Paradyn Parallel Performance Tools

# MRNet API Programmer's Guide

Release 3.0
August 2010

MRNet Project
www.paradyn.org/mrnet

Paradyn Project
www.paradyn.org

Computer Sciences Department
University of Wisconsin
Madison, WI 53706-1685
paradyn@cs.wisc.edu

# 1. INTRODUCTION

MRNet is a customizable, high-throughput communication software system for parallel tools and applications with a master/slave architecture. MRNet reduces the cost of these tools' activities by incorporating a tree-based overlay network (TBON) of processes between the tool's front-end and back-ends. MRNet uses the TBON to distribute many important tool communication and computation activities, reducing analysis time and keeping tool front-end loads manageable.

MRNet-based tools send data between front-end and back-ends on logical flows of data called streams. MRNet internal processes use filters to synchronize and aggregate data sent to the tool's front-end. Using filters to manipulate data in parallel as it passes through the network, MRNet can efficiently compute averages, sums, and other more complex aggregations on back-end data.

Several features make MRNet especially well-suited as a general facility for building scalable parallel tools:

- *Flexible organization*. MRNet does not dictate the organization of the TBON. MRNet process organization is specified in a configuration file that can specify common network overlays like k-ary and k-nomial trees, or custom layouts tailored to the system(s) running the tool. For example, MRNet internal processes can be allocated to dedicated system nodes or co-located with tool back-end and application processes.
- *Scalable, flexible data aggregation*. MRNet's built-in filters provide efficient computation of averages, sums, concatenation, and other common data reductions. Custom filters can be loaded dynamically into the network to perform tool-specific aggregation operations.
- *High-bandwidth communication.* MRNet transfers data within the tool system using an efficient, packed binary representation. Zero-copy data paths are used whenever possible to reduce the cost of transferring data through internal processes.
- *Scalable multicast.* As the number of back-ends increases, serialization when sending control requests limits the scalability of existing tools. MRNet supports efficient message multicast to reduce the cost of issuing control requests from the tool front-end to its back-ends.
- *Multiple concurrent data channels*. MRNet supports multiple logical streams of data between tool components. Data aggregation and message multicast takes place within the context of a data stream, and multiple operations (both upward and downward) can be active simultaneously.

# 2. ABSTRACTIONS

The MRNet distribution has two main components: `libmrnet`, a library that is linked into a tool's front-end and back-end components, and `mrnet_commnode`, a program that runs on intermediate nodes interposed between the application front-end and back-ends. `libmrnet` exports an API (see **"C++ API Reference" on page 8**) that enables I/O interaction between the front-end and groups of back-ends via MRNet. The primary purpose of `mrnet_commnode` is to distribute data processing functionality across multiple computer hosts and to implement efficient and scalable group communications. In addition, there is another component, `libmrnet_lightweight`, which exports an API (see **"C API Reference" on page 23**) that enables I/O interaction between the front-end and groups of "lightweight" back-ends via MRNet. The following sub-sections describe the lower-level components of the MRNet API in more detail.

## 2.1 End-Points

An MRNet end-point represents a tool or application process. In particular, they represent the back-end processes (i.e., leaf processes) in the tree overlay. The front-end can communicate in a unicast or multicast fashion with one or more of these end-points as described below.

## 2.2 Communicators

MRNet uses communicators to represent groups of end-points. Like communicators in MPI, MRNet communicators provide a handle that identifies a set of end-points for point-to-point, multicast or broadcast communications. MPI applications typically have a non-hierarchical layout of potentially identical processes. In contrast, MRNet enforces a tree-like layout of all processes, rooted at the front-end. Accordingly, MRNet communicators are created and managed by the front-end, and communication is only allowed between a front-end and its back-ends. As such, back-ends cannot interact with each other directly using the MRNet API.

## 2.3 Streams

A stream is a logical channel that connects the front-end to the end-points of a communicator. All MRNet communication uses the stream abstraction. Streams carry data packets downstream, from the front-end toward the back-ends, and upstream, from the back-ends toward the front-end. Streams are expected to carry data of a specific type, allowing data aggregation operations to be associated with a stream. The type is specified using a format string (see **Appendix E: "Format Strings" on page 40**) similar to those used in C formatted I/O primitives (e.g., a packet whose data is described by the format string "`%d %d %f %s`" contains two integers followed by a float then a character string). MRNet expands the standard format string specification to allow for description of arrays.

## 2.4 Filters

Data aggregation is the process of merging multiple input data packets and transforming them into one or more output packets. Though it is not necessary for the aggregation to result in less or even different data, aggregations that reduce or modify data values are most common. MRNet uses data filters to aggregate data packets. Filters specify an operation to perform and the type of

the data expected on the bound stream. Filter instances are bound to a stream at stream creation. MRNet uses two types of filters: synchronization filters and transformation filters. Synchronization filters organize data packets from downstream nodes into synchronized waves of data packets, while transformation filters operate on the synchronized data packets yielding one or more output packets. A distinction between synchronization and transformation filters is that synchronization filters are independent of the packet data type, but transformation filters operate on packets of a specific type.

Synchronization filters operate on data flowing upstream in the network, receiving packets one at a time and outputting packets only when the specified synchronization criteria has been met. Synchronization filters provide a mechanism to deal with the asynchronous arrival of packets from children nodes. The synchronizer collects packets and typically aligns them into waves, passing an entire wave onward at the same time. Therefore, synchronization filters do no data transformation and can operate on packets in a type-independent fashion. MRNet currently supports three synchronization modes:

- *Wait For All*: wait for a complete wave (i.e., a packet from every child node) before producing output packets (SFILTER_WAITFORALL)
- *Do Not Wait:* output packets immediately (SFILTER_DONTWAIT)
- *Timeout* : output packets after 'timeout' milliseconds (SFILTER_TIMEOUT), or when a complete wave has been accumulated. The timeout period begins upon receipt of the first packet since the filter last produced output. The timeout value in milliseconds can be set using `Stream::set_FilterParameters`. Note that this timeout value is used at each level of the tree - a timeout value of 100ms combined with a tree of depth three should produce outputs at the front-end approximately 300ms after a packet is sent from a back-end. The default timeout value is 0ms. If you use SFILTER_TIMEOUT without setting a non-zero timeout value, it will behave similar to SFILTER_DONTWAIT.

Transformation filters can be used on both upstream and downstream data flows. Transformation filters input a group of synchronized packets, and combine data from multiple packets by performing an aggregation that yields one or more new data packets. Data packets produced by a transformation filter can be forwarded in either direction on a Stream by placing them in the appropriate output set. Since transformation filters are expected to perform computational operations on data packets, there is a type requirement for the data packets to be passed to this type of filter: the data format string of the stream's packets and the filter must be the same. Transformation operations must be synchronous, but are able to maintain state from one execution to the next. MRNet provides several transformation filters that should be of general use:

- *Basic scalar operations on characters/integers/floats*: minimum (TFILTER_MIN), maximum (TFILTER_MAX), summation (TFILTER_SUM), average (TFILTER_AVG)
- *Concatenation*: operation that inputs n scalars and outputs a vector of length n of the same base type (TFILTER_ARRAY_CONCAT)

**Appendix D: "Adding New Filters" on page 38** describes facilities for adding new user-defined transformation and synchronization filters.

# 3. A SIMPLE EXAMPLE

## 3.1 The MRNet Interface

A complete description of the MRNet API is in **"C++ API Reference" on page 8** and **"C API Reference" on page 23**. This section offers a brief overview only. Using `libmrnet`, a tool can leverage a system of internal processes, instances of the `mrnet_commnode` program, as a communication substrate. After instantiation of the MRNet network (discussed in "MRNet Instantiation" on page 6), the front-end and back-end processes are connected by the internal processes. The connection topology and host assignment of these processes is determined by a configuration file, thus the geometry of MRNet's process tree can be customized to suit the physical topology of the underlying hardware resources. While MRNet can generate a variety of standard topologies, users can easily specify their own topologies; see **Appendix C: "Process-Tree Topologies" on page 36** for further discussion.

