

PCクラスタワークショップ in 神戸

# 日立のテクニカルコンピューティングへの 取り組み

2011/2/18

株式会社 日立製作所  
中央研究所  
清水 正明

**HITACHI**  
Inspire the Next

Copyright © Hitachi,Ltd.2011 All rights reserved 1

---

## 目 次

---

**1** 日立テクニカルサーバラインナップ

---

**2** 日立サーバラインナップ

---

**3** GPUコンピューティングへの取り組み

---

**4** SC' 10 日立展示

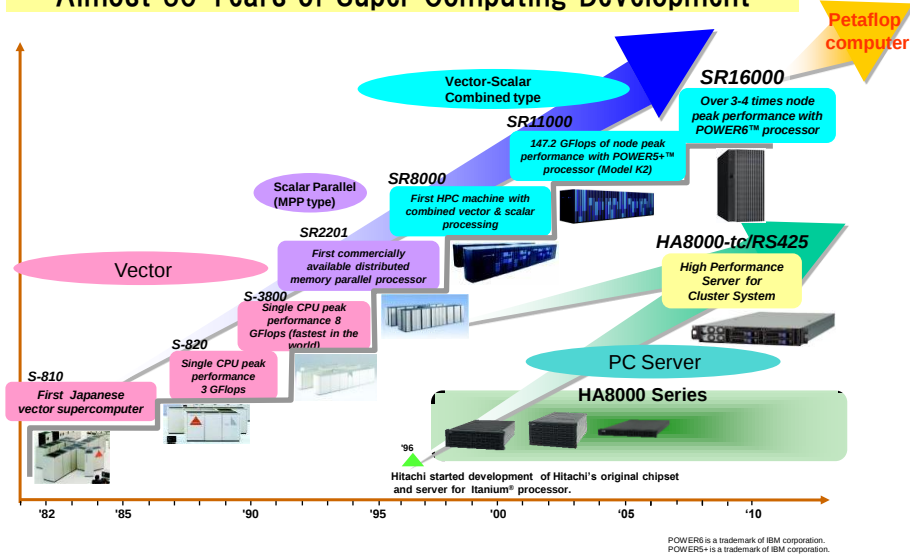
---

Copyright © Hitachi,Ltd.2011 All rights reserved 2

## 1-1

## 日立テクニカルサーバ：History & Future

### Almost 30 Years of Super Computing Development



Copyright © Hitachi, Ltd. 2011 All rights reserved 3

## 1-2

## 日立テクニカルサーバ ラインアップ

### PCクラスからスカラSMPまでラインアップ拡充

- 最新POWER7™プロセッサ32way搭載SR16000
- Xeon™ 2ソケットプロセッサ
- 今後の最新cpuにも対応予定



Copyright © Hitachi, Ltd. 2011 All rights reserved 4

電力性能比・価格性能比に優れた  
POWER7搭載のSR16000シリーズ次世代機 第一弾



日立開発・製造機 SR16000/XM1

**日立開発・製造:**

IBM社との戦略的アライアンスに基づく、日立開発・製造のPOWER7搭載サーバ。EP8000/750と共通プラットフォーム

**電力性能比の向上:**

SR16000モデルL1/L2と同等の32way SMP構成をPOWER7 4ソケットで実現。ノード消費電力は約1/3に大幅に削減

**抜群の価格性能比:**

POWER7の圧倒的な性能と、戦略的な価格付けにより、価格競争力を強化!

