Portable Run-Time Support for Dynamic Object-Oriented Parallel Processing[1]

Andrew S. Grimshaw
Jon B. Weissman
W. Timothy Strayer

Abstract

Mentat is an object-oriented parallel processing system designed to simplify the task of writing portable parallel programs for parallel machines and workstation networks. The Mentat compiler and run-time system work together to automatically manage the communication and synchronization between objects. The run-time system marshals member function arguments, schedules objects on processors, and dynamically constructs and executes large grain data dependence graphs. In this paper we present the Mentat run-time system. We focus on three aspects — the software architecture, including the interface to the compiler and the structure and interaction of the principle components of the run-time system; the run-time overhead on a component by component basis for two platforms, a Sun SparcStation 2 and an Intel Paragon; and an analysis of the minimum granularity required for application programs to overcome the run-time overhead.

Keywords: object-oriented, parallel processing, dataflow, distributed memory, MIMD

1. Introduction

The object-oriented paradigm has proven to be a powerful tool for managing complexity in the development of software for sequential computers, and its power is being exploited in the more complex domain of software for parallel machines as well [5][6][8][22][32][40]. Software for parallel machines must provide efficient run-time support if it is to achieve acceptable performance. This is particularly true for distributed-memory MIMD machines where communication and synchronization costs can be large. Given that high performance is the *raison d'etre* of parallel computing, performance cannot be allowed to suffer due to run-time overhead. Therefore, if the object-oriented approach is to be successfully applied to parallel systems, efficient run-time support must be provided.

Mentat is an object-oriented parallel processing system designed to simplify the task of writing portable parallel applications software [20][22][23][44]. The fundamental objectives of Mentat are to (1) provide easy-to-use parallelism, (2) facilitate the portability of applications across a wide range of platforms, and (3) achieve good performance. The first two objectives are addressed through Mentat's underlying object-oriented approach: high-level abstractions mask the complex aspects of parallel programming, including communication, synchronization, and scheduling. The question is whether Mentat — or any parallel processing system — can meet the first two objectives and not sacrifice the third. How Mentat dynamically supports its object-oriented programming language, and the run-time costs incurred, are the topic of this discussion.

1. This work is partially funded by NSF grants ASC-9201822 and CDA-8922545-01, and NASA NGT-50970.

Mentat has two primary components: the Mentat Programming Language (MPL) and the Mentat run-time system (RTS). The MPL is an object-oriented programming language based on C++. The granule of computation is the class member function. The programmer is responsible for identifying those object classes whose member functions are of sufficient computational weight to allow efficient parallel execution. MPL programs manipulate instances of these special classes by invoking member functions on them just as if they were instances of C++ classes.

The compiler and run-time system work together to ensure that the data and control dependencies between Mentat class instances are automatically detected and managed without programmer intervention. The underlying assumption is that the programmer's strength is in making decisions about granularity and partitioning, while the compiler together with the run-time system can better manage communication, synchronization, and scheduling. This simplifies the task of writing parallel programs.

Mentat differs from other systems in several ways.[2] First, Mentat is the only system in the literature that combines the object-oriented paradigm with coarse grain dataflow. Further, Mentat dynamically detects and manages data dependencies as they develop, rather than at compile time. This facilitates larger grain computation by eliminating many small grain control actors which are inefficient in a distributed-memory environment. Second, Mentat supports both task and data-parallelism, not just data-parallelism as in [8][15][18][28][33]. Third, Mentat operates over a spectrum of architectures, from loosely coupled heterogeneous networks of workstations, to tightly coupled multicomputers. Finally, Mentat has a scalable, distributed, control mechanism. This includes both the mechanism used to construct and modify program graphs at runtime as well as the scheduling mechanism. We must point out, however, that Mentat is not suitable for all applications: Mentat is useful for medium to coarse grain applications. If the computation granularity of an application is too small, performance will suffer.

Performance aspects of Mentat are presented in [20][23]. For a general overview of Mentat and of the Mentat parallel processing philosophy see [22]. In this paper we present the Mentat run-time system which supports the Mentat Programming Language. Our objectives are threefold: first, to describe the inner workings of the run-time system, its software architecture, and the structure and interaction of its components, second, to present the run-time costs incurred by Mentat applications on two platforms, and third, to provide insight into the minimum computation granularity required on those platforms. We begin with background information on the macro dataflow model of computation and the Mentat Programming Language. Macro dataflow is the model used by Mentat, and implemented by the run-time system. The run-time system architecture is our next focus, starting with the virtual machine structure, a description of the services provided, and their design. We follow that with a sketch of the unique implementation aspects and performance of two of the run-time system implementations, on a network of Sun SparcStation 2 workstations and an Intel Paragon.

2. Background

2.1 Macro Dataflow

The macro dataflow (MDF [19]) model is a medium grain, data-driven computation model inspired by dataflow [2][16][41][47]. Recall that in dataflow, programs are directed graphs where the vertices are computational primitives (e.g., add, subtract, compare, etc.) called *actors*, the edges, or *arcs*, model data dependencies, and *tokens* carry data along the edges between the actors. An actor is a function that maps

2. A more complete discussion of related work appears in Section 5.

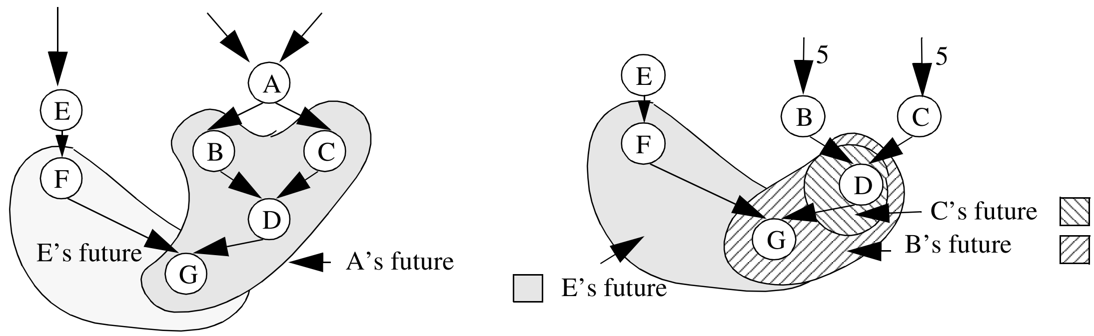


Figure 1. (a) A macro dataflow subgraph. The future of the actor A is shown. (b) New futures after actor A from (a) returns a value.

inputs to outputs. Dataflow is data-driven in that programs are self-synchronized by data motion. An actor may only execute when all of the required data, in the form of tokens, has arrived.

Macro dataflow differs from traditional dataflow in three ways. First, the computation granularity is larger than in traditional dataflow [3][7][10]. Actors are high-level functions such as matrix-multiply specified in a high-level language, not primitive operations such as add. The required granularity in the Mentat MDF model varies from platform to platform, but is in the thousands of instructions rather than tens of instructions. In Section 4.4, we present a detailed analysis of the required granularity. Second, some actors, called *persistent actors*, may maintain state between invocations. Sets of persistent actors may share the same state. The actors that share the same state are executed in mutual exclusion, in a monitor-like fashion. Finally, program graphs are not fixed at compile-time; instead, program graphs are constructed at run-time by observing the data dependencies as execution unfolds.

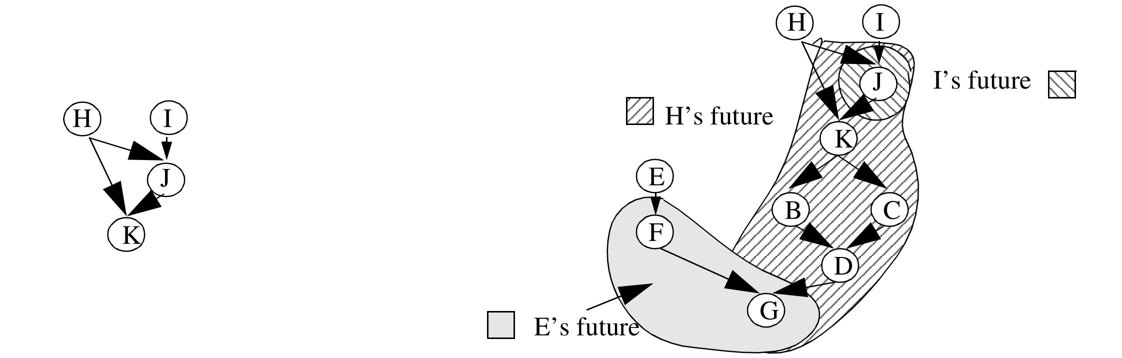
Program graphs are represented in MDF using *futures*.[3] A future represents the future of the computation with respect to a particular actor at a particular instant in time. In this respect MDF futures are similar to a parallel form of a continuation [1]. For example, consider the program graph fragment of Figure 1(a). *A*'s future is shown by the shaded area enclosing the actors (B, C, D, G). The future of *A* includes all computations that are data dependent on the result of the computation that *A* performs. Although actors *E* and *F* are in the same program graph as *A*, they are not in *A*'s future, nor is *A* in their future.

Within the MDF execution model, actors such as *A* may execute when their tokens arrive. *A*'s future is passed along with the tokens.When the actor completes, one of two things happens: the actor returns a value that is transmitted to each direct descendent in its future, or the actor elaborates itself into a subgraph. In either case, modifications need to be made to the program graph, either to reflect the completion of the actor or to include the new subgraph. Because the modifications require changing *A*'s future only, the modifications can be made locally. Other processors, such as those executing *E* or *F*, need not be notified.

Suppose that *A* returns the value 5. Since *A* has two output arcs, *A*'s future is broken into two futures which, along with the value 5, are forwarded to *B* and *C*. The new state is shown in Figure 1(b). *B* and *C* are now enabled and may execute since they have a token (value) on each input arc.

