U.S. ATLAS Tier 3 Task Force

March 27, 2009

Raymond Brock[1*], Doug Benjamin[2**], Gustaaf Brooijmans[3], Sergei Chekanov[4**], Jim Cochran[5], Michael Ernst[6], Amir Farbin[7], Marco Mambelli[8**], Bruce Mellado[9], Mark Neubauer[10], Flera Rizatdinova[11], Paul Tipton[12], and Gordon Watts[13]

[1]*Michigan State University,* [2] *Duke University,* [3]*Columbia University,* [4]*Argonne National Laboratory,* [5]*Iowa State University,* [6]*Brookhaven National Laboratory,* [7]*University of Texas at Arlington,* [8]*University of Chicago,* [9]*University of Wisconsin,* [10]*University of Illinois,* [11]*Oklahoma State University,* [12]*Yale University,* [13]*University of Washington*

[*]*chair,* [**]*expert member*

Contents

1 Introduction 4

2 Executive Summary 8
- 2.1 Observations . . . 10
- 2.2 Recommendations . . . 11
  - 2.2.1 A U.S. Strategy for Tier 3 Computing . . . 12
  - 2.2.2 Some Technical Jobs to Do . . . 13
  - 2.2.3 Forming a Partnership with the Universities . . . 14
  - 2.2.4 Policies and Numbers . . . 14
- 2.3 A Future Narrative . . . 15
  - 2.3.1 Group A . . . 15
- 2.4 Conclusion . . . 20

3 Definitions and Assumptions 21
- 3.1 The ATLAS Event Data Model . . . 21
  - 3.1.1 ATLAS Tiered Computing Centers . . . 21
  - 3.1.2 ATLAS Data Formats . . . 24
  - 3.1.3 Derived Physics Data . . . 29
  - 3.1.4 Responsibilities of the Tier 1 and Tier 2 Centers . . . 31
- 3.2 Analysis Model . . . 32

4 The Use Cases 36
- 4.1 Steady State Data Distribution. . . . 38
- 4.2 Dataset Creation. . . . 38
- 4.3 Monte Carlo Production. . . . 41
- 4.4 Chaotic Data Analysis. . . . 43
- 4.5 Chaotic User Analysis Use Cases. . . . 43
  - 4.5.1 Intensive Computing Use Cases . . . 45
  - 4.5.2 Use Cases: Conclusion . . . 47

5 The Tevatron Experience 50
- 5.1 Desconstruction of a DØ Analysis . . . 50
  - 5.1.1 A Story: DØ Infrastructure Evolution . . . 51
  - 5.1.2 The Story Continues: DØ Data Formats . . . 53
  - 5.1.3 A Happy Ending: A DØ Analysis . . . 55
- 5.2 A CDF Analysis . . . 58
  - 5.2.1 Evolution of CDF Analysis Computing . . . 60

6 Modeling **61**
6.1 The Calculation 61
6.2 Example Calculation: Monte Carlo Production 63
6.3 Input Parameters 64
6.4 Estimating Required Monte Carlo Production Resources 66
6.5 Modeling Analysis 67
6.6 Conclusions 71

7 Tier 3 Task Force Recommendations **72**
7.1 Potential U.S. ATLAS Tier 3 Strategies 74
7.1.1 A Flexible Tier 3 U.S. ATLAS System: Four Kinds of Centers 74
7.1.2 Tier 3 Architectures 76
7.2 Revisiting the Use Cases 79
7.2.1 Distributed Data Management and Compute Elements 79
7.2.2 Value-Added From a T3gs System 80
Data Format Production 81
Monte Carlo Production 81
Summary 83
7.2.3 Value-Added From a T3g System 83
Focused Signal-Background Analyses 84
Conclusions 85
7.2.4 Technical Recommendations 86
7.3 Tier 3 Support Strategies 88
7.3.1 Summary of the T3gs and T3g Idea 89
7.4 Participatory U.S. ATLAS Cluster Program 89

8 Conclusions **91**
8.1 An Exciting Particle Physics Mission is Guaranteed 92
8.1.1 The University Community is Key 92

9 References **95**

Appendices **97**

A Charge to the Task Force **97**

B Original Whitepaper **98**
B.1 Reaction to the White Paper 103

C ATLAS Glossary **104**

D Typical Hardware and SI2k Specifications **105**

E Characterization of Tier 3 Sites **107**
E.1 T3gs 107
E.1.1 The University of Illinois T3gs Project 109
The Network 111
The Classes 112
E.2 T3g 115
E.2.1 The ANL T3g Project 118
E.2.2 The Duke T3g Project 120

F Survey of non-U.S. ATLAS Tier 3 Strategies **122**
F.1 United Kingdom 122
F.2 Canada 122
F.3 Netherlands 123
F.4 Spain 123
F.5 Germany 124
F.6 Italy 124
F.7 France 124

G Survey of CMS/ALICE Tier 3 Strategies **125**
G.1 CMS 125
G.2 ALICE 126

H Survey of U.S. University ATLAS Computing **127**

1 Introduction

Everything about the LHC is huge. In addition to sheer physical size, ATLAS will produce a torrent of data so vast as to flood any single computer system. So, consistent with the international nature of High Energy Physics, these data must be distributed around the world for primary reconstruction and for the multiple—and repeated—stages of processing necessary to decrease its overall bulk to a reasonable size.

While this reduction effort will be significant, it is relatively straightforward, compared to the extraction of scientific results: physics analysis never goes as planned. Mistakes are made. Detector calibrations and corrections challenge the cleverest analysts. False starts and dead ends accompany good ideas and brilliant breakthroughs. Collaborations and individuals are stimulated by the potential for discovery and motivated by intense competition. As a result, pushing technical limits and stretching policy boundaries have both been a part of life during large-scale physics analyses. Experiment and laboratory administrators must strike a delicate balance between not discouraging fresh—even anarchical—approaches to computing, while not invalidating carefully reasoned planning.

The scale of data and numbers of people involved in the LHC significantly increases the stress on processing, storage, network capabilities, *and human organization* over those faced by the Tevatron experiments. Even in their mature years, predicting and implementing workable long term production and analysis strategies for CDF and DØ were very difficult. The need to react to jumps in instantaneous and integrated luminosity, maturing and new analysis techniques, and repeated revolutions in technology was often humbling. Despite impressive planning, experience within ATLAS computing will similarly confront surprises and the need to react quickly to both setbacks and opportunities. This reaction can either be difficult—because of rigid structures—or efficient—because of designed-in adaptability.

Observation 1 *Challenges to efficient LHC physics analysis are likely to be greater than imagined and so "flexible" and "nimble" should continue to be the guiding principles in the design of computing infrastructure.*

The starting point of this data-deluge is a 200 Hz bytestream of 1.6 MB raw data records flowing from the High Level Trigger (HLT) — almost 30 TB per day. The destination is a reduced dataset on a physicist's desktop somewhere in the ATLAS universe which is suitable for productive analysis. Ultimately, such data-reduction schemes have to satisfy a human-scale question such as: "How long are you willing to wait for a full analysis pass through your dataset?"

A quick calculation: on most disk systems, the fastest evaluation of a `ROOTtuple` is the I/O limitation of about 10 MBps. If we presume a human impatience scale of about an hour, just reading through a dataset and plotting should fit that duration. As a round number, if we presume a year's accumulation of a rare signal plus background amounts to only a million events, then for this quick example, that final data format has to be about 40 kB/event— raw records need to be squeezed into packages 2% of their original size, and the total event sample from HLT to desktop has to be reduced by a factor of 300,000 without loss of crucial information.

How this is envisioned to take place has been described many times in memos and presentations. But, incredibly, it's still an unsettled situation when it comes to the human factor, at the end of the chain—the campus-sized analysis, where the actual Science originates. In point of fact, the simple example above is unrealistic: a million event sample as an object of analysis is undersized. So, in most cases, simple "desktop" analyses will not be so simple and the dataset sizes are likely to be many TB. The human scale of approximately an hour is still about right, so the number of processors per node and multiple I/O threads will be significant. There is an experienced-based obsevation, however, which is borne out in experiment after experiment which fights against this overall data bulk:

Observation 2 *Physicists often reduce dataset sizes in order to bring as much data, as near to their desktop as is feasible, as often as is required.*

This effort to bring data close to the analyzer is understandable as the best way to control the inevitable, unpredictable inefficiencies in dealing with remote batch systems serving many customers. Starting, stopping, restarting, lossy dataset transfer, and remote monitoring are all important real-time needs which are best accomplished with local control. So, that's the question: what tasks can be done most efficiently and economically on university campuses, and what tasks must be relegated to "the grid" and remote facilities.

This document is an attempt to characterize the particular, important, last link in the chain of "tiered" computing from the ATLAS Computing Model, namely the Tier 3 level which has typically been presumed to be a university-based—and university-owned— system for local users. Recent evolution of the ATLAS Analysis Model and the Event Data Model have significantly changed the relationships among the three U.S.-based computing tiers and we found it meaningless to describe the Tier 3 experience without adopting a model for the Tier 2 responsibilities. In trying to understand the needs and the desires of university analyzers, we are motivated by **Observation 1** and guided by **Observation 2**.

Because this is a subject which is likely to be of interest outside of the expert ATLAS community, there has been a concerted effort to be complete in preparing this document and to draw into one place numbers, policies, and procedures which are currently scattered in presentations, twikis, and memos. We anticipate that the readership will include people not connected directly with ATLAS and perhaps unfamiliar with jargon and specifics and so we've also included a glossary (Appendix C defining and characterizing ATLAS-specific terms and labels. In fact, this information was so dispersed and scattered through websites, talks on Indico, in memos and reports, that we make our first recommendation[1]

Recommendation 9: ATLAS computing and analysis policies, existing resource amounts, targeted resource quantities, data format targets, times for data reduction, etc.: basically all parameters and rules should be in one place. A policy should be considered "official" only when updated at a single twiki page. One repository should define official reality and should be updated when that reality changes.

The Executive Summary, Section 2, enumerates all of the Observations and Recommendations and is structured to stand alone. Reading it will summarize the conclusions and through a narrative, a sense of how this might play out in the future as a fictitious group makes use of the Task Force results.

The bulk of this large report provides the justification for the recommendations and many details which follow in five parts: **Section 3: Definitions and Assumptions**, **Section 4: Use Cases**, **Section 5: The Tevatron Experience**, **Section 6: Modeling**, and **Section 7: Recommendations**. Appendices present results of other, similar systems in and outside of ATLAS as well as examples of U.S. ATLAS Tier 3 systems, user survey results, and other technical information.

We believe that there are compelling quantitative reasons to design a set of computing "Tier 3" clusters for the use of U.S. ATLAS university groups. No less important than the quantitative reasoning for this conclusion are intangible, programmatic reasons why we believe this to be the case. We

[1] Throughout the text, the Recommendations are numbered according to their relative importance, which is the order in which they appear in the Executive Summary, Section 2. By contrast, the Observations are numbered in the order in which they appear in the text.

will make both arguments below.

Finally, a note about dates used in this report. There are many lists of anticipated luminosities, numbers of cpus, storage-commitments, etc. which have all been predicated on a 2008 startup of LHC collisions and so are all out of date. We presume that they are out of date by +1 year for our purposes. For example, current obligations for "2010" we presume will be operational for actual-2011. We have taken our charge (see Section A in the Appendix) to cover a period in the future where ATLAS data-taking and analysis are at a relatively stable stage and we have defined that to be a year in which 10fb^{-1} of physics data are taken. Another Task Force is considering the situation appropriate to the first year or so of data-taking where conditions will be rapidly changing and actual physics analysis will be less important than calibration, alignment, bug-fixing, and disaster-detection. The first time this comes up in the text, we will remind the reader that "2010" really is meant to imply roughly "2011."

2 Executive Summary

This report summarizes the investigation of the Tier 3 Task Force convened by U.S. ATLAS management during the summer of 2008. The charge is presented in Appendix A. Basically, it asked for recommendations in three areas:

1. Use Cases
   (a) Typical workflows for physicists analyzing ATLAS data from their home institutions should be enumerated. This needs to be inclusive, but not in excruciating detailed. It should be defined from within the ATLAS computing/analysis models, the existing sets of Tier 2 centers, and their expected evolutions.
   These are enumerated in Section 4.
   (b) If there are particular requirements in early running, related to detector commissioning and/or special low-luminosity considerations, this should be noted.
   See below.
   (c) If particular ATLAS institutions have subsystem responsibilities not covered by the existing Tier 1/2 deployment, this should be noted. Is the previous whitepaper relevant?
   We believe that, while there are subsystems (e.g., the Muon Project at the University of Michigan, within the AGL-T2 center) which do have a special relationship with a Tier 2, none have emerged since deployment. The previous whitepaper is addressed in Appendix B.
2. Generic Tier 3 Configurations.
   (a) Some Tier 3's may be very significant because of special infrastructure availabilities and some Tier 3's maybe relatively modest. Is there only 1 kind of Tier 3 center, or are their possible functional distinctions which might characterize roles for some Tier 3's that might not be necessary for others? Description of "classes" of Tier 3 centers, if relevant, should be made.
   This is addressed in Section 7.1.
   (b) Support needs and suggestions for possible support models should be considered.
   This is addressed in Section 7.3.
3. Funding.

(a) This is not part of the US ATLAS Operations budget, so funding must come out of the institutes through core funding or local sources. We would like to make it easier for institutes to secure funding for ATLAS computing–this can only happen if it fits in the DOE and NSF budgets (precedent: the amount of funding groups got for computing equipment in Tevatron experiments) and it must fit in the overall US ATLAS model.

(b) For the latter, we have to make the case that the existing Tier 1/2 centers are not enough.
This is addressed in Sections 5 and 6.

(c) Perhaps a recommendation can be justified for an estimated amount needed for a viable Tier 3 cluster.
This is addressed in Appendix E

Subsequent to the formation of this task force, a separate group was charged with evaluating the resource needs for the first year or so of data-taking. Consequently, we ignored 1.(b.) above and focused our attention on some future period in which scientific-quality data are being produced. We arbitrarily chose the first 10fb^{-1} year as the benchmark.

It is important to note that the Computing Model has been somewhat fluid. This is especially true in the responsibilities asked of the Tier 2 centers (in the U.S.). While this is hinted at in the text, an example of this is in the data-caching responsibilities. When the U.S. Tier 2 centers were established, the "Derived Physics Data" (DPD) formats had not been integrated into the ATLAS analysis model and so where to store what formats and how much of each format is to be stored at Tier 2s has not been finalized. This same situation holds with respect to the production of some of the lesser formats themselves. So, how to integrate Tier 3 analysis centers into an overall fabric of still-evolving Tier 2 centers is a moving target. We would note that while some of this will naturally evolve, the time for making decisions on some of these matters is past due.

Through our investigation we summarize our conclusions in two formats: Observations and Recommendations. "Observations" are meant to be modest alerts to circumstances, ideas, concerns, and possibilities in order to motivate discussion among the U.S. ATLAS leadership.

So, the following list our Observations in the order in which they appear in the text[2]:

[2] Observations are numbered in the order in which they appear in the text.

2.1 Observations

Observation 1 *Challenges to efficient LHC physics analysis are likely to be greater than imagined and so "flexible" and "nimble" should continue to be the guiding principles in the design of computing infrastructure.*

(page 4)

Observation 2 *Physicists often reduce dataset sizes in order to bring as much data, as near to their desktop as is feasible, as often as is required.*

(page 5)

Observation 3 *The entire DPD production chain (D^1PD , D^2PD , and D^3PD) is to be an essential feature of the analysis sequence. And yet the lack of experience in producing DPDs through the whole chain is difficult to understand. Reliable timings are unavailable, for example. Storing both AODs and D1PDs at Tier 2s seems redundant, but there is yet no guidance on which, how much, when, how the AOD format storage and the DPD storage and production is to be arranged. The ultimate storage load on the Tier 2s is therefore unevaluated (see below). (Note, the performance DPD—dDPD—will be the major data format in early running and is not a part of the concern here.)*

(page 35)

Observation 4 *The Tier 2 systems' responsibilities are tremendously significant. Should we discover an underestimate in CPU, storage, or network needs of ATLAS as a whole, the analysis needs of U.S. university physics community will be adversely affected.*

(page 49)

Observation 5 *Is there any reason to think that the first 20 years of the ATLAS computing experience will be any less astonishing? Is it wise to design tightly to current expectations, as if the future will be a continuous extrapolation of the present? If history is at all a reliable guide, it argues for the most flexible, most modular, and least rigidly structured systems consistent with 2008 technology and budgets.*

(page 51)

Observation 6 *Physics analysis moves fast, at a rate which is often more rapid than can be tolerated by a rigid computing structure or system management. Analyzers will sometimes take matters into their own hands when a bureaucracy is perceived to be in the way.*

(page 55)

Observation 7 *Full-scale, precision analyses will be a huge load on the Tier 2 structure from the perspective of computation and file-access. Monitoring and resubmitting failed jobs will surely continue to be a serious complication for analyzers. If history is a guide, current predictions of how this maps to the ATLAS analysis future are sure to be underestimated.*

(page 58)

Observation 8 *Should ATLAS-wide production needs be more than the Tier 2 centers can provide, the only flexibility is to "eat" away at the 50% of the Tier 2 resources nominally reserved for U.S. user analysis. One has to ask what the likelihood is of such an outcome and whether U.S. ATLAS analysis could survive the effects of such a result.*

(page 74)

Observation 9 *It may be possible for university groups to confederate with one another, from one campus to another, or even across department and disciplinary boundaries within a single campus. For some Tier 3 tasks, such arrangements may work well. We know of no functioning arrangements at the time of this writing, but we believe that efforts are underway to create them on a few campuses.*

(page 79)

Observation 10 *The technical (and social) challenges are enormous and in order for the LHC Mission to succeed—and it must succeed—the U.S. community has to be fully equipped and fully staffed in order to meet those challenges.*

(page 92)

2.2 Recommendations

In addition to our Observations, we make several Recommendations pursuant to the Charge. The list of Recommendations—in rank order of their importance—are below. The numbering in the text corresponds to the rank ordering here.

Apart from **Recommendation 9** above, all of the Task Force recommendations appear in Section 7 beginning on page 72[3].

[3]Throughout the text, the Recommendations are numbered according to their relative importance, which is the order in which they appear in the Executive Summary, Section 2.

2.2.1 A U.S. Strategy for Tier 3 Computing

The story told in Section 5 (page 50) plus the modeling described in Section 6 (page 61) suggest to us that for the U.S., the ATLAS Computing Model is possibly too rigid— that relying on the Tier 2 cloud alone might reduce U.S. analysis capabilities. In order to add flexibility and a degree of nimbleness required in order to react to surprises, we recommend the characterization of four kinds of Tier 3 systems for the U.S.

We do not expect that these systems should be created overnight. Rather, we propose a characterization of each and a terminology so that each group, in negotiation with its home institution, U.S. ATLAS management, and their individual funding agencies might target the kind of computing systems they anticipate will best fit their group's analysis plans and so that all of the stakeholders will understand the implications of each choice.

Accordingly, **Recommendations 1-5** are a group which, when taken together, provide the minimal structure from which Tier 3 systems could be deployed over the next few years.

Recommendation 1: With past history as a guide and with prudent concern for the challenge and uncertainties of ATLAS analysis, the *structured* U.S. ATLAS computing infrastructure should be deeper than the Tier 2 centers. A flexible and nimble infrastructure would include strategically extending some data production, Monte Carlo simulation, and analysis into the U.S. ATLAS Tier 3 sector. (page 74)

Recommendation 2: The strategy for building a flexible U.S. ATLAS Tier 3 system should be built around a mix of 4 possible Tier 3 architectures: T3gs, T3g, T3w, and T3af. Each is based on a separate architecture and each would correspond to a group's infrastructure capabilities. Each leverages specific analysis advantages and/or potential ATLAS-wide failover recovery. They are specifically defined in Section 7.1.2. (page 76)

Recommendation 2 is the heart of the Task Force conclusions. Much more can be read about the details in Section 7.1.2. Table 21 on page 76 summarizes and contrasts the different Tier 3 architectures.

Recommendation 3: In order to support a Tier 3 subscription service, without a significant support load or the need to expose itself to the ATLAS data catalog, a particular DQ2 relationship must be established with a named Tier 2 center, or some site which can support the DQ2 site services on its behalf. This breaks the "ubiquity" of Tier 2s — here, a particular Tier 3 would have a particular relationship with a named Tier 2. This dual-capability (limited exposure of a site's file catalog and a subscription-like functionality) has been colloquially referred to as "outsourcing" DQ2 site services. (page 86)

Recommendation 4: U.S. ATLAS should establish a U.S. ATLAS Tier 3 Professional, a system administration staff position tasked to 1) assist in person the creation of any Tier 3 system; 2) act as a named on-call resource for local administrators; and 3) to lead and moderate an active, mutually supportive user group. (page 89)

Recommendation 5: In order to qualify for the above U.S. ATLAS Tier 3 support, U.S. ATLAS Tier 3 institutions must agree to 1) supply a named individual responsible on campus for their system and 2) adhere to a minimal set of software and hardware requirements as determined by the U.S. ATLAS Tier 3 Professional. (page 89)

2.2.2 Some Technical Jobs to Do

The unique nature of Tier 3s is that they are private. Funds will come hard and groups will maintain policy control over their systems. While the T3gs systems might occasionally be deployed on behalf of ATLAS as a whole, it would be a group's decision when and how long to make that contribution. This means that, in addition to the modifications to DQ2 described in Recommendation 3, more control over local `pAthena` functionality is required.

Recommendation 6: We recommend that the recent addition of `pAthena` local control-functionality be maintained, and possibly extended to allow for more convenient control and access/monitoring of the Tier 3 site configuration by local administrators. (page 87)

The ability to reliably transfer large datasets to and from Tier 3s is essential. We have tried to identify a target for bandwidth and suggest that sites be brought to this standard along with their individual evolution to their target Tier 3 kind. A big job would be to guarantee the target bandwidth from all Tier 3s to the entire Tier 2 cloud. A more reasonable approach might be to take advantage of regional and resource infrastructure which might make targeting particular Tier 3-Tier 2 connectivity at the target bandwidth.

Recommendation 7: Sustained bandwidth of approximately 20MBps is probably required for moving TB sized files between Tier 2 and Tier 3 locations and it should be the goal that every campus or lab group establish such capability within a few years. This requires a high level of cooperation and planning among U.S. ATLAS computing, national network administrators, and campus administrators. Note: it might be useful and prudent to tune bandwidth between *particular* Tier 3 locations and *particular* Tier 2 centers rather than to set a national standard which might be difficult to meet. Note that the Resource Allocation Committee will have authority over the large-scale movement of data and any large scale caching of Tier 3 generated files into the Tier 1 or Tier 2 clouds. (page 88)

2.2.3 Forming a Partnership with the Universities

One reason to not just put all U.S. ATLAS Tier 3 funds into one or more national labs is that U.S. ATLAS physicists will benefit by having an identified, hardware presence on their campuses. Another reason is that with non-recurring contributions from universities to their local Tier 3 sites might substantially leverage U.S. funding agencies and result in more computing. The LHC has been a newsworthy venture so far and many universities have demonstrated their interest in their faculty participation. We believe that this interest is worthy of recognition.

Recommendation 8: Enhancement of U.S. ATLAS institutions' Tier 3 capabilities is essential and should be built around the short and long-term analysis strategies of each U.S. group. This enhancement should be proposal-based and target specific goals. In order to leverage local support, we recommend that U.S. ATLAS leadership create a named partnership or collaborative program for universities which undertake to match contributions with NSF and DOE toward identifiable U.S. ATLAS computing on their campuses. Public recognition of this collaboration should express U.S. ATLAS's gratitude for their administration's support and offer occasional educational and informational opportunities for university administrative partners such as annual meetings, mailings, video conferences, hosted CERN visits, and so on. (page 91)

2.2.4 Policies and Numbers

In the course of putting together this document, it became clear that policy and important quantitative information about existing, pledged, and targeted resources, timings, benchmarks, etc. was spread all over the web. The Computing TDR [11] is the go-to document for ATLAS policy—except when it's not! Most information exists in memos, which supersede other memos and in Indico where management representatives have given talks in various meetings. All Task Forces have something to say about "documentation" and this one is no different:

Recommendation 9: ATLAS computing and analysis policies, existing resource amounts, targeted resource quantities, data format targets, times for data reduction, etc.: basically all parameters and rules should be in one place. A policy should be considered "official" only when updated at a single twiki page. One repository should define official reality and should be updated when that reality changes. (page 6)

2.3 A Future Narrative

> *Prediction is very difficult, especially about the future.*
> Credited to Niels Bohr.
>
> *The best way to predict the future is to invent it.*
> Alan Kay, Stanford Engineering, Volume 1, Number 1, Autumn 1989, pg 1-6.

The recommendations presented here require that a number of moving parts come together: parts identifiable now plus those which will only be apparent as the analyzers begin to wrestle physics out of the raw data and have to improvise.

In order to imagine how things might evolve, we present a "story"— a fictitious, future narrative about a medium sized university group benefiting from the implementation of our recommendations—and how U.S. ATLAS and ATLAS as whole benefited from experience gained by significant, local resources. It begins as a run-up to the first 10fb^{-1} year, probably not sooner than 2012, which is the benchmark timeframe of this report.

2.3.1 Group A

When the LHC started in the fall of 2009, Group A was completing its U.S. based accelerator physics program with a group size of four faculty, two post docs, and three graduate student lines which were incrementally transitioning to ATLAS projects. Theirs was one of the dozen U.S. ATLAS institutions which had indicated that they had no Tier 3 capability to the 2008 survey accompanying this document.

Their department was unable to support much research computing and so it had always relied on its host laboratory for analysis, sending resources to the lab in exchange for fair-share batch privileges. They also stationed their post docs and students at their lab, so there was really no reason to build any campus-based computing prior to the advent of the LHC.

Replicating their traditional lab-based presence at CERN was too expensive and so they determined to build an analysis capability at home in order to serve their larger department-based group. With their funding agency, U.S. ATLAS, and their university they established a multi-year plan based on the Tier 3 model.

After considering the various options, [Table 21 on page 76], Group A chose to start at the T3w [Recommendation 2, page 76] level for all of their seats on campus and at CERN for the two or three of their staff who would be resident in Geneva. This meant that they needed modern, high-speed workstations and sufficient storage capacity to hold multiple TB of `ROOTtuples` on each station, at each locale.

As their support model, they charged two people sharing the responsibility, a post doc and a graduate student, rotating among their group [Recommendation 5, page 89]. The post doc was the official Group A contact to the "T3w Working Group." Both attended its weekly phone meetings with the ATLAS Analysis Support Manager (AASM), who was familiar with Group A's configuration and the people involved [Recommendation 4, page 89]. This familiarity came as a result of her initial campus visit when they set up the T3w stations and she helped them to work out a plan for their data and software management. Their setup was similar to other groups, as Group A had specifically purchased hardware and implemented software according to the T3w Working Group menu. The AASM personally helped them to create a "T3w out of the box" with DVD setups for most services and software to create the standard ATLAS environment.

Through the early years of ATLAS running, the group had the following commitments to ATLAS:

1. Physics analysis using early data. They were members of two physics groups which initially involved making `ROOTtuple` based comparisons between data and ATLAS-wide Monte Carlo files.

2. Calibration of a single subsystem. This project required them to stage pDPD file selections through the Tier 2 cloud, bringing to their individual stations only those which were absolutely necessary along with

a few hundred GB of ESDs and appropriate flattened conditions and trigger database records.

3. Validation of fast Monte Carlo against both data and the full simulation. This required small statistics running of various parameterized full simulation as well as iteratively producing fast Monte Carlo, all through the grid.

With multiple fb^{-1} of collider data immanent, it was time for the group to begin to prepare for larger scale responsibilities to their physics groups, including continued Monte Carlo development. The increased statistics uncovered issues requiring detailed and precise tuning and data-Monte Carlo comparison and multiple passes of high statistics, fast simulation. The U.S. Tier 2 cloud was not capable of supporting this and all of the other high statistics efforts in a timely way since production tasks and Simulation tasks were taking 80% of the Tier 2 resources [See Section 6].

Group A's solution was to continue their planned evolution in collaboration with their university (which contributed one-time funds), their funding agency, and U.S. ATLAS to create—in stages—a T3g system on their campus. They decided to begin with an implementation of the ANL T3g (E.2.1) model, as their department server room was capable of accommodating 6 high-end towers, which would provide a batch system of close to 15 TB of storage using the new 3TB disks available since 2011. This system was to grow over the next 3 years according to an agreed-upon budget. Eventually, it would become a 15 tower batch worker cluster with more than 200 cores and nearly 30TB of storage, given the 8 core cpus which were becoming standard.

Plans were also under way to lease space in a campus-wide computing building where they would eventually evolve to a Duke-like T3g, rack-based cluster of higher density, recycling their towers to individual stations (E.2.2)[4]. In this way, as ATLAS's data load increased, and the demands on simulation and eventual systematic uncertainty iteration increased, Group A was able to be a uniquely productive contributer to their physics subgroups—able to respond to requests for simulation and eventually high statistics systematic uncertainty analyses...all in-house, and all on quick notice.

This evolution could not have happened without a robust network. When, they started in 2010, their ability to transfer data from the Tier 2 cloud was averaging a few Mbps. In preparation for this evolution, U.S. ATLAS had co-

[4]Group A determined that a move to a T3gs cluster was likely to be more than what their needs required

operated with Internet 2 and their campus network administrators to tune a direct, point to point connection between their department server room and the Tier 2 site in an adjacent state [Recommendation 7, page 88] . The measurements made in late 2010 showed a handful of choke points where state and university router systems needed both firmware and hardware updates. Over time, the group could count on 10-20 MBps (80-160Mbps) throughput when they needed it. With the subscription data service implemented only a few years previously, they could transfer 1-2 TB in a day or so from that nearby Tier 2 site [Recommendation 3, page 86].

Local resources were key in a number of ways to U.S. ATLAS, and eventually all of ATLAS.

- The experience gained by their post doc-graduate student teams of local administrators, evolved in time to one which required only minimal intervention and attention from the AASM.
- While the already mature "T3g Working Group" meetings had made Group A's transition from T3w to T3g relatively straightforward, each university group, including theirs, seemed to produce one person who was particularly expert at one or another T3g issue.
- Their experience with in-house fast Monte Carlo simulation led one of their post docs to create a unique event-by-event, dynamic staging of memory and core usage in order to deal with the increasingly problematic pile-up digitization times in fast simulation[5]. His access to local, high density computing was the essential ingredient to solving this frustrating problem. Nobody had predicted this breakthrough.

All in all, Group A became an independent, capable computing and analysis center as a result of the following aspects of the Tier 3 Project:

- The common vocabulary of "T3w," "T3g," and "T3gs" meant that they were able to communicate efficiently with U.S. ATLAS and their agencies and all knew what the parameters were.
- The creation of the AASM position within U.S. ATLAS and the active phone meetings—in exchange for Group A's commitment of a small portion of a post doc/student's-worth of effort meant that they were able to build a largely independent, largely common T3w, and then

[5] A fanciful scenario designed as a placeholder for the breakthroughs which will be made by university physicists with access to sufficiently robust, local resources

T3g system. It also meant that their post docs gained valuable experience, which fed back to U.S. ATLAS when they went on to faculty positions of their own.

- The flexibility gained by implementing the "outsourcing" ability to their nearby Tier 2 of the Data Management Services (DQ2) services meant that when they needed it, about once a month, Group A could transfer files of multiple TBs without requiring human resubmission for failed transfers. It also meant that their modest storage element system was not responsible for also supporting outside requests for cached data [Recommendation 3, page 86].

- The Internet 2 support was invaluable in engineering their point-to-point connectivity to a particular Tier 2 site to a level where they could expect to transfer multiple TB files in a timescale of about a day.

