ATLAS Dataflows and STEP09

Graeme Stewart

2009-05-12

University of Glasgow | Department of Physics & Astronomy

Objectives for ATLAS

- Make a parallel test of all principle ATLAS offline computing activities
  - At nominal data taking rates
- We use the ATLAS Computing Model and follow it as closely as we can
  - Reprocessing and reconstruction will happen in T1s
    - Simulation will be pushed out to the T2s
  - T2s will balance their resources between simulation (50%) and user analysis (50%)
- Plan to run 1-14 June

ATLAS Activities

- Data export from CERN to T1s and T2s
- Reprocessing exercise at T1s
  - Tape reading and writing
  - Post-reprocessing data export
- Simulation Activities
  - Geant 4 at Tier-2s
  - Reconstruction at Tier-1s
- User analysis challenge
  - Tier-2s for AOD and DPD analysis
- Please see the twiki:

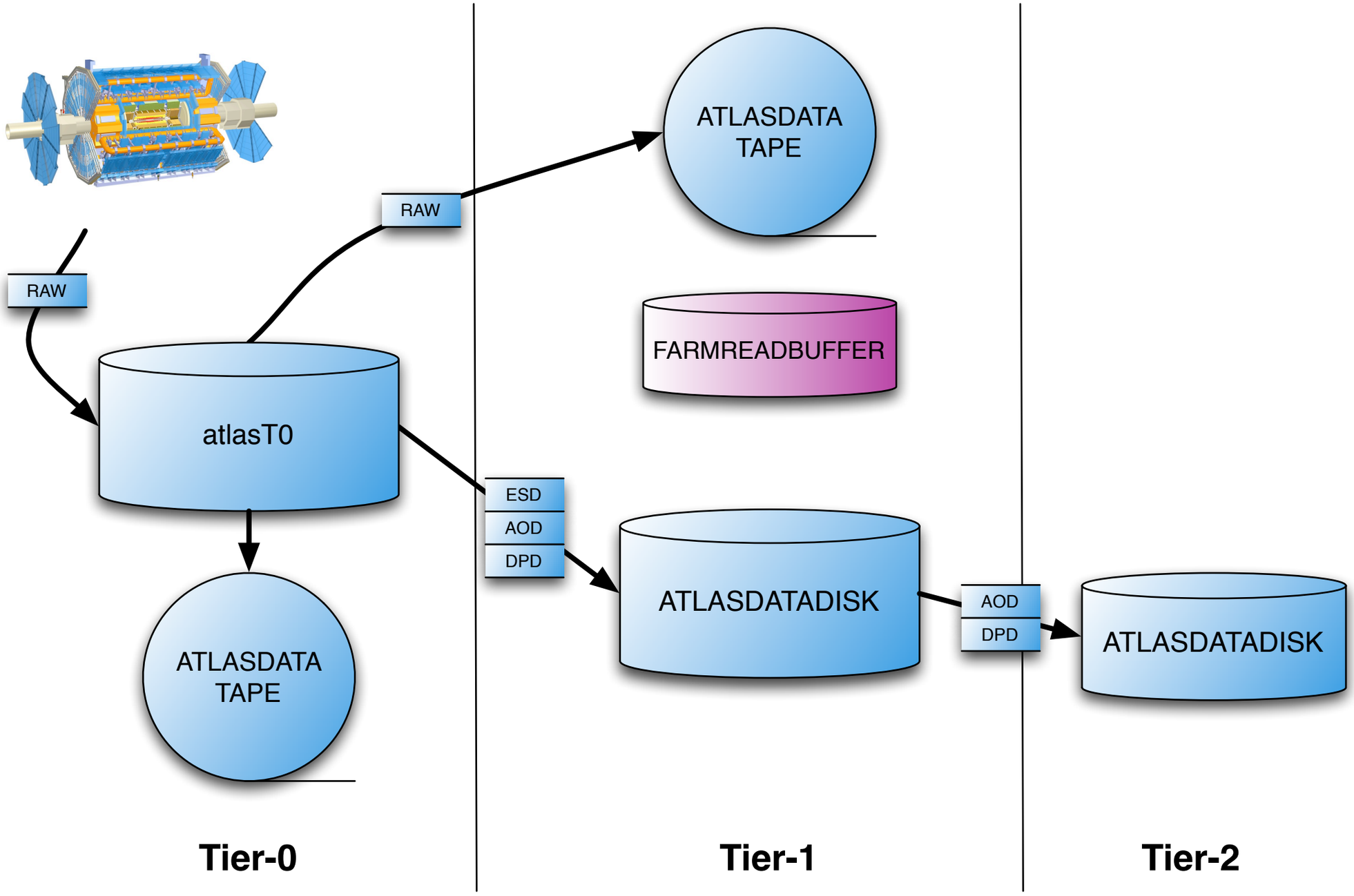
  https://twiki.cern.ch/twiki/bin/view/Atlas/Step09

