JLDGの現状と計画

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
2013年3月5,6日
於：富士ソフトアキバプラザ

筑波大学計算科学研究センター
吉江友照

序：計算素粒子物理のデータ共有

□ HPCに於けるデータ管理・活用の2つの側面

- **研究グループ内でのデータマネージメント・共有**
  - ✓ 複数の研究機関に属する複数の研究者が複数のスパコンを利用
  - ✓ どの機関でも同じデータ（構造）が見える
  - ✓ 研究機関をまたぐユーザー認証・管理システム
- **コミュニティの資産としてのデータ蓄積とその活用**

□計算素粒子物理コミュニティーのデータグリッド

- **JLDG : Japan Lattice Data Grid**
  - ✓ 各機関のスパコンで生成されるデータを、機関を跨いで共有
  - ✓ 2008実運用開始, 2011からHPCI戦略プログラム分野5で
- **ILDG : International Lattice Data Grid**
  - ✓ 基礎データ (QCD 配位)をコミュニティで蓄積・利用

JLDG team と budget

- JLDG team
  - **天笠(筑波), 建部(筑波), 浮田(筑波), 吉江(筑波),松古(KEK), 外川(大阪), 石川(広島), 武田(金沢),駒(沼津高専), 實本(東京), 青木･青山･山崎(名古屋), (株)日立ソリューションズ東日本**
- former collaborator
  - **宇川彰(筑波), 佐藤三久(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－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」**

