IP-SAN 統合 PC クラスタにおける NPB 実行性能比較

Comparison of NPB Execution Performance on IP-SAN consolidated PC Cluster

神坂 紀久子 † 山口 実靖 ‡ 小口 正人 † 喜連川 優 §
Kikuko Kamisaka Saneyasu Yamaguchi Masato Oguchi Masaru Kitsuregawa

1. はじめに

近年，大規模な科学技術計算やデータマイニング処理等は，ハードウェア性能の飛躍的向上と低価格化により，分散メモリ型の並列計算機である PC クラスタにおいて実行することが一般的になった．

一方，計算機で処理されるデータ量の急激な増大に伴い，ストレージを統合する技術である SAN(Storage Area Network) が注目を集めている [1]．PC クラスタにおけるストレージを SAN で接続することによって，分散されたストレージに存在する大容量データを容易に管理することができる．

SAN を使用した一般的な PC クラスタでは，クライアントとサーバ間のフロントエンドとディスクアクセスを行うバックエンドのネットワークが分離されている．そこで，分離して存在するそれらのネットワークを統合した IP-SAN 統合 PC クラスタを実現することによって，運用管理負荷およびネットワーク構築コストを軽減することが可能になる．しかし，双方のネットワークを統合すると，ストレージアクセスとノード間通信は，同一の IP ネットワーク経由でデータが転送されるため，ストレージアクセスのバルクデータにより，ノード間通信が多大な影響を受け，並列分散処理実行時の性能が大幅に劣化する可能性がある．

そこで，本稿では，IP-SAN 統合 PC クラスタを構築し，フロントエンドトラヒックとバックエンドトラヒックの統合が PC クラスタの性能に与える影響を，I/O 処理を伴う並列計算ベンチマークを用いて評価した．

2. IP-SAN 統合 PC クラスタ

従来，PC クラスタを含む並列計算システムは，各サーバにストレージが直接接続された DAS を使用した PC クラスタが広く使用されていた（図 1）．しかし，DAS はディスクごとにデータが分散されているため，ストレージを効率良く使用することができず，管理も容易でない．

それに対し，近年 PC クラスタは，ストレージ統合を目的として，高速な Fibre Channel を使用した FC-SAN が用いられるようになった（図 2）．しかし，FC を用いる SAN (FC-SAN) では，FC 用のスイッチやインタフェースなどのハードウェアが高価であり，導入，管理コストがかかるという問題がある．特に大規模な PC クラスタを構築する場合などには，可能な限りネットワークやストレージのコストを抑えることが重要になる．

そこで，Gigabit/10Gibabit Ethernet の登場により，IP ネットワークを使用してストレージの導入および管理コストを軽減することができる IP-SAN が提案され，

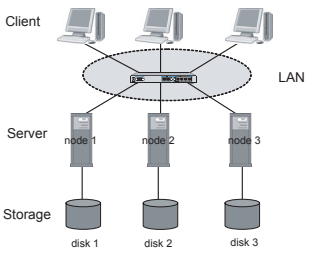

図 1: DAS を用いた PC クラスタ

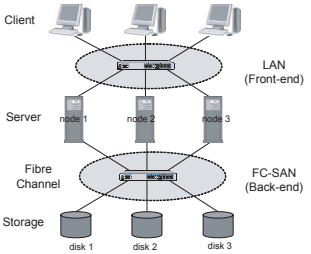

図 2: FC-SAN を用いた PC クラスタ

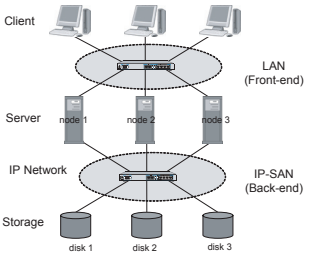

図 3: IP-SAN を用いた PC クラスタ

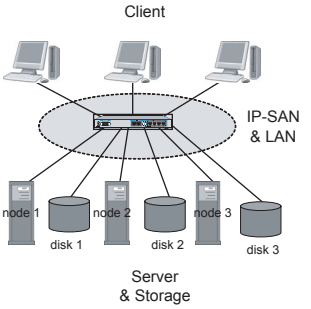

図 4: IP-SAN 統合 PC クラスタ

徐々に普及し始めている．IP-SAN は TCP/IP プロトコルと Ethernet で構築される SAN であり，ハードウェアが安価で管理技術者も多いため，汎用ネットワークの PC クラスタを構築することが可能になる．IP-SAN で使用される技術で代表的なものに，2003 年 2 月に IETF により正式承認された iSCSI (Internet SCSI) プロトコルがある [2]．iSCSI では，TCP/IP パケットの中に SCSI コマンドをカプセル化することによって，ノード (イニシエータ) とストレージ (ターゲット) 間で，ブロックレベルのデータ転送を行う通信プロトコルである．これにより，iSCSI は標準の SCSI 命令体系を使用したまま，IP ネットワークを介したストレージへのデータ転送を可能にする．

