M×N Communication and Parallel Interpolation in CCSM3 Using the Model Coupling Toolkit

Robert Jacob, Jay Larson, Everest Ong

Argonne National Laboratory, Argonne, IL

ABSTRACT

The Model Coupling Toolkit (MCT) is a software library for constructing parallel coupled models from individual parallel models. MCT was created to address the challenges of creating a parallel coupler for the Community Climate System Model (CCSM). Each of the submodels that make up CCSM is a separate parallel application with its own domain decomposition, running on its own set of processors. This application contains multiple instances of the MxN problem, the problem of transferring data between two parallel programs running on disjoint sets of processors. CCSM also requires efficient data transfer to facilitate its interpolation algorithms. MCT was created as a generalized solution to handle these and other common functions in parallel coupled models. Here we describe MCT's implementation of the data transfer infrastructure needed for a parallel coupled model. The performance of MCT scales satisfactorily as processors are added to the system. However, the types of decompositions used in the submodels can affect performance. MCT's infrastructure provides a flexible and high-performing set of tools for enabling interoperability between parallel applications.

1. Introduction

A growing trend in high-performance scientific computing is the creation of new applications for a multidisciplinary problem by combining two or more separate applications from individual disciplines. One field that has pioneered this approach is climate modeling. A climate model usually contains multiple submodels that simulate the behavior of physical subsystems such as the global atmosphere, the global ocean, the land surface, and sea ice. Each model is produced by practitioners of a subdiscipline in the atmospheric and oceanic sciences. A coupled climate model is created by combining component[1] models and allowing them to mutually provide boundary conditions for each other.

Like other high-performance scientific applications, climate models are implemented as parallel programs operating on physically distributed data. Coupled climate models are a combination of individual distributed-memory parallel programs.

Version 3 of the Community Climate System Model (CCSM) (Collins *et al.*, 2005) is an example of a state-of-the-art parallel coupled climate model. CCSM is a collection of high-performance applications that simulate the interaction of the Earth's ocean and atmosphere, its land surface, and sea ice. Although the components of climate models are physically three-dimensional systems, their common interface is a two-dimensional surface. The coupling problem amounts to representing the physical fluxes across the two-dimensional interface in a consistent and coordinated way. CCSM's architecture is a "hub-and-spokes" model as shown in Figure 1. The exchange of information across the surface and overall time integration of the system are controlled by a fifth application called the coupler (the hub in Fig. 1). All models send and receive data only with the coupler and not with each other. This approach provides a convenient central point of control for dealing with important scientific requirements in the model, such as enforcing global conservation of energy exchanged between the models and compensating for different time steps in the course of simulating a day.

When creating a climate model like CCSM from separate parallel codes, the standard approach in the community has been to combine them under a single, ad hoc software framework. The new version of CCSM standardizes its framework with a newly designed library called cpl6. Cpl6 marks a major advance over previous versions of the CCSM coupler because of its modular design and allowance for distributed-memory parallelism in the coupler itself. Cpl6 and the design of CCSM are described further in a companion paper

[1] A "component" in this paper is a submodel of a coupled system and does not refer to a component programming model.

(Craig *et al.*, 2005).

For building parallel coupled models, cpl6 users a new software library called the Model Coupling Toolkit (MCT). Although written to address the coupling needs of an earth science model, MCT is a general-purpose library that can be used to couple any models exchanging data represented on a numerical grid, whether structured or unstructured. An overview of MCT is described in another companion paper (Larson *et al.*, 2005). Here we focus on how MCT solves the largest problem posed by the CCSM coupler: its parallel data transfer needs. Section 2 describes the parallel data transfer characteristics of CCSM and the problems that motivated the creation of MCT. Section 3 describes MCT's solutions to these problems. Section 4 examines the performance of MCT's data transfer methods. We conclude in Section 5 with a discussion of MCT's role in other applications.

2. Parallel Communication and CCSM's Coupler

The component models in CCSM typically have different numerical methods for solving their respective system of partial differential equations, and these methods may employ different types of numerical grids or grids with different resolutions. The physical component models in recent versions of CCSM use distributed-memory parallelism and domain decomposition to distribute grid points in the horizontal (north-south and east-west) direction.

The Message Passing Interface (MPI, version 1) (Message Passing Interface Forum, 1994) is typically used to communicate data *within* each component as required by their numerical methods. This intramodel communication has typically been handled by libraries or methods chosen or developed by the individual model development teams. MCT address the *model-coupler*, or intermodel, parallel communication requirements of CCSM and the intramodel communication requirements of a distributed-memory parallel coupler. In addition, MCT satisfies other requirements identified for the new coupler, such as extensibility and generalization of the model-coupler interface. Further discussion of these requirements and how they are met by MCT and cpl6 can be found in the companion papers (Larson *et al.*, 2005; Craig *et al.*, 2005).

2.1. Model-Coupler Communication

In previous versions of CCSM, MPI was used to communicate data between the models and the coupler (along the spokes in Figure 1), often with direct calls to the send and receive functions of the MPI library.

Prior to the current release, the coupler itself was not a distributed memory parallel application. It was a separate executable that used OpenMP threading in some floating-point-intensive portions of the code but ran as a

single MPI process. Communication to the coupler was achieved by a model first gathering data to its MPI root processor and then sending it to the coupler's single MPI process in a single message. Since the future development path of CCSM (*Community Climate System Model Science Plan (2004-2008)* www.ccsm.ucar.edu/management/sciplan2004-2008.pdf) points to increasing horizontal resolution (currently at about 250 kilometers) and increasing the number of physical processes simulated, this one-processor communication point was going to become more of a bottleneck.

Since all data goes through the coupler, the coupler must maintain a representation of each model's numerical grid. In a parallel coupler, a decomposition of that grid over the coupler's processors is required. With each model running on its own M processors and the coupler running on N processors, CCSM with a parallel coupler contains multiple examples of the "MxN problem." The MxN problem is the transfer of a distributed data object from a module running on M processes to another running on N processes (see http://www.cs.indiana.edu/feberta/mxn for a summary). The exact pattern of communication will change as M and N change, as they do when load balancing CCSM for different problem sizes and hardware systems, and will also be different for each model-coupler pair. A solution that was both efficient and scalable was needed. MCT provides a general solution to this problem by deriving a set of point-to-point communications that transfers the data with a minimum number of messages (Section 3.2). With MCT, the new coupler in CCSM, cpl6, is now a distributed-memory parallel application.

2.2. Intracoupler Communication

An important function of the coupler is interpolating, or mapping, data between the numerical grids of the models as data is routed through the coupler. In CCSM, interpolation is performed as a matrix-vector multiply (see Section 3.3). Each two-dimensional grid (for the atmosphere, ocean, land, or sea-ice models) is unrolled into a vector, one for the source grid and one for the destination. The matrix of mapping weights contains a row for each source element and a column for each destination. The two-dimensional grids have between 10^4 and 10^5 grid points, but most of the matrix elements are zero, so that interpolation is a sparse-matrix-vector multiply. In CCSM, the grids are fixed and regular, and the nonzero elements are computed once offline by using the SCRIP program (Jones, 1998) and read in at run time.

In the previous versions of the coupler, interpolation was trivial because all the data on the source and destination grids and the matrix elements were held in a single address space. A parallel coupler requires distributing the matrix elements accounting for the decomposition of the two grids. Given that the grids of the ocean and atmosphere model, for example, are decom-

posed arbitrarily in the coupler, there is no guarantee that all the atmosphere data on the atmosphere grid needed to complete an interpolation onto the ocean grid will be colocated on the same processor. Thus, although the matrix elements need to be distributed only at initialization, source data, which is updated each time-step, needs to be redistributed each time-step before interpolation. MCT also provides methods for these parallel communication needs.

3. Parallel Communication with MCT

The introduction of distributed-memory parallelism to the CCSM coupler created new parallel communication needs, both for data transfer between models and the coupler and for parallel data interpolation within the coupler.

Other software packages contain general solutions for communication between parallel data structures, including PAWS (Beckman *et al.*, 1998; Keahey *et al.*, 2001), CUMULVS (Geist *et al.*, 1997), MetaChaos (Ranganathan *et al.*, 1996; Edjlali *et al.*, 1997) and its successor InterComm (Lee and Sussman, 2004). There are also domain-specific or domain-inspired solutions such as Roccom (Jiao *et al.*, 2003), the Distributed Data Broker (DDB) (Drummond *et al.*, 2001), and the Flexible Modeling System (www.gfdl.noaa.gov/fms). Others have based their solution on the Common Object Request Broker Architecture (CORBA), such as MCEL (Bettencourt, 2002) or proposed extensions to CORBA such as GridCCM (Perez *et al.*, 2003) and Pardis (Keahey and Gannon, 1997). In view of the requirements above, however, these packages were not entirely suitable as the basis for a parallel coupler in CCSM. Some, such as CUMULVS and MCEL, did not provide MxN data transfer at the time this project began. Others lack a Fortran interface, introduce a major new package dependency such as PVM (Geist *et al.*, 1994), handle only certain parallel data types, or do not provide interpolation abilities.

