ESnet
ENERGY SCIENCES NETWORK

ESnet: Advanced Networking for Data-Intensive Science

William Johnston, Eli Dart, Chin Guok
Energy Sciences Network (ESnet)
Lawrence Berkeley National Laboratory

ASCAC, March 2019

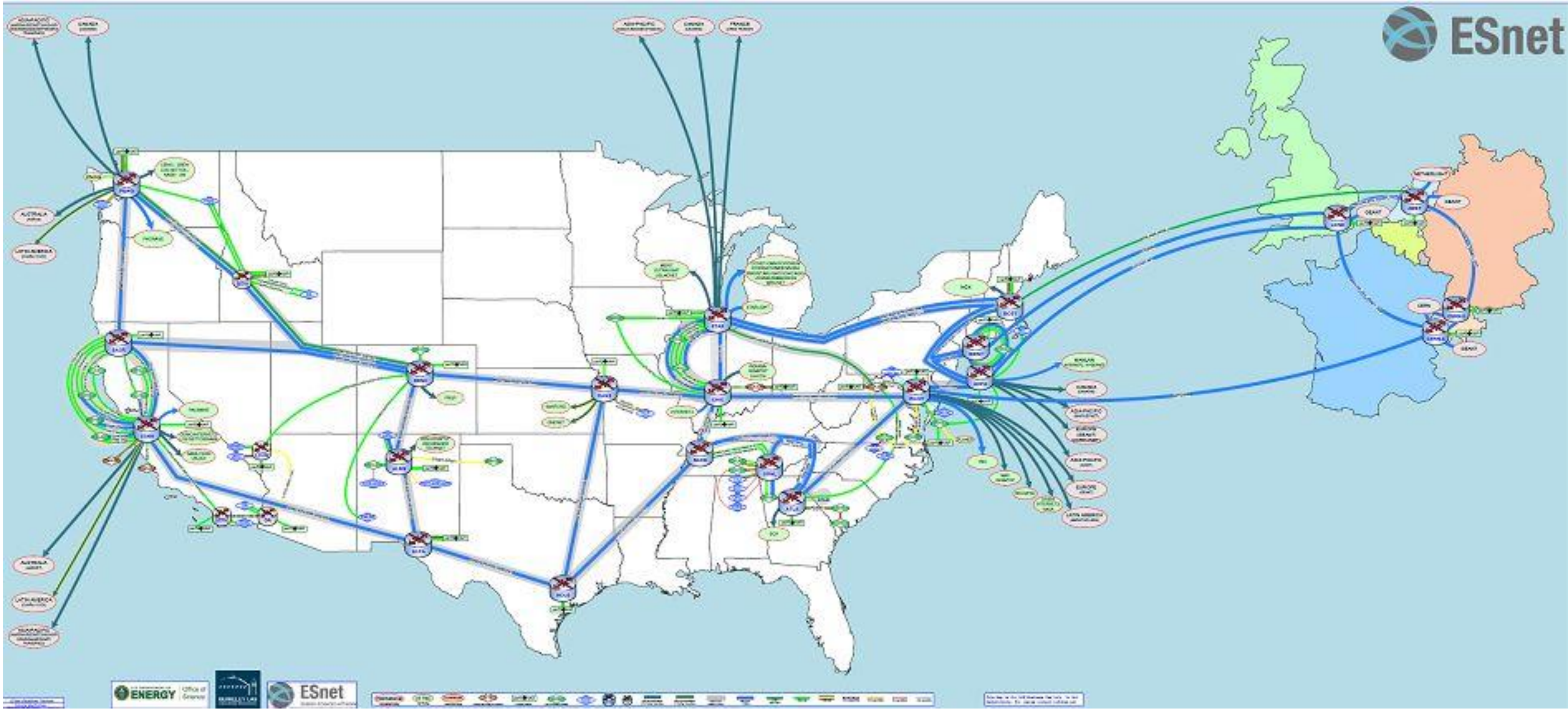
U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

