GPUによる汎用計算を目指して

GPUのための並列計算モデル

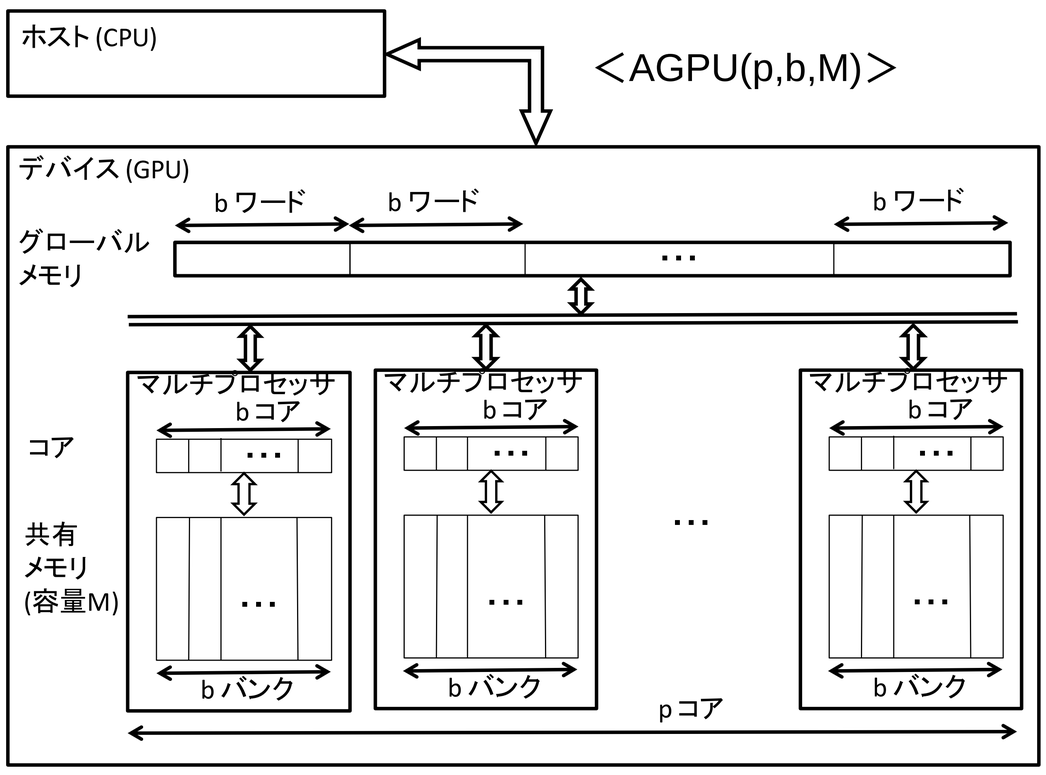
どんな研究？

GPUは元々はグラフィックス処理専用のプロセッサであったが，効率の良い並列処理を行えることからグラフィックス処理以外の様々な処理にGPUが使用され始めている．しかし，これまではGPU向けアルゴリズムに対して，簡便にアルゴリズムの良さを評価することができなかった．そこで，GPU向けアルゴリズムを汎用的かつ簡便に評価できるような並列計算モデルを提案する．

何がわかる？

提案するAGPUモデルは，GPU向けアルゴリズムを評価するための並列計算モデルであり，GPUアーキテクチャの特徴を良く捉えたモデルとなっている．本モデルを用いることで，GPU向けアルゴリズムに対し，実際にプログラムを実行することなく容易に漸近解析（様々な条件での大まかな挙動の解析）が行える．これにより，GPU向けアルゴリズムを容易に設計・評価できるようになる．

提案する並列計算モデル



アーキテクチャ

–実際のGPUと同等のアーキテクチャ
–3つのパラメータ p, b, M を利用して記述される
- p: コア総数
- b: マルチプロセッサごとのコア数
- M: マルチプロセッサごとの共有メモリ容量

–グローバルメモリのブロックサイズを b とする
- 1回のグローバルメモリアクセスで最大 b ワードを取得できる

–共有メモリのバンク数を b とする
- 1回の共有メモリアクセスで最大 b ワードを取得できる

評価基準（計算時間）

- 時間計算量：マルチプロセッサあたりの実行命令数
- I/O計算量：グローバルメモリアクセス回数 (全マルチプロセッサの合計)

