Building and Using a Fault-Tolerant MPI Implementation

Graham E. Fagg and Jack J. Dongarra

International Journal of High Performance Computing Applications 2004 18: 353
DOI: 10.1177/1094342004046052

BUILDING AND USING A FAULT–TOLERANT MPI IMPLEMENTATION

Graham E. Fagg [1,2]
Jack J. Dongarra[2]

Abstract

In this paper we discuss the design and use of a fault-tolerant MPI (FT-MPI) that handles process failures in a way beyond that of the original MPI static process model. FT-MPI allows the semantics and associated modes of failures to be explicitly controlled by an application via a modified functionality within the standard MPI 1.2 API. Given is an overview of the FT-MPI semantics, architecture design, example usage and sample applications. A short discussion is given on the consequences of designing a fault-tolerant MPI both in terms of how such an implementation handles failures at multiple levels internally as well as how existing applications can use new features while still remaining within the MPI standard.

Key words: Fault tolerant, message passing, parallel computing, MPI

1 Introduction

MPI (Snir et al., 1998) is the current standard message passing system used to build high performance applications for both clusters and dedicated MPP systems. Initially MPI was designed to allow for very high efficiency and thus performance on a number of early 1990s MPPs, which at the time had limited OS runtime support. This led to the current MPI design of a static process model. This model was possible to implement for MPP vendors, easy to program for, and more importantly something that could be agreed upon by a standards committee. The second version of MPI standard known as MPI-2 (Gropp et al., 2000) did include some support for dynamic process control, although this was limited to the creation of new MPI process groups with separate communicators. These new processes could not be merged with previously existing communicators to form intra-communicators needed for a seamless single application model and were limited to a special set of extended collectives (group) communications.

The MPI static process model suffices for small numbers of distributed nodes within the currently emerging masses of clusters and several hundred nodes of dedicated MPPs. Beyond these sizes the mean time between failure (MTBF) of CPU nodes becomes a factor. As attempts to build the next generation Peta-flop systems advance, this situation will only become more adverse as individual node reliability becomes outweighted by orders of magnitude increase in node numbers and hence node failures. Current GRID (Foster and Kesselman, 1999a) technologies such as GLOBUS (Foster and Kesselman, 1999b) also provide for middleware services such as naming and resource discovery that are robust and handle expected failures gracefully. Unfortunately the MPI message passing library for Globus, MPICH-G (Foster and Karonis, 1998), is not expected to handle loss of MPI processes or partitioning of networks gracefully and failures still lead to pathological failure of applications unless special precautions are taken. Such application checkpointing is discussed further in the next section.

The aim of FT-MPI is to build a fault-tolerant MPI implementation that can survive failures, while offering the application developer a range of recovery options other than just returning to some previous checkpointed state. FT-MPI is built on the HARNESS (Beck et al., 1999)

[1]HIGH PERFORMANCE COMPUTING CENTER STUTTGART ALLMANDRING 30, D-70550 STUTTGART, GERMANY

[2]DEPARTMENT OF COMPUTER SCIENCE, SUITE 413, 1122 VOLUNTEER BLVD., UNIVERSITY OF TENNESSEE, KNOXVILLE, TN 37996-3450, USA

The International Journal of High Performance Computing Applications, Volume 18, No. 3, Fall 2004, pp. 353–361
DOI: 10.1177/1094342004046052

meta-computing system, and is meant to be used as the HARNESS default application level message passing interface. Its design allows it to be easily ported to other GRID environments by porting its modular services that are implemented in the form of short-lived daemons.

