ATL-SOFT-2004-007
CERN-LHCC-2004-037/G-085
V1.2
10 January 2005

THE ATLAS COMPUTING MODEL

Prepared by: D. Adams, D. Barberis, C. Bee, R. Hawkings, S. Jarp, R. Jones[1], D. Malon, L. Poggioli, G. Poulard, D. Quarrie, T. Wenaus

on behalf of the ATLAS Collaboration

Abstract: The ATLAS Offline Computing Model is described. The main emphasis is on the steady state, when normal running is established. The data flow from the output of the ATLAS trigger system through processing and analysis stages is analysed, in order to estimate the computing resources, in terms of CPU power, disk and tape storage and network bandwidth, which will be necessary to guarantee speedy access to ATLAS data to all members of the Collaboration. Data Challenges and the commissioning runs are used to prototype the Computing Model and test the infrastructure before the start of LHC operation.
The initial planning for the early stages of data-taking is also presented. In this phase, a greater degree of access to the unprocessed or partially processed raw data is envisaged.

[1] Chair and contact person: Roger.Jones@cern.ch

1 Introduction

The primary purpose of this document is to present the current ideas on the steady-state ATLAS offline computing model, to detail the tests performed to validate that model and to give the current best estimate of the resources required. The model extends from the primary event store, which is where events selected by the trigger are recorded, to the analyst at a remote university. Consideration is also given to computing issues in the DAQ and the High Level Trigger (HLT) farm, i.e. prior to the primary event store. The information presented will provide input to the LHCC review in January 2005, and also inform the LCG and ATLAS Computing TDRs and the construction of the Computing MoUs in 2005.

Also considered is the model for the commissioning of the computing system with real data. This will require enhanced access to raw and nearly-raw data for calibration, algorithm development etc. This short-term redeployment of the resources that will eventually provide the steady-state solution has clear implications for the resource providers.

The ideas and estimates presented here have evolved from previous studies, including the ATLAS computing resource and cost estimates developed using the MONARC hierarchical model [1] for the CERN LHC Computing Review [2]. The advent of the World Wide Grid triggered new ideas on how to organize the ATLAS system as a virtual worldwide distributed computing facility.

The main requirement on the Computing Model is to enable all members of the ATLAS Collaboration speedy access to all reconstructed data for analysis during the data-taking period, and appropriate access to raw data for organised monitoring, calibration and alignment activities. This document outlines a model that makes substantial use of Grid Computing concepts, thereby allowing the same level of data access, and making available the same amount of computing resources, to all members of the ATLAS Collaboration.

2 Executive Summary

The ATLAS computing model embraces the Grid paradigm and a high degree of decentralisation and sharing of computing resources. However, as different computer facilities are better suited to different roles, a degree of hierarchy, with distinct roles at each level, remains. This should not obscure the fact that all of the roles described are vital and must receive due weight. The required level of computing resources means that off-site facilities will be vital to the operation of ATLAS in a way that was not the case for previous CERN-based experiments.

The primary event processing occurs at CERN in a Tier-0 Facility. The RAW data is archived at CERN and copied (along with the primary processed data) to the Tier-1 facilities around the world. These facilities archive the RAW data, provide the reprocessing capacity, provide access to the various processed versions and allow scheduled analysis of the processed data by physics analysis groups. Derived datasets produced by the physics groups are copied to the Tier-2 facilities for further analysis. The Tier-2 facilities also provide the simulation capacity for the experiment, with the simulated data housed at Tier-1s. In addition, Tier-2 centres will provide analysis facilities and some will provide the capacity to produce calibrations based on processing some raw data. A CERN Analysis Facility provides an additional analysis capacity, with an important role in the calibration and algorithmic development work.

It is assumed that each of the facilities is responsible for funding the required resources for its declared role. This is in contrast with some earlier views of the financial organisation. In terms of organisation, ATLAS will negotiate relationships between Tier-1s and Tier-2s and also between Tier-1s themselves to try to optimise the smooth running of the system in terms of data transfer, balanced storage and network topologies.

The computing model gives rise to estimates of required resources that may be used to design the various facilities. It is not assumed that all Tier-1s or Tier-2s will be of the same size. However, the ratio of disk, tape and CPU resources required is implied in each case. An example of the resources required is given in Table 1, which shows the estimated resources for one full year of data taking in 2008 (to which the requirements for any initial running in 2007 need to be added).

	CPU(MSI2k)	Tape (PB)	Disk (PB)
CERN Tier-0	4.1	4.2	0.35
CERN AF	2.2	0.4	1.6
Sum of Tier-1's	18.0	6.5	12.3
Sum of Tier-2's	16.2	0.0	6.9
Total	40.5	11.1	21.2

Table 1: The estimated resources required for one full year of data taking with the 2008/2009 expected luminosity and live time.

3 Offline Computing Model

3.1 Input parameters for proton-proton collisions

Input parameters for the offline Computing Model are derived from the information contained in the "HLT/DAQ TDR" [3] and the "Physics TDR" [4]. All input parameters are to be considered as reference numbers, our best estimates at the moment. In the following sections, the ATLAS Computing Model is first worked out for one year of steady-state operation of proton-proton data taking. Later chapters deal with the commissioning of the system and the resource requirements as a function of time.

3.1.1 LHC design parameters and trigger rate

E = 14 TeV (two 7 TeV proton beams)
L = $0.5*10^{33}$ $cm^{-2}s^{-1}$ in 2007,
$2*10^{33}$ $cm^{-2}s^{-1}$ in 2008 and 2009,
10^{34} $cm^{-2}s^{-1}$ (design luminosity) from 2010 onwards
σ = 100 mb = 10^{-25} cm^2
Collision rate = $L \cdot \sigma = 10^9$ Hz p-p collisions at design luminosity
Trigger rate = 200 Hz independent of the luminosity.
It is assumed that the trigger thresholds and selection conditions will be adjusted continually so as to maximize the physics reach of the experiment.

3.1.2 Types of data

RAW	Real Raw Data, as recorded after the HLT
SIM	Simulated Data
DRD	Derived Reconstruction Data
ESD	Event Summary Data (after reconstruction)
AOD	Physics Analysis Object Data
DPD	Derived Physics Data (analogous with today's n-tuples)
TAG	Event tags, short event summaries primarily for event selection

The following assumptions are used to calculate the storage and computing resources:

Item	Unit	Value
Raw Data Size	MB	1.6
ESD Size	MB	0.5
AOD Size	kB	100
TAG Size	kB	1
Sim. Data Size	MB	2,0
Sim. ESD Size	MB	0,5
Time/Reco 1ev	kSI2k-sec	15
Time/Simu 1ev	kSI2k-sec	100
Time/Analyse 1ev	kSI2k-sec	0.5
Event rate after EF	Hz	200
Operation time	seconds/day	50000
Operation time	days/year	200
Operation time (07)	days/year	100
Event statistics	events/day	10^7
Event statistics	events/year	$2 \cdot 10^9$

The assumed processing times are projections based on those for the current code, in the light of planned future improvements and known inefficiencies, and are for the running conditions in 2008 and 2009.

Between data without pile-up and the pile-up for design luminosity, the event sizes are seen to grow by 50% and the processing time by 75%; this information is used in the resource evolution projections. At present, all processing-time numbers are higher than assumed here. For the event sizes, the first prototype of the AOD is smaller than assumed here (but has yet to be tested in terms of the required functionality for analysis). The other data formats are presently larger than the target size, but the target sizes are believed to be achievable.

3.2 Event Store

The physics event store holds a number of successively derived event representations, beginning with raw or simulated data and progressing through reconstruction into more streamlined event representations suitable for analysis. Constituent components are described in the following paragraphs.

RAW Data: RAW data are events as output by the Event Filter (EF, the final stage of the HLT) for reconstruction. The model assumes an event size of 1.6 megabytes, arriving at an output rate of 200 Hz (including 20 Hz of calibration trigger data).

Events arrive from the Event Filter in "bytestream" format, reflecting the format in which data are delivered from the detector, rather than in any object-oriented representation. Events will be transferred from the EF to the Tier-0 in files of at most 2 GB. Each file will contain events belonging to a single run (corresponding to a prolonged period of data-taking using the same trigger selections on the same fill in the accelerator), but the events in each file will not be consecutive nor ordered [5].

Event Summary Data (ESD): ESD refers to event data written as the output of the reconstruction process. ESD is intermediate in size between RAW and Analysis Object Data (see below). Its content is intended to make access to RAW data unnecessary for most physics applications other than for some calibration or re-reconstruction. ESD has an object-oriented representation, and is stored in POOL ROOT files. The target size is 500 kilobytes per event [6].

Analysis Object Data (AOD): AOD is a reduced event representation, derived from ESD, suitable for analysis. It contains physics objects and other elements of analysis interest. The target size is 100 kilobytes per event. It has an object-oriented representation, and is stored in POOL ROOT files.

Tag Data (TAG): TAG data are event-level metadata — thumbnail information about events to support efficient identification and selection of events of interest to a given analysis. To facilitate queries for event selection, TAG data are stored in a relational database. The assumed average size is 1 kilobyte per event.

Derived Physics Data (DPD): DPD is an n-tuple-style representation of event data for end-user analysis and histogramming. The inclusion of DPD in the computing model is an acknowledgment of the common practice by physicists of building subsamples in a format suitable for direct analysis and display by means of standard analysis tools (PAW, ROOT, JAS etc.), though software providers certainly expect that analysis, histogramming, and display via standard tools will be possible with AOD as input.

Simulated Event Data (SIM): SIM refers to a range of data types, beginning with generator events (e.g., from Pythia or similar programs) through simulation of interactions with the detector (e.g., Geant4 hits) and of detector response (digitization). It may also include pileup, the superposition of minimum bias events, or the simulation of cavern background. Events may be stored after any of these processing stages. The storage technology of choice is POOL ROOT files. Digitised events may alternatively be stored in bytestream format for trigger studies or for emulation of data coming from the Event Filter. Simulated events are often somewhat larger than RAW events (approximately 2 megabytes in size), in part because they usually retain Monte Carlo "truth" information.

Other formats are allowed in the software and processing model that are not included in the baseline. For example, the Derived Reconstruction Data (DRD) is an option being considered for the early phase of data taking for a subset of the data. It consists of raw data augmented with partially reconstructed objects to allow easy calibration and optimisation of detector code. This format is only of use if the trade-off between storage cost and CPU to derive the partial-reconstruction is in the favour of storage. This in turn depends on the sample size required and the number of times the sample is passed-over by the detector groups.

3.3 The Tier Structure and the Roles of the Various Tiers

While the ATLAS computing model is very much Grid-based, there still remain distinct roles for different facilities that may be characterised in the following ways. It is to be stressed that all are important and make an invaluable contribution to ATLAS, and a sensible balance between the resources in the various Tiers is essential for the operation of the computing model.

3.3.1 Tier-0 at CERN

The Tier-0 facility at CERN is responsible for the archival 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 somewhat slower 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 described below.

