1R－04　ＰＶＭによる中規模クラスタシステムの負荷分散と効率について

黒住祥祐　　平石裕実　　竹内　勉　　大本　英徹
京都産業大学工学部

1．　まえがき

多数のＰＣ（パソコン）をＬＡＮで接続した分散処理はこれまでも広く行なわれている。しかし、クラスタシステムとして、実用化されているものは、数台から数十台のものが多い。

ＰＣの高性能化とＬＡＮの高速化により、量産効果の高いこれらの製品を使ったクラスタマシンに注目する。本文では、約５００台のＰＣをＰＶＭ（Parallel Virtual Machine）によりクラスタシステムとして構成し、その負荷分散と効率について報告する。

2．　ＫＳＵ　Ｌｉｎｕｘ　クラスタ

本学では１９９９年４月から５０８台のＰＣを７教室に設備し、全学の情報教育に活用している。ＣＰＵ４００ＭＨｚ、メモリ１２８ＭＢ、ディスク８ＧＢ、ＬＡＮ１００Base－Tx のＰＣに Linux と NT のＯＳを搭載している。通常は授業使用と自由使用により、１台のＰＣとして使用する。この運用と保守のために、ＬＡＮにより集中的に電源投入切断と状態監視のほか、ＯＳの更新も可能である。クラスタサーバ、ＮＦＳサーバ、ＷＥＢサーバなどは常時稼動している。

この管理機能は学内のすべてのＰＣからも利用資格があれば利用できる。授業時間は教育用に優先であり、ほぼ１：２の割合で Linux の方が少ないが、夜間は全台を Linux として立ち上げることができる。これらをＰＶＭによりクラスタシステムとし、最大５０８台の Linux クラスタを可能としている。

3．　クラスタシステムのベンチマーク

本クラスタシステムは専用ではなく、ユーザや利用ＯＳが勝手に変わり、不意に電源が切られるような過酷な条件でクラスタシステムを構成する。短時間で全ＰＣの状況を把握し、ＰＶＭの動作効率がどの程度であるかを知るためにベンチマークプログラムを作成した。ベンチマークの機能は

（1）全台のＯＳ利用状況一覧
（2）ＰＶＭの開始と終了の処理時間予測
（3）クラスタシステム処理能力の予測

である。

Load Share and Efficiency for Medium Scale Cluster System using PVM
Yoshisuke KUROZUMI, Hiromi HIRAISHI, Tsutomu TAKEUCHI and Eitetsu OOMOTO
Kyoto Sangyo University Faculty of Engineering

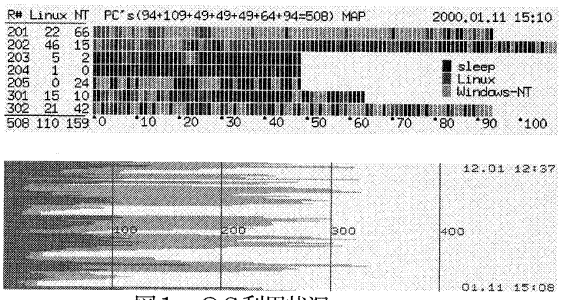

図１　ＯＳ利用状況

５０８台のＯＳ利用状況の一覧を図１に示す。各 PC は未使用，Linux，NT の３状態のいずれかである。Linux として使われていれば、そのホスト名により PVM を立ち上げる。

PVM の開始と終了は PVM コマンドで行い、普通は１台あたり約０．１秒で復帰する。しかし、この間に PC の電源断などの状態変更があると、PVM は数分間のタイムオーバまで待ち、全体で数１０分もかかることがある。この状況をチェックすることができる。

クラスタシステムの処理能力は CPU、メモリ、ディスク、ネットワークの総合性能である。従来からある並列処理用のベンチマークは連立方程式計算や流体解析であり、明らかに本システムのような条件では高性能は望めない。本システムが得意なのは、完全並列に近い問題とか、台数や負荷の変動に簡単に対応できる問題に限られる。そこで、つぎのベンチマークを考えた。

（1）　素数計算
（2）　全メモリ検索
（3）　同時転送

これらの処理を各種のパラメータで実行した結果をつぎに紹介する。

4．素数計算

素数計算は理論と応用に魅力ある問題であり、多くの試みがなされている。ここでは、数ｎまでの素数を求め、その個数を並列処理により計算する。素数を求める方法にエラトステネス法があり、１台のＰＣによる処理時間はｎ＝１０＊＊９までの素数で約２００秒である。処理時間は n*log n に比例することが知られている。これ以上になると３２ビットのＰＣでは桁数とメモリ制限で計算困難となる。

