

東京大学 HPCI 共用ストレージにおける オープンソースソフトウェアの活用

原田 浩¹ 蛭原 純² 藤本 大輔² 實本 英之² 平野 光敏¹

概要：HPCI 共用ストレージは Gfarm ファイルシステムを採用した大規模分散並列ファイルサーバである。東京大学柏キャンパスと理化学研究所計算科学研究機構に分散設置され、保守ベンダー、サーバベンダーとともに異なるマルチベンダーシステムでもある。このような大規模分散・マルチベンダーシステムを円滑かつ効率的に運用するには、分散環境に対応した、ベンダーフリーな管理・運用ツールの導入が不可欠である。HPCI 共用ストレージの資源提供機関の一つである東京大学情報基盤センターでは ZABBIX、Ganglia、Puppet 等のオープンソースソフトウェアを積極的に導入することによって、マルチベンダー構成の大規模分散システムである HPCI 共用ストレージを一元的に管理・運用可能な体制を構築した。

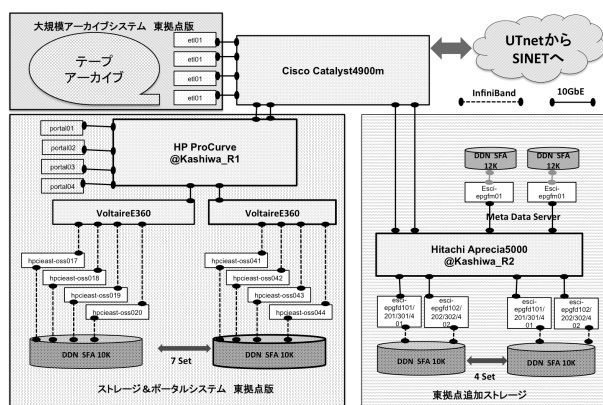


図 1 東京大学 HPCI 共用ストレージ構成

1. はじめに

東京大学情報基盤センターでは、2012 年 9 月の HPCI 運用開始から HPCI 共用ストレージ [1] の資源提供と運用を行っている。運用しているハードウェア資源の概要を図 1 に示す。HPCI 共用ストレージは、東京大学柏キャンパスと理化学研究所計算科学研究機構（以下 AICS）に分散設置された分散並列ファイルサーバである。東京大学・AICS 設置機材の合計でサーバ計算機 60 台、ストレージシステム 23 セット、データ容量 21.5PB にもなる大規模システムでもある。サーバ計算機 28 台は CentOS を採用している。さらに東京大学・AICS 設置機材ではサーバベンダー、保守ベンダー共に異なるマルチベンダーシステムでもある。HPCI 共用ストレージは上記のような大規模システムであ

るが、東京大学ではハードウェアベンダーの運用・保守関係者は常駐せず、ハードウェア保守契約のみで運用している。このような大規模な分散マルチベンダーシステムを円滑に運用するため、東京大学ではオープンソースソフトウェアを積極的に活用することとした。

2. 運用課題

HPCI 共用ストレージは、2012 年 9 月の本稼働開始に先立ち 2012 年初頭から 2 度にわたるシステム構築・運用のリハーサルを行った。リハーサル運用の結果、以下のような課題が明らかとなった。

2.1 ユーザ管理

HPCI 共用ストレージの使用を許可された課題とユーザに関する情報は HPCI 申請支援システムから入手する。採択課題・終了時だけではなく、各課題にユーザが追加・削除された場合もユーザの追加・削除等の業務が頻繁に発生する。HPCI 申請支援システムから入手したユーザ情報を複数システムから構成される 60 台ものサーバ計算機に迅速に展開する必要がある。さらにログインサーバでは、通常のアカウント管理に加えて、ユーザごとに HPCI 共用ストレージのマウントポイントを作成する等の作業が必要となる。

2.2 Gfarm 保守

HPCI 共用ストレージは、東京大学設置機材だけでもメタデータサーバ 2 台、データサーバ 32 台に及ぶ。さらにストレージユーザ用ログインサーバ 4 台も東京大学に設置