- Finally, their gratitude to their administration's participation was shown through U.S. ATLAS's sponsorship of an annual video/in-person morning of ATLAS results, broadcast to university administrators from CERN and around the U.S. [Recommendation 8, page 91]. This familiarity with ATLAS physics was part of the reason that Group A's campus was able to raise awareness of the need for separate, for-lease space for research group high density computing housing, which Group A eventually took advantage of with 2 racks of computing, and RAID-based storage elements, implementing the Duke Model T3g.

The Tier 3 plan as presented above supports the flexibility that U.S. ATLAS will surely need as the unknowns become major issues once data start to flow. Just as for DØ and CDF, we expect that technology will take leaps and physicists will break out of normal procedure and create new, better ways to do things [Recommendation 1, page 74 and Section 5].

Here, Group A moves from a lab-centric, active group to a significant, campus-centric group: from a standing start to a T3w, to a significant T3g—all in basically two grant cycles. The story is perhaps optimistic, and is one of many ways in which a group might choose a "T3w," "T3g," "T3gs," or "T3af" future. But, it's a conceivable scenario for how each U.S. ATLAS institution's plans—tailored to their circumstances, planned with their agencies, and matching their particular goals—can lead to years of inspired analysis. This is how the foundation for decades of "Flexible" and "Nimble" will be built.

2.4 Conclusion

U.S. ATLAS (and CMS) face enormous challenges over the next 20 years at LHC. These include commissioning the detectors, especially those components for which U.S. physicists have been responsible; following through on the data handling, production, and reduction pledges; maintaining the sort of on-site presence which seems always to be necessary in order to be "in the know" in HEP experiments; incredibly, aggressively pursue upgrades for the 2012 timeframe, as well as the SuperLHC timeframe; and finally, participating in the physics analysis at a level commensurate with the U.S. talent and investment. Of all of these significant challenges, the last one is the hardest.

The physics rewards at the LHC are enormous—millennial in scope. The U.S. investment has been significant—hundreds of millions of dollars already with nearly half of the experimental community involved in ATLAS and CMS alone. This project will span entire careers of young physicists who are now post docs and assistant professors.

One way to handcuff progress and dilute the sort of physics analysis leadership that we expect from U.S. HEP at LHC would be to inadvertently put ourselves on a path where computing is either inadequate for the jobs at hand, or too limited to take advantage of new technologies and analysis strategies which *will* come along.

In what follows we have attempted to suggest, in part through Tevatron narratives, and in part by confronting the Tier 2 responsibilities, that more flexibility is needed. The best way to avoid such limitations is to plan for as capable a computing structure, as deeply as possible.

This is a leverage for the U.S. LHC physics mission in two ways: First, it will help to provide failover should the overall system find itself resource-limited. Second, it will provide the ability to test and deploy new ideas, new technologies, and new strategies.

"Flexible" and "Nimble" are the best guides to unleashing imaginative solutions to the coming ATLAS computing and analysis challenges over the next 20 years. Less than this commitment may hinder the U.S. physics mission to one of followers, rather than leaders.

3 Definitions and Assumptions

The current picture of ATLAS analysis in the U.S. largely follows the ATLAS model with the caveat that the U.S. computing plan provides for more data to be stored on-shore than for other nations.

3.1 The ATLAS Event Data Model

The Event Data Model (EDM) [6, 10, 11] is still a fluid concept, and if experience in other large collider experiments is a guide, will continue to evolve long after analysis begins in earnest. The amounts of data are vastly larger than any previous scale and the number of simultaneous analyzers is also considerably larger than any prior experience. This motivates our emphasis on 'flexibility" and "nimbleness."

Data flow from the HLT to the Tier 0 center will be at 200 Hz, independent of luminosity. So, for the purposes of this discussion, we can ignore instantaneous or integrated luminosity in our calculations of event data accumulation[6]. For a year of $\pi \times 10^7$ s, an annual event accumulation is about 6×10^9 per year, but for our calculations, we use the more conservatively rounded, annual accumulation of 2×10^9 events.

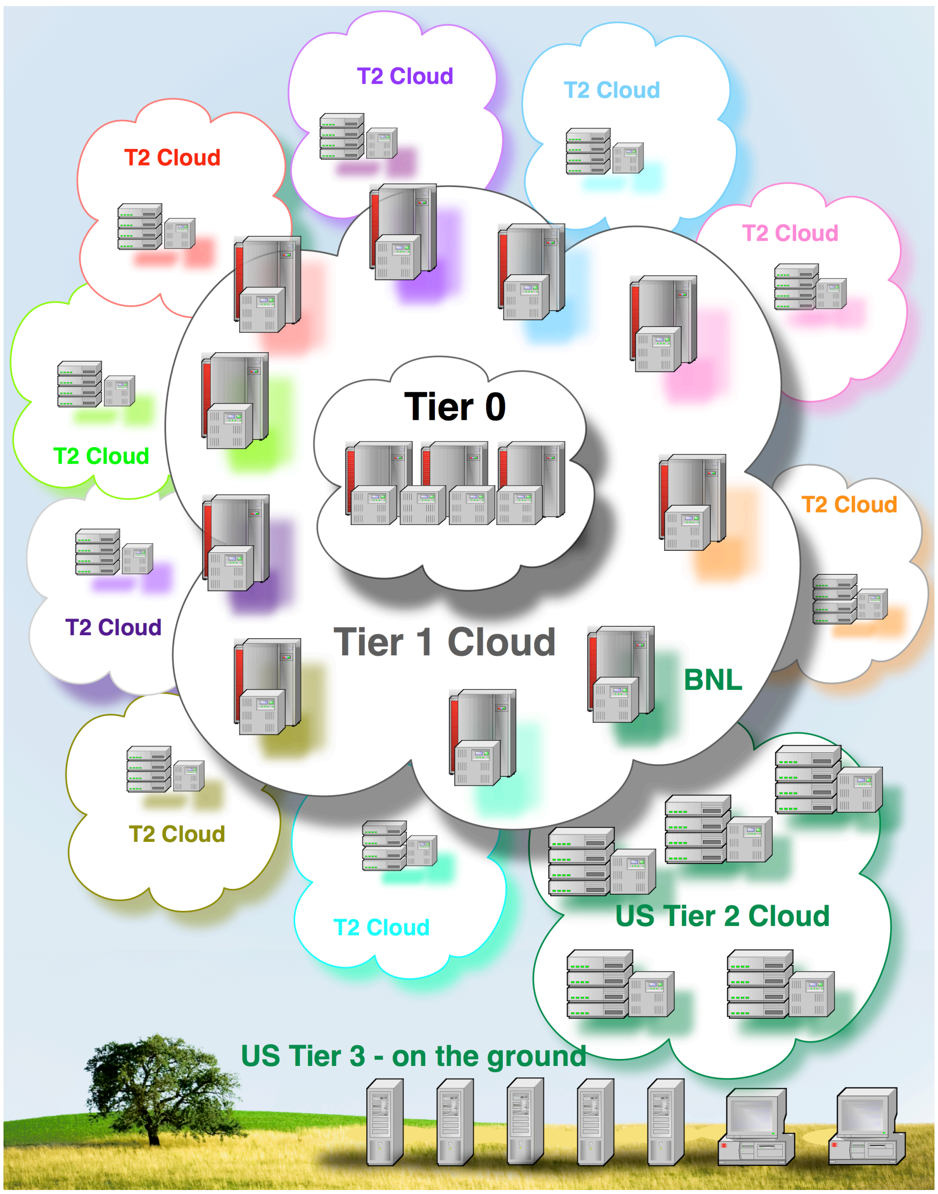
3.1.1 ATLAS Tiered Computing Centers

The production chain for ATLAS data is described below, but it consists of the successive reduction of data from RAW to manageable sizes, suitable for repeated analysis. This reduction is performed at increasing detail through an international array of Tiered computing centers. There are ten national computing hubs called Tier 1 centers in the U.S., Canada, Korea, Germany, the United Kingdom, France, Italy, Scandinavia, the Netherlands, and Spain. Around each Tier 1 center are arrayed a set of Tier 2 and Tier 3 clusters. This logical arrangement is graphically suggested in Figure 1.

Tier 1 and Tier 2 centers are ATLAS-obligated resources and the tasks which they perform are defined by ATLAS computing and physics management. For example, Tier 1 centers have responsibilities for production tasks which are ATLAS-wide, in addition to reprocessing and other responsibilities. Tier 2 centers are required to provide a minimum of 50% of their resources to ATLAS-directed effort and the other 50% to their national ATLAS computing needs.

[6] This is not strictly correct when we discuss Monte Carlo production where inclusion of pileup is highly dependent on the instantaneous luminosity and so we include it.

Figure 1: The ATLAS worldwide computing structure is a collection of "clouds" within which data are shared. Each Tier 2 cloud is logically connected to its national Tier 1 center, and in turn all of the Tier 1 centers form a cloud logically connected to the single Tier 0 center at CERN. The Tier 3 sites are "grounded," below the clouds, and not a part of their nation's Tier 2 clusters.



In the United States, the Tier 1 center is at Brookhaven National Laboratory and the five Tier 2 centers are located at: Boston University and Harvard University; The University of Michigan and Michigan State University; the University of Texas at Arlington, University of Oklahoma, Langston University, and the University of New Mexico; the University of Chicago and Indiana University; and The Stanford Linear Accelerator Center. Table 1 shows the current U.S. pledges for computing and storage for the BNL Tier 1 center, while Table 2 lists the pledges for the U.S. Tier 2 centers. (Here is the reminder: in this table and future tables, the years are presumed to be one year offset from what's shown.) Appendix D defines the SI2k benchmarking standard and lists values for popular processors. As a comparison,

Table 1: Tier 1 U.S. pledges to ATLAS [9]. Remember, these projections assumed a 2008 LHC startup and are considered for this study to be 1 year offset.

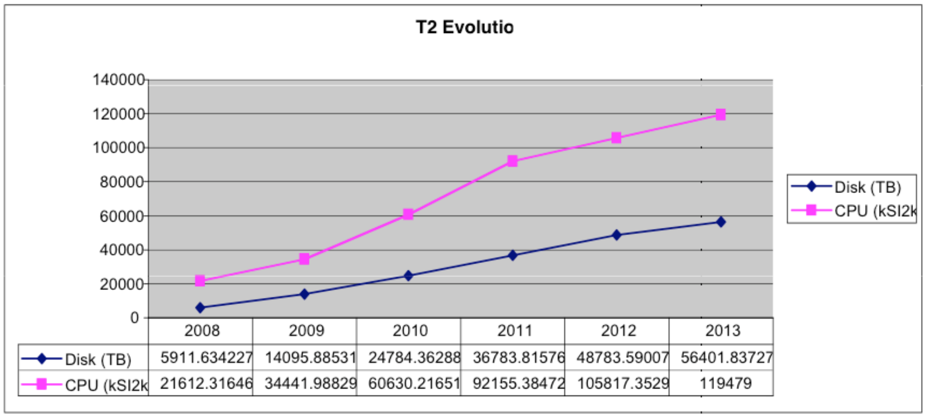
US Pledge to wLCG	2007	2008	2009	2010	2011
CPU (kSI2k)	2,560	4,844	7,337	12,765	18,194
Disk (TB)	1,000	3,136	5,822	11,637	16,509
Tape (TB)	603	1,715	3,277	6,286	9,820

Table 2: Tier 2 centers' targets of CPU and disk storage [9]

Tier 2	resource	2007	2008	2009	2010	2011
Northeast Tier 2	CPU (kSI2k)	394	665	1,049	1,592	1,966
	Disk (TB)	103	244	445	727	1,024
ATLAS Great Lakes	CPU (kSI2k)	581	965	1,406	1,670	2,032
	Disk (TB)	155	322	542	709	914
Midwest Tier 2	CPU (kSI2k)	826	1,112	978	1,262	1,785
	Disk (TB)	213	282	358	362	512
SLAC Tier 2	CPU (kSI2k)	550	820	1,202	1,191	1,685
	Disk (TB)	228	462	794	1,034	1,462
Southwest Tier 2	CPU (kSI2k)	998	1,386	1,734	1,966	2,514
	Disk (TB)	143	256	328	650	1,103
Total U.S. Tier 2	CPU (kSI2k)	3,348	4,947	6,367	7,681	9,982
	Disk (TB)	842	1,567	2,467	3,482	5,015

Table 22 in Appendix D on page 105 shows the computing capabilities of

Figure 2: ATLAS Worldwide Tier 2 evolution.



a few recently used processors and disk systems. Notice that the U.S. Tier 2 system as a whole will constitute approximately 10MSI2k units of computing, or more than 7,000 job slots and more than 5PB of storage. At a single location, this combined capability would amount to more than 20 full racks of typical 8 processor nodes—nearly 1/2MW of heat production—and more than 30 racks of 3U Dell PVMD1000 enclosures. Hence, part of the reasoning behind distributing Tier 2 resources among many locations.

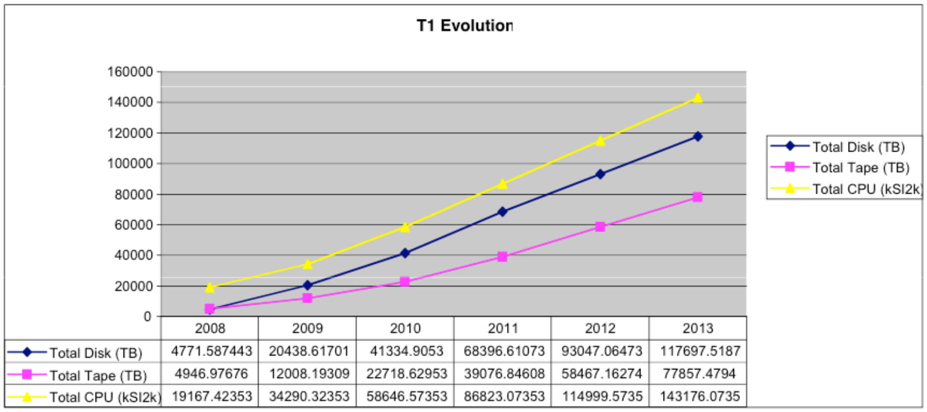
As for ATLAS as a whole, Figures 2 [13] and Figure 3 [13] show the evolution of the collaboration's capabilities over time[7] For our set-point of 10fb^{-1}, the 2010 numbers are relevant.

3.1.2 ATLAS Data Formats

The trip from RAW data to the physicist desktop is one of successively reducing the contents and the numbers of the event records. The deeper one follows this reduction, the smaller the total event sizes are and the more specialized is the audience. The newly formulated analysis guidance specify that the lowest order event formats should be analyzable by the highest

[7]Note added in "proof" in March 2009: new expectations for both global Tier 1 and Tier 2 evolutions were announced at the Software Week in March 2009. The Tier 2 expectations are essentially the same for computing, while storage expectations are lower by about 10% for 2010, converging to those plotted here by 2013. The Tier 1 expectations for both processing and storage are now roughly 20% lower in 2010, closing the gap to the original benchmarks plotted here by 2013.

Figure 3: ATLAS Worldwide Tier 1 evolution.



	2008	2009	2010	2011	2012	2013
Total Disk (TB)	4771.587443	20438.61701	41334.9053	68396.61073	93047.06473	117697.5187
Total Tape (TB)	4946.97676	12008.19309	22718.62953	39076.84608	58467.16274	77857.4794
Total CPU (kSI2k)	19167.42353	34290.32353	58646.57353	86823.07353	114999.5735	143176.0735

level software tools, such as `Athena` .

The features of each data format which are important for this discussion are these:

RAW data A fraction of the streamed raw data is sent to each Tier 1 site, destined for tape storage. RAW data are then redundantly stored within the Tier 1 international cloud. As currently configured, filtering by stream is done at this stage.

ESD data The Event Summary Data (ESD), bounded by the filtered streams, and are sent to each Tier 1 for tape storage. The U.S. Tier 1 center at Brookhaven National Laboratory (BNL) will uniquely store 100% of the ESDs on disk. They contain reconstructed information, including calorimeter cell data (for example as much as ∼270 kB/event for top events), tracking information (∼200 kB/event for top events), and full trigger information. Other Tier 1 centers store a fraction of the total.

AOD data The Analysis Object Data (AOD) is a summary of the ESD information and for ATLAS data and event records are bounded by the same stream boundaries as the RAW and ESD formats. It is currently larger than anticipated by about 20% and the expectation is that it will be reduced. The AODs were not designed to contain calorimeter cell data (although at writing, electromagnetic [EM] object cell information is included), nor hit details, nor full trigger information. The AODs (and ESDs and D^{n}PDs for $n > 2$ are accessible from within the

`Athena` framework, and also from within `ROOT` like structured Ntuples using `AthenaROOTAccess` in Linux. Figure 4 sketches the data flow from T0 through to the Tier 2 centers.

TAGs The TAGs are event-level metadata descriptions which come with pointers to the `POOL` file-resident data. They are meant to facilitate event selection.

Table 3 shows the target record sizes for the various data formats, while Table 4 shows the recent size of the two major formats for five different streams [17]. Obviously, reaching the target sizes is not complete and we can see where focus is required by looking at the contents of one of the FDR2 AOD, shown in Table 5 [7].

Table 3: Data formats for ATLAS and quantities used in this analysis.

Format	Target Range	Current	Used	1 Year Dataset
RAW	1.6 MB		1.6 MB	1600 TB
ESD	0.5 MB	0.7 MB	0.5 MB	500 TB
MC ESD	0.5 MB		0.5 MB	500 TB
AOD	0.1 MB	0.17 MB	0.150 MB	100 TB
TAG	1 kB		1 kB	1 TB

Table 4: The sizes per event (in kB) of various streams for the v13 ESD and AOD formats.

Container	ESD	AOD
eg	742	162
jet	748	163
express	?	172
minbias	425	32
muons/B	737	176
Total	> 2MB	426

The AOD "workhorse" data format is targeted at approximately 100 kB/event in size. In principle, if resource limitations were nonexistent, one could do almost all ATLAS analysis on the AODs. But, four more simple calculations show that this is not possible if one reads every event:

Figure 4: The flow of data from SDX1 through the CERN-based T0; the set of Tier 1 centers; and, through the BNL Tier 1, the U.S. Tier 2s.

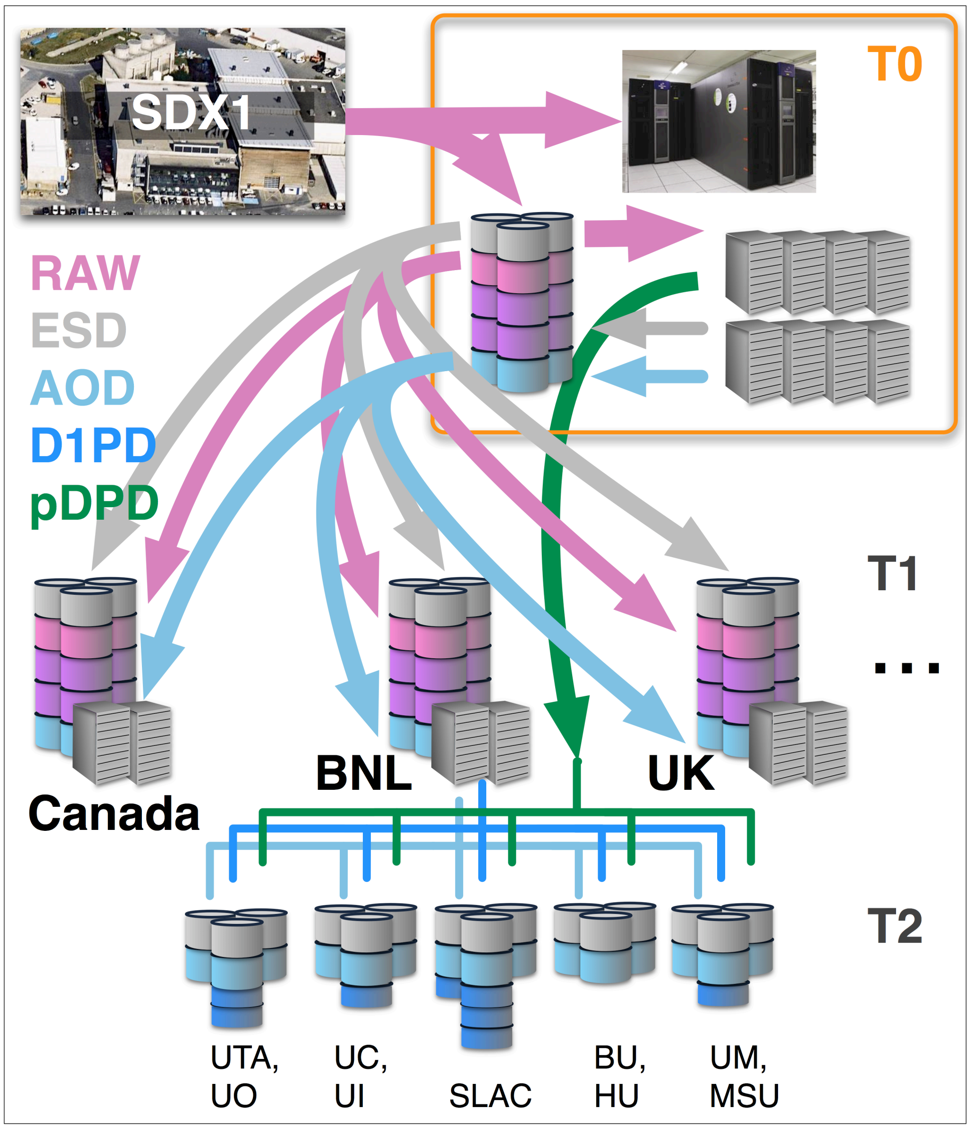


Table 5: The contents (in kB) of the FDR2 AOD, totaling 166kB.

Trig	InDet	Calo	Jet	Eg	Muon	Tau	EMT	EID	MET
62	20	25	25	3	7	2	15	4	3

1. If the AOD were the only format available and we take **Observation 2** seriously, then transferring it to a university site is problematic. First, it would require at least a 100 TB Storage Element (SE) system at the university end—the equivalent of 10 Dell MD1000 enclosures with 10 1TB drives each, which is an entire rack of SE and server units—about a $60k investment.

2. But, even if this raw storage capacity existed, the actual transfer of 100 TB of data assuming a 1 Gbps dedicated optical connection would still be limited by the few-hundred MBps disk Read/Write speeds of even a high-end RAID system. The transfer would take roughly 2 weeks. Realistic, sustained overall data transfer within the ATLAS world is currently considerably less than a fraction of a 1 Gbps network. Without dedicated fiber links, data transfer rates are unacceptably low—a few MBps— in many areas.

3. Even if a university researcher relied on a large, remote site for calculations with the AOD dataset, one still faces unacceptable analysis limitations. If we assume a high-end RAID Read rate of 200 MBps each, that `Athena` is capable of reading at disk-access speeds, and only a trivial calculational requirement of 1 ms/event (such as only plotting histograms), then a remote dedicated cluster of 100 cores (about 12 nodes) would require essentially a whole day to go through the entire AOD. Obviously, for most analysis tasks, a higher calculation load is required. Dedication of 100 job slots in multiple, continuous 24 hour blocks to single university user analyses at a remote Tier 1 or Tier 2 site would be a significant commitment. Plus, most analysis tasks require considerable more computation. For reference, a 20 ms calculation on a single node would process only 3% of the sample in a whole week per core.

4. One idea is that the system of Tier 2 clusters is simply used to reduce an AOD into something much smaller for subsequent analysis. If in this example, the task was to analyze the AOD and only write a 10 kB

ROOTtuple as a quick skim of 2 ms/event, this would still require about 5 core-weeks to produce.

Although whole-dataset AOD analyses are obviously more suited for Tier 2s, relying solely on AODs is not sensible. The ways out of this problem are well-known and applied in all HEP experiments. The first step is the early filtering of events into streams. These can be based on a variety of criteria and can be either inclusive (with the same events repeated in multiple streams) or exclusive (with no data replication). The ATLAS plan for streams is only a few years old and is still under review. But, roughly, there are expected to be pure physics streams, probably based on trigger designation, and a handful of calibration streams—as many as 4-7 of the former and 3-4 of the latter. The plan calls for the streams to be built in the front-end of the production process at the SFO on the HLT output and implemented early. This Streaming Study Group [12] recommends that the mental picture should be one of a stream being a stand-alone experiment. Obviously, cross-stream analyses must be possible and the careful accounting of luminosity and duplicate event counting is always present.

3.1.3 Derived Physics Data

Even with streamed data splitting, there is still so much information that further reduction is necessary. This is a relatively recent conclusion for ATLAS and details were eventually fleshed out with the publication of the *Analysis Model Report* of January, 2008. [6] Here a plan was suggested which includes the introduction of Derived Data Physics (DPD) data formats, a concept which is obvious in principle, but complicated in practice. Three progressively more specialized DPDs are envisioned:

D^1PD Also called the "Primary DPD," this is a format which is envisioned to be unique to 10-12 different groups, probably a skim (see below) of the AOD according to trigger stream, with minimal analysis. The guidelines are that the sum of all D^1PD should equal the total AOD volume. Early in the run, 80% of the D^1PD size is expected to be devoted to the "performance DPD" (called pDPD here), with the remaining 20% divided among approximately 10 physics DPDs. Estimates of the sizes of future fraction of pDPD to total vary and we will eventually presume that ultimately 20% of the total will be for pDPD.

D^2PD The secondary DPD format is undefined at this writing, but generally thought to be the stage at which significant analysis is performed at the

Athena level, according to the physics group need. It is anticipated to be designed to particular physics or performance groups' requirements and will likely be augmented with calculated and derived quantities and be slightly bigger than the D^1PD from which it was made. So, its creation will be longer and the files will be larger, perhaps as much as 10% or so.

D^3PD The tertiary DPD is envisioned to be lightweight and as a flat ROOTtuple, intentionally portable. Predictions of its size vary, but it's likely to be something of order $1/3 \times D^1$PD . Practice shows that for the same information in the file, the D^3PDs are smaller and faster to analyze than the POOL based formats.

pDPD The "performance DPD" is designed to facilitate commissioning tasks and early calibration and data quality development. It is currently built directly from the ESDs and contains information not passed through to AODs.

private ROOTtuple Of course, users will likely make their own ROOTtuple formats. While D^3PDs will be official, everyone will produce private ROOTtuples for their own use.

Table 3 adds DPD entries with reasonable guesses for their respective event record sizes.

Table 6: DPD formats and size estimates. N.B. The DPD current amounts are from [18] and are approximations to FDR $t\bar{t}$ data and are just presented as a snapshot and not to be taken literally.

Format	Target Range	Current	Used	1 Year Dataset
D^1PD	$\sim 1/4\times$ AOD	31 kB	25 kB	25 TB
D^2PD	$\sim 1.1\times D^1$PD	18 kB	30 kB	30 TB
D^3PD	$\sim 1/3\times D^1$PD	5 kB	6 kB	6 TB
pDPD	$\sim 0.1\times$ AOD	NA	?	?

The same kit and storage technologies that were used to create the AthenaROOTAccess approach to AOD analysis, made it possible to use the same approach for derived data. The D^1PD and D^2PD formats are directly analyzable with Athena as they are POOL based, while as a flat ROOTtuple, the D^3PD will not be POOL based. There is considerable uncertainty surrounding most important aspects of the DPD concept and include critical questions like: What will be the content of each layer of DPD format? Where they will be produced? Where each DPD dataset be stored? How often they will be produced?

These uncertainties affect how we evaluate the potential efficacy and configuration of possible Tier 3 systems. The FDR2 exercise did not fully explore the space of DPDs and users tended to produce flat `ROOTtuples` directly from the AODs, so the whole concept is both conceptually and operationally untested at this point.

Ultimately, something like the above DPD scenario will be realized and so we chose to presume it will occur as advertised and had to make choices on the various sizes, which ATLAS resource would make and store which format, and how often. Figure 5 shows the staging of the various formats.

3.1.4 Responsibilities of the Tier 1 and Tier 2 Centers

With the above capabilities and data formats, the responsibilities of the U.S. Tier 1 and Tier 2 centers can be sketched.

The responsibilities[8] of the U.S. ATLAS Tier 1 site at Brookhaven National Laboratory include:

- Reliable storage of complete sets of ESD (current on disk plus previous version on tape), AOD, Ntuples, and TAGs on disk plus a fraction of RAW data as well as all U.S. generated RDO (Raw Data Objects) data: Monte Carlo, and Primary data. The fraction of RAW varies from site to site, but is anticipated to be roughly 10% per Tier 1. The fraction of ESDs varies from site to site and is expected to average 20% per Tier 1. However, the U.S. Tier 1 is designed to hold 100% of the ESD data in two copies. 100% of two copies of the AODs are expected to be stored at all Tier 1 sites.

- Anticipated, but not determined yet: 100% of all D^1PD are to be stored at all Tier 1 sites.

- Provide CPU for managed ATLAS-wide production

- CPU and storage for ATLAS-wide reprocessing of RAW data[9]

[8] Abstracted from [9] and [11].

[9] Reprocessing is planned to take place in two ways: within the first couple of months of T0 distribution, more reliable calibrations and alignments are expected to be available and so they will be applied in a global reprocessing at each Tier 1. Next, perhaps annually, but certainly at some later time still better calibrations or methods are expected to be available and one or more reprocessings will again take place.

- Provide CPU for regional and local production of large samples through `Panda`
- Provide CPU for user analysis through `pAthena`
- Provide CPU for interactive `Athena` for testing/software development

The responsibilities of the Tier 2 Cloud in the U.S. include:

- Reliable storage of RAW, ESDs, AODs, and TAGs on disk for Monte Carlo and Primary Data. The fractions of RAW and ESD formats will be trace amounts for debugging and code development. The fraction of AODs on Tier 2 sites in the U.S. is not determined: during early running, 100% of AODs are expected. During long-term, stable running approximately 1/3 of all AODs are expected to be distributed across the U.S. Tier 2 Cloud.
- Anticipated, but not determined yet: the hope is that multiple copies of all D^1PD are to be distributed across the entire U.S. Tier 2 Cloud, so that multiple sites might hold the same data.
- Not determined yet: what fraction of D^2PD data will be available.
- 50% of CPU resources are centrally managed for Monte Carlo production and other ATLAS-wide responsibilities.
- An undetermined fraction of CPU resources are likely to be detailed to D^2PD and D^3PD production.

