Practical Issues Associated With 9K MTUs

I2/NLANR Joint Techs, Miami, 4 Feb 2003

Joe St Sauver, Ph.D. (joe@oregon.uoregon.edu)
Director, User Services and
Network Applications
University of Oregon Computing Center

http://darkwing.uoregon.edu/~joe/jumbos/

Introduction

- I became interested in so-called “jumbo frames” in conjunction with running UO’s Usenet News servers, having heard many wonderful things about how they might improve the performance of my boxes.
- I’ve learned (the hard way) that jumbo frames can be a difficult technology to deploy in the wide area for a variety of reasons. We’ll talk about those reasons in the remainder of this talk.

Talk Timing/Length

- This talk is probably longer than it should be for the allotted time (particularly right before lunch).
- We'll cover what we can until it is time for lunch, then we'll quit wherever we're at (I promise). Chow comes first. :-)
- I've built these slides with sufficient detail that they should be self-explanatory if studied independently *post hoc*.

"Sell me on jumbo frames!?!"

- Let me make this absolutely clear: I'm not here to "sell you" on doing jumbo frames -- when all is said and done, you might (or you might not) want to do jumbo frames. Only you can make that decision.
- I do want you to know about practical issues associated with trying to do jumbo frames, practical issues that may impact your decision about the issue.
- Let's begin by reviewing frame sizes.

Section 1. Frame Sizes

Normal ethernet frames

- Normal standards-compliant IEEE-defined* ethernet frames have a maximum MTU of 1500 bytes (plus 18 additional bytes of header/trailer for srcaddr, dstaddr, length/type, and checksum).

  * http://standards.ieee.org/getieee802/download/802.3-2002.pdf at 3.1.1, 4.4.2.1, 4.4.2.3, and 4.4.2.4

A sidenote on frame size nomenclature

- It is common to see normal ethernet frame sizes quoted both as 1500 (w/o headers) and 1518 (with headers)
- Some vendors do unusual things; e.g., Juniper talks about 1514 rather than 1518 (excluding just the 4 byte FCS of ethernet frames when specifying MTUs; see http://www.juniper.net/techpubs/software/junos/junos56/swconfig56-interfaces/html/interfaces-physical-config5.html)

Ethernet frames larger than 1518 bytes DO exist...

- All how-do-you-want-to-count-em issues aside, frames larger than 1518 do exist...
- For example, 802.1Q/802.3ac tagging increases the size by 4 bytes to 1522 bytes
- Another example: Cisco InterLink Switch Frame Format takes the max encapsulated ethernet frame size out to 1548 bytes
- Frames of this sort just slightly >1518 are called “baby giant” or “baby jumbo” frames

And of course non-ethernet frames may be larger still:

- -- FDDI IP MTU of 4352 bytes (per RFC1390)/4470 (in practice)
  -- Standard POS links with 16 bit CRCs typically have maximum receive unit (MRU) values of 4470; with CRC-32, 9180 octets.
  -- ATM (Cisco default of 4470, 9180 per RFC2225)
  -- Fibre Channel (RFC2625): 65,280, etc.

You will also see ethernet MTUs less than 1500 bytes...

- Normal 1500 byte ethernet MTUs can get reduced by a variety of events, for example they can become reduced when you tunnel traffic using PPPOE, a GRE tunnel, or some other sort of encapsulation:

  -- PPPOE (RFC2516), as currently used by many dialup and broadband ISPs): 1500 byte MTU's become 1492 bytes

  -- GRE tunnels (RFC2784): 1500-->1476

9K MTUs (“jumbo frames”)

- And then there are frames that are six times the size of normal ethernet frames (9180 bytes long), so-called “jumbo frames,” the target of today’s talk.
- 9180 is also noteworthy because it is the MTU of the Abilene backbone

Some benefits of jumbo frames

- Reduced fragmentation overhead (which translates to lower CPU overhead on hosts)
- More aggressive TCP dynamics, leading to greater throughput and better response to certain types of loss.
- See:
  http://sd.wareonearth.com/~phil/jumbo.html
  http://www.psc.edu/~mathis/MTU/
  http://www.sdsc.edu/10GigE/

Section 2. Are Jumbo Frames Actually Seen "In the Wild" on Abilene?

