Writing Programs that Run EveryWare on the Computational Grid *

University of Tennessee Technical Report Number UT-CS-99-420
http://nws.npaci.edu/EveryWare

Rich Wolski † John Brevik ‡ Chandra Krintz § Graziano Obertelli ¶
Neil Spring ‖ Alan Su **

August 11, 1999

1 Introduction

Increasingly, the high-performance computing community is blending parallel and distributed computing technologies to meet its performance needs. A new architecture, known as *The Computational Grid* [12], has recently been proposed which frames the software infrastructure that is required to implement high-performance applications using widely dispersed computational resources. The goal of a Computational Grid is to aggregate ensembles of shared, heterogeneous, and distributed resources (potentially controlled by separate organizations) to provide computational "power" to an application program. Applications should be able to draw compute cycles, network bandwidth, and storage capacity seamlessly from the Grid[1] in a way analogous to the way in which household appliances draw electrical power from a power utility.

To realize this vision, the application programming environment must be able to

- leverage *all* potentially useful resources that the user can access,

*Supported by the National Partnership for Advanced Computational Infrastructure (NPACI), NSF grant ASC-9701333, Advanced Research Projects Agency/ITO under contract #N66001-97-C-8531, and caffeine.

†University of Tennessee – email: rich@cs.utk.edu
‡University of Calfornia, Berkeley – email: brevik@math.berkeley.edu
§University of Calfornia, San Diego – email: ckrintz@cs.ucsd.edu
¶University of Calfornia, San Diego – email: graziano@cs.ucsd.edu
‖University of Washington – email: nspring@cs.washington.edu
**University of California, San Diego – email: alsu@cs.ucsd.edu

[1]We will capitalize the word "Grid" when referring to "Computational Grid" throughout this paper.

- exploit the heterogeneity of the resource pool to the program's advantage, and
- manage the effects of dynamically changing resource performance characteristics caused by contention, reconfiguration, and federated administration.

Researchers have developed several innovative and powerful software infrastructures to support the Grid paradigm [11, 22, 41, 6] and several pilot projects [18, 26, 33] have been launched to investigate the efficacy of Grid computing. Each of these technologies, however, is designed assuming there will be a single, unifying Grid infrastructure installed and functioning on all available resources. We note, however, that the resource pool available to an application is generally *specific to its user* and not a particular infrastructure. **A single user may have access to resources owned by different organizations which do not agree to support a single, unifying software base on all resources**.

In this paper, we describe the design of *EveryWare* – a user-level software toolkit for writing Grid programs consisting of three separate software components:

- a **portable *lingua franca*** that is designed to allow processes using different infrastructures and operating systems to communicate,
- a set of **performance forecasting libraries** that enable an application to make short-term resource and application performance predictions in near-real time, and
- a **distributed state exchange** service that allows application components to manage and synchronize program state in a dynamic environment.

The goal is to allow a user to write Grid programs that combine the best features of different Grid infrastructures such as Globus [11], Legion [22], Condor [41], or NetSolve [6] as well as the native functionality provided by Java, Windows NT, and Unix to the performance advantage of the application. EveryWare is implemented as a portable set of libraries and processes that can "glue" different locally-available infrastructures together so that a program may draw upon these resources seamlessly. If sophisticated systems such as Globus [11], Legion [22], or Condor [41] are available, the EveryWare program must be able to use the features provided by those systems effectively. If only basic operating system functionality is present, however, an EveryWare program should be able to extract what ever functionality it can, realizing that these resources may be less effective than those supporting better infrastructure. The ability to leverage all resources accessible by the user, regardless of installed infrastructure, makes it possible to draw those resources from separate administrative domains. That is, to treat the user's accessible resources as a Computational Grid.

We have implemented a prototype toolkit to test the efficacy of the EveryWare approach. In an experiment entered as a contestant in the High-Performance Computing Challenge [25] at SC98 in November of 1998, we were able to use this prototype to leverage Globus [11], Legion [22], Condor [41], NetSolve [6] Grid computing infrastructures, the Java [19] language and execution environment, native Windows NT [44], and native Unix systems ***simultaneously*** in a single, globally distributed application. The application, a program that searches for Ramsey Number counter examples, does not use exhaustive search, but rather requires careful dynamic scheduling to avoid substantial communication overheads. Moreover, by focusing on enhancing the interoperability of the resources in our pool, we were able to combine the Tera MTA[42] and the NT Supercluster[34] - two unique and powerful resources - with a raft of other, more commonly available systems including parallel supercomputers, PC-based workstations, shared-memory multiprocessors, and Java-enabled desk-top browsers. With non-dedicated access to all resources, under extremely heavy load conditions, the EveryWare application was able to sustain supercomputer performance levels over long periods of time. As such, the Ramsey Number Search application using EveryWare represents an example of a true Grid program - the computational "power" of all resources that were available to the application's user was assessed, managed, and delivered to the application.

In detailing our Computational Grid experiences, this paper makes three important contributions.

- It demonstrates, quantitatively, the potential power of globally distributed Grid computing.
- It details the real-world experiences we gained using most of the relevant distributed computing technology available to us in the fall of 1998.
- It describes a programming model and methodology for writing Grid programs.

We endeavor to measure, quantitatively, the degree to which the Ramsey Number Search application was able to leverage a world-wide Computational Grid by combining disparate software infrastructures. To use EveryWare effectively, we adopted a programming methodology in which computational clients request program management functions (scheduling, persistent storage management, event synchronization, etc.) from application-specific distributed services. We believe that this programming model will allow newly-developed Grid applications to meet both the performance and robustness needs of Grid users.

In [12] (page 18), the authors define the criteria for the Computational Grid as the provision of *consistent*, *dependable*, *pervasive*, and *inexpensive* compute power to the end user. We present the results of our study in terms of four quantitative metrics:

1. the sustained computational performance of the entire application,
2. the difference between the performance variability exhibited by the underlying resources and the performance variability exhibited by the application itself,
3. the overall duration of continuous program execution, and
4. the number of different resource types used by the application.

Metric (1) describes the degree to which the EveryWare application was able to achieve "high-performance" using a Computational Grid. Application performance is a key motivating factor for Computational Grids that differentiates it from other distributed computing efforts. Metric (2) demonstrates how well EveryWare was able to hide underlying performance variance from the application's user. Application performance must remain *consistent* and stable even when the deliverable performance of individual resources varies or users will turn to other, lower performance execution platforms. Metric (3) describes the *dependability* and robustness of the EveryWare application to host and network failure, and metric (4) quantifies the degree of *pervasiveness* we were able to achieve with EveryWare. For the Computational Grid to be pervasive, it must be able to include as many of the available resources and resource types as possible. We believe that all four of these metrics are necessary to describe and quantify the degree to which an application achieves the goals (described in [12]) of Grid computing. In addition, the EveryWare application we implemented at SC98 used no dedicated resources and required no privileged access. As the EveryWare application was a guest on the systems it used, the expense associated with its execution was limited to the cost of a "guest" login on each system. That is, it was *inexpensive* to execute since it was able to harvest standard, guest-level priority cycles.

In the next section we describe Computational Grid computing and motivate the design of EveryWare. In Section 3 we detail the functionality of the EveryWare toolkit and describes the programming model it implements. Section 4 discusses the Ramsey Number Search application we used in this experiment and in Section 5, we detail the performance results we were able to obtain in terms of the four metrics described above. Section 6 describes the relationship of EveryWare to related research, and we conclude in Section 7 with a description of future research directions.

2 Computing with Computational Grids

The goal of EveryWare is to enable the construction of true Grid programs – ones which draw computational power seamlessly from a dynamically changing resource pool. Since the field is evolving, a single definition of "Computational Grid" has yet to be universally adopted [2]. In this work, we will use the following definition.

Computational Grid A heterogenous, shared, and federated collection of computational resources that are linked together by a network that supports interprocess communication.

By "shared" we mean that it is impractical to dedicate all of the resources in a Computational Grid to a single application for an appreciable amount of time. The term "federated" means that each resource is expected to have a local resource manager, and no single overarching resource management policy can be imposed on all resources.

The resources housed at the National Partnership for Advanced Computational Infrastructure (NPACI) and National Computational Science Alliance (NCSA) constitute examples of Computational Grids under our definition. At these centers, machines and storage devices of various types are internetworked. Each resource is managed by its own resource manager (e.g. batch scheduler, interactive priority mechanism, etc.) and it is not generally possible to dedicate all resources (and the network links that interconnect them) at either site to a single application. Moreover, it is possible to combine NPACI and NCSA resources together to form a larger Computational Grid, that has the same characteristics. In this larger case, it is not even possible to mandate that a uniform software infrastructure be present at all potentially useful execution sites.

[2] In [12], the authors define Computational Grids in terms of a set of criteria that must be met. We address these criteria in our work, but prefer the definition provided herein for the purpose of illustration

To effectively exploit a Computational Grid, a program must be able to

1. *leverage concurrency* for execution performance,
2. *manage the resource heterogeneity* of the underlying system to the advantage of the program,
3. *run ubiquitously* on all of the resources its user may legally access, and
4. *tolerate dynamic performance variation* caused by contention, resource failure, resource acquisition, and local management policy.

