計算機アーキテクチャ 142−24
ハイパフォーマンス 85−24
コンピューティング
（2001. 3. 9）

# CRAY-T3E/1200E における NAS Parallel Benchmark

黒川原佳† 井口寧†† 松澤照男††

本報告では，CRAY-T3E から CRAY-T3E/1200E へのハードウェアのバージョンアップ (CPU のクロック速度と通信速度の向上) によるシステムの全体性能を検討する．CPU クロック速度の向上では，システムのリニアな性能向上は望めない．また，通信速度の向上が含まれるため，システムの全体性能に様々な影響があらわれる．アプリケーションプログラムを実行する上での性能向上に対する複合的な要因を評価する必要がある．検討に用いるベンチマークプログラムは，様々な評価が行なえる NAS Parallel Benchmark (NPB) を用いた．NPB は実際の科学技術計算を行なった場合の評価も行なえる．その結果，CPU クロック速度と通信速度の向上によって平均 40 % 程度の性能向上が得られたことが分かった．

# NAS Parallel Benchmark on the CRAY-T3E/1200E

Motoyoshi Kurokawa,† Yasushi Inoguchi†† and Teruo Matsuzawa††

In this paper, we discuss the system performance between CRAY-T3E and T3E/1200E. CRAY-T3E/1200E is a new version of T3E and it has twice speed of CPU clock and improved communication performance. System performance is not linearly improved with CPU clock speed. Various influences appear to the system performance because the improvement of the communication performance is also included. It is necessary to examine various factors to evaluate system performance the from benchmark codes. NAS Parallel Benchmark (NAS), which is able to evaluate various features, is used as a benchmark program. NPB can evaluate system performance based on real-world science and technology computations. AS a result, it is shown that the system performance is improved about 40 % by upgraded hardware.

## 1. はじめに

本報告では，CRAY-T3E から CRAY-T3E/1200E へのハードウェアのバージョンアップ (CPU のクロック速度と通信性能の向上) による性能への影響を検討する．クロックアップだけでは全体性能をリニアに性能向上させることは難しい．CPU の演算速度には，データ供給が伴うため CPU のみのクロック速度の向上では，データ供給を行なうメモリ性能が相対的に低くなることになる．また，通信性能もクロックアップで演算性能が上昇するため，通信性能も同等に上昇させることで相対的な性能低下は防げるが，通信性能を相対的に向上させることは非常に困難である．ハードウェアのバージョンアップによる性能向上は，様々な要因が考えられるため，様々な評価を行なえるベンチマークを用いる必要がある．

† 北陸先端科学技術大学院大学 情報科学研究科 博士後期課程
School of Information Science, JAIST
†† 北陸先端科学技術大学院大学 情報科学センター
Center for Information Science, JAIST

並列計算機の性能評価に用いるベンチマークプログラムは，プログラムの性質や調べたい性能に合わせて様々存在する．中でもよく用いられるのは，TOP500 の評価に用いられている Linpack や NAS Parallel Benchmark(NPB)[1)~3)] である．Linpack は，連立一次方程式の直接解法を基本としたものである．しかし，Linpack は並列計算機のある一面を評価していると考えられる．本報告では，より総合的な判断を行なうため，様々なベンチマークプログラムの集合体である NPB を用いた．NPB は，Computational Fluid Dynamics(CFD) の研究分野において必要な数値計算をほぼ全て含んでおり，他分野の数値計算の多くを内包している．

ハードウェアのバージョンアップは，PC クラスタ等のようなコモディティな部品を用いる場合に良く行なわれる．PC クラスタの利点は，各部品単位の変更が容易に行なえる．この性質は，ベクトル型計算機を除く，昨今の MPP システムにも十分に適用出来ると思われる．MPP システムのバージョンアップは，現状では比較的稀れなケースであると考えられるが，今後

は，MPP システムの構成部品のコモディティ化が急速に進むと考えられる．しかし，PC のハードウェアのバージョンアップとは，通信機構やメモリ機構の構造が大きく異なるため一致しないと考えられる．MPP システムのハードウェアをバージョンアップした場合，アプリケーション毎にどのような性能向上やボトルネックが存在するかを的確に知る必要がある．

本報告では，CRAY-T3E から CRAY-T3E/1200E へのハードウェアのバージョンアップによる性能向上の様々な影響を評価するため，様々なベンチマークプログラムの集合体である NPB V2.3 を用いて詳細に検討した．

