Research Infrastructures and Networks

Xavier Espinal (CERN)

13th GÉANT-SIG Next Generation Networks, 7th Dec 2023

HL-LHC

- HL-LHC network traffic will be dominated by a) RAW data export from CERN to the T1s and b) data reprocessing activities.
- ATLAS and CMS experiments will both produce ~350 PB of RAW data per year.
- Traffic from CERN to the T1s for RAW data export will be ~400 Gbps per experiment on quasi-real time. Estimate 7M seconds/year of LHC data taking.
- Estimate of extra 100 Gbps per experiment for other data formats (eg. user analysis oriented)
- Alice and LHCb estimates are of 100Gbps per experiment.

HL-LHC

(estimate) Network bandwidth needs per T1 region

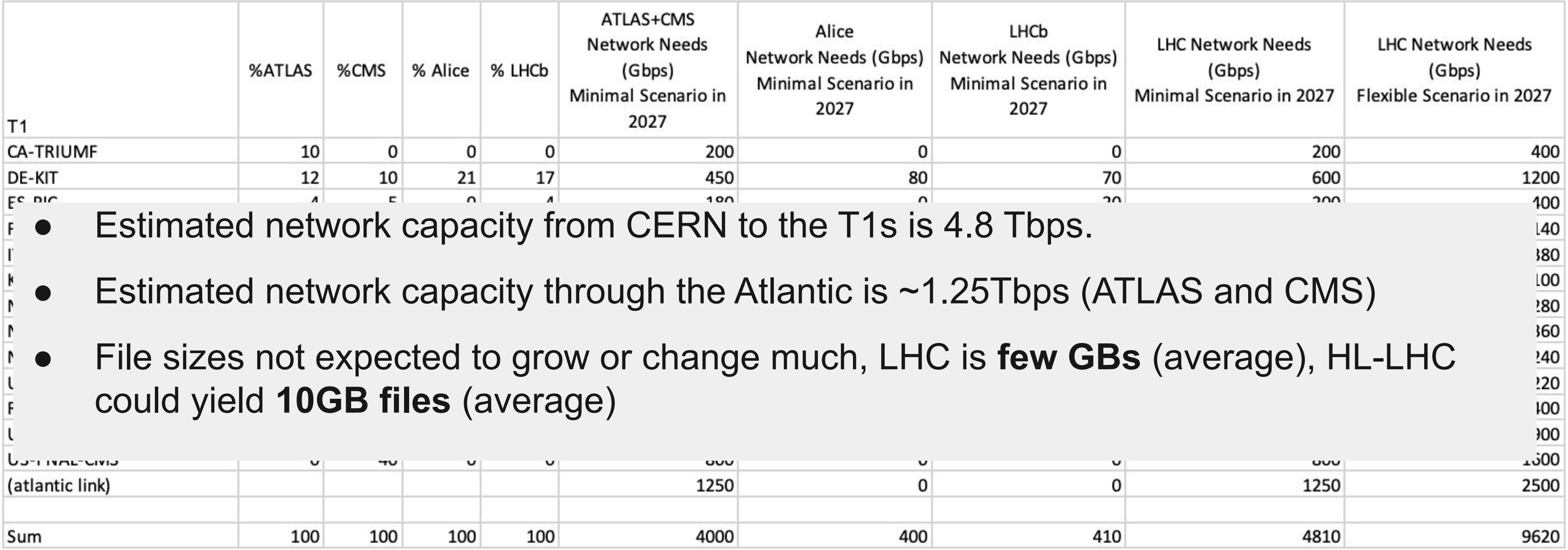
x4 included (to deal with burst and overprovision)

T1	%ATLAS	%CMS	% Alice	% LHCb	ATLAS+CMS Network Needs (Gbps) Minimal Scenario in 2027	Alice Network Needs (Gbps) Minimal Scenario in 2027	LHCb Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027
CA-TRIUMF	10	0	0	0	200	0	0	200	400
DE-KIT	12	10	21	17	450	80	70	600	1200
ES-PIC	4	5	0	4	180	0	20	200	400
FR-CCIN2P3	13	10	14	15	450	60	60	570	1140
IT-INFN-CNAF	9	15	26	24	480	110	100	690	1380
KR-KISTI-GSDC	0	0	12	0	0	50	0	50	100
NDGF	6	0	8	0	110	30	0	140	280
NL-T1	7	0	3	8	140	10	30	180	360
NRC-KI-T1	3	0	13	5	50	50	20	120	240
UK-T1-RAL	15	10	3	27	490	10	110	610	1220
RU-JINR-T1	0	10	0	0	200	0	0	200	400
US-T1-BNL	23	0	0	0	450	0	0	450	900
US-FNAL-CMS	0	40	0	0	800	0	0	800	1600
(atlantic link)					1250	0	0	1250	2500
Sum	100	100	100	100	4000	400	410	4810	9620

