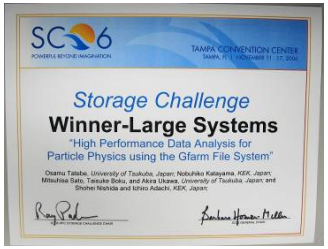
OpenSFS APAC Lustre User Group 2013 Tokyo
October 17, 2013

Gfarm: Present Status and Future Evolution

Osamu Tatebe
University of Tsukuba

Gfarm file system

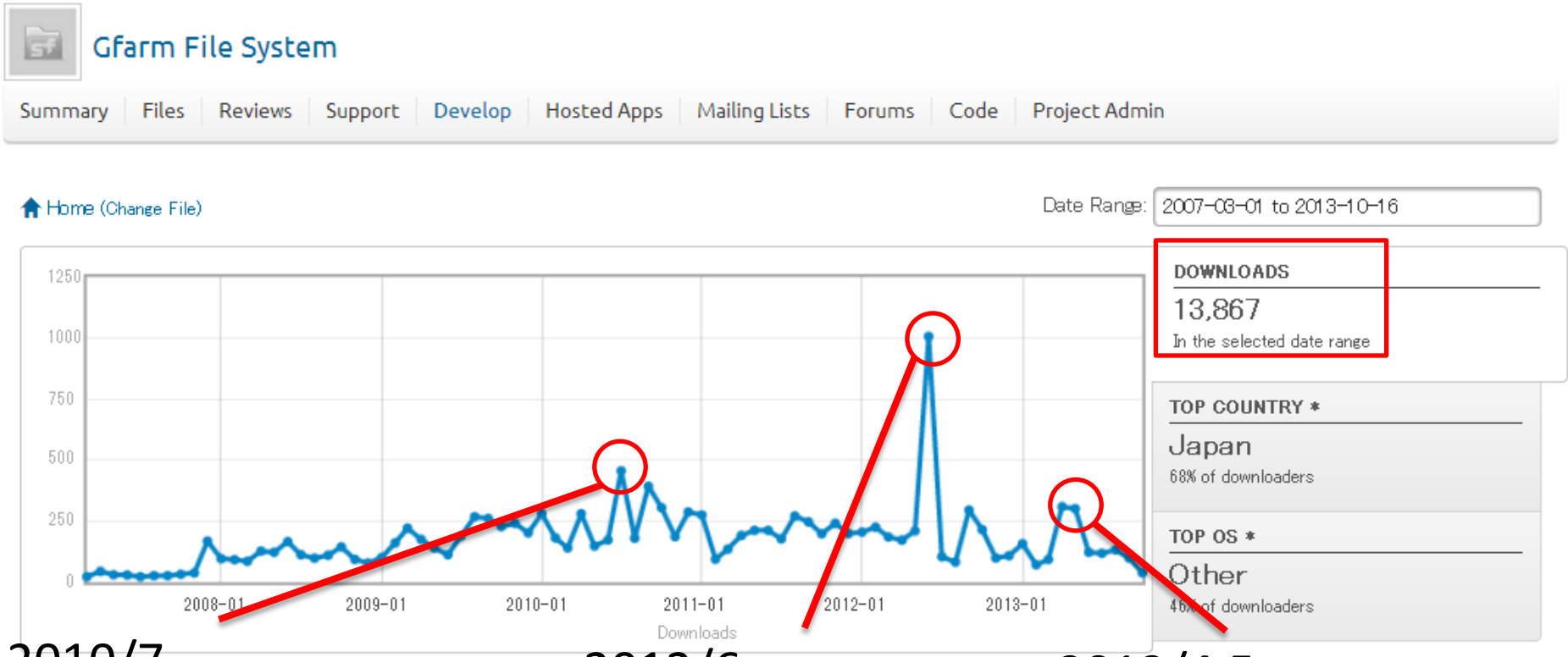


- Award-winning file system since 2000
  - Distributed infrastructure award in SC03
  - Most Innovative Use of Storage In Support of Science Award in SC05
  - Winner – Large Systems in HPC Storage Challenge in SC06
- Open Source distributed file system
  - http://sf.net/projects/gfarm/
- Supported by NPO OSS Tsukuba Support Center
- Features
  - Scaled-out performance in wide area
  - Data access locality, file replica
  - No single point of failure
    - Automatic file replica creation in case of storage failure
    - Hot stand-by MDS

