

Microway InfiniScope™  
InfiniBand Diagnostic Tool

Paul G. Howard, Ph.D.

Chief Scientist

Microway, Inc.

April 2007



InfiniScope<sup>TM</sup>, available from Microway, is a graphical diagnostic tool for InfiniBand clusters. The top section of the display (see Figure 1) shows all connections between host HCAs and switches and between ports on different switches. It also uses colors and shapes to show the current traffic level (the transmit bandwidth) on each port. The chart in the lower section shows the recent traffic history of a single HCA or switch port, of a switch, or of all hosts taken collectively.

In Figure 1, the chart shows the total traffic originating at all hosts for the last minute and a half. At this time the cluster was idle with only a little background traffic; most of it was to and from the master node, where the subnet manager was running. InfiniScope can also show the traffic for all ports on one switch, or for a single port as in Figure 2.

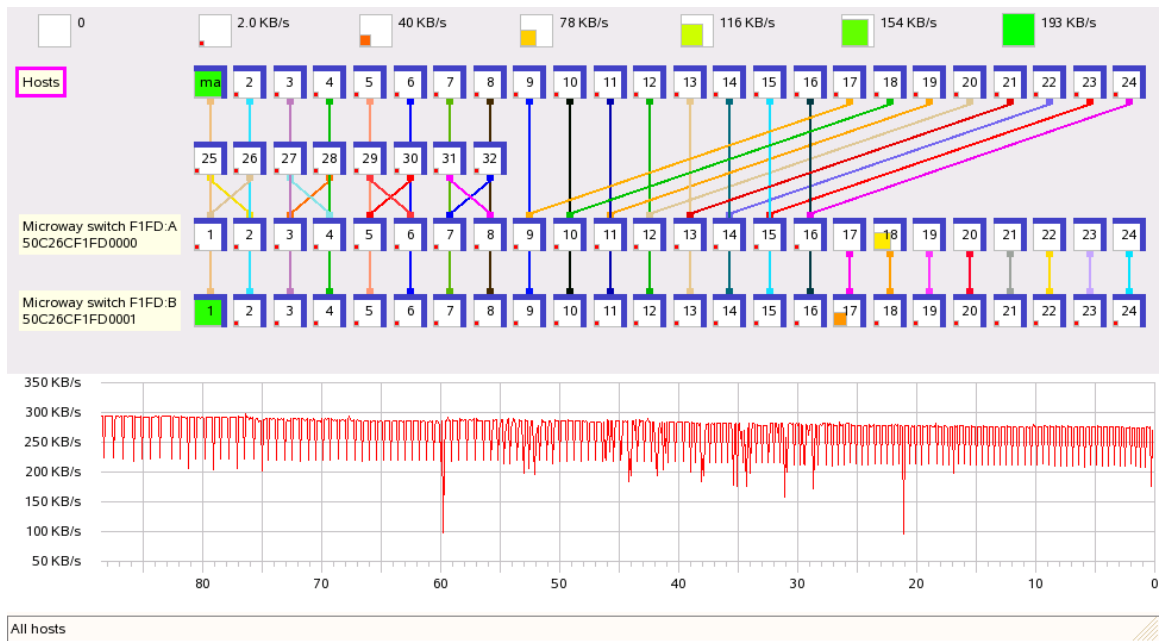


Figure 1: The InfiniScope display. The cluster has 32 nodes and a 36-port Microway FasTree<sup>TM</sup> switch. Internally the switch consists of two 24-port switches connected by 6 internal lanes (ports 19–24). In this cluster there are two additional external cables between the two switches, at ports 17 and 18. The traffic shown here is background traffic, when the cluster is idle.

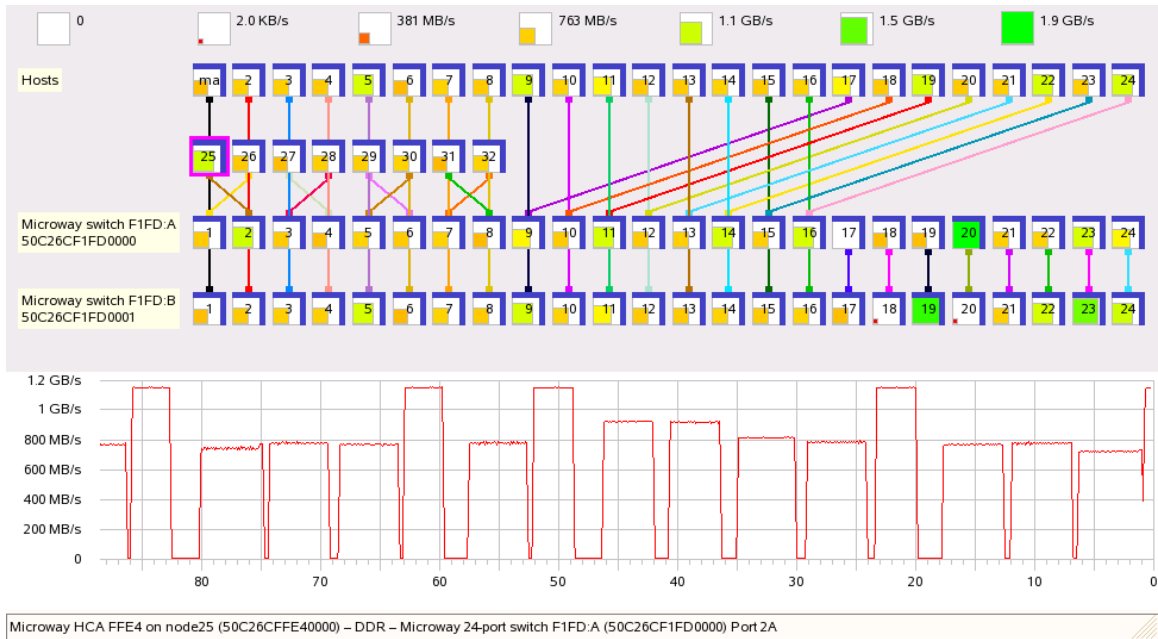


Figure 2: Traffic on a single port. The chart shows the transmit bandwidth from node 25.

## Cluster administration

InfiniScope can be used by cluster administrators to ensure that a cluster is running properly. All connections are shown as colored lines connecting two ports, so the administrator can quickly check that the cluster is cabled correctly. InfiniScope automatically keeps a permanent record of all connections, including the global ID and port number of every port on every switch and HCA, so it is possible to tear down the cluster and restore it to exactly the same state.

