

REMORA v1.0

REsource MOnitoring for Remote Applications
Document Revision 0.1 - Work In Progress
September 29, 2015

Antonio Gómez Iglesias, Carlos Rosales Fernández
`agomez@tacc.utexas.edu`, `carlos@tacc.utexas.edu`

High Performance Computing
Texas Advanced Computing Center
The University of Texas at Austin

Copyright 2015 The University of Texas at Austin.

Abstract

Remora is a tool to monitor runtime resource utilization. As of version 1.0 the suite contains code designed to test:

- Memory utilization in CPUs, and Xeon Phi co-processors
- CPU utilization
- IO utilization for the Lustre file system
- NUMA properties
- Network topology

Throughout the text we have used text boxes to highlight important information. These boxes look like this:

These gray boxes contain highlighted material for each section and chapter.

Our intention is to create a simple runtime resource monitoring tools that provide both simple to understand high level information to the user, as well as detailed statistics for in-depth analysis. We welcome all feedback. Feedback that includes suggestions for improvement in the usability, reliability, and accuracy of this tool is particularly welcome.

Version 1.0 is the first public version of Remora.

Contents

1	Installation	2
2	Design and Implementation	3
2.1	Statistics Collected	3
3	Using Remora	7
3.1	Collecting Data	7
3.2	Post-Processing	7
3.3	Execution Customization	7
	References	8

1 Installation

The Performance Analysis Workbench is simple to install. In order to use all of its features you will need GNU Make and a C compiler. If testing the Xeon Phi you will also need a compiler and with support for compilation of native code for the Xeon Phi - currently limited to Intel 13+ compilers.

First of all obtain the latest version tarball, currently `remora-1.0.tar.gz`, and expand it in a convenient location in your system:

```
tar xzvf ./remora-1.0.tar.gz
```

Alternatively clone the git repository with the latest development version:

```
git clone https://github.com/TACC/remora
```

This will create a top level directory called `remora`, with subdirectories `/docs` and `extra`. Change directory to the top level of `remora` and edit `install.sh` to reflect your choice of installation directory, build type, and xltop configuration. The variables to modify are:

Variable	Description	Default
REMORA_DIR	Absolute path to installation directory	.
XLTOP_PORT	Port number that xltop is listening in	xxxx
PHI_BUILD	Enable (1) or disable (0) Xeon Phi build	0

Once these variables are set the tool can be installed by using:

```
./install.sh
```

The executable tests will be placed in the `remora/bin` directory unless the `REMORA_DIR=/install/dir/location/` has been modified at the top of the installation script.

2 Design and Implementation

Remora is designed with ease of use in mind. Our goal is to create a usage model that users can take advantage of with minimal effort. While there are models where the tool could run transparently, just by loading a module that would change the environment to collect all the information at runtime, we decided to opt for a model where the user loads the module, and then changes the submission script to invoke our tool by prepending its name to the actual command that has to be analyzed. This has the advantage of preventing unnecessary overhead when the module is loaded and runs that do not require instrumentation are submitted. It also simplifies the data collection for serial jobs and scripts, which do not use an MPI launcher that can be easily hijacked or modified. For example, if the original command is `mpirun ./myparallelapp`, the new command will be `remora mpirun ./myparallelapp`. A more complicated scenario is when the user wants to run a set of different scripts in the same job. In that case, the user will have to put all the commands regarding the execution in a shell script (i.e. a shell script called `myjob`). Then, all the user has to do is to call the script as follows `remora ./myjob`. The tool can be used in a batch script or interactively in the command prompt.

In its current implementation the tool generates a flat ssh tree with a single connection from the master node in the execution to all other nodes. This connection initiates a background process that collects the statistics for each node with a frequency specified by the user. The remote background tasks execute a loop over all processes owned by the user, and aggregate the data before committing it to file.

For runs involving a Xeon Phi co-processor the background task is pinned to the last core (assumed to own hardware threads 0, 241, 242 and 243) since in most execution modes this will avoid interference with the application during runtime and minimize impact on the execution time. The source file `mic_affinity.c` can be modified in order to change this setting.

As previously stated, REMORA allows users to monitor the statistics as they are being generated. All the collected data is written to files accessible by the users in real time.

2.1 Statistics Collected

REMORA collects a set of statistics that are useful in many different scenarios when profiling an application. The data collected by REMORA consists of:

- Detailed timing of the application.

- CPU utilization.
- Memory utilization.
- NUMA information.
- I/O information (file system load and Lustre traffic).
- Network information (topology and InfiniBand traffic).

Dynamic information is collected every `REMORA_PERIOD` seconds. The following describes the data collected in more detail.

CPU The application reports the average CPU usage of the last second (independently of the value specified for `REMORA_PERIOD`). This information is very important for applications that use OpenMP, where it is possible to easily analyze how the cores are being used. It also allows to check for a correct pinning of threads to the cores: not pinning processes could lead to threads floating between cores, which will be show up in this report. MPI applications can also benefit from this information.

Memory One of the most recurring questions for HPC users is "how can I know how much memory my job is using?". Trying to answer that question, REMORA collects the most relevant statistics regarding memory usage every `REMORA_PERIOD` seconds:

- Virtual Memory (and Max Virtual Memory): this is a very important value as the OOM (out of memory) killer will use it to kill the application if needed.
- Resident Memory (and Max Resident Memory): physical memory used by the application.
- Shared memory: applications have access to shared memory by means of `/dev/shm`. Any file that is put there counts towards the memory used by the application, so the application reports this usage.
- Total free memory: this will take into account the memory not being used by the application, the libraries needed by the application, and the OS.

