Hardware Specifications and Use Case Description for

J2-DDC Routing System

Revision 1.1

Author: Tuan Duong

AT&T

Revision History

Revision	Date	Author	Description
1.0	9/10/2019	Tuan Duong	OCP-Draft 1
1.1	1/14/2020	Tuan Duong	Added Recommened Management and Fabric Connectivity Map for small, medium, larger deployments

Table of Contents

1. License **(OCP CLA Option)**

Contributions to this Specification are made under the terms and conditions set forth in Open Compute Project Contribution License Agreement (“OCP CLA”) (“Contribution License”) by:

[AT&T]

Usage of this Specification is governed by the terms and conditions set forth in **[Open Compute Project Hardware License – Permissive (“OCPHL Permissive”),] (“Specification License”).**

Note: The following clarifications, which distinguish technology licensed in the Contribution License and/or Specification License from those technologies merely referenced (but not licensed), were accepted by the Incubation Committee of the OCP:

[None].

NOTWITHSTANDING THE FOREGOING LICENSES, THIS SPECIFICATION IS PROVIDED BY OCP "AS IS" AND OCP EXPRESSLY DISCLAIMS ANY WARRANTIES (EXPRESS, IMPLIED, OR OTHERWISE), INCLUDING IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, FITNESS FOR A PARTICULAR PURPOSE, OR TITLE, RELATED TO THE SPECIFICATION. NOTICE IS HEREBY GIVEN, THAT OTHER RIGHTS NOT GRANTED AS SET FORTH ABOVE, INCLUDING WITHOUT LIMITATION, RIGHTS OF THIRD PARTIES WHO DID NOT EXECUTE THE ABOVE LICENSES, MAY BE IMPLICATED BY THE IMPLEMENTATION OF OR COMPLIANCE WITH THIS SPECIFICATION. OCP IS NOT RESPONSIBLE FOR IDENTIFYING RIGHTS FOR WHICH A LICENSE MAY BE REQUIRED IN ORDER TO IMPLEMENT THIS SPECIFICATION. THE ENTIRE RISK AS TO IMPLEMENTING OR OTHERWISE USING THE SPECIFICATION IS ASSUMED BY YOU. IN NO EVENT WILL OCP BE LIABLE TO YOU FOR ANY MONETARY DAMAGES WITH RESPECT TO ANY CLAIMS RELATED TO, OR ARISING OUT OF YOUR USE OF THIS SPECIFICATION, INCLUDING BUT NOT LIMITED TO ANY LIABILITY FOR LOST PROFITS OR ANY CONSEQUENTIAL, INCIDENTAL, INDIRECT, SPECIAL OR PUNITIVE DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS SPECIFICATION, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING NEGLIGENCE), OR OTHERWISE, AND EVEN IF OCP HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

2. Scope

This document describes the high level technical specifications for the Broadcom Jericho 2 based white box elements in Disaggregated Distributed Chassis Routing System architecture (DDC-RS). It also describes possible field of use for such systems. The goal is to provide the contextual background for the ODM partner to use as guidelines to make sound design decisions in building the components for the DDC Routing System.

Because DDC-RS is new to the industry, not all requirements will be known at the outset. As the design and requirements are being worked at the network, system, and software level, learnings will be gained that may drive additional hardware support or capabilities.

The document lists key requirements and constraints that the design must meet but leaves room for options for innovation for the hardware manufacturer in the design/implementation/manufacturing process.

3. Traditional Modular Chassis Routing System (MC-RS)

It is helpful to study the traditional modular chassis routing system architecture first to see that the DDC-RS is an evolutionary architecture.

Figure 1, shows a front and back view of a finished product view of a typical modular chassis routing system from a major OEM.

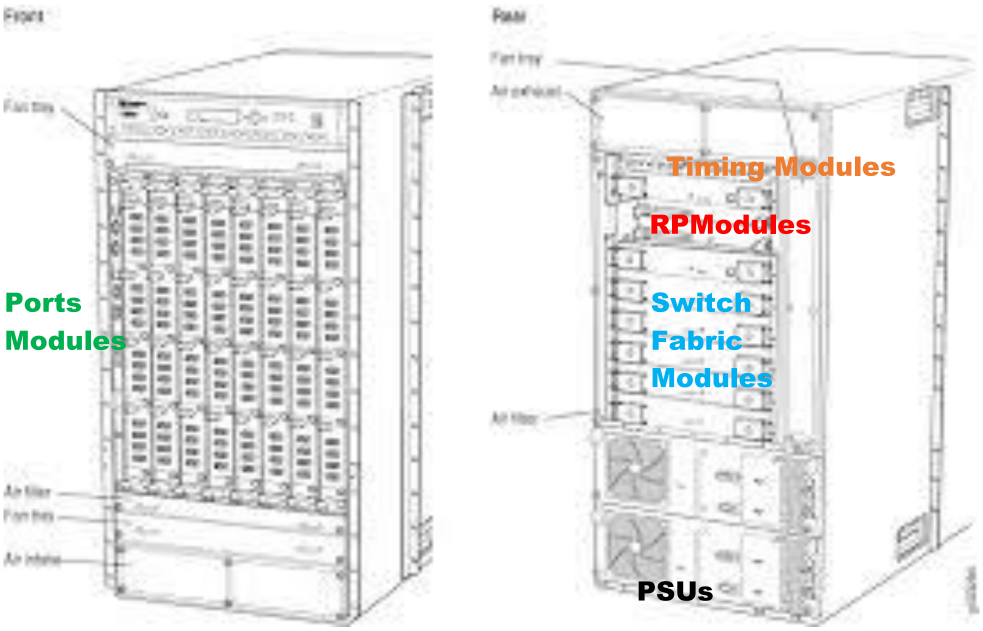


Figure1: Front and Back View of a Modular Chassis Routing System from an OEM.

For example, this modular system supports 8 slots to hold line cards (Ports) in the front. In the back, it supports up to 6 switch fabric modules, 2 Route Processor modules, and 2 Timing Modules. The Chassis provides redundant PSUs and FAN FRUs which are hidden inside the chassis. All these components are packaged in a modular chassis.

