ATLAS Note draft
03 August 2006 (V1.2)

ATLAS Distributed Data Management Operations

D.Barberis, J.Chudoba, S.Jezequel, J.Kennedy, A.Klimentov,
D.Liko, P.Nevski, A.Olszewski, L.Perini, G.Poulard

ATL-SOFT-PUB-2006-006
18 September 2006

ATLAS Distributed Data Management (DDM) service is developped for data transfer between ATLAS sites and for data cataloguing. The Data Management Software (SW) is based on DQ2 (practically all DDM core services are DQ2 based [1]) and end-users tools (aka dq2_get package developped by T.Maeno et al). In this paper we address the issues of DDM day-by-day operation, DDM operations team organization, roles and responsibilities of Tier-1s and Tier-2s DDM coordinators.

1 Introduction

The overall organization of the management of the ATLAS data is fully described in the ATLAS Computing Model [2]. It is worthwhile to recall briefly that in this model the data will be distributed to the ATLAS Computing Tiers sites which play a different role:

- Tier-0 at CERN is responsible for the archiving and distribution of the primary RAW data received from the event Filter. It provides the prompt reconstruction of the calibration and express streams and the first pass processing of the primary event stream. The derived datasets (ESD, primary AOD and TAG sets) are distributed from the Tier-0 to the Tier-1 facilities.
- Tier-1s take responsibility to host and provide long-term access and archiving of a subset of the RAW data. They also undertake to provide the capacity to perform the reprocessing of the RAW data under their curation, and to provide ATLAS-wide access to the derived ESD, AOD and TAG datasets. The Tier-1s also undertake to host a secondary low-latency copy of the ESD, AOD and TAG samples from another Tier-1 and the simulated data samples from Tier-2 facilities. The currently number of ATLAS Tier-1s is 10.

- Tier-2 facilities may take a ranges of significant roles such as providing calibration constants, simulation and analysis. They will provide all of the required simulation capacity. As mentioned previously simulated data will be hosted at the associated Tier-1. Typically few (3-4) Tier-2s will be associated with each Tier-1.

Clearly the movement of the data between these Tiers is of a great importance and the role of the Distributed Data Management is mostly to provide a service for data cataloguing and data transfer between ATLAS sites.
The second generation of ATLAS DDM SW (DQ2) is described in more details in [3], we just briefly remind the technical highlights. DQ2 has moved to dataset based approach. The dataset can be defined as an aggregation of files plus associated metadata. There is also the concept of datablocks, a frozen (permanently immutable) aggregation of files for the purposes of distributing. It is important to mention that DQ2 global services have no global physical file replica catalog, and include global dataset repository and global dataset location catalog. DQ2 local site services (per *GRID*/site/tier) provide logical-to-physical file name mapping. The implementations of this catalog are *GRID* specific. Currently all local catalogues are deployed per ATLAS site/storage element (SE). The key features of DQ2 are dataset subscription and notification. Any site can subscribe to dataset. The new version of dataset is automatically made available on site. All managed data movement in the system is automated using the subscription system. When content of dataset is modified, the sites subscribing to it are notified and data is moved accordingly.
DQ2 central services are installed on Tier-0. DQ2 Tier-1 site services are installed on a dedicated machine (so-called VO box) to run data management services and other services. The detailed requirements for the VO Box and the components ATLAS plans to use on it are described in [4].

2 DDM operations team structure and responsibilities

The DDM Operations team was set up within ATLAS Computing Operations project in Feb 2006. The team was actively involved in DDM/DQ2 deployment, testing and installation, DDM/prodSys integration and DDM central facilities maintenance.
Now with the stable DDM/DQ2 release, we are bringing the system into production, and we need a clear definition of DDM operations group organization and responsibilities. The issues related to Tier-1-Tier-2 association are described in detail in [5]. In this note we address the day-by-day DDM operations and Tier-1/Tier-2 DDM coordinators responsibilities.
In the following, we will not differentiate Tier-2 and Tier-3 and call them Tier-2 (for DDM ops Tier-2 and Tier-3 have the same type of computing activities but Tier-3 ones are controlled by the local administator whereas Tier-2s are managed centrally by ATLAS).
The DDM operations team must include qualified ATLAS collaborators, having basic knowledge of *python* language, *GRID* architecture, *GRID* data management, monitoring tools, and ATLAS DDM components.