The MRNet API contains Network, EndPoint, Communicator, and Stream objects that a tool's front-end and back-end use for communication. The Network object is used to instantiate the MRNet network and access EndPoint objects that represent available tool back-ends. The Communicator object is a container for groups of end-points, and Stream objects are used to send data to the EndPoints in a Communicator.

```
1    front_end_main(...) {
2          Network * net;
3          Communicator * comm;
4          Stream * stream;
5          PacketPtr packet;
6          int tag = FirstApplicationTag;
7          float result;
8
9          net = Network::CreateNetworkFE(topology_file, backend_exe, argv);
10         comm = net->get_BroadcastCommunicator( );
11         stream = net->new_Stream(comm, TFILTER_SUM, SFILTER_WAITFORALL);
12         stream->send(tag, "%d", SUM_INIT);
13         stream->recv(&tag, packet)
14         packet->unpack("%f", &result);
15   }
```

**Figure 1: MRNet Front-End Sample Code**

A simplified version of code from an example tool front-end is shown in **Figure 1: MRNet Front-End Sample Code**. In the front-end code, after some variable definitions in lines 2-6, an instance of the MRNet network is created on line 9 using the topology specified in topology_file. In line 10, the newly created Network object is queried for an auto-generated broadcast communicator that contains all available end-points. In line 11, this Communicator is used to establish a Stream that will use a built-in filter that finds the summation of the data sent upstream. The front-end then sends one or more initialization messages to the backends; in our example code on line 12, we broadcast an integer initializer on the new stream. The tag parameter is an application-specific value denoting the nature of the message being transmitted. After the send operation, the

front-end performs a blocking stream receive at line 13. This call returns a tag and a packet. Finally, line 14 calls unpack to deserialize the floating point value contained in packet.

```
1    back_end_main(int argc, char** argv) {
2          Stream * stream;
3          PacketPtr packet;
4          int val, tag;
5          float random_float = (float) random( );
6
7          Network * net = Network::CreateNetworkBE(argc,argv);
8          net->recv(&tag, packet, &stream);
9          packet->unpack("%d", &val );
10         if( val == SUM_INIT )
11               stream->send(tag, "%f", random_float);
12   }
```

**Figure 2: MRNet Back-End Sample Code**

**Figure 2: MRNet Back-End Sample Code** shows the code for the back-end that reciprocates the actions of the front-end. Each tool back-end first connects to the MRNet network in line 5, using the back-end version of the Network constructor that receives its arguments via the program argument vector (`argc/argv`). While the front-end makes a stream-specific receive call, the back-ends use a stream-anonymous network receive that returns the tag sent by the front-end, the packet containing the actual data sent, and a stream object representing the stream that the front-end has established. Finally, each back-end sends a scalar floating point value upstream toward the front-end.

A complete example of MRNet code can be found below in **Appendix B: "A Complete Example: Integer Addition" on page 31**.

## 3.2 MRNet Instantiation

While conceptually simple, creating and connecting the internal processes is complicated by interactions with the various job scheduling systems. In the simplest environments, we can launch jobs manually using facilities like `rsh` or `ssh`. In more complex environments, it is necessary to submit all requests to a job management system. In this case, we are constrained by the operations provided by the job manager (and these vary from system to system). We currently support two modes of instantiating MRNet-based tools.

In the first mode of process instantiation, MRNet creates the internal and back-end processes, using the specified MRNet topology configuration to determine the hosts on which the components should be located. First, the front-end consults the configuration and uses a remote shell program to create internal processes for the first level of the communication tree on the appropriate hosts. Upon instantiation, the newly created processes establish a network connection to the process that created it. The first activity on this connection is a message from parent to child containing the portion of the configuration relevant to that child. The child then uses this information to begin instantiation of the sub-tree rooted at that child. When a sub-tree has been established, the root of that sub-tree sends a report to its parent containing the end-points accessible via that sub-tree. Each internal node establishes its children processes and their respective connections

sequentially. However, since the various processes are expected to run on different compute nodes, sub-trees in different branches of the network are created concurrently, maximizing the efficiency of network instantiation.

In the second mode of process instantiation, MRNet relies on a process management system to create some of the MRNet processes. This mode accommodates tools that require their back-ends to create, monitor, and control other processes. For example, IBM's POE uses environment variables to pass information, such as the process' rank within the application's global MPI communicator, to the MPI run-time library in each application process. In cases like this, MRNet cannot provide back-end processes with the environment necessary to start MPI application processes. As a result, MRNet creates its internal processes recursively as in the first instantiation mode, but does not instantiate any back-end processes. MRNet then waits for the tool back-ends to be started by the process management system to ensure they have the environment needed to create application processes successfully. To allow back-ends to connect to the MRNet network, information such as process host names and connection port numbers must be provided to the back-ends. This information can be provided via the environment, using shared filesystems or other information services as available on the target system. To collect the necessary information, the front-end can use the MRNet API methods for discovering the network topology details. This mode of process instantiation is referred to as "back-end attach mode". We show how to construct a tool that requires back-end attach in `$MRNET_ROOT/Examples/NoBackEndInstantiation`.

# 4. THE MRNET API

Standard MRNet relies on the back-end nodes supporting C++ libraries. However, we have also created a lightweight backend library with a pure C interface. The instantiation process is the same and both methods of process instantation are supported, although the API interface is slightly different.

## 4.1 C++ API Reference

All classes are included in the `MRN` namespace. For this discussion, we do not explicitly include reference to the namespace; for example, when we reference the class `Network`, we are implying the class `MRN::Network`.

In MRNet, there are five top-level classes: `Network`, `NetworkTopology`, `Communicator`, `Stream`, and `Packet`. The `Network` class primarily contains methods for instantiating and destroying MRNet process trees. The `NetworkTopology` class represents the interface for discovering details about the topology of an instantiated network. Application back-ends are referred to as end-points, and the `Communicator` class is used to reference a group of end-points. A communicator is used to establish a `Stream` for unicast, multicast, or broadcast communications via the MRNet infrastructure. The `Packet` class encapsulates the data packets that are sent on a stream. The public members of these classes are detailed below.

### 4.1.1 Class Network

The corresponding lightweight backend API class is **"Class Network" on page 23**.

```
Network * Network::CreateNetworkFE(
      const char * topology,
      const char * backend_exe,
      const char ** backend_argv,
      const std::map< std::string, std::string>* attrs=NULL,
      bool rank_backends = true,
      bool using_memory_buffer = false );
```

The front-end constructor method that is used to instantiate the MRNet process tree. `topology` is the path to a configuration file that describes the desired process tree topology.

`backend_exe` is the path to the executable to be used for the application's back-end processes. `backend_argv` is a null terminated list of arguments to pass to the back-end application upon creation. If `backend_exe` is NULL, no back-end processes will be started, and the leaves of the topology specified by topology will be instances of `mrnet_commnode`.

`attrs` is a pointer to a map of attribute-value string pairs. `attrs` is currently used only on Cray XT platforms; for other platforms, it takes the default value of NULL. On Cray XT, when communication or back-end processes of the MRNet tree are to be co-located with application processes, `attrs` must contain a string pair that maps the string "apid" to a valid

ALPS application id string, which is a unique identifier for an application process started using ALPS `aprun`.

`rank_backends` indicates whether the back-end process ranks should begin at 0, similar to MPI rank numbering, and defaults to `true`.

If `using_memory_buffer` is set to `true` (default is `false`), the topology parameter is actually a pointer to a memory buffer containing the specification, rather than the name of a file.

When this function completes without error, all MRNet processes specified in the topology will have been instantiated. You may use `Network::has_error` to check for successful completion. The explicit use of the `Network` constructor is now deprecated.

**`Network * Network::CreateNetworkBE( int argc, char ** argv );`**

The back-end constructor method that is used when the process is started due to a front-end network instantiation. MRNet automatically passes the necessary information to the back-end process using the program argument vector (`argc/argv`) by inserting it after the user-specified arguments. The explicit use of the `Network` constructor is now deprecated.

In the "back-end attach" mode of network instantiation, where the back-end is not launched directly by MRNet, the back-end program must construct a suitable argument vector. Typically, the front-end program will obtain information about the leaf `mrnet_commnode` processes using the `NetworkTopology` class, and pass this information to back-ends using external communication channels (e.g., a shared file system). The back-ends choose a leaf process as a parent, and use that parent's host, port, and rank information to attach. Each back-end must choose a unique value for its local rank; this value must be larger than any of the ranks of the processes in the existing network. The following code shows how to construct a valid argument vector:

```
char parHostname[64], myHostname[64], parPort[6], parRank[6], myRank[6];
// fill parent data here using info from front-end
gethostname( myHostname, 64 );
sprintf( myRank, "%d", <unique rank> );
be_argc = 6;
char* be_argv[be_argc];
be_argv[0] = argv[0];
be_argv[1] = parHostname;
be_argv[2] = parPort;
be_argv[3] = parRank;
be_argv[4] = myHostname;
be_argv[5] = myRank;
```

**`void Network::~Network();`**

`Network::~Network` tears down the MRNet process tree when the `Network` object is deleted. Note that `Network::shutdown_Network` is deprecated.

**void Network::waitfor_ShutDown();**

Network::waitfor_ShutDown can be used by back-ends to block until the network has been shut down by the front-end.

**bool Network::is_ShutDown();**

Back-ends use this method to query if the network has been shut down; returns true if it has been shut down, false otherwise.

**bool Network::set_FailureRecovery( bool enable );**

Network::set_FailureRecovery is used by a front-end to control whether internal communication processes and back-ends will automatically re-connect to a new parent when their parent terminates unexpectedly. By default, failure recovery is enabled and processes will reconnect. Call this method with enable set to false to turn off automatic failure recovery. This method returns true if the setting has been applied successfully, false otherwise.

**bool Network::has_Error( );**

Network::has_error returns true if an error has occured during the last call to a Network method. Network::print_error can be used to print a message describing the exact error.

**ErrorCode Network::get_Error( );**

Network::get_Error returns an ErrorCode for an error that occured during the last call to a Network method. Network::get_ErrorStr can be used to retrieve a message string describing the error.

**const char * Network::get_ErrorStr( ErrorCode code );**

Network::get_ErrorStr returns a character string describing the error indicated by code.

**void Network::print_error( const char * error_msg );**

Network::print_error prints a message to stderr describing the last error encountered during a Network method. It first prints the null-terminated string error_msg followed by a colon, then the actual error message followed by a newline.

**std::string Network::get_LocalHostName();**

Network::get_LocalHostName returns the name of the host on which the local MRNet process is running.

**Port Network::get_LocalPort();**

Network::get_LocalPort returns the listening port of the local MRNet process.

