ESnet
ENERGY SCIENCES NETWORK

R&E Upgrades for HL-LHC

Dale W. Carder

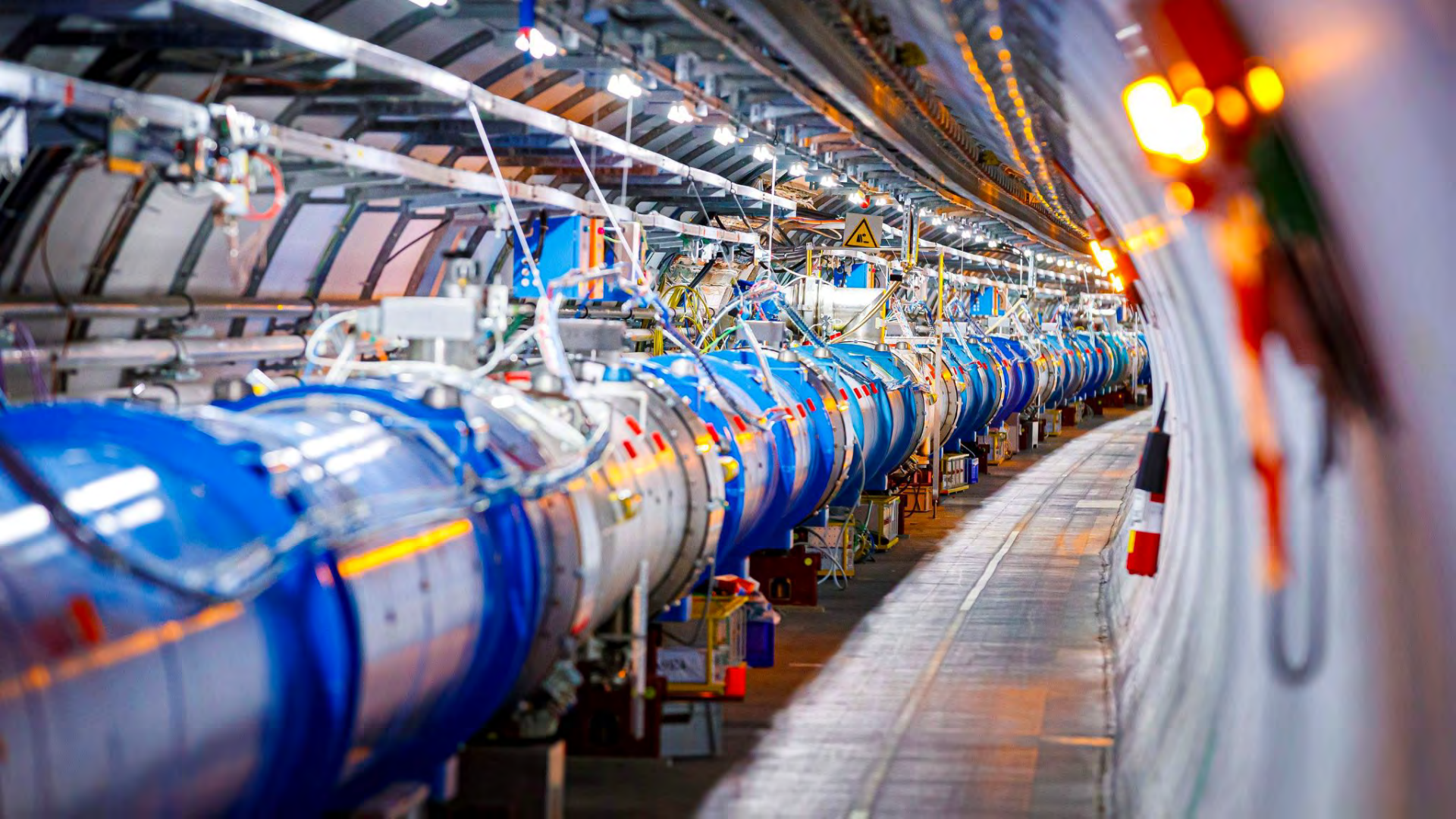
ESnet Network Engineering