The proposed ATLAS DDM operations team structure :

- DDM operation coordinator (DDMops)
- Central DDM operations group (part of the group members are based at CERN, but a good fraction of the group works (will work) in ATLAS Laboratories and Universities)

- Regional DDM operation coordinators (usually one per Tier-1 or one per country)
- Regional DDM operations group (it will include people from Tier-1 as well as from Tier-2s associated with the Tier-1)

DDMops is responsible for ATLAS data transfer[1)] between Tier-0 and Tier-1s. DDMops is also responsible for proper data cataloguing and management. Basically, DDMops team must keep data integrity and provide routine data transfer, data monitoring and data access to the ATLAS physics community. The data transfer priorities are defined by the collaboration management, data preparation and physics coordinators (see section 3). The data transfer between the pit and Tier-0 is not the responsibility of DDMops group. DDMops and central operations team are responsible for :

- day-by-day data transfer between Tier-0 and Tier-1s
  - data transfer monitoring and control
  - data rerouting in case of Tier-1/Tier-2 or transfer channels instability
  - resolving data transfer errors and providing first line expertise
  - reporting data transfer problems to the Computing Operations Coordinator
- deployment of DDM/DQ2 releases (it is important to mention that data processing and MC production have high priority and new DDM/DQ2 version deployment must be agreed on between ProdSys and Data processing coordinators)
- 24/7 support of central DDM operations facilities
  - production server at CERN
  - DQ2 client
  - central databases
- Support DDM Savannah users requests portal
- keep data integrity, in particular clean obsolete datasets and files entries from LFC/LRC and DDM catalogues
- Help ATLAS users to transfer data
- Communication with services developers and providers (both ATLAS and WLCG) and CERN Networking personnel (NetOps)

Regional DDM operations group includes Regional DDM operations coordinator and people working for Tier-2s associated with the particular Tier-1. The group structure and organization can be different from region to region, but each group must provide help to ATLAS physics community, manage ATLAS DDM SW and do DDM/DQ2 deployment. Regional DDMops and his team are responsible for :

- day-by-day data transfer between Tier-1 and Tier-2s
  - data transfer monitoring and control
  - resolving data transfer errors and providing first line expertise

[1)] Real RAW, AOD, ESD data distribution from the Tier-0 to Tier-1 and Tier-2s

- day-by-day data transfer between Tier-1 and Tier-0 (together with central DDM team)
- data exchange between Tier-1s during RAW data reprocessing
- DDM/DQ2 SW installation and maintenance on VO boxes
- keep data integrity in particular clean obsolete datasets and files records from DDM and LFC/LRC
- Help to ATLAS users to transfer data, in particular for the physicists from geographically closed Universities and Laboratories

DDM operation is considered as a primary job for the above teams during :

- ATLAS data-taking
- ATLAS global tests

Primary means that during data-taking and tests the operation team supports ATLAS Computing Shifts and provides help in case of data-transfer failure, contacts networking and computing experts to solve the problems with data transfer channels, storage system, disk space, etc, daily checks data transfer logs and status, defines tools necessary for DDM operation and monitoring, and communicates with DDM/DQ2 core team.

Tables 1 and 3 (Appendix 1) show responsibles for DDM central operation and ATLAS Tier-1 DDM operations team.

2.1 Funding and manpower issues

The formation of a Data Management team within a cloud (Tier-1/Tier-2s system), although a central ATLAS service, should be seen as the responsibility of the cloud members themselves and it is expected that the member sites (or associated funding bodies) provide the majority of the required manpower.

Breaking the DDM responsibilities down on the cloud level gives us smaller units of responsibility and closer contact to the associated sites. What is more, overlap with the responsibilities of other clouds will allow us to develop tools and practices in common while ensuring that clouds maintain their control over their data management.

It is estimated that initially 1.7-1.9 FTE equivalents would be needed to support DDM operations for each Tier-1 cloud; these operations include supporting and monitoring all previously described services and also 3D database replication. This manpower level will decrease to 1 FTE (over several people to guarantee full coverage) after a couple of years, as the automation and robustness of the tools will improve.

