

Use of the IPv6 Flow Label for WLCG Packet Marking

Dale W. Carder - LBNL / ESnet (presenter)

Tim Chown - Jisc

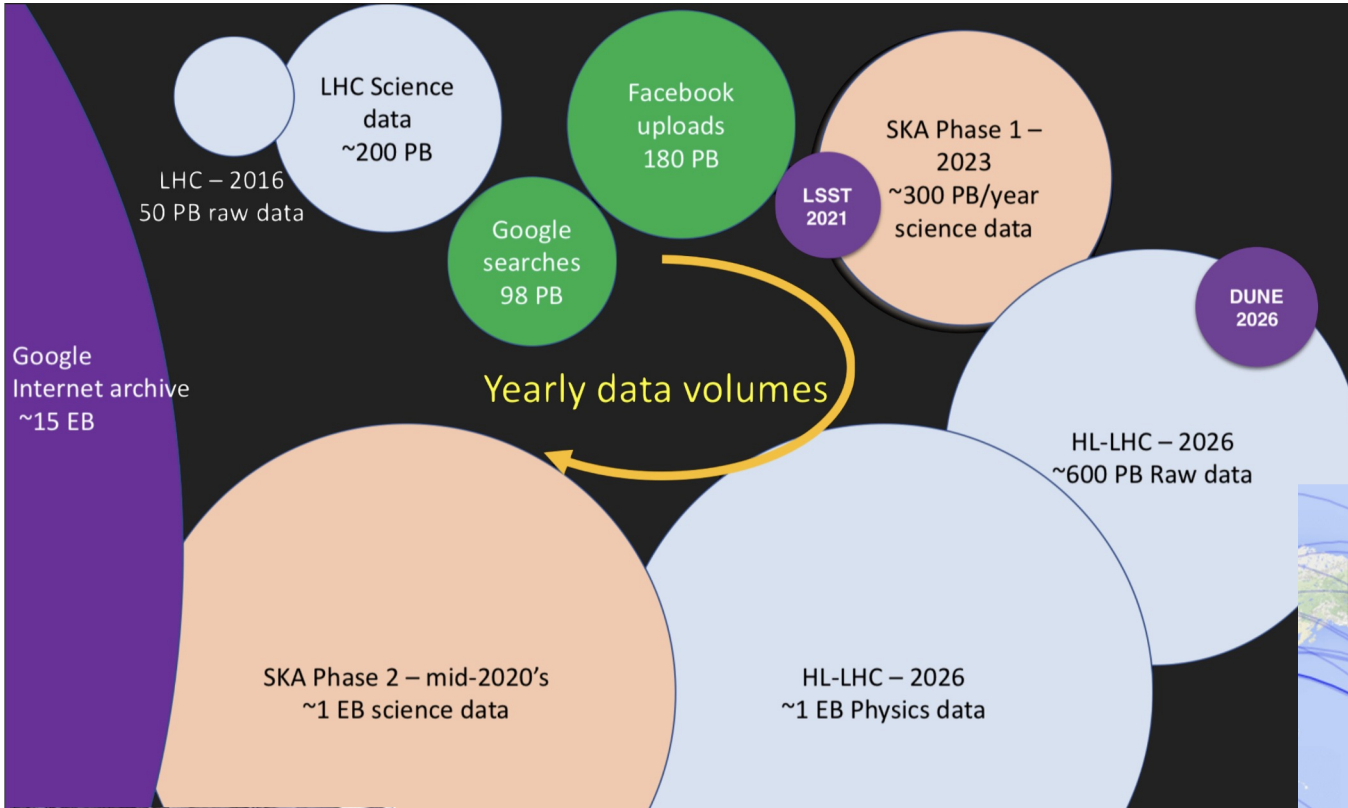
Shawn McKee - University of Michigan

Marian Babik - CERN

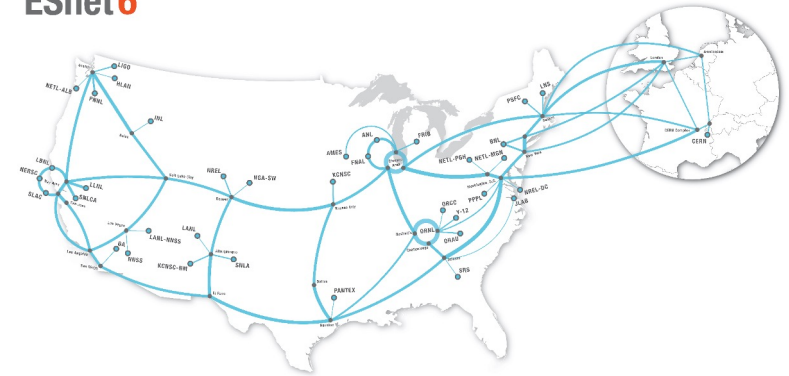
draft-cc-v6ops-wlcg-flow-label-marking

IETF 117, San Francisco, 25 July 2023

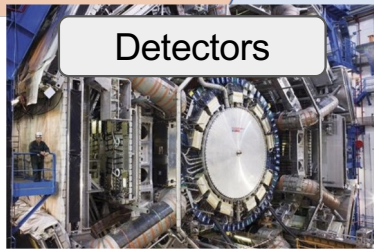
Infrastructure distributed worldwide on purpose-built networks



ESnet6



Accelerators



Detectors



Datasets

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 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 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

Discussion