

# Tier 3 Data Access Considerations

Rob Gardner

3-9-10

US ATLAS Facilities Meeting

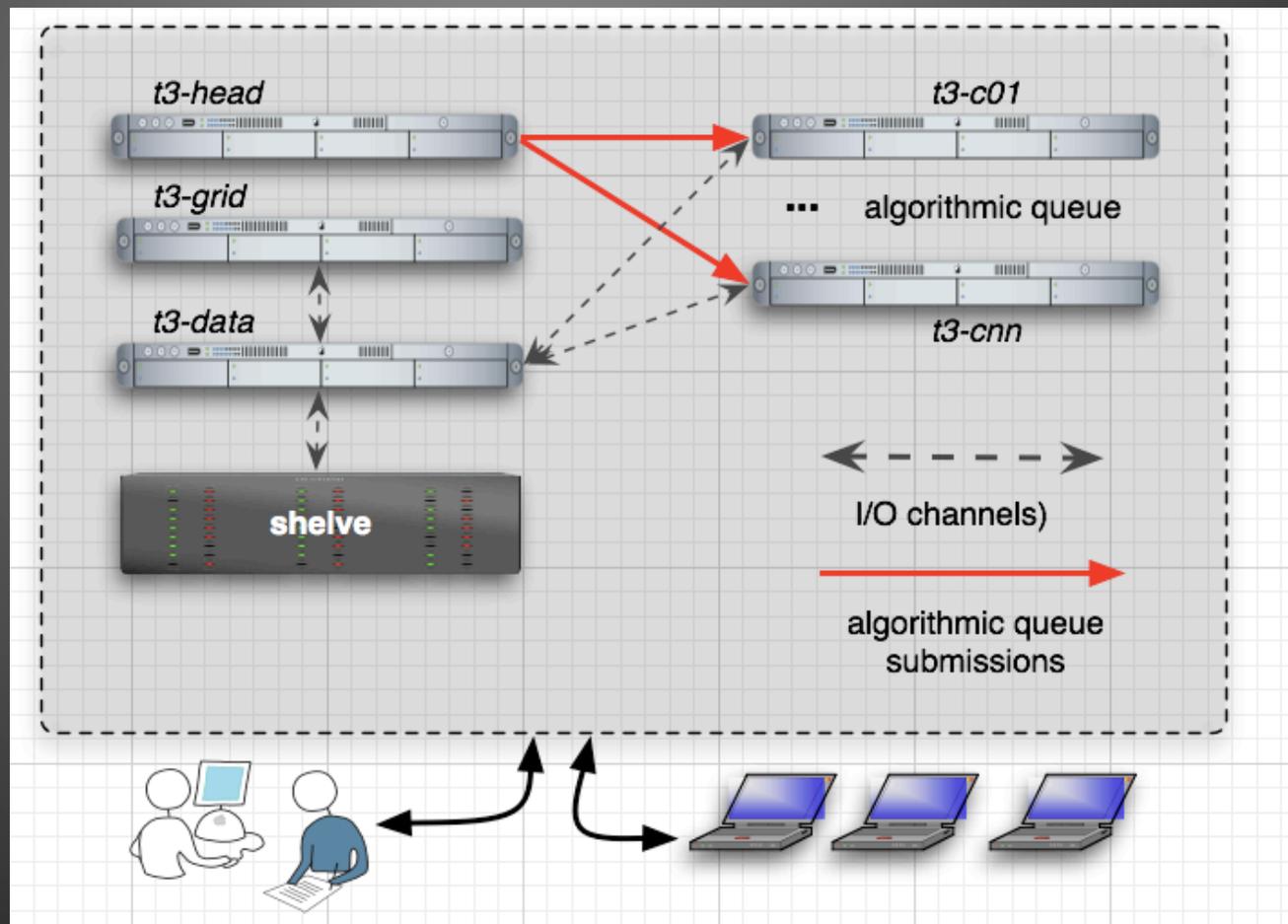
OSG All Hands - Fermilab

# General T3 Data Access Considerations

- Tier 3 systems come in various sizes and architecture
  - Cluster of worker nodes providing job slots only (A)
  - Cluster of worker nodes that have significant local disk (B)
- In Case A
  - typically a network attached storage system perhaps providing a distributed file system (GPFS, Lustre)
  - or storage system (xroot data servers + master with redirector and name space server)
- In Case B
  - Worker nodes can be stacked with lots of disk relatively cheaply
  - Access therefore potentially simplified: data is local, possibly fewer services needed
  - Job scheduling with data location service, PROOF
  - Nodes run jobs and provide data service

