ハイパフォーマンス
コンピューティング 91−18
（２００２．８．２２）

AIST スーパークラスタ構想

工藤知宏† 児玉祐悦†
建部修見† 関口智嗣†

我々は数千プロセッサからなる AIST スーパークラスタの導入を計画している。AIST スーパークラスタは、Linpack ベンチマークで 10 ～ 20 TFLOPS 程度以上の性能の達成を目標としており、主として科学技術計算に用い、グリッドシステムの一部として運用される。本報告では、計画の概要についてのべ、現時点での大規模クラスタの構築においてどのような事項を考慮しなくてはならないかについて考察する。

Basic Concept of AIST Super Cluster

Tomohiro Kudoh ,† Yuetsu Kodama ,† Osamu Tatebe †
and Satoshi Sekiguchi†

We have a plan to introduce a cluster system called the AIST Super Cluster, which consists of thousands of computing processors. The target Linpack performance of the AIST Supercluster is more than 10 or 20 TFLOPS. It will be operated as a part of a grid system and will be mainly used for scientific applications. In this report, we show a basic design concept of the cluster and discuss critical issues in designing today's cluster systems.

1. はじめに

近年、大量生産されるサーバ型の計算機を多数用いて構成するクラスタシステムが、高性能計算環境を構成する手段として注目を集めている。

現在、世界最高性能の計算機は地球シミュレータ [1),2)] で、35.86 TFLOPS の Linpack 性能を持っている。地球シミュレータは、5120 台のスーパーコンピュータを高速なクロスバネットワークで結合した構成をとる。

このようなスーパーコンピュータを用いた高性能計算機は、比較的安定して高い性能を得ることができるが、プロセッサやネットワークなどの主要部品は大量に生産されるものではないため、性能あたりのコストは高価である。

地球シミュレータのプロセッサは 8 GFLOPS の単体性能を持つが、近年のマイクロプロセッサはその 1/2 程度の性能を持っており、多数のサーバ機をつないで構成したクラスタは高性能な環境を比較的安価に構築できる。

しかし、数千のプロセッサからなる大規模なクラスタシステムは、安定稼働や管理の容易性、使い勝手などの点で地球シミュレータのようなシステムに比べて多くの問題を抱えている。

産業技術総合研究所では、科学技術計算に用いるために大規模かつ長時間の数値計算を行うクラスタシステム「AIST スーパークラスタ」の導入を計画している。AIST スーパークラスタは、10 ～ 20 TFLOPS の Linpack 性能の達成を目標としており、多数の計算ノードと、それをサポートするインタラクティブノード、ストレージ機器、ネットワークなどからなるクラスタシステムである。このクラスタは主に科学技術計算に用いられ、グリッドシステムの一部として運用される。必要に応じてシステム全体で、あるいは小さな単位に分割して用いる。

このような大規模クラスタの構築はほとんど例がないため、その設計はクラスタシステムに固有の様々な問題点や、大規模システムで生起すると予想される問題を考慮して行う必要がある。

これには、

- クラスタ全体にわたってプロセスの実行、スケジューリングなどを管理するソフトウェアシステムを持つこと
- クラスタを効率よく使用する通信ソフトウェアやプログラミング環境を持つこと
- グリッドシステムを構成する要素クラスタとして外部ネットワークとの間の十分な通信バンド幅を持つこと
- 故障の発生をなるべく小さくすること
- 一時的（トランジェント）なエラーに対処できるソフトウェアアーキテクチャを採用すること
- システム全体の電源の入切や、故障検出などを集中して管理する機構を持つこと

などがある。これらの事項に対処した設計をすると共に、クラスタ運用開始後もアプリケーションプログラマにとって使い易い環境の構築を継続的に進める必要がある。次節以降では、AIST スーパークラスタの計画の概要について述べると共に設計上考慮すべき事項について議論する。

† 産業技術総合研究所グリッド研究センター
Grid Technology Research Center, AIST

	単体プロセッサピーク性能 (GFLOPS)	単体プロセッサ Linpack性能 (GFLOPS)	20TFLOPS に必要なプロセッサ数
2002年7月	5.00	2.50	16000
2003年1月	6.30	3.15	12699
2003年7月	7.94	3.97	10079
2004年1月	10.00	5.00	8000
2004年7月	12.60	6.30	6350
2005年1月	15.87	7.94	5040

