LHCb requirements

Concezio Bozzi

INFN Ferrara

LHCONE/LHCOPN workshop

CERN January 13th 2020

Outline

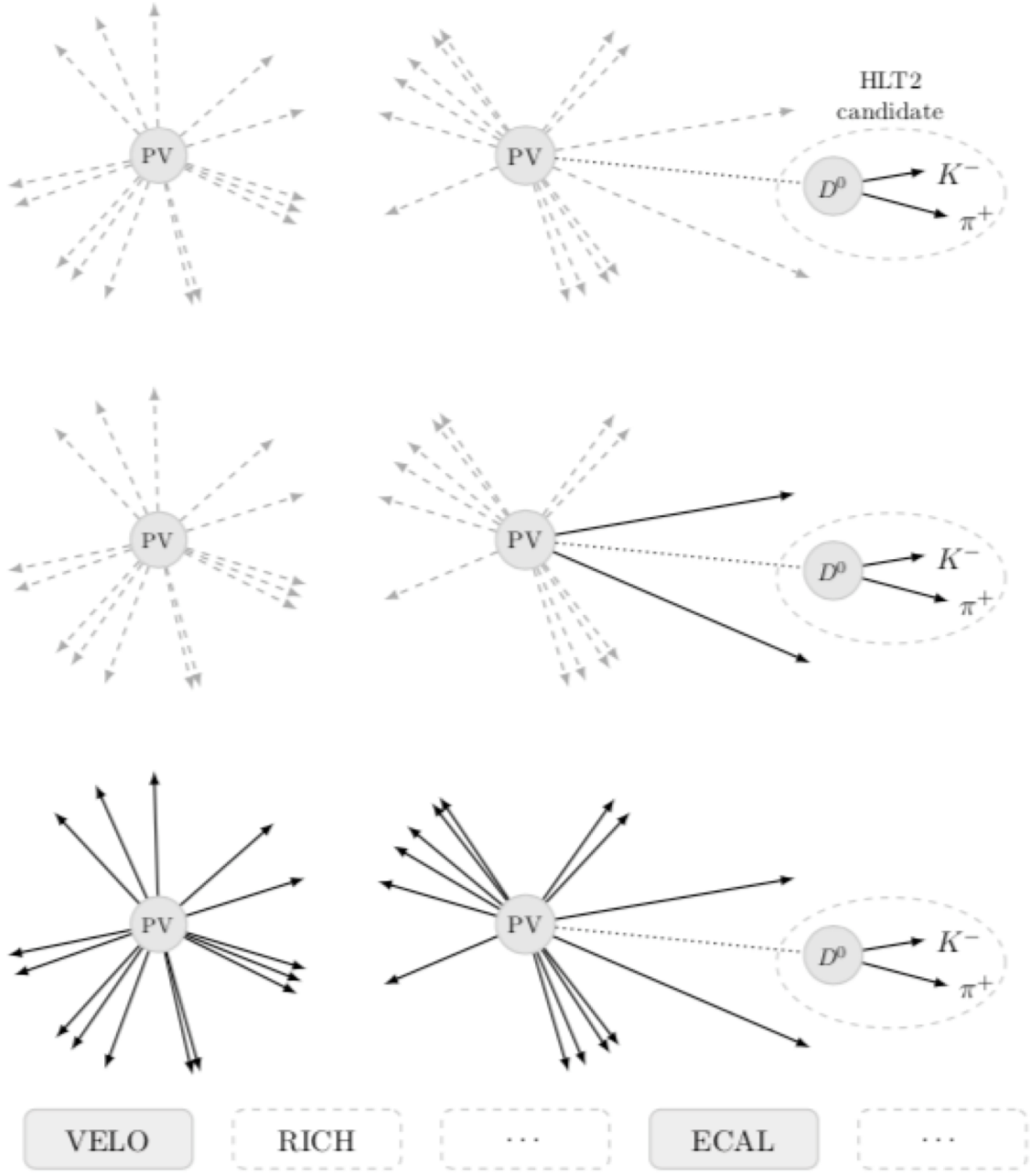
- The LHCb dataflow
- Network utilization during Run2
- The LHCb Run3 upgrade and impact on network requirements
- Final comments

Data streams from the LHCb detector

- Data from the LHCb detector organised in 3 streams
  - **FULL:** «classic» stream, consisting of subdetector RAW banks, events to be promptly reconstructed at Tier0 or Tier1 sites
  - **TURBO:** new concept implemented in Run2. Reconstruction output from HLT farm is saved. Data ready to be analysed, no further processing needed
  - **TURCAL:** calibration stream, with both reconstruction output and (some) RAW banks

