Internet Week 2011 仮想化DAY

-組み合わせで作るクラウドシステム-

最新技術動向 GlusterFS

2011/12/1

(株)NTTPCコミュニケーションズ

1. 発表者の紹介
2. 分散ファイルシステムとは
3. GlusterFS概論
4. GlusterFSの最新動向
5. GlusterFSの今後(を占う)
6. まとめ
7. 参考

1. 発表者の紹介
2. 分散ファイルシステムとは
3. GlusterFS概論
4. GlusterFSの最新動向
5. GlusterFSの今後(を占う)
6. まとめ
7. 参考

高橋　敬祐 (TAKAHASHI Keisuke)

(個人用)Twitter ID : @keithseahus 

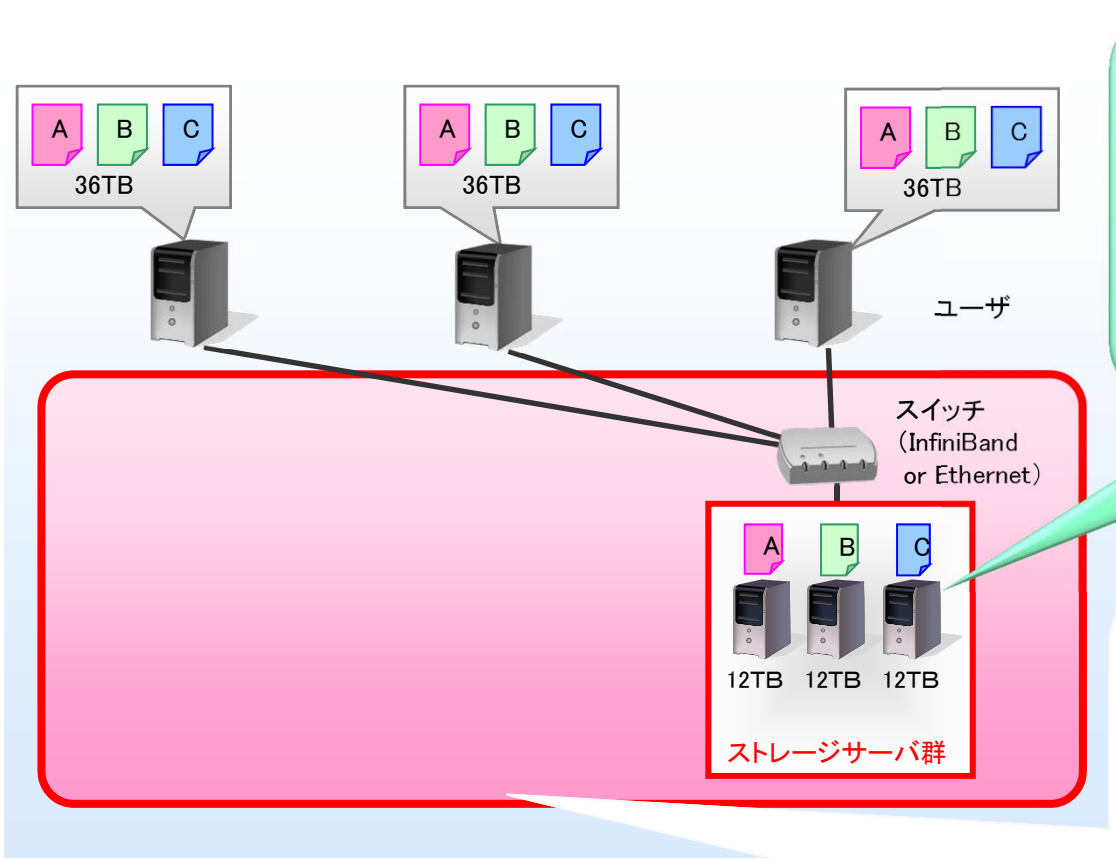
- NTTPCコミュニケーションズ在籍 (2006年～)
  - Webマイニング系開発
  - 分散FS特にGlusterFSの調査，研究，それを利用した開発及びOSS活動
  - 宇宙航空関連実証実験及び開発
  - その他先端技術調査

- この界隈で特に関係の深い人
  - 伊藤氏 (@thatsdone) 
  - STEC 伊藤氏 (@HarrisonIto) 
  - NTTPC 中冨 (@nakacya) 
  - NTTPC 高田 (@mikiT_T) 

- 出入りしている勉強会
  - VIOPS
  - クラウドネットワーク研究会
  - クラウドストレージ研究会
  - Tokyo Erlang Workshop
  - その他

1. 発表者の紹介
2. 分散ファイルシステムとは
3. GlusterFS概論
4. GlusterFSの最新動向
5. GlusterFSの今後(を占う)
6. まとめ
7. 参考

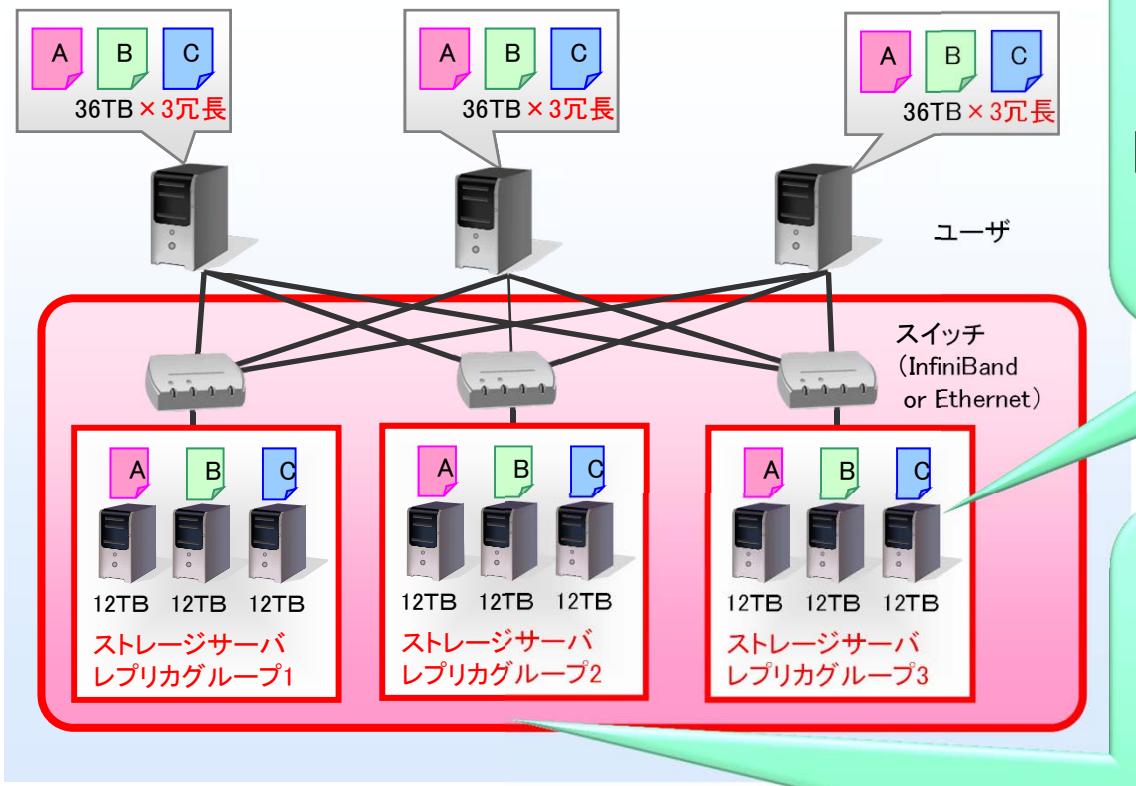
アーキテクチャ



単一のファイルシステムが、複数の物理コンピュータノードに分散して存在する。

レプリケーションやフォールトトラレンスなど、分散FS特有の機能を提供するものが多い。

アーキテクチャ



単一のファイルシステムが、複数の物理コンピュータノードに分散して存在する。

レプリケーションやフォールトトラレンスなど、分散FS特有の機能を提供するものが多い。

ストレージとしての分類と適用領域

DAS
- JBOD
- SCSIストレージ
- eSATA

SAN
- FC-SAN
- IP-SAN
- iSCSI

分散FS
RedHat GFS, Ceph, Sheepdog, ZFS, NetApp WAFL, HP LeftHandなど

NAS
- NFS
- CIFS

分散FS
Coda, Lustre, PVFS, OpenAFS, Gfarm, GlusterFS, Microsoft DFS など

オブジェクトストレージ

分散FS Google File System, HDFS, MogileFS, Amazon S3 (on Dynamo), Swift, kumofsなど

火付け役となったGoogle File System

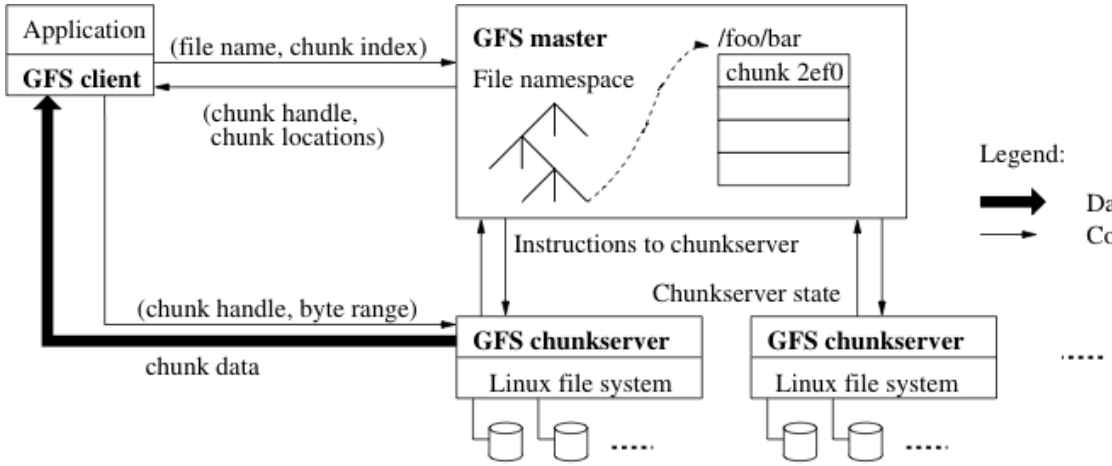


Figure 1: GFS Architecture

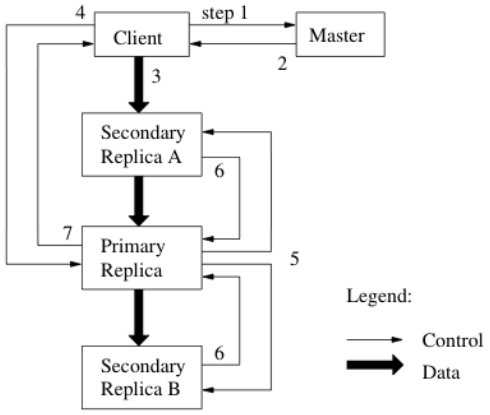


Figure 2: Write Control and Data Flow

- 2003年10月 19th ACM Symposiumにて発表。
- 2007年頃 論文がインターネット公開されて話題に。
- 追記型，チャンク，レプリケーション，MapReduceへの最適化が特徴。
- Hadoop HDFSにこれに近いものが実装される。

