スパコンのネットワークをサクサクにする

ULP-HPC のインターコネクト技術

Interconnect Technologies for Next-generation HPC Systems

鯉渕 道紘　Michihiro KOIBUCHI　　藤原 一毅　Ikki FUJIWARA

どんな研究？

スーパーコンピュータの正体は、ネットワークでつながった数千台の小さなマシンです。マシンがどんなに速くても、ネットワークが遅かったらスパコンは性能を発揮できません。また、数千台のマシンをつなぐ、1,000キロメートルを超えるケーブルも悩みの種です。私たちは、スパコンのネットワークを「**もっと速く**」「**もっとスリムに**」する方法を探究しています。

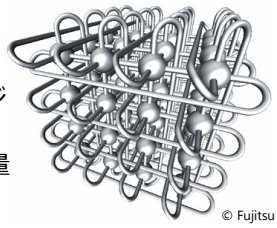
何がわかる？

ネットワークの「低遅延化」と「省資源化」を両立するために、私たちは2つの方法を考えています。ひとつは、今まで規則的につないでいたマシン同士を**ランダムにつなぐ**こと。もうひとつは、ケーブルの代わりに**レーザービーム**を使うこと。これらの方法により、次世代のスパコンに求められる高性能ネットワークをシンプルに実現できることがわかりました。

状況設定

- 従来のスパコンのネットワークの限界
  - 計算速度と通信速度の乖離
  - 大きなスイッチ遅延と規則的トポロジ → 通信遅延が小さくならない
  - ケーブルだけで 100 トンを超える物量 → 故障しても修理できない

広帯域指向／重厚長大な設計



▲ 3D Torusトポロジ（京コンピュータ）

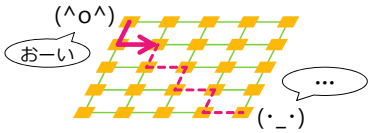
- 次世代スパコンに求められるネットワークの要件
  - ノード数　～10,000
  - 通信遅延　1 マイクロ秒以下
  - メッセージサイズ　3KB 未満

小さなメッセージを／頻繁に／遠いところへ／早く届けたい！

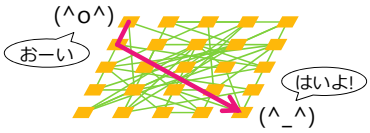
▲ 2,400kmに及ぶ配線（地球シミュレータ）

研究内容（1）

■ ランダムトポロジ



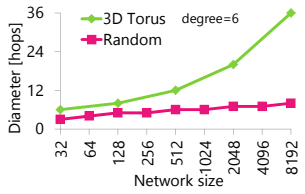
▲ 規則的なトポロジ：
メッセージが届くのに時間がかかる



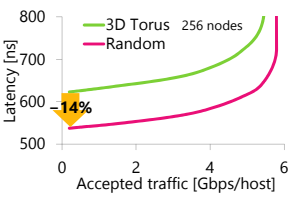
▲ ランダムトポロジ：
近道があるのでメッセージがすぐ届く

- ノード間をランダムにつなぐと、スモールワールド効果により大きなネットワークほど劇的に距離が縮まる
  - 通信遅延も減少
  - スループット・耐故障性が向上
  - 乱数による性能差は無視できる
- ただし、ケーブルの総延長は増大
  - ランダムリンクの長さを制限
  - ラック配置を最適化
- 既存ネットワークのケーブルをランダムに差し替えるだけでも同様の効果
  - これならケーブルが増えない

▼ 最大距離（縦軸）とノード数（横軸）



▼ 遅延（縦軸）とスループット（横軸）



発表論文

[1] Michihiro Koibuchi, Hiroki Matsutani, Hideharu Amano, D. Frank Hsu, Henri Casanova: "A Case for Random Shortcut Topologies for HPC Interconnects", The 39th International Symposium on Computer Architecture (ISCA), Jun. 2012.

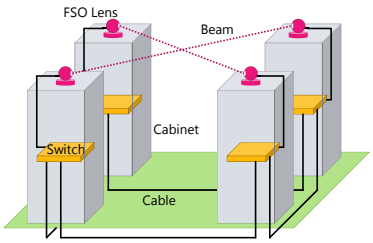
[2] Michihiro Koibuchi, Ikki Fujiwara, Hiroki Matsutani, Henri Casanova: "Layout-conscious Random Topologies for HPC Off-chip Interconnects", The 19th International Symposium on High-Performance Computer Architecture (HPCA), Feb. 2013.

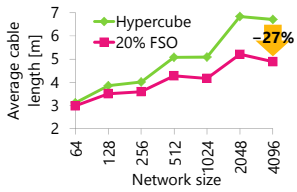
[3] 河野隆太, 藤原一毅, 松谷宏紀, 天野英晴, 鯉渕道紘: "ホストから複数リンクを用いた低遅延ネットワークトポロジ", 電子情報通信学会 コンピュータシステム研究会, 2013年1月.

[4] 藤原一毅, 鯉渕道紘: "ランダムなネットワークトポロジのラック配置最適化に関する研究", 電子情報通信学会論文誌 J96-D(8), 2013年8月.

研究内容（2）

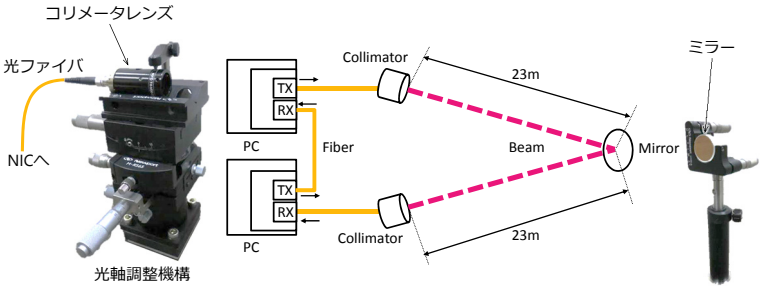
■ 光空間無線通信（FSO）





▲ ラック間ケーブルの 20% を FSO に置き換えると、平均ケーブル長が 27% 減る

- マシンラックの上にレンズを置き、ラック間を光ビームで接続
  - 10GBASE-LR／40GBASE-LR4 の赤外光を利用（λ=1310nm）
- 46m 離れたレンズ間でミラーを介した通信実験
  - **40Gbps** ワイヤレートで長時間安定した通信を確認
- ビームの向きを変えるだけでトポロジを変えられる
  - ランダムトポロジ化、パーティショニング、故障回復



発表論文

[5] 鯉渕道紘, 藤原一毅, 長谷川洋平, 橋本陽一, 松谷宏紀, 天野英晴: "光空間リンクを用いた省配線・可変トポロジであるHPC相互結合網", 情報処理学会 HPC/ARC研究会, 2012年12月.

連絡先：鯉渕道紘 ／ 国立情報学研究所 アーキテクチャ科学研究系　准教授

TEL : 03-4212-2575　FAX : 03-4212-2120　Email : koibuchi@nii.ac.jp