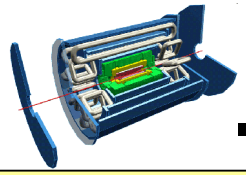
Comments on the ATLAS Computing Model

GridKa T1-T2 Workshop, 19/20.10.2005

Günter Duckeck, LMU München

- Computing Model Intro
- Resources for Tier-1 & Tier-2
- Networking
- Fine-grained storage split-up
- Basic storage I/O requirements

ATLAS Offline Computing



~Pb/sec

Event Builder

10 GB/sec

Event Filter
~7.5MSI2k

Some data for calibration and monitoring to institutes

Calibrations flow back

400 MB/sec

T0 ~5MSI2k

Tier 0

- ~5 Pb/year
- No simulation

Tier 1

US Regional Centre

Dutch Regional Centre

French Regional Centre

German Regional Centre (GridKa)

- ~2MSI2k/T1
- ~2 Pb/year/T1

Tier 2

LRZ/RZG Tier2
~500kSI2k

Desy Tier2
~500kSI2k

Fr/Wu Tier2
~500kSI2k

- In total ~600 TB/year

≥622Mb/s links

LMU Munich

MPI Munich

Uni D ...

Uni M...

Tier 3

PC (2004) =
~1 kSpecInt2k

Physics data cache

Desktop

Baseline assumptions

- Assumptions:
  - 200 days running in 2008 and 2009 at 50% efficiency (107 sec live)
  - 25 days running in 2007 (2.5x106 sec live)
  - Events recorded are rate limited in all cases – luminosity only affects data size and data processing time
  - Luminosity:
    - $0.5*10^{33}$ cm-2s-1 in 2007
    - $2*10^{33}$ cm-2s-1 in 2008 and 2009
    - 10^{34} cm-2s-1 (design luminosity) from 2010 onwards
- Hierarchy
  - Tier-0 has raw+calibration data+first-pass ESD
  - CERN Analysis Facility has AOD, ESD and RAW samples
  - Tier-1s hold RAW data and derived samples and ‘shadow’ the ESD for another Tier-1
  - Tier-1s also house simulated data
  - Tier-1s provide reprocessing for their RAW and scheduled access to full ESD samples
  - Tier-2s provide access to AOD and group Derived Physics Datasets and do the simulation

Data Volumes and Data Types

- **RAW data** for primary reco at Tier-0 (and Tier-1 for reproc)
  - 1.6 MB/event, $2*10^9$ ev/year, 3.2 PB/year
  - 1 copy at Tier-0, 1 copy distributed over ~10 Tier-1 (on tape)
- **ESD** (event summary data, reco objects + raw data subset), for physics-group analysis at Tier-1
  - 0.5 MB/event, 1 PB/year
  - 2 copies distributed over ~10 Tier-1 on disk
- **AOD** (analysis object data, reconstructed physics objects: jets, leptons, etc) for user analysis at Tier-2
  - 0.1 MB/event, 180 TB/year
  - 1 copy at each Tier-1 and 1 copy shared among ~3 Tier-2 centers
- **TAG** (TAG data, basic event-level info) for fast skimming
  - 1 kB/event, 2 TB/year, each T-1/T-2 center
- Same structure for **simulated** data, size **~20%** of real data

ESD/AOD Streaming Model

- ESD:
  - Single stream
  - 2 copies distributed over 10 Tier-1 sites ⇒ 20% ESD per Tier-1
- AOD (*still under discussion ...*):
  - Baseline proposal is for non-overlapping streams (about 10)
    ⇒ each event appears only once
  - Primary motivation is to reduce number of files containing events of interest
  - Physics channels can appear in multiple streams depending on mapping to trigger decisions
  - primary physicist interface to set of events of interest is the *event collection* not the *stream*
  - 3 Tier-2 share full set of AOD

Tier-1/Tier-2 Tasks

- Tier-1: 10 centers in major countries worldwide
  - Physics- & calibration groups ``organized'' analysis of ESD data
  - Calibration-group ``organized'' analysis of ESD data (and Raw data)
  - ATLAS wide re-processing of RAW data 1-2 x / year
  - Main repository for ESD, AOD (real and simulated)
  - No User-level analysis!
- Tier-2: ~30 centers distributed worldwide
  - User-level ``chaotic'' analysis of AOD data
  - Organized Simulation production by ATLAS & Physics groups
  - Analysis of group and user data
  - Repository for AOD, group data sets and some user data

