3 GPUコンピューティングへの取り組み

- 日立のGPGPUへの取り組み
- HPCシステムとアプリケーションの性能

3-1 日立のGPGPUへの取組み(1)

計算科学を用いた研究開発分野でGPU利用が拡大中
研究所を中心に技術交流会を定期的に開催

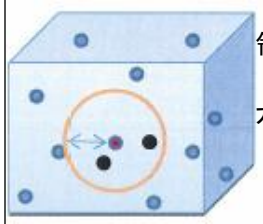
<利用分野(検討中含む)>

- 原子炉炉心解析
- 火力・原子力発電の蒸気タービン流れ解析
- ボイラ燃焼効率解析
- 粒子線治療シミュレーション
- 材料物性・ナノシミュレーション
- 機械(熱流体，構造，振動)
- 電磁場
- ライフサイエンス
- 金融(実効金利計算)
- 他

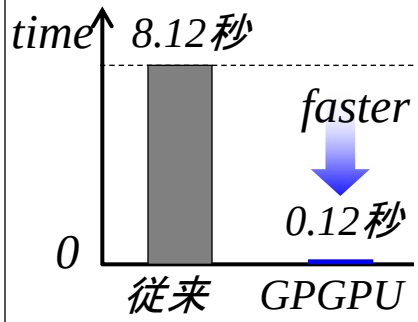
日立のGPGPUへの取組み(2)

