HOKUSAI GreatWaveシステムのベンチマーク

理化学研究所 情報基盤センター
2015/6/19 和光

理研シンポジウム2015

Outline

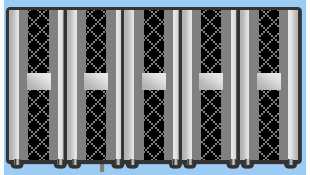
- HOKUSAI GreatWave systemの概要
  - CPU, memory, network
- ベンチマーク
  - 姫野ベンチマーク
  - NAS Parallel Benchmarks
  - IMB
  - 量子化学計算(SMASH)

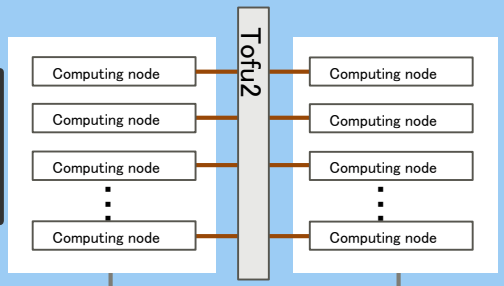
Outline of HOKUSAI GreatWave (GW) system

Computing system

Massively Parallel Computer (MPC)

FX100 (Fujitsu, 1,080 nodes)
Total peak performance: 1.0PFLOPS (1TFLOPS/node)
Total memory: 34TB (32GB/node)





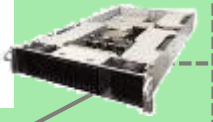
Application Computing Server (ACS)

ACS with Large memory (ACSL)

2 nodes
CPU: Intel Xeon *4 /node
Peak Performance: 1.2TFLOPS/node
Memory: 1TB/node

ACS with GPU (ACSG)

30 nodes
CPU: Intel Xeon *2 /node
Peak performance: 0.88TFLOPS/node
Memory: 64GB/node



With GPGPU x4 / node

High speed network (InfiniBand FDR)

Ethernet Network

Front end servers
Login node x4

MDS | MDS | OSS | OSS

MDT | OST

Online Storage
Actual capacity: 2.2PB

File Server | File Server | Tape Control | Tape Control

Primary Storage | Secondary Storage

Hierarchical Storage
Tape Capacity: 7.9PB (uncompressed)

RIKEN Network

Users

GreatWave Massively Parallel Computer (GW-MPC)

- CPU: SPARC™XIfx (1.975Hz)
  - 1,080 nodes, 32 cores/node, 34,560 cores
  (1ノードに、2つのコアメモリグループ(CMG))
- Peak performance
  - 1.092 PFLOPS (RICC: 0.0982 PFLOPS)
    - 1.975 Hz x 16演算 x 32 cores x 1,080
  - 1.011 TFLOPS/node (RICC: 0.0937 TFLOPS/node)
- メモリ: 32GB/node
  - BW: 480 GB/s/node (RICC: 51.16 GB/s/node)
  - B/F: 0.47 B/FLOP (RICC: 0.54 B/FLOPS)
- 通信: 6次元メッシュ/トーラス
  - FDR InfiniBand
  - 12.5GB/s x 双方向 (ノード間) (RICC: 2 GB/s)

計測したベンチマーク

- 現状でどういう状態かを確認することを目的
  - チューニングなどは行わずに測定
- 姫野ベンチマーク
  - ポアソン方程式をヤコビの反復法で解く
  - 主にメモリバンド幅の性能に依存
- NAS Parallel Benchmarks (NPB)
  - NASA Ames Research Centerで開発された航空宇宙関連のシミュレーションのコアルーティンから抜き出されたベンチマーク
  - 科学技術計算に必要なさまざまな性能を評価
- IMB
  - ネットワークの通信性能を測定
- 量子化学計算(SMASH)のベンチマーク
- 他の発表者によるベンチマーク
  - 安藤さん：分子動力学(GENESIS)
  - 北澤さん：京チェックスイート

姫野ベンチマーク

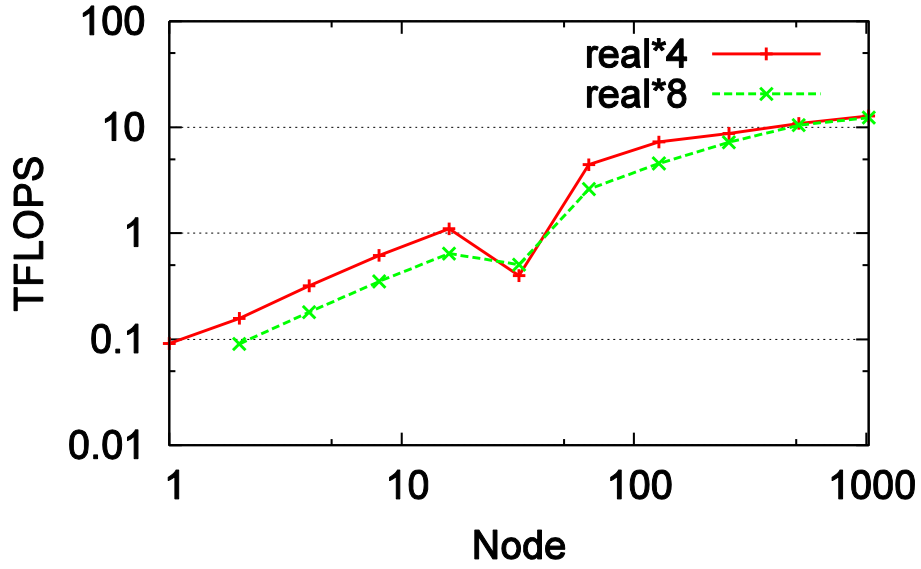
- 測定条件
  - 行列サイズはXL(1024x512x512)を用いる
  - フラットMPI、MPI+openMP(-Kopenmp)、MPI+自動並列化(-Kparallel)
  - 最適化オプション : -Kfast
  - 並列数とグリッドの分割

1(1x1x1), 2(1x1x2), 4(1x2x2), 8(2x2x2), 16(2x2x4), 32(2x4x4), 64(4x4x4), 128(4x4x8), 256(4x8x8), 512(8x8x8), 1,024(8x8x16), 2,048(8x16x16), 4,096(16x16x16), 8,192(32x16x16), 16,384(32x32x16), 32,768(32x64x16)

