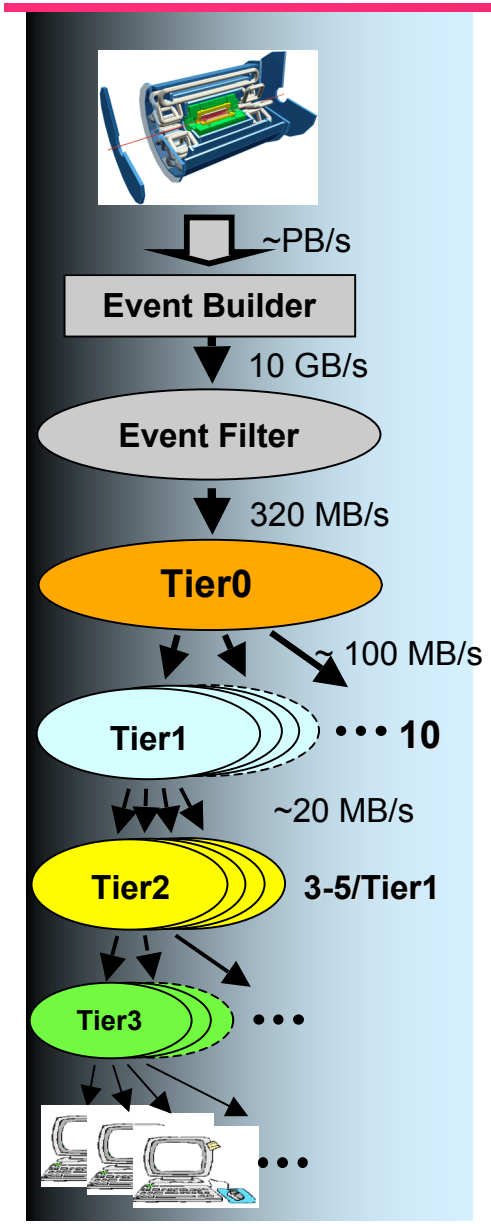
ATLAS Megatable

Dario Barberis

CERN & Genoa University/INFN

Data replication and distribution



In order to provide a reasonable level of data access for analysis, it is necessary to replicate the ESD, AOD and TAGs to Tier-1s and Tier-2s.

RAW:

- Original data at Tier-0
- Complete replica distributed among all Tier-1
  - Randomized dataset to make reprocessing more efficient

ESD:

- ESDs produced by primary reconstruction reside at Tier-0 and are exported to 2 Tier-1s
- Subsequent versions of ESDs, produced at Tier-1s (each one processing its own RAW), are stored locally and replicated to another Tier-1, to have globally 2 copies on disk

AOD:

- Completely replicated at each Tier-1
- Partially replicated to Tier-2s (~1/3 – 1/4 in each Tier-2) so as to have at least a complete set in the Tier-2s associated to each Tier-1
  - Every Tier-2 specifies which datasets are most interesting for their reference community; the rest are distributed according to capacity

TAG:

- TAG databases are replicated to all Tier-1s (Oracle)
- Partial replicas of the TAG will be distributed to Tier-2 as Root files
  - Each Tier-2 will have at least all Root files of the TAGs that correspond to the AODs stored there

Samples of events of all types can be stored anywhere, compatibly with available disk capacity, for particular analysis studies or for software (algorithm) development.

Computing Model and Resources

- The ESD size kept increasing during the last few years, as the ability to (partially) reprocess the events starting from the ESD needs more input information (calorimeter cells, InDet hits)
  - We think that, as we understand the detector better, we will be able to reduce the amount of extra information in the ESD
- We have therefore to provide for larger ESD in early years, reaching design value for higher luminosity in 2010:
  - Still can only allow ~1 MB/event in 2007/8 (instead of 0.5 MB/event original design)
- Having now a much better detector geometry representation than in the past, and operating the default simulation over $|\eta|<6$ (instead of $|\eta|<3$), the simulation time per event increased from 100 kSI2k-sec/event to 400 kSI2k-sec/event on average
  - We think that for many channels we will be able to use shower parameterisation in the calorimeters, but its performance is still under test
- In the first years of operation, there will be the need to tune calibration and reconstruction algorithms on real data; we have therefore increased the available CPU for user reconstruction with a decreasing profile with time
- We have also increased the level of simulation from 20% to 30% of the raw data rate following the LHCC recommendations
- The new resource calculation takes into account the agreed machine schedule for the first years of operation, the above changes in ATLAS input numbers and the global envelope of resource pledges according to the WLCG MoU

