B-029

クラスタをメモリ資源として利用するための動的メモリ提供システムの提案

A Dynamic Memory Allocation System to Utilize a Cluster as a Memory Resource

三浦望† 緑川博子† 甲斐宗徳†

Nozomu Miura Hiroko Midorikawa Munenori Kai

1．まえがき

筆者らはローカルの物理メモリサイズに制限されることなく，クラスタの遠隔ノードのメモリを利用し仮想的に大容量のメモリ空間を提供するシステム，DLM(Distributed Large Memory)を構築してきた[1].

通常の OS では，アプリケーションプログラムがローカルメモリサイズ以上のメモリを利用する場合，ハードディスク内のスワップ領域を用いて不足分を補っていた．しかし，このようなハードディスクへのアクセスは非常に低速であるため，実用には向かない．一方，近年の高速ネットワークによりローカルハードディスクに比べ遠隔メモリへのアクセスのほうが高速になってきている．ハードディスクへのスワップを遠隔ノードのメモリへ展開させるようにしたものが DLM である．

筆者らはマルチクライアント向け分散型大容量メモリシステム DLM-M(DLM for Multi-client)[2]を構築し，複数のクライアントに対しローカルメモリサイズを超えるプログラムの実行を可能とするシステムを提供してきた．このシステムでは，各クライアントが使用する遠隔ホスト名(メモリサーバ)とその遠隔ホストでの使用メモリサイズを明示的に指定するため，複数クライアント間の調整は行われておらず，特定のメモリサーバへの負荷の集中などが起こる可能性もあった．

そこで，クラスタ全体でメモリサーバの割付や負荷の分散などを効率的に行うために，クラスタ全体の負荷状況を判断し適切なメモリサーバ自動割付を行う，動的メモリ提供システムを提案する．本提案システムでは，クライアントから遠隔メモリサーバを明示的に指定するのではなく，アプリケーションが必要とする総メモリサイズだけを指定するようになっている．

本報告では，システムの設計と初期実装について述べる．

2．マルチクライアント向け分散型大容量メモリ

マルチクライアント向け分散型大容量メモリ(DLM-M)は逐次向けアプリケーション用に開発された，マルチクライアント対応メモリ提供システムである．図 1 のように，クライアントアプリケーションの計算を行うノードを計算ノード，遠隔メモリを提供するノードをメモリサーバと言う．それぞれに計算プロセスとメモリサーバプロセスが実行されており，メモリサーバプロセスはマルチクライアントに対応しているので計算プロセスの要求に応じ，専用の処理プロセスを必要に応じて生成する．ユーザのアプリケーションコードは計算プロセスの計算スレッド内で行われる．DLM-M には以下の特徴がある．

- 各メモリサーバプロセスの起動をオフラインで明示的に行う
- クライアントアプリケーションが使用するメモリサーバを，設定ファイル(図 2)を用いてユーザが明示的に指定する．

設定ファイルは，利用するメモリサーバ名 (または IP アドレス)，利用するメモリサイズ (MB) を含んでいる．

他のクライアントの利用状況によらず，固定的な記述であるため，クラスタ内で同じメモリサーバへ複数クライアントからのアクセスが集中し，処理速度の低下を招く場合や，未使用のメモリサーバが存在するなどの非効率な状況が起こりうる．

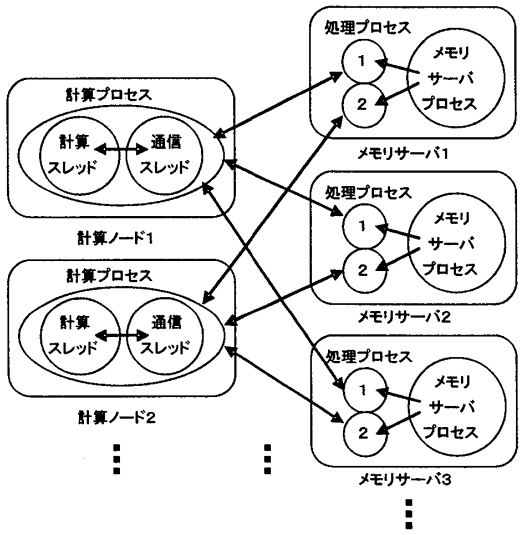

図 1　DLM-M システム構成

```
calnode     2048   // 2GB  計算ノード
memhost1    2048   // 2GB  メモリサーバ1
memhost2    4096   // 4GB  メモリサーバ2
memhost3    4096   // 4GB  メモリサーバ3
```

図 2　ユーザアプリケーション設定ファイル(dlmm.conf)

3．メモリサーバ自動割付システムの設計

新たに提案するメモリサーバ自動割付システムでは，メモリサーバを自動で選定しクライアントに通知する管理プロセスを導入する(図 3)．管理プロセスの起動しているノードを管理ノードとここで呼ぶ．

メモリサーバ自動割付システムは

- メモリサーバプロセスは管理プロセスの起動時に遠隔生成される

†成蹊大学工学研究科情報処理専攻，Graduate School of Engineering, Seikei University

・ 計算プロセスの要求によりメモリサーバを自動選定する

という特徴がある．

メモリサーバ自動割付システムはメモリサーバプロセスの状況を監視し，クラスタ全体で効率的な資源の利用ができるように適切なメモリサーバをクライアントへ通知する．

計算プロセスは自動選定されたメモリサーバを用いて従来の DLM-M の機構を利用し，クライアントアプリケーションを実行する．

4．メモリ自動割付システムの実装

新たに導入した管理プロセスと計算プロセス／メモリサーバプロセスのそれぞれの通信には UDP を用いた．これは，計算ノードとメモリサーバの通信に比べ，通信量が少なく，現在運用を想定している LAN 環境においてはパケットの欠損の可能性が低いと考えるためである．計算プロセスとメモリサーバプロセス間の通信には従来どおり TCP を用いている．

