CMS Network and Booting

M. Dobson

on behalf of CMS DAQ

Overview

This presentation needs and follows the common ATLAS + CMS general presentation to avoid repeating content.

- ClientID support in CMS
- ClientID known issues
- Micro Services for network boot
- Next Steps

ClientID support in CMS

- CMS ATCA Specifics documented in EDMS note 2735323: “Custom hardware network interface specification for Phase-2 CMS”
  - https://edms.cern.ch/document/2735323/
  - **Shows requirements on HW developers** for their boards
  - Has all references to ATCA specification
  - Note: a mistake in the ClientID part compared to original proposal
    - Shelf ID does NOT contain ATCA, e.g.: USC55-S1A10-10
  - Update to the note in preparation

ClientID support: DNS

- 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
- DNS Aliases can be added for easier use by sub-detectors
  - e.g. ATCA-TRACKER-ECM-1
- Streamline most used IP names (when there is typically one)
  - Controller: ATCA-USC55-S1A10-10-XX-CTRL-1
    becomes ATCA-USC55-S1A10-10-XX-CTRL
  - Could do the same for the RMCP Shelf Manager IP

ClientID support: production

- The ClientID proposal (as shown in previous presentation) is implemented in CMS
  - Already supported by our DHCP servers
- Will use the ClientID field in the IT network database (LanDB) when registering devices
  - As soon as available (end of 2023)
    - In the mean time, at usage stage rely on the naming convention
- When building DHCP server configuration today
  - Take ATCA devices registered in the CMS network from LanDB
  - Build ClientID from the naming convention
  - Insert next to the PCs using MAC Addresses