Other work has met these requirements to different degrees. AppLeS [4] (*App*lication *Le*vel *S*cheduling) agents have been able to meet these requirements in environments where a single infrastructure is present and the scheduling agent does not experience resource failure. An AppLeS agent dynamically evaluates the performance that all available resources can deliver to its application, and crafts a schedule that maximizes the application's overall execution performance. EveryWare supports this principle in environments where a single scheduling agent can access all available resources without forming a performance bottleneck (Requirement (4)), but also extends it to wide-area lossy environments in which several infrastructures may be available. Note also that the AppLeS agent is a specialized application component that performs a single application management function: scheduling. EveryWare allows an application programmer to generalize this notion to other application management functions in the form of application-specific services. Using the EveryWare toolkit, these services can be implemented for a variety of Grid infrastructures (Requirement (3)) to meet the performance and robustness needs for widely distributed Grid applications. The MPI (Message Passing Interface) [10], and PVM (Parallel Virtual Machine) [17] implementations for networked systems allow distributed clusters of machines to programmed as a single, "virtual" parallel machine (meeting Requirement (1)). In addition, portable implementations that do not require privileged (super-user) access for installation or execution [23, 17] are available (Requirement (3)). However, they do not manage resource heterogeneity on behalf the program nor do they expose it to the programmer so that it may be managed explicitly. Grid computing systems such as Globus [11], Legion [22], Condor [41], and HPC-Java [24] include support for resource heterogeneity as well, but they are not available ubiquitously (Requirement (3)). As they gain in popularity, we anticipate these systems to be more widely installed and maintained. However we note that their level of sophistication makes porting them to new and experimental environments labor intensive. As new architectures become available, it is important to be able to incorporate them into a Grid quickly. Dynamically schedulable programs that are capable of tolerating resource performance fluctuations have been developed by the Autopilot [38], Winner [2] and MARS [16] groups (Requirement (4)). Most of these systems rely on a centralized scheduler for each application, however, making, them sensitive to network and host failure. If the scheduler fails or becomes disconnected from the rest of the application, the program is disabled. In addition, having a single scheduling agent impedes scalability as communication with the scheduler becomes a performance bottleneck.

EveryWare is designed as a portable "toolkit" for linking together program components running in different environments. Individual program components may use what ever locally available infrastructure is present. In addition, we provide a low-level "bare-bones" implementation that is designed to use only basic operating system functionality. In this way, an EveryWare application does not assume any one, single infrastructure or operating system will be accessible from every resource. Borrowing from the AppLeS [4] project, EveryWare applications characterize all resources in terms of their quantifiable impact on application performance. In this way, heterogeneity is expressed as the difference in deliverable performance to each application. The EveryWare toolkit includes support for process replication and performance forecasting so that an EveryWare application can adapt to dynamically changing resource conditions. We leverage the Network Weather Service [46, 45] forecasting facilities to provide both heterogeneity management and adaptive prediction functionalities.

3 The EveryWare Toolkit

To realize the performance offered by the Grid computing paradigm, a program must be ubiquitous, adaptive, robust, and scalable. Ubiquity is required because the resources are federated. The owners of the resources allow them to participate in the Grid, but maintain ultimate authority over their use. As such, the resource pool may change without notice. Resources may be added, removed, replaced, or upgraded by their owners. In addition, it is difficult to ensure that all resources owners will install and maintain a single, unifying Grid software infrastructure. If the program is not compatible with all potentially available Grid infrastructures, operating systems, and hardware architectures it will not be able to draw some of the "power" that the Grid can provide. Adaptivity is required to ensure performance. If the resource pool is changing, or the performance of the resources are fluctuating due to contention, the program must be able to choose the most profitable resource combination from the resources that are available at any given time. Similarly, if resources become unavailable due to owner-initiated reclamation, excessive load, or failure, the program must be able to make progress. Scalability, in a Grid setting, allows the program to use resources efficiently. The greater degree to which the program can be dispersed, the greater the flexibility the Grid system has in being able to meet the program's performance needs.

The EveryWare toolkit is composed of three separate software components: a portable *lingua franca* that allows processes using different infrastructures and operating systems to communicate, a set of performance forecasting services and libraries that enable an application to make short-term resource and application performance predictions in near-real time, and a distributed state exchange service that allows application components to manage and synchronize program state in a dynamic environment. Figure 1 depicts the relationship between these components. Application components that are written to use different Grid infrastructure features can communicate amongst themselves, with the EveryWare state exchange service, and with other multi-infrastructure services such as the Network Weather Service [46] using the *lingua franca*. Dynamic forecasting libraries (small triangles in the figure) can be loaded with application components directly. These libraries, in conjunction with the performance forecasts provided by the NWS, permit the program to anticipate performance changes and adapt execution accordingly. The distributed state-exchange services provide a mechanism for synchronizing and and replicating important program state to ensure robustness and scalability.

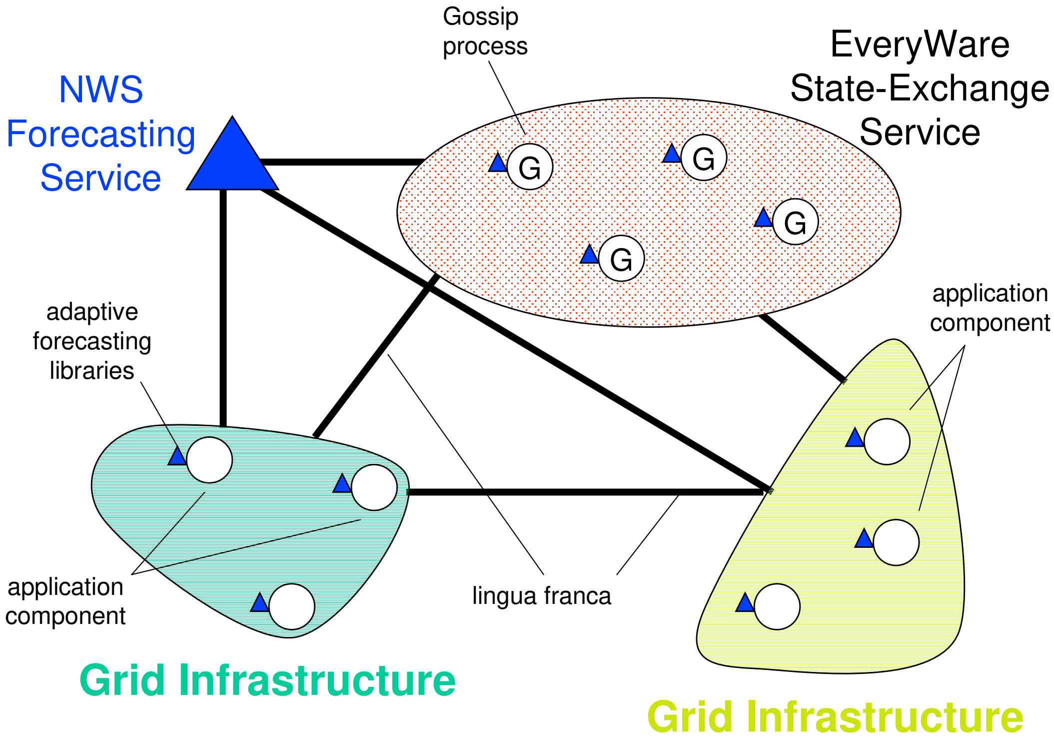


Figure 1: **EveryWare Components**

The toolkit we have implemented is strictly a prototype designed to expose the relevant programming issues. As such, we do not describe the specific APIs supported by each component (we expect them to change dramatically in our future implementations). Rather, in this section, we motivate and describe the functionality of each EveryWare component and discuss our overall implementation strategy. Our intention is to use the prototype first to implement a variety of applications so that we may determine what functionality is required, and then to provide a "user-friendly" implementation of EveryWare for public release.

3.1 *Lingua Franca*

We implemented the *lingua franca* using C and TCP/IP sockets. To ensure portability, we tried to limit the implementation to use only the most "vanilla" features of these two technologies. For example, we did not use non-blocking socket I/O nor did we rely upon keep-alive signals to inform the system about end-to-end communication failure. In our experience, the semantics associated with these two useful features are vendor, and in some cases, operating system release-level specific. We tried to avoid controlling the portability of EveryWare through C preprocessor flags whenever possible so that the system could be ported quickly to new architectures and environments.

Similarly, we chose not to rely upon XDR [31] for data type conversion for fear that it would not be readily available in all environments. Another important decision was to strictly limit our use of signals. Unix signal semantics are somewhat detailed and we did not want to hinder the portability to non-Unix environments (e.g. Java and Window NT). More immediately, many of the currently available Grid communication infrastructures such as Legion [22] and Nexus [14] take over the user-level signal mechanisms to facilitate message delivery. Lastly, we avoided the use of threads throughout the architecture as differences in thread semantics and thread implementation quality has been a source of incompatibility in many of our previous Grid computing efforts.

Above the socket level, we implemented rudimentary packet semantics to enable message typing and delineate record boundaries within each stream-oriented TCP communication. Our approach takes its inspiration from the publicly available implementation of `netperf` [27]. However, the actual implementation of the messaging layer comes directly from the current Network Weather Service (NWS) [46], where it has been stress-tested in a variety of Grid computing environments.

3.2 Forecasting Services

We also borrowed and enhanced the NWS forecasting modules for EveryWare. To make performance forecasts, the NWS applies a set of light-weight time series forecasting methods and dynamically chooses the technique that yields the greatest forecasting accuracy over time (see [45] for a complete description of the NWS forecasting methodology). The NWS collects performance measurements from Grid computing resources (CPUs, networks, etc.) and uses these forecasting techniques to predict short-term resource availability. For EveryWare, however, we needed to be able to predict the time required to perform arbitrary but repetitive program events. Our strategy was to manually instrument the various EveryWare components and application modules with timing primitives, and then pass the timing information to the forecasting modules to make predictions. We refer to this process as *dynamic benchmarking* as it uses benchmark techniques (e.g. timed program events) perturbed by ambient load conditions to make performance predictions.

