INAF

ISTITUTO NAZIONALE DI ASTROFISICA
NATIONAL INSTITUTE FOR ASTROPHYSICS

LFAA SPS Network

G. Comoretto

Rev. 2.2

30/01/2020

DOCUMENT HISTORY

Revision	Date of Issue	Engineering Change Number	Comments
01	2020-30-01	-	First Release
02.2	2020-28-02		

DOCUMENT SOFTWARE

	Package	Version	Filename
Wordprocessor	Libreoffice		SKA_LFAA_RACK_network.v2-2.odt
Block diagrams			
Other			

CONTRIBUTOR DETAILS

Designation	Name	Organisation
Primary Author and Responsible Organisation	Gianni Comoretto	INAF
Contributors	Jader Monari	INAF
	Cristian Albanese	Sanitas
	Sandro Pastore	Sanitas
	Alessio Magro	University of Malta
	Riccardo Chiello	University of Oxford
	Carolina Belli	INAF

Copyright

Document owner	INAF Osservatorio Astrofisico di Arcetri
	This document is written for internal use in the SKA project

Table of contents

1 Introduction

1.1 Purpose of the document

This document includes some initial considerations on how to:

- organize the network structure in the LFAA SPS, both for the control network and for the data transport network
- assign MAC and IP addresses
- manage the LFAA network.

This is just a starting point, any correction or modification is welcome.

The document is based on the LFAA ADD [RD1] and SPS DDD [RD2] documents, which provide a general network architecture, and details this description to a level suitable for actual implementation.

1.2 Scope of the document

Document is part of the detailed specification for the SPS management software. The software interacts with the MCCS control software, but implementation of the MCCS network is not considered. The SPS network interacts with the CSP receiving ports, and the appropriate interface must be included in the LFAA-SPS ICD [RD3].

1.3 Intended audience

This document assumes pre-knowledge of the LFAA SPS architecture, and of the implementation of the relative components (cabinet, Cabinet Management Board, Subrack Management Board, Tile Processing Module). Relevant aspects are recalled when appropriate.

1.4 Document overview

Section 2 describes the problem, and the requirements that a network architecture must satisfy.

Section 3 describes the proposed detailed solution

1.5 Reference documents

RD1.	SKA-TEL-LFAA-0200026: LFAA Architecture Design Document
RD2.	SKA-TEL-LFAA-0500035: LFAA SPS Detailed Design Document
RD3.	100-000000-004-03: LFAA to CSP Interface Control Document

2 Problem and requirements

LFAA network is described in RD1 and RD2. The overall network structure is shown in Figure 1.

SPS is organized into 256 SPS cabinets, each one implementing the beamforming electronics for two stations. 148 cabinets are hosted in the Central Processing Facility, and the remaining 108 into 36 Remote Processing Facilities. Cabinets are grouped into groups of 4 (CPF) or 3 (RPFs) providing some backup routing and the possibility of reducing network complexity.

