

100G Deployment

Challenges & Lessons Learned from the ANI Prototype & SC11

Chris Tracy, Network Engineer

ESnet Engineering Group





Outline

- Background on ANI
- 100Gbps optical transmission over the WAN
- Pluggable optics
- Testing, measurement, debugging, fault isolation
- Interoperability
- Case Study
- References



Advanced Networking Initiative (ANI)

100GigE prototype circuits



01/11/11

3



Advanced Networking Initiative (ANI)

~13,000mi dark fiber



01/11/11

3



Advanced Networking Initiative (ANI)

experimental testbed



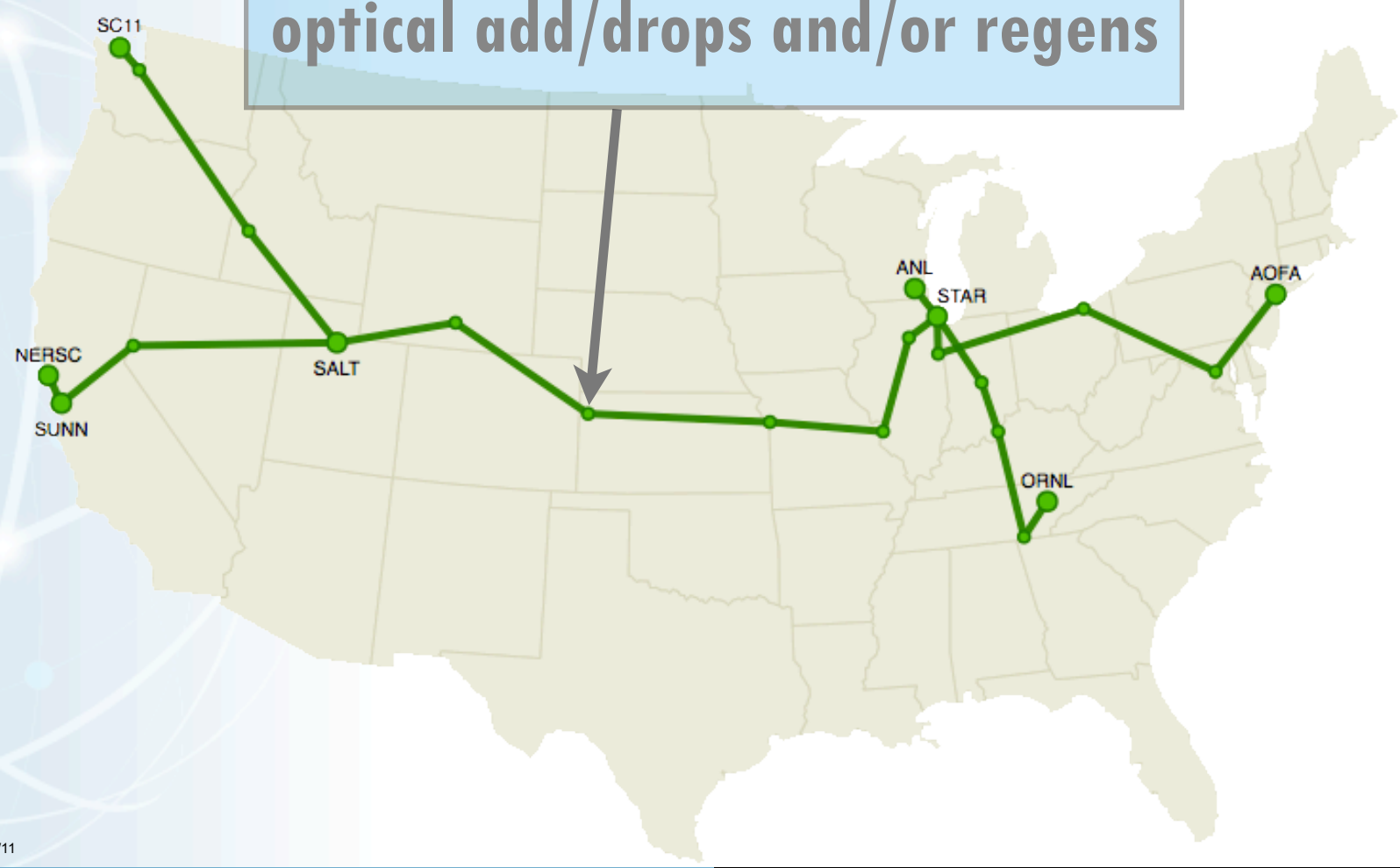
01/11/11

3



Advanced Networking Initiative (ANI) Seven 100GigE Circuits for SC11

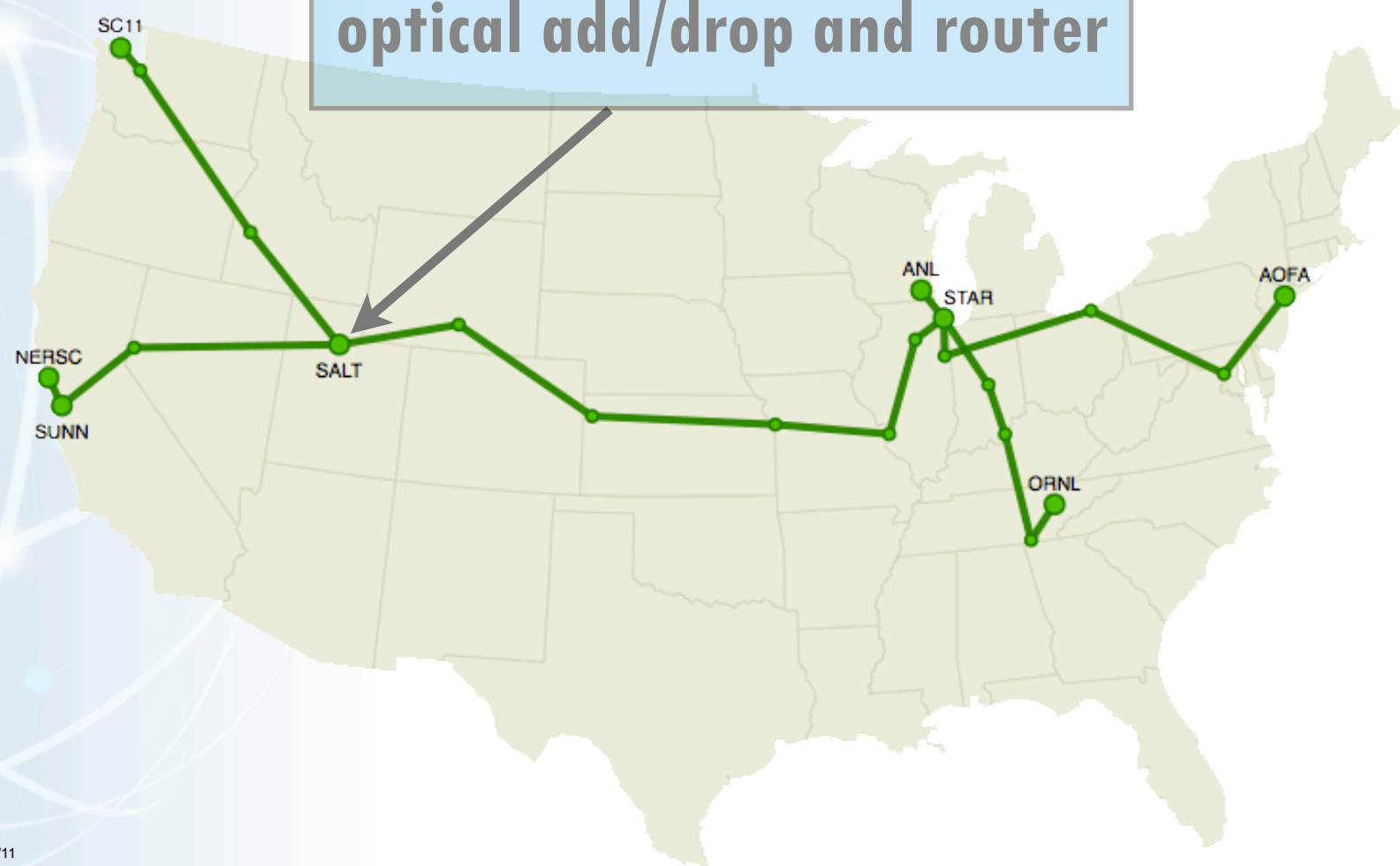
optical add/drops and/or regens





Advanced Networking Initiative (ANI) Seven 100GigE Circuits for SC11

optical add/drop and router





100G Optical Transmission over the WAN

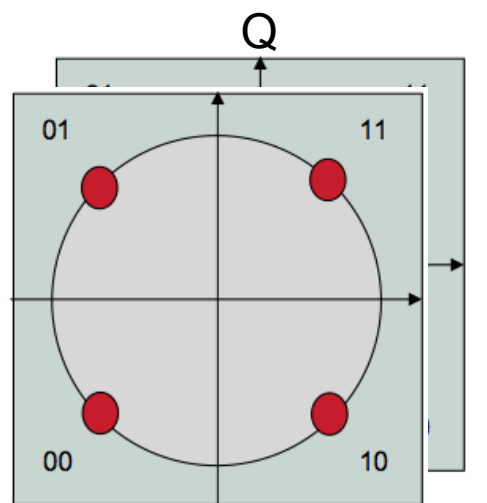
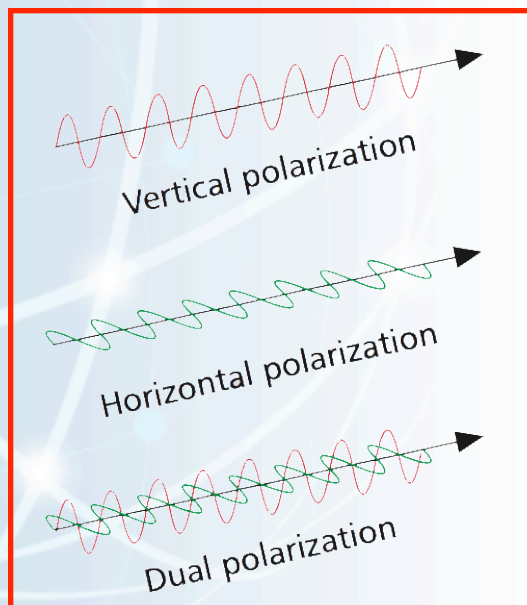
System Perspective

- 100G Ethernet client signals - IEEE P802.3ba [2]
 - on core-facing *and* customer/peer-facing edge ports
 - we have standardized on 100GBASE-LR10 CFPs
- 100G client signals mapped into Optical Transport Unit 4 (OTU4)
 - ITU-T G.709 [3] for encapsulating 100GigE into OTU4
 - includes mandatory Forward Error Correction (FEC)
- transported using dual polarization-quadrature phase shift keying (DP-QPSK) technology with coherent detection [4]

