HL-LHC network needs and data transfer challenges

This document summarises the current (2021) understanding of the HL-LHC network needs and proposes a schedule and targets for data transfer challenge. This commissioning activity should assess the readiness of the infrastructure in preparation for HL-LHC. We focus on two use cases, which will drive the network needs at the time of HL-LHC, according to the currently known data and computing models of the experiments:

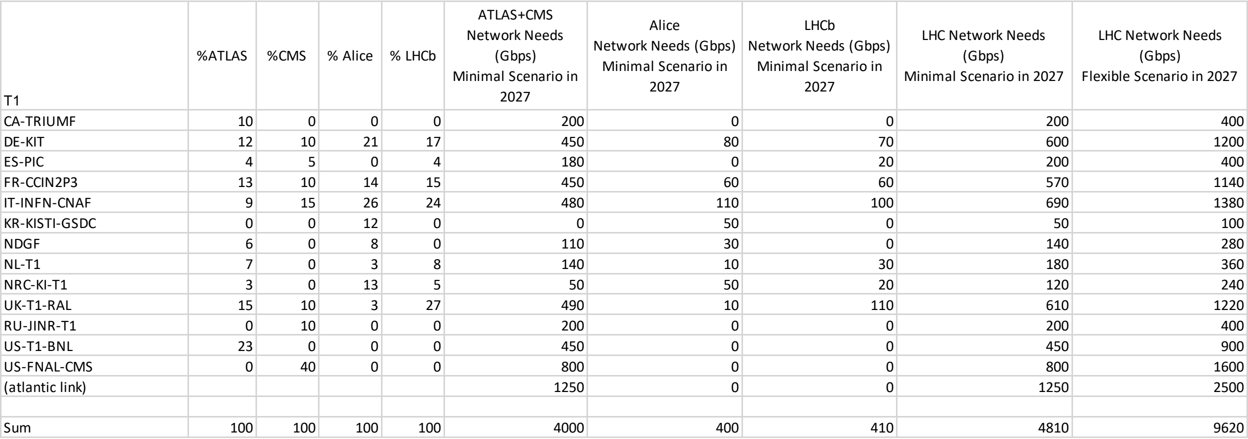
1) Export of RAW data from CERN to the T1s
2) Data Reprocessing

At the time of HL-LHC , the ATLAS and CMS activities will account for most of the network use in WLCG. We however include also the estimates from Alice and LHCb. The estimates have large uncertainties and should be treated as best known current approximations. They will be refined in the next few years following the HL-LHC computing TDRs preparation process.

Export of RAW data from CERN to the T1s. At HL-LHC, both the ATLAS and CMS experiments will produce ~350 PB of RAW data per year. The traffic from CERN to the T1s for RAW data export will be ~400 Gbps per experiment as we want to export in quasi-real time (7M seconds of LHC data taking per year). Those numbers do not include other data formats and represent a flat usage. We estimate we should include an additional 100 Gbps per experiment to account for those data formats. For Alice and LHCb, we estimate around 100Gbps per experiment, based on the Run-3 computing models.

The network infrastructure from CERN to the T1s needs to be sized to absorb such traffic. Based on current experience we will need a factor x2 in network capacity to absorb bursts of use (the above numbers refer to a flat use). We need another factor x2 overprovisioning to avoid running the network at full occupancy and suffering the operational problems that occur from that.

We estimate therefore the need for **4.8 Tbps** of network capacity from **CERN to the T1s** by the time of HL-LHC. Each T1 will provide a different fraction of the RAW data storage capacity and will therefore need to absorb a different fraction of this traffic. In Table 1 a breakdown per T1 is provided, based on the 2020 WLCG tape pledges. **~1.25Tbps** will be needed across the Atlantic to cover both the needs of ATLAS and CMS.



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

Table 1: network bandwidth needs per T1 (or region)

Reprocessing. The data at the T1 needs to be staged from tape and exported to the T2s for processing (assuming most of that processing happens outside the T1). To estimate the network needs, we consider a scenario where 100% of the data collected in the year and stored at a specific T1 is reprocessed in less than three months. Because 3 months ~= 7M seconds the table above can be used to set the targets for T1 to T2s traffic as well.

