How Much Headroom Do You Need?

Bandwidth and High Performance Networks

Joe St Sauver, Ph.D.
(joe@uoregon.edu or joe@internet2.edu)
University of Oregon and Internet2

Joint Techs, Lincoln Nebraska
July 21st-23rd, 2008

http://www.uoregon.edu/~joe/how-much-headroom

Disclaimer: All opinions expressed are solely those of the author and do not necessarily represent the opinions of any other entity.

1. Introduction

The Origin Of This Talk

- On June 20th, 2008 Russ Hobby wrote to the Internet2 Network Technical Advisory Committee (NTAC) mailing list, commenting:

> *One of the main things that has set R&E network apart from the commodity Internet has been the loading of the network. R&E networks have generally been lightly loaded with a lot of bandwidth headroom so that applications would never see bandwidth as a constraint. There has always been resources for new applications to burst above the the average traffic. Usually we would move to the next new, faster technology before network contention became an issue.*
>
> *During these times IP traffic within the R&E community was generally the only thing that was on the links of the network. Things are changing. Other services, such as commercial*

The Origin Of This Talk (continued)

peering and dynamic circuits are starting to share those network links. This means that bandwidth available for R&E IP traffic can now vary in continuous amounts rather then full circuit bandwidth increments. If we are to maintain the advantage of R&E networks having the bandwidth headroom, management of the bandwidth for the IP network will have to be more closely managed than in the past.

However we have never really defined what "lightly loaded" was or what "sufficient bandwidth overhead" is. If we are to maintain superior IP network services in light of a more closely managed overall bandwidth, perhaps we should define a convention for the R&E community.

What do people think? Is it an issue?

I Thought Russ Asked An Excellent Question

- And apparently so did a number of other folks on the NTAC (https://wiki.internet2.edu/confluence/display/ntac/Home) mailing list, because a lively discussion ensued, one which I enjoyed participating in.

- Since I'd spoken up, had previously talked about capacity-planning (see "Capacity Planning and System and Network Security," http://www.uoregon.edu/~joe/i2-cap-plan/), and because this turned out to be such an interesting topic, Marla from MCNC asked me to pull together this talk for Joint Techs.

- I was happy to do so, because I believe that answering Russ's question correctly is **key** to **fully realizing** the value of each site's Internet2 connectivity, while also insuring that that connectivity remains capable of **reliably delivering high performance**.

Picking a Correct Utilization Target Is Key

- While you want to take full advantage of your Internet2 connectivity, the goal should NOT be to see your connection pegged or "flat topped".

- If you ARE seeing your connections to Internet2 in that state, that's a sign that you may be under-provisioned and you may want to consider adding additional capacity (but I don't think anyone's currently flattopping).

- On the other hand, it does no good to have unduly restrictive usage practices which result in high performance connections languishing substantially underutilized. You can't "save" any bandwidth you're not currently using, so if you have capacity, and you have a reasonable use for that capacity, well, you might as well use it -- **as long as you don't end up congesting during peak periods.**[6]