## 2. CRAY-T3E システム

CRAY-T3E と CRAY-T3E/1200E について概略を述べる．CRAY-T3E は，CRAY Inc.(旧 CRAY Research) が開発した MPP システムである [4]．CRAY-T3E/1200E は，CPU と通信速度を向上させたものである．基本的なシステムの特徴は，以下の通りである．

(1) 1 ノード辺り 1 PE
(2) CPU は，Alpha 21164 (EV5)
(3) 外部レジスタ・セット (E-Register) によるリモートアクセス (グローバル・アドレッシング)
(4) Stream Buffer によるローカル・データアクセス
(5) 三次元 Bi-directional トーラス結合

CRAY-T3E は，1 ノード辺り 1 Processing Element (PE) である．Alpha 21164(EV5) を CPU とし，三次キャッシュを持たない．代わりにベクトルパイプランの概念を取り込んだ Stream Buffer と呼ばれる定ストライド (4 段) のパイプライン型 (6 本) のバッファを持ち，二次キャッシュのミスヒットの先読みによるデータ転送機能を有する．しかし，Stream Buffer の使用は，Stream Buffer 内部のデータとキャッシュあるいはメインメモリ内のデータとの不整合性が発生する場合があり，データを間接参照する場合に注意が必要となる．また，E-Register を介してグローバル・アドレス空間に存在するデータ (ローカル/リモート) の全てにアクセス出来る．インターコネクト・ネットワークは，三次元のトーラス構造のネットワークトポロジである．

CRAY-T3E と上位モデルである CRAY-T3E /1200E は，システムの特徴は同一である．CRAY-T3E と CRAY-T3E/1200E とのシステムの性能差を表 1 と各 PE の概要を図 1 に示す．

CPU のクロック速度向上によって演算性能と CPU 内部のキャッシュ速度が 100% 向上した．しかし，メインメモリへの帯域等の CPU 外の性能は同一である．ネットワーク性能は，通信帯域の増加によって約 45 % 向上した．また，CRAY-T3E システムの通信機構は，PE(CPU) の動作と密接に関連したシステムであるため，通常の PC 等で通信性能を向上させた場合とは全く異なるものである．

ベンチマークプログラムをコンパイルする際のオプションは，全て "-O3 -dp" とした．

## 3. NAS Parallel Benchmark

NAS Parallel Benchmark (NPB) について概略を述べる．NPB は，NASA Ames Reseearch Center の NAS(Numerical Aerodynamics Simulation) Lab. で開発された熱流体関連の科学技術計算のベンチマークである．(現在の最新バージョンは 2.3)

NPB は，5 つの主要アルゴリズムのカーネルと 3 つの数値流体計算コード (アプリケーションコード：圧縮性流体を擬似的に計算) からなる．それぞれ、並列計算コードと逐次計算コードからなり，並列計算コードは，Fortran 77 と MPI，逐次計算コードは，Fortran 77 で記述されている．また、計算サイズも 4 種類用意されており，用途に応じた計算サイズを用いることが出来る．計測結果として，Mop/s(Mega OPeration per Second) 値が得られる．EP と IS を除いたものは，ほぼ MFlop/s と同等の値である．EP の Operation は，乱数生成数であり，IS の Operation は，整数の数である．そのため，EP と IS に関しては，他のベンチマークコードから得られる Mop/s 値と区別する必要がある．ベンチマークコードは，問題サイズ (メモリサイズ) の大きさによって CLASS 分けがなされている．本報告で用いた CLASS は，問題サイズが小さい順に CLASS W，A，B，C である．以下に各ベンチマークコードの概要を示す．

- 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により求めるプログラムである．非構造格子を用いた流体アプリケーションで良く用いられる．計算は，メモリ帯域を必要とする．同一長のベクトルデータの通信と内積を得るための 1 データの通信が行なわれる．
  - **EP** 乗算合同法によって一様，正規乱数を生成するプログラムである．モンテカルロ法でよく用いられ，並列で解く際に通信がほとんど発生しない．そのため，浮動小数点演算性能のみを示す．
  - **FT** FFT を用いて三次元偏微分方程式を解くプ