Data persistency

- Different levels of persistency:
  - FULL and TURCAL: the full event is persisted
  - TURBO: selective persistency, ranging from candidate firing the trigger to the entire event, optionally including some RAW subdetector data banks



Data streams from the LHCb detector

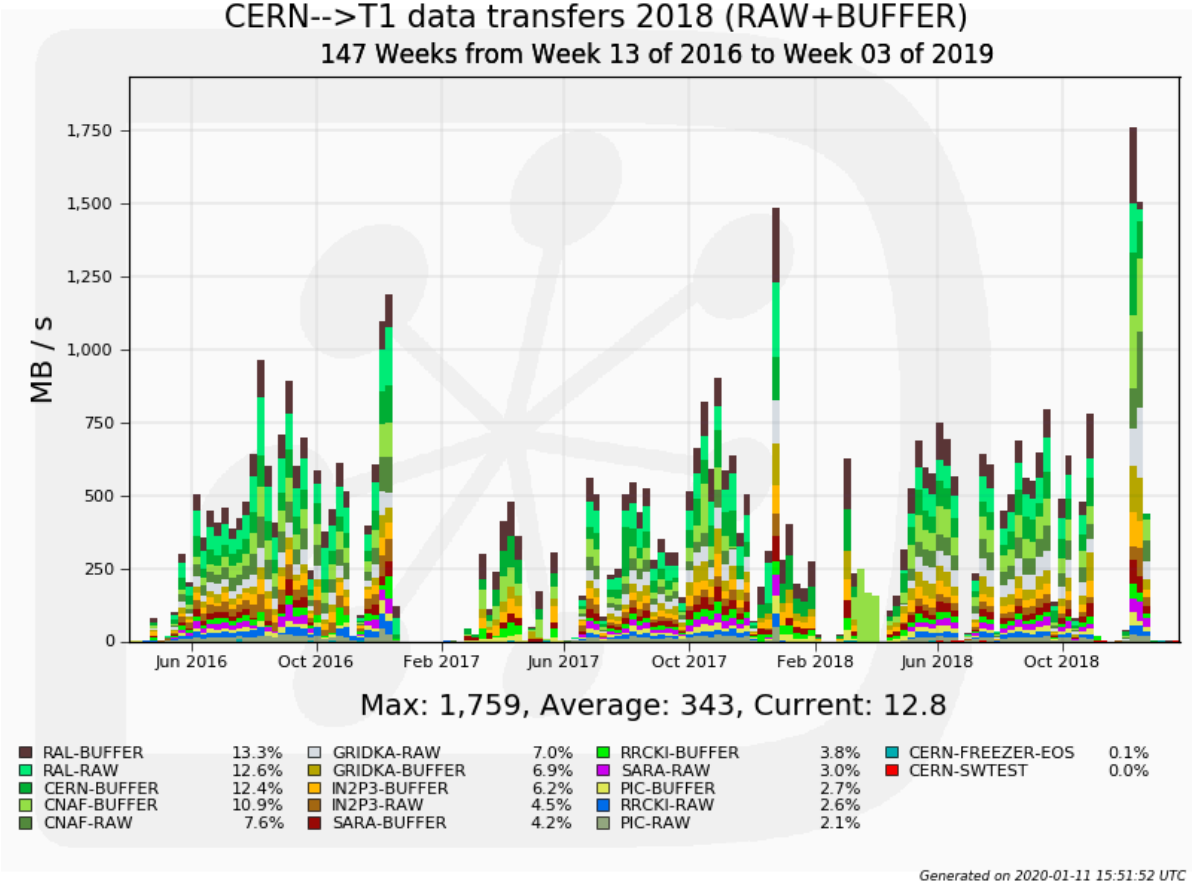
- Due to selective persistency, emphasis has shifted from trigger rate (Hz) to bandwidth (bytes/s)
  - save less information and give more rate for a given bandwidth!
- Example of rate and bandwidth division for 2018 data taking

stream	event size (kB)	event rate (kHz)	rate fraction	throughput (GB/s)	bandwidth fraction
FULL	70	7.0	65%	0.49	75%
Turbo	35 (*)	3.1	29%	0.11	17%
TurCal	85	0.6	6%	0.05	8%
total	61	10.8	100%	0.65	100%

(*) Turbo event size is an average. It ranges from a few kB (minimal persistence) to full event size

Network utilization CERN→Tier1 sites

- The vast majority of transfers from the Tier0 to Tier1 sites is due to moving the three (FULL, TURBO, TURCAL) data streams
- Transfers to Tier1
  - BUFFER for further processing e.g.
    - Reconstruction of FULL stream
    - Reformatting of TURBO stream
  - RAW for tape archival
    - 1 copy for the entire Tier1 ensemble (1 copy at CERN)
- Average ~0.6GB/s, peak up to 1.75GB/s (ion running)

