性能要求表精査

丸山直也

2013年8月6日

メモリバンド幅

- 現状の問題点
  - 数値の根拠があいまい
  - 演算数はアーキテクチャによらずほぼ一定だが（数学関数を除く）、メモリへのアクセスは厳密にはメモリアーキテクチャ／アルゴリズムに依存
- 対策
  - アクセス量をアルゴリズムから導出もしくは実測

メモリアクセス量記載事項

- メモリアクセス量
  - ロード・ストアサイズ
- 積算根拠
  - 実測もしくはアルゴリズムから算出
- 測定条件（実測の場合）
  - 測定システム
  - 問題条件設定
- 実測方法
  - プロファイラ等によりキャッシュとDRAMの間の転送量を計測

参考情報： 京でのお手軽計測方法

- 京／FX10の簡易プロファイラを利用
  - “hwm” オプションを指定
  - “Mem throughput_chip” x ”Elapsed” ➔ アクセス量

実行

```
fipp -C -Ihwm -d profile-dir-path mpiexec ./a.out
```

表示

```
fipppx -A --Ihwm -d profile-directory-path
```

例

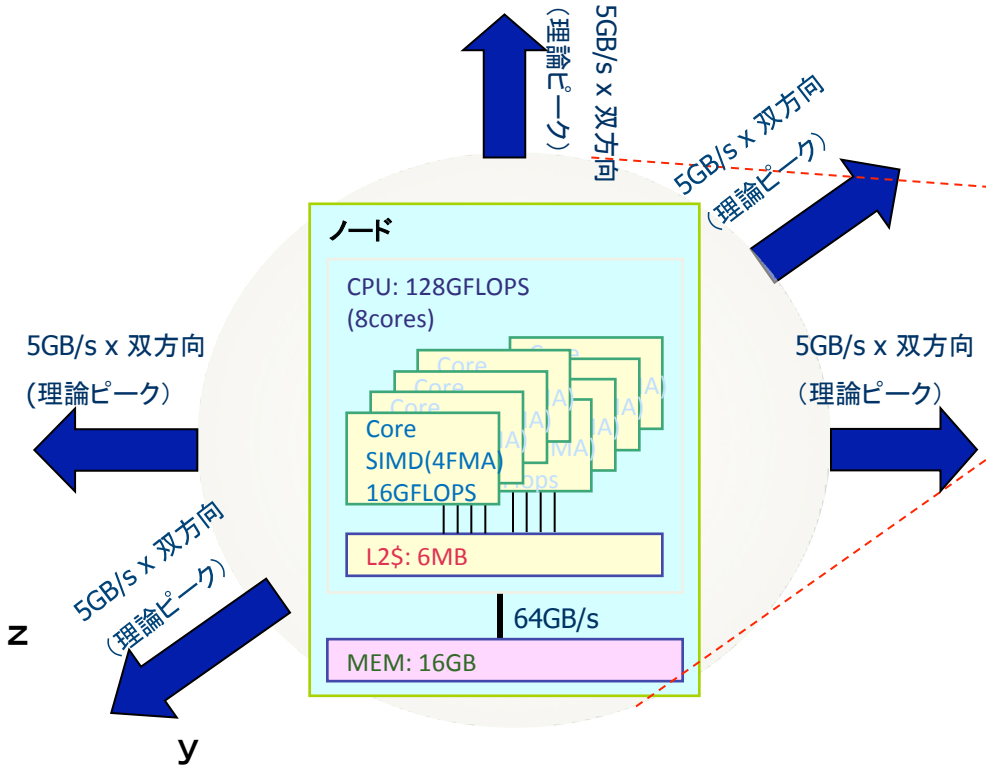
```
********************************************************************************
       Application - performance monitors
********************************************************************************

      Elapsed(s)       MFLOPS MFLOPS/PEAK(%)           MIPS   MIPS/PEAK(%)
   ---------------------------------------------------------------------------
       59.4789     29169.0471        12.3313     63239.3209   53.4694   Application
   ---------------------------------------------------------------------------
       59.4789     29169.0471        12.3313     63239.3209    53.4694   Process      0

                   Mem throughput   Mem throughput
      Elapsed(s)   _chip(MB/S)      /PEAK(%)         SIMD(%)
   ---------------------------------------------------------
       59.4789     20049.8082        23.5017         0.1537   Application
   ---------------------------------------------------------
       59.4789     20049.8082        23.5017         0.1537   Process      0
```

