Experiment Data Flows

Roger W L Jones

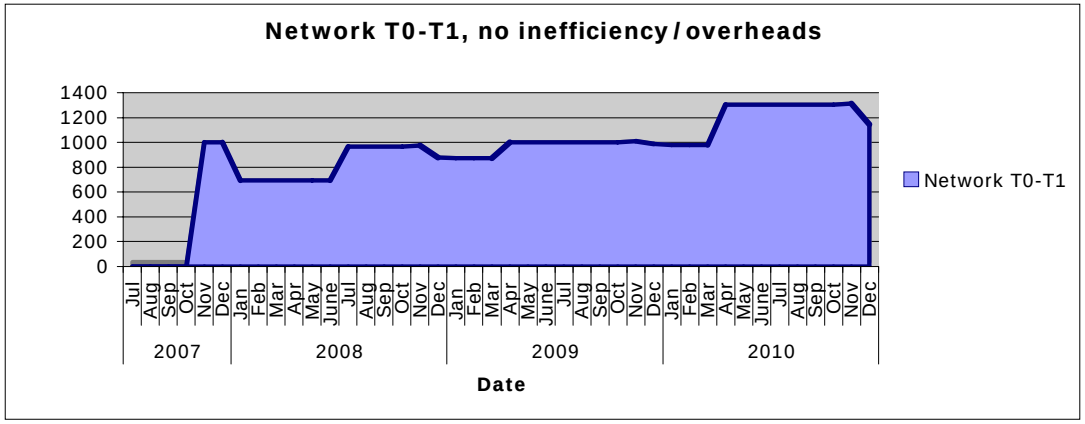
FTS Workshop, SARA 18 Oct 06

Outwards From CERN

- The raw data arrives into the input buffer @ CERN in ~10 streams
- It is then shipped to the 10 Tier 1s as soon as possible
  - The data is shipped on a ‘round robin’ basis (no preferred sites)
  - The transfer is to MSS
- The data is also stored in Castor
  - Should be pinned to disk for ~48 hours for calibration and processing
  - Is the disk the buffer or Castor?
  - Should be visible from the CERN Analysis Facility

Outwards from CERN (2)

- **The data are processed in 48 hours (< 5 days)**
  - **ESD, AOD and (file based) TAG shipped to Tier 1s (t1d1)**
  - **Should go to Tier 1 that holds the corresponding RAW and its partner**
  - **If there is a problem, buffer or ship from partner Tier 1**
- **TAG also merged with relational database**
- **Outward flow of conditions database updates**



Does not yet include ~100MB/s overhead for BNL full ESD copy

Inwards to CERN

- Small but vital calibration data stream from T1s and T2s
- Small fraction of the simulated data (with hits) from the custodial Tier 1 to the CAF (t0d1)
- A larger fraction of the derived physics datasets (DPD) and AOD (t0d1)

Into the Tier 1

- **From CERN:**
  - RAW data (t1d0, but small pre-assigned fraction t1d1)
  - The corresponding ESD, AOD, (file-based) TAG (all t1d1)
    - TAG merged with relational database
  - The ESD, AOD & TAG from partner Tier 1 (t0d1)
  - Conditions data
- **From other T1s**
  - ESD from reprocessing at partner T1 (t0d1)
  - DPD, AOD & TAG from reprocessing at all other T1s (t0d1)
- **From T2s**
  - Hits/RAW, ESD, AOD & TAG from simulation at associated T2s (t1d1)
  - Small amounts of DPD from analysis at T2s

Outward Tier 1- Tier 1

- New ESD, AOD and (file based) TAG produced from the ‘custodial’ local raw data
- Substantial quantities of DPD sets produced every few weeks
- Rare transfers to restore lost files

ATLAS "average" T1 Internal Data Flow (2008)

RAW
ESD (2x)
AODm (10x)