Data Export from CERN

- Export of data at nominal rates from CERN
  - 200Hz trigger rate, 50k seconds of beam a day
- RAW goes as MoU share
- ESD goes as 2x MoU share
- AOD and DPD, 100% goes to all T1s

	Data Types	Size/Event
RAW	Detector outputs	1.6MB
ESD	Event Summary Data	1.0MB
AOD	Analysis Object Data	0.2MB
DPD	Derived Physics/Performance Data	0.2MB

T0 Export Picture



T0→T1 Rates and Volumes

Rates (MB/s)

Tier-1	MoU Share	RAW	ESD	AOD+DPD	ATLASDATATAPE	ATLASDATADISK	Total
ASGC	0.05	9.3	11.6	46.3	9.3	57.9	67.1
BNL	0.25	46.3	57.9	46.3	46.3	104.2	150.5
CNAF	0.05	9.3	11.6	46.3	9.3	57.9	67.1
FZK	0.10	18.5	23.1	46.3	18.5	69.4	88.0
LYON	0.15	27.8	34.7	46.3	27.8	81.0	108.8
NDGF	0.05	9.3	11.6	46.3	9.3	57.9	67.1
PIC	0.05	9.3	11.6	46.3	9.3	57.9	67.1
RAL	0.10	18.5	23.1	46.3	18.5	69.4	88.0
SARA	0.15	27.8	34.7	46.3	27.8	81.0	108.8
TRIUMF	0.05	9.3	11.6	46.3	9.3	57.9	67.1
Total	1.00	185.2	231.5	463.0	185.2	694.4	879.6

STEP09 Volume (TB in 2 weeks)