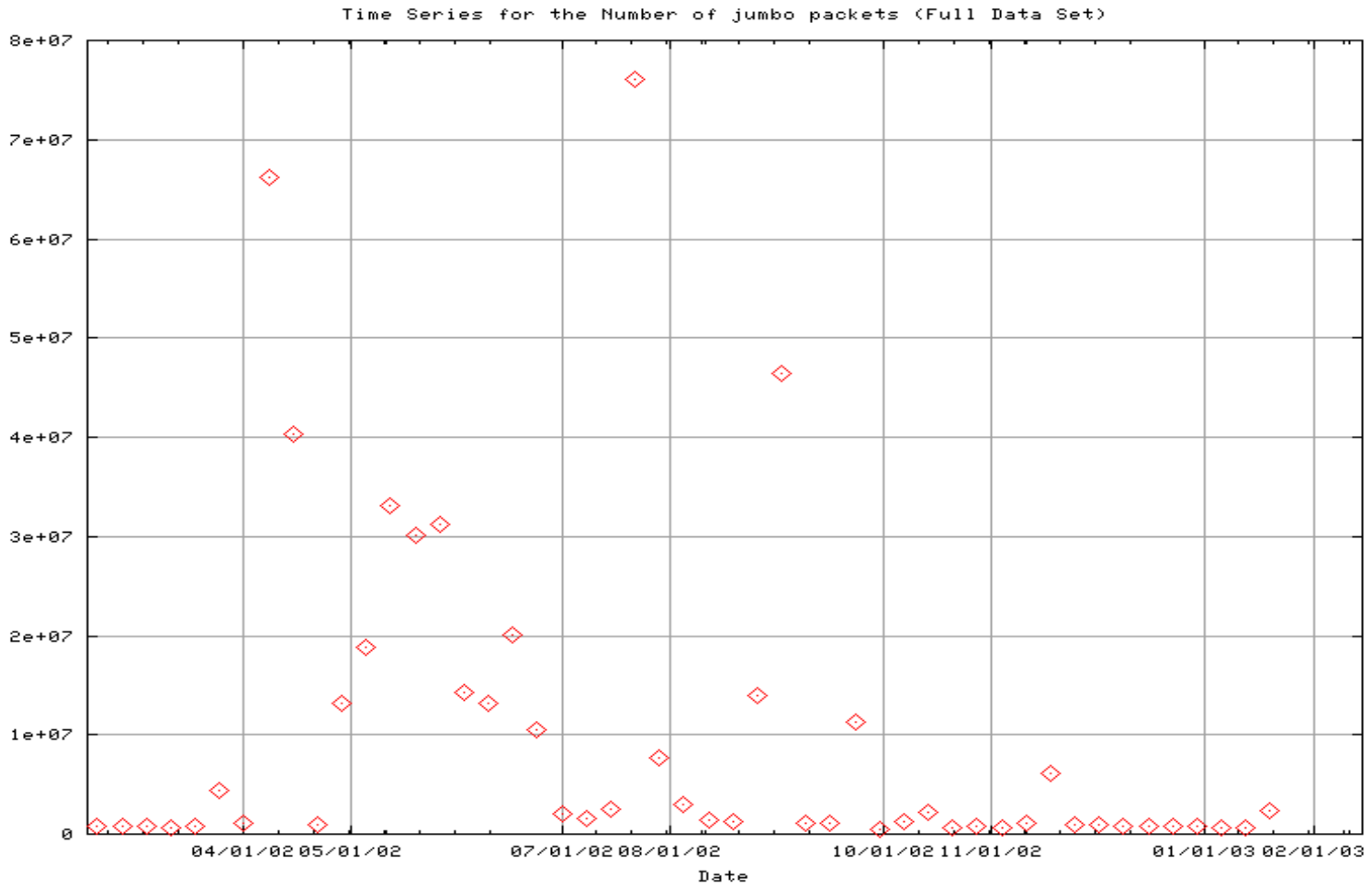
The light's green, but...

- The Abilene backbone supports jumbo frames on all nodes under normal operational conditions [one link was recently temporarily constrained to 8192 due to a multicast bug]
- Jumbo frames have been publicly endorsed by I2 (e.g., see: http://www.internet2.edu/presentations/spring02/20020508-HENP-Corbato.ppt)
- But how much jumbo frame traffic are we actually seeing on Abilene? Virtually none.

I2 Netflow Packet Size Data

- For example, if you check http://netflow.internet2.edu/weekly/20030113/#full_packsizes you'll see that out of 144.3G packets, only 704.4K packets were larger than 1500 octets ("<0.00%" of all packets) during that week.
- We really don't know if those packets are 4470 or 9180 octets or … but at one level, that detail really doesn't matter -- what is key is that there's virtually nothing >1500.

And jumbo frame traffic levels have been routinely low...



http://netflow.internet2.edu/weekly/longit/jumbo-packets.png

Putting the pieces together:

- If we believe:
  -- the Abilene backbone itself (and I2 as an organization) support jumbo frames and
  -- jumbo frames are generally a good idea
  -- but we aren't seeing widespread use of jumbo frames at the current time and
  -- use of jumbo frames doesn't appear to be trending up in any systematic way…

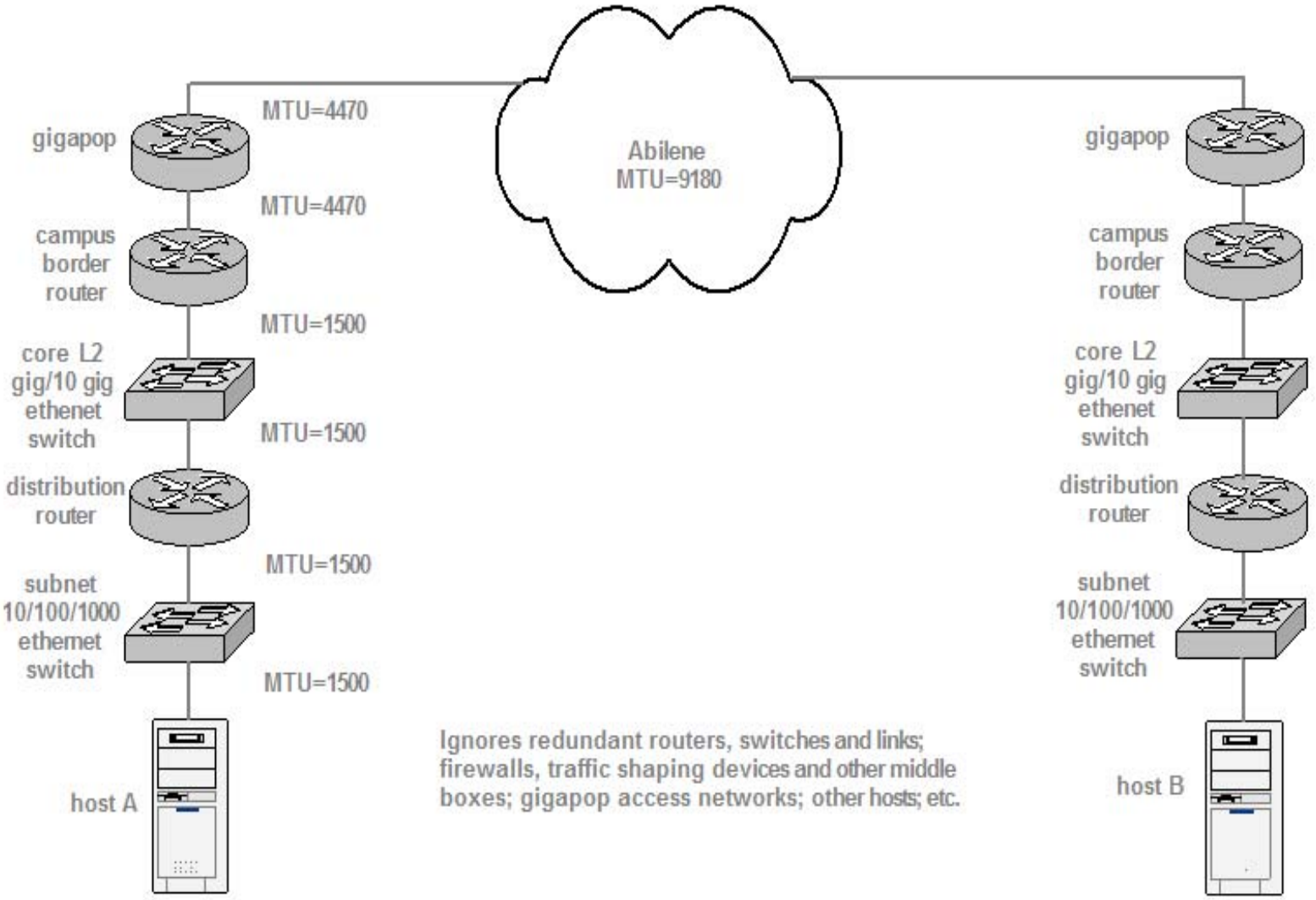
  It is then reasonable to assume that a systematic practical problem exists.

