



High Performance and Productivity  
NEXXUS 4820 ベンチマーク

スケーラブルシステムズ株式会社




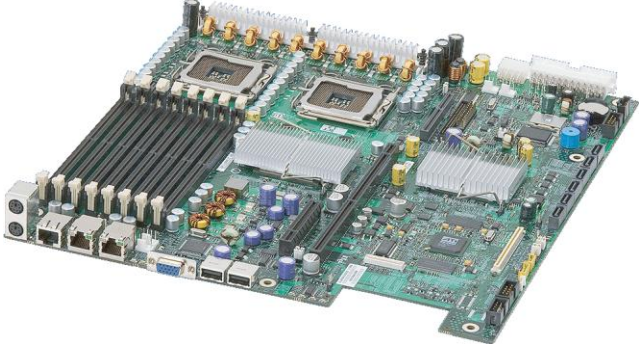
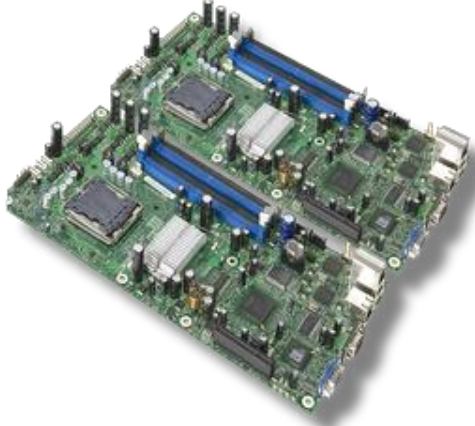
# NEXXUS 4000ベンチマーク

- 2つのタイプのチップセットが選択可能
  - NEXXUS 4820AL
    - デュアルソケット / SE5000AL
  - NEXXUS 4820PT
    - シングルソケット / S3000PT
- MPIベースのアプリケーション
  - シングルソケット + InfiniBand構成での性能の確認
  - 複数のMPI実装での性能の確認




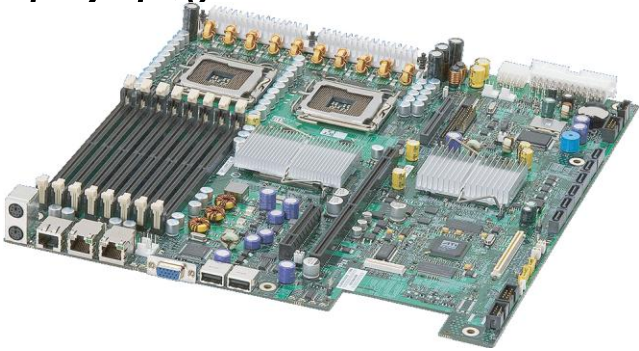


# NEXXUS 4000 パーソナルクラスタ

モデルタイプ	NEXXUS 4820 <sup>®</sup> AL	NEXXUS 4820 <sup>®</sup> PT
チップセット	SE5000AL	S3000PT
プロセッサ	Intel <sup>®</sup> Xeon <sup>®</sup> 5100/5300	Intel <sup>®</sup> Xeon <sup>®</sup> 3000/3200 Intel <sup>®</sup> Core <sup>™</sup> Duo/Quad
 <p>プロセッサとチップセットの選択が可能なブレード構成</p>	<p>1333MHzのデュアルソケット、独立バスアーキテクチャ</p> 	<p>シングルソケットHPC向けアーキテクチャ</p> 
オペレーティングシステム	Microsoft Windows Compute Cluster Server 2003、RedHat、SuSE などのLinuxディストリビューション	



# NEXXUS 4000 パーソナルクラスタ

モデルタイプ	NEXXUS 4820 <sup>®</sup> AL	<ul style="list-style-type: none"><li>• 1333MHz 独立したデュアルバスアーキテクチャ</li><li>• インテルが提供する様々な付加機能 (iAMTやVT、メモリミラーなど) を利用可能</li><li>• 通常のサーバと同じ機能、性能をさらに高いコストパフォーマンスで提供</li><li>• 3.0GHzまでのプロセッサをサポート</li><li>• 静穏性を高める低電圧 Xeon 5148も搭載可能</li></ul>
チップセット	SE5000AL	
プロセッサ	Intel <sup>®</sup> Xeon <sup>®</sup> 5100/5300	
 <p>プロセッサとチップセットの選択が可能なブレード構成</p>	1333MHzのデュアルソケット、独立バスアーキテクチャ 	
オペレーティングシステム	Microsoft Windows Compute Cluster Server 2003、RedHat、SuSE などのLinuxディストリビューション	



# NEXXUS 4000 パーソナルクラスタ

## モデルタイプ

チップセット

プロセッサ



プロセッサとチップセットの選択が可能なブレード構成

オペレーティングシステム

- NEXXUS 4820ALよりも更に高いコストパフォーマンスを実現
- MPIベースのアプリケーションでは、シングルソケット+高速インターコネクトでの高い性能の実現
- DDR2 667MHzメモリによるHPCアプリケーションの高速実行

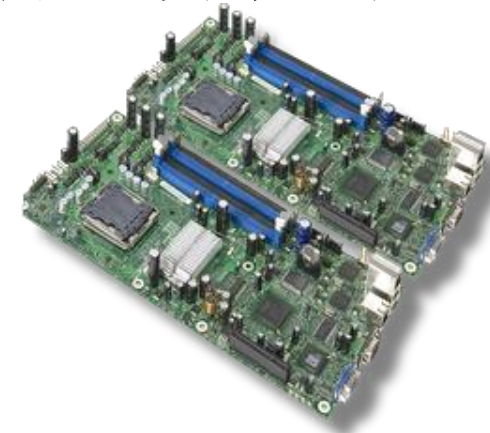
Microsoft Windows Compute Cluster Server 2003、RedHat、SuSEなどのLinuxディストリビューション

## NEXXUS 4820® PT

S3000PT

Intel® Xeon® 3000/3200  
Intel® Core™ Duo/Quad

シングルソケットHPC向けアーキテクチャ





# NAS Parallel Benchmark

- EP The Embarrassingly Parallel
- MG Multigrid
- CG Conjugate Gradient
- FT 3-D FFT PDE
- IS Integer Sort
- LU LU Decomposition
- SP Scalar Pentadiagonal
- BT Block Tridiagonal



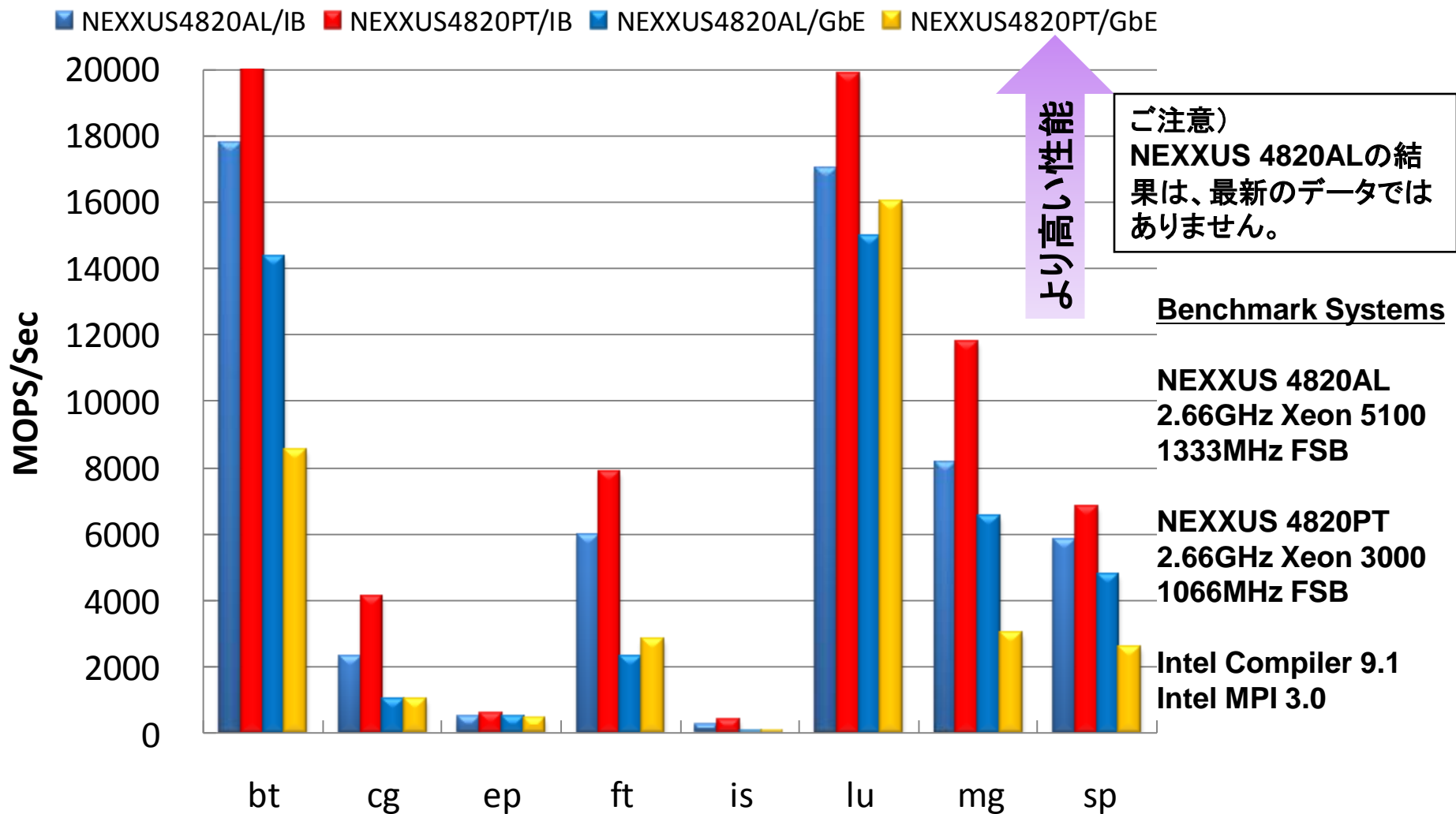
# ベンチマークコードの特徴

NPB	説明	通信の特徴
EP	指定されたスキームに従い生成された、多数のガウス分布に従う擬似乱数を用いたモンテカル法を用いた応用プログラムに良く見られる並列計算の特性を持つ	データ転送をほとんど必要としない
MG	3次元ポアソン方程式を簡略化したマルチグリッド法で解く	高度な構造的通信を必要とする
CG	大規模で正値対称な疎行列の、最小固有値の近似値を共役勾配法を用いて解き、非構造格子を用いたアプリケーションで一般的な並列計算の特性を持つ	不規則な長距離通信をテストし、スパースの行列-ベクトル積を行なう
FT	3次元偏微分方程式をFFTを用いて解き、スペクトル法を使用したアプリケーションの典型的な並列計算の評価となる	長距離通信性能のテスト
IS	パーティクル法を使用したアプリケーションにおいて重要なソートの性能を評価している。物理における、パーティクルをあるセルに割り当てて、パーティクルがセルから流れ出るかどうかを見る、particle-in-cell法のアプリケーションなどの並列計算の特性を見ることに適している	このベンチマークは整数演算スピードと通信性能の両方をテストを行い、浮動小数点演算は含まない
LU	LU分解を実際には行わず、5x5のブロックをもつ上下三角行列システムをSSOR(Symmetric Successive Over-Relaxation)法で解く	代表的な陰解法
SP	非優位対角なスカラ5重対角方程式を解く	
BT	非優位対角な5x5のブロックサイズをもつブロック3重対角方程式を解く	



