

Some ESnet Observations on Using and Managing OSCARS Point-to-Point Circuits

LHCONE / LHCOPN meeting, May 3-4, 2012, KTH, Stockholm

William Johnston, et al
Energy Sciences Network (ESnet)



- OSCARS was introduced into ESnet as a proto-production service in early 2007, by mid-late 2008 was a supported production service
- The DICE collaboration (DANTE, Internet2, ESnet) had a prototype of the inter-domain control protocol (IDCP) working by late 2008 – early 2009
- The service has managed all LHCOPN Tier 0 – Tier 1 traffic since mid-2008 in ESnet, most T2-T1 traffic since 2009
- A lot of what OSCARS is about from ESnet's point of view is capacity management
- A lot of what OSCARS is about from the user's point of view is capacity guarantees

OSCARS Virtual Circuit Service Goals



The general goal of OSCARS is to

- Allow users to request guaranteed bandwidth between specific end points for specific period of time
 - User request is via Web Services interface (for programs) or a Web browser interface (for users)
 - The assigned end-to-end path through the network is called a virtual circuit (VC)

Goals that have arisen through user experience with OSCARS include:

➤ Flexible service semantics

- e.g. allow a user to exceed the requested bandwidth, if the path has idle capacity – even if that capacity is committed (but unused)
- Rich service semantics
 - E.g. provide for several variants of requesting a circuit with a backup, the most stringent of which is a guaranteed backup circuit on a physically diverse path

The environment of large-scale science is inherently multi-domain

- OSCARS must interoperate with similar services in other network domains in order to set up cross-domain, end-to-end virtual circuits
 - In this context OSCARS is an InterDomain [virtual circuit service] Controller (“IDC”)

- Basic
 - User requests VC b/w, start time, duration, and endpoints
 - The endpoint for L3VC is the source and dest. IP addr.
 - The endpoint for L2VC is the domain:node:port:link - e.g. “esnet:chi-sl-sdn1:xe-4/0/0:xe-4/0/0.2370” on a Juniper router, where “port” is the physical interface and “link” is the sub-interface where the VLAN tag is defined)
 - Explicit, diverse (where possible) backup paths may be requested
 - This doubles the b/w request
- VCs are rate-limited to the b/w requested, but are permitted to burst above the allocated b/w if unused bandwidth is available on the path
- Currently the VC, in-allocation packet priority is set to high, out-of-allocation (burst) packet priority is set to low, this leaves a middle priority for non-OSCARS traffic (e.g. best effort IP)
 - In the future VC priorities and over allocated b/w packet priorities will be settable
- In combination, these semantics turn out to provide powerful capabilities