Input taken from the DC2024 WLCG workshop: https://indico.cern.ch/event/1307338/

HL-LHC

(estimate) Network bandwidth needs per T1 region

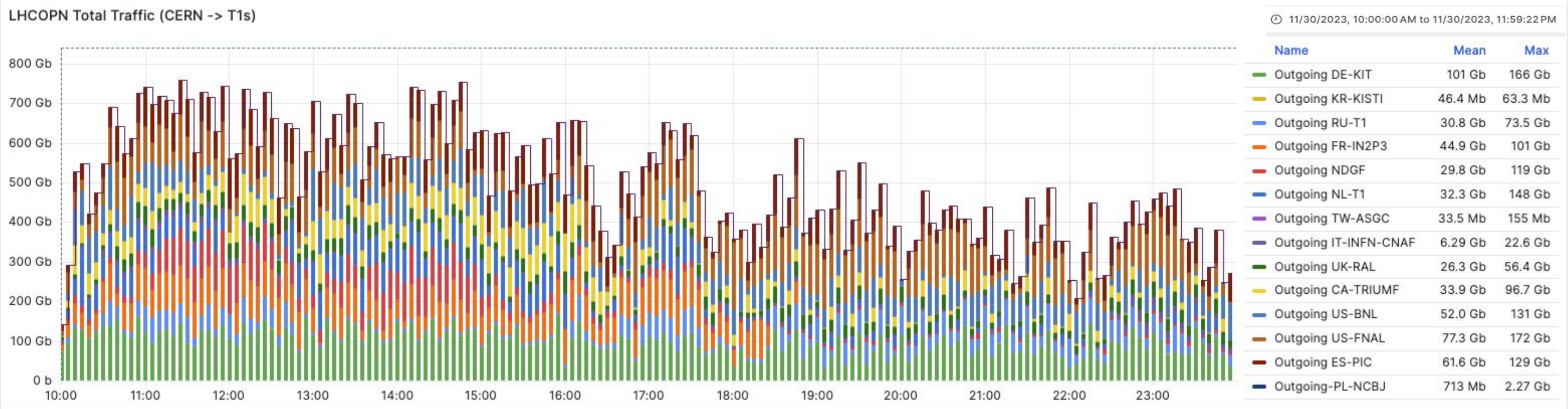


T1	%ATLAS	%CMS	% Alice	% LHCb	ATLAS+CMS Network Needs (Gbps) Minimal Scenario in 2027	Alice Network Needs (Gbps) Minimal Scenario in 2027	LHCb Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027
CA-TRIUMF	10	0	0	0	200	0	0	200	400
DE-KIT	12	10	21	17	450	80	70	600	1200
ES-PIC	4	5	0	4	180	0	20	200	400
US-FNAL-CMS	0	40	0	0	800	0	0	800	1600
(atlantic link)					1250	0	0	1250	2500
Sum	100	100	100	100	4000	400	410	4810	9620

- Estimated network capacity from CERN to the T1s is 4.8 Tbps.
- Estimated network capacity through the Atlantic is ~1.25Tbps (ATLAS and CMS)
- File sizes not expected to grow or change much, LHC is **few GBs** (average), HL-LHC could yield **10GB files** (average)

