超流動OS開発用超並列マシンシミュレータ

3T－8

平野 聡　田沼 均　須崎 有康　一杉 裕志　塚本 亨治
電子技術総合研究所

1　はじめに

超並列システム用オペレーティングシステム「超流動OS」[7]開発のために作成した、10000プロセッサ程度の規模の超並列システムのシミュレータERSEの概要について述べる。ERSEは超流動OSの機能要素や並列アプリケーションを実行し、ネットワークの転送特性を反映した仮想時間による性能測定環境を提供する。

多数のプロセッサを備える超並列システムのシステム開発には、並列システム・シミュレータが欠かせない。目的によりシミュレータの提供する機能は以下のように異なる。

1. ネットワークの性能測定用。プログラムのトレースから大規模ネットワークの転送特性を評価する。プログラムの実行はできない。例、INSIGHT[3]。
2. OS、アプリケーションの開発用。プログラムの実行が可能である。
   (a) インストラクションレベルのシミュレータ。OSの核やコンパイラを開発するために用いる。プログラムの実行は低速であるため大規模なシステムのシミュレーションは不可能。
   (b) プログラムレベルのシミュレータ。プログラムはネイティブコードで高速実行する代わりに、対象マシンのインストラクションの実行をエミュレートする機能はない。
      i. ネットワーク・エミュレーションを行なわないシミュレータ。プログラムの機能レベルの開発を目的とする。例、COSの開発環境[5]。
      ii. ネットワーク・エミュレーション機能を有するシミュレータ。プログラムの機能レベルの開発に加えて、ネットワークの転送特性を含めた性能評価が可能なもの。

ERSE (ETL RWC-1 Simulation Environment)の対象マシンのひとつはRWC-1[2]である。OSやアプリケーションの開発では、RWC-1システム全体を含む大規模なシミュレーションを行なう必要がある。そのシミュレータは、

- PE数が多量であっても、許容できる処理時間でプログラムの実行が行なえること。
- RWC-1のRICAアーキテクチャやネットワークトポロジ(CCCB網[2])の性能上の特徴を反映した性能評価が可能なこと。

を満たす必要がある。

そこで、ERSEは上記分類で2-b-iiのように、ネットワーク・エミュレーション機能を有し、プログラムレベルでの実行と仮想時間による性能測定を行なうシミュレータとした。ユーザはC++による並列オブジェクトの集合として開発・評価したいプログラムを記述し、ERSE上で実行する。

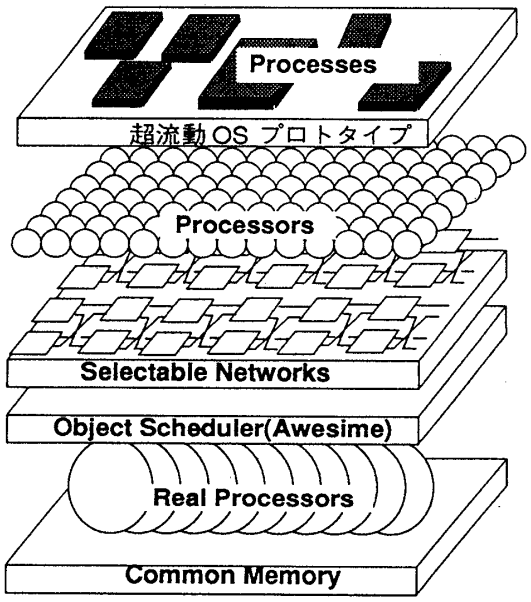

図1:ERSEの全体構成

2　ERSEの全体構成

ERSEは、プログラムに対してRICAやCCCB網の一部の機能を提供する。プログラムは、ERSEを動作させる計算機のネイティブコードで動作するため、インストラクションレベルのシミュレーションを行なうシミュレータと比較して高速である。

ERSEの構成を図1に示す。各要素の役割は以下の通り。

プロセス ユーザのプログラムはERSEのプロセスとなる。プロセスは並列オブジェクト(スレッド)の集合である。RWC-1のシステム記述言語であるMPC++[4]との互換性、移行性のためにマルチスレッド実行が可能なC++で記述する。マルチスレッド化はユーザのクラスが後述のAwesimeのクラスを継承することにより実現される。

超流動OSプロトタイプ 超流動OSプロトタイプはプロセス・スケジューラ、性能評価用の大域的仮想仮想記憶機構、管理情報共有機構「MetaShare」を含んでいる。これらもERSEのプロセスとして実装されてい

MPP Simulation Environment for "Fluid", An OS for Massively Parallel Systems

Hirano S.,Tanuma H.,Suzaki K.,Ichisugi Y.,Tsukamoto M., Electrotechnical Laboratory, Japan

る。プロセス・スケジューラはプロセスをPE空間に割り付ける。

(注:RWC-1にはRWCPのSCOREカーネルがのる)

プロセッサ 1台から10000台程度の仮想プロセッサ(PE)である。プロセスには複数のPEが割り当てられる。プロセス内の各スレッドにはsend/receiveによるネットワーク・アクセスのためのAPIが提供される。

選択可能なネットワーク ネットワークをエミュレートする。ネットワークはシステムオブジェクト群として実装されている。ネットワークトポロジは実行時に選択可能である。現在二次元トーラスとRWC-1のCCCB網が実装されている。

