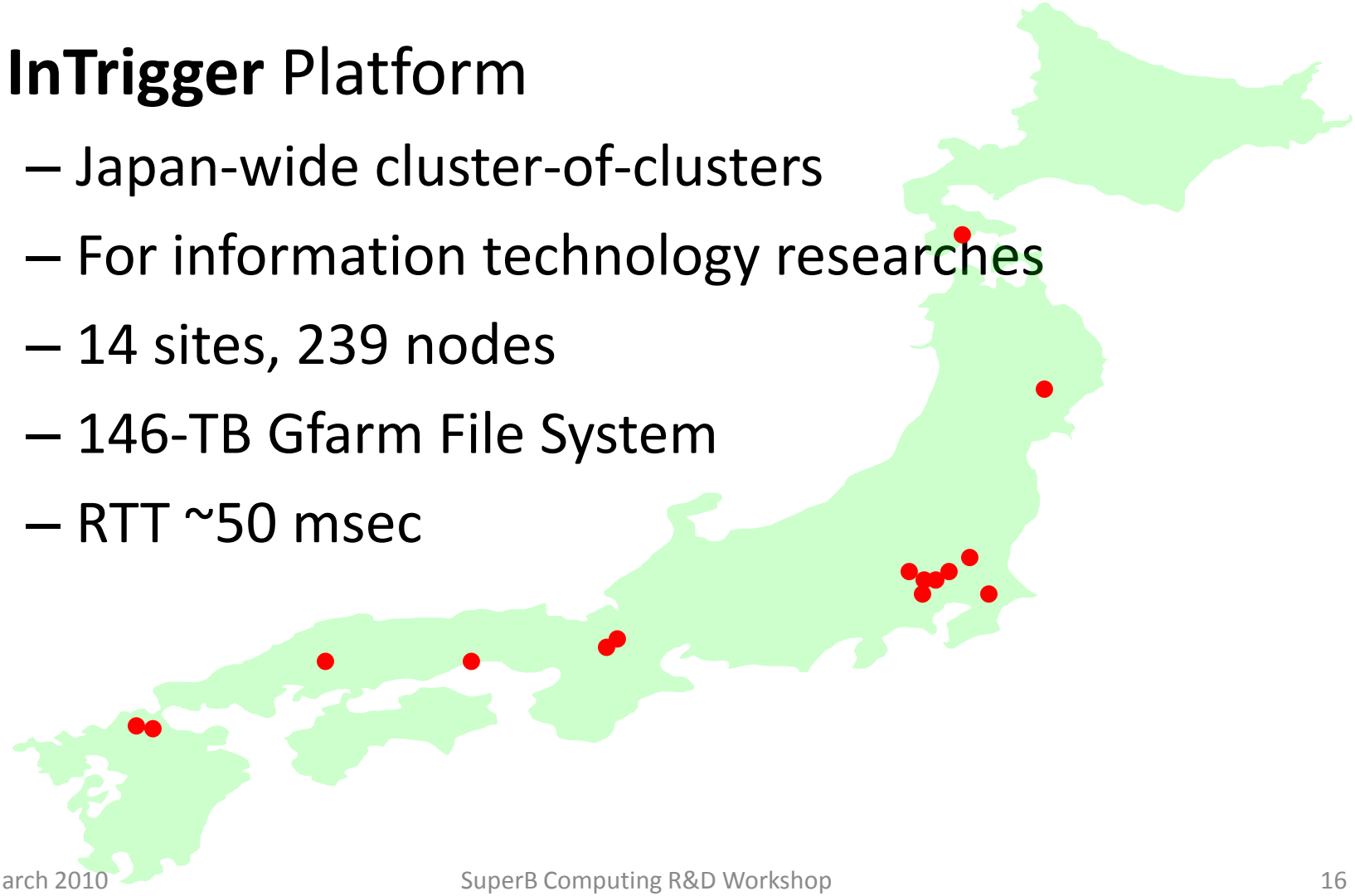
- **Data archive operation**
  - Large-scale QCD data stored in a nationwide distributed file system can be accessed directly through Web and GridFTP