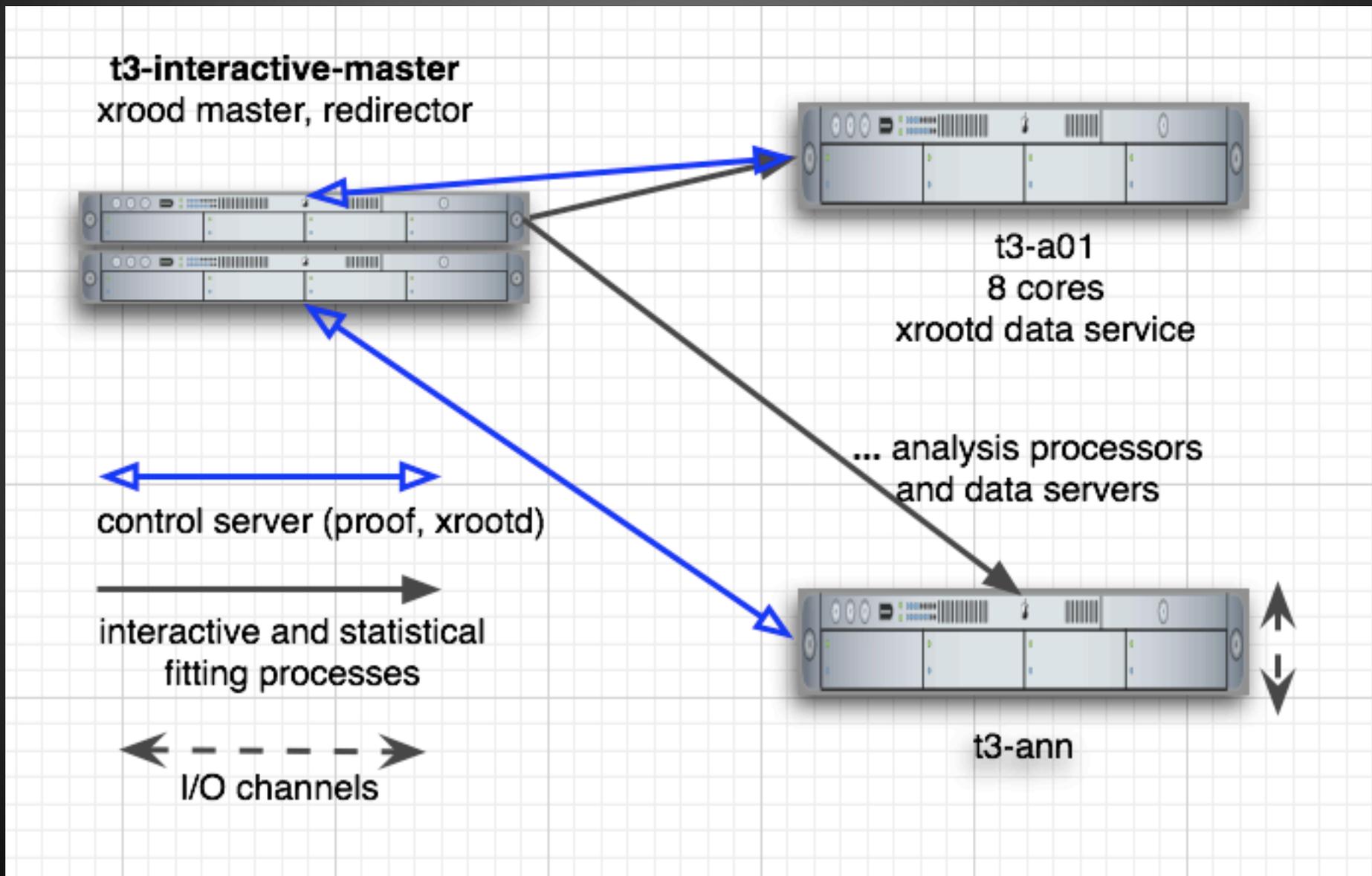
# Type A

- Thin worker nodes (1U, lightly “disked”)
- Eg: storage system - “storage node” +  $\geq 1$  SAS attached shelves
- Filesystem (Lustre/GPFS) or xrootd