The border of each port is colored blue or red according as the connection is double data rate (DDR) or single data rate (SDR). In Figure 3 ports 5A and 6A of switch F1FD and the corresponding HCA ports on nodes 30 and 29 came up at SDR, indicated by the red marking on the ports; fixing this is usually as easy as reseating the connectors, but you have to know that there is an issue.

Similarly, connections that come up with lane width 1X (instead of 4X) are indicated by a heavy flashing yellow and black line. In Figure 4 the connection between

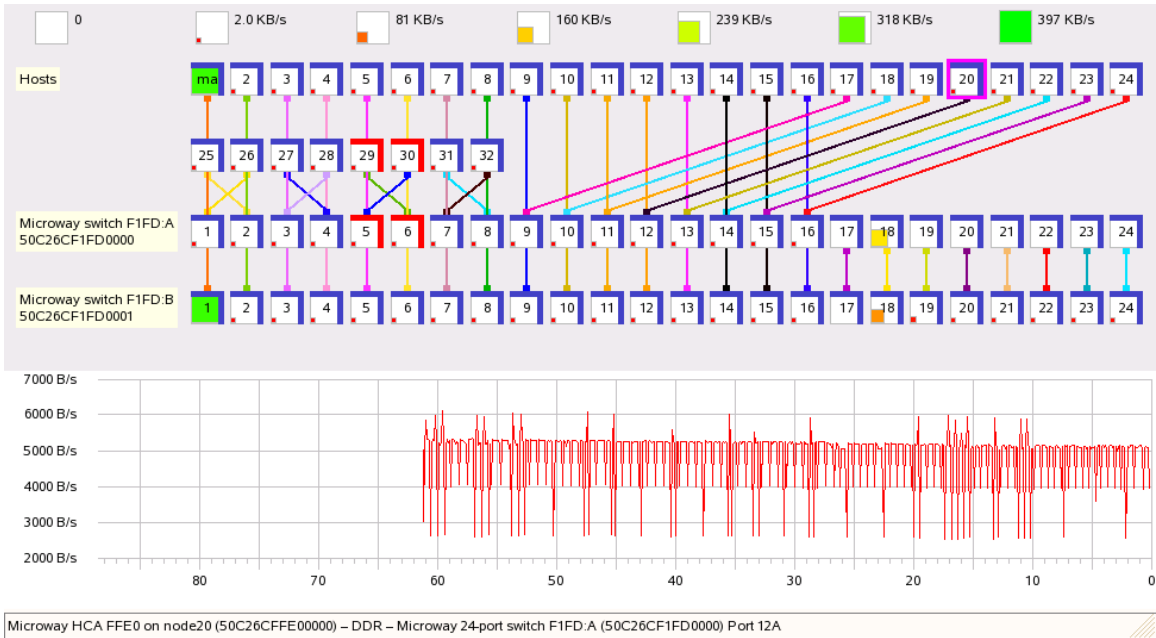


Figure 3: SDR connections. The connection between node 29 and port 6A and the connection between node 30 and port 5A came up at SDR.

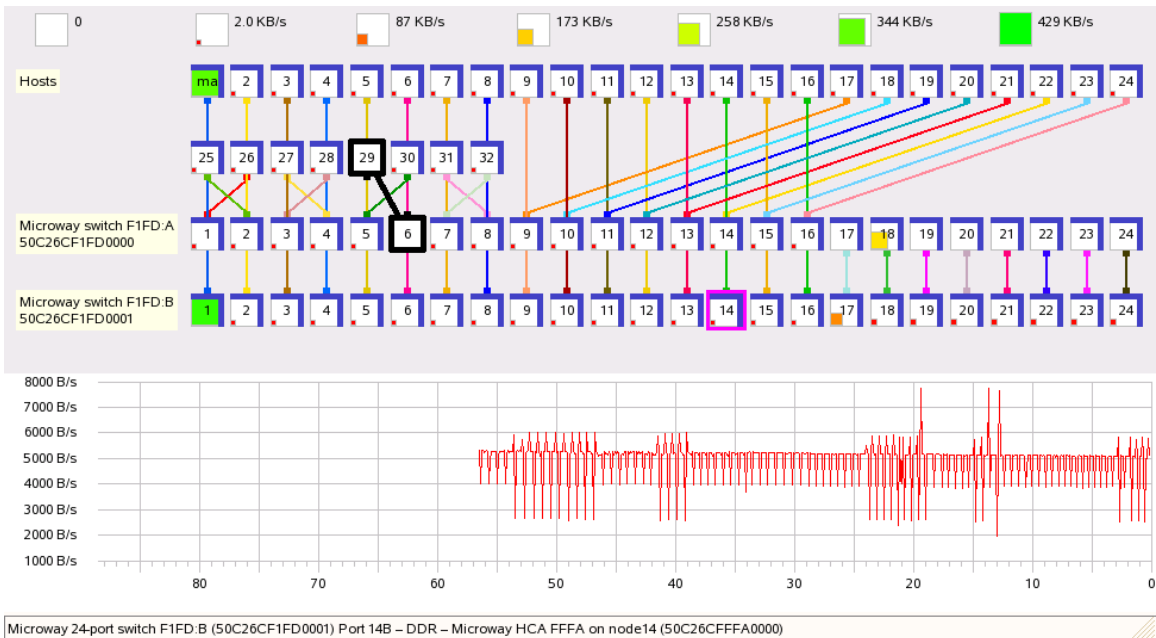


Figure 4: A 1X connection, between node 29 and port 6A. On the display the heavy boxes and the line connecting them flash yellow and black.

