CCS External Review
February 19, 2014

Gfarm distributed file system

Osamu Tatebe
University of Tsukuba

Gfarm file system

STORCLOUD

SC05 STORCLOUD CHALLENGE AWARD

Presented to

National Institute of Advanced Industrial Science and Technology

for

Most Innovative Use of Storage In Support of Science

SC06

POWERFUL BEYOND IMAGINATION

TAMPA CONVENTION CENTER

TAMPA, FL | NOVEMBER 11-17, 2006

Storage Challenge

Winner-Large Systems

"High Performance Data Analysis for Particle Physics using the Gfarm File System"

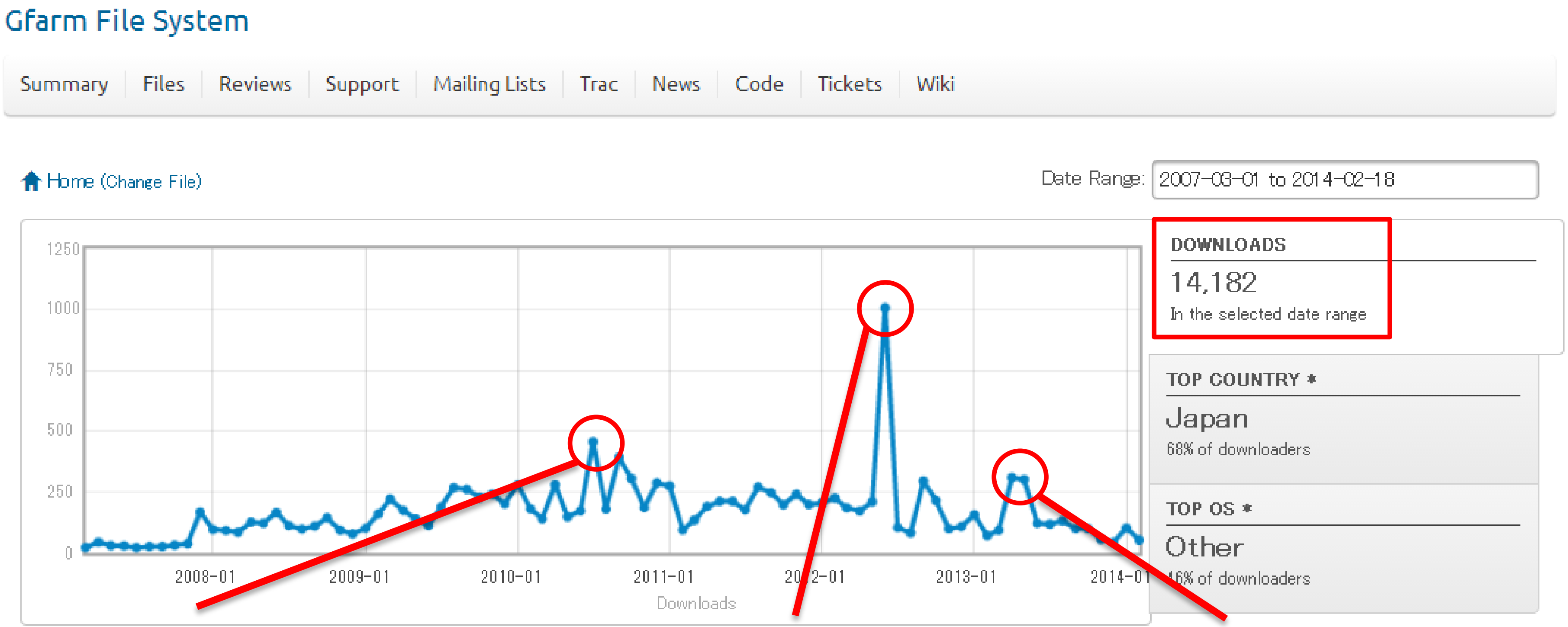
Osamu Tatebe, *University of Tsukuba, Japan*; Nobuhiko Katayama, *KEK, Japan*; Mitsuhisa Sato, Taisuke Boku, and Akira Ukawa, *University of Tsukuba, Japan*; and Shohei Nishida and Ichiro Adachi, *KEK, Japan*;

- 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

- 14,182 downloads since March, 2007



Gfarm File System

Summary | Files | Reviews | Support | Mailing Lists | Trac | News | Code | Tickets | Wiki

Home (Change File)