そこで、m台のＰＣで並列処理するため数ｎをm分割し、それぞれのＰＣで素数の個数を求め合計する。この方法で処理した結果を図２に示す。

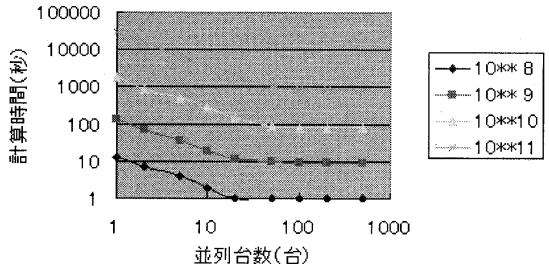

図２　並列台数と計算時間

横軸はＰＣの台数（１－５００台）で縦軸は並列処理による計算時間（秒）である。数ｎの範囲は１０＊＊８から１０＊＊１１までである。ＰＣの台数が２０－５０台までは順調に台数に比例して性能が向上するが、そのあとは台数を増してもほとんど性能向上はない。

クラスタシステムには転送時間を含めＰＶＭなどのクラスタＯＳのオーバヘッドが常に存在する。この問題ではオーバヘッドはつぎのようになる。

分割時間＋送信時間＋素数計算時間＋受信時間
２％　　４％　　９０％　　４％

この例の合計約１０％のオーバヘッドでは上のような現象は起きない。

エラトステネス法では大量のメモリを使用する。 数ｎが１０＊＊９であれば１２５MB、１０＊＊１２であれば１２ＧＢ必要であり、並列処理により、データを分散させる必要がある。

たとえば、１０＊＊９までの計算で、等分割１０台で並列処理し、900000000-1000000000 の間の素数を計算すると１２．５MBのメモリを使い、処理時間は２１秒となる。分割数を大きくし、最後の分割範囲のみに注目し、プロットしたグラフを図３に示す。

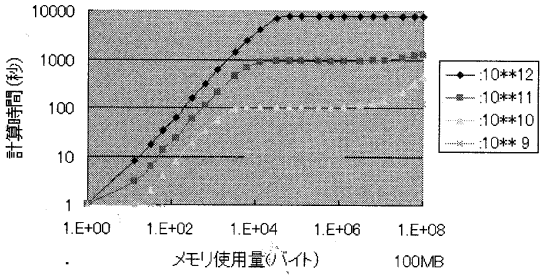

図３　メモリ使用量と計算時間

横軸はメモリ使用量（１K－１００MB）で縦軸は１台のＰＣによる計算時間である。数値の範囲は１０＊＊９から１０＊＊１２までである。メモリ使用量が５００－５００００Ｂ（５０KB）まではメモリ使用量に比例して計算時間が増加する。しかし、そこから３M－１００MBまでは計算時間は増加しない。その後、メモリ使用量が増すと、またまた増加に転じる。

最初の増加部分はキャッシュでの処理であり、つぎのフラット部分はキャッシュとメモリでの処理であり、最後の増加部分はディスクスワップ処理となる。大量のメモリをとびとびにアクセスする問題ではこの例のような現象となり、２次元配列の参照でも起こりうる。

このようにＰＣ台数に対して処理時間が非線形になる問題では、台数が多ければ速くなるとは限らず、最高効率を与える最小台数が存在することになる。これを求める負荷分散アルゴリズムを開発すれば大規模クラスタを多ユーザが効果的に分散使用できるであろう。

５．全メモリ検索

クラスタシステムの効用にメモリの拡大がある。１台のＰＣは１２８MBのメモリであるが、５００台あれば、６０ＧＢを超える容量となる。たとえば、５０ＧＢの文章を全ＰＣのメモリ上に作成し、単純な全文検索したところ４秒でヒット件数の結果を得た。１台１１０MBの検索は２秒であり、５００台のクラスタオーバヘッドが２秒であった。

この処理は並列化効率のよい例で、５０ＧＢのメモリを持つ１台のＰＣでは１０００秒であり、２５０倍の効率となる。単純全文検索ではなく、処理時間の多い複雑な画像検索ではさらに効率アップとなる。

６．同時転送

クラスタシステムでは、ＰＣ間のデータ転送時間が問題である。１００Base-Tx のＬＡＮで接続しているため、少量データを瞬間的に送ることは不得意で、１パケットあたり約５m秒の転送時間である。LAN は本質的にシリアルとなるためパケット数に比例して処理時間は増加する。

また、並列処理の効果をあげるためには、同時に処理終了にすることである。このとき、最後にサーバに結果を集めるために直列処理となり、数バイトのデータでも５００台で約２秒必要である。大量データを転送して終了するときは、サーバに５００台のＰＣから一斉にデータを送ると衝突のため極端に遅くなることがあり、転送時間分ずらす工夫が必要となる。

７．あとがき

中規模クラスタシステムの利用状況を述べた。ＰＣやＬＡＮはますます高性能化するであろう。しかもコストパーフォマンスは改善され、普及するに違いない。

クラスタシステムとしては、ラックマウントＰＣを組み込んだ専用クラスタと本システムのように、各台が単独にも使われる共用クラスタが普及すると思われる。とくに共用クラスタでは、突然の電源投入切断や負荷の激変があり、これに耐えしかもこれを活用できるクラスタＯＳを開発したいし、また、その出現を期待する。

参考文献

[1] Al Geist, et al “PVM : Parallel Virtual Machine” MIT Press,1994

[2] Rajkumar Buyya “High Performance Cluster Computing” Prentice Hall,1999