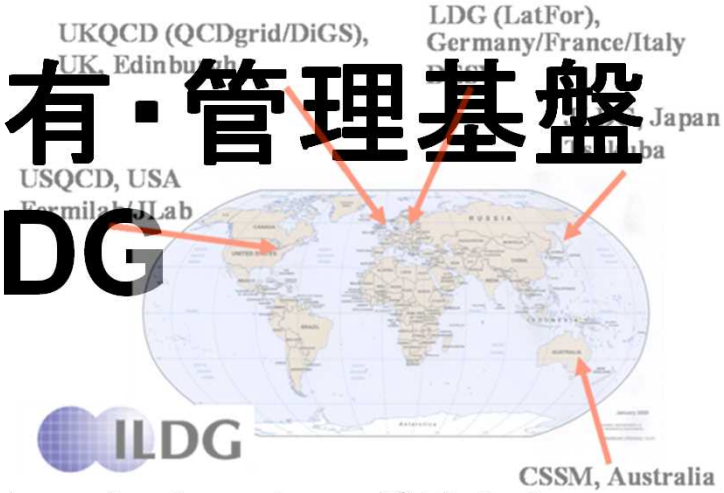
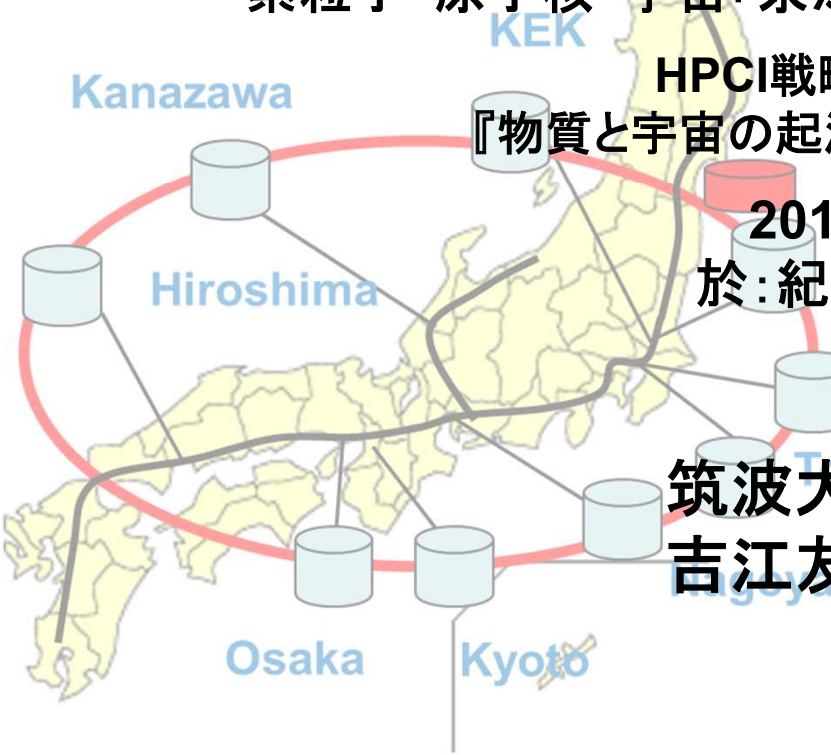
格子QCDデータ共有・管理基盤
JLDG/ILDG



素粒子・原子核・宇宙「京からポスト京に向けて」シンポジウム

HPCI戦略プログラム分野5
『物質と宇宙の起源と構造』全体シンポジウム

2015年3月11日
於：紀尾井フォーラム



筑波大学計算科学研究センター
吉江友照

計算素粒子物理学(Lattice QCD)と関連分野の為のデータグリッド

JLDG: Japan Lattice Data Grid
ILDG: International Lattice Data Grid

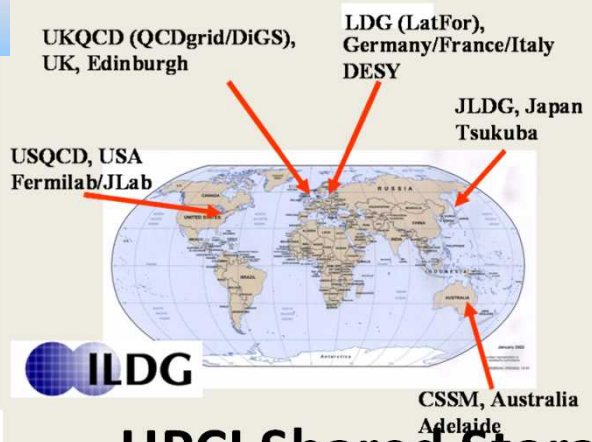
1. **JLDGと関連システム: Overview**
2. **JLDG: システムと利用状況**
3. **HPCI 共用ストレージ･JLDG 連携システム**
4. **JLDG の今後**

JLDGと関連システム: Overview

Provide flat file system for 9 sites in Japan

Backbone: SINET4 1Gbps L3-VPN (NII with KEK support)

budget by HPCI Strategic Program Filed 5 since 2011



Japan Lattice Data Grid

ILDG

part of data can be accessed from ILDG

HPCI Shared Storage

11.5PB Tokyo(Kashiwa)

10.0PB Riken-AICS

fast parallel file copy

gfarm

- Kanazawa: 100TB down
- KEK: 1130TB
- CCS
- Tsukuba: 2800TB
- Riken (Wako): 120TB new
- Tokyo (Kashiwa): 470TB
- Hiroshima: 230TB
- Nagoya: 300TB
- Osaka: 300TB
- Kyoto: 120TB restart

storage: 5.4PB (9sites,30servers)