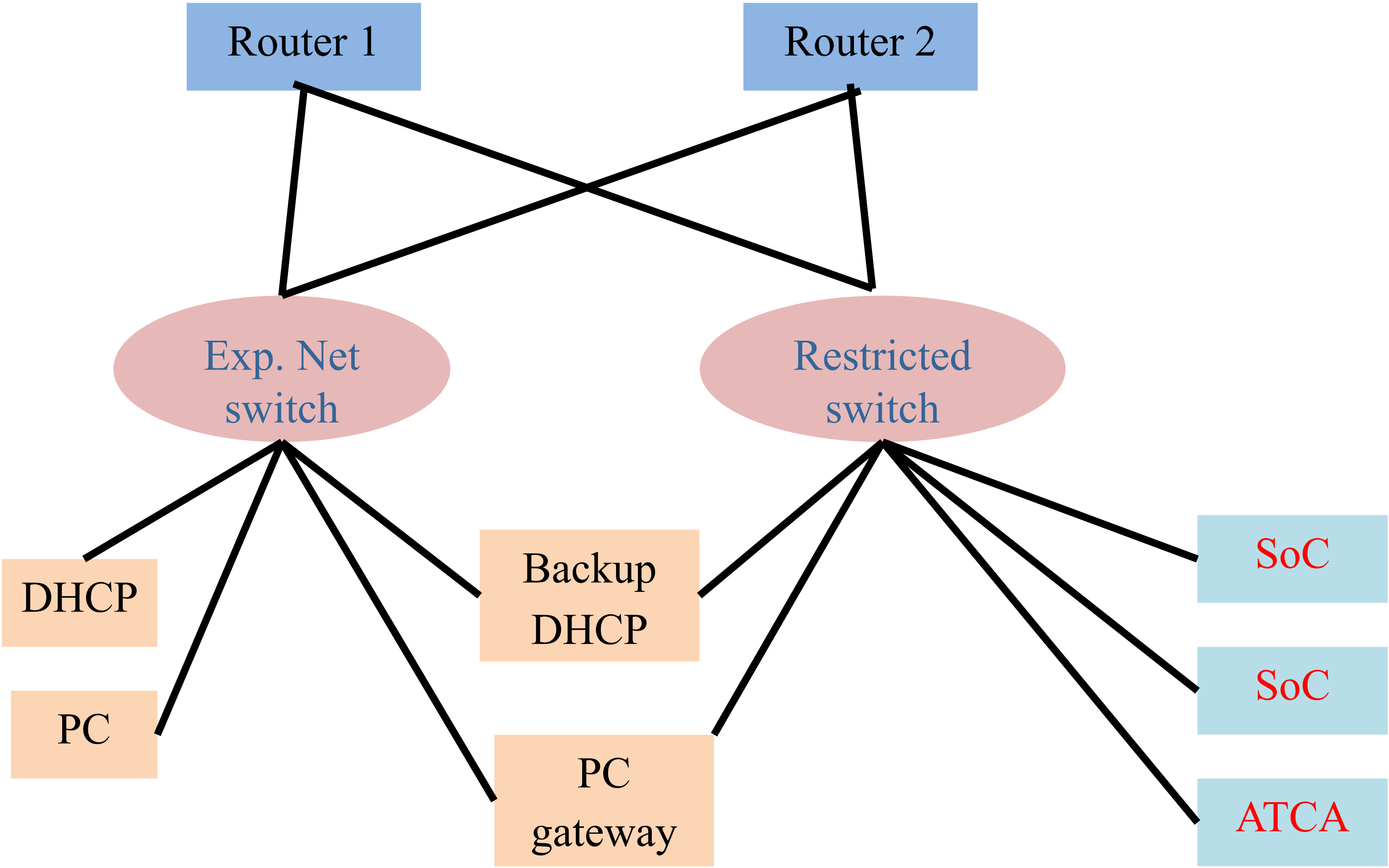
ClientID as seen in dhcpd.conf

```
host atca-scx5-c2e32-09-01-ctrl { host-identifier option dhcp-client-identifier ff:00:00:00:00:00:02:00:00:31:5a:48:50:4d:2e:33:2d:31:cd:53:43:58:35:2d:43:32:45:33:32:2d:30:39:00:00:00:00:00:00:00:00:00:01:c0:01; fixed-address 10.176.73.192; option host-name atca-scx5-c2e32-09-01-ctrl; option subnet-mask 255.255.255.128; option broadcast-address 10.176.73.255; option routers 10.176.73.129; }
host atca-scx5-c2e32-09-01-ipmc { host-identifier option dhcp-client-identifier
....
host atca-scx5-c2e32-09-07-sw { host-identifier option dhcp-client-identifier ff:00:00:00:00:00:02:00:00:31:5a:48:50:4d:2e:33:2d:31:cd:53:43:58:35:2d:43:32:45:33:32:2d:30:39:00:00:00:00:00:00:00:00:00:07:cf:01; fixed-address 10.176.73.206; option host-name atca-scx5-c2e32-09-07-sw; option subnet-mask 255.255.255.128; option broadcast-address 10.176.73.255; option routers 10.176.73.129; }
....
host atca-scx5-c2e32-09-shmm-0 { host-identifier option dhcp-client-identifier ff:00:00:00:00:00:02:00:00:31:5a:48:50:4d:2e:33:2d:31:cd:53:43:58:35:2d:43:32:45:33:32:2d:30:39:00:00:00:00:00:00:00:00:00:03:00:00:00; fixed-address 10.176.73.175; option host-name atca-scx5-c2e32-09-shmm-0; option subnet-mask 255.255.255.128; option broadcast-address 10.176.73.255; option routers 10.176.73.129; }
host atca-scx5-c2e32-09-shmm-1 { host-identifier option dhcp-client-identifier
...
host ctrl-s2c16-06-01 { hardware ethernet 6C:2B:59:B4:FC:DB; fixed-address 10.176.59.10; option host-name ctrl-s2c16-06-01; option subnet-mask 255.255.255.128; option broadcast-address 10.176.59.127; option routers 10.176.59.1; }
...
```

ClientID support: known issues 1

ClientID DHCP requests are dropped by the Juniper routers

- Found issue a couple of weeks ago
- Confirmed last week: identifies them as bad
- Reproduce in lab to test without breaking production
- Workaround to run DHCP on gateway PC (Control Hub for uTCA)



ClientID support: known issues 2

- As mentioned by Petr yesterday, Linux kernel clientID implementation is broken
  - Really in the kernel (before mounting the RootFS). It is NOT dhclient !
  - Need to investigate how to overcome this
    - Currently for Labs use features of dnsmasq DHCP server
    - Can simply pass info for static config and mounting (IP, netmask, GW, server IP, path)
- Once RootFS is mounted, all services are then started, including network
  - Here one can use ClientID as dhclient has working implementation
  - Two options:
    - Use software similar to SIPL to get the ShelfID and Slot from IPMC
    - Pass the ClientID on the kernel command line from U-Boot
  - Second option is the preferred one (less implementation work)