■GPGPU技術に関し、学術系～産業系アプリの先行評価・提案中

(1)分子動力学

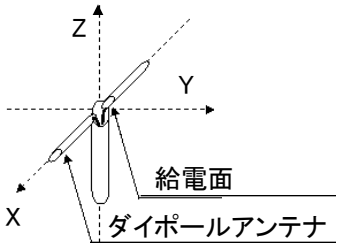


近傍の分子リストの作成処理

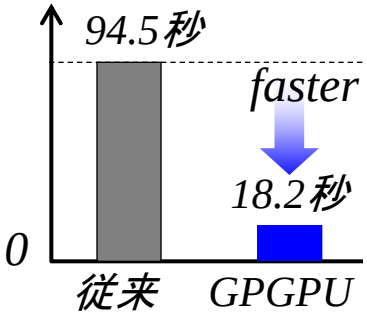


加速率＝約80倍

(2)電磁場解析

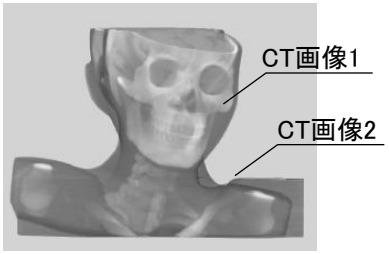


マクスウェル方程式による3次元解析

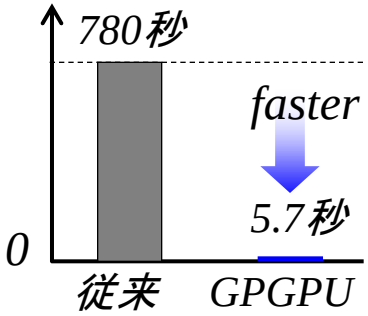


約5.6倍

(3)医用画像処理

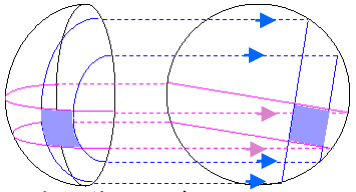


複数医用画像の重ね合せ位置決定

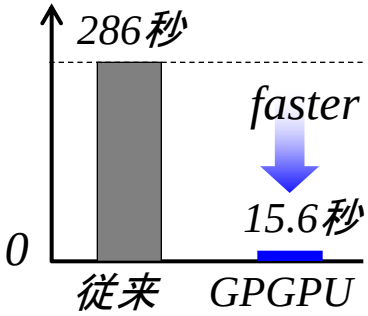


約140倍

(4)監視画像処理



魚眼レンズ画像を平面図に変換

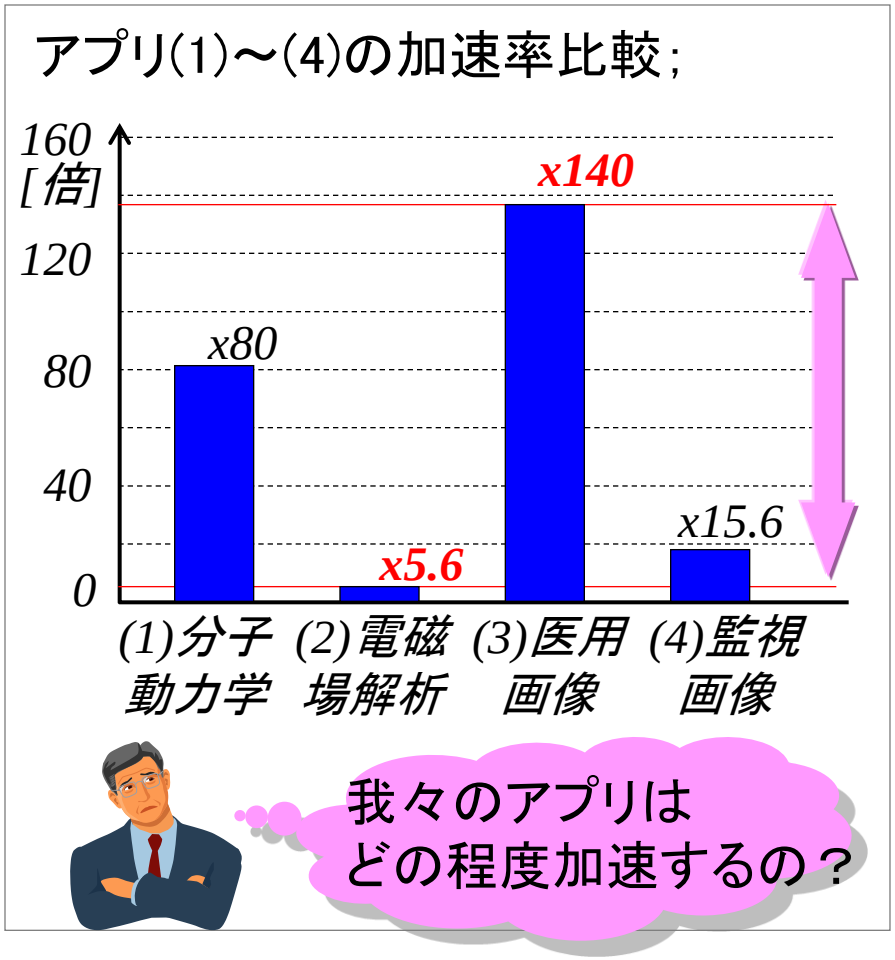


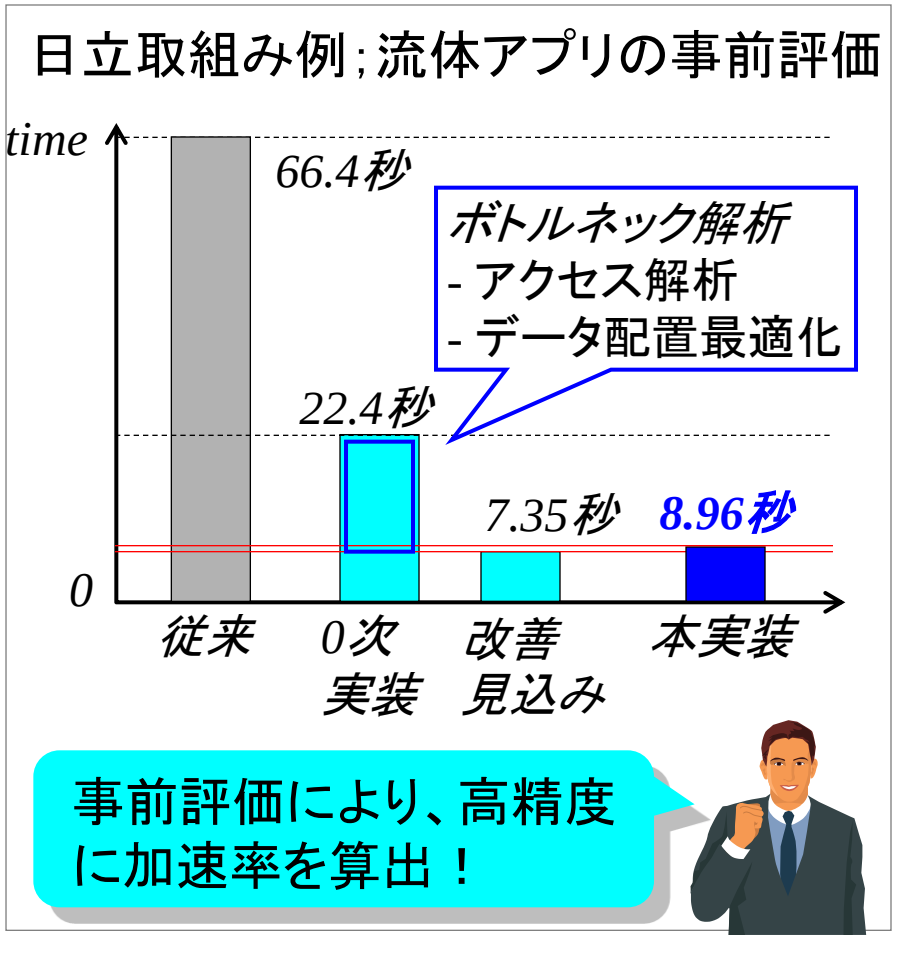
約18倍

学術系 ⟷ 産業系

3-3 GPGPUの性質と日立の取り組み

- GPGPU性質：アプリによって加速率に大差。投資判断が難しい。
- 日立取組：業務アプリを解析し、投資前に加速率を評価可能に。





◆社内にはGPUユーザ多数
◆利用技術・最適化技術も蓄積中
◆ソリューションメニューも整備
