Data and Network challenges in preparation of HL-LHC

DOMA meeting - 23rd of September 2020

edoardo.martelli@cern.ch

Data production in HL-LHC

CMS and ATLAS will produce ~350PB of raw data per year, running for 100 days/year

The traffic from CERN to all the T1s will be ~400Gbps for 100 days per year per experiment, as it will export in quasi-real time.
- FNAL (40% of resources for CMS) will import at ~150 Gbps
- BNL (25% of resources for ATLAS) will import at ~100 Gbps

Therefore it is estimated the need for **4Tbps of network capacity from CERN to the T1s** by the time of HL-LHC, of which **~1.5Tbps will be needed across the Atlantic** to cover the needs of ATLAS and CMS

Needed transatlantic capacity

By the time of HL-LHC, we will need **to demonstrate our capability to transfer ~1Tbps across the Atlantic**

Note that today there is headroom between the transatlantic network needs and its capacity. This headroom will shrink considerably for HL-LHC as the transatlantic capacity will not increase by a factor 50 as the data volume.

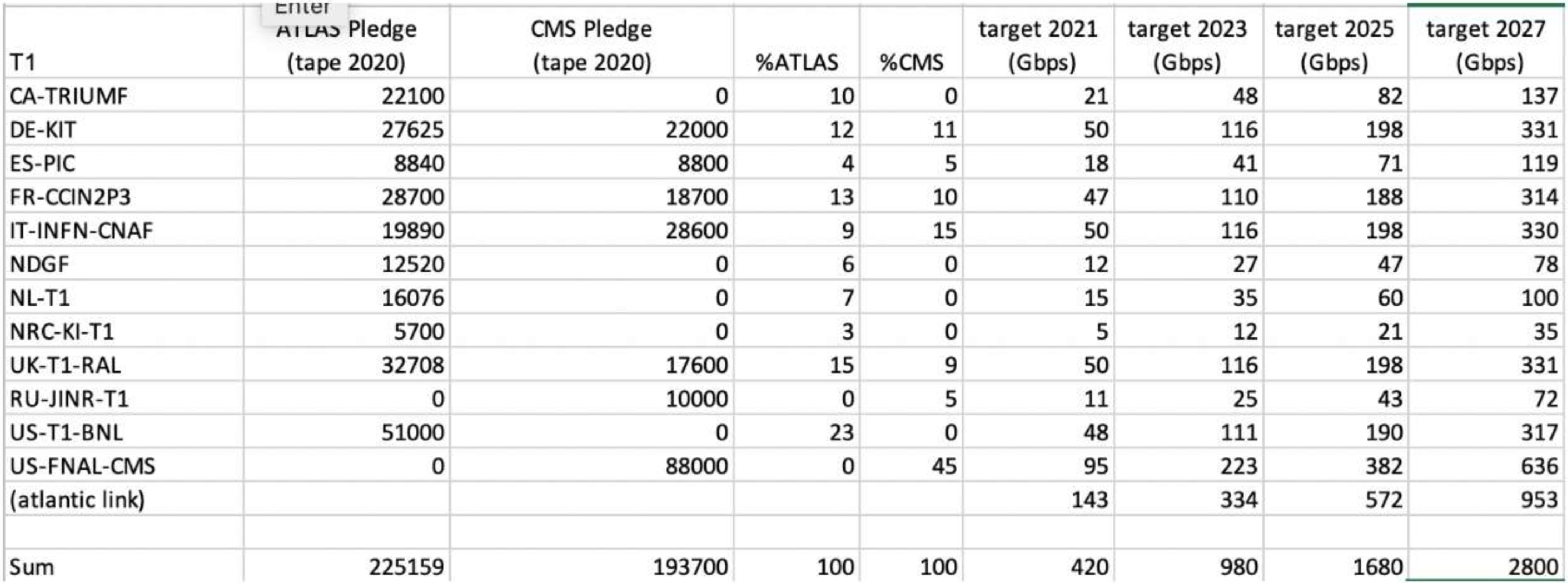
Data and Network challenges

The challenges could consist in **demonstrating the capability to transfer an increasing volume of data over the next years to reach the production transfer target**, sustained for a few days, by the start of HL-LHC in 2027.

