Adaptive MPI Manual
The best contact path for licensing issues is by e-mail to kale@illinois.edu or send correspondence to:

Prof. L. V. Kale  
Dept. of Computer Science  
University of Illinois  
201 N. Goodwin Ave  
Urbana, Illinois 61801 USA  
FAX: (217) 244-6500

Electronic documents will include a direct link to the official Charm++ page at http://charm.cs.illinois.edu/

http://charm.cs.illinois.edu/
# Contents

1 Introduction ....................................................... 3
   1.1 Overview .................................................. 3
   1.2 Terminology ................................................. 6

2 Charm++ .......................................................... 6

3 AMPI ............................................................. 7
   3.1 AMPI Compliance to MPI Standards .......................... 7
   3.2 Name for Main Program ...................................... 8
      3.2.1 Fortran .................................................. 8
      3.2.2 C or C++ ............................................... 8
   3.3 Global Variable Privatization ............................... 8
      3.3.1 Automatic Globals Swapping ............................ 9
      3.3.2 Manual Change ......................................... 9
      3.3.3 Source-to-source Transformation ....................... 11
      3.3.4 TLS-Globals ........................................... 12
   3.4 Extensions for Migrations ................................... 12
      3.4.1 Registering User Data .................................. 13
      3.4.2 Migration ............................................... 13
      3.4.3 Packing/Unpacking Thread Data ......................... 14
   3.5 Extensions for Checkpointing ............................... 17
   3.6 Extensions for Memory Efficiency ......................... 18
   3.7 Extensions for Interoperability ............................ 18
   3.8 Extensions for Sequential Re-run of a Parallel Node .... 20
   3.9 User Defined Initial Mapping ............................... 20
   3.10 Performance Visualization ................................. 21
   3.11 Compiling AMPI Programs .................................. 21

A Installing AMPI .................................................. 21

B Building and Running AMPI Programs ......................... 22
   B.1 Building .................................................... 22
   B.2 Running ..................................................... 22
1 Introduction

This manual describes Adaptive MPI (AMPI), which is an implementation of the MPI-2.2 standard on top of Charm++. Charm++ is a C++-based parallel programming library being developed by Prof. Laxmikant (Sanjay) Kale and his students since 1992 at the University of Illinois at Urbana-Champaign.

We first describe the philosophy behind Adaptive MPI. Then we give a brief introduction to Charm++ and rationale for AMPI. We then describe AMPI in detail. Finally we summarize the changes required for existing MPI codes to run with AMPI. Appendices contain the details of installing AMPI, building and running AMPI programs.

1.1 Overview

Developing parallel Computational Science and Engineering (CSE) applications is a complex task. One has to implement the right physics, develop or choose and code appropriate numerical methods, decide and implement the proper input and output data formats, perform visualizations, and be concerned with correctness and efficiency of the programs. It becomes even more complex for multi-physics coupled simulations, many of which are dynamic and adaptively refined so that load imbalance becomes a major challenge. In addition to imbalance caused by dynamic program behavior, hardware factors such as latencies, variability, and failures must be tolerated by applications. Our philosophy is to lessen the burden of application developers by providing advanced programming paradigms and versatile runtime systems that can handle many common programming and performance concerns automatically and let application programmers focus on the actual application content.

Many of these concerns can be addressed using the processor virtualization and over-decomposition philosophy of Charm++. Thus, the developer only sees virtual processors and lets the runtime system deal with underlying physical processors. This is implemented in AMPI by mapping MPI ranks to Charm++ user-level threads as illustrated in Figure 1. As an immediate and simple benefit, the programmer can use as many virtual processors ("MPI ranks") as the problem can be easily decomposed to. For example, suppose the problem domain has $n \times 2^n$ parts that can be easily distributed but programming for general number of MPI processes is burdensome, then the developer can have $n \times 2^n$ virtual processors on any number of physical ones using AMPI.

![Figure 1: MPI processes are implemented as user-level threads in AMPI](image)

AMPI’s execution model consists of multiple user-level threads per process and, typically, one process per physical processor. The Charm++ scheduler coordinates execution of these threads (also called Virtual Processors or VPs) and controls execution as shown in Figure 2. These VPs can also migrate between processors for the purpose of load balancing or other reasons. The number of VPs per processor specifies

---

1 Currently, MPI-2.2’s PMPI profiling interface is missing from AMPI and the MPI-3.1 standard is under active development, though we already support non-blocking and neighborhood collectives.
the virtualization ratio (degree of over-decomposition). For example, in Figure 2 virtualization ratio is four (there are four VPs per each processor). Figure 3 show how the problem domain is over-decomposed in AMPI's VPs as opposed to other MPI implementations.

Another benefit of virtualization is communication and computation overlap, which is automatically realized in AMPI without programming effort. Techniques such as software pipelining require significant programming effort to achieve this goal and improve performance. However, one can use AMPI to have more virtual processors than physical processors to overlap communication and computation. Each time a VP is blocked for communication, the Charm++ scheduler picks the next VP among those that are ready to execute. In this manner, while some of the VPs of a physical processor are waiting for a message to arrive, others can continue their execution. Thus, performance improves without any changes to the application source code.

Another potential benefit is that of better cache utilization. With over-decomposition, a smaller subdomain is accessed by a VP repeatedly in different function calls before getting blocked by communication and switching to another VP. That smaller subdomain may fit into cache if over-decomposition is enough. This concept is illustrated in Figure 4 where each AMPI subdomain (such as 12) is smaller than corresponding MPI subdomain (such as 3) and may fit into cache memory. Thus, there is a potential performance improvement without changing the source code.
One important concern is that of load imbalance. New generation parallel applications are dynamically varying, meaning that processors’ load is shifting during execution. In a dynamic simulation application such as rocket simulation, burning solid fuel, sub-scaling for a certain part of the mesh, crack propagation, particle flows all contribute to load imbalance. A centralized load balancing strategy built into an application is impractical since each individual module is developed mostly independently by various developers. In addition, embedding a load balancing strategy in the code complicates it greatly, and programming effort increases significantly. The runtime system is uniquely positioned to deal with load imbalance. Figure 5 shows the runtime system migrating a VP after detecting load imbalance. This domain may correspond to a weather forecast model where there is a storm cell in the top-left quadrant, which requires more computation to simulate. AMPI will then migrate VP 13 to balance the division of work across processors and improve performance. Note that incorporating this sort of load balancing inside the application code may take a lot of effort and complicate the code.

