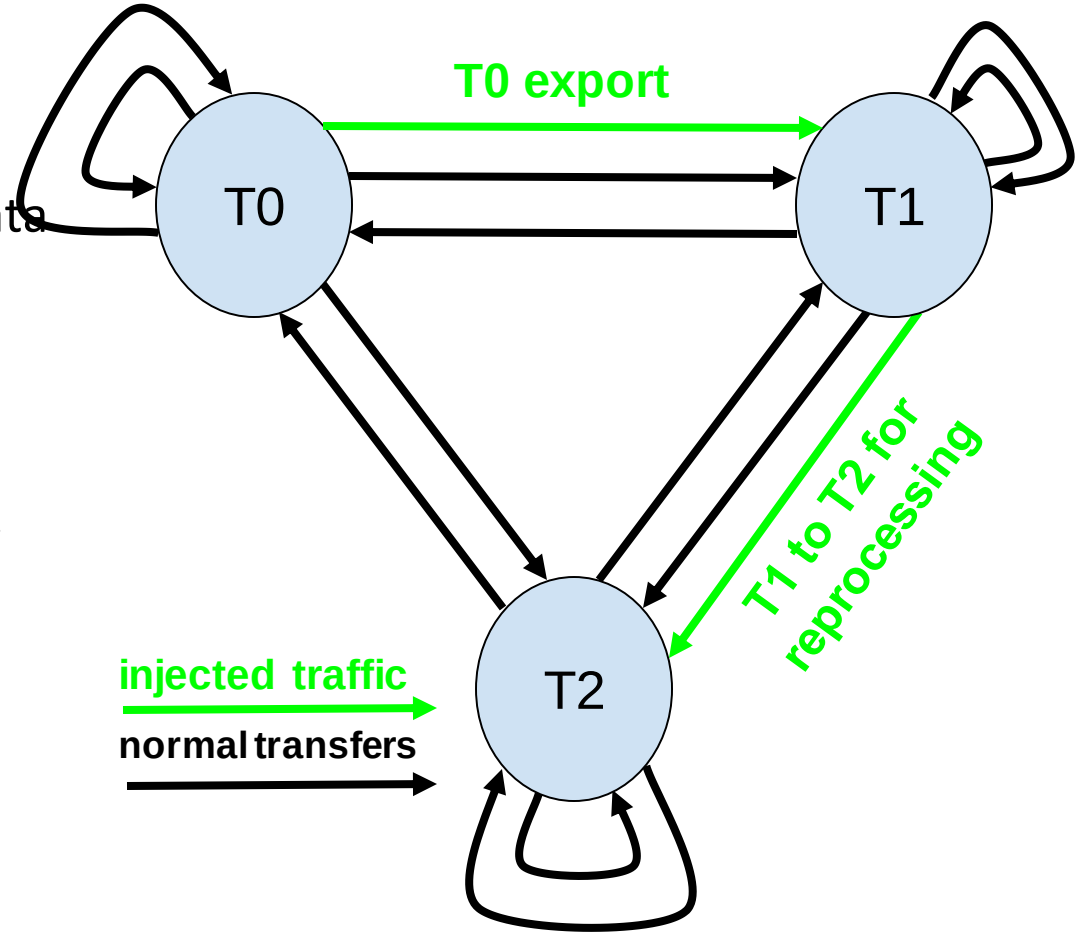
ATLAS plans for Data Challenges 2024

2023-11-09

Petr Vokáč

Model for HL-LHC Data Challenges

- Considers use-cases driving network needs
  - Export of 350PB RAW data to T1
  - Data Reprocessing mostly at T2
- Quasi-Realtime T0 export assuming ~3 months data taking per year
  - 400Gbps for RAW data
  - additional 100Gbps for other formats
  - 2x to absorb bursts, 2x overprovisioning
  - 2Tbps estimated for ATLAS (4.8Tbps all LHC)
- Similar assumptions / bandwidth for reprocessing
- Minimal vs. flexible scenario
  - cover existing flexibility built into WFMS
  - factor 2x ⇒ 4Tbps (T0→T1) + 4Tbps (T1→T2)
- Not considered MC production, derived date re-creation, data consolidation, recovery, …