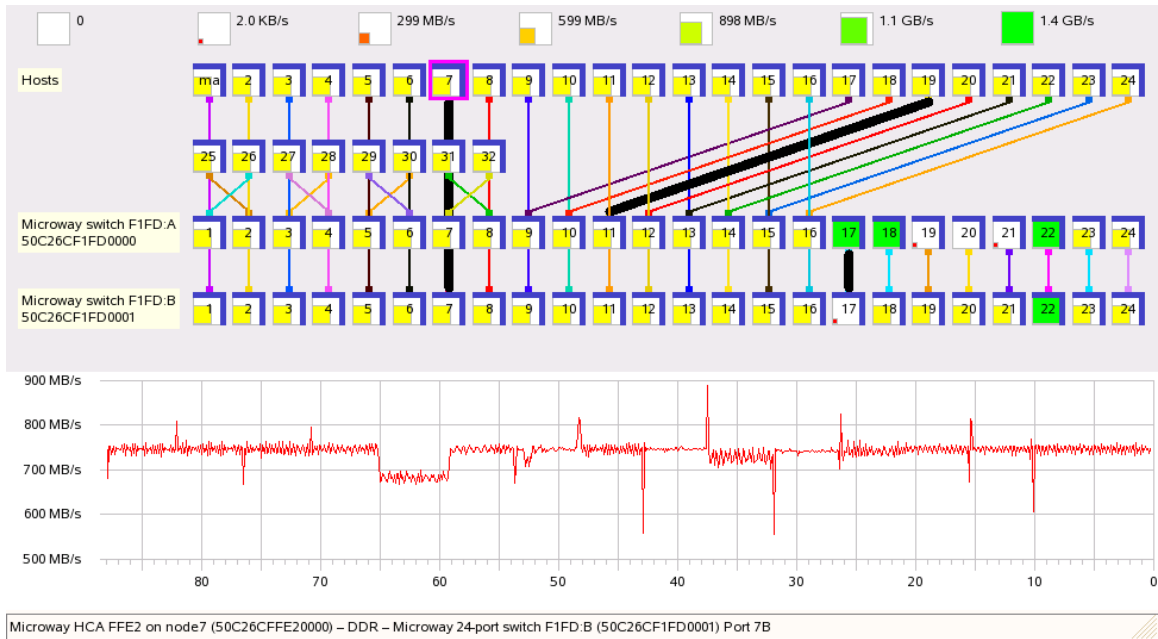


Figure 5: Host-to-host path. Data between nodes 7 and 19 travels from the HCA on node 7 to port 7B, then from port 17B to port 17A, and finally from port 11A to the HCA on node 19.

node 29 and port 6A is a 1X connection. Again, fixing the problem often merely requires reseating the connectors, but again you have to know that there is an issue.

InfiniScope can also show the data path between any two hosts, as assigned by the subnet manager, as shown in Figure 5.

## *flop*, the fabric loading program

*flop*, an MPI program to heavily load the fabric with traffic, is included with InfiniScope. In its standard mode it sends 5,000,000,000 bytes from each host to some other randomly selected host, then repeats the process with a different receiving partner. Eventually every host sends to every other host. Each pass takes about 5 seconds, and InfiniScope clearly shows the 5-second periodicity.

In Figure 6, *flop* is running and InfiniScope shows that there is no traffic in either direction between ports 17A and 17B; although there is a connection, it is not being used. When that cable between the two switches was added, the subnet

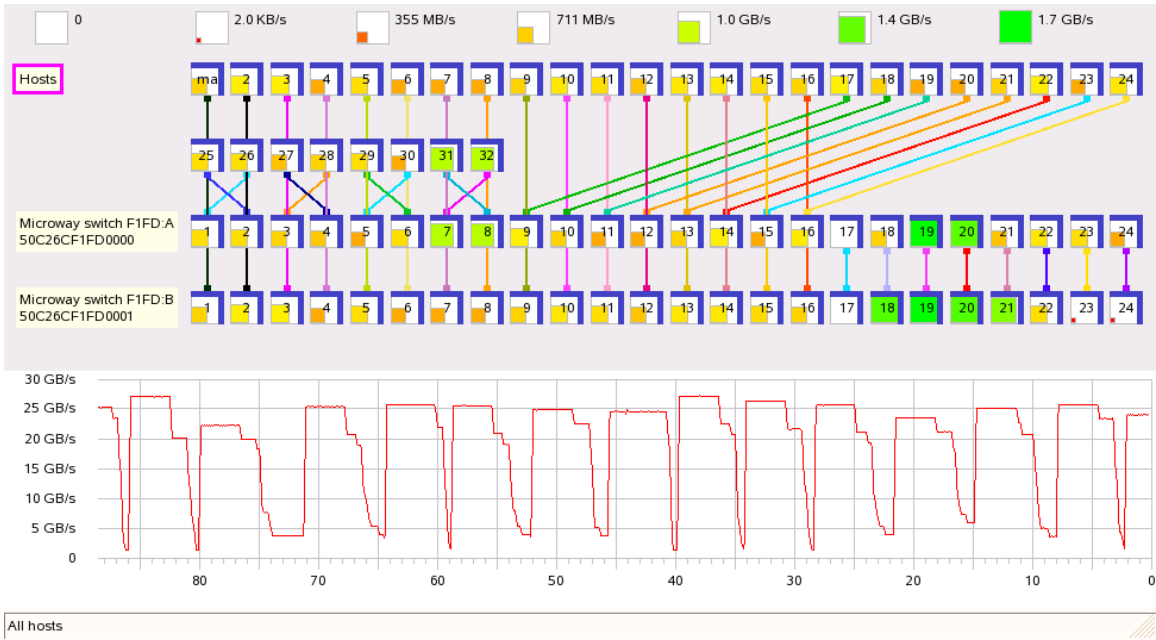


Figure 6: *flop* with an inactive connection between ports 17A and 17B. There is no colored square inside the boxes for those two ports, indicating that no data is flowing.

