L-025

インターネット回線下での”SRFS on Ehter”の性能評価

大川 博文† 藤田 直行†
Hirofumi Ookawa Naoyuki Fujita

1. まえがき

近年、SuperSINET やつくば WAN 等のように、インターネットバックボーン網が 10Gbps の帯域を持つようになり、また、ギガビットイーサネット(GbE)技術の普及に伴い、端末用の GbE インタフェイスも非常に安価になるなど、ネットワーク利用環境は急激に広帯域化している。

一方、HPC システムの処理性能の向上によって、その処理対象は巨大化、細密化しており、出力ファイルサイズは非常に巨大なものとなっている。

しかしながら、UNIX 上での遠隔ファイルシステムとして一般的である NFS(Network File System)では、ウィンドウサイズなどのチューニング無しには、その広帯域を十分に利用しきれないため、遠隔地からの HPC システムの利用はいまだ困難なままである。

本研究所の有する HPC システムでは、ノード間を接続するクロスバネットワーク上で高速な分散ファイルシステムを実現するための仕組みとして、SRFS(Shared Rapid File System)を用いている。これをベースとしてイーサネット対応を加えたものが、”SRFS on Ehter”である。[1][2]

今回、遠隔ファイルシステムとして”SRFS on Ether”を用い、宇宙航空研究開発機構(以下、JAXA(調布))と防災科学研究所(以下、防災研(つくば))間の、インターネット接続回線下における、ファイル IO 性能の測定を行った。

本稿では、その実験環境と測定方法を述べ、”SRFS on Ether”の性能を評価する。

2. 実験環境

実験を行ったネットワーク環境を図 1 に示す。

JAXA(調布)は、SuperSINET ルータが収容されている三鷹国立天文台に、1Gbps の商用 Ehternet サービスで接続しており、SuperSINET のバックボーンは 10Gbps の帯域を有している。一方、防災研(つくば)はつくば WAN に参加しており、つくば WAN のバックボーンも同じく 10Gbps の帯域を持つ。SuperSINET とつくば WAN は 1Gbps で相互接続している。

両機関ともにインターネット収容ルータから端末まで、1000Base-SX または 1000Base-T の媒体により接続した。

実験環境として、端末にグローバルアドレスを設定してインターネットに直接接続する環境と、実利用環境を想定して、プライベートアドレス空間を VPN ルータを介して接続する環境とを設定した。

両機関に設置した端末の役割とスペックを表 1 に示す。

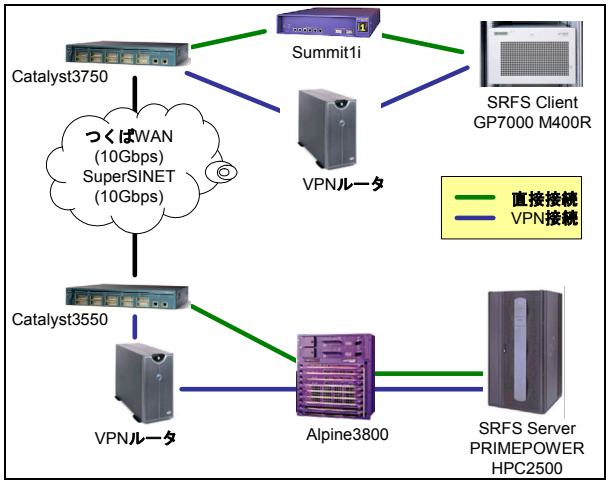

図 1 ネットワーク構成

表 1 サーバ／クライアントスペック

	SRFS Server	SRFS Client
機種名	PRIMEPOWER HPC2500	Fujitsu/PFU GP7000 M400
CPU	SPARC64-V x 8	SPARC64-III x 4
CPU Clock	1.3 GHz	300 MHz
System Clock	260 MHz	74 MHz
Memory	16384 MB	3072 MB
OS	Solaris8	Solaris8
networkI/F	1000Base-SX	1000Base-T

3. 測定結果

(1) トラフィックジェネレータを用いた測定(UDP)

トラフィックジェネレータ(NXS100G)を用いて UDP スループットを測定した。表 2 に直接接続下での測定結果を示す。

表 2 NXS100G の測定結果

frame size [Bytes]	防災研(つくば) →JAXA(調布) [Mbps]	JAXA(調布)→ 防災研(つくば) [Mbps]
64	923	925
128	973	951
256	984	959
512	988	964
1024	981	967
1280	986	969
1518	983	972