図 1 プロセッサの性能予測と 20TFLOPS の Linpack 性能に必要なプロセッサ数

2. 期待される性能と構成

AIST スーパークラスタが対象とする計算問題の種類は多岐にわたるため、個々の計算ノードに高い性能が求められると同時に、ノード間ネットワークの性能、メモリ性能、ストレージ性能も高いことが求められる。

高性能計算の性能指標には Linpack ベンチマークが広く用いられている。AIST スーパークラスタでは、10 ～ 20 TFLOPS 以上の Linpack 性能の達成を目標としている。

計算ノードの個々のプロセッサのパフォーマンスについては、現時点で Intel Xeon などが 5 GFLOPS 程度の理論ピーク性能を持っている。

我々の測定では 2.2 GHz の Intel Xeon(理論ピーク性能 4.4 GFLOPS) 単体での Linpack 性能は 2.4 GFLOPS であった。ここから、単一プロセッサでは Linpack 性能はピーク値の半分程度になることが予想できる。

多数の計算ノードを用いれば理論性能はプロセッサの性能にプロセッサ数を乗じた値になるが、実際の計算での性能はこれよりも低くなる。数千もの計算プロセッサを用いたクラスタシステムでの Linpack 性能についての正確な知見はないが、おおむね単一プロセッサの Linpack 性能にプロセッサ数を乗じた値の 5 割から 7 割程度が期待できると考えられる。ここでは 5 割を仮定する。

これらの仮定からは、20 TFLOPS の Linpack 性能を実現するために必要なプロセッサ数 nPU は、

$$nPU = \frac{20\ TFLOPS}{peak\_perf\_of\_single\_PU * 0.5 * 0.5}$$

で与えられる。

また、現在のところプロセッサの性能向上は Moore の法則に従っており（あるいは Moore の法則に従うべく開発が行われており）1 年半で 2 倍の性能向上が期待できる。

これらから単純に計算すると、20 TFLOPS の性能を得るために必要なプロセッサ数は図 1 のようになる。

また、Linpack 性能のみが高くとも想定される様々な計算問題で十分な性能が得られるシステムであるとは限らない。そのため、ネットワーク性能などについては個々に規定する必要がある。

3. ハードウェアの構成

3.1 全 体 構 成

大規模クラスタでは、個々のノードの構成により、全体の大きさ、性能、消費電力が左右される。ノードには、マザーボード、電源、周辺装置などが付随するため、同一のプロセッサを同一数使用した場合、一筐体に格納されるプロセッサ数が多い方が消費電力は小さくなると考えられる。

8000 プロセッサを持つクラスタシステムを想定すると、計算ノードの選択と全体構成は以下のようになる。

1U 2PU SMP 現在、2 プロセッサが 1U の大きさの筐体に格納された SMP サーバ機が広く用いられている。1U のサーバ機は 1 ラックに 40 ノード、80 プロセッサ程度が格納可能である。この場合、8000 プロセッサを格納するのに 100 本のラックが必要ということになる。この構成の Intel Xeon を用いたサーバ機の 1 台あたりの最大消費電力は 250W 程度であり、全体での消費電力は計算ノードのみで 1000kW となる。

4U 8PU SMP 8PU 程度の SMP を 4U 筐体で構成することができれば、実装密度としては前項と同程度となる。筐体数が少なくなるため全体の消費電力は前項の構成と比べて若干少なくなると考えられる。

ブレード型 このほかに、ネットワークや周辺機器を格納するラックが加わる。ブレード型の構成をとればさらに実装密度を上げられると考えられるが、現在のところブレード型は比較的性能の低いプロセッサを用いたものが主流である。ブレード型を用いれば、設置面積は上述の構成よりも有利である。

既存のアーキテクチャをそのまま用いると消費電力が非常に大きくなってしまう。消費電力を押さえることが今後の大規模クラスタシステムの鍵であり、AIST スーパークラスタでは、実現可能と思われるターゲットとして、システム全体で 1000kW 以下に押さえることを想定している。

