100PFlopsを実現するためのハードウェアとその課題

～戦略的高性能計算システム開発に関するワークショップ～

2010/08/02

株式会社 日立製作所
情報・通信システム社
エンタープライズサーバ事業部　第一サーバ本部　第三部

鬼塚　久幸

HITACHI
Inspire the Next

Contents

uVALUE

HITACHI
Inspire the Next

1
自己紹介

uVALUE

1.　自己紹介

HITACHI
Inspire the Next

- 所属・氏名
  日立製作所　エンタープライズサーバ事業部
  鬼塚　久幸　(Hisayuki Onizuka)

- 担当業務
  入社後、
  ・SR11000 H1、J1、K1モデルの開発
  ・T2K(東京大学)のサーバ開発
  などを担当。

  現在は、
  HPC向けPCクラスタの開発に従事。

HITACHI
Inspire the Next

2

5年後のスーパーコンピュータ像

uVALUE

2-1. CPU

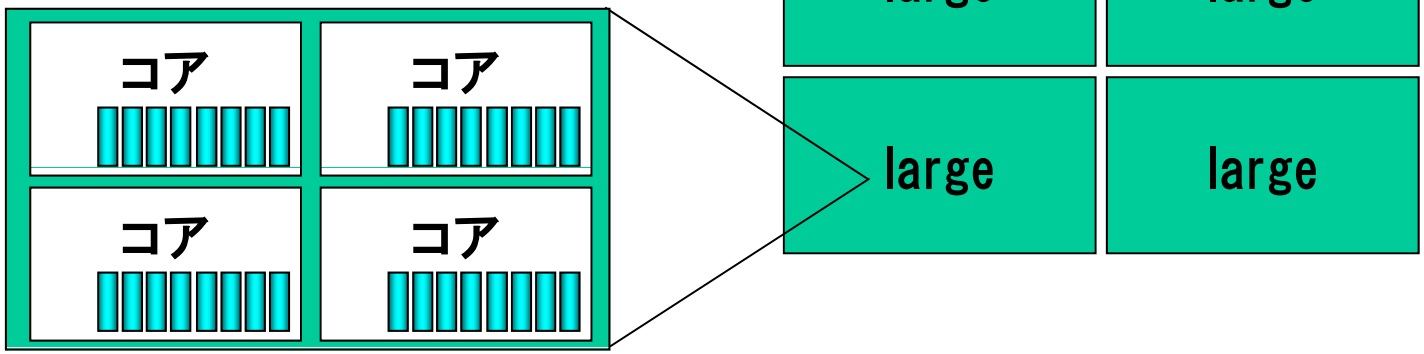
HITACHI
Inspire the Next

- 2014年でのCPU動向

Intel 2014年にはメインストリームCPU (Haswell後継のRockwell)に
第4世代のLarrabeeが統合されるとの噂。*

<予想されるスペック>
Large CPU:4コア×8FPU＝32FPU
Small CPU:16コア×32FPU＝512FPU
→ 2.0GHzだとすると1.088Tflops程度か？

100Pflopsでは、5万ノード@2socket/ノード
の規模。



*出典：後藤弘茂 http://pc.watch.impress.co.jp/docs/column/kaigai/20091222_338322.html

2-2. 5年後のセンターマシンの概要

HITACHI
Inspire the Next

T2K(東大)

- CPU
  - 36.8 [Gflops]@2.3GHz
    Quad-Core Opteron 8356

27倍

5年後の規模観

- CPU
  - 1.088[Tflops]

5年後のスペックをCPUの比から算出すると、約27倍の性能向上。これを単純にメモリ容量・メモリバンド幅、ネットワークバンド幅に適用すると・・・

- メモリ
  - 8[GB/socket]
  (・一部 32[GB/socket])
  - 10.6[GB/s] (B/F＝0.3)

- ネットワーク
  - 1.25[GB/s/socket]

- メモリ
  - 216[GB/socket]
  (・864[GB/socket])
  - 286.2[GB/s]