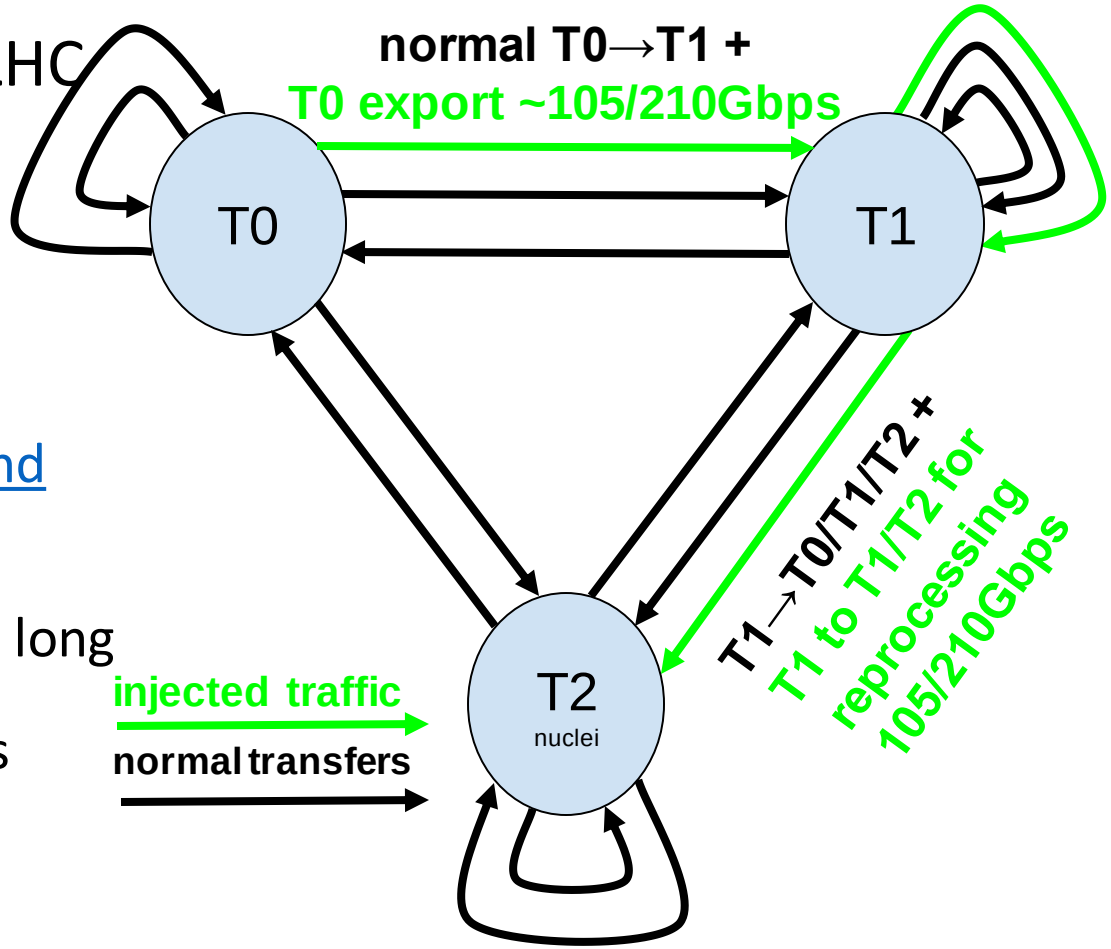


HL-LHC Network Needs (DC21 planning)

T1 Sites (T0 export / T1→T2 reco)	HL-LHC Minimal Scenario [Gbps]	HL-LHC Flexible Scenario [Gbps]	DC27 (100%) [Gbps]	DC26 (~~60~~→50%) [Gbps]	DC24 (25%) [Gbps]	DC24 ATLAS [Gbps]	DC24 CMS [Gbps]	DC24 Alice [Gbps]	DC24 LHCb [Gbps]	~~DC23 (30%) [Gbps]~~	DC21 (10%) [Gbps]
CA-TRIUMF	200	400	100	60	30	30	0	0	0	~~30~~	10
DE-KIT	600	1200	300	180	80	32	26	11	11	~~90~~	30
ES-PIC	200	400	100	60	30	13	13	0	3	~~30~~	10
FR-CCIN2P3	570	1140	290	170	70	33	21	7	9	~~90~~	30
IT-INFN-CNAF	690	1380	350	210	90	24	35	14	16	~~100~~	30
KR-KISTI-GSDC	50	100	30	20	10	0	0	10	0	~~10~~	0
NDGF	140	280	70	40	20	16	0	4	0	~~20~~	10
NL-T1	180	360	90	50	20	15	0	1	4	~~30~~	10
NRC-KI-T1	120	240	60	40	20	8	0	8	4	~~20~~	10
UK-T1-RAL	610	1220	310	180	80	39	21	1	18	~~90~~	30
RU-JINR-T1	200	400	100	60	30	0	30	0	0	~~30~~	10
US-T1-BNL	450	900	230	140	60	60	0	0	0	~~70~~	20
US-FNAL-CMS	800	1600	400	240	100	0	100	0	0	~~120~~	40
(transatlantic link)	1250	2500	630	380	160	60	100	0	0	~~190~~	60
Sum	**4810**	**9620**	**2430**	**1450**	**640**	**270**	**246**	**56**	**65**	~~**730**~~	**240**

DC21 minimal / flexible scenario

- Minor differences with respect to HL-LHC Data Challenges model
  - T1→T1 transfers in addition to T1→T2
  - transfers don't follow strictly hieratical Tx
  - more realistic for minimal scenario
- Details and recommendations
  - [WLCG Data Challenge 2021 description and conclusions]
  - data injection period too long
  - flexible target not reached for sufficiently long time
  - flexible could / should utilize all directions
  - monitoring issues / improvements

