パケットアナライザ:DR Giga-Analyzer

中野理[1]　鳥居健一[1]　吉田昇一[1]　柳沢敏孝[1]　水口健二[1]　生田祐吉[1]　陣崎明[2]
下見淳一郎[2]　玉造潤史[3]　中村誠[4]　稲葉真理[5]
平木敬[5]

1) (株) 富士通コンピュータテクノロジーズ 神奈川県川崎市上小田中 4-1-1
2) (株) 富士通研究所 神奈川県川崎市上小田中 4-1-1
3) 東京大学大学院理学系研究科 東京都文京区本郷 7-3-1
4) 東京大学情報基盤センター 東京都文京区弥生 2-11-16
5) 東京大学大学院情報理工学系研究科 東京都文京区本郷 7-3-1

概 要

DR Giga-Analyzer は 1 Gbps の通信に対し高精度タイムスタンプを付加したログを 2 時間以上にわたり収集可能なパケットアナライザであり、通信の両端点のログをつきあわることで遠距離 2 点間の TCP/IP 通信の振る舞いを解析することを目的としている。本稿では、 DR Giga Analyzer のシステムデザインおよびハードウェア・ソフトウェア構成について延べる。

1　はじめに

近年ネットワークの高速化は目ざましく、計算機のネットワークインターフェイスとしてはギガビット・イーサネット (GbE) が主流となってきている。国内では SuperSINET に代表されるように 10 ギガビット・イーサネット (10GbE) が大域ネットワークのバックボーンとして採用され、また海底ケーブルの整備により日米あるいは日韓間においても 600Mbps 〜 2.4 Gbps 程度の帯域をもつ通信は日常的なものとなりつつある。

我々は、遠距離高帯域のネットワークを十分に活用し実験・観測科学研究プロジェクトが巨大データを遠隔研究施設間で共用することを目標とするデータレゼボワール・システムを提案し、プロトタイプモデルを実装、性能評価を行ってきた [1, 2, 3, 4, 5]。データレゼボワール・システムは、ストライプして格納した巨大データを iSCSI/パラレルストリーム転送することにより、高速・高バンド幅転送を実現しているが実際のネットワークで性能を計測すると予備実験より実際の性能が悪いことが観察された。たとえば、2 ポートプログラマブル ネットワークインターフェースカード Comet i-NIC(COMmunication Enterprising Technology intelligent-NIC. http://www.comet-can/jp/) を GbE L2 ブリッジとして用いて、入力パケットをバッファ上で遅延し指定した確率でパケットを破棄することで通信遅延及びパケット損失のある 遠距離高速ネットワーク環境をエミュレートするシステムを作成したが、現実のネットワークでの性能は、疑似遠距離ネットワークに比較し、平均性能が悪く、また性能のばらつきも非常に多かった。

実際のネットワークでのふるまい、特に 超高速 TCP/IP 通信におけるパケットロスの影響を詳細に調べるためには両端で別々に記録したパケットログの対応をとる必要があるため標準時刻同期タイムスタンプを付加することのできるパケットアナライザを作成することにした。DR Giga-Analyzer は以下の要件を満たす。

- 標準時刻同期のタイムスタンプをパケットログに付加する事
- 100nsec タイムスタンプ解析度をもつ事
- 1Gbps のフルワイヤキャプチャが可能である事