4.4P(82%) used

66M files stored

as of March 10, 2015

T. Yoshie for JLDG team

Operation Policy

- Any lattice QCD collaborations in Japan can use the JLDG without charge with no quota
- can store data of any type (configurations, quark propagators, etc.) if these data should be shared within collaboration

- JLDG team discusses everything of JLDG
  - developers (computer scientists at Tsukuba)
  - representatives of JLDG sites
  - representatives of collaborations

- No JLDG specific budgets
  - part of physics/computer science budget is devoted

JLDG team と budget

- **JLDG team:　26名（11機関＋企業）**
  - 建部,天笠,山﨑,吉江(筑波),松古(KEK),外川,鎌野,石井(大阪),石川 (広島),武田(金沢),
    實本(東京),青木,青山,三浦(名古屋),青木,福村(京都),渡邊,土井(理研),
    駒,住吉(沼津高専),滝脇(国立天文台),
    三上,金野,佐々木,首藤,木村(日立ソリューションズ東日本)

- former collaborator
  - 宇川,佐藤,浮田(筑波)

- budget
  - 日本学術振興会先端研究拠点事業「計算素粒子物理学の国際研究ネットワークの形成」
  - 国立情報学研究所CSI 委託事業「グリッド・認証技術による大規模データ計算資源の連携基盤の構築」
  - 国立情報学研究所「e-science 研究分野の振興を支援するCSI 委託事業」の研究課題「計算素粒子物理学の高度データ共有基盤JLDG の構築」及び「計算素粒子物理学のデータ共有基盤JLDGの高度化」
  - 新学術領域・素核宇宙融合「分野横断アルゴリズムと計算機シミュレーション」
  - 最先端研究基盤整備事業業「eーサイエンス実現のためのシステム統合・連携ソフトウェアの高度利用促進」
  - **HPCI戦略プログラム分野5「物質と宇宙の起源と構造」**

Progress of JLDG 2005-2014 : Chronology

- ✓ **2005/11 Start of Development**
- ✓ **2007/03 Prototype implementation on 5 sites**
- ✓ **2008/06 Official start of operation, connected to ILDG**
- ✓ **2009/12 Research groups started to store daily research data (user/group access control)**
- ✓ **2011/12 FUSE mount (gfarm2fs) operation started**
- ✓ **2012/06 Two new sites joined JLDG**
- ✓ **2013/12 Cooperation with HPCI Shared Storage mount both JLDG and HPCI SS file systems, copy files in multiple streams**
- ✓ **2014/03 More 2 sites joined JLDG**

HPCI: national project started in 2011 for constructing High Performance Computing Infrastructure.

JLDGの利用形態

Japan Lattice Data Grid

SR16000@KEK

KEK

1130TB

ある拠点のスパコンで基礎データを生成し JLDG に格納
別拠点のスパコンで読み出し、物理量を計算

どの拠点からも同一のファイルシステム
拠点のアカウントに依存しないユーザー管理
（グリッド証明書に基づく仿想組織）

HA-PACS@U.Tsukuba

データの一元管理

Kanazawa

100TB
down

CCS

Tsukuba

2800TB

Hiroshima

230TB

gfarm

Riken (Wako)

120TB new

Tokyo (Kashiwa)