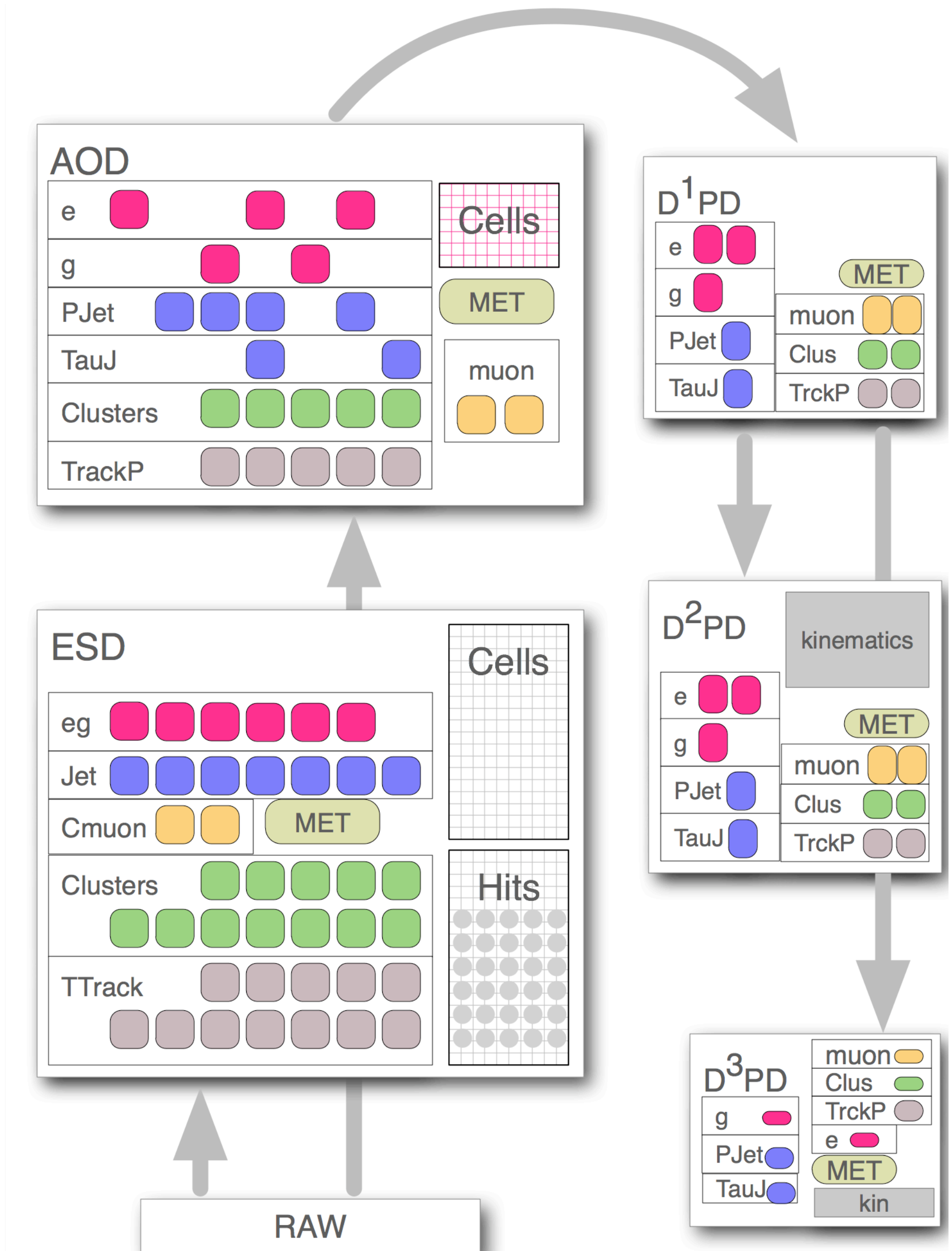
Notice, that the location of DPD production and storage is not yet determined.

3.2 Analysis Model

The Analysis Model for ATLAS has evolved over time and now has settled on the following order of events, although it is still largely untested. It is difficult to find a single, definitive description of what is to be done where and what is to be stored where, but it is clear that the Tier 2 sites are integral to the plan and that they take on new tasks and storage responsibilities.

In order to describe the production flow, we enumerate the various operations which can be performed on a data record, transforming an input file to an output file:

Figure 5: The path of an ATLAS event record from ESD through the last flat `ROOTtuple`, D^3PD stage. The chain shown strictly follows the Analysis Model [6], but the possibility exists that it might be advantageous to produce D^3PDs , for example, from D^1PDs or AODs.



Skim, SK Unwanted events in an input file are eliminated and desired events are written to the output...as a selection. Example: skimming files for particular trigger patterns.

Thin, TH Record by record, various objects within events are eliminated and the remainder of each record is written to the output file. Example: thinning files to retain only the highest quality muon fit.

Slim, SL Record by record, information within particular object containers are eliminated in the output events. Example: Detailed fit information is removed from tracks.

Augment, AU Record by record, user information is added to the output file. Examples: invariant masses are calculated and stored for particular electron pair clusters. Database information is stored when running an `Athena` job via `pAthena`. Note, **AU** is not an official ATLAS nomenclature. It is added here for completeness to represent an important aspect of data production at the D^2PD stage (see below).

Figure 5 visually suggests the notions of thinning, slimming, and augmentation. We assume that the responsibilities for data flow are according to the following routes:

1. **RAW → ESD: produced at T0**. The RAW and ESD data are collected at the T0 site, written to tape, and distributed around the world to the 10 Tier 1 centers in such a fashion that two complete copies of the ESDs exist within the Tier 1 cloud.

2. **ESD → AOD: produced at T0**. As shown in Fig. 5, the production of AOD is a matter of slimming and thinning (not skimming). For example, the detailed cell and tracking containers are eliminated. Currently, in fact, the cell information is slimmed to retain those which are associated with electrons.

3. **ESD → TAG: produced at the T0**. Likewise, the TAGs are produced with the ESDs and follow them to the Tier 1 sites with the AOD files.

4. **ESD → pDPD: produced at the T0**. This will be the primary, early-years path for commissioning and early calibration development.

5. **AOD → D^1PD : produced at the Tier 1**. The current plan is that for early running, only a handful of D^1PDs will be produced, and

probably remade often. After calibrations are understood and physics-quality data are beginning to reliably flow from CERN, the plan calls for about a dozen D^1PDs to be produced according to the various inclusive streams. The content of the D^1PDs is not determined and they have not featured prominently in the FDR exercises. It is expected that their content will be determined by the physics groups with the controlling interests in the various streams themselves and that their production will be a responsibility of those groups to keep them identical, world wide.

6. **D^1PD → D^2PD : produced at the Tier 2**. The fate of the secondary DPDs is less clear. They are again expected to be the province of the physics groups, but it is possible that subgroups may become active in the production of specialty formats. While they are likely to be further skimmed, thinned, and slimmed, a central feature of the secondary DPDs is that they will will be "decorated" with specialized user data. They may, then, be larger data records than their parents, but since they will presumably be skimmed, the overall data sizes may not be significantly larger. These will likely be very different, subgroup to subgroup.

7. **D^2PD → D^3PD : produced at the Tier 2**. The flat `ROOTtuple` data sets will be the province of the individual physicist. They will be the only format not included in a `POOL` storage. It is not clear where they will be stored and whether ATLAS will have responsibility for their evolutions.

Observation 3 *The entire DPD production chain (D^1PD , D^2PD , and D^3PD) is to be an essential feature of the analysis sequence. And yet the lack of experience in producing DPDs through the whole chain is difficult to understand. Reliable timings are unavailable, for example. Storing both AODs and D1PDs at Tier 2s seems redundant, but there is yet no guidance on which, how much, when, how the AOD format storage and the DPD storage and production is to be arranged. The ultimate storage load on the Tier 2s is therefore unevaluated (see below). (Note, the performance DPD—dDPD—will be the major data format in early running and is not a part of the concern here.)*

4 The Use Cases

The data reductions steps, copying operations, and data creation stages are a finite set. In this section we outline in graphical and tabular form the most significant examples, using definitions found in Table 7. The data production chain is pictured in Figure 6 for reference. (Note that the 2 month reprocessing at the Tier 1 centers is not shown on this figure for simplicity. Also, note that it assumes that AOD and ESD production happen as a chained sequence. This is not yet finalized as the AODs may be produced as a separate step from cached ESDs, or the whole RAW→ESD→AOD sequence might be one large step.) In general, the operations fall into four

Table 7: Operations or transformations used in the Use Case enumeration and the simulation in Section 6.

transformation/ definition	abbrev.	comments
Skim	SK	Elimination of unwanted events.
Thin	TH	Elimination of objects within records.
Slim	SL	Elimination of information within objects, within events.
Augment	AU	Addition of derived quantities within event records.
Copy	C	File transfer from one tier to another over the grid or directly.
Tier 1	T1	A general Tier 1 site.
Tier 2	T2	A general Tier 2 site.
Tier 3	T3	A general Tier 3 site.
Tier 2 Cloud	T2CL	The entirety of the Tier 2 cluster set.
Histogram	hist	The production of histograms as a final output of a transformation.
Text	txt	The production of an ASCII file as the final output of a transformation.
Special	sp	A special format.

broad categories: Steady State Data Distribution; Dataset Creation; Monte Carlo Production; and Chaotic Data Analysis.

Figure 6: The production stages from the HLT through the D^{3}PD as originally envisioned. The yellow data formats are `POOL` based, while the pink D^{3}PD is a flat `ROOTtuple`. (Following A. Shibata.)

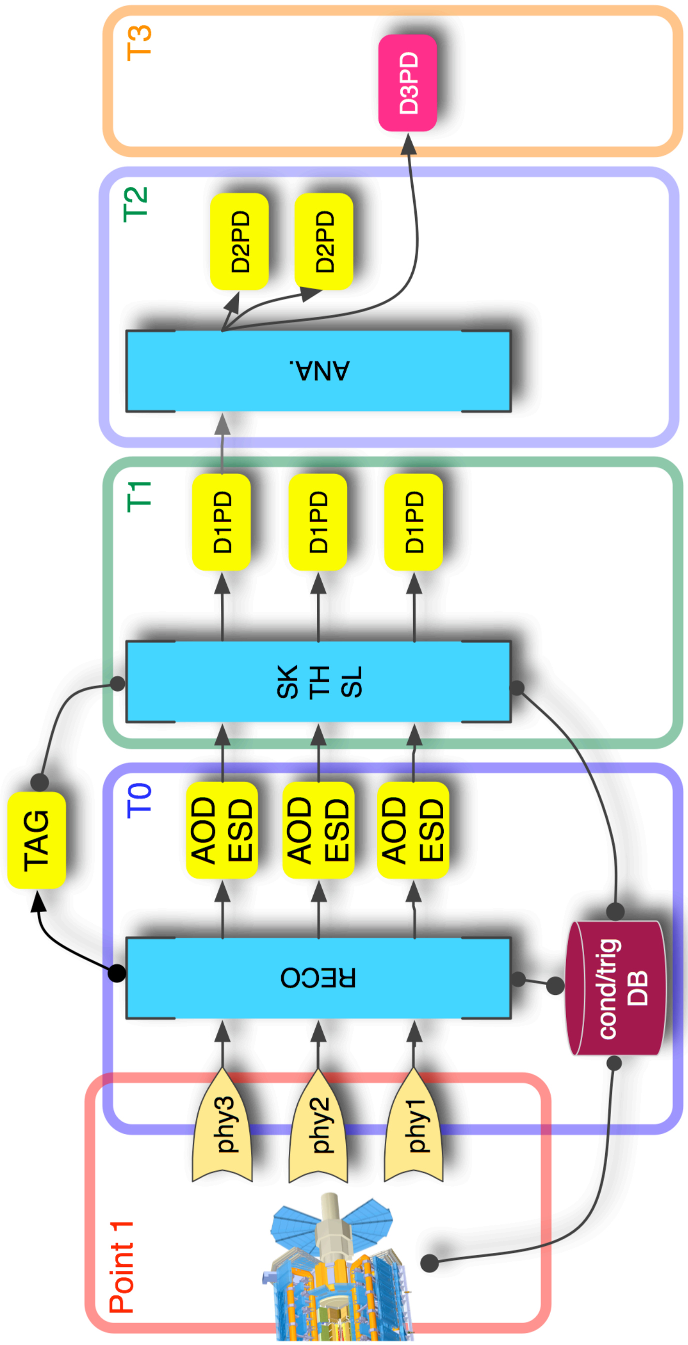


Table 8: The Steady State Data Distribution Use Cases. In most cases, this is a Copy operation involving Primary formats.

	data in:	data out:	from:	to:	by:	trans:	who:
	ESD	ESD	T0	T1	T0	C	
P1	AOD	AOD	T0	T1	T0	C	all groups
P2	AOD	AOD	T1	T2	T1	C	all groups
P3	AOD	D1	T1	T1,T2	T1	SK, SL, TH	all groups
P4	ESD	pDPD	T0,T1	T2,T3	T0,T1	SK, SL, TH, AU	all groups

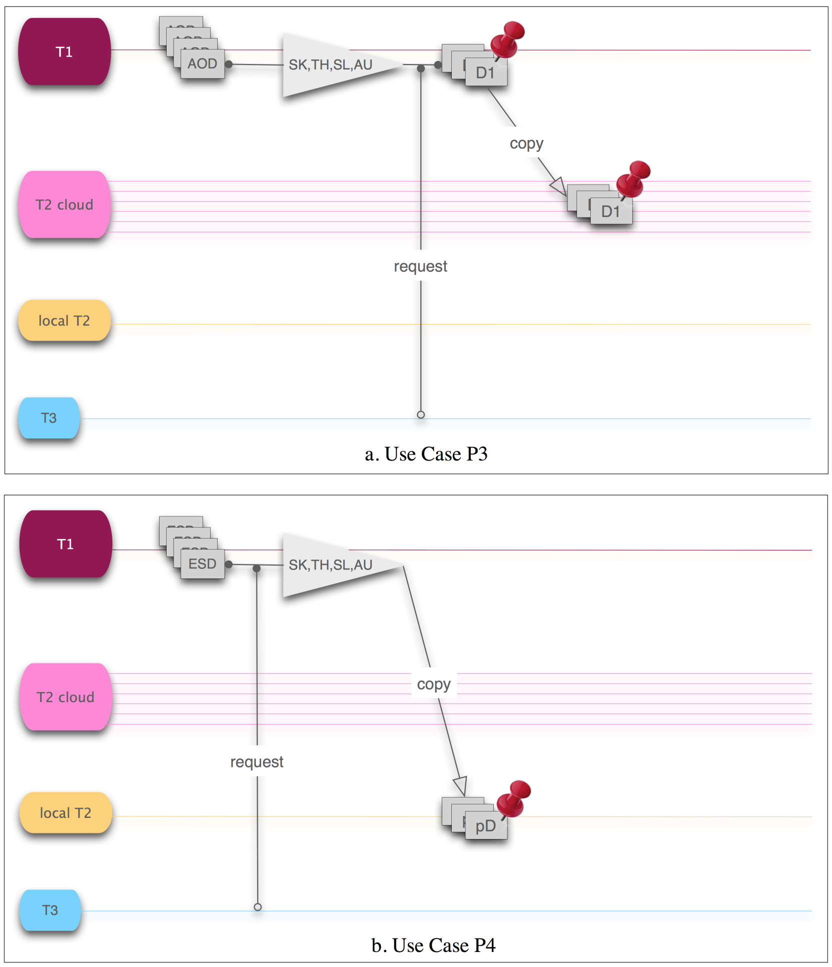
4.1 Steady State Data Distribution.

A number of operations automatically flow from the T0 center at CERN, pushing data to the Tier 1's. The ESD, AOD, and TAGs are T0 responsibilities and are cached at the Tier 1 centers (along with RAW). The D^1PD format is subsequently created at the Tier 1 from the ESDs. Table 8 lists the operations, including the point of origin, destination, actual computational responsibility, as well as the group responsible for the operation. As a graphical representation, Figure 22 shows Use Case P3, corresponding to the production of D^1PD and its subsequent distribution to the Tier 2 cloud. Use Case P4, the production of a performance DPD (pDPD) would be identical, except that it will likely be produced form the ESD format, rather than the AOD.

4.2 Dataset Creation.

Dataset creation at the Tier 2 centers could be a major responsibility and will involved parallel management of all of the ATLAS world Tier 2 centers. While final decisions are yet to be made about the size, source, and roles for the D^2PD and D^3PD , the current plan suggests that their production and, in the case of the D^2PD , storage are Tier 2 responsibilities from locally cached D^1PDs . Table 9 enumerates the likely Use Cases involving these formats and Figure 8 pictures the two important cases ("C1" and "C2") for creation and a possible storage and transfer operation for both D^2PD and D^3PD . The current analysis model is not clear on where the D^2PD and D^3PD will be produced. The D^2PD is a serious analysis task and will possibly take significant time and require substantial reserved space for the outputs. It is

Figure 7: The following figure format will be used extensively in what follows. It is meant to quickly convey a picture of the movement of data, the transformations applied, and the triggers for events among the computing tiers. The primary DPD production path, shown for Use Case P3. The performance DPD production, Use Case P4, would be identical, except it is likely to be made from the ESD. P4 can originate at the T1, as shown, for reprocessed data, or from the T0 for early data.



a. Use Case P3

b. Use Case P4

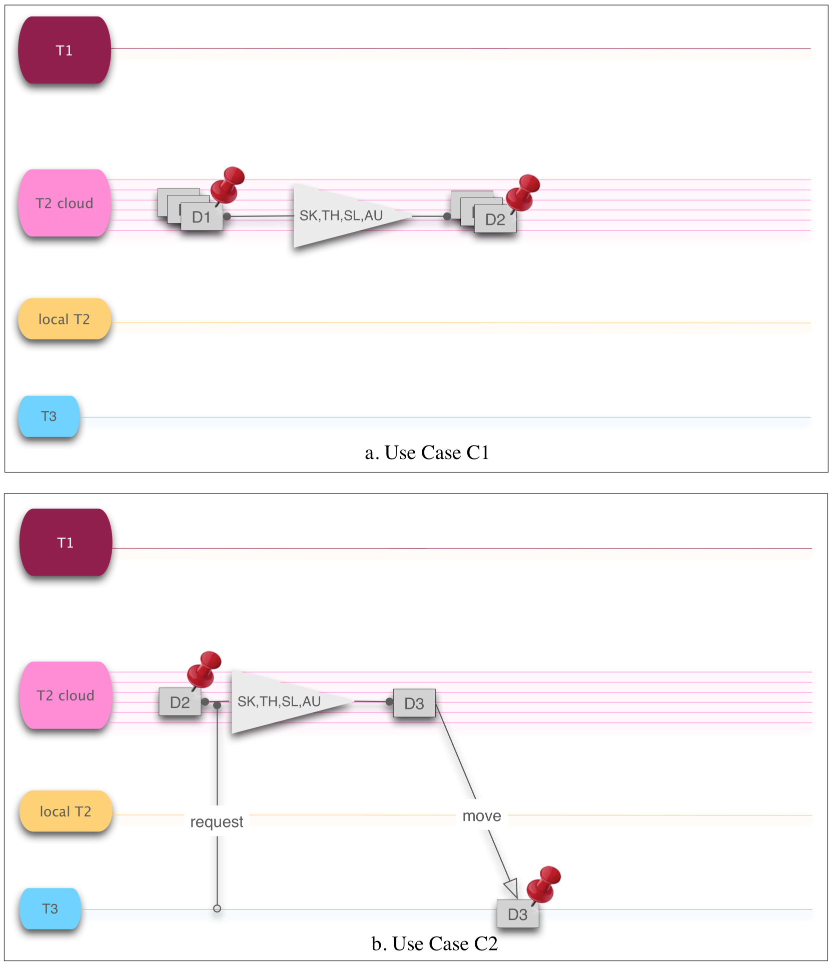
also not clear how often these formats will be produced, but most estimates are on the order of every month. As Table 9 suggests, the responsibility for defining the contents and the frequency of production of the D^2PD is likely to be that of the relevant physics groups. The Use Cases (C1 and 2 for dataset "Creation") are both a part of the normal production process, but also include the likelihood of episodic and chaotic D^3PD creation.

Table 9: The Steady State Data Format Creation Use Cases. In addition, a Fixing use case has been included.

	data in:	data out:	from:	to:	by:	trans:	who:
C1	D^1PD	D^2PD	T2	T2CL	T2CL	SK,SL, TH, AU	all subgroups
C2	D^2PD	D^3PD	T2CL	T2CL	T2CL	SK,SL, TH, AU	particular subgroups
F	D^1PD	D^2PD	T2CL	T2CL	T2CL	SK,SL, TH, AU	particular groups

The D^3PD datasets will likely be episodically produced, rather than as a part of the continuous production process. It is not expected that they will require permanent storage at the Tier 2s, but that they will be pulled from Tier 2s after their production back to the home Tier 3 from which the request was initiated.

Figure 8: The D^2PD and D^3PD production paths. In 8a., the secondary DPD is shown produced in the Tier 2 cloud and stored their for its lifetime. In 8b., the tertiary DPD is produced in the Tier 2 cloud, on demand from users and brought back to the requester at his/her institutional Tier 3 center.



a. Use Case C1

b. Use Case C2

4.3 Monte Carlo Production.

Monte Carlo production is a special case. While the actual simulation tasks are relegated to the Tier 2 centers, the physics generator inputs are strictly controlled at the Tier 1 centers. The Tier 2s move simulated, digitized data back to Tier 1, which in turn would serve it back as if it were real data. So, Monte Carlo data move in two directions. During data-taking, this will all be going on simultaneously with real data movement.

Monte Carlo production comes in multiple levels of sophistication, from a full GEANT simulation through to a fast, parameterized version. Experiments in the past have taken different approaches to this effort. The LAr calorimeter-based DØ experiment relies almost solely on full GEANT simulation, while CDF uses a faster approach. The ATLAS experiment's complexity, however, prohibits reliance on full-simulation for more than a fraction of the dataset.

Table 10: The Monte Carlo Production Use Case.

	data in:	data out:	from:	to:	by:	trans:	who:
M1		sp	T1	T2	T1	AU, C	RAC
M2	sp	RDO	T2	T1	T1	AU,C	grid

For the purpose of this study, the ATLAS Monte Carlo full simulation ("Full") takes place in four stages: Generation, Simulation, Digitization, and Reconstruction. Because resources are precious and mistakes are costly, there is a considerable bureaucracy surrounding the officially sanctioned Monte Carlo (MC) generation steps:

- Generation. The generators for MC come from many sources. The large, general purpose generators PYTHIA and HERWIG are used to produce stable particles as the inputs to GEANT, already taking care of the promptly decaying particles. Both have different hadronization models and implementations and so having two is sometimes important. While both have physics models built in, one is not limited to those program's choices of parameters or reactions as they both can serve as vehicles for taking more specialized, theoretically oriented particle physics generators' outputs as their input to hadronization engines. The end result, in any case, is a set of relatively stable particles in standard HepMC format, suitable for passing to the detector simulation. The Generator stage in the U.S. is handled by the Tier 1 center at Brookhaven.

- Simulation. By far, the bulk of the computational effort is in the simulation stage during which the Generated particles are stepped through the modeled detector material, depositing energy, decaying, and scattering. The control over the computational effort is considerable, where "knowing when to stop" is a critical parameter for slow particles. This has been tuned and is relatively stable. To set the scale, where Generation of a single event may take small fractions of a second, Simulation is many minutes on modern CPUs. The Simulation stage is executed at the Tier 2 centers and will dominate much of the ATLAS obligated resources for the life of the experiment.

- Digitization. The energy depositions must be "digitized" in order to create outputs which look like those of the real data outputs, the eventual Raw Data Output (RDO) files. At this stage, noise is added as well as the problematic "pile-up" of overlaid minimum bias events from multiple interactions. This latter overlay is according to a luminosity-dependent algorithm and is problematic, both from the point of view of the additional effort required for computing (as much as 2-10 times the time it takes to generate bare events, ignoring pile-up), and because the model for pile-up will only really be understood when real data arrive. The Digitization stage is also done at the Tier 2 centers.

- Reconstruction. Both the HLT and event reconstruction are run on the RDO files, with the latter identical in format to real data. The RDOs are converted to byte-stream format and sent back to the Tier 1. Currently, the Reconstruction step happens at the Tier 2s, and the subsequent data would then be restored back on the Tier 2s as described above.

Figure 9 shows a graphical representation of the event generation and the simulation use cases from Table 10 in Appendix **??**. Once, the byte-stream data are cached at the Tier 1 center, then data production of the regular formats happens as normal, but with the Tier 1s taking the T0 role in the creation of the ESD, AOD, and TAG formats.

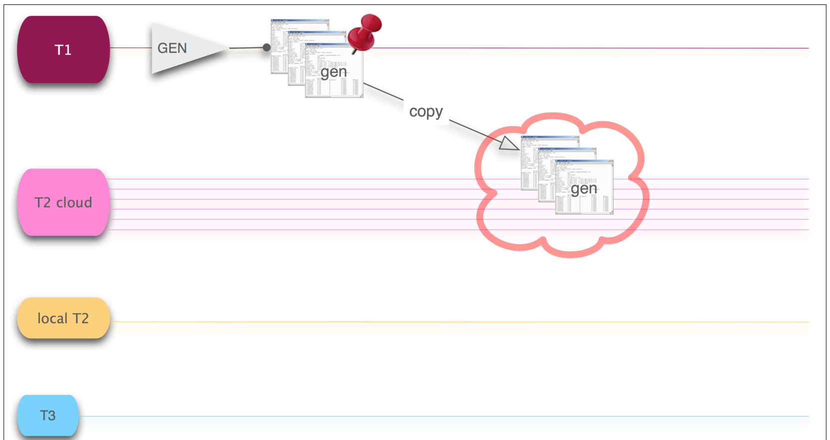
4.4 Chaotic Data Analysis.

The actual hands-on analysis is predictably disorganized and personal and is expected to be done at the physicist workstation near the person doing the work.

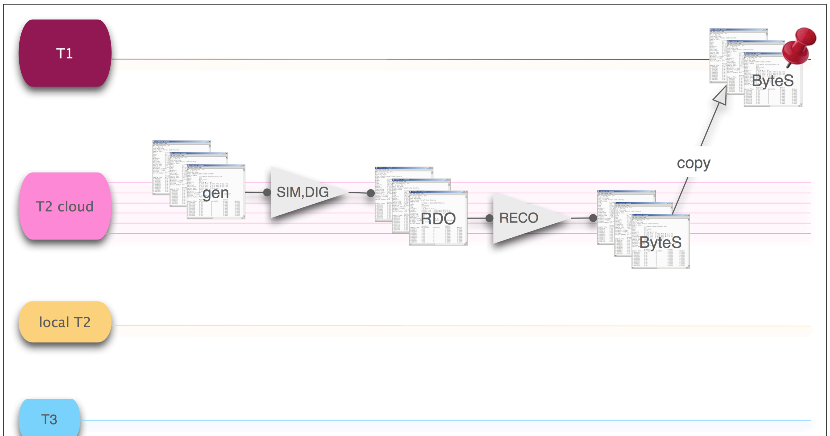
4.5 Chaotic User Analysis Use Cases.

The naively anticipated Use Cases for Tier 3 centers is that they submit jobs to the Grid for `ROOTtuple` creation and bring them back to the Tier 3

Figure 9: The Use Cases M1 and M2. Note that Monte Carlo data are copied back to the Tier 1 centers as the primary way to make them available to the whole collaboration. In some cases, some Tier 2 centers may have sufficient bandwidth to provide that availability themselves.



a. Use Case M1



b. Use Case M2

for "chaotic" analysis. These tasks would be likely inspection of data for irregularities, performance of various verification tests, signal-background comparisons, and Monte-Carlo-data comparison. Each of these will likely require repeated, reapplication of the use case when various weighting factors are determined and applied, and/or selections are refined and applied.

Table 11: The Chaotic Analysis Use Cases.

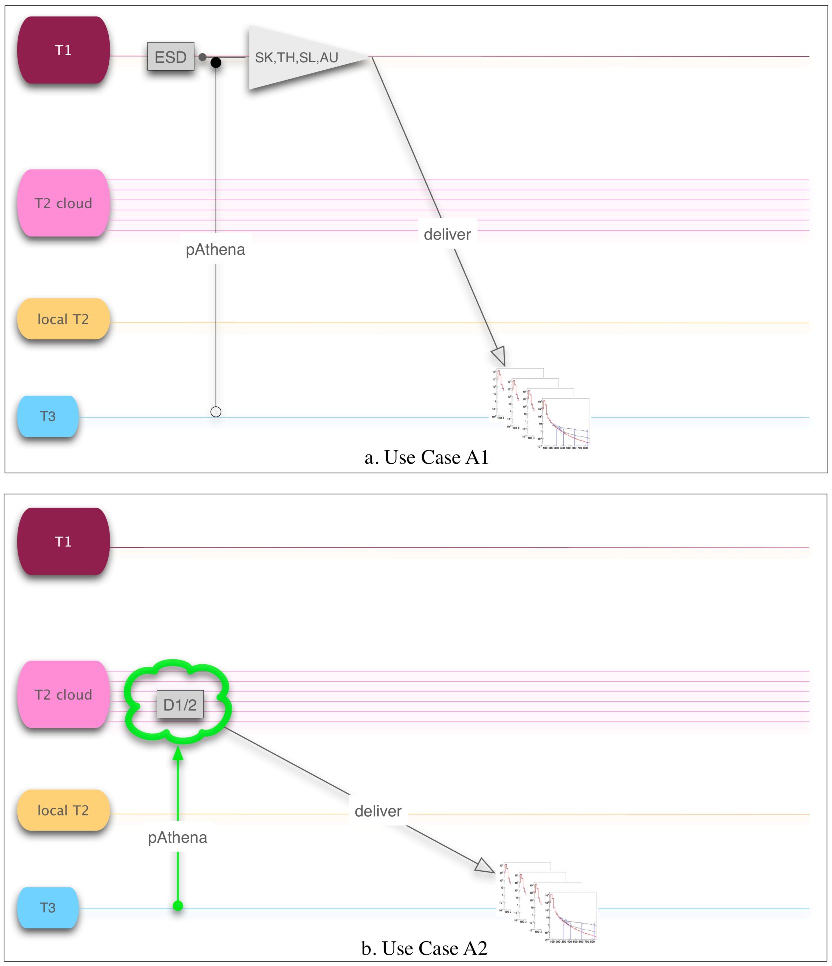
	data in:	data out:	from:	to:	by:	trans:	who:
A1	ESD	hist	T1	T3	T1,T2	SK, AU	analyzer
A2	D^2PD	hist	T2CL	T3	T2CL	SK	analyzer
A3	D^3PD	hist, txt	T3	T3	T3	AU, CH	analyzer
A4	D^3PD	hist, txt	T3	T3	T2CL	AU	analyzer
A5	AOD	hist	T2CL	T3	T2CL	SK	analyzer

It will almost never happen that an analyzer will submit a job to the grid to produce a `ROOTtuple`, bring it back, and then spend weeks working on only that file. Iteration will be required and therefore round-trip speed will be a premium consideration. The Use Cases envisioned for the naive use of Tier 3 centers are shown in Table 11. Use Case A1 ("A" for Analysis) is the situation in which an analyzer needs access to information only stored on the ESD. This could be for cell or hit level analysis, but also include the situation in which database access is required, and the quantities obtained are then added to a D^3PD for local analysis. Use Case A2 is a true Grid analysis—the paradigm analysis case envisioned for universities—where a user submits a `pAthena` request to the Grid for processing within the Tier 2 cloud, the job runs in multiple locations corresponding to the instructions and the data locations, and the results are returned in the form of histograms or a flat `ROOTtuple` to the Tier 3 for further analysis. (Of course, the result could be a D^3PD file as well, which would be Use Case C2.) Figure 10 shows these cases in pictorial form. Use Case A3 is the personal iterative analysis of `ROOTtuples` in order to produce plots.

4.5.1 Intensive Computing Use Cases

Use Case A4 is interesting as a computational challenge, but also as an historical example of how good ideas can greatly impact a Computing Plan. These sorts of projects were not imaginable even a decade ago and yet they are now ubiquitous in HEP analyses in which small signals best observed to be distinct from large backgrounds only through correlated kinematical distributions using a variety of multivariate techniques. Most familiar are Neural Network calculations, but on the rise are examples of so-called Matrix Element analyses. The latter are computationally intensive as they involve taking a measured event and comparing it to all of the ways that such

Figure 10: The Analysis Use Cases A1 and A2 both involved Grid-based recovery of flat `ROOTtuples` from analyses carried on at either the Tier 1 or the Tier 2 cloud. Use Case A5 is identical in principle to A2, with AOD substituted for D^1PD or D^2PD .



a. Use Case A1

b. Use Case A2

an event with its kinematical characteristics could have been produced by simulated events going all the way back to the “matrix element.” Each data event, then is mimicked by millions of simulated events which are suitably smeared for detector effects with unobserved final state variables numerically integrated over the multibody phase spaces. For top quark physics, this can be many final state jets and hence, many integrations.

CDF reports anecdotally that a recent Electroweak Top Quark search required CPU-centuries to analyze using this technique, while a current estimate within DØ for a 3fb^{-1}top mass determination requires 225×10^3CPU-h (just under a CPU-century). To cope with this impossible situation, DØ has instituted special Grid-based queues in order to farm these calculations to external sites, and relieve the central analysis facility (CAB, see below) from the task. Flexibility in both experiments made these analyses tractable.