Date Range: 2007-03-01 to 2014-02-18

DOWNLOADS

14,182

In the selected date range

TOP COUNTRY *

Japan

68% of downloaders

TOP OS *

Other

16% of downloaders

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

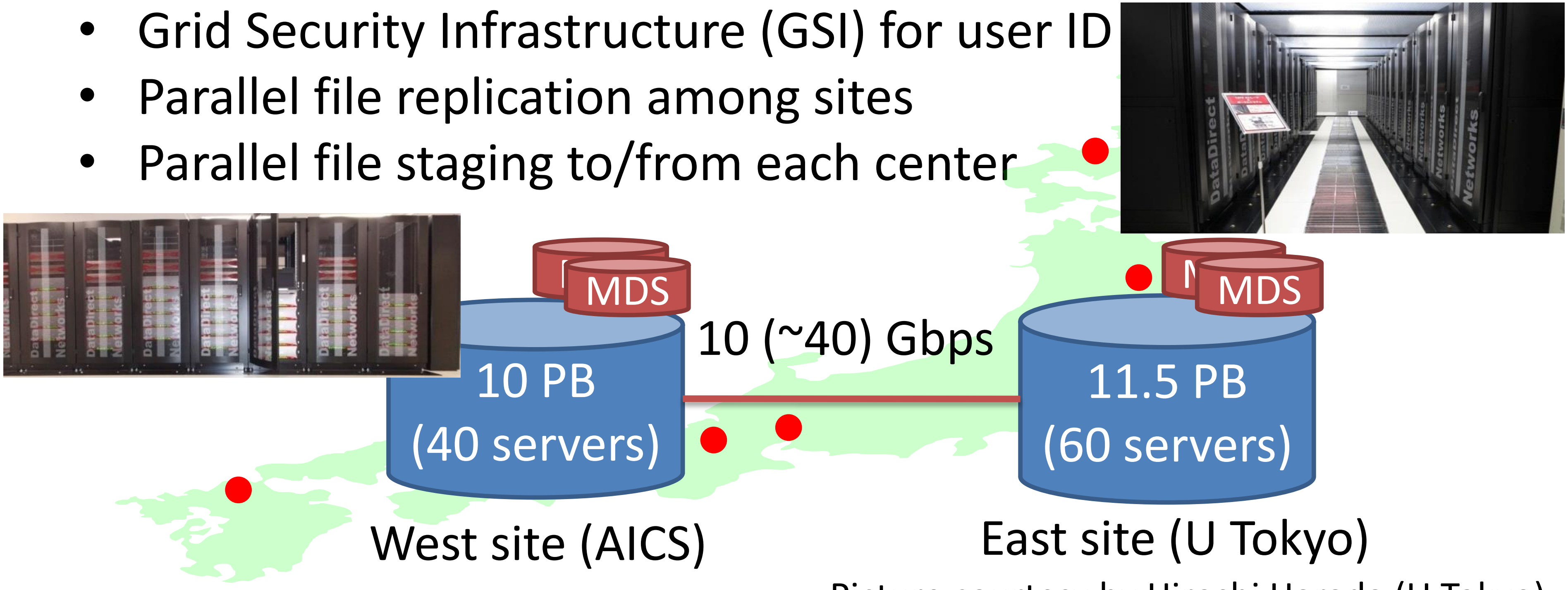
Collaboration

- Japan Lattice Data Grid (JLDG)
  - Division of Particle Physics, Division of Computational Informatics
  - KEK, Osaka Univ, Hiroshima Univ, Nagoya Univ, Kyoto Univ, Kanazawa Univ, RIKEN, Univ Tokyo
- GPV/JMA Archive
  - Division of Global Environmental Science, Division of Computational Informatics
- HPCI Shared Storage
  - RIKEN, Nine National Universities, AIST, JAMSTEC, ...
- NICT Science Cloud
  - NICT

HPCI SHARED STORAGE

HPCI Shared Storage

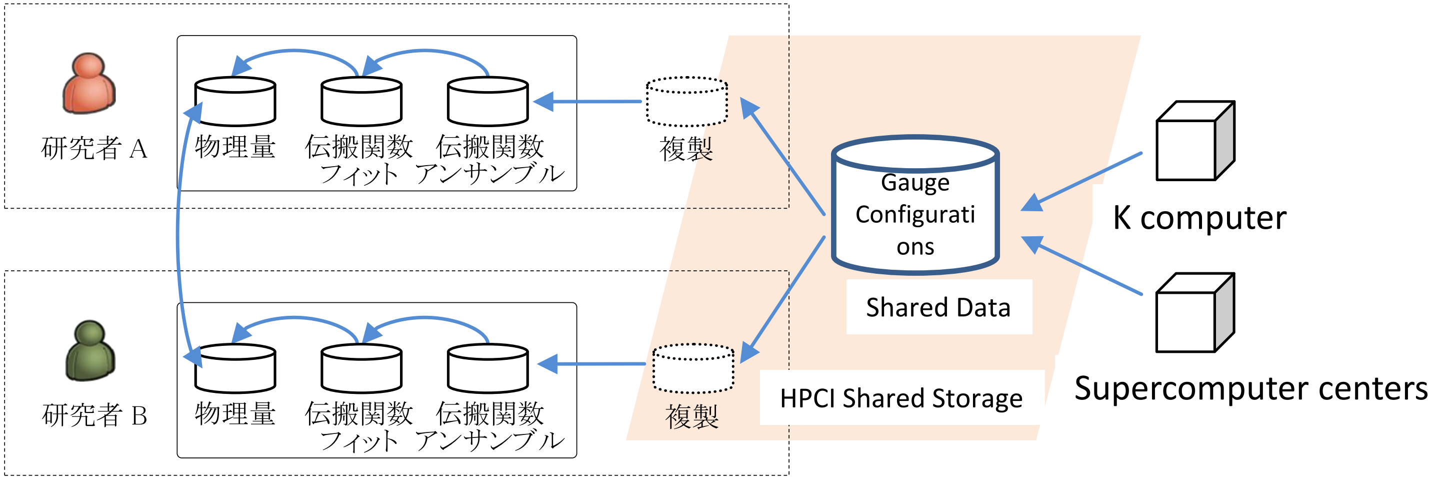
- HPCI – High Performance Computing Infrastructure
  - “K”, Hokkaido, Tohoku, Tsukuba, Tokyo, Titech, Nagoya, Kyoto, Osaka, Kyushu, RIKEN, JAMSTEC, AIST
- A 20PB single distributed file system consisting East and West sites
- Grid Security Infrastructure (GSI) for user ID
- Parallel file replication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



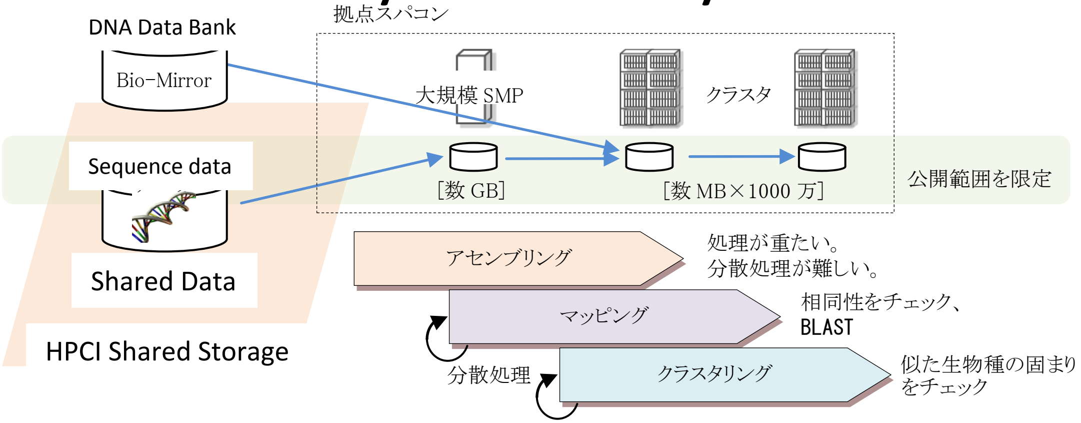
Picture courtesy by Hiroshi Harada (U Tokyo)

