大規模データ解析に向けた高速ストレージの性能評価

中部大学工学部　大嶋晃敏

概要

近年、観測分野やシミュレーション分野では、システムの大規模化と観測機器の高性能化に伴って、データ量の増大とその処理に大きな負荷がかかるようになってきている。 スーパーコンピューターの共同利用では、システムあたりの利用者数が百数十名にも達することもあり、生成されるデータ量も年間数ペタバイト (PB) におよぶ。一方で、観測分野においても観測精度の向上に伴って、データ量の増大が著しく、データ保管とその処理のためのストレージの大規模化と高性能化が喫緊の課題となっている。本報告では、高性能コンピューティング (HPC) 分野で利用されている理論通信帯域 56 Gbps の Infiniband FDR を用いて大規模データ処理システムに向けた試作機の構築と性能評価試験を行なった。

1　はじめに

いずれの分野においても増大するデータ量とその効率的な処理は解決すべき課題であるが、天文観測や物理実験の分野でもプロジェクトの大型化と観測技術の先進化に伴い同様の課題が存在する。例えば CERN の大型粒子加速器 LHC では、データ選別後のデータ量が年間 15 PB に達するといわれる。LHC の場合は、世界各地の計算機資源とストレージ資源をデータグリッドに組み込むことにより [1]、大量のデータへのアクセスを可能にしているが、一般にこれほど大規模なシステム構築は多額の費用を要するため不可能である。また装置の高性能化や汎用計算機の高性能化により、超大型観測に限らず中・小規模の実験・観測・計算科学おいても、データ量の増大は無視することができなくなっている。このように多くの分野にとって、大量のデータを安全に長期間保管し、効率良く処理することは、そのプロジェクトの重要な部分を占めている。

2　ノード間通信技術

最新のスーパーコンピューターランキング TOP500 によると、ノード間通信に用いられる技術は、Gigabit Ethernet(1 G) が 25 %、10Gigabit Ethernet(10 G) が 15 %、Infiniband が 44 %、それ以外は各メーカー独自の通信技術が占めている [2]。このように Infiniband[3] はこれまで HPC 分野で広く用いられてきたが、その特徴は主に広い通信帯域と低遅延性である。これは RMDA(Remote Direct Memory Access) により遠隔メモリに直接アクセスすることで、OS は通信に関する部分を可能な限り CPU を介さず行うことで実現されている。Infiniband ネットワークは、主として HCA(Host Channel Adapter) と呼ばれるインターフェースカードとスイッチで構成される、いわゆるスイッチ型ファブリックによるネットワークである。Infiniband で接続される計算機器には HCA が用いられるが、本報告での性能試験でも各ファイルサーバに HCA を用いた。Infiniband ネットワークでは、主に IPoIB 方式と RDMA 方式のどちらかでデータの送受信が行われるが、IPoIB は RDMA と異なりデータ送信用のバッファを用意する必要がある。なお本報告で行なったストレージ性能評価では IPoIB を用いている。HCA 等は infiniband スイッチにより 1 つのサブネット内で、アドレスやルーティングと共にサブネットマネージャーと呼ばれるサービスによって管理される。

IBTA(Infiniband Trade Association) による仕様 Infiniband 1.2.1 では SDR(2.5 Gbps)、DDR(5 Gbps)、QDR(10 Gbps) が、Infiniband1.3 では FDR(14 Gbps)、EDR(26 Gbps) が決められている。また IBTA が公表しているロードマップ

(図.1) によると 2017 年までに 1 レーン当たりの帯域が 50 Gbps の HDR が予定されている。Infiniband1.2.1 と Infiniband1.3 の仕様上の大きな違いはエンコード方式である。Infiniband1.2.1 では、8b/10b が用いられていたが、Infiniband1.3 から 64b/66b が用いられることにより、効率が 80 % から 97 % に上がっている。本報告で用いているインフィニバンドは最新の FDR(図.2) である [4]。

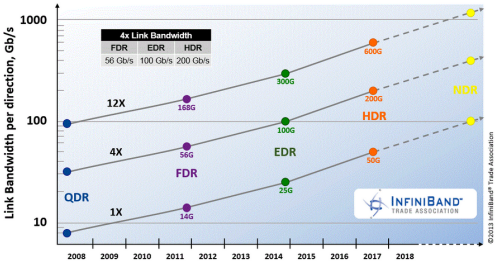

図 1　IBTA による Infiniband のロードマップ。

図 2　今回の評価試験で使用した Mellanox 社の Infiniband FDR 製品。

これまで筆者が扱ってきた Infiniband(DDR、QDR、FDR) の、RDMA を用いたノード間通信帯域 (双方向) の性能比較データを図.3 に示す。ベンチマークテストは米オハイオ州立大学の NBCL(Network-Based Computing Laboratory) の OSU Micro-benchmarks を用いた [5]。

