

ATLAS TDAQ System Administration: an overview and evolution

ISGC 2013

Christopher Jon Lee

*University of Johannesburg, South Africa
CERN*

*for and on behalf of the
ATLAS TDAQ SysAdmin team.*

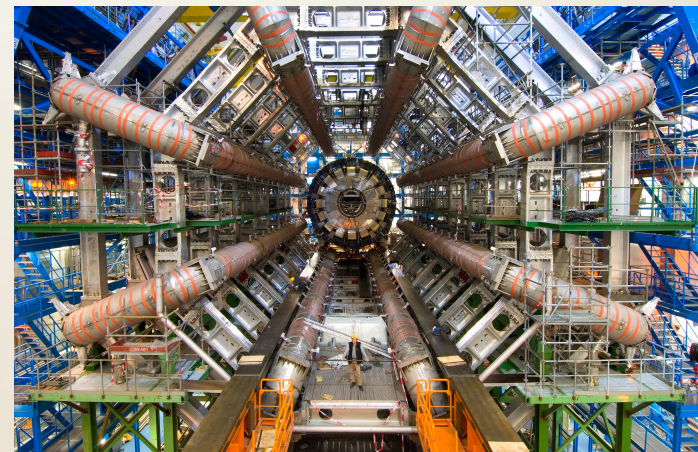
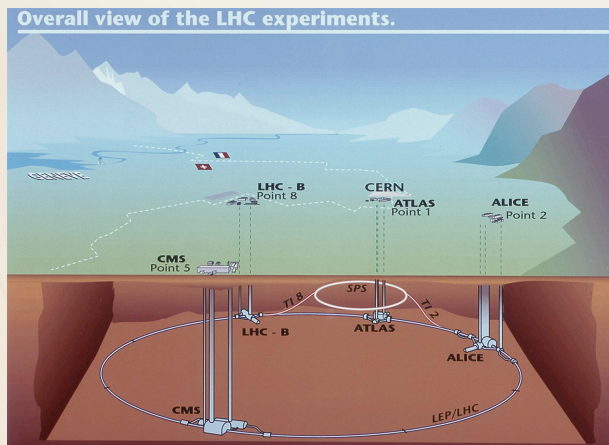


UNIVERSITY
OF
JOHANNESBURG

Introduction

LHC & ATLAS

- ◆ Large Hadron Collider, an accelerator, ~100 m underground
- ◆ 27 kilometres in circumference
- ◆ Hadrons 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 from these collisions are recorded by the **Trigger and Data Acquisition system**
- ◆ After 3 years of LHC runs, Long Shutdown 1 (LS1) has started



Introduction

Trigger and Data Acquisition

- ◆ Large online computing farm, used to process the data readout from ~100 million channels
- ◆ Ancillary functions (monitoring, control, etc.)
- ◆ ATLAS Point 1 – Counting Rooms
 - ◆ approximately 100 m underground, in close proximity to the detector (USA15)
 - ◆ on the surface near to the ATLAS Control Centre (P1, SDX1 & SCR)
- ◆ In General Public Network (GPN)
 - ◆ laboratory for software development, prior to implementation into P1, recently been commissioned (TestBed)



The Racks

- ◆ USA15:
 - ◆ 220 racks deployed over 2 floors
 - ◆ 2009, ~70% filled, 1MW
 - ◆ 2013, > 90% filled, 1.29MW*
 - ◆ 2.5MW of cooling can be provided
- ◆ SDX1:
 - ◆ 120 racks deployed over 2 floors
 - ◆ 2009, ~50% filled, 385 kW*
 - ◆ 2013, ~91% filled, 709 kW*
- ◆ TestBed:
 - ◆ 22 racks over 1 floor
 - ◆ 68% filled, 60kW of power. 11 Racks by TDAQ
 - ◆ 100 kW of Cooling can be provided
- ◆ Possibly an increase in hosts to deal with future requirements