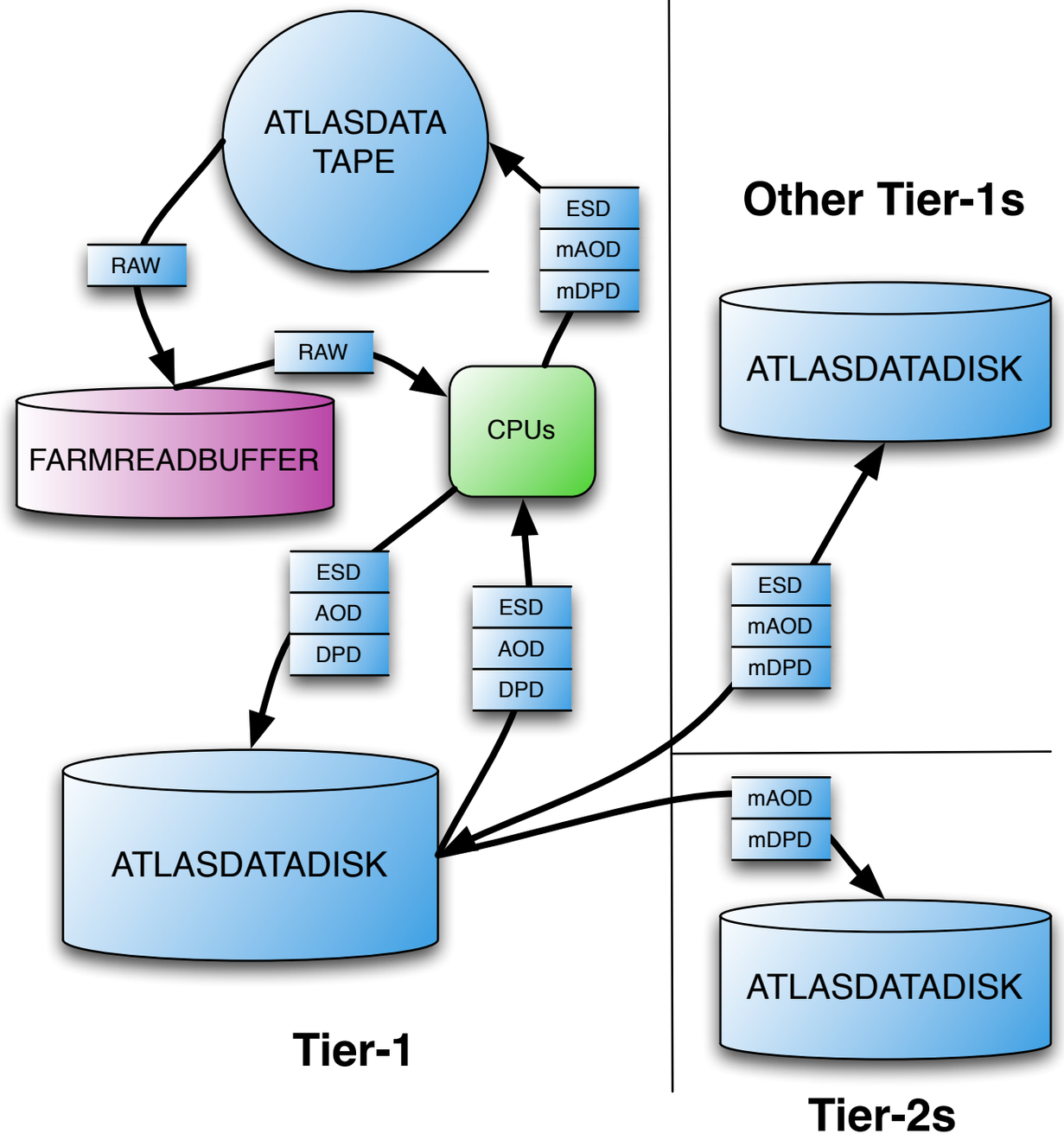
Tier-1	MoU Share	RAW	ESD	AOD+DPD	ATLASDATATAPE	ATLASDATADISK	Total
ASGC	0.05	11.2	14.0	56.0	11.2	70.0	81.2
BNL	0.25	56.0	70.0	56.0	56.0	126.0	182.0
CNAF	0.05	11.2	14.0	56.0	11.2	70.0	81.2
FZK	0.10	22.4	28.0	56.0	22.4	84.0	106.4
LYON	0.15	33.6	42.0	56.0	33.6	98.0	131.6
NDGF	0.05	11.2	14.0	56.0	11.2	70.0	81.2
PIC	0.05	11.2	14.0	56.0	11.2	70.0	81.2
RAL	0.10	22.4	28.0	56.0	22.4	84.0	106.4
SARA	0.15	33.6	42.0	56.0	33.6	98.0	131.6
TRIUMF	0.05	11.2	14.0	56.0	11.2	70.0	81.2
Total	1.00	224.0	280.0	560.0	224.0	840.0	1,064.0

Reprocessing

- T1 Reprocessing involves the pre-stage of RAW from ATLASDATATAPE to the read buffer
- Outputs are the normal ATLAS data products: ESD, AOD, DPD
  - Plus histograms, TAGS and other data quality metrics
- Final outputs are written back to tape
  - This is after suitable merging to increase file sizes
  - Outputs are then distributed to other T1s and T2s

