Performance of the Bryan-Fritsch numerical model on parallel computers

Adapted from: Bryan, G. H., 2002: Analysis of the thermodynamic structure of squall lines as revealed by numerical simulations. Ph.D. thesis, *The Pennsylvania State University*.

2.8 Evaluation of parallel performance

The numerical model was tested on two parallel computing systems. One is a Linux cluster at the Center for Academic Computing at The Pennsylvania State University, hereafter referred to as Lion-X. This machine has sixty-four 500 MHz Intel processors in thirty-two dual-processor nodes. There are two network devices to handle inter-processor communication: a fast ethernet switch, which can transfer data at 100 Mb/s; and a Myricom Myrinet switch, rated at 1.28 Gb/s. Results from both network devices are presented in this section.

The second computing device is an IBM SP cluster at the National Center for Atmospheric Research, hereafter referred to as Blackforest. This machine has more than 1,000 Power-3 IBM processors, each with a clock speed of 375 MHz. The Power-3 processors can perform up to 4 operations per cycle, and hence have an overall speed up to four times greater than the Intel processors of the Linux cluster. The communications device of Blackforest is rated at more than 1 Gb/s. Blackforest is composed of 4-processor nodes with 2 GB of shared memory per node. Based on this hardware setup, two tests were performed on Blackforest: one in which MPI was used exclusively to transfer information between processors; and a second in which OpenMP was used to parallelize intra-node processing, and MPI was used to transfer data between nodes.

2.8.1 Fixed-domain test

Two tests were used to evaluate the parallel performance of the numerical model. The first is a fixed-domain test, in which the number of grid points in the domain is kept constant as processors are added. For this test, as the number of processors is increased, each processor has fewer grid points. The goal of using more processors in this setup is to decrease the time

required to run the simulation; hence, this type of test would be important for real-time modeling centers, such as the National Centers for Environmental Prediction.

A two-hour simulation of a supercell thunderstorm was used for the fixed-domain test. The domain has 128x128 grid points in the horizontal and 40 vertical levels for runs on blackforest, and 120x120x40 grid points for runs on Lion-X. The initial conditions are identical to those used by Weisman and Rotunno (2000), except only the "1/4 circle" hodograph case is considered. Timing results are listed in Table 2.3. The results are also presented graphically in Fig. 2.17, where "speedup" is the ratio of time required to run the simulation on one processor to the time required to run on *n* processors,

$$speedup = \frac{\text{time on 1 processor}}{\text{time on } n \text{ processors}} \ .$$

For a perfectly parallelized code, the speedup would be identical to the number of processors used, i.e., ideal performance on Fig. 2.17a would be a straight line with slope of 1 (the thick gray line). Results from Lion-X (the solid lines in Fig. 2.17) show a clear advantage of the faster Myrinet communication device over the roughly 10 times slower ethernet device. In fact, results are nearly perfect out to 32 processors, and excellent out to 60 processors. A plot of parallel efficiency is provided in panel 2.17b, where

$$parallel\ efficiency = 100 \times \frac{speedup}{n} \ .$$

Ideally, a parallel efficiency of 100% is desired. The results using the Myrinet communication device show greater than 90% efficiency on Lion-X, which is an exceptionally good result.

Results on Blackforest do not show the same spectacular results: in fact, parallel efficiency is between 50%-70% with 64 processors, and drops to about 40% with 128 processors. Further analysis of the runs on Blackforest reveals that the model spends a greater percentage of time waiting for communications to complete as compared to runs on Lion-X, despite the fact that the communication device of Blackforest is rated at about the same speed as the Myrinet device of Lion-X. However, the processors of Blackforest are considerably faster (almost 4 times as fast) as those of Lion-X. It is concluded that Lion-X shows superior parallel performance *for this particular test* since communications on Lion-X are completed while calculations are being performed (since non-blocking communications are used with MPI to overlap communications and computations); in contrast, the faster Blackforest processors

complete the computations in considerably less time – and before communications are complete – and then sit idle until communications finish. It could therefore be concluded that this experiment was not ideally designed to test the advantages of the IBM-SP, and perhaps the parallel performance of Blackforest in Fig. 2.17 is underrated. Furthermore, it is noted that the time required to complete this simulation with 128 processors on Blackforest is about 3 minutes (see Table 2.3); clearly, this number of processors is unnecessary for this particular simulation. Overall however, the parallel performance of the code can be considered successful. Note, for example, that greater than 90% parallel efficiency is achieved with 16 processors on Blackforest.