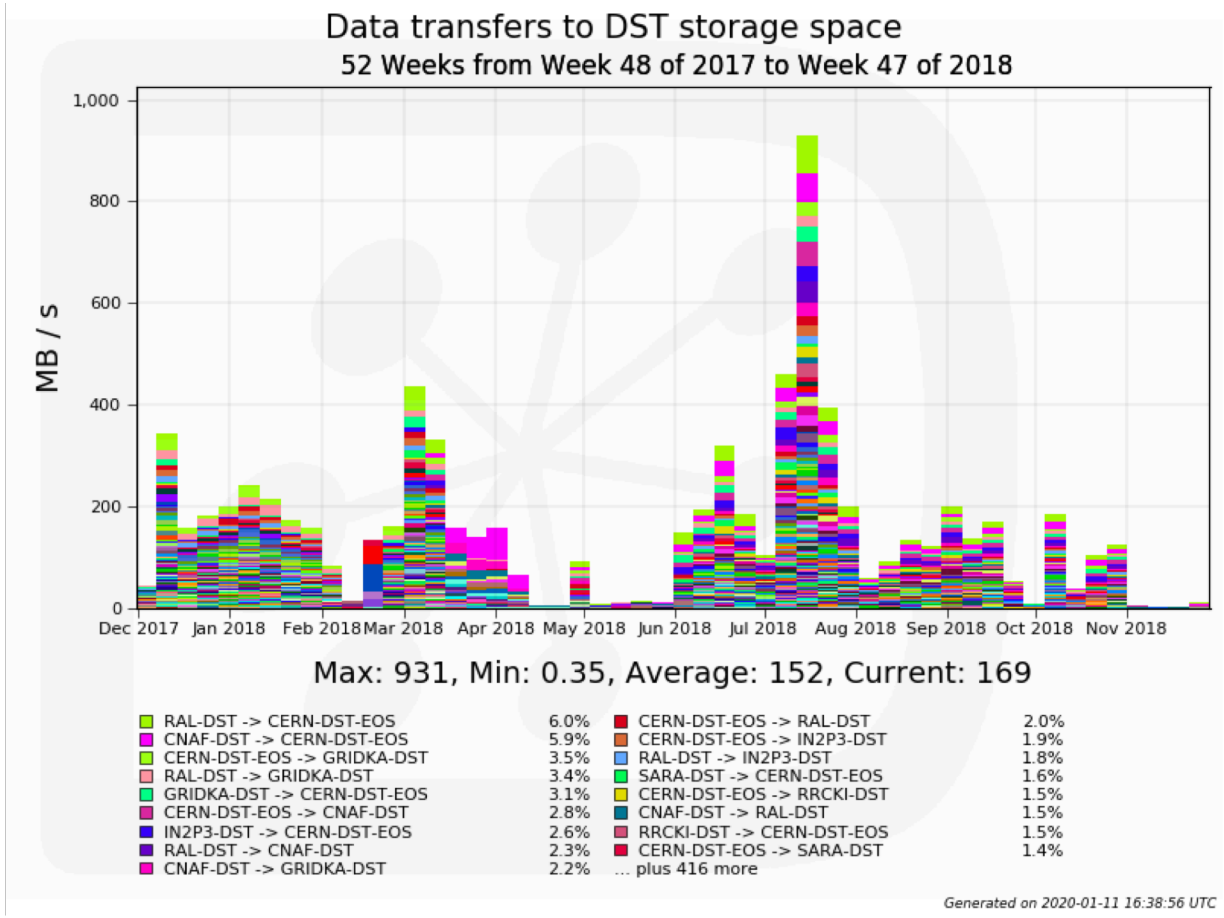


Reconstruction / Stripping

- Reconstruction of FULL is performed at Tier1s (80% of events) and Tier0 (20%)
  - Output as RDST files
  - saved on tape ARCHIVE (1 copy only)
- TURBO does not need to be reconstructed, but only reformatted. Same T0/1 share
- No event re-reconstruction!
  - Alignment and calibration performed online on the trigger farm and applied on HLT
- RDST files are «stripped» according to selection criteria specific to each analysis. Stripping takes place at the same site as reconstruction. Output as
  - DST: full event information; stripping = event filtering
  - mDST: selective persistency; stripping = filtering + slimming
    - The offline equivalent of the TURBO stream
- (m)DST files are merged and grouped in O(10) streams and
  - Stored on tape ARCHIVE (1 copy) and DST disk
  - Replicated to DST disk on either another Tier1 or a Tier2 with disk (Tier2D)
    - 3 copies in total

