Parallel Navier-Stokes Computations on Shared and Distributed Memory Architectures

M. Ehtesham Hayder
Institute for Computational Mechanics in Propulsion
Lewis Research Center
Cleveland, Ohio

D.N. Jayasimha
Ohio State University
Columbus, Ohio

and

Sasi Kumar Pillay
Lewis Research Center
Cleveland, Ohio

Prepared for the
33rd Aerospace Sciences Meeting and Exhibit
sponsored by the American Institute of Aeronautics and Astronautics
Reno, Nevada, January 9–12, 1995
Parallel Navier-Stokes Computations on Shared and Distributed Memory Architectures

M. Ehtesham Hayder
Institute for Computational Mechanics in Propulsion
Ohio Aerospace Institute
NASA Lewis Research Center, Cleveland, OH

D. N. Jayasimha*
Department of Computer Science
The Ohio State University, Columbus, OH

Sasi Kumar Pillay
Chief, Computer Graphics Branch
Computer Services Division
NASA Lewis Research Center, Cleveland, OH

Abstract

We study a high order finite difference scheme to solve the time accurate flow field of a jet using the compressible Navier-Stokes equations. As part of our ongoing efforts, we have implemented our numerical model on three parallel computing platforms to study the computational, communication, and scalability characteristics. The platforms chosen for this study are a cluster of workstations connected through fast networks (the LACE experimental testbed at NASA Lewis), a shared memory multiprocessor (the Cray YMP), and a distributed memory multiprocessor (the IBM SP1). Our focus in this study is on the LACE testbed. We present some results for the Cray YMP and the IBM SP1 mainly for comparison purposes. On the LACE testbed, we study: (a) the communication characteristics of Ethernet, FDDI and the ALLNODE networks and (b) the overheads induced by the PVM message passing library used for parallelizing the application. We demonstrate that clustering of workstations is effective and has the potential to be computationally competitive with supercomputers at a fraction of the cost.

Key words: Navier-Stokes computations, parallel computing, cluster computing.

1. Introduction

Numerical simulations can play an important role in examining the physical processes associated with many important practical problems. The suppression of jet exhaust noise is one such problem which will have a great impact on the success of the High Speed Civil Transport plane (HSCT). The radiated sound emanating from the jet can be computed by solving the full (time-dependent) compressible Navier-Stokes equations. This computation can, however, be very expensive and time consuming. The difficulty can be partially overcome by limiting the solution domain to the near field where the jet is nonlinear and then using acoustic analogy [e.g., Lighthill\(^1\)] to relate the far-field noise to the near-field sources. The latter technique requires obtaining the time-dependent flow field. In this study we concentrate on such flow fields near the nozzle exit.
The advent of massively parallel processors and clustered workstations gives the opportunity for scientists to parallelise their computationally intensive codes and reduce turnaround time. Recognising this, a number of researchers\(^2\)\(^3\)\(^4\)\(^5\) have studied CFD applications on specific parallel architectures. Our goal is to implement a particular CFD application code which uses a high order finite difference scheme on a variety of parallel computing platforms to study the computational, communication, and scalability characteristics. The platforms chosen for this study are a cluster of workstations connected through fast networks (the LACE\(^6\) experimental testbed at NASA Lewis), a shared memory multiprocessor (the Cray YMP), and a distributed memory multiprocessor (the IBM SP1). Clustered workstations such as the LACE are becoming increasingly popular because they show promise as a low cost alternative to expensive supercomputers and massively parallel processors. We have followed a pragmatic approach by parallelising only those segments of the algorithm which are computationally intensive (identified using a profiler). Such an approach can reduce the time for parallelising "dusty deck" sequential codes. Our distributed memory implementation on the LACE uses PVM\(^7\), a portable message passing library developed at the Oak Ridge National Laboratory.

In the next two sections we briefly discuss the governing equations and the numerical model. Section 4 has a discussion of the parallel architectures used in the study and the parallelisation schemes. In Section 5 we present an analytical model for computation and communication costs and discuss simulation results on the three platforms used in this study.

