

# 分散共有ファイルシステム

## Gfarm

---

株式会社ベストシステムズ

西 克也

(NPOつくばOSS技術支援センター理事)

# Gfarmファイルシステム

- 2000年より研究開発を続けている
  - 2003年, 国際会議SC03でDistributed Infrastructure Award受賞
  - 2005年, 国際会議SC05でMost Innovative Use of Storage In Support of Science Award受賞
  - 2006年, 国際会議SC06でHPC Storage Challenge優勝

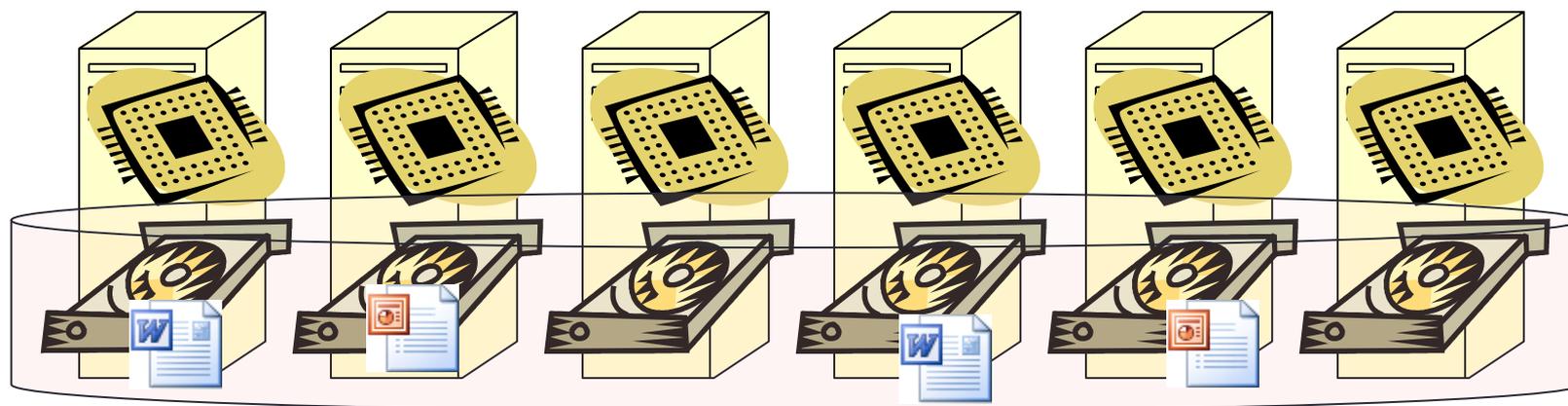


- オープンソース広域分散ファイルシステム  
<http://sf.net/projects/gfarm/>

- **広域**で性能が**スケールアウト**するファイルシステム
  - ファイルサーバ, クライアント追加によるスケールアウト
  - ローカル(近いサーバへの)アクセス優先, ファイル複製
  - 単一障害点なし
    - **自動ファイル複製作成機能**
      - ファイルシステムノード障害時も運用停止無し
    - **ホットスタンバイMDSサーバ**
      - **同期・非同期メタデータ複製機能**

