Tier 3(g) Cluster Design and Recommendations

Doug Benjamin
Duke University

Tier 3g design/Philosophy

- Design a system to be flexible and simple to setup (1 person < 1 week)
- Simple to operate - < 0.25 FTE to maintain
- Scalable with Data volumes
- Fast - Process 1 TB of data over night
- Relatively inexpensive
  - Run only the needed services/process
  - Devote most resources to CPU's and Disk
- Using common tools will make it easier for all of us
  - Easier to develop a self supporting community.

How do you want to use your Tier 3?

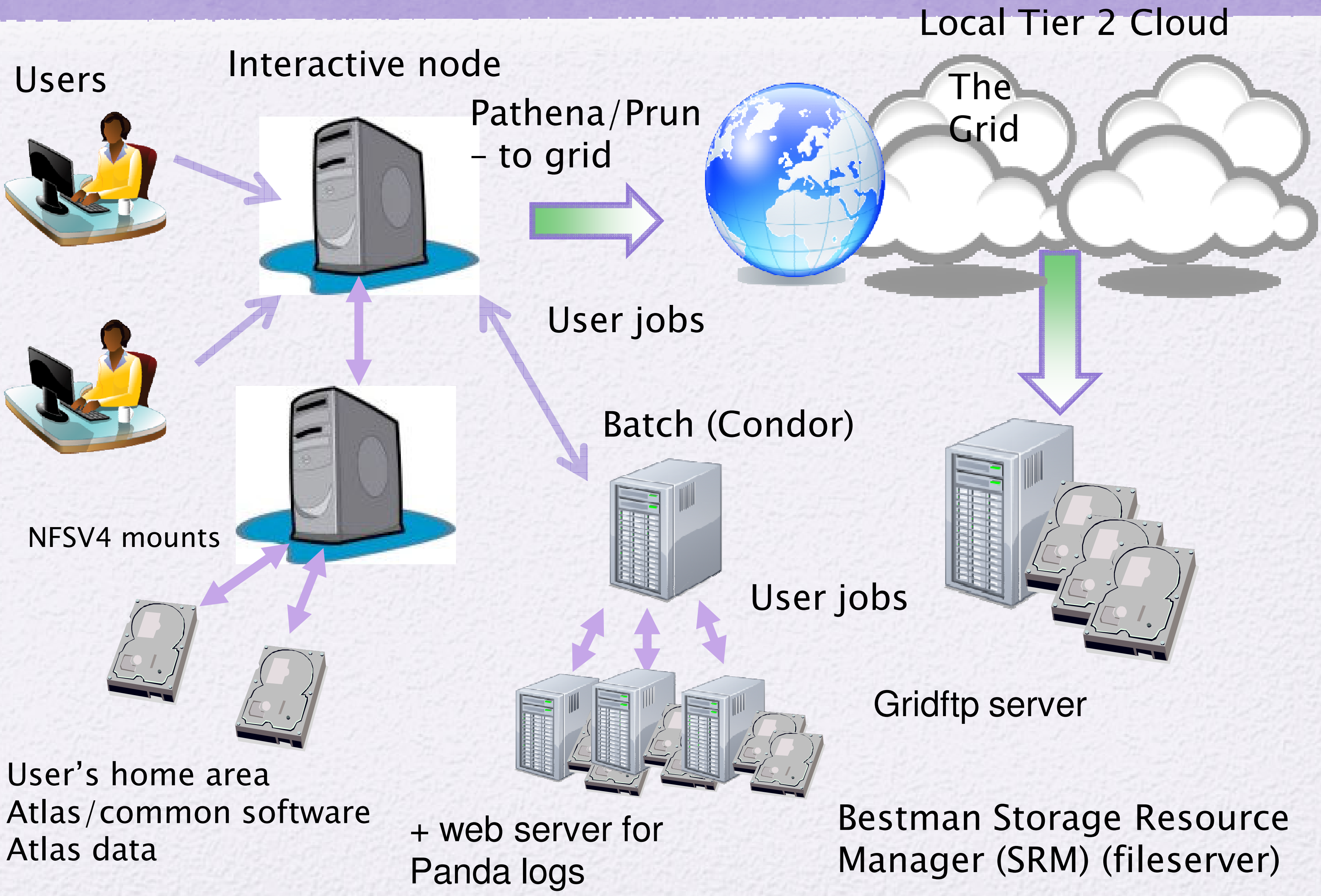
- Do you want to generate events?
  - Implies more CPU power than disk space
- Do you want to do AOD analysis or make within Athena? -> implies many cores
- Do you want analyze mostly ntuples at your Tier 3?

How much Disk do you need?

- Mostly ntuple analysis at Tier 3g

  Example (1 fb^{-1}) calculation from Ayana Arce – Duke Univ
- [1.2 TB] To store one stream of single lepton signal data (real data, no truth) AODs in the absence of QCD fakes, we would need about 1.2 GB per inverse pb. This estimate comes from the most recent topmix sample, which has unweighted Ws, Zs ,dibosons, and top. The trigger efficiency is not taken into account (there are real leptons so this is a factor of order 1).
- [5 TB] Multiply by a factor of 2ish (optimistic) for QCD fakes (in the trigger). Multiply by ~2 to keep both lepton flavors.
- [2.5 TB] The semi-official top dAODs made from the non-fake samples seem to be 1/2 as large. (Note that these dAODs *add* EDM objects (more jet collections, the top EDM) but skim on reco-level leptons.
- [1.2 TB] Divide by ~2 again if we strip off more collections? Note, this requires running our own jobs on the grid to get AOD-equivalents
- [0.3 TB] Divide by ~4 if we keep absurdly detailed ROOT files instead
- **[1 TB]** Multiply output format by X for fully simulated signal MC samples. Neglect truth-level samples. Still, X=3 is probably optimistic.
- (Assumes ntuple analysis at Tier 3 and efficient use of GRID)

Tier 3g configuration



Local Tier 2 Cloud

The Grid

Users

Interactive node

Pathena/Prun – to grid

User jobs

Batch (Condor)

