分散並列処理のためのプラットフォーム Lemuria の評価

4 Z − 6

斎藤 彰一　　大久保 英嗣
立命館大学理工学部情報学科

1 はじめに

我々は, 分散環境上で並列処理を行うためのプラットフォームである Lemuria を開発している [1]. Lemuria は, 分散共有メモリに基づいた分散並列処理システムである.

複数のマシンで構成される分散環境において, 利用されていないマシンなどの資源を用いて処理の分散化を行うことは, 処理そのものの高速化が実現されるだけでなく, システム全体の有効利用を図ることができる. Lemuria は, そのような分散環境を容易に実現することを目的としている

Lemuria は, 種々のシステム上でも動作可能とするために, 各々のシステムを変更することなく実現されている. ユーザレベルで動作する複数のデーモンとユーザプロセスにリンクするライブラリによって構成され, 既存のオペレーティングシステム上で動作するミドルウェアである.

本稿では, SPLASH2[2] による, Lemuria の分散共有メモリシステムの評価について述べる.

2 Lemuria

Lemuria は, 1) マシン内; 2) 複数のマシンによって構成されたクラスタ (Lemuria Cluster: 以下 Cluster), 3) クラスタ間, の 3 つの階層を持つ構造になっている. Lemuria の構成を図 1に示す. Cluster は, イーサネットの同一セグメントに属するマシン群によって構成される. Cluster 間の通信はすべて Reflector と呼ばれるデーモンによって中継される. Reflector は, 中継時にページデータのキャッシュと通信の取りまとめを行う. 各マシンには, Lemuriad と呼ばれるデーモンが配置され, マシン内の分散共有メモリの一貫性制御と, Cluster 内の他のマシンとの通信を行う. Lemuria で発生する通信は, これらの階層毎に Lemuriad と Reflector によってまとめられて行われる. 階層を越えて直接行われることはない. Lemuria の分散共有メモリシステムに実装されている一貫性制御プロトコルは, Write-Invalidate 方式である.

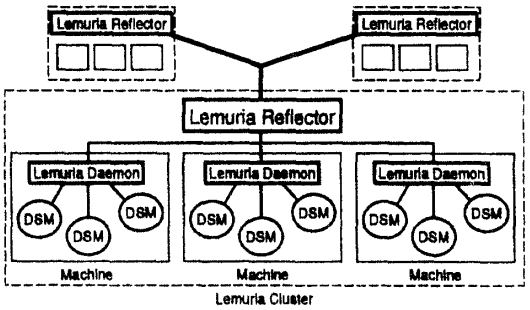

図 1: Lemuria の構成

3 実験環境

実験に使用した計算機は, SONY 製 News5000(主記憶 32Mbytes) である. ネットワークはイーサネット (10M) によって接続され, 1 セグメントに 19 台ずつ配置されている. さらに, ワークステーションルータによって 4 セグメントが接続されている (図 2参照). この構成が 4 組あり, 合計 16 セグメントから構成されている. 実験では, これらのセグメント毎に, 1 つの Reflector を配置し, それを Cluster としている.

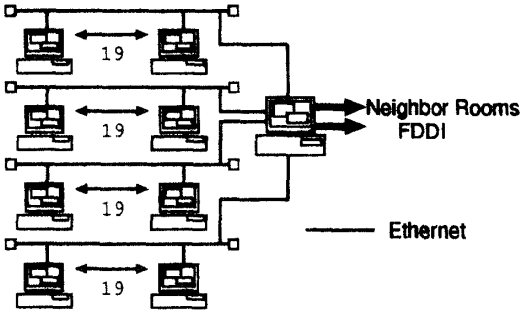

図 2: 実験ネットワークの構成

4 マシンの台数効果

Lemuria によるマシンの台数効果の測定について述べる. 実験ネットワークは, 4 つの Cluster を構成し, マシンを各 Cluster に均等に配置している. 測定に使用

An Evaluation of Lemuria, A Platform for Distributed Parallel Processing
Shoichi Saito and Eiji Okubo
Department of Computer Science, Faculty of Science and Engineering, Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu, Shiga 525, Japan

したアプリケーションは, SPLASH2 の FFT(データ数 $262{,}144 = 2^{18}$) を使用した. マシン数は, 8 台, 16 台, 32 台, 64 台の各々について測定を行った. 測定結果を図 3 に示す