Input taken from the DC2024 WLCG workshop: https://indico.cern.ch/event/1307338/

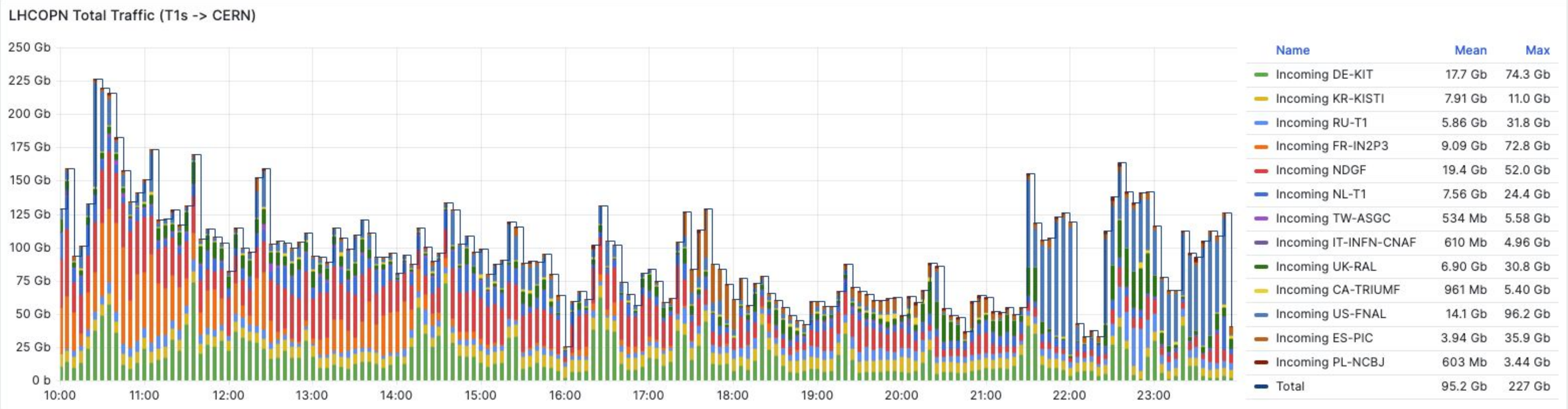
Pre-Data Challenge + Production (30-Nov-2023)



LHCOPN Total Traffic (CERN -> T1s)

11/30/2023, 10:00:00 AM to 11/30/2023, 11:59:22 PM

Name	Mean	Max
Outgoing DE-KIT	101 Gb	166 Gb
Outgoing KR-KISTI	46.4 Mb	63.3 Mb
Outgoing RU-T1	30.8 Gb	73.5 Gb
Outgoing FR-IN2P3	44.9 Gb	101 Gb
Outgoing NDGF	29.8 Gb	119 Gb
Outgoing NL-T1	32.3 Gb	148 Gb
Outgoing TW-ASGC	33.5 Mb	155 Mb
Outgoing IT-INFN-CNAF	6.29 Gb	22.6 Gb
Outgoing UK-RAL	26.3 Gb	56.4 Gb
Outgoing CA-TRIUMF	33.9 Gb	96.7 Gb
Outgoing US-BNL	52.0 Gb	131 Gb
Outgoing US-FNAL	77.3 Gb	172 Gb
Outgoing ES-PIC	61.6 Gb	129 Gb
Outgoing-PL-NCBJ	713 Mb	2.27 Gb