---

3. MDF futures should not be confused with Multilisp futures [27]. Multilisp futures represent a promise to deliver a value in the future.



(a) A's elaborated subgraph

(b) New graph

Figure 2. Actor elaboration. The actor A from Figure 1 may elaborate into the subgraph of (a). The resulting graph is shown in (b). This figure also illustrates an optimization with respect to futures. Although *H* and *I* have the same future, a single copy of this future is sent from *H* to *J* resulting in a smaller message.

Alternatively, *A* may elaborate itself into an arbitrary subgraph. Subgraph elaboration takes place when an actor's implementation is realized using other actors connected together to form a small subgraph. Which actors, and how they are connected together, is a language issue and will be discussed later. For now, suppose that *A* elaborates into the subgraph shown in Figure 2(a). The new state is shown in Figure 2(b). In this case the graph has grown, rather than contracted as in Figure 1(b). The point is that in both cases only *A*'s future needs modification; neither *E*, nor any other actor need be notified of the change.

2.2 The Mentat Programming Language

Rather than invent a new language for writing parallel programs, the Mentat Programming Language (MPL) is an extension of the object-oriented language C++. The extensions allow the programmer to provide granularity information to the compiler and run-time system.

The most important extension to C++ is the keyword "mentat" as a prefix to class definitions, as shown in Figure 3. This keyword indicates to the compiler that the member functions of the class are of sufficient computational weight to be worth executing concurrently in parallel. Mentat classes are defined to be either *regular* or *persistent*. The distinction reflects the two different types of actors in MDF. Regular Mentat classes are stateless, and their member functions can be thought of as pure functions in the sense that they maintain no state information between invocations. As a consequence, the run-time system may

```
mentat class bar {
// private member functions and variables
public:
   int op1(int,int);
   int op2(int, int);
};
```

Figure 3. A Mentat class definition. Without the keyword "mentat", it is a legitimate C++ class definition.

instantiate a new instance of a regular Mentat class to service each invocation of a member function from that class, even while other instances of the same class already exist.

Persistent Mentat classes, on the other hand, do maintain state information between member function invocations. Since state must be maintained, each member function invocation on a persistent Mentat object is served by the same instance of the object.

Instances of Mentat classes are called *Mentat objects*. Each Mentat object possesses a unique name, an address space, and a single thread of control. Mentat objects are *logically* address space-disjoint. The current implementation of Mentat objects is address space-disjoint, although a thread-based implementation is planned. Because Mentat objects are address space-disjoint, all communication is via member function invocation and is accomplished by RPC. Because Mentat objects have a single thread of control, they have monitor-like properties. In particular, only one member function may be executing at a time on a particular persistent object. The single thread and disjoint address space provide constrained access to contained variables and state manipulated by the object, and prevent race conditions.

Variables whose classes are Mentat classes are analogous to variables that are pointers. They are not an instance of the class; rather, they name or point to an instance. We call these variables *Mentat variables*. As with pointers, Mentat variables are initially *unbound* (they do not name an instance) and must be explicitly *bound*. A bound Mentat variable names a specific Mentat object. Unlike pointers, when an unbound Mentat variable is used and a member function is invoked, it is not necessarily an error. If the variable names a regular Mentat class, the underlying system instantiates a new Mentat object to service the member function invocation. On the other hand, instances of persistent Mentat classes must be bound explicitly through other language mechanisms not discussed here.

2.2.1. Member Function Invocation

Member function invocation on Mentat objects is syntactically the same as for C++ objects. Semantically, however, there are three important differences. First, Mentat member function invocations are non-blocking, providing parallel execution of member functions when data dependencies permit. Second, each invocation of a regular Mentat object member function causes the instantiation of a new object to service the request. This, combined with non-blocking invocation, means that many instances of a regular class member function can be executing concurrently. Finally, Mentat member functions are always call-by-value because the model assumes distributed memory. All parameters are physically copied to the destination object. Similarly, return values are by-value. Pointers and references may be used as formal parameters and as results, but the effect is that the memory object to which the pointer points is copied. Variable size arguments are supported as well, since they are convenient for implementing class libraries (e.g., matrix algebra classes).

2.2.2. The Return-to-Future Mechanism

Mentat member functions use the return to future `rtf` to return values. The value returned is forwarded to all Mentat object member function invocations that are data dependent on the result, and to the caller if necessary.

While there are many similarities between the C `return` and the MPL `rtf`, they differ in three significant ways. First, a `return` returns data to the caller. An `rtf` may or may not return data to the caller depending on the data dependencies of the program. If the caller does not use the result locally, then the caller does not receive a copy. This reduces communication overhead. Second, a C `return` signals the end of the computation in a function, while an `rtf` does not. An `rtf` indicates only that the result is available.

Since each Mentat object has its own thread of control, additional computation may be performed after the `rtf`, e.g., to update state information or to communicate with other objects. By making the result available as soon as possible, we permit data dependent computations to proceed concurrently with the local computation that follows the `rtf`. This is analogous to send-ahead in message passing systems. Third, in C, before a function can `return` a value, the value must be available. This is *not* the case with an `rtf`. When a Mentat object member function is invoked, the caller does not block; rather, we ensure that the results are forwarded wherever they are needed. Thus, a member function may `rtf` a "value" that is the result of another Mentat object member function that has not yet been completed, or perhaps even begun execution, as in Figure 2(b).

2.2.3. The mselect/maccept Statement

The MPL `mselect/maccept` statement is modeled on the Ada `select/accept`. The programmer may specify which member functions are candidates for execution by including them in the `mselect`. As in Ada, the entries may be protected with a guard that must evaluate to true at run-time in order for the member function to be a candidate for execution. Unlike Ada, rendezvous semantics are not supported and the caller does not block. The `mselect/maccept` statement gives the programmer the ability to control access to the object by specifying conditions for access. It also permits the construction of more elaborate and explicit synchronization constructs than are implicit in the language.

2.2.4. Task and Data Parallelism in Mentat

Task parallelism is realized using the language features just described. Instances of Mentat classes are created by users and used much as their C++ counterparts. For each instance a thread and a logically disjoint address space is created. The compiler generates code to marshall arguments and to construct macro dataflow graphs at run-time. The nodes in the graphs are member function invocations, the arcs model the formal parameters, and tokens represent the actual parameters.

Data parallelism in Mentat applications is currently realized manually using the task parallelism features and encapsulation. The basic idea is simple. A data parallel object is implemented using a persistent Mentat class that encapsulates several persistent Mentat objects. Member functions of the class are implemented by calling the same member function on the contained persistent Mentat objects which contain the data for the class. For example, a Mentat matrix class has been implemented by decomposing the matrix into *k* sub-matrices [34]. Each sub-matrix is implemented using a persistent Mentat object. Member functions that manipulate the whole matrix are treated as task parallel invocations by the caller. Internally they are implemented by appropriately manipulating the sub-matrices. This is known as intra-object parallelism. The result is a program graph elaboration as described earlier in which the program subgraph has a node for each contained Mentat object. This results in a combined task- and data-parallel program.

Similar techniques can be used for a wide-range of data structures, not just regular, dense, two-dimensional arrays. The difficulty with this method for realizing data parallelism is that while it is straight-forward for the programmer, it is somewhat tedious. The programmer is responsible for managing the iteration space, data domain decomposition, and the exchange of boundary regions for some problems. The iteration must be performed over the contained Mentat objects and their contained data elements.

To eliminate the need to hand generate this rather tedious code we are exploring the incorporation of data parallel language features into the Mentat programming language [46]. The language extensions are similar to those being developed by Gannon [32], Quinn [28], Larus [31] and others. Our work is significantly different in that the result will be a language that combines task and data parallelism in a seamless fashion, while the above languages are data parallel only.

3. The Run-Time System (RTS)

MPL programs are executed on a virtual macro dataflow machine implemented by the RTS. Each Mentat class member function corresponds to an MDF actor, and each formal parameter corresponds to an incoming arc for that actor. Tokens correspond to the actual parameters of the member function invocation. Token matching is the process of matching tokens (messages) that are the actual arguments of the same Mentat object member function invocation. The RTS constructs program graphs, manages communication and synchronization, performs token matching, performs object instantiation and scheduling, and allows selective message reception to support an ADA-like select/accept semantics. The virtual machine abstracts the platform-specific details so that the MPL code is portable across various MIMD architectures. The run-time system is currently running on several systems: the Intel iPSC/860 and Intel Paragon, both using NX/2 [29], the IBM SP-2, and networks of Sun, HP, Silicon Graphics, and IBM RS/6000 workstations using UDP packets and Unix sockets.

3.1 The Virtual Machine

The virtual macro dataflow machine provides a service interface to the compiled MPL code. The services are divided into two groups, those library routines that are linked with application objects, and those that are provided by independent daemon Mentat objects. The library services include communication, dataflow detection, and token matching for persistent Mentat objects. Object scheduling and instantiation is provided by instantiation managers (IM’s), and unbound token matching by the token matching units (TMU’s). The IM’s and TMU’s are daemons.

Internally, the virtual machine uses a classic layered approach. The top level provides an interface through which the compiler-generated code interacts with the virtual machine. The interface in turn is implemented as machine-independent modules. At the bottom level are the architecture- and operating system-specific modules where all platform-specific code has been isolated. The platform-specific modules include the communication system MMPS (the Modular Message Passing System [25]), and the object loader used by the instantiation manager. By isolating platform-dependent code we simplify both software maintenance and the task of porting Mentat to new platforms.

Each Mentat object and daemon has a unique name, an address space, and a thread of control. Communication between objects is solely through the message passing system - there is no shared memory.