表 1　性能比較

| | CRAY-T3E | CRAY-T3E/1200E |
|---|---|---|
| CPU | EV5 (**300 MHz**) | EV5 (**600 MHz**) |
| Primary cache size (Bandwidth) | 4KB (**4.8 GB/s**) | 4KB (**9.6 GB/s**) |
| Secondary cache size (Bandwidth) | 96KB (**4.8 GB/s**) | 96KB (**9.6 GB/s**) |
| Main Memory | 64 MB | 512 MB |
| Interconnect Bandwidth | **450MB/s** | **650MB/s** |

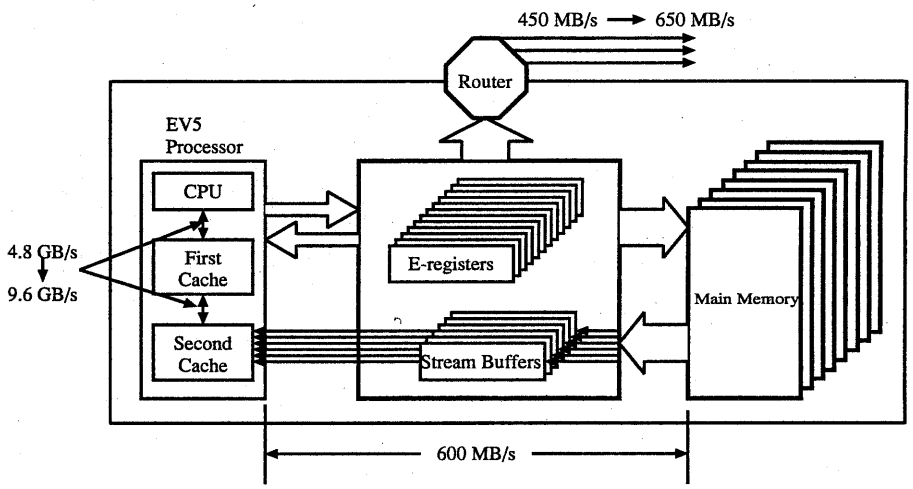


図 1　PE 内の概要

ログラムである．乱流の DNS (Direct Numerical Simulation) 解析をスペクトル法で行なう際に良く用いられる．FFT を各次元毎に解いていくため配列を持ち替える．特に FT は複素数型の配列を MPI_Alltoall で通信するため，通信負荷が非常に大きい．

**IS**　大規模な整数値ソートを行うプログラムである．粒子法等でよく用いられる．粒子を再び適切なセルに割り当てるためのソートを行なう．MPI_Alltoallv の負荷が通信の多くを占める．このベンチマークのみ C 言語で記述されている．

**MG**　三次元 Poisson 方程式を Multigrid 法によって求めるプログラムである．非圧縮流体計算中に現れる Poisson 方程式を解くために用いられる．メッセージ長が非一様な通信を行なうが，計算負荷は高くない．

- 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 法を用いて解くプログラムである．

**SP**　5 重対角方程式をスカラー ADI 法を用いて解くプログラムである．BT, SP ともに物理量等の格子面データの送受信を行ない，一部のデータ通信に関して通信隠蔽が行なわれている．演算量は，CLASS W を除いて，BT が多い．

**LU**　上下三角行列を対称 SOR 法を用いて解くプログラムである．並列化にはパイプライン処理が用いられ，通信処理の隠蔽が行なわれる．CLASS W では最も演算量が多い．

## 4. 結　　果

### 4.1 逐次計算コードと並列計算コード

オーバーヘッドを示すため，図 2 に CRAY-T3E/1200E を用いた CLASS A での逐次計算コードの実行結果と並列計算コードを 1 PE での実行結果の性能差による並列化を示した．図の縦軸は Mop/s, 横軸は各ベンチマークとする．

CRAY-T3E/1200E では，IS を除いて逐次計算コードの方が，並列計算コードの方が多少性能が高い．逐次計算コードと並列計算コードでは，プログラムコードが同等ではないため，並列化のためのオーバーヘッドがあらわれている．CG と FT は，並列化のためのオーバーヘッドが低く，逐次計算コードと並列計算

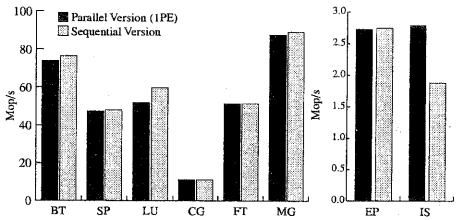

図 2　逐次コードと並列コードの性能差

コードの差が僅かである．IS は，並列計算コードの方が高速であり，メモリコピーに MPI_Alltoallv を用いる方が同等の機能を逐次計算コードによって実現するよりメモリのアクセス効率が高いと推測されるためである．.