また、設置場所は 100 本強のラックを設置することを想定して $500m^2$ とした。

このような大規模システムでは、保守性や冷却を十分に考慮して配置、配線を行うことが重要である。

3.2 計算ノードの構成

現段階では使用するプロセッサは決定していないが、Intel Xeon などの IA32 プロセッサ、Itanium 2 などの IA64 プロセッサ、Opteron, Power4 などのプロセッサを使用することが考えられる。

現時点でクラスタを構築する場合の計算ノードの例として考えられる、Intel Xeon プロセッサを用いた 1U 2PU SMP サーバ機の仕様は、おおよそ以下のようになる。

プロセッサ Intel Xeon 2.4GHz × 2
メモリ 4 GByte
ハードディスク 125 GByte × 4
ネットワーク マザーボードに 1000BASE-T ポートが 2 組
I/O スロット PCI 64bit/66MHz または PCI-X 64bit/133MHz が 1 ～ 2 組

プロセッサのパフォーマンスは、前述のように 1.5 年で 2 倍程度、メモリ容量も同様に 1.5 年で 2 倍程度向上する。ハードディスクの容量は 1 年で 2 倍程度向上する。性能が向上しても価格はほぼ一定にとどまる傾向があるため、クラスタの購入時期により現在の構成よりも高い性能のサーバ機を同一価格で得られることになる。

より多いプロセッサによる SMP 構成とすれば、一筐体に格納されるプロセッサ数を多くでき、ノードあたりのメモリ量を相対的に増やすことができる。しかし、メモリバンド幅が全体性能の隘路となることがないよう注意する必要がある。

また、32bit のアドレス空間しか持たないプロセッサではプロセスがアクセスできるメモリ空間は 4 GByte に制限される☆。しかし、32bit のプロセッサである Intel Xeon などでは、物理メモリアドレスは 36bit であり、ノードあたりでは最大 64 GByte のメモリを実装できることになる。

3.3 ネットワークの構成

計算性能を重視したクラスタでは、計算時に用いるネットワークと、管理やデータ転送用に用いるネットワークは分離することが望ましい。

計算用ネットワークは、システムの性能を左右する重要なネットワークであり、通信の上位レイヤに負担をかけることなく広いバンド幅を提供できなくてはならない。このためにはユーザレベルの RDMA(Remote DMA) 通信を効率的に用いることができる必要があり、パケットを廃棄しないこと、十分な信頼性を持つことなどが求められる。

並列システムの経験則として、実効性能が高い並列システムを構築するにはノードの計算性能 1 MFLOPS あたりノードが入出力する通信リンクのバンド幅が 1 Mbps 程度 (単方向あたり) 必要であるとされており、現在のノードの性能からは少なくとも 3 Gbps 以上、できれば 10 Gbps 程度の実効通信性能を持つネットワークが望まれる。また、ネットワーク全体のトポロジについても、

- 単一リンクのバンド幅だけでなく、システム全体でランダムなノード間通信が行われる際にも、十分なバンド幅を持つこと。
- システムをいくつかの構成単位に分割して使用することが考えられるので、そのような分割を行った場合に、なるべく小さな分割単位まで、構成単位内ノード間のトラフィックが他の構成単位に影響を与えないこと。すなわち、いわゆるクローズドパーティションを構成できること。

を考慮する必要がある。

現在、この種のネットワークとしては Myrinet[3] や Quadrics QsNet[4] などが多く用いられているが、InfiniBand[5] も利用可能になりつつあり、どのようなネットワークがもっとも適しているかを、性能、コストの両面から検討する必要がある。

管理、データ転送用ネットワークとしては、最近のサーバタイプの計算機は本体に 1000BASE-T のポートを持つものが多いためこれを用いることが考えられる。千台規模のノードをつなぐネットワークとなるため、十分な実用性能が得られるようなトポロジ、スイッチ装置を用いる必要がある。

