

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

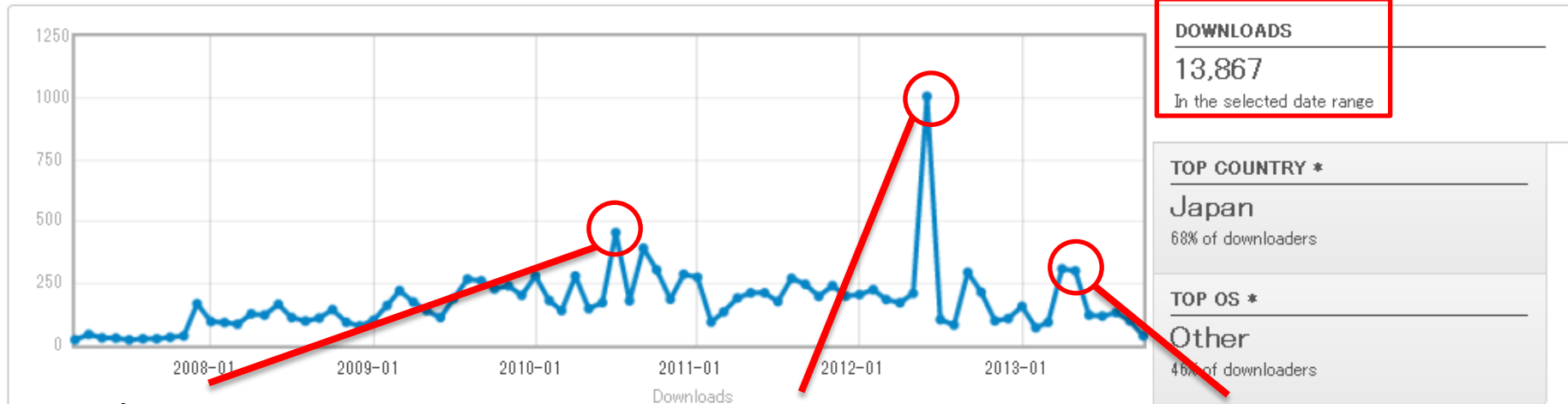
- 13,867 downloads since March, 2007

 **Gfarm File System**

Summary | Files | Reviews | Support | **Develop** | Hosted Apps | Mailing Lists | Forums | Code | Project Admin

[Home \(Change File\)](#)