470TB

SX9@U.Osaka

Nagoya

300TB

Osaka

300TB

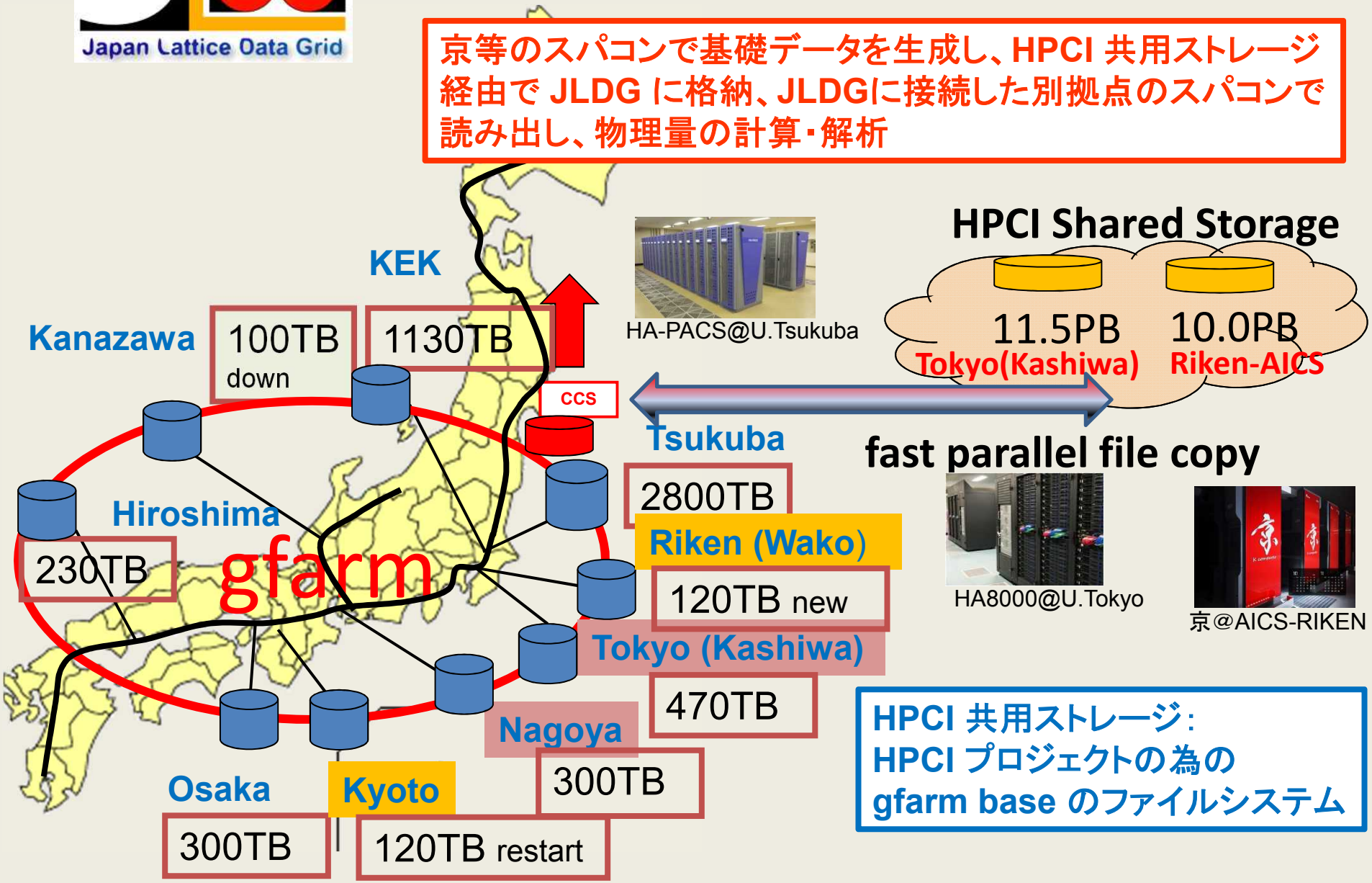
Kyoto

120TB restart

HPCI共用ストレージ･JLDG連携システム

Japan Lattice Data Grid

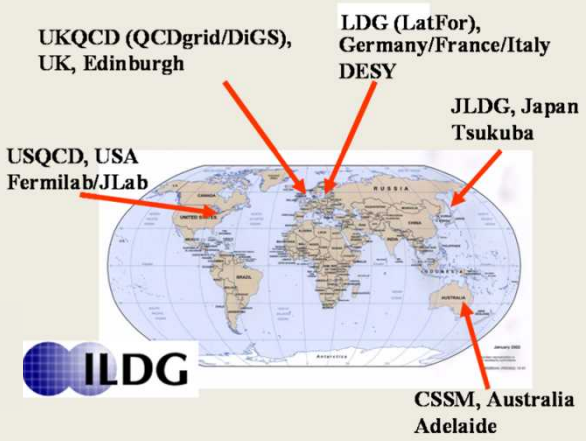
京等のスパコンで基礎データを生成し、HPCI 共用ストレージ経由で JLDG に格納、JLDGに接続した別拠点のスパコンで読み出し、物理量の計算･解析



HPCI 共用ストレージ:
HPCI プロジェクトの為の
gfarm base のファイルシステム

ILDGとの連携

ILDG: International Lattice Data Grid
5つの地域グリッド間でQCD配位（基礎データ）を共有する為の Grid of Data Grids