ossTsukuba

downloads

- 13,867 downloads since March, 2007



2010/7
Version 2.3.2, 2.4.0
456 downloads

2012/6
HPCI installation etc
1,007 downloads

2013/4,5
Version 2.5.8
610 downloads

HPCI SHARED STORAGE

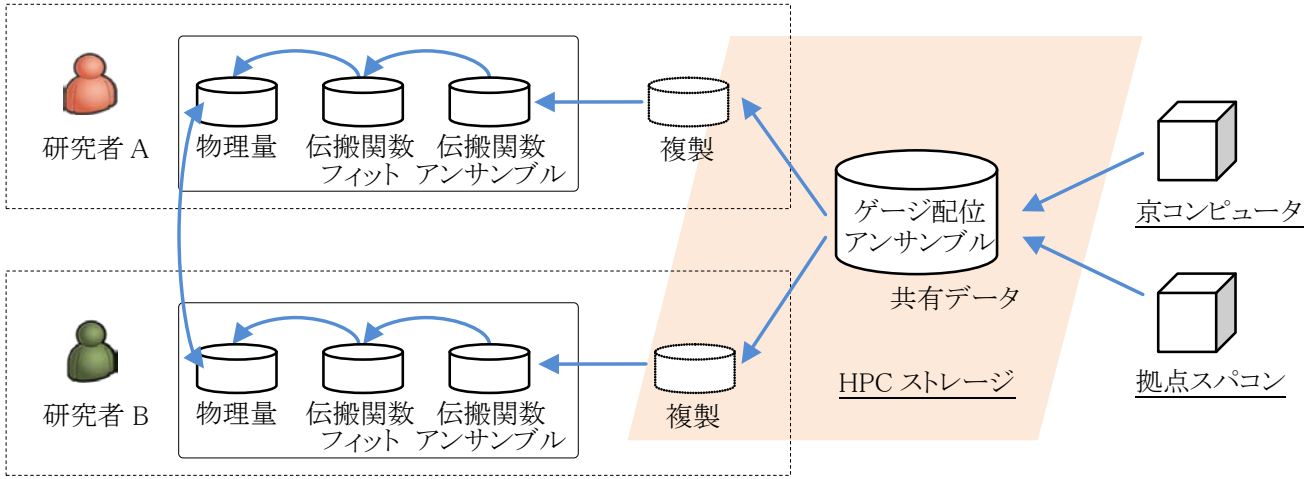
HPCI Shared Storage

- HPCI – High Performance Computing Infrastructure
  - “K”, Hokkaido, Tohoku, Tsukuba, Tokyo, Titech, Nagoya, Kyoto, Osaka, Kyushu, RIKEN, JAMSTEC, AIST
- A 20PB single distributed file system consisting East and West sites
- Grid Security Infrastructure (GSI) for user ID
- Parallel file replication among sites
- Parallel file staging to/from each center

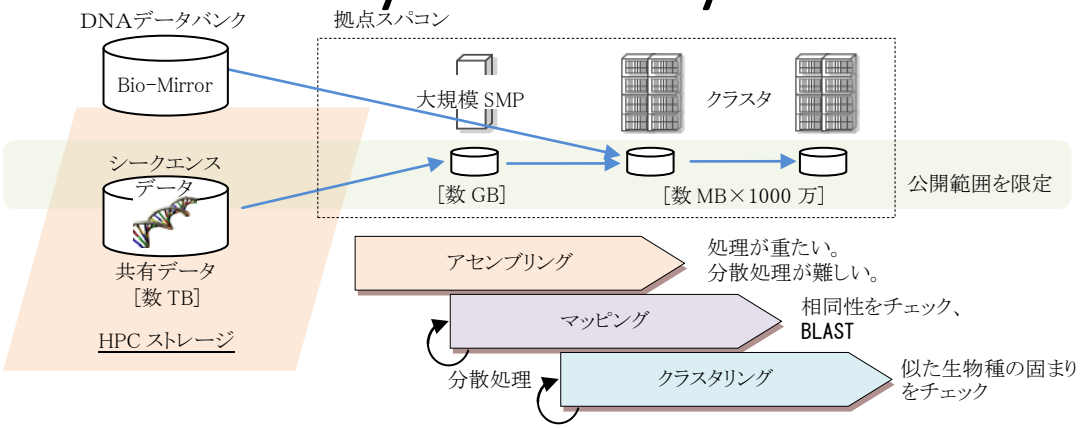
Picture courtesy by Hiroshi Harada (U Tokyo)