We could foresee **milestones as 15% of the target 2021, 35% in 2023, 60% in 2025 and 100% in 2027.**
This could be adjusted based on the growth plan of the NRENS.
The same should be done between CERN and a subset (or all) the T1s in Europe, scaled by the size of the T1 wrt BNL and FNAL.

The 2021 target is important because it provides a baseline, but also it allows us to commission our capability to transfer data at a higher rate for special periods of Run-3.

Expected T1s incoming traffic (T0->T1s)



T1	ATLAS Pledge (tape 2020)	CMS Pledge (tape 2020)	%ATLAS	%CMS	target 2021 (Gbps)	target 2023 (Gbps)	target 2025 (Gbps)	target 2027 (Gbps)
CA-TRIUMF	22100	0	10	0	21	48	82	137
DE-KIT	27625	22000	12	11	50	116	198	331
ES-PIC	8840	8800	4	5	18	41	71	119
FR-CCIN2P3	28700	18700	13	10	47	110	188	314
IT-INFN-CNAF	19890	28600	9	15	50	116	198	330
NDGF	12520	0	6	0	12	27	47	78
NL-T1	16076	0	7	0	15	35	60	100
NRC-KI-T1	5700	0	3	0	5	12	21	35
UK-T1-RAL	32708	17600	15	9	50	116	198	331
RU-JINR-T1	0	10000	0	5	11	25	43	72
US-T1-BNL	51000	0	23	0	48	111	190	317
US-FNAL-CMS	0	88000	0	45	95	223	382	636
(atlantic link)					143	334	572	953
Sum	225159	193700	100	100	420	980	1680	2800

T1-T2: Reprocessing at T2s

The data at the T1 needs to be staged from tape and exported to the T2s for processing

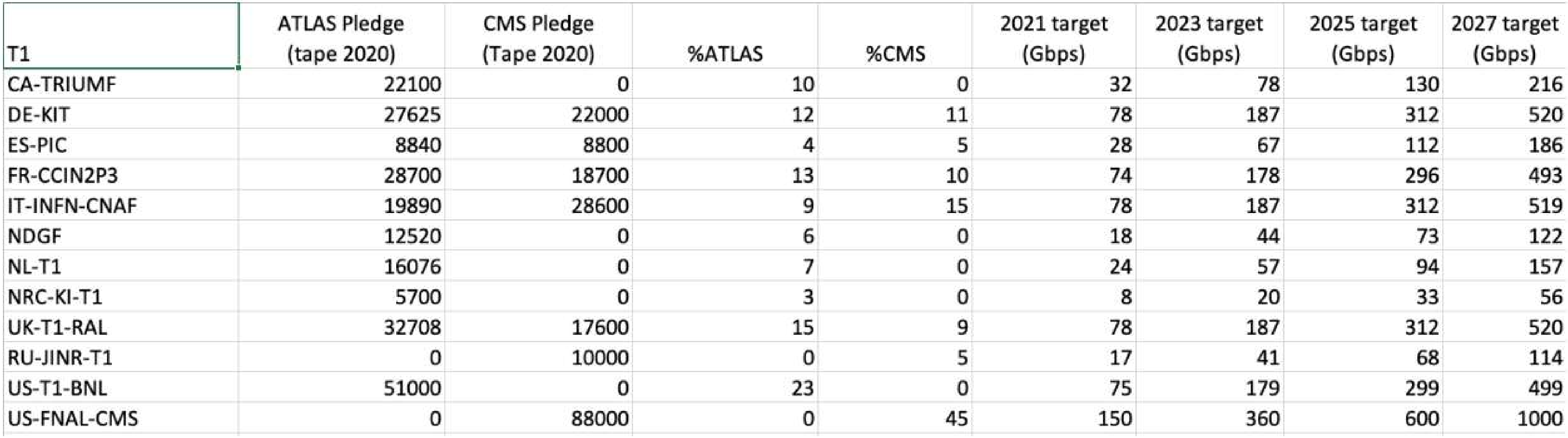
The target is to be able to reprocess 100% of the data collected in the year and stored at a specific T1 in less than three months.

