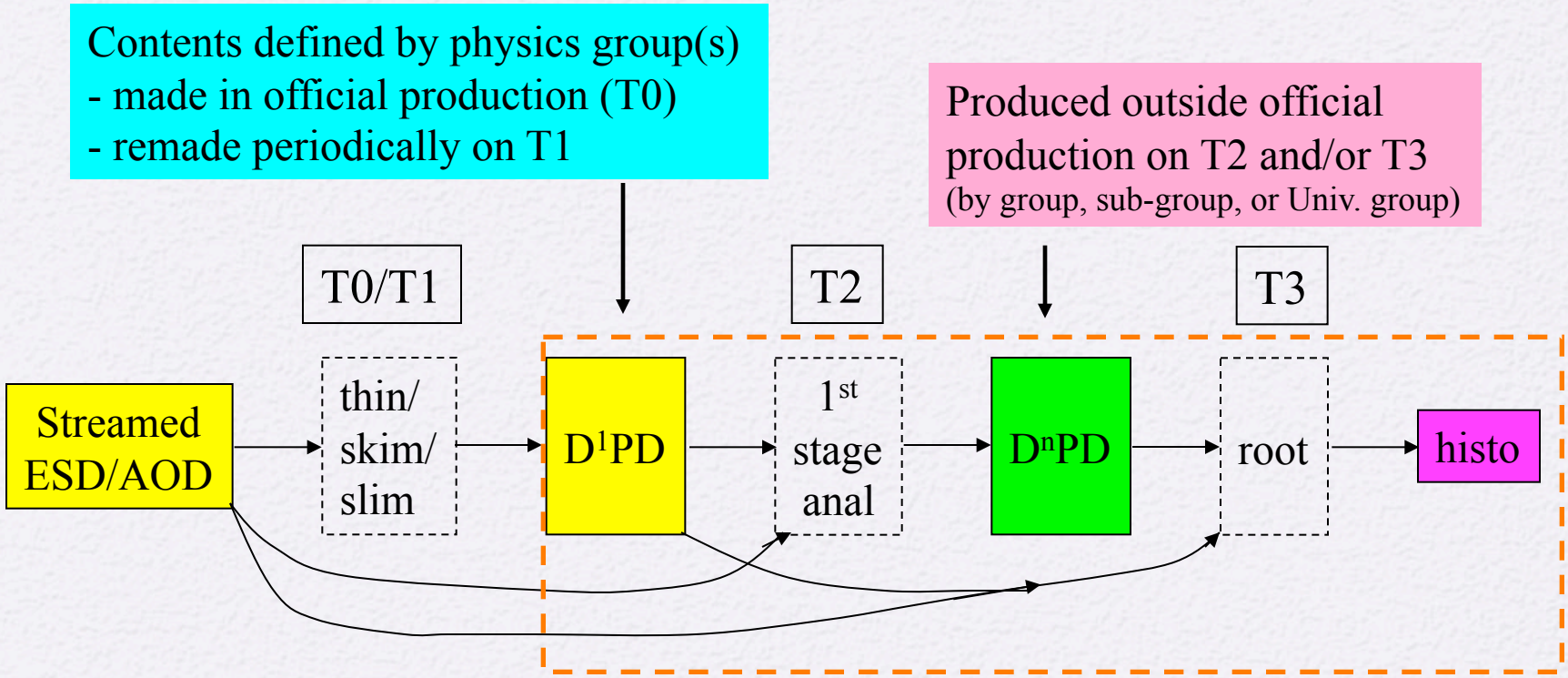
Tier 3 architecture

Doug Benjamin
Duke University

ATLAS Analysis Model – analyzer view



ESD/AOD, D^1PD, D^2PD - POOL based

D^3PD - flat ntuple

Jim Cochran's slide about the Analysis Model

Tier 3's are where people will do a lot of their analysis. "Form must follow function"

Types of Tier 3's

- Tier 3 gs (grid services)
  - Part of US Panda Cloud – accepts Panda jobs
  - Requires significant labor to keep production quality (at least 0.5 FTE - talented system admin).
- Tier 3 w (workstation) (Interactive)
  - Interactive workstation with Atlas Software
  - No batch system - Can submit Pathena or Prun grid jobs
  - All Atlas data retrieved using client tools (dq2-get)
- Tier 3g (most common type)
  - Interactive nodes – submit jobs to grid or local batch
  - Same functionality as Tier 3w
  - Atlas Data through Grid Storage Element and **data subscription**

Tier 3g design/Philosopy

- 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 to support
  - Easier to develop a self supporting community.