Date Range: 2007-03-01 to 2013-10-16



2010/7

Version 2.3.2, 2.4.0

456 downloads

2012/6

HPCI installation etc

1,007 downloads

2013/4,5

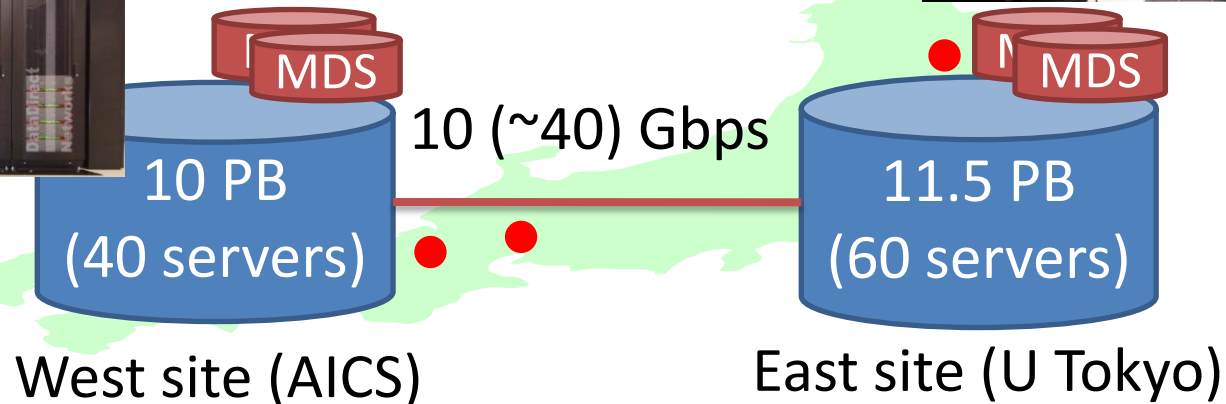
Version 2.5.8

610 downloads

HPCI SHARED STORAGE

HPCI Shared Storage

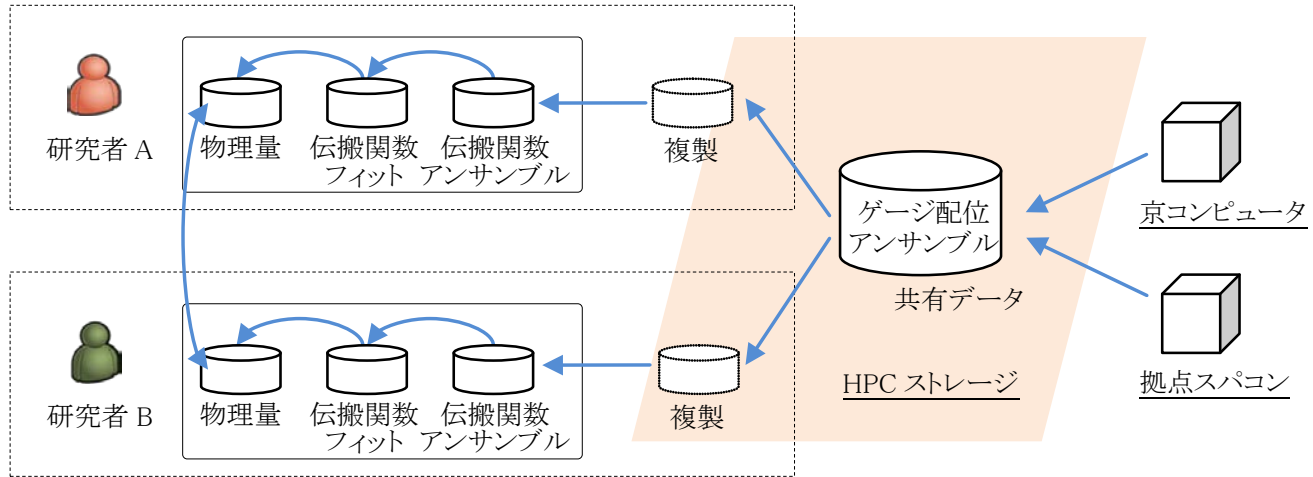
- HPCI – High Performance Computing Infrastructure
 - “K”, Hokkaido, Tohoku, Tsukuba, Tokyo, Titech, Nagoya, Kyoto, Osaka, Kyushu, RIKEN, JAMSTEC, AIST
- A 20PB single distributed file system consisting East and West sites
- Grid Security Infrastructure (GSI) for user ID
- Parallel file replication among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



Picture courtesy by Hiroshi Harada (U Tokyo)

