CERN

Global Networking Challenges for the Coming Decade

Tony Cass, CHEP 2024, 23rd October 2024
With thanks to Carmen Misa Moreira, Edoardo Martelli, Marian Babik, Enzo Capone, Bruno Hoeft, Lars Fischer, Pepe Flix, David Kelsey & Shawn McKee

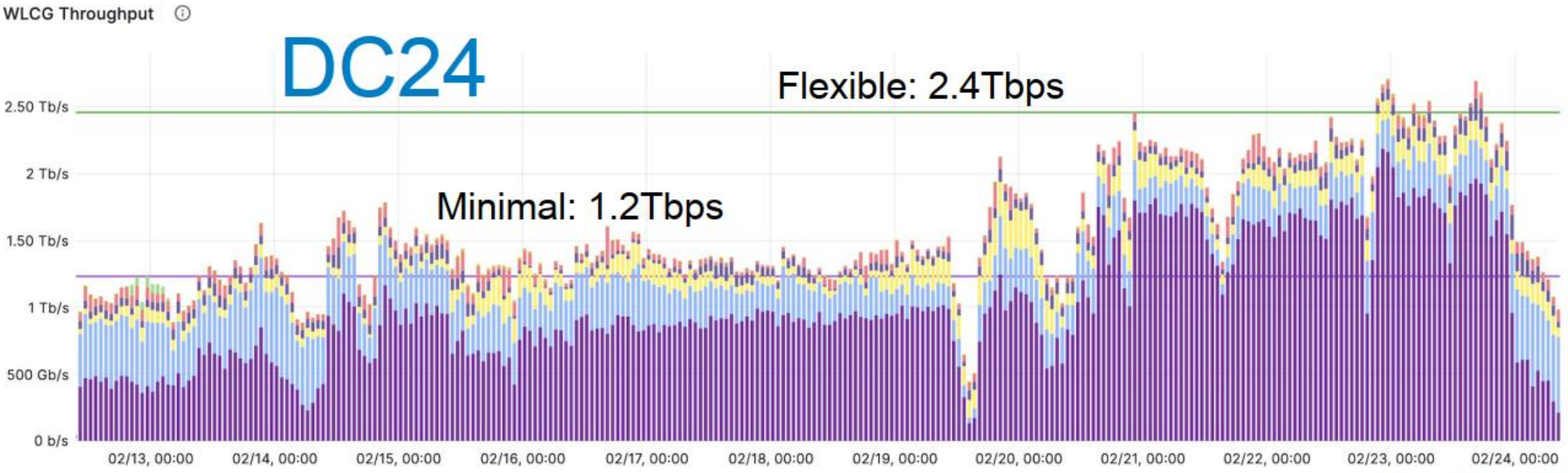
Nothing here about…

- Geopolitical risks
  - Undersea cables
  - Climate change mitigations
    - … but HEP is well set here
- Computing landscape changes
  - Cloud services
  - HPC centres

156. Integration of the Barcelona Supercomputing Center for CMS computing: towards large scale production
Dr Josep, Flix (CIEMAT - PIC)
5/11/23, 11:30 AM

CHEP 2023

… so what am I going to talk about?

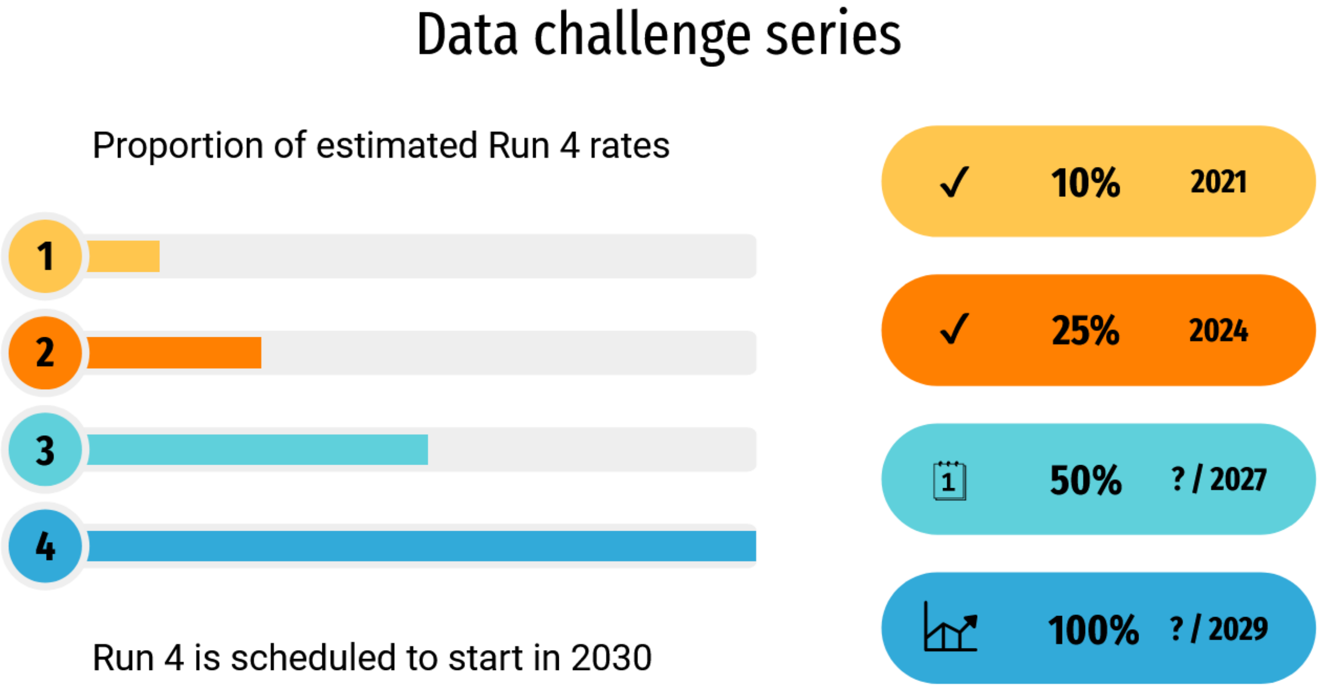


	max	avg	current
Data Challenge	2.19 Tb/s	1.02 Tb/s	211 Gb/s
atlas	625 Gb/s	304 Gb/s	567 Gb/s
alice xrootd	349 Gb/s	115 Gb/s	71.4 Gb/s
cms xrootd	191 Gb/s	67.4 Gb/s	42.7 Gb/s
cms	271 Gb/s	57.2 Gb/s	75.0 Gb/s
belle	38.9 Gb/s	9.45 Gb/s	17.1 Gb/s

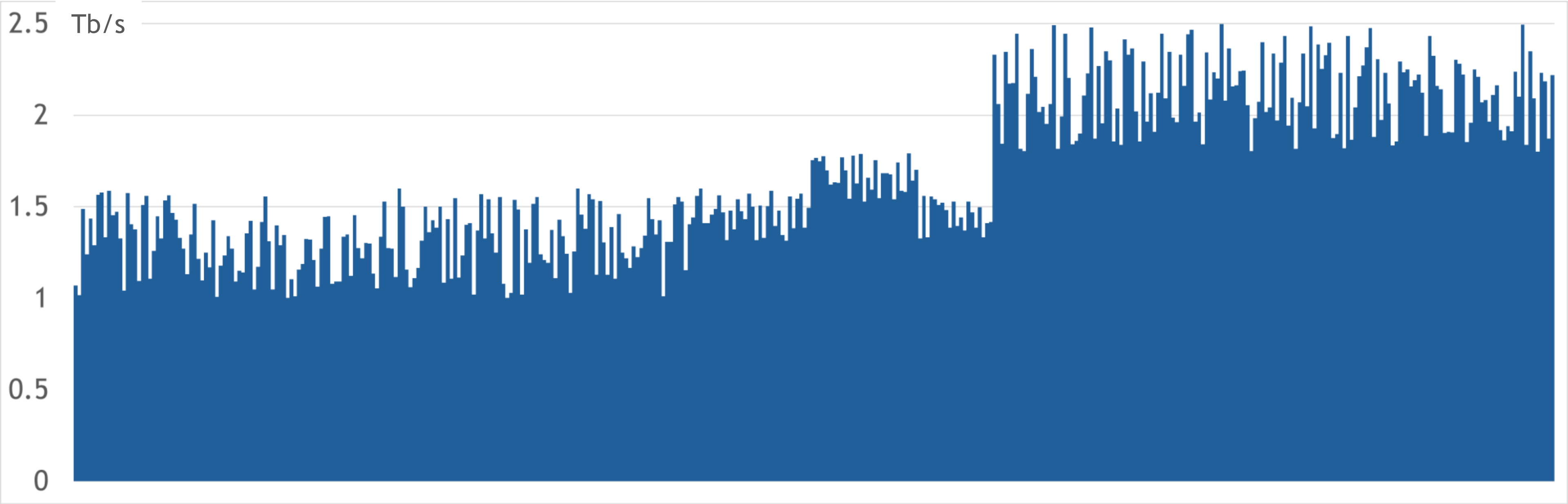
Katy in Hamburg: “There are other bottlenecks than network bandwidth”

