PathScale™

AN INTRODUCTION TO THE PATHSCALE OPTIPATH™ MPI ACCELERATION TOOLS

BRYAN O'SULLIVAN
Lead Engineer, Tools Group
PathScale, Inc.

Abstract

Whether you are starting out or a seasoned expert in clustered high performance computing, if you are concerned about achieving the best performance, a new class of performance analysis software may provide the help you are looking for.

1. The Cluster Performance Problem

Clusters of commodity systems are becoming the dominant platform for high performance computing (HPC), making up more than half of the TOP500 list of the world's fastest supercomputers[1].

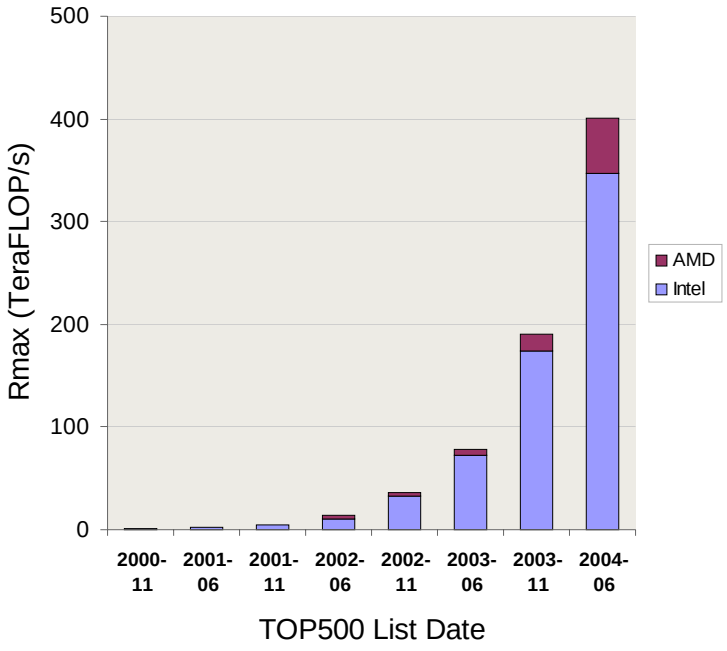


Chart 1. Growth in R_{max} of commodity clusters

Scientists and engineers use clusters to split up an application into a number of cooperating elements, working in parallel on small chunks of the overall problem. These elements are distributed across the individual computers in a cluster, and communicate using the Message Passing Interface (MPI) standard.

The State of the Art

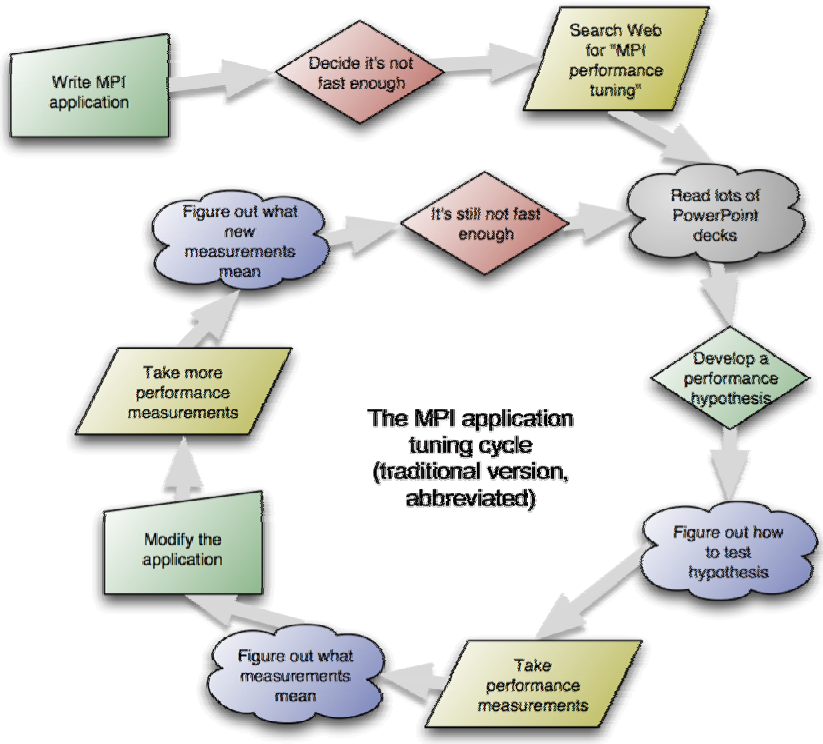
Within the HPC community, achieving good performance is acknowledged to be a difficult task. It requires expertise, time, and resources. It is particularly difficult to tune applications for commodity

[1] Source: TOP500, June 2004-http://www.top500.org/

clusters, as there are few suitable performance tools available; most of those that exist are aimed at standalone systems, not HPC clusters.

