スーパーコンピュータにおける電力性能最適化フレームワークの評価

井上弘士　〇小野 貴継

九州大学大学院システム情報科学研究院

2018年5月11日

九州大学

内容

- HPCにおけるインターコネクション・ネットワークの消費電力
- 消費電力削減技術
- 既存のインターコネクト・シミュレータ
- 低電力モードをサポートするシミュレータ（TraceRP）の開発
- 電力と性能の評価
- まとめ

ネットワークの消費電力

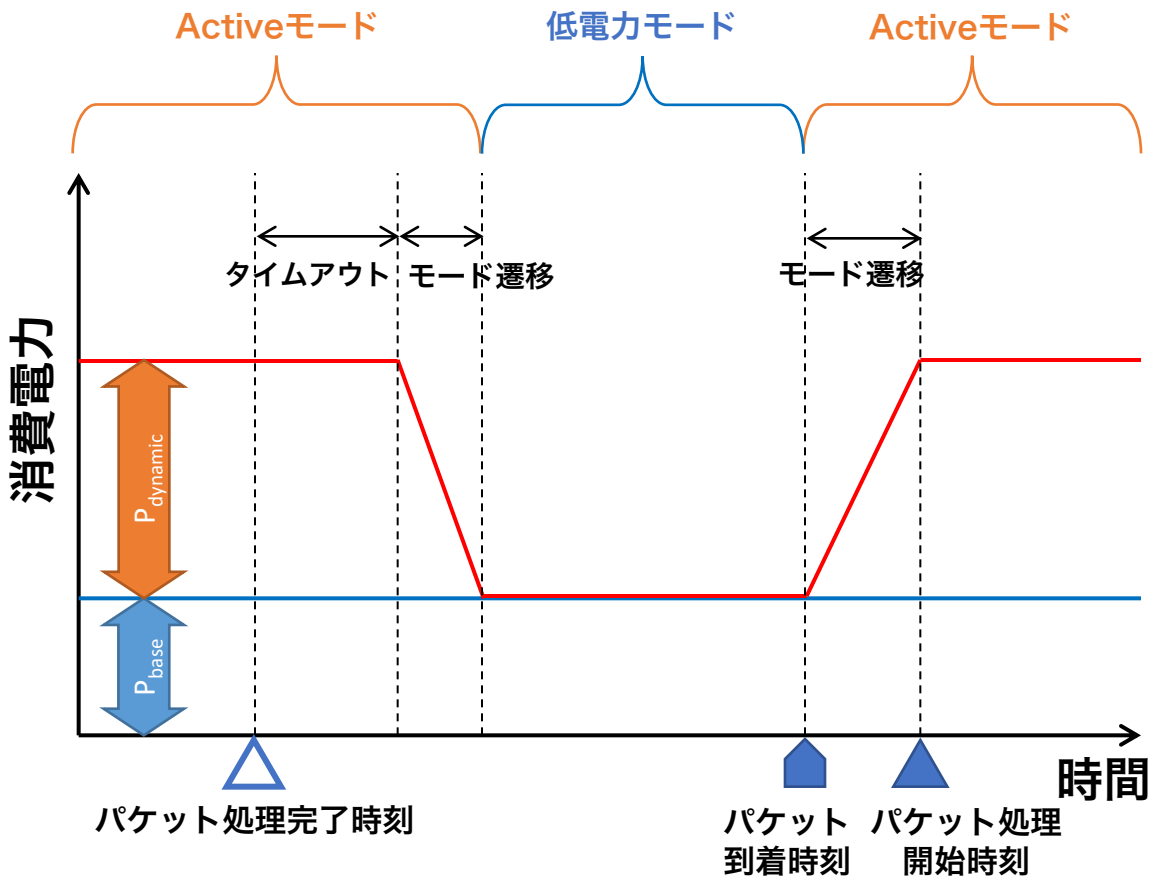
- インターコネクション・ネットワーク省電力化が課題
  - ネットワークの電力は最大でHPCシステム全体の33%*

- ネットワークの電力
  - リンクの電力が支配的
  - 通信の有無に関わらず一定の電力を消費

ネットワークの消費電力削減技術が必要

*Kogge, P. M.: Architectural Challenges at the Exascale Fron- tier, *Simulating the Future: Using One Million Cores and Beyond (invited talk)* (2008)

消費電力削減技術（On/Offリンク）



既存のインターコネクトシミュレータと課題

- ネットワークの性能や電力を見積もる手段
  - 一般にシミュレータを利用
- 既存のシミュレータ
  - 性能推定（例：BigSim*, Nsim**）
  - 性能/電力推定（例：Dimemas***）
- 既存の性能/電力シミュレータの課題
  - HPCシステムの利用において頻出する状況（マルチジョブ実行）を未サポート

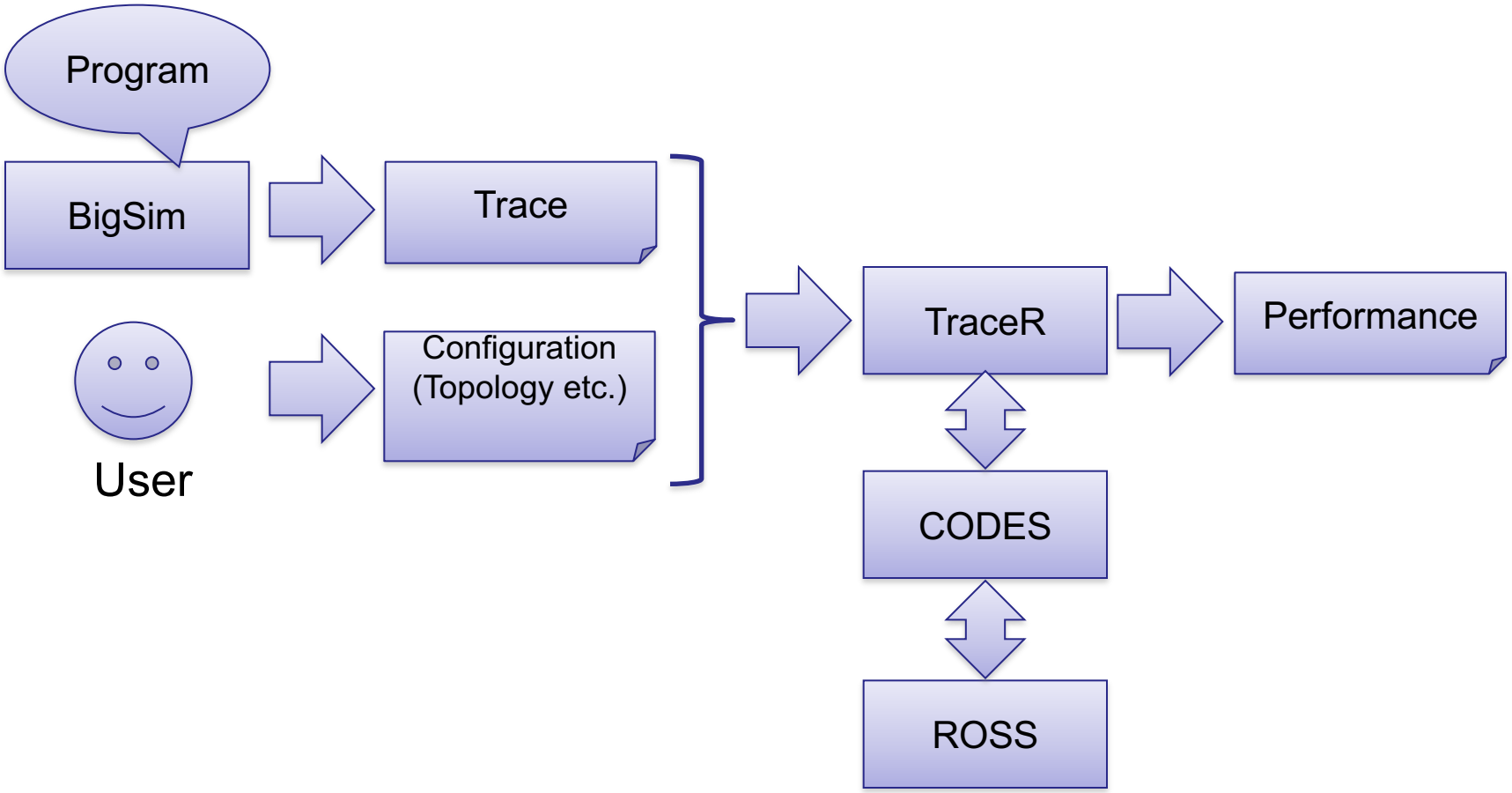
On/Offリンクとマルチジョブをサポートするシミュレータが必要

*Zheng, G., et al. : BigSim: a parallel simulator for performance prediction of extremely large parallel machines, *18th International Parallel and Distributed Processing Symposium, 2004. Proceedings.* (2004).
**Hideki, M., et al. : NSIM: An Interconnection Network Simulator for Extreme-Scale Parallel Computers, *IEICE Transactions on Information and Systems*, Vol. E94.D, No. 12, pp. 2298–2308 (2011).
***Saravanan, K., et al. : Power/Performance Evaluation of Energy Efficient Ethernet (EEE) for High Performance Computing, *Proceedings of the 2013 IEEE International Symposium on Performance Analysis of Systems and Software*, pp. 205–214 (2013).