†独立行政法人 宇宙航空研究開発機構,JAXA

ネットワークの素性能を測定するため、測定値には Ethernet のオーバーヘッド(ヘッダ情報)を含めた。

機器のインタフェイスの問題で、VPN 接続下での測定は行えなかった。

(2) TCP スループットの測定

Solaris8 標準の ping、traceroute コマンドを用いて、RTT を測定した。また、ネットワーク性能測定ツールとして広く用いられている netperf を用いて TCP スループットを測定した。ファイル IO 性能の測定では、OS の TCP ウィンドウサイズを 128KB としているため、netperf での測定でも受信ソケットサイズおよび送信ソケットサイズのオプションを 128KB と固定し、メッセージサイズをパラメータとして測定を行った。

RTT の測定結果は、直接接続下で平均 3.3 [msec]、VPN 接続下では、平均 4.0 [msec] であった。「スループット = ウィンドウサイズ / RTT」の関係より、理論上のスループットとしては、それぞれ、310[Mbps]、256[Mbps]となる。

netperfによる実測結果では、それぞれ最大、252.6[Mbps]、146.8[Mbps]であった(図２、図３)。

(3) ファイル IO 性能の測定

NFS、”SRFS on Ether”の両ファイルシステムにおいて、IO 長をパラメータとして、read/write のシステムコールを計測した。NFS には noac オプションを設定した。

直接接続下と VPN 接続下のそれぞれのネットワーク構成におけるファイル IO 性能を、図２、図３に示す。

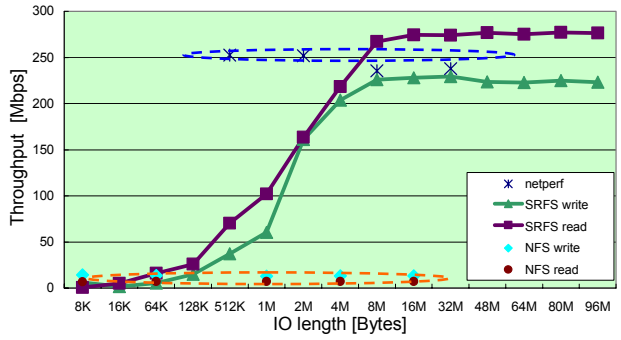

図２　　直接接続下での IO 性能

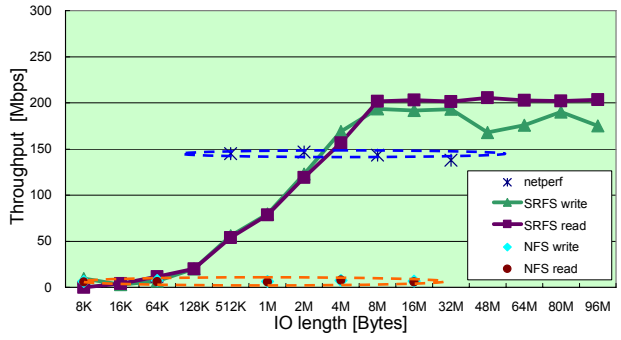

図３　　VPN 接続下での IO 性能

4. まとめ

トラフィックジェネレータによる UDP 性能の測定結果と、RTT の測定結果より、JAXA(調布)ー防災研(つくば)間のネットワーク環境は、媒体性能にほぼ達するほどの広帯域が確保されており、また、ネットワーク的にも WAN としては比較的近い位置にある(RTT=3～4 [msec])と言える。

しかし、ファイル IO 性能の測定では、NFS を利用した場合 10-20Mbps 程度にとどまった。これは、TCP ウィンドウサイズのチューニングを行っていないためでもあるが、noac オプションを設定したために、属性情報の交換のオーバーヘッドが如実に現れたと考えられる。

対して”SRFS on Ehter”では、インターネットに直接接続した環境で、200-260Mbps、VPN 接続下を通す環境下でも 180-200Mbps の性能を計測している。

ここで、スループットが netperf の結果を上回ったのは、”SRFS on Ether”が IO 長によって自動的に行っている多重化が有効に機能しているためである(TCP ウィンドウサイズを 1MB とし、最大 8 多重と設定している)。

今後、”SRFS on Ehter”を HPC システムの遠隔利用のための共有ファイルシステムとして展開していくために、Solaris8 以外のプラットフォームへの対応を求めるとともに、「広帯域であるが遅延の大きなネットワーク環境(Long Fat-pipe Network,LFN)」などでの実証が必要である。

5. 謝辞

本実験の実施に際し、実験環境の構築、提供をいただいた防災科学研究所の佐藤様に深く感謝いたします。また、情報提供、実験作業にご協力いただいた富士通株式会社の矢澤様、甲斐様ならびに皆様に深謝いたします。

参考文献

[1] Naoyuki Fujita and Hirofumi Ohkawa, ”Storage Devices, Local File System and Crossbar Network File System Characteristics, and 1 Terabyte File IO Benchmark on the “Numerical Simulator III” ‘, 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies, 2003

[2] Naoyuki Fujita, ‘Shared Rapid File System on Ethernet’, http://romulus.gsfc.nasa.gov/msst/conf2004/index.html, 2004