While not an early running period calculational technique, Matrix Element calculations will almost certainly be a fact of life in ATLAS and the CPU cycles necessary in order to handle these calculations will be required from somewhere—and at levels which dwarf the Tevatron experience. In fact, with the leap in computing capability envisioned for ATLAS, even more exciting (read "terrifying") computational analysis techniques may become as important to ATLAS as the Matrix Element technique has become to DØ and CDF.

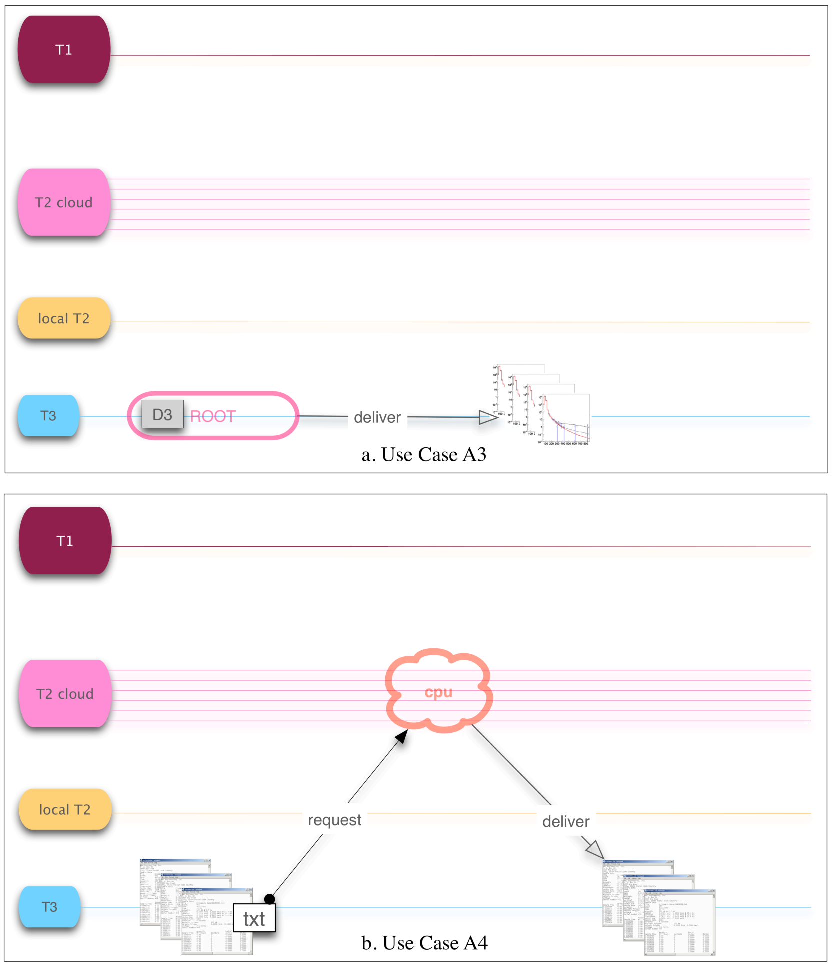
Other types of computationally intensive tasks similar in spirit to Neural Nets and Matrix Element calculations are becoming more prevalent: as computing capabilities go up, physicists think of ways to push these capabilities to the limit and thereby accomplish new things. Among other "meta" computing (analyses of analyses?) techniques are the generation of ensembles of pseudo-experiments, primarily for the study of systematic uncertainties and the critical sophisticated techniques for properly combining many multivariate analyses such as the COLLIE program within DØ. These are all similar in spirit: little or no data in and out, but literally cpu-centuries of computation in between.

Figure 11 suggests the Tier 2 cloud as the most likely source of computing for these calculations. In addition to Matrix Element analyses, enhanced fitting techniques are also extremely intensive calculations, many 100's of hours for a statistically limited analysis in DØ. These analyses are all basically the same in nature: almost no input (typically a small flat `ROOTtuple` or even a text file), almost no output, and essentially no network load. Just CPU cycles for hours on end.

4.5.2 Use Cases: Conclusion

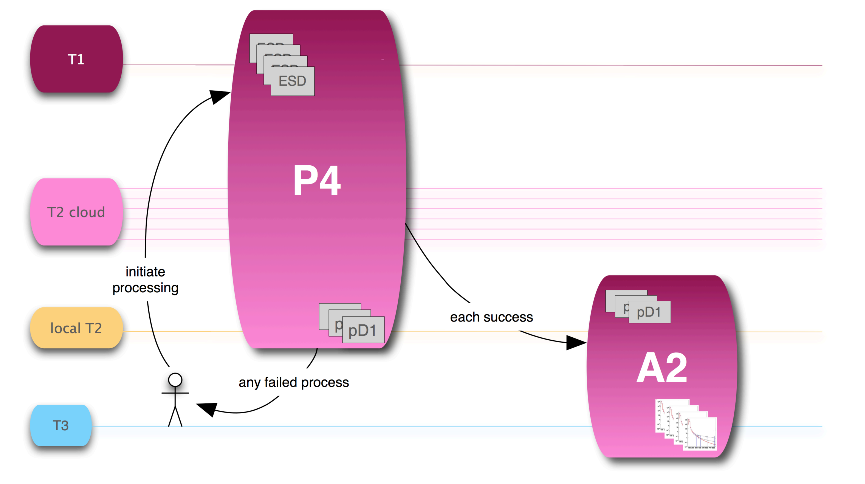
Any physics analysis (a Project) can be put together as as combination of the above Use Cases. For example, the Project of taking ESDs and creating `ROOTtuple` sets from them is a combination of use case P4 plus A2, as shown in Figure 12.

Figure 11: The Analysis Use Cases A3 and A4 involving naive, truely chaotic local analysis of `ROOTtuples` and the CPU intensive Matrix Element or fitting calculations requiring hundreds of hours of CPU cycles.



Taken together the Use Cases circumscribe a sobering set of responsibilities. Each can be characterized by the amount of computing and storage resources required and the network capabilities necessary to transport them around the world and across the country. It is a complicated dance which mixes the HLT heartbeat of continuous data flow from T0 through the Tier 2s (P1-P4, C1 and C2, and M1-M2) with the asynchronous personal needs

Figure 12: A Project is a combination of Use Cases. Here, a user pushes through enough jobs sufficient to create performance DPDs and then subsequently, flat `ROOTtuple` sets on his/her desktop. This is a combination of Use Cases P4 and A2.



(A1-A4) involving the Tier 2s and Tier 3s.

The unpredictability of A1-4 in both magnitude as well as frequency is where one aspect of where the cautionary "flexible" and "nimble" warning originates. The other, more critical, aspect is the astonishing burden that is placed on the Tier 2 centers. In the current plan, the Tier 2 centers form the critical junction, serving both the experiment as a whole through Monte Carlo production and critical dataset creation, but also their crucial connection to their local, national communities. Miscalculation in any part of their infrastructure—CPU capacity, disk storage capacity and availability, and network bandwidth and reliability—and the national analysis efforts will suffer, as the overall ATLAS-wide responsibilities are too significant to ignore.

Observation 4 *The Tier 2 systems' responsibilities are tremendously significant. Should we discover an underestimate in CPU, storage, or network needs of ATLAS as a whole, the analysis needs of U.S. university physics community will be adversely affected.*

5 The Tevatron Experience

If the past is any guide, any 2008 characterization of the ATLAS analysis model will not survive, unmodified. In fact, if the past is used as a model, "flexibility" should be an essential design criterion and an essential administrative guide. We have two experiences which are the most similar to the ATLAS situation: DØ and CDF.

5.1 Desconstruction of a DØ Analysis

> *...the scale of the software development effort for Run II is quite comparable to that of Run I. In Run II the system will again include multiple platforms of at least three currently supported flavors of UNIX and very likely some version of the NT operating system as well by the end of Run II.* "Run II Computing and Software Plan for the DØ Experiment," 1997.

NT?? Predicting the future is hard and when the future is a mixture of moving technologies, good ideas from physicists, and surprising problems, even experienced and well-meaning planners can miss the mark. DØ and CDF form our only experience with large, hadron collider analysis efforts. In many ways, they had to invent many of the approaches which we now take for granted and they certainly lived through at least four revolutions in computing: the ubiquity of OO software (necessitating rewriting of all code); the emergence of inexpensive, commodity computer clusters (necessitating the abandonment of large, expensive-maintenance, SMP [Shared-Memory multi-Processor] machines); the availability of distributed disk servers and management systems like dCache (encouraging the abandonment of tape-based storage systems for real-time analysis); and of course the development of high speed networking and switching technologies (creating the wholly new concept of grid computing).

Add to this mix of individual revolutions the invention and perfecting of the World Wide Web as, first a cute method of sharing flat information files, and now as an essential means of not only sharing information but controlling it. One looks back with amazement at the lifetime of these two 20 year old experiments and what they've witnessed and endured. Each has had to respond to the various evolutionary and revolutionary changes by reinventing what was presumed to be The Plan for how computing would be managed in the next phase. Responses were not always pretty and in many cases were pushed by users against entrenched technology, organizational, and management choices.

Observation 5 *Is there any reason to think that the first 20 years of the ATLAS computing experience will be any less astonishing? Is it wise to design tightly to current expectations, as if the future will be a continuous extrapolation of the present? If history is at all a reliable guide, it argues for the most flexible, most modular, and least rigidly structured systems consistent with 2008 technology and budgets.*

In order to set the scale, Table 12 from Boehnlein [2] should be sobering. It shows an experienced projection of the DØ expectations for computing against the actual situation a decade later. These 1997 projections were done with the entirety of Run I tevatron experience in hand. And yet, with all of that wisdom, crucial quantities were underestimated. Especially surprising should be the large increase in required analysis disk and the difficult increase in reconstruction times. The former was surely due to the user need for on-demand event processing (notice the reduction of tape storage per year over expectation), which in turn was a result of improved analysis techniques and probably the repeated analysis that comes from systematics-dominated signals. The latter was due to an overly optimistic expectation for just how difficult tracking would be in an busy, event-overlapped environment. Of course, the explosion of remote site computing was again, a user need which was largely accommodated by funding opportunities. We emphasize that this was an honest appraisal of what happens in a research environment, with constantly shifting ground. That this many different revolutions and surprises could be managed and brought to bear on enormous analysis tasks is a testimony to both the skill of the Fermilab computing professionals and flexibility that was eventually built into the experiments' tiers.

5.1.1 A Story: DØ Infrastructure Evolution

It is perhaps instructive to compare the DØ "tiers" with those planned for ATLAS and Table 13 shows the closest analogy to the planned ATLAS tiered system. The Reconstruction farm is a stand-alone facility doing basically one thing: taking raw data and processing it into the usable data formats suitable for DØ physics analysis. This includes preparing the 27 exclusive skims, which are then later combined into 14 logical skims. The Reconstruction farm is essentially identical in spirit to the ATLAS T0 center.

Table 12: Comparison of the 1997 Computing Plans for the DØ experiment looked at from 2006 [2].

	1997 projections	2006 actual
Peak (average) data rate (Hz)	50 (20)	100(35)
Events collected	600M/year	1500M/year
Raw Data Size (kB.event)	250	250
Reconstructed Data size(kB/event)	100	80
User format (kB/event)	1	40
Tape Storage	280 TB/year	1.6 PB on tape
Tape reads/writes (weekly)		30 TB/7TB
Analysis/cache disk	7 TB/year	220 TB
Reconstruction time (GHz-s/event)	2.0	50
User analysis times (GHz-s/event)	?	1
User analysis weekly reads	?	3B events
Primary reconstruction farm size (THz)	0.6	2.4 THz
Central analysis farm size (GHz)	0.6	2.2 THz
Remote resources (GHz)	?	$\sim$ 2.5THz

Table 13: The DØ experiment "tiered" computing clusters and the closest ATLAS analogs.

	reconstruction farm	CAB cluster	CLuED0 cluster
DØ	400 nodes dedicated batch	1126 nodes, 2 clusters 5198 job slots batch	500 nodes interactive & batch
ATLAS	$\sim$ Tier 1?	$\sim$ Tier 2's?	$\sim$ Tier 3's?

The CAB (Central Analysis Backend) was, like the whole analysis effort, added after the fact. The need for a commodity-processor batch system was not envisioned originally and had to be created after much user demand and growing costs of maintenance for the entrenched SMP system. As a batch-only, large computational and storage cluster, CAB is essentially functionally similar to the ATLAS Tier 2 systems.

Finally, CLuED0 ("Clustered Linux Environment for D0") is an interactive cluster which is a user-owned, user-managed desktop system which has home directories, a fair-share disk storage system, and limited batch queues. It has a special relationship with CAB, as there is an integrated set of scripting tools which facilitate automatic submission of batch jobs from CLuED0 to CAB. CLuED0 matches very closely the idea behind the ATLAS Tier 3 tier, as both a locally-owned hardware system, and because of the problematic nature of user-generated support.

Neither CLuED0 (which came first) nor CAB were planned in the sense in which they evolved. This was both for technical and financial reasons which probably could not have been foreseen. Each faced initial resistance, as they were not in the original planning and because they required modifications to maintenance and security strategies. CLuED0 in particular was a grass-roots creation which created interesting challenges to management. It was necessary, and so the independent analyzers prevailed and it is the primary physicist platform today. It should be noted that CLuED0 has a much tighter system management structure now than it did when it was first created. Its success is in direct proportion to the eventual buy-in by the Fermilab Computing Division and experiment management. Expert system management evolved along with the original, "renegade" user-creators and everyone is very satisfied now.

5.1.2 The Story Continues: DØ Data Formats

Evolution of data format within DØ was a complicated story as well. There was a "DST" format, which is somewhat like the ATLAS ESD in scope, but more like the AOD as it was expected to be the "workhorse" format, one step from `PAW` ntuples. However, it was too unwieldy for many purposes, and people kept inventing their own, smaller, closer-to-them formats which led each physics group into different, non-overlapping directions. (Remember **Observation 2.**) What grew instead was the TMB ("thumbnail") format from a TAG-like object of 5kB per event, to 20, and then 70kB/event. TMBs are the paths that analyzers use in order to obtain cell/hit information.

Table 14 shows the DØ data formats and a close match to their ATLAS

Table 14: The DØ experiment data formats and the closest ATLAS analogs.

	RAW	DST	TMB	CAF
DØ	1MB	100 kB	70 kB	40 kB
ATLAS	RAW	$\sim$ ESD	$\sim$ AOD	$\sim$ D^1PD

counterparts. One could argue about the ESD designation in favor of AOD as the closest to the DØ TMB. One argument in favor is an important one: the TMB contains hit/cell information which makes on-the-fly reprocessing (called "fixing" in DØ parlance) possible. Currently, the smallest format in ATLAS in which this can be done is the ESD, although even this plan is evolving within ATLAS as some cell-level electron information is kept within the AODs, so Table 14 assigns them as analogs. The growth in size of the TMB in DØ was, in part, the need to include this information, which is not present in the CAF format. That the CAF and TMB data are in parallel available allows for "re-CAFing" based on fixing, without a whole experiment-wide preprocessing.

But, going hand-in-hand with the TMB evolution was the need to condense the many independent data structures into a common form. Each physics group had evolved its own `PAW` and eventually `ROOTtuple` structures which greatly inhibited collaboration. While data formats were common at an initial state, the actual group-level selection and analysis took place at the `ROOT` level and were the domains of the physics groups themselves. People "voted with their feet" to find the fastest analysis path, which pointed directly to home-grown formats. In 2005, by management fiat, a common CAF[10] ("Common Analysis Format") structure was designed and imposed on the physics and analysis groups, after considerable wasted time. To go along with the CAF data format, the CAFe ("CAF-environment") set of tools, was created, tailored to the available hardware making common tasks simple. The whole structure is an OO, `ROOT` -based TTree structure, now common at a deep level among the physics analysis groups.

None of the above were in the original DØ analysis plans. The original TMB was supposed to be lightweight, and not suitable for full physics analysis. It was too small, but it got larger in time but eventually the unpacking step was too slow for interactive analysis. The DST was meant to be for

[10] Note, there are two uses of the acronym "CAF". The Common Analysis Format refers to the DØ data format, while the Central Analysis Facility refers to the CDF batch cluster, described below. We presume that the context will distinguish these two CAFs

analysis, but it was too big. The analysis hardware was meant to be a large, SGI, SMP batch system with satellite NT workstations for user `ntuple` analysis. However, maintenance and upgrade costs were prohibitive and locking into a single vendor technology meant that taking advantage of increasing processor speeds of commodity PC's was impossible.

So, neither the hardware nor the thoughtfully produced software plans were sufficient for DØ analysis needs and the analyzers sometimes had to move faster than the bureaucracy was able to respond . Out of that was born CAF, TMB, CLuED0, and CAB. Laboratory and experiment support came around and the DØ analysis system is now robust, flexible, and responsive to the unexpected breakthroughs in analysis techniques.

Observation 6 *Physics analysis moves fast, at a rate which is often more rapid than can be tolerated by a rigid computing structure or system management. Analyzers will sometimes take matters into their own hands when a bureaucracy is perceived to be in the way.*

5.1.3 A Happy Ending: A DØ Analysis

One of the computationally intensive analyses at hadron colliders is that of the current attempts to detect the signal for Electroweak production of single top quark events over an enormous background. The signal is the production of high a p_T lepton, significant missing energy, one (or two) tagged B mesons, and 2-3 high-p_T jets and so the signal looks exactly like some $t\bar{t}$ channels, QCD production of W bosons plus heavy flavor, and misidentified "normal" QCD jet production. The cross section at the Tevatron for this process is approximately 3 fb and at the LHC it is 100 times that. At the Tevatron both the uncertainties in the signal and some of the background determinations are statistically limited. At the LHC, most measurements will be systematics dominated, placing an even higher burden on the computing necessary to perform these analyses.

As a measurement dominated by backgrounds and heavily dependent on event topology, considerable effort goes into generating signal and background samples from full-event Monte Carlo and relying on data for other backgrounds. This requires considerable skimming projects in order to select the samples appropriate for data-Monte Carlo comparison, tuning weightings, and tuning topological and kinematical cuts. The separate reactions required include: a separate skim for QCD backgrounds which come from the same original data as the signal[11], but with nearly orthogonal selections; individual generated signal samples for each final state topology; and the generation of 45 separate Monte-Carlo backgrounds. Table 15 shows the complete set of numbers of files, numbers of events, and numbers of submitted jobs in order to make a single, complete pass through the whole sample. This exercise, during about a year long period, happened just about every month.

Compounding the juggling of files and datasets, there were two separate reconstruction program versions to cover the whole time period over which this measurement is taking place. All of this work was done on the CAB, and because of the number of jobs required, it took the graduate students about a day to get the events successfully through the system, and about two days to put the whole package together for comparison with the data.

This kind of human-intensive activity is often lost in the prediction of what is involved in a large-scale analysis. The realities of sharing of queues, the vagaries of network reliability, mistakes, and time-outs when simultaneous reads of input files lead to clock times which are considerably longer than just a naive calculation of CPU times for any such project. Figure 13

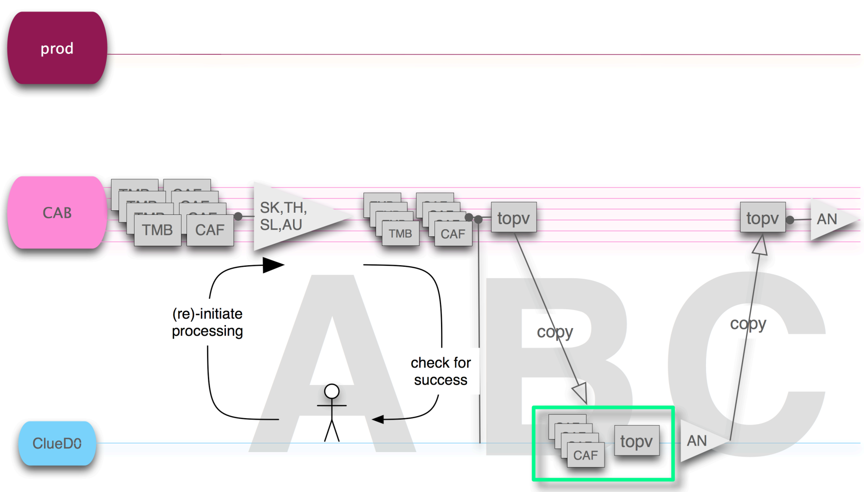
Table 15: The numbers of files, jobs, and events processed each time the DØ single top anaysis is run through a re-selection round. This happened almost every month during the early analysis design, and has happened even at a mature analysis stage: during the DØ internal review toward publication.

source	files	events	jobs
data	96k	1600M	2400
QCD background	96k	1600M	2400
signal MC	25.6k	200M	2400
bckgnd MC	12k	120M	560
total	240k	3B	8000

shows a sketch of this single analysis. The step "A" is what was just described: the over-and-over submission of 8000 job requests to the CAB involving the access to 240,000 files...monthly. The rest of this analysis, "B" and "C" in the figure, involve the regular chaotic analysis—on the DØ "Tier 3" of CLuED0—of manipulating cuts, displays, selections, and Monte Carlo

[11] An early, but significant modification in top quark analysis was the decision to use data, and to not rely on simulation, to estimate the QCD backgrounds in top quark analyses. It is a perfect example of the physics driving an analysis in an unanticipated direction, thereby impacting computing.

Figure 13: The Project for the DØ single top quark analysis includes a number of steps. Especially time-consuming and computationally intensive, was the skimming within the (enclosed) CLuED0-CAB grid. Even in that tight environment, failed jobs, timeouts, etc. required continuous monitoring and job resubmission.



data comparisons. During the later stages of this analysis, a separate set of files (the "topovars" in the figure) are refined and submitted back to the CAB for the extensive Boosted Decision Tree analysis. Typically, these decision tree analyses take about 10 hours per job, for approximately 500 jobs submitted.

The bottom line to this story is the reality of an unusually intense analysis is that:

- Thousands of jobs submitted;
- on a periodic basis;
- involving hundreds of thousands of files and billions of events;

- with a very person-intensive monitoring and resubmission;
- and an I/O non-intesive, but computationally significant fitting exercise is not unusual.

Note well: this extensive package of projects is before the first systematic uncertainty has been probed. And, this is for one of a hundred analyses within just DØ.

Observation 7 *Full-scale, precision analyses will be a huge load on the Tier 2 structure from the perspective of computation and file-access. Monitoring and resubmitting failed jobs will surely continue to be a serious complication for analyzers. If history is a guide, current predictions of how this maps to the ATLAS analysis future are sure to be underestimated.*

5.2 A CDF Analysis

As is the case with DØ, there are many computationally intensive analyses in CDF, including the search for single top quark production in ppbar collisions at the Tevatron. One particularly intensive CDF analysis is the search for Higgs boson production and decay into W boson pairs which both subsequently undergo leptonic decay [1]. Although the background and possible signal contributions will be quite different at the LHC (e.g. $gg \to H \times 100$, $gg \to WW$ non-negligible) leading to different analysis challenges, there are important lessons to be learned in terms of the computing challenges and types of processing steps which will be involved. Like the single top analysis, experimentalists are confronted with finding a very rare signal possibly buried beneath a mountain of Standard Model (SM) background from many different sources, the dominant of which looks sufficiently like the signal that we need multivariate techniques to statistically distinguish the two. The limits we obtain are perceptibly impacted by our systematic uncertainties and so a thorough treatment of them using computationally intensive pseudo-experiments is required.

The central processing starts with PB of raw data and necessary Monte Carlo samples on tape. A large production farm runs managed production (reconstruction) on these which creates data containing high-level objects like tracks, jets, muons, EM clusters, etc. analogous to the ATLAS ESD/AOD. This data is that further processed into one of two "standard" `ROOT` Object-based formats called Stntuple which contained even higher level objects convenient for analysis. In 3fb^{-1}, the total size of the Stntuple we worked with (high pt electron, muon, and jet streams) amounted to tens of TB. We further processed the Stnttuples to skim, thin, and augment with derived information based upon refined calibrations the data into a custom (by the analyzers) ROOT I/O-based format we called Dbntuples. These Dbntuples were approximately a TBs in total and drove a number of heavy diboson analyzes (*WW, WZ, ZZ*). Finally, the Dbntuples were processed into a ROOTtupleformat for plotting, MVA input, and systematic variations in analogy to the ATLAS D^3PD format and anticipated usage. These "summary ntuples" amounted to tens of GB in size and were the samples we worked most frequently with and also generated most frequently.

The reconstruction and Stntuple generation we centrally managed on dedicated resources. We did all Dbntuple and summary ntuplegeneration, limit calculations via pseudo-experiments, systematic variations, MVA calculation (Matrix Element) and neural net training on the CDF Central Analysis Facility (CAF) at Fermilab using our own resource shares that were based upon equal-share rules. Its important to point out that the central production was very rare (say 1-2 times per year at most) while the later stages of the processing were done very frequently, in some cases a few times per week. In addition, some of this later processing is almost exclusively computational (e.g. limit calculation or Matrix Element calculation that can take approximately a minute per event) such that it is does not require high bandwidth access to data handling services. In fact, running on the CAF which has such high bandwidth access to data is a waste of precious resources since batch slots are limited. Every effort was make to avoid wasting these resources.

A lesson here is that there are likely to be lots of processing steps in the analyses (the CDF approach here is far from ideal) and the later steps will need to be done many times. The resources required vary wildly, from skimming/thinning/slimming-like jobs requiring high-bandwidth access to data handling services to tasks that are purely computational but very substantial nonetheless. The ATLAS Tier 3s can play an important role in ensuring that the very substantial later stages of analysis processing happen close to the analyzers rather than taking up precious Tier 2 resources because there is no other recourse. It is also very important that any estimation of computing requirements accounts for these later processing steps because even though they involve much less data than the AODs, they have potentially huge multipliers.

5.2.1 Evolution of CDF Analysis Computing

In many respects, CDF Run II analysis computing evolved independently in a way analogous to DØ, indicative of the common challenges each experiment faced. Before late 2001, CDF computing was mired in the use of a large SGI SMP machine which served interactive login, batch jobs, and data handling for the collaboration. It became increasingly clear that this model did not scale, with a large number of users (hundreds) both running internet browsers on the SMP and other interactive uses and trying to analyze the increasing large volume of data and simulation that CDF was generating. A review of the analysis computing was undertaken within the collaboration and a new model based on a large farm of commodity (→ cheap) hardware running Linux and operated in batch mode (insulated from interactive use) emerged. In addition, several hundred TB of commodity TB file servers operating as a cache-layer (running dCache) in front of the Enstore-based tape system was deployed.

At the time, standard GRID tools were emerging but were at such an early stage as to be essentially unable on the scale the CDF collaboration required. In response to this situation and a growing need for usable analysis computing to analyze the CDF data set, a custom job management system for submission, authentication, and sandboxing based on kerberos-aware python was developed. This approach was initially ridiculed by many in both CDF and also DØas being arcane, simplistic, and "going down a road we've been down before with other custom projects." Being physicists interested in getting our physics done and not computer scientists focused on elegance and longevity, we did what it took to make the system work for doing physics. Thus was born the CDF Central Analysis Facility (CAF) and it worked (and continues to work). In my respects, one can argue that it represents the first production GRID in operation. In terms of data handling, we employed dCache as a cache layer in front of the Enstore tape system, with SAM later added but used only for its data cataloging services. Dzero followed suit with the CAB and used SAM as it was designed to be used (i.e. a data handling system). The CDF CAF and analysis model has evolved significantly since then, toward more standard GRID software like Condor-G (and encapsulated glide-in capabilities).

Of course, GRID tools like those available with Open Science Grid (OSG) and employed by U.S. ATLAS are far more evolved then back in 2001 when the CDF computing model was reworked. The lesson here is that physicists will do what is takes to have robust access to data and get their physics done. It is also worth noting that GRID monitoring was a deficiency throughout.

Again, custom tools based on python and RRD had to be developed within CDF to provide users the information they require. This information goes beyond simple status information. Historical information was very much needed, mostly for planning purposes but also, of course, for debugging problems. For example, we attempted to provide an estimate of future job completion time based on current system load but also past history of execution times. The biggest complaint users had was in the spirit of the following: "I've been able to run my jobs in a week over the last month, but now it is taking several weeks to complete my job and I have to give a presentation in Physics Group X on Friday..." The ability of physicists to plan is very important to what we do, and adequate monitoring capabilities is critical to achieving this end.

In summary, the sooner that the full computing model can be exercised with realistic use cases and at the required scale, the better to avoid unforeseen deficiencies requiring a deviation from the baseline computing model to get physics done. In many respects, the work of this Task Force and the recommendations therein are driven by a desire to exercise the analysis computing model as thoroughly as possible, design in flexibility where possible, develop contingency for unforeseen circumstances, and broaden the knowledge base for analysis computing of collaboration as a whole.

6 Modeling

In order to explore the degrees of freedom inherent within the U.S. ATLAS structure, we have performed some simulation within acknowledged parameter variations. We do not expect that these calculations are precise. They are meant to give an impression of whether the system is flexible against reasonable extrapolation to the unknowns which are inherent in this kind of research. Where possible, we justify our parameters. Where not, we try to motivate them with appropriate caution.

Our model assumes that that the responsibilities listed in Section 3.1.4 and our focus will be on Monte Carlo Production, presumed to be solely a Tier 2 responsibility.

6.1 The Calculation

The deployment of ATLAS's Computing Model has yielded a complicated multi-tier system composed of hundreds of GRID sites scattered around the world. We have made an attempt to balance the sophistication of our model of this system against the goals of our calculations so that our results may be easily understood, yet are quantitatively accurate. Therefore we employ several inherit simplifications in our model:

- We perform a calculation, not a simulation.
- This calculation is steady-state, representing a snap-shot of the load on the computing systems.
- We choose the total run-time of specific series of jobs as our figure of merit.

There are four basic components in our model:

1. A *resource* is class or tier of sites. For example, all Tier 1 sites are considered one resource. For our model, the most important parameter associated with resource is the CPU cycles it provides, measured in kSI2K.
2. A *transformation* is a processing step with specific inputs and outputs. For example, AOD→ DPD is one transformation. Many parameters are associated with a transform, including number of input/output events, processing CPU (in kSI2K sec) required per event, and the per event input/output data size.
3. A *chain* is a series of transformations where the output of one step is the input to the next. For example, the Monte Carlo production chain consists of Nothing → GEN → SIM → DIGI → ESD/AOD → D^1PD .
4. Since ATLAS reserves a fraction of certain resources for production activity, we also introduce the concept of *queues* for each resource. A queue is a fixed fraction of the CPU at a resource coupled with a scheme for sharing this CPU with transformations (more details below). Every resource specifies what queues it offers. Every transformation specifies which resources and queues it will use.

The critical feature of the computing system which our model must reproduce is the sharing of resources between all transformations. Clearly, the more transformations running in the system, the more time it will take for every transformation to complete. We ensure reproduction of this behavior in the calculations behind our model, which is the result of the following sequence of steps:

1. User specifies the resources.
2. User specifies the chains running in the system. Each chain consists of a set of transforms.
3. Each transform calculates how much kSI2K sec of CPU it requires to complete.
4. Transforms are collected from chains, and assigned to the specified resources/queues.
5. Queues assign a fraction of their CPU to each transform.
   - An *analysis* queue divides CPU evenly between all transforms.
   - A *production* queue gives each transform CPU resources which are proportional to the kSI2K sec required to complete the transform. The effect is that all transforms in a given queue will take the same time to complete.
6. Transforms divide the required kSI2K sec of CPU by the CPU provided to them in order to calculate how long they will take to complete. Disk read/write times added to this time by properly comparing the I/O rates (based on the CPU processing rate and the input/output file sizes) with the maximal single job IO rates (assumed to be 10 MB/sec)[12].
7. In order to estimate the data-flow between resources (eg Tier 1 $\rightarrow$ Tier 2), chains note when sequential transformations are executed on different resources, and report the minimum transfer rate necessary in order to not stall processing at either resource. We assume that sufficient bandwidth is available for so that transfers are not stalled.
8. Chains pull results from transforms, producing a summary.