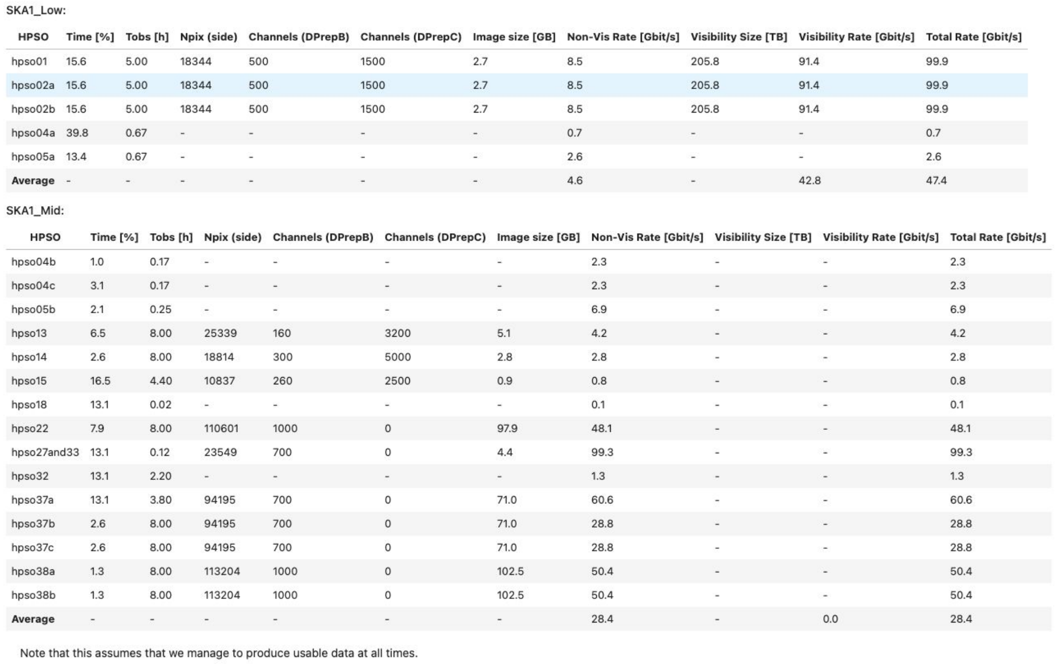


LHCOPN Total Traffic (T1s -> CERN)

Name	Mean	Max
Incoming DE-KIT	17.7 Gb	74.3 Gb
Incoming KR-KISTI	7.91 Gb	11.0 Gb
Incoming RU-T1	5.86 Gb	31.8 Gb
Incoming FR-IN2P3	9.09 Gb	72.8 Gb
Incoming NDGF	19.4 Gb	52.0 Gb
Incoming NL-T1	7.56 Gb	24.4 Gb
Incoming TW-ASGC	534 Mb	5.58 Gb
Incoming IT-INFN-CNAF	610 Mb	4.96 Gb
Incoming UK-RAL	6.90 Gb	30.8 Gb
Incoming CA-TRIUMF	961 Mb	5.40 Gb
Incoming US-FNAL	14.1 Gb	96.2 Gb
Incoming ES-PIC	3.94 Gb	35.9 Gb
Incoming PL-NCBJ	603 Mb	3.44 Gb
Total	95.2 Gb	227 Gb

SKA

- Combined SKA expected **traffic derived from data products is 200Gbps**:
    - Considering both observatories SKA Low in Australia (100Gbps) and SKA High in South Africa (100Gbps)
- SKAO **data volume estimate is of 700 PB/year**, 2 scenarios envisaged: data pre-placement or move compute to data
- SKA full operations expected by 2028, but data transfers and live system from **2026**



SKA1_Low:

HPSO	Time [%]	Tobs [h]	Npix (side)	Channels (DPrepB)	Channels (DPrepC)	Image size [GB]	Non-Vis Rate [Gbit/s]	Visibility Size [TB]	Visibility Rate [Gbit/s]	Total Rate [Gbit/s]
hpso01	15.6	5.00	18344	500	1500	2.7	8.5	205.8	91.4	99.9
hpso02a	15.6	5.00	18344	500	1500	2.7	8.5	205.8	91.4	99.9
hpso02b	15.6	5.00	18344	500	1500	2.7	8.5	205.8	91.4	99.9
hpso04a	39.8	0.67	-	-	-	-	0.7	-	-	0.7
hpso05a	13.4	0.67	-	-	-	-	2.6	-	-	2.6
Average	-	-	-	-	-	-	4.6	-	42.8	47.4