**Rank Network::get_LocalRank();**

Network::get_LocalRank returns the rank of the local MRNet process.

```
int Network::load_FilterFunc( const char * so_file, const char* func );
```

This method, used for loading new filter operations into the Network is conveniently similar to the conventional `dlopen` facilities for opening a shared object and dynamically loading symbols defined within.

`so_file` is the path to a shared object file that contains the filter function to be loaded and `func_name` is the name of the function to be loaded.

On success, `Network::load_FilterFunc` returns the id of the newly loaded filter which may be used in subsequent calls to `Network::new_Stream`. A value of -1 is returned on failure.

```
int Network::recv(
      int * tag,
      PacketPtr & packet,
      Stream ** stream,
      bool blocking = true );
```

`Network::recv` is used to invoke a stream-anonymous receive operation. Any packet available (i.e., addressed to any stream) will be returned (in roughly FIFO order).

`otag` will be filled in with the integer tag value that was passed by the corresponding `Stream::send` operation. `packet` is the packet that was received. A pointer to the stream to which the packet was addressed will be returned in `stream`.

`blocking` is used to signal whether this call should block or return if data is not immediately available; it defaults to a blocking call.

A return value of -1 indicates an error, 0 indicates no packets were available, and 1 indicates success.

```
bool Network::enable_PerformanceData(
      perfdata_metric_t metric,
      perfdata_context_t context );
```

`Network::enable_PerformanceData` uses `Stream::enable_PerformanceData` to start the recording of performance data of the specified `metric` type for the given `context` on all streams. Returns `true` on success, `false` otherwise. **Appendix F: "MRNet Stream Performance Data" on page 41** describes the supported metric and context types. See `Stream::enable_PerformanceData` for additional details.

```
bool Network::disable_PerformanceData(
      perfdata_metric_t metric,
      perfdata_context_t context );
```

`Network::disable_PerformanceData` stops the recording of performance data of the specified `metric` type for the given `context` on all streams. Returns `true` on success, `false` otherwise. See `Stream::disable_PerformanceData` for additional details.

```
bool Network::collect_PerformanceData(
      std::map< int, rank_perfdata_map > & results,
      perfdata_metric_t metric,
      perfdata_context_t context,
      int aggr_filter_id = TFILTER_ARRAY_CONCAT );
```

`Network::collect_PerformanceData` collects the performance data of the specified `metric` type for the given `context` on all streams. The performance data of each stream is passed through the transformation filter identified by `aggr_filter_id`. The data for all streams is stored within the map `results`, keyed by stream identifier. Returns `true` on success, `false` otherwise. See `Stream::collect_PerformanceData` for additional details.

```
void Network::print_PerformanceData(
      perfdata_metric_t metric,
      perfdata_context_t context );
```

`Network::enable_PerformanceData` uses `Stream::print_PerformanceData` to print recorded performance data of the specified `metric` type for the given `context` on all streams. Data is printed to the MRNet log files. See `Stream::print_PerformanceData` for additional details.

```
unsigned int Network::num_EventsPending( );
```

`Network::num_EventsPending` returns the number of pending events available for retrieval using `Network::next_Event`.

```
Event * Network::next_Event( );
```

This method returns a pointer the next pending `Event`, or `NULL` if no events are available. Each event has an associated `EventClass`, one of `Event::DATA_EVENT`, `Event::TOPOLOGY_EVENT`, or `Event::ERROR_EVENT`, that can be queried using `Event::get_Class`. Similarly, each event has an associated `EventType` that can be queried using `Event::get_Type`.

```
void Network::clear_Events( );
```

This method clears all pending events.

```
bool Network::register_EventCallback(
        EventClass eclass,
        EventType etyp,
        evt_cb_func cb_func,
        void * cb_func_data );
```

`Network::register_EventCallback` allows users to register a callback function to be called when events are generated.

`eclass` should be set to one of `Event::DATA_EVENT`, `Event::TOPOLOGY_EVENT`, or `Event::ERROR_EVENT`.

`etyp` should be set to either `Event::EVENT_TYPE_ALL`, to have the function called when any event within the specified `EventClass` occurs, or one of the valid class-specific `EventType` values (see the classes `DataEvent`, `TopologyEvent`, and `ErrorEvent` in `"mrnet/Event.h"` for the class-specific types).

The type `evt_cb_func` is defined as '`void (*evt_cb_fn)( Event* e, void* cb_data )`'. All user-defined callback functions must use the same function prototype. When an event occurs, all callbacks registered for that type of event will be called. Each function is passed a pointer to the `Event`, and the value of the auxiliary data pointer `cb_func_data` given at registration, which may be `NULL`.

```
void Network::remove_EventCallback(
        evt_cb_func cb_func,
        EventClass eclass,
        EventType etyp );
```

This method removes `cb_func` from the list of functions to be called for the specified `EventClass` and `EventType`. If `eclass` is given as `Event::EVENT_CLASS_ALL`, the function will be removed for all events. `etyp` can be given as `Event::EVENT_TYPE_ALL` to remove the function for all types of events in the given `eclass`.

```
void Network::remove_EventCallbacks(
        EventClass eclass,
        EventType etyp );
```

This method removes all functions to be called for the specified `EventClass` and `EventType`. If `eclass` is given as `Event::EVENT_CLASS_ALL`, all callback functions will be removed for all events. `etyp` can be given as `Event::EVENT_TYPE_ALL` to remove all functions registered for all types of events in the given `eclass`.

**int Network::get_EventNotificationFd( EventClass eclass );**

`Network::get_EventNotificationFd` returns a file descriptor that can be used with `select` or `poll` to receive notification of interesting DATA, TOPOLOGY, or ERROR events.

`eclass` should be set to one of `Event::DATA_EVENT`, `Event::TOPOLOGY_EVENT`, or `Event::ERROR_EVENT`. `Event::DATA_EVENT` can be used by both front-end and back-end processes to provide notification that one or more data packets have been received. `Event::TOPOLOGY_EVENT` and `Event::ERROR_EVENT` can only be used by front-end processes, and provide notification when the front-end observes a change in network topology or an error, respectively.

When the file descriptor has data available (for reading), you should call `Network::clear_EventNotificationFd` before taking action on the notification. When notifications are no longer needed, use `Network::close_EventNotificationFd`.

NOTE: this functionality is not available on Windows platforms.

**void Network::clear_EventNotificationFd( EventClass eclass );**

This method resets the event notification file descriptor returned from `Network::get_EventNotificationFd`. `eclass` should be set to one of `Event::DATA_EVENT`, `Event::TOPOLOGY_EVENT`, or `Event::ERROR_EVENT`.

NOTE: this functionality is not available on Windows platforms.

**void Network::close_EventNotificationFd( EventClass eclass );**

This method closes the event notification file descriptor returned from `Network::get_EventNotificationFd`. `eclass` should be set to one of `Event::DATA_EVENT`, `Event::TOPOLOGY_EVENT`, or `Event::ERROR_EVENT`.

NOTE: this functionality is not available on Windows platforms.

**bool is_LocalNodeChild( ) const;**

**bool is_LocalNodeParent( ) const;**

**bool is_LocalNodeInternal( ) const;**

**bool is_LocalNodeFrontEnd( ) const;**

**bool is_LocalNodeBackEnd( ) const;**

These methods return `true` if the local process is of the specified type, `false` otherwise.

### 4.1.2 Class NetworkTopology

Instances of `NetworkTopology` are network specific, so they are created when a `Network` is instantiated. MRNet API users should not need to create their own `NetworkTopology` instances.

The corresponding lightweight backend API class is **"Class NetworkTopology" on page 24**.

**NetworkTopology * Network::get_NetworkTopology( );**

`Network::get_NetworkTopology` is used to retrieve a pointer to the underlying `NetworkTopology` instance of a `Network`.

**unsigned int NetworkTopology::get_NumNodes( );**

This method returns the total number of nodes in the tree topology, including front-end, internal, and back-end processes.

**NetworkTopology::Node * NetworkTopology::find_Node( Rank node_rank );**

This method returns a pointer to the tree node with rank equal to `node_rank`, or NULL if not found.

**NetworkTopology::Node * NetworkTopology::get_Root( );**

This method returns a pointer to the root node of the tree, or NULL if not found.

**void NetworkTopology::get_Leaves(**
**std::vector<NetworkTopology::Node * > & leaves );**

This method fills the `leaves` vector with pointers to the leaf nodes in the topology. In the case where back-end processes are not started when the network is instantiated, a front-end process can use this function to retrieve information about the leaf internal processes to which the back-ends should attach.

**void NetworkTopology::get_BackEndNodes(**
**std::set< NetworkTopology::Node * > & nodes );**

This method fills a set with pointers to all back-end process tree nodes. Note that this method is unsafe to use while the network topology is in flux, as is the case during the "back-end attach" mode of MRNet tree instantiation.

**void NetworkTopology::get_ParentNodes(**
**std::set<NetworkTopology::Node * > & nodes );**

This method fills a set with pointers to all tree nodes that are parents (i.e., those nodes having at least one child).

**void NetworkTopology::get_OrphanNodes(**
**std::set< NetworkTopology::Node * > & nodes );**

This method fills a set with pointers to all tree nodes that have no parent due to a failure.

```
void NetworkTopology::get_TreeStatistics(
      unsigned int & num_nodes,
      unsigned int & depth,
      unsigned int & min_fanout,
      unsigned int & max_fanout,
      double & avg_fanout,
      double & stddev_fanout );
```

This method provides users statistics about the tree topology by setting the passed parameters.