評価基準（メモリ使用量）

- マルチプロセッサあたりの共有メモリ使用量（ワード数）
- グローバルメモリ使用量（ワード数）

GPU向けアルゴリズムの設計・評価例

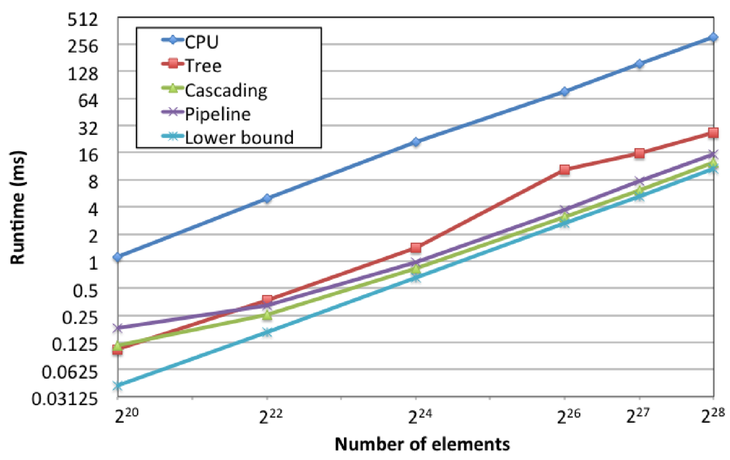
リダクション（総和計算）アルゴリズムの設計・評価

＜データ数n >> p の時の計算時間＞

	従来手法		提案手法
アルゴリズム（非可換演算可否）	Tree-based（対応）	Cascading（非対応）	Pipeline（対応）
時間計算量	O(n log b /p)	O(n/p)	O(n/p)
I/O計算量	O(n/b)	O(n/b)	O(n/b)

最適でない　最適

＜tesla C1060 による処理時間＞



AGPUモデルによる計算時間評価は
実際の処理時間の良い指標となる

マルチスレッディングの効果見積もり

- GPU向けアルゴリズムではマルチスレッディング（1つのコアが複数の処理を時分割で切り替えながら計算すること）をうまく活用することが高速実行のキーポイントとなっている
  ⇒ マルチスレッディングを活用することにより，グローバルメモリアクセスの待ち時間（レイテンシ）を短縮（隠ぺい）できる
- AGPUモデルを用いて，マルチスレッディングの効果を見積もることができる
  - レイテンシ隠ぺいの効果 ⇒ “多重度”が大きいと効果大
  - マルチスレッディングを考慮した時間計算量（右列参照）

最速プレフィックスサムアルゴリズムの評価

入力 T [2 5 4 1]
↓ 加算によるプレフィックスサム
出力 U [0 2 7 11]

- 計算時間 (@AGPU(v,b,M))
  –時間計算量： $\mathrm{O}\left(\left(\frac{n}{v}+\frac{v}{b^2}\right)\left(\min\left(\left\lceil\frac{b}{\alpha}\right\rceil,\alpha\right)+\frac{\log b}{\alpha}\right)\right)$
  ⇒αの増加と共に小さくなり，α=Ω(b)では $\mathrm{O}\left(\frac{n}{v}+\frac{v}{b^2}\right)$
  –I/O計算量： $\frac{3n}{b}+\mathrm{O}\left(\frac{p}{b}\right)$
- メモリ使用量 (@AGPU(v,b,M))
  –共有メモリ使用量： $\mathrm{O}(\alpha b)$ ワード
  –グローバルメモリ使用量： $2n+\frac{v}{b}$ ワード
- 多重度 (@AGPU(p,b,M))
  – $\mathbf{M} := \min\left\{\frac{v}{p}, \mathrm{O}\left(\frac{M}{\alpha b}\right)\right\}$ vを十分に大きくすれば $\mathbf{M} := \mathrm{O}\left(\frac{M}{\alpha b}\right)$
- 考察
  –αはチューニングにより決められる値であり，時間計算量と多重度のトレードオフになっている

マルチスレッディングを考慮した時間計算量

AGPU(v,b,M)上のアルゴリズム
- 1コアが1スレッドを処理する
- 時間計算量： $S(v)$

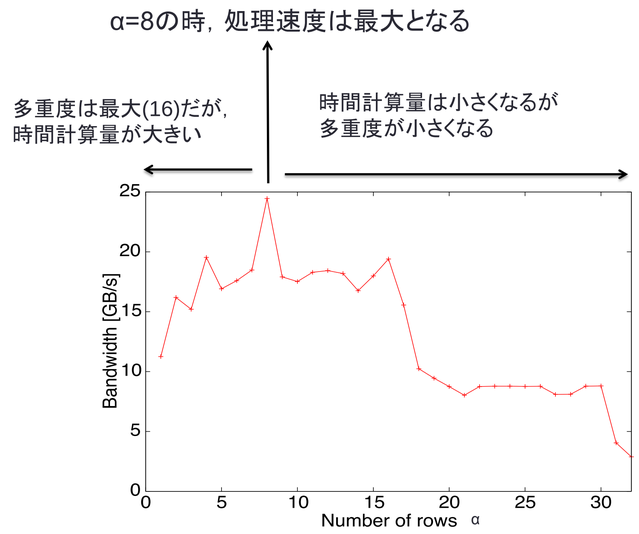
AGPU(p,b,M)上で実行
- 1コアが v/p スレッドを処理する
- 時間計算量： $T(p) \le \frac{v}{p} S(v)$

時間計算量はv/p(コア数の比)に比例する

※ I/O計算量はコア数を変えても変化しない

＜tesla C1060 によるプレフィックスサムの処理時間＞

α=8の時，処理速度は最大となる



連絡先：定兼　邦彦／ 国立情報学研究所 情報学プリンシプル研究系　准教授 sada@nii.ac.jp