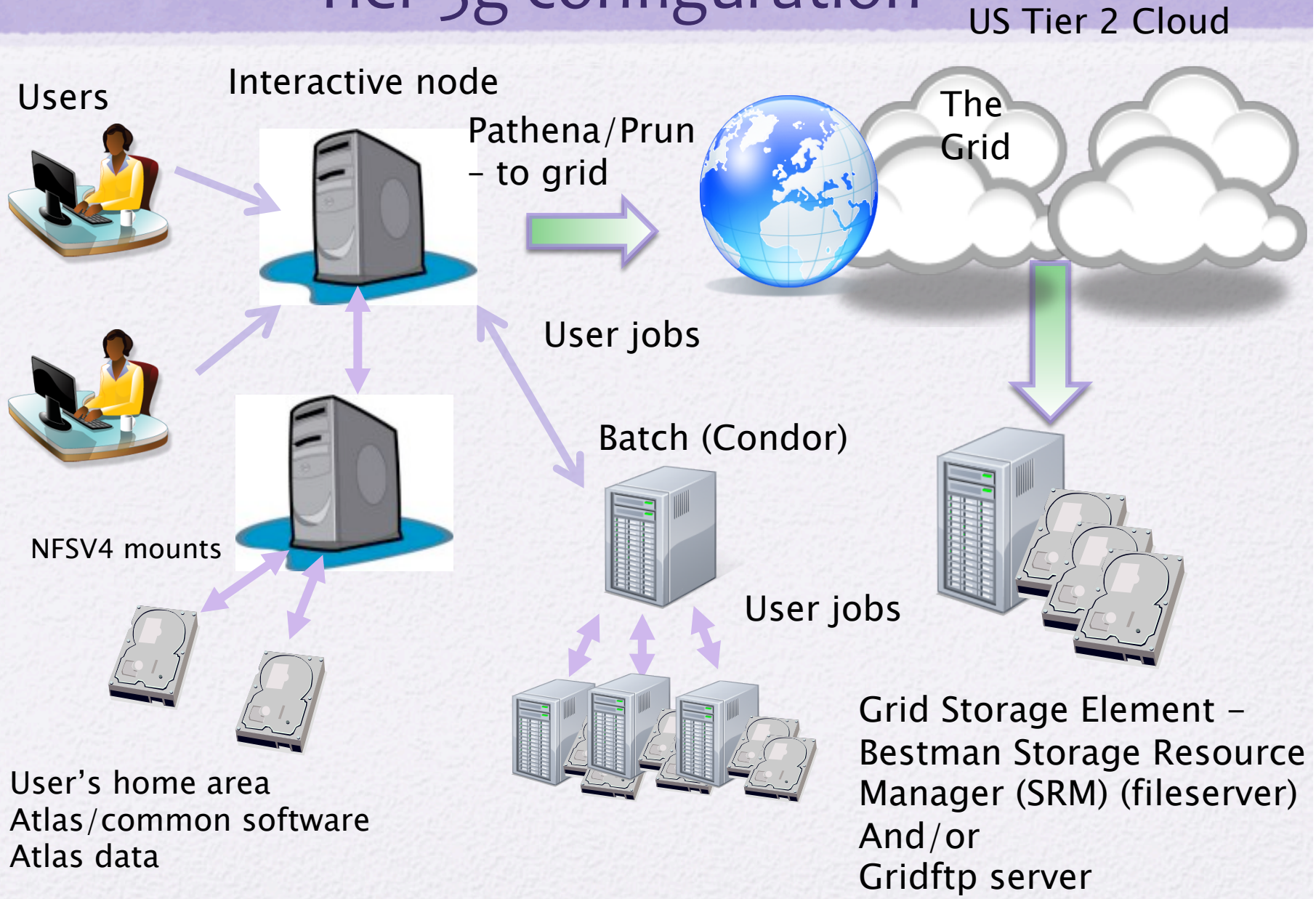
Tier 3 integration

- Tier 3 installation instructions before Jan '10
- Starting to bring on line Storage Elements.
  - Need know limits of Atlas DDM *before* the crunch
  - The more sites the better
- Tier 3's are connected to the Tier 1-2 cloud through the Storage Elements – focus of integration with the US Atlas computing facilities.
- Once ARRA computing funds arrive will focus effort on sites starting from scratch initially