# Gfarmファイルシステムの構成

- 一般的なPCのローカルディスクを束ねる
- ユーザには、共有ファイルシステムとしてみえる
- 複数のディスクに分散してデータを保持

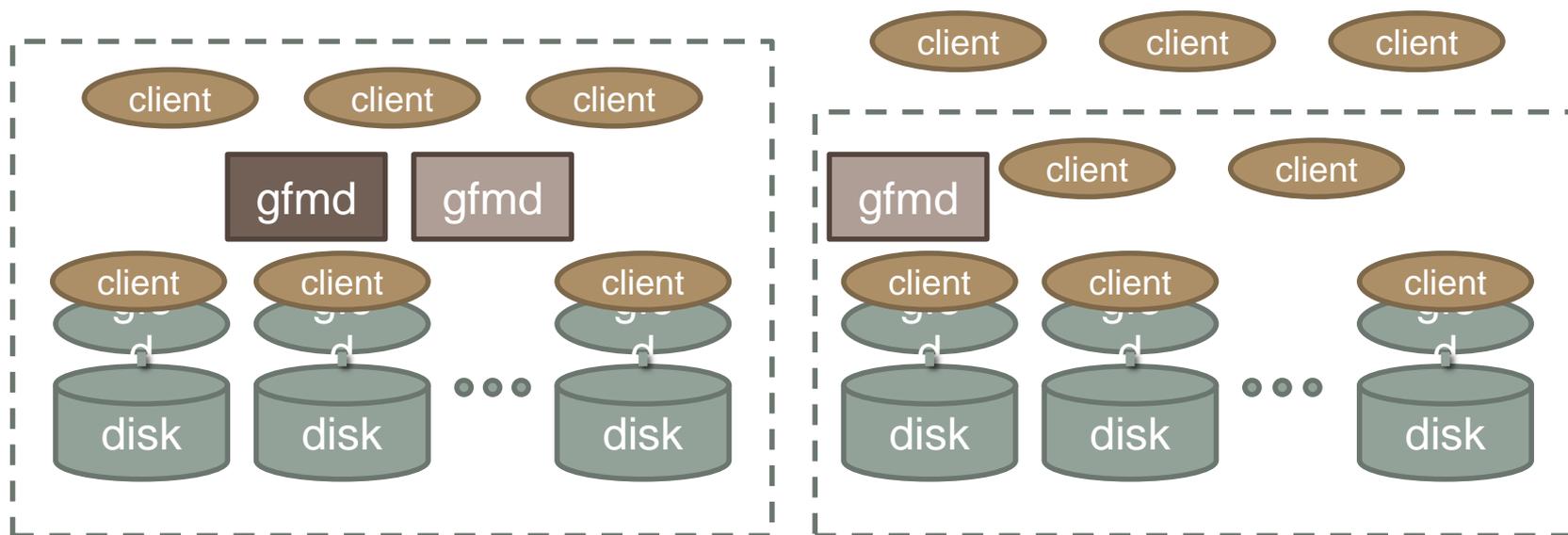


Gfarmファイルシステム



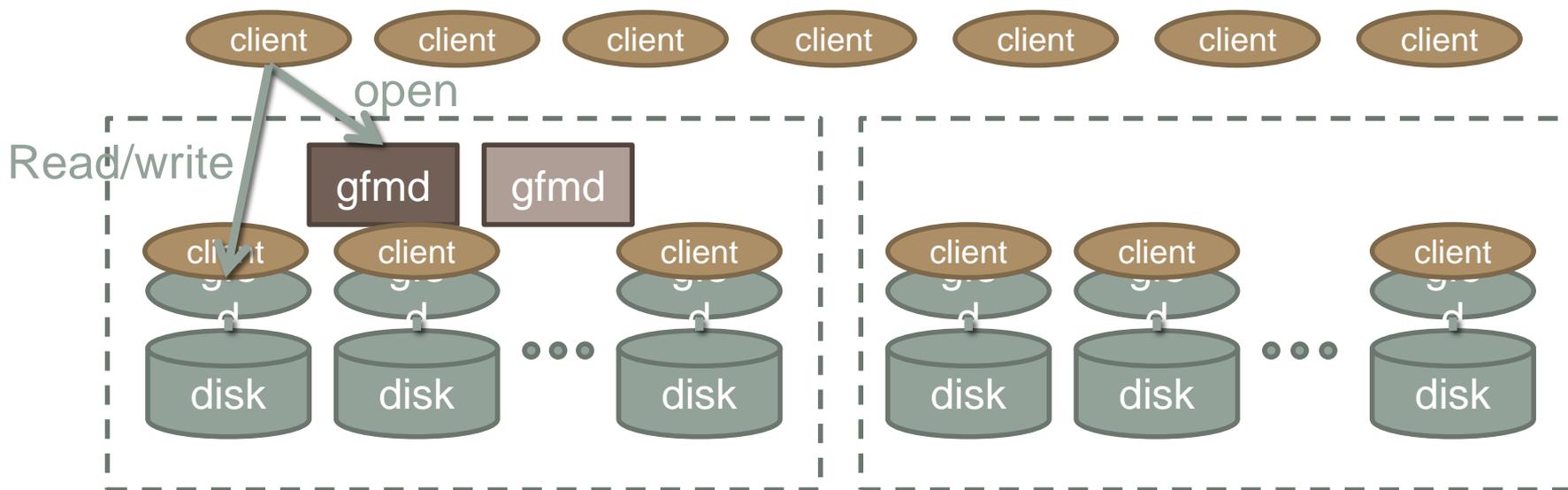
# 構成

- メタデータサーバ (active-standby可) (gfmd)
- 多数ファイルシステムノード (gfsd)
- 多数クライアント
  - ファイルシステムノードと同ノードとして, 分散データ処理!



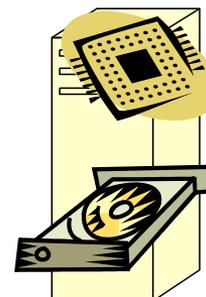
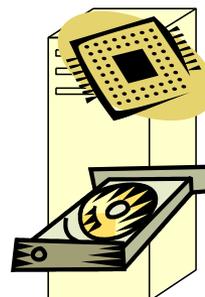
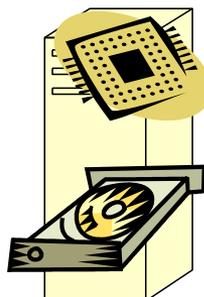
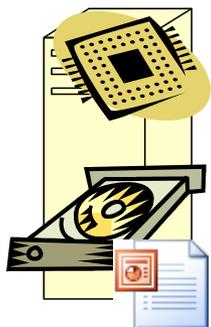
# スケールアウトする構成

- メタデータサーバにはopen, close時だけアクセス
- データアクセスは直接近いファイルシステムノードに
  - ファイルアクセスの分散
- メタデータサーバの処理能力限界まではアクセス性能はスケールアウト



