ATLAS TDAQ System Administration: evolution and re-design

CHEP 2015

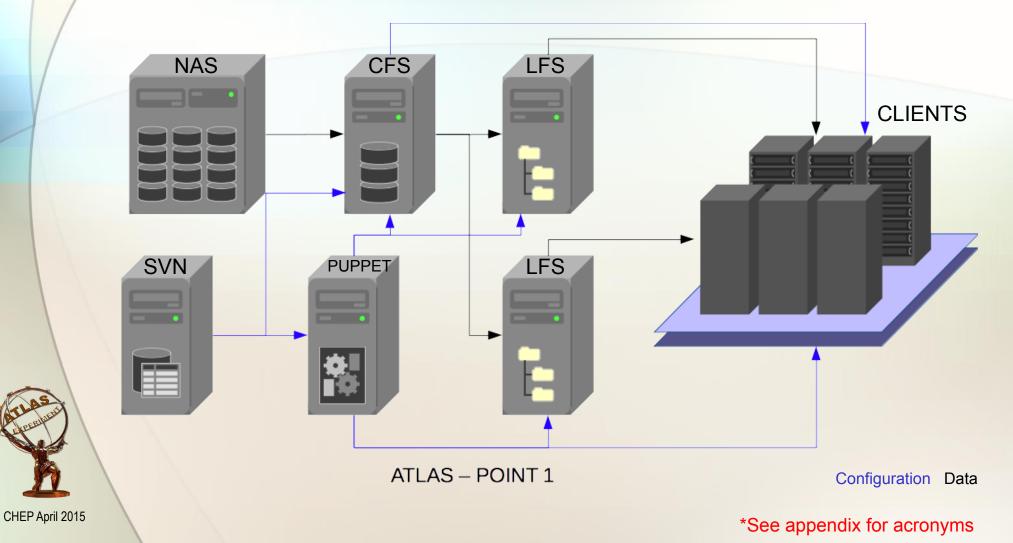
ERN

Christopher Jon Lee University of Johannesburg, South Africa CERN

for and on behalf of the ATLAS TDAQ SysAdmin team

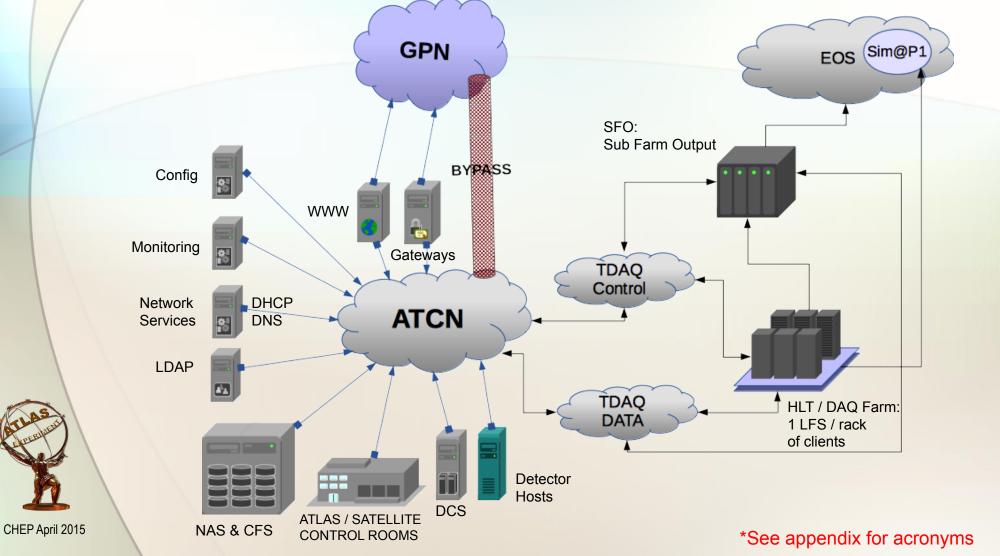
Overview

After 3 years of LHC beam (Run1), 2 years of upgrades (LS1), Restart has begun...



Overview

After 3 years of LHC beam (Run1), 2 years of upgrades (LS1), Restart has begun...