**中規模SMPノード・クラスタ:**

SRシリーズのコンセプトを受継ぐデザイン。中規模SMPノードのクラスタシステムによりスループット指向の中規模システムに最適

## 2 日立サーバラインアップ

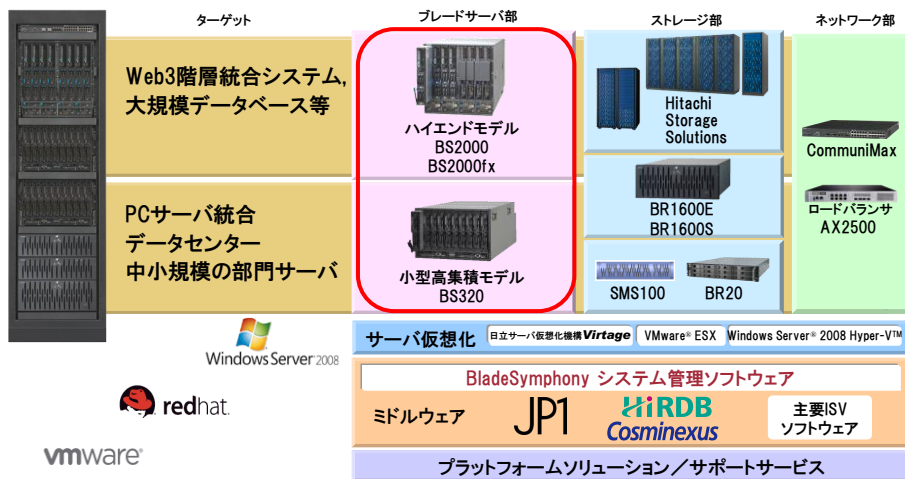
- ・ブレードサーバ
- ・ラックマウントサーバ

## 2-1

## BladeSymphony ラインアップ

ブレードサーバを核に、ストレージ、ネットワーク、管理ソフトウェアを一体化した  
統合サービスプラットフォーム BladeSymphony

- 各製品、充実のラインナップで、用途に応じた製品を提供
- 仮想化環境やソリューションを含めたシステム提供も可能



Copyright © Hitachi, Ltd. 2011 All rights reserved 7

## 2-2

## ハイエンドモデル BS2000/BS2000fx

高性能・高信頼志向のシステム向け

**BS2000**  
**BS2000fx** NEW

### ■ 仮想統合を実現する高信頼スケーラブル・ブレードサーバ

仮想化による集約、高速処理に適応した性能・拡張性

- サーバブレード間SMP接続(64cores MAX、メモリ 1TBMAX)
- I/Oスロット拡張装置(64スロット MAX)

日立サーバ仮想化機構 **Virtage** 標準搭載(\*1)

メインフレームの高信頼・高可用化技術を継承

業界最高レベルの高効率電源

- CSCI Gold基準適合, 80 PLUS® GOLD認証取得(\*2)

基幹システムの長期安定稼働を支援

- Eタイプ

ハードウェア長期保守対応

- ロングライフサポートサービス7年/10年(\*3)



最大8ブレード/10U

- \*1: Essentialモデル
- \*2: 電源負荷50%時の変換効率92%を実現
- \*3: BS2000 Eタイプにて提供

Copyright © Hitachi, Ltd. 2011 All rights reserved 8

## 2-3

## 小型高集積モデル BS320

より軽く、より小さく 高密度実装を追求 **BS320**

## ■ 幅広い用途に対応する高集積・省電力ブレードサーバ

- ・ 高さ6U(約27cm)に最大10ブレード搭載可能
- ・ 最大重量約98kg/シャーシの軽量設計
- ・ 用途に応じた多彩なサーバブレードをラインアップ
- ・ 日立サーバ仮想化機構 **Virtage** に対応<sup>(※1)</sup>
- ・ 高効率電源 (CSCI Silver基準適合, 80 PLUS® SILVER認証取得<sup>(※2)</sup>)
- ・ ハードウェア長期保守対応 (ロングライフサポートサービス: 7年)



標準サーバブレード SAN専用サーバブレード HDD拡張サーバブレード PCI拡張サーバブレード ロードレギュレーションサーバブレード



\*1: PCI拡張サーバブレード **Virtage** モデルで提供  
 \*2: 負荷50%時の変換効率89%以上を実現

Copyright © Hitachi, Ltd. 2011 All rights reserved 9

## 2-4

## HA8000ラインアップ

(4Processor)

HA8000/RS440



出荷開始時期: 2010/11/30

Xeon(X7560/X7550/E7540/E7520)

- RAID追加機能
- 2.5型HDD 容量アップ
- SSD
- ロングライフサポート
- ENERGY STARモデル

(2Processor)

HA8000/TS20



出荷開始時期: 2010/11/30

Xeon(X5670/E5640/E5620/E5503)

- RAID追加機能
- 2.5型HDD 15krpm追加
- 3.5型 2TB SATA HDD追加
- 動的パワーキャッピング
- SSD

HA8000/RS220



出荷開始時期: 2010/11/30

Xeon(X5680/X5670/E5640/E5560/L5630/E5503)

- CPU追加
- RAID追加機能
- 2.5型HDD 15krpm追加
- 3.5型 2TB SATA HDD追加
- 動的パワーキャッピング
- SSD
- AC200V
- 低電圧メモリ