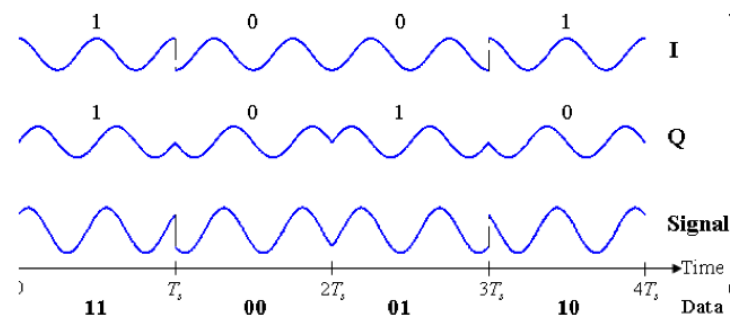
100G Optical Transmission over the WAN

Dual polarization-quadrature phase shift keying (DP-QPSK)

- DP: two independent optical signals, same frequency, orthogonal
 - single transmit laser, each signal carries half of the data
 - two polarizations → lower modulation rate → reduce optical bandwidth
 - allows 100G payload (plus overhead) to fit into 50GHz of spectrum



phase-amplitude constellation
2 bits per symbol



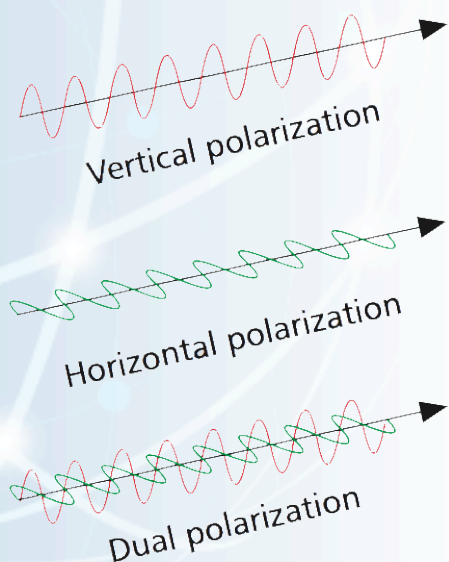
signals added together to form QPSK signal

source: [1] and [4]

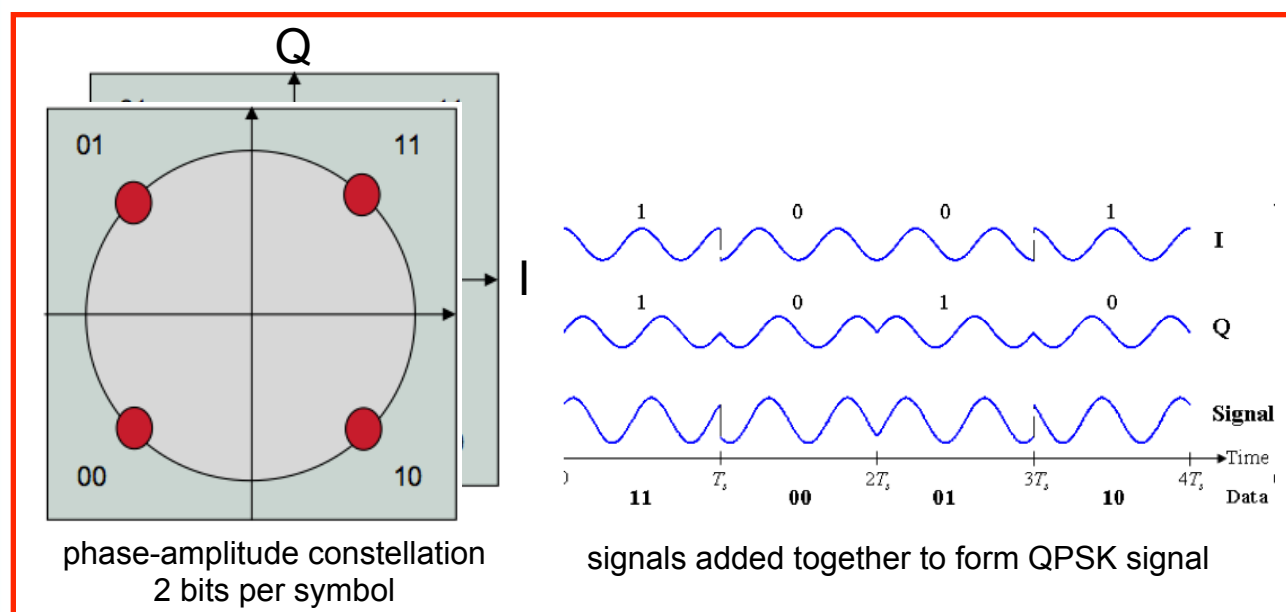
100G Optical Transmission over the WAN

Dual polarization-quadrature phase shift keying (DP-QPSK)

- QPSK: encode data by changing the phase of the optical carrier
 - compare to on-off keying (OOK), intensity modulation → '0'=off, '1'=on
 - further reduces the symbol rate by half, sends twice as much data
- Together, DP and QPSK reduce required rate by a factor of 4



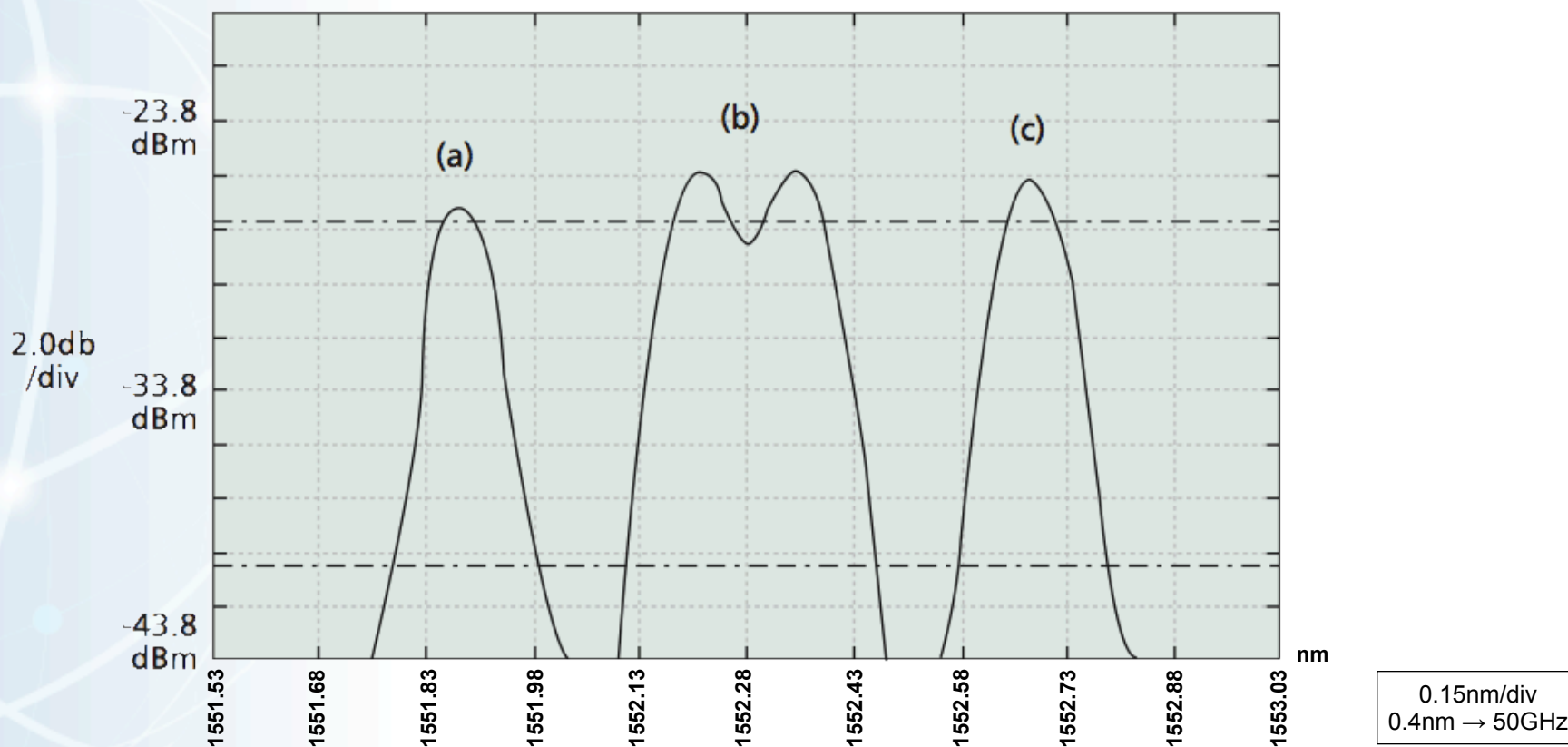
01/11/11



source: [1] and [4]

7

Spectral Efficiency: 10 vs 40 vs 100Gb/s



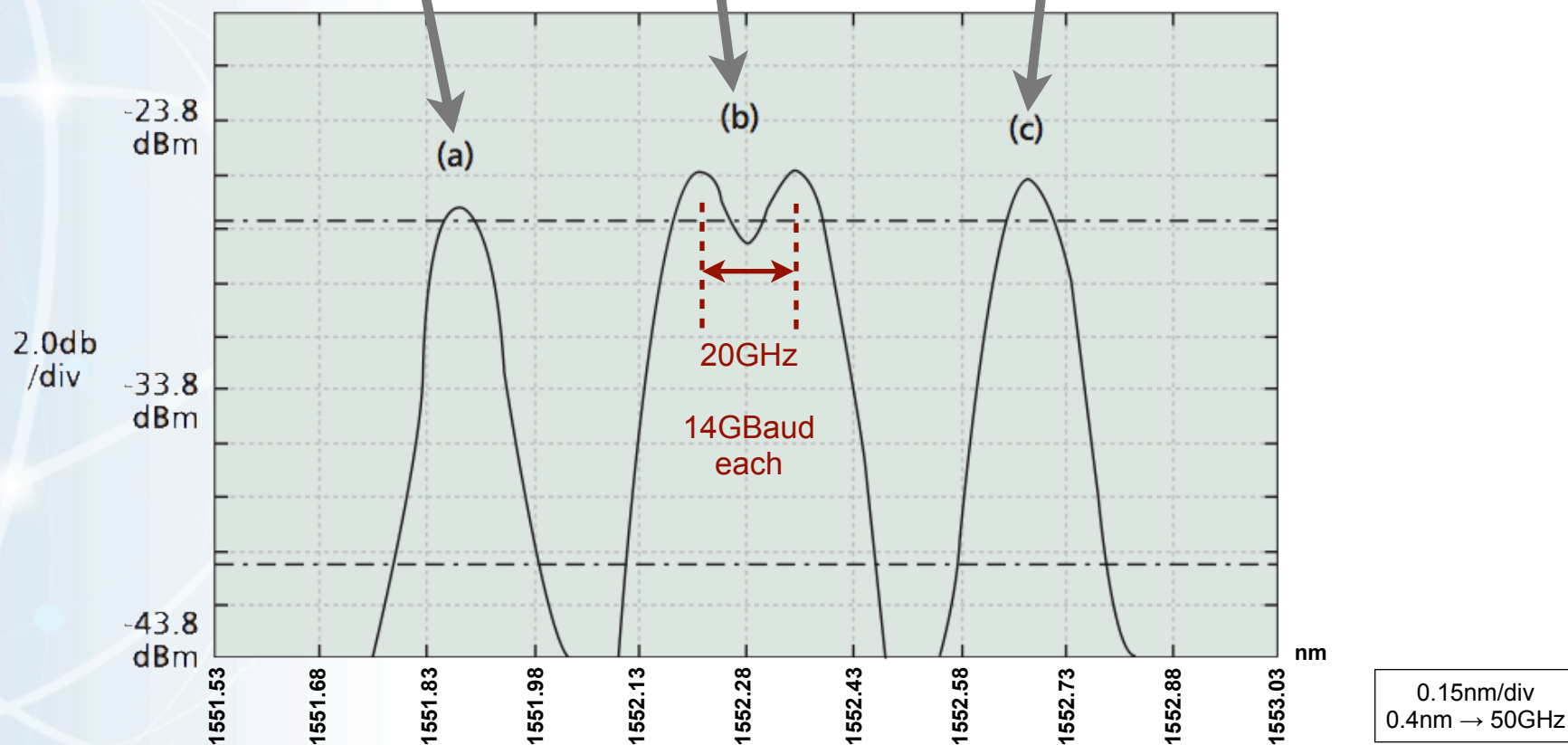
Source: [1] Roberts, Beckett, Boertjes, Berthold and Laperle (2010, July)