（事前評価からサポート）

◆GPU対応製品（PCIe x8,x16 搭載）HA8000 他
販売中

◆GPU搭載した大規模クラスタ(HPCシステム)
検討中

3 GPUコンピューティングへの取り組み

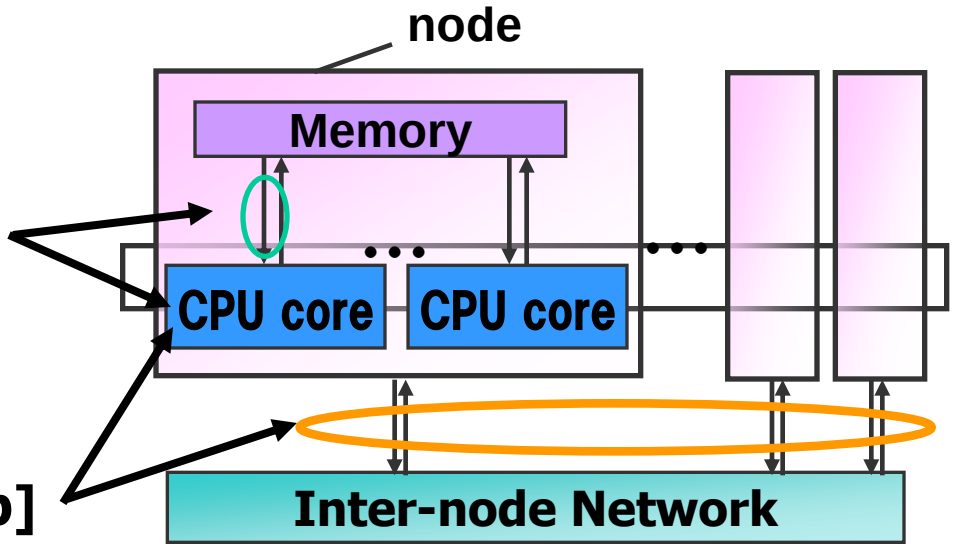
- 日立のGPGPUへの取り組み
- HPCシステムとアプリケーションの性能

3-5 システムバランスとアプリケーションの性能

アプリケーションの実効性能(効率)を以下の2点から定量評価

(1) ピーク演算性能に対する
メモリバンド幅 [Byte/flop]

(2) ピーク演算性能に対する
ネットワークバンド幅 [Byte/flop]



Example of high performance sever

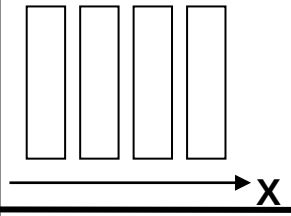
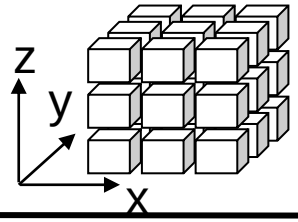
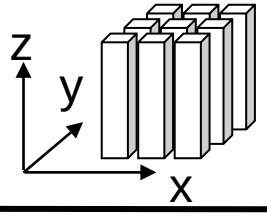
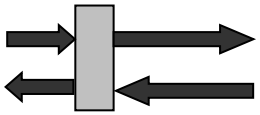
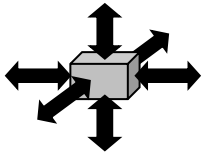
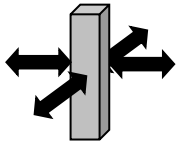
(1),(2)の数値を変化させて実効性能への影響を見る (シミュレーション)
⇒ アプリケーションが求めるシステムバランスを求める