There are many different load balancing strategies built into Charm++ that can be selected by an AMPI application developer. Among those, some may fit better for a particular application depending on its characteristics. Moreover, one can write a new load balancer, best suited for an application, by the simple API provided inside Charm++ infrastructure. Our approach is based on actual measurement of load information at runtime, and on migrating computations from heavily loaded to lightly loaded processors. For this approach to be effective, we need the computation to be split into pieces many more in number than available processors. This allows us to flexibly map and re-map these computational pieces to available processors. This approach is usually called “multi-domain decomposition”.

Charm++, which we use as a runtime system layer for the work described here, simplifies our approach.
It embeds an elaborate performance tracing mechanism, a suite of plug-in load balancing strategies, infrastructure for defining and migrating computational load, and is interoperable with other programming paradigms.

1.2 Terminology

Module A module refers to either a complete program or a library with an orchestrator subroutine\(^2\). An orchestrator subroutine specifies the main control flow of the module by calling various subroutines from the associated library and does not usually have much state associated with it.

Thread A thread is a lightweight process that owns a stack and machine registers including a program counter, but shares code and data with other threads within the same address space. If the underlying operating system recognizes a thread, it is known as a kernel thread, otherwise it is known as a user-thread. A context-switch between threads refers to suspending one thread’s execution and transferring control to another thread. Kernel threads typically have higher context switching costs than user-threads because of operating system overheads. The policy implemented by the underlying system for transferring control between threads is known as its thread scheduling policy. Scheduling policy for kernel threads is determined by the operating system, and is often more inflexible than user-threads. Scheduling policy is said to be non-preemptive if a context-switch occurs only when the currently running thread willingly asks to be suspended, otherwise it is said to be preemptive. AMPI threads are non-preemptive user-level threads.

Object An object is just a blob of memory on which certain computations can be performed. The memory is referred to as an object’s state, and the set of computations that can be performed on the object is called the interface of the object.

2 Charm++

Charm++ is an object-oriented parallel programming library for C++. It differs from traditional message passing programming libraries (such as MPI) in that Charm++ is “message-driven”. Message-driven parallel programs do not block the processor waiting for a message to be received. Instead, each message carries with itself a computation that the processor performs on arrival of that message. The underlying runtime system of Charm++ is called Converse, which implements a “scheduler” that chooses which message to schedule next (message-scheduling in Charm++ involves locating the object for which the message is intended, and executing the computation specified in the incoming message on that object). A parallel object in Charm++ is a C++ object on which a certain computations can be asked to be performed from remote processors.

Charm++ programs exhibit latency tolerance since the scheduler always picks up the next available message rather than waiting for a particular message to arrive. They also tend to be modular, because of their object-based nature. Most importantly, Charm++ programs can be dynamically load balanced, because the messages are directed at objects and not at processors; thus allowing the runtime system to migrate the objects from heavily loaded processors to lightly loaded processors. It is this feature of Charm++ that we utilize for AMPI.

Since many CSE applications are originally written using MPI, one would have to do a complete rewrite if they were to be converted to Charm++ to take advantage of dynamic load balancing and other Charm++ features. This is indeed impractical. However, Converse – the runtime system of Charm++ – supports interoperability between different parallel programming paradigms such as parallel objects and threads. Using this feature, we developed AMPI, which is described in more detail in the next section.

\(^2\)Like many software engineering terms, this term is overused, and unfortunately clashes with Fortran90 module that denotes a program unit. We specifically refer to the later as “Fortran 90 module” to avoid confusion.
3 AMPI

AMPI utilizes the dynamic load balancing and other capabilities of Charm++ by associating a “user-level” thread with each Charm++ migratable object. User’s code runs inside this thread, so that it can issue blocking receive calls similar to MPI, and still present the underlying scheduler an opportunity to schedule other computations on the same processor. The runtime system keeps track of the computational loads of each thread as well as the communication graph between AMPI threads, and can migrate these threads in order to balance the overall load while simultaneously minimizing communication overhead.

3.1 AMPI Compliance to MPI Standards

Currently AMPI supports the MPI-2.2 standard, with preliminary support for some MPI-3.1 features and a collection of extensions explained in detail in this manual. One-sided communication calls in MPI-2 and MPI-3 are implemented, but they do not yet take advantage of RMA features. Non-blocking collectives have been defined in AMPI since before MPI-3.0’s adoption of them. Also ROMIO\(^3\) has been integrated into AMPI to support parallel I/O features. Build AMPI with --with-romio and link your application with -lmpiromio to take advantage of this library.

The following MPI basic datatypes are supported in AMPI. (Some are not available in the Fortran binding. Refer to the MPI Standard for details.)

\[
\begin{array}{llll}
\text{MPI_DATATYPE_NULL} & \text{MPI_BYTE} & \text{MPI_UNSIGNED_LONG} & \text{MPI_LONG_DOUBLE_INT} \\
\text{MPI_DOUBLE} & \text{MPI_PACKED} & \text{MPI_LONG_DOUBLE} & \text{MPI_2FLOAT} \\
\text{MPI_INT} & \text{MPI_SHORT} & \text{MPI_FLOAT_INT} & \text{MPI_2DOUBLE} \\
\text{MPI_FLOAT} & \text{MPI_LONG} & \text{MPI_DOUBLE_INT} & \text{MPI_LB} \\
\text{MPI_COMPLEX} & \text{MPI_UNSIGNED_CHAR} & \text{MPI_LONG_INT} & \text{MPI_UB} \\
\text{MPI_LOGICAL} & \text{MPI_UNSIGNED_SHORT} & \text{MPI_2INT} & \\
\text{MPI_CHAR} & \text{MPI_UNSIGNED} & \text{MPI_SHORT_INT} & \\
\end{array}
\]

The following MPI reduction operations are supported in AMPI.

\[
\begin{array}{llllllll}
\text{MPI_MAX} & \text{MPI_MIN} & \text{MPI_SUM} & \text{MPI_PROD} & \text{MPI_MAXLOC} & \text{MPI_MINLOC} & \text{MPI_REPLACE} \\
\text{MPI_LAND} & \text{MPI_LOR} & \text{MPI_LXOR} & \text{MPI_BAND} & \text{MPI_BOR} & \text{MPI_BXOR} & \text{MPI_NO_OP} \\
\end{array}
\]