Usage Scenario

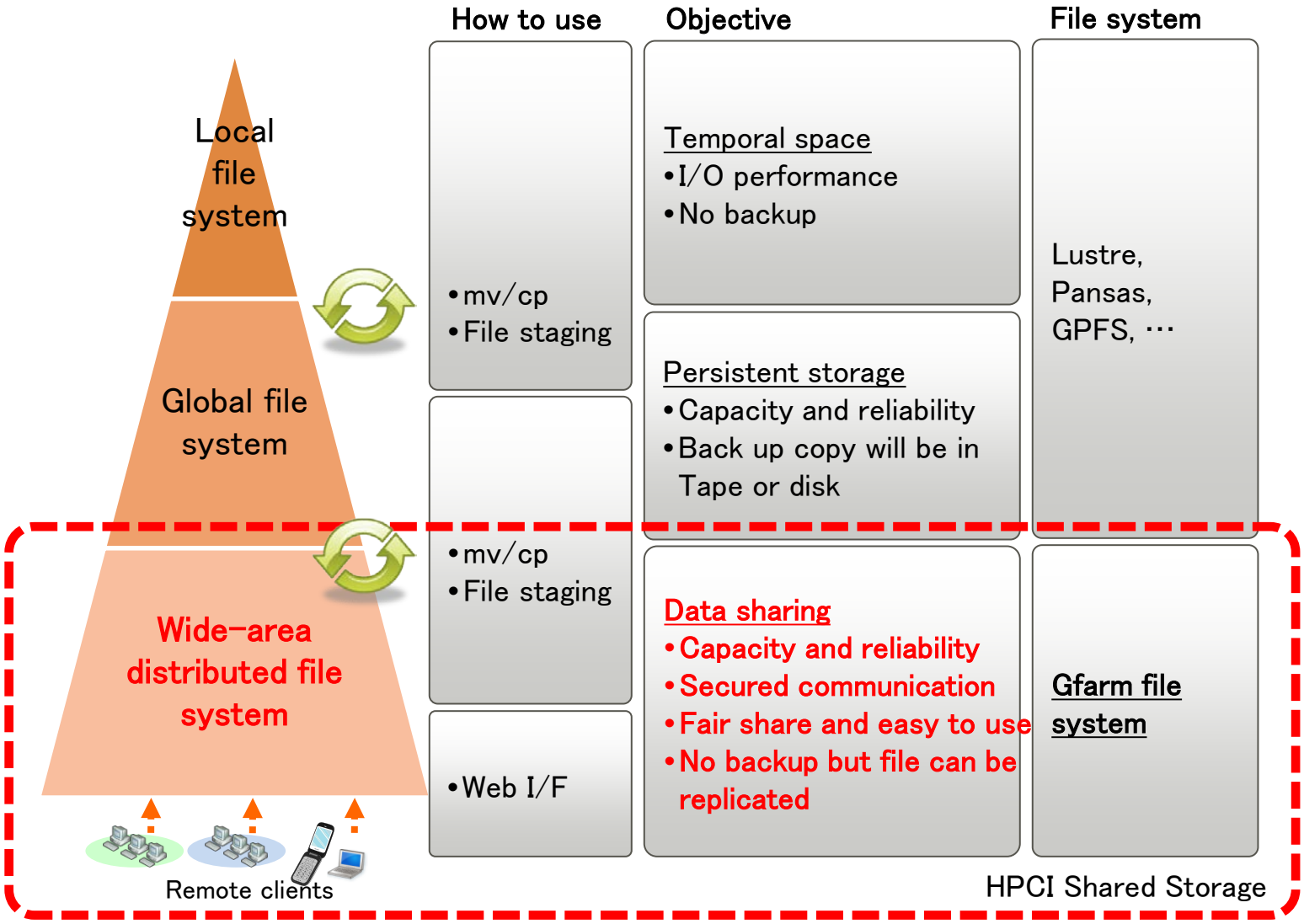
- QCD



- Large scale simulation (ex. cosmic simulation)
  - Simulation data obtained by the K computer, will be analyzed and visualized by University's supercomputer
- Life Science