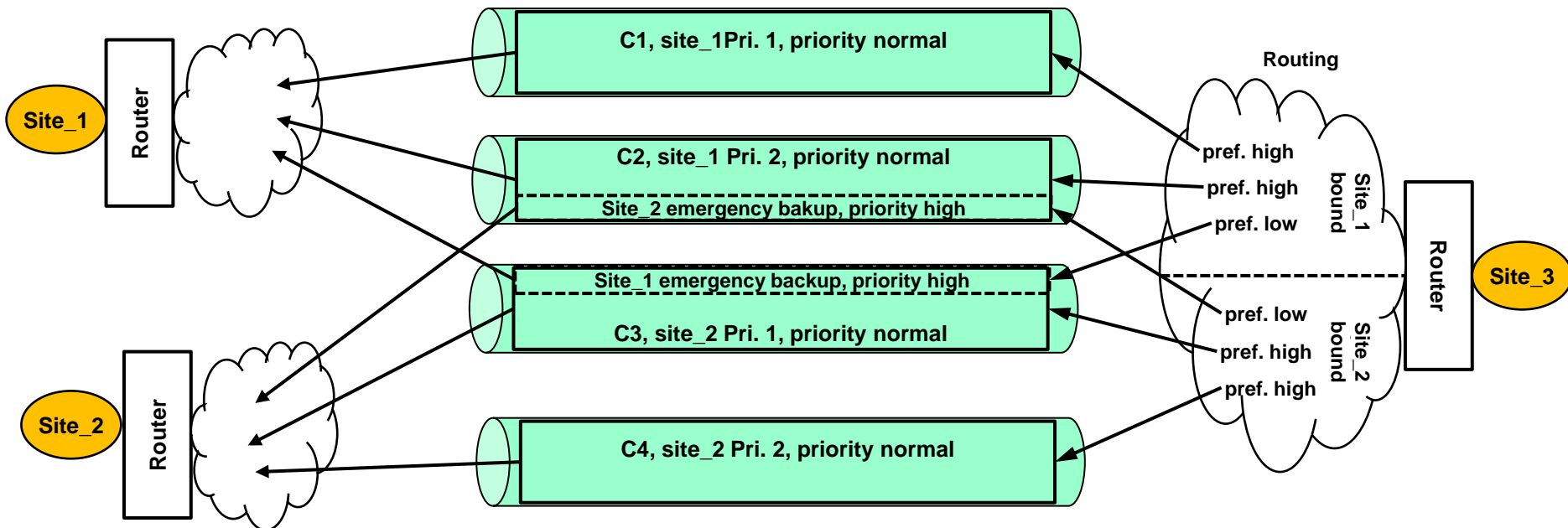
OSCARS Example

Hypothetical re-purposing strategy for backup in the event of both primary circuits any one T1 down at the same time

Six OSCARS circuits on four physical paths, two primary for each of site_1 and site_2, and one secondary for each site

The secondary / emergency circuits get an agreed upon bandwidth by virtue of using an elevated priority

- These circuits are managed by routing instances at the ends and are only used if both primary circuits to the site are down
- When not in use, the primary traffic gets the full circuit bandwidth



End User View of Circuits – How They Use Them

- Who are the “users?”
 - Sites, for the most part
- How are the circuits used?
 - End system to end system, IP
 - Almost never – very hard unless private address space used
 - » Using public address space can result in leaking routes
 - » Using private address space with multi-homed hosts risks allowing backdoors into secure networks
 - End system to end system, x over VLAN
 - Relatively common
 - Interesting example: RDMA over VLAN likely to be popular in the future
 - » SC11 demo of 40G RDMA over WAN was very successful
 - » CPU load for RDMA is a small fraction that of IP
 - » The guaranteed network of circuits (zero loss, no reordering, etc.) required by non-IP protocols like RDMA fits nicely with circuit services (RDMA performs very poorly on best effort networks)
 - Point-to-point connection between routing instance – e.g. BGP at the end points
 - Essentially this is how all current circuits are used: from one site router to another site router
 - » Typically site-to-site or advertise subnets that host clusters, e.g., LHC analysis or data management clusters

Currently active OSCARS circuits



Name	Description	Capacity*			Capacity*
es.net-789	FNAL - DUKE, VLAN 633	1.00G	es.net-1852	FNAL - USLHCNET backup, VLAN 3501	3.00G
es.net-790	FNAL - ASGC, VLAN 3120	1.00G	es.net-1854	BNL-USLHCNET, VLAN 3524	8.50G
es.net-792	FNAL - UTK, VLAN 3140	1.00G	es.net-1875	FNAL - UNL, VLAN 3550	1.00G
es.net-809	FNAL - USLHCNET, VLAN 3500	8.50G	es.net-1910	Terapaths Development	1.00G
es.net-817	FNAL - MIT, VLAN 3001	1.00G	es.net-1944	JGI-NERSC C, VLAN 2	2.90G
es.net-819	FNAL - IN2P3, VLAN 3111	1.00G	es.net-1945	JGI-NERSC B, VLAN 10	100M
es.net-842	FNAL - PURDUE, VLAN 3113	1.00G	es.net-1999	JGI-NERSC A, VLAN 96	3.00G
es.net-843	FNAL - ULTRALIGHT, VLAN 3101	1.00G	es.net-2146	FNAL - UMN Soudan primary, VLAN 3130	500M
es.net-902	BNL - BU (VLAN 3001)	1.00G	es.net-2430	LBL - SDSC, VLAN 41 - 3811, 1 Gbps	1.00G
es.net-913	BNL - TRIUMF (VLAN 2608)	1.00G	es.net-2641	CESNET	1.00G
es.net-914	BNL - USLHCNET backup, VLAN 3514	3.00G	es.net-2642	CERN BNL Primary	8.50G
es.net-1810	FNAL - USLHCNET, VLAN 3506	8.50G	es.net-2707	BNL ATLAS UOC	1.00G
es.net-1812	FNAL - SARA - DEKIT, VLAN 2613	1.00G	es.net-2708	BNL ATLAS SLAC	1.00G
es.net-1814	FNAL - TIFR, VLAN 2499	1.00G	es.net-2804	FNAL - UMN NOvA, VLAN 3140	500M
es.net-1817	FNAL - UFL, VLAN 1302	1.00G	es.net-2892	Circuit between GPN and NASA/ARC for ARC to EROS connectivity	200M
es.net-1842	FNAL - UCSD, VLAN 3503	1.00G	es.net-2904	Circuit for L2 connectivity between PacketDesign and ANI r	100M
es.net-1849	BNL-AGLT2 (UMich / Mich State), VLAN 3102	1.00G	es.net-2930	ANI 100G Testbed Researcher Access	100M

