

Tier-3 Implementation Committee (T3IC)

**US ATLAS Distributed Computing Workshop
SLAC**

April 7, 2013

Mark Neubauer

University of Illinois at Urbana-Champaign

Jason Nielsen

University of California at Santa-Cruz



Tier-3 Implementation Committee



- T3IC formed in November 2013 by US ATLAS Ops Program
- Need for such a committee driven primarily by
 - Desire from the US ATLAS Operations Program to formulate a specific **plan-of-action** to implement the core T3TF report recommendations, with estimates of any Ops-supported resources required
 - ❖ Technological solutions discussion in the T3TF Report was deliberately vague to leave that work to T3IC where expertise in relevant areas is concentrated
 - Desire from the funding agencies to receive a **clear message** from US ATLAS regarding needs and how future Tier-3 funding should be spent
- Chairs: Mark Neubauer (Illinois) and Jason Nielsen (UCSC)
- Members: D. Benjamin, K. De, M. Ernst, R. Gardner, A. Taffard, E. Varnes, T. Wenaus.
- Ex-officio: J. Cochran, S. Rajagopalan (Ops), C. Brock (US IB)

Completed report in April/May (2014)



T3IC Charge (Needs)



- To carry out a comprehensive study that proposes a cost-effective implementation plan to address the T3 challenges. [...] Among the questions and issues you should address are:
 - Provide a best estimate of the computing capacities required to satisfy the physics analysis activities in the U.S. over the next five years
 - Address how far the existing T3 infrastructure goes to accomplish these goals and how to make better use of all existing resources to support the U.S. physics analysis requirements
 - Address the incremental capacities that are needed to provide adequate support for U.S. physics analysis



T3IC Charge (Plans)



- [...] Among the questions and issues you should address are:
 - Identify potential implementation plans that address the T3 needs of U.S. ATLAS physicists and evaluate their cost-effectiveness. Your evaluation of the cost-effectiveness should take into account and identify any synergies, efficiencies, institutional or laboratory leveraging, potential for existing or additional funding sources, and possibly other intangibles. A final ranked comparison table should summarize the cost effectiveness of these plans.
 - Identify how the T3 computing resources and personnel would be managed for (centrally, institutionally, etc) for each plan. This is important for any solutions that require Operations Program funds.
 - Identify what can be accomplished within the current Operations Program budget guidance and prioritize additional requests if supplemental funding materializes (from Operations or other sources).



T3IC Activities



- Primary activities thus far have been consideration of
 - Estimate of resources required for physics analysis
 - How extra resources help satisfy the analysis needs
 - Development of candidate implementation plans
 - Comparison of plans (e.g. by cost effectiveness, degree to which they satisfy Tier-3 user needs, ...)
 - Identification of tools/technologies that facilitate Plans
- 1st three charges
- This talk
- Other talks in this session



Plan 1



- Refresh of the current Tier-3 “system”
 - In a climate of constrained budgets, this would be accomplished with a \$ amount of the same order as the procurement done with past ARRA + MRI awards
 - Hardware bought after this year at same \$ amount would likely cover the factor of 2-3 extra resources we estimate we will need to achieve Run 2 physics goals.
 - Need to check the scaling of \$/CPU and \$/TB to confirm this
 - +Tier3InABox? Would address concern about lack of sys admin expertise at some sites



Plan 2



- Re-purpose existing Tier3 hardware and “just use the facilities” (→beyond-pledge resources)
 - This possible plan was raised by a panelist at DOE/NSF review in Columbia
 - In our last T3IC meeting, we agreed that it is important to stand up such a plan for comparison to the others.



Plan 3



- Each willing institution should receive funds to locally deploy a modest system for interactive and light batch use.
 - Last T3IC meeting we agreed to include a Local Resource Unit (LRU) in the report to specify (resources, hardware, present \$\$\$)
 - Appropriate scaling for group size and/or scope of analysis activities
 - The rough LRU size might be 100 CPUs and 5 TB of storage (per user? group?) with remote data access through FAX
 - dual Intel E5-2670 (32 LC) w/ 64 GB → \$6.5k
 - Batch horsepower is in the facilities in the form of beyond-pledge resources and tools are provided to utilize those resources from the local systems
 - Leverage FAX and FAXbox for shared storage
 - ATLASConnect/Condor flocking
 - Panda queues/priority scheme
-



Plan 4



- Plan 3 + a few large Tier-3s shared equally by all of US ATLAS
 - *Tier3zilla(s)*: A small number of large Tier-3s awarded to institutions on a competitive basis after a US ATLAS-wide call and expression-of-interest



Plan 4 & The Tier3zilla(s)