オブジェクトスケジューラ 並列オブジェクトの仮想時間スケジューリングにはC++のマルチスレッド・ライブラリであるAwesime[1]を改造して用いている。

実プロセッサ群 ERSEはUnix上の1プロセスである。

ERSEは、SunOS4の場合は1台、Solaris2の場合は複数の実プロセッサを使用してオブジェクトを実行する。

共有メモリ ERSE用のメモリ。アプリケーションのデバッグや評価に便利なよう、全てのプロセス、オブジェクトからアクセス可能な共有メモリである。

他に、ERSEの構成を定義するファイルがあり、使用するネットワークトポロジ、ネットワークの転送速度、バッファ長などの定義を与える。

ERSEの実行時には上記の構成要素全てを一緒にリンクし、Unixの一つのプロセスとして実行する。現在ERSEは、Sunのシングル及びマルチプロセッサのワークステーションとCRAY CS6400 (Solaris2互換機)上で動作している。1024PE構成のRWC-1をシミュレートするERSEは約200MBのメモリを消費する。我々が使用しているCRAYの場合、実メモリは4GBであるため、装備する32台のプロセッサを使用して16000PE程度までのシミュレーションが可能である。デバッグにはGDBを用いる。

3 ERSEの実行モデル

性能評価のため、オブジェクト(スレッド)はERSEが提供する仮想時間を基準として動作する。ある仮想時間に動作可能なオブジェクトの集合をイベントセットと呼ぶ。オブジェクトスケジューラはイベントセット内のオブジェクトを実プロセッサに割り付け実行する。イベントセットが空になると仮想時間が進行し次のイベントセットに移行する。即ち、ERSEの仮想時間管理は集中型であり、見込み実行やロールバックは行なわない。

オブジェクトが発生する非同期メッセージは網に送られ、通信先のオブジェクトに到着する。メッセージを待っていた受け側オブジェクトは動作を開始する。この間にメッセージパッシングに伴う仮想時間が経過する。オブジェクトは自分自身が必要な(RWC-1で消費されるであろう)仮想時間の間、実行権を放棄することにより仮想時間を消費する。つまり、ネットワーク部分の仮想時間の進行はERSEが行なうが、ユーザプログラムについてはコード中に仮想時間を消費する簡単なコードを埋め込む必要がある。これはプログラムがネイティブ・コードで動作するために必要な処置である。

4 ERSEのネットワークの構成

RWC-1のネットワークであるCCCB網はX-Y平面をバンヤン結合、Y-Z平面をキューブ結合とする。直接網であるため、ルータはPEに内蔵される。ルータは入力4本、出力4本(PE-網のリンク、X,Y,Z次元方向のリンク)を有する。Store & Forwardデッドロックを防ぐため、PE用リンク以外の各リンクは3系統の入力バッファを有しリンクを共有している。リンクの使用権の競合は(現在の実装では) FIFOアービタによって解決される。RWC-1は網からPEへ入るリンクに4種類のプライオリティ付メッセージキューを設けているが、ERSEでは実装の都合上、各スレッドにメッセージキューを設けている。オブジェクトの発生するメッセージは1ワードから16ワードの大きさのパケットに分解されて網に投入される。ルータ内ではアービトレーションに1仮想時間かかり、リンクの通過にパケット長×1仮想時間かかる。パケットの転送はCut & Throughであるため、リンクの先のバッファが空の場合はすぐに次のルーティングが始まる。ブロックする場合は待たされる。

各ノードにおけるパケットの衝突の頻度、バッファの平均使用量等の統計情報の測定が可能である。

5 おわりに

以上、超流動OSの開発用に作成した超並列マシンのシミュレータERSEの概要について述べた。ERSEは既に1(実)プロセッサ上で動作しており、大域的仮想仮想記憶の評価等に使用している[6]。複数プロセッサを使用する版は現在作成中である。

謝辞 本研究の一部はRWC計画の一環として「超並列システムアーキテクチャに関する研究」で行なわれたものである。

参考文献

[1] D. Grunwald. A Users Guide to Awesime: An Object Oriented Parallel Programming and Simulation System. *Univ. of Colorado Boulder Tech. Report CU-CS-552-91*, 1991.

[2] 坂井修一, 岡本一晃, 横田隆史ほか. 超並列計算機RWC-1の基本構想. *JSPP'93*, pp. 87–94, 1993.

[3] 柴村, 久我, 末吉. 超並列計算機のための相互結合網シミュレータ. 情報処理学会論文誌, Vol. 5, No. 4, pp. 589–599, 1994.

[4] 石川裕, 堀敦司ほか. 並列プログラミング言語MPC++の実現. *JSPP'94*, pp. 105–112, 1994.

[5] 村山正之, 斉藤信男. 協調的オペレーティングシステムCOSの開発環境. *JSPP'94*, pp. 357–364, 1994.

[6] 平野聡, 一杉裕志, 田沼均, 須崎有康, 塚本亨治. 大域的仮想仮想記憶(GVVM)のRWC-1上での性能予測. 情報処理学会研究報告 *94-OS-65 (SWoPP'94)*, pp. 121–128, 1994.

[7] 平野聡, 田沼均, 須崎有康, 濱崎陽一, 塚本享治. 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*, pp. 17–24, 1993.