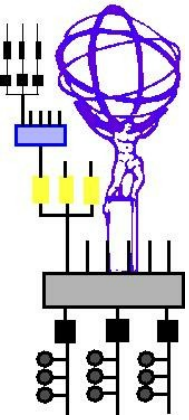



# *System And Network Administration for SoCs*

**D. A. Scannicchio &**
**M. Dobson**
**on behalf of the**
**ATLAS TDAQ & CMS DAQ**
**SysAdmin Teams**

# Content

- CNIC Rules
- OS Administration
- Network Administration
- Network address proposal
  - ★ HPM.3 Site Type proposal, Client ID, DNS Names
- Containers
- How far have we come since last Workshop
- Next Steps
- Conclusions

# General (reminder)

This presentation covers System and Network Administration on Experiment Networks (would be applicable to the Technical Network also), and will concentrate on the specific case of SoC devices.

- All CERN users have agreed to Operational Circular N°5 (OC5) and the subsidiary rules and guidelines defining the use of CERN Computing facilities (includes the experiment facilities)
- The Computing and Network Infrastructure for Controls (CNIC) security policy defines the Experiment domain rules:
  - ★ https://edms.cern.ch/document/584092
  - ★ Experiment networks are **isolated** from the General Purpose Network
  - ★ Dedicated **Gateway PCs** are used to access the Experiment Networks
  - ★ Communication between domains are **restricted** with Firewall Rules
  - ★ All systems must be **centrally managed**
  - ★ Regular updates and capability to apply **security patches**
  - ★ Only **Approved systems** connected to the network
  - ★ **Network segregation** and isolation used for restricted connectivity
  - ★ Responsibility could be delegated to the owner of the devices

# Operating Systems at CERN

- Currently supported OSs at CERN
  - ★ CentOS 7 & RHEL 7 : EOL 30 June 2024
    - Extended Lifecycle support for special cases (e.g. experiments): 30 June 2026
  - ★ ALMA/RHEL 8: EOL 31 May 2029
  - ★ ALMA/RHEL 9: EOL 31 May 2032
  - ★ End of RH Site License 31 May 2029
- Future Linux Committee
  - ★ Gathered requirements for future linux from all stakeholders, including experiments
  - ★ Options as presented in IT Linux talk

See presentation by Alex for IT-Linux team

# Operating System Administration

- By experiment TDAQ SysAdmins

- The currently supported Linux OSes
  - ★ In ATCN, moving to ALMA 9 over the EYETS 23/24
  - ★ In CMS network: mainly CC7, RHEL 8 for DAQ Event Building & HV, ALMA 9 for core services. Subsystems can move to ALMA 9 during EYETS 23/24 if desired
  - ★ All are installed, configured and managed by the corresponding SysAdmin team
- Embedded systems' OSes (e.g. Raspberry PIs, Topic, ARM, ATCA) are mostly not supported by CERN IT and therefore not by SysAdmins either
  - ★ Exception is ARM (aarch 64)
- Embedded Linux devices are currently isolated
  - ★ Centrally maintained node acting as a Gateway, connected to the experiment network AND to an isolated "private" switch
  - ★ Or network equivalent (see later)
- Aim is to ensure security for all the experiment network systems
  - ★ Protecting experiment devices from the embedded systems and vice versa
- Appropriate policies on regular updates and/or security updates need to be defined to lift restrictions

# OS management and updates

- Inside ATLAS/CMS CERN Linux repos are mirrored and snapshotted
  - ★ Prevent unwanted updates
  - ★ Know exactly which version is being used
  - ★ Know which nodes are aligned and how
  - ★ Some software is linked to the kernel version (e.g. TDAQ drivers)
- Usually OS are updated yearly (during YETS)
  - ★ Unless major vulnerabilities and update is requested by Computing Security Officer
- Updates/upgrades are tested on test nodes before applying on productions
  - ★ Normally tested on test beds or selected machine, and pushed to production a couple of weeks later
- Procedure is speeded up in case of a security vulnerability
  - ★ Could be done within a day or so (happened in the past)