For example, we used the forecasting modules and dynamic benchmarking to predict the response time of each EveryWare state-exchange server. We instrumented each server to record the time required to get a response to a request made to each of the other servers, for each message type. To do so, we identified each place in the server code where a request-response pair occurred, and tagged each of these "events" with an identifier consisting of address where the request was serviced, and the message type of the request. By forecasting how quickly a server would respond to each type of message, we were able to dynamically adjust the message time-out interval to account for ambient network and CPU load conditions. This dynamic time-out discovery proved crucial to overall program stability. Using the alternative of statically determined time-outs, the system frequently misjudged the availability (or lack thereof) of the different EveryWare state-management servers causing needless retries and dynamic reconfigurations (see subsection 3.3 below for a discussion of EveryWare state-exchange functionality).

In general, the forecasting services and dynamic benchmarking allow both the EveryWare toolkit, and the application using it, to dynamically adapt itself to changing load and performance response conditions. We trimmed down and adapted the NWS forecasting subsystem so that it may be loaded as a library with application and EveryWare code. We also added a tagging methodology so that arbitrary program events could be identified and benchmarked. We used standard timing mechanisms available on each system to generate time stamps and event timings. However, we anticipate that more sophisticated profiling systems such as Paradyn [32] and Pablo [9] could be incorporated to yield higher-fidelity measurements.

3.3 Distributed State Exchange Service

To function in the current Grid computing environments, a program must be robust with respect to resource performance failure while at the same time able to leverage a variety of different target architectures. EveryWare provides a distributed state exchange service that can be used in conjunction with application-level checkpointing to ensure robustness. EveryWare state-exchange servers (called *Gossips*) allow application processes to register for state synchronization. The synchronizing application component must register a contact address, a unique message type, and a function that allows a *Gossip* to compare the "freshness" of two different messages having the same type. All application components wishing to use Gossip service must also export a state-update method for each message type they wish to synchronize.

Once registered, an application component periodically receives a request from a *Gossip* process to send a fresh copy of its current state (identified by message type). The *Gossip* compares that state (using the previously registered comparator function) with the latest state message received from other application components. When the *Gossip* detects that a particular message is out-of-date, it sends a fresh state update to the application component that originated the out-of-date message.

To allow the system to scale, we rely on three assumptions. First, that the *Gossip* processes cooperate as a distributed service. Second, that the number of application components wishing to synchronize is small. Lastly, that the granularity of synchronization events is relatively coarse. Cooperation between *Gossip* processes is required so that the workload associated with the synchronization protocol may be evenly distributed. *Gossips* dynamically partition the responsibility for querying and updating application components amongst themselves. For the SC98 experiment, we stationed several *Gossips* at well-known addresses around the country. When an application component registered, it was assigned a responsible *Gossip* within the pool of available *Gossips* whose job it was to keep that component synchronized.

In addition, we allowed the *Gossip* pool to fluctuate. New *Gossip* processes registered themselves with one of the well-known sites and were announced to all other functioning *Gossips*. Within the *Gossip* pool, we used the NWS clique protocol [46] (a token-passing protocol based on leader-election [15, 1]) to manage network partitioning and *Gossip* failure. The clique protocol allows a clique of processes to dynamically partition itself into subcliques (due to network or host failure) and then merge when conditions permit. The EveryWare *Gossip* pool uses this protocol to reconfigure itself and rebalance the synchronization load dynamically in response to changing conditions.

The assumptions about synchronization count and granularity are more restrictive. Because each *Gossip* does a pair-wise comparison of application component state, N^2 comparisons are required for N application components. Moreover, if the overhead associated with state synchronization cannot be amortized by useful computation, performance will suffer. We believe that the prototype state-exchange protocol can be substantially optimized, (or replaced by a more sophisticated mechanism) and careful engineering can reduce the cost of state synchronization over what we were able to achieve. However, we hasten to acknowledge that not all applications or application classes will be able to use EveryWare effectively for Grid computation. Indeed, it is an interesting and open research question as to whether large-scale, tightly synchronized application implementations will be able to extract performance from Computational Grids, particularly if the Grid resource performance fluctuates as much as we have typically observed [47, 45]. EveryWare does not allow any application to become an effective Grid application. Rather, it facilitates the deployment of applications whose characteristics are Grid suitable so that they may draw computational power ubiquitously from a set of fluctuating resources.

Similarly, the consistency model required by the application program dramatically affects its suitability as an EveryWare application, in particular, and as a Grid application in general. The development of a high-performance state replication facilities that implement tight bounds on consistency is an active area of research. EveryWare does not attempt to solve the distributed state consistency problem for all consistency models. Rather, it specifies the inclusion of replication and synchronization facilities as a constituent service. For the application that describe in the next Section (Section 4), we implemented a loosely consistent service based on the *Gossip* protocol. Other, more tightly synchronized services can be incorporated, each with its own performance characteristics. We note, however, that applications having tight consistency constraints are, in general, difficult to distribute while maintaining acceptable performance levels. EveryWare is not intended to change the suitability of these programs with respect to Grid computing, but rather enables their implementation and deployment at what ever performance level they can attain.

3.4 The EveryWare Programming Model

An EveryWare application is structured as a set of *computational clients* that request run-time management services from a set of *application-specific servers*. Figure 2 depicts the structure of an application. Application clients (denoted "A" in the figure) can execute in a number of different environments, such as NetSolve [6], Globus [11], Legion [22], Condor [41], etc. They communicate with application-specific scheduling servers (marked "S" in the figure) to receive scheduling directives dynamically. Persistent state managers tuned for the application (denoted by "P" in the figure) control and protect any program state that must survive host or network failure. Application performance logging servers (marked "L") allow arbitrary messages to be logged by the application. Finally, all application components use the EveryWare *Gossip* service (marked "G") to synchronize state. To anticipate load changes, the various application components consult the Network Weather Service (NWS) — a distributed dynamic performance forecasting service for Computational Grids [46, 45, 35].

This application architecture offers several advantages. First, the overall program can be constructed incrementally. Since the EveryWare toolkit is robust, new clients and servers can be added, old ones removed or improved, etc. while the application itself continues to execute. Since we do not have to restart the application every time we wish to add a new program component, we can improve and evolve the running application dynamically. Another advantage is that it

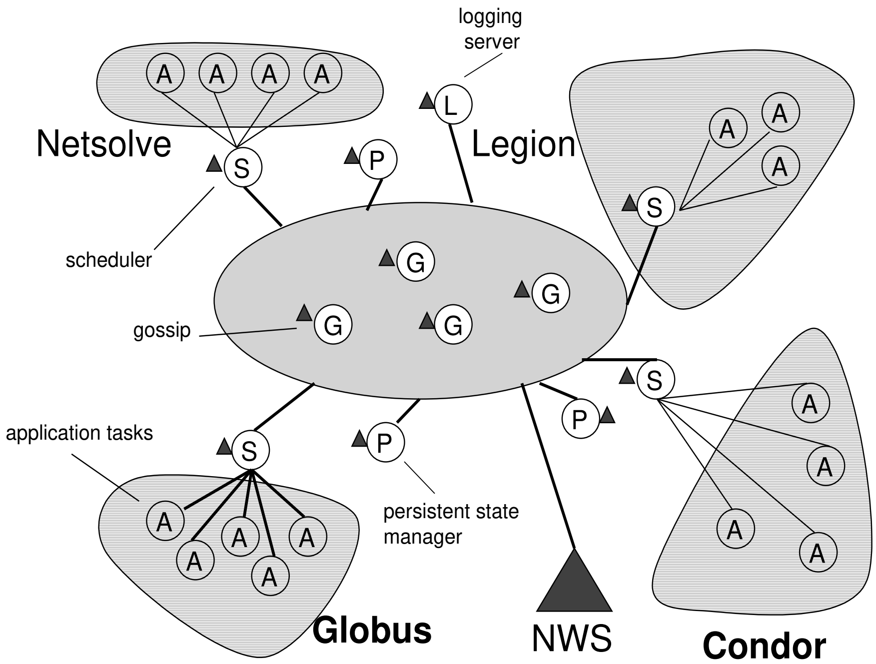


Figure 2: **EveryWare Application Structure**

allows us to implement infrastructure-specific clients that can get the best possible performance by running in "native" mode. Since the clients need only speak the protocol required by each server, we do not need to put a complete software veneer between the computational code and the native infrastructure.

Note that the EveryWare programming model fundamentally different from that used by most procedure oriented Grid infrastructures such as NetSolve [6], Legion [22], and CORBA [36]. These infrastructures typically support applications structured as a single controlling client that makes method or remote-procedure calls to remote computational servers (or remote objects, if the infrastructure is object-oriented). Under the EveryWare programming model, computation is centered at the clients and program control is coordinated by a set of cooperating application-specific servers. This novel application structure allows EveryWare applications is meet the "consistent" and "dependable" qualitative criteria described in [12] by offering greater scalability and robustness than a single-client approach.

4 Example Application: Ramsey Number Search

The application we chose to implement to test the effectiveness of EveryWare attempts to improve the known bounds of classical Ramsey numbers. The n^{th} *classical* or *symmetric Ramsey number* $R_n = R_{n,n}$ is the smallest number k such that any complete two-colored graph on k vertices must contain a complete one-colored subgraph on n of its vertices. It can be proven in a few minutes that $R_3 = 6$; it is already a non-trivial result that $R_4 = 18$, and the exact values of higher R_n are unknown.

Observe that to show that a certain number j is a lower bound for R_n, one might try to produce a particular two-colored complete graph on $(j-1)$ vertices that has no one-colored complete subgraph on any n of its vertices. We will refer to such a graph as a "counter-example" for the n^{th} Ramsey number. Our goal was to find new lower bounds for Ramsey numbers by finding counter-examples.

This application was especially attractive as a first test of EveryWare because of its loose synchronization requirements and its resistance to exhaustive search. For example, if one wishes to find a new lower bound for R_5, one must search in the space of complete two-colored graphs on 43 vertices, since the known lower bound is currently 43 ([37]). Since such a graph has $\binom{43}{2} = 903$ edges, there are $2^{903} > 10^{270}$ different two-colored graphs on 43 vertices. Even if one could examine 10^{12} configurations every second, an exhaustive search would take over 10^{250} years.