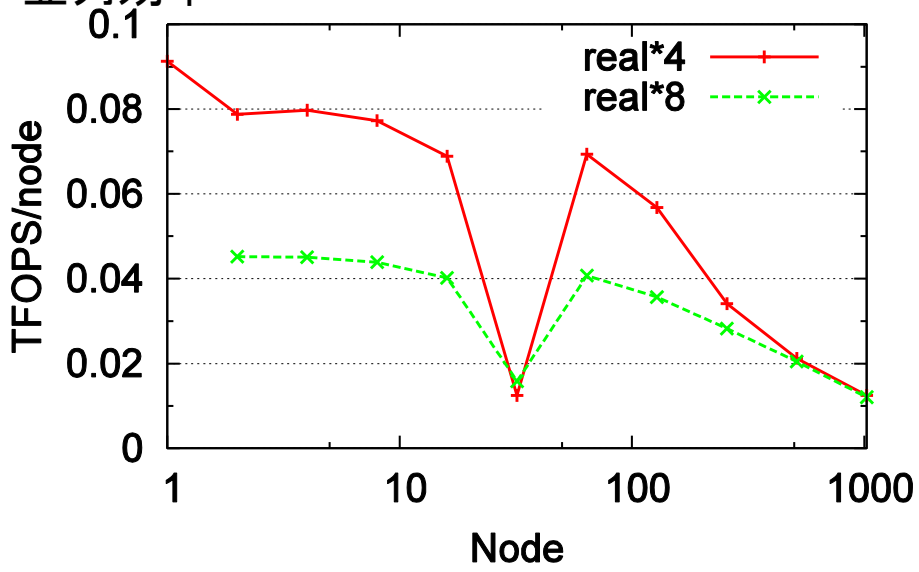
姫野ベンチマークの結果1
フラットMPI並列

- ノード間もノード内の32コアもMPI並列
- MPIプロセス数[nodex32]

スケーラビリティ



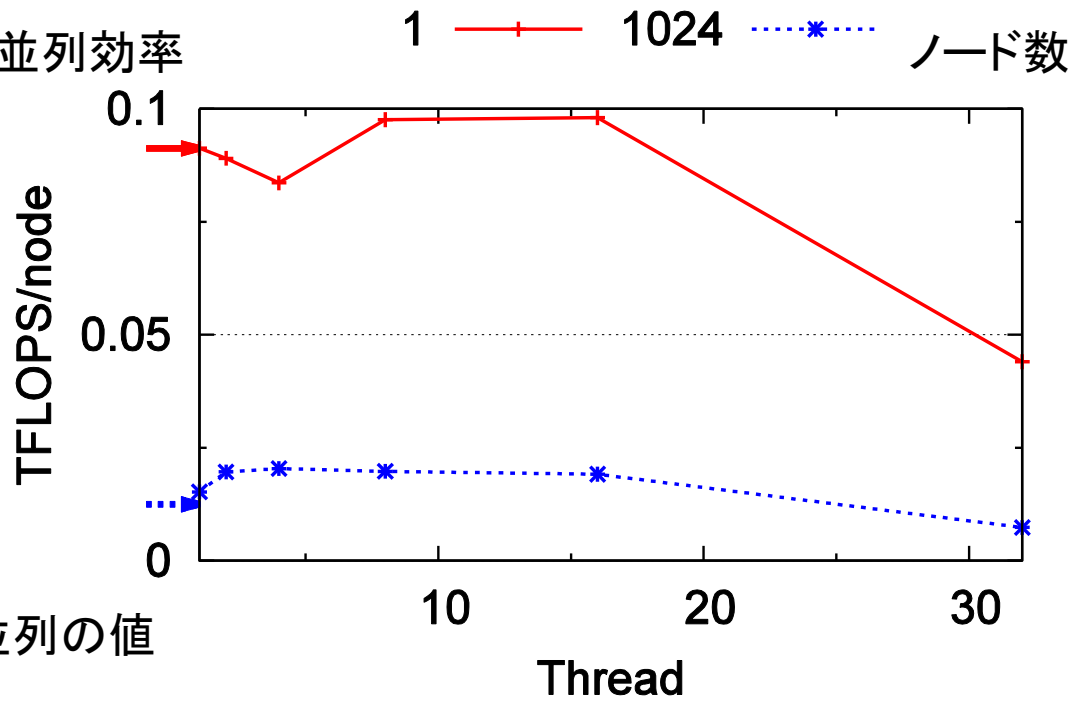
並列効率



1ノードで理論性能の9%程度で、バンド幅からすると少し低い
32ノードでは性能が大きく落ちるが、メモリ関連の問題か?
倍精度は1ノードでは半分の性能で、1,024ノードでは同程度

姫野ベンチマークの結果2
スレッド並列数による性能の影響(MPI+openMP)

- ノード数（コア数）固定でスレッド並列数を変えて測定
- MPIプロセス数[nodex(32/thread)]x openMPスレッド数[thread]