- Institutions that receive project/NSF/DOE \$ for a large Tier3 will need to share the resources with the rest of US ATLAS. Why would any group want to do this? Policy proposal :
 - Hosting group gets priority use up to a fraction F and they provide fair share, opportunistic access. Possibilities:
 - $F = 1$ (too good of a deal for hosting group)
 - $0 < F < 1$
 - $F = 0$ (too unattractive to a hosting group)
- Support is an open question for any Tier3zilla, since anything beyond support of the core infrastructure and functionality would make hosting such a system less attractive



Plan 5



- Plan 3 + some large-ish Tier-3s, each shared among a specific set of US ATLAS institutions
 - *Tier3consortia*: Formation of Tier-3 “federations” or “consortia” which are large (but not likely to be Tier3zilla sized given funding constraints) flexible, pooled resources shared by multiple institutions

Some comments

- Not sure how this would work in practice. Need some elaboration from proponents. Might be merged with Plan 4
 - Consortia to be encouraged (+ T3TF-recommended). But...
 - How would we formalize this to be sure that No Group Is Left Behind?
 - Gerrymandering? Informal agreements, self-organizing?
 - In the limit of a small number of these, we are back to the Tier3zilla model, but without ability to down-select effectively
-



Some non-\$\$ Plans Comparisons



- Plans 1 and 2 could be seen as ~“baselines” (maybe closer to “strawmen”...) that still benefits from new technologies
- Plan 3 addresses issues of
 - non-optimal distribution of resources across US groups
 - more total batch than needed, given those available in facilities
 - aging hardware
 - poor economy of scale
- Plans 4 & 5 address what Plan 3 addresses but additionally that some institutes bring in-kind “stuff” (hardware, people, expertise) to US ATLAS which **should be embraced**



Additionally, Plans 3,4,5 include...



- At least one resource for all of US ATLAS with
 - Interactive login access
 - ATLAS and Grid software and tools
 - Local batch system to schedule jobs onto a modest pool of “local” CPU resources
 - Local storage with quotas that is accessible both from batch and interactive jobs
- ➔ BNL an obvious site for this and serves as a template if another lab would also like to support such a resource for US ATLAS
- Agile, for elastic expansion/reallocation of resources
- FAX/FAXBox and Rucio cache, to alleviate local storage crunch
- PandDA



Motivations for Plans 4,5



Given beyond-pledge resources, why do we need Plans 4,5?

1. Enhance ATLAS campus presence
2. Get some cost-sharing (resources at highly-reduced cost)
 - more resources for US ATLAS overall
3. Access to additional resources (facilitated by Agile infrastructure, ATLAS Connect) during intensive need



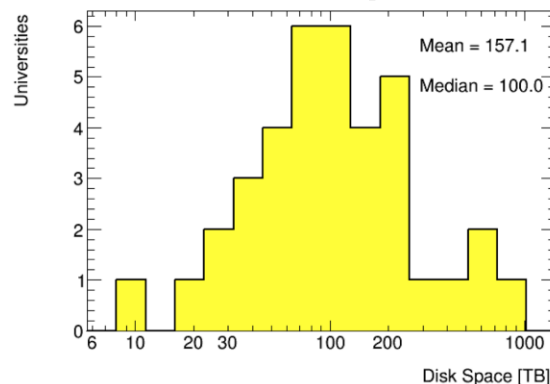
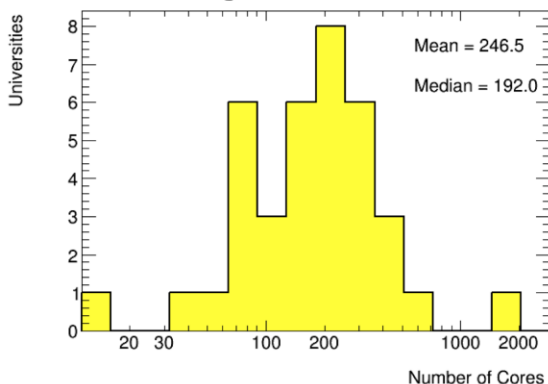
Bonus Material



US ATLAS Tier-3s



- 90% of US universities consider Tier-3 computing as critical to their group's productivity in ATLAS
- The current deployment of the US Tier 3 computing was purchased using ARRA, MRI and university funds



- The majority of equipment was purchased in 2010, w/o replacement
- Two committees reviewed the Tier-3s and produced a report
 - 2009 Tier-3 Task Force. Chair: Chip Brock
 - 2013 Tier-3 Task Force: Chairs: Jianming Qian and Gabriella Sciolla



Resource Estimate



- To address the first three points in the charge regarding the required resources for physics analysis, Doug, Kaushik, Mark, and Anyes were asked to provide more information on their typical analysis workflow. In particular,
 - 1) *required (or actual) turnaround time (wall clock time)*
 - 2) *location of input*
 - 3) *size and location of output*
 - 4) *typical number of remote or local job slots used*
 - 5) *additional requirements or constraints on remote vs. local data**and, for each of the following analysis stages:*
 - 1) *skim of group D3PD to secondary D3PD (or small ntuple)*
 - 2) *loop over small ntuple to tune cuts, etc., and make plots*
 - Several of these are now in the T3IC Google Drive area
-