An efficient parallel application scales almost linearly; when run on ten CPUs, it will run almost ten times faster than on one. Good scaling is difficult to achieve, even on a modest number of CPUs. Not only will an application fail to approach the peak advertised performance of its cluster, the performance curve quickly levels off—and frequently even drops—as the size of the cluster increases.

In fact, it is so difficult to scale MPI application performance that managers of clusters at HPC facilities often limit their users to running applications on no more than 16 or 32 CPUs at a time. Using larger numbers of CPUs yields so little benefit for many applications that the extra compute power is effectively wasted.



2. Bringing Expertise to Bear

At PathScale, we already deliver EKOPath™, the highest-performing C, C++ and Fortran compilers available in the 64-bit commodity HPC market, and we have introduced InfiniPath™, the world's highest-performing HPC interconnect. In addition, we understand that developing a fast application requires knowing where and why it is not performing. Our solution to this knowledge gap is the OptiPath™ MPI Acceleration Tools.

3. Introducing the OptiPath™ MPI Acceleration Tools

The approach behind the OptiPath tools is simple, to liberate you from the need to take months-long detours into becoming an MPI performance tuning expert. The tools support the most popular 32- and 64-bit HPC platforms, the most widely used Linux™ distributions, and a variety of cluster interconnects.

Most importantly, the OptiPath tools guide you straight to the root causes of your application's performance problems, and start you on the road to fixing them.

- The OptiPath tools present you with a ranked list of the performance problems they find.
- For each problem, the tools display graphics indicating when, and how much, the problem occurs.
- The OptiPath tools pinpoint every location within your code where a problem occurs, ranked by severity, and annotated with complete context.
- The tools provide explanations for why such problems frequently occur, and suggestions for effectively fixing them.

This eliminates much of the head scratching, manual data gathering, and repetition of traditional performance analysis, by automating the parts of the process that ought not to require human intervention. This frees you up to apply your ingenuity to solving the performance problem.

In addition, the OptiPath tools are structured to make light work of the performance analysis and tuning process. They provide a powerful graphical user interface, and complete control from the command line for the times when you want to write scripts.

The tools make it easy for you to organize your performance experiments, by collecting them into notebooks.

Eliminate Drudgery

A seasoned developer of parallel applications has a “toolbox” of techniques for tuning an MPI application to perform better. Here are a few typical approaches to finding performance problems.

- Time the application's performance at different cluster sizes, and plot the speedup achieved for each size. For an untuned application, this curve will quickly flatten out at about 4 CPUs.
- Profile the serial portions of the code, to find and fix per-CPU bottlenecks. This is often sufficient to bring the cluster size where an application starts “losing steam” from 4 CPUs up to perhaps 16.
- Instrument the application to measure the amount of time spent computing and communicating. Compare the ratio of these two values at different cluster sizes. Find the communication hot spots as the cluster size grows, and fix them.

Each of these techniques requires a substantial amount of manual work—instrument the application, catalog performance numbers, plot charts, tweak the application's behavior, try again. It is difficult to apply them blindly; you have to develop a body of experience to know which methods to try, and which numbers are significant.

How do you approach the performance tuning process with the OptiPath tools?

- Tell the OptiPath tools just once how to run your application, and how to change a few variables that might affect performance—such as cluster size, model resolution, or the data sets to use.

- Follow the tools' suggestions for the kinds of experiments to run; for example “run at cluster sizes 2, 4, 8, 16, and 32.” With a few clicks, you can be running a batch of experiments. The OptiPath tools will inform you of their progress. If you need to, you can run more experiments later, or rerun the same ones to check their consistency.

- Examine the analyses. You'll be brought straight to the worst problems, right down to the exact lines of code affected. You'll be presented with your application's performance trends as the number of CPUs, or your model resolution, change. You can read about why the problem may be occurring, and what you can do to fix it.

- Once you have made some changes to your application, click once to rerun the entire suite of experiments, then compare the “before” and “after” analyses.

