Myrinet2000 PC-clusters Performance

Roberto Innocente Carlo Carloni CalameOlumide Sunday Adewale

April 15, 2002

In this short write-up we report on the performance of some PC-clusters with Myrinet2000 (also known as Myrinet 3) network cards.

Table of Contents

Introduction

This report gives a comparative analysis of some results obtained on three different clusters. The first cluster is made of Pentium III@550Mhz nodes and Myrinet 2 cards, the second has Pentium III@933Mhz nodes with Myrinet 2000 cards, while the third cluster, that we are benchmarking, is made of Pentium 4@1.7Ghz nodes with Myrinet2000 boards and switch.

The GM software we used was customised in order to exploit the performance of the Intel 860 chipset as this was left out in the GM software.

Configuration

The nodes are equipped with two Intel Pentium P4 (Willamette) CPUs with 256k L2 cache, running at 1.7 Ghz, with an Intel 860 chipset, RIMM PC800 memory and 2 PCI 64 bits / 66 Mhz slots. The Myrinet boards are PCI64C with 2M of RAM and a Lanai 9.2 processor runnig at 200 Mhz. The switch is a 16 port Myrinet 2000 switch (SW16M).

GM software

GM is the proprietary protocol and API used by Myricom on their boards.

As it comes, it is configured to used at most 256 bytes tranfers between the chipset and the Myrinet board. With the 64-bit PCI bus, as it exists on our nodes, this is not sufficient to exploit the available bus bandwidth (it will perform

just 32 data transfers before to relinquish the bus). With these short burst transfers, we were able to measure a PCI bandwidth of only 147MB per second. By increasing the PCI burst length to 512 bytes (64 transfers), we obtained a PCI bandwith of 227 MB/s reading from the board and of 315 MB/s writing to the Myrinet board. (you need to patch the drivers/linux/gm/gmarch.c file)

Testing tools

The bandwidth and latency measures were made using the standard

```
gm_allsize
```

Myricom script, coming with GM software. On one node, the command was invoked in slave mode with

```
./gm_allsize --slave --size=20 </tt>
```

, on the other, it was invoked with

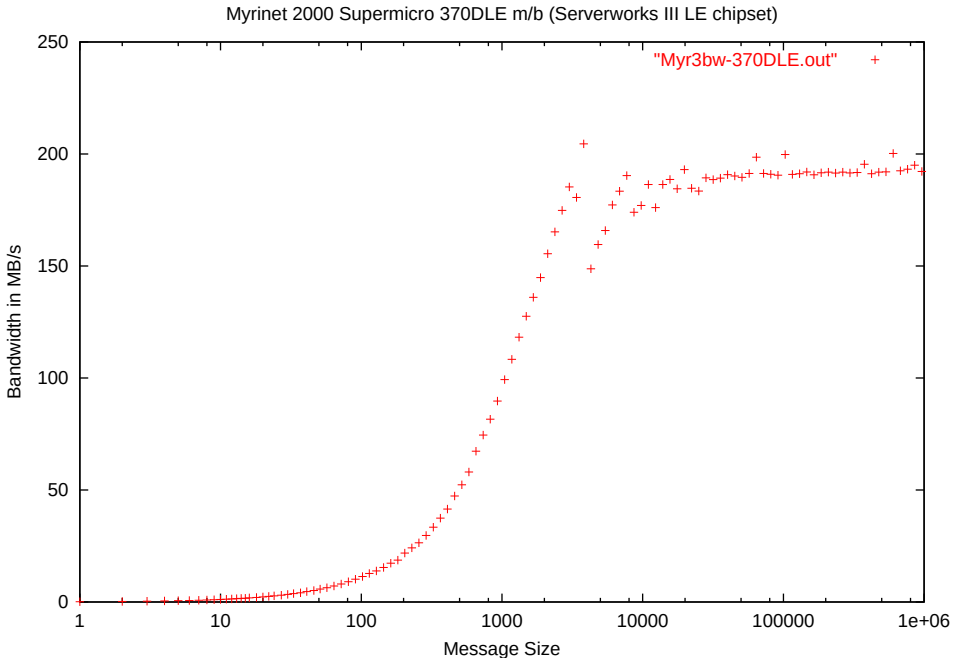
```
./gm_allsize -u -bw -h node1 --size=20 --geometric </tt>.
```