3.2 System Services

There are five basic services provided by the run-time system: (1) object naming and basic communication, (2) dataflow detection, (3) token matching for persistent objects, (4) token matching for regular objects, and (5) object scheduling and instantiation. As discussed earlier, the first three are implemented by libraries and the latter two are implemented by daemons.

The run-time libraries are linked to each Mentat application and to the daemon objects. The library interface is the virtual machine used by the MPL compiler. Naturally, the interface is object-oriented, and consists of a set of class definitions. Instances of these classes are manipulated by the compiler. The most important classes are `computation_instance`, `mentat_object`, `mentat_message`, and `predicate_manager`.

```
class mentat_object {
   object_name i_name;
public:
   CIP invoke_fn(int fn_number,int arg_count,arg_struct ...);
   // invoke_fn marshals the arguments. fn_number indicates the
   // function, arg_count the number of arguments
   void create(); // instantiate new back-end
   void destroy();// destroy the back-end
};
```

Figure 4.Partial interface of the front-end `mentat_object` class.

computation_instance

The RTS keeps track of Mentat object member function invocations at run-time using c*omputation instances*, which correspond to nodes in a MDF program graph. They contain the name of the Mentat object invoked, the number of the invoked member function, the *computation tag* that uniquely identifies the computation, a list of the arguments (either values or pointers to other computation instances that will provide the values), and a successor list (also computation instances). A computation instance contains sufficient information to acquire the value that is the result of the operation. A CIP is a computation instance pointer.

mentat_object

The RTS implementation of Mentat objects consists of two components, (1) the front-end class `mentat_object` that contains the name of a Mentat object (process) and is the handle with which operations on the Mentat object are performed, and (2) a back-end server object process that contains the Mentat object's state and performs the member functions. Mentat class member function names are mapped by the compiler to function numbers, thus a `pop()` operation on a queue might be mapped to function number 104. Member function invocation involves using the front-end as a surrogate for the back-end server object. The front-end `mentat_objects` are essentially object names and a set of member functions used to communicate with the back-end server. The compiler generates the server loops that implement the back-ends and code to manipulate `mentat_objects`. Three member functions defined on every `mentat_object`, `invoke_fn()`, `create()` and `destroy()` are shown in Figure 4.

The member function `invoke_fn()` is called when a Mentat object member function is invoked. For example, to invoke `A.pop()` the compiler would emit a `A.invoke_fn(104,0)` to call function 104 on object `A` with no arguments. `Invoke_fn()` creates a new computation instance for the computation, (i.e., a new program graph node is created), and marshals the arguments, both actual arguments (line 3, Figure 5(a)), and arguments that are computation instances (line 5, Figure 5(a)). If an argument is a computation instance, `invoke_fn()` adds an arc from the argument to the new computation instance it is constructing. The functions `create()` and `destroy()` are called when a persistent Mentat object is instantiated and destroyed respectively.

mentat_message

A `mentat_message` is a message that contains the destination object name, the destination function number, an argument number, a computation tag, a future, and some data. Mentat messages have the property that if the destination object name is not bound, i.e., it does not name a specific instance, the message will be sent to a token matching unit (TMU).

```
bar A,B,C;
int w,x,y;
w = A.op1(4,5);
x = B.op1(6,7);
y = C.op1(w,x);
rtf(y);
```

(a) Draw an arc from A.op1() and B.op1() to C.op1().

```
bar A;
int w,x,y;
w = A.op1(4,5);
y = w +1;
```

(b) `w` is used in a strict expression, block at wait for value.

Figure 5. Two uses of result variables. In this example, bar is a regular Mentat class.

predicate_manager

A `predicate_manager` specifies a predicate which is a set of member functions that are candidates for execution. The compiler constructs `predicate_managers` for each `mselect/maccept` in the program. The `predicate_manager` class implements the token matching function for bound Mentat objects.

3.2.1. Naming and Basic Communication

Naming and basic communication services are provided by the Modular Message Passing System (MMPS) [25]. MMPS provides C++ class-based naming and message passing services, message construction, synchronous and asynchronous send, and blocking and non-blocking receive.

A MMPS-name is an address that names a specific object (process) on a specific processor. It contains information to allow the implementation to communicate with the object using the underlying host operating system's IPC primitives. For example, on UNIX systems using UDP datagrams, a MMPS name contains an index into a host table that contains 30 character hostnames (IP format), and a port number. On the Intel iPSC/860 and Intel Paragon, a MMPS name contains an integer host identifier and an integer process identifier. The use of an address in the MMPS name means that named objects may not migrate — a restriction that we have not found burdensome to date. We expect to change this soon, though, to allow objects to migrate to support fault tolerance and better resource utilization.

3.2.2. Run-Time Dataflow Detection

The objective of run-time dataflow detection is to dynamically detect and manage data dependencies between Mentat object invocations, mapping the resulting dependence graph onto a macro dataflow program graph. The data dependencies between Mentat object function invocations correspond to arcs in the MDF program graph.

The dataflow detection library routines monitor the use of certain variables (called *result variables*) at run-time to produce data dependence graphs. The basic idea is to monitor the use of Mentat objects, and the use of the results of Mentat object member function invocations. Informally, if at run-time we observe a variable `w` (Figure 5) being used on the left-hand side of a Mentat object member function invocation, we mark `w` as *delayed* and monitor all uses of `w`. Whenever `w` is delayed and `w` is used as an argument to a Mentat object member function invocation, we construct an arc from the invocation that generated `w` to the consumer of `w`. If `w` is not delayed, we use its value directly. Whenever `w` is used in a *strict* expression, we start the computation that computes `w`, and block waiting for the answer. A strict expression is an expression in which we must have the value in order to proceed. For example, on line 4 of Figure 5(b) the plus operator is strict, we must have the value of w to proceed.

More formally, let A be a Mentat object with a member function

```
int operation1(int,int)
```

A *Mentat expression* is one in which the outermost function invocation is an invocation of a Mentat member function, e.g., the right-hand side of

```
x = A.operation1(4,5);
```

A Mentat expression may be nested inside of another Mentat expression, e.g.,

```
x = A.operation1(5,A.operation1(4,4));
```

The right-hand side of every *Mentat assignment statement* is a Mentat expression, e.g.,

```
x = A.operation(4,5);
```

A *result variable* (RV) is a variable that occurs on the left-hand side of a Mentat assignment statement, e.g., w in Figure 5. It has a delayed value if the most recent assignment statement to it was a computation instance and the actual value for the computation instance has not been resolved. An RV has an actual value if it has a value that may be used. To detect data dependence at run-time we monitor all uses of result variables, both on the left- and right-hand sides.

Each RV has a state that is either *delayed* or *actual*. We define the *result variable set* (RVS) to be the set of all result variables that have a delayed value. Membership in RVS varies during the course of object execution. We define the *potential result variable set* (PRV) to be the set of all result variables. A variable may be a member of the PRV set and never be a member of result variable set. Membership in the PRV set is determined at compile time.

The run-time system performs run-time dataflow detection by maintaining a table of the addresses of the members of the result variable set called the RV_TABLE. There are two cases to consider. If the result variable is not in the RVS then it is not in the RV_TABLE and it's value is already stored at it's address in memory. If the result variable address is in the RVS then it's value is depends on a Mentat expression not yet evaluated, and the result variable is stored in the RV_TABLE. Each RV_TABLE entry contains the address of a result variable, and a pointer to a computation instance. The computation instance corresponds to the most recent assignment to the variable.

There are four functions of interest that operate on the RV_TABLE:

```
RV_INSERT((char*)rv_address, computation_instance* node);
RV_DELETE((char*) rv_address);
force();
RESOLVE((char*) rv_address, int size);
```

The function RV_INSERT() creates an entry in the RV_TABLE for the result variable pointed to by rv_address with a computation instance pointed to by node. If an entry already existed for rv_address, it is overwritten. Thus, we are implementing the single assignment rule using the RV_TABLE and computation instances. RV_INSERT() is the mechanism for adding a PRV to the RVS.

The function RV_DELETE() deletes the RV_TABLE entry associated with rv_address if one exists. Before the entry is deleted, its associated computation instance is *decoupled*. By decoupling the computation instance the compiler tells the RTS that this computation instance will never occur on the right-hand side of an expression again. This is the mechanism for removing a PRV from RVS.

The function `force()` examines all of the locally created computation instances. If the computation instance has not previously been evaluated, i.e., its arguments have not been sent, then a future is constructed and the arguments are sent to the object that is being invoked. Once the arguments have been sent, the computation instance is marked as evaluated. If the computation instance has no actual arguments, then the corresponding member function invocation will receive its arguments from its predecessors in the program graph. AT this point, the computation instance will be marked as evaluated.

The function `RESOLVE()` is called when the user program requires a value for a result variable. This is the case when a strict expression is encountered. If an entry in the `RV_TABLE` exists for `rv_address`, `RESOLVE()` calls `force()`, and blocks until the result is available. Otherwise the value is known and simply returned. Once the result is available, `RESOLVE`() places the result into the memory to which the `rv_address` points.

In Figure 5 two program fragments were presented to illustrate blocking versus non-blocking member function invocation. The MPL translations for these code fragments are shown in Figure 7(a) and Figure 7(b), respectively. In Figure 7(c), the left-hand figure illustrates the program subgraph state before execution of the code fragment of (a). The actor labeled `F` executes the code fragment, creates the subgraph containing `A`, `B`, and `C`, and replaces itself with the subgraph. The result is the right-hand figure.

In Figure 7(d) there is no graph elaboration. Instead, because `w` is used in a strict expression, the compiler emits a `RESOLVE` to acquire the delayed value of `w`. The effect in this case is the execution of a short subgraph and a blocking RPC semantics.