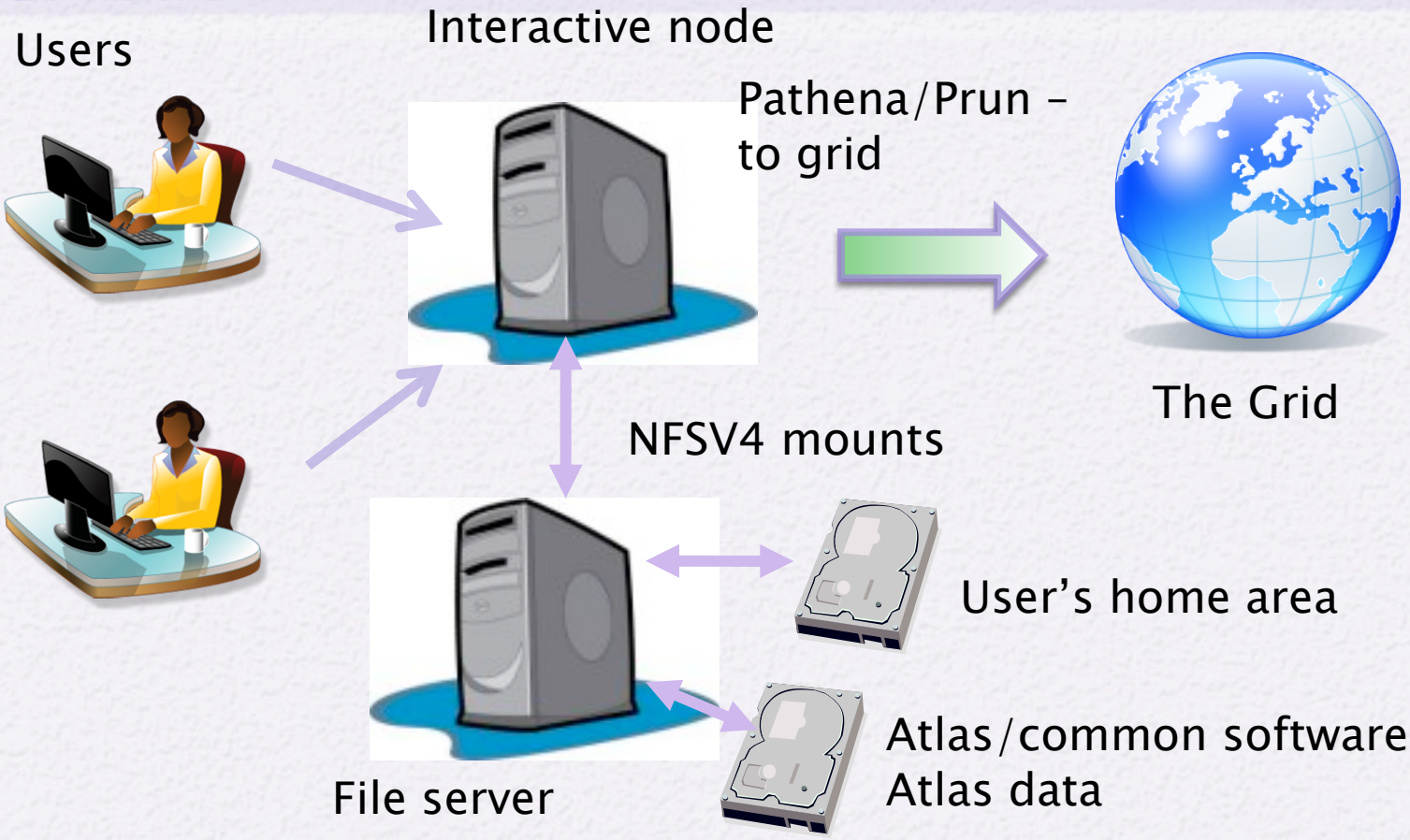
Tier 3's Integration phases

- Initial phase (now until Jan 2010)
  - Develop design and initial instructions
    - Instructions and description being placed on ANL wiki https://atlaswww.hep.anl.gov/twiki/bin/view/Tier3Setup/WebHome
    - Rik is in the process organizing the wiki to make it easier to follow
  - Goal to have vetted/tested (by many US Atlas members) instructions by January
- Next Phase (implementation phase) (~ March 2010)
  - New sites – As ARRA funds arrive new sites will be assembled
  - Existing sites - use the instructions to add services/grow

Tier 3g configuration



Tier 3g – Interactive computing

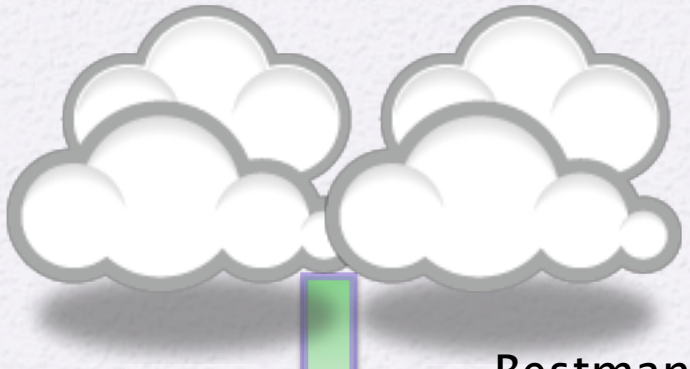


- Design Mostly done – Instructions on wiki – being refined
- Learning how people are using the instructions and adapting them accordingly