The Tier-0 must provide an extremely high availability and response time in the case of errors. In the event of prolonged down-time, first-pass processing and calibration must be taken over by the Tier-1 facilities described below. To account for failures and network outages, a disk buffer corresponding to about 5 days of data production will be required for the data flowing into the Tier-0. A smaller output buffer will be required in case of failures in the transfer of the derived datasets offsite (although switching to an alternate Tier-1 destination must be possible in the system).

3.3.2 Tier-1 Facilities

Approximately 10 Tier-1 facilities are planned world-wide that will serve ATLAS. They take responsibility to host and provide long-term access and archival of a subset of the RAW data (on average $1/10^{th}$ each). 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, with the most up-to-date version of the data available with short latency (‘on disk’) and the previous version available but perhaps with a longer latency (‘on tape’). The Tier-1s also undertake to host a secondary low-latency copy of the current ESD, AOD and TAG samples from another Tier-1, and the simulated data samples from Tier-2 facilities to improve access and provide fail-over. All of the datasets hosted are considered to be for the collaboration as a whole, and the storage and CPU pledged to be funded by the Tier-1 for that purpose.

The Tier-1s must allow access to and provide capacity to analyse all of the hosted samples, and will provide part of the calibration processing capacity. Modest RAW data samples must be available at short latency to allow calibration and algorithmic development. They will also host some of the physics working group DPD samples.

Tier-1 facilities are expected to have a high level of service in terms of availability and response time. Given the vital role in receiving the raw data and reprocessing, down-times in excess of 12 hours become problematic in terms of catching up with processing and with the storage elsewhere of RAW data. The fact that the ESD will be copied to two sites (see 3.4) reduces somewhat the reliance on a given Tier-1 for short periods.

3.3.3 Tier-2 Facilities

Tier-2 facilities may take a range of significant roles in ATLAS such as providing calibration constants, simulation and analysis. This range of roles will result in different sizes of the facilities. Tier-2 facilities also provide analysis capacity for physics working groups and subgroups. This analysis activity is generally chaotic in nature. They typically will host one third of the available current primary AOD and the full TAG samples. They will also host some of the physics group DPD samples, most likely in accordance with local interest. In addition, they will provide all of the required simulation capacity for the experiment (but with the simulated data typically migrated to the Tier-1 unless general on-demand access can be ensured at the site). Agreements on the primary host for the data from a given Tier-2 will be

negotiated, although some flexibility will be required in the case of access problems. The relationships formed will be influenced by the ATLAS organisational plans and by the networking topology available. The primary host arrangement will help the planning of network links and may well follow the arrangements within a region for Grid operations and user support.

The Tier-2s will also host modest samples of RAW and ESD data for code development. Some Tier-2s may take significant role in calibration following the local detector interests and involvements.

The level of service in terms of availability and response time expected of a Tier-2 is lower than for a Tier-1 (unless it chooses to host the simulated data it generates).

3.3.4 CERN Analysis Facility

The CERN analysis facility is, as the name suggests, primarily devoted to analysis, supporting a relatively large user community. It will also provide an important platform for calibration and code development. It will be particularly useful for user access to RAW data, given its co-location with the Tier-0 facility.

The CERN analysis facility is expected to have a level of service comparable to the Tier-0, given its key role in calibration and alignment and the requirement that these activities introduce minimal latency in the first-pass data processing.

3.4 Data Flow

The source of the input real data for the computing model is primarily the Event Filter (EF). Data passing directly from the online to offsite facilities for monitoring and calibration purposes will be discussed only briefly, as they have little impact on the total resources required, and also require further clarification. While the possibility of other locations for part of the EF is to be retained, the baseline assumption is that the EF resides at the ATLAS pit. Other arrangements have little impact on the computing model except on the network requirements from the ATLAS pit area. The input data to the EF will require approximately 10x10 Gbps links with very high reliability (and a large disk buffer in case of failures). The output data requires an average 320MB/s (3Gbps) link connecting it to the first-pass processing facility. Remote event filtering would require upwards of 10Gbps to the remote site, the precise bandwidth depending on the fraction of the Event Filter load migrated away from the ATLAS pit.

While the option of streaming data (see Appendix A) at the EF should be retained, the baseline model assumes a single primary stream containing all physics events flowing from the Event Filter to Tier-0. Several other auxiliary streams are also planned, the most important of which is a calibration hot-line containing calibration trigger events (which would most likely include certain physics event classes). This stream is required to produce calibrations of sufficient quality to allow a useful first-pass processing of the main stream with minimum latency. A working target (which remains to be shown to be achievable) is to process 50% of the data within 8 hours and 90% within 24 hours.

Two other auxiliary streams are planned. The first is an express-line of physics triggers containing about 5% of the full data rate. These will allow both the tuning of physics and detector algorithms and also a rapid alert on some high-profile physics triggers. It is to be stressed that any physics based on this stream must be validated with the ‘standard’ versions of the events in the primary physics stream. However, such a hot-line should lead to improved reconstruction. It is intended to make much of the early raw-data access in the model point to this and the calibration streams. The fractional rate of the express stream will vary with time, and will be discussed in the context of the commissioning.

The last minor stream contains pathological events, for instance those that fail in the event filter. These may pass the standard Tier-0 processing, but if not they will attract the attention of the development team.

On arrival at the input-disk buffer of the first-pass processing facility (henceforth known as Tier-0) at the input disk buffer, the raw data file:

a) is copied to Castor tape at CERN;

b) is copied to permanent mass storage in one of the Tier-1s;

c) calibration and alignment procedures are run on the corresponding calibration stream events;

d) the express stream is reconstructed with the best-estimate calibrations available;

e) once appropriate calibrations are in place, first-pass reconstruction ('prompt' reconstruction) is run on the primary event stream (containing all physics triggers), and the derived sets archived into Castor (these are known as the 'primary' data sets, subsequent reprocessing giving rise to better versions that supersede them);

f) **two** instances of the derived ESD are exported to external Tier-1 facilities; each Tier-1 site assumes principal responsibility for its fraction of such data, and retains a replica of another equal fraction of the ESD for which another Tier-1 site is principally responsible. Tier-1 sites make current ESD available on disk.[2] ESD distribution from CERN occurs at completion of first-pass reconstruction processing of each file. As physics applications may need to navigate from ESD to RAW data, it is convenient to use the same placement rules for ESD as for RAW, i.e., if a site hosts specific RAW events, then it also hosts the corresponding ESD. The proposed "one file in, one file out" model for ESD production jobs makes achieving such correspondence simpler.

g) the derived AOD is archived via the CERN analysis facility and an instance is shipped to **each** of the external Tier-1s (a full copy at each Tier-1);

h) the AOD copy at each Tier-1 is replicated and shared between the associated Tier-2 facilities;

i) the derived TAG is archived into Castor and an instance is copied to **each** Tier-1. These copies are then replicated to each Tier-2 in full.

Step b), the transfer of the RAW data to external Tier-1 facilities, is an important requirement. These sites are the primary data sources for any later re-reconstruction of that data, and serve as the principal sources of CPU resources for any such reprocessing. It not only allows reprocessing of data, asynchronous with data-taking, it also allows additional capacity to be employed if there is a backlog of first-pass processing at the Tier-0. Note that this implies a degree of control over the Tier-1 environment and processing that is comparable to that at the Tier-0.

Selected ESD will also be copied to Tier-2 sites for specialized purposes. Resource estimates reflect this fact, but the models and policies by which this replication may be accomplished are negotiated among Tier-1 centres and their associated Tier-2 sites.

The AOD and TAG distribution models are similar, but employ different replication infrastructure because TAG data are database-resident. AOD and TAG distribution from CERN occur upon completion of first-pass reconstruction processing of each run.

3.5 First-pass Processing

The assumed input to the first-pass processing is bytestream RAW data and the output is ESD. It is assumed that in normal operations the first-pass processing is conducted on the CERN Tier-0 facility, although the Tier-1 facilities could provide additional capacity in exceptional circumstances.

We make the following assumptions about output from the Event Filter and input to first-pass reconstruction:

- Event Filter processors send their outputs to one of 30-50 SubFarm Output managers (SFOs).
- Events are written to files in bytestream format by SFOs.
- SFOs are equivalent to one another, and do not sort physics events by trigger or type.

[2] At least one Tier-1 site proposes to host the entire ESD. This is not precluded, but the site would nonetheless, like every other Tier-1, assume principal responsibility for its agreed fraction of the ESD.

- In normal operation, SFOs will fill a file to a specified size or event count threshold, then close the file and open a new one.
- Files are the unit of transfer from the Event Filter to the Tier-0 centre.
- Files are eligible for transfer as soon as they are filled and closed.
- The data acquisition system assigns run numbers and event numbers, and provides event timestamps.
- At run boundaries, files written by SFOs are closed and transferred to the Tier-0 centre: no RAW data file from the Event Filter contains events from more than one run.

Note that this process does not ensure that the events within a file are temporally ordered.

3.5.1 Rates, latency, and buffering

An output rate of 200 Hz is approximately 4 Hz per SFO. At 1.6 megabytes per event, each SFO will fill a 2-gigabyte file with approximately 1250 events every 5 minutes. This in itself sets a minimum time before processing can proceed for any stream. However, this merely sets the latency for the 'prompt' reconstruction; for the primary stream, the latencies will be set by the time until calibration, alignment, and other conditions data are available to Tier-0 processors as discussed in section 3.9.3. Data arrives at the Tier-0 at a rate of 320 megabytes per second. A disk pool requires approximately ~25 terabytes for each day of buffer capacity it is proposed to provide. Proposed criteria for overall production latency are:

- The express and calibration data streams reconstructed with less than 8 hours' latency;
- 90% of primary data stream reconstructed within 48 hours, the bulk beginning after approximately 24 hours.

The system is assumed to have an input disk buffer of 127 TB, which corresponds to approximately five days of data-taking.

The processing of the calibration and alignment data, which crucially sets the latency for the bulk processing, is discussed in section 3.9.

It should be noted that the first-pass processing provides the opportunity for more sophisticated filtering and compression/data reduction of the RAW data sample. As confidence is gained with the processing chain and the understanding of the detector, this may be taken advantage of to reduce the RAW data stored at the remote sites (and consequently the volume of derived data). However, for the baseline model we do not assume such a reduction. In the baseline model, the RAW data distributed to the Tier-1s is a straight copy of the data received from the Event Filter, and will often be shipped offsite before the first-pass processing has occurred.

3.5.2 First-pass ESD production

First-pass ESD production takes place at the Tier-0 centre. The unit of ESD production is the run, as defined by the ATLAS data acquisition system. ESD production begins as soon as RAW data files and appropriate calibration and conditions data arrive at the Tier-0 centre. The Tier-0 centre provides processing resources sufficient to reconstruct events at the rate at which they arrive from the Event Filter. These resources are dedicated to ATLAS event reconstruction during periods of data taking. The current estimate of the CPU required is 3000 kSI2k (approximately 15 kSI2k-seconds per event times 200 events/second).