`num_nodes` is the total number of tree nodes (same as the value returned by `NetworkTopology::get_NumNodes`), `depth` is the depth of the tree (i.e., the maximum path length from root to any leaf), `min_fanout` is the minimum number of children of any parent node, `max_fanout` is the maximum number of children of any parent node, `avg_fanout` is the average number of children across all parent nodes, and `stddev_fanout` is the standard deviation in number of children across all parent nodes.

```
void NetworkTopology::print_TopologyFile( const char * filename );
```

This method will create (or overwrite) the specified topology file `filename` using the current state of this `NetworkTopology` object.

```
void NetworkTopology::print_DOTGraph( const char * filename );
```

This method will create (or overwrite) the specified dot graph file `filename` using the current state of this `NetworkTopology` object.

```
std::string NetworkTopology::Node::get_HostName();
```

This method returns a string identifying the hostname of the tree node.

```
Port NetworkTopology::Node::get_Port();
```

This method returns the listening port of the tree node.

```
Rank NetworkTopology::Node::get_Rank();
```

This method returns the unique rank of the tree node.

```
Rank NetworkTopology::Node::get_Parent();
```

This method returns the rank of the tree node's parent.

```
const std::set< NetworkTopology::Node * > &
NetworkTopology::Node::get_Children();
```

This method returns a set containing pointers to the children of the tree node, and is useful for navigating through the tree.

**unsigned int NetworkTopology::Node::get_NumChildren();**

This method returns the number of children of the tree node.

**unsigned int NetworkTopology::Node::find_SubTreeHeight();**

This method returns the height of the subtree rooted at this `NetworkTopology` node.

### 4.1.3 Class Communicator

Instances of `Communicator` are network specific, so their creation methods are functions of an instantiated `Network` object. There is no corresponding lightweight backend class.

**Communicator * Network::new_Communicator();**

This method returns a pointer to a new `Communicator` object. The object contains no end-points. Use `Communicator::add_EndPoint` to populate the communicator.

**Communicator * Network::new_Communicator( Communicator & comm );**

This method returns a pointer to a new `Communicator` object that contains the same set of end-points contained in `comm`.

```
Communicator * Network::new_Communicator(
      std::set< CommunicationNode * > & endpoints );
```

This method returns a pointer to a new `Communicator` object that contains the provided set of end-points.

**Communicator * Network::new_Communicator( std::set< Rank > & endpoints );**

This method returns a pointer to a new `Communicator` object that contains the set of end-points corresponding to processes whose ranks are provided in the passed set.

**Communicator * Network::get_BroadcastCommunicator( );**

This method returns a pointer to a broadcast `Communicator` containing all the end-points available in the system at the time the function is called.

Multiple calls to this method return the same pointer to the `Communicator` object created at network instantiation. If the network topology changes, as can occur when starting back-ends separately, the object will be updated to reflect the additions or deletions. This object should not be deleted.

**bool Communicator::add_EndPoint( Rank ep_rank );**

This method is used to add an existing end-point with rank `ep_rank` to the set contained by this `Communicator`.

If the set of end-points in the communicator already contains the new end-point, the function returns success. This method fails if there exists no end-point defined by `ep_rank`. This method returns `true` on success, `false` on failure.

```
bool Communicator::add_EndPoint( CommunicationNode * endpoint );
```

This method is similar to `add_EndPoint` above except that it takes a pointer to a `CommunicationNode` object instead of a rank. Success and failure conditions are exactly as stated above. This method returns `true` on success and `false` on failure.

```
const std::set< CommunicationNode * > & Communicator::get_EndPoints( );
```

Returns a reference to the set of `CommunicationNode` pointers comprising the end-points in the communicator.

```
std::string CommunicationNode::get_HostName( );
```

Returns a character string identifying the hostname of the end-point represented by this `CommunicationNode`.

```
Port CommunicationNode::get_Port( );
```

Returns the listening port of the end-point represented by this `CommunicationNode`.

```
Rank CommunicationNode::get_Rank( );
```

Returns the rank of the end-point represented by this `CommunicationNode`.

### 4.1.4 Class Stream

Instances of `Stream` are network specific, so their creation methods are functions of an instantiated `Network` object. The corresponding lightweight backend API class is **"Class Stream" on page 25**.

MRNet provides two types of streams, homegenous and heterogeneous. Homogenous streams use the same filters at every process participating in the stream, while heterogeneous streams allow for different filters to be used at different processes.

```
Stream * Network::new_Stream(
      Communicator * comm,
      int up_transfilter_id = TFILTER_NULL,
      int up_syncfilter_id = SFILTER_WAITFORALL,
      int down_transfilter_id = TFILTER_NULL );
```

This version of `Network::new_Stream` is used to create a homogenous `Stream` object attached to the end-points specified by a Communicator object `comm`.

`up_transfilter_id` specifies the transformation filter to apply to data flowing upstream from the application back-ends toward the front-end; the default value is `TFILTER_NULL`.

`up_syncfilter_id` specifies the synchronization filter to apply to upstream packets; the default value is `SFILTER_WAITFORALL`.

`down_transfilter_id` allows the user to specify a filter to apply to downstream data flows; the default value is `TFILTER_NULL`.

```
Stream * Network::new_Stream(
      Communicator * comm,
      std::string us_filters,
      std::string sync_filters,
      std::string ds_filters );
```

This version of `Network::new_Stream` is used to creae a heterogeneous `Stream` object. Users specify where packet filters are placed within the tree. Like the homogenous version of `Network::new_Stream`, the end-points are specified by the `comm` argument.

Strings are used to specify the filter placements, with the following syntax: "`filter_id => rank; [filter_id => rank; ...]`". If "`*`" is specified as the `rank` for an assignment, the filter will be assigned to all ranks that have not already been assigned. If a rank within `comm` is not assigned a filter, it will use the default filter. See `$MRNET_ROOT/Examples/HeterogeneousFilters` for an example of using `Network::new_Stream` to specify different filter types to be used within the same stream.

`us_filters` specifies the transformation filters to apply to data flowing upstream from the application back-ends toward the front-end.

`sync_filters` specifies the synchronization filters to apply to upstream packets.

`ds_filters` allows the user to specify filters to apply to downstream data flows.

Note that more than one filter should not be assigned to a single rank in any of these strings.

```
Stream * Network::get_Stream( unsigned int id );
```

Returns a pointer to the `Stream` identified by `id`, or `NULL` on failure.

```
unsigned int Stream::get_Id( );
```

Returns the integer identifier for this `Stream`.

```
const std::set< Rank > & Stream::get_EndPoints( );
```

Returns the set of end-point ranks for this `Stream`.

```
unsigned int Stream::size( );
```

Returns an integer indicating the number of end-points for this `Stream`.

```
bool Stream::is_ShutDown( );
```

For use by back-ends only, this method returns `true` if the front-end has deleted this `Stream`, `false` otherwise.

```
int Stream::send( int tag, const char * format_string, ... );
```

Invokes a data send operation on the calling `Stream`. `tag` is an integer that identifies the data in the packet. `format_string` is a format string describing the data in the packet (See **Appen-**

**dix E: "Format Strings" on page 40** for a full description.) On success, `Stream::send` returns 0; on failure -1.

NOTE: `tag` must have a value greather than or equal to the constant `FirstApplicationTag` defined by MRNet (`#include "mrnet/Types.h"`). Tag values less than `FirstApplicationTag` are reserved for internal MRNet use.

**`int Stream::flush();`**

Commits a flush of all packets currently buffered by this `Stream`. A successful return value of 1 indicates that all packets on the calling stream have been passed to the operating system for network transmission.

**`int Stream::recv( int * tag, PacketPtr & packet, bool blocking = true );`**

Invokes a stream receive operation. Packets received by the calling `Stream` will be returned by this method, one-at-a-time, in FIFO order.

`tag` will be filled in with the integer tag value that was passed by the corresponding `Stream::send` operation. `packet` is set to point to the received packet.

`blocking` determines whether the receive should block or return if data is not immediately available; it defaults to a blocking call.

A return value of -1 indicates an error, 0 indicates no packets were available, and 1 indicates success.

**`int Stream::get_DataNotificationFd( );`**

`Stream::get_DataNotificationFd` returns a file descriptor that can be used with `select` or `poll` to receive notification that data has arrived for a stream.

When the file descriptor has data available (for reading), you should call `Stream::clear_DataNotificationFd` before taking action on the notification. When notifications are no longer needed, use `Stream::close_DataNotificationFd`.

NOTE: this functionality is not available on Windows platforms.

**`void Stream::clear_DataNotificationFd( );`**

This method resets the data notification file descriptor returned from `Stream::get_DataNotificationFd`.

NOTE: this functionality is not available on Windows platforms.

**`void Stream::close_DataNotificationFd( );`**

This method closes the data notification file descriptor returned from `Stream::get_DataNotificationFd`.