**NB: The "capacity" reflects a lower bound guarantee since OSCARS semantics allow bursting over the requested bandwidth if capacity is available*

Spectrum Network Monitor System Monitors OSCARS Circuits

Navigation

Explorer Locater Users

Name ▲

- My SPECTRUM 3 1 2
 - Global Collections
 - Global Collection Hierarchy
 - Configuration Manager (3)
 - eHealth Manager (1)
 - VPN Manager
 - sage (0x4000000) 3 1 2
 - Enterprise VPN Manager
 - Service Management (3)
 - TopOrg
 - Universe (6) 3 1 1
 - CHIC Hub (8)
 - CLEV Hub (2)
 - Multicast Pingables (169)
 - NEWY Hub (6)
 - SUNN Hub (11)
 - WASH Hub (7)
 - World
 - Correlation Manager
 - LostFound
 - MPLS Transport Manager (7)
 - anl-mr1 (1)
 - aofa-sdn1 (9)
 - bnl-mr1 (5)
 - chic-sdn2 (1)
 - fnal-mr1 (12)
 - star-cr1 (1)
 - OSCARS_ES_NET-638 (1)
 - OSCARS_ES_NET-638 ...
 - star-sdn1 (7)
 - Multicast Manager (24) 1
 - Policy Manager
 - QoS Manager
 - Remote Operations Manager
 - Secure Domain Manager
 - Telco EMS Manager

Contents: OSCARS_ES_NET-638 of type MplsPath


Alarms Topology List Events Information

Filter: Displaying 8 of 8

Condition	Name	Network Address	Secure Domain	Manufacturer	Model Class	MAC Address	Type	Landscape
Normal	aofa-cr2	134.55.200.100	Directly Managed	Juniper Netw...	Switch-Router	00:a0:a5:61:...	MX480	sage (0x4000000)
Normal	bnl-mr1	134.55.200.66	Directly Managed	Cisco	Switch-Router	00:13:5f:e1:...	Cat6509	sage (0x4000000)
Normal	newy-sdn1	134.55.200.30	Directly Managed	Juniper Netw...	Switch-Router	00:a0:a5:61:...	MX960	sage (0x4000000)
Normal	wash-sdn2	134.55.200.76	Directly Managed	Juniper Netw...	Switch-Router	00:a0:a5:61:...	MX960	sage (0x4000000)
Normal	clev-sdn1	134.55.200.54	Directly Managed	Juniper Netw...	Switch-Router	00:a0:a5:61:...	MX960	sage (0x4000000)
Normal	star-sdn1	134.55.200.96	Directly Managed	Juniper Netw...	Switch-Router	00:a0:a5:61:...	MX960	sage (0x4000000)
Normal	chic-sdn2	134.55.200.98	Directly Managed	Juniper Netw...	Switch-Router	00:a0:a5:61:...	MX960	sage (0x4000000)
Normal	star-cr1	134.55.200.95	Directly Managed	Juniper Netw...	Switch-Router	00:a0:a5:61:...	MX480	sage (0x4000000)

Component Detail: OSCARS_ES_NET-638 of type MplsPath

Information Host Configuration Root Cause Interfaces Performance Neighbors Alarms Events Attributes