The following are AMPI extensions to the MPI standard, which will be explained in detail in this manual. All AMPI extensions to the MPI standard are prefixed with AMPI rather than MPI. All extensions are available in C, C++, and Fortran, with the exception of AMPI_Command_argument_count and AMPI_Get_command_argument which are only available in Fortran.

\[
\begin{array}{llll}
\text{AMPI_Migrate} & \text{AMPI_Register_pup} & \text{AMPI_Get_pup_data} \\
\text{AMPI_Migrate_to_pe} & \text{AMPI_Comm_set_migratable} & \text{AMPI_Evacuate} \\
\text{AMPI_Load_set_value} & \text{AMPI_Load_start_measure} & \text{AMPI_Load_stop_measure} \\
\text{AMPI_Iget} & \text{AMPI_Iget_wait} & \text{AMPI_Iget_data} \\
\text{AMPI_Iget_free} & \text{AMPI_Type_is_contiguous} & \text{AMPI_Register_main} \\
\text{AMPI_Alltoall_iget} & \text{AMPI_Alltoall_medium} & \text{AMPI_Alltoall_long} \\
\text{AMPI_Yield} & \text{AMPI_Suspend} & \text{AMPI_Resume} \\
\text{AMPI_Register_just_migrated} & \text{AMPI_Register_about_to_migrate} \\
\text{AMPI_Command_argument_count} & \text{AMPI_Get_command_argument} \\
\end{array}
\]

AMPI also provides a set of built-in attributes on all communicators and windows to find the PE or Node number that a virtual processor is currently running on, as well as the total number of PEs and Nodes in the job. The built-in attributes are AMPI_MY_PE, AMPI_MY_NODE, AMPI_NUM_PES, and AMPI_NUM_NODES. These attributes are accessible from any rank by calling MPI_Comm_get_attr.

\(^3\)http://www-unix.mcs.anl.gov/romio/
3.2 Name for Main Program

To convert an existing program to use AMPI, the main function or program may need to be renamed. The changes should be made as follows:

3.2.1 Fortran

You must declare the main program as a subroutine called “MPI_MAIN”. Do not declare the main subroutine as a program because it will never be called by the AMPI runtime.

```
program pgm -> subroutine MPI_Main
  ...
end program -> end subroutine
```

3.2.2 C or C++

The main function can be left as is, if mpi.h is included before the main function. This header file has a preprocessor macro that renames main, and the renamed version is called by the AMPI runtime by each thread.

3.3 Global Variable Privatization

For the before-mentioned benefits to be effective, one needs to map multiple user-level threads onto each processor. Traditional MPI programs assume that the entire processor is allocated to themselves, and that only one thread of control exists within the process’s address space. So, they may use global and static variables in the program. However, global and static variables are problematic for multi-threaded environments such as AMPI or OpenMP. This is because there is a single instance of those variables so they will be shared among different threads in the single address space and a wrong result may be produced by the program. Figure 6 shows an example of a multi-threaded application with two threads in a single process. `var` is a global or static variable in this example. Thread 1 assigns a value to it, then it gets blocked for communication and another thread can continue. Thereby, thread 2 is scheduled next and accesses `var` which is wrong. Semantics of this program needs separate instances of `var` for each of the threads. That is where the need arises to make some transformations to the original MPI program in order to run correctly with AMPI. Note, this is the only change necessary to run an MPI program with AMPI, that the program be thread-safe and have no global variables whose values differ across different MPI ranks.

The basic transformation needed to port the MPI program to AMPI is privatization of global variables. With the MPI process model, each MPI node can keep a copy of its own “permanent variables” – variables

```
<table>
<thead>
<tr>
<th>Thread 1</th>
<th>Thread 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>var = myid</code> (1)</td>
<td><code>var = myid</code> (2)</td>
</tr>
<tr>
<td><code>MPI_Recv()</code></td>
<td><code>MPI_Recv()</code></td>
</tr>
<tr>
<td>(block...)</td>
<td>(block...)</td>
</tr>
<tr>
<td><code>b = var</code></td>
<td><code>b = var</code></td>
</tr>
</tbody>
</table>
```

Figure 6: Global or static variables are an issue for AMPI

4Typical Fortran MPI programs contain three types of global variables.

1. Global variables that are “read-only”. These are either parameters that are set at compile-time. Or other variables that are read as input or set at the beginning of the program and do not change during execution. It is not necessary to privatize such variables.

2. Global variables that are used as temporary buffers. These are variables that are used temporarily to store values to
that are accessible from more than one subroutine without passing them as arguments. Module variables, “saved” subroutine local variables, and common blocks in Fortran90 belong to this category. If such a program is executed without privatization on AMPI, all the AMPI threads that reside on one processor will access the same copy of such variables, which is clearly not the desired semantics. To ensure correct execution of the original source program, it is necessary to make such variables “private” to individual threads. We provide two choices: automatic global swapping and manual code modification.

3.3.1 Automatic Globals Swapping

Thanks to the ELF Object Format, we have successfully automated the procedure of switching the set of user global variables when switching thread contexts. Executable and Linkable Format (ELF) is a common standard file format for Object Files in Unix-like operating systems. ELF maintains a Global Offset Table (GOT) for globals so it is possible to switch GOT contents at thread context-switch by the runtime system.

The only thing that the user needs to do is to set flag `-swapglobals` at compile and link time (e.g. “ampicc -o prog prog.c -swapglobals”). It does not need any change to the source code and works with any language (C, C++, Fortran, etc). However, it does not handle static variables and has a context switching overhead that grows with the number of global variables. Currently, this feature only works on x86 and x86_64 (e.g. amd64) platforms that fully support ELF. Thus, it may not work on PPC or Itanium, or on some microkernels such as Catamount. When this feature does not work for you, you can try other ways of handling global or static variables, which are detailed in the following sections.

3.3.2 Manual Change

We have employed a strategy of argument passing to do this privatization transformation. That is, the global variables are bunched together in a single user-defined type, which is allocated by each thread dynamically. Then a pointer to this type is passed from subroutine to subroutine as an argument. Since the subroutine arguments are passed on the stack, which is not shared across all threads, each subroutine when executing within a thread operates on a private copy of the global variables.