dwcarder@es.net

Internet2 TechEx '23
2023-09-20

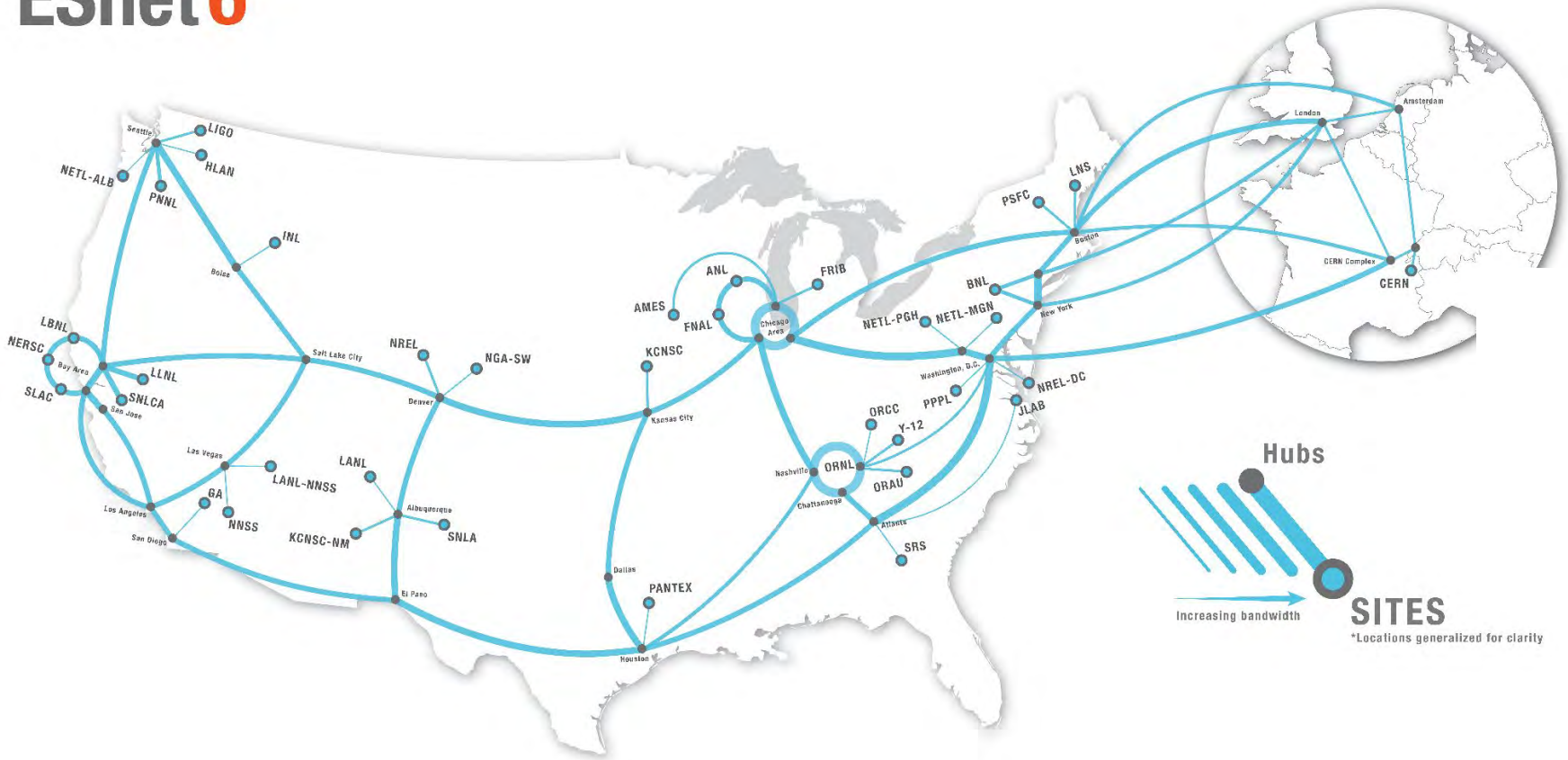
U.S. DEPARTMENT OF ENERGY
Office of Science