Storage structure of HPCI Shared Storage



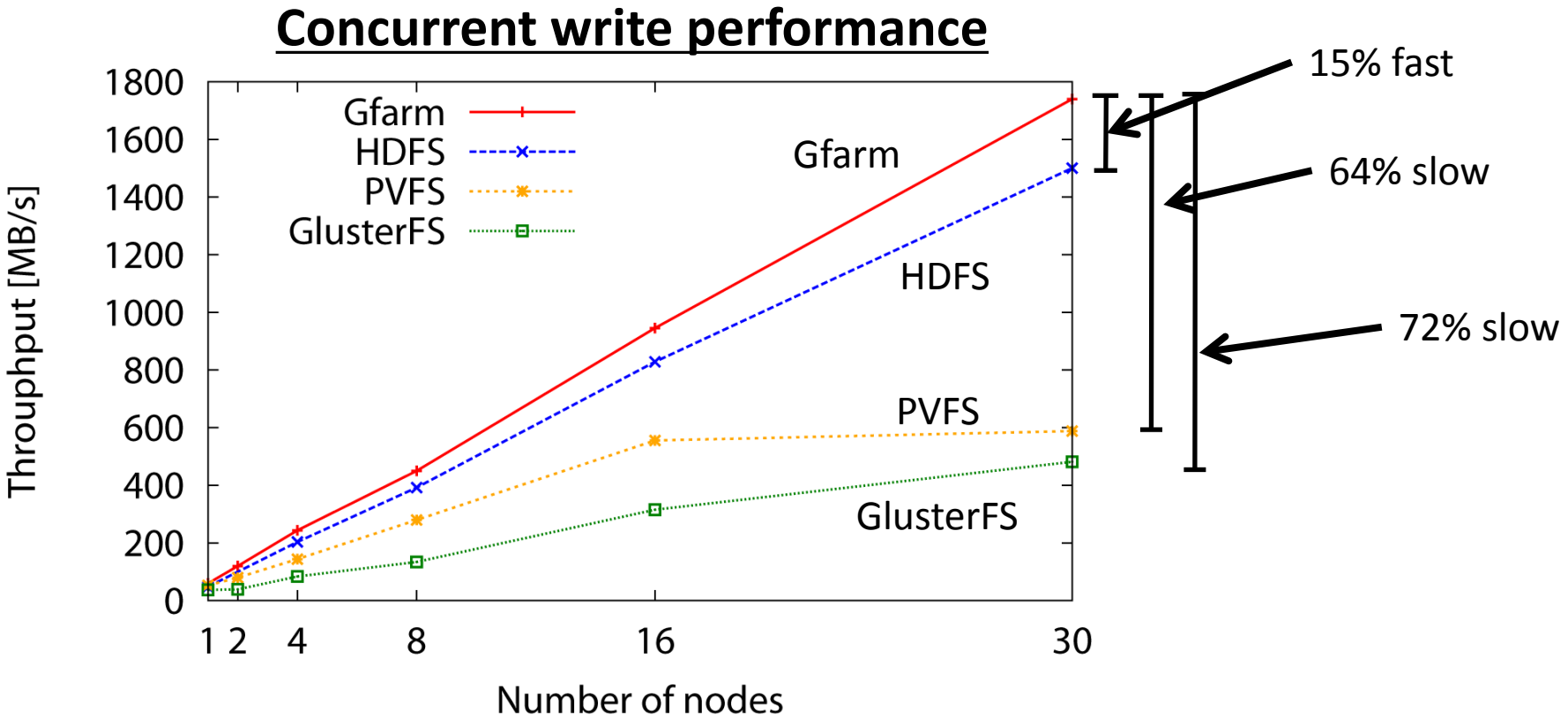
How to use HPCI shared storage

% **mount.hpci** **# mount command**
Update proxy certificate for gfarm2fs
timeleft : 167:50:40 (7.0 days)
Mount GfarmFS on /gfarm/hp120273/tatebe
% **df -H /gfarm/hp120273/tatebe**
Filesystem Size Used Avail Use% Mounted on
fuse 23P 2.9P 20P 14% /gfarm/hp120273/tatebe
% **cd /gfarm/hp120273/tatebe**
% **gfpcopy –P /work/CSI/tatebe/data . #parallel copy command**
….
copied_file_num: 10
copied_file_size: 6553600000
total_throughput: 70.233735 MB/s
total_time: 93.311284 sec.
% **gfncopy –s 2 data** **#specify # replicas**
（file repilcas are automatically created on background）

PARALLEL AND DISTRIBUTED DATA ANALYSIS

Hadoop Gfarm plugin [Mikami, Ohta, Tatebe, IEEE/ACM Grid 2011]