6.2 Example Calculation: Monte Carlo Production

Figure 14 shows the output of the modeling of the Monte Carlo chain which consists of five transformations:

[12] Our model can also account for maximal site disk input/output rate and addition CPU processing required for turning persistent/compressed data into transient/uncompressed data. Presently these factors are assumed to be accounted for in other parameters and IO no per site IO limit is imposed.

1. Nothing → Generated Events,
2. Generated Events → Simulated Events,
3. Simulated Events → Digitized Events,
4. Digitized Events → Reconstructed Events (AOD/ESD),
5. AOD → Primary Derived Physics Data (D^1PD).

The first and last transformations are run on Tier 1 production queues, the remainder are run on Tier 2 production queues. In the shown example, 100% of Tier 1 resources are allocated to the production queue which is also populated with the re-processing chain (not shown). 80% of the Tier 2 are allocated for production. The various parameters which are input into this calculation are presented in the following sections.

Each transform reports the CPU required (in kSI2K sec) and provided (in kSI2K), the input/output data size (in KB), and the total time required to run. Note that because production queue allocation described in the previous section, all transformation running on the same resource take approximately the same time[13]. Since we assume that all steps of the chain are running simultaneously, the "Chain Max" parameter, which is the running time for the slowest transform, is the total time for the chain to complete. If each transform was run after completion of the previous step, "Chain Total", which is the sum of all running times, would be the total time for the chain to complete. Finally the flow volume/rate parameters reflect how much data is moved between resources and the required rate in order to not stall any transformation. In the example, the Tier 1 → Tier 2 flow reflects movement of generated data, and the Tier 2 → Tier 1 reflects the movement of AOD back to Tier 1 (for D^1PD production).

6.3 Input Parameters

Table 16 summarizes the some of parameters which were used for modeling of Monte Carlo production. The most relevant are the simulated number of events (product of the annual recorded dataset and the fraction simulated) and the per event time for each step of the simulation chain. Note that since pile-up events are mixed into the Monte Carlo during digitization, this time must be multiplied by an instantaneous luminosity-dependent factor.

[13]The model also accounts for the time required to read/write data. This additional time, which is typically small for non-analysis jobs, is not accounted for when queues provide

Table 16: Various parameters used in the simulation and later in the text.

quantity	value used	high	low	comments
LHC year	2010	2011	n.a.	assume 2008 start
Ins. $\mathcal{L}$ $\mathrm{cm}^{-2}\mathrm{s}^{-1}$	2×10^{33}	3.5×10^{33}	10^{33}	Garoby, LHCC 08
annual $\int \mathcal{L}dt$ fb^{-1}	10	?	?	rounded from 12
annual dataset	2×10^9 events	?	?	[9]
sim. time	1990 kSI2K s ($t\bar{t}$)	2850 kSI2K s γj	1030 kSI2K s $W \to \mu$	[19]
dig. time	29.1 kSI2K s ($t\bar{t}$)	29.2 kSI2K s j	23.1kSI2K s $W \to \mu$	[19]
reco. time	47.4 kSI2K s ($t\bar{t}$)	78.4 kSI2K s j	8.07 kSI2K s $W \to e$	[19]
digitization pileup factor	3.5	5.8	2.3	[19]
fraction of full dataset for full sim	0.1	0.2	na.	
factor rel. to full sim. for $t\bar{t}$	0.05 (ATLFAST-II)	0.38 (fG4)	0.004 (ATLFAST-IIF)	[19]
$\mathrm{D^1PD} \to \mathrm{D^2PD}$	0.5 kSI2K s	?	?	[18]
$\mathrm{D^2PD} \to \mathrm{D^3PD}$	0.5 kSI2K s	?	?	[18]
disk R/W	100 MBps	200 MBps	10 MBps	S. McKee private
sustained network	50 MBps	100 MBps	10 MBps	S. McKee private
fraction of data in pDPD	20%			
# primary DPD	10			
# subgroups	5			
average CPU	1.4 kSI2K units	2	NA	
total ATLAS Tier 2 computing	60.63MSI2k			[13]

```
Monte Carlo:
    (Nothing)--> [Generation (Monte Carlo)]--> (Gen)
      NEvents:  200000000.0  CPU Needed:  46000000.0  CPU Provided:  31.7
      In:  0.0 ( 0.0 ) Out:  10.0 ( 10.0 )
          Total Time: 16.796 ( 16.8 ) days, IO/CPU Fraction: 0.0
    (Gen)--> [Simulation (Monte Carlo)]--> (Sim)
      NEvents:  200000000.0  CPU Needed:  400000000000.0  CPU Provided:  45894.9
      In:  10.0 ( 10.0 ) Out:  2000.0 ( 2000.0 )
          Total Time: 101.0 ( 202.08 ) days, IO/CPU Fraction: 0.0
    (Sim)--> [Digitization (Monte Carlo)]--> (Digi)
      NEvents:  200000000.0  CPU Needed:  13340000000.0  CPU Provided:  1530.6
      In:  2000.0 ( 2000.0 ) Out:  2000.0 ( 2000.0 )
          Total Time: 101.2 ( 202.41 ) days, IO/CPU Fraction: 0.0
    (Digi)--> [SimReconstruction (Monte Carlo)]--> (SimESDAOD)
      NEvents:  200000000.0  CPU Needed:  9400000000.0  CPU Provided:  1078.5
      In:  2000.0 ( 200.0 ) Out:  1000.0 ( 100.0 )
          Total Time: 100.9 ( 201.8 ) days, IO/CPU Fraction: 0.0
    (AOD)--> [AOD-> \d Making (Monte Carlo)]--> (\d)
      NEvents:  200000000.0  CPU Needed:  1120000000.0  CPU Provided:  617.5
      In:  150.0 ( 150.0 ) Out:  150.0 ( 150.0 )
          Total Time: 21.002 ( 252.03 ) days, IO/CPU Fraction: 0.0
Chain Max: 101.21 ( 252.03 ) days, Chain Total: 340.94 ( 875.1 ) days,
           IO/CPU Fraction: 0.0 ( 0.0 )
Flow Volume (TB):  {'Tier2->Tier1': 27.9
                    'Tier1->Tier2': 1.86}
Flow Rate (MB/sec):  {'Tier2->Tier1': 3.36,
                      'Tier1->Tier2': 0.22}
```

Figure 14: Example output from the Monte Carlo chain.

6.4 Estimating Required Monte Carlo Production Resources

In order to demonstrate the relative importance of various input parameters, table 17 lists several illustrative calculations of various Monte Carlo production scenarios. The calculated figure of merit, which is reported in the last column, is the minimum number days required for the full Monte Carlo production pass. Comparing calculation 1 and 2, we see that luminosity dependence of digitization (described above) is negligible for luminosities up to 10^{33}. Calculation 3 shows that roughly 20% recorded ATLAS data can be fully simulated in one year, provided 50% of Tier 2 resources are dedicated to Monte Carlo production. In comparison, 100% of the recorded data can be fast-simulated in less than one-half of a year with the same resources (calculation 4). Therefore, as calculation 5 shows, ATLAS can perform 10% full simulation, 100% fast simulation with 50% of Tier 2 resource dedicated to production. Finally, calculations 6 to 9 illustrate that more than 90%

CPU to transforms. This small inconsistency results in nearly negligible difference between transform run times in production queues.

of Tier 2 resources will be required for production for 10% full simulation, 300% fast simulation, a scenario which some may argue is more in line with realistic physics analysis needs.

Table 17: Illustrative calculations described in the text.

Calculation	Tier 2 Production Fraction	Simulation Fraction	Fast Simulation Fraction	Luminosity	Time (days)
1	50%	10%	0%	1×10^{32}	159
2	50%	10%	0%	1×10^{33}	162
3	50%	20%	0%	1×10^{33}	323
4	50%	0%	100%	1×10^{33}	166
5	50%	10%	100%	1×10^{33}	328
6	50%	10%	300%	1×10^{33}	660
7	75%	10%	300%	1×10^{33}	443
8	90%	10%	300%	1×10^{33}	371
9	100%	10%	300%	1×10^{33}	336

In order to properly estimate the fraction of Tier 2 resources necessary for simulation production, we ran our calculation repeatedly, scanning the Tier 2 production fraction, and the full and fast simulated fraction of the collected data (for the year 2010). Figure 15 shows minimal percent of all ATLAS Tier 2 CPU resources required to be able to simulate a given full and fast fraction of collected data in one year.

6.5 Modeling Analysis

While our model of the ATLAS computing systems can reliably handle simultaneously running of a variety of analysis chains, we found it difficult to guess what analysis models will be chosen by ATLAS physicists, how many of every type of analysis will be running at a given time, and what resources would be required for the steps of such analyses. Without a running experiment, it is nearly impossible to build a model of all ATLAS analysis activity.

In order to simplify the problem, we designed a single illustrative analysis chain based on DPD-making and ROOT-analysis studies performed by Akira Shibata [18] summarized in Table 18. The most important behavior observed in these studies is that the event processing rate for a given DPD making job is a function of size of event data read/written. The more data

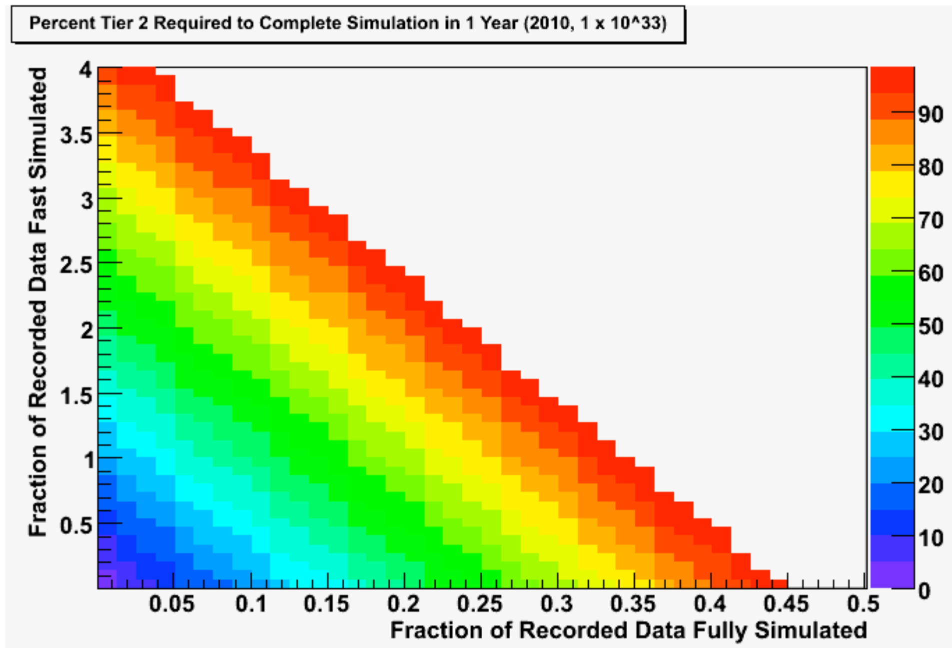


Figure 15: Percentage of Tier 2 CPU required for simulation production as function of fraction of 2010 recorded data which is fully and fast simulated.

required for a job, the more time required to read that data and the more operations performed on those data. In addition, ROOT analysis was found to be approximately 20 times faster on D^{3}PD (flat-ntuple) versus `POOL` based DPDs, with a large dependence on the language, compiler, and framework employed in the analysis software.

Based on these findings, we constructed an analysis chain consisting of the following transformations:

1. D^{1}PD $\rightarrow$ D^{2}PD : The D^{1}PD is 25 KB/event, and contains 10% of all recorded, full and fast simulated data. We assume 10$ full simulation, 300% fast simulation. The outputed D^{2}PD contains augmented information, resulting in a size of 30 KB/event, but no additional skimming or thinning. This step most closely corresponds to the Top D^{1}PD entry of Table 18, which was found to run at 3 Hz, independent of input (AOD or D^{1}PD).

2. D^{2}PD $\rightarrow$ D^{3}PD : The output is 10 KB/event and no skimming is applied. This step most closely corresponds to the Top D^{3}PD entry of

Table 18: Summary of DPD making studies performed.

DPD Output Name	DPD Output (KB)	AOD Input Rate (Hz)	D^1PD Input Rate (Hz)
None	0	96	255
Very Small D^3PD	0.37	84	198
Small D^3PD	0.71	43	63
Top D^3PD	4.9	14	N/A
Very Small D^2PD	1	10	10
Small D^2PD	18.7	8	10
Top D^1PD	31.4	3	3

Table 18. However since the output is larger (10 KB/event rather than 4.9 KB/event) we estimate an event processing rate of 10 Hz for this step.

3. $D^3PD \rightarrow$ Plots: considering the various rate found in [18], we believe that 10000 Hz is a fair estimate of event processing rate for this step.

While this particular set of transformations may not represented a likely analysis chain, we hope that the analysis load on the ATLAS computing system is well represented when we allow for multiple instances of this chain to occupy the system.

Our primary goal is to estimate the number of analyzers which Tier 2s can support. Based on results of Section 6.4, we assume 80% of Tier 2 resources will dedicated to Monte Carlo production, and the remainder be available for analysis. Then we consider 2 scenarios:

- *Independent*: Every analyzer runs every step of the chain.
- *Cooperative*: Analyzers cooperate, sharing DPDs when possible.

Figure 16 plots the time taken for one iteration of the analysis chain as a function of number of simultaneous analyzers, assuming all analyzers work independently. Considering that D^1PDs will be made monthly, this iteration time must be less than 30 days. If we consider multiple iterations and other concerns, 10 days is likely a more reasonable time between availability of D^1PDs and an analyst's extraction of their first "final" plots. Our model therefore shows that ATLAS can only support about 10 independent analyzers. Note that in this scenario, $D^1PD \rightarrow D^2PD$ is the most time consuming task. Because of the analysis queue resource sharing with the 2 other transformations, one-third of the 20% of tier 2 analysis resources are dedicated to $D^1PD \rightarrow D^2PD$ jobs. If the other transforms could be moved to other resources (e.g. Tier 3s), then the Tier 2s could support 30 different $D^1PD \rightarrow D^2PD$ transforms which would complete in 10 days.

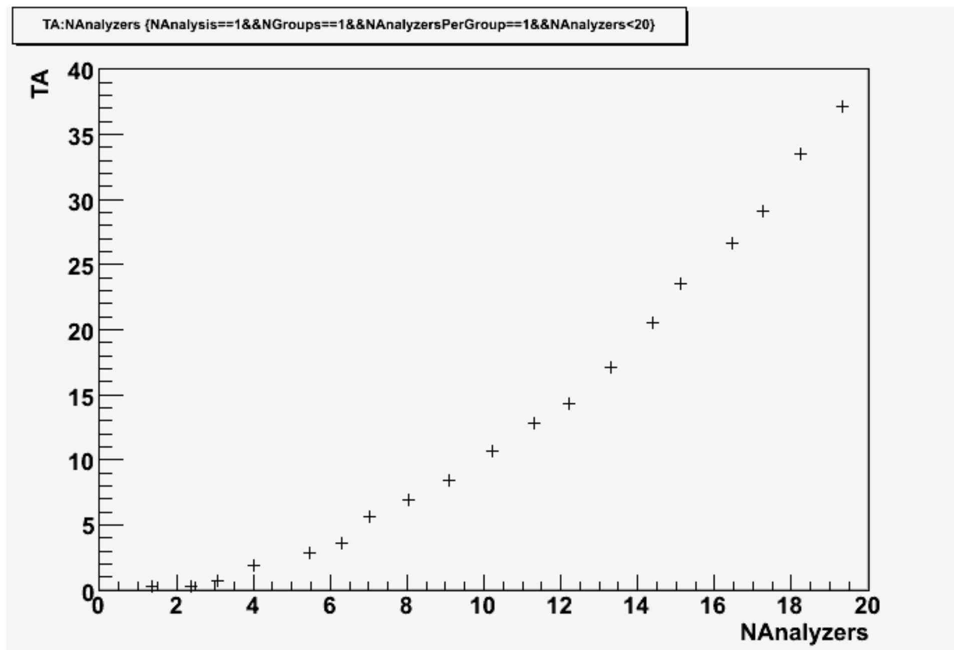


Figure 16: Time required (in days) for a single analysis iteration as function of the total number of analyzers, assuming every analyzer works independently. Here, only 20% of the Tier 2 resources are available for analysis.

Clearly the cooperative scenario is more realistic. For our modeling of this scenario, we imagine that 10 ATLAS analysis groups will process D^1PDs into D^2PDs , resulting in 10 different D^2PDs in all. 5 separate sub-groups will then process each D^2PDs into D^3PDs , resulting in 50 different D^3PDs . Finally, 10 analyzers will use each D^3PD to make plots, resulting in 500 analyzers in all. In order to study the number of analyzers the system can support, we scan the number of groups, sub-groups per group, and analyzers per sub-group, keeping the 10:50:500 relative ratio. The results are shown in Figure 17. We now find that 800 cooperative analyzers can co-exist on the computing system, if they can wait 10 days for their first plots.

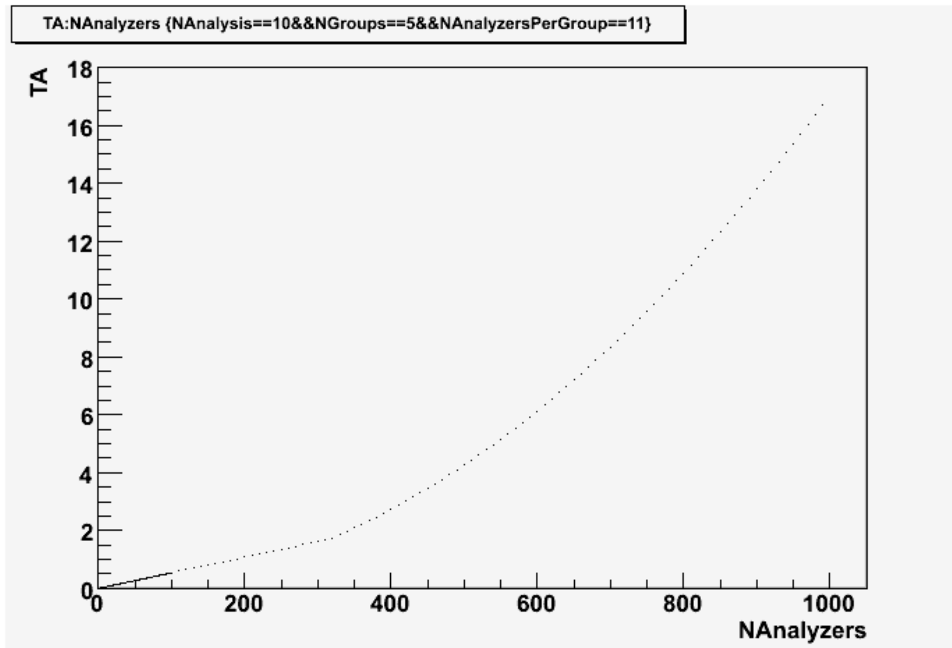


Figure 17: Time required (in days) for a single analysis iteration as function of total number of analyzers, assuming analyzer works cooperatively. Here, only 20% of the Tier 2 resources are available for analysis.

6.6 Conclusions

Our modeling leads us to several observations:

- Dedicating 50% of Tier 2s to Monte Carlo (MC) production will at best allow 10% (100%) of one year's worth of recorded data to be fully (fast) simulated within a year. We are likely to need to dedicate a larger fraction of Tier 2s to MC production in order to have the multiple iterations of MC production necessary for simulation tuning and the statistics required for extracting measurements.

- Assuming we dedicate 80% of Tier 2s to MC production (leaving 20% for analysis) and assuming that 1/3 of analysis resources are dedicated to transforms which read D^1PDs and produce D^2PDs or D^3PDs , it would take 10 days for 10 such transforms to simultaneously run through their input. Effectively, each physics or performance group can only run through its D^1PD once or twice a month.

- Placing all analysis and MC production activity at Tier 2s provides very little headroom for contingencies.

While the size of individual Tier 3s may be small, the number of Tier 3s sites will rather large. Therefore, it is not difficult to work out scenarios where roughly equivalent total resources are available at Tier 3s and Tier 2s. For example, the ATLAS 2010 Tier 2 CPU is equivalent to 100 ATLAS institutions with Tier 3s composed of 60 kSi2K each (roughly 40 cores or 5 eight-core boxes). The impact of so much additional computing capacity is game-changing. Clearly Tier 3s would be used for analysis tasks, therefore leaving more Tier 2 capacity for physics or performance groups to run through their D^1PDs . But they may also assume a significant fraction of MC production responsibilities, thereby leaving even more room for analysis on Tier 2s.

7 Tier 3 Task Force Recommendations

The two example Tevatron analyses present a picture of thousands of job requests, involving access to many thousands of files, done on a periodic basis—as much as monthly for some. Extrapolating these experiences into the ATLAS world, one is impressed with the amount of computing that might be asked of the Tier 2 centers as the active source of data and only significant production, analysis, and Monte Carlo job slots.

This is further attention-getting when one accounts for a major computing difference between CDF or DØ and ATLAS: Many Tevatron Standard Model measurements are statistically limited—either signal or background or both—and so the determination of systematic uncertainties is bounded by the event sample sizes. Statistics will not be a burden at LHC in almost all measurements, and so considerably more scrutiny of detector behavior, model parameter excursions, and background uncertainties will be required. Clearly, this has ramifications on computation. Data sets will be used repeatedly as sources of actual or fake backgrounds and multiple, specialized Monte Carlo samples will be required to explore parameter spaces of resolution and theoretical terms. The more data are collected, the more deeply this scrutiny will go.

This leads to the question: what would be the result of unpredicted periodic or even a permanent increase in the already extensive Tier 2 burden? Experience at the Tevatron suggests a number of ways in which this might occur, any one of which would have significant implications for U.S. analysis.

1. For example, could more full simulation Monte Carlo be required than currently anticipated? If so, Section 6 suggests that this will become a serious issue.

2. Could major errors occur within large Monte Carlo samples necessitating emergency regeneration ATLAS-wide? Both of these Monte Carlo surprises have happened more than once in the Tevatron experiments[14]. In the ATLAS model, redoing significant samples is almost a reprocessing-level production task, from source to re-production of the D^1PD to D^3PD formats, experiment-wide.

3. Could there be more turnover in D^1PD or D^2PD analysis than anticipated? Under the current scheme, a major regeneration of data from the AOD level necessitates a whole chain of production regeneration—all D^1PD and all D^2PD samples, and probably even D^3PD samples of which there might be hundreds or thousands in a mature experiment.

4. Reprocessing of the entire dataset is anticipated in ATLAS and this is prudent. The DØ experience was that extended reprocessing resources were sometimes underestimated and that the Monte Carlo production capability of the experiment was considerably reduced during reprocessing since MC resources were pressed into service for weeks at a time. Such an event within ATLAS would translate into the Tier 2 centers taking on some of the Tier 1 roles, at the cost of user analysis, D^2PD , D^3PD , and Monte Carlo production.

From the simulation studies presented in Section 6 we see that the required Tier 2 resources could be considerable and that the 50% fraction of Tier 2 resources for "analysis" may be at best, fragile. For realistic assumptions about the fraction of full-simulation and fast simulation, not only is analysis capability arguably at risk, that flexibility that we believe is important is potentially nonexistent if Tier 2s are the terminal significant production and analysis tier.

Previous experience at the Tevatron should motivate a computing model for the U.S. that is built around the ability to manipulate the various pieces into new roles, demanded by the circumstances. In contrast, the current vision of Tier 3 centers is of a set of independent and relatively low-capacity campus sites following the philosophy that the Tier 2s and user facilities

[14] Famous was an incident with the usage of the Monte Carlo generator ALPGEN in the W/Z plus jets mode—a random number seed was misused by many users at the Tevatron and emergency re-simulation was required for this important signal/background reaction.

at the Tier 1 and elsewhere will be the computing engines of first and last resort.

Observation 8 *Should ATLAS-wide production needs be more than the Tier 2 centers can provide, the only flexibility is to "eat" away at the 50% of the Tier 2 resources nominally reserved for U.S. user analysis. One has to ask what the likelihood is of such an outcome and whether U.S. ATLAS analysis could survive the effects of such a result.*

Recommendation 1: With past history as a guide and with prudent concern for the challenge and uncertainties of ATLAS analysis, the *structured* U.S. ATLAS computing infrastructure should be deeper than the Tier 2 centers. A flexible and nimble infrastructure would include strategically extending some data production, Monte Carlo simulation, and analysis into the U.S. ATLAS Tier 3 sector.

7.1 Potential U.S. ATLAS Tier 3 Strategies

7.1.1 A Flexible Tier 3 U.S. ATLAS System: Four Kinds of Centers

The tiered computing model is the most flexible structure currently conceivable to process, reprocess, distill, disseminate, and analyze ATLAS data. However, as our calculations in Section 6 suggest, the Tier 2 centers themselves may not be sufficient to reliably serve as the primary analysis engine for 400 U.S. physicists.

Are there uncertainties in these calculations?—There almost certainly are. But we conclude that the risks are too high to behave as if this issue is unlikely—especially in light of the history of these enormous experiments' and the way in which adapting to circumstances became a persistent fact of life. The third tier can be an important component to buffer the U.S. ATLAS analysis system from unforeseen, future problems. In fact, it can be developed to significantly leverage U.S. ATLAS physicists' contributions to their physics groups while providing what might be that missing, but crucial flexibility.

The current situation is not very healthy. Appendix H reports the results of a survey done of all U.S. ATLAS institutions regarding their available Tier 3 resources for ATLAS. These are summarized in Figure 18.

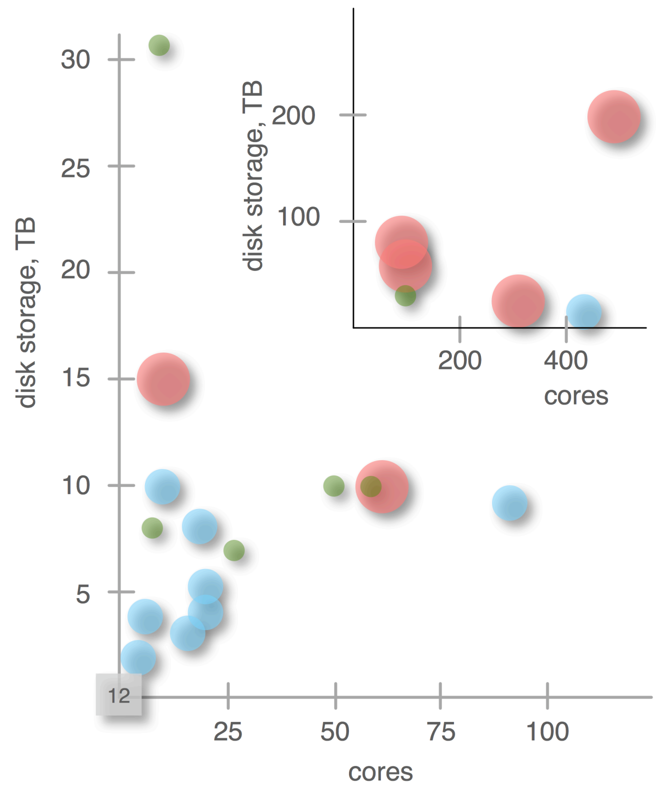


Figure 18: U.S. ATLAS Tier 3 resources during late fall, 2008 at universities. The size of the circles represent the rated connectivity to the outside world: small green, 100Mbps; medium blue, 1Gbps; and large red, 10Gbps. The inset shows some significant Tier 3 centers, most of which are associated with existing Tier 2 centers. No effort was made in this figure to account for varying speeds of the processors, see Appendix H for more details. Also, note that there are 12 institutions with no Tier 3 capability.

We envision the Tier 3 level as possibly presenting two faces to the Grid:

- The first presence is one in which it fully participates as both consumer and provider of computing services to the ATLAS Virtual Organization (VO), whether cached data or computer processing or both. Simultaneously, it would provide large-scale analysis or Monte Carlo capability for members of its local VO.
- The second presence is one of being just a consumer within a local VO, enjoying access on demand to data sets, but without the responsibility

and resource load of serving any ATLAS needs.

These two Grid relationships mark a crucial distinction as the latter— if possible— creates a significant scientific presence for a university group without a burdensome maintenance load. But, they do so within the important boundary condition: Tier 3 sites are by definition funded by "private" means: university and grant contributions. The local users control access, policy, and usage of their Tier 3 facilities.

We call Tier 3 centers with the first of these Grid relationships Grid-Responsible Tier 3 Centers and the second, Grid-Active Tier 3 Centers. While there are technical distinctions between them (see below), the basic difference is perhaps best thought of as the VO that they serve: Grid-Responsible Tier 3 Centers can, *if they choose*, serve the U.S. ATLAS community as a whole while Grid-Active Tier 3 Centers serve only the local community which owns them.

Recommendation 2: The strategy for building a flexible U.S. ATLAS Tier 3 system should be built around a mix of 4 possible Tier 3 architectures: T3gs, T3g, T3w, and T3af. Each is based on a separate architecture and each would correspond to a group's infrastructure capabilities. Each leverages specific analysis advantages and/or potential ATLAS-wide failover recovery. They are specifically defined in Section 7.1.2.

7.1.2 Tier 3 Architectures

The 4 Tier 3 architectures are the following:

1. Tier 3 with Grid Services, "T3gs" A Tier 3 center is a campus-based cluster with grid resources sufficient to support `pAthena` job queues and DQ2 clients. They are distinct from Tier 2s in that they may choose to allow members of the U.S. ATLAS VO job access, but definitely provide privileged access to the groups which own the resources. Any U.S. ATLAS group with the minimum Tier 3 resources (see below) can become a Tier 3. The reality is that a broad spectrum of "Tier 3 centers" already exists within U.S. ATLAS. For some groups, for example, those with Tier 2 centers on their campuses, space, power, and air handling supply enough capacity to support both the Tier 2 needs and university-owned clusters. Each of the eight Tier 2 university groups, plus SLAC and the University of Wisconsin (which benefits from the CMS Tier 2 center on its campus) have those capabilities now. We

have attempted to characterize a minimal T3gs cluster in Appendix E on page 107, including pricing.

2. Tier 3 with Grid data access, "T3g" A Tier 3 center of this sort could be a desktop cluster, or a small batch cluster, with storage sufficient to support large datasets. It would be a DQ2 client, but share DQ2 site services and catalog access with a particular, named Tier 2 center in order to support data subscriptions. It should be possible to submit `pAthena` jobs from within the cluster to the outside world, but also to itself and not expose itself to analysis jobs from the outside. We have attempted to characterize a minimal T3g cluster in Appendix E on page 107, including pricing.

3. Tier 3 workstations, "T3w" This center refers to a set of unclustered workstations individually running OSG, DQ2 client, and `ROOT` software. It would essentially be only capable of `ROOTtuple` analysis on modest sized datasets and submitting `pAthena` jobs for processing and storage elsewhere (which could be within the Tier 2 cloud, or, of course, the new T3gs cloud).

4. Tier 3 hosted at a national Analysis Facility, "T3af" This would involve a special arrangement with either a large T3gs or a National Laboratory Analysis Facility, such as the proposed Brookhaven Analysis Facility (BAF) [8]. The model might be one or both of two strategies: 1) universities could ship university-stickered hardware to the AF or 2) universities could spend against an existing purchasing account created for that purpose to the AF. The CDF arrangement at Fermilab is an example of the latter where groups would purchase approved equipment configurations to be housed in the CDF CAF in exchanged for fair-share computing privileges in proportion to their contribution.