part of data can be accessed from ILDG

Japan Lattice Data Grid

KEK
1130TB

Kanazawa
100TB
down

CCS

Tsukuba
2800TB

Riken (Wako)
120TB new

Hiroshima
230TB

gfarm

Tokyo (Kashiwa)
470TB

Nagoya
300TB

Osaka
300TB

Kyoto
120TB restart

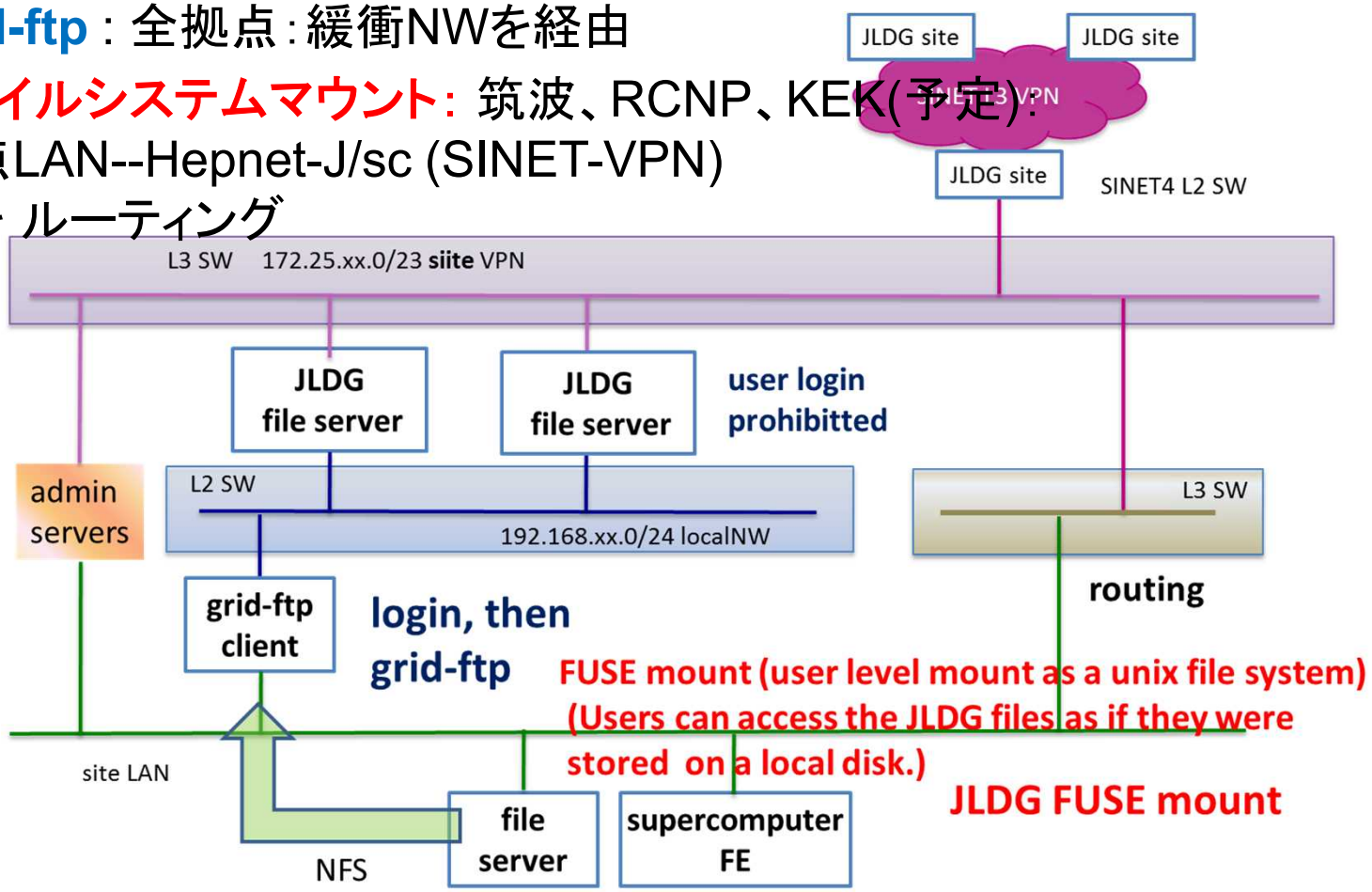
QCD配位をJLDGの特定のディレクトリに、特定のフォーマットで格納

ILDGからアクセス

JLDGはILDGの日本地域グリッドとして機能

JLDG:システムと利用状況

- **grid-ftp** : 全拠点 : 緩衝NWを経由
- **ファイルシステムマウント:** 筑波、RCNP、KEK(予定):
  拠点LAN--Hepnet-J/sc (SINET-VPN)
  間を ルーティング



クライアントにログインし、グリッド証明書から代理証明書を生成

```
jldg-fr3[103]% grid-proxy-init
Your identity: /C=JP/O=Japan Lattice Data Grid/OU=pacscs/CN=Tomoteru Yoshie
Enter GRID pass phrase for this identity:
Creating proxy ..................................... Done
Your proxy is valid until: Tue Mar 10 16:58:51 2015
```

```
jldg-fr3[114]% ls -l RC64x64-hM-002090
-rw-r--r-- 1 yoshie LATTICE 9663677344 Apr 30  2013 RC64x64-hM-002090
```

約9GB のファイルを grid-ftp で JLDG に投入

```
jldg-fr3[115]% uberftp jldg-fs9
220 jldg-fs9 GridFTP Server 3.41 (gcc64, 1330711604-80) [Globus Toolkit 5.0.5] ready.
230 User nobody logged in.
UberFTP (2.8)> cd /gfarm/pacscs/LATTICE/yoshie
UberFTP (2.8)> put RC64x64-hM-002090
RC64x64-hM-002090: 9663677344 bytes in 5 Minutes 27.489634 Seconds (28.141 MB/s)

jldg-fr3[134]% gfwhere /gfarm/pacscs/LATTICE/yoshie/RC64x64-hM-002090
jldg-fs17-sc scjldgkek07 jldg-fs14-sc
```

オリジナルと複製が3サーバに作成される

JLDGをマウントして、linux のコマンドでファイルを操作

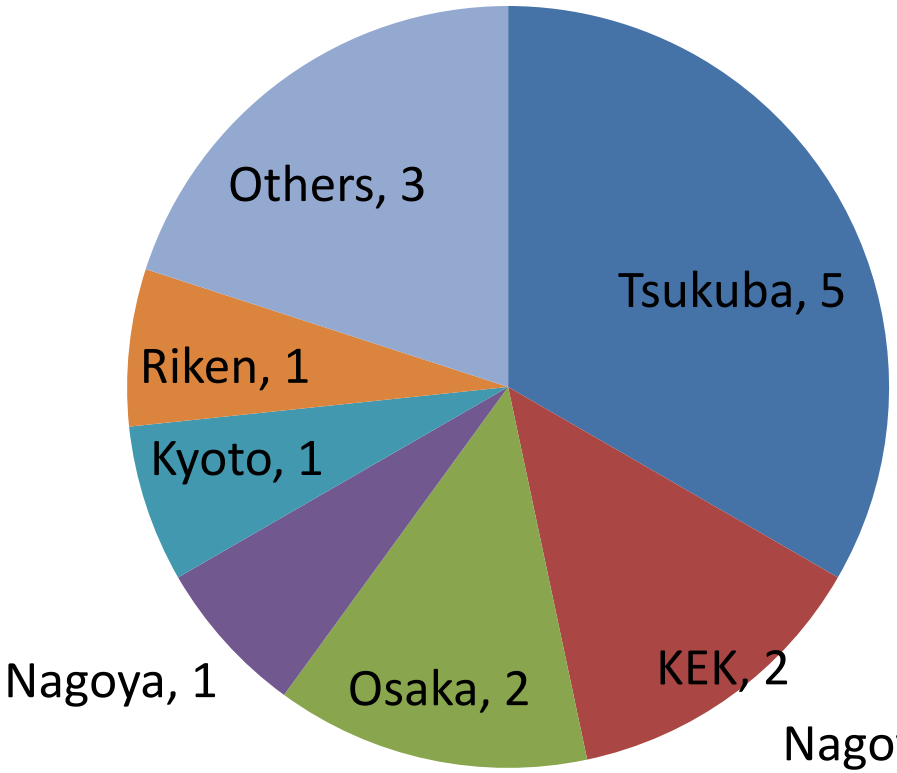
```
[yoshie@hapacs-2 ~]$ gfarm2fs /tmp/yoshie
[yoshie@hapacs-2 ~]$ cd /tmp/yoshie/gfarm/pacscs/LATTICE/yoshie/
[yoshie@hapacs-2 yoshie]$ ls -l RC64x6-hM-002090
-rw-rw-r-- 1 yoshie 70001 9663677344  3月  9 12:40 2015 RC64x64-hM-002090

[yoshie@hapacs-2 yoshie]$ cp RC64x64-hM-002090 .
[yoshie@hapacs-2 yoshie]$ ls –l RC64x64-hM-002090
-rw-r--r-- 1 yoshie WMFQCD 9663677344  3月  9 12:52 2015 RC64x64-hM-002090
```

