GWD-TYPE Jon Crowcroft, Cambridge
Category:Informationa;
GHPN [if applicable: Revised DATE]

1. Grid Issues with Network Infrastructure

1.1 Status of This Memo

This memo provides information to the Grid community on topics in the area of high performance network research that the network community feel need attention. It does not define any standards or technical recommendations.] Distribution is unlimited.

1.2 Copyright Notice

Abstract

Grid Issues with Network Infrastructure that the network community might prioritize – there is a sister document which contains discussion of network problems that the Grid community perceive as critical.

1.3 Contents

Packet templates make Code complexity a lot lower in the common case even for a big protocol like TCP or SCTP. "User space" v. kernel myths - in this author's experience it is still today really worth getting people to put transports into the kernel - reasons include independent failure of application and protocol as well as good control of end system resources. It ain't that

This document summarizes networking issues identified by the Grid community.

1.4 Author Information

J. Crowcroft (Editor), Computer Laboratory, University of Cambridge

2. Introduction

The Grid High-Performance Networking (GHPN) Research Group focuses on the relationship between network research and Grid application and infrastructure development. The vice-versa relationship between the two communities is addressed by two documents, each of it describing the relation from the particular view of either group. This document summarizes Grid issues identified by the Network community.

2.1 Scope and Background

Grids are built by user communities to offer an infrastructure helping members to solve their specific problems. Hence, the geographical topology of the Grid depends on the distribution of the community members. Though there might be a strong relation between the entities building a virtual organization, a Grid still consists of resources owned by different, typically independent organizations. Heterogeneity of resources and policies is a fundamental result of this. Grid services and applications therefore sometimes experience a quite different resource behavior than expected. Similarly, a distributed infrastructure with ambitious service demands puts stress on the capabilities of the interconnecting network more than other environments. Grid applications therefore often identify existing bootlenecks, either caused by conceptual or implementation specific problems, or missing service capabilities. Some of these issues are listed below.

This is a second draft contribution for a document for the GHPNRG http://www.ggf.org/6_DATA/gridhigh.htm which is meant to list topics that the network community is working on and is sometimes asked questions about by folks who make intensive use of networks, such as Global GRID Forum people.

3. Congestion Control (contrariwise: see QoS)

3.1 Slow Start

Is this always necessary? no, but beware of ISPs who mandate it, and if you think you can use less than recent history rather than recent measurements, look at the Congestion Manager and TCP PCB state shearing work first!

3.2 Congestion Control

This is not optional in a non QoS network (which is just about any network) - adaption is mandatory

3.3 AIMD and Equation Based

AIMD is not the only solution to a fair, convergent control rule for congestion avoidance and control. Other solution are around - Rate based, using loss, or ECN feedback, can work to be TCP fair, but not generate the characteristic Saw Tooth.

3.4 Assumptions and errors

Most *connections* do not behave like the Padhye equation, but most bytes are shipped on a small number of connections , and do - c.f. Mice and Elephants.

The jury is still out on whether there are non greedy TCP flows (ones who do not have infinite sources of data at any moment)

3.5 Ack Clocking

Acknowledgements clock new data into the network - aside from rare (mainly only on wireless nets) ack compression, this provides a rough “conservation” law for data. It is not a viable approach for unidirectional (e.g. multicast) applications.

3.6 RMT and Unicast

Reliable Multicast Transport protocols (PGM, ALC) use a variety of techniques to mimic TCP mainly.

3.7 Mobile and Congestion Control

Mobile nodes experience temporary indications of loss *and* congestion during a hand-off. People have proposed mechanisms for indicating whether these are “true” or chimera.