NOTE: this functionality is not available on Windows platforms.

```
int Stream::set_FilterParameters(
      FilterType ftype,
      const char *format_str, ... ) const;
```

`Stream::set_FilterParameters` allows users to dynamically configure the operation of a stream transformation filter by passing arbitrary data in a similar fashion to `Stream::send`. When the filter executes, the passed data is available as a `PacketPtr` parameter to the filter, and the filter can extract the configuration settings.

ftype should be given as `FILTER_UPSTREAM_SYNC` to configure the synchronization filter, `FILTER_UPSTREAM_TRANS` for upstream transformation filter and `FILTER_DOWNSTREAM_TRANS` for downstream transformation filter.

```
int Stream::set_FilterParameters(
      const char *format_str,
      va_list params,
      FilterType ftype ) const;
```

This method is the same as the previous method except for the filter configuration parameters are given in the `va_list` form.

```
bool Stream::enable_PerformanceData(
      perfdata_metric_t metric,
      perfdata_context_t context );
```

`Stream::enable_PerformanceData` starts recording performance data for the specified `metric` type for the given `context`. Returns `true` on success, `false` otherwise. **Appendix F: "MRNet Stream Performance Data" on page 41** describes the metric and context types.

```
bool Stream::disable_PerformanceData(
      perfdata_metric_t metric,
      perfdata_context_t context );
```

`Stream::disable_PerformanceData` stops recording performance data for the specified `metric` type for the given `context`. Previously recorded data is not discarded, so that it can be retrieved with `Stream::collect_PerformanceData`. Users can enable/disable recording for a particular `metric` and `context` any number of times before collecting the results. Returns `true` on success, `false` otherwise.

```
bool Stream::collect_PerformanceData(
      rank_perfdata_map & results,
      perfdata_metric_t metric,
      perfdata_context_t context,
      int aggr_filter_id = TFILTER_ARRAY_CONCAT );
```

`Stream::collect_PerformanceData` collects the recorded performance data for the specified `metric` type for the given `context`. The collected data is returned in a

rank_perfdata_map, which associates individual node ranks to a std::vector< perf_data_t > containing the recorded data instances. After collection, the recorded data at each nodeis discarded. Returns true on success, false otherwise.

Users can aggregate the recorded data across nodes by specifying a transformation filter with aggr_filter_id. Note that only the built-in filter types of TFILTER_SUM, TFILTER_MIN, TFILTER_MAX, TFILTER_AVG, and TFILTER_ARRAY_CONCAT are supported. By default, performance data from each node is concatenated, and results contains every recorded data instance for each node. If the summary aggregation filters are used, results will contain a single associated pair. The rank for this pair is equal to -1 (number of aggregated ranks), and the vector contains one or more aggregated instances.

```
void Stream::print_PerformanceData(
      perfdata_context_t metric,
      perfdata_context_t context );
```

Stream::print_PerformanceData prints recorded performance data of the specified metric type for the given context. At each rank, the data is printed to the MRNet log files and then discarded.

### 4.1.5 Class Packet

A Packet encapsulates a set of formatted data elements sent on a stream. Packets are created using a format string (e.g., "%s %d" describes a null-terminated string followed by a 32-bit integer, and the packet is said to contain two *data elements*). MRNet front-end and back-end processes do not create instances of Packet; instead they are automatically produced from the formatted data passed to Stream::send. **Appendix E: "Format Strings" on page 40** contains the full listing of data types that can be sent in a Packet.

When receiving a packet via Stream::recv or Network::recv, the Packet instance is stored within a PacketPtr object. PacketPtr is a class based on the Boost library shared_ptr class, and helps with memory management of packets. A PacketPtr can be assumed to be equivalent to "Packet *", and all operations on packets require use of PacketPtr.

The corresponding lightweight backend API class is **"Class Packet" on page 26**.

```
int Packet::get_Tag( );
```

Returns the integer tag associated with this Packet.

```
unsigned short Packet::get_StreamId( );
```

Returns the stream id associated with this Packet.

```
const char * Packet::get_FormatString( );
```

Returns the character string specifying the data format of this Packet.

**void Packet::unpack( const char * format_string, ... );**

Extracts data contained within this `Packet` according to the `format_string`, which must match that of the packet. The function arguments following `format_string` should be pointers to the appropriate types of each data item. For string and array data types, new memory buffers to hold the data will be allocated using `malloc`, and it is the user's responsibility to `free` these strings and arrays. Note that for array data elements, an extra argument must be passed to hold each array's length.

**void Packet::set_DestroyData( bool destroy );**

This method can be used to tell MRNet whether or not to deallocate the string and array data members of a `Packet`. If `destroy` is `true`, string and array data members will be deallocated using `free` when the `Packet` destructor is executed - this assumes they were allocated using `malloc`. The default behavior for user-generated packets is not to deallocate (`false`). Turning on deallocation is useful in filter code that must allocate strings or arrays for output packets, which cannot be freed before the filter function returns.

## 4.2 C API Reference

In the MRNet lightweight back-end library, the MRNet C++ classes are mimicked for ease of use. With the exception of constructors/destructors, API calls in standard MRNet can be translated to their lightweight versions according to the following pattern:

```
return_type class::function_name( param1_type param1, ...);
```

translates to:

```
return_type class_function_name(
      class class_object,
      param1_type param1, ...);
```

### 4.2.1 Class Network

**Network_t * Network_CreateNetworkBE( int argc, char ** argv );**

The back-end constructor method. MRNet automatically passes the necessary information to the back-end process using the program argument vector (`argc/argv`) by inserting it after the user specified arguments. See **"Network * Network::CreateNetworkBE( int argc, char ** argv );" on page 9** for more information on the required arguments.

**void delete_Network_t( Network_t * network );**

`delete_Network_t` acts as a destructor for the `Network_t` object and cleans up internal structures before freeing the `Network_t` pointer.

**void Network_waitfor_ShutDown( Network_t * network );**

`Network_waitfor_ShutDown` blocks until the network has been shut down.

```
char Network_is_ShutDown( Network_t * network );
```

Returns `true` if the network has been shut down.

```
char* Network_get_LocalHostName( Network_t * network );
```

`Network_get_LocalHostName` returns the name of the host where the process is running.

```
Port Network_get_LocalPort( Network_t * network );
```

`Network_get_LocalPort` returns the listening port of the local process.

```
Rank Network_get_LocalRank( Network_t * network );
```

`Network_get_LocalRank` returns the rank of the local process.

```
int Network_recv(
      Network_t * network,
      int otag,
      Packet_t * packet,
      Stream_t * stream );
```

`Network_recv` is used to invoke a stream-anonymous receive operation. Any packet available (i.e., addressed to any stream) will be returned in roughly FIFO order.

`otag` will be filled in with the integer tag value that was passed by the corresponding `Stream_send` operation. `packet` is the packet that was received. A pointer to the `Stream_t` to which the packet was addressed will be returned in `stream`.

In standard MRNet, `Network::recv` had an additional parameter, `blocking`, to indicate whether this call should block or return if data is not immediately available. However, because the lightweight back-ends are single-threaded, there is only the blocking option; therefore this parameter has been omitted.

A return value of -1 indicates an error and 1 indictes a success.

### 4.2.2 Class NetworkTopology

```
NetworkTopology_t * Network_get_NetworkTopology( Network_t * network );
```

`Network_get_NetworkTopology` is used to retrieve a pointer to the underlying `NetworkTopology_t` instance within `network`.

```
Node_t * NetworkTopology_find_Node(
      NetworkTopology_t * net_top,
      Rank node_rank );
```

This method returns a pointer to the topology node with rank equal to `node_rank`, or `NULL` if no match is found.

```
Node_t * NetworkTopology_get_Root( NetworkTopology_t * net_top );
```

This method returns a pointer to the root node of the tree, or `NULL` if not found.

```
char * NetworkTopology_Node_get_HostName( Node_t * node );
```

This method returns a string identifying the hostname of the `node`.

```
Port NetworkTopology_Node_get_Port( Node_t * node );
```

This method returns the listening port of the `node`.

```
Rank NetworkTopology_Node_get_Rank( Node_t * node );
```

This method returns the rank of the `node`.

```
Rank NetworkTopology_Node_get_Parent( Node_t * node );
```

This method returns the rank of the node's parent.

```
unsigned int NetworkTopology_Node_find_SubTreeHeight( Node_t * node );
```

This method returns the height of the sub-tree rooted at the `node`.

### 4.2.3 Class Stream

```
void delete_Stream_t( Stream_t * stream );
```

`delete_Stream_t` acts as a destructor for the `Stream_t` object and cleans up internal structures before freeing the `Stream_t` pointer.

```
Stream_t * Network_get_Stream( Network_t * network, unsigned int id );
```

`Network_get_Stream` returns a pointer to a `Stream_t` identified by `id`, or `NULL` on failure.

```
unsigned int Stream_get_Id( Stream_t * stream );
```

This method returns the integer identifier for this `Stream_t`.

```
int Stream_send(
      Stream_t * stream,
      int tag,
      const char * format_string, ... );
```

This method sends data on `stream`. `tag` is an integer that identifies the data to be sent by the stream. `format_string` is a format string describing the types of the data elements (see **Appendix E: "Format Strings" on page 40** for a full description.) On success, `Stream_send` returns 0; on failure, -1.

NOTE: `tag` must have a value greater than or equal to the constant `FirstApplicationTag` defined by MRNet (`#include "mrnet_lightweight/Types.h"`). Tag values less than `FirstApplicationTag` are reserved for internal MRNet use.

```
int Stream_flush( Stream_t * stream );
```

This operation is currently not required in lightweight MRNet, as `Stream_send` will deliver the data for network transmission. This method will always return the value 1 for success.

```
int Stream_recv(
      Stream_t * stream,
      int * tag,
      Packet_t * packet );
```

`Stream_recv` invokes a stream-specific receive operation. Packets addressed to the passed `stream` will be returned, one-at-a-time, in FIFO order.

