1．並列計算機アーキテクトからみた計算機クラスタ

森　眞一郎・富田　眞治［京都大学］

はじめに

今日，並列処理は研究対象の時代から実用化の時代へと向かいつつあり，多くの並列計算機が最先端のスーパーコンピュータとして活躍している[1)]．

一方で，このような超並列あるいは高並列計算機とは，一線を画して実用的な並列度の並列計算機の活躍も目立ってきている．企業の部門レベルや大学の学科クラスの組織でも入手可能な数十台規模のプロセッサを持った部門サーバ的なミッドレンジの並列計算機である．その中でも，最近注目を集めているのが「計算機クラスタ」である（**図-1**参照）．

計算機クラスタは，パーソナルコンピュータ（PC）やワークステーション（WS），場合によってはPC/WSサーバといった小規模並列計算機を，汎用のネットワークで接続した一種の並列計算機と見なすことができる．プロセッサならびにネットワーク性能の急速な伸びを受けて，コストパフォーマンスに優れた並列計算機としての地位を築きつつある．

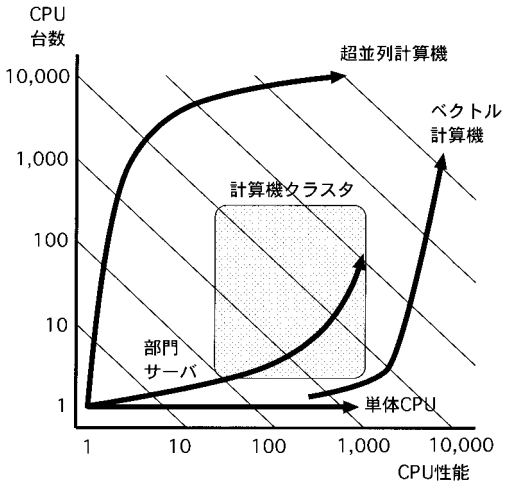

図-1　過去10年間の計算機性能の推移傾向

以下，本稿では計算機クラスタ台頭の背景（**図-2**参照）について説明し，次に並列計算機としての計算機クラスタの特徴を紹介した後，具体的な計算機クラスタの研究／開発事例を紹介する．

計算機クラスタ台頭の背景

・並列計算機の登場

専用のハードウェア（HW）を駆使して，汎用のマイクロプロセッサを並列接続する並列計算機の研究は1970年代初頭に始まり，スーパーコンピュータ並みの性能を目指して，個々の計算機ごとに最適化された低レイテンシ，高スループットのネットワーク系ならびにメモリ系HWの研究／開発が行われた．汎用プロセッサの低価格化に伴い，1980年代前半には研究用の多くの並列計算機が開発され，各プロセッサごとにメモリを分散配置しメッセージパッシングによってプロセッサ間通信を行う**メッセージパッシング型並列計算機**や，プロセッサとは独立にメモリを集中配置し，この共有メモリを介してプロセッサ間の通信を行う**（集中）共有メモリ型並列計算機**（最近は，Symmetric MultiProcessor（SMP）と呼ばれることが多い）が登場した．

1985年頃からは，集中配置した共有メモリシステムにおけるスケーラビリティの問題を解決する手法として，主記憶は各プロセッサに分散配置するが，プログラマに対してはHWで共有メモリ環境を提供する**分散共有メモリ型並列計算機**が登場する．汎用計算機の市場の一部がRISC型のマイクロプロセッサを搭載したワークステーションに移り始めた時期でもあり，並列計算機で使用するプロセッサのコストパフォーマンスが向上した．

・計算機クラスタの幕開け

1990年前後には，多くの並列計算機ベンダが登場し商用の並列計算機が数多く開発された．また，数千～数万台のCPUを搭載する超並列計算機の研究が始まった時期である．