Spectral Efficiency: 10 vs 40 vs 100Gb/s

100G coherent dual-polarization dual-carrier

10G single-polarization single-carrier

40G coherent dual-polarization single-carrier



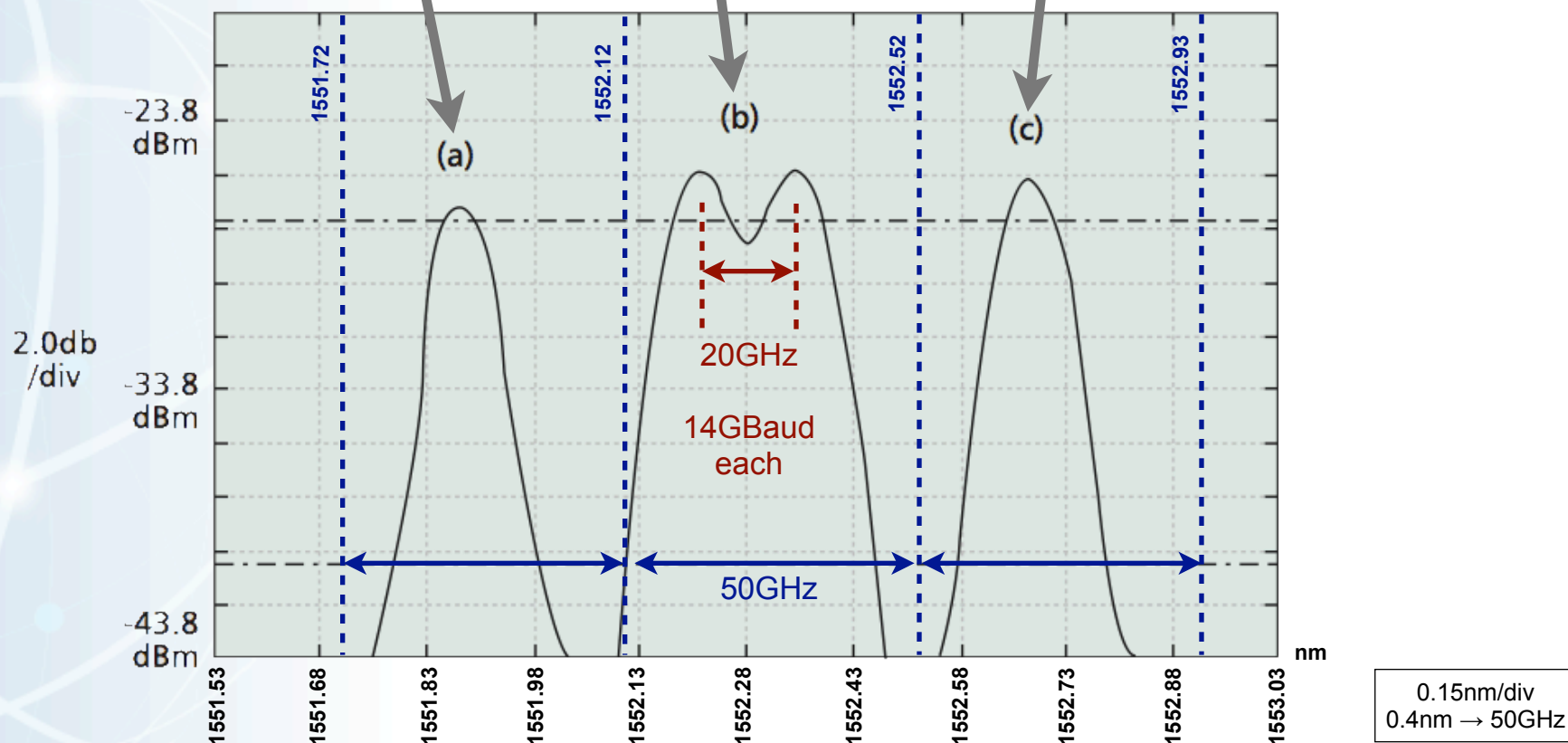
Source: [1] Roberts, Beckett, Boertjes, Berthold and Laperle (2010, July)

Spectral Efficiency: 10 vs 40 vs 100Gb/s

100G coherent dual-polarization dual-carrier

10G single-polarization single-carrier

40G coherent dual-polarization single-carrier



Source: [1] Roberts, Beckett, Boertjes, Berthold and Laperle (2010, July)



100G Optical Transmission over the WAN

Coherent Detection (on the receiver) - Breakthrough Technology

- Technology originally developed circa 1980s - [1] and [7]
 - advances in digital signal processing (# of gates, operating frequency) allowed coherent detection to emerge as disruptive technology in optical communications
- Offers significant improvement in noise tolerance over conventional direct detection schemes
- Able to compensate for propagation impairments such as chromatic dispersion (CD) and polarization mode dispersion (PMD)
 - dispersion compensating fibers are no longer needed in long-haul applications
- high-speed A/D samples incoming analog components
 - DSP ASIC to apply numerical adaptations, recover signal



100G Optical Transmission over the WAN

Coherent Detection - Another Breakthrough

- works like a radio receiver - source: [1], [4], and [7]
 - use a strong local oscillator tuned to the frequency of interest
- has provided a breakthrough in optical filtering capabilities
 - tunable optical filters exist but have never been cost-effective
- important impacts of this technology:
 - colorless ROADMs are [finally] becoming available by using coherent optical filtering - complete software reprogrammability
 - no longer need fixed channel filters (arrayed waveguide gratings)
 - no longer important to stay aligned to a 50 or 100 GHz ITU grid
 - spectrum can be used more efficiently, system is more flexible
 - wavelength selective elements would need to become flexible



100G Optical Transmission over the WAN

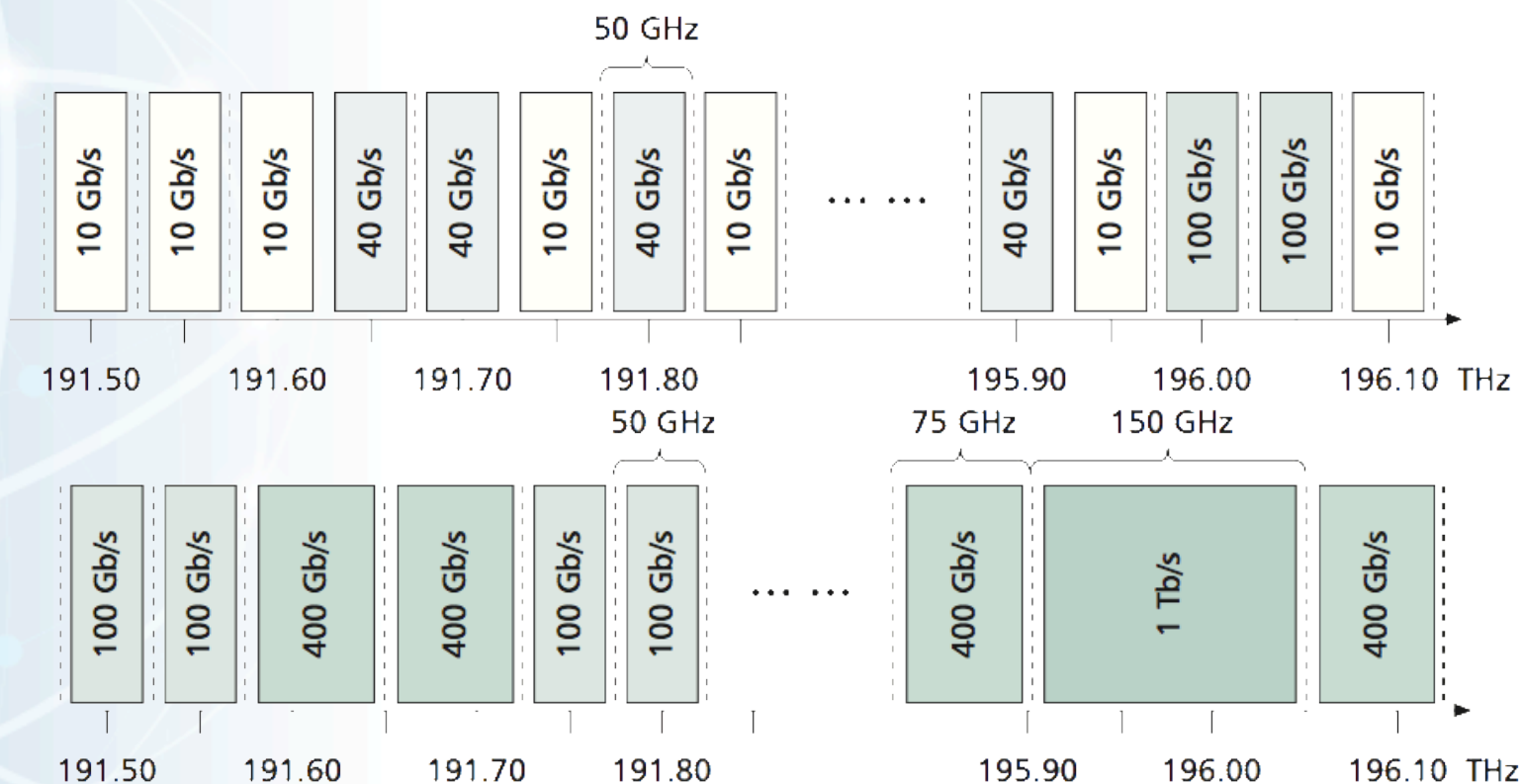
Amp placement and fiber span length has become very important

- 100G transmission for a given channel of optical spectrum is approaching Shannon limit
 - going higher than 100Gb/s, at the same distance, in same amount of spectrum, requires improvement of SNR (or just use more spectrum)
- Even if we just want to stay at 100Gb/s, reach is limited by OSNR
 - EDFAs contribute noise - Amplified Spontaneous Emission (ASE)
 - on a long segment consisting of cascaded EDFAs, a few very long (high-loss) fiber runs degrades the system's SNR
 - could break up long spans and place more amplifiers for less loss between amps, or possibly deploy Raman, to improve SNR
 - otherwise, extra 100G re-gens could be required to get the desired reach - might not get "advertised" reach if amp spacing is not ideal

Flexible Grid Concept

At rates $>100\text{Gb/s}$, 50GHz spacing would require high SNR, will limit reach

- Future-proof network by allowing channels to occupy more spectrum



Source: [8] Gringeri, Basch, Shukla, Egorov and Xia (2010, July)

CFP Pluggable Optics - LR4 or LR10?