1 Hz
85K f/day
720 MB/s

Tape

RAW
1.6 GB/file
0.02 Hz
1.7K f/day
32 MB/s
2.7 TB/day

RAW
ESD2
AODm2
0.044 Hz
3.74K f/day
44 MB/s
3.66 TB/day

Tier-0

disk buffer

AODm1	AODm2
500 MB/file	500 MB/file
0.04 Hz	0.04 Hz
3.4K f/day	3.4K f/day
20 MB/s	20 MB/s
1.6 TB/day	1.6 TB/day

ESD1	AODm1
0.5 GB/file	500 MB/file
0.02 Hz	0.04 Hz
3.4K f/day	3.4K f/day
20 MB/s	20 MB/s
1.6 TB/day	1.6 TB/day

RAW	AOD2
1.6 GB/file	10 MB/file
0.02 Hz	0.2 Hz
1.7K f/day	17K f/day
32 MB/s	2 MB/s
2.7 TB/day	0.16 TB/day

ESD2	AOD2	AODm2
0.5 GB/file	10 MB/file	500 MB/file
0.02 Hz	0.2 Hz	0.004 Hz
3.4K f/day	17K f/day	0.34K f/day
20 MB/s	2 MB/s	2 MB/s
1.6 TB/day	0.16 TB/day	0.16 TB/day

Each Tier-2

CPU farm

Plus simulation and analysis data flow

ESD2	AODm2
0.5 GB/file	500 MB/file
0.02 Hz	0.036 Hz
1.7K f/day	3.1K f/day
10 MB/s	18 MB/s
0.8 TB/day	1.44 TB/day

ESD2	AODm2
0.5 GB/file	500 MB/file
0.02 Hz	0.004 Hz
3.4K f/day	0.34K f/day
20 MB/s	2 MB/s
1.6 TB/day	0.16 TB/day

ESD2	AODm2
0.5 GB/file	500 MB/file
0.02 Hz	0.036 Hz
3.4K f/day	3.1K f/day
20 MB/s	18 MB/s
0.8 TB/day	1.44 TB/day