A new job is launched for each RAW data file arriving at the Tier-0 centre. Each ESD production job takes a single RAW event data file in bytestream format as input and produces a single file of reconstructed events in a POOL ROOT file as output. With the current projection of 500 kilobytes per event in the Event Summary Data (ESD), a 2-gigabyte input file of 1250 1.6-megabyte RAW events yields a 625-megabyte output file of reconstructed (ESD) events as output.

3.5.3 First-pass AOD production

Production of Analysis Object Data (AOD) from ESD is a lightweight process in terms of CPU resources, extracting and deriving physics information from the bulk output of reconstruction (ESD) for use in analysis. An AOD production job in principle takes one or more ESD files in POOL ROOT format as input, and produces one or more POOL ROOT files containing AOD as output. Current estimates propose an AOD size of 100 kilobytes per event.

AOD must be derivable from ESD without reference to RAW data, but one might imagine concatenating {RAW→ESD→AOD} production into a single job for the sake of efficiency. While the database and control framework infrastructure support such concatenation, the current model separates ESD from AOD production. The reason is that job concatenation results in a large number of small files: a concatenated job that takes a single 2-gigabyte RAW event file as input, while producing a 625-megabyte ESD file, would produce only a 125-megabyte AOD file, even if only one output AOD stream is written. If AOD output is written to multiple streams, AOD files are likely to average approximately 12 megabyte in size. The proposed model for first-pass AOD production is therefore to run it as a separate step at the Tier-0 centre, using on the order of 50 ESD files as input to each AOD production job.

As AOD events will be read many times more often than ESD and RAW data, AOD events are physically clustered on output by trigger or physics channel or other criteria that reflect analysis access patterns. This means that an AOD production job, unlike an ESD production job, produces many output files. The streaming model is that each AOD event is written to exactly one stream: AOD output streams comprise a disjoint partition of the run. All streams produced in first-pass reconstruction share the same definition of AOD. On the order of 10 streams are anticipated in first-pass reconstruction.

It is of course true that some events are of interest to more than one physics working group. Such events are nonetheless written exactly once, to avoid complications for analyses that cross stream boundaries (e.g., Have I already seen this event in another stream?). Streams should be thought of as heuristically-based data access optimisations: the idea is to try to reduce the number of files that need to be touched in an average analysis, not to produce perfect samples for every analysis. More specialized sample building will take place at Tier-1 and Tier-2 centres. Every Tier-1 centre receives a complete copy of the AOD — all of the streams. The streams merely control which events are close to which other events in which files.

3.5.4 TAG production

AOD production jobs simultaneously write event-level metadata (event TAGs), along with “pointers” to POOL file-resident event data, for later import into relational databases. The purpose of such event collections is to support event selection (both within and across streams), and later direct navigation to exactly the events that satisfy a predicate (e.g., a cut) on TAG attributes.

To avoid concurrency control issues during first-pass reconstruction, TAGs are written by AOD production jobs to files as POOL “explicit collections,” rather than directly to a relational database. These file-resident collections from the many AOD production jobs involved in first-pass run reconstruction are subsequently concatenated and imported into relational tables that may be indexed to support query processing in less than linear time.

While each event is written to exactly one AOD stream, references to the event and corresponding event-level metadata may be written to more than one TAG collection. In this manner, physics working groups may, for example, build collections of references to events corresponding to (possibly overlapping) samples of interest, without writing event data multiple times. A master collection (“all events”) will serve as a global TAG database. Collections corresponding to the AOD streams, but also collections that span stream boundaries, will be built during first-pass reconstruction.

Standard utilities provided by the database group make it possible to analyse the events in such collections by following the pointers to the corresponding events at sites that host the corresponding event data (e.g.,

all Tier-1 sites for AOD), or, alternatively, to run extraction jobs that iterate over all events in such a collection and extract (copy) the corresponding data into personal files that contain those events and no others, perhaps for shipment to smaller-scale facilities or personal computers. Such extraction is expected to be common at Tier-1 and Tier-2 sites.

3.6 Reprocessing

The current model assumes that the new data will be reprocessed approximately 2-3 months after acquisition using the same software version but improved calibration and alignments. These will be obtained from continued study of the calibration stream data and also of the first-pass ESD. It is this 'offline' calibration process that sets the timescale for the reprocessing. A second reprocessing of the complete dataset, including the data from previous years, is envisaged at the end of data taking each year, using up-to-date algorithms and calibrations. The reprocessing will probably take place more frequently during the first couple years. In some cases it may be possible to reprocess starting from ESD rather than going back to the raw data.

It is assumed that the bulk reprocessing occurs at the Tier-1 facilities, and that the dominant access to RAW data will be through this scheduled and read-occasionally process. It is possible that the EF farm could be pressed into service for this activity; we believe that the architecture of the Tier-0 and EF farms should allow the repartitioning of resources between the two. However, such dual use is not to be assumed in the baseline computing model.

3.7 Data Analysis

The types of event data (SIM/RAW, ESD, AOD, TAG, DPD) are described in an earlier section. All the useful events acquired from the detector and those events from large production Monte Carlo samples will be reconstructed to produce ESD, AOD and TAG using the production process described in the previous section. In principle, a user wishing to perform data analysis can access any of this data but, in practice, the available resources (CPU, disk and bandwidth) for any one job will limit access to a small fraction. This is true not because most of the resources are dedicated to production but because there are many such analysts each typically submitting a series of jobs in an iterative analysis process.

It should be noted in the following that analysis in 'local' Tier-3 facilities (which may in practice be a fraction of a facility otherwise regarded as a Tier-2 or even a Tier-1), where the resource allocation and sharing is completely under local control for a local community, is not included. This may form an appreciable additional resource in the overall ATLAS computing.

3.7.1 Analysis procedures and data flow

The resources required by analysis jobs will vary widely and a combination of physics priority and fair share will be used to allocate resources and thus determine which jobs run and when. On the communal ATLAS Tier-2 resources, each user will be assigned a monthly resource quota that can be extended with approval from a physics group. Jobs exceeding the quota assigned to an individual user can be submitted to the production system for processing in a more controlled manner with priority assigned by a physics group. All large-scale access to Tier-1 resources must be arranged through physics or detector groups, given the resource implications. A Grid computing system will enable processing to take place at remote sites and even multiple sites for a single job. The Grid model also makes the system extensible: non-ATLAS Grid resources can easily be utilized when available and needed.

The goal of the data organisation is to enable users to identify the input dataset of interest and then to enable the processing system to gain efficient access to the associated data. Here "dataset" refers to some collection of data, possibly, but not necessarily, all the data in a single file or collection of files. These datasets are catalogued along with metadata specifying their content (ESD, AOD, …) and bookkeeping data specifying their provenance and quality to enable users to make selections. The provenance also enables users to discover if the output dataset they intend to create already exists or at least appears in the catalogue.

The datasets appearing in the metadata catalogue are typically virtual, i.e. they have no specific location (e.g. list of files) holding their data. Instead there is a dataset replica catalogue, which provides the mapping to one or more concrete replicas for each virtual dataset. The processing system may choose from these the concrete dataset whose location provides the most efficient data access. The job description should also include the required content to ensure that the input dataset is suitable for processing. If the input dataset includes unneeded content, it may be replaced with a sub-dataset removing the need to stage unused data.

The POOL collection files make use of the ROOT infrastructure and so the AOD and ESD event headers and objects they reference are directly accessible once the files holding these objects are available. The distributed analysis system will typically stage all required files on a local disk before starting a processing job. If the data are sparse, i.e. the dataset does not reference all the data in the files, the system may copy the referenced data (e.g. selected events) into new files to avoid transferring data that will not be accessed. A new concrete dataset may be formed from these files. This dataset is a concrete replica equivalent to the original. The new files and dataset may be catalogued for future processing.

Both ESD and AOD are stored in POOL event collection files and are processed using the ATLAS software framework, Athena. The TAG data are stored in relational tables as event attributes for these pool collections. The decreasing event size in the event model allow an analyst with a given set of resources, to process a much large number of AOD events than ESD or RAW events. In addition, the AOD is likely to be more accessible with a full copy at each Tier-1 site and large samples at Tier-2 sites. An analyst beginning with a sample containing a very large number of events can issue a query against the TAG data to select a subset of events for processing using AOD or ESD.

A typical analysis scenario might begin with the physicist issuing a query against a very large tag dataset, e.g. the latest reconstruction of all data taken to date. For example, the query might be for events with three leptons and missing transverse energy above some threshold. The result of this query is used to define a dataset with the AOD information for these events. The analyst could then provide an Athena algorithm to make further event selection by refining the electron quality or missing transverse energy calculations. The new output dataset might be used to create an n-tuple for further analysis or the AOD data for the selected events could be copied into new files. A subset of particularly striking events identified in one of these samples could be used to construct a dataset that includes the ESD and perhaps even RAW data for these events. The physicist might then redo the electron reconstruction for these events and then use it to create a new AOD collection or n-tuple.

An actual analysis would be much more complicated with steps being repeated and the addition of Monte Carlo signal and background samples. Large data samples (say 0.1 TB and larger) will be processed using the distributed analysis system where the user specifies an input data dataset and query or algorithm (also known as “transformation”) to apply to this dataset and the processing system generates an output dataset. Each dataset may include event data and/or summary (histogram, n-tuple…) data. An event dataset may be represented in many ways: a deep copy of the included data, a copy of the relevant event headers, the tokens for these event headers, a list of event identifiers with reference to another dataset, or simply references to the transformation and input dataset (virtual data). The processing system decides which is most appropriate and where to place the associated data possibly with some guidance from the user. This enables the system to balance usage of the different resources (processing, storage and network).

3.7.2 Resource Model for Analysis

For the purposes of estimation of the required resources for analysis, the analysis activity is divided into two components. The first is a scheduled activity run through the working groups, analysing the ESD and other samples and extracting new TAG selections and working group enhanced AOD sets or n-tuple equivalents. The jobs involved would be developed at Tier-2 sites using small sub-samples in a chaotic manner, but would be approved for running over the large data sets by physics group organisers. It is assumed there are ~20 physics groups at any given time, and that each will run over the full sample four times in each year. It is also assumed that only two of these runs will be retained, one current and one previous.

The second class of user analysis is chaotic in nature and run by individuals. It is mainly undertaken in the Tier-2 facilities, and includes direct analysis of AOD and small ESD sets and analysis of DPD. It is estimated that for the analysis of DPD, some 25 passes over 1% of the events collected each year would require only 92 SI2k per user. (It should be kept in mind that the majority of the user analysis work is to be done in Tier-3s.) Assuming the user reconstructs one thousandth of the physics events once per year, this requires a more substantial 1.8 kSI2k per user. It is assumed the user also undertakes CPU-intensive work requiring an additional 12.8 kSI2k per user, equivalent to the community of 600 people running private simulations each year equal to 20% of the data-taking rate. Such private simulation is in fact observed in some experiments, although it must be stressed that all samples to be used in published papers must become part of the official simulation sets and have known provenance. It is assumed that each user requires 1 TB of storage by 2007/2008, with a similar amount archived.