NFSV4 mounts

User jobs

Gridftp server

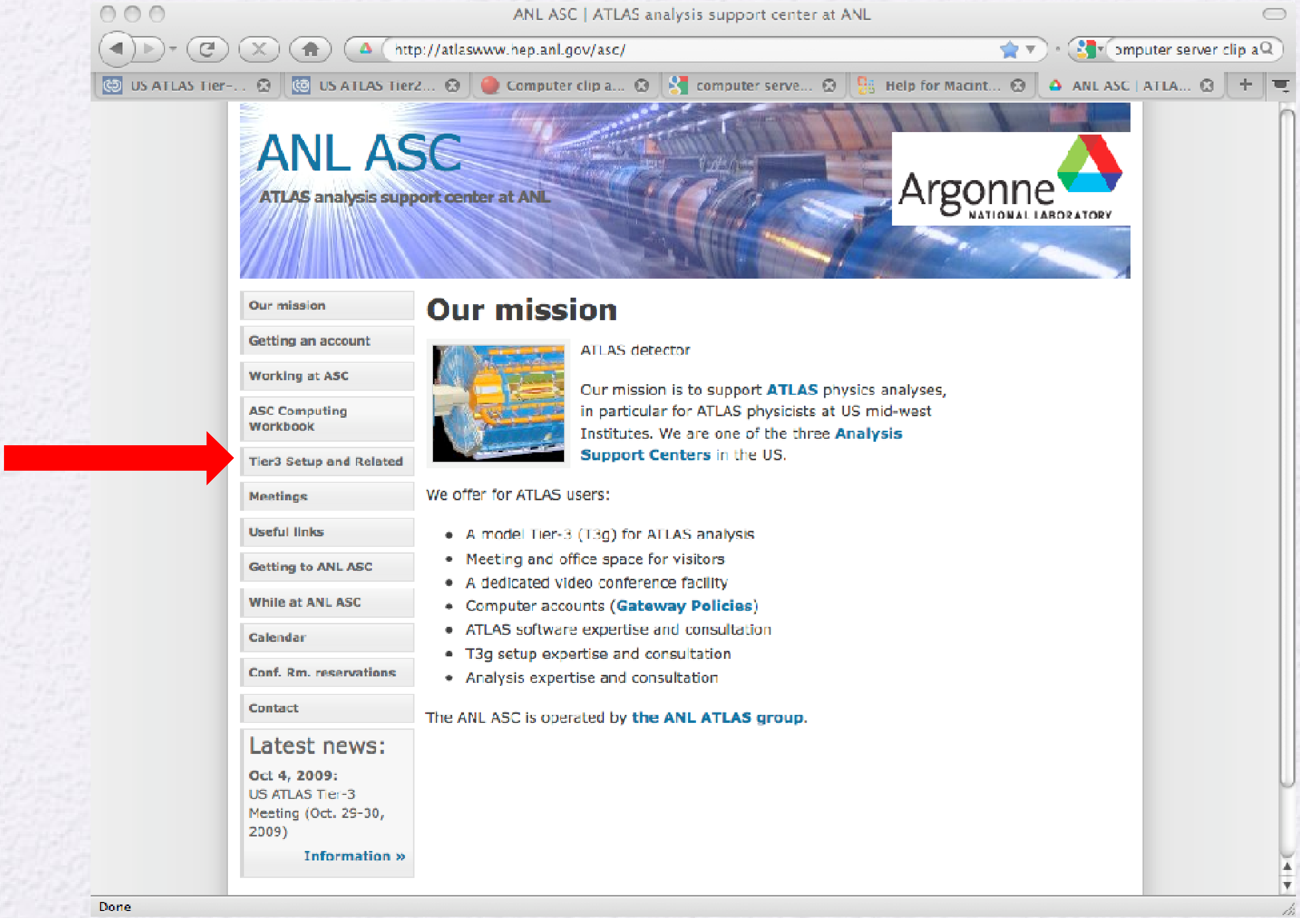
User's home area
Atlas/common software
Atlas data

+ web server for
Panda logs

Bestman Storage Resource
Manager (SRM) (fileserver)

Where to find details

- Tier 3 configuration wiki currently at ANL
  https://atlaswww.hep.anl.gov/twiki/bin/view/UsAtlasTier3/Tier3gSetupGuide



Our mission

ATLAS detector

Our mission is to support **ATLAS** physics analyses, in particular for ATLAS physicists at US mid-west Institutes. We are one of the three **Analysis Support Centers** in the US.

We offer for ATLAS users:

- A model Tier-3 (T3g) for ATLAS analysis
- Meeting and office space for visitors
- A dedicated video conference facility
- Computer accounts (**Gateway Policies**)
- ATLAS software expertise and consultation
- T3g setup expertise and consultation
- Analysis expertise and consultation