LHC schedule used for resource calculations

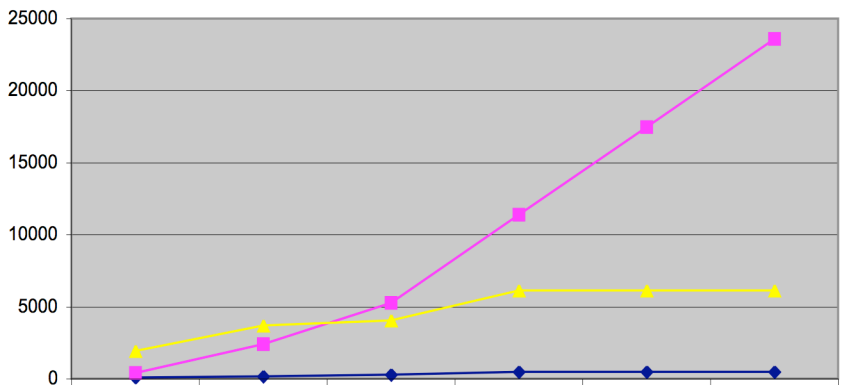
year	*energy*	*luminosity*	*physics beam time*
2007	**450+450 GeV**	**$5x10^{30}$**	**protons - 26 days at 30% overall efficiency ➔ $0.7*10^6$ seconds**
2008	**7+7 TeV**	**$0.5x10^{33}$**	**protons - starting beginning July ➔ $4*10^6$ seconds** **ions - end of run - 5 days at 50% overall efficiency ➔ $0.2*10^6$ seconds**
2009	**7+7 TeV**	**$1x10^{33}$**	**protons:50% better than 2008 ➔ $6*10^6$ seconds** **ions: 20 days of beam at 50% efficiency ➔10^6 seconds**
2010	**7+7 TeV**	**$1x10^{34}$**	**TDR targets:** **protons: ➔ 10^7 seconds** **ions: ➔ 10^6 seconds**

Total ATLAS Requirements start 2008

	CPU (MSI2k)	Disk (PB)	Tape (PB)
Tier-0	3.7	0.15	2.4
CAF	2.1	1.0	0.4
Sum of Tier-1s	18.1	9.9	7.7
Sum of Tier-2s	17.5	7.7	0.0