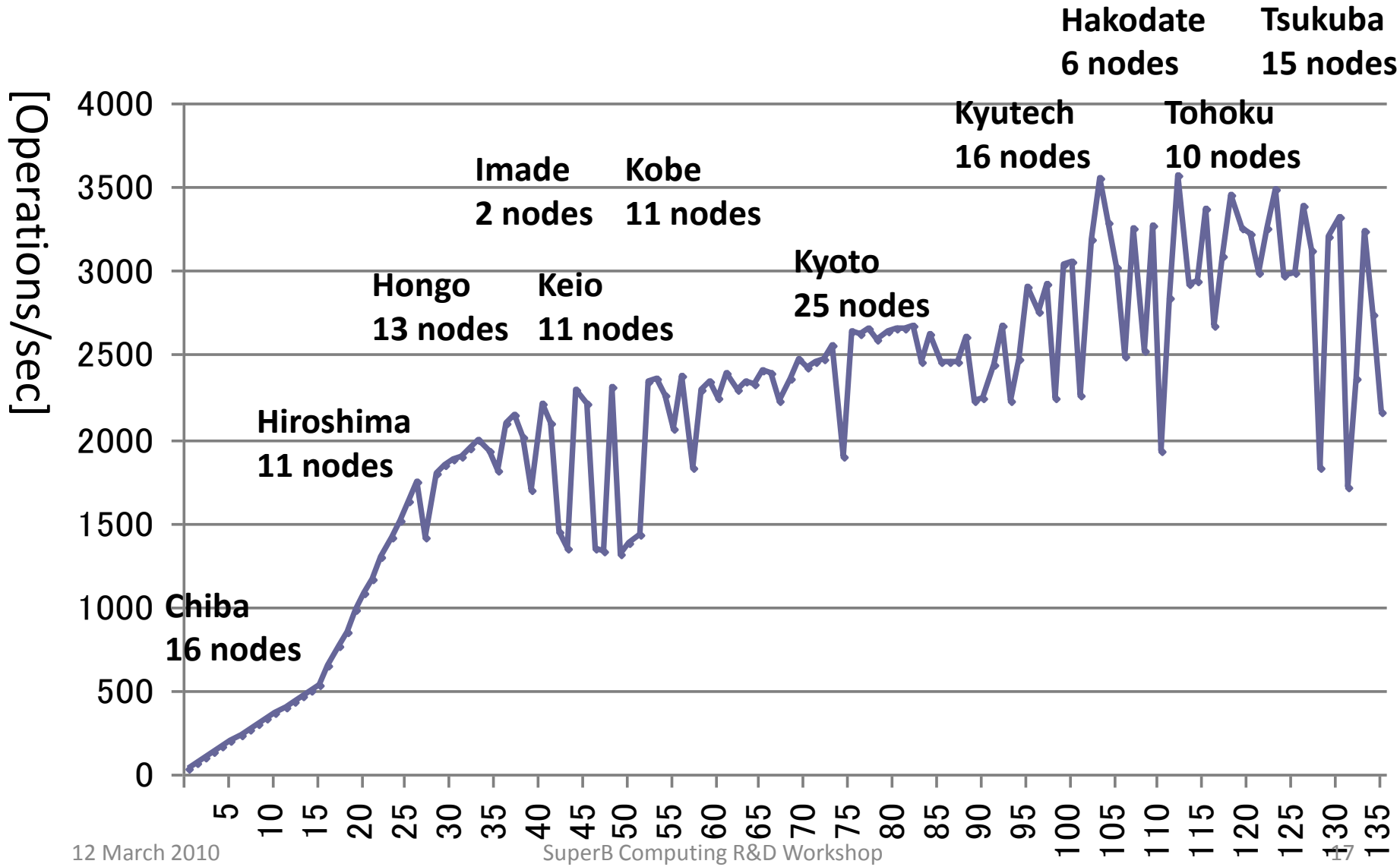
Performance of Gfarm in geographically-distributed clusters

