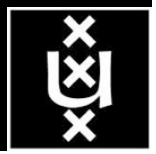
SN6-GN2 Packet Reordering Tests

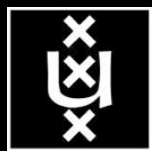
Hans Blom, Paola Grosso, Cees de Laat

University of Amsterdam



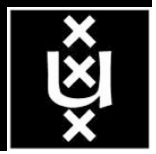
Motivation

- Migrating connections from GEANT1 to GEANT2 network.
- At the POPs GEANT2 has Juniper, M40, M160 or T640
- In January 2006 SURFnet moved to the new setup, where the new connection goes via a Juniper M160.
- The hardware architecture of the M160 makes this router prone to packet reordering. SURFnet asked us to try to quantify this effect for this particular connection
- many studies on the effect have been performed by others, see (use Google :-) :
  - M.Przybylski, B. Belter, A. Binczewski, *Shall we worry about Packet re-ordering?*



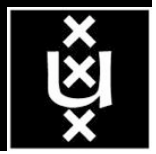
Background

- When packets reach the OC-192 interface on the Juniper M160 they are directed to four processors. Under certain circumstances the order in which the packets will leave the fabric after the processing does not match the arrival order, thus causing *packet reordering*.
- Some authors claim that the reordering is not related to the amount of traffic reaching the interface, reordering is the effect of packets of different size reaching the interface, and creating queue in the fabric,
- Others indicate that the reordering depends on the load on the interface that in most cases needs to be at least 73% of the total OC-192 capacity for the re-ordering to happen. These authors indicate that the effect of reordering is also different for the various classes of traffic.



Experiment

- 2 Hosts, one in SURFnet and one in Belgium
- SURFnet node:
  - 2.4 GHz Intel XEON processor
  - 1 GB memory
  - Debian Linux with a 2.6.14 kernel
  - 1 Gbit/s nic
- BELNET
  - 500 MHz Pentium III processor
  - 400 MB memory
  - 2.6.10 - gentoo kernel
  - 1 Gbit/s nic
- Iperf and UDPmon
  - Iperf, *http://dast.nlanr.net/Projects/Iperf/*
  - UDPmon, *http://www.hep.man.ac.uk/u/rich/net/*



Path

1 Gbit/s
10 Gbit/s
2.5 Gbit/s

The path between the two hosts in the GEANT2 setup:

1 Gi3-15-5.Bangbang.Amsterdam1A.surf.net (145.125.80.61) 0.711 ms 0.798 ms 0.887 ms

2 145.125.80.14 (145.125.80.14) 0.940 ms 1.108 ms 0.897 ms

3 AF-500.XSR01.Amsterdam1A.surf.net (145.145.80.9) 0.905 ms 1.102 ms 0.911 ms

4 gi6-0-2.ar5.amsterdam1.surf.net (145.145.166.21) 0.715 ms 0.538 ms 0.603 ms

5 PO6-0.CR1.Amsterdam1.surf.net (145.145.162.1) 0.687 ms 0.609 ms 0.621 ms

*6 PO0-0.BR1.Amsterdam1.surf.net (145.145.166.34) 0.820 ms 1.522 ms **

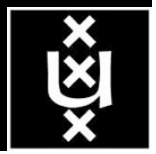
7 62.40.124.157 (62.40.124.157) 3.984 ms 0.608 ms 0.614 ms