This scheme is demonstrated in the following examples. The original Fortran90 code contains a module `shareddata`. This module is used in the main program and a subroutine `subA`.

```fortran
!FORTRAN EXAMPLE
MODULE shareddata
   INTEGER :: myrank
   DOUBLE PRECISION :: xyz(100)
END MODULE

SUBROUTINE MPI_MAIN
   USE shareddata
   include 'mpif.h'
   INTEGER :: i, ierr
   CALL MPI_Init(ierr)
   CALL MPI_Comm_rank(MPI_COMM_WORLD, myrank, ierr)
   DO i = 1, 100
      xyz(i) = i + myrank
   END DO
   CALL subA
   CALL MPI_Finalize(ierr)
END PROGRAM
```

be accessible across subroutines. These variables have a characteristic that there is no blocking call such as `MPI_recv` between the time the variable is set and the time it is ever used. It is not necessary to privatize such variables either.

3. True global variables. These are used across subroutines that contain blocking receives and therefore the possibility of a context switch between the definition and use of the variable. These variables need to be privatized.
SUBROUTINE subA
  USE shareddata
  INTEGER :: i
  DO i = 1, 100
    xyz(i) = xyz(i) + 1.0
  END DO
END SUBROUTINE

//C Example
#include <mpi.h>

int myrank;
double xyz[100];

void subA();
int main(int argc, char** argv){
  int i;
  MPI_Init(&argc, &argv);
  MPI_Comm_rank(MPI_COMM_WORLD, &myrank);
  for(i=0;i<100;i++)
    xyz[i] = i + myrank;
  subA();
  MPI_Finalize();
}

void subA(){
  int i;
  for(i=0;i<100;i++)
    xyz[i] = xyz[i] + 1.0;
}

AMPI executes the main subroutine inside a user-level thread as a subroutine.
Now we transform this program using the argument passing strategy. We first group the shared data into a user-defined type.

!FORTRAN EXAMPLE
MODULE shareddata
  TYPE chunk
    INTEGER :: myrank
    DOUBLE PRECISION :: xyz(100)
  END TYPE
END MODULE

//C Example
struct shareddata{
  int myrank;
  double xyz[100];
};

Now we modify the main subroutine to dynamically allocate this data and change the references to them. Subroutine subA is then modified to take this data as argument.

!FORTRAN EXAMPLE
SUBROUTINE MPI_Main
    USE shareddata
    USE AMPI
    INTEGER :: i, ierr
    TYPE(chunk), pointer :: c
    CALL MPI_Init(ierr)
    ALLOCATE(c)
    CALL MPI_Comm_rank(MPI_COMM_WORLD, c%myrank, ierr)
    DO i = 1, 100
        c%xyz(i) = i + c%myrank
    END DO
    CALL subA(c)
    CALL MPI_Finalize(ierr)
END SUBROUTINE

SUBROUTINE subA(c)
    USE shareddata
    TYPE(chunk) :: c
    INTEGER :: i
    DO i = 1, 100
        c%xyz(i) = c%xyz(i) + 1.0
    END DO
END SUBROUTINE

//C Example
void MPI_Main{
    int i,ierr;
    struct shareddata *c;
    ierr = MPI_Init();
    c = (struct shareddata*)malloc(sizeof(struct shareddata));
    ierr = MPI_Comm_rank(MPI_COMM_WORLD, c.myrank);
    for(i=0;i<100;i++)
        c.xyz[i] = i + c.myrank;
    subA(c);
    ierr = MPI_Finalize();
}

void subA(struct shareddata *c){
    int i;
    for(i=0;i<100;i++)
        c.xyz[i] = c.xyz[i] + 1.0;
}

With these changes, the above program can be made thread-safe. Note that it is not really necessary to dynamically allocate chunk. One could have declared it as a local variable in subroutine MPI_Main. (Or for a small example such as this, one could have just removed the shareddata module, and instead declared both variables xyz and myrank as local variables). This is indeed a good idea if shared data are small in size. For large shared data, it would be better to do heap allocation because in AMPI, the stack sizes are fixed at the beginning (and can be specified from the command line) and stacks do not grow dynamically.

3.3.3 Source-to-source Transformation

Another approach is to do the changes described in the previous scheme automatically. It means that we can use a tool to transform the source code to move global or static variables in an object and pass them
Table 1: Portability of current implementations of three privatization schemes. “Yes” means we have implemented this technique. “Maybe” indicates there are no theoretical problems, but no implementation exists. “No” indicates the technique is impossible on this platform.

3.3.4 TLS-Globals

Thread Local Store (TLS) was originally employed in kernel threads to localize variables and provide thread safety. It can be used by annotating global/static variables in C/C++ with `__thread` in the source code. Thus, those variables will have one instance per extant thread. This keyword is not an official extension of the C language, though compiler writers are encouraged to implement this feature. Currently, the ELF file format supports Thread Local Storage.

It handles both global and static variables and has no context-switching overhead. Context-switching is just changing the TLS segment register to point to the thread’s local copy. However, although it is popular, it is not supported by all compilers. Currently, Charm++ supports it for x86/x86_64 platforms. A modified gfortran is also available to use this feature upon request. To use TLS-Globals, one has to add `__thread` before all global variables. For the example above, the following changes to the code handles the global variables:

```c
__thread int myrank;
__thread double xyz[100];
```

The runtime system also should know that TLS-Globals is used at compile time:

```c
ampiCC -o example example.C -tlsglobals
```

Table 1 shows portability of different schemes.

3.4 Extensions for Migrations

AMPI provides fully automated support for migrating MPI ranks between nodes of a system without any application-specific code at all. We do so using a memory allocator, Isomalloc, that allocates memory per user-level thread to globally unique virtual memory addresses. This means that every PE in the system reserves slices of virtual memory for all other threads, allowing transparent migration of stacks and pointers into memory (Isomalloc requires 64-bit virtual memory addresses and support from the operating system for mapping memory to arbitrary virtual addresses). Applications only need to link with Isomalloc to enable automatic migratability, using `-memory isomalloc`.

For systems that do not support Isomalloc and for users that wish to have more fine-grain control over which application data structures will be copied at migration time, we have added a few calls to AMPI. These include the ability to register thread-specific data with the run-time system, to pack and unpack all of the thread’s data, and to express willingness to migrate.