`tag` will be filled in with the integer tag value that was passed by the corresponding `Stream::send` operation. `packet` is the received `Packet_t`.

Unlike the standard C++ `Stream::recv`, `Stream_recv` will always block if data is not immediately available.

A return value of -1 indicates an error and 1 indicates success.

### 4.2.4 Class Packet

When receiving a packet, it is stored within a `Packet_t` object. Note that standard MRNet makes use of the `PacketPtr` object, which is based on the Boost library `shared_ptr` class. However, in the lightweight back-end library, pointers to `Packet_t` objects are used instead.

```
int Packet_get_Tag( Packet_t * packet );
```

This method returns the integer tag associated with `packet`.

```
unsigned short Packet_get_StreamId( Packet_t * packet );
```

This method returns the stream id associated with `packet`.

```
char* Packet_get_FormatString( Packet_t * packet );
```

This method returns the character string specifying the data format of `packet`.

```
void Packet_unpack(
      Packet_t * packet,
      const char * format_string, ... );
```

This method extracts data elements contained within `packet` according to the `format_string`, which must match that of `packet`. The function arguments following `format_string` should be pointers to the appropriate types of each data element. For string and array data types, new memory bufffers to hold the data will be allocated using `malloc`, and it is the user's responsibility to `free` these strings and arrays. Note that for array data elements, an extra argument must be passed to hold each array's length.

# APPENDIX A: BUILDING AND TESTING MRNET

For this discussion, `$MRNET_ROOT` is the location of the top-level directory of the MRNet distribution and `$MRNET_ARCH` is a string describing the platform (OS and architecture) as discovered by the configure process. For the installation instructions, it is assumed that the current working directory is `$MRNET_ROOT`.

## A.1: Supported Platforms and Compilers

MRNet has been developed to be highly portable; we expect it to run properly on all common Unix-based as well as Windows platforms. This being said, we have successfully built and tested MRNet on the following systems:

- Linux: x86, x86_64, power64, ia64, CrayXT
- Solaris 8, 9: sparc32
- AIX 5.2, 5.3: power32
- Windows: x86

## A.2: System Requirements

MRNet requires GNU make for building on UNIX/Linux systems. Our build system attempts to use native system compilers where available. For building on Windows systems, Visual Studio 2005 solution/project files are available, as are pre-compiled libraries and binaries.

## A.3: Build Configuration

MRNet uses GNU autoconf to discover the platform specific configuration parameters. The script that does this auto-configuration is called `configure`.

```
UNIX>  ./configure --help
```

shows all possible options of the command. Below, we display the MRNet-specific ones:

| | |
|---|---|
| `--enable-shared` | Build shared library versions of MRNet and XPlat |
| `--enable-debug` | Build MRNet and XPlat with debug information |
| `--enable-verbosebuild` | Show build actions (useful for debugging build problems) |
| `--with-startup=METHOD` | Choose tree instantiation method: "ssh" (default), or "cray_xt" (Cray XT systems)<br>For Cray XT only, when co-locating MRNet processes with an already running application launched using ALPS. |
| `--with-alpstoolhelp-lib=DIR` | Specify `DIR` as the absolute path to the directory containing the `libalps` library. |
| `--with-alpstoolhelp-inc=DIR` | Specify `DIR` as the absolute path to the directory containing the `libalps.h` header file. |

`./configure` without any options should give reasonable results, but the user may specify certain options. For example,

```
UNIX> ./configure CXX=g++
```

instructs the configure script to use `g++` for the C++ compiler.

## A.4: Compilation and Installation

To build MRNet:

```
UNIX>  make
```

After a successful build, the following files will be present:

- `$MRNET_ROOT/lib/$MRNET_ARCH/libmrnet`: MRNet API library
- `$MRNET_ROOT/lib/$MRNET_ARCH/libmrnet_lightweight`: MRNet lightweight back-end API library
- `$MRNET_ROOT/lib/$MRNET_ARCH/libxplat`: Cross-platform API library that exports platform dependent routines to MRNet
- `$MRNET_ROOT/lib/$MRNET_ARCH/libxplat_lightweight`: Cross-platform API library that exports platform dependent routines to MRNet, for use with the lightweight back-end library
- `$MRNET_ROOT/bin/$MRNET_ARCH/mrnet_commnode`: MRNet internal process executable

- `$MRNET_ROOT/bin/$MRNET_ARCH/mrnet_topgen`: MRNet topology file generator

To build the MRNet tests and examples:

```
UNIX>  make tests
UNIX>  make examples
```

The tests and examples consist of front-end and back-end programs, and custom filter libraries:

- `$MRNET_ROOT/bin/$MRNET_ARCH/*_[FE,FE_lightweight,BE,BE_lightweight]`: Front-end and back-end programs. Not all lightweight tests and examples require a separate front-end; in these cases, the standard front-end should be used with the lightweight back-end.
- `$MRNET_ROOT/bin/$MRNET_ARCH/mrnet_tests.sh`: A shell script that runs the test programs and checks for errors in an automated fashion.
- `$MRNET_ROOT/lib/$MRNET_ARCH/*Filter[s].so`: test and example filter libraries.

To install the MRNet components (i.e., the executables, libraries, and headers) to the directories specified during `configure`. If `--prefix` is not provided to `configure`, the default install locations are `/usr/local/{bin,lib,include}/`:

```
UNIX>  make install
```

To install the MRNet tests or examples:

```
UNIX>  make install-tests
UNIX>  make install-examples
```

If your system does not provide the C++ Boost headers (normally installed in `/usr/include/boost`), we provide the subset of Boost header files necessary for building MRNet. To install these headers:

```
UNIX>  make install-boost
```

## A.5: Testing the Code

The shell script, `mrnet_tests.sh` is placed in the binary directory with the other executables during the building of the MRNet tests as described above. This script can be used to run the MRNet test programs and check their output for errors. The script is used as follows:

```
UNIX> mrnet_tests.sh [ -l | -r <hostfile> | -a <hostfile> ]
                     [ -f <sharedobject> ] [ -lightweight ]
```

The `-l` flag is used to run all tests using only topologies that create processes on the local machine (note: running all the tests locally can take quite a while - anywhere from 30 minutes to an hour depending on the machine capabilities). The `-r` flag runs tests using remote machines specified in the file whose name immediately follows this flag. To run tests both locally and remotely, use the `-a` flag and specify a hostfile to use. To run the programs that test MRNet's ability to dynamically load filters, you must specify the absolute location of the shared object `test_DynamicFilters.so` produced when the tests were built. The `-lightweight` flag is used to run tests with both the standard back-ends and the lightweight back-ends.

## A.6: Bugs, Questions, and Comments

MRNet is maintained by the Paradyn Project, University of Wisconsin-Madison. Comments and feedback whether positive or negative are encouraged.

Please report bugs to paradyn@cs.wisc.edu. Bug fixes as patches are also welcome.

The MRNet webpage is http://www.paradyn.org/mrnet/

# APPENDIX B: A COMPLETE EXAMPLE: INTEGER ADDITION

The source code for the example contained in this appendix can be found in `$MRNET_ROOT/Examples/IntegerAddition`. All examples can be built by typing '`make`' from within the `$MRNET_ROOT/Examples` directory.

## B.1: A Complete MRNet Front-End

```
1    #include "mrnet/MRNet.h"
2    #include "IntegerAddition.h"
3    using namespace MRN;
4
5    int main(int argc, char **argv)
6    {
7          int send_val=32, recv_val=0;
8          int tag, retval;
9          PacketPtr p;
10         if( argc != 4 ){
11               printf("Usage: %s topology be_exe so_file\n", argv[0]);
12               exit(-1);
13         }
14         const char * topology_file = argv[1];
15         const char * be_exe = argv[2];
16         const char * so_file = argv[3];
17         const char * argv=NULL;
18
19         // Instantiates the MRNet internal nodes, using the organization
20         // in "topology_file," and the specified back-end application
21         Network * network = Network::CreateNetworkFE( topology_file,
22                                                       be_exe, &argv  );
23
24         // Make sure path to "so_file" is in LD_LIBRARY_PATH
25         int filter_id = network->load_FilterFunc( so_file, "IntegerAdd" );
26         if( filter_id == -1 ){
27               printf( "Network::load_FilterFunc() failure\n");
28               delete network;
29               return -1;
30         }
31
32         // A Broadcast communicator contains all the back-ends
33         Communicator * comm_BC = network->get_BroadcastCommunicator( );
34
35         // Create a stream that uses Integer_Add filter for aggregation
36         Stream * stream = network->new_Stream( comm_BC, filter_id,
37                                                SFILTER_WAITFORALL);
38         int num_backends = comm_BC->get_EndPoints().size();
39
```

```
40          // Broadcast a control message to back-ends to send us "num_iters"
41          // waves of integers
42          tag = PROT_SUM;
43          unsigned int num_iters=5;
44          if( stream->send( tag, "%d %d", send_val, num_iters ) == -1 ){
45                printf("stream::send() failure\n");
46                return -1;
47          }
48          if( stream->flush( ) == -1 ){
49                printf("stream::flush() failure\n");
50                return -1;
51          }
52
53          // We expect "num_iters" aggregated responses from all back-ends
54          for( unsigned int i=0; i<num_iters; i++ ){
55                retval = stream->recv(&tag, p);
56                if( retval == -1){
57                      //recv error
58                      return -1;
59                }
60                if( p->unpack( "%d", &recv_val ) == -1 ){
61                      printf("stream::unpack() failure\n");
62                      return -1;
63                }
64                if( recv_val != num_backends * i * send_val ){
65                      printf("Iteration %d: Failure!\n", i);
66                }
67                else{
68                      printf("Iteration %d: Success! recv_val(%d) == %d\n",
69                            i, recv_val, send_val*i*num_backends );
70                }
71          }
72
73          if(stream->send(PROT_EXIT, "") == -1){
74                printf("stream::send(exit) failure\n");
75                return -1;
76          }
77          if(stream->flush() == -1){
78                printf("stream::flush() failure\n");
79                return -1;
80          }
81
82          // Network destruction will exit all processes
83          delete network;
84          return 0;
85    }
```