目次

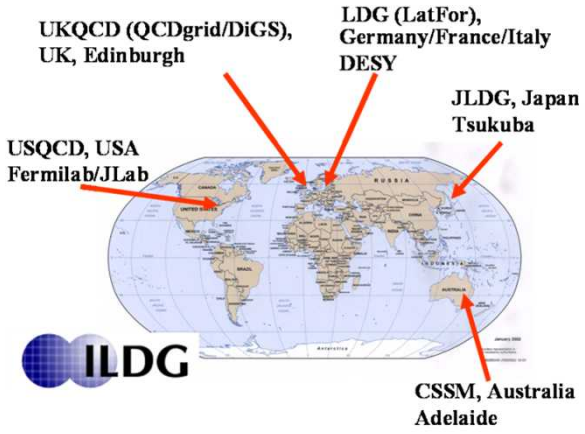
- 序: 計算素粒子物理のデータ共有

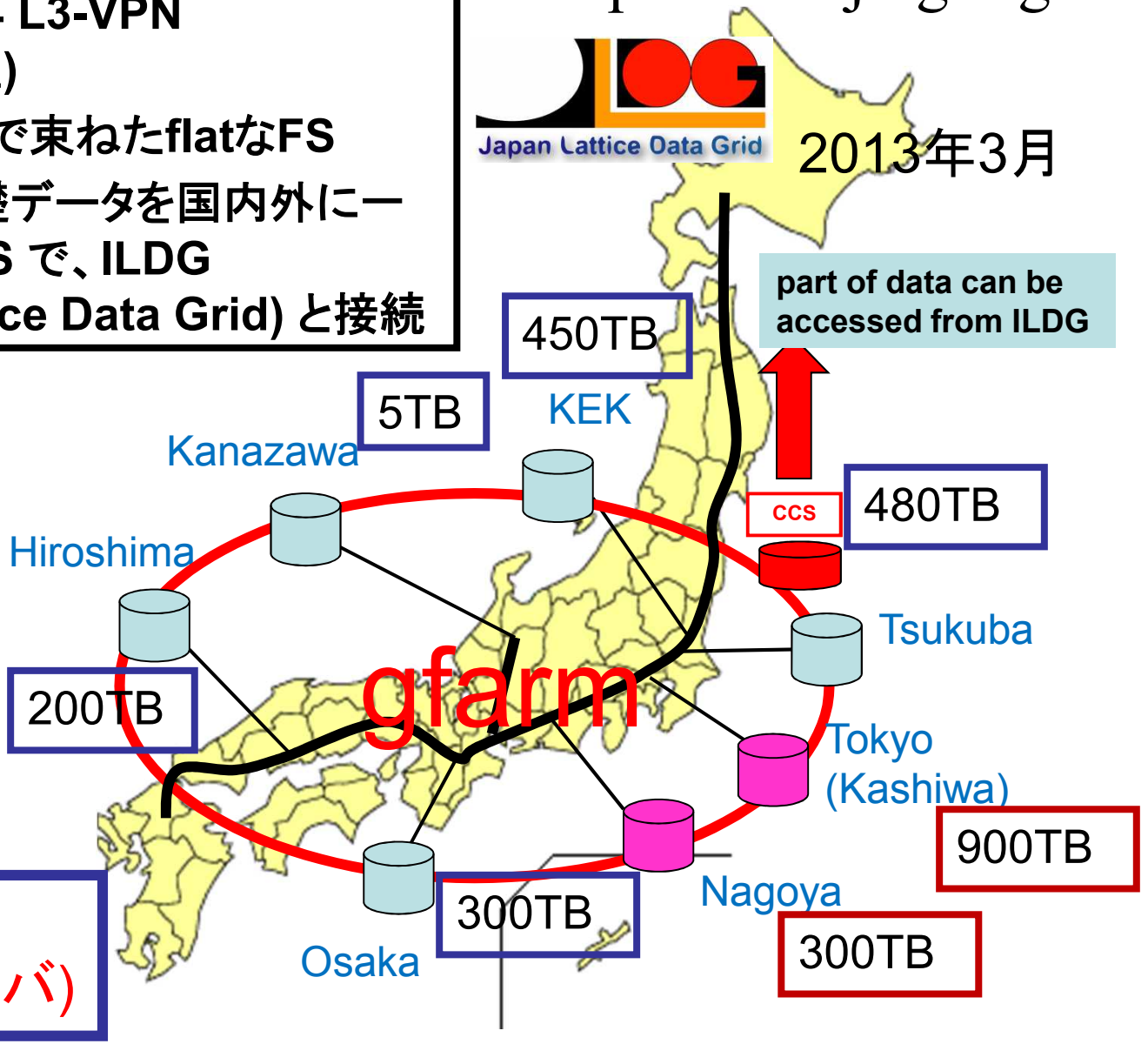
- JLDGの概要
- 利用シーン
- 運用・利用状況
- 今年度の進展と来年度以降の可能性
- 長期戦略

JLDGの概要:システム

- **Backbone: SINET4 L3-VPN (NII 提供, KEK 管理)**
- **7拠点のFSをgfarmで束ねたflatなFS**
- **Lattice QCD の基礎データを国内外に一般公開。筑波大 CCS で、ILDG (International Lattice Data Grid) と接続**