- 2 時間以上キャプチャ可能である事
- キャプチャデータ操作が容易である事
- 汎用品を使うことで製作コストを下げる事

本稿では、DR Giga-Analzyer のシステム構成、諸元、適用範囲について延べる。

2 構成

DR Giga-Analyzer の構成を図 1 に示す。

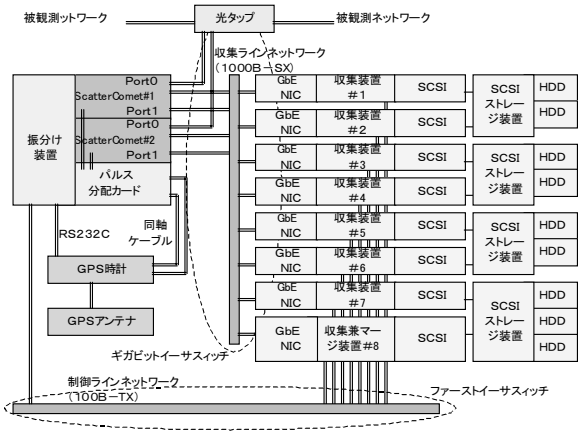

図 1: DR Giga-Analyzer の構成

DR Giga-Analyzer は、1 台の振分け装置と 8 台の収集装置、および光タップ、 GPS 時計、GbE スイッチから成る。また、振分け装置には、Scatter Comet と呼ぶ Comet i-NIC を搭載する。なお、振分け装置本体および収集装置本体は汎用の IA サーバを採用し、OS は Linux 2.4.18、ファイルシステムは EXT2(,3) を採用している。

2.1 フルワイヤーレートキャプチャー

本 DR Giga-Analyzer の対象とするメディアはギガビットイーサーネット (1000B-SX) である。即ち、フルワイヤレート双方向で 2Gbps、 最短パケット (64Byte 長) では 2,976,000pps 以上のキャプチャ能力が必要となる。現行市販されているサーバ性能単体では、この条件を満足するデータ格納は不可能であるため、収集側に 8 台のサーバを用意し、一旦、パケット振り分け装置がパケットを受け取り、タイムスタンプを付加したうえで収集用サーバに振り分け分配、収集用サーバが受けとったデータを保存することで分散収集格納する手法を採用することとした。具体的には被観測ネットワーク上の双方向 (上り/下り) のパケットは、光タップにより分岐され、それぞれ、振り分け装置の ScatterComet に受信される。Scatter Comet 内部には、予め各収集装置宛ての UDP カプセリングヘッダを用意しておき、ラウンドロビン方式で受信パケットを 8 台の収集用サーバにフォワーディングする。収集用サーバは、受信したフォワーディングパケットからカプセリングヘッダを取り除き、ディスク装置に書き込んで行く。

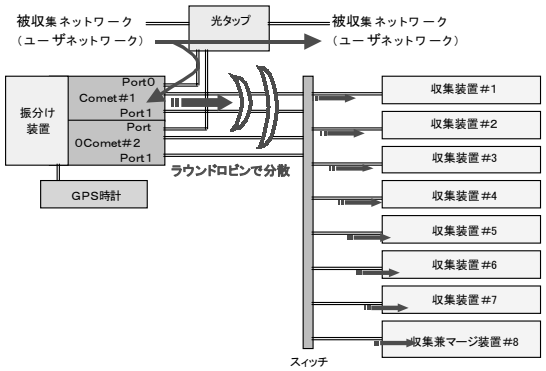

図 2: キャプチャ動作

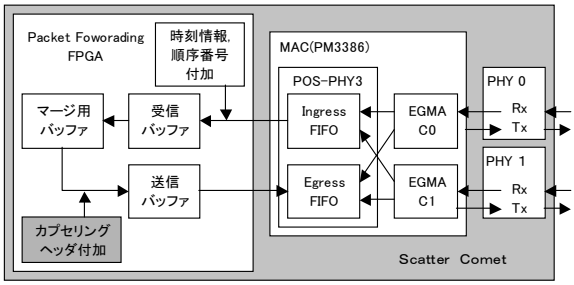

図 3: Scatter Comet ブロック図

上記基本設計に加え、安定したフルワイヤ収集、特に連続した短パケットの収集を実現するために以下の手法を採用した。

Interrupt Coalescing

従来の NIC ではパケット受信毎に割込みが起こっていたが高速ネットワーク用 NIC では高速トラフィックによるシステムの負担を減らす為、

複数パケットの送受信を1回の割込みで通知する“Interrupt Coalescing” が使われるようになってきている。本装置でも この Interrupt Coalescing を採用している。

フォワーディングパケットのマージ

上記 Coalescing パラメタをチューニングしても、短パケット収集性能は、64Byte 時で 60Mbps,122,700pps 程度であった。そこで、Scatter Comet から収集装置にフォワーディングする際、短パケットであれば複数パケットをマージしてフォワーディングする事により、収集装置側の受信パケット数を減らすことにした。