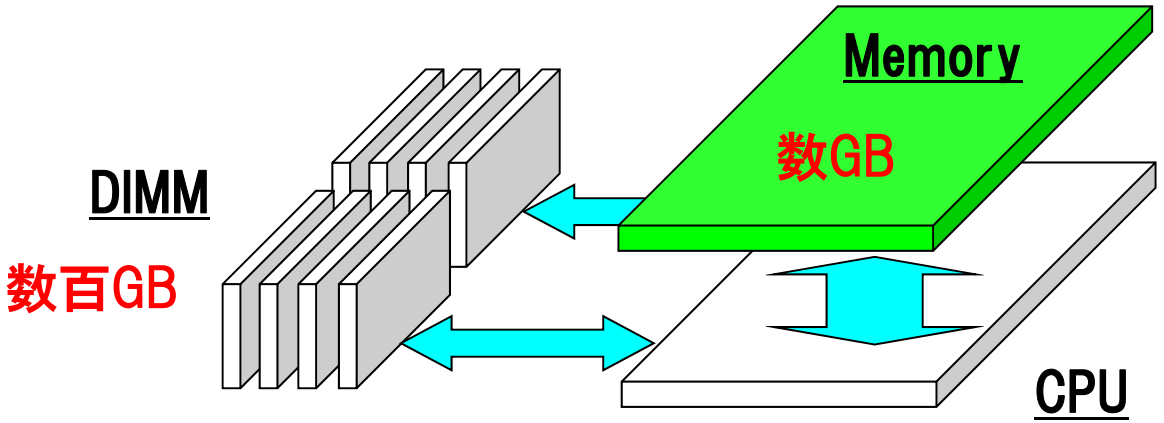
- ネットワーク
  - 33.75[GB/s/socket]

2-3. メモリ

HITACHI
Inspire the Next

B/Fを維持するためには、

- 主記憶の一部をCPUの上にチップ・スタックする方法。
  シリコン貫通ビアでスタック・メモリと接合。
  →基板上の配線制約を受けないことが利点で、チャネル数を増やせるが
  スタックメモリとしては、8層でせいぜい数GB程度がいいところ。
  メモリ容量が限られることが問題。米国UHPC(DARPA)では16GB。*