2 Checkpoint and Roll Back Versus Replication Techniques

The first method that attempted to make MPI applications fault-tolerant was through the use of checkpointing and roll back. Co-Check MPI (Stellner, 1996) from the Technical University of Munich was the first MPI implementation built that used the Condor library for checkpointing an entire MPI application. In this implementation, all processes would flush their message queues to avoid in-flight messages getting lost, and then they would all synchronously checkpoint. At some later stage, if either an error occurred or a task was forced to migrate to assist load balancing, the entire MPI application would be rolled back to the last complete checkpoint and be restarted. This system's main drawback was the need for the entire application having to checkpoint synchronously, which depending on the application and its size could become expensive in terms of time (with potential scaling problems). A secondary consideration was that they had to implement a new version of MPI known as tuMPI as updating MPICH was considered too difficult.

Another system that also uses checkpointing, but at a much lower level, is Starfish MPI (Agbaria and Friedman, 1999). Unlike Co-Check MPI, which relies on Condor, Starfish MPI uses its own distributed system to provide built-in checkpointing. The main difference with Co-Check MPI is how it handles communication and state changes which are managed by Starfish using strict atomic group communication protocols built upon the Ensemble system, and thus avoids the message flush protocol of Co-Check. Being a more recent project, StarFish supports faster networking interfaces than tuMPI.

The project closest to FT-MPI known to the authors is the Implicit Fault Tolerance MPI project MPI-FT (Louca et al., 1998) by Paraskevas Evripidou of Cyprus University. This project supports several master–slave models where all communicators are built from grids that contain "spare" processes. These spare processes are utilized when there is a failure. To avoid loss of message data between the master and slaves, all messages are copied to an observer process, which can reproduce lost messages in the event of any failures. This system appears only to support SPMD style computation and has a high overhead for every message and considerable memory needs for the observer process for long running applications. This system is not a full checkpoint system in that it assumes any data (or state) can be rebuilt using just the knowledge of any passed messages, which might not be the case for non-deterministic unstable solvers.

MPICH-V (Bosilca et al., 2002) from Université de Paris Sud, France is a mix of uncoordinated checkpointing and distributed message logging. The message logging is pessimistic; thus they guarantee that a consistent state can be reached from any local set of process checkpoints at the cost of increased message logging. MPICH-V uses multiple message storage (observers) known as Channel Memories (CM) to provide message logging. Process level checkpointing is handled by multiple servers known as Checkpoint Servers (CS). The distributed nature of the checkpointing and message logging allows the system to scale, depending on the number of spare nodes available to act as CM and CS servers. Ping-pong performance of MPICH-V compared to MPICH-p4 is around 50%, although application performance is usually much better. In the case of the NAS BP benchmark the overhead for MPICH-V compared to MPICH over P4 varies between 6% and 20%. Handling of a failure is automatic and transparent to the user, although currently only master–slave or SPMD applications are supported.

FT-MPI has much lower overheads compared to the above checkpointing systems, and thus much higher potential performance. These benefits do, however, have consequences. An application using FT-MPI has to be designed to take advantage of its fault-tolerant features as shown in the next section, although this extra work can be trivial depending on the structure of the application. If an application needs a high level of fault tolerance where node loss would equal data loss then the application has to be designed to perform some level of user-directed checkpointing. FT-MPI does allow for atomic communications much like Starfish, but unlike Starfish, the level of correctness can be varied for individual communicators. This provides users the ability to fine tune for coherency or performance as system and application conditions dictate. An additional advantage of FT-MPI over many systems is that checkpointing can be performed at the user level and the entire application does not need to be stopped and rescheduled as with process-level checkpointing.

Currently GRID application efforts such as GrADS (Berman et al., 2001) primarily focus on gaining high performance from GRIDs rather than handling failures, although current efforts at the University of Tennessee (Petitet et al., 2001) involve checkpointing distributed applications to improve fault tolerance. Unlike the above checkpointing systems that rely on local disks for checkpointed data storage, the current GRADS effort is experimenting with replicated distributed storage built on top of the IBP (Plank et al., 1999) system to improve both availability and performance. This system is also a user-level checkpointing scheme rather than process-level and thus would benefit from avoiding rescheduling as provided by FT-MPI.

