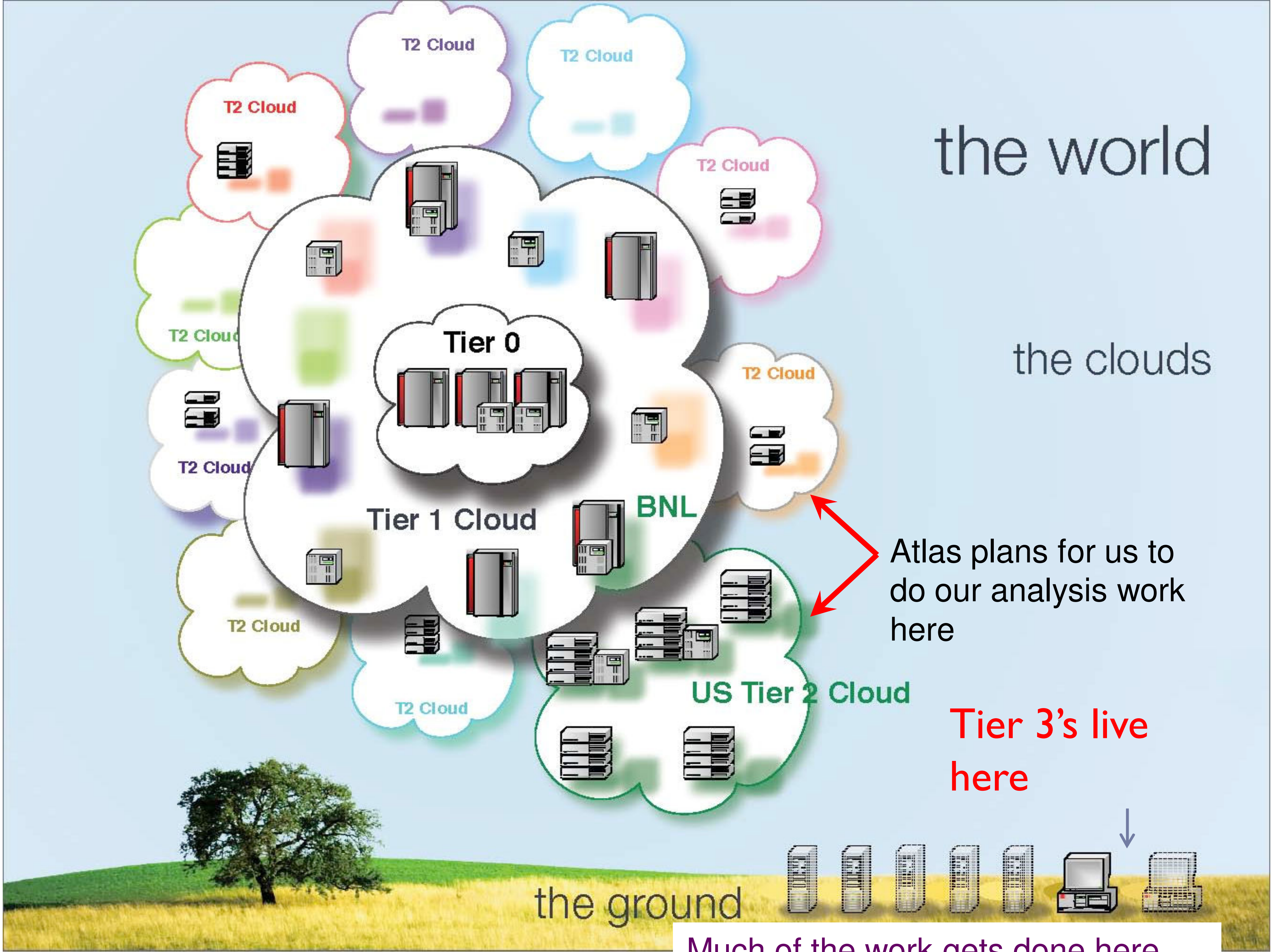
Tier 3 Computing

Doug Benjamin
Duke University



the world

the clouds

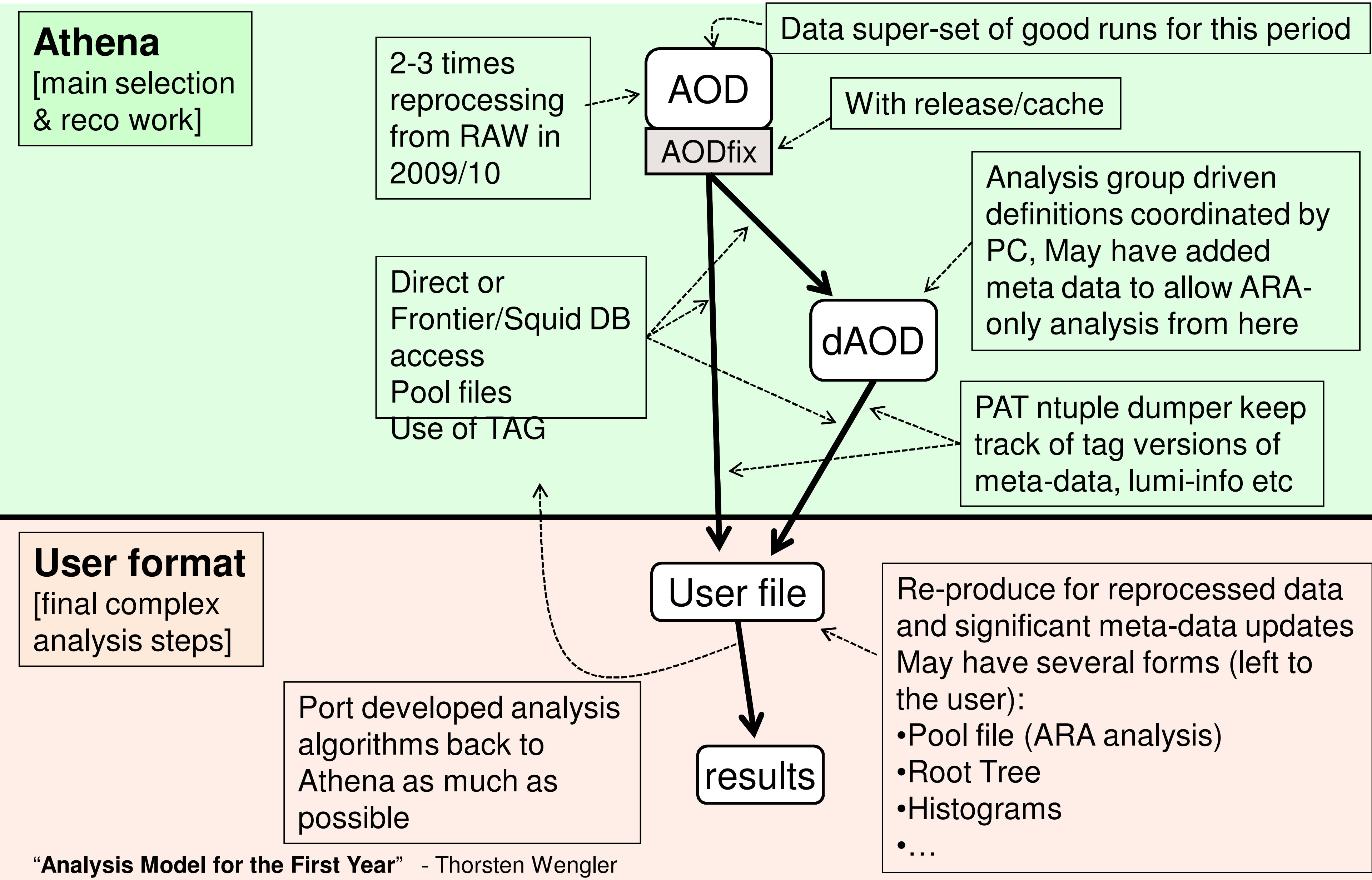
the ground

Atlas plans for us to do our analysis work here

Tier 3's live here

Much of the work gets done here

Skeleton physics analysis model 09/10



Athena
[main selection & reco work]

Data super-set of good runs for this period

2-3 times reprocessing from RAW in 2009/10

AOD

AODfix

With release/cache

Analysis group driven definitions coordinated by PC, May have added meta data to allow ARA-only analysis from here