Figure 2, shows the major functional components that make up this modular system in slightly more details. These include the following functional blocks:

1. **Line card (LC) Modules** for service ports with a small embedded CPU for distributed control and management of module. In the back there are pins/connectors to the backplane for power and connections to the switching Fabric.
2. **Switching Fabric (FM) Modules** provide the non-blocking bandwidth to switch traffic from one line card to another line card.
3. **Timing Module** provide synchronized timing across line cards and/or switching fabric as needed. It can also provide synchronized clock source across line cards for precision time stamping.
4. **Route Processor modules** are the redundant centralized master computing resources where the Network Operating System runs on to control the behavior of the entire system.
5. **Internal Ethernet Management Switching Module** which is not visible in Figure 1, are used internal control and management communication between the RP and the other modules. In Figure 2, these are depicted as the two purple colored blocks labeled Ethernet {SW0, SW1}.
6. ***Other potential communication paths** are I2C and GPIO PCB traces for low level control of silicon components on each module which are not shown.

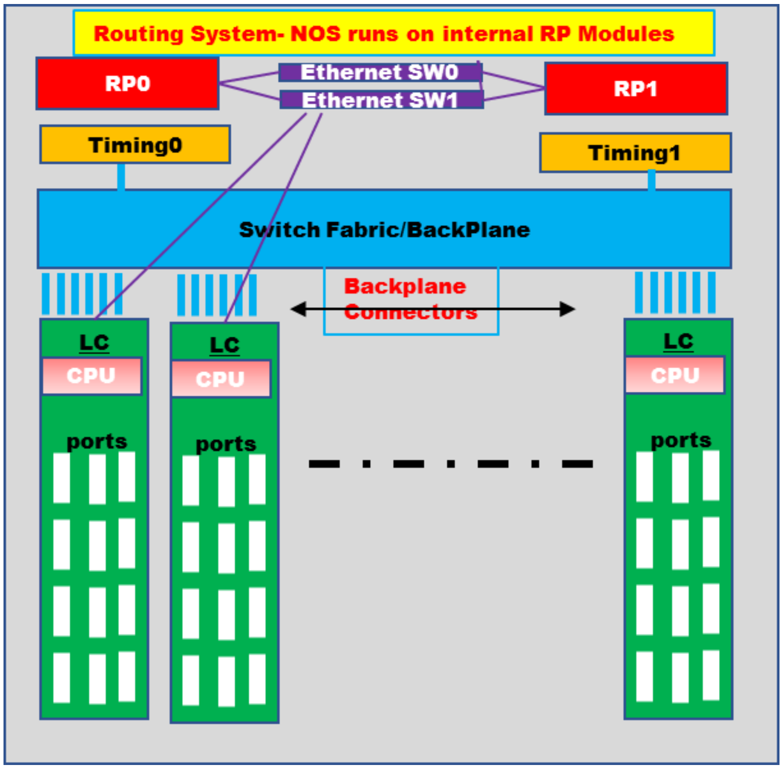


Figure2: Functional Block Diagram of a Modular Chassis Routing System.

In the past, this modular chassis architecture is a cost-effective way to build a larger system because it can make use of copper traces between LC and Switch Fabric. The challenge of a modular system is the mechanical designs (which has Intellectual Property) to solve thermal and spatial challenges and “hot” OIR of modules within the system and still achieve “hitless forwarding”.

The downside is it is a proprietary design and the software that controls such system is also proprietary. With the advent of higher speeds, shrinking process technology, and the push toward lower power per bit, and the trade-off between copper backplane traces versus optical backplane are close to the inflection point.

It is also simpler from a mechanical design standpoint to build individual modules and interconnect them in a Leaf-Spine CLOS topology using optical interconnect. The trade-off here is the cost of the optical interconnect, rack deployment, non-shared power, and managing a CLOS topology of N-independently functioning elements while achieve the desired traffic forwarding performance in a coherent and consistent fashion.

The evolutionary technology is a Distributed Disaggregated Chassis architecture that interconnect like a CLOS but function as a **SINGLE Routing system**.

4. Distributed Disaggregated Chassis Routing System (DDC-RS)

In a Distributed Disaggregated Chassis Routing System, the functional components that were shown in Figure 2 are now redistributed differently as physical boxes or virtualized or containerized functions. Figure 3 shows a high-level view of how such a system can be built.

In this illustration, the DDC-RS consists of a number of fixed-RU Pizza boxes that are provide ports and a number of fixed-RU Pizza boxes that provide Fabric connectivity. The LC (Line Cards) are connected to the FM (Fabric Modules) in a non-blocking CLOS model via external high-speed optics. Each FM and LC module has an embedded CPU that runs NOS software subsystems for distributed control and management of the respective module and communicates to the “Master” NOS for this system running on RP0/RP1. The purple colored external Ethernet switch and connections provide this distributed communications channel. As illustrated, this routing system is implemented over 4 close proximity cabinets.