3 FT-MPI Semantics

Current semantics of MPI indicate that a failure of an MPI process or communication causes all communicators associated with them to become *invalid*. As the standard provides no method to reinstate them (and it is unclear if we can even *free* them), we are left with the problem that this causes MPI_COMM_WORLD itself to become invalid and thus the entire MPI application will grind to a halt.

FT-MPI extends the MPI communicator states from {valid, invalid} to a range {FT_OK, FT_DETECTED, FT_RECOVER, FT_RECOVERED, FT_FAILED}. In essence this becomes {OK, PROBLEM, FAILED}, with the other states mainly of interest to the internal fault recovery algorithm of FT_MPI. Processes also have typical states of {OK, FAILED} which FT-MPI replaces with {OK, Unavailable, Joining, Failed}. The *Unavailable* state includes unknown, unreachable or "we have not voted to remove it yet" states. A communicator changes its state when either an MPI process changes its state, or a communication within that communicator fails for some reason. Some details of failure detection are given in Section 4.1.

The typical MPI semantics is from OK to Failed, which then causes an application abort. By allowing the communicator to be in an intermediate state we allow the application the ability to decide how to alter the communicator and its state, as well as how communication within the intermediate state behaves.

3.1 FAILURE MODES

On detecting a failure within a communicator, that communicator is marked as having a probable error. Immediately as this occurs the underlying system sends a state update to all other processes involved in that communicator. If the error was a communication error, not all communicators are forced to be updated; if it was a process exit then all communicators that include this process are changed. Note that this might not be all current communicators as we support MPI-2 dynamic tasks and thus multiple MPI_COMM_WORLDS.

How the system behaves depends on the communicator failure mode chosen by the application. The mode has two parts, one for the communication behavior and one for the how the communicator reforms, if at all.

3.2 COMMUNICATOR AND COMMUNICATION HANDLING

Once a communicator has an error state it can only recover by rebuilding itself, using a modified version of one of the MPI communicator build functions such as MPI_Comm_{create, split or dup}. Under these functions the will still be the same as if there had been no error, or else new communicator will follow the following semantics depending on its failure mode:

- SHRINK. The communicator is reduced so that the data structure is contiguous. The ranks of the processes are changed, forcing the application to recall MPI_COMM_RANK.
- BLANK. This is the same as SHRINK, except that the communicator can now contain gaps to be filled in later. Communicating with a gap will cause an invalid rank error. Note also that calling MPI_COMM_SIZE will return the extent of the communicator, not the number of valid processes within it.
- REBUILD. Most complex mode that forces the creation of new processes to fill any gaps until the size is the same as the extent. The new processes can either be places in the empty ranks, or the communicator can be shrunk and the remaining processes filled at the end. This is used for applications that require a certain size to execute as in power of two FFT solvers.
- ABORT. This is a mode which affects the application immediately when an error is detected and forces a graceful abort. The user is unable to trap this. If the application needs to avoid this they must set all communicators to one of the above communicator modes.

Communications within the communicator are controlled by a message mode, which can be either of the following.

1. NOP. No operations on error, i.e., no user level message operations are allowed and all simply return an error code. This is used to allow an application to return from any point in the code to a state where it can take appropriate action as soon as possible.
2. CONT. All communication that is NOT to the affected/failed node can continue as normal. Attempts to communicate with a failed node will return errors until the communicator state is reset.

The user discovers any errors from the return code of any MPI call, with a new fault indicated by MPI_ERR_OTHER. Details as to the nature and specifics of an error are available though the cached attributes interface in MPI as discussed in Section 3.4 below.