2. Avoiding Congestion During Peak Usage Periods

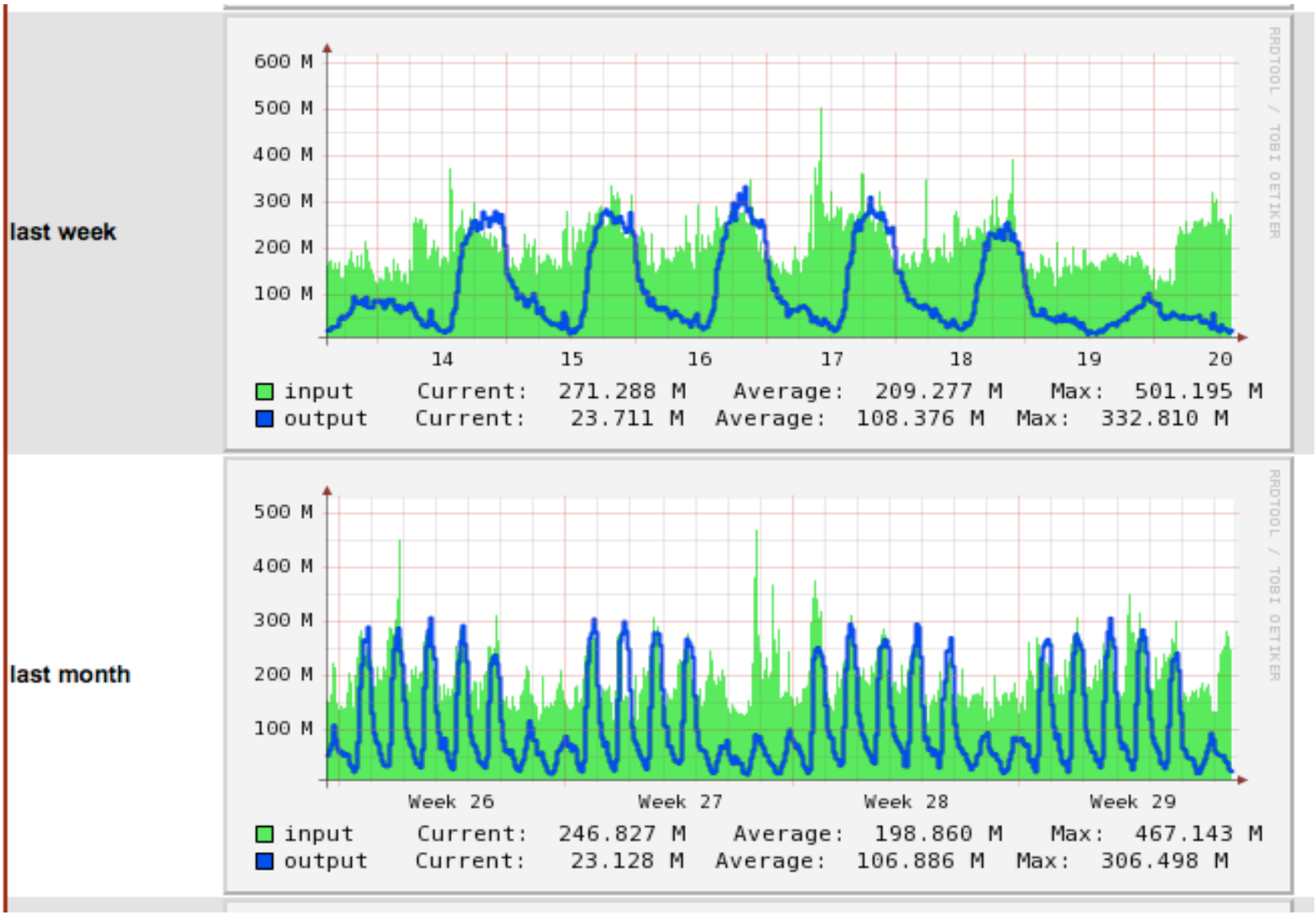
Way Back When, Some People Thought High Performance Networks Required QoS

- Everybody remember the **Qbone**? At that time, we "knew" that some applications (such as **voice over IP**, or **network video**) might be **jitter sensitive**, and there was a substantive desire to insure that those at-risk applications were sheltered from the impact of other types of traffic via quality of service (QoS).
- We also re-learned that **TCP/IP would back off, hard,** in the face of congestion, and that that could dramatically reduce **throughput**
- But we had lots more to learn, too. For example, we learned that premium **QoS was hard to deploy**. Fortunately, we also learned we didn't actually need it as long as the network doesn't congest even during **peaking periods.**

Peaking Periods

- It would be a wonderful world if network traffic were constant and invariant. Given perfectly **level demand**, pretty much anyone could determine the amount of capacity they'd need, and cost effectively procure just that amount.
- But network loads aren't constant -- they tend to be irregular, and have **peaks and troughs**. We can't just provision capacity for the **average** of those peaks and troughs, because if we did that, we'd have more capacity than we need for the "trough" times, but less than we'd need for "peak" times. **We want to make sure we have enough capacity to accommodate [mostlall] of those peaks.**
- The height of the peaks we see will also vary from day to day, or season to season. For example, the height of workday peak loads may be higher than the peak seen on weekends, and peak loads may be higher during the school year than during vacation periods. Regretably, we can't dial connections up and down to correspond to those time-varying demand profiles.

Time Varying Loads: When's The Weekend?