---

5[http://www.eclipse.org/photran](http://www.eclipse.org/photran)

6[http://rosecompiler.org/](http://rosecompiler.org/)
3.4.1 Registering User Data

When the AMPI runtime system decides that load imbalance exists within the application, it will invoke one of its internal load balancing strategies, which determines the new mapping of AMPI ranks so as to balance the load. Then the AMPI runtime packs up the rank’s state and moves it to its new home processor. AMPI packs up any internal data in use by the rank, including the thread’s stack in use. This means that the local variables declared in subroutines in a rank, which are created on stack, are automatically packed up by the AMPI runtime system. However, it has no way of knowing what other data are in use by the rank. Thus upon starting execution, a rank needs to notify the system about the data that it is going to use (apart from local variables). Even with the data registration, AMPI cannot determine what size the data is, or whether the registered data contains pointers to other places in memory. For this purpose, a packing subroutine also needs to be provided to the AMPI runtime system along with registered data. (See next section for writing packing subroutines.) The call provided by AMPI for doing this is **AMPI_Register_pup**. This function takes three arguments: a data item to be transported along with the rank, the pack subroutine, and a pointer to an integer which denotes the registration identifier. In C/C++ programs, it may be necessary to use this integer value after migration completes and control returns to the rank with the function **AMPI_Get_pup_data**.

3.4.2 Migration

The AMPI runtime system could detect load imbalance by itself and invoke the load balancing strategy. However, since the application code is going to pack/unpack the rank’s data, writing the pack subroutine will be complicated if migrations occur at a stage unknown to the application. For example, if the system decides to migrate a rank while it is in initialization stage (say, reading input files), application code will have to keep track of how much data it has read, what files are open etc. Typically, since initialization occurs only once in the beginning, load imbalance at that stage would not matter much. Therefore, we want the demand to perform load balance check to be initiated by the application.

AMPI provides a subroutine **AMPI_Migrate(MPI_Info hints)**; for this purpose. Each rank periodically calls **AMPI_Migrate**. Typical CSE applications are iterative and perform multiple time-steps. One should call **AMPI_Migrate** in each rank at the end of some fixed number of timesteps. The frequency of **AMPI_Migrate** should be determined by a tradeoff between conflicting factors such as the load balancing overhead, and performance degradation caused by load imbalance. In some other applications, where application suspects that load imbalance may have occurred, as in the case of adaptive mesh refinement; it would be more effective if it performs a couple of timesteps before telling the system to re-map ranks. This will give the AMPI runtime system some time to collect the new load and communication statistics upon which it bases its migration decisions. Note that **AMPI_Migrate** does NOT tell the system to migrate the rank, but merely tells the system to check the load balance after all the ranks call **AMPI_Migrate**. To migrate the rank or not is decided only by the system’s load balancing strategy.

Essentially, a call to **AMPI_Migrate** signifies to the runtime system that the application has reached a point at which it is safe to serialize the local state. Knowing this, the runtime system can act in several ways. The MPI_Info object taken as a parameter by **AMPI_Migrate** gives users a way to influence the runtime system’s decision-making and behavior. Users should set the MPI_Info key "ampi_load_balance" to one of the following values: "sync", "async", or "false". Synchronous load balancing assumes that the application is already at a synchronization point. Asynchronous load balancing does not assume this, and a value of "false" is the same as not requesting load balancing at all.

// Setup
MPI_Info hints;
MPI_Info_create(&hints);
MPI_Info_set(hints, "ampi_load_balance", "sync");

...
for (int iter=0; iter < max_iters; iter++) {
    // Time step work ...
    if (iter % lb_freq == 0)
        AMPI_Migrate(hints);
}

We also provide callbacks that user code can register with the runtime system to be invoked just before and right after migration: AMPI_Register_about_to_migrate and AMPI_Register_just_migrated respectively. We also provide routines for starting and stopping load measurements, and for users to explicitly set the load value of a rank using the following: AMPI_Load_start_measure, AMPI_Load_stop_measure, and AMPI_Load_set_value. And since AMPI builds on top of Charm++, users can experiment with the suite of load balancing strategies included with Charm++, as well as write their own strategies based on user-level information and heuristics.

3.4.3 Packing/Unpacking Thread Data

Once the AMPI runtime system decides which ranks to send to which processors, it calls the specified pack subroutine for that rank, with the rank-specific data that was registered with the system using AMPI_Register_pup. If an AMPI application uses Isomalloc, then the system will define the Pack/Unpack routines for the user. This section explains how a subroutine should be written for performing explicit pack/unpack.

There are three steps for transporting the rank’s data to another processor. First, the system calls a subroutine to get the size of the buffer required to pack the rank’s data. This is called the “sizing” step. In the next step, which is called immediately afterward on the source processor, the system allocates the required buffer and calls the subroutine to pack the rank’s data into that buffer. This is called the “packing” step. This packed data is then sent as a message to the destination processor, where first a rank is created (along with the thread) and a subroutine is called to unpack the rank’s data from the buffer. This is called the “unpacking” step.

Though the above description mentions three subroutines called by the AMPI runtime system, it is possible to actually write a single subroutine that will perform all the three tasks. This is achieved using something we call a “pupper”. A pupper is an external subroutine that is passed to the rank’s pack-unpack-sizing subroutine, and this subroutine, when called in different phases performs different tasks. An example will make this clear:

Suppose the user data, chunk, is defined as a derived type in Fortran90:

```
MODULE chunkmod
    INTEGER, parameter :: nx=4, ny=4, tchunks=16
    TYPE, PUBLIC :: chunk
        REAL(KIND=8) t(22,22)
        INTEGER xidx, yidx
        REAL(KIND=8), dimension(400):: bxm, bxp, bym, byp
    END TYPE chunk
END MODULE
```

Then the pack-unpack subroutine chunkpup for this chunk module is written as: 
SUBROUTINE chunkpup(p, c)
    USE pupmod
    USE chunkmod
    IMPLICIT NONE
    INTEGER :: p
    TYPE(chunk) :: c

    call pup(p, c%t)
    call pup(p, c%xidx)
    call pup(p, c%yidx)
    call pup(p, c%bxm)
    call pup(p, c%bxp)
    call pup(p, c%bym)
    call pup(p, c%byp)

END SUBROUTINE

There are several things to note in this example. First, the same subroutine \texttt{pup} (declared in module \texttt{pupmod}) is called to size/pack/unpack any type of data. This is possible because of procedure overloading possible in Fortran90. Second is the integer argument \texttt{p}. It is this argument that specifies whether this invocation of subroutine \texttt{chunkpup} is sizing, packing or unpacking. Third, the integer parameters declared in the type \texttt{chunk} need not be packed or unpacked since they are guaranteed to be constants and thus available on any processor.