2. Governing Equations

We solve the Navier-Stokes and the Euler equations to compute flow fields of an axisymmetric jet. The Navier-Stokes equations for such flows in polar coordinates can be written as

\[
F = r \begin{pmatrix}
\rho u \\
\rho u^2 - \tau_{zz} + p \\
\rho u v - \tau_{vr} \\
\rho u v - u \tau_{zz} - v \tau_{vr} - \kappa T_z
\end{pmatrix}
\]

\[
G = r \begin{pmatrix}
\rho v \\
\rho uv - \tau_{rr} \\
\rho v^2 - \tau_{rr} + p \\
\rho v H - u \tau_{rr} - v \tau_{rr} - \kappa T_r
\end{pmatrix}
\]

\[
S = \begin{pmatrix}
0 \\
0 \\
p - \tau_{r\theta}
\end{pmatrix}
\]

\[F \text{ and } G \text{ are the fluxes in the } x \text{ and } r \text{ directions respectively, and } S \text{ is the source term that arises in the cylindrical polar coordinates, and } \tau_{ij} \text{ are the shear stresses. One obtains the Euler equations from the above equations, by setting } \kappa = 0.\]

3. Numerical Model

We use the fourth order MacCormack scheme, due to Gottlieb and Turkel\(^8\), to solve the Navier-Stokes and the Euler equations. For the present computations, the operator \(L\) in the equation \(LQ = S\) or equivalently \(Q_t + F_z + G_r = S\) is split into two one-dimensional operators and the scheme is applied to these split operators. For the one dimension model/split equation \(Q_t + F_z = S\), we express the predictor step with forward differences as

\[Q_i^{n+1} = \frac{1}{2}[Q_i^n + Q_i^n + \frac{\Delta t}{6\Delta x}\{7(F_{i+1} - F_i^n) - (F_{i+2} - F_{i+1}^n)\} + \Delta t S_i]\]

and the corrector step with backward differences as

\[Q_i^{n+1} = \frac{1}{2}[Q_i^n + Q_i^n + \frac{\Delta t}{6\Delta x}\{7(F_{i-1} - F_{i-2}) - (F_{i-1} - F_{i-2})\} + \Delta t S_i]\]

This scheme becomes fourth-order accurate in the spatial derivatives when alternated with symmetric variants\(^8\). We also use the standard MacCormack scheme in our tests. This scheme is second order accurate in both space and time and can be written as

\[Q_i^{n+1} = \frac{1}{2}[Q_i^n + Q_i^n + \frac{\Delta t}{\Delta x}(F_{i+1} - F_{i-1}) + \Delta t S_i]\]

We define \(L_1\) as a one dimensional operator with a forward difference in the predictor and a backward difference in the corrector. Its symmetric variant \(L_2\) uses a backward difference in the predictor and a forward difference in the corrector. For our computations, the one dimensional sweeps are arranged as
\[ Q^{n+1} = L_{1x}L_{1y}Q^n \]
\[ Q^{n+2} = L_{2x}L_{2y}Q^{n+1} \]

This scheme is used for the interior points. In order to advance the scheme near boundaries the fluxes are extrapolated outside the domain to artificial points using a cubic extrapolation to compute the solution on the boundary. At the outflow boundary, we then solve the following set of equations to get the solution at the new time for all boundary points.

\[
\begin{align*}
    p_t - \rho c u_t &= 0 \\
    p_t + \rho c u_t &= R_3 \\
    p_t - \rho c v_t &= R_3 \\
    v_t &= R_4
\end{align*}
\]

where \( R_i \) is determined by which variables are specified and which are not. Whenever, the combination is not specified, \( R_i \) is just those spatial derivatives that come from the Navier-Stokes equations. Thus, \( R_i \) contains viscous contributions even though the basic format is based on inviscid characteristic theory. This framework of outflow boundary condition implementation is discussed by Hayder and Turkel. Due to the presence of a singularity, special care is needed to treat the centerline. Here, we briefly outline one of the formulations for the centerline used in three dimensional calculations. In this treatment, equations corresponding to the Cartesian coordinates are solved at the centerline. Away from the centerline variables are solved in polar coordinates. We use \( r^2 = x^2 + y^2 \), \( \tan \theta = y/x \), \( z = r \cos \theta \) and \( y = r \sin \theta \) to derive the equations at the centerline. The conservation equation for the Cartesian variables is