評価アプリケーション

4種類の並列アプリについて評価を実施

並列化スキームとプロセス間通信パターンは以下の通り

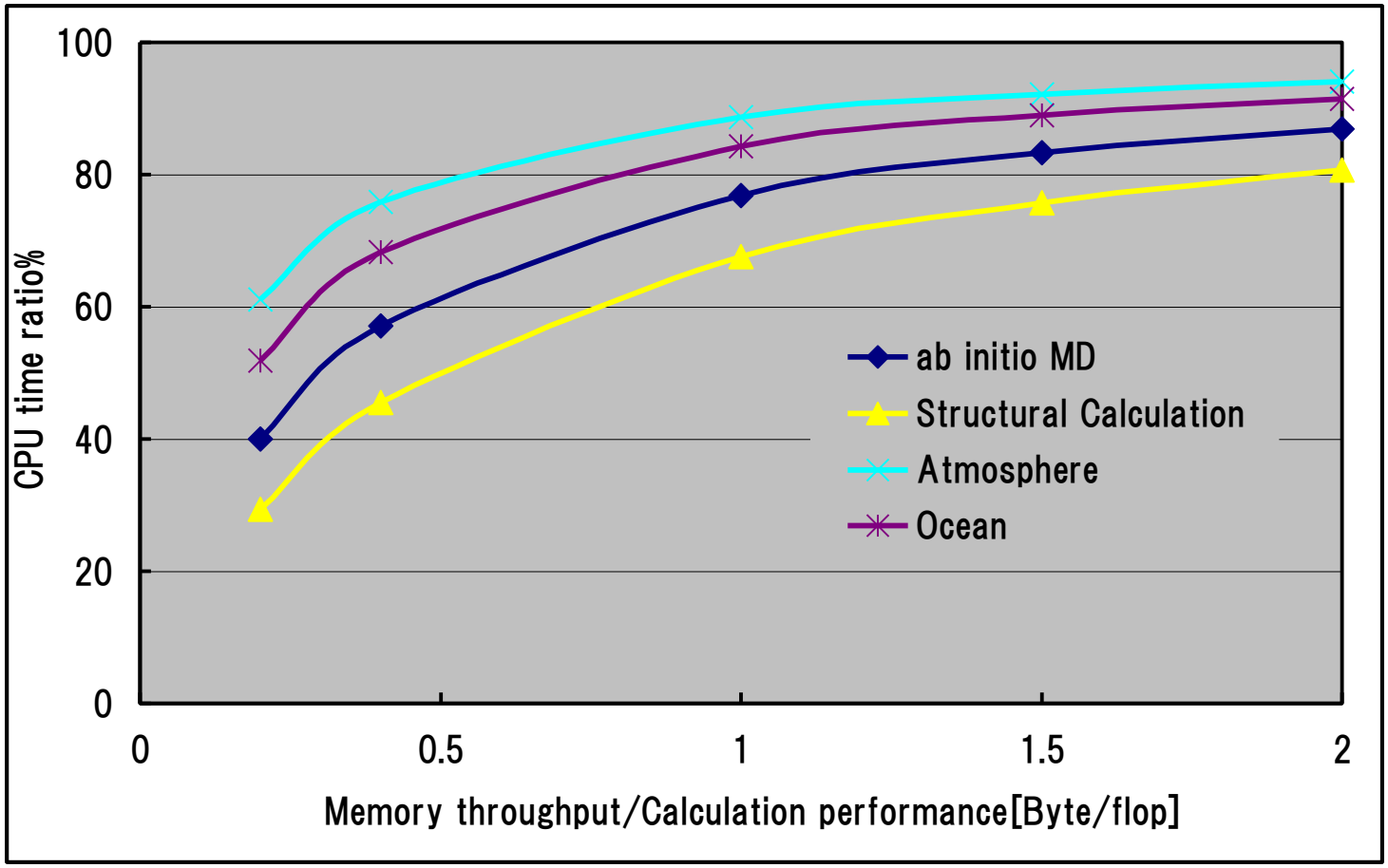
No.	Application	Calculation method	Partition	type
1	Ab initio MD	FFT, DGEM	Band Energy	1
2	Structural Calculation	Finite Element Method	3-Dim. space	2
3	Atmosphere	Difference Method	2-Dim.	3
4	Ocean	Difference Method	2-Dim.	3

Type	1	2	3
Partition			
Communication Pattern			
MPI function	MPI_allreduce (MPI_sum)	MP_send, MPI_recv MPI_allreduce(MPI_sum)	MP_send, MPI_recv

3-7 メモリバンド幅と演算性能のバランス

(メモリバンド幅 GB/s) / (演算性能 GFlop/s) > 0.4[Byte/flop]