Micro-Services for network boot

- Set of services each running in its own container, to help boot devices over the network (DNS/DHCP, TFTP, NFS)
- Used as part of the CMS DAQ gitlab-CI pipeline and automated testing (see Karen’s presentation on Thursday morning)

- Gitlab Project for this:
  - https://gitlab.cern.ch/hardware/network-services

- Pre-requisites (on lab server)
  - Install docker and docker-compose
  - Create `tftpboot` and `nfsshare` folders (take note of paths to these)
  - Clone the repo above
  - For NTP install `chrony` locally

Micro-Services: configuration

- Add NFS and TFTP server directories to `docker-compose.yml`
- Configure the Chrony server using the `chrony/chrony.conf` config file example from git + the NTP servers in your network
- If you have `firewalld` running, open up ports to the services
- Describe your Lab configuration
  - Server, and Boards, with their info
  - In the file `boot-config.yml`
    (example provided as `boot-config-example.yml`)
- Build the services config files

```
python config-builder.py
```

Micro-Services: startup

- Start the services

```
docker-compose -f docker-compose.yml up
```

- Stop the services

```
docker-compose -f docker-compose.yml down
```

Feedback Welcome

Next steps 1

Service infrastructure:

- Fix the Linux Kernel ClientID issue
- Micro-Services
  - Syslog server
  - NTP server in container
- TFTP & NFS server integration:
  - TFTP server for firmwares and Kernel
  - RootFS over NFS
  - Read/Write overlay
- Puppet for final configuration steps:
  - What is needed ?
  - How to integrate into current puppet management ?

Next steps 2

Structural considerations for supporting thousands of devices:

- TFTP & NFS server integration:
  - Hierarchical structure for storing board type, board specific and maybe sub-system specific parts
  - Access rights for this structure to the maintainers of the boards
    - Could have server per sub-system at the beginning, simplified hierarchy
- Look into Split boot v2 for general usage
  - Keep the netbooted part
  - Use the two stage for the FSBL, PMU etc…
- Scaling studies
  - Bandwidth needs
  - Bootup time for full cluster
  - Split of NFS servers in multiple “zones”

Questions ?

Backup slides

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 (over serial line from laptop)
  - 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
    - See later slides for warning on difference between Physical and Logical slot numbers
  - 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 following slides)

ClientID in practice... 1/3

- The fixed crate part of the Client Identifier would look like:
  - | ff | 00:00:00:00 | 00:02 | 00:00:31:5a |
    | TYPE | IAID | DUID-EN | PICMG IANA ID |
  - | 48:50:4d:2e:33 | 2d:31 |
    | “HPM.3” identifier | “-1” HPM.3 DUID version |
  - | ce |
    | Shelf address type/length |
  - | 55:53:43:35:35:2d:53:31:41:31:30:2d:31:30:00:00:00:00:00:00|
    | “USC55-S1A10-10” Chassis Address or Shelf Address in ASCII |
  - | 00 |
    | Shelf Type |

ClientID Slot Location

- Primary Site Type: 00h for board level end-point (HPM.3 spec)
- Primary Site Number = Physical Slot Number of board
  - See warning on Physical versus Logical slot numbering later in presentation !
- Secondary Site Type: suggest to use OEM range (C0h to CFh)
- Secondary Site Number: suggest starting at 1 and allowing multiple (ignore 0)

End Point	Primary Site Type	Primary Site Number	Secondary Site Type	Secondary Site Number
Shelf Manager	03h	00h RMCP 01h Priv 1 02h Priv 2	00h	00h
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

Fixed by Spec.

