JPARC 実験データの共通計算機システムへの転送

八代茂夫

高エネルギー加速器研究機構 共通基盤研究施設 計算科学センター

概要

KEKCC のストレージシステム HPSS にアクセスするために用意されている各種のインターフェイスとそれぞれの特徴を述べる。次に HPSS への KEK 内からのアクセスおよび約 70km 離れた JPARC 実験施設からのアクセスの性能の測定結果を報告し、HPSS を利用するにあたってのインターフェイス選択の目安を示す。

1 はじめに

高エネルギー加速器研究機構(KEK)で行われる実験のデータを解析することを目的とするシステムである共通計算機システム(KEKCC)[1] は 2009 年 3 月に更新された。今回導入されたシステムでは従来からのプロジェクトの利用に加えて、新しいプロジェクトである JPARC 実験により生成されるデータの保管および解析を主目的の 1 つとしている。KEKCC はつくば地区に設置されており、JPARC の実験施設は直線距離で約 70km 離れた東海地区にあるので、この間でデータ転送をおこなう。KEKCC のストレージへのデータ転送にあたっては高速かつ安定的に転送することが課題である。

データ転送に利用する可能性のあるアクセスインターフェイスを検討し、その転送性能の測定を行なったでの報告する。

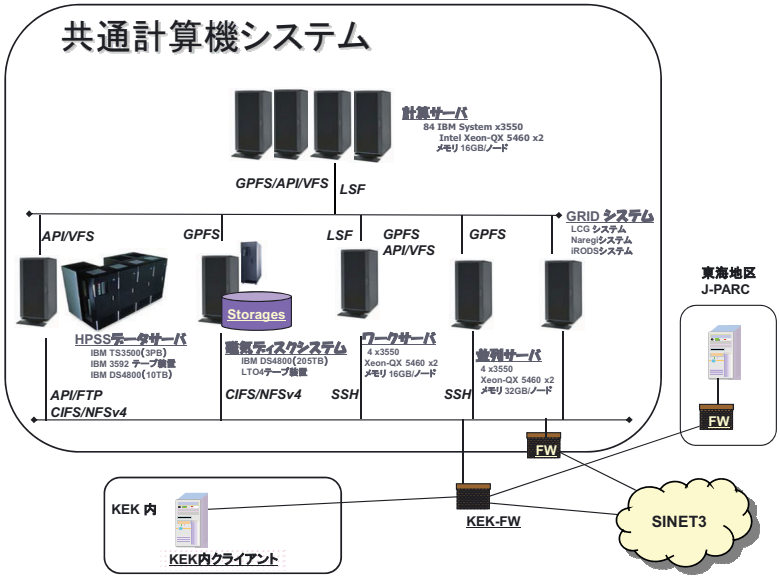

図 1. 共通計算機システムの概略図

2　大容量ストレージシステム HPSS

2.1　共通計算機システムの概要

計算サーバ、ワークサーバ、並列サーバ、GRID システム、磁気ディスクシステム、および実験データを蓄積する大容量ストレージシステム(HPSS データサーバ)からなる。大容量ストレージシステムは、磁気テープライブラリ、キャッシュディスク、サーバ群により構成され、ソフトウエア High Performance Storage System (HPSS)[2] で管理される。図 1 はシステムの構成図である。

2.2　HPSS アクセスのインターフェイス

HPSS により管理される大容量ストレージシステムは最大容量 10PB の磁気テープライブラリ装置、10TB のキャッシュディスク装置、サーバ群により構成されている。図 2 はシステムの概略図である。

HPSS でのデータの書き込みは、先ずキャシュディスクに対して行われ、一定時間経過後に磁気テープに転送される。読み出し時はキャシュディスクにデータがある場合には、そのデータがクライアントに送られる。磁気テープにある場合には、キャシュディスクに転送された後に送られる。KEKCC ではキャシュディスクへの転送が始まると同時にクライアントに送られる設定にしている。データ転送は HPSS の core サーバ、mover サーバや VFS サーバあるいはワークサーバや IRODS サーバを通じて行なわれる。