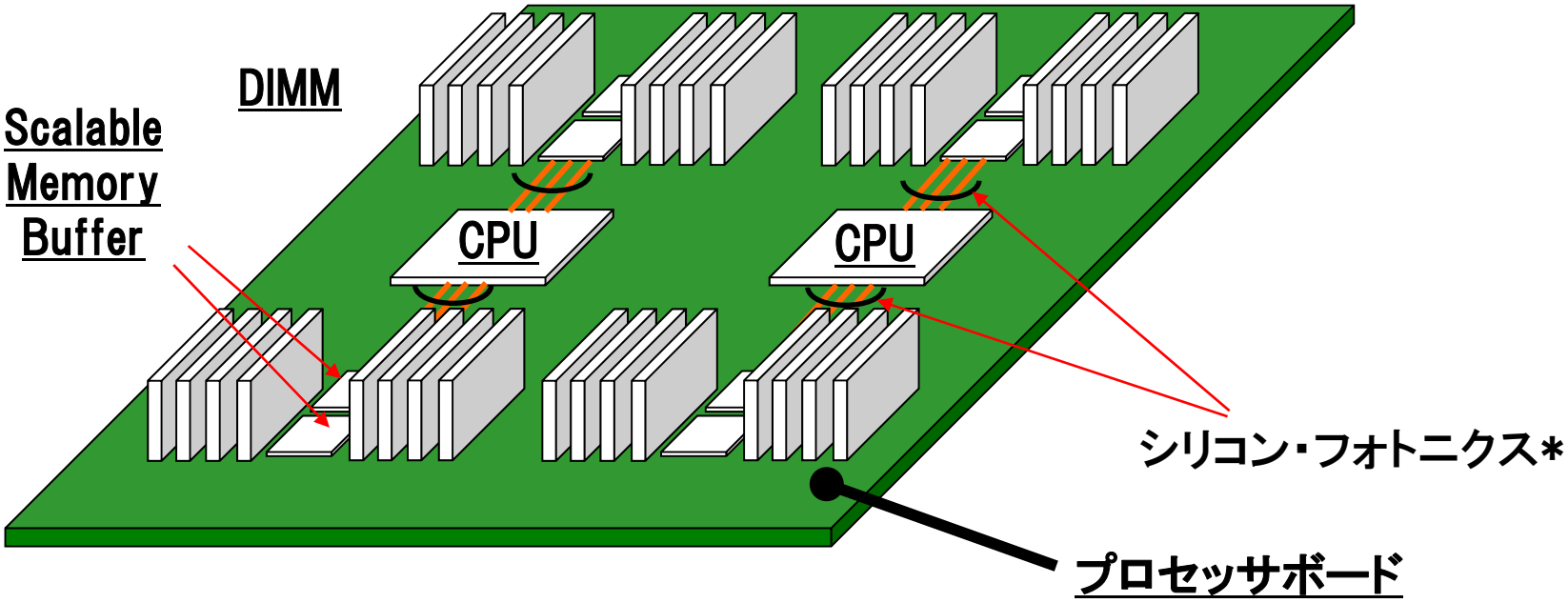
- 各プラットフォームのメモリ拡張ソリューションを使ってチャネル数を増やす。

*出典: DARPA http://www.er.doe.gov/ascr/Research/CS/DARPA%20exascale%20-%20hardware%20(2008).pdf

2-3. メモリ

HITACHI
Inspire the Next

- 各プラットフォームのメモリ拡張ソリューションを使ってチャネル数を増やす。
  DDR4(ディファレンシャル方式): 1chあたり、51.2GB/s (pinあたり、6.4Gbps)
  8chで409.6[GB/s]を実現可能。 32[GB/本]×16[本]＝512[GB/socket]



*出典: http://pc.watch.impress.co.jp/docs/news/20100728_383742.html

これでなんとか各種メモリ性能は救えるがいくつか問題も・・・

2-3. メモリ

HITACHI
Inspire the Next

- 電気信号の限界
  SMB-CPU間の帯域をPCBで実現する方法。やはり光しかないか？
  PCB上で光伝送を実現する技術開発(シリコンフォトニクス)。コストは？

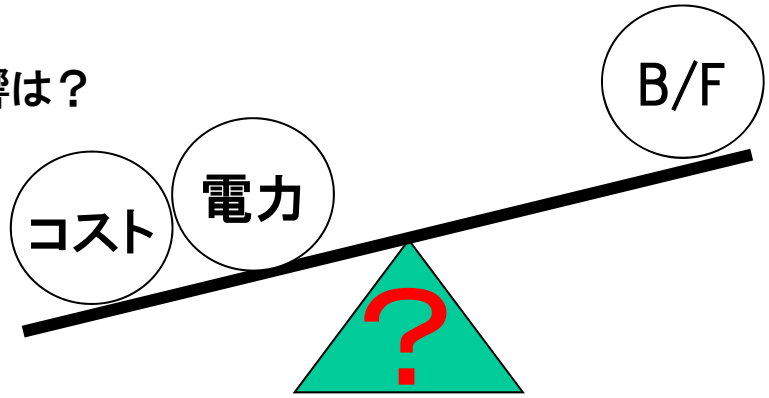
- DDR4(ディファレンシャル方式)での配線層数
  ピン数、配線数は2倍。層数増加によるコストは？ DIMMパッケージの大きさは？

- SMB (Scalable Memory Buffer)の消費電力
  Nehalem-EX世代のSMBであるMill Brookの消費電力は7～8[W]程度。
  仮に、消費電力がこのままなら、5年後のスペックを達成するには、ノード当たり8個程度必要。
  →64[W]程度。
  このメモリ性能を実現するために、システム全体で新たに
  64[W]×5万ノード＝3.2[MW]必要となる。
  当然、1M＝1億円とすると、運用コスト(3.2億円)も余計に掛かる。