*estimated at ~6.5 kW/rack

SDX1 :: LEVEL 2 :: Rows 2 & 3

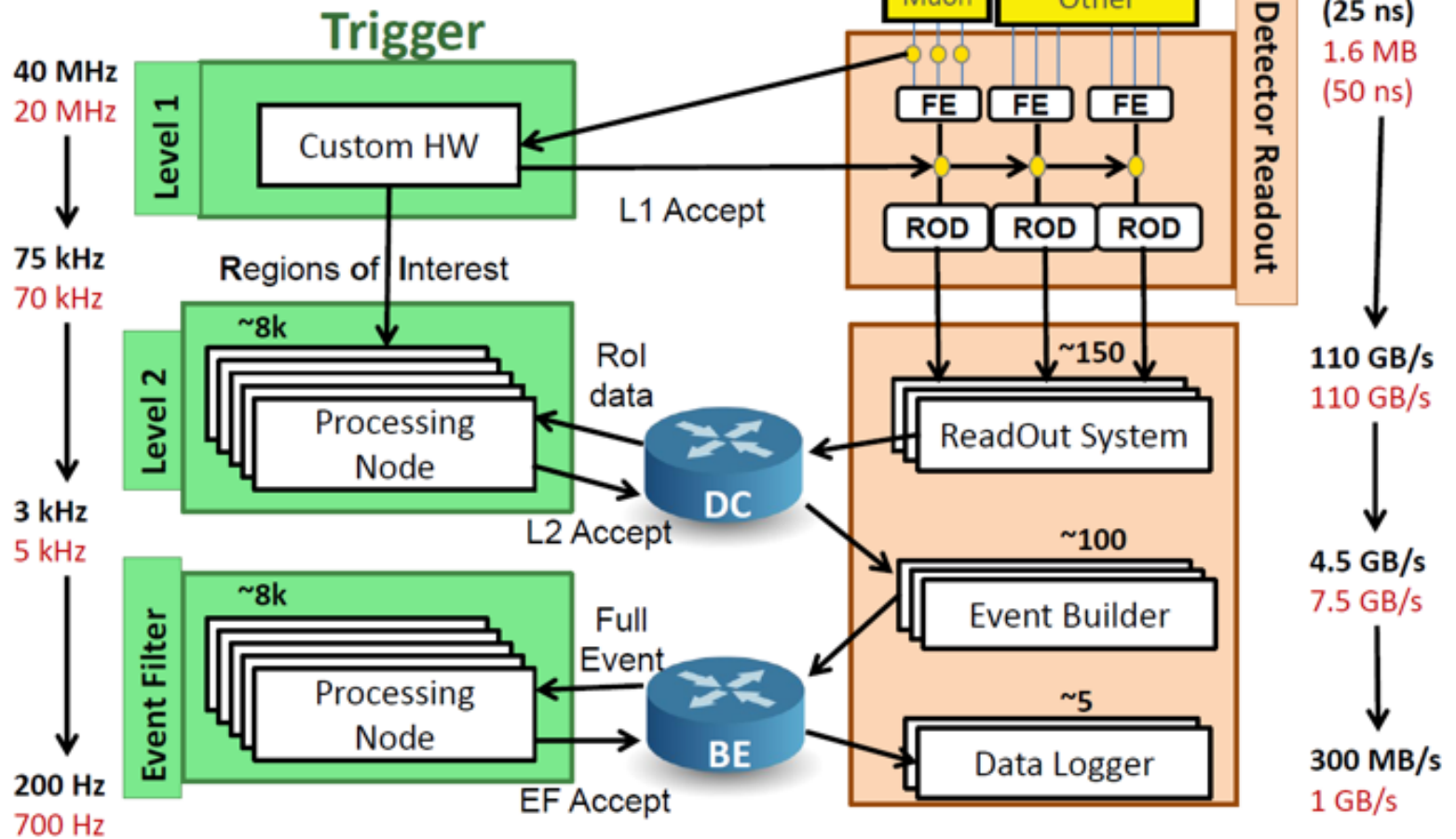
C.J.Lee - UJ, CERN - ATLAS TDAQ SysAdmin

TDAQ Design Overview

TDAQ Today

Design
(2012 - avg)