また、図.4 と図.5 では、Gigabit Ethernet と Infiniband FDR(IPoIB) の通信帯域および通信遅延時間の性能比較それぞれ示している。

3　分散ファイルシステム

近年、インターネット上で生成される文書・画像・動画などといったデータ量の増加は著しく、データセンターにはそれに対応できる高い拡張性と効率的な保管に優れたストレージシステムが求められている。同様に、データ量の急激な増加は、HPC 分野でも起こっており、計算性能・通信性能への要求が高い分、一般的なデータセンターよりも状況は深刻な場合がある。

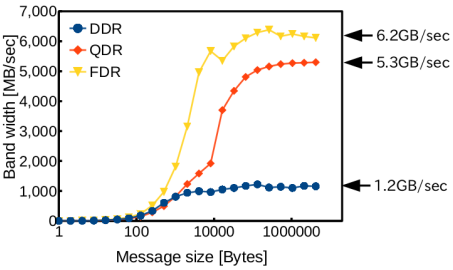

図　3　Infiniband　DDR,QDR,FDR　の RDMA を用いた場合の双方向通信性能。

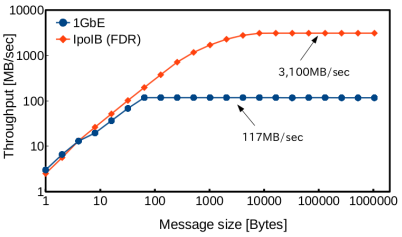

図 4　Gigabit Ethernet と IPoIB による Infiniband FDR の通信制能比較。

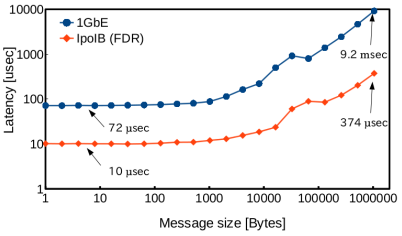

図 5　Gigabit Ethernet と IPoIB による Infiniband FDR の通信遅延時間の比較。

このような問題の解決策の 1 つとして、分散ファイルシステムが広く採用されている。現在、世の中には数多くの分散ファイルシステムが存在するが、その代表が Luster File Sysytem[6] である。Lustre はその高性能さから既に多くの HPC 分野で分散ストレージとして利用されている。本報告では、Gluster[7] を用いた分散ファイルシステムの性能評価を行なった。Gluster も代表的な分散ファイルシステムであるが、メタデータサーバがないことが特

徴であり、これによるシングルポイント障害が発生しない点が Lustre との大きな違いである。Gluster も他の分散ファイルシステムと同様に、拡張性が高く数 PB まで拡張が可能であり、既に数 PB 規模での商用利用の実績がある。システムの運用面では、ノードの追加や容量拡張を無停止で行うことができ、障害発生時にも運用を継続しながら復旧作業が可能であるなど利点が多い。ノード間通信には、Infiniband を使用することができ、RDMA を用いることで広帯域で低遅延のシステムが構築できる。また、TCP/IP によるシステム構築も可能で Gigabit Ethernet や 10Gigabit Ethernet、IPoIB を用いることができる。Gluster には、様々なファイル分散方式があり、ファイル実体を複数のストレージに分散する「distributed」(今回はこれを採用)、分散と同時にレプリケーションを作成する「distributed-replicate」、冗長性は無いが高速性能を実現する「distributed-stripe」、拠点間でのレプリケーションが可能な「distributed-replicate + Geo-replication」などが利用できる。

4 Gluster ファイルシステム性能試験

今回構築した分散ファイルシステム (図.6) は、図.7 に示すように各 40TB の RAID 領域を持つファイルサーバ 3 台を Infiniband FDR で接続し、IPoIB を用いた Infiniband ネットワーク上で、120TB の単一ボリュームとして構成されたファイルシステムである。このボリュームは、全てのサーバから同等に見えており、これに対して I/O 性能の評価を行なった。各サーバには、10 core の Intel Xeon CPU 2.5 GHz と 64 GB が搭載されている。

ファイルシステムの性能指標として、主に Throughput [MB/s] と IOPS の測定を実施した。用いた性能評価プログラムは、fio、iozine[10]、IOR[8],[9] である。図.8 に示すように、fio を用いた 4 KB 単位の IOPS 性能をローカル RAID と Gluster ボリュームとで比較している。これによると、分散ファイルシステムが苦手としているブロックサイズの小さい I/O 性能がよく現れている。次に、4 kB と 512 kB の I/O 単位について Random I/O