HA8000/RS210



出荷開始時期: 2010/11/30

Xeon(X5670/E5640/E5620/L5630/E5503)

- CPU追加
- RAID追加機能
- 2.5型HDD 15krpm追加
- 3.5型 2TB SATA HDD追加
- 動的パワーキャッピング
- SSD
- AC200V
- 低電圧メモリ

(1Processor)

HA8000/SS10



出荷開始時期: 2010/11/30

Core i3-540/Pentium G6950

- RAID追加機能
- 3.5型 2TB SATA HDD追加

HA8000/RS110



出荷開始時期: 2010/11/30

Xeon(X3480/X3470/X3460/X3430)

- CPU追加
- RAID追加機能
- 3.5型 2TB SATA HDD追加

HA8000/TS10



出荷開始時期: 2010/11/30

Xeon(X3480/X3470/X3460/X3430)

- CPU追加
- RAID追加機能
- 3.5型 2TB SATA HDD追加

Copyright © Hitachi, Ltd. 2011 All rights reserved 10

## 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

Copyright © Hitachi, Ltd. 2011 All rights reserved <sup>11</sup>11

### 3-1

#### 日立のGPGPUへの取り組み(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

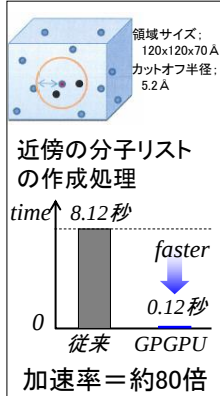
Copyright © Hitachi, Ltd. 2011 All rights reserved 12

## 3-2

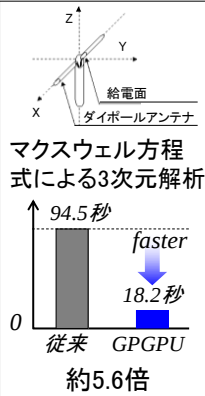
## 日立のGPGPUへの取組み(2)

### ■GPGPU技術に関し、学術系～産業系アプリの先行評価・提案中

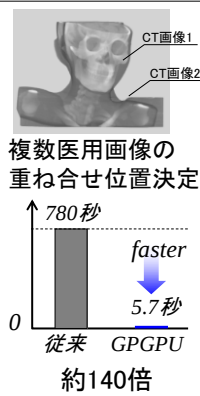
#### (1)分子動力学



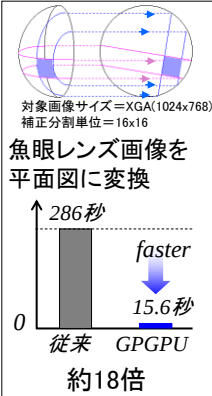
#### (2)電磁場解析



#### (3)医用画像処理



#### (4)監視画像処理



←学術系

産業系→

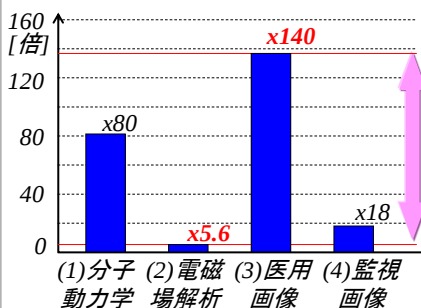
Copyright © Hitachi, Ltd. 2011 All rights reserved 13

## 3-3

## GPGPUの性質と日立の取組み

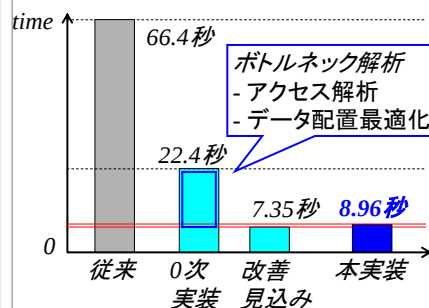
- GPGPU性質:アプリによって加速率に大差。投資判断が難しい。
- 日立取組:業務アプリを解析し、投資前に加速率を評価可能に。

アプリ(1)～(4)の加速率比較;



