

HPCI 共用ストレージの運用

實本 英之¹ 塙 敏博¹ 原田 浩² 中 誠一郎¹ 建部 修見³ 平川 学² 藤本 大輔¹ 蛭原 純¹
宮寄 洋¹ 佐島 浩之¹

概要：概要：HPCI 共用ストレージは、HPCI 参加機関に対して容量約 21.5PB を提供する、高速広域ファイルシステムである。HPCI 共用ストレージの構成および、利用本格化から現在までの実際の運用状況を紹介する。また、現状の運用を踏まえた将来的な利用予測についても述べる。

1. はじめに

革新的ハイパフォーマンス・コンピューティング・インフラ (HPCI)[1] では、データ共有基盤として共有ストレージを提供している。この共有ストレージは、広域ファイル共有システム Gfarm[4] を利用したもので、東京大学柏キャンパス、理化学研究所計算科学研究機構 (AICS) に総容量 21.5PB を分散配置し運用を開始した。運用後、東京工業大学を新たに拠点として加えている。この共有ストレージは、HPCI 認証基盤を用いており、HPCI ユーザーはどのインフラ上のどの計算資源からとも同一ユーザーとしてアクセスすることが可能である。これにより、多数のユーザーで構成され、複数の拠点に跨り計算資源を利用するような、大規模な研究プロジェクトにおいてデータの共有を容易にすることができる。

2. HPCI 共用ストレージの特徴

東京大学、AICS に分散設置されている HPCI 共用ストレージの構成概要をそれぞれ、図 1、図 2 に示す。

総容量 21.5PB もの領域を、長期間安定的にサービスを行うため、メタデータの多重化、ファイルデータ/ストレージコントローラの二重化、RAID6 の利用等を行っており、ほとんどの物理障害に対しても活性交換による連続運用が可能である。さらに、各種障害に関しては、自動検知システムをもち、ハードウェア障害や復旧状態に関しては、関係者間で即座に共有することが可能になっている。システム監視においては、多拠点に跨った設備であるため保守業者、サーバ計算機メーカー、サーバ OS 等が異なった状況に対応可能な監視ツールが必要となっている。この

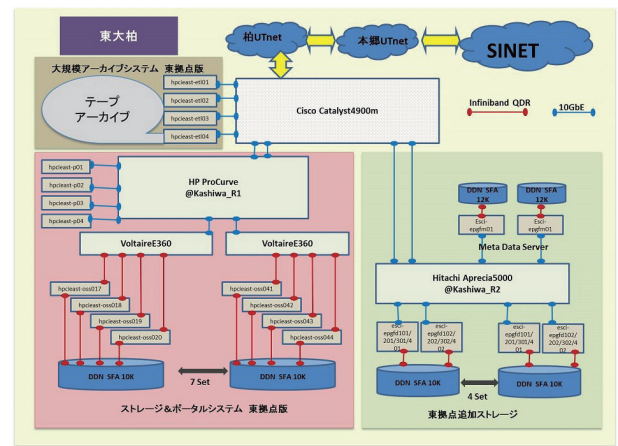


図 1 HPCI 共用ストレージ東京大学設置機材

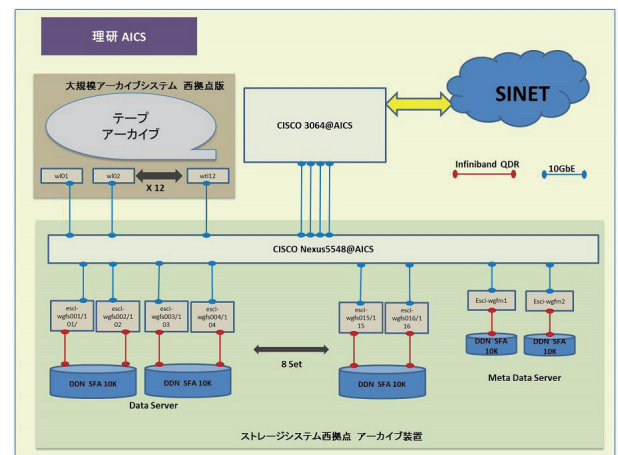


図 2 HPCI 共用ストレージ AICS 設置機材

ため、1) 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、2) Well-known なプロトコルを用いて、複数機関から安全に情報共有可能、3) CentOS 等のフリー OS 上でも動作実績があり実運用に耐える、を満たすツールを選定した。これを満たすものとして、ZABBIX[2] に