出典: Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, Google The Google File System, *SOSP'03, October 19–22, 2003, Bolton Landing, New York, USA.*

分散FSの変遷：クラウド以前

DAS

SAN

分散FS
RedHat GFS, Ceph, Sheepdog, ZFS, NetApp WAFL, HP LeftHandなど

NAS
実装しやすく
扱いやすい
NASが主流。

分散FS
Coda, Lustre, PVFS, OpenAFS, Gfarm, GlusterFS, Microsoft DFS など

オブジェクトストレージ

分散FS Google File System, HDFS, MogileFS, Amazon S3 (on Dynamo), Swift, kumofsなど

分散FSの変遷：クラウド時代

DAS

SAN

データセンタ事業者向けにSANの利用が増加。

分散FS
RedHat GFS, Ceph, Sheepdog, ZFS, NetApp WAFL, HP LeftHandなど

NAS

プライベート・クラウド向けとして、SANだけでなくNASも根強い。

分散FS
Coda, Lustre, PVFS, OpenAFS, Gfarm, GlusterFS, Microsoft DFS など

パブリック・クラウド向けにオブジェクトストレージとしての実装が台頭。

オブジェクトストレージ

分散FS Google File System, HDFS, MogileFS, Amazon S3 (on Dynamo), Swift, kumofsなど

1. 発表者の紹介
2. 分散ファイルシステムとは
3. GlusterFS概論
4. GlusterFSの最新動向
5. GlusterFSの今後(を占う)
6. まとめ
7. 参考

分散FSの中での位置付け

DAS

SAN

分散FS
RedHat GFS, Ceph, Sheepdog, ZFS, NetApp WAFL, HP LeftHandなど

NAS

GlusterFSは、NASとして利用可能なファイルストレージ。

分散FS
Coda, Lustre, PVFS, OpenAFS, Gfarm, GlusterFS, Microsoft DFSなど

オブジェクトストレージ

分散FS Google File System, HDFS, MogileFS, Amazon S3 (on Dynamo), Swift, kumofsなど

GlusterFS開発以前

Management

Hitesh Chellani, Acting CEO

Hitesh has been leading overall strategy and busi
with AB. Prior to Gluster he was at California Digit
and systems engineering. During his tenure there
deployed the 'Thunder' supercomputer at Lawren
the second fastest supercomputer in the world. Pri
Dubai UAE where he managed business develop
engineer for Tata Unisys in India. He holds a Bach

Anand Babu (AB) Periasamy, CTO

As CTO and Co-founder, AB sets the vision and st
CTO of California Digital Corporation, where his w
supercomputing class performance. He drove the
enterprise data centers and helped close strategic
world's second fastest Supercomputer code name
AB also serves on the board of "Free Software Fou
other Free Software projects like GNU FreeIPMI (I
(Gratuitous ARP Daemon), bios-config (edit/replic
GNU Hurd) and Hymn/PlayFair (iTunes ripper), GN
Google talk), and Freehoo (Scheme extensible me
Engineering degree from Annamalai University, T

TOP 10 Sites for June 2004

For more information about the sites and systems in the list, click on the links or view the complete list.

Rank	Site	Computer
1	The Earth Simulator Center Japan	Earth-Simulator NEC
2	Lawrence Livermore National Laboratory United States	Thunder - Intel Itanium2 Tiger4 1.4GHz - Quadrics California Digital Corporation
3	Los Alamos National Laboratory United State	ASCI Q - AlphaServer SC45, 1.25 GHz Hewlett Packard
4	IBM - Roche United State	
5	NCSA United State	
6	ECMWF United Kingd	
7	Institute of P (RIKEN) Japan	
8	IBM Thomas United State	
9	Pacific North United State	
10	Shanghai Su China	

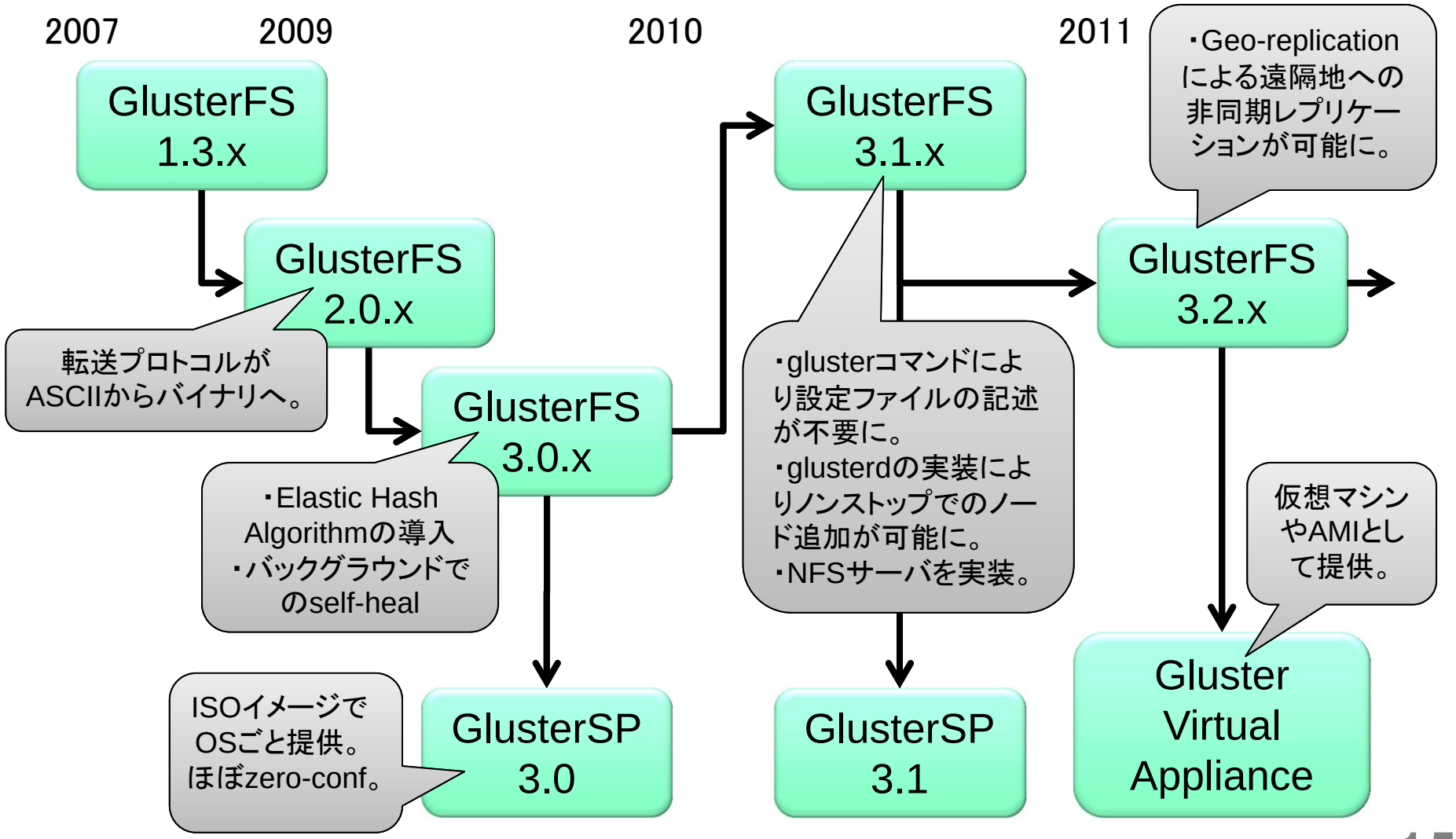
Home ▸ Sites ▸ Lawrence Livermore National Laboratory

Thunder

Details | Performance/Linpack Data | **Ranking History**

List	Rank	Rmax (GFlops)
11/2008	177	19940
06/2008	90	19940
11/2007	47	19940
06/2007	34	19940
11/2006	19	19940
06/2006	14	19940
11/2005	11	19940
06/2005	7	19940
11/2004	5	19940
06/2004	2	19940

GlusterFS 1.3～3.2



GlusterFSの仕様と要件

仕様

記述言語	C
ライセンス	GPL v3, AGPL v3

OS及びハードウェア要件

	サーバ	クライアント
OS要件	Linux (RHEL系5, 6推奨)	Linux, Windows(CIFS)
ハードウェア要件	64bit Intelアーキテクチャ	64bit Intelアーキテクチャ

ネットワーク要件

	サーバ～クライアント サーバ～サーバ	サーバ～遠隔サーバ (Geo-replication)
ギガビット・イーサネット	○	○
10ギガビット・イーサネット	○	○
InfiniBand RDMA	○	
InfiniBand IPoIB	○	○
インターネット		○

GlusterFS 10の特徴

1. ペタバイト・スケール
2. 機能ごとのモジュール化
3. ゼロ・シングルポイント障害
4. 選べるレプリケーション方式
5. NFSサーバ
6. 簡単・スピーディーな構築
7. 管理・監視コマンド
8. InfiniBand RDMA
9. 豊富な導入実績
10. クラウド基盤連携

ペタバイト・スケール

- 16エクサバイト/1ボリューム(ソースコードより)。

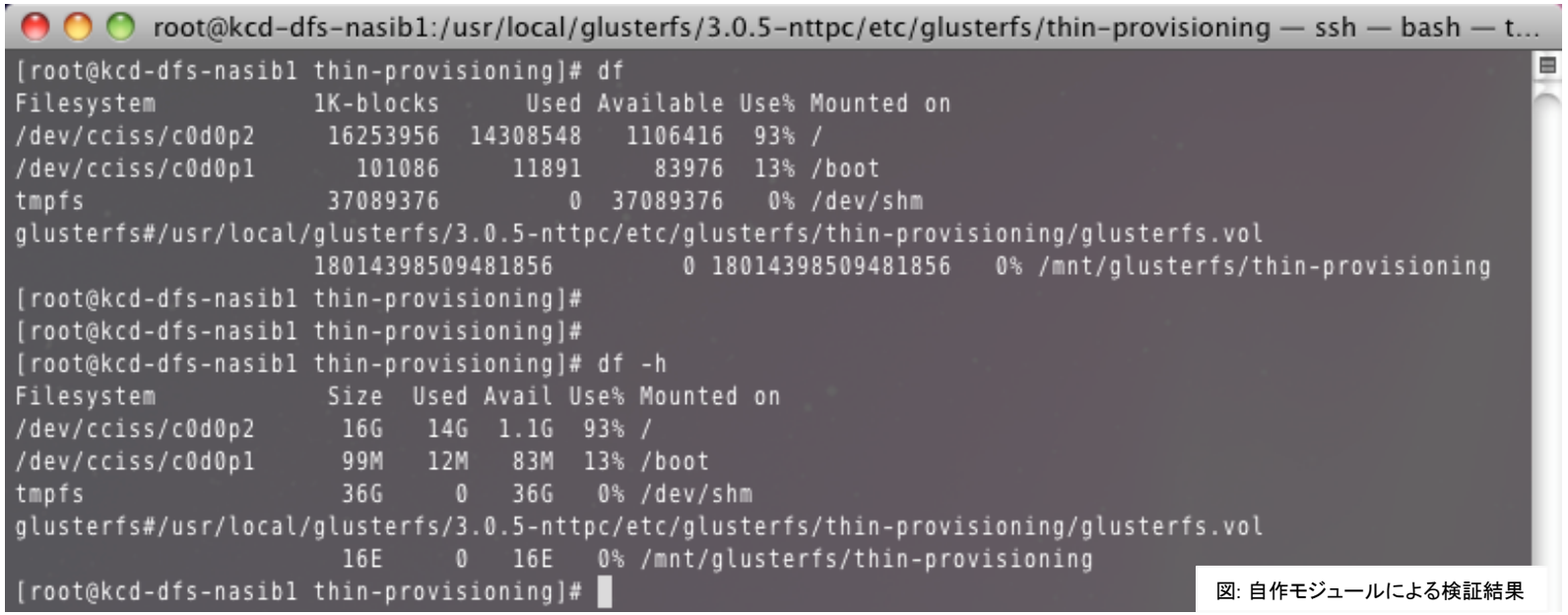


図: 自作モジュールによる検証結果

- 商用利用実績としては数ペタバイトの企業も。
- 無停止でノード追加(＝容量拡張)が可能。
- 扱えるファイル数は2^128個/1ボリューム

機能ごとのモジュール化

- GlusterFSの機能をモジュールに隠蔽。
- 機能追加が容易。
- モジュールの動的ローディングにより、構成変更が容易。
- バグの影響範囲を小さくすることができる。