Section 3. Understanding the Absence of Jumbo Frames on Abilene

Rule #1:

- **The smallest MTU used by any device in a given network path determines the maximum MTU (the MTU ceiling) for all traffic travelling along that path.**
- This principle dominates ANY effort to deploy jumbo frames.
- Consider, for example, a typical idealized conceptual network interconnecting host A and host B across Abilene….

Idealized conceptual network



So, in our hypothetical conceptual network...

- Even though the Abilene backbone can support **9180** byte MTU traffic, and
- Even though our hypothetical router-to-router links are able to support at least **4470** byte MTU traffic,
- The default 1500 byte MTU of the ethernet switches and the ethernet NIC in our hypothetical network means our traffic will have a maximum frame size of **1500** bytes.

And this doesn't even consider the guys on the other end...

- …who will likely also have one or more network devices in the path that use an MTU of 1500 (or less).
- Of course, since Rule #1 applies from end to end, even after you fix your network to cleanly pass jumbo frames, if your collaborators haven't, you will still be constrained to normal frame MTUs to those hosts.

Digging In Systematically

- If we want to discover the choke points I2 users face in doing jumbo frames, we need to dig in systematically.
- The first possible culprit lies at the Gigapop/Abilene direct connector level.

Section 4. The Gigapop (and Abilene Direct Connector) Level

Could the problem be at the Gigapop/direct connector Level?

- We know that the Abilene backbone is jumbo frame enabled, so the binding constraint shouldn't be found there.
- Could the problem actually be at the Gigapop/Abilene connector level?

Gigapops and Abilene direct connectors: critical gatekeepers for many downstream users

- Gigapops and direct connections to Abilene are particularly worthy of attention because they represent a critical "common point of potential failure" relevant to all downstream folks who connect via their facilities (e.g., a single Gigapop that isn't jumbo enabled can preclude use of jumbo frames for hundreds of thousands of downstream customers).

The Internet2 Router Proxy

- We used the http://loadrunner.uits.iu.edu/~routerproxy/abilene/ to investigate the interface MTUs of Abilene connectors. (v4 and v6 MTUs are explicitly broken out only when they differ for the same site)

No way to do this without naming names

- We mention specific Gigapops and connectors by name in the following section, true. That may be viewed by some as "pointing fingers," but that's not the goal. The goal is to isolate/fix MTU chokepoints.
- If it makes you feel any better, the Oregon Gigapop is right in there with many of the rest of you, NOT jumbo clean, either.
- I throw the first stone at myself. <*bonk*>

Abilene connector MTUs

- Connectors are listed in the order shown in the Abilene Core Node Router Proxy output. Down interfaces are omitted.
- Atlanta:
  -- POS 0/0 (SOX OC48): 9180
  -- POS 3/0 (UFL OC12): 4470
  -- POS 3/1 (SFGP/AMPATH OC12): 4470
  -- POS 5/2 (USF OC3): 4470
  -- ATM 7/0 (MS State OC3): 4470

More connector MTUs… (1)

- Chicago Next Generation:
  -- GE-0/3/0 (Starlight 10Gig): 9192
  -- GE-0/3/0.103 (Starlight): 9174
  -- GE-0/3/0.104 (Surfnet): 1500
  -- GE-0/3/0.111 (NREN): 4470
  -- GE-0/3/0.121 (CERN 1Gbps): 9174
  -- GE-0/3/0.135 (CANet/Winnepeg): 9174
  -- GE-0/3/0.144 (CANet/Toronto): 9174
  -- GE-0/3/0.515 (CERN 10Gbps): 9174
  -- GE-1/0/0.0 (MREN): 2450

More connector MTUs… (2)

- Chicago Next Generation (cont.):
  -- SO-2/1/0 (WISCREN OC12): 9192
  -- SO-2/1/1.0 (ESNET OC12): 9180
  -- SO-2/1/2.0 (Nysernet OC12): 9180
- Denver:
  -- POS 3/0 (Arizona State OC3): 4470
  -- POS 3/1 (New Mexico OC3): 4470