CMS suggestion

ClientID in practice... 2/3

- The dynamic part of the Client Identifier would look like:

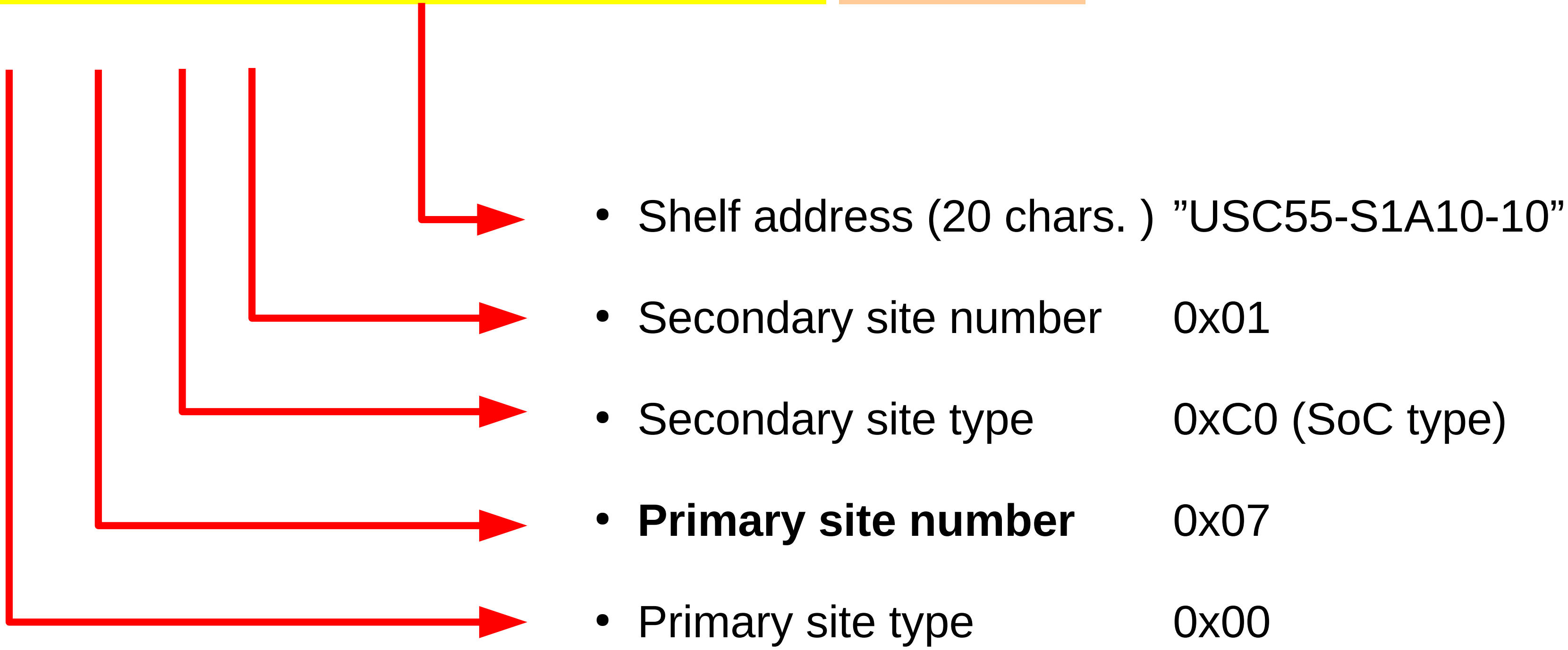
  - The Slot Location part for shelf managers:
    - RMCP: 03:00:00:00;
    - Slot 1: 03:01:00:00;
    - Slot 2: 03:02:00:00

  - The Slot Location part for board elements (e.g. physical slot 7):
    - IPMC: 00:07:00:00
    - SoC: 00:07:C0:01
    - Switch: 00:07:CF:01
    - FPGA: 00:07:C1:01

Secondary Site type and number as suggested by CMS

ClientID in practice... 3/3

- The full Client Identifier would look like:
  - host-identifier option dhcp-client-identifier ff:00:00:00:00:00:02:00:00:31:5a:48:50:4d:2e:33:2d:31:ce:55:53:43:35:35:2d:53:31:41:31:30:2d:31:30:00:00:00:00:00:00:00:00:00:07:C0:01;



Shelf address (20 chars.)	”USC55-S1A10-10”
Secondary site number	0x01
Secondary site type	0xC0 (SoC type)
Primary site number	0x07
Primary site type	0x00

DNS Names proposal

- 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
- DNS Aliases can be added for easier use by sub-detectors
  - e.g. ATCA-TRACKER-ECM-1