What is ESnet

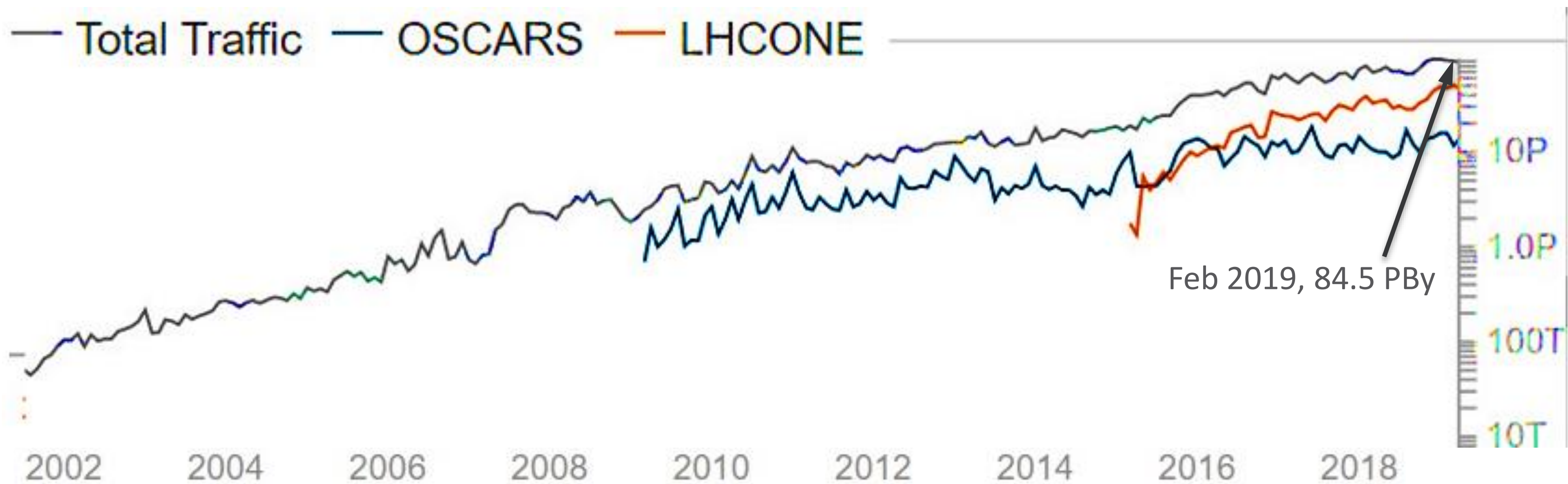
- The Energy Sciences Network (ESnet) is a high-performance, unclassified network built to support scientific research
- Funded by the U.S. Department of Energy's Office of Science (SC) and managed by Lawrence Berkeley National Laboratory
- Provides services to more than 50 DOE research sites, including the entire National Laboratory system, its supercomputing facilities, and its major scientific instruments, as well as to a dozen or so US University High Energy Physics groups
- Connects to 140 research and commercial networks, permitting DOE-funded scientists to productively collaborate with partners around the world

Connectivity and capacity are key for science



- ESnet5 ca late 2017: Multiple 100 Gbit/sec paths with redundancy
- ESnet core network covers US and Europe
- Connectivity to all major US and international research and education (R&E)sites