3.4 2 次 記 憶

計算ノードにはハードディスクが装着されるが、このディスクは基本的には計算中に一時的にそのノードで用いるデータを置くスクラッチ領域として用いる。

システムの運用には、ユーザ領域などを置く、複数のノード間で参照できる共有ファイルシステムが必要である。

☆ OS の構成によりさらに少なくなることがある

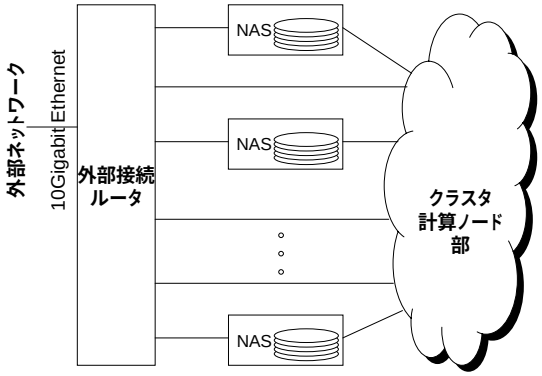

図 2 外部ネットワーク接続

この共有ファイルシステムは NAS を用いて構成するのがもっとも簡単であるが、ノード数が多いので、システム全体に十分な性能を供給することができるよう考慮する必要がある。システムを分割して使用することが多いことを考慮すると、ある程度局所性のある NAS 構成とし、NAS 装置間のデータの同期をとる仕組みを用意するなどの方法も考えられる。

また、各ノードの二次記憶を用いたクラスタファイルシステムにより共有ファイルシステムを構築することも考えられる。しかし、共有ファイルシステムは、大容量データの格納だけでなく、コンパイルなど日常的なファイル処理にも用いるため、それらの処理に対する十分な性能とファイルシステム自身の十分な信頼性が求められる。このような性能をどのようにして確保するかについて、現在のクラスタファイルシステムが解決しなくてはならない課題は多いと思われる。

3.5 その他の機器

ユーザがログインして計算の準備、モニタリングなどを行うインタラクティブノードが相当数必要である。

ユーザは、外部からネットワークを介して本クラスタシステムを使用する。外部ネットワークとの接続には、管理・データ転送用ネットワークにつながる 10Gbit Ethernet(10GBASE-LR) のポートをシステム全体で持つことを想定している。この際、外部と共有ファイルシステムとの間で大量のデータ転送が行われることが想定されるので、共有ファイルシステムとの間のバンド幅が十分に確保されることが望ましい。

複数の NAS を用いて共有ファイルシステムを構築する場合、例えば図 2 のような構成が考えられる。

4. ソフトウェアの構成

クラスタシステムは、本来単体で動作するように設計された計算機をネットワークで接続したものであるため、安定して効率よくクラスタを運用するソフトウェアを用いることが重要である。

AIST スーパークラスタでは、多くのクラスタシステムで安定稼働の実績がある SCore システムソフトウェア [6),7)] を用いる。

産業技術総合研究所では、SCore 上で動作する PM 通信ライブラリ [8),9)] を用いたプログラムを有しており、本クラスタシステムでもこれらのプログラムが動作する必要がある。PM は、通信に際してオペレーティングシステムの介在によるオーバヘッドを排除し

たユーザレベルゼロコピー通信 (Remote DMA) を用いる[10]。MPI が PM 上で利用可能で、C および C++ を用いた並列プログラムを動作させることができる。

一方、科学技術計算用途では C,C++ に加えて FORTRAN(77, 90) で記述されたプログラムが多く、また、並列プログラミングには MPI を用いてプロセス間通信をするモデルと、OpenMP を用いて共有メモリ型プログラミングをするモデルの両者が用いられる。これらについてもサポートされる必要がある。

また、効率的な並列プログラムの開発, 効率的なクラスタ利用のために並列デバッガ、パフォーマンス測定ツールが備えられていることも必要である。

さらに、本クラスタはグリッドシステムの一部として運用されるため、少なくとも Globus[11](現在の最新バージョンは 2.0[12]) が動作することが求められる。AIST スーパクラスタが備えるグリッドサービス [13] の詳細は未定だが、NPACI の TeraGrid[14] などとも協調しながら配備されることになる。