Geographically-distributed platform

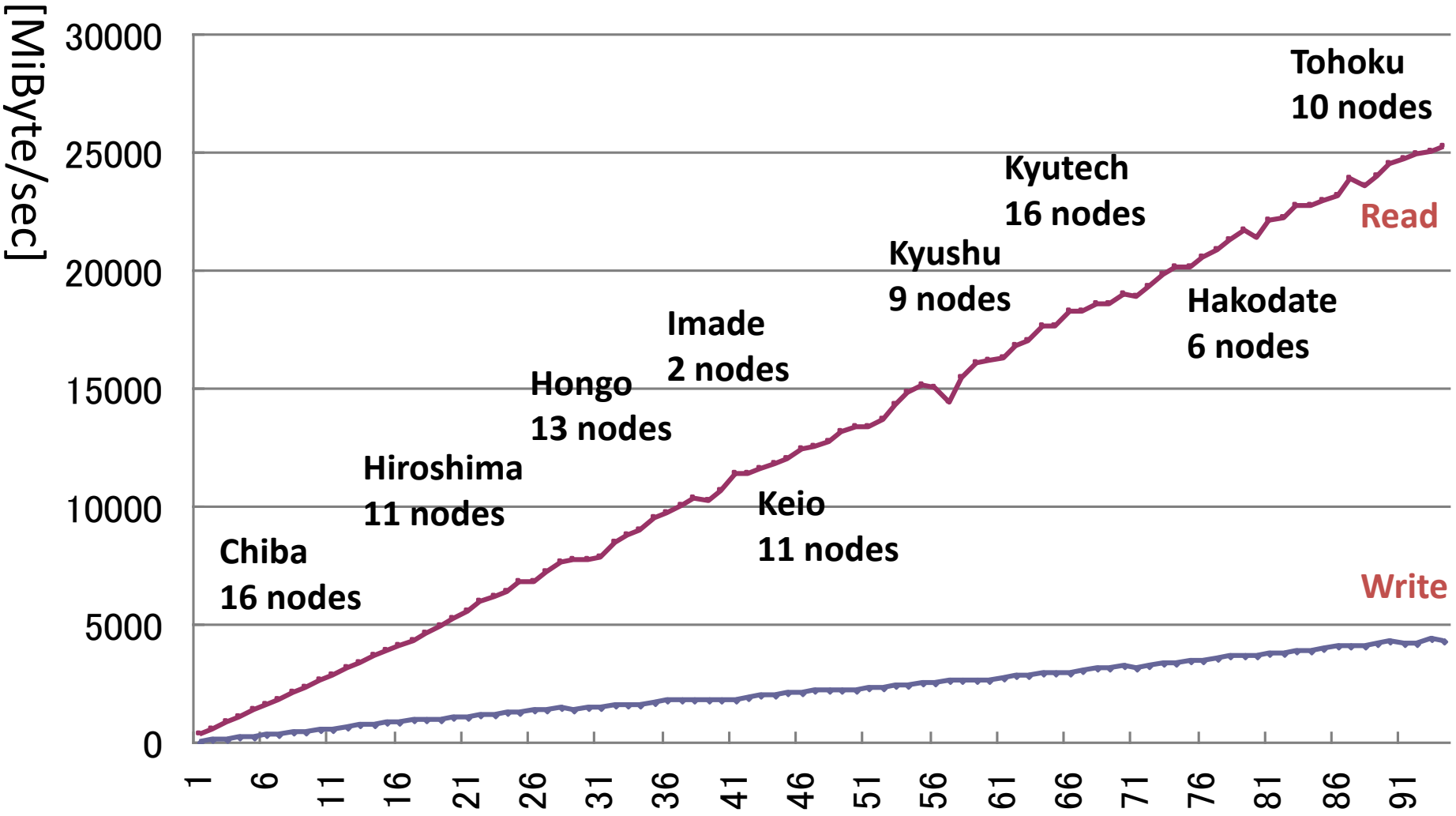
- **InTrigger** Platform
  - Japan-wide cluster-of-clusters
  - For information technology researches
  - 14 sites, 239 nodes
  - 146-TB Gfarm File System
  - RTT ~50 msec



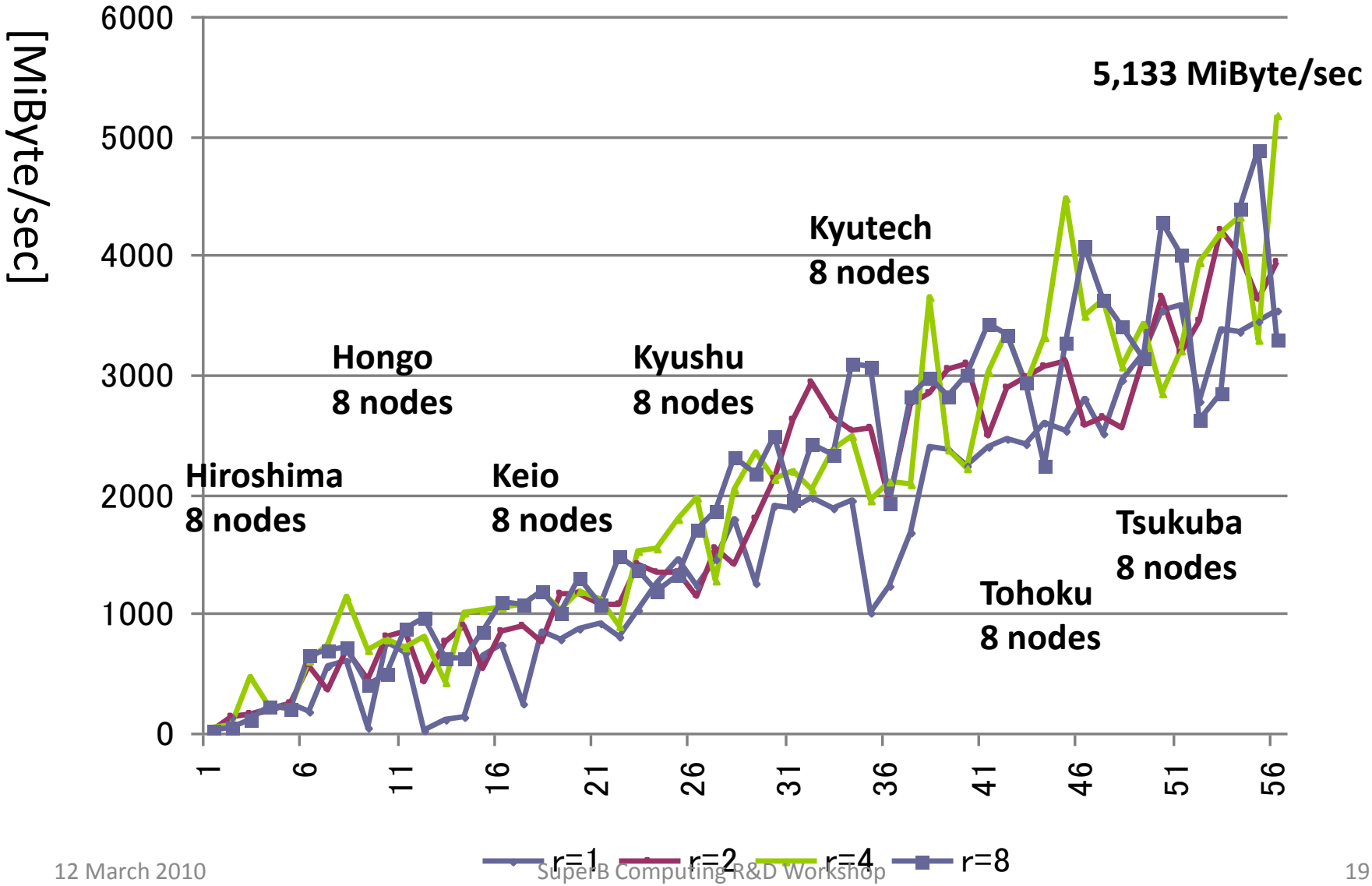
Metadata operation (mkdir)