3

OS Upgrade

- Scientific Linux CERN
 - Only supported Linux OS in use
 - Full support direct from on-site experts
 - All Linux machines are now running SLC6
 - will remain the OS (Major) version for Run 2

CÉRM

- Windows OS
 - Used by Detector Control System (DCS) for one specific application
 - SLC6 hosts Windows VM
 - Windows VM is managed by DCS



During beam, NO changes are made to the running system

Local boot vs Net boot

- 664 Local boot Servers, DCS, TDAQ/SysAdmin Infrastructure
- 2392 Net boot HLT Farm, Read Out Systems, Single Board Computers, etc
- Local boot (Standard installation with boot from disk)
 - provisioning by PXE + Kickstart + 🔬 puppet labs
 - DHCP + PXE provided by an LFS from Configuration Database (See slides below)
 - template-based kickstart files
- "Net Boot" via PXE
 - the more components one has in a system, the greater the risk of failure, So... reduce any components that are not "needed"
 - in ATLAS, extensive use of PCs with no operating system on disk
- Each reboot, is essentially a fresh clean OS

Advantages:

- ease of maintenance
- reproducibility on a large scale
- reduced HW replacement times

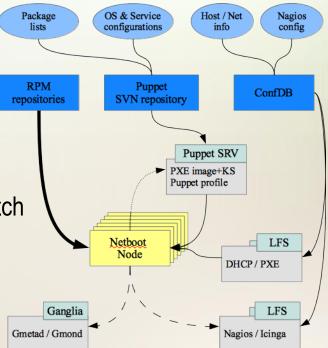
Disadvantages:

- requires ad-hoc development and support
- not suitable for running servers
- less flexible

Netbooted – Redesign for SLC6

A completely new netbooting system compared to SLC 4 & 5

- based on <u>NFSroot</u> and customized to our needs
 - only R/W areas are kept in RAM (e.g.: /etc, /var, ...)
 - "bind-mounts" overlaid on R/O NFS mount of / from the LFS
 - gives the users more free RAM for running their apps
- Image created by puppet in a chrooted environment
- NO "Golden Image"
 - always able to rebuild from versioned config
- Support for old hardware
 - 32 bit non-PAE kernel provided and maintained by CERN IT (on a best effort basis)
 - ELF image for non-PXE clients requires private patch of mknbi package



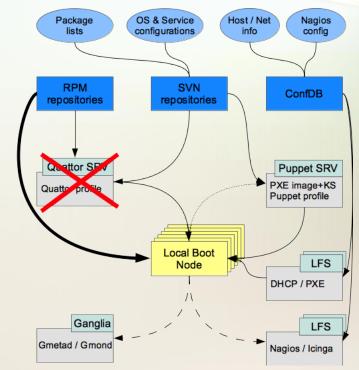


CHEP April 2015

C.J.Lee - UJ, CERN - ATLAS TDAQ SysAdmin

Configuration Management Systems

- With such a large cluster of machines CMS's are the only way to sanely control what happens on machines in a large farm
- Quattor was retired as it was becoming obsolescent
- A puppet chosen over Chef and CF Engine
 - previous experience by existing Sys Admins
 - during LS1 CERN IT also adopted Appropriate
 - WLCG applications mostly puppetised
- All systems are Puppet controlled
 - local boot : Daemon run, every 30 min
 - net boot: puppet --apply via hourly cron job
 - no need to reboot in order to apply simple configuration changes
 - Very similar system between the two
 - code re-usability
 - easier to maintain





- ConfDB is our core Configuration DataBase
 - PHP based web UI, Python for utilities and REST API
 - configuration "state" database of all systems
 - DHCP details operational status
 - boot type / parameters
 - entire system other than ConfDB is maintained by "code"
 - interface between CERN IT Databases and Appropriate
 - web UI provides various tools to ease cluster management ssh, IPMI etc
- Included an OS release system for SLC6+
 - different release versions can run on different machines
 - useful for testing new versions and/or revert changes in case of problems
- More functionality added to REST API, used by puppet and other tools e.g.
 - network interface configuration (including bonding)
 - getting machine status in puppet and configure it accordingly (e.g. TDAQ or Sim@P1*)