In this view, the activities of the CERN Analysis Facility are seen to be those of a large Tier-2, but with a higher-than-usual number of ATLAS users (~100), but without the simulation responsibilities required of a normal Tier-2. We envisage ~30 Tier-2 facilities of various sizes, with an active physics community of ~600 users accessing the non-CERN facilities.

3.7.3 Distributed analysis system

The distributed analysis system is based on web services and enables users to submit jobs from any location with processing to take place at remote sites. It provides the means to return a description of the output dataset and to enable the user to access quickly the associated summary data. Complete results are available after the job finishes and partial results are available during processing. The system ensures that all jobs, output datasets and associated provenance information (including transformations) are recorded in the catalogues. In addition, users have the opportunity to assign metadata and annotations to datasets as well as jobs and transformations to aid in future selection.

The distributed analysis system will typically split a user job request into a collection of subjobs, usually by splitting the input dataset. The results (output datasets) of the subjobs will be merged to form the overall output dataset. The distributed analysis system will decide how to split and merge datasets taking into account available resources and user requirements such as response time. Some fraction of resources will be dedicated to interactive analysis that enables users to examine (at least partial) results 10-100 seconds after job submission.

3.8 Simulation Process

Simulated data are assumed to be produced in the external Tier-2 facilities. Once produced, the simulated data must be available for the whole collaboration on an essentially 24/7 basis, as for real data. This requirement suggests that the simulated data should be concentrated at the Tier-1 facilities unless the lower tiers can guarantee the required level of access. However, it is assumed that all of the required derived datasets (ESD, AOD and TAG) are produced together at the same site, and then transported to their eventual storage location.

If the produced simulated data are to be shipped to the local storage site, this would contribute to the required connectivity to the regional Tier-1. If it is not shipped offsite, the bandwidth must also support the replication offsite on demand when the simulated data are analysed, or else the local facility must also support (and be credited for) the processing of ATLAS analysis jobs from all regions. The relative merits of moving jobs to the data or data to the jobs are a key issue in the Grid, and the optimal strategy depends on the requested data volume, the CPU requirement of the job and the available network interconnect. In general, all the factors tend to suggest that the storage and analysis of simulated data are best handled through the Tier-1 facilities by default, although some larger Tier-2 facilities may wish to share this load, with appropriate credit.

In addition to the official datasets that are officially requested by ATLAS and available to all, there will be additional samples that will be generated locally to test and optimise the simulation procedures and support local analyses. These only enter in the computing model in terms of their resource requirements, and are

accounted in the resource requirements per user in the analysis section.

3.9 Alignment and Calibration

Calibration and alignment processing refers to the processes that generate 'non-event' data that are needed for the reconstruction of ATLAS event data, including processing in the trigger/event filter system, prompt reconstruction and subsequent later reconstruction passes. This 'non-event' data (i.e. calibration or alignment files) are generally produced by processing some raw data from one or more sub-detectors, rather than being raw data itself, so e.g. Detector Control Systems (DCS) data are not included here. The input raw data can be in the event stream (either normal physics events or special calibration triggers) or can be processed directly in the subdetector readout systems. The output calibration and alignment data will be stored in the conditions database, and may be fed back to the online system for use in subsequent data-taking, as well as being used for later reconstruction passes.

Calibration and alignment activities impact the computing model in several ways. Some calibration will be performed online, and require dedicated triggers, CPU and disk resources for the storage of intermediate data, which will be provided by the event filter farm or a separate dedicated online farm. Other calibration processing will be carried out using the recorded raw data before prompt reconstruction of that data can begin, introducing significant latency in the prompt reconstruction at Tier 0. Further processing will be performed using the output of prompt reconstruction, requiring access to AOD, ESD and in some cases even RAW data, and leading to improved calibration data that must be distributed for subsequent reconstruction passes and user data analysis.

3.9.1 Types of processing

Various types of calibration and alignment processing can be distinguished:

1. Processing directly in the subdetector readout system (the RODs). In this case, the processing is done using partial event fragments from one subdetector only, and these raw data fragments do not need to be passed up through the standard ATLAS DAQ chain into the event stream (except for debugging). This mode of operation can be used in dedicated standalone calibration runs, or using special triggers during normal physics data-taking.
2. Processing in the EF system, with algorithms either consuming dedicated calibration triggers (identified in the level 1 trigger or HLT), or 'spying' on physics events as part of the normal processing. In particular, an algorithm running at the end of a chain of event filter algorithms would have access to all the reconstructed information (e.g. tracks) produced during event filter processing, which may be an ideal point to perform some types of calibration or monitoring tasks. If the calibration events are identified at level 1 or 2, the event filter architecture allows such events to be sent to dedicated sub-farms, or even for remote processing at outside institutes.
3. Processing after the event filter, but before prompt reconstruction. Event bytestream RAW data files will be copied from the event filter to the Tier-0 input buffer disk as soon as they are ready, and could then be processed by dedicated calibration tasks running in advance of prompt reconstruction. This could be done using part of the Tier-0 resources, or event files could also be sent to remote institutes for processing, the calibration results being sent back for use in later prompt reconstruction, provided the latency and network reliability issues can be kept under control.
4. Processing offline after prompt reconstruction. This would most likely run on outside Tier-1 or Tier-2 centres associated with the subdetector calibration communities, leaving CERN computing resources free to concentrate on other tasks. RAW data, ESD and AOD will all be distributed outside CERN, though data from more than one centre would be needed to process a complete sample due to the 'round robin' distribution of RAW and ESD to Tier-1 centres.

All of these processing types will be used by one or more of the ATLAS subdetectors; the detailed calibration plans for each subdetector are still evolving. The present emphasis is on understanding the subdetector requirements, and ensuring they are compatible with the various constraints imposed by the different types of online and offline processing.

3.9.2 Calibration streams

As discussed above, the output from the event filter will consist of four main streams: the principal physics stream, an express stream of 'discovery-type' physics, a calibration stream, and a diagnostic stream of pathological events. A first outline and incomplete proposal for the calibration stream is given below, though the contents will evolve towards and beyond the start of data taking:

- An inner detector alignment stream, with 10-100 Hz of reconstructed track information (not raw data), processed in the event filter, and amounting to a maximum of 4 MB/second.
- A LAr electromagnetic calorimeter calibration stream, with 50 Hz of inclusive electron candidates identified in the event filter. All five time samples of the electromagnetic calorimeter would be written out, but only for the region around the electron candidate, amounting to 50 kB/event or 2.5 MB/second.
- A muon calibration stream, taking a muon chamber region of interest identified at level 1 and outputting muon (MDT and trigger chamber) hit data for autocalibration at the full level 1 rate of O(10 kHz), corresponding to 6 MB/second. This may have a significant effect on the muon level 1 trigger for physics data taking, and may only be done e.g. for a few hours each week.
- Inclusive high p_T electrons and muons selected by the event filter at 20 Hz, with full event readout (32 MB/second). All these events will also be written in the primary physics stream, but duplicating these data into separate calibration streams will greatly facilitate efficient access for global detector debugging, calibration and reconstruction tuning. This will be especially important during initial running, and it is anticipated that this stream would gradually be phased out as data-taking advances and experience is gained with handling the primary physics stream through event collections.

These streams sum to a total data rate of about 45 MB/second, dominated by the inclusive high p_T leptons, corresponding to 13% of the total bandwidth out of the event filter (200 Hz of 1.6 MB events). The RAW data for all these streams corresponds to 450 TB/year, not counting ESD and subsequent reprocessing passes (which will be frequent at least at the beginning). It is clear only a fraction of this data will be able to be kept on disk at Tier-0, and priority will have to be given to the most recent data. However, this should be acceptable as most of the data is only needed for short-term calibration and debugging activities that should be complete in a few days or weeks, at least once the initial start-up phase is completed.

3.9.3 Prompt reconstruction latency

Prompt reconstruction latency refers to the time between the data being taken and it being processed through all stages so as to be ready for user analysis. Assuming that event filter output nodes (SFOs) write 2 GB files, each one will fill and close such a file every five minutes or so, and this should be transferred to the Tier-0 input buffer disk within a further few minutes. The reconstruction and ESD production for each file will be done on a single processor, and is expected to take around five hours, with AOD production introducing a small extra latency.

The latency incurred in the preparation of conditions data is much more difficult to assess. Several subdetectors have calibration constants that are expected to change significantly for each LHC fill, and these will be re-determined every fill or every day using dedicated calibration tasks running from the raw data or independently of the event stream (e.g. optical muon and inner detector optical alignment systems). It seems unlikely that the calibration processing to determine the first-pass calibration constants can be completed much sooner than 24 hours after the end of the fill (this may involve some preliminary reconstruction on dedicated calibration samples followed by verification on independent samples). A global optimisation is needed amongst all subdetectors to see what would be gained by a target of 12, 24 or 48 hours, balanced against the need for increased disk buffer storage at Tier-0 to avoid staging all the data to tape and then back in again for prompt reconstruction. The express physics stream would probably be processed more quickly, perhaps using the constants derived from the previous fill. Such a fast turnaround would also provide a sample of data useful for rapid data quality monitoring using fully reconstructed events. In general, considering possible failures in the calibration process and the potential problems introduced by delayed or off-site first-pass processing, a buffer corresponding to five days of input RAW

data will be required for ATLAS.

3.9.4 Offline calibration and alignment

In principle, offline calibration and alignment processing is no different to any other type of physics analysis activity and could be treated as such. In practice, many calibration activities will need access to large event samples of ESD or even RAW data, and so will involve resource-intensive passes through large amounts of data on the Tier-1s or even the Tier-0 facility. Such activities will have to be carefully planned and managed in a similar way to bulk physics group productions. At present, the offline calibration processing needs are not sufficiently well understood, though the recent definitions of the contents of DRD, ESD and AOD should help subdetectors in defining what processing tasks they need to do, and how they will accomplish them.

3.10 Heavy Ion data

It is foreseen that ATLAS will take data with Heavy Ion (initially Pb-Pb) collisions for approximately one month per year, starting in late 2008[7]. Here we assume that the Heavy Ion data-taking period will last 10^6 seconds (50000 seconds/day for 20 days) each year. The trigger rate is effectively limited by the available bandwidth between the HLT and Tier-0 buffers (320 MB/s) and the event size (5 MB) to ~50 Hz. We take this as a reference number for the purpose of this document.

The Computing Model for Heavy Ion data is still not completely worked out. The simplest assumption is that the data flow and processing pattern would be exactly the same as for p-p interactions. In this case, Heavy Ion data require an addition of 10% to storage needs and a larger increase to CPU requirements in all computing centres.

Another possibility is that Heavy Ion data are distributed only to the subset of Tier-1s hosting the communities of physicists most interested in those data, and processed only there. This model necessitates a study of the available network bandwidth to those computing centres and significantly increased resources to be made available in those centres, although of course the total amount of required resources would remain the same as in the “simple” model above.

4 Commissioning the System

The data processing in the very early phase of data taking will be rather different. While the distribution and access to the data should be well-prepared and debugged by the various data challenges, there will still be a requirement for heightened access to raw data to produce the primary calibrations and to optimise the reconstruction algorithms in the light of the inevitable surprises thrown up by real data. The access to raw data is envisaged in two formats, RAW files and (if sensible) DRD.