Therefore, we must use heuristic techniques to control the search process. Note that this combinatoric complexity makes exhaustive search techniques like those employed in cryptographic factoring [28, 5] ineffective. Rather, the process of counter-example identification is related to distributed "branch-and-bound" state-space searching.

4.1 Application Clients

Our goal was to create a dynamically changing population of computational processes executing different heuristics. Heuristic design is an active area of research in combinatorics [37]. As such, we designed the application to be able to incorporate different heuristic algorithms concurrently, each of which implemented as a single application client. The clients would then use the *lingua franca* to communicate with a set of application servers to receive scheduling directives and state management services.

The heuristics that we used all involved *directed search*, by which we mean the following: On the search space of two-colored complete graphs of a particular size, there is a numerical "score" which assigns to each graph the degree to which it fails to be a counter-example in some suitable sense. There is also a set of manipulations called "moves" (transformations) that one can perform on a particular graph to produce other graphs. The algorithm, then, is roughly to start with an arbitrary graph and perform a sequence of moves with a view toward lowering the score by each successive move. Note that in any such heuristic, it is necessary to provide some possibility of making a move which worsens the score; otherwise, there is the danger that the search will get trapped at a local minimum which is not a global minimum.

In our case, the score assigned to a two-colored graph is simply the number of "violations," or complete one-colored subgraphs on n vertices that it possesses; thus a graph is a counter-example if and only if its score is 0.

Various algorithms employed used slightly different definitions for their moves. The simplest and most common was to change the color of a single edge. Thus, for a graph on 43 vertices possessing 903 edges, there are 903 possible moves that can be made from any given graph. In other algorithms, a move comprised changing the colors of 3 edges. Still other algorithms worked in restricted search spaces which partitioned the edges and only considered those graphs for which all the edges in any given partition were the same color; in such a case a move comprised changing the colors of all the edges within a particular partition.

The two classes of search heuristics employed were those based on *tabu search* [37] and *simulated annealing*. In a tabu search, the algorithm keeps a list (the tabu list) of a fixed length recording the most recent moves that have been made. From a given configuration, it examines all moves not in the tabu list, finds the one that gives the lowest score, and makes and records this move. The tabu list is in place to avoid loops; in practice, some element of randomness is necessary in order to avoid large loops. We employed two variants of the tabu search, namely one that allowed a particular move to be made no more than twice on the list and another that allowed a particular move onto the list if its last appearance was with a different predecessor.

The simulated annealing heuristic mimics the physical behavior of a mass as it undergoes cooling; in this case, the score of a configuration is analogous to the temperature of the mass. Generally, from a given configuration the algorithm chooses a move at random and makes the move if it results in a lower score; otherwise, it rejects the move and chooses another at random from the same configuration. However, the algorithm will accept a random move, regardless of the resulting score, with a small probability that decreases as the score drops; this randomness has the effect of keeping the algorithm from getting trapped in a local minimum.

4.1.1 Scheduling Service

To schedule the EveryWare Ramsey Number application, we use a collection of cooperating, but independent scheduling servers to control application execution dynamically. Each computational client periodically contacts a scheduling server and reports its algorithm type, the IP address of the machine on which it is running, the progress it has made since it last made a scheduling decision, and the amount of time that has elapsed since its last contact. Servers are programmed to issue different control directives based on the type of algorithm the client is executing, how much progress the client has made, and the most recent computational rate of the client.

In addition, the scheduling servers are responsible for migrating work. Clients report the number of monochromatic cliques in the graph they are testing when they check in. If the number is low the server will ask the client for a copy of the graph it is currently considering. If it is high, the server sends the client a better graph and directs it to continue from a different point in the search space. The clients are programmed to randomize their starting point in different ways to prevent the system from dwelling irrevocably in a local minimum. In addition, the thresholds for identifying a “good” graph (one with a low monochromatic clique count), a bad one, and the number of times a good one can be migrated to serve as a new starting point in the search space, are tunable parameters.

The schedulers also make decisions based on dynamic performance forecasting information. If a scheduler predicts that a client will be slow based on previous performance, it may choose to migrate that client's current workload to a machine that it predicts will be faster. Rather than basing that prediction solely on the last performance measurement for each client, the scheduler uses the NWS lightweight forecasting facilities to make its predictions. Note that this methodology is inspired by some of our previous work in building application-level schedulers (AppLeS) [39, 4]. AppLeS is an agent-based approach in which each application is fitted with a customized application scheduler that dynamically manages its execution. For the Ramsey Number Search application, however, a single scheduling agent would have been insufficient to control the entire application, both because it would limit the scalability of the application and because the agent constitutes a single-point-of-failure. We designed an application-specific scheduling service that forms organized and robust, but dynamically changing groups of cooperating processes that can make progress if and when the network partitions. As such, we term this type of scheduling Organized Robust AutoNomous Group Scheduling (ORANGS). ORANGS and AppLeS are, indeed, similar in that they use NWS performance forecasts to make application-specific scheduling decisions. However, the distributed and robust nature of the ORANGS service made it a more appropriate choice for the Ramsey Number Search application.

Notice that, for the Ramsey Number search application, the scheduling service considers the use of *all* available resources. When an application client checks in with a scheduling server, the server evaluates the client in terms of the performance it will be able to deliver to the application (using the forecasting services) and decides on the amount and type of work that client should receive. In *all* cases, the Ramsey Number search clients receive some amount of work to perform. For other applications, however, the scheduling service may decide that the use of a particular resource will hinder rather than aid performance and, hence, should be excluded. Therefore, while resource selection is not an issue for Ramsey Number search, the EveryWare programming model supports its implementation.

Schedulers within the scheduling service communicate non-persistent state amongst themselves via the *Gossip* service. In particular, the IP addresses and port numbers of all servers are circulated so that new server instances can be added dynamically. Clients are furnished with a list of active servers when they make contact so that they can contact alternates in the event of a failed server communication. Similarly, scheduling servers learn of different *Gossip* servers, persistent state managers, and logging servers via *Gossip* updates.

4.1.2 Persistent State Management Service

To improve robustness, we identify three classes of program state within the application: local, volatile-but-replicated, and persistent. Local state is state that can be lost by the application due to machine or network failure (e.g. local variables within each computational client). Volatile-but-replicated state is passed between processes as a result of *Gossip* updates, but it is not written to persistent storage. For example, the up-to-date list of active servers is volatile-but-replicated state. Persistent state must survive the loss of all active processes in the application. The largest counter example that the application has yet to find, for example, is check-pointed as persistent state.

We use a separate persistent state service for three reasons. First, we want to limit the size of the file system footprint left by the application. Many sites restrict the amount of disk storage a guest user may acquire. By separating the persistent storage functionality, we are able to dynamically schedule the application's disk usage according to available capacities.

Secondly, we want to ensure that persistent state is ultimately stored in "trusted" environments. For example, we maintained a persistent state server at the San Diego Supercomputer Center because we were assured of regular tape back-ups and industrial quality file system security.

Lastly, we are able to implement run-time sanity checks on all persistent state accesses. If a process attempts to store a counter example, for example, the persistent state manager first checks to make sure the stored object is, indeed, a Ramsey counter example for the given problem size.

To implement this functionality, all persistent state objects must be typed. For each persistent type used in the program, the state manager needs a set of sanity-checks (performed when an object is accessed) and a comparator operator so that the state may be synchronized by the *Gossip* service. We acknowledge that developing this functionality for all Grid applications may not be possible. However, we note that many Computational Grid infrastructures currently support mechanisms that can be used to implement the state management functionality we require for Ramsey Number search. For example, the sanity checks performed by the state manager were implemented, primarily, to prevent errant or malicious processes from damaging program state. Instead, Globus authentication mechanisms [13] could be used to provide access control so that only trusted processes may modify persistent state. Similarly, the Legion class management system [29] tracks object instances in a way that could be used to identify stale state. We wanted to ensure that all application components (computational clients and application-specific servers) would be portable to any environment so we did not choose to rest any of the application's functionality on a particular infrastructure. Future versions of the Ramsey Search application may relax this restriction to further benefit from maturing Computational Grid technologies.

4.1.3 Logging Service

To track the performance of the application dynamically, we implemented a distributed logging service. Scheduling servers base their decisions, in part, on performance information they receive from each computational client. Before the information is discarded, it is forwarded to a logging server so that it can be recorded. Having a separate service, again, allows us to limit and control the storage load generated by the application. For example, the National Partnership for Advanced Computational Infrastructure (NPACI) loaned our group a pair of file servers so that we could capture a performance log that spanned the time of the conference.

As with the persistent state managers and the scheduling servers, the logging servers register themselves with the *Gossip* service. Any application process wishing to log performance information learns of a logging server through the server list that is circulated. The logging servers do not register a state synchronization function, however. They use the *Gossip* service only to join the running application.

5 Results

To test the efficacy of our approach, we deployed the Ramsey Number search application on a globally distributed set of resources during SC98. As part of the test, we entered EveryWare in the High-performance Computing Challenge [25] (an annual competition held during the conference) as we believed that the fluctuating loads generated by our competitors would test the capabilities of our system vigorously.

We instrumented each application client to maintain a running count of the computational operations it performs so that we could monitor the performance of Ramsey Number search application. The bulk of the work in each of the heuristics (see Section 4) are integer test and arithmetic instructions. Since each heuristic has an execution profile that depends largely on the point in the search space where it is searching, we were unable to rely on static instruction count estimates. Instead, we inserted counters into each client after every integer test and arithmetic operation. Since the ratio of instrumentation code to computational code is essentially one-to-one (one integer increment for every integer operation) the performance estimates we report are conservative. Moreover, we do not include any instrumentation instructions in the operation counts nor do we count the instructions in the client interface to EveryWare - only "useful" work delivered to the application is counted. Similarly, we include all communication delays incurred by the clients in the elapsed timings. The computational rates we report include all of the overheads imposed by our software architecture and the ambient loading conditions experienced by the program during SC98. That is, all of the results we report in this section are conservative estimates of the sustained performance *delivered* to the application during the experiment.