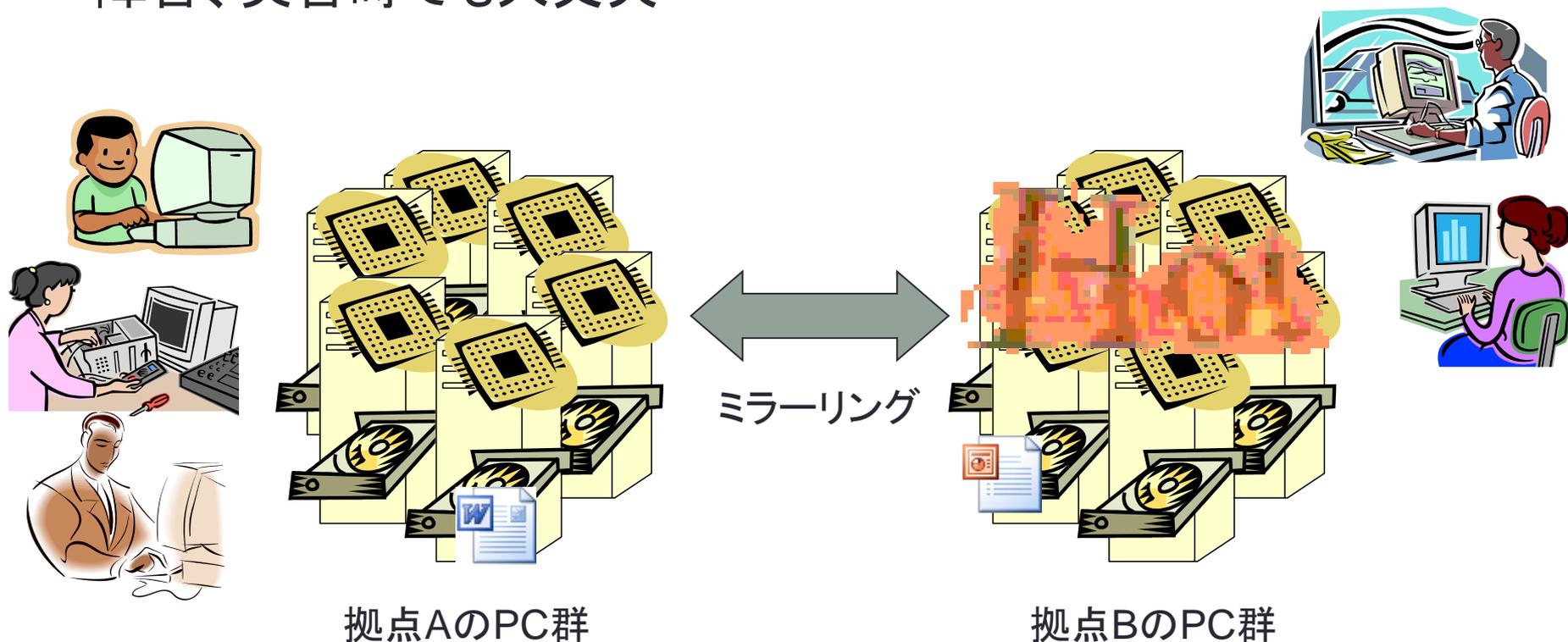
# 利用例(1): 組織内の共有ファイルシステム

- ファイルシステムの容量を運用中に増加
- ファイル複製の数を運用中に増やして、ホットスポットの回避と、信頼性の向上



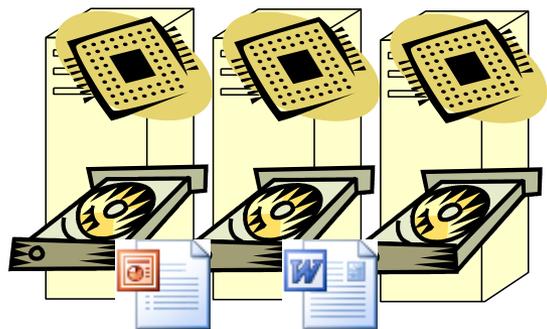
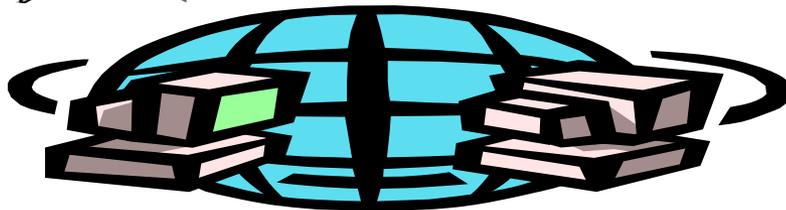
# 利用例(2): 拠点間でのデータ共有

- ミラーリングを行い、それぞれの拠点に保持されたデータをアクセス
  - データが近くにあるため高速なアクセス
- 障害、災害時でも大丈夫

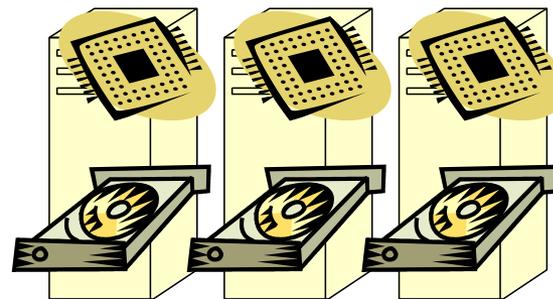


# 利用例(3): 遠隔のファイル格納サービス

- ファイルの複製を地理的に離れた場所に保持することにより、高信頼なサービスを実現



データセンターA



データセンターB



# ダウンロード数

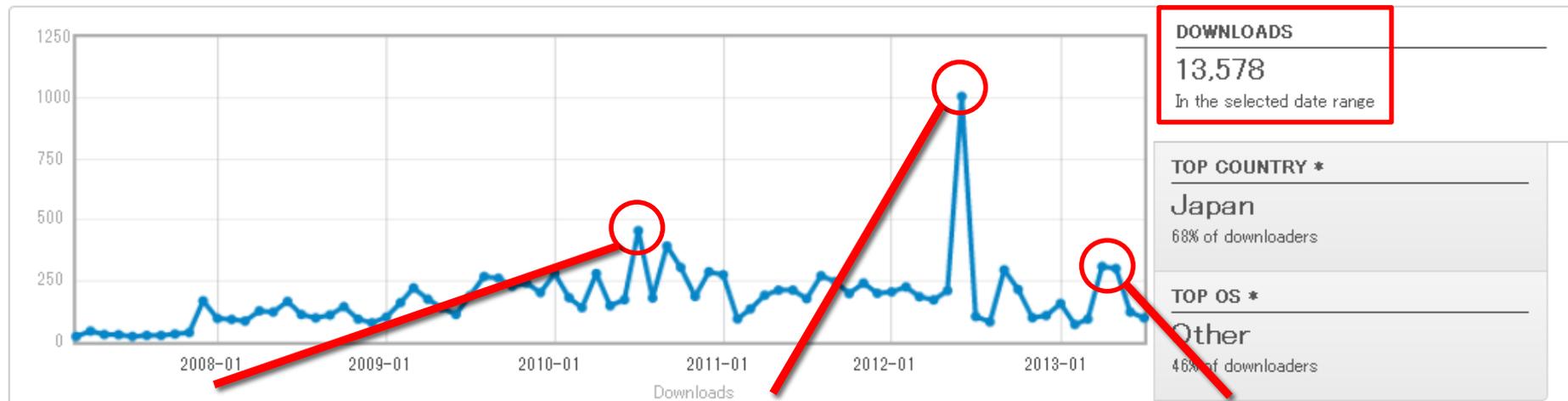
- SourceForgeに移してから13,578ダウンロード

## Gfarm File System

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

Home (Change File)

