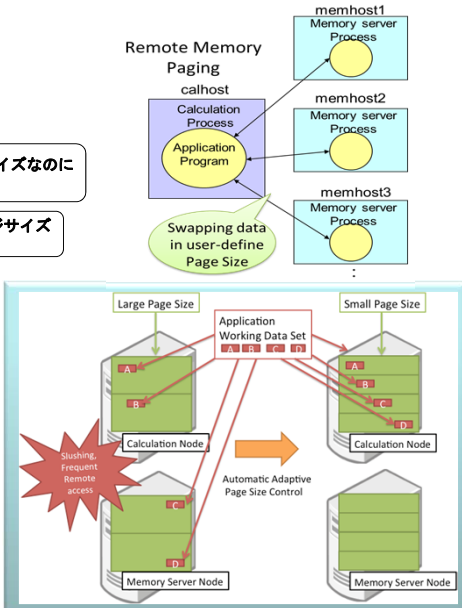
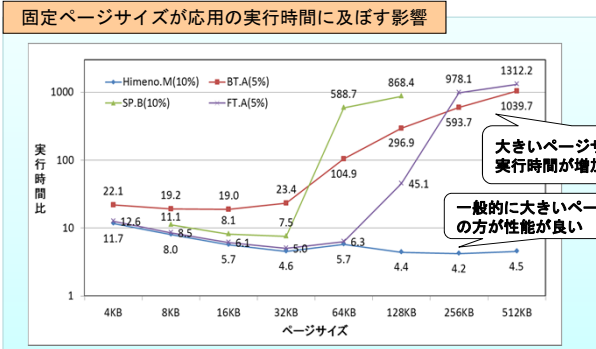


リモートページングのためのページサイズ自動調整機構 — ループ文におけるワーキングセット推定と選択的制御の導入 —

古尾谷 歩, 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。



スワップページサイズ自動調整機構の基本的なアイデア
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、ローカルメモリに収まるような適切なページサイズ(DLMページサイズ)にすることにより、スラッシングを防止する。

定義
ワーキングセット (WS) : ある処理区間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)

ページサイズ自動調整機構

ワーキングセットの推定
自動調整機構では、プログラム中のすべてのループ文の前後にスワップインしたページ枚数 (**WSページカウント**) と実行時間を記録するための2つの関数を挿入して、両関数に挟まれた区間毎に、ワーキングセットを見積もる。この値とローカルメモリサイズにある現在のページサイズによるページ数 (**ローカルページカウント**) を比較して、ページサイズの変更を行う。

選択的制御
応用実行中、最大実行時間をもつループ区間の値を常に更新し、前回の実行時間が、この値の1%以下の実行時間であるループ区間は、ページ変更や時間計測を省略する。

計測の効率化
前回の計測区間と同じ区間を計測する場合で、かつその区間のページサイズの変更がない場合は、時間計測を省略する。

ページサイズの変更の基本方針 :

- WSページカウント > ローカルページカウント**
スラッシングが起きていると推定し、以下のターゲットサイズにできるだけ近いページサイズに、次回の実行で変更する。

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

PS_{next}: 目標ページサイズ LM: ローカルメモリサイズ WS: ワーキングセットページカウント

- WSページカウント ≤ ローカルページカウント**
スラッシングは起きていないと推定し、大きなサイズによる効率的な通信のために、ページサイズを2倍にする。

計測関数の自動挿入と入れ子への対処
ユーザによる手動で計測関数を挿入することなく、プログラム中の全ループの前後に計測関数を挿入しても、選択的制御と計測効率化により、重要なループだけが、ページサイズ制御に生かされる。入れ子ループにも対応している。

swapi_countstart(int id);
前回の該当IDの実行時間により計測するか否かを判断する。前回推奨されたページサイズに変更する。計測する場合にはWSページカウントと実行時間計測を開始する

swapi_refresh(int id);
WSページカウントにより次回の該当IDのページサイズと実行時間を記録する。外側のループIDの指定ページサイズに戻す

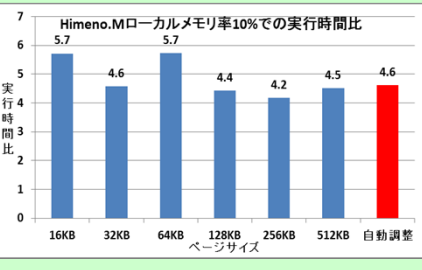
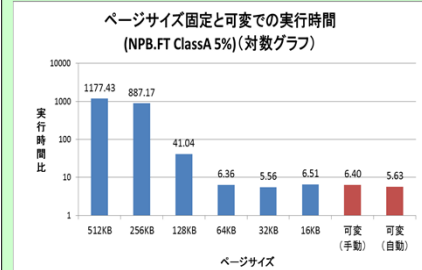
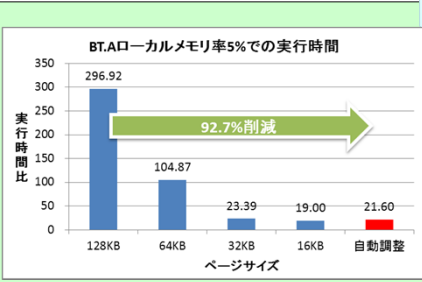
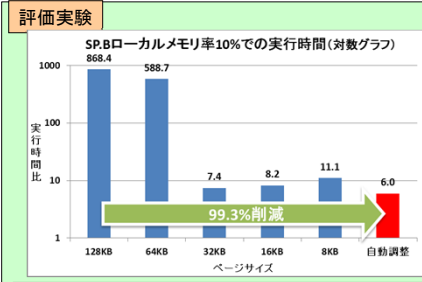
使用例

```

swapi_countstart(0);
for (i=0; i<M; i++) {
  swapi_countstart(1);
  for (j=0; j<N; j++) {
    some statements
  }
  swapi_refresh(1);
}
swapi_countstart(0);
    
```

複数箇所計測する場合には
引数として計測箇所のIDを渡す

自動調整機構は swapi_countstart(ID) と swapi_refresh(ID) 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



NPBのSPクラスB, BTクラスA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。FT.Aは規定通り6回である。いずれも、ローカルメモリ100%で逐次実行した場合との実行比

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