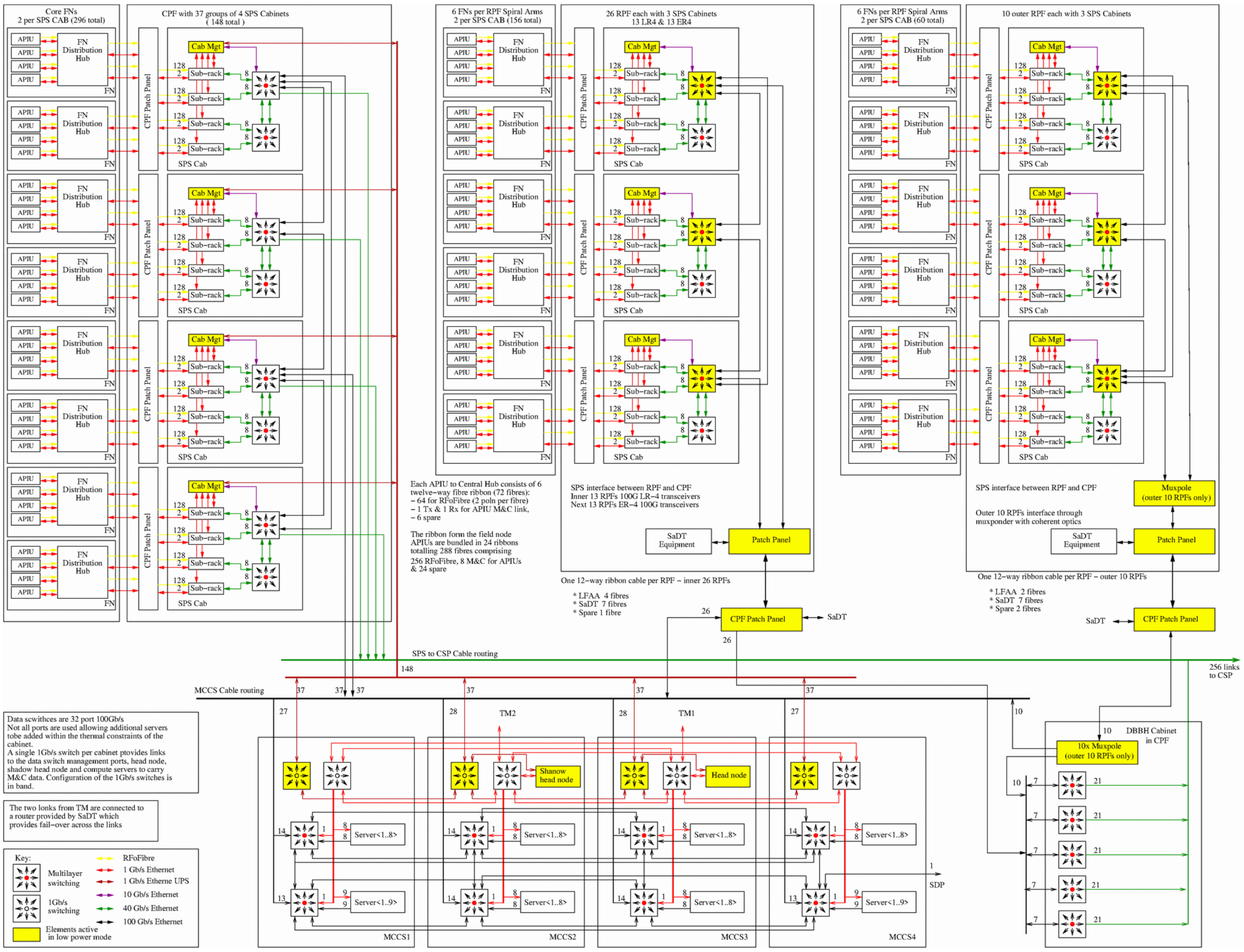
The network serves two purposes:

- Local Monitor and Control: MCCS monitors the status of all equipment, and provides calibration, pointing, and sequencing commands to control the system and to execute the astronomic observations. Expected data rate is modest, even when 8192 individual TPMs must be controlled.
- Data flow: partial beams are produced by each TPM, summed together into a station beam, and transmitted to CSP. Calibration spigots and transient buffer samples are also sent from TPMs to the MCCS. Data rate is high inside each cabinet (23 Gb/s per TPM), between the cabinet and CSP (23 Gb/s per cabinet) and between the cabinet and MCCS (22-27 Gb/s per cabinet).

Data flow uses a 40G/100G network. Two high speed switches are present in each cabinet. One of the switches is connected to MCCS and to SPS, using two separate links.

LMC flow uses a standard 1Gb network inside each cabinet. This network is connected to MCCS both with a 1Gb link (active in any condition) and through the Data network when the high speed switches are active.

Front Node equipment is seen as a single node per station (TBC), which is physically connected through a 1Gb optical SFP connection. The node is part of the cabinet net but is controlled directly by MCCS. SPS acts only as a routing path.