Usage Scenario

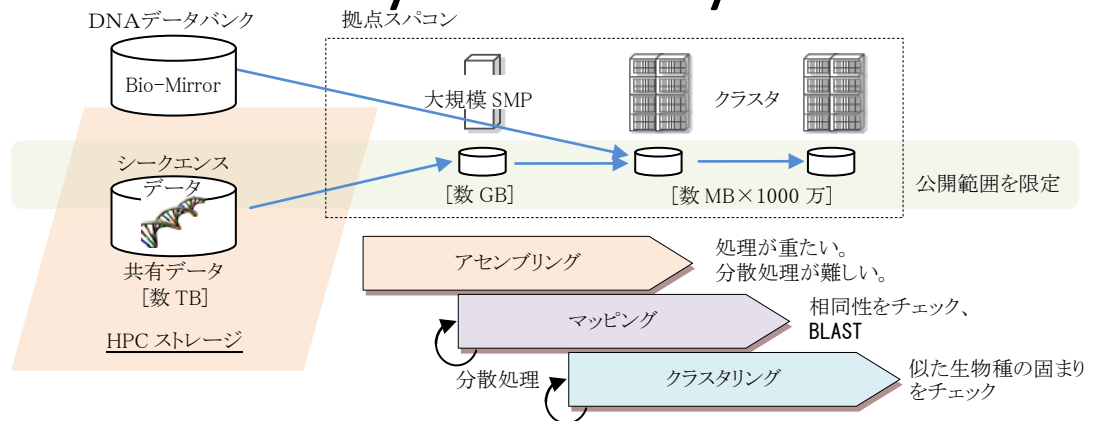
- QCD



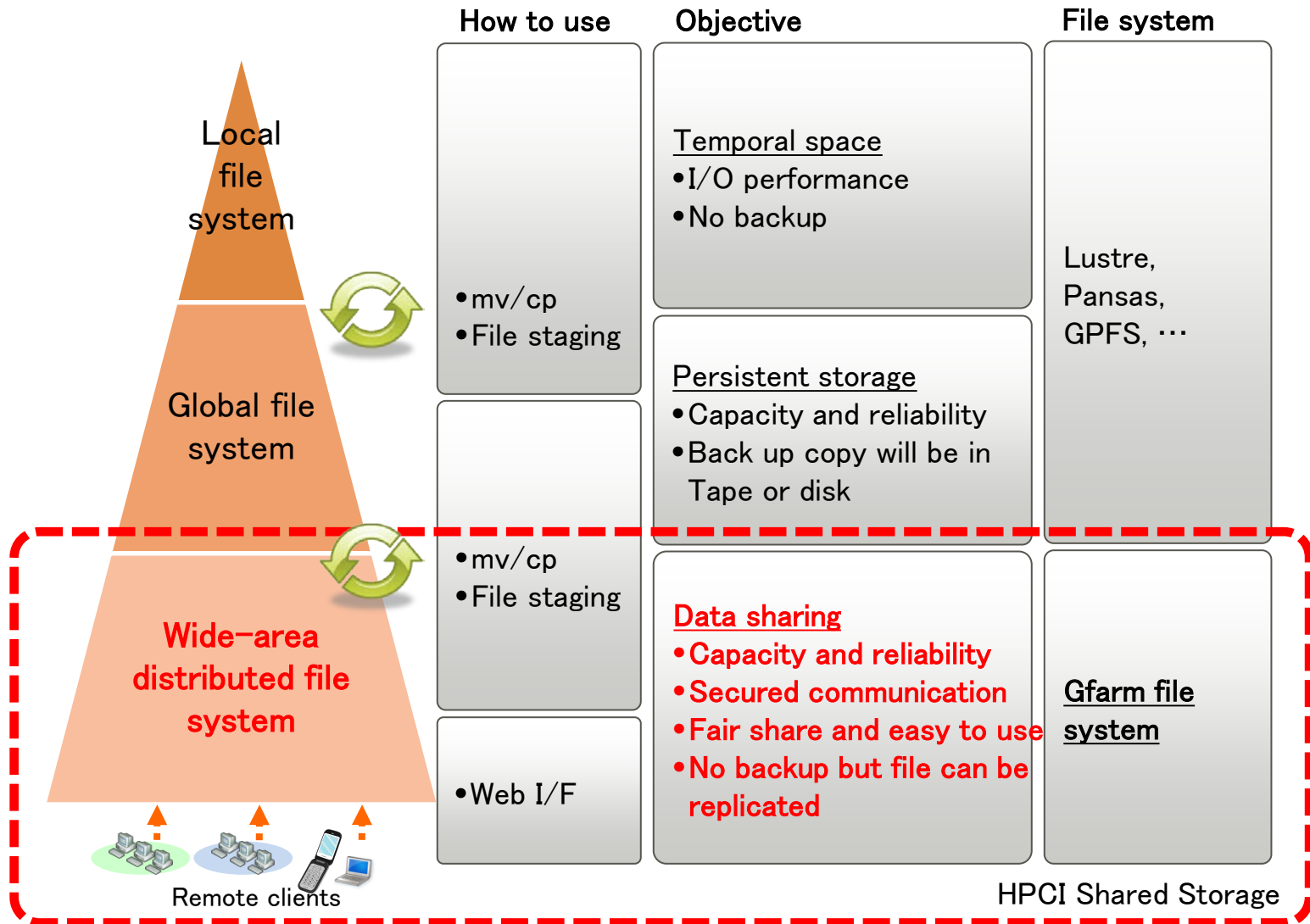
- Large scale simulation (ex. cosmic simulation)