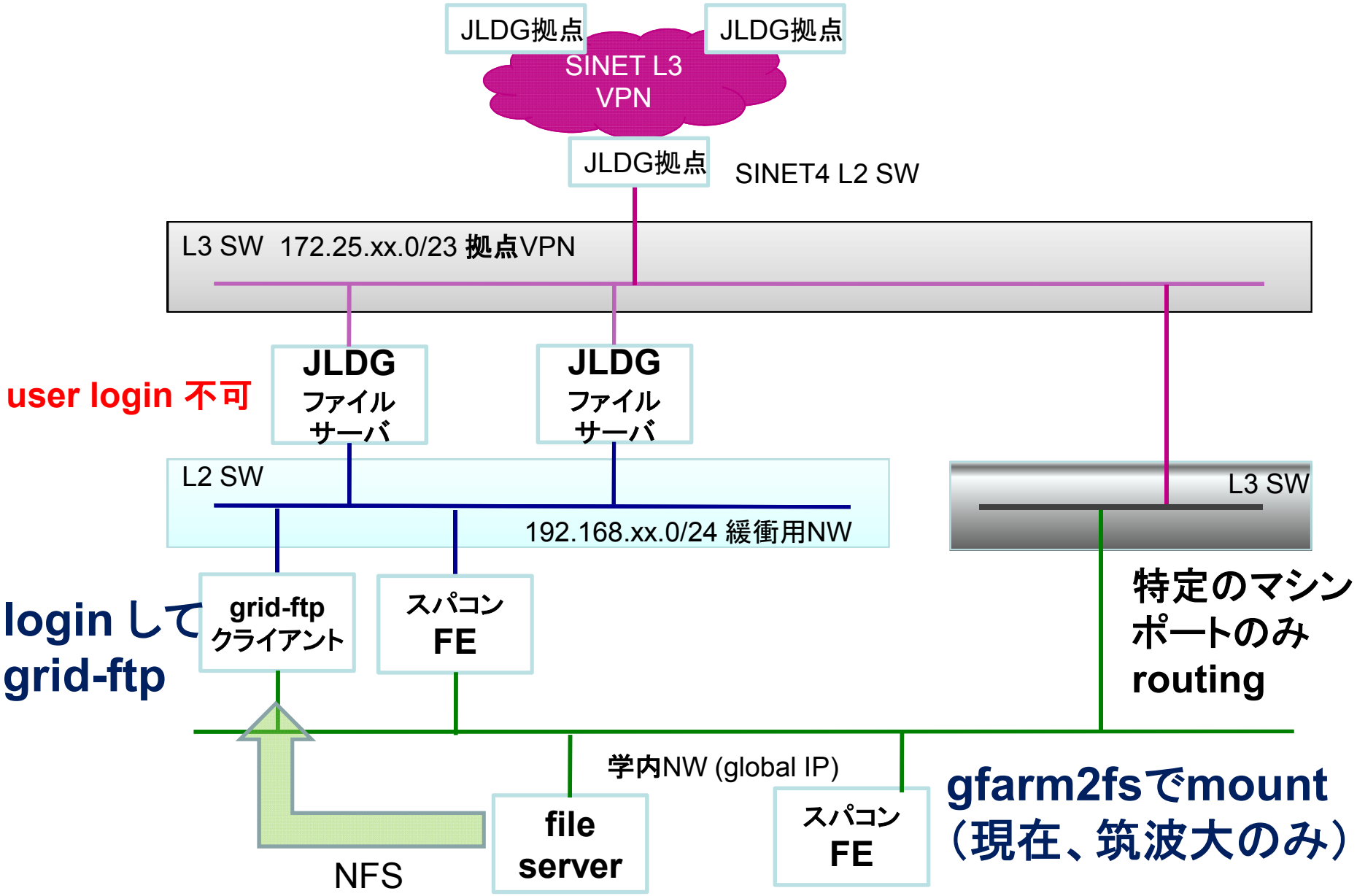
http://www.jldg.org/





総計2.6PB
(7機関,20サーバ)

JLDGの概要：ネットワーク



利用シーン（1）

- grid-ftp base: **uberftp (interactive/non-interactive)**

KEK SR16K FE で

```
ntcf01c01p03[14]% uberftp scjldg05.sc.kek.jp
UberFTP> cd /gfarm/pacscs/junk
UberFTP> put config-001010
config-001010: 1073741824 bytes in 1.509481 Seconds (678.379 MB/s)

jldg-fr3[101]% gfwhere /gfarm/pacscs/junk/config-001010
hn-oss45 scjldgkek05

jldg-fr3[104]% uberftp jldg-fs9
UberFTP> cd /gfarm/pacscs/junk
UberFTP> get config-001010
config-001010: 1073741824 bytes in 22.778923 Seconds (44.954 MB/s)
```

1GBのファイルをput

KEKと東大にファイルが作られた

筑波大でget

結構速い！

参考： **KEK—Tsukuba scp: 4.5MB/s**

利用シーン(2)

◘ fuse-mount : unix ファイルシステムとしてアクセス

ファイルシステムとしてマウント

```
[yoshie@hapacs-2 ~]$ gfarm2fs /tmp/yoshie
[yoshie@hapacs-2 ~]$ df
gfarm2fs          2367209577084 236921516892 2130288060192  11% /tmp/yoshie
[yoshie@hapacs-2 junk]$ cd /tmp/yoshie/gfarm/pacscs/junk
[yoshie@hapacs-2 junk]$ ls
[yoshie@hapacs-2 junk]$ cp /work/WMFQCD/yoshie/Dummy/config-0010* .
```

HA-PACS でJLDGに書き込み

```
flare24[195]% cd /tmp/yoshie/gfarm/pacscs/junk
flare24[196]% ls -l
total 2359296
-rw-r--r-- 1 yoshie 70001 268435456 Mar  1 14:16 config-001010
-rw-r--r-- 1 yoshie 70001 268435456 Mar  1 14:16 config-001020
-rw-r--r-- 1 yoshie 70001 268435456 Mar  1 14:16 config-001030
-rw-r--r-- 1 yoshie 70001 268435456 Mar  1 14:16 config-001040
-rw-r--r-- 1 yoshie 70001 268435456 Mar  1 14:16 config-001050
-rw-r--r-- 1 yoshie 70001 268435456 Mar  1 14:16 config-001060
```