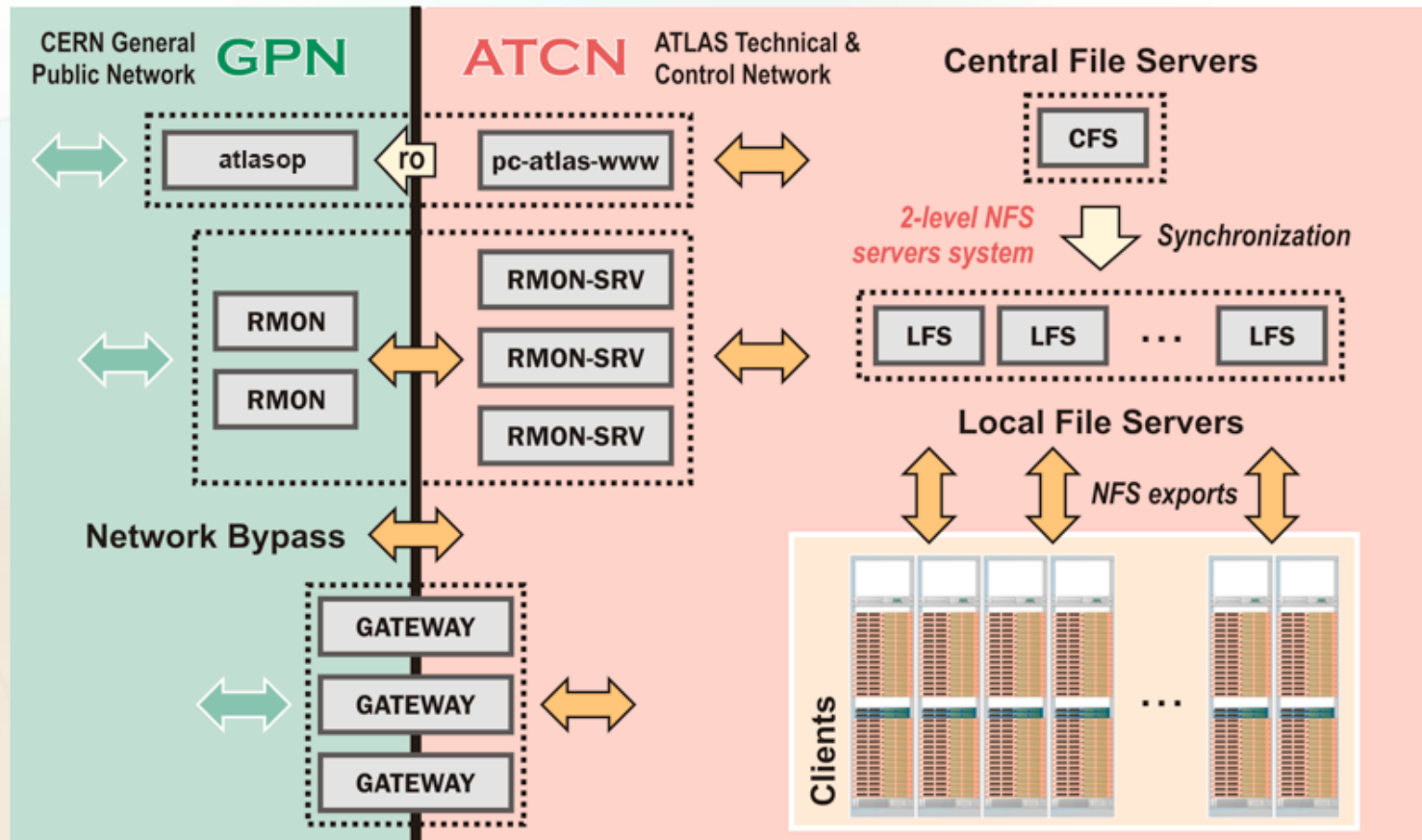
DAQ



* Image made by Nicoletta Garelli - Athens Trigger Workshop



ATLAS Point1 Functional Layout



Redundancy

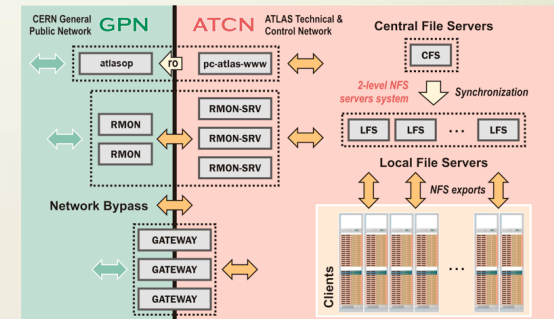
- ◆ Can't afford downtime & miss out on events provided by the LHC
- ◆ Two centralised UPS lines with diesel backup generators
- ◆ Independent UPS lines available to SDX1 for mission critical equipment
- ◆ Redundant configuration for critical services
 - ◆ DNS/DHCP/NTP
 - ◆ LDAP
 - ◆ Domain Controller
 - ◆ LFS'
- ◆ Network Attached Storage, serves the most critical NFS and CIFS areas
 - ◆ NetApp 3140, it has 12 TB over 84 HDD's in RAID DP, dual head system and redundant fibre channels
 - ◆ can survive the loss of one head, any one FC link, one FC interface on a shelf - just not of a whole disk shelf
- ◆ CERN's Tivoli for critical data
- ◆ Subversion for code and configurations



Centralised & Local Storage Systems

- ◆ CFS Linux node - no direct exports:
 - ◆ Trigger/DAQ and Offline Software installation
 - ◆ Coordination of synchronisations to LFS'
- ◆ NetApp Central Filer - NFS and selected CIFS exports :
 - ◆ user home directories, DAQ software distribution
 - ◆ Nagios RRD* files
 - ◆ node configuration area
 - ◆ DAQ configuration area
 - ◆ static and dynamic Web content
 - ◆ dedicated file exchange area for gateways
- ◆ LFS
 - ◆ provides boot services and NFS exports for clients
 - ◆ synced from NetApp

ATLAS Point1 Functional Layout



Clients

Local Boot

- ◆ Provisioning by PXE + Kickstart
 - ◆ DHCP + PXE provided by an LFS from ConfDB* information
 - ◆ template-based kickstart files
- ◆ Quattor
 - ◆ CERN standard Configuration Management Tool
 - ◆ production system, managing 237 hosts in the Online Farm
 - ◆ tight control on installed packages
 - ◆ lack of flexibility for complex configuration/services dependencies
 - ◆ multiple languages for implementing modules

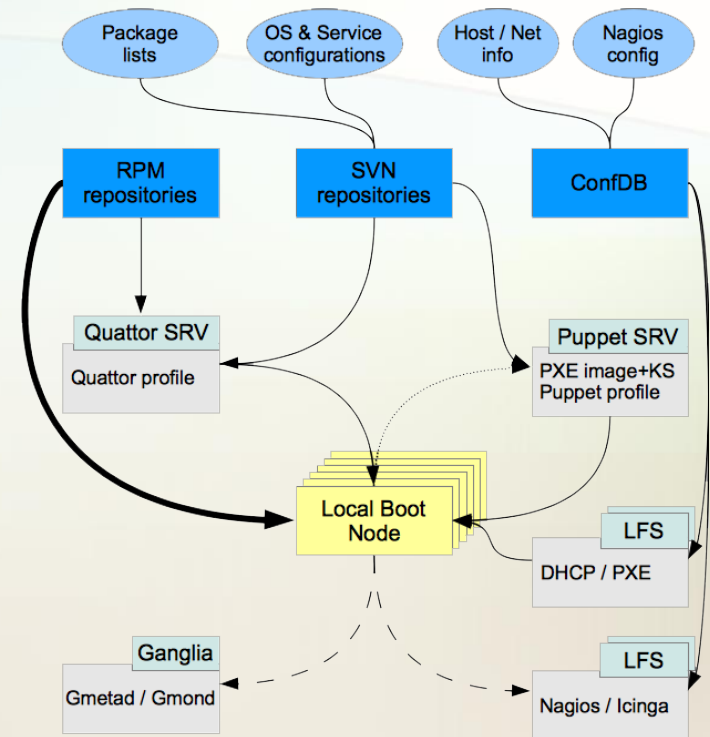
*See slide 14



Clients

The Puppet Master

- ◆ Puppet
 - ◆ widespread industry adoption, active development
 - ◆ full featured, highly flexible
 - ◆ focus on consistency and idempotency*
 - ◆ gentler learning curve
 - ◆ In production:
 - ◆ 25 complex servers in Point 1
 - ◆ entire Testbed
 - ◆ Complements Quattor
 - ◆ Migrate to Puppet and SLC6
 - ◆ our manifest code base has grown to ~15000 LOC
 - ◆ Puppet is being adopted by CERN IT and the other experiments