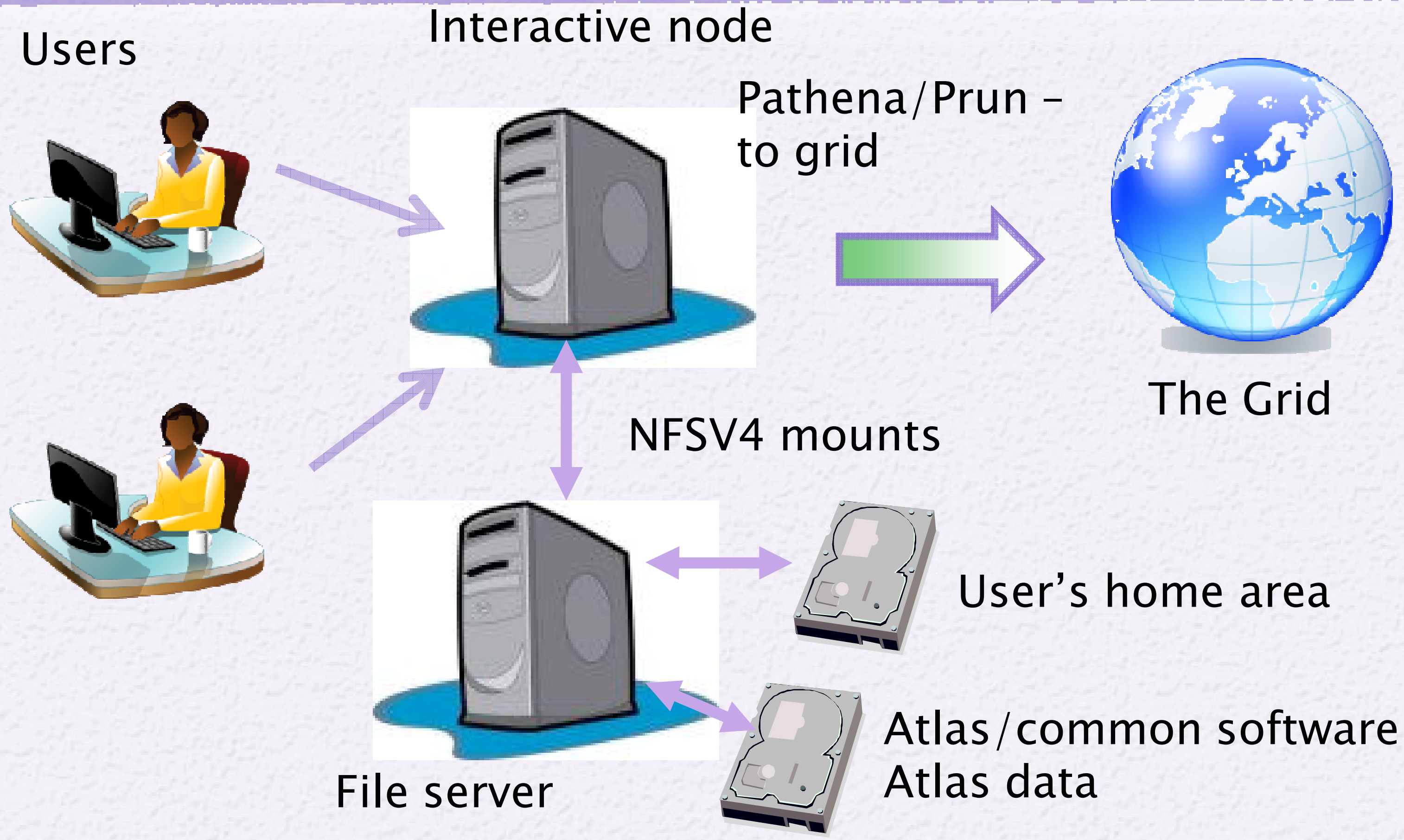
The ANL ASC is operated by **the ANL ATLAS group**.

Latest news:

Oct 4, 2009:
US ATLAS Tier-3 Meeting (Oct. 29-30, 2009)

Information »

Tier 3g – Interactive computing



Users

Interactive node

Pathena/Prun –
to grid

The Grid

NFSV4 mounts

User’s home area

Atlas/common software
Atlas data

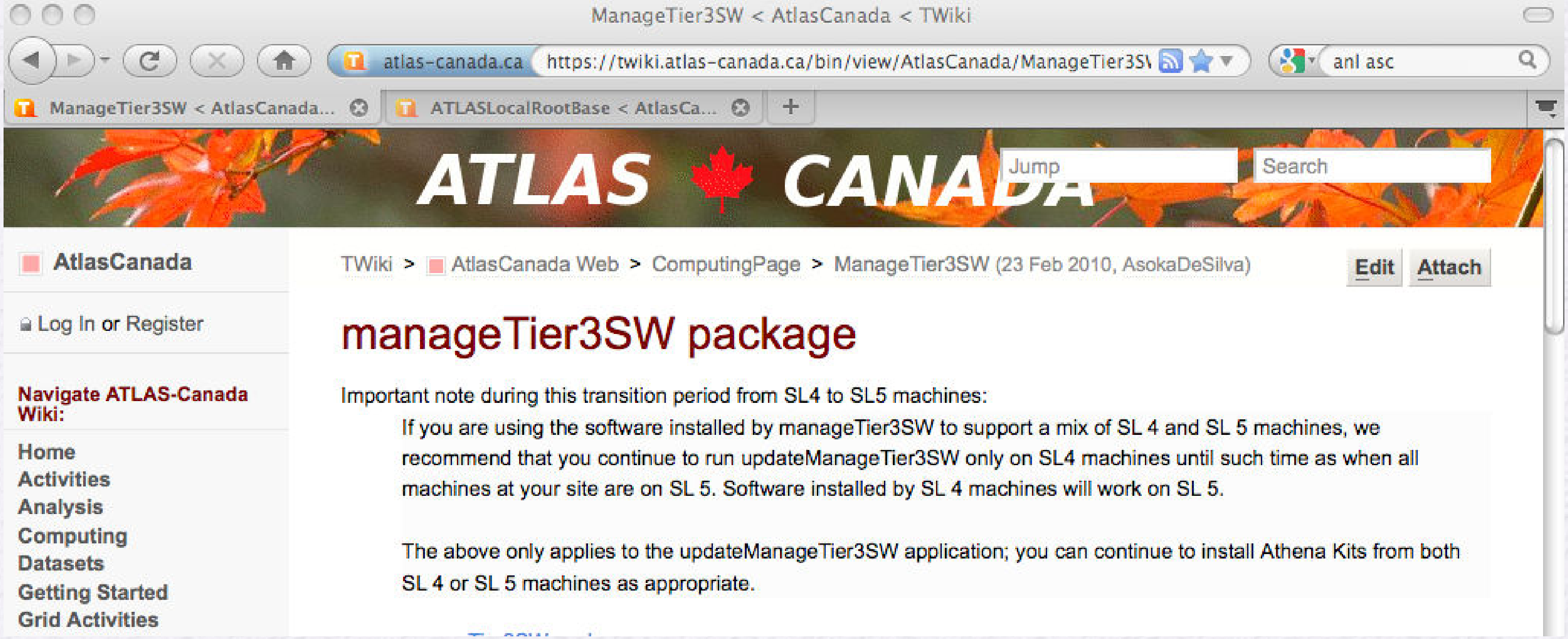
File server

Common User environment (next slide)
Atlas software installed (two methods)
  manageTier3SW
Web file system CVMFS

Atlas Code installation

- NFS file server
  - ManageTier3 SW package (Asoka DeSilva Triumf)

https://twiki.atlas-canada.ca/bin/view/AtlasCanada/ManageTier3SW


ManageTier3SW < AtlasCanada < TWiki
atlas-canada.ca
https://twiki.atlas-canada.ca/bin/view/AtlasCanada/ManageTier3S\
anl asc
ManageTier3SW < AtlasCanada...
ATLASLocalRootBase < AtlasCa...
ATLAS CANADA
Jump
Search
AtlasCanada
Log In or Register
Navigate ATLAS-Canada Wiki:
Home
Activities
Analysis
Computing
Datasets
Getting Started
Grid Activities
TWiki > AtlasCanada Web > ComputingPage > ManageTier3SW (23 Feb 2010, AsokaDeSilva)
Edit
Attach