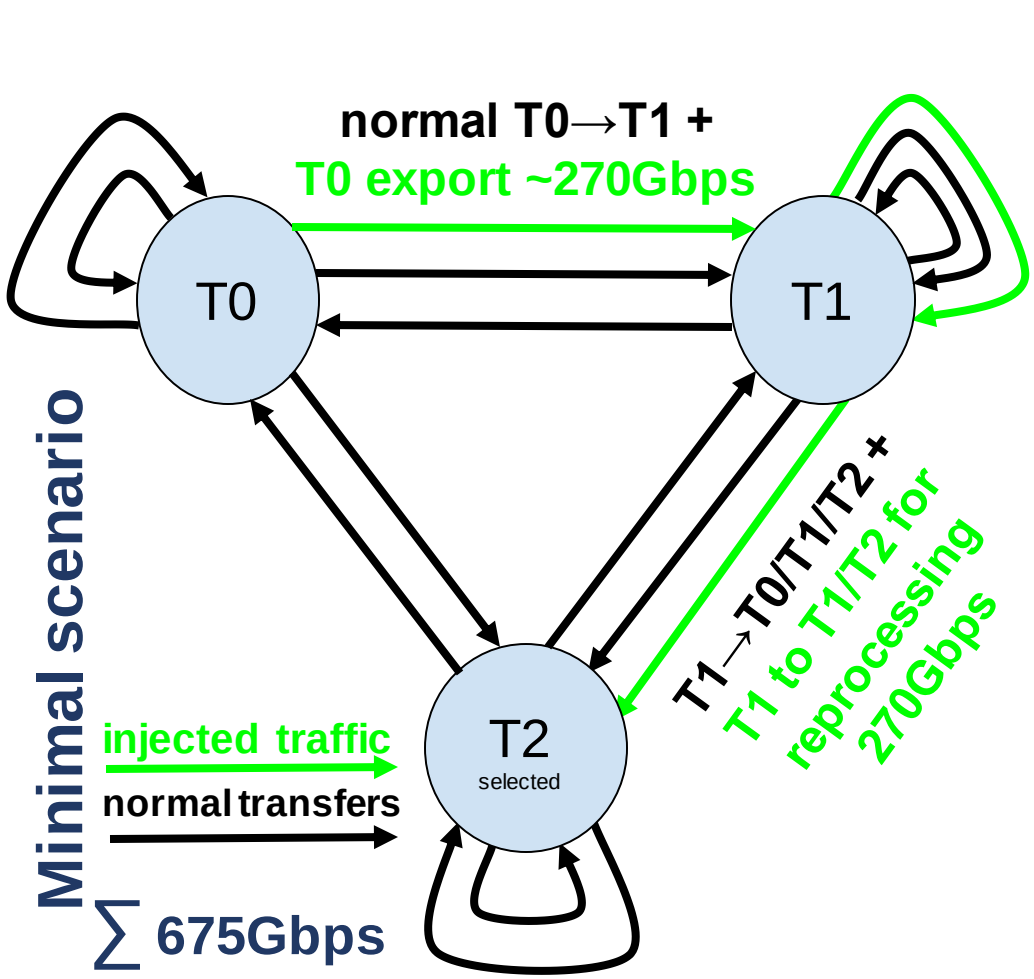


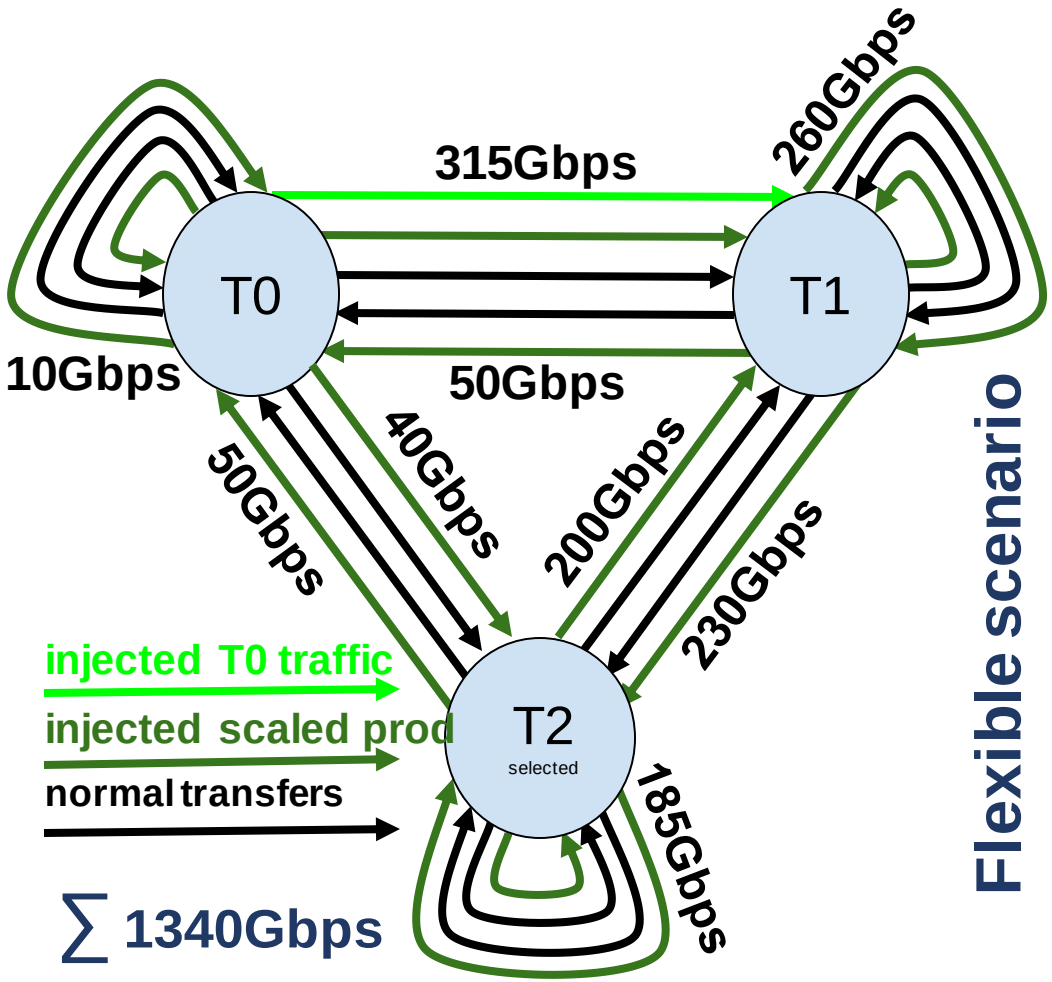
DC24 Models

- Minimal scenario
  - use DC21 model (list of T2 destinations slightly different)
  - T1 ingress: inject traffic from **T0 to T1** to reach **ATLAS target rates**
    - total T1 ingress higher – additional normal T1→T1 and T2→T1
  - T1 egress: inject traffic from T1 to T1 and selected T2
    - total **T1 egress to T0/T1/T2** should reach **ATLAS target rates**
- Flexible scenario
  - scale all production input/output transfer up to 2x ATLAS min. rates for T1 egress
  - Inject scaled traffic between all sites (T0/T1/T2 as source & destination)
  - real data (08/2022 – 08/2023 production in/out averages) to model transfer patterns
  - Tx ingress must match Tx egress, total ingress match egress

T1 Sites target rates for minimal scenario	DC24 ATLAS [Gbps]
CA-TRIUMF	30
DE-KIT	32
ES-PIC	13
FR-CCIN2P3	33
IT-INFN-CNAF	24
KR-KISTI-GSDC	0
NDGF	16
NL-T1	15
NRC-KI-T1	8
UK-T1-RAL	39
RU-JINR-T1	0
US-T1-BNL	60
US-FNAL-CMS	0
(transatlantic link)	60
Sum	270