A few other functions are provided in module \texttt{pupmod}. These functions provide more control over the packing/unpacking process. Suppose one modifies the \texttt{chunk} type to include allocatable data or pointers that are allocated dynamically at runtime. In this case, when chunk is packed, these allocated data structures should be deallocated after copying them to buffers, and when chunk is unpacked, these data structures should be allocated before copying them from the buffers. For this purpose, one needs to know whether the invocation of \texttt{chunkpup} is a packing one or unpacking one. For this purpose, the \texttt{pupmod} module provides functions \texttt{fpup_isdeleting}(\texttt{fpup_isunpacking}). These functions return logical value \texttt{.TRUE.} if the invocation is for packing (unpacking), and \texttt{.FALSE.} otherwise. The following example demonstrates this:

Suppose the type \texttt{dchunk} is declared as:

```fortran
!FORTRAN EXAMPLE
MODULE dchunkmod
    TYPE, PUBLIC :: dchunk
        INTEGER :: asize
        REAL(KIND=8), pointer :: xarr(:), yarr(:)
    END TYPE dchunk
END MODULE
```

```c
//C Example
struct dchunk{
```
int asize;
double* xarr, *yarr;
);

Then the pack-unpack subroutine is written as:

!FORTRAN EXAMPLE
SUBROUTINE dchunkpup(p, c)
    USE pupmod
    USE dchunkmod
    IMPLICIT NONE
    INTEGER :: p
    TYPE(dchunk) :: c

    pup(p, c%asize)

    IF (fpup_isunpacking(p)) THEN !! if invocation is for unpacking
        allocate(c%xarr(c%asize))
        ALLOCATE(c%yarr(c%asize))
    ENDIF

    pup(p, c%xarr)
    pup(p, c%yarr)

    IF (fpup_isdeleting(p)) THEN !! if invocation is for packing
        DEALLOCATE(c%xarr)
        DEALLOCATE(c%yarr)
    ENDIF

END SUBROUTINE

//C Example
void dchunkpup(pup_er p, struct dchunk c){
    pup_int(p, c.asize);
    if(pup_isUnpacking(p)){
        c.xarr = (double *)malloc(sizeof(double)*c.asize);
        c.yarr = (double *)malloc(sizeof(double)*c.asize);
    }
    pup_doubles(p, c.xarr, c.asize);
    pup_doubles(p, c.yarr, c.asize);
    if(pup_isPacking(p)){
        free(c.xarr);
        free(c.yarr);
    }
}

One more function fpup_issizing is also available in module pupmod that returns .TRUE. when the invocation is a sizing one. In practice one almost never needs to use it.

Charm++ also provides higher-level PUP routines for C++ STL data structures and Fortran90 data types. The STL PUP routines will deduce the size of the structure automatically, so that the size of the data does not have to be passed in to the PUP routine. This facilitates writing PUP routines for large pre-existing codebases. To use it, simply include pup_stl.h in the user code. For modern Fortran with pointers and allocatable data types, AMPI provides a similarly automated PUP interface called apup. User
code can include pupmod and then call apup() on any array (pointer or allocatable, multi-dimensional) of built-in types (character, short, int, long, real, double, complex, double complex, logical) and the runtime will deduce the size and shape of the array, including unassociated and NULL pointers. Here is the dchunk example from earlier, written to use the apup interface:

!FORTRAN EXAMPLE
SUBROUTINE dchunkpup(p, c)
  USE pupmod
  USE dchunkmod
  IMPLICIT NONE
  INTEGER :: p
  TYPE(dchunk) :: c

  !! no need for asize
  !! no isunpacking allocation necessary

  apup(p, c%xarr)
apup(p, c%yarr)

  !! no isdeleting deallocation necessary
END SUBROUTINE

3.5 Extensions for Checkpointing

The pack-unpack subroutines written for migrations make sure that the current state of the program is correctly packed (serialized) so that it can be restarted on a different processor. Using the same subroutines, it is also possible to save the state of the program to disk, so that if the program were to crash abruptly, or if the allocated time for the program expires before completing execution, the program can be restarted from the previously checkpointed state. Thus, the pack-unpack subroutines act as the key facility for checkpointing in addition to their usual role for migration. Just as in load balancing, no application specific code is required when using Isomalloc: the AMPI runtime takes care of all the details involved in migrating data.

To perform a checkpoint in an AMPI program, all you have to do is make a call to int AMPI_Migrate(MPI_Info hints) with an MPI_Info object that specifies how you would like to checkpoint. Checkpointing can be thought of as migrating AMPI ranks to storage. Users set the checkpointing policy on an MPI_Info object’s "ampi_checkpoint" key to one of the following values: "to_file=directory_name", "in_memory", or "false".

Checkpointing to file tells the runtime system to save checkpoints in a given directory. (Typically, in an iterative program, the iteration number, converted to a character string, can serve as a checkpoint directory name.) This directory is created, and the entire state of the program is checkpointed to this directory. One can restart the program from the checkpointed state (using the same, more, or fewer physical processors than were checkpointed with) by specifying "+restart directory_name" on the command-line.

Checkpointing in memory allows applications to transparently tolerate failures online. The checkpointing scheme used here is a double in-memory checkpoint, in which virtual processors exchange checkpoints pairwise across nodes in each other’s memory such that if one node fails, that failed node’s AMPI ranks can be restarted by its buddy once the failure is detected by the runtime system. As long as no two buddy nodes fail in the same checkpointing interval, the system can restart online without intervention from the user (provided the job scheduler does not revoke its allocation). Any load imbalance resulting from the restart can then be managed by the runtime system. Use of this scheme is illustrated in the code snippet below.

// Setup
MPI_Info hints;
MPI_Info_create(&hints);
MPI_Info_set(hints, "ampi_checkpoint", "in_memory");

... // Main time-stepping loop
for (int iter=0; iter < max_iters; iter++) {
    // Time step work ...
    if (iter % chkpt_freq == 0)
        AMPI_Migrate(hints);
}

A value of "false" results in no checkpoint being done that step. Note that AMPI_Migrate is a collective function, meaning every virtual processor in the program needs to call this subroutine with the same MPI_Info object. The checkpointing capabilities of AMPI are powered by the Charm++ runtime system. For more information about checkpoint/restart mechanisms please refer to the Charm++ manual ??.