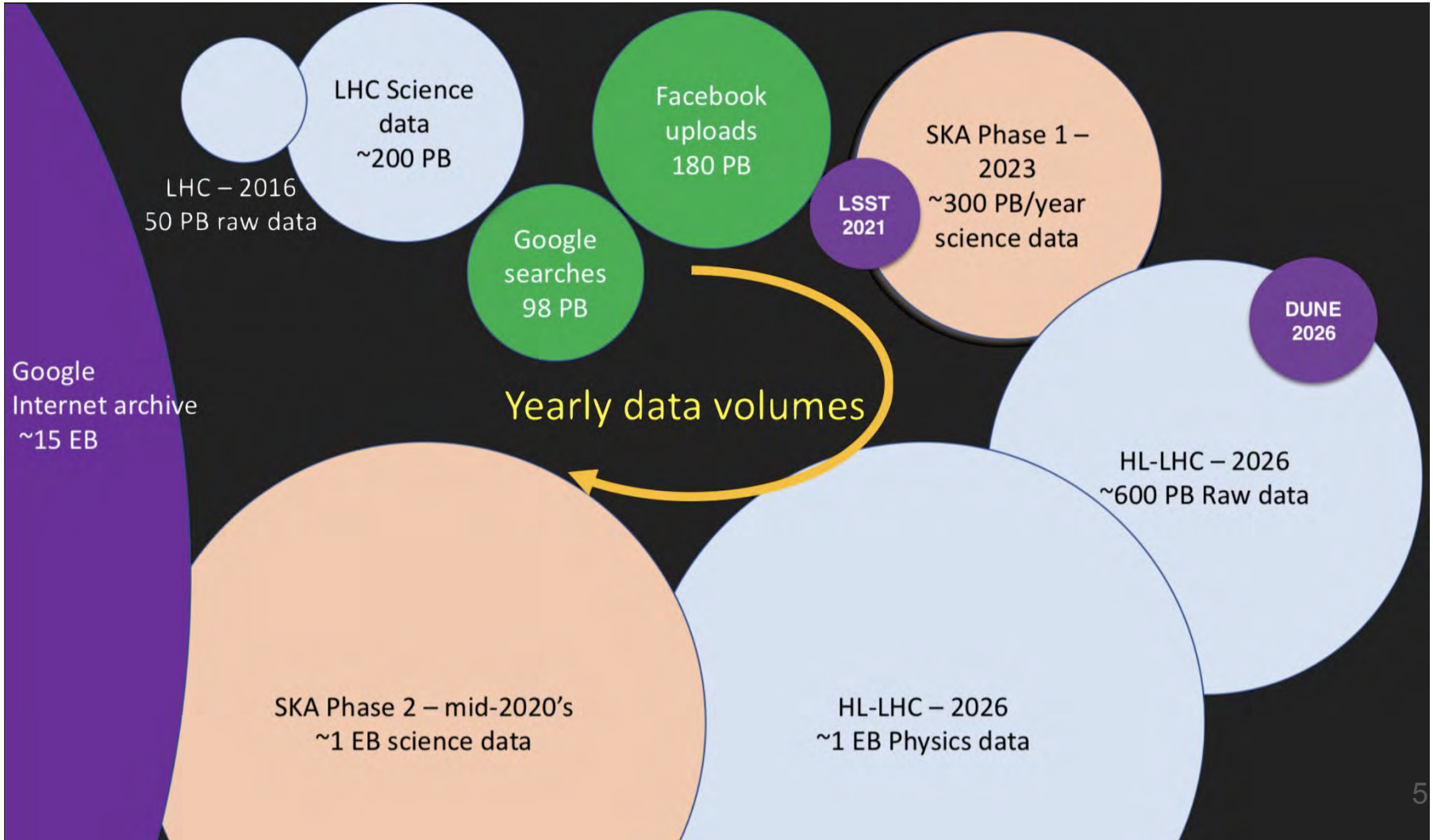
BERKELEY LAB




CMS
LHCb
ALICE
ATLAS

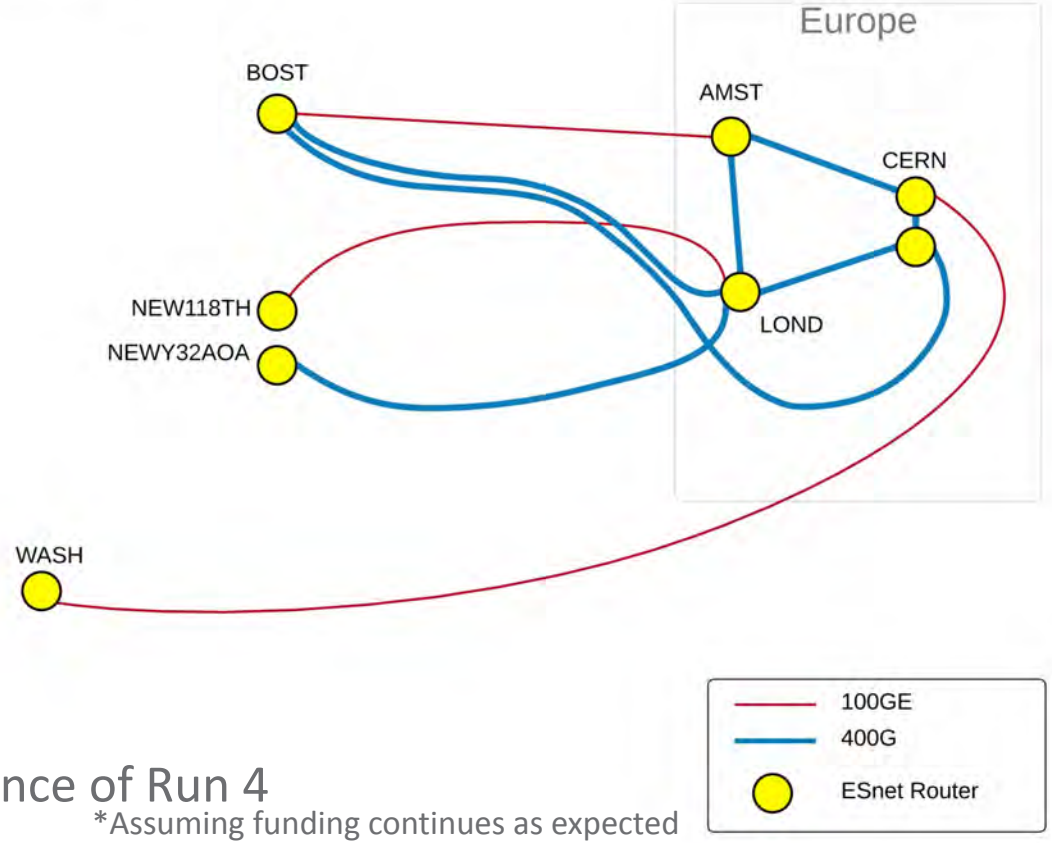
ESnet6


Seattle
LIGO
HLAN
NETL-ALB
PNNL
INL
Boise
LBNL
NERSC
Bay Area
LLNL
SLAC
SNLCA
San Jose
Salt Lake City
NREL
NGA-SW
Denver
Las Vegas
LANL-NNSS
GA
NNSS
Los Angeles
San Diego
LANL
Albuquerque
SNLA
KCNSC-NM
El Paso
Dallas
PANTEX
Houston
KCNSC
Kansas City
AMES
ANL
FNAL
FRIB
Chicago Area
Nashville
ORNL
ORCC
Y-12
ORAU
Chattanooga
Atlanta
SRS
NETL-PGH
NETL-MGN
Washington, D.C.
PPPL
NREL-DC
JLAB
PSFC
LNS
Boston
BNL
New York
London
Amsterdam
CERN Complex
CERN
Hubs
Increasing bandwidth
SITES
*Locations generalized for clarity


LHC – 2016
50 PB raw data
LHC Science data
~200 PB
Facebook uploads
180 PB
Google searches
98 PB
LSST 2021
SKA Phase 1 – 2023
~300 PB/year science data
DUNE 2026
Google Internet archive
~15 EB
Yearly data volumes
HL-LHC – 2026
~600 PB Raw data
SKA Phase 2 – mid-2020's
~1 EB science data
HL-LHC – 2026
~1 EB Physics data

Trans-Atlantic & EU ring upgrades

- Currently underway:
  - ✅ 400G New York - London
  - 400G Boston - London
  - 400G Boston - CERN
  - 400G Europe Ring
- Trans-Atlantic capacity targets
  - 800G aggregate now
  - 1.5T in Q4 2023
  - ...
  - **3.2T* in 2027**, well in advance of Run 4



