

# ADC Development and R&D

Simone Campana

CERN IT-ES

# ATLAS Computing Model

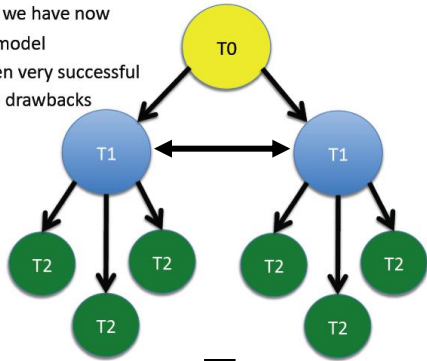
- Global computing within a hierarchy of Tiered centers organized in (primarily) regional 'clouds'
  - *Optimize data flow in an environment of expensive, limited-bandwidth networks*
- Cloud model has limitations that are avoidable in the era of abundant networking
  - Cloud boundaries negatively impact efficient data processing, data placement and access, disk space usage, network bandwidth, ...
- **Exploit ability of our networks to move data effectively *between* as well as *within* clouds**

# Operational Experience

- Our Tier 1s have been very effective – as expected – as flexible, efficient processing and data serving centers
- A lesson of experience is that Tier 2s are extremely effective and flexible as well – and not just as an appendage of a Tier 1
- Greater equivalency between Tier 1s and Tier 2s can lead to fuller and more efficient resource utilization
- **Trend today and for the future: flattening of the hierarchy**

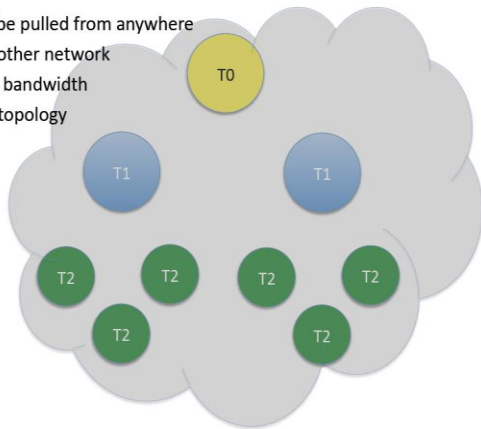
# Steps toward a new model

- This is what we have now
- It is a push model
- And has been very successful
- But has also drawbacks



*Nordic Grid didn't deploy MONARC model and NDGF model is very close to one we are considering now*

- Data can be pulled from anywhere
- Needs another network
- Not more bandwidth
- Different topology



- PD2P: Dynamic Data Placement

- Automatically spread in-demand data across resources

- Tier2D: Tier2s directly connected to Tier 1s, Tier 2Ds and CERN

## Evolve the model; not a discontinuous change

Network monitoring becomes essential

Tier2 connectivity vital; T2s becomes T2Ds when capable and validated

Utilize a hybrid approach of strategic pre-placement and dynamic usage-driven placement; continuously optimize