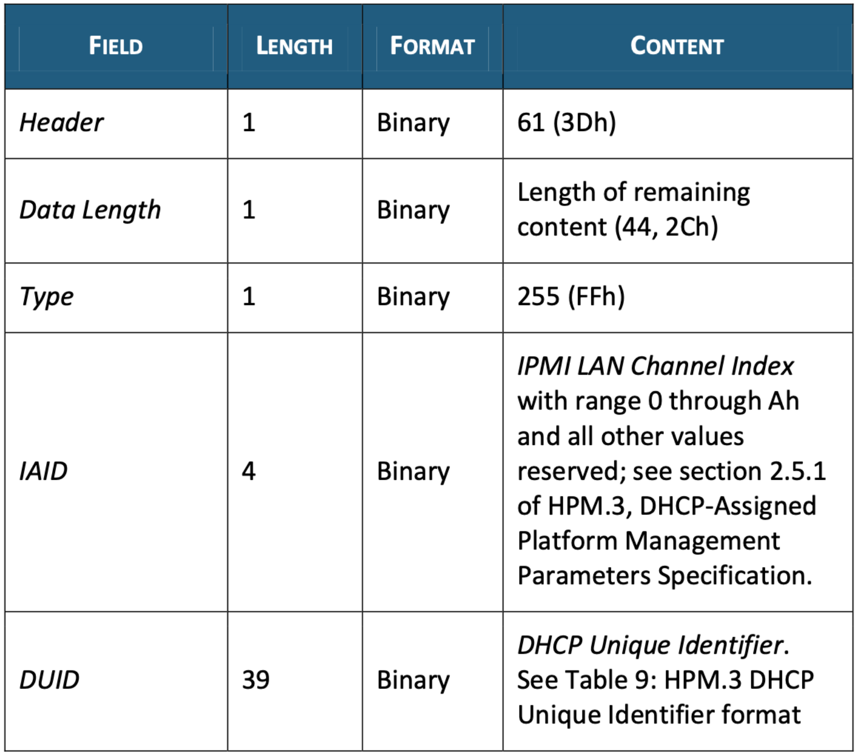
ClientID content

- RFC 4361 gives this format:

```
Code  Len  Type  IAID              DUID
+----+----+-----+----+----+----+----+----+----+---
| 61 | n  | 255 | i1 | i2 | i3 | i4 | d1 | d2 |...
+----+----+-----+----+----+----+----+----+----+---
```

  - Code is 61, Type 255,
  - IAID and DUID are “Identity Association Identifier” and “DHCP Unique Identifier” as originally specified in RFC 3315 (DHCPv6)
  - The other parameters depend on the exact sub-format used (e.g. HPM.3)
- The “**Pigeon Point Chassis Manager User Guide (Jun 20, 2018)**” picks up the HPM.3 DHCP spec. in a nice table.