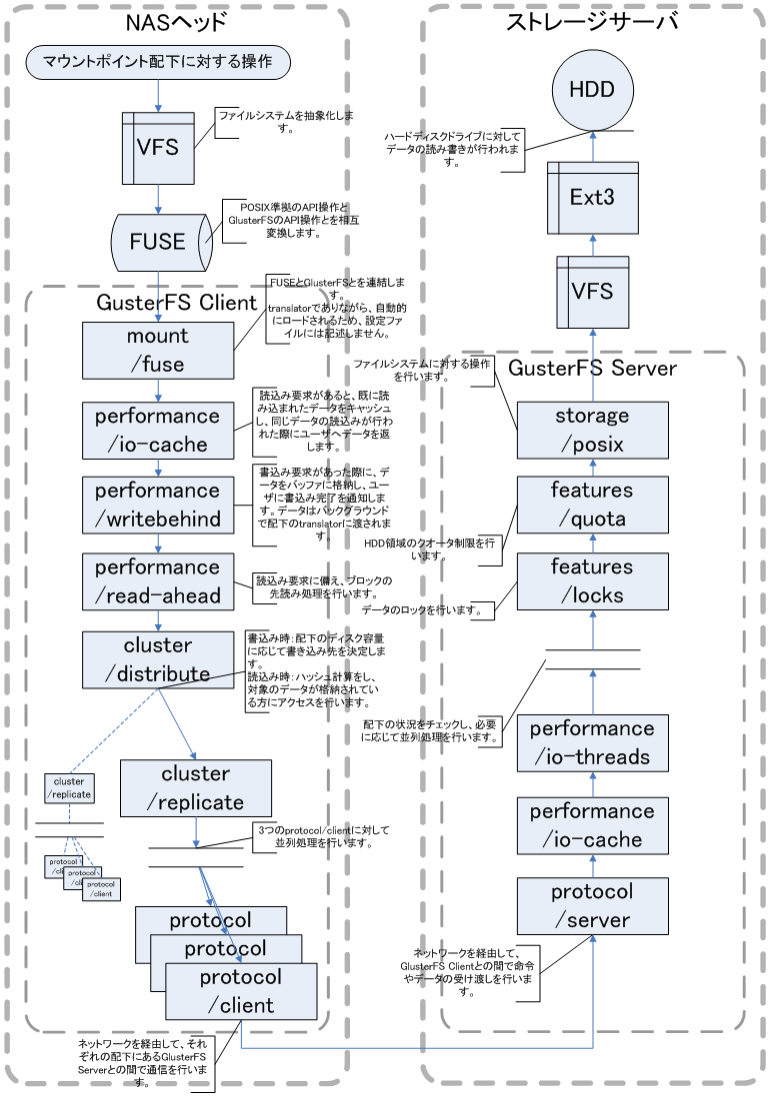
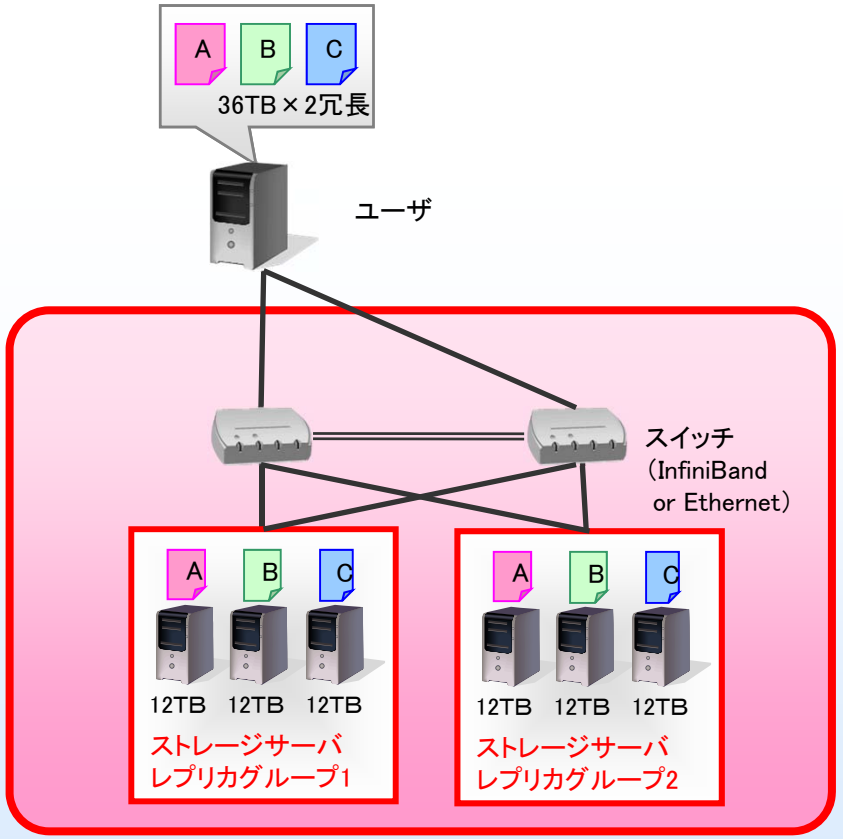


図: GlusterFS 2.0でのモジュール構成例

ゼロ・シングルポイント障害

- 多くの場合にシングルポイントとなる中央サーバが存在しない。
- シングルポイントが無い構成が取れる。
- レプリカ障害時のフェイルオーバも(当然)可能。



選べるレプリケーション方式

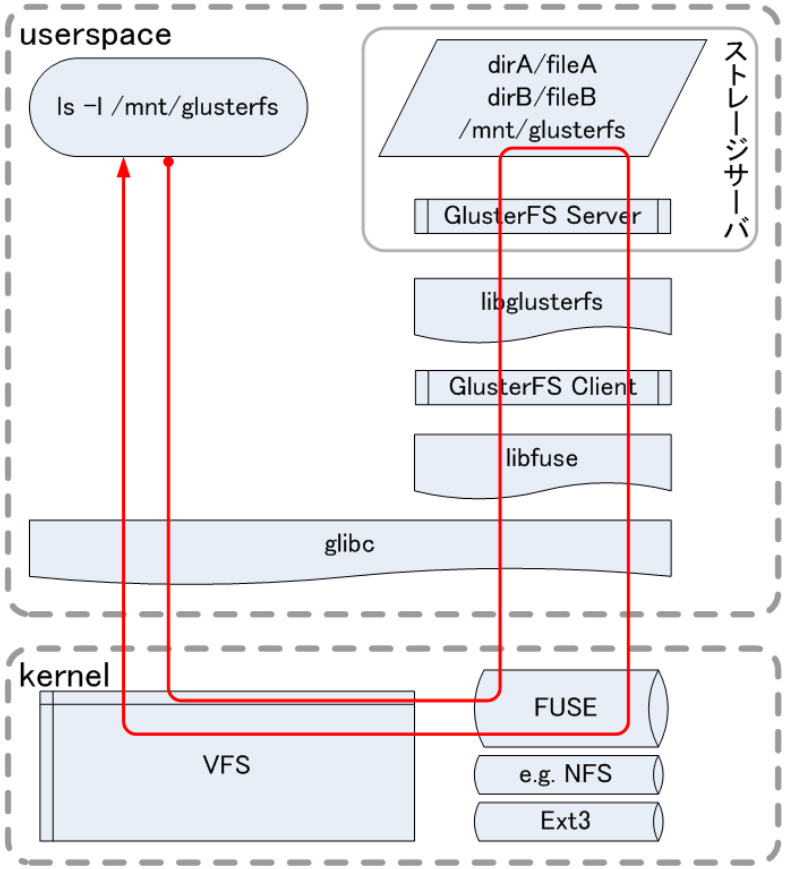
	同期 レプリケーション	非同期 レプリケーション
長所	確実なレプリケーション	帯域幅に影響されない
短所	十分な帯域が必要	確実性を保証できない
使用方法	ボリューム作成時に設定	Geo-replicationを設定

組み合わせて使用することも可能。

NFSサーバ

- GlusterFSが独自に実装。
- 直接GlusterFSプロトコルでストレージ側と通信するサーバ。
- FUSEを使用しない。
- クライアント側でのFUSE及びGlusterFSのインストールが不要。
- クライアントではカーネルのバッファ・キャッシュが有効に。