3.6 Extensions for Memory Efficiency

MPI functions usually require the user to preallocate the data buffers needed before the functions being called. For unblocking communication primitives, sometimes the user would like to do lazy memory allocation until the data actually arrives, which gives the opportunities to write more memory efficient programs. We provide a set of AMPI functions as an extension to the standard MPI-2 one-sided calls, where we provide a split phase MPI_Get called AMPI_Iget. AMPI_Iget preserves the similar semantics as MPI_Get except that no user buffer is provided to hold incoming data. AMPI_Iget_wait will block until the requested data arrives and runtime system takes care to allocate space, do appropriate unpacking based on data type, and return. AMPI_Iget_free lets the runtime system free the resources being used for this get request including the data buffer. Finally, AMPI_Iget_data is the routine used to access the data.

int AMPI_Iget(MPI_Aint orgdisp, int orgcnt, MPI_Datatype orgtype, int rank,
              MPI_Aint targdisp, int targcnt, MPI_Datatype targtype, MPI_Win win,
              MPI_Request *request);

int AMPI_Iget_wait(MPI_Request *request, MPI_Status *status, MPI_Win win);

int AMPI_Iget_free(MPI_Request *request, MPI_Status *status, MPI_Win win);

int AMPI_Iget_data(void *data, MPI_Status status);

3.7 Extensions for Interoperability

Interoperability between different modules is essential for coding coupled simulations. In this extension to AMPI, each MPI application module runs within its own group of user-level threads distributed over the physical parallel machine. In order to let AMPI know which ranks are to be created, and in what order, a top level registration routine needs to be written. A real-world example will make this clear. We have an MPI code for fluids and another MPI code for solids, both with their main programs, then we first transform each individual code to run correctly under AMPI as standalone codes. Aside from the global and static variable privatization transformations needed, this also involves making the main program into a subroutine and naming it MPI_Main.
Thus now, we have two MPIs, one for the fluids code and one for the solids code. We now make these codes co-exist within the same executable, by first renaming these MPIs as Fluids_Main and Solids_Main\(^7\) writing a subroutine called MPI_Setup.

```fortran
SUBROUTINE MPI_Setup
    USE ampi
    CALL AMPI_Register_main(Solids_Main)
    CALL AMPI_Register_main(Fluids_Main)
END SUBROUTINE
```

This subroutine is called from the internal initialization routines of AMPI and tells AMPI how many numbers of distinct modules exist, and which orchestrator subroutines they execute.

The number of ranks to create for each module is specified on the command line when an AMPI program is run. Appendix B explains how AMPI programs are run, and how to specify the number of ranks (+vp option). In the above case, suppose one wants to create 128 ranks of Solids and 64 ranks of Fluids on 32 physical processors, one would specify those with multiple +vp options on the command line as:

```bash
> charmrun gen1.x +p 32 +vp 128 +vp 64
```

This will ensure that multiple modules representing different complete applications can co-exist within the same executable. They can also continue to communicate among their own ranks using the same AMPI function calls to send and receive with communicator argument as MPI_COMM_WORLD. But this would be completely useless if these individual applications cannot communicate with each other, which is essential for building efficient coupled codes. For this purpose, we have extended the AMPI functionality to allow multiple “COMM_WORLDS”; one for each application. These world communicators form a “communicator universe”: an array of communicators aptly called MPI_COMM_UNIVERSE. This array of communicators is indexed \([1\. \ldots \text{MPI_MAX_COMM}]\). In the current implementation, MPI_MAX_COMM is 8, that is, maximum of 8 applications can co-exist within the same executable.

The order of these COMM_WORLDS within MPI_COMM_UNIVERSE is determined by the order in which individual applications are registered in MPI_Setup.

Thus, in the above example, the communicator for the Solids module would be MPI_COMM_UNIVERSE(1) and communicator for Fluids module would be MPI_COMM_UNIVERSE(2).

Now any rank within one application can communicate with any rank in the other application using the familiar send or receive AMPI calls by specifying the appropriate communicator and the rank number within that communicator in the call. For example if a Solids rank number 36 wants to send data to rank number 47 within the Fluids module, it calls:

```fortran
CALL MPI_Send(InitialTime, 1, MPI_Double_Precision, tag, 47, MPI_Comm_Universe(Fluids_Comm), ierr)
```

```c
int Fluids_Comm = 2;
ierr = MPI_Send(InitialTime, 1, MPI_DOUBLE, tag, 47, MPI_Comm_UNIVERSE(Fluids_Comm));
```

\(^7\)Currently, we assume that the interface code, which does mapping and interpolation among the boundary values of Fluids and Solids domain, is integrated with one of Fluids and Solids.
The Fluids rank has to issue a corresponding receive call to receive this data:

```fortran
!FORTRAN EXAMPLE
INTEGER , PARAMETER :: Solids_Comm = 1
CALL MPI_Recv(InitialTime, 1, MPI_Double_Precision, tag,
               36, MPI_Comm_Universe(Solids_Comm), stat, ierr)
```

```c
int Solids_Comm = 1;
ierr = MPI_Recv(InitialTime, 1, MPI_DOUBLE, tag,
                 36, MPI_Comm_Universe(Solids_Comm), &stat);
```

### 3.8 Extensions for Sequential Re-run of a Parallel Node

In some scenarios, a sequential re-run of a parallel node is desired. One example is instruction-level accurate architecture simulations, in which case the user may wish to repeat the execution of a node in a parallel run in the sequential simulator. AMPI provides support for such needs by logging the change in the MPI environment on a certain processors. To activate the feature, build AMPI module with variable “AMPIMSGLOG” defined, like the following command in charm directory. (Linking with zlib “-lz” might be required with this, for generating compressed log file.)

```
> ./build AMPI netlrt-lin-x86_64 -DAMPIMSGLOG
```

The feature is used in two phases: writing (logging) the environment and repeating the run. The first logging phase is invoked by a parallel run of the AMPI program with some additional command line options.

```
> ./charmrun ./pgm +p4 +vp4 +msgLogWrite +msgLogRank 2 +msgLogFilename "msg2.log"
```

In the above example, a parallel run with 4 processors and 4 VPs will be executed, and the changes in the MPI environment of processor 2 (also VP 2, starting from 0) will get logged into diskfile "msg2.log".