T1 Reprocessing Workflow

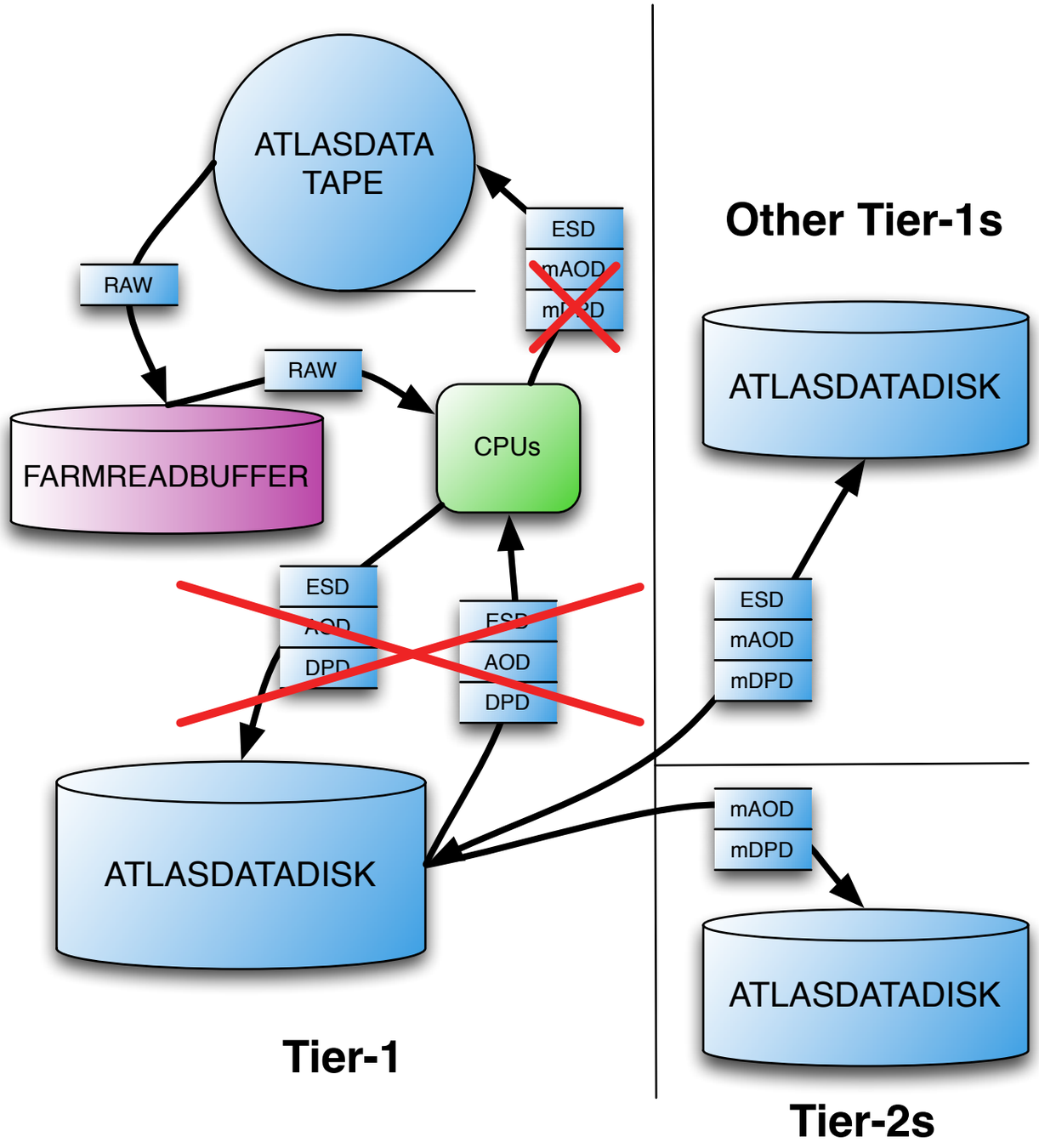
- Here mAOD/mDPD means merged files



Reprocessing in STEP09

- We considered re-running the whole Spring Reprocessing campaign
  - But this is too complex an exercise, involving over 10k tasks, different processing and merging steps
  - It's difficult to monitor and develop metrics for
- Instead we have a simpler 'pseudo-reprocessing'
  - This runs a RAW→ESD job
  - We will read (pre-stage) the RAW from tape and write the ESD back to tape
  - *This will write a realistic number of files back to tape, but each file will be smaller than expected when beam data is reprocessed*
  - Ergo: the metric for success is in files/hour, not GB/hour