3.8 Economics, Fairness etc

Congestion control results in an approximately fair distribution of bottleneck bandwidth - this may not be great if you paid more to get a fat pipe to the net. But, you are probably nearer the core and have every right to ask the ISP to upgrade their bottlenecks anyhow and the people that paid less should be bottlenecked at **their** access links in that case. So?

http://www.psc.edu/networking/tcp_friendly.html

3.9 Observed Traffic

Observsations (see many IMW papers) are that traffic is currently mainly made up of mice (small, slow) flows and elephants (large, fast, long) flows at the individual 5-tuple level, and at the POP aggregate level.

4. Routing

Priorities for good routing system design are:

1. Fast Forwarding
   Packet classification and switched routers have come a long way recently - we are unlikely in the software world to beat the h/w in core routers, but we can compete nicely in access devices - certainly, there is no reason why a small cluster couldn;t make a good 10Gbps router - but there’s every reason why a PCI bus machine maxes out at 1Gbps!
2. Faster Convergence

   Routers and links fail. the job of OSPF/ISIS and BGP is to find the alternate paths quickly - in reality they take a whole to converge - IGPs take a while (despite being mainly link state nowadays) because link failure detection is NOT obvious - sometimes you have to count missed HELLO packets (since some links don’t generate an explicit clock). BGP convergence is a joke. But there are smart people on the case.
3. Theory and practice

   Most the problems with implementing routing protocols are those of classic distributed (p2p/autonomous) algorithms: dealing with bugs in other peoples implementations - it takes a

good programmer about 3 months to do a full OSPF. It then takes around 3 years to put in all the defences.

4. Better (multi-path, multi-metric) routing

   Equal cost Multipath OSPF and QOSPF have been dreamt up - are they used a lot? multipath in limited cases appears to work quite well. Multimetric relies on good understanding of traffic engineering and economics, and to date, hasn't seen the light of day. Note that also, in terrestrial tier one networks, end-to-end delays are approaching transmission delays, so asking for a delay (or jitter) bound is getting fairly pointless - asking for a throughput guarantee is a good idea, but doesn't need clever routing!

5. Does MPLS Help? No, not one bit.

   Policies are hard - BGP allows one to express unilateral policies to the planet. this is cute (the same idea could be used for policy management of other resources like CPUs in the GRID) however, it results in difficulties in computing global choices (esp Multihoming) - there are fixes.

   http://www.potaroo.net/
   http://www.telstra.net/gih
   NANOG
   See also Overlays (e.g. RON, and "underlay" routing in planetlab).

5. Packet Sizes

Go faster LANs have always pushed the MTU up - since ATM LANs (remember the fore asx100) we tried 9280 byte packets, and enjoyed things. But the GRID is global, so the MTU is that of the weakest link. Most stuff is on 100BaseT somewhere on the path so we aren't likely to see more than the occasional special case non 1500 byte path. However, with path MTU discovery, we get that auto-magically

5.1 5 Multicast MSS is a real problem:)

Sub-IP packet size is a consideration - some systems (ATM) break packets into tiny little pieces, then apply various level 2 schemes to these pieces (e.g. rate/congestion control) – most these are anathema to good performance.

http://www.nlanr.net/NA/Learn/packetsizes.html
http://www.faqs.org/rfcs/rfc1191.html
etc

6. Overlays

Overlays and P2p (e.g. Pastry, CAN, Chord, Tapastry, etc) are becoming commonplace - the routing overlay du jour is probably RON from MIT - these (at best) are an auto-magic way of configuring a set of Tunnels (IPinIP, GRE etc). I.e. they build you VPNs In fact routing overlays may be a problem if there is more than one of them (see SIGCOMM 2003 paper on selfish routing). But there are moves afoot to provide one (e.g. see SIGCOMM paper on underlays).

