The Organic Grid: Self-Organizing Computation on a Peer-to-Peer Network

Arjav J. Chakravarti, Gerald Baumgartner, Mario Lauria

***Abstract*— Desktop grids have been used to perform some of the largest computations in the world and have the potential to grow by several more orders of magnitude. However, current approaches to utilizing desktop resources require either centralized servers or extensive knowledge of the underlying system, limiting their scalability.**

We propose a new design for desktop grids that relies on a self-organizing, fully decentralized approach to the organization of the computation. Our approach, called the Organic Grid, is a radical departure from current approaches and is modeled after the way complex biological systems organize themselves. Similarly to current desktop grids, a large computational task is broken down into sufficiently small subtasks. Each subtask is encapsulated into a mobile agent, which is then released on the grid and discovers computational resources using autonomous behavior. In the process of "colonization" of available resources, the judicious design of the agent behavior produces the emergence of crucial properties of the computation that can be tailored to specific classes of applications.

We demonstrate this concept with a reduced-scale proof-of-concept implementation that executes a data-intensive independent-task application on a set of heterogeneous, geographically distributed machines. We present a detailed exploration of the design space of our system and a performance evaluation of our implementation using metrics appropriate for assessing self-organizing desktop grids.

***Index Terms*— Mobile agents, grid scheduling, self-organizing computation.**

I. Introduction

Many scientific fields, such as genomics, phylogenetics, astrophysics, geophysics, computational neuroscience, or bioinformatics, require massive computational power and resources, which might exceed those available on a single supercomputer. There are two drastically different approaches for harnessing the combined resources of a distributed collection of machines: large-scale desktop-based master-worker schemes and more traditional computational grid schemes.

Some of the largest computations in the world have been carried out on collections of PCs and workstations over the Internet. Tera-flop levels of computational power have been achieved by systems composed of heterogeneous computing resources that number in the hundreds-of-thousands to the millions. This extreme form of distributed computing is often called *internet computing*, and has allowed scientists to run applications at unprecedented scales at a comparably modest cost. The desktop-based platforms on which Internet-scale computations are carried out are often referred to as *desktop grids*. In analogy to computational grids [1], [2], these collections of distributed machines are glued together by a layer of middleware software that provides the illusion of a single system [3], [4], [5]. While impressive, these efforts only use a tiny fraction of the desktops connected to the Internet. Order of magnitude improvements could be achieved if novel systems of organization of the computation were to be introduced that overcome the limits of present systems.

A number of large-scale systems are based on variants of the master/workers model [6], [3], [7], [4], [5], [8], [9], [10], [11], [12], [13]. The fact that some of these systems have resulted in commercial enterprises shows the level of technical maturity reached by the technology. However, the obtainable computing power is constrained by the performance of the master (especially for data-intensive applications) and by the difficulty of deploying the supporting software on a large number of workers. Since networks cannot be assumed to be reliable, large desktop grids are designed for independent task applications with relatively long-running individual tasks.

By contrast, research on traditional grid scheduling has focused on algorithms to determine an optimal computation schedule based on the assumption that sufficiently detailed and up to date knowledge of the system state is available to a single entity (the metascheduler) [14], [15], [16], [17]. While this approach results in a very efficient utilization of the resources, it does not scale to large numbers of machines. Maintaining a global view of the system becomes prohibitively expensive and unreliable networks might even make it impossible.

To summarize, the existing approaches to harnessing machine resources represent different design strategies, as shown in Table I. Traditional grid approaches or Condor decide to limit the size of the system and assume a fairly reliable network in exchange for being able to run arbitrary tasks, such as MPI tasks. Desktop grid approaches restrict the type of the application to independent (or nearly independent) tasks of fairly large task granularity in exchange for being able to run on very large numbers of machines with potentially unreliable network connections. The best of both worlds, arbitrary tasks and large numbers of machines, is not possible because the central task scheduler would become a bottleneck.

Manuscript received June 15, 2004; revised October 14, 2004. This work was partially supported by the Ohio Supercomputer Center grants PAS0036-1 and PAS0121-1.

A.J. Chakravarti is with The MathWorks, Inc., Natick, MA 01760, USA. This work was done while pursuing a Ph.D. at The Ohio State University. Email: Arjav.Chakravarti@mathworks.com

G. Baumgartner is with the Department of Computer Science, Louisiana State University, Baton Rouge, LA 70803, USA. Email: gb@csc.lsu.edu

M. Lauria is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA. Email: lauria@cse.ohio-state.edu

	Large desktop grids (e.g., BOINC)	Small desktop grids (Condor)	Traditional grids (e.g., Globus)	Organic Grid
Network	large, unreliable	small, reliable	small, reliable	large, unreliable
Task granularity	large	medium to large	medium to large	medium to large
Task model	independent task	any	any	any
Task scheduling	centralized	centralized	centralized	decentralized

TABLE I

CLASSIFICATION OF APPROACHES TO LARGE-SCALE COMPUTATION

We present a new approach to grid computing, called the Organic Grid, that does not have the restrictions of either of the existing approaches. By using a decentralized, adaptive scheduling scheme, we attempt to allow arbitrary tasks to be run on large numbers of machines or in conditions with unreliable networks. Our approach can be used to broaden the class of applications that can be run on a large desktop grid, or to extend a traditional grid computing approach to machines with unreliable connections. The tradeoff of our approach is that the distributed scheduling scheme may not result in as good resource usage as with a centralized scheduler.

The Organic Grid project is an effort to redesign from scratch the infrastructure for distributed computation on desktop grids. Our middleware represents a radical departure from current grid or Peer-to-Peer concepts, and does not rely on existing grid technology. In designing our Organic Grid infrastructure we have tried to address the following questions:

- What is the best model of utilization of a system based on the harvesting of idle cycles of hundreds-of-thousands to millions of PCs?
- How should the system be designed in order to make it consistent with the grid computing ideals of computation as a ubiquitous and easily accessible utility?

Nature provides numerous examples of complex systems comprising millions of organisms that organize themselves in an autonomous, adaptive way to produce complex patterns. In these systems, the emergence of complex patterns derives from the superposition of a large number of interactions between organisms that have relatively simple behavior. In order to apply this approach to the task of organizing computation over complex systems such as desktop grids, one would have to devise a way of breaking a large computation into small autonomous chunks, and then endowing each chunk with the appropriate behavior.

Our approach is to encapsulate computation and behavior into mobile agents. A similar concept was first explored by Montresor et al. [18] in a project showing how an ant algorithm could be used to solve the problem of dispersing tasks uniformly over a network. In our approach, the behavior is designed to produce desirable patterns of execution according to current grid engineering principles. More specifically, the pattern of computation resulting from the synthetic behavior of our agents reflects some general concepts about autonomous grid scheduling protocols studied by Kreaseck et al. [19]. Our approach extends previous results by showing i) how the basic concepts can be extended to accommodate highly dynamic systems, and ii) a practical implementation of these concepts.

One consequence of the encapsulation of behavior and computation into agents is that they can be easily customized for different classes of applications. Another desirable consequence is that the underlying support infrastructure for our system is extremely simple. Therefore, our approach naturally lends itself to a true peer-to-peer implementation, where each node can be at the same time provider and user of the computing utility infrastructure. Our scheme can be easily adapted to the case where the source of computation (the node initiating a computing job) is different from the source of the data.

The main contributions of this paper are: i) the description of a new organization principle for desktop grids which combines biologically inspired models of organization, autonomous scheduling, and strongly mobile agents, ii) the demonstration of these principles as a working proof-of-concept prototype, iii) a detailed exploration of the design space of our system, and iv) the performance evaluation of our design using metrics appropriate for assessing self-organizing desktop grids.

