Crashing Spacecraft and Crossing Continents: Summary of a Network Performance Investigation

Pekka Savola
CSC - Scientific Computing Ltd
PL405, 02101 Espoo
Finland
psavola@funet.fi

John Chevers
DANTE
City House, 126-130 Hills Road
Cambridge CB2 1PQ
UK
john.chevers@dante.org.uk

Abstract— **The GÉANT2 project established a Performance Enhancement and Response Team (PERT) process [1] in 2004 to assist network users in troubleshooting network performance issues and provide information on how to optimize network performance. This paper presents a summary and conclusions of a PERT case where the aim was to transfer data from an eVLBI radio telescope in Chile to the Netherlands across academic IP backbones. As that path goes through highly utilized, bandwidth-challenged networks, involves a significant number of operators and networks, and has a very high (300+ ms) round-trip time, engineering a high and reliable transfer rate proved to be a challenge.**

The goal of this paper is to raise awareness of the PERT process so that more users could use it to solve their performance problems and to describe general observations which are applicable to other similar usage scenarios.

Keywords: performance, PERT, eVLBI

I. Introduction

European Space Agency's SMART-1 spacecraft was scheduled to crash on Moon in September 2006, and a telescope in Chile was best placed to record the events and send the data for analysis to Joint Institute for VLBI in Europe (JIVE) in the Netherlands [2]. GÉANT2 PERT investigated performance from August to October 2006. We present findings and general conclusions from this activity.

A. Characteristics of the Case

The path from Chile to the Netherlands went through the following networks: TIGO (the observatory), University of Conception (UdeC), Chile NREN (REUNA), Latin American backbone (RedCLARA), European NREN (GÉANT2), the Netherlands NREN (SURFnet) and the Joint Institute for VLBI in Europe (JIVE).

The path had the following general characteristics:

1) very long round-trip time (310 ms),
2) most of the path is lossless and congestion-free, and
3) head-end of the path is lossy and congested.

In particular, international backbones from Brazil through the Netherlands were uncongested and lossless. REUNA in Chile was more bandwidth-constrained and multiple rate limiters had been applied: the academic network's uplink was rate-limited to a contracted rate of 90 Mbit/s. REUNA's ATM backbone was 155 Mbit/s and "download" direction was relatively heavily utilized; "upload" capacity was available though. Another rate-limiter (5 Mbit/s) was applied at the university border router.

On the other hand, measurements from a server in REUNA's offices resulted in very good performance, pinpointing the problem to either REUNA's backbone or the campus network.

In theory, after making an exception to the 5 Mbit/s rate-limiter, uplink capacity should have been adequate. Only about 15-25 Mbit/s TCP performance could be established; we'll next discuss some reasons why.

B. Limitations of the Case

After the exception in 5 Mbit/s rate-limit, UDP performance was typically good at night; lossless or almost lossless 40 Mbit/s upload could be achieved. At the same time, TCP performance was still not very good, varying between 10-25 Mbit/s. The following issues were noted:

1) The 90 Mbit/s rate-limiter between REUNA and RedCLARA had a very small (in the order of kilobytes) burst size tolerance. This had detrimental impact on large flows. The burst size was first raised and then the limiter removed completely.
2) The 5 Mbit/s rate-limiter was accidentally re-applied again after a while.
3) Traffic shaping to a 20-25 Mbit/s rate at the sending host was found to improve performance and performance stability as it constrained TCP's burstiness.
4) Linux kernel upgrade from 2.6.7 to 2.6.18 introduced features such as receive-side TCP buffer autotuning, sensible default buffer values, congestion control algorithm selection and a number of TCP bug fixes. Especially TCP receive-side buffer autotuning was found to stabilize performance.
5) Enabling SACK improved performance under loss.

II. Conclusions

A. Generic

Some of the generic findings were the following:

1) Users and operators should become more familar with PERT knowledge base [3],
2) Performance issues should be reported, and

3) Many administrative domains make debugging more difficult.

PERT KB is a collection of performance-related information and folklore. Taking a look at it and improving it (Wiki registration needed) should help a lot when facing performance issues.

The more administrative domains are involved and the more complex setup is employed, the more difficult and time-consuming the performance evaluation and improvement process will be; reserving multiple months to performance tests is not a bad idea!

A particular issue with multiple networks is that problem isolation is often very challenging. Typically it is required to have access to BWCTL [4] or iperf servers in a network to make measurements. Often access to use these servers is restricted; having more liberal policies might help a lot. Further, access to network usage graphs or other statistics (e.g., as envisioned by the perfSONAR project [5]) is very helpful in evaluating the amount of cross-traffic especially if the performance tests exhibit loss or other issues.

B. Administrative

In European networks, having gigabit or ten gigabit access or backbones is commonplace. In other parts of the world this is not necessarily the case; whole countries might still be behind low-capacity, congested links of the order of 100 Mbit/s or even less. In some cases there may be a mismatch of expectations versus reality. Such a network may have difficulties in obtaining funding for upgrades and high-bandwidth users should be willing to pay significantly more for higher capacity. More funding should be provided to the network bottlenecks or the users need to adjust their expectations to match the capacity of the network.

