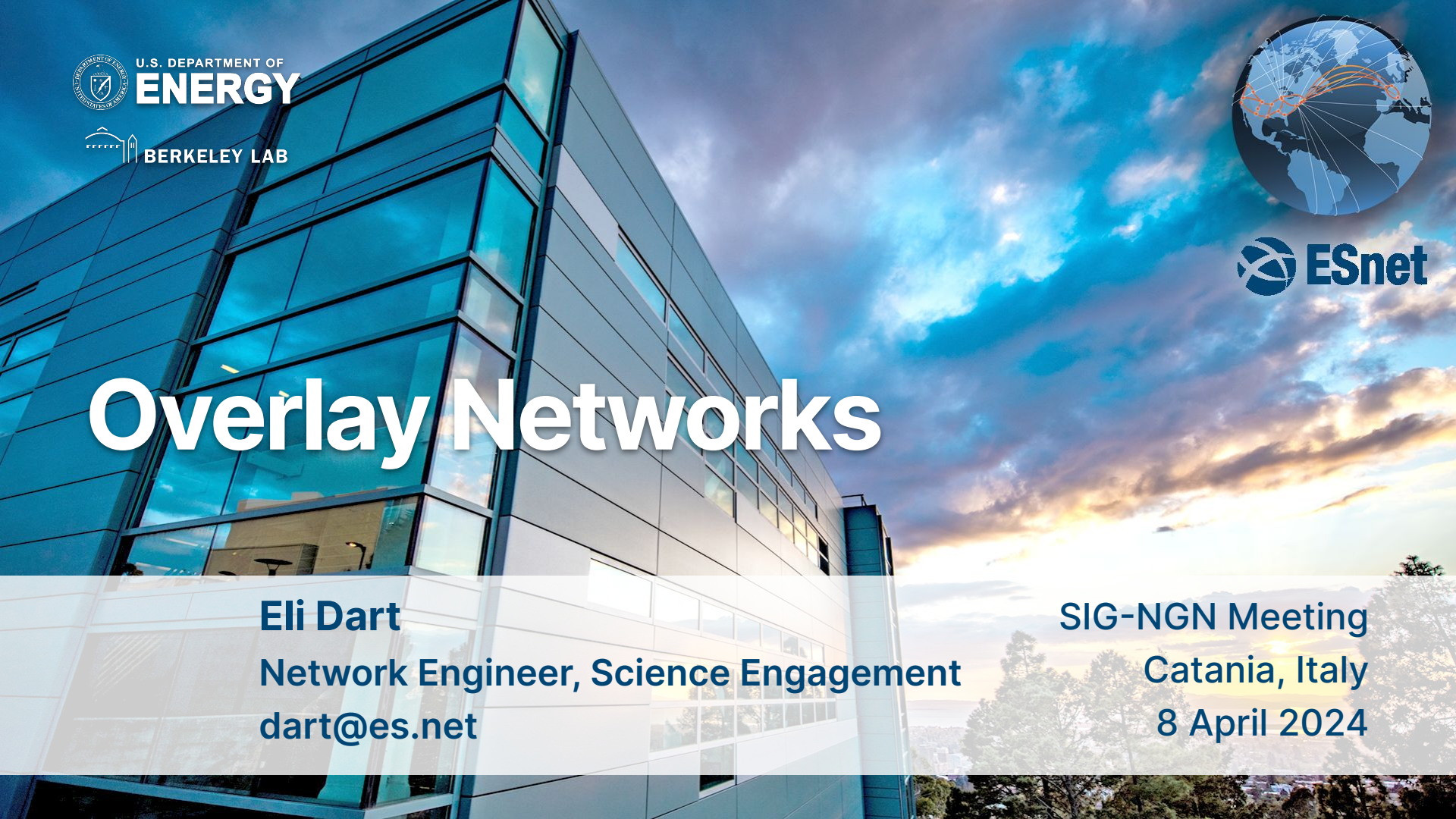


U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

ESnet

Overlay Networks

Eli Dart

Network Engineer, Science Engagement

dart@es.net

SIG-NGN Meeting

Catania, Italy

8 April 2024

Outline

- Overlay networks - Layer2 vs. Layer3
- LHCOPN vs. LHCONE comparison
- Should all science collaborations have an overlay?
- Discussion

Overlay Networks - What Are They?