Average Tier-1 requirements (2008)

	Disk (TB)	Tape (TB)
Raw	43	304
ESD (current)	257	90
ESD (previous)	129	90
AOD	283	36
TAG	3	0
Calibration	240	0
MC RAW	0	80
MC ESD (current)	57	20
MC ESD (previous)	29	20
AOD Simulation	63	8
Tag Simulation	1	0
Group User Data	126	0
Total	**1231**	**648**

	Tier-1
	CPU (kSI2k)
Reconstruction	450
Calibration	50
Analysis	1300
Simulation	0
Total	1800

Average Tier-2 requirements (2008)

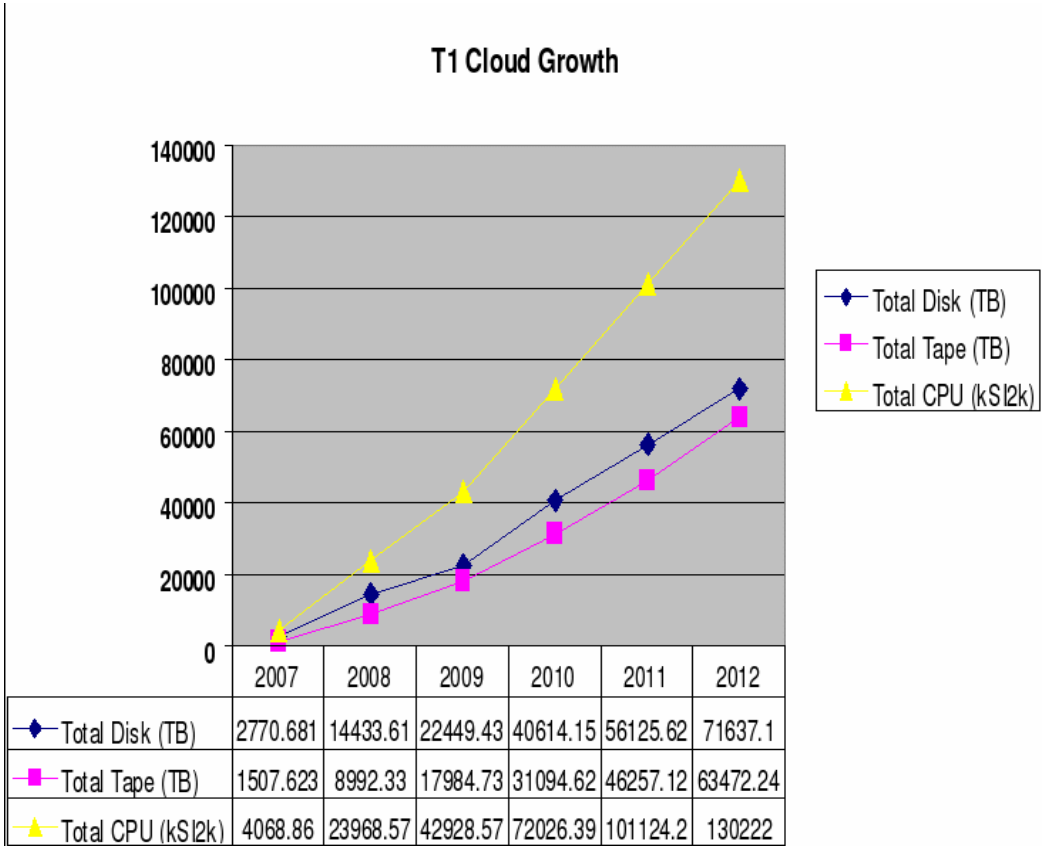
	total (TB)
Raw	1.4
General ESD	12.9
AOD	85.7
TAG	2.6
	0
RAW Sim	0
ESD Sim	5.7
AOD Sim	19
Tag Sim	0.6
User Group	41.9
User Data	60.5
Total	230.3