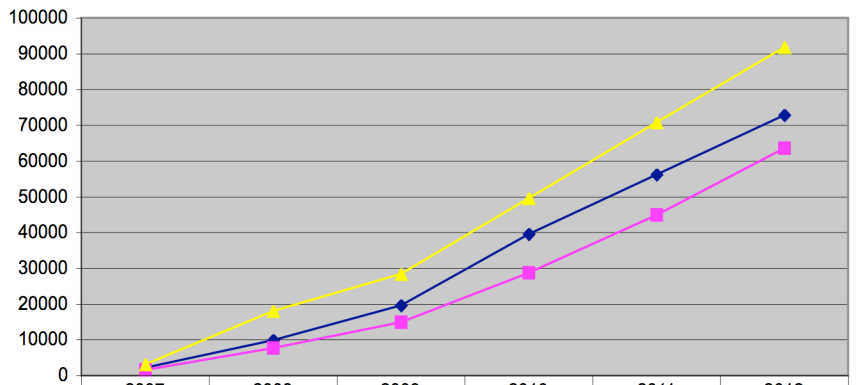
Evolution

New T0 Evolution



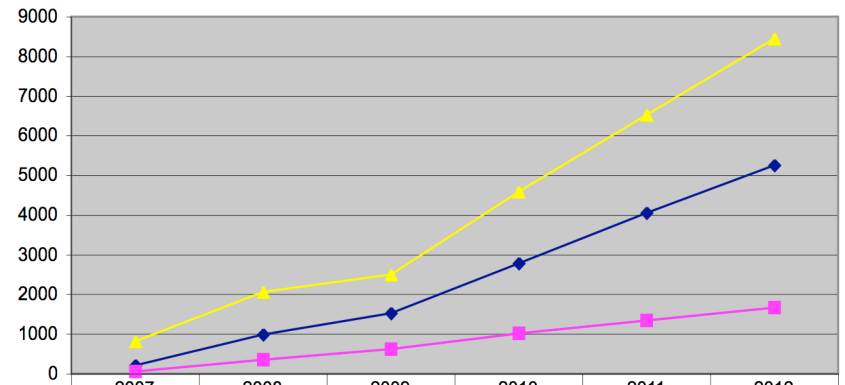
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	75.14785714	152.4621429	277.3242857	472.3528571	472.3528571	472.3528571
Total Tape (TB)	381.3075	2381.711	5267.2345	11371.158	17475.0815	23579.005
Total CPU (kSI2k)	1910	3705	4058	6105	6105	6105

New T1 Evolution



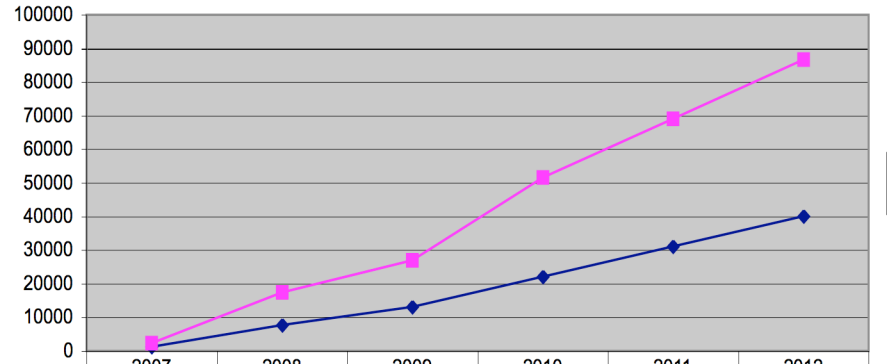
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	2157.0332	9938.696929	19686.41793	39487.79764	56190.82307	72893.8485
Total Tape (TB)	1543.186667	7693.996427	14949.57676	28698.0172	44929.67775	63644.55841
Total CPU (kSI2k)	3173.323529	18122.83529	28423.02353	49573.22353	70723.42353	91873.62353

New CAF Evolution



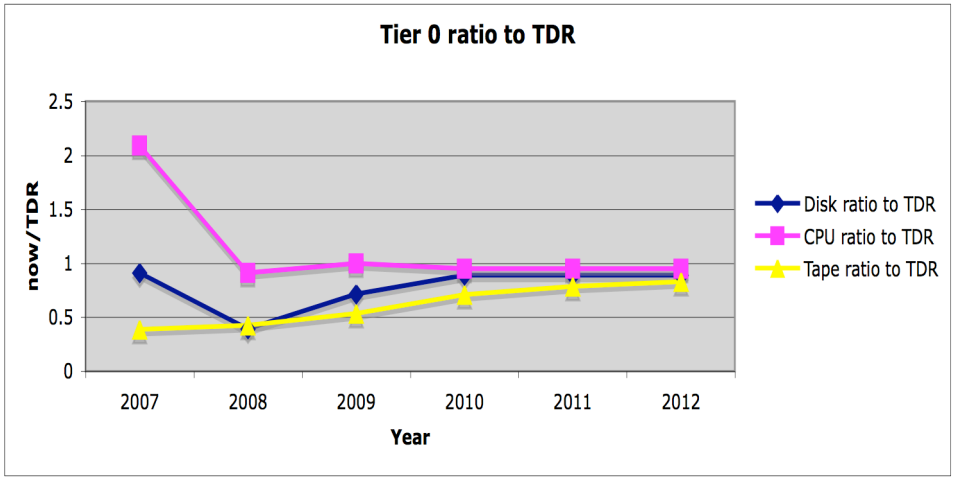
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	212.2436607	986.3915464	1529.026057	2777.498914	4047.976771	5255.197486
Total Tape (TB)	57.3206625	356.5720482	625.1016482	1017.151648	1342.801648	1668.451648
Total CPU (kSI2k)	821	2069	2502	4596	6523	8450