2013 T3TF Report



- Six recommendations, paraphrased below:
 - ① Support the LT3Cs and make strong case for continued LT3C funding
 - ② Provide mechanisms for Tier-3 jobs to expand onto resources outside of the LT3Cs when the LT3C resources are fully utilized. Invest in technologies that give Tier-3 functionality to institutions w/o an LT3C
 - ③ Support the use of WAN data access for Tier-3s
 - ④ Provide all users a sufficient amount of guaranteed storage space located at Tier-1/Tier-2s (where the batch computing resources are)
 - ⑤ Provide capability for users to direct, upon submission, their output from Grid jobs back to their local storage. Output should be retrievable with minimal delay after the batch job is completed
 - ⑥ Provide documentation and organize comprehensive tutorials to train interested users on how to take advantage of new analysis resource
 - Report also included needs estimates and possible solutions
-



Resource Estimate (cont.)



- To scale to 100 fb^{-1} of 14 TeV data, we need best estimates
 - To be useful, estimates should be accurate to $\sim 20\text{-}50\%$.
 - We guess the scaling will be less than a simple x4 factor, but probably more than x2.
 - What are some things we could do differently with x4 data?
 - We will need a multiplicative factor for the number of workflows, including information on how they are shared among analyzers
 - ❖ A rough way to do this is to count the number of analyses being performed in the US, which is not easy. We have a count of the number of notes with US contacts; maybe this is close enough
 - What about MC generation, toyMC (pseudo-exp's for stats interp?), etc?
 - ❖ MC production on Tier-3 is probably small, given ATLAS central prod for MC
 - ❖ Toy MC may not be small, also others like ME calculation, MVA training, etc
 - ❖ Could lead to non-linear scaling w/ lumi. Decided to do linear estimate 1st
 - Analysis model changes from AMMSG need to be considered carefully



Resource Estimate (cont.)



- In Dec 5 meeting, two of these workflows were discussed
 - UC Irvine (Anyes): Detailed breakdown on a SUSY multilepton analysis
 - Analysis workflow shared among 5-6 UCI analyzers who produce different results
 - 1st step is to produce skims from ~200 TB of NTUP_SUSYSKIM input
 - Output of these grid jobs is ~0.8 TB, 60% of which is MC; Output brought to T3
 - 2nd step is to produce a private ntuple for plotting
 - Desire ~2 hr turn-around, so that multiple passes can be made per day, requiring several 100 parallel jobs using 100 CPU-hours to accomplish
 - 3rd and final step is the plotting or limit calculation step, always on the Tier-3
 - Duke (Doug): Overview of Top/SM dilepton analysis (AIDA)
 - This analysis also performs a skim down to private ntuples that are about 6% of the NTUP_COMMON size. This skimming takes 800 CPU-hours on the Grid.
 - Then the derived ntuple analysis on the Tier 3 takes about 6 CPU-hours. Input/output dataset sizes will be provided soon.
 - Kaushik get SUSY jet+MET input; Mark on UIUC analysis flow
-



Resource Estimate (cont.)



- Some very preliminary conclusions based on workflows studied
 - **Space at the Tier-3 is not a limitation** now or in the future. Access, latency, and reliability bigger concerns to users than volume
 - **Computing power is the limiting factor**. At the moment, the analysis framework is tuned so that the turnaround time is at the upper limit of convenience.
 - ❖ With 4x the data, it may not be possible to achieve a reasonable turnaround time on the Tier-3 alone. This implies that either the analysis needs to be more efficient or we need to find computing cycles outside of the Tier-3.
 - One concern expressed about moving some jobs off of the Tier-3 into other facilities: losing a lot in **reliability** and workflow **control** (e.g. checking files for failures, following up on lost files, etc.).
- We need a wider range of analysis workflows before drawing preliminary conclusions about resource estimates



Satisfying Analysis Needs



- Need to show how extra resources (beyond current Tier-3 resources) help satisfy the analysis needs.
- The workflows will point to specific cases, but we should also include the laundry list of enabling technologies. It could include some “best practices” examples of how those technologies can be used to enable analysis
- This list will be used to motivate in part the computing requirements, when the technologies imply a change from current resources