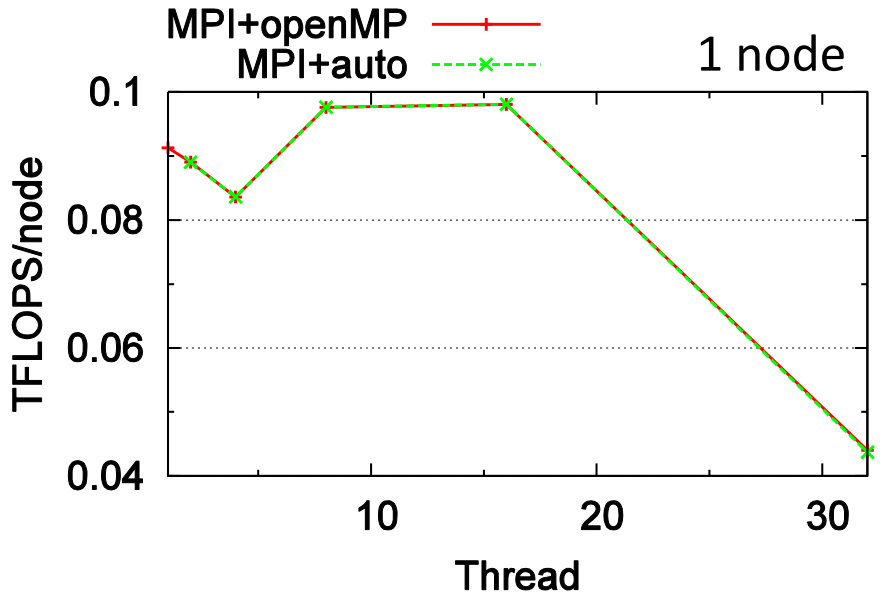
矢印はフラット並列の値

32スレッド並列は性能が相当落ちる

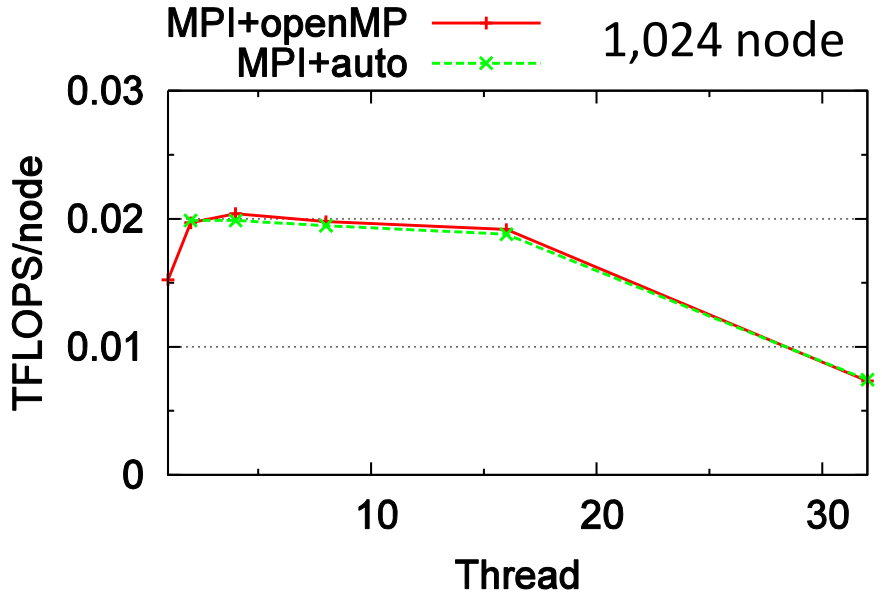
姫野ベンチマークの結果3
MPI+openMPとMPI+自動並列の比較

- ノード数（コア数）固定でスレッド並列数を変えて測定
- MPI+openMPとMPI+自動並列で比較

並列効率



並列効率



自動並列化でもopenMP並列化とほぼ同じ性能

NAS Parallel Benchmark (NPB)

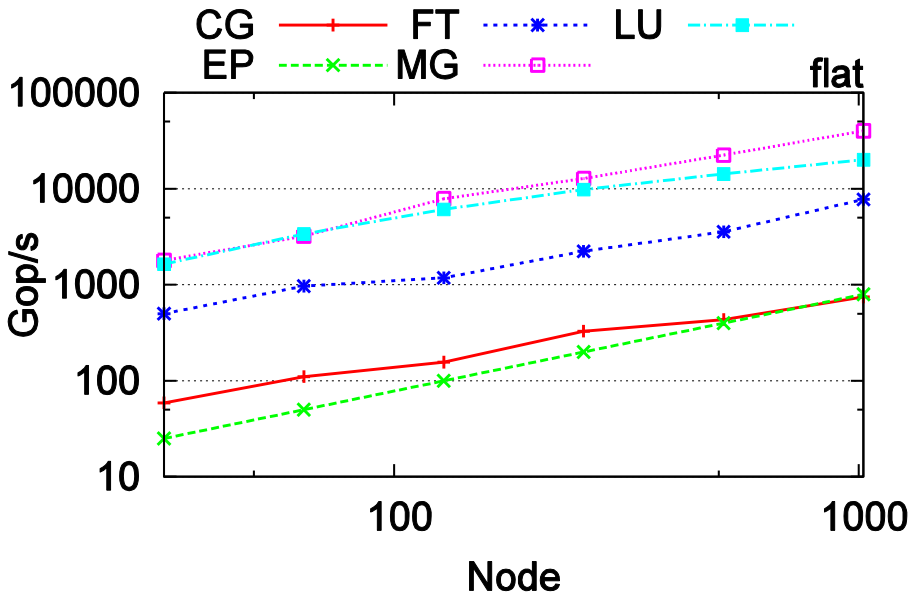
- カーネルベンチマーク
  - CG：正値対称大規模疎行列の最小固有値を共役勾配法で求める
    - 通信量が多い
  - EP：乗算合同法による一様、正規乱数を生成
    - 通信がほとんど発生しない
  - FT：FFTを用いた3次元偏微分方程式の解法
    - 大域通信性能の測定
  - IS：大規模整数ソート
  - MG：マルチグリッド法のカーネル
    - 階層的な隣接通信
- アプリケーションベンチマーク
  - BT：ブロック3重対角方程式をADI法で解く
  - SP：5重対角方程式をスカラーADI法で解く
  - LU：上下三角行列を対称SOR法で解く
    - 純粋な隣接通信でMGに比べて通信が多い

すべてCLASS Eで測定
CG,EP,FT,MG,LUについて
性能評価

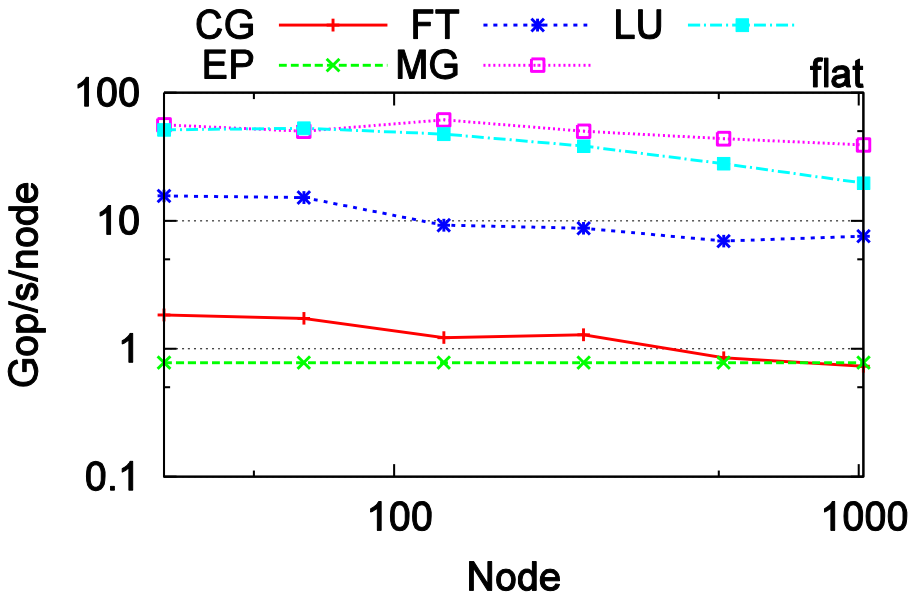
NPB 1
フラットMPI並列

MPIプロセス数[nodex32]