*Assuming funding continues as expected

HL-LHC Timeline & Milestones

- Data Challenge '24
  - February 2024
  - Run at ~25% of HL-LHC target workload
  - May double the traffic to some sites
- Data Challenge '26
  - double the traffic again
  - 50% the target workload
- Data Challenge '28
  - double the traffic again
  - 100% of the workload
- 2029 Production begins

ESnet

1:1 Conversations w/ University & Regionals

- Gathering and helping synchronize plans from
  - Individual PI's
  - Departmental Support Staff
  - Campus IT & CIO
  - Regional Networks
  - R&E Exchange points

- Coordinating transition to 400G connections
  - avoiding premature costs

ESnet

Current R&D Efforts

PerfSonar

- Worldwide deployment of 5.x
- ESnet nodes are **IPv6-only**

SciTags

- IPv6 Packet Marking

  https://www.scitags.org/

SENSE

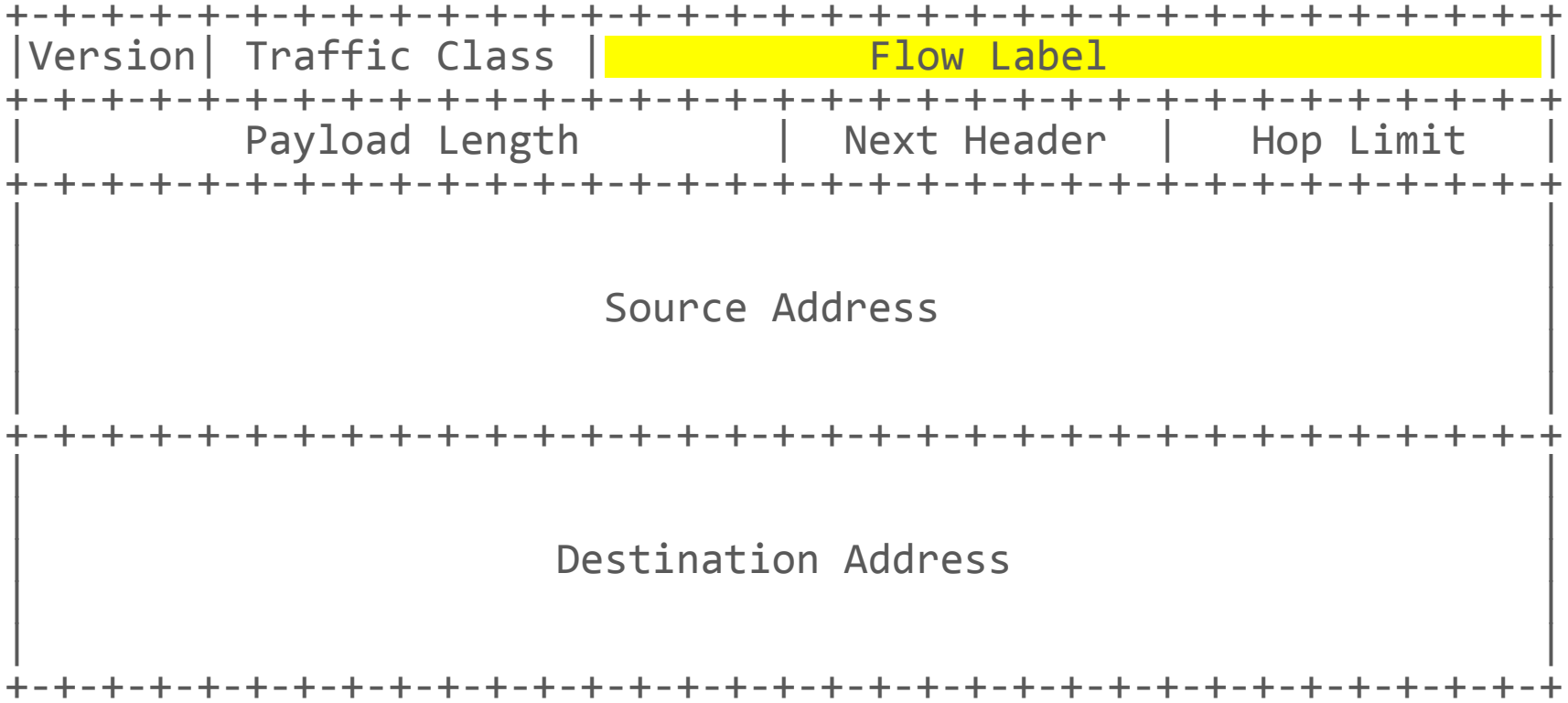
- Dynamic resource provisioning & traffic engineering