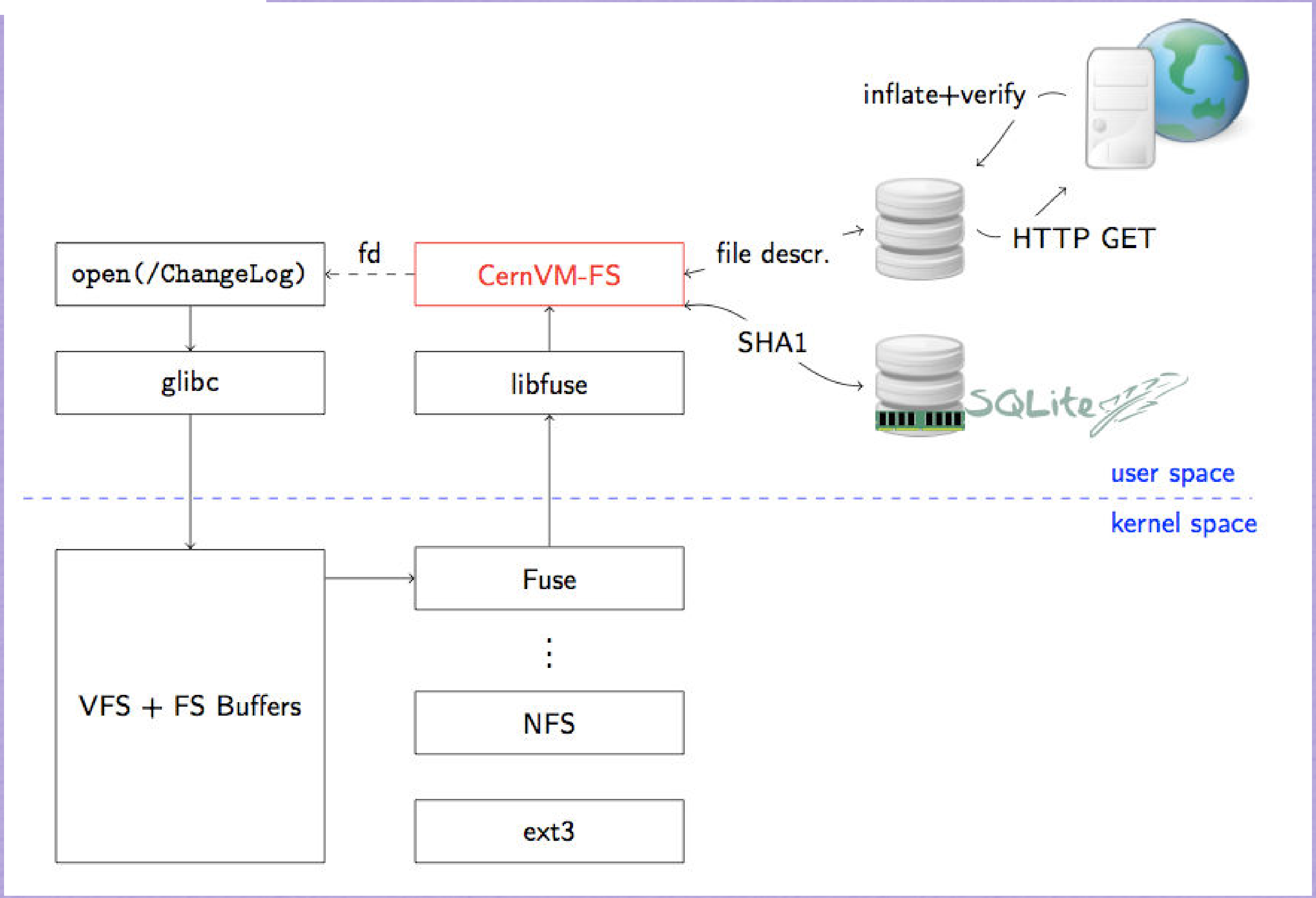
manageTier3SW package

Important note during this transition period from SL4 to SL5 machines:

> If you are using the software installed by manageTier3SW to support a mix of SL 4 and SL 5 machines, we recommend that you continue to run updateManageTier3SW only on SL4 machines until such time as when all machines at your site are on SL 5. Software installed by SL 4 machines will work on SL 5.
>
> The above only applies to the updateManageTier3SW application; you can continue to install Athena Kits from both SL 4 or SL 5 machines as appropriate.

Well tested straight forward to use

CVMFS 0.2 (v2)


inflate+verify
HTTP GET
open(/ChangeLog)
fd
CernVM-FS
file descr.
SHA1
SQLite
glibc
libfuse
user space
kernel space
VFS + FS Buffers
Fuse
⋮
NFS
ext3

NFS V4 vs CVMFS Comparison Athena Compilations

Rik Yoshida (ANL)

Dell R710: 8 cores (16 hyperthreaded)