Usage Scenario

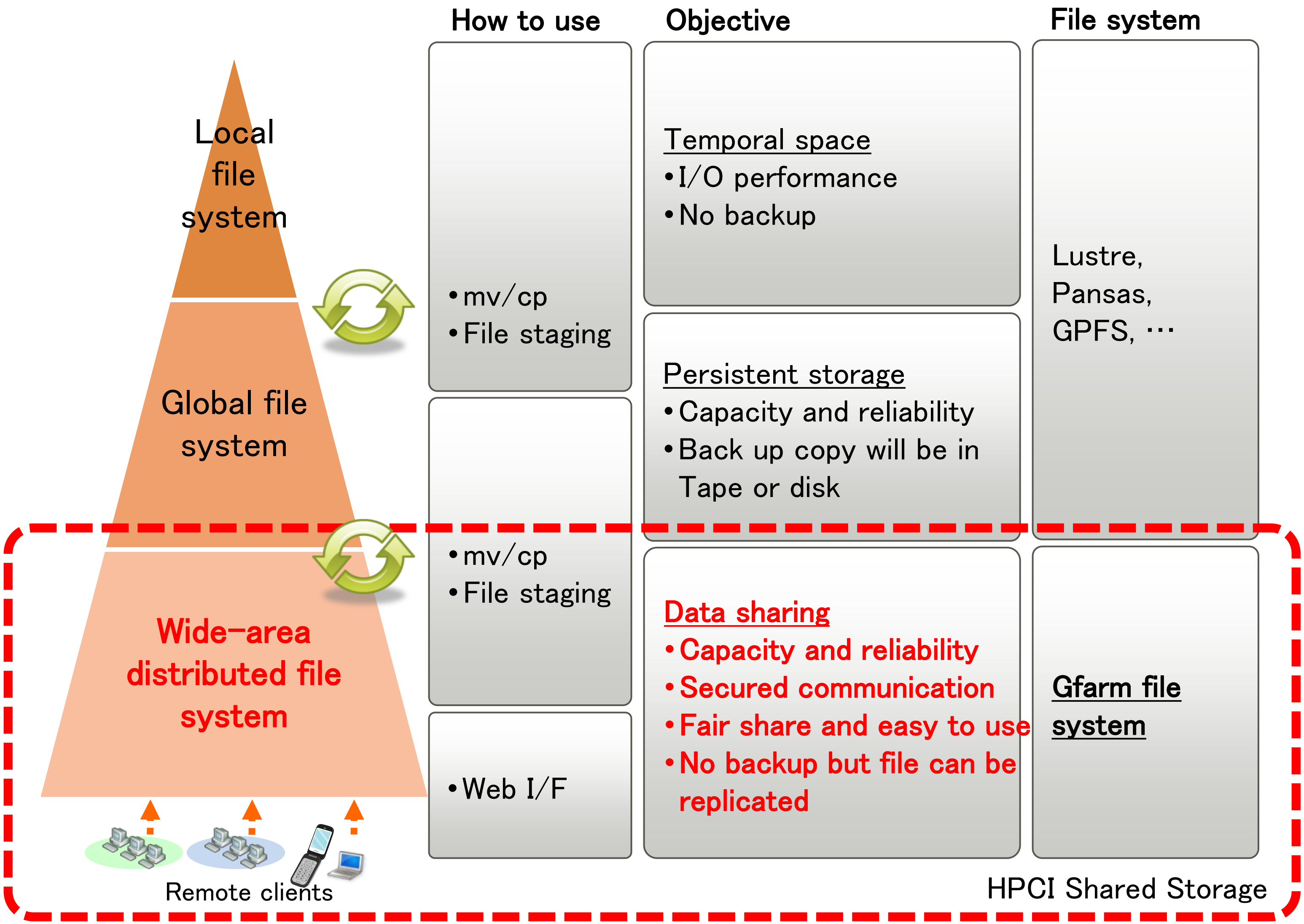
- QCD



- 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

- mv/cp
- File staging

- mv/cp
- File staging

- Web I/F

Objective

Temporal space

- I/O performance
- No backup

Persistent storage

- Capacity and reliability
- Back up copy will be in Tape or disk

Data sharing

- Capacity and reliability
- Secured communication
- Fair share and easy to use
- No backup but file can be replicated