	CPU (kSI2k)
Reconstruction	65
Simulation	180
Analysis	290
Total	**540**

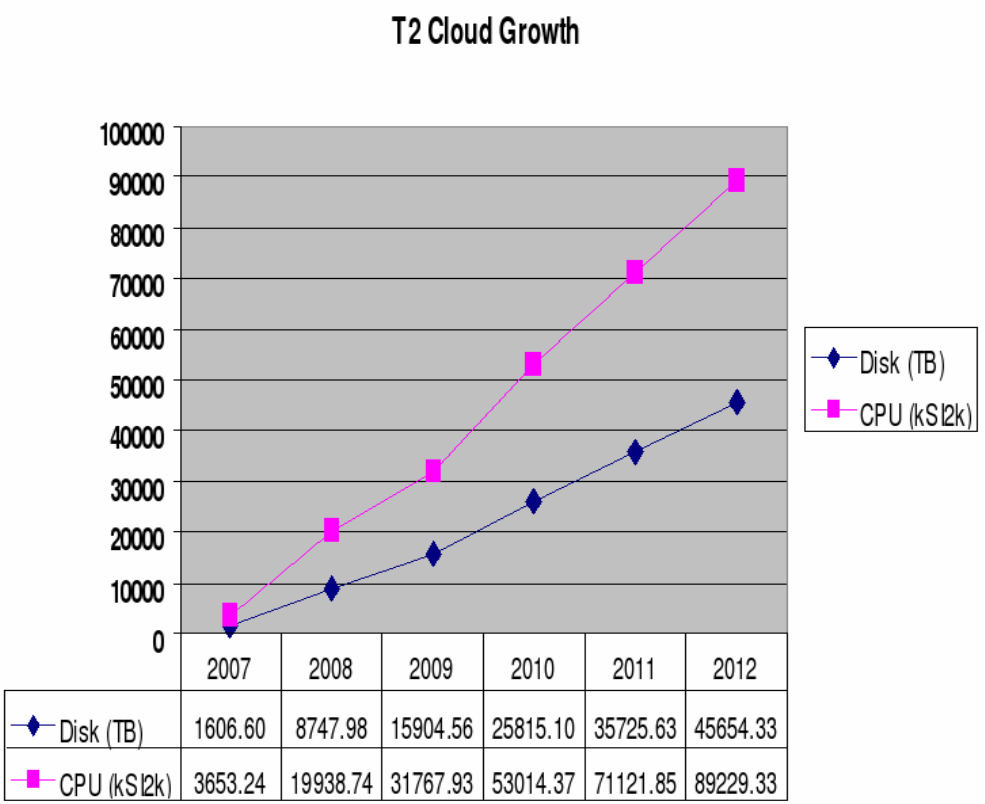
- Proportional scaling not required. Can have T-2 focused more on
  - Simulation = CPU
  - AOD analysis=disk
  - User Analysis=both

Tier-1/2 growth

- Resource needs basically proportional to accumulated data
- slight kink in 2009 due to projected high-lumi running:
  Trigger rate const but larger events