OSCARS_ES_NET-638 [set](#)
MplsPath

General Information

Creation Time	Ingress Device
Condition	Egress Device
ID	Notes

Path Hops - OSCARS_ES_NET-638 of type MplsPath - SPECTRUM OneClick

File View Help

Filter: Displaying 8 of 8

Hop	Device Condition	Device	Device IP	Incoming IF Co...	Incoming IF	Outgoing IF Condition	Outgoing IF
1	Normal	star-cr1	134.55.200.95		star-sdn1_xe-1/0/0.0	Normal	star-cr1_xe-1/0/0.0
2	Normal	star-sdn1	134.55.200.96	Normal	star-sdn1_xe-1/0/0.0	Normal	star-sdn1_xe-8/0/0.0
3	Normal	chic-sdn2	134.55.200.98	Normal	chic-sdn2_xe-0/2/0.0	Normal	chic-sdn2_xe-7/0/0.0
4	Normal	clev-sdn1	134.55.200.54	Normal	clev-sdn1_xe-7/1/0.0	Normal	clev-sdn1_xe-1/2/0.0
5	Normal	wash-sdn2	134.55.200.76	Normal	wash-sdn2_xe-1/1/0.0	Normal	wash-sdn2_xe-2/0/0.0
6	Normal	newy-sdn1	134.55.200.30	Normal	newy-sdn1_xe-0/0/0.0	Normal	newy-sdn1_xe-2/1/0.0
7	Normal	aofa-cr2	134.55.200.100	Normal	aofa-cr2_xe-2/1/0.0	Normal	aofa-cr2_xe-2/0/0.0
8	Normal	bnl-mr1	134.55.200.66	Normal	bnl-mr1 Te2/1		

End User View of Circuits – How They Keep Track of Them



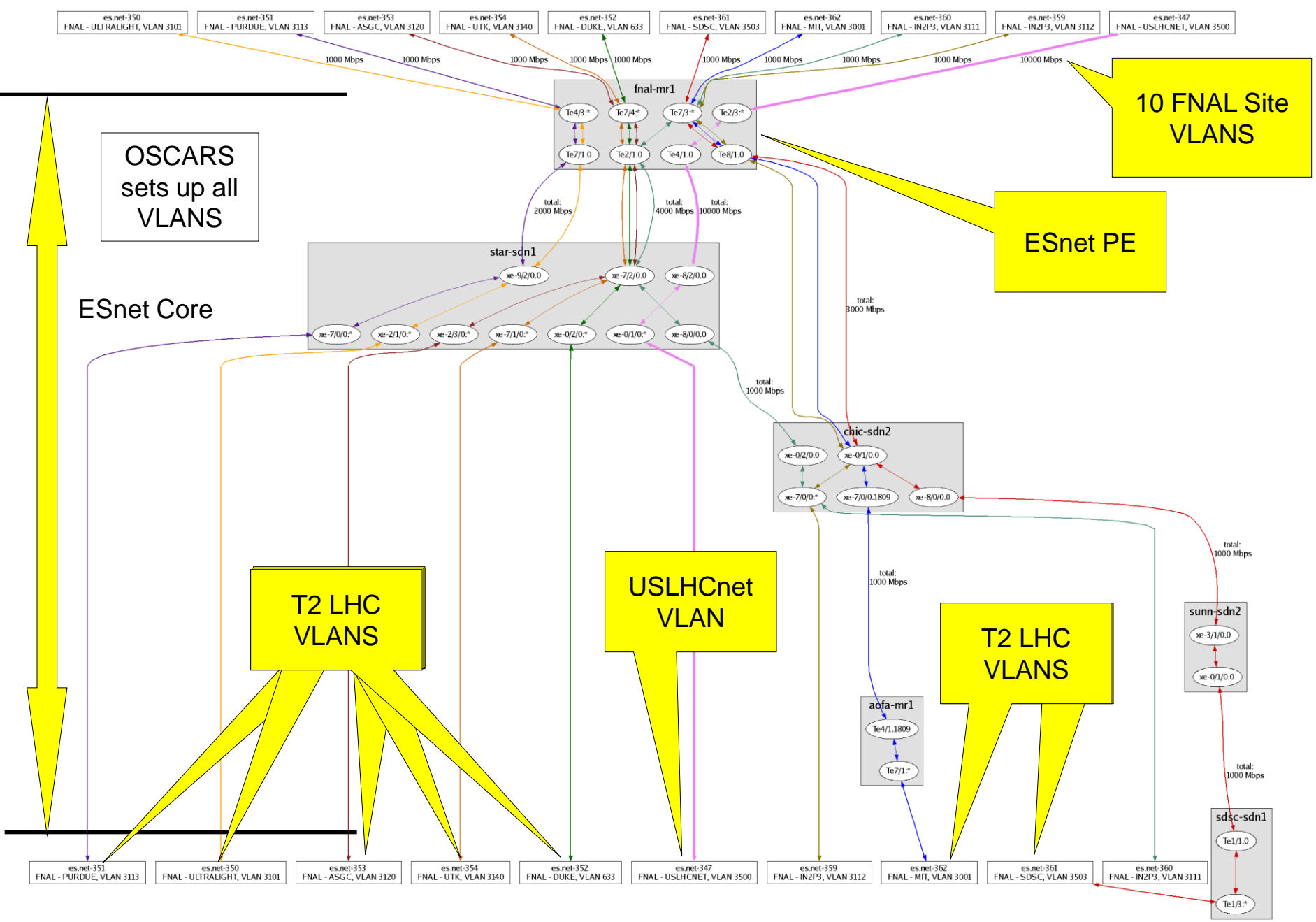
- Circuits are both ephemeral and long-lived
 - All of the circuits shown below are long-lived
 - Around SC conference time ,for example, lots of short term circuits are set up
- The circuits are requested through an OSCARS account that can be used by the end users, however since the “site-coordinator” (LAN admin type) has to build the connection from host to border router, the site coordinator typically make the circuit request to ESnet
 - Exceptions to this rule are site automated tools like LambdaStation and TeraPaths that create their own circuits
 - The site coordinator role has privileged access to OSCARS and can modify/terminate all circuits made to site regardless of who made the reservation