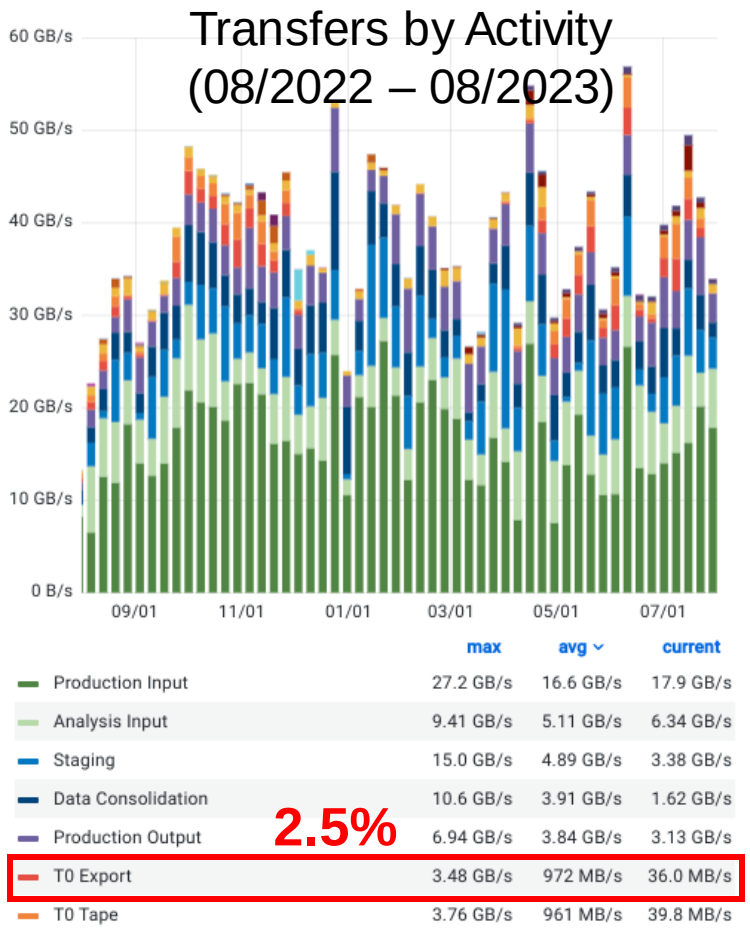
DC24 minimal vs. flexible scenario



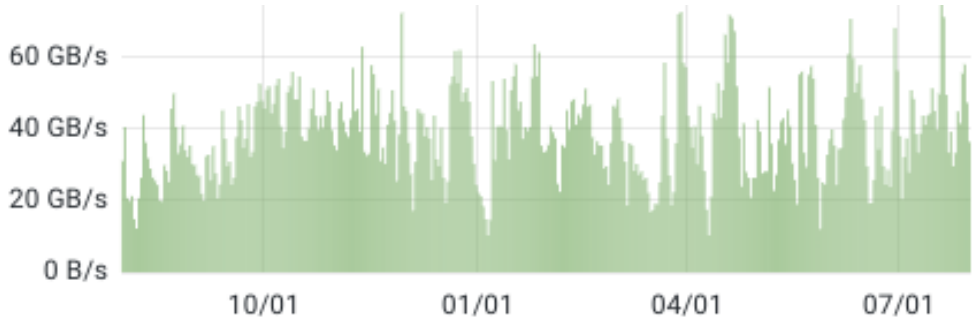


HL-LHC Model vs. Transfer Throughput Reality

- ATLAS transfers a lot of data
  - T0 Export ~7.5Gbps is just a small fraction of all TPC transfers ~285Gbps
    - factor two higher rate (~15Gbps) during data taking periods
  - T0 Export ~2.6% vs. production input/output ~67.9%
  - 28TB data transferred 08/2022 – 08/2023
- HL-LHC 350TB RAW ⇒ ~ 90Gbps average T0 export
  - 12x more RAW data compared to last year T0 Export volume
  - scaling current **average throughput** by this factor leads to **~3.5Tbps for HL-LHC**
  - close to overprovisioning assumed in HL-LHC DC Model
    - transfers with saturated links makes ops life difficult
    - factor two to cover peaks ⇒ ~7.0Tbps required by ATLAS
- Current **ATLAS** transfer patterns **needs flexible model**



How challenging is DC24

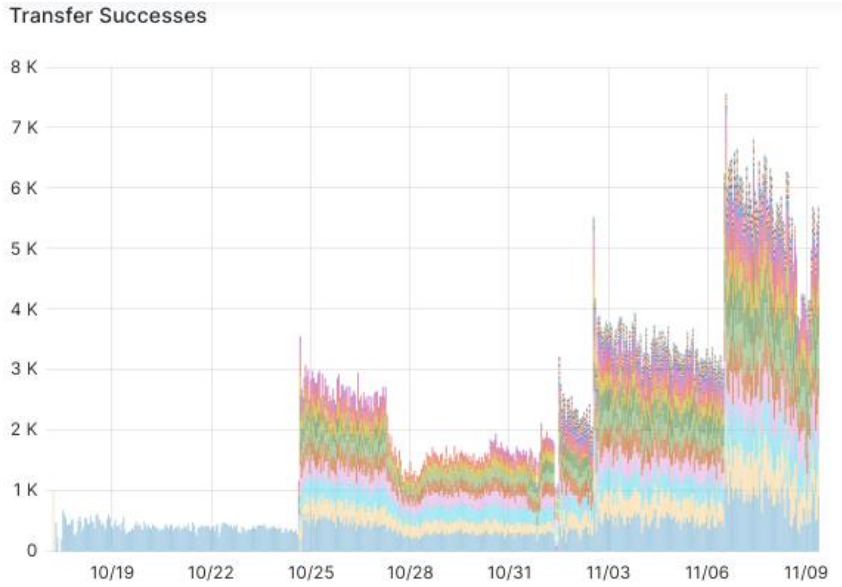


- Average transfer throughput **~285Gbps**
  - minimal scenario **675Gbps**, flexible scenario **1340Gbps**
    - final sum for flexible scenario depend on list of participating T2
    - throughput up to 1600Gbps if all sites participate in DC24
  - hourly peaks up to 900Gbps
  - daily averages fluctuate by factor 2-3
  - transfers from NDGF already above minimal scenario
  - our model allows us to estimate transatlantic throughput
- DC24 injected throughput for minimal model at same level as daily fluctuations
  - challenging to hit **target rate** without going over or **falling short**