¹ 東京大学 情報システム部情報基盤課

² 東京大学 情報基盤センター

されている。これら 38 台ものサーバ計算機において短時間で Gfarm システムの保守・更新作業を行う必要がある。システムのメンテナンス中にハードウェア障害対応等の作業時間を考慮すると、長くとも 1 時間程度で全サーバの Gfarm の保守・更新作業を確実に実行できる必要がある。

2.3 セキュリティツール管理

HPCI 共用ストレージは、インターネットを通じて HPCI 参加機関からアクセスされる設計となっている。そのため HPCI 共用ストレージはインターネットからの不正侵入等の脅威に耐える高いセキュリティを備えたシステムであることが求められる。セキュリティツールの管理・運用についてもメンテナンス時に全サーバに対して迅速な保守、更新が求められる。

2.4 ストレージ監視

HPCI 共用ストレージは、既に述べたようにマルチベンダー・分散システムである。運用・管理を行うエンジニアも複数機関から構成される。このような環境では、複数の組織からネットワークを通じてシステムを一元管理できるツールの導入が必須である。

3. オープンソース活用

HPCI 共用ストレージの構築・運用リハーサルから、HPCI 共用ストレージのような大規模分散システムを円滑かつ効率的に運用するには、以下の条件を満たす管理・運用ツールの導入が必須となる。

- サーバベンダー、ストレージベンダー、保守ベンダーに依存しないツールであること
- https 等のプロトコルを用いて、複数機関から安全にネットワーク越しに情報共有できること
- CentOS 等フリーな OS 上でも動作実績があり、実運用に耐えるツールであること

東京大学情報基盤センターでは、本稼働に備えリハーサル環境で見つかった問題点を解決するため、以下のオープンソースソフトウェアを導入し分散環境に対応した効率的な管理・運用体制を構築した。

3.1 Puppet

Puppet[3] は、Puppet Labs が配布するオープンソースの構成管理ツールである。”Puppet manifests”と呼ばれるファイルにシステム構成情報を記述しておく、対象となるシステムを自動で構成できるようになる。HPCI 共用ストレージでは、以下の業務に Puppet を活用している。

3.1.1 ユーザ管理

HPCI 共用ストレージでは、ユーザの認証方法として GSI 認証を採用している。以下の手順で、共用ストレージにユーザ登録を行う。

● アカウント作成

共用ストレージに新規ユーザを登録する際には、共用ストレージすべての機器に対して、適切なディレクトリ作成および権限設定を行う必要がある。煩雑になりがちなこれらの作業を、Puppet を利用して自動化している。HPCI 申請支援システムにアクセスし、新規ユーザ情報を含む CSV ファイルをダウンロードする。既存ユーザを登録している CSV ファイルにマージして、共用ストレージを利用するユーザ全員の情報を持つ CSV ファイル (以下、ユーザ情報 CSV ファイル) を作成する。共用ストレージ機器のユーザ登録は、すべて Puppet 経由で行っている。Ruby スクリプトを利用して、ユーザ情報 CSV ファイルから Puppet 設定定義ファイルに変換する。アカウントを登録する共用ストレージ機器で、Puppet によりユーザの UNIX アカウント生成、グループ生成、Gfarm マウントディレクトリ生成を自動化した。

● grid-mapfile 配布

GSI 認証では、利用可能ユーザー一覧となるテキストファイル grid-mapfile を利用している。HPCI ユーザのうち、グリッド証明書が既に発行されているユーザのリスト (以下、グリッド証明書ユーザリスト) を、HPCI 事務局の指定サーバから subversion で一時間に一度取得する。差分ユーザリストから grid-mapfile を作成し、ユーザがアクセスする可能性のある共用ストレージ機器すべてに配布する。

● Gfarm ユーザ領域作成

前項で作成した、差分ユーザリストから、Gfarm 設定コマンドを生成し、各ユーザについての Gfarm 環境を設定する。Gfarm ユーザ作成、Gfarm グループ作成、Gfarm ディレクトリ作成、Quota 設定が自動化される。