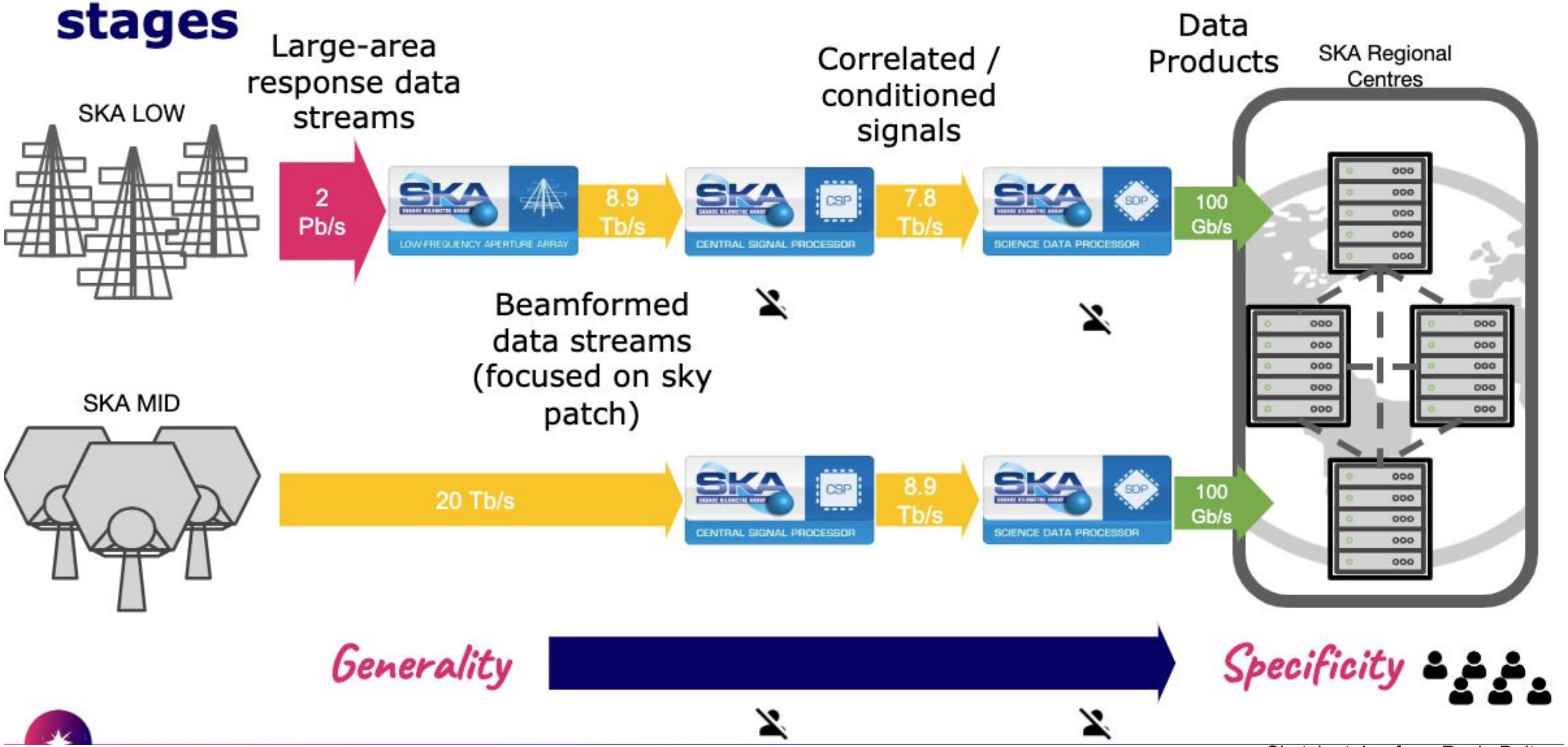
SKA1_Mid:

HPSO	Time [%]	Tobs [h]	Npix (side)	Channels (DPrepB)	Channels (DPrepC)	Image size [GB]	Non-Vis Rate [Gbit/s]	Visibility Size [TB]	Visibility Rate [Gbit/s]	Total Rate [Gbit/s]
hpso04b	1.0	0.17	-	-	-	-	2.3	-	-	2.3
hpso04c	3.1	0.17	-	-	-	-	2.3	-	-	2.3
hpso05b	2.1	0.25	-	-	-	-	6.9	-	-	6.9
hpso13	6.5	8.00	25339	160	3200	5.1	4.2	-	-	4.2
hpso14	2.6	8.00	18814	300	5000	2.8	2.8	-	-	2.8
hpso15	16.5	4.40	10837	260	2500	0.9	0.8	-	-	0.8
hpso18	13.1	0.02	-	-	-	-	0.1	-	-	0.1
hpso22	7.9	8.00	110601	1000	0	97.9	48.1	-	-	48.1
hpso27and33	13.1	0.12	23549	700	0	4.4	99.3	-	-	99.3
hpso32	13.1	2.20	-	-	-	-	1.3	-	-	1.3
hpso37a	13.1	3.80	94195	700	0	71.0	60.6	-	-	60.6
hpso37b	2.6	8.00	94195	700	0	71.0	28.8	-	-	28.8
hpso37c	2.6	8.00	94195	700	0	71.0	28.8	-	-	28.8
hpso38a	1.3	8.00	113204	1000	0	102.5	50.4	-	-	50.4
hpso38b	1.3	8.00	113204	1000	0	102.5	50.4	-	-	50.4
Average	-	-	-	-	-	-	28.4	-	0.0	28.4

Note that this assumes that we manage to produce usable data at all times.