The purpose of this work is the initial exploration of a novel concept, and as such it is not intended to give a quantitative assessment of all aspects and implications of our new approach. In particular, detailed evaluations of scalability, degree of tolerance to faults, adaptivity to rapidly changing systems, or security issues have been left for future studies.

II. BACKGROUND AND RELATED WORK

A. Peer-to-Peer and Internet Computing

The goal of utilizing the CPU cycles of idle machines was first realized by the Worm project [20] at Xerox PARC. Further progress was made by academic projects such as Condor [8]. The growth of the Internet made large-scale efforts like GIMPS [7], SETI@home [3] and folding@home [4] feasible. Recently, commercial solutions such as Entropia [5] and United Devices [21] have also been developed.

The idea of combining Internet and peer-to-peer computing is attractive because of the potential for almost unlimited computational power, low cost, ease and universality of access — the dream of a true computational grid. Among the technical challenges posed by such an architecture, scheduling is one of the most formidable — how to organize computation on a highly dynamic system at a planetary scale while relying on a negligible amount of knowledge about its state.

B. Scheduling

Decentralized scheduling is a field that has recently attracted considerable attention. Two-level scheduling schemes have been considered [22], [23], but these are not scalable enough for the Internet. In the scheduling heuristic described by Leangsuksun et al. [24], every machine attempts to map tasks on to itself as well as its K best neighbors. This appears to require that each machine have an estimate of the execution time of subtasks on each of its neighbors, as well as of the bandwidth of the links to these other machines. It is not clear that this information will be available in large-scale and dynamic environments.

G-Commerce was a study of dynamic resource allocation on a grid in terms of computational market economies in which applications must buy resources at a market price influenced by demand [25]. While conceptually decentralized, if implemented this scheme would require the equivalent of centralized commodity markets (or banks, auction houses, etc.) where offer and demand meet, and commodity prices can be determined.

Recently, a new autonomous and decentralized approach to scheduling has been proposed to address specifically the needs of large grid and peer-to-peer platforms. In this bandwidth-centric protocol, the computation is organized around a tree-structured overlay network with the origin of the tasks at the root [19]. Each node in the system sends tasks to and receives results from its K best neighbors, according to bandwidth constraints. One shortcoming of this scheme is that the structure of the tree, and consequently the performance of the system, depends completely on the initial structure of the overlay network. This lack of dynamism is bound to affect the performance of the scheme and might also limit the number of machines that can participate in a computation.

C. Self-Organization of Complex Systems

The organization of many complex biological and social systems has been explained in terms of the aggregations of a large number of autonomous entities that behave according to simple rules. According to this theory, complicated patterns can emerge from the interplay of many agents — despite the simplicity of the rules [26], [27]. The existence of this mechanism, often referred to as *emergence*, has been proposed to explain patterns such as shell motifs, animal coats, neural structures, and social behavior. In particular, certain complex behaviors of social insects such as ants and bees have been studied in detail, and their applications to the solution of specific computer science problems has been proposed [18], [28]. In a departure from the methodological approach followed in previous projects, we did not try to accurately reproduce a naturally occurring behavior. Rather, we started with a problem and then designed a completely artificial behavior that would result in a satisfactory solution to it. Our work was inspired by a particular version of the emergence principle called Local Activation, Long-range Inhibition (LALI), which was recently shown to be responsible for the formation of a complex pattern using a clever experiment on ants [29].

D. Strongly Mobile Agents

To make progress in the presence of frequent reclamations of desktop machines, current systems rely on different forms of checkpointing: automatic, e.g., SETI@home, or voluntary, e.g., Legion. The storage and computational overheads of checkpointing put constraints on the design of a system. To avoid this drawback, desktop grids need to support the asynchronous and transparent migration of processes across machine boundaries.

Mobile agents [30] have relocation autonomy. These agents offer a flexible means of distributing data and code around a network, of dynamically moving between hosts as resource availability varies, and of carrying multiple threads of execution to simultaneously perform computation, decentralized scheduling, and communication with other agents. There have been some previous attempts to use mobile agents for grid computing or distributed computing [31], [32], [33], [34].

The majority of the mobile agent systems that have been developed until now are Java-based. However, the execution model of the Java Virtual Machine does not permit an agent to access its execution state, which is why Java-based mobility libraries can only provide *weak mobility* [35]. Weak mobility forces programmers to use a difficult programming style.

By contrast, agent systems with *strong mobility* provide the abstraction that the execution of the agent is uninterrupted, even as its location changes. Applications where agents migrate from host to host while communicating with one another, are severely restricted by the absence of strong mobility. Strong mobility also allows programmers to use a far more natural programming style.

The ability of a system to support the migration of an agent at any time by an external thread, is termed *forced mobility*. This is essential in desktop grid systems, because owners need to be able to reclaim their resources. Forced mobility is difficult to implement without strong mobility.

We provide strong and forced mobility for the full Java programming language by using a preprocessor that translates an extension of Java with strong mobility into weakly mobile Java code that explicitly maintains the execution state for all threads as a mobile data structure [36], [37]. For the target weakly mobile code we currently use IBM's Aglets framework [38]. The generated weakly mobile code maintains a movable execution state for each thread at all times.

III. Autonomic Scheduling

A. Overview

One of the works that inspired our project was the bandwidth-centric protocol proposed by Kreaseck et al. [19], in which a grid computation is organized around a tree-structured overlay network with the origin of the tasks at the root. A tree overlay network represents a natural and intuitive way of distributing tasks and collecting results. The drawback of the original scheme is that the performance and the degree of utilization of the system depend entirely on the initial assignment of the overlay network.

In contrast, we have developed our systems to be adaptive in the absence of any knowledge about machine configurations, connection bandwidths, network topology, and assuming only a minimal amount of initial information. While our scheme is also based on a tree, our overlay network keeps changing to adapt to system conditions. Our tree adaptation mechanism is driven by the perceived performance of a node's children, measured passively as part of the ongoing computation [39]. From the point of view of network topology, our system starts with a small amount of knowledge in the form of a "friends list", and then keeps building its own overlay network on the fly. Information from each node's "friends list" is shared with other nodes so the initial configuration of the lists is not critical. The only assumption we rely upon is that a "friends list" is available initially on each node to prime the system; solutions for the construction of such lists have been developed in the context of peer-to-peer file-sharing [40], [41] and will not be addressed in this paper.

The Local Activation, Long-range Inhibition (LALI) rule is based on two types of interactions: a positive, reinforcing one that works over a short range, and a negative, destructive one that works over longer distances. We retain the LALI principle but in a different form: we use a definition of distance which is based on a performance-based metric. In our experiment, distance is based on the perceived throughput which is some function of communication bandwidth and computational throughput. Nodes are initially recruited using the "friends list" in a way that is completely oblivious of distance, therefore propagating computation on distant nodes with same probability as close ones. During the course of the computation agents behavior encourages the propagation of computation among well-connected nodes while discouraging the inclusion of distant (i.e. less responsive) agents.

The methodology we followed to design the agent behavior is as follows. We selected a tree-structured overlay network as the desirable pattern of execution. We then empirically determined the simplest behavior that would organize the communication and task distribution among mobile agents according to that pattern. We then augmented the basic behavior in a way that introduced other desirable properties. With the total computation time as the performance metric, every addition to the basic scheme was separately evaluated and its contribution to total performance, quantitatively assessed.

One such property is the continuous monitoring of the performance of the child nodes. We assumed that no knowledge is initially available on the system, instead passive feedback from child nodes is used to measure their effective performance, e.g., the product of computational speed and communication bandwidth.