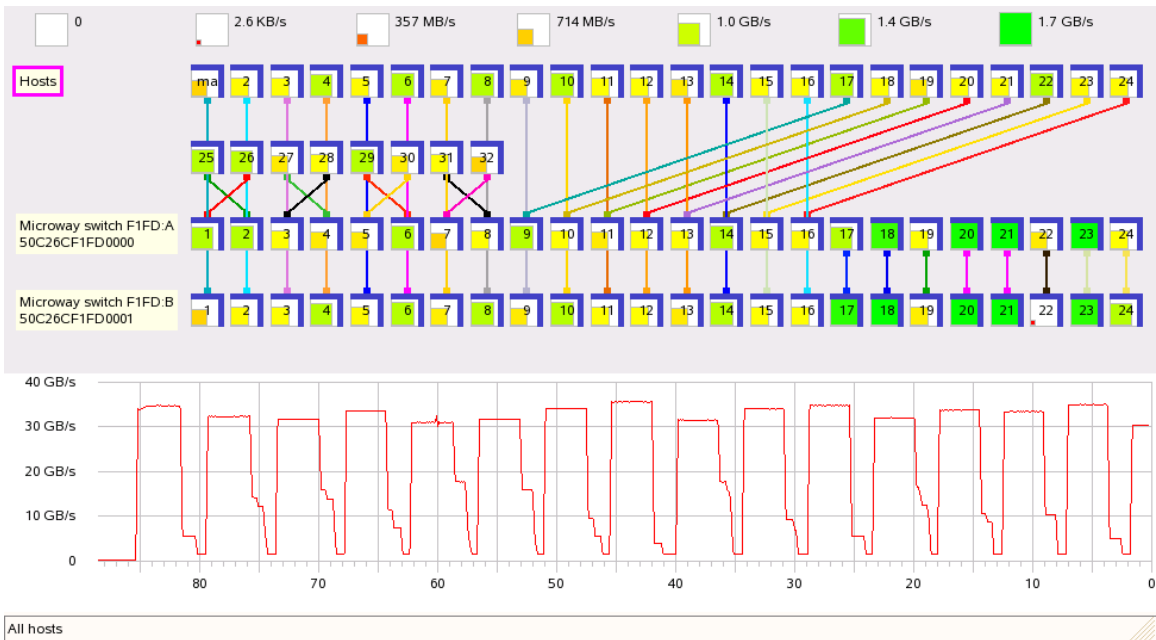


Figure 7: *flop* after restarting the subnet manager. The port 17A and 17B boxes contain green squares, indicating traffic on the connection.



Figure 8: Two instances of *flop* are running at the same time, leading to a highly irregular traffic pattern.

manager recognized the connection, but did not re-route any paths through it, since there were already seven other connections between the two switches. The inactive connection puts extra pressure on the other seven connections, limiting the overall cluster bandwidth to about 25 MB/s. The problem is easily solved (by restarting the subnet manager), but as before, you have to know that there is a problem. After a subnet manager restart, all connections between the two switches became active. As seen in Figure 7 the overall bandwidth went up to over 30 MB/s.

Another otherwise hard-to-spot problem is when multiple programs are running at the same time using the same ports. Even though all programs work correctly, they all suffer a performance loss. In InfiniScope this problem manifests itself in irregular performance graphs. The same symptoms can appear when a program was not properly terminated on one or more nodes. Figure 8 shows what can happen when two instances of *flop* are running.

A problem that comes up occasionally is accidentally attempting to run the subnet manager on more than one host, which results in running the primary subnet manager

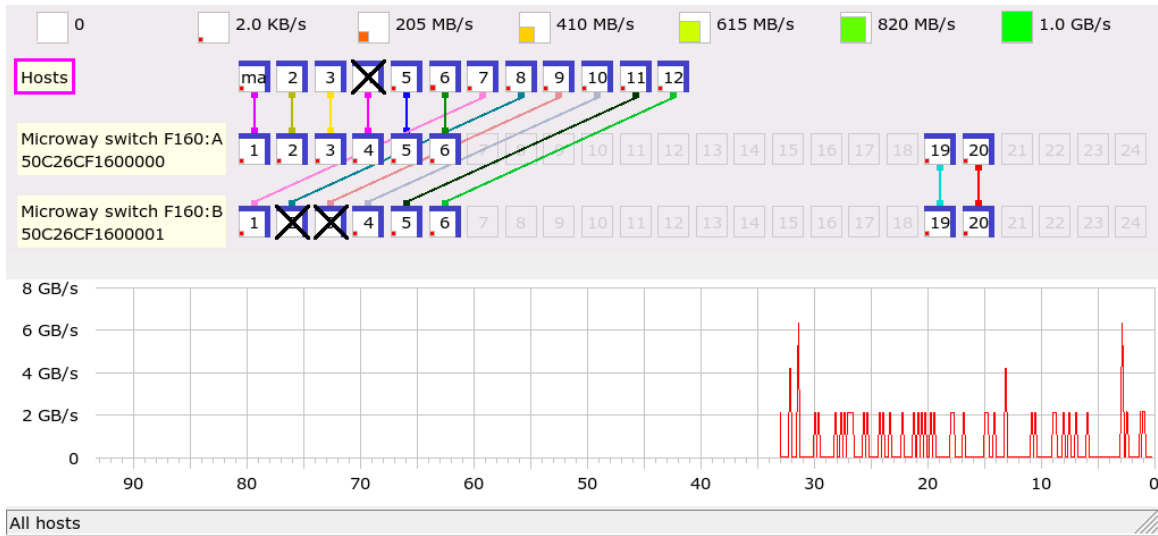


Figure 9: Errors occurred on node 4 and ports 2B and 3B. This figure is based on a 12-node cluster with a 48-port FasTree switch (internally two independent 24-port switches).

on the wrong host. InfiniScope does not detect this problem directly, but it usually gives a clue by reporting meaningless connections.

Hardware errors are rare and usually self-correcting, but if there is an issue with a switch or HCA, you would like to know about it. While InfiniScope is running, it continually checks each port, and reports errors by error type on the console and in the log file, as well as indicating them on the display, as shown in Figure 9.

## Profiling applications

InfiniScope can be used to profile the traffic patterns of your MPI applications. You can see the data injection rate from each host or from all hosts collectively. You can also monitor the traffic on switch-to-switch connections to identify cases of insufficient bandwidth. For instance, Figure 10 shows that even the heavy traffic load imposed by *flop* does not saturate the 8 links between the two switches.

Many MPI programs have their own characteristic traffic fingerprint, showing both the periodicity of the application and the possible communication bottlenecks.