Virtualisation

- Mostly CORE and TEST systems, NOT for DAQ/HLT
- No cloud-like approach
 - no shared storage, privileged simplicity
- Instead of a single redundant system, rely on multiplicity of systems
 - 50 VMs in Point 1
 - ~100 DCS
- Currently running as VM's
 - gateways
 - DCS Windows services
 - icinga
 - public Nodes

- domain controllers
- development web servers

51 VMs in GPN

~2700 Sim@P1 *

- > J puppet
- LDAP

- ATLAS K
- For new VH hardware, tested and really happy with the results of FlashCache
 - huge improvement on disk IO vs cost



Simulation @ P1



- Sim@P1 is the opportunistic use of the existing TDAQ HLT farm as a grid cluster
 - allow non-utilised resources to be exploited for ATLAS Prod jobs
- Virtual machines on top of the HLT machines, acting as computing nodes interconnected through a Virtual LAN on the data network
 - VMs and VLANs isolate the offline computing nodes network
 - no interferences with ATCN & DCS
 - security
- Communicate with the outside world (GPN) via a logically separated link to CERN IT
 - ACLs are in place for allowing only traffic towards the needed CERN services (Castor / EOS, Condor, etc.)
- More than 1300 nodes of the HLT farm are now able to run Monte Carlo, high CPU intensive jobs
 - EVGEN and SIMULATION only, by design
- Produced 1.7 billion Monte Carlo events since 1st Jan 2014
 - limited by downtimes during LS1 (cooling, power interventions etc)
- Switching between states is controlled by ATLAS control room
- For more information, see our poster on the topic
 - "Design, Results, Evolution and Status of ATLAS simulation in Point 1". Poster by Franco Brasolin

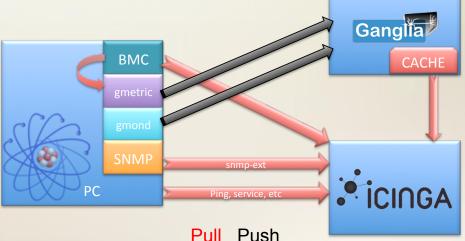




- Ganglia as collector for performance/health metrics, high scalability with rrdcache
- Ganglia-web provides advanced user interface to historical data in RRDs
- Icinga replaces Nagios
 - provides active checks and alerting
 - 🦻 can use Ganglia data
 - can reuse Nagios plugins and much of Nagios configuration

HW monitoring via IPMI

- Complete rewrite during LS1, work still in progress
- New version based on OpenIPMI, previous based on IPMItool
 - unique sensor ID's vs (unstable) sensor names
 - better performance with SDR caching
- Local readout fed to Ganglia
- ICINGA monitors SEL, Sensor OK state, specific values via Ganglia
- IPMI varies with vendor, type, version... always catching up



Monitoring & logging implementation

• icinga

- One VM for core systems:
 ~570 nodes, ~5000 checks
- One PC for farm systems:
 ~2200 nodes, ~31000 checks
- Users receiving status notifications
- Testing Icinga 2 for distributed scheduling, configuration and performance

System Logs management

- Rsyslog on all machines, also as collector for remote (replaces syslog-ng)
- Remote logging to LFS or central syslog servers
 - net boot: 2 day retention (local), 30 day on LFS
 - local boot: 30 day retention
 - exposed systems: 12 weeks
- Remote logging for security critical servers to CERN IT
- Investigating central collection & analysis tools: Splunk, ELSA, logstash + elastic search