  - **(semi-)automatic** adaptable **scaling of injected data volume?**
  - monitoring / scaling throughput for individual T1(?)

Minimal scenario		
Site	Ingress increase factor	Egress increase factor
CERN-PROD	1.0	7.8
BNL-ATLAS	4.7	2.3
FZK-LCG2	3.7	2.6
IN2P3-CC	3.6	2.2
INFN-T1	4.2	2.8
NDGF-T1	2.4	1.0
SARA-MATRIX	3.2	2.2
pic	4.3	3.2
RAL-LCG2	3.4	1.6
RRC-KI-T1	5.4	5.3
TRIUMF-LCG2	4.7	2.8
T2 average	2.2	1.0

Data injection

- Backfill on top of normal production transfers
  - use real data already stored at site
  - datasets with unique files and average size 3GB (check if exists at each site)
  - limit number of required deletions (number of transferred files)
  - average HL-LHC filesize is expected to be bigger compared to current average
- Injected volume
  - scaled yearly average (08/2022–08/2023) of production input/output
  - per-site egress factor coming our minimal / flexible model
  - additional throughput calculated for each link
    - transfers withing same site set to 0, compensated by increasing to other sites
    - source site → destination site: Gbps to be injected
  - minimal model: 226 links, flexible (with 16x T2): 703 links
    - number of links up to #sites^2
    - links with negligible transfer throughput can removed and compensated
- Every 15mins inject new Rucio rules with 2hour lifetime
  - ensure stable throughput rate for injected data (1h FTS monitoring resolution)
  - rule injection tool tuned for hundreds rules within 15min interval
  - this needs to be (stress) tested