5. 保 守 性

AIST スーパークラスタのように大規模なクラスタシステムでは、保守性について考慮した構成とすることが重要である。

5.1 集中管理

日常のメンテナンスを容易にするため、以下のような操作は一カ所で集中して行えることが求められる。

(1) 個々のノードのリセット、BIOS の設定及びアップデート、電源の投入、断など
(2) OS のインストール、アップデート、ディスクイメージあるいはファイルシステムの配布
(3) 環境モニタ、故障検出

クラスタは、通常のサーバ機の集合の形態をとることがふつうであるため、このような集中管理を実現するためには、ハードウェア、ソフトウェア共に対応の検討が必要である。

5.2 故障対応とエージング

クラスタシステムでは、一般に導入初期にネットワークやハードディスクなどに多くの故障が発生する。このため、システムの安定化のために、設置前にネットワークカード、ネットワークスイッチ、ハードディスクなどのエージングを行う。

一方、保守コストを考慮し、故障頻度が少なくなると考えられる保証期間後の故障に対しては、全体の 1 ～ 2 %程度の予備機器により運用することを予定している。

6. おわりに

クラスタは、高い計算能力を安価に得られる手法として注目を集めているが、現在のところユーザにとっては従来のスーパーコンピュータなどと比べて使い勝手がいいとはいえない。また、製造元・開発者の異なる様々な装置とソフトウェアの組み合わせにより構成されるため、実際に運用したときに十分な性能が得られないことがある。AIST スーパークラスタは、これまでにない大規模なクラスタシステムであり、既存システムにない問題が表面化する可能性もある。

このため、導入後も、ユーザのプログラムの開発、改良及び関連機器の接続等に関する技術的な相談に速やかに応じられる体制を用意するとともに、オペレーティングシステム、SCore、コンパイラ、ネットワークなどについてのチューニングを実地に行い、使い易く性能の高いシステムを構築していく必要がある。

謝辞

本報告は産業技術総合研究所クラスタ検討ワーキンググループでの多くの議論に基づいている。また、SCore およびクラスタの構成法について東京大学大学院情報理工学系研究科の石川裕助教授に多くの助言を頂いた。

参 考 文 献

1) 谷啓二, 横川三津夫. 地球シミュレータ計画 i. 情報処理, Vol41,No3, 2000.
2) 横川三津夫, 谷啓二. 地球シミュレータ計画 ii. 情報処理, Vol41,No4, 2000.
3) Myricom, Inc. *http://www.myri.com/*.
4) Fabrizio Petrini, Wuchun Feng, Adolfy Hoisie, Salvador Coll, and Eitan Frachtenberg. The quadrics network (qsnet): High-performance clustering technology. In *Hot Interconnects 9*, August 2001.
5) InfiniBand Trade Association. *http://www.infinibandta.org/*.
6) A. Hori, Y. Ishikawa, H. Konaka, M. Maeda and T. Tomokiyo. “Overview of Massively Parallel Operating System Kernel SCore”. Technical report, RWCP, 1993. TR-93003.
7) 堀敦史, 手塚宏史, 石川裕. ギャングスケジューリングの PC クラスタ上での実装. In *HPC*, 第 67-14 巻, pp. 79–84, October 1997.
8) 手塚宏史, 堀敦史, 石川裕. ワークステーションクラスタ用通信ライブラリ PM の設計と実装. 並列処理シンポジウム JSPP’96. 情報処理学会, June 1996.
9) Hiroshi Tezuka, Atsushi Hori, Yutaka Ishikawa, and Mitsuhisa Sato. PM: An Operating System Coordinated High Performance Communication Library. In *High-Performance Computing and Networking ’97*, 1997.
10) 手塚, 堀, O’Carroll, 原田, 石川. ピンダウンキャッシュを用いたユーザレベルゼロコピー通信. 情報処理学会研究報告. 情報処理学会, August 1997.
11) Ian Foster and Carl Kesselman. Globus: A metacomputing infrastructure toolkit. *Intl J. Supercomputer Applications*, Vol. 11, No. 2, pp. 115–128, 1997.
12) *Globus Toolkit 2 release*. http://www.globus.org/gt2/.
13) *Open Grid Services Architecture*. http://www.globus.org/ogsa/.
14) *TeraGrid*. http://www.teragrid.org/.