STEP09 Reprocessing



- Do T1→T1/2 data export using 'Functional test' infrastructure

Reprocessing Targets

Tier-1	RAW to 'Reprocess' (TB)	Raw Files	ESD Files	Tape Read Rate (MB/s)	Tape Write Rate (MB/s)	Tape Write Rate (Files/hr)	Tape Volume (TB)
ASGC	56	35,000	35,000	46.3	4.1	104.2	4.9
BNL	280	175,000	175,000	231.5	20.3	520.8	24.5
CNAF	56	35,000	35,000	46.3	4.1	104.2	4.9
FZK	112	70,000	70,000	92.6	8.1	208.3	9.8
LYON	168	105,000	105,000	138.9	12.2	312.5	14.7
NDGF	56	35,000	35,000	46.3	4.1	104.2	4.9
PIC	56	35,000	35,000	46.3	4.1	104.2	4.9
RAL	112	70,000	70,000	92.6	8.1	208.3	9.8
SARA	168	105,000	105,000	138.9	12.2	312.5	14.7
TRIUMF	56	35,000	35,000	46.3	4.1	104.2	4.9
Total	1,120	700,000	700,000	925.9	81.0	2,083.3	98

- The reprocessing rate we aim for is x5 the nominal data taking rate
- ESD produced from cosmics is much smaller than for beam data (average 140MB files c.f. 1.0GB)

Data Re-distribution

- Once Tier-1s have reprocessed data, they need to distribute it across the grid
- This data flows as the original data from the T0 (but no RAW)
  - ESD to 1 partner T1
  - AOD and DPD to all T1s, with further distribution to T2s
- Because of the special nature of the reprocessing test, we decided to do this through the functional test framework

T1-T1 Rates

Tier-1	14 Days RAW (TB)	My ESD (TB)	My AOD +DPD (TB)	Export Volume (TB)	Import Volume (TB)	ATLASDATADISK (TB)	Export Rate (MB/s)	Import Rate (MB/s)
ASGC	11.2	7	2.8	32.2	60.2	70	26.6	49.8
BNL	56	35	14	161	77	126	133.1	63.7
CNAF	11.2	7	2.8	32.2	60.2	70	26.6	49.8
FZK	22.4	14	5.6	64.4	64.4	84	53.2	53.2
LYON	33.6	21	8.4	96.6	68.6	98	79.9	56.7
NDGF	11.2	7	2.8	32.2	60.2	70	26.6	49.8
PIC	11.2	7	2.8	32.2	60.2	70	26.6	49.8
RAL	22.4	14	5.6	64.4	64.4	84	53.2	53.2
SARA	33.6	21	8.4	96.6	68.6	98	79.9	56.7
TRIUMF	11.2	7	2.8	32.2	60.2	70	26.6	49.8
Total	224	140	56	644	644	840	532.4	532.4