Each APIU to Central Hub consists of 6 twelve−way fibre ribbon (72 fibres):
− 64 for RFoFibre (2 poln per fibre)
− 1 Tx & 1 Rx for APIU M&C link,
− 6 spare

The ribbon form the field node APIUs are bundled in 24 ribbons totalling 288 fibres comprising 256 RFoFibre, 8 M&C for APIUs & 24 spare

SPS interface between RPF and CPF
Inner 13 RPFs 100G LR−4 transceivers
Next 13 RPFs ER−4 100G transceivers

One 12−way ribbon cable per RPF − inner 26 RPFs

* LFAA 4 fibres
* SaDT 7 fibres
* Spare 1 fibre

SPS interface between RPF and CPF

Outer 10 RPFs interface through muxponder with coherent optics

One 12−way ribbon cable per RPF − outer 10 RPFs

* LFAA 2 fibres
* SaDT 7 fibres
* Spare 2 fibres

Data scwithces are 32 port 100Gb/s
Not all ports are used allowing additional servers tobe added within the thermal constraints of the cabinet.
A single 1Gb/s switch per cabinet ptovides links to the data switch management ports, head node, shadow head node and compute servers to carry M&C data. Configuration of the 1Gb/s switches is in band.

The two lonks from TM are connected to a router provided by SaDT which provides fail−over across the links

Figure 1 General architecture of LFAA network

The physical structure of the cabinet network is shown in figure 2

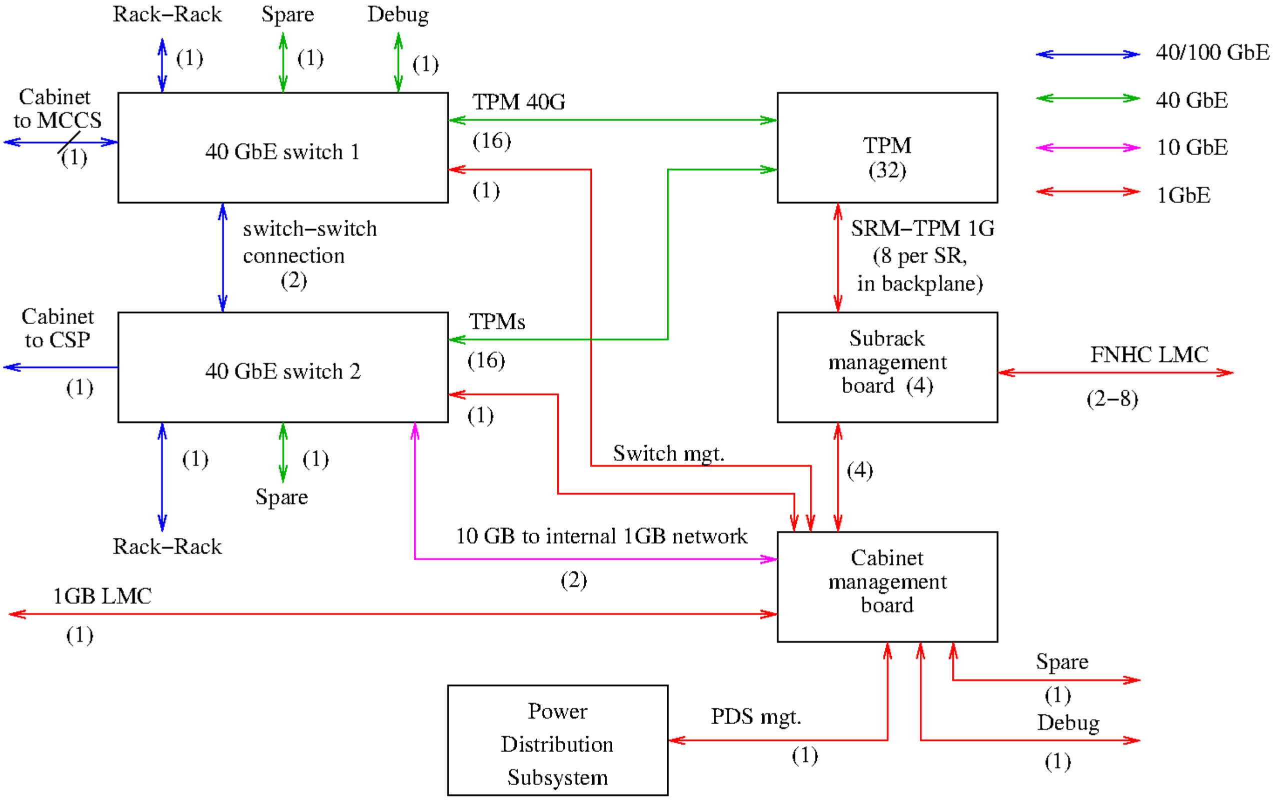


Figure 2 Cabinet physical network

2.1 Requirement rationale

MCCS must be able to correctly address resources for each antenna and station. These resources are identified by their IP address, on the software side, and by their physical location in the CPF or RPF, on the hardware side. The physical to software mapping must be stable and predictable.

