

“Practical” Guide to GlusterFS



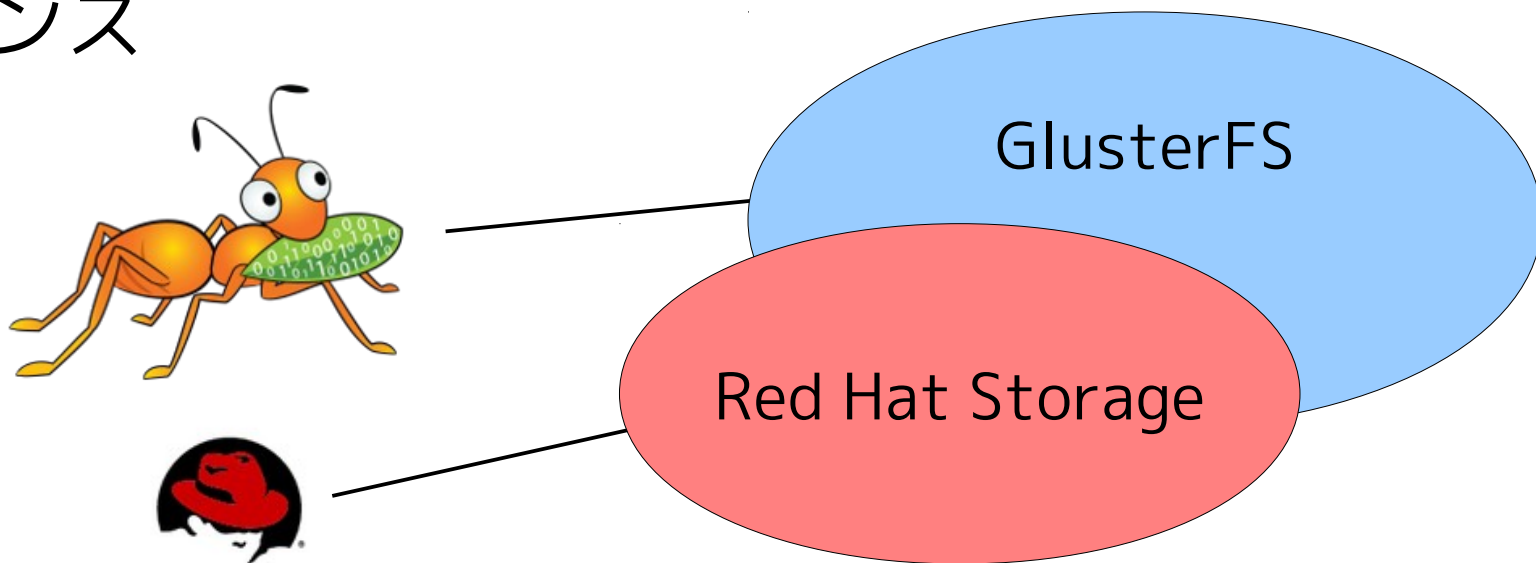
2012/07/20
Etsuji Nakai
Senior Solution Architect
Cloud Evangelist
Red Hat K.K.

GlusterFSとは

- コモディティハードウェアで、スケールアウト型分散ファイルシステムを構築するソフトウェア
- Linuxが動く所であればどこでも利用可能
 - オンプレミスでの利用
 - クラウド上での利用
- 複数APIからのアクセスに対応
 - REST APIで保存したファイルをHadoopで処理するなど可能

GlusterFSとRed Hat Storageの関係

- GlusterFSは、コミュニティメンバーによるオープンソースプロジェクトとして開発を継続
- Red Hat Storage (RHS) は、コミュニティ版をアップストリームとする安定版に、RHEL/XFS/管理ツールなどを組み合わせたソフトウェアアプライアンス



