

## 6.12 NAS Parallel Benchmarks 2.0, MPI versions

NAS Parallel Benchmarks (NPB) 2.0 [7] currently includes five of the original eight benchmark problems, two of which are kernel benchmarks (FT and MG) and three which are computational fluid dynamics (CFD) application benchmarks (LU, SP, and BT). Results were obtained for the CFD application benchmarks. The benchmarks are based on FORTRAN 77 and the MPI message passing standard. Table 25 shows the various problem sizes for the NAS parallel benchmarks. DAISy runs the Class A problem size. Table 26 shows the standard operation count for the NPB with a Class A problem size and the MFLOPS results for the DAISy cluster, with the CRAY Y-MP/1 being the standard.

Benchmark Code	Class A	Class B	Class C
Embarrassingly Parallel (EP)	$2^{28}$	$2^{30}$	$2^{32}$
Multigrid (MG)	$256^3$	$256^3$	$512^3$
Conjugate Gradient (CG)	14000	75000	150000
3-D FFT PDE (FT)	$256^2 \times 128$	$512 \times 256^2$	$512^3$
Integer Sort (IS)	$2^{23}$	$2^{25}$	$2^{27}$
LU Solver (LU)	$64^3$	$102^3$	$162^3$
Pentadiagonal Solver (SP)	$64^3$	$102^3$	$162^3$
Block Tridiagonal Solver (BT)	$64^3$	$102^3$	$162^3$

Table 25. NAS Parallel Benchmarks Problem Sizes.

From D. Bailey, T. Harris, W. Saphir, R. van der Wijngaart, A. Woo, and M. Yarrow's "The NAS Parallel Benchmarks 2.0" [1995] [4].

LU is a simulated CFD application which uses symmetric successive over-relaxation (SSOR) to solve a block lower triangular-block upper triangular system of equations resulting from an un-factored implicit finite-difference discretization of the Navier-Stokes equations in three dimensions. SP and BT are simulated CFD applications that solve systems of equations resulting from an approximately factored implicit finite-difference discretization of the Navier-Stokes equations. BT solves block-tridiagonal systems of  $5 \times 5$  blocks; SP solves scalar pentadiagonal systems resulting from full diagonalization of the approximately factored scheme.

Benchmark Name	Nominal Size, Class A	CRAY Y-MP/1		p5-90/16 100baseT	
		Operation Count (x10 <sup>9</sup> )	MFLOPS	MFLOPS total	MFLOPS per process
Embarrassingly Parallel (EP)	2 <sup>28</sup>	26.68	211	NA	NA
Multigrid (MG)	256 <sup>3</sup>	3.905	176	NA	NA
Conjugate Gradient (CG)	14000	1.508	127	NA	NA
3-D FFT PDE (FT)	256 <sup>2</sup> x128	5.631	196	NA	NA
Integer Sort (IS)	2 <sup>23</sup>	0.7812	68	NA	NA
LU Solver (LU)	64 <sup>3</sup>	64.57	194	56.01	3.5
Pentadiagonal Solver (SP)	64 <sup>3</sup>	102	216	20.5	1.28
Block Tridiagonal Solver (BT)	64 <sup>3</sup>	181.3	229	73.77	4.61

Table 26. NAS Parallel Benchmarks Standard Operation Counts.  
From S. Saini, and D. H. Baile's "NAS Parallel Benchmark Results" [1995] [4].

### 6.12.1 Application Benchmark: LU

The LU benchmark code requires a power-of-two number of processors. A 2-D partitioning of the grid onto processors occurs by halving the grid repeatedly in the first dimensions, alternately  $x$  and then  $y$ , until all power-of-two processors are assigned, resulting in vertical pencil-like grid partitions on the individual processors. The ordering of point based operations constituting the SSOR procedure proceeds on diagonals which progressively sweep from one corner on a given  $z$  plane to the opposite corner of the same  $z$  plane, thereupon proceeding to the next  $z$  plane. Communication of partition boundary data occurs after completion of computational on all diagonals that contact an adjacent partition. This constitutes a diagonal pipelining method and is called a "wavefront" method. It results in a relatively large number of small communications of 5 words each. Table 27 shows the approximate sustained performance per dollar for the Class A LU benchmark.