\[ W_t + F_x + G_y + H_z = 0 \]

or

\[ W_t + \frac{(Fy_\theta - Gx_\theta)}{r} + \frac{(Gx_r - Fy_r)\theta}{r} + H_z = 0 \]

or

\[ (rW)_t + (r\tilde{F})_x + (\tilde{G})_\theta + (rH)_z = 0 \]

where \( \tilde{F} = F \cos \theta + G \sin \theta \) and \( \tilde{G} = G \cos \theta - F \sin \theta \). Because both \( F \) and \( G \) are single valued, \( F_\theta \) and \( G_\theta \) are zero at the centerline. Then using the L'Hospital rule at \( r=0 \) we get

\[ W_t + (F + G_\theta) \cos \theta + (G - F_\theta) \sin \theta + H_z = 0 \]

The above equation is solved at the centerline in three dimensional calculations. For axisymmetric calculations, we have options to use the L'Hospital rule or extrapolations to get the quantities at the centerline. Our numerical model is three dimensional. However, to maintain low computational costs, in this study we obtained only two dimensional results. Further discussion of our numerical model is given in Hayder et al. and Mankbadi et al.11

4. Parallel Implementations

4.1 Network of Workstations

The LACE (Lewis Advanced Cluster Environment) is a cluster of 33 IBM RS6000 workstations (node0 – node32). Node 0 is the file server. All the nodes are connected through two Ethernet networks (10 Mbits/sec). In addition, nodes 00–16 are connected through the much faster FDDI interface (100 Mbits/sec) and nodes 17–32 are connected through IBM's ALLNODE prototype switch (capable of a peak throughput of 512 Mbits/sec). The ALLNODE switch is an Omega interconnection network capable of providing multiple contentionless paths between the nodes of the cluster.

Communication among the processors is through message passing. We used the popular PVM (Parallel Virtual Machine) message passing library (version 3.2.2) to implement our parallel program. We partitioned the domain along the \( z \) direction only. Our partitioning scheme was static since there was no explicit need to keep long vectors given the architecture of the RS6000. The LACE test bed is upgraded periodically. Our results are for RS6000/560 processors. We will report results with upgraded hardware and software in subsequent studies.

4.2 Shared Memory Architecture

We used the Cray YMP/8, which has eight vector processors, for this study. The Cray YMP/8 has a peak rating of approximately 2.7 GigaFLOPS. It offers a single address space and the communication between processes executing on different processors is through shared variables. Using DOALL directives, we partitioned the domain along the orthogonal direction of the sweep to keep the vectors lengths large and to avoid non-stride access for most of the variables.
4.3 Distributed Memory Multiprocessor

In this study, we also used the IBM SP1 which is a distributed memory, message passing multiprocessor. We would like to point out that our test bed is the machine at NASA Lewis in which each processor is a RS6000/370. This system has the original SP1 processors and software upgrades for SP2. The nodes of the SP1 are interconnected through a variation of the Omega network. This network permits multiple contentionless paths between nodes like the ALLNODE switch. We used the MPL library for parallel implementation of our numerical model on the SP1.