参考：GlusterFS FUSE コールシーケンス



簡単・スピーディーな構築

1. rpmのインストール
2. peerの登録
3. volumeの作成とスタート
4. クライアントからのマウント

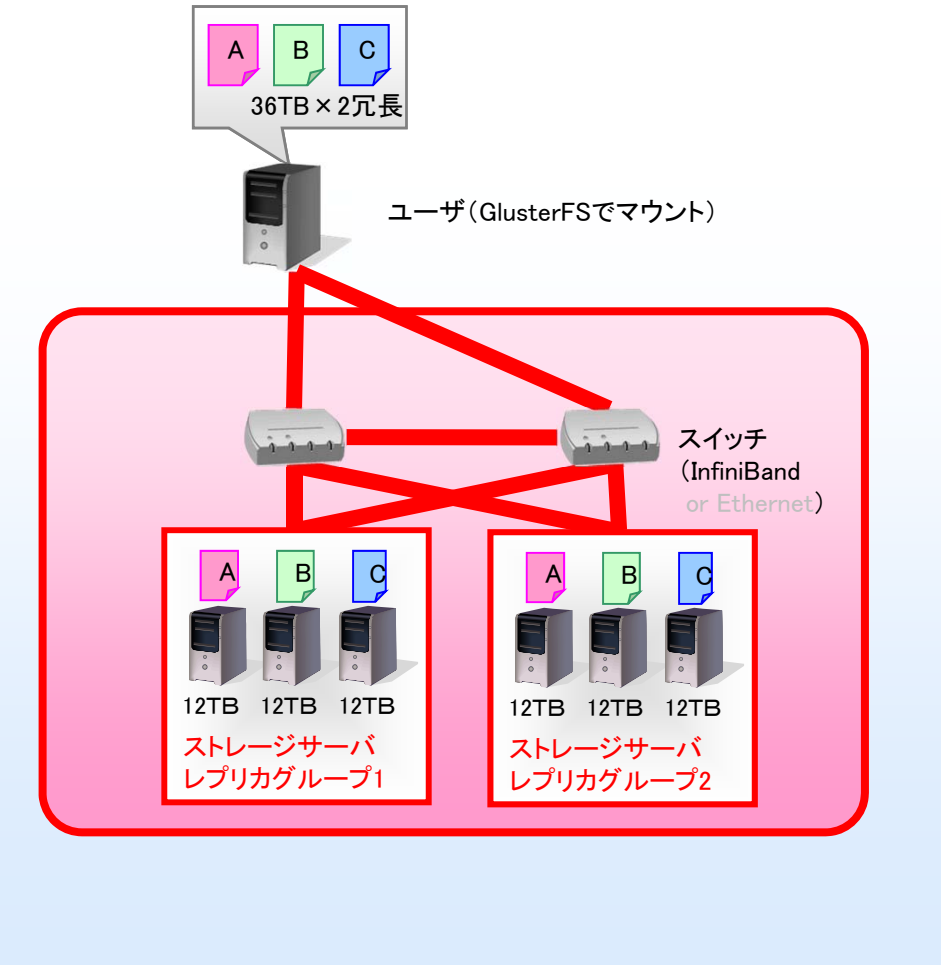
```
# rpmbuild -ta glusterfs-3.2.x.tar.gz --without rdma
# rpm -Uvh /usr/src/redhat/RPMS/x86_64/glusterfs-*
# for i in {peer1 peer2 peer3 peer4}; do gluster peer probe $i; done
# gluster volume create volname replica 2 peer1:/path/to/brick peer2:/path/to/brick peer3:/path/to/brick peer4:/path/to/brick
# gluster volume start volname
# mount -t nfs -o hard,intr,nosuid peer1:/volname /mnt/nfs
(# mount -t glusterfs localhost:/volname /mnt/glusterfs)
```

管理・監視コマンド

- 管理コマンド
  - gluster peer {probe|status|detach} ...
  - gluster volume {create|info|start|stop|delete|rename|set} ...
  - gluster volume {add-brick|remove-brick|replace-brick} ...
  - gluster volume rebalance ...
  - gluster volume log {filename|locate|rotate} ...
- 監視コマンド
  - gluster volume top ... {open|read|write} ...
  - gluster volume top ... {opendir|readdir} ...
  - gluster volume top ... {read-perf|write-perf} ...

InfiniBand RDMA

- NFSの場合は、NFSサーバ～ストレージ間。
- GlusterFSの場合は、クライアント～ストレージ間。
- TCPではなくRDMAを使用することで、低レイテンシを実現。

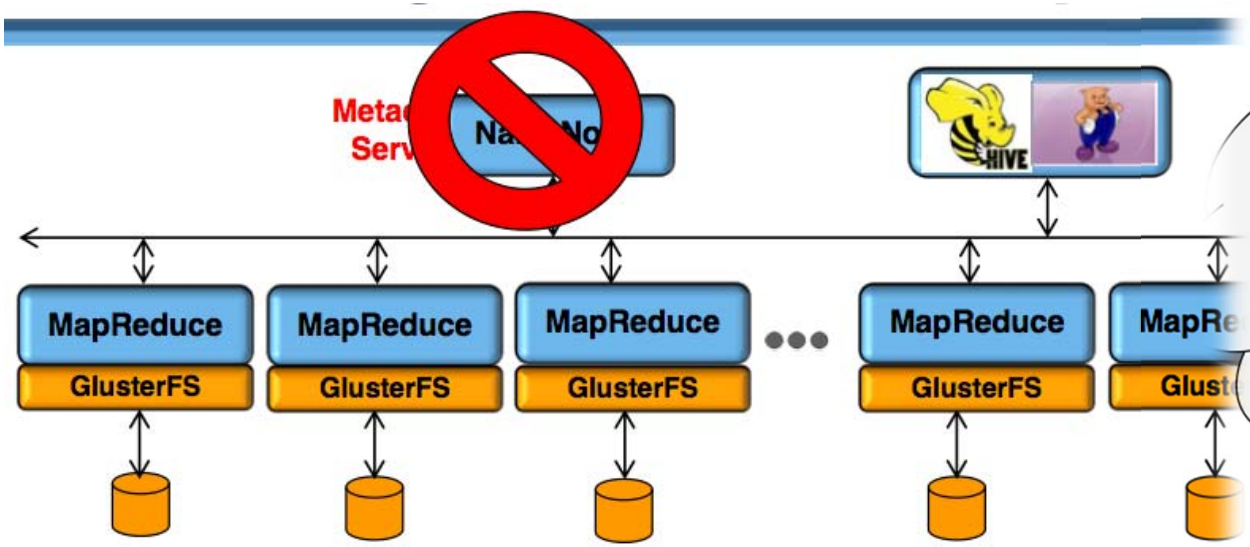


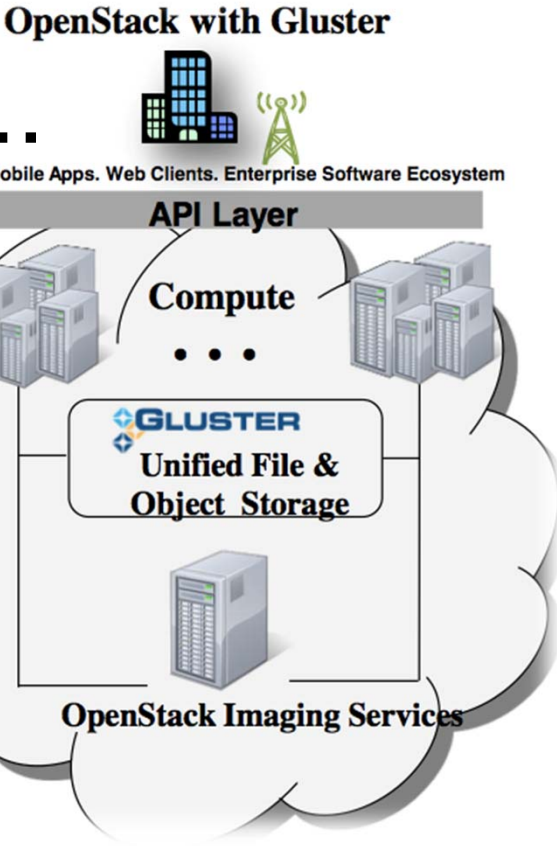
豊富な導入実績

- 音楽配信
- 動画配信
- 広告配信
- エンタープライズ向けクラウドソリューション
- 医療
- etc

クラウド基盤連携

- HadoopのHDFSの代替として…
  - HDFSとの共存も可能
- OpenStackのSwiftの代替として…





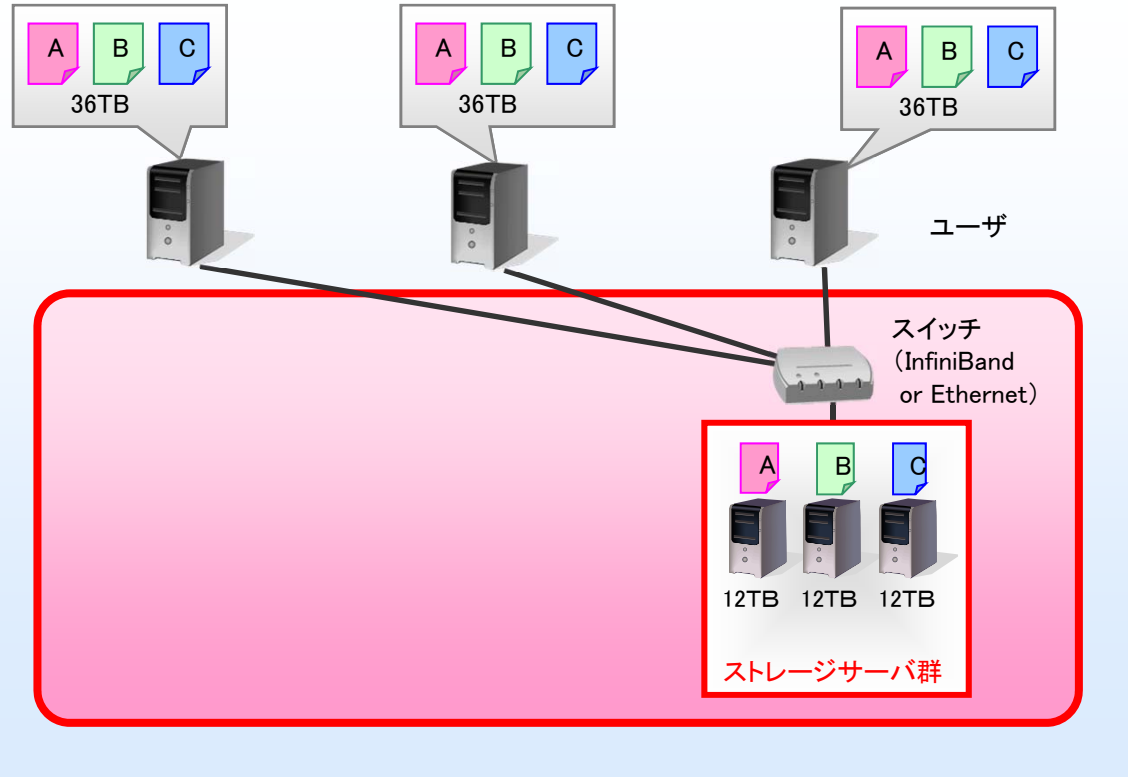
GlusterFS（又は分散FS）のメリット

- ソフトウェア開発者
  - データアクセス方式を選べる（POSIX準拠，RESTful API）。
  - 教育用にも適している。
  - モジュール開発のロマン。
- システム運用者
  - 導入や操作が簡単。
  - 筐体故障に強い。
  - 専用の監視コマンドがある。
  - サーバの追加による広帯域化を検討することができる。
- サービス企画者
  - スモールスタートが可能。
  - ベンダ・ロックインを回避できる。
  - 増設時点で最適なハードウェアを選ぶことができる。
  - クラウド上に構築することもできる（Gluster Virtual Appliance）。
  - NFSを使用することで責任分界点を明確にできる。
  - 必要に応じてDRサービスを追加することができる（Geo-replication）。
- 営業
  - ストレージのGB単価を大幅に下げられる。
  - 様々な導入実績を紹介することができる。
  - クラウド時代に合ったキャッチーな提案ができる。