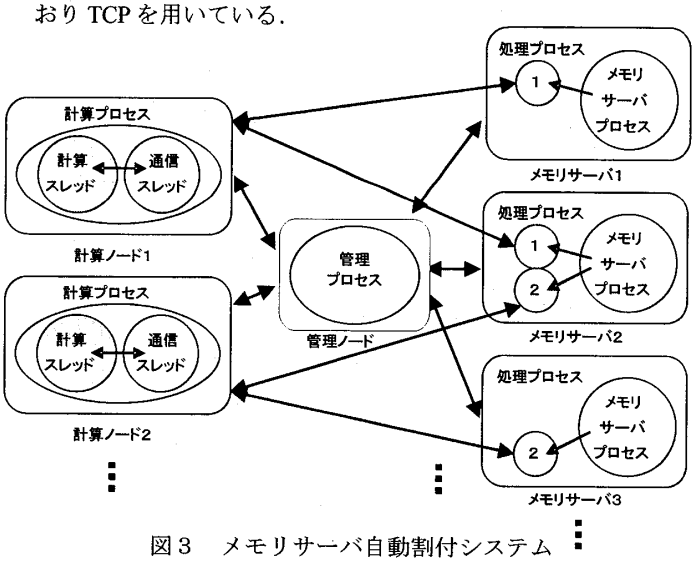

図３　メモリサーバ自動割付システム

4.1 管理プロセス

管理プロセスの起動コマンドは図４ のように引数として設定ファイル（dlmm-admin.conf）を渡す．DLM-M の設定ファイルと同様にメモリサーバ名と利用するメモリサイズを記述する．この設定ファイルに従って，管理プロセスは各メモリサーバプロセスの遠隔起動を行う．すべてのメモリサーバプロセスから起動完了メッセージ受信すると，管理プロセス内部にメモリサーバリストを構築し，クライアントアプリケーションからの要求に応答できる状態となる．メモリサーバリストは各メモリサーバの利用可能メモリサイズ，現在の使用メモリサイズ，サービス中のクライアント数などを保持する．

4.2 クライアントアプリケーションの実行

クライアントアプリケーションは図４のように，引数で管理ノード名(-s)，利用する総メモリサイズ(-m)，そしてローカルメモリサイズ(-l)を指定する．ユーザにはどの遠隔メモリサーバを利用しているかなどの詳細は隠されている．実装上は DLM-M の初期化関数内で以下に述べる処理が行われる．

クライアントアプリケーションが実行される計算プロセスは，実行に必要な遠隔メモリのサイズ(利用する総メモリサイズからローカルメモリサイズを差し引いた値)を管理ノードへ要求する．管理ノードは要求を受け取ると，要求メモリ量に応じ，メモリサーバリストから条件の良いメモリサーバを選択し計算プロセスへホスト名と利用可能メモリサイズを送信する．そして，管理プロセス内部に保持しているメモリサーバリストの情報を更新する．現在の選定基準は単純に，サービス中のクライアント数が少ないものとしている．

計算プロセスは使用可能メモリサーバデータを受け取ると，指定メモリサーバと接続を確立し遠隔メモリを使用して計算スレッドでユーザプログラムを実行する．それ以降，計算プロセスは管理プロセスを介さず直接メモリサーバプロセスとデータのやり取りを行う．

クライアントアプリケーションコードの実行が終了すると，計算プロセスは終了処理として管理プロセスへ終了通知を送信する．管理プロセスは通知を受け取りメモリサーバリストの情報を更新する．

```
DLM-M Admin 起動コマンド>
    dlmm_admin   dlmm-admin.conf
クライアントの実行 >
    User_program  -- -s admin_host –m 1000 –l 400
```

図４　DLM-M Admin とクライアントの起動

5. 初期実装の動作確認

メモリ自動割付システムの初期実装の動作確認をローカルメモリ 1GB のノード 14 個からなる小規模のクラスタにおいて，11 メモリサーバ，1 管理ノード，2 計算ノードとして，1.4GB のメモリを必要とするアプリケーションプログラムを複数稼動させて動作を確認した．

6. おわりに

今回はシステムの提案と初期実装について行った．現在，サーバ割り付けルールは処理中クライアントの少ないメモリサーバからの割り当てを採用している．しかし，今後は CPU の使用率やネットワークの混雑具合なども考慮したメモリ割り付け手法を構築していく予定である．

また，現在は単一クラスタ内におけるメモリサーバの選定のみを行っているが，WAN で接続されたクラスタ群から適切なクラスタを選定し，そのクラスタ内でのメモリサーバの割り付け，さらに計算ノードの割り付けを行うことも視野に入れている．これにより，クライアントアプリケーションの実行時の状況に応じて，ユーザには見えない形で，適切な資源を選び効率的な実行を可能にする．

参考文献

[1]緑川, 黒川, 姫野：“遠隔メモリを利用する分散大容量メモリシステム DLM の設計と 10GbEthernet における初期性能評価”, 情報処理学会論文誌コンピューティングシステム, Vol.1, No.3, pp.136-157（2008,12）

[2]齋藤, 緑川, 甲斐：“マルチクライアント向け分散型大容量メモリシステム DLM-M の設計と実装”,情報科学技術フォーラム FIT2008, FIT 論文集,C-003, pp.199-200,（2008,9）