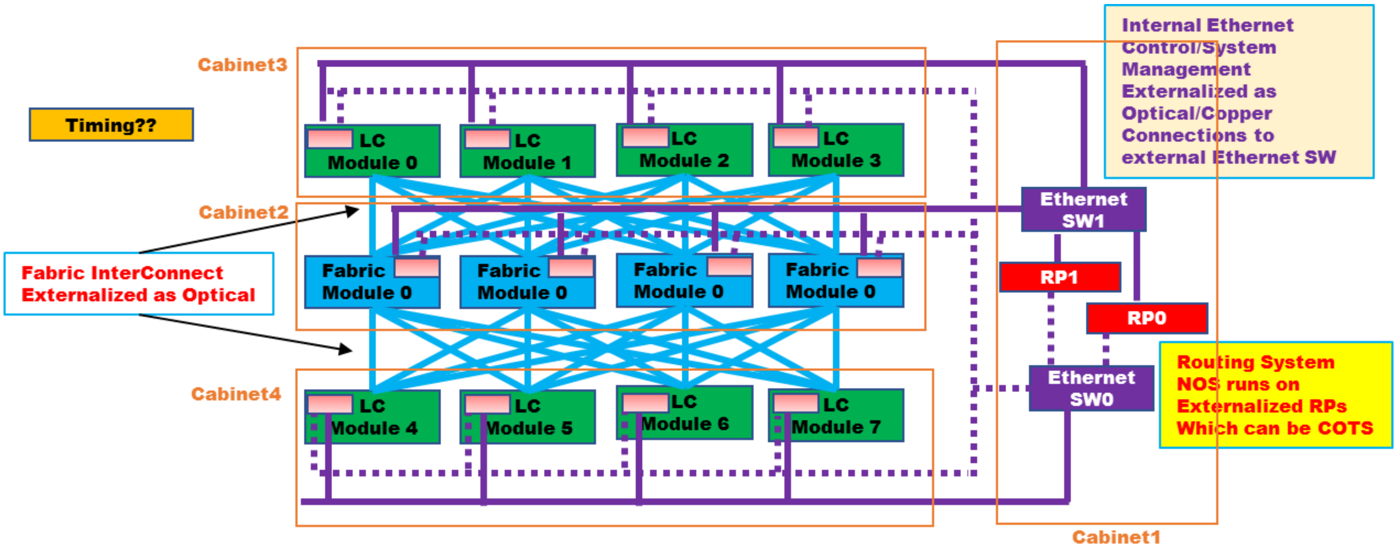


Figure 3: High-Level View of a DDC Routing System

Two things to note in the DDC-RS architecture that is different from the MC-RS architecture:

1. Bullet #6 in the MC-RS does not exist between the RP boxes and the FM or LC boxes.
2. At this point, it is not clear how the Timing function that existed in the Modular Chassis will have a role or be implemented in a DDC-RS architecture. ***This item will need further investigation***. (In the first generation of white boxes for the DDC, we will not address the synchronization requirements to transport 1588 PTP packets across the DDC cluster. We will address this issue in the next generation of J2 WB for the DDC)

The DDC-RS architecture exhibits the following attributes which are trade-offs relative to the MC-RS:

- ➢ Such a system can be spread out into a wider space to reduce the mechanical design challenges to address thermal and spatial challenges. However, there is a cost associated with a cabinet footprint.
- ➢ It does not have the constraints of only so many line cards per chassis. It lends itself well to a scale-out architecture.
- ➢ The NOS now needs to deal with a higher level of distributed systems and potential error conditions associated with External connections.
- ➢ Connections which were once internal PCB traces are now externalized via transceiver connections. There are many of these connections and associated transceiver costs and points of failure and errors. ***As such, we need to build cost-effective hardware hooks for software to leverage to help detect, troubleshoot, identify these potential failures. (e.g. Beaconing)***
- ➢ There are opportunities to design the components of the DDC-RS optimized for the environment they will be deployed. (e.g. Cabinet space clearance, Cooling, Powering)

5. Distinction of DDC-RS from Traditional Data Center Leaf-Spine

Both designs use the CLOS topology. However, the distinction is in how the traffic is spread over the fabric as it ingresses one leaf and egress different leaf over the Spine/Fabric layer.

Traditional Data Center Leaf-Spine	**DDC-RS**
Over-subscription of Leaf to Spine BW	*Equal or Over-Provisioned for Fabric to Leaf Ports BW.*
Under normal operations (non-failure), potentially Blocking or subject to "Elephant" Flows	*Under normal operations, completely non-blocking and **NOT** subject to "Elephant" Flows.*
Packets transiting the Spine-Layer Fabric are variable sized Ethernet Frames	*Packets transiting the Fabric, depending on technology/implementation, are:* *1. Proprietary formatted fixed sized cells or* *2. Proprietary formatted variable sized Frames.*
The logic of spreading of traffic over different Spine links are typically done via control-plane by the NOS which are programmed as routes into the HW tables of the forwarding ASIC.	*The Logic of spreading of traffic over the available Fabric ports are done at the Microcode level of the Forwarding ASIC and is transparent to the NOS. Implementation involves proprietary sophisticated Credits/Tokens mechanisms.*
In the DC Leaf-Spine design, each leaf and spine typically runs a NOS that is independent of each other. As traffic enters each element, the NOS in each element makes an	*In the DDC-RS, all elements { Port modules and Fabric modules} operates under the control of a single NOS. The function of the NOS can be distributed among the*

independent Traffic Load Balancing algorithm. The drawback is that each element lacks a global view so the algorithm is at best locally optimized. Thu, some implementations address this problem by using an SDN-controller approach that has a global view and coordinates the programming of the forwarding tables in the ASIC of ***each element*** *in the CLOS to achieve a global optimization.*	*elements of the CLOS but still under control of one "Brain". From this perspective, the Traffic Load Balancing algorithm is always "globally" optimized through this single routing system. From the end-to-end network perspective, the traffic is just passing through one router – from one port to another port on a different line card or module.*

6. Field Of Use (FOU) Description

Figure 4 below illustrates a high-level diagram of the different network layers and where the P-Core and PE traditionally operate. Typically for the P-Core use case, the nodes will be deployed in pairs per location for redundancy. The diagram currently shows separate boxes for the different PE applications such as L3-VPN, L2-VPN, Internet, and possibly others. Customers typically is singly connected to a PE. So a PE system needs more redundancy than a P-Core. The combination of PE applications into fewer number of boxes will heavily dependent upon NOS implementation and performance and scale requirements of the applications. It is expected that the same hardware will be used regardless of NOS features or implementation. The DDC-RS is can replace the traditional OEM Routers currently used in these FOUs.