P2P: are slightly different - they do content sharing and have cute index/search/replication strategies varying from mind-numbingly stupid (napster, gnutella) to very cute (CAN, Pastry). They have problems with Locality and Metrics so are not the tool for the job for low latency file access....in trying to mitigate this , they (and overlay routing substrates) use ping and

pathchar to try to find proximal nodes: c.f. limitations of Ping/Pathchar convergence when not native (errors/confidence)

Peer-to-Peer Harnessing the Power of Disruptive Technologies

Edited by Andy Oram, March 2001, 0-596-00110-X

7. QoS (contrariwise: see Congestion Control)

QoS would be a nice thing. THere are many fine papers on QoS, but few describe anything anyone has deployed:-)

Parameters typically include

- Throughout
- Delay
- Availability
- Some people add security/integrity
- Some people also mention loss...
  Threats: Theft and Denial of Service

Protection is really what people want - If I send x bps to site S, what y bps will be received, ow much d later?

To guarantee y=x, and d is minimised, you need:

- Admission Control (so we are not sharing as we would if we adapted under congestion control)
- Scheduling (so we do not experience arbitrary queueing delays)
- Re-routing may also need to be controlled and pre-empted: alternate routes (also known, unfortunately as protection paths) may be needed if we want QoS to include availability as well as throughput guarantees and delay bounds.

8. Network Structure

"edge", "core", etc is a myth :- in the global net the average traffic path includes 7 ASs - most inter-domain traffic traverses heavily used Internet Exchange points (e.g. London) where capacity only just about matches demand, whereas core networks are often "over-provisioned" (UK academic net now runs at <5% utilisation).

Aggregation is a technique to scale traffic management for QoS - by only managing classes of aggregates of flows, we get to reduce the state and signaling/management overhead for it. VPNs/tunnels of course are aggregation techniques, as are things that treat packet differently on subfields like DSCP, port etc etc

SLAs are around already despite non widespread QoS - however, SLAs are only intra-ISP to my knowledge (some Internet Exchanges offer SLAs but end 2 end SLAs are as scarce as dragons).

9. Economics - are important here again as you can imagine!

An Engineering Approach to Computer Networking
Keshav, 1997, Addison-Wesley Pub Co; ISBN: 0201634422
or
Internet QoS: Architectures and Mechanisms for Quality of Service
by Zheng Wang, 2001, Morgan Kaufmann Publishers; ISBN: 1558606084

10. Multicast

10.1 Tier 1 routing works. Most ISPs run core native multicast

- Interdomain only just limps (its getting better...MSDP Problems, App Relay Solutions
- RMT - we have some candidate protocols for reliable multicast - nothing as solid as 1988 TCP quite yet tho.
- Address Allocation and Directories are not great yet, hence beacons and so on.
- Access Network are in bad shape...e.g.
- DSLAMs dont do IGMP snooping
- Cable dont do IGMP snooping
- Dialup cant hack it at all

10.2 Does IPv6 Help (don't laugh!) - yes it might!

Developing IP Multicast Networks: The Definitive Guide to

Designing and Deploying CISCO IP Multicast Networks

by Beau Williamson, 2000, Cisco Press; ISBN: 157870077

and

Multicast Communication: Protocols, Programming, and Applications

by Ralph Wittmann, Martina Zitterbart

Morgan Kaufmann Publishers; ISBN: 1558606459

11. Operating Systems

Linux, Solaris etc...there's a lot we could say here - lots of things can and should be configured - see www.psc.edu for a LOT Of help.

Zero copy stacks:- we'd all like this - zero copy receive is hard;

RDMA is not obviously the answer

Interrupts (self selecting NICs) we should minimises these if we want TCP to go to 10Gbps on a reasonable processor - there are nice techniques - these are configuarable socket buffer considerations -there are lots!
Protection and scheduling domains - if we could get away from OSs that confused these , life would be easier!

If all these were auto-magically set, life would be a lot easier.

W Richard Stevens, TCP/IP Illustrated, All Volumes.
and
Understanding the Linux Kernel,
D.P. Bovet and M. Cesati, O'Reilly, 2001,
ISBN 0-596-00002-2

12. Layer 2 Considerations

12.1 Layer 2 NBMA nets - lots - a pain

> Layer 2 shared media nets - was decreasing due to switched ether, now increasing due to wireless. Switching and routing re-cursed - layer 2 switching and routing usually makes life HARDER for the IP engineer. Flow and congestion control re-cursed - layer 2 reliability and flow control almost ALWAYS make life worse for the IP and TCP engineer.