- 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

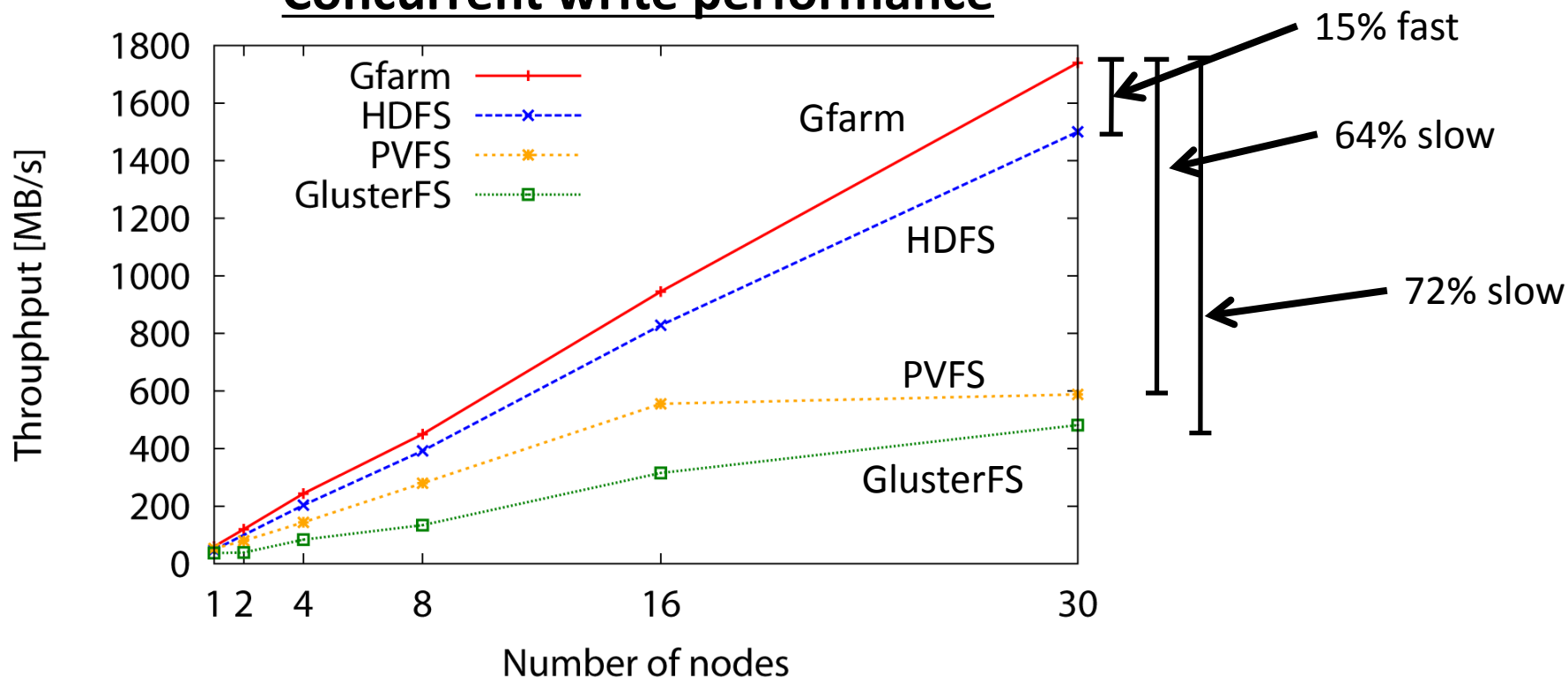
```
% mount.hpci # mount command
Update proxy certificate for gfarm2fs
timeleft : 167:50:40 (7.0 days)
Mount GfarmFS on /gfarm/hp120273/tatebe
% df -H /gfarm/hp120273/tatebe
Filesystem      Size  Used Avail Use% Mounted on
fuse            23P   2.9P  20P  14% /gfarm/hp120273/tatebe
% cd /gfarm/hp120273/tatebe
% gfcopy -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.
% gfcopy -s 2 data #specify # replicas
(file replicas 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

Concurrent write performance



Pwrake workflow engine

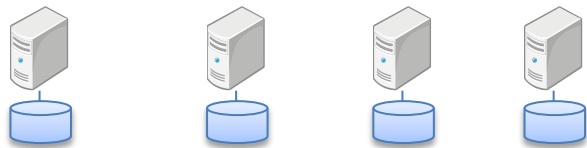