GlusterFSの制約

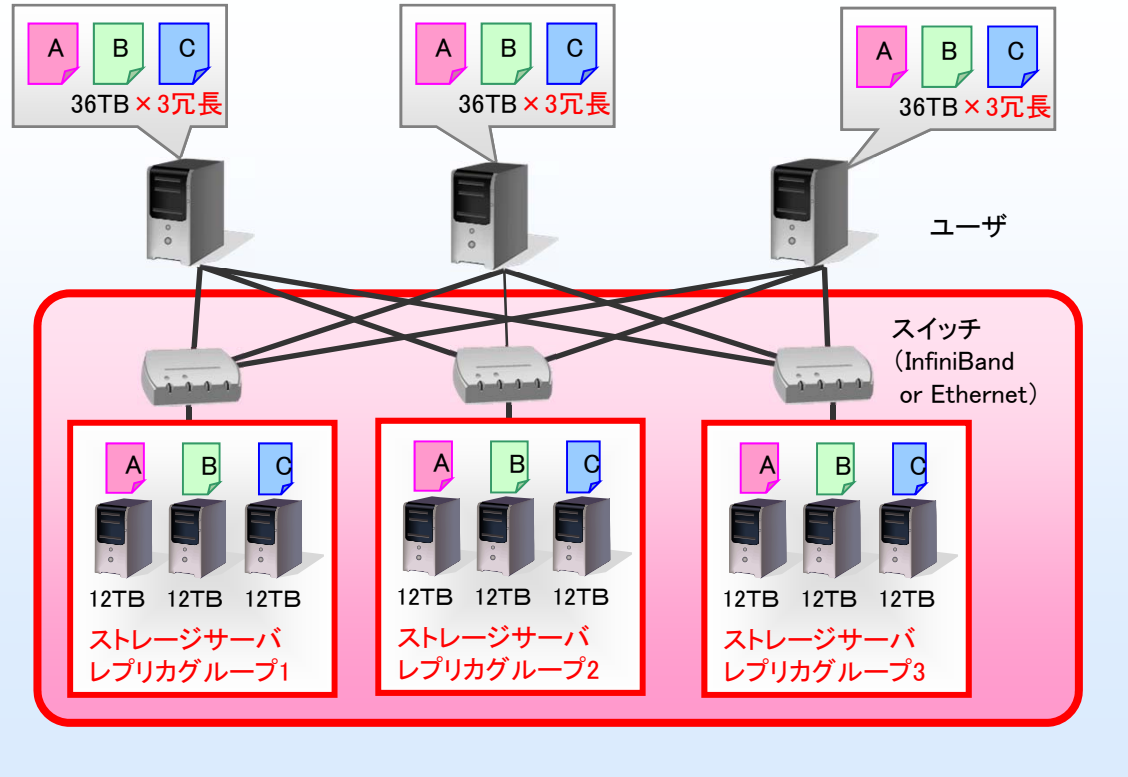
- 提供していない機能等（例）
  - マルチテナンシー（UID/GIDとパーミッション，Quotaはある）
  - シン・プロビジョニング
  - 重複排除
  - データ・コラプションのチェック
  - 暗号化（データ，経路※） ※Geo-replicationは経路暗号化対応
  - NFS ネットワーク・ロック・マネージャ (NLM)
  - ブロックストレージ
  - QoS
  - パリティベースの分散
  - 仕様書・設計書（笑）
- 特別な運用が必要な操作
  - ボリュームサイズ縮小
  - レプリカ数・ストライプ数の変更
  - インターコネクトのプロトコル変更（TCP <-> RDMA）
  - 既存クラスタの一部を別クラスタに移動（古い設定ファイルに注意）
- アーキテクチャ上の制約
  - Hash計算なので偏りは発生し得る
  - 特定操作によりシステム全体に負荷がかかる
  - replicaやstripeのグルーピングは固定的
  - Split Brain問題は残る
  - NAT越えはできない（※Geo-replicationを除く）
  - FUSE特有の制約（O_DIRECTやmmapが使えないなど。GlusterFSプロトコルの場合。）
  - RDMAはzero-copyではない
  - TCPポート数に起因する事実上のノード数制限

distribute構成



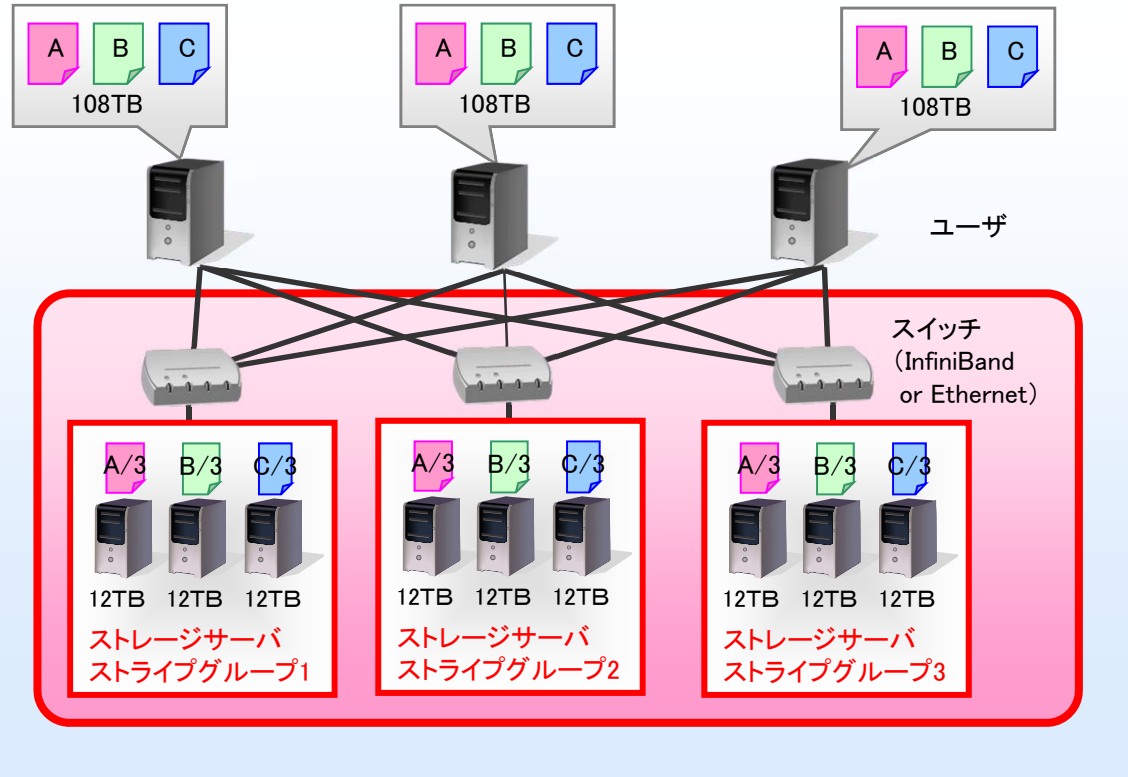
- 最もプリミティブな構成。
- キャッシュ等、容量が小さく、最悪消えても問題が無いようなデータ向き。
- SSD等を利用することでハードウェアの信頼性が高められれば、通常利用も可能かも？

distributed-replicate構成



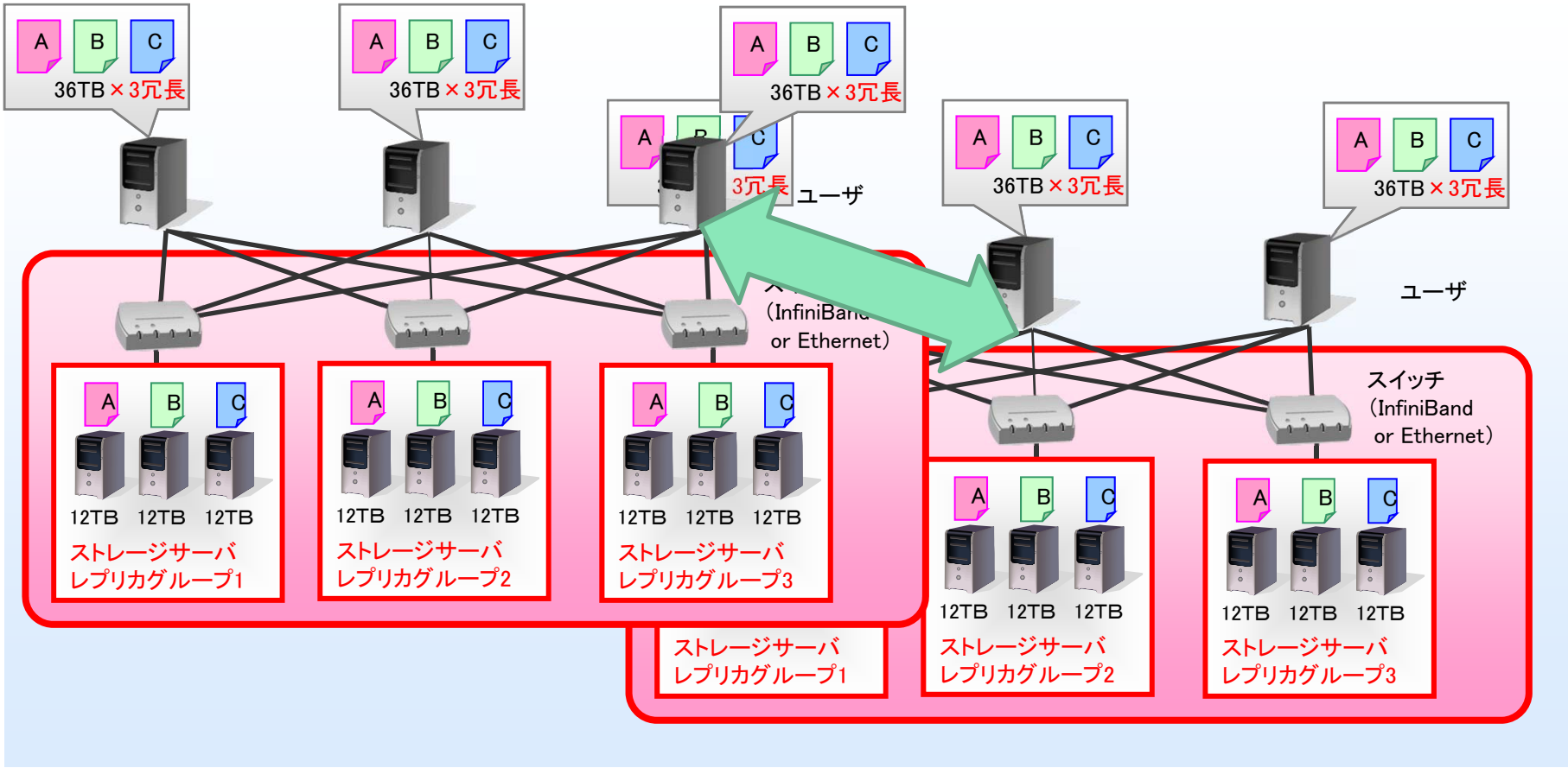
- 最もポピュラーな構成。
- 同期レプリケーションは、書き込み時にレプリカの数に比例して帯域を圧迫することに注意。
- また、レプリカの数に比例してハードウェア，ラック代，電気代がかかることにも注意。