   It is important to note that in CDF this arragement was a quota system and not a strict partition between collaboration-wide and University-owned resources. Here is a concrete example to illustrate the arrangement. Assume that CDF has 1000 batch slots for collaboration-wide access configured to give equal share to each CDF member. University X has money and a perceived need for computing resources to do analysis beyond that provided by the CAF. However, they either do not have the infrastructure, expertise, security/policy control and/or desire to deploy a computing cluster to satisfy their perceieved need. They buy 100 CPUs (batch slots) worth of hardware in compliance with the hardware requirements for CAF system administration and send it FNAL to be incorporated into the CAF. The Condor-based batch system in the CAF is configured such that the total number of batch slots available to the entire collaboration is now 1000+100=1100 but University X gets *immediate* access to up to 100 batch slots *in addition* to their equal share of the 1000 collaboration-wide slots. Note that this is a win-win for both the collaboration and University X. University X effectively gets a 100 CPU cluster that they pay for without having to worry about system adminstration (nor power/cooling in the CDF model). The collaboration as a whole gets use of their hardware when they are not using it up to their quota. Despite best intentions, no group uses 100% of their hardware resources over long periods of time for physics.

Among these:

- As noted, a handful of T3gs sites already exist as significant centers associated already with U.S. Tier 2 locations.
- T3w represents what some have assumed to be a typical Tier 3 center.
- T3af is intentionally similar to the CDF Central Analysis Facility now.
- T3g is new and is perhaps closest in function to the DØ CLuED0 desktop cluster.

Each of these sites is distinct from one another and each serves a distinct purpose. Each is scalable from within, and any T3w or T3g could be upgraded or evolved into the next, more capable site. Groups could formulate a multi-year plan with their universities and their funding agencies to pursue a specific development path, starting with T3w and becoming T3g, for example.

A major concern for all groups would be the level of support required of them. In Section 7.3 below, we make recommendations about that important issue. But, before that, we review examples of the broadening of the Use Cases outlined in Section 4 which become possible with an array of Tier 3 centers as described above.

Observation 9 *It may be possible for university groups to confederate with one another, from one campus to another, or even across department and disciplinary boundaries within a single campus. For some Tier 3 tasks, such arrangements may work well. We know of no functioning arrangements at the* *time of this writing, but we believe that efforts are underway to create them on a few campuses..*

7.2 Revisiting the Use Cases

Because of all of the possible surprises outlined above, a U.S. production system which terminates with the Tier 2 cloud is neither flexible, nor nimble. With the four kinds of Tier 3 centers described above, this deficiency can be addressed, and if we plan this over years, we can react to the unknown conditions that the LHC will present to us.

To that end, we can roughly delineate the boundaries around the two larger Tier 3 centers and indicate their capabilities by expanding on the use case discussion from Section 4.

7.2.1 Distributed Data Management and Compute Elements

As ATLAS accumulates data, the benefits on having local analysis capability increases (more control, no reliance on external networking, storage, and processing resources, no competition), but the computing burden also increases (more CPU, more storage, and the need to bring the data to the local site). Development of a local site can evolve, starting with modest CPU power and modest storage, increasing both as funds and needs dictate. However, sufficient access to the large datasets is the make-or-break requirement which will permit the development of Tier 3 clusters capable of significant, local ATLAS computing. Data access includes two minimal requirements.

1. Connectivity from the campus to the source of the data must be reliable and of sufficient bandwidth in order to support the migration of files in the TB range. Currently, it appears that “Physics Building” to Tier 2 cloud or T1 experiences vary widely: some anecdotally report few 10’s MBps sustained transfer rates, others report only a few MBps transfers. Evening this out is both a national ATLAS issue and also a local university concern: apart from regional, state, and national networks, connectivity can be compromised within campuses and at campus boundaries. In order for substantial on-campus analysis, 10’s of MBps transfers are likely to be required by the time of the 10fb^{-1}period covered in this report.

2. The Distributed Data Management (DDM) system within ATLAS is complicated and technical. Access to the data essentially requires sophisticated tools on both ends: from the data request to the satisfaction of a request. Following Mambelli [14], access to ATLAS data can follow a successively more sophisticated set of configurations as suggested in Table 19. Each step involves more difficult installation and maintenance.

Table 19: The hierarchical list of possible storage configurations (c*n*ddm) and job execution compute elements (c*n*je) within ATLAS [14].

configuration	comments
c0ddm:	no locally managed storage, relying on external SE
c1ddm:	SE only (ATLAS visible files are elsewhere
c2ddm:	DQ2 endpoint + SE (site services & LFC outsourced)
c3ddm:	DQ2 site services + endpoint + SE (LFC outsourced)
c4ddm:	LFC + DQ2 site services + endpoint + software
c0je:	No grid computing elements
clje:	Grid computing elements
c2je:	Grid CE + Panda support

c4ddm plus c2je is a conventional Tier 2 setup. c2ddm is currently the existing DDM arrangement at the University of Chicago Tier 3. Notice that the considerable benefit of the c2ddm configuration is the ability to make use of subscription services to data and the consequent recovery and retry failover mechanisms built into DQ2 site services transfer agents.

Finally, a site's computing element (CE) configuration can range from a single workstation or laptop capable of only running `ROOT` to a site which supports worker nodes responsive to Panda pilots within a full Panda configuration. The simple hierarchical range of CE are also shown in Table 19. A c0je would only be capable of running `ROOT` and local Linux software; a c1je site would have benefit of grid-installed, ATLAS software updates and be capable of submitting pAthena jobs to the grid; and c2je sites would be able of supporting pAthena computing on their site.

7.2.2 Value-Added From a T3gs System

While not attempting to be prescriptive, we believe that we can illustrate the flexibility that becomes available with T3gs system. For the purposes of illustration, we assume that such a system is rack-based, with 40 nodes of 8-processor-class computing and 10's of TB of storage elements. Further, we presume connectivity to the outside ATLAS world through at least a 1Gbps fiber network, if not a shared 10Gbps network. We illustrated two sorts of value-added capabilities: data production and Monte Carlo production.

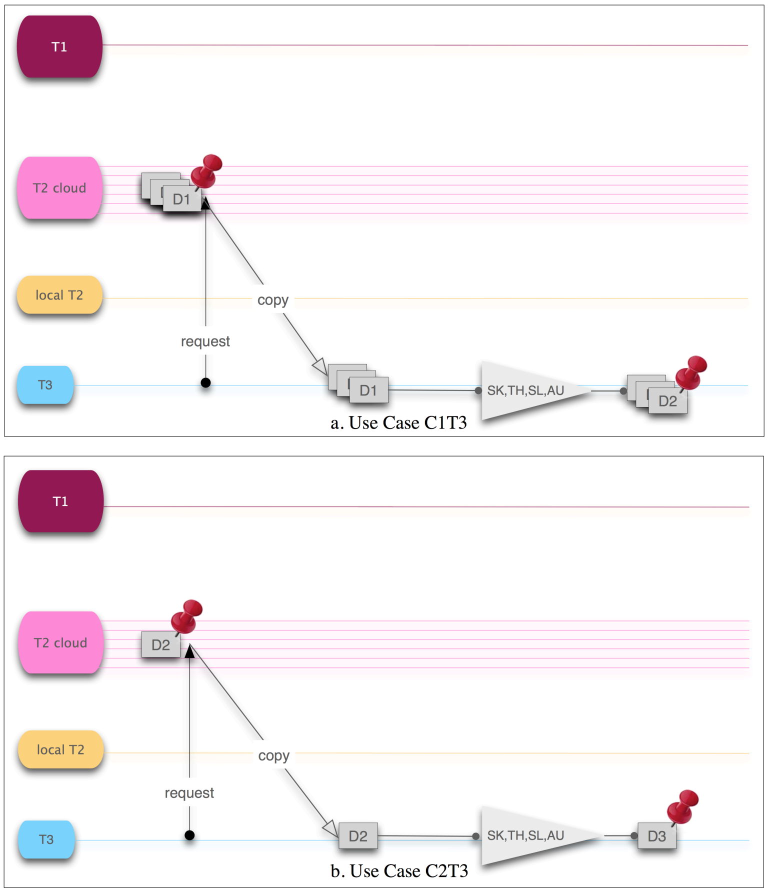
Such a site would be a combination of at least: c3ddm or c4ddm SE and c2je CE from Table 19.

Data Format Production An example production use case we consider is the ability to produce D^2PD from D^1PD datasets for a full stream in a reasonable time according to the parameters of Tables 6 and 16. This use case involves copying and temporarily caching a full stream of 1.6×10^8 events of D^1PD , or 4 TB in total. At a sustained data transfer rate of 50 MBps, this would require approximately 24 hours. For good 2008 transfer rates of 10 MBps, this would require roughly 5 days for one full stream. Notice, that this is a future capability, already reached on ATLAS systems in a non-production environment. In 2008 terms, average transfer rates are roughly 5 times slower, as shown in the "low" column of Table 16. ANL has observed sustained transfer rates from the MWTier 2 of >20MBps, but a factor of 10 or so slower in transfer from BNL. Figure 19 illustrates the Use Cases for such a production task, as well as a similar use case for processing D^3PD from D^2PD .

Once cached, using 0.5 kSI2k-s to process to D^2PD , would require approximately 900 node-days producing an output dataset of 5 TB, and a consequent up-transfer time of another day at 50 MBps. For one full rack of nodes, the processing time would be approximately 3 kSI2k-d, or about 2 clock-days for a 2008 modern CPU. For a group needing enhanced production capability or a redo of production in an emergency situation, this is a reasonable wait time. The total storage would be less than 10 TB total, and while network requirements are siginificant, even if the efficiency of transfer is much less than 100%, the quick calculation illustrated here suggests a serial processing-transfer, when in fact, these would be done in parallel so that the slowest rate would be the actual clock-span for the whole project. In this case, transfer could even be only 50% efficient before it would dominate the overall project.

Monte Carlo Production As a contrast, we also can characterize a powerful Monte Carlo use case, here, with the idea that local physicists at a

Figure 19: Use Cases C1 and C2 are here augmented to include T3gs contributions: C1T3 and C2T3.



university with a T3gs would be utilizing their private resources in support of the physics group of interest to their local VO. Large-scale, full simulation is so significant a task, it is likely best left to the Tier 2 clouds to perform for intensive processes such as $t\bar{t}$. However, it is almost certain that "signal-sized" Monte Carlo production—full or a fast simulation—will be necessary, given the paucity of ATLAS-wide Monte Carlo and the burdens facing Tier 2 simulation. The only way for a group to explore systematic effects, theoretical parameter ranges—or even to fix a mistake, is the existence of a nimble Monte Carlo facility tuned and directed to the physics group's needs. Control of such a facility would allow any U.S. university group to contribute in a crucial way to their international physics groups.

In order to illustrate, we choose a “signal-sized” sample appropriate to our mid-range $t\bar{t}$ physics set in a 10fb^{-1}setting. The CSC Note [5] describes a lepton-plus-jets sample size for muons and electrons of about 6000 events. We'll presume a factor of 3 to account for background generation and a looser acceptance for purposes of illustration. The CSC note was for 100pb^{-1}, and so we scale up for our scenario of 10fb^{-1}and these two factors suggest a Monte Carlo generation exercise of 1.8×10^6 events.

A group might be interested in either/both full simulation for this set, or a fast simulation. For our modeling (Section 6) we presumed ATLFAST-II, and do so here. Under these conditions, this dataset could be fully-simulated in a full rack of processors in about 130 kSI2k-days and fast-simulated in less than a single kSI2k-week. For 2008 processors, this would be about 3-clock months for full, and less than 5 clock-days for the fast simulation. If pileup is included for the instantaneous luminosity presumed, then, this full simulation exercise would require 3.5 times these amounts.

This probably sets a limit for what a single T3gs could do for full simulation, but multiple fast simulations for “signal-sized” samples would be an important resource for most physics groups and an important contribution for any so-capable U.S. university group.

The data transfer for the produced samples is not so different from the D^{1}PD and D^{2}PD samples in the Production example. If full RAW, ESD, AOD, D^{1}PD data formats are produced, then they could be transfered back to the cloud in less than a day using the presumption of 50 MBps sustained transfer.

Summary T3gs system consisting of approximately a half to full rack of 8 processor nodes, 10's of TB of storage, and a reliable network capability could be a welcome production fail-over capability for DPD processing, and a crucial and unique contribution to any physics group effort requiring significant simulation. This would be welcome within all physics groups.

7.2.3 Value-Added From a T3g System

The T3gs idea involves a significant commitment by a university site as the hardware involved at that level would require special infrastructure. The T3g idea is meant to be a system capable of supporting significant computing contributions, yet still fit within an office environment and with minimum maintenance. The boundary conditions for such a cluster would include:

1. Local access to datasets of sufficient size to support full analyses of average complexity at the AOD, D^{2}PD , and D^{3}PD level.
2. Sufficient CPU power to locally produce small Monte Carlo datasets.
3. Local access to ESD datasets of sufficient statistical precision in order to create/debug/tune analyses for eventual grid submission for detailed detector studies.
4. Involve only a "consumer" relationship to the ATLAS grid: data cached on a T3g site should be invisible and inaccessible from the grid and that CPUs supporting local T3g analyses should be unavailable for grid use.
5. Long-scale, repetitive operations should not require repeated human intervention. This is especially true of large file transfers and losing jobs at unknown locations within the grid. Anecdotally, submission to the grid leads to approximately 10% failure rates.
6. Processing should be 100% reliable, which argues strongly for local control.
7. Support required of local users should be minimal.
8. Database hosting (such as the LHC File Catalog, LFC and transfer database) should be minimal or nonexistent.
9. Special airhandling and power should not be required.

Such a site would be a combination of c2ddm SE and c1je or c2je CE configurations.

Focused Signal-Background Analyses One of the crucial aspects of analysis is quick turnaround and full awareness of the state of any submitted job. "Quick" is in the eye of the beholder, of course, but the rule of thumb of about a single day's processing should still hold for large, but local jobs.

Colleagues at Argonne National Laboratory have begun to construct a Tier 3 (PC Farm, "PCF") which currently contains 3, 8 core tower PCs with 8GB RAM and 2 TB of internal drives in a batch cluster of condor slaves. Their benchmark analysis is an inclusive γ production sample with $p_T(\gamma) >$80 GeV and their experience is that 4.5pb^{-1} results in workable `ROOTtuples` of 1.5 GB. With assumptions that signal and background samples are equal, that Monte Carlo is generated at twice the signal size, and that the analysis task is to produce augmented D^3PD s from AODs requires 20 TB of storage, about 4 TB of which is signal. Similarly, inclusive jet analyses with E_T >400-500 GeV requires 40 TB of storage. These analyses serve as a high end examples as D^3PD analysis would be less demanding.

Their benchmarking suggests that full processing through the signal sample with 10 towers of Dual Core AMD Opteron 280, 2.4 GHz processors would take approximately 48 hours. The question of how to get the data to the site for such analyses is an important one. Experience suggests that issuing a `dq2_get` command for datasets approaching a TB would require too much human nursing of resubmitting requests and bookkeeping. The other alternative is for the site to run the full DQ2 site services and catalog, which is a significant effort and commitment. An ideal situation for a modest installation would be to rely on a Tier 2 site to host the catalog and DQ2 site services on their behalf. Then such a site could issue subscription requests and the data would arrive with automatic re-starting and bookkeeping. This intermediate solution has been colloquially dubbed "DQ2-lite" and is functioning at the University of Chicago.

Transfer of the full 4 TB signal dataset would require about 24 hours at a sustained 50 MBps rate, which is adequate.

Conclusions The definitions of these two kinds of Tier 3 clusters: T3gs and T3g are meant to be different in size and infrastructure; the capabilities they would provide to their local users (and to ATLAS as a whole); the services that they would host; and the subsequent support requirements demanded of each.

We have attempted to benchmark roughly minimal starting points for each kind of cluster and Appendix E on page 107 lists examples and current pricing for each. These would be significant enhancements the university capabilities, but for relatively modest costs. Table 20 summarizes parameters that might roughly distinguish them according to the benchmarks described in Appendix E. Note that "modest cost" is a relative term for the T3gs system as there are significant infrastructure costs for a rack of computing which would produce 10's of KW of heat. Depending on the existing networking infrastructure, in order to be most productive even a T3g system might require university contributions—or even state contributions—to guarantee necessary bandwidth.

Figure 20 shows how the benchmark characterizations of the T3gs and T3g capabilities map onto the storage-core space Figure 18. The Orange region roughly shows the T3g space, while the green, the T3gs space. The

Table 20: Approximate characterization of the T3gs and T3g-sized clusters

service/resource	T3gs	T3g
cores	~ 168	~ 80
storage	~ 24 TB	~ 20 TB
cost	~ $80k	~ $30k

white region includes the current U.S. sites with 24 fitting below the T3g capability band. The sites shown on the figure are just copied from Figure 18 onto the new scale. Obviously, the U.S. now has 8 sites which are already in the T3gs or T3g state.

7.2.4 Technical Recommendations

In order to support the services described for the T3gs and T3g systems—in particular, c2ddm and c1je—the following technical and organizational decisions should be considered: The "outsourcing" of DQ2 Site Services, databases, and large catalogs requires some changes to DQ2 and the permission of privileged relationships with some particular Tier 2 centers.

Recommendation 3: In order to support a Tier 3 subscription service, without a significant support load or the need to expose itself to the ATLAS data catalog, a particular DQ2 relationship must be established with a named Tier 2 center, or some site which can support the DQ2 site services on its behalf. This breaks the "ubiquity" of Tier 2s — here, a particular Tier 3 would have a particular relationship with a named Tier 2. This dual-capability (limited exposure of a site's file catalog and a subscription-like functionality) has been colloquially referred to as "outsourcing" DQ2 site services.

It is desirable to run pAthena jobs wholly within a T3gs or T3g site, without allowing outside jobs to be run on that site.

In order to give the user the same interface at the Tier 3s as compared to Tier 1/2s for job submission, control, and monitoring, it is desirable to configure the Tier 3s for `pAthena` services. The nature of Tier 3s is that they have local control over their how their resources are allocated (which users can run jobs on their site, how much share is given to different activities like analysis and production, etc.). The functionality to handle aspects of this control has been added by the Panda developers very recently.

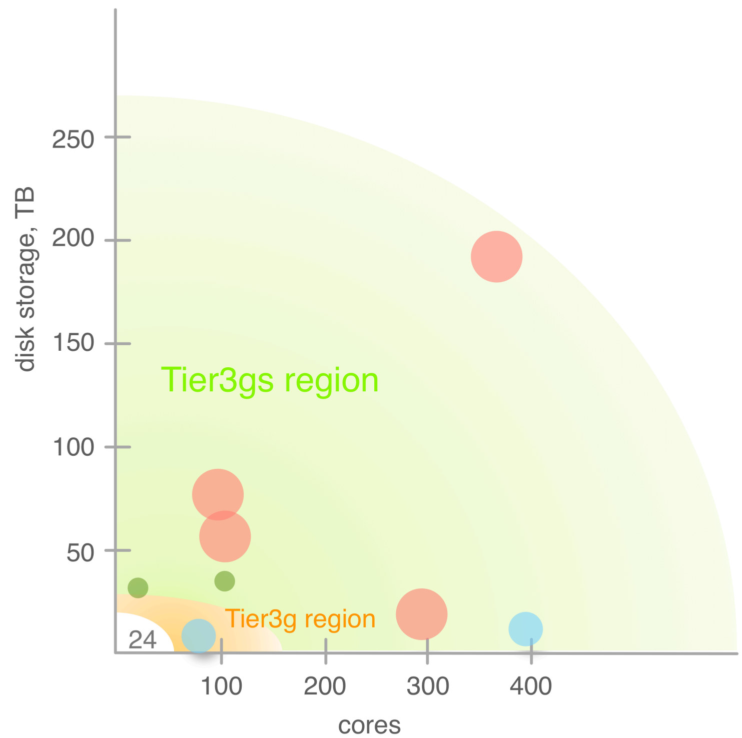


Figure 20: The 8 sites with greater than 8 cores and 30TB of disk space are mapped onto a storage-core space which is the scale of the inset in Figure 18. The Orange region corresponds roughly to the capability of the benchmark (and above) T3g systems, while the Green region corresponds roughly to the capability of the benchmark (and above) T3gs systems.

Recommendation 6: We recommend that the recent addition of `pAthena` local control-functionality be maintained, and possibly extended to allow for more convenient control and access/monitoring of the Tier 3 site configuration by local administrators.

In this way, policy can be easily implemented and monitored without technical hurdles and the resource configuration across the U.S. ATLAS is generalized.

Access to the data is the go-no-go necessity for both T3gs and T3g. Currently, bandwidth is uneven between university sites and the Tier 2s or Tier 1, ranging from a few MBps to tens of MBps. The above simple analyses suggest that working files will be in the few TB range, as much as 4TB for the simple T3g example. Roughly, 2TB would take about 24 hours to transfer at a sustained 20MBps rate (an ideal figure, ignoring restarts, header information, etc). This we take as a benchmark goal for each university site for the 2010-2011 timeframe of this report. Note, we are not making a recommendation about all universities and all possible Tier 2 sites. We have in mind a targeted goal for each campus: a point-to-point, tuning between each T3g or T3gs and the particular Tier 2 center from which episodic, large data-file transfer would occasionally be required.

Recommendation 7: Sustained bandwidth of approximately 20MBps is probably required for moving TB sized files between Tier 2 and Tier 3 locations and it should be the goal that every campus or lab group establish such capability within a few years. This requires a high level of cooperation and planning among U.S. ATLAS computing, national network administrators, and campus administrators. Note: it might be useful and prudent to tune bandwidth between *particular* Tier 3 locations and *particular* Tier 2 centers rather than to set a national standard which might be difficult to meet. Note that the Resource Allocation Committee will have authority over the large-scale movement of data and any large scale caching of Tier 3 generated files into the Tier 1 or Tier 2 clouds.

7.3 Tier 3 Support Strategies

An essential component of the recommended strategy here is the creation of a centralized support structure. The considerable obstacle to creating and sustaining a campus-based computing cluster is the continuing support required. While the definition of T3w clusters is meant to reduce this burden, it does not eliminate it. Even for Tier 3 centers, full-time system support is often a deal-breaker for any single university group.

Rather than presume or encourage individual system administration lines in continuing grants, we recommend the establishment of a centrally-located, U.S. ATLAS funded support system consisting of personnel who will travel to sites to assist in bringing them to functionality and be available for consultation if and when problems develop. We do not mean a help-desk. Rather, we presume a named crew of system support professionals who will establish personal relationships with their university clients and perhaps even campus network administrators. We believe that this investment is well-worth the funds required and will help to establish a coherent administrative structure across the U.S. ATLAS community and serve to develop a savvy set of physicist-system administrators as well. Without such support, the only thing that will be consistently usable for most U.S. institutions will be the T3af model, and in turn, essentially no campus presence.

The DØ collaboration, and subsequently CDF maintained a world-wide user-network of site administrators (physicists and computer professionals) and one or two Fermilab Computing Division (CD) experts to first, install, and second, maintain the highly complex Sequential Access to Metadata (SAM) DDM system. Many installations of SAM were unique to individual sites because of administrative and technical firewalls and often a CD expert would travel to the site, assist in the installation, and then continue the personal connection into the maintenance phase. Weekly or biweekly phone meetings kept this group together for years. It is precisely this sort of arrangement that we envision here.

Recommendation 4: U.S. ATLAS should establish a U.S. ATLAS Tier 3 Professional, a system administration staff position tasked to 1) assist in person the creation of any Tier 3 system; 2) act as a named on-call resource for local administrators; and 3) to lead and moderate an active, mutually supportive user group.

Recommendation 5: In order to qualify for the above U.S. ATLAS Tier 3 support, U.S. ATLAS Tier 3 institutions must agree to 1) supply a named individual responsible on campus for their system and 2) adhere to a minimal set of software and hardware requirements as determined by the U.S. ATLAS Tier 3 Professional.

7.3.1 Summary of the T3gs and T3g Idea

Appendix E presents lists of all server, storage, network, and software necessary to create examples of each of the Tier 3 center types described above. Table 21 compares the "T3 quartet" for a number of criteria. These are approximate and refer to entry level configurations.

7.4 Participatory U.S. ATLAS Cluster Program

The LHC is a very well-known scientific program and most campuses are aware of their participation and proud of it. Many institutions are welcome

Table 21: Some of the approximate "Tier 3 quartet" functionalities and requirements are listed for relative comparison purposes.

	T3w	T3g	T3gs	T3af
stands for:	workstation	grid access	grid services	analysis facility
approx. number of cores	$\sim 8-32$	> 80	> 168	limited by agreement
format	towers	towers [ANL model, see E.2.1] or rack [Duke model, see E.2.2]	rack	rack
storage capacity	$\sim$ few TB	$>$ 20TB	$>$ 30TB	limited by agreement
clustered? batch?	no or minimal	yes, headless workers	yes, headless workers	yes headless workers
interactive `ROOTtuple` analysis?	yes	no	no	no
MC, e.g. $t\bar{t}$	few hundreds ATLFAST in hours; millions generator in hours	few thousands ATLFAST in days; millions generator in hours	few millions ATLFAST in days; many millions generator in hours	few millions ATLFAST in days; many millions generator in hours
data production capability?	no	no	yes	yes
support level	owner/ group	group	group/dept professional	lab professional
network rating	100Mbps	$\geq$ 1 Gbps	10Gbps	10Gbps
software, services	ROOT Athena	ROOT, OSG Athena Local Resource Manager (e.g. Condor, PBS, ArCond[see E.2.1]) DQ2 endpoint "outsourced" catalog, subscription	ROOT, OSG Local Resource Manager (e.g. Condor, PBS) robust network file system (e.g. dCache, xRootd) DQ2 site services	same as T3gs
specialized cooling/ power	none	none (towers) CRAC (rack)	CRAC 10's kW	facility
costs	$\geq$ 20k	$\geq$ 30k	$\geq$ 80k	negotiated

to proposals for one-time support of significant research programs, while reluctant to support programs which might imply a future long-term commitment, such as support personnel.

We recommend the initiation of a program of recognition among U.S. ATLAS, both NSF and DOE, and universities which choose to participate in one-time, or periodic capitalization of campus clusters or centers. Such a contribution to ATLAS should be treated as a **substantial collaboration** and a program of recognition should be established to certify any institution's investment in the ATLAS scientific mission. Institutional membership in such a program would presumably take the form of a match against Agency support and would form a quantitative value-added to the establishment of campus-based computing, as opposed to simply an Agency allocation to one of the national laboratories. It should also be acknowledged in ways which enhance the campus' access to ATLAS outreach materials, ATLAS TV participation, visits from ATLAS scientists, and offers of hosting of university administrators at CERN and/or other U.S. ATLAS sites of interest and/or programs of interest. In short:

Recommendation 8: Enhancement of U.S. ATLAS institutions' Tier 3 capabilities is essential and should be built around the short and long-term analysis strategies of each U.S. group. This enhancement should be proposal-based and target specific goals. In order to leverage local support, we recommend that U.S. ATLAS leadership create a named partnership or collaborative program for universities which undertake to match contributions with NSF and DOE toward identifiable U.S. ATLAS computing on their campuses. Public recognition of this collaboration should express U.S. ATLAS's gratitude for their administration's support and offer occasional educational and informational opportunities for university administrative partners such as annual meetings, mailings, video conferences, hosted CERN visits, and so on.

8 Conclusions

There are both quantitative and qualitative reasons to support a robust, university-based ATLAS computing structure. Here, we make the hopefully obvious observations about how U.S. ATLAS will succeed: through a well-supported, robust, academic HEP program.

8.1 An Exciting Particle Physics Mission is Guaranteed

The U.S. HEP community faces an enormous challenge in the coming years. At this writing, two long-standing laboratories have changed their missions from HEP to materials science. The flagship U.S. HEP laboratory is nearing the end of its 25 year old collider program with an uncertain future—not for lack of important science, but because of budget constraints. The vast majority of U.S. university elementary particle physicists will be working at off-shore facilities for a number of years, perhaps decades.

Ironically, in this period of reduced support, the physics opportunities have never been more significant! Either the Standard Model will play out to its advertised conclusion and obligate us to the unraveling of its extension, the existence of which is necessary for internal consistency. This will lead to new physics. Or, after decades of resisting abuse, the Standard Model will finally break at the LHC—obviously, leading us to new physics. This is the ultimate No-Lose Theorem: we are on the verge of a revolution in High Energy Physics.

Everyone reading this document in 2009 has spent essentially his or her entire career within this model which under any scenario now faces a extension or a complete overhaul. This is not the time for a weakened academic High Energy Physics program! The ATLAS, CMS, and LHCb communities must make every second at the LHC count.

Observation 10 *The technical (and social) challenges are enormous and in order for the LHC Mission to succeed—and it must succeed—the U.S. community has to be fully equipped and fully staffed in order to meet those challenges.*

8.1.1 The University Community is Key

The 50 year history of U.S. HEP has been driven by the vibrancy and technical expertise of its university community. The LHC era can either enhance that presence, or it can weaken it. One sure way to weaken it is for U.S. physics departments to conclude that, because of the abandonment of U.S. based beams, HEP is no longer worth the considerable investment that all major universities have made in faculty appointments and facilities. The way that the LHC era can enhance HEP at U.S. universities is by making a virtue out of a necessity. The CERN laboratory, while enormous, cannot support the kind of on-site presence that many U.S. groups have been accustomed to for decades. So, most of U.S. LHC high energy physicists will be on their campuses—for some departments, maybe the for the first time.

This increased campus presence *could be a good thing for the field*. A better thing for the field will be the growth of tangible, on-campus facilities as a part of the U.S. ATLAS program, writ large. This argues for a strategy which seeks to enhance local computing, especially if such a strategy can leverage local matching contributions, thereby enhancing U.S. ATLAS capabilities as a whole.

Universities overwhelmingly house the imagination engines which will drive ATLAS physics analysis. The sheer distance and prohibitive costs demand that the U.S. ATLAS analysis effort will be spread among the 40 or so institutions. In order for the scientific mission to succeed, a strong university analysis effort will have to be structured and maintained for the duration. This has its benefits as well as its challenges. The challenges are obvious: cooperative code development across distances is always difficult. It places a burden on documentation and what will seem to be a slower pace than in the past where hallway conversation frequently served as the means of disseminating patch releases and providing help. Video conferencing and other collaborative tools will undoubtedly develop out of necessity.

But, the benefits are surprisingly substantial. Traditionally, most universities posted students and postdocs at the host lab. Faculty traveled frequently, often weekly. HEP presence within academic departments was often a source of concern and bewilderment to colleagues, complicating hiring, promotion, and resource allocation. The LHC will probably result in more HEP personnel posted on campuses and if we “play this right,” HEP as an academic discipline could benefit.

The unprecedented publicity—overwhelmingly positive, even in the face of the September incident—has caught the attention of the public and university communities many of whom were pleased to discover that they had physicists engaged in this exciting enterprise. The opportunity for campus-based awareness of our science in the short-term and the long term, is unprecedented. Campus-based facilities serving the overall ATLAS analysis effort in quantitatively tangible ways could become a source of pride and a spirit of collaboration among U.S. high energy physicists, their departments, and their administrations.

The formula is simple: a strong campus-based, university HEP presence serves the LHC scientific mission. Therefore, nurturing the health of the HEP academic system should be a sensible component of any resource allocation strategy.

As a mission-preserving strategy, this sentiment should argue for strong, participatory, and tangible Tier 3 presence throughout the LHC experience. When coupled with the quantitative and strategic arguments above, the conclusion should be clear: an enhanced campus computing presence, developed over time—evolving as ATLAS proceeds down its still-developing path—will be an important component to U.S. ATLAS's scientific success.

9 References

References

[1] T. Aaltonen et al. Search for a Higgs Boson Decaying to Two *W* Bosons at CDF, 2008.

[2] Amber Boehnlein. DØ computing model, 2006. Presentation at HCP 2006.

[3] A. Chekanov and R. Yoshida. A PC farm for ATLAS TIER3 analysis, ATL-COM-GEN-2009-016, 2009.

[4] S. Chekanov and R. Yoshida. THE ARCOND PROJECT, ANL ASC (2008).

[5] ATLAS Collaboration. Expected performance of the atlas experiment, detector, trigger and physics, 2008. CERN-OPEN-2008-020.