How data comes to Tier 3g's

As T3 Storage Elements come online and are tested – will be added to Atlas Data Management System (DDM) as Tiers of Atlas –

US Tier 2 Cloud



Data will come from any Tier 2 site

Bestman Storage Resource Manager (SRM) (fileserver)

Perhaps – Gridftp sufficient

Installation Instructions in wiki (your comments encouraged)

Recent through put test with ANL SE – (> 500 Mb/s)

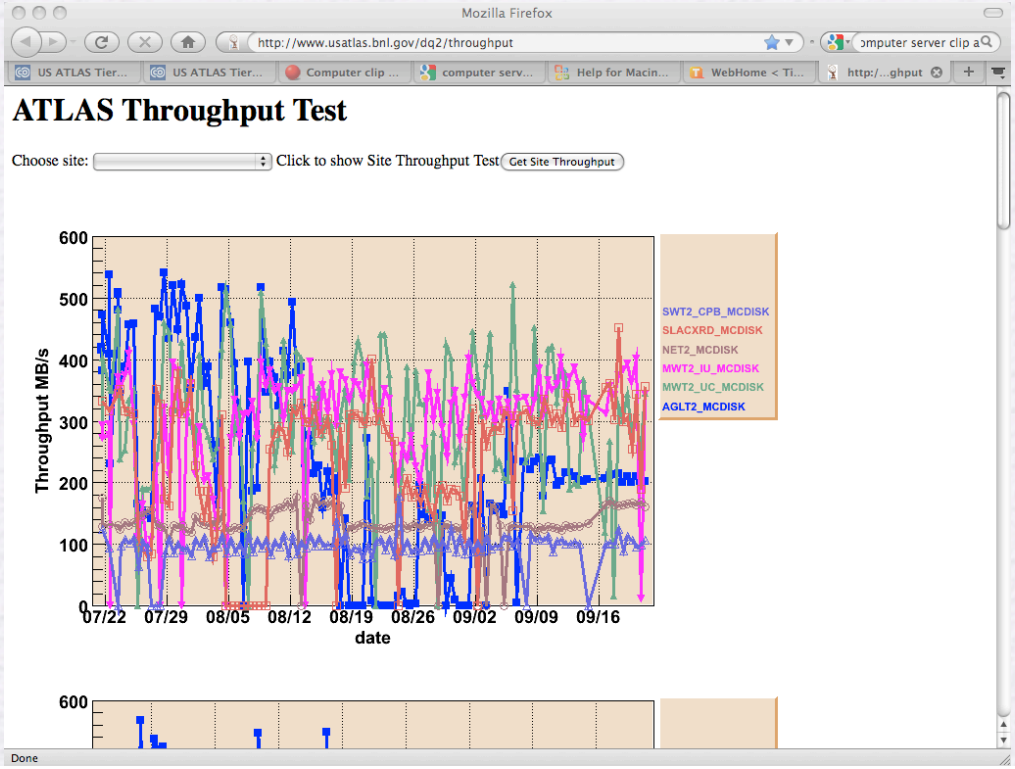
Shows $1200 PC (Intel i7 chip/ X58 chipset/ SL5.3) can be a SE for a small T3.

4 T3g's in Tiers of Atlas (Duke, ANL, Penn and SMU)
-Part of throughput testing
-Asked other T3g's to setup their SE's
(all are welcome/ encouraged to setup SE)