186 OK 0/0/0 WARNING	0/0/0 UNREACHABLE 0/0/0 UNR	ONOWN 0 PENDING 3/189 TOTAL					FICINGA	0.01/10.01/0.485 s 0.00/10.00.
General E	stosg+02-00	Can nothita		OK 2013-03-68 15:32:52	7d 15h 40m 25e	10	HTTP OK: HTTP/1.1 200 OK - 569 bytes in 0.005 second response firm	
Home		net/squidport		OK 2013-03-08 15:23:34	114d 20h 56m 44s	10	TCP OK - E:001 second response time on port 3128	
Documentation 🗮 🔛		notisah		OK 2013-03-08 15:31.49	171d 1h 31m 12s	10	SSH OK - OpenSSH_5.3 (protecti 1.99)	
Search:	atiosgw02-01	📴 natihita		OK 2013-03-68 15:32:52	8d 15h 40m 25s	10	HTTP OK: HTTP/1.1 200 OK - 569 bytes in 0.006 second response firm	
		netisputport		OK 2013-03-08 15:25:45	114d 20h 48m 16a	10	TOP OK - 8:006 second response time on port 3128	
U/tv4 D		netlash		OK 2013-03-66 15:31:52	115d 3h 24m 27s	10	SSH OK - OpenSSH_5.3 (protecti 1.99)	
Tactical Overview Host Detail	at angw02-02	Carlina 🔁		OK 2013-03-06 15:32:51	8d 15h 40m 26s	10	HTTP CK: HTTP/1.1 200 CK - 569 bytes in 0.005 second response firm	
Service Detail		notiveritien		OK 2013-03-08 15-29-41	23d Oh 11m 52a	10	TCP OK - £:005 second response time on port 3128	
Hostgroup Overview Hostgroup Summary		netlash		OK 2015-03-08 15:31:56	23d Oh 14m 5a	10	SSH OK - OpenSSH_5.3 (protocol 1.99)	
Servicegroup Overview	ationgs/04-00	Contribite		OK 2013-03-08 15:22:38	35d 3h 55m 7a	10	HTTP OK: HTTP/1.1 200 OK - 569 bytes in 0.008 second response firm	
Servicegroup Summary Status Map		netispidport		OK 2013-03-08 15:25:48	35d 3h 52m 58s	10	TCP OK - 6:003 second response time on port 3128	
roblems D		netiash		OK 2013-03-08 15:29:45	10d th 45m 1a	10	SSH OK - OpenSSH 5.3 (protecti 1.99)	
Service Problems	at an ap-lb-1	gangiakou_ide	80	OK 2013-03-08 15:32.48	27d 11h 46m 27s	15	OK System check - DK (1) - OK cox, idle = 99.8 %	
Unhandled Services Host Problems	and any or t	ganglakpu_we	-	OK 2013-03-08 15:32:53	3d 20h 56m 25s	1/10	OK System check - OK (1) -OK cpu_wie = 0.9 %	
Unhandled Hosts				OK 2013-03-08 15:29:49	36d 4h 19m 8s	10	OK System check - OK: (1) - OK bpt_Me = US 5 OK System check - OK: (1) - OK part_max_used = 45.5 %	
All Unhandled Problems Network Outgoes		gangla/part_max_used		OK 2013-03-08 15:23:45	6d 9h 22m 15e	10	HTTP OK: HTTP/1.1 302 Found - 404 bytes in 0.014 second response 1	ine .
ten 1		nethtps		OK 2013-03-08 15:25:56	9d 10h 52m 51s	10	HTTP OK: HTTP/1.1 302 Found - 430 bytes in 0.112 second response 1	
Comments		netiseh		OK 2013-03-08 15:32:12	36d 4h 37m 7s	10	SSH OK - OpenSSH_4.3 (protocel 2.0)	
Downtime								