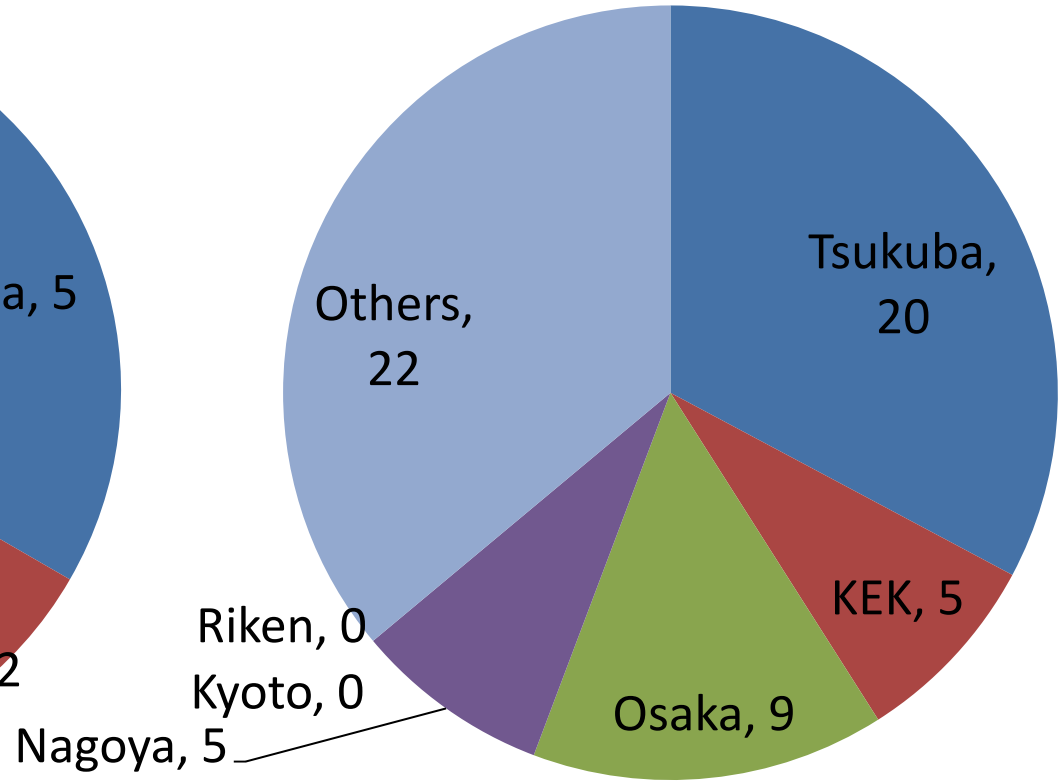
JLDGファイルシステムの特徴
Linux ファイルシステムと比較して

- ディスクスペースの追加が容易
  - ファイルサーバを購入し、システムに組み込み
- 機関（拠点）とは独立なユーザー・グループ管理
  - JLDG用のユーザー証明書：グリッド証明書/電子証明書
  - ユーザーが所属するグループ管理：仮想組織
- 自動ファイル複製とデータ保全
  - JLDG では、3サーバ上に3つの複製(オリジナル含む)を保持
  - システム障害等で全複製が失われたケース O(10)
- 柔軟なアクセス制御
  - ファイル/ディレクトリに対するユーザー・グループ単位のアクセス制限
  - 特定ユーザーに個別のアクセス権限を与える事も可能
- ファイルの高速並列コピー
- “silent” data corruption の検知
  - ファイル（と複製)作成時に、自動で md5sum を計算・比較