我々のアプリは  
どの程度加速するの？

日立取組み例:流体アプリの事前評価



事前評価により、高精度  
に加速率を算出！



Copyright © Hitachi, Ltd. 2011 All rights reserved 14

- ◆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

Copyright © Hitachi,Ltd.2011 All rights reserved 15

## 3 GPUコンピューティングへの取り組み

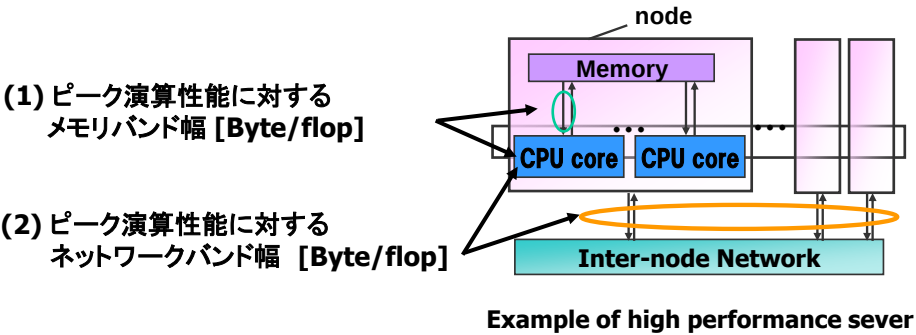
- ・日立のGPGPUへの取り組み
- ・HPCシステムとアプリケーションの性能

Copyright © Hitachi,Ltd.2011 All rights reserved <sup>16</sup>16

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),(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

3-6

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

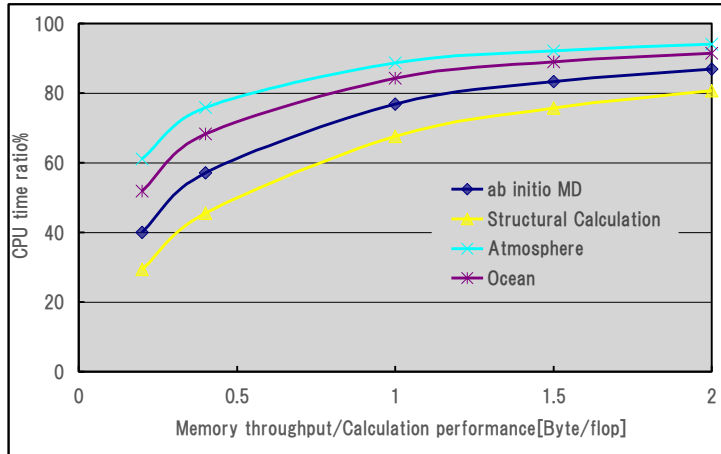
| No. | Application            | Calculation method    | Partition    | type |
|-----|------------------------|-----------------------|--------------|------|
| 1   | Ab initio MD           | FFT, DGEM             | Band Energy  | 1    |
| 2   | Structural Calculation | Finite Element Method | 3-Dim. 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_rcv<br>MPI_allreduce(MPI_sum) | MP_send, MPI_rcv |

### 3-7 メモリバンド幅と演算性能のバランス

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

- CPU time ratio becomes lower.  $\rightarrow 0.2 \sim 0.4 [\text{Byte/flop}]$
- Better to keep more than 1.0 [Byte/flop]

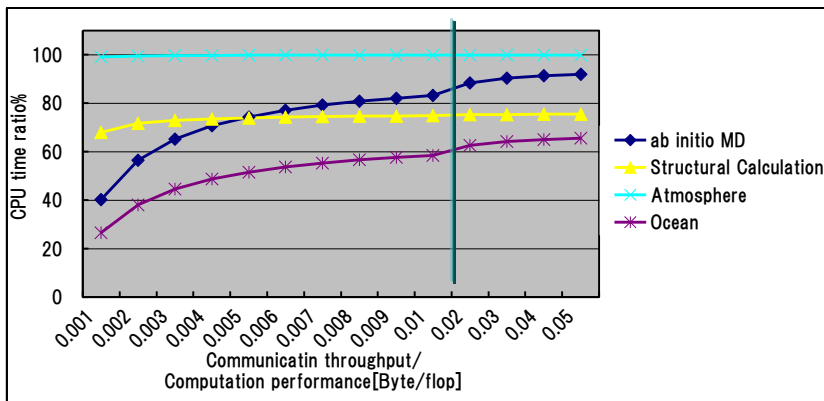


Copyright © Hitachi, Ltd. 2011 All rights reserved 19

### 3-8 ネットワーク性能と演算性能のバランス

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

- 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

Copyright © Hitachi, Ltd. 2011 All rights reserved 20

### ◆アプリケーションの要請

$$\frac{(\text{メモリバンド幅 GB/s})}{(\text{演算性能 GFlop/s})} > 0.4 [\text{Byte/flop}]$$

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

### ◆マルチGPUシステムのバランス

$$\frac{(\text{メモリバンド幅 GB/s})}{(\text{演算性能 GFlop/s})} = 0.25 [\text{Byte/flop}]$$