This example illustrates the basic concepts used to detect data dependence at run-time. Although these particular code fragments were straight-line code, more complicated fragments, such as those that contain loops, conditionals, and multiple scopes, are handled in the same fashion.

3.2.3. Token Matching

In the macro dataflow model, as in pure dataflow, tokens belonging to a particular computation must be matched [41][47]. For example, in Figure 6(a), the tokens containing the values 4 and 5 must be matched.

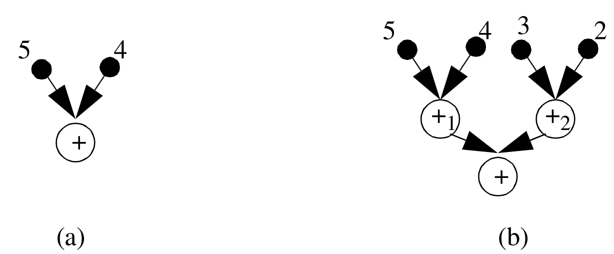


Figure 6. Token matching example. An actor may fire when there are tokens on each input arc.

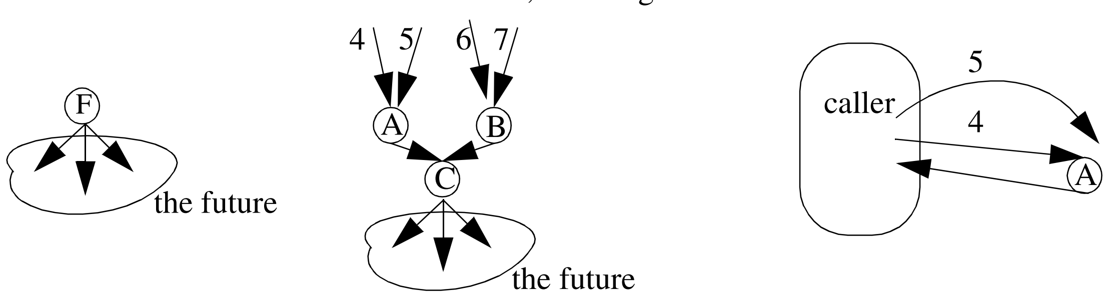
When both tokens are available, and they have been matched, we say the actor is enabled and may fire (execute). The matching process is complicated by the fact that there may be more than one instance of an actor in a particular program graph. For example, in Figure 6(b) there are three + actors and four tokens, two for the first operand of +, 5 and 3, and two for the second operand, 4 and 2. The problem is correctly matching the 5 and 4 to the $+_1$ execution and the 3 and 2 to the $+_2$ execution. The problem is further complicated by the fact that there may be multiple instances of $+_1$ executed, for example, as a loop unrolls, or for multiple parallel invocations of a function.

```
bar A;
int w,x,y;
//w = A.op1(4,5);
// Add w to RVS, create a node in the subgraph, marshall arguments.
(*RV_INSERT((&w)))=A.invoke_fn(101,2, // function 101 takes two arguments
   ICON_TO_ARG(4),                   // marshal the first argument
   ICON_TO_ARG(5));                  // marshal the second argument

//x = B.op1(6,7);
(*RV_INSERT((&x)))= B.invoke_fn(101,2,ICON_TO_ARG(6),ICON_TO_ARG(7));

//y = C.op1(w,x);
(*RV_INSERT((&y)))=C.invoke_fn(101,2,PRV_TO_ARG(&w,sizeof(w)),PRV_TO_ARG(&
x,sizeof(w)));

// rtf(y);
// elaborate the current actor into the constructed subgraph
rtf(PRV_TO_ARG(&y,sizeof(y)));
```

(a) Code transformation for Figure 5(a)

```
bar A;
int w,x,y;
//w = A.op1(4,5);
// Add w to RVS, create a node in the subgraph, marshall arguments.
(*RV_INSERT((&w)))=A.invoke_fn(101,2,ICON_TO_ARG(4),ICON_TO_ARG(5));

//y = w +1;
y = (RESOLVE(&w),w) + 1; // force() and wait for the value
```

(b) Code transformation for Figure 5(b). Control flow blocks waiting for the result of the member function invocation, resulting in an RPC-like behavior.



(c) Initial graph and elaboration for fragment (a)

(d) Graph for (b).

Figure 7. Code transformations and generated graphs for code fragments of Figure 5. ICON_TO_ARG and PRV_TO_ARG are marshalling functions for integer constants and PRV's respectively. PRV_TO_ARG marshals the argument if the RV is actual. If the RV is delayed, it constructs an arg_struct that points to the computation_instance that will generate the value.

The problem of identifying which tokens belong together has been solved in dataflow using token coloring [41][47]. Each actor invocation is assigned a unique color and then tokens destined for that actor instance are marked with that color. The problem then reduces to matching tokens of the same color. A similar coloring scheme is used in the Mentat RTS. Each computation is assigned a unique computation tag. The computation tag is formed from the MMPS name of the caller and an integer. Each invocation made by a particular caller is assigned a unique integer value used in forming the tag (we use a counter). Once the tokens are matched the actor may be scheduled for execution.

In Mentat, the token matching problem has two components, token matching for bound persistent Mentat objects whose location is known, and token matching for unbound regular object member function invocations. The primary difference is that in the former case all of the tokens can be sent directly to the object, where matching the tokens can easily be done in the objects local address space. In the latter case the location of the regular object is unknown — in fact it does not yet exist, and will not exist until its tokens have been matched. In this case, the tokens must be matched first by another agent, the token matching unit. The regular object token matching problem is complicated by the fact that the tokens may be generated on different processors in a distributed memory system, requiring the use of a distributed algorithm. The realization of token matching for bound and unbound tokens is provided by the predicate manager and the token matching unit respectively.

3.2.4. Token matching for persistent objects

Token matching for persistent objects is performed by the predicate manager (PM). The predicate manager is a library linked into all Mentat applications and objects. In addition to token matching, the predicate manager supports the MPL `mselect/maccept` statement. For example, the compiler can construct a predicate that says that it is interested in messages for member functions 101, 103, and 105, where 101, 103, and 105 take one, three and two arguments respectively. The predicate manager examines its database of received messages to determine if it has any complete 101, 103, or 105 *work units*. A work unit is a set of messages that all share the same computation tag, i.e., they correspond to the different arguments of the same instance of a member function invocation. A work unit is complete when all argument messages have arrived.

Work units are stored in a structure known as the *actor list*. The actor list contains an entry for each separate invocation on a member function. This is shown in Figure 8. For example, member function 105 has two entries in its actor list. The work unit corresponding to the first entry is complete. The collection of messages into work units corresponds to token matching in the MDF model.

The predicate management interface to the compiler allows the compiler to specify an ordered list of functions. Each function in the list corresponds to an accept in a `mselect/maccept`. The compiler generates code for a select/accept such that when the `mselect/maccept` is encountered at run-time, the guards are evaluated to determine the ordered list of functions with the list ordered by priority. This is later illustrated in Figure 9.

At run-time when an `mselect/maccept` is encountered the list of functions is generated and a call to `block_predicate` is made. First, `block_predicate` traverses the ordered list of functions. For each function it finds the actor list for that function and scans the actor list for a complete work unit. If a complete work unit is found, the messages corresponding to the arguments are returned and the work unit deleted. At this point the compiler generated code invokes the appropriate object member function using the values contained in the messages as arguments. If no complete work unit is found then the next function in the ordered function list is similarly checked. This continues until a completed work unit is found or the list

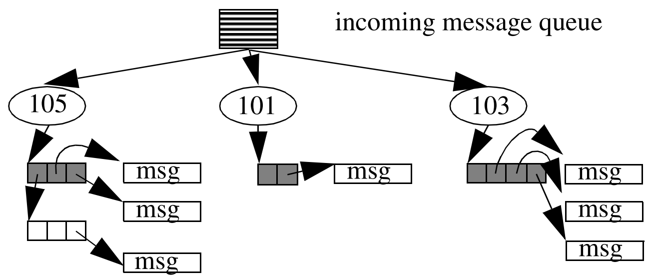


Figure 8. The predicate manager maintains its message database as a set of linked lists of *work units*. Incoming messages are forwarded to the correct list based on the function number. The complete work uni are shaded.

is exhausted. If none is found, *block_predicate* blocks on message arrival. The code fragment and transformation shown in Figure 9 illustrates how the predicate manager is used to support the `mselect/maccept` semantics.

3.2.5. Token matching for regular objects

The token matching unit (TMU) is responsible for matching tokens for regular object member function invocations. The basic problem is similar to that of token matching for bound objects discussed earlier. The difference is that there is no bound object to which to send the tokens where matching can occur, since the object that will perform the member function does not yet exist.

Token matching is accomplished using a distributed algorithm. The number of TMUs grows with system size so that the token matching does not become a bottleneck. On Unix machines, there is one token matching unit (TMU) for each processor in the system. On multicomputers such as the Intel Paragon, a TMU is assigned to each processor cluster. The algorithm is both simple and scalable, and has four stages that are executed for each regular Mentat object member function invocation: unbound token routing, token matching, instance acquisition, and token delivery.

During unbound token routing, tokens which are generated at Mentat objects and are destined for an unbound regular object member function invocation are routed to the same TMU. This TMU is determined using a hash function on the computation tag of the member function invocation instance. This is similar to the way hashing is done on token colors in dynamic dataflow [16][41]. The hash functions are very simple and uniformly distribute the tokens to the TMU's at run-time. A result of using a hash function to route tokens is that each token typically makes one hop in the first stage of the algorithm.