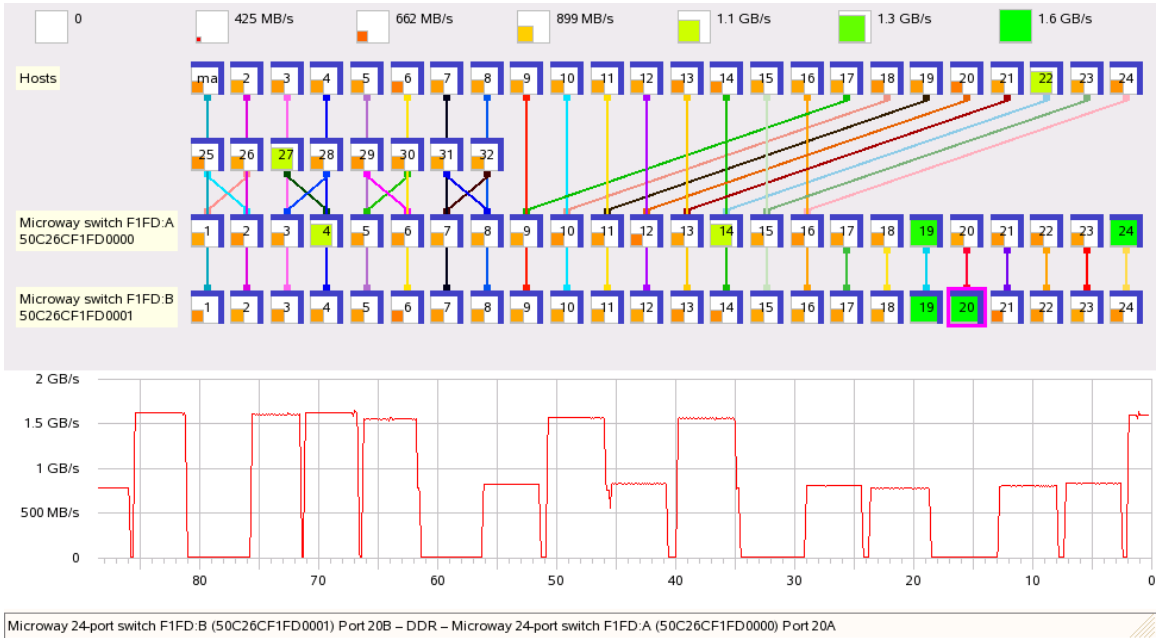


Figure 10: *flop* does not saturate the 8 links between the two switches.

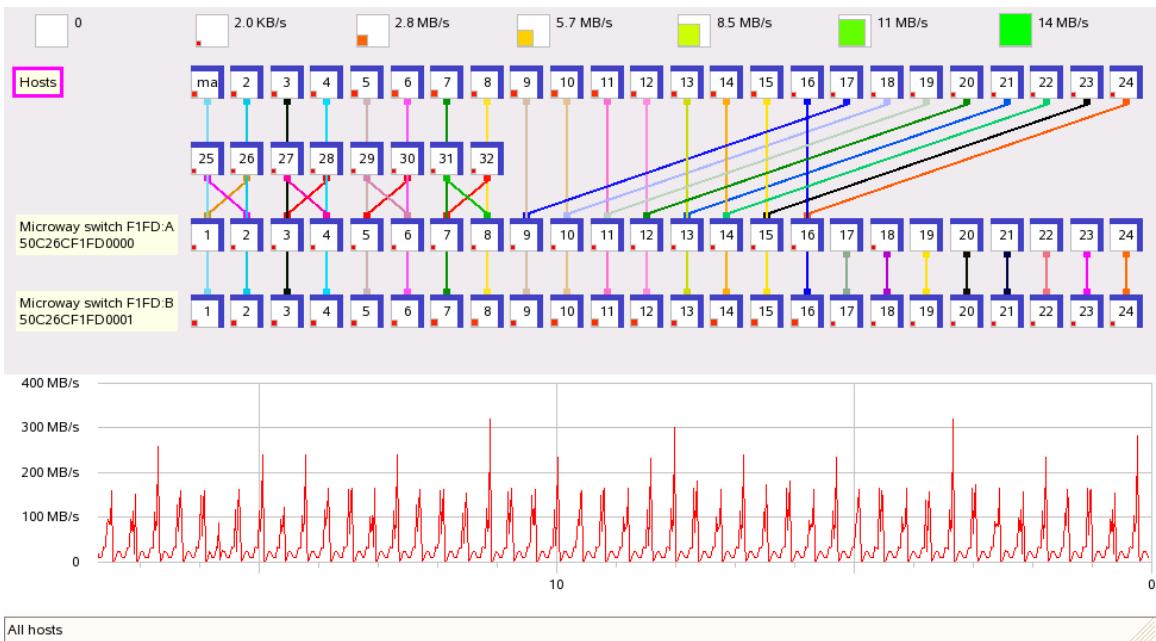


Figure 11: The NAS LU benchmark, C problem size, 16 processes. Each iteration takes about 0.4 second, and the maximum overall cluster bandwidth is only about 300 MB/s.

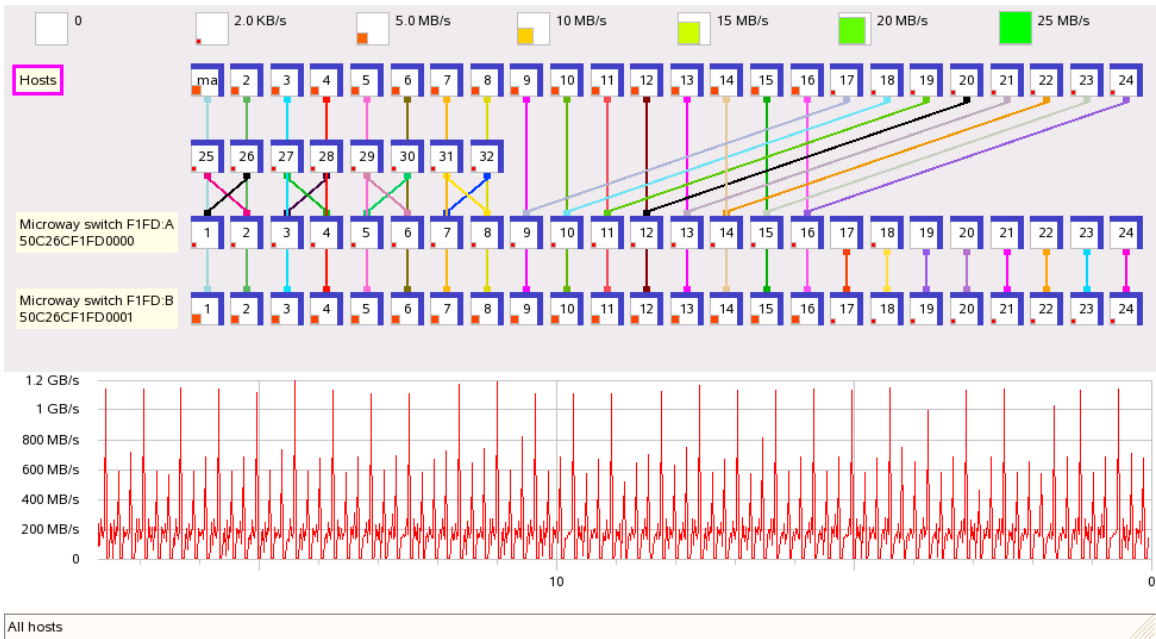


Figure 12: The NAS SP benchmark, C problem size, 16 processes. Each iteration takes about 0.7 second, and the maximum overall cluster bandwidth is about 1.2 GB/s.

Figure 11 shows the traffic pattern for the LU benchmark (“C” problem size, 16 processes) in the NAS test suite, and Figure 12 shows the pattern for the SP benchmark (also “C” problem size, 16 processes).

