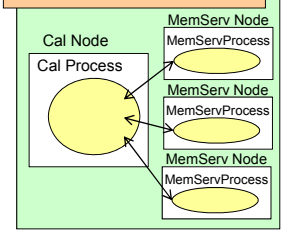
遠隔メモリアクセスのための スワップページサイズ自動調整機構の初期評価

内山 丞，緑川 博子（成蹊大）

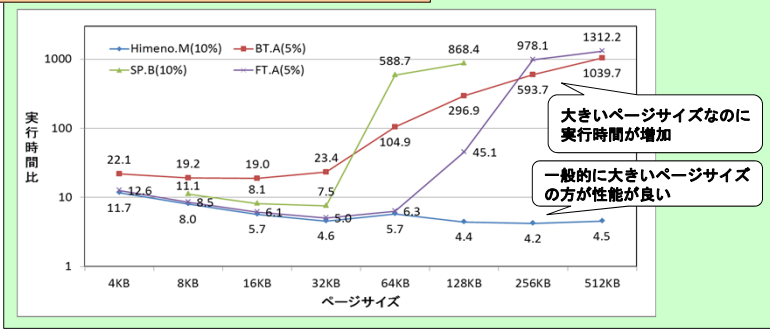
分散大容量メモリシステム (Distributed Large Memory System)

ネットワークで結ばれたコンピュータの物理メモリを通信によって利用し，
逐次処理用に**仮想的に大容量のメモリ空間**を提供するシステム．

DLMシステム構成イメージ



固定ページサイズが応用の実行時間に及ぼす影響



スワップ処理

メモリサーバ上(MemServ Node)に割り付けたデータへアクセスする場合
→必要なデータを持つメモリサーバ(MemServNode)からそのデータを含むDLMページを計算ノード(Cal Node)へ持ってきて(スワップイン)，
不必要と思われるDLMページをメモリサーバ(MemServNode)へ送る(スワップアウト)．

スワップページサイズ自動調整機構

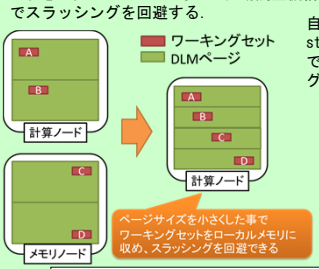
ローカルメモリサイズと応用の計算コアのワーキングセットに応じて、適切なDLMページサイズにプログラム実行時に動的に自動調整を行う．
これにより，不適切なページサイズによる過度なページスワッピング（スラッシング）を回避し、
DLMシステムにおける応用プログラム実行性能を適切に維持する

定義

ワーキングセット（WS）：ある処理の中で使用されるデータの集合
ローカルメモリ率：ローカルノードにあるメモリ量／応用プログラムが使用する全メモリ量
DLMページサイズ：DLMシステムが使用する独自の転送単位（OSページサイズの整数倍）

原因と対処

実行時間が爆発的に増加する場合には，ループ文の中で必要なデータをローカルノードが保持しきれないために、同一ページが繰り返しスワップされてしまう，**スラッシング**と呼ばれる状態が発生している．そこで自動調整機構がページサイズを適切な大きさにしてやることでスラッシングを回避する．



自動調整機構ではID（任意の整数）を引数とするstartとendの2つの関数を用意している．本機構では各IDの両関数に挟まれた区間毎に，ワーキングセットを見積もり，ページサイズの変更を行う．

WSの見積り手法：ループ文内でスワップインされたページの枚数をおよそのWSとする

ページサイズの変更：
ページサイズを小さくする場合には下記の式を用いて目標ページサイズに最も近いサイズまで一度に変更する．大きくする場合には目標ページの方向に変更するが、一度に二倍までに制限をかけている。

$$PS_{next} = LM / WS$$

PSnext：目標ページサイズ　LM：ローカルメモリサイズ　WS：およそのワーキングセット

使用方法

ユーザはループ文の前後に「start(int id);」「end(int id);」を挿入するだけで自動調整機構を利用できる．
本機構では引数のidに応じて、個別に情報の記録とDLMページサイズの動的変更を行う．

使用例

```
start(0);
for (i=0;i<x;i++) {
      for (j=0;j<y;j++) {
            for (k=0;k<z;k++) {
//何らかの処理
            }
      }
}
end(0);
start(1);
for (i=0;i<x;i++) {
      for (j=0;j<y;j++) {
            for (k=0;k<z;k++) {
//何らかの処理
```

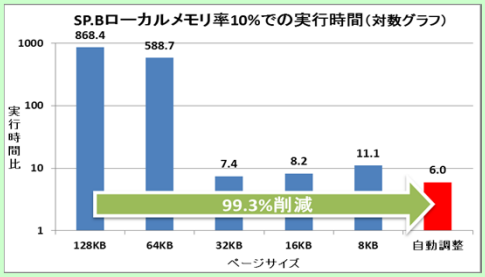
複数個所で計測する場合には引数として計測箇所のIDを渡す

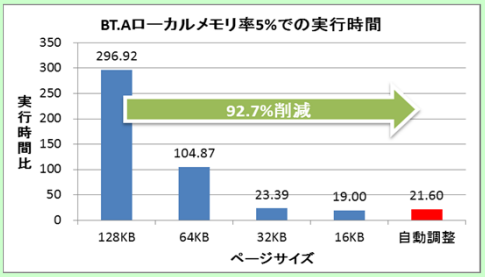
自動調整機構はstart関数とend関数に挟まれた部分をIDにより区別し、個別にページサイズの変更を行う

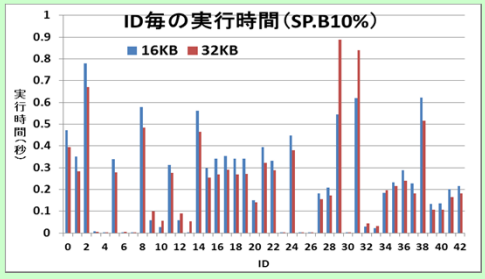
従来のページ管理表		自動調整機構独自のページ管理表	
各種情報	スワップインカウント	id	PSnext
--------	12364	0	16384
--------	25	1	65536

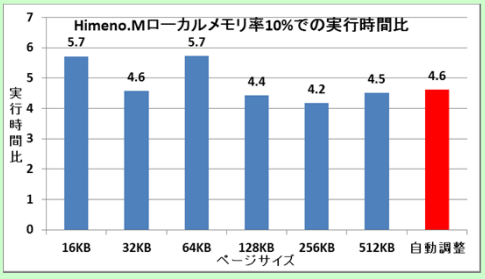
評価実験

以下のグラフはNPBのSPクラスB，BTクラスBにおいてイテレーションを10回(本来は200回)に変更し，
各ページサイズ固定の場合と128KB～16KBの間で自動調整を行った場合の実行時間のグラフである．









実験環境：東京大学情報基盤センター，T2K(ha8000)
Network：40Gbps，20Gbps，Memory：20GB/node x 2nodes