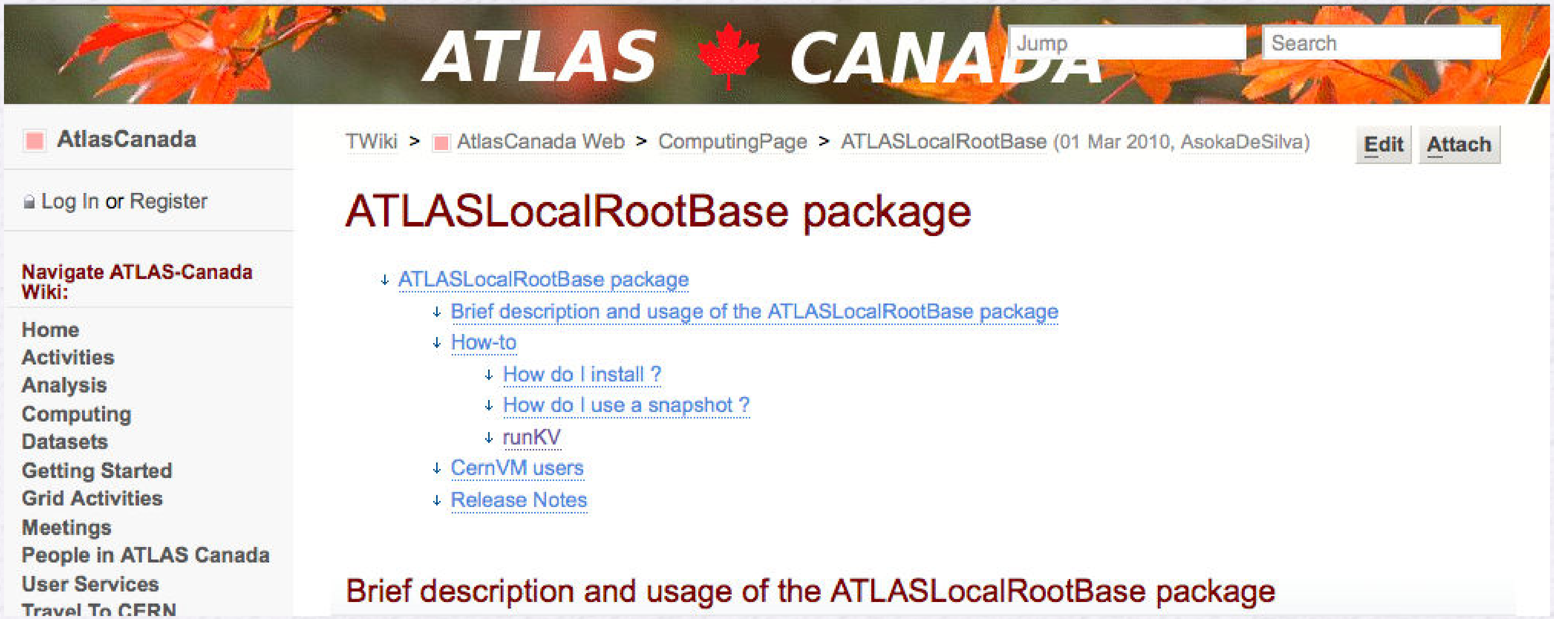
No. Simultaneous Condor jobs:	1	4	8	14
NFS4	7 min	15 min	60 min	
CVMFS2	7 min		8 min	11 min

Tier 3 User environment

ATLASLocalRootBase

https://twiki.atlas-canada.ca/bin/view/AtlasCanada/ATLASLocalRootBase

- Can easily setup a tested suite of software needed for work in a Tier 3



ATLASLocalRootBase package

Brief description and usage of the ATLASLocalRootBase package

Developed by Asoka DeSilva

Tier 3g User environment

- ANL ASC cluster configured with this User environment.
- User guide contains information
  https://atlaswww.hep.anl.gov/twiki/bin/view/UsAtlasTier3/Tier3gUsersGuide

Add to your .bashrc file

```
export ATLAS_LOCAL_ROOT_BASE=/export/share/atlas/ATLASLocalRootBase
alias setupATLAS='source ${ATLAS_LOCAL_ROOT_BASE}/user/atlasLocalSetup.sh'
```

- Interactive use (to setup an athena version)

```
setupATLAS     # Some info output after this command
localSetupGcc --gccVersion=gcc432_x86_64_slc5  # Sets the compiler version
export ATLAS_TEST_AREA=<some area>/15.6.6  # defines your test area (note vers. #)
source /export/home/atlasadmin/temp/setupScripts/setupAtlasProduction_15.6.6.sh
```

- To see what other software is available

```
    showVersions
```

Tier 3g User environment(2)

- User environment inside a shell script.

```
export ATLAS_LOCAL_ROOT_BASE=/export/share/atlas/ATLASLocalRootBase
source ${ATLAS_LOCAL_ROOT_BASE}/user/atlasLocalSetup.sh
source ${ATLAS_LOCAL_ROOT_BASE}/packageSetups/atlasLocalGccSetup.sh --gccVersion=gcc432_x86_64_slc5
export ATLAS_TEST_AREA=<some area>/15.6.6
source /export/home/atlasadmin/temp/setupScripts/setupAtlasProduction_15.6.6.sh
```

- Root inside a shell script

```
export ATLAS_LOCAL_ROOT_BASE=/export/share/atlas/ATLASLocalRootBase
source ${ATLAS_LOCAL_ROOT_BASE}/user/atlasLocalSetup.sh
source ${ATLAS_LOCAL_ROOT_BASE}/packageSetups/atlasLocalGccSetup.sh --gccVersion=gcc432_x86_64_slc5
source ${ATLAS_LOCAL_ROOT_BASE}/packageSetups/atlasLocalPythonSetup.sh --pythonVersion=2.5.2
source ${ATLAS_LOCAL_ROOT_BASE}/packageSetups/atlasLocalROOTSetup.sh –rootVersion=5.26.00-slc5-gcc4.3
```

How data comes to Tier 3g’s

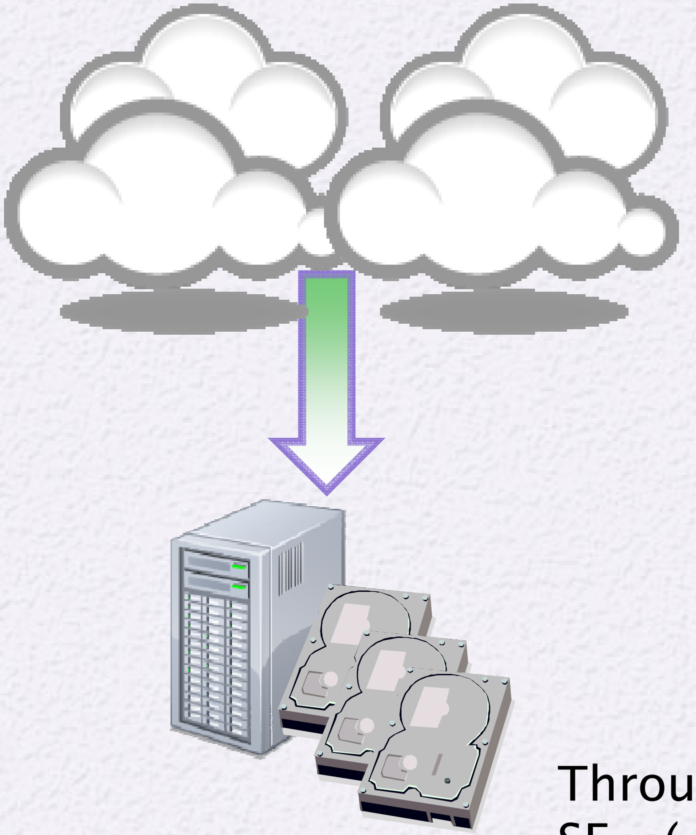
Two methods
• Enhanced dq2-get
(uses fts channel)
(available fairly soon)