Once the tokens have arrived at the designated TMU they are stored in a token database. The database is organized in a manner similar to that used in the predicate manager. Tokens are collected into unmatched work units. Because tokens may both be large and unmatched for a long time, tokens may be stored on disk when the memory allocation for the TMU is exhausted.

When all of the tokens for a particular computation tag have arrived, the TMU issues an *instantiation* request to the local instantiation manager (IM). The TMU does not block on the instantiation request. Instead, it continues to receive unmatched tokens, and issues additional instantiation requests. Thus, several instantiation requests may be outstanding at any given time.

```
mselect {
            : maccept op1(int arg1,int arg2);
             break;
   (z>5)    : maccept op2(int arg1,int arg2);
            break;
};
```

(a) MPL `mselect/maccept` statement. The `maccept op2()` is guarded by (z>5).

```
{
int pred_number;
mentat_message *msg[2];
predicate_manager pred(2);
pred.enable_operation(0, 101, 2);
   // The first parameter is the result that will be returned if
   // operation 101 is accepted. The second parameter is the function
   // identifier, and the third indicates the number of arguments to 101.

if (z>5) pred.enable_operation(1, 102, 2);
   // The guard (z>5) must be true for operation 102 to be enabled.
pred_number=pred.block_predicate(msg);
 switch (pred_number) {
 case 0 : { // op1(int,int) has been called, service it.
   int arg1 = RESOLVE_MSG(int, msg[0]);
   // RESOLVE_MSG is a macro that convers a message to the indicated type.
   int arg2 = RESOLVE_MSG(int, msg[1]);
   op1(arg1,arg2);
   delete msg[0];
   delete msg[1]; }
   break;
 case 1 : { // op2(int,int) has been called, service it.
   int arg1 = RESOLVE_MSG(int, msg[0]);
   int arg2 = RESOLVE_MSG(int, msg[1];
   op2(arg1,arg2);
   delete msg[0];
   delete msg[1]; }
   break;
 }
}
```

(b) Translation of (a) generated by the MPL compiler. Each enable operation tells the predicate manager that calls on that function are to be accepted.

Figure 9. Code transformations for run-time support of `mselect/maccept`.

Token delivery begins once the IM has replied to the instantiation request. The instantiation manager returns a bound Mentat object name. The named object has been selected by the instantiation manager to service the member function. The TMU extracts from the token database the tokens for the computation and forwards them to the named object. Thus, each token is transported two hops, once to the TMU, and once from the TMU to the object. When the tokens arrive at the bound object, they are re-matched using the predicate manager, and the member function is executed to completion.

3.2.6. Scheduling and Instantiation

The instantiation managers (IM's) perform four basic functions, Mentat object placement (scheduling), Mentat object instantiation, binding and name services, and general configuration and status information services. The first two services, scheduling and instantiation, are the most important and the most interesting. The last two, binding and status information services, are primarily bookkeeping, e.g., reporting to the user the name of the objects running on a particular processor. Here we will confine our attention to scheduling and instantiation. A more detailed description can be found in [24].

The basic scheduling problem is to assign Mentat objects to processors in such a manner that total execution time of the application is minimized. The problem is complicated by three facts. First, nothing is known about the future resource requirements of the object being scheduled, or of the application as a whole. Second, the communication patterns of the application and the precedence relations between objects are unknown. Third, the current global state of the system is unknown, e.g., what are the utilizations and queue lengths of the processors. This third problem is further complicated by the fact that in distributed systems environments, there may be other users whose resource demands cannot be anticipated.

There is a rich literature on scheduling in distributed systems [13][17][26]. The general scheduling problem is NP-hard. Thus, much of the work in scheduling is based upon heuristics. Our scheduler, FALCON (Fully Automatic Load Coordinator for Networks) [24] is heuristic and based upon the work of Eager, Lazowska and Zahorjan [17] who developed a model for sender-initiated adaptive load sharing for homogeneous distributed systems. Their model uses system state information to describe the way in which the load in the system is distributed among its components.

Scheduling decisions are reached using a distributed algorithm. Each node (or virtual node in the case of multicomputers) has an instantiation manager and a token matching unit. The scheduling decision on each instantiation manager consists of two sub-decisions: (1) determining whether to process a task locally or remotely (transfer policy), and (2) determining to which node a task selected for transfer should be sent (location policy).

The transfer policy that we have selected is a *threshold* policy: a distributed, adaptive policy in which each node of the system uses only local state information to make its decisions. No exchange of state information is required in deciding whether to transfer a task. A task originating at a node is accepted for processing if the local state of the system is below some threshold. Otherwise, an attempt is made to transfer that task invocation request to another node. Only newly received tasks are eligible for transfer.

In order to avoid instability, where nodes are devoting all of their time to transferring tasks and none of their time to processing them, we employ a simple control policy that places a *transfer limit*, on the number of times a task may be transferred. When the transfer limit is reached the instantiation manager must accept the task. The transfer limit is set in a configuration file.

The location policy is invoked if the transfer policy does not accept the task for local instantiation or if a location hint specifies a different node for execution of the task. The three location policy algorithms that we have implemented are *random*, *round-robin* and *best-most-recently*.

Once the task location is determined, the instantiation of a persistent Mentat object is accomplished by loading a new copy of the executable. For the instantiation of regular objects, the IM tries to *reuse* an existing object of the same class, i.e., it does not load a new instance. Instead it returns the name of the existing object. This saves the time required to perform the load and exploits the fact the regular object instances are *stateless*.

4. Implementation

As of January 1995 Mentat has been implemented on ten platforms. The implementations fall into two categories, Unix-based workstation networks, the Sun 3, Sun 4, Silicon Graphics, HP, and IBM RS/

6000; and distributed memory MIMD multicomputers, the TMC CM-5, IBM SP-2, Intel iPSC/2,[4] Intel iPSC/860 Gamma, and the Intel Paragon. Within each category the implementations are similar, although there are some significant differences between the multicomputers. Rather than describe each implementation in detail, we present the core features of the Unix and multicomputer implementations. We then present the performance of the run-time system primitives on two well-known and representative platforms, the Sun SparcStation 2, and the Intel Paragon.

4.1 Unix Workstations

The multitasking Unix implementation of the run-time system is by far the most complex of our implementations, primarily because it requires the most system-specific code; the operating system communication support is very weak, interrupts of various forms must be managed, and the process model, while offering many options, differs in subtle ways from platform to platform.

Mentat objects are implemented by Unix processes that are forked by the instantiation manager. Multiple Mentat objects may reside on each Unix host and may execute concurrently. All communication between objects, including intra-host communication, is implemented in MMPS using UDP packets and an interrupt-driven, stop-and-wait protocol.

4.2 Multicomputers

The four multicomputers to which we have ported Mentat differ in several important respects — yet they are very similar in terms of their communication support; all provide for asynchronous, guaranteed delivery of arbitrary size messages. The differences lie in their level of process support. They fall into a spectrum from no process support on the TMC CM-5, to almost Unix-like process support on the iPSC/2 and Paragon, to full process support on the IBM SP-2.

On the CM-5 there is one user process per processor and all processors must execute the same executable image.[5] Thus, there is no support for dynamically loading different Mentat object executables on different processors, nor for multiple Mentat objects per processor. Instead all Mentat objects, the instantiation manager, the token matching unit, and the main program, must be linked into one large executable. A switch statement is then used to select which object to execute.

At the other extreme are the IBM SP-2, Intel iPSC/2 and Paragon which have multitasking operating systems. The iPSC/2 supports up to twenty processes per processor, and the capability to load new executables. Thus, the iPSC/2 execution model is very similar to the Unix model. The Paragon nodes run OSF-1 and has complete multitasking support.

In the middle is the Intel iPSC/860 Gamma. The Gamma supports exactly one process per processor. However, the executable image on the processor may be changed by issuing a load command to the operating system. Since both the Gamma and the CM-5 allow only one process per processor, a different process mapping paradigm is adopted.

4.3 Performance

Overhead is the friction of parallel processing. The performance of Mentat applications depends upon the overhead of the Mentat run-time system calls needed for program execution. Good performance

4. The iPSC/2 and the CM-5 are currently unsupported.
5. The CM-5 is actually multitasking, but a single user application is restricted to a single task per node.

requires that the run-time overhead be kept low. The performance of each Mentat run-time system component is presented for a network of Sun SparcStation 2 workstations and the Intel Paragon. Two types of run-time system overhead are measured: library calls and service requests. A library call is a function that is performed by the run-time system on behalf of the user application. Most library calls are simple function calls that operate in the local address space of the caller without any communication. A service request is a special type of library call that requires participation of a Mentat daemon process and requires communication. All measurements were taken when the machine or network usage was as low as possible.

We have divided the functions into two types: **primitives** and **composites**. The primitives include communication, dataflow detection, token and predicate management, and scheduling, and correspond to run-time system features already discussed. The composites contain a number of primitive operations and capture all of the overhead terms for three different scenarios. Below are descriptions of the overhead terms, and how they were measured, followed by the data for the two platforms.

<u>Communication</u>

- New and delete a `mentat_message` — Allocate a `mentat_message`, initialize its members, and delete it.
- Message transport — Send a message containing one byte of data. There is a 192 byte header in the Unix implementation, and a 124 byte header on the Paragon.The time is measured by starting a timer, sending the message from A to B, on receipt B sends a message back to A, on receipt A stops the timer. The result is then divided by two to determine the one-way cost.
- Message send — Send a one byte message but don't wait for delivery. This is measured by starting a timer, sending from A to B using the asynchronous send capability, and stopping the timer. The message has been delivered to the communication system (MMPS) only. MMPS will asynchronously transport the message.
- Communication bandwidth — the per byte communication bandwidth. Computed as the message transport time for an *N* byte message minus latency (1 byte message transport) divided by the message size *N*.