The data could be streamed directly to the processing centres or buffered at the T1 and transferred in a burst. This has different implications on the storage needs at T1s and T2s, the balance with CPUs and the network needs.

A T1 will need to commission its capability to stream an aggregated 1Tbps to the T2s. The 1Tbps T1 egress capacity is the target for 2027 for a 40% T1 serving only one experiment (e.g. FNAL). The targets for the other T1s can be derived from there.

Also for T1-T2 challenges, intermediate targets should be defined and challenged at the time of the experiment's reprocessing and derivation campaigns (e.g. through the data carousel) in Run-3.

Expected T1s outgoing traffic (T1->T2s)



T1	ATLAS Pledge (tape 2020)	CMS Pledge (Tape 2020)	%ATLAS	%CMS	2021 target (Gbps)	2023 target (Gbps)	2025 target (Gbps)	2027 target (Gbps)
CA-TRIUMF	22100	0	10	0	32	78	130	216
DE-KIT	27625	22000	12	11	78	187	312	520
ES-PIC	8840	8800	4	5	28	67	112	186
FR-CCIN2P3	28700	18700	13	10	74	178	296	493
IT-INFN-CNAF	19890	28600	9	15	78	187	312	519
NDGF	12520	0	6	0	18	44	73	122
NL-T1	16076	0	7	0	24	57	94	157
NRC-KI-T1	5700	0	3	0	8	20	33	56
UK-T1-RAL	32708	17600	15	9	78	187	312	520
RU-JINR-T1	0	10000	0	5	17	41	68	114
US-T1-BNL	51000	0	23	0	75	179	299	499
US-FNAL-CMS	0	88000	0	45	150	360	600	1000

Reprocessing at HPCs

HPC will also be used.

The use case where an HPC would provide an allocation of 5k nodes (128 cores each) for many days capable to process 10kHz of events, implies **demonstrating the capability to stream 1Tbps of data into a HPC**

Intermediate targets should be defined also for this case

Network R&D: tagging and shaping

The data challenges will happen in parallel to production activities. We need the capability to **mark the traffic for different activities by tagging the packets** with the source experiment and application purpose (see Research Network Technology WG Packet Marking activity).

We should focus on the scheduled traffic (asynchronous, storage to storage, via FTS) as that will be the bulk of the network utilisation. The focus should be instrumenting the T1s and largest T2s which participate to the challenges. This will be the foundation to understand the possible future needs for traffic shaping and orchestration. **Traffic shaping for transfer speed optimization** is the second achievable target of the RNTWG.

Network R&D: bandwidth provisioning

The challenges should follow but also **drive the expansion of the network capacity**. At the same time, we will need to acquire access to extra network capacity to proceed with the challenges and the R&Ds on **dynamic provision of additional bandwidth**

The third subgroup of the RNTWG is focused on
Network Orchestration

Other projects like SENSE, NOTED, DTNs, AUTOGOLE could enable an efficient use of existing bandwidth and deliver extra bandwidth when more is needed