Another property is continuous, on-the-fly adaptation using the restructuring algorithm presented in Section III-D. Basically, the overlay tree is incrementally restructured while the computation is in progress by pushing fast nodes up towards the root of the tree. Other functions that were found to be critical for performance were the automatic determination of parameters such as prefetching and task size, the detection of cycles, the detection of dead nodes and the end of the computation.

In this paper we focus on the solution to one particular problem: the scheduling of the independent, identical subtasks of an independent-task application (ITA) whose data initially resides at one location. The size of individual subtasks and of their results is large, and so transfer times cannot be neglected. The application that we have used for our experiments is NCBI's nucleotide-nucleotide sequence comparison tool BLAST [42].

```
receive request for s subtasks from node c
// c may be the node itself
if subtask_list.size>=s
  c.send_subtasks(s)
else
  c.send_subtasks(subtask_list.size)
  outstanding_subtask_queue.
    add(c,s--subtask_list.size)
  parent.
    send_request(outstanding_subtask_queue.
                 total_subtasks)
```

Fig. 1. Behavior of Node on Receiving Request

Our choice of using an ITA for our proof-of-concept implementation follows a common practice in grid scheduling research. However our scheme is general enough to accommodate other classes of applications. In a recent article we have demonstrated using a fault-tolerant implementation of Cannon's matrix multiplication algorithm that our scheduling scheme can be adapted to applications with communicating tasks [43], [44].

B. Basic Agent Design

A large computational task is encapsulated in a strongly mobile agent. This task should be divisible into a number of independent subtasks. A user starts the computation agent on his/her machine. One thread of the agent begins executing subtasks sequentially. The agent is also prepared to receive requests for work from other machines. If the machine has any uncomputed subtasks, and receives a request for work from another machine, it sends a clone of itself to the requesting machine. The requester is now this machine's *child*.

The clone asks its parent for a certain number of subtasks to work on, s. A thread begins to compute the subtasks. Other threads are created — when required — to communicate with the parent or other machines. When work requests are received, the agent dispatches its own clone to the requester. The computation spreads in this manner. The topology of the resulting overlay network is a tree with the originating machine at the root node.

An agent requests its parent for more work when it has executed its own subtasks. Even if the parent does not have the requested number of subtasks, it will respond and send its child what it can. The parent keeps a record of the number of subtasks that remain to be sent, and sends a request to its own parent. Every time a node of the tree obtains r results, either computed by itself or obtained from a child, it sends them to its parent. This message includes a request for all pending subtasks. This can be seen in Figures 1 and 2.

C. Maintenance of Child-lists

Each node has up to c active children, and up to p potential children. Ideally, $c + p$ is chosen so as to strike a balance between a tree that is too deep (long delays in data propagation) and one that is too wide (inefficient distribution of data).

```
receive t subtasks from parent
subtask_list.add(t)
if outstanding_subtask_queue.
     total_subtasks>=t
  <send t subtasks to nodes in
   outstanding_subtask_queue>
else
  <send outstanding_subtask_queue.
   total_subtasks subtasks to nodes in
   outstanding_subtask_queue>
// may include subtasks for node itself
```

Fig. 2. Behavior of Node on Receiving Subtasks

```
receive feedback from node c
if child_list.contains(c)
  child_list.update_rank(c)
else
  child_list.add(c)
  if child_list.size>MAX_CHILD_LIST_SIZE
    sc:=child_list.slowest
    child_list.remove(sc)
    old_child_list.add(sc)
    inverted_child_list:=inv(child_list)
    sc.send_ancestor_list(inverted_child_list)
```

Fig. 3. Behavior of Parent Node on Receiving Feedback

The active children are ranked on the basis of their performance. The performance metric is application-dependent. For an ITA, a child is evaluated on the basis of the rate at which it sends in results. When a child sends r results, the node measures the time-interval since the last time it sent r results. The final result-rate of this child is calculated as an average of the last R such time-intervals. This ranking is a reflection of the performance of not just a child, but of the entire subtree with the child node at its root.

Potential children are the ones which the current node has not yet been able to evaluate. A potential child is added to the active child-list once it has sent enough results to the current node. If the node now has more than c children, the slowest child, sc, is removed from the child-list. As described below, sc is then given a list of other nodes, which it can contact to try and get back into the tree. The current node keeps a record of the last o former children, and sc is now placed in this list. Nodes are purged from this list once a sufficient, user-defined time period elapses. During that interval of time, messages from sc will be ignored. This avoids thrashing and excessive dynamism in the tree. The pseudo-code for the maintenance of child-lists has been presented in Figure 3.

D. Restructuring of the Overlay Network

The topology of the overlay network is a tree, and it is desirable for the best-performing nodes to be close to the root. In the case of an ITA, both computational speed and link bandwidth contribute to a node's effective performance. Having well connected nodes close to the top enhances the extraction of subtasks from the root and minimizes the communication delay between the root and the best nodes. Therefore the overlay network is constantly being restructured so that the nodes with the highest throughput migrate toward the root, pushing those with low throughput towards the leaves.

```
receive node b from node c
if old_child_list.not_contains(b)
  potential_child_list.add(b)
  c.send_accept_child(b)
else
  c.send_reject_child(b)
```

Fig. 4. Behavior of Parent Node on Receiving Propagated Child

```
receive accept_child(b) from parent
// a request was earlier made to parent
// about node b
b.send_ancestor_list(ancestor_list)
// b will now contact parent directly
```

Fig. 5. Behavior of Child Node on Receiving Positive Response

A node periodically informs its parent about its best-performing child. The parent then checks whether its grandchild is present in its list of former children. If not, it adds the grandchild to its list of potential children and tells this node that it is willing to consider the grandchild. The node then instructs its child to contact its grandparent directly. If the contact ends in a promotion, the entire subtree with the child node at its root will move one level higher in the tree. This constant restructuring results in fast nodes percolating towards the root of the tree and has been detailed in Figures 4 and 5. The checking of a promising child against a list of former children prevents the occurrence of trashing due to consecutive promotions and demotions of the same node.

When a node updates its child-list and decides to remove its slowest child, sc, it does not simply discard the child. It prepares a list of its children in descending order of performance, i.e., slowest node first. The list is sent to sc, which attempts to contact those nodes in turn. Since the first nodes that are contacted are the slower ones, the tree is sought to be kept balanced. The actions of a node on receipt of a new list of ancestors are in Figure 6.

E. Size of Result Burst

Each agent of an ITA ranks its children on the basis of the time taken to send some results to this node. The time required to obtain just one result-burst, or a result-burst of size 1, might not be a good measure of the performance of a child. Nodes might make poor decisions about which children to keep and discard. The child propagation algorithm benefits from using the average of R result-burst intervals and from setting r, the result-burst burst size, to be greater than 1. A better measure for the performance of a child is the time taken by a node to obtain $r*(R+1)$ results. However, r and R should not be set to very large values because the overlay network would take too much time to take form and to get updated.

F. Fault Tolerance

If the parent of a node were to become inaccessible due to machine or link failures, the node and its own descendants would be disconnected from the tree. The application might require that a node remain in the tree at all times. In this scenario, the node must be able to contact its parent's ancestors. Every node keeps a (constant size) list of a of its

```
receive message from parent
ancestor_list := message.ancestor_list
if parent != ancestor_list.last
  parent:=ancestor_list.last
```

Fig. 6. Behavior of Node on Receiving new Ancestor-List

```
while true
  send message to parent
  if <unable to contact parent>
    ancestor_list.remove(parent)
    if ancestor_list.size = 0
      <find-new-parent or self-destruct>
    parent := ancestor_list.last
```

Fig. 7. Fault Tolerance — Contacting Ancestors

ancestors. This list is updated every time its parent sends it a message. The updates to the ancestor-list take into account the possibility of the topology of the overlay network changing frequently.