WS で読み出し

どんなコマンドでも使える (emacs で編集など), プログラムから読み書きできる。

利用シーン（3）

gfpcopy: 高速並列コピー

```
[yoshie@hapacs-2 Dummy]$ gfpcopy -p ./ gfarm:///gfarm/pacscs/junk
copied_file_num: 100
copied_file_size: 26843545600
total_throughput: 29.481515 MB/s
total_time: 910.521250 sec.
```

256MB, 100 file を JLDG にコピー
30MB/s: そこそこ？

```
gfwhere -r . | grep -v '^$' | less
flare24[252]% gfwhere -r . | grep -v '^$' | less
gfarm://mds1.jldg.org:11001/gfarm/pacscs/junk/Dummy/config-001150:
jldg-fs9-sc hn-oss45
gfarm://mds1.jldg.org:11001/gfarm/pacscs/junk/Dummy/config-001160:
scjldgkek05 scjldgkek06
gfarm://mds1.jldg.org:11001/gfarm/pacscs/junk/Dummy/config-001170:
jldgnagfs0-s hn-oss47
……
gfarm://mds1.jldg.org:11001/gfarm/pacscs/junk/Dummy/config-001230:
jldghu02 hn-oss47
```

あちこちに書かれる

JLDGの運用･利用状況

- セキュリティポリシー
  - 各拠点（大学等）のネットワーク管理主体の承認
- 管理グループ
  - 拠点の代表･利用グループの代表
  - ポリシーの改訂から、HW/SWの保守･管理まで
  - ユーザー認証
- ユーザー･グループ管理
  - private CA によるユーザー認証と仮想組織管理
  - LQCDと関連分野の研究者は誰でも

- 利用状況
  - 研究グループ: 10
  - ユーザー数: 67 + 65(ILDG経由利用)
  - データ量: 243TB, 12M files （内、公開データ 70K 件）
  - 通信量･回数: 記録していない
  - JLDGを利用した研究成果発表数：66件

今年度の進展

- 東大柏拠点(900TB)、名大KMI拠点(300TB)の新設
- 既設拠点の増強
  - 広島(200TB)、筑波(100TB)、KEK(400TB)、大阪(300TB)
- 可用性･安定性の向上
  - メタデータサーバの更新と二重化
    - メモリ不足によるダウンが殆どなくなった
  - gfarm の update
    - ファイル投入時の複製の自動作成が安定に
  - 監視システムの導入（作業中）
- 管理･運用･保守を一部外注
- 利便性の向上
  - KEK スパコン(A)フロントエンドからのJLDG利用
  - 筑波大CCS HA-PACS, WS群 でのJLDG利用

