遠隔メモリを利用する大容量メモリシステム DLM と DLM コンパイラ

緑川 博子† 黒川 原佳‡ 姫野 龍太郎§

1. はじめに

64bit の OS や CPU の普及により，桁違いに大きなアドレス空間を使えるようになってきた．しかし，１台のコンピュータで提供できる物理メモリにはスロット数などのハードウエア制約もあり，大容量物理メモリ搭載マシンは非常に高価になる．物理メモリ不足時には swap 領域（通常，ローカルディスクのファイル）を大容量化することにより，大データを扱うプログラムの実行は可能だが，実メモリにデータが全て収まる場合に比べ非常に低速になる．

このように１台のホストの搭載メモリ量に制限がある環境でも，ネットワークにつながれたホストの比較的小容量の遠隔メモリを集めて，仮想的に大容量メモリとして利用することが可能な分散型大容量メモリシステム DLM (Distributed Large Memory) を構築した[1]．物理メモリを超える大容量データを扱いたいときには，ユーザが従来の C 逐次プログラムに，dlm 宣言を付加するだけで（図２参照)， dlm 指定データを遠隔メモリ上に分散されたデータとして,ユーザには透過的に,プログラム実行を行う．このためのプログラム自動変換のための DLM コンパイラも構築した．

１GbE および 10GbE 結合クラスタでは，通常 OS の swap ファイル利用時に比べ，DLM では 5～10 倍の性能向上が得られた[1]．本稿ではさらに，ローカルメモリのみを使う通常プログラムと遠隔メモリを一部用いた DLM プログラムとで，どの程度の性能低下があるか，10GbE 結合のクラスタ上で評価した．

2. DLM システムの基本構造

DLM では，外部遠隔メモリをあたかもローカル物理メモリの swap 領域（DLM-swap 領域と呼ぶことにする）のように利用する．従来の OS に組み込まれた OS swap 機構では，実メモリが足りないときには通常，ローカルハードディスクなどのファイルを swap 領域として用いるが，DLM swap 機構は，遠隔メモリとデータを転送して，直接ローカルメモリにマップ，アンマップする．したがって，DLM システムの内部処理には，ハードディスクなどへの直接的ファイルアクセス処理は存在しない．

ただし，DLM システムの DLM-swap 処理は，OS の swap 処理を全て取って代わることを意図して設計されたものではなく，全く独立に動いている．すなわち従来の OS が用いていた swap デバイス（ローカルファイルシステム）を直接，遠隔メモリデバイスに置き換えるというような研究とは異なっている[3]．DLM では，場合によっては（dlm 宣言指定なしデータサイズが実メモリ量を超えたときなど)，OS の swap 領域（ファイル）と DLM-swap 領域（遠隔メモリ）が同時に存在することも可能である．この形態は DLM 本来の使い方としては性能上，推奨された形ではないが，ユーザがプログラム全体サイズやローカルメモリサイズを理解していなくても安全に動かすことができる，移植性の点でメリットはある．

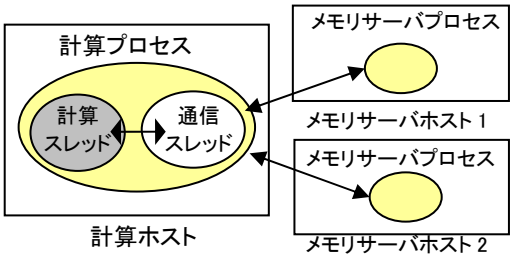

図１ DLM システムの構成 :

図１は，DLM システムの概念図で，ユーザ逐次プログラムコードは計算ホスト内の計算スレッドで行われる．プログラム実行開始時にメモリサーバプロセスが１台以上のメモリサーバホストに，通信スレッドが計算ホストにそれぞれ自動生成され，実行中の遠隔メモリと計算ホストのメモリとの間のデータ転送を行う．終了時にはこれらは自動的に終了処理される．いずれもユーザには透過的に行われる．

† 成蹊大学 理工学部 情報科学科 Department of Computer and Information Science, Seikei University
‡ 理化学研究所 情報基盤センターAdvance Center for Computing and Communication, RIKEN
§ 理化学研究所 次世代計算科学研究開発プログラム Research Program for Computational Science, RIKEN

3. ローカルメモリ使用プログラムと DLM の比較

実験では，表１の理化学研究所次世代計算科学研究開発プログラムのクラスタ（CSLM）を使用した．

表１　10GbEthernet クラスタ　(CSLM)