## InfiniScope can be used with MPI Link-Checker™

InfiniScope can be used in conjunction with MPI Link-Checker™, also available from Microway. Figure 13 shows part of a typical traffic pattern as MPI Link-Checker runs through its series of tests. Figure 14 shows typical output from MPI Link-Checker for the same cluster.

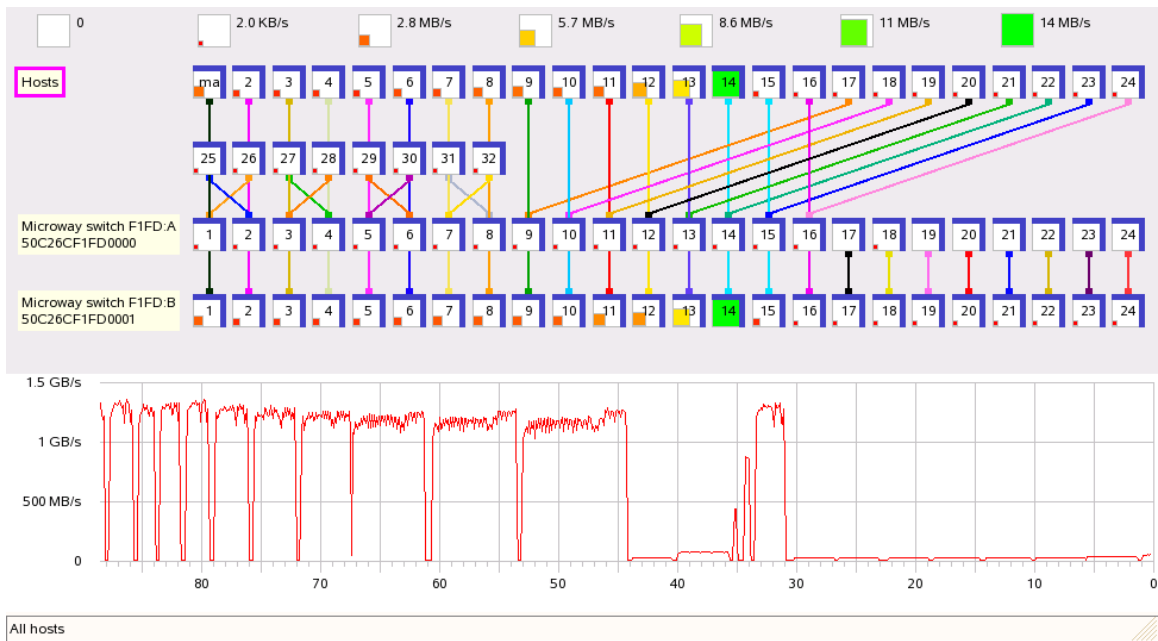


Figure 13: MPI Link-Checker traffic.

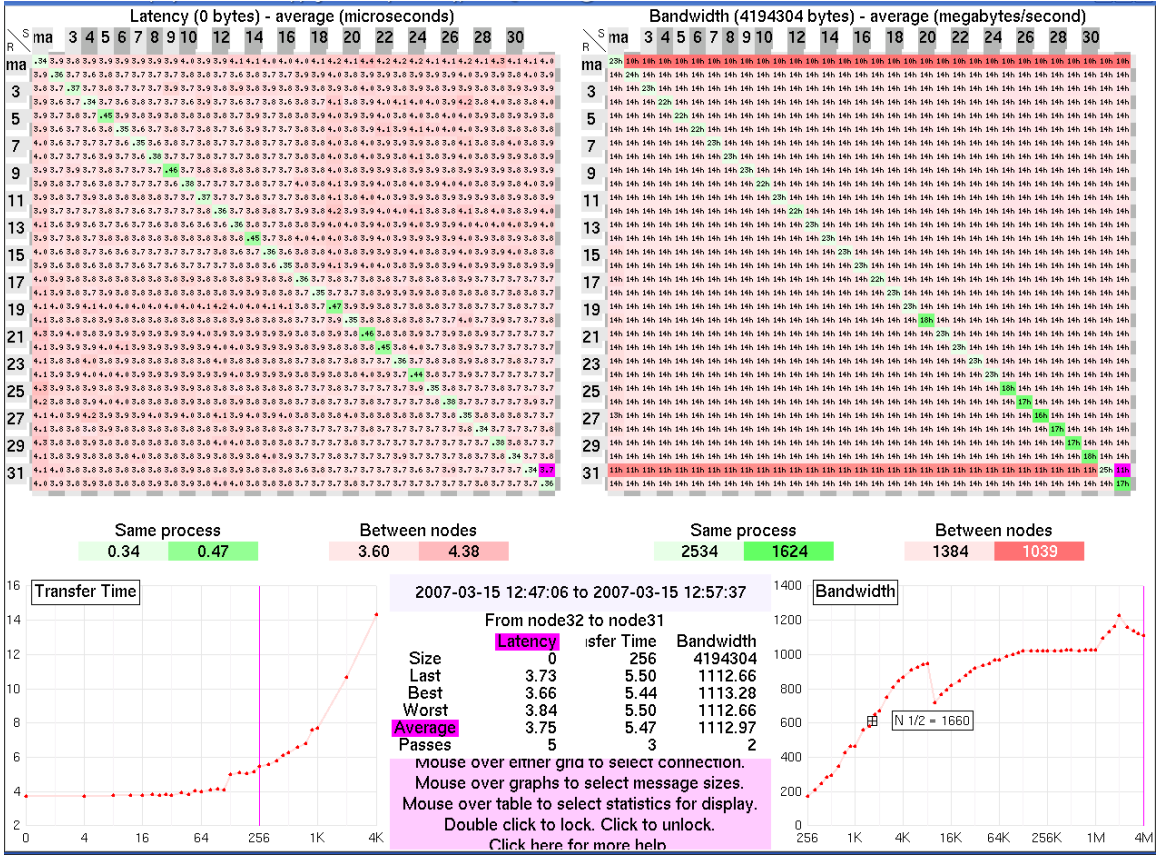


Figure 14: MPI Link-Checker screenshot for the 32-node cluster.