### 4.2 CPU の性能差

CPU のクロックアップによる性能向上を示すために，CRAY-T3E と CRAY-T3E/1200E 上で並列計算コードを 1 PE で実行し，PE の単体性能の向上を検討する．各ベンチマークプログラムの結果を図 3 に示す．左縦軸は Mop/s，右縦軸は性能比 (Performance ratio)，

$$\text{Performance ratio} = \frac{\text{Mop/s/PE(CRAY-T3E/1200E)}}{\text{Mop/s/PE(CRAY-T3E)}} - 1.0 \quad (1)$$

横軸に各ベンチマークプログラムを示す．

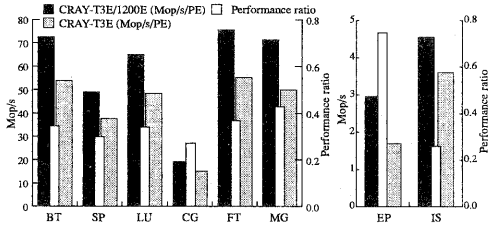

図 3　Computation Performance on the 1 PE (CLASS W)

EP は，メモリ帯域の影響が少ないため，PE 単体の性能向上が高い．EP は，CPU のクロックアップの効果が現れやすいベンチマークであると言える．CG と IS が他の場合と較べて，性能向上が 30 % 以下となり，性能向上が低い．この原因は，メインメモリ帯域の性能が上がっていないためと推測される．CG と IS を除いた場合では，FT と MG が 40 % 程度であり，BT，SP，LU が 35 % 程度となった．EP を除いた平均的な PE の性能向上は，およそ 35 % である．

CPU の演算性能が向上し，メインメモリの帯域が向上しない場合，データロードに対する影響が大きいベンチマークでは性能向上が難しいと考えられる．図 3 の結果からメモリ帯域の影響が比較的大きいと思われる CG と IS を用いてメモリ帯域の影響を検討する．図 4 に CG，図 5 に IS を示し，各 PE 数での問題の大きさ (CLASS) を横軸に，性能 (Mop/s/PE) を縦軸に取ったグラフを示す．

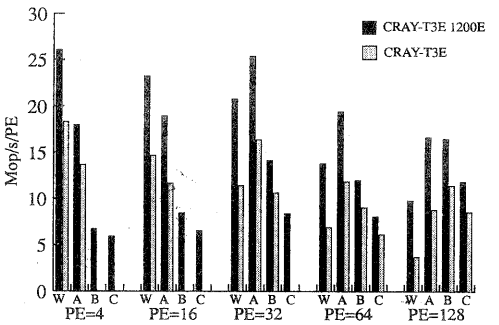

図 4　Kernel code CG

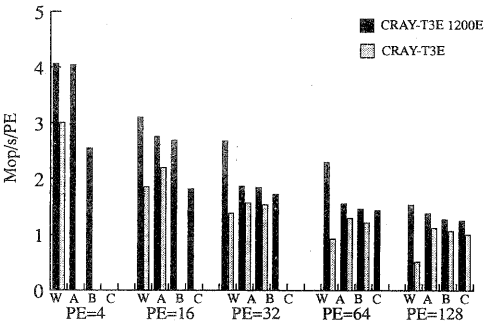

図 5　Kernel code IS

CG は，多くの場合で Mop/s/PE 値が W > A > B > C の順となり，問題サイズが大きくなるに従って性能が低下した．また，問題サイズと PE 数によって性能にピークがあらわれ，その位置は，CRAY-T3E と CRAY-T3E/1200E で一致した．性能ピークは，キャッシュの影響であると考えられる．キャッシュは，本報告のバージョンアップによって，容量は変化しないが，速度はクロック周波数分向上した．係数行列の配列データは，キャッシュに収まるデータサイズではない．しかし，CLASS A の 32 PE の場合ではワークベクトルの総量は二次キャッシュサイズに非常に近いサイズとなるため，ワークベクトルの演算に関しては，CPU のクロックアップが大きく影響したと考えられる．その結果，キャッシュが有効に機能している間は CRAY-T3E との性能差が非常に大きい．性能向