Date Range: 2007-03-01 to 2013-07-25



2010/7  
Version 2.3.2, 2.4.0  
456 downloads

2012/6  
HPCI導入など  
1,007 downloads

2013/4,5  
Version 2.5.8  
610 downloads

# 最新機能・状況紹介

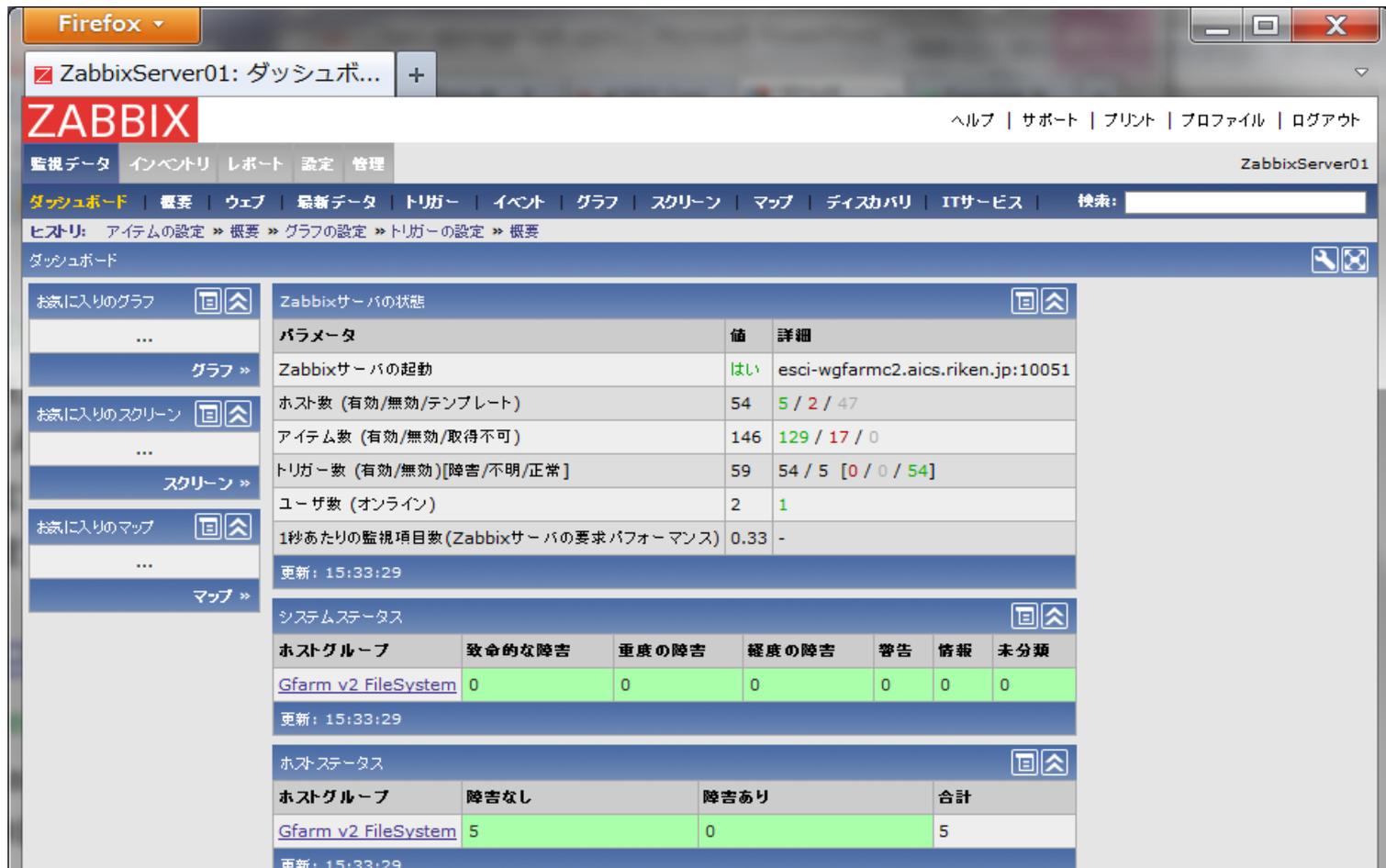
---

# 複製数自動維持 [Gfarm 2.5.8]

- 複製数はディレクトリ、ファイル単位で指定可能
- 正常時は作成時に指定数作成される
- ファイルシステムノード障害発生時、複製指定数変更時などに自動的に複製を裏で作成

# Gfarmファイルシステム運用監視

- Zabbixプラグインにより、各サーバを監視



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 navigation bar includes "ヘルプ | サポート | プリント | プロファイル | ログアウト" and "ZabbixServer01". The left sidebar shows navigation options like "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", "ディスクパリ", "ITサービス", and a search bar. The main content area is divided into several sections:

- お気に入りのグラフ**: A list of favorite graphs with a "グラフ" link.
- お気に入りのスクリーン**: A list of favorite screens with a "スクリーン" link.
- お気に入りのマップ**: A list of favorite maps with a "マップ" link.
- Zabbixサーバの状態**: A table showing server parameters.
- システムステータス**: A table showing system status for "Gfarm v2 FileSystem".
- ホストステータス**: A table showing host status for "Gfarm v2 FileSystem".

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	-

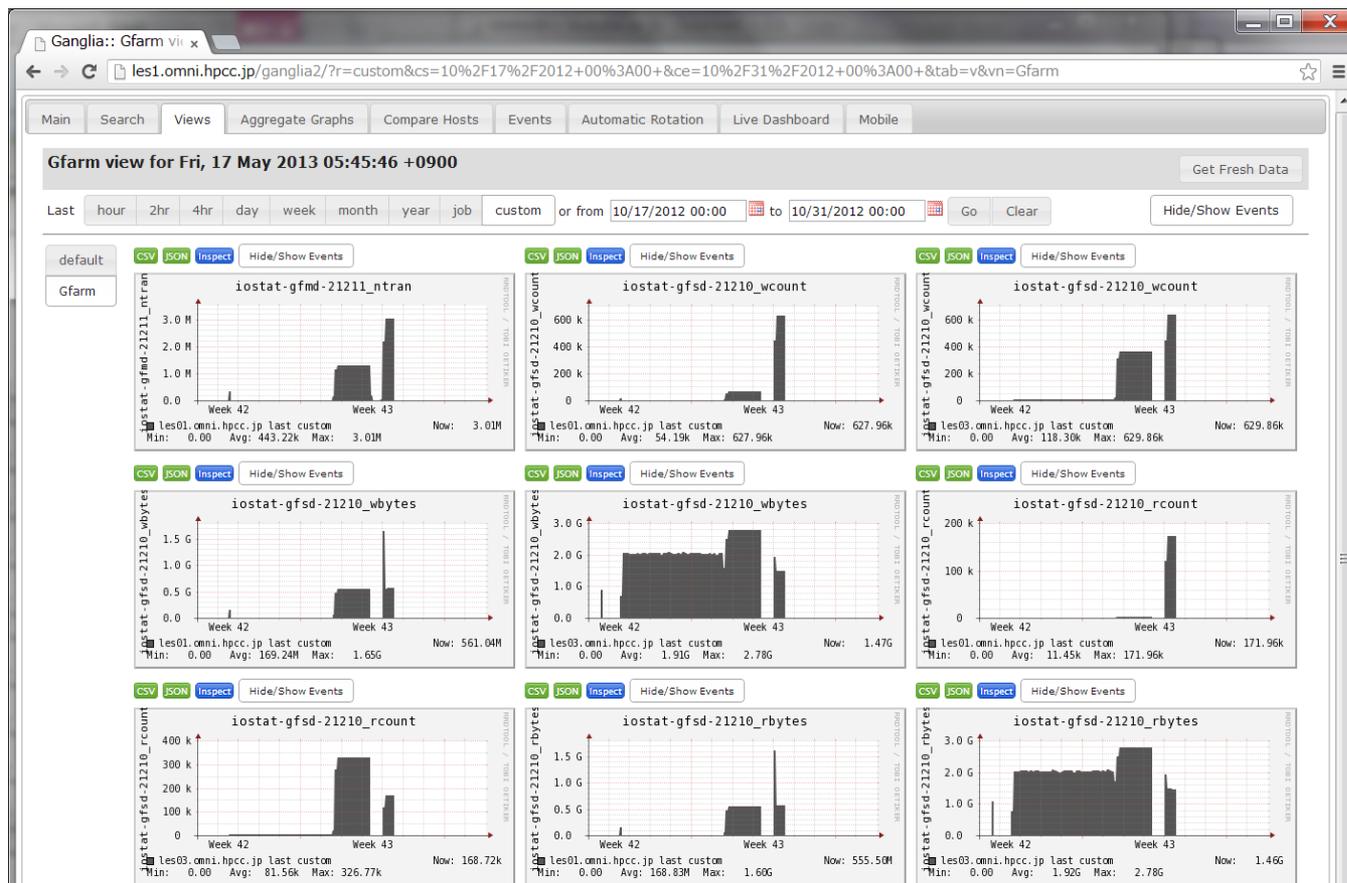
ホストグループ	致命的な障害	重度の障害	軽度の障害	警告	情報	未分類
Gfarm v2 FileSystem	0	0	0	0	0	0