5. Results

5.1 Analytical Model

We will first provide a model to estimate the computation and communication costs. Note that in one time step there is one corrector and one predictor step in each of the coordinate directions. Each of the predictor and corrector steps has a forward and a backward sweep. Let us examine the communication cost for two dimensional computations with $P_z$ partitions in the $z$ direction. We divide the cost of data communications across partition boundaries into two parts, $C_s$ and $C_f$. $C_s$ is associated with the communication needed to compute stresses or heat fluxes at the boundary points. This will be zero for the Euler calculations. $C_f$ is associated with the transfer of fluxes at the domain boundaries, as required by the forward and the backward sweeps. This would be same for both the Euler and the Navier-Stokes calculations. For this partitioning, we get

$$C_s, z = 2n_s [N_r t_z + t_s]$$

$$C_f, z = n_s [S_c N_r t_z + t_s]$$

$$C_s, r = 2 (n_r - 1) [N_r t_z + t_s]$$

$$C_f, r = 0$$

where $n_s$ is the number of neighboring variables needed for stress and heat flux computations. Subscripts $z$ and $r$ in the above set of equations refer the direction of the sweep. While velocities at neighboring points in both directions are needed for stress computations, temperatures in only the sweep direction are needed for heat flux computations. Thus, for two dimensional computations, $n_s$ is three in the sweep direction and two in the other direction. Let $N_r$ be the number of grid points in the radial direction, $t_z$ and $t_r$ be the startup and vector transfer rates (time per element) respectively, $n_s$ be the number of conservation variables (4 in two dimensions), and $S_c$ be the width of the stencil measured from the point where variables are updated. This is one for the standard (second order) MacCormack scheme and 2 for the fourth order extension. $S_c$ would increase for higher order schemes. The coefficient two in above equations for $C_s$ comes from the fact that data exchanges in both directions across the partition boundary are needed for stress computations. On the contrary, since the algorithm uses one sided differences of fluxes, the communications of fluxes are only one direction at any time. The total communication times for one time step per processor is then

$$T_{cm} = 2[4n_s - 2(N_r t_z + t_z) + n_s (S_c N_r t_z + t_z)]$$

This is expected to rise as the surface area in the partitions increase. It is easily seen that the total computation time per processor is

$$T_{cp} = \frac{N_r}{P_z} N_z T_c$$

where $T_c$ is the computation time per grid point and $N_z$ is the total number of grid points in the $z$ direction. It should be noted that as the size of the problem changes, $T_c$ could also change based on the architecture of the machine. As the size of the problem increases, there will be an increased amount of cache transfers and misses. This would increase the overhead associated with the computation. The total time spent per processor is then

$$T < T_{cm} + T_{cp}$$

The inequality is a result of some overlapped computations and communications. As the number of processor ($P_z$) increases, $T_{cm}$ will become significant compared to $T_{cp}$.

5.2 Numerical results

We consider a jet with the mean inflow profile

$$\bar{U}_r = U_\infty + (U_c - U_\infty)g_r$$

$$\bar{T}_r = T_c + (T_\infty - T_c)g_r + \frac{\gamma - 1}{2} M_c^2(1 - g_r)g_r$$

$$g_r = \frac{1}{2} [1 + \tanh \left(\frac{r - \gamma}{\theta}\right)]$$

where $\theta$ is the momentum thickness. The subscripts $c$ and $\infty$ refer to the centerline and free stream values respectively. At inflow, we assume the radial velocity is zero and the static pressure is constant. The standard size of our computational domain is 50 radii in the axial direction and 5 radii in the radial direction. We excite the inflow profile at location $r$ and time $t$ as

$$U(r, t) = \bar{U}(r) + \epsilon Re(\bar{U} e^{i \omega_S t})$$
\[ P(r, t) = \hat{P}(r) + \epsilon \text{Re}(\hat{\rho} e^{i\nu S_i t}) \]
\[ \rho(r, t) = \hat{\rho}(r) + \epsilon \text{Re}(\hat{\rho} e^{i\nu S_i t}) \]
\[ V(r, t) = \epsilon \text{Re}(\hat{V} e^{i\nu S_i t}) \]

\( \hat{U}, \hat{V}, \hat{\rho} \) and \( \hat{P} \) are the eigenfunctions of the linearized equations with the same mean flow profile, \( \epsilon \) is the excitation level and \( S_i \) is the Strouhal number.