File system

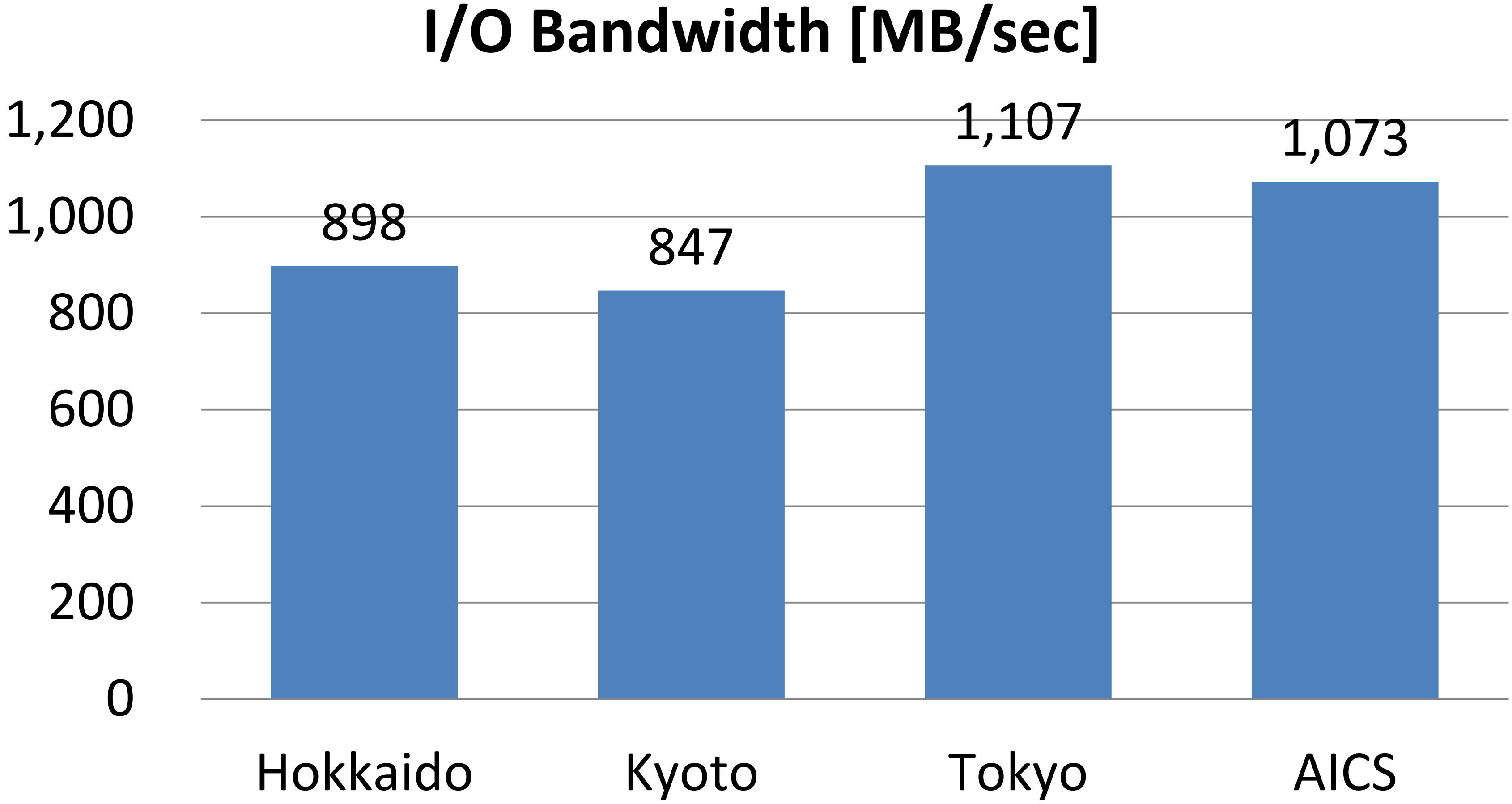
Lustre, Pansas, GPFS, …

Gfarm file system

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)
```

Initial Performance Result



File copy performance of 300 1GB files

Related Work

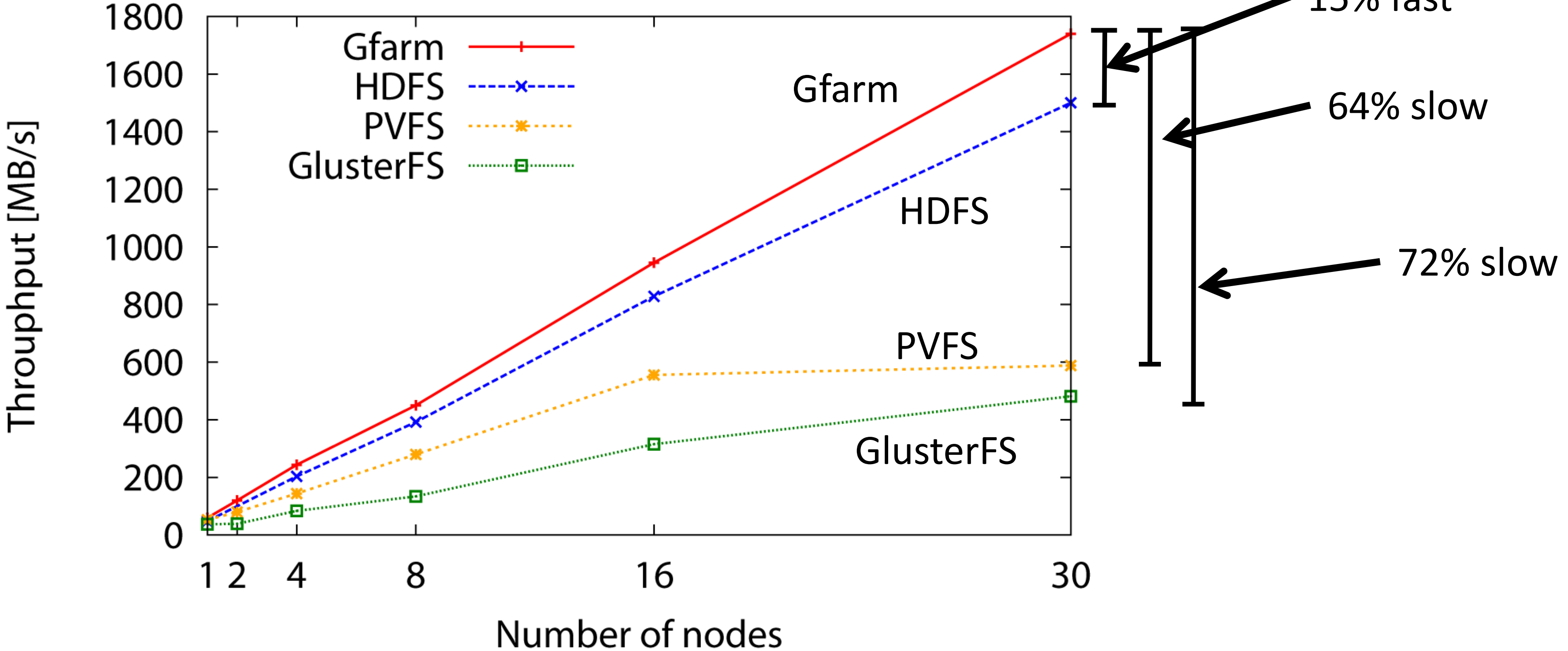
- XSEDE-Wide File System (GPFS)
  - Planned, but not in operation yet
- DEISA Global File System
  - Multicluster GPFS
    - RZG, LRZ, BSC, JSC, EPSS, HLRS, …
    - Site name included in the path name – no distribution transparency
      - files cannot be replicated across sites
  - PRACE does not provide global file system
    - Limitation of operation systems that can mount