CERN

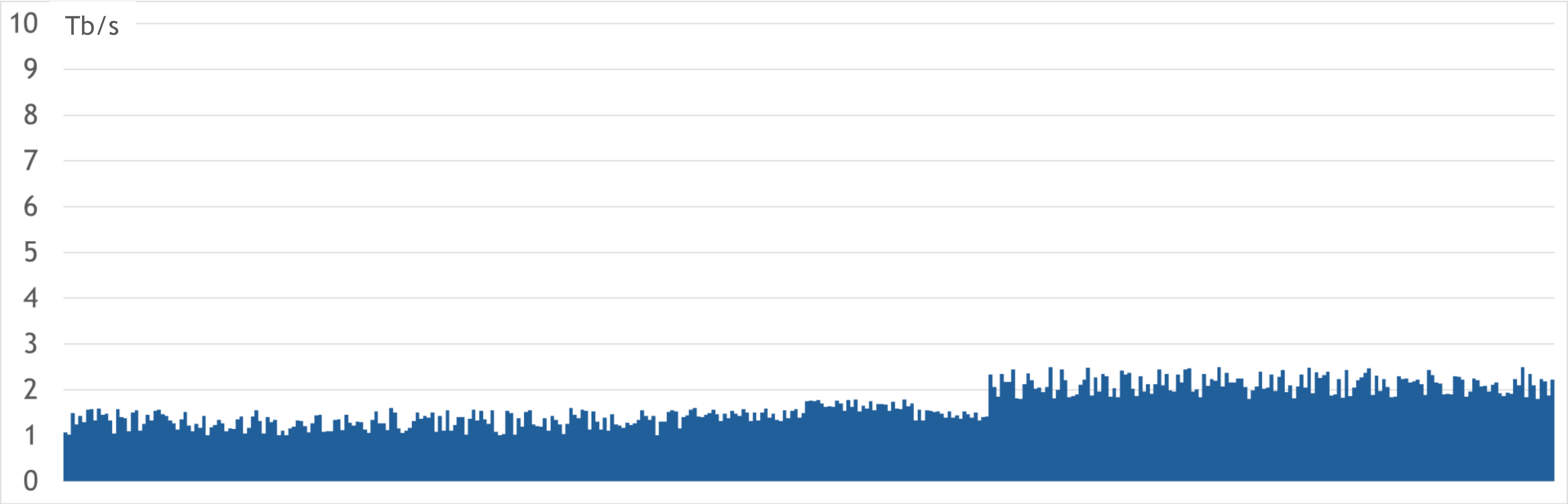
What next?


Data challenge series
Proportion of estimated Run 4 rates
1
2
3
4
Run 4 is scheduled to start in 2030
✓ 10% 2021
✓ 25% 2024
1 50% ? / 2027
100% ? / 2029

~DC24

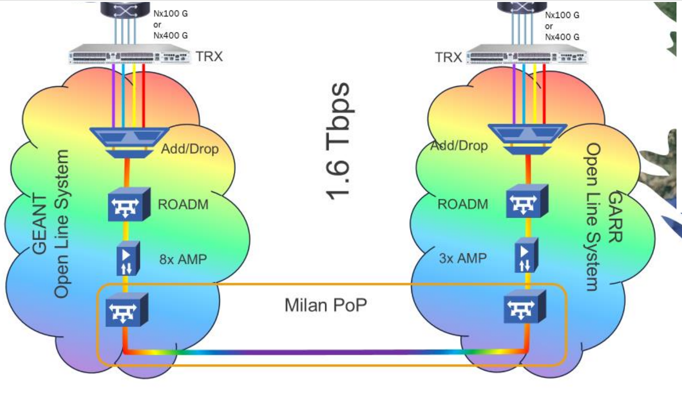


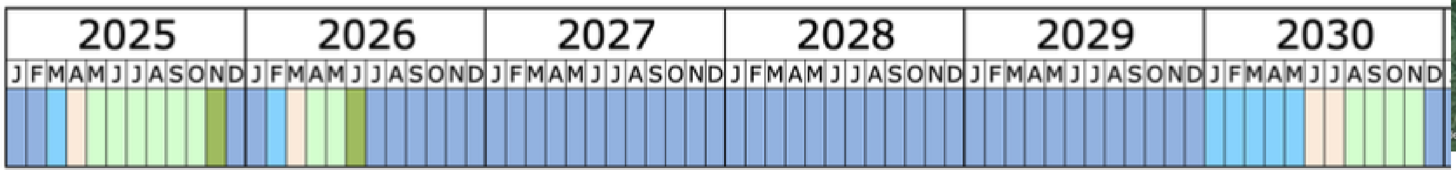
~DC24 using DC29's y-axis



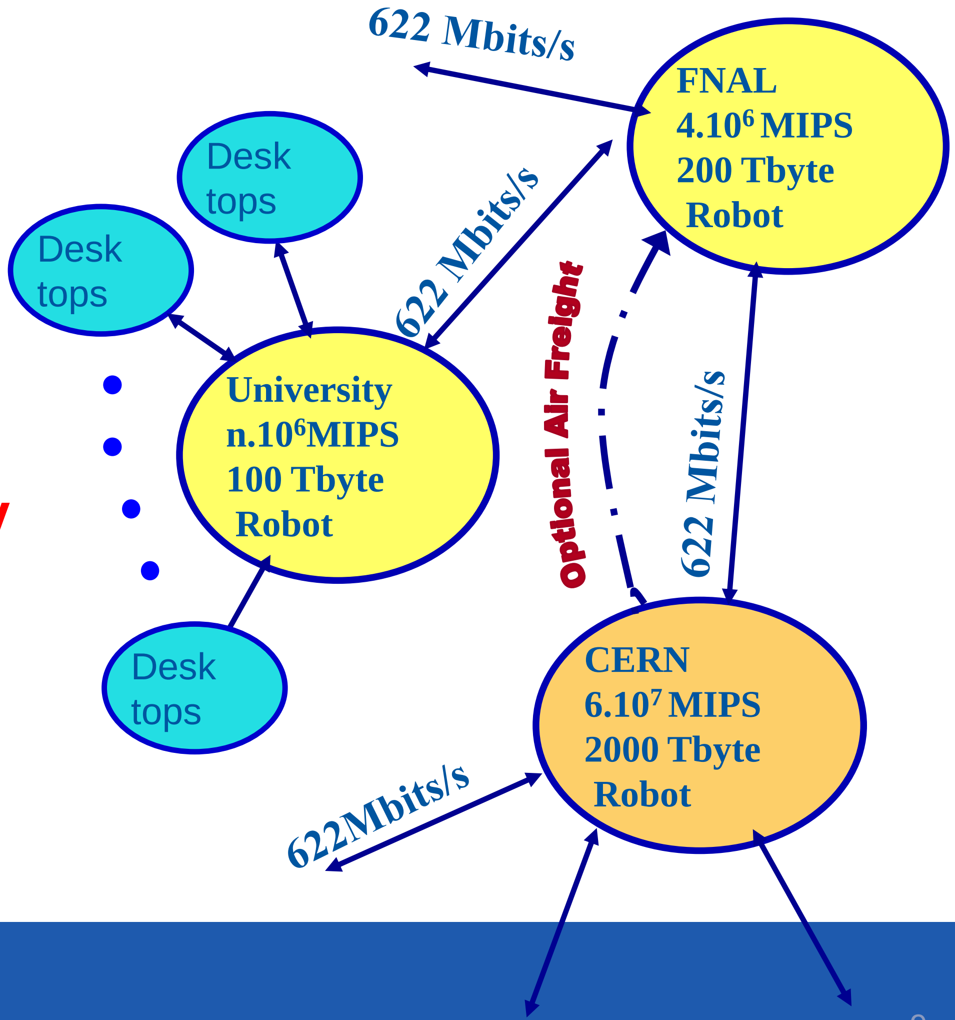
What challenges?