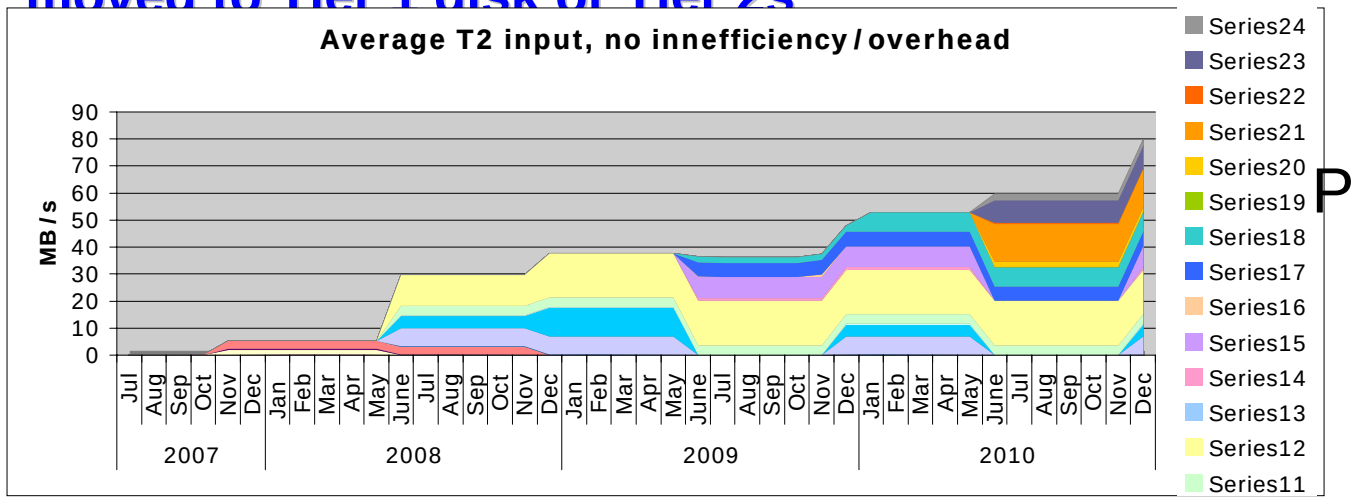
Other Tier-1s

disk storage

Other Tier-1s

Tier 1 to Tier 2

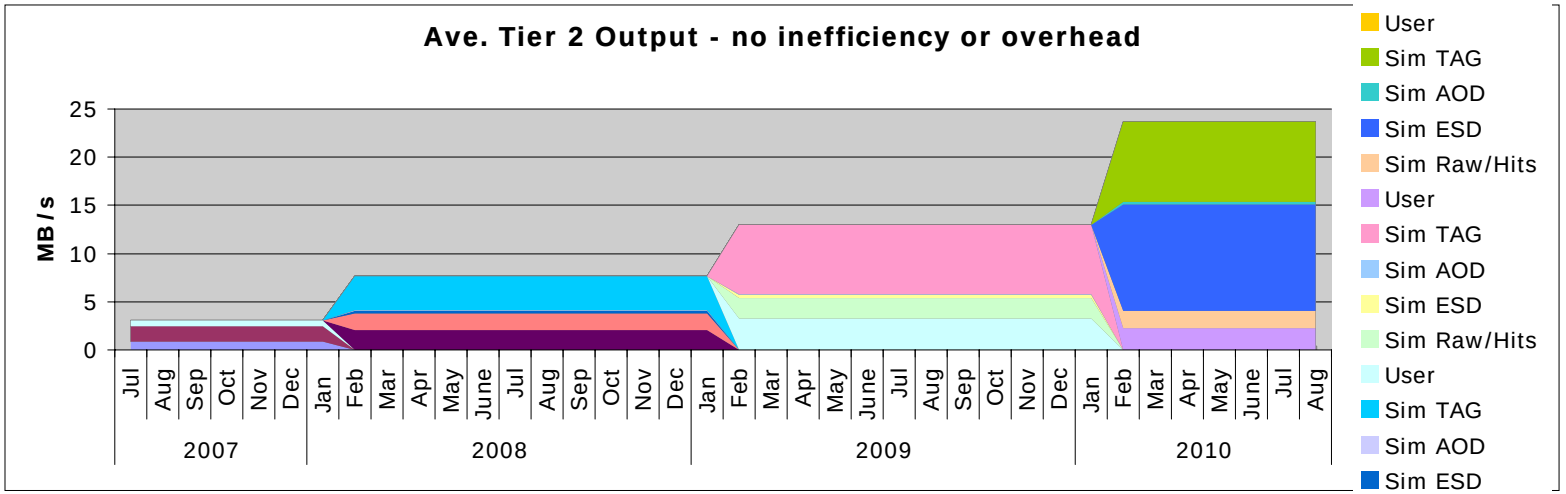
- Tier 2s are of many different sizes
- Transfer is normally to associated cloud of Tier 2s
- In some cases, the request from another Tier 1 will go via the 'local' Tier 1
- Small traffic of pre-subscribed set of raw data and ESD
- Similar traffic of later RAW/ESD selections being moved to Tier 1 disk or Tier 2s



Prelim!