Functional tests with dc_inject

Transfer rates table for sites

- T0/T1/T2 nuclei automatically included in DC24
  - **T2 sites can opt-in / opt-out from DC24**
- Table with DC24 details for sites
  - expected throughput, deletions and space
    - for minimal and flexible scenario
    - no additional space requirements from sites
      - it's up to Rucio to have sufficient space usable for DC
  - sites
    - **verify & provide details about WAN**
      - bandwidth available for ATLAS
      - not to clog upstream network with injected data
    - should provide upstream network monitoring (campaign)
  - red rows – sites not yet included (must opt-in)
- all-cloud mailing list
  - finalize list of participating sites (link)
  - response from few clouds still missing

ATLAS DC24 transfer rates

Final T2 ingress/egress depends on number of participating T2 sites and might be in given range
rows in red color: sites must explicitly ask be included in DC24 (details will be sent to all-clouds list)
(preliminary version: 20231103)
Deletion rates are calculated from ingress bandwidth assuming 3GB average filesize)

Table: DC24 (src: ingress / egress)			Site WAN (Gb/s)		DC24 minimal scenario				DC24 flexible scenario			
			Total (Gb/s)	Usable by ATLAS	T0 Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)	T0 Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)
Site	Tier	Cloud				∑ ingress	∑ egress			∑ ingress	∑ egress	
CERN-PROD	T0	CERN	2100	911	270.0	27.9	291.3	0 (0k)	270.0	93.1 - 112.2	363.1	884 (13k)
T0 summary					270.0	27.9	291.3		270.0	93.1 - 112.2	363.1	
BNL-ATLAS	T1	US	400	400	60.0	82.2	60.0	764 (11k)	60.0	107.5 - 119.6	120.0	1089 (15k)
FZK-LCG2	T1	DE	400	162	32.0	61.7	32.0	431 (6k)	32.0	86.3 - 100.3	64.0	911 (13k)
IN2P3-CC	T1	FR	200	93	33.0	53.3	33.0	413 (6k)	33.0	81.6 - 95.8	66.0	861 (12k)
INFN-T1	T1	IT	300	81	24.0	39.5	24.0	319 (5k)	24.0	54.8 - 64.0	48.0	588 (8k)
NDGF-T1	T1	ND	200	157	16.0	30.7	21.8	151 (2k)	16.0	77.9 - 96.6	32.0	842 (12k)
SARA-MATRIX	T1	NL	400	291	15.0	30.4	15.0	192 (3k)	15.0	54.4 - 66.0	30.0	604 (9k)
pic	T1	ES	200	89	13.0	21.4	13.0	170 (2k)	13.0	29.1 - 34.4	26.0	319 (5k)
RAL-LCG2	T1	UK	400	196	39.0	60.6	39.0	464 (7k)	39.0	88.5 - 100.1	78.0	861 (12k)
RRC-KI-T1 (no active T0 export)	T1	RU	200	79	8.0	13.4	8.0	109 (2k)	8.0	15.1 - 17.2	16.0	160 (2k)
TRIUMF-LCG2	T1	CA	100	100	30.0	45.9	30.0	403 (6k)	30.0	60.8 - 69.7	60.0	643 (9k)
T1 summary					270.0	439.3	275.8		270.0	655.9 - 763.8	540.0	
CA-VICTORIA-WESTGRID-T2	T2	CA	100	100		5.8 - 7.5	1.5 - 1.5	24 (0k)		3.6 - 13.4	1.5 - 11.0	104 (1k)
Australia-ATLAS	T2	CA	20	20		0.1 - 0.2	0.4 - 0.4	0 (0k)		0.1 - 2.4	0.4 - 2.6	25 (0k)
CA-WATERLOO-T2	T2	CA	40	40		2.0 - 2.4	1.2 - 1.2	7 (0k)		1.3 - 9.8	1.2 - 9.3	90 (1k)
CA-SFU-T2	T2	CA	100	100		5.9 - 7.7	5.7 - 5.7	45 (1k)		43.0 - 61.9	41.4 - 41.4	616 (9k)