A child sends a message to its parent — the a-th node in its ancestor-list. If it is unable to contact the parent, it sends a message to the $(a-1)$-th node in that list. This goes on until an ancestor responds to this node's request. The ancestor becomes the parent of the current node and normal operation resumes.

If a node's ancestor-list goes down to size 0, it attempts to obtain the address of some other agent by checking its data distribution and communication overlays. If these are the same as the scheduling tree, the node has no means of obtaining any more work to do. The mobile agent informs the agent environment that no useful work is being done by this machine, before self-destructing. The environment begins to send out requests for work to a list of friends. The pseudo-code for the fault tolerance algorithm is in Figure 7.

In order to recover from the loss of tasks by failing nodes, every node keeps track of unfinished subtasks that were sent to children. If a child requests additional work and no new task can be obtained from the parent, unfinished tasks are handed out again.

G. Cycles in the Overlay Network

Even though the scheduling overlay network should be a tree, failures could cause the formation of a cycle of nodes. The system recovers from this situation by having each node examine its ancestor list on receiving it from its parent. If a node finds itself in that list, it knows that a cycle has occurred. The node attempts to break the cycle by obtaining the address of some other agent on its data distribution or communication overlays. However, if these are identical to the scheduling overlay, the node will be starved of work. If the agent is starved of work for more than a specified time, it self-destructs.

H. Termination

The root of the tree is informed when the computational task has been completed. It sends a termination message to each of its actual, potential and former children. The computation agent on the root then self-destructs. The children of the root do the same. Termination messages spread down to the leaves and the computation terminates. There are two scenarios in which termination could be incomplete:

- A termination message might not reach a node. The situation is the same as that described in Subsection III-F.
- Consider that computation agents are executing on nodes $n1$ and $n2$. $n1$ receives a termination message, but $n2$ does not because of a failure. The agent on $n1$ destroys itself. $n1$ now sends request messages to its friends. If one of these is $n2$, a clone of $n2$'s agent is sent to $n1$. An unchecked spread of computation will not occur because agents send out clones only if they do not have any uncomputed subtasks. $n1$ and $n2$ will eventually run out of subtasks and destroy themselves as explained in Subsection III-F.

I. Self-adjustment of Task List Size

A node always requests a certain number of subtasks and obtains their results before requesting more subtasks to work on. The size of a subtask is simply an estimation of the smallest unit of work that every machine on the peer-to-peer network should be able to compute in a time that the user considers reasonable; scheduling should not be inordinately slow on account of subtasks that take a long time to compute. However, in an ITA-type application, the utilization of a high-performance machine may be poor because it is only requesting a fixed number of subtasks at a time.

A node may request more subtasks in order to increase the utilization of its resources and to improve the system computation-to-data ratio. A node requests a certain number of subtasks, t, that it will compute itself. Once it has finished computing the t subtasks, it compares the average time to compute a subtask on this run to that of the previous run. Depending on whether it performed better, worse or about the same, the node requests $i(t)$, $d(t)$ or t subtasks for its next run, where $i(t) > t$ and $d(t) < t$.

J. Prefetching

A potential cause of slowdown in the basic scheduling scheme described earlier, is the delay at each node due to its waiting for new subtasks. This is because it needs to wait while its requests propagate up the tree to the root and subtasks propagate down the tree to the node.

It might be beneficial to use prefetching to reduce the time that a node waits for subtasks. A node determines that it should request t subtasks from its parent. It also makes an optimistic prediction of how many subtasks it might require in future by using the i function that is used for self-adjustment. $i(t)$ subtasks are then requested from the parent. When a node finishes computing one set of subtasks, more subtasks are readily available for it to work on, even as a request is submitted to the parent. This interleaving of computation and communication reduces the time for which a node is idle.

While prefetching will reduce the delay in obtaining new subtasks to work on, it also increases the amount of data that needs to be transferred at a time from the root to the current node, thus increasing the synchronization delay and data

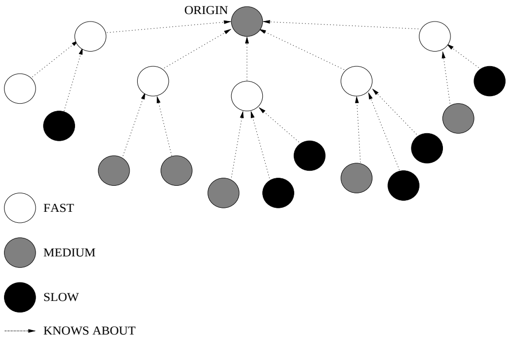


Fig. 8. Good Configuration with A Priori Knowledge

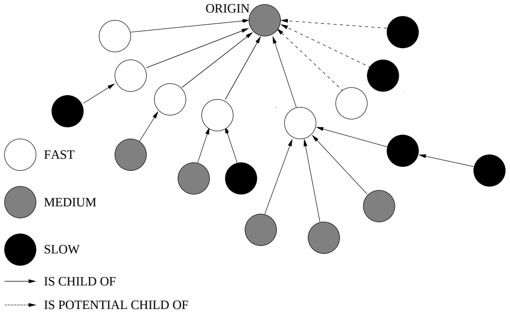


Fig. 9. Final Node Organization, Result-burst size=3, Good Initial Configuration

Parameter Name	Parameter Value
Maximum children	5
Maximum potential children	5
Result-burst size	3
Self-adjustment	linear
Number of subtasks initially requested	1
Child-propagation	On

TABLE II
ORIGINAL PARAMETERS

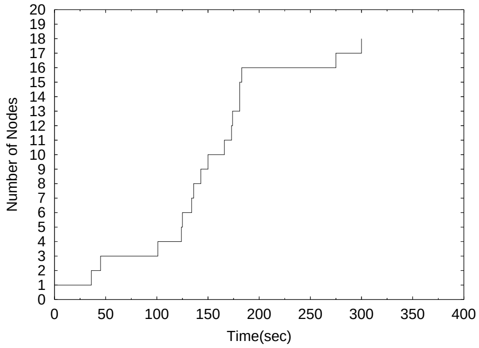


Fig. 10. Code Ramp-up

transfer time. This is why excessively aggressive prefetching will result in a performance degradation.

IV. Measurements

We have conducted experiments to evaluate the performance of each aspect of our scheduling scheme. The experiments were performed on a cluster of eighteen heterogeneous machines at different locations around Ohio. The machines ran the Aglets weak mobility agent environment on top of either Linux or Solaris.

The application we used to test our system was the gene sequence similarity search tool, NCBI's nucleotide-nucleotide BLAST [42] — a representative independent-task application. The mobile agents started up a BLAST executable to perform the actual computation. The task was to match a 256KB sequence against 320 data chunks, each of size 512KB. Each subtask was to match the sequence against one chunk. Chunks flow down the overlay tree whereas results flow up to the root. An agent cannot migrate during the execution of the BLAST code; since our experiments do not require strong mobility, this limitation is irrelevant to our measurements.

All eighteen machines would have offered good performance as they all had fast connections to the Internet, high processor speeds and large memories. In order to obtain more heterogeneity in their performance, we introduced delays in the application code so that we could simulate the effect of slower machines and slower network connections. We divided the machines into fast, medium and slow categories by introducing delays in the application code.