T1-T2 Rates

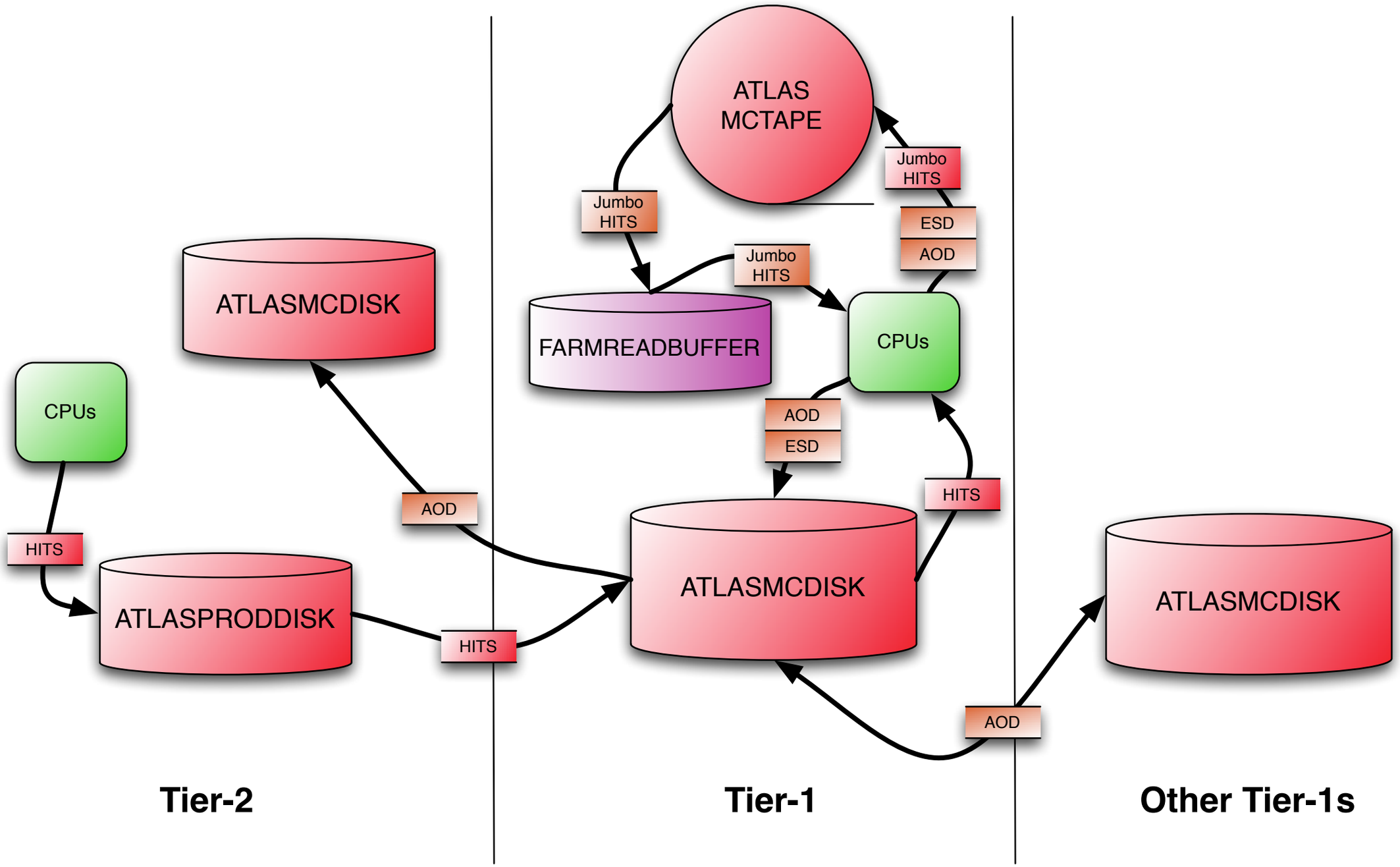
Tier-1	AOD+DPD Copies to T2s	Total Exported Data TB	Data Export Rate (MB/s)
ASGC	0.5	56	46.3
BNL	3	336	277.8
CNAF	1	112	92.6
FZK	2	224	185.2
LYON	2	224	185.2
NDGF	0	0	0.0
PIC	1	112	92.6
RAL	2	224	185.2
SARA	0.5	56	46.3
TRIUMF	1	112	92.6
Total	13	1456	1,203.7

- Computing model foresaw 1 copy of AOD+DPD in a cloud
- However T2 clouds vary hugely in size and many clouds export more than one copy
  - Conversely some clouds are seriously short of T2 capacity

Monte Carlo Production

- We should be able to run real monte-carlo jobs during step (part of the mc09 campaign)
  - G4 simulation done in Tier-2s, producing small HITS files
  - Small HITS uploaded to T1
  - Small HITS merged into jumbo HITS (written to MCTAPE)
  - Then reconstructed to AOD (and some ESD)
- Target rates will be *lower* than many clouds run in peak production because of other activities