praguelcg2	T2	DE	100	100		6.9 - 8.9	2.3 - 2.3	50 (1k)		16.9 - 22.9	15.5 - 15.5	197 (3k)
MPPMU	T2	DE				2.6 - 3.3	1.3 - 1.3	10 (0k)		1.7 - 9.4	1.3 - 9.1	82 (1k)
wuppertalprod	T2	DE	10	9		4.6 - 5.9	1.8 - 1.8	32 (0k)		9.9 - 10.0	6.4 - 8.8	74 (1k)
DESY-ZN	T2	DE	40	40		6.4 - 8.4	1.9 - 1.9	48 (1k)		14.3 - 19.2	12.4 - 12.4	163 (2k)
DESY-HH	T2	DE	100	100		9.1 - 10.0	1.9 - 1.9	49 (1k)		9.9 - 10.0	5.4 - 7.2	48 (1k)
UNI-FREIBURG	T2	DE				2.6 - 3.3	1.7 - 1.7	9 (0k)		1.8 - 11.3	1.7 - 11.6	101 (1k)
CYFRONET-LCG2	T2	DE	10	10		2.6 - 3.2	1.2 - 1.2	9 (0k)		1.7 - 9.9	1.2 - 9.4	90 (1k)
GoeGrid	T2	DE				5.2 - 6.6	1.2 - 1.2	20 (0k)		3.3 - 11.5	1.2 - 9.0	86 (1k)
IEPSAS-Kosice	T2	DE				1.1 - 1.3	0.4 - 0.4	4 (0k)		0.8 - 3.7	0.4 - 3.3	31 (0k)
LRZ-LMU	T2	DE				2.4 - 3.0	1.8 - 1.8	8 (0k)		1.6 - 12.9	1.8 - 12.5	120 (2k)
CSCS-LCG2	T2	DE	100	100		5.6 - 7.2	3.0 - 3.0	22 (0k)		3.6 - 22.3	3.0 - 21.3	198 (3k)
FMPhI-UNIBA	T2	DE				0.9 - 1.0	0.5 - 0.5	2 (0k)		0.7 - 4.2	0.5 - 4.0	37 (1k)
SAMPA	T2	ES	9	9		1.1 - 1.5	0.9 - 0.9	6 (0k)		0.6 - 8.1	0.9 - 7.3	79 (1k)
UAM-LCG2	T2	ES	10	10		0.7 - 0.9	0.4 - 0.4	4 (0k)		3.2 - 4.5	2.9 - 2.9	42 (1k)
ifae	T2	ES	200	200		2.7 - 3.4	0.7 - 0.7	11 (0k)		1.6 - 5.7	0.7 - 4.8	44 (1k)
NCG-INGRID-PT	T2	ES	9	9		0.5 - 0.7	0.2 - 0.2	2 (0k)		0.4 - 2.3	0.2 - 1.8	20 (0k)
IFIC-LCG2	T2	ES				4.1 - 5.3	2.0 - 2.0	17 (0k)		2.5 - 14.2	2.0 - 13.2	124 (2k)
EELA-UTFSM	T2	ES	10	10		0.2 - 0.2	0.3 - 0.3	1 (0k)		0.1 - 2.3	0.3 - 2.3	23 (0k)
TOKYO-LCG2	T2	FR	40	40		16.5 - 21.7	5.5 - 5.5	127 (2k)		30.0 - 39.7	29.8 - 29.8	317 (5k)
RO-07-NIPNE	T2	FR	100	100		4.3 - 5.4	2.6 - 2.6	29 (0k)		18.7 - 26.3	18.4 - 18.4	249 (4k)
BEIJING-LCG2	T2	FR	20	20		0.0 - 0.0	0.2 - 0.2	0 (0k)		0.0 - 1.5	0.2 - 1.3	15 (0k)
HK-LCG2	T2	FR				0.1 - 0.1	0.3 - 0.3	0 (0k)		0.1 - 2.3	0.3 - 2.3	23 (0k)
GRIF	T2	FR	100	100		7.2 - 9.4	4.2 - 4.2	32 (0k)		4.1 - 36.1	4.2 - 33.2	339 (5k)
IN2P3-LPC	T2	FR	100	100		2.4 - 3.0	1.5 - 1.5	14 (0k)		1.6 - 10.0	1.5 - 8.9	117 (2k)
IN2P3-LAPP	T2	FR	20	20		4.8 - 5.9	2.7 - 2.7	27 (0k)		16.1 - 19.9	13.6 - 15.1	174 (2k)
IN2P3-CPPM	T2	FR	100	100		2.5 - 3.2	1.6 - 1.6	17 (0k)		10.0 - 10.0	7.3 - 9.9	89 (1k)
INFN-MILANO-ATLASC	T2	IT	10	10		2.1 - 2.7	1.7 - 1.7	9 (0k)		1.2 - 10.0	1.7 - 10.0	94 (1k)
INFN-NAPOLI-ATLAS	T2	IT	100	100		3.8 - 4.8	2.2 - 2.2	24 (0k)		15.7 - 21.9	14.9 - 14.9	205 (3k)
INFN-ROMA1	T2	IT	10	10		2.5 - 3.2	1.1 - 1.1	17 (0k)		7.7 - 9.9	6.6 - 7.0	88 (1k)
INFN-FRASCATI	T2	IT	10	10		2.1 - 2.6	1.0 - 1.0	7 (0k)		1.5 - 8.6	1.0 - 7.9	75 (1k)
SE-SNIC-T2	T2	ND				0.0 - 0.0	0.0 - 0.0	0 (0k)		0.0 - 0.1	0.0 - 0.1	1 (0k)
UNIBE-LHEP	T2	ND				0.0 - 0.0	0.0 - 0.0	0 (0k)		0.0 - 0.0	0.0 - 0.0	0 (0k)
NIKHEF-ELPROD (no tape)	T1	NL	1000	1000		6.8 - 9.1	3.1 - 3.1	32 (0k)		3.7 - 21.5	3.1 - 21.8	188 (3k)
TECHNION-HEP	T2	NL				4.0 - 5.0	1.5 - 1.5	13 (0k)		2.8 - 13.6	1.5 - 11.6	115 (2k)
TR-10-ULAKBIM	T2	NL	9	9		0.2 - 0.3	0.9 - 0.9	1 (0k)		0.2 - 7.6	0.9 - 7.1	79 (1k)