LR10 seems to be gaining popularity and vendor support

Does your equipment support third-party optics?

- Similar to 1G & 10G pluggables, vendors sell certified optics
- 3rd party may not work (or be supported) by your equipment

Digital Diagnostic Monitoring (optical power monitoring)

- may or may not be supported, especially with uncertified optics
- if supported - may be aggregate, per-lane, or both



Interface	Wavelength	Cost	Fiber Type	Connector	Reach	Link Budget	Comment
100GBase-LR4	4 x 25G WDM lanes 1294.53-1310.19nm	\$\$\$	SMF	SC / LC	10 km	7.3 dB	IEEE Standard [2]
100GBase-LR10	10 x 10G WDM lanes 1521-1597nm	\$\$	SMF	SC / LC	4-10 km*	5.4 dB	non-standard, no 10:4 gearbox, 10x10 MSA [5]
100GBase-SR10	10 x 10G parallel MMF 840-860nm	\$	parallel MMFs	MPO / MTP	100 / 150 m OM3/OM4 MMF	8.3 dB	IEEE Standard [2]

* reach may vary by model & transmit rate



Pluggable Optics - Optical Power

Field Testing of 100G Transceivers

- CFPs will appear quite hot with conventional power meters
- testing spectrum compliance and per-lane optical power with parallel optics on SMF requires use of an Optical Spectrum Analyzer
- some power meters may be “tunable” to measure different λ 's (consider passband width, marginal for spectrum compliance)
- every CFP will have a slightly different Tx lane transmit profile
 - maximum reach over dark fiber may vary slightly depending on this profile



Interface	Lanes	Average Launch Power Each Lane (min / max)	Total Average Launch Power (max)	Source
100GBase-LR4	4	-4.3 to 4.5 dBm	+10.5 dBm	[2]
100GBase-LR10	10	-5.8 to 3.0 dBm	+13 dBm	[5]

Pluggable Optics - Other Considerations

100G Fiber Patches

- LR10: can generally* be connected back-to-back without attenuation
- SR10: MPO-terminated parallel MMF for short patches (we haven't tried this)
 - going through patch panels would be messy
 - break-out using 1xMPO to 10xSC octopus cables?

CXPs

- high-density, targeting MMF connections
- to be interoperable with CFPs
- active optical cables



source: [6]



Testing and Measurement

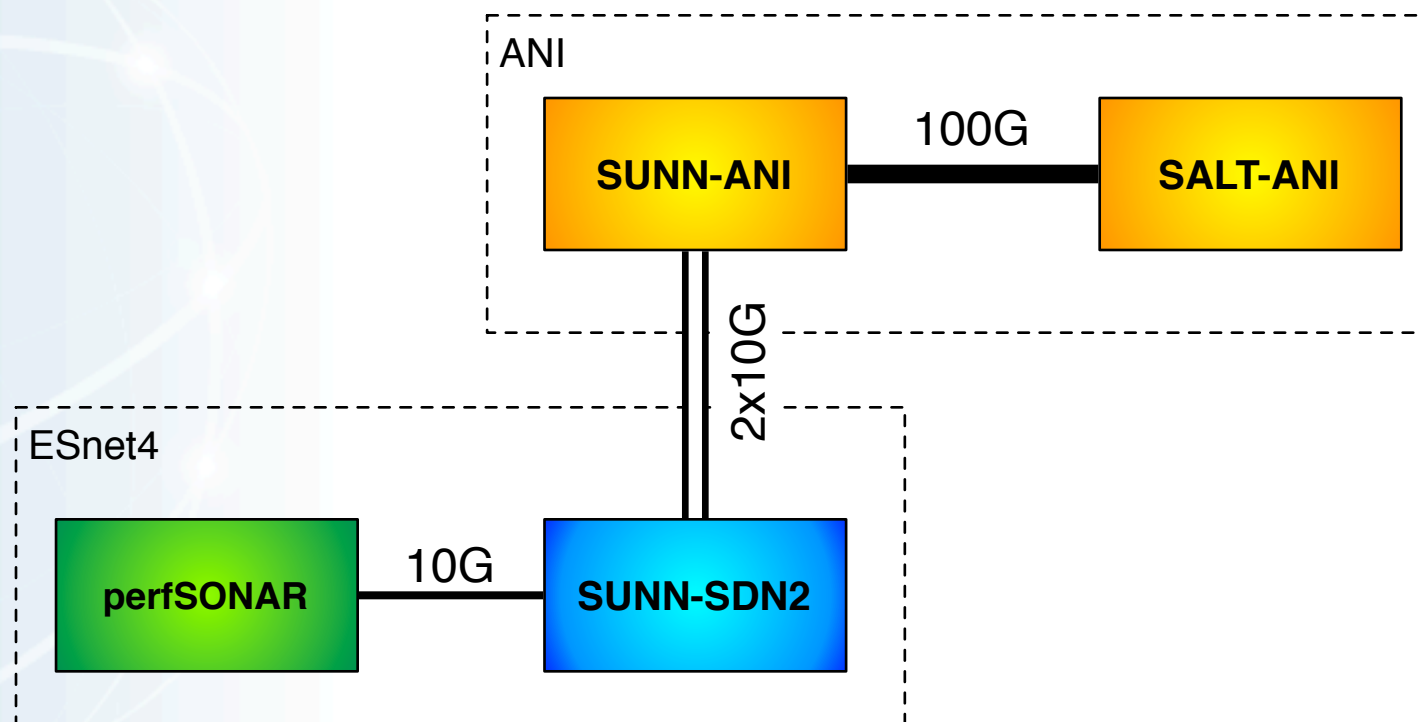
October 3rd, 2011, 14:07 Pacific

- First pings across transcontinental 100GigE between two routers
- 41 days left before the start of SC11
 - Want to give users as much time as possible to tune their applications
- Ping is a good first step, need more testing before hand-off to users
 - Transport system shows clean FEC, low BER, optical layer clean
 - We want to drive links at 100% utilization and measure 0 drops
 - Validate QoS, throughput, latency, loss, interoperability, etc.
- How are we going to test seven 100GigE circuits?
 - At this point, work on building/deploying hosts capable of sourcing >10Gbps of traffic had begun, still preliminary
 - Do we need a 100GigE hardware tester?



Testing and Measurement - UDP flows

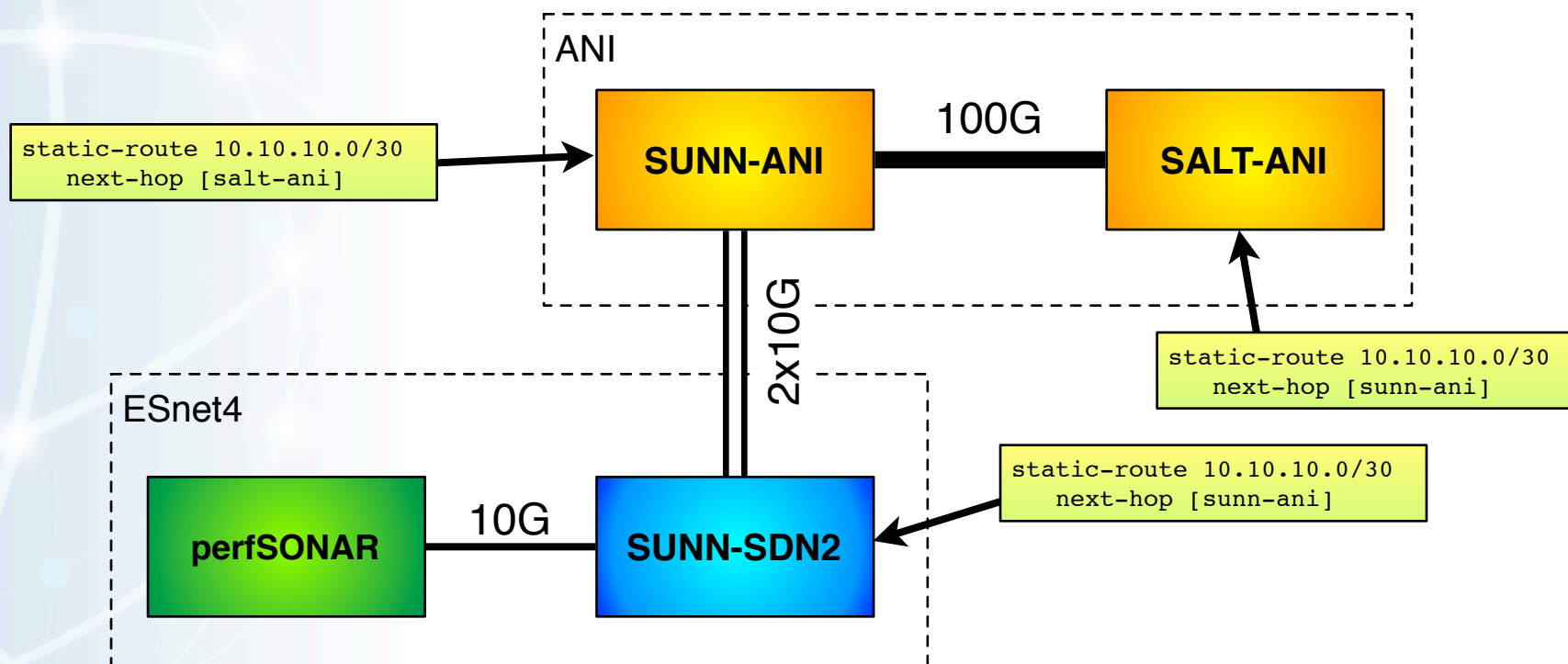
Could we leverage the 10Gbps systems we already have deployed?



Testing and Measurement - UDP flows

Could we leverage the 10Gbps systems we already have deployed?