3.3 POINT-TO-POINT VERSUS COLLECTIVE CORRECTNESS

Although collective operations pertain to point-to-point operations in most cases, extra care has been taken in implementing the collective operations so that if an error occurs during an operation, the result of the operation the operation is aborted.

```
/* pre-defined key value */
key = FT_MPI_LIST_NUM_FAILED;  /* key for finding number of failure events */
key2 = FT_MPI_LIST_FAILED;       /* key for getting pointer to failures in a list */

rc= MPI_func (comm…)
If (rc==MPI_ERR_OTHER) {
       rc = MPI_Comm_get_attr (comm, key, &num_failed, &flag);
       rc = MPI_Comm_get_attr (comm, key2, &failed_ptr, &flag);
       for (i=0;i<num_failed;i++)
              printf("failure %d was rank %d\n", i+1, failed_ptr[i]);
}
```

Example 1. Checking for order of failures

```
key = FT_MPI_COM_NUM_FAILED; /* key for finding how many individual ranks failed */
key2 = FT_MPI_COM_FAILED;   /* key for accessing complete failure map of a communicator */

rc= MPI_Send (----, com);
If (rc==MPI_ERR_OTHER) {
       rc = MPI_Comm_get_attr (comm, key, &num_failed, &flag);
       rc = MPI_Comm_get_attr (comm, key2, &failed_ptr, &flag);
       /* check list of failures */
       failed_how_many_times = failed_ptr [rank];
       }
```

Example 2. Accessing failures via process RANK

Broadcast, gather and all-gather demonstrate this perfectly. In broadcast, even if there is a failure of a receiving node, the receiving nodes still receive the same data, i.e., the same end result for the surviving nodes. Gather and all-gather are different in that the result depends on if the problematic nodes sent data to the gatherer/root or not. In the case of gather, the root might or might not have gaps in the result. For the all2all operation, which typically uses a ring algorithm, it is possible that some nodes may have complete information and others incomplete. Thus, for operations that require multiple node input as in gather/reduce type operations any failure causes all nodes to return an error code, rather than possibly invalid data. Currently an addition flag controls how strictly the above rule is enforced by utilizing an extra barrier call at the end of the collective call if required.

3.4 FT-MPI NOTIFICATION OF FAILURES

The MPI standard does not indicate how errors are reported beyond standard return codes and error classes to provide additional information. Without altering the meaning of the standard, FT-MPI utilizes these mechanisms so that applications that have been adapted to FT-MPI still compile and link correctly on other MPI implementations.

To remain within the standard FT-MPI notifies the application with a single return code MPI_ERR_OTHER that an error has occurred and then makes additional information available via the attribute caching mechanism. A human readable form of the failure is also provided via a MPI error class using the MPI error string function.

Two forms of essentially the same information are made available to the application. The first form returns the error information for a complete communicator in terms of the number of failures per rank since the last recovery. The second form returns the failed ranks in the order that they were detected *locally*. This ordering is only consistent globally in terms of the total failures not the ordering reported at each node unless the FTMPI_NOTIFIER daemon is used to force ordering of events.

3.5 FT-MPI BASIC USAGE

Simple usage of FT-MPI would be in the form of an error check and then some corrective action such as a communicator rebuild. A typical code fragment is shown in Example 3, where on an error the communicator is simply rebuilt and reused.

Some types of computation, such as SPMD master–worker codes, only need the error checking in the master

```
rc= MPI_Send (----, com);
If (rc==MPI_ERR_OTHER) {
       MPI_Comm_dup (com, newcom);/* collective recovery occurs here! */
       MPI_Comm_free (com);
       com = newcom;
      }
/* continue.. */
```

Example 3. Simple FT-MPI send usage

```
rc = MPI_Bcast ( initial_work…);
if(rc==MPI_ERR_OTHER)reclaim_lost_work(…);
  while ( ! all_work_done) {
  if (work_allocated) {
    rc = MPI_Recv ( buf, ans_size, result_dt,
                    MPI_ANY_SOURCE,   MPI_ANY_TAG, comm, &status);
    if (rc==MPI_SUCCESS) {
                          handle_work (buf);
                          free_worker (status.MPI_SOURCE);
                          all_work_done--;
                          }
    else {
           reclaim_lost_work(status.MPI_SOURCE);
         if (no_surviving_workers) { /* ! do something ! */ }
        }
  } /* work allocated */
/* Get a new worker as we must have received a result or a death */
  rank=get_free_worker_and_allocate_work();
  if (rank) {
    rc = MPI_Send (… rank… );
    if (rc==MPI_OTHER_ERR) reclaim_lost_work (rank);
    if (no_surviving_workers) { /* ! do something ! */ }
  } /* if free worker */
} /* while work to do */
```