スケーラビリティ



並列効率



- EPは予想通り並列性能が高い
- MGも並列性能が高く、LUは100ノードを超えると少し落ちてくる
- FTは100ノード前後で少し落ちるが、それ以上はあまり落ちない
- CGは並列性能がよくない

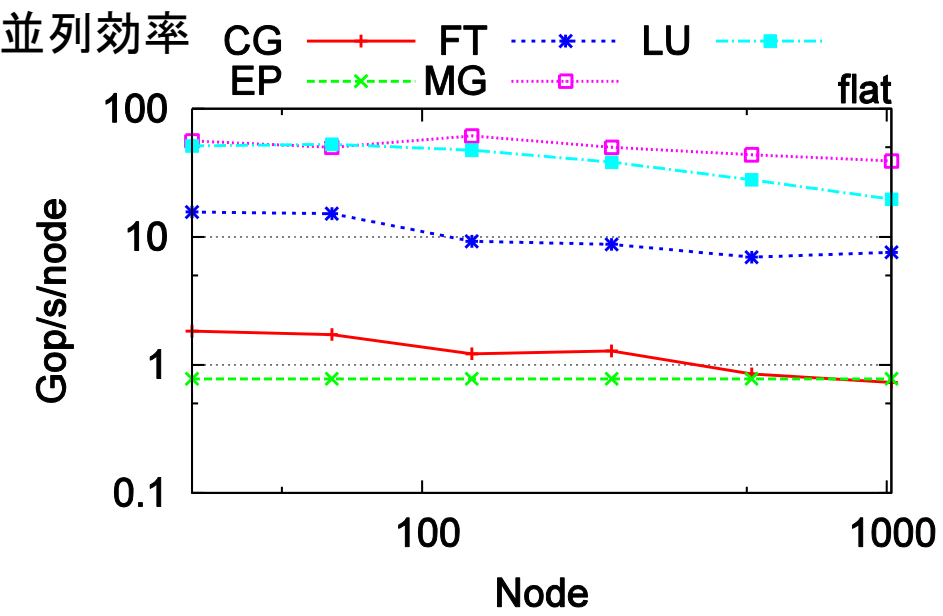
大規模並列でも優れた並列性能が得られた

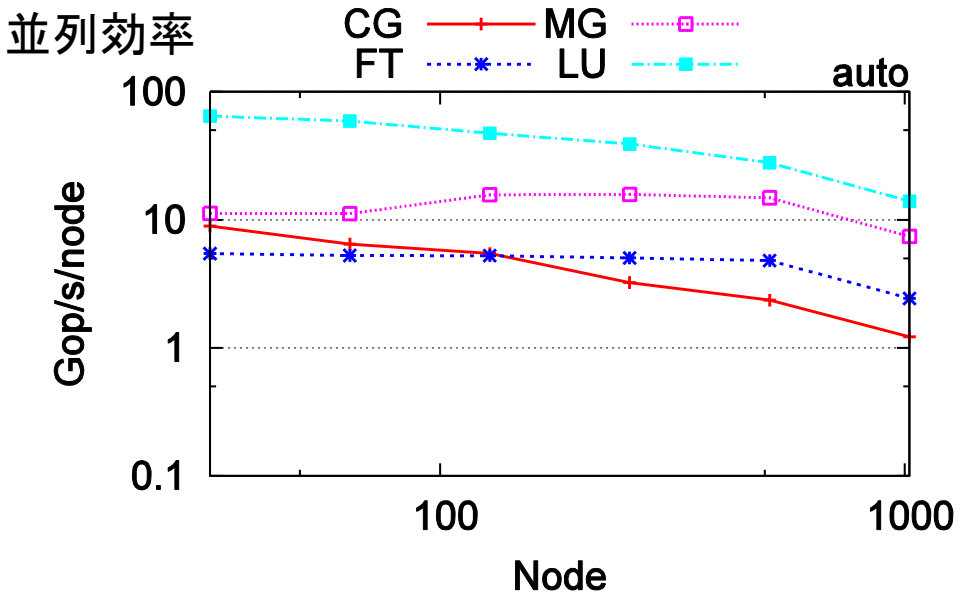
NPB 2
MPI+自動並列(16スレッド並列)による性能の影響

MPIプロセス数[nodex32]

MPIプロセス数[nodex2]
+自動並列スレッド数[16]

EPは性能が大きく落ちるので除外





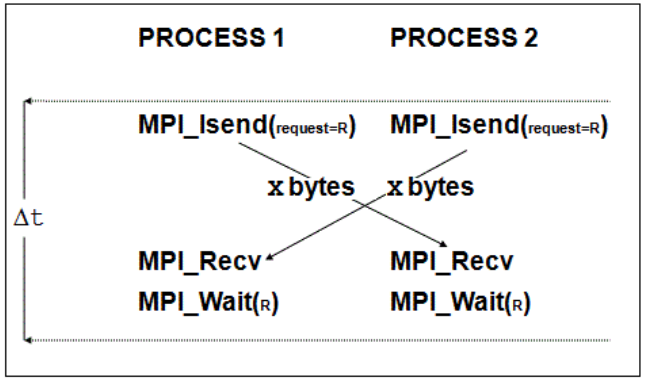
自動並列化をすることによって、
フラットMPIより速くなることもある

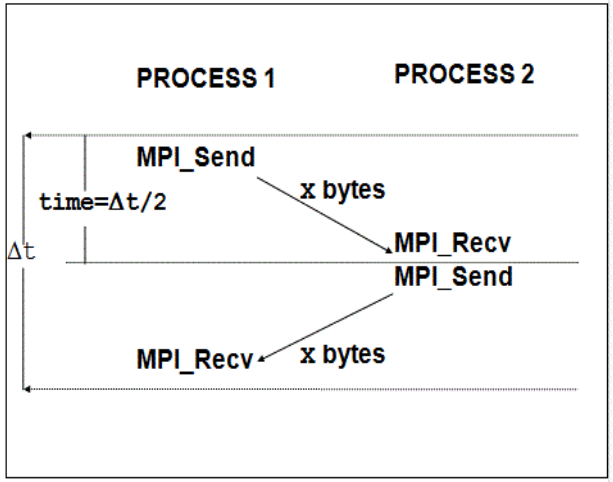
Intel MPI Benchmark (IMB)