As shown in Figure 11, the nodes were initially organized randomly. The dotted arrows indicate the directions in which request messages for work were sent to friends. The only thing a machine knew about a friend was its URL. We ran the computation with the parameters described in Table II. Linear self-adjustment means that the increasing and decreasing functions of the number of subtasks requested at each node are linear. The time required for the code and the first subtask to arrive at the different nodes can be seen in Figure 10. This is the same for all the experiments.

A. Comparison with Knowledge-based Scheme

The purpose of these tests is to evaluate the quality of the configuration which is autonomously determined by our scheme for different initial conditions.

Two experiments were conducted using the parameters in Table II. In the first, we manually created a good initial configuration assuming a priori knowledge of system parameters. We then ran the application, and verified that the final configuration did not substantially depart from the initial one. We consider a good configuration to be one in which fast nodes are nearer the root. Figures 8 and 9 represent the start and end of this experiment. The final tree configuration shows that fast nodes are kept near the root and that the system is constantly re-evaluating every node for possible relocation (as shown by the three rightmost children which are under evaluation by the root).

We began the second experiment with the completely random configuration shown in Figure 11. The resulting configuration shown in Figure 12 is substantially similar to the good configurations of the previous experiment; if the

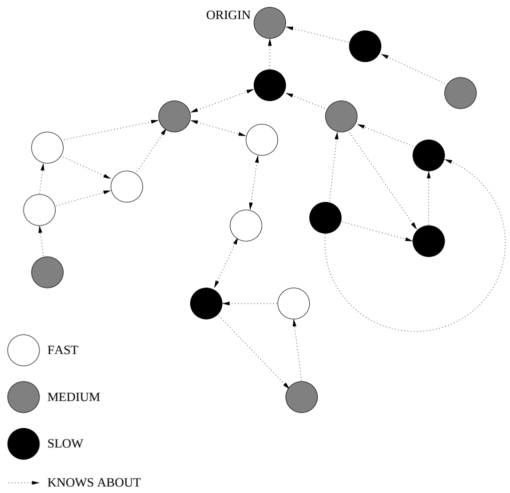


Fig. 11. Random Configuration of Machines

Configuration	Running Time (sec)
original	2294
good	1781

TABLE III
EFFECT OF PRIOR KNOWLEDGE

execution time had been longer, the migration towards the root of the two fast nodes at depths 2 and 3 would have been complete.

B. Effect of Child Propagation

We performed our computation with the child-propagation aspect of the scheduling scheme disabled. Comparisons of the running times and topologies are in Table IV and Figures 12 and 13. The child-propagation mechanism results in a 32% improvement in the running time. The reason for this improvement is the difference in the topologies. With child-propagation turned on, the best-performing nodes are closer to the root. Subtasks and results travel to and from these nodes at a faster rate, thus improving system throughput and preventing the root from becoming a bottleneck. This mechanism is the most effective aspect of our scheduling scheme.

C. Result-burst size

The experimental setup in Table II was again used. We then ran the experiment with different result-burst sizes. The running times have been tabulated in Table V. The child evaluations that are made by nodes on the basis of one result are poor. The nodes' child-lists change frequently and are far from ideal, as in Figure 14.

Scheme	Running Time (sec)
With	2294
Without	3035

TABLE IV
EFFECT OF CHILD PROPAGATION

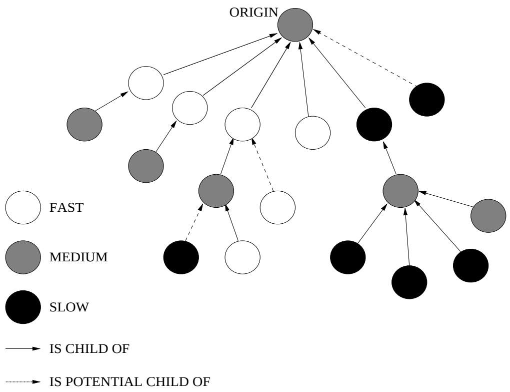


Fig. 12. Final Node Organization, Result-burst size=3, With Child Propagation

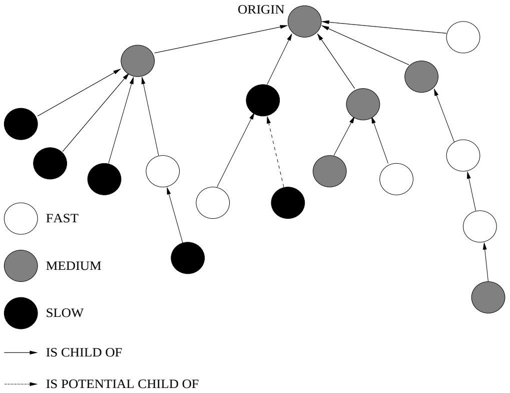


Fig. 13. Final Node Organization, Result-burst size=3, No Child Propagation

There is a qualitative improvement in the child-lists as the result-burst size increases. The structure of the resulting overlay networks for result-burst sizes 3 and 5 are in Figures 12 and 15. However, with very large result-bursts, it takes longer for the tree overlay to form and adapt, thus slowing down the experiment. This can be seen in Figure 16.

D. Prefetching and Initial Task Size

The data ramp-up time is the time required for subtasks to reach every single node. Prefetching has a positive effect on this. The minimum number of subtasks that each node requests also affects the data ramp-up. The greater this number, the greater the amount of data that needs to be sent to each node, and the slower the data ramp-up. This can be seen in Table VI and Figures 17, 18, 19, 20 and 21.

Prefetching does improves the ramp-up, but of paramount importance is its effect on the overall running time of an experiment. This is also closely related to the minimum number of subtasks requested by each node. Prefetching improves system throughput when the minimum number of subtasks requested is one. As the minimum number of subtasks requested by a node increases, more data needs to be transferred at a time

No. of Subtasks	Ramp-up Time (sec)	Ramp-up Time (sec)	Running Time (sec)	Running Time (sec)
	Prefetching	No prefetching	Prefetching	No prefetching
1	406	590	2308	2520
2	825	979	2302	2190
5	939	1575	2584	2197

TABLE VI
EFFECT OF PREFETCHING AND MINIMUM NUMBER OF SUBTASKS

Result-burst Size	Running Time (sec)
1	3050
3	2294
5	2320
8	3020

TABLE V
EFFECT OF RESULT-BURST SIZE

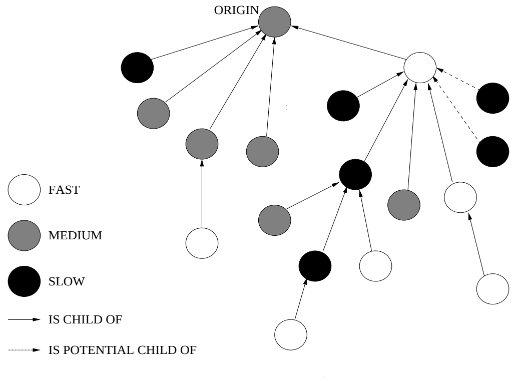


Fig. 14. Node Organization, Result-burst size=1

from the root to this node, and the effect of prefetching becomes negligible. As this number increases further, prefetching actually causes a degradation in throughput. Table VI and Figure 22 summarize these results.

E. Self-Adjustment

We ran an experiment using the configuration in Table II and then did the same using constant and exponential self-adjustment functions instead of the linear one. The data ramp-ups have been compared in Table VII and Figure 23. The ramp-up with exponential self-adjustment is appreciably faster than that with linear or constant self-adjustment. The aggressive approach performs better because nodes prefetch a larger amount of subtasks, and subtasks quickly reach the nodes farthest from the root.