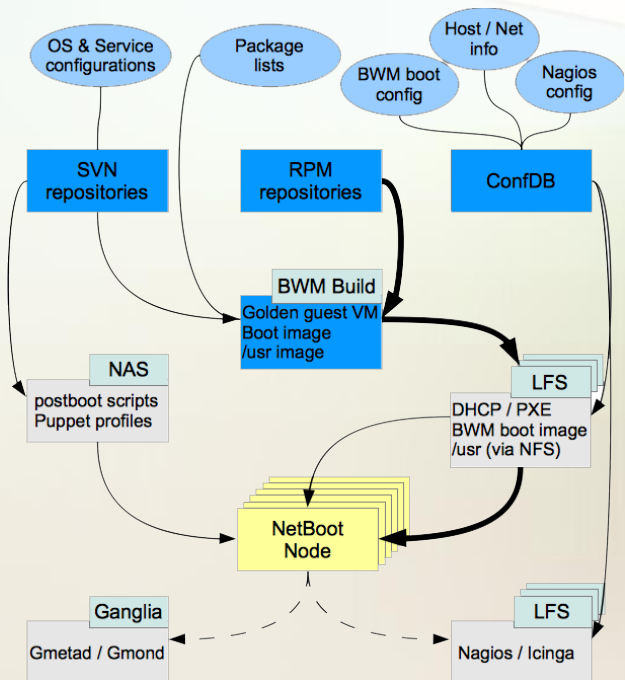


*unchanged in value following multiplication by itself;



Look Mom ... No disk

- ◆ 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
- ◆ “NetBooted” via PXE
- ◆ Advantages:
 - ◆ ease of maintenance
 - ◆ reproducibility on a large scale
 - ◆ reduced "installation" times
- ◆ Disadvantages:
 - ◆ requires ad-hoc development and support
 - ◆ not suitable for running servers



Clients

Netbooted Nodes

- ◆ ~ 2350 Nodes booted Scientific Linux CERN5 OS via PXE
- ◆ 75 Local File Servers (LFS), provide DHCP, PXE and TFTP for booting, /usr read-only directory via NFS
- ◆ Configuration of DHCP, PXE and boot parameters are provided by our in house built ConfDB*
- ◆ Boot With Me Tool (BWM)
 - ◆ generates PXE boot images (kernel + RAMdisk root) and /usr
 - ◆ uses reference SLC5 VM image as source
- ◆ BWM post-boot script system
 - ◆ hierarchy of shell scripts, configures servers, disk, and NFS mounts
 - ◆ store on a central Network-Attached Storage, executed by the client
- ◆ Subversion is used to keep track of image configurations and post-boot scripts

*See slide 14



Clients

Puppet again...

- ◆ Problems with BWM post-boot
 - ◆ no permanent changes to a working system
 - ◆ needs to be rebooted to apply configuration changes
 - ◆ bash scripts not easily maintainable
- ◆ Move net-booted to Puppet as well
 - ◆ manifests hosted on central NFS
 - ◆ **no** server
 - ◆ no daemon, run via cron job
 - ◆ no need to reboot node for changes to take affect
- ◆ Minor problems on low resource machines and TDAQ /ROS drivers



Tools

Configuration Database

- ◆ Management of large number of NetBooted nodes is far from trivial
- ◆ ConfDB was developed
 - ◆ python backend, using a MySQL database
 - ◆ web based GUI
- ◆ Keeps record of the base system configurations
- ◆ Greatly help to speed up routine tasks:
 - ◆ registering clients, with data extracted from CERN's network DataBase (LanDB)
 - ◆ configuring kernel boot options
 - ◆ client to LFS assignment and migration
 - ◆ deploying DHCP and NAGIOS configurations
 - ◆ issuing IPMI and system commands on multiple nodes in parallel



Tools ConfDB UI

Cluster: All

ConfDB GUI ADMINISTRATIVE INTERFACE

- 🔍 **Devices**
 - » Edit Devices
 - » Move Clients
 - » Add Devices
- 🔍 **Deployment**
 - » DHCP
 - » NAGIOS
- 🔍 **Operations**
 - » IPMI commands
 - » SSH commands
- 🔍 **Boot Images / OS**
 - » Boot Images List
 - » Add Boot Image
 - » Boot Options List
 - » Boot Option Add
- 🔍 **Nagios**
 - » Services List
 - » Service Add
 - » Templates List
 - » Template Add
 - » Users List
 - » User Add
 - » Groups List
 - » Group Add
- 🔍 **SEL**
 - » History
- 🔍 **Maintenance**
 - » Maint. Operations

Hostname: xpu-66 Host type: Clients 🔍 Search + Advanced search 📄 Commit changes

Search results:

- pc-tdq-xpu-66001
- pc-tdq-xpu-66002
- pc-tdq-xpu-66003
- pc-tdq-xpu-66004
- pc-tdq-xpu-66005
- pc-tdq-xpu-66006
- pc-tdq-xpu-66007
- pc-tdq-xpu-66008
- pc-tdq-xpu-66009
- pc-tdq-xpu-66010
- pc-tdq-xpu-66011
- pc-tdq-xpu-66012
- pc-tdq-xpu-66013
- pc-tdq-xpu-66014
- pc-tdq-xpu-66015
- pc-tdq-xpu-66016
- pc-tdq-xpu-66017
- pc-tdq-xpu-66018
- pc-tdq-xpu-66019
- pc-tdq-xpu-66020
- pc-tdq-xpu-66021
- pc-tdq-xpu-66022
- pc-tdq-xpu-66023
- pc-tdq-xpu-66024
- pc-tdq-xpu-66025
- pc-tdq-xpu-66026
- pc-tdq-xpu-66027
- pc-tdq-xpu-66028
- pc-tdq-xpu-66029
- pc-tdq-xpu-66030
- pc-tdq-xpu-66031
- pc-tdq-xpu-66032
- pc-tdq-xpu-66033
- pc-tdq-xpu-66034
- pc-tdq-xpu-66035
- pc-tdq-xpu-66036
- pc-tdq-xpu-66037
- pc-tdq-xpu-66038
- pc-tdq-xpu-66039
- pc-tdq-xpu-66040