JLDGのシステム状況・利用状況

研究グループとユーザー



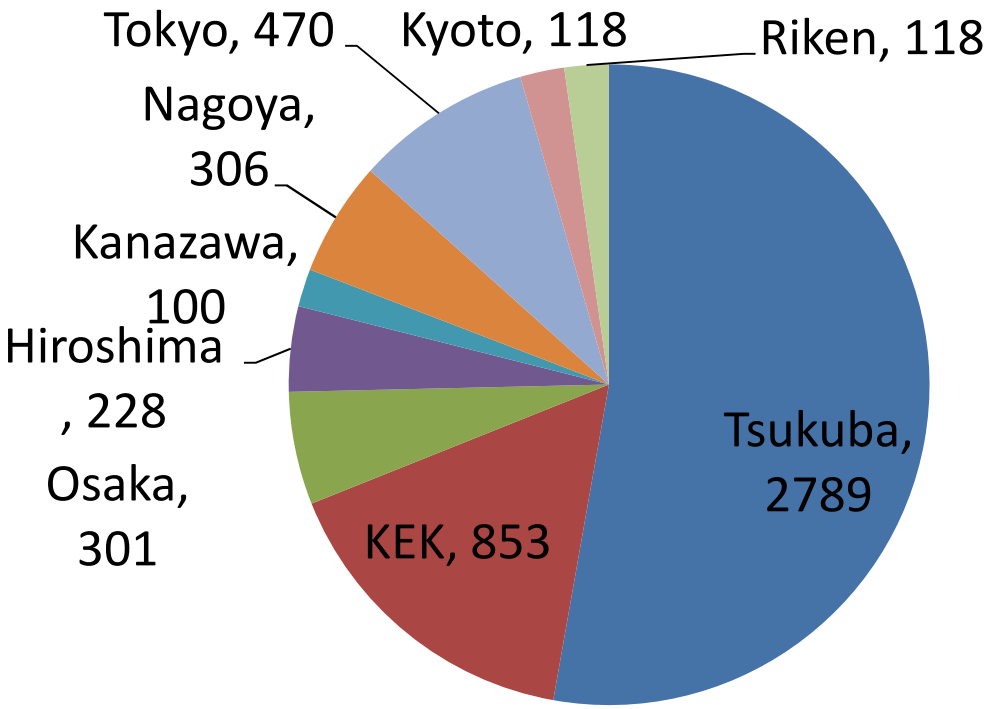


JLDG利用研究グループ数
（主管拠点毎）
計15グループ (Astro 1 含む）

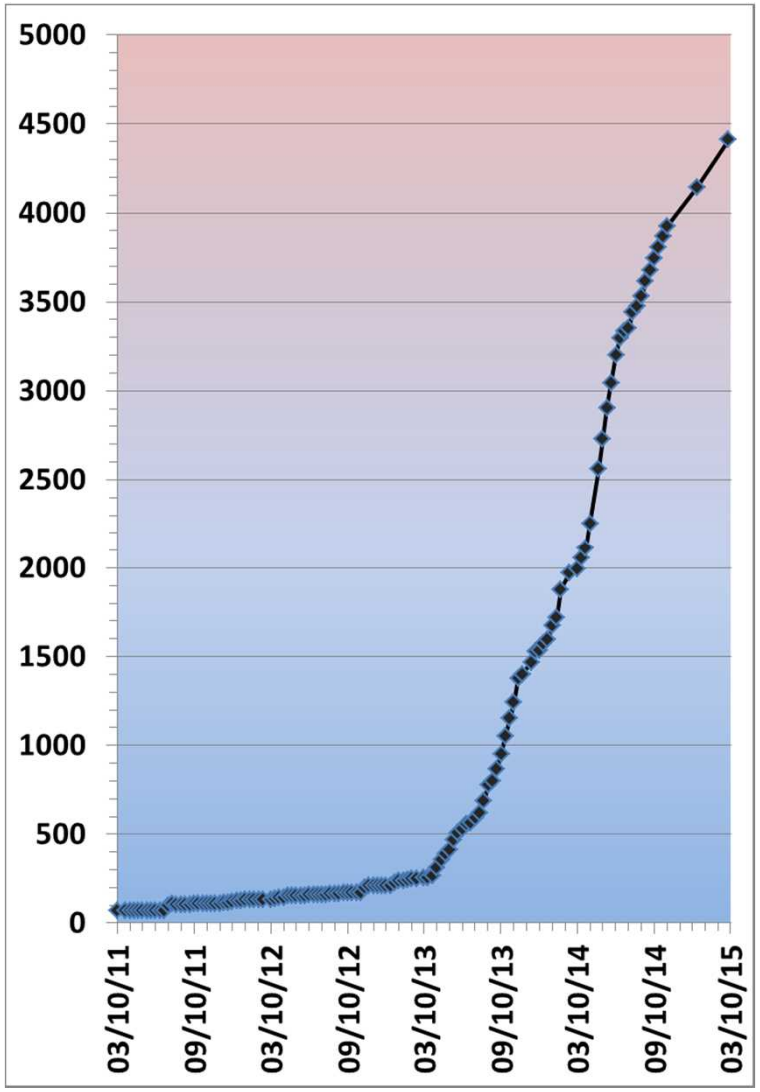
7グループ (2011/11時点）

ユーザ数
（登録時の所属毎）
計61ユーザー

52ユーザー(2011/11時点)



ディスクスペース設置量(TB)
計5283TB
(as of March 9, 2015)



使用量推移(TB)
4.4TB (66M files)
130TB (3M files) (2011/11 時点)

ロデータ共有・管理のためJLDGを利用した研究の成果発表件数

発表年	成果発表数
2007	3
2008	4
2009	7
2010	11
2011	20
2012	16
2013	30
2014 (9月まで)	7
Total	98

JLDG (ILDG 含む)自身の発表件数: 3件

HPCI共用ストレージ・JLDG 連携システム

両データグリッド間の
並列高速データコピー

gfpcopy

Tokyo(Kashiwa)　Riken-AICS

HPCI 共用ストレージ

Kanazawa
Hiroshima
JLDG
Tsukuba
Riken (Wako)
Tokyo (Kashiwa)
Nagoya
Osaka
Kyoto

mount

mount

mount
grid-ftp

jldghpci
.ccs.tsukuba.ac.jp　連携システムホスト

代理証明書
委譲(自動)

GSI認証
ログイン

JLDG クライアント

代理証明書
コピー(手動)