上は，キャッシュの影響が少ない問題サイズが大きい場合約 25 % である．

IS では，問題サイズと PE 数の増加に伴って性能は低下した．また，CLASS W では，CRAY-T3E と CRAY-T3E/1200E において逆の傾向が見られた．IS は，4.1 節で見られるように MPI ライブラリの性能に大きく依存したため，CRAY-T3E と CRAY-T3E/1200E での MPI ライブラリやコンパイラの性能差と考えることが出来る．CLASS W を除く問題サイズで性能向上は，約 20 % である．

### 4.3 PE 間通信性能

CPU の性能向上は前節で明らかになった．その結果を元に PE 間通信の性能向上を検討する．各ベンチマークの結果を並べるだけでは，CPU の性能向上と通信性能が両方含まれた状態があらわされる．CPU の性能比 ($P_{CPU}$) が明らかであるため，各ベンチマークを並列化する際に，演算量が前節の CLASS W と同程度になる問題サイズと PE 数を用いた場合の性能比 ($P_{tmp}$) から既知の CPU 性能向上の割合を相殺することで通信性能比 ($P_{comm}$) による全体性能の向上を類推することができる．

$$P_{comm} = P_{tmp} - P_{CPU} \qquad (2)$$

例えば，BT を CLASS A，16 PE で実行する場合の演算量が，BT の CLASS W，1 PE で実行する際の演算量とほぼ同一である場合，図 3 に示した CLASS W，1 PE の CPU の性能向上比 $P_{CPU}$ が得られ，CLASS A，16 PE を計算することで，CRAY-T3E と CRAY-T3E/1200E の性能比 $P_{tmp}$ を求めることが出来るため，$P_{tmp} - P_{CPU}$ を取ることで通信性能比 $P_{comm}$ を検討できる．

通信処理の振る舞いが明確なベンチマークプログラムを用いる．LU は，通信隠蔽処理を行なっているため，理論的には通信時間が陽にあらわれない．また，IS も MPI_Alltoallv の挙動が明確でない．また，EP は，並列実行時にも通信が発生しない．よって，PE 間通信性能を検討するため，LU，IS，EP を除いたベンチマークプログラムを用いる．用いた問題サイズは CLASS B とし，PE 数は CLASS W の 1 PE のメモリワークサイズと同程度の量となる数である．用いた PE 数は，BT で 100 PE，SP で 25 PE，CG で 64 PE，FT で 64 PE，MG で 64 PE である．図 に選択したベンチマークプログラムの各 PE 数での性能を示す．左縦軸に Mop/s/PE，右縦軸に Performance ratio，横軸に各ベンチマークプログラムと PE 数を示す．また，図 7 に通信性能比 $P_{comm}$ を示す．

FT の通信性能向上率が最も高い．FT は複素型の大きな配列データを多く通信する．通信帯域の向上が最も有効に働いたと考えられる．通信性能の向上率はおよそ通信帯域の性能向上率と等しい．その他の場合では，およそ 3 ～ 10 % 程度の通信性能向上が見られた．FT を除く場合では，通信回数が多いため，通信性能の多くは通信帯域ではなくネットワークのレイテンシーに依存するものと考えられる．平均的なアプリケーションプログラムを実行した場合の通信性能の向上は，およそ 5 % 前後と推測される．

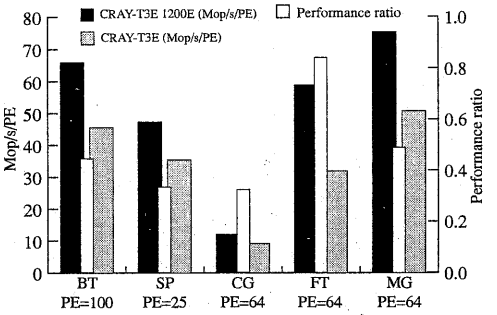


図 6 Communication Performance (CLASS B)