These are *predicted* computing needs *within* SKAO. Data generation output rates between <1 to 100 Gbps on the fractions of time assumed

SKA Regional Centres: SKAO data processing stages



SKA LOW

Large-area response data streams

2 Pb/s

8.9 Tb/s

Correlated / conditioned signals

7.8 Tb/s

Data Products

100 Gb/s

SKA Regional Centres

Beamformed data streams (focused on sky patch)

SKA MID

20 Tb/s

8.9 Tb/s

100 Gb/s

Generality

Specificity

Sketch stolen from Rosie Bolton

CTA

- Two Telescope arrays. North: La Palma (Spain) and South: Paranal (Chile).
- Raw Data Volume: **~2PB/year/site => 50PB** in the first decade to be transmitted off-site
  - **Raw file sizes O(few GB)**, but smaller size for derived data products, eg. processed data available to users via a Science Archive.
  - Data will be stored in a Hot version and two Cold versions (=on tape, with 300km physical separation)
- Four off-site data centres: PIC Barcelona, DESY-Zeuthen, CSCS Lugano and INAF/INFN Frascati
  - Between the four off-site data centre a **minimum of 10 Gbps bandwidth must be available** for data replication purposes.
  - A redundant network connectivity of each data centre to their local NREN is also recommended.
- Last mile connectivity:
  - North array - RedIRIS, South array - ESO REUNA Chilean

Rubin Observatory - LSST

- Rubin Observatory data flows from Chile desert to SLAC. Few hours later is shipped to European sites, processed and sent back to SLAC (150ms latency).
- **Raw Data**: 20 TB/night, 300 nights/year (+5 PB extra every year), totalizing ~170PB (year 10)
    - A subset of data products is replicated to about 12 Data Access Centers around the world, network connectivity of most of those sites is not yet good enough.
    - The data release is composed of: raw images, calibrated images and the astronomical catalog data which is ingested into a multi-PB relational database.
    - **Reprocessing** once a year (whole raw dataset)
- **Challenge**: Majority of **small size files O(MB)** sent across the Atlantic over a **high-latency network**
    - Astronomy projects typically store their data this way: 1 FITS file per CCD in the focal plane of the camera (200 CCDs in Rubin’s camera)
    - HTTP/3 protocol uses UDP, which helps in transferring small files over high latency networks. But it requires complicity of network providers and sites to allows these flows. Could this be explored?
- **R&D activities**:
    - LSST will benefit from the ongoing work in the LHC community understanding network flows, packet marking.

Network R&D ideas

- **Packet Marking**. Capability to *paint* the traffic for the main data workflows. Biggest part of data transit will continue being asynchronous, TPC-based transfers *driven* by Rucio and *executed* by FTS.

- **Network Awareness** system to reduce potential needs of over-provisioning. Requires stateful, prompt monitoring of network links among sites.

⇒ *Both needed to further explore traffic shaping and network/traffic orchestration possibilities*

- Network **provisioning** methods to boost efficiency by combining networks (eg. NOTED project)
  - eg. Leverage LHCOPN and LHCONE
  - eg. Leverage ESNET (US) with GEANT and other RENs

- **Jumbo Frames**:
  - Should JF be explored as possible “standard”? Our use case match its purpose. But mismatches between routers lead to efficiency drop, no consensus yet.
  - Currently JF have the same MTU range as in the Gigabit Ethernet era: 1500 bits (simple) and 9000 bits (Jumbo). Technology evolved, shouldn’t other ranges be explored? eg. IPv6 supports 65k packets out of the box

- TCP vs. **UDP** revisit? Big portion of our transfers are large volumes push-style? What will be the costs/drawbacks of UDP retransmissions?

Next steps

- WLCG Data Challenge planned for early 2024: *the multi-Tbps challenge*
- Data Lake models (Rucio+FTS) being adopted or under serious consideration in Europe's largest Research Infrastructures
  - Expect growing importance of Content Delivery and Latency hiding mechanisms (caching services)
  - R+D efforts on packet marking and traffic shaping. Will benefit from these common models and tools
- Will our networks be ready to cater for data transfers needs in few years? Not only for Raw data but also analysis/user driven workflows?
- Are we active enough to foresee the right tools to react on the network infrastructure, eg. QoS, traffic shaping, prioritisation?
- Should be promote stronger global coordination?
  - Aiming for agreed roadmaps on network usage recommendations (blocksize, MTU,..), tools (monitor, QoS,) joint R&D efforts (shaping, marking)... and promote coordinated network specific tests?

CERN