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

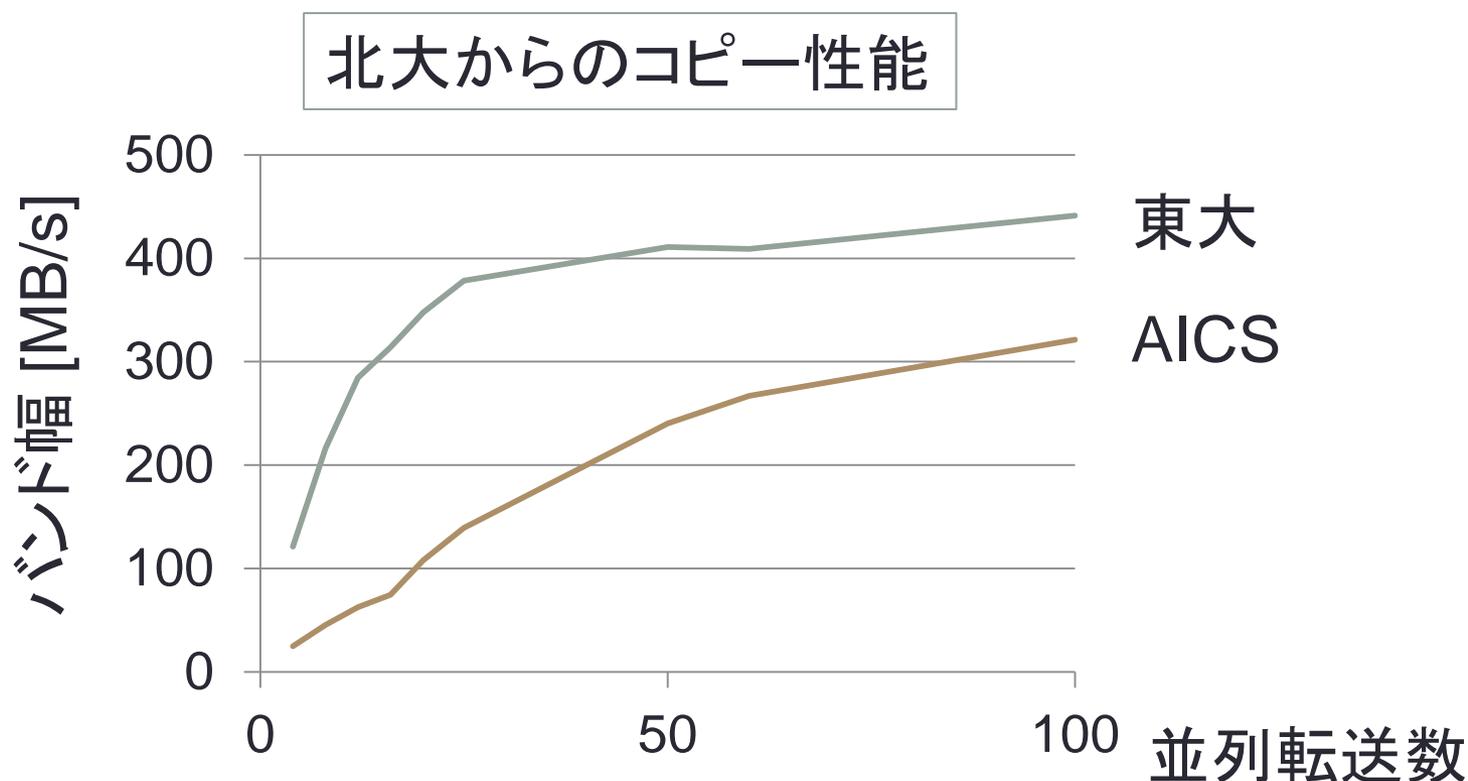
# 性能モニタリング [Gfarm 2.5.8]

- GangliaプラグインによるIOPS、バンド幅のリアルタイム性能モニタリング



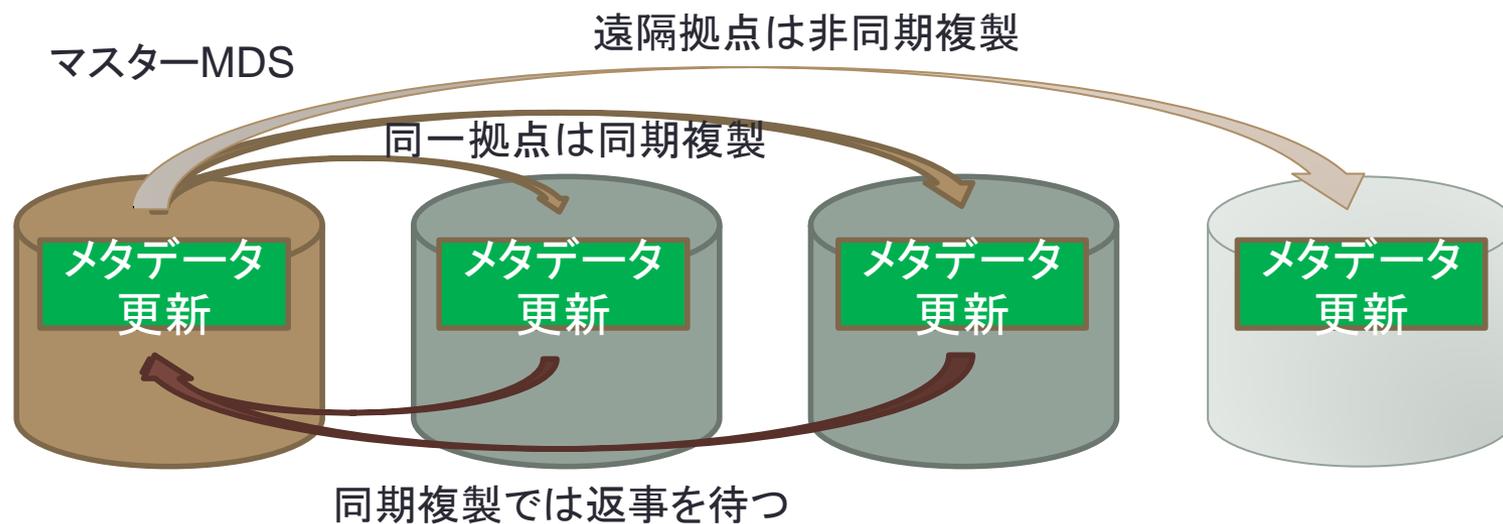
# 高速コピーコマンド [Gfarm 2.5.5]