Direct or Frontier/Squid DB access
Pool files
Use of TAG

dAOD

PAT ntuple dumper keep track of tag versions of meta-data, lumi-info etc

User format
[final complex analysis steps]

User file

Re-produce for reprocessed data and significant meta-data updates
May have several forms (left to the user):

- Pool file (ARA analysis)
- Root Tree
- Histograms
- …

Port developed analysis algorithms back to Athena as much as possible

results

“**Analysis Model for the First Year**” - Thorsten Wengler

Types of Tier 3's

- Tier 3 gs (grid services)
  - Same services as Tier 2 - Can accept Panda jobs
- Tier 3 w (workstation)
  - Interactive workstation with Atlas Software
- Tier 3 af (Analysis Facility)
  - Located at National Lab, University groups can add CPU or purchase storage, Useable by all US Atlas members with Fair share – Like CDF CAF
- Tier 3 g (grid aware) (most common type)
  - Local batch system, Grid aware storage, local storage, Atlas software, interactive nodes, can send jobs to grid

Some goals of this talk:

- All US ATLAS Institutes thinking about what to do next about the computing resources under their control (Tier 3 resources) to maximize their usefulness.
- You may:
  - already have a working ATLAS Tier 3. Thinking about expansion.
  - have some computing infrastructure but not set up for ATLAS analysis.
  - have no Tier 3, but applied for and/or received funding for one.
  - be thinking about investing in an Analysis Facility at BNL or SLAC.
- A year ago, there was little planning or organization in the use of T3 resources.
  - DB (plus some volunteers) has began to set the direction for the T3 resources since last year. A lot of work has already been done.
  - Rik Yoshida has officially jointed the effort (he was working on Tier 3 issues for some time already).
  - He and I are working closely together. Our aim is to organize the T3 efforts for the maximum benefit to all US ATLAS institutes.

T3 roadmap

- Now:
  - Prerequisites
    - Understand how people are likely to do analysis
    - Keep in mind technical parameters of the US ATLAS facilities
  - Survey the T3 technical solutions already available. (Already done to a large extent)
- Very soon (next month or so)
  - Build up a (set of) recommended configurations and instructions for setting up T3(g) (already underway).
    - Must be easy to setup and maintain (<<1 FTE)
    - Must allow for evolution. (Not all desirable features will be initially available)
    - Consider the setup of existing T3s
  - Build up a support structure for T3s.
    - There will likely be only a small core (~1 FTE) of explicit support people.
    - A T3 community which is self-supporting must be built up. We will need as much standardization as we can get.
  - Start building (or extending) T3s:
    - This is primarily to be done by each institute from the T3 instructions.
    - Probably start with one or two "guinea pigs".
  - Define the Analysis Facilities;
    - What the costs are.
    - What you will get.
  - Start a program of T3 improvements (some effort already beginning).
    - Ease of deployment and maintenance (e.g.VM)
    - Addition of desirable features (e.g. data management)

Slide borrowed from Rik Yoshida – Tier 2/Tier 3 meeting talk

Your T3 resources

- Each institute will have to decide how to allocate their T3 resources.
- The basic choices:
  - Analysis Facilities: you will be able to contribute to AFs in exchange for a guaranteed access to processing power and disk.
  - T3g: if starting from scratch, you could build a pretty powerful system starting from several 10's of k$. Will need ~1 FTE-week to build but maintenance should be << 1 FTE.
  - T3gs: this is basically a miniature T2: will need sizable funding and manpower commitment. Maintenance will require 0.5-1.0 (expert) FTE.
  - Of course you might choose to have both a stake in AF and a T3g(s).
- Not easy to decide what is optimal.
- As you know, we currently only have the rough outlines of plans in most areas. Given the many unknowns and diverse situations of the institutes, it's not possible, nor desirable to formulate specific plans without close consultation with all institutes.
- So, Rik and I have started to contact all US Atlas institutions (not via e-mail, but either in person or on the phone) to discuss each groups particular situation.
  - Met with 15 institutions last week, will meet with 12 more next week (27 out of 42 institutions)
  - We will contact the rest of people (we will call and setup an appointment)
- Designate a contact person who will have given some thought to the following..