Example 4. FT-MPI Master-Worker code

code if the user is willing to accept the master as the only point of failure. Example 4 shows how complex a master code can become. In this example the communicator mode is BLANK and communications mode is CONT. The master keeps track of work allocated, and on an error just reallocates the work to any "free" surviving processes. Note that the code has to check to see if there are any surviving workers remaining after each death is detected.

3.6 FT-MPI USAGE WITHIN EXISTING MESSAGE PASSING LIBRARIES

Many real-world parallel applications use numeric libraries, such as ScaLAPACK (Blackford et al., 1997) and PETSc (Balay et al., 2000), which themselves use MPI internally through multiple layers. Altering such libraries by changing each occurrence of each MPI call is impractical and error prone.

A more elegant solution is to use the MPI error handling functions to automatically handle the errors for the application. When combined with the long jump mechanism in the C language this can provide a very simple solution to many classes of error handling. A typical program flow for an application is given in Figure 1. If the application already contains user-level checkpointing then only the initial startup section of the code needs to be altered. The flow within a normal process would proceed as follows:

1. MPI_Init would indicate if the process was started normally via MPIRUN or was a restarted node within an application.

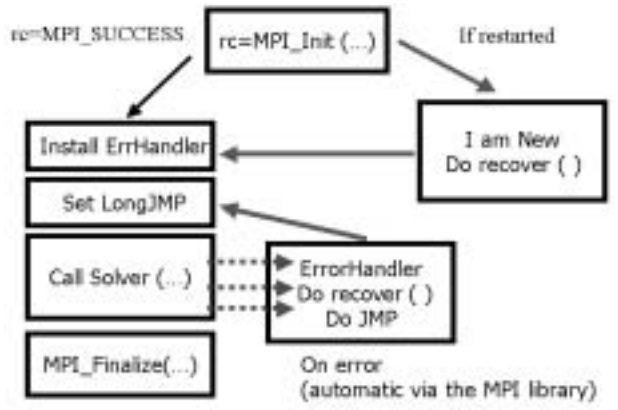

Fig. 1 Flow control in a typical FT-MPI application using MPI Error Handlers.

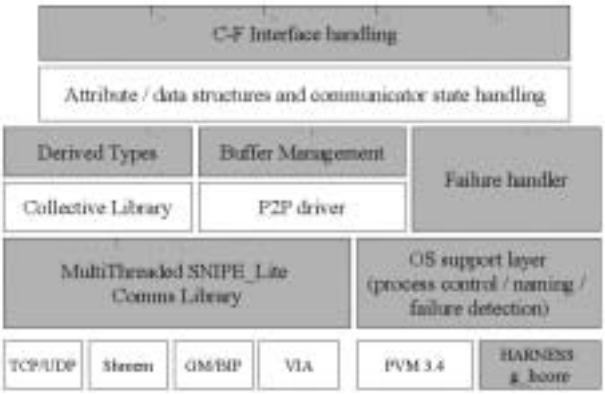

Fig. 1 Overall structure of the FT-MPI implementation.

```
ehf = (MPI_Handler_function *) (&errhandleruserfunc); /* get handle to my error handler */

MPI_Errhandler_create (ehf, &errh);          /* create MPI handle to my function */
MPI_Errhandler_get (MCW, &errh_org);                 /* get original MPI handler */
MPI_Errhandler_free (&errh_org);
MPI_Errhandler_set (MCW, errh);              /* replace default with my function */
```

Example 5. Installing an error handler under MPI

2. If the process is normal, then the application would install the MPI error handler that they wrote as shown in code Example 5.
3. The process would set a long jump so that it could return to the top level functions where it can correctly manage program flow during a recovery. This is required as a failure could be many levels of function calls later.
4. The code would call the numeric library containing MPI calls (i.e., a parallel solver).
5. If completed successfully the code would enter MPI_finalize and terminate normally.