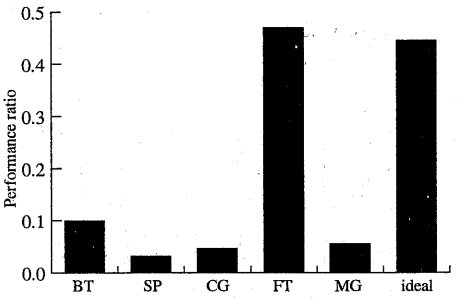


図 7 Communication Performance ratio

### 4.4 バージョンアップの効果

実際に科学技術計算のアプリケーションプログラムがどの程度の性能向上が得られるかを NPB によって示す．用いたベンチマークプログラムの種類は，得られる Mop/s 値の種類が同一 (Floating Point) であるものを用いた．図 8 に CLASS W を示し，図 9 に CLASS A を示す．図は，横軸に各ベンチマークと PE 数を示し，左縦軸に Mop/s/PE，右縦軸に CRAY-T3E と CRAY-T3E/1200E の性能比を示した．

前節までに PE の単体性能の性能向上と PE 間通信の性能向上を検討した．PE 単体の性能向上は，35 % 程度，通信の性能向上は 5 % 程度と推測された．NPB を用いたシステム全体の性能向上は，40 % 程度であると推測される．

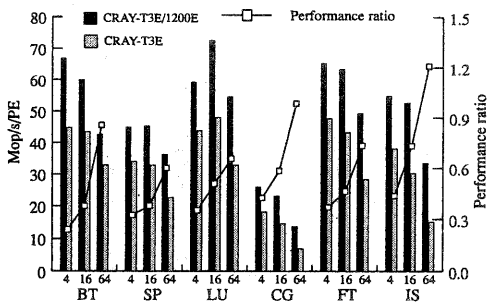


図 8 Performance for CLASS W

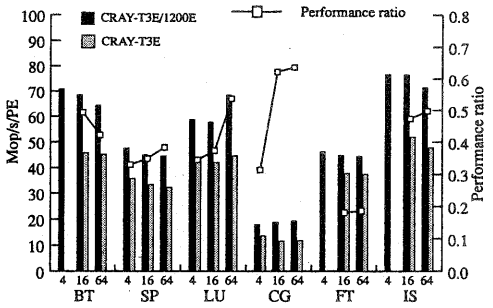


図 9 Performance for CLASS A

図 8 では，PE 数の増加に従って，性能比が右上がりに向上した．これは，図 4 で見られたキャッシュの影響が大きくなっていることが上げられる．PE 数が少ない場合，30 ～ 40 % 程度の性能比が見られた．

図 9 では，PE 数の増加によって CLASS W ほどの性能比の向上は見られない．CG，FT を除く性能比は，35 ～ 50 % 程度である．FT の性能比は一段低く，20 % 程度である．FT は，CLASS A では通信依存が大きくなるため，その程度の性能比になったと推測される．図 5 でも問題量が大きくなった場合にはその程度の性能比を示した．また，CG は，16，64 PE において高い性能比が得られた．これも 図 4 で見られたキャッシュの影響と見られる．

## 5. おわりに

本報告では，MPP システムの構成部品のバージョンアップというケースについて，その効果を NPB を用いて検討した．一般的に CPU のクロックアップによる性能向上は，メモリ帯域に大きく依存する．CRAY-T3E/1200E システムでは，PE の単体性能で多くのベンチマークコードで約 30 % 以上の性能向上が得られた．しかし，メモリ帯域に依存するベンチマークコードでは，30 % 以下の性能向上しか得られなかった．また，CPU のクロックアップによる性能向上の影響が大きいキャッシュにデータが収まる場合には，CPU のクロックアップ分の性能向上が得られた．

通信性能の向上を検討した場合，通信データ量が非常に大きい場合，通信性能の向上がそのままあらわれた．しかし，通信データ量が少ない場合では，3 ～ 10 % 程度の性能向上にとどまった．

MPP システムトータルとしての性能向上は，35 ～ 50 % 程度の性能向上が得られた．しかし，通信性能が大きく影響するアプリケーションでは，システムのトータル性能は，20 % 程度の向上にとどまった．

## 参 考 文 献

1) http://www.nas.nasa.gov/Software/NPB/
2) Browning,D.："NAS Kernels Survey Report", *Report RND-92-003* (1992)
3) Bailey,D., Barszcz,E., Barton,J., Browning,D., Carter,R., Dagum,L., Fatoohi,R., Fineberg,S., Frederickson,P., Lasinski,T., Schreiber,R., Simon,H., Venkatakrishnaan,V., Weeratunga,S.："THE NAS PARALLEL BENCHMARKS", *RNR Technical Report RNR-94-007* (1994)
4) http://www.cray.com/products/systems/crayt3e/