Tier 2 to Tier 1

- **Mainly simulated data**

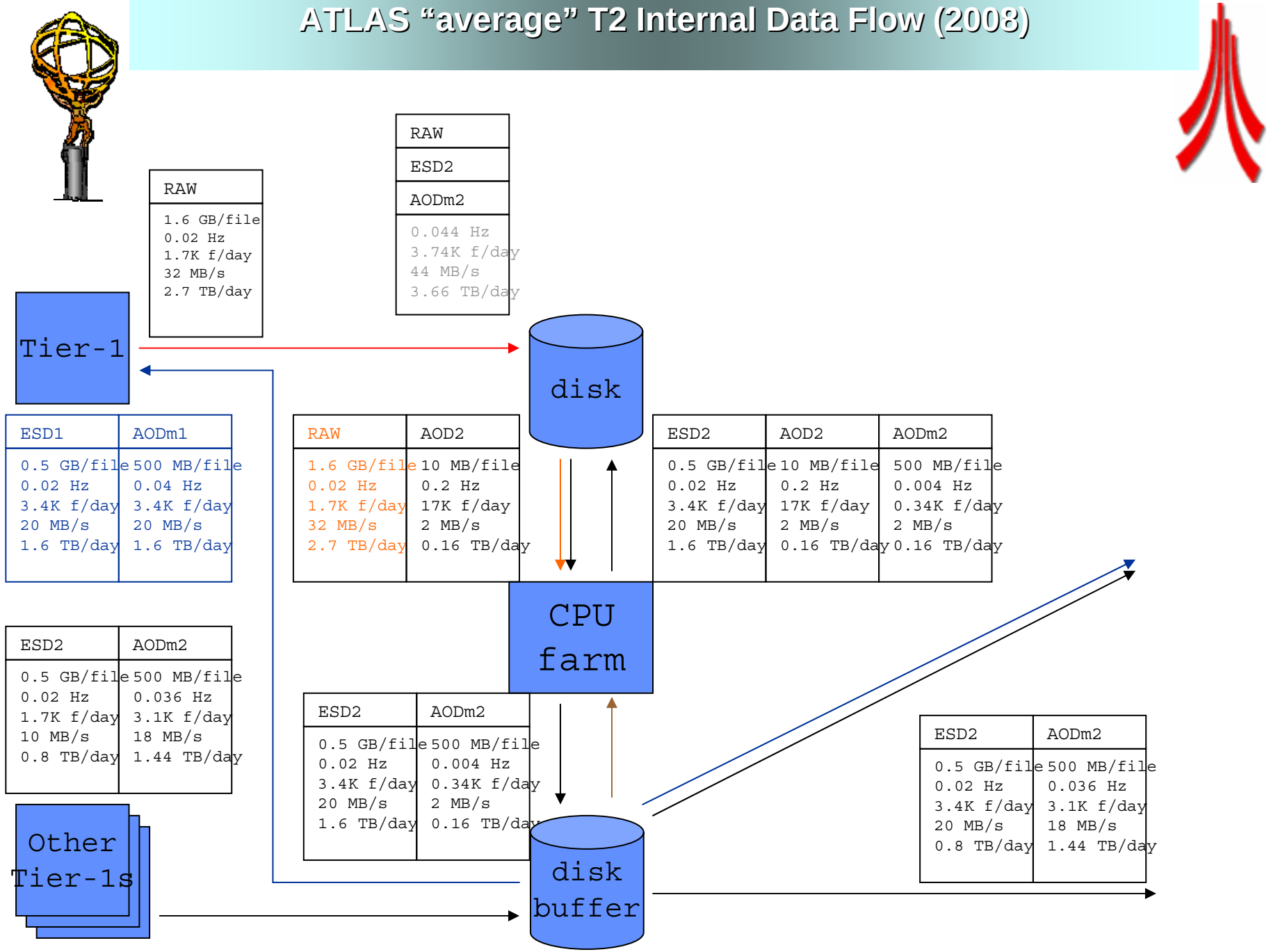


Prelim!

Data Placement

- **Allocation of data to Tier 2s 'democratic'**
- **Placement of data in T2s done centrally**
  - **All subscriptions done centrally**
  - **T2 can express preferences**
- **Data will be clumped at selected sites**
- **Some data (e.g. small RAW & ESD samples) requested for T2s**
  - **Generally deep copies made as part of group analysis passes at Tier 1s (scheduled)**

ATLAS “average” T2 Internal Data Flow (2008)



How do other experiments differ? CMS

- **CMS are very similar in T0-T1 movements**
- **CMS do not have ‘partnered’ Tier 1s**
- **Biggest difference in the data movement to Tier 2s**
  - **Data is called on demand**
  - **Disk is all cache, file lifetime 30 days**
  - **This reduces the disk size, but has bigger bandwidth requirements**

CMS Dataflows

- **Basic summary:**
  - **T0 -> T1: OPN, moderate traffic (per T1), reliability vital**
  - **T1 <-> T1: OPN, large traffic, reliability important**
  - **T1 -> T2: off-OPN, very large traffic (per T1), many channels**
    - **Possibly the most challenging case for file transfer**
  - **(T2 -> T1): small**
- **T0 -> T1:**
  - **'Nominal T1' receives ~50MB/s (raw throughput) from T0, 100 days per year**
  - **FEVT (RAW+RECO) from T0 reconstruction farm**
  - **'pseudo-online flow'; guaranteed reliability (all the way to T1 tape) is essential**
  - **Note that many T1 have already exceeded this rate during CSA06**