As will be seen in the next section, the steady-state model has considerable capacity for analysis and detector/physics group files. There is also a considerable planned capacity for analysis and optimisation work in the CERN analysis facility. It is envisaged that in the early stages of data-taking, much of this is taken up with a deep copy of the express and calibration stream data. For the initial weeks, the express data may be as upwards of 20 Hz, but it is clear that averaged over the first year, it must be less than this. If this averages at 10 Hz over the full year, and we assume we require two processing versions to be retained at any time at the CERN analysis facility, this translates to 620 TB of disk.

It is also assumed that there will be considerable reprocessing of these special streams. The CPU involved must not be underestimated. For example, to process the sample 10 times in 6 months would require a CPU capacity of 1.1 MSI2k (approximately 1000 current processors). This is before any real analysis is considered. Given the resource requirements, even reprocessing this complete smaller sample will have to be scheduled and organised through the physics/computing management.

Groups must therefore assess carefully the required sample sizes for a given task. If these are small enough, they can be replicated to Tier-2 sites and processed in a more ad hoc manner there. Some level of ad hoc reprocessing will of course be possible on the CERN Analysis Facility.

The CERN Analysis Facility resources are determined in the computing model by a steady-state mixture of activities that includes AOD-based and ESD-based analysis and steady-state calibration and algorithmic development activities. This gives 1.1 PB of disk, 0.58 PB of tape and 1.7 MSI2k processing power for the initial year of data taking. This resource will initially be used far more for the sort of RAW-data based activity described in sections 3.6 and 3.9, but must make a planned transition to the steady state through the first year. If the RAW data activities continue in the large scale for longer, the work must move to be shared by other facilities. The Tier-1 facilities will also provide calibration and algorithmic development facilities throughout, but these will be limited by the high demands placed on the available CPU by reprocessing and ESD analysis.

There is considerable flexibility in the software chain in the format and storage mode of the output datasets. For example, in the unlikely event of navigation between ESD and RAW proving problematic when stored in separate files, they could be written to the same file. As this has major resource implications if it were adopted as a general practice, this would have to be for a done for a finite time and on a subset of the data. Another option that may help the initial commissioning process is to produce DRD, which is essentially RAW data plus selected ESD objects. This data format could be used the commissioning of some detectors where the overhead of repeatedly producing ESD from RAW is high and the cost of storage of copies of RAW+ESD would be prohibitive. In general, the aim is to retain flexibility for the early stage of data taking in both the software and processing chain and in the use of the resources available.

5 Resource Requirements

5.1 Resource Requirements for One Year of Data-Taking

The primary purpose of the computing model exercise is to describe the model and estimate the computing resources required for a full year of data taking in 2008, with the inputs as described in section 3.1. Two main areas of computing activity have to be considered:

- Large-scale production, e.g. reconstruction, MC simulation. Two processing passes are assumed for the dataset in a year; the model is as described above in section 3.6.
- Data analysis.

The data analysis is more uncertain in its requirements, but a plausible scenario has been outlined in section 3.7. It should be stressed that the model and the resource requirements below do not include the personal analysis performed at the 'Tier-3' resources (or their equivalent share of Tier-2s and Tier-1s). The user load includes a CPU-intensive term that is in aggregate equal to the simulation of 20% of the data rate. (This may be full simulation, fast simulation, reconstruction or other CPU-intensive work), another that describes the analysis of the working group DPDs in a chaotic fashion, and a final contribution for limited event re-reconstruction. Given the many different analyses to be performed and the various data formats required, these are by necessity approximations.

It is important to note that that the predictions that follow include allowance for inefficiencies in the usage of CPU and disk resources. For scheduled CPU usage, the efficiency is assumed to be 85%, while for chaotic usage it falls to 60%. The disk storage efficiency is taken to be 70%, while tape storage efficiencies are assumed to be 100%.

The model assumes there will be 10 Tier-1s, which will be of different sizes. We assume there are on average three Tier-2 facilities for each Tier-1; again, there will be a range of sizes of Tier-2. The CERN Analysis Facility is taken to have 100 active ATLAS users (in addition to the active 600 users associated with the Tier-2s), five times the size of a nominal Tier-2.

The simulation is (along with the analysis load) a key driver of the Tier-2 resource requirements. It also has an effect on the Tier-1 storage and CPU requirements, as they host the simulated data. Many previous experiments have been severely limited in their capacity to produce fully simulated data, and this has hampered their physics output. Table 2 and Table 3 show the components that depend on the percentage of the data that is fully simulated as separate items for the case of 20% and 100%. Full simulation of 20% of the full data set ($4x10^8$ events simulated per year) would be barely acceptable.

	CPU (MSI2k)	Tape (PB)	Disk (PB)
CERN T0			
Simulation	0	0.0	0.0
Other	4	4.2	0.4
CERN AF			
Simulation	0	0.1	0.4
Other	2	0.4	1.3
Tier 1			
Simulation	2.8	1.3	1.7
Other	15.2	5.2	10.6
Tier 2			
Simulation	5.6	0.0	1.0
Other	10.6	0.0	5.9

Table 2: The projected resources for handling the 2008 data alone, assuming full simulation of 20% of the data rate.

	CPU (MSI2k)	Tape (PB)	Disk (PB)
CERN T0			
Simulation	0	0.0	0.0
Other	4	4.2	0.4
CERN AF			
Simulation	0	0.3	1.8
Other	2	0.4	1.3
Tier 1			
Simulation	14.0	6.4	8.6
Other	15.2	5.2	10.6
Tier 2			
Simulation	28.1	0.0	5.2
Other	10.6	0.0	5.9

Table 3 The projected resources for handling the 2008 data alone, assuming full simulation of 100% of the data rate.

It has been assumed that the Tier-1 facilities store on average one tenth of the RAW data and follow the processing model outlined in section 3. It is further assumed that the raw data are stored on tape/slow access media, with only a small subset remaining on disk to allow software development and testing[3]. The actual Tier-1 and Tier-2 capacity might be larger, and indeed shared with other experiments. Here only that part that is visible to and accessible by all ATLAS members, is taken into account, and would be credited in the ATLAS accounting.
The detailed use of resources in the various Tiers is given in more detail in Appendix C: Details of the Resource Usage in the Tiers.

5.2 Ramp-up and Resource Requirement Evolution

Clearly, the system described by Table 3 will not be constructed in its entirety by the start of data-taking in

[3] This defines one extreme position, assuming that almost no reprocessing will be possible at CERN. In reality, CERN should be able to provide a large resource for reprocessing during non-running periods.

2007/2008. From a cost point-of-view, the best way to purchase the required resources would be 'just in time'. However, the early period of data-taking will doubtless require much more reprocessing and less compact data representations than in the mature steady-state, as discussed in the section on Commissioning above. There is therefore a requirement for early installation of both CPU and disk capacity.

It is therefore proposed that by the end of 2006 a capacity sufficient to handle the data from first running under the conditions described in section 3.1, needs to be in place, with a similar ramping of the Tier-1 facilities. During 2007, an additional capacity required for 2008 should be bought. In 2008, an additional full-year of capacity should be bought, including the additional archive storage medium/tape required to cope with the growing dataset. This would lead to a capacity, installed by the start of 2008, capable of storing the 2007 and 2008 data as shown in Table 4; the table assumes that only 20% of the data rate is fully simulated.

	CPU(MSI2k)	Tape (PB)	Disk (PB)
CERN Tier-0	4.1	6.2	0.35
CERN AF	2.8	0.6	1.8
Sum of Tier-1's	26.5	10.1	15.5
Sum of Tier-2's	21.1	0.0	10.1
Total	54.5	16.9	27.8

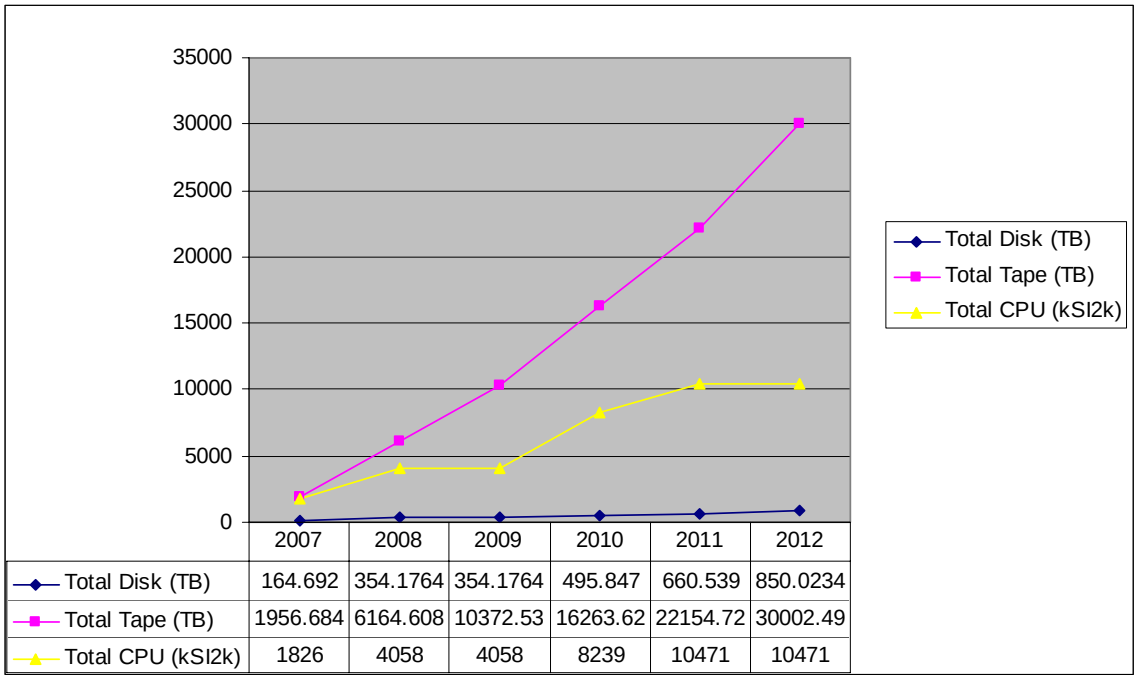
Table 4: The projected total resources required at the start of 2008 for the case when 20% of the data rate is fully simulated.

For the Tier-2s, a slightly later growth in capacity, following the integrated luminosity, is conceivable provided that the resource-hungry learning-phase is mainly consuming resources in Tiers 0 and 1. However, algorithmic improvements and calibration activity will require also considerable resources early in the project. As a consequence, we have assumed the same ramp-up for the Tier-2s as for the higher Tiers.

Once the initial system is built, there will for several years be a linear growth in the CPU required for processing, as the initial datasets will require reprocessing as algorithms and calibration techniques improve. In later years, subsets of useful data may be identified to be retained/reprocessed, and some data may be rendered obsolete. However, for the near future, the assumption of linear growth is reasonable. For storage, the situation is more complex. The requirement exceeds a linear growth if old processing versions are not to be overwritten. On the other hand, as the experiment matures, increases in compression and selectivity over the stored data may reduce the storage requirements.