All devices have a serial number, which must be visible, in human and machine readable format, in a label sited in an easily visible location (SKA-SYS_REQ-2573). Devices which can be addressed on the net have one or multiple ports, each one with an unique MAC address. At least one port per device has a factory set MAC address, which is also reported on a label. Some ports (e.g. high speed interfaces on FPGAs) have configurable MAC address, which is arbitrarily assigned in firmware, by the control software.

If a device has multiple ports, usually one will be considered the “main” one. This is the one which allows an external controller to check (or set) the remaining ports.

The MAC address and IP address of each interface must be defined in a unique way, after power up. A list of serial numbers and MAC addresses will be maintained in a central database, as part of the Logistics Analysis and Configuration structure. This is however a very error prone procedure, if it is the only way to map physical locations to hardware. So an independent way to assign IP addresses is highly desirable.

The IP/MAC assignment procedure must:

- minimize operator work
- minimize possibility of errors

- guarantee that no deadlock situations occur
- allow exchanging of identical boards without significant operator work
- allow unique addressing based on geographic position within cabinet
- minimize (or avoid) identifications based on the MAC address. MAC address in the Configuration structure will be cross checked, and inconsistencies reported (and resolved).

The operator must follow a simple and easy procedure when installing or moving the boards, with a way to check that everything is correct.

In [RD2] it is suggested that the SPS network is split into separate class D subnets, with each subnet including a group of 4 (or 3) cabinets. As there are more than 100 IP nodes per cabinet, it is more appropriate to have a single network per cabinet. Separating the LMC and Data networks in each cabinet is also advisable, to prevent data flooding from a defective FPGA to stop the LMC functionalities. This implies that cross network (level 3) routing is implmented in the CMB and in the cabinet switches.

2.2 Requirement list

A tentative requirement list for the network structure is:

1. Each element in the system shall be marked with a label displaying a unique Serial Number and Part Number. The label should normally be in an easily visible location but may also include embedded identification for items such as embedded firmware/software.
2. Each element which can be addressed only via Ethernet shall have a principal interface with a unique MAC address, factory set.
3. MAC address of the principal interface shall be reported on a label. The label should normally be in an easily visible location.
4. Labels shall be both human and machine readable.
5. All interfaces in each element shall have a MAC address than can be inquired from the main Ethernet port, or set by command on the main Ethernet port, or set using a separate mechanism which allows geographic addressing.
6. All elements which do not have an hardwired geographic addressing shall have a simple way to identify itself.
7. Operator settable identification hardware shall have a simple visual feedback, e.g a display unit
8. Network shall be partitioned at cabinet level, to avoid spread of network errors
9. Network configuration can be completely managed inside each cabinet. Cabinet Management Board provides base configuration without relying on MCCS, if the MCCS is not available, and checks/adjusts it when communication with MCCS is re-established

3 IP/MAC address assignment mechanism

The following scheme is proposed, to implement the above requirements. Additional, implementation dependent requirements are:

- Separate level 2 networks are present in each cabinet for LMC (1G net) and beamforming (40G net)
- LMC net should be accessible when network switches are offline
- A malfunctioning network interface should not block network outside its subnet
- IP numbers should be assigned locally based on geographic information (physical position within cabinet, cabinet ID)
- Network topology should not rely on specific wiring: no changes should occur by exchanging two identical cables on switch ports
  - This has some limitations. Ports to FNHC may be enabled on specific interfaces. 40G ports to TPM in one station must be connected to the same switch, to avoid traffic bottlenecks. TPMs use only one specific 40G port each.
- The cabinet 1G network must be accessible from MCCS (and viceversa).
- TPMs (40G net) should have access to DAQ service in MCCS, and to the CSP port
- CSP port is on the cabinet subnet (desirable). Alternatively, it is set by MCCS together with adequate routing information in the routing switches
- For RPFs, routing must be added to properly route MCCS and CSP connections through the single high speed link.
- TPM's generate dummy short packets (directed to CSP or MCCS DAQ) to keep MAC/IP table updated in the switches. This allows cables to be connected arbitrarily to TPMs, with no need for switch management for MAC/IP-to-port mapping. These packets must be discarded by the receiver, or by the net
- Improperly formed or badly addressed packets should be discarded by the switches