- Gfpcopy – 多数ファイルを並列に転送することにより、遠距離からのファイルコピーを高速化



# ホットスタンバイMDS [Gfarm 2.5.0]

- マスターMDSにおいてメタデータ更新
  - スレーブMDSに転送
  - ジャーナルファイルに保持
- 同期複製の場合は、スレーブMDSからの返事を待つ
- 非同期複製はディザスタリカバリのため



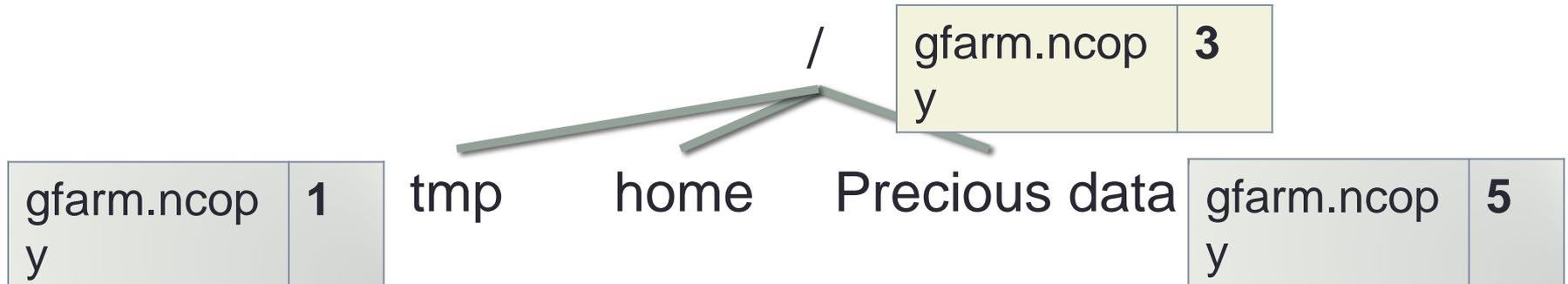
# 拡張ACL [Gfarm 2.4.2]

- POSIX 1003.1e DRAFT 17をベース
- 所有者、グループ、otherだけではなく、特定ユーザ、特定グループでrwxを指定可

# ファイル自動複製機能 [Gfarm 2.5.8]

- (祖先の)ディレクトリの拡張属性で複製数を指定
- Close時、更新時に自動的にファイル複製を作成

% **gfncopy** -s **3** /



# クォータによる利用制限 [Gfarm 2.3.1]

- See [doc/quota.en](http://doc/quota.en)
- 管理者 (gfarmadm) が設定可能
- ユーザ, グループごと
  - 利用容量, ファイル数の制限
  - ファイルによる制限と複製も考慮した物理制限
  - ハードリミットと猶予期間のあるソフトリミット
- ファイルオープン時にチェック
  - 注意: 越えたら作成できないが, 既にオープンしているファイルは容量制限を超えることが可能

# XML拡張属性 [Gfarm 2.3.0]

- 通常の拡張属性に加え, XMLをvalueとする  
% **gfxattr** **-x** -s -f value.xml filename xmlattr
- Xpathによるdirectory treeのXML拡張属性の検索  
% **gfindxmlattr** [-d depth] XPath path

# Samba VFS for Gfarm

- Gfarm2fsを利用しなくてもSambaからGfarmを利用するためのモジュール

# Debian packaging

- Squeezeのパッケージへの取り込み



The screenshot shows the Debian website interface for the source package 'gfarm' in the 'squeeze' release. The page title is 'ソースパッケージ: gfarm (2.3.0-5)'. It lists several binary packages built from this source: 'gfarm-client', 'gfarm-doc', 'gfmd', 'gfsd', 'libgfarm-dev', and 'libgfarm0'. A legend at the bottom indicates that packages with a red circle icon are architecture-dependent, while those with a green diamond icon are architecture-independent. On the right side, there are sections for 'gfarm に関するリンク', 'Debian の資源', 'メンテナ', and '外部の資源'.

debian

検索 ソースパッケージ名  [すべてのオ](#)  
[ブション](#)

>> Debian >> パッケージ >> squeeze (testing) >> ソース >> net >> gfarm [ squeeze ] [ sid ]

## ソースパッケージ: gfarm (2.3.0-5)

以下のバイナリパッケージがこのソースパッケージからビルドされています。

- [gfarm-client](#)  
Gfarm clients
- [gfarm-doc](#)  
Documentation for the Gfarm filesystem
- [gfmd](#)  
Gfarm metadata server
- [gfsd](#)  
Gfarm filesystem daemon
- [libgfarm-dev](#)  
Development files for the Gfarm filesystem
- [libgfarm0](#)  
Runtime libraries for the Gfarm filesystem

### その他の gfarm 関連パッケージ

 構築依存  構築依存 (アーキテクチャ非依存)

### gfarm に関するリンク

Debian の資源:

- [バグ報告](#)
- [開発者情報 \(PTS\)](#)
- [Debian での変更履歴](#)
- [著作権ファイル](#)

メンテナ:

- [NIIBE Yutaka \(QA ページ\)](#)
- [Osamu Tatebe \(QA ページ\)](#)
- [Noriyuki SODA \(QA ページ\)](#)

外部の資源:

- [ホームページ \[datafarm.apgrid.org\]](#)