- Ethernet bandwidths keep increasing
- There are new te
- There is plenty

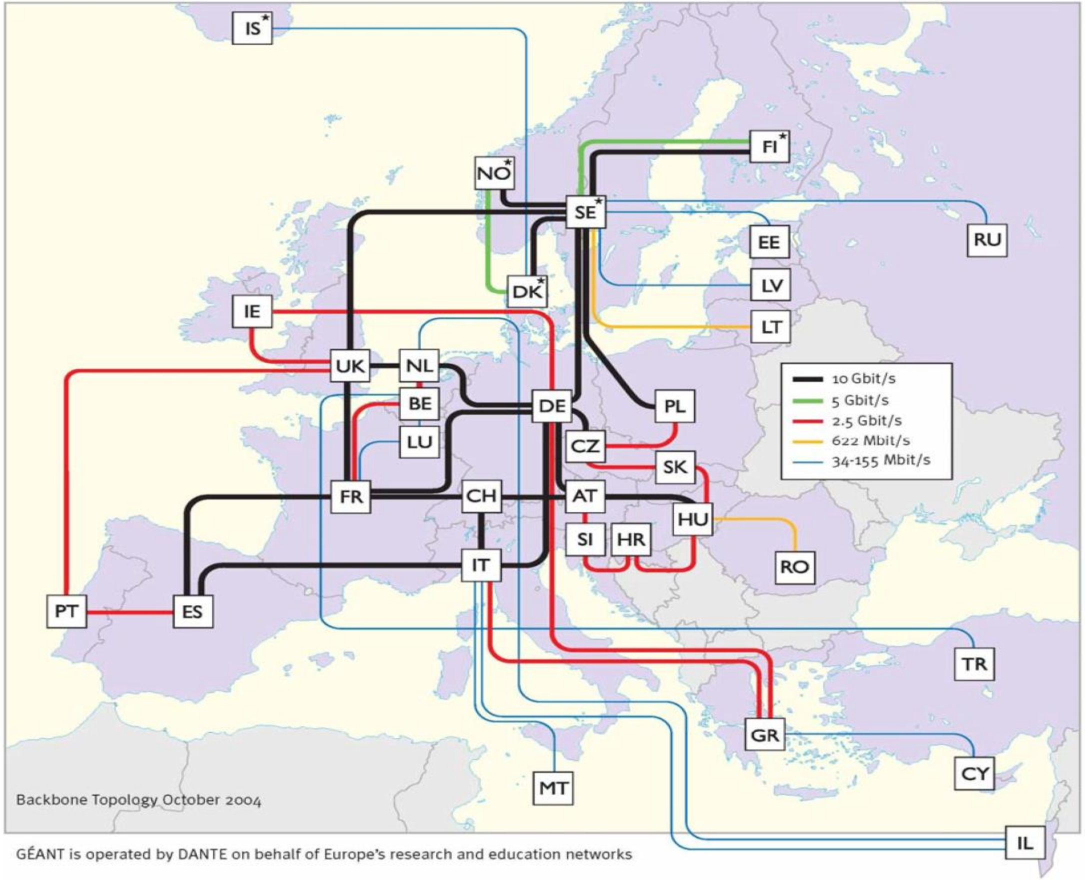




1999 predictions for 2006 network capability

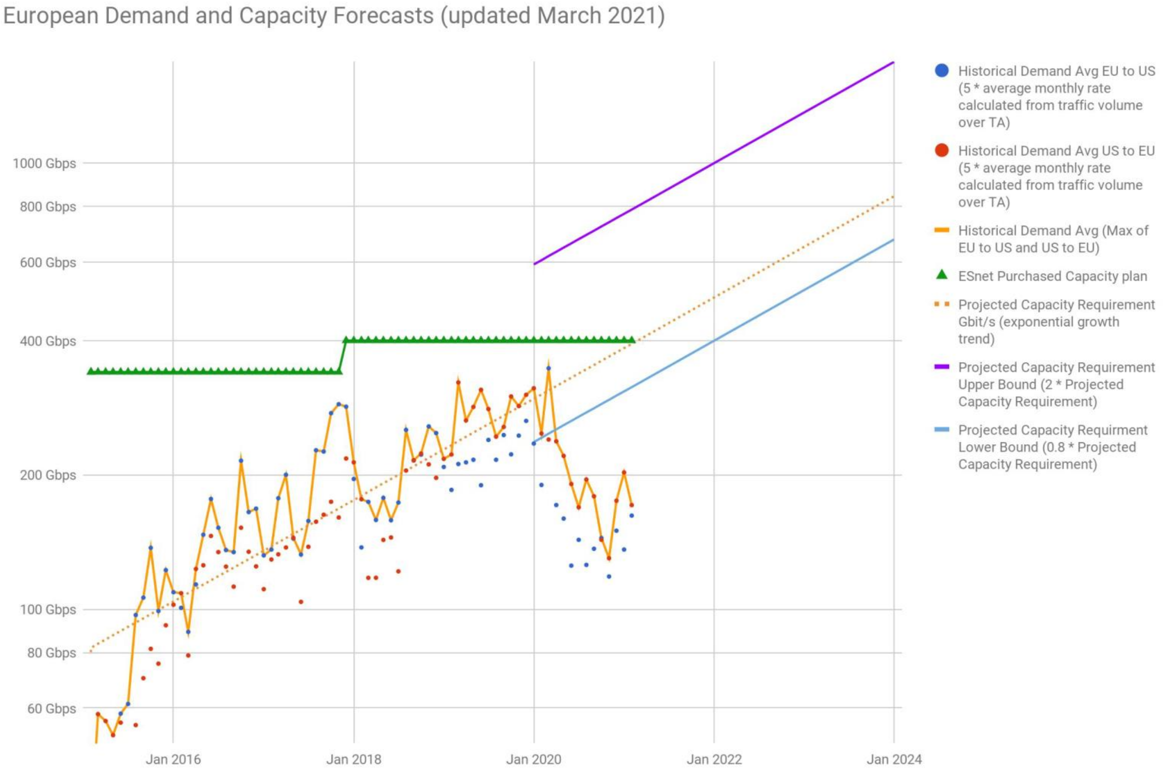


2005 reality:



What challenges?

- It hasn't always been plain sailing

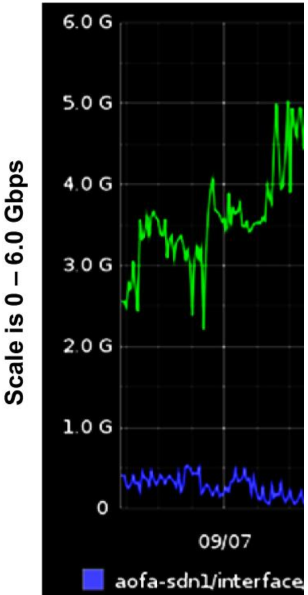


CERN

Early LHC data transfers worried NRENS!

The Nee

- GÉANT obs
  connection v
  - headed for
- ESnet obser



The Need for Traffic Engineering – Example

- This high degree of parallelism means that the largest host-host data flow rate is only about 2 Mbps, but in aggregate this data mover farm is doing 860 Mbps (seven day average) and has moved 65 TBytes of data
  - this also makes it hard to identify the sites involved by looking at all of the data flows at the peering point – nothing stands out as an obvious culprit
- THE ISSUE:
- This clever physics group is consuming 60% of the available bandwidth on the primary U.S. – Europe general R&E IP network link – for weeks at a time!
- This is obviously an unsustainable situation and this is the sort of thing that will force the R&E network operators to mark such traffic on the general IP network as scavenger to ensure other uses of the network

What challenges?

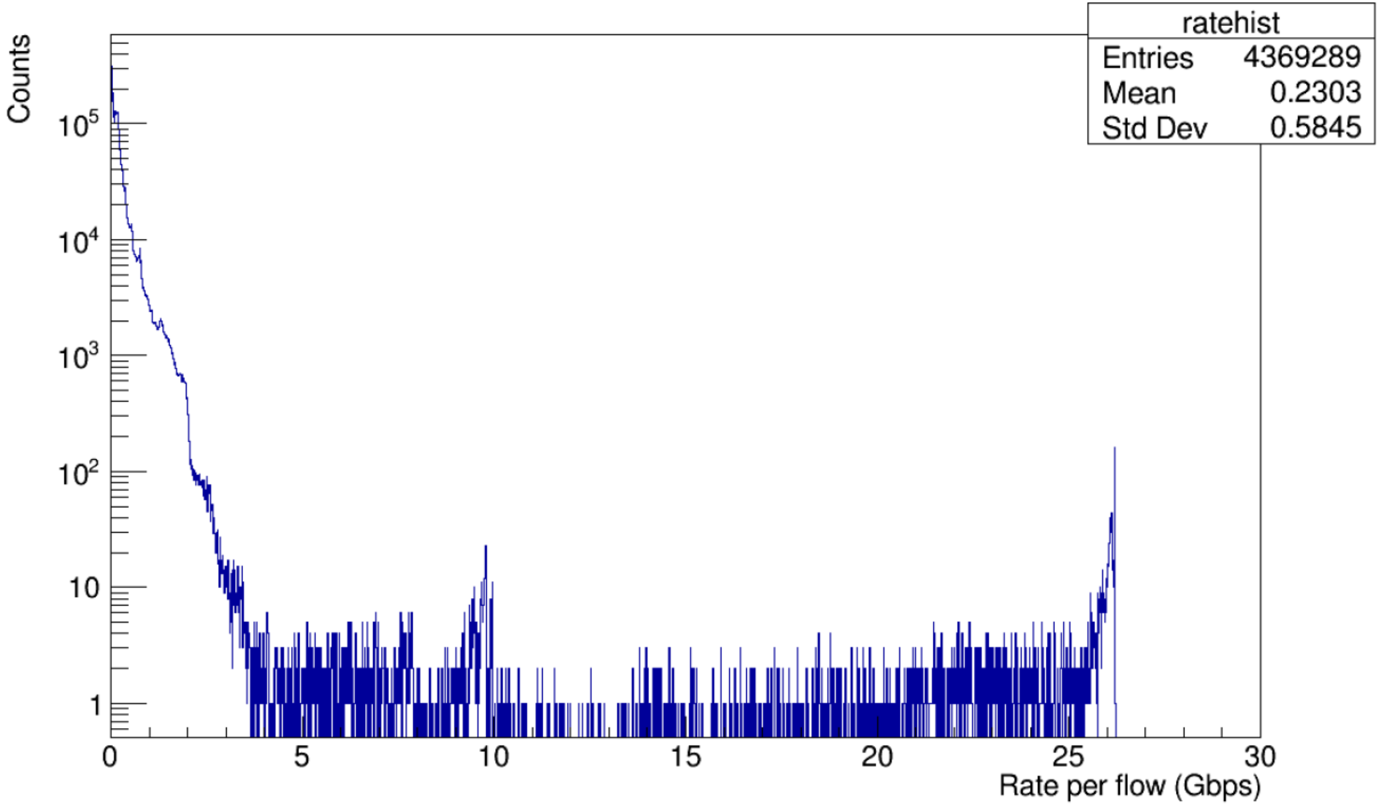
- It hasn’t always been plain sailing
- Our data transfers are (mostly) inefficient …
  - … and we don’t understand them