End User View of Circuits – How They Keep Track of Them



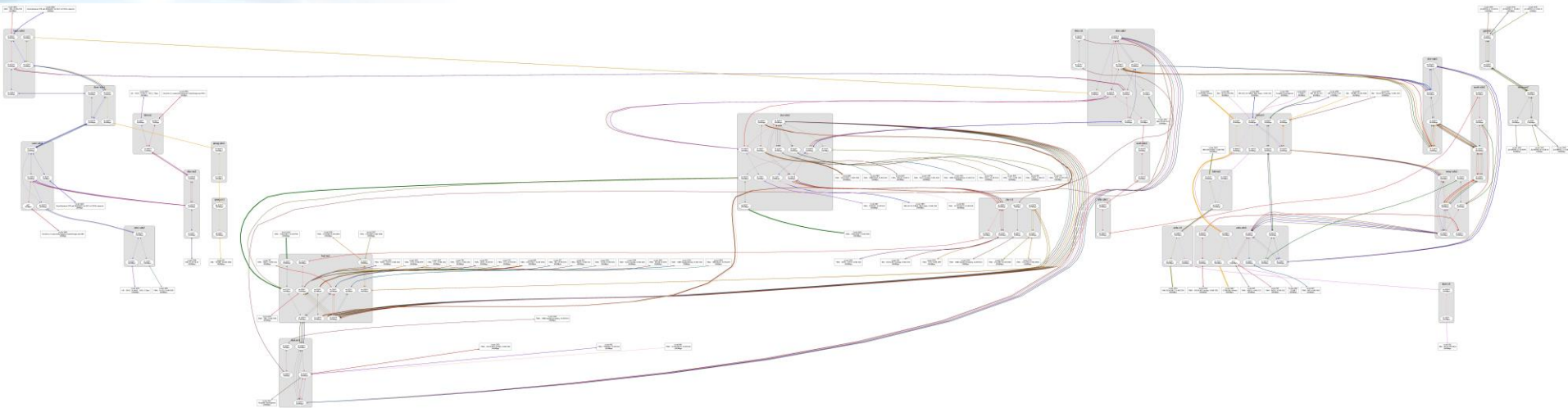
- Only ESnet has a consistent, global view of the circuit connectivity
 - This lead to building a tool that automatically generated circuit maps that shows
 - End points
 - Intermediate devices
 - While global maps are generated, most sites are interested in the circuits that connect to their site, and so site-specific maps are also generated
 - Maps are mostly useful on a workstation where they can easily “browsed”