KEKCC でサポートしている HPSS アクセスのインターフェイスを表 1 に示す。POSIX 準拠 I/O 関数(API)、Parallel FTP(pftp)、Kerberos ftp(kftp)、およびファイルシステムインターフェイスの VFS が HPSS により用意されている。

KEKCC では更に VFS を経由して SSH, i Rule Oriented Data Systems(iRODS)[3] , gridftp, および CIFS でアクセスできる環境を構築した。SSH によるファイル転送には scp, sftp, SSHfs[4] , WinSCP などが利用できる。また API によるアプリケーションである hpsscat や hpssput などのファイル転送ができる機能も用意されている。API、pftp や hssput/hpsscat の利用には HPSS パッケージの導入が必要である。kftp は Linux のディストリビューションに含まれる Kerberos の設定を行なえば利用できる。iRODS の利用にはクライアントパッケージ

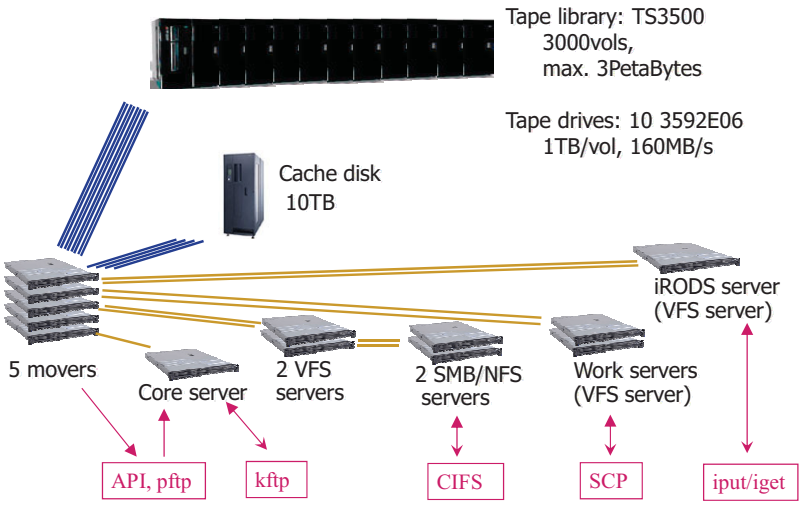

図 2. 大容量ストレージシステム HPSS

の導入が必要である。

利用者はクライアントのおかれている環境に合わせて最適な方法を選択して利用できる。ファイアウオールを通過するか､NAT を経由するか、高速な転送を求めるか、ネットワークの遅延の程度、ユーザインターフェイスの好みなどを考慮して選択する。表 1 に示した FW との相性「難あり」は、通過させる必要のあるポートが相当数になるインターフェイスである。通過させるには 、FW の運用方針の確認が必要である。

表 1. HPSS アクセスのインターフェイス

インターフェイス	特徴	FW 相性	NAT 経由
POSIX 準拠 I/O 関数	C の関数	難あり	不可
hssput/hpsscat	ファイル転送	難あり	不可
Parallel FTP(pftp)	ファイル転送	難あり	不可
Kerberos ftp(kftp)	ファイル転送	良	可
VFS	ファイルシステム		
SSH	VFS 経由	良	可
iRODS	VFS 経由	難あり	可
gridftp	VFS 経由		
CIFS	VFS 経由	良	可

API および pftp は core server になされた要求に対して mover からデータが転送されるので、一般的に高速な転送が可能である。しかし第 3 者転送を行なうので NAT 経由では処理できない。