# Operating System for ARM

- Support for ARM (aarch 64), upstream support by OS
  - ★ 32 bit not supported
  - ★ Packages and images made available, CERN packages built
- Kernel ?
  - ★ Can the RHEL/ALMA kernel be used ?
    - ➢ Demonstrated to work for simple design/test boards
    - ➢ However RHEL 9 removed support for Xilinx HW between 9.1 and 9.2
      - Need to go back to stock kernel: being investigated now
    - ➢ How to port all needed device drivers? How to keep up with Vivado/Petalinux updates?
  - ★ Petalinux Kernel: used by most/all people at the moment
    - ➢ What mitigation might be needed (kernel config, other…) to use it unprotected in exp. net. ?
- Root FS
  - ★ Use the RHEL/ALMA one
    - ➢ Proven to work
  - ★ NFS mounted root FS?
    - ➢ Potentially read only, with read/write overlay for each host
  - ★ Puppet based configuration steps
    - ➢ Used as only mechanism ? (Yes for ATLAS, very high probability for CMS)
    - ➢ Puppet apply for well defined synchronisation points (boot) and no run time interference

# Netbooted OS

- ATLAS already use netbooted machines extensively
  - ★ Kernel and ramdisk loaded via UEFI, DHCP, TFTP
  - ★ Unique shared root filesystem mounted from the local file server
    - ➢ Read/Write overlay on read-only root FS
    - ➢ System finishes its configuration in the RW areas by running Puppet at boot time (then once an hour)
  - ★ Versioning for the root filesystem, kernel and initial ramdisk together
  - ★ Able to go forward and back as needed
- Netbooting of SoCs is nearly as easy
  - ★ SDcard to hold only FSBL and U-boot, rest can be obtained remotely:
    - ➢ Shown by a few people, including Kareen on Thursday
    - ➢ Even more versatile with partial PS configuration. Complexity in development/understanding ? Needs better understanding.
  - ★ Can be extended for the bit map files & HW description
    - ➢ To include bit map files, hardware description files etc…
  - ★ Versioning and auto upgrade/downgrade has also been shown
  - ★ Backup to QSPI or eMMC can also be used as backup/failover
  - ★ Advantages: **file system accessible even when SoC is down**