OSCARS Managed External Circuit Topology at FNAL, June 2008



All OSCARS Circuits

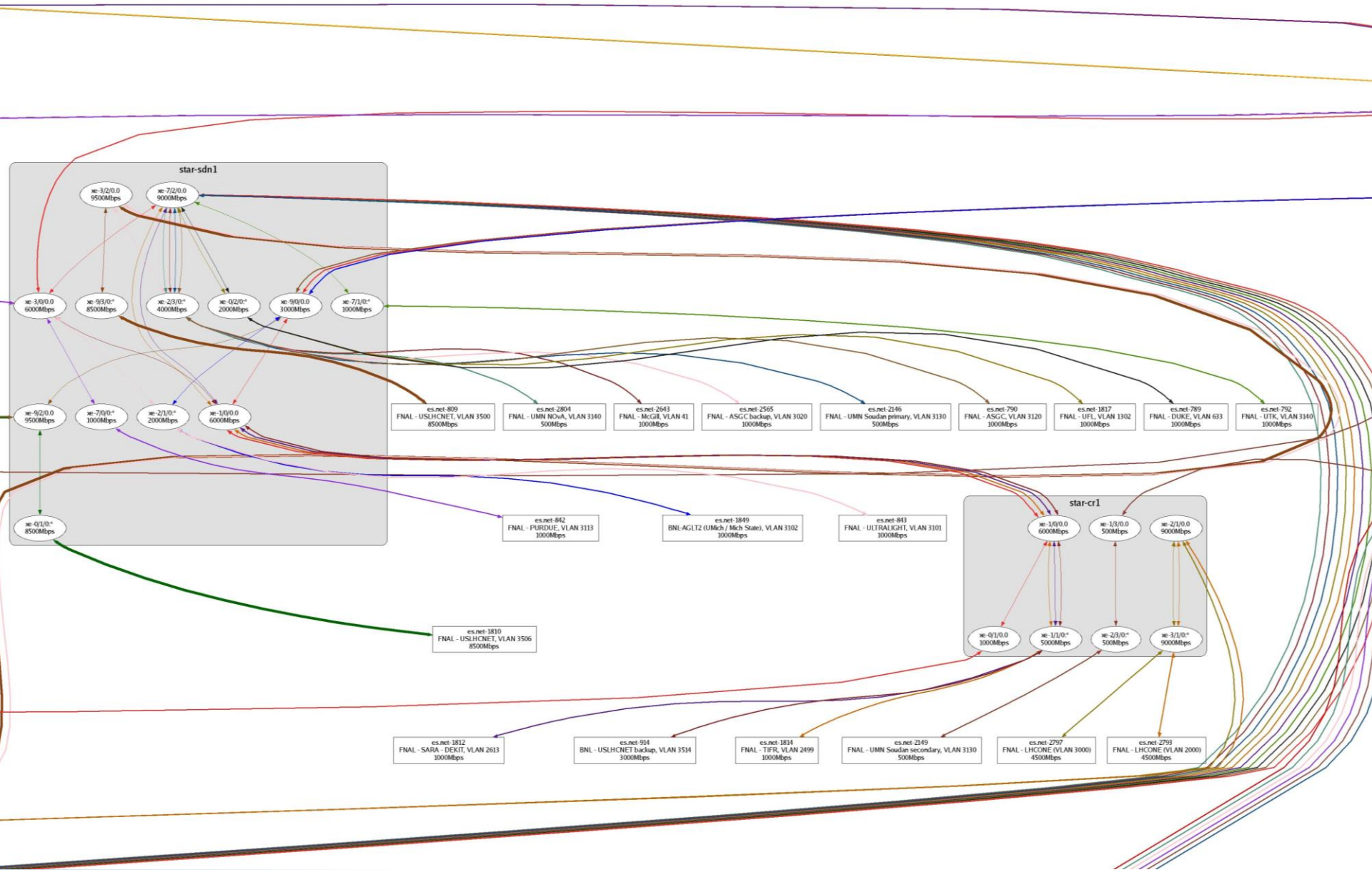
- Circuit maps are automatically generated for each site that has a circuit endpoint