他、安定した収集性能を確保する為、長時間占有するプロセスの平滑化等のチューニングも実施している。また、長時間のキャプチャリングに対しては、第一に汎用のサーバとディスク装置を用いる事により、スケーラブルな最大キャプチャ時間の展開が可能な構成としている。また、ヘッダトレースモードを用意しており、パケットのフルキャプチャの他、受信したパケットの先頭から 64Byte, 128Byte, 256Byte のみをキャプチャできる。ディスク容量 2TB では、フルパケットトレースで約 2.3 時間、64Byte ヘッダトレースでは最大 40 時間超のフルワイヤレートキャプチャが可能となっている。

2.2 タイムスタンプ精度

ギガビットイーサネット環境では、1500Byte 長パケットは 12 μ sec で、64Byte 長パケットであれば 0.5 μ sec 間隔で送受信される。 また、異なる 2 点間のネットワークトラフィック解析を行なう場合、2 点間で同期した時刻関係が把握できる事が望ましい。これらを満足する為、DR Goga-Analyzer では GPS 時計が出力する標準時刻と同期したタイムスタンプを付加する。

GPS 時計:LS-20K はその内部に 1ppm 水晶を持ち、標準時刻に対する精度は、GPS 衛星捕捉時、カタログ値 10 μ sec, 実精度 100nsec を有する。また、Scatter Comet 内には 40MHz/50ppm 水晶を持ち、ソフトウェア的に設定されたある時点の標準時刻をこの自クロック (40MHz/50ppm) で計時している。ここで、GPS 時計出力の標準時に同期した 1msec 間隔

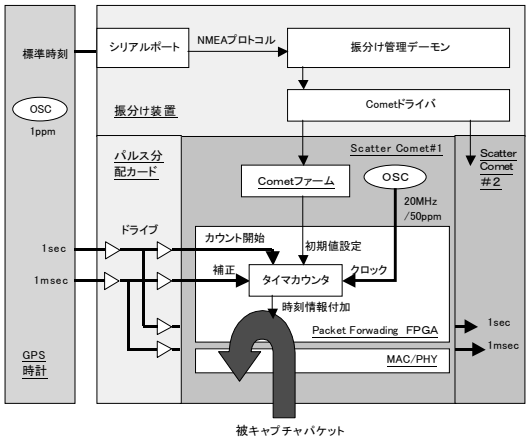

図 4: タイムスタンプの付加

パルスを Scatter Comet に接続し、このパルスにより Scatter Comet 内時刻を補正させる事により、標準時と同期した高精度時刻情報を有する事が可能である。Scatter Comet が受信パケットを受け付けると、その高精度時刻情報は専用ハードウェアにより付加される。また、受信時の到着ジッタを抑える為、受信回路 (MAC) 近端、受信バッファ動作の影響を受けない位置で時刻情報の付加を行なう様、配慮している。これらの構造により、標準時刻に対する実精度 100nsec、計時解析度 25nsec のタイムスタンプを可能としている。

2.3 分散記録されたパケット情報のマージ

各収集装置で分散して記録されたパケット情報は、時刻情報および別途付加されている受信順序番号により整順され、1 つのストリームデータとして結合 (マージ) され出力される。

2.4 tcpdump フォーマット出力

キャプチャされたパケットデータは上述の高精度時刻情報とともに、tcpdump フォーマットで出力される。これにより、様々な tcpdump 出力形式の恩恵を得る事が出来る。また、μsec 精度でのパケット送信時刻の観測により、ミクロな振舞を解析可能となっている。

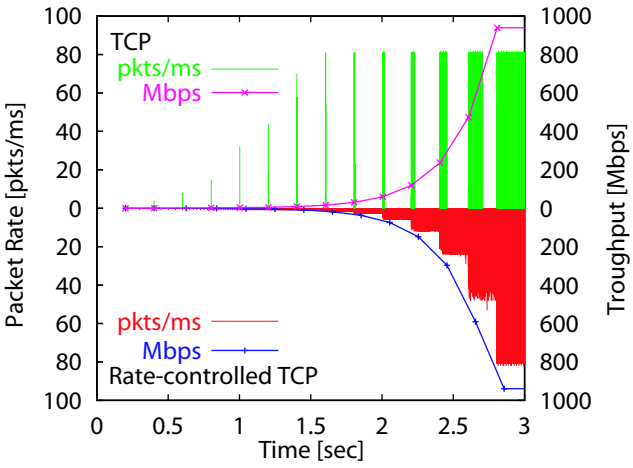

図 5: TCP/IP 通信の解析例

3 諸元

DR Giga-Analyzer は以下に示す 2 タイプが作成されている。初代機が専用のディスク装置を有し大容量化していたのに対し、次代機は、収集装置に 1U サーバ/73GB HDD 搭載可能なサーバを選択し、小型化を図っている。