- An overlay is a virtual network built on top of another network.
- Several ways to do this, including:
  - Separate routing instance (Layer 3)
  - Ethernet frame service (Layer 2)
  - Optical spectrum (not going to cover this here)
- Each of these has different costs/benefits
  - Rigidity vs. flexibility
  - Hard, relaxed, or zero QoS guarantees
  - Local vs. non-local reasoning about config and policy (who has to do what in order for all this to work?)

Layer2 Overlays

- Most of the recent (past 20 years) R&E experiments with Layer2 overlays have provided Ethernet frame service at the network edge
  - Switched VLANs (NLR FrameNet, various SCinet experiments, others)
  - OSCARS and friends (AL2S, SENSE, others)
  - Some experiments with WAN InfiniBand also (Longbow)
  - Not covering ATM here
- Somehow or other, wide area network is made to look like a single Ethernet network (one broadcast domain)
  - Interdomain demarc is a VLAN tag

Layer2 Overlays: What Do I Gain?

- There is a lot of simplicity from the end host perspective
  - I can pick whatever address space I want (so long as everyone involved agrees)
  - I have an ARP entry for everyone I'm communicating with
- I can use LAN protocols without WAN/Firewall concerns
  - Filesystem mount
  - RDMA/RoCE/IBoE - but only if there's no packet loss
- Security risk is often reduced
  - No way into the Layer2 overlay from intermediate devices in the path
  - You can easily enumerate who can land packets on you (if you trust The Other End)

Layer2 Overlays: What Do I Gain? (2)

- Technologies exist for providing hard QoS guarantees (if you use L2 MPLS LSPs)
  - Explicit path
  - Bandwidth allocation with enforcement
  - Fine-grained QoS
- Just those three things have a *lot* of power
  - Support network research, smart grid controls experiments, etc.
  - Bandwidth guarantees + explicit paths → LHCOPN
- This is why ESnet has OSCARS, our Layer2 circuit service - it works well in many different cases

Layer2 Overlays: What Do I Lose?

- Very brittle
  - Depending on technology, less resilience/rerouting/etc. when compared to routed IP (Layer3) - especially interdomain
  - If it works, it works. Mess one thing up and it often fails hard
- Most troubleshooting tools (e.g. ping, traceroute) are layer3 constructs
  - Layer2 failures are often silent
  - Binding 10.x.x.x addresses to each hop is sometimes the only option for troubleshooting WAN VLANs
  - Comparing packet ingress and egress counters for each component/provider in the path is tedious and opaque
- Simple VLANs have little to no QoS, and oversubscription/congestion is opaque

Layer2 Overlays: What Do I Lose? (2)

- Some security risk is increased
  - Layer3/Layer4 filters often not available or not applied (except on the end hosts - increased end host burden)
  - Trust The Other Side more, esp. with LAN protocols
- Scheduling oversubscription may not match usage
  - QoS guarantees are great, but sharing becomes harder
  - Easy for an underutilized network to be “full” from a scheduling perspective if QoS guarantees are honored
- Vastly increased host burden if hosts use Layer2 circuits directly
  - End host must reason about the WAN
  - Very different than “default gateway”
  - ESnet SENSE project tries to mitigate this

Layer3 Overlays

- Separate VRF (Virtual Routing and Forwarding) instance
- Virtual network with a routing table, routing policy, set of interfaces, etc.
  - Includes definition of what prefixes are “in” or allowed
- Allows for different types of traffic (e.g. traffic for prefixes learned from types of BGP peers) to be handled separately
- Commonplace in modern networks, but not required

Layer3 Overlays: What Do I Gain?

- Ability to apply separate policy to sets of peers, prefixes, interfaces
- More flexible than Layer2 - still have IP and all its power
  - Benefits of increased policy control while keeping layer3
  - Troubleshooting tools still work
  - Rerouting around failures still works in “normal” ways
- Widely supported in available hardware

Layer3 Overlays: What Do I Lose?

- Less in the way of traffic engineering control when compared to Layer2
  - Fewer QoS options
  - More like a “network” and less like a point-to-point service
- Everyone connected to it has to do the work of connecting to it
  - Differentiate between overlays at the edge
  - Hosts are not typically aware of Layer3 overlays, so network policy burden increases (more on this later)
  - Non-local reasoning (“which overlay gets me to what service/host/site/domain/whatever?”)