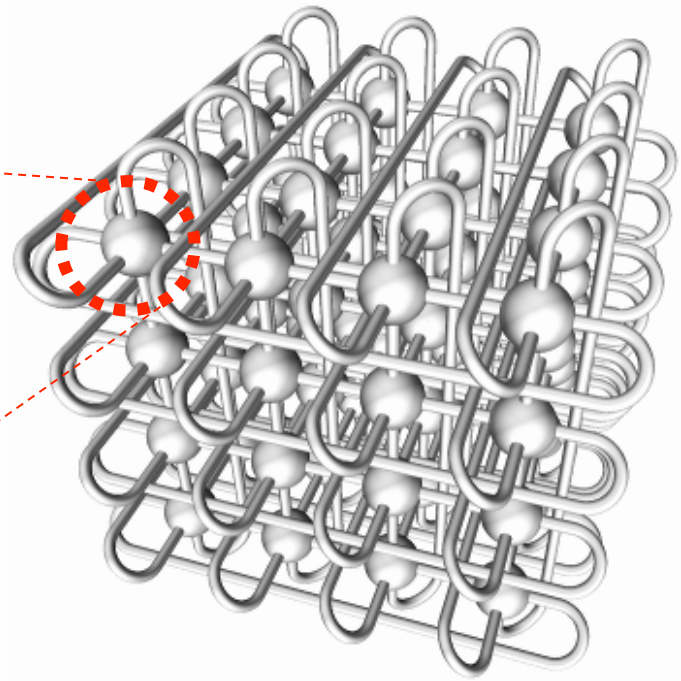
ネットワークアーキテクチャ用語集

- トポロジー
  - 直接網（トーラスなど）
  - 間接網（ファットツリーなど）
- レイテンシ
  - ネットワーク直径（最大ホップ数）
- バンド幅
  - インジェクションバンド幅
    - 各ノードからの総バンド幅
  - バイセクションバンド幅
    - 全体を半分に切った2集合間の総バンド幅
- フルバイセクション
  - インジェクションバンド幅*0.5=バイセクションバンド幅

京ネットワーク構成

- 計算ノードの構成
  - CPU（8コア）　：1個
  - ICC（インターコネクト用LSI)　：1個
  - メモリ　：16GB
- インターコネクトの構成
  - ユーザービューは3次元トーラス
  - 帯域：3次元の正負各方向にそれぞれ 5GB/s x 2（双方向）【理論ピーク】
  - バイセクションバンド幅 30TB（双方向）
  - ケーブル：約200,000本，約1000km