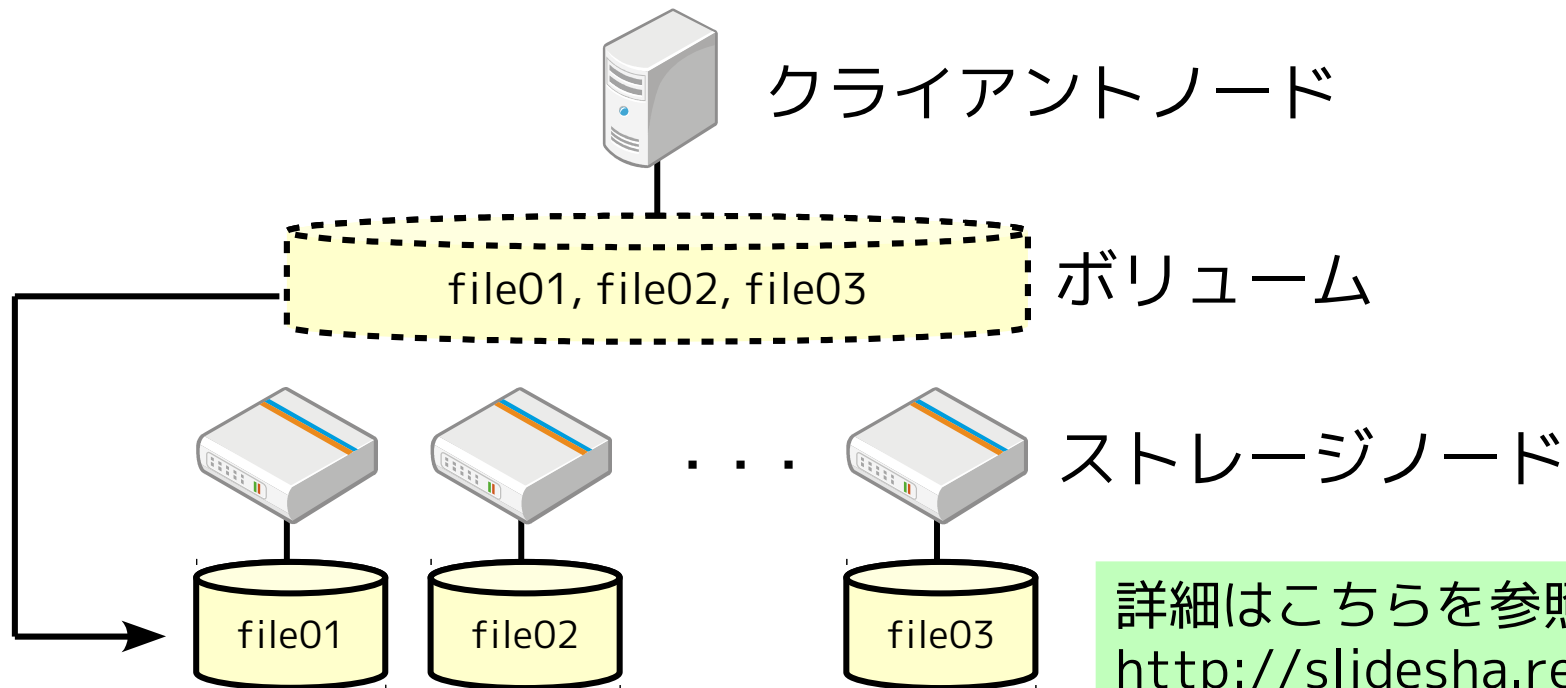


- NFS Lock対応 / Memory Leak Fix / ボリュームの動的縮小 / Replication数の動的変更 etc....

<http://download.gluster.org/pub/gluster/glusterfs/>

メタデータノードを持たないアーキテクチャ

- ファイル名のハッシュで保存するノードを決定
- ディレクトリ毎に分散ハッシュテーブルを用意
ーハッシュレンジなどの情報は拡張属性に保存
- レプリケーション、ストライピングなども可能



詳細はこちらを参照
<http://slidesha.re/LvEusu>

GlusterFSの使い所

- 数百TBの大容量ファイルサーバが必要な場合
 - Near-line / Backup / Archive Storage
- 多数のクライアントからの同時アクセスに対するパフォーマンス耐性が必要な場合
- HPCクラスタのバックエンドストレージ
 - IO Intensiveな計算には向かない（目安は “4～5K IOPS/node” 以下）
- 複数ロケーションに同じデータを配置する場合
 - GeoReplicationを利用

GlusterFSを使わない方がよいケース

- High Transaction / IO Intensive
 - 特にRDBのバックエンド
 - 「mv処理のAtomicity」などのコーナケースに依存するアプリケーションにも要注意
- 同一ファイルへの書き込み競合が多い場合
 - 整合性を確保するためのロック機構のオーバーヘッドはそれなりに大きい

「SAN Storageの安価な代替」とは考えないこと

(参考) 典型的なHW構成

- Nehalem-EX以降の2ソケットサーバ
- Disk
 - SAS Disk (RAID6) / Archive目的ならSATAでも可
 - ファイルシステムのi-node サイズは、512B推奨
- Memory
 - 16GB(Generic) / 32GB(HPC) / 48GB(Archive)
- Network
 - 10GbE or 1GbE (HPCでは10GbE推奨)
 - Bondingで冗長化

その他の気になるキーワード

- 仮想化ハイパーバイザのバックエンドストレージ
 - RHEV / KVM
 - VMware
- CIFS/Samba
- InfiniBand
 - TCP/IP over IB
 - RDMA
- IoDrive / SSD

個別にご相談ください



Thank you