5.1 Sustained Execution Performance

As a Computational Grid experiment, we wanted to determine if we could obtain high application performance levels from widely distributed, heavily used, and non-dedicated computational resources. In Figure 3, we show the sustained execution performance of the entire application during the twelve-hour period including and immediately preceding the judging of our High-performance Computing Challenge entry at at SC98 on November 12, 1998[3]. The $x - axis$ shows the time of day, Pacific Standard Time [4], and the $y - axis$ shows the average computational rate over a five-minute time period. The highest rate that the application was able to sustain was 2.39 billion integer operations between 09:51 and 09:56 during a test an hour before the competition (right-hand side of the graph). The judging for the competition itself (which required a "live" demonstration) began at 11:00. As several competing projects were being judged simultaneously, and many of our competitors were using the same resources we were using, the networks interlinking the resources suddenly experienced a sharp load increase. Moreover, many of the competing projects required dedicated access for their demonstration. Since we deliberately did not request dedicated access, our application suddenly lost computational power (as resources we claimed by

[3] We demonstrated the system for a panel of judges between 11:00 AM and 11:30 AM PST.

[4] SC98 was held in Orlando, Florida which is in the Eastern time zone. Our logging and report facilities, primarily located at stable sites on the west coast, used Pacific Standard Time. As such, we report all time-of-day values in PST.

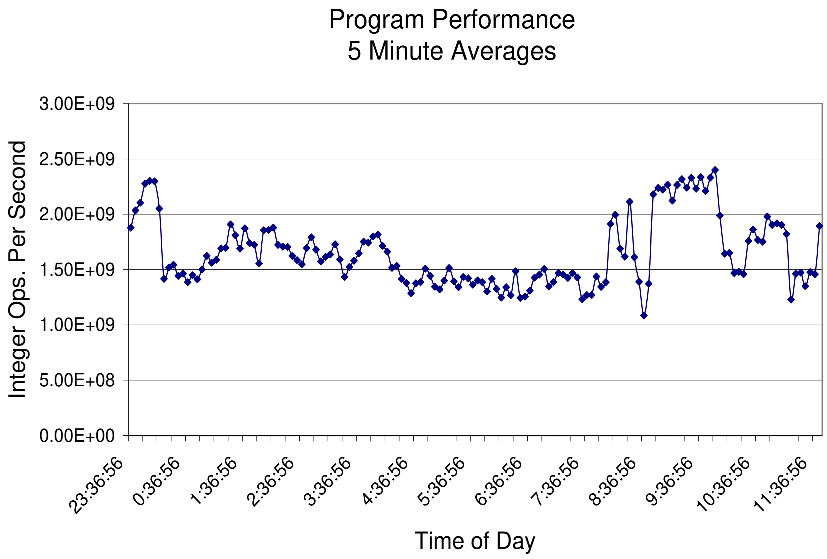


Figure 3: **Application Speed**

other applications) as the communication overheads rose (due to increased communication load). The sustained performance dropped to 1.1 billion operations as a result. The application was able to adapt to the performance loss and reorganize itself so that by 11:10 (when the demonstration actually took place), the sustained performance had climbed to 2.0 billion operations per second.

This performance profile clearly demonstrates the potential power of Computational Grid computing. With non-dedicated access, under extremely heavy load conditions, the EveryWare application was able to sustain supercomputer performance levels.

In Figure 4, we show the number of hosts used during the same time period. In this figure, each data point represents the number hosts checking in during the corresponding five-minute period[5]. Note that the maximum host count (266) occurs at 23:51 as we ran a large scale test of the system the night before the competition. However, the maximum host count does not correspond to the maximum sustained rate. While we were able to incorporate many new and powerful resources on the morning of the competition, we lost some of the workstations that were loaned to us by Condor during the night. Also, these host count numbers are based on unique IP addresses (and not process id) making them very conservative. Since some systems use the same

[5]The maximum time between check-ins for any computational client was set to five minutes during the test.

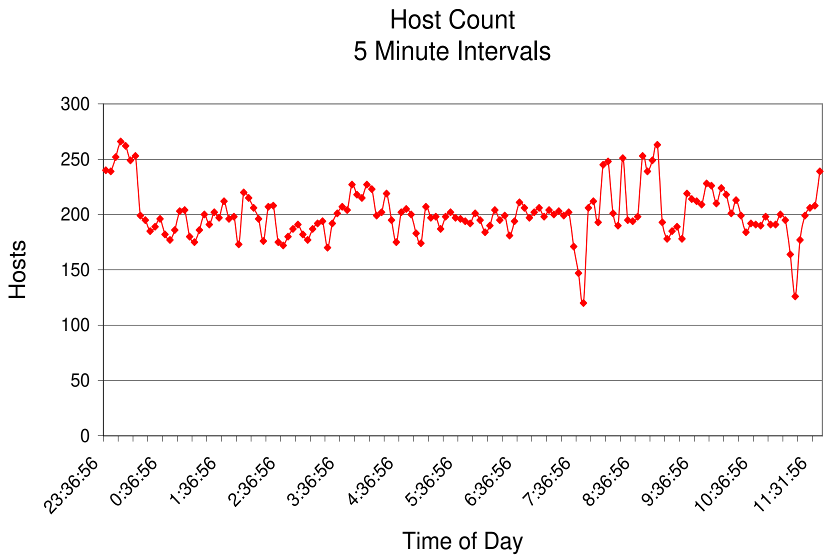
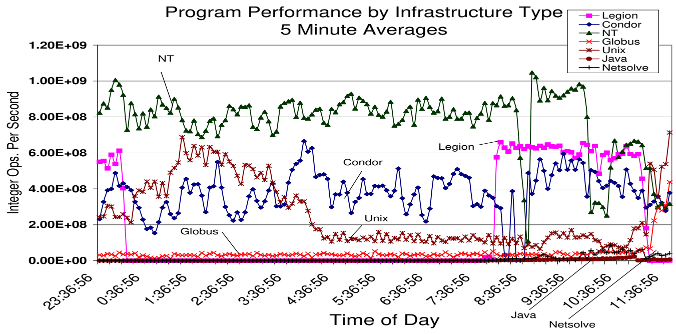


Figure 4: **Application Host Count**

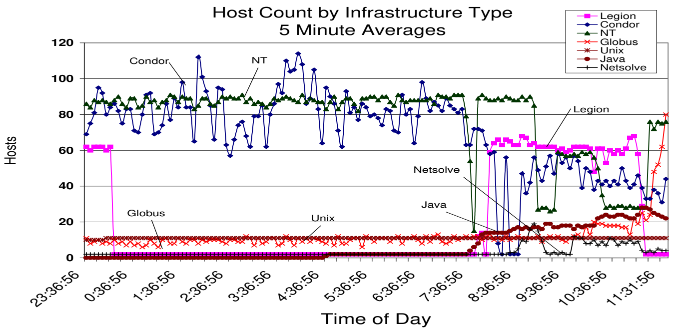
IP address for all hosts (e.g. the NT Supercluster) the actual host population was much higher. However, we could not distinguish between multiple processes on different hosts with the same IP address, and multiple process restarts due to eviction for the combined host population. As a result, we report the more conservative estimates.

5.2 Performance Response

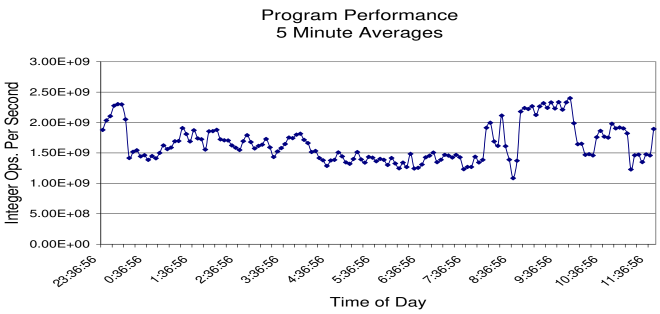
We also wanted to measure the smoothness of the performance response the application was able to obtain from the Computational Grid. For the Grid vision to be implemented, an application must be able to draw "power" uniformly from the Computational Grid as a whole despite fluctuations and variability in the performance of the constituent resources. In Figures 5 and 6 we compare the overall performance response obtained by the application (graph (c) in both figures) with the performance and resource availability provided by each infrastructure. Figure 5 makes this comparison on a linear scale and Figure 6 shows the same data on a log scale so that the wide range of performance variability may be observed. In Figures 5a and 6a we detail the number of cycles we were able to successfully deliver from each Grid infrastructure during the twelve hours leading up to the competition. Similarly, in Figure 5b, we show the host availability from each



(a)



(b)



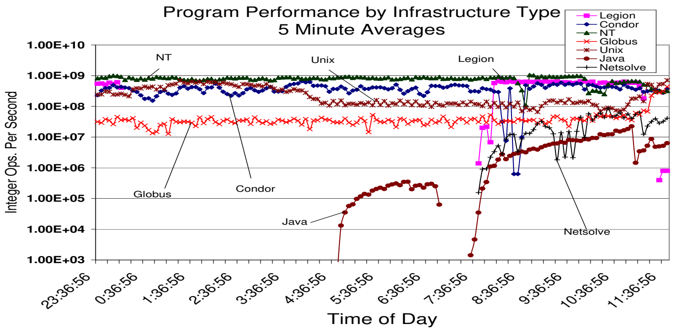
(c)

Figure 5: Sustained Processing Rate by Infrastructure (a), Host Count by Infrastructure (b), and Total Sustained Rate (c)

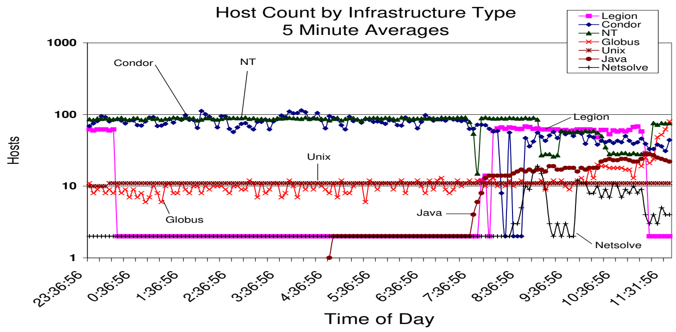
infrastructure for the same time period. Together, these graphs provide insight into the diversity of the resources we used in the SC98 experiment.