- Data subscription
  - SRM/gridftp server part of DDM Tiers of Atlas

Bestman Storage Resource Manager (SRM) (fileserver)

- Sites in DDM ToA will tested frequently
- Troublesome sites will be blacklisted (no data) extra support load

US Tier 2 Cloud



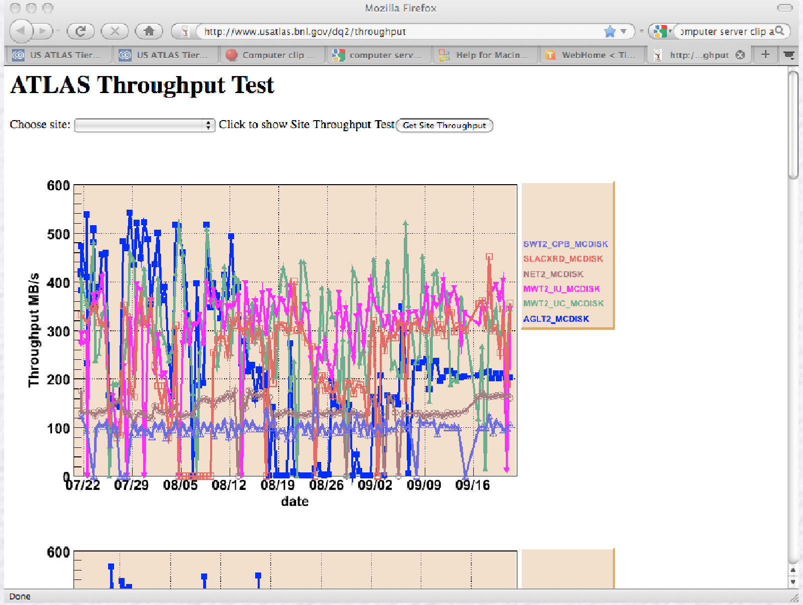
Data will come from **any** Tier 2 site

Throughput test with ANL SE – (> 500 Mb/s)

Shows $1200 PC (Intel i7 chip/ X58 chipset/ SL5.3) can be a SE for a small T3.

Storage Element installation/testing

- Instructions for Bestman-Gateway SRM https://atlaswww.hep.anl.gov/twiki/bin/view/Tier3Setup/SetupSE
- Gridftp only instructions coming. (Can use existing instructions)
- Through put testing and instructions http://www.usatlas.bnl.gov/dq2/throughput (testing graphs) https://atlaswww.hep.anl.gov/twiki/bin/view/Tier3Setup/ThroughputCleanup



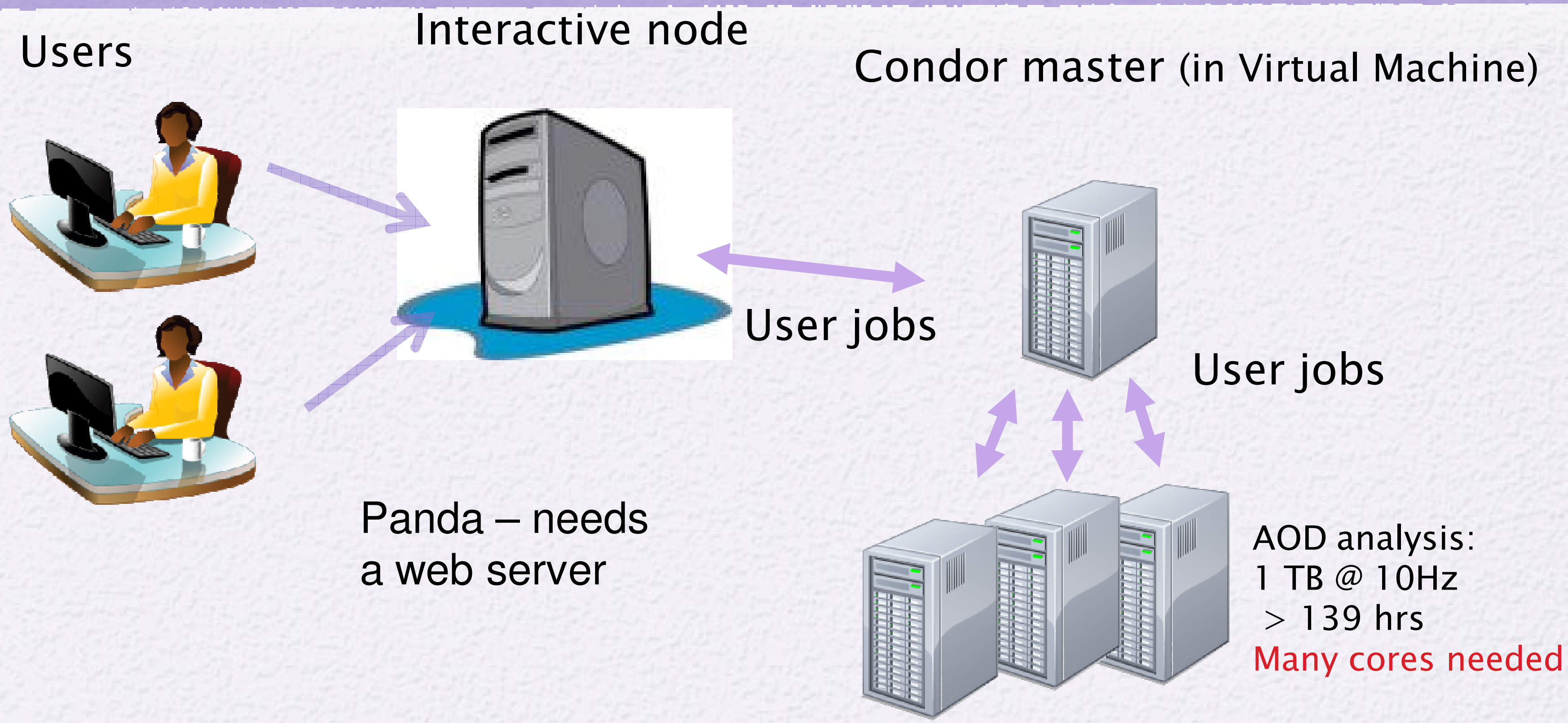
ATLAS Throughput Test

Choose site: Click to show Site Throughput Test

Implications of ToA

- Your site will be both a data sink and data source
- You will need to have Atlas DDM tests run at your site on a fixed cycle (appropriate for T3's)
- File remove implies remove in database before remove at Site or Dark data
- You can be black listed if you fail too many DDM tests. -> No data
- Must provide good quality of service