## B.2: A Complete MRNet Back-End

```
1    #include "mrnet/MRNet.h"
2    #include "IntegerAddition.h"
3
4    using namespace MRN;
5
6    int main(int argc, char **argv)
7    {
8          Stream * stream=NULL;
9          PacketPtr p;
10         int tag=0, recv_val=0, num_iters=0;
11         Network * network = Network::CreateNetworkBE( argc, argv );
12         do {
13               if ( network->recv(&tag, p, &stream) != 1){
14                     fprintf(stderr, "stream::recv() failure\n");
15                     return -1;
16               }
17               switch(tag){
18               case PROT_SUM:
19                     p->unpack( "%d %d", &recv_val, &num_iters );
20
21                     // Send num_iters waves of integers
22                     for( unsigned int i=0; i<num_iters; i++ ){
23                           if( stream->send(tag, "%d", recv_val*i) == -1 ){
24                                 printf("stream::send(%%d) failure\n");
25                                 return -1;
26                           }
27                           if( stream->flush( ) == -1 ){
28                                 printf("stream::flush() failure\n");
29                                 return -1;
30                           }
31                     }
32                     break;
33               case PROT_EXIT:
34                     printf("Processing PROT_EXIT ...\n");
35                     break;
36               default:
37                     printf("Unknown Protocol: %d\n", tag);
38                     break;
39               }
40         } while ( tag != PROT_EXIT );
41
42         network->waitfor_ShutDown();
43         delete network;
44         return 0;
45   }
```

## B.3: A Complete MRNet Lightweight Back-End

```
1    #include "mrnet_lightweight/MRNet.h"
2    #include "IntegerAddition_lightweight.h"
3
4    int main(int argc, char **argv)
5    {
6          Stream_t * stream;
7          Packet_t* p = (Packet_t*)malloc(sizeof(Packet_t));
8          int tag=0, recv_val=0, num_iters=0;
9          Network_t * net = Network_CreateNetworkBE( argc, argv );
10         do {
11               if( Network_recv(net, &tag, p, &stream) != 1 ) {
12                     printf("BE: stream::recv() failure\n");
13                     break;
14               }
15               switch(tag) {
16               case PROT_SUM:
17                     Packet_unpack(p, "%d %d", &recv_val, &num_iters );
18                     // Send num_iters waves of integers
19                     unsigned int i;
20                     for( i=0; i<num_iters; i++ ) {
21                           printf("BE: Sending wave %u ...\n", i);
22                           if( Stream_send(stream,tag, "%d",
23                                                recv_val*i) == -1 ){
24                                 printf("BE: stream::send(%%d) failure\n");
25                                 tag = PROT_EXIT;
26                                 break;
27                           }
28                           if( Stream_flush(stream ) == -1 ){
29                                 printf("BE: stream::flush() failure\n");
30                                 tag = PROT_EXIT;
31                                 break;
32                           }
33                           sleep(2); // stagger sends
34                     }
35                     break;
36               case PROT_EXIT:
37                     if( Stream_send(stream,tag, "%d", 0) == -1 ) {
38                           printf("BE: stream::send(%%s) failure\n");
39                           break;
40                     }
41                     if( Stream_flush(stream ) == -1 ) {
42                           printf("BE: stream::flush() failure\n");
43                     }
44                     break;
45
```

```
46              default:
47                    fprintf(stderr, "BE: Unknown Protocol: %d\n", tag);
48                    tag = PROT_EXIT;
49                    break;
50              }
51        } while ( tag != PROT_EXIT );
52
53        if ( p != NULL )
54              free (p);
55
56        Network_waitfor_ShutDown(net);
57        delete_Network_t(net);
58        return 0;
59  }
```

## B.4: A MRNet Filter: Integer Addition

```
1   extern "C" {
2
3   //Must declare the format of data expected by the filter
4   const char * IntegerAdd_format_string = "%d";
5   void IntegerAdd( std::vector< PacketPtr > & packets_in,
6                    std::vector< PacketPtr > & packets_out,
7                    std::vector< PacketPtr > & /* packets_out_reverse */,
8                    void ** /* filter state */,
9                    PacketPtr & /* configuration parameters */,
10                   TopologyLocalInfo & /* local topology information */)
11  {
12        int sum = 0;
13
14        for( unsigned int i = 0; i < packets_in.size( ); i++ ) {
15              PacketPtr cur_packet = packets_in[i];
16              int val;
17              cur_packet->unpack("%d", &val);
18              sum += val;
19        }
20
21        PacketPtr new_packet ( new Packet(packets_in[0]->get_StreamId(),
22                                packets_in[0]->get_Tag(), "%d", sum ) );
23        packets_out.push_back( new_packet );
24  }
25
26  } /* extern "C" */
```

# APPENDIX C: PROCESS-TREE TOPOLOGIES

MRNet allows a tool to specify a node allocation and process connectivity tailored to its computation and communication requirements and to the system where the tool will run. Choosing an appropriate MRNet configuration can be difficult due to the complexity of the tool's own activity and its interaction with the system. This section describes how users define their own process topologies, and the `mrnet_topgen` utility provided by MRNet to facilitate generation of topology specification files.

## C.1: Topology File Format

The first parameter to `Network::CreateNetworkFE` is the name of an MRNet topology file. This file defines the topological layout of the front-end, internal, and back-end MRNet processes. In the syntax of the topology file, the `hostname:id` tuple represents a process with instance `id` running on `hostname`. It is important to note that the instance is used to distinguish processes on the same host, and does not reflect a port or process rank. A line in the topology file has the form:

```
hostname1:0 => hostname1:1 hostname1:2 ;
```

meaning a process on `hostname1` with instance id `0` has two children, with instance ids `1` and `2`, running on the same host. MRNet will parse the topology file without error if the file properly defines a tree in the mathematical sense (i.e. a tree must have a single root, no cycles, full connection, and no node can be its own descendant). Please note that the hostname associated with the root of the topology must match the host where the front-end process is run, or a run-time error will occur.

NOTE: A single topology specification line may span multiple physical lines to improve readability. For example:

```
hostname1:0 =>
                  hostname1:1
                  hostname1:2
                  ;
```

## C.2: An Example Topology File

```
nutmeg:0 => c01:0 c02:0 c03:0 c04:0 ;
c03:0 => c05:0 ;
c04:0 => c06:0 c07:0 c08:0 c09:0 ;
#       nutmeg
#          |
#       -------
#     /  |   |  \
#  c01  c02  c03  c04
#              |    |
#            c05    |
#               -------
#             /  |   |  \
#          c06  c07 c08  c09
```

## C.3: Topology File Generator

MRNet provides a topology generator program that supports generation of balanced and k-nomial topologies using simple specifications, and arbitrary topologies with a more complex specification that fully enumerates the topology fan-outs at each level of the tree. After MRNet is built, this program can be found at `$MRNET_ROOT/bin/$MRNET_ARCH/mrnet_topgen`. The usage can be obtained by running `mrnet_topgen` without arguments.

The generator program uses host lists that specify available hosts and the maximum number of processes to place on each host. The format for the host list is one host specification per line, where each specification is of the form hostname[:num_slots]. If the number of process slots is not given with the host, the generator program assumes only one process should be placed on the host. Additionally, if the same hostname is given on multiple lines, the number of processes that can be placed on the host is the summation of the process slot counts for all lines. An example host list file follows:

```
host1:4
host2
host3:2
host2
```

The above host list file results in three hosts being available for topology process placement, with `host1` having four available slots, and `host2` and `host3` each having two available slots. The generator program also allows users to specify different host lists for the placement of internal communication processes and back-end processes (see the `mrnet_topgen` usage for more information).

Some MRNet front-end programs may wish to generate a topology at run-time. To support this requirement, MRNet provides three API classes: `BalancedTree`, `KnomialTree`, and `GenericTree` that front-end programs may use directly to generate any topology that can be produced by `mrnet_topgen`. Not surprisingly, `mrnet_topgen` is built upon these classes, and its source code (`$MRNET_ROOT/tests/config_generator.C`) can serve as a reference for front-end programs wishing to use these classes.