During the execution if an error occurred, the FT-MPI runtime library would catch it and as soon as the program enters a MPI routine, flow control would be passed to the MPI error handler the user provided in 2 above. At this point the user's application could block on a communicator create/duplicate function after which they would probably load the user-level checkpoint data. After recovery they would then jump back to the top level of the application, reset the jump and then continue as per item 3 above.

A restarted process would discover from the MPI_Init function that it was restarted and would then load any recovery data rather than initial data, install the error handler and continue as a normal process.

4 FT_MPI Implementation Details

FT-MPI is a partial MPI-2 implementation. It currently contains support for both C and Fortran interfaces, all the MPI-1.2 function calls required to run both the PSTSWM (Worley et al., 1995) and BLAS (Choi et al., 1995) applications. BLAS is supported so that ScaLAPACK (Blackford et al., 1997) applications can be tested. Currently only some of the dynamic process control functions from MPI-2 are supported.

The current implementation is built as a number of layers as shown in Figure 2. Operating system support is provided by either PVM or the C HARNESS *G_HCORE*.

Although point-to-point communication is provided by a modified SNIPE_Lite communication library taken from the SNIPE project (Fagg et al., 1999).

A number of components have been extensively optimized, including derived data types (Fagg et al., 2001) and message buffers and collective communications (Vadhiyar et al., 2001).

4.1 FAILURE DETECTION

It is important to note that the failure handler shown in Figure 2 gets notification of failures from both the point-to-point communications libraries, as well as the OS support layer. In the case of communication errors, the notify is usually started by the communication library detecting a point-to-point message not being delivered to a failed party rather than the failed parties OS layer detecting the failure. The handler is responsible for notifying all tasks of errors as they occur by injecting notify messages into the send message queues ahead of user-level messages. An additional daemon known as the FTMPI_NOTIFER can be used to guarantee ordered delivery of failure notification messages and thus aid in complex debugging.

The failure handler within the FTMPI run-time library relies on the conservation of event messages from the underlying system to build a coherent system state during recovery. A consequence of this is that temporary bi-sectioning of the network between G_HCORE startup daemons can lead to some processes being marked as failed; thus the sum of living tasks and failure events will remain constant.

4.2 LOW-LEVEL MESSAGE HANDLING

Many MPI message passing libraries employ multiple message delivery schemes which vary with message size to provide a balance between performance, unexpected message buffering memory requirements and blocking semantics. GM, for example, switches between eager (always send) and rendezvous modes as the message size increases.

FT-MPI uses eager for performance on all blocking sends and switches to a token-based system for large non-blocking messages. As with the failure detection, the handling of communication during failures relies on a guaranteed delivery of flow control messages and failure events.

During a failure all processes flush communications with all existing communication contexts. They complete all pending operations involving a remote process, until either they have received a flow control message indicating that the process is entering a global state rebuild or a failure event for that process is received. Thus the number of flow control stop messages and death events

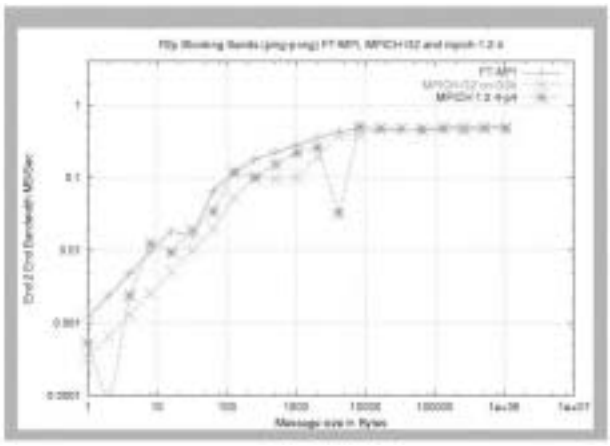

Fig. 3 Point-to-point message performance of FT-MPI compared to various MPICH versions.