JINR-LCG2	T2	RU	100	100		1.9 - 2.4	0.8 - 0.8	8 (0k)		1.1 - 6.1	0.8 - 5.6	53 (1k)
RU-Protvino-IHEP	T2	RU	20	20		0.4 - 0.5	0.4 - 0.4	1 (0k)		0.4 - 3.5	0.4 - 3.2	33 (0k)
UKI-LT2-RHUL	T2	UK	10	10		0.9 - 1.1	0.3 - 0.3	3 (0k)		0.6 - 1.9	0.3 - 1.6	14 (0k)
UKI-NORTHGRID-MAN-HEP	T2	UK	40	40		8.4 - 10.8	2.7 - 2.7	61 (1k)		19.0 - 25.6	18.3 - 18.3	217 (3k)
UKI-SOUTHGRID-RALPP	T2	UK	20	20		0.8 - 0.9	0.6 - 0.6	2 (0k)		0.6 - 5.1	0.6 - 4.7	48 (1k)
UKI-SCOTGRID-GLASGOW	T2	UK	20	20		2.5 - 3.2	1.3 - 1.3	10 (0k)		1.6 - 10.2	1.3 - 9.3	92 (1k)
UKI-LT2-QMUL	T2	UK				7.2 - 8.8	2.9 - 2.9	23 (0k)		5.0 - 19.7	2.9 - 19.7	156 (2k)
UKI-SCOTGRID-ECDF	T2	UK				0.8 - 1.0	0.5 - 0.5	3 (0k)		0.6 - 3.7	0.5 - 3.6	33 (0k)
UKI-NORTHGRID-LANCS-HEP	T2	UK	40	40		5.5 - 6.8	4.4 - 4.4	18 (0k)		3.8 - 35.5	4.4 - 32.0	336 (5k)
UKI-NORTHGRID-LIV-HEP	T2	UK				0.7 - 0.9	0.4 - 0.4	3 (0k)		0.5 - 3.2	0.4 - 2.8	29 (0k)
Taiwan-LCG2 (no tape)	T1	TW	20	20		3.5 - 4.1	1.7 - 1.7	9 (0k)		2.6 - 12.2	1.7 - 10.8	101 (1k)
NET2	T2	US	10	10		0.0 - 0.0	0.0 - 0.0	0 (0k)		0.0 - 0.0	0.0 - 0.0	0 (0k)
SWT2_CPB	T2	US	100	100		9.7 - 12.1	8.5 - 8.5	59 (1k)		58.8 - 83.7	60.7 - 60.7	815 (12k)
AGLT2	T2	US	100	100		9.9 - 12.7	7.0 - 7.0	70 (1k)		47.4 - 67.0	49.3 - 49.3	642 (9k)
OU_OSCER_ATLAS	T2	US	100	100		1.2 - 1.6	0.6 - 0.6	5 (0k)		0.7 - 5.4	0.6 - 4.8	49 (1k)
MWT2	T2	US	200	200		24.2 - 32.0	9.9 - 9.9	193 (3k)		60.0 - 82.0	67.2 - 67.2	720 (10k)
BU_NESE	T2	US				2.8 - 3.7	2.5 - 2.5	14 (0k)		1.5 - 16.7	2.5 - 17.7	161 (2k)
BU_ATLAS_Tier2	T2	US				0.1 - 0.1	0.3 - 0.3	0 (0k)		0.1 - 0.4	0.3 - 0.6	4 (0k)
T2 summary						213.1	107.2			574 - 759	420 - 732	
Summary						680.4	674.2			1323 - 1635	1323 - 1635	

Transfer throughput tests schedule

- Each test should run for at least **48 hours** at the target rates
- Ensure that the timing of the tests aligns with other **experiments** (CMS)
  - **stress network** with same / similar tests **at same time**
  - minimal scenario: T0→T1 export … fine for ATLAS
  - minimal scenario: T1→T2 … fine for ATLAS
  - additional traffic injected in their xroot federation … no corresponding ATLAS test
    - Either continue with "full minimal scenario" (T0 export + "reco") or flexible scenario
  - flexible scenario … fine for ATLAS
- A lot of other proposals for DC24
  - which tests can be run in parallel
  - which tests must be executed in sequence

Next steps

- November 2023
  - finalize list of participating T2 sites (missing response from some clouds)
  - per-link rule injection and deletion rates (scale tests)
- December 2023
  - test dc_inject tool – keep defined transfer rate on given link
- January 2024
  - verify available space at participating sites
- Before DC24
  - configure as many sites as possible to support transfers with tokens
  - configure as many sites as possible to support fireflies
    - currently possible only on storages dedicated to one VO