その一方で，TCP/IPベースのネットワークで接続

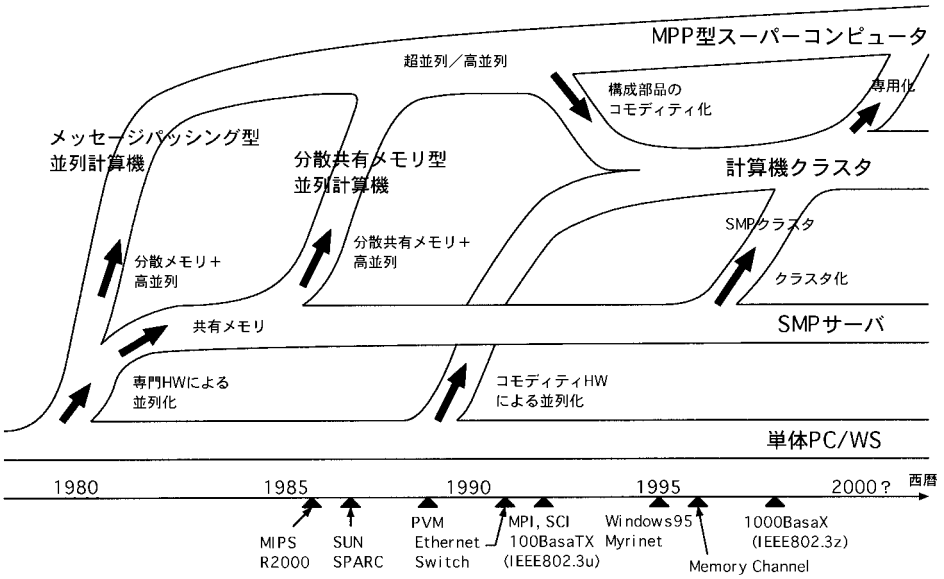

図-2　計算機クラスタへの道程

された複数の計算機の集合体を，仮想的に1台の大きな並列計算機として見せるためのソフトウェアメカニズムであるPVM（Parallel Virtual Machine）が1989年に開発された[2]．PVMは個々の計算機上の通信デーモンプロセスを介してメッセージ交換を行うメッセージ通信ライブラリであり，特別なHWを一切必要とせず，公開されたソフトウェアをインストールするだけで仮想的な並列処理環境が構築できた．そのため，軽い気持ちで並列処理を楽しみたいユーザに支持され広く普及した．これが計算機クラスタの幕開けである．

その数年後には，IBMや富士通などのハイエンドの並列計算機メーカが，自社製のUNIXワークステーションを専用高速ネットワークで接続したメッセージパッシング型の並列計算機[3), 4)]を市場に投入した．この時点で，「計算機クラスタ」と呼ぶべき物理的な計算機（群）が登場した訳である．

• 計算機クラスタの急成長

1995年前後には，イーサネットスイッチや100Mbit Ethernetなどの技術が一般にも普及し，比較的安価に高性能なネットワークの構築が可能になった．さらに，Myricom社のMyrinet[5]，DEC社のMemory Channel[6]，IEEE1596規格のSCI[7]など，計算機クラスタを指向した専用の高速ネットワークが登場した．これにより，専用の並列計算機並みのスループット（数Gbps）を持つネットワークの構築も可能となった．一方で，PCの量産効果による価格低下と，PC向けプロセッサの急速な性能向上により，コストパフォーマンスの優れた計算機クラスタの実現が可能となった．この時期の計算機クラスタの特徴は，クラスタを構成するPC/WS間接続のネットワークをインターネット接続のためのネットワークとは別に設けたシステムが開発された点である．

一方，ソフトウェアの方では，アプリケーションプログラムインタフェース（API）として，多くの並列計算機メーカや研究者らで組織されたMPI Forumが策定したMessage Passing Interface（MPI）[8]が広く普及した時期である．これにより並列プログラムの記述性／移植性が著しく改善され，多くのプログラマにとって並列処理の敷居が一段と低くなった．また，プロセッサ間通信に伴うソフトウェアのオーバヘッドを軽減／削除する試みが多く行われた時期でもある．ユーザプロセスが直接ハードウェアを操作するユーザレベル通信の実現やゼロコピー通信などがその例である．LinuxやFreeBSDなどのフリーなOSの普及が，これらの試みに拍車を加え，安定かつ高性能な計算機クラスタ向けのフリーなソフトウェアが入手可能になったのである．