We also compared the running times of the three runs which are in Table VII. Interestingly, the run with the exponential self-adjustment performed poorly with respect to the other runs. This is due to nodes prefetching extremely large numbers of subtasks. Nodes now spend more time waiting for their requests to be satisfied, resulting in a degradation in the throughput at that node.

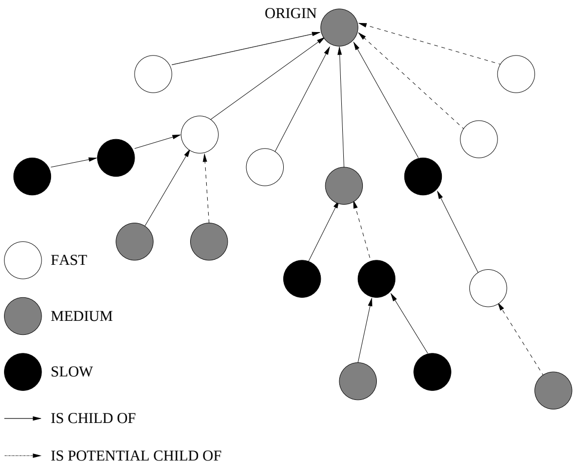


Fig. 15. Node Organization, Result-burst size=5

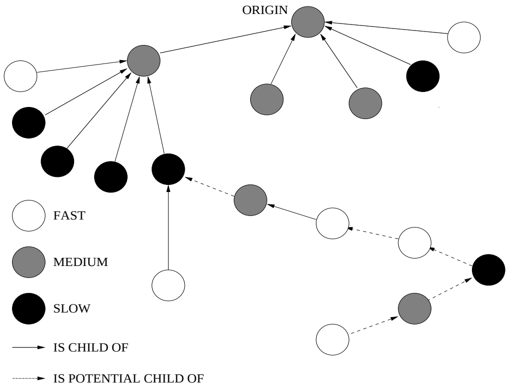


Fig. 16. Node Organization, Result-burst size=8

The linear case was expected to perform better than the constant one, but the observed difference was insignificant. We expect this difference to be more pronounced with longer experimental runs and a larger number of subtasks.

F. Number of children

We experimented with different child-list sizes and found that the data ramp-up time with the maximum number of children set to 5 was less than that with the maximum number of children set to 10 or 20. These results are in Table VIII. The root is able to take on more children in the latter cases and the spread of subtasks to nodes that were originally far from the root takes less time.

Instead of exhibiting better performance, the runs where

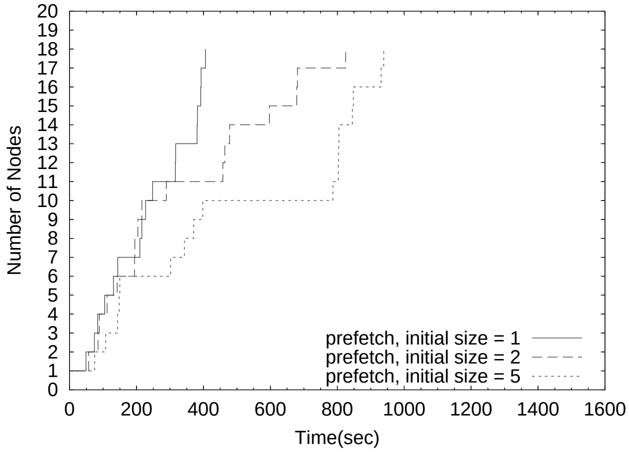


Fig. 17. Effect of Minimum Number of Subtasks on Data Ramp-up with Prefetching

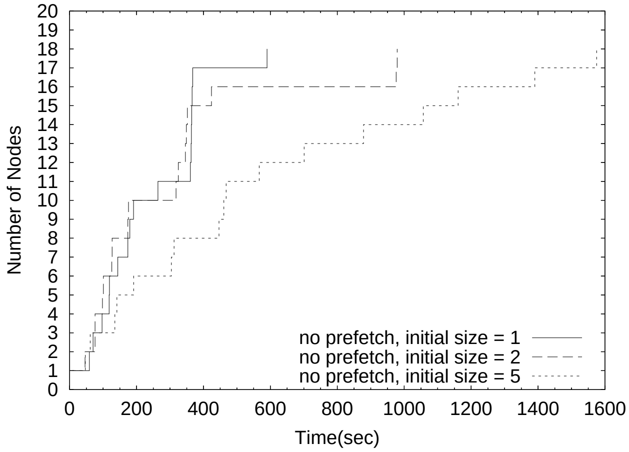


Fig. 18. Effect of Minimum Number of Subtasks on Data Ramp-up without Prefetching

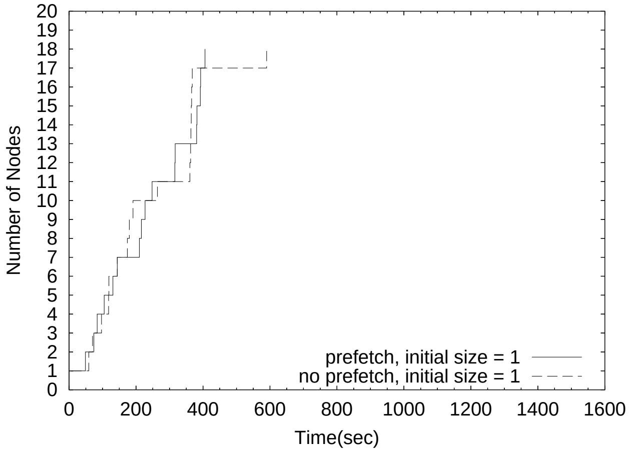


Fig. 19. Effect of Prefetching on Data Ramp-up with Minimum Number of Subtasks = 1

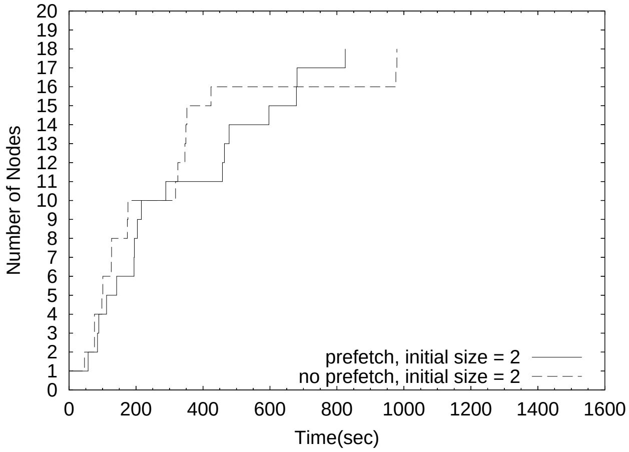


Fig. 20. Effect of Prefetching on Data Ramp-up with Minimum Number of Subtasks = 2

large numbers of children were allowed, had approximately the same total running time as the run with the maximum number of children set to 5. This is because children have to wait for a longer time for their requests to be satisfied.

In order to obtain a better idea of the effect of several children waiting for their requests to be satisfied, we ran two experiments: one with the good initial configuration of Figure 8, and the other using a star topology — every non-root node was adjacent to the root at the beginning of the experiment itself. The maximum sizes of the child-lists were set to 5 and 20, respectively. Since the overlay networks were already organized such that there would be little change in their topology as the computation progressed, there was minimal impact of these changes on the overall running time. The effect of the size of the child-list was then clearly observed as in Table IX. Similar results were observed even when the child-propagation mechanisms were turned off.