[6] D. Constanzo, I. Hinchliffe, and S. Menke. Analysis model report, 2008. draft 1.4.

[7] Davide Costanzo. Event data model, 2008. ATLAS Week 04.11.2008.

[8] Michael Ernst. Plans for us facilities support for physics analysis, 2008. U.S. ATLAS Institutional Board Meeting, Simon Fraser University.

[9] Michael Ernst. U.s. atlas computing facilities, 2008. DOE/NSF U.S. LHC Software and Computing Review, Irvine, CA, 4-7, February 2008.

[10] D. Adams et al.. The atlas computing model, 2004. ATL-SOFT-2004-007, CERN-LHC-2004-037/G-085, v1.2.

[11] ATLAS Computing Group. Atlas computing technical design report, 2005. ATLAS TDR-017, CERN LHCC-2005-022.

[12] Data Streaming Study Group. Data streaming in atlas, 2007. https://twiki.cern.ch/twiki/pub/Atlas/DataStreamingReport/DataStreamingReport-loc.pdf.

[13] Roger Jones. Resource issues, 2007. ATLAS Software Week, November 4, 2008.

[14] Marco Mambelli. Tier 3 configurations which are technically possible: description and implications, 2009. private communication.

[15] Shawn McKee. Accounting p2: Scaling factors, 2007. http://www.usatlas.bnl.gov/twiki/bin/view/Admins/AccountingP2.

[16] Condor Project. THE CONDOR PROJECT.

[17] Jim Shank. Atlas computing resources, 2008. ATLAS Week 02.14.2008.

[18] Akira Shibata. Root analysis and implications to analysis model in atlas, 2008. ATLAS Software Week, November 2, 2008.

[19] K. Assamagan *et al.*. The atlas monte carlo project, 2009. draft.

Appendices

A Charge to the Task Force

The charge was electronically received on July 31, 2008 and was the following:

```
   US ATLAS and ATLAS have been formulating ideas and policy on
Tier 3 computing for a number of years now.  There was a white paper
from the US in August 2006 [ref., attached copy] and a task force
for ATLAS [ref.  ATLAS Twiki:
https://twiki.cern.ch/twiki/bin/view/Atlas/Tier3TaskForce ] that
ended early 2008.  Since then, the US ATLAS computing model and
perhaps more importantly, the Analysis Model have become clearer,
though both are still evolving.  We would like to revisit (revise,
update or rewrite) the US white paper taking into account these
recent developments.

   1.  Use Cases.  Typical workflows for physicists analyzing ATLAS
data from their home institutions should be enumerated.  This needs
to be inclusive, but not in excruciating detailed.  It should be
defined from within the ATLAS computing/analysis models, the existing
sets of Tier 2 centers, and their expected evolutions.  If there
are particular requirements in early running, related to detector
commissioning and/or special low-luminosity considerations, this
should be noted.  If particular ATLAS institutions have subsystem
responsibilities not covered by the existing Tier 1/2 deployment,
this should be noted.  Is the previous whitepaper relevant?

   2.  Characterizations of generic Tier 3 configurations.  Some
Tier 3's may be very significant because of special infrastructure
availabilities and some Tier 3's maybe relatively modest.  Is there
only 1 kind of Tier 3 center, or are their possible functional distinctions
which might characterize roles for some Tier 3's that might not
be necessary for others?  Description of "classes" of Tier 3 centers,
if relevant, should be made.  Support needs and suggestions for
possible support models should be considered.

   3.  Funding.  This is not part of the US ATLAS Operations budget,
```

```
so funding must come out of the institutes through core funding
or local sources.  We would like to make it easier for institutes
to secure funding for ATLAS computing--this can only happen if it
fits in the DOE and NSF budgets ( precedent:  the amount of funding
groups got for computing equipment in Tevatron experiments) and
it must fit in the overall US ATLAS model.  For the latter, we have
to make the case that the existing Tier 1/2 centers are not enough.
Perhaps a recommendation can be justified for an estimated $ amount
needed for a viable Tier 3 cluster -- something like X+n*Y's
where n = number of active physicists.

   The report should be completed in the form of a written document
which can both function for internal US ATLAS reference and as a
whitepaper for Agency consideration.  To that end, it might refer
to appendices for technical details and include an executive summary.
This is a US ATLAS study and if it differs in significant ways from
previous US ATLAS recommendations and/or worldwide ATLAS circumstances,
this should be noted.

   Please try to complete your work by October 1, 2008.
```

B Original Whitepaper

The Task Force was asked to react to the 2006 Whitepaper[15]. Note: no authorship is identified for this document.

US ATLAS Tier-3 Whitepaper
Version 8
Aug. 8, 2006

The US ATLAS project has been asked to define the scope and role of Tier-3 resources (facilities or "centers") within the existing ATLAS computing model and US ATLAS computing activities and facilities. This document attempts to address these questions by describing Tier-3 resources generally, and their relationship to the US ATLAS Software and Computing Project.

Originally the tiered computing model came out of MONARC (see http://monarc.web.cern.ch/MONARC/) work and was predicated upon the

[15] In order to embed the Whitepaper into this document, it was transcribed from its pdf image.

network being a scarce resource. In this model the tiered hierarchy ranged from the Tier-0 (CERN) down to the desktop or workstation (Tier 3). The focus on defining the roles of each tiered component has evolved with the initial emphasis on the Tier-0 (CERN) and Tier-1 (National centers) definition and roles. The various LHC projects, including ATLAS, then evolved the tiered hierarchy to include Tier-2's (Regional centers) as part of their projects (Hoffman committee final report, CERN/LHCC/2001-004).

Tier-3's, on the other hand, have (implicitly and sometime explicitly) been defined as whatever an institution could construct to support their Physics goals using institutional and otherwise leveraged resources and therefore have not been considered to be part of the official U.S. ATLAS Research Program computing resources nor under their control. We believe that this continues to be the case for Tier-3s, namely that *Tier-3s are not officially part of the US ATLAS Research Program*, meaning there is no formal MOU process to designate sites as Tier-3s and no formal control of the program over the Tier-3 resources. Tier-3's are the responsibility of individual institutions to define, fund, deploy and support.

However, having noted this, we must also recognize that Tier-3's must exist and will have implications for how our computing model should support US ATLAS physicists. Tier-3 users will want to access data and simulations and will want to enable their Tier-3 resources to support their analysis and simulation work. Tier 3's are an important resource for U.S. physicists to analyze LHC data.

One important question is to what extent the Research Program should support Tier-3's? For example, would we require that Tier-2 centers provide wide-area file-systems that Tier-3's can access? What level of software install support could Tier-3 expect (if any)?

This document will define how Tier-3's should best interact with the US ATLAS (and ATLAS) computing model, detail the conditions under which Tier-3s can expect some level of support and set reasonable expectations for the scope and support of US ATLAS Tier-3 sites.

Tier 3's in the ATLAS/US ATLAS Computing Model

The ATLAS computing model describes a hierarchical distributed virtual computing facility within which are defined Tier-1 and Tier-2 computing centers having certain specific MOU agreed roles and capacities to be used for the benefit and at the direction of ATLAS as a whole. The U.S. ATLAS Research Program management, together with international ATLAS, decides how these MOU pledged resources are used. This is accomplished in the U.S.

Resource Allocation Committee (RAC)[16]. In this model the primary functions of the Tier-1 are to host and provide long term storage for, access to and re-reconstruction of a subset of the ATLAS RAW data (20% in the case of the US Tier-1), provide access to ESD, AOD and TAG data sets and support the analysis of these data sets. The primary functions of the Tier-2's are simulation (they provide the bulk of simulation for ATLAS), calibration, chaotic analysis for a subset of analysis groups and hosting of AOD, TAG and some physics group samples.

US ATLAS has acted to establish compute capacity beyond the capacity it has pledged to meet the obligation of international ATLAS to be used specifically for the benefit US ATLAS physicists. This US ATLAS specific computing is located at the Tier-1 and Tier-2's making use of the infrastructure and operational expertise required there anyway, at a scale of 50% (for the Tier-1) of the level of the capacity being pledged to international ATLAS. US ATLAS decides how these resources are used by means of the Resource Allocation Committee, not the local Tier-1 or Tier -2's or international ATLAS.

Tier-3 sites are institution-level non-ATLAS or US ATLAS funded or controlled centers/clusters which wish to participate in ATLAS computing, presumably most frequently in support of the particular interests of local physicists (physicists at the local Tier-3 decide how these resources are used). These are clusters of computers which can vary widely in size. It should be noted that substantial institutional funding to originate such clusters is potentially available, and that they could make a real contribution to the impact of US ATLAS on the overall ATLAS physics output. As such, there is considerable value in providing some level of technical support to these sites.

Support issues (financial, technical expertise, services)

- Individual ATLAS institutions are expected, out of their local resources, to buy individual physicist's equipment, laptops, desktops, printers, etc.
- An individual physicist's share of the ATLAS and US ATLAS resources (at Tier- 1 and Tier-2's) in combination with modest local computing resources (which could be just a modern desktop machine for each physicist) should be sufficient to accomplish required computing tasks for ATLAS and for effective participation in physics analysis.

[16] http://www.usatlas.bnl.gov/twiki/bin/view/AtlasSoftware/ResourceAllocationCommittee

- The Tier-1 and Tier-2's have as primary responsibilities to support such analysis by their users with capacity shares and priorities being established by the RAC for US ATLAS controlled resources together with international ATLAS management for the resources pledged to ATLAS as a whole.
- Sites having significant institutional or base grant-funded computing centers or clusters are encouraged to use them for analysis or other ATLAS computing activities.
- Support from the Tier-1 and Tier-2's to such Tier-3 centers in terms of expertise (install, configure, tune, troubleshooting of ATLAS releases and the OSG stack) and services (data storage, data serving, etc.) follows from responsibility to support the US ATLAS user community. This support would have to be limited to Tier 3 sites with standard ATLAS operating systems.
- Larger Tier-3 sites should be or should become participants in OSG and so get additional technical support via that path.

Part of our task is to set reasonable expectations for the size and scope of Tier-3 centers. We recognize that there will likely be extremely large variances in the amount of computing power and storage at US ATLAS Tier-3 sites. One could reasonably define a Tier-3 as anything a US ATLAS institution so designates, larger than a single machine. We fully expect that some Tier-3 sites may have resources to rival a Tier-2 (or perhaps even the Tier-1!). Our goal is not to constrain the definition of a Tier-3 but to determine a reasonable capability for a Tier-3.

The typical scaling from the MONARC model was to assume that the Tier-0 would provide about 1/3 of the total resources for an LHC project and the integrated Tier-1?s would provide about 1/3 with the last 1/3 provided by the integrated power of the global Tier-2's. In the US ATLAS case this implied that the five Tier-2's would each contribute roughly 1/5 of the Tier-1. Although Tier-3's may be any size, we expect most of them to be smaller than a Tier-2.

Alternatively we could estimate a suggested Tier-3 capacity by determining the type of activities a Tier-3 would be expected to support and scale accordingly. This is perhaps the best means of determining what a "typical" Tier-3 requires in computing power, network connectivity and storage.

We envision the following to be typical examples of uses of a Tier 3:

- Interactive analysis of Ntuples. This requires no direct connection to the ESD or AOD, but it does require access to the data when these Ntuples are generated.
- Development of analysis code. This would motivate a local copy of a small number (perhaps a few thousand) of ESD, AOD, or RAW events. It would be desirable for at least some fraction of these events to be complete "vertical slices"—having the RAW, ESD, AOD and TAG for the same events.
- Running small local test jobs before submitting larger jobs to the Tier-1 or Tier-2 via the grid. This would motivate similar sized copies of the data as above. It also motivates having access to at least the appropriate subset of the TAGs at the Tier- 3, because this is the same selection mechanism that will be used when the full scale job is run,
- Running skimming jobs of the Tier-1 and Tier-2's via the grid, and copying the skimmed AOD (or rarely ESD) back to the Tier-3 for further analysis. The output of this skim must be a very small subset of the AOD of order a few percent.
- Analyzing the above skimmed data via Athena.
- Production of MC samples of special interest to the local institution.
- For larger Tier-3 centers, opening those resources to ATLAS managed production as well as individual ATLAS physicists via OSG Grid interfaces and the ATLAS VO authentication, authorization and accounting infrastructure. Guidance for establishing policies for queue priorities and/or storage may be discussed in the RAC.

These use cases can be met by large or small clusters at Tier-3 centers with the standard OSG software suite installed as well as ATLAS releases, the ATLAS Distributed Data Management end user tools (DQ2), and potentially TAG databases or files. This is a well established process at the U.S. Tier-1 and Tier-2 sites (though some problems are still being worked out) and we expect that support for installing these software suites will be the extent of U.S. Research Program support at Tier-3 centers.

Summary

- Some local compute resources, beyond Tier-1 and Tier-2, are required to do physics analysis in ATLAS.
- These resources are termed Tier-3 and could be as small as a modern desktop computer on each physicist's desk, or as large as Linux farm, perhaps operated as part of a shared facility from an institution's own resources.
- Resources outside of the U.S. ATLAS Research Program are sometimes available for Tier-3 centers. A small amount of HEP Core Program money can sometimes leverage a large amount of other funding for Tier-3 centers. Decisions on when it is useful to spend Core money in this way will have to be considered on a case by case basis.
- Support for Tier-3 centers can be accommodated in the U.S. Research Program provided the Tier-3 centers are part of the Open Science Grid and that they provide access those resources with appropriate priority settings to US ATLAS via the VO authentication, authorization and accounting infrastructure.

B.1 Reaction to the White Paper

The Charge to the Task Force asked whether the White Paper of 2006 was still relevant. This was discussed in one meeting of the Task Force and the conclusions were the following:

- The White Paper was written before some major changes to the ATLAS Analysis model were formulated, in particular the designation of the DPD formats has some (potentially positive) benefits for university-scale computing centers in that some skimming and thinning would already have been done in the process of producing D^2PD or D^3PDs .
- The tasks assigned to Tier 3s above are an appropriate minimum set of capabilities for U.S. ATLAS campus-based physicists.
- The Tier 2 simulation burden and the apparent tightness in the analysis resource structure led the Task Force to conclude that a deeper structure will likely be necessary.
- The White Paper is correct in noting that Tier 3 centers are locally controlled.

- The Task Force felt that it would be useful to characterize classes of Tier 3 centers in order to establish a vocabulary and to quantify goals that university groups might seek to reach in the building up of their groups.
- The support model envisioned by the White Paper would probably not be sufficient in order to build out most university groups' capabilities from T3w→T3g or T3g→T3gs, etc.

C ATLAS Glossary

A glossary is provided which has been taken from a variety of ATLAS sources including the TDR [11] as well as [6] and `https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasGlossary`.

AOD Analysis Object Data: a summary of the reconstructed event, with sufficient information for user analysis. Thought to not necessarily contain cell-level information.

Athena ATLAS offline software framework (also the Greek goddess of wisdom).

AUG Augment: Not an ATLAS official term, but used here to represent the addition of derived or calculated information within the DPD production process.

Castor CERN Advanced STORage Manager—the T0 tape system (also the mortal twin of Pollux).

dCache Disk Cache System which transparently manages the storage and exchange of data spread across various storage devices.

DM Data Management.

DPD Derived Physics Data: Subset—or enhancement—of ESDs or AODs.

D^1PD Primary DPD, or "D-one": largely a skim of the AOD.

D^2PD Secondary DPD, or "D-two": a further skim, thin, and/or slim of the AOD, but also with derived information added, or "augmented".

D^3PD Tertiary DPD, or "D-three": The final official stage of DPD, a flat `ROOTtuple`.

DQ DonQuijote, the distributed data manager (also the chivalrous man from La Mancha).

EDM Event Data Model.

EF Event filter: the third trigger level.

FDR Full Dress Rehearsal: A test of the entire software and analysis chain with simulated data.

HLT High Level Trigger: the level2 and level3 (EF).

OSG Open Science Grid.

PanDA Production ANd Distributed Analysis system: the system for distributing job and storage "tokens" within the Tier structure for reserving computing resources and data sets.

pAthena a python script for access to OSG resources via the Panda system.

POOL Pool Of persistent Objects for LHC: data management for managing multiple files.

skimming Process of removing entire event records, aka selection.

slimming Process of reducing the stored information from within containers within event records.

TAG Very brief event-level metadata meant for selection, not for analysis.

TDR Technical Design Report.

thinning Process of excluding containers or object from within event records.

VO Virtual Organization.

D Typical Hardware and SI2k Specifications

The standard LHC benchmark for comparing computing element capability has been the "SpecInt" unit which is used to periodically evaluate contemporary processors for their integer based performance. This has been a more accurate measure over floating point benchmarks for ATLAS software.The LHC history to date has used a standard established in the year

2000, called SI2000 or SI2k. However, this is now obsolete in the industry and the new standard, SPECInt2006, seems to be nonlinearly related to the SI2k measurements. So, comparing the future with respect to the past will be somewhat cumbersome as manufacturers are not "past dating" their modern equipment to the SI2k measure and to date new processors have not been re-standardized by anyone else.

For modern processors of the 3GHz variety, a standard unit is multiples of approximately 1000 SI2k units, or 1kSI2k. Various sites have attempted to measure this quantity themselves. Figure 22 [15] shows a collection of measurements for standard processors in use now.

Table 22: Estimates of SI2k values collected from various sources for popular processors. From [15].

processor	nickname	Padova	HEP	HEPIX	OSG	BNL
Intel X5355	clovertown	2755	1322	1413	2178	
Intel E5345	clovertown	1190	1267	1889		
Intel E5335	clovertown	2123			1678	
Intel 5160	woodcrest	3161	1505	1602	2420	
Intel 5440	harpertown					2264
Opteron 270		1282	941	1056	1452	1270
Opteron 2214		1352	965	1097	1518	
Opteron 2216						1625
Opteron 2218		1648	1193	1347	1827	1625
Opteron 285		1692	1225	1383	1787	
Opteron 280		1549	1121	1266	1683	
Xeon 3.2 Hz		1516	855			1290
Xeon 3.06 Hz		1427	1166	1402	1169	945
Xeon 2.8 GHz					1123	
Xeon 2.4 GHz			1055	1264	911	747
PIII 1.25 GHz		611	299	319	501	
Opteron 275		1389	1005	1135	1521	1341

E Characterization of Tier 3 Sites

E.1 T3gs

Tier 3gs systems are meant to be substantial clusters with the same services as a Tier 2 center, but with fewer computing and storage elements. As a benchmark, we take as a generic example a single, 42U-rack system as shown in Figure 21.

This generic system can be built of standard components using currently available capacities and pricing as shown in Table 23. This strawman system would produce about 10kW of heat and so cooling infrastructure would be required. Such a system would provide approximately 320kSI2k processing.

Figure 21: Generic single-rack T3gs system.

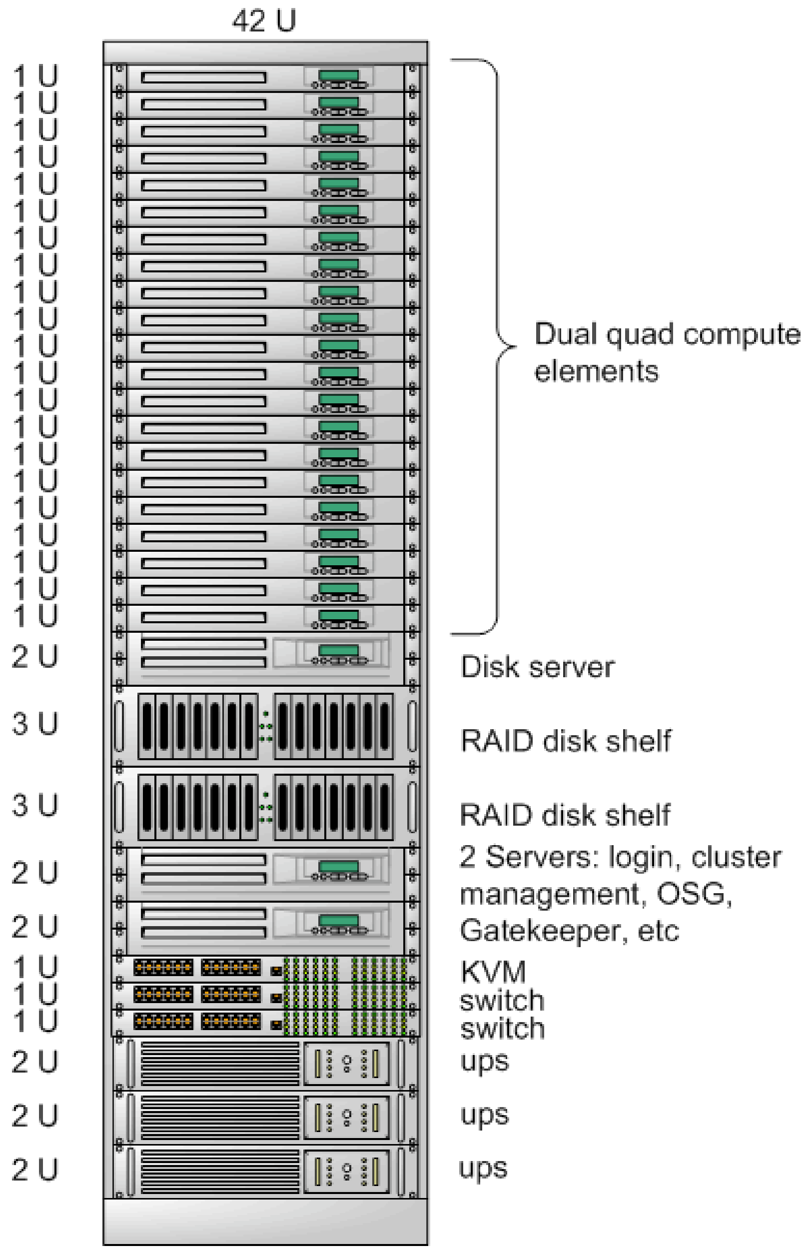


Table 23: Strawman T3gs system designed to fit in one, 42U rack with maximum processing and storage possible. Other systems are certainly possible. At added expense and slightly reduced capability, but with considerable simplification in cabling, etc., a blade-based system would fit in a rack as well. Such a system would provide approximately 320kSI2k processing.

component	typical model	quantity	unit cost, k$
UPS	DELL	3	1.0
switch	DELL PowerConnect 48GbE, portmanaged	2	1.5
servers	DELL PE2950 E5440 processor, 2.83GHz, 32GB RAM, 250GB drive	3	4.2
compute elements	DELL PE1950 E5440 processor, 2.83GHz, 16GB RAM, 250GB drive	21	2.4
storage elements	DELL MD1000	2 (24TB, usable)	5.4
KVM	Belkin	1	1.3
rack			1
total cost			$82.1k

E.1.1 The University of Illinois T3gs Project

As described in this Report, a primary motivation for development of Tier-3 sites is to provide enhanced flexibility within the US ATLAS computing GRID. This flexibility is not only to utilize the significant university and laboratory-based resources to increase the overall computing capacity for steady-state operations, but is also to made available additional resources in times of intensive need (e.g. data (re)processing or Monte Carlo simulation) and to avoid utilization of precious Tier-2 resources for jobs that could be done just as easily at a local site (e.g. D3PD analysis, systematic uncertainty evaluations, pseudo-experiment generation, NN training, Matrix Element calculation, etc). Properly configured and supported Tier-3 centers provide natural points of expansion of the overall ATLAS computing capacity. The process of deploying a Tier-3 site that is integrated into the ATLAS computing model also has the benefit of distributing the knowledge of scientific computing around the collaboration, which has its own intrinsic value.

The T3gs is the most flexible of the recommended Tier-3 sites and can be thought of as functionally (not hardware or capacity) equivalent to a Tier-2 site. The T3gs is distinct from a Tier-2 in that it is locally funded and hence its resources are under compile local policy control. However, the T3gs supports pAthena job submission, DQ2-based data handling, and possibly its own LFC instance, and can therefore is flexible enough to be used in general ATLAS production, as a need arises. To be usable in this manner, it is expected that a T3gs site have sufficient administration to be robust and posses substantial CPU, storage, and network capabilities.

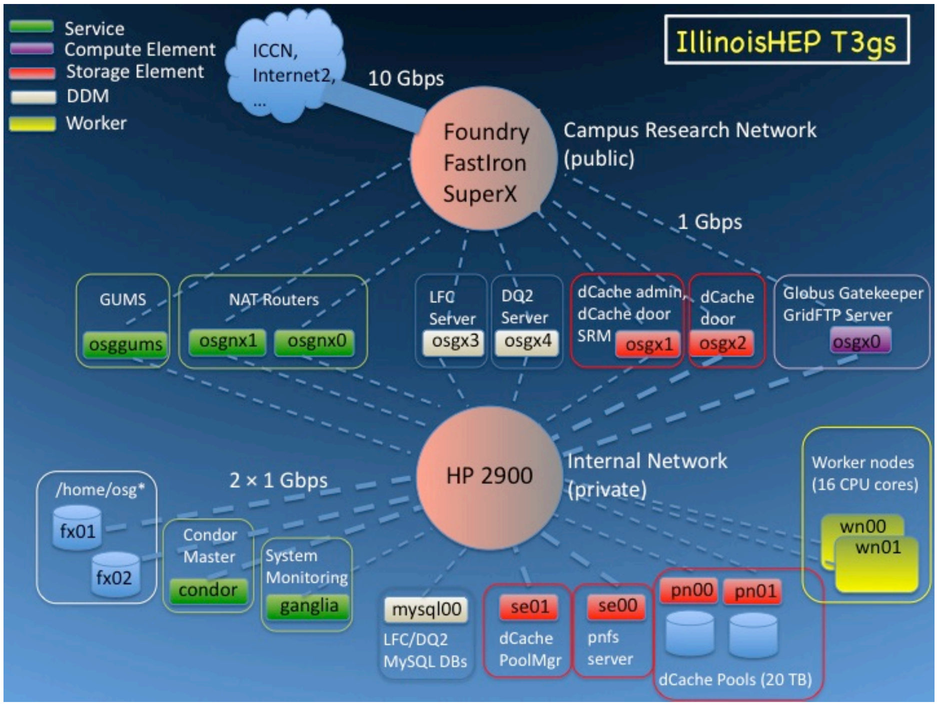
The Illinois Group has deployed a T3gs system. This is the IllinoisHEP OSG Grid site which has been operational (however, not will full T3gs services until recently) since February 2008. The hardware is located in Loomis Laboratory of Physics at the University of Illinois in a room with sufficient cooling and power to support several racks worth of hardware. This site also has a direct connection to a "Research Network" which avoids the campus firewalls (potential bottlenecks) and provides 10 Gbps connecting to ICCN and Internet2.

Rather than focus on a large scale deployment of CPU and disk resources at the onset, the approach has been to deploy the a small amount of CPU and disk resources and focus on getting the required required services to work with the rest of US ATLAS computing. In this way, the resource utilization can be monitored under typical usage to look for bottlenecks and problematic components. This deployment has been accomplished at the time of writing, with much more testing to be done.

The purpose of this appendix is to detail the current configuration of the Illinois T3gs site, primarily from a hardware and services perspective. No attempt is made to detail that installation process in getting this to work, as this information will be documented elsewhere. This is also not to be read as a recommended hardware configuration, as many of the nodes will be upgraded once the system is in operation.

The IllinoisHEP site currently consists of 19 nodes divided among 5 classes of machines. These are characterized as Service, CE, SE, DDM and WN. These machine are interconnected using two network switches, one of which serves the public internet, the other the private, internal only network. The CPUs are all Intel based (Pentium III and up) with memory from 1GB up to 16GB. All nodes run SL 4.7 except one which is SL 5.1). Some nodes use SCSI disks; others are SATA. Some disks are JBOD, others are in hardware RAID subsystems using RAID 5 subsets. Everything is connected

Figure 22: The IllinoisHEP T3gs



to a Raritan Paragon KVM system. An overview is shown in Figure **??**.

The Network The network switches used are Foundry FastIron SuperX and an HP 2900. The SuperX connects to the campus core at 10Ge (and thus ICCN, ESnet, etc at 10 Gbps). This switch serves those nodes on the public network only. It has over 128 Gigabit ports but only 8 of these are used by the Tier-3 nodes. These ports are on a campus VLAN called the Campus Research Network (CRN). This network completely by passes the campus firewall systems (restricted to about 3 Gb), thus increasing the potential throughput to many Gbps (up to a 10 Gbps). This switch is provided, controlled and maintained by the Illinois campus network group (CITES) and thus the site administrator has no ability to modify any of its configuration. This prevents bonding NICs on the public nodes.

The second switch is an HP2900 and is used solely for internal private network connections. It has 48 Gigabit ports and two 10 Gbps ports for expansion. This switch belongs to us and is thus under local administrator control (important for bonding).

Eight of the nodes in the Tier-3 are dual NICs and connected to both the public and private networks while the other 11 nodes are only connected to the private network. All network connections are Gigabit. Most, but not all, of the nodes on the private network use bonding to connect two NICs to the internal switch, doubling the bandwidth.

The Classes Our nodes reside in 5 different classes referred to as Service, CE, SE, DDM and Worker. Service nodes are those which provide needed services to all the nodes in the other classes, such as file serving (NFS), Condor, GUMS, SQL, NTP, DNS, NIS, etc. The CE class is the compute element; SE is the Storage Element; DDM is the Distributed Data Management (LFC/DQ2); WN are the Worker Nodes. Each is interconnected in various ways.

Service Class (7 nodes: 2 public, 5 private)
Node names: fx00, fx01, osggums, osgnx0, osgnx1, condor, ganglia

The two nodes, fx00 and fx01, are file servers for all the nodes at the site. They serve via NFS various file systems that reside on two old (make that very old) Promise RM8000 subsystems. The servers themselves are 2U dual Intel Xeon (2 GHz with HT), 2 GB and Adaptec 39160 SCSI controllers and dual Gb NICs. The file systems reside on RAID 5 subsets and are initialized as ext3. This nodes are connected only to the private network with both NICs bonded to the HP2900. These nodes run SL4.7. These nodes are also the NIS master/slave for all the other nodes. NIS, though not very secure was easy to setup. It is only accessible on the private network and locked to only our sites nodes.

FX00 serves the following:

```
/usr/local     Usual, plus condor executables, test scripts, etc
/home/atlas    Home areas for users
/home/osguser  Home areas for service accounts (usatlas1, etc)
```

FX01 serves the following:

```
/home/osg/WN                Worker node VDT installation ($OSG_GRID)
/home/osgstore/app          Applications ($APP)
/home/osgstore/data         Data area      ($DATA, $TMP)
/home/osgstore/gsiftp       GSI ftp area
/home/osgstore/site-read    Site read
```

`/home/osgstore/site-write` Site write

The node osggums is the GUMS server and provides the authentication service for all the nodes at the site. This node is a 1U dual Intel Xeon (2.66 GHz with HT), 2 GB memory, SATA drives and dual Gb NICs. It is connected to both networks, however all communication for the authentication service if via the public network. This is because the host certificate is registered for osggums.hep.uiuc.edu and you cannot have two certificates. Thus you an only perform this service via the public network. The private network is to allow this node to NFS mount files systems such as /usr/local, /home/atlas, /home/osguser. This nodes runs SL 4.7. Gums was installed with VDT 1.10. It has its own set of CAs installed on a local disk with Gums.

The two nodes (osgnx0, osgnx1) are NAT routers which provide connections for all nodes on the private internal only network to the public network. This allows updates to take places as well as data transfers for the dCache system and worker nodes. These nodes are 2U dual Pentium III (1 GHz), 1GB of memory and two 1 Gbps NICs. One of the NICs is connected to the public network; the other to the private network. They run SL 4.7 and use IPTABLES to provide the NAT service.

The node condor is the Condor master. It is 1U dual Xeon (2.66 GHz with HT), 2 GB memory, SATA drives and dual Gb NICs. It is only connected to the private network, so both NICs are bonded to the HP2900. This node runs SL 4.7 and currently Condor 7.1.0.