3.1 Network structure

Each cabinet has a local subnet, with netmask /24, in the 10.x.x.x local address space. This may be further partitioned into 2 subnets with netmask /25, for the 1G local management and control and for the 40/100G fast net. Bits 14:23 of the address derive in a direct and deterministic way from the cabinet ID. Simplest solution is assuming that these bits are the 10 bit cabinet ID. Bits 8:13 of the address are assigned in the SKA-LOW network.

Routing is performed by creating L3 subnets in the two 40/100 Gb switches, operated as a single entity.

Cabinets are uniquely identified by a 10 bit ID, in the range 000 to 999 (to be representable as a 3 digit decimal number), which is reported on a label on the cabinet chassis. Consecutive cabinets in a cabinet row, or in a RPF, have consecutive cabinet numbers.

Each cabinet has a net address space of 8 bits, in the 10.x.x.x private network space. Bits 8:13 of the net address are fixed, while bits 14:23 correspond to the cabinet. The net space shall be partitioned into two separate nets:

- 10.x.x.0/23 for LMC (mainly using 1GbE connections)

- 10.x.x.128/23 for high speed data.

Addresses in the LMC net are assigned by the CMB, which implements a DHCP server. DHCP clients are identified by their hostname. Addresses within a subrack are assigned applying an offset form the main SMB IP address, assigned via DHCP.

MCCS has its own network structure. Cabinet and MCCS nets are connected using two physical routes:

- Direct link with CMB. This link is always active
- Link to the network switches. This link becomes operative after the switches are properly configured

CMB obtains an IP address on this network using DHCP. IP address for the second link can e assigned using DHCP or directly by MCCS, through the CMB.

IP nodes on the FNCH and on the CSP have IP address in the cabinet LMC and data networks, respectively.

The proposed network structure is shown in figure 3.