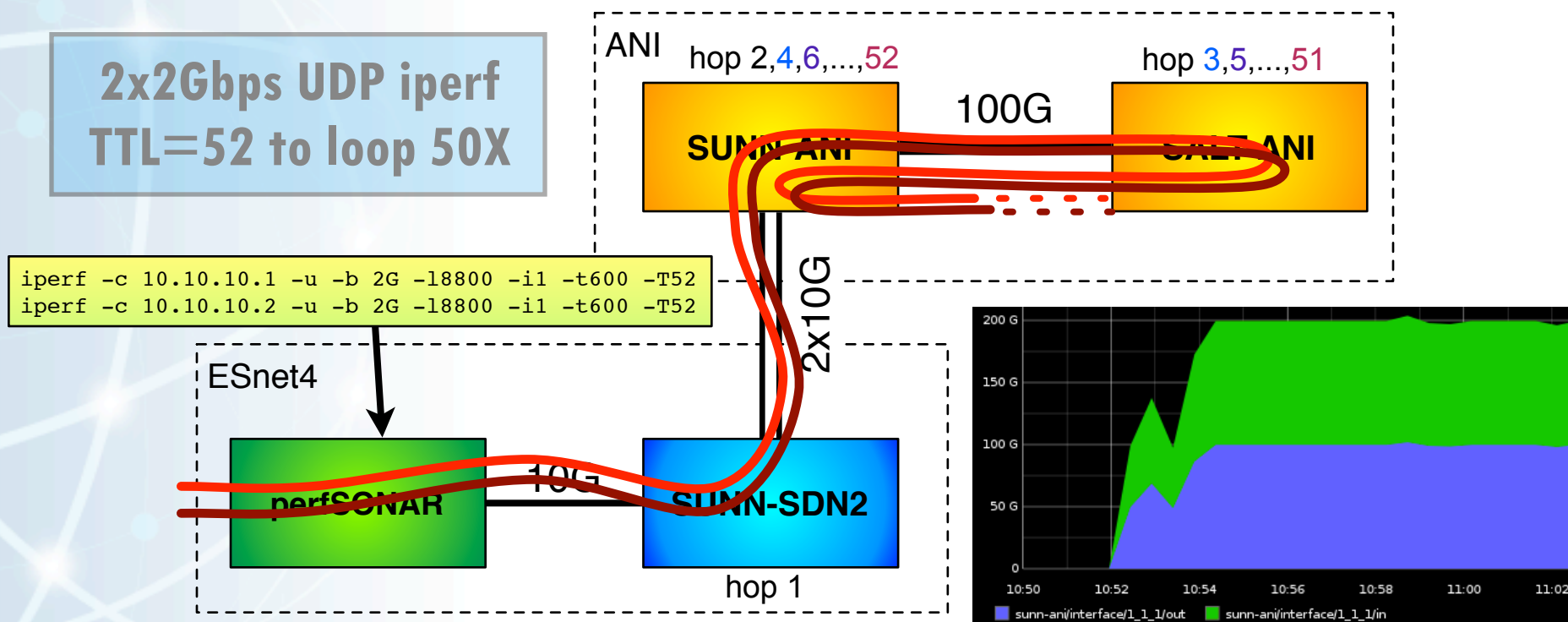
- let's have some fun with routing loops



Testing and Measurement - UDP flows

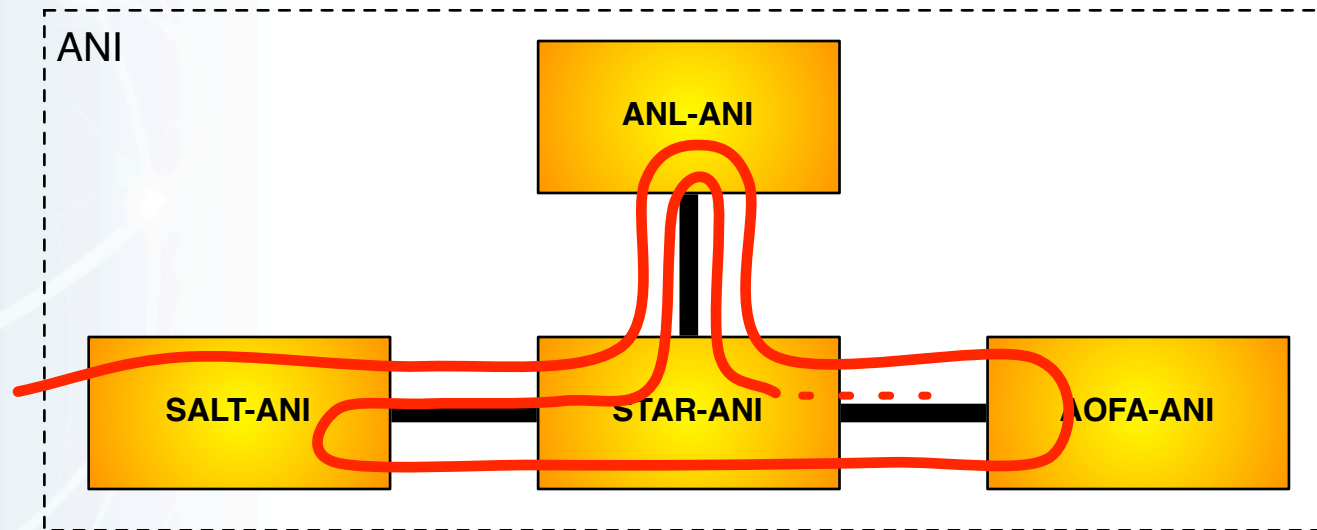
Could we leverage the 10Gbps systems we already have deployed?

- let's have some fun with routing loops and carefully chosen TTLs



Testing and Measurement - UDP flows

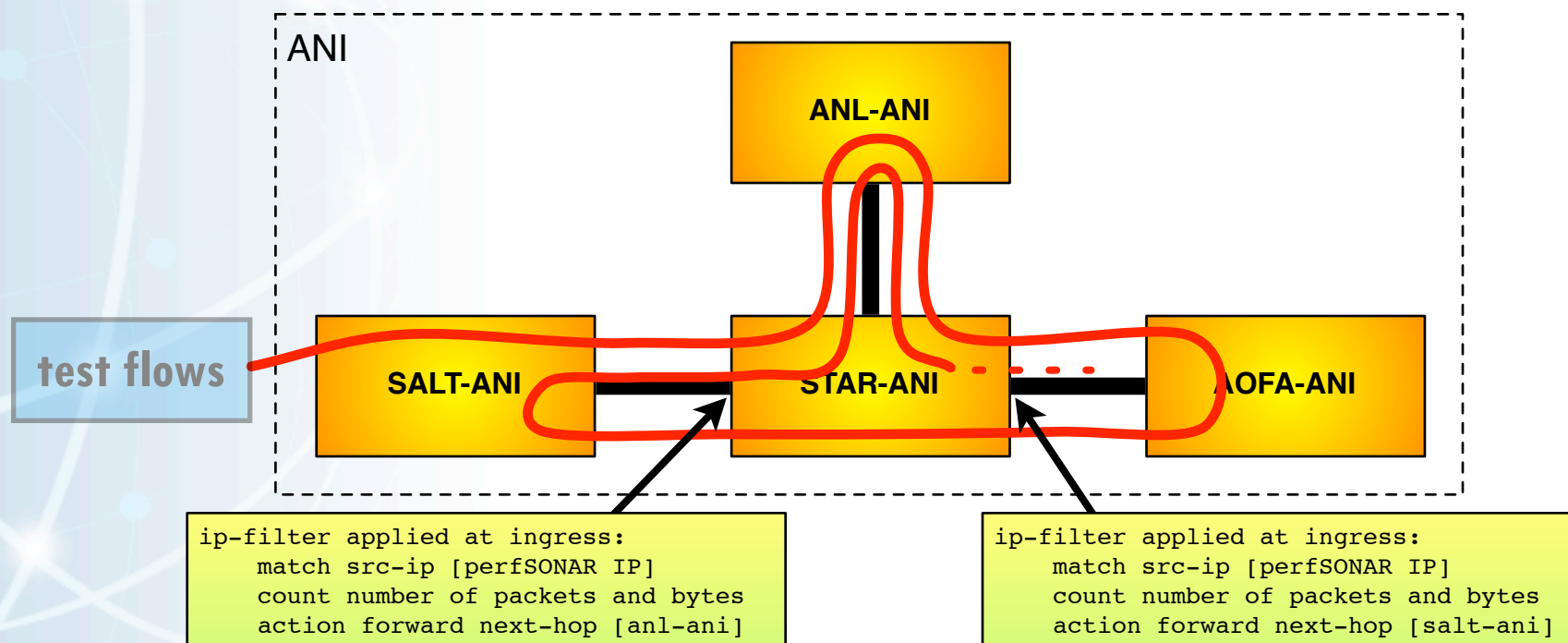
This was a quick solution to saturate the link, can we improve it?



Testing and Measurement - UDP flows

This was a quick solution to saturate the link, can we improve it?

- use policy-based routing for more complex loops - still testing UDP
- firewall ACLs for counting packets / bytes to measure loss

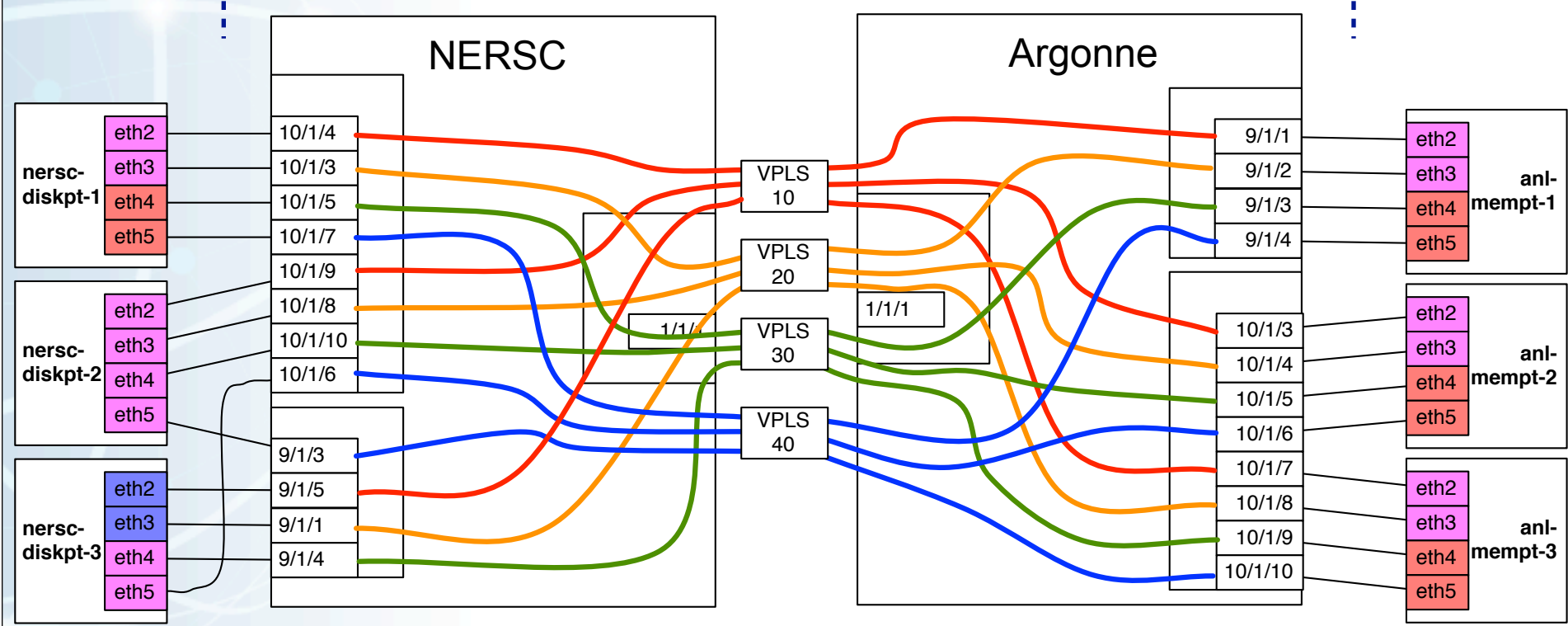




Testing and Measurement - TCP flows

Deployed systems on ANI experimental testbed - source/sink 4x10Gbps

48.6ms RTT, 97.9Gbps aggregate TCP throughput* with 10 TCP streams



01/11/11

* thanks to Eric Pouyol, Brian Tierney, and others

19

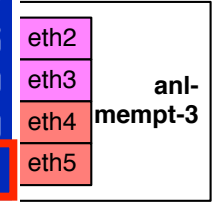
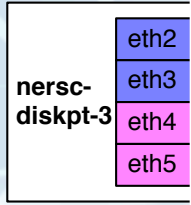
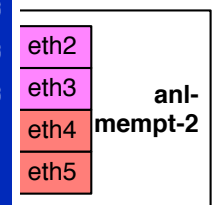
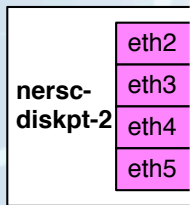
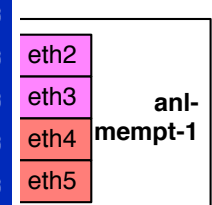
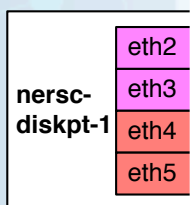