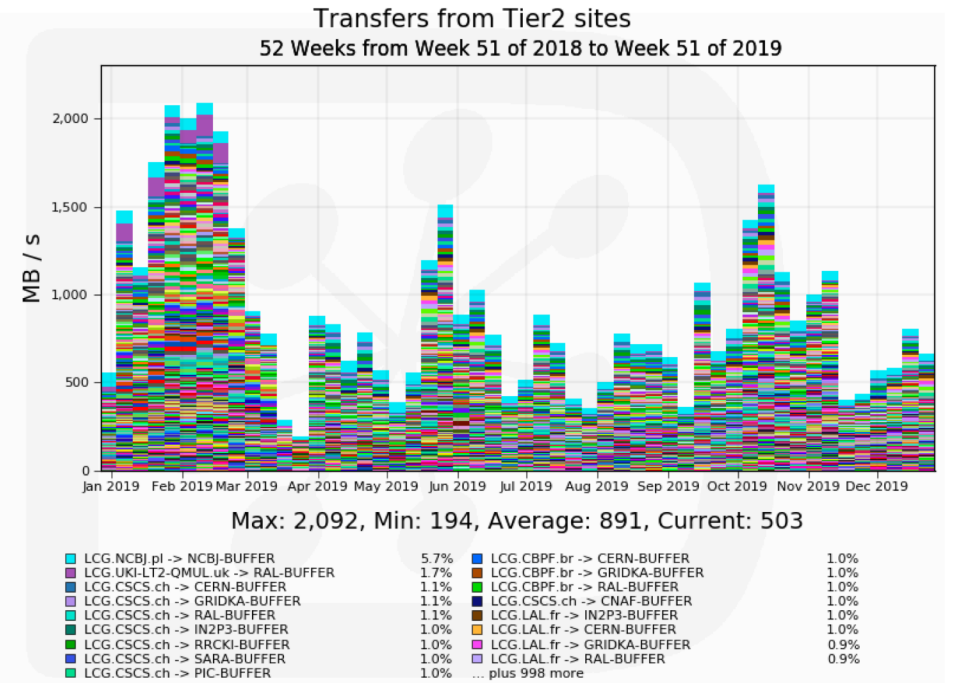
Bandwidth utilisation, stripping

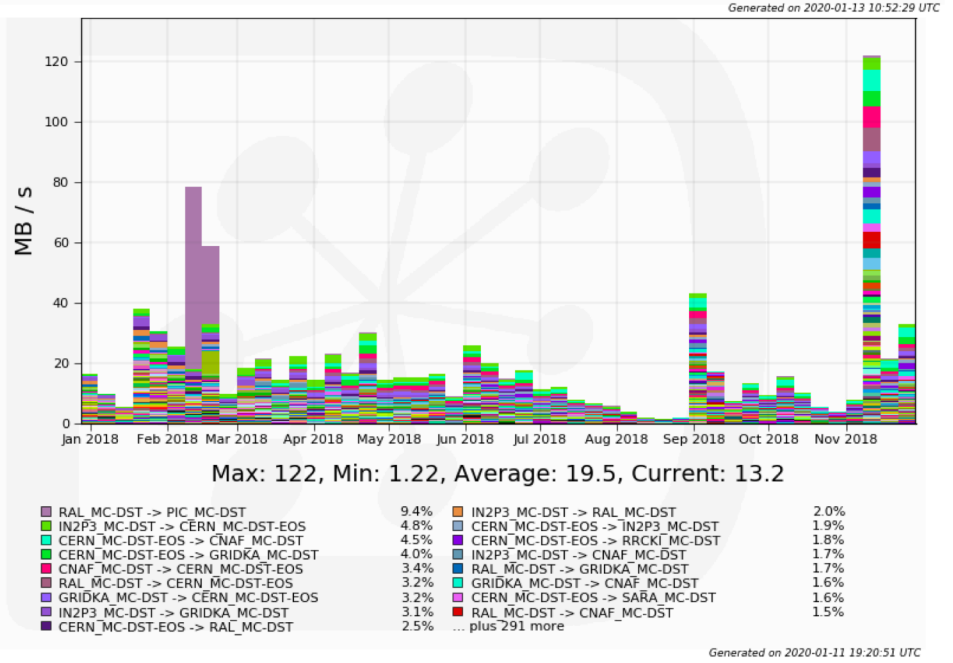
- Stripping output saved to DST disk space on Tier0, Tier1 and Tier2D sites
- Average bandwidth ~200MB/s, peaks at up to 900MB/s
- First-pass stripping is concurrent with data taking
- Full re-stripping or incremental stripping may occur during shut-downs
  - Staging from tape required. 50% of the RAW size is transfered over the network