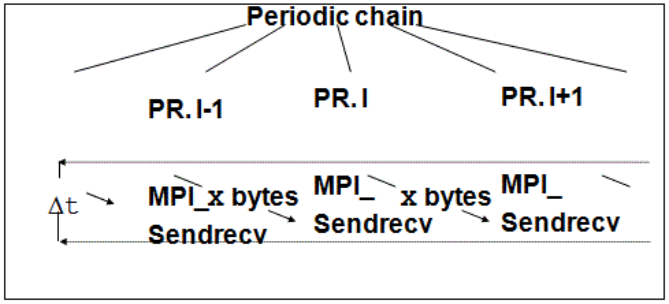
- IMBとは、Intel が提供しているMPI通信に係る性能測定のベンチマーク集
  - https://software.intel.com/en-us/articles/intel-mpi-benchmarks
- 今回は帯域など絶対評価が可能なMPI-1のベンチマークについて測定
  - PingPong, PingPing, Barrier, Exchange, SendRecv, Multi-PingPing
  - PingPong, PingPingは2ノード利用
  - Barrier, Exchange, SendRecv,Multi-PingPingは1024ノード利用

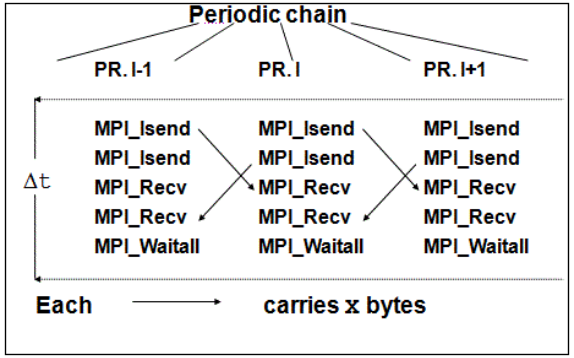
ベンチマークの概要

- PingPing
- PingPong
- SendRecV
- Exchange
- Multi-PingPing







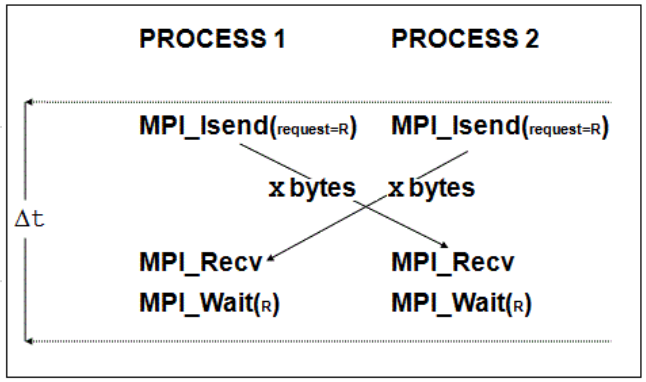


Intel MPI Benchmarks User Guide and Methodology Description より

バリアの測定時間と ハードウェアバリア機能の障害

- 1,024ノードでのバリアの測定時間
  - 3月18日(ハードウェアバリア機能あり)
    - 25.51us
  - 4月28日(ハードウェアバリア機能なし)
    - 134.78 us
    - 今回のベンチマークはこの環境で測定

PingPing



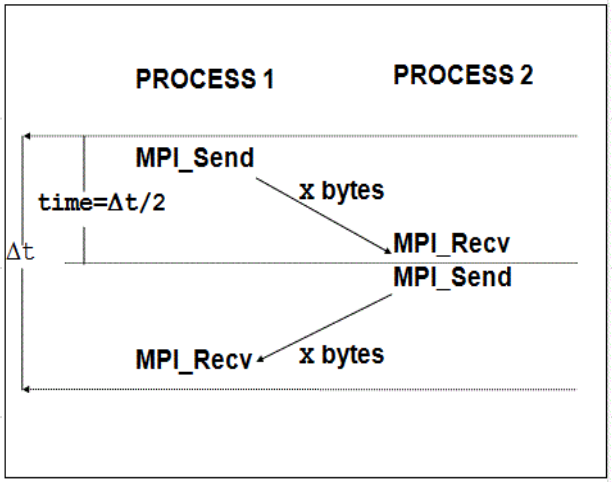
10GB/s

80%程度の性能
(理論性能は12.5GB/s)