    - PRACE does not assume to use multiple sites
- Distant access to Lustre File System
  - Many researches in TeraGrid
    - Increase the number of pending requests
  - SC11 paper showed the performance could not be simply improved in private 100Gbps network

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



15% fast

64% slow

72% slow

Pwrake workflow engine

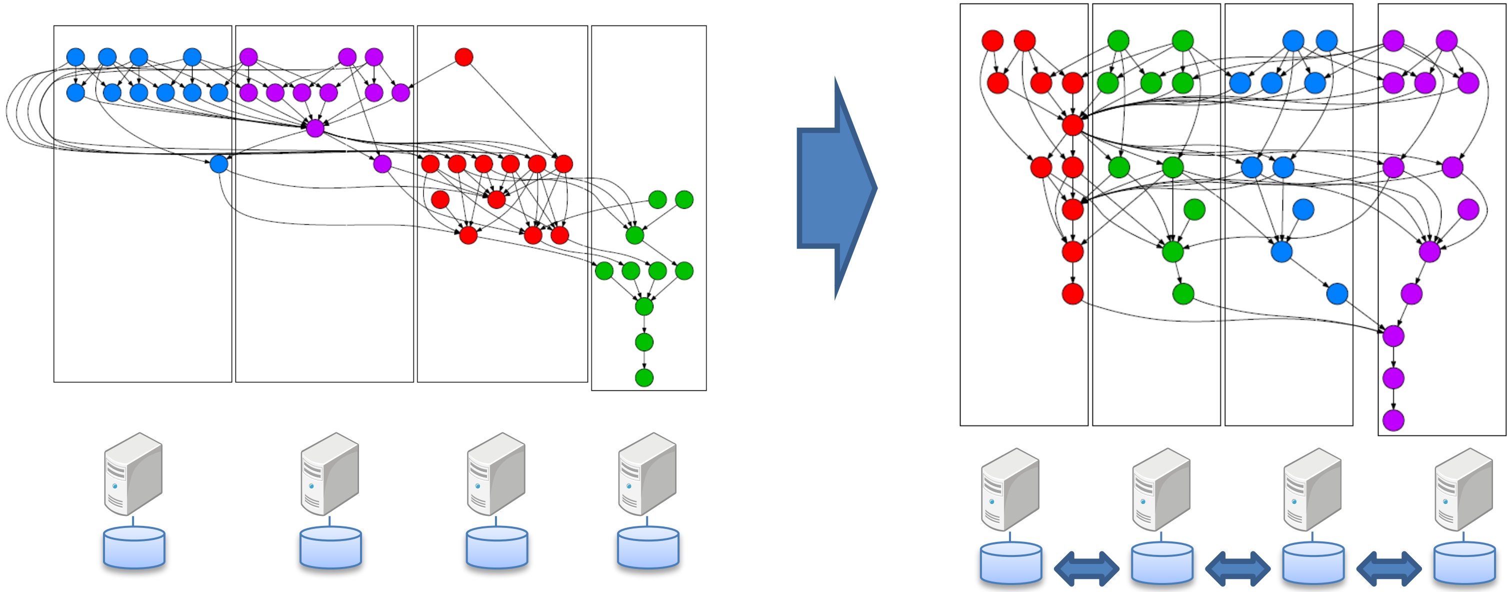
- Rake extension – parallel and distributed workflow language and execution engine
- http://github.com/masa16/Pwrake/
- Gfarm file system support
  - Automatic mount/umount of Gfarm file system