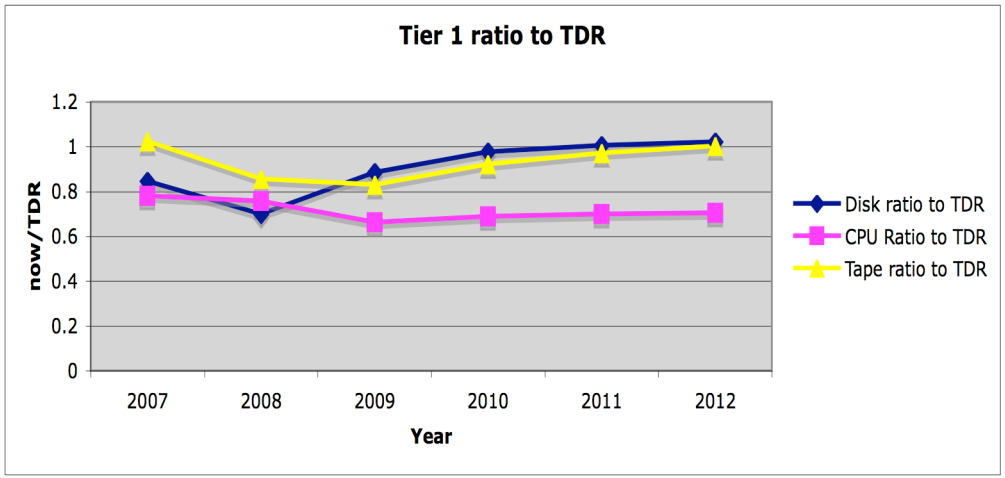
New T2 Evolution

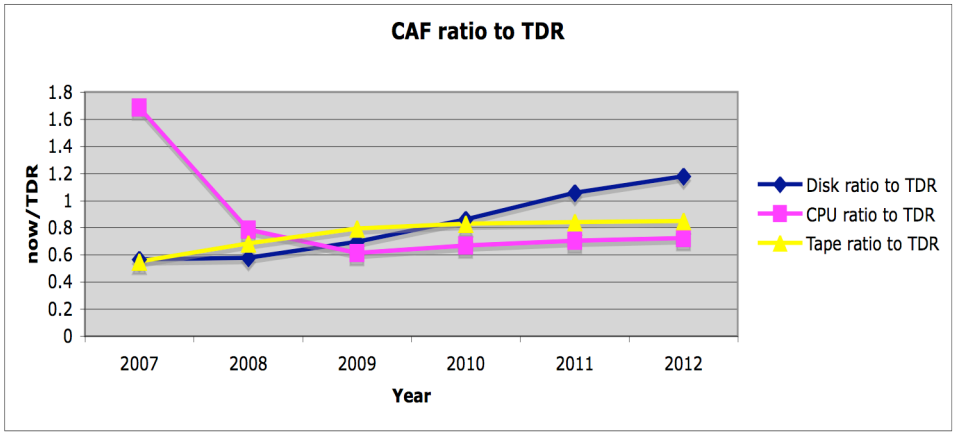


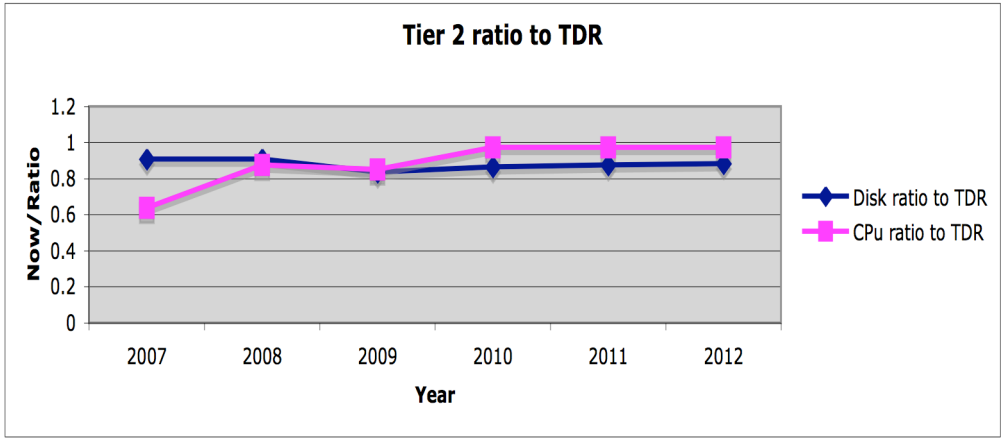
	2007	2008	2009	2010	2011	2012
Disk (TB)	1259.040486	7744.368955	13112.03563	22132.30423	31091.45139	40050.91999
CPU (kSI2k)	2336.108333	17494.50644	26972.75589	51544.63737	69128.41886	86712.20034

Ratio to Computing TDR









NB: there was a mistake in the Tier-0/CAF CPU requirements for 2007 in the Computing TDR (subsequently corrected already in 2005): CPU for calibrations does not scale with the length of the data-taking period

Observations on computing resources

- Data storage requirements generally fall with reduced live-time (obviously)
- CPU does not fall as much
  - CERN CPU determined by rate and calibration requirements
  - More calibration and optimisation is needed for 2007 data
- Higher than hoped simulation time per event
- Tier-1s see significant reductions
  - Cumulative effect of less data on reprocessing
- Tier-2s see a small initial fall but are bigger after 2009
- There is an argument for spreading the gain and the pain with Tier-1s by introducing more flexibility in the model:
  - Tier-1s can now produce simulated data when not fully busy with reprocessing

Megatable inputs

- Disk space is the resource that limits our total computing capacity right now
  - This is true for both Tier-1s and Tier-2s
- We distribute RAW and 1st-pass processed ESD data to Tier-1s approximately according to their pledged DISK capacities
  - After removing the disk space for AODs (100 kB/event)
  - BNL in addition gets a full set of ESD data
- Our data distribution model requires a coupling between Tier-1s
  - As we keep 2 copies of the most recent version of ESD data on disk (and one on tape at the production site)
  - Reprocessed ESDs also get exchanged between paired Tier-1s of similar capacity
    - BNL⇔IN2P3CC+FZK, NIKHEF/SARA⇔ASGC+TRIUMF, RAL⇔CNAF, PIC⇔NDGF
- The Tier-2s receive from "their" Tier-1 a fraction of AOD (and smaller fractions of ESD and RAW) compatible with their DISK size
  - Tier-2s in European countries with a Tier-1 get globally a full set of AOD
  - Some "large" Tier-2s also get a full set of AOD
- CPU capacity at Tier-2s is split between analysis (proportional to the data they hold on disk) and simulation
  - Simulation output is transferred to Tier-1s that have storage capacity