Testing and Measurement - TCP flows

Deployed systems on ANI experimental testbed - source/sink 4x10Gbps

48.6ms RTT, 97.9Gbps aggregate TCP throughput* with 10 TCP streams

nersc-diskpt-1-v4012:	1179.1875 MB /	1.00 sec =	9891.8010 Mbps	0 retrans
nersc-diskpt-1-v4013:	1179.2500 MB /	1.00 sec =	9888.4787 Mbps	0 retrans
nersc-diskpt-1-v4014:	1179.1875 MB /	1.00 sec =	9891.1482 Mbps	0 retrans
nersc-diskpt-1-v4015:	1179.1250 MB /	1.00 sec =	9891.1581 Mbps	0 retrans
nersc-diskpt-2-v4012:	1179.2500 MB /	1.00 sec =	9891.9494 Mbps	0 retrans
nersc-diskpt-2-v4013:	1179.0625 MB /	1.00 sec =	9891.1580 Mbps	0 retrans
nersc-diskpt-2-v4014:	1179.3750 MB /	1.00 sec =	9893.1365 Mbps	0 retrans
nersc-diskpt-2-v4015:	1179.1250 MB /	1.00 sec =	9891.0690 Mbps	0 retrans
nersc-diskpt-3-v4014:	1121.8750 MB /	1.00 sec =	9410.9602 Mbps	0 retrans
nersc-diskpt-3-v4015:	1121.8750 MB /	1.00 sec =	9410.9884 Mbps	0 retrans



	Input	Output
Octets	18462079	12387383345
Packets	184615	1369129
Errors	0	0
Utilization (% of port capacity)	0.17	99.31



Testing and Measurement

	Advantages	Disadvantages
Hardware Testers	<ul style="list-style-type: none">• stable & robust, tests many protocols• 40/100Gbps interface speeds• test single streams > 10Gbps• accurate measurement of loss/reordering• very precise control of packet content• can source/sink high bandwidth flows consisting of small packets	<ul style="list-style-type: none">• TCP implementation <i>may</i> not support any congestion control algorithm, not a good indicator of how an end host would perform• generally cannot run user applications that interface to the test equipment• relatively expensive
PC-based Testers	<ul style="list-style-type: none">• run user applications (data transfer such as GridFTP, bbFTP, scp)• real-world TCP implementation<ul style="list-style-type: none">- pluggable TCP congestion control algo.• choice of various measurement and analysis applications• depending on exact configuration, packet capture capabilities	<ul style="list-style-type: none">• at 100G, requires careful build and tuning• host issues due to NIC driver, kernel, interfering user/system processes• line-rate flows will have a limit on smallest packet size• possibility of bugs in measurement applications• accurate measurement problematic<ul style="list-style-type: none">- implement counting on hardware



Debugging and Fault Isolation

Tracking down loss

- Relatively easy to track down if there are accurate counters
 - and the ability to filter on some particular test traffic (e.g. 5-tuple)
- Difficult when there is background traffic and equipment does not allow packet counting via ACLs, etc.
- Important to understand where *all* of the “Drop” counters are

Tracking down re-ordering

- More difficult (especially if you don't have a hardware tester)
- Some open-source test tools can report on sequence errors or mis-orders, but there is room for improvement
- Can capture TCP packet dumps, analyze with tcptrace



Debugging and Fault Isolation

Using loopbacks

- Loopbacks can be very useful in certain situations
- Bring up facility loops facing a layer-3 router on transport equipment
 - ping broadcast address of layer-3 interface
 - check for obvious signs of trouble
 - helpful to verify patching
- If wide-area link is down, bring up loops at termination points to verify patching between layer-3 and transport gear
 - Place loops at re-gen points, check for link up/down on router



Interoperability

Between Layer-2/3 Gear and Optical Transport Equipment

- no issues have arisen
- transport layer is expected to be transparent, so far, it has been

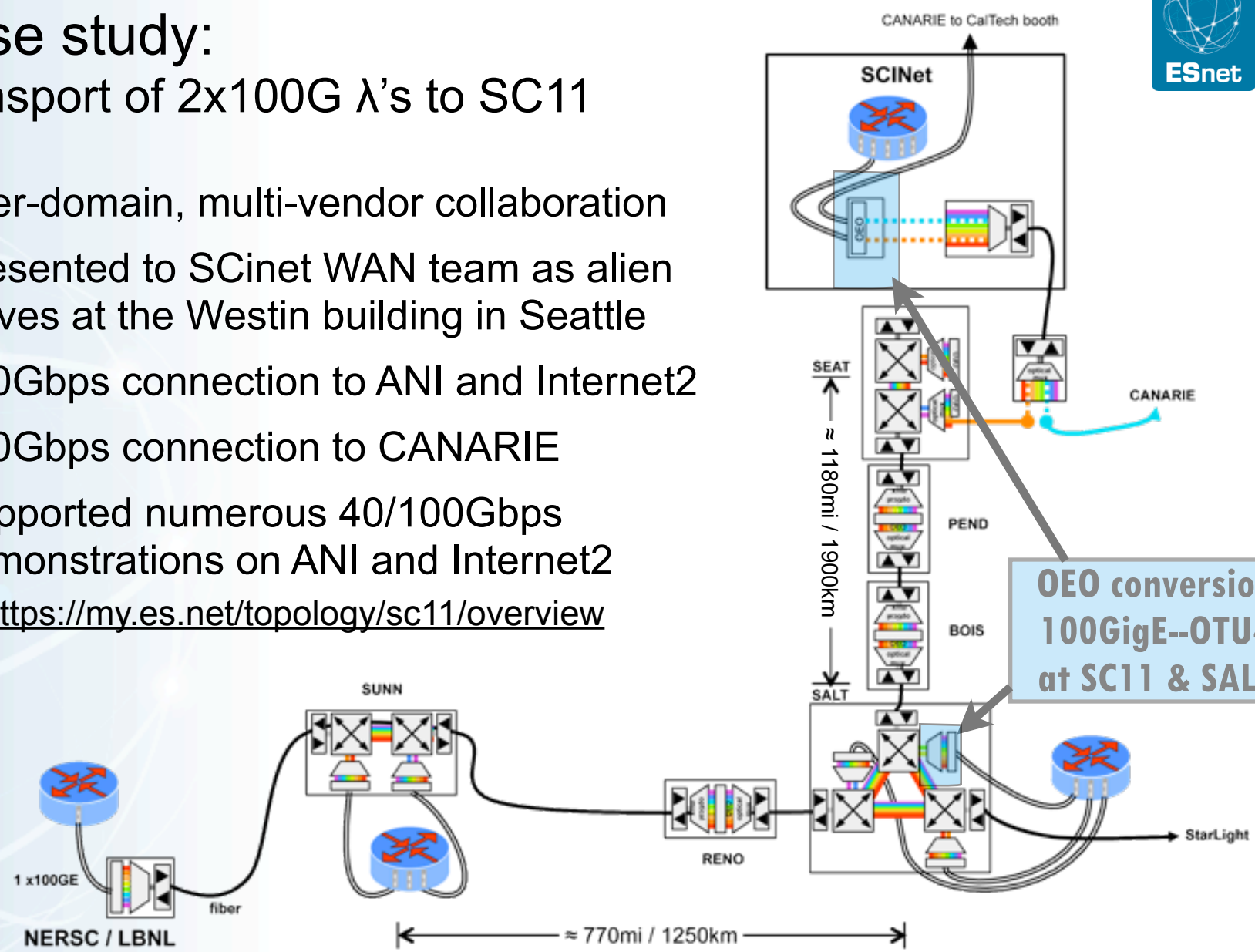
Between Layer-2/3 Equipment and other Layer-2/3 Equipment

- Remember, 100GigE is still very new
- Vendors have chosen slightly different ways of supporting 100G
 - not necessarily optimizing for one single line-rate flow
 - in some cases these have led to challenges in achieving desired performance
 - re-working the way the system is logically configured or physically cabled can potentially have a huge impact
- It is very important to understand exactly how high-bandwidth flows transit your equipment



Case study: Transport of 2x100G λ 's to SC11

- Inter-domain, multi-vendor collaboration
- Presented to SCinet WAN team as alien waves at the Westin building in Seattle
- 100Gbps connection to ANI and Internet2
- 100Gbps connection to CANARIE
- Supported numerous 40/100Gbps demonstrations on ANI and Internet2
 - <https://my.es.net/topology/sc11/overview>