Minimalistic vs Flexible Scenario: for both use cases the network needs calculated above refer to a minimalistic scenario where the network is considered and used as a scarce resource. The experiment computing models evolved between LHC Run-1 and Run-2 to leverage more the network to obtain flexibility in workload scheduling, increase efficiency and turnaround time of processing and analysis activities. If we want to continue with such a model we estimate an additional x2 network capacity is needed. Table 1 shows both scenarios, which should be considered as the fork in between the network planning must occur. In the flexible scenario, we expect the largest T1s (KIT, IN2P3, CNAF, RAL, BNL, FNAL) to be connected with CERN and the T2s (at least at regional level) through a ~1Tbps network (more for FNAL). For the other T1s we expect ~500Gbps. The numbers from the flexible scenario are coherent with what was presented at the LHCONE/LHCOPN meeting in early 2020. The minimal scenario is coherent with the numbers presented at the ESNET planning of summer 2020.

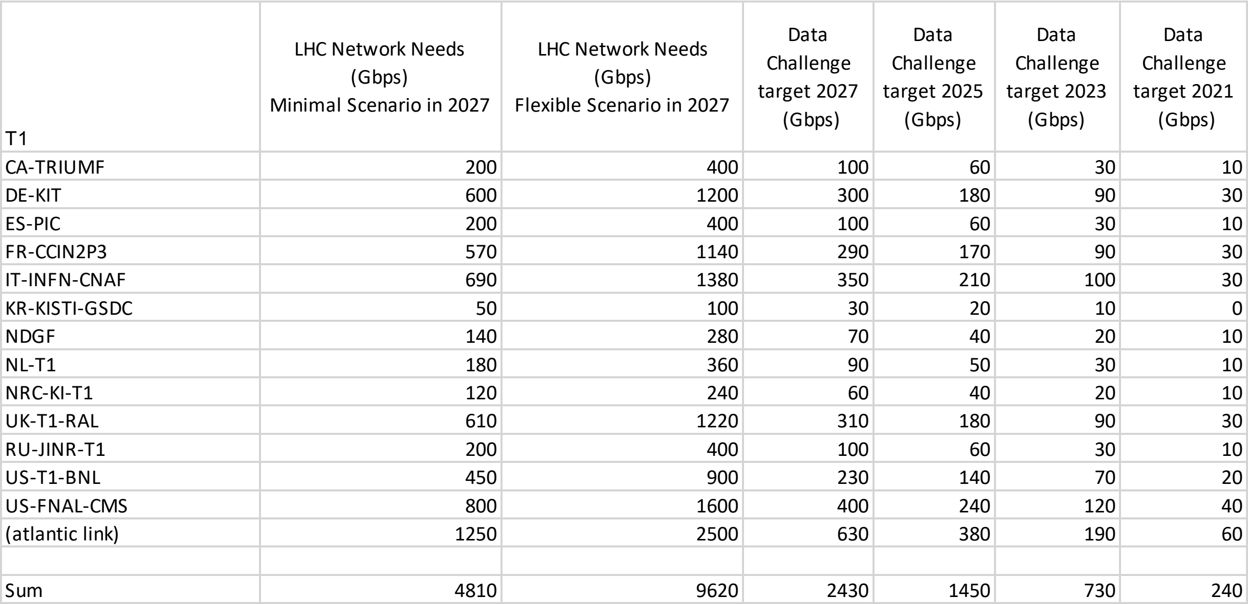
Data Challenges

By the time of HL-LHC, we plan to demonstrate our capability to fill ~50% of the full bandwidth required in the minimal scenario with production-like traffic (storage to storage, using the third party copy protocols and data management services used in production by the experiments).

Our challenge will consist in demonstrating our 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 foresee milestones as 10% of the target 2021, 30% in 2023, 60% in 2025 and 100% in 2027. This could be modified based on the growth plan of the NRENS.

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. For example, it would allow archiving with less delay the HI data to the T1s. Table 2 shows the data challenge target rates both for T0 data export to the T1s and T1 exports to the T2s for reprocessing.