代理証明書
ダウンロード

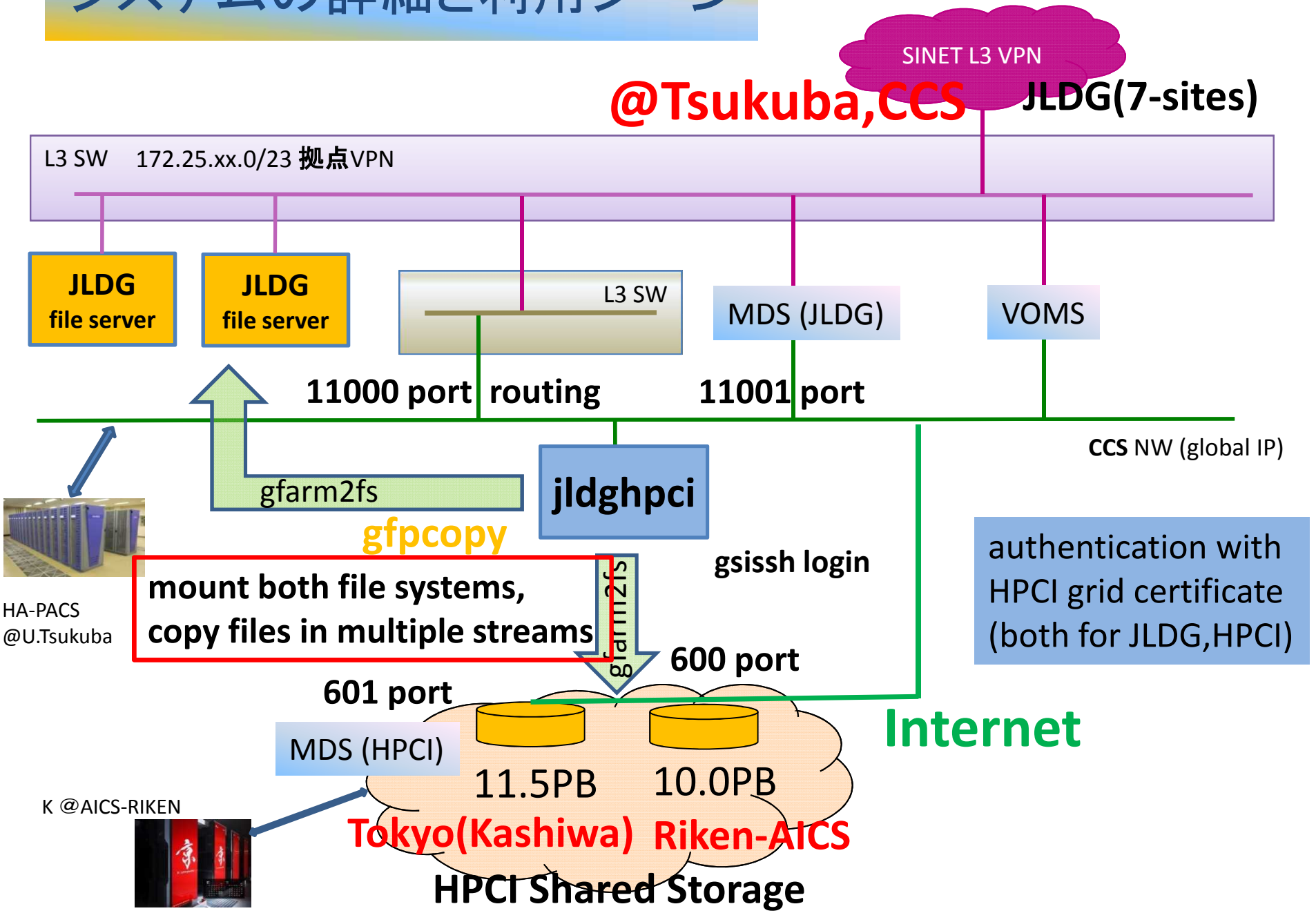
又は、
証明書本体
ダウンロード

HPCI電子証明書
発行システム

HPCI-SS　JLDG 共、HPCI 電子証明書による認証

HPCI利用研究課題「HPCI共用ストレージ・JLDG連携」

システムの詳細と利用シーン



HPCI電子証明書から代理証明書を生成

```
lyra1.ccs.tsukuba.ac.jp[106]% grid-proxy-init
Your identity: /C=JP/O=NII/OU=HPCI/CN=Tomoteru%40Yoshie[hpci000151]
Enter GRID pass phrase for this identity:
Creating proxy ..................................................... Done
Your proxy is valid until: Tue Mar 10 17:19:19 2015
```

連携システムに gsissh でログイン

```
lyra1.ccs.tsukuba.ac.jp[107]% gsissh jldghpci
Last login: Wed Mar  4 13:04:52 2015 from lyra1.ccs.tsukuba.ac.jp
```

設定ファイルを指定して HPCI-SS JLDG を順次マウント

```
[yoshie@jldghpci ~]$ mount.gfarm2fs /etc/gfarm2.conf-hpci ~/HPCI gfarmfs_root=/
Mount GfarmFS on /home/yoshie/HPCI
[yoshie@jldghpci ~]$ mount.gfarm2fs /etc/gfarm2.conf-jldg ~/JLDG gfarmfs_root=/
Mount GfarmFS on /home/yoshie/JLDG
```

確かに！

```
[yoshie@jldghpci ~]$ df -H
Filesystem            Size  Used Avail Use% Mounted on
……..
gfarm2fs               23P   15P  8.1P  65% /home/yoshie/HPCI
gfarm2fs              5.4P  4.5P  933T  83% /home/yoshie/JLDG
```

