広域・広帯域ネットワークにおける IP-SAN プロトコルのためのスループット最大化機構

Automatic Parallelism Tuning Mechanism for IP-SAN Protocols in Long-Fat Networks

西島 孝通 [1] 井上 史斗 [2] 大崎 博之 [2] 今瀬 真 [2]
Nishijima Takamichi Fumito Inoue Hiroyuki Ohsaki Makoto Imase

大阪大学 基礎工学部 情報科学科 [1]
Department of Information and Computer Sciences, School of Engineering Science, Osaka University
大阪大学 大学院情報科学研究科 情報ネットワーク学専攻 [2]
Department of Informatiion Networking, Graduate School of Information Science and Technology, Osaka University

1 はじめに

近年、ネットワークの広帯域・広域化に伴い、IP ネットワーク上で SAN を構築する IP-SAN が広まっている。現在、IP-SAN を実現するプロトコルとして、iSCSI や iFCP など複数のプロトコルが使用されている。これらの IP-SAN プロトコルでは、実際のデータ転送は TCP によって行われている。しかし TCP には、広帯域・広域ネットワークにおいて、スループットが低下するという既知の問題がある。これまで、IP-SAN プロトコルのスループット向上手法はいくつか提案されているが、それらは個別の IP-SAN プロトコルごとの対処法であった。しかし、スループットの低下は IP-SAN プロトコル自体の問題ではなく、トランスポート層プロトコルである TCP 自体の問題である。このため、IP-SAN プロトコルに依存しない、汎用的な解決法が望ましいと考えられる。

2 ブロックデバイスレイヤにおける並列 TCP コネクション数調整機構 BDL-APT

本稿では、IP-SAN プロトコルに依存しないスループット向上手法として、ブロックデバイスレイヤにおける並列 TCP コネクション数調整機構 BDL-APT (Block Device Layer for Automatic Parallelism Tuning) を提案する (図 1)。BDL-APT は、オペレーティングシステムにおけるブロックデバイスと、IP-SAN プロトコルの中間のレイヤで動作する。ネットワーク環境に応じて、並列 TCP コネクション数 (= IP-SAN プロトコルのセッション数) を調整することにより、IP-SAN のスループットを最大化する。

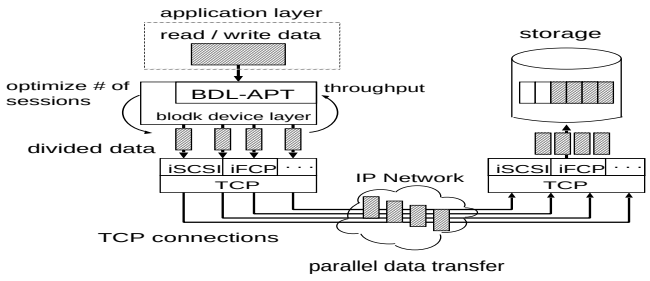

図 1: ブロックデバイスレイヤにおける並列 TCP コネクション数調整機構 BDL-APT の概要

BDL-APT は、ブロックデバイスレイヤにおいて、ストレージに対するストレージアクセス要求を分割し、複数の IP-SAN セッションに分割する。この時、IP-SAN セッションの多重度を調整することにより、結果として、使用される並列 TCP コネクション数が調整されることになる。並列 TCP コネクション数の調整方法は、既存の並列 TCP コネクション数調整機構 APT (Automatic Parallelism Tuning) [1] をそのまま利用する。それぞれの IP-SAN セッションにおける、ストレージアクセス時の応答時間およびスループットを計測し、これらの情報をもとに使用する IP-SAN セッション数を調整する。

3 性能評価

BDL-APT を、Linux オペレーティングシステムの MD (Multiple Disks) レイヤの一つとして実装した。MD とは、Linux オペレーティングシステムにおけるソフトウェア RAID の一実装である。

実験では、IP-SAN プロトコルとして、Linux オペレーティングシステムの NBD (Network Block Device) を用いた。NBD サーバと NBD クライアントを、ネットワークエミュレータ (dummynet) を介して接続した。NBD クライアントから NBD サーバに対して連続的にデータ転送を行い、その時の NBD セッション数およびスループットの時間的変動を計測した。NBD サーバおよびクライアントの TCP ソケットバッファサイズを 512 [Kbyte] とした。ネットワークエミュレータの帯域を 1 [Gbit/s]、遅延を 10 [ms] とした。

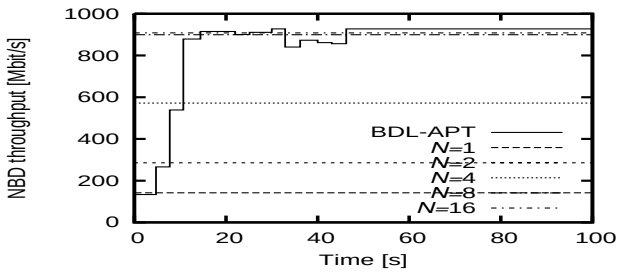

図 2: 実装した BDL-APT の結果: IP-SAN プロトコルとして NBD を用いた時のスループットの時間的変動

図 2 に、BDL-APT を用いた時のスループットの時間的変動を示す。比較のため、NBD セッション数を 1、2、4、8、16 にそれぞれ固定した時のスループットもあわせて示している。この結果より、転送開始から 45 [s] 程度で、BDL-APT の NBD セッション数が最適化され、スループットが 927 [Mbit/s] に収束していることが分かる。なお、最適化後の NBD セッション数は 11 であった。これより、BDL-APT は、ネットワークの状況に応じて NBD セッション数を調整することにより、ネットワーク資源を有効に利用できていることが確認できた。

参考文献

[1] T. Ito, H. Ohsaki, and M. Imase, “GridFTP-APT: Automatic parallelism tuning mechanism for GridFTP in long-fat networks,” *IEICE Transactions on Communications*, vol. E91-B, pp. 3925–3936, Dec. 2008.