Read/Write N Separate 1GiB Data



Read Shared 1GiB Data



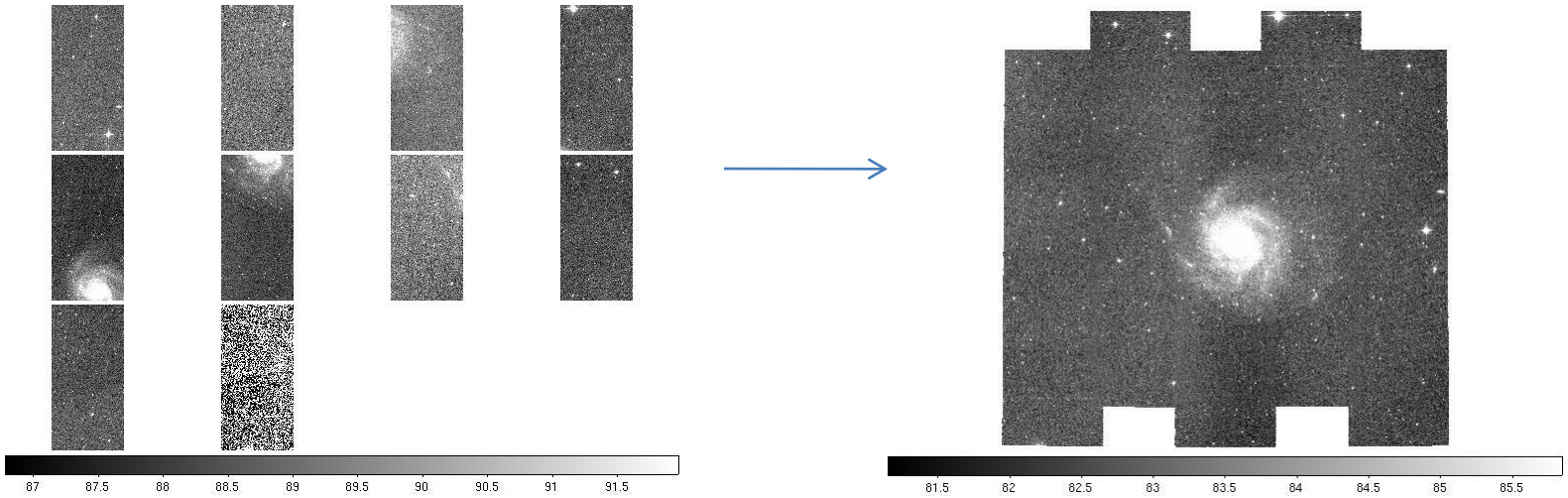
Summary of wide-area performance of Gfarm

- In geographically-distributed clusters
- Metadata operation (mkdir)
  - 3,570 ops/sec
- I/O performance
  - 4,370 MiB/sec of write bandwidth
  - 25,170 MiB/sec of read bandwidth
    - by 94 nodes that scales in wide area
  - 5,133 MiB/sec of bandwidth for reading a shared data
    - by 56 nodes that also scales in wide area