Monte Carlo simulation

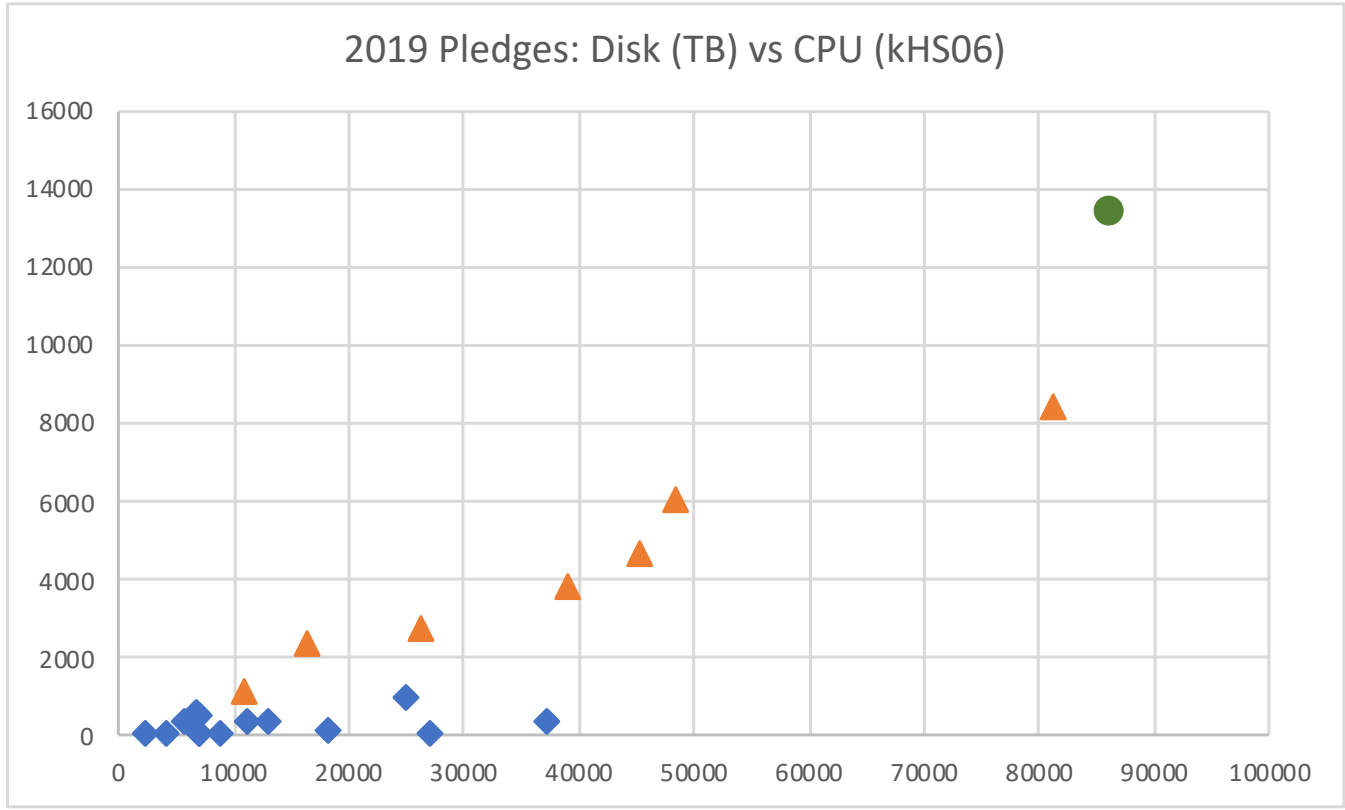
- No input data required. Starting from random seed!
  - Pile-up significantly smaller than GPDs
- Simulation runs everywhere
  - It accounts for 90% of grid jobs, nearly constant throughout the year
- Simulation is produced at
  - Tier0, Tier1 (50%): no network transfer
  - Tier2 (50%): network throughput depends on several factors:
    - MC reconstruction / filtering run locally or somewhere else
    - amount of fast simulation
- About 500MB/s average from Tier2 to Tier0,1 for full simulation, 2GB/s for fast simulation
- Simulation reconstruction is heavily filtered
  - E.g. 40B events simulated in 2019 but only 1.2PB logical volume added
- MC is stored on MC_DST disk and ARCHIVE
- One replica of MC_DST (two disk copies)
  - average throughput below 100MB/s
- Total throughput: O(0.6-2GB/s) depending on fast simulation





Data placement for physics analysis

- Data distribution model quite simple
- Jobs run where data is
  - Mostly at Tier0 and Tier1s
- Number of sites with data relatively small
- Well-balanced CPU and disk resources
  - Grid user jobs are given the highest priority anyway
- No need for caches, pre-placement, etc
- No impact on network