3.1.2 Gfarm 保守

HPCI 共用ストレージでは、広域分散ファイルシステムとして、Sourceforge で公開されているオープンソースソフトウェアである、Gfarm を利用している。共用ストレージの利用にあたって、日々運用を続ける中で生じた問題点および改善点を、常に Gfarm 開発元へフィードバックを行い、運用側と開発側が密に協調した形で、より安定したサービスを提供するよう努力を続けている。

共用ストレージでは、これまで表 1 に示す日程で Gfarm の更新を行ってきた。当初、Gfarm の更新については、作業手順マニュアルを作成し、一台ずつマニュアルに従って設定作業を行ってきた。しかし、更新が頻繁になるにつれ、作業時間の短縮と、複雑かつ多岐に渡る設定について作業漏れを事前に防ぐ対策が必要となり、設定作業を Puppet により自動化した。自動化により数分で全サーバの Gfarm システムを確実に更新することが可能となり、メンテナン

ス時間の削減に貢献した。

3.1.3 セキュリティツール配布

共用ストレージのセキュリティ設定にあたり、以下の項目に留意している。

対象の明確化 どの機器のセキュリティを保守するか明確にする

設定内容の明確化 どのように機器を設定するか明確にする

利用ツールの明確化 どのセキュリティツールのどのバージョンを利用するか明確にする

当初、セキュリティツールの配布および設定は、作業手順書を作成して、設定及び更新作業を行ってきた。Puppet を用いることにより設定内容が明確になり、作業漏れも少なくなり確実に多数の機材にセキュリティツールを導入することが可能になった。

Puppet を利用してソフトウェアおよび設定ファイルを一元管理しているセキュリティツールを以下に示す。

- chkrootkit
- denyhosts
- fail2ban
- logwatch
- openssh
- rkhunter
- tripwire

3.1.4 その他

上記の他に保守作業全般に渡る以下の作業を Puppet により自動化している。

- 必要なディレクトリ作成およびパーミッション設定
- Gfarm ソースファイル取得および展開
- Gfarm 及び関連ソフトウェアコンパイルおよびインストール
- 各種設定ファイル設定
- 共用ストレージ全機器について各種証明書の一元管理
- 各種 PC ファームウェアの一元管理

表 1 共用ストレージにおける Gfarm 更新履歴

2012/05/08	gfarm2.5.5
2012/07/20	gfarm-r6436
2012/08/07	gfarm-r6498
2012/08/22	gfarm-r6582
2012/08/27	gfarm-2.5.6
2012/09/03	gfarm-2.5.7
2012/10/11	gfarm-2.5.7.1
2012/11/02	gfarm-2.5.7.2
2012/12/27	gfarm-2.5.7.3
2013/03/27	gfarm-2.5.8-rc3
2013/04/24	gfarm-2.5.8
2013/05/27	gfarm-2.5.8.1
2013/08/26	gfarm-2.5.8.2
2013/10/04	gfarm-2.5.8.3



図 2 ZABBIX によるシステム監視

- 一般ユーザ及び管理ユーザの一元管理

3.2 ZABBIX

ZABBIX[4] は ZABBIX SIA が開発・配布を行っているシステム監視ツールである。ウェブインターフェイスをサポートし、ネットワーク経由で大規模システムの監視を行うことが出来る。HPCI 共用ストレージでは、システム監視に ZABBIX を利用している。監視画面の例を図 2 に示す。サーバ管理として一般的な監視項目に加え、Gfarm システム固有の監視項目を gfarm-zabbix プラグインにより補い、さらにサイト固有の項目については、独自の監視プラグインを作成し、監視を行っている。ZABBIX を用いることにより、ウェブブラウザから東京大学・AICS 両設置機材を一元的に監視できる体制を構築した。

一般的な監視項目は以下の通りである。

- ネットワーク流量グラフ
- ARP 監視
- 障害全般監視
- 各種サーバアラート検出

gfarm-zabbix プラグインを利用した監視項目は以下の通りである。

- Gfarm 関連プロセス監視
サイト固有の監視項目については以下の通りである。
- 新規共用ストレージユーザ登録検出
- グリッド証明書の CRL が更新されているか検出