Slide borrowed from Rik Yoshida – Tier 2/Tier 3 meeting talk

The needs of your institutes

- Do you know how the people in your group will use to do analysis?
- Some sample questions.
  - Where do you plan to do your interactive computing?
    - □ Athena code development before Grid submission.
    - □ Root sessions to run on the output of your athena jobs.
    - Are you counting on lxplus or acas? Will you need your own resources such as a local T3 or a share in an Analysis Facility (T3AF)?
    - **Did you know that BNL will be reducing the number of general slots by 80%?**
  - You will use the Grid to do main athena processing.
    - How stretched will the T2 (and T1) analysis queues be?
    - If they are oversubscribed, where will you do "medium sized" jobs?
      - □ Analysis Facilities? Buy share?
      - □ Local T3? Build one that's usable for TB sized processing.
  - Do you have atypical needs for your T3?
    - Access to raw data and conditions DB?
    - Test MC generation?
  - If you have a T3 or a cluster already:
    - What are your limitations? Memory/core? Networking?
    - Have you actually tried to run ATLAS applications at realistic scale on your setup?
- Many of these questions are unanswerable—but considering questions like this will help you in deciding what to do next.

Slide borrowed from Rik Yoshida – Tier 2/Tier 3 meeting talk

Tier 3gs and Tier 3w

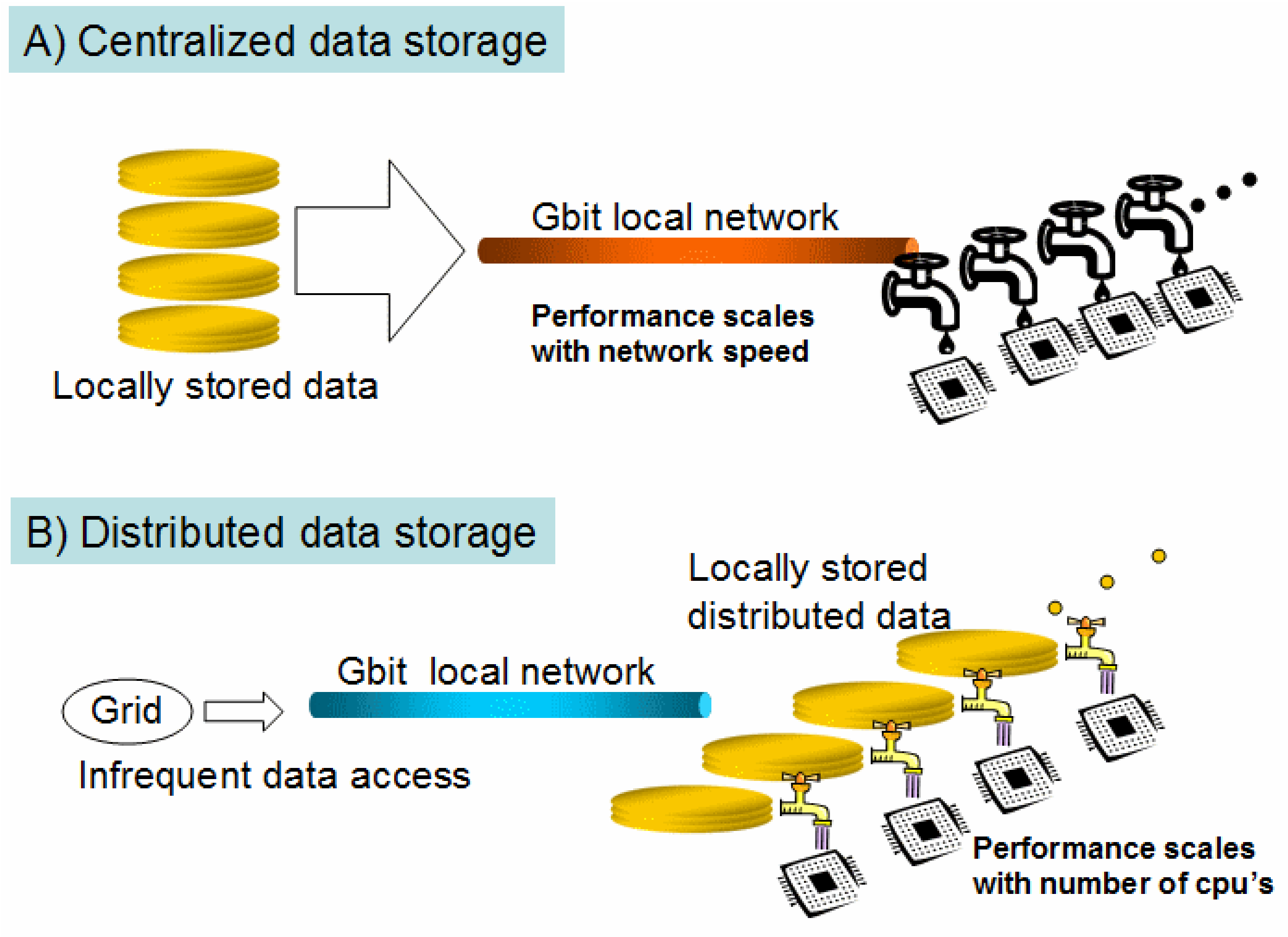
- Tier 3 gs (grid services)
  - Same services as Tier 2 - Tier 2 "Lite"
  - Requires significant labor to keep production quality (at least 0.5 FTE - talented system admin).
  - Only a small fraction (~1%) Panda jobs run at Tier 3gs
- Tier 3 w (workstation)
  - Interactive workstation with Atlas Software
    - Atlas code on machine or served from local NFS fileserver
  - No batch system
  - Can submit Pathena or Prun grid jobs
  - All Atlas data retrieved using client tools (dq2-get)