Run3 LHCb Upgrade

- With the upgrade conditions several factors need to be applied
  - Luminosity $4*10^{32}$ $cm^{-2}s^{-1}$ to $2x10^{33}$ $cm^{-2}s^{-1}$
  - HLT efficiency increase because of removal of L0 hardware trigger
  - Raw event size increase due to pileup, according to simulation
- Without any changes the HLT output rate would increase in Run 3 to 17.4 GB/s

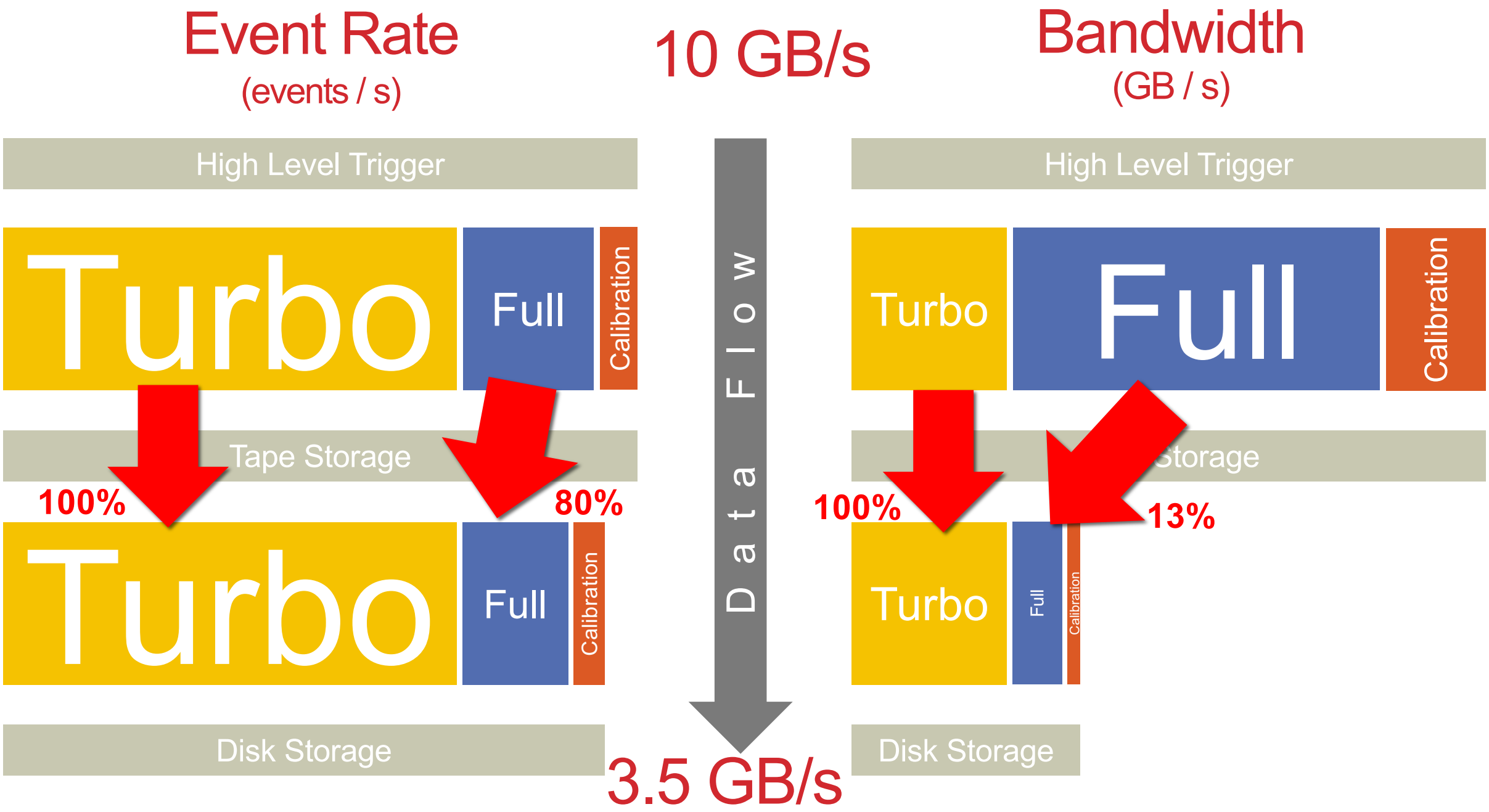
	Run 2 (GB/s)	Lumi	No L0	Raw size	Run 3 (GB/s)
Full	0.49	x5	x2	x3	14.7
Turbo	0.11	x5	x2	x1	1.1
Calibration	0.05	x5	x2	x3	1.6
Total	0.66				17.4