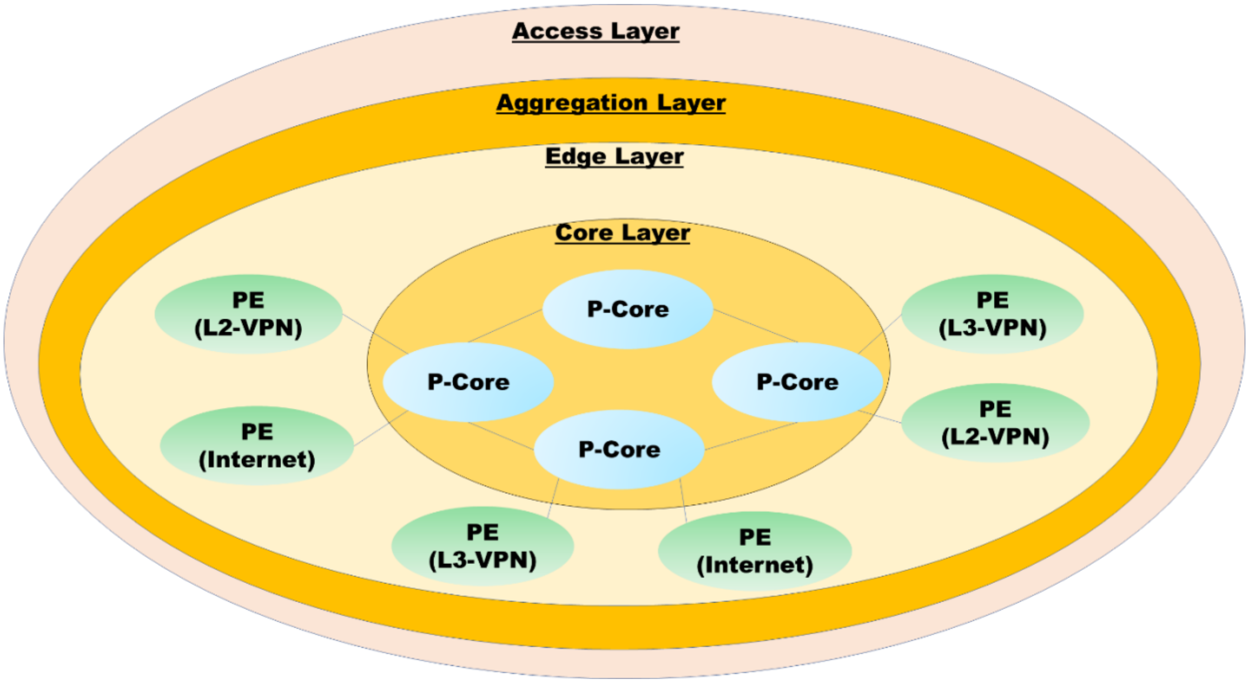


Figure 4 –Potential Field of Use for DDC-RS

7. DDC-RS Hardware Components

After many months of design discussions with Broadcom and the NOS partner, the team concluded that the Broadcom's family of chips {Jericho 2 with built-in HBM deep buffer, the OP2 external

TCAM/Statistical Processor, Ramon Fabric Chip} would meet the required scale and performance requirements for the P-Core and PE use cases. These chips will be used as the main merchant silicon forwarding ASIC in the DDC-RS hardware Components.

This section specifies the high-level description of the hardware components that can be used to build a DDC-RS. More details specification of each components will be described in other sections. ODM can make other hardware models/variations as long as they meet the key requirements in this specification to work in a DDC cluster.

The following table lists current components in the DDC-RS:

Component Name / "Acronym"	Component Specification
DCP100	*{2-RU, EIA 19", Max Depth 30"} with 40x100G QSFP28 (for ports) and 13x400G QSFP-DD (for fabric) pluggables.*
DCP400	*{2-RU, EIA 19", Max Depth 30"} with 10x400G QSFP-DD (for ports) and 13x400G QSFP-DD (for fabric) pluggables.*
DCF48	*{2-RU, EIA 19", Max Depth 30"} with 48x400G QSFP-DD (for fabric) pluggables.*
DCF24	*{1-RU, EIA 19", Max Depth 30"} with 24x400 QSFP-DD (for fabric) pluggables.*
DCM	*{1-RU, EIA 19", Max Depth 30"} with 48x10G/1G SFP and 6x100G QSFP28 Pluggables.*
DCC	*{1-2RU, EIA 19", Max Depth 30"} TP76200/TP450 Level 3 Compliant COTS server with NICs to support required number of 100G interfaces.*

8. DCM and DCC Specifications

This document gives the specifications for the DCP100, DCP400, DCF48 and DCF24. The DCM and DCC are out of scope. For reference, these components are readily available from vendors today and the model number is given here.

DCM = Acton/Edgcore AS5916-54XL

DCC = HP DL380P Gen10 Skylake Server {NEBS Level 3 Compliant}

9. Physical Design Constraints

With the proper cabling the following consideration must be made for the physical platform dimension (does not include handles or SFP). All cabling must be front accessible and adhere to AT&T bend radius standards as specified in ATT-TP76300, section J part 2.10.

- **Width:** 19" rack mount EIA cabinet standard with 4 post mounting.
- **Depth:** Maximum 30" depth to allow for Cabling and Power clearance in a 42" deep cabinet.
- **Height:** 2RU or 1RU.

- **AirFlow:** Front to back.
- **Access:** Front Access for fibers and cables. Rear Access for Power and FAN FRUs.
- **Temperature:** Range {20C to + 50C} Ambient. Typical 26C.
- **Environmental Spaces:** AT&T -{CCS, GTS, MOW Tele-Houses}

The objective of the designs for the DDC-RS components is to be able to mount them in the AT&T standard cabinet which are EIA- 19", 42" deep and 42RU high. This cabinet dimensions are used in the three environmental spaces {CCS, GTS, MOW} that is targeted for these use cases. Figure 5 illustrates the typical cabinet and **minimum** mounting clearances required to allow for cabling and power cables and PDUs. For example, the front rails need to be located minimum 6 inches back from the front of the cabinet. For this use case, it may require 7 inches to allow for sufficient fiber bend radius or fiber densities for example. The rear mounting rails need to have 8 inches minimum. For a 30" deep piece of equipment, this would mean the front mounting ears are flushed and the rear mounting rails needs to be able to adjust to 27 inches {= 42-7-8}. That is why the rear mounting rails need to be adjustable from {26" – 30"}.