ESnet traffic increases about 10x every 4 years

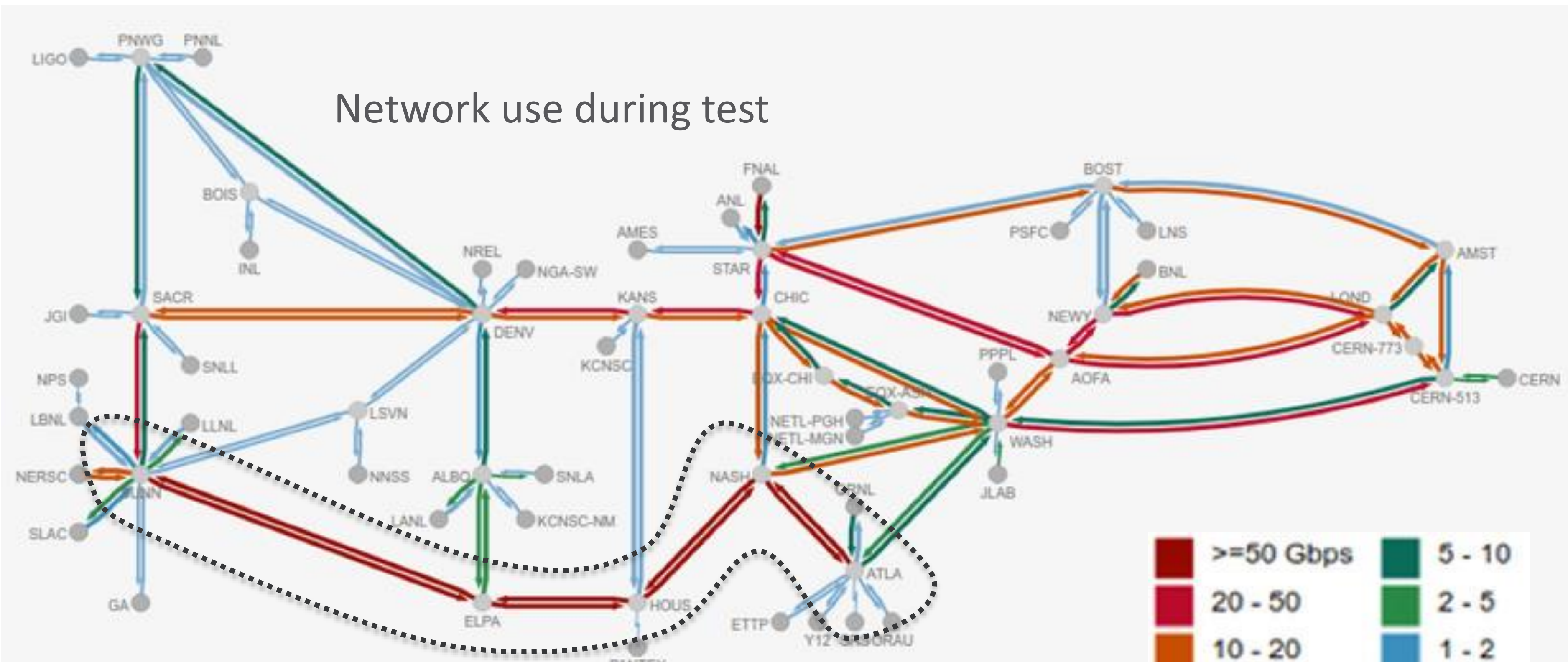


(note the log scale)

- ESnet's exponential traffic growth is driven by the rising tide of data produced by global collaborations that can involve thousands of researchers, specialized facilities like the Large Hadron Collider and digital sky surveys, and more powerful supercomputers

Superfacility use of network drives higher data transfer speed requirements

- Testing for connecting SLAC's LCLS II linear collider coherent X-ray facility directly to DOE supercomputers in 2020
  - ESnet configured a 5000 mile, 100Gb loop from SLAC, and allowed up to 80Gb tests
  - Software and network-attached storage systems from DOE SBIR funded Zettar were tested at SLAC and moved **1 Petabyte in 29 hours over the 5000 miles**