Also, a significant number of performance issues are caused by software or hardware configurations of equipment on the path (e.g., underperforming firewalls or CPU-based routers, existence or configuration of rate-limiters, etc.), examining the configuration might help in isolating problems much more quickly. Unfortunately many network operators are reluctant to give out this information even to the PERT teams.

C. Technical

1) Long RTT and Packet Loss: Very long RTT requires multiple megabytes of socket buffers in order to transport a high amount of data quickly. In lossless networks this is relatively straightforward, but as experienced here, even very small loss has dramatic performance for two reasons:

- Enlarging buffers even further could compensate for single losses (always data in flight) but this makes the problem worse in networks that cannot support resulting higher rates due to congestion, and
- Signalling the packet loss and reaction takes (proportionally) much longer and may lead to "silent" periods in data transmission.

For example, if you needed to transfer 4MB real-time data every second (32 Mbit/s), a packet loss burst at 30 ms would require 34 Mbit/s capacity, while at 300 ms RTT would require about 80 Mbit/s to compensate for the silent period caused by the loss.

2) TCP Performance on Lossless Links: Another observed phenomenon was that UDP performance tests might indicate lossless or almost lossless (e.g., just one or two losses in a 2 minute test run) performance, even up to 80 Mbit/s, while the TCP performance was still poor at 10-25 Mbit/s.

At the time of writing, this is still being investigated. Very likely this is caused by a performance bottleneck at a TCP stateful firewall or router somewhere in the path (and similar inspection is not done with UDP). Such a big difference cannot be explained by TCP's more burstier nature compared to UDP. Indeed, firewalls or underperforming (software) routers have often been identified as performance bottlenecks in the past.

3) Rate-limiters and Configuration: Especially in bandwidth-challenged parts of the network, rate-limiters are usually used to promote fairness among users. Identifying the existence of (especially TCP) rate-limiters can often be tedious work, particularly if the network is suffering from other issues (such as congestion loss). Proper configuration of limiters (e.g., adequate burst sizes) is also important and difficult to identify.

4) Would a Separate Connection Have Helped?: A separate connection (e.g., a lambda) from TIGO in Chile to an adequately provisioned network (straight to RedCLARA PoP in Brazil) would likely have helped to avoid performance bottlenecks in the university campus and national backbone, but the cost would likely have been higher than upgrading the backbone in Chile (in progress as of this writing). As such, in a situation like this, money seems better spent in upgrading the main network instead of building separate connections.

5) Application and Transport Protocols: A challenging, atypical environment such as experienced stresses the limits of transport protocols and applications. To gain optimal performance, transport protocols may need to be tuned and applications designed with the environment in mind.

III. Acknowledgements

Pekka Savola was visiting and funded by SWITCH while working on the PERT case. Many people from the PERT duty case manager shift participated in the investigation of the case; in particular, Simon Leinen's contributions are acknowledged. We also thank Sergio Sobarzo of TIGO, Sandra Jaque of REUNA, and Arpad Szomoru of JIVE in resolving the case.

References

[1] DANTE. GÉANT2 Performance Enhancement and Response Team. [Online]. Available: http://www.geant2.net/server/show/nav.874

[2] J. Chevers. PUB-06-151v3: Report Describing Research Network Support for eVLBI Tracking of the SMART-1 Spacecraft.

[3] GÉANT2 PERT Knowledge Base. [Online]. Available: http://kb.pert.switch.ch

[4] Internet2. Bandwidth Test Controller (BWCTL). [Online]. Available: http://e2epi.internet2.edu/bwctl/

[5] PERFormance Service Oriented Network monitoring ARchitecture (perfSONAR) project. [Online]. Available: http://www.perfsonar.net

IV. Vitae

A. Pekka Savola

Pekka Savola graduated from Helsinki University of Technology in April 2003, and is continuing with post-graduate studies. On the side, he has worked for CSC - Scientific Computing for over 5 years after other work in the networking field. He has been very active in the IETF during the last couple of years, mostly in areas related to IPv6, multicast and routing, authoring over a couple of dozen internet-drafts and otherwise contributing. He is a member of IETF operations and addressing directorates, and a co-chair of the v6ops working group.

B. John Chevers

John Chevers completed a BSc in Physics and an MSc in Optoelectronics at the University of London. His PhD from the University of Strathclyde used Time-Resolved Fluorescence Spectroscopy to study Sol-Gel Glasses. Subsequently he worked as a Scientist and Project Coordinator at Marconi Optical Components and Bookham Technology, producing optoelectronic modules for the telecom industry.

John joined DANTE in August 2004 as a Project Manager for user support. He is also project co-ordinator for the ORIENT link to China and an activity leader in the Radio Astronomy EXPReS project.