Data is collected from `/proc/<pid>/status` for all of these except shared memory, which can be obtained from `/dev/shm`. Memory usage for all user processes is aggregated and written to a single file per node involved in the execution. At the end of the run the maximum values for memory utilization (and minimum value of total free memory) are aggregated into a single file.

When Xeon Phi co-processors are used as part of a symmetric execution model, each Xeon Phi is treated as a separate node and the same memory information is collected from Phi and from host CPU. Individual files are maintained for each node and Phi and the per node aggregated summary is provided individually for nodes and Xeon Phi devices, since their available memory tends to be different.

NUMA As it is well known, NUMA (Non-Uniform Memory Access) can have a large impact on the overall performance of an application. Sometimes small changes in the code can lead to large improvements once it has been discovered that NUMA was imposing a penalty over the application. Our tool reports how memory is being used in each socket and it also collects the number of NUMA hits and misses. The information is extracted from the numastat tool: numastat is called only once on each collecting period; the output of numastat is then analyzed and several different fields are used for the statistics:

- Number of hits: total number of memory access hits.
- Number of misses: total number of memory access misses.
- Number of hits in the current node: if the data that the application was looking for is in the same of node where the core requesting that data is located, it produces a hit in the current NUMA node.
- Number of hits in the other node: when the data required is in cache, but in the other NUMA node.
- Total memory free/used on each node.

xltop [1]. xltop is a Lustre monitor developed at TACC that can be freely downloaded and installed. It provides statistics about file system utilization, including I/O requests per second and read and write volumes in MB/s. Originally, xltop was a continuous monitoring client that connects to other tools (xltop-master, xltop-servd, and xltop-clusd) to retrieve the information statistics from every MDS (metadata server) and OSS (Object Storage Server). This is a tool used by administrators to specifically look for very heavy I/O applications that can damage the stability of the file system. We have modified the original xltop code so that it can be invoked by any user to only check for the Lustre I/O statistics of the job being analyzed. Also, while xltop originally used ncurses to show the results, we have now implemented a version that collects the statistics in a file in order to simplify the analysis of the results.

Lustre Stats Our tool collects information regarding the data transferred by Lustre on each node used by the job while running. Normally, these statistics do not provide much information to the users. However, they are very useful if there was a problem in the file system while the job was being analyzed, as the number of messages dropped will significantly increase. The following Lustre information is collected:

- Number of currently active Lustre messages. It also includes a highwater mark of this value.
- Messages sent/received: total number of Lustre messages sent/received by the current node.

- Messages dropped: number of Lustre messages that failed to be delivered to the destination.
- Bytes sent/received: total number of bytes sent/received on Lustre messages.

InfiniBand Packets Number of packets transmitted using InfiniBand. This data can be used to get extra information regarding how the communication in parallel applications takes place. In particular, the time series can be used to correlate high network activity levels with sections of the code, and those sections can be revised for possible optimization.

3 Using Remora

3.1 Collecting Data

After making sure the remora `bin` directory is in your `PATH` and the remora `lib` directory is in your `LD_LIBRARY_PATH`

3.2 Post-Processing

All the data is collected in a set of files with the statistics organized in columns. Users can take those files and run any postprocessing tool that they develop. However, for simplicity, REMORA already provides a plotting script that takes all the statistics generated during collection and generates a number of plots. These plots show the most relevant information previously collected and represent a visual alternative while analyzing the results. This is a Python script that requires Numpy, matplotlib and blockdiag. The script can be called from the batch script or from the login nodes after the job has finished. In this second scenario, the script requires the id of the job to be analyzed as argument.

3.3 Execution Customization

REMORA is configurable in terms of the amount and type of data collected, but sensible defaults are provided to simplify its use. By default the statistics are collected every ten seconds.

REMORA provides two different running modes and it also allows the users to specify how frequently the data is collected. A verbose mode is provided mostly for troubleshooting and should not be used by default. The behavior of the application is controlled via three environment variables:

- `REMORA.PERIOD`: Time in seconds between consecutive data records. This is the time from the end of a collection event until the start of the next collection event. Depending on the platform where the tool is running, the overhead introduced by the application can make the duration of the collection event to vary, in which case there will be less data points in the collected results than expected. However, in the systems that we have tested the overhead of the application is small enough that the total number of collection points (CP) is almost equal to $CP = ET/RP$ where ET is the execution time (in seconds) and RP is the period (in seconds).

- `REMORA_MODE`: this variable accepts two possible values (`FULL` and `BASIC`). The `FULL` mode runs all the tests that the tool allows. The `BASIC` mode only reports memory and cpu usage. This is the recommended mode when the users know that the application of interest does not create problems in the distributed file system. The tool that checks for the file system load [1] is the main contributor to the small overhead introduced by the tool. When this report is not required, the overall overhead of the tool is very small.
- `REMORA_VERBOSE`: Enable (1) or disable (0) verbose mode. Default is disabled.

Bibliography

- [1] xltop is a continuous Lustre monitor with batch scheduler intergation:
<https://github.com/jhammond/xltop>