Part of the MCT approach was to use only languages and libraries already present in CCSM: MPI and Fortran90. MCT is written entirely in Fortran90. Fortran90 supports derived types and allows grouping public and private subroutines and functions into "modules." (This paper will occasionally refer to the modules as "classes" and the subroutines as "methods" even though Fortran90 is not a true object-oriented language.) Since CCSM was already using MPI to pass data between between models and the coupler (as a single, root-to-root message), MCT provides an MPI-style double-sided message-passing model for solving the MxN problem: MCT calls replace the single MPI calls of the nonparallel cpl5 coupler at the same locations in each model's code (see Section 3.2). MPI is still used underneath MCT as the communication layer. These choices allowed MCT to provide solutions to

CCSM's needs without impacting the model's portability.

3.1. Data Storage and Decomposition

To cope with the wide variety of data types in CCSM, MCT introduces a standard data representation called the AttributeVector.[2] The AttributeVector is a Fortran90 derived type that consists of a two-dimensional array of reals and one of integers. The first index refers to the "Attribute" corresponding to the physical quantity being stored, such as temperature, wind, or humidity. The second index refers to the value of each attribute at a physical grid point. All the fields sent to or received from the coupler can each be contained in an instance of the AttributeVector. In CCSM3, the AttributeVector is the datatype exchanged between the coupler and other models, and AttributeVectors are the parallel data type used within the coupler. AttributeVectors are *locally sized*: they contain just enough space to hold the data local to a processor in a model's decomposition. Methods that operate on AttributeVector, such as communication methods equivalent to an MPI broadcast and gather, operate on all the attributes at once: the temperature, wind speed, and other values in the AttributeVector are sent in one message to their destinations.

The GlobalSegmentMap or GSMap is the MCT datatype for describing a decomposition of a numerical grid or the portion of a grid used in coupling. The datatype is *global* because it contains a description of the decomposition of the entire grid. After initialization, which is typically done collectively with each processor describing its portion of the grid, the returned data type is identical on all processors. Thus, each processor can inquire the processor id of any point in the numerical grid.

The GSMap is defined by numbering all the points in the numerical grid to be described. A very simple example of a grid with 20 numbered points distributed over two processors is shown in Figure 2. The GSMap data type values for this decomposition are as follows:

```
ngseg:   2        ! total number of segments
gsize:   20       ! total number of points
start:   1 , 11   ! value of starting point for each
                  !                           segment
length:  10, 10   ! length of each segment.
pe_loc:  0 , 1    ! MPI rank of processor containing
                  !                      each segment.
```

Using the three integer arrays `start`, `length`, and `pe_loc`, one can describe

[2]We shall use the following typographic conventions. References to class or Fortran90 module names are indicated with classname. File names, subroutine names, and other parts of source code are indicated as `subroutine`.

the decomposition of any grid, structured or unstructured, provided the grid points can be conceptually numbered sequentially. The GSMap is efficient at describing block decompositions because that type is used most often in climate modeling.

Before data is sent to the coupler, values must be copied from the model's internal datatype into the correct location in the AttributeVector. The correct location is determined by the implied relationship between an AttributeVector and the GSMap. The implied relationship between local memory in the AttributeVector and the local portion of the GSMap is shown in Figure 3 for the points on processor 1 of Figure 2. Because the data types of individual models in CCSM are varied and unknown to MCT, MCT serializes the *grid* of the model. MCT then uses the implicit relationship between the memory of the AttributeVector and the locally owned portion of the GlobalSegmentMap shown in Figure 3 to map data between processor-grid point spaces. The GSMap method GlobalToLocal translates between global values of locally owned grid points and memory indices in the AttributeVector. The MCT user has the responsibility for copying data between gridded data stored in their model's internal data types and the AttributeVector according to the relationship in Figure 3. By generalizing the decomposition of the grid instead of trying to generalize the model's unknown internal parallel data type, MCT gains great flexibility at the cost of a single data copy.

3.2. Communication Schedules and M×N Transfer

In order to send data between two models, an AttributeVector with the same number of Attributes and related to the same numerical grid *with the same gird-point numbering scheme* must exist on each side of the communication. The local size of the data may be different depending on the different decompositions. For example, the CCSM atmosphere model may decompose its grid over 64 processors while the coupler decomposes its representation of the atmosphere's grid over 16 processors. The model and coupler will then have two different GSMaps, and the local size of the AttributeVector for "atmosphere-to-coupler" data will be different, but MCT can still transfer the data so long as the atmosphere grid is numbered the same way in each model.

Given two decompositions of the same numbered numerical grid specified in two GSMaps, one can easily build a mapping between the location of one grid point on a processor to its location on another processor. The set of all these mappings forms a customized MxN routing table. This table can be used by a processor to indicate the destination or source processor for each of its data points in the alternative decomposition. In MCT, this assembled table is stored in another Fortran90 derived type called the Router.

Router initialization is a form of "handshaking," where two models learn how to exchange data with each other in parallel. The algorithm has two phases. First, the models exchange their GSMaps. This is a synchronization point between models. The received GSMap is then broadcast to each processor of that model. At that point, each processor has the GSMap for each side of the communication and can build its local Router in parallel with the other processors. This is the usual method for CCSM, where the models are each on disjoint sets of processors. MCT also provides methods to send a GSMap asynchronously and to initialize a Router if the two GSMaps are already available. Since the grids in a climate model such as CCSM are fixed, Router initialization is typically done once at startup.

After initialization, the Router can be read two ways: as both a list of *local* memory indices of the local AttributeVector and a list of MPI processor ranks to which they must be sent, and as a list of *local* memory indices to receive data from a list of processors. The Router is a two-way map from one decomposition to another, and therefore the same Router data can be used for an MxN send or receive.

Figure 4 illustrates a portion of a Router for processor 0 in a system with two components spread over six processors. On the first two processors, the grid has been given a one-dimensional decomposition, while on the the last four the same grid has been given a two-dimensional decomposition. The Router allows processor 0 to know that during a send, it must send four points to processor 2 and during a receive it will receive four points from processor 2. Processor 0's Router also contains information about the shared points on processor 3 (not shown).

The Router also contains an integer ID uniquely identifying the model intended to be the partner in any communication. This integer ID is used to look up the other model's processors and their global MPI rank using a lookup table created as part of MCT initialization.

Once a Router has been initialized and an AttributeVector has been filled with new data to be sent, the M×N transfer is accomplished with a matched pair of calls analogous to MPI message passing:

```
MCT_Send(Model1_AttributeVector, Model1_Router)
```

and on the receive side:

```
MCT_Recv(Model2_AttributeVector, Model2_Router).
```

In MCT, the AttributeVector takes the place of the buffer address in MPI communication routines. The Router, "pointing" to another model and its decomposition of the grid, takes the place of the destination/source MPI rank.

To avoid latency costs, `MCT_send` packs all attribute values (temperature, wind, etc.) into one message for a given set of grid points destined for a processor. The `MCT_send` blocks until the underlying MPI sends are complete. MCT also provides nonblocking versions, `MCT_Isend` and `MCT_Irecv`, for asynchronous MxN data transfer.

If the Router indicates a given processor must send or receive a message from/to more than one processor, these messages are posted together through successive calls to `MPI_Isend` and `MPI_Irecv` (the blocking `MCT_send` includes a call to `MPI_Waitall`). Posting several nonblocking calls at once can, in principle, be faster than explicity scheduling matching pairs of sends and receives (Gropp *et al.*, 1999).

Although two sides of a communication with a Router must reference the same numbering scheme for a grid, the two grids may differ on the total number of points. This flexibility is useful in CCSM where the land model uses the same global grid as the atmosphere but allocates storage only for land points. As long as those points have the same index in their respective GSMaps, a Router can still be constructed and MCT can exchange gridded data between the two models. A similar situation occurs in the ocean model, which has a grid defined over the entire globe but allocates memory only for the ocean points.

3.3. Communication Support for Interpolation

In previous versions of the CCSM coupler, interpolation of data was trivial because the single-node coupler held all data points for two grids and their interpolation weights in the same memory. The introduction of distributed-memory parallelism to the coupler creates new requirements for parallel data transfer within the coupler.