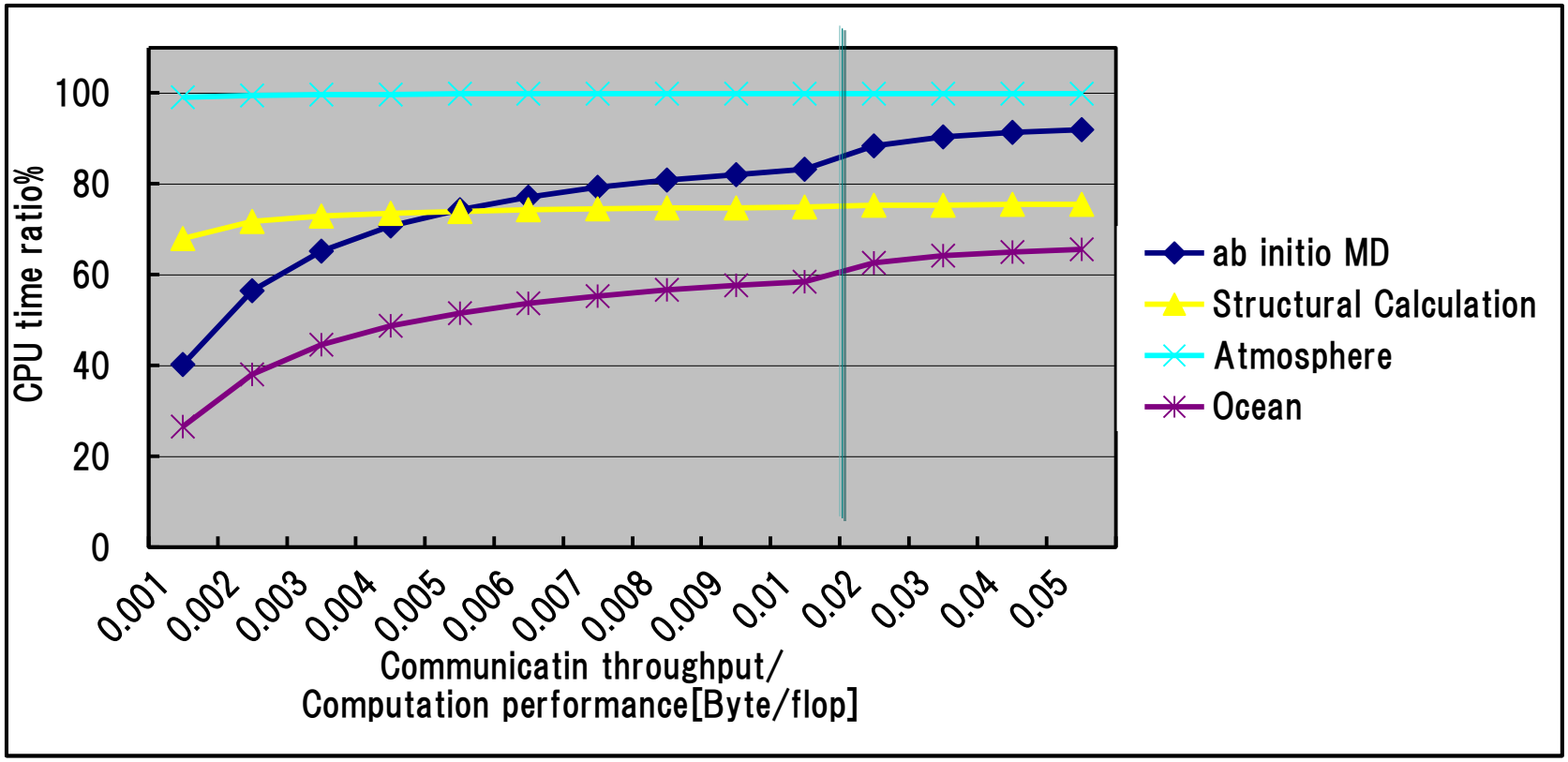
- **CPU time ratio becomes lower.** ⟹ **0.2 ~ 0.4[Byte/flop]**
- **Better to keep more than 1.0 [Byte/flop]**



3-8 ネットワーク性能と演算性能のバランス

(ネットワークバンド幅 GB/s) / (演算性能 GFlop/s) > 0.02 [Byte/flop]

- **The ratio of the communication time depends on the application.**
- **Better to keep more than 0.02 [Byte/flop]**



グラフは Memory throughput/Calculation performance[Byte/flop] = 0.4 の場合

◆アプリケーションの要請

(メモリバンド幅 GB/s) / (演算性能 GFlop/s) > 0.4[Byte/flop]

(ネットワークバンド幅 GB/s) / (演算性能 GFlop/s)

> 0.02 [Byte/flop]

◆マルチGPUシステムのバランス

(メモリバンド幅 GB/s) / (演算性能 GFlop/s) = 0.25[Byte/flop]

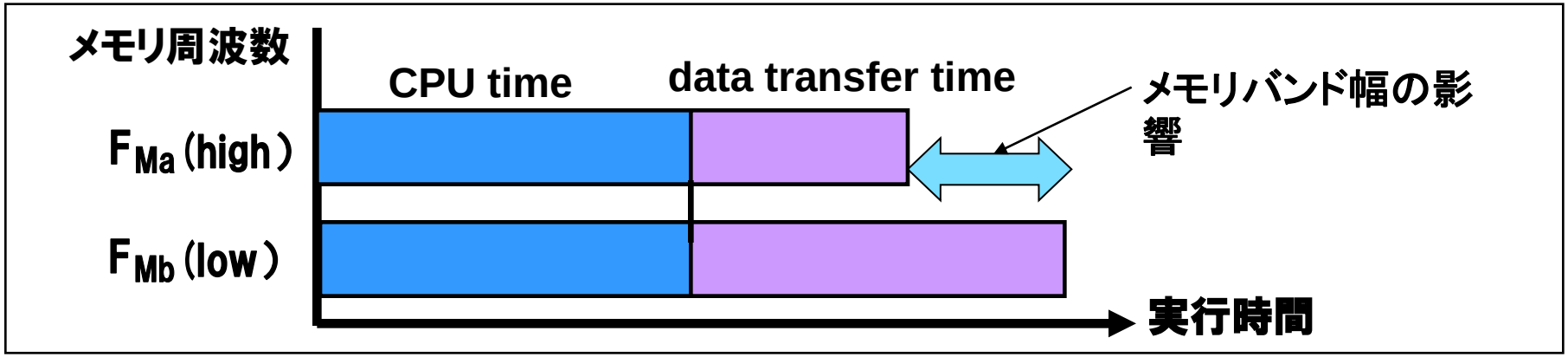
(ネットワークバンド幅 GB/s) / (演算性能 GFlop/s)