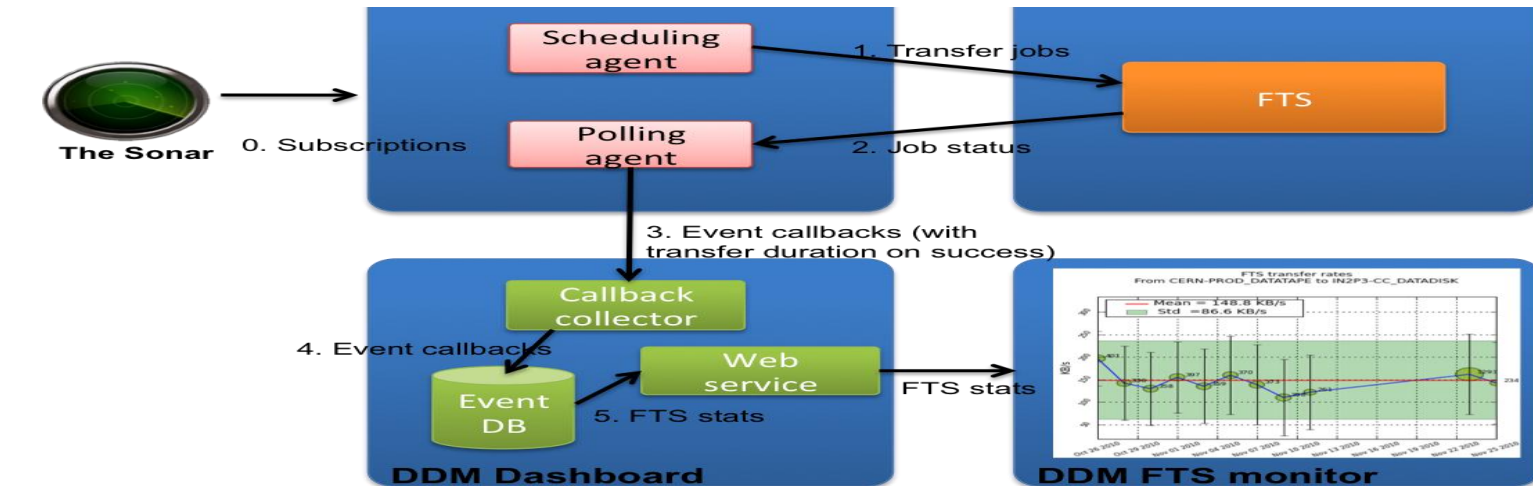
# Growing Importance of Tier 2s

- Nearly half are now T2Ds – can service data requests across ATLAS
  - Target is all Tier 2s are T2Ds
- T2Ds can join clouds other than their home cloud – multi-cloud mode – so clouds are fluid and flexible rather than static to respond to production needs
  - eg. grow a cloud beyond its regional bounds to pump a high priority task through quickly
  - T1s can join another cloud in the same way
- T2s as primary replica repositories under study
- T2s hosting more and more group production activities, using dedicated group space at T2s

# Evolution of Data Placement

- Move towards intelligent caching of data rather than solely planned placement
- Decrease number of *primary* replicas.  $N$  *primary* replicas are always guaranteed
- 2011 began with no planned data placement at Tier2s, only dynamic caching
  - T2 analysis activities dropped; strategic planned placement reintroduced to complement caching
- Use data popularity and data age information to regulate number of replicas ATLAS wide
- Use planned placement for the data samples, formats known to be popular (D3PDs, AODs)
- Allow users to drive placement themselves (DaTRI)
- Use both Tier1s and Tier2Ds as data sources