As discussed in Section 2.2, interpolation is performed as a sparse-matrix vector multiply. For example, interpolating data from the atmosphere model's grid to ocean model's grid is

$$\overbrace{\begin{pmatrix} o_1 & o_2 & \dots & o_m \end{pmatrix}}^{m \text{ ocean grid points}} = \overbrace{\begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix}}^{n \text{ atmosphere grid points}} \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1m} \\ w_{21} & w_{22} & \dots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nm} \end{pmatrix}.$$

The atmosphere vector $\mathbf{A}$ is the input, or source, while the ocean vector $\mathbf{O}$ is the output, or destination. The matrix $\mathbf{W}$ is sparse, and in CCSM only the nonzero elements and their locations are stored. MCT provides a datatype called the SparseMatrix to store the weights and their row and column indices using an AttributeVector. MCT also provides a matrix-AttributeVector mul-

tiply method that takes advantage of the storage order in the AttributeVector. All attributes for a grid point are interpolated before moving to the next point, thereby allowing cache reuse of the attribute data (Larson *et al.*, 2005).

In the parallel coupler, both the source and destination vectors have a decomposition that is set independently of any consideration for interpolation. Some additional communication is required to get the weights, source, and destination points all on the same processor. MCT's tools for doing this are based on the GSMap and AttributeVector. First, just as the GSMap defined the decomposition of a numerical grid based on a sequential numbering of its points, the vectors on either side of the matrix equation above are also serializations of these numerical grids. MCT assumes these two serializations are the same: the row-number in the matrix corresponds to the index value of the grid numbering used in the GSMap for the source grid's decomposition, and the column-number is the index value of the destination grid. This relationship is crucial to MCT's support for parallel interpolation.

3.3.1) Initialization of Parallel Interpolation

The first task is to distribute the nonzero interpolation weights from the matrix $\mathbf{W}$, which in CCSM are calculated offline by using SCRIP and read in from files. CCSM does not yet use parallel I/O, and in most cases input data is read from the root node and scattered. MCT provides methods to derive a decomposition for the sparse-matrix elements and scatter weights into this decomposition. The source and destination vector decompositions imply two possible decompositions of the matrix elements: by the decomposition of the rows (source) or by the columns (destination). Both decompositions are equally valid, and MCT provides methods for completing the sparse matrix multiply in each case.

In a destination-based decomposition, entire columns of nonzero weights are scattered to the coupler processors according to the destination vector's decomposition, as shown in Figure 5a. MCT's AttributeVector_Scatter is used for this operation. This scatter method takes an AttributeVector on one processor that contains values for all the points of a grid and sends them to locally-sized AttributeVectors on each processor according to the decomposition described in the destination vector's GSMap. This is analogous to `MPI_ScatterV`. In this case, however, the GSMap is used to determine the destination instead of the MPI rank. This scatter is performed once at startup in CCSM.

The additional communication required to make the interpolation data-local in this case brings the necessary source points to each processor. The necessary source points are the elements from $\mathbf{A}$ needed to calculate each

point in **O** local to a processor. These points can be determined from the nonzero elements of the scattered sparse matrix. Since MCT stores the nonzero matrix element along with its row and column number, the column-scattered SparseMatrix can be examined to collect which corresponding row points are needed on that processor as shown in Figure 5b. This information, a set of row numbers that are also grid-point numbers and a processor rank, is sufficient to build a second GSMap for the source grid. The MCT routine SparseMatrixToXGlobalSegMap derives a GSMap for **A** (or **X**) from the column-decomposed **W**. Constructing this GSMap is also a one-time initialization cost in the model. Note that this derived GSMap may contain duplicated points as shown in Figure 5b because the same physical source point may be needed for multiple destination points that may be on different processors. The MCT GSMap allows this situation, and MCT Routers and transfer methods can move data between duplicated and distinct decompositions. These structures and methods allow MCT to position exactly the points neeeded for interpolation with a minimum in both the number and size of messages.

3.3.2) Parallel Interpolation

Before each interpolation, the data in **A** must be moved from its coupler-defined default decomposition to the sparse-matrix derived decomposition. MCT provides the Rearranger class for this kind of data transfer. Rearranger transfers data within a group of processors between two different decompositions. This approach is in contrast to `MCT_Send` and `MCT_Recv`, which move data between disjoint sets of processors. Rearranger takes data in one AttributeVector with an associated GSMap and rearranges it into another AttributeVector associated with a different GSMap all within a group of processors.

In the CCSM coupler, two copies of **A** are necessary because the coupler performs additional computations with atmosphere data where duplicate points, such as those in the sparse-matrix derived decomposition, would be problematic. If this were not the case, the atmosphere could send data directly to the sparse-matrix derived decomposition using a Router, and the rearrange step could be skipped.

A complete atmosphere-ocean parallel interpolation with matrix elements scattered according to the ocean decomposition proceeds as follows. First, the atmosphere data is sent to the coupler by using the MxN `MCT_Send` and is stored in an AttributeVector. A call to Rearrange rearranges the atmosphere data into a second AttributeVector with the interpolation-ready decomposition derived from the scattered matrix elements. This AttributeVector is the input in a call to the data-local MCT AttributeVector-SparseMatrix multiply

routine. This call interpolates data for all attributes and stores the output in an ocean-resolution AttributeVector, which can then be used in the coupler for additional computation or passed to the ocean model.

Using the same parts of MCT, the interpolation can be done with the matrix weights distributed according to the the *source* vector's decomposition. In this case, the source data is left in place, and a portion of the interpolation is performed on each processor. An intermediate AttributeVector and GSMap are created to hold the output, which now may be only a portion of the final destination point's value. An optional argument to the Rearrange call adds the partial sums while forming the final output AttributeVector.

4. Results

We investigated the performance of MCT routines involved in parallel communication and interpolation. The platform used was a Linux cluster called "Jazz," located at Argonne National Laboratory. Jazz contains 350 nodes each with a single 2.4 GHz Pentium Xeon processor and either 2 GB or 1 GB of RAM. The processors are connected via Myrinet 2000. Portland Group Fortran was used for compilation.

For the timings below, we used a test application that exercises MCT datatypes in the context of a real climate simulation consisting of an atmosphere model, an ocean model, and a coupler. The atmosphere model uses the same horizontal grid as CCSM3, which covers the globe with 64 latitudes and 128 longitudes. The ocean model also uses the same grid as the CCSM3 ocean model, which contains 384 latitude and 320 longitude points covering the globe. As in CCSM, the mapping weights were calculated offline by using SCRIP. The grid-point numbering scheme for both grids is as shown in Figure 2; the southernmost point on the western boundary is point number 1, and subsequent points are numbered from west to east and south to north. As in CCSM3, the atmosphere, ocean, and coupler run on distinct sets of processors, and the coupler contains a decomposition and storage for data on both the atmosphere and ocean grids. In most cases, the number of coupler processors is varied from the 2 or 8 processors used in the released version of CCSM3 to 32 processors.

4.1. Router Initialization

Since the Router plays such an important role in MCT parallel communications, we timed the initialization of a Router under several possible decompositions. The GSMap description of a decomposition can differ based on both the number of processors and the "strategy." A typical decomposition in the atmosphere can be imagined by placing the grid shown in Figure 2 over a map of the world. The decomposition shown in Figure 2 is called "latitu-

dinal" (or by-row) strategy because it divides the world into latitude bands. With the grid numbering shown in Figure 2, latitudinal decomposition would result in one segment per processor. The orthogonal decomposition to Figure 2 is "longitudinal" (or by-column) and, with the same grid numbering scheme, would yield one segment per latitude per processor. A combination of those two strategies, dividing the grid into latitude-longitude squares, is called "checkerboard" (row-column), or Cartesian (Figure 4, right-hand side), and would yield one segment per latitude owned by a processor. A 2x2 checkerboard decomposition of the 64x128 atmosphere grid would yield a total of 128 segments, 32 for each processor. The GSMap description of the decomposition differs for each strategy and each processor count.

Since the Router is constructed between two GSMaps, the full performance space is four dimensional, two for the decomposition strategy and two for the processor count. A small portion of that space is considered in Figure 6. Figure 6 shows time to construct a Router in the coupler between the coupler's decomposition (local GSMap) and the atmosphere's decomposition (remote GSMap) of the atmosphere grid. This Router is used by the coupler to send or receive data to and from the atmosphere. The timing considers only the calculation of the Router after the GSMaps have been exchanged. The total time for 10 calls to the Router initialization routine was measured and the maximum time over all processors plotted.