TraceRPの開発要件と方針

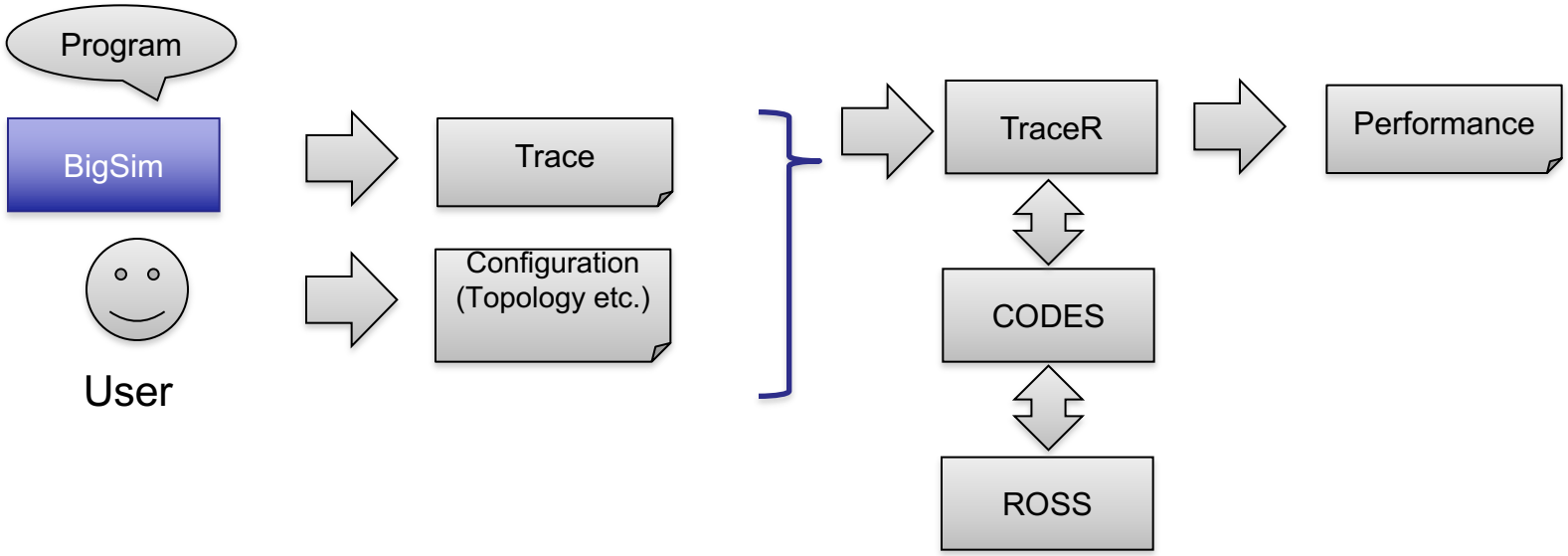
- 要件
  - 性能およびリンクの消費電力推定
  - 主要なトポロジに対応
  - マルチジョブサポート
  - On/Offリンクサポート
- 方針
  - マルチジョブに対応しているTraceRをベースに開発

TraceR



Jain, N., Bhatele, A., White, S., Gamblin, T. and Kale, L. V.: Evaluating HPC Networks via Simulation of Parallel Work- loads, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 14:1–14:12 (2016).

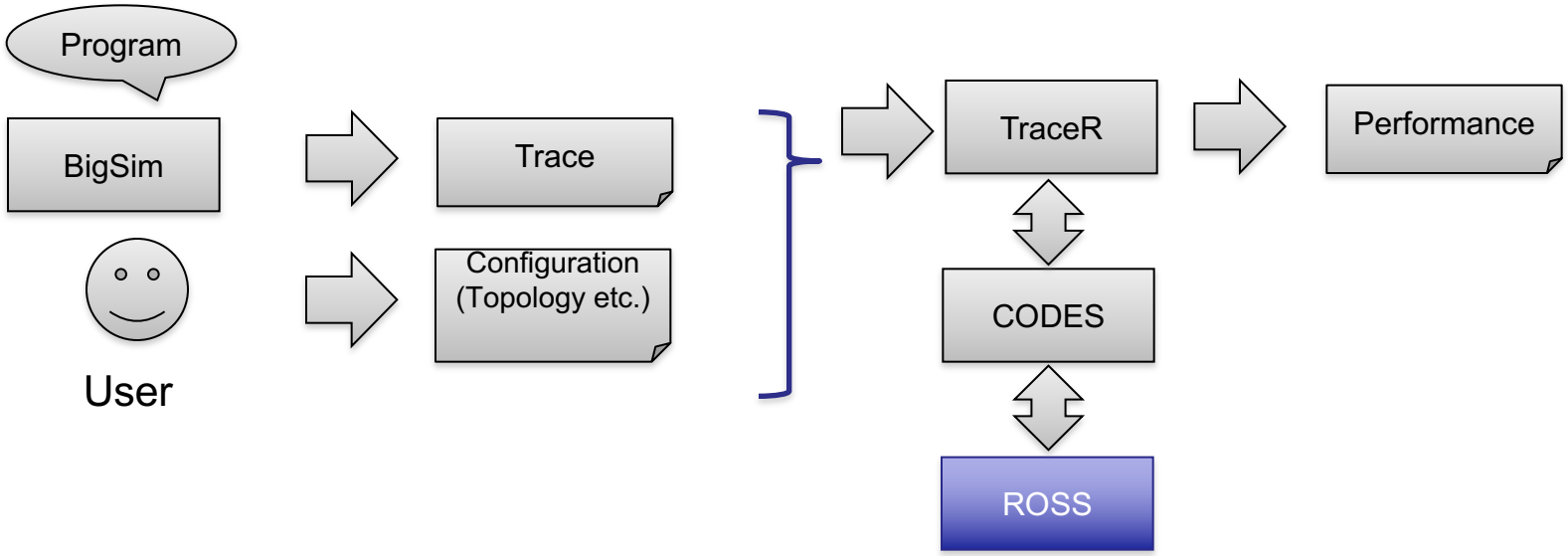
TraceR



- BigSim*
  - アプリケーション・プログラムの通信パタンを取得
  - CHARM++の仮想化機能を利用することで物理コア数よりも多くのプロセスを実行可能
  - 大規模ネットワークの通信トレース生成を実現

*Zheng, G., Kakulapati, G. and Kale´, L. V.: BigSim: a parallel simulator for performance prediction of extremely large parallel machines, *18th International Parallel and Distributed Processing Symposium, 2004. Proceedings.* (2004).