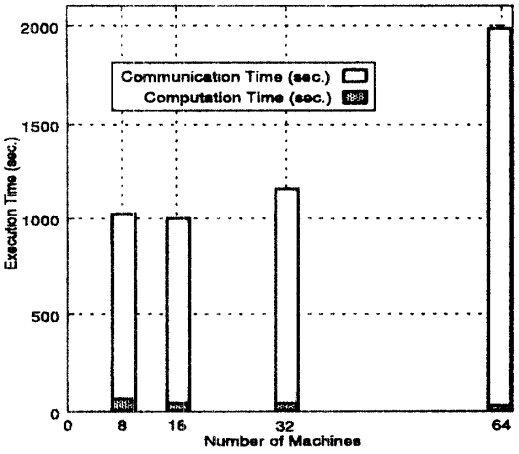

図 3: FFT

本実験では, 台数効果は現れていない. 原因としては, 処理中に全データのコピーが数回発生していることが上げられる. このために, ページ転送に要する時間が処理時間の 90 %以上に達した. しかし, 64 台での処理時間は 8 台での処理時間の約 2 倍に留まっている. これは通信の取りまとめによって, ページの供給を行うマシンへのページ要求の数が減少したことと, ページのキャッシュによって, ページの供給を行うマシンが複数になったためである. これらは Reflector の効果によるものである.

5 Cluster 分割による速度変化

Lemuria の Cluster はイーサネットのセグメント毎に構成されている. 従って, Cluster に属するマシン数が増加した場合に, ネットワークの使用が増加し輻輳を発生される場合がある. そのため, 適度なマシン数によって Cluster を構成する必要がある.

処理を行うマシン数を固定して, Cluster を構成するマシン数と Cluster 数を変化させた場合の, 処理速度の変化について述べる. なお, すべての Cluster は, 同じ数のマシンによって構成されている. 図 4は, 各マシン数において, Cluster 数が最小の場合の結果を 1 として (例えば, 64 台の場合は 4Clusters), Cluster 数増化による処理時間の比の変化の割合を表したものである.

同一マシン数で処理を行った場合でも, Cluster 毎のマシン数によって処理時間に大きな差が見られた. Cluster 数を増加させることで Reflector によるページのキャッシュや通信の取りまとめの効果が増加しているものと考

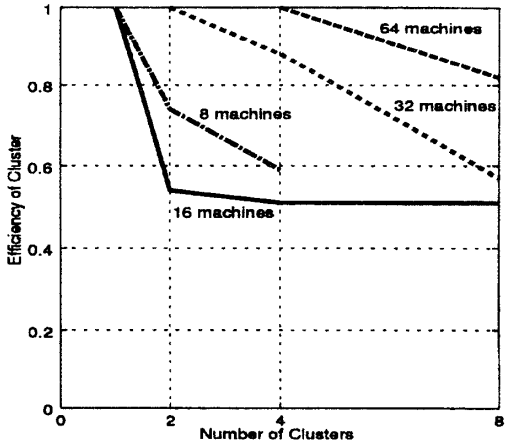

図 4: Cluster 分割による速度変化

えられる. また, Cluster 内での通信量と Cluster 間の通信量の関係によっても変化すると考えられる. どちらかが多くなっても当該ネットワークが輻輳を発生させ通信時間の増加を招き, 結果として処理時間が増加すると考えられる. この結果から, 処理に使用するマシン数によって, Cluster 数と Cluster を構成するマシン数を決定することができる.

6 おわりに

本稿では, SPLASH2 による Lemuria の分散共有メモリシステムの評価について述べた. Reflector や Cluster 分割によって処理が高速になることについて示した. しかし, FFT では, 処理時間に対して通信時間の占める割合が大きく, 台数効果が現れていない. 他のアプリケーションについても評価を行い, 必要ならばこの点に関して改良を検討する.

また, 今後のネットワークは, スイッチングハブを中心とした 100M イーサネットや, ATM などに変わると思われる. 現状の Lemuria の構成はシェアード型のイーサネットを前提としたものである. 今後, これらのスイッチ型ネットワーク上での Cluster や Reflector の構成を検討し, 改良を行い再度評価を行う予定である.

参考文献

[1] 斎藤 彰一, 大久保 英嗣: 分散並列処理のためのプラットフォーム Lemuria の構成, 電子情報通信学会技術研究報告 Vol. 96, No. 33, pp. 31–36 (1996).

[2] Steven Cameron Woo, Moriyoshi Ohara, Evan Torrie, Jaswinder Pal Singh, and Anoop Gupta: The SPLASH-2 Programs: Characterization and Methodological Considerations. In Proceedings of the 22nd International Symposium on Computer Architecture, pp 24–36 (1995).