G4 Simulation Workflow



STEP09 G4 Simulation Volumes

Tier-1	T2 Cores	50% Prod Share	Event Rate (1000s/G4 event)	STEP09 Events	HITS Volume TB	ESD Volume TB	AOD Volume TB	ATLASMCDISK (TB)	ATLASMCTAPE (TB)
ASGC	500	250	0.25	302,400	0.60	0.06	0.08	4.14	0.60
BNL	5,500	2,750	2.75	3,326,400	6.65	0.67	0.83	10.80	6.65
CNAF	1,200	600	0.6	725,760	1.45	0.15	0.18	5.07	1.45
FZK	4,000	2,000	2	2,419,200	4.84	0.48	0.60	8.80	4.84
LYON	4,000	2,000	2	2,419,200	4.84	0.48	0.60	8.80	4.84
NDGF	2,000	1,000	1	1,209,600	2.42	0.24	0.30	6.14	2.42
PIC	500	250	0.25	302,400	0.60	0.06	0.08	4.14	0.60
RAL	4,000	2,000	2	2,419,200	4.84	0.48	0.60	8.80	4.84
SARA	500	250	0.25	302,400	0.60	0.06	0.08	4.14	0.60
TRIUMF	800	400	0.4	483,840	0.97	0.10	0.12	4.54	0.97
Total	23,000	11,500	11.5	13,910,400	27.82	2.78	3.48	65.38	27.82

- ATLAS G4 simulation takes 1000s/event on a modern core
- T2 cores are estimated from current ATLAS production
- T1s must also reconstruct their HITS volume - this will be (re)done from mc08 data

Analysis Challenge

- Analysis in ATLAS mainly happens in Tier-2s
- 50% of Tier-2 CPU should be allocated to user analysis activities
  - We currently see 30% activity in at least some T2s
- In addition, some Tier-1s have attached analysis facilities
  - During STEP09 we have to check that this analysis activity does not disrupt scheduled Tier-1 activity

Test Framework

- ATLAS has developed 'hammercloud' test framework which has been run at all Tier-2s since last year
- This has greatly helped Tier-2s prepare for analysis
- During STEP09 we will ramp this activity upwards:
  - 4 different AOD analyses
  - Constant flow of jobs (load generator)
  - Use both the WMS and PanDA backends in EGEE
    - So we ask for /atlas/Role=pilot to be supported to test this

Tier-2 ATLAS Shares

Activity	VOMS Role	ATLAS Batch System Share
Production	/atlas/Role=production	50%
Analysis (pilot based)	/atlas/Role=pilot	25%
Analysis (WMS submission)	/atlas	25%

- We would like T2s who support this to provide feedback on how well balance between activities works during STEP09
  - e.g., jobs run/queued, CPU efficiencies each day

T1 STEP09 Requirements

- Put it all together and...

STEP09 Storage Totals							
Tier-1	**ATLASMCDISK (TB)**	**ATLASMCTAPE (TB)**	**ATLASDATADISK (TB)**	**ATLASDATATAPE (TB)**	**Network In (MB/s)**	**Network Out (MB/s)**	**Total Network (MB/s)**
ASGC	4.1	0.6	140	16	117	73	190
BNL	10.8	6.7	252	81	214	411	625
CNAF	5.1	1.5	140	16	117	119	236
FZK	8.8	4.8	168	32	141	238	380
LYON	8.8	4.8	196	48	166	265	431
NDGF	6.1	2.4	140	16	117	27	144
PIC	4.1	0.6	140	16	117	119	236
RAL	8.8	4.8	168	32	141	238	380
SARA	4.1	0.6	196	48	166	126	292
TRIUMF	4.5	1.0	140	16	117	119	236
Total	65	28	1,680	322	1,412	1,736	3,148