8 62.40.124.162 (62.40.124.162) 4.017 ms 3.949 ms 4.005 ms

9 oc192.m160.core.science.belnet.net (193.191.1.1) 4.028 ms 48.850 ms 4.060 ms

10 oc48.m20.access.science.belnet.net (193.191.1.66) 4.123 ms 4.163 ms 4.121 ms

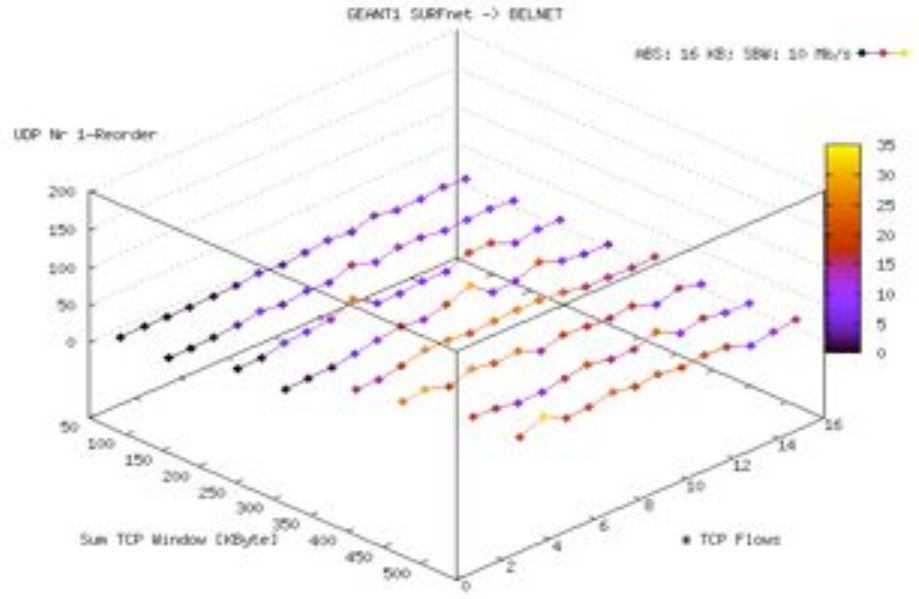
11 faro.belnet.be (193.190.198.61) 4.122 ms 4.053 ms 4.051 ms