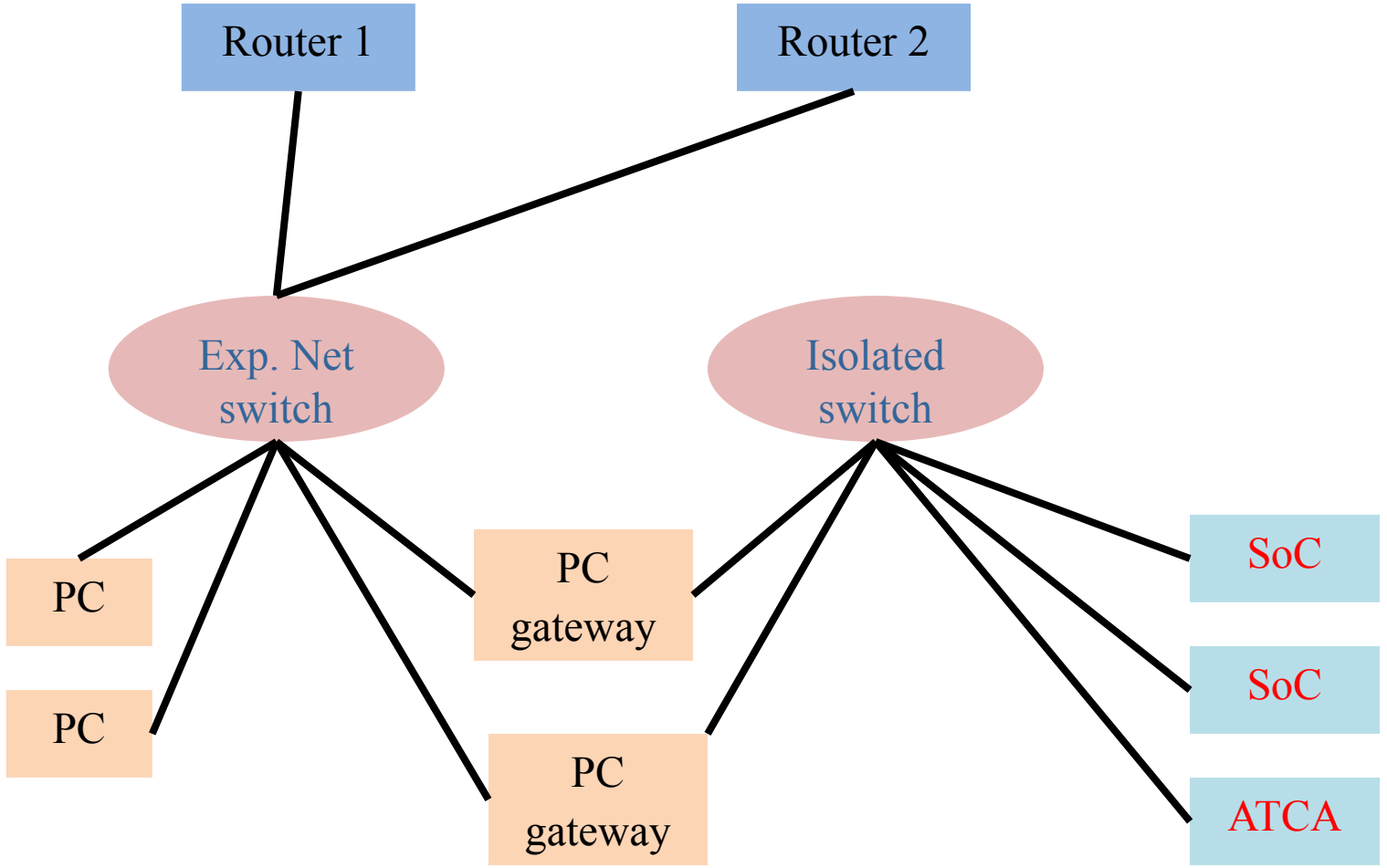
# Network Administration

- The CNIC Security Policy also defined network aspects
- All network devices (mainly Ethernet) must be registered in the CERN Network database (LanDB)
  - ★ No unregistered private networks (e.g. 192.168.X.Y) are allowed
  - ★ All private interfaces have to be registered
  - ★ No Network Address Translation (NAT) allowed
  - ★ Experiment networks are private 10.X.Y.Z subnets
  - ★ Network isolation:
    - ➢ Isolated networks can be registered as Experiment network subnets (e.g 10.X.Y.Z)
    - ➢ Restricted experiment subnets via Access Control Lists are more versatile. They are implemented via Inter Domain or Control Sets in LanDB (CERN Network database).
- Network infrastructure and devices can be managed by IT Networking or by experiments
  - ★ In practice most experiments asked IT networking to manage their networks
    - ➢ The case for ATLAS and CMS, the accelerator, etc…
  - ★ Monitoring tools and HW replacement also taken care of by IT networking
  - ★ However, DHCP/DNS essential services managed by TDAQ SysAdmins
    - ➢ Using the information from the IT network database