Bandwidth for Data Challenges

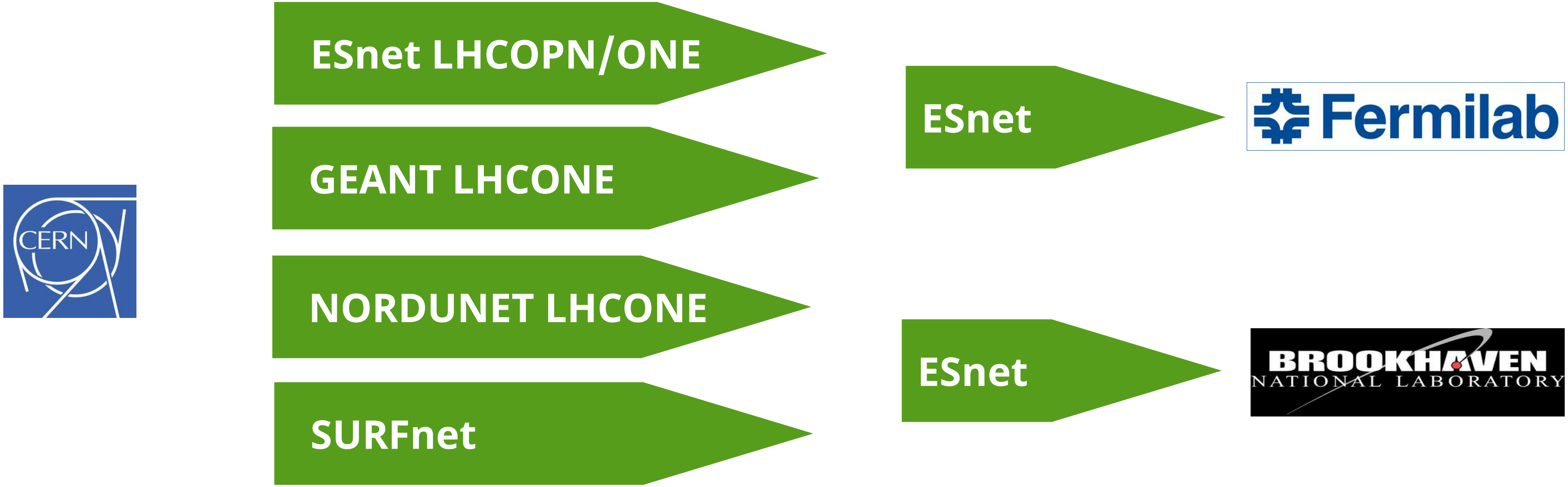
Data challenges could be run using extra bandwidth where already available. For example by pairing LHCOPN links with LHCONE. Or complementing the ESnet capacity over the Atlantic with the capacity offered by GEANT and others.

We could also lease network lines of given capacity between T0 and T1s. An example is the CERN-Amsterdam 400Gbps line and we could think about others, discussing how to share the cost of the lease.

At the same time, we should explore the possibility of **leasing a network connection offered by large cloud providers** and compare the cost with the previous cases. This case would require large bandwidth connections to a Cloud providers on both ends

Transatlantic challenge: T0 to US T1s

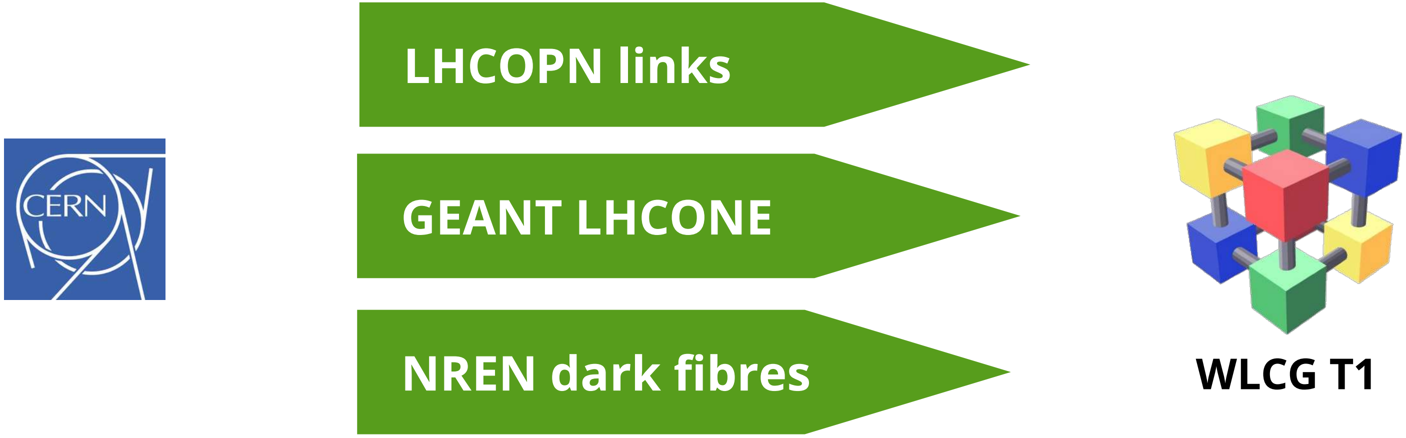
- Could be done aggregating several transatlantic links of the different LHCONE providers (7-800Gbps today)
- Bandwidth at T1s needs to be improved (today limited to 200Gbps)
- Target: 1Tbps
- Necessary to involve Experiments and Service managers to generate enough data transfers to fill the network