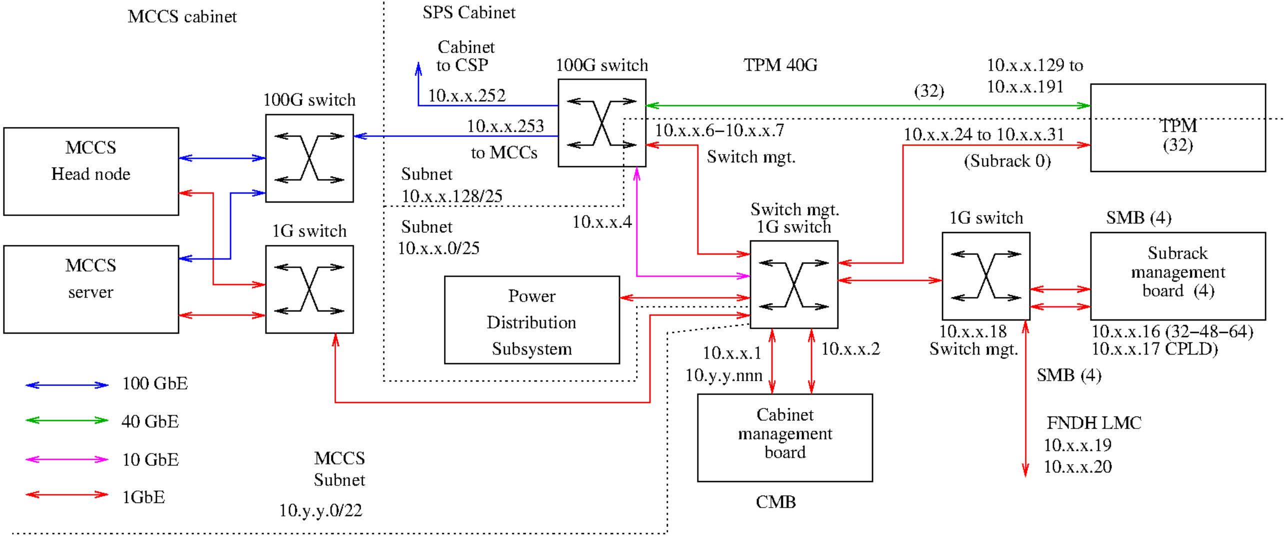


Figure 3 Logical partition of the cabinet network

3.2 IP address assignment: cabinet

Network management inside the cabinet is performed by CMB. Each CMB has a front panel device, which specifies a 10 digit number, interfaced via I2C. The operator sets this number on the front panel, using a combination of pushbuttons. The number is displayed in decimal on a 3-digit display. The number should match a corresponding cabinet number, in the cabinet frame, ad is set consecutively for adjacent cabinets. In this way it is immediate to check that numbers are assigned correctly.

Cabinet sets its hostname as “cabinetNNN”, with “NNN” the cabinet number (leading zeros included). On power-on the cabinet performs a DHCP query based on its hostname, on the 1Gb dedicated interface to MCCS. The on-board Ethernet switch is configured to have the switch port to the MCCS as part of a VLAN, using a shared physical interface to the CMB CPU. This latter port is shared with the cabinet 1Gb subnet, and since DHCP requests are performed separately for the

CMB (on the MCCS wide net) and for the other cabinet elements (on the cabinet LMC net) the two nets must be segregated.

At the same time the CMB starts a DHCP server on the cabinet 1Gb net. Net address includes a base address in the 10.x.x.x private network space, and 10 bit cabinet number. Variable portion of base address is 6 bit long (address bits 8:13), and is specified in a configuration file. MCCS may set or change this base address.

The PDU is the only device available at startup. CMB starts each device in the cabinet by enabling the corresponding power outlet in the PDU. Each device configures its network control port placing a DHCP request on the cabinet LMC bus, exposing its hostname. Internal configuration of each device (in particular sub-racks) is performed autonomously.

Other elements in the cabinet net can be identified by hostname. Network switches may have their hostname (on the management port) changed to “switch0” and “switch1”, with “switch0” the one having a direct connection to MCCS.

The cabinet PDS can have a hostname “pbs”. If not possible, the PBS MAC address is used. The same scheme is used for all other addressable elements in the cabinet, outside the subracks.

The FNCH equipment can be identified either by MAC address, or also setting its hostname. A good candidate could be “stationNNN” with NNN the station number. Cabinet to station mapping is fixed, and kept in a table distributed to all CMBs.

IP address in the CMB switch and CMB CPLD are set directly by the CMB CPU.

Each CMB/SMB board has a factory assigned MAC address. This is reported on a machine (Q-code or bar code) & human readable label. MAC address is factory set, is associated with the board serial number and is never changed by the operator. It can be changed as part of board reconditioning, but this is equivalent to replace it. When the board is installed, board ID and MAC are machine read (using an optical scanner) and associated to the ID of the cabinet or of the subrack crate (also scanned). This is used mainly for maintenance. Front panel ID number is set by the operator using the pushbutton interface. A mechanism to avoid unintentional change shall be provided (e.g a mechanical lid or a long push of an “enable” button).

3.3 IP assignment: subrack

SMBs use a similar mechanism to determine its hostname. Each SMB has a (smaller, or less populated) similar front panel device, with one digit. Digit may assume values from “0” to “3”, which must correspond to the subrack position in the cabinet. During initialization, SMB CPU sets its hostname “subrack0” to “subrack3”, and configures its main IP address using DHCP. Then CMB sets the IP address of its components using fixed offsets from its main address.

Each board inside the subrack has also a simple way to geographically identify it.

- TPMs are identified using a I2C connection on the backplane. The I2C connection allows to sense the TPM presence, and for setting up board parameters, including IP addresses.
- FNCH performs a DHCP request to the CMB as the link becomes active

The SMB uses three addresses, for the CPU, the CPLD and the internal 1G switch. The switch is not specifically configured, as it provides no routing/management functions.

Each TPM has two ports on the LMC net, for the TPM CPU and CPLD respectively. The network source parameters (source port, MAC and IP addresses) for the ports on the data network are set by

SMB on start up. Destination parameters are set using the standard LMC programming interface by MCCS.