Tier 3 G
(most common Tier 3)

- Interactive nodes
- Can submit grid jobs.
- Batch system w/ worker nodes (Condor)
- Atlas Code available (at least kit releases)
- Currently client tools used for fetch data (dq2-get)

  DDM being modified to automatically deliver data to Tier 3
- Storage can be one of two types (sites can have both)
  - Located on the worker nodes
    - Bare disks on workers – ANL PC Farm
    - XROOTD
  - Located in dedicated file servers (NFS/ XROOTD)
- Grid Storage Element Bestman –gateway with gridftp

Why choose distributed data storage?



A) Centralized data storage

Locally stored data

Gbit local network

Performance scales with network speed

B) Distributed data storage

Grid

Infrequent data access

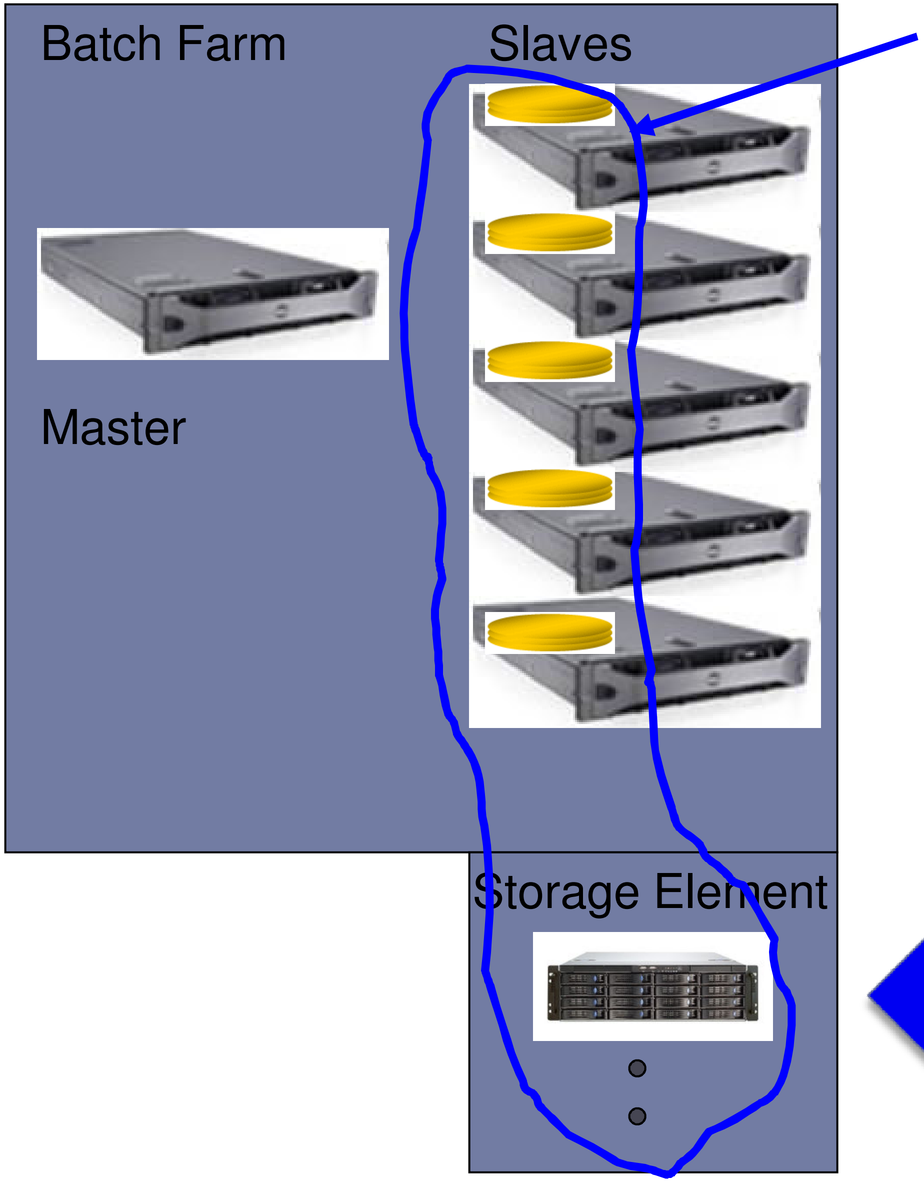
Gbit local network

Locally stored distributed data

Performance scales with number of cpu's

Since most Tier 3's will not have 10 Gbe between storage and worker nodes - distributed data storage makes sense

Basic Configuration of a T3g



XrootD forms a single file system for the disks in Slaves and

1) Uses the SE as a “mass storage” from which data sets are copied to the slave disks.
2) Distributes the files in a data set evenly among the slave disks.
3) Keeps track of files-disk correlation to allow the Arcond program to submit the batch jobs to the nodes with local files.