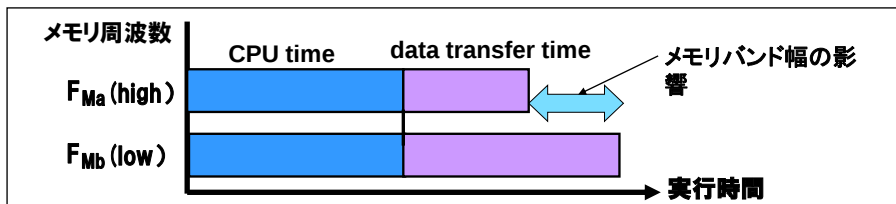
$$\frac{(\text{ネットワークバンド幅 GB/s})}{(\text{演算性能 GFlop/s})} = 0.004 [\text{Byte/flop}]$$

### ◆実際にアプリケーションの性能はどうなるか？

Copyright © Hitachi, Ltd. 2011 All rights reserved 21

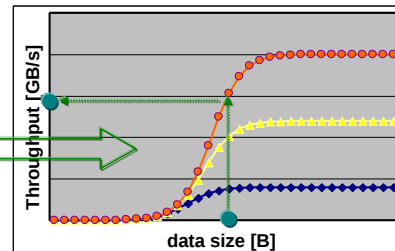
## 3-10 GPUシステムでのアプリの性能推定

**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

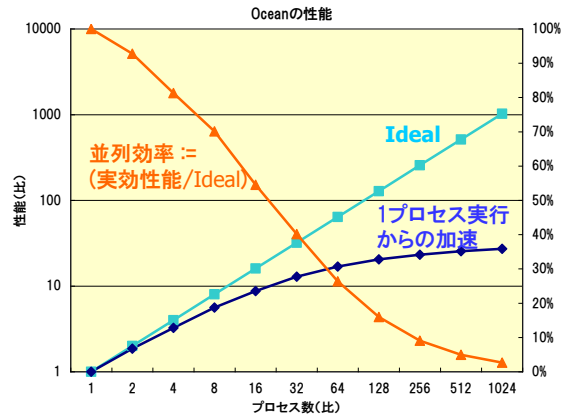


Copyright © Hitachi, Ltd. 2011 All rights reserved 22

## 3-11

## 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  $\Rightarrow$  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  $\Rightarrow$  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  $\Rightarrow$  全実行時間の9.4%
- ・ 演算効率は B/F から推測  $\Rightarrow$  3.3%