All communication times are based on MMPS[25]. All times are the average of 1000 iterations.

<u>Dataflow detection</u>

- New and delete a computation instance — Allocate a computation instance, initialize its members, and destroy it.
- Add an arc — Add an arc between two computation instances.
- Construct a future — Construct a future from the computation instances. Two cases are given, for a short RPC-like future, and for an unrolled loop with one hundred actor invocations.
- RV_TABLE insert — We measure the time required to insert an entry into the RV_TABLE. Because the table is implemented by a hash table, the insertion cost is constant since the number of entries are typically few.
- RV_TABLE look up — Look up an entry in the RV_TABLE.

<u>Token & predicate management</u>

- Block predicate — We measure the time required to search the message database. All of the necessary messages have already arrived. Thus, there is no blocking waiting for a message. However, the time to probe the message system to determine if a message has arrived is included. Two predicates were enabled, the match was found for the first predicate. The cost includes the token matching.

Scheduling

- Object instantiation — The time required to create a new Mentat object. This includes all scheduling time and the time to load the executable. Because measurements were done on an idle system these results should be viewed as lower bounds, not expected values. Note that the time will normally vary depending on the number of transfers required for scheduling. Subsequent loads of the same executable from the same instantiation manager are often much faster due to file caching by the operating system.

- Object destruction — The time required to destroy a Mentat object.

Composite costs

- Null RPC on a persistent object — Perform a blocking RPC that takes an integer parameter and returns an integer. This includes the time to construct the graph, transport the tokens, match the tokens at the invoked object, and return the result to the caller. No computation is performed.

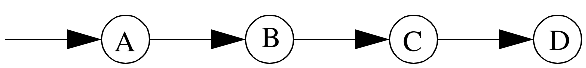
- Null RPC on a regular object — Here we measured the time for a single blocking RPC call on a regular mentat object. This includes the time to construct the graph, transport the tokens to the TMU, match the tokens at the TMU, have the TMU acquire an instance (via IM), transport the tokens from the TMU to the object, match the tokens at the object, and return the result to the caller.

- Total Mentat overhead — Null RPC time - 2×(message transport time). This captures Mentat overhead versus hand-written send/receive.

- Pipelined RPC on persistent objects — Four persistent objects are instantiated and then called as shown below, see Figure 10(a). The member functions perform no computation and return a value. The average time per member function invocation is determined by dividing the elapsed time by 100, e.g., by 4×25.

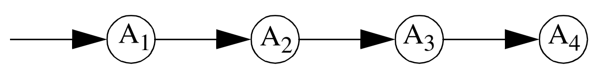
- Pipelined RPC on regular objects — This is similar to the above except that *A* is regular object, see Figure 10(b). The A_i are instances of *A* — some requests may be served by the same instance.

```
int i,j;
timer.start();
for (i 0;i<25;i++) {
   j= A.op1(i);
   j= B.op1(j);
   j= C.op1(j);
   j= D.op1(j);
   D.sink (j);
}
//wait for last computation
if (D.done());
timer.stop();
```



(a)

```
int i,j;
timer.start();
for (i 0;i<25;i++) {
   j= A.op1(i);
   j= A.op1(j);
   j= A.op1(j);
   j= A.op1(j);
   D.sink(j);
}
// wait for last computation
if (D.done());
timer.stop();
```



(b)

Figure 10

Environments

Sun SparcStation 2: The Sun configuration consisted of a collection of 8 Sparc2s connected by ethernet. Each Sparc (Sun 4) processor runs at 40 MHz and was configured with 32 MBytes of real memory. The run-time system performance is presented in Table 1.

Category	*Functions*	Cost uSec	
		Sun Network	**Intel Paragon**
Communication	**new and delete a message**	16	16
	message transport	1730	260
	message send	566	60[a]
	bandwidth (Kbytes/sec)	710	3930
Dataflow detection	**new and delete computation instance**	21	15
	add arc	3	3
	construct a future - short	57	6
	construct a future - 100 actors	1684	910
	RV_TABLE insert	13	20
	RV_TABLE lookup	10	20
Token management	**block predicate**	170	48
Scheduling	**object instantiation**	235,800	621,000
	object destruction	2400	300
Composites	**single null RPC - persistent**	3960	1280
	pipelined null RPC - persistent	775	413
	single null RPC - regular	12,320	520,800
	pipelined null RPC - regular	2950	10,210
	total Mentat overhead	500	760

a. The latency is for a 0-byte Mentat message that includes over 100 bytes of header and is therefore higher than the minimum latency reported elsewhere.

Table 1. Performance of Mentat RTS on Sun Network and Intel Paragon

Intel Paragon: The Paragon performance data was collected on a 56-node Intel Paragon at the Jet Propulsion Laboratory using an 8 node partition. Each i860 processor runs at 40 MHz and contains 32 MBytes of real memory. The run-time system performance is also shown in Table 1.

4.4 Performance Observations

On the Sun network, the cost of communication clearly dominates all other overhead terms. This is no surprise given the UDP/IP implementation. The cost of dynamic graph construction, monitoring RV's, creating computation instances, adding arcs, and constructing short future lists, is quite small (120 uS) when compared to the message transport costs. This number is obtained by adding up the component costs for these operations in Table 1.

On the Intel Paragon, communication does not dominate to nearly the same degree. Communication is an order of magnitude faster than on the Sun 4. Once again, this is not a surprise as the Paragon has an optimized communication system and special communication hardware. The result is that the RTS software overhead, which remains fairly constant between these two platforms, is the dominant overhead on the Paragon. On both platforms, object instantiation is also very slow due to the overhead of process creation in these environments — fork/exec in Unix and load in NX. The cost for regular object RPC and pipelining includes the cost of object instantiation while the cost for persistent object RPC and pipelining does not. Hence, the cost of these operations for regular objects is higher. On the Paragon the cost of regular object RPC is even higher since there is currently no regular object reuse on this platform.

With low level performance numbers in hand we can now address the question: Is Mentat efficient? Mentat introduces overhead that will not be experienced by a hand-coded implementation. This raises several new questions. What is the penalty versus hand-coded? Given the overhead costs what are the granularity requirements for Mentat objects? And finally, are there any performance benefits obtained by using Mentat? Below we answer these in turn.

<u>What is the performance penalty?</u> From Table 1 we see that the cost of performing an RPC "by hand" on the Suns is 2×(message_transport + new and delete message) = 2×(1730+31) = 3522 μS. The Mentat time is 3960 μS, a 12% increase. Whether this is acceptable depends on the application, though even a hand-coded application must be fairly coarse-grain to achieve good performance. On the Paragon the picture is worse. The "by hand" time is 2×(260+16) = 552 μS versus 1280 μS for Mentat, a 132% increase in overhead. This means that the minimum grain size for Mentat applications is significantly larger on the Paragon than for hand-coded applications. This is significant for applications that require a small grain size. As to inherently fine-grained problems, it is important to note that the ratio between processor performance (MFLOPS) and inter-processor communication bandwidth is expected to increase over time, making inherently fine-grained problems more difficult to execute on MIMD machines.

<u>What are the granularity requirements for Mentat applications?</u> The granularity requirements depend on the Mentat application. To give some insight into this problem, we consider a classic fan-out/fan-in computation with k workers implemented as a k-ary tree. This is a simple scenario that we use to determine the magnitude of the minimum granularity. More complex computations with a less regular structure would require a more detailed critical path analysis but would likely be within an order of magnitude of this value. To determine the granularity requirements for this problem, we compute the minimum grain size needed to break even. The break-even point is where the sequential time and parallel time are equal. That is, where the benefits of parallel execution begin to offset the overhead. Beyond this point, it is profitable to exploit parallelism. The minimum acceptable granularity should not be confused with the desired computation granularity. Mentat applications often exhibit a computation granularity much higher than the minimum in order to obtain the best possible performance. We will assume that we have idle processors and an idle network, i.e., that other users will not impact performance.

In Figure 11, we show a time-line for the two worker case with the major cost components included. The minimum number of workers needed to realize any performance gain if all real work is done by the workers is two; if one worker is used then we pay all of the overhead and receive none of the benefit of parallel execution. The critical path contains the labelled primitives. The total execution time is the time on the critical path:

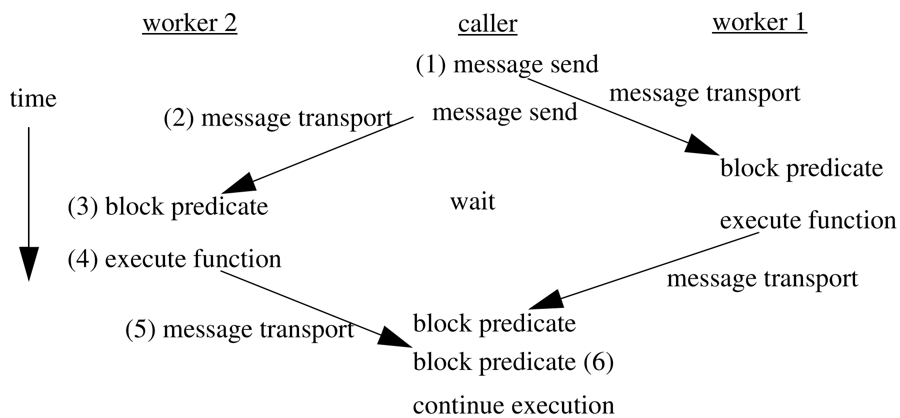


Figure 11

$$\underset{(1)}{\text{message send}} + \underset{(2,5)}{2\times\text{message transport}} + \underset{(4)}{\text{function time}} + \underset{(3,6)}{2\times\text{block predicate}}$$

The asynchronous message sends are sequential from the caller with the last message send included in message transport cost (2). If we assume that the function time is the same for both workers, then break even is achieved for two workers when:

2×(function time) = message send + 2×(message transport + block predicate) + function time

The left-hand side is the sequential time and the right-hand side is the parallel time computed above. Solving for function time yields:

function time = message send + 2×(message transport + block predicate)

This generalizes to k workers, break even is achieved when:

k×(function time) = *k*-1×message send + 2×(message transport + block predicate) + function time

function time = (*k*-1×message send + 2×(message transport + block predicate)) / (*k*-1)

This analysis makes the simplifying assumption that message transport cost is independent of the number of workers. Often this is not the case due to contention for communication bandwidth. The product of the function time and the processor rate (in MFLOPs) gives the minimum granularity of the break even point in units of floating point instructions. The Sun 4 used in the tests is a 7 MFLOP machine and the i860 is a 80 MFLOP peak performance processor. The i860 peak is not realizable on typical codes. A rate of 10 MFLOPS on the i860 is considered good [36]. We assume 7 and 10 MFLOPs for the Sun 4 and i860 respectively. Using the data from Table 1 and these floating point rates we obtain the function time and minimum grain sizes for *k*=2 and *k*=4 as shown in Table 2.