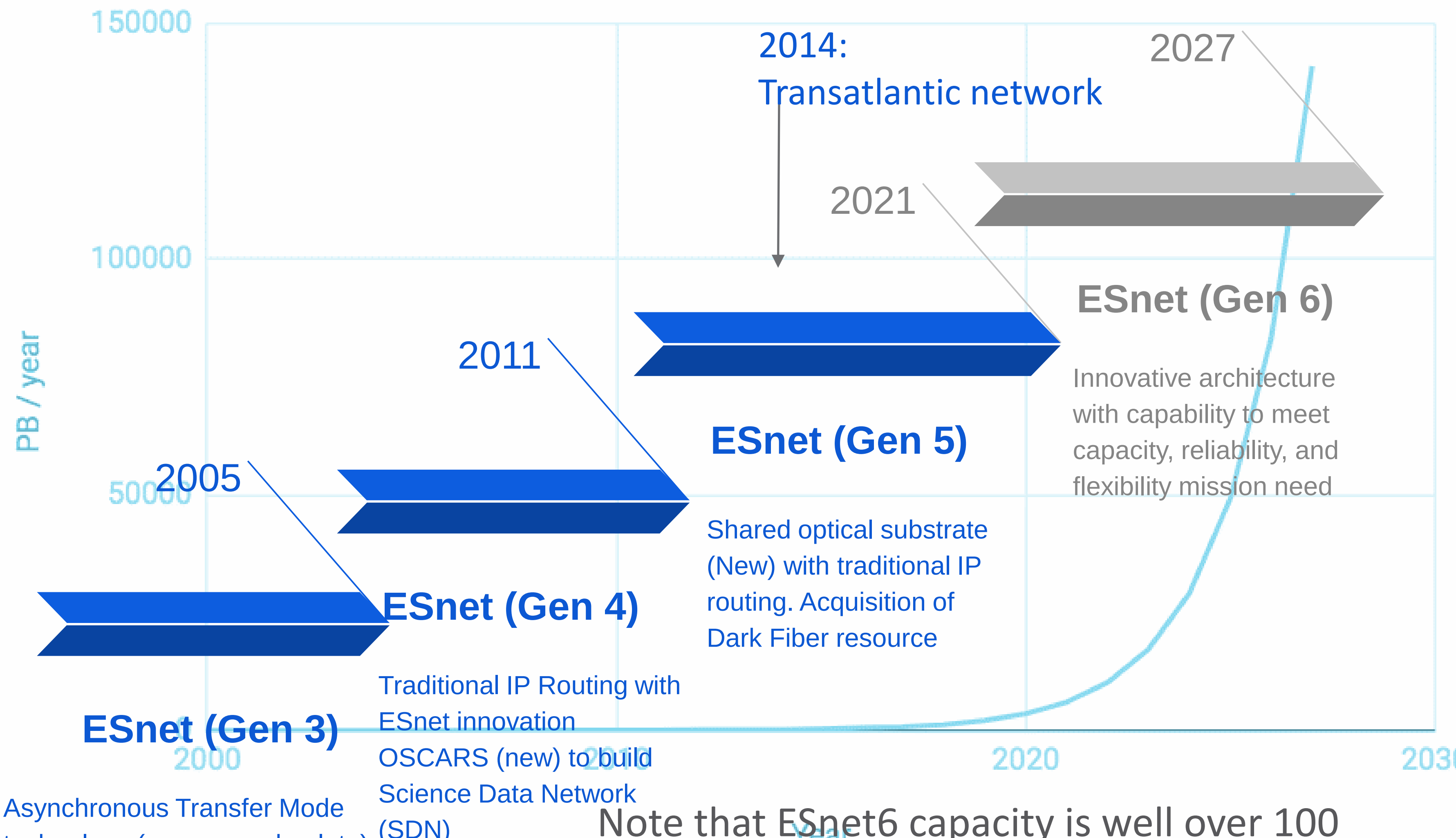


ESnet6 – the next generation

- **“Hollow core” based on Esnet owned fiber network where ESnet manages entire capacity of the fiber**
  – Packet and optical services with full monitoring
  – Scalable allocation of optical capacity
  – Full mesh connectivity between all sites for data path protection and restoration
  – Software driven API for automatic bandwidth allocation
- **“Service edge” – where the users connect**
  – Standard and customized (e.g. encapsulation based) user connections with quality of service guarantees
  – Automated management and dynamic creation of user requested services
  – Per-flow monitoring
  – High speed per packet filtering and forwarding to enforce security policies
  – Programable interfaces to support emerging and new Software Defined Networking functions
- ESnet on-going AI based research into automated error detection and flow characterization will be incorporated into ESnet 6