Specifically, Condor supports a dynamic loan-and-reclaim resource usage model. Users agree to loan idle workstations to the Condor system for use by other processes. When a user-specified keyboard activity or load threshold is exceeded, the resource is declared busy and any Condor jobs that are running at the time are evicted. Note that Condor processing power and host count fluctuated through the night and then fell off as the day began in Wisconsin and user activity caused their workstations to be reclaimed. For Java, the performance trajectory was the opposite. We fitted the Java applets with the necessary logging features at approximately 4:30 AM, although we had a small number of test hosts running before then. At approximately 8:00 AM, we announced the availability of the Java implementation and solicited participation from "friendly" sites. In addition, we began to execute the Java applet using HotJava [20] on workstations that had been brought to SC98 for general use by conference attendees. At about the same time, Legion (which had been down since approximately midnight) became available again and the application immediately began to take advantage of the newly available resources. Our Globus utilization, however, was low until just after the competition ended at 11:30 AM, when it suddenly spiked. The Globus group entered the High-performance Computing Challenge with two separate entries. As we did not request dedicated access or special access priority for the demonstration, our application was able to leverage these resources only after higher-priority Globus processes finished. NetSolve gave us access to the student workstation laboratories and several resources in the Innovative Computing Laboratory at the University of Tennessee. We detected a bug in the performance logging portion of the NetSolve implementation at approximately 8:00 AM, hence we have no reliable performance numbers to report for the period before then. The bulk of the NT hosts we were able to leverage came from the Superclusters [40] located at the National Computational Science Alliance (NCSA) and in the the Computer Systems Architecture Group [7] (CSAG) located at the University of California, San Diego. These systems used batch queues to provide space-shared access to their processors. Unix host count remained relatively constant throughout the experiment, but performance jumped at the end as the Tera MTA (the fastest Unix host) was added to the resource pool.

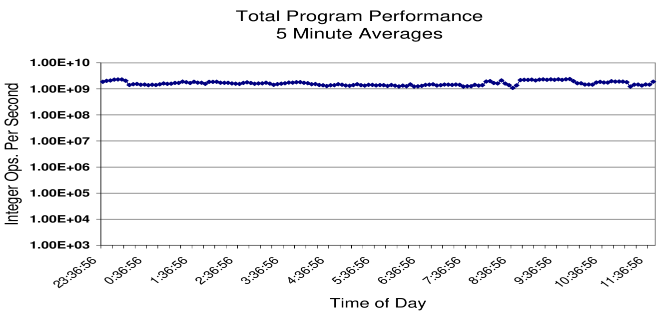
In Figure 5c we reproduce Figure 3 for the purpose of comparison. Figure 6c shows this same data on a log scale. By comparing graphs (a) and (b) to (c) on each scale we expose the degree to which EveryWare was able to realize the Computational Grid paradigm. **Despite fluctuations in the deliverable performance and host availability provided by each infrastructure, the application itself was able to draw power from the overall resource pool relatively uniformly**. As such, we believe the EveryWare example constitutes the first application to be written that successfully demonstrates the potential of high-performance Computational Grid computing. It is one of the first examples of a truly adaptive Grid program. For this accomplishment, the EveryWare experiment was awarded "Best Acceleration" at SC98 by the High-performance Computing Challenge panel of judges.



(a)



(b)



(c)

Figure 6: Log Scale – Sustained Processing Rate by Infrastructure (a), Host Count by Infrastructure (b), and Total Sustained Rate (c)

5.3 Aggregate Performance

Figure 7 shows the total number of integer operations the application was able to obtain during the twelve hours before the competition (on a log scale). With the exception of Java and NetSolve, all

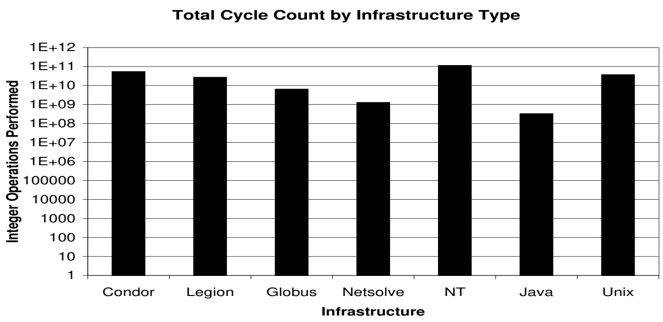
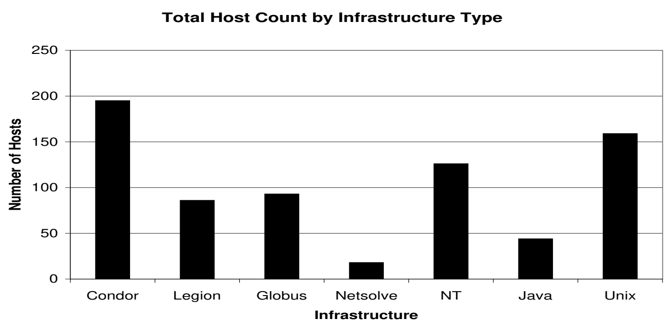


Figure 7: **Total Cycle Count by Infrastructure**

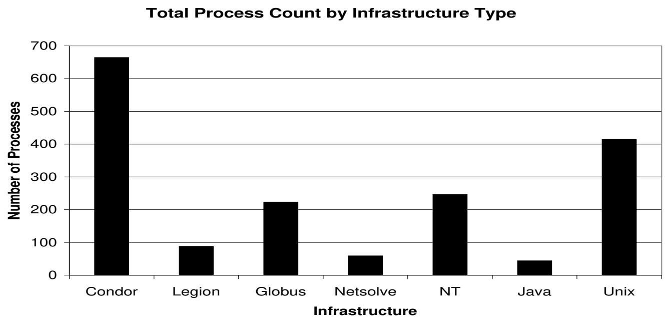
infrastructures were within an order of magnitude in terms of the cycles they delivered. Interpreted Java applet performance was typically between 3 and 5 times slower than native binary execution, and the NetSolve computational servers were shared by other NetSolve jobs and student projects.

5.4 Robustness

We also wished to measure the robustness of our approach. High-performance computer users often complain about application sensitivity to resource failure in distributed environments. Figure 8a shows the total number of hosts controlled by each infrastructure that were used by the application during the twelve hours leading up to the competition and Figure 8b shows the number of processes the system used during the same period. Comparing the two gives an indication of the process failure and restart rate during the experiment. Each computational client was programmed to run indefinitely so, in the absence of process failure, the number of processes would equal the number of hosts. We implemented several "ad-hoc" process restart mechanisms for the environments in which they were not automatic. However, most of the process restarts were due either to deliberate termination on our part while debugging, or dynamic resource reclamation by resource owners. On the Condor system, we ran each computational client as a "vanilla" job which is terminated without notice when the resource on which it is running is reclaimed, and subsequently restarted when another suitable resource is free. It is interesting that, despite the midweek daytime usage, process restart due to resource reclamation was relatively infrequent in the Condor environment during the experiment. The Globus comparison illustrates the power if the GRAM interface [11]. Globus allows all processes to be launched and terminated through a single GRAM request. During the time leading up to the competition, we were improving and



(a)



(b)

Figure 8: Total Host Count by Infrastructure (a), Total Process Count by Infrastructure (b)

debugging our Globus implementation. Having a single control point allowed us to restart large batches of processes easily. Under Legion, the concept of process is not defined. Instead, class "instances" move between blocked and running states (and vice versa) so we simply report the number of instances we used during the demonstration. As a result this level of process restart activity is an estimate. The numbers are accurate for the Globus, Condor, and Unix environments but somewhat ambiguous for the other infrastructures. Despite the level of process failure we were able to detect, however, we were able to obtain the sustained processing rates shown in Figure 3 during the same time period.

Indeed, EveryWare and the application design we used proved to be quite robust. In Figure 9 we show host counts over five-minute intervals during the 17 days prior to the judging on November 12. Some portion of the application was executing, more or less continuously during the entire period (there are no horizontal gaps in the figure). As we concentrated our initial efforts on developing the EveryWare toolkit and new Ramsey search heuristics, we did not add performance logging to the running system until October 26. The program had actually been running continuously since early June of 1998, however we only have performance data dating from the end of October. Note that we were able to add, and then completely revise, the performance

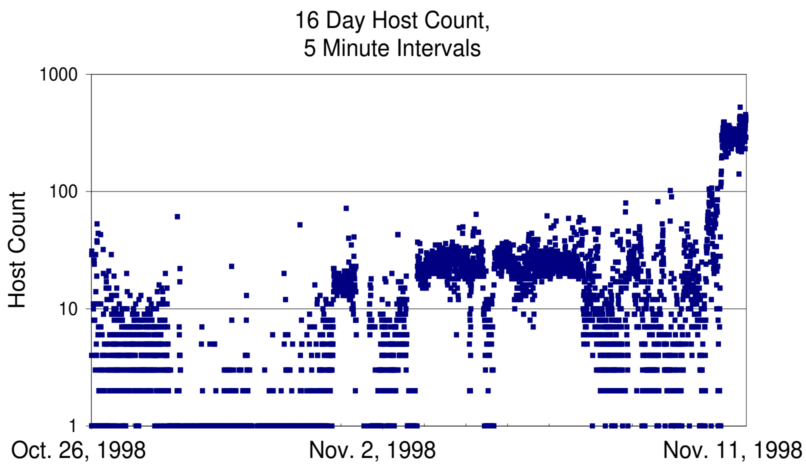


Figure 9: **Sixteen-day Host Counts**

logging service while the program was in execution.