Computer System	# of Proc.	Memory	Time in seconds	Ratio to CRAY Y-MP/1	List Price Million Dollars	Performance per Million Dollars	Date
CRAY Y-MP	1	NA	333.5	1	NA	NA	Aug-92
Convex SPP1000	32	4 GB	126	2.65	2.5	1.06	Mar-95
CRAY J916	16	2 GB	47.59	7.01	1.05	6.67	Jul-95
CRAY T3D	1024	64 MB/PE	7.09	47.04	3.6	13.07	Mar-95
DEC Alpha Server 8400 5/300	12	2 GB	79.13	4.21	0.718	5.87	Oct-95
IBM RS/6000 SP Wide-node1 (67Mhz)	128	128 MB/PE	15.2	21.94	5.08	4.32	Mar-95
IBM RS/6000 SP Wide-node2 (77Mhz)	64	128 MB/PE	19.2	17.37	5.74	3.03	Oct-95
IBM RS/6000 SP Thin-node2 (67Mhz)	128	64MB/PE	15.9	20.97	3.48	6.03	Mar-95
SGI PC XL (75Mhz)	16	2 GB	65.3	5.11	0.895	5.71	Jun-94
SGI PC XL (90Mhz)	16	2 GB	65.9	5.06	1.02	4.96	May-95
DAISy	16	64 MB/node	2897.49	0.12	0.06	1.92	Nov-95

Table 27. Approximate sustained performance per dollar for Class A LU benchmark.  
From S. Saini, and D. H. Baile's "NAS Parallel Benchmark Results" [1995] [4].

### 6.12.2 Application Benchmark: SP and BT

The SP and BT algorithms have a structure similar to the LU algorithm: Each solves three sets of uncoupled systems of equations, first in the  $x$ , then in the  $y$ , and finally in the  $z$  direction. These systems are scalar pentadiagonal in the SP code, and block triadiagonal with 5x5 blocks in the BT code.

The implementations of the SP and BT solve these systems using a multi-partition scheme. In the multi-partition algorithm each processor is responsible for several disjoint sub-blocks of points ("cells") of

the grid. The cells are arranged such that for each direction of the line solve phase the cells belonging to a certain processor will be evenly distributed along the direction of solution. This allows each processor to perform useful work throughout a line solve, instead of being forced to wait for the partial solution to a line from another processor before beginning work. Additionally, the information from a cell is not sent to the next processor until all sections of linear equation systems handled in this cell have been solved. Therefore, the granularity of communications is kept large and fewer messages are sent.

Both the SP and BT codes require a square number of processors. Tables 28 and 29 show the approximate sustained performance per dollar for Class A SP and BT benchmarks respectively.

Computer System	# of Proc.	Memory	Time in seconds	Ratio to CRAY Y-MP/1	List Price Million Dollars	Performance per Million Dollars	Date
CRAY Y-MP	1	NA	471.5	1	NA	1	Aug-92
Convex SPP1000	64	4 GB	102	4.62	2.5	1.84	Mar-95
CRAY J916	16	2 GB	77.54	6.08	1.05	5.79	Jul-95
CRAY T3D	1024	64 MB/PE	5.41	87.15	3.6	24.21	Mar-95
DEC Alpha Server 8400 5/300	12	2 GB	102.75	4.59	0.718	6.39	Oct-95
IBM RS/6000 SP Wide-node1 (67Mhz)	128	128 MB/PE	18.7	25.21	5.08	4.96	Mar-95
IBM RS/6000 SP Wide-node2 (77Mhz)	64	128 MB/PE	26.46	17.82	5.74	3.1	Oct-95
IBM RS/6000 SP Thin-node2 (67Mhz)	128	64MB/PE	20.6	22.89	3.48	6.58	Mar-95
SGI PC XL (75Mhz)	16	2 GB	67.2	7.02	0.895	7.84	Jun-94
SGI PC XL (90Mhz)	16	2 GB	63.18	7.46	1.02	7.32	May-95
DAISy	16	64 MB/node	3883.83	0.12	0.06	2.02	Nov-95

Table 28. Approximate sustained performance per dollar for Class A SP benchmark. From S. Saini, and D. H. Baile's "NAS Parallel Benchmark Results" [1995] [4].

Computer System	# of Proc.	Memory	Time in seconds	Ratio to CRAY Y-MP/1	List Price Million Dollars	Performance per Million Dollars	Date
CRAY Y-MP	1	NA	792.4	1	NA	1	Aug-92
Convex SPP1000	64	4 GB	78	10.16	2.5	4.06	Mar-95
CRAY J916	16	2 GB	98.8	8.02	1.05	7.64	Jul-95
CRAY T3D	1024	64 MB/PE	4.56	173.77	3.6	48.27	Mar-95
DEC Alpha Server 8400 5/300	12	2 GB	103.47	7.66	0.718	10.67	Oct-95
IBM RS/6000 SP Wide-node1 (67Mhz)	128	128 MB/PE	20.1	39.42	5.08	7.76	Mar-95
IBM RS/6000 SP Wide-node2 (77Mhz)	64	128 MB/PE	29.01	27.31	5.74	4.76	Oct-95
IBM RS/6000 SP Thin-node2 (67Mhz)	128	64MB/PE	20.8	38.1	3.48	10.95	Mar-95
SGI PC XL (75Mhz)	16	2 GB	91.8	8.63	0.895	9.64	Jun-94
SGI PC XL (90Mhz)	16	2 GB	80.2	9.88	1.02	9.69	May-95
DAISy	16	64 MB/node	2641.61	0.30	0.06	5.0	Nov-95

Table 29. Approximate sustained performance per dollar for Class A BT benchmark. From S. Saini, and D. H. Baile's "NAS Parallel Benchmark Results" [1995] [4].