Figure 6 shows the combined effect of changes to the two GSMaps involved in a Router initialization on the time to calculate the Router. For the remote GSMap received from the atmosphere, the number of atmosphere processors is fixed at 16, but the strategy varies between latitudinal (circle), longitudinal (square), and checkerboard (diamond). The grid point numbering scheme is as shown in Figure 2. For 16 atmosphere processors the total number of segments for latitudinal is 16, for longitudinal it is 1024 (64×16), and for the checkerboard decomposition it is 256. For the coupler's local GSMap, the strategy is fixed at latitudinal, but the number of processors, and thus the total number of segments, varies from 1 to 32.

The Router initialization algorithm searches first over all segments in the local GSMap looking for the ones it owns. For each segment it owns, it then loops over all the segments of the remote map looking for a match. Scaling is nearly linear for all decomposition combinations between 2 and 8 processors because the local segment length decreases as processors are added. Some superlinear scaling occurs from 8 to 16 processors for the latitudinal decomposition because at 16 processors the coupler and atmosphere GSMaps are exactly the same and it is easy to find matching segments. The longitudinal remote GSMap is the slowest because this decomposition is completely orthogonal to both the grid-point numbering scheme used for the grid and

the latitudinal decomposition used in the coupler that is building the Router. The combination of 32 possible segments in the local decomposition and 1,024 segments in the remote decomposition leads to the longest initialization time. This suggests that, whenever possible, one should choose a grid-point numbering scheme and decomposition strategy that minimizes the total number of segments. Also, decompositions with very different strategies lead to increased initialization times.

4.2. M×N Data Transfer

The performance of `MCT_Send` in the atmosphere component of the MCT test program is shown in Figure 7. In our test model, the atmosphere AttributeVector has 17 total fields to send, a typical number for a climate model like CCSM3. The atmosphere grid contains 8,192 points. Assuming 64-bit real number representation, roughly a megabyte of information must be sent out of the atmosphere each coupling time step (every simulated hour in CCSM3). The focus on the send, rather than the receive, operation is arbitrary; similar results are obtained if the receive is measured (not shown).

As in the Router initialization, the performance space is four dimensional. The performance of an M×N data transfer between components on disjoint sets of processors will depend on the decomposition strategy and the number of processors on each side of the transfer. Both the per-message size and the number of messages can change if the decomposition or number of processors changes. A portion of this performance space is examined in this section.

Figure 7a shows the cumulative time for 100 calls to `MCT_Send` from an atmosphere on 16 processors to a coupler with a varying number of receiving processors. A barrier was placed before the send to decrease the effects from load imbalance in the rest of the MCT test application. The minimum over all processors is shown. The coupler decomposition strategy is fixed at latitudinal, while the atmosphere uses two decompositions, latitudinal (circles) and checkerboard (squares).

For the processor counts considered, as the number of receiving processors increases, the total time decreases. In the latitudinal case, the decomposition strategy is the same as in the coupler, and performance is more predictable than if the send is from a checkerboard decomposition. As in the Router initialization (Section 4.1), when the decompositions are very different, there is an increased cost.

Figure 7b show the results for the same measurement except that the atmosphere is fixed at 32 instead of 16 processors. For this case, communication time again decreases as the number of processors on the receive side increases. When comparing the two latitudinal curves, there is a minimum when the number of processors and the decomposition strategies match ex-

actly (16 processors for Fig. 7a and 32 for Fig. 7b). The 32-processor send measurement is faster at all receive processor counts than is the 16 processor send, which shows that MCT can realize some performance gains when processors are added to either side of an M×N communication.

The departure of the curves from ideal speedup is caused by accumulated latency costs. While the size of each message decreases as the number of processors increase, there are also more messages to send. The maximum number of messages will equal the product of the number of processors on each side of the communication. This maximum is reached when the decomposition strategies on each side are orthogonal. The checkerboard decomposition contains some orthogonality to the receiver's latitudinal decomposition, and this probably accounts for the different shapes of the checkerboard and latitudinal curves in Figure 7. At the extreme case of very high numbers of processors and very different decompositions, the number of messages could saturate the available bandwidth. In that case, it may be necessary for the coupled model programmer to adjust the decomposition on one side to match the other. If the decompositions, and M and N, are identical, only one message per processor pair is required.

The latency cost of sending a message was anticipated in the design of MCT's parallel transfer routines. Instead of sending one message for each attribute in an AttributeVector, MCT_Send first copies all the grid points and fields destined for a processor from the AttributeVector into an internal buffer. This copy cost was also measured and was between 1 and 2 orders of magnitude less than the transmission costs shown in Figure 7 at all configurations tested. While the copy cost increased as the number of messages increased, it was still an order of magnitude less than the smallest transmission time shown on Figure 7.

4.3. Parallel Interpolation

The performance of MCT's parallel interpolation functions is presented in Figures 8 and 9. Section 3.3 explained how parallel interpolation is divided into two phases: a rearrangement of data for the interpolation and the matrix-multiply doing the interpolation. This section considers the cost of both phases and how they vary for number of processors and decompositions.

While the previous measurements required two components, interpolation is performed entirely in one component, in this case the coupler of the MCT test application. As in the previous sections, the decomposition strategy is still an important factor in performance. Figure 8 shows the performance of the parallel interpolation routine when transforming data from the atmosphere to the ocean grid subject to changes in the decomposition of the atmosphere grid and the number of processors in the coupler. The destina-

tion ocean grid decomposition is fixed at latitudinal in all cases. The solid line shows the total cost of the compute-only part of the interpolation for 10 calls to the MCT parallel interpolation subroutine for three different decompositions of the source atmosphere grid. All three decompositions have the same cost because, after rearrangement, the computation is identical. The maximum over all processors is plotted in Figure 8 and shows some departure from ideal scaling because the work load is distributed according to the ocean decomposition and not according to the number of matrix weights and therefore there is a load imbalance. The average time over all processors, shown for the latitudinal decomposition only in Figure 8 (other decompositions are similar), scales linearly.

The cost of the communication of data between the atmosphere's user determined decomposition and the matrix-multiply derived decomposition within the coupler is plotted by using dashed lines in Figure 8. Some of the behavior seen in previous sections appears again for this measurement. The latitudinal decomposition has the smallest communication cost because it matches the sparse matrix-derived decomposition, which in turn was based on the latitudinal decomposition of the destination ocean grid. Rearranging data from a longitudinal decomposition to the sparse-matrix derived decomposition has the highest communication costs, and the checkerboard decomposition is close to the cost of the longitudinal. As for the send performance (Section 4.2), the difference in communication costs is caused by the difference between the decompositions.

The cost of communication is around an order of magnitude less than the interpolation at nearly all processor counts. The ratio of computation to communication decreases as the processor count increases because the amount of work remains the same while the total number of messages—and their associated latency costs—increases. Like the Router and MCT_Send, the Rearranger keeps communication costs to a minimum by sending only the points necessary to complete the interpolation and sending all attributes in a single message. The Rearranger performs an in-memory copy for those points located on the same processor in the two different decompositions so processors can avoid sending MPI messages to themselves.

MCT allows the order of communication and computation to be reversed (Section 3.3) so that communication can be performed on the coarser grid. The reason for including this capability is shown in Figure 9. Figure 9 shows the cost of the communication and computation phases of interpolating data from the ocean to the atmosphere grid for two different orderings of communication and computation. "Multiply then Rearrange" means that first interpolation is performed by using whatever ocean points are available in the user-defined latitudinal decomposition (solid line with circles) and then

partial sums are reduced to their resident process IDs on the destination grid decomposition (dashed line with circles). The total cost of the parallel interpolation is the sum of the two measurements. As above, the communication is an order of magnitude smaller than the computation for all processor counts.

For the "Rearrange then Multiply" case, ocean data is first rearranged into a sparse-matrix derived decomposition, and then the interpolation is performed. The costs of communication and computation in this case are comparable, and the total cost of parallel interpolation is nearly double the "Multiply then Rearrange" case. The reason rests with the ratio of the number of points in the atmosphere and ocean grids, which is nearly a factor of 10. In "Rearrange then Multiply" a large amount of ocean data must be communicated between coupler processors, while in "Multiply then Rearrange" a smaller amount of atmosphere data needs to be communicated. For the ocean-to-atmosphere interpolation, there is a performance advantage to performing a partial interpolation of data from the ocean to the atmosphere grid and then rearranging and summing the partial results to form the final atmosphere resolution data. This performance advantage was first noted in the Parallel Climate Model (PCM) (Bettge *et al.*, 2001), and PCM's unique coupler contained specially written routines to perform a similar operation (Tony Craig, personal communication, 2002). MCT's general routines for parallel interpolation matches or exceeds the performance of PCM's custom-written routines (Ong *et al.*, 2002).