A comparison of the MPI-only approach (long dashed lines in Fig. 2.17) to that using both MPI and OpenMP simultaneously (short dashed lines) reveals greater performance for the MPI-only approach. This result was surprising, since it was expected that the lower communication requirements of the OpenMP approach would result in enhanced performance. It was found that using OpenMP for shared-memory parallelization within the 4-processor nodes of Blackforest caused the decrease in performance. It is unclear, at this point, why greater efficiency could not be achieved using OpenMP, and further study will be carried out in the future.

2.8.2 Scaled-domain test

The second type of test used to evaluate the parallel performance of the numerical model is a "scaled-domain" test, in which each processor has the same number of grid points as the number of processors are increased. Thus, the overall domain during this test becomes larger as processors are added, but the work required *per processor* remains roughly the same. It is important to note that a scaled-domain test is more relevant to this study than the fixed-domain test presented in the last section, because the goal is to simulate squall lines with very high resolution, which requires very large numbers of grid points (e.g., 10^7 to 10^8 grid points).

The sub-domains for this test each have 100x100x72 grid points; this is roughly the number of grid points per processor that is required to simulate mesoscale convective systems (e.g., squall lines) with very high resolution. For example, using these sub-domain dimensions with a grid spacing of 125 m and with 128 processors results in a numerical domain with horizontal dimensions of 200 km x 100 km and a vertical extent of 18 km; such a domain is suitable for the simulation of deep convective phenomena (squall lines, supercells, bow echoes).

Finally, it is noted that this number of grid points requires about 200 MB of memory per processor, which is about half of the maximum memory available per processor on Blackforest.

Timing results for a model integration of 100 time steps are presented in Table 2.4 and Fig. 2.18. For this test, an ideal result would be an identical time as the number of processors are increased, or a perfectly horizontal line on Fig. 2.18 (see thick gray lines). Results for both Lion-X and Blackforest show a marked increase from one processor to four processors, and then nearly constant timing results after four processors. It is assumed that the decrease in performance from one to four processor occurs because of the increased memory access requirements; in other words, the four processor runs have more grid points *per node* than the one processor runs. Beyond four processors, the number of grid points per node remains the same.

Considering only the performances for at least four processors, the results using Lion-X show a degradation in performance out to 60 processors. Specifically, the 60 processor simulations require about 10% more time than the 4 processor simulations. This performance is considered acceptable, and can be used as a benchmark to estimate timing for longer simulations. For Blackforest, the performance is nearly ideal from 4 processors to 128 processors. (The increase in performance from 4 to 16 processors for the MPI-only run may be an anomaly.) This result proves the utility of the numerical model for extremely large domain simulations (the 128 processor run has 9.2×10^7 grid points), as well as the robustness of the parallelization technique.

Table 2.3. Timing results from the "fixed-domain" test. The number of horizontal grid points per processor (g.p./proc.) is shown, along with the time (in seconds) required to complete the simulation.

processors	Lion-X, ethernet		Lion-X, myrinet		Blackforest MPI		Blackforest MPI-OMP	
	g.p/proc.	time (s)	g.p/proc.	time (s)	g.p/proc.	time (s)	g.p/proc.	time (s)
1	120x120	28091.7	120x120	26903.3	128x128	9384.7	128x128	9393.0
4	60x60	7022.6	60x60	6726.3	64x64	2536.6	64x64	2564.3
8	60x30	3795.8	60x30	3374.6	64x32	1165.7	64x32	1262.3
16	30x30	2201.1	30x30	1723.5	32x32	634.9	32x32	655.3
24	30x20	1569.1	30x20	1134.2				
32	30x15	1219.4	30x15	8544.6	32x16	356.2	32x16	383.4
48	20x15	941.2	20x15	591.4				
60	20x12	782.8	20x12	460.8				
64					16x16	221.6	16x16	273.8
128					16x8	175.7	16x8	188.4

Table 2.4. Timing results from the "scaled-domain" test. The number of horizontal grid points per processor (g.p./proc.) is shown, along with the time (in seconds) required to complete the simulation.