Tier 3g – Batch/ Distributed computing



Users

Interactive node

Condor master (in Virtual Machine)

User jobs

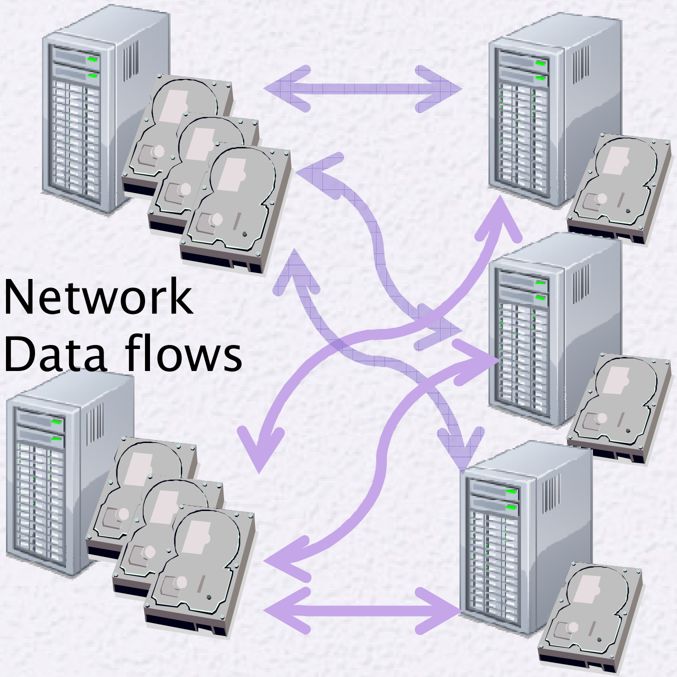
User jobs

Panda – needs
a web server

AOD analysis:
1 TB @ 10Hz
 > 139 hrs
Many cores needed

- Common user interface to batch system simplifies users' work
- Panda being testing in Tier 3g (Duke and ANL ASC)
    - Torre is writing the instructions now
- ANL has developed such an interface **ARCOND**
    - Well tested on their system
    - Will need to be adapted for Xrootd storage

Tier 3g – Data storage options

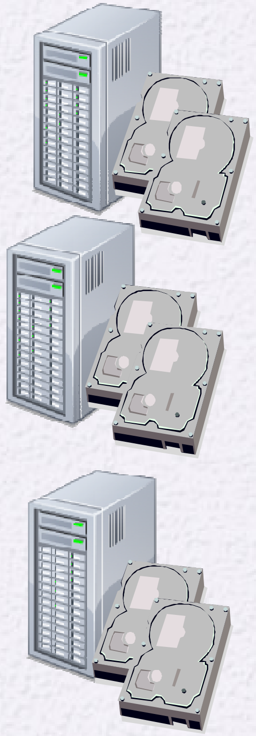


Network
Data flows

File servers

Worker nodes with
little local storage

Storage on worker nodes



XRootD can be used to manage either type of storage

Ntuple processing in Batch

- Ran Sergei Chekanov's standalone ntuple analysis
  - ~20 M events in 100 files (random order)
  - Data in xrootd system on two nodes
  - 14 jobs at a time

		Processing Node (Events/sec)	
Node containing data files	Avg rate per job	ascwrk0	ascwrk1
	ascwrk0	3480	4770
	ascwrk1	4786	5626

Network processing
ganglia 100 MB/s
No I/O wait seen
Data node only serving data

Local processing
(effect of # disks seen)
I/O wait seen on ascwrk0

ascwrk0 – 4 disk raid 5
ascwrk1 – 6 disk software Raid 6

Athena AOD to Ntuple: 100k events

Dell R710: 8 cores (16HT)

Number Simultaneous Condor Jobs	1	4	8	14
NFS4	11 min	12 min	14 min	19 min
Local Disk	15 min	13 min	14 min	19 min
XDR local disk	11 min	11 min	13 min	18 min
XDR* remote disk	14 min	16 min	43 min	

* 2 jobs out of 13 jobs had a read error

Instructions from the beginning

- How to install OS software (kickstart file)