# 16並列での実行性能

## NAS Parallel Benchmark/Class B



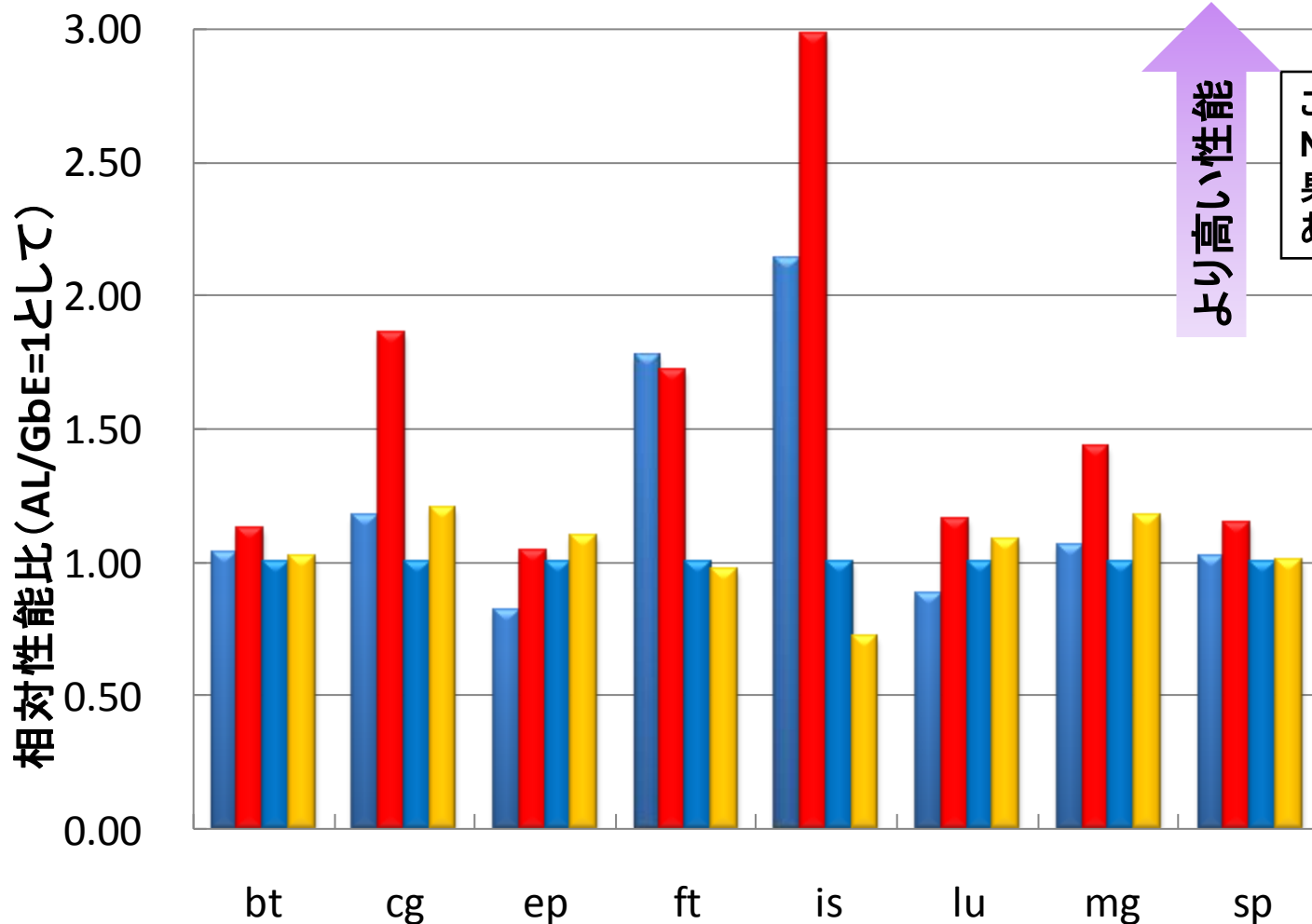




# 16並列での実行性能

## NAS Parallel Benchmark/Class B

■ NEXXUS4820AL/IB ■ NEXXUS4820PT/IB ■ NEXXUS4820AL/GbE ■ NEXXUS4820PT/GbE



より高い性能

ご注意)  
NEXXUS 4820ALの結果は、最新のデータではありません。

### Benchmark Systems

**NEXXUS 4820AL**  
2.66GHz Xeon 5100  
1333MHz FSB

**NEXXUS 4820PT**  
2.66GHz Xeon 3000  
1066MHz FSB

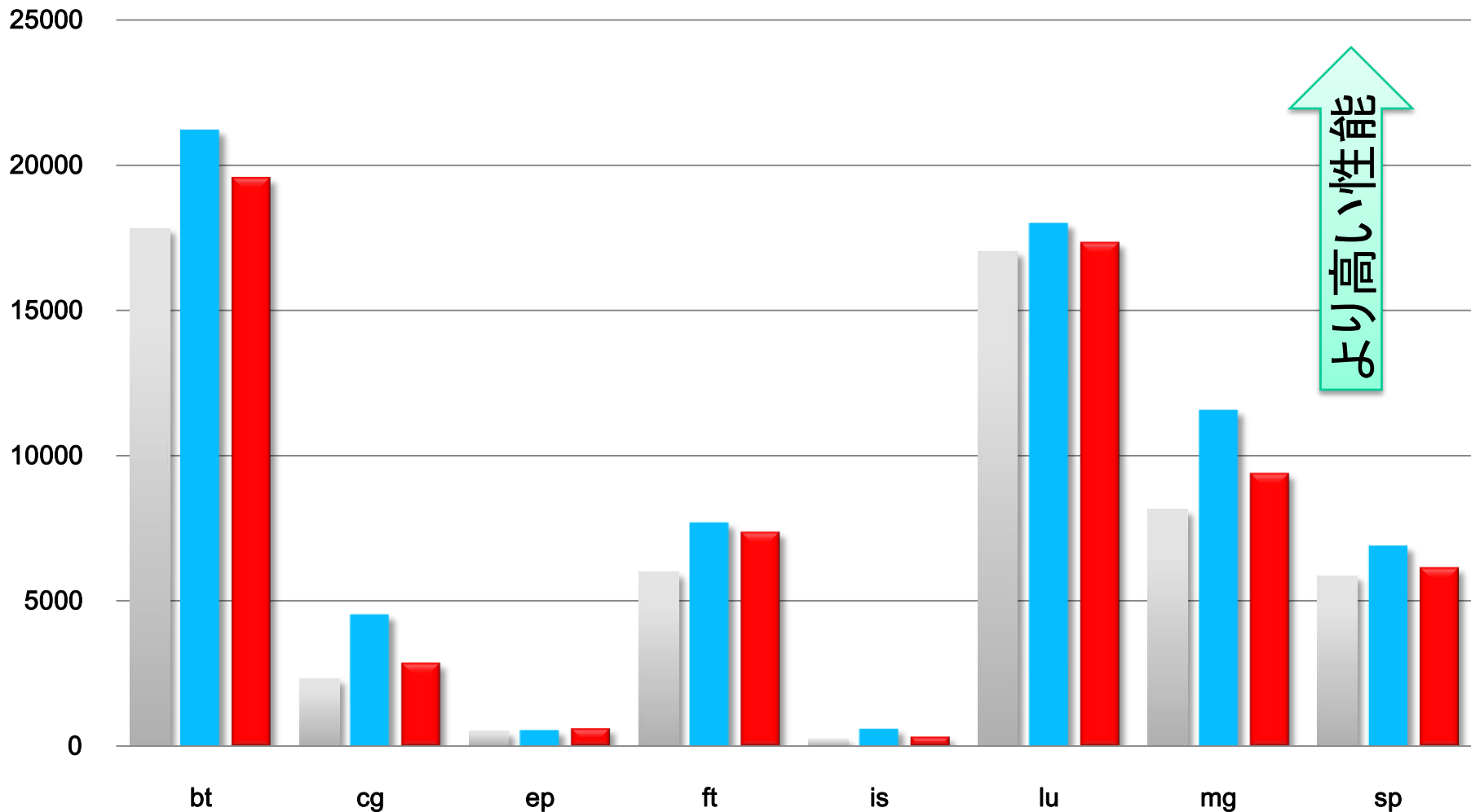
Intel Compiler 9.1  
Intel MPI 3.0



最新データ

# NAS Parallel Benchmark 16並列実行性能

■ NEXXUS AL(2.66GHz Xeon 5150) ■ NEXXUS PT(2.66GHz Core2Duo) ■ R1400 (3.0GHz Xeon 5160)

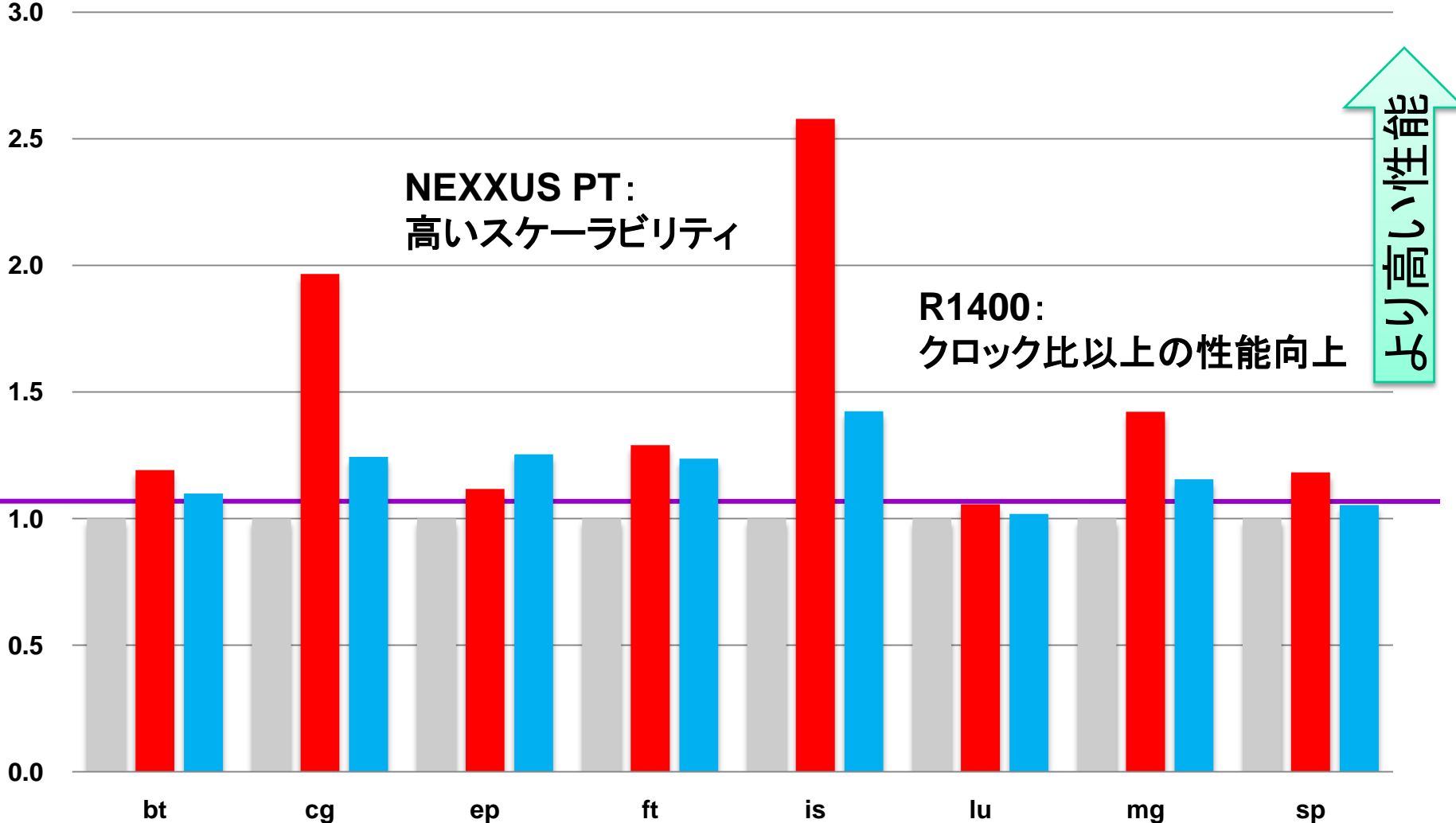


より高い性能



# NPB - 16並列実行性能 NEXXUS AL に対する相対性能比

■ NEXXUS AL(2.66GHz Xeon 5150) ■ NEXXUS PT(2.66GHz Core2Duo) ■ R1400 (3.0GHz Xeon 5160)