# Network monitoring: sonar



Site Name	SrcSite	SrcCloud	SrcTier	DetSite	DetCloud	DetTier	AvgBRS(MB/s)	EvS	AvgBRM(MB/s)	EvM	AvgBRL(MB/s)	EvL	Prio
AGLT2_1o_AUSTRALIA-ATLAS	AGLT2	US	T2D	Australia-ATLAS	CA	T2	0.41+/-0.02	10	5.54+/-1.72	110	0.00+/-0.00	0	2
AGLT2_1o_BEIJING-LCG2	AGLT2	US	T2D	BEIJING-LCG2	FR	T2D	0.24+/-0.04	10	1.51+/-0.38	11	2.46+/-0.85	301	5
AGLT2_1o_BNL-OSG2	AGLT2	US	T2D	BNL-ATLAS	US	T1	1.76+/-1.53	8	19.62+/-11.40	1204	38.94+/-15.20	293	8
AGLT2_1o_CA-ALBERTA-WESTGRID-T2	AGLT2	US	T2D	CA-ALBERTA-WESTGRID-T2	CA	T2	0.91+/-0.38	11	7.94+/-2.61	10	16.29+/-6.70	10	5
AGLT2_1o_CA-MCGILL-CLUMEQ-T2	AGLT2	US	T2D	CA-MCGILL-CLUMEQ-T2	CA	T2	0.89+/-0.04	10	6.39+/-1.05	10	22.22+/-3.84	10	5
AGLT2_1o_CA-SCINET-T2	AGLT2	US	T2D	CA-SCINET-T2	CA	T2D	0.78+/-0.03	10	4.90+/-1.02	10	15.47+/-0.80	10	5
AGLT2_1o_CA-VICTORIA-WESTGRID-T2	AGLT2	US	T2D	CA-VICTORIA-WESTGRID-T2	CA	T2D	1.48+/-0.20	10	10.56+/-0.98	10	41.14+/-8.44	10	5
AGLT2_1o_CERN-PROD	AGLT2	US	T2D	CERN-PROD	CERN	T0	1.10+/-0.18	10	6.21+/-2.13	10	15.99+/-6.83	10	7
AGLT2_1o_CSCS-LCG2	AGLT2	US	T2D	CSCS-LCG2	DE	T2D	0.44+/-0.11	5	1.31+/-0.23	5	11.57+/-4.94	5	5
AGLT2_1o_CSTCDIE	AGLT2	US	T2D	csTCDie	NL	T3	0.57+/-0.08	10	4.14+/-0.98	10	0.00+/-0.00	0	0
AGLT2_1o_CYFRONET-LCG2	AGLT2	US	T2D	CYFRONET-LCG2	DE	T2	0.09+/-0.01	5	0.92+/-0.35	5	4.79+/-2.55	5	5
AGLT2_1o_DESY-HH	AGLT2	US	T2D	DESY-HH	DE	T2D	0.66+/-0.04	5	3.34+/-0.50	5	27.85+/-4.29	5	5
AGLT2_1o_DESY-ZN	AGLT2	US	T2D	DESY-ZN	DE	T2D	0.47+/-0.12	5	0.91+/-0.15	5	7.42+/-0.80	5	5
AGLT2_1o_FZK-LCG2	AGLT2	US	T2D	FZK-LCG2	DE	T1	0.11+/-0.00	5	1.40+/-1.28	6	7.21+/-5.26	249	7
AGLT2_1o_GOEGRID	AGLT2	US	T2D	GoeGrid	DE	T2D	0.13+/-0.00	5	0.79+/-0.01	5	4.80+/-0.47	5	5
AGLT2_1o_GRIF-IRFU	AGLT2	US	T2D	GRIF-IRFU	FR	T2	0.25+/-0.03	10	0.36+/-0.08	10	0.40+/-0.06	10	5
AGLT2_1o_GRIF-LAL	AGLT2	US	T2D	GRIF-LAL	FR	T2D	0.93+/-0.02	5	4.76+/-0.71	5	19.40+/-4.76	5	5
AGLT2_1o_GRIF-LPNHE	AGLT2	US	T2D	GRIF-LPNHE	FR	T2D	0.58+/-0.05	5	1.12+/-0.92	10	3.99+/-7.41	10	5
AGLT2_1o_HEPHY-UIBK	AGLT2	US	T2D	HEPHY-UIBK	DE	T2	0.15+/-0.02	5	2.62+/-0.18	5	6.88+/-1.46	5	5
AGLT2_1o_IFAE	AGLT2	US	T2D	ifae	ES	T2D	0.57+/-0.21	5	5.80+/-0.98	5	10.41+/-6.07	5	5
AGLT2_1o_IFIC-LCG2	AGLT2	US	T2D	IFIC-LCG2	ES	T2D	0.28+/-0.07	5	1.87+/-0.84	5	4.57+/-3.13	5	5

<http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview#currentView=Sonar>

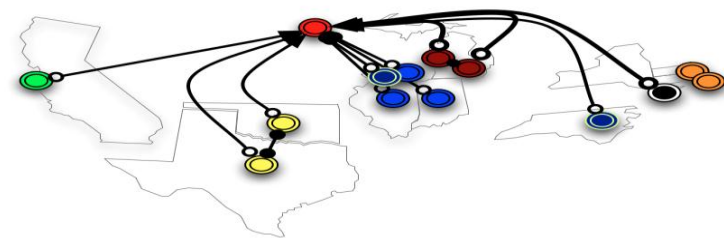
# Network Monitoring: perfSONAR

- ATLAS suggests the deployment of perfSONAR at all T1s and T2s.
  - Point to point network monitoring based on iperf.
  - Latency and throughput
- Details in:  
<https://twiki.cern.ch/twiki/bin/view/LHCONE/SiteList>
- We consider network monitoring a key aspect of the infrastructure
  - We asked WLCG to support this initiative as part of the project. Very positive feedback.