安価なハードウェア（HW）とフリーなソフトウェア（SW）を使ったパーソナル並列計算機としての計算機クラスタが登場した時期といってもよい．

• 計算機クラスタの発展

現在は，複数のプロセッサを搭載したSMP型のWSやPCが比較的容易に入手可能になっており，これらをベースにしたSMPクラスタが登場している．ハイエンドのSMPサーバを提供する計算機メーカも，さ

らなるスケーラビリティを追究して，SMPサーバを汎用の高速ネットワークで接続するSMPクラスタを発表し始めている．計算機クラスタの性能は，すでにハイエンドなスーパーコンピュータの領域に到達しつつあるといってもよい．ソフトウェア環境も本格的なマルチユーザ対応のものが整い始めている．

今後は，さらなるHWの高性能化とプログラムの記述性向上を目指して，分散共有メモリを提供する計算機クラスタが普及してくるものと考えられる．分散共有メモリを提供する計算機クラスタは，現在活発な研究が行われているが，SWのみで仮想的な共有メモリイメージを提供する方法（本特集「2．分散共有メモリに基づく計算機クラスタ」の記事参照）と，HWで実際に共有メモリを実現するアプローチがある[9), 10)]．後者は，従来の分散共有メモリ型並列計算機で培われた技術を応用するもので，ハードウェア資源の仮想化や通信レイテンシの軽減／隠蔽のためにネットワークやメモリ回りに強力なHWサポートがあるのが特徴である．このクラスの計算機クラスタでは，従来の並列計算機とのアーキテクチャ上の差異が再びなくなってくるものと考えられる．

計算機クラスタの特徴

このように，計算機クラスタは並列計算機の構成手法の1つとして地位を確立してきた．では，従来の並列計算機とは何が違うのであろうか．以下では，独自に開発された専用HW/SWによって構成される「生粋の並列計算機」の代表としての超並列計算機（MPP）と，基本的にコモディティHW/SWによって構成される「計算機クラスタ」を，いくつかの項目について比較することで計算機クラスタの特徴づけを行う（**図-3**参照）．

• CPUの性能

基本的には，MPPも計算機クラスタも市販のプロセッサを使用するという点で同一であるが，最近のプロセッサは商業的な戦略から，ハイエンド向けとローエンド向けで若干の差を出す製品があるのでこの差が出る可能性がある．ただし，システムの開発期間の問題が絡むと，後述する通り計算機クラスタの方が有利になる場合もある．

• ネットワーク・メモリ系の性能

MPPはその製品向けに専用LSIの開発を行うなど，並列処理のために最適化したメモリ系やネットワーク系のハードウェアを備えるのに対して，計算機クラスタではコモディティなPC/WSを使用するため，HWのレイテンシに関してMPPより劣る傾向がある．HWスループットに関しては，計算機クラスタもMPPに近い性能を出し始めている．

• I/Oのスケーラビリティ

計算機クラスタは個々のPC/WSが独自にI/O装置を持っており，プロセッサ台数に比例してI/O能力がスケールアップする．これに対してMPPでは，システム全体に1個，あるいは，ある程度のプロセッサ台数ごとに1個のI/O装置を持つ構成が多い．1台のI/O装置の性能を比較すると，一般にMPP向けの方が性能がよいものが多い（価格もそれなりに高い）が，I/Oのスループットに関しては数で勝負の計算機クラスタに負ける傾向がある．