As an example a team may be formed from an expert (data-guru) who would be 75-100% dedicated to data management and O (3) other members who would give 30% of their time to DDM matters.

The formation of such a cloud based DDM team would ensure that the cloud runs smoother and is more configurable than would be possible if the data management was provided by one ATLAS wide service.

We are supposed to have fully operational DDM team by the time of ATLAS Offline and computing commissioning.

3 ATLAS data transfer and datasets subscription policy

ATLAS data transfer policy is defined in the Computing TDR [2]. In the next sections we try to define responsibilities of Tier-1 and Tier-2 personnel. In general, ATLAS users can subscribe for reasonable amount of data (organized by datasets [6]) and ask for data transfer between ATLAS Tiers or from the particular Tier to the personal computing facilities. The bulk data transfer can be requested by data preparation and physics groups; at the same time, ATLAS will have predefined data transfer (RAW, AOD, ESD and MC data) between Tiers as described in [2]. In case of resources saturation, the data transfer priority is given to the ATLAS data transfer, which can be redefined by ATLAS computing management. The data transfer prioritites between Tiers can be redefined by data preparation and physics group coordinators.
The ATLAS DDM operations team must constantly monitor the data transfer and report any ambiguities to CMB.

4 Tier-1 DDM operations

Regional DDM ops teams work together with the central team. The responsibilities are similar, though for the regional team one of the main priorities is to run DDM services in a stable way on Tier-1 and associated Tier-2s, monitor data transfer and assure that all (according to the computing model) data are transmitted to the centers. The data volume to be handled by each Tier-1 and data transmission rates are shown on Fig.1.

The relations between ATLAS Tiers are described as “cloud model”, where each Tier-1 provides services for a group of Tier-2s. Data transmission between Tier-1 and Tier-2s are described in the next section, it is important to mention that the Regional DDM Ops team can choose the local DDM deployment model, for example VO box installation for some of Tier-2s.

5 Tier-2 DDM operations

The data handling policy is described in more detail in [2].

5.1 Data at Tier-2 sites

We assume that Tier-2 sites will store :

- for long periods (years) :
  - Some Raw/ESD real data for small analysis: according to local requests
  - All production AODs from real data: 200TB/year
    (can be shared among Tier-2s in Tier-1 cloud)

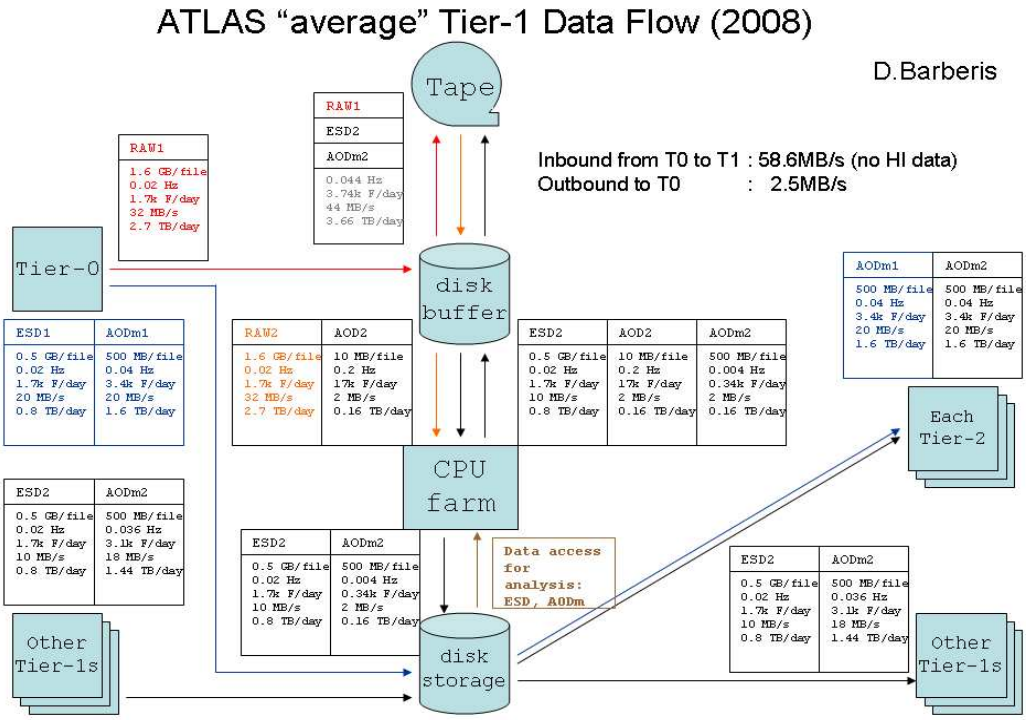

Figure 1: ATLAS "average" Tier-1 data flow (without MC data)

