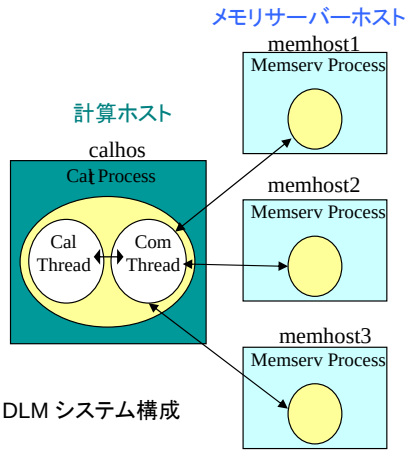
分散大容量メモリシステムDLM の初期性能評価

緑川 博子, 小山 浩生（成蹊大）， 黒川 原佳， 姫野 龍太郎（理研）

現状： 64BitOSの普及 広大なメモリ空間 X86_64(AMD64, Intel64)現実装で 256tebibytes(2^{48})
大容量データプログラムの実行： 物理メモリ、swapファイルサイズで制限を受ける

DLM システム ネットワークで結ばれたコンピュータの遠隔物理メモリを集め， **仮想的な大規模メモリを構築**
→ 大規模データを扱うプログラムの実行が可能



DLM システム構成

DLM設定ファイル
使用ホストと使用メモリ容量を指定する

hostfile

```
calhost    2048  // 2GB
memhost1 8192  // 8GB
memhost2  4096  // 4GB
memhost3  4096  // 4GB
memhost4  4096  // 4GB
   :
```

先頭行はプログラム実行ホスト
DLMに使用するメモリサイズを指定

2行目以降はメモリサーバに使用する遠隔
ホストとDLMに提供できるメモリサイズ

hostfileの先頭4行を用い，
計算ホストとメモリサーバ3台
を使用する指定例

プログラム実行コマンド例

実行例 prog -- -n 4 –f hostfile

DLM コンパイラ，DLMライブラリ 通常プログラムに**dlm**を加えるだけで利用可能

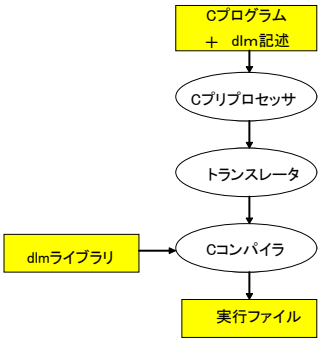
コンパイル例 dlmc test0.c -ldlm

DLMプログラム例1（行列ベクトル積 matv.c）
DLM静的割り当て例

```
#include <stdio.h>
#define N 16384  // total memory 2^31B + 2^15 B, 約2GiB

dlm double a[N][N], x[N], y[N];  // DLM使用

int main(int argc, char *argv[])
{ int i,j;
  double temp;
  //  行列aを初期化
  for(i = 0; i < N; i++)
    for(j = 0; j <N; j++) a[i][j] = i;
  //   ベクトルxを初期化
  for(i = 0; i < N; i++)    x[i] = i;

  //  a[N][N]*x[N]=y[N] 計算
  for(i = 0; i < N; i++){
    temp = 0;
    for(j = 0; j <N; j++)  temp += a[i][j]*x[j];
    y[i] = temp;
  }
  return 0;
}
```

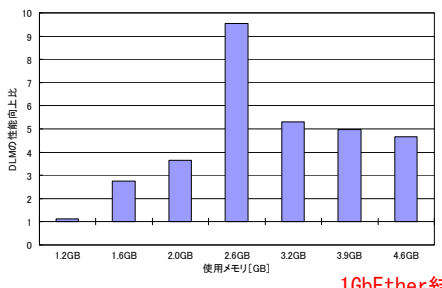
DLMプログラム例2（一次元配列書き込み test0.c ）
DLM動的割り当て例

```
#include <stdio.h>
#include <stdlib.h>
#define N ((long int)900000000)   // 3.6G = 900M * sizeof(int)=4

int main(int argc,char *argv[])
{
  double time;
  int *array;
  int temp;
  unsigned long int i;

  array=(int *)dlm_alloc(sizeof(int)*N);

  for(i=0;i<N;i++)     array[i]=i;  // sequential access
  for(i=0;i<N;i+=1024)     array[i]=-1;  // page access
  return 0;
}
```

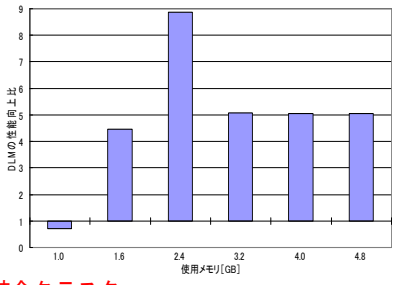


DLM Cコンパイラの構成

性能評価 DLMのSWAPFile使用時に対する速度向上比



行列ベクトル積 matv.c



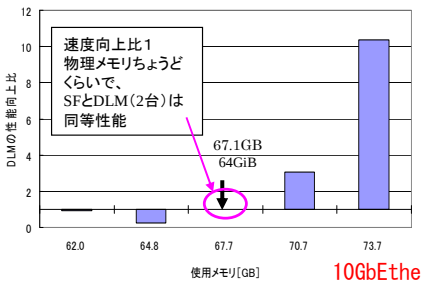
1次元配列アクセス test0.c

1GbEther結合クラスタ （swap領域 4GB）

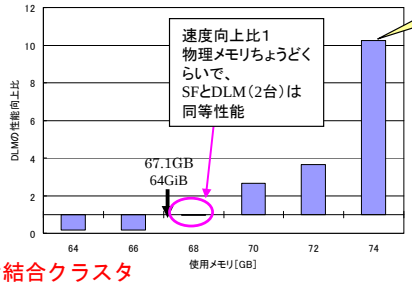
Cluster	HP ML150G2 x 8 Nodes
Node CPU	Xeon 2.8GHz x 2CPU HyperThread
Node Memory	1GByte (1GiB) (L2cache 1MB/CPU)
OS	Linux kernel2.6.20-1.2320.fc5 x86_64
Compiler	gcc version 4.1.1 20070105
Network	1GbE
Switch	CentreCom GS924GT(1GbE Switch)
Hard Disk	ST3808110AS 80GB S-ATA2 3Gbps 7200rpm,8MB cache, seek time 11ms(ave)

1GbEther結合クラスタ

物理メモリに対するswap使用比が15%程度で，性能は10倍を超える



行列ベクトル積 matv.c



1次元配列アクセス test0.c

10GbEther結合クラスタ

10GbEther結合クラスタ (CSLM) (swap領域10GB)

Cluster	HP DL585 G2 x 5 Nodes
Node CPU	Opteron 2.8GHz x 4 (8Cores)
Node Memory	64GByte (64GiB)
OS	Linux kernel 2.6.9-42 x86_64
Compiler	gcc version 3.4.6
Network	10GbE protocol (Myri-10G)
Hard Disk	SAS 147GB 10krpm 2台RAID1 Smart array 5i HP 431958-B21 (SAS 147GB, 10krpm, TransRate 300MBps, seektime 4(Ave),8.1(Max)ms)

1GbitEthernetクラスタ： 遠隔メモリ/搭載物理メモリのサイズ比が200%で、swap使用時の約5倍の性能が得られた
10GbitEthernetクラスタ： 遠隔メモリ/搭載物理メモリのサイズ比が15%で、swap使用時の10倍以上の性能が得られた