図 6　性能評価試験で構築した分散ファイルシステム。

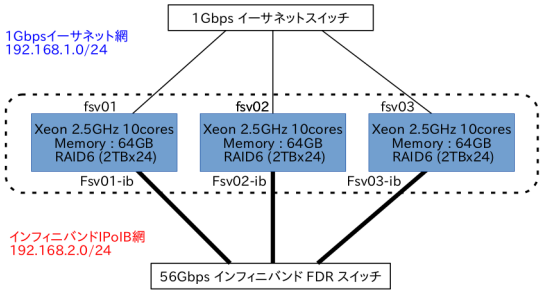

図 7　Gluster 分散ファイルシステムのネットワーク構成。

と Sequential I/O の Throughput 性能を実測した結果を図.9 に示す。また、I/O 単位を 1 MB として 128 GB の大きなファイルについて Throughput 性能も図.10 に示す。これらの結果から、比較的大きな I/O 単位で高い性能が出ていることが分かる。

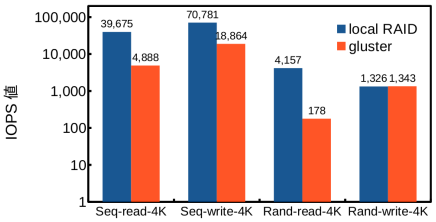

図 8　fio による IOPS 性能。5GB のファイルを 4kB のブロックサイズで読み書き。

ここまで用いた fio、iozone は単独サーバ上で行う性能評価プログラムであるが、現実のシステムを想定し、全サーバからの I/O 性能評価を行なうた

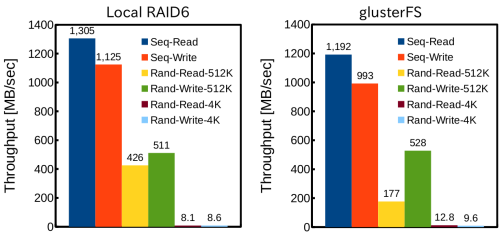

図 9　fio による Throughput 性能。5GB のファイルを 4kB および 512kB ブロックサイズで読み書き。

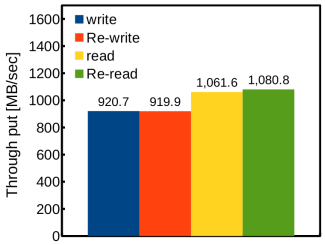

図 10　iozone を用いた Throughput 性能。128GB のファイルを 1MB のブロックサイズで読み書き。

めに、IOR を用いた MPI 並列プロセスからの I/O 性能の評価を行なった。IOR で生成した I/O プロセスは各サーバ 10 プロセスで、プロセス数を 1 から 30 まで順次増やしながら性能を測定した。図.11 によると、Write 性能は安定して約 2.5 GB/s 出ているのに対し、Read 性能は 20 プロセスあたりから約 1 GB/s に下がっている。

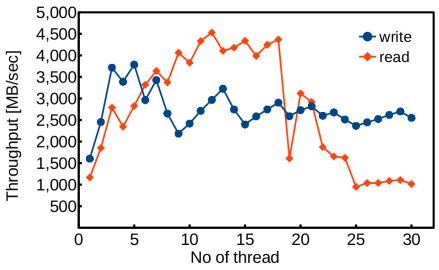

図 11　MPI による並列プロセスからの I/O 性能。

5　まとめ

本報告では、大規模データの高速ストレージを念頭に、Infiniband FDR を用いた Gluster 分散ファイルシステムを一つの解決策として紹介した。性能評価にさざまな I/O パターンを試み、Sequential I/O では常時 1 GB/s 程度の良好な性能を示すことが確認できた。また分散ファイルシステムの性質として、ランダム I/O に性能劣化が生じると予想されたが、本報告で実施した性能評価でも確認された。ただしメタデータサーバのない Gluster を用いていることを考慮すると、今回構築したシステムは総じて良い性能を示したと思われる。今後は、実際の利用に近い環境で性能評価を行う予定である。

6　謝辞

本報告で使用したファイルサーバと Infiniband FDR は、平成 25 年度国立天文台大学支援経費で購入したものである。またファイルサーバで使用したハードディスクは、国立天文台天文シミュレーションプロジェクトから寄贈頂いたものである。

参考文献

[1] http://wlcg.web.cern.ch/
[2] http://www.top500.org/
[3] Solving the NVGRE Performance Challenge, October 2013, Mellanox technologies
[4] Infiniband Trade Association, http://www.infinibandta.org/
[5] http://mvapich.cse.ohio-state.edu/benchmarks/
[6] http://wiki.lustre.org/index.php/Main_Page
[7] An Introduction to Gluster Architecture, versions 3.1.x, white paper, www.gluster.com
[8] IOR: I/O Performance Benchmark
[9] Using IOR to Analyze the I/O performance for HPC Platforms, H. Shan and J. Shalf, CUG Proceedings 2007.
[10] Iozone Filesystem Benchmark, W.D. Norcott, http://www.iozone.org