超並列OS「超流動OS」の超分散化の考察

2H－7

平野　聡　田沼　均　須崎　有康
一杉　裕志　丹羽　竜哉　塚本　享治

電子技術総合研究所

概要　複数の超並列システムを高速ネットワーク(WAN)を介して統合して制御することにより、より大きな問題を高速に解いたり、ネットワークに対してより高度なサービスを提供可能な超並列超分散システムを開発することを計画している。そのため、我々が開発している超並列システムのためのオペレーティングシステムである超流動OSを超分散化することを検討する。ネットワーク内に複数存在するプロセッサ空間を統合し、プロセス割り付けや負荷分散を行う。

1 はじめに

我々は超並列システムのためのオペレーティングシステム「超流動OS」を開発している[1,2]。超流動OSは1000台から100万台に及ぶ台数のプロセッサを備える超並列システム上でアプリケーションプログラムを効率よく実行する技術を開発することを目的とする。これまでに、並列仮想記憶、負荷分散、ソフトウェアの実行環境への自律的適応等の研究を行なってきた。

大規模な計算を行う利用者はより大きな処理能力を駆使したいと望むが、数万台以上のプロセッサを備える超並列システムを実際に構築することはほとんどの利用者にとって難しい状況である。

一方で、光通信やATM等のネットワーク技術の進歩により、近い将来に広域ネットワークでギガビット程度の大きなスループットを得られる見通しとなった。OZ++[3]のような分散処理環境も開発されており、地球上に分散する多くのサーバが協調して高度なサービスを提供する超分散システムが高速ネットワーク上に構築されて行くであろう。より高度なサービスを高速に処理するためは、超並列システムを高速ネットワークで多数結合し統合管理する超並列超分散処理が有望である。そこで、我々は超流動OSの超分散化を構想している。本論文は構想の前段階として、我々が行なってきた研究テーマのうち資源管理を超分散化することを検討する。

Extension of the Fluid Operating System for Massively Parallel and Distributed Processing.
Hirano, Tanuma, Suzaki, Ichisugi, Niwa, Tukamoto, Electro technical Laboratory

2 超流動OSについて

超流動OSの主要な構成要素を以下に示す[2]。

- □ OSポリシー -- 複数プロセスの混合実行による実行効率の向上を目指す。
  - ● 管理情報共有機構MetaShare -- 負荷分散木(LBT)によりシステムワイドの負荷分散を行う。
  - ● 大域的仮想仮想記憶GVVM -- メモリの自動負荷分散を行う。
  - ● プロセス割当法のTSS化 -- プロセスをプロセッサに割り付ける際にTSSと組み合わせることにより、実行効率の向上と応答時間の短縮を図る。
- □ 開発・実行環境 -- あるプロセス内のソフトウェアが実行時の環境や与えられたデータの性質に自律的に適合して性能向上を目指す。
  - ● 最適アルゴリズム自動選択法AASM -- 与えられたデータの性質に最も適合するアルゴリズムを自動選択する。
  - ● 最適並列度推定法 -- データの大きさに最も適合する並列度を自動選択する。
- □ 言語 -- 並列、分散システムを容易に記述することが可能な言語記述系を開発する。

これらの構成要素は基本原理の確認を終え実証用の実装を進めている段階にある。

3 超流動OSの超分散化

地理的に離れた多数の超並列システムを1.2Gb/s程度の高速ネットワーク複数本を介して統合管理することを考える。超流動OSは、効率のよい資源管理を行うために以下の機能を備えていな

ければならない。

- 分散プロセス割り付け
  利用者によって投入されたプロセスは、実行に必要な数のプロセッサとプロセッサ空間上の形状を要求する。超流動OSは投入された超並列システム内部だけではなく、他システムも含めて最も実行効率がよいことが期待されるシステム上の空き領域を発見してプロセスを割り付ける必要がある。
- 複数システムに跨ったプロセス割り付け
  上記のプロセス割り付けで空き領域が見つからなかった場合、あるいは、要求されたプロセッサ台数がどのシステムのプロセッサ台数よりも多かった場合、プロセスの性質によってはひとつのプロセスを複数の超並列システムの空き領域に跨って割り付けても十分な性能が得られる可能性がある。例えば、データのクラスタリングが可能で、クラスタ化された並列システムやワークステーション・クラスタでも性能が得られる問題である。システムに跨ったプロセスの割り付けが有効か否かはその時点でのネットワークの混雑度にも依存する。超流動OSはプロセスと協調して割り付け場所を判断する。それぞれのシステムが同じプログラムのオブジェクトを有することと、ネットワークを介して交換されるメッセージに互換性があることが必用である。ローカルネットワーク間でメッセージをブリッジするハードウェアは性能向上に貢献するであろう。
- 同期したスケジューリング
  複数システムに跨って割り付けられたプロセスは効率のよいメッセージ交換のためそれぞれのシステム上で同一時刻のタイムスライスで実行されなければならない。そのため、システム間で同期可能なスケジューラが必用である。ただし、同期するのは特定プロセッサだけでよい。超流動OSのTSSスケジューラはプロセスを仮想マシン上のプロセッサ空間に割り付け、複数の仮想マシンをタイムスライス毎に順に実行することによってプロセッサ利用率と応答性のよいTSSを実現する。仮想マシンの対象を複数システムに拡張することによりシステム間で同期したスケジューリングとプロセッサ割り付けを行うことを検討している。
- メモリの負荷分散
  超流動OSの大域的仮想仮想記憶GVVMは使用頻度の低いページをメモリ負荷の軽いプロセッサにページアウトすることによってメモリの負荷分散を行う。転送されたメモリは再びページインされるまで使用されないメモリであるため異機種間で転送を行なっても何ら問題はない。従ってGVVMを並列分散に拡張することは比較的容易であろう。
- プロセス、スレッドの負荷分散
  プロセスの内部が異機種に転送して実行可能なオブジェクトによって構成されている場合、広い範囲のシステムを対象としてプロセスの負荷分散が可能である。負荷分散木LBTによる負荷分散法は低い探索コストで負荷の低いプロセッサ領域をより近くに発見することを目指している。複数システムを対象にLBTの分散データ構造を構築するよう拡張することにより、システムに跨った負荷分散を行うことを検討している。

4 終わりに

負荷分散で考慮するべきことは多い。プロセッサの負荷、メモリの負荷、動かそうとしているプロセスの性質、回りのプロセスの性質等がある。広大、不均質で、かつ広域ネットワークの込み具合次第で「近さ」が変化するプロセッサ空間を対象として複雑な負荷分散戦略を予め記述することは不可能であろう。従って、経験を蓄積し負荷分散戦略を学習してゆく仕組みの開発が超並列超分散の大きなテーマである。

謝辞 本研究の一部はリアルワールドコンピューティングプログラムによって行われたものである。

参考文献

[1] 平野、田沼、須崎、濱崎、塚本, 超並列システム用オペレーティングシステム「超流動OS」の構想、情報処理学会研究報告 93-OS-58
[2] http://www.etl.go.jp/Organization/Bunsan/Fluid/Fluid.html
[3] http://www.etl.go.jp/Organization/Bunsan/OZ/OZ.html