XRootD can run on discrete file servers or in a distributed data configuration

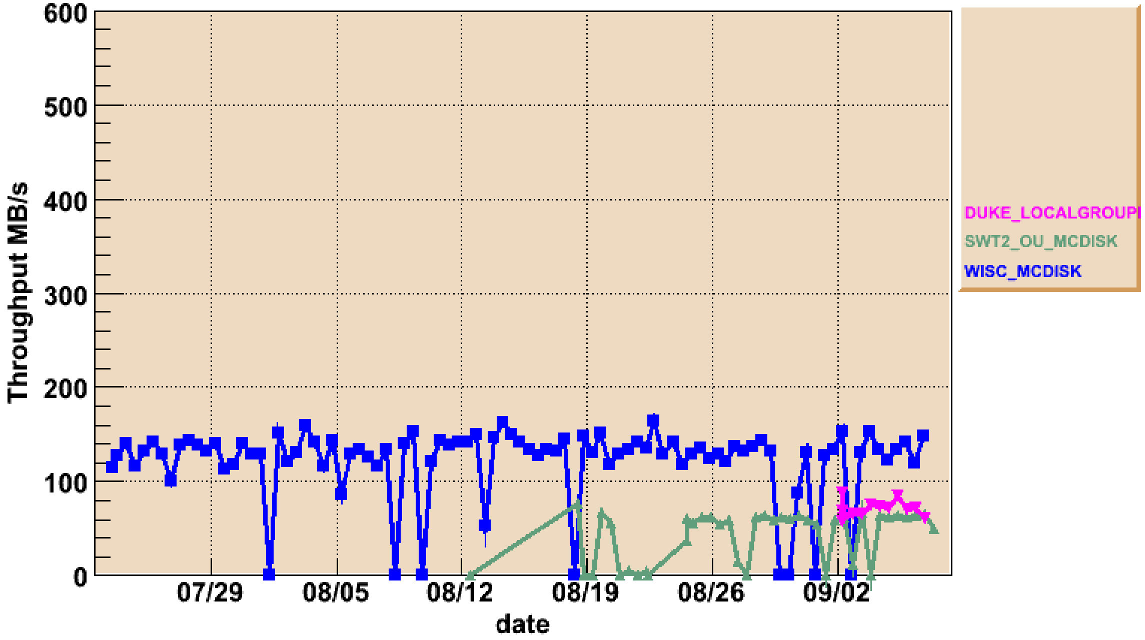
Data from grid goes here

Tier 3 issues

- Monitoring of Data storage
  - Tier 3's will have finite amount of storage - Storage will be like a cache – Need to monitor data usage patterns to determine Data longevity at site. (clean up old data)
  - Storage system performance monitoring
  - Work in progress (XRootD, Proof PQ2 tools can help)
- Data Access (how get the data to your site)
  - ***Atlas agreed last week to change DQ2 to allow for data subscriptions.***
- Database Access at Tier 3's
  - ***Atlas agreed to support SQUID/Frontier to provide excellent Conditions DB access to Tier 3's***