  - Data aware job scheduling
- Masahiro Tanaka, Osamu Tatebe, "**Pwrake: A parallel and distributed flexible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

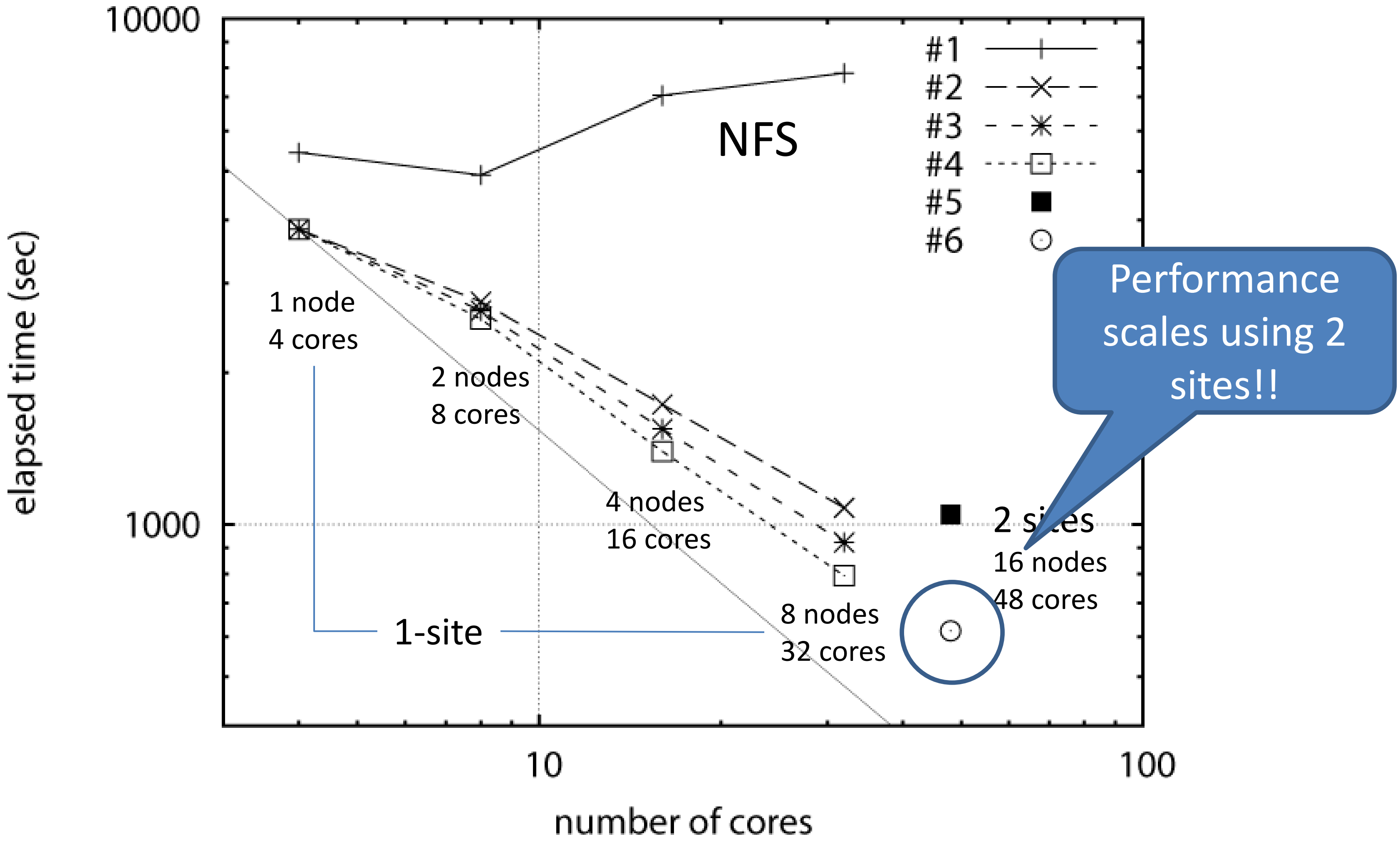
Multi-constraint graph partitioning



Load imbalance happens in each parallel step

Reduce 14% of data transfer
Improve 31% of performance

Performance result of Montage astronomical data analysis workflow



Performance scales using 2 sites!!

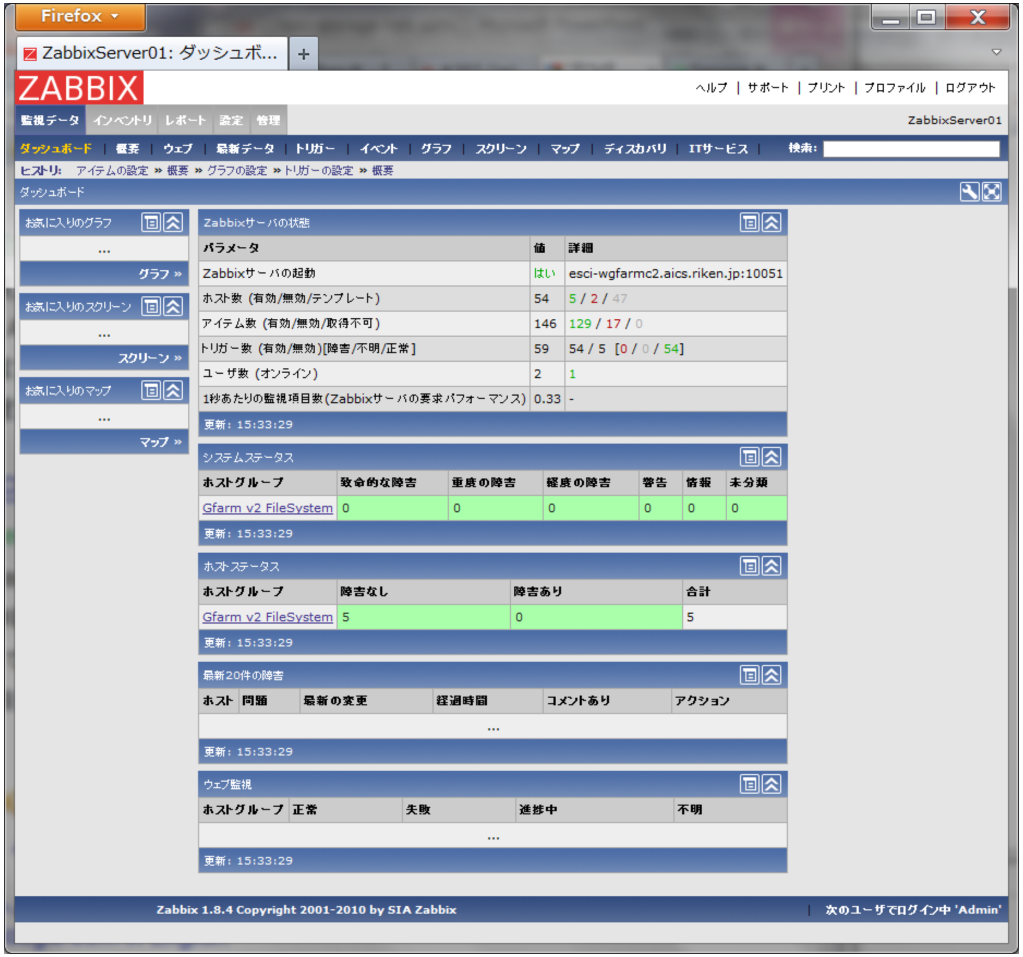
SELECTED FEATURES

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 server failure, and changing # replicas
  - # replicas can be specified in each directory