Hostname: pc-tdq-xpu-66015 LanDB Nagios Hardware DB

MACs: 00-26-6C-FA-C6-F0, 00-26-6C-FA-C6-F1, 00-26-6C-FA-C6-F3 **Manufacturer:** DELL **Model:** POWEREDGE C6100

Rack: Y.08-04.D1 [66] **Position in Rack:** U18 **Building:** 3178 **Floor:** 1W **Room:** 0804 **Host Group:** Point 1

Service Tag: **OS Version:** Net_SLC5_64 **Description:**

Ipmi Type: ipmi20 **Nagios Server:** pc-tdq-lfs-066 **Config server:** Choose...

BMC Specification: 37_1_1.26_2.0_20569_55 Include in DHCP relay list Don't include in DHCP Sync Nagios home directory Net Booted

PC Type: pc **Netboot Server:** pc-tdq-lfs-066

Boot parameters: Open

» root=/dev/ram0 ramdisk=131072 ip=dhcp selinux=0

NICs

» Type: control, Name: pc-tdq-xpu-66015, IP: 10.146.95.45, MAC: 00-26-6C-FA-C6-F0, Netmask: 255.255.255.0, Gateway: 10.146.95.1, Network domain: ATLAS

» Type: dc2, Name: pc-tdq-xpu-66015-dc2, Alias: pc-tdq-xpu-66015-ef2-vlan12, IP: 10.150.60.29, Netmask: 255.255.0.0, Gateway: 10.150.1.1, Network domain: ATLAS, Vlan ID: 12

» Type: ef2, Name: pc-tdq-xpu-66015-ef2, IP: 10.151.43.49, MAC: 00-26-6C-FA-C6-F1, Netmask: 255.255.255.0, Gateway: 10.151.43.1, Network domain: ATLAS


» Type: mgmt, Name: pc-tdq-xpu-66015-mgmt, IP: 10.146.95.44, MAC: 00-26-6C-FA-C6-F3, Netmask: 255.255.255.0, Gateway: 10.146.95.1, Network domain: ATLAS

Templates: Open

» BASIC-XPU

» INTERFACE_UP!to,ctrl0,ef2,vlan12"

✖ Delete host(s) 🔄 Update host info from LanDB 📄 Commit changes



ATLAS
EXPERIMENT

22/02/2013

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

15

Tools

OS Repository Management

- ◆ Simple custom-made system for managing "time-frozen" snapshots of CERN package repositories
- ◆ Sufficient functionality through 2012, controlled upgrades, (theoretical) rollback capability
- ◆ We want more flexible functionality and easier management
 - ◆ partial upgrades, client status reporting etc.
 - ◆ may adopt a third-party open source tool, e.g. Pulp
- ◆ Rollback/versionlock are possible in principle but not easy: Puppet +yum does not offer the same detailed control as Quattor+SPMA



Monitoring Nagios

- ◆ Nagios has been used since 2007
- ◆ Primarily monitors:
 - ◆ the health status of the OS
 - ◆ the hardware, selected services and network components
 - ◆ provides alerting for critical events
- ◆ Separate Nagios server instance on each of the LFS nodes
- ◆ Feeding data to a central RRD* storage and to a single MySQL Cluster
- ◆ A web-based interface was developed in-house
- ◆ Significant development effort into hardware status', via IPMI
- ◆ Configurations integrated in ConfDB
- ◆ Over 5 years, machines monitored by NAGIOS increased from ~1500 to ~3000 in Point 1

* *round-robin database*



Monitoring

| GROUPS | TOTAL | ONLINE | OFFLINE | BROKEN | RESERVED |
|---------------|-------------|-------------|-----------|-----------|-----------|
| ⊕ Gateways | 6 | 6 | 0 | 0 | 0 |
| ⊕ WebServers | 2 | 2 | 0 | 0 | 0 |
| ⊕ FileServers | 2 | 2 | 0 | 0 | 0 |
| ⊕ DNS | 2 | 2 | 0 | 0 | 0 |
| ⊕ CFS | 1 | 1 | 0 | 0 | 0 |
| ⊕ LDAP | 4 | 4 | 0 | 0 | 0 |
| ⊕ MYSQL | 5 | 5 | 0 | 0 | 0 |
| ⊕ VH | 2 | 2 | 0 | 0 | 0 |
| ⊕ ACR | 128 | 119 | 0 | 0 | 9 |
| ⊕ SCR | 49 | 33 | 16 | 0 | 0 |
| ⊕ TDQ | 2186 | 2168 | 2 | 9 | 7 |
| ⊕ LFS | 74 | 73 | 0 | 0 | 1 |
| ⊕ ONL | 33 | 33 | 0 | 0 | 0 |
| ⊕ AMS | 7 | 7 | 0 | 0 | 0 |
| ⊕ MON | 32 | 32 | 0 | 0 | 0 |
| ⊕ GMON | 6 | 5 | 0 | 1 | 0 |
| ⊕ ROS | 157 | 156 | 1 | 0 | 0 |
| ⊕ SFI | 48 | 48 | 0 | 0 | 0 |
| ⊕ SFO | 9 | 9 | 0 | 0 | 0 |
| ⊕ DC | 12 | 12 | 0 | 0 | 0 |
| ⊕ L2SV | 8 | 8 | 0 | 0 | 0 |
| ⊕ XPU | 1208 | 1196 | 0 | 8 | 4 |
| ⊕ EF | 448 | 448 | 0 | 0 | 0 |
| ⊕ PRESERIES | 131 | 129 | 1 | 0 | 1 |
| ⊕ RMON SRVs | 3 | 3 | 0 | 0 | 0 |
| ⊕ NET-MON | 7 | 7 | 0 | 0 | 0 |
| ⊕ SYS | 3 | 2 | 0 | 0 | 1 |
| ⊕ SBC | 161 | 153 | 5 | 2 | 1 |
| ⊕ PUB | 14 | 12 | 2 | 0 | 0 |
| ⊕ DCS | 102 | 98 | 4 | 0 | 0 |
| ⊕ MU-CALSRV | 2 | 2 | 0 | 0 | 0 |
| ⊕ SWITCH | 100 | 100 | 0 | 0 | 0 |
| ⊕ OTHERS | 156 | 132 | 21 | 0 | 3 |
| TOTAL | 2926 | 2841 | 50 | 11 | 20 |