# Data Federations

Slides by  
R. Gardner

- ATLAS and CMS started an R&D on storage federations 2 years ago.
- Currently, the only usable technology is xrootd
- Basically all storages offer LAN and WAN access capabilities through xrootd



# Xrootd Federation R&D to Production

- Positive experience in USATLAS
- Will extend to pilot sites in Europe.
  - UK and DE volunteered to try regional federations
- Use cases of increasing complexity
  1. Fallback solution for I/O issues at with the local storage
  2. Brokering at sites with no complete data sample

# Federations and WAN access

- Obviously, need good WAN access performance



<http://ivukotic.web.cern.ch/ivukotic/WAN/index.asp>

- TTreeCache increases performance in avg of 3x
- Studies ongoing.

# Evolving Distributed Data Management: The Next-Generation DDM System

- Objective: streamline & simplify the system based on operational experience and new developments
  - Eliminate unused features, focus on performance and scalability of critical functions, add new needed capabilities, track evolution of middleware, leverage newly available technologies
- Impact: DDM has many dependent clients, from end users to distributed analysis to the production system
  - Boundary condition: no interruption to ongoing DDM operations (present version supported through transition)
  - The largest single ADC development project for the near future
- Status: conceptual model for the next-gen system 'Rucio' documented and currently under discussion with the many user communities
- Timeline:
  - Architecture document & detailed design ~Jan 2012
  - Iterative development throughout 2012
  - Validation and deployment for production in 2013

# SRM and protocols

- ATLAS is in line with the outcome of the WLCG TEG:
  - SRM will still be needed as interface to Mass Storage
  - In general, the access to storage will rely less and less on SRM
  - ATLAS already by-passes SRM whenever possible (i.e. local access)
- Development in pipeline
  - ATLAS tools being able to leverage other protocols for storage access
    - gridFTP w/o SRM, xrootd, http
  - Grid tools being able to leverage other protocols for storage access
    - gridFTP is already there, need FTS-3 for others.

# CVMFS

- Software installation, database releases and conditions data
  - CERNVM File System as a common technology
- ATLAS invites all sites to migrate to CVMFS ASAP
  - Sometime soon, no more need of HOTDISK
  - Thanks to CVMFS we can run on nightlies on the Grid
- Very soon, ATLAS SW installation will move away from WMS and use Panda
  - No more need of special high priority queues

# AGIS

- ATLAS is developing a Information Service (AGIS)
  - Used for some use-cases already
- AGIS caches information from various sources
  - Including the BDII
- Correctness of infos in the BDII is very important
  - Queue status
  - Disk Space
  - Pledges and HEPSPEC benchmarks

# Multicores

- Multicore or whole-node scheduling coming
  - 64 bit reco memory footprint & AthenaMP
  - local batch system scaling - otherwise #slots blows up
  - cloud computing - whole-node (VM) by definition
  - Reduce number of transient files and access to storage
- but not there yet ... : deployment
  - dedicated multi-core resources wasted when only serial jobs
  - serial-to-Ncore is expensive – drain last job
  - Ok for steady stream of Ncore jobs
- But not there yet... : software
  - AthenaMP for Reco and G4sim, but not digi (prevents MC reco usage)
  - File merging is still serialized ...
- Need to get more experience. Not ready for production.

# gLexec

- ATLAS si not particularly fan of gLexec ...
  - We believe it creates more problems than it solves
- Nevertheless, WLCG Security TEG provided quite clear indications
  - In addition we can leverage the identity switching to optimize resource usage
- Two possibilities under evolution:
  - handle gLexec from the ATLAS pilot
    - Original idea, was fragile but many problems could be fixed by now. Re-testing.
  - Handle gLexec through glideinWMS
    - It works, but introduce a considerable overlay and complication
- If gLexec is broken at your site today, we will not complain...