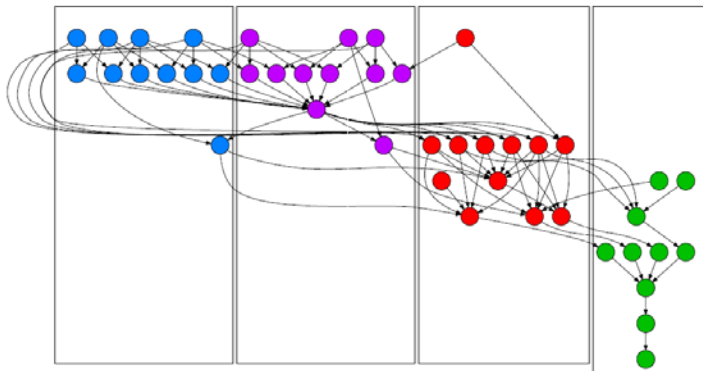
- Rake extension – parallel and distributed workflow language and execution engine
- <http://github.com/masa16/Pwrake/>
- Gfarm file system support
 - Automatic mount/umount of Gfarm file system
 - Data aware job scheduling
- Masahiro Tanaka, Osamu Tatebe, "**Pwrake: A parallel and distributed flexible 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]

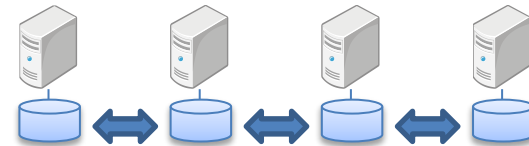
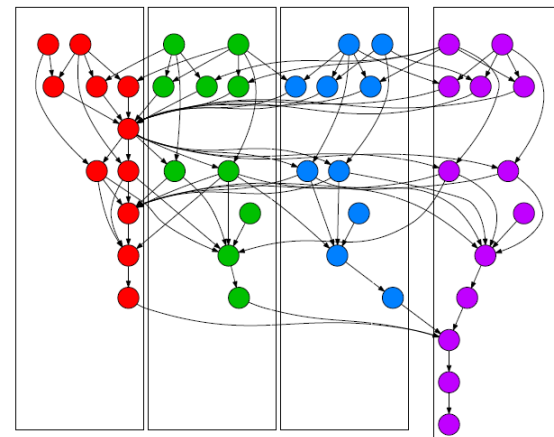
job scheduling by **multi-constraint graph partitioning** to **minimize data transfer** and **maximize parallel job executions**

Simple graph partitioning



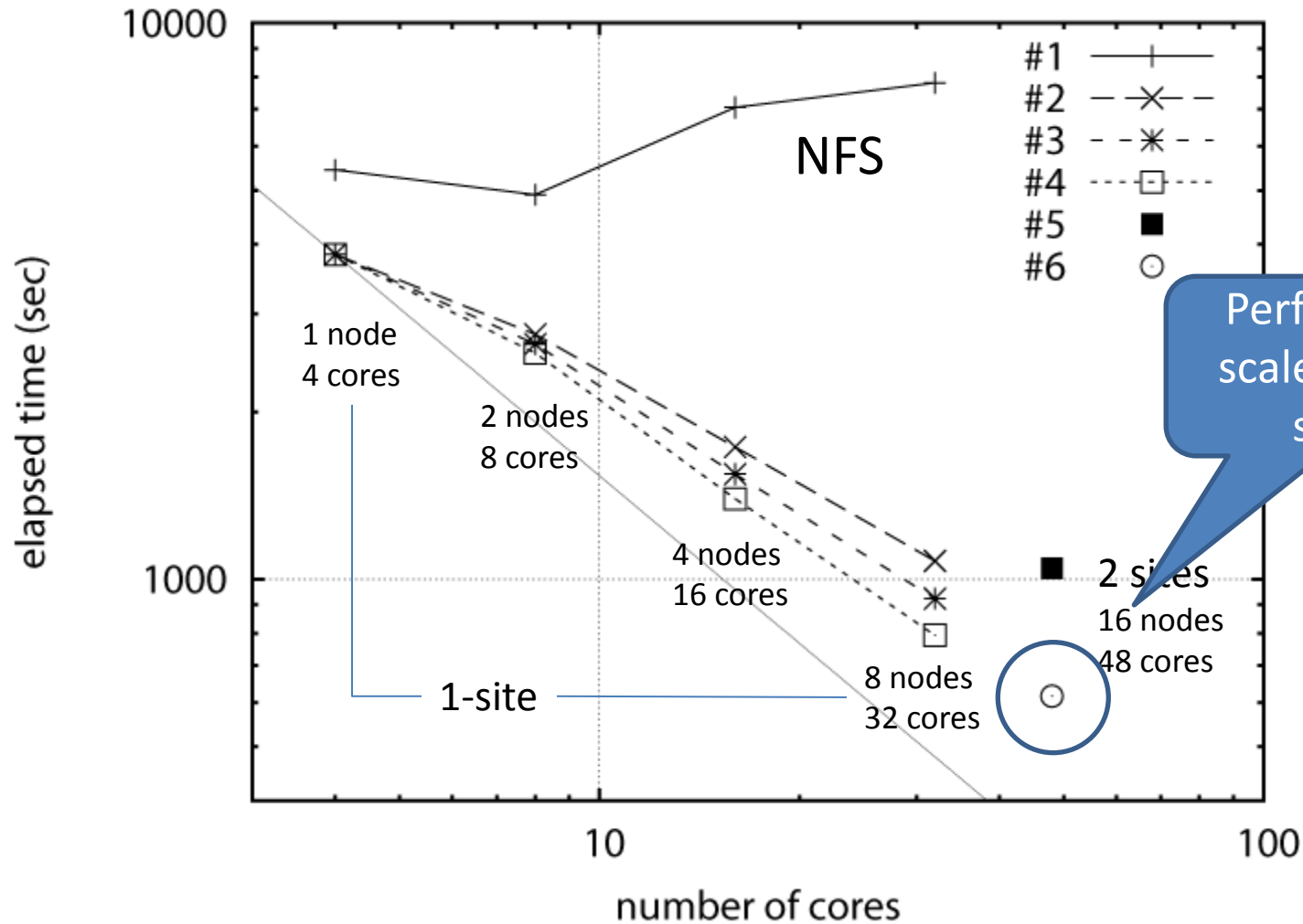
Load imbalance happens in each parallel step

