InfiniPath: A New High Speed, Low Latency Cluster Interconnect

Greg Lindahl
PathScale, Inc.
lindahl@pathscale.com

The HPEC community has had considerable success using commodity technologies in embedded situations. Although embedded computing has radically different power and other constraints, some commodity technologies turn out to give excellent performance in embedded situations, without the embedded market having to fund all the R&D that went into creating the commodity product.

High speed, low latency interconnects are a technology that HPEC shares with ordinary High Performance Computing (HPC). InfiniBand is a new interconnect technology which is becoming popular in the HPC world. The PathScale InfiniPath host adaptor provides industry-leading low latency, high scalability, and high performance. It is suitable for low power, low chip count deployments. It consists of a single, 4-watt chip, and connects directly to AMD Opteron and Athlon64 cpus over the HyperTransport bus. Other embedded cpus using the HyperTransport bus, such as the Broadcom SiByte line, some PowerPC chipsets, and Transmeta's cpus, are potential targets for future development.

Our initial software implementation includes a TCP/IP-over-InfiniBand driver and an MPI library. We are also working on implementing a full set of InfiniBand protocols, based on the OpenIB software stack.

MPI performance details include:

1. 8-byte MPI latency of 1.32 microseconds, which is 1/2 to 1/3 the latency of competing interconnects
2. Peak bandwidth of 952 megabytes/second (uni-directional), and 1850 megabytes/second (bi-directional) (both streaming)
3. N1/2, the message size at which 1/2 of the peak bandwidth is achieved, of 385 bytes (streaming).
4. Good scaling on SMP nodes as more cpus are added.

TCP/IP performance has a peak bandwidth of 583 megabytes/second, with a one-way latency of 6.7 microseconds.

Benchmark details

MPI benchmarks were run on May 9, 2005 at PathScale's Customer Benchmark Center cluster -- 16 Microway Navion HTX Series servers with a total of 32 AMD 2.6 GHz Opteron processors (2 GB of memory per processor) on an Iwill DK8-HTX motherboard running Linux 2.6.11-1.14_FC3smp. Nodes are connected through an InfiniCon 9024 24-port InfiniBand switch using one meter cables. For these computations, 1 Megabyte = 10^6 bytes. Latency and Bandwidth were computed using the Ohio State benchmarks osu_lat, osu_bw, and osu_bibw. These bandwidth programs measure “streaming” bandwidth and not “ping-pong” bandwidth; the latency computation times a round-trip of a short packet and divides it by 2 to get a 1-way latency.

TCP/IP performance numbers were generated using netperf.