Process Info Performance Info	at as op-to-2	ide ganglaktpu_ide		OK 2013-03-08 15:32:58	27d 11h 46m 25s	16	OK System check - OK: (1) -OK cpu_idle = \$4.0 %	
Scheduling Queue		gangilaktpu_wio		OK 2013-03-08 15:32:49	46 1h 57m 11s	1/10	OK System check - DK: (1) – DK cpu_wie = 0.0 %	
porting D		gangla/part_max_used		OK 2013-03-68 15:31:00	36d 17h 34m 12s	10	OK System check - OK: (1) -OK part_max_used = 29.2 %	
Trends Availability		notihita		OK 2013-03-68 15:29:57	9d 13h 34m 10s	10	HTTP CK: HTTP/1.1 302 Found - 414 bytes in 0.004 second response t	
Averability Alert Histogram		netihtps netiseh		OK 2013-03-68 15:32:27 OK 2013-03-68 15:32:28	7d 3h 20m 50s 68d 3h 1m 0s	10	HTTP OK: HTTP/1.1 302 Found - 440 bytes in 0.025 second response to	ine
Alert History							SSH OK - OpenSSH_4.3 (protocel 2.0)	
Alert Summary Notifications	inaatds v04	Canglatov_ide		OK 2013-03-08 15:32:18	27d 11h 46m 25s	15	OK System check - OK: (1) -OK cpu_idle = 99.8 %	
Event Log		gangila/cpu_wio	1	OK 2013-03-08 15:32:12	3d 11h 29m 11s	1/10	OK System check - DK: (1) - DK opu_wia = 1.4 %	
nfiguration a		gangla/part_max_used		OK 2013-03-08 15:30:05	36d 22h 20m 37s	10	OK System check - OK: (1) -OK part_max_used = 59.4 %	
View Config		(pm/ping		OK 2013-03-08 14:43:57	65d 3h 57m 34s	10	FPING OK - insetdor/04-mgmt (cosi-0%, rta=8.730000 me)	
		netidheperv	_ ₽ ₿(SITICAL 2013-03-08 15:29:13	3d 20h 57m 10s	32	CRITICAL: No DHCPOFFERs were received.	
		notisah		OK 2013-03-68 15:32:26	65d Oh 7m 23a	10	SSH OK - OpenSSH_4.3 (protocsi 2.0)	
	localhost	Current Users		OK 2013-03-68 15:30:05	189d 7h 15m 10s		USERS CK - 0 users currently logged in	
		Icinga Startup Delay		OK 2013-03-08 15:29:01	189d 7h 4m 54a	54	OK licings started with 1 seconds delay	
		Root Partition		OK 2013-03-08 15:31:11	189d 7h 15m 10s	14	DISK CK - free space: / 19901 MB (91% inde=99%):	
		Swep Usage		OK 2013-03-08 15:30:11	189d 7h 15m 10s	14	SIMAP OK - 100% free (\$189 MB out of \$199 MB)	
		Tatal Processes		OK 2013-03-08 15:29:05	36d 15h 17m 51a	5/4	PROCS OK: 148 processes with STATE = RS2DT	
		notiash		OK 2015-03-08 15:32:30	189d 7h 7m 48s	10	SSH OK - OpenSSH_4.3 (protacel 2.0)	
	po-atlas-origon-02	🥜 👎 📴 netissh		XITCAL 2013-03-08 15:28:15	79d 8h 30m 40s	13	CRITICAL - Socket timeout after 10 seconds	
	polatas-origon-04	🖉 🛡 🗁 netissh		SITICAL 2015-03-08 15:30:12	94 5h 24m 28s	18	CRITICAL - Socket timeout after 10 seconds	
	po-atas-mon-01	(cm/ping		OK 2013-03-08 15:22:35	54d 16h 52m 56e	10	FPING OK - pc-atlas-mon-01-mpmt (oss=0%, rta=58 700000 ms)	
	PC-000-00-001	Contract of the second					(The on-position of High (our os, niess (0000 High	