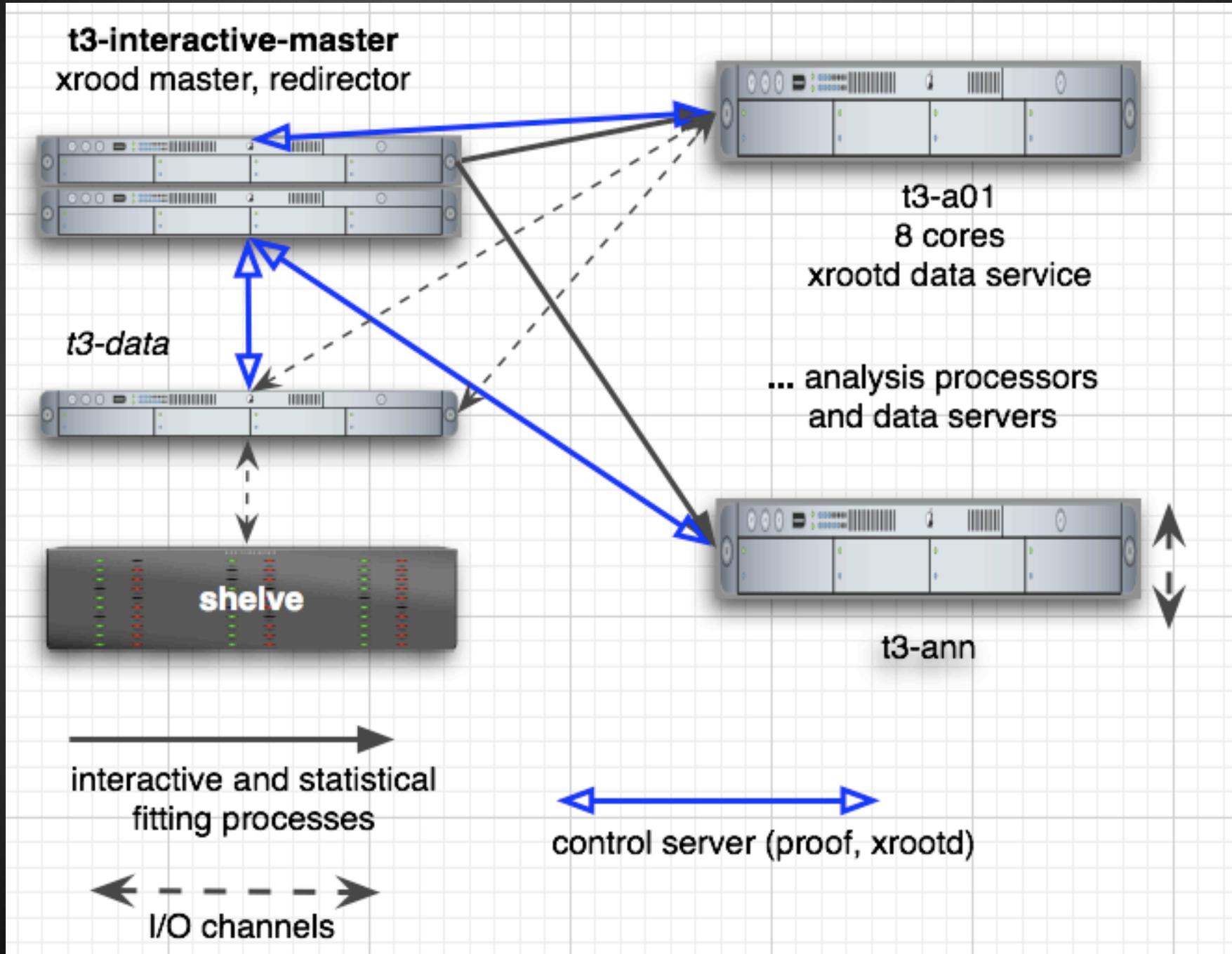


# Type B

- Worker-local-storage-rich nodes (eg. for xrootd)