# APPENDIX D: ADDING NEW FILTERS

## D.1: Defining an MRNet Filter

A filter function has the following signature:

```
void filter_name(
      std::vector< PacketPtr > & packets_in,
      std::vector< PacketPtr > & packets_out,
      std::vector< PacketPtr > & packets_out_reverse,
      void ** filter_state,
      PacketPtr & config_params,
      const TopologyLocalInfo & topol_info );
```

`packets_in` is a reference to a vector of packets serving as input to the filter function. `packets_out` is a reference to a vector into which output packets should be placed. When packets need to be sent in the reverse direction on the stream, `packets_out_reverse` can be used instead of `packets_out`. Both `packets_out` and `packets_out_reverse` can be used simultaneously. `filter_state` may be used to define and maintain state specific to a filter instance. `config_params` is a reference to a `PacketPtr` containing the current configuration settings for the filter instance, as can be set using `Stream::set_FilterParameters`. Finally, `topol_info` provides information that can be used by filters to determine the local process's placement in the topology, as well as access to the local `Network` object.

For each filter function defined in a shared object file, there must be a const char * symbol named by the string formed by the concatenation of the filter function name and the suffix "`_format_string`". For instance, if the filter function is named `my_filter_func`, the shared object must define a symbol "`const char* my_filter_func_format_string`". The value of this string will be the MRNet format string describing the format of data that the filter can operate on. A value of "" denotes that the filter can operate on data of arbitrary value.

## D.2: Fault-Tolerant Filters

MRNet automatically recovers from failures of internal tree processes (i.e., those processes that are not the front-end (root) or back-ends (leaves)). As part of the recovery, MRNet will extract filter state from the children of a failed process and pass that state as input to each child's newly chosen parent. If you have a filter that maintains persistent state using `filter_state`, you can provide an additional function within the shared object for your filter that MRNet may use to extract the state. The name of this extraction function should be the same as the filter name with the suffix "`_get_state`" appended. For instance, if the filter function is named my_filter_func, the extraction function should be named `my_filter_func_get_state`.

A filter state extraction function has the following signature:

```
PacketPtr filter_name_get_state( void ** filter_state, int stream_id );
```

`filter_state` is a pointer to the state defined by the filter for the stream identified by `stream_id`. This function should extract the necessary state and return a packet that can be passed

as input to the filter function. Since the packet will be processed as a normal input packet for the filter, it's format must match that expected by the filter. A fault-tolerant filter example is provided in `$MRNET_ROOT/Examples/FaultRecovery`.

## D.3: Creating and Using MRNet Filter Shared Object Files

Since we use the C facility `dlopen` to dynamically load new filter functions, all C++ symbols must be exported. That is, the filter function, format string, and state extraction function definitions must fall within the statements:

```
extern "C" {
```

and

```
}
```

The file that contains the filter functions and format strings must be compiled into a valid shared object. For example, with the GNU C++ compiler on ELF systems, the options "`-fPIC -shared -rdynamic`" can be used. Please refer to your compiler documentation for the appropriate options for other compilers. You can also refer to the setting of the `SOFLAGS` variable in `$MRNET_ROOT/Examples/Makefile` to see the options chosen during the configure process for compiling the Example filter libraries.

Additionally, front-end and back-end programs that will dynamically load filters must be built with compiler options that notify the linker to export global symbols (for GNU compilers, you can use "`-Wl,-E`").

# APPENDIX E: FORMAT STRINGS

After the % character that introduces a conversion, there may be a number of flag characters. `u`, `h`, l, and a are special modifiers meaning unsigned, short, long and array, respectivley. The full set of conversions are:

**Table 1: Format String Conversions**

| | |
|---|---|
| `c` | signed 8-bit character |
| `uc` | unsigned 8-bit character |
| `ac` | array of signed 8-bit characters |
| `auc` | array of unsigned 8-bit characters |
| `hd` | signed 16-bit decimal integer |
| `uhd` | unsigned 16-bit decimal integer |
| `ahd` | array of signed 16-bit decimal integers |
| `auhd` | array of unsigned 16-bit decimal integers |
| `d` | signed 32-bit decimal integer |
| `ud` | unsigned 32-bit decimal integer |
| `ad` | array of signed 32-bit decimal integers |
| `aud` | array of unsigned 32-bit decimal integers |
| `ld` | signed 64-bit decimal integer |
| `uld` | unsigned 64-bit decimal integer |
| `ald` | array of signed 64-bit decimal integers |
| `auld` | array of unsigned 64-bit decimal integers |
| `f` | 32-bit floating-point number |
| `af` | array of 32-bit floating-point numbers |
| `lf` | 64-bit floating-point number |
| `alf` | array of 64-bit floating-point numbers |
| `s` | null-terminated character string |
| `as` | array of null-terminated character strings |

NOTE: All array format specifiers, "`a*`", require an implicit length parameter of type `unsigned int` to be given: e.g., `send("%af", float_array_pointer, float_array_length)`

# APPENDIX F: MRNET STREAM PERFORMANCE DATA

The primary abstraction for communication and data processing within MRNet is the stream, so performance metrics and contexts are associated with actions on a particular stream.

All data is recorded as instances of a `perf_data_t`, which is simply a union type that can hold a 64-bit signed integer, a 64-bit unsigned integer, or a double precision float. As shown below, the data values can be accessed using the `i`, `u`, or `d` union fields.

```
typedef union { int64_t i; uint64_t u; double d; } perfdata_t;
```

Metrics define the type of performance data to record. The supported metric types are:

* `PERFDATA_MET_NUM_BYTES` : count of data bytes (uint64_t)
* `PERFDATA_MET_NUM_PKTS` : count of data packets (uint64_t)
* `PERFDATA_MET_ELAPSED_SEC` : elapsed seconds (double)
* `PERFDATA_MET_CPU_USR_PCT` : percent CPU utilization by user (double)
* `PERFDATA_MET_CPU_USR_PCT` : percent CPU utilization by system (double)
* `PERFDATA_MET_MEM_VIRT_KB` : virtual memory footprint in KB (double)
* `PERFDATA_MET_MEM_PHYS_KB` : physical memory footprint in KB (double)

Contexts specify when to record data. The supported contexts are:

* `PERFDATA_CTX_SEND` : when data is sent
* `PERFDATA_CTX_RECV` : when data is received
* `PERFDATA_CTX_FILT_IN` : before executing transformation filter
* `PERFDATA_CTX_FILT_OUT` : after executing transformation filter
* `PERFDATA_CTX_NONE` : when data is collected

**Table 2** shows which metrics are valid for a given context. When a metric is valid for only `CTX_FILT_OUT`, the metric is actually recorded through a combination of measurements at `CTX_FILT_IN` and `CTX_FILT_OUT`. When a metric is valid for only `CTX_NONE`, the data is only recorded at the time it is collected. An example MRNet application that makes use of the Stream performance data collection facilities is provided in `$MRNET_ROOT/Examples/PerformanceData`.

NOTE: `MET_CPU_USR_PCT`, `MET_CPU_SYS_PCT`, `MET_MEM_VIRT_KB`, and `MET_MEM_PHYS_KB` are currently only supported for Linux.

**Table 2: Metric-Context Compatibility Matrix**

| | CTX_SEND | CTX_RECV | CTX_FILT_IN | CTX_FILT_OUT | CTX_NONE |
|---|---|---|---|---|---|
| **MET_NUM_BYTES** | yes | yes | yes | yes | no |
| **MET_NUM_PKTS** | yes | yes | yes | yes | no |
| **MET_ELAPSED_SEC** | no | no | no | yes | no |
| **MET_CPU_USR_PCT** | no | no | no | yes | no |
| **MET_CPU_SYS_PCT** | no | no | no | yes | no |
| **MET_MEM_VIRT_KB** | no | no | no | no | yes |
| **MET_MEM_PHYS_KB** | no | no | no | no | yes |

# APPENDIX G: ENVIRONMENT VARIABLES

**Table 3: Environment Variables**

| | |
|---|---|
| XPLAT_RSH<br>XPLAT_RSH_ARGS | Set `XPLAT_RSH` to the name of the remote shell program to use for remote process execution. Default is '`ssh`'. `XPLAT_RSH_ARGS` can be used to pass shell-specific options to the remote shell. |
| XPLAT_REMCMD | If it is necessary to run the remote shell program with a utility such as `runauth` to non-interactively authenticate the unattended remote process, that command may be specified using `XPLAT_REMCMD`.<br><br>NOTE: Each MRNet process that needs to start remote processes checks its environment for the remote shell to use. Thus, XPLAT_RSH and related variables must be set in the environment for all MRNet front-end and communication processes. The easiest method of ensuring the environment is set correctly is to define `XPLAT_RSH` within the login scripts for the user's default shell. |
| XPLAT_RESOLVE_HOSTS | Tell XPlat to perform DNS resolution of hostnames and IP addresses by setting the variable to '1'. Default is '1'. |
| XPLAT_RESOLVE_CANONICAL | When `XPLAT_RESOLVE_HOSTS` is '1', setting `XPLAT_RESOLVE_CANONICAL` to '1' will tell XPlat to try to resolve all hostnames to their canonical DNS format. Default is '0'. |
| MRNET_OUTPUT_LEVEL | Set the debug output level (valid values are 1-5, default is 1). Level 1 will only log warning/error messages, level 3 provides fairly detailed function execution logging, and level 5 will produce every log message that MRNet generates. |
| MRNET_DEBUG_LOG_DIRECTORY | Specify the absolute path to the directory to store MRNet log files. If not set, the first existing directory from the following list is used:<br>• $HOME/mrnet-log<br>• /tmp |
| MRN_COMM_PATH | If `mrnet_commnode` is not in your path by default, you can specify the full path using this variable. |