5. Conclusions

MCT provides efficient, scalable datatypes and subroutines for the parallel communication problems found in multiphysics parallel coupled models such as the Community Climate System Model. If data is stored in a AttributeVector, the numerical grid is numbered sequentially, and the decomposition is described with a GlobalSegMap, MCT contains several methods to accomplish parallel data transfers in inter- and intramodel communication.

MCT takes a different approach from other M×N communication solutions by serializing the numerical grid used in the model instead of the memory space of the model's parallel data type. The implicit relationship between the grid's decomposition and the AttributeVector allows MCT to handle any decomposition of any grid at the cost of a data copy and some effort by the programmer to determine the mapping between a model's internal data structures and the AttributeVector.

MCT provides several methods to scatter interpolation weights stored in files for CCSM and read in on the root node of CCSM's coupler. If future versions of CCSM perform parallel I/O or calculate interpolation weights

in parallel, the results could be stored in the same SparseMatrix, and the rest of the parallel interpolation methods would still be applicable. Because the specific choice of interpolation method affects the scientific output of a model, MCT currently leaves the calculation of the interpolation weights to the application developer, but future versions will include some support for calculating these weights from an MCT GeneralGrid.

MCT performance was measured on a Linux cluster, and good scaling was observed for the tested parallel methods. Performance is better if two decompositions of the same grid have a similar strategy. This result suggests that the designers of a coupled system should be aware of the decompositions used in the component models and incorporate this information in their coupling strategy to maximize coupled system performance. Data transfer costs of the parallel interpolation routines can grow if the problem size remains fixed while the number of processors increases. This cost can be avoided if the component performing the mapping does not need the source data for other purposes. MCT allows communication on either side of the interpolation calculation, which can lower the overall time required when interpolating from a high- to low-resolution grid.

The overhead of MCT in CCSM is difficult to measure without examining the full performance of this complex parallel application. A performance study of CCSM is beyond the scope of this paper. Experience with MCT within CCSM3, however, shows that the overhead of copying in and out of AttributeVectors is less than 1% of the total time to simulate a day of climate interaction. This low overhead is because coupling is relatively infrequent and the rest of the model's activity, such as radiation calculations and fluid dynamics, is very time consuming. Other costs of MCT do not count as overhead because they are performing critical functions of the coupler. The copy costs can be avoided if an application adopts AttributeVectors as their internal datatype. The total cost of using MCT's parallel interpolation algorithms was as good as or better than hand-written versions for one coupled model (Ong *et al.*, 2002).

Besides the lack of a common datatype, the biggest barrier to constructing more coupled systems in the earth sciences and in other fields is that candidate submodels are seldom developed with coupling in mind. The CCSM architecture is a way to build coupled systems with such applications, but the cost is multiple executables and less flexibility in integration schemes (concurrent vs. sequential). The most significant benefit of the Earth System Modeling Framework (Hill *et al.*, 2004) effort is that it is encouraging many modeling groups to refactor their codes for coupling. These groups are cleanly separating initialization and runtime methods and data structures. Once this refactoring is complete, it will be straightforward to construct new

coupled models using libraries currently in operational use such as FMS, MCEL, cpl6, and MCT.

Although MCT was created to solve the problems of a parallel coupler for CCSM, MCT contains no earth-science-specific assumptions and could be used to couple any two models that use a numerical grid. MCT also does not require the use of a coupler and can be used to implement direct inter-component coupling. An MCT_Send can be called directly from an atmosphere to an ocean model provided the ocean model has a decomposition and AttributeVector for the atmosphere data. MCT's parallel methods can also be used for sequentially coupled systems and mixed sequential-concurrent systems such as the Fast Ocean Atmosphere Model (Jacob *et al.*, 2001).

Because MCT is based on MPI and follows the MPI programming model, MCT can support Grid (Foster and Kesselman, 1998) computing by linking to a Grid-enabled version of MPI such as MPICH-G2 (Karonis *et al.*, 2003). A Grid-enabled version of MCT was used to couple a regional-scale coupled ocean-atmosphere system across a Grid (Dan Schaffer, personal communication, 2003). Future versions of MCT will include some OpenMP directives for the computationally intensive parts of the code such as the matrix-multiply routines and data transfer schemes, which can account for "masking" of unnecessary grid points.

The robustness of MCT's data transfer routines has already been demonstrated in a real-world application. Over 10,000 years of climate simulation have been performed with CCSM3 (Lawrence Buja, personal comunication, 2004). All the data transfer between models and interpolation in the coupler used MCT's M×N data transfer and parallel interpolation routines.

MCT is available at http://www.mcs.anl.gov/mct.

Acknowledgments We thank Tony Craig, Brian Kauffman, Tom Bettge, John Michalakes, Jace Mogill, and Ian Foster for valuable discussions during development of MCT. We also thank Jace Mogill for his improvements to the algorithm for Router initialization. We thank Michael Tobis for providing valuable comments on an early version of this paper. This work was supported by the Climate Change Research Division subprogram of the Office of Biological & Environmental Research, Office of Science, U.S. Department of Energy through the Climate Change Prediction Program (CCPP), the Accelerated Climate Prediction Initiative (ACPI-*Avant Garde*), and the Scientific Discovery through Advanced Computing (SciDAC) Program, under Contract W-31-109-ENG-38. We gratefully acknowledge use of "Jazz," a 350-node computing cluster operated by the Mathematics and Computer Science Division at Argonne National Laboratory as part of its Laboratory Computing Resource Center.

REFERENCES

Beckman, P. H., P. K. Fasel, and W. F. Humphrey, 1998: Efficient Coupling of Parallel Applcations Using PAWS. In *Proc. 7th IEEE International Symposium on High Performance Distributed Computation.*

Bettencourt, M. T., 2002: Distributed Model Coupling Framework. In *Proc. 11th IEEE Symposium on High Performance Distributed Computing*, pp. 284–290.

Bettge, T., A. Craig, R. James, V. Wayland, and G. Strand, 2001: The DOE Parallel Climate Model (PCM): The Computational Highway and Backroads. In V. N. Alexandrov, J. J. Dongarra,, and C. J. K. Tan (Eds.), *Proc. International Conference on Computational Science (ICCS) 2001*, Volume 2073 of *Lecture Notes in Computer Science*, Berlin, pp. 148–156. Springer-Verlag.

Collins, W. D., M. Blackmon, C. Bitz, G. Bonan, C. Bretherton, J. A. Carton, P. Chang, S. Doney, J. J. Hack, J. T. Kiehl, T. Henderson, W. G. Large, D. McKenna, B. D. Santner, and R. D. Smith, 2005: The Community Climate System Model: CCSM3. *J. Climate,*, to be submitted.

Craig, A. P., R. L. Jacob, B. G. Kauffman, T. Bettge, J. Larson, E. Ong, C. Ding, and H. He, 2005: Cpl6: The New Extensible High-Performance Parallel Coupler for the Community Climate System Model. *Int. J. High Perf. Comp. App.,*, this issue.

Drummond, L. A., J. Demmel, C. R. Mechose, H. Robinson, K. Sklower, and J. A. Spahr, 2001: A Data Broker for Distirbuted Computing Environments. In V. N. Alexandrov, J. J. Dongarra,, and C. J. K. Tan (Eds.), *Proc. 2001 International Conference on Computational Science*, pp. 31–40. Springer-Verlag.

Edjlali, G., A. Sussman, and J. Saltz, 1997: Interoperability of Data-Parallel Runtime Libraries. In *International Parallel Processing Symposium*, Geneva, Switzerland. IEEE Computer Society Press.

Foster, I., and C. Kesselman, 1998: *The GRID: Blueprint for a New Computing Infrastructure.* Morgan-Kaufmann.

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

Geist, G. A., J. A. Kohl, and P. M. Papadopoulos, 1997: CUMULVS: Providing Fault Tolerance, Visualization and Steering of Parallel Applications. *Int. J. High Perf. Comp. App.,* **11**(3), 224–236.

Gropp, W., E. Lusk, and A. Skjellum, 1999: *Using MPI: Portable Parallel Programmin with the Message-Passing Interface, second edition.* MIT Press.

Hill, C., C. DeLuca, V. Balaji, M. Suarez, A. da Silva, and the ESMF Joint Specification Team, 2004: The Architecture of the Earth System Modeling Framework. *Comp. in Science and Engineering,* **6**, 12–28.

Jacob, R., C. Schafer, I. Foster, M. Tobis, and J. Anderson, 2001: Computational Design and Performance of the Fast Ocean Atmosphere Model. In V. N. Alexandrov, J. J. Dongarra,, and C. J. K. Tan (Eds.), *Proc. 2001 International Conference on Computational Science*, pp. 175–184. Springer-Verlag.