```
% gfncopy –s 3 /home/tatebe
```

Gfarm zabbix plugin


Firefox
ZabbixServer01: ダッシュボ...
ZABBIX
ヘルプ | サポート | プリント | プロファイル | ログアウト
監視データ
インベントリ
レポート
設定
管理
ZabbixServer01
ダッシュボード | 概要 | ウェブ | 最新データ | トリガー | イベント | グラフ | スクリーン | マップ | ディスカバリ | ITサービス |
検索:
ヒストリ: アイテムの設定 » 概要 » グラフの設定 » トリガーの設定 » 概要
ダッシュボード
お気に入りのグラフ
...
グラフ »
お気に入りのスクリーン
...
スクリーン »
お気に入りのマップ
...
マップ »

Zabbixサーバの状態

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

システムステータス

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

ホストステータス

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

最新20件の障害

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

ウェブ監視

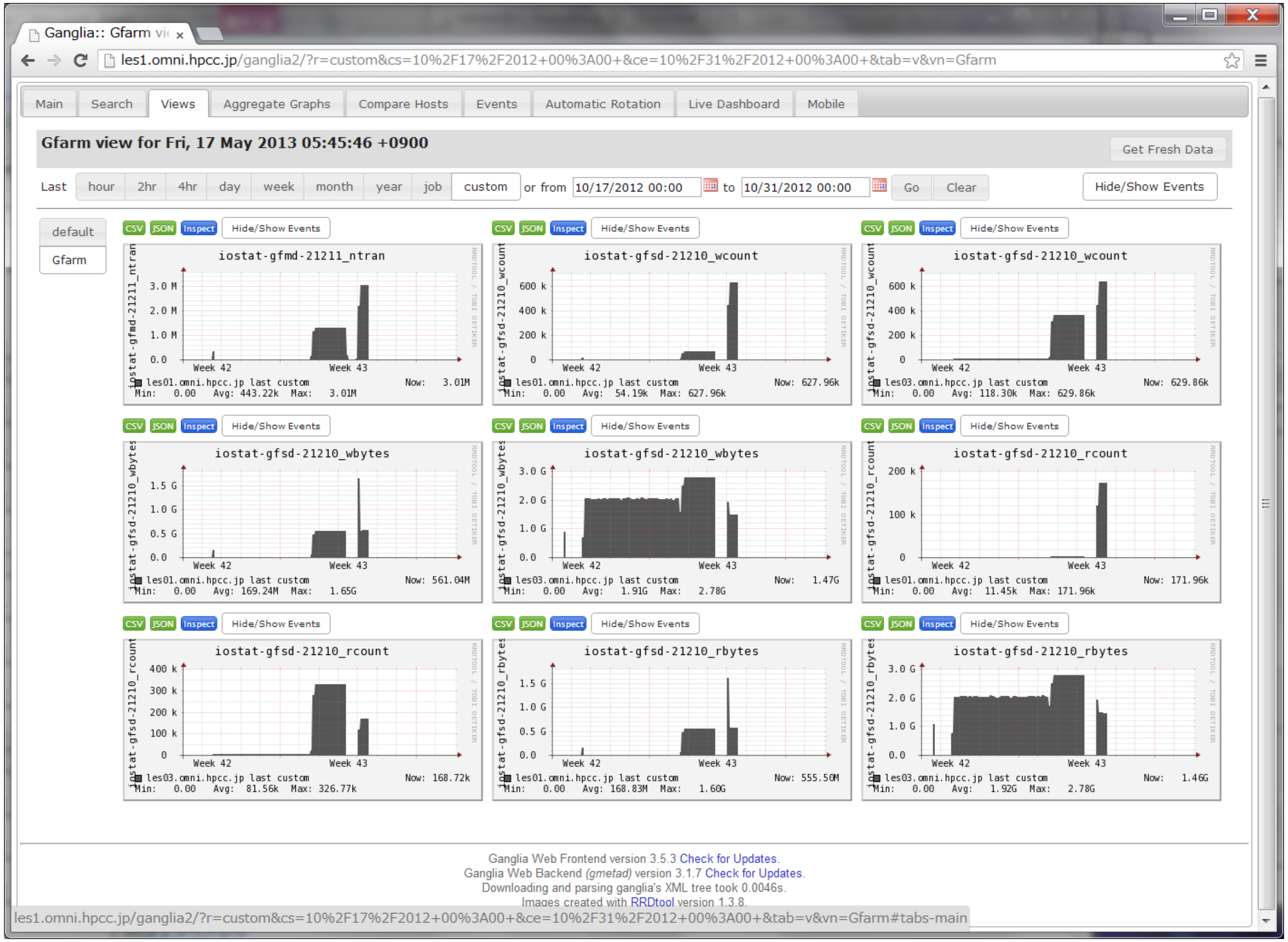
ホストグループ	正常	失敗	進捗中	不明
...				

更新: 15:33:29

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

Realtime server monitoring and automatic ticket issue

Gfarm ganglia plugin


Ganglia:: Gfarm vi
Main
Search
Views
Aggregate Graphs
Compare Hosts
Events
Automatic Rotation
Live Dashboard
Mobile
Get Fresh Data
Last
hour
2hr
4hr
day
week
month
year
job
custom
Go
Clear
Hide/Show Events
default
Gfarm
CSV
JSON
Inspect
Hide/Show Events
iostat-gfmd-21211_ntran
3.0 M
2.0 M
1.0 M
0.0
Week 42
Week 43
les01.omni.hpcc.jp last custom Now: 3.01M
Min: 0.00 Avg: 443.22k Max: 3.01M
iostat-gfsd-21210_wcount
600 k
400 k
200 k
0
les01.omni.hpcc.jp last custom Now: 627.96k
Min: 0.00 Avg: 54.19k Max: 627.96k
iostat-gfsd-21210_wcount
les03.omni.hpcc.jp last custom Now: 629.86k
Min: 0.00 Avg: 118.30k Max: 629.86k
iostat-gfsd-21210_wbytes
1.5 G
1.0 G
0.5 G
0.0
les01.omni.hpcc.jp last custom Now: 561.04M
Min: 0.00 Avg: 169.24M Max: 1.65G
iostat-gfsd-21210_wbytes
3.0 G
2.0 G
1.0 G
0.0
les03.omni.hpcc.jp last custom Now: 1.47G