Aside: Summary of ATLAS use of Tape

1. Write RAW data from CERN to DATATAPE
2. Write all products from T1 reprocessing to DATATAPE
3. Write merged HITS to MCTAPE
4. Write reconstructed outputs (AOD, DPD, RDO) to MCTAPE

- N.B. MCTAPE and DATATAPE are separate space tokens and the stage-out buffer should also be separate

T1 STEP09 Tape Requirements

Tier-1	Tape Reading Rate (MB/s)	Tape Writing Rate (MB/s)	Files/Hour	If the files were 2GB then write rate (MB/s) would be...
ASGC	47	14	104	58
BNL	237	72	521	289
CNAF	47	15	104	58
FZK	97	31	208	116
LYON	143	44	313	174
NDGF	48	15	104	58
PIC	47	14	104	58
RAL	97	31	208	116
SARA	139	40	313	174
TRIUMF	47	14	104	58
Total	949	289	2,083	1,157

- This is lower than the nominal rate we might expect if all activities were happening at once
  - However, we would not do a reconstruction and reprocessing campaign at the same time
- Will help establish current tape system limits
- N.B. Writing to tape must take priority over reading

T1 Space Token Summary

Space Token	Size	Purpose
ATLASDATADISK	40%	
ATLASMCDISK	35%	
ATLASGROUPDISK	10%	Group Analysis Outputs
ATLASSCRATCHDISK	10%	Transfer Buffer
Buffer Space (*TAPE + Read)	10%	

- T1s should deploy ~50% of their 2009 pledge now
  - Retain flexibility for the run...
- Extra disk can be deployed into buffer space

T2 Space Token Summary

Space Token	Size	Purpose
ATLASDATADISK	30%	
ATLASMCDISK	25%	
ATLASGROUPDISK	20%	Group Analysis Outputs
ATLASSCRATCHDISK	20%	Default User Analysis Output Buffer
ATLASPRODDISK	5%	Production buffer

- You may also have non-pedged resources in ATLASLOCALGROUPDISK

Open Issues

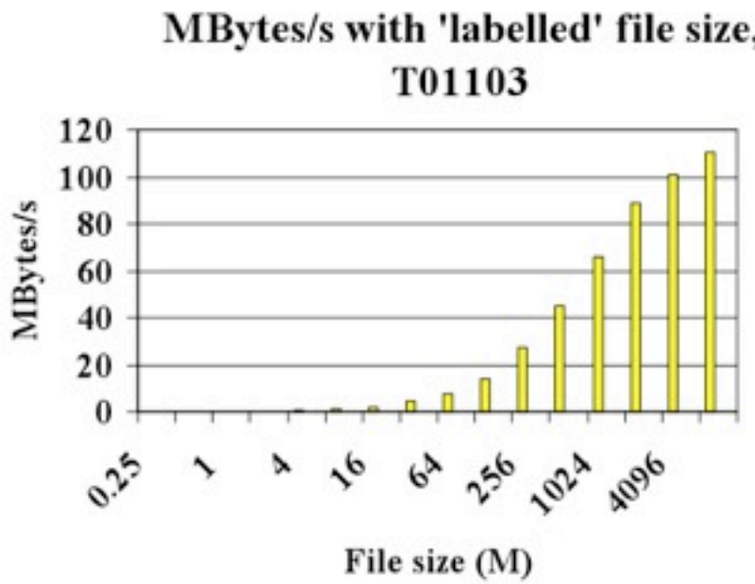
- Will ASGC be ready?
- Reduced capacity at IN2P3, plus scheduled tape backend downtime
- Metrics
  - Watch this space...

Backup Slides

Tape Speed/File Size

File Size (MB)	Write Speed (MB/s)	Files/Hour
256	17	239
1024	45	158
2048	65	114
4096	90	79

Compressed data transfer rate, write



Use compression, 1K=1024, 1M=1048576

dd if=/dev/zero ibs=80 of=/dev/nst0 obs=80 count=2

dd if=/dev/zero ibs=256k of=/dev/nst0 obs=256k count=64 (16 MB file)

dd if=/dev/zero ibs=80 of=/dev/nst0 obs=80 count=2