		初代機	次代機(小型化)
構成機器	振分け装置	**Fujitsu Primergy C200** (PentiumIII 1.26GHz, 512MB x2, Scatter Comet x2 搭載, RedHat 7.3(Kernel 2.4.18), EXT2)	**DELL PowerEdge2650** (Xeon 2.8GHz, 256MB x2, Scatter Comet x2 搭載, RedHat 7.3(Kernel 2.4.18), EXT3)
	収集装置	**DELL PowerEdge1650 x8** (PentiumIII 1.4GHz x2, 512MB x2, Intel PRO1000/MF 搭載, RedHat 7.3(Kernel 2.4.18), EXT2) **DELL PowerVault 220S x4** (HDD 計 2TB)	**Appro 1224SX x8** (Xeon 2.0GHz x2, 512MN x2, Intel PRO1000/MF 搭載, HDD 18GB x4, RedHat 7.3(Kernel 2.4.18), EXT3)
	光タップ	**昭和電線電纜 マルチモード光カプラタップ**(分岐比 50:50)	
	GPS 時計	**白山工業 LS-20K**(カタログ値 10μsec 精度, 実精度 100ns)	
	GbE Switch	**Extreme Networks Summit 5i**	
	FE Switch	**Extreme Networks Summit 24e2**	
キャプチャ性能		ギガビットイーサネットフルワイヤキャプチャ	
キャプチャ容量		2TB	384GB
最大キャプチャ時間		フルワイヤレート時 Full Trace 2.3 h 64B Header Trace 最大 41.9 h	フルワイヤレート時 Full Trace 0.46 h 64B Header Trace 最大 8.29 h
サイズ、電源容量		36U 強, 最大定格計 5395W	13U 強, 最大定格計 3744W

図 6: 諸元

4 まとめ

DR Giga-Analyzer は前述の通り、GPS より高精度な一意な時刻を得られる為、大陸間ネットワークデータ通信に於いてもエンド-エンドの TCP/IP の振る舞いを十分に解析する事が出来、高速ネットワーク基盤整備の検証及びボトルネック解析等に活用されている。

5 謝辞

本研究は文部科学省科学技術振興調整費先導的研究基盤整備「科学技術研究向け超高速ネットワーク基盤整備」および科学技術振興事業団 CREST による研究領域「情報社会を支える新しい高性能情報処理技術」研究課題「ディペンダブル情報処理基盤」で補助された。

参考文献

[1] K. Hiraki, M. Inaba, J. Tamatsukuri, R. Kurusu, Y. Ikuta, H. Koga, A. Zinzaki, “Data Reservoir: Utilization of Multi-Gigabit Backbone Network for Data-Intensive Research”, SC2002, Nov. 2002. http://www.sc-2002.org/paperpdfs/pap.pap327.pdf

[2] R. Kurusu, M. Sakamoto, Y. Ikuta, K. Hiraki, M. Inaba, J. Ta matsukuri, H. Koga, A. Zinzaki, “Data Reservoir, Multi-Gigabit Data Transfer Facility, Its Design and Implementation”, Proc. PDCAT, pp. 100-108, Sept. 2002.

[3] K. Hiraki, M. Inaba, J. Tamatsukuri, R. Kurusu, Y. Ikuta, H. Koga, A. Zinzaki, “Data Reservoir: A New Approach to Data-Intensive Scientific Computation”, Proc. ISPAN, pp. 269-274, May 2002.

[4] M. Nakamura, M. Inaba, K. Hiraki, “Fast Ethernet is sometimes faster than Gigabit Ethernet on LFN — Observation of congestion control of TCP streams”, Proc. PDCS, pp. 854-859, Nov. 2003.

[5] M. Nakamura, M. Inaba, K. Hiraki, “End-node transmission rate control kind to intermediate routers towards 10Gbps era”, PFLDnet 2004, Argonne, IL, Feb. 2004.