Min: 0.00 Avg: 1.91G Max: 2.78G
iostat-gfsd-21210_rcount
200 k
100 k
0
les01.omni.hpcc.jp last custom Now: 171.96k
Min: 0.00 Avg: 11.45k Max: 171.96k
iostat-gfsd-21210_rcount
400 k
300 k
200 k
100 k
0
les03.omni.hpcc.jp last custom Now: 168.72k
Min: 0.00 Avg: 81.56k Max: 326.77k
iostat-gfsd-21210_rbytes
1.5 G
1.0 G
0.5 G
0.0
les01.omni.hpcc.jp last custom Now: 555.50M
Min: 0.00 Avg: 168.83M Max: 1.60G
iostat-gfsd-21210_rbytes
3.0 G
2.0 G
1.0 G
0.0
les03.omni.hpcc.jp last custom Now: 1.46G
Min: 0.00 Avg: 1.92G Max: 2.78G
RRDTOOL / TOBI OETIKER

les1.omni.hpcc.jp/ganglia2/?r=custom&cs=10%2F17%2F2012+00%3A00+&ce=10%2F31%2F2012+00%3A00+&tab=v&vn=Gfarm

Gfarm view for Fri, 17 May 2013 05:45:46 +0900

or from 10/17/2012 00:00 to 10/31/2012 00:00

Ganglia Web Frontend version 3.5.3 Check for Updates.
Ganglia Web Backend *(gmetad)* version 3.1.7 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0046s.
Images created with RRDtool version 1.3.8.

les1.omni.hpcc.jp/ganglia2/?r=custom&cs=10%2F17%2F2012+00%3A00+&ce=10%2F31%2F2012+00%3A00+&tab=v&vn=Gfarm#tabs-main

Realtime IOPS and bandwidth monitoring

NPO Tsukuba OSS Support Center

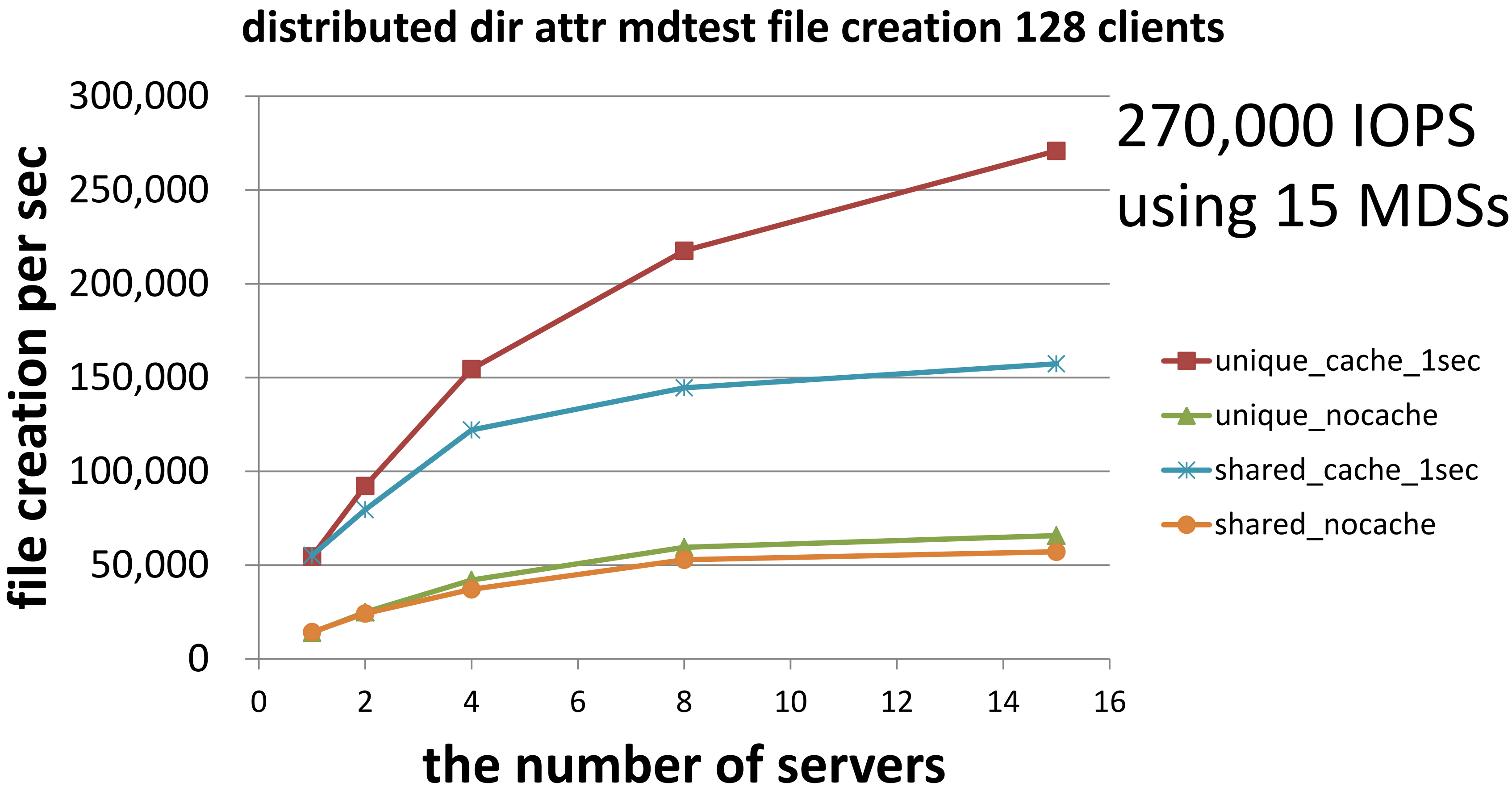
- http://oss-tsukuba.org/
- Established in Apr, 2013
- Gfarm software support
- Inaugural symposium on Aug 28, 2013

FUTURE EVOLUTION

Gfarm 2.6

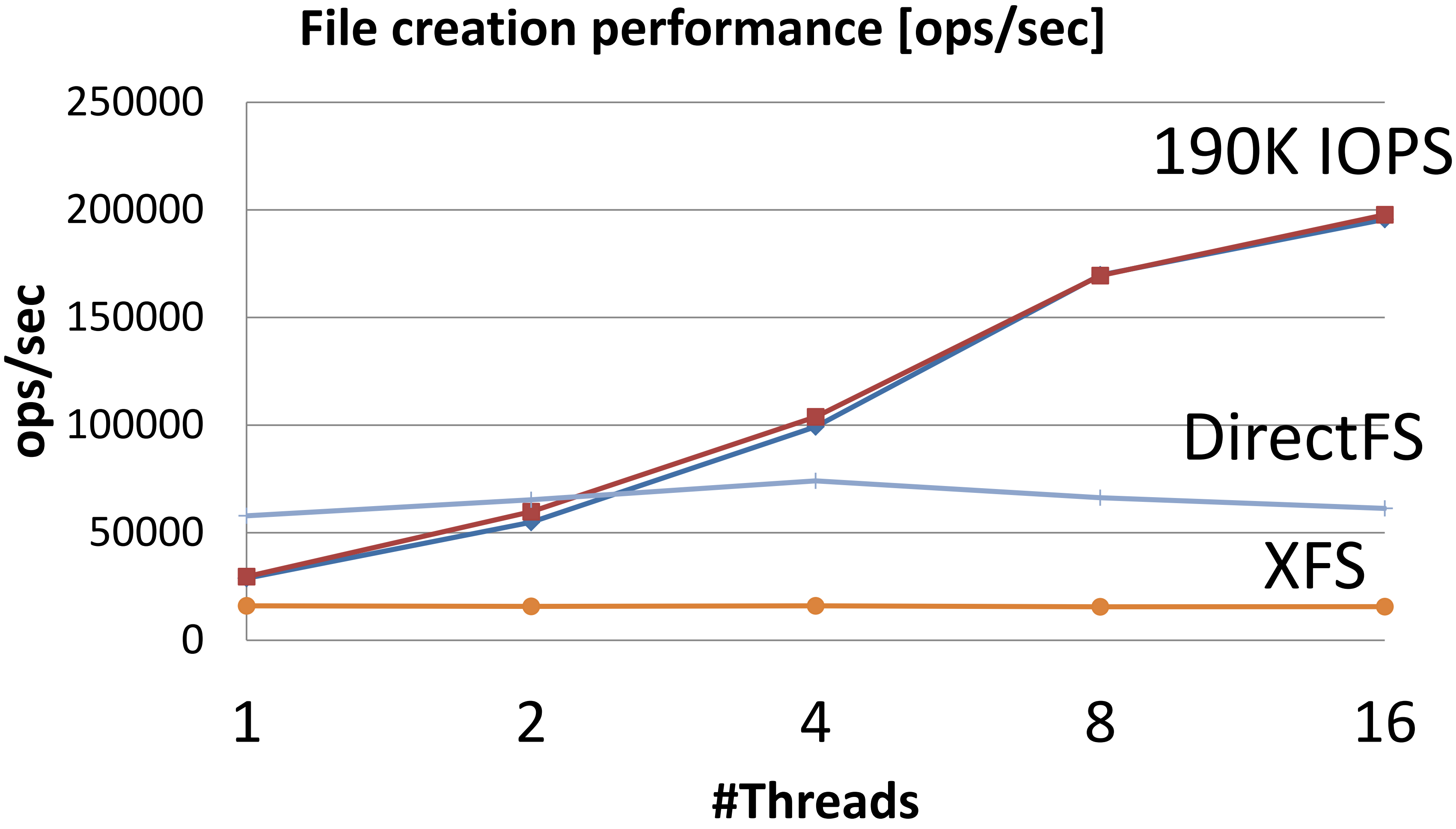
- Will be released Q1, 2014
- Functionality to specify replica location to be created
- Transparent MDS failover
- Performance improvement of gfpcopy

R&D of distributed MDS [JST/CREST]



270,000 IOPS using 15 MDSs

R&D of object storage for ioDrive [JST/CREST]



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