= 0.004 [Byte/flop]

◆実際にアプリケーションの性能はどうなるか？

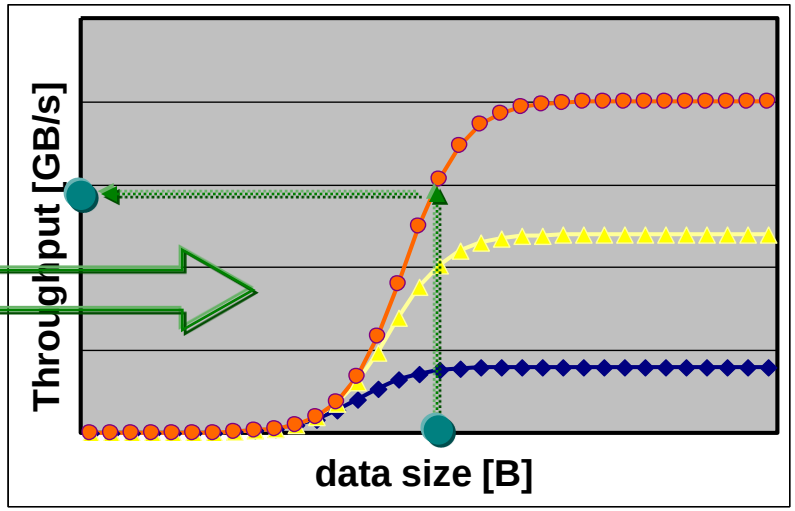
3-10 GPUシステムでのアプリの性能推定

演算時間：GPUシステムのB/Fより実効効率を計算
サーバのメモリ周波数を変化させて実行時間を測定
実行時間をCPU時間とデータ転送時間に分解



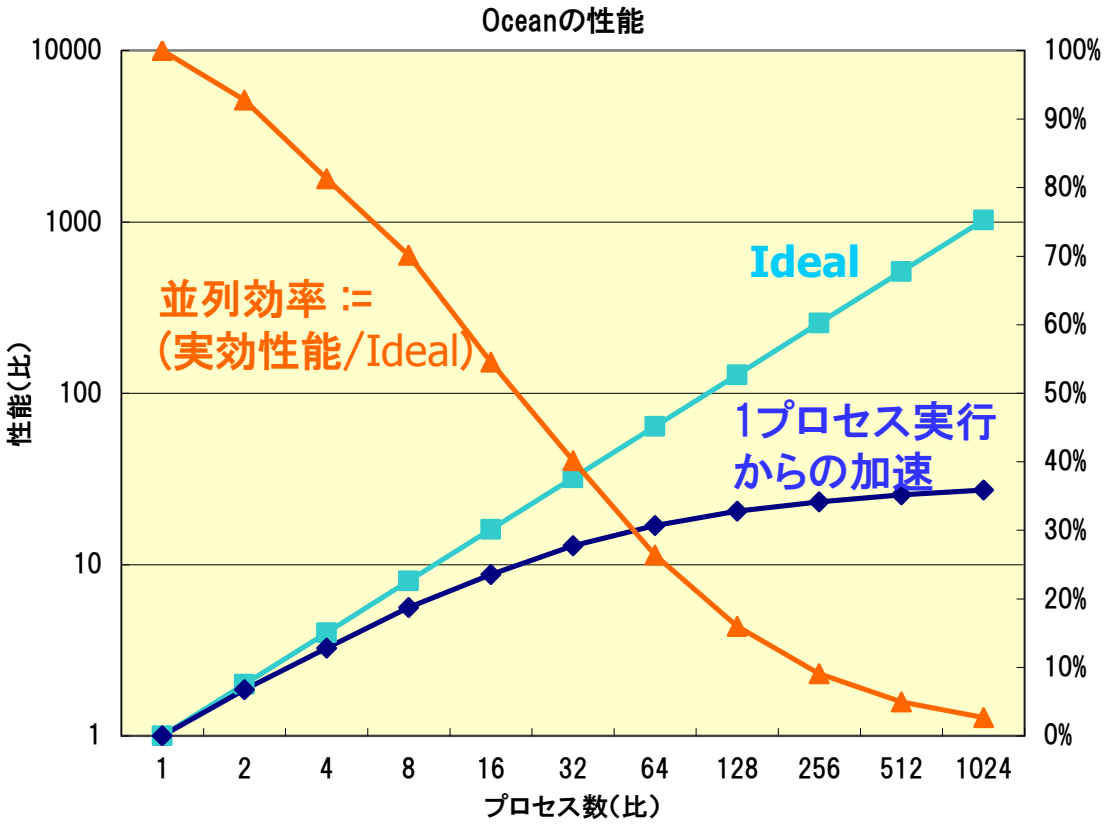
通信時間：PCクラスタで並列実行して
通信プロファイルを取得

個々の通信に対して
通信量から通信時間を
グラフから求める



Ocean の並列性能[推定]