Performance charts

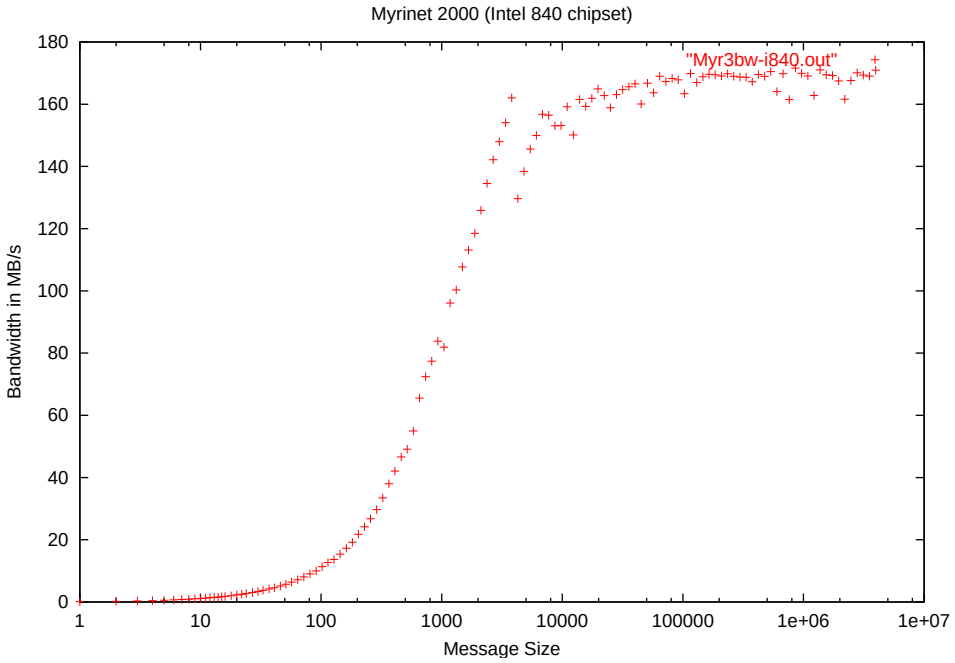
We insert also data on bandwidth and latency obtained on a Supermicro 370DLE motherboard with the Serverworks III LE chipset we had on loan some time ago. This motherboard has a PCI 64 bit/33 Mhz slot which was used to install the Myrinet board.

Bandwidth charts

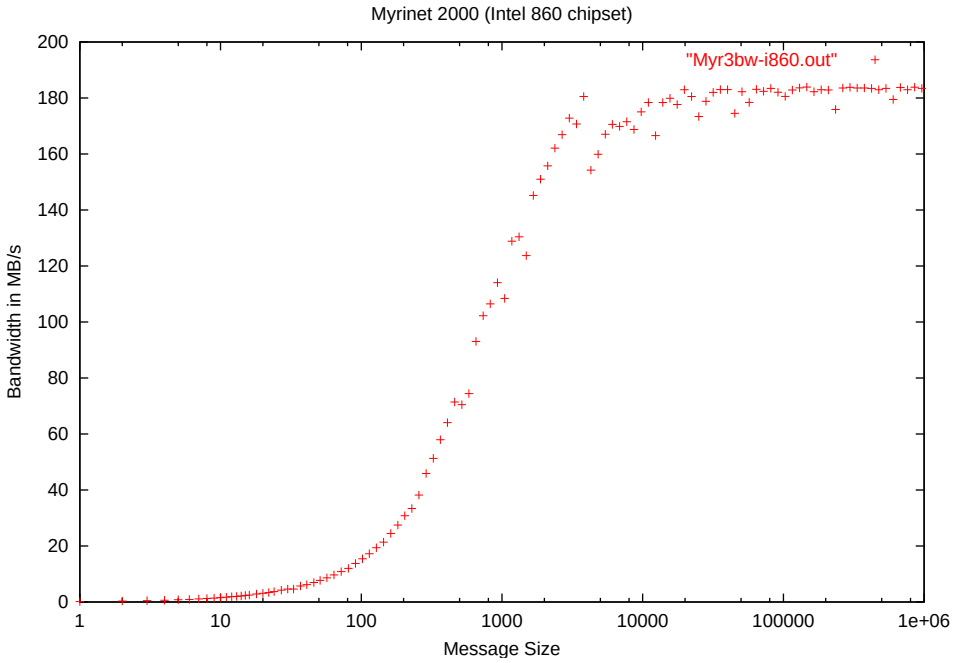
- Bandwidth with Server Works III LE chipset at <Myr3bw-370DLE.png>