More connector MTUs… (3)

- Denver Next Generation:
  -- SO-1/1/1.0 (Arizona): 4470 (v4)
  9180 (v6)
  -- SO-1/1/2.0 (Oregon OC3): 9180
  -- SO-1/1/3.0 (Utah OC3): 4470 (v4)
  9180 (v6)
  -- SO-1/2/0.0 (New Mexico): 9180
  -- SO-1/2/1.0 (Qwest Lab): 4470 (v4)
  9180 (v6)
  -- SO-2/0/1.0 (Front Range): 9180

More connector MTUs… (4)

- Houston Next Generation:
  -- SO-1/0/0.0 (Texas Tech): 4470 (v4) 9180 (v6)
  -- SO-1/0/1.0 (UT Dallas/SWMed): 9180
  -- SO-1/0/2.0 (Texas Gigapop): 4470 (v4) 9180 (v6)
  -- SO-1/0/3.0 (N. Texas Gigapop): 4470 (v4) 9180 (v6)
  -- SO-1/1/0.0 (Tulane): 4470 (v4) 9180 (v6)
  -- SO-1/1/1.0 (LAnet): 4470 (v4) 9180 (v6)

More connector MTUs… (5)

- Houston Next Generation (cont.):
  -- AT-2/3/0.18 (Texas Austin): 4470
  -- AT-2/3/0.222 (Texas El Paso): 4470
  -- AT-2/3/0.6481 (SWRI): 4470
  -- AT-2/3/0.7202 (FL A&M): 4470
- Indianapolis Next Generation:
  -- SO-1/0/0.0 (OARNet): 9180
  -- SO-1/2/0.0 (U Louisville): 4470
  -- AT-2/0/0.6 (vBNS v6 only): 4470
  -- AT-2/0/0.35 (Kreonet KR): 4470

More connector MTUs… (6)

- Indianapolis Next Generation (cont.):
  -- AT-2/0/0.145 (vBNS v4 only): 4470
  -- AT-2/0/0.293 (ESNet): 4470
  -- AT-2/0/0.297 (NISN): 4470
  -- AT-2/0/0.668 (DREN): 4470
  -- AT-2/0/0.1842 (USGS): 4470
  -- AT-2/0/0.2603 (Nordunet): 4470
  -- AT-2/0/0.3425 (6tap v6 only): 4470
  -- AT-2/0/0.3662 (HARNET): 4470
  -- AT-2/0/0.6939 (Hurricane v6 only): 4470

More connector MTUs… (7)

- Indianapolis Next Generation (cont. 2):
  AT-2/0/0.7539 (TAnet TW): 4470
  AT-2/0/0.7660 (APAN Tokyo): 4470
  AT-2/0/0.9405 (CERnet CN): 4470
  SO-2/1/0.0 (Northern Lights): 9180
  SO-2/1/1.0 (Indiana Gigapop): 9180
  SO-2/1/2.77 (Qwest): 4470 (v4) 9180 (v6)
  SO-2/1/2.512 (Merit): 4470
  SO-2/1/3.0 (NCSA): 9180

More connector MTUs… (8)

- Kansas City M5:
  AT-0/1/1.101 (Iowa State): 4470
- Kansas City Next Generation:
  SO-1/0/0.0 (Great Plains): 9180
  SO-1/0/1.0 (OneNet): 4470
  SO-1/1/0.0 (Memphis): 4470 (v4) 9180 (v6)
- Los Angeles:
  POS 2/0 (DARPA Supernet): 4470
  ATM 5/0.1 (Calren2 South OC12): 4470
  ATM 5/0.2 (CUDI OC12, Tijuana): 9180
  GE-0/1/0.0 (CalREN 10GE): 1500==>9180

More connector MTUs… (9)

- New York:
  -- POS 1/0 (DANTE-GEANT): 4470
  -- POS 4/0 (HEAnet IE): 4470
  -- POS 5/0 (ESnet): 4470
  -- POS 5/2 (DANTE-GTREN): 4470
  -- ATM 7/3.1 (HEAnet IE): 4470
- New York Next Generation:
  -- SO-0/1/0.0 (IEEAF OC192): 9176
  -- SO-1/0/0.0 (SINET OC48): 9180
  -- SO-1/1/0.0 (WPI): 9180

More connector MTUs… (10)

- New York Next Generation (cont.):
  -- SO-1/1/1.0 (Rutgers): 9180
  -- SO-1/1/2.0 (Nysernet): 9180
  -- SO-1/2/0.0 (IEEAF OC12): 9176
  -- SO-1/2/2.0 (Nordunet): 4470
  -- GE-2/1/2.0 (ESNet): 9000
  -- SO-2/3/0.0 (NOX OC48): 9180
- Sunnyvale:
  -- ATM 0/0.9 (GEMnet): 4470

More connector MTUs… (11)

- Sunnyvale Next Generation:
  -- SO-1/2/0.0 (SingAREN): 4470
  -- SO-1/2/1.0 (Oregon OC3): 4470==>9180
  -- SO-1/2/3.0 (WIDE v6 only): 4470
  -- AT-1/3/1.24 (NREN ARC): 4470
  -- AT-1/3/1.25 (NREN DX): 4470
  -- AT-1/3/1.293 (ESNet): 4470
  -- AT-1/3/1.297 (NISN): 4470
  -- AT-1/3/1.668 (DREN 668): 4470
  -- AT-1/3/1.1842 (USGS): 4470

More connector MTUs… (12)