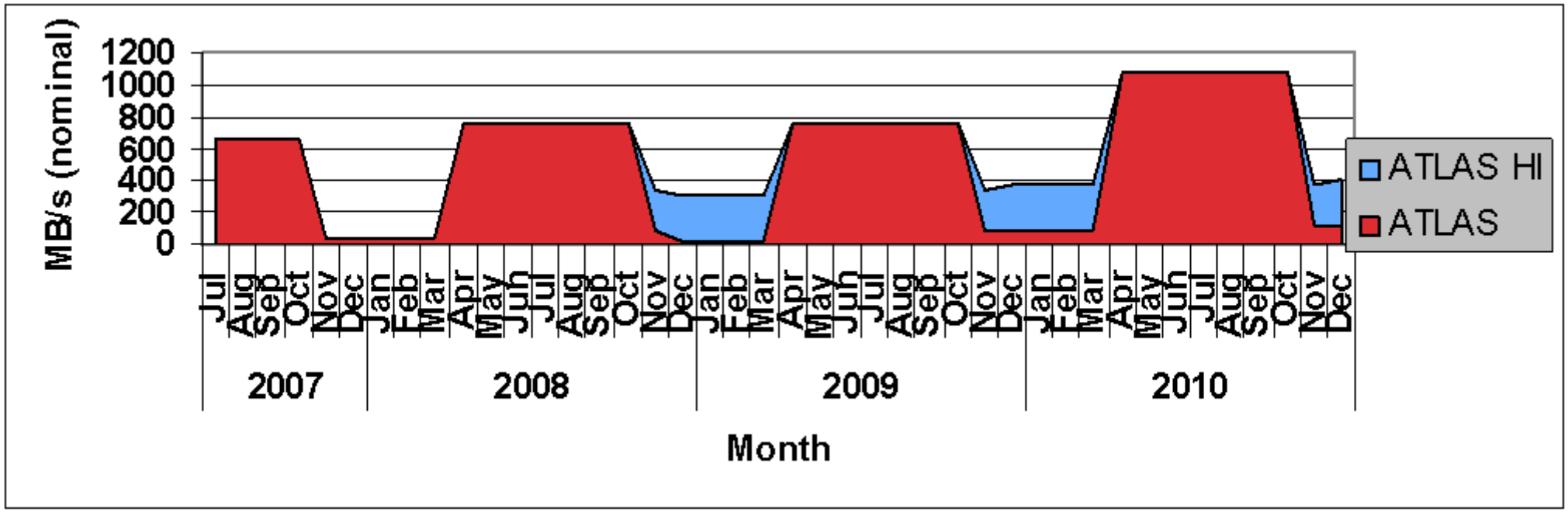
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	2770.681	14433.61	22449.43	40614.15	56125.62	71637.1
Total Tape (TB)	1507.623	8992.33	17984.73	31094.62	46257.12	63472.24
Total CPU (kSI2k)	4068.86	23968.57	42928.57	72026.39	101124.2	130222



	2007	2008	2009	2010	2011	2012
Disk (TB)	1606.60	8747.98	15904.56	25815.10	35725.63	45654.33
CPU (kSI2k)	3653.24	19938.74	31767.93	53014.37	71121.85	89229.33