European challenge: T0 to T1s

- Also for Tier1s in Europe several links could be aggregated, leveraging the large availability of dark-fibres

- Target: 1Tbps



LHCOPN links

GEANT LHCONE

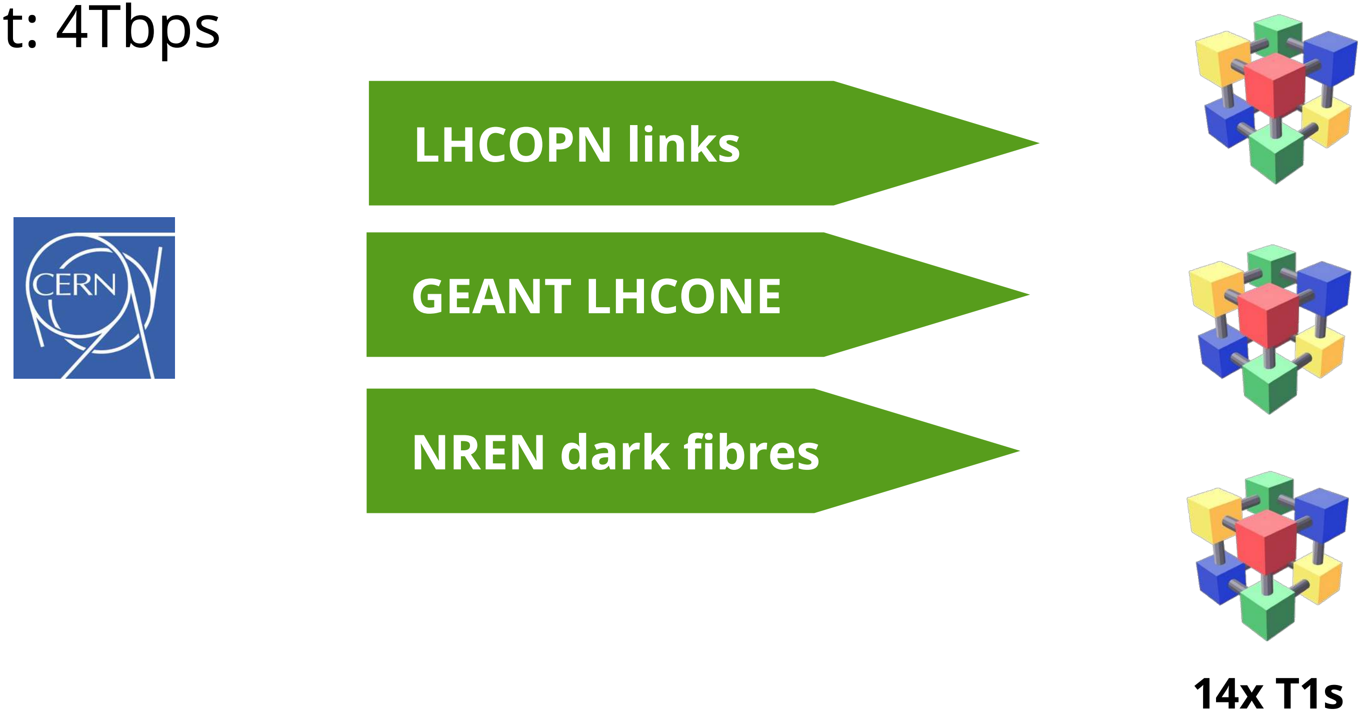
NREN dark fibres

WLCG T1

E.g. soon available 400Gbps test link between CERN and NL-T1

T0 challenge: T0 to all T1s

- The Tier0 needs to demonstrate the capacity to stream data at > 1Tbps
- Target: 4Tbps