  - All MC AOD: 40TB/year [2)]
    (can be shared among Tier-2s in Tier-1 cloud)
  - All TAG: 2TB/year
  - Conditions data: several TB

- for shorter periods (buffer)
  - MC Raw/ESD:
    1600/5+500/5 = 400TB * (say 1 month of processing) *1/10 /30 = 1TB [3)]

5.2 Data transfers

Based on the above assumptions we will have the following transfers between Tier-1 and Tier-2 sites:[4)]

- Transfers from Tier-1 to Tier-2s
  - Some Raw/ESD real data: for small analysis

[2)]The current estimation for "raw" AODs is 200 TB (2 versions), MC will produce 20% of raw AODs, 200TB*20%=40 TB

[3)]ESD event size is 1600kB, AOD size 500kB, number of events in year 2008 - 1000M, this leads to a total of 1600TB ESD and 500TB AOD per year. 20% events for the estimate of MC needs leads to 400 TB. This MC needs to be produced at Tier2's (we assume ATLAS has 30 of them, so a factor of 1/30 per site) and in a period of 1 year (so a factor of 1/10 for 1 month buffer). Production of the second version of ESD/AOD would require increase in the buffer size

[4)]the transfer between Tier-0 and Tier-2s can be done in some predefined cases; we are not discussing this, because there is no final agreement how it will be organized

  - The accepted share (usually 1/3) of production AOD from real data
  - The accepted share (usually 1/3) of production AOD from MC data
  - All TAG data
  - Conditions data

- Transfers from Tier-2s to Tier-1
  - All MC Raw/ESD/AOD
  - Calibrations

At Tier-2 sites, probably only the most recent version of the AOD should be available. When the reprocessing is done, there will be a copy of new AOD data from Tier-1 to Tier-2 sites.

5.3 Storage and Data management

Tier-2s pledge for ATLAS some resources which should be made available for the whole Atlas community. Additional restrictions may come from the competition between the ATLAS production and common user sharing of resources - this will be handled by ATLAS policy and separate queues/disk catalogs for different groups of users. The group policy will be implemented based on VOMS groups and roles attributes of user authentication and authorization system on the *GRID*.

The production data transfers will in general be coordinated by the Regional DDM ops team. The tasks of this team will also include booking and managing of the pledged storage resources at Tier-1 and Tier-2 sites. Only the members of ATLAS DDM ops team should be allowed to write to reserved production space. They will be also responsible for removing old versions/obsolete data or the data transfered out to Tier-1 site from production space at Tier-2s with the agreement from authorities responsible for availability and quality of data, e.g. Data Preparation and Physics Coordinators.

The subscriptions to datasets will be done using a DQ2 client program. Later on the DQ2 system will pass transfer tasks to regional FTS servers for execution. These servers will have the capability to prioritize transfer tasks depending on the VOMS authorization attributes of the person submitting the task.

The control of the timing of transfers over FTS channels between Tier-1 and Tier-2 centers has been proposed in the following way :

- Tier-1 controls Tier-2 $\rightarrow$ Tier-1 channel
- Tier-2 controls Tier-1 $\rightarrow$ Tier-2 channel (if such a special channel is configured)
- Transfers from other Tier-1/Tier-2s are handled by an inclusive 'star' channel and cannot be controlled on an individual basis.