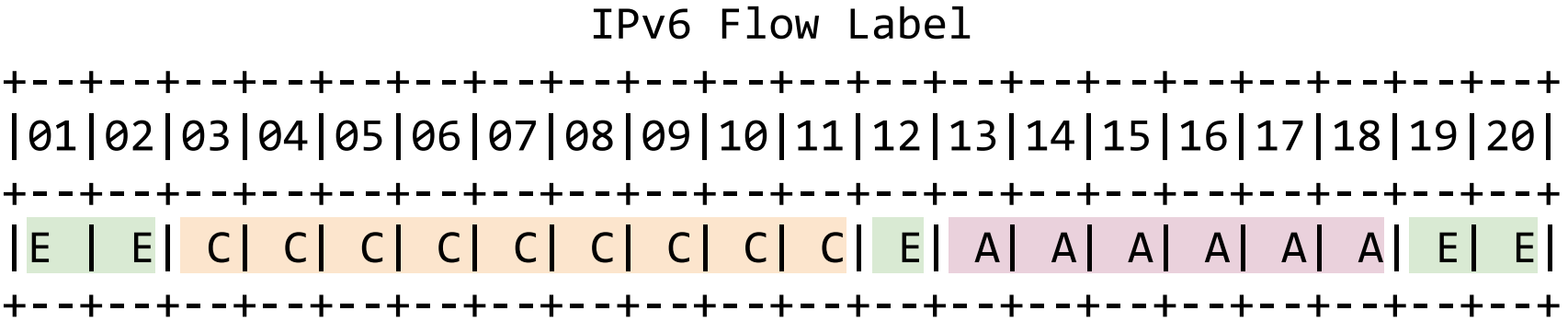
Traffic Pacing

- Test out BBRv3 congestion control & packet pacing

ESnet

IPv6 Header


Version
Traffic Class
Flow Label
Payload Length
Next Header
Hop Limit
Source Address
Destination Address



- (C) Community identifier: "Who are you affiliated with?"
- (A) Activity identifier: "What are you doing within your community?"
- (E) Entropy bits sprinkled throughout
  - set at random once per network flow for the duration of its lifetime.

