

QsNet^{II} Network Bandwidth

Authors: Duncan Roweth and John Taylor

The full configuration of a QsNet^{II} network provides equal numbers of links at each stage of the network. These links are bi-directional and are rated at approximately 940 MBytes/sec after protocol.

Process to process communication bandwidths are restricted (by the PCI-X interface) to between 820 and 910 MBytes/sec depending upon the chipset and the communication pattern.

A full QsNet^{II} network has roughly twice the peak bi-section bandwidth as it has host adapter bandwidth. For example in a 1024-way network there are 1024 bi-directional links between the node switches and the top switches, a total of 960 GB/sec in each direction. This compares with an aggregate host adapter bandwidth of 840GB/sec.

Extra capacity in the top of the switch is good in that it reduces contention, but it increases the cost and the overall effect on performance will depend upon how the network is used.

In order to quantify these effects we ran a number of performance tests on a 1024-way network, using first 512 and then 996 nodes. The tests were as follows.

- tping A bi-directional mirrored ping in which process 0 communicates with process n-1, process 1 with process n-2, etc, so that all traffic goes through the top switch. The tping program uses libelan TPORT message passing in the same way as MPI.
- gex A global exchange test in which every process sends data to every other process. The gex program uses libelan puts in the same way as Shmem.
- beff The Pallas effective bandwidth benchmark

We ran these tests with 4, 6 and 8 top switches to test the effect of reduced network bi-section bandwidth on these applications.

The results on 512 nodes were as follows:

Top Switches	tping		gex		beff	
4 (50%)	769	90%	729	90%	76533	91%
6 (75%)	853	99%	800	99%	81767	98%
8 (1000)	859	100%	812	100%	83753	100%

Table1: 512 nodes

The results on 996 nodes were as follows:

Top Switches	tping		gex		beff	
4 (50%)	717	84%	696	87%	144049	93%
6 (75%)	844	98%	780	97%	150483	97%
8 (1000)	857	100%	803	100%	155687	100%

Table2: 996 nodes

In the [Figure1](#) below we plot the bi-section bandwidth as a function of number of nodes for two 1024 systems with different PCI-X chipsets measured with the tping test above.

The blue line shows the performance of an HP RX2600 based system with a full network (8 top switches). The red lines show the performance of an Intel IA64 system with 4 top switches (50%) and 8 top switches (100%).

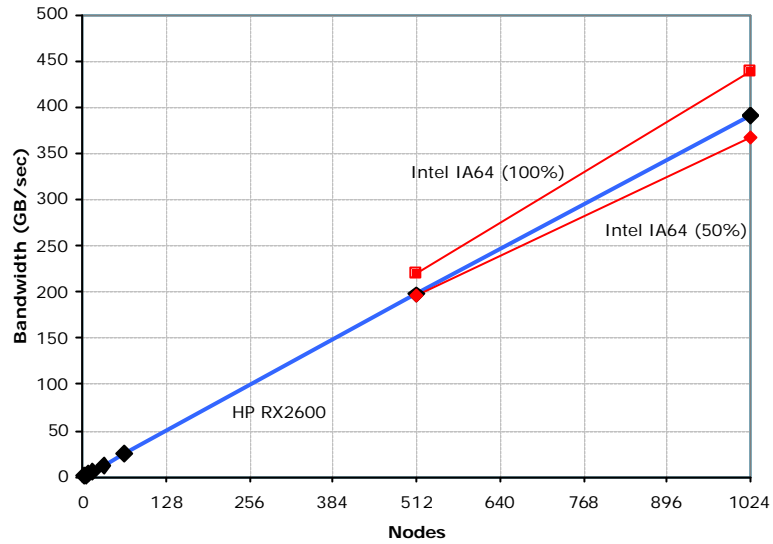


Fig 1: Bi-section bandwidth

Both systems show excellent scalability, bi-section bandwidth rising with the number of nodes.

Conclusions

Our tests show that there are performance benefits of having a full network, but that the extent of the benefit depends upon the application. The benefit of the extra top switches increases with the number of nodes, as one would expect.

Our test applications are communication bound; we would expect the effects of reducing the excess network bandwidth to be less for applications that are less communication intensive.

The choice of whether to use a full network or a reduced bandwidth network becomes a trade-off between price and performance. A full bandwidth 1024-way network is approximately 17% more expensive than the half-bandwidth network if all the cables are copper. If fibre links are used between the node switches and the top switches then the premium for a full bandwidth network rises to approx 27%. Half bandwidth networks offer significant savings in terms of cost and complexity