Processors	Lion-X, ethernet		Lion-X, myrinet		Blackforest MPI		Blackforest MPI-OMP	
	g.p/proc.	time (s)	g.p/proc.	time (s)	g.p/proc.	time (s)	g.p/proc.	time (s)
1	100x100	3030.7	100x100	3073.9	100x100	1335.8	100x100	1337.7
4	100x100	3687.0	100x100	3602.8	100x100	2486.0	100x100	2350.0
16	100x100	3792.4	100x100	3665.4	100x100	2428.4	100x100	2389.2
32	100x100	3822.6	100x100	3695.0	100x100		100x100	
60	100x100	4065.9	100x100	3860.1	100x100		100x100	
64	100x100		100x100		100x100	2465.0	100x100	2485.5
128	100x100		100x100		100x100	2552.9	100x100	2581.1

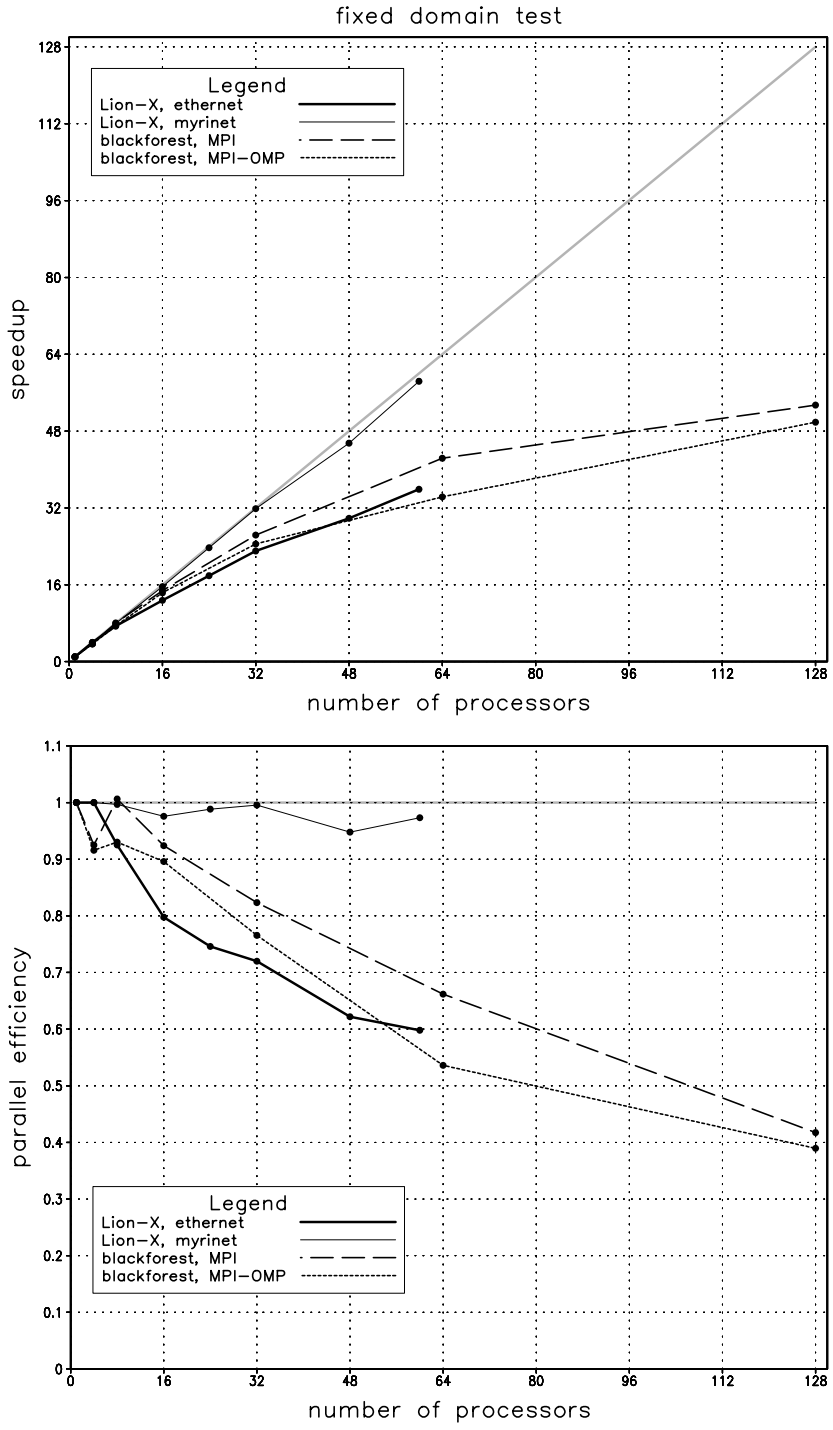

Fig. 2.17. Parallel performance for the "fixed-domain" test: (a, top) speedup versus number of processors; and (b, bottom) parallel