DC24 Bandwidth Per Flow

Each entry represents one complete flow

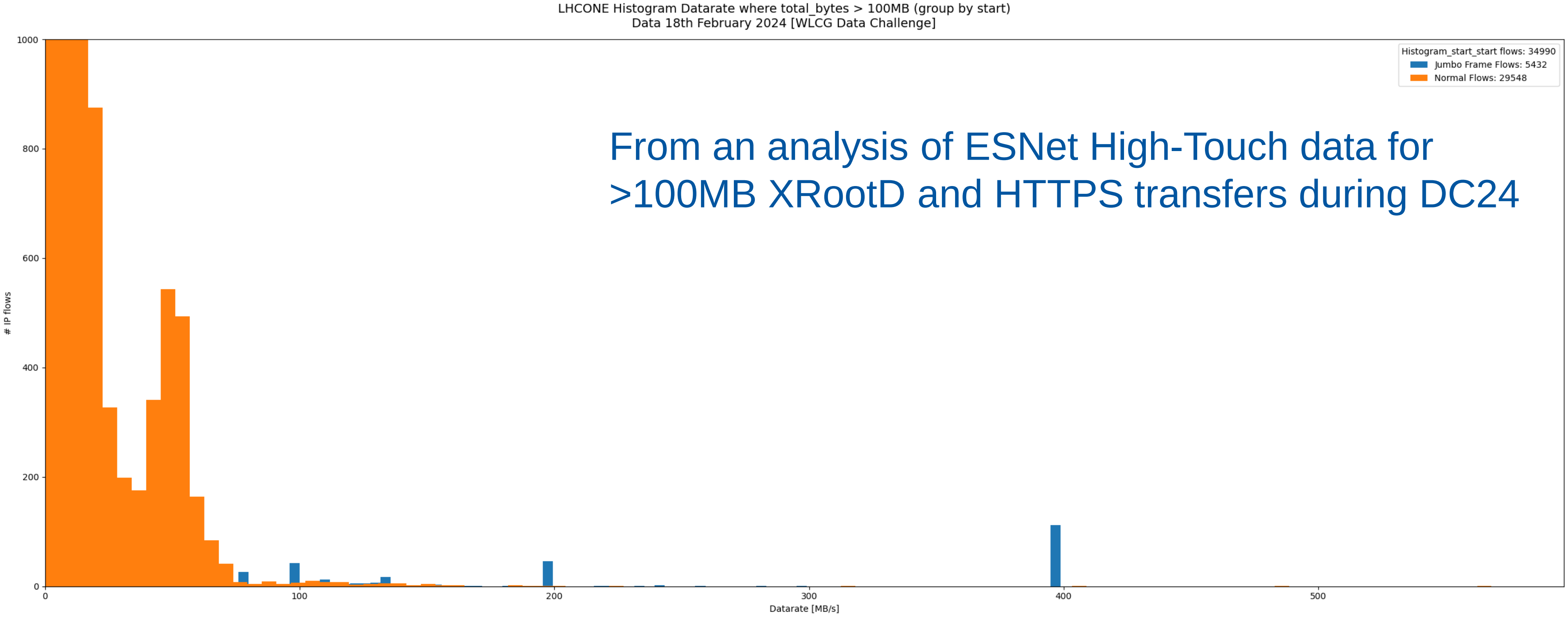
Peaked at very low bandwidth

Average is 230 Mbps



Plot by Tristan Sullivan from analysis of Esnet High-Touch data collection (Yatish Kumar)

Jumbo Frames



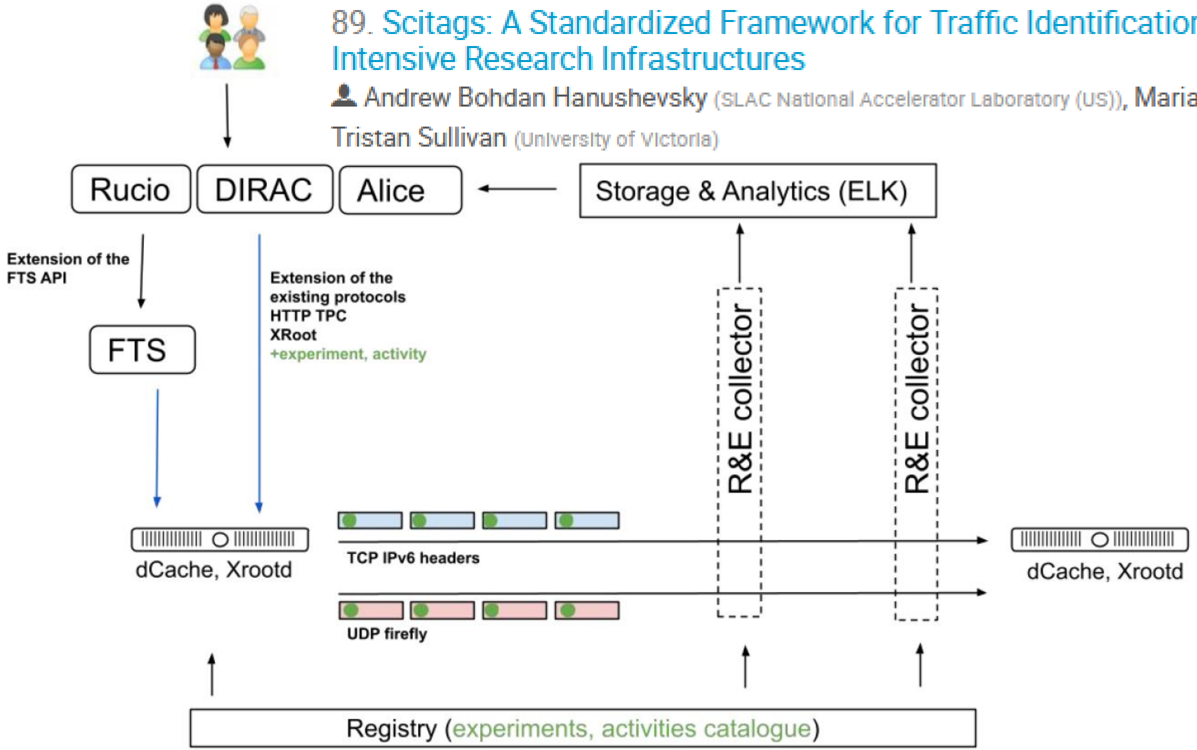
From an analysis of ESNet High-Touch data for >100MB XRootD and HTTPS transfers during DC24

Understanding Traffic Flows

How scitags work

89. Scitags: A Standardized Framework for Traffic Identification and Network Visibility in Data-Intensive Research Infrastructures

Andrew Bohdan Hanushevsky (SLAC National Accelerator Laboratory (US)), Marian Babik (CERN), Tristan Sullivan (University of Victoria)



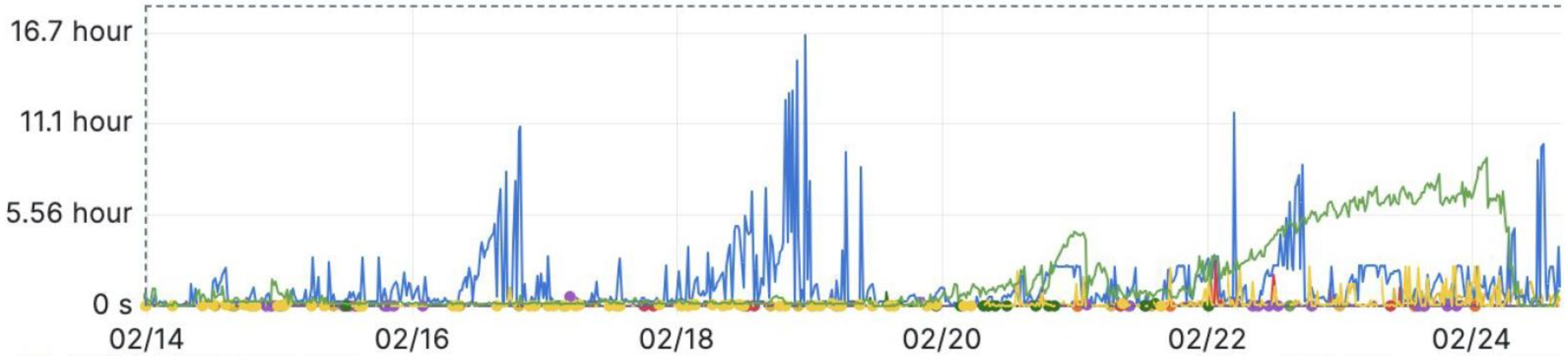
DC24: CERN EOS CMS

Max duration of flows split by Exp/Activity

First week had a lot of “fat” flows from production activity (but none from DC)