Monitoring Ganglia

- ◆ 2011: we introduced Ganglia for performance monitoring and trending
- ◆ We are currently evaluating on a smaller scale system
- ◆ Used primarily for special purpose nodes
 - ◆ local boot nodes
 - ◆ monitoring, Online and Output (SFO) sub-farms
- ◆ Single central server
- ◆ RRD* caching daemon used for I/O performance
- ◆ Ganglia is used by many Grid facilities and will be by LHCb as well
- ◆ Integration with Icinga

* *round-robin database*

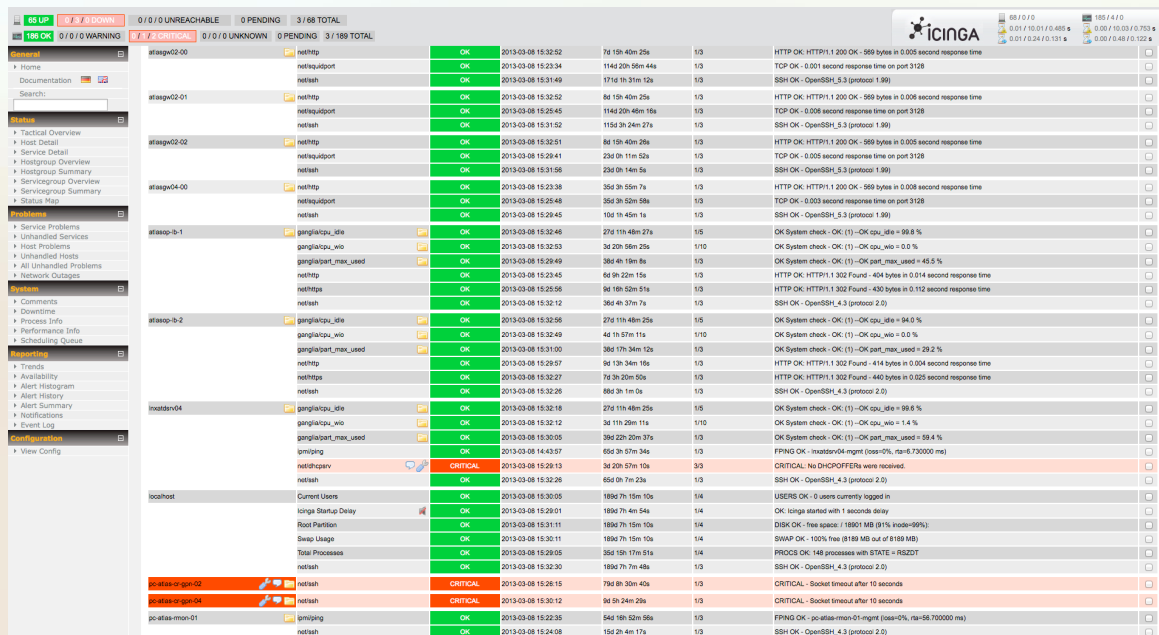




ICINGA and Gearman

Monitoring

- ◆ Icinga+Gearman provide active checks with distributed scheduling
- ◆ Alerting
- ◆ Icinga+Gearman adopted by CMS and LHCb
- ◆ Icinga can reuse Nagios plugins, and much of Nagios configuration