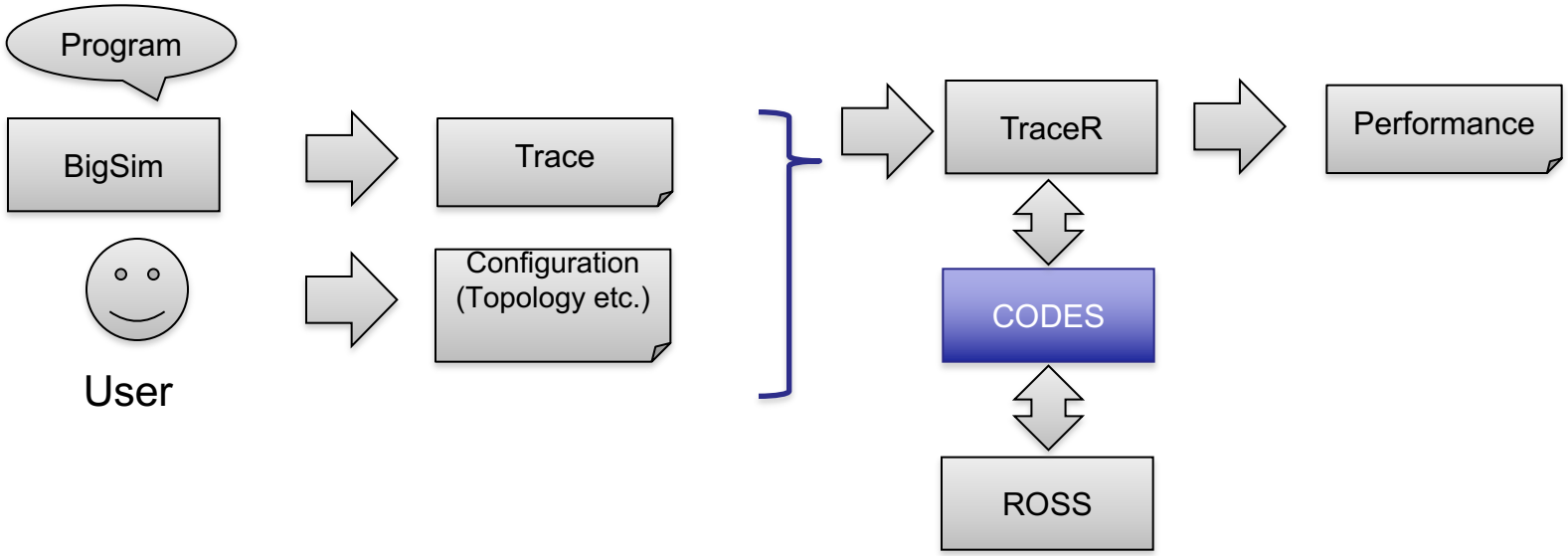
TraceR



- ROSS*
  - 大規模並列離散事象シミュレータ
  - プロセス間で分散される論理プロセスを定義しタイムスタンプを有するイベントをスケジュールすることが可能

*Carothers, C. D., Bauer, D. and Pearce, S.: ROSS: a high- performance, low memory, modular time warp system, *Proceedings Fourteenth Workshop on Parallel and Distributed Simulation*, pp. 53–60 (2000).

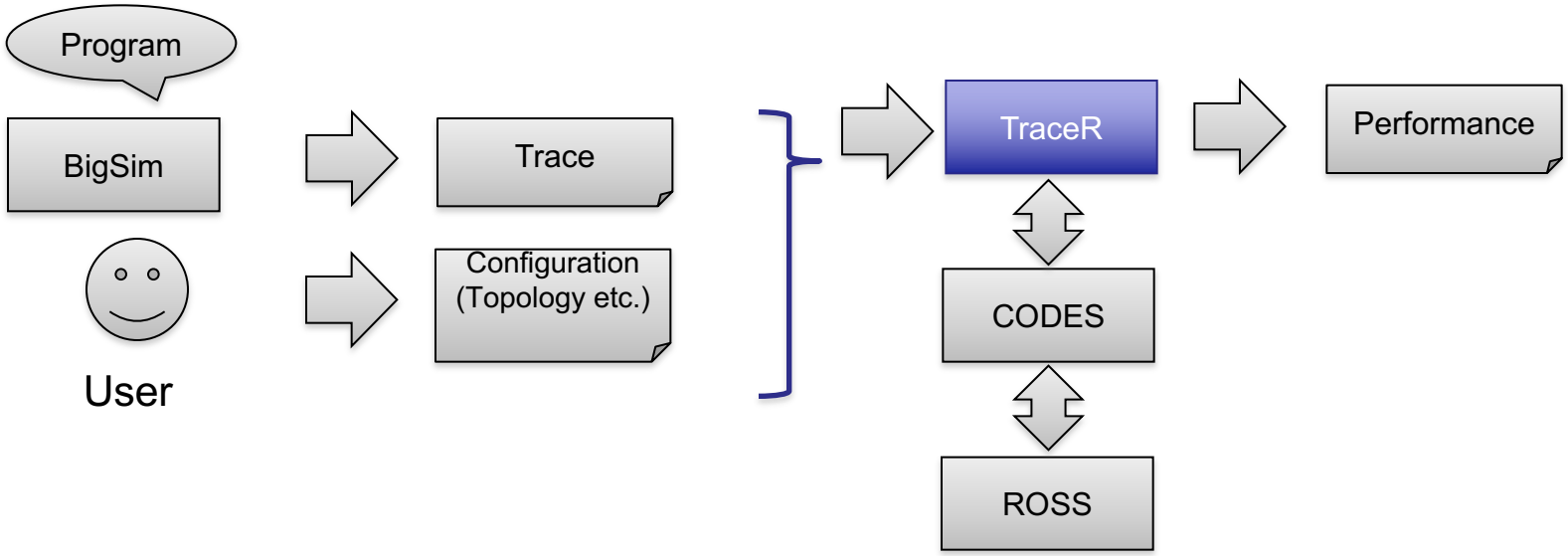
TraceR



- CODES*
  - HPC向けストレージやネットワークを対象としたシミュレータ
  - Model-netはネットワークタイプやリンクの帯域幅，レイテンシやパケットサイズなどを定義可能

*Mubarak, M., Carothers, C. D., Ross, R. B. and Carns, P.: Enabling Parallel Simulation of Large-Scale HPC Network Systems, *IEEE Trans. Parallel Distrib. Syst.*, Vol. 28, No. 1, pp. 87–100 (2017).