- Sunnyvale Next Generation (cont.):
  -- AT-1/3/1.6360 (Hawaii via DREN): 4470
  -- AT-1/3/1.7170 (DREN 7170): 9180
  -- SO-2/0/0.0 (Calren North OC12): 4470
  (v4) 9180 (v6)
- Seattle:
  -- POS 4/0 (PNW): 9180
- Seattle Next Generation:
  -- GE-1/0/0.0 (Pacific Wave): 1500
  -- SO-1/2/0.0 (Hawaii): 4470

More connector MTUs… (13)

- Washington DC Next Generation:
  -- SO-1/0/0.100 (MAX OC48): 9180
  -- SO-1/1/0.0 (Drexel): 4470 (v4) 9180 (v6)
  -- SO-1/1/1.0 (Delaware): 9180
  -- SO-1/3/0.0 (PSC): 9180
  -- SO-2/0/0.0 (NCNI/MCNC): 4470 (v4)
  9180 (v6)
  -- SO-2/1/1.0 (Network Virginia): 4470
  -- SO-2/1/2.0 (MAGPI): 9180

More connector MTUs… (14)

- Washington DC Next Generation (cont.):
  -- AT-2/2/0 (UMD NGIX): 9192
  -- AT-2/2/0.1 (NISN): 4470
  -- AT-2/2/0.2 (vBNS): 4470
  -- AT-2/2/0.3 (DREN): 4470
  -- AT-2/2/0.4 (vBNS v6 only): 4470 (v4)
  9180 (v6)
  -- AT-2/2/0.5 (USGS): 4470
  -- AT-2/2/0.7 (DREN): 9000
  -- SO-3/0/0.0 (DARPA Supernet): 9180

An aside about I2 International MOU Partners using StarTap

- Traffic that's strictly between StarTap participants isn't reflected in the I2 Netflow weekly reports packet size summaries, but many I2 folks peer at StarTap or do material work with StarTap connected folks. If that's you, you may also want to investigate relevant StarTap participant MTUs. Try: http://loadrunner.uits.iu.edu/~routerproxy/startap/ (we won't use that data here today)

I2 IPv4 Gigapop (and I2 direct connector) attachment MTU summary...

MTU	*Site count*
9180 (or above)	29 (27.1%)
9000<-->9176	9 (8.41%)
4470	66 (61.7%)
2450	1 (0.93%)
1500	2 (1.86%)

	107

What that summary tells us...

- Clearly, at least as of 1/29/2003, many Gigapops (and Abilene direct connectors) are NOT able to support true 9180 byte jumbo frames for their users.
- HOWEVER, all but a couple of Gigapops/Abilene direct connectors DO connect to I2 at some MTU larger than 1500, so MTU issues at the Gigapop/ connector router or ATM switch are not enough to explain "no >1500 MTU traffic."

Ye Olde Opaque Gigapop/Connector

- An old problem: while we can look at each I2 Gigapop/direct connector's interface MTU, we really don't know much about what sits behind that router interface or ATM interface (e.g., in most cases, internal architectures are somewhat opaque).
- For example, the I2 participant-facing-side of a gigapop router might connect to a L2 ethernet switch using a 1500 byte MTU, death for any jumbo frame initiative.

Probing for Gigapop MTUs

- While you can find traceroute gateways at some Internet2 schools, none of those gateways allow you to launch arbitrary size ping packets with the don't fragment bit set.
- The Cisco CLI extended ping and extended traceroute commands offer the functionality we want, but that command is only available to users with EXEC privileges on the router of interest.

However, if the path from an Abilene host is jumbo clean...

- Some Unix and W2K ping commands allow the user to specify both a payload length and to set don't fragment, e.g.:

```
% ping -M do -s 1472 foo.bar.edu (Linux)
c:\ ping -f -n 1 -l 1472 foo.bar.edu (W2K)
```

  If your path into Abilene is jumbo clean, this allow you to do quite a bit of detective work, teasing out the MTU's of remote network devices on paths of interest.
- Tracepath is also a very convenient tool for this

But I2 paths aren't necessarily symmetric

- I should mention that I2 paths are often asymmetric for a variety of reasons relating to costs, traffic capacity on circuits, active BGP routing management, politics, chance, etc. This problem is only becoming more common as institutions work to build out more sophisticated multihomed networks. [see Hank Nussbacher's "Asymmetry of Internet2" at http://www.internet-2.org.il/i2-asymmetry/sld001.htm]

Why asymmetry can matter for jumbo frames

- Asymmetric routing maters for those interested in jumbo frames because even if you have a jumbo-clean path in one direction, reciprocal traffic flowing in the opposite direction may flow via a totally different set of devices, and those devices may (or may NOT) support jumbo frames.

An example of I2 asymmetry:

- traceroute to www.washington.edu from UO

1 ge-4-2.uonet2-gw.uoregon.edu (128.223.142.3) 0.607 ms
2 ge-0-0-0.0.uonet8-gw.uoregon.edu (128.223.2.8) 0.566 ms
3 ge-0-0.core1.eug.oregon-gigapop.net (198.32.163.149) 0.435 ms
4 eug-snva.oregon-gigapop.net (198.32.163.10) 17.168 ms
5 snva-snvang.abilene.ucaid.edu (198.32.11.122) 13.046 ms
6 sttl-snva.abilene.ucaid.edu (198.32.8.9) 31.786 ms
7 sttl-sttlng.abilene.ucaid.edu (198.32.11.125) 31.151 ms
8 hnsp1-wes-so-5-0-0-0.pnw-gigapop.net (198.48.91.77) 31.230 ms
9 uwbr1-GE3-0.cac.washington.edu (198.107.151.51) 21.078 ms
10 dirtdevil-V24.cac.washington.edu (140.142.154.15) 19.722 ms
11 www4.cac.washington.edu (140.142.15.233) 19.151 ms

