Myrinet eXpress (MX): Is your Interconnect Smart?

Patrick Geoffray
Myricom, Inc.
patrick@myri.com

Abstract

Ever since the first prototypes more than ten years ago, the high-performance interconnect Myrinet has been designed around a revolutionizing concept: scalability lives on the edges. In other words, the switching element of the fabric must be kept very simple and the processing power moved to the network interfaces (NICs).

Myrinet NICs are built around a programmable RISC processor that operates directly in the data path. Myrinet eXpress (MX), a new communication layer, was specifically designed to harness this capacity of customizing the behavior of the hardware in order to improve the performance of common software interfaces such as MPI and Ethernet emulation.

MX embeds functions into the NIC to assist and improve the communication protocols on the host, while providing a very short critical path. MX opportunistically runs part of the host protocol closest to the data path, decoupling the progression of the protocol from the execution of the host application. The results show not only an impressive low-level point-to-point performance, but also a sharp improvement of higher-level metrics such as overlapping of communication, computation, and collective communication for MPI. Similar results are seen for high-throughput, low-overhead IP communication.

We will present the overall design of MX, along with the key technical choices we made and the problems we encountered. We will also present various results to show the impact on benchmarks and real world HPC applications.