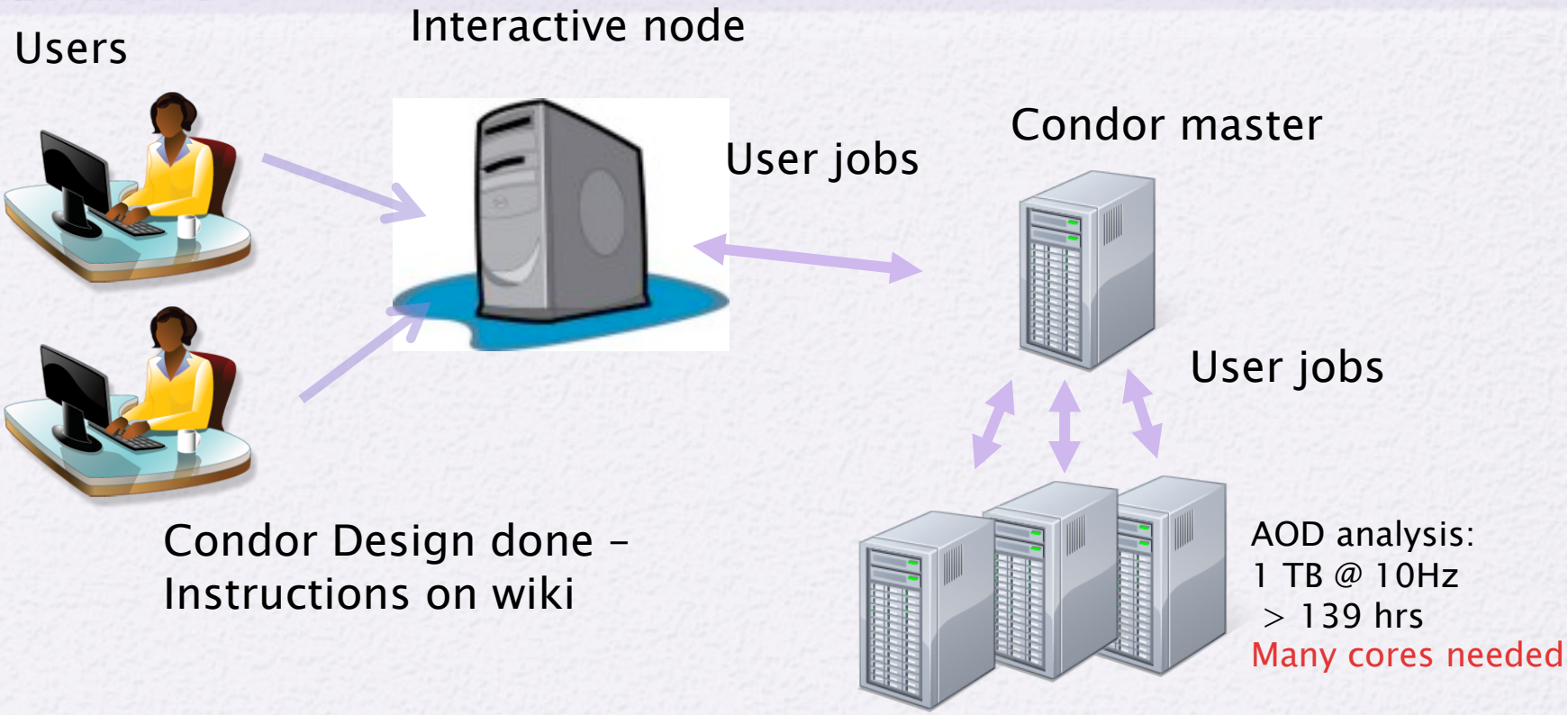
Storage Element installation/testing

- Instructions for Bestman-Gateway SRM
  https://atlaswww.hep.anl.gov/twiki/bin/view/Tier3Setup/SetupSE

- Through put testing and instructions
  http://www.usatlas.bnl.gov/dq2/throughput (testing graphs)
  https://atlaswww.hep.anl.gov/twiki/bin/view/Tier3Setup/ThroughputCleanup



Tier 3g – Batch/ Distributed computing



✧ Common user interface to batch system simplifies users' work

✧ ANL has developed such an interface **ARCOND**

- ✧Well tested on their system
- ✧Will need to be adapted for other Tier 3 sites

✧ Latest verison of Condor 7.4.0 has mechanism that can be used as a local site mover.