Overlays Case Study: LHCOPN vs. LHCONE

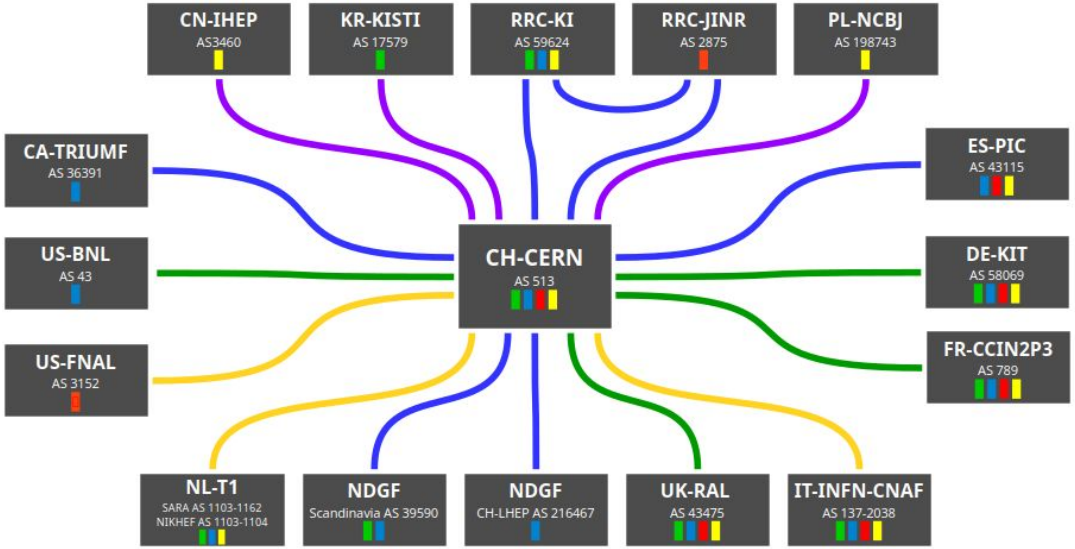
- LHCOPN: LHC Optical Private Network
  - Layer2 circuits between CERN and LHC Tier1 centers
  - For distribution of data from CERN to Tier1s, not for other traffic
  - Originally conceived as optical circuits.
- LHCONE: LHC Open Network Environment
  - Layer3 overlay network for WLCG site to site traffic
  - Allows policy control over LHC traffic (more on this in a moment)
- Each of these has different costs and benefits

LHCOPN History: Why?

- At the time (early to mid 2000s, before LHC Run 1) network was considered most likely to fail of network, servers, storage, software
- Needed guaranteed bandwidth between CERN and Tier1s
- Multiple paths with no fate sharing
  - Redundancy → guaranteed (highly probable?) availability
- Protected minimum bandwidth allocation from CERN for each Tier1
- Conceived as point to point circuits - much of it is still operated this way today

LHCOPN Map: 7 April 2024

LHCOPN



https://twiki.cern.ch/twiki/bin/view/LHCOPN/OverallNetworkMaps

ESnet

LHCOPN in ESnet

- ESnet implementation used (and still uses) OSCARS circuits (MPLS LSPs with QoS, explicit path, etc, providing Ethernet frame delivery service on a VLAN tag at ESnet edge)
- Layer3 addresses bound to Layer2 circuit endpoints, BGP session between site interfaces across the circuit
  - Allows for normal routing inside the sites at the edges of the LHCOPN circuit
  - If the circuit fails, the BGP session drops and traffic re-routes (gives some Layer3 advantages + Layer2 advantages)
- Bandwidth guarantees provide guaranteed minimum service with burst capacity
- Explicit paths ensure no fate sharing between primary, secondary, and tertiary paths
  - (But these are all separate circuits - not scalable to a large mesh of sites)