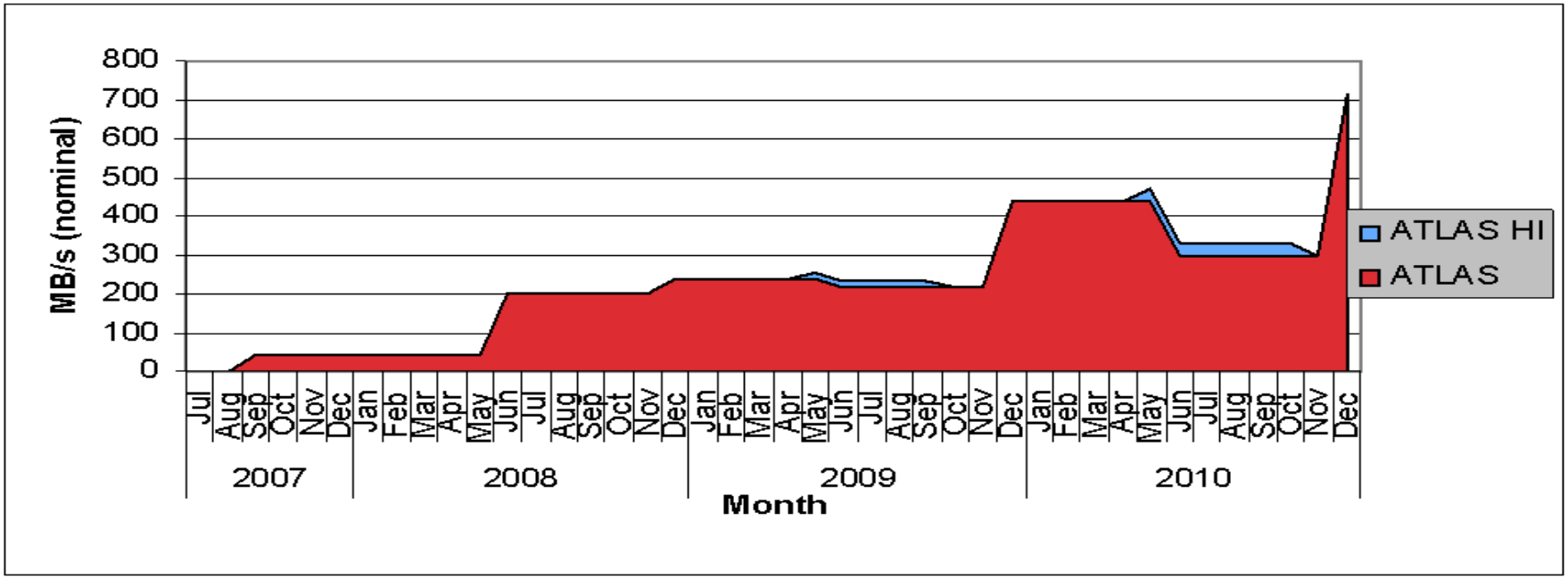
Networking Tier-0 --> Tier-1

- Traffic from T0 to average Tier-1 is ~75MB/s raw in 2008
- With LCG headroom, efficiency and recovery factors for service challenges this is 3.5Gb/sec



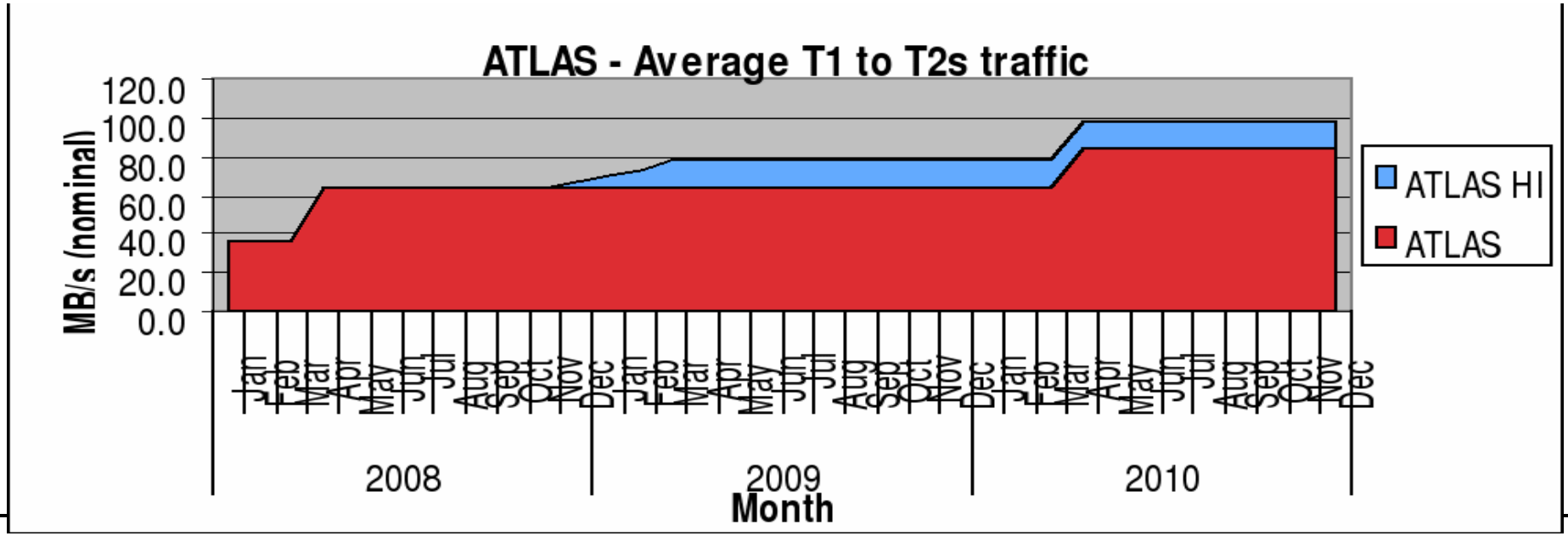
Networking Tier-1 --> Tier-1

- Significant traffic of ESD and AOD from reprocessing between T1s
  - 250MB/sec raw for average T1
  - Rising to ~2Gb/sec for average T1 after usual factors