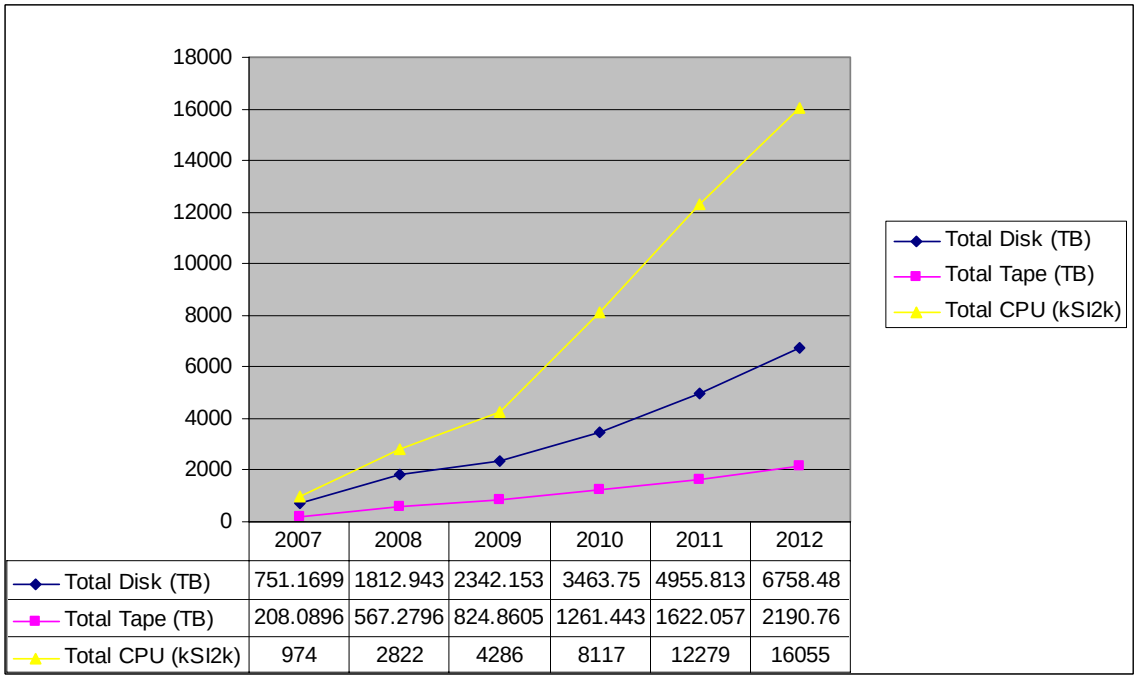
The projections do not include the replacement of resources, as this depends crucially on the history of the sites at the start of the project.

Heavy-Ion running is effectively a whole new experiment. The event sizes and processing times are uncertain, and it is not obvious that all Tier-facilities will participate. The requirements for heavy-ions running are not considered in the present exercise.



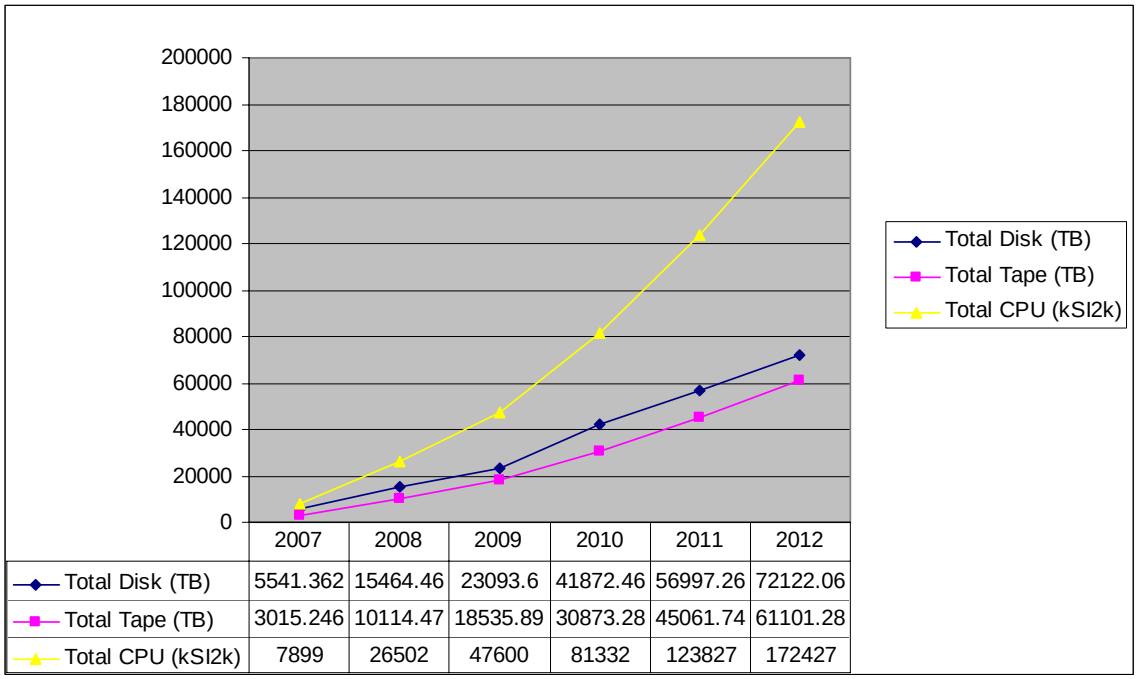
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	164.692	354.1764	354.1764	495.847	660.539	850.0234
Total Tape (TB)	1956.684	6164.608	10372.53	16263.62	22154.72	30002.49
Total CPU (kSI2k)	1826	4058	4058	8239	10471	10471

Figure 1: The projected growth in ATLAS Tier-0 resources with time.



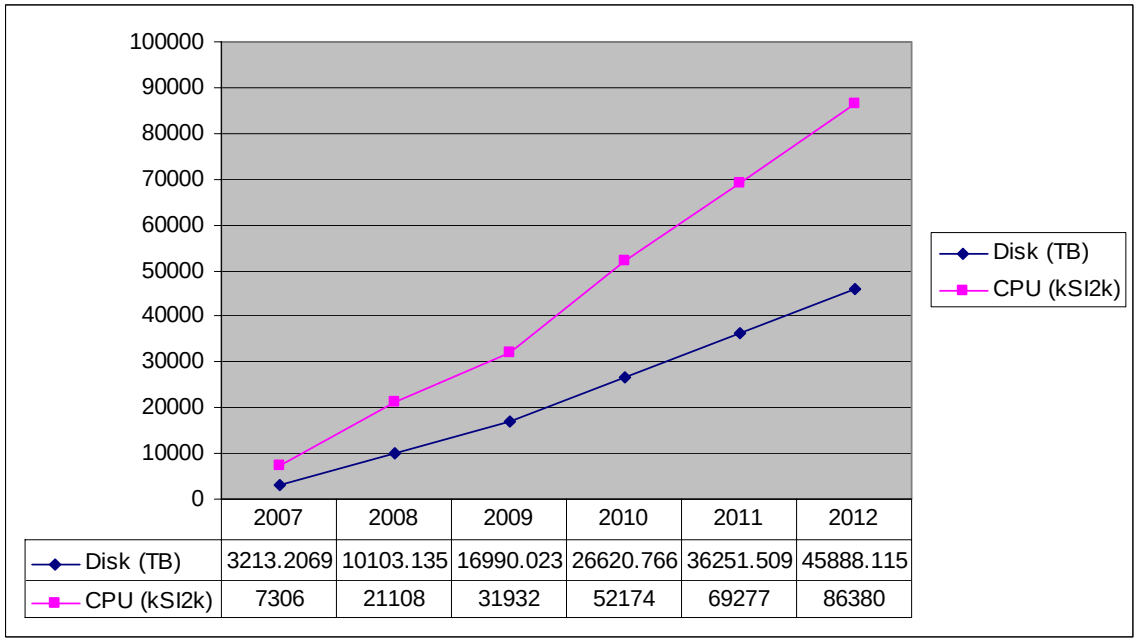
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	751.1699	1812.943	2342.153	3463.75	4955.813	6758.48
Total Tape (TB)	208.0896	567.2796	824.8605	1261.443	1622.057	2190.76
Total CPU (kSI2k)	974	2822	4286	8117	12279	16055

Figure 2: The projected growth in the ATLAS CERN Analysis Facility



	2007	2008	2009	2010	2011	2012
Total Disk (TB)	5541.362	15464.46	23093.6	41872.46	56997.26	72122.06
Total Tape (TB)	3015.246	10114.47	18535.89	30873.28	45061.74	61101.28
Total CPU (kSI2k)	7899	26502	47600	81332	123827	172427

Figure 3: The projected growth in the capacity of the combined ATLAS Tier-1 facilities.



	2007	2008	2009	2010	2011	2012
Disk (TB)	3213.2069	10103.135	16990.023	26620.766	36251.509	45888.115
CPU (kSI2k)	7306	21108	31932	52174	69277	86380

Figure 4: The projected growth of the combined ATLAS Tier-2 facilities. No repurchase effects are included.

6 Networking Requirements

The required bandwidth out of CERN is essentially set by the migration of data to the external Tier-1 facilities. There is a need for semi-immediate (real time) transfer of data offsite.

It should be noted that the future cost evolution for bandwidth depends on many external factors. Thus far, we have benefited from uncharged access to bandwidth provided under external funding. It is to be hoped that this continues, but as we are dependent on bandwidth, we are carrying a large potential source of financial risk.

The traffic is based on the planned flow of RAW, ESD and AOD data from CERN to the Tier-1; this is assumed to be direct point-to-point. The raw required bandwidth averaged over the data-taking period is ~71 MB/s for ATLAS production. This is expected to rise to 100 MB/s in 2010.

Source	**Inbound from CERN (MB/s)**	**Outbound to CERN (MB/s)**
RAW	30.4	
ESD Versions	20	1.41
AOD versions	18	0.28
TAG Versions	0.18	0
Group DPD		0.81
Total CERN/AverageTier-1	68.58	2.51

Table 5: The decomposition of the traffic between CERN and an average sized Tier-1 in 2008-2009.

An estimate has also been made of the ATLAS-only traffic between the Tier-1 facilities. This is comprised of the copying of 'back-up' samples from another Tier-1 for the various data processing versions. It also includes the transfer of analysis group derived datasets. The average raw data volume to be shifted is ~52 MB/s, comparable with the CERN-Tier-1 requirement. This is expected to rise to 63 MB/s in 2010. The actual traffic from other sources, such as the dataflow associated with job submission is not included, but believed to be relatively small. Naturally, factors for efficiency of bandwidth usage and contention need to be added.

Source	**Inbound from other Tier-1s (MB/s)**	**Outbound to other Tier-1s (MB/s)**
ESD Versions	12	14
AOD versions	20.8	2.08
TAG Versions	0.21	0.02
Group DPD		4
Total Tier-1 to Tier-1	33	19

Table 6: The decomposition of the Tier-1 to Tier-1 traffic for an average sized Tier-1 in 2008-2009.