Cluster	HP DL585 G2 x 5 Nodes
Node CPU	DualCore AMD Opteron(8220SE) 2.8GHz x 4 (8Cores)
Node Memory	64GiByte(67.1GB)
OS	Linux　kernel 2.6.9-42　x86_64
Network	10GbEthernet protocol (Myri-10G)
Hard Disk	SAS　147GB 10krpm 2 台　RAID1 Smart array 5i, HP 431958-B21 (TransRate 300MBps, seektime 4(Ave)8(max)ms)

```
#define N   ( (long int) 8000000000)
dlm int array[N];    dlm 宣言　このデータに DLM を使用
int main(int argc,char *argv[])
{
 unsigned long int i,j;
 for(i=0;i<N;i++)    array[i]=i;    // ①1st sequential access
 for(i=0;i<N;i+=1024) array[i]=-1;    // ② per page access
 for(j=0;j<N;j++) array[j]=j;    // ③ 2nd sequential access
 return 0;
}
```

図２　DLM プログラム例（静的宣言,一次元配列例）

図２のような，8G 個要素の整数配列(32GB) の①初回連続アクセス，②1 ページ(4KB)毎離散アクセス，③再度の連続アクセスを含むプログラム例の結果を示す．32GB の dlm データのうち，ローカルメモリと遠隔メモリの割合を変化させて，すべてがローカルメモリ（通常）の場合との速度比を計測した．

図３は①～③各処理における遠隔メモリとの swap 回数を示す（DLM ページサイズ 32KB）．①～③はデータ領域先頭からの書き込みで，①はローカルメモリアクセスに続き，遠隔メモリアクセスが起こり，全体に占める遠隔メモリの割合に比例して遠隔ページとの swap 回数が増える．②は OS のページ単位（4KB）に１回の書き込みを行う．DLM ではこの場合 DLM ページの連続 swap が起き，遠隔メモリ使用割合によらず一定回数(DLM ページが 32KB ならば 8 回書き込みに１回)の swap が起こる．③も swap out 済みデータへの再アクセスなので最初から swap が発生する．図４は DLM ページサイズを変化させた場合の①の性能（TCP 使用時）を示す．1 回の転送サイズは増えるが，ページサイズが大きくなるほど swap 回数が減り，性能は高くなる．DLM の性能はキャッシュと同様にメモリアクセス局所性によるが，図４の場合，DLM ページサイズが 32KB の時，遠隔メモリアクセス頻度が全体の 5%以下ならば性能は 80%以上，20%程度ならば性能は半分程度であった．

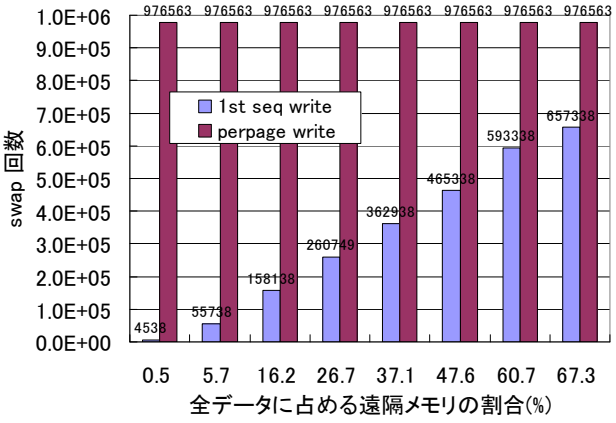

図３　処理①②③swap 回数（②③は同数で一定）

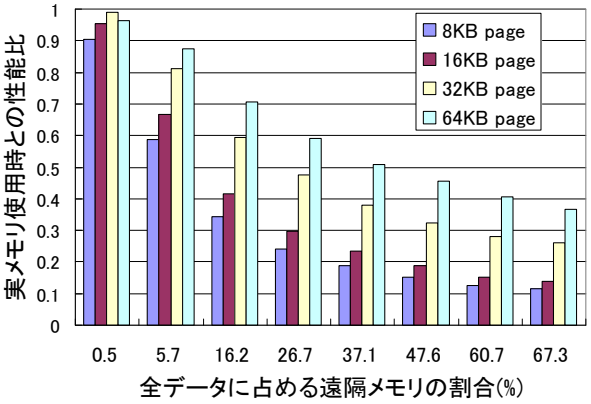

図４　①における DLM ページサイズ（TCP）の影響

4. おわりに

現在，NPB など応用プログラムにおける性能を調査中であるが[2]，メモリアクセス局所性が存在することが多く，搭載メモリを越える容量のデータを扱う処理に DLM が利用できる可能性を示した．

参　考　文　献

[1] 緑川,小山,黒川,姫野，"分散大容量メモリシステム DLM の設計と DLM コンパイラの構築"，信学会 CPSY,信学技報 Vol.102, No.398, pp.29-34, Dec.2007

[2] 緑川,小山,黒川,姫野，"遠隔メモリを利用する大容量メモリシステム DLM とコンパイラ"，情報処理学会, HPC 研究会資料 HPC115-7,　May.2008

[3] 北村，松葉，石川，“大規模メモリ空間の利用を支援する遠隔スワップメモリシステム，” 情報処理学会研究報告, 2007-HPC-111(21), pp.121-126, Aug. 2007.