Second week was different, some DC flows took hours to finish

Max Duration Received per Exp/Act



cms - Functional Test	2 s	3.05 min	18.8 min
cms - Debug	0 s	2.93 min	52.5 min
cms - Data rebalancing	0 s	12.4 min	2.42 hour
cms - Data Challenge	1 s	1.59 hour	8.99 hour

Technical Spec for Packet Marking

Packet Marking via the use of the IPv6 Flow Label

```
                                   IPv6 Header
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Traffic Class |           Flow Label                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Payload Length        |  Next Header  |   Hop Limit   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

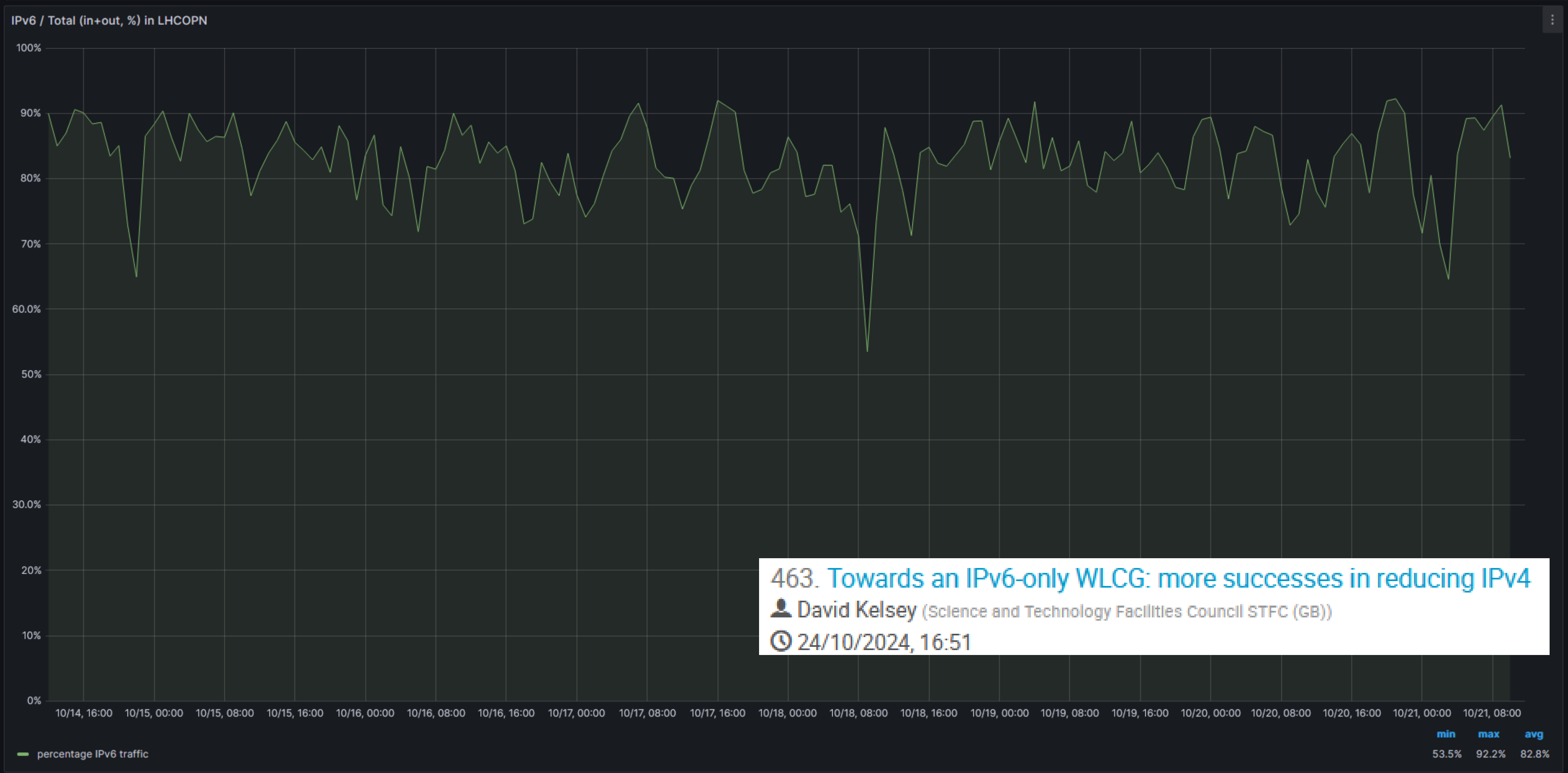
```
                          Flow Label
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|01|02|03|04|05|06|07|08|09|10|11|12|13|14|15|16|17|18|19|20|
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|E | E| C| C| C| C| C| C| C| C| C| E| A| A| A| A| A| A| E| E|
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
```

- (C) Community identifier: "Who are you affiliated with?"
- (A) Activity identifier: "What are you doing within your community?"
- (E) Entropy bits sprinkled throughout

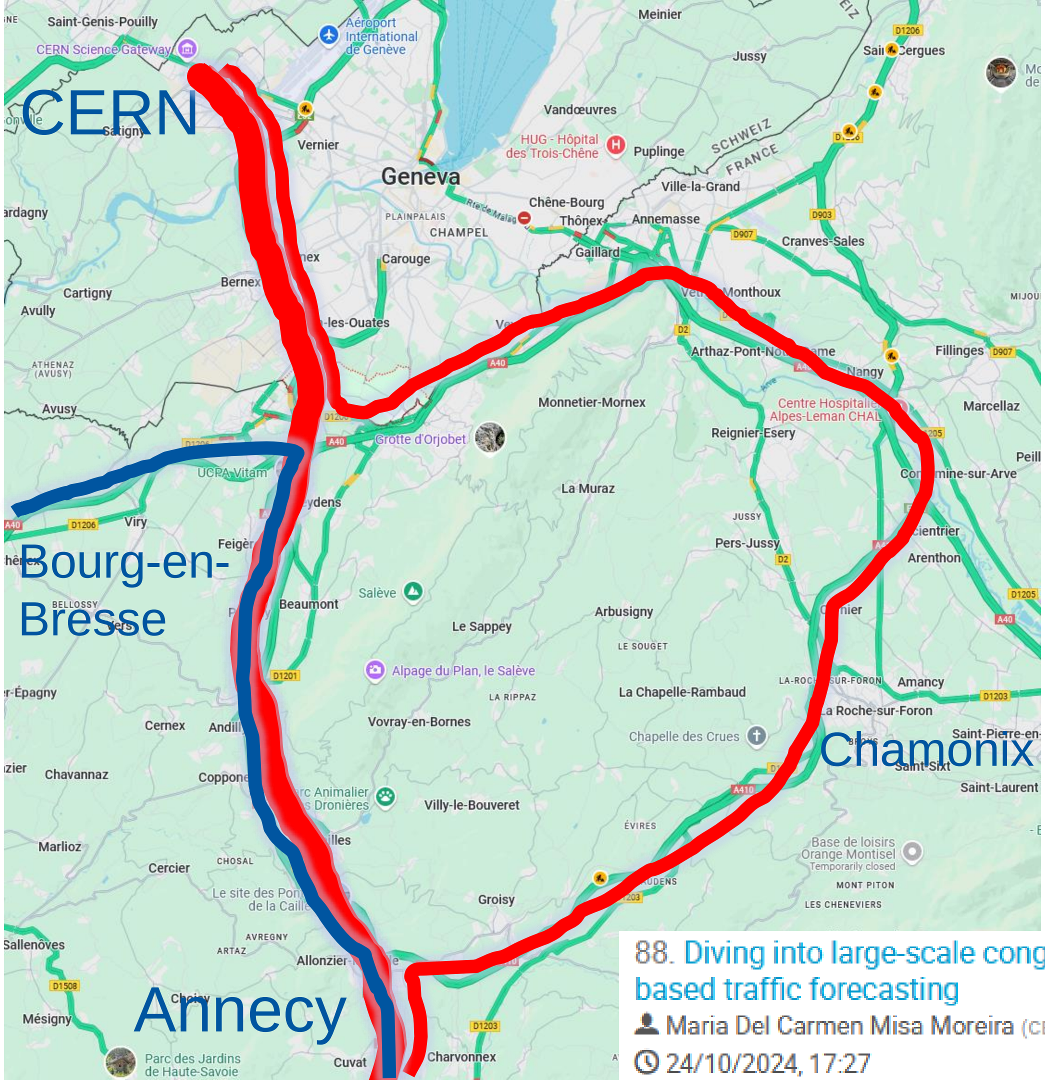
IETF RFC-Informational Draft is available with more details

Started exploring HbH option as an alternative (eBPF-PDM, eBPF-extHeaders)

But this only works with IPv6!



Managing Traffic Flows: NOTED

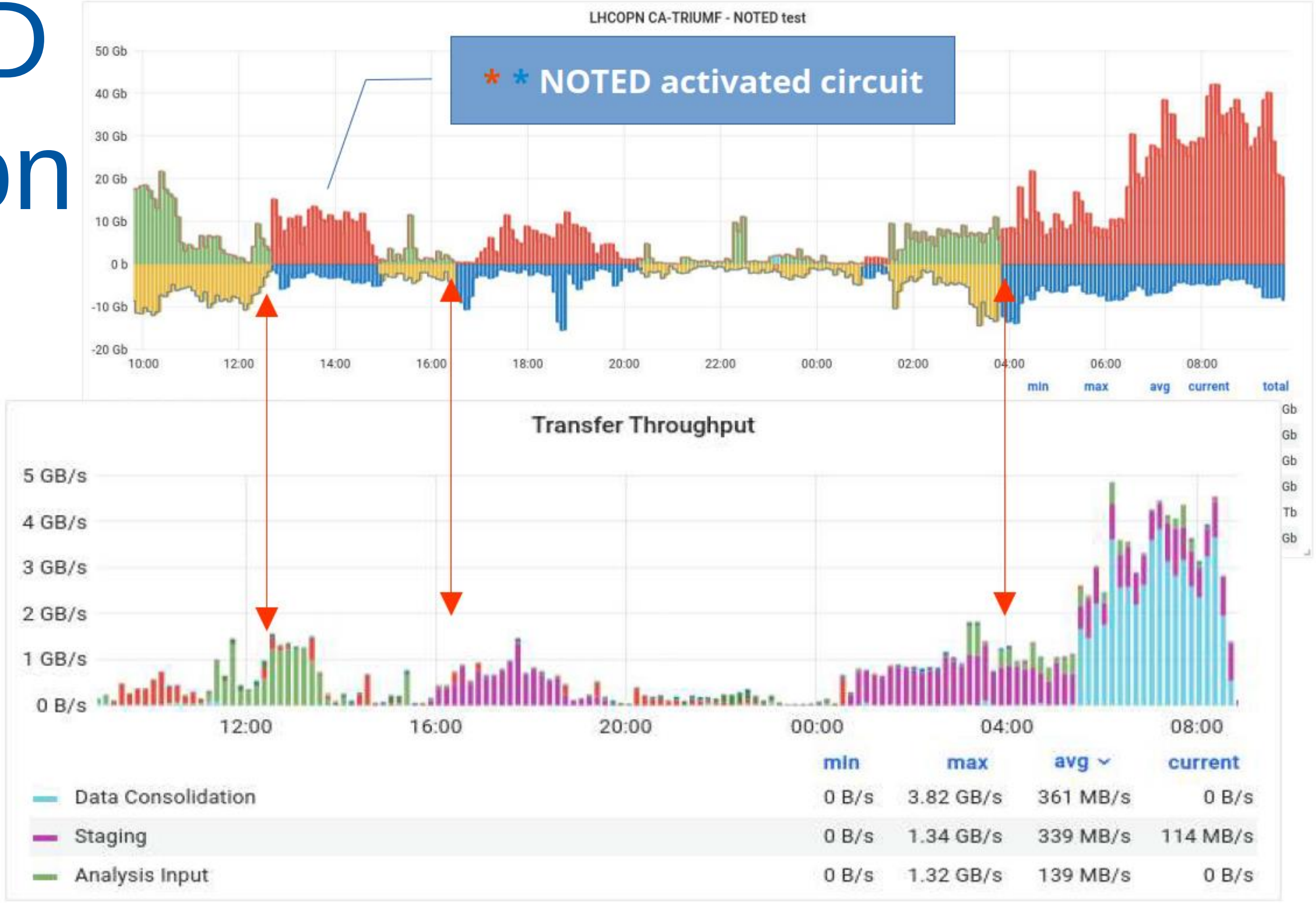


88. Diving into large-scale congestion with NOTED as a network controller and machine learning-based traffic forecasting

Maria Del Carmen Misa Moreira (CERN)

24/10/2024, 17:27

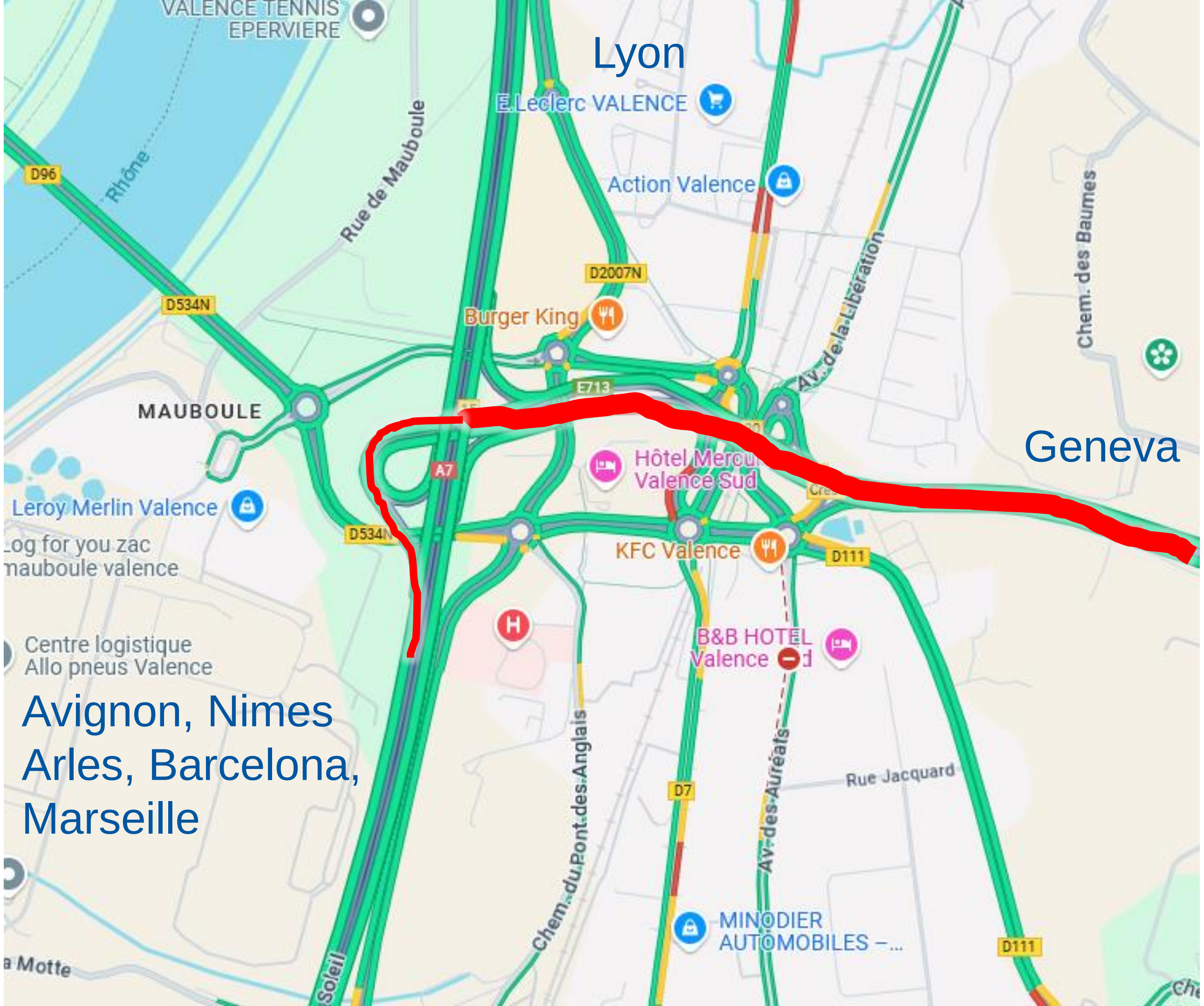
NOTED in action



Packet pacing

Congestion Protocols

BBRv3



What challenges?

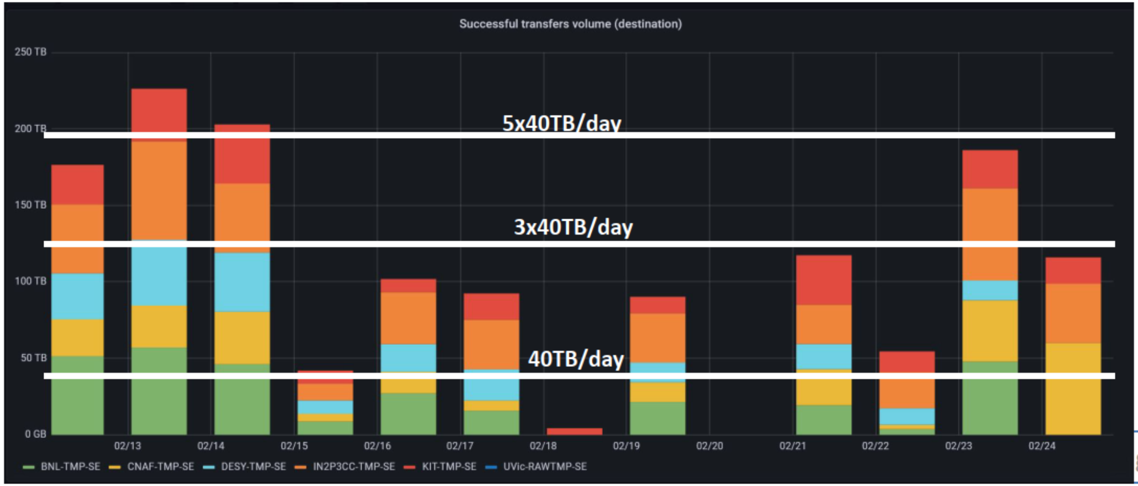
- It hasn't always been plain sailing
- Our data transfers are (mostly) inefficient …
  - … and we don't understand them