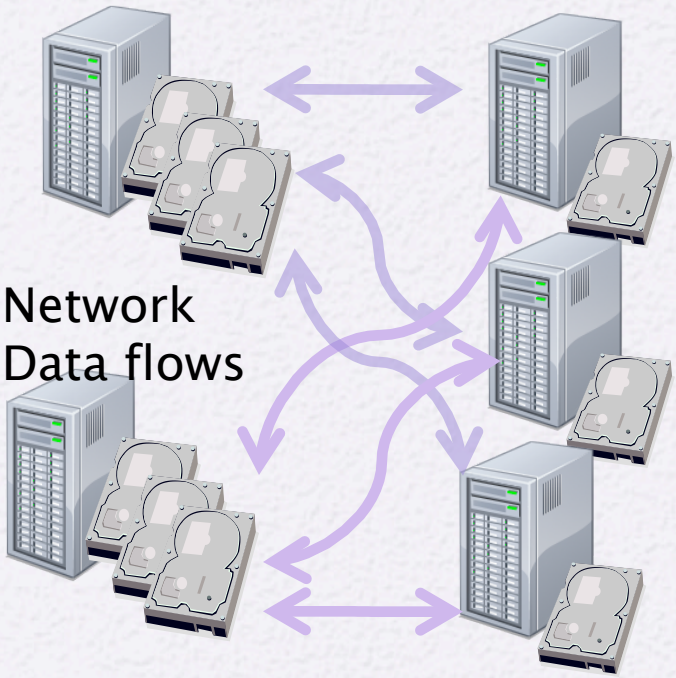
Batch system / storage configuration

- Condor installation/ configuration instructions
  - DB/RK made two trips to Madison Wi. Worked with Condor and OSG team - Instructions developed during the trip.
    - VDT will provide yum repository for Condor
    - Condor Team and Duke/ANL will collaborate on testing of new file transfer code in Condor 7.4.0
  - OSG wrote first pass of XRootD installation instructions

    https://twiki.grid.iu.edu/bin/view/ReleaseDocumentation/XrootdStandAlone
- NFSV4 instructions/scripts
  - On web this weekend (have script to configure NFS mounts already)