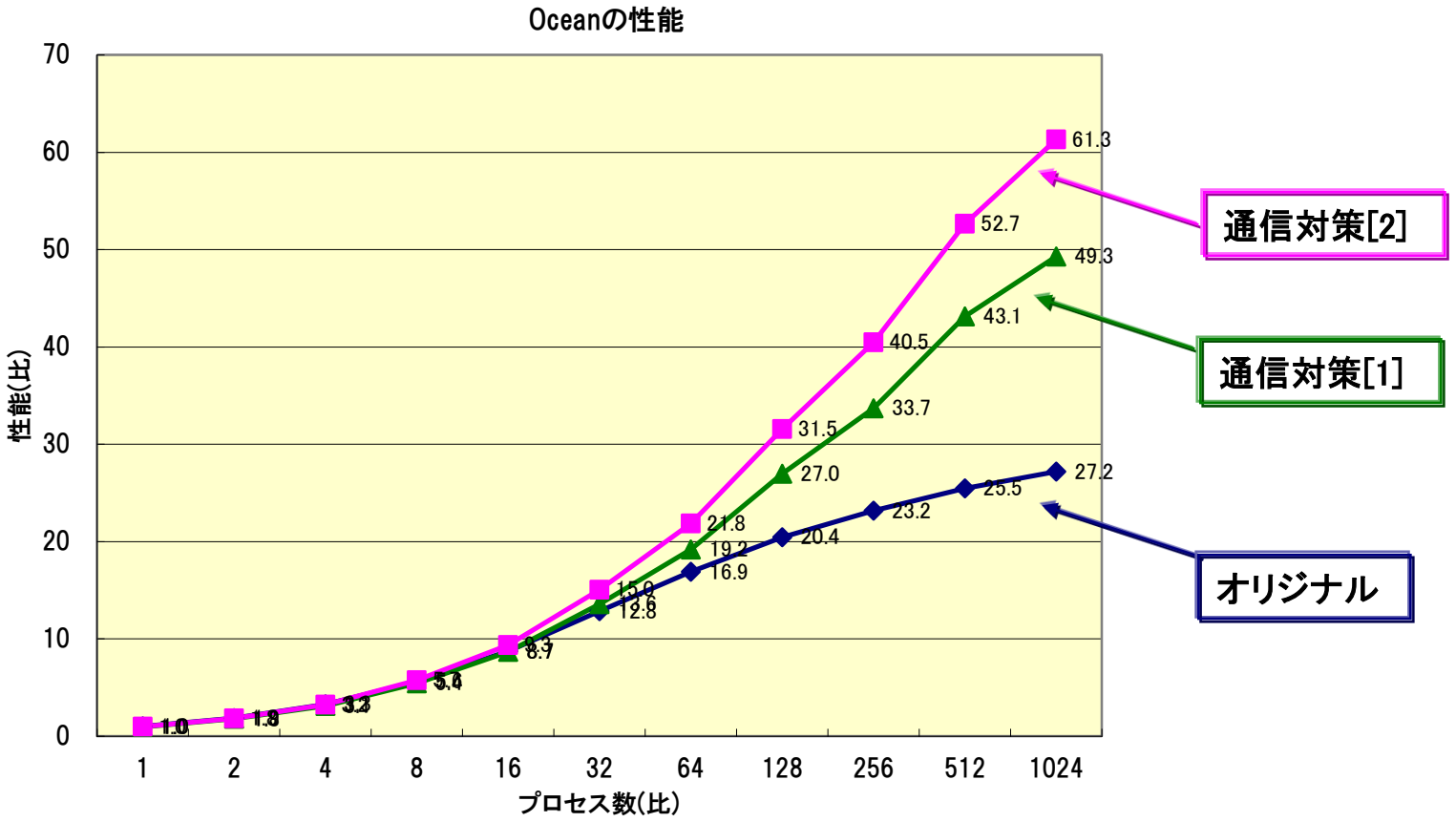
- 同じ規模の問題を x−方向、y−方向の順で分割を繰り返す ⇒ strong scaling
- 1プロセスのメモリ使用量がGPUに収まる最小の並列数を基準（グラフのプロセス数（比）=1）
  プロセス数（比）=1 のメモリ使用量 2.6GB ⇒ S2050 で利用可能な最大値
- プロセス数（比）=1のときの通信時間 ⇒ 全実行時間の9.4%
- 演算効率は B/Fから推測 ⇒ 3.3%