ESnet

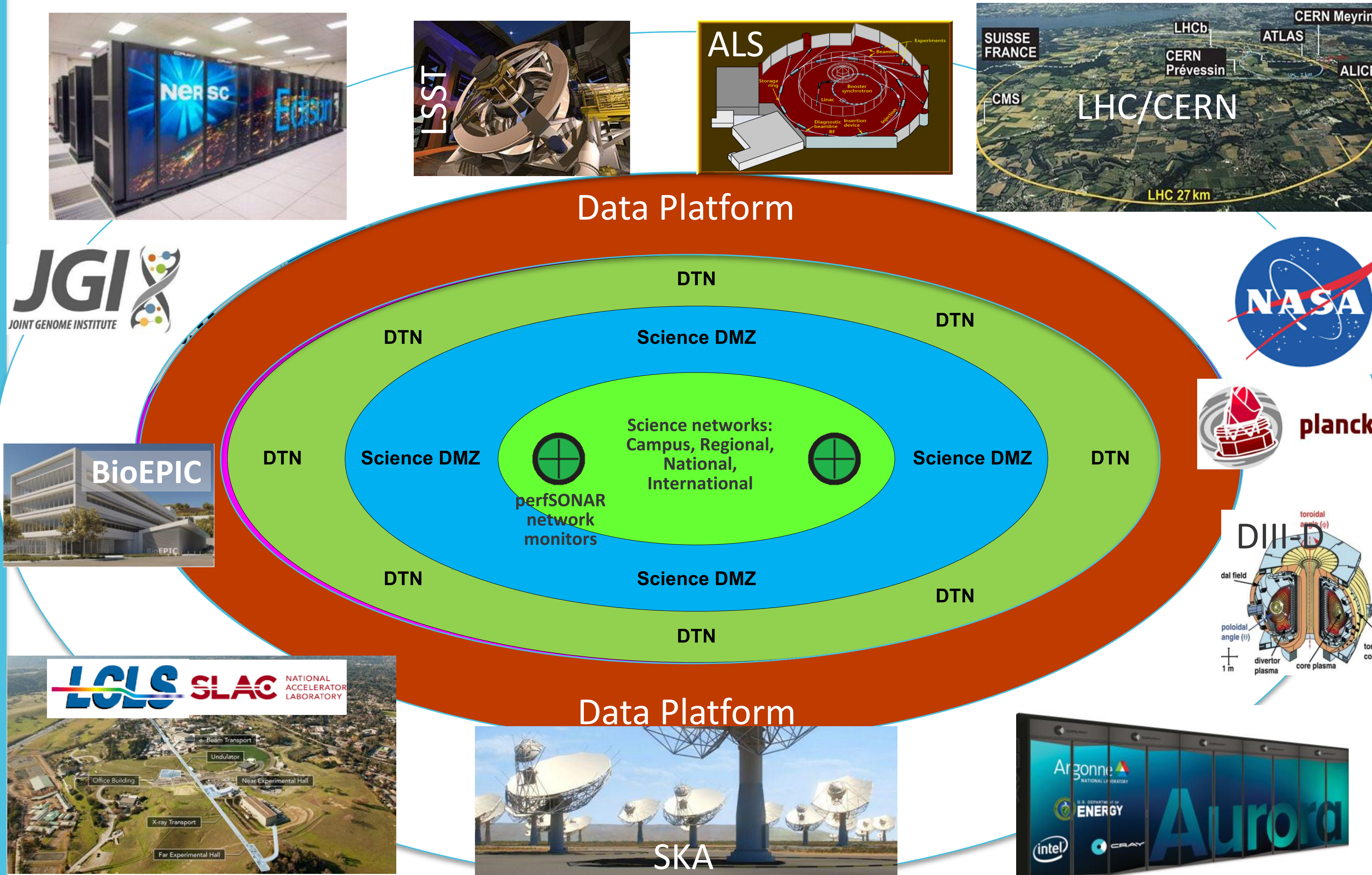
Each major upgrade transforms the facility with innovative, cutting edge technologies



Asynchronous Transfer Mode technology (new, now obsolete), provided by Carrier Sprint

Note that ESnet6 capacity is well over 100 Exabytes/yr, which will be needed if multiple SC superfacilities like the LCLS move a petabyte/day.