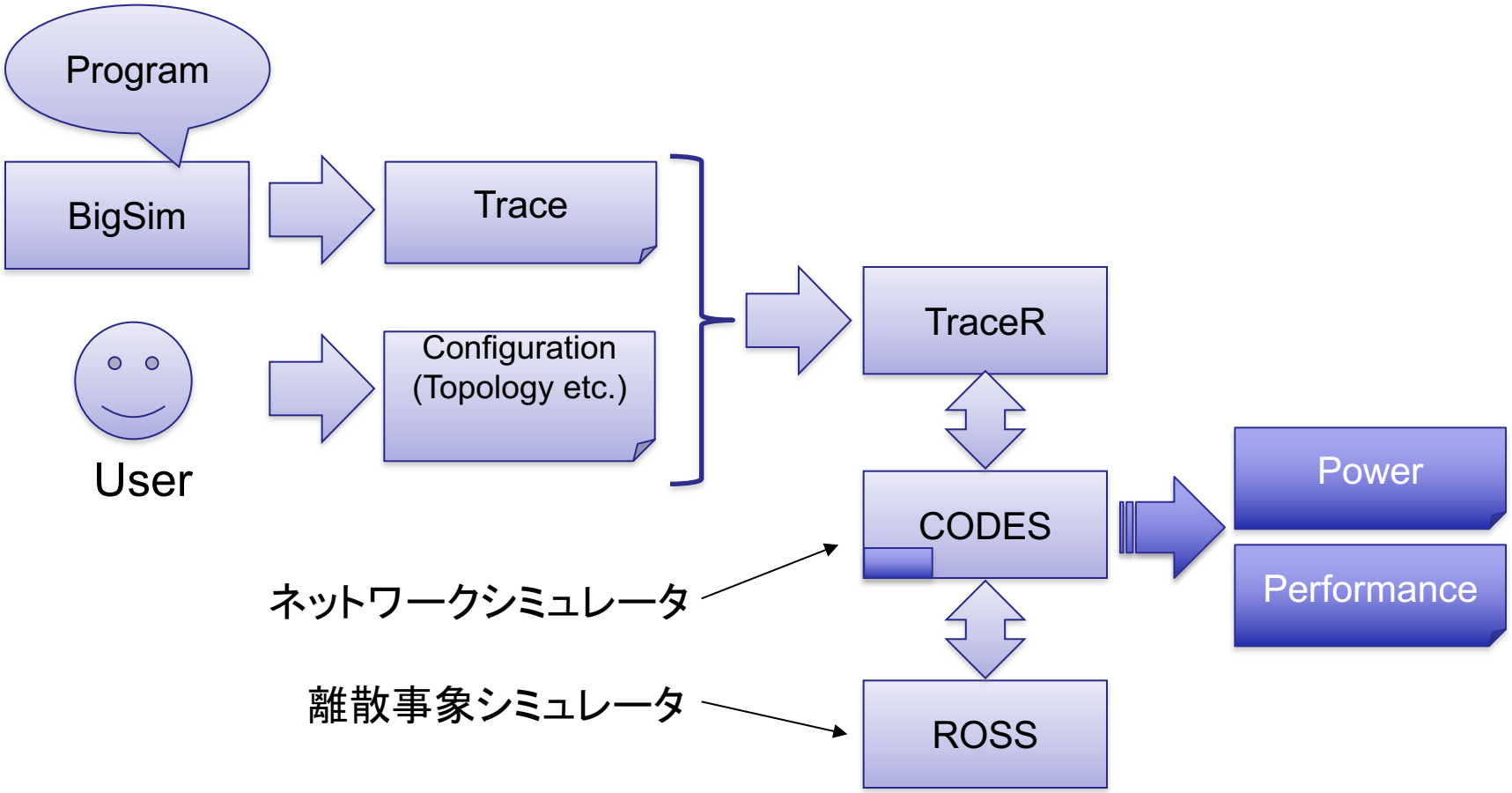
TraceR



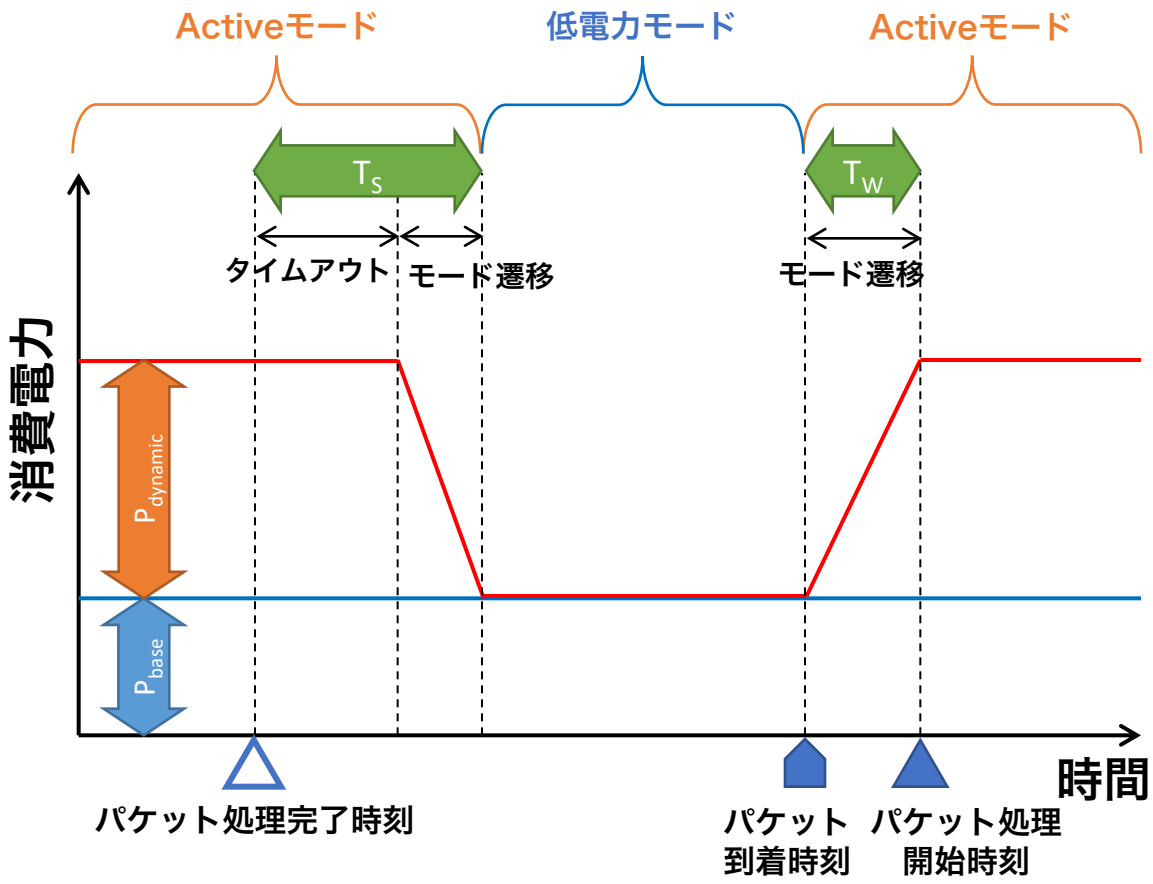
- TraceR*
  - マルチジョブやタスクマッピング機能
  - 通信トレース取得回数削減のためメッセージサイズ可変

*Jain, N., Bhatele, A., White, S., Gamblin, T. and Kale, L. V.: Evaluating HPC Networks via Simulation of Parallel Work- loads, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 14:1–14:12 (2016).

TraceRPの実装イメージ



TraceRPがサポートする低消費電力化技術



TraceRPの実装と検証

- 前提
  - モード遷移中はActiveモード時の電力を消費すると仮定
  - Up/Downリンクそれぞれで On/Offリンクを制御
  - On/Offリンク制御回路の電力は含まない
- 実装
  - CODESのソースコードを修正
  - トポロジ：Fat-tree，Torus，Dragonfly
  - TsとTwは可変
  - Offの時間を測定
  - 低電力モード時に通信が発生するとTwの遅延を再現
- 検証
  - 先行研究において提案されたモデル式と比較検証

モデルを用いた検証

- ReviriegoらのOn/Offリンクを対象とした電力モデル*

全ポートの電力　最大負荷時の電力

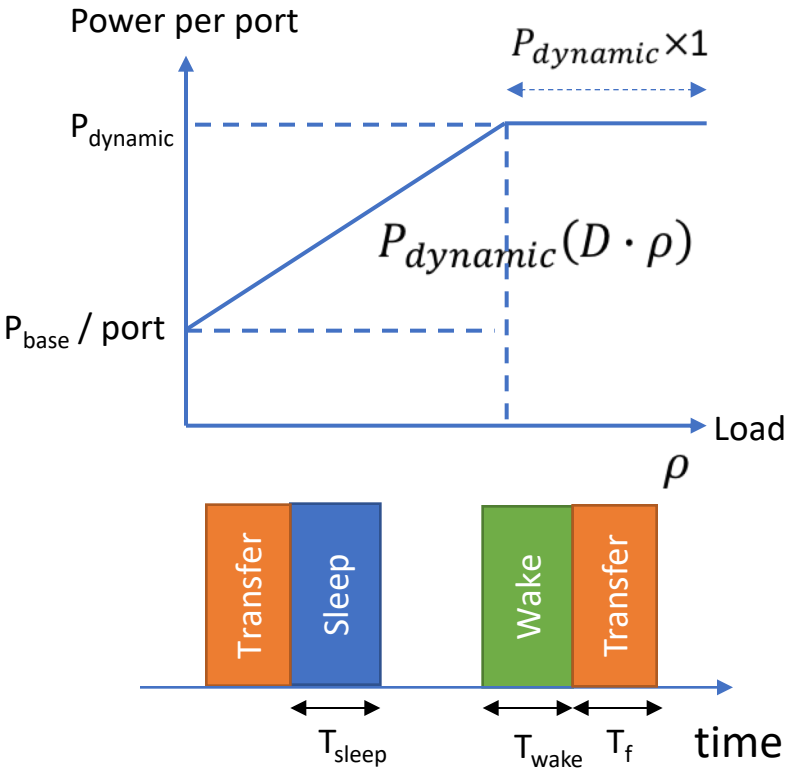
$$P = P_{base} + P_{dynamic} \sum_{i=1}^{N_{port}} \min(1, D \cdot \rho_i)$$

負荷がない場合の電力　負荷

パケット処理時間

$$D = \frac{T_f}{T_{wake} + T_{sleep} + T_f}$$

Activeモードに遷移する時間　低電力モードに遷移するまでの時間

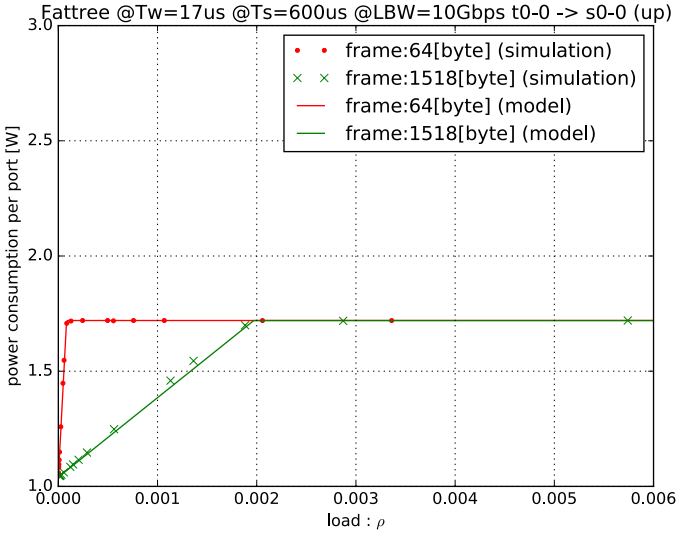


*P. Reviriego *et al.*, "An energy consumption model for Energy Efficient Ethernet switches," *2012 International Conference on High Performance Computing & Simulation (HPCS)*, Madrid, 2012, pp. 98-104.