The bandwidth required between a Tier-1 and Tier-2s will depend very much on the size of the Tier-2s concerned, but in terms of the files to made available for general use at each it will be typically below

10 MB/s. The traffic associated with user jobs is again not included, and may well be the dominant term in this case.

Remote Event Filter processing is not considered in this baseline model, but if it is to occur then a significant additional bandwidth upwards from 10 Gb/s will be needed between the ATLAS pit and the remote site, the actual capacity required depending on the Event Filter load displaced.

7 Summary

The baseline steady-state ATLAS Computing Model for the early years of data-taking has been outlined. It presumes a well-planned, organized and maintained Virtual Distributed Computing Facility. Such a facility will support ATLAS production, simulation and analysis. A considerable degree of central organisation will be required in the analysis activities given the volume of data to be processed in a world-wide system.

Consideration has also been given to the commissioning of the system with first-data. It is essential that a considerable degree of flexibility both in data formats and in resource usage be retained to ensure rapid analysis of the early data and a swift understanding and calibration of the detector.

This paper has also summarized the present understanding of the required resources. The overall costs are only be indicative due to large uncertainties in most of the numbers used to perform these calculations, and pricing regimes differ considerably with location.

The final system will be described in the Computing TDR. Our continuing Data and Service Challenge activities are designed to refine the understanding of the computing model, and also to commission the required tools well in advance of the first data-taking.

Appendix A: On Streaming

Whether events should be streamed or routed by physics channel or trigger, and at which processing stage, is a complex question, one to which different high-energy physics experiments have arrived at different answers. Some have argued that if we stream at all, we should stream as early in our processing as possible; others have argued that perhaps we should not stream at all. There are several intermediate possibilities as well.

In the baseline model, four streams exist at the input to the Tier-0: the primary physics stream containing all physics events; an express stream containing a subset of events; the calibration stream; and the diagnostic stream. The primary physics stream and calibration stream will be distributed democratically world-wide. A similar distribution may occur for the express stream. The diagnostic stream would remain at CERN.

If, as is our aim, the overwhelming majority of ATLAS analyses may be accomplished using ESD and its derived data products (ideally, with AOD and beyond), then the primary use for RAW data are as input to re-reconstruction. In this case, it is likely that entire runs would be re-reconstructed — not just events of a particular type — so that having streamed RAW data by type would not provide any advantage. Because streaming RAW data out of the Event Filter by type can lead to significant file size imbalance (due to widely varying rates), and because streaming by type poses additional challenges to the ATLAS democratic data distribution model (Which event types are reconstructed first? Which event types are sent to my Tier-1?), we opt herein for a model in which RAW data are not streamed by type. It should be noted, however, that if the Event Filter is capable of routing physics events both to physics and to calibration streams, then the capability to support streaming RAW events will likely be available should ATLAS choose to amend this strategy.

ATLAS offline database software is agnostic about processing stages, and can as easily write multiple ESD streams as multiple AOD streams. The choice is a matter of policy; amending the choice would not require substantial changes to core software infrastructure. The choice we propose is to stream AOD by type, but not ESD. The reason for not streaming ESD is to retain the one-to-one “file in, file out” correspondence to RAW data, which is itself not streamed by type, simplifying provenance tracking, with clear assignment of primary responsibility for hosting the corresponding events to a specific Tier-1 site. A RAW event file contains whatever events came out of an Event Filter SFO in a given interval of time, independent of type, and a corresponding ESD file contains reconstructed data for exactly those events.

Appendix B: Data Challenges and Tests of the Computing Model

B.1 Data Challenge 2

The principal purpose of ATLAS Data Challenge 2 (DC2) is to deploy and test a prototype of the ATLAS computing model, both as input to the ATLAS Computing Technical Design Report and as an opportunity to identify possible shortcomings in ATLAS plans in time for adjustments before data-taking begins. Event data flow from the Event Filter through reconstruction to worldwide data distribution is central to any such exercise.

An additional test that would be very useful cannot easily be done by ATLAS alone; that is, the effect of network traffic from the other LHC experiments out of CERN at the same time as our slice tests. This might be most easily achieved if, for example, ALICE is running tests of the off-site transport of their data. If this could be co-ordinated with one or more of our slice tests, this would be a good simulation of the competition between ATLAS and the likely traffic from CMS and LHCb, or indeed the network contention during heavy-ion running.

B.2 Data Challenge 3

Data Challenge 3 (DC3) will be run between the end of 2005 and the first half of 2006. DC3 will be the last occasion to test the ATLAS Computing Model on a large scale before the start of data taking, therefore is has to be seen as the dress-rehearsal of the system. Several aspects of the Computing Model, although supported by the current software, were not included in the DC2 operation, due to the concurrent Combined Test Beam (CTB) operation in 2004.In particular, the following items, on which experience has been gathered in the CTB, will have to be integrated after the end of DC2 into the global processing suite:

- interactive analysis framework
- online operation of the Athena framework;
- monitoring of data flow and algorithmic software;
- trigger reconstruction and selection algorithms (both for level 2 and the event filter);
- detector misalignment in the simulation;
- detector alignment and calibration procedures in the main data flow;
- conditions database;
- detector inefficiency.

This integration work will take place between the end of 2004 and summer 2005: the aim is to have a system as complete as possible for DC3.

B.3 Commissioning runs

It is envisaged in the global ATLAS schedule to start taking data with cosmic ray triggers as soon as enough components of the ATLAS detector will be in place in the ATLAS pit. According to the current estimates this operation will start some time in 2005. By that time, the offline computing infrastructure will have to be ready to absorb and process the data in real time. These data, of increasing complexity and rate, will be used as further, more realistic, tests of the Computing Model before the start of LHC operations in 2007.

Appendix C: Details of the Resource Usage in the Tiers

This section briefly details the distribution of the resources in each Tier as a function of data format and activity.
Table 7 to Table 14 give the decomposition of the storage and CPU required in the various Tiers for one full year of data-taking.

	Disk (TB)	Shelf.Tape (TB)
Raw	0	3040
ESD	0	1000
Buffer	114	0
Calibration	240	168
Total	**354**	**4208**

Table 7: The storage requirements at the CERN Tier-0 as a function of data format**.**

	CPU (MSI2k)
Reconstruction	3.5
Calibration	0.5
Analysis	0.0
Total	**4.1**

Table 8: The CPU requirements at the CERN Tier-0 as a function of activity.

	Disk (TB)	Tape (TB)
Raw	241	0
ESD (current)	229	0
ESD (previous)	0	18
AOD (current)	257	0
AOD (previous)	0	4
TAG (current)	3	0
TAG (previous)	0	2
MC ESD (current)	286	0
MC ESD (previous)	0	4
MC AOD (current)	57	0
MC AOD (previous)	0	40
MC Tag (current)	0.6	0
MC Tag (previous)	0	0.4
Calibration	240	168
User Data	303	212
Total	**1615**	**448**

Table 9: The storage requirements at the CERN Analysis Facility as a function of data format**.**

	CPU (MSI2k)
Reconstruction	0.2
Calibration	0.5
Analysis	1.5
Total	**2.2**

Table 10: The CPU requirements at the CERN Analysis Facility as a function of activity.

	Disk (TB)	Tape (TB)
Raw	430	3040
ESD (current)	2570	900
ESD (previous)	1290	900
AOD	2830	360
TAG	30	0
Calibration	2400	0
MC RAW	0	800
MC ESD (current)	570	200
MC ESD (previous)	290	200
AOD Simulation	630	80
Tag Simulation	10	0
Group User Data	1260	0
Total	**12300**	**6480**

Table 11: The storage requirements at the Tier-1 facilities as a function of data format**.**

	CPU (MSI2k)
Reconstruction	4.5
Calibration	0.5
Analysis	12.9
Total	**18.0**

Table 12: The CPU requirements at the Tier-1 facilities as a function of activity**.**

	Disk (TB)
Raw	43.4
General ESD (curr.)	385.7
General ESD (prev..)	0.0
AOD	2571.4
TAG	77.1
RAW Sim	0.0
ESD Sim (curr.)	171.4
ESD Sim (prev.)	0.0
AOD Sim	571.4
Tag Sim	17.1
User Group	1257.1
User Data	1815.3
Total	**6910.1**

Table 13: The storage requirements at the Tier-2 facilities as a function of data format**.**

	CPU (MSI2k)
Reconstruction	2.0
Simulation	5.4
Analysis	8.8
Total	**16.2**

Table 14: The CPU requirements at the Tier-2 facilities as a function of activity.

Appendix D: On Costs

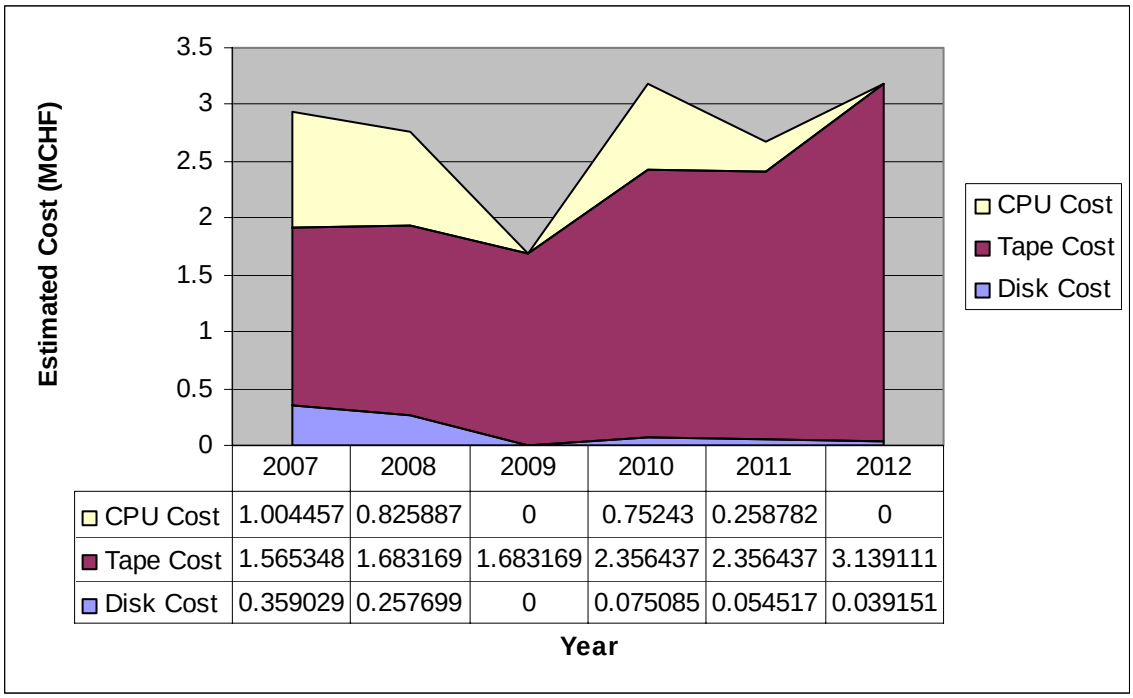
The detailed costs of each site depend both on the history of the resource purchases (which will determine the replacement purchase schedule and costs), and on the local market and purchasing rules. Any predicted cost profile can at best be indicative.