- More global data-intensive science collaborations
  - LHC traffic won't always be dominant
  - Life will be more complicated

Belle II Data Challenge 2024 within WLCG DC24

Main goal: Emulate data transfer conditions in a Belle II high-lumi scenario.

Transfers from KEK to RAW Data Centers according to our distribution schema (estimated 40TB/day to be distributed with the following share 30%BNL, 20%CNAF, 15% IN2P3CC, 15%UVic, 10%DESY, 10%KIT)

- Min - The target speed to achieve is 3x3.7Gbit/s = **11.1 Gbit/s**
- Max - The target speed to achieve is 5x3.7Gbit/s = **18.5 Gbit/s**



Tests fully succeeded. Results will be present at Poster Session of CHEP2024 in Krakow

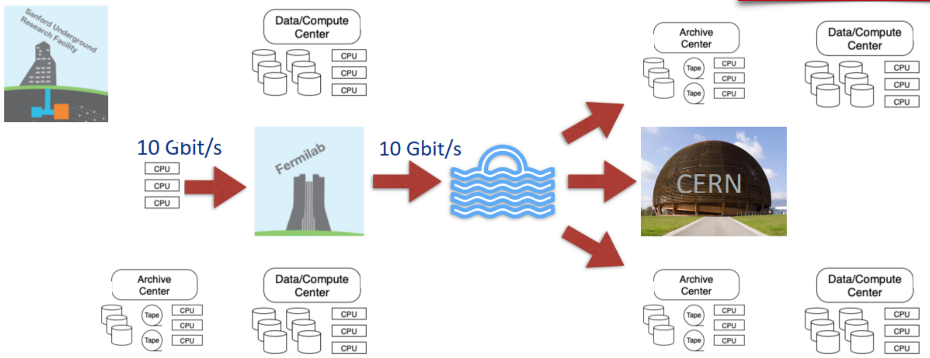
https://indico.cern.ch/event/1338689/contributions/6010887/

Courtesy Silvio Pardi
University Federico II and INFN, Naples

DEEP UNDERGROUND NEUTRINO EXPERIMENT

DUNE Involvement in WLCG Data Challenge 24

“FD” Raw Data
archival storage



not to scale, not a technical design, it’s just a cartoon

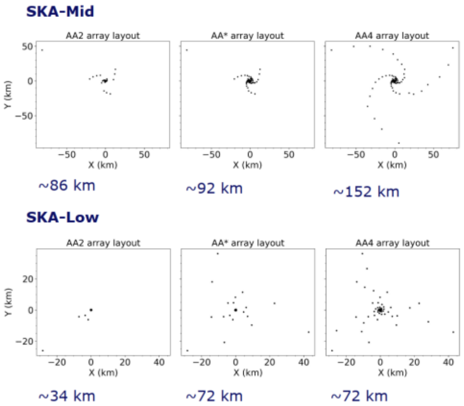
- 30 PB/year limit to archival storage for the Far Detector
  - translates to 10 Gbit/s from SURF to FNAL
  - replicate that “FD” raw data to archival storage facilities around the world
  - replicate the “FD” raw data to disk storage elements around the world for prompt access from compute elements
- job processing of the raw data drastically reduces the data volume of derived datasets transferred back to RSEs

Courtesy Michael Kirby, FNAL

Staged delivery: timeline and layouts

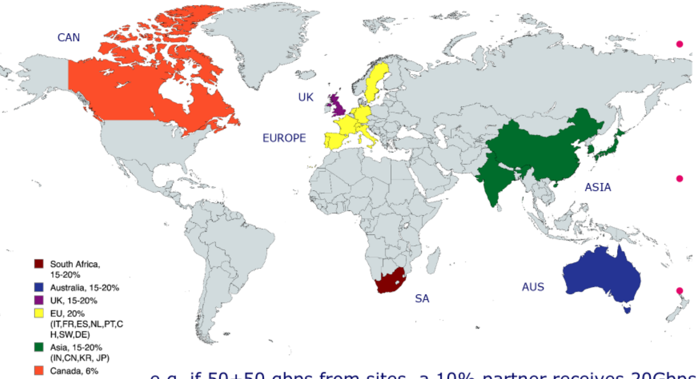
Milestone event (earliest)		SKA-Mid (end date)	SKA-Low (end date)
AA0.5	4 dishes 6 stations	2025 May	2024 Nov
AA1	8 dishes 18 stations	2026 Apr	2025 Nov
AA2	64 dishes 64 stations	2027 Mar	2026 Oct
AA*	144 dishes 307 stations	2027 Dec	2028 Jan
Operations Readiness Review		2028 Apr	2028 Apr
AA4	197 dishes 512 stations	TBD	TBD

Not just building up in terms of array size, but also capabilities