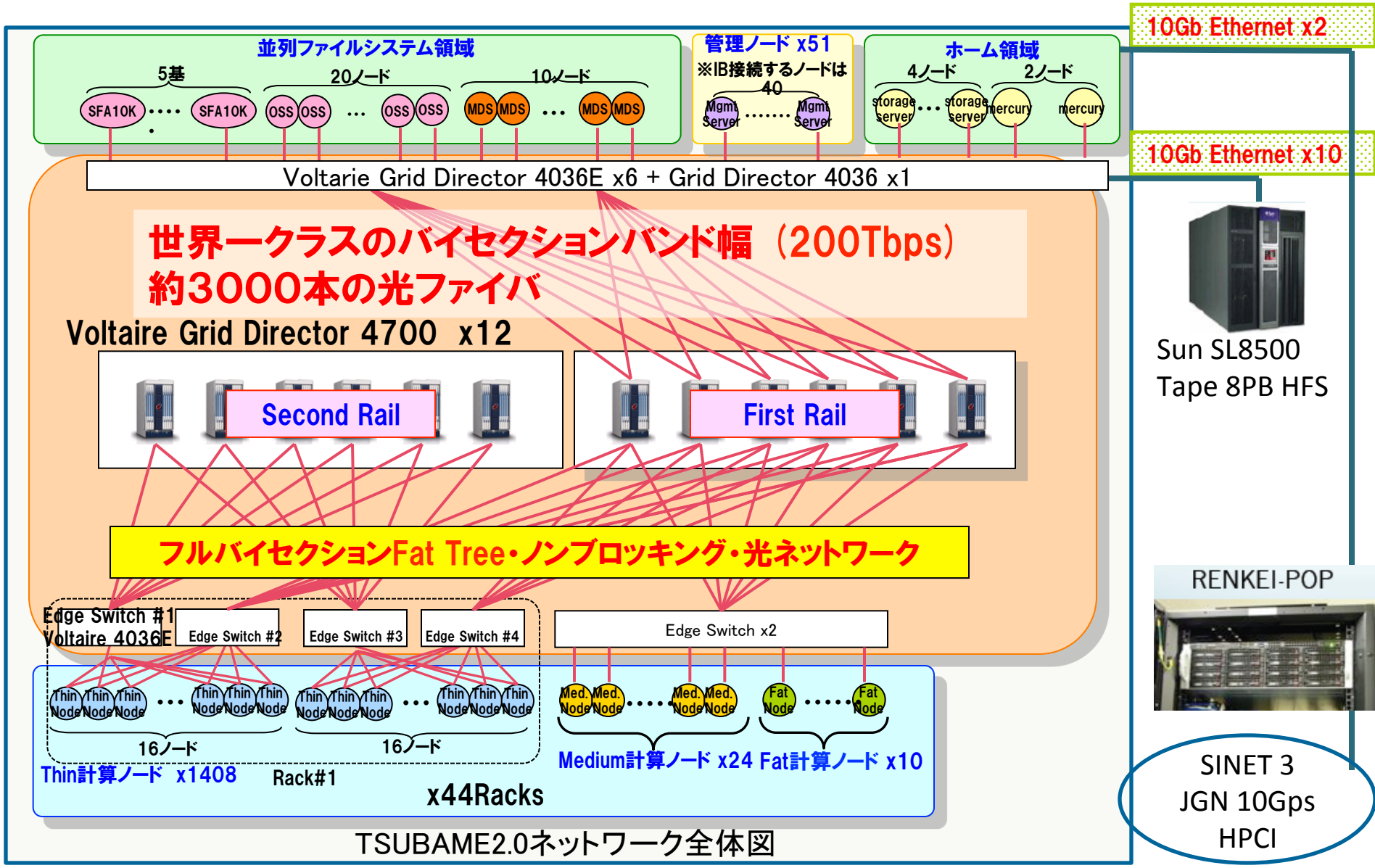




3次元トーラスのイメージ

提供：富士通（株）

TSUBAME2.0ノード間相互結合網



TSUBAME2.0ネットワーク全体図

想定されるエクサシステム構成
ノード間インターコネクト

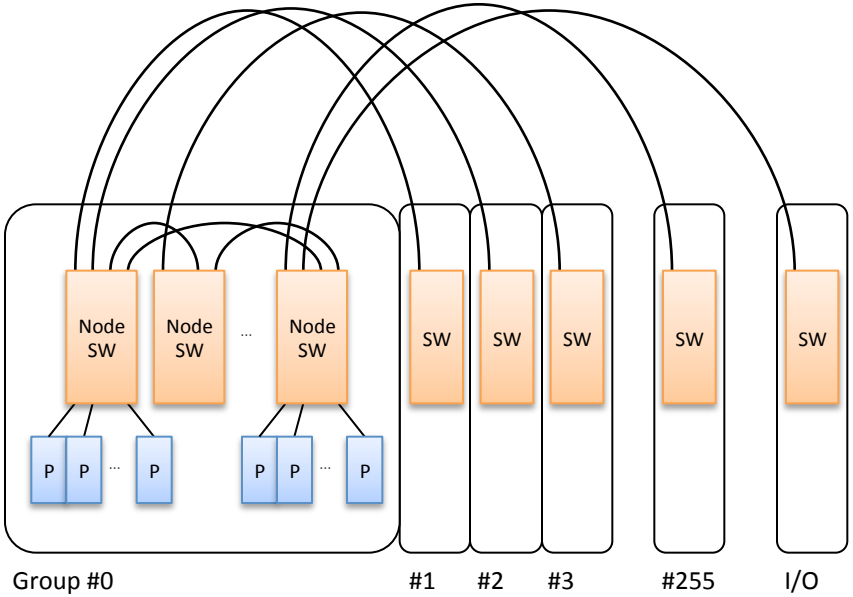
High-radix NW (Dragonflyの例)