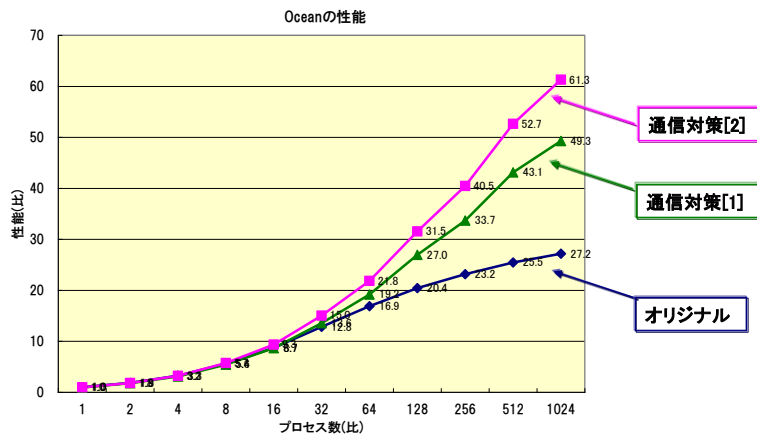


Copyright © Hitachi, Ltd. 2011 All rights reserved 23

## 3-12

## 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、InfiniBand を追加して2方向の隣接間通信を同時実行

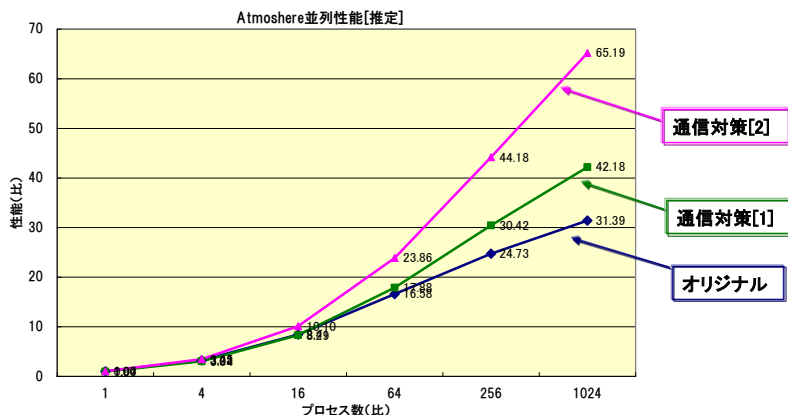


Copyright © Hitachi, Ltd. 2011 All rights reserved 24

### 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 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、InfiniBandを追加して2方向の隣接間通信を同時実行

Copyright © Hitachi, Ltd. 2011 All rights reserved 25

### 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

$$\frac{\text{メモリバンド幅 GB/s}}{\text{演算性能 GFlop/s}} = 0.25 [\text{Byte/flop}]$$

メモリ性能バランスはPCサーバよりやや低め。効率が良い(80%)

$$\frac{\text{ネットワークバンド幅 GB/s}}{\text{演算性能 GFlop/s}} = 0.004 [\text{Byte/flop}]$$

ネットワーク性能が相対的に低く見える。レイテンシ > 20 μs

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)

Copyright © Hitachi, Ltd. 2011 All rights reserved 26

## 4 SC' 10 日立展示

Copyright © Hitachi,Ltd.2011 All rights reserved 27

### 4-1 日立ブース



SC10 11/15-11/18  
NewOrleans



Copyright © Hitachi,Ltd.2011 All rights reserved 28

## 4-2

## 次世代サーバボード



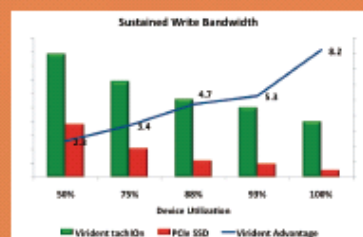
Copyright © Hitachi, Ltd. 2011 All rights reserved 29

## 4-3

### Clustered PCIe SSD

#### Virident tachION™ eliminates I/O bottleneck

- Leading SSD performance with datacenter reliability
  - Over 300K sustained 4k IOPs
- Compact half size PCIe card
- Lower power for reduced TCO
- Enterprise class reliability
  - Hardware RAID
  - Field replaceable modules
- Modular design with field upgradeable FLASH



Total Read : **41.6GB/s**  
(1.3GB/s x 32)  
With 32nodes in a 38U rack

Copyright © Hitachi, Ltd. 2011 All rights reserved 30

## Clustered GPU

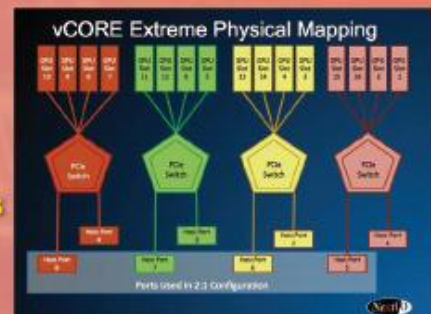


8GB/s  
(PCIe 2.0 x16)

Total Performance: **35.2TFLOPS**  
(GPU : 0.5TFLOPS x 64)  
(CPU : 0.14TFLOPS x16)  
With 16nodes + 4GPU unit  
in a 38U rack

### NextIO vCORE™ Extreme

Scale-Out GPU Computing Solution for HPC  
Scalable from 2 GPU's per server to 16 GPU's per  
server. Supports 16 GPU's and up to 8 servers  
connections in 3U



vCORE Extreme Physical Mapping

Ports labeled H 2:1 Configuration

Copyright © Hitachi,Ltd.2011 All rights reserved 31

# HITACHI

## Inspire the Next

Copyright © Hitachi,Ltd.2011 All rights reserved 32