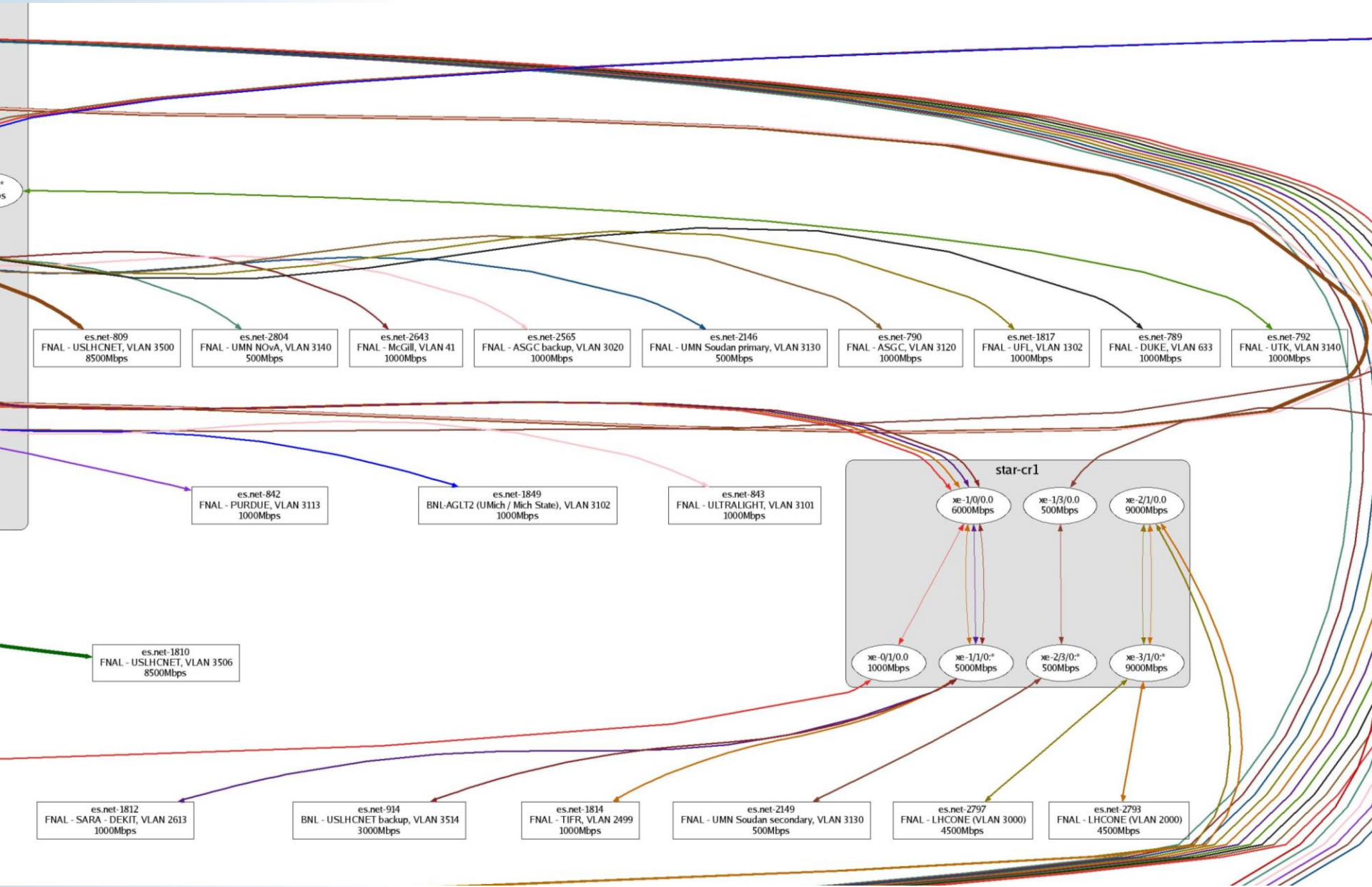
The circuits in and around Fermilab (U.S. CMS Tier1) Shows OPN circuits and Tier 2 circuits coming to Fermi