- B/Fと電力・コストはトレードオフ。
  B/Fをハードで手を抜いた場合、アプリ開発への影響は？

米国UHPC(DARPA)ですら、なんとか B/F＝0.16 (※)
を実現。

※メモリバンド幅:88[GW/s], CPU性能:4.5[Tflops]
1[Word]=8[Byte]で計算



2-4. ネットワーク

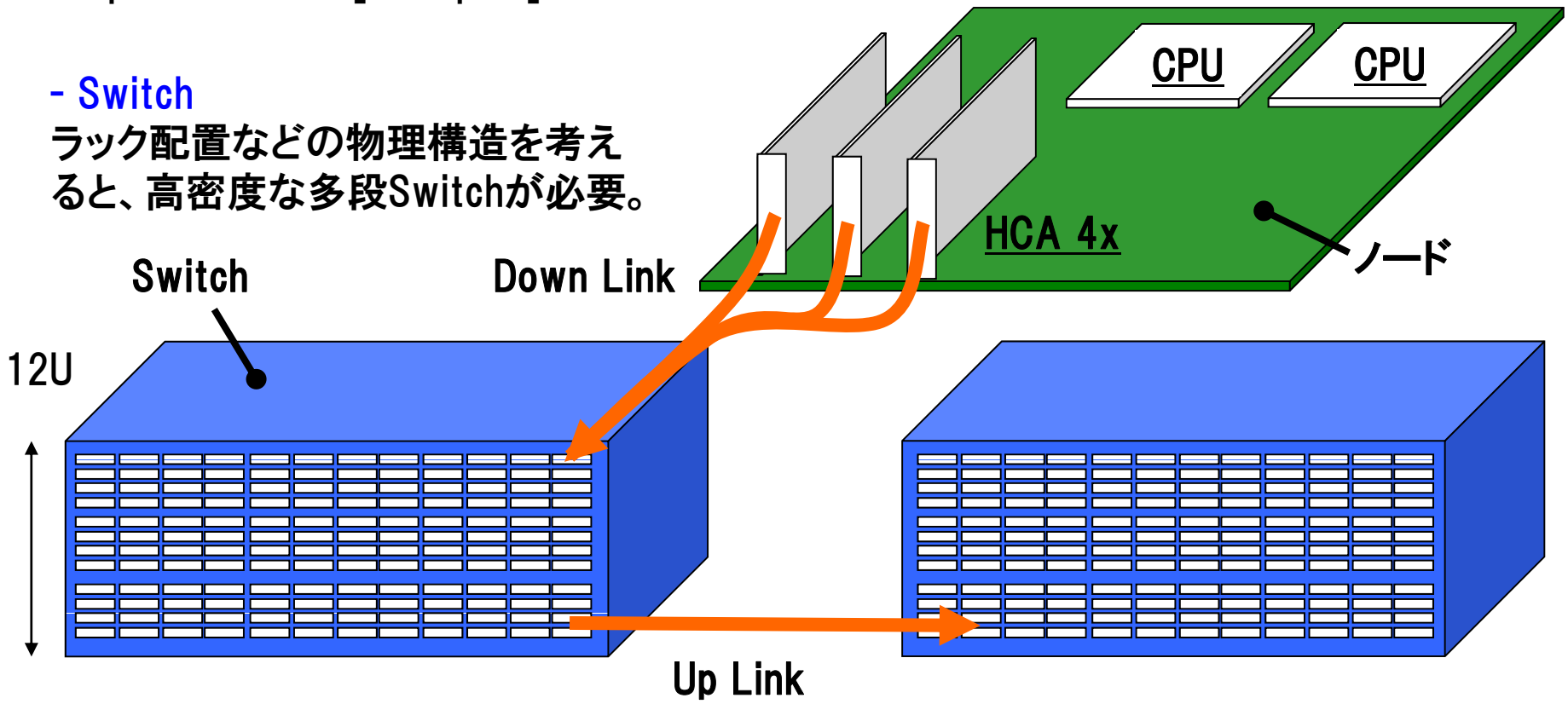
HITACHI
Inspire the Next

- InfiniBand® HDR 4xを3本で、33.75×2(socket)＝67.5[GB/s]を実現可能。
  * Hexadecimal Data Rate

144[CXP port/switch]＝12[CXP port/linecard]×12[linecard]
・Down Link： 72[CXP port]＝72[ノード]×3[QSFP port]
・Up Link　： 72[CXP port]