5.5 Resource Diversity

For the Computational Grid paradigm to succeed, all useful resources must be accessible by the application. Metaphorically speaking, all profitable methods of power generation must be usable by any power consumer. Figure 10 compares the deliverable performance from the fastest host controlled by each infrastructure. The values not only benchmark our code on various architectures, but provide insight into the wide range of resource options we were able to leverage during the experiment. In each case, we attempted to use the native, vendor-specific C compiler (as opposed to GNU `gcc`) with all optimization capabilities enabled. On the left-hand side of the figure, we compare the best performance from each infrastructure. The fastest Unix machine was the Tera MTA [42]. We report only the single processor performance, however the Tera was also able to automatically parallelize the code and achieve an almost linear speed-up on two processors. The fastest NT-based machine was was located at the University of Wisconsin, but we are unable to determine its architectural characteristics. An unknown participant downloaded the NT binary from the EveryWare home page when we announced that the system was operational

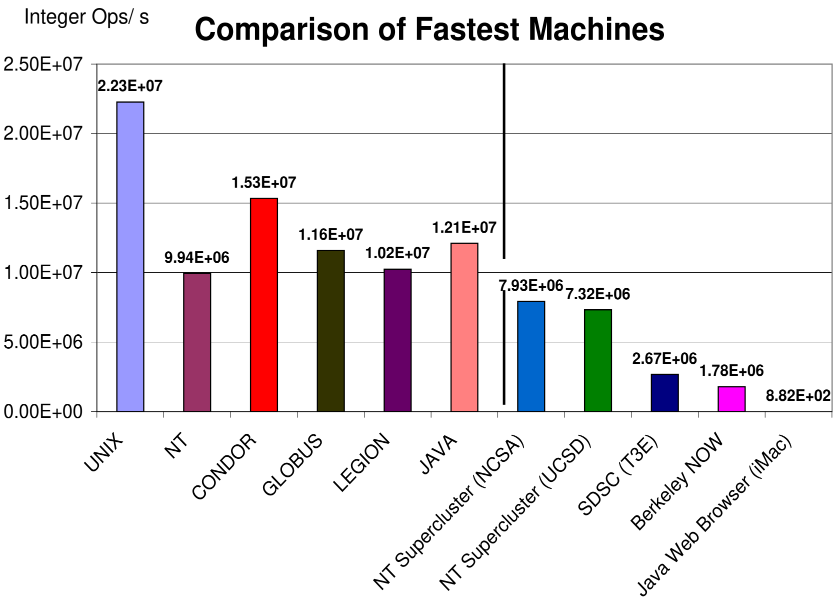


Figure 10: **Host Speeds**

on Wednesday morning. The fastest Condor machine was a Pentium P6 running Solaris was also located at the University of Wisconsin. Single processor Pentium P6 performance was particularly good (second only to the Tera) for the integer-oriented search heuristics we developed. The fastest Legion host was a Digital Equipment Corporation Alpha processor running Red Hat Linux, located at the University of Virginia and the fastest Globus machine was an experimental Convex V class host located at the Convex development facility in Richardson, Texas. Surprisingly, the fastest Java execution was faster than the fastest NT, Legion, and Globus machines. An unknown participant at Kansas State University loaded the applet and executed it using Microsoft Corporation's Internet Explorer (the IP address is for a 300Mhz dual-processor (Pentium II) machine running NT with Explorer). We speculate that a student used some form of just-in-time compilation technology to achieve the execution performance depicted in the figure, although we are unable to ascertain how this performance level was reached.

On the right-hand side of the figure, we show the best single-processor performance of other interesting and popular machines. The NT Superclusters at UCSD and NCSA generated almost identical per-node processing rates. A single node of the Cray T3E located at the San Diego Supercomputer Center was able to run only slightly faster than a single node of the Berkeley NOW [8]. This comparison surprised us since the T3E is space shared (meaning that each process had exclusive access to its processor once it made it through the batch queue) and the NOW (which is timeshared) was heavily loaded. On the far right-hand side of the figure, we show the speed of a publicly accessible Apple iMac workstation located in a coffee shop on the UCSD campus which is typical of the interpreted Java performance we were able to achieve.

In addition to detailing the relative performance of different architectures and infrastructures, Figure 10 demonstrates the utility of EveryWare. It would not have been possible to include experimental (and powerful) resources such as the Tera MTA and the NT Superclusters without the EveryWare toolkit. At the time of the experiment, none of the existing Grid infrastructures had been ported to either architecture. We were able to port EveryWare to both systems quickly (under 30 minutes for the Tera) allowing us to couple them with other, more conventional hosts that did support some form of Grid infrastructure. By providing execution ubiquity, EveryWare was able to leverage resources that no other Grid computing infrastructure could access. As such, **the Ramsey Number Search application is the first program to couple the Tera MTA, both NT Superclusters, and the Berkeley NOW with parallel supercomputers such as the Cray T3E, workstations, and desktop web browsers**. We were able to successfully span the spectrum of available computing platforms with EveryWare.

6 Related Work

EveryWare is a toolkit that allows an application to leverage dynamically a variety of high- and low-level execution environments for performance. As such, it shares common goals with many of the infrastructures it leverages. It is similar to Globus [11] in that application components communicate via different well-defined protocols to obtain Grid "service." EveryWare extends this notion, however, by providing a way for the Grid programmer to develop application-specific protocols and services so that the application (and not just the underlying infrastructure) can be robust and ubiquitous.

EveryWare supports information hiding and location transparency in the same way object-oriented systems such as Legion [22] and CORBA [36] do. Application components can be implemented as objects that communicate via remote method invocations. A key difference is that an EveryWare program can also use messaging protocols at the same time to leverage external facilities.

EveryWare complements the functionality provided by Condor [41] by providing a robust messaging layer. In the SC98 experiment, we deployed the Ramsey Number Search application as a "vanilla" program within the Condor environment. Vanilla Condor programs do not use the checkpoint and restart facilities nor do they have their system calls redirected. This implementation strategy allowed a single program to span multiple Condor pools, each containing a different architecture type. At present, Condor's checkpointing facilities restrict migration to a single architecture type, but we wished to leverage all of the architecture types in any resource pool configured to use Condor. In [3], the authors use Condor to deliver a large number of cycles over an extended period to a high-throughput application. Our effort differs from this experiment, both in its scope and in its goal. We were able to embed the Condor pools available to us in a larger, more heterogeneous Computational Grid. Moreover, we were using Condor to increase the performance of the application as measured by turn-around time. The application we chose arrives at its answers faster with the addition of resources if those resources are carefully scheduled.

EveryWare shares the notion of brokered invocation with NetSolve [6], NINF [33], and NEOS [30]. A scheduling entity chooses resources to use and manages application execution while the application executes. For robustness and performance, however, we chose to implement an application-specific scheduling service for the Ramsey Number Search application. Having a single scheduling agent for a globally distributed application presents both a performance bottle-neck and a single-point-of-failure for the application. The state exchange facilities in EveryWare make it possible to build a distributed scheduling service to avoid these problems. AppLeS [4] takes a similar single-agent approach to NetSolve, but the agent is tailored to meet the application's performance needs. We extend the application-specific scheduling concept developed for AppLeS by building a high-performance, robust, and distributed scheduling service for the application.

Prophet [43] is a system designed to schedule SPMD applications on networks of computers. Originally designed for the Mentat [21] parallel and distributed computing system for local area deployment, its designers have recently extended it to wide area settings. Unlike EveryWare, however, it relies on a single ubiquitous infrastructure. Moreover, it does not support the robustness features and state-exchange functionality provided as part of EveryWare. that

7 Conclusions and Future Work

By leveraging a heterogeneous collection of Grid software and hardware resources, dynamically forecasting future resource performance levels, and employing relatively simple distributed state management techniques, EveryWare has enabled the first application implementation that meets the requirements for Computational Grid computing. In [12](page 18) the authors describe the criteria that a Computational Grid must fulfill as the provision of *pervasive*, *dependable*, *consistent*, and *inexpensive* computing.

- **Pervasive** — At SC98, we were able to use EveryWare to execute a high-performance, globally distributed program on machines ranging from the Tera MTA to a web browser located in a campus coffee shop at UCSD.

- **Dependable** — The Ramsey Number Search application ran continuously from early June, 1998, until the High-Performance Computing Challenge on November 12, 1998.

- **Consistent** — During the twelve hours leading up to the competition itself, the application was able to draw uniform compute power from resources with widely varying availability and performance profiles.

- **Inexpensive** — All of the resources used by the Ramsey Number Search application were non-dedicated and accessed via a non-privileged user login.

To our knowledge, EveryWare is the first Grid software effort that has been able to successfully meet these criteria, and to demonstrate the degree to which they are met quantitatively.

We plan to study how EveryWare can be used to implement other Grid applications as part of our future efforts. In particular, we plan to use it to build Grid versions of a medical imaging code written at the University of Tennessee, and a data mining application from the University of Torino. We also plan to extend ORANGS to include storage scheduling directives and memory constraints. Finally, we plan to leverage our experience with EveryWare to build new Network Weather Service sensors for different Grid infrastructures.