- Design and Implement Gfarm-Hadoop plugin to access POSIX compatible Gfarm file system from Hadoop apps
- Compare with HDFS, PVFS and GlusterFS



Pwrake workflow engine

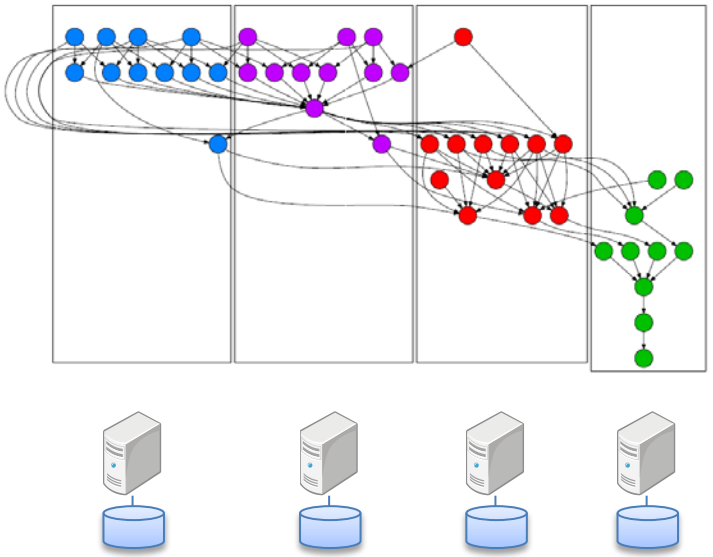
- Rake extension – parallel and distributed workflow language and execution engine
- http://github.com/masa16/Pwrake/
- Gfarm file system support
  - Automatic mount/umount of Gfarm file system
  - Data aware job scheduling

- Masahiro Tanaka, Osamu Tatebe, "**Pwrake: A parallel and distributed flexible workflow management tool for wide-area data intensive computing**", Proceedings of ACM International Symposium on High Performance Distributed Computing (HPDC), pp.356-359, 2010
- Masahiro Tanaka and Osamu Tatebe , "**Workflow Scheduling to Minimize Data Movement using Multi-constraint Graph Partitioning**", Proceedings of IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 2012 (to appear)

Data aware workflow scheduling [Tanaka, Tatebe, CCGrid 2012]

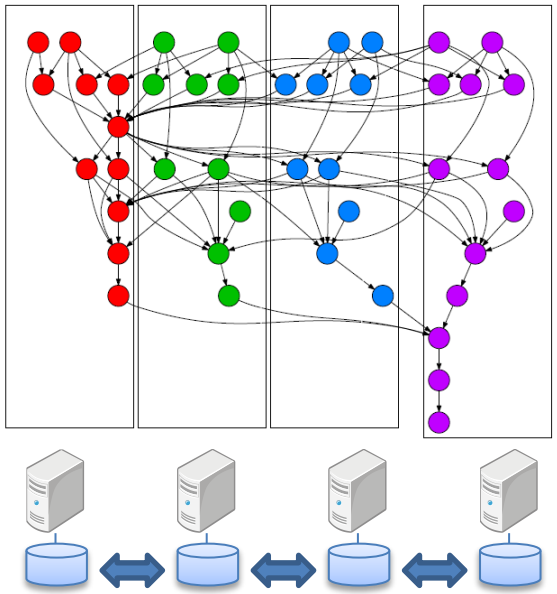
job scheduling by **multi-constraint graph partitioning** to minimize data transfer and maximize parallel job executions