Jon.Crowcroft@cl.cam.ac.uk

Signaling (implicit, explicit) is just painful.

802.11 - in its glory:

http://www.apple.com/ibook/wireless.html

General discussion of slow lossy links:

http://www.ietf.org/html.charters/pilc-charter.html

12.2 WAP horrors - see web for many stories
GPRS - see:

http://www.cl.cam.ac.uk/Research/SRG/netos/coms/index.html

Other end of "Spectrum", see http://www.cis.ohio-state.edu/~jain/refs/opt_refs.htm (includes Raj Jain's own list of hot topics!)

13. Light v. Heavyweight Protocols

13.1 Header prediction.

Packet templates make Code complexity a lot lower in the common case even for a big protocol like TCP or SCTP. "User space" v. kernel myths - in this author's experience it is still today really worth getting people to put transports into the kernel - reasons include independent failure of application and protocol as well as good control of end system resources. It ain't that hard and user space will just almost never be as fast. It is true that one day, we will have novel OS structures that make user space stacks work well - this is true in our Computer Laboratory, but not in the wild, yet.

Computer Networks, A Systems Approach Peterson and Davie, Morgan Kaufmann, 1996, ISBN 1-55860-368-9 (2nd ed. too)

14. Macroscopic Traffic and System Considerations

Self similarity, so? traffic is self similar (i.e. arrivals are not i.i.d) - this doesn't actually matter much (there is a horizon effect)

Traffic Phase Effects: p2p (IP router, multiparty applications etc) have a tendency (like clocks on a wooden door, or fireflies in the mekong delta) to synchronise :- this is a bad thing

14.1 Flash Crowds
e.g. genome publication of new result followed by simultaneous dbase search with similar queries from lots of different places...

14.2 11.3 Asymmetry

Many things in the net are asymmetric - see ADSL lines, see client-server, master-slave, see most NAT boxes. See BGP paths. beware - assumptions about symmetry (e.g. deriving 1 way delay from RTT) are often wildly wrong. Asymmetry also breaks all kinds of middle box snooping behaviour.

15. Security Considerations

Security is under consideration in several sections, particularly denial and theft of service, as well as network Performance impact of different use and misuse.

Author Information

Jon Crowcroft
Computer Laboratory
University of Cambridge
Gates Building
17 JJ Thomson Avenue,
Cambridge CB3 0FD
UK

Jon.Crowcroft@cl.cam.ac.uk

Glossary

TLA= Three Letter Acronym

Intellectual Property Statement

Full Copyright Notice

References

The Art of Computer Systems Performance Analysis

Raj Jain, 1991, Wiley, ISBN 0-471-50336-3

Web Protocols and Practice
B. Krishnamurthy & J. Rexford,
Addison Wesley, 2001, ISBN 0-201-710885

Security Engineering, Ross Anderson, 2001 Wiley & Sons; ISBN: 0471389226

Global Reference:

ACM CCR 25th Anniversary Edition,
ACM SIGCOMM CCR, Volume 25, No.1 January 1995,
ISSN #: 0146-4833
http://www.acm.org/sigcomm/ccr/archive/ccr-toc/ccr-toc-95.html

J. Sterbenz, J. Touch,

"High-Speed Networking: A Systematic Approach to

High-Bandwidth Low-Latency Communication," John Wiley & Sons, April

2001, ISBN: 0471330361.

http://catalog.wiley.com/remtitle.cgi?isbn=0471330361&country=826

"Computational Grids: The Future of High-Performance Distributed Computing,"
eds. I. Foster and C. Kesselman, Morgan Kaufmann, ISBN 1-55860-475-8, July 1998.
http://www.mkp.com/grids/