Enabling the data ecosystem where science is done



+ The goal is that the users see only the data platform for data access
+ Its building blocks enable transparent, high speed data movement

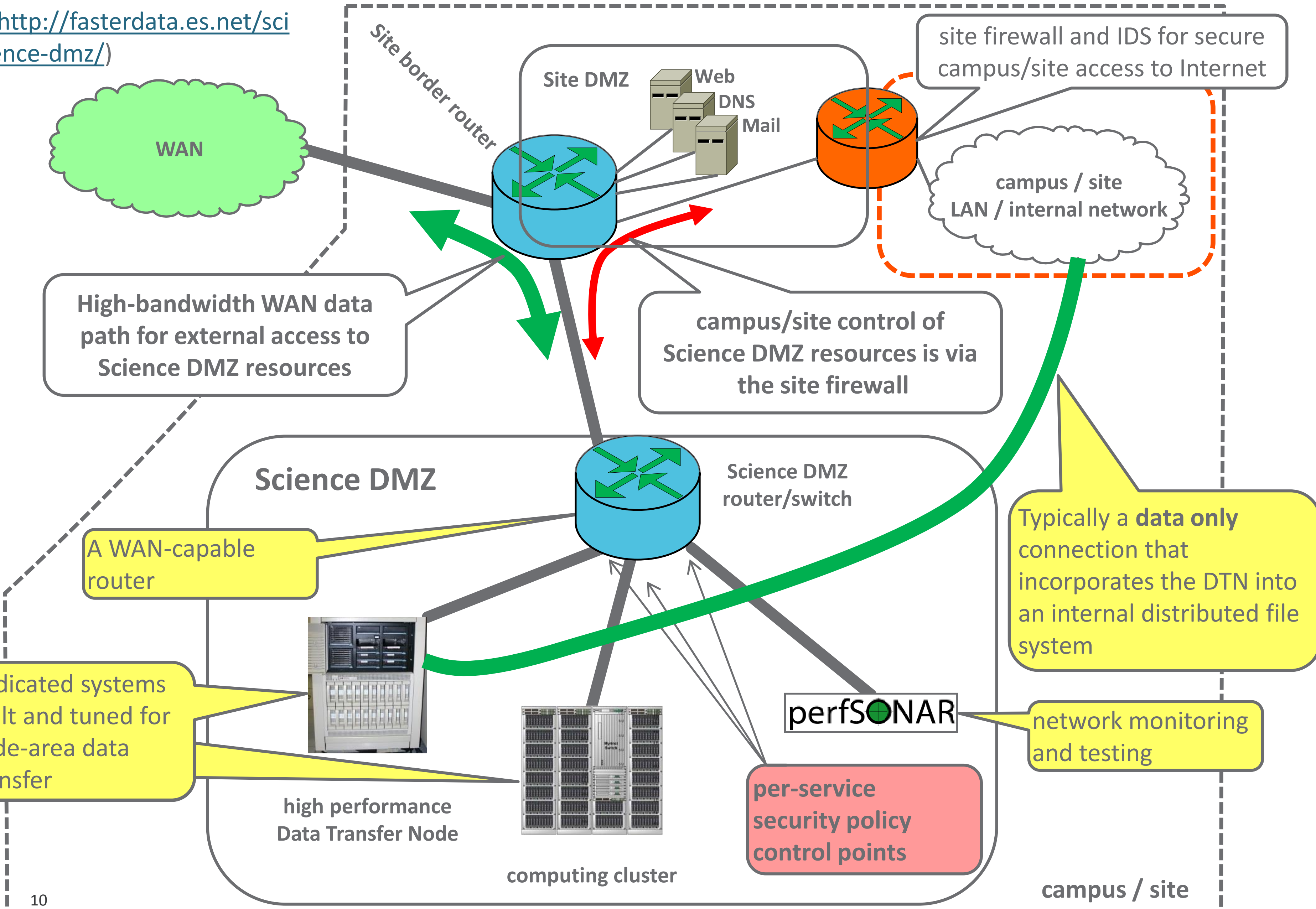
ESnet

Building blocks for the Data Platform - 1

- **Science DMZ**
  - **A network architecture that enables securely connecting site high performance data servers directly to the wide area network**
  - Located on the site “DMZ” – a piece of the campus network that is outside the campus firewall (which is a performance bottleneck for large data transfers)
  - Designed for high speed remote data transfers and strong security
  - Most DOE sites, many NFS funded sites, and some European sites now use the Science DMZ architecture (several hundred, overall)
- A recent innovation is to add a “**Data Portal**” capability
  - A Web frontend that lets the user request data from Science DMZ storage and directs that data to a local system.
  - The user never directly interacts with the Science DMZ components which increases ease of use and security
  - Similar in concept to Globus Online
  - See “The Modern Research Data Portal: a design pattern for networked, data-intensive science” https://peerj.com/articles/cs-144/

The Science DMZ

(http://fasterdata.es.net/science-dmz/)



Building blocks for the Data Platform - 2

- **Data Transfer Nodes (DTNs)**
  - **Purpose-built systems dedicated to wide area data transfer**
    - **This involves a lot of software and hardware not typically found on disk servers (see fasterdata.es.net)**
  - DTN has access to local site storage - e.g. a connection to a local storage infrastructure such as a SAN, or the direct mount of a high-speed parallel filesystem such as Lustre or GPFS
  - Runs software tools designed for high-speed data transfer to remote systems
    - typical software packages include GridFTP and its service-oriented descendent Globus Online, discipline-specific tools such as the LHC's XRootd, and versions of default toolsets such as SSH/SCP with high-performance patches applied
    - Therefore, **not all DTNs can talk to each other, but DTNs serving a science collaboration typically run the same software and can communicate with other DTNs supporting the collaboration world wide**