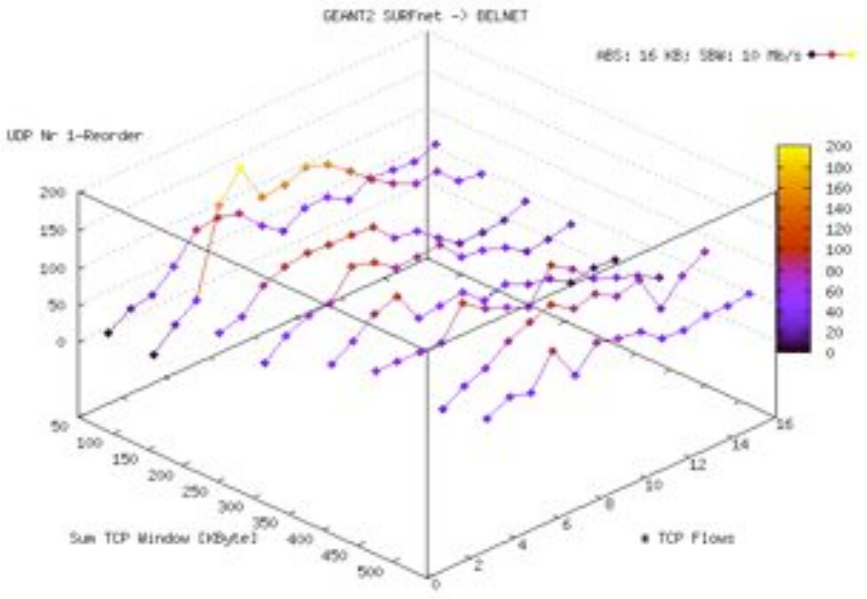


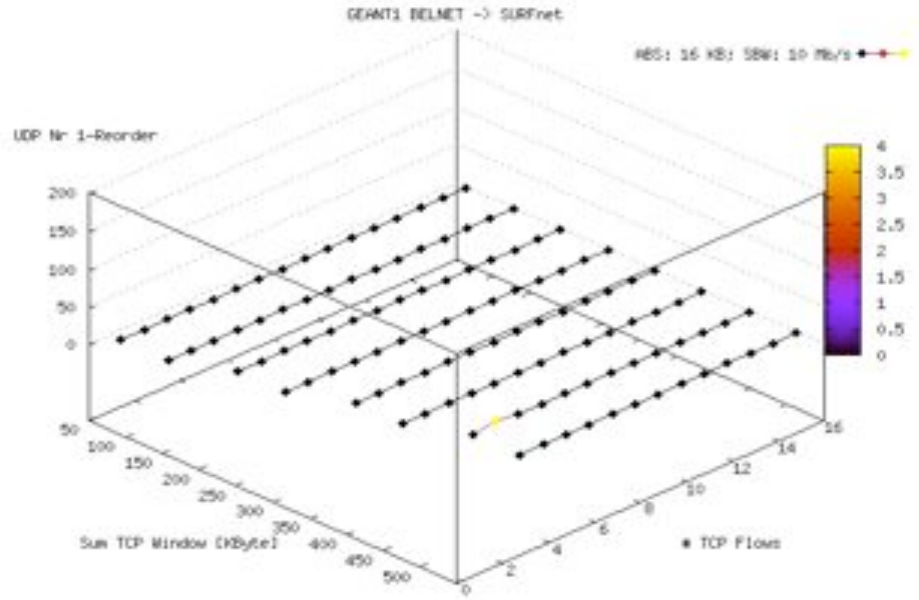
TCP throughput

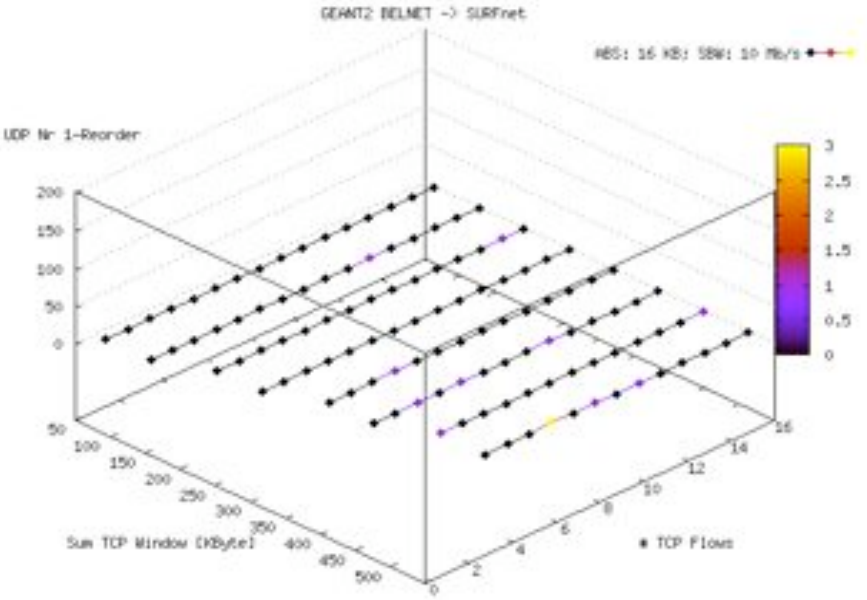
- We report the TCP throughput achieved in tests running at the same time as the UDP stream of 10 Mbit/s. We can see that:
  - there is no significant difference in the throughput between the setup via GEANT and GEANT2.
  - there is no significant difference between the two traffic directions.
- In both cases the maximum throughput that can be achieved is in the order of ~250 Mbit/s.
- The BELNET node has performance problems