Thanks to D Newbold

CMS Dataflows

- **T1 <-> T1**
  - Replication of new data (e.g. new AOD) between T1 on OPN
  - Need to bound the time taken for distribution of new data (otherwise need two AOD copies on disk for extended period)
    - Currently assume 14 days for replication
  - Peak rate for nominal T1 ~150MB/s (raw throughput)
    - NB: Low duty cycle; what does this imply?
- **T1 -> T2**
  - Serving of AOD / RECO data to T2 centres
  - NB: Online copy of RECO data is held at one T1 only
    - Implies *many-to-many* T1 <-> T2 transfers
  - Challenging use case for file transfer tools, but is a *hard requirement*, fundamental to the computing model
  - Important to have the possibility of many <-> many model for general transfers (AOD) to ensure robustness

CMS Dataflows

- **T1 -> T2 (cont):**
  - N_{T1} = 7, N_{T2} = 25 (including some federated T2 sites)
  - Essential to form robust strategies for channel management, etc
- **T2 data rates**
  - Important to understand the use case: T2 traffic is bursty
    - Driven by ad hoc demand for data for analysis
    - Key metric is 'time to download a complete dataset', since this defines productivity
  - Average rates are useful for T1 planning, but not for T2 / FTS
  - E.g. average rate for nominal T1 ~120MB/s, nominal T2 ~35MB/s
  - Reality: T2 will want to transfer at wire speed in bursts (100MB/s+)
  - Reality: T1 should be able to sustain such a pattern (for small number of T2) against background traffic
  - Take account of this when specifying and testing file transfers

LHCb

- **LHCb again have similar movement patterns between T0 and T1**
  - **Their stripping corresponds to the ATLAS group based analysis**
  - **Their analysis is at Tier 1, and so 'on demand'**
- **Bandwidth requirement is a lot lower**
- **Tier 2 usage is very different**
  - **Simulation only**
    - **Only movement of simulated data to Tier 1**
    - **In two cases, simulated data goes to CERN**

LHCb - CERN-Tier 1 Processing

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

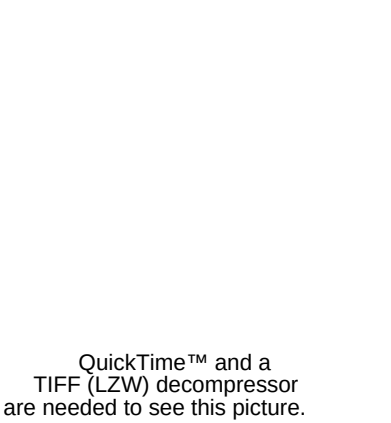
LHCb - CERN-Tier 1 Reprocessing

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

LHCb CERN-Tier 1
Stripping @ CERN

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

LHCb - Tier 1 data receiving



LHCb Tier 1 Stripping/data taking

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

LHCb Tier 1 Reprocessing

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

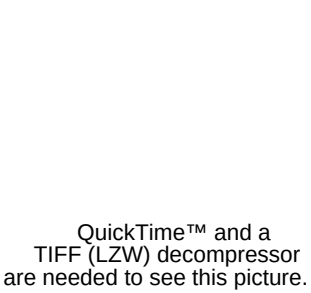
Two Months

LHCb Tier 1 re-Stripping

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

One month, twice a year

LHCb Tier 1 Analysis



ALICE

- **Sorry to ALICE, I did not have time to get detailed slides**
  - **These comments are based on earlier ‘megatables’ and do not have the new schedule**
- **Rather like the ATLAS flows, but with CERN having more Tier 1 aspects**
- **About 1/6 of T0-T1 traffic**
- **Little T1-T1 traffic**
- **Substantial T2→T1 traffic.bigger T1→T2 traffic (similar to ATLAS)**