Building blocks for the Data Platform - 3

- **Monitoring and testing** the network end to end is only way to keep multi-domain, international scale networks error-free, which is essential for high speed data transfer
  - **perfSONAR** provides a standardize way to test, measure, export, catalogue, and access performance data from many different network domains (service providers, campuses, etc.) – i.e. it is a multi-domain monitoring system
    - perfSONAR is a community effort to
      - define network management data exchange protocols, and
      - standardized measurement data formats, gathering, and archiving
  - perfSONAR is deployed extensively throughout LHC related networks, in international networks, and at end sites – there are thousands of instances deployed
  - www.perfsonar.net

LHCONE: Not all big data traffic is suitable for the general Internet

- As the LHC ramped up to first production operation, ESnet monitoring detected several transatlantic network paths serving the R&E community were being congested
- Finding the cause was not trivial because it turned out to be LHC data analysis groups moving data with GridFTP using dozens of parallel data transfers, so no one end system stood out in the monitoring
- ESnet engaged CERN on how to deal with this, and CERN set up a study group to characterize the problem
- CERN, ESnet, and Internet2 to set up a working group to make recommendations on how to address this issue
  - ESnet engineers proposed a **network overlay approach where the paths used by the overlay were explicitly under control of network operators**
    - In other words, the paths could be easily configured by network engineers not to interfere with general R&E traffic in their domain
    - Access to the overlay was limited to high energy physics projects, which also provided a modicum of security
- The result is called **LHCONE** and carries most of the LHC data worldwide

 – See http://lhcone.web.cern.ch

ESnet

- This is not a slide, it is a copy of the working map of LHCONE
- It indicates the scale and geographic scope (world wide) of the overlay
- Most of the petabytes/week that move during LHC data analysis use LHCONE
    - Half of all ESnet traffic is on LHCONE paths within ESnet

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