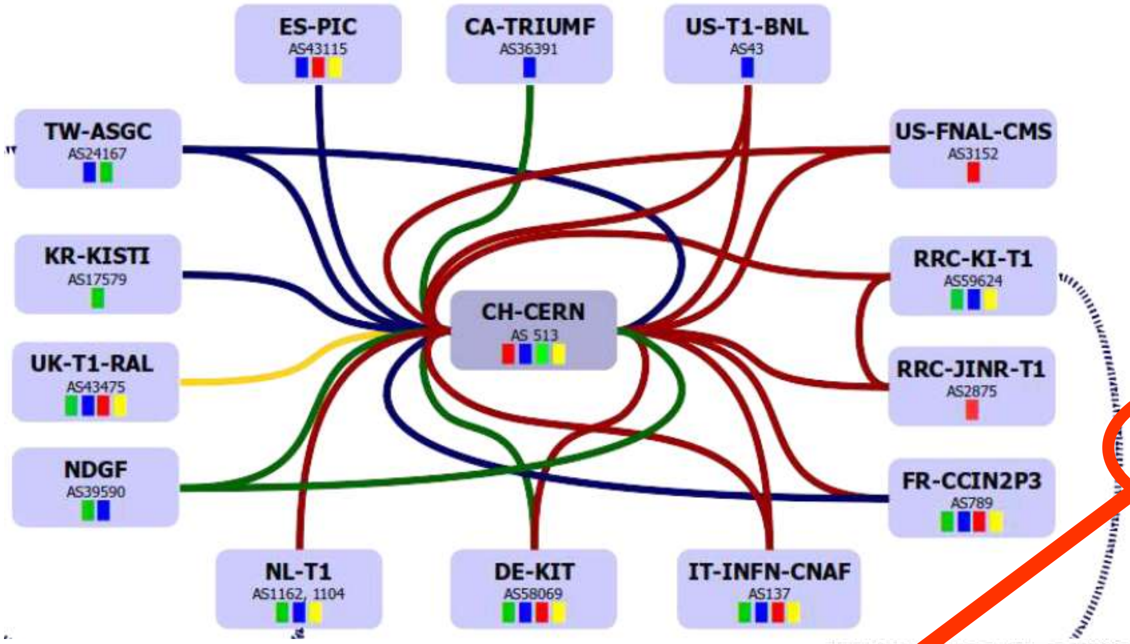
LHCOPN links

GEANT LHCONE

NREN dark fibres

14x T1s

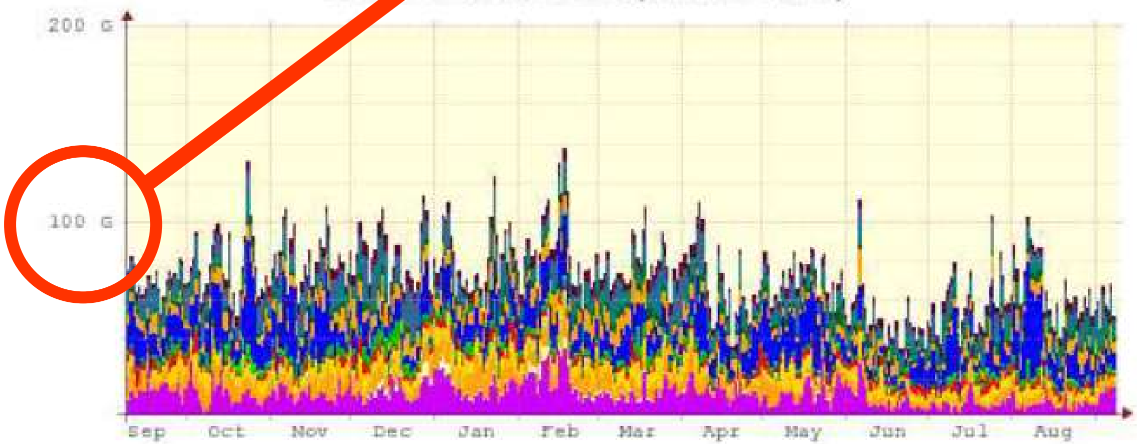
T0-T1 bandwidth is already available



Numbers