# System Administration Today

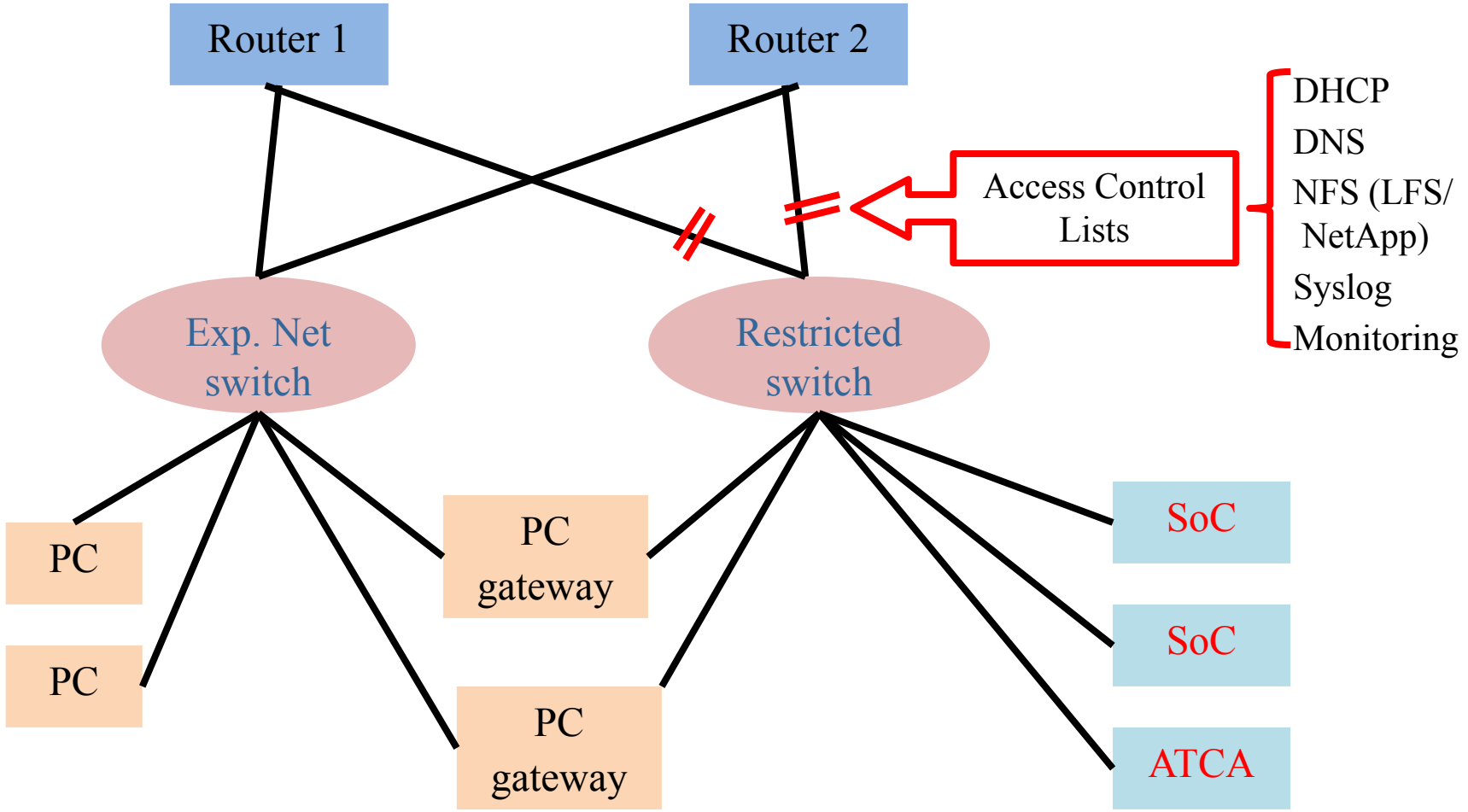
- Central Management by Experiment System Administration teams
  - ★ Manage the CERN supported OSes
  - ★ A Configuration Management System is used
    - ➢ Puppet for both ATLAS and CMS (also used by CERN), others in other areas
  - ★ User/role management also available throughout the system
  - ★ Authentication via Kerberos (not entirely or not yet, LDAP used for some cases)
- For devices with non managed OSes
  - ★ Network Isolated
  - ★ Managed Gateway nodes to access them
  - ★ Additional effort to configure the isolated network
  - ★ Monitoring systems have to be adapted to check isolated devices
  - ★ Striving to ensure security for the whole system
- Today split between the two categories is very much >95% for centrally managed versus <5% for non managed