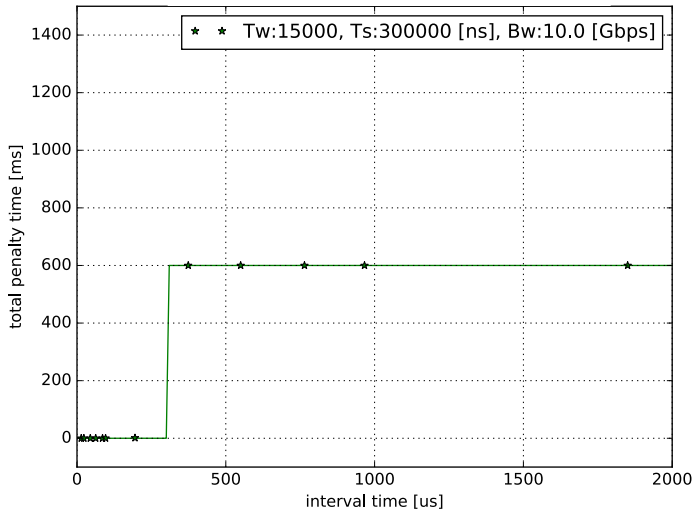
検証方法と結果

- 電力検証方法
  - 電力モデルと比較
    - ポートあたりの電力を推定
  - ネットワーク構成
    - スイッチに2つのターミナルを接続
  - 通信プログラム
    - ping-pong
- レイテンシ検証方法
  - 通信間隔とTsを変更
    - 理論増加分と比較
  - ネットワーク構成
    - スイッチに2つのターミナルを接続
  - 通信プログラム
    - ping-pong

電力の検証結果



レイテンシの検証結果



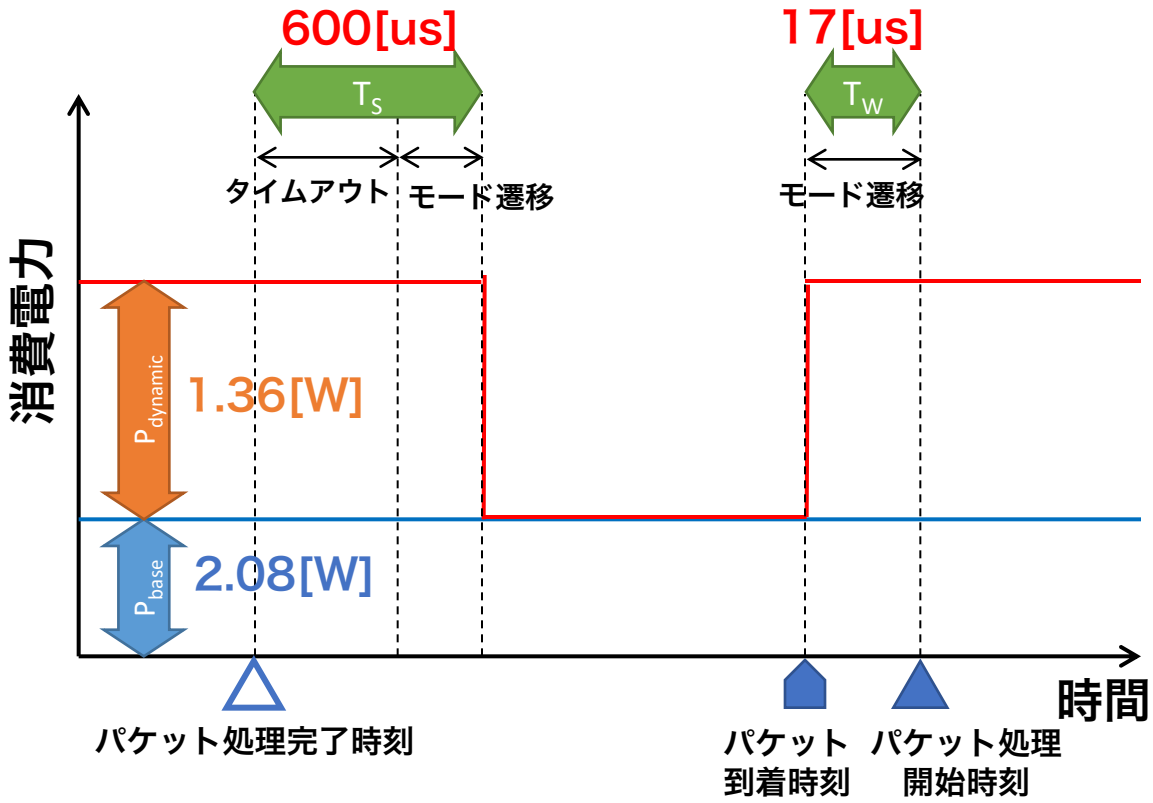
電力と性能の評価

- 想定
  - MPIプログラムを8,192プロセスで実行することを想定
- ベンチマークプログラム
  - near-neighbor：非構造型メッシュ通信プログラム
  - permutation：行列積と行列転置プログラム
  - qbox：第一原理分子動力学におけるMPI集団通信
  - stencil4d：ステンシル計算プログラム
  - a2a：All-to-All通信

評価における構成とパラメータ

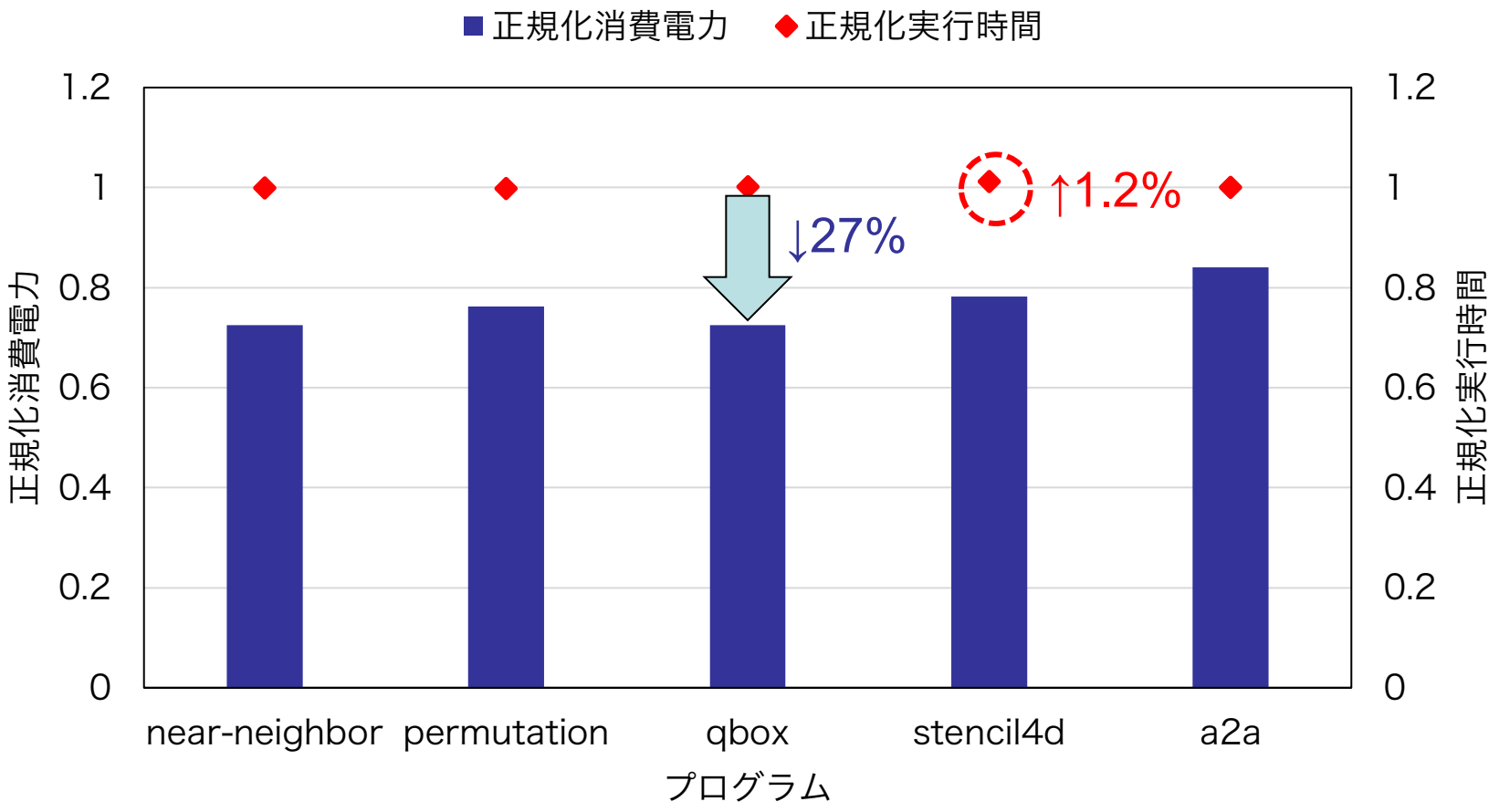
トポロジ	Fat-tree	Torus	Dragonfly
ノード数	288	256	342
エンドポイント数	32	32	32
スイッチあたりのポート数	24	6	12
スイッチあたりのノード数	12	1	3
ネットワーク構成	3-level	3D torsu 4x8x8	グループあたり6スイッチ 計19グループ