8 Acknowledgements

It is impossible to acknowledge and thank adequately all of the people and organizations that helped make the EveryWare demonstration at SC98 a success. As such, we miserably fail in the attempt by expressing our gratitude to the AppLeS group at UCSD for enduring weeks of maniacal behavior. In particular, we thank Fran Berman for her moral support during the effort, and Marcio Faerman, Walfredo Cirne, and Dimitri Zagorod for launching EveryWare on every conceivable public email and Java workstation at SC98 itself. We thank NPACI for supporting our High-performance Challenge entry in every way and, in particular, Mike Gannis for enthusiastically making the NPACI booth at SC98 ground-zero for EveryWare. Rob Pennington at NCSA left no stops unpulled on the NT Supercluster so that we could run and run fast, and Charlie Catlett, once again, made it all happen at "The Alliance." We inadequately thank Miron Livny (the progenitor of Condor and the University of Wisconsin) for first suggesting and then insisting that EveryWare happen. Henri Casanova, at UCSD, single-handedly ported EveryWare to NetSolve after an off-handed mention of the project was carelessly made by a project member within his range of hearing. Steve Fitzgerald, at Cal State Northridge and ISI/USC introduced us to the finer and more subtle pleasures of Globus, as did Greg Lindahl for analogously hedonistic experiences with Legion. Brent Gorda and Ken Sedgewick at MetaExchange Corporation donated entirely too much time, space, coffee, good will, more coffee, sound advice, and patience to the effort. Allen Downey and the Colby Supercomputer Center provided us with cycles, encouragement, and more encouragement. Cosimo Anglano of Dipartimento di Informatica, Università di Torino provided us with intercontinental capabilities and tremendously spirited support. Lastly, we thank EveryOne who participated anonymously via our web interface and downloads. We may not know who you are, but we know your IP addresses, and we thank you for helping us through them.

References

[1] H. Abu-Amara and J. Lokre. Election in asynchronous complete networks with intermittent lin k failures. *IEEE Transactions on Computers*, 43(7):778–788, 1994.

[2] O. Arndt, B. Freisleben, T. Kielmann, and F. Thilo. Scheduling parallel applications in networks of mixed uniprocessor/multiprocessor workstations. In *Proceedings of ISCA 11th Conference on Parallel and Distributed Computing*, September 1998.

[3] J. Basney, M. Livny, and T. Tannenbaum. Deploying a high throughput computing cluster. In R. Buyya, editor, *High Performance Cluster Computing*, volume 1, chapter 5. Prentice Hall, 1999.

[4] F. Berman, R. Wolski, S. Figueira, J. Schopf, and G. Shao. Application level scheduling on distributed heterogeneous networks. In *Proceedings of Supercomputing 1996*, 1996.

[5] The bovine rc5-64 project – `http://distributed.net/rc5/`.

[6] H. Casanova and J. Dongarra. NetSolve: A Network Server for Solving Computational Science Problems. *The International Journal of Supercomputer Applications and High Performance Computing*, 1997.

[7] The concurrent systems architecture group – `http://www-csag.ucsd.edu/`.

[8] D. Culler, A. Arpaci-Dusseau, R. Arpaci-Dusseau, B. Chun, S. Lumetta, A. Mainwaring, R. Martin, C. Yoshikawa, and F. Wong. Parallel computing on the berkeley now. In *to appear in JSPP'97 (9th Joint Symposium on Parallel Processing)*, 1997. available from `http://now.CS.Berkeley.EDU/Papers2`.

[9] L. DeRose, Y. Zhang, and D. Reed. Svpablo: A multi-language performance analysis system. In *Proceedings of 10th International Conference on Computer Performance Evaluation*, September 1998.

[10] M. P. I. Forum. Mpi: A message-passing interface standard. Technical Report CS-94-230, University of Tennessee, Knoxville, 1994.

[11] I. Foster and C. Kesselman. Globus: A metacomputing infrastructure toolkit. *International Journal of Supercomputer Applications*, 1997.

[12] I. Foster and C. Kesselman. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers, Inc., 1998.

[13] I. Foster, C. Kesselman, G. Tsudik, and S. Tuecke. Security architecture for computational grids. In *Proc. 5th ACM Conference on Computer and Communications Security Conference*, pages 83–92, 1998.

[14] I. Foster, C. Kesselman, and S. Tuecke. The nexus approach to integrating multithreading and communication. *Journal of Parallel and Distributed Computing*, 1996.

[15] H. Garcia-Molina. Elections in a distributed computing system. *IEEE Transactions on Computers*, C-31(1):49–59, Jan 1982.

[16] J. Gehrinf and A. Reinfeld. Mars - a framework for minimizing the job execution time in a metacomputing environment. *Proceedings of Future general Computer Systems*, 1996.

[17] A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, and V. Sunderam. *PVM: Parallel Virtual Machine A Users' Guide and Tutorial for Networked Parallel Computing*. MIT Press, 1994.

[18] Gusto webpage at `http://www-fp.globus.org/testbeds/`.

[19] J. Gosling and H. McGilton. The java language environment white paper, 1996.

[20] J. Gosling and H. McGilton. The java language environment white paper; chapter 10, 1996.

[21] A. Grimshaw. Easy-to-use object-oriented parallel programming with mentat. *IEEE Computer*, May 1993.

[22] A. S. Grimshaw, W. A. Wulf, J. C. French, A. C. Weaver, and P. F. Reynolds. Legion: The next logical step toward a nationwide virtual computer. Technical Report CS-94-21, University of Virginia, 1994.

[23] W. Gropp, E. Lusk, N. Doss, and A. Skjellum. A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Computing*, 22(6):789–828, Sept. 1996.

[24] T. Haupt, E. Akarsu, G. Fox, and W. Furmanski. Web based metacomputing. Technical Report SCCS-834, Syracuse University Northeast Parallel Architectures Center, 1999. available from `http://www.npac.syr.edu/techreports/html/0800/abs-0834.html`.

[25] The high-performance computing challenge at sc98 – `http://www.supercomp.org/sc98/hpcc/`.

[26] The nasa information power grid – `http://science.nas.nasa.gov/pubs/nasnews/97/09/ipg.html`.

[27] R. Jones. `http://www.cup.hp.com/netperf/netperfpage.html`. Netperf: a network performance monitoring tool.

[28] A. Lenstra and M. Manasse. Factoring by electronic mail. In *Advances in Cryptology – EUROCRYPT '89*, pages 355–371, 1990.

[29] M. J. Lewis and A. S. A. Grimshaw. Using dynamic configurability to support object-oriented programming languages and systems in legion. Technical Report CS-96-19, University of Virginia, 1996.

[30] J. M. M. Ferris, M. Mesnier. Neos and condor: Solving optimization problems over the internet. Technical Report ANL/MCS-P708-0398, Argonne National Laboratory, March 1998. available from http://www-fp.mcs.anl.gov/otc/Guide/TechReports/index.html.

[31] S. Microsystems. Xdr: External data representation, 1987. ARPA Working Group Requests for Comment DDN Network Information Center, SRI International, Menlo Park, CA, RFC-1014.

[32] B. Miller, M. Callaghan, J. Cargille, J. Hollingsworth, R. Irvin, K. Karavanic, K. Kunchithapadam, and T. Newhall. The paradyn parallel performance measurement tools. *IEEE Computer*, 28(11):37–46, Nov. 1995.

[33] H. Nakada, H. Takagi, S. Matsuoka, U. Nagashima, M. Sato, and S. Sekiguchi. Utilizing the metaserver architecture in the ninf global computing system. In *High-Performance Computing and Networking '98, LNCS 1401*, pages 607–616, 1998.

[34] The nt supercluster at ncsa – http://www.ncsa.uiuc.edu/general/cc/ntcluster/.

[35] The network weather service home page – http://nws.npaci.edu.

[36] "OMG". The complete formal/98-07-01: The corba/iiop 2.2 specification, 1998.

[37] S. Radziszowski. Small ramsey numbers. In *Dynamic Survey DS1 – Electronic Journal of Combinatorics*, volume 1, page 28, 1994.

[38] R. L. Ribler, J. S. Vetter, H. Simitci, and D. A. Reed. Autopilot: Adaptive control of distributed applications. In *Proc. 7th IEEE Symp. on High Performance Distributed Computing*, Aug 1998.

[39] N. Spring and R. Wolski. Application level scheduling: Gene sequence library comparison. In *Proceedings of ACM International Conference on Supercomputing 1998*, July 1998.

[40] N. SuperCluster. verb+http://www.ncsa.uiuc.edu/general/cc/ntcluster+.

[41] T. Tannenbaum and M. Litzkow. The condor distributed processing system. *Dr. Dobbs Journal*, February 1995.

[42] The tera mta – http://www.tera.com.

[43] J. Weissman and X. Zhao. Scheduling parallel applications in distributed networks. *Concurrency: Practice and Experience*, 1(1), 1998.

[44] Microsoft windows nt operating system – http://www.microsoft.com/ntserver/nts/techdetails/overview/wpglobal.asp/.

[45] R. Wolski. Dynamically forecasting network performance using the network weather service. *Cluster Computing*, 1998. also available from http://www.cs.utk.edu/~rich/publications/nws-tr.ps.gz.

[46] R. Wolski, N. Spring, and J. Hayes. The network weather service: A distributed resource performance forecasting service for metacomputing. *Future Generation Computer Systems (to appear)*, 1999. available from http://www.cs.utk.edu/~rich/publications/nws-arch.ps.gz.

[47] R. Wolski, N. Spring, and J. Hayes. Predicting the cpu availability of time-shared unix systems on the computational grid. In *Proc. 8th IEEE Symp. on High Performance Distributed Computing*, 1999. available from http://www.cs.utk.edu/~rich/publications/nws-cpu.ps.gz.