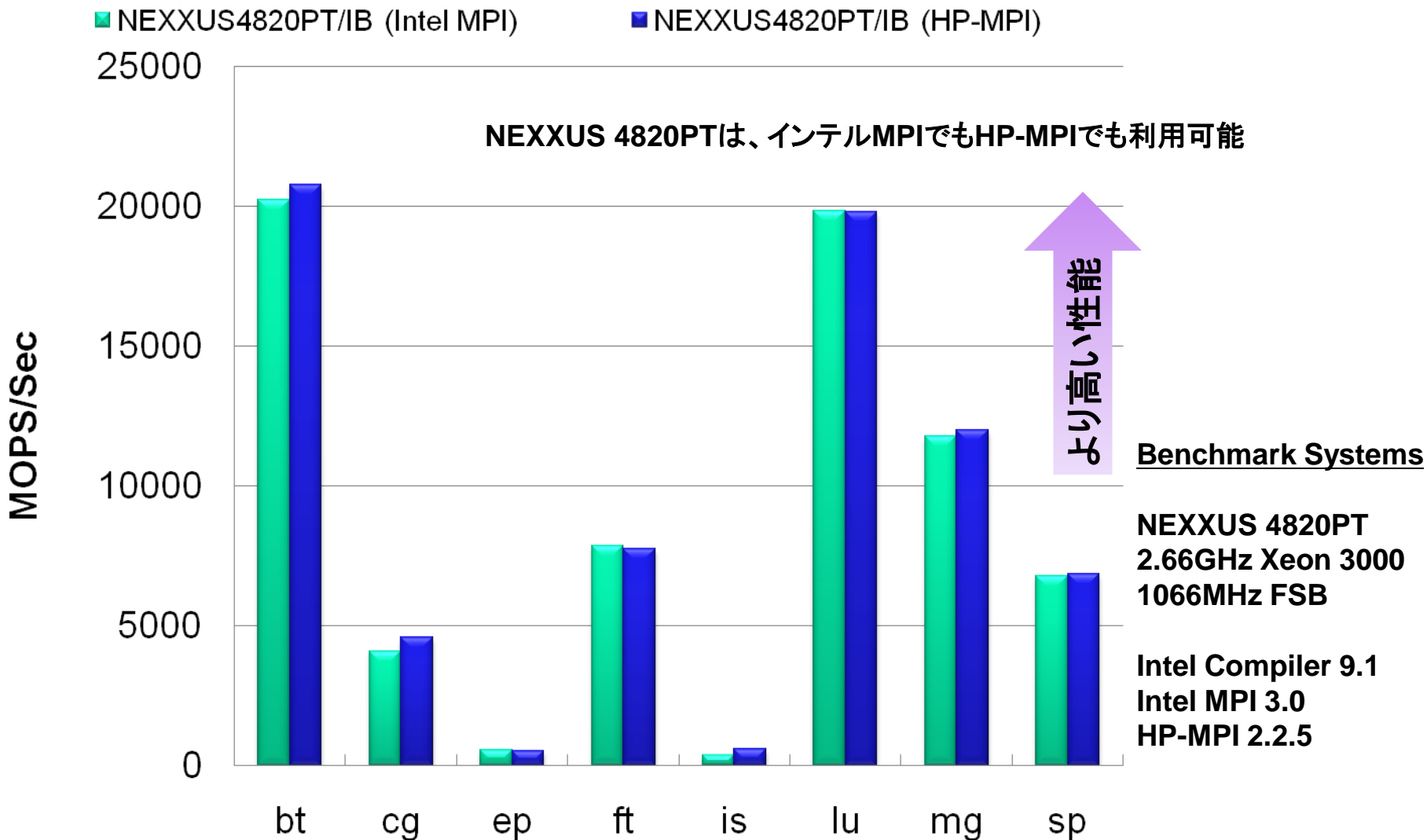
## ベンチマークの結果について

- NPBの結果は、各テストの通信特性と計算アルゴリズムの特性を反映した結果となっている
- InfiniBand搭載のNEXXUS 4820/PTは、ほぼ全てのテストで最も良い性能を示す
  - MPIなどのメッセージパッシングによる分散並列処理においては、シングルソケットを利用し、ローカルなメモリバンド幅とノード間的高速通信を利用するシステム構成は効率的である
  - S3000PTの開発の目的の正しさをこのベンチマークでは証明している



# MPI実装での実行性能比較

## NAS Parallel Benchmark/Class B





# NEXXUS 4000 WindowsCCS ベンチマーク

- ベンチマークシステム
  - NEXXUS 4820AL (ALと表記)
    - デュアルソケット / SE5000AL
  - NEXXUS 4820PT (PTと表記)
    - シングルソケット / S3000PT
- MPIベースのアプリケーション
  - NAS Parallel Benchmark
  - InfiniBand構成での性能の確認
  - Linuxアプリケーションのポーティングテスト





# NEXXUS 4000 WindowsCCS ベンチマーク

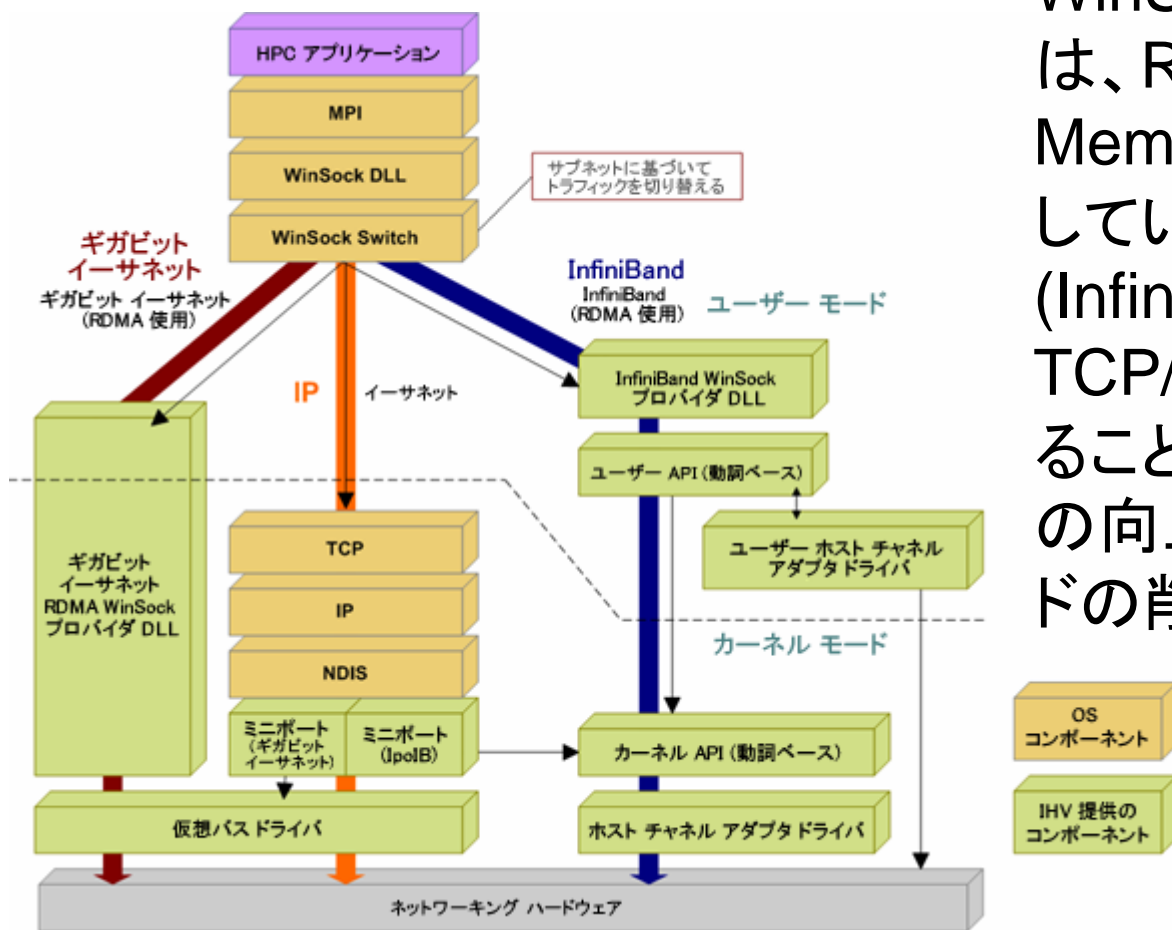
- インテル Visual Fortran コンパイラ  
ー 9.1 Windows版スタンダード・エ  
ディション
- Microsoft Message Passing  
Interface (MPI) ソフトウェア (MS  
MPI)
  - WinSock Direct トポロジを利用





# WinSock Direct トポロジ

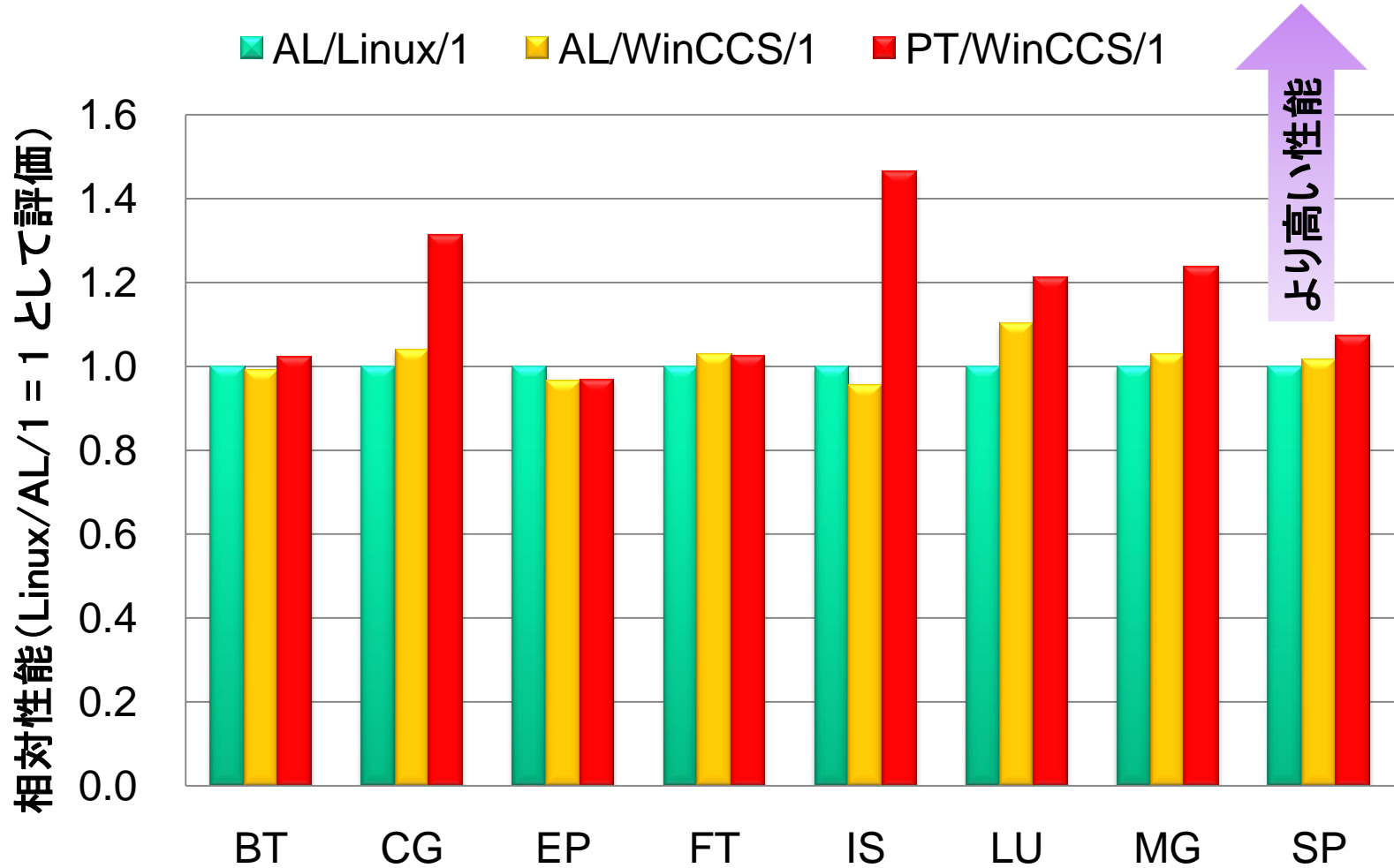
WinSock Direct プロトコルは、RDMA (Remote Direct Memory Access) をサポートしているハードウェア (InfiniBandなど) を使って TCP/IP スタックをバイパスすることにより、パフォーマンスの向上と CPU オーバーヘッドの削減を実現