Event size:
Turbo/FULL ~0.167

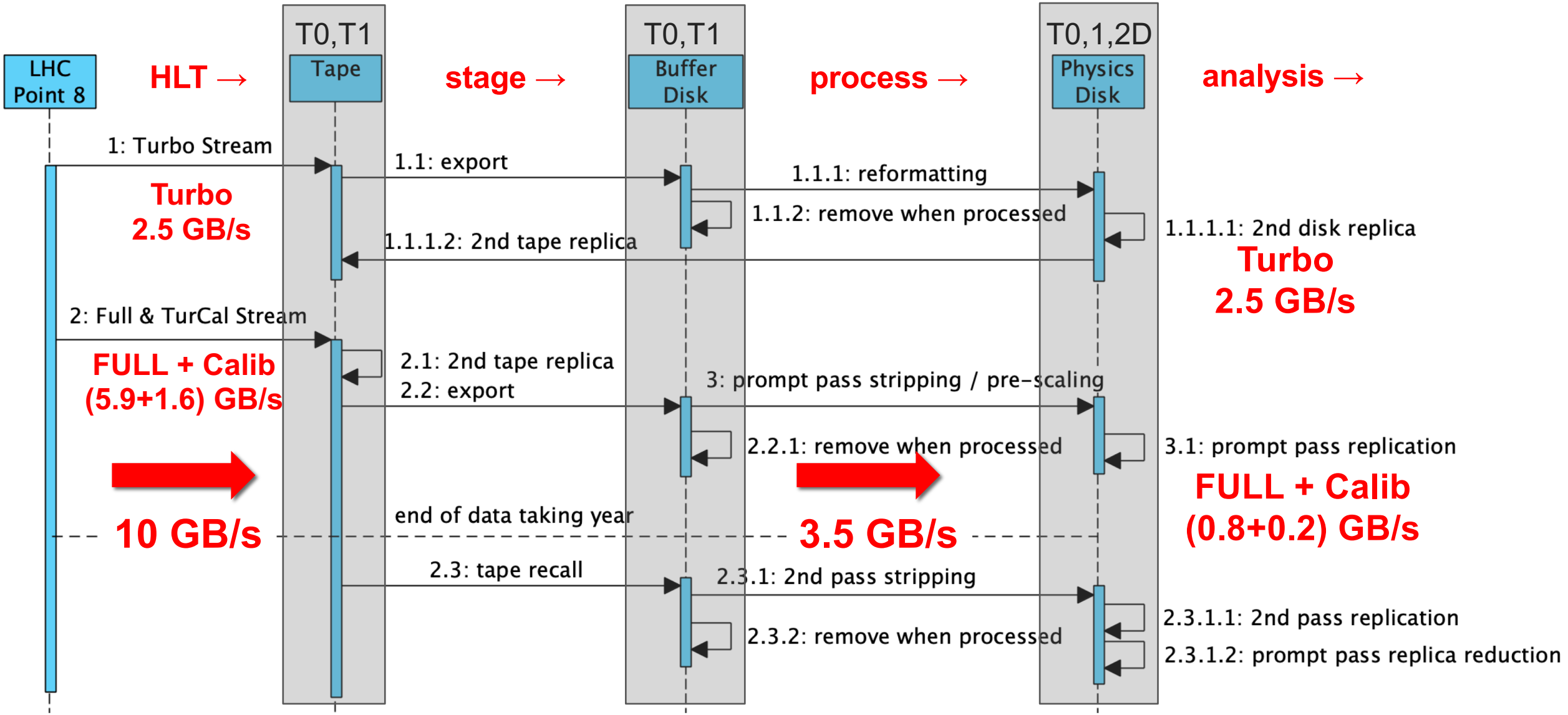
Mitigation of the HLT output bandwidth

- Moving a larger fraction of the physics programme to TURBO decreases the output bandwidth and preserves the physics reach of LHCb
- About 60% of the physics selections currently on FULL migrating to TURBO
  - Massive migration, not trivial!
- Logical bandwidth to tape: 10 GB/s
- Logical bandwidth to disk reduced to 3.5GB/s by stripping FULL and TURCAL more aggressively (select substantial fraction but slim by factor 6)
- This gives requirements of O(100PB) tape and O(50PB) disk per data taking year

		Throughput to tape		Throughput to disk	
stream	rate fraction	throughput (GB/s)	bandwidth fraction	throughput (GB/s)	bandwidth fraction
FULL	26%	5.9	59%	0.8	22%
Turbo	68%	2.5	25%	2.5	72%
TurCal	6%	1.6	16%	0.2	6%
total	100%	10.0	100%	3.5	100%