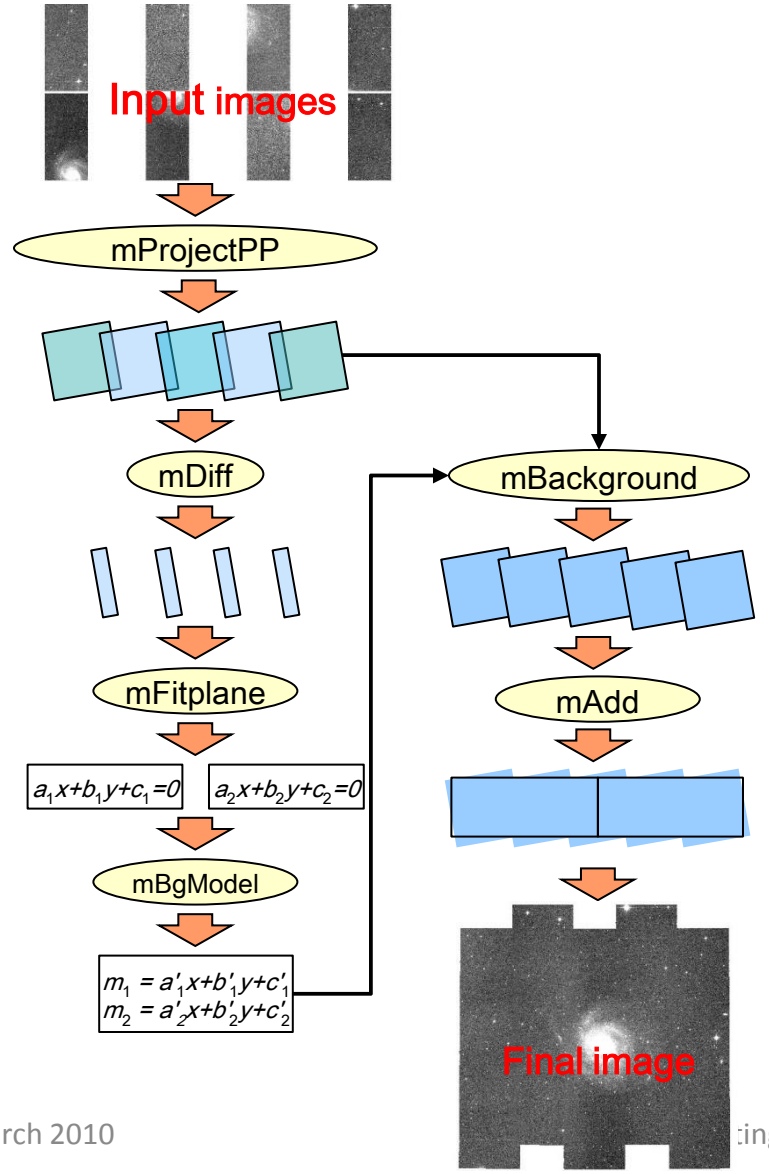
Case Study in Astronomy

Montage : Tool for Astronomical Image Processing

- A tool for producing a custom mosaic image from multiple shots of images.
  - http://montage.ipac.caltech.edu/