Jiao, X., M. T. Campbell, and M. T. Heath, 2003: Roccom: An Object-Oriented, Data Centric Software Integration Framework for Multiphysics Simulations. In *Proc. of the 17th Annual ACM International Conference on Supercomputing.*

Jones, P. W., 1998: A User's Guide for SCRIP: A Spherical Coordinate Remapping and Interpolation Package. , Los Alamos National Laboratory, Los Alamos, NM.

Karonis, N., B. Toonen, and I. Foster, 2003: MPICH-G2: A Grid-Enabled Implementation of the Message Passing Interface. *J. Parallel and Distributed Comp.,* **63**(5), 551–563.

Keahey, K., P. Fasel, and S. Mniszewski, 2001: PAWS: Collective Interactions and Data Transfers. In *Proc. High Performance Distributed Computing Conference*, San Francisco, CA.

Keahey, K., and D. Gannon, 1997: PARDIS: A Parallel Approach to CORBA. In *Proc. High Performance Distributed Computing Conference*, Portland, OR, pp. 31–39.

Larson, J., R. Jacob, and E. Ong, 2005: The Model Coupling Toolkit: A New Fortran90 Toolkit for Building Multi-Physics Parallel Coupled Models. *Int. J. High Perf. Comp. App.*,, this issue.

Lee, J., and A. Sussman, 2004: Efficient Communication between Parallel Programs with InterComm. CS-TR-4557 and UMIACS-TR-2004-04, University of Maryland, Department of Computer Science and UMIACS.

Message Passing Interface Forum, 1994: MPI: Message-Passing Interface Standard. *Int. J. Supercomputer App. and High Perf. Comp.,* **8**(3/4), 159–416.

Ong, E., J. Larson, and R. Jacob, 2002: A Real Application of the Model Coupling Toolkit. In C. J. K. Tan, J. J. Dongarra, A. G. Hoekstra,, and P. M. A. Sloot (Eds.), *Proc. 2002 International Conference on Computational Science*, Volume 2330 of *Lecture Notes in Computer Science*, Berlin, pp. 748–757. Springer-Verlag.

Perez, C., T. Priol, and A. Ribes, 2003: A Parallel COBRA Component Model for Numerical Code Coupling. *Int. J. High Perf. Comp. App.,* **17**(4), 417–429.

Ranganathan, M., A. Acharya, G. Edjlali, A. Sussman, and J. Saltz, 1996: Runtime Coupling of Data-Parallel Programs. In *Proc. 1996 International Conference on Supercomputing*, Philadelphia, PA.

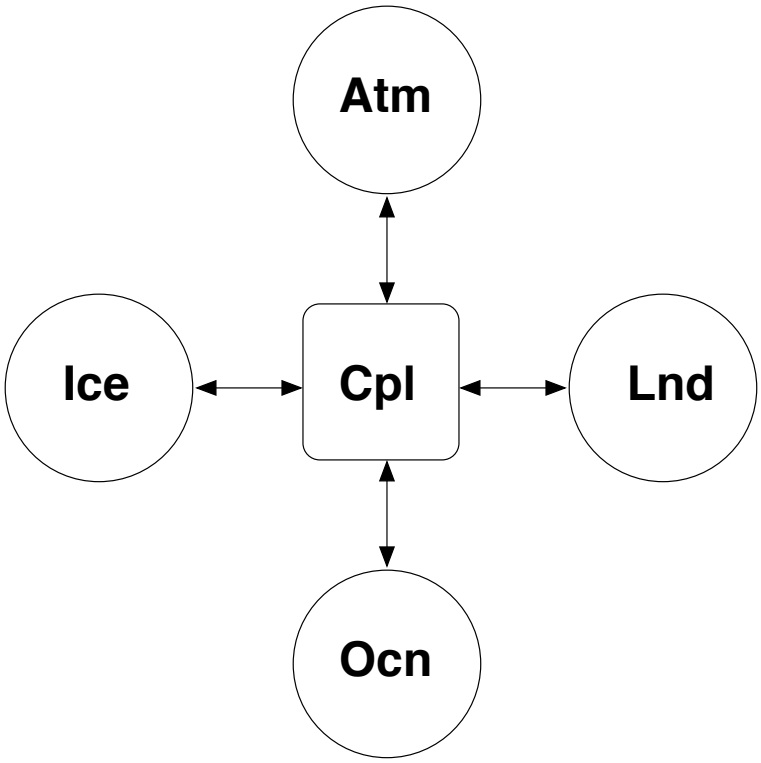

Figure 1: The "hub-and-spokes" execution model of the Community Climate System Model. CCSM contains 5 separate executables: an atmosphere model (Atm), ocean model (Ocn), land model (Lnd), sea-ice model (Ice), and a coupler (Cpl).

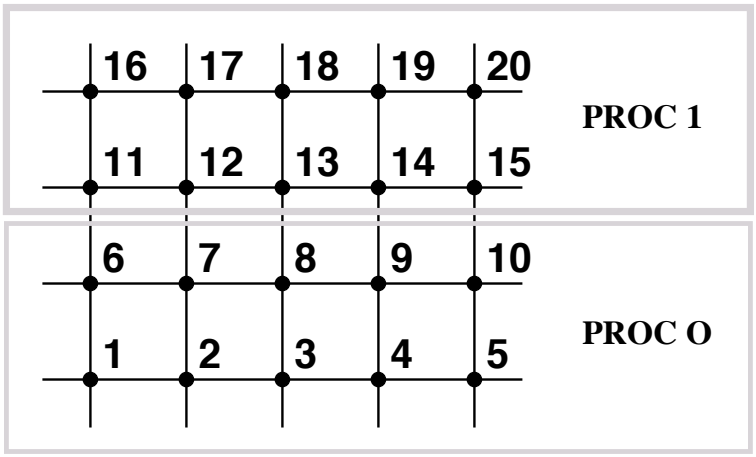

Figure 2: 2-rows x 1-column decomposition of a numbered grid.

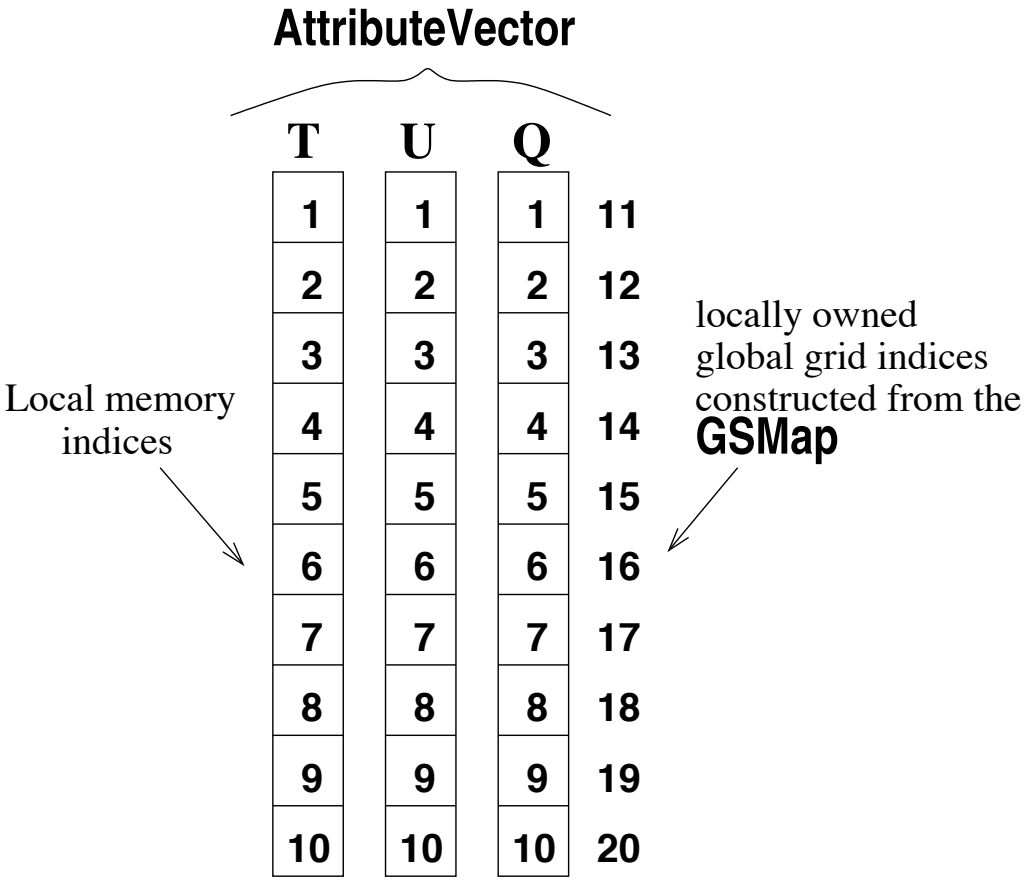

Figure 3: Relationship between local memory indices in an AttributeVector and the grid point numbering contained in the GSMap for the points owned by processor 1 in Fig. 2. In this example, the AVect is being used to store three variables: temperature (T), east-west velocity (U), and specific humidity (Q).