- traceroute to www.uoregon.edu from UW

1 astrovac-V11.cac.washington.edu (140.142.15.161) 1 ms
2 uwbr1-GE2-1.cac.washington.edu (140.142.154.23) 0 ms
3 core1-wes-ge-1-0-0-0.pnw-gigapop.net (198.107.151.119) 1 ms
4 core1-pdx-so-0-0-0-0.pnw-gigapop.net (198.107.144.18) 5 ms
5 prs1-pdx-FE2-0.pnw.gigapop.net (198.107.144.78) 4 ms
6 198.107.144.90 (198.107.144.90) 11 ms
7 ptck-core2-gw.nero.net (207.98.64.138) 4 ms
8 eugn-core2-gw.nero.net (207.98.64.1) 10 ms
9 eugn-car1-gw.nero.net (207.98.64.165) 7 ms
10 uo1-gw.nero.net (207.98.64.34) 21 ms
11 ge-1-1.uonet2-gw.uoregon.edu (128.223.2.2) 21 ms
12 darkwing.uoregon.edu (128.223.142.13) 20 ms

Paths aren't necessarily stable, nor is "I1" jumbo clean...

- Even if we get a clean jumbo capable path today, there is no guarantee that that path won't shift to a new (non-jumbo-clean) path on a temporary or permanent basis tomorrow… or even from I2 to "I1."
- The availability of 9180 MTU paths in the commodity Internet (e.g., other than over Abilene) is an open question; no identified commodity ISP at this time offers jumbo clean transit.

Action Item?

- Notwithstanding all that, if I may slip into non-directive Minnesotan speak for a sec,

  “Ya know, some guys might think that it would be a good thing if Gigapops and direct connectors tried to pass jumbo frames cleanly, if folks got a chance to look at that sometime and wanted to play around with that a little -- but it could be worse, can’t complain.”

Section 5. Jumbo Frames at the Abilene Participant or Campus Level

Let's Assume The Gigapops Are Okay

- In order to move this along, and having beaten on the Gigapop operators enough, let's pretend that the Gigapops are all set with respect to jumbo frames, and move on down to the campus/Internet2 participant level. [Getting a path jumbo clean is similar to performance tuning a host in that as you remove one bottleneck, another one will often pop up.]

Campus jumbo frame issues...

- When it comes to campus jumbo frame "roadblocks," the problems most likely to arise are one (or all) of the following:
  1) non-jumbo capable router interfaces
  2) non-jumbo-capable gig switches in the campus core or at the subnet level
  3) dominance of 100Mbps/10Mbps ethernet and lackof MTU concurrence on a subnet
  4) reluctance toward making major changes throughout the campus just to facilitate a a non-essential specialized technology

1) Non-jumbo capable router interfaces

- When you try to turn up jumbo frames on a interface of one of your routers, you may be dismayed to find out that some of those interfaces simply won't support 9K frames.

Examples of MTU-limited router interfaces

- Cisco 3GE for the GSR only supports frames up to 2450 bytes (http://www.cisco.com/warp/public/cc/pd/rt/12000/prodlit/thpge_ds.htm)
- Cisco PA-GE (for the 7100 and 7200VXR) only supports frames up to 4476 bytes (http://www.cisco.com/univercd/cc/td/doc/product/core/7200vx/portadpt/ether_pa/pa_ge/2696.pdf)

Examples of MTU-limited router interfaces (cont.)

- Cisco GEIP (e.g., for Cisco 7500s) support MTUs up to 4470 (http://www.cisco.com/univercd/cc/td/doc/product/software/ios111/cc111/geip.htm); the GEIP+, 4476 (http://www.cisco.com/en/US/products/hw/routers/ps359/products_module_installation_guide_chapter09186a008007e5c1.html -- you juts gotta love those Cisco URLs (and small MTUs))

So how do I “fix” those non-jumbo capable interfaces?

- “Fixing” MTU-impaired router interfaces usually is an exercise in purchasing replacement equipment.
- Ironic note: experimental projects (such as trying to do jumbo frames) are often deployed on otherwise unneeded “surplus” legacy equipment, which is often precisely the sort of equipment least likely to have jumbo capable interfaces!

2) Non-Jumbo-Capable Core and Subnet Ethernet Switches

- There are many very popular ethernet switches on the market that do NOT support jumbo frames.
- **Non-jumbo-capable ethernet switches in the campus core and at the subnet level are probably the single biggest reason why it is rare to find campus path MTUs greater than 1500 bytes.**
- Replacements can be purchased, but they usually aren't cheap.

Relative costs (jumbo- and non-jumbo capable) of switches

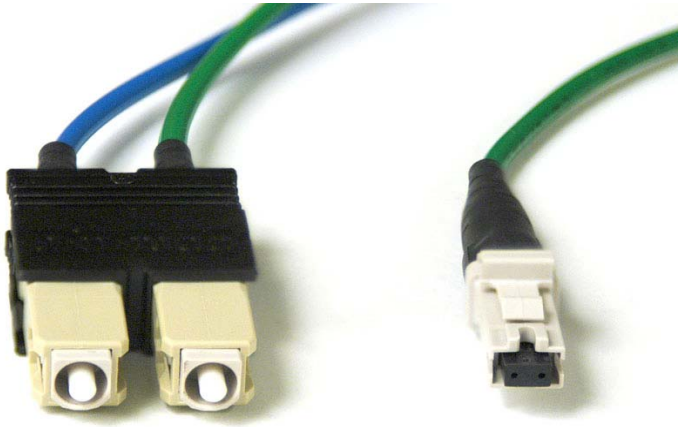
- HP Procurve 4000M switches, NOT jumbo frame capable, are less than $1500 for the chassis (complete with 40 10/100 ports you can use to fill out a 2nd 4000M somewhere else). 1xGig SX modules go for <$350; ditto 100/1000 baseTX gig copper modules.
- If all you need is a small gig copper switch, you can even get an 8 port Netgear GS508T for less than $550!