The significances of the different environmental spaces is the NEBS compliant requirements, more precisely for AT&T, TP76200/TP450 Level1/Level3 requirements. This will be spelled out in more details in another section.

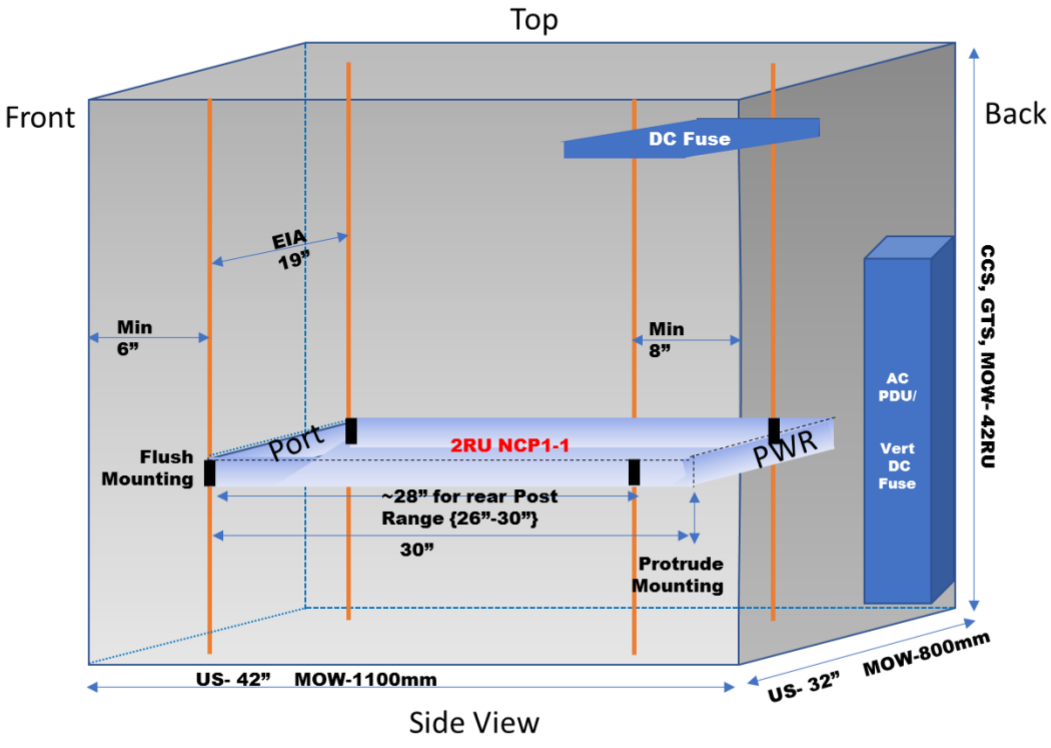


Figure 5: AT&T Standard 42 RU Cabinet, Mounting, Clearance Guidelines.

10. Hardware Compliance Requirements

The components used in the J2 DDC-RS needs to meet the following requirements.

- AT&T TP76200 (Issue 20) & TP76450 (v17) for Level 3.

- Copies of this document and general information about AT&T's environmental equipment standards can be found at https://ebiznet.sbc.com/sbcnebs/

System Functional Block Diagram for DCP100

Figure 6 shows a block diagram of the major sub-system components for the DCP100 40x100G + 13x400G module: CPU, BMC, MAC, FAN, PSU, Interfaces and Interface types.

Reverse GearBoxes:

The reverse gear boxes (BCM 81724) are needed to take one 56G PAM4 Serdes from the J2 and break them down to them down to 2 x 28G NRZ serdes to drive the 100G QSPF28 optics or 4x25G QSFP28 optics with break-out. The reverse gear boxes can also operate in pass-through mode to support 10G interfaces using 4x10G QSFP+ optics. When it is operating in this mode, the 56G PAM4 serdes from the J2 MAC is actually operating at 12G NRZ.

Retimers

Additional Retimers (BCM 81358) may be needed to improve serdes signal integrity due to the complexity of this module chip placement and longer PCB traces.

Front Panel Ports Form Factor and Power Constraints:

- There are 13x400G QSFP-DD ports supporting up to 12W optics for fabric connections.
- There are 40x100G QSFP28 ports supporting up to 5W optics for service ports connections.
- When a 4x10G QSFP+ is plugged into designated ports, then DCP100 can also support 10G interfaces. However, some of the 100G QSFP28 ports will be blocked (unusable as a result). The serdes traces between the J2 MAC and the reverse gearboxes must be designed to support this mode. If only 4x10G QSFP+ optics are used in this module, then a maximum of 80x10G interfaces can be supported.
- A combination of 10G and 100G interfaces should also be supported.

MAC, External TCAM & Statistical Engine:

The DCP100 module will use the BCM88690 with on die 8GB HBM deep buffer memory and the external OP2 BCM 16K processor to support enhanced route scale and statistics which can be used be NOS for different purposes depending upon the use case.

Common Craft & Management:

Other sub-system components such as BMC, Craft and Management interfaces and 2 digits 7 segment LED displays are common to the DCP100, DCP400 and DCF and will covered together in a separate section.