# Isolated Switch



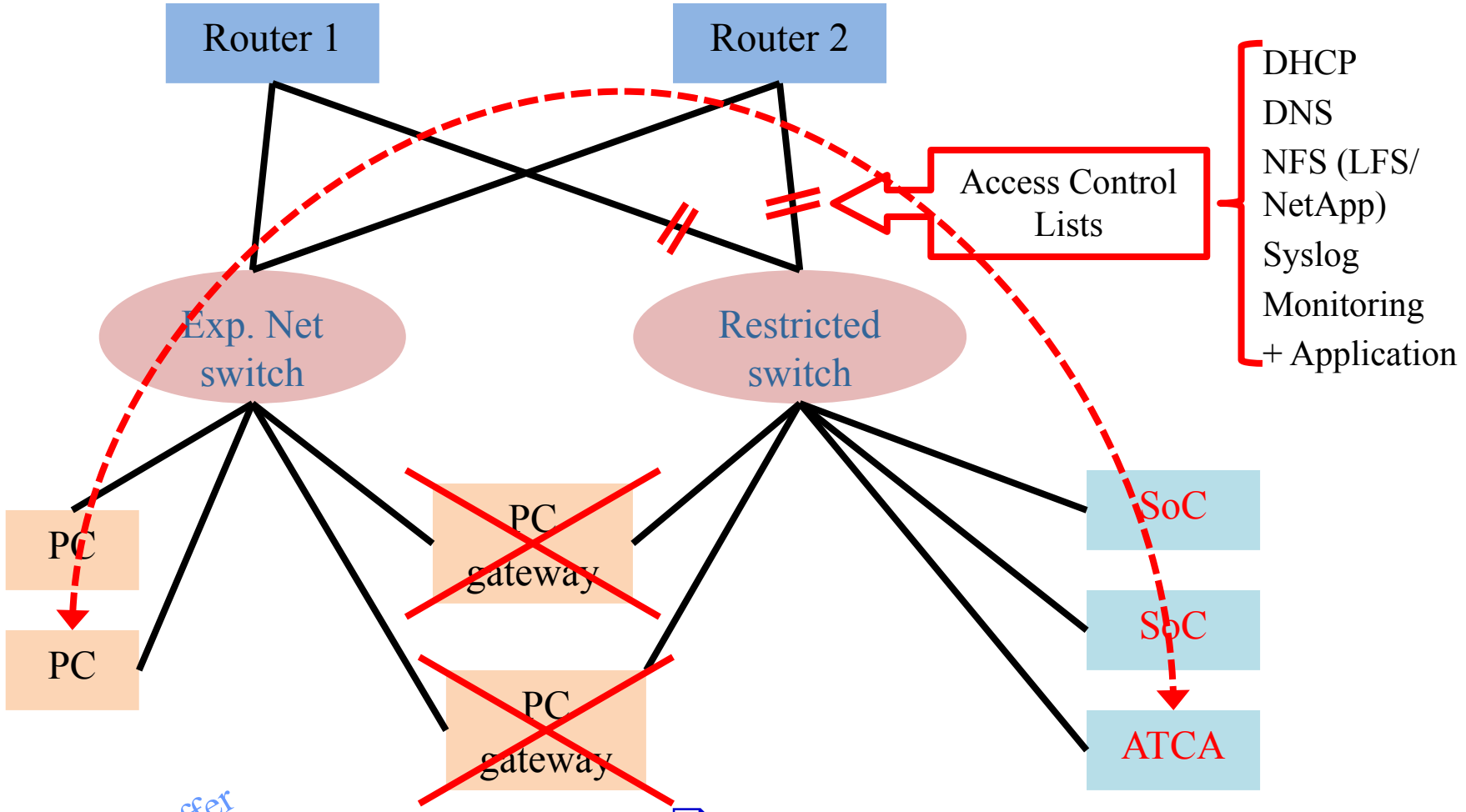

- Pros
  - ★ Full isolation
- Cons
  - ★ Network config, remote monitoring etc… (see previous slides)

# Restricted switch: core services




- Pros
  - ★ Easier core services integration
  - ★ Restrictions still apply
- Cons
  - ★ GW PC still needed for application level access

# Restricted switch: limited comm.




- Pros
  - ★ ~~Gateway~~ PC not compulsory
  - ★ Selective connectivity
  - ★ Can use 10Gb or more
- Cons
  - ★ Potential traffic bottleneck
  - ★ Potential traffic DoS

# Run 4 Global Challenges

- Proportion of SoC (maybe unsupported Linux devices) increases dramatically for Phase-2
  - ★ Order 150 ATCA crates (both ATLAS/CMS) with ~10 boards each (~1500 ATCA boards)
  - ★ Each ATCA board has at least an IPMC (could be Zynq based, probably with FreeRTOS)
  - ★ Each board will likely have a Zynq or ComExpress controller
  - ★ Potentially more Zynq devices on the core FPGAs
- Network isolation could be very cumbersome
  - ★ For all the reasons stated earlier, exacerbated by sheer number
  - ★ Gateway PCs could be a significant number
- Restricted Network are probably the viable alternative
  - ★ Restrictions but with many core services available
  - ★ Potentially getting rid of Gateway PC and allow selective connectivity
  - ★ Gateway PC may need to be replaced by “Buffer/Supervisor” PC (see later)