2.3 インターフェイスの使用例

それぞれのインターフェイスについて、使用例を簡単に示す。

- Hpsscat/hpssput によるファイル転送

  $ hpssput /hpss/ce_g/cc/yashiro/test/outfile local_file

  $ hpsscat /hpss/ce_g/cc/yashiro/test/remote_file > outfile

- Parallel FTP によるファイル転送

  $ /opt/hpss/bin/pftp_client -v hco01.cc.kek.jp 4021

  cd /hpss/ce_g/cc/yashiro/test/

  ftp> put local_file

  ftp> get remote_file

  bye

- Kerberos ftp によるファイル転送

  $ kftp

  ftp> put local_file

  ftp> get remote_file

  bye

- Scp によるファイル転送

  $ scp local_file hpss.cc.kek.jp: /hpss/ce_g/cc/yashiro/test/outfile

  $ scp hpss.cc.kek.jp: /hpss/ce_g/cc/yashiro/test/remote_file outfile

- Sftp によるファイル転送

  $ /usr/bin/sftp -v hpss.cc.kek.jp

  cd /hpss/ce_g/cc/yashiro/test/

  ftp> put local_file

ftp> get remote_file

bye

- SSHfs マウントとファイル転送

  $ sshfs hpss.cc.kek.jp: /hpss/ce_g/cc/yashiro/test/ ~/mnt

  $ cp local_file ~/mnt/outfile

  $ cp ~/mnt/remote_file outfile

- iRODS によるファイル転送

  $ iput -f local_file outfile

  $ iget -f remote_file

3 転送性能

3.1 測定条件

つくば地区の KEK LAN から KEKCC の HPSS にアクセスする場合の転送性能と、東海地区の JLAN からアクセスする場合の転送性能を測定した。RTT はそれぞれ約 0.9ms、10ms であった。

クライアント計算機の CPU は XEON X5450 を 2 CPU、メモリーを 4GB 搭載した計算機で、OS は CentOS5.3, kernel 2.6.18 である。利用者が実際に使用することを前提に測定するので、特別なチューニング等を行わないことにした。

HPSS は運用中に測定した。HPSS のバージョンは 6.2.2, core サーバは IBM p550 (POWER6 3.5GHz 2Core 4CPU, 8GB)で OS は AIX5.3、mover サーバは IBM p5 520 (POWER5+ 1.65GHz 2Core 4CPU, 2GB)で OS は AIX5.3、VFS サーバは IBM x3650 (Intel Xeon-QX5460 3.16GHz 4Core, 8GB) で OS は RHEL4、IRODS サーバは IBM x3650 で OS は RHEL5、ワークサーバは IBM x3550 (Intel Xeon-QX5460 3.16GHz 4Core 2CPU, 16GB) で OS は RHEL5 である。

HPSS ではキャシュディスク容量やマイグレーションポリシーを適切に設定することにより、キャシュディスク領域が不足せず、利用者のアクセス中に磁気テープへのアクセスの発生が最少になるよう調整できる。この状況で利用することが推奨されている。この場合には、クライアントとキャシュディスクとの間のデータ転送になり、その転送性能が重要になる。これを今回の性能測定の対象とした。

測定したインターフェイスは kftp、pftp、hpssput/hpsscat、iRODS、scp である。測定に使用したのは 907MB の圧縮の効かないファイルである。

3.2 転送性能と LAN アダプタの関係

表 2 に転送性能の測定結果を示す。iput および iget は iRODS のファイル転送のコマンドである。

つくば地区の計算機は Intel 80003ES2LAN Gigabit Ethernet Controller およびプラネックスコニュニケーションズ(株)の GN-1200TW2 で測定した。表では前者を GbE1、後者を GbE2 と表わしている。Planex は Intel より性能が劣った。送信の場合には高性能の pftp および hpssput の性能が 64MB/s あたりで抑えられている。受信の場合にはインターフェイスによっては 3 分 1 以下との非常に悪い。