The table values are computed easily. Consider *k*=2 for the Sun network. The minimum function granularity for the Sun 4 is:

function time = 556 + 2×(1730+170) = 4356 μS

Since the Sun 4 is a 7 MFLOP machine, the minimum grain size in floating point operations is 7×4356 or 30,492 floating point operations. The other table values are computed in a similar fashion. As expected, the minimum granularity is much larger for the Sun 4 due to the high communication cost on the network.

Environment	Function time in usec		Grain size in fp instructions	
	k=2	*k=4*	*k=2*	*k=4*
Sun SparcStation 2	**4356**	**1823**	**30492**	**12761**
Intel Paragon	**676**	**265**	**6760**	**2650**

Table 2. Minimum Granularity on the Sun network and Intel Paragon

Performance benefits: A final comment on performance is in order. Mentat performance is not always worse than hand-coded. On the contrary, we have observed that Mentat performance can be superior to hand-coded performance. At first this might seem counter-intuitive. The reason is that Mentat often exploits concurrency that a programmer may not attempt to exploit because of the increased complexity.

Two examples illustrate this. First, there are data-parallel applications that iteratively perform a work phase, in which an operation is carried out in parallel on the data set, and a reduction phase, where the results are, for example, sorted and merged. The natural way to code this by hand is to perform the data-parallel operation for the i^{th} iteration, collect the results, perform the next data-parallel operation, and so on. Note the synchronization point at the end of each iteration. Often the Mentat code will not synchronize at the end of each iteration. Instead, the compiler will generate code that results in the pipelined execution of the data-parallel operation and the reduction phase. Thus, the data-parallel operation and the reduction are performed concurrently. Further, if the data-parallel operations require a data dependent variable amount of time then the workers will finish at different times. In a hand-coded implementation the next iteration begins only when all workers have completed. Once again, the Mentat code may be less synchronous, synchronizing only at the end of all iterations. Thus if different workers are slower for different iterations, then the less synchronous code will complete more quickly.

A second example is pipelined data filtering and feature extraction applications [38]. When constructing pipelined applications by hand a great deal of effort goes into balancing the amount of work performed in each stage because the throughput is limited by the slowest stage. If the computation requirements for a stage are data dependent then balancing the work by hand is difficult, and often not performed. This can be overcome by using regular Mentat objects. The system instantiates a new instance for each instance of each stage. Thus, if one iteration is slower, additional processor resources are utilized so that the next iteration may begin as soon as the data is available, keeping the pipe full, and preventing a "bubble" from forming. While this can be done by hand, it often is not.

5. Related Work

Work related to our effort falls into four categories, other object-oriented parallel processing systems, other compiler-based distributed memory systems, other portable parallel processing systems, and other large grain dataflow systems.

5.1 Parallel Object-Oriented Systems

In the object-oriented parallel processing domain Mentat differs from systems such as Parmacs [5], Presto [6], and Jade [30] (shared memory object-based systems) in its ability to easily execute on both shared memory MIMD and distributed memory MIMD architectures, as well as hybrids. pC++ [8] and

Paragon [15] on the other hand are data-parallel derivatives of C++. Mentat accommodates both functional and data-parallelism, often within the same program. ESP [40] is perhaps the most similar of the parallel object-oriented systems. It, too, is a high-performance extension to C++ that supports both functional and data-parallelism. What distinguishes Mentat is our compiler support. In ESP remote invocations either return values or futures. If a value is returned, then a blocking RPC is performed. If a future is returned, it must be treated differently. Futures may not be passed to other remote invocations, limiting the amount of parallelism. Finally, ESP supports only fixed size arguments (except strings). This makes the construction of general purpose library classes, e.g., matrix operators, difficult.

There are also a large number of object-based systems and languages for distributed systems [4][14], as opposed to parallel systems. Mentat differs from these and other distributed object based systems in our objectives, we strive for performance via parallelism rather than distributed execution.

5.2 Compiled Distributed Memory Systems

Until recently, there were few results for compiled — as opposed to hand-coded — applications on distributed memory machines. There are now several active projects in this area, Fortran-D [18], HP Fortran [33], Dataparallel C [37], Paragon [15], and the Inspector/executor [48] model to name a few. These are primarily data-parallel languages with [48] tied to a data-parallel model of computation. What differentiates our work from theirs is that Mentat exploits opportunities for both functional and data-parallelism. Further, in Mentat, parallelism, and the communication and synchronization constructs that are required, is dynamically detected and managed. Most other systems statically determine the communication and synchronization requirements of the program at compile-time.

5.3 Portable Systems

Applications portability across parallel architectures is an objective of many projects. Examples include PVM [43], Linda [11][12], the Argonne P4 macros [9], Fortran D [18] and Parti [35]. Our effort shares with these and other projects the basic idea of providing a portable virtual machine to the programmer. The primary difference is the level of the abstraction. Low-level abstractions (e.g., send/receive, and in/out) such as in [9][11][43] require the programmer to operate at the assembly language level of parallelism. This makes writing parallel programs more difficult. Others [7][18][28][39] share our philosophy of providing a higher level language interface in order to simplify applications development. Parti [35] is limited to runtime data level parallelism. What differentiates our work from other high-level portable systems is that we support both functional and data-parallelism as well as support the object-oriented paradigm.

5.4 Large Grain Dataflow

Mentat and macro dataflow (MDF) differ from other coarse-grain dataflow systems such as CDF [3], Hence [7], and Code/Rope [10] in several ways. First, MDF graphs are dynamic rather than static. Second, some MDF actors maintain state internally, rather than copying and circulating tokens to maintain state. Third, the MDF program graphs are constructed at runtime using compiler-generated code. Thus, the programmer is not responsible for generating the program graphs using a graphical interface as in [7][10]. Finally, the MDF model has been implemented on *both* shared-memory and distributed-memory MIMD machines, unlike [10] which are restricted to shared-memory only.

6. Summary and Future Work

The application of the object-oriented programming paradigm to parallel computing depends on the efficient implementation of a supporting parallel run-time system. The Mentat run-time system provides a portable run-time environment for the execution of parallel object-oriented programs. The run-time system supports parallel object-oriented computing via a virtual macro dataflow machine that provides services for object instantiation and scheduling, `select/accept` management, remote member function invocation, dynamic data dependence detection and management, and object scheduling and instantiation. The run-time system achieves portability by using a layered virtual machine model that hides platform-specific details from the user and the compiler. To date the run-time system has been implemented on platforms that include networks of Sun 3s, Sun 4s, HPs, IBM RS/6000s, and Silicon Graphics workstations, as well as multicomputers such as the Intel iPSC/860, Intel Paragon, and the IBM SP-2.

We presented the overhead costs for the dominant run-time services on two platforms, the Sun SparcStation 2 and an Intel Paragon. Performance on applications is good [20][22][23] for problems of sufficient granularity and leads us to the conclusion that the run-time system overhead is tolerable for many applications. However, the overhead is too high for those applications that require fine-grain communication or that have been coded to require fine-grain communication.

Future work on the Mentat run-time system falls into three categories, porting to additional platforms, generating a threads-based implementation, and providing heterogeneous metasystem support. The quest to port to an expanding list of platforms continues. We have set our sights in the near term on the DEC alpha. In the multicomputer domain we are targeting the Cray T3D. Toward this end we are further isolating machine, operating system, and compiler dependencies so that porting will become even easier.

In a thread-based implementation of the run-time system Mentat, objects will no longer necessarily be physically address space disjoint. Communication between Mentat objects within the same address space will use messages implemented using shared memory. There are several advantages to a thread-based implementation: intra-host communication is much faster, shared memory multiprocessor (e.g., KSR or multiprocessor Sparc) implementations will be much faster than is possible with disjoint address spaces, and limitations on Mentat applications caused by underlying unitasking operating systems (e.g., iPSC/860 and CM-5) will be eliminated. This will enable Mentat to support more efficient execution of finer-grained applications. A significant disadvantage is that many user-defined library routines are not re-entrant. This can cause timing-dependent bugs that do not exist in the current implementation.

A metasystem is a collection of heterogeneous hosts, scalar workstations, vector processors, SIMD machines, and shared and distributed memory MIMD machines connected by a multilayer, heterogeneous interconnection network. Our objective is to integrate these hosts into a system that provides the illusion of one large virtual machine to users. As part of the Mentat metasystem testbed project [21] we are extending the Mentat run-time system into a heterogeneous environment. Issues that must be addressed include data alignment, data coercion, and scheduling, in particular, automatic problem decomposition and placement based on computation granularity and application communication topology [45].