Multi-constraint graph partitioning



Reduce **14%** of data transfer
Improve **31%** of performance

Performance result of Montage astronomical data analysis workflow



Performance scales using 2 sites!!

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
 - % **gfncoy** -s 3 /home/tatebe

Gfarm zabbix plugin

The screenshot displays the Zabbix web interface in a Firefox browser window. The page title is "ZabbixServer01: ダッシュボード". The main content area shows the "Zabbixサーバの状態" (Zabbix Server Status) section, which includes a table of parameters and their values. Below this, there are sections for "システムステータス" (System Status), "ホストステータス" (Host Status), "最新20件の障害" (Latest 20 Incidents), and "ウェブ監視" (Web Monitoring).

パラメータ	値	詳細
Zabbixサーバの起動	はい	esci-wgfarmc2.aics.riken.jp:10051
ホスト数 (有効/無効/テンプレート)	54	5 / 2 / 47
アイテム数 (有効/無効/取得不可)	146	129 / 17 / 0
トリガー数 (有効/無効)[障害/不明/正常]	59	54 / 5 [0 / 0 / 54]
ユーザ数 (オンライン)	2	1
1秒あたりの監視項目数 (Zabbixサーバの要求パフォーマンス)	0.33	-

更新: 15:33:29

ホストグループ	致命的な障害	重度の障害	軽度の障害	警告	情報	未分類
Gfarm v2 FileSystem	0	0	0	0	0	0

更新: 15:33:29

ホストグループ	障害なし	障害あり	合計
Gfarm v2 FileSystem	5	0	5

更新: 15:33:29

ホスト	問題	最新の変更	経過時間	コメントあり	アクション
...					

更新: 15:33:29

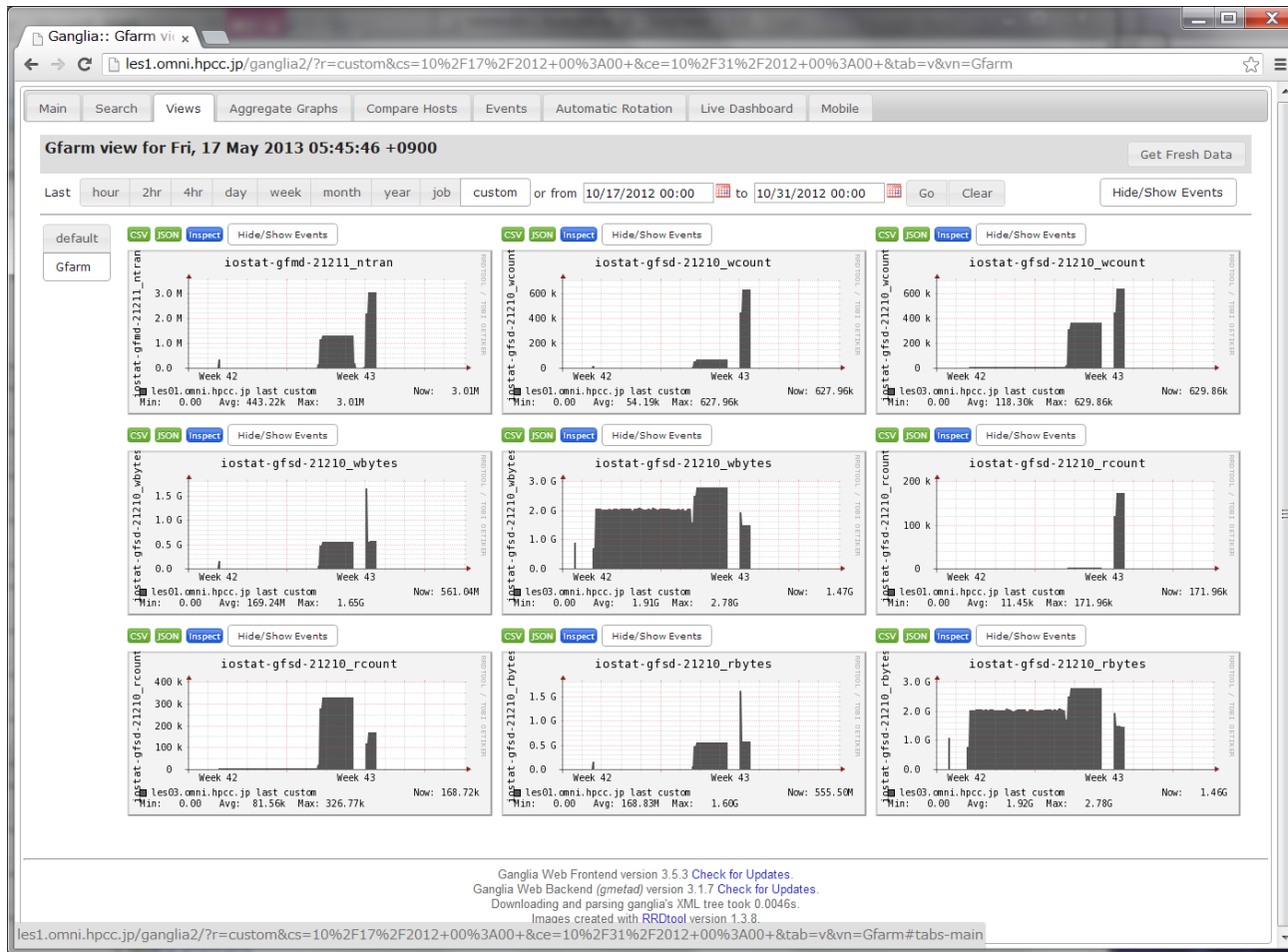
ホストグループ	正常	失敗	進捗中	不明
...				

更新: 15:33:29

Zabbix 1.8.4 Copyright 2001-2010 by SIA Zabbix | 次のユーザーでログイン中 'Admin'

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