- Switch

ラック配置などの物理構造を考えると、高密度な多段Switchが必要。



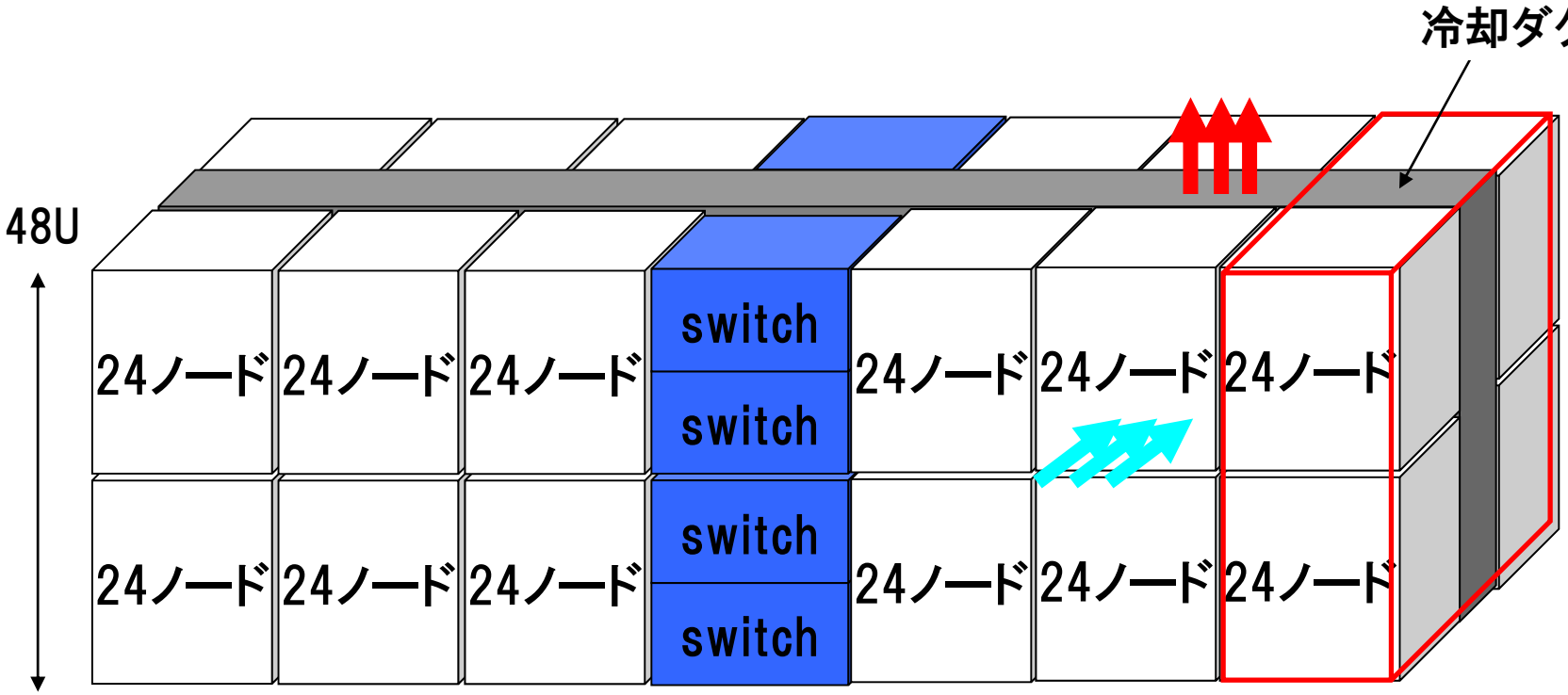
uVALUE

2-5. ラッキング

HITACHI
Inspire the Next

- ラッキング

2socket/1Uサーバを96[ノード/ラック]に搭載するとして、



100[PFlops]＝5万[ノード]×2[socket/ノード]×1[Tflops/socket]
5万ノード＝96[ノード/ラック]×520[ラック]

計算ノードだけで、520ラックも必要！

HITACHI
Inspire the Next

3

10年後に向けて

uVALUE

3.　10年後に向けて

HITACHI
Inspire the Next

・5年後のハードウェアは、現在のサーバ開発のスキームの延長で、大きな課題はあるもののなんとかやっていけそうな範囲ではあるが、10年後は有りモノでのハードウェアでの実現は難しい。

・エクサFlopsを追求するとどうしてもメモリバンド幅がなおざりになってしまう。米国UHPC(DARPA)でのエクサFlopsマシンにリーズナブルにスケーリングすると思われるアプリは、HPL、WRF、AVUS(※) 他、数種類程度。

※ HPL：High Performance LINPACK
　WRF：Weather Research and Forcasting code
　AVUS：Air Vehicle Unstructured Solver

・低いB/Fに寛容なアプリ開発のスキームも今後必要になってくるのでは？

・5年後のペタFlopsマシンは、ハード・ソフト両面でエクサFlops開発に向けた重要な存在。
5年後のペタFlopsマシンでの、アプリにおけるスケーリングの研究は、次に繋げるために今後必要になってくるはず。


uVALUE