- Bandwidth with Intel 840 chipset at <Myr3bw-i840.png>

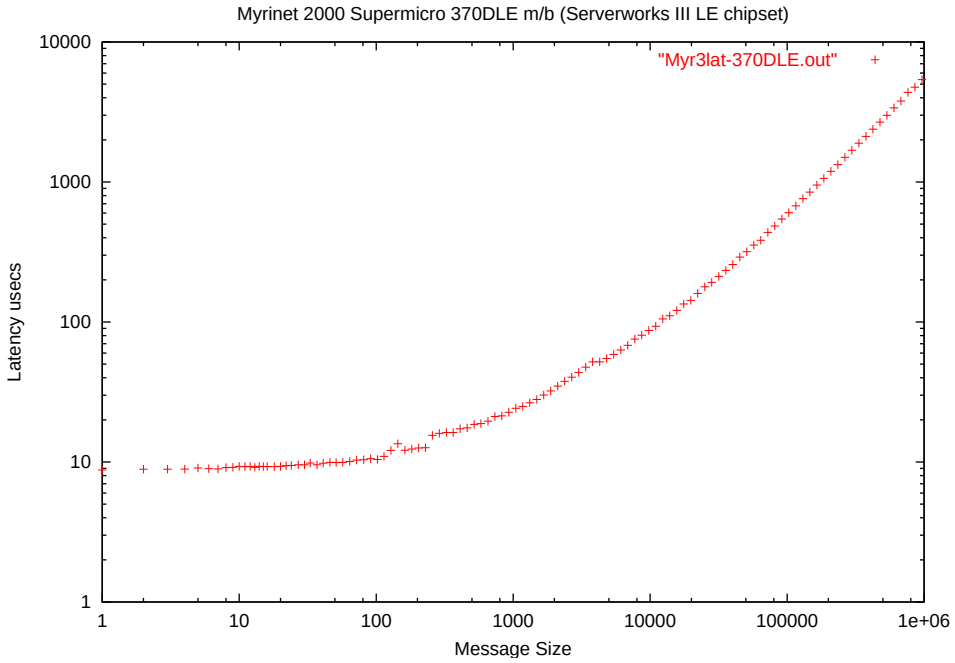


- Bandwidth with Intel 860 chipset at <Myr3bw-i860.png>