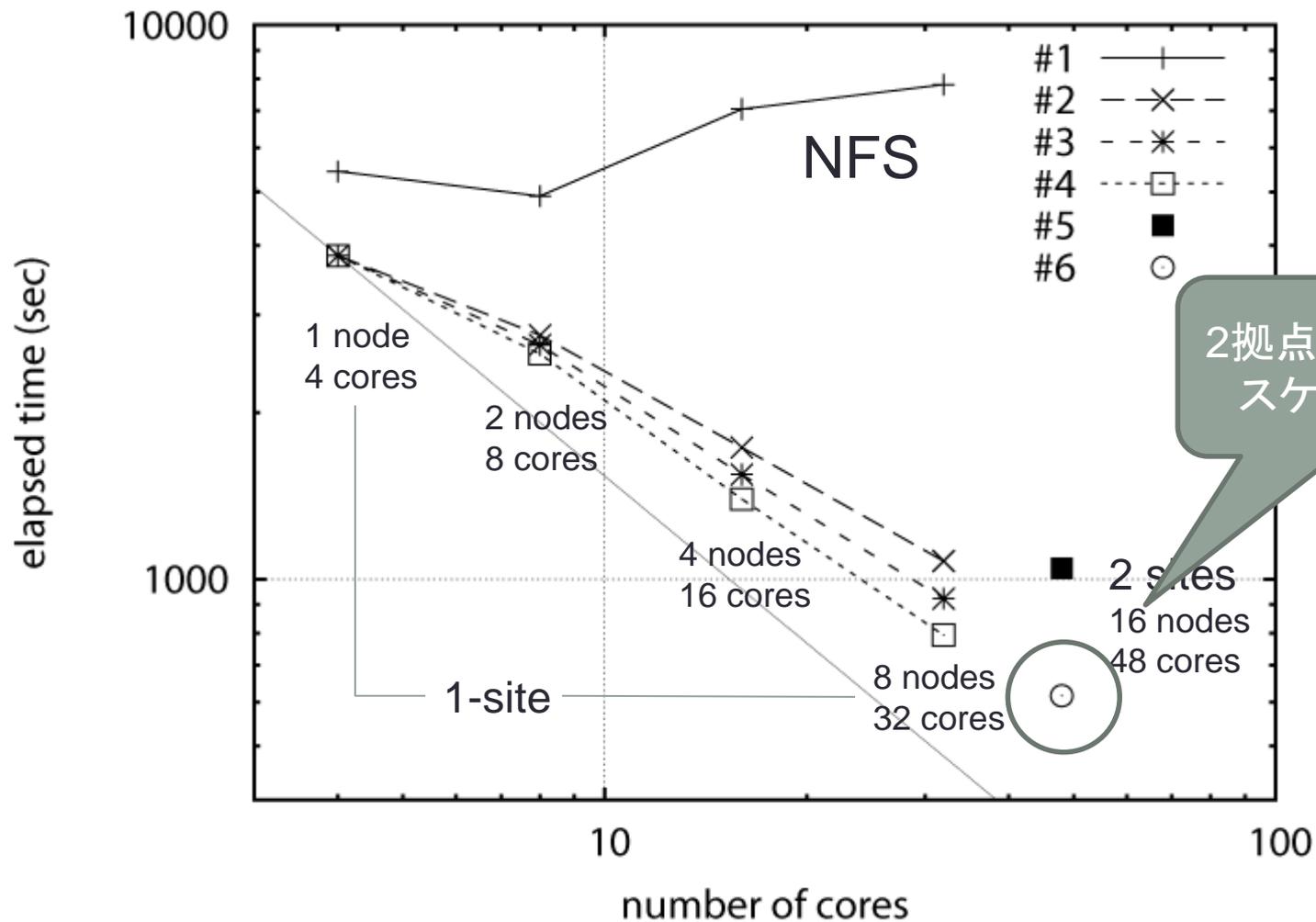
# 分散並列処理

---

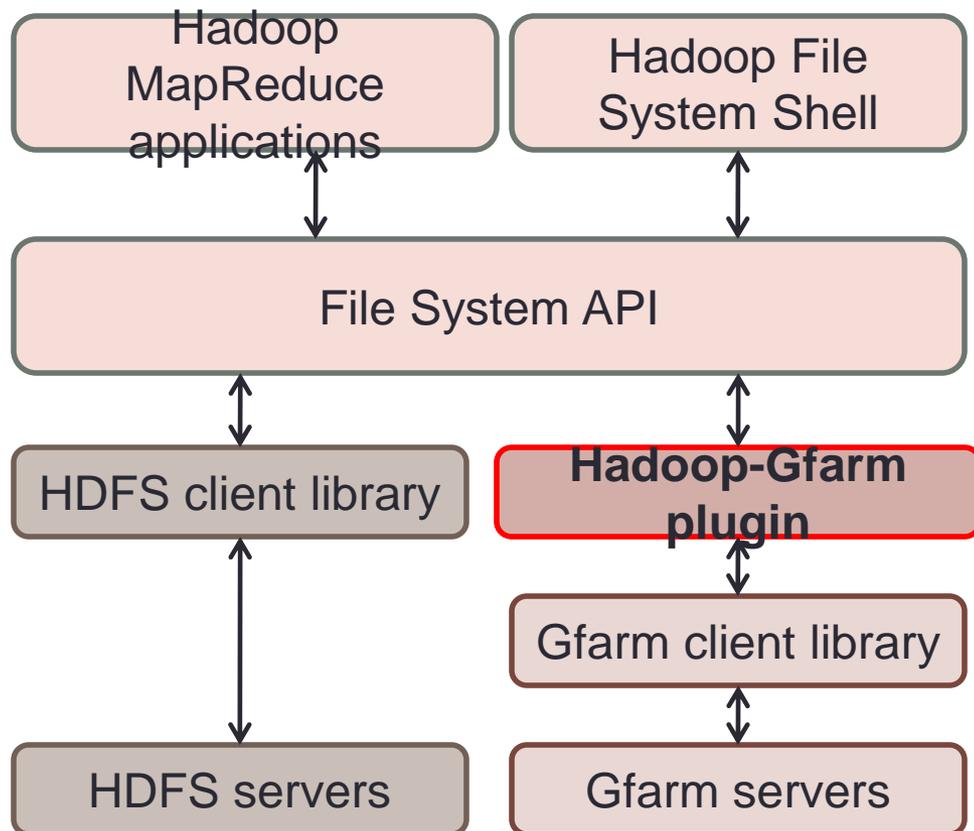
# Pwrakeワークフローエンジン

- Rakeを拡張。並列分散ワークフロー言語, 実行エンジンに
- <http://github.com/masa16/Pwrake/>
- Gfarmファイルシステムにおける拡張
  - 自動的にgfarm2fsでマウント, アンマウント
  - ファイルの所在を考慮したジョブスケジューリング
- 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)