# Type A-B Combo



# ATLAS T3 Distributed Storage WG

- Technical area to evaluate and make recommendations for distributed storage systems
- Technologies to be evaluated: xrootd, Lustre & GPFS as indicated interests from ATLAS Tier 3 workshop
- Charge:
  - Document the current usage in Atlas in Tier 2 and Tier 3 sites
  - Determine and make available best practices guidelines.
  - Develop suggestion for deployment at all Tier 3 sites.
  - Propose test metrics for the considered design and tabulate the results
- Leader (s) - two sub-groups formed
  - Santiago Gonzalez de la Hoz (Valencia): Lustre & GPFS
  - Rob Gardner (Chicago): xrootd

# Milestones

- week 3: site survey results, best practices wiki
- week 6: Preliminary configuration recommendation
- week 6-9: testing phase
- week 12: Final recommendation of Tier 3 configuration and documentation (wiki)

# WP1 Distributed Storage (Lustre & GPFS)

(leader Santiago González de la Hoz, IFIC-Valencia)



- Membership

- **LUSTRE:**

- UAM-MADRID Tier 2 (Juan Jose Pardo and Miguel Gila)
    - LIP-COIMBRA Tier 2 (Miguel Oliveira, Helmut )
    - BONN-Physikalisches Institut (Simon Nderitu)
    - DALLAS-Southern Methodist University (Justin Ross)
    - IFIC-VALENCIA (Javier Sánchez and Álvaro Fernández)
    - ISRAEL T2/T3 Federation-Weizmann Institute, Tel Aviv University, The Technion (Lorne Levinson and Pierre Choukroun)
    - DESY (Yves Kemp and Martin Gasthuber)
    - U. OKLAHOMA (Horst Severini)

- **GPFS:**

- Edinburgh (Wahid Bhimji)
    - Italian sites (Gianpaolo Carlino and Fulvio Galeazzi)

- **DATA ACCESS:**

- CERN (Andrea Sciaba)

## Status

- **First goal:** To have a real overview of technologies, configuration (HW ad SW) at various sites using the Lustre/ GPFS File System and the current usage in ATLAS
  - Milestone 1: site survey result, Best practices wiki
- For that:
  - A twiki page (**LustreTier3**) has been done linked on AtlasTier3 general twiki:
    - <https://twiki.cern.ch/twiki/bin/view/Atlas/LustreTier3>
  - A **survey form/questionnaire** for Lustre has been done
    - <http://spreadsheets.google.com/viewform?formkey=dFVFQkFFczdORDY2bC1raTRkd21hN1E6MA>
    - We have already first results for all sites sites
  - A **survey form/questionnaire** for GPFS has been done
    - <http://spreadsheets.google.com/viewform?hl=en&formkey=dGdiMU5aaJNvYnNSRktoOWhSQ3V5aWc6MA>
  - Some **twiki pages** with current Lustre and GPFS configuration in each site has been updated and linked on LustreTier3 twiki page.

## Status

- **The survey form** has questions about the current configuration (general issues, metadirectory, disk servers, clients, etc) in order to have a real overview of the technologies.

Timestamp	Which is the name of your institute/university?	Which Tier2 do you have associated? (i.e. IFIC-LGG2)	Are you sharing Lustre with your Tier2 or another infrastructure (i.e. department)?	Is your Tier3 a Grid or non-grid site?	How many users are using your Tier3?	How many groups are using your Tier3?	Do you have any comment?	Are you using a different machine for MGS and MDT?
2/19/2010 19:14:50	Southern Methodist University	US-SWT2	NON	Grid with Storage Element with SRM (Bestman)	~20	1		both on same
2/22/2010 18:25:32	Israel Tier-2/Tier3 Federation (Weizmann Institute, Tel Aviv University, The Technion)	Israel Tier-2/Tier3 Federation (Weizmann Institute, Tel Aviv University, The Technion)	YES	Yes, the Tier2	35	three	We use Lustre for both local storage and SRM storage via STORM. For the time being we have DCache SE in production and we are in process of migrating to Storm-Lustre	different machines.
2/23/2010 14:05:03	Physikalisches Institut, Universität Bonn S:stname.UNI-BONN	No associated	NON	yes, dCache-SRM. We are installing Storm-Lustre	70	Atlas, IL, Zeus, (to be used by IceCube)		NO