3.4 Addresses required in the cabinet networks

Each cabinet requires a number of IP addresses on the local LMC cabinet net, on the local data network ((126 addresses available on each net) and on the MCCS network. This latter is not yet fully defined, but no particular assumptions are required.

3.4.1 Addresses required on the LMC cabinet network:

Each CMB has

- 1 base address on the cabinet LMC network
- I address for the board CPLD
- 1 address for management of the board network switch

The cabinet needs also addresses for

- 2 (minimum 1 per station) to 8 (total available ports) addresses to FNCH
- PDU
- two 40G switches for switch management
- The switches need also an address to implement a gateway to the MCCS network.
- 1 spare address for debug (possibility to use an external PC for debug/monitor)

Each subrack (4 total) has

- 1 base address on the cabinet LMC network
- I address for the board CPLD
- 1 address for management of the board network switch

Each TPM (32 total) requires the following IP addresses:

- Control port (1G)

3.4.2 Addresses required on the cabinet data network

At the cabinet level:

- 1 address for the CSP ingest port
- 1 address for the gateway to the MCCS network
- 1 spare address for debug (possibility to use an external PC for debug/monitor)

Each TPM requires

- 40G MAC: 2 (1 per FPGA)

TPMs require a MAC (and possibly an IP) address on a number of service interfaces (backplane, FPGA-to-FPGA). While these interfaces use Ethernet, they are in no way exposed to the external networks. So fixed addresses (e.g. a private network, with static MAC and IP numbers, the same for all subracks) can be used with no conflicts.

3.4.3 Addresses required on the MCCS network

- One address for the CMB
- One address for the switch gateway.

3.5 External connections

Rack-rack connections group together 4 cabinets (for CPF) or 3 cabinets (for RPFs).

1G LMC is part of a flat control network connected to MCCS. It allows conrolling the CMBs even when the switches are not powered.

FNHC LMC is one (or up to 4, to be determined) optical link to the front node electronics. Cabinet provides just a physical link (and routing) to these devices, but does not control them.

Cabinet-to-MCCS and Cabinet-to-CSP links could be aggregated between 2 or 3 cabinets, using rack-to-rack connections. It is advantageous to aggregate these physical ports on a virtual network.

3.6 IP address summary

The requirements above are summarized in the following table

Device	N.	LMC ports	Data ports	Total LMC ports	Total data ports
CMB	1	3		3	
PDU	1	1		1	
40 /100 Gb Switch	2	1		2	
40 /100 Gb Switch gateway	1	1	1	1	1
FNCH (per station)	1-4	2		2-8	
SMB	4	3		12	
TPM	32	1	2	32	64
CSP port	1		1		1
Debug (external PC with interface)	1	1	1	1	1
Total				**60**	**67**

4 IP addressing example

An example of a possible SPS addressing scheme is shown. It is assumed that SPS network starts at 10.20.0.0/14. Cabinet 153 would have assigned the following addresses:

CMB CPU	10.20.153.1
CMB CPLD	10.20.153.2

CMB switch mgt	10.20.153.3
VLAN gateway interface	10.20.153.126
PDS	10.20.153.4
40/100 GB switch #1 mgt.	10.20.153.5
40/100 GB switch #2 mgt.	10.20.153.6
SMB-0 CPU	10.20.153.16
SMB-0 CPLD	10.20.153.17
SMB-0 switch mgt	10.20.153.18
TPM-0-0 1GB port	10.20.153.23
TPM-0-7 1GB port	10.20.153.31
FNHC-0-0	10.20.153.8
FNHC-0-1	10.20.153.9
SMB-1	10.20.153.32
SMB-2	10.20.153.48
SMB-3	10.20.153.56
TPM-0-0 40GB port #1	10.20.153.129
TPM-0-0 40GB port #2	10.20.153.130
TPM-0-1 40GB port #1	10.20.153.131
TPM-0-1 40GB port #2	10.20.153.132
TPM-3-7 40GB port #1	10.20.153.191
TPM-3-7 40GB port #2	10.20.153.192
Interface to CSP (if required)	10.20.153.252
VLAN gateway interface	10.20.153.254

CSP link may have an address in this subnet (this means that each CSP port has a different subnet), or may have assigned a different subnet, with routing table set up accordingly in the 40G switches.