¹ 東京大学
² 理化学研究所
³ 筑波大学

よる稼働状況監視と Ganglia[3] によるアクセス状況監視を導入した。これにより、HPCI 運用関係者は共用ストレージ機材全体の動作状況を一元的に把握可能である。また、高性能なファイル入出力を実現するために、InfiniBand 4xQDR による高帯域内部ネットワークの採用すると共に、メタデータサーバにおける書き込みバイトブロックの最適化も行っている。1GiB ファイル 300 個程度のファイルを HPCI 構成機関である、北海道大学、東京大学、京都大学から 100 並列程度で書き込みした結果、北海道大学で 450MB/s、京都大学で 900MB/s 程度の性能を得ている。これは地理的な影響による性能低下である。なお、京都大学では AICS と東京大学双方への書き込みがほぼ同じ性能であった。東京大学からの書き込みは距離による性能低下が最も少なく 1100MB/s の転送帯域を実現している。

3. 運用

本項では、2013 年 11 月以降、2014 年 10 月現在までの運用まとめについて報告する。

3.1 総容量

2013 年 11 月から 2014 年 10 月までのストレージ利用量を図 3 に示す。これは、Gfarm ファイルシステムによる複製の有無、それぞれに関しての利用量となっている。本格利用が始まり 1 年を経過した 2013 年の年末より、ストレージへの書き込みが一斉に開始されており、それまでアプリケーションのプロトタイプ実行や、デバッグ、チューニングを行っていたユーザが本格的な HPCI 利用を開始したことが見て取れる。以降、利用量は順調に伸び、2014 年 10 月において、全体のおおよそ 69%にあたる 15PB が利用された。また、総利用量のうち、ほぼ 50% ずつを AICS-東大で分け合っていることが確認されており、レプリケーションの多くが AICS、東大間で行われていることがわっている。図 4 は増加量が安定した 2014 年 4 月からのストレージ利用量を対数補間したものを示している。これに基づく、2015 年 7 月には現在と同程度の冗長度で複製を行うことが難しくなり、2018 年 8 月には複製の無い状態でもストレージを使い切ることが予測できる。