On/Offリンクのパラメータ

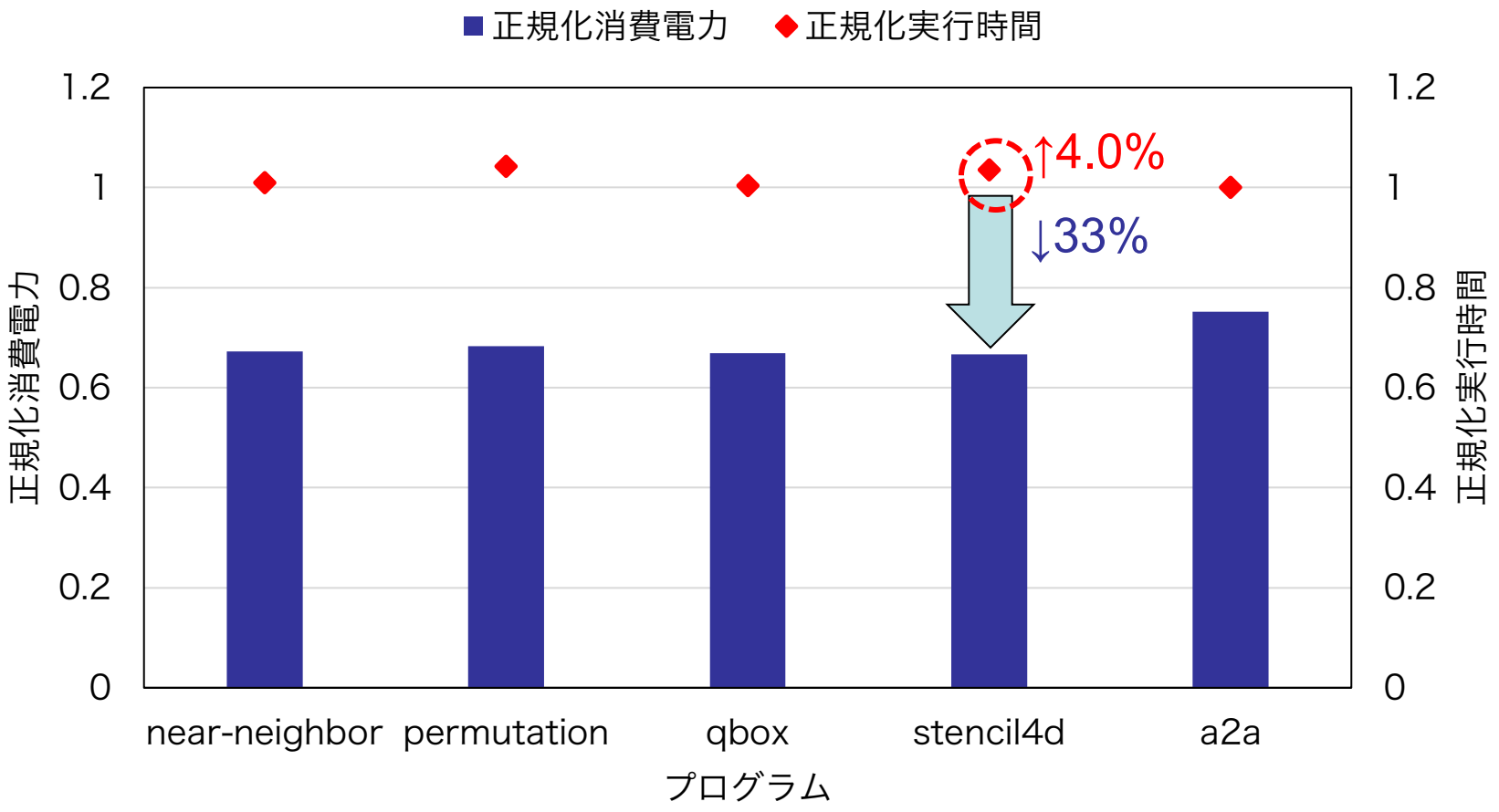


10BASE-TのEnergy Efficient Ether対応スイッチに基づき決定

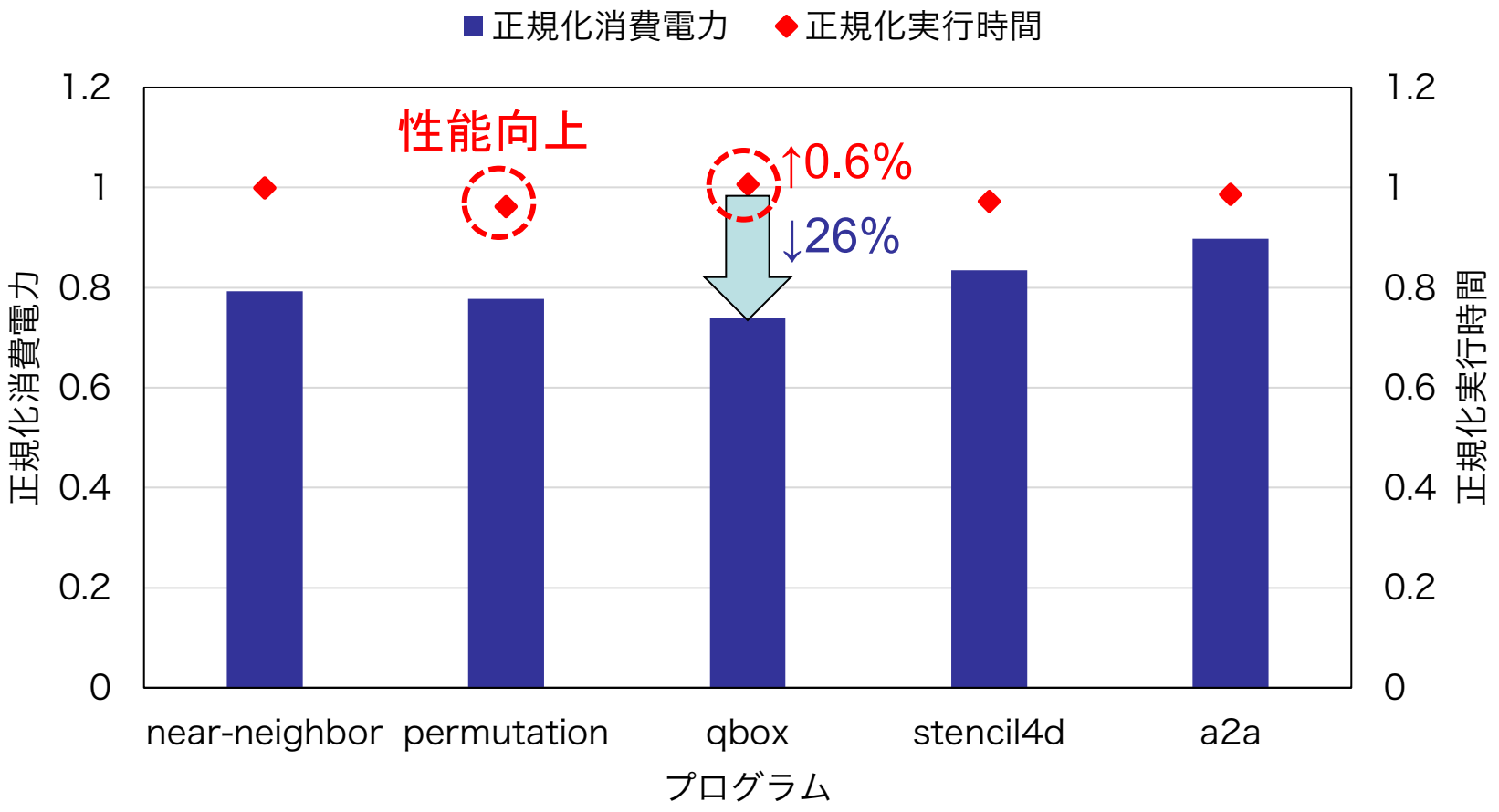
評価結果（シングルジョブ, Fat-tree）



評価結果（シングルジョブ, Torus）

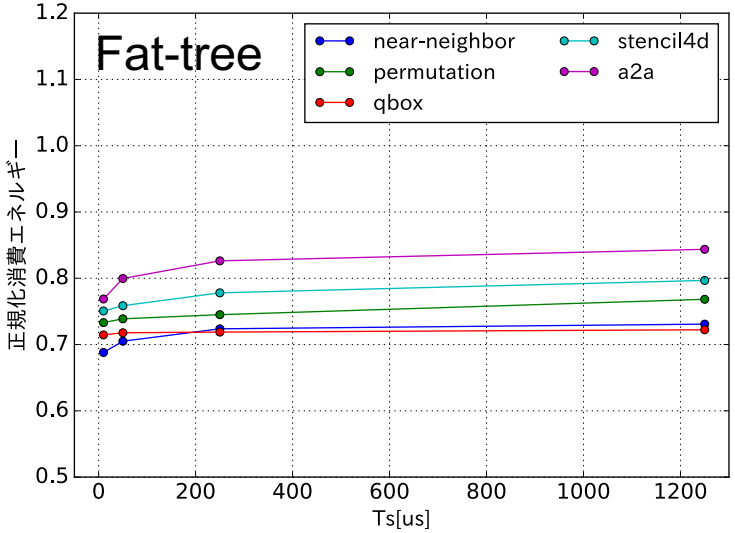


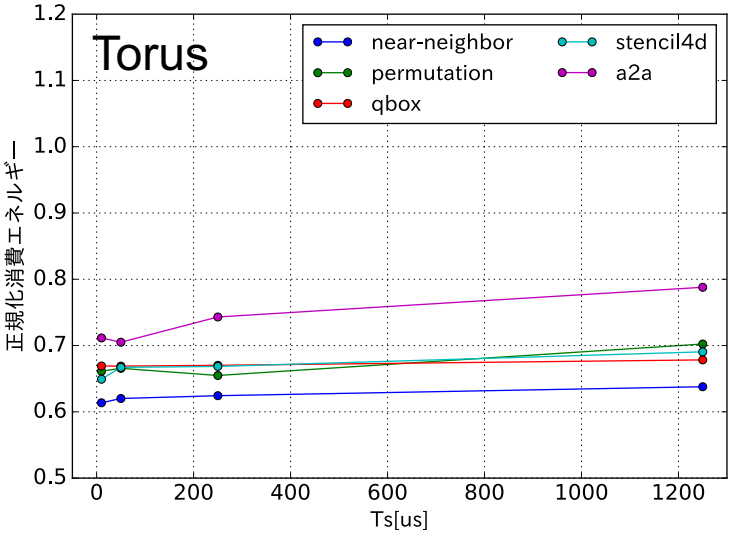
評価結果（シングルジョブ, Dragonfly）