of open connections must match the number of pre-failure open connections. This allows all/any processes in an eager send to always complete as their target guarantees emptying the pipe before entering the global recovery state, thus avoiding any deadlocks.

5 FT-MPI Performance

Figure 3 shows the performance of FT-MPI for point-to-point messages compared to MPICH-p4 and MPICH-G2 under Globus 2.0. Further performance information can be obtained from Fagg et al. (2001) and Vadhiyar et al. (2001). As was stated in Section 2, the performance of FT-MPI is not hindered by fault handling. Any additional costs of fault tolerance only occur at applications startup, during a failure recovery and during shutdown.

6 Conclusions

FT-MPI is an attempt to provide application programmers with different methods of dealing with failures within MPI application than just checkpoint and restart. It is hoped that by experimenting with FT-MPI, new application methodologies and algorithms will be developed to allow for both high performance and the survivability required by both unreliable GRIDs and the next generation of terra-flop and beyond machines. FT-MPI in itself is already proving to be a useful vehicle for experimenting with self-tuning collective communications, distributed control algorithms, various dynamic library download methods and improved sparse data handling subsystems, as well as being the default MPI implementation for the HARNESS project.

Future work in the FT-MPI library system will concentrate on developing a number of drop-in library templates or skeletons to simplify the construction of fault-tolerant applications.

BIOGRAPHIES

Graham Fagg received his BSc in Computer Science and Cybernetics from the University of Reading (UK) in 1991 and a PhD in Computer Science in 1998. From 1991 to 1993, he worked on CASE tools for interconnecting array processors and Transputer MIMD systems. From 1994 to 1995 he was a research assistant at the Cluster Computing Laboratory at the University of Reading working on code generation tools for group communications. From 1996 to 2001 he worked as a senior research associate and then a Research Assistant Professor at the University of Tennessee. From 2001 to 2002 he was a visiting guest scientist at the High Performance Computing Center Stuttgart (HLRS). Currently he is a Research Associate Professor at the University of Tennessee. His current research interests include distributed scheduling, resource management, performance prediction, benchmarking, cluster management tools, parallel and distributed IO and high-speed networking. He is currently involved in the development of a number of metacomputing and GRID middle-ware systems including SNIPE/2, MPI_Connect, HARNESS, Open MPI, and a process level fault-tolerant MPI implementation (FT-MPI).

Jack Dongarra holds an appointment as University Distinguished Professor of Computer Science in the Computer Science Department at the University of Tennessee and is an Adjunct R&D Participant in the Computer Science and Mathematics Division at Oak Ridge National Laboratory (ORNL) and an Adjunct Professor in Computer Science at Rice University. He specializes in numerical algorithms in linear algebra, parallel computing, and the use of advanced-computer architectures, programming methodology, and tools for parallel computers. His research includes the development, testing and documentation of high quality mathematical software. He has contributed to the design and implementation of the following open source software packages and systems: EISPACK, LINPACK, the BLAS, LAPACK, ScaLAPACK, Netlib, PVM, MPI, NetSolve, Top500, ATLAS, and PAPI. He has published approximately 200 articles, papers, reports and technical memoranda and he is co-author of several books. He is a Fellow of the AAAS, ACM, and the IEEE and a member of the National Academy of Engineering.

References

Agbaria, A. and Friedman, R. 1999. Starfish: fault-tolerant dynamic MPI programs on clusters of workstations. In *Proceedings of the 8th IEEE International Symposium on High Performance Distributed Computing*, Redondo, Beach, CA.