| Host | Service | Status | Last Check | Duration | Attempts | Output |
|----------------|------------------------------|----------|---------------------|------------------|----------|---|
| atstgw02-00 | nethttp | OK | 2013-03-08 15:32:52 | 7s 19m 45m 25s | 1/3 | HTTP OK: HTTP/1.1 200 OK - 989 bytes in 0.005 second response time |
| | nettcp | OK | 2013-03-08 15:23:34 | 114d 20h 56m 44s | 1/3 | TCP OK - 0.001 second response time on port 3128 |
| | netssh | OK | 2013-03-08 15:31:49 | 171s 1h 31m 12s | 1/3 | SSH OK - OpenSSH_5.3 (protocol 1.99) |
| atstgw02-01 | nethttp | OK | 2013-03-08 15:32:52 | 8d 15h 40m 25s | 1/3 | HTTP OK: HTTP/1.1 200 OK - 989 bytes in 0.006 second response time |
| | nettcp | OK | 2013-03-08 15:25:45 | 114d 20h 46m 16s | 1/3 | TCP OK - 0.006 second response time on port 3128 |
| | netssh | OK | 2013-03-08 15:31:52 | 115d 3h 24m 27s | 1/3 | SSH OK - OpenSSH_5.3 (protocol 1.99) |
| atstgw02-02 | nethttp | OK | 2013-03-08 15:32:51 | 8d 15h 40m 26s | 1/3 | HTTP OK: HTTP/1.1 200 OK - 989 bytes in 0.005 second response time |
| | nettcp | OK | 2013-03-08 15:29:41 | 23d 0h 11m 52s | 1/3 | TCP OK - 0.005 second response time on port 3128 |
| | netssh | OK | 2013-03-08 15:31:56 | 23d 0h 14m 5s | 1/3 | SSH OK - OpenSSH_5.3 (protocol 1.99) |
| atstgw04-00 | nethttp | OK | 2013-03-08 15:23:38 | 35d 3h 55m 7s | 1/3 | HTTP OK: HTTP/1.1 200 OK - 989 bytes in 0.008 second response time |
| | nettcp | OK | 2013-03-08 15:25:48 | 35d 3h 52m 58s | 1/3 | TCP OK - 0.003 second response time on port 3128 |
| | netssh | OK | 2013-03-08 15:28:45 | 10s 1h 46m 1s | 1/3 | SSH OK - OpenSSH_5.3 (protocol 1.99) |
| atstgw-0-1 | ganglia_cpu_util | OK | 2013-03-08 15:32:48 | 27d 11h 48m 37s | 1/5 | OK System check - OK (1) - OK cpu_util = 88.8 % |
| | ganglia_cpu_wio | OK | 2013-03-08 15:32:53 | 3d 20h 58m 25s | 1/10 | OK System check - OK (1) - OK cpu_wio = 0.0 % |
| | ganglia_part_max_used | OK | 2013-03-08 15:29:49 | 26d 4h 19m 8s | 1/3 | OK System check - OK (1) - OK part_max_used = 45.5 % |
| atstgw-0-2 | nethttp | OK | 2013-03-08 15:23:45 | 6d 9h 22m 15s | 1/3 | HTTP OK: HTTP/1.1 302 Found - 404 bytes in 0.014 second response time |
| | nettcp | OK | 2013-03-08 15:25:56 | 9d 18h 52m 51s | 1/3 | HTTP OK: HTTP/1.1 302 Found - 430 bytes in 0.112 second response time |
| | netssh | OK | 2013-03-08 15:32:12 | 36d 4h 37m 7s | 1/3 | SSH OK - OpenSSH_4.3 (protocol 2.0) |
| instantiate04 | ganglia_cpu_util | OK | 2013-03-08 15:32:56 | 27d 11h 48m 25s | 1/5 | OK System check - OK (1) - OK cpu_util = 84.0 % |
| | ganglia_cpu_wio | OK | 2013-03-08 15:32:49 | 4d 1h 57m 11s | 1/10 | OK System check - OK (1) - OK cpu_wio = 0.0 % |
| | ganglia_part_max_used | OK | 2013-03-08 15:31:00 | 36d 17h 34m 12s | 1/3 | OK System check - OK (1) - OK part_max_used = 29.2 % |
| instantiate04 | nethttp | OK | 2013-03-08 15:29:07 | 8d 13h 34m 18s | 1/3 | HTTP OK: HTTP/1.1 302 Found - 414 bytes in 0.004 second response time |
| | nettcp | OK | 2013-03-08 15:32:27 | 7d 3h 20m 50s | 1/3 | HTTP OK: HTTP/1.1 302 Found - 440 bytes in 0.025 second response time |
| | netssh | OK | 2013-03-08 15:32:28 | 86d 3h 1m 5s | 1/3 | SSH OK - OpenSSH_4.3 (protocol 2.0) |
| pcatlas-mon-01 | ganglia_cpu_util | OK | 2013-03-08 15:32:18 | 27d 11h 48m 25s | 1/5 | OK System check - OK (1) - OK cpu_util = 88.8 % |
| | ganglia_cpu_wio | OK | 2013-03-08 15:32:13 | 3d 11h 26m 11s | 1/10 | OK System check - OK (1) - OK cpu_wio = 1.4 % |
| | ganglia_part_max_used | OK | 2013-03-08 15:30:05 | 26d 22h 20m 37s | 1/3 | OK System check - OK (1) - OK part_max_used = 58.4 % |
| instantiate04 | instantiate04-mgmt (over-0%) | OK | 2013-03-08 14:43:07 | 65d 3h 57m 34s | 1/3 | FFMPEG OK - instantiate04-mgmt (over-0%, run=6.730000 ms) |
| | net@qserv | CRITICAL | 2013-03-08 15:29:13 | 3d 20h 57m 10s | 3/3 | CRITICAL: No DHCP OFFERS were received. |
| | netssh | OK | 2013-03-08 15:32:28 | 65d 0h 7m 23s | 1/3 | SSH OK - OpenSSH_4.3 (protocol 2.0) |
| instantiate04 | Current Users | OK | 2013-03-08 15:30:05 | 189d 7h 15m 10s | 1/4 | USERS OK - 0 users currently logged in |
| | login Start-Up Delay | OK | 2013-03-08 15:29:01 | 189d 7h 4m 54s | 1/4 | OK: login started with 1 seconds delay |
| | Root Partition | OK | 2013-03-08 15:31:11 | 189d 7h 15m 10s | 1/4 | DISK OK - free space / 18901 MB (91% inode=90%): |
| instantiate04 | Swap Usage | OK | 2013-03-08 15:30:11 | 189d 7h 15m 10s | 1/4 | SWAP OK - 100% free (8189 MB out of 8189 MB) |
| | Top Processes | OK | 2013-03-08 15:29:05 | 25d 18h 17m 51s | 1/4 | PROC_OK: 148 processes with STATE=ISIDZDT |
| | netssh | OK | 2013-03-08 15:32:20 | 189d 7h 4m | 1/3 | SSH OK - OpenSSH_4.3 (protocol 2.0) |
| instantiate04 | netssh | CRITICAL | 2013-03-08 15:28:15 | 79d 8h 30m 45s | 1/3 | CRITICAL: - Socket timeout after 10 seconds |
| | netssh | CRITICAL | 2013-03-08 15:30:12 | 9d 5h 24m 29s | 1/3 | CRITICAL: - Socket timeout after 10 seconds |
| | instantiate04-mgmt | OK | 2013-03-08 15:22:35 | 84d 18h 52m 56s | 1/3 | FFMPEG OK - pcatlas-mon-01-mgmt (inuse=0%, run=58.700000 ms) |
| instantiate04 | netssh | OK | 2013-03-08 15:24:08 | 15d 2h 4m 17s | 1/3 | SSH OK - OpenSSH_4.3 (protocol 2.0) |