  https://atlaswww.hep.anl.gov/twiki/bin/view/UsAtlasTier3/MakingKickstartUSB

- Will provide LDAP instructions for User account managment

  https://atlaswww.hep.anl.gov/twiki/bin/view/UsAtlasTier3/SetupLDAPserver

- Yushu Yao will help with cluster management (Puppet) -> instructions comming

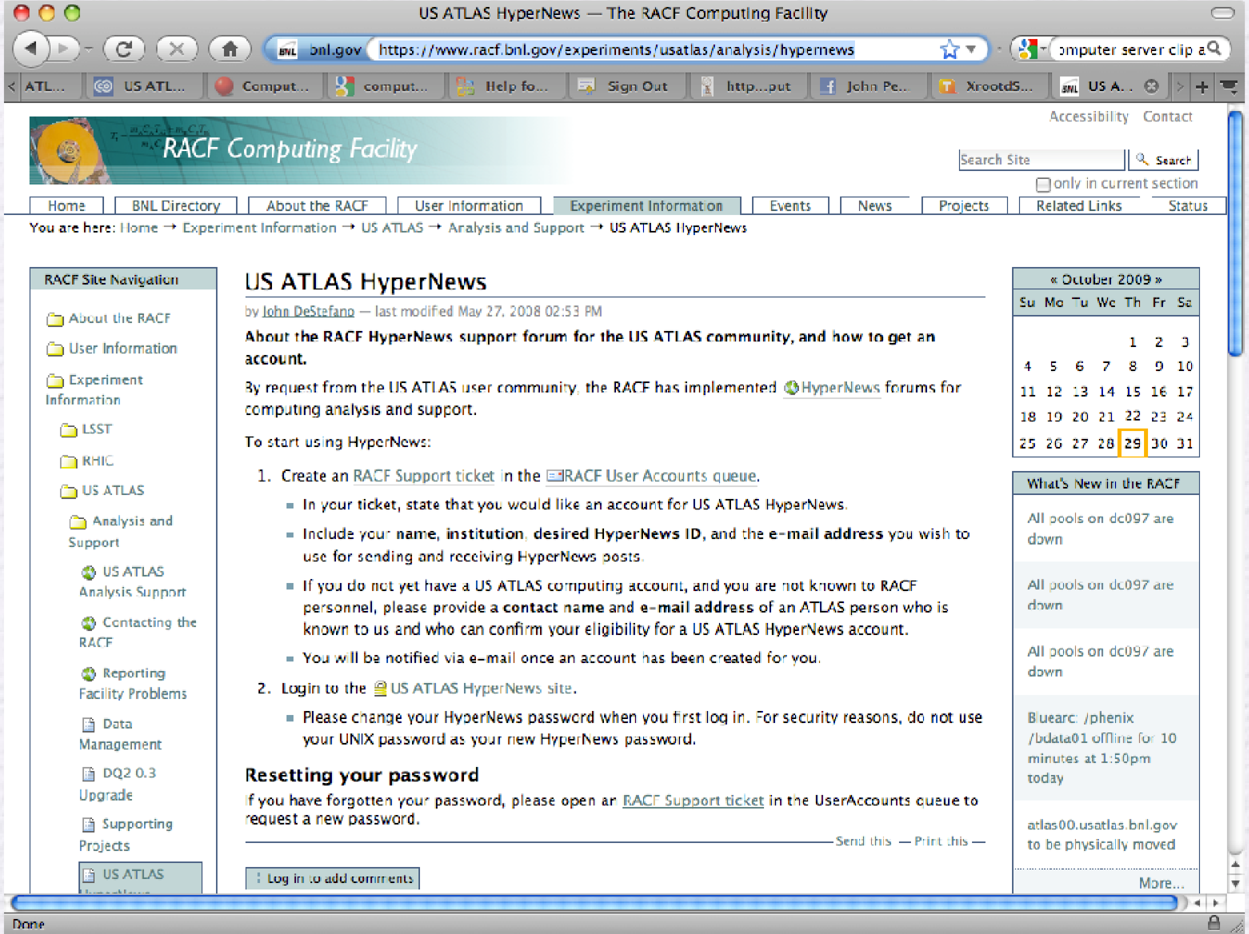
- Virtual Machines are used for Head node services

- https://atlaswww.hep.anl.gov/twiki/bin/view/UsAtlasTier3/CreateVirtualMachines

Tier 3 Hypernews

- Tier 3's will be community supported
  - US Atlas Hypernews - HN-Tier3Support@bnl.gov

    https://www.racf.bnl.gov/experiments/usatlas/analysis/hypernews



You are here: Home → Experiment Information → US ATLAS → Analysis and Support → US ATLAS HyperNews

RACF Site Navigation

- About the RACF
- User Information
- Experiment Information
  - LSST
  - RHIC
  - US ATLAS
    - Analysis and Support
      - US ATLAS Analysis Support
      - Contacting the RACF
      - Reporting Facility Problems
      - Data Management
      - DQ2 0.3 Upgrade
      - Supporting Projects
      - US ATLAS HyperNews

US ATLAS HyperNews

by John DeStefano — last modified May 27, 2008 02:53 PM

About the RACF HyperNews support forum for the US ATLAS community, and how to get an account.

By request from the US ATLAS user community, the RACF has implemented HyperNews forums for computing analysis and support.

To start using HyperNews:

1. Create an RACF Support ticket in the RACF User Accounts queue.
   - In your ticket, state that you would like an account for US ATLAS HyperNews.
   - Include your **name**, **institution**, **desired HyperNews ID**, and the **e-mail address** you wish to use for sending and receiving HyperNews posts.
   - If you do not yet have a US ATLAS computing account, and you are not known to RACF personnel, please provide a **contact name** and **e-mail address** of an ATLAS person who is known to us and who can confirm your eligibility for a US ATLAS HyperNews account.
   - You will be notified via e-mail once an account has been created for you.
2. Login to the US ATLAS HyperNews site.
   - Please change your HyperNews password when you first log in. For security reasons, do not use your UNIX password as your new HyperNews password.

Resetting your password

If you have forgotten your password, please open an RACF Support ticket in the UserAccounts queue to request a new password.

Send this — Print this —

Log in to add comments

« October 2009 »

Su	Mo	Tu	We	Th	Fr	Sa
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

What's New in the RACF

All pools on dc097 are down

All pools on dc097 are down

All pools on dc097 are down

Bluearc /phenix /bdata01 offline for 10 minutes at 1:50pm today

atlas00.usatlas.bnl.gov to be physically moved

More...

Hardware details

- Dell Atlas Pricing hard to beat
- Interactive/Batch nodes
  - AGT2 found Intel E5520 optimum price point
  - 24-36 GB RAM sufficient (too like RAM – get swapping)
  - For Disk heavy nodes Dell R-710 (6-8 disks) (2U)
    - ~ (8 – 2.5" 0.5 TB disks, E5520, 24GB RAM) (4 TB Raw)
    - ~ (6 1 TB disks, E5520, 24GB) (6 TB Raw)
    - ~ (6 0.5 TB disks, E5520, 24GB) (3 TB Raw)
    - ~ (6 2 TB disks, E5520, 24GB) (12 TB Raw)

  (Note- Raid will reduce the amount of space)

Hardware details(2)

- Storage Options
  - Dell 710 w/ E5520 , 24GB RAM
    - 6 2 TB (12 TB Raw) – (see Atlas Portal for price)
  - Dell R710 / Perc6E raid card / MD 1000 Disk shelf
    - 15 1 TB drives (15 TB Raw) (server + (MD1000))
    - 15 2 TB drives (30 TB Raw) (server + (MD1000))
    - Room for expansion (3 Disk shelves per Perc6 raid card)

- Services node –
  - Dell R410 (12 GB RAM , 4 0.5TB disks Raid10) -

- Interactive node –
  - (6 1 TB disks, E5520, 24GB) (6 TB Raw)

Conclusions

- Groups should think about their own analysis needs. Tier 3's should be designed accordingly
- Many of the installation/configuration instructions written and online
  - Other instructions will be online shortly
  - Using scripts when possible to make life easier