並列コピーの実行例

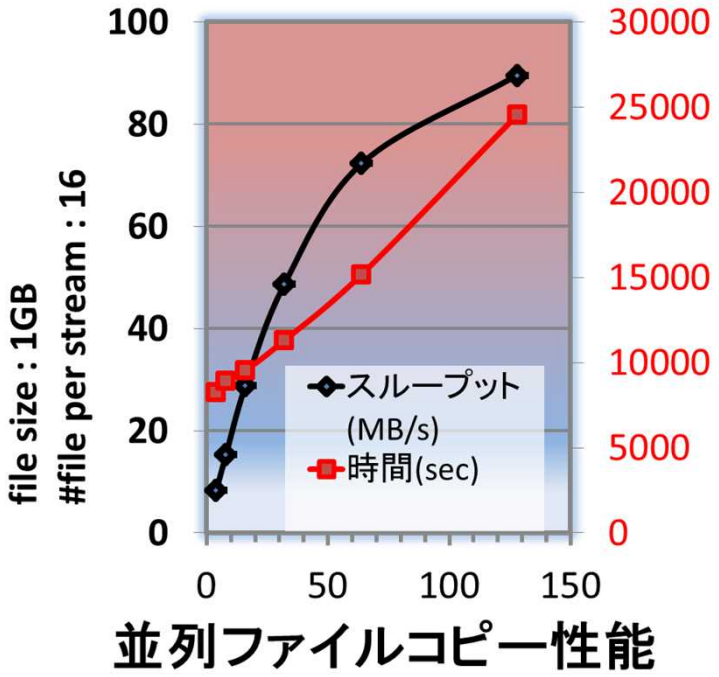
```
jldghpci % gfpcopy -P -j 128  ~/HPCI/home/hp120108/hpci000151/ConfData-2048 ¥
                                        ~/JLDG/gfarm/pacscs/hpci/
[OK]COPY, 1.04MB/s(1.04e+03s): gfarm://esci-epgfm01.cspp.cc.u-tokyo.ac.jp:601/ho
me/hp120108/hpci000151/ConfData-2048/config-0977.dat(esci-wgfs001.aics.riken.jp:
600) -> gfarm://mds2.jldg.org:11001/gfarm/pacscs/hpci/ConfData-2048/config-0977.
dat(jldgnagfs0-s:11000)
.....
[OK]COPY, 0.476MB/s(2.26e+03s): gfarm://esci-epgfm01.cspp.cc.u-tokyo.ac.jp:601/h
ome/hp120108/hpci000151/ConfData-2048/config-1648.dat(esci-epgfd205.cspp.cc.u-
tokyo.ac.jp:600) -> gfarm://mds2.jldg.org:11001/gfarm/pacscs/hpci/ConfData-2048/config-
1648.dat(jldghu03:11000)
.....
copied_file_size: 2199023255552
total_throughput: 89.562374 MB/s
total_time: 24552.980815 sec.
```

gfpcopy で並列コピーを開始

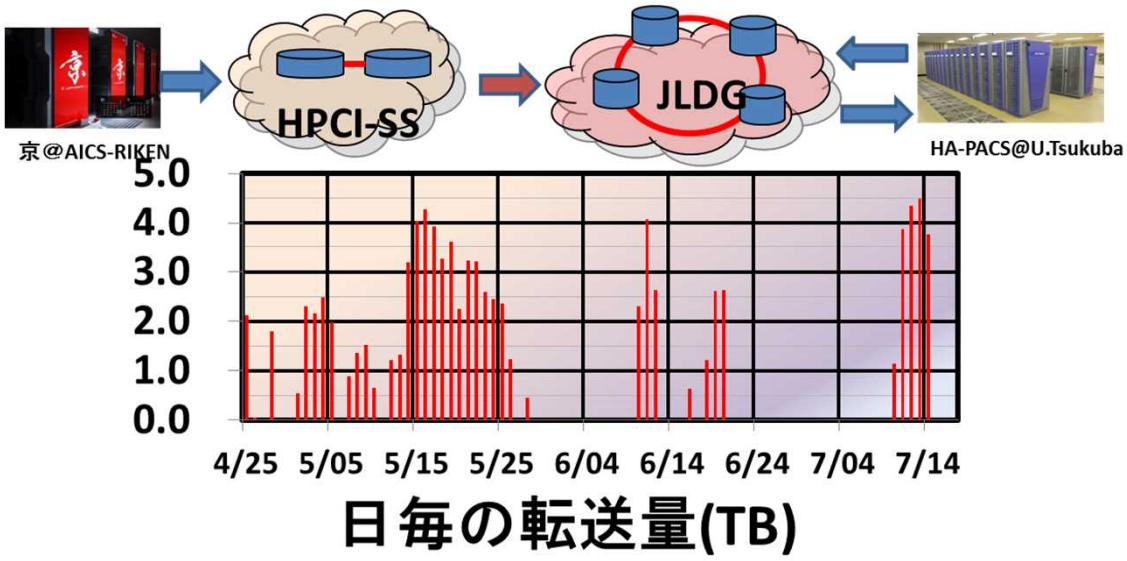
HPCI (Riken-AICS)→JLDG (Nagoya)

HPCI (U.Tokyo)→JLDG (Hiroshima)

性能と利用状況



利用状況(典型例)



多量の大きなファイルを並列コピー: 90MB/s (1Gbps NW)

実際の研究に使われている: max 4TB/day

JLDGの今後

- JLDG は、国内の計算素粒子物理研究のインフラとなっている
  - ✓ 運用の継続、システムの拡充は継続したい

- 計画: 現システムの延長上
  - ✓ 利便性の向上 : 各拠点でファイルシステムマウントの実現 （KEK 作業開始
  - ✓ SINET5 への移行 （一部拠点 NW 1Gbps → 10 Gbps へ）
  - ✓ 安定性･可用性の向上
    - 古い管理機器の更新、管理機器の二重化
    - ファイル複製の作成法の見直し
    - 障害検知と管理者への通知
  - ✓ 安全性の向上 : セキュリティ維持の為の方法の整理とルーチン化

- 懸念 : ディスク使用量の急激な増大
  - ✓ ファイルサーバを需要にあわせて追加できるか : 予算の獲得ができるか
  - ✓ 使用量の抑制が必要か（自主的に、quota 制限）
  - ✓ 来年度末までの見込み : 1.6PB 増設 （大半はHPCI予算ではない）

- 維持･管理体制 : ・・・・・