# ATCA Specifics: network addresses

- ATCA spec. foresees usage of Client ID based DHCP
  - ★ Each Shelf is identified by a Shelf address (an arbitrary <=20 character long ASCII string):
    - Should be unique in a DHCP domain
    - For us it could be “Building”, “Rack” and “U” (always unique): USC55-S1A10-10
    - Could drop “Building” in experiment networks
    - Configuration of shelf done when installed (script to run over serial line from laptop)
  - ★ IPMC gets this info (FRU data) from Shelf Manager
  - ★ ATCA boards are located in the shelf by the Physical Slot Number
    - Primary site type is 00h and primary site number is Physical slot number
  - ★ Secondary site type and secondary site number can be used to identify sub-elements of an ATCA board (e.g. AMC Modules):
    - Could be used to identify Zynq, FPGAs, Switches
    - Propose to use values from OEM range (see next slide)
  - ★ ComExpress or Zynq controllers can get this info from IPMC:
    - See talk by Petr on Thursday and the talk by Ralf in previous SoC Workshop
- Suggest to implement LLDP on end-points:
  - ★ Very useful for network debugging, especially during commissioning
    - Done in ATLAS & CMS for Run 3 in DAQ data network, very useful

# Secondary Site type/number proposal

- Primary Site Type is 00h for a board level end-point (HPM.3 spec)
- Primary Site Number is then the Physical Slot Number of the board
  - ★ See warning by Petr on Physical versus Logical versus what you see !
- Secondary Site Type offers multiple categories
  - ★ Suggest not to use the RTM one (even for the SRTM or CMS DAQ RTM)
  - ★ Suggest to use OEM range (C0h to CFh)
- Secondary Site Numbers
  - ★ Suggest starting at 1 and allowing multiple (will not use 0 with OEM type)

| End Point | Primary Site Type | Primary Site Number | Secondary Site Type | | Secondary Site Number |
|---|---|---|---|---|---|
| Board IPMC | 00h | XXh | 00h | | 00h |
| Board Switch | 00h | XXh | CFh | Last OEM value | YYh |
| Board Controller | 00h | XXh | C0h | First OEM value | YYh |
| Workhorse FPGA | 00h | XXh | C1h | Second OEM value | YYh |
| Other | 00h | XXh | CNh | 1 < N < F | YYh |

# Example Client IDs for end-points

- Example using the above proposal

- Crate located in CMS underground service cavern (USC55) in rack S1A10 in U10
  - ★ Shelf Address: `USC55-S1A10-10 or`
    `55:53:43:35:35:2d:53:31:41:31:30:2d:31:30:00:00:00:00:00:00`

- Shelf manager Client IDs)
  - ★ RMCP: `ShelfAddress + 00 + 03:00:00:00`
  - ★ Slot1: `ShelfAddress + 00 + 03:01:00:00`
  - ★ Slot2: `ShelfAddress + 00 + 03:02:00:00`

- Board level Client IDs for a board in slot 7 (1st Hub slot)
  - ★ IPMC: `ShelfAddress + 00 + 00:07:00:00`
  - ★ ZynQ: `ShelfAddress + 00 + 00:07:C0:01`
  - ★ Switch: `ShelfAddress + 00 + 00:07:CF:01`
  - ★ FPGA: `ShelfAddress + 00 + 00:07:C1:01`

# DNS Names proposal (Phase-2)

- Geographical
  - ★ Based on Shelf Address, Slot, Function and Index/Number
  - ★ Prepended by “ATCA”
  - ★ Shelf Manager:
    - ➢ ATCA-USC55-S1A10-10-SHMM-0/1/2
  - ★ IPMC:
    - ➢ ATCA-USC55-S1A10-10-XX-IPMC
  - ★ Switch:
    - ➢ ATCA-USC55-S1A10-10-XX-SW
  - ★ Controller (Zynq or Com-e):
    - ➢ ATCA-USC55-S1A10-10-XX-CTRL-1/2/3
  - ★ FPGA:
    - ➢ ATCA-USC55-S1A10-10-XX-FPGA-1/2/3
  - ★ ATLAS not using this kind of naming convention today
- DNS Aliases can be added for easier use by sub-detectors
  - ★ e.g. ATCA-TRACKER-ECM-1
- For end-points which cannot do DHCP
  - ★ SoC could do request on behalf of end-point and communicate IP back:
    - ➢ e.g. reverse lookup using naming convention