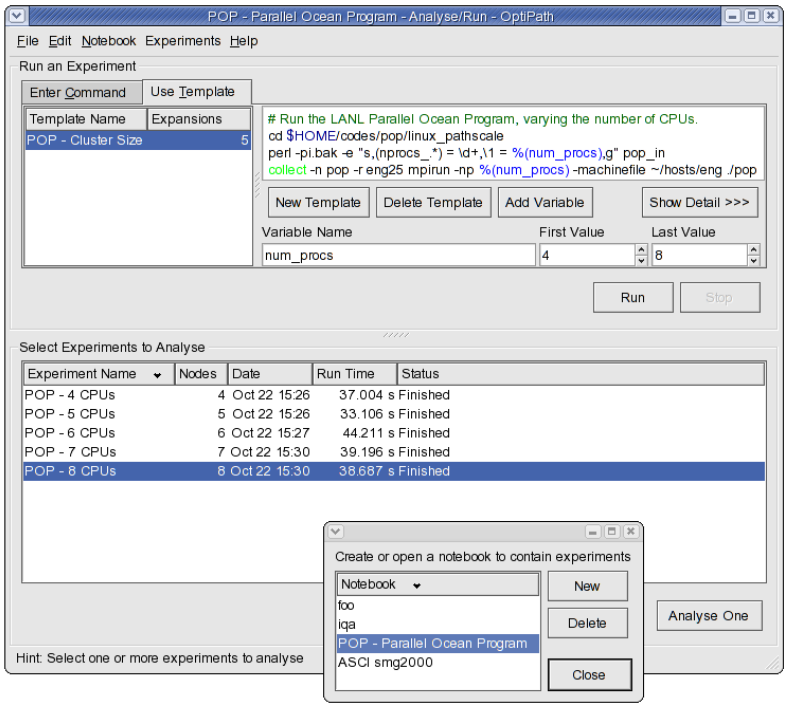


Illustration 2: Notebook and experiment management

4. The Tools in Action

The OptiPath tools present a powerful graphical user interface that lets you run, analyze, and manage your performance experiments.

Managing Experiments

Keeping track of your performance experiments is easy with the OptiPath tools. They present a graphical mechanism for keeping your related experiments grouped together, running new experiments, and analyzing collections of experiments.

When you need to run a series of experiments, the OptiPath GUI helps you to automate this process, and makes it easy for you to express what is changing from one experiment to the next.

Experiment Analysis

The OptiPath tools handle the analysis of individual and multiple experiments differently.

For a single experiment, the OptiPath tools give you details on that one experiment, displaying graphs of its performance over time.

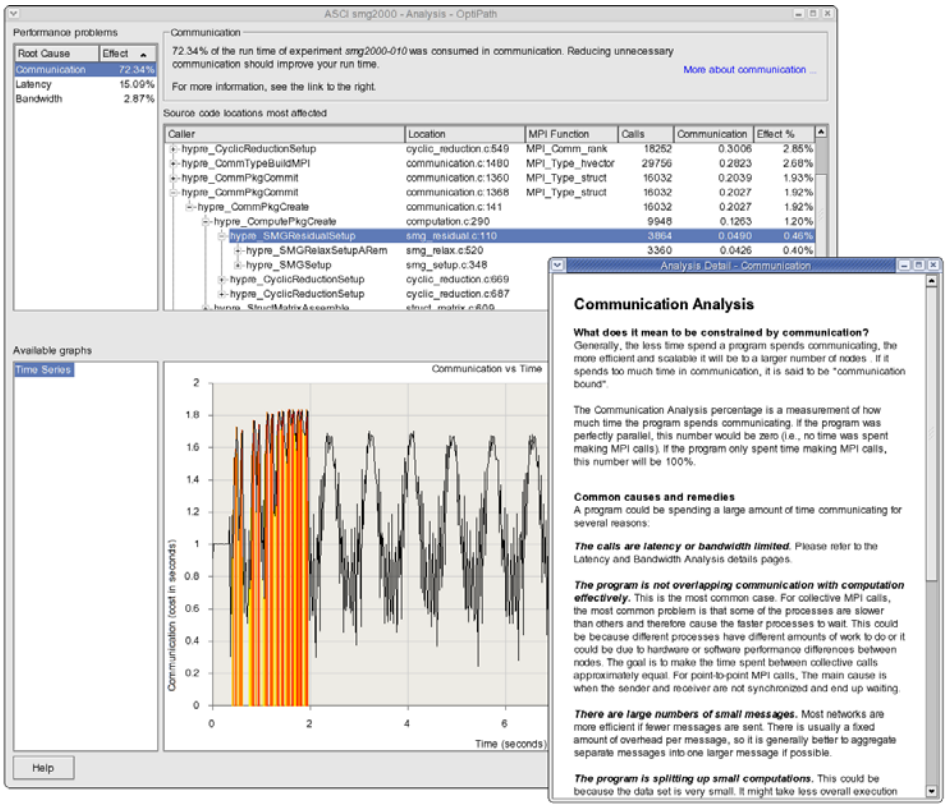


Illustration 2. Detailed analysis information

When you analyze multiple experiments, the OptiPath tools identify problems and trends across all of those experiments.

In each case, the tools present you with the root causes of your performance problems; the locations within your source code where those problems occur and graphical data that let you visualize the problems.

In addition, the OptiPath tools provide you with details regarding the likely causes of each problem, along with approaches you can use for solving these problems.

5. Architecture of the OptiPath MPI Acceleration Tools

The OptiPath tools guide you straight to the root causes of your MPI application's most pressing performance problems. How do they do this? Internally, the OptiPath tools are divided into four major modules.