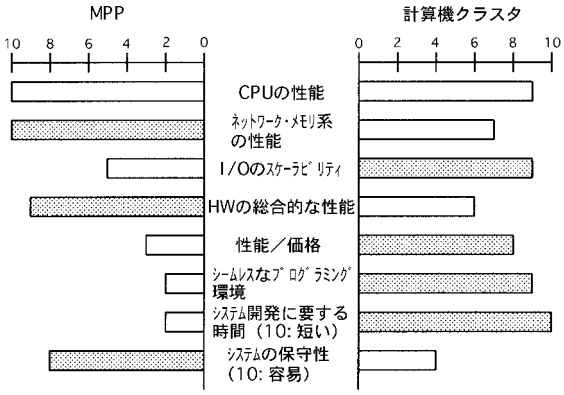

図-3　MPPと計算機クラスタの比較

• HWの総合的な性能

両者のHW設計の基本方針をあえて区別するならば，MPPが「N台のCPUを使って，1CPUのときの計算時間の1/Nの時間で計算できること」に対して，計算機クラスタは「N台の計算機を使って1台のときに現実的な時間で処理できるデータサイズのN倍のサイズのデータを扱うこと」である．一般に，前者の要求を満足するには，細部に渡って性能に余裕を持たせたHW設計が必要である．したがって，ピーク性能が同一であっても総合的な性能はMPPの方が優れている．

• コストパフォーマンス

プロセッサ100台規模までのシステムで比較すると，MPPと計算機クラスタの瞬間最大性能は比肩しているので，量産効果の高い計算機クラスタが単価当たりのピーク性能が高くコストパフォーマンスに優れるといえる．しかし，MPPは基本コストは高いが，台数に比例するプロセッサ1台当たりの価格は計算機クラスタに比べて一般に低いので，仮にプロセッサ1000台規模の計算機クラスタを実現した場合には，コストパフォーマンスの逆転が起こる可能性がある．また，ピーク性能に近い性能を引き出せないアプリケーションに対しては，計算機クラスタはMPPに比べて性能の低下傾向が顕著であり，この場合にもコストパフォーマンスが逆転する可能性がある．

• プログラミング環境

MPPは，従来より個々の並列計算機がそれぞれの

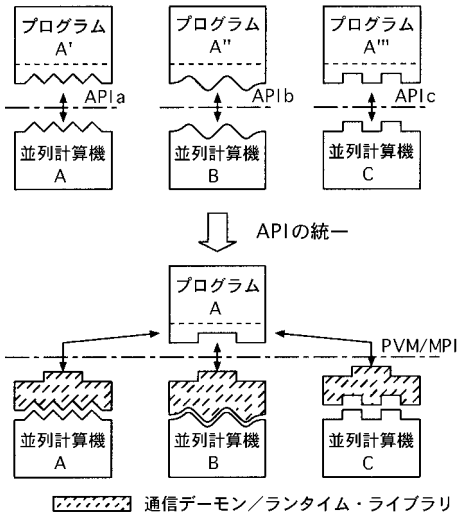

図-4　APIの統一による移植性の向上

HW性能を存分に発揮するための独自のAPIを提供していた．そのため，APIの修得などの初期オーバヘッドが大きく，また，ある計算機でよいアプリケーションが開発されても別の計算機への移植が困難といった問題があった（**図-4**参照）．これに対して計算機クラスタでは，それ自体が登場した背景にPVMやMPIといった標準的なAPIが存在したために，機種依存の少ない効率的なプログラミング環境が提供されている☆．

また，MPPのOSは自社開発したものが多く，一般ユーザが並列処理へ向けた最初の一歩を躊躇する原因にもなっていた．これに対して計算機クラスタは（基本的には）市販のOSそのものを使用するため，日常的に使っているPC/WSとまったく同一の環境が使用できるという利点があり，並列処理の敷居を下げるという貢献をしている．

・システム開発に要する期間

一般に，MPPが2年から3年かけて開発されるのに対して，計算機クラスタを構成するWS/PCは3カ月から半年ごとに新製品が登場するのが現状である．そのため，同時期に入手可能なMPPと計算機クラスタでプロセッサの世代が違うという現象も発生する．

・システム保守のコスト

MPPはHW/SWの安定性が高くメンテナンスも委託できるのに対して，計算機クラスタはHW/SWの安定性が比較的低く，基本的に自分でメンテナンスが必要であるため保守コストが高くなる．

以上，著者の主観を交えてMPPと計算機クラスタの比較を行った．計算機クラスタがどんなものか大雑把なイメージは持っていただけたと思う．以下では，より具体的な計算機クラスタの研究／開発動向を紹介する．

☆ 最近の並列計算機は，最初からMPIを実装したものが多いが，MPIを使うとMPPのHW性能が十分に発揮できずに計算機クラスタとの有意差が出せないという問題がある．