During the data challenges, additional traffic will be injected in the system to complement the production activity. The sum of the production traffic and the additional traffic should reach the target. The data challenges will use the production services of the WLCG sites and experiments.



T1	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027	Data Challenge target 2027 (Gbps)	Data Challenge target 2025 (Gbps)	Data Challenge target 2023 (Gbps)	Data Challenge target 2021 (Gbps)
CA-TRIUMF	200	400	100	60	30	10
DE-KIT	600	1200	300	180	90	30
ES-PIC	200	400	100	60	30	10
FR-CCIN2P3	570	1140	290	170	90	30
IT-INFN-CNAF	690	1380	350	210	100	30
KR-KISTI-GSDC	50	100	30	20	10	0
NDGF	140	280	70	40	20	10
NL-T1	180	360	90	50	30	10
NRC-KI-T1	120	240	60	40	20	10
UK-T1-RAL	610	1220	310	180	90	30
RU-JINR-T1	200	400	100	60	30	10
US-T1-BNL	450	900	230	140	70	20
US-FNAL-CMS	800	1600	400	240	120	40
(atlantic link)	1250	2500	630	380	190	60
Sum	4810	9620	2430	1450	730	240

Table 2: data challenge target rates.

R&D activities

Several R&D activities complement and support the data challenges program. They should be carried on in parallel to the above described data challenges program and/or in support to it. The outcomes should be integrated progressively in the network services so that they can be commissioned at scale in future challenges.

R&D on network capabilities. Challenging and/or novel aspects of our use case are: the bandwidth requests can match or over-match capacity on some routes and some segments and the aggregate of requests can match or overwhelm the available capacity. A stateful network management system could reduce many of the needs for over-provisioning and address the challenges above. Such a system requires detailed monitoring information along network paths and at sites, and can evaluate the impact of further significant allocations. Key features include:

- Handling multiple requests taking policy and priority (a paradigm "to be defined") into account
- Giving weight to performance/throughput, load balancing, good use of site resources, organizational and geographical preferences in assigning paths; eventually - a hierarchy of objectives and constraints in a multi-objective optimization strategy
- Identification, diversion and assignment to alternate, additional, or privileged paths when available, OR otherwise constraining the allocations not to impede existing best effort traffic on shared routes
- Deciding how to deal with the constraints as real-time requests keep coming in, via: Queueing or real-time adjustments of allocations, with notifications to and from the client workflow/data-management system
- Setting break-points on taking back capacity when the application does not well-use the allocation(s) it has been given.

As part of this system, we need the capability to mark the traffic for different activities by “annotating” packets. 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 to is the second achievable target and traffic orchestration is the longer term one. The HPC use case for example could be one that would benefit largely from traffic shaping (the bandwidth needs could be considerable but limited in time) and the R&D could look in that direction for the first demonstrator. As part of this R&D theme, there are several US and European projects trying out the technologies to provide bandwidth on demand for point to point connections. Depending on flexibility and cost, the transfer scheduler (i.e. FTS) could request a specific guaranteed bandwidth between two sites. In combination with the current model of having statically provisioned network connections, a more dynamic provisioning model *might* be more cost effective, compared to a model where we have to significantly overprovision our network connections.

R&D on network provisioning. The challenges should follow but also drive the expansion of the network capacity. At the same time, we might need to acquire access to extra network capacity to proceed with the challenges and the R&Ds. There are several possibilities:

The NOTED project can allow a more efficient use of multiple network paths from different providers. This would support complementing, for example, the LHCOPN bandwidth with LHCONE. We discussed for example complementing as a test the ESNET capacity over the Atlantic with the one offered by GEANT and other RENs. The same could be done also between CERN and other T1s in Europe. For transatlantic, there is an existing opportunity with the Advanced North Atlantic (ANA) consortium. This supports a wide range of uses by the global R&E community and we can ask what part of ANA can possibly be devoted to specified "high priority" uses, as long as we understand the limits of what we will use for the rest without completely blocking the other traffic.

We can lease network lines of given capacity between T0 and T1s. An example is the CERN-SARA available 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 case. For example in the US there are plans to attempt to reach 500Gbit/sec in 2021 inside a cloud. It is relatively inexpensive and does not interfere with production activities.