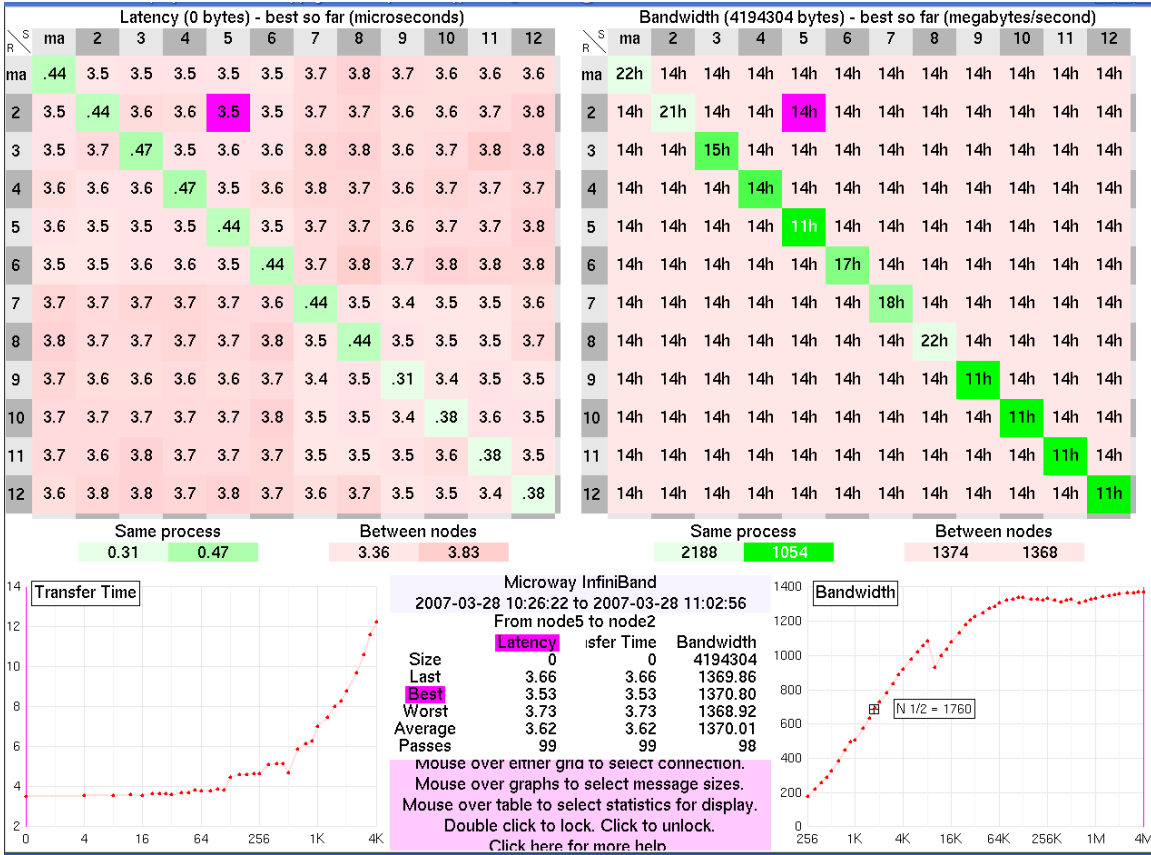


Figure 15: A view of the case study cluster using MPI Link-Checker.

## A case study

Here is a brief case study showing one way that InfiniScope can be used. The cluster being observed has 12 hosts and two InfiniBand switches. (The switches are two 24-port switches in the same case, available from Microway as a 48-port FasTree™ switch.) Six of the hosts are connected to each of the switches. The goal is to determine how many cables should be used to connect the two switches. Figure 15, a screenshot of MPI Link-Checker, shows bandwidths and latencies between all pairs of hosts on this cluster.

The time for *flop* to complete one iteration should be about 5 seconds (5,000,000,000 bytes at about 1 GB/s, the highest transmit rate that can be expected when a number of HCAs are all simultaneously transmitting and receiving). The total data injection



Figure 16: Case study to determine how much bandwidth is needed between two switches. *flop* is running, and there is one cable between switch F160:A and F160:B.

rate should be about 12 GB/s. *flop* pairs up senders and receivers randomly, so some pairings cause more data to travel between the switches than others. Figure 16 shows that with one cable between the two switches, some passes take about 5 seconds at a rate of 12 GB/s. Other less favorable passes take 10 seconds at a rate of 8 GB/s falling to 4 GB/s, or even 15 seconds at a rate of 4 GB/s. Figure 17 shows that with a second cable (and after a subnet manager restart to force the second cable to be used), more of the passes take 5 seconds, and most of the slower passes start at 10 GB/s, but still a number of passes take 7 or 8 seconds. Performance is still limited by the available inter-switch bandwidth. With a third cable (see Figure 18), most of the passes take about 5 seconds, and the rate is close to 12 GB/s for all passes. Three cables are sufficient. Increasing the number of cables to four (Figure 19) or even six (Figure 20) produces no significant improvement in performance. InfiniScope shows that for this application that loads the fabric heavily with a random traffic pattern, three cables are sufficient to connect the two switches.

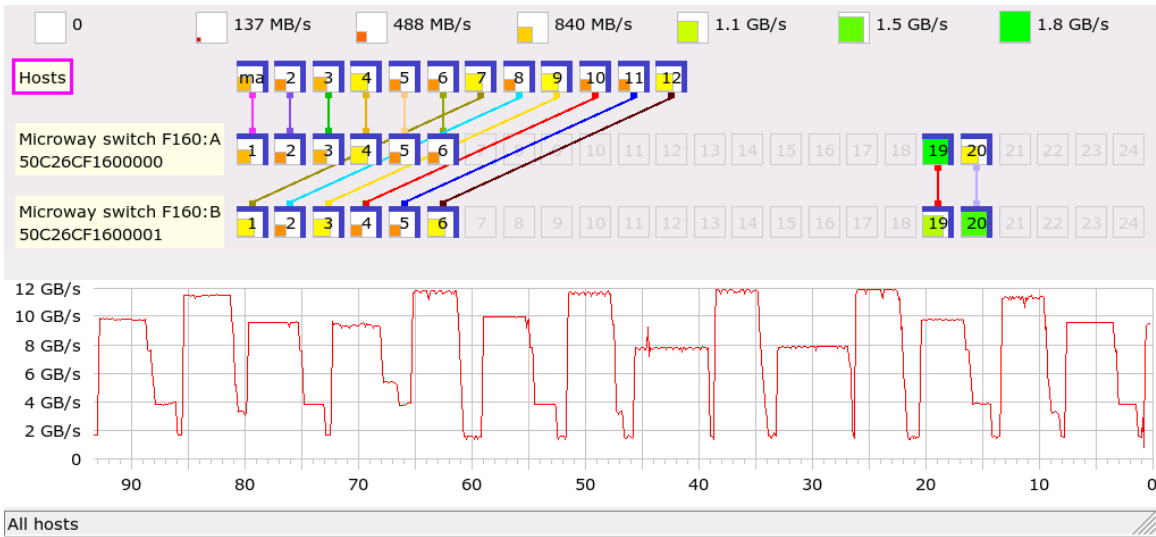


Figure 17: Case study with 2 interswitch cables, after restarting the subnet manager.