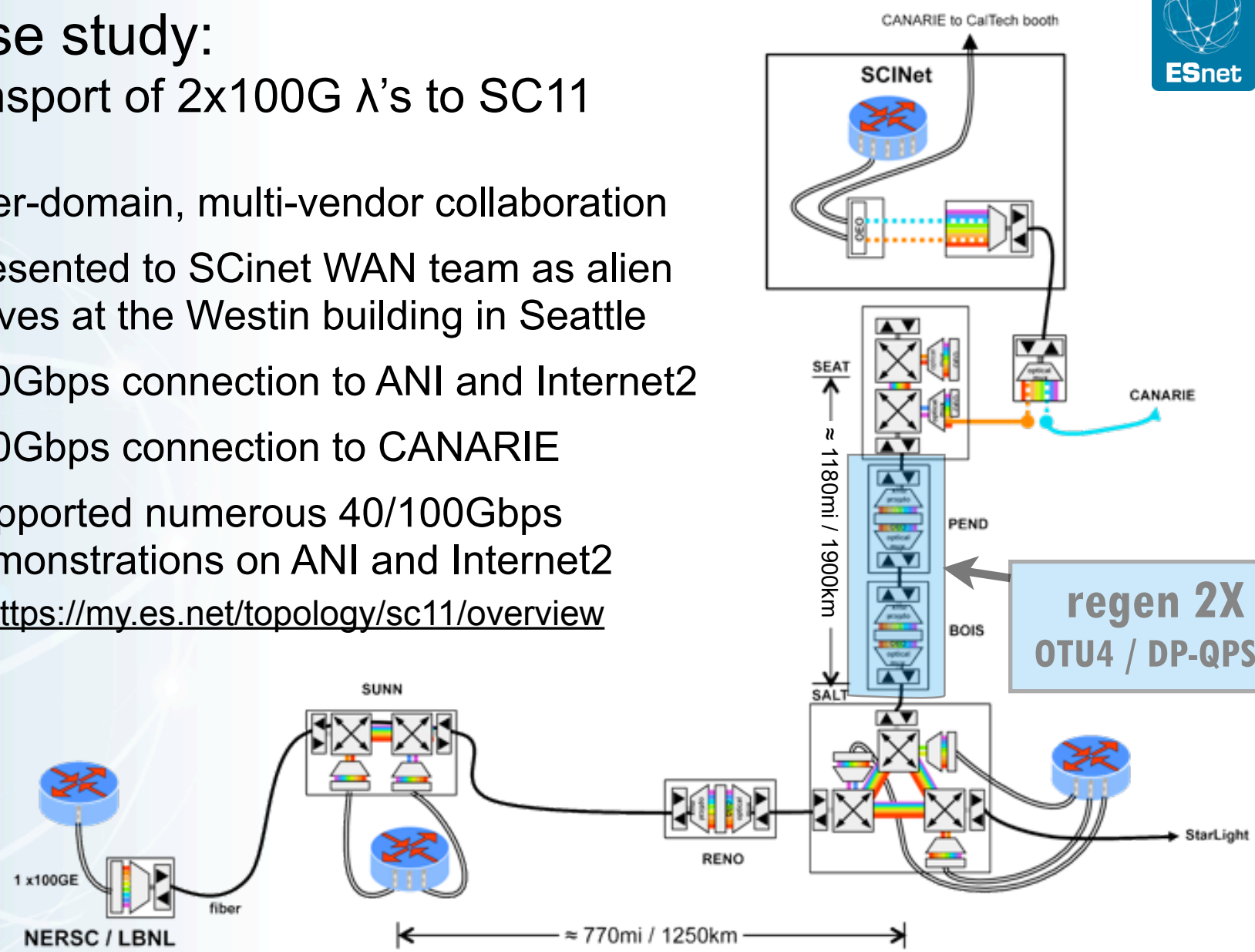


01/11/11



Case study: Transport of 2x100G λ 's to SC11

- Inter-domain, multi-vendor collaboration
- Presented to SCinet WAN team as alien waves at the Westin building in Seattle
- 100Gbps connection to ANI and Internet2
- 100Gbps connection to CANARIE
- Supported numerous 40/100Gbps demonstrations on ANI and Internet2
 - <https://my.es.net/topology/sc11/overview>



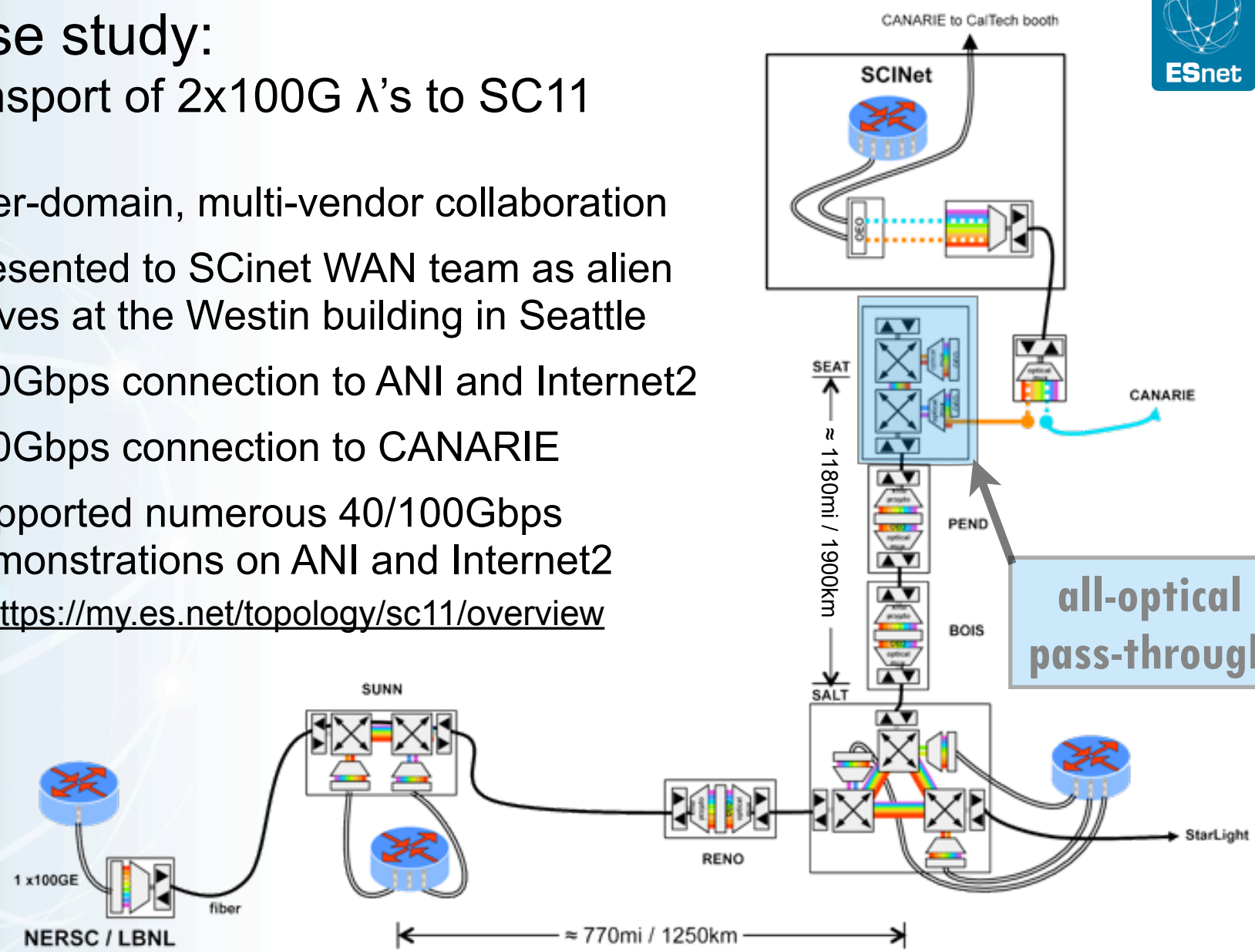
01/11/11

24

Case study:

Transport of 2x100G λ 's to SC11

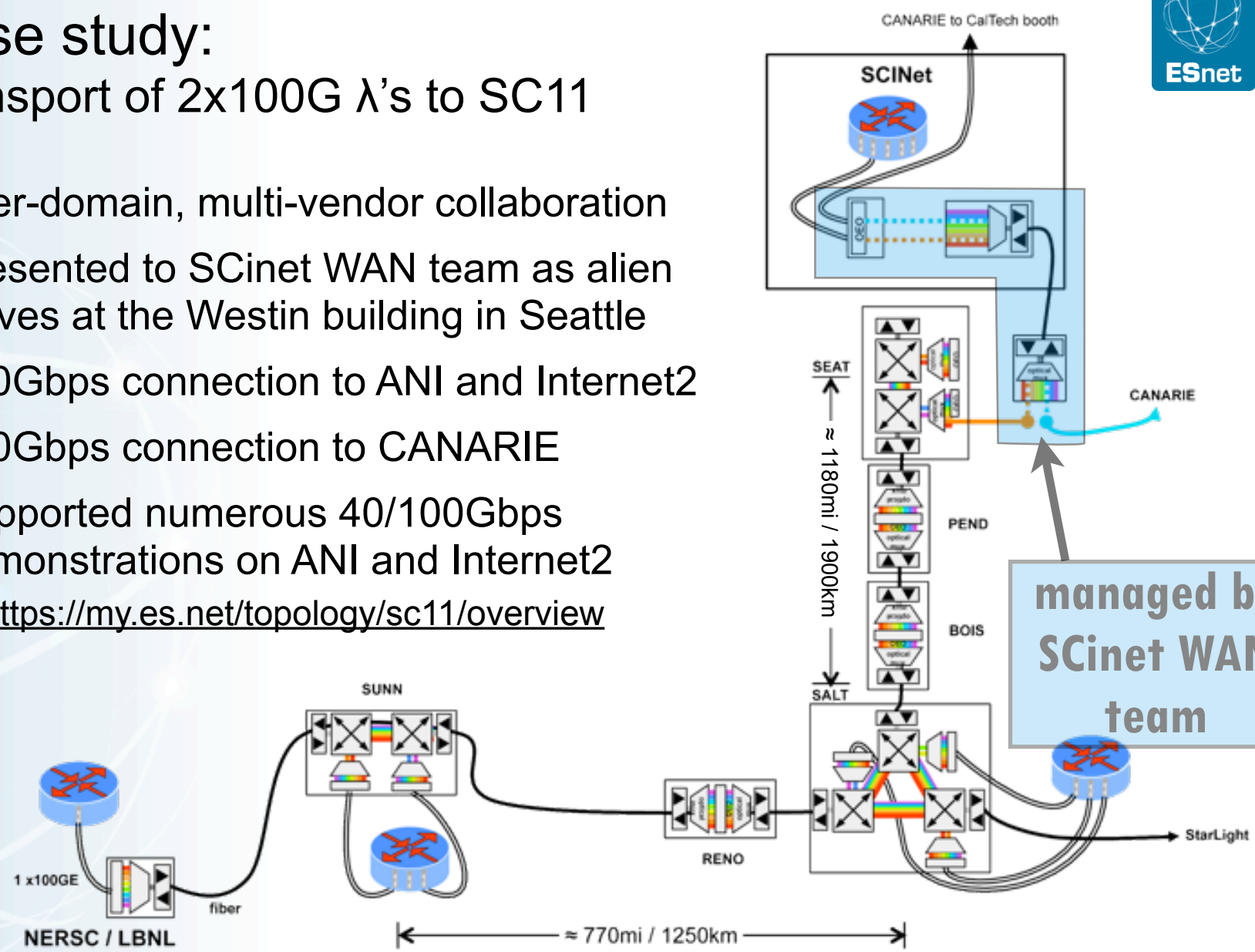
- Inter-domain, multi-vendor collaboration
- Presented to SCinet WAN team as alien waves at the Westin building in Seattle
- 100Gbps connection to ANI and Internet2
- 100Gbps connection to CANARIE
- Supported numerous 40/100Gbps demonstrations on ANI and Internet2
 - <https://my.es.net/topology/sc11/overview>