The trace collector runs a performance experiment with very low overhead, gathers data, and returns the data to your workstation. Unlike other MPI performance tools, the trace collector does not force you to recompile or relink your application. The trace collector does not rely on shared filesystem access, so it is suitable for use in secured clusters. It is compatible with major batch schedulers, such as Torque.

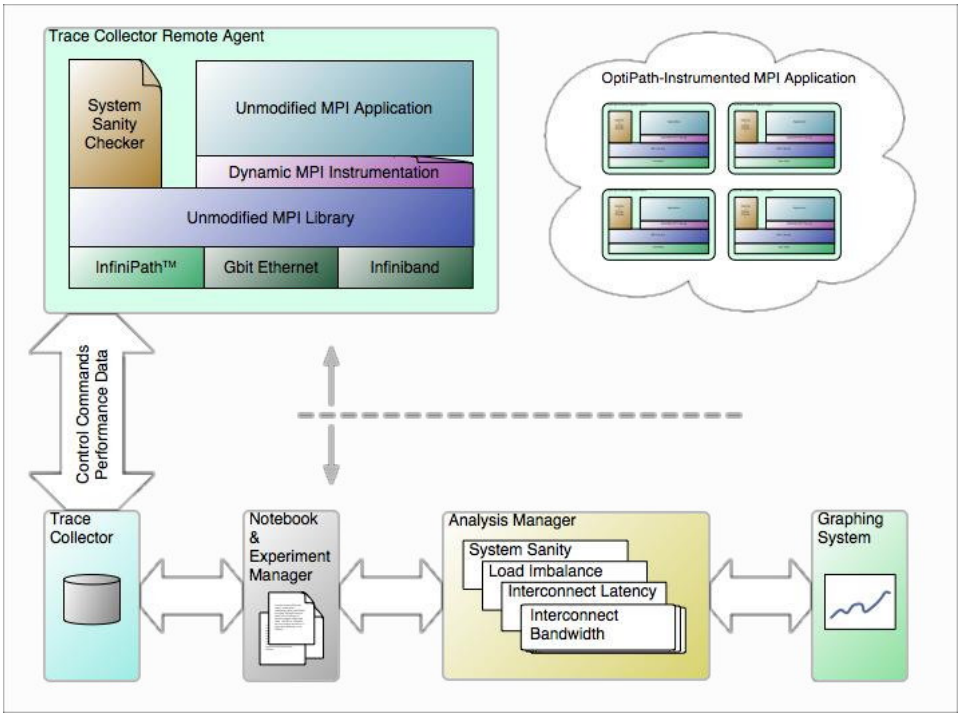


Illustration 3. Architecture of the OptiPath MPI Acceleration Tools

The notebook and experiment manager is responsible for your performance data. It maintains a history of the experiments you have run, so you can rerun complex batches of experiments with a single click. It also pre-caches and saves analyses, to give you faster first-time access to the analysis of large, intensive computations.

The analysis manager maintains the engines that perform individual analyses. Our analysis architecture achieves accuracy without sacrificing speed.

The graphing system provides visual interpretations of both time series and trend data. It identifies repeats, variations, and missing values in data, and automatically chooses the most appropriate information to present.

Software and Hardware Compatibility

The OptiPath tools are designed to be flexible. They support both AMD and Intel's 32-bit x86 (Athlon, Pentium, Xeon) and 64-bit x86-64 (Athlon 64, Opteron, Xeon 64) processor lines. In addition, the OptiPath tools running on one supported processor platform can run applications on, and analyze data from, any other.

The tools support the two major branches of Linux distributions used in HPC—Red Hat Enterprise and Fedora Core, and SuSE Enterprise and SuSE Professional. In addition, they support two of the most widely used batch scheduling systems, Sun Grid Engine and Torque (PBS).

The OptiPath tools work with PathScale's InfiniPath interconnect, Fast and Gigabit Ethernet, and most Infiniband host adapters and switches.

Early Access to the OptiPath MPI Acceleration Tools

PathScale is currently seeking beta testers for the OptiPath tools. We are particularly interested in a wide range of experience levels, from scientists and engineers taking their first steps with clustered HPC applications to seasoned MPI performance analysis experts.

To register for the beta program, visit PathScale's website at http://www.pathscale.com.

Conclusion

PathScale's OptiPath MPI Acceleration Tools guide you to the root causes of performance problems in your MPI applications. They offer a unique approach to performance analysis, giving you source-level details of your problems, backed up with the graphics, information, and context you need to address each problem and move on.

For More Information

Additional information on PathScale, Inc. and its products can be obtained by visiting http://www.pathscale.com on the World Wide Web, or contacting:

PathScale, Inc.
477 N. Mathilda Avenue
Sunnyvale, CA. 94085 USA

Phone: (408) 746-9100
Fax: (408) 746-9150
www.pathscale.com