Networking and Tier—2s

- Tier-2 to Tier-1 networking requirements presumably low
  - 2-3 x / year AOD T-1 --> T-2
  - Continuous phys. group sets T-1 --> T-2
  - Continuous simul data T-2 --> T-1 (3.6 MB/s)
  - without job traffic ~17.5MB/s for ‘average’ T2
  - ~850Mb/sec after eff factors



Comment on ATLAS storage needs

- In ATLAS CM only two categories:
  - disk = low latency access
  - tape = archive/backup; access <= 2x/year
- More fine-grained levels in practice:
  - 'online-access' needed for data with event-navigation (seek in file)
    - Current AOD/ESD
    - Rest of “disk” can go to HSM (=disk-cache+tape)
- Tape-needs for archive/backup explicitly specified in CM
  - for large fraction of data multiple copies easily available worldwide
    - no need for extra backup copy, restore via network

Example split at Tier-1

	total size (TB)	online disk (TB)	HSM (disk/tape) (TB)	archive Tape (TB)
Raw	304	43	0	304
ESD (current)	257	257	0	90
ESD (previous)	129	0	129	90
AOD	283	56	227	36
TAG	3	3	0	0
Calibration	240	0	240	0
MC RAW	80	0	0	80
MC ESD (current)	57	57	0	20
MC ESD (previous)	29	0	29	20
AOD Simulation	63	13	50	8
Tag Simulation	1	1	0	0
Group User Data	126	0	126	90
Total	**1572**	**430**	**801**	**738**

- Only ~1/3 of “Tier-1 disk” actually “online” disk for event-seek
- 2/3 should be ok for HSM

Example split at Tier-2

	total (TB)	online disk (TB)	HSM (disk/tape) (TB)	archive Tape (TB)
Raw	1	0	1	
General ESD (curr	13	0	13	
AOD	86	86	0	
TAG	3	3	0	
ESD Sim (curr.)	6	0	6	
AOD Sim	19	19	0	
Tag Sim	1	1	0	
User Group	42	0	42	42
User Data	61	0	61	61
Total	**230**	**108**	**122**	**102**

- ~1/2 of “Tier-2 disk” is“online disk” for event-seek
- Tape archive for (precious) user data might be useful

Data access profile (GD)

- No details in ATLAS CM so far
- Take few simple scenarios to get baseline numbers:
  - Tier-1 RAWD reprocessing, using full available CPU capacity:
    - I/O = CPU-cap / proc-time * event-size = 1800 / 15 * 1.6 = 190 MB/s (from tape)
  - Tier-1 ESD Analysis, using full available CPU capacity:
    - I/O = CPU-cap / proc-time * event-size = 1800 / 0.5 * 0.5 = 1800 MB/s (from disk)
    - when distributed equally over 200 TB: 9 MB/s per TB Filesys
  - Tier-2 AOD Analysis, using full available CPU capacity:
    - I/O = CPU-cap / proc-time * event-size = 535/ 0.5 * 0.1 = 110 MB/s (from disk)
    - when distributed equally over 70 TB: 1.6 MB/s per TB Filesys
  - Tier-2 AOD “Skimming” (1/40 sec / AOD), using full available CPU capacity:
    - I/O = CPU-cap / proc-time * event-size = 535/ 0.025 * 0.1 = 2100 MB/s (from disk)
    - when distributed equally over 70 TB: 110 MB/s per TB Filesys

Comments on uncertainties (GD)

- Assumed processing times and data sizes:

Item	Unit	Spec Value	Current
Raw Data Size	MB	1.6	?
ESD Size	MB	0.5	1.2
AOD Size	kB	100	50
TAG Size	kB	1	
Sim. Data Size	MB	2,0	
Sim. ESD Size	MB	0,5	1.2
Time/Reco 1ev	kSI2k-sec	15	22
Time/Simu 1ev	kSI2k-sec	100	250-800
Time/Analyse 1ev	kSI2k-sec	0.5	0.1-0.2

Comments on uncertainties - 2 (GD)

- User analysis really on AOD only?
  - presumably not in the beginning ...
- Exclusive streaming or overlapping physics streams ?
- Simulation set only 20% of real data ?(CMS=100%)
- two reconstruction passes only ?
  - presumably not in the beginning ...
- Job-to-Data (now) or Data-to-Job (Networking...)