計算機クラスタの研究／開発動向

・ベンダ事例

[超並列クラスタ系]

IPMのSPシリーズや，富士通のAP3000は超並列のスーパーコンピュータとして，早くから市場に投入されている計算機クラスタである．計算ノードには市販のWS用に開発されたものをほぼそのままの形で採用し，複数ノードを専用の筐体に納めたものを専用の高速ネットワークで接続する形態をとる．計算ノードがモジュール化されているため，基本的にはモジュールの交換だけで最新CPUの性能を享受することができる．したがって，WS用の新しいモジュールが開発される頻度で，システムの更新が可能になっている．

[SMPクラスタ系]

多くのベンダが次世代サーバとして，計算機クラスタを市場に投入し始めている．その多くは，最大構成でシステム全体のプロセッサ数が256台程度の高並列計算機の一角を占めている．

SUN Microsystems社は，Ultra Enterpriseサーバを1ノードとして最大4ノードでクラスタを構成するシステムを提供している．ネットワークカードとしてはSbus用のSCIカードを開発し，4ポートのSCIスイッチで接続している．

DEC社（現Compaq社）は，自社開発のMemory Channel IIを使った製品の供給を行っている．計算ノードとしては，同社のAlphaチップを最大120CPU搭載しこれらを階層的なクロスバ網で接続したSMPサーバを使用する．Memory Channelは，共有データの書込みに関して若干特殊な操作が必要であるが，共有メモリに準ずるユーザインタフェースを提供しており，その機能を利用して細粒度の通信にも対応できるソフトウェアによる分散共有メモリ環境の開発を行っている．

Sequent社はPentiumPro 4台のSMPクラスタをSCI準拠のリンク（Link速度8Gbps，複数リンク構成）でリング状に接続するNUMA-Qを供給している．ハードウェアによる完全な分散共有メモリ環境を実現し，細粒度並列処理を実現する．ネットワークカードがNUMA-Qにかなり特化されているが，ネットワーク自体はSCIベースの標準品である．類似のシステムとして，Data General社のAV25000がある．

・研究所，大学関係

[究極のお手軽クラスタ]

米国のロスアラモス国立研究所が開発した計算機

クラスタAvalonは，コモディテイ・ハードウェアとフリーソフトウェアだけで作られており，DECのAlpha 68台にフリーなPC UNIXの1つであるLinuxを搭載し，通常のFastEthernetとGigabitEthernetで接続したシステムである．手作りでシステムを組み立て，15万ドルの投資で19GFLOPS[☆2]を達成している．

[超並列計算機の代替としての計算機クラスタ]

我が国の新情報処理開発機構（RWC）で開発されているMyrinetベースの計算機クラスタ[11]は，超並列計算機と比較して遜色のない性能を出すことを目標としてる．Avalonが基本的にHW/SWのすべてを既製品で構成したのに対して，RWCのクラスタでは，HWはすべてコモディティ部品であるが，NICカードのドライバ，OS，通信ライブラリ，言語処理系といったSW全体に渡って独自に開発／改良したシステムである．後述のNOWプロジェクトのActive Messages，イリノイ大学のFast Messages[12]と肩を並べて著名な並列通信ライブラリPMを開発している．詳細は本特集「5．RWCPにおけるクラスタ開発記」などを参照されたい．

[計算機クラスタによるスケーラブルI/O]

カリフォルニア大学バークレイ校では，1990年代の初頭からNetwork of Workstations（NOW）の研究が行われている[13]．どちらかというと，OSやディスク回りの研究が中心である．現在は，140台のWS（OSはSolaris）と35台のPC（OSはPC UNIXあるいはWindowsNT）をMyrinetベースのネットワークで接続している．システム内には1000台近いハードディスクがあり，独自に開発したサーバレス・ファイルシステム xFSを使用してdisk-to-diskソートのベンチマークで1997年4月時点の世界記録を達成している．

xFSは，「ローカルディスク上のデータよりも，他の計算機の主記憶にあるデータの方が速くアクセスできる」という発想のもとに，データを使用するクライアントの主記憶をグローバルな共有ファイルキャッシュとして使用するcooperative cachingや，複数計算機のディスクにまたがってストライピングを行うソフトウェアRAIDの手法などを応用することでスケーラブルで可用性の高いファイルシステムを実現している．