# Cloud Computing

- Buzz word or powerful tool?
- Could be the latter if it offers the uniform homogeneity across computing resources that grids never quite gave us
- Potential to complement and augment the grid
- Adoption of virtualization by facilities has begun
- Academic clouds already exist, commercial clouds may (sooner or later) be cost-effective for peak processing
  - CERN has one! lxcloud
- Funding agencies have started asking us about cloud utilization
- **Prudent course: cloud-enable our processing so we can evaluate utility of clouds and be able to leverage them**
  - Answer for ourselves: are they useful? Will they be?
  - Effective integration with storage is a central issue

# ATLAS Cloud Computing R&D

- Condor Cloud Scheduler (U Victoria)
  - Leverage Condor to run PanDA pilots on CERNVM VMs, no PanDA changes required
  - Cloud Scheduler manages VM creation and proxy based authorization/authentication
  - Proof of concept working, needs pilot factory for real-world use
- Cloud Factories (CERN)
  - Understand how to run PanDA sites in the cloud
  - Manage VMs directly using a cloud API
  - Running example on CERN's Ixcloud with CERN SE data access
  - Target use case: a personal cloud factory
    - Eg. Amazon + credit card = analysis done by conference deadline

# Cloud Futures

- Many of the activities are reaching a point where we can start getting feedback from users
  - Should focus in the next months on eliminating options, and determine what we can deliver in production
- Cloud Storage
  - This is the hard part. Some free S3 endpoints are just coming online, so effective R&D is only starting now.
  - Looking forward to good progress in caching (Xrootd/HDFS in cloud) and DDM S3 Evaluation (Rucio incubator proposal)
- Support the grid sites who want to offer private cloud resources
  - Develop guidelines, best practices
  - Good examples already, e.g. LxCloud, PIC, BNL, and others.

# NoSQL (ops, Structured Storage)

## □ Relational database management systems

- Vertical scalability (“scale up”)
- Few powerful nodes
- Shared state
- Explicit partitioning
- Resistant hardware
- ACID
- Implicit queries (*WHAT*)

## □ Structured storage

- Horizontal scalability (“scale out”)
- Lots of interconnected low cost nodes
- Shared nothing architecture
- Implicit partitioning
- Reliability in software
- BASE
- Explicit data pipeline (*HOW*)

### Main problems addressed:

1. There is an upper limit of processing power you can put in a single node
2. Explicit partitioning can be cumbersome
3. Relaxation of ACID properties can be necessary
4. Query plans need information about the data contents

# Comparison of technologies

	MongoDB	Cassandra	Hadoop/HBase
<b>Installation/ Configuration</b>	Download, unpack, run	Download, unpack, configure, run	Distribution, Complex config
<b>Buffered read 256</b>	250'000/sec	180'000/sec	150'000/sec
<b>Random read 256</b>	20'000/sec	20'000/sec	20'000/sec
<b>Relaxed write 256</b>	10'000/sec	19'000/sec	9'000/sec
<b>Durable Write 256</b>	2'500/sec	9'000/sec	6'000/sec
<b>Analytics</b>	Limited MapReduce	Hadoop MapReduce	MapReduce, Pig, Hive
<b>Durability support</b>	Full	Full	Full
<b>Native API</b>	Binary JSON	Java	Java
<b>Generic API</b>	None	Thrift	Thrift, REST

**ATLAS and the WLCG Database TEG decided for Hadoop**

# NoSQL: “Big data processing”

- DDM use cases
  - disk pool access, log file analysis, trace mining, file sharing, accounting, search
  - Leveraging an existing Hadoop installation at CERN
- Panda use cases
  - Data Mining and Monitoring
  - Currently running on Cassandra@BNL
    - will be ported to Hadoop@CERN

# Summary

- Operations is Key, but some amount of computing development and R&D is essential to keep up with the state-of-the-art in scalability and efficiency
- In the short term we expect to see some concrete deliverables in production
- Work is still ongoing in all of these activities – some activities are closer to wrapping up than others.
- Thank you to all who help prepare these slides!