And in comparison...

- The best/least expensive jumbo-capable replacement we could find for a 3Com 9300 (e.g., providing us with a dozen SX ports), was an Extreme Summit 5i, at nearly $10K:

And that doesn't include replacement fiber jumpers

- Add to that the cost of purchasing a stock of MTRJ-to-SC fiber jumpers (all our NICs are SC, as were the ports on the old 9300, while the Extreme used MTRJ connectors).



Want more info on some jumbo capable gigabit switches?

- -- Cisco Cat 5K or 6x00 series (www.cisco.com/warp/public/473/148.pdf)
  -- Extreme Summit 5i (www.extremenetworks.com/libraries/prodpdfs/products/summit5i.asp)
  -- Foundry FastIron 400 (www.foundrynet.com/products/123wiringcloset/fastiron/FIx00.html)
  -- Nortel Alteon 180 (www.nortelnetworks.com/products/01/alteon/webswitch/prodlit.html)

You'll probably need more than just one jumbo-capable switch

- Even you get a jumbo capable switch installed for a given subnet, you still need to insure that ALL upstream ethernet switches, including any switches in your campus core, are ALSO jumbo frame capable [unless you plan to do something really ugly like taking traffic directly from a jumbo capable subnet switch directly to your campus border router, bypassing your normal campus network infrastructure entirely. Ugh.]

Purchase timing

- As you look at potentially replacing an existing campus core gig switch with one that is jumbo capable, timing may be an issue. That is, there may be reluctance to buy replacement core gigabit switches right now when 10gig switches are almost (but not quite) ready for prime time. See, e.g., www.nwfusion.com/news/2002/120210gig.html
- This is also a period when budgets for capital equipment purchases may be tight...

3) 100Mbps, 10Mbps ethernet and subnet MTUs

- A more subtle fact impacting jumbo frame deployment at the campus level is that jumbo frames are rarely supported on 10 or 100Mbps ethernet links. This is relevant because at most campuses:
  -- relatively few hosts are gigabit attached
  -- gigabit hosts often live on the same subnet as 10Mbps or 100Mbps hosts
  -- things get tricky if all hosts on a subnet fail to agree on a common MTU

Cleaning up the neighborhood

- Faced with that reality, the most common option is probably to create a separate gigabit-only jumbo frame subnet, which usually means somebody's going to have to renumber unless you've been very lucky/ systematic in assigning IP addresses.
- You may also need additional gigabit router interfaces (assuming you want to keep the legacy 10/100 hosts downstream of a gigabit uplink).

4) “If it isn’t broken…”

- The final potential killer roadblock at the campus level is reluctance on the part of many network engineers to screw around with a stable production network just so a few systems can begin [trying] to use a perceived “non-essential” feature.
- You should also be prepared to be asked, “Well, who else on I2 that you work with is using jumbo frames at this point, anyhow?” [the classic chicken-and-egg question that also dogged IP multicast and IPv6 rollout]

Section 6. Empirical Test of Internet2 Participant MTUs

Internet2 Participant MTUs

- All that discussion aside, "*How many I2 participants appear to have routine >1500 MTU connectivity, for example to their primary web server www.<whatever>.edu?"*
- Courtesy of Bill Owens and Nysernet, tests were done from ATM-connected Debian box [with at least a 4470 byte-clean path to Abilene] to over 211 Internet2 participant main web sites.

On the choice of primary web servers as an MTU test target

- We know that some may question our choice of the institution's primary web server as our MTU test target -- such a box may not have any need for jumbo frames, for example. True. However, it does provide a convenient, centrally maintained, universally available "important" host to test. (We'd gladly test other better-connected hosts if we knew they existed!)

It's a 1500 byte MTU world out there...

- The most noteworthy thing we found is that none of the tested hosts could accept >1500 byte frames.
- **Copies of the MTU tests for each I2 participant domain are available at darkwing.uoregon.edu/~joe/tracepath/**
- In some cases, because an upstream gigapop or connector was already clamped at 1500, we really can't tell if that participant would otherwise be able to do >1500 byte frames.

Typical tracepath test

- **tracepath www.indiana.edu**

```
1?: [LOCALHOST] pmtu 9180
1:  199.109.33.1 (199.109.33.1) 2.530ms
2:  199.109.33.1 (199.109.33.1) asymm  1  2.455ms
    pmtu 4470
3:  roc-m10-nyc-m20.nysernet.net (199.109.5.53)
    asymm 4 23.164ms
4:  buf-m20-roc-m10.nysernet.net (199.109.6.2) asymm
    5 24.608ms
5:  abilene-chin-buf-m20.nysernet.net (199.109.2.2)
    asymm 6 36.977ms
6:  iplsng-chinng.abilene.ucaid.edu (198.32.8.77)
    asymm  7  40.751ms
7:  ul-abilene.indiana.gigapop.net (192.12.206.250)
    asymm  8  40.998ms
8:  ul-abilene.indiana.gigapop.net (192.12.206.250)
    40.754ms pmtu 1500
9:  192.12.206.73 (192.12.206.73) asymm 10 40.895ms
10:  wcc6-gw.ucs.indiana.edu (129.79.8.6) 58.161ms
11:  lux.ucs.indiana.edu (129.79.78.4) 41.580ms reached
     Resume: pmtu 1500 hops 11 back 11
```

Unusual cases

- In doing our tests, we ran into some unusual cases (e.g., commodity routes pref'd over I2 routes, complete filtering of ICMP, etc.)
- If tracepath didn't complete, or if tracepath returned unusual results, we manually probed further using traceroute and ping. In most cases, we were able to verify that the site would accept 1500 byte packets with don't fragment set, but would reject 1501 byte packets with don't fragment set.