Table 8: HPM.3 DHCP Client ID

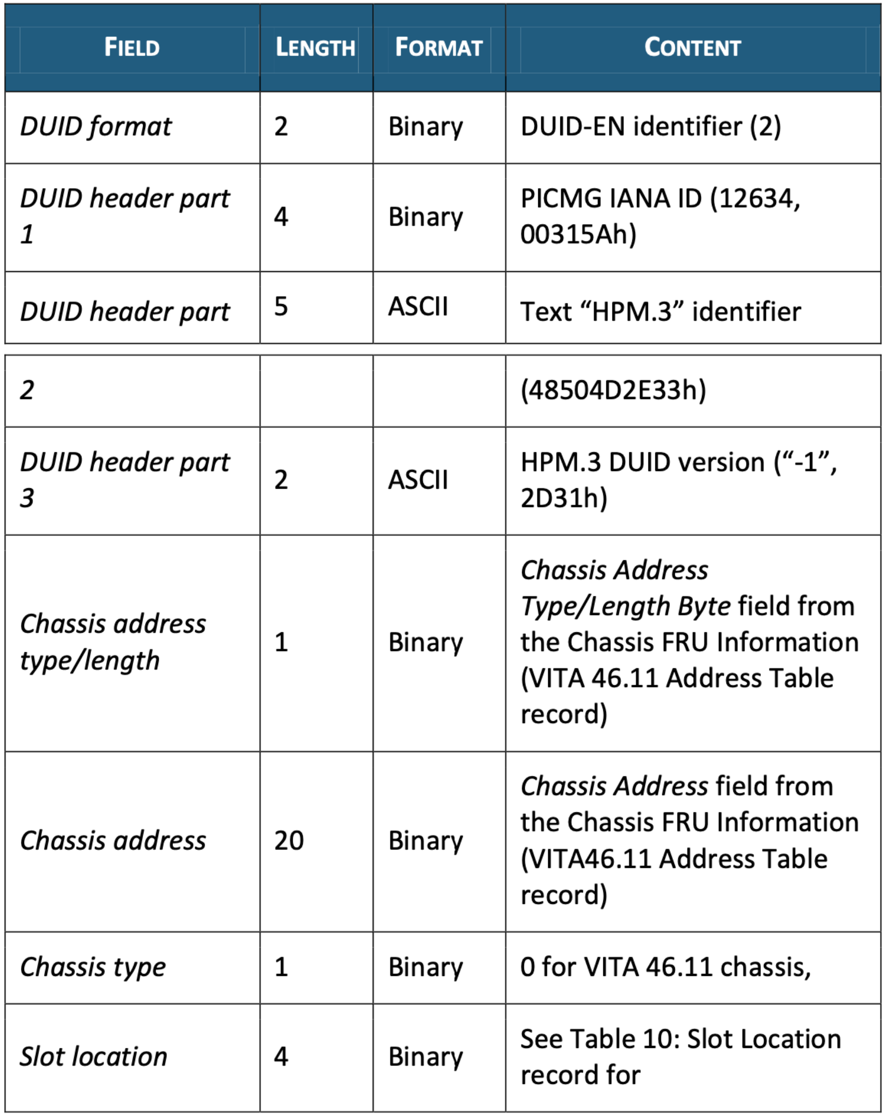


Field	Length	Format	Content
Header	1	Binary	61 (3Dh)
Data Length	1	Binary	Length of remaining content (44, 2Ch)
Type	1	Binary	255 (FFh)
IAID	4	Binary	*IPMI LAN Channel Index* with range 0 through Ah and all other values reserved; see section 2.5.1 of HPM.3, DHCP-Assigned Platform Management Parameters Specification.
DUID	39	Binary	*DHCP Unique Identifier*. See Table 9: HPM.3 DHCP Unique Identifier format

HPM.3 DHCP Unique Identifier

- The “**Pigeon Point Chassis Manager User Guide (Jun 20, 2018)**” gives the HPM.3 DUID content:
  - The Chassis address and address type are given in the chassis FRU
  - The Chassis address uniquely identifies the chassis (unique in the domain). It is 20 bytes binary or ASCII (depends on the type field).
  - Suggestion is to use ASCII for the chassis address (human readable)
  - Chassis Type is 00h for ATCA
  - The slot location identifies exactly the location in the chassis (see next slides)