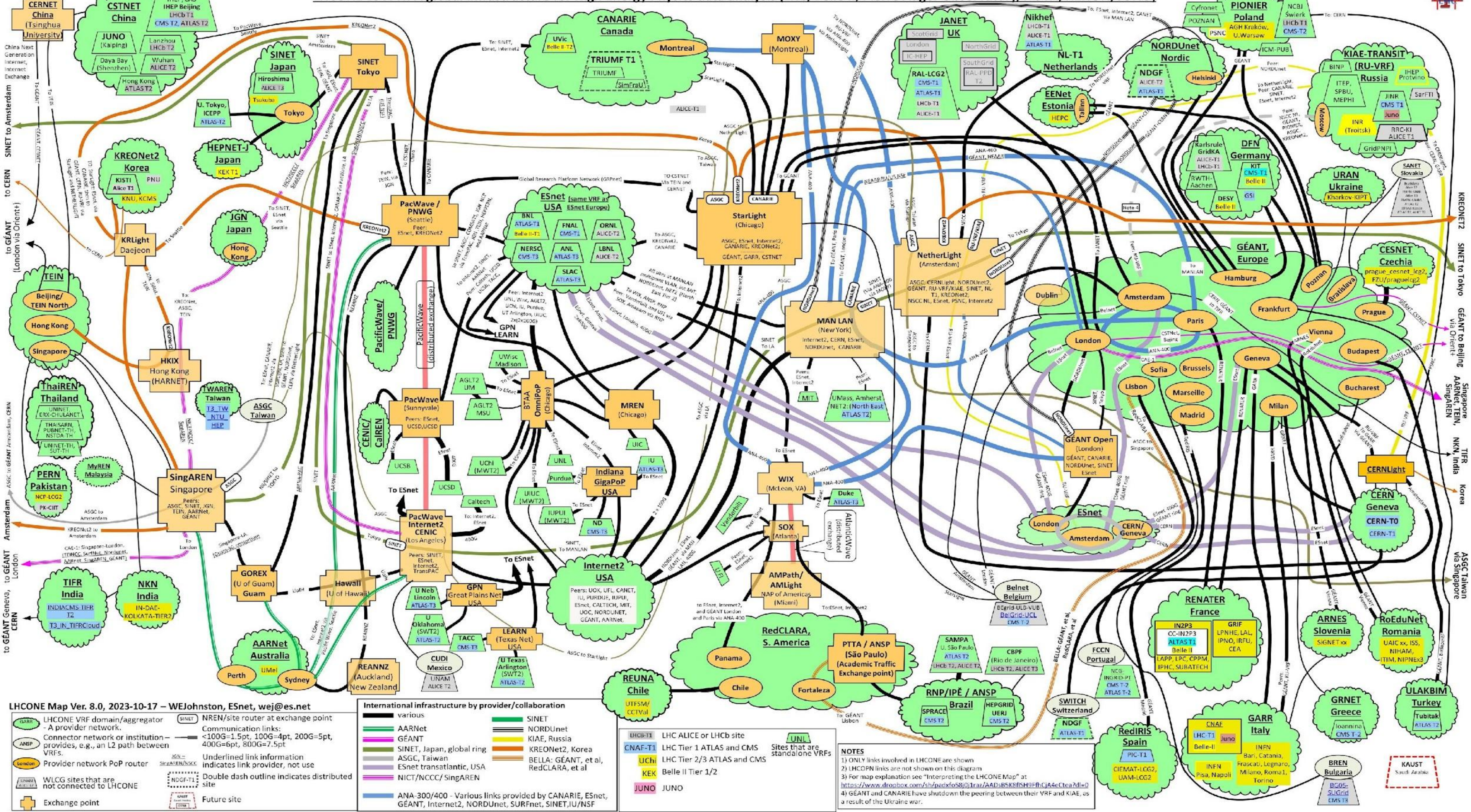
LHCONE History: Why?

- Policy: routing based on AUP, funding, etc.
  - ***Keep LHC traffic on specific circuits or paths***
  - Keep non-LHC traffic off of specific circuits or paths
- Security: keep “LHC” separate from “the Internet”
  - Prefix filtering provides some security benefits
  - Allows for lower-friction security engineering, similar to the Science DMZ model, at some sites
  - If only LHC infrastructure uses LHCONE, then dorm rooms, home broadband, and other sources of malicious traffic aren’t able to use the fast path
- Sociology: participants are Part Of LHCONE, which gives leverage with campus IT

LHCONE Structure

- Layer3 overlay network interconnecting WLCG (Worldwide LHC Computing Grid) sites on multiple continents
  - Gets past the "N^2 mesh of Layer2 circuits" problem
  - LHCONE routing table contains only prefixes devoted to LHC (security benefits)
- Sites announce the prefixes for their WLCG hosts/services into LHCONE so others can reach them via LHCONE

LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)