Unlike the first run, the re-run is a sequential program, so it is not invoked by charmrun (and omitting charmrun options like +p4 and +vp4), and additional command line options are required as well.

```
> ./pgm +msgLogRead +msgLogRank 2 +msgLogFilename "msg2.log"
```

### 3.9 User Defined Initial Mapping

You can define the initial mapping of virtual processors (vp) to physical processors (p) as a runtime option. You can choose from predefined initial mappings or define your own mappings. The following predefined mappings are available:

**Round Robin** This mapping scheme maps virtual processor to physical processor in round-robin fashion, i.e. if there are 8 virtual processors and 2 physical processors then virtual processors indexed 0,2,4,6 will be mapped to physical processor 0 and virtual processors indexed 1,3,5,7 will be mapped to physical processor 1.

```
> ./charmrun ./hello +p2 +vp8 +mapping RR_MAP
```

**Block Mapping** This mapping scheme maps virtual processors to physical processor in ranks, i.e. if there are 8 virtual processors and 2 physical processors then virtual processors indexed 0,1,2,3 will be mapped to physical processor 0 and virtual processors indexed 4,5,6,7 will be mapped to physical processor 1.

```
> ./charmrun ./hello +p2 +vp8 +mapping BLOCK_MAP
```
Proportional Mapping This scheme takes the processing capability of physical processors into account for mapping virtual processors to physical processors, i.e. if there are 2 processors running at different frequencies, then the number of virtual processors mapped to processors will be in proportion to their processing power. To make the load balancing framework aware of the heterogeneity of the system, the flag \textit{+LBTestPESpeed} should also be used.

\begin{verbatim}
> ./charmrun ./hello +p2 +vp8 +mapping PROP_MAP
> ./charmrun ./hello +p2 +vp8 +mapping PROP_MAP +balancer GreedyLB +LBTestPESpeed
\end{verbatim}

If you want to define your own mapping scheme, please contact us for assistance.

3.10 Performance Visualization

AMPI users can take advantage of Charm++’s tracing framework and associated performance visualization tool, Projections. Projections provides a number of different views of performance data that help users diagnose performance issues.

AMPI defines the following extensions for tracing support:

\begin{verbatim}
AMPI_Trace_begin
AMPI_Trace_end
AMPI_Trace_register_function_name
AMPI_Trace_register_function_id
AMPI_Trace_start_function_name
AMPI_Trace_start_function_id
AMPI_Trace_end_function_name
AMPI_Trace_end_function_id
\end{verbatim}

See the Projections manual for information on how to visualize traces.

3.11 Compiling AMPI Programs

Charm++ provides a cross-platform compile-and-link script called \texttt{charmc} to compile C, C++, Fortran, Charm++ and AMPI programs. This script resides in the \texttt{bin} subdirectory in the Charm++ installation directory. The main purpose of this script is to deal with the differences of various compiler names and command-line options across various machines on which Charm++ runs. It is recommended that \texttt{charmc} or the AMPI compiler scripts be used to compile and link AMPI programs. One major advantage of using these is that one does not have to specify which libraries are to be linked for ensuring that C++ and Fortran90 codes are linked together correctly. Appropriate libraries required for linking such modules together are known to \texttt{charmc} for various machines.

In spite of the platform-neutral syntax of \texttt{charmc}, one may have to specify some platform-specific options for compiling and building AMPI codes. Fortunately, if \texttt{charmc} does not recognize any particular options on its command line, it promptly passes it to all the individual compilers and linkers it invokes to compile the program. See the appendix for more details on building and running AMPI programs.

A Installing AMPI

AMPI is included in the source distribution of Charm++. To get the latest sources from PPL, visit: http://charm.cs.illinois.edu/software and follow the download links. Then build Charm++ and AMPI from source.

The build script for Charm++ is called \texttt{build}. The syntax for this script is:

\begin{verbatim}
> build <target> <version> <opts>
\end{verbatim}

For building AMPI (which also includes building Charm++ and other libraries needed by AMPI), specify \texttt{<target>} to be \texttt{AMPI}. And \texttt{<opts>} are command line options passed to the \texttt{charmc} compile script. Common compile time options such as \texttt{-g}, \texttt{-O}, \texttt{-Ipath}, \texttt{-Lpath}, \texttt{-llib} are accepted.

To build a debugging version of AMPI, use the option: \texttt{-g}. To build a production version of AMPI, use the option: \texttt{--with-production}. 21
<version> depends on the machine, operating system, and the underlying communication library one wants to use for running AMPI programs. See the charm/README file for details on picking the proper version. Following is an example of how to build a production version of AMPI under linux and ethernet environment, with debugging symbols produced:

> build AMPI netlrlts-linux-x86_64 --with-production -g

## B Building and Running AMPI Programs

### B.1 Building

Charm++ provides a compiler called charmc in your charm/bin/ directory. You can use this compiler to build your AMPI program the same way as other compilers like cc. To build an AMPI program, the command line option `-language ampi` should be applied, which the AMPI compiler wrappers in charm/bin do automatically. All the command line flags that you would use for other compilers can be used with the AMPI compilers the same way. For example:

> ampicc -c pgm.c -O3
> ampif90 -c pgm.f90 -O0 -g
> charmc -language ampi -o pgm pgm.o -lm -O3

To use Isomalloc for transparently migrating user heap data, link with `-memory isomalloc`. To use a Charm++ load balancer, link a startegy or a suite of strategies in with `-module LBs`. For example:

> ampicc pgm.c -o pgm -O3 -memory isomalloc -module CommonLBs

### B.2 Running

The Charm++ distribution contains a script called `charmrun` that makes the job of running AMPI programs portable and easier across all parallel machines supported by Charm++. `charmrun` is copied to a directory where an AMPI program is built using `charmc`. It takes a command line parameter specifying number of processors, and the name of the program followed by AMPI options (such as number of VPs to create, and the stack size of every VP) and the program arguments. A typical invocation of an AMPI program `pgm` with `charmrun` is:

> charmrun +p16 ./pgm +vp64

Here, the AMPI program `pgm` is run on 16 physical processors with 64 total virtual ranks (which will be mapped 4 per processor initially).

To run with load balancing, specify a load balancing strategy. If Address Space Layout Randomization is enabled on your target system, you may need to add the flag `+isomalloc.sync` when running with migration. You can also specify the size of user-level thread’s stack using the `+tcharm_stacksize` option, which can be used to decrease the size of the stack that must be migrated, as in the following example:

> charmrun +p16 ./pgm +vp128 +tcharm_stacksize 32K +balancer RefineLB