UDP packet re-ordering

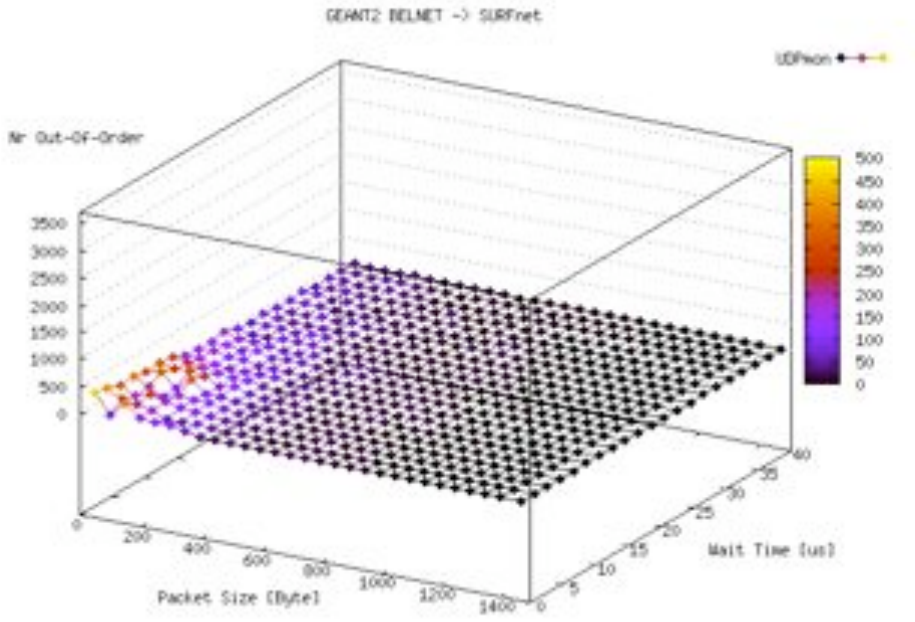


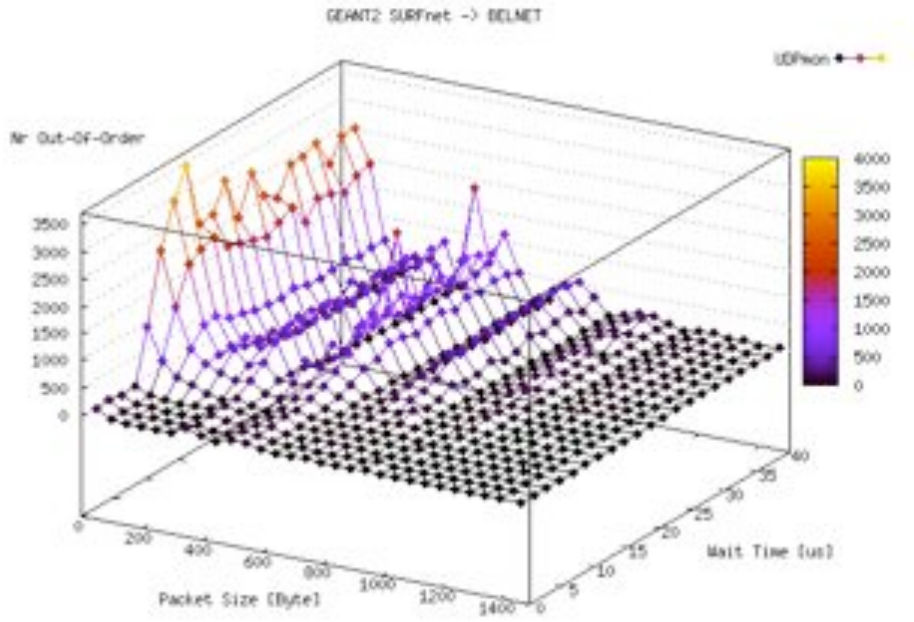






UDPmon packet re-ordering





Conclusions

- We found re-ordering in the end to end link.
- We found it hard to derive a definite conclusion on the magnitude of packet reordering in the SURFnet - GEANT2 setup.
- We cannot clearly disentangle the host from the network effects with current setup.
- With Iperf tests we have seen an increase in reordered packets in the GEANT2 setup both for TCP and UDP traffic, but this is only in the direction SURFnet → BELNET. Especially for UDP the increase GN1->GN2 has been dramatic: at times a factor of 6.
- We explain this asymmetry because of the difference in performance between the two test nodes: the packet sent out by the BELNET node have a larger inter-packet gap due to the "slower" processor than the ones sent out by the somewhat more powerful SURFnet node. The larger gap makes re-ordering less likely to happen in the direction BELNET → SURFnet, given the packets will have time to leave the four processors in the router fabric in order.
- We have also seen reordered packets with UDPmon tests and long distance tests to StarLight.
- If as some authors claim, it is really necessary to saturate the links to the M160 to quantify the reordering, the current nodes are not adequate and we will need additional test nodes.

Questions ?