Montage Workflow

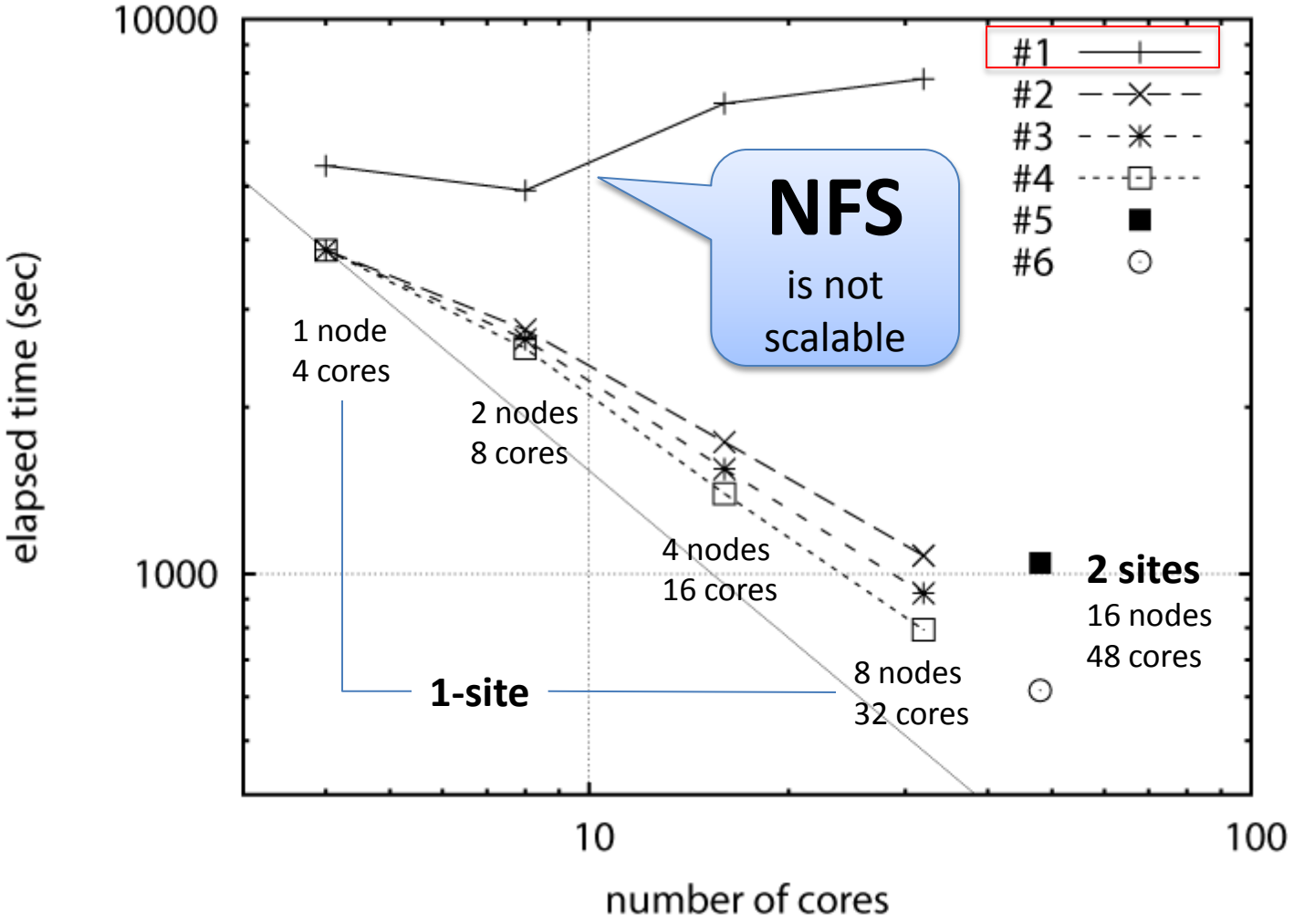


- 1 image : 1 process
- Parallel processes for multiple images
- Complicated workflow
- Programs are not modified for Gfarm