V. Conclusions and Future Work

We have designed an autonomic scheduling algorithm in which multi-threaded agents with strong mobility form a tree-structured overlay network. The structure of this tree is varied dynamically such that the nodes that currently exhibit good performance are brought closer to the root, thus improving the performance of the system.

We have described experiments with scheduling a massively parallel application whose data initially resides at one location and whose subtasks have considerable data transfer times. The experiments were conducted on a set of machines distributed across Ohio. While this paper concentrated on a scheduling scheme for independent-task applications, we are experimenting with adapting the algorithm for a wide class of applications. Recent results show that our approach can be adapted to communicating applications, such as Cannon-style matrix multiplication [43], [44].

It is our intention to present a desktop grid application developer with a simple application programming interface that will allow him/her to customize the scheduling scheme to the characteristics of an application. A prototype of this has already been implemented.

An important problem that we will address in future is the initial assignment of the friend-list. There has been some research on the problem of assigning friend-lists [40], [41], and we will consider how best to apply this to our own work.

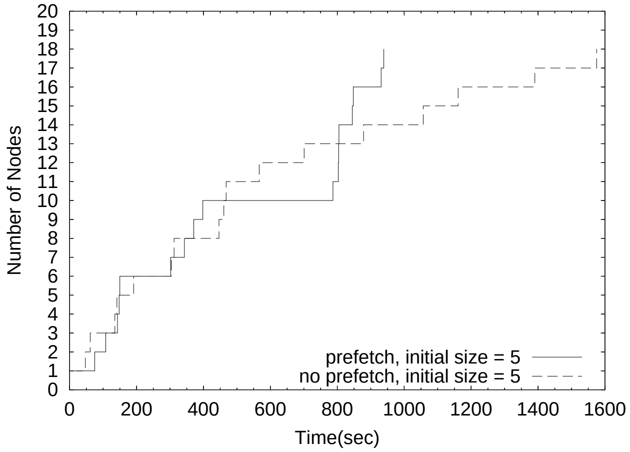


Fig. 21. Effect of Prefetching on Data Ramp-up with Minimum Number of Subtasks = 5

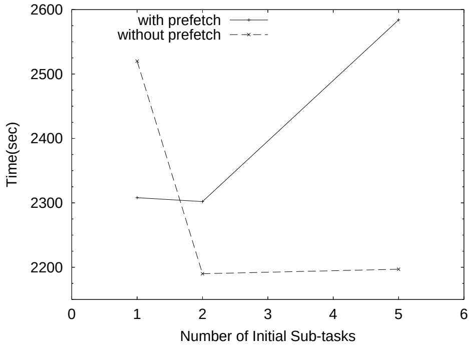


Fig. 22. Effect of Prefetching and Min. No. of Subtasks

The experimental platform was a set of 18 heterogeneous machines. In future, we plan to harness the computing power of idle machines across the Internet by running a mobile agent platform inside a screen saver in order to create a desktop grid of a scale of the tens or hundreds of thousands. Researchers will then be free to deploy scientific applications on this system.

Self-adjustment Function	Ramp-up Time (sec)	Running Time (sec)
Linear	1068	2302
Constant	1142	2308
Exponential	681	2584

TABLE VII
EFFECT OF SELF-ADJUSTMENT FUNCTION

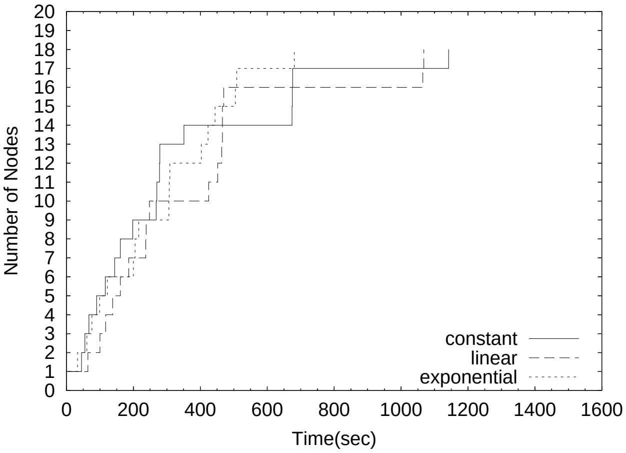


Fig. 23. Effect of Self-adjustment Function on Data Ramp-up Time

REFERENCES

[1] I. Foster, C. Kesselman, and S. Tuecke, “The anatomy of the Grid: Enabling scalable virtual organizations,” *International Journal of High Performance Computing Applications*, vol. 15, no. 3, 2001.

[2] I. Foster, C. Kesselman, J.Nick, and S. Tuecke, “The physiology of the Grid: An open Grid services architecture for distributed systems integration,” 2002, http://www.globus.org/research/papers.html. [Online]. Available: http://www.globus.org/research/papers.html

[3] SETI@home. [Online]. Available: http://setiathome.ssl.berkeley.edu

[4] folding@home. [Online]. Available: http://folding.stanford.edu

[5] A. A. Chien, B. Calder, S. Elbert, and K. Bhatia, “Entropia: architecture and performance of an enterprise desktop grid system,” *Journal of Parallel and Distributed Computing*, vol. 63, no. 5, pp. 597–610, 2003.

[6] B. O. I. for Network Computing (BOINC), http://boinc.berkeley.edu/. [Online]. Available: http://boinc.berkeley.edu/

[7] G. Woltman. [Online]. Available: http://www.mersenne.org/prime.htm

[8] M. Litzkow, M. Livny, and M. Mutka, “Condor — a hunter of idle workstations,” in *Proceedings of the 8th International Conference of Distributed Computing Systems*, June 1988.

[9] M. Maheswaran, S. Ali, H. J. Siegel, D. A. Hensgen, and R. F. Freund, “Dynamic matching and scheduling of a class of independent tasks onto heterogeneous computing systems,” in *Proceedings of the 8th Heterogeneous Computing Workshop*, Apr. 1999, pp. 30–44.

[10] E. Heymann, M. A. Senar, E. Luque, and M. Livny, “Adaptive scheduling for master-worker applications on the computational grid,” in *Proceedings of the First International Workshop on Grid Computing*, 2000.

[11] T. Kindberg, A. Sahiner, and Y. Paker, “Adaptive Parallelism under Equus,” in *Proceedings of the 2nd International Workshop on Configurable Distributed Systems*, Mar. 1994, pp. 172–184.

[12] D. Buaklee, G. Tracy, M. K. Vernon, and S. Wright, “Near-optimal adaptive control of a large grid application,” in *Proceedings of the International Conference on Supercomputing*, June 2002.

[13] N. T. Karonis, B. Toonen, and I. Foster, “MPICH-G2: A grid-enabled implementation of the message passing interface,” *Journal of Parallel and Distributed Computing*, vol. 63, no. 5, pp. 551–563, 2003.

[14] A. S. Grimshaw and W. A. Wulf, “The legion vision of a worldwide virtual computer,” *Communications of the ACM*, Jan. 1997.

[15] F. Berman, R. Wolski, H. Casanova, W. Cirne, H. Dail, M. Faerman, S. Figueira, J. Hayes, G. Obertelli, J. Schopf, G. Shao, S. Smallen, N. Spring, A. Su, and D. Zagorodnov, “Adaptive computing on the grid using AppLeS,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 14, no. 4, pp. 369–382, 2003.

[16] D. Abramson, J. Giddy, and L. Kotler, “High performance parametric modeling with Nimrod/G: Killer application for the global grid?” in *Proceedings of International Parallel and Distributed Processing Symposium*, May 2000, pp. 520–528.