来年度以降の可能性

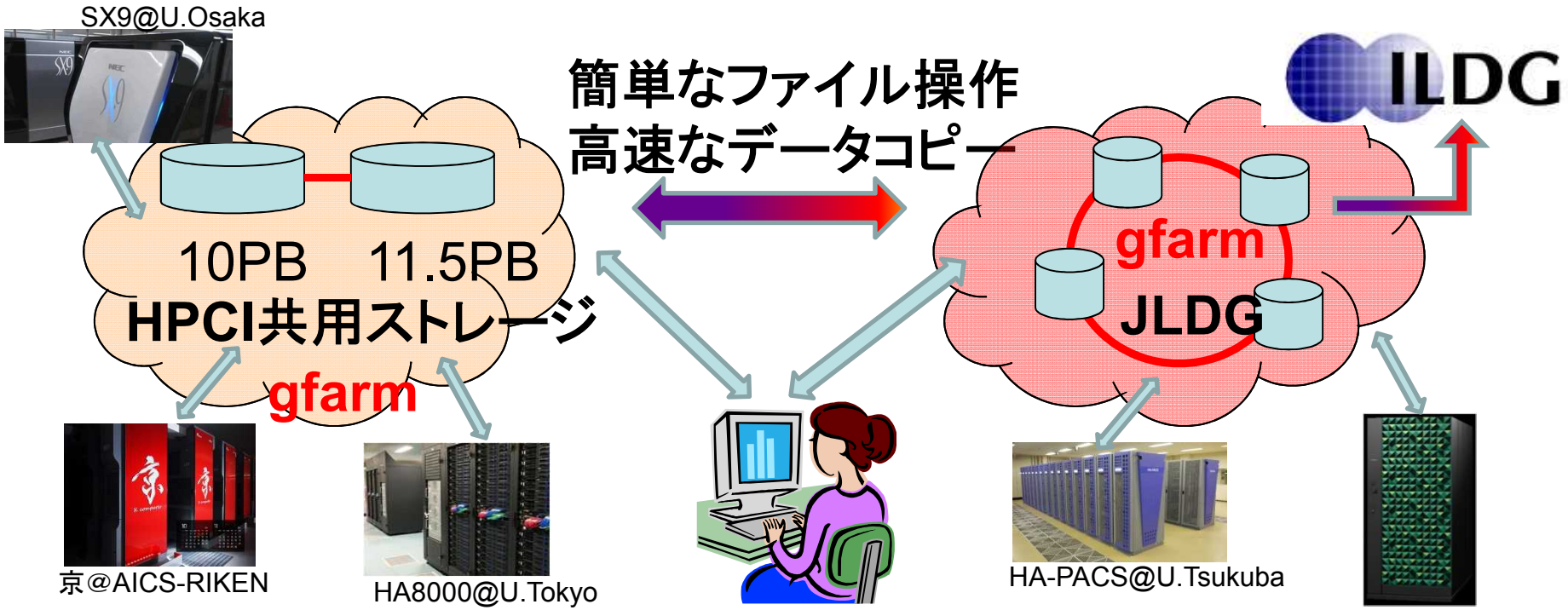
- 可用性･安定性のさらなる向上
  - 管理機器の２重化
    - 特にILDGとの連携用のサーバ
  - ファイル複製の見直し
    - 東西に複製をつくる、など

皆様からのご意見
アイディアを！

- 利便性のさらなる向上
  - JLDGファイルシステムマウントの他機関での実現
    - まずKEKから？
  - 拠点の新設（私案)： 京都YITP, AICS, 理研和光
  - HPCI共用ストレージとの連携

- 分野5の他分野（原子核、宇宙）での利用
  - ニーズはありますか

HPCI共用ストレージとの連携



京などのスパコンで生成した基礎データを共用ストレージや
JLDGに格納、他のスパコンで物理量を計算、WSで二次解析

京を中核とするHPCI利用課題で
課題継続期間のみ利用可

コミュニティの研究者･グループ
データの長期の共有･蓄積
ILDG経由で国内外へ一般公開

- **HPCI共用ストレージとJLDGの連携システムを構築･提供**
- **データ共有･蓄積する際の利便性向上**

スパコン写真は、各々のWeb-siteより転載

HPCI共用ストレージ･JLDG連携

HPCIシステム利用研究課題として推進中

- 技術検討と方針の決定
  - HPCI共用ストレージとJLDGを同時mountできるclient machine を筑波大学計算科学研究センターに設置
  - 両gfarm FS用のconfig file を環境変数で指定し、2回mount
  - gfpcopyコマンドによるHPCI共用ストレージとJLDG間の高速コピー
  - HPCI共用ストレージとJLDGで共通のgrid証明書が必要
  - HPCI電子証明書をJLDGでの認証に使用、証明書のダウンロード許可を (helpdeskを通して)依頼中
- 機材の設置: 済み
  - Dell PowerEdge R420, Xeon E5-2470(2.30GHz,8core)x4, 64GB .
- 予備実験
  - JLDG client を setupし、config-file を入れ替えるだけで、HPCI共用ストレージがmountできる事を確認
- システム構築, 機能･性能検証, 環境整備, 運用開始
  - HPCI 電子証明書による gsi-openssh

長期戦略

- JLDGはコミュニティインフラとして維持すべきか？
  - 各スパコン拠点のディスク容量の巨大化
  - HPCI共用ストレージの運用開始
  - 長期のデータ共有･蓄積の手段としての意義

- どう維持し発展させるか
  - ネットワーク構成の見直し
    - VPN上のシステムを継続するか、グローバルIPに移行するか
    - 性能と運用の手間のバランス
  - 予算（HPCIの次にどうなるか）
  - 体制（開発･構築段階から、継続運用の段階に）