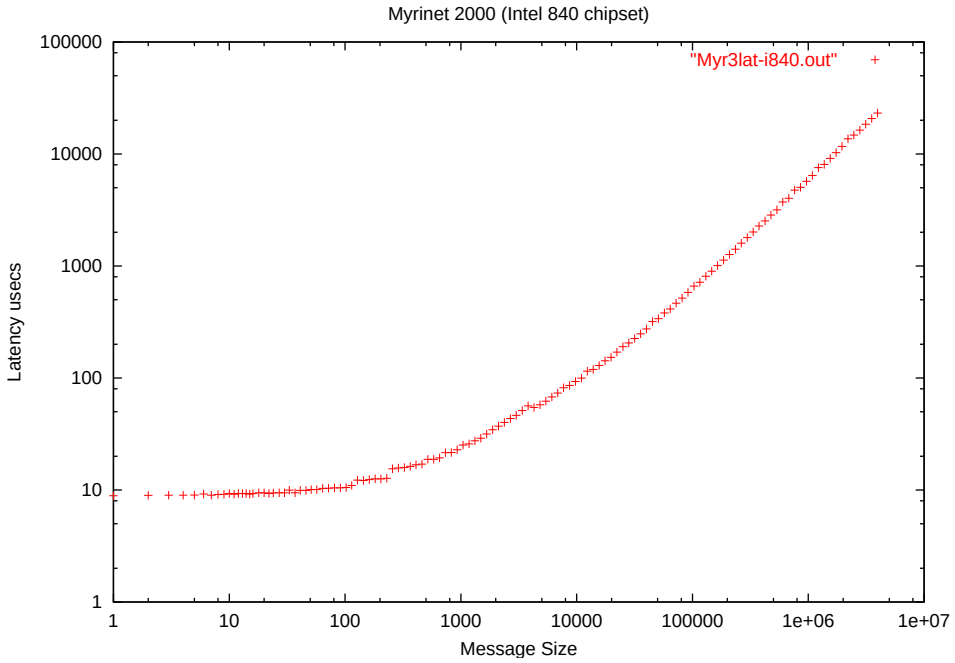


Latency charts

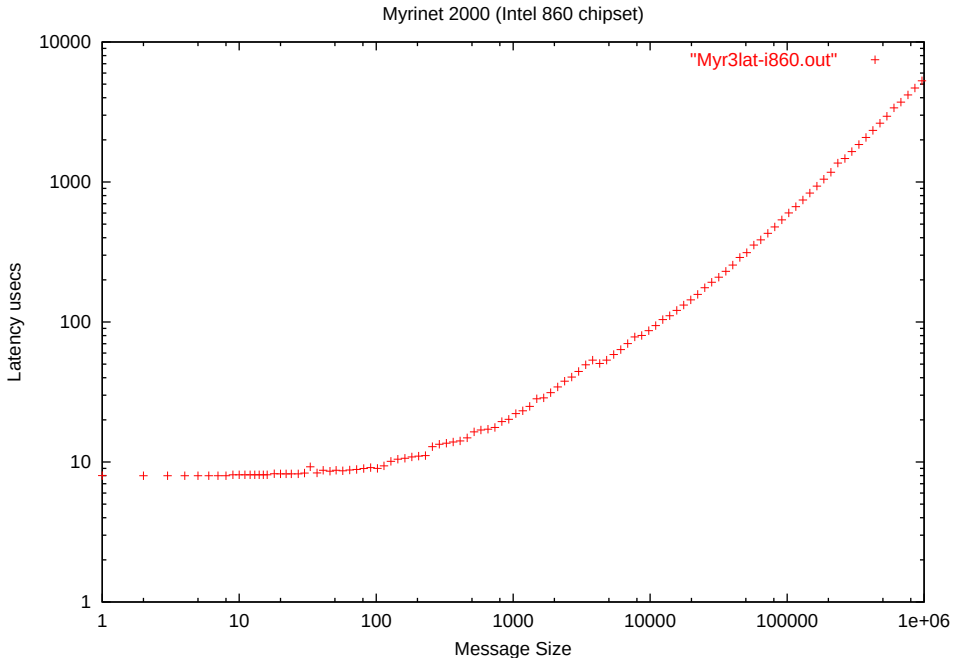
- Latency with ServerWorks III LE chipset at <Myr3lat-370DLE.png>