Acknowledgments

This research was performed in part using the Intel Gamma and Paragon operated by Caltech on behalf of the Concurrent Supercomputing Consortium. Access to this facility was provided by the Jet Propulsion Laboratory. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

7. References

[1] H. Abelson, G. J. Sussman, and J. Sussman, *Structure and Interpretation of Computer Programs*, The MIT Press, Cambridge Massachusetts, 1985.

[2] T. Agerwala and Arvind, "Data Flow Systems," *IEEE Computer*, vol. 15, no. 2, pp. 10-13, February, 1982.

[3] R. F. Babb, "Parallel Processing with Large-Grain Data Flow Techniques," *IEEE Computer*, pp. 55-61, July, 1984.

[4] H. Bal, J. Steiner, and A. Tanenbaum, "Programming Languages for Distributed Computing Systems," *ACM Computing Surveys*, pp. 261-322, vol. 21, no. 3, Sept. 1989.

[5] B. Beck, "Shared Memory Parallel Programming in C++," *IEEE Software*, 7(4) pp. 38-48, July, 1990.

[6] B. N. Bershad, E. D. Lazowska, and H. M. Levy, "Presto: A System for Object-Oriented Parallel Programming," *Software - Practice and Experience*, 18(8), pp. 713-732, August, 1988.

[7] A. Beguelin et al., "HeNCE: Graphical Development Tools for Network-Based Concurrent Computing," *Proceedings SHPCC-92*, pp. 129-136, Williamsburg, VA, May, 1992.

[8] F. Bodin, et. al., "Distributed pC++: Basic Ideas for an Object Parallel Language," *Proceedings Object-Oriented Numerics Conference*, pp. 1-24,Sunriver, Oregon, April 25-27, 1993.

[9] J. Boyle et al., *Portable Programs for Parallel Processors*, Holt, Rinehart and Winston, New York, 1987.

[10] J. C. Browne, T. Lee, and J. Werth, "Experimental Evaluation of a Reusability-Oriented Parallel Programming Environment," IEEE Transactions on Software Engineering, pp. 111-120, vol. 16, no. 2, Feb., 1990.

[11] N. Carriero and D. Gelernter, "Linda in Context," *Communications of the ACM*, vol. 32, no. 4, pp. 444-458, April, 1989.

[12] N. Carriero, D. Gelernter, and T.G. Mattson, "Linda in Heterogeneous Computing Environments," *Proceedings of WHP 92 Workshop on Heterogeneous Processing*, IEEE Press, pp. 43-46, Beverly Hills, CA, March, 1992.

[13] T. L. Casavant, and J. G. Kuhl, 'A Taxonomy of Scheduling in General-Purpose Distributed Computing Systems,' *IEEE Transactions on Software Engineering*, vol. 14, pp. 141-154, February, 1988.

[14] R. Chin and S. Chanson, "Distributed Object-Based Programming Systems," *ACM Computing Surveys*, pp. 91-127, vol. 23, no. 1, March., 1991.

[15] A.L.Cheung, and A.P. Reeves, "High Performance Computing on a Cluster of Workstations," *Proceedings of the First Symposium on High-Performance Distributed Computing*, pp. 152-160, Syracuse, NY, Sept., 1992.

[16] J. Dennis, "First Version of a Data Flow Procedure Language," MIT TR-673, May, 1975.

[17] D.L. Eager, E. D. Lazowska, and J. Zahorjan, "Adaptive Load Sharing in Homogeneous Distributed Systems", *IEEE Transactions on Software Engineering*, vol. 12, pp. 662-675, May, 1986.

[18] G. C. Fox, et al., "Fortran D Language Specifications," Technical Report SCCS 42c, NPAC, Syracuse University, Syracuse, NY.

[19] A. S. Grimshaw, "The Mentat Computation Model - Data-Driven Support for Dynamic Object-Oriented Parallel Processing," Computer Science Technical Report, CS-93-30, University of Virginia, May, 1993.

[20] A. S. Grimshaw, E. A. West, and W.R. Pearson, "No Pain and Gain! - Experiences with Mentat on Biological Application," *Concurrency: Practice & Experience*, pp. 309-328, Vol. 5, issue 4, July, 1993.

[21] A.S. Grimshaw, J.B. Weissman, E.A. West, and E. Loyot, "Metasystems: An Approach Combining Parallel Processing and Heterogeneous Distributed Computing Systems," *Journal of Parallel and Distributed Computing*, Vol. 21, No. 3, June 1994.

[22] A. S. Grimshaw, "Easy to Use Object-Oriented Parallel Programming with Mentat," *IEEE Computer*, pp. 39-51, May, 1993.

[23] A. S. Grimshaw, W. T. Strayer, and P. Narayan, "Dynamic Object-Oriented Parallel Processing," *IEEE Parallel & Distributed Technology: Systems & Applications*, pp. 33-47, May, 1993.

[24] A. S. Grimshaw and V. E. Vivas, "FALCON: A Distributed Scheduler for MIMD Architectures", *Proceed-*

ings of the Symposium on Experiences with Distributed and Multiprocessor Systems, pp. 149-163, Atlanta, GA, March, 1991.

[25] A. S. Grimshaw, D. Mack, and W.T. Strayer, “MMPS: Portable Message Passing Support for Parallel Computing,” *Proceedings of the Fifth Distributed Memory Computing Conference*, pp. 784-789, Charleston, SC., April 9-12, 1990.

[26] A. Hac, “Load Balancing in Distributed Systems: A Summary”, *Performance Evaluation Review*, ACM, vol. 16, pp. 17-25, February 1989.

[27] R. H. Halstead Jr., “Multilisp: A Language for Concurrent Symbolic Computation,” *ACM Transactions on Programming Languages and Systems*, pp. 501-538, vol. 7, no. 4, October, 1985.

[28] P. J. Hatcher, et al, “Data-Parallel Programming on MIMD Computers,” *IEEE Transactions on Parallel and Distributed Systems,* vol. 2, no. 3, pp. 377-383.

[29] Intel Corporation, “iPSC/2 USER’S GUIDE,” Intel Scientific Computers, Beaverton, OR, March, 1988.

[30] M.S. Lam and C. Rinard, “Coarse-Grain Parallel Programming in Jade,” *3rd ACM SIGPLAN Symposium on Principle and Practice of Parallel Programming*, pp. 94-105, Williamburg, VA 1991.

[31] J. R. Larus, B. Richards, and G. Viswanathan, ”C**: A Large-Grain, Object-Oriented, Data-Parallel Programming Language,” UW Technical Report #1126, Computer Sciences Department, University of Wisconsin, November 1992.

[32] J. K. Lee and D. Gannon, “Object Oriented Parallel Programming Experiments and Results,” *Proceedings of Supercomputing ‘91*, pp. 273-282, Albuquerque, NM, 1991.

[33] D. B. Loveman, “High Performance Fortran,”*IEEE Parallel & Distributed Technology: Systems & Applications*, vol. 1, no. 1, pp. 25-42, February, 1993.

[34] L. MacCallum and A.S. Grimshaw, ‘‘A Parallel Object-Oriented Linear Algebra Library,’’ *Proceedings Object-Oriented Numerics Conference*, pp. 233-249, Sunriver, Oregon, April 25-27, 1994.

[35] R. Mirchandaney, et al., “Principles of Runtime Support for Parallel Processors”, *International Conference on Supercomputing*, St. Malo France, July 1988.

[36] S. A. Moyer, “Performance of the IPSC/860 Node Architecture,” Computer Science Technical Report, IPC-91-09, May 1991

[37] N. Nedeljkovic, and M.J. Quinn, “Data-Parallel Programming on a Network of Heterogeneous Workstations,” *Proceedings of the First Symposium on High-Performance Distributed Computing*, pp. 28-36, Syracuse, NY, Sept., 1992.

[38] L. Sparks, et al., “Large-Scale Retrieval of Atmospheric Parameters from Remote Sounding Data,” in the Technical Digest on Optical Remote Sensing of the Atmosphere, 1991 (Optical Society of America, Washington, D.C., 1991), Vol. 18, pp. 75-77.

[39] W. Shu and L.V. Kale, “Chare Kernel - a Runtime Support System for Parallel Computations,” *Journal of Parallel and Distributed Computing*, No. 11, pp. 198-211, 1991.

[40] S. K. Smith, et al., “Experimental Systems Project at MCC,” MCC Technical Report Number: ACA-ESP-089-89, March 2, 1989.

[41] V. P. Srini, “An Architectural Comparison of Dataflow Systems,” *IEEE Computer*, pp. 68-88, March, 1986.

[42] B. Stroustrup, “What is Object-Oriented Programming?*” IEEE Software*, pp. 10-20, May, 1988.

[43] V.S. Sunderam, “PVM: A framework for parallel distributed computing,” *Concurrency: Practice and Experience*, vol. 2(4), pp. 315-339, December, 1990.

[44] J. B. Weissman, A. S. Grimshaw, and R. Ferraro, “Parallel Object-Oriented Computation Applied to a Finite Element Problem,” *Scientific Computing*, Vol 2, pp 133-144, 1994.

[45] .J.B. Weissman and A.S. Grimshaw, “Network Partitioning of Data Parallel Computations,” *Proceedings of the Third International Symposium on High-Performance Distributed Computing*, August 1994.

[46] E.A. West and A. S. Grimshaw,”Braid: Integrating Task and Data Parallelism,” *Proceedings of Frontiers* ‘95, February, 1995.

[47] A. H. Veen, "Dataflow Machine Architecture," *ACM Computing Surveys*, pp. 365-396, vol. 18, no. 4, December, 1986.

[48] J. Wu et al, "Distributed Memory Compiler Design for Sparse Problems," ICASE Report 91-13, January 1991.