Tier 3g – Data storage options

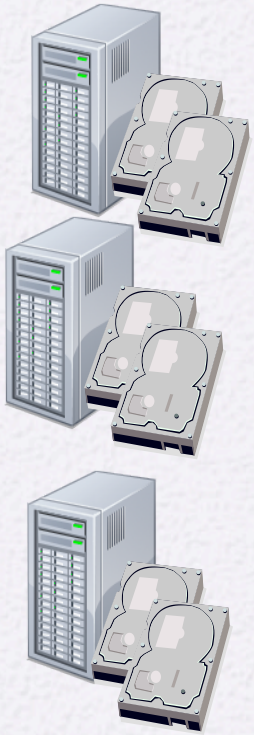
Storage on dedicated storage servers



File servers

Worker nodes with
little local storage

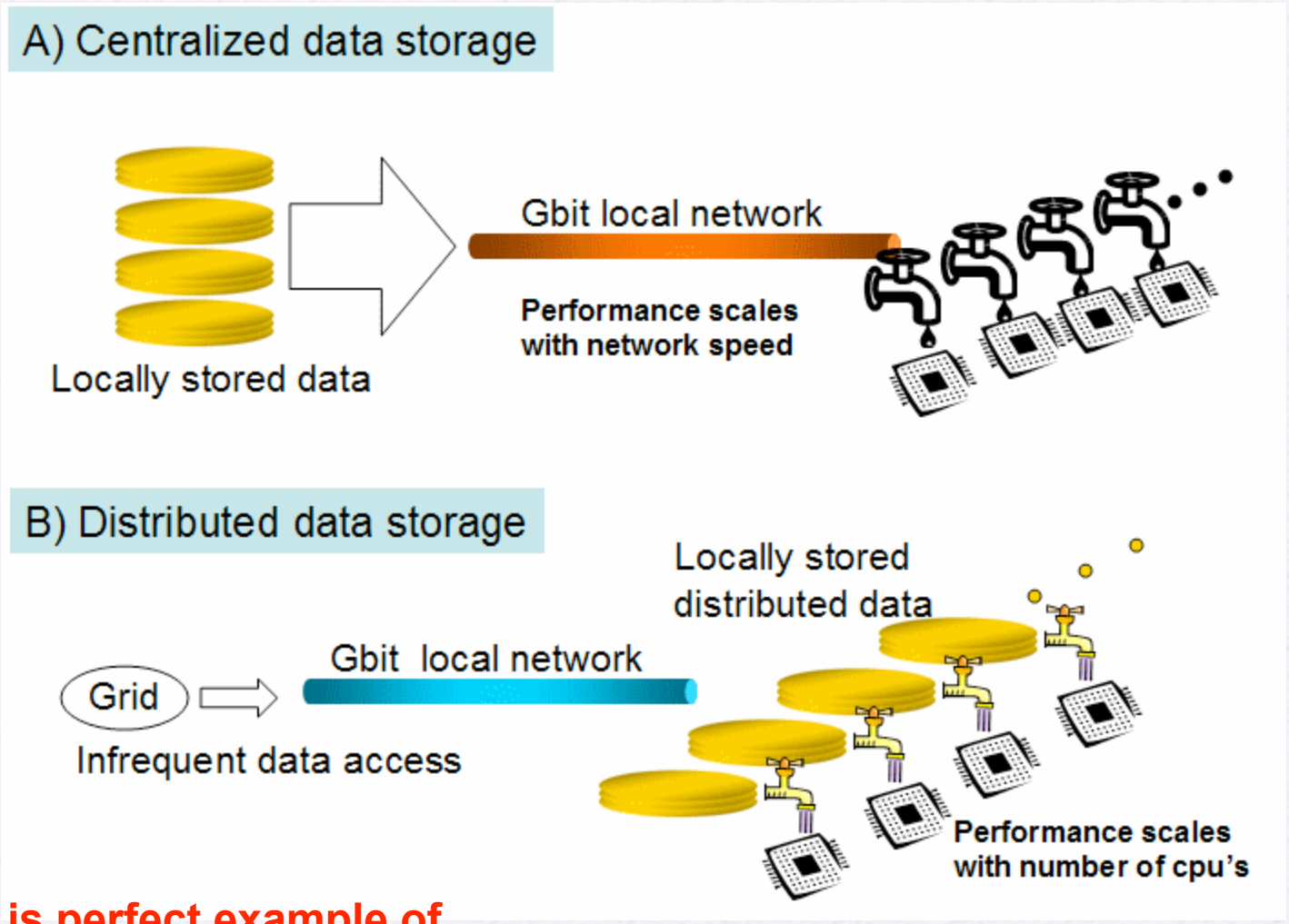
Storage on worker nodes



XRootD can be used to manage either type of storage

Draft of XRootD installation instructions exist

Distributed data advantages (Data stored on worker nodes)



Proof is perfect example of the advantages

Disk on worker node cheaper than in dedicated file servers

Where to find details

- Tier 3 configuration wiki currently at ANL

https://atlaswww.hep.anl.gov/twiki/bin/view/UsAtlasTier3/Tier3gSetupGuide

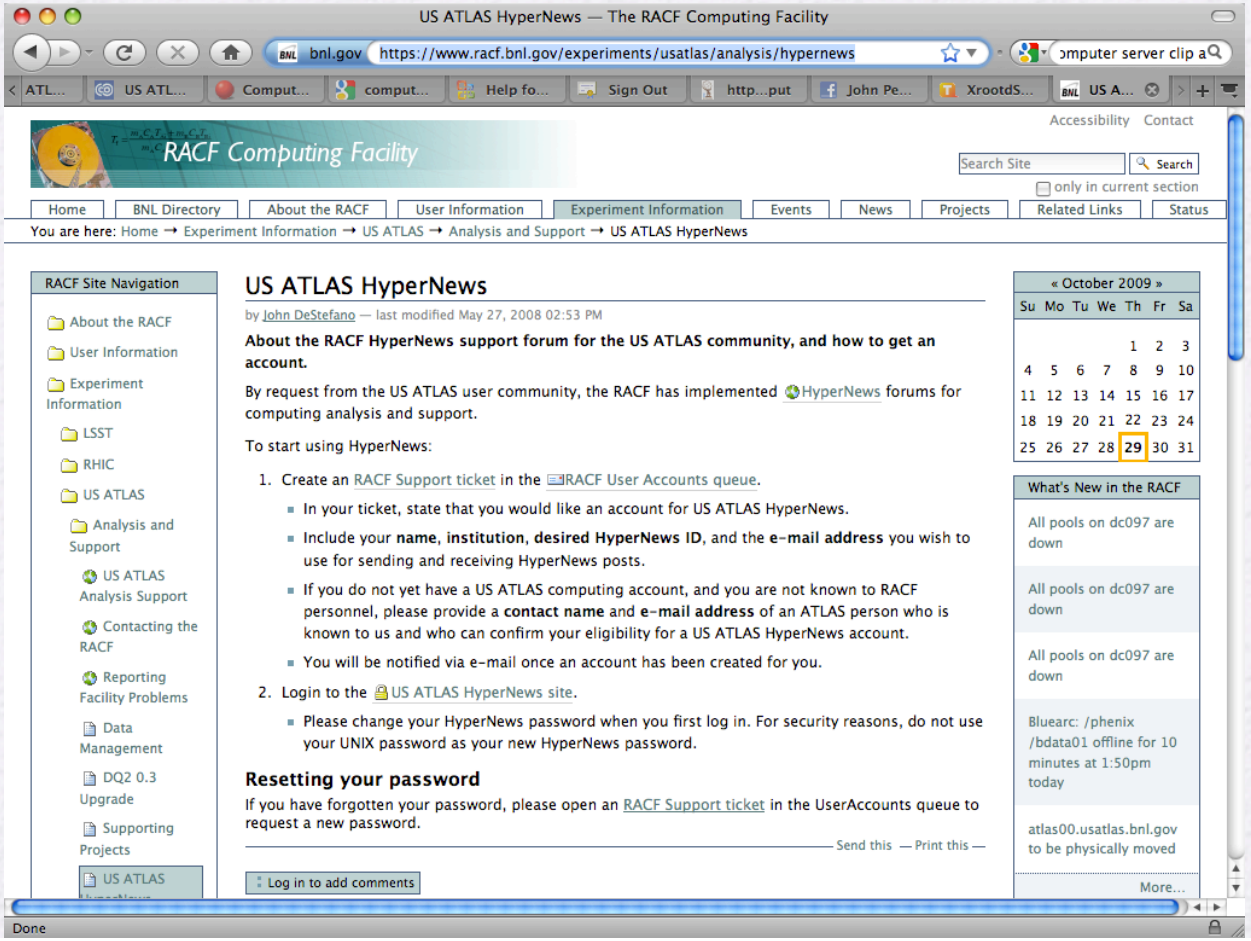
  - Rik has begun to reorganize the wiki
    - Improve the flow of the instructions
    - Make it easier for users to us
    - In draft form *(personal opinion – ready for comments and editing by users)*