Event Rate
(events / s)
10 GB/s
Bandwidth
(GB / s)
High Level Trigger
Turbo
Full
Calibration
Tape Storage
100%
80%
Turbo
Full
Calibration
Disk Storage
Data Flow
3.5 GB/s
High Level Trigger
Turbo
Full
Calibration
Storage
100%
13%
Turbo
Full
Calibration
Disk Storage

Data Processing Workflow per Data Taking Year



Monte-Carlo production in Run3 onwards

- Amount of events to be simulated scales with integrated luminosity
- Limit CPU by increasing usage of fast simulations
  - But this has a big impact on network traffic
- Limit storage and network usage by
  - Filtering in generation and stripping
  - Saving output in mDST format (x40 size reduction)
- As a result, expect to generate a volume of O(10 PBs) of simulated data per year
  - 1/3 is kept on (MC_DST) disk, the rest is parked on tape
  - One disk replica is made, this gives an estimation of O(1 GB/s) network traffic
- If MC reconstruction is split and fast simulation dominates, then transfers of simulation output from Tier2 sites becomes dominant
  - O(5-10GB/s) as a ballpark estimate, to be further discussed

Consequences on network utilization

- The data distribution model will not change
- Scale throughputs measured in Run2 by the appropriate factors
- Total ballpark estimate, with O(20%) contingency:

20GB/s on average

Data type	Source	Destination	Run2 (GB/s)		Run3/4 (GB/s)	
			average	peak	average	peak
Primary datasets	T0	T1	0.6	1.7	10	20
Stripping output	T0, T1	T0, T1, T2D	0.2	0.9	1.7	7.5
Monte Carlo	All sites	T0, T1, T2D	0.6	2	5	10
TOTAL			**1.5**		**17**	

General considerations

- The current trend of allowing users to run docker/singularity images could impact network utilization, since this ultimately requires downloading the image on each worker node. That should be carefully thought about

- In terms of features, could a minimal QoS per user of the network be introduced?
  - Network is the only resource for which there is no pledge nor fairshare

- Our main concern for network in future is bandwidth availability
  - Non-LHC users are coming with large requirements

- Monitoring and performance: we regard ourselves as just users; of course we are willing to help with requirements/use cases/providing info/etc

Final remarks

- The fast and reliable network provided to us in the past years is at the basis of our successful computing operations and ultimately of the physics productivity of LHCb
- Running jobs where the data is → big impact on network usage
- We are not big consumers right now, however we will increase our network usage by a factor 10 in Run3 and beyond
- Given the increasing role of Tier2 sites, it would be good to consolidate network towards them
- Network QoS would be very useful
- More structured monitoring would be welcome

backup

Run3 Computing model in a nutshell

- LHCb Upgrade computing model accommodates a trigger output BW of 10 GB/s
  - Massive usage of novel event selection (Turbo) and event size reduction (selective persistence) techniques
  - Save the full bandwidth on cheap storage
  - Reduce by more than a factor of 2 disk requirements using the above techniques
- CPU needs dominated by MC production
  - Massive use of faster simulation techniques
- In summary:
  - Substantial reduction of expensive resources
  - Maintain the full breadth of the physics programme
  - Flexible: incorporate future technology advancements

LHCb Run3 Computing Model assumptions						
L (cm^{-2} s^{-1})	2×10^{33}					
Pileup	6					
Running time (s)	5×10^6 (2.5×10^6 in 2021)					
Integrated luminosity	10 fb^{-1} (5 fb^{-1} in 2021)					
Trigger rate fraction (%)	26 / 68 / 6 Full/Turbo/TurCal					
Logical bandwidth to tape (GB/s)	10 (5.9 / 2.5 / 1.6 Full/Turbo/TurCal)					
Logical bandwidth to disk (GB/s)	3.5 (0.8 / 2.5 / 0.2 Full/Turbo/TurCal)					
Ratio Turbo/FULL event size	16.7%					
Ratio full/fast/param. MC	40:40:20					
HS06.s per event for full/fast/param. MC [a]	1200 / 400 / 20					
Number or MC events[b]	2.3×10^9 / fb^{-1} / year					
Data replicas on tape	2 (1 for derived data)					
Data replicas on disk	2 (Turbo); 3 (Full, TurCal)					
MC replicas on tape	1 (MDST)					
MC replicas on disk	0.3 (MDST, 30% of the total dataset)					
Resource requirements						
WLCG Year	Disk (PB)		Tape (PB)		CPU (kHS06)	
2021	66	1.1	142	1.5	863	1.4
2022	111	1.7	243	1.7	1579	1.8
2023	159	1.4	345	1.4	2753	1.7
2024	165	1.0	348	1.0	3467	1.3
2025	171	1.0	351	1.0	3267	0.9

[a] corresponding to 120, 40, 2s on a 10HS06 computing core

[b] simulation of year N starts in year N+1