T3g: functionalities and project status

R. Yoshida (ANL)

Introduction

- Tier3 is your resource to do ATLAS analysis. You're in charge of how you want to set up and use it.
- Still, it's clearly in no one's interest for everyone to go off in random directions. Doug Benjamin and Rik Yoshida have been asked by US ATLAS to coordinate the T3 activities within US ATLAS.
- Core realities:
  - Most T3's have very limited manpower for setup and operation. There will be up to ~30 independent T3's in additions to those associated with T2's and T1.
  - US ATLAS T3 support has very limited manpower (officially 0.5 FTE).
  - A useful ATLAS T3 cluster is a relatively complex system.

The Plan, and where we are on the plan

- The manpower at most US ATLAS T3's are not sufficient to support a full Grid site. Most T3's should be consumers of ATLAS data (delivered from the Grid). It will offer neither cpu cycles nor data to those outside the T3. (This type of T3 is called "T3g" in the US, and referred to as "non-Grid T3" ATLAS wide.)
- The T3's should be set up in alike a manner as possible. This will minimize the effort in several ways.
  - Substantially lessen the initial load on T3 coordination. In fact, the only way to cope with the number of sites that need to come up in a short time.
  - Build a community of T3 admins who will have a shared experience and can begin to support each other.
- As of today we have:
  - A model T3g at ANLASC.
  - Detailed step-by-step instructions on how to build a T3g in being made.
  - The instructions comprise the "phase 1" services that will be sufficient for people to build an immediately usable T3g. This is what we are here to discuss at this meeting.

The Plan, and where we are on the Plan

- The setup of US ATLAS T3's is an integral part of ATLAS planning.
  - In cooperation with Atlas Distributed Computing (ADC) we've convened ATLAS T3 working groups in several topics to arrive at a set of recommendations for ATLAS T3's—with a large amount of US ATLAS input.
  - The setup instructions being presented here conforms to—and benefits from—the international ATLAS T3 recommendations and efforts that went into it.
  - The plan is that the T3g presented here will also be the basis for ATLAS standard for "non-grid" T3's.
- The User Interface at T3:
  - The user interface for T3 analyzers must be easy to use and largely familiar to the majority of those people already working at lxplus, bnl, etc.
  - Should also be as uniform as possible.
  - ATLAS is in the process of recommending AtlasLocalRootBase UI for T3's.
    - Has been in use in Canada for some time.
    - We've tested it in the context of an ANL Analysis Jamboree in March with great success.
    - Is easy to understand and use.
    - Part of our recommendations.

What can you do at a T3g?

- Run Athena jobs interactively on small data samples.
- Submit jobs to Grid using pathena (or prun) and retrieve the output
- Get substantial amount (several TB) ATLAS data to a local storage and keep them.
- Analyze, using athena or root, a large (TB) data sets in a short time (~1day) in an local batch system
- Generate and reconstruct Monte Carlo samples locally.
- Run root jobs interactively for final steps of the analysis.

The T3g design presented here will enable you to do all of these things with
- Budget beginning with several 10's of k$'s.
- Manpower ~ 1 FTE-week for setup and < 0.25 FTE for maintenance. (we hope)

Setting up a cluster

Some preliminary preparation

- Setting up a cluster and administering it is much more involved than, say, installing linux on a desktop. If you can get experienced help at your institute, you should do so.
- A person with a clear responsibilities for the T3 cluster is needed—cannot be a group responsibility.

1. Assign one person from your group (he will need to be an ATLAS member), and a backup, to the T3g setup effort. If at all possible, the same persons should be responsible for T3g administration (atlasadmin) when T3g is operational.
2. Having a backup person will be important. Although the maintenance tasks is envisioned to be light, some of these will have to be done daily or weekly, or it may not be able to wait until the admin returns. Think about rotation of responsibilities after a while.
3. It's recommended that the data management (fetching of large data sets from the grid as well as management of the batch analysis data area) is also handled by the atlasadmin.

People you need to know

You will need to establish contact with following people.

- Department or university will probably have a sys. admin who manages computers in your environment already; bring him/her into the discussion from the beginning. He/she may be able to actively participate in the setup; or take a part of the responsibility for running the cluster. Effort has been made to separate the “root” type tasks from the non-privileged “atlasadmin” tasks to make this easier. In any case he/she needs to stay informed.
- Your University will have a person who is responsible for networking. He/she also needs to be contacted but after you have an initial decision about the size and scope of your cluster. You will need to obtain IP addresses for your cluster as well as to discuss any connectivity issues that might come up with this person.
- Space, Power and Cooling; depending on the size of your installation, you will need to take into consideration space, power and cooling needs for your cluster. Probably your department sys. admin will be able to help you on these issues. Typically there is another set of people to contact about infrastructure; the contact needs to happen after the initial decision about the size of the cluster is made.