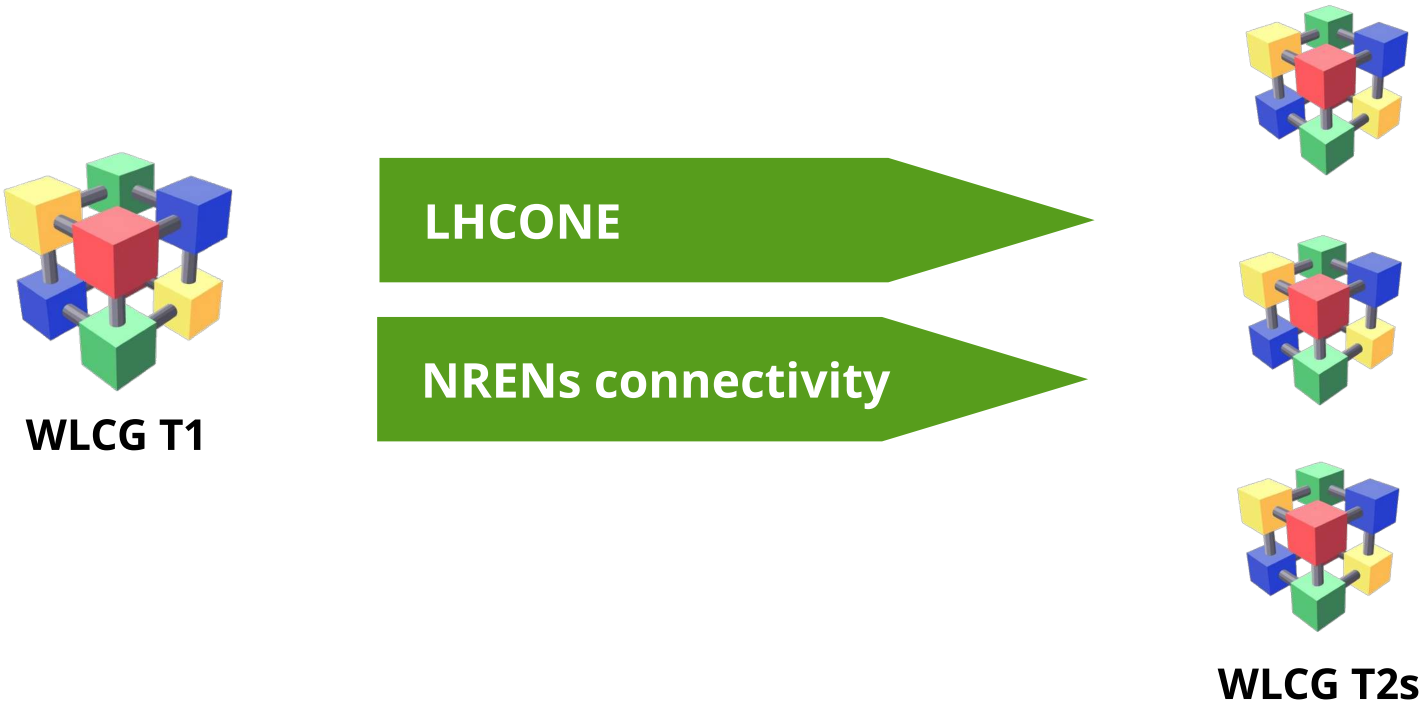
- 14 Tier1s + 1 Tier0
- 12 countries in 3 continents
- Dual stack IPv4-IPv6
- 1.1Tbps to the Tier0

LHCONE TOTAL Traffic (CERN -> Tiers1)