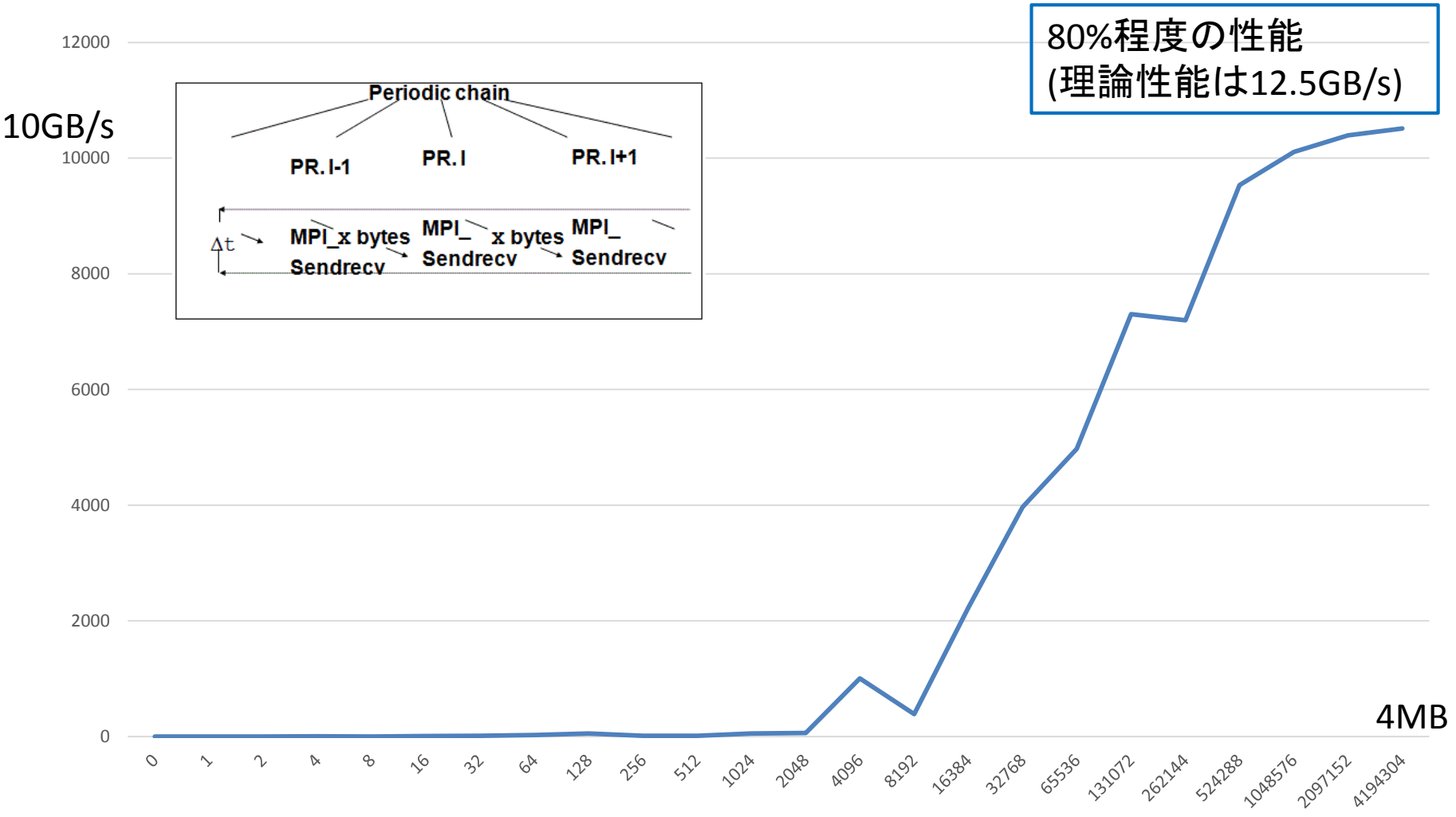
PingPong

10GB/s

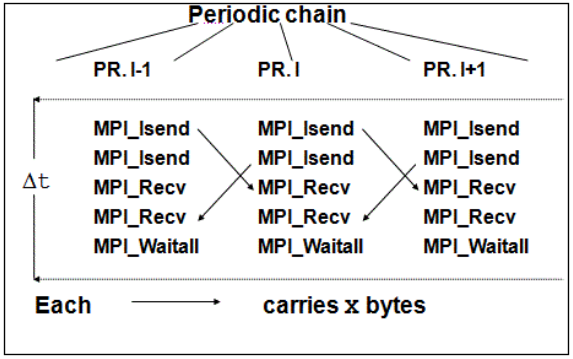
80%程度の性能
(理論性能は12.5GB/s)



SendRecv


80%程度の性能
(理論性能は12.5GB/s)
10GB/s
12000
10000
8000
6000
4000
2000
0
Periodic chain
PR. I-1
PR. I
PR. I+1
Δt
MPI_x bytes
Sendrecv
MPI_ x bytes
Sendrecv
MPI_
Sendrecv
4MB
0
1
2
4
8
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
131072
262144
524288
1048576
2097152
4194304