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: 11 January 2024

D. Carder
Energy Sciences Network
T. Chown
Jisc
S. McKee
University of Michigan
M. Babik
CERN

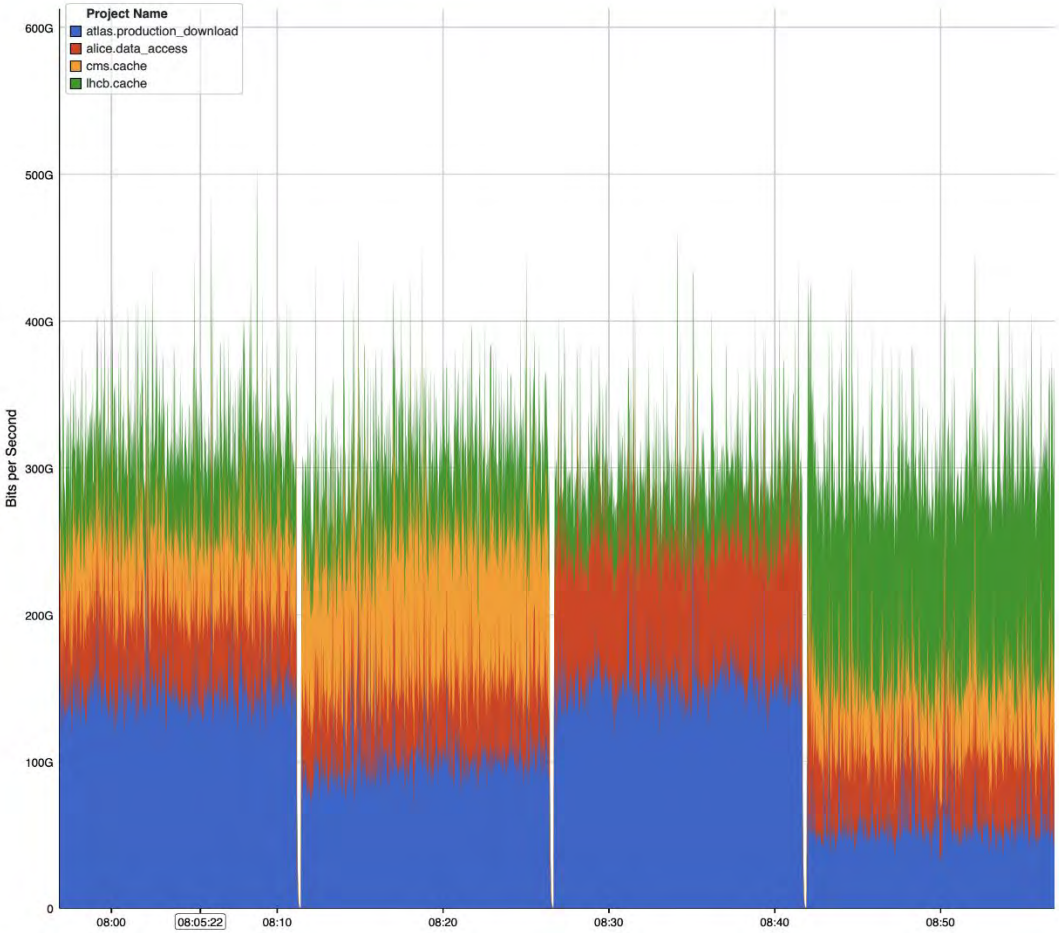
10 July 2023

Use of the IPv6 Flow Label for WLCG Packet Marking
draft-cc-v6ops-wlcg-flow-label-marking-02

Abstract

This document describes an experimentally deployed approach currently used within the Worldwide Large Hadron Collider Computing Grid (WLCG) to mark packets with their project (experiment) and application. The marking uses the 20-bit IPv6 Flow Label in each packet, with 15 bits used for semantics (community and activity) and 5 bits for entropy. Alternatives, in particular use of IPv6 Extension Headers (EH), were considered but found to not be practical. The WLCG is one of the largest worldwide research communities and has adopted IPv6 heavily for movement of many hundreds of PB of data annually, with the

ESNet-CERN SciTags (IPv6 Flow Labels)



supplemental material

SciTags Rationale

- Complex workflows used by multiple data-intensive science communities
  - ~1.4M x86 cores across ~170 sites w/ ~1.6 EB of storage
  - Individual network flows usually small, but can aggregate to many 10's Gbit/s
- Traffic on purpose-built networks (LHCOPN, LHCONE) as well as R&E Networks
  - **Predominantly IPv6,** working towards **IPv6 exclusively**
- Mark packets to identify traffic owner/purpose.
  - Coarse definitions of community/activity provides insight *in aggregate*
- Track data transfers with *existing* network flow monitoring (IPFIX & sFlow)
  - Quantify global behavior and analyse tradeoffs at scale
    - ex: dataset & storage placement, job scheduling
- Potential future use for traffic engineering

Discussion on IETF Compliance

- [RFC6437] interoperate as entropy into ECMP / LACP hash functions
- [RFC6437] **RECOMMENDED** that hosts use a discrete uniform distribution
- [RFC8200] treat these packets in the network as a single flow
- [RFC7098] server load balancing. Minimally a 2-tuple w/ source address
  - (generally out of scope for our use cases)

[RFC6437] && [RFC3697] "Router performance SHOULD NOT be dependent on the distribution of the Flow Label values. Especially, the Flow Label bits alone make poor material for a hash key."

[RFC6438] intermediate routers using ECMP or LAG "MUST minimally include the 3-tuple {dest addr, source addr, flow label}"

Alternatives considered & discussed in the IETF draft

- Hop-by-hop options
  - highly problematic
  - potential for drops outside of a limited domain
- Destination options
  - buried deeper, not as easy to expose via IPFIX
  - socket API issues, potential for future work?
- Source address prefix/bit colouring
  - it's a hack
- Marking in payload
  - can't, it's encrypted
- Tokens / Path signals
  - emerging area
- Firefly
  - flow marking via separate, in-band telemetry packets
  - parallel effort, work in progress