Summary

- LS1 was anything but a "shutdown" for our Team
- We have streamlined and improved the Point1 system
 - many more tasks are now fully automated, with very little human intervention needed
- Puppet, cleaner and easier to maintain
- Monitoring much more comprehensive than before
 - provides many more checks, still rapidly evolving
- Still Investigating Open LDAP (2.4.39) issues with opening/closing connections.
- Hoping Run 2 will provide us some "quiet time" to clean up



Have you tried turning it off and on again?

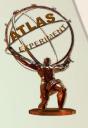
Have you tried reconfiguring the primary power coupling?

Glossary

- ATCN: Atlas Technical and Control Network
- CFS: Central File Server
- DCS: Detector Control Systems
- GPN: General Public Network
- HLT: High Level Trigger
- LFS: Local File Server
- LHC: Large Hadron Collider
- LS1: Long Shutdown 1
- NAS: Network Attached Storage
- PXE: Preboot eXecution Environment
- ROS: Read Out System
- SBC: Single Board Computer
- SLC: Scientific Linux: Cern edition
- SVN: Subversion
- TDAQ: Trigger and Data Acquisition
- WLCG: Worldwide LHC Computing Grid



BACKUPS / SPARE



Introduction LHC & ATLAS

- Large Hadron Collider, an accelerator, ~100m underground
- 27 km in circumference
- Protons are accelerated in opposite directions at 4 TeV
- Smashing together in the center of ATLAS, one of 7 experiments
- 600 million collisions per second
- Data about these collisions are recorded by the Trigger and Data Acquisition system



erall view of the LHC experiments

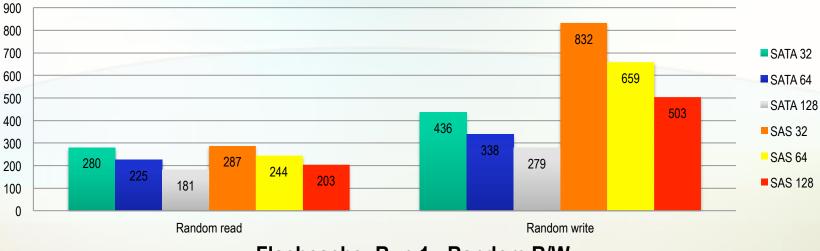
"Private networks" now managed by IT

- Private networks for us:
 - isolated Networks within the ATCN
 - high voltage power supplies
 - ATCA / VME crates
- Use of unregistered private networks is against CERN IT security policies
- Unmanaged switches are not supported anymore (= no spares)
- Integration with CERN IT
 - standardised system
 - no RFC1918 networks
 - all devices on the network registered and traceable
 - increased security
 - increased management
 - > 4 hour piquet-like support from CERN IT

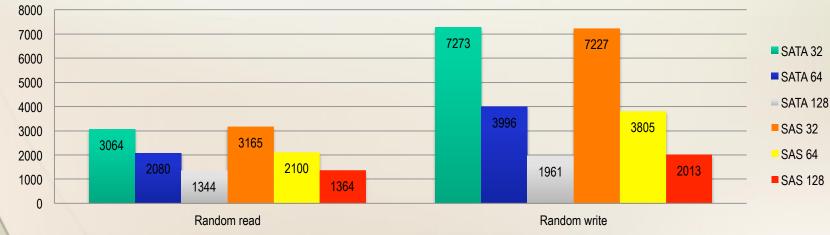


FlashCache tests

Normal - RAID5 - Random R/W



Flachcache -Run 1 - Random R/W



CHEP April 2015

C.J.Lee - UJ, CERN - ATLAS TDAQ SysAdmin

Sim@p1 -> TDAQ Mode

- Switching from control room shifter through a WEB gui when a LHC beam stop longer than 24h is foreseen
- Fast and automated:
 - from sim@p1 to TDAQ ~ 12 min
 - from TDAQ to sim@p1 ~ 1h
 - emergency switch from sim@p1 to TDAQ: 100 s



Monitoring system upgrade

Nagios v2 + Custom UI

- Old production system: stable, but complex
- Multiple standalone Nagios servers
- Central configuration from ConfDB
- Central storage in MySQL/cluster and RRDs
- High I/O load on MySQL and NetApp can become a bottleneck
- Custom WebUI requires maintenance
- Nagios v2 is obsolete

Nagios v3 + Custom UI

- With the LFS SLC6 migration complete, this was put into production
- One single MySQL server replaced the 4 machines cluster used for storing Nagios data
 - better performance
 - easier maintenance

Atlas Control Room

- Completion of the plan for PCoverIP migration (Desk remote technology over TCP/IP)
 - KVM (keyboard, video, mouse) from SDX1 to ACR over network
 - each machine has 1 or 2 PCoverIP cards (depending on number of screens)
 - each desk has 1 terminal client
- A joint collaboration between OPM and TDAQ NetAdmins+SysAdmins
- Full redundancy
 - > 2 switches in SDX1 and 2 switches in SCX1.
 - Cards and Terminals have dual connections
- Less clutter
 - 2 optical fibers between SDX1 and SCX1, providing two independent connections, replaced ~100 copper cables
- Delayed updating existing systems as current available market hardware provides no redundancy and no major improvements



Satellite Control Room

- Experts base of operations
 - provides similar workspaces as in ACR
- Advanced debug tools allowed
- extensive use of IMACS and Mac MINI running SLC6
 - allowed for long lasting, system that can be replaced

CRD (KDE)

- Control Room Desktop
- Provides tools required per "seat" in the ACR
- No direct terminal access
 - controlled and authenticated access to terminal windows
- This implementation needed a version of KDE not available in SLC6

22