Tier 3 Support/Help

- US Atlas provides ½ person for T3 Support
- Open Science Grid (OSG) will help
- Our software providers are helping (Condor Team, XrootD team)
  - Rik and Doug just spent 2 days working with OSG and Condor team to setup and provide documentation for US Atlas Tier 3's.
- Tier 3's will be community supported
  - US Atlas Hypernews - **HN-Tier3Support@bnl.gov**
  - US Atlas Tier 3 trouble ticket at BNL
    - USAtlasTier3RT-RACF-USAtlasTier3@bnl.gov

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



Tier 3 Self organization

- As a result of recent Tier 3 meeting – many people volunteered to help. 
- Effort organized into 7 areas
  - Documentation
    - Jim Cochran and Iowa State will follow instructions at ISU
    - Paul Keener configure things at Penn
    - Tom Rockwell will start at MSU
    - Craig Blocker (Brandeis) will use on test cluster at ANL
  - Benchmarking (evaluate / develop)
    - Waruna Fernando (OSU)
    - Alden Stradling (UTA)

Tier 3 self organization (page 2)

- Hardware configuration
  - Justin Ross (SMU) test blades
  - Sergey Panitkin (BNL)
  - Waruna Fernando (OSU)
  - Tom Rockwell (MSU)
- Cluster Management
  - Tom Rockwell (MSU)
  - Rob Gardner (Chicago)
  - Justin Ross (SMU)
  - Marco Mambelli is surveying OSG sites for their solution

Tier 3 efforts (Grid Storage Management)

- Grid Storage management

  (Do we need SRM + Gridftp or Gridftp alone)

  - Duke, ANL, Penn and SMU are up.
    - Recent FTS change caused the sites to fail until reconfigured. (need better notification mechanism)
  - Iowa State and Ohio State are setting up SE Element
  - Duke and ANL will test subscription.
  - Iowa State has agreed to test Panda job output to Tier 3 SRM
  - Working to add additional sites when possible.

Tier 3 Future…

- Continue to investigate technologies to make Tier 3's more efficient and easier to manage
  - Virtual machine effort continues.
    - Collaboration between BNL, Duke, OSU, LBL, UTA, CERN
    - Bestman-gateway SRM running in VM at BNL as proof of concept. (Will tune it for performance)
    - XRootD redirector next
    - Condor part of CERNVM
- ANL will setup integration and test clusters
  - Integration cluster – small Tier 3 – test code there before recommending upgrades in rest of Tier 3 cloud. (including two virtual clusters – Duke & Dell contributed hardware)

Tier 3 efforts (Virtualization / Web file system)

- Virtualization
  - Based on Xen and CERNVM
    - Will investigate other hypervisors if needed
  - Virtualization will allow better standardization (reduce overall support load)
  - Volunteers: Sergey Panitkin (BNL), Waruna Fernando (OSU), Yushuo Yao (LBNL), Alden Stradling (UTA), DB (Duke)
- Web file system
  - Yushuo Yao (LBNL) has RPM – tested at Univ Washington.
  - Looks promising for solution of Atlas software
  - Works on VM or physical machine
  - Need to evaluate space management issues
  - Squid cache's at Tier 3 will be required for best performance

Conclusions

- Tier 3's are *our* tool for data analysis
  - Encouraging user feedback on the design, configuration and implementation
  - Several people have volunteered to help on tasks now.
- Tier 3's will be community supported
  - We should standardize as much as possible
    - Use common tools (modify existing tools to make them more inclusive)
- Next 5 months are planning on a flurry of Tier 3 activity.
  - User involvement now will pay off very soon
  - Thanks for excellent attendance Tier 3 meeting shows people are getting involved