Tier-1—Tier-2 Associations

- ASGC
  - Tier-2s in Taiwan and Melbourne
  - Receives simulation from Tokyo and Beijing
  - Can send AODs etc. to Tokyo and Beijing
- BNL
  - All Tier-2s in the USA
- CNAF
  - All Tier-2s in Italy
- FZK
  - All Tier-2s in Germany
  - CSCS, Krakow, Prague, Innsbruck
- IN2P3-CC
  - All Tier-2s in France
  - Romania
  - Sends AODs etc. to Tokyo and Beijing
- NDGF
  - NDGF (Tier-2) and Ljubljana
- NIKHEF/SARA
  - Tier-2s in Russia and Israel
  - Receives simulation from NorthGrid (UK)
  - [Receives any surplus simulation from Prague]
- PIC
  - All Tier-2s in Spain and Lisbon
- RAL
  - Sends AODs etc. to all Tier-2s in the UK
  - Receives simulation from all Tier-2s in the UK except NorthGrid
- TRIUMF
  - All Tier-2s in Canada (not yet official, therefore not in the Megatable)

Remarks

- All data belong to the ATLAS Collaboration and are distributed according to the needs of the Collaboration
  - Including the needs of all collaborating institutions
- The data rates from simulation production are generally much lower than those for real data
  - Therefore it does not really matter in which direction we transfer them from the point of view of network capacity
  - We try to optimise the usage of storage facilities by introducing flexibility in simulation production and storage
- As storage capacities will evolve at different rates for each computing centre in the future, our data distribution pattern may be revised, at least on a yearly basis

Tier-3s and other resources

- Some Grid/ATLAS sites are not part of the official list of Tier-2s but will be able to contribute to simulation production
  - None of them will be a large facility
- From the point of view of our infrastructure, they are not different from Tier-2 sites
  - Their simulation production will be directed to the most convenient Tier-1 site
  - Data transfer to these sites will be driven by local needs (low rates)
- Only really open point is the situation in Latin America
  - "Planned" Tier-2 in Brasil
  - "Planned" Tier-3s in Argentina, Chile and Colombia

Megatable summary (2008)

TOTALS			T0=>T1	T2=>T1	T1=>T1	T1=>T2	T1<=T1	Storage for T2 TByte			Storage for T1 TByte		
	Total Tape Tby	Eff. Disk Tbyte	MByte/s	MByte/s av	MByte/s in	MByte/s av	MByte/s out	Tape1-Disk	Tape1-Disk	Tape0-Disk	Tape1-Disk	Tape1-Disk	Tape0-Disk
ASGC	534.3	848.9	72.1	45.8	75.8	25.4	84.9	211.8	80.7	158.6	161.7	80.1	274.9
BNL	1344.0	3710.5	240.2	92.4	278.7	225.7	228.0	427.5	162.9	320.0	504.0	249.6	1864.8
CNAF	464.4	761.4	70.7	35.8	74.3	53.2	82.7	165.7	63.1	124.1	157.5	78.0	267.8
FZK	482.1	869.6	91.7	23.9	97.0	91.2	115.8	110.3	42.1	82.6	220.5	109.2	374.9
IN2P3	1313.0	1928.1	112.7	139.2	119.8	94.9	148.9	643.8	245.3	482.0	283.5	140.4	482.0
NDGF	172.7	362.2	56.7	0.0	59.2	0.0	60.6	0.0	0.0	0.0	115.5	57.2	196.4
NIKHEF	653.2	1142.4	109.2	38.3	116.0	60.0	143.3	177.4	67.6	132.8	273.0	135.2	464.1
PIC	360.9	582.2	56.7	29.5	59.2	38.6	60.6	136.3	51.9	102.0	115.5	57.2	196.4
RAL	894.7	1410.1	105.7	78.6	112.2	27.9	137.8	363.6	138.6	272.2	262.5	130.0	446.3
TRIUMF	166.4	349.0	55.3	0.0	57.6	0.0	58.4	0.0	0.0	0.0	111.3	55.1	189.2
SUM	6385.6	11964.2	970.9	483.5	1049.7	617.0	1121.1	2236.4	852.2	1674.2	2205.0	1092.0	4756.5

- Add ~80 MB/s from CERN to BNL for their full set of ESD data
- Canadian Tier-2s are not official yet and therefore not counted; they will approximately balance the size of the TRIUMF Tier-1
- Israel is also not counted
- Prague goes entirely to FZK