3.3 Ganglia

Ganglia[5] はクラスタ計算機、Grid 環境向けに開発された監視システムである。HPCI 共用ストレージでは、サイト全体のネットワークトラフィック及び Gfarm メタデータサーバ/ファイルサーバの処理件数の監視に利用している。Ganglia を用いることにより、東京大学・AICS 両機関



図 3 Ganglia によるトラフィック監視

HPCI 共用ストレージの統計監視が可能になった。監視画面の例を図 3 に示す。

3.4 その他のツール

トラブルチケットシステムについては Gfarm 開発元へのレポートは trac[6] を利用し、内部の細かい作業については fossil[7] を利用している。trac によるチケットシステム画面の例を図 4 に示す。fossil の wiki に運用上気がついた点をその都度メモ書きしておき、運用ドキュメントとして Sphinx[8] 形式のファイルでまとめてゆく。Sphinx を利用して、単一ドキュメントファイルから Web サーバ参照用コンテンツと印刷用ドキュメントを生成している。

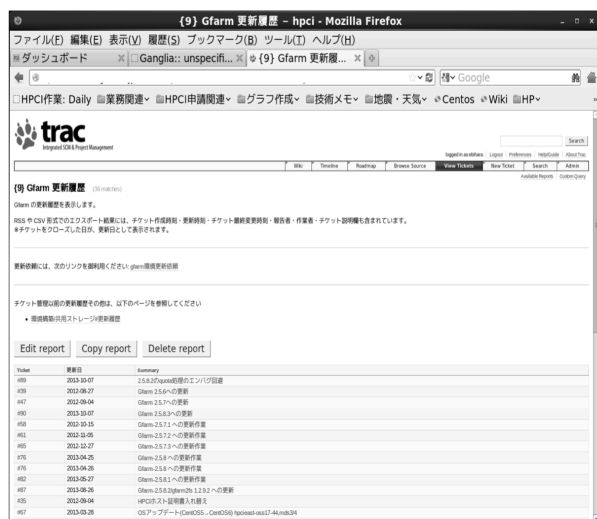


図 4 trac によるチケット管理

4. まとめ

東京大学・AICS が HPCI に資源提供している HPCI 共用ストレージは、マルチベンダー機材で構成された大規模

分散ファイルシステムである。本稼働に先立つ構築・運用リハーサル経験から、HPCI 共用ストレージを円滑かつ効率的に運用するには、分散環境に対応したベンダーフリーなツールの導入が必須であると判断した。東京大学情報基盤センターでは、分散環境に対応したベンダーフリーなツールとして Puppet、ZABBIX、Ganglia 等のオープンソースソフトウェアを積極的に導入・活用した。これらオープンソースソフトウェアの活用によって、HPCI 共用ストレージの一元的管理・運用体制を構築した。2012 年 9 月 28 日の HPCI の運用開始から 1 年以上の実運用を継続しているが、ほぼ計画通りの運用実績を達成し、2013 年 10 月 21 日現在で、56 課題、439 ユーザに対して HPCI 共用ストレージを提供している。今後もオープンソースソフトウェアの導入を推進し、より効率的な運用環境の構築を目指すとともに、安定稼働の継続によってユーザの研究活動をサポートする所存である。

謝辞 本取り組みは文部科学省「HPCI の運用」の支援を受けている。

参考文献

- [1] 東京大学 情報基盤センター, HPCI 関連サービス, <http://www.cc.u-tokyo.ac.jp/other/hpci/>.
- [2] Gfarm File System, <http://datafarm.apgrid.org/index.en.html>.
- [3] Puppet Open Source, <http://puppetlabs.com/puppet/puppet-open-source>.
- [4] ZABBIX, <http://www.zabbix.com>.
- [5] Ganglia Monitoring System, <http://ganglia.sourceforge.net>.
- [6] The Trac Project, <http://trac.edgewall.org>.
- [7] Fossil, <http://www.fossil-scm.org>.
- [8] SPHINX Python Documentation Generator, <http://sphinx-doc.org>.