図 3 は，IP-SAN を用いて構築した PC クラスタの例である．一般に，SAN を用いた PC クラスタでは，並列計算におけるノード間通信はフロントエンドで行われ，データの I/O 処理に伴うストレージアクセスはバックエンドで行われる．この場合，各ネットワークはスイッチで分離されているため，ノードからストレージへのバルクデータと，並列演算処理のためにノード間で通信されるデータは混在することはない．

そこで本稿では，運用管理の効率化とネットワーク構築コスト削減を目的として，図 4 のような IP-SAN 統合 PC クラスタを実現する．IP-SAN 統合 PC クラスタによって，ネットワークの構築コストや管理コストが軽減できるだけでなく，クライアントからのバックエンド管理も容易になる．

† お茶の水女子大学
Ochanomizu University
‡ 工学院大学
Kogakuin University
§ 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo

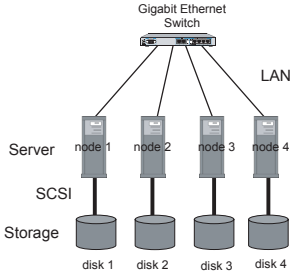

図 5: 実験環境 1

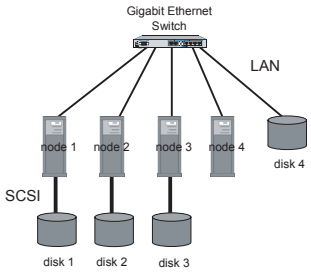

図 6: 実験環境 2

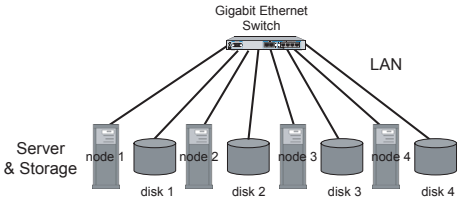

図 7: 実験環境 3

表 1: 性能評価実験環境：使用計算機

OS	initiator : Linux 2.4.18-3 target : Linux 2.4.18-3
CPU	Intel Pentium 4 CPU 1500MHz
Main Memory	384MB
HDD	36GB SCSI HDD
NIC	Intel(R)PRO/1000 MT

3. IP-SAN 統合 PC クラスタにおける IO 性能の基礎実験

本稿では，フロントエンドとバックエンドのネットワークを統合した IP-SAN 統合 PC クラスタを構築し，I/O 処理を伴うベンチマークを使用して，基本的な並列演算性能を評価した．ベンチマークには NAS Parallel Benchmark (NPB) Ver.2.4[3] を使用し，MPI ライブラリとして MPICH2-1.0.3[4] を使用している．

3.1 実験環境

実験に用いたシステム環境を表 1 に示す．本実験において実行した並列分散処理では，ノード数を 4 として以下の場合の性能を評価する．

- 4 台の全ノードが各々のローカル SCSI ディスクにアクセスした場合（図 5）
- 1 台のノードのみ iSCSI ディスクを介してアクセスした場合（図 6）
- 4 台の全ノードがそれぞれ異なるディスクに iSCSI を介してアクセスした場合（図 7）

iSCSI を使用する場合は，図 6, 7 に示すように，イニシエータとターゲットは各ノードと同じ Gigabit Ethernet スイッチで接続し，IP-SAN 統合 PC クラスタを実現している．ただし，本稿の実験環境である図 6, 7 においては，イニシエータとターゲットは 1 対 1 接続になっており，各ノード (イニシエータ) は特定のストレージデバイス (ターゲット) に接続する構成となっている．iSCSI の実装には，ニューハンプシャー大学 InterOperability Lab[5] が提供しているオープンソースの実装 (UNH-iSCSI Ver. 1.5.2) を用いた．