We propose the following schema of responsibility for dataset subscription in Tier-1 $\leftrightarrow$ Tier-2 data transfers:

- Transfers from Tier-1 to Tier-2s

- *Small* Raw/ESD samples from real/MC data: for private analysis.
  Users subscribe on their own. These subscriptions will go to a separate storage area so to prevent abuse of the storage space allocated for ATLAS production data.
- Real AOD from production.
  Tier-2 sites get an opportunity to say which data they are interested in by contacting DDM representatives either directly or by using some Web page interface with dataset availability presentation and dataset preference selection mechanism. DDM team reps would take these preferences into account as much as they can. They must however assure that all of the AOD data will be distributed among T2 sites sharing a single AOD sample copy.
- TAG data, Conditions data.
  These are subscribed for Tier-2s by DDM ops team.

- Transfers from Tier-2s to Tier-1
  - All MC Raw/ESD/AOD.
    The produced datasets are automatically registered in ATLAS central DDM DB. The subscription for the datasets needs to be automated. DDM ops will control datasets subscriptions, data transfer perfomance and data integrity.
  - Calibrations.
    If there is production of calibration data at Tier-2, they should take care of validation and distribution of the data to central repository.

6 Acknowledgements

We would like to thank Drs. S.Campana, R.Gardner, A. de Salvo, O.Smirnova, H.Severini and T.Wenaus for reading and discussing this document.

7 Appendix A : ATLAS DDM operations team members (Jul 2006)

Table 1 and Table 3 show the responsibles [5)] for DDM central operation and ATLAS Tier-1 DDM operations team.

Coordinator	Alexei Klimentov
Members	Simone Campana, Jiri Chudoba, Wensheng Deng, Cunfeng Feng, Stephane Jezequel, Z.Liang, Pavel Nevski and Oxana Smirnova
DB support	Yuri Smirnov, Jason Smith
Monitoring	Tomasz Wlodek
End-Users Tools	Tadashi Maeno and Dietrich Liko (as Distributed Analysis coordinator)

Table 1: ATLAS central team DDM operations

[5)] All people from the lists below have many other responsibilities and *NONE* of them works 100% for DDM operations

Tier-1	DDM Coordinator	DDM operations team
ASGC	Jason Shih	*TBS*
BNL	Wensheng Deng	Hironori Ito, Zhao Xin, Marty Dippel, Kristy Kallback-Rose
CERN	Alexei Klimentov	Pavel Nevski, Cunfeng Feng
CNAF	Guido Negri	*TBS*
FZK	John Kennedy	Jiri Chudoba, Andrzej Olszewski
LYON	Stephane Jezequel	Ghita Rahal
NG	*TBS*	*TBS*
PIC	Xavier Espinal	Mireia Dosil
RAL	Catalin Condurache	*TBS*
SARA	Jiri Chudoba	*TBS*
TRIUMF	Rod Walker	Denice Deatrich, Reda Tafirout

Table 2: ATLAS Regional DDM ops teams

US ATLAS	Alexei Klimentov

Table 3: ATLAS Regional DDM operations coordinators

References

[1] M.Branco, D.Cameron, P.Salgado "*https://uimon.cern.ch/twiki/bin/view/Atlas/DDM*". ATLAS DDM/DQ2 WiKi page

[2] "ATLAS Computing Technical Design Report". 4 July 2005

[3] M. Branco, D. Cameron, T. Wenaus. A Scalable Distributed Data Management System for ATLAS. Computing In High Energy Physics 2006 (CHEP 06), Mumbai, India, February 2006

[4] M.Branco, D.Cameron, P.Salgado "*https://twiki.cern.ch/twiki/bin/view/Atlas/DDMVOBoxRequirements*". ATLAS DDM VOBox Requirements.

[5] D. Barberis "ATLAS Tier-1-Tier-2 Associations". ATLAS note 8 June 2006;

[6] "ATLAS Dataset Definition". ATLAS SWING internal note 8 Mar 2006;

Glossary

CASTOR	Hierarchical storage system. CERN made
CMB	ATLAS Computing Management Board
DDM	Distributed Data Management
DQ2	Don Quixote de la Mancha- roman by Miguel de Cervantes, also ATLAS DDM SW
DPM	Disk Pool Manager
ESD	Event Summary Data
MC	Monte-Carlo, detector and physics simulation SW
SE	Storage Element
SW	SoftWare
TDR	Technical Design Report
VOBox	dedicated machine at each Tier-1 site (and sometimes Tier-2s) to run data management services and other services