# Lustre & GPFS

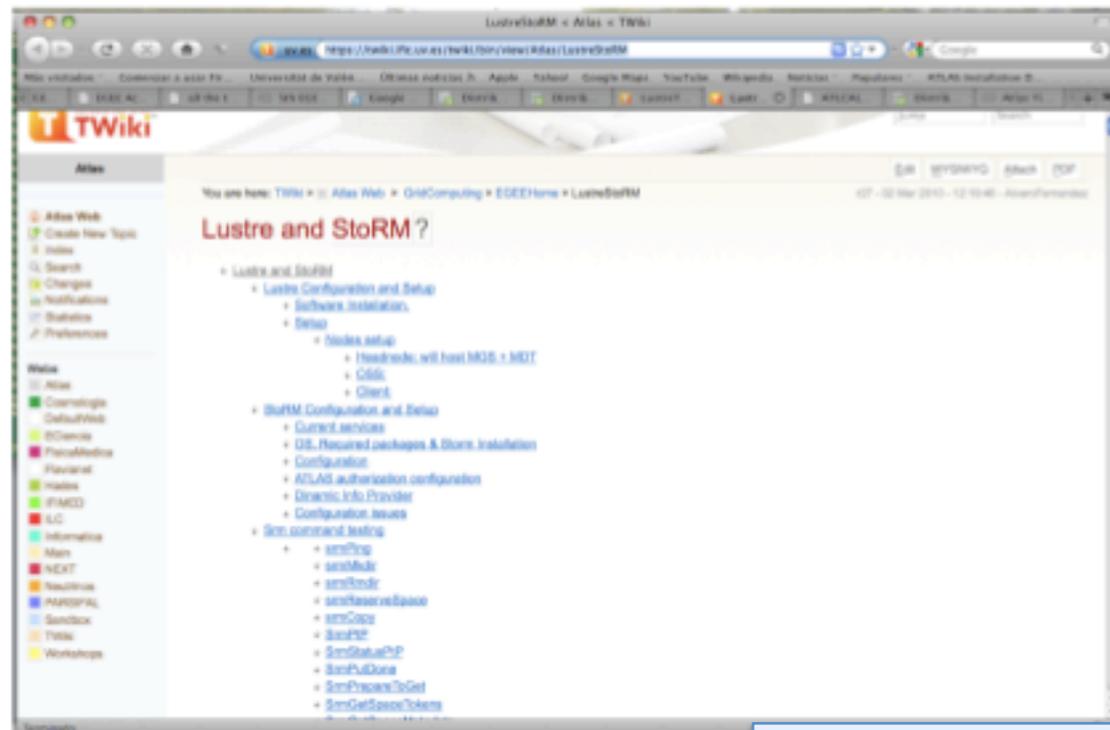
Are you using HA meta-data server?	Which connectivity are you using?	How much memory do you have in your MDS?	How many processors/cores do you have in your MDS?	How many file systems are you using (MDTs)?	Are you using Lustre pool features?	Are you using quotas?	Are you using users ACL's?	What disk capacity do you have in Lustre (In Total)?
not there yet, hope to purchase the hardware soon.	1 Gbps	24 GB	2xquad core intel	1	no	no	no	240TB
NO, we are not yet at Lustre V2. We are at Lustre V1.6	1 Gbps	16GB	2 x Quad-Core AMD Opteron(tm) Processor 2384	one	No	No	yes	Weizmann site: 112TB usable Tel Aviv Univ: 80TB usable Technion: 80TB usable
NO	1 Gbps	32GB	2 x quad core Intel Xeon CPU @ 3.00 GHz	1	NO	YES	YES	200TB