xe-3/1/0.0
500Mbps



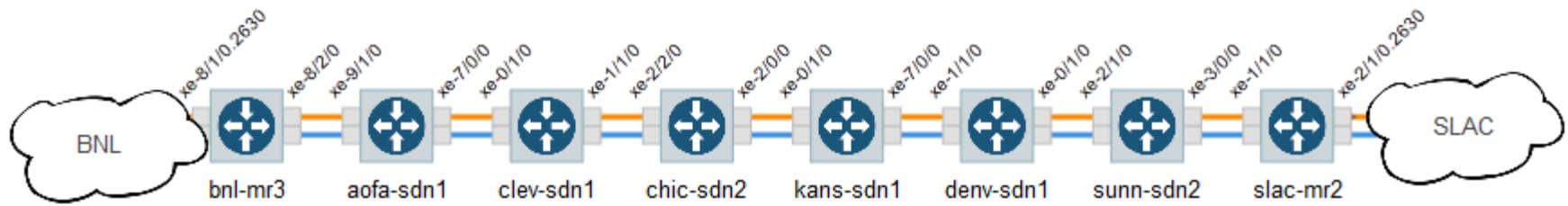
The circuits in and around Fermilab (U.S. CMS Tier1) Shows OPN circuits and Tier 2 circuits coming to Fermi



Monitoring is Done Automatically



es.net-2708 | BNL ATLAS SLAC | 10-13-2011 To 10-14-2020



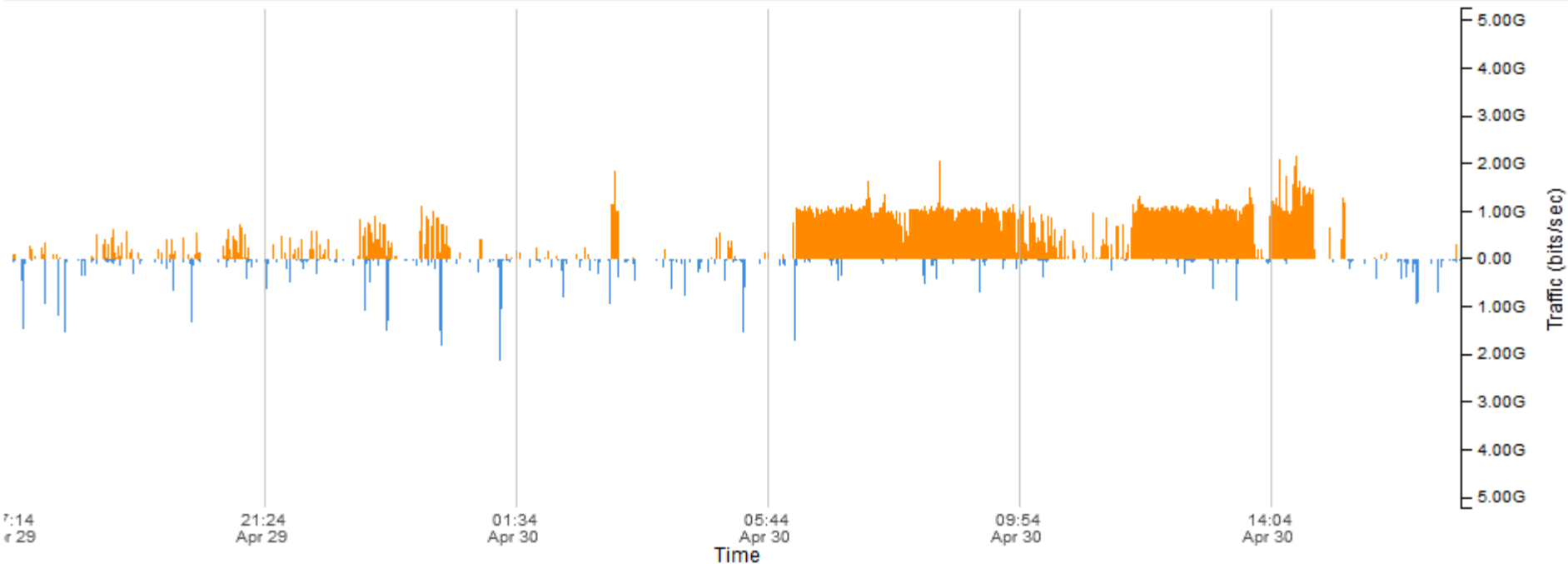
Circuit Traffic [Refresh](#)

1day [Last hour](#)

[A to Z Delivered](#) [Z to A Delivered](#)

Circuit Capacity 1.00G

[Reset zoom](#)



Monitoring is Done Automatically



es.net-1854 | BNL-USLHCNET, VLAN 3524 | 02-18-2010 To 02-17-2020



Circuit Traffic [Refresh](#)

1day [Last hour](#)

[A to Z Delivered](#) [Z to A Delivered](#)

Circuit Capacity 8.50G

[Reset zoom](#)