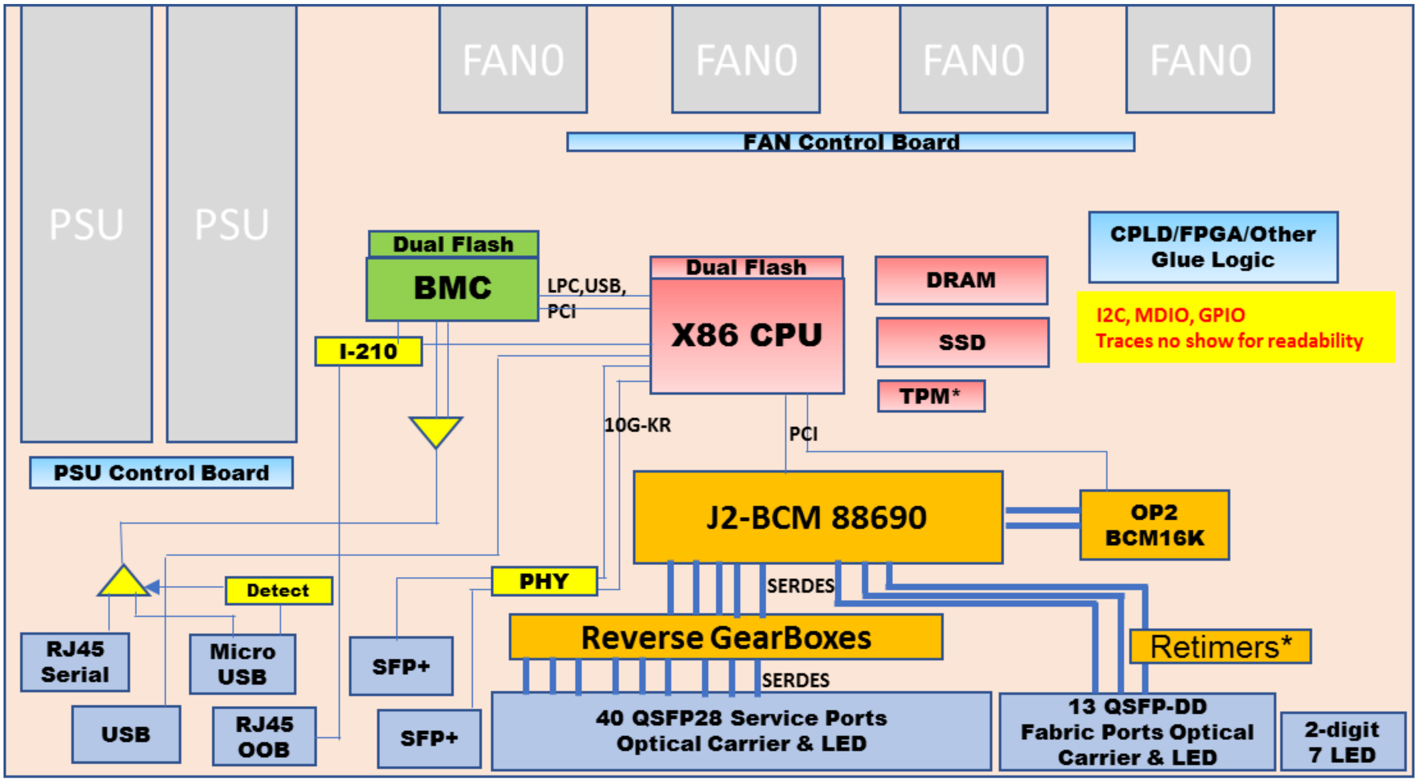
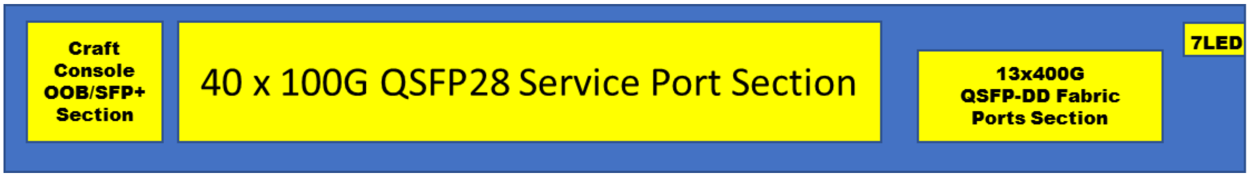


Figure 6a – DCP100 Functional Block diagram

Front Panel Layout:

Recommended front panel layout for DCP100.


Craft Console OOB/SFP+ Section
40 x 100G QSFP28 Service Port Section
13x400G QSFP-DD Fabric Ports Section
7LED

4x10G and 4x25G Break-out Specification for DCP100.

DCP100 will need to support 4x10G or 4x25G QSFP break-out optics to provide connectivity to legacy equipment for various Field of Use. The design of the RGB (Reverse Gear Box) should be such that it allows maximum usage and flexibility.
Example1: a RGB will take in 8x50G PAM4 Serdes from the MAC to provide up to 4 QSFP28 100G ports. When one of these QSFP28 is used in 4x10G or 4x25G breakout, then only one QSFP28 100G port should be blocked.
Example2: The other two QSFP28 can still operate as 100G or as one 4x10G or 4x25G break out with one of the ports blocked.

11. System Functional Block Diagram for DCP400

Figure 7 shows a block diagram of the major sub-system components for the DCP400 10x400G + 13x400G module: CPU, BMC, MAC, FAN, PSU, Interfaces and Interface types.

There are no reverse gearboxes needed in the DCP400 because the ports are all 400G. This is a simpler hardware layout than the DCP100 which has more components. The front panel is also simpler due to a fewer number of front-panel ports.

Front Panel Ports Form Factor and Power Constraints:

There are 10x400G QSFP-DD with up to 15W power for services and 13x400G with up to 12W power for Fabric connections.

External TCAM & Statistical Engine:

The DCP400 module will use the BCM88690 with on die 8GB HBM deep buffer memory and the external OP2 processor to support enhanced route scale and statistics similar to DCP100.

Common Craft & Management:

Other sub-system components such as BMC, Craft and Management interfaces and 2 digits 7 segment LED displays are common to the DCP100, DCP400 and DCF and will covered together in a separate section.