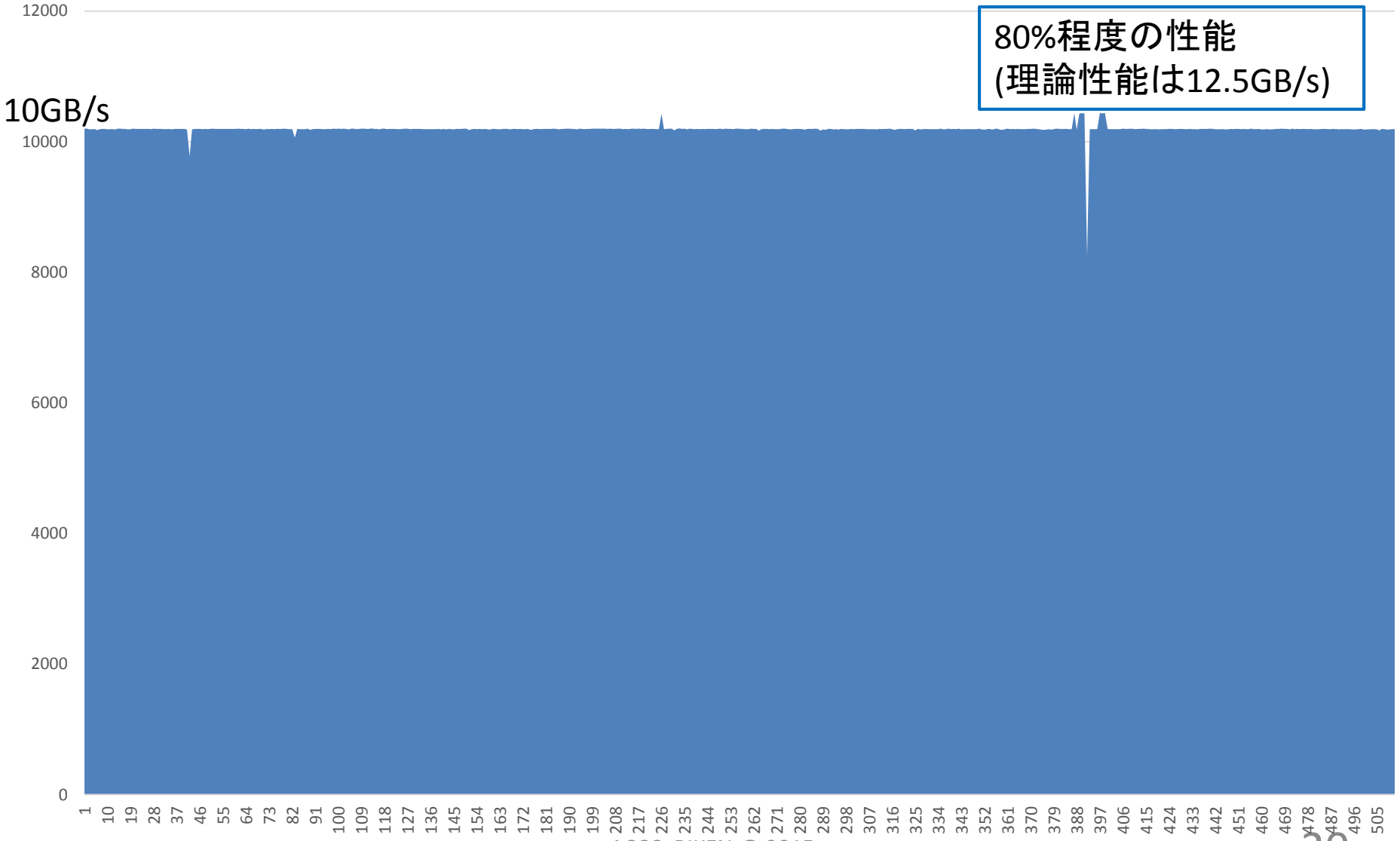
Exchange



80%程度の性能
(理論性能は12.5GB/s)

Multi-PingPing

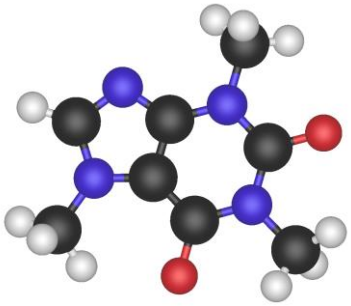
4MBでの測定



量子化学計算のベンチマーク

- 量子化学計算はHOKUSAIでもpopular
  - 但し、スケーリングの良い量子化学計算プログラムは少ない。
  - FX100向けにチューニングされているとなると皆無
- 分子研の石村和也氏作SMASHを用いたベンチマーク
  - http://sourceforge.net/projects/smash-qc/
  - 概要：http://www.slideshare.net/NakataMaho/hpcs2015
  - DFTなどの計算を行う第一原理量子化学計算プログラム
  - よいスケーリング
    - 一万コアでもスケール、構造最適化もスケール
    - SIMD化は難しい
  - 数百原子くらいまで計算可能

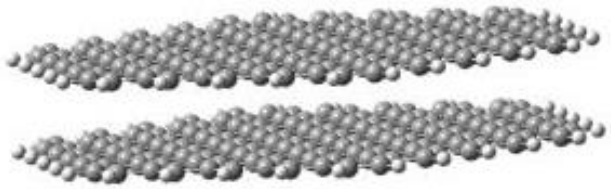
量子化学計算とは?



- 分子を第一原理的に解く
  - HΨ=EΨ
- 必要なパラメータ
  - 核の座標、核の電荷、電子数
  - 近似法(I) : 平均場近似～配置間相互作用(FullCI, Exact Diagonalization)まで様々
  - 近似法(II) : 基底関数
- よく聞かれる Jargon
  - SCF(Hartree-Fock or 平均場), DFT (密度汎関数法)
  - 分子軌道法(MO), Kohn-Sham方程式 (DFT)
  - B3LYP汎関数 (DFT)

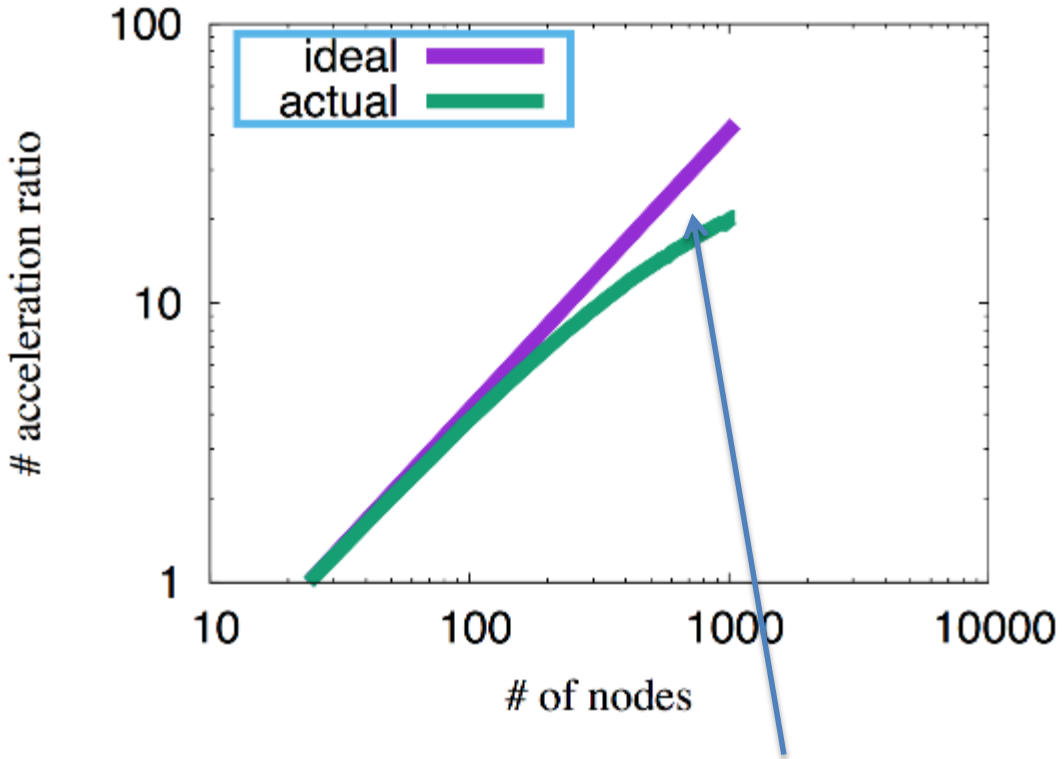
ベンチマーク結果

1056ノードで、24ノードの20倍(45%)の効率



計算機: HOKUSAI (MPC) Fujitsu FX100
分子 : $(C_{150}\ H_{30})_2$ 360原子
基底関数: cc-pVDZ (4500基底)
計算方法 : B3LYP
SCFサイクル数 :16
並列ノード内:OpenMP ノード間:MPI
オプション: -Kfast
京での結果:石村氏による

京より1.58倍高速
(両方とも12288コアでの比較)