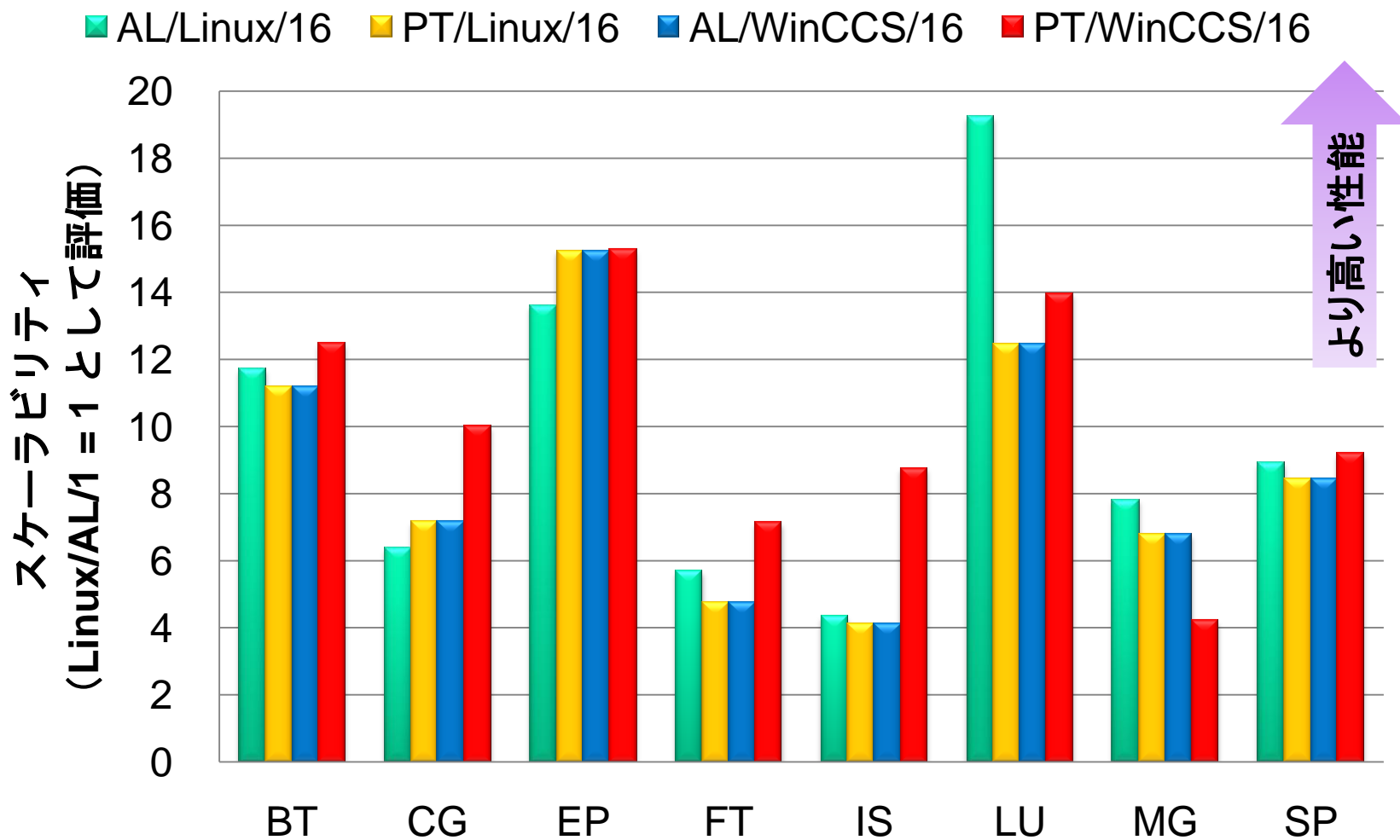


# NEXXUS 4000 WindowsCCS ベンチマーク





# NEXXUS 4000 WindowsCCS ベンチマーク





# NEXXUS 4000 WindowsCCS ベンチマーク

- ベンチマーク結果
  - NEXXUS 4000でのNAS Parallel Benchmarkでは、Windows CCSは、いくつかのテストを除いて、Linuxでのテストを上回る性能を示しています。
  - NEXXUS 4000のシングルソケットモデルは、一般的なデュアルソケットのサーバ以上の性能を示しています。また、Windows CCSでもその性能面での優位性を示しています。
- Windows CCS+NEXXUS 4000は、高速計算が必要なワークロードに関して、新しいオプションを提供します。

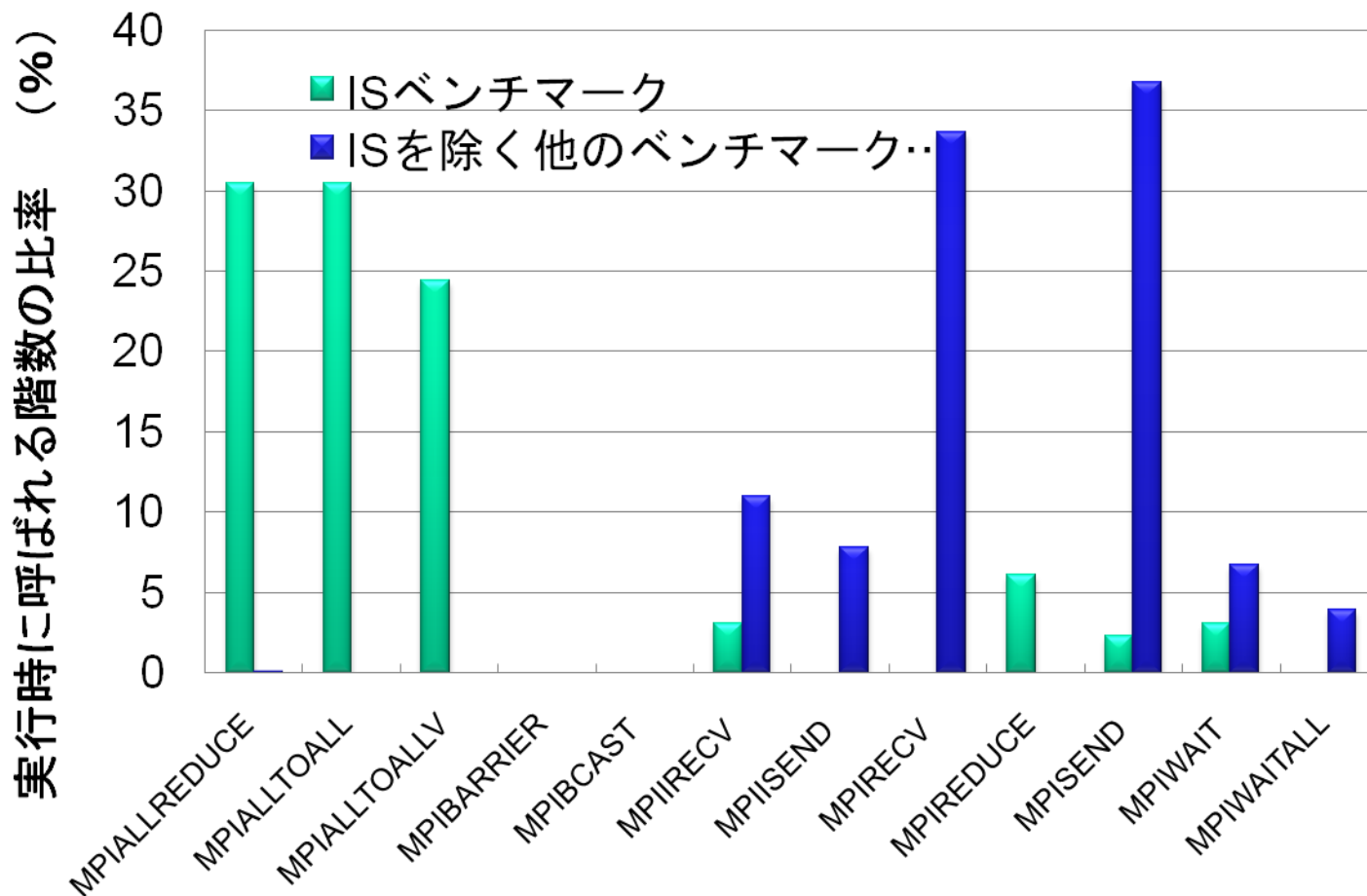


## 参考資料

- NAS Parallel Benchmark でのMPI通信の統計的な解析を行った論文が複数あり、その中で、NAS Parallel Benchmark の各プログラム中での利用MPIライブラリについて、調査が行われています。
- NAS Parallel Benchmarkについては、ISは他のプログラムとその実行特性が大きく異なることがこれらの調査でも明らかになっています。
  - Communication Characteristics in the NAS Parallel Benchmarks
    - Ahmad Faraj Xin Yuan
    - Department of Computer Science, Florida State University, Tallahassee, FL 32306
  - The Use of the MPI Communication Library in the NAS Parallel Benchmarks
    - Theodore B. Tabe, Member, IEEE Computer Society, and Quentin F. Stout, Senior Member, IEEE Computer Society



# MNAS Parallel Benchmarkでの通信特性



ISベンチマークは、他のアプリケーションと大きく違った通信パターンを持ち、また各通信時のメッセージサイズの変動が大きいという特徴を持ちます。



# Cluster OpenMP プログラム コンパイルと実行例

## クラスタ間共有データの定義

```
$ cat -n cpi.c
 1 #include <omp.h>
 2 #include <stdio.h>
 3 #include <time.h>
 4 static int num_steps = 1000000;
 5 double step;
 6 #pragma intel omp sharable(num_steps)
 7 #pragma intel omp sharable(step)
 8 int main ()
 9 {
10 int i, nthreads;
11 double start_time, stop_time;
12 double x, pi, sum = 0.0;
13 #pragma intel omp sharable(sum)
14 step = 1.0/(double) num_steps;
15 #pragma omp parallel private(x)
16 {
17     nthreads = omp_get_num_threads(); // 実行時間関数によるスレッド数の取得
18 #pragma omp for reduction(+:sum) // "for" ワークシェア構文
19     for (i=0;i< num_steps; i++){ // privateとreduction指示句
20         x = (i+0.5)*step; // の指定
21         sum = sum + 4.0/(1.0+x*x);
22     }
23 }
24 pi = step * sum;
25 printf("%5d Threads : The value of PI is %10.7f¥n",nthreads,pi);
26 }
27
```