Tier 3 issues - Networking

- Networking
  - We often take networking for granted – yet it needs to optimized for efficient transfers. - implies interaction with Campus Network admin.
  - Internet 2 has agreed to help us (Thanks!)

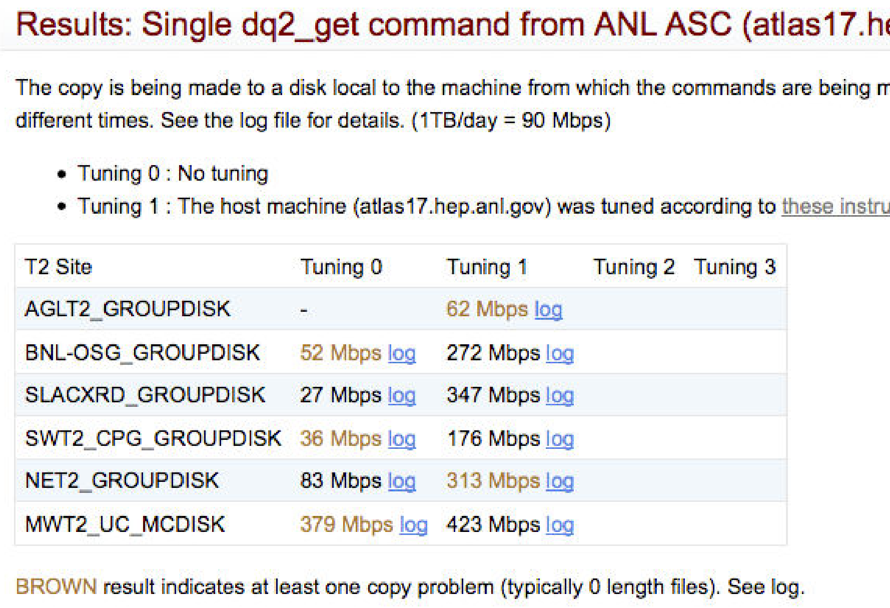


Automated Throughput test to Tier 3 site (Disk to Disk test)

Will add more sites (ANL this week)

Tier 3 Networking – Tuning/dq2 client

- Doug Benjamin and Rik Yoshida tested copy rates to Duke and using dq2 client from various sites with US Cloud
  - https://atlaswww.hep.anl.gov/twiki/bin/view/ASC/Dq2_getStressTest



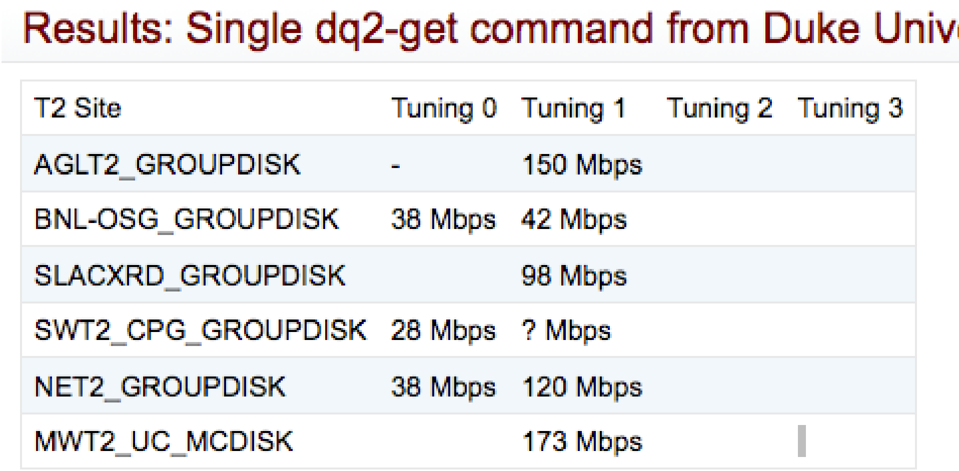
Results: Single dq2_get command from ANL ASC (atlas17.hep.anl.gov)