東海地区の計算機は Planex で測定した。Intel に代えると性能が向上する可能性がある。可能性を検討するためにつくば地区の Intel を搭載した計算機で、iproute パッケージの tc コマンドで 10ms の遅延を加えて測定した結果が表 2 の GbE1+ の値である。この値を東海地区の Planex での値と比較すると、特に scp での HPSS からの読

み出しは大きな性能向上を期待できる。

3.3 性能の比較検討

つくば地区からのアクセスでは、HPSS によって提供される pftp、hpssput、kftp の性能が 64MB/s～97 MB/s と非常に良い。高速な転送を求めるなら、これらのインターフェイルが好ましい。

一方、東海地区からのアクセスになると、pftp、hpssput、kftp の高性能が期待できなくなる。HPSS にファイルを送る場合には scp、iRODS、pftp の性能が良い。NAT 下のクライアントの場合には pftp は使えないので scp あるいは iRODS になる。HPSS からファイルを受け取る場合には iRODS あるいは kftp の性能が良い。

表 2. HPSS の転送性能 (MB/s)

GbE1 は Intel 80003ES2LAN アダプタ、GbE1+は同アダプタで 10ms の遅延を付加、GbE2 は Planex GN-1200TW2 アダプタ

	インターフェイス	つくば			東海
		GbE1	GbE2	GbE1+	GbE2
HPSS への書き込み	kftp-put	64.0	63.0	10.0	11.0
	pftp-put	97.3	64.3	18.1	26.6
	hpssput	89.5	61.3	15.9	19.7
	iput	23.5	23.0	24.0	23.2
	scp	32.4	28.4	25.2	31.3
HPSS からの読み出し	kftp-get	83.0	26.0	18.0	20.0
	pftp-get	86.5	38.7	14.6	17.0
	hpsscat	85.2	24.5	14.1	16.6
	iget	16.1	18.3	21.1	17.6
	scp	31.3	22.7	13.0	3.8

なお、iRODS については飯田好美氏の報告[5] が予定されている。

4 最後に

今回は利用者が余分な負担なく使いながら、ある程度の性能を得ることのできるインターフェイスを探ることを目的とした。そのためにパラメータのチューニングも行なわず、標準値を用いた。例えば SSH はデータの暗号化を選択できて、データの性質と暗号方式の組み合わせにより性能が大きく変わる。暗号方式を適切に選択すればより良い性能が得られる可能性がある。iRODS では通信の並列度を変更できるが、標準的な設定に任せた。Pftp の get では “setpb 4MB” オプションで性能が向上する可能性がある。東海地区からのファイル送信に更なる転送速度を求めるならチューニングを検討する余地がある。

計算機の性能、LAN アダプタによっても結果が大きく変わる。今後東海地区の計算機の LAN アダプタを変更して再測定をする予定である。また、計算機の性能による影響度を、それぞれのインターフェイスについて調査したい。

HPSS をファイルシステムで扱うには、VFS、CIFS、NFS がある。それぞれ一長一短がある。しかし SSHfs を利用すればサーバ側の設定なしに利用できる。通信には sftp を使っているので、インターネット越しの利用もセキュアにできる。ファイルシステムでのアクセスを希望する場合には検討の価値がある。

5 謝辞

HPSS 環境の構築に関して計算科学センターの佐々木節氏、飯田好美氏、および日本 IBM の伊藤義彦氏、玉井千恵子氏、山本智実氏をはじめとする方々に感謝します。J-PARC との接続について計算科学センターの真鍋篤氏、鈴木聡氏、鈴木次郎氏の多大なる協力に感謝します。

SSHfs および tc についてヒントを下さった日本原子力研究開発機構 J-PARC センターの石川弘之氏に感謝します。

参考文献

[1] KEKCC, http://kekcc.kek.jp/

[2] HPSS, http://www.hpss-collaboration.org/

[3] iRODS, https://www.irods.org/

[4] SSHfs, http://fuse.sourceforge.net/sshfs.html

[5] 飯田好美、iRODS を用いたデータ管理システムの導入、第５分科会 5-007