distributed-stripe構成



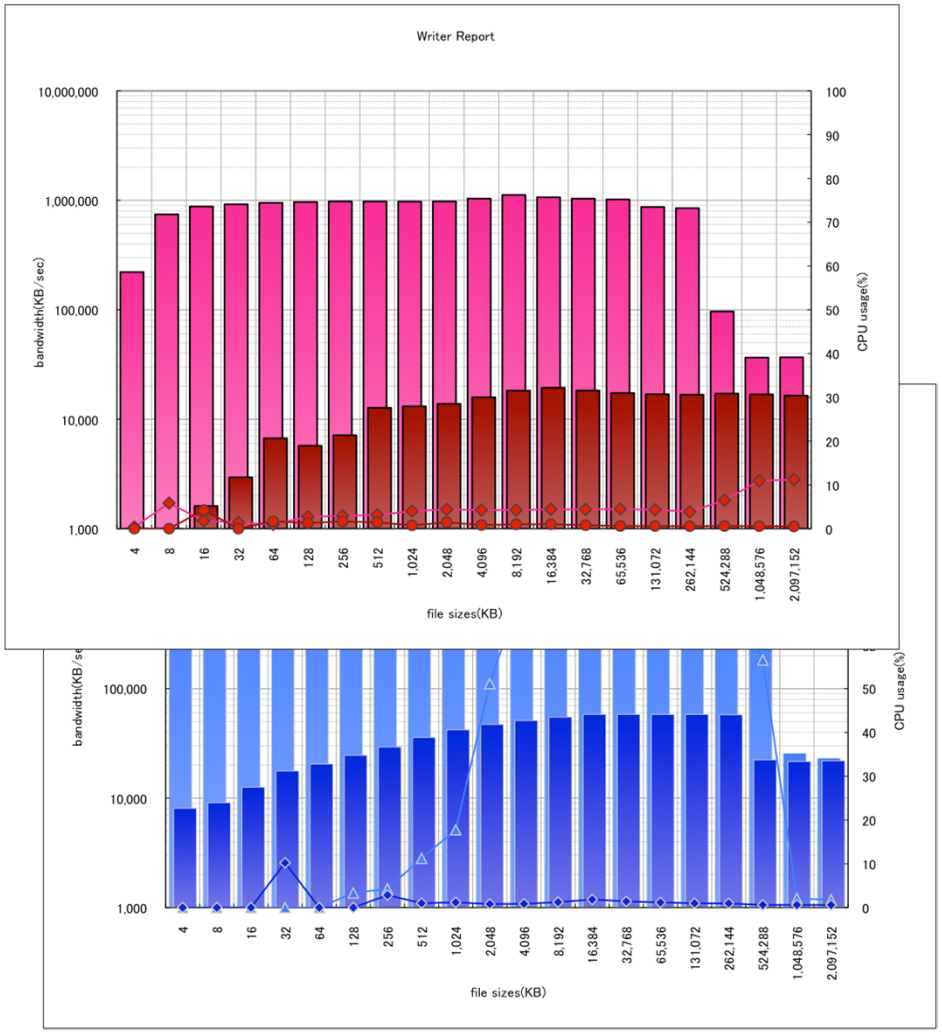
- 冗長性よりも速度を稼ぎたい場合に有効。
- しかし、大抵の場合は、コネクティビティにお金をかけた方がランニングコストが安く済むはず。
- RAID10構成は現状はとれない(3.0系までは出来た)。

distributed-replicate + Geo-replication構成



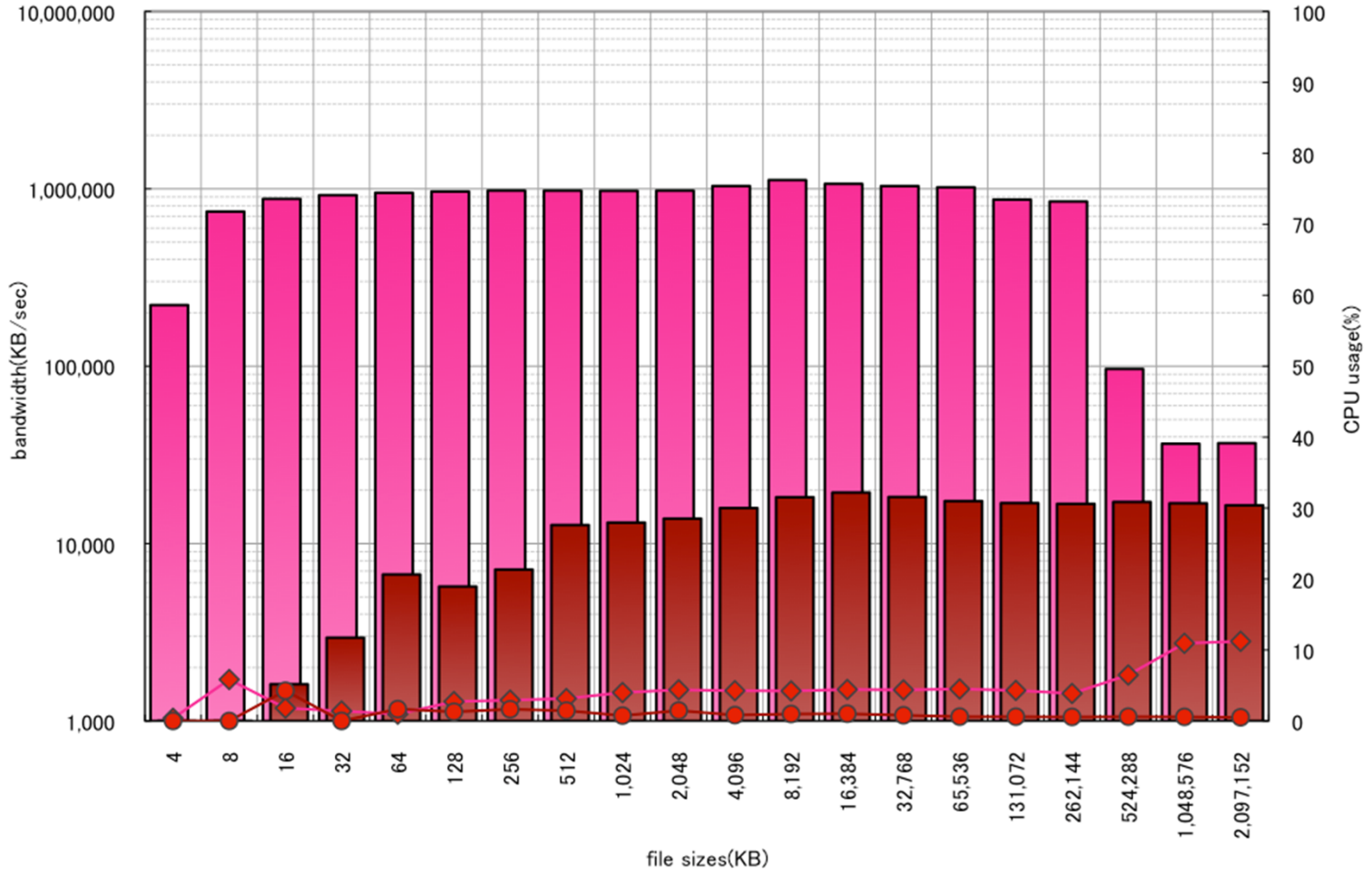
- 環境
  - distributed-replicate構成(replica 3)
  - ギガビット・イーサネット越しのNFSマウント
  - ネットワークはすべてギガビット・イーサネット
  - クライアント：HP DL360 G6
  - ストレージ：HP DL180 G5
- ベンチマーク
  - IOzone 3.326
  - ファイルサイズ：4KB～2GB
  - レコードサイズ：4KB～16MB
  - 2種類のベンチマーク
    - 特定オプション (-c -e -I -o -G -+r -+D) の有無
    - iozone -az -i 0 -I -2 ... -w -+w -+q 0 ... [特定オプション]
  - 各10回計測。
  - 各ファイルサイズごとの代表値を使用。本ベンチマークにおける代表値は、各レコードサイズの算術平均値（Bandwidth, CPU使用率）。なお、結果の分類及び有意水準の設定は行っていない。

グラフの説明

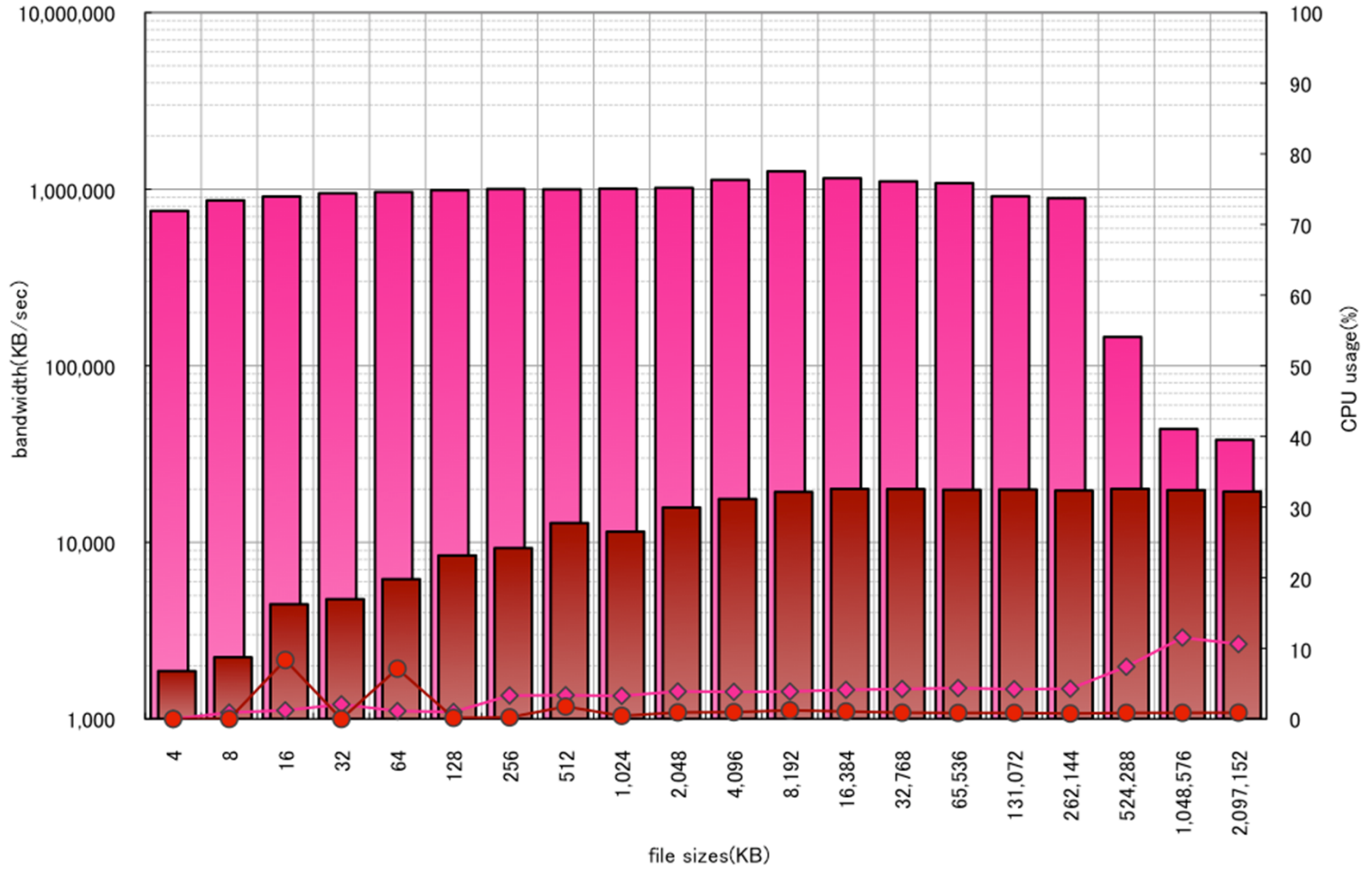


- 赤系は書き込み系。青系は読み込み系。
- 濃い色：特定オプションあり，淡い色：特定オプションなし
- 対数グラフ。
- x軸：ファイルサイズ(KBytes)，4KB～2GB
- 左y軸：Bandwidth (KBytes/sec)，棒グラフ，1MB/sec～10GB/sec
- 右y軸：CPU使用率(%)，折れ線グラフ，0%～100%