// OpenMP実行時間関数呼び出し  
// のためのヘッダファイルの指定

## OpenMP実行時間関数

// OpenMPサンプルプログラム:  
// 並列実行領域の設定  
  
// 実行時間関数によるスレッド数の取得  
// "for" ワークシェア構文  
// privateとreduction指示句

## コンパイルとメッセージ

```
$ icc -cluster-openmp -O -xT cpi.c
cpi.c(18) : (col. 1) remark: OpenMP DEFINED LOOP WAS PARALLELIZED.
cpi.c(15) : (col. 1) remark: OpenMP DEFINED REGION WAS PARALLELIZED.
$ cat kmp_cluster.ini
--hostlist=node0,node1 --processes=2 --process_threads=2 --no_heartbeat --startup_timeout=500
$ ./a.out
4 Threads : The value of PI is 3.1415927
```

## 並列実行処理環境の設定



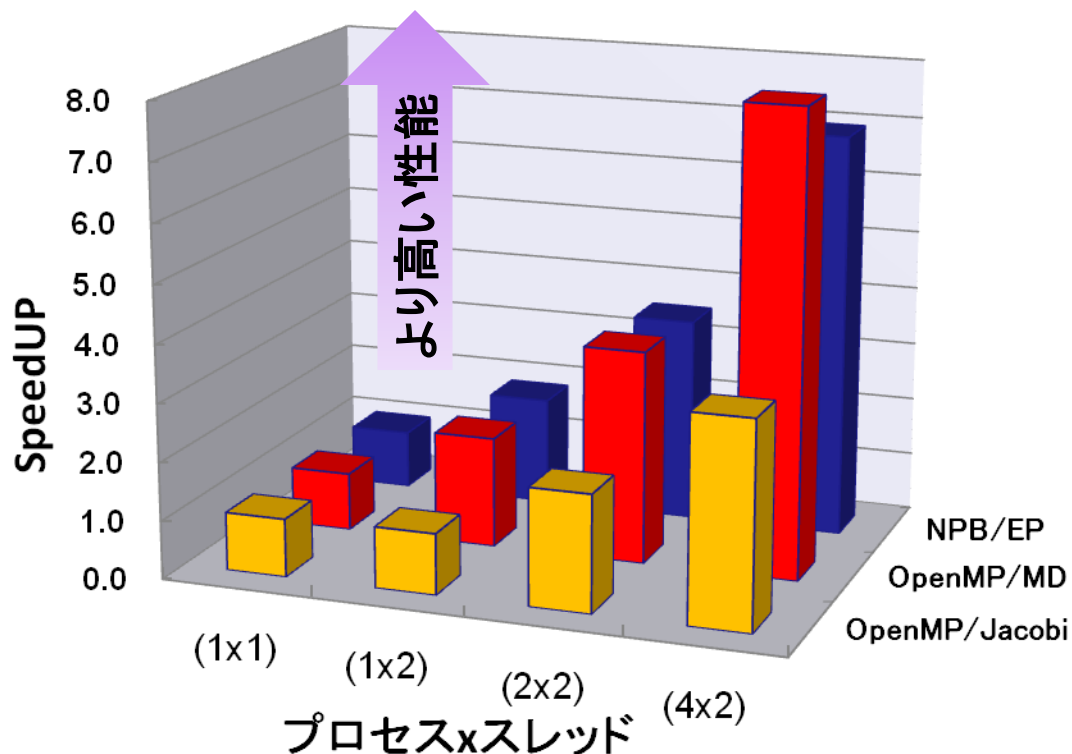
# Cluster OpenMP プログラム スケーラビリティサンプル

## ベンチマークシステム

- NEXXUS 4820-PT
- 2.66GHz/1066MHz FSB/16GB Memory/InfiniBand

## プログラムサンプル

- NAS Parallel Benchmark / EP ベンチマーク
- OpenMPサンプルプログラム (分子動力学サンプル、nparts=10000で実行)
- OpenMPサンプル (Jacobi法サンプル、5000x5000)