[17] I. Taylor, M. Shields, and I. Wang, *Grid Resource Management*. Kluwer, June 2003, ch. 1 - Resource Management of Triana P2P Services.

[18] A. Montresor, H. Meling, and O. Babaoglu, “Messor: Load-balancing through a swarm of autonomous agents,” in *Proceedings of 1st Workshop on Agent and Peer-to-Peer Systems*, July 2002.

[19] B. Kreaseck, L. Carter, H. Casanova, and J. Ferrante, “Autonomous protocols for bandwidth-centric scheduling of independent-task applications,” in *Proceedings of the International Parallel and Distributed Processing Symposium*, Apr. 2003, pp. 23–25.

[20] J. F. Shoch and J. A. Hupp, “The ”worm” programs — early experience with a distributed computation,” *Communications of the ACM*, Mar. 1982.

[21] U. Devices. [Online]. Available: http://www.ud.com

[22] H. James, K. Hawick, and P. Coddington, “Scheduling independent tasks on metacomputing systems,” in *Proceedings of Parallel and Distributed Computing Systems*, Aug. 1999.

[23] J. Santoso, G. D. van Albada, B. A. A. Nazief, and P. M. A. Sloot, “Hierarchical job scheduling for clusters of workstations,” in *Proceedings*

Max. No. of Children	Time (sec)
5	1068
10	760
20	778

TABLE VIII
EFFECT OF NO. OF CHILDREN ON DATA RAMP-UP

Max. No. of Children	Time (sec)
5	1781
20	2041

TABLE IX
EFFECT OF NO. OF CHILDREN ON RUNNING TIME

of the 6th annual conference of the Advanced School for Computing and Imaging, June 2000, pp. 99–105.

[24] C. Leangsuksun, J. Potter, and S. Scott, "Dynamic task mapping algorithms for a distributed heterogeneous computing environment," in *Proceedings of the Heterogeneous Computing Workshop*, Apr. 1995, pp. 30–34.

[25] R. Wolski, J. Plank, J. Brevik, and T. Bryan, "Analyzing market-based resource allocation strategies for the computational grid," *International Journal of High-performance Computing Applications*, vol. 15, no. 3, 2001.

[26] A. Turing, "The chemical basis of morphogenesis," in *Philos. Trans. R. Soc. London*, no. 237 B, 1952, pp. 37–72.

[27] A. Gierer and H. Meinhardt, "A theory of biological pattern formation," in *Kybernetik*, no. 12, 1972, pp. 30–39.

[28] E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, Santa Fe Institute Studies in the Sciences of Complexity, 1999.

[29] G. Theraulaz, E. Bonabeau, S. C. Nicolis, R. V. Sol, V. Fourcassi, S. Blanco, R. Fournier, J.-L. Joly, P. Fernndez, A. Grimal, P. Dalle, and J.-L. Deneubourg, "Spatial patterns in ant colonies," in *PNAS*, vol. 99, no. 15, 2002, pp. 9645–9649.

[30] D. B. Lange and M. Oshima, "Seven good reasons for mobile agents," *Communications of the ACM*, Mar. 1999.

[31] J. Bradshaw, N. Suri, A. J. Cañas, R. Davis, K. M. Ford, R. R. Hoffman, R. Jeffers, and T. Reichherzer, "Terraforming cyberspace," in *Computer*. IEEE, July 2001, vol. 34(7).

[32] O. Rana and D. Walker, "The Agent Grid: Agent-based resource integration in PSEs," in *16th IMACS World Congress on Scientific Computation, Applied Mathematics and Simulation*, Lausanne, Switzerland, August 2000.

[33] B. Overeinder, N. Wijngaards, M. van Steen, and F. Brazier, "Multi-agent support for Internet-scale Grid management," in *AISB'02 Symposium on AI and Grid Computing*, O. Rana and M. Schroeder, Eds., April 2002, pp. 18–22.

[34] R. Ghanea-Hercock, J. Collis, and D. Ndumu, "Co-operating mobile agents for distributed parallel processing," in *Third International Conference on Autonomous Agents (AA '99)*. Mineapolis, MN: ACM Press, May 1999, pp. 398–399.

[35] G. Cugola, C. Ghezzi, G. P. Picco, and G. Vigna, "Analyzing mobile code languages," in *Mobile Object Systems: Towards the Programmable Internet*, 1996.

[36] A. J. Chakravarti, X. Wang, J. O. Hallstrom, and G. Baumgartner, "Implementation of strong mobility for multi-threaded agents in Java," in *Proceedings of the International Conference on Parallel Processing*. IEEE Computer Society, Oct. 2003.

[37] ——, "Implementation of strong mobility for multi-threaded agents in Java," Dept. of Computer and Information Science, The Ohio State University, Tech. Rep. OSU-CISRC-2/03-TR06, Feb. 2003.

[38] D. B. Lange and M. Oshima, *Programming & Deploying Mobile Agents with Java Aglets*. Addison-Wesley, 1998.

[39] A. J. Chakravarti, G. Baumgartner, and M. Lauria, "The Organic Grid: Self-organizing computation on a peer-to-peer network," Dept. of Computer and Information Science, The Ohio State University, Tech. Rep. OSU-CISRC-10/03-TR55, Oct. 2003.

[40] Gnutella. [Online]. Available: http://www.gnutella.com

[41] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A scalable content addressable network," in *Proceedings of ACM SIGCOMM'01*, 2001.

[42] Basic Local Alignment Search Tool. [Online]. Available: http://www.ncbi.nlm.nih.gov/BLAST/

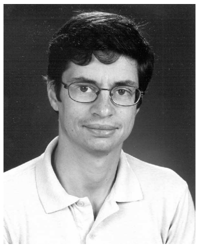
[43] A. J. Chakravarti, G. Baumgartner, and M. Lauria, "Application-specific Scheduling for The Organic Grid," Dept. of Computer and Information Science, The Ohio State University, Tech. Rep. OSU-CISRC-4/04-TR23, April 2004.

[44] ——, "Application-specific scheduling for the organic grid," in *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing (GRID 2004)*, Pittsburgh, November 2004.

Arjav J. Chakravarti received the B.E. degree in Electronics Engineering from the University of Mumbai, India, in 2001, and the M.S. and Ph.D. degrees in Computer and Information Science from The Ohio State University in 2004. His research interests include distributed computing, autonomic computing, mobile agents, and grid computing. He is currently working on distributed systems development at The MathWorks, Inc.

Gerald Baumgartner received the Dipl. Ing. degree from the University of Linz, Austria, and M.S. and Ph.D. degrees from Purdue University, all in computer science. He began his academic career at The Ohio State University in 1997 and is currently visiting the Department of Computer Science at Louisiana State University.

His research interest includes compiler optimizations, the design and implementation of domain-specific and object-oriented languages, desktop grids, and development and testing tools for object-oriented and embedded systems programming.



Mario Lauria earned a Laurea degree in Electrical Engineering and a Ph.D. in Computer Science from the Università di Napoli "Federico II", and a M.S. in Computer Science from the University of Illinois at Urbana-Champaign. Following completion of his Ph.D. in 1997 he spent a year at UIUC and one at the University of California, San Diego, as a post-doctoral associate working on cluster architecture. In March 2000 he joined the Ohio State University in Columbus, OH, as an Assistant Professor in the Department of Computer Science and Engineering; since 2001 he holds a joint appointment in the department of Biomedical Informatics.

His research interests include scalable cluster I/O, data intensive computing on the grid, high performance computational biology, biologically inspired models of computation, cellular computation. He is the recipient of a Fulbright scholarship and of a NATO Advanced Science Fellowship.