レイテンシ

隣接最短: 2hop, 100ns/dir
隣接最長: 5hop, 1000ns/dir (全ノードでメッシュを構築した時の最悪部分)
最も遠いノード間の通信: 5hop,1000ns/dir

バンド幅

ノードからのinjection帯域: 25〜50 GB/s
2ノード間の通信帯域: 25〜50 GB/s
バイセクションバンド幅: 2.5〜5.0 PB/s



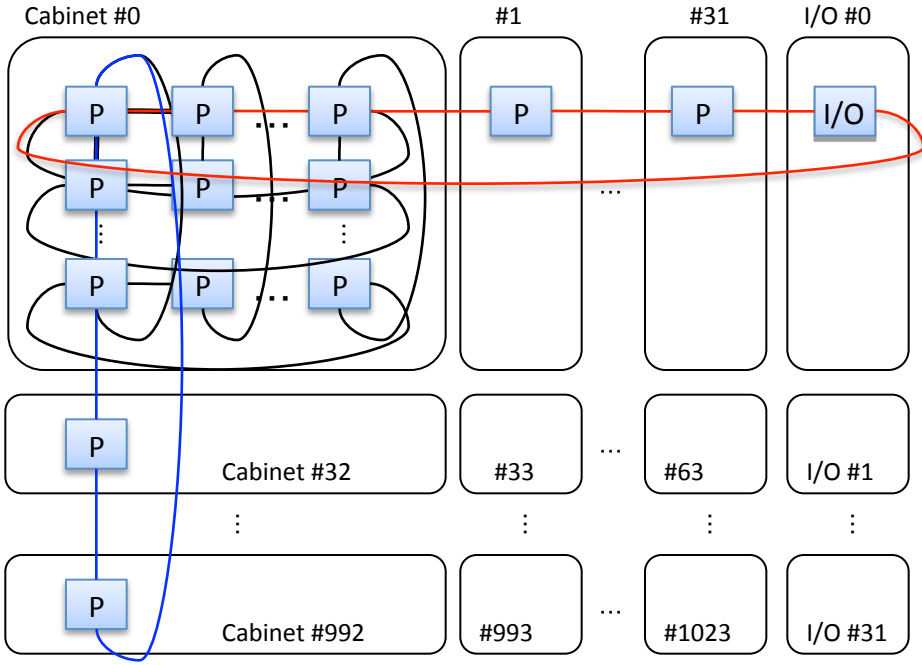
Low-radix NW (4Dトーラスの例)

レイテンシ

隣接最短: 1hop, 50ns/dir
隣接最長: 1hop, 100ns/dir (全ノードでメッシュを構築した時の最悪部分)
最も遠いノード間の通信: ~50hop, ~5000ns/dir

バンド幅

ノードからのinjection帯域: 50〜100 GB/s
2ノード間の通信帯域: 6.3 〜 12.5 GB/s (minimal routingの場合)
バイセクションバンド幅: 0.3 〜 0.5 PB/s



複数cabinetでgroupを形成、別にI/O用のラックを接続

アプリケーションのネットワーク要求

- アーキテクチャによらない記述とする
- 例
  - 3次元領域分割、隣接＋リダクション、隣接サイズ: 毎イテレーションノードあたりXバイト、毎イテレーションYバイトのリダクション
- 記載項目
  - 問題分割
  - 通信パターン
  - 通信量モデル
  - 通信にかけられる時間

ネットワーク記載項目

- 問題分割方法
- イテレーション毎の通信パターン
  - 隣接、集団通信（MPIの集団通信API名）、など
  - 集団通信の場合はMPIコミュニケータの構成（MPI_COMM_WORLDかもしくは分割したコミュニケーターか）
- パターン毎の通信サイズモデル
- 時間制約
  - アプリケーション実行全体時間の内、通信に要する時間制約

性能数値精査

- ミニアプリ化における調査結果（プロファイル）との比較
  - 提供アプリほぼすべてについて基本的な性能プロファイルを取得
  - 計算量、メモリアクセス量、メモリ使用量、通信パターン、ファイルI/O
- AICS運用技術部門におけるこれまでの知見