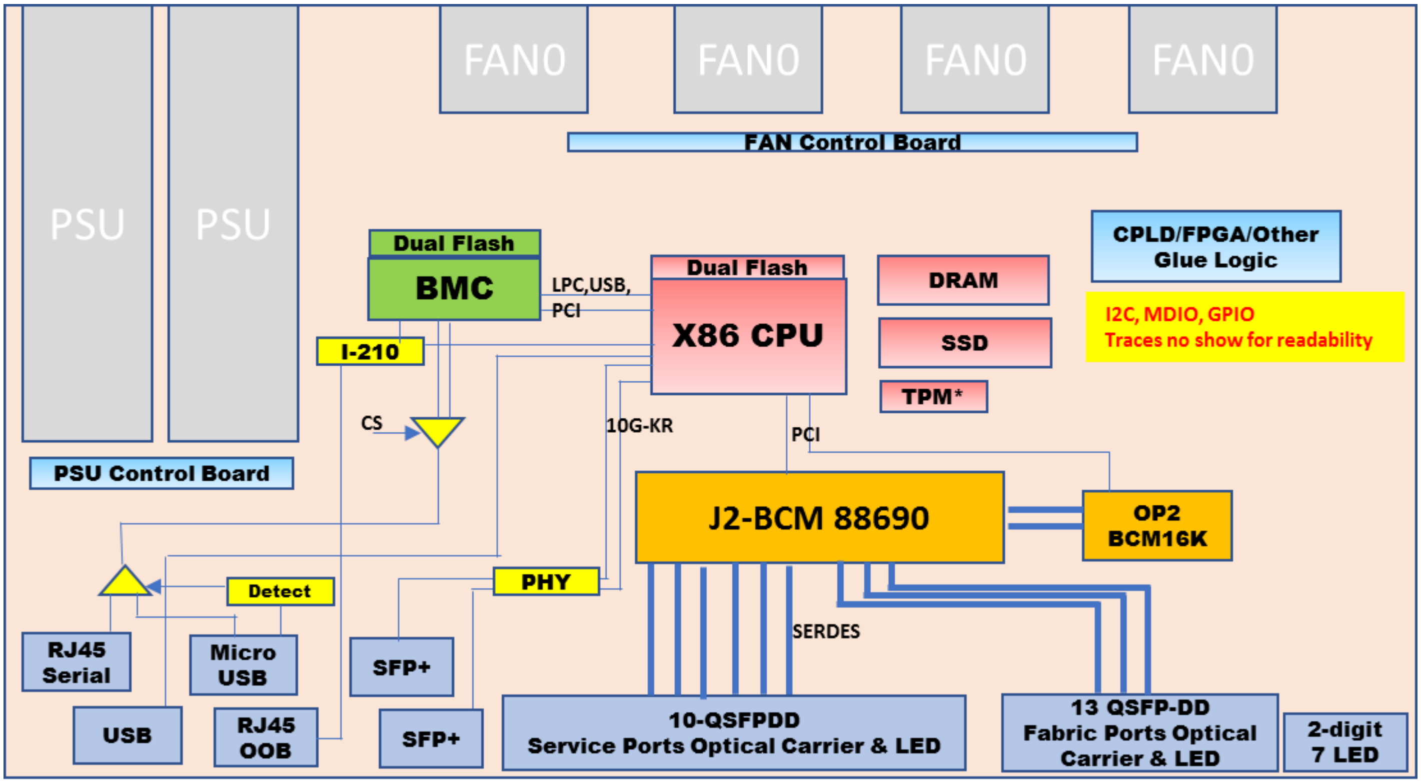
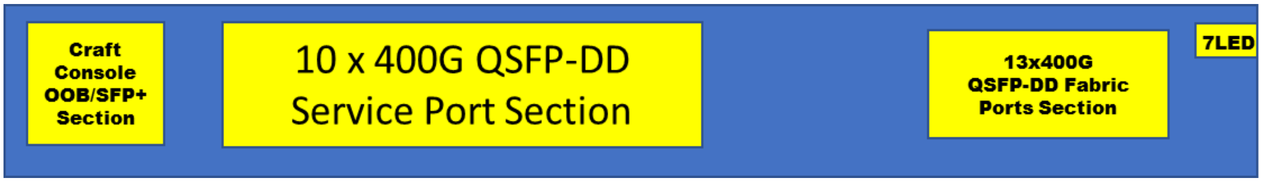


Figure 7a – DCP400 Functional Block diagram

Front Panel Layout:

Suggested front panel layout of DCP400.



12. System Functional Block Diagram for DCF48

Figure 8 shows a block diagram of the major sub-system components for the DCF48 : CPU, BMC, Fabric Chip, FAN, PSU, Interfaces and Interface types.

Front Panel Ports Form Factor and Power Constraints:

There are 48x400G QSFP-DD with up to 12W power for Fabric connections to the DCPxxx modules. Since this is a Fabric Module, there are no service ports supported on this module.

Ramon Fabric

The DCF module will use two Ramon (BCM88790) chips to support up to 48x400G QSFP-DD optics for fabric connections to NCP modules. The 4x56G PAM4 Serdes from each Ramon chip will be interleaved into each 400G QSFP-DD front panel port, thereby the fabric traffic is automatically spread over both Ramon Fabric chips.

Retimers

Additional Retimers (BCM 81358) may be needed to improve serdes signal integrity due to the complexity of this module chip placement and longer PCB traces.

Common Craft & Management:

Other sub-system components such as BMC, Craft and Management interfaces and 2 digits 7 segment LED displays are common to the DCP100, DCP400 and DCF and will covered together in a separate section.