Courtesy Ian Collier, Rosie Bolton & Shari Breen

SKA Regional Centre Broad Distribution: Fair Share (if ~50 Gbps per SKAO site)



- Roughly, 6 global zones of equivalent size (Canada smaller)
- **Distribute two base copies** of each data product to different countries, and perhaps insist to different regions
- Average incoming rate per (20%) region not more than 2x20 Gbit/s = 40Gbit/s (~2x6 Gbit/s for Canada)
- **Modelling assumes average 100 Gbit/s out of SA and AUS**

e.g. if 50+50 gbps from sites, a 10% partner receives 20Gbps data (200 TBytes per day, 70 PBytes per year)

SKA estimated data rates*

***these numbers should be used as a guide only - email Shari.Breen@skao.int for further information about ongoing work**

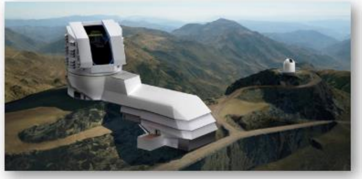
- Numbers refer to data to be delivered to the science community via the SRCNet

Milestone	Year	Primary activity	Estimated data rate	
			Low	Mid
AA2 • 64 Mid dishes • 64 Low stations	2026 - 2027	**Science Verification** - observed in dedicated ~week long blocks + single observations interspersed throughout. A higher rate of raw data products will be included at this stage.	1.5 PB/week^ 20 Gbps	2 PB/week^ 27 Gbps
AA* • 144 Mid dishes • 307 Low stations	2027 - 2029	**Science Verification** - observed in dedicated ~week long blocks + single observations interspersed throughout. A higher rate of raw data products will be included at this stage.	5 PB/week^ 66 Gbps	9 PB/week^ 119 Gbps
AA* • 144 Mid dishes • 307 Low stations	2029 +	**Operations** - Observation cycles, starting with shared risk observing, building to successful science observations ~90% of the time	173 PB/year 44 Gbps	280 PB/year 72 Gbps
Target is to deliver the SKA Baseline Design but the details of this transition between AA* and AA4 are TBD				
AA4 • 197 Mid dishes • 512 Low stations	2030 +	**Operations** - full SKA baseline design	216 PB/year 55 Gbps	400 PB/year 100 Gbps