Security

Remote Access Subsystems

- ◆ Security at Point 1 is of utmost importance
- ◆ Access to the ATLAS Technical and Control Network (ATCN) is highly restricted
- ◆ Only allowed via one of the following gateway systems:
 - ◆ ATLAS Point 1, allowing expert users access to their restricted machines via SSH or SCP protocols
 - ◆ ATLAS Remote Monitoring System provides the graphical terminal services required for organising the remote participation in the ATLAS sub-detector monitoring shifts
 - ◆ ATLAS DCS Windows Terminal Servers, allowing experts access to the Detector Control System
- ◆ Host and network based accounting, security monitoring and intrusion prevention systems are configured on all the gateways



Security



Role Based Access Control

- ◆ Own user database in the form of an LDAP server based on OpenLDAP software
- ◆ Standalone but for consistency it is synchronised with CERN IT
- ◆ Slave Windows Domain Controller using the CERN NICE credentials.
- ◆ Local service accounts, authentication (passwords) in LDAP
- ◆ User based authentication, NOT group based authentication
- ◆ Role Based Access Control (RBAC) authorisation system
- ◆ ~360 unique roles in a hierarchical structure and ~4250 Users
- ◆ However..
 - ◆ shifter and expert roles automatically enabled/disabled for the person on shift / on call
 - ◆ time-limited remote access ONLY on case-by-case authorisation by Shift Leader



Virtualisation

- ◆ Adopted for specific services, not planning to use for the full scale farm
- ◆ Machines currently running as VM's
 - ◆ Gateways
 - ◆ Domain Controllers
 - ◆ Few Windows services
 - ◆ Development Web servers
 - ◆ Core Nagios servers
 - ◆ Puppet & Quattor
 - ◆ Public Nodes
- ◆ Xen on Gateways (Bastion Hosts) KVM with SLC6 for everything else
- ◆ Currently :
 - ◆ 35 Virtual machines in Point 1
 - ◆ 11 in GPN
- ◆ Expected:
 - ◆ Additional ~100 DCS
 - ◆ ~1500 transient VM instances for Sim@P1, on netbooted nodes



Sim @ P1

- ◆ HLT Farm would be mostly idle during Long shutdown 1 (LS1)
- ◆ Use it for ATLAS simulations jobs considerable fraction (>10%) of ATLAS Grid
- ◆ Support from BNL during the setup phases
- ◆ Support from NetAdmin for dedicated network
- ◆ Preserve security of ATCN by isolating VMs and VLANs
- ◆ Tests are ongoing in TDAQ Lab4 Test Bed

| # | Type | CPU Cores: non-HT / HT | Memory | Local disk |
|-------------|--------------|---------------------------|--------------|---------------|
| 341 | Dell PE 1950 | 8 / 8 (*) | 16 GB | 80 GB |
| 320 | Dell PE 6100 | 8 / 16 | 24 GB | 250 GB |
| 832 | Dell PE 6100 | 12 / 24 | 24 GB | 250 GB |
| 1493 | Total | 15272 / 27816 | 33 TB | 315 TB |

(*) HT not enabled



Sim @ P1

The Plan



- ◆ *OpenStack*, as an overlay infrastructure
 - ◆ provides necessary management of VM resources
 - ◆ support & control of physical hosts remain with TDAQ
 - ◆ delegate VM Farm support to Offline Operations
- ◆ Easy to quickly switch from HLT ↔ GRID
 - ◆ i.e.: during LS1: monthly full-scale test of TDAQ sw upgrade
- ◆ BNL, CERN IT, CMS
 - ◆ sharing experiences
 - ◆ support if needed
 - ◆ ATLAS is already, successfully using BNL cloud resources



Conclusion

- ◆ 3 years of LHC run, points to a **good, linear scalability** of the LFS-based architecture, **OK overall**, but to be kept under control
- ◆ Netboot limited functionality, great flexibility in OS migration
- ◆ Localboot management is improving with puppet
- ◆ The monitoring system will see significant improvements
- ◆ Overall, the current architecture is **sound and performing well**, and the possible **changes** would increase complexity and **not give drastic improvements**
- ◆ We have decided to focus on trying to **simplify and streamline** the system where possible, making it **more robust and maintainable**
- ◆ The experience in cloud and virtualisation from the Simulations in Point1 project can be useful for future evolutions



BACKUP



Sim @ P1

SDX1@P1: Full Scale System