Case study: Transport of 2x100G λ 's to SC11

- Inter-domain, multi-vendor collaboration
- Presented to SCinet WAN team as alien waves at the Westin building in Seattle
- 100Gbps connection to ANI and Internet2
- 100Gbps connection to CANARIE
- Supported numerous 40/100Gbps demonstrations on ANI and Internet2
 - <https://my.es.net/topology/sc11/overview>

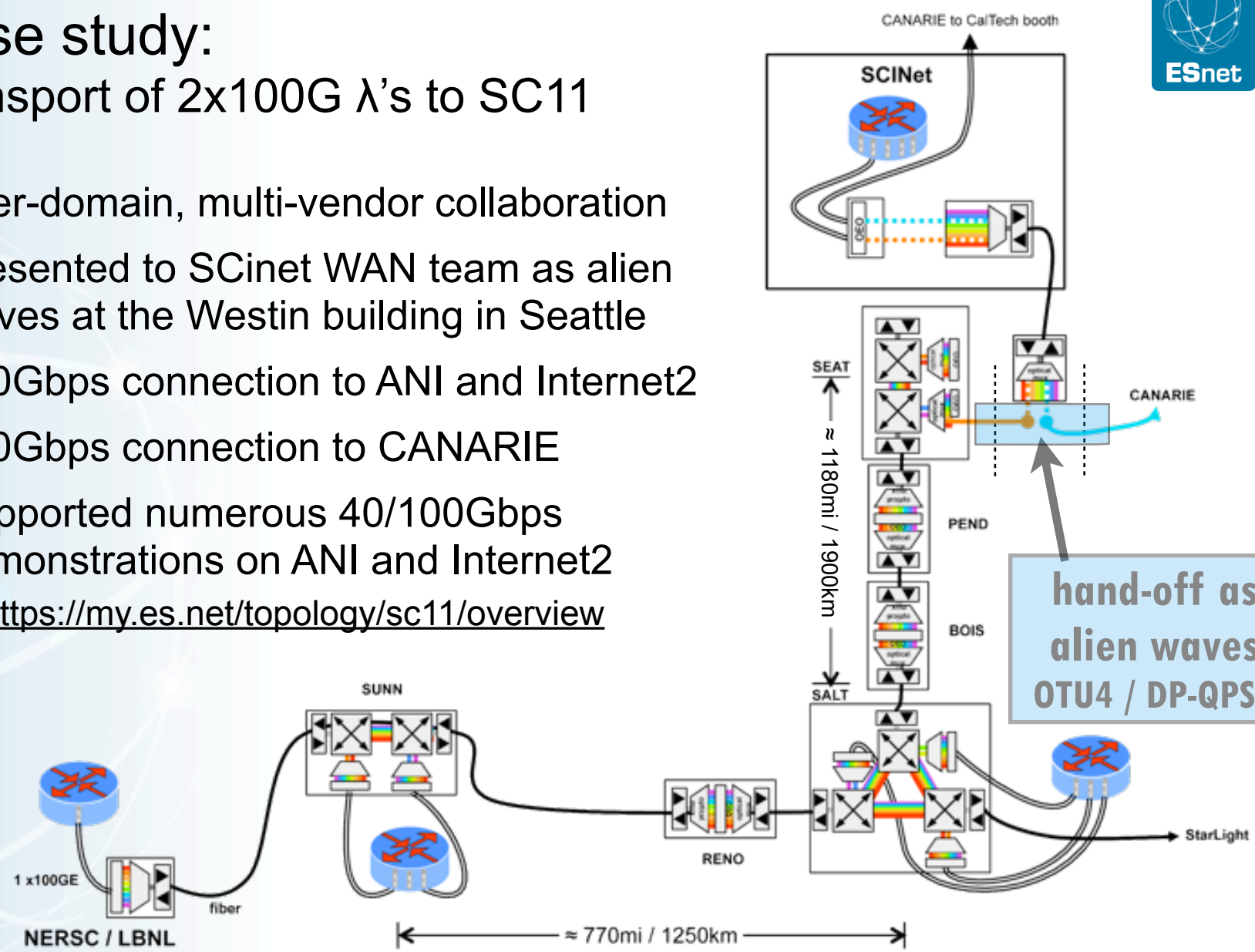


01/11/11



Case study: Transport of 2x100G λ 's to SC11

- Inter-domain, multi-vendor collaboration
- Presented to SCinet WAN team as alien waves at the Westin building in Seattle
- 100Gbps connection to ANI and Internet2
- 100Gbps connection to CANARIE
- Supported numerous 40/100Gbps demonstrations on ANI and Internet2
 - <https://my.es.net/topology/sc11/overview>



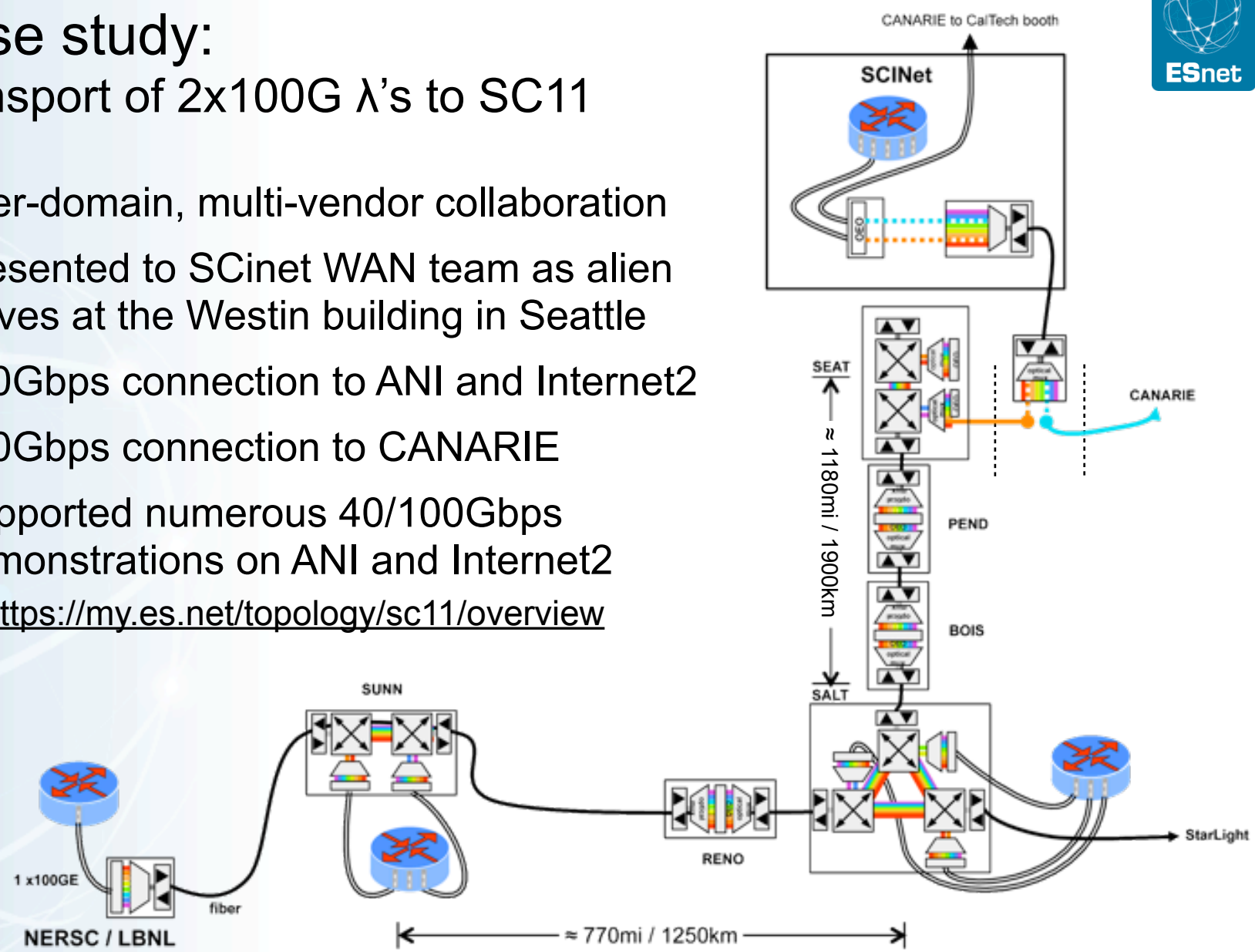
01/11/11

24



Case study: Transport of 2x100G λ 's to SC11

- Inter-domain, multi-vendor collaboration
- Presented to SCinet WAN team as alien waves at the Westin building in Seattle
- 100Gbps connection to ANI and Internet2
- 100Gbps connection to CANARIE
- Supported numerous 40/100Gbps demonstrations on ANI and Internet2
 - <https://my.es.net/topology/sc11/overview>



01/11/11

24



Future Work

Near-term: Mixed 10G NRZ with 40G/100G DP-QPSK

- These technologies do not necessarily play nice together [9]
- Kevin McGrattan's talk at Clemson [10] has a lot of detailed background on this topic
- Understand impact in terms of:
 - reach penalty on 100G channels
 - colorless ROADM deployment (with non-coherent wavelengths)

Deployment of Directionless and Colorless ROADM components

Long-term: Flexible Grid?

- Flexible wavelength selectable components
- Super channels at 400Gb/s or 1Tb/s



References

- [1] Roberts, K., Beckett, D., Boertjes, D., Berthold, J., Laperle, C. (2010, July). 100G and Beyond with Digital Coherent Signal Processing. *Communications Magazine, IEEE*, 48(7), 62-69. Retrieved January 16, 2011, from IEEE Xplore database.
- [2] IEEE P802.3ba. <http://standards.ieee.org/getieee802/download/802.3ba-2010.pdf>
- [3] ITU-T G.709. <http://www.itu.int/rec/T-REC-G.709/en>
- [4] OIF-FD-100G-DWDM-01.0 - 100G Ultra Long Haul DWDM Framework Document (June 2009). <http://www.oiforum.com/public/documents/OIF-FD-100G-DWDM-01.0.pdf>
- [5] 10x10 MSA Revision 2.4. <http://www.10x10msa.org/documents/MSA%20Technical%20Rev2-4.pdf>
- [6] Finisar C.wire. <http://finisar.com/products/active-cables/C.wire>
- [7] Digital Coherent Receiver Technology for 100-Gb/s Optical Transport Systems. <http://www.fujitsu.com/downloads/MAG/vol46-1/paper18.pdf>
- [8] Gringeri, S., Basch, B., Shukla, V., Egorov, R., Xia, T. (2010, July). Flexible Architectures for Optical Transport Nodes and Networks. *Communications Magazine, IEEE*, 48(7), 40-50. Retrieved January 16, 2011, from IEEE Xplore database.
- [9] Magill, P. (2010, September). 100G Coherent trials and deployments: AT&T plans and perspective. *ECOC2010, 36th European Conference and Exhibition on Optical Communication*. Retrieved January 16, 2011, from IEEE Xplore database.
- [10] McGrattan, K. Considerations in Migrating DWDM Networks to 100G. <http://events.internet2.edu/2011/jt-clemson/agenda.cfm?go=session&id=10001587>



Questions?

Thanks!