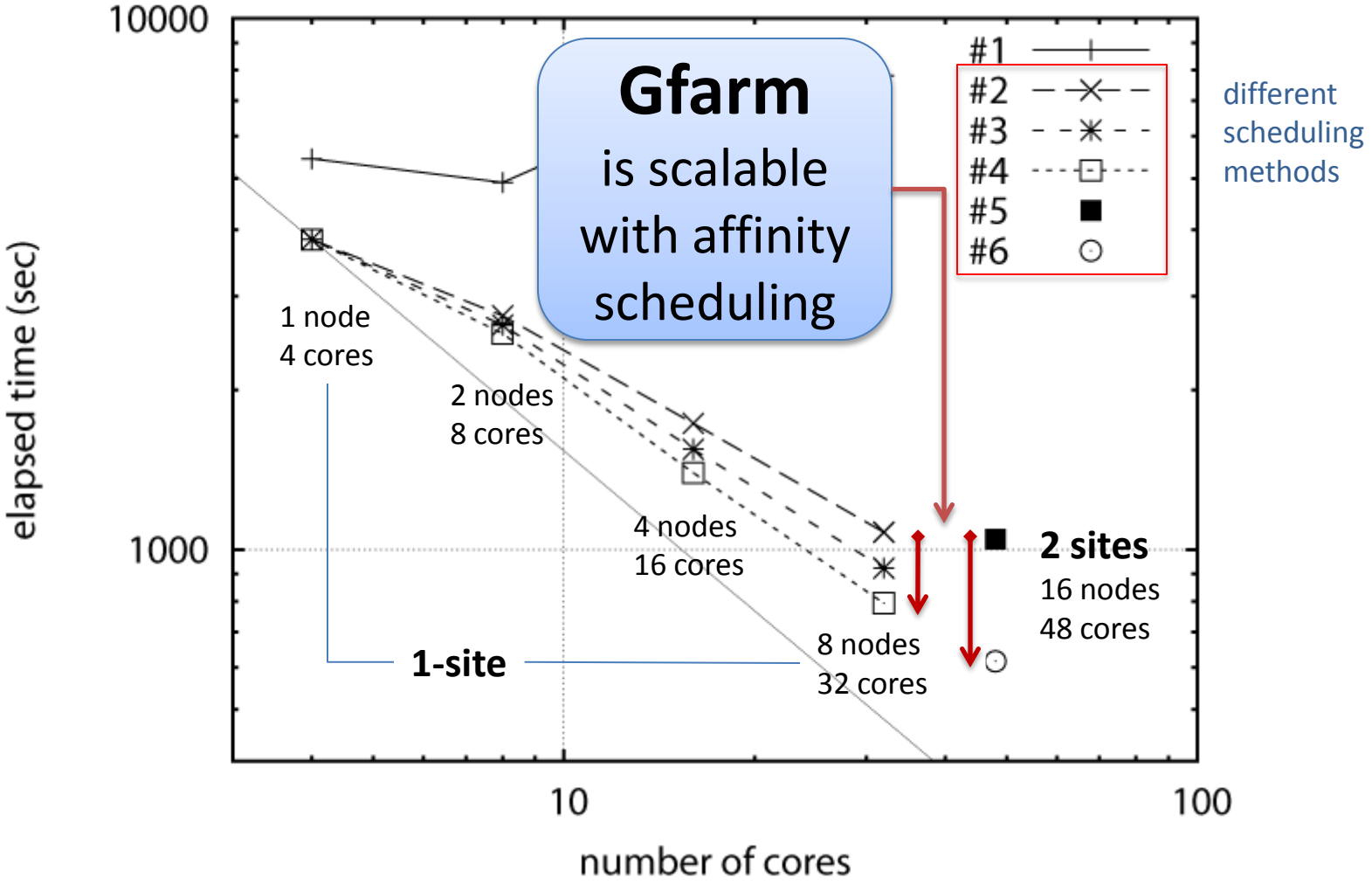
Development of Workflow Tool

- Requirements for workflow tool:
  - Affinity Scheduling
  - Powerful description of workflow
    - Loops & Conditions as well as Task dependency
- Rake
  - Ruby version of **make**
  - Powerful description than **Makefile**
- **Pwrake**
  - **Parallel workflow extension for Rake**
    - Parallel execution on remote hosts
    - **Plug-able Scheduling method: enables Affinity Scheduling**