efficiency for the same data. The thick gray line represents ideal performance; the thick solid line is output from the Linux cluster (Lion-X) using ethernet communications; the thin solid line is output from Lion-X using myrinet communications; the long dashed line is output from the IBM-SP (Blackforest) using only MPI; and the short dashed line is output from Blackforest using OpenMP within nodes and MPI between nodes.

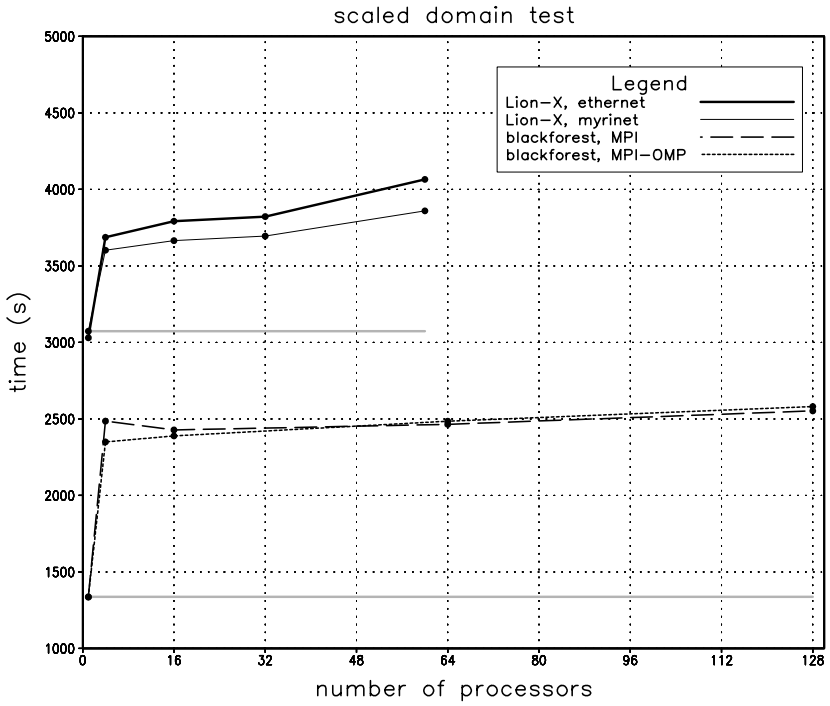

Fig. 2.18. Parallel performance for the "scaled-domain" test, in which the number of grid points per processor remains the same. Total time for the simulations is plotted versus number of processors used. Line styles are the same as for Fig. 2.17.