Setting up a Cluster (cont)

- Campus computer security officer: there is someone responsible for the local computer security. He/she needs to be contacted early on in the cluster set up process.
- "Nearest" Tier2: Although primary support for T3g will not be through a Tier 2, it is nevertheless necessary to establish contact with a nearby US ATLAS Tier 2. Your T3g, for example, will rely on certain Tier2 provided services such as web proxy caching for certain T3 functionalities.
- You will most likely be purchasing your equipment from Dell. There is a Dell Representative that serves your campus.
- Last but not least, your primary support will be the US ATLAS T3 coordination.

Deciding on your T3g Configuration

- What are the analysis needs of your group ?
  - What are your group members working on? Take a poll of your group.
    - What type of data: ESD, AOD, dESD, dAOD, ntuple, other
    - What type of processing: athena analysis, fast mc, full mc, root, other?
    - how much data do you need locally at any given time? what's the refresh rate on the data you need?
  - Probably the answers are uncertain enough that you will want a setup which is generic enough to handle most things well. However—if you find that you will need to do something specialized (e.g. a lot of full MC generation), you should discuss with the T3 coordinators about the appropriate configuration.
- Is your T3g going to be a part of an existing cluster (e.g. campus research cluster)?
  - No. Proceed with the T3g instruction provided.
  - Only the user accounting will be shared. In this case some isolated parts of the T3g instructions will need to be modified. Otherwise proceed with the T3g instructions provided.
  - Yes. In this case, it's advisable to have an in-depth discussion with the T3 coordinators. Many things will need to be tailored to the existing structure.
- What is your budget ?
  - You now need to match the desires with the reality of the budget. You will probably need at least one iteration to come to the final hardware configuration.
  1. Make a preliminary hardware configuration
  2. Using the preliminary configuration, estimate the associated costs. These include space, power and cooling needs as well as possible network related costs. You will need to discuss this with the relevant contacts.
  3. Make the final configuration taking the associated costs into account.

Deciding on hardware: general remarks

- Guidelines presented here are meant to minimize the setup and maintenance efforts require while still having a good computing performance.
  - Our aim was to try to have 1 FTE-week setup and << 1 FTE maintenance.
  - It's always possible to spend effort instead of money and obtain a more powerful cluster by going outside our recommendations. We think such efforts can give you up to ~10% more CPU power and perhaps up to ~20% more disk space.
  - Difficult to estimate the manpower cost of such an effort. Depends largely on the expertise and the commitment of the person doing it.
  - If you want to do a very custom setup, we'll do our best to support you, but please understand that the support for the recommended setup will have priority.

The computing hardware

- Three class of machines for nominal T3g (most of you)
  - 2 Service nodes:
    - 1 server for: NFS, Data Gateway/buffer, Cluster Monitoring, Cluster Management
    - 1 server for: Batch Management, Data Management, User Management, Web data buffer
  - Interactive nodes (one or more):
    - User login, interactive analysis, submission to local batch and Grid.
    - local user storage area.
  - Batch nodes (one or more—two or more for a meaningful batch system):
    - Parallel batch processing queues.
    - Storage space for data.
- Depending on you needs you might add
  - Storage nodes for data.
- For a very light installation, you can consider an interactive only cluster.
  - Service nodes, in this case will most likely only 1 server (not all services will be needed) or even be a part of an interactive node.

How many Interactive and Batch nodes?

Interactive nodes:

- Processing power:
  - Baseline Dell R710 has 8 cores (hyperthreaded so it “looks” like 16 cores)
- User disk space:
  - Can house up to 8 TB of disks (more normal configuration 4 TB)
- A single R710 node will likely provide resources for up to ~8 relatively intensive interactive users.

Batch nodes:

- “As many as you can afford”
- An R710 will allow one user to run 16 parallel jobs at a time.

If and when additional funding become available, you can add batch (or interactive) nodes as appropriate.

Step by step guide (up to the point where you actually place the hardware order)

1. Decide on the persons responsible.
2. Poll your colleagues and find out their needs/expectations for your T3
3. Decide on what you think will be the hardware configuration.
   - Guidelines from this talk.
   - Detailed recommendations on the Twiki.
4. Discuss the hardware configuration with the T3 coordinators.
5. Contact your local sys. admin and have a first discussion of your intentions.
6. Contact the network person on your campus and have a first discussion.
   - Understand the (in principle) network speed to where your T3g will be. If it is lower than 1 Gbps, find out what it will take to increase it.
   - Understand on which network you will be placing your cluster. Is it on the public internet or is it a campus network of some kind.
   - Understand if you can obtain IP addresses for all of your nodes. You need to understand if you will need to build a private network for your cluster if it's not possible to get each node an IP address. Remember that you may expand your cluster at a later date.

Step by step guide (up to the point where you actually place the hardware order)

7. Understand the infrastructure needs. Discuss with your sys admin and infrastructure people.
   - Space for the equipment
   - Power, cooling needs (information on the Twiki)
   - Noise control

   Remember to take the possibility of expansion in mind.
8. Evaluate the non computing equipment expenses.
9. Re-do the hardware configuration in light of what you have learned.
   - If very much different from the initial config.. need to iterate.
10. Re discuss with the T3 coordinators.
11. Contact your local cyber security person and discuss your plans. OSG security team can help you in this discussion.
12. If all went well, place the order.