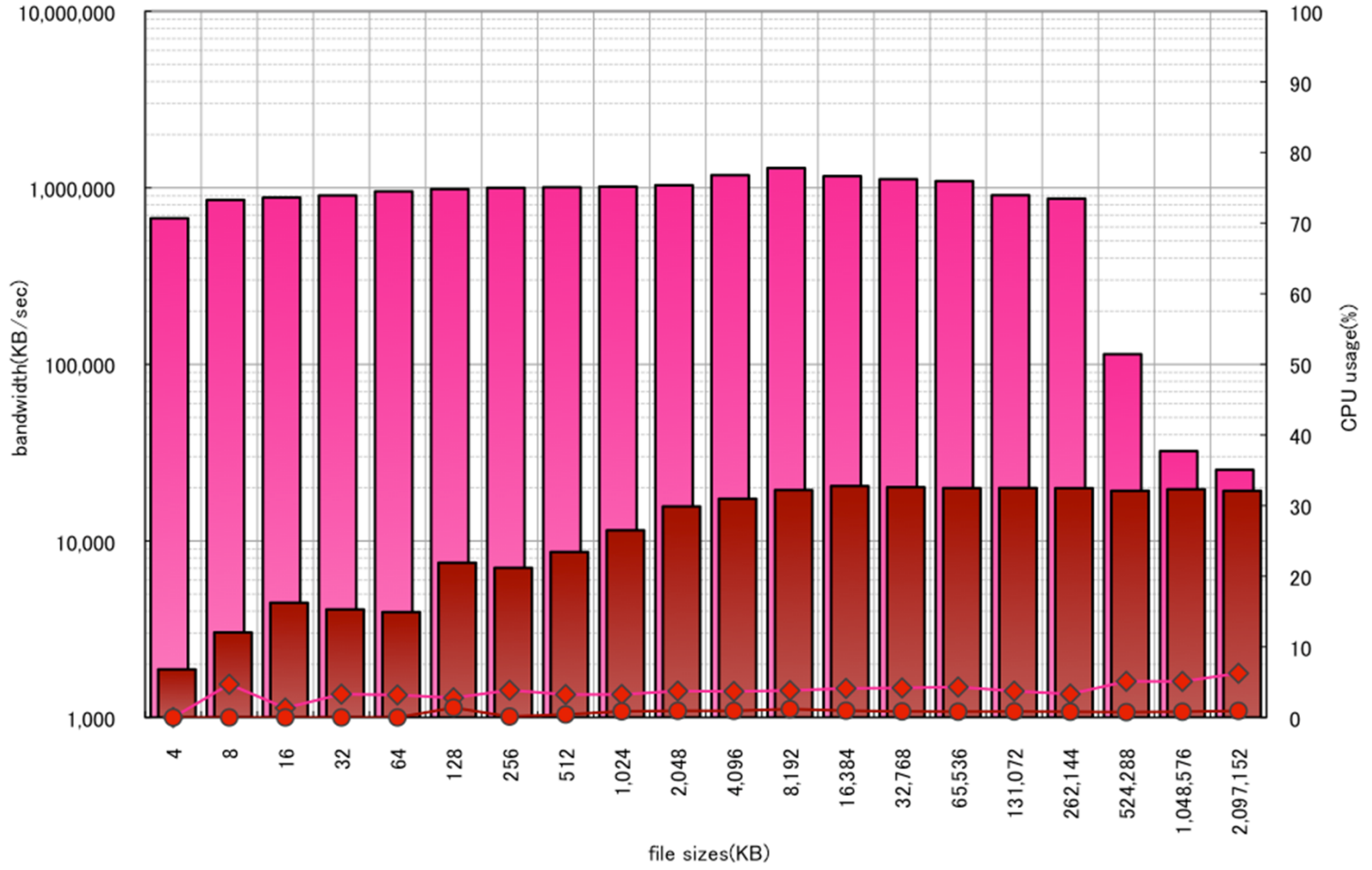
シーケンシャル書き込み



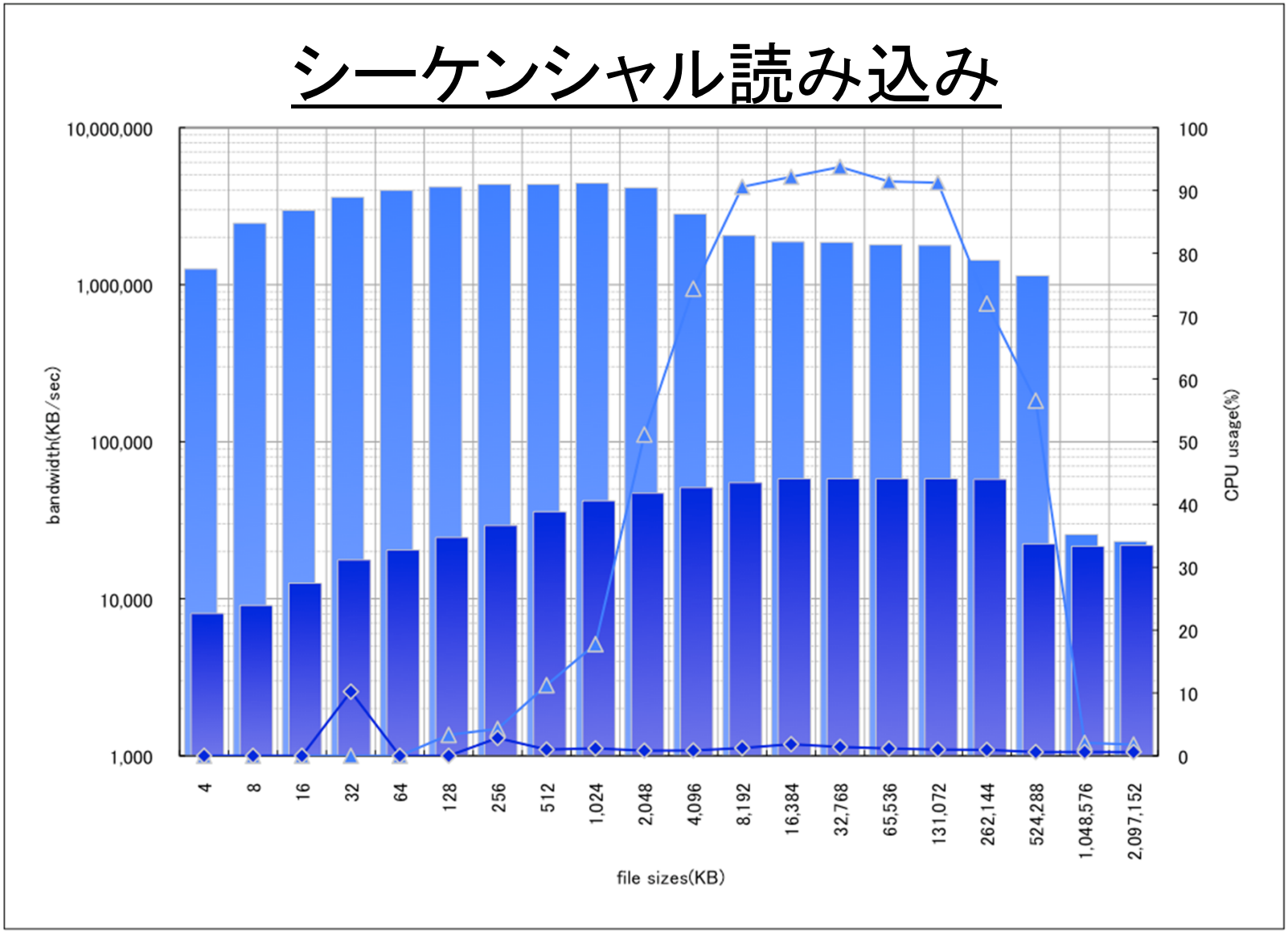
シーケンシャル上書き



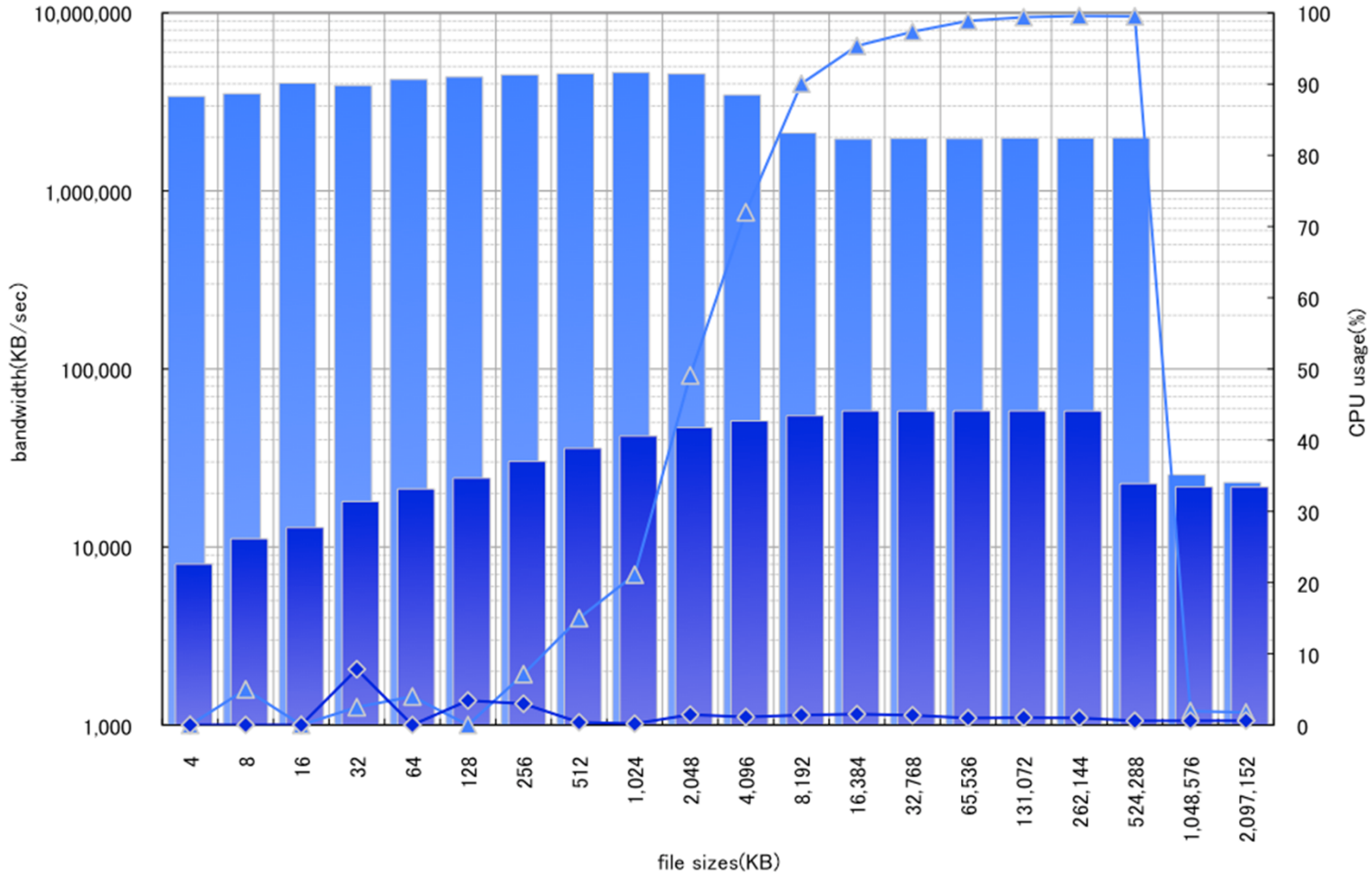
ランダム書き込み



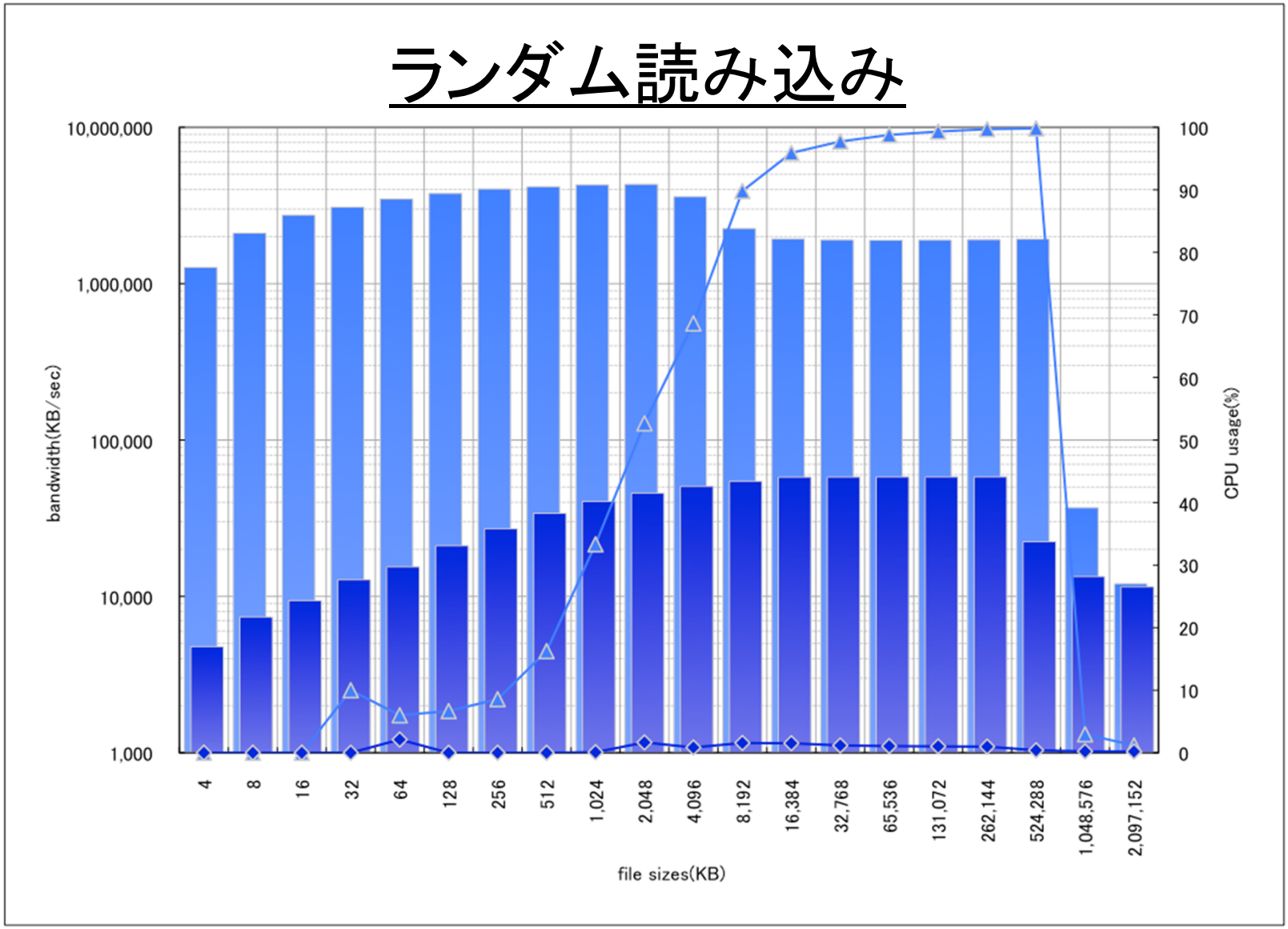
シーケンシャル読み込み



シーケンシャル再読み込み



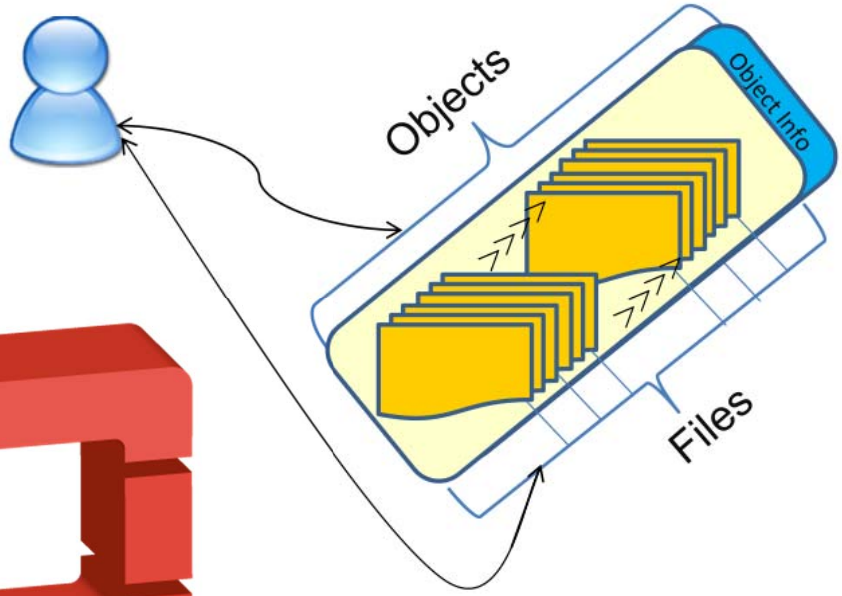
ランダム読み込み



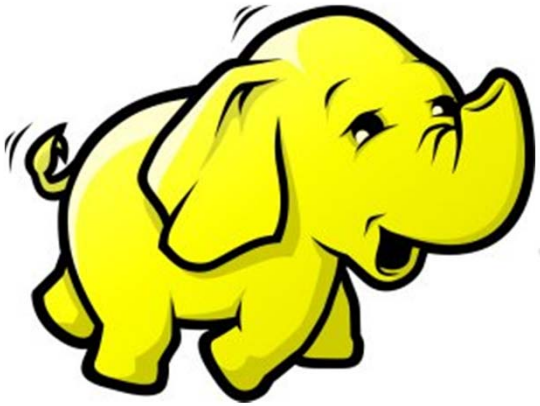
1. 発表者の紹介
2. 分散ファイルシステムとは
3. GlusterFS概論
4. GlusterFSの最新動向
5. GlusterFSの今後(を占う)
6. まとめ
7. 参考

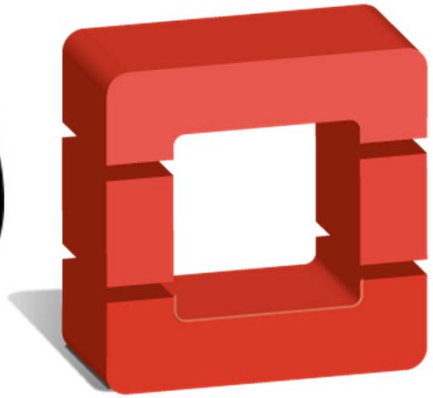
GlusterFS 3.3qa版

- オブジェクトストレージの実装



- クラウド基盤連携





openstack™

GlusterFSの最新動向

1. 発表者の紹介
2. 分散ファイルシステムとは
3. GlusterFS概論
4. GlusterFSの最新動向
5. GlusterFSの今後(を占う)
6. まとめ
7. 参考

- HekaFS(旧CloudFS)とのマージ？

- 迫り来る強力なライバルたち？

1. 発表者の紹介
2. 分散ファイルシステムとは
3. GlusterFS概論
4. GlusterFSの最新動向
5. GlusterFSの今後(を占う)
6. まとめ
7. 参考

まとめ

- GlusterFSの概要と最新動向をお伝えしました。
- **完成度が高く特徴的**でありながら、**継続的な発展**が期待できる分散FSです。
- HadoopやOpenStackと連携できるため、クラウドシステムを「**組み合わせて作る**」のに適しています。
- **扱いやすくラーニングコストも低い**ため、ヒューマンリソースを他の複雑なシステムに向けることができます。
- **拡張性が高い**ため、プログラマにとっても面白いテーマだと思います。