Cost-effectiveness



- The “cost-effectiveness” criterion likely boils down to bang/\$\$.
- The baseline expectation is that this would be the cost of extra hardware, with little or no additional personnel cost
- To 1st order, about how much additional resources should go into L3TCs for Run 2 (to 2nd-order, exactly where and when) vs. using the beyond-pledge resources (BPRs) at the facilities
 - Of course, BPRs are already being used for “US physics”, but more for official MC production and not so much for T3-like workflows
 - Since the BPRs exist right now and could be brought into play with ~zero additional cost, a concern is that all plans that maximize cost-effectiveness might favor BPRs exclusively as the “T3 solution” if they do not take into account some less “tangible” aspects of the L3TCs (most notably, in-kind contributions, convenience, and accessibility)
 - ❖ Needs to be defined and quantified before constructing the set of plans



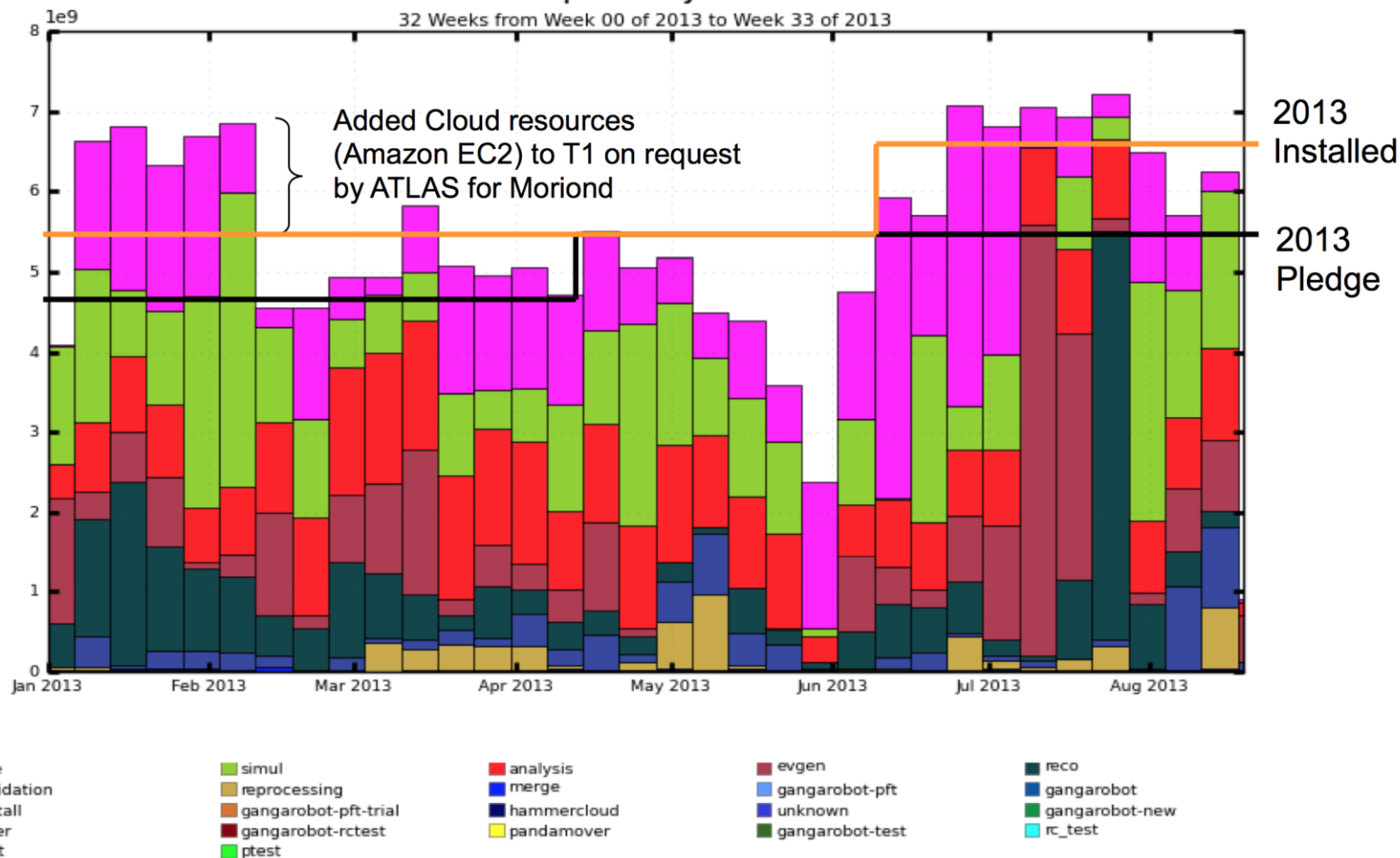
Tier-1 Usage (2013)



dashboard

CPU consumption All Jobs in seconds

32 Weeks from Week 00 of 2013 to Week 33 of 2013





Tier-2 Utilization (Jan-Aug 2013)



dashboard

CPU consumption All Jobs in seconds

32 Weeks from Week 00 of 2013 to Week 33 of 2013