# Montage天文ワークフローによる 性能評価



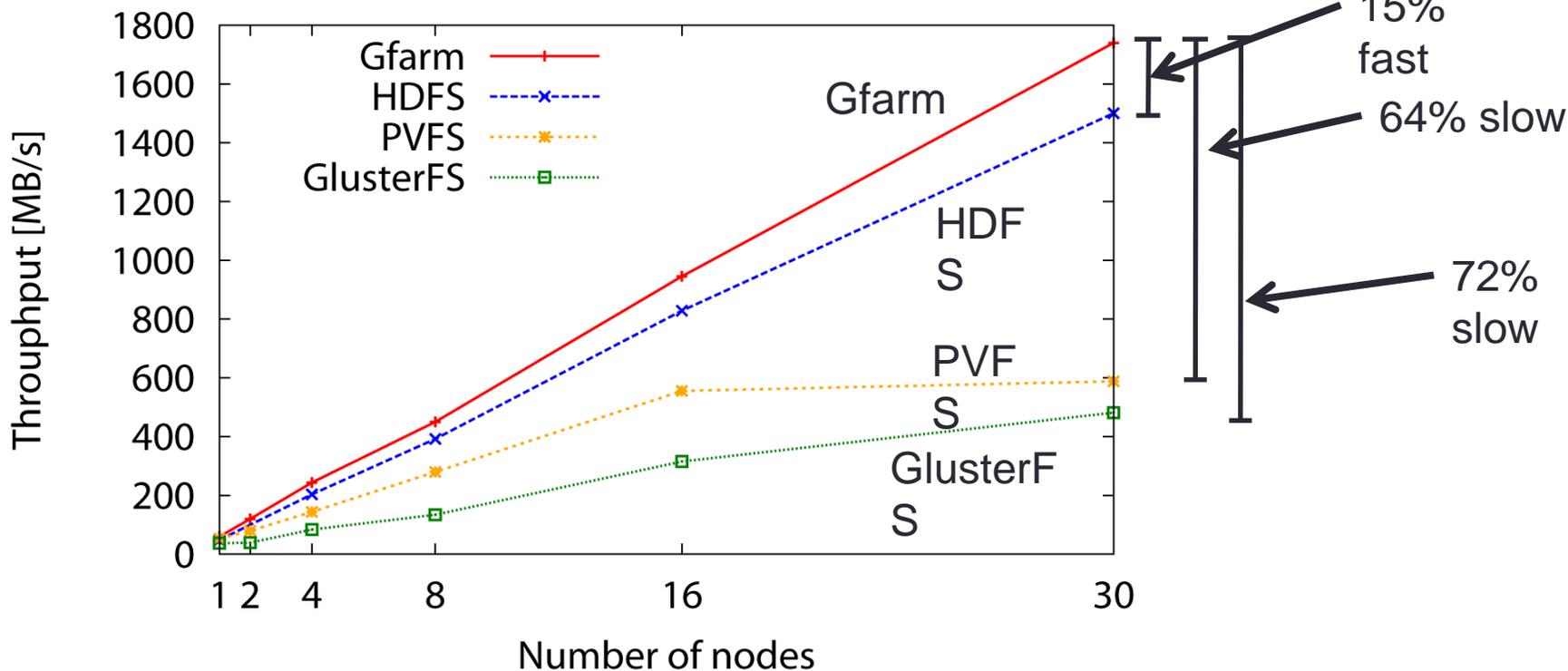
# Hadoop-Gfarmプラグイン



- Hadoop から **Gfarm URL** で Gfarm へアクセスするためのプラグイン
- <http://sf.net/projects/gfarm/>
- JNIによりHadoopからGfarmのクライアントライブラリを呼んでいる
- Hadoopアプリケーションは**ファイルの格納位置を考慮してスケジューリング**

# Hadoop MapReduceによるHDFSとGfarmの性能比較

書き込み性能



**HDFSを凌ぐ性能！**

# 今後の予定

- Gfarm 2.6.0を今秋～冬にリリース予定
  - 自動ファイル複製作成場所指定機能
  - フェイルオーバの高度化
- 広域分散超大規模データ処理
  - あらゆる分野のe-サイエンス(Data-Intensive Science)の促進

# サポート体制

---

# NPO法人設立

- 名称：特定非営利活動法人つくばOSS技術支援センター
- 所在地：茨城県つくば市
- 役員：
  - 理事長：建部 修見（筑波大学）
  - 理事3名＋監事1名
- 目的：
  - Gfarmを中核とするOSSの普及・促進
  - Gfarmを中核とするOSSのサポート
  - Gfarmコミュニティの運営

# 会員種別

- 正会員

- この法人の目的に賛同して入会した個人
- 年会費:1万円／口
- 議決権あり

- 賛助会員

- この法人の目的に賛同し、活動を支援するために入会した個人及び団体
- 法人賛助会員年会費:5万円／口
- 個人賛助会員年会費:5千円／口

# サポート料金

- オープンソースであるGfarmのサポート
- サポートへの加入は法人賛助会員の資格が必須
  - ゴールド会員以上の資格
- サポートはチケット制: 1チケット5万円
- サポートチケット販売方法:
  - 次の種類から選択して購入して頂けます。単一チケットの販売はございません。
    - 4チケット 20万円
    - 12チケット 50万円
    - 25チケット 100万円
    - チケットは購入時から1年間有効
- サポート加入のメリット
  - 迅速な対応、応答時間が設定されています。
  - クローズまでの対応
  - Webによる24時間受け付け

# NPO設立シンポジウム

- 日時:2013年9月19日木曜日午後1:30より
- 場所:東京 赤坂見附 SRAグループ本社ビル
- 受付登録:<http://kokucheese.com/event/index/107013/>
- プログラム
  - 13:30 - 13:45 NPO設立の経緯と今後の期待
    - 監事 高杉 英利
  - 13:45 - 14:25 講演1「NICTサイエンスクラウドとPwroke/Gfarmによるビッグデータ処理」
    - 村田健史(情報通信研究機構)
  - 14:25 - 15:05 講演2「HPCI共用ストレージにおけるGfarmの運用と性能」
    - 原田浩(東京大学)
  - 15:05 - 15:20 休憩
  - 15:20 - 15:50 GfarmインストールHOWTO
    - 江波均(株式会社ベストシステムズ)
  - 15:50 - 16:20 GfarmとNPOの活動
    - 建部 修見(筑波大学)
  - 16:20 - 17:30 パネルディスカッション「NPOへの期待」
    - モデレータ: 西
    - パネリスト: 村田、原田、大野木、藤波、建部
  - 17:30 - 閉会
  - 18:00 - 20:00 懇親会(会場別)

# お問い合わせ

Office[at]oss-tsukuba.org  
<http://www.oss-tsukuba.org>