1. 発表者の紹介
2. 分散ファイルシステムとは
3. GlusterFS概論
4. GlusterFSの最新動向
5. GlusterFSの今後(を占う)
6. まとめ
7. 参考

- Gluster Support
  - Red Hat Networkによるサポートへ移行中。

- NTTPC Gluster Support
  - Glusterの正式な代理店。
  - ライセンスの販売取り次ぎ。
  - 日本語でのサポート。
  - 日本語での構築支援。
  - storage-contact @ nttpc.co.jp まで。

- 過去の活動
  - 第一回GlusterFS座談会
    - 2011/09/14 於 株式会社プリファードインフラストラクチャー様
    - スライド公開中（プログラマ／ソフトウェア技術者向け）。
    - Ustream（録画）公開中。-> http://www.ustream.tv/channel/glusterfs
    - Togetterあり。-> http://togetter.com/li/188183
  - 第？回社内Lightning Talk Nite
    - 2011/11/2 於 (株)NTTPCコミュニケーションズ
    - スライド公開中（Web系プログラマ向け）。
- 今後の活動予定
  - 第四回クラウドストレージ研究会
    - 2011/12/8 於 ビットアイルセミナールーム（天王洲）
    - より技術的な詳細を紹介予定（インフラエンジニア向け？）
  - 第二回GlusterFS座談会
    - 日程未定

- 公開資料
  - http://www.slideshare.net/keithseahus

- 日本GlusterFSユーザー会
  - http://groups.google.com/group/gluster-ja

※上記に関しては発表者のGluster Community Memberとしての個人的活動になります。各関係企業は関与をしておりません。

1. 発表者の紹介
2. 分散ファイルシステムとは
3. GlusterFS概論
4. GlusterFSの最新動向
5. GlusterFSの今後(を占う)
6. まとめ
7. 参考

ありがとうございました