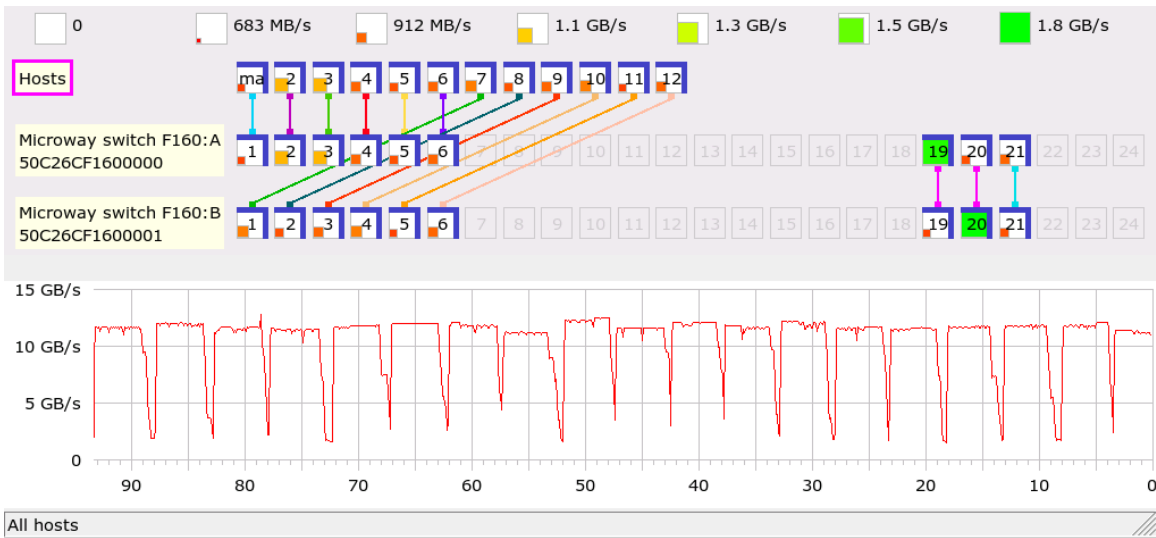


Figure 18: Case study with 3 interswitch cables.

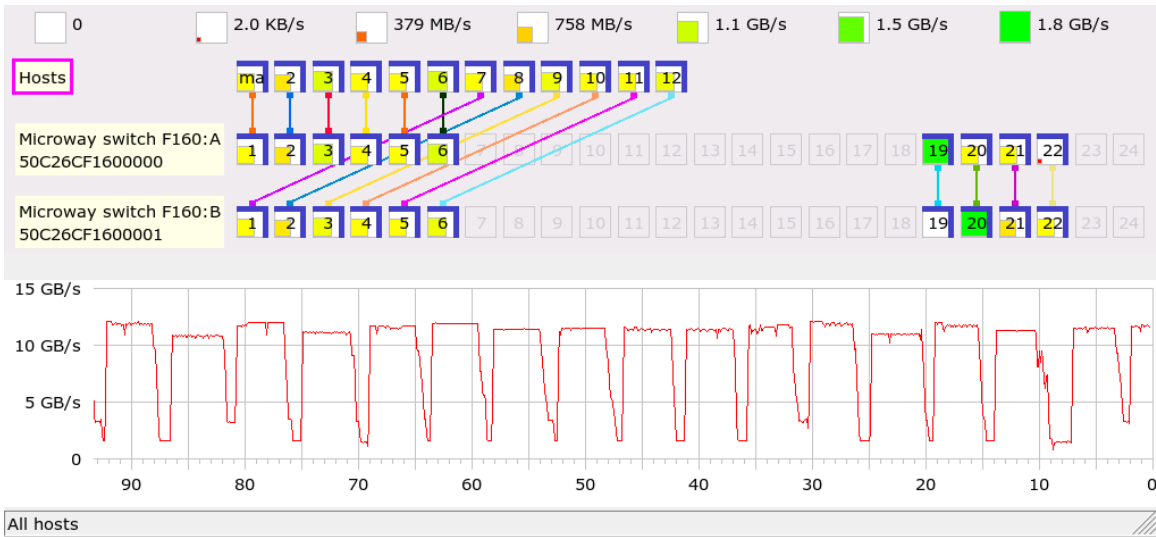


Figure 19: Case study with 4 interswitch cables.



Figure 20: Case study with 6 interswitch cables.

## About Microway

Microway's reputation as the world leader in innovative solutions for High Performance Technical Computing has been unchallenged since 1982, when our software made it possible to use the 8087 math coprocessor in the IBM PC. Our products consistently receive excellent reviews, our prices are competitive, and our service and technical support are outstanding. Microway's top notch Research and Development staff keeps you on the leading edge of technology with timely, powerful new products. At Microway our customers are treated as our most valuable resource, which is why our customer base remains strong and continues to grow.

Microway's products include clusters for High Performance Computing, FasTree™ InfiniBand-based switches, TriCom™ multi-function HCAs, WhisperStation™ silent workstations, the NodeWatch™ remote hardware monitoring system, and ServaStor™ storage solutions. Our clusters incorporate AMD dual-core Opterons, Intel dual- and quad-core Xeons, DRC FPGA coprocessors, and Mellanox silicon.

Designed and developed in-house, Microway software includes MCMS™ cluster management tools, InfiniScope™ InfiniBand diagnostic software, and MPI Link-Checker™ and MPI Fast-Check™ MPI diagnostic tools. Microway's Linux-based clusters and data solutions are used by customers in life sciences, academia, enterprise and government research laboratories.

The technical staff at Microway is qualified to assist you in benchmarking and speeding up your existing code and enhancing your present software and hardware investment. Our staff has over 50 years combined experience in designing Linux cluster configurations. We offer white papers on our web site at [www.microway.com](http://www.microway.com), as well as technical documentation of the hardware and software we design and integrate. To design your next custom system or cluster, please call our Sales Department at +1 (508) 746-7341. Our Technical Support Department can be reached at the same number or via email at [tech@microway.com](mailto:tech@microway.com).

For more than 25 years, the employees at Microway have earned our reputation for

excellence. We are proud of this reputation and totally committed to designing innovative products that provide state of the art solutions required to keep our customers on the leading edge of technology.