See http://dc-snmp.wcc.grnoc.iu.edu/i2net/

"Headroom Requirements" and Statistics

- Given all that, if you look closely, the "how much headroom do I need to avoid congestion" question is really a **statistical** one. After all, network headroom is effectively just "**buffer capacity**" designed to accommodate peaking traffic flows.
- In most cases, assuming you're economically constrained, you want "enough" capacity to handle the demands you see some "large fraction of the time," accepting that in some small residual fraction of cases you WILL congest (with all the negative things that may imply for a typical mix of TCP and UDP traffic).
- How big is a "large fraction of the time?" Well, you get to pick, either explicitly or implicitly. A "large fraction of the time" might be 80% of the time, or 95% of the time, or 99.999% of the time, but there will **ALWAYS** be **some** probability that you'll run out of capacity at some time unless your wide area link has more capacity than some downstream chokepoint.

The Implications of Your Choice…

- If you buy too little capacity, performance during peaking periods will be poor; but if you buy more capacity than you need, that excess capacity will do nothing for you, except cost you money which you could have spent on something else. So, if "money's no object," buy "lots." [wouldn't that be nice?]

- If "money **is** tight," or you can tolerate at least some possibility of congestion, or you can explicitly shed load if congestion appears imminent, buy "less." Of course, these days money ***is*** tight for many of us, so there may be connectors who have to make do with the amount of capacity they can afford, rather than the amount of capacity they might need or want.

- Other factors may also impact your provisioning decisions.

R&E *AND* Commodity Internet Links Probably Need A "Balanced Build Out"

- **You can't field services or applications which go fast only to Internet2** (while going slower everywhere else) because most applications generally aren't "network aware."
- As a result, you need funding to support BOTH research and education network links AND commodity Internet links. Funding expended on commodity Internet capacity can obviously impact the amount of money available for R&E network capacity.
- **The Internet2's Commercial Peering Service provides one way of leveraging excess capacity you might have on a comparatively lightly loaded R&E connection, allowing you to use some of that excess capacity to accommodate commodity Internet commercial traffic at no incremental cost.**
- But why would sites have excess capacity on some of their R&E links?

Fixed and Limited Number of Capacity Options

- You might want to purchase 783.74Mbps worth of R&E capacity, or 1.8Gbps worth of R&E capacity, but network capacity options can't be "dialed in" to those sort of arbitrary values. Your choices for non-legacy Internet2 IP (packet only) connectivity are only:

  -- 1 Gbps: $250,000/year
  -- 2.5 Gbps $340,000/year
  -- 10 Gbps $480,000/year

- Because of the structure of that pricing, it doesn't make sense to purchase multiple 1Gbps or multiple 2.5Gbps (e.g., the cost of 2x1Gbps to get 2Gbps worth of capacity is greater than the cost of just buying one 2.5Gbps, and similarly the cost of 2x2.5Gbps to get 5Gbps is greater than the cost of just buying one 10Gbps link)
- The result of that reality is that you may only need 3Gbps, but you may find that it is more cost effective to buy one 10Gbps link, and that might leave you with 7Gbps worth of "excess" capacity.

Headroom and "Low Utilization"

- Note the tension that exists -- headroom is good when it comes to providing buffer capacity to handle peak loads, but as the amount of headroom increases, average utilization will tend to go down.
- If you are budget constrained, financial types might look at "low" average utilization levels and say, "Eh -- look at how low our utilization is! We're buying 10Gbps worth of capacity, but on average we're only using 1 Gbps worth of that capacity. We're buying 'too much' capacity!"
- As we've just discussed, a number of factors can contribute to what looks like "excess" capacity or "low utilization" levels, while in reality the "low utilization" may just be a side effect of cost effectively getting the capacity that's actually required, or provisioning capacity required to accommodate peaking loads.
- It is going to be important for technical staff to make sure that non-technical staff understand and correctly interpret this issue.

Symmetric Inbound and Outbound Capacity

- Higher education traffic profiles also often aren't perfectly balanced in- and out-bound. At most schools, **traffic to the school from the Internet dominates traffic from the school to the Internet.** Put another way, most schools are so-called "eyeball" networks, not "content" networks.
- It would be great if inbound and outbound capacity could be independently provisioned, but that's generally not an option for high capacity research and education links -- the link you buy will have one (symmetric) level of in- and out-bound capacity.
- This is yet another factor that contributes to lower average utilization, particularly lower average **outbound** utilization, than one might expect.
- However, that imbalance also provides a real **opportunity**: you may be able to donate hosting to software mirrors, or other outbound traffic uses, with virtually no incremental cost.

3. Atypical Traffic Sources

But The Problem Isn't Predictable Loads…

- The things that often ultimately drives headroom levels is the desire to be ready to accommodate and service those **UN-predictable** loads.

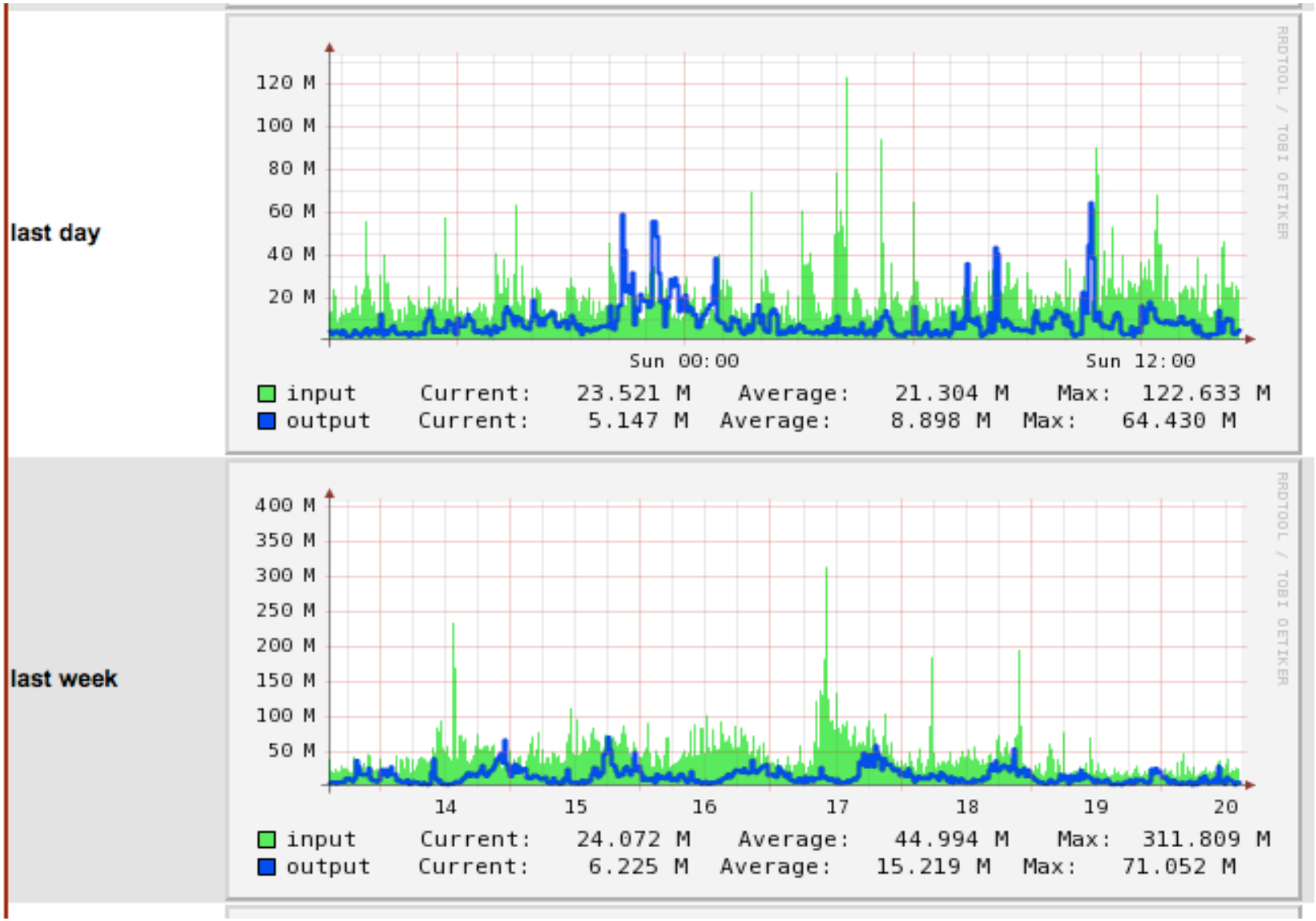
- These UN-predictable load sources may include things like:

  -- high performance point sources
  -- individual users acting *en masse*
  -- failover of redundant links
  -- DDoS traffic

High Performance Point Sources

- One factor that really complicates picking the right capacity for a link, and which we all worry about, is what might be called "high performance point sources," or **individual systems** which can, in and of themselves, have a material impact on aggregate traffic levels. Servers of that sort may have gig or even ten gig connectivity, as well as the CPU, memory and I/O performance needed to take full advantage of their high speed interfaces
- When individual systems can source gigabit (or multi-gigabit!) traffic levels, your ability to count on the aggregation of thousands or tens of thousands of comparatively small bandwidth users to "balance each other out" disappears.
- **Expect to see "spike-ier," more-random-looking, harder-to-plan-for demand patterns**.
- Do your best to "know your load" (anytime someone gets a gig or ten gig connection on campus, chat with them a little, eh?)

Spikey, Hard to Predict Loads



See http://dc-snmp.wcc.grnoc.iu.edu/i2net/

Individual Users, Acting *En Masse*

- Individual users acting suddenly acting *en masse* as part of online crowds can also generate atypical loads.

- A hypothetical example: everyone watching (in real time) World Cup soccer video coverage of a particularly dramatic game

- FWIW: We **still** have a lot of work to do when it comes to getting most users tuned for optimal end-to-end performance (see chart on the next slide). The fact that most users **aren't** "individually dangerous" high performance point sources is actually sort of a bad sign, I think.

We Still Have Our Work Cut Out For Us

Table 1. Selected Points from Distribution Graphs (Bulk TCPs)

Percentile	Throughput (b/s)	Durations (s)	Size (octets)
1	1.391M	1	10.05M
5	1.491M	7	10.50M
10	1.621M	14	11.08M
50	3.410M	58	18.90M
90	16.35M	59	56.33M
95	28.85M	59	78.60M
99	81.03M	59	183.1M
99.9	169.4M	59	381.1M
99.99	931.9M	115	1.232G
99.999	1.026G	131	3.112G
100	5.800G	132	6.688G

See: http://netflow.internet2.edu/weekly/20080707/

Another Worry: Failover of Redundant Links

- Network elements can fail or be damaged from time to time. For example, buried fiber run may be accidentally cut by a backhoe, or a transceiver may fail
- When an incident of that sort happens, if a network has been built with redundant paths, the traffic that was previously flowing over the now-damaged link will fail onto the redundant link.
- But obviously, in order for that single remaining connection to be large enough to be able to carry all that traffic without experiencing congestion, **each link** of the redundant pair of connections must be large enough to **carry ALL** the load -- but that means that during normal times, **each connection should routinely operate at "half" (or less) of its potential capacity**.
- This is one reason why many carriers routinely insure that they never run at more than 50% utilization.

DDoS Traffic

- Attack traffic, such as distributed denial of service (DDoS) traffic, also complicates headroom planning, particularly since some miscreants may be willing and able to source as much traffic as it takes to take ANY site of their choice off the air.

- It is hard to know how to plan the amount of capacity you might want to thwart a DDoS, except to say that "more is generally better" when it comes to weathering traffic-based network attacks.

4. Conclusions

So How Much Headroom *Should* You Have?

- You tell me! :-) Some (dramatically varying) answers might include:

  -- "Enough to prevent congestion during peaking periods" (cue statistical models here)

  -- "Keep utilization below 50%" (to protect against shifting loads from failing links)

  -- "Given the uncertainties associated with atypical loads sources, buy as much capacity as you can afford"

  -- "Don't worry about utilization/headroom unless/until you actually observe performance problems, or receive complaints."

Other Recommendations to Consider

- T**ry pushing your Internet2 connection harder/running it hotter than you currently may be**. Many sites are extremely conservative when it comes to how much traffic they carry on their Internet2 links and how much headroom they attempt to reserve, perhaps unnecessarily so. Push the pedal a little harder.
- In particular, **if you have been holding back from trying the Internet2 Commercial Peering Service, you may want to give it another look**. If you're so tight on bandwidth that you really don't think you can -- should you be thinking about an upgrade?
- Recognize that **fear of negatively impacting host throughput** is probably the most common reason why people don't want to run their Internet2 connection overly hot, yet based on empirical metrics, we've still got a *long* way to go. For tips on going faster see PSC's excellent "Enabling High Performance Data Transfers," http://www.psc.edu/networking/projects/tcptune/

Thanks for the Chance to Talk Today

- Are there any questions?