[共有メモリ環境を目指して]

ソフトウェアによる共有仮想記憶システム（SVM）を提案したプリンストン大学のK. Liらのグループは4-Way PentiumProをベースにしたSMPクラスタSHRIMPを開発している[14]．SHRIMPでは，ネットワークインタフェースを仮想化し，プロテクションを緩めずにユーザレベル通信を可能にする機構として，仮想アドレス空間にネットワークインタフェースのチャネルをマッピングするVirtual Memory Mapped Channelを提案し，これを実現するための専用ネットワークインタフェースカードを開発している．2つのプロセス間で，いったんチャネルを設定すると，それ以後のプロセス間通信はそのチャネルに対応する領域を共有メモリとして使用することができ，通信のたびにシステムコールを介す必要がなくなる．

[☆2] 数値の上では64プロセッサのOrigin 2000に匹敵する性能．

おわりに

スーパーコンピュータといえばベクトル計算機であった10年前，「安価な汎用マイクロプロセッサをたくさんつないで，ベクトル計算機に匹敵する性能を出す」というのが並列計算機のうたい文句であった．そして今日，「PCあるいはWSを適度につないで専用並列計算機に匹敵する性能を出す」とうたっているのが計算機クラスタである．前者の場合，「逐次処理」から「並列処理」という真に技術的な変革があったのに対し，後者は「使いやすさ」や「コストパフォーマンス」の追究という，並列処理技術の「発展」から「成熟」への変革であるといえる．

計算機クラスタの普及によって並列処理の敷居が下がり，より多くのプログラマが並列処理を体験する機会をもったことは非常に有意義である．このようなブレークスルーによって並列処理が一部の研究者のための技術ではなく，ごく一般の当たり前の技術となることを願う次第である．

参考文献

1）富田眞治: 並列コンピュータ工学, 昭晃堂（1996）.
2）Geist, A. et al.: PVM: Parallel Virtual Machine A Users' Guide and Tutorial for Networked Parallel Computing, MIT Press.
3）Agerwala, T. et al.: SP2 System Architecture, IBM SYSTEM JOURNAL, Vol. 34, No.2（1995）.
4）AP3000ご紹介資料, 富士通株式会社HPC本部（1995）.
5）Boden, N. J. et al.: Myrinet: A Gigabit-per-Second Local Area Network, IEEE Micro, pp.29-36（Feb. 1995）.
6）Fillo, M. and Gillet, R. B.: Architecture and Implementation of MEMORY CHANNEL2, DIGITAL Technical Journal（online）（28 Aug. 1997）.
（http://www.digital.com/info/DTJP03/DTJP03HM.HTM）
7）IEEE Standard for Scalable Coherence Interface（SCI）, IEEE Std 1596-1992.
8）MPI Forum, MPI: A Message Passing Interface Standard（1995）.
9）Mori, S. et al.: A Distributed Shared Memory Multiprocessor: ASURA－Memory and Cache Architectures－, Proc. of SUPERCOMPUTING 1993, pp.740-749（Nov. 1993）.
10）青木他: 共有メモリベースのシームレスな並列計算機環境を実現するオペレーティングシステムの構想, 情報処理学会研究報告, 97-OS-74, pp.195-200（Feb. 1997）.
11）手塚, 堀, 石川: 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, JSPP '96, pp.41-48（June 1996）.
12）High Performance Virtual Machinesホームページ
（http://www-csag.cs.uiuc.edu/projects/hpvm.html）
13）Anderson, T. E. et al.: A Case for NOW（Networks of Workstations）, IEEE Micro, pp.54-64（Feb. 1995）.
14）Blumrich, M. A. et al.: Virtual Memory Mapped Network Interface for the SHRIMP Multicomputer, Proc. of Int'l Symp. on Comp. Arch., pp.142-153（1994）.

（平成10年9月28日受付）