Usually single controller (SoC) so shortcut is to drop “-1” on the end

# Support for ClientID by IT

- Under discussion since last SoC Workshop

- Support for ClientID accepted in the network database (LanDB)
  - ★ Property of a device (like MAC address)
  - ★ Linked to an interface
  - ★ Soap methods to add information and query (rest API later)
  - ★ Should be going **production before the end of the year**

- NO support for ClientID in the CERN DHCP servers
  - ★ Question of security
    - ➢ Currently based mainly on MAC address verification/banning
  - ★ Not a major problem
    - ➢ At the experimental site, TDAQ SysAdmins manage DHCP servers with info extracted from LanDB
    - ➢ In the labs, usually have private DHCP servers (can also be populated from LanDB)

# ATCA network connectivity

- Network connectivity needs
  - ★ Each board has  a 1Gb connection (used for Zynq/IPMC)
    - ➢ Should go via the Hub Slot Switch
  - ★ Do boards “require” (not want) higher bandwidth links ? Usage/Need ?
  - ★ CMS plans a 1Gb per board with 2x10Gb uplinks (redundancy ?)
  - ★ Front/back panel links not foreseen in CMS => requires external switch (space, power)
- Overall bandwidth needs (NEED INPUT)
  - ★ 1500 boards at 1Gb => 1.5Tb/s !!! (no mention of 10Gb)
  - ★ Aggregate links with row level switches or one chassis switch
  - ★ Where to ?
    - ➢ Boot, syslog, NFS servers: probably local per sub-system and/or switch
    - ➢ Monitoring: where to ? How much ? Local aggregation points by sub-system ?
    - ➢ Configuration data: buffer/supervisor PC that gets the data from the DB and feeds the boards with information at high speed (10-100Gb)
    - ➢ Control: buffer/supervisor PC to act as intermediary controller for many crates/boards
  - ★ Need to understand needs to design the infrastructure ! Hence test systems.

# Containers: a few comments

- Currently ATLAS/CMS do not use containers online
  - ★ Currently hypervisors and VMs are used (same install procedure as for physical machines)
- Containers are nice for Application development/deployment
  - ★ TDAQ SysAdmins understand this
- Completely new model to integrate into infrastructure
  - ★ Being investigated by TDAQ SysAdmin teams
  - ★ The build of container images gives users full flexibility
    - ➢ More than they would usually get on bare metal !
    - ➢ Could include SW which is not known or with security bugs which TDAQ SysAdmins have no control over (compare to bare metal)
  - ★ Security assessment of the images is more complex to evaluate
    - ➢ CERN level service would be nice : **proposal to IT being assessed in the next few months**
  - ★ Will require significant work to be able to come to a production service
  - ★ Kubernetes being investigated by some ATLAS & CMS groups at proof of concept level
    - ➢ Some people have ideas about running container on SoC !

# Proposals: advances since last WS

- Support for an ARM OS
  - ★ No support for Arm32 from IT due to upstream support not existing
  - ★ Arm64 support for CentOS/RHEL/ALMA 8 and 9
  - ★ Links with CentOS community Special Interest Groups ?
  - ★ Central build infrastructure: two initiatives
    - ➢ Build infrastructure for FPGA etc… (mentioned earlier this week by ATS sector and supported by EP)
    - ➢ Build infrastructure for Gitlab-CI with bare metal: being setup by IT, with new HW on order
- Discussions with IT networking to support ClientID
  - ★ Available in LanDB by end of this year
- Discussions with sub-detectors HW producers to understand network requirements
  - ★ Number of connections
  - ★ Of which type (1Gb/10Gb)
  - ★ Via a Hub slot switch/front panel

More or less understood
Some worries about X2O...