表 2: NPB の Class と問題サイズ

Class	Size	Mbytes written
A	64 × 64 × 64	419.43
B	102 × 102 × 102	1697.93

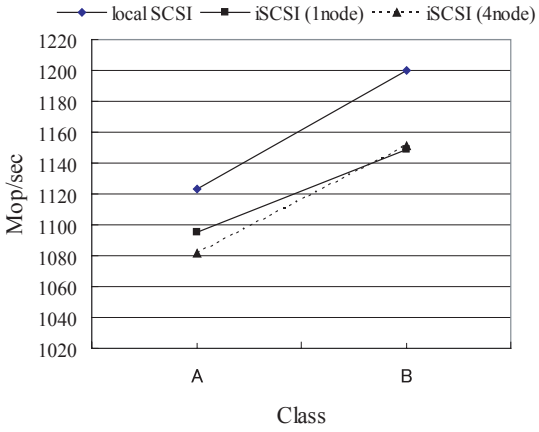

図 8: NPB I/O の Mops 値

3.2 I/O 処理を伴うアプリケーションの性能評価

本実験では，I/O 処理を伴うベンチマークを含む並列計算ベンチマークとして，NAS Parallel Benchmark (NPB) Ver.2.4 を用いた．NAS Parallel Benchmarks は NASA Ames Research Center で開発された，航空関連の流体シミュレーションの実行性能を並列コンピュータ上で評価するベンチマークである．NPB は，5 つの Parallel Kernel Benchmarks と 3 つの Parallel CFD（Computational Fluid Dynamics）Application Benchmarks から構成されている．本稿の実験で使用した NPB 2.4 は，非優位対角な 5 × 5 ブロックサイズの三重対角方程式の解法である BT (Block Tri-diagonal) に対してのみ，大量の I/O 処理を行うアプリケーション実行時の性能を測定することができる．よって本実験では，BT を使用して全体実行時間と Mops (Millions of Operations Per Second) 値を測定した．Mops 値は 1 秒間あたりの 100 万演算数である．実行した NPB の Class，配列サイズ，write 処理されるデータサイズを表 2 に示す．表 2 に示した問題サイズの反復回数はいずれも 200 であり，NPB の実行オプションは並列 I/O を実行する epio を使用している．測定は 5 回行い，その平均値を評価した．

図 8 は，すべてローカルの SCSI ディスクを使用した場合，iSCSI ディスクを 1 台あるいは 4 台使用した場合の Mops 値を示している．Class A において，ローカル SCSI ディスクを使用した場合は 1123 であるのに対し，4 台の iSCSI ディスクを使用した場合は 1081 と低い値が得られた．Class B においてもローカル SCSI ディスクが 1200 に対し，iSCSI ディスクが 4 台では 1152 という結果が得られた．また，iSCSI ディスクが 1 台の場合と 4 台の場合では，Class A ではやや 4 台の場合の方が演算

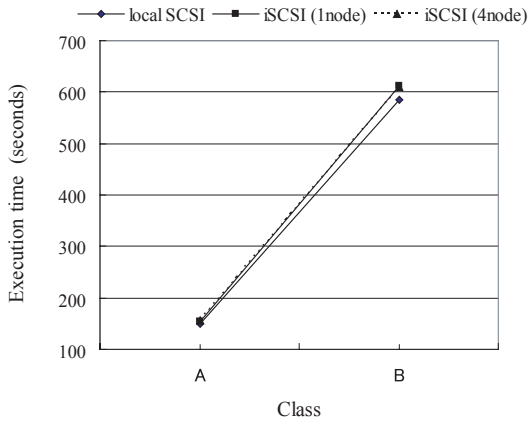

図 9: NPB I/O の実行時間

性能が低いが，Class B では性能に大きな差はみられなかった．

図 9 は，すべてローカルの SCSI ディスクを使用した場合，iSCSI ディスクを 1 台あるいは 4 台使用した場合の NPB の実行時間を示している．Class A の問題サイズにおいては，ローカル SCSI ディスク，iSCSI ディスクともに実行時間に差はなく，ほぼ同じ値が得られた．また，iSCSI ディスクの台数を 1 台から 4 台に変化させた場合においても，実行時間に大きな変化はなかった．Class A より問題サイズが大きい Class B の場合には，iSCSI ディスクの場合はローカルディスクよりやや実行時間が大きい程度であった．

これは，本実験の環境において IP-SAN 統合 PC クラスタは，イニシエータとターゲットが 1 対 1 に接続されており，各ノードの I/O が直接衝突することがないためであると考えられる．よって，IP-SAN 統合 PC クラスタは，実行時間の点においては各ノードのローカルストレージを用いた場合の性能に近ける可能性のあることがわかった．

4. まとめと今後の課題

本稿では，SAN を使用した PC クラスタにおいて，運用管理の効率化とネットワーク構築コスト削減を目的として，分散して存在するクライアント，サーバ，ストレージをネットワーク統合する IP-SAN 統合 PC クラスタ環境を実現した．また，フロントエンドトラヒックとバックエンドトラヒックの統合が，PC クラスタにおける並列処理アプリケーションの実行性能に与える影響を評価するため，I/O 処理を伴う並列計算ベンチマークを使用して基本的な測定を行った．その結果，IP-SAN 統合 PC クラスタにおいて，各ノードの I/O が直接衝突することがない設計の場合には，実行時間の点において，各ノードのローカルストレージを用いた場合の性能に近ける可能性のあることがわかった．すなわち，クラスタの構成法とアプリケーションの種類によっては，IP-SAN 統合 PC クラスタは望ましい性能を発揮することができると考えられる．

今後の課題として，IP-SAN 統合 PC クラスタを設計する際に，各ノードとストレージを 1 対 1 接続するのではなく，特定のストレージに複数のサーバがアクセスするような環境を構築する．それにより，ノード間通信とストレージアクセスが完全に混在する環境において，どの程度並列分散処理性能が劣化するかを本稿の実験環境と比較して調査する．

謝辞

本研究は一部，文部科学省科学研究費特定領域研究課題番号 13224014 によるものである．

参考文献

[1] Storage Networking Industry Association, http://www.snia.org/.
[2] iSCSI Draft, http://www.ietf.org/rfc/rfc3720.txt.
[3] NAS Parallel Benchmark (NPB), http://www.nas.nasa.gov/Software/NPB .
[4] MPICH2, http://www-unix.mcs.anl.gov/mpi/mpich .
[5] InterOperability Lab in the University of New Hampshire, http://www.iol.unh.edu/.