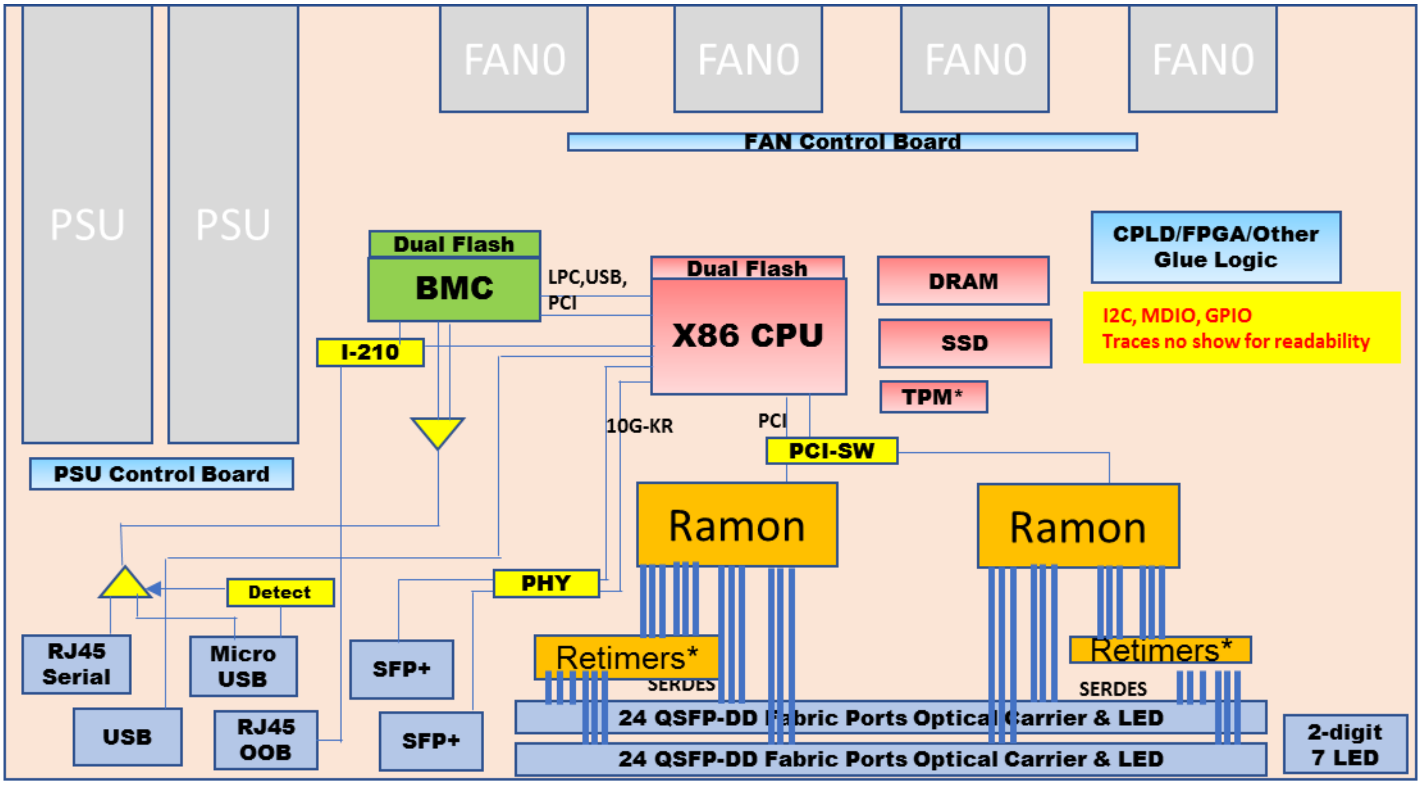
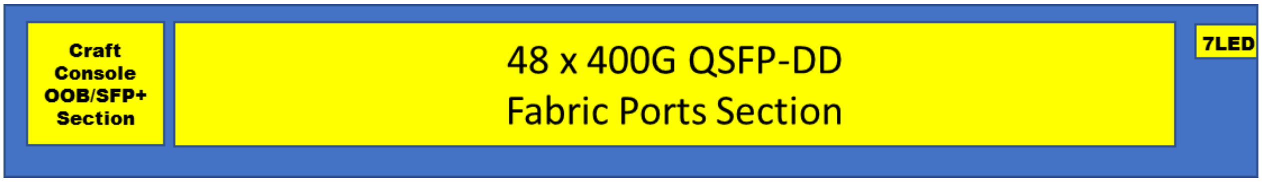


Figure 8 –DCF Functional Block diagram

Front Panel Layout:

Recommended front panel layout for DCF.



13. Port Numbering Specifications

AT&T numbering standard for white boxes starts from zero {0,1,2,….}, increasing from Left to Right. This applies to Ports, FAN and PSU and other FRU (Field Replaceable Units). Numbering starting from Zero also applies to break out ports which is more dependent upon software implementation as opposed to hardware and silk screening implementations.

Manufacturer has a degree of freedom for the numbering with respect to vertical grouping. For examples, the following schemes are acceptable.

NOTE: The port numbering illustration shown in these tables does not reflect the actual number ports specified for the Open CSGR. It is just to illustrate the acceptable numbering scheme.

0	2	4	6	8	10	12	14	16	18
1	3	5	7	9	11	13	15	17	19
20	22	24	26	28	30	32	34	36	38
21	23	25	27	29	31	33	35	37	39

Table 1: 2 grouping numbering scheme: Upper/lower port grouping. Sequentially Numbered Upper grouping then followed by Lower grouping.

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39

Table 2: 4 rows numbering scheme: Sequentially number each row then move to next row.

0	4	8	12	16	20	24	28	32	36
1	5	9	13	17	21	25	29	33	37
2	6	10	14	18	22	26	30	34	38
3	7	11	15	19	23	27	31	35	39

Table 3: 1 Grouping numbering scheme. Sequentially number the ports in a column then move to next column.

As shown in the system block diagram, the numbering for the SFP groups of ports should start from 0 and groups of QSFP form factor ports should also start from 0. The table below illustrates this concept.

0	2	4	6	8	10	12	14	16	18		0	2	4
1	3	5	7	9	11	13	15	17	19		1	3	~~5~~

Table 4: 0-19 are ports of one physical grouping characteristics (e.g. SFP Pluggables). 0-5 are ports of different physical grouping characteristics (Example: QSFP-Pluggables)

14. Common Design Components Across DCP100, DCP400, DCF48, DCF24

As show in Figures 6-8, there are many sub-components of the hardware that are common across the DCP100, DCP400, DCF48, DCF24. These include the following:

- PSU FRU
- Fan FRU
- The Craft, Management, and Console interfaces
- The 7 segment 2-digits LED display/beaconing
- BMC,
- X86 host CPU

15. Fan Module Specifications

The DDC-RS environment will have rear access. Fans needs to be Redundant, field replaceable and hot swappable for the DCP100, DCP400, DCF48, DCF24. Fans are accessible from rear panel.

16. Power Supply Specifications

- ➢ DC PSU accept nominal -48V DC Power.
- ➢ AC PSU operates from 200-240V 50-60Hz input range.
- ➢ The power supplies shall meet the 80 Plus Gