Thinking About Lambda-Based Network Architectures and Your Applications

Internet2 Member Meeting

8:45-10AM, September 20th, 2005

Philadelphia, Pennsylvania

Joe St Sauver, Ph.D. (joe@uoregon.edu)

University of Oregon Computing Center

http://ww.uoregon.edu/~joe/lambdas/

I. Introduction

Good Morning!

- It is a pleasure to be here in Philadelphia today, and to have the opportunity to talk with you a little about lambda based network architectures and your applications.
- This talk was originally scheduled for later in the day, but since that slot would have overlaped with another optical networking talk, we're doing this talk now, instead.
- It's currently 5:45AM in my normal Pacific time zone, so if I look sleepy, please be patient. :-)

The Audience for Today's Talk

- This talk has a fairly strategic focus, and is really meant for those who have been trying to decide how National Lambda Rail (NLR) or similar national scale optical networking initiatives will fit with their institutional and regional networking requirements. That group likely includes:
  - -- institutional executive members
  - -- network leads, and
  - -- application-oriented people.

  I'll try to include a little something for everyone, with some stuff probably too simple, and some too complex.
- Because some may refer to this talk after the fact, and because we also have netcast participants and audience members for whom English may not be their primary language, I've tried to prepare this talk in some detail so as to make it easy for those folks to follow along.

Where I'm "Coming From"

- This talk is not about campus, metro, regional, or international optical networks. Issues of pivotal importance to national optical networks may be completely irrelevant to optical networks at other scales.
- My time horizon is two to three years. Wonderful things may happen farther out, but I'm primarily interested in what's happening in the immediately foreseeable future.
- I'm very concrete and applied: what's the specific real problem that we've identified which we're trying to solve?
- I believe in eating the pork chop that's already on your plate before you go back for 4 more from the buffet:
  If someone says they need OC192 (10Gbps) service, have they already demonstrated the ability to effectively load an OC48 (2.4Gbps)? If they already have an OC48 but it is largely idle, why not see what they can do with that, first?

Where I'm "Coming From" (continued)

- Ongoing projects are more interesting to me than brief one-off special projects or demonstrations. If you're going to work hard, I believe it makes sense to spend that effort building something strategic, something that will last. Create the Panama Canal, not an ice sculpture.
- Make decisions about projects with a twenty year duration carefully; you'll need to feed that baby until (s)he's an adult.
- Solutions must scale to handle anticipated target audiences (and more). Pay attention to step functions.
- Assume that budgets are limited, and money does matter. What's the business case?
- I like the simplest solution that will work.
- I tend to resist artificial urgency and ignore peer pressure.

My perspective may or may not be consistent with yours…

Speaking of Perspectives: A Disclaimer

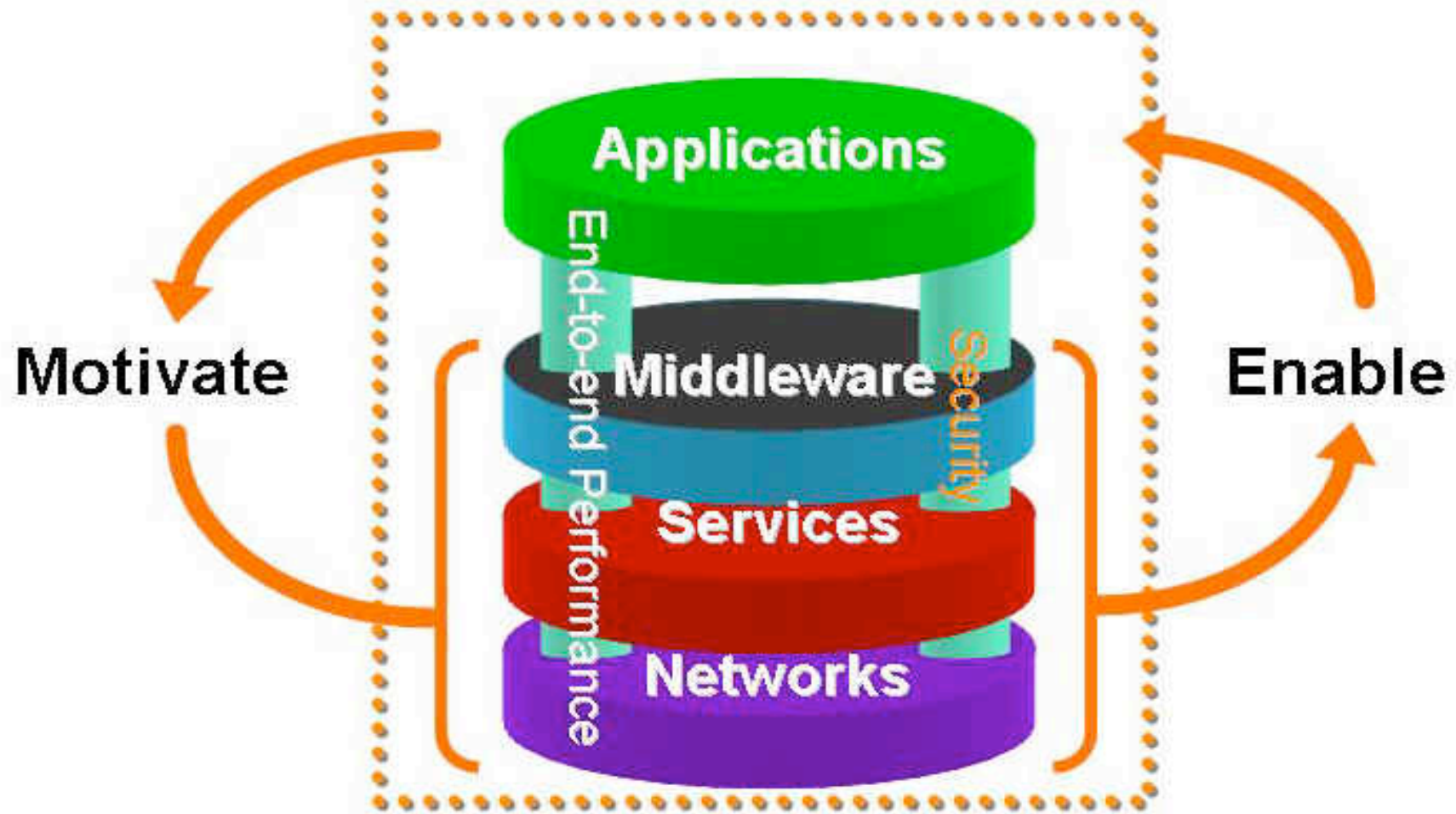
- The University of Oregon is not currently a member of National Lambda Rail, so my perspective is that of a 3rd party/outsider.
- The views expressed in this talk are solely my own, and should NOT be taken as expressing those of Internet2, NLR, the University of Oregon, the Oregon Gigapop or any other entity.
- National scale optical networking is in flux. Even by the time this meeting is over, this talk will be out of date.
- Do not make any decisions based just on what I'll share during this talk; do your own due diligence and make up your own mind when it comes to the issues discussed.

II. Applications and Advanced Networks

Application "Fit" and Advanced Networks

- We believe that if you want to make effective use of advanced networks such as Abilene (or now NLR) you really should spend time thinking about how your prospective applications "fit" with those networks.
- If you don't think about application fit, you may build (or connect to) an absolutely splendid network only to see little (if anything) ever happen over that facility.
- Those who remember the NSF HPC connections program will remember that a key component of applying for funding for a vBNS or Abilene connection was identification of specific applications that would actually use those new connections.
- "Applications should motivate new networks, and networks should enable new applications."

The Application-Driven Network Deployment Process



Source: http://www.internet2.edu/resources/Internet2-Overview-2.ppt at slide 15
Used with permission

My Interest in Networked Applications and Advanced Networks Isn't New

- Back in the Spring of 1999, I spent some time thinking about what sort of applications might work well on Internet2, culminating in a piece I wrote called "Writing Applications for Internet2" (see http://cc.uoregon.edu/cnews/spring1999/writing_i2_applications.html); that article was subsequently adapted for national audiences and redistributed by NLANR/DAST (see http://dast.nlanr.net/Guides/WritingApps/)
- My goal at that time was to make sure that our local users understood the constraints that might impact what they could do with the (then-new) Internet2, and to also help them begin thinking about what applications might fit, and work well, when run over our new connectivity.

The Constraints We Foresaw in 1999 For Internet2

The constraints mentioned in that article were fairly simple:

- One end of the application needed to be homed at UO, running from our network with access to Internet2
- The other end of the application needed to be at a site that also had live high performance connectivity
- The application should (ideally) have characteristics which would take advantage of Internet2's unique capabilities
- The application should be able to differentiate between high performance connections and commodity Internet connections
- Applications should be ongoing, or time critical
- Applications shouldn't be for commercial purposes, nor should they involve classified data

Based On Those Constraints…

We made some predictions about what might work well:

- "Pull" network applications where you can narrowly focus the networks from which information is being retrieved
- "Push" network applications where you can narrowly focus the networks to which information is being sent
- Prearranged server-talking-to-server applications such as NNTP (USENET News), or World Wide Web cache hierarchies
- Applications using multicast
- Applications used by a relatively small number of technically competent trusted users working with large datasets
- Applications which open many parallel network streams to diverse locations (continued)

Examples of Applications Which We Predicted Would Work Well Over Internet2 (continued)

- Applications where there is a large discrepancy between bandwidth available via commodity network connectivity and bandwidth available via high performance networks (e.g., overseas sites in many regions, provided that the overseas site has access to high-performance network connectivity)

Retrospectively…

- Those predictions weren't too far off the mark, and today we routinely make effective use of our Internet2 connection, loading it to target utilization levels.
- Can we now offer a similar prescription for lambda-based networks, such as NLR?
- A first step is probably to begin with a brief backgrounder on National Lambda Rail, for those who may not be familiar with it.
- Heck, for that matter, what's a lambda, and how is it different from what we're used to on Internet2?

III. NLR Backgrounder

Lambdas Defined

- A lambda is a specific wavelength, or "color of light," in a wave division multiplexing (WDM) system, running over fiber optic links. Think of this as being kin to using a prism to break the white light that might normally flow over fiberinto different colors, each of which can be used to carry information independently of what's going on "in" the other colors.
- By using WDM technology, the amount of traffic that a fiber optic link can carry is multiplied, perhaps to forty times its original capacity. Conceptually, where once a piece of fiber had room for only one channel of network traffic, you can now think of that same piece of fiber as supporting forty parallel independent channels of information, each on its own "lambda" or color of light, with the net result being that one pair of fiber can suddenly act as if it were forty.

"Why Does WDM Gear Always Generate 40 Waves?"

- Sometimes the question comes up of, "Why does WDM gear always provide 40 wavelengths?" The answer, of course, is that it doesn't.
- You can purchase dense wave division multiplexing (DWDM) gear that can yield 80 or 160 or even 320 wavelengths from a piece of fiber, or coarse wave division multiplexing (CWDM) gear that only gives you a 8 or even fewer channels.
- The higher density gear, because it allows you to cram more channels onto a piece of fiber and because it is built to tighter tolerances, generally costs more than the coarse, lower channel count, WDM gear.
- The optronics used for NLR, however, does happen to be 40 channel gear.

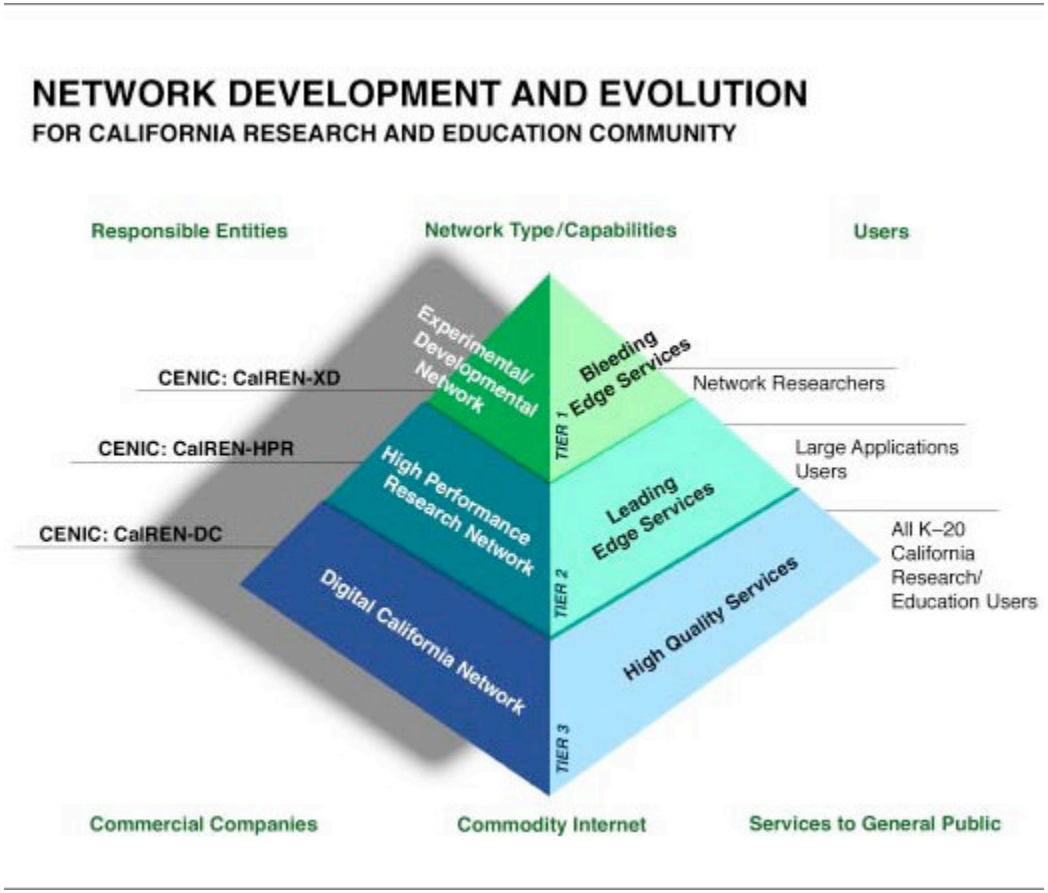
Dedicated Circuits vs. Shared Capacity

- The relative abundance that's associated with WDM makes it possible for us to begin potentially thinking on a national or International scale about dedicated circuits rather than just the shared (or "statistically multiplexed") network capacity that's typical of packet switched networks such as the Internet, or Abilene.
- While it would not make sense for you to set up a lambda just to distribute a web page from someone's web server in New York to a browser in Texas, or to use a lambda to distribute an email message from someone in California to someone in Florida, maybe there will be times when it might make sense to give someone "their own lambda" rather than having them share network capacity with other users. We'll see!
- So how about NLR in particular?

NLR: Born in the Golden State

- Understanding NLR means understanding its roots and original role… CENIC's CALREN, the California research and education network, envisioned three tiers of network service for its constituencies:
  1) Ubiquitous regular Internet service,
  2) High performance production research and education network access (e.g., Internet2/Abilene access), needed by/of interest to a smaller set of users, such as physical scientists working with large datasets, and
  3) Experimental access to a "breakable" cutting-edge network, offering services needed by an even smaller set of extremely advanced users, such as computer scientists doing bleeding edge network research
- It is that third category of network service that has evolved into NLR.

The Three-Tier CENIC CALREN Pyramid



Source: http://www.cenic.org/calren/index.htm
used with permission

Additional Factors Motivating NLR's Emergence

- CANARIE, the Canadian research and education network, became an articulate advocate for the simplicity and cost-effectiveness of customer-owned fiber networks
- Gigapops continued to add customers, including state K12 networks ("SEGP"'s), which incented both upgrades to Abilene connections and the creation of regional optical networks, key components of the current NLR model
- More regional fiber was deployed than was needed; wave division multiplexing caused a national bandwidth surplus
- It became possible to swap excess capacity in one region to get capacity on another route for just the cost of hardware
- By purchasing a few additional fiber links, you could tie all those regional networks into a unified national network
- The Internet financial bubble burst, making the needed residual fiber potentially cheap to acquire

Additional Motivating Factors (cont.)

- The Cisco GSR routers that were originally used on Internet2 got replaced with Juniper T640's; after a bit, Cisco released its new uber-router, the CRS-1, and wanted to re-engage the higher ed R&E networking community
- TheQuilt drove commodity Internet prices down about as low as they could go; the only thing that would be cheaper would be settlement free peering. Settlement free peering required the ability to cost-effectively haul commodity Internet traffic to multiple locations nationally.
- Abilene's conditions of use foreclosed some opportunities; for example, Internet2 was limited in its work with federal mission networks. A new network could be AUP free.
- There was concern over being "locked in" to one network provider (Qwest) for all high performance R&E networking.

Additional Factors (cont. 2)

- The supercomputing community hit a slump and needed to reinvent themselves; grids were born. High performance links were integral to interconnecting those clusters (much as the original vBNS linked traditional supercomputer sites)
- Big science embarked on projects which would generate prodigious amounts of data, data which would need to be wheeled around the country and to/from overseas.
- The engineering folks wanted to do something new and fun
- Some folks who were "late to the party" when Internet2 first got started were highly interested and motivated and determined to not miss out the second time around.
- The U.S. developed a "lambda gap" vis-à-vis Europe
- Abilene lost its "elite" cachet (even K12 had access!) and no longer served a winnowing function for research funding

And So NLR Was Born…

- An optical network that was to be many things to many different constituencies, including coming to have some roles far-removed from it's original Californian pyramid capstone niche.
- For the record, NLR's official goals were/are:
  - Support experimental and production networks
  - Foster networking research
  - Promote next generation applications
  - Facilitate interconnectivity among high performance research and education networks

www.nlr.net/presentations/SC2004_TWW_Slides.htm
(slide 31)

Current NLR Higher Ed Members (*All* Are Consortial)

- CENIC
- CIC
- Cornell (with plans which include other universities in NY state**)
- Duke Univ, representing a coalition of NC universities
- Florida Lambda Rail
- Internet2
- Lonestar Education and Research Network
- Louisiana Board of Regents
- Mid-Atlantic Terascale Partnership and the VA Tech Foundation
- Oklahoma State Board of Regents
- Pittsburgh Supercomputing Center and the Univ of Pittsburgh
- PNW Gigapop
- Southern Light Rail
- UCAR, representing a coalition of universities and government agencies from Colorado, Wyoming, and Utah
- Univ of New Mexico, on behalf of the State of New Mexico

** http://www.news.cornell.edu/Chronicle/04/6.10.04/LambdaRail.html

Today's Interest in NLR

- Those consortia represent a lot of I2 member sites. Interest is NLR today is strong for a variety of reasons, including:
  -- vendors and next generation network evangelists have put great emphasis on the importance and long term potential of lambda-based architectures
  -- a number of consortia have made material multi-year financial commitments to be able to participate in National Lambda Rail (NLR), typically $5 million over five years
  -- a handful of well-funded federal projects running over NLR have received substantial publicity
  -- there have been ongoing discussions concerning the merger of NLR with Internet2, and routine presentations about NLR (and HOPI, and FiberCo) at Internet2 events,
  -- having seen Abilene effectively displace the vBNS, some people may believe that NLR will play a similar role vis-à-vis Abilene, and worry about how that might affect them

NLR On My Mind…

- Regardless of whether or not NLR eventually becomes the "new Abilene" (or at least a substrate upon which Abilene runs), NLR has already come to occupy something of a "displacing" role. By this I mean that while NLR probably did not mean to do so, NLR has come to preoccupy Internet2 "thought space," as well as consuming Internet2 (and member) political, financial, managerial and technical resources that might have been directed otherwise, absent discussions about/work on NLR.
- Assuming NLR is our intended collective top priority, and we're crystal clear about what NLR can (or can't) do for us, that's great. If that's not the case, there should be more dialog.
- Part of that process will be thinking carefully about the new capabilities we want from lambda-based networks.

IV. General Capabilities

NLR: Premium Quality of Service (QoS)?

- For example, is traffic sent cross-country via a dedicated lambda somehow "better" than best-effort traffic sent via an uncongested (but shared) Abilene connection?
  -- Will we see lower latency?
  -- Less jitter?
  -- Less packet loss?
  -- Higher throughput?
  -- Lowered probability of a disruptive network outage?
  Is NLR at root a wide area premium QoS project?
  [Y'all may know how much I "love" QoS…]
- Have we identified current or projected applications that need network characteristics not already available on Abilene? (remember that Abilene is an extremely well engineered and well run network, and sets a technical standard that will be very difficult to materially surpass)

If Not Better-Than-Best-Effort Traffic, Maybe We're Looking for Bandwidth That's Above What Abilene Offers?

- If NLR is not about better-than-best-effort service, then what is it about?
- Is it about providing relief for traffic levels that cannot be accommodated by the already available Abilene connections, including 10GigE/OC192 connections? For example, will the "default" NLR connection not be a single 10Gig pipe, but some aggregate of two, three or more? Are traffic levels necessitating those sort of pipes already discernable, or known to be coming in the foreseeable future? (If so, E2EPI has been a success!)
- Or is it a matter of carrying that sort of bulk traffic over lambda-based connections at a lower cost, or more flexibly, than current Abilene 10 gigabit connections?
- We'll talk about that more later.

Commodity Internet/"Commercial" Traffic?

- There are other possibilities.
- Is an important role for NLR the carrying of traffic that can't be carried over Abilene for policy reasons?
- For example, the Abilene Conditions of Use ("COU") (see http://abilene.internet2.edu/policies/cou.html) states
  > "Abilene generally is not for classified, proprietary, unrelated commercial, recreational, or personal purposes."
- There is at least one existing NLR project that explicitly includes traffic of this type (commodity internet traffic on the Pacific Wave Extensible Peering project).

'Mission Network' Traffic?

- Related to commodity internet/commercial traffic (in terms of having COU-limited access to Abilene) is mission network traffic. [Mission networks are the high-performance networks run by federal agencies in support of their scientific research programs such as the Department of Energy's ESNet, DOD's DREN, NASA's NREN, etc.]
- Mission networks connecting to Abilene do NOT see the full set of routes that regular higher ed connectors get (see http://abilene.internet2.edu/policies/fed.html).
- That restrictive routing policy limits the usefulness of Abilene for mission-network-connected agencies, and may motivate interest by at least some of those agencies in AUP-free alternatives such as NLR.
- **Many** NLR projects involve mission network-related sites.

Lambda-based Networks and Local Policy Issues

- The commodity Internet constraint and the mission network constraint just mentioned are examples of policy-driven Internet2-level network limitations, but they may not be the only policy-driven problems which NLR may be used to overcome -- there may also be local policy artifacts.
- For example, it is easy to overlook the extent to which local perimeter firewalls (or other mandated "middleboxes") can cause problems for some applications, particularly if you're trying hard to go fast or do something innovative. It will often be virtually impossible to get an exemption from site-wide security policies for conventional connections.
- On the other hand, if you're bringing in a *lambda*, that lambda will both have a different security risk profile and may not even be able to be handled by available firewalls. Thus, it may be exempted from normal security mandates.

Coverage in Tough-to-Reach Areas?

- NLR could have been a way to tackle other issues, too.
- For example, NLR might have been a solution for some Internet2 members in geographically challenged parts of the country (e.g., our Northern Tier friends in the Dakotas, for example).
- Hmm… maybe, but remember that in NLR's case, the network footprint closely follows the existing Abilene map, with access network issues generally remaining the responsibility of a regional networking entity rather than being handled directly. NLR wasn't meant to fix the "Northern Tier" problem (although who knows what may become possible in the future).
- See http://www.ntnc.org/default.htm for more information about the Northern Tier Network Consortium.

Research Conducted Via the Network vs. Networking Research

- I would be remiss if I did not acknowledge that NLR does not exist solely for the purpose of serving those doing research via the network (such as those working with supercomputers, or physicists moving experimental data). Another major role is support for research about networking.

  Quoting Tom West:

  > "NLR is uniquely dedicated to network research. In fact, in our bylaws, we are committed to providing at least half of the capacity on the infrastructure for network research."
  >
  > http://www.taborcommunications.com/hpcwire/hpcwireWWW/04/1110/108776.html

Experimenting on Production Networks

- Most computer science networking experiments can be run on the Internet (or over Abilene) without disrupting normal production traffic. Some experiments, however, are radical enough that they have the potential to go awry and interfere with production traffic.
- When Abilene was first created, there was hope among computer scientists that it might remain a "breakable" network capable of supporting extreme network experimentation, but Abilene quickly became a production network upon which we all depended, and thus too mission-critical to potentially put at risk.
- Given that, one possible niche for a national lambda-based network would be as breakable infrastructure upon which risky experimentation can (finally) occur.
- Recall NLR's original role in the CALREN service pyramid

But Is A National Scale Breakable Lambda-Based Experimental Network What's Needed?

- When thinking about a breakable network testbed, the question that needs to be asked is, "Does such a network need to actually have a national footprint? Or could the same experiments be done in a testbed lab located at a single site, or perhaps on a state-scale or regional-scale optical network? Does that testbed need to be in the ground/at real facilities or could that sort of work be handled satisfactorily with reels of fiber looped back through WDM gear in a warehouse, instead?
- Is it sufficient for a national scale network testbed facility to be at the lambda level, or are we still "too high up the stack"? Will critical research involving long haul optics, for example, actually require the ability to work at layer 0, in ways that (once again) might be incompatible with production traffic running over that same glass?

General Possibilities vs. Specific Applications

- The preceding are all general possibilities relating to national optical networking.
- While it is fine to talk about general possibilities for NLR, when access to NLR becomes more broadly available, how, specifically, will lambda-based architectures likely end up being used?
- One approach to seeing what's well-suited to NLR is to take a look at how NLR is currently being used by early adopters, looking perhaps for common application themes or characteristics.

V. Current NLR Layer 1 Projects

Public NLR Layer 1 Projects

- There are a number of publicly identified NLR layer one (lambda-based) testbed projects at this time (see http://www.nlr.net/supported.html). They are:

  1) The Extensible TeraScale Facility (TeraGrid)
  2) OptIPuter
  3) DOE UltraScience Net
  4) Pacific Wave Extensible Peering Project
  5) Internet2 HOPI project

- Some additional projects not mentioned on that page include Cheetah and regional initiatives using NLR waves
- NLR also provided/will provide wavelengths for SC2004- and SC2005-related activities

The Sept 12th-14th 2005 NASA Meeting

- With respect to information about current applications, the timing of my talk is fortuitous: there was an invitation-only NASA meeting just earlier this month, at which roadmaps for many NLR projects were discussed. See: "Optical Networks Testbed Workshop 2" http://www.nren.nasa.gov/workshop8/
- **If you end up looking at only one presentation from that workshop, make it Robert Feurstein (Level3)'s: "A Commercial View of Optical Networking In the Near Future," http://www.nren.nasa.gov/workshop8/ppt/Level3_ONT2_7_v1.ppt (also known as the "Poppycock/Forgeddabout It/ Hooey/Malarkey" talk)**

1) Extensible TeraScale Facility (TeraGrid)

- The TeraGrid site describes its project as:
  "TeraGrid is an open scientific discovery infrastructure combining leadership class resources at eight partner sites to create an integrated, persistent computational resource. Deployment of TeraGrid was completed in September 2004, bringing over 40 teraflops of computing power and nearly 2 petabytes of rotating storage, and specialized data analysis and visualization resources into production, interconnected at 10-30 gigabits/second via a dedicated national network."
  (http://www.teragrid.org/about/)
- This is a major project:
  "U.S. computing grid gets $148 million boost"
  http://news.com.com/2100-7337_3-5841788.html

TeraGrid Sites and Lambdas

- http://www.teragrid.org/i/TG_10-20-04_1280.jpg shows a hub-and-spoke network architecture centered on Argonne, with radials running:
  -- Argonne-PSC
  -- Argonne-TACC (Univ of Texas Austin)
  -- Argonne-{Purdue,IU}-ORNL
  -- Argonne-Caltech-SDSC
  -- Argonne-NCSA
- Lambdas used by TeraGrid (per Tom West/SC2004's www.nlr.net/presentations/SC2004_TWW_Slides.htm):
  -- 3 Chicago-Pittsburgh
  -- 1 Chicago-Austin
  -- 1 Chicago-ORNL

Salient Characteristics of the TeraGrid

- One of the useful things about looking at existing testbed applications is that maybe as we look at them can see some common themes emerge:
  -- Lambdas were used as "glue" to stitch together regional optical networks
  -- Lambdas were allocated persistently (rather than dynamically) on NLR
  -- Primarily research via the network; not network research
  -- Supercomputing-related
  -- DOE-related; NSF-funded
  -- Uses a hub-and-spoke architecture
  -- Has some long runs (e.g., Chicago to San Diego)
  -- Has multiple lambdas used for at least one path (Chicago to Pittsburgh)
  -- Has at least one lambda shared across multiple end sites

2) OptIPuter

- OptIPuter defined: www.calit2.net/presentations/lsmarr/2005/SMARR-OpenHouse-OptIPuterAHMJan05.ppt – The OptIPuter is: "Optical networking, Internet Protocol, Computer Storage, Processing and Visualization Technologies
  - Dedicated Light-pipe (One or More 1-10 Gbps WAN Lambdas)
  - Links Linux Cluster End Points With 1-10 Gbps per Node
  - Clusters Optimized for Storage, Visualization, and Computing
  - Does NOT Require TCP Transport Layer Protocol
  - Exploring Both Intelligent Routers and Passive Switches

  "Applications Drivers:
  - Interactive Collaborative Visualization of Large Remote Data Objects: Earth and Ocean Sciences; Biomedical Imaging"
- $13.5 million in NSF funding over five years (beginning 2002) http://ucsdnews.ucsd.edu/newsrel/science/Optiputer.htm
- See also "OptIPuter Roadmap Summary 2006-2010," www.nren.nasa.gov/workshop8/ppt/OptIPuter_ONT2_7_v1.ppt

OptIPuter Sites and Lambdas

- CAVEwave Press Release www.evl.uic.edu/core.php?mod=4&type=4&indi=298
- Slide 6 of http://www.optiputer.net/events/ppts/DEFANTI-OptIPuter-AHMOpenHouse-28Jan2005.ppt shows OptIPuter nodes at Chicago, Kansas City, Denver, Salt Lake City, Seattle, Sunnyvale and Los Angeles (all along NLR path)
- Lambdas used (per Tom West's SC2004 talk):
  1 Chicago-Seattle
  1 Seattle-UCSD
- See also http://www.startap.net/translight/

Salient Characteristics of The OptIPuter

- Mambretti and DeFanti state that 'For the OptIPuter, the "Network" is A Large Scale, Distributed System Bus and Distributed Control Architecture; A "Backplane" Based on Dynamically Provisioned Datapaths' (OptIPuter Roadmap Summary 2006-2010 at slide 2)
- Persistent lambda allocation (although project apparently has great ongoing interest in dynamic light paths)
- Production traffic oriented
- Supercomputing-related
- NASA-related; NSF-funded
- Point-to-point/linear architecture
- Eastern termination of architecture at Chicago is interesting, perhaps reflecting international collaborations and reinforcing termination of transatlantic circuits in Chicago rather than NYC

3) DOE UltraScience Net

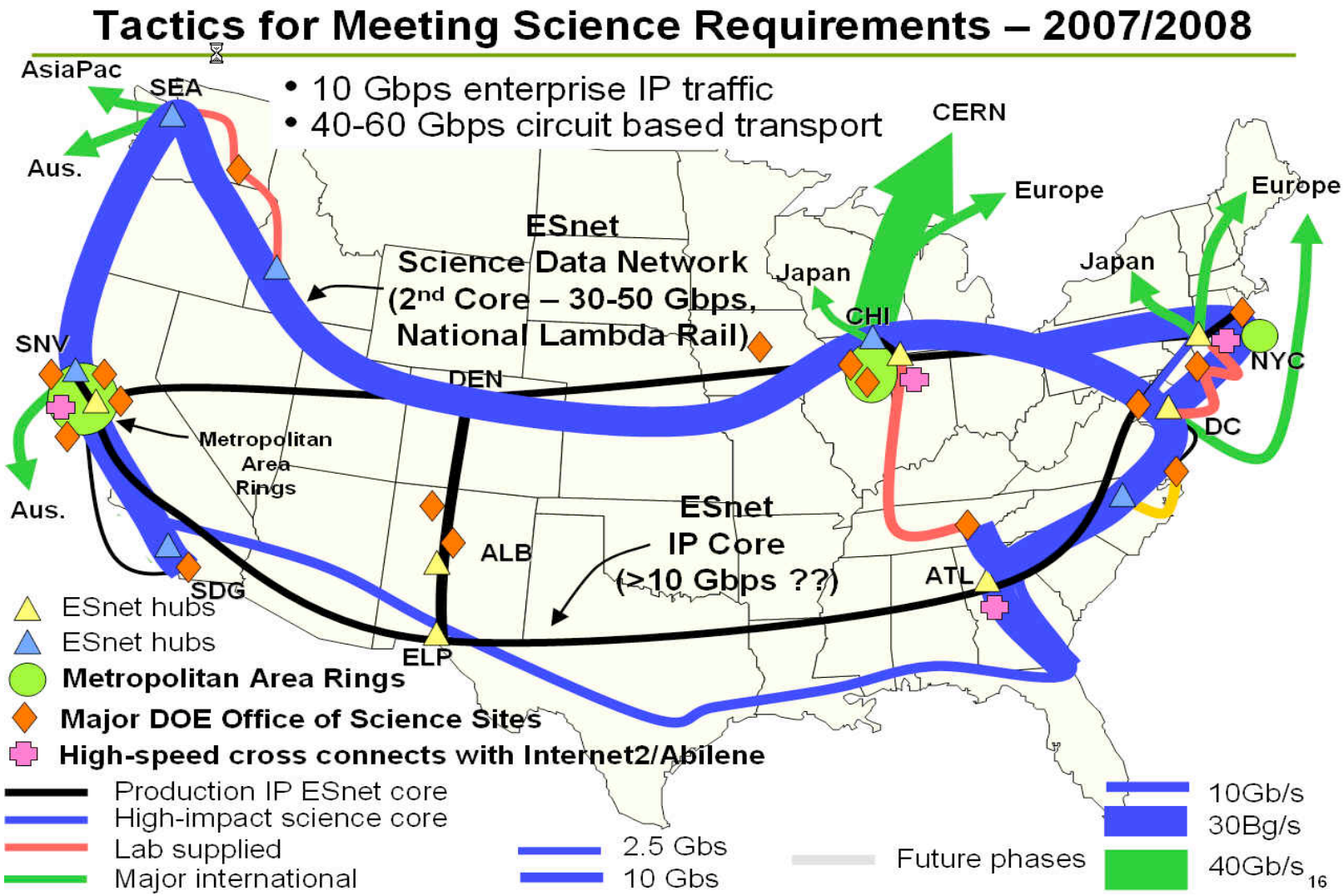
- What is DOE UltraScience Net?
  "The UltraNet is a research network funded by DOE Office of Science that provides a cross-country testbed consisting of multiple wavelengths provisioned on-demand. It provides the capabilties of:
  -- end-to-end on-demand dedicated paths at lambda and sub-lambda resolution
  -- packet switching at multiple OC192 rates
  -- collections of hybrid paths provided on demand."
  http://www.csm.ornl.gov/ultranet/summary.html

- See also: "DOE Ultra Science Net – In a Nutshell"
  http://www.nren.nasa.gov/workshop8/ppt/USN_ONT2_7_v1.ppt

Some Salient Characteristics of The DOE UltraScience Network

- Persistent lambda allocation from NLR for the service, but dynamic (on-demand) path allocation
- Research via the network, also research on networking (e.g., see www.sc.doe.gov/ascr/billwingstalk.ppt at pp. 9)
- Obviously a DOE project
- Lambdas used (per Tom West/SC2004 talk):
  2 Chicago-Seattle
  1 Seattle-Sunnyvale
- See also planned DOE Science Data Network core, "one component of a new three part ESNet network architecture, with that SDN intended for: large, high-speed science data flows; multiply connecting MAN rings for protection against hub failure; a platform for provisioned, guaranteed bandwidth circuits; alternate path for production IP traffic"

ESNet Science Data Network NLR-Related Plans



•http://www.internet2.edu/presentations/jtsaltlake/20040214-ESnetUpdate-Johnston.ppt – used with permission

4) Pacific Wave Extensible Peering Project

- Distributed AUP-free bilateral Internet peering point (Seattle and LA): http://www.pacificwave.net/about.html and http://www.cenic.org/projects/pacificwave/about.htm
- For those not familiar with peering points/exchange points, these are facilities where network service providers or ISPs can bring connections so that they can exchange customer traffic (and only customer traffic) with other participants on an as-arranged basis, often without financial settlements. A list of exchange points is at: http://www.ep.net/ep-main.html
- Peers at the Pacific Wave Extensible Peering Project (per www.cenic.org/projects/pacificwave/participants.htm):
  -- LA: Abilene, Calren, LosNetos, Qatar Foundation
  -- Seattle: Abilene, Canarie, Comcast, DREN, ESNet, Gemnet, Kreonet2, Microsoft, PNWGP, Peer1, PointShare, SingaREN, TANET2

Pacific Wave Extensible Peering Project Salient Characteristics

- A lambda was used as an interconnect fabric to glue the two exchange points together
- The required lambda was persistently allocated
- The project is production-traffic-oriented
- AUP free (edu, governmental, commercial partner traffic)
- Some network utilization data is publicly available; see: http://cricket.cenic.org/grapher.cgi?target=%2Futilization%2Fcenic-backbone%2Fpacific-wave See also http://stryper.uits.iu.edu/transpac2/
- Lambdas used (per Tom West's SC2004 talk): 1 Seattle-Los Angeles
- See also Pacific Wave's eastern analog, Atlantic Wave: http://www.nitrd.gov/subcommittee/lsn/jet/conferences/20050517/20050517_sobieski.pdf

5) Internet2 HOPI project

- "The Hybrid Optical and Packet Infrastructure (HOPI) project is examining a hybrid of packet and circuit switched infrastructures and how to combine them into a coherent and scalable architecture for next generation networks. The HOPI testbed utilizes facilities from Internet2 and the National Lambda Rail (NLR) to model these future architectures." (see http://networks.internet2.edu/hopi/)
- See also the HOPI testbed whitepaper linked from the HOPI web site, http://networks.internet2.edu/hopi/
- Differs from some of the other projects in that it is focused on research ABOUT networking, not research taking place via the network
- HOPI has one wavelength over the full NLR footprint, as well as some other resources

6) Cheetah

- "1 10 [G] wave Raleigh to Atlanta – Cheetah – MATP – U.VA" (http://www.nlr.net/docs/NLR.quarterly.status.report.200503.pdf)
- "LambdaRail Connection to Propel Va. Research Universities Into Future" referring to Cheetah and the creation of the MidAtlantic Terascale Partnership node in McLean VA at http://www.virginia.edu/topnews/03_22_2005/lamda.html
- "CHEETAH: Circuit-switched High-speed End-to-End Transport ArcHitecture," www.ece.virginia.edu/~mv/pdf-files/opticomm2003.pdf
- See also: http://www.nren.nasa.gov/workshop8/pps/09.B.CHEETAH_Habib.ppt

7) "Regional Projects"

- "11 additional 10G waves supporting a variety of projects at regional levels – FLR [Florida Lambda Rail] and CENIC/PNWGP"

  http://www.nlr.net/docs/NLR.quarterly.status.report.200503.pdf

8) Waves for Supercomputing

- "8 short terms 10 G waves for SC2004 Conference/ Exposition in November 2004" http://www.nlr.net/docs/ NLR.quarterly.status.report.200503.pdf

- And it appears that NLR waves will also be back at SC2005 in Seattle in November; see…

  http://www.nlr.net/sc05/

  http://www-iepm.slac.stanford.edu/monitoring/bulk/ sc2005/sc05-waves.jpg

Changes to NLR L1 Projects/Circuits

- Regardless of whether you're a current user of NLR facilities, or just curious, you may want to know what changes are happening to the NLR network infrastructure.
- Subscription to the NLR operations mailing list itself is closed/limited to NLR participants, but if you're so inclined anyone can review the National Lambda Rail Weekly Report archives at http://noc.nlr.net/nlrwr/
- Those reports provide an excellent overview of where NLR is at operationally, and make it easy to track changes which may be occurring
- Given Hurricane Katrina, some NLR work in progress may understandably take longer than it otherwise would on the "southern" half of the NLR build, still underway.

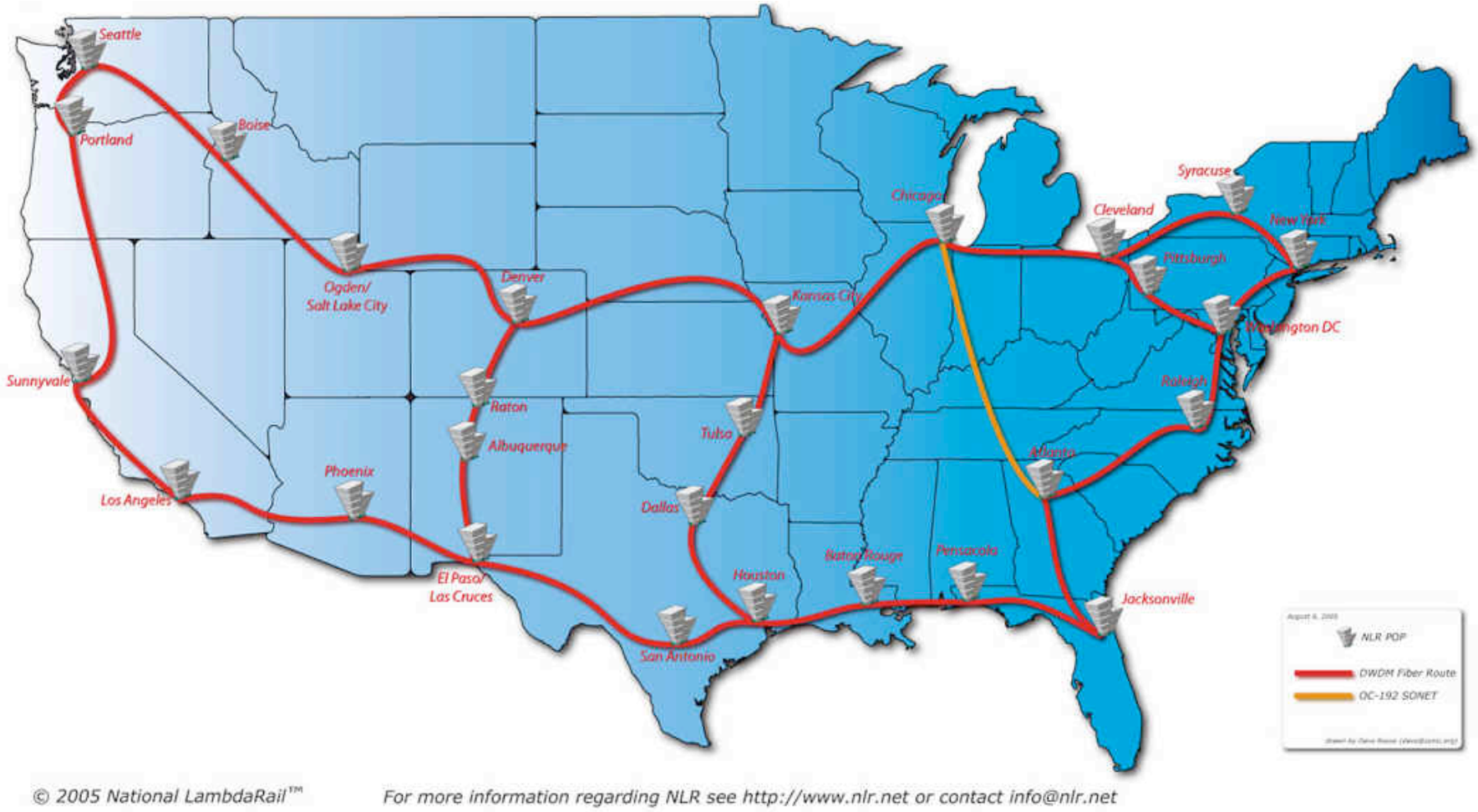
VI. NLR Native L2 and L3 Services

The NLR L2 and L3 Services

- In addition to the specific special projects mentioned in the preceding section (all basically L1 based), NLR also offers ubiquitous NLR layer two and layer three services to NLR participants. Those services represent a minimum commitment of two of the five pre-defined full footprint NLR waves:
  1) NLR Layer 2 service
  2) NLR Layer 3 service
  3) HOPI wave
  4) hot spare
  5) Wave in support of network research projects (being equipped by Cisco's Academic Research and Technology Group)

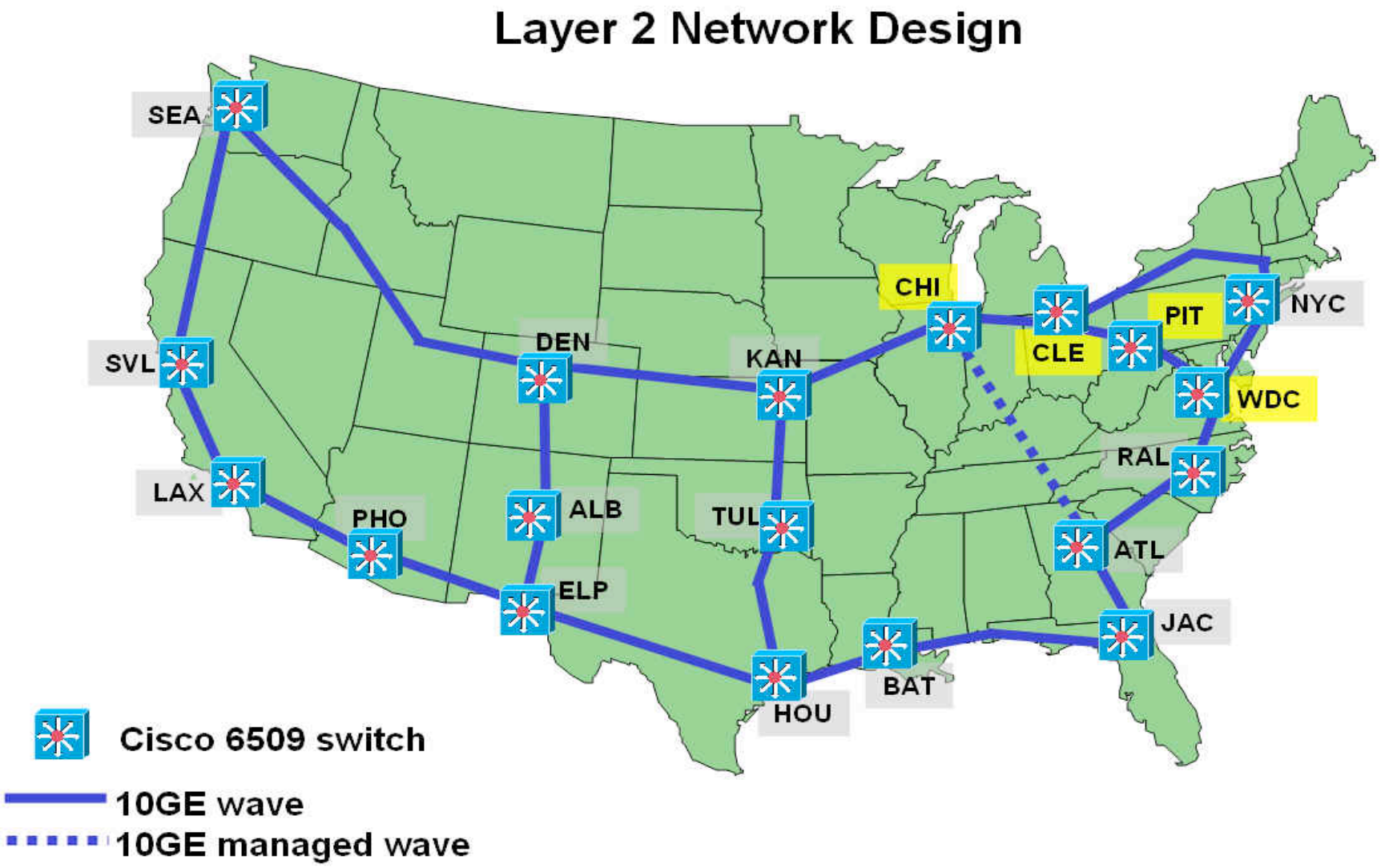
  www.nlr.net/docs/NLR.quarterly.status.report.200503.pdf

The Commonly Seen Map of NLR: Many L1 POPs



http://www.nlr.net/images/NLR-Map-large.jpg
Image credit: National Lambda Rail, used with permission.

The Less Commonly Seen NLR L2 Map: Fewer Nodes



http://www.internet2.edu/presentations/jtsaltlake/20050213-NLR-Cotter.ppt
used with permission

What Is the NLR L2 Service?

- Caren Litvanyi's talk "National Lambda Rail Layer 2 and 3 Networks Update" (http://www.internet2.edu/presentations/jtvancouver/20050717-NLR-Litvanyi.ppt) is excellent and provides the best description… Excerpts include:
- "Provide circuit-like options for users who can't use, can't afford, or don't need, a 10G Layer1 wave."
- "MTU can be standard, jumbo, or custom"
- "Physical connection will initially be a 1 Gbps LX connection over singlemode fiber, which the member connects or arranges to connect."
- "One 1GE connection to the layer 2 network is part of NLR membership. Another for L3 is optional."

What Is the NLR L2 Service? (cont.)

- Continuing to quote Litvanyi…
  "Initial Services:

  "--Dedicated Point to Point Ethernet – VLAN between 2 members with dedicated bandwidth from sub 1G to multiple 1G.

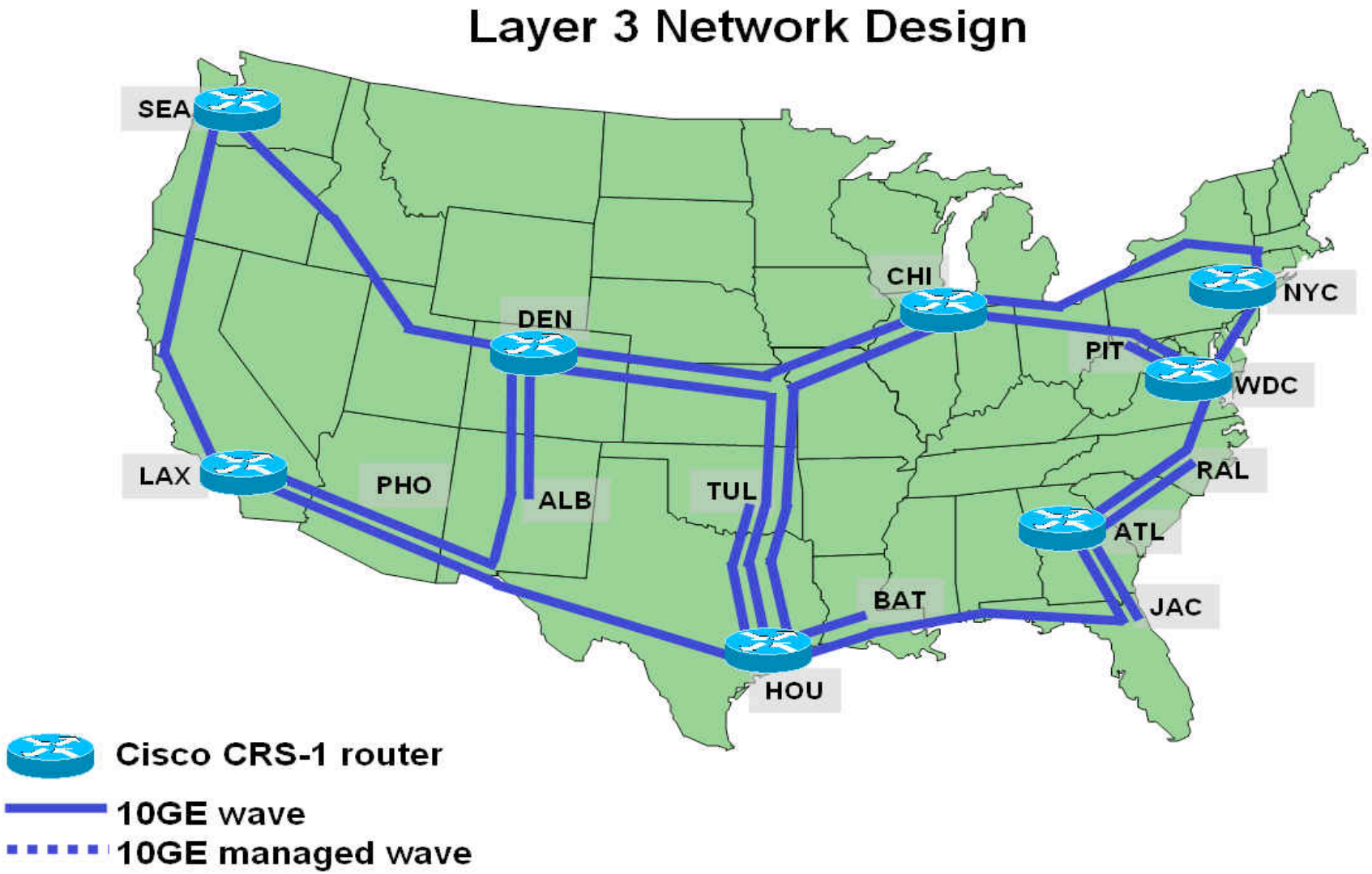
  "--Best Effort Point to Multipoint – Multipoint VLAN with no dedicated bandwidth.

  "--National Peering Fabric – Create a national distributed exchange point, with a single broadcast domain for all members. This can be run on the native vlan.
  This is experimental, and the service may morph."
- Litvanyi's talk includes a list of NLR L2 street addresses (can be helpful in planning fiber build requirements)

Some Thoughts About NLR L2 Service

- NLR L2 service is likely to be the most popular NLR production service among the pragmatic folks out there:
  -- it is bundled with membership at no additional cost
  -- the participant-side switch will be affordable
  -- the L2 service has finer grained provisioning that is most appropriate to likely load levels
- Hypothetical question: assume NLR participant wants to nail up point to point L2 VLAN with participant at CHI with dedicated 1Gbps bandwidth. Later, ten additional participants ALSO want to obtained dedicated 1 Gbps VLANs to CHI across some common part of the NLR L2 shared wave. What's the plan? Will multiple NLR lambdas be devoted to handle that shared L2 service load? Will some of that traffic get engineered off the hot link? Will additional service requests just be declined?

Another Less Commonly Seen Map…
The NLR L3 Map, With Just 8 Routing Nodes



http://www.internet2.edu/presentations/jtsaltlake/20050213-NLR-Cotter.ppt
used with permission

What Is NLR L3 Service?

- Again quoting Litvanyi's "National Lambda Rail Layer 2 and 3 Networks Update"…
- "Physical connection will be a 10 Gbps Ethernet (1310nm) connection over singlemode fiber, which the member connects or arranges to connect."
- "One connection directly to the layer 3 network is part of NLR membership, a backup 1Gbps VLAN through the layer 2 network is optional and included."

Random Notes About NLR L3 Service

- Probably obvious, but….
  Total $ Cost to NLR for each L3 routing node >>
  Total $ Cost to NLR for each L2 switching node >>
  Total $ Cost to NLR for each L1 lambda access POP
  (e.g., higher layer site also have the lower layer equipment)
- Demand for L3 service may be limited: 10Gbps routers and router interfaces don't come cheap.
- L3 participant backhaul will burn incremental lambdas; current L3 stubs shown on the map are: ALB <==> DEN, TUL <==> HOU, BAT <==> HOU, JAC <==> ATL, RAL <==> ATL, PIT <==> WDC. There will be more.
- Default L3 access link speed (10Gbps) is equal to the core network speed (10Gbps); implicitly, any L3 participant has sufficient access capacity to saturate the shared L3 core.
- NLR was assigned AS19401 for its use on 2005-05-31

Abilene and NLR L2/L3 Geographical Matrix

Site	Abilene Router	NLR CSR-1 Node	L3 Stub	L2 Node
Atlanta	X	X	n/a	X
Chicago	X	X	n/a	X
DC	X	X	n/a	X
Denver	X	X	n/a	X
Houston	X	X	n/a	X
Indianapolis	*X*	*NO*	*NO*	*NO*
Kansas City	*X*	*NO*	*NO*	*X*
LA	X	X	n/a	X
New York	X	X	n/a	X
Seattle	X	X	n/a	X
Sunnyvale	*X*	*NO*	*NO*	*X*
Albuquerque	NO	NO	X	X
Baton Rouge	NO	NO	X	X
Jacksonville	NO	NO	X	X
Pittsburgh	NO	NO	X	X
Raleigh	NO	NO	X	X
Tulsa	NO	NO	X	X
Cleveland	NO	NO	NO	X
El Paso	NO	NO	NO	X
Phoenix	NO	NO	NO	X

VII. So Let's Come Back to The Classic High Bandwidth Point-to-Point Traffic Scenario

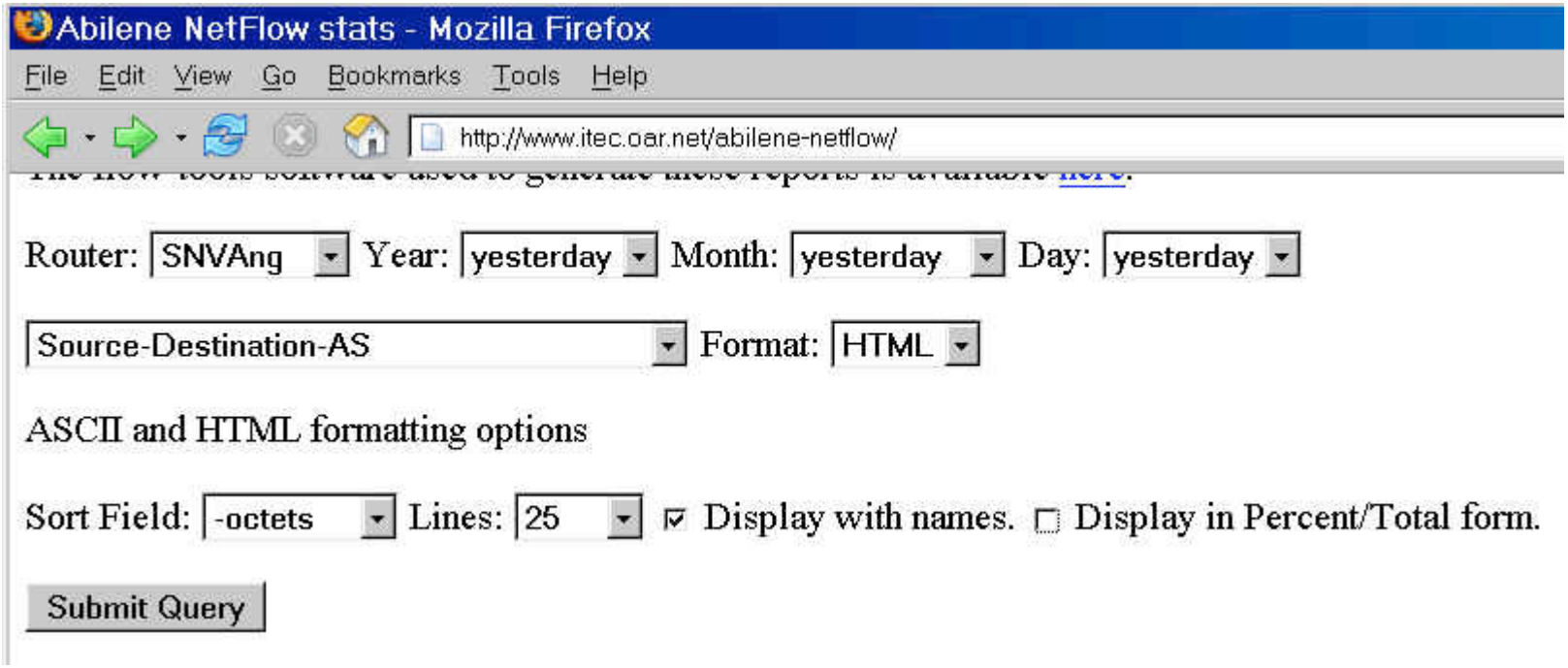
Sustained High Bandwidth Point-to-Point Traffic

- If you're facing sustained high bandwidth point-to-point traffic, that is usually pointed to as the classic example of when you might want to use a dedicated lambda to bypass the normal Abilene core.
- Qualifying traffic is:
  -- **NOT** necessarily the FASTEST flows on Abilene (why? because those flows, while achieving gigabit or near gigabit speeds, may only be of short duration)
  -- **NOR** are you just looking for a SINGLE large flow that transfers the most data per day (some applications may employ multiple parallel flows, or be "chatty," repeatedly opening and closing sessions, or there may be multiple applications concurrently talking between two sites, flows which when aggregated represent more traffic than any individual large flow).

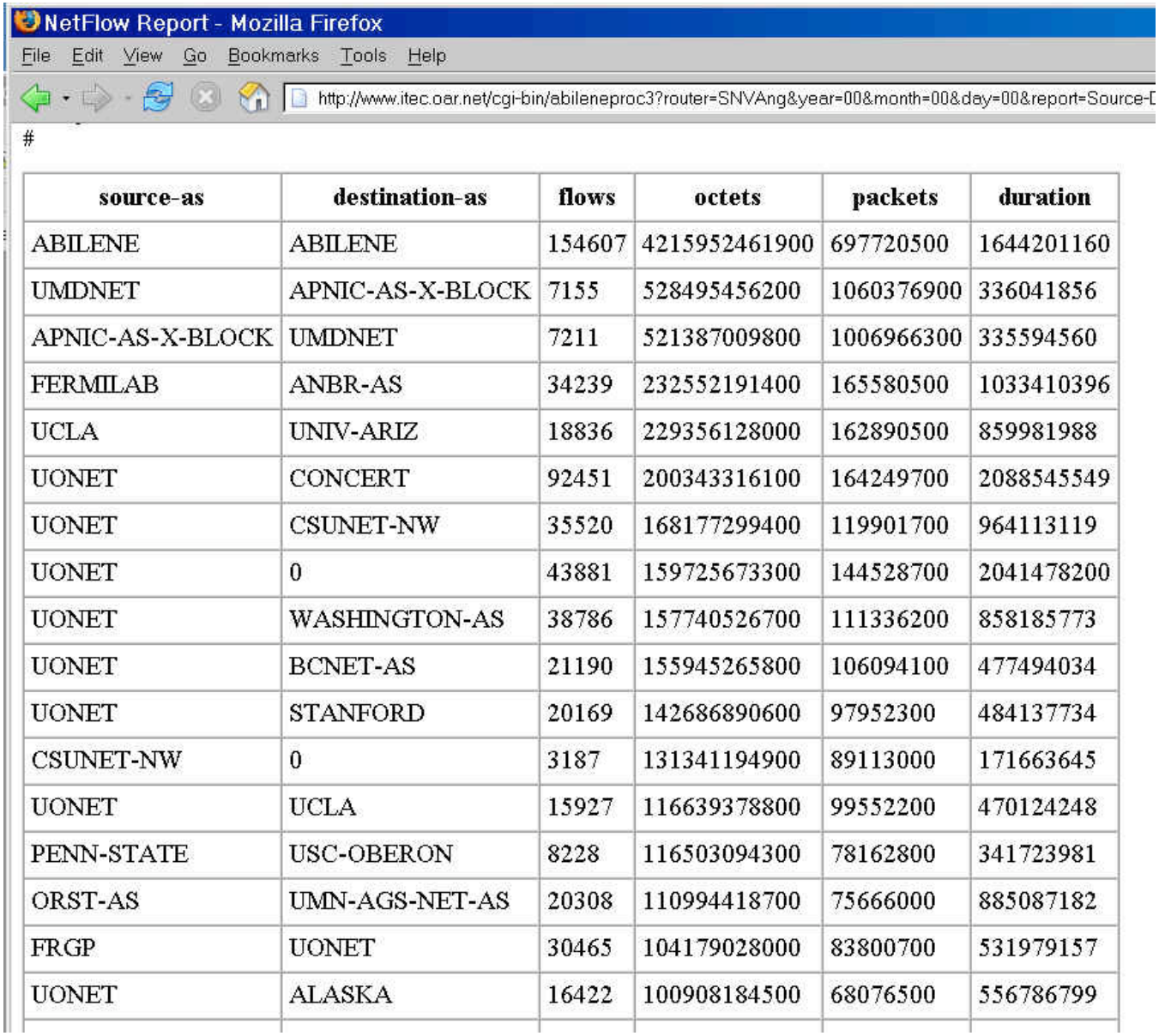
Identifying Potential Site Pairs for Lambda Bypass

- Okay then… so how do we spot candidate traffic which we might want to move off the Abilene core?
- First step in the process is basically the same one involved in hunting for commodity peering opportunities: analyze existing source X destination traffic matrices, looking for the hottest source-destination traffic pairs.
- Internet2 kindly provides netflow data, including per-node top source-destination aggregates. That data is usually available for each Abilene routing node.
- For example, we can look at what's happening at Sunnyvale (we'll only look at one day's worth of data; in reality, you'd obviously want to look at a much longer period to develop baselines)…

The Abilene Netflow Web Interface



Sample Output



source-as	destination-as	flows	octets	packets	duration
ABILENE	ABILENE	154607	4215952461900	697720500	1644201160
UMDNET	APNIC-AS-X-BLOCK	7155	528495456200	1060376900	336041856
APNIC-AS-X-BLOCK	UMDNET	7211	521387009800	1006966300	335594560
FERMILAB	ANBR-AS	34239	232552191400	165580500	1033410396
UCLA	UNIV-ARIZ	18836	229356128000	162890500	859981988
UONET	CONCERT	92451	200343316100	164249700	2088545549
UONET	CSUNET-NW	35520	168177299400	119901700	964113119
UONET	0	43881	159725673300	144528700	2041478200
UONET	WASHINGTON-AS	38786	157740526700	111336200	858185773
UONET	BCNET-AS	21190	155945265800	106094100	477494034
UONET	STANFORD	20169	142686890600	97952300	484137734
CSUNET-NW	0	3187	131341194900	89113000	171663645
UONET	UCLA	15927	116639378800	99552200	470124248
PENN-STATE	USC-OBERON	8228	116503094300	78162800	341723981
ORST-AS	UMN-AGS-NET-AS	20308	110994418700	75666000	885087182
FRGP	UONET	30465	104179028000	83800700	531979157
UONET	ALASKA	16422	100908184500	68076500	556786799

Percents rather than really big numbers…

```
# records:                  109178
# first-flow:               1126051201 Wed Sep  7 00:00:01 2005
# last-flow:                1126137597 Wed Sep  7 23:59:57 2005
# now:                      1126170674 Thu Sep  8 09:11:14 2005
#
# mode:                     streaming
# compress:                 off
# byte order:               little
# stream version:           3
# export version:           5
#
```

source-as	destination-as	flows	octets	packets	duration
11537	11537	0.380791	26.558490	3.930691	0.612359
27	2500	0.017622	3.329269	5.973759	0.125154
2500	27	0.017760	3.284489	5.672864	0.124987
3152	1251	0.084329	1.464968	0.932817	0.384879
52	1706	0.046392	1.444834	0.917663	0.320288
3582	81	0.227703	1.262067	0.925320	0.777849
3582	2152	0.087484	1.059437	0.675480	0.359070
3582	0	0.108077	1.006196	0.814220	0.760319
3582	73	0.095528	0.993690	0.627226	0.319619
3582	271	0.052190	0.982381	0.597694	0.177836
3582	32	0.049675	0.898859	0.551826	0.180310

Some Thoughts on That Sample Traffic Data…

- For Sunnyvale, for this day, the top source-destination pair (>26% of octets) is obviously intra-Abilene traffic (presumably iperf measurement traffic).
- It would probably not be a good idea to move traffic that's specifically designed to characterize the Abilene network onto a network other than Abilene. Some things you just need to leave where they are. :-)
- Excluding measurement traffic, nothing else jumps out at us at the same order of magnitude… ~3% of traffic seen at that site (the next highest traffic pairing) is probably not enough to justify pulling that traffic out of the shared Abilene path for those nodes, especially since the Abilene backbone itself is still uncongested.
- The lack of promising opportunities for bypass shouldn't be surprising since traffic normally isn't highly localized.

And Even 10% of 3Gbps Wouldn't Be All That Much

- If you assume that…
  -- the Abilene core as shown on the Abilene weather map is running *maybe* 3Gbps on its hottest leg
  -- an absurdly high estimate for the level of flow locality (or point-to-point concentration) might be 10% of that, excluding iperf traffic (remember, reality is ~3%)
  -- the unit of granularity for bypass circuits is a gigabit…
  THEN you really don't have much hope for discovering a set of ripe existing gigabit-worthy bypass opportunities: 10% of 3Gbs is just 300 Mbps
- Yeah, 300 Mbps isn't peanuts, but it also isn't anything that the existing Abilene core can't handle, and it seems a shame to "waste" a gig (or even 10gig!) circuit on just 300Mbps worth of traffic when the existing infrastructure can handle it without breaking a sweat.

What About From The Perspective of an Individual Connector?

- Even if it doesn't make sense from Abilene's point of view to bother diverting a few hundred Mbps onto NLR, what about from the perspective on an individual connector? For example, what if an Abilene OC12 (622 Mbps) connector was "flat-topping" during at least part of the day? Should they try diverting traffic onto NLR, bypassing/offloading their hypothetical current Abilene OC12 connection, *or* should they upgrade that regular Abilene connection to GigE, OC48, or 10GigE/OC192?
- The issue is largely economic – NLR costs a minimum of $5 million over 5 years, while the incremental cost of going to even 10GigE/OC192 from OC12 is just ($480,000/yr-$240,000/yr), or $1.2 million over 5 years. If you as a connector need more capacity, just upgrade your existing Abilene circuit.

ASNs vs. Larger Aggregates

- The analysis mentioned on the preceding pages was done on an autonomous system by autonomous system (ASN x ASN) basis. [If you're not familiar with ASNs, see http://darkwing.uoregon.edu/~joe/one-pager-asn.pdf for a brief overview.] At least in the case of NLR lambdas, ASNs may be too fine a level of aggregation.
- Given the consortial nature of many NLR connections, it may make more sense to analyze traffic data at the NLR-connection X NLR-connection level instead.
- We keep coming back to the problem, though, that core Abilene traffic levels, while non-trivial, just aren't high enough to justify the effort of pruning off existing flows.

"What About Those Anticipated Huge Physics Data Flows I Keep Hearing About?"

- If you're thinking of the huge flows that are expected to be coming in from CERN, those will be handled by NLR all right, but via the DOE Science Data mission network described earlier in this talk. I'm fully confident that they've got things well in hand to handle that traffic, ditto virtually any other commonly mentioned mega data flows.
- If you know an example of one that's NOT already being anticipated and provided for, I'd love to hear about it.

VIII. The Paradox of Relative Resource Abundance

One Wavelength? Plenty.
Forty Wavelengths? Not Enough.

- Abilene currently runs on just one wavelength – 10 Gbps -- and that's enough, at least for now.
- NLR, on the other hand, will have forty wavelengths -- 400 Gbps -- but because of the way those wavelengths may get allocated, that may not be "enough" (virtually from the get go).
- It would thus be correct, in a very Zen sort of way, to talk about it being both very early, and possibly in some ways already "too late," when it comes to getting involved with NLR.

Do The Math…

- We start with 40 waves, half reserved for network research
- Of the remaining 20, at LEAST four were allocated "at birth" (shared L2 service, shared L3 service, HOPI, 1 hot spare) -- 16 are left after that. (I say "at least 4" because L2 service may be so popular that it could need multiple lambdas.)
- There are 15 known NLR participants already. If each participant wanted even *one* full-footprint non-research lambda for its own projects, well…
- Some projects use multiple parallel waves across a common path, or long resource-intensive transcontinental waves; other participants need to have L3 connections backhauled to the nearest L3 router node, etc.
- Add additional new Fednet/Int'l/Commercial participants…
- Before you know it, you're out of waves, at least at some locations, and you're just getting going.

"What About The Southern Route?"

- Whenever things look tight this way, folks always look at the redundant connectivity engineered into the system – in NLR's case, "What about the (still being completed) Southern Route?" I assert that it would be a really bad idea to book your backup capacity for production traffic. Gear fails. Backhoes eat fiber. Hurricanes flood POPs. Disgruntled employees burn down data centers. You really want redundant capacity to handle misfortunes.
- So, if my capacity analysis is correct, I believe NLR should either be looking at higher density WDM gear (to get more waves onto their existing glass), higher bandwidth interfaces (so they can avoid parallel 10 gig link scenarios) or if it is cheaper, they should be thinking about preparing to acquire and light additional fiber.
- Or you could redefine what's "network research" :-)

NLR May Also Have Pricing Issues

- I suspect NLR might run into pricing issues, too. It is really hard to get pricing right so that capacity get efficiently used.
- Too high? Capacity lies idle. No one uses the resource.
- Too low? Capacity gets allocated inefficiently and gobbled up prematurely (and in extreme cases, you don't generate enough revenue to purchase the next increment of capacity you may need).
- NLR may have a tough price point to hit:
  -- assume NLR costs $100 million invested over 5 years to build, or $20 million/year
  -- ($20 M/yr) / 40 waves ==> $500K/wave/yr (asset value)
  -- **But you can get an Abilene 10Gig for less, $480K/year**
- Complications: $480K/year is ongoing; NLR investment probably has a life > 5 years; time value of money isn't considered; unclear how lambdas will be priced; etc.

Speaking of Pricing Waves…

- Pricing lambdas might also be a funny thing. You could talk about charging a flat fee for a full footprint wave, and selling only that, but that's pretty inelegant and inefficient
- You could charge on a per-lambda route-mile basis. Short haul customers and the folks back east would love that – bills would be miniscule there. Folks looking at the vast open distances found in the west, however, would howl like coyotes at the bills they'd get.
- Another alternative would be to adopt postalized pricing, and charge a flat rate to nail up a lambda between any two points on NLR. This is simple, and great for the west, but one that would "overcharge" short haul users.
- Other options include charging the actual cost of providing each particular facility (tedious), or auctioning lambdas (that could get ugly in competitive markets).

Hypothetical

- NLR is AUP free (e.g., commercial traffic is allowed)
- Assume university X purchases a full footprint wave "cheap"
- Said university is entrepreneurial, and uses that wave to construct the core of a commercial Internet backbone, perhaps initially "camouflaged" as "just" an Internet service for alumni, a national scale student network "training environment," whatever.
- Said commercial Internet backbone, run by university X, now generates significant revenue, enough to underwrite new numerical compute cluster, student legal P2P music program, new faculty parking structure, you name it.
- This scenario will not happen because: _____ _____?

IX. Conclusions

- If you're a typical Abilene site, you probably don't need NLR (you may want NLR, and that's great, or having NLR may help you get research funding, but you probably won't need NLR to handle typical application traffic)
- Abilene 10GigE/OC192 connections are a real bargain
- If you have special policy-driven circumstances (e.g., you're on a federal mission network, or you want to do interesting things with commercial Internet traffic), NLR's probably the best thing since sliced bread.
- If you're a computer scientist who actually wants to do research about networking, "NLR's" original purpose, NLR is just waiting for you. :-)
- We may shortly see some very interesting capacity gyrations and economic phenomena occur.
- It will also be fascinating to see what happens if NLR and Internet2 merge, or I2 picks NLR for the "next Abilene"

Thanks For The Chance to Talk Today!

- Are there any questions?