Simple graph partitioning



Load imbalance happens in each parallel step

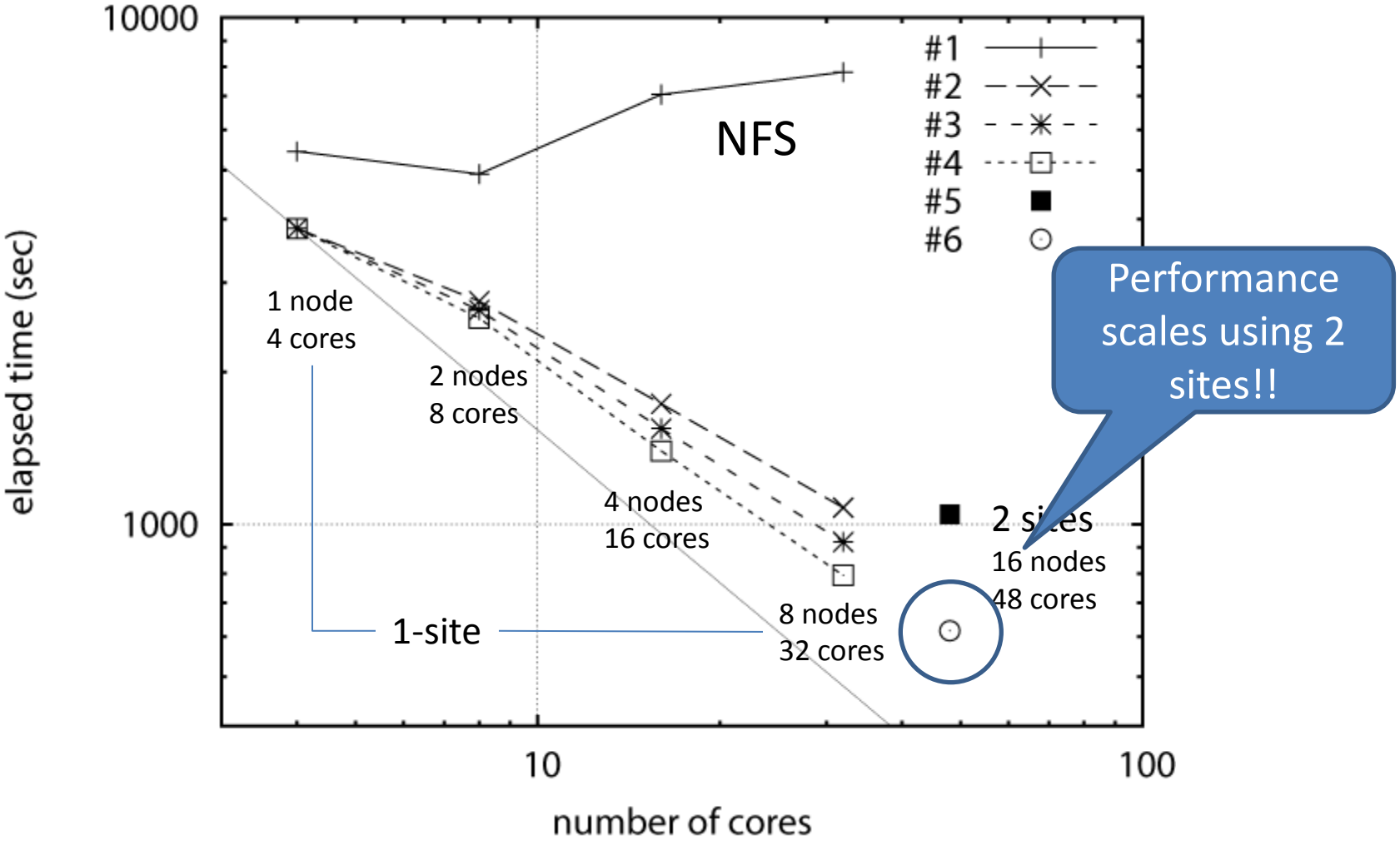
Multi-constraint graph partitioning



Reduce 14% of data transfer

Improve 31% of performance

Performance result of Montage astronomical data analysis workflow



SELECTED FEATURES

Consistency check and repair

- Consistency check and repair at MDS startup
- Consistency check and repair at file server startup in parallel
- # replicas is automatically maintained in case of file creation, file server failure, and changing # replicas
  - # replicas can be specified in each directory

    % **gfncopy** –s 3 /home/tatebe

Gfarm zabbix plugin

Realtime server monitoring and automatic ticket issue

Gfarm ganglia plugin

Realtime IOPS and bandwidth monitoring

NPO Tsukuba OSS Support Center

- http://oss-tsukuba.org/
- Established in Apr, 2013
- Gfarm software support
- Inaugural symposium on Aug 28, 2013

FUTURE EVOLUTION

Gfarm 2.6

- Will be released Q1, 2014
- Functionality to specify replica location to be created
- Transparent MDS failover
- Performance improvement of gfpcopy

Summary

- Gfarm file system
  - Developed since 2000, O(14,000) downloads
  - HPCI Shared Storage, NICT Science Cloud, Japan Lattice Data Grid (JLDG), companies
- > 1 GB/s parallel copy performance
- Hadoop MapReduce, Workflow, MPI-IO
- Management tools
- NPO OSS Tsukuba Support Center
- R&D for distributed MDS and object store