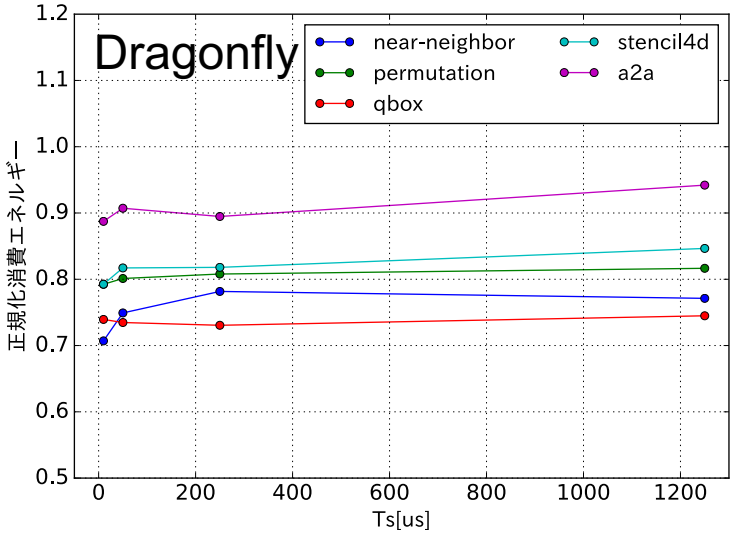


評価結果(Tsが消費エネルギーに与える影響)

- 目的
  - Tsと消費エネルギーの関係
- パラメータ
  - Tw=17
  - Ts=10, 50, 250, 1250

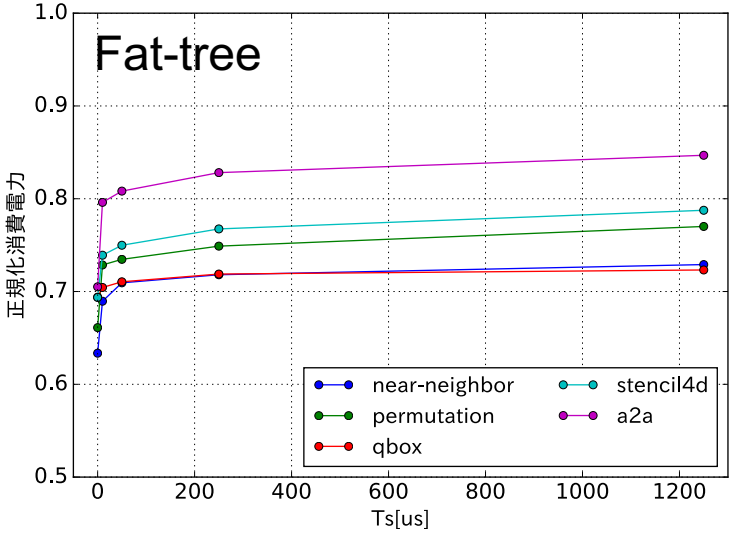


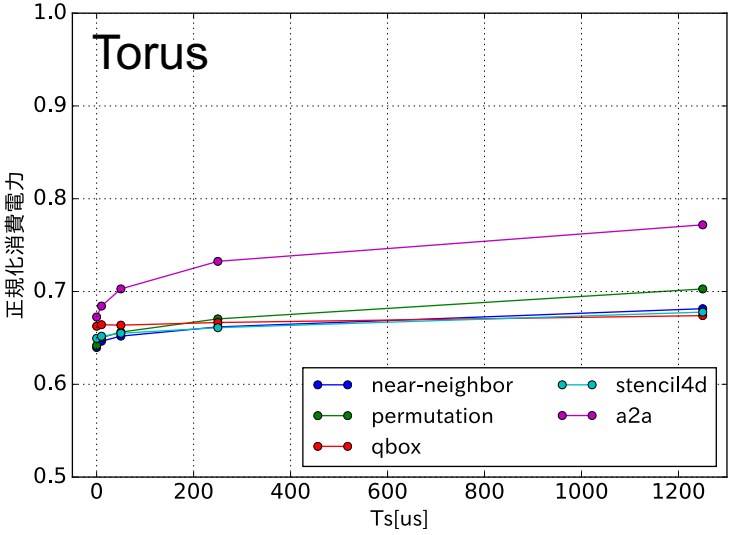


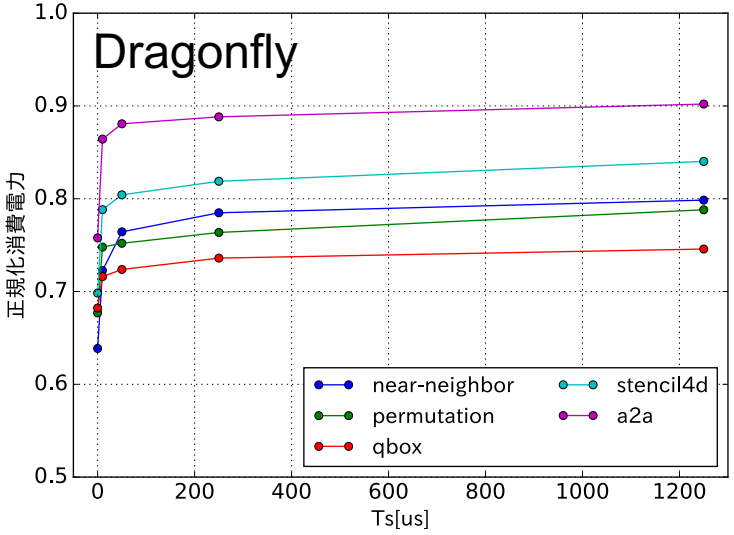


評価結果(Tsが消費電力に与える影響)