- Latency with Intel 840 chipset at <Myr3lat-i840.png>

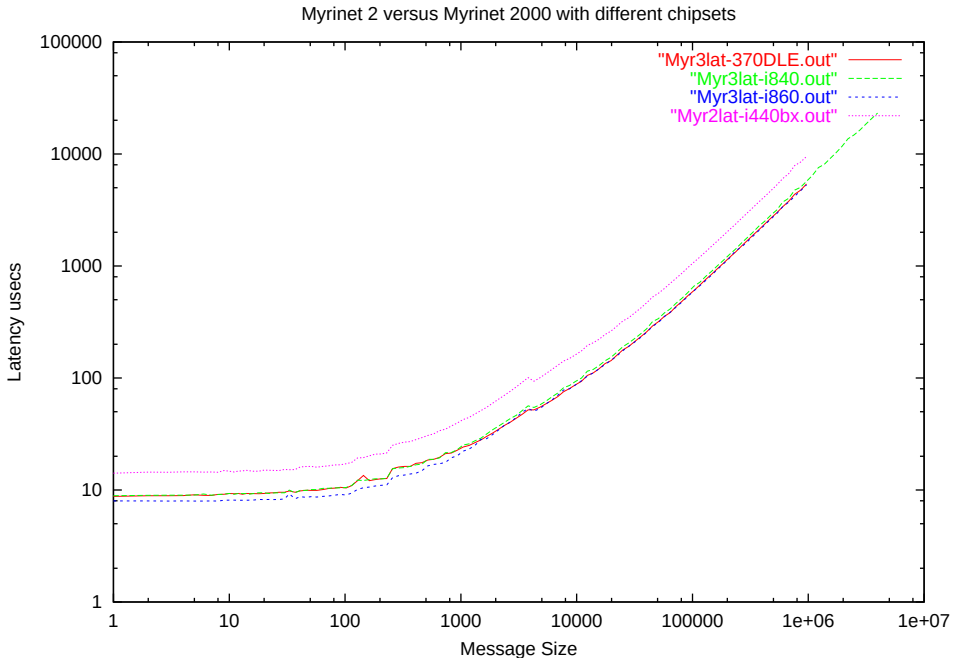


- Latency with Intel 860 chipset at <Myr3lat-i860.png>

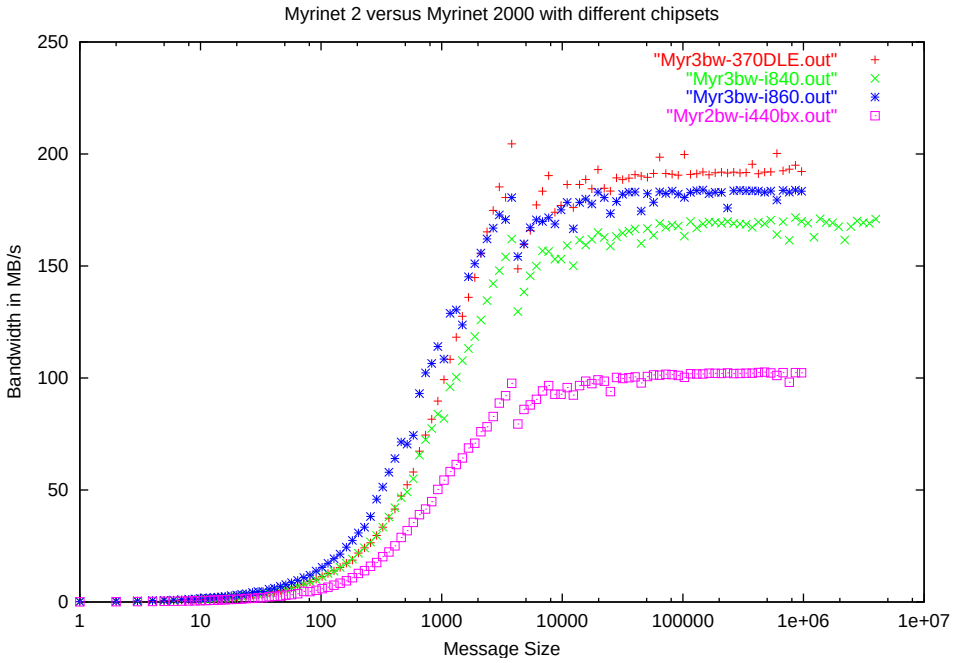


Summary comparisons

- Latency at <comparison_lat.png>



- Bandwidth at <comparison_bw.png>



Conclusions

We have seen that while there is a substantial difference of performance between the Myrinet 2 and Myrinet 2000 boards, there is no significant difference in performance on the different platforms we tested using the Myrinet 2000 boards.

We have also seen that there is no significant minimum latency difference (the one of a minimum length message) between the two Myrinet generations because of the overhead of the PCI bus.