We consider a case with \( \mathcal{U}_x = \frac{1}{4}, \mathcal{U}_z = \frac{1}{2}, \) momentum thickness, \( \theta = \frac{1}{8} \) and Strouhal number, \( S_i = \frac{1}{4} \). The jet center Mach number is 1.5 while the Reynolds number based on the jet diameter is 1.2 million. Our present mean flow and eigen function profiles are same as those in Scott et al. In Figure 1 we show a contour plot of axial momentum from the solution of the Navier Stokes equations. A grid of size 100x250 was used in this computation. This result was obtained after about 16,000 time steps. For all other results in this paper, we simulated the same physical problem on the same grid, but ran for 5000 time steps.

We optimised our code on the LACE to reduce computation time. In the process of optimization, we developed several versions of our code. We will refer the original version of the code as version-1. In version-2, we replaced exponentiations in the code by multiplications (i.e. \( a^2 \) is replaced by a times \( a \), etc.). Then we rearranged dimensions of the arrays to ensure \( \text{stride-1} \) access whenever possible. We call the modified code version-3. We also reduced number of divisions and used a single COMMON block instead of many COMMON blocks. We will refer these versions as version-4 and version-5 respectively. In Figure 2 we compare the execution times of different versions of our numerical models on a single processor. It can be easily noted that \( \text{stride-1} \) access was the single most important optimization and it reduced the computation time considerably. Our partitioning scheme resulted in a good load balance among the processors. In Figure 3, we show the computation times in different processors for our solutions of the Navier-Stokes equations on sixteen processors of the IBM SP1. Results shown in this figure include the computation time and the execution time for the MPL library calls, but exclude any communication wait or transfer time. To study communication characteristics in distributed computing, we developed two additional versions of our code. In version-6 some computations are overlapped with communications. And finally, we divided interprocessor messages in shorter packets to avoid bursty traffic. This is our version-7. In Figure 4 we compare execution times on multiple processors connected by different networks. It is noted that as the number of processor increases, performance of the cluster degrades for the Ethernet due to traffic congestions. For large numbers of processors, in fact execution time increases as the number of processors increases. The break even point for the Ethernet was about 8-10 processors. Both the ALLNODE switch and the FDDI network show sublinear speedups. We observe that the high speed of FDDI is balanced by slower multiple contention free links in the ALLNODE switch. In Figure 5, we analyse the communication characteristics of different networks. The communication time we report is really the sum of the actual communication time, which is not overlapped with the computation time, and the waiting time by a processor for messages. Communication times for the ALLNODE and the FDDI did not change very much for the number of processors used in this study. On the other hand, Ethernet shows a significant increase of communication times with a large number of processors. For the Navier-Stokes computations with the ALLNODE switch with 16 processors, communication time is comparable to the computation and PVM setup time (PVM set up time includes execution times for all PVM library calls, including packing and unpacking. This time does not include any communication wait and transfer time). Relative communication times are lower for the Euler computations. In Figure 6, we show similar results for the standard (second order) MacCormack scheme. Since the stencil size is now three points wide in any particular direction, instead of five points for the fourth order extension, there is less communication. As a result, Ethernet performed relatively better. However, we note that communication times are comparable to the computation and PVM setup time for more than eight processors. We should point out that all our results in the study except those in Figure 6, were obtained with using the fourth order MacCormack scheme.

In version-6, we tried to optimize our numerical model further by doing some computations overlapped with interprocessor communications. Such an approach should reduce communication times. However, this modification also increases the complexity of the numerical model and overhead associated with loop setups. Our results with overlapped computation and communication along with the results with version-7 are shown in Figure 7. We observe that overlapped computation and communication did not improve the performance of our numerical model. With shorter packet size (in version-7), the Ethernet performed better, while we notice degradation for simulations using the ALLNODE switch. The apparent ineffectiveness of the optimizations in versions 6 and 7 is likely due to the nature of communication requirements of our numerical model, i.e., our algorithm. We used our numerical model of version 4 to obtain execution times on the IBM-SP1. As indicated earlier, we used the MPL library for our parallel implementations. Our results are shown in Figure 8. For 16 processors, the speed up was about 13 for the Euler and 14 for the
Navier-Stokes computations. We would like to point out that processors in the SP1 are not exactly same as those in the LACE cluster. As shown in this figure, communication times were negligible compared to execution times. Because of the fast communication, we expect to see significant speedups with additional processors.