Location of the bottlenecks

- While it is sometimes possible to determine the location of the bottle neck based on tracepath output (at the participant/campus level, or at the gigapop level, for example), in many cases a lack of rDNS data for hosts in the path can make this tricky to do right.
- Rather than provide a summary of gigapop/host bottlenecks, we encourage you to look at the data for individual sites that are relevant to your own collaborations.

Noted in passing: filtering ICMP

- In doing our test, we noticed that some folks are “protecting” their users from ICMP (RFC792) messages by filtering (or rate limiting) ICMP echo/echo reply, ICMP destination unreachable, ICMP time exceeded, etc.
- Yes, I know that SANs and others have encouraged sites to adopt a restrictive policy with respect to ICMP traffic, but if you block ICMP, you WILL break stuff.

Filtering ICMP and PMTUD

- "Path MTU Discovery and Filtering ICMP"
  http://alive.znep.com/~marcs/mtu/
  does an excellent job of laying out one issue that broadly filtering ICMP can cause.

  We will talk further about PMTUD in the next section of this talk.

7. Jumbo Frames at The Host Level

Not all network paths are equal

- While it would be nice if all (or even many) network paths on Abilene were jumbo frame capable, the reality is that many will not be for the foreseeable future.
- However, let's assume that because of concerted efforts, some interesting paths will become jumbo capable end-to-end.
- How then, if we are to do jumbo frames, how does a host determine what MTU should be used with which path?

Which MTU to use...

- Systems can simply send frames no larger than the smallest maximum size allowed per RFC879 (e.g., 576 bytes). [Before you laugh, this is what Windows 2000 does if you disable PMTU discovery!] But this doesn't help us do jumbo frames….
- A maximum segment size can be specified at the time a connection is setup (RFC793). [Doesn't really help with jumbo frames]
- Systems can (try to) do RFC1191 PMTUD.

RFC1191 Path MTU discovery

- ‘The basic idea is that a source host initially assumes that the PMTU of a path is the (known) MTU of its first hop, and sends all datagrams on that path with the DF bit set. If any of the datagrams are too large to be forwarded without fragmentation by some router along the path, that router will discard them and return ICMP Destination Unreachable messages with a code meaning "fragmentation needed and DF set" [7]. Upon receipt of such a message (henceforth called a "Datagram Too Big" message), the source host reduces its assumed PMTU for the path. The PMTU discovery process ends when the host's estimate of the PMTU is low enough that its datagrams can be delivered without fragmentation.”

RFC1191, November 1990

PMTUD-related blackholes

- PMTUD doesn't always work. For instance, if PMTUD is attempted but a site filters the destination unreachable messages used by PMTUD, a black hole condition may arise.
- PMTUD black hole detection may ameliorate this condition (but in doing so we act to suppress a symptom rather than cure the underlying disease condition).

Problems with PMTUD

- A variety of problems with Path MTU discovery are discussed in RFC2923, “TCP Problems with Path MTU Discovery.”
- These problems are not just a hypothetical or theoretical concern; see, for example:

  http://www.netheaven.com/pmtulist.html
  http://home.earthlink.net/~jaymzh666/mss/

PMTUD security issues

- Moreover (as was mentioned in RFC1191 itself, it was clearly known that the PMTUD mechanism has a fundamental vulnerability to DOS attacks due to the unauthenticated nature of ICMP messages. [e.g., bad guys could force all traffic to fragment using a tiny MTU (e.g., 68 bytes), or force your MTU very high to try to create a blackhole]
- draft-etienne-secure-pmtud-00.txt (expired May 2, 2002)?

Host gigabit ethernet jumbo frame hardware/OS issues

- Besides generic issues relating to PMTU discovery, a fundamental question is “Do popular host hardware platforms and operating systems support jumbo frames?”

Jumbo frames under Solaris

- Sun gigabit adapters often try to make a virtue out of supporting “Standard ethernet frame size (1518 bytes)” [Sun Gigabit Ethernet/P 2.0 Adapter] or say something like “The Sun GigaSwift Ethernet adapter is interoperable with existing Ethernet equipment assuming standard Ethernet minimum and maximum frame size…”
- See: www.sun.com/products-n-solutions/hardware/docs/Network_Connectivity/SunGigabit_Ethernet/

Aftermarket jumbo-capable gigabit cards for Solaris

- www.syskonnect.com/syskonnect/products/sk-98xx.htm (for driver info see www.syskonnect.com/ syskonnect/support/driver/d0102_driver.html)
- www.antares.com/ethernet/ethernet.htm

DEC/Compaq/HP Alphaservers and OpenVMS

- http://h18000.www1.hp.com/products/quickspecs/10479_na/10479_na.HTML says “when connected point-to-point with another cooperating NIC or switch, the PCI-to-Gigabit Ethernet NICs can transfer Jumbo Frames of up to 9,000 bytes in length...”
- As always, hardware, firmware and OS restrictions may apply

Linux and Windows 2000

- Linux and W2K supports jumbos nicely
- Many vendors make jumbo capable NICs with Linux and Windows 2000 driver support including Syskonnect, Intel, 3Com, Netgear and others.
- http://www.syskonnect.com/syskonnect/news/testresults/rep1.pdf

Continuing the discussion...

- If you are interested in working on this topic further, a mailing list is available; to subscribe, send email to

  majordomo@lists.uoregon.edu

  with a message body reading

  subscribe jumbo-clean

Special thanks to...

- -- Bill Owens and Nysernet for their support of the tracepath measurements
  -- Dave Meyer, Dale Smith and Jose Dominguez here at the UO CC for all their patience/help with my many odd projects.
  -- Joanne Hugi, my boss and the Associate Vice President for Information Services at UO, for her encouragement and for her ongoing support of the Oregon Gigapop, Oregon's connection to Internet2.

Questions?