The copy is being made to a disk local to the machine from which the commands are being made. The commands are issued at different days of the week and at different times. See the log file for details. (1TB/day = 90 Mbps)

- Tuning 0 : No tuning
- Tuning 1 : The host machine (atlas17.hep.anl.gov) was tuned according to these instructions.

T2 Site	Tuning 0	Tuning 1	Tuning 2	Tuning 3
AGLT2_GROUPDISK	-	62 Mbps log		
BNL-OSG_GROUPDISK	52 Mbps log	272 Mbps log		
SLACXRD_GROUPDISK	27 Mbps log	347 Mbps log		
SWT2_CPG_GROUPDISK	36 Mbps log	176 Mbps log		
NET2_GROUPDISK	83 Mbps log	313 Mbps log		
MWT2_UC_MCDISK	379 Mbps log	423 Mbps log		

BROWN result indicates at least one copy problem (typically 0 length files). See log.



Results: Single dq2-get command from Duke Univ

T2 Site	Tuning 0	Tuning 1	Tuning 2	Tuning 3
AGLT2_GROUPDISK	-	150 Mbps		
BNL-OSG_GROUPDISK	38 Mbps	42 Mbps		
SLACXRD_GROUPDISK		98 Mbps		
SWT2_CPG_GROUPDISK	28 Mbps	? Mbps		
NET2_GROUPDISK	38 Mbps	120 Mbps		
MWT2_UC_MCDISK		173 Mbps		

Tier 3 issues - Support

- Support

  (goal – maintain a Teir 3g w/ < 0.25 FTE Postdoc/grad stud.)

  (less than 1 week Full time to setup a Tier 3)

  - Likely less than ~ 1 FTE assigned to support Tier 3's.
  - OSG would like to help
  - Atlas Canada - they support 1 configuration – have good tool kit.
  - Standardization should help here.
  - Since there will be only a small amount of formal support for T3s for the foreseeable future, the T3 community must become self-sustaining.

Tier 3g configuration instructions and getting help

- Tier 3g configuration details and instructions in Tier 3 wiki at ANL and BNL:
  - https://atlaswww.hep.anl.gov/twiki/bin/view/Tier3Setup/WebHome
  - https://www.usatlas.bnl.gov/twiki/bin/view/Admins/Tier3Setup
- US Atlas Hypernews - **HN-Tier3Support@bnl.gov**
- US Atlas Tier 3 trouble ticket at BNL USAtlasTier3

  RT-RACF-USAtlasTier3@bnl.gov
- If all else fails contact us:
  - Doug Benjamin - US Atlas Tier 3 technical support lead (benjamin@phy.duke.edu)
  - Rik Yoshida – US Atlas Tier 3 coordinator (Rik.Yoshida@anl.gov)

Future Tier 3 improvements

- Tier 3 Virtualization "Tier 3 in Box"
  - BNL has provided a test cluster for Virtualization studies and development (Xen based)
  - Using CERNVM and Xen build VM for worker node
  - Some Tier 3's already using VM's (OSG, Duke).
  - reduce support load globally and locally
  - Provide better security for Tier 3's

- Will look at other technologies to make Tier 3 as effective as possible.

Conclusions

- Data is coming. So is some money --- We need to start setting up our Tier 3's sooner than later.

- We should configure the Tier 3's in a manner to be the most effective.

- Tier 3's are a collaborative effort. We will need your help.

- Since the Atlas Analysis model is evolving – Tier 3's must be adaptable and nibble