The following projections are based on the LCG cost estimations [8] for the CERN Tier-0. They assume all equipment required for 2007 is bought in 2007. The estimates were for the mid-year cost, but we are making the optimistic assumption that they apply at the start of year. No repurchasing is included.

For the years after 2010, a Moore's Law extrapolation is applied to the disk and CPU costs, while the tape costs are fixed until 2012.

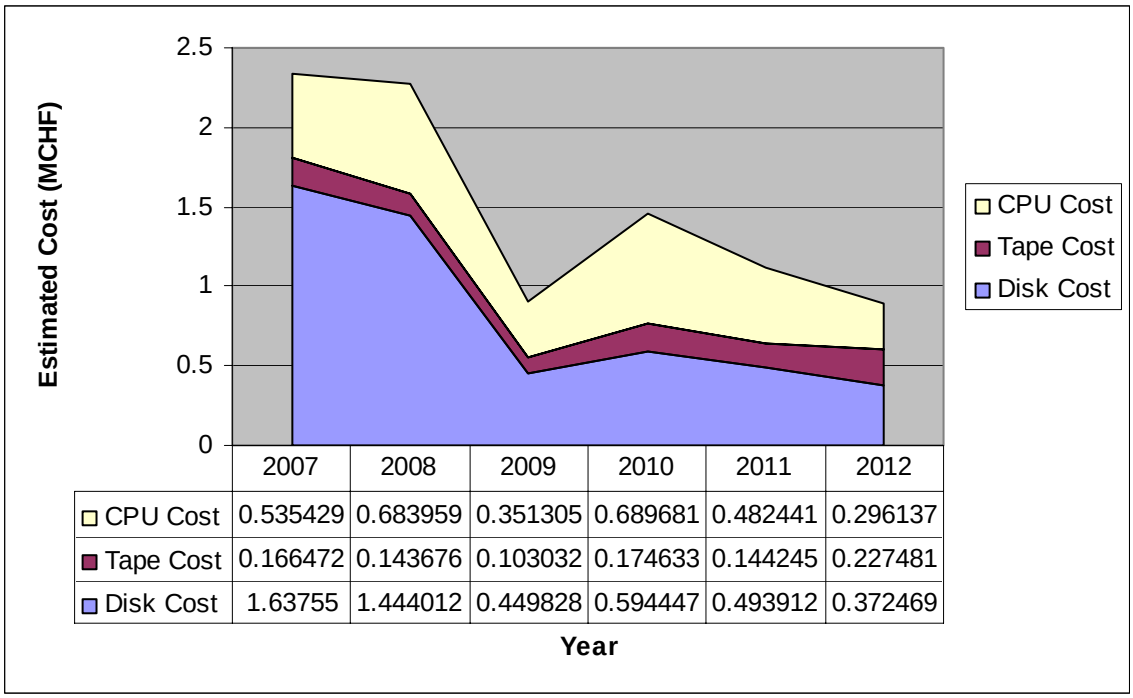
No explicit staffing costs are included in the projections.

The resultant purchase profile for the ATLAS Tier-0 is given in Figure 5 and that of the ATLAS component of the CERN Analysis facility in Figure 6. For the cost profile for the combined ATLAS Tier-1 Facilities is given in Figure 7 and that of the combined Tier-2s in Figure 8.



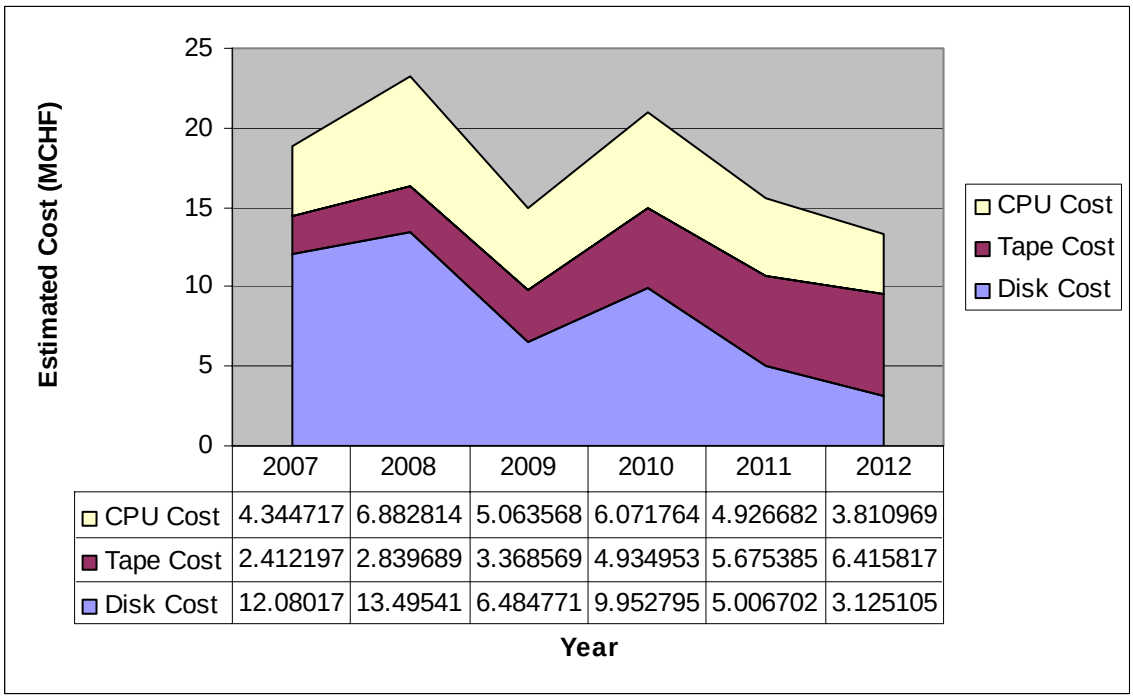
	2007	2008	2009	2010	2011	2012
CPU Cost	1.004457	0.825887	0	0.75243	0.258782	0
Tape Cost	1.565348	1.683169	1.683169	2.356437	2.356437	3.139111
Disk Cost	0.359029	0.257699	0	0.075085	0.054517	0.039151

Figure 5: The projected cost profile for the ATLAS component of the CERN Tier-0.



	2007	2008	2009	2010	2011	2012
CPU Cost	0.535429	0.683959	0.351305	0.689681	0.482441	0.296137
Tape Cost	0.166472	0.143676	0.103032	0.174633	0.144245	0.227481
Disk Cost	1.63755	1.444012	0.449828	0.594447	0.493912	0.372469

Figure 6: The projected cost profile for the ATLAS component of the CERN Analysis Facility.



	2007	2008	2009	2010	2011	2012
CPU Cost	4.344717	6.882814	5.063568	6.071764	4.926682	3.810969
Tape Cost	2.412197	2.839689	3.368569	4.934953	5.675385	6.415817
Disk Cost	12.08017	13.49541	6.484771	9.952795	5.006702	3.125105

Figure 7: The projected cost profile for the combined ATLAS Tier-1 facilities.

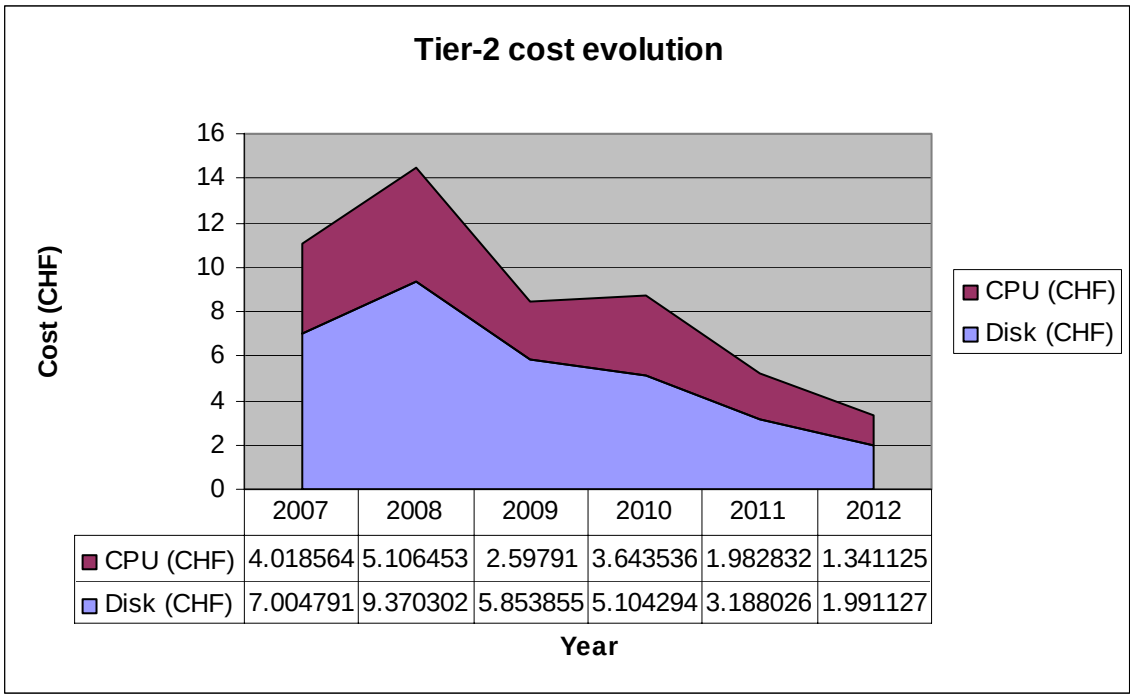


Figure 8: The projected cost profile for the combined ATLAS Tier-2 facilities.

References

[1] MONARC Phase 2 Report, M. Aderholz *et al*., CERN-LCB-2000-001, March 2000.

[2] Report of the Steering Group of the LHC Computing Review, H. Hoffmann *et al*., CERN-LHCC-2001-004, March 2001.

[3] ATLAS High-Level Trigger, Data Acquisition and Controls Technical Design Report, ATLAS Collaboration, CERN-LHCC-2003-022, June 2003.

[4] ATLAS Detector and Physics Performance Technical Design Report, ATLAS Collaboration, CERN-LHCC-1999-14 and CERN-LHCC-1999-15, June 1999.

[5] Report of the ATLAS Online – Tier-0 Task Force, D.G. Charlton *et al*., ATL-GEN-2004-002, December 2004.

[6] Report of the ATLAS AOD/ESD Definition Task Force, D. Rousseau *et al*., ATL-SOFT-2004-006, December 2004.

[7] Heavy Ion Physics with the ATLAS Detector, ATLAS Collaboration, CERN-LHCC-2004-009, March 2004.

[8] LCG Note, B. Panzer-Steindel, CERN-LCG-PEB-2004-20; LCG Note, B. Panzer-Steindel, CERN-LCG-PEB-2004-21.