How many disk servers are you using?	How many OSTs by disk server are you using?	How many disks has an OST?	What kind of disk redundancy is using the OST?	How much memory do you have by disk server?	How many processors/cores do you have by disk server?	Are you using another kind of connectivity?	Which connectivity are using your disk servers?	How many network interfaces do you have?
6	10	on a RAIDed array.	Hardware (Nissan SATAbeast) Raid 6	24	2 quad core intel	no	1 Gbps	4
weo at each site	Weizmann 16 Tel Aviv 10 Technion 10	10 as RAID-6 8+2	HW RAID6	16GB	2 x quadcore Intel (R) Xeon(R) CPU NOE5410 @ 2.33 GHz	No	10 Gbps	one
6	2	12	RAID6	16GB	2 x quad core Intel Xeon CPU @ 3.00 GHz	NO		

How many clients (nodes and cores) do you have with access to Lustre?	What kind of clients are accessing to Lustre?	Which connectivity are using your clients?	Do you have read-write access?
today, 28 clients with 224 cores very soon, 124 clients with 992 cores	WN	1 Gbps	yes
Weizmann 73 nodes, 500 cores Tel Aviv 23 nodes, 184 cores Technion 36 nodes, 288	WN, UI, SE,	1 Gbps	yes

## Status

- **The institute twiki page** has information about the current Lustre/GPFS installation.
  - For instance Lustre+StoRM installation at IFIC (<https://twiki.ific.uv.es/twiki/bin/view/Atlas/LustreStoRM>)



## Status

- **Whit that information we have reached the first milestone and goal.**
- Later, using the survey questionnaire and twiki pages, **we must think about the best practices guidelines and suggestions for deployment at a Tier3 sites**
  - What is different among us?
  - What do we have in common?
  - ....?
- **Next Step** (Milestone 2): Preliminary configuration recommendation

# Xrootd T3 team

- Team forming
  - XrootD with PROOF and batch queues overlaid (Neng Xu)
  - OSU operates a PROOF farm with xrootd (Waruna Fernando, Harris Kagan)
  - NYU operates a PROOF farm with xrootd (Attila Krasznahorkay)
  - ANL and Duke are testing xrootd for batch queues (Doug Benjamin and Rik Yoshida)
  - Tier 2 centers: Y. Wei (SLAC) & Patrick McGuigan (UTA)
  - Distributed xrootd @ Tier 2 (Sarah Williams, MWT2)
  - PROOF farm on xrootd with batch queues overlaid at DESY Wolfgang Ehrenfeld
  - Laura Sargsyan, Sarkis Mkoyan (Yerevan)
  - Privet Vladimir (Dubna)
  - Andrea Sciaba (CERN)
  - Xrootd development team: Andy Hanushevsky, Wei Yang, Fabrizio
- Wiki: <https://twiki.cern.ch/twiki/bin/view/Atlas/XrootdTier3>

# Xrootd discussions

- Need to take an inventory across ATLAS
  - Probably include Tier 2 centers
- Which deployment features to include in survey?
  - Number of data server nodes
  - Number of computing cores served
- What about import/export area?
  - Note there are very few stand-alone t3 sites so difficult to say.
  - ANL test cluster will probably be setup in this way - has to do with how the data gets populated; also depends on whether t3 is attached to a t2. also depends on development of dq2-get extensions
- Preliminary survey:
  - <http://www.surveymonkey.com/s/atlasxrootd>

# Xrootd discussions

- What are their plans with integration w/ ATLAS DDM tools, SRM, etc
  - note there is another working group to study this
  - Consider SRM and data import questions secondary
  - Focus on reading access with data already present; performance
- Reliability of loaded systems
- Installation and configuration options for an "ATLAS-standard" xrootd Tier 3
  - In association with PROOF
  - Source of packages: from ROOT; from VDT; managing patches
  - Fabrizio Furano:
    - working on a simple xrootd install and configuration; maximum simplicity.
    - Additional features can be incorporated easily, such as federating sites. repo has tags of CVS at slac and a scripted build system.

# Xrootd discussions

- Testing and validation
  - Stand-alone testing
    - Varying #clients, data distribution for various configurations
  - Hammer cloud testing
  - ANL testing xrdcp from same node
    - expect to have results shortly
  - Massimo - discussion w/ Dan for HC tests. Could send a couple of datasets.
  - HC-Panda jobs, using mini-pilot factory. For sites already with a CE it should be more straightforward.
  - Two new T3 queues created at ANL and Duke