Balay, S., Gropp, W.D., McInnes, L.C., and Smith, B.F. 2000. *PETSc 2.0 Users Manual* Argonne National Laboratory, ANL-95/11, Revision 2.0.29.

Beck, M. et al. 1999. HARNESS: a next generation distributed virtual machine. *Journal of Future Generation Computer Systems*, 15(5–6):571–582.

Berman, F. et al. 2001. The GrADS Project. *International Journal of High Performance Computing Applications*, 15(4): 327–344.

Blackford, S. et al. 1997. ScaLAPACK: a linear algebra library for message-passing computers. In *Proceedings of 1997 SIAM Conference on Parallel Processing*, Minneapolis, MN.

Bosilca, G. et al. 2002. MPICH-V: toward a scalable fault-tolerant MPI for volatile nodes. In *Proceedings of SuperComputing 2002*, Baltimore, MD.

Choi, J., Dongarra, J., Ostrouchov, S., Petitet, A., Walker, D., and Whaley, R. 1995. A Proposal for a Set of Parallel Basic Linear Algebra Subprograms, *LAPACK Working Note #100*, CS-95-292, May.

Fagg, G.E., Moore, K., and Dongarra, J.J. 1999. Scalable networked information processing environment (SNIPE). *Journal of Future Generation Computer Systems*, (15): 571–582.

Fagg, G., Bukovsky, A., and Dongarra, J. 2001. HARNESS and fault-tolerant MPI. *Parallel Computing*, 27(11):1479–1496.

Foster, I. and Karonis, N., 1998. A Grid enabled MPI: message passing in heterogeneous distributed computing systems. In *Proceedings of SuperComputing 98 (SC98)*, Orlando, FL.

Foster, I. and Kesselmann, C. 1999a. *The GRID: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, San Mateo, CA.

Foster, I. and Kesselman, C. (editors). 1999b. The Globus Toolkit. In *The GRID: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, San Mateo, CA, pp. 259–278.

Gropp, W., Lusk, E., and Thakur, R. 2000. *Using MPI-2: Advanced Features of the Message Passing Interface*, 1st edition, MIT Press, Cambridge, MA.

Louca, S., Neophytou, N., Lachanas, A., Evripidou, P. 1998. MPI-FT: a portable fault tolerance scheme for MPI. In *Proceedings of PDPTA '98 International Conference*, Las Vegas, NV.

Petitet, A., Blackford, S., Dongarra, J., Ellis, B., Fagg, G., Roche, K., and Vadhiyar, S. 2001. Numerical libraries and the Grid. *International Journal of High Performance Computing Applications*, 15(4):359–374.

Plank, J.S., Beck, M., Elwasif, W.R., Moore, T., Swany, M., and Wolski, R. 1999. The Internet Backplane Protocol: Storage in the Network. In *NetStore99: The Network Storage Symposium*, Seattle, WA.

Snir, M., Otto, S., Huss-Lederman, S., Walker, D., and Dongarra, J. 1998. *MPI – The Complete Reference. The MPI Core*, Vol. 1, 2nd edition.

Stellner, G. 1996. CoCheck: checkpointing and process migration for MPI. In *Proceedings of the International Parallel Processing Symposium*, April, Honolulu, HI, pp. 526–531.

Vadhiyar, S.S., Fagg, G.E., and Dongarra, J.J. 2001. Performance modeling for self-adapting collective communications for MPI. In *LACSI Symposium 2001*, October 15–18, Santa Fe, NM.

Worley, P.H., Foster, I.T., and Toonen, B. 1995. Algorithm comparison and benchmarking using a parallel spectral transform shallow water model. In *Proceedings of the 6th Workshop on Parallel Processing in Meteorology*, G.-R. Hoffmann and N. Kreitz, editors, World Scientific, Singapore, pp. 277–289.