^Data rates refer to dedicated Science Verification observing weeks, not an average over a year

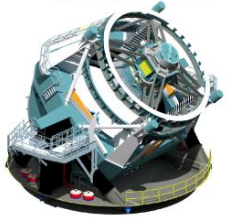
Observatory overview

SITE



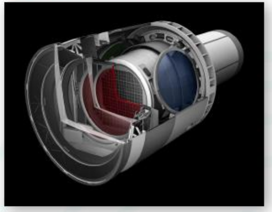
southern hemisphere | 2647m a.s.l. | stable air | clear sky | dark nights | good infrastructure

TELESCOPE



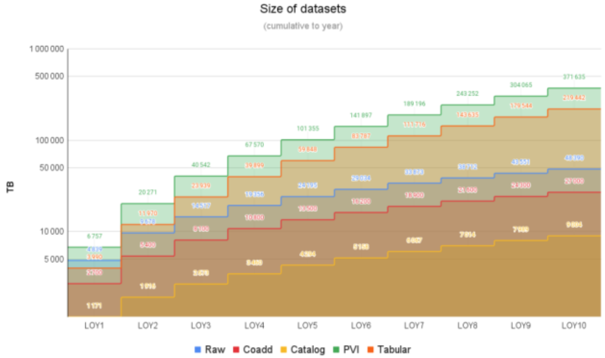
main mirror ⌀ 8.4 m (effective 6.4 m) | large aperture: f/1.234 | wide field of view | 350 ton | compact | to be repositioned about 3M times over 10 years of operations

CAMERA



3.2 G pixels | ⌀ 1.65 m | 3.7 m long | 3 ton | 3 lenses | 3.5° field of view | 9.6 deg² | 6 filters *ugrizy* | 320-1050 nm

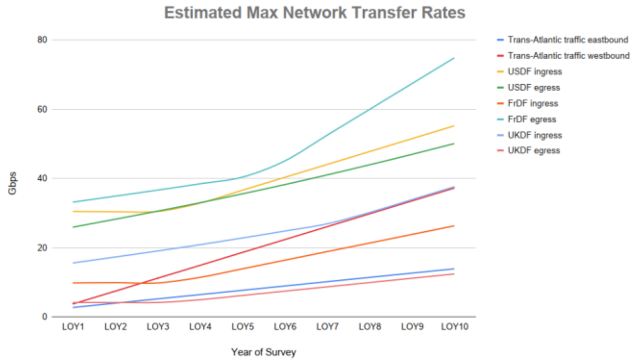
Cumulative data volume



~0.5 EB of data by the end of the survey by 2035

raw image data (~50 PB)

Projected data transfer rates

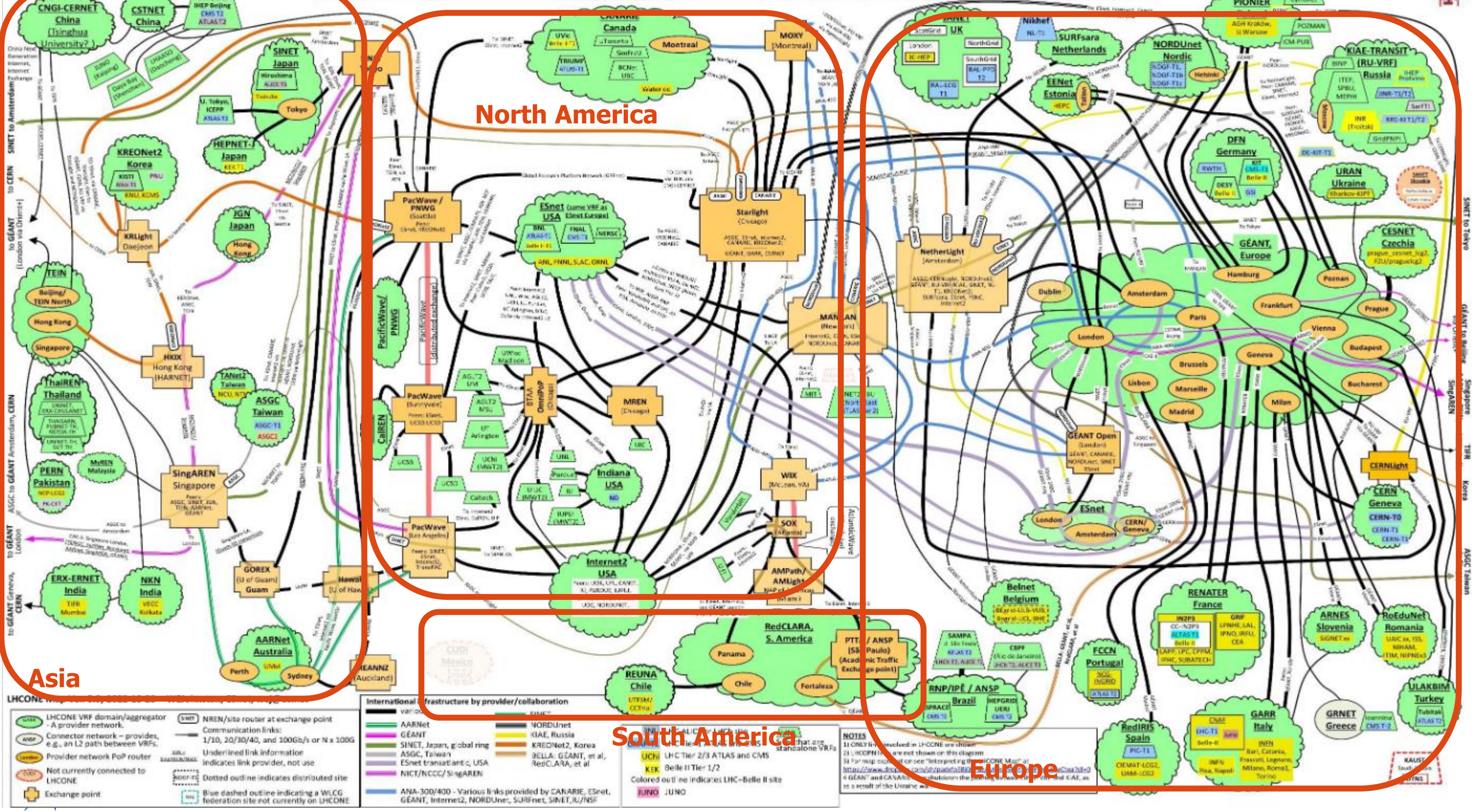


These estimations make some assumptions that we may need to revisit as we learn how data reprocessing will proceed in real-life conditions

Adapted from R. Dubois, Rubin Observatory

More global science – More complication

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



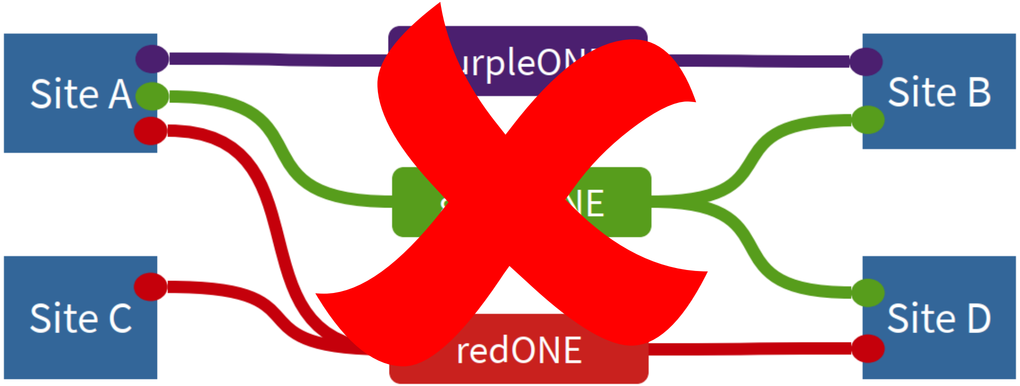
LHCONE in a multi-science world

multiple “LHCONEs”

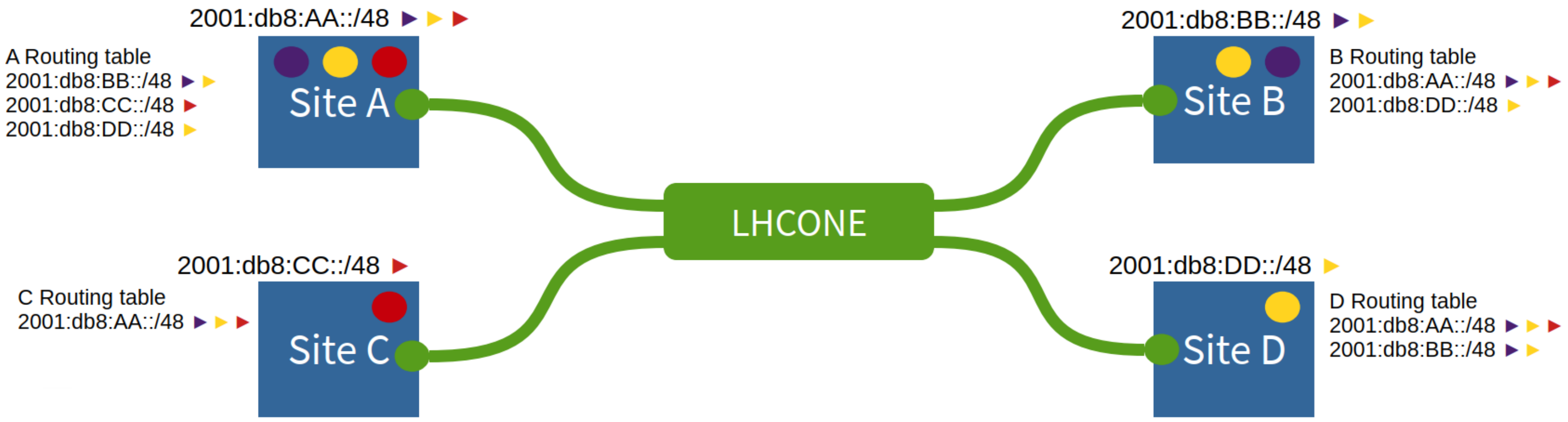
Each site joins only the VPNs of the groups it is collaborating with (e.g. ATLAS-ONE, CMS-ONE, DUNE-ONE, BelleII-ONE…)

- **Major Benefit**: reduced exposure of data-centre/Science-DMZ to other sites

- **Major Challenge**: how to correctly route traffic into VPNs at sites that join several of them? Operational complexity



Network Routing Wizadry



Summary

- Wide area networking will continue to deliver quality services for the HEP community into the HL-LHC era.
- But
  - we need to (re)learn how to transfer data efficiently,
  - we need to understand and perhaps manage traffic flows,
  - IPv4 has to go, and
  - life will be more complicated in a multi-science world.

CERN