# プログラミングの生産性の向上

- 対話処理での開発環境

パーソナルクラスタ

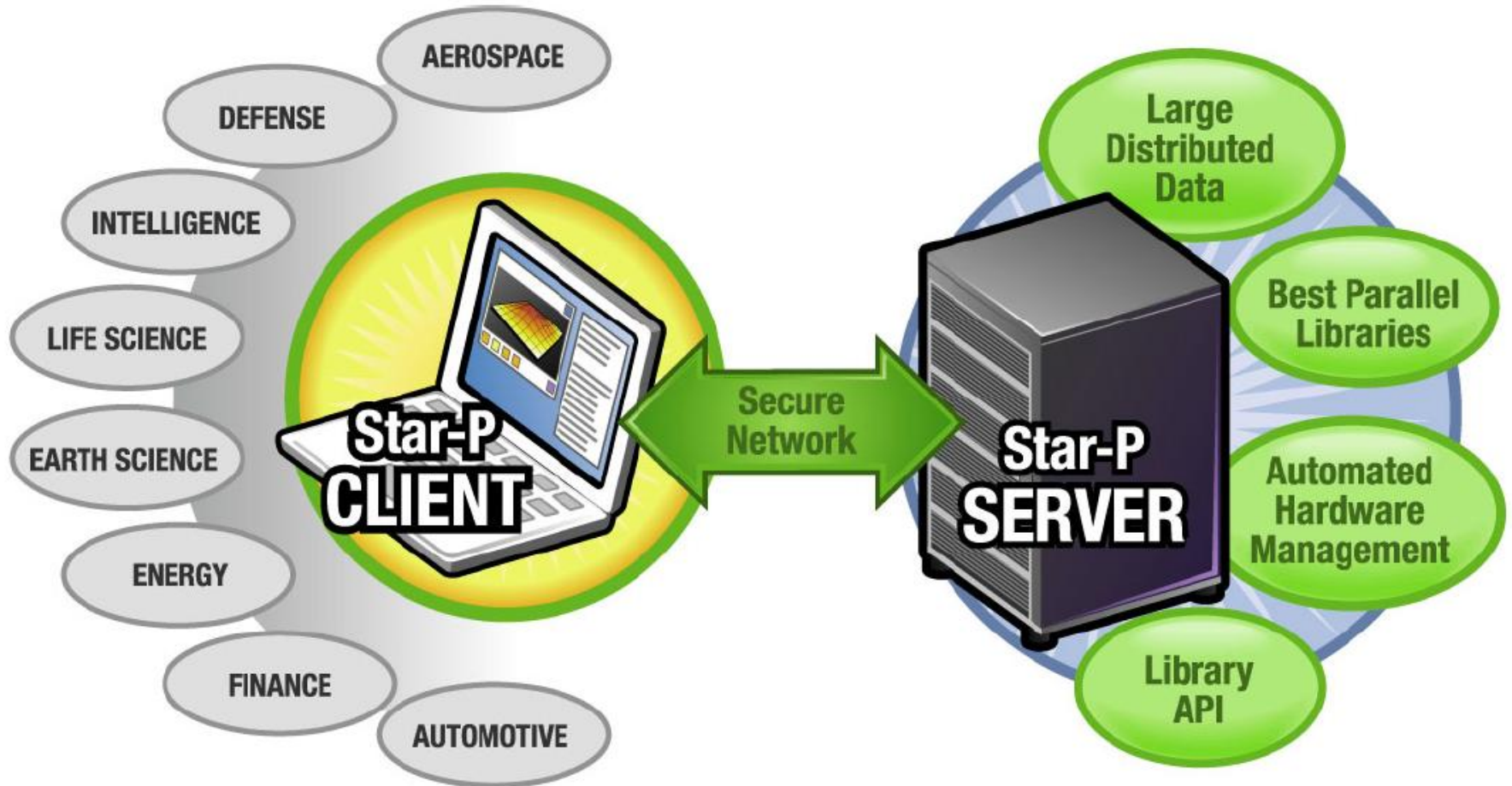






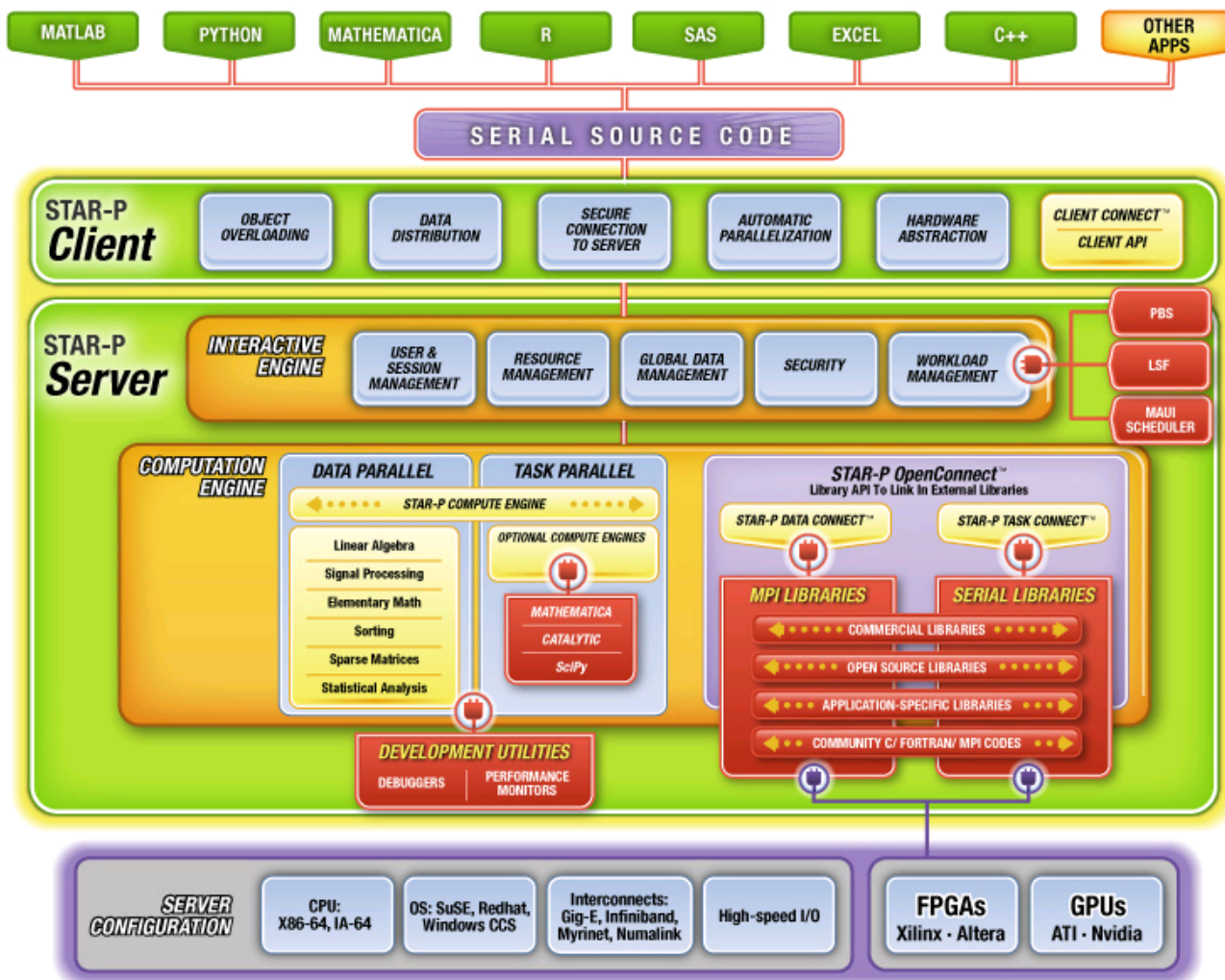
# Star-Pベンチマーク

- Star-P 並列計算プラットフォーム





# Star-P アーキテクチャ





# Program Sample and Performance

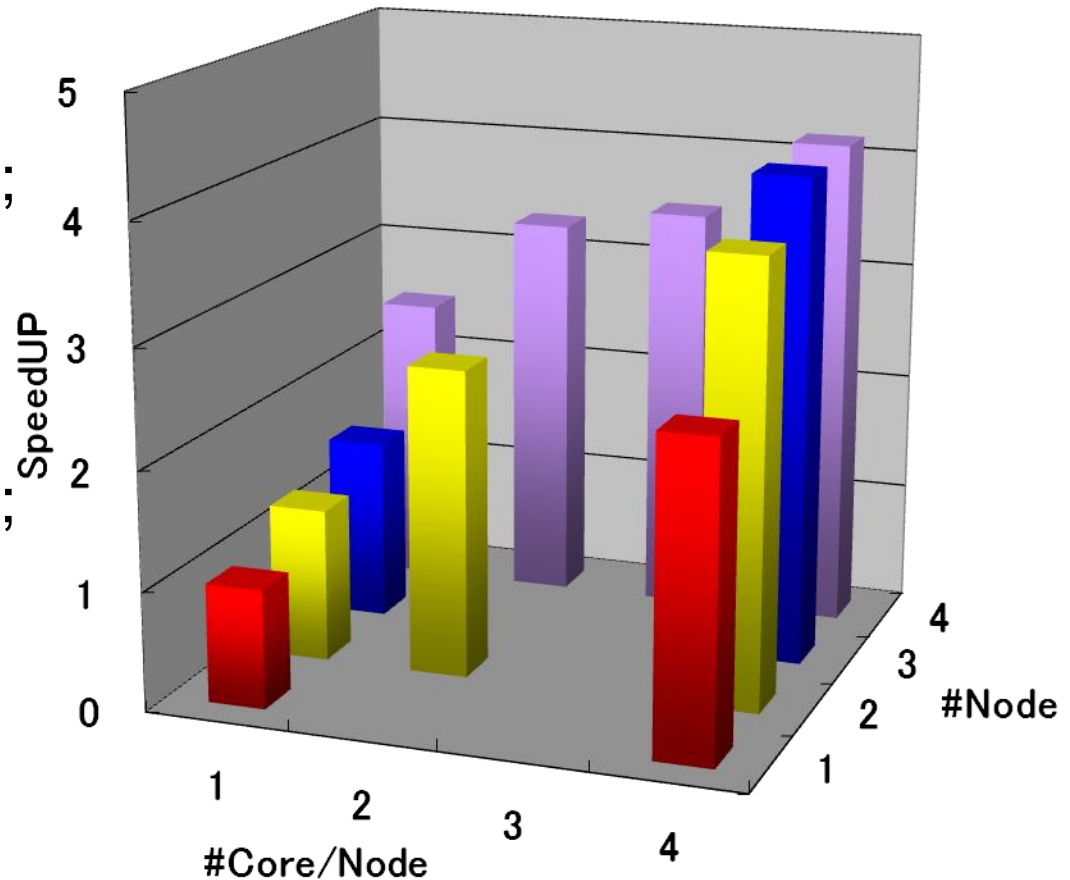
- NxN Inverse Matrix

## MATLAB

```
n = 7000;  
a = rand(n); b = rand(n);  
tic;a*b;inv(a);fft(a);toc
```

## Star-P

```
n = 7000*p;  
a = rand(n); b = rand(n);  
tic;a*b;inv(a);fft(a);toc
```





# Program Sample and Performance

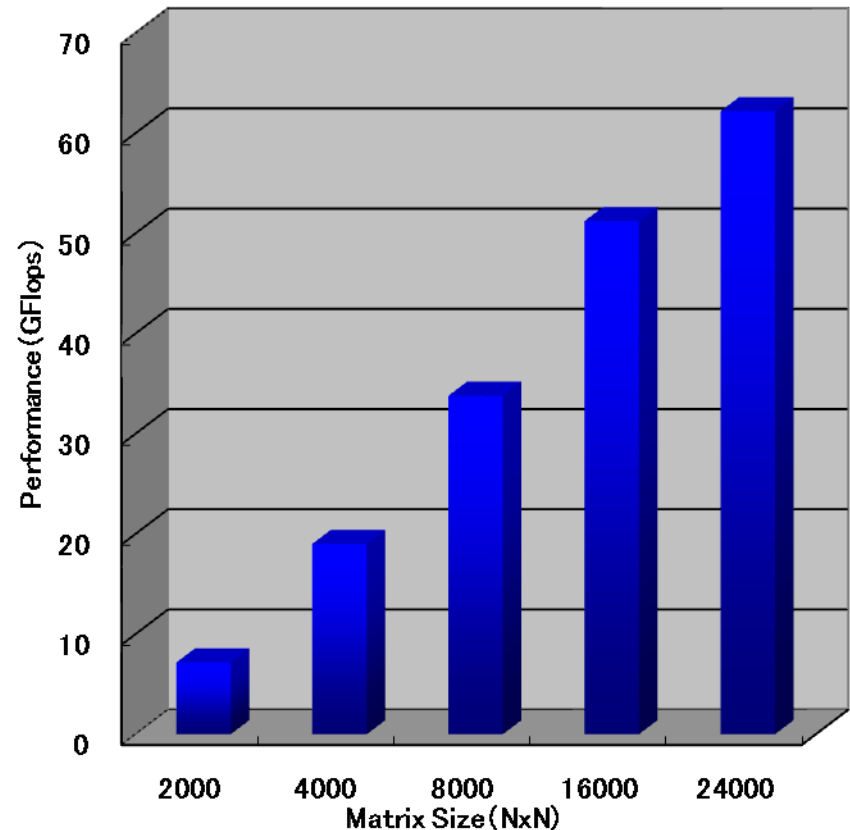
- NxN Matrix Multiply

## MATLAB

```
clear;  
n = 24000;  
a = rand(n); b = rand(n);  
tic;a*b;toc  
ppwhos a
```

## Star-P

```
clear;  
n = 24000*p;  
a = rand(n); b = rand(n);  
tic;a*b;toc  
ppwhos a
```



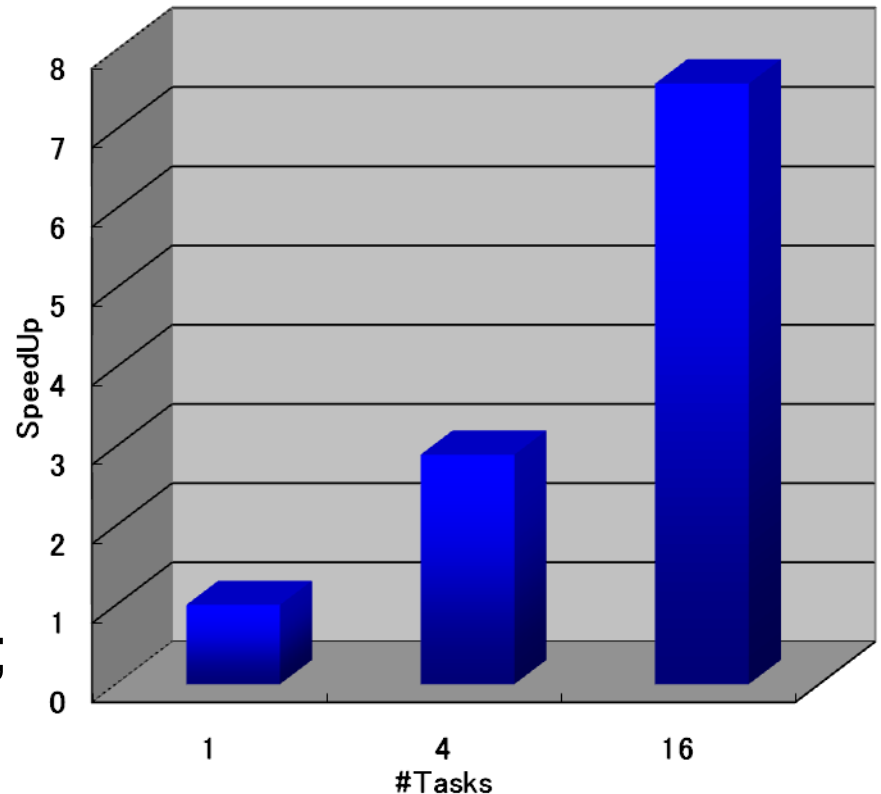


# Program Sample and Performance

- Multiple NxN Inverse Matrix

## Star-P

```
b=rand(1000,1000,40);  
a=rand(1000,1000,40);  
tic;  
for l=1:40  
    b(:,:,l)=inv( a(:,:,l) );  
end;  
toc  
a=rand(1000,1000,40*p);  
% b=rand(1000,1000,40*p);  
tic; b=ppeval('inv', a); toc
```





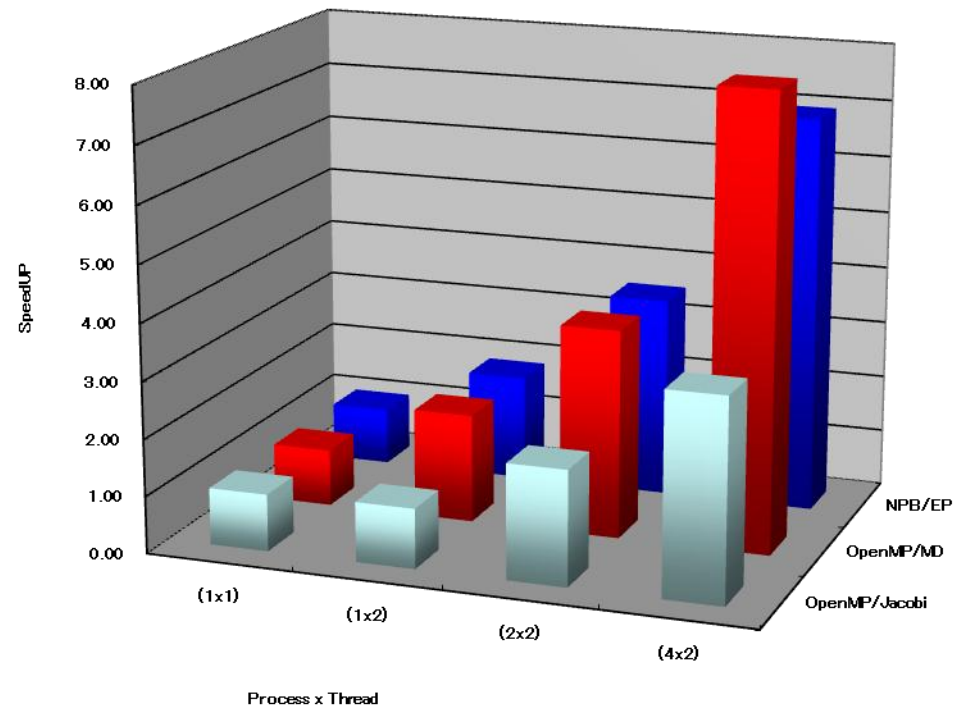
# Cluster OpenMP Program Scalability Study

## Benchmark System

- NEXXUS 4820-PT
- 2.66GHz/1066MHz FSB/16GB Memory/InfiniBand

## Program sample

- NAS Parallel Benchmark / EP Benchmark
- OpenMP Sample program (MD, nparts=10000)
- OpenMP Sample program (Jacobi, 5000x5000)





# ISVアプリケーション

- MPIベースの商用アプリケーションの性能は非常に重要
- NEXXUS 4820PT+InfiniBandオプションは、このような商用アプリケーションでも高い性能が期待できる
- 性能比較
  - 公開されているデュアルプロセッサ構成のシステムとの比較
  - 同一並列度での性能比較