Reprocessing at T2s

- T1s can aggregate existing and available bandwidth
- Target: 1Tbps



LHCONE

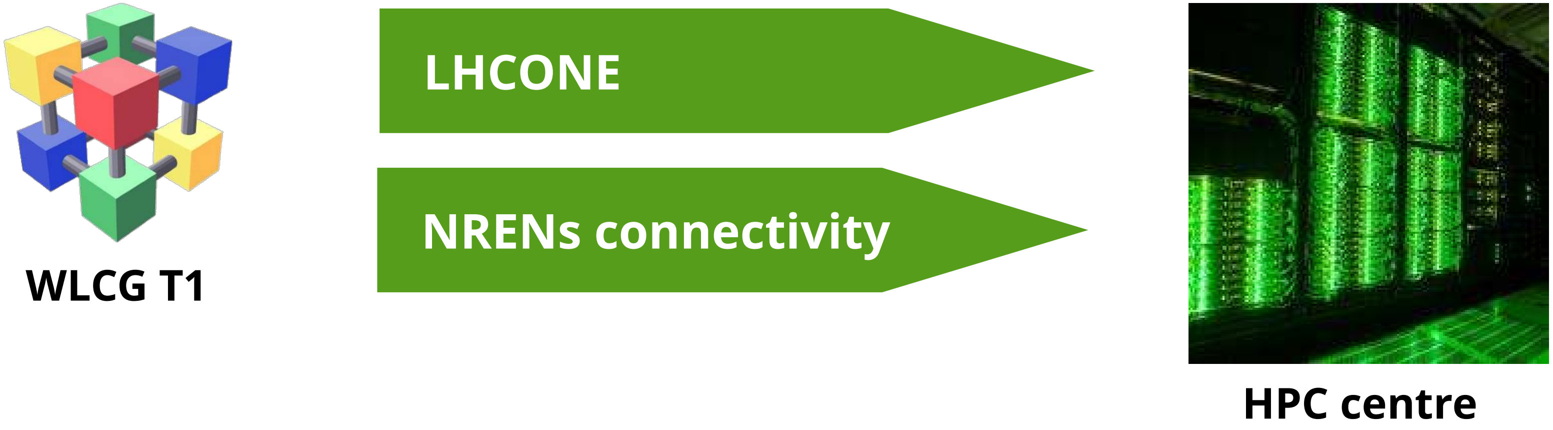
NRENs connectivity

WLCG T1

WLCG T2s

HPC challenges

- Connect HPC centres to R&E networks at large bandwidth
- HPC centres not always equipped with large bandwidth network access
- Target: 1Tbps



LHCONE

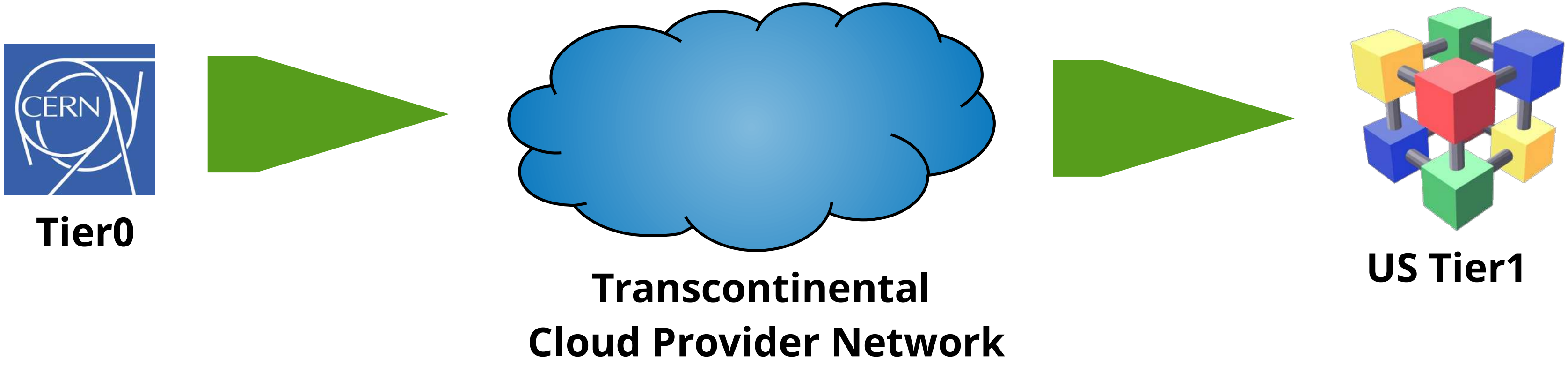
NRENs connectivity

WLCG T1

HPC centre

Bandwidth from cloud providers

- Explore the possibility to use the transcontinental networks of large Cloud providers on demand
- Necessary to have pre-provisioned large connections to Cloud Providers
- Target: compare costs



What’s next

- Involve Experiments and Storage-Transfer service managers
- Define set of meaningful data challenges
- Asses the aggregated network capacity of existing and upcoming storage
- Agree on the schedule of the challenges

References

https://docs.google.com/document/d/1sVnfkUS_7uh892eTtHUnPPcbpEaYwyCxcUEWAZtmAQA

Questions or comments?