# Proposals 2: advances since last WS

- Feed the SoC Interest group twiki with:
  - ★ Tools, tool chains
  - ★ Compilers
  - ★ Instructions for using/building etc…
  - ★ Document working (or not) solutions
- Use the SoC Interest Group mailing list
  - ★ For sharing knowledge
  - ★ For community help
- Test setups
  - ★ Provided and maintained by SysAdmins
  - ★ Where supported tools/ideas/solutions can be tested
- Define collaboration between SysAdmin teams
  - ★ Mutualize effort
  - ★ Pool knowledge
  - ★ Avoid re-inventing the wheel

SoC Interest Group setup

Twiki has information

Mutualising effort between SysAdmins groups

Test setups being finalised

# Next steps

- Not listed in any special order, may be tackled out of order

- Setup ClientID support on experiment test setups
  - ★ Test implementation done by IT in LanDB
  - ★ **Workaround in place for CMS**, based on naming convention (see next talk)
- Setup root FS with puppet configuration running at boot
  - ★ With NFS in RW mode: **working**
  - ★ With RO root FS and RW overlay
- Are people interested in VM/docker image with central services for labs or bastion hosts ?
  - ★ DHCP with ClientID, TFTP, NFS, syslog server, DNS, NTP : **shown by Kareen on Thursday morning, more news on next presentation**
- Integration with different types of boards
  - ★ Investigate requirements
  - ★ Understand issues
  - ★ Look at the kernel issues (see previous slides)

# Words of Caution

- An attempt at clarification ...

- What is applicable to Phase-1 ?
  - ★ MAC addresses and not Client ID
  - ★ Network isolation (no support for ARM) in both proposed ways

- What is applicable to Phase-2 ?
  - ★ Site Number proposal, Client ID, Naming convention
  - ★ ARM support (likely netboot)
  - ★ Puppet configuration

- Differences between ATLAS/CMS
  - ★ On timescale of Phase-2, should be more or less in line on all covered topics
  - ★ Specific exception would be mentioned as we progress
  - ★ Different role out/testing timescales (very manpower dependent)
  - ★ We are collaborating to try to benefit from the available manpower

# Breaking news !

- GitLab - New ARM Runners centrally offered OTG0144823

- ARM Instance Runners
  - ★ Git service would like to share an important update regarding our GitLab Runners infrastructure.
  - ★ As of Friday 6th October 2023, Git Service will start offering new centrally ARM runners based on Kubernetes.
  - ★ Initially, these runners will run with limited capacity, allowing a total maximum of 6 concurrent jobs, and they will run non-privileged workflows (docker commands are unlikely to work). This concurrency will be increased over time as the demand increases, and privileged workflows will be considered in a later stage.

- The documentation about usage of GitLab ARM Runners can be found at General considerations for K8s+executer ARM Runners

- GitLab documentation in general is: https://gitlab.docs.cern.ch/

# Conclusions

- Progress is being made on the TDAQ SysAdmin side
  - ★ Booting of SoC
  - ★ ClientID soon ready for use, workaround already in place for CMS
  - ★ OS support:
    - ➢ Stock kernel to be investigated
  - ★ Setup and use/support of test labs
  - ★ Containers usage being investigated

- You may be interested to subscribe to an IT Mattermost channel (ARM64): https://mattermost.web.cern.ch/it-dep/channels/aarch64-arm64

- Thank you

# Backup

# Network addresses up to now…

- Standard PCs/uTCA based HW:
  - ★ Names (in CMS) are functional & geographical
    - ➢ {type/function}-{Rack}-{U-position}-{index/slot}
      - ✓ mch-s2f05-20-01
      - ✓ amc-s2f05-20-[01..12]
  - ★ Names (in ATLAS) are functional
    - ➢ {type}-{detector}-{function}
  - ★ Either IP/Name assigned based on MAC address using DHCP or RARP (uTCA)
    - ➢ Needs update to change PC/board in case of failure
    - ➢ Centrally managed by TDAQ SysAdmin Teams
    - ➢ ATLAS has this one only
  - ★ Or using System Manager (based on IPMI board insert/removal) triggers
    - ➢ Needs description of crate layouts
    - ➢ No MAC address needed
    - ➢ Geographical names important
    - ➢ No update on failure
    - ➢ Layout description manage[illegible]y sub-detector