### 6.13 Parallel Seismic Inverse Problem

The DAISy cluster has been used to calculate an inverse problem in seismic tomography. The project goal [28] is to demonstrate a parallel seismic inverse code that runs scalably on inexpensive IBM compatible platforms, incorporating a modular design that separates the parallel algorithm from the specific model used for seismic imaging. The seismic data generated by means of impacts on the earth's surface, consists of timings between generation and reception. The data can be inverted through a tomographic scheme to give a three-dimensional picture of the local rock velocity.

The algorithm is a hybrid of bisection ray tracing and a P-wave Huygens' principle approach and parallelizes in an embarrassingly parallel manner. The algorithm has an adjustable parameter that controls the resolution of the resulting 3D velocity distribution. High resolutions will require ~1 sec between communications, while lower resolutions require ~.01 sec. Though, this code is embarrassingly parallel, it is ideal to test the sensitivity of the cluster to network latency.

Figure 19 shows a three-dimensional rendered image of the subterranean galleries of the "Lucky Friday" silver mine located in Northern Idaho. For acceptable tomographic feature prediction 1 sec to .1 sec is required per task (on the DAISy 90 MHz Pentium). This is useful as a check on the inverse model because the topography of the mine tunnels are measured.

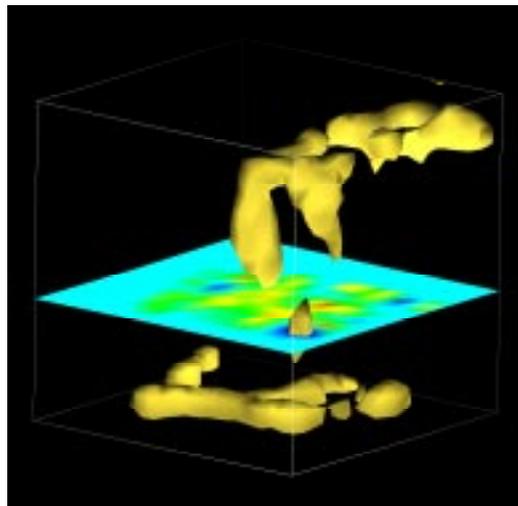


Figure 19. Parallel Seismic Inverse Model. This is the tomographic rendering from seismic data for the "Lucky Friday" silver mine in Northern Idaho. The gold features accurately predict the known locations of the mine galleries. The blue plane is an orthogonal slice through the observation volume. Colors on this plane indicate the effective "sound" velocity of the rock: red is faster; blue is slower.

Figure 20. shows the execution time for the parallel seismic inverse model on various platforms. All runs used the same source code and the GNU C++ compiler (G++) for the native OS without optimization. The PC cluster performs admirably against the considerably more costly workstations.

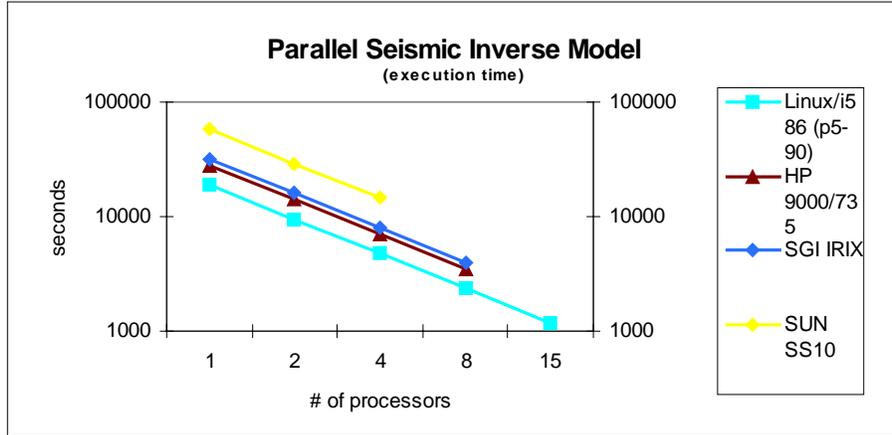


Figure 20. Execution time for Parallel Seismic Inverse Model.