対角化はN^3だが並列効率が悪い

対角化を除くと**効率は90%**を超える

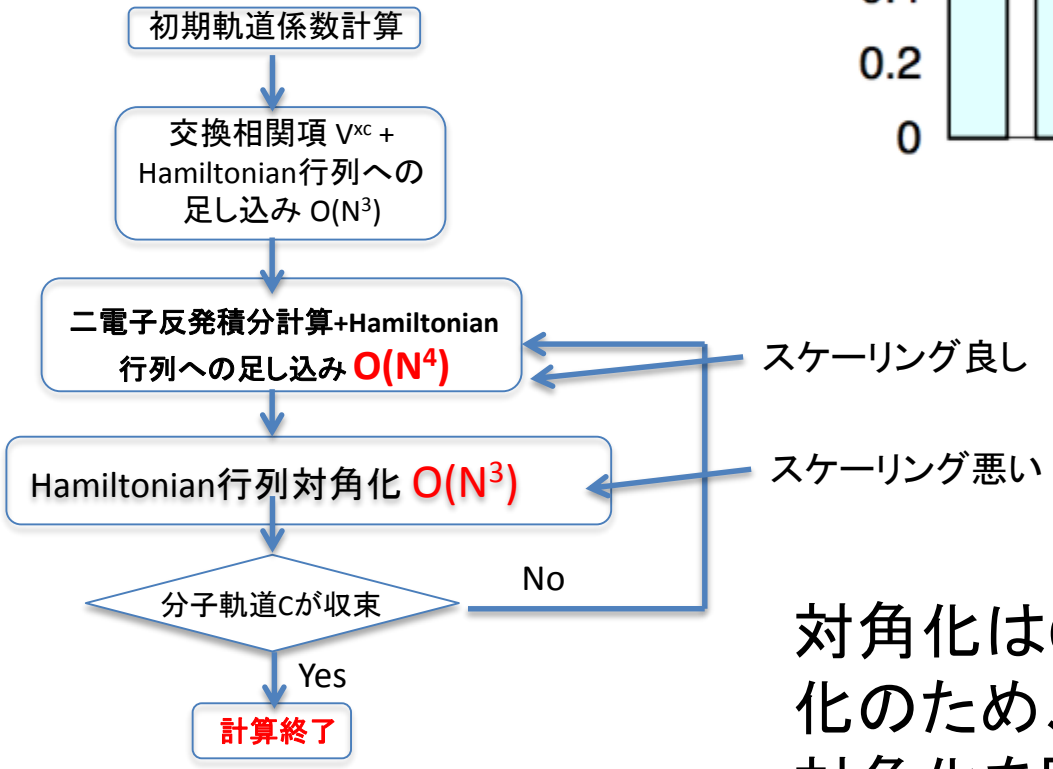
ベンチマーク結果:詳細

$$FC = \epsilon SC$$

$$F_{\mu\nu} = H_{\mu\nu} + V_{\mu\nu}^{XC} + d\sum_{i,\lambda,\sigma} 2C_{\lambda i}C_{\sigma i}\{2(\mu\nu|\lambda\sigma) - (\mu\lambda|\nu\sigma)\}$$

2電子反発積分

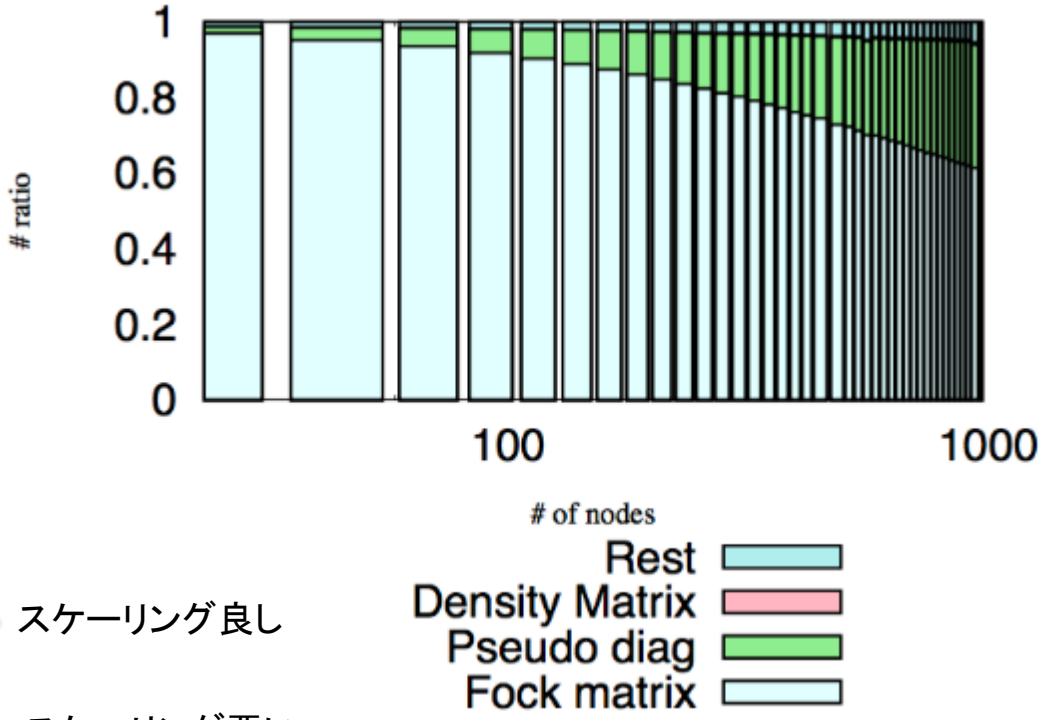
$$(\mu\nu|\lambda\sigma) = \int dr_1 \int dr_2 \phi_\mu(r_1)\phi_\nu(r_1)\frac{1}{r_{12}}\phi_{\lambda(}r_2)\phi_\sigma(r_2)$$



スケーリング良し

スケーリング悪い

※収束しないこともある



対角化はO(N3)だが小さい行列の対角化のため、並列効率が悪い
対角化を除くと効率は90%を超える

まとめ

- 姫野ベンチ
  - メモリバンド幅に対して少し落ちる性能
  - 32スレッド並列は性能が大幅に落ちる
  - 最適なスレッド並列数は、2,4,8,16で状況によって変わる
- NAS Parallel Benchmarks
  - 1000ノード程度の大規模並列でも優れた並列性能
  - 自動並列を試してみる価値がある
- IMB
  - 理論性能の8割程度の通信性能
- 量子化学計算(SMASH)
  - 1056ノードで、24ノードの20倍(45%)の効率
  - 同じコア数で京より1.58倍高速