On the Cray-YMP, initially we ran the model with only the Autotasking option on and obtained speedups of slightly above one. We then identified the computationally intensive portions of the code using a profiler, performed subroutine inlining, and partitioned data structures explicitly into shared and private variables. With these modifications, using eight processors, we obtained speedups of more than 6 for both the Euler and the Navier Stokes computations. Our results on the Cray YMP are shown in Figure 9.

We compare results of three computing platforms in Figure 10.1 and 10.2. For the LACE cluster, we show results with the ALLNODE switch. Execution times in both the LACE cluster and the IBM SP1 are likely to go down with additional processors. Because of the fast network, speed ups for the IBM SP1 are likely to be higher. The LACE cluster with 16 processors is about one and a half to two times slower than a single processor Cray YMP. The performance of the IBM SP1 appear to be similar. Both the LACE cluster and the IBM SP1 are going through upgrades and performances are going to improve. Parallel computers, specially the network of workstations are becoming viable and cost effective alternatives to traditional supercomputing. For cluster computing, one should try to use a fast network. Otherwise the communication times can become significant and in fact the execution time can rise as processors are added beyond a certain number. In our tests with large number of processors, the ALLNODE performed well and the FDDI was also comparable. On the other hand, the Ethernet performed poorly for such cases. We plan to follow up our present work with studies of other parallel computing platforms including the Cray T3D and examinations of other numerical models.

Acknowledgements

The centerline formulation for three dimensional calculations presented in this paper was implemented by the first author following suggestions of Prof. Eli Turkel. We would like to thank Mr. Dale Hubler for his assistance in using the LACE and the IBM-SP1. We are grateful to Dr. Lou Povinelli for his support for this research in the ICOMP program.

6. References


Figure 1: Axial momentum in an excited axisymmetric jet

Figure 2: Execution time on a single processor RS6000/560

Figure 3: Computation times on sixteen RS6000/370 processors IBM SP-2
Figure 4.1: Navier Stokes execution time on the LACE

Figure 4.2: Euler execution time on the LACE
Figure 5.1: Navier Stokes computations on the LACE

Figure 5.2: Euler computations on the LACE
Figure 6.1: Navier Stokes computations by the standard MacCormack scheme

Figure 6.2: Euler computations by the standard MacCormack scheme
Figure 7.1: Communication optimisations for the Navier Stokes computations

Figure 7.2: Communication optimisations for the Euler computations
Figure 8: Execution times on the IBM-SP2 (with RS6000/270 processors)

Figure 9: Execution times on the Cray YMP
Figure 10.1: Navier-Stokes execution times

Figure 10.2: Euler execution times
Parallel Navier-Stokes Computations on Shared and Distributed Memory Architectures

M. Ehtesham Hayder, D.N. Jayasimha, and Sasi Kumar Pillay

National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135-3191

We study a high order finite difference scheme to solve the time accurate flow field of a jet using the compressible Navier-Stokes equations. As part of our ongoing efforts, we have implemented our numerical model on three parallel computing platforms to study the computational, communication, and scalability characteristics. The platforms chosen for this study are a cluster of workstations connected through fast networks (the LACE experimental testbed at NASA Lewis), a shared memory multiprocessor (the Cray YMP), and a distributed memory multiprocessor (the IBM SP1). Our focus in this study is on the LACE testbed. We present some results for the Cray YMP and the IBM SP1 mainly for comparison purposes. On the LACE testbed, we study: (a) the communication characteristics of Ethernet, FDDI and the ALLNODE networks and (b) the overheads induced by the PVM message passing library used for parallelizing the application. We demonstrate that clustering of workstations is effective and has the potential to be computationally competitive with supercomputers at a fraction of the cost.