Ocean の並列性能[推定]の改善

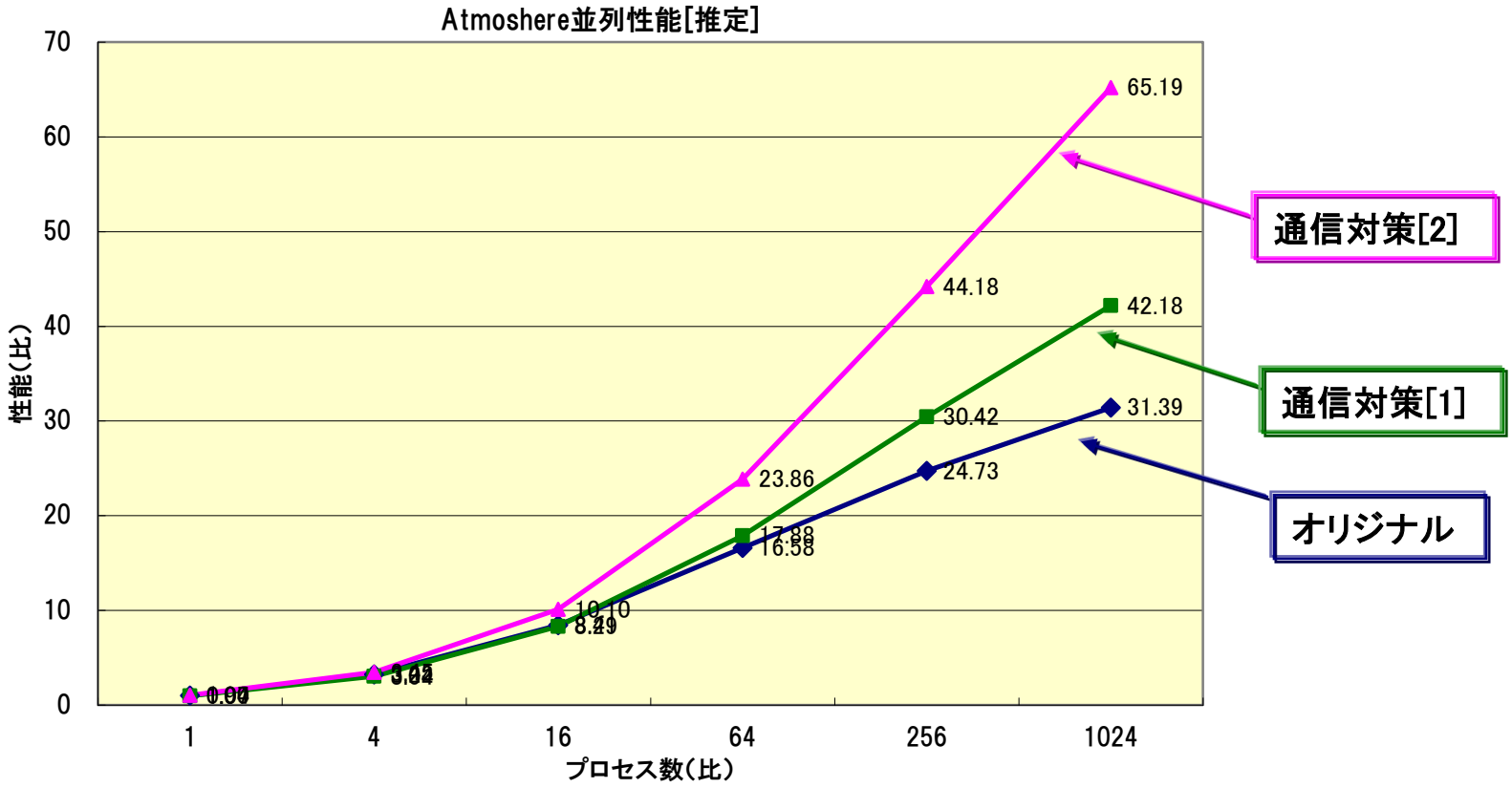
[1] 通信アルゴリズムによる対策
隣接通信する境界面を多層化して通信回数を削減

[2] 資源増強による対策
さらに、InfiniBannd を追加して2方向の隣接間通信を同時実行



3-13 Atmospher の並列性能[推定]

・プロセス数(比)=1 のメモリ使用量 1.8GB
・演算効率は B/Fから推測 ⇒ 3.8%、プロセス数(比)=1 のとき通信時間は16%



[1] 通信アルゴリズムによる対策
隣接通信する境界面を多層化して通信回数を削減

[2] 資源増強による対策
さらに、InfiniBannd を追加して2方向の隣接間通信を同時実行

3-14 まとめ：マルチGPUシステムとアプリの性能

◆マルチGPUシステムの特徴

(メモリバンド幅 GB/s)／(演算性能 GFlop/s) = 0.25[Byte/flop]

メモリ性能バランスはPCサーバよりやや低め。効率は良い(80%)

(ネットワークバンド幅 GB/s)／(演算性能 GFlop/s) = 0.004 [Byte/flop]

ネットワーク性能が相対的に低く見える。レイテンシ > 20μs

GPU Direct は*データ長 > 16KB* で効果大

◆アプリケーションの並列実行性能

・GPUのメモリを最大に使用した weak scaling ではネットワークの弱さは目立たない

・strong scaling でのスケーラビリティ劣化は早い

今回の評価では 16GPUで約10倍加速、以後急速に劣化

・strong scaling でのスケーラビリティを保つには努力が必要

演算に隠蔽できれば良い

転送データ長が大きい場合はパイプライン化

転送データ長が小さい場合は通信回数の削減

演算数が増えても通信回数削減を検討(shadow領域の多層化など)