# ベンチマークシステム

## NEXXUS 4820-PT

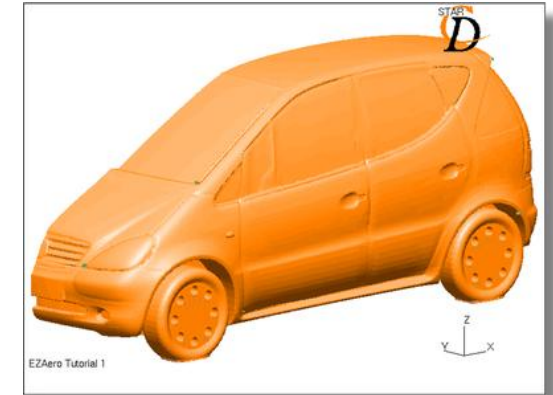
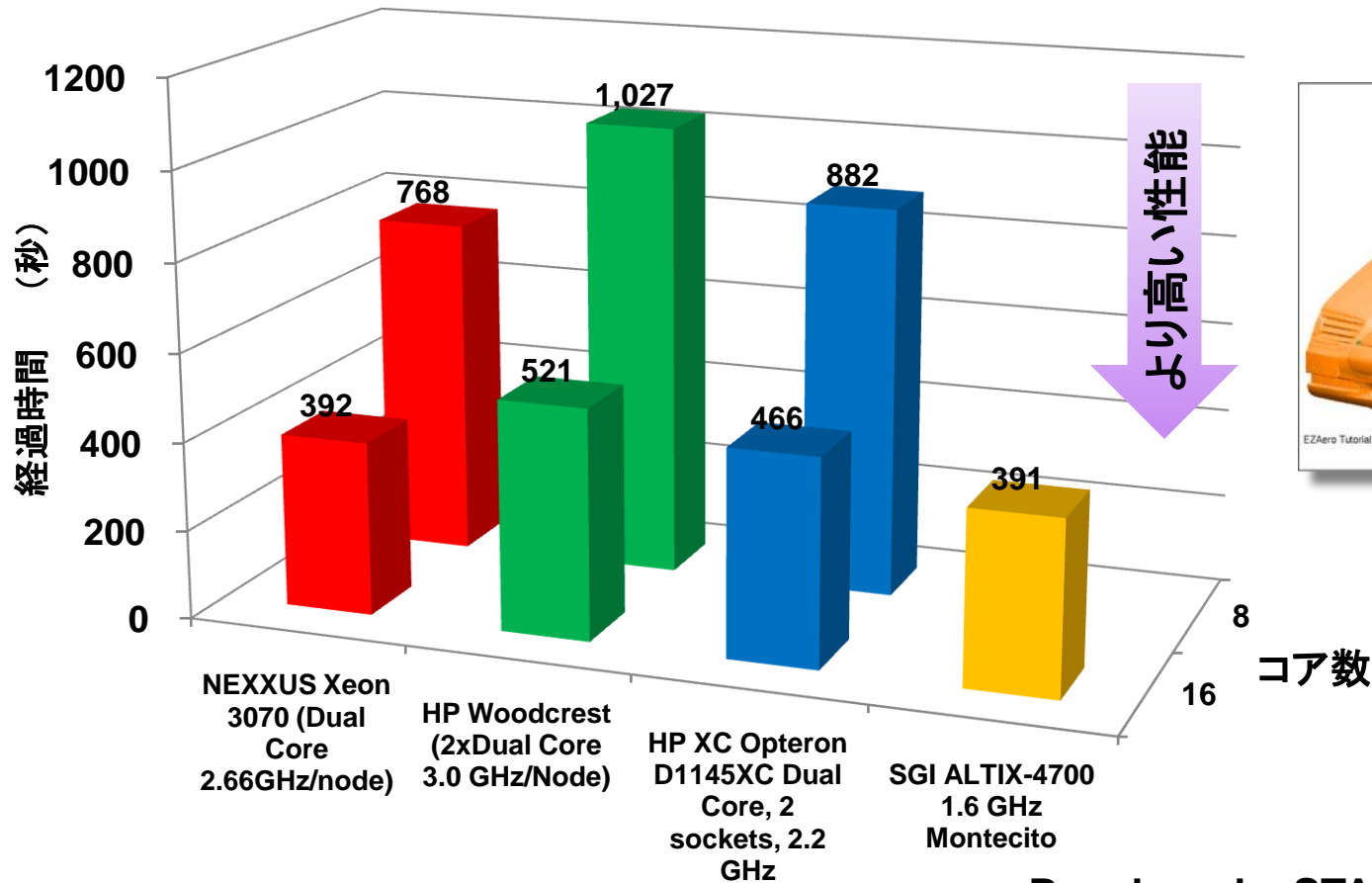
- 8 Processors、16 Cores (Intel Xeon 3000 Processor 2.66GHz 2x4M 1066MHz)
- 16 GB DDR2 533 ECC/Reg.
- 8 x HDD 80GB, 7200rpm, 4 MB, SATA II 2.5"
- Nexxus Chassis
- Integrated KVM/USB Switch
- Integrated 16 Port GigE Switch
- Integrated 8 Port InfiniBand Switch







# STAR-CD ベンチマーク

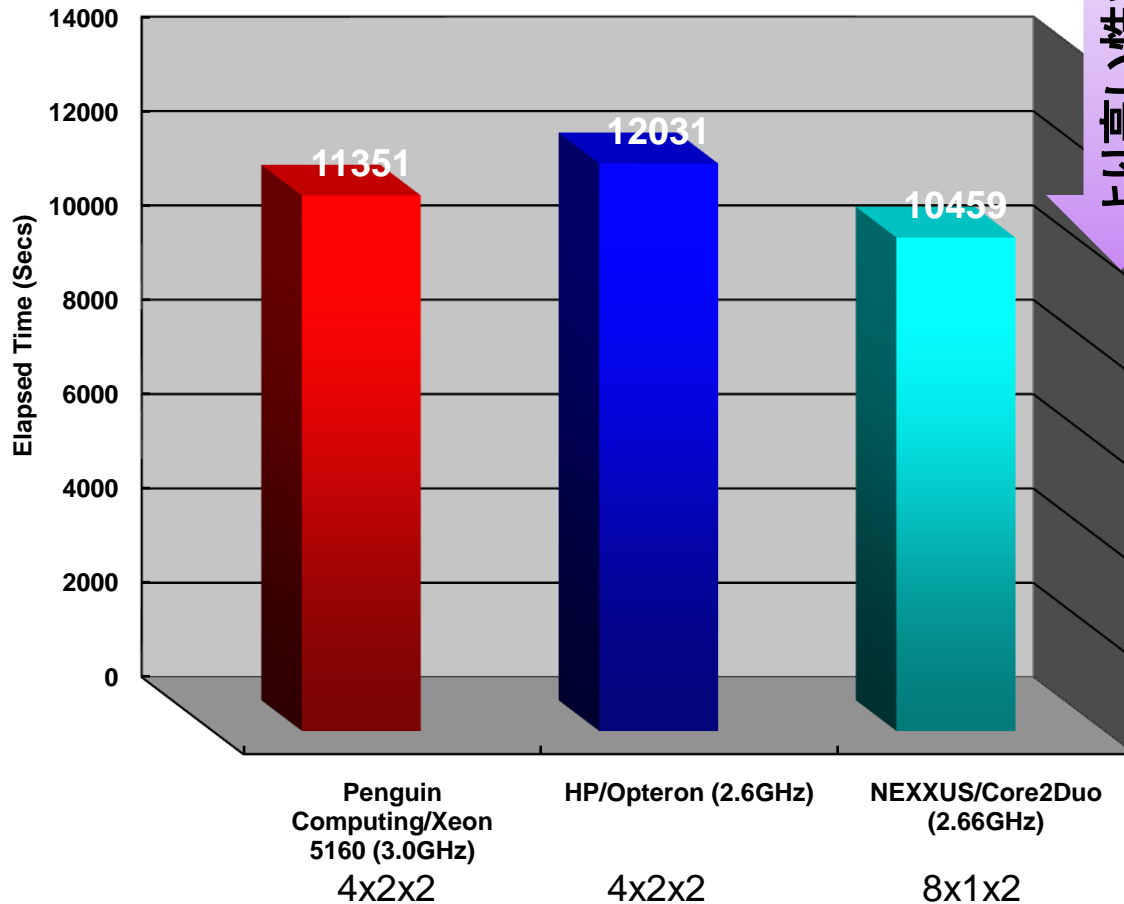


Benchmarks STAR-CD V3240/V3260  
A-Class DATASET

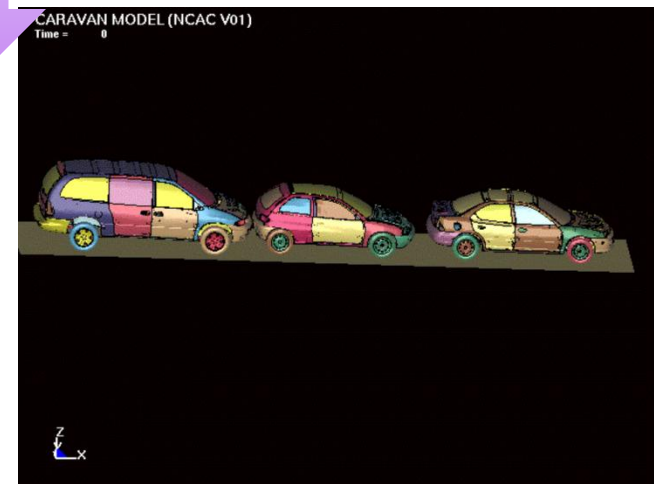
<http://www.cd-adapco.com/products/STAR-CD/performance/320/aclass32.html>