LHCOPN vs. LHCONE Comparison

- LHCOPN is a narrowly-scoped capability deployed for a specific purpose
- Guaranteed service between CERN and Tier1s
- Uses Layer2 circuits with QoS, with the advantages and disadvantages that come with it
  - Bandwidth guarantees (usually) work
  - Explicit paths are valuable for eliminating fate sharing
  - Failures in the middle are invisible to the ends - they just see packet loss or BGP session transition
  - Explicit paths require explicit human re-engineering as the network evolves over time

LHCOPN vs. LHCONE Comparison (2)

- LHCONE is a mini-Internet for LHC
- Large number of connected sites
- Sites exchange data freely, and in fact rely on the ability to do so (e.g. remote I/O, arbitrary data transfers as dictated by automated workflows)
- Broader scope invites increased complexity (but also provides increased utility)
- Both provide security benefits
  - Some measure of control over what traffic sent to and received from LHC vs. Internet
  - LHCONE's broader scope reduces that benefit

Lessons Learned

- LHCOPN is straightforward, but comes with some challenges.
  - Difficult for ends to reason about the middle (as described)
  - Does not scale beyond a small number of sites, even with hub-and-spoke instead of mesh
  - Moves, adds, and changes are heavy for network operators
- LHCONE is a much more complicated thing
- Lots of things learned from 10+ years of LHCONE

Lessons Learned: LHCONE

- It has been very difficult to use the same address space for LHCONE and “normal” networking
  - Constraints on what jobs run on which machines
    - Easier to put Belle-II and others in LHCONE than to separate them out on the hosts
  - Source-based and destination-based routing to conform to AUP → policy routing
    - Operational complexity at the end sites
    - Requires more sophisticated/expensive hardware
  - Dedicated address space (e.g IPv6) for an overlay would solve some of this, but would come with its own costs
    - e.g. need for a central coordinating authority
    - Additional complexity on end hosts
- A network architecture that enforces application AUPs has significant end-host implications

Lessons Learned: LHCONE (2)

- Significant value in separating LHC traffic from general Internet traffic
  - Allows network engineers to select the best circuit (e.g. for high bandwidth)
  - E.g Allows some networks to make use of specific circuits with specific funding
  - Security benefits for site engineering
- Significant sociological value in technical connectivity to a major international construct
  - Gives some sites leverage with Campus IT
  - Collaboration meetings provide community cohesion
- However, everyone must connect
  - Difficult when there are many small sites in a community
  - Without broad consensus, many benefits are lost

Lessons Learned: LHCONE (3)

- Difficult to attach multi-tenant community resources to LHCONE (!!)
  - Multi-program sites/facilities - LHC is one user among many
  - They don't want to differentiate between jobs by IP address
- Lots of collaboration and coordination required
  - Institutions that use LHCONE need to have network engineering staff who understand this stuff, or others to mentor them
  - As do their up-stream providers!
  - Changes at one site affect others (e.g. rogue traffic: Bruno and MOC spent a bunch of time on this [*even though the volume of traffic was tiny compared to overall - not sure it was worth the effort*])
  - Coordination of policy and operations is critical for success
- Currently not feasible to extend LHCONE into Cloud

A Note On Performance

- It's easy to say "Layer2 performs better" or "LHCONE is necessary for performance" - not true
  - Modern devices forward equally well at Layer2 and Layer3
- If a path is meticulously engineered, it will almost always perform better than "if it pings it's good"
- ***Overlays are often routed around enterprise firewalls***
  - Protect/defend the overlay with performant security technologies instead of enterprise firewalls
  - This is essentially the Science DMZ model, which works fine with destination-based Layer3 (no overlay)

Overlays Everywhere - EveryONE?

- So: should new collaborations build overlay networks?
  - As usual, I answer this with “it depends”
- If a collaboration is approaching data networking from scratch, I would encourage them to consider well before creating their own LHCONE-like overlay
  - How many sites? Will they all connect?
  - What is the data workflow? How many different data workflows?
  - Do you have a central coordinating authority?
  - How much network engineering expertise does your community have?
  - How much cohesion is there between different sites?
- Fundamentally: Are the benefits worth the cost?
  - It can be a great thing, but it’s not free


U.S. DEPARTMENT OF
ENERGY
BERKELEY LAB
ESnet
Thanks!
Eli Dart
dart@es.net
https://my.es.net/
https://www.es.net/
https://fasterdata.es.net/