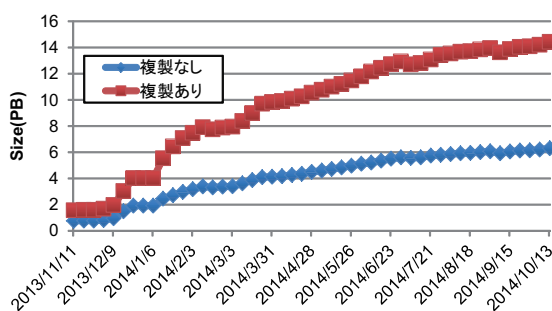


図 3 共用ストレージ利用量推移

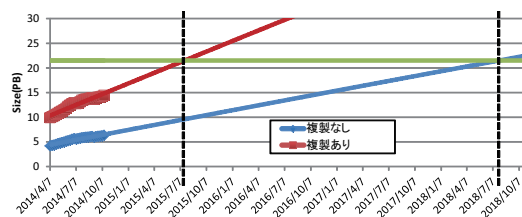


図 4 共用ストレージ利用量予測

3.2 ファイルサイズ

図 5 は 2013 年 11 月から 2014 年 10 月における、ファイルの平均サイズを示している。基本的に増加傾向であるが、大きな減少がおおよそ 2 度確認された。1 度目は 2014 年 3 月で、これは年度末のプロジェクト終了に合わせたデータ削除による影響である。2 度目の 2014 年 9 月については、ストレージ利用量増加に伴う容量圧迫のため、ユーザにファイル整理、テープドライブの利用を要望した時期と重なっている。つまり、ユーザは比較的サイズの大きいファイルを削除したと考えられる。続いて 2014 年 10 月におけるサイズ毎のファイル数を図 6 に、サイズ毎のストレージ利用量を図 7 に示した。1MB 未満の小さなファイルが大量にある一方、1GB 以上のファイルが利用量全体の 70% 近くを占めていることがわかる。

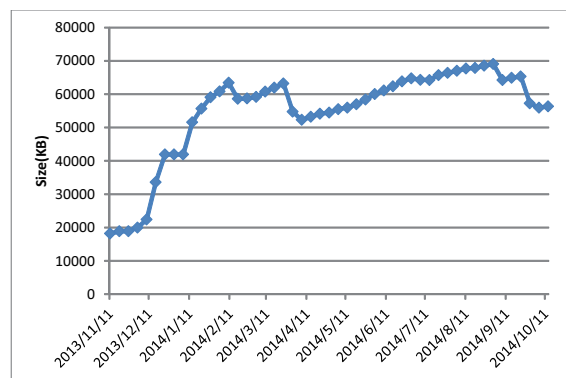


図 5 ファイルの平均サイズ

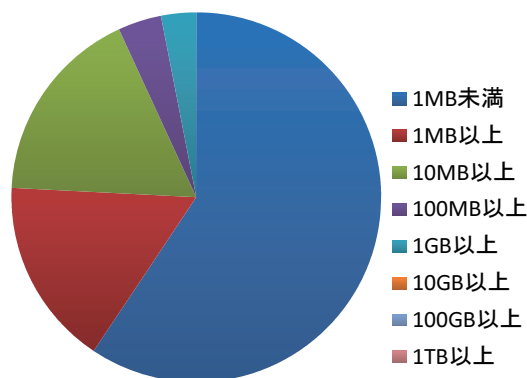


図 6 サイズ毎のファイル数

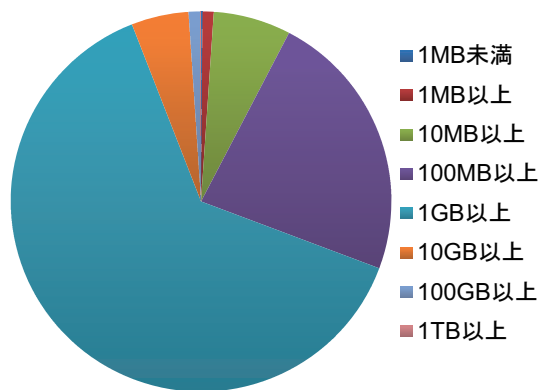


図 7 サイズ毎のストレージ利用量

4. まとめ

HPCI 共用ストレージについて 2014 年度の利用状況を紹介した。また、このデータを元に、今後のストレージ利用量を推測した結果、2015 年 7 月には複製が作成しがたくなり、2018 年 7 月には全てのファイルが複製を持たない状態でも、空き容量がなくなると予想される。

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] 革新的ハイパフォーマンス・コンピューティング・インフラ, http://www.mext.go.jp/a_menu/kaihatu/jouhou/hpci/1307375.htm
- [2] ZABBIX, <http://www.zabbix.com>
- [3] Ganglia Monitoring System, <http://ganglia.sourceforge.net>
- [4] Osamu Tatebe, Kohei Hiraga, Noriyuki Soda, Gfarm Grid File System, New Generation Computing, Ohmsha, Ltd and Springer, Vol.28, pp. 257-275, 2010