- 目的
  - Tsによる電力削減の限界を調査
- パラメータ
  - Tw=0

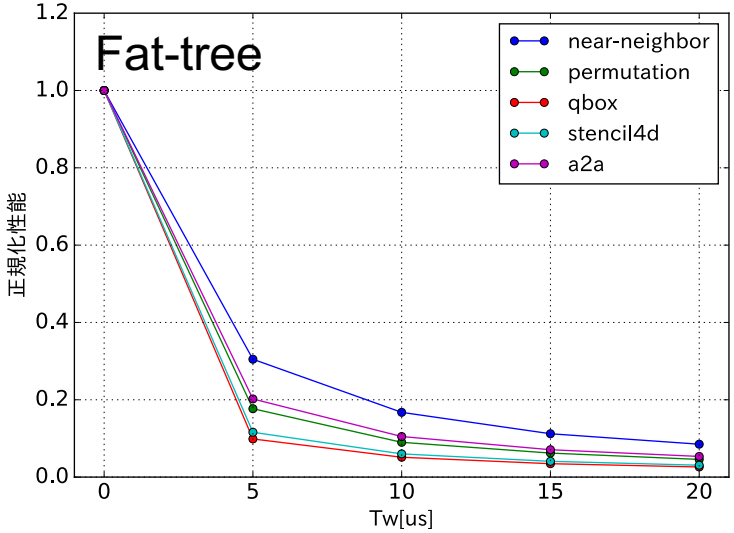


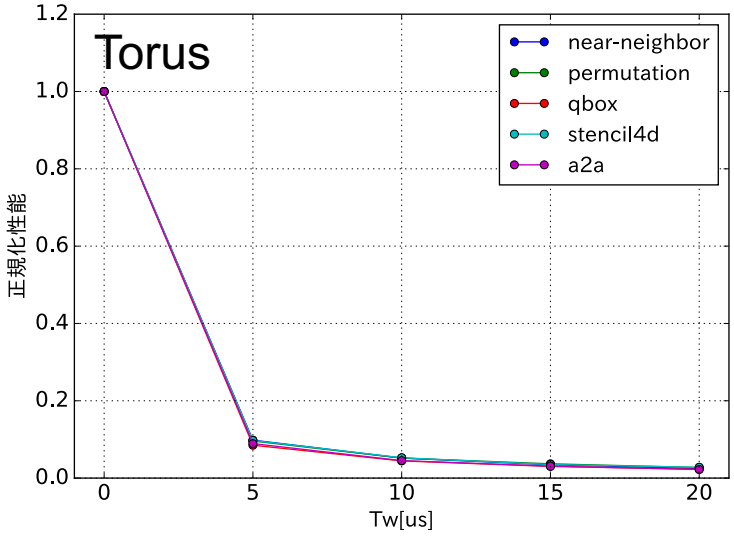


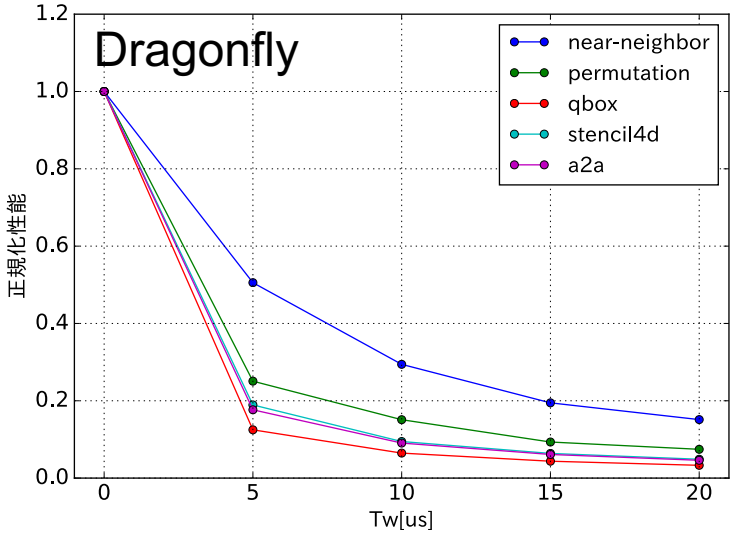


評価結果(Twが性能に与える影響)

- 目的
  - Twによる性能低下の限界を調査
- パラメータ
  - Ts=0

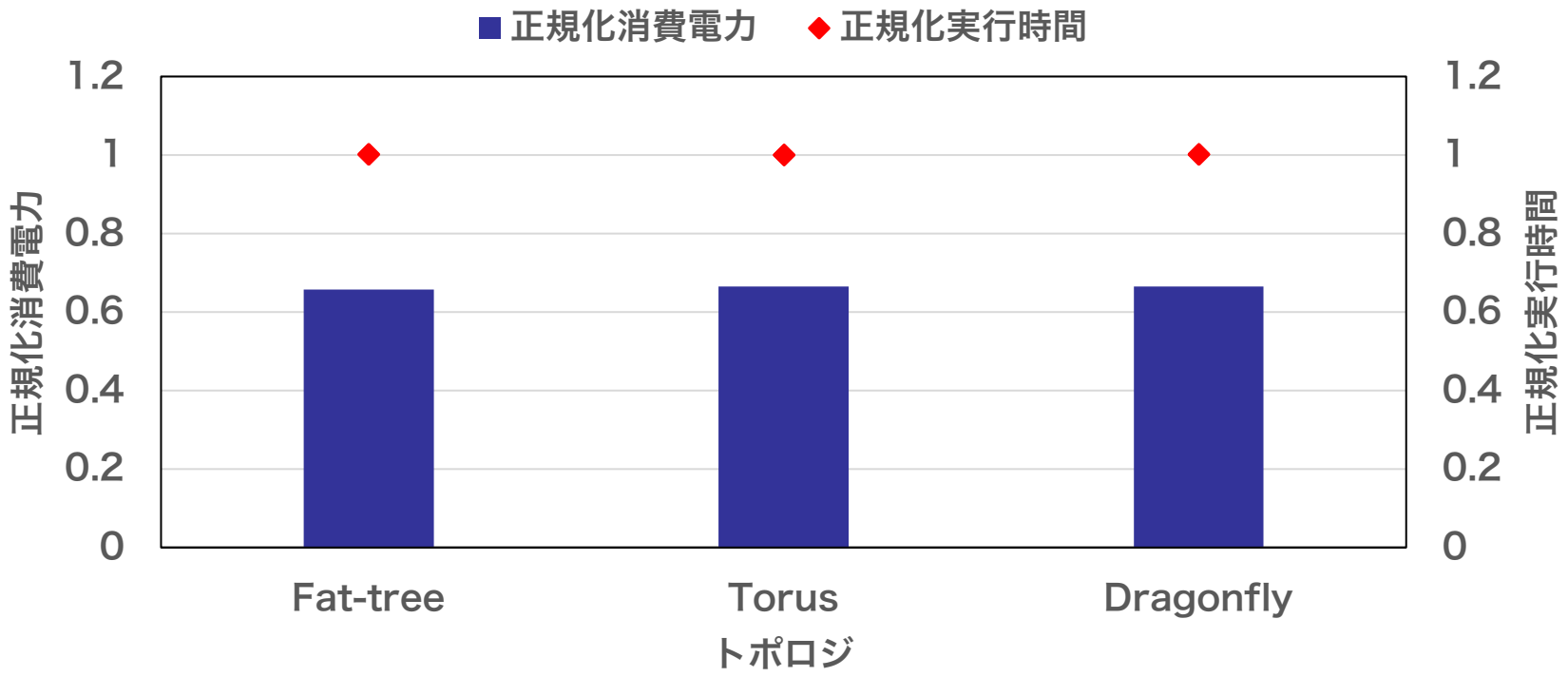






マルチジョブの評価

- 5つのプログラムを同時に実行
- 1つのプログラムあたり8,192プロセス



まとめと今後の予定

- まとめ
  - ネットワークの低電力化に対する要求
  - 電力削減技術（On/Offリンク）適用時の消費電力と性能を見積もるシミュレータが必要
  - シミュレータ（TraceRP）を開発
    - Fat-tree，Torus，Dragonflyに対応
    - マルチジョブに対応
    - On/Offリンクの主要パラメータを変更可能
- 今後の予定
  - ネットワーク規模を拡大して実験
  - オープンソースとしてシミュレータを公開