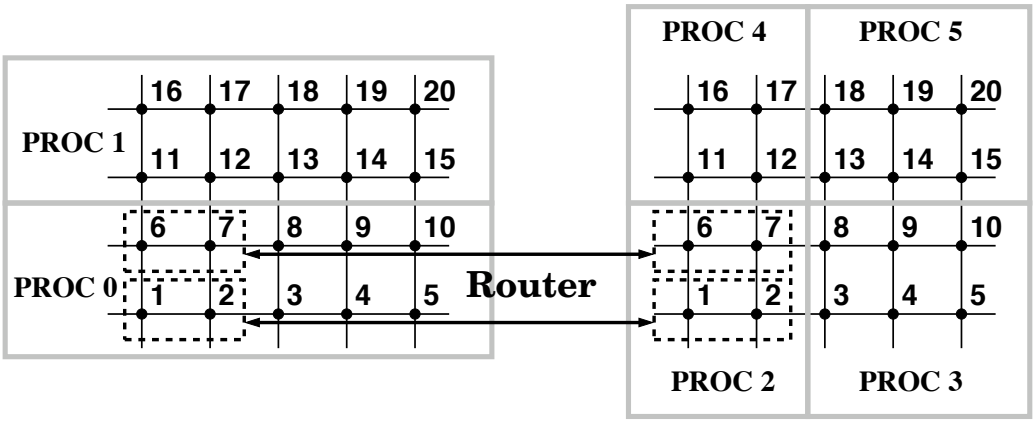

Figure 4: Schematic showing how an MCT Router maps points between two decompositions of the same grid.

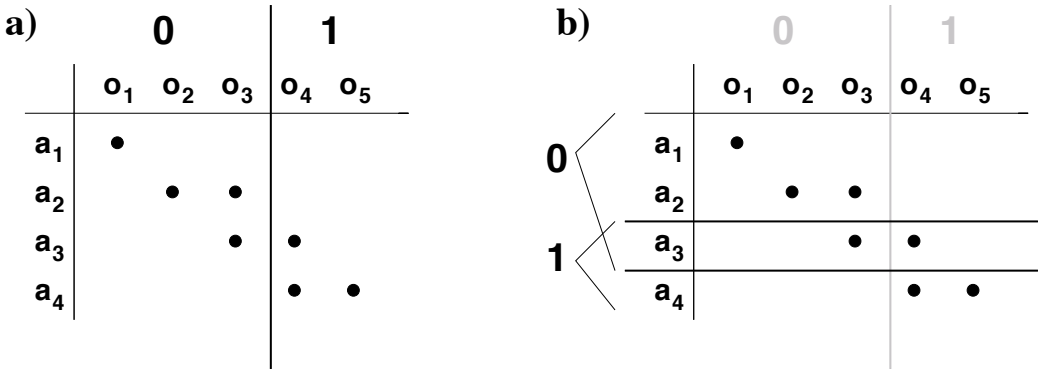

Figure 5: Steps in determining the decompositions involved in parallel interpolation.

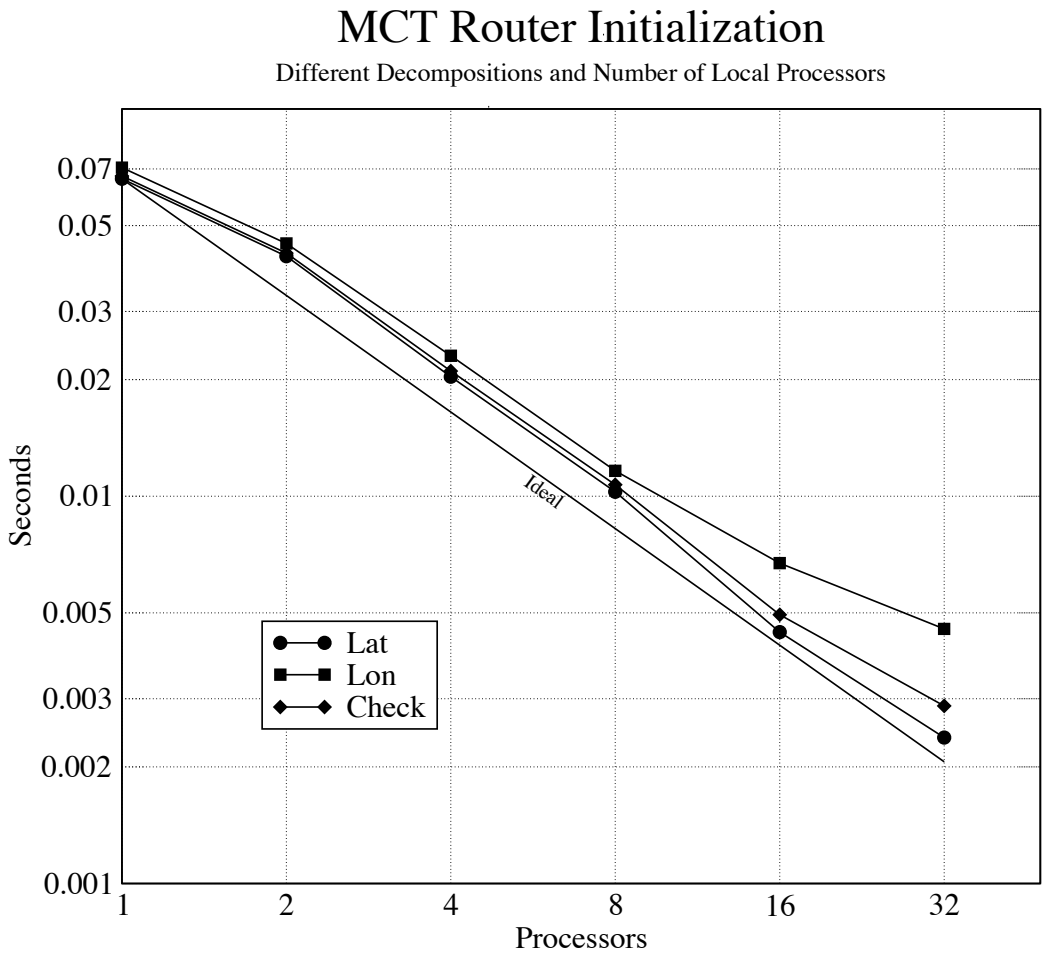

Figure 6: Total time for 10 calls to the MCT Router initialization routine for different pairs of GlobalSegMaps.

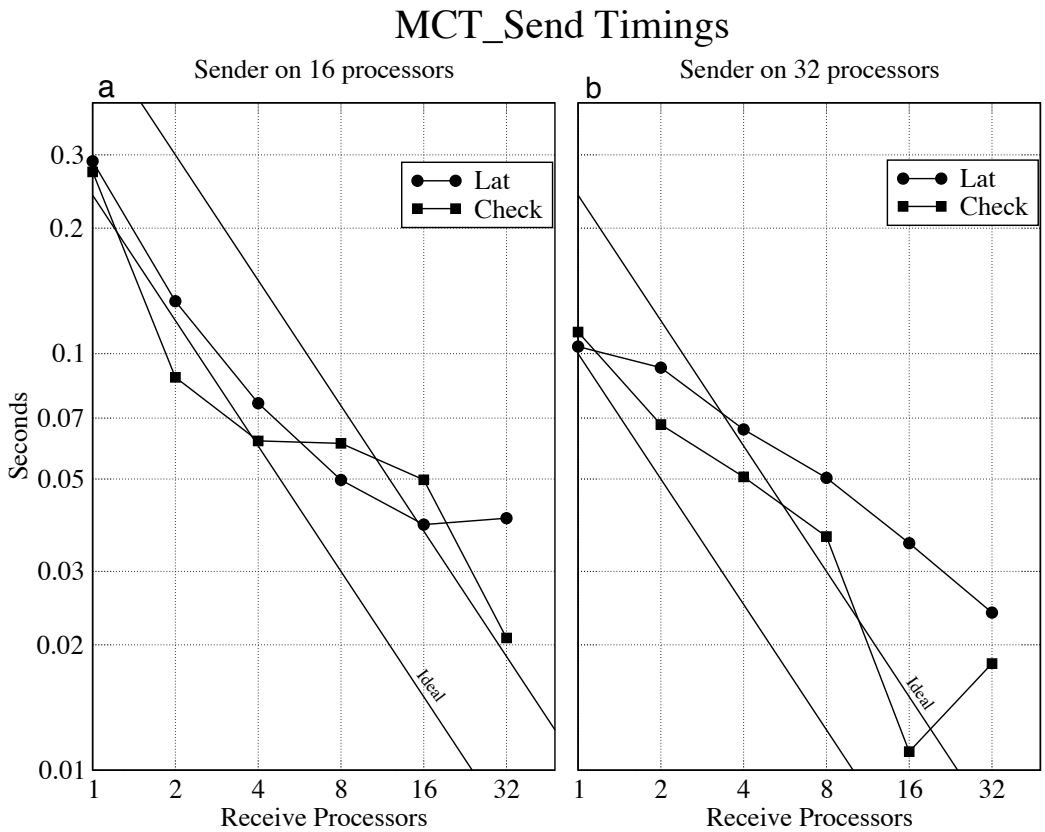

Figure 7: Total time for 100 M×N sends of data from one model to another. The Sender is given two different decompositions and either 16 (a) or 32 (b) processors. The Receiver decomposition is fixed (latitudinal), but the processor count varies from 1 to 32.