# LS-DYNA ベンチマーク 16 プロセッサコアベンチマーク



より高い性能



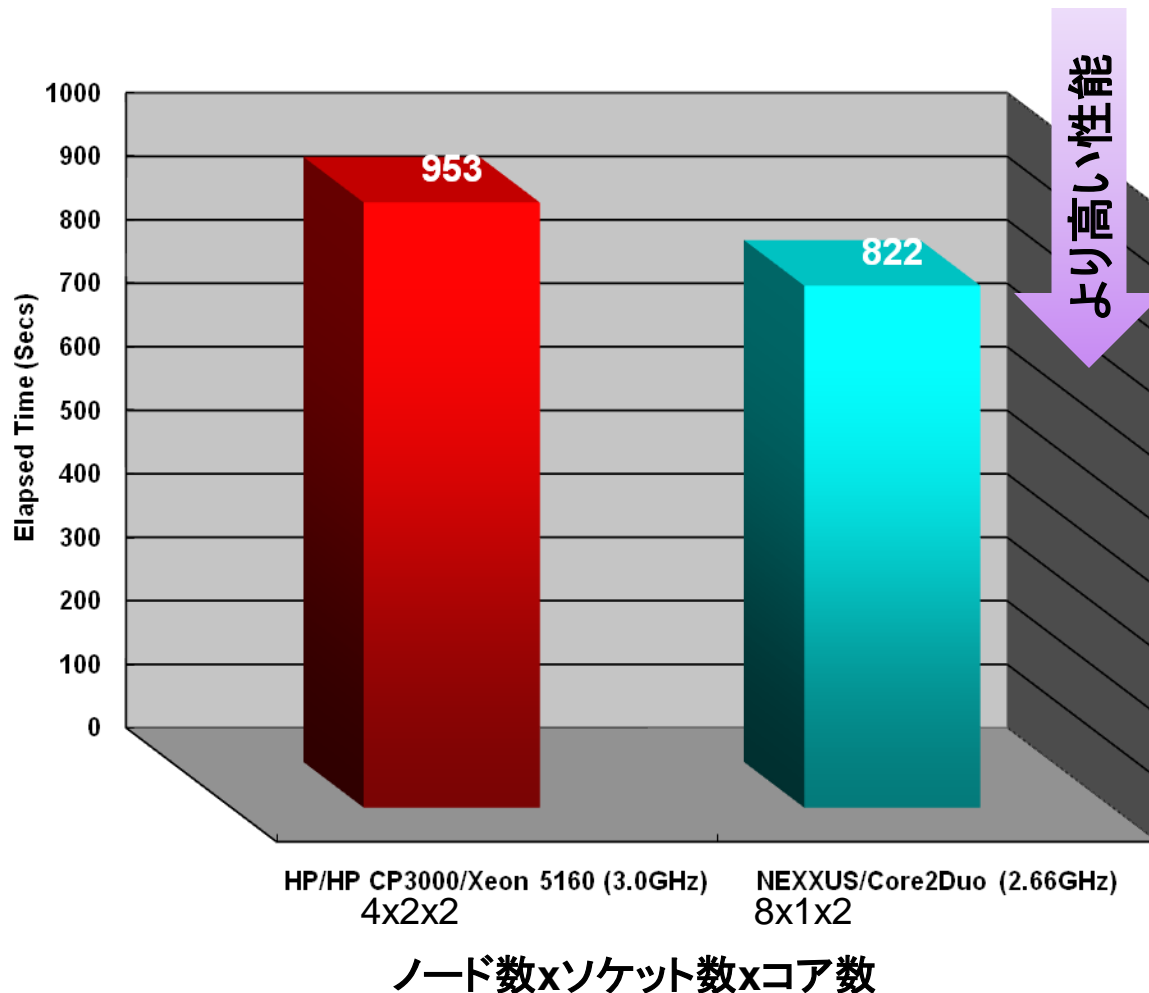
ノード数xソケット数xコア数

LS-DYNA 971  
3 Vehicle Collision

<http://www.topcrunch.org/>



# LS-DYNA ベンチマーク 16 プロセッサコアベンチマーク



LS-DYNA 971  
neon\_refined\_revised  
<http://www.topcrunch.org/>

スケーラブルシステムズ株式会社



# HPL ベンチマーク結果

マトリックスサイズ

GFLOPS値

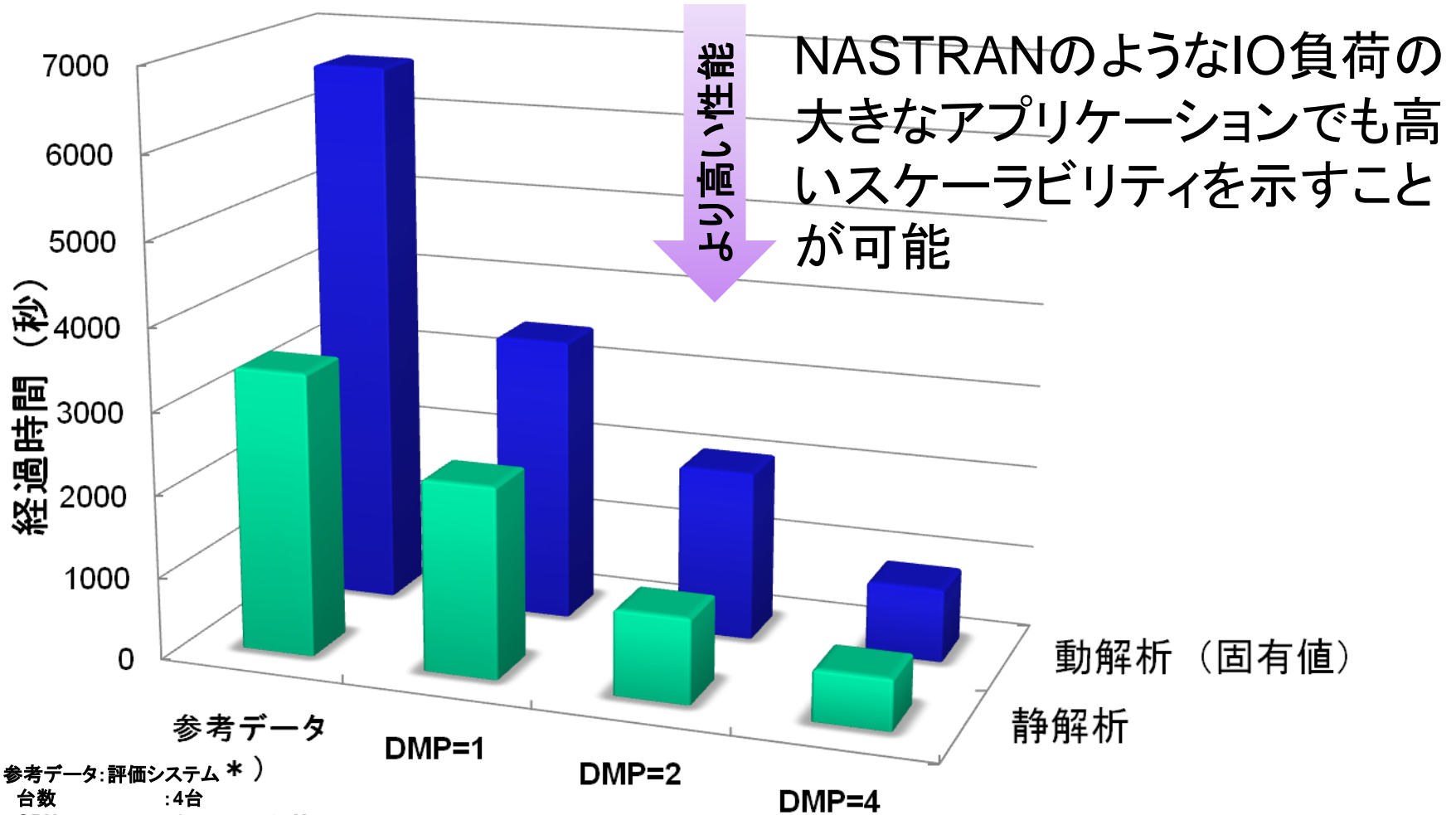
Computer (Full Precision)	Number of Procs or Cores	$R_{max}$ GFlop/s	$N_{max}$ Order	$N_{1/2}$ Order	$N_{Peak}$ GFlop/s
HP AlphaServer GS1280 7/1300 (1.3 GHz)	64	142.8	122500		166.4
NEC SX-6/16M2 (1.77ns)	16	142.8	51200	2048	144
Cray X-1 (800 MHz)	12	142.4	73728	7040	153.6
SGI Origin 2000 (195 MHz)	480	141.2	108864	13312	187
HITACHI SR8000-E1/16(300MHz)					
SGI 1100 Cluster (Dual Pentium III, 1 GHz)					
IBM SP 8 nodes (375 MHz POWER3 High)					
CRAY X1 (800 MHz, 12 procs)					
IBM S80s (450 MHz, SP switch)					
IBM eServer pSeries 690 Turbo(1.1 GHz)					
Compaq ES40/EV6 AlphaServer SC					
Fujitsu VPP500/100(10nsec)					
HP Superdome (750 MHz)	64	133.82	138888		192
IBM SP 32 nodes (375 MHz POWER3 Thin)	128	132.75	107000	15400	192
hp server rp8400 (750 MHz, HyperPlex)	64	132.71	137808	21384	192
hp server rp8400 (750 MHz, 1000bT)	64	132.69	165456	29268	192
Sun Fire 15K (1050MHz/8MB E5)	80	132.6	96116	14000	168.0
Intel Itanium 2 1.3 GHz	32	132.5	73400		166.4
Dell PowerEdge HPC(Dual Pentium III, 1 GHz)	400	131.0	130000	65000	400
Fujitsu VPP500/96 (10nsec)	96	129.5	49728	12430	154
Fujitsu VPP700/64 (7nsec)	64	129.5	115200	12800	141
Paragon XP/S MP(1024 Nodes, OS=SUNMOS S1.6)	3072	127.1	86000	17800	154
NEC SX-8/8 (2 GHz)	8	126.2	30720		128
NEC SX-5/16 (4 nsec)	16	125.8	55296		128

T/V	M	NB	P	Q	Time	Gflops
W10C2L4	40000	112	2	8	325.98	1.309e+02
$  Ax-b  _{\infty} / (eps *   A  _1 * N) =$						0.0161855 ..... PASSED
$  Ax-b  _{\infty} / (eps *   A  _1 *   x  _1) =$						0.0147993 ..... PASSED
$  Ax-b  _{\infty} / (eps *   A  _{\infty} *   x  _{\infty}) =$						0.0029062 ..... PASSED

サイズ40000という非常に小さなサイズでも高い性能を示す



# NX NASTRANベンチマーク結果



参考データ: 評価システム \* )  
台数 : 4台  
CPU : Opteron 246 X2  
Memory : 8GB  
HDD : 600GB/PC  
OS : SUSE10.1 x86-64SMP

動解析 (固有値)

静解析

より高い性能

NASTRANのようなIO負荷の大きなアプリケーションでも高いスケーラビリティを示すことが可能



## ベンチマーク結果について

- NASTRANのようなIO負荷の大きなアプリケーションでも、シングルソケットのノードであれば、ディスクを占有することが出来るため、複数ノードでのスケーラビリティを示すことが可能となっている
- 高価なストレージシステムを導入することなく、高速のスクラッチ領域をもった解析システムの構築が可能となる
- シンプルなシステム構成のため、ボトルネックの把握とその対応が容易



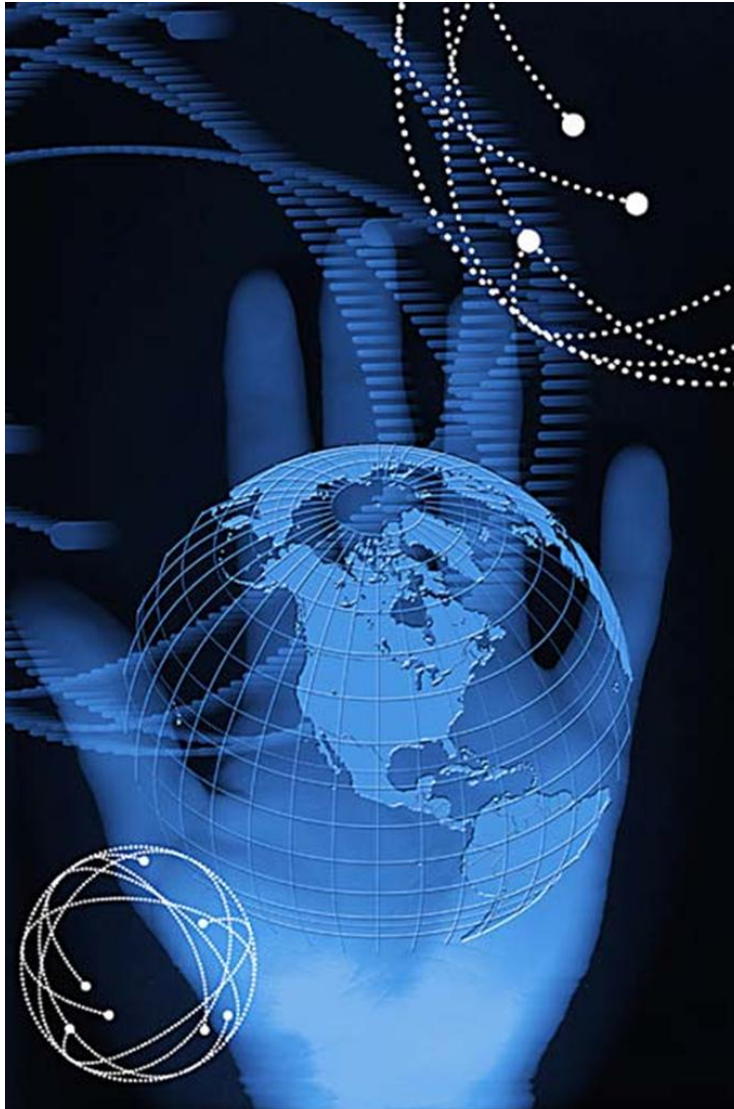
## この資料について

この資料の無断での引用、転載を禁じます。  
社名、製品名などは、一般に各社の商標または登録商標です。なお、本文中では、特に®、TMマークは明記していません。

In general, the name of the company and the product name, etc. are the trademarks or, registered trademarks of each company.

Copyright Scalable Systems Co., Ltd. , 2007. Unauthorized use is strictly forbidden.

8/15/2007





さらに詳しい情報や最新情報は.....

ホームページにて公開しています。ホームページには、お問い合わせ窓口も開設しておりますので、ご利用ください。

コンサルテーション

<http://www.sstc.co.jp>

製品技術

<http://www.hp2c.biz>

8/15/2007