Table 9: HPM.3 DHCP Unique Identifier format



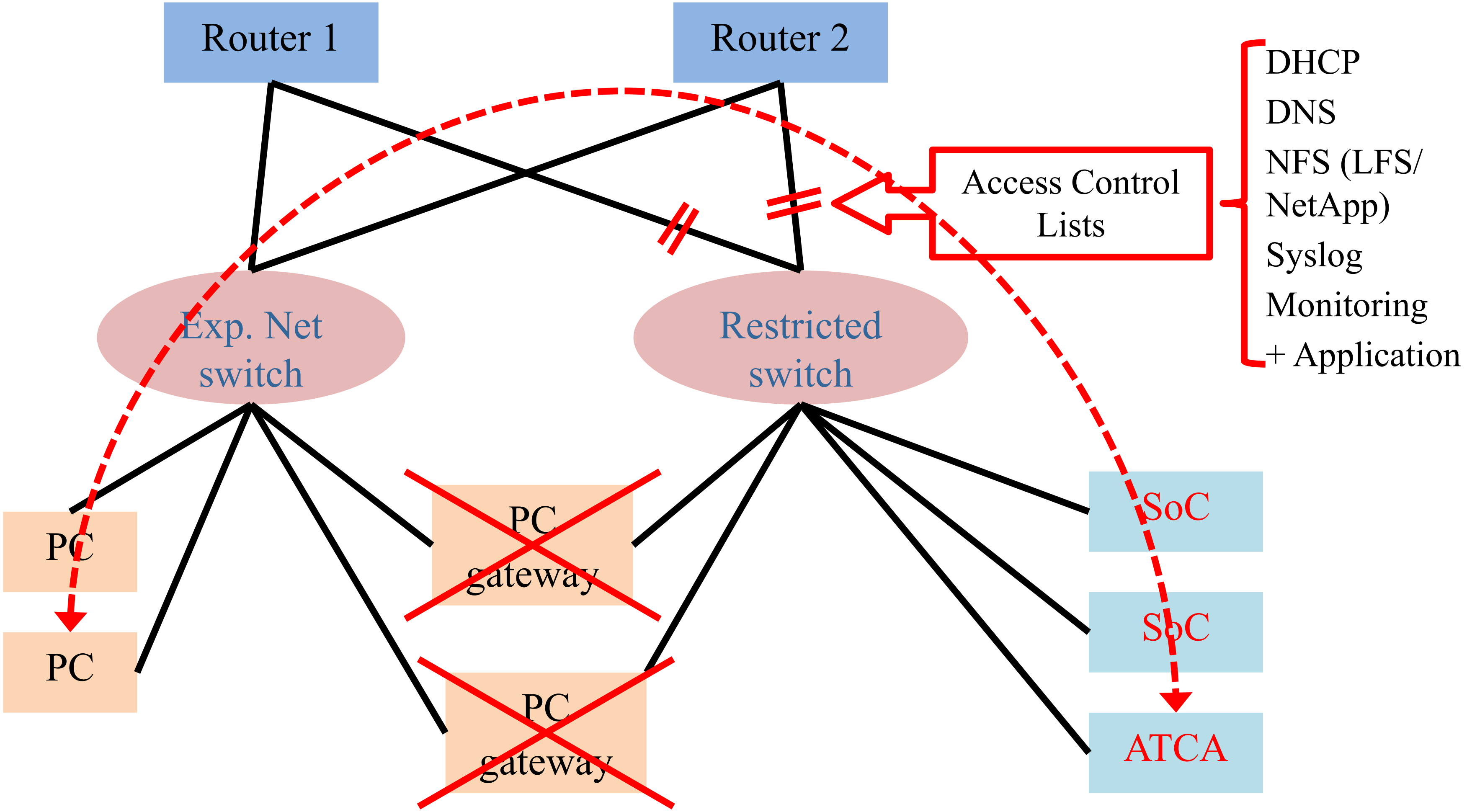
Field	Length	Format	Content
DUID format	2	Binary	DUID-EN identifier (2)
DUID header part 1	4	Binary	PICMG IANA ID (12634, 00315Ah)
DUID header part 2	5	ASCII	Text “HPM.3” identifier (48504D2E33h)
DUID header part 3	2	ASCII	HPM.3 DUID version (“-1”, 2D31h)
Chassis address type/length	1	Binary	*Chassis Address Type/Length Byte* field from the Chassis FRU Information (VITA 46.11 Address Table record)
Chassis address	20	Binary	*Chassis Address* field from the Chassis FRU Information (VITA46.11 Address Table record)
Chassis type	1	Binary	0 for VITA 46.11 chassis,
Slot location	4	Binary	See Table 10: Slot Location record for

HPM.3 ClientID benefits

- The HPM.3 ClientID uses geographical addressing:
  - It does not represent the piece of HW (as for MAC address) but location of device: e.g. IPMC of board in slot 10 of shelf X
  - Replacement of boards/components on failure require no intervention in DHCP
  - The intervention required is on crate replacement where the Shelf Address has to be reset inside the shelf FRU. The “**PICMG HPM.3 R2.0, DHCP-Assigned Platform Management Parameters, May 4, 2016**”, section 2.6, paragraph 74 specifies:

    “The content of an xTCA Shelf Address (with its corresponding Shelf Address Type/Length Byte) is typically specified when a Shelf is provisioned or re-provisioned.”
- The HPM.3 ClientID is hierarchical and structured (see later)
- Within a crate, all values are constant except Slot Location

Networking: selective connectivity



- Selective connectivity, isolation of sub-systems
- Gateway PC may become server for buffering, log collection, etc… with high BW connection (100Gb) to ATCA crate switch