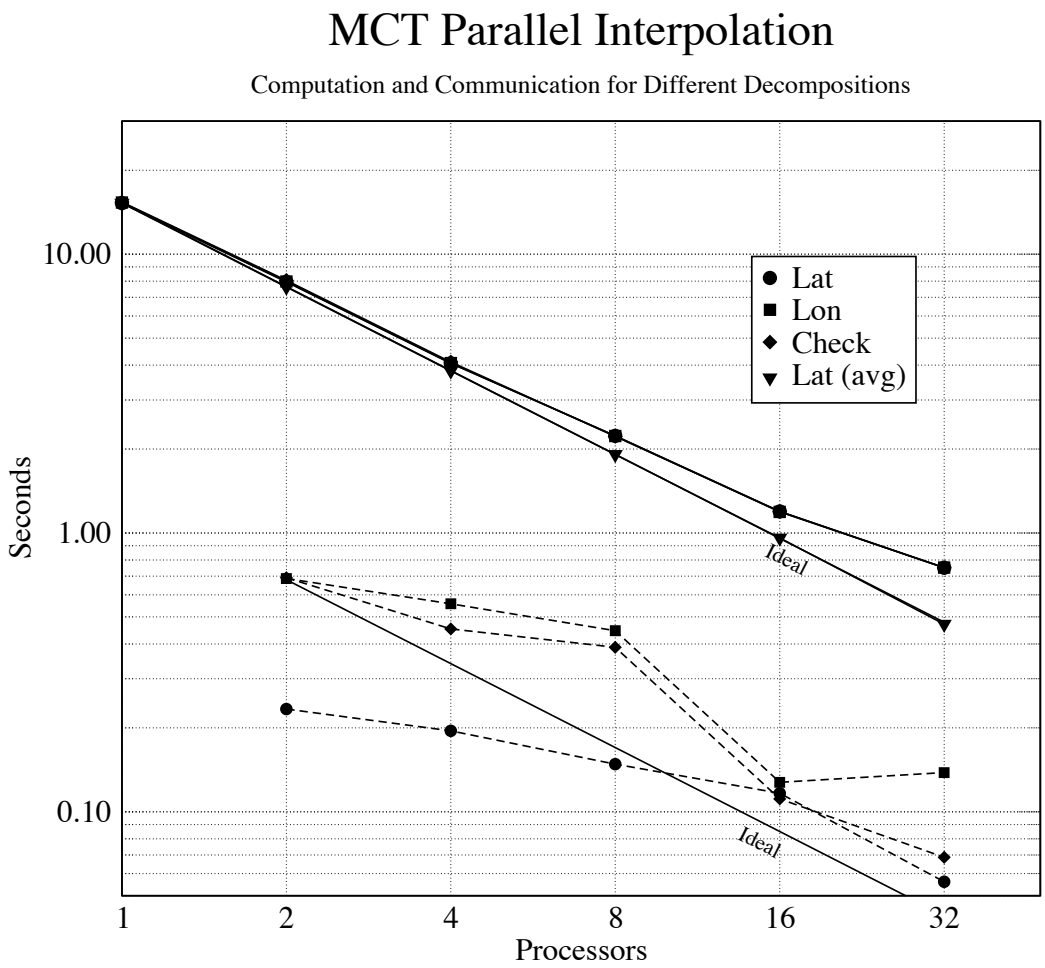

Figure 8: Total time for 10 calls to MCT parallel interpolation routine with different decompositions of the source grid (see text). The destination grid decomposition is fixed at latitudinal. Source grid resolution is lower than the destination grid. Both the computation (solid) and communication (dashed) costs associated with interpolation are shown. For computation cost, the maximum over all processors is plotted. For the latitudinal decomposition, the average cost is also shown (triangles). Ideal scaling curves are provided for comparison.

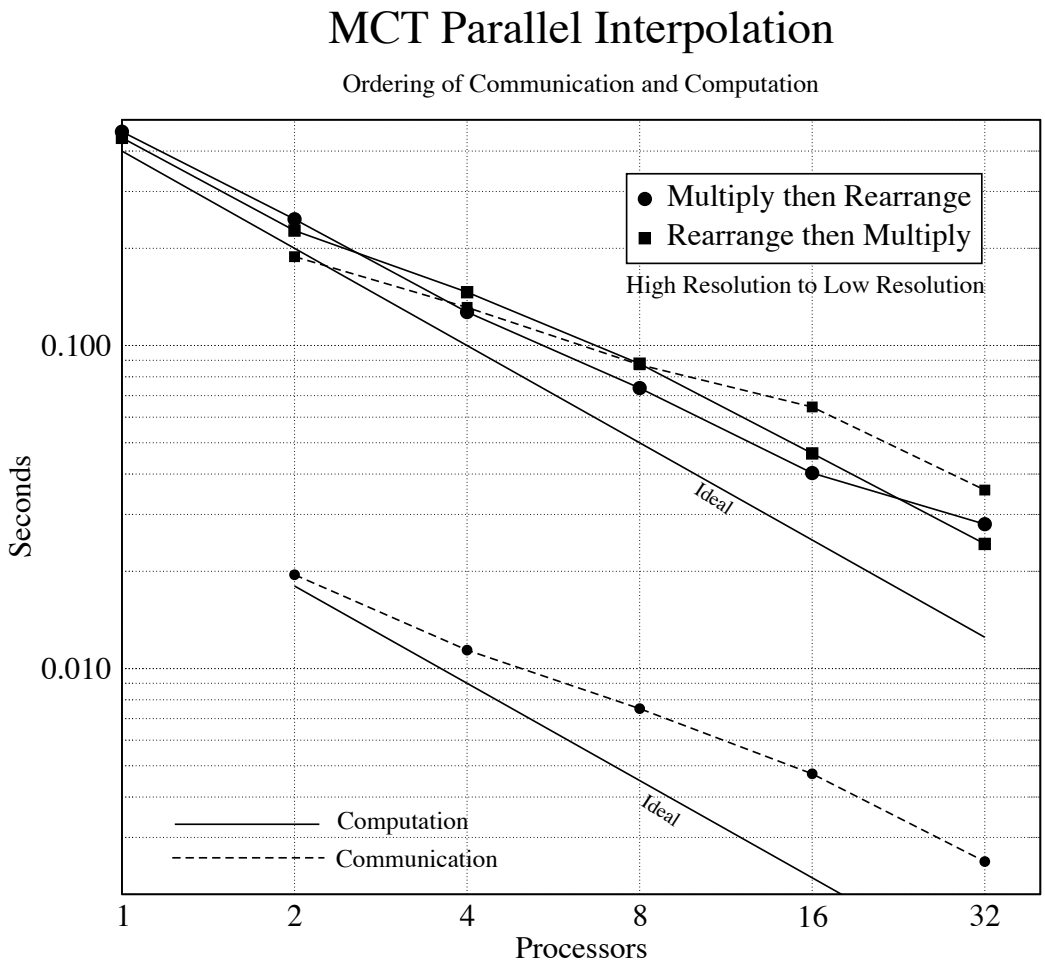

Figure 9: Total time for 10 calls to MCT parallel interpolation routine with different ordering of communication and computation. Source grid is higher resolution than destination grid. Both the computation (solid) and communication (dashed) costs associated with interpolation are shown. Ideal scaling curves are also provided for comparison.

The submitted manuscript has been created by the University of Chicago as Operator of Argonne National Laboratory ("Argonne") under Contract No. W-31-109-ENG-38 with the U.S. Department of Energy. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.