Performance of Montage workflow



Performance of Montage workflow



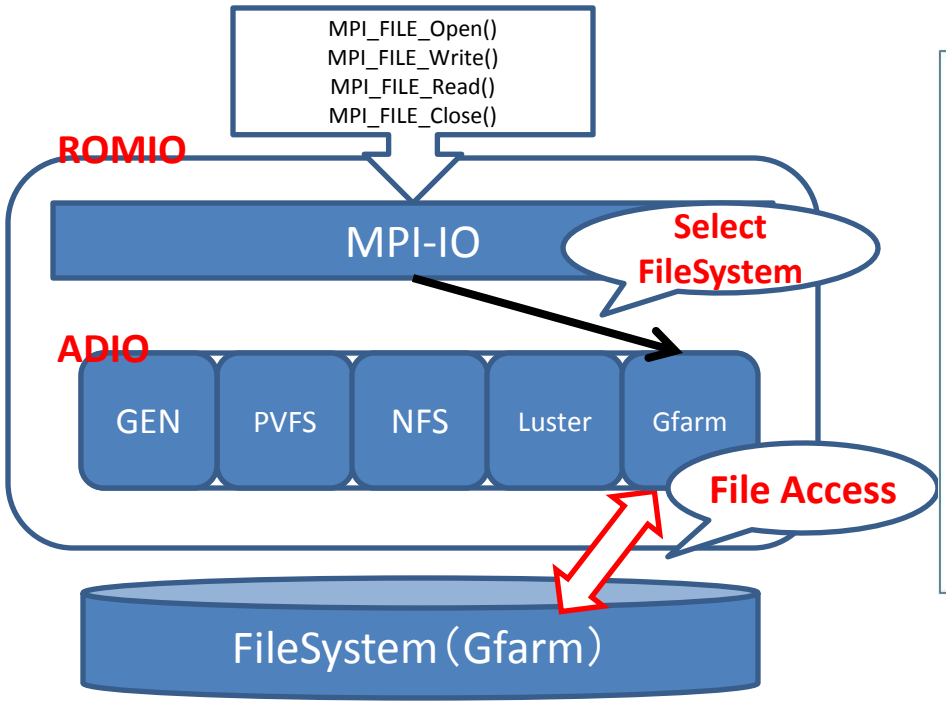
Other Researches on Gfarm

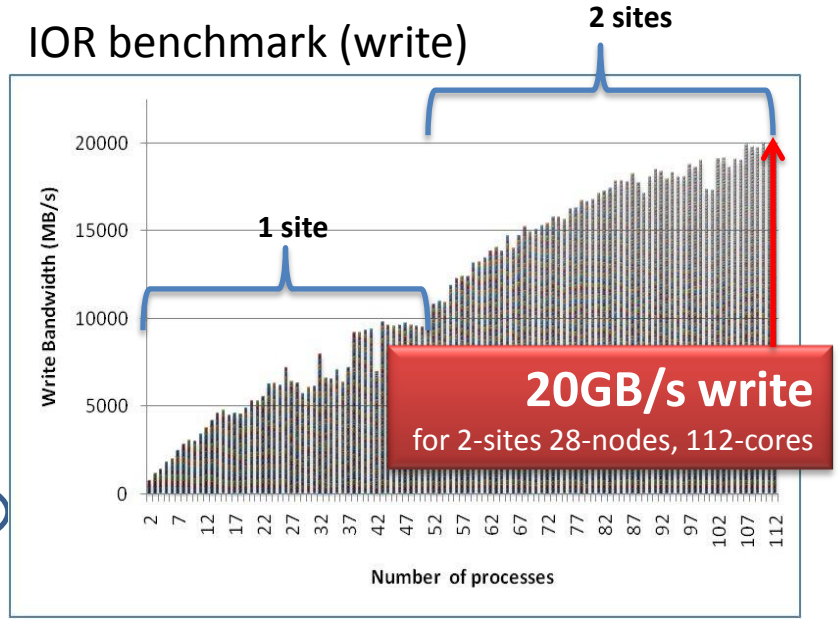
Current Researches on Gfarm

- Fast Replication between clusters (K.Suzuki)
- Multiple Metadata Server (K.Hiraga)
- MPI-IO (H.Kimura)
  - Scalable parallel write by File-view
- Gfarm in Cloud (K.Kobayashi)
  - Amazon EC2, Eucalyptus
- Hadoop-Gfarm (S.Mikami)
  - Gfarm FS instead of HDFS

Scalable MPI-IO by Gfarm

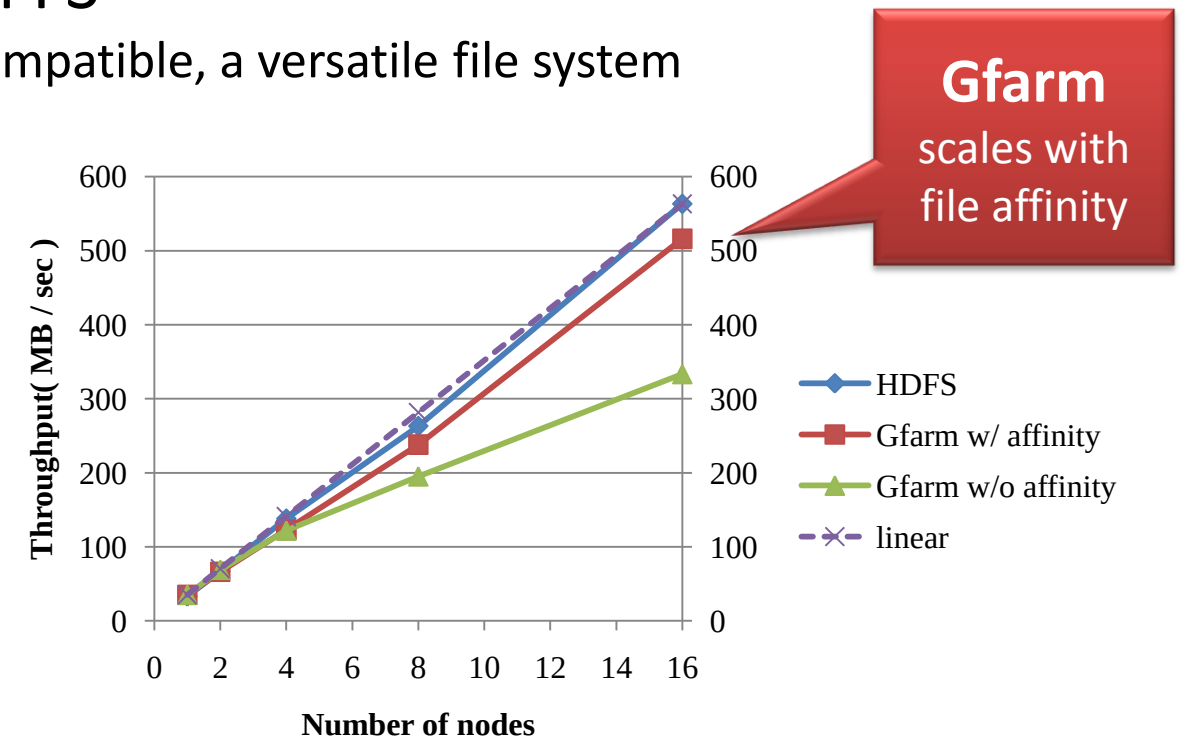
- Implement ADIO (Abstract-device interface for parallel I/O) for Gfarm
- File view :
  - Single File of MPI-IO <=> Distributed Files in Gfarm





Hadoop-Gfarm Plug-in

- instead of HDFS (Hadoop Distributed File System)
  - POSIX incompatible, specific for MapReduce
- use Gfarm FS
  - POSIX compatible, a versatile file system



Conclusions

- Gfarm File System
  - High availability, high performance, and cost effective open-source wide-area distributed Grid file system
- Belle experiment data processing
  - 24.0 GB/s with 704 nodes
- Performance in geographically distributed clusters
  - 5,133 MiB/sec of bandwidth for reading a shared data by 56 nodes that also scales in wide area
- Astronomical image processing
  - Scales with File Affinity Scheduling
- Various researches on Gfarm are ongoing.