The node ganglia is our Ganglia server. It is a 2U Pentium III (1 GHz), 1 GB memory and a single Gb NIC. It is attached only to the private network. It runs SL5.1 because that is what the newest version of Ganglia requires.

CE (Compute Element) class (1 node: 1 public)
Node names: osgx0

This node is the Compute Element. It is a 2U dual Xeon (3.0 GHz with HT), 2 GB memory, four Gbps NICs (3 in use) and local SCSI disks. It connects to the public network with a single Gb NIC but to the private with 2 bonded Gb NICs. It runs SL4.7 and currently has VDT 1.10 installed on a local disk. Its has it own set of CAs installed on this local disk. This CE area, /home/osg/CE, is NFS exported to the other nodes in the site (such as WN and DDM) so we have one synchronized copy of CAs updated on on the CE. This node NFS mounts the /home/atlas, /home/osguser and /home/osgstore from the file servers. Please note that /home/osgstore/tmp ($WNTMP) is a local disk on each node to avoid high NFS traffic for this temp space.

SE (Storage Element) class (6 nodes: 2 public, 4 private)
Node names: osgx1, osgx2, se00, se01, pn00, pn01

These nodes comprise the dCache based storage element. The two nodes on the public network are osgx1 and osgx2. The node osgx1 is the admin, http, srm and a door, The node osgx2 is a door only. The other four nodes, se00, se01, pn00, pn01 are only connected to the private network. The node se00 is the Poolmanager; se01 is the pnfs server; pn00 and pn01 are pool nodes.

- OSGX1: 1U dual Xeon (2.6 GHz with HT), 2 GB memory, SATA, four Gbps NICs
- OSGX2: 2U dual Pentium III (1 Ghz), 1 GB memory, SCSI, four Gbps NICs
- SE0xx: 2U dual Xeon (2.0Ghz with HT), 2GB memory, SCSI, two Gbps NICs
- PNxx: 2U dual Pentium III (1 Ghz), 1 GB memory, SCSI, one Gbps NIC, Adaptec 29160

The RAID subsystem attached to the PN nodes is a Promise VTrak M610p SCSI/SATA. It has 16 1.5 TB Seagate disks, split into two RAID5 subsets of 8 drives each. The R5 is then broken up into four 2 TB logical disks and one 1.5 TB logical disk. Each pool node then has 5 pools, 9.5 TB. The dCache then has 19 TB of usable space in pools.

WN (Worker Nodes) class (2 nodes: 2 private)
Node names: wn00, wn01

These nodes are the worker nodes for the site. The IllinoisHEP T3gs has only two of these at present, however this is a simple point of expansion. These nodes are dual quad core Intel Xeon (2.33 GHz), 16GB memory, SATA and two Gb NICs. Only one NIC is connected to the private network on the HP2900. These nodes mount the /home/osg/WN area from fx01 and the /home/osg/CE area from osgx0 (for the CAs).

DDM (Distributed Data Management) class (3 nodes: 2 public, 1 private)

Node names: osgx3, osgx4, mysql00

This DDM class is our LFC and DQ2 servers with a MySQL server for their databases.

The MySQL server, mysql00, is a SQL server for the LFC and DQ2 databases. It has MySQL 5.1 installed on an SL4.7 system. The node is a dual Pentium III (1 GHz) with 1GB memory, SCSI and one Gb NIC. It is attached only to the private network. The databases are currently stored on a JBOD SCSI disk but need to be on a RAID system to help from loosing these databases.

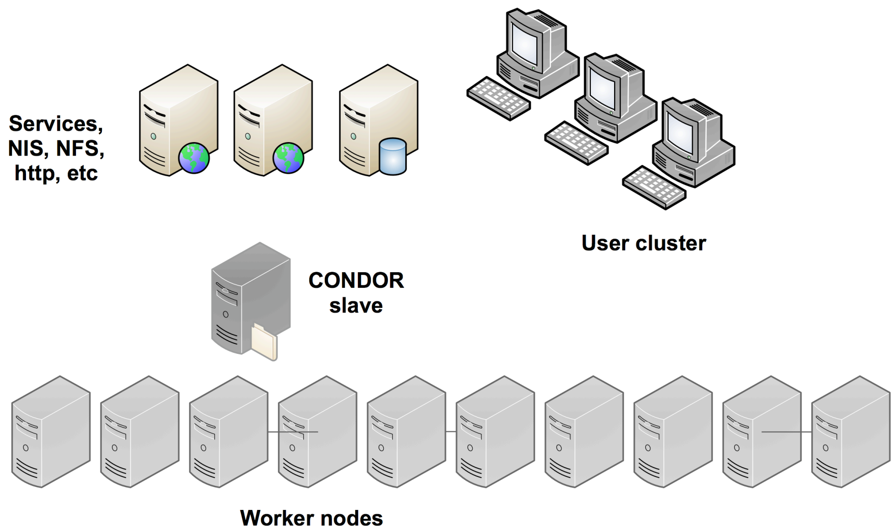
The node osgx3 is the LFC server. It is a dual Pentium III (1 GHz), 1 GB memory, SCSI and two Gb NICs. One NIC is on the public network and the other private network. This node is still being configured for use.

The node osgx4 is the DQ2 server. It is a dual Pentium III (1 GHz), 1 GB memory, SCSI and two Gb NICs. One NIC is on the public network and the other private network. This node is still being configured for use.

E.2 T3g

In contrast to the T3gs, the T3g concept is one which can be housed in an institution without special infrastructure for cooling. As such, it is tower-based and the towers themselves could be housed in a single geographical location as befits a department's computing facilities. Or, the individual towers could be distributed throughout a group's office/lab areas.

Figure 23: Generic tower-based T3g system.



As a benchmark T3g we chose a 10 tower system with 2 TB per tower. The processors chosen are Intel hypertown class, but the server modules are commodity PCs. Scaling up from this minimal system can be envisioned in a variety of directions: more memory for worker towers, more storage, more capable server nodes, etc. However, a medium-sized group would be able to do significant analysis with this system. Such a system would provide approximately 120kSI2k processing.

Notice that if department infrastructure allows, a T3g system could be built into shared racks. Such a system with RAID storage could be scaled down from the T3gs model shown in Figure 21. An example is outlined in Table 25. Such a system would provide approximately 160kSI2k processing. Note, it would add a heat load of approximately 5KW to whatever existing infrastructure existed previously.

Table 24: Minimal strawman T3g system. Such a system would provide approximately 120kSI2k processing.

component	typical model	quantity	unit cost, k$
switch	Cisco 1GB	1	2.5
worker towers	Intel-based E5410 2.33GHz, 2 TB storage 8GB RAM	10	2.0
server elements	DELL PE1950 E5440 processor, 2.83MHz, 16GB RAM, 250GB drive	4	0.5
total cost			$24.5k

Table 25: Strawman T3g system designed to fit into an already existing rack. Other systems are certainly possible. At added expense and slightly reduced capability, but with considerable simplification in cabling, etc., a blade-based system would fit in a rack as well. Such a system would provide approximately 160kSI2k processing.

component	typical model	quantity	unit cost, k$
UPS	DELL	1	1.0
switch	DELL PowerConnect 48GbE, portmanaged	1	1.5
servers	DELL PE2950 E5440 processor, 2.83GHz, 32GB RAM, 250GB drive	1	4.2
compute elements	DELL PE1950 E5440 processor, 2.83GHz, 16GB RAM, 250GB drive	10	2.4
storage elements	DELL MD1000	2 (24TB, usable)	5.4
total cost			$41.5k

E.2.1 The ANL T3g Project

Unlike the traditional cluster where a file storage server is separate from compute nodes, each node in the ANL PC farm (PCF) cluster acts as both storage and compute node. This cluster configuration is possible due to: 1) the relatively recent advent of cheap and large disk drives; 2) availability of commodity multicore processors; 3) availability of the Condor batch system [16]; 4) the particular nature of the jobs. Specifically, it is assumed that the jobs run over all available data in sequential order, i.e. no chaotic access to the data located on different nodes need be supported (although random access to data within one single Linux box is possible). The type of analysis job to be used (e.g. Athena, ROOT, PyROOT, Python) is unrestricted. There are several characteristics of such design [3]:

- Low cost since there are no expensive disk servers and switches required. Off-the-shelf hardware with open source Linux is used;
- Small effort in management, easy to set up and maintain;
- Scalable. Any number of slaves with more file storage can be added;
- Low network load. No I/O bottleneck, since only locally stored data on each PC are processed, thus the vast bulk of the I/O takes place within the nodes and not through NFS. The network is used only for job submission or retrievals;
- Robust. A broken disk or CPU affects only a small fraction of data which can be recovered using the Grid and the DQ2 client software independent of the rest of the system.

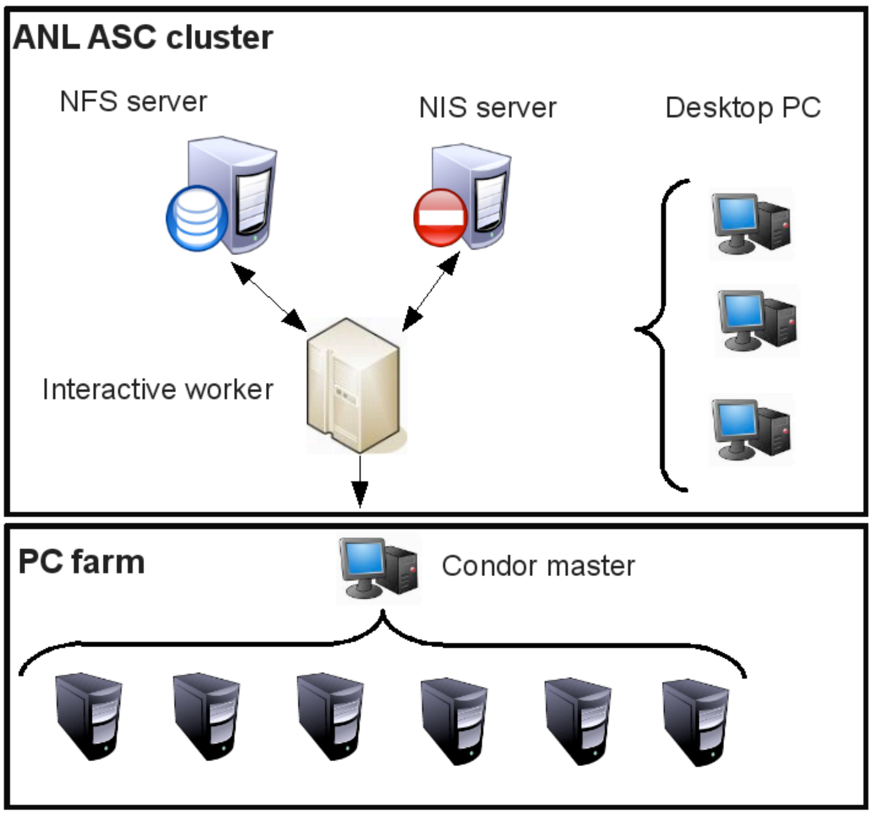


Figure 24: A schematic representation of the existing ANL ASC cluster with planned upgrade by adding a PC farm.

At present, the ANL PCF consists of 3 nodes (24 cores, 6TB file storage), which will be expanded to 10 nodes (80 cores, 20 TB file storage) in future. Figure 24 shows a schematic representation of the existing ANL ASC cluster. The most relevant part for ATLAS data analysis consists of:

- A main worker server that connects the PCF nodes via NFS file storage using a 1 Gb network switch.

- A server that serves as a Condor master for the PCF. This computer is used as a backup interactive worker node.
- Three "slave" nodes with each having 8 computational cores and 2TB for local data storage. The overall number of CPU cores was 24.

Each Condor slave has 8-core processor based on the Intel(R) Xeon(R) CPU E5405 @ 2.33GHz, with 8 GB RAM and 2 TB SATA hard drives. The system disk holding SL4.6 operating system, ATLAS release and the Condor home directory is 250 GB big. The price per box was about $2k.

To run in parallel over data samples (AOD, DPD's, ROOT ntuples) scattered over multiple Linux boxes, the ArCond (ARgonne CONDor) program [4] developed at the ANL ASC was used. The ArCond package was designed as a front-end of Condor [16] for data data discovery, job submission, retrieval, merging of output ROOT files, and downloading data on each PC node in parallel using the DQ2 program. Since September 2008, about 5000 athena and compiled C++/ROOT jobs were processed on 24 cores with a very low fault rate (less than 0.01% per job).

To benchmark the PC farm, a fraction of the JF17 data sample was used (inclusive jets with $P_T(jet) > 17$ GeV). 1.5M events were stored 5500 AOD files (240 GB). The data sample was copied to the PC farm from the Grid and divided equally among the three slave computer nodes. For our benchmark tests, the JF17 sample above can be processed for about 180 minutes. Thus, the benchmark test indicates that processing of 210k/h by a single core is possible. For the designed goal of 80 cores, 1.7M AOD events per hour should be possible to process. This indicates that the desired level of performance has been achieved.

The benchmark was also done for a significantly reduced file format, ROOT ntuples, which are similar to physics DPDs. For this test, the complete JF17 sample (about 8M events) was converted into ROOT ntuples. The ntuples were divided equally among 3 computers nodes. The processing time for a compiled C++/ROOT program submitted using the ArCond was about 4 min for each core. Assuming that this file format will be used for final analysis, the processing of ROOT ntuples corresponding to 10 fb^{-1} of data should take about 3h.

More detailed description of the ANL PC farm setup can be found in [3].

E.2.2 The Duke T3g Project

From Doug Benjamin: Some Tier 3 setups will be comprised of rack mount hardware due to the existing policies at the University or a desire to have the hardware in the most space efficient configuration. This section describes such a configuration and will contain information on services and hardware required.

This Tier 3 configuration is comprised of interactive nodes where the users can connect. The majority of computing capacity is provided by worker nodes managed by the Condor batch system. Condor requires a master node which will handle job scheduling. Common software including Atlas software can be stored on a NFS file server with mounts on all worker nodes. Storage can either be distributed amongst the worker nodes, located in large RAID file servers or a mixture of both. The storage should be managed by the xrootd system and requires a redirector node. The local data should be cataloged and monitored with a MySQL database located on a database node. A Tom cat web server can be used to display the monitoring information on the web. Given the advances in computer virtualization, virtual machines can be used instead of dedicated hardware. Depending on the network policies of Campus network, the Tier 3 may be connected via managed switches.

The interactive node or nodes depending on the number of users actively using the Tier 3 are used for local code and analysis macro development, use of Atlas client tools for grid job submission (`pAthena`), Atlas DDM client tools (dq2-get, dq2-ls etc). The interactive node(s) will be used for job submission to the local batch system or an advanced analysis platform such as local Proof farm. It is expected that the user home areas are located on a NFS file server with some level of data security through RAID disk hardware and have an established backup policy and practice to provide some reasonable level of data security. These interactive nodes need to have sufficient CPU capacity and system memory to be use full for the users.

The storage subsystem can either be distributed amongst the worker node, located in large file servers using RAID 5/6 or a mixture of the two. xrootd should be used to manage most of the storage. By using xrootd, the Tier 3 users and administrators have straight forward and efficient means of using and managing the disk capacity. This will become ever more important as storage capacity grows to 50-100 TB range. xrootd contains monitoring that is stored a MySQL database for later retrieval and presentation. The MySQL database can also be used to locate which file server contains a given file. This will provide efficient processing of the data as the jobs can be moved to the data when the data is located on the worker nodes. When the storage is located on dedicated file servers xrootd provides an efficient and robust method for serving the data to the compute servers and has several advantages over NFS. When the data is not located on the worker nodes, then the delivered across the computer network. xrootd can be used with Bestman-gateway and xrootdfs to provide a SRM V2 compliant storage system. When coupled with dq2-get, it is straight forward to automatically distribute datasets across data nodes in the xrootd system.

The computer network fabric within the Tier 3 should sufficiently robust to anticipated load. Storage located on dedicated file servers require more network capacity than storage located worker nodes. Tier 3 sites may find it advantageous to use managed switches. This allows Campus network administrators to route traffic from the Tier 3 cluster to the edge of the campus and onto research networks (like Internet 2).

F Survey of non-U.S. ATLAS Tier 3 Strategies

F.1 United Kingdom

Typically each University group has a local compute/disk cluster, probably 100-300 CPUs and a few tens of terabytes. These are not funded by our agency (STFC) as ”Tier-3s”, but rather as general computing support for groups. We also have a weird setup in the UK (maybe the US is similar) that every Tier-2 is split across several (O(4)) university sites. So we all have some Tier-2 machines as well as our ”Tier-3” machines at each university. In addition there is a (relatively) new UK phenomenon that universities mainly now have some sizeable campus facilities (O(1000) cores) [...] . But every UK group is different. We have no centralised support for Tier-3’s from ATLAS/STFC - each group has a computer manager and typically an ATLAS computing support expert who will look after the local ”Tier-3”. Relative sizes: My guess would be that until recently they have been similar, but in future the T2 is likely to be bigger. and from a different source: the ’Tier 3’ capacity in the UK is mainly a reserved share for the UK users on the Tier 2s (not declared as part of the ATLAS pledge).

F.2 Canada

In no way is our Canadian Tier-2 infrastructure (hardware) and perhaps most importantly the personnel support that will be required so run these facilities sorted out. What we have talked about so far is not to have a dedicated ”Tier-3” center at a few geographic location in Canada but try instead to make sure that each institute can have local computing infrastructure to do these kind of things. [...] Now, most institutes in Canada have already some O(100) cores mini-cluster already, so these should be used as Tier-3s.

The other thing is, I believe in Canada that it is mostly implied that many of the institute's local computing resources (call them Tier-3s if you want) will not be grid sites, we just don't have and can't afford (at least right now) the expertise that would be needed at 11 different institutions to achieve that. It's one thing to imagine a large "Tier-3s" for each institution and it is another thing to secure enough support for all this computing infrastructure. [...] If you define the "Tier-3 system manager" as a postdoc at an institute, yes, he/she would be able to get help from the TRIUMF user support personnel which was hired as part of the Tier-1 center. That also applies with the Tier-2 personnel, if they have any questions, they should contact and work with the people at the Tier-1. in Canada we are [...] "reserving" extra resources for Canadian usage off our Tier-2s. If you were to, say, assume O(100 cores) per institutes, that means O(1000 cores) of "Tier-3s" institutional hardware for all of Canada. We are requesting for our Tier-2s a total of 1.6k/2.6k/5k kSI2k for 2008/09/10 for all of Canada. Say, one core is 3 kSI2k, that means about 500/900/1,600 cores.

F.3 Netherlands

Only one Tier-3 is foreseen, and it will be at NIKHEF with a direct link to the Tier-1 center Support would be provided by the same people

F.4 Spain

IFIC Valencia has a Tier-3, but most other institutions do not. It is documented. The Tier3 and Tier2 are tightly coupled. Their PCs hang in the same rack, software installation is shared up to a level, the CGI /lustre sytem of the storage element is used also for the Tier3. They are independent at the funding and ownership level: Tier2 resources are owned by ATLAS and payed for by the Tier2 project. The Tier3 is funded separately, and dedicated to users at IFIC. User support (to complain about failed ATHENA installations, for tutorials, etc.) is provided by the Tier3 project. There is not really [a policy from the funding agency]. I believe the Tier3 projects will not become the standard approach in Spain. Most institutes will have to finance their Tier3 from the normal ATLAS project. Tier3 is an order of magnitude smaller in term of CPU [than the Tier-2]. Interactive analysis requiring ATHENA is supposed to be performed in the Tier3, but batch analyses should be submitted to the global ATLAS computing system. In case of total failure of the distributed computing model, one could envisage the possibility to boost the Tier3 resources and perform our analyses ”at home”, but this is not the default scenario.

F.5 Germany

[Tier-3’s] are university specific centers that do not have an official responsibility. Actually they can be quite big, certainly of the size of a ’normal’ Tier 2. Their funding, however, is in most cases a onetime issue without a guaranteed continuous support. We are trying to establish that a Tier 3 system manager get software support from experts at Tier 1 or 2 centers. [The funding policy w.r.t. the funding agency] is a very complicated issue, the German funding system is in no way ’normal’ - if not to say it contradicts any reasonable strategy. As a result funding of Tier 3s (and to some degree even Tier 2s) is ad - hoc and depends on the willingness of the university. In the future [the cpu power of Tier-3’s] may be several times the power of standard ATLAS Tier 2s. This will be very difficult to predict.

F.6 Italy

Funding for T3’s is not official from INFN - the money comes out of university or groups base funding. The T1 in Bologna (CNAF) and the T2 in Rome help with support of the T3’s - installation of the ATLAS software and also for middleware support. There is a T3 co-located at the T2 Rome center (which is large - 50 boxes, 200 cores, 24 TB - slowly increasing in size). CPU resources are shared with others in physics as well as biology (!!). Currently used mostly for MC production and analysis.

F.7 France

A general overview of the French Tx system. France research is organized around labs, not so much universities. Most are T2’s with interactive capability (example: I used Lyon while I was at Marseille more than I used the T3 at Marseile). In Paris several institutions have gotten together and and created a T2/T3 - central management that makes some geographical sense.

G Survey of CMS/ALICE Tier 3 Strategies

G.1 CMS

The support of US CMS Tier-3s is shared between OSG, US CMS Grid Services, DISUN, and self-supporting. Efforts within the US CMS Grid Services include:

- Operations Support, Integration, Interoperability,
- Participate in Middleware development, integration, support: Glidein workload management system, security, accounting, information.

Support from OSG includes

- Providing common software and services across many diverse communities.
- Helps site administrators in installation/configuration, usage, security, support.
- Contributes to the WLCG in an equivalent fashion to EGEE.
- Peers with the EGEE to make things work better for ATLAS and CMS and the WLCG in general. OSG also increasingly works with TeraGrid.

The data flow to the Tier3 is strongly tied to their data format. It seems that their AODs are directly readable by root. They have not considered the creation of derived data formats via slimming, skimming, etc... They move data around with their FeDex system and get all the AODs. Their Tier2s will not have all the AODs. Their Tier2s will get the data according to the physics needs of the community clustered around them.

They have approved a plan to build an Analysis Facility at Fermi Lab, similar to what it's done in BNL. The LHC Physics Center (LPC) is developing a Computing Analysis Facility (CAF). The LPC-CAF is a Tier 3 facility and, as such has no specific responsibilities for CMS Operations. Specific responsibilities to CMS for data processing and Monte Carlo are carried out by the US Tier1 and Tier2 centers. The purpose of a Tier3 Center is to bridge the gap between physics analysis capabilities provided worldwide by Tier2s and the individual physicists desktop. While the LPC-CAF will be a significant facility, it is likely that the aggregate needs of US CMS physicists will, at some point, exceed its capability. With this in mind, some ground rules have been proposed for initial use of the LPC-CAF and provide an outline of how priorities would be set.

G.2 ALICE

In ALICE they have a simple (and even a bit ”simplistic”) approach to T3s. For us there is no ”essential” difference between T1, T2 and T3s, but only a gradation:

- T1 have MSS, sign the WLCG MoU and abide to the conditions laid out there;
- T2 do not have MSS (or better are not required to provide MSS, they welcome T2 with MSS if any would exist), sign the WLCG MoU and abide to the conditions laid out there;
- T3 do not sign the WLCG MoU, do not have custodial role, but nevertheless have the same software setup of other centres and participate to the global activities in the same way. In this case their resources are accounted in the global contribution of the FA to the ALICE computing.

Their model is much more a cloud than a hierarchical grid. They only make sure that reconstruction passes and ordered analysis are preferentially executed on T1s for question of data locality. If a centre does not integrate in the ALICE distributed computing environment (AliEn VO-Box and free access to ALL ALICE jobs), they do not complain, but they do not guarantee any support and they do not account these resources in the contribution of the corresponding FAs.

The strength of their model is that everybody profit from the smallest T3 added, because it becomes part of the global grid. A maintained VO-Box may be a high threshold for some centers, but it justifies their effort to support them in return. Remember that they are understaffed and very-severely under funded, and their experience with ”opportunistic” resources is very bad. Usually these are not worth the effort.

A grey area is when a centre decides to install a ”standard” (as far as this exist) ALICE / Proof facility. In this case, at least in principle, they should account this contribution if and only if all ALICE users (in principle) could ask an account there. In practice they have not yet defined a precise policy however, because they want to encourage the usage of Proof that they have found to be extremely useful with their experience with the CERN Analysis Facility. So they do not want to hamper this with strict rules from the beginning.

The weakness of their model is that it assumes a well-working grid. Indeed their grid is working fairly well. For a global view see http://pcalimonitor.cern.ch/map.jsp . The other weak spot is political. When asking resources to FA's, ALICE physicists cannot say that this would profit the national community directly, but that it will improve the global computing infrastructure and, therefore, indirectly, it will help also the national community.

They "support" this "redisitributive" model with their computing rules, which you can find here.

http://aliceinfo.cern.ch/Offline/General-Information/Offline-Policy.html

H Survey of U.S. University ATLAS Computing

A survey of all U.S. ATLAS institutions was undertaken to ascertain the amount of computing, storage, and networking resources available. Figure 18 was drawn from the tables of results in the pages that follow.
The results of the survey are summarized in Table 26.

Table 26: The respondents to the survey of 2008. The responses in the first row totaled 23, which included two national laboratories. The number in the second row with no Tier 3 clusters is 12.

	institution
responded and were included	ANL; Columbia University; Duke University; Iowa State University; Louisiana Tech University; University of Oregon; SUNY, Stony Brook; University of South Carolina; University of Indiana; University of Chicago; Southern Methodist University; Oklahoma University; University of Illinois; Michigan State University; Tufts University; LBNL; University of Texas, Dallas; University of Wisconsin; University of Texas, Arlington; University of Massachusetts; University of Michigan; Boston University; Harvard University; and Hampton University.
responded "no current Tier 3"	University of Washington; Oklahoma State University; University of California, Irvine; University of California, Santa Cruz; University of Arizona; Brandeis University; University of Iowa; Yale University; Northern Illinois University; California State University, Fresno; U. New Mexico; and BNL.
did not respond	University of Pennsylvania; SUNY Albany; MIT; NYU; Ohio State University; University of Pittsburgh; and SLAC.

INSTITUTION	ANL	Columbia	Duke	Iowa State U	Louisiana Tech	University of Oregon	SUNY SB	U. of South Carolina	Indiana U	University of Chicago	SMU
Do you have T3 cluster (yes/no)	yes	yes	yes	yes	yes	yes	not officially	not officially	yes	yes	yes
FTE to serve the T3		0	8	0.07-->0.24	0.5	0.05	0.1	0.05	0.25	0.25	0.2
HARDWARE:											
Number of computers in the T3 cluster	8 PC. 20 cores	20	WN-5, SN-2, FS-1	10	16, 1, 1	2	WN 3, SN 1	2,3,1	51 (48/1/2)	28	30
(worker nodes/server nodes/file servers)	3 servers (with 4 cores),										
	rest are desktops										
Number and type of CPU	Dual Core AMD Opteron(tm)	50 3 GHz cores	4 Opteron 275,	18 (~3 GHz), 1 GB/cpu	(64) 2.33 GHz Intel	4 cores, Intel, 2.4GHz	26 cores, Intel, 2 GHz	6 (4-Intel Xeon	90 1 core 2 GHz Intel	44 Dual Core AMD	60 Pentium 4
	Processor 280		2 Opteron 2218 ,		Xeon 64bit			2.66GHz, 2-Intel	~20 cores 2.2-3 GHz	Opteron(tm)	
			8 Xeon E5430					Xeon X5365 3GHz)	mixed AMD and Intel	Processor 285	
SI2K total units			34.8 k SI2K		1 635 per processor	3500	24*2000	24k		157256	~100,000
Disk storage (TB)	4	10	3	8	10	2	6.8	4	9	80	10
Tape storage (TB)	none	0	0	0	0	0	0	0	0	0	0
Network connectivity	1 Gb	100 Mbps	Campus – Lambda net	GB internal	10 GbE	2.5Gbs Oregon to	Offsite ~ 4-5 MB	Gbit, internet 2	1 GBps	10G	150 Mb(Internet 1),
			2 Gbe to campus net to department switch	(74 port switch), internet 2 external		Internet 2	locally Gigabit				150 Mb(Internet 2)
SOFTWARE:											
Is your T3 cluster in the GRID? (yes/no)	no, but condor is used for	no	yes	yes, OSG	yes	no	no	no	yes	no, but have gridftp	yes
	local runs									providing DQ2 endpoint	
Cluster Monitoring system (for ex. Ganglia)	no automatic procedure	no	no			no	no	no	Ganglia	Ganglia, Nagios	Nagios
Which method has been used to install the	no automatic procedure	none	RPMS		PXE	none	manual		Rocks	"Cloner" - from ACT	RedHat Kickstart
cluster? (PXE, OSCAR,…)											
OTHER:											
any known future purchases	will be expanded to 100 cores	10 dual quad-core	5 WN (8 cores) in 2009	15K for CPU			no plans	1 (2 intel xeon x5365	small number of		
	with 30-40 Tb within 1-2 years	20 TB disk	5-8 TB storage based					3GHz, 3TB)	fast cores		
		(on direct 1 Gbps to cern)	on Xrootd, and								
			Xrootd server								

OU	Illinois	MSU	Tufts	LBNL	UT Dallas	U. Wisconsin-Madison	UTA	U. Massachusetts-Amherst	U of Michigan	BU	Harvard U.	Hampton U.
yes	yes	yes	yes, shared	yes	yes	yes	yes	yes	yes	T2 with some T3	yes	yes
			with University							capabilities		
0.25	1	0.1	2.5	1	0.3	1	0	0.05	0.3	0.05	1	0.5
33, 3, 3	12	33 (30,2,1)	40/1/1	10	1 gateway,19 workers	12/5/20	50	8	100/3/7	4096 shared with T2,	2/8/96	32/2/2
										dedicated 500 cores		
7 P-III, 46 Xeon/P-4	Intel, from 1 GHz to	2x Xeon e5430	40 blades each with	10	dual quad cores	8 Intel 2.66Ghz cores	100 ZEON	11 2GHz CPUs	70 dual core Athlon		4096 x Intel®	2x32 quad-core AMD64.
	Quad Core	(2.66 GHz)	2 E5440 2.83 GHz		one 2.66 GHz, 19	per machine	2.4 GHz		25 dual quad-core Xeon		Xeon(R) CPU	2.2 GHz
			quad cores		2.33 GHz				5 dual dual-core Opteron		2.33 GHZ	
		30*8*1.4k=336k		250K			1000	110	770k	1M dedicated fot T3	10	
										11M shared with T2		
8	10	15	8	15	6.2	150	20	10	62	200		36
0	0	0	0	100	0		0	0		0		16x400 GB
	10GigE to campus core	1Gb/s, 10Gb/s	Infiniband interconnect/	Connected to	Gigabit between	1Gb/s	100 MB/s	Internet 2	10Gb/s	10 Gb/s	10 Gb/s	DS3 (44.7Mb/s)
	ICCN Esnet, 1 GigE to servers	campus network + spare capacity of optical network	Gigabit Ethernet to campus network	Tier 1	nodes, Internet2							
yes	yes	Yes - OSG	No	Yes	Yes	Yes	Yes	No	yes,co-hosted w AGLT2	T2 is on grid, T3-?	yes	not yet, but planned
Ganglia	Ganglia, Gratia, MonaLisa	Ganglia		Yes	?	Ganglia	Ganglia	Ganglia	Ganglia/Cacti/ITAs sistant	Ganglia, Nagios, others	Ganglia	to be determined
RHEL5.2 CD	Pacman	Rocks	RedHat 5.2 with LSF		?	PXE	Rocks	manual	PXE	custom	custom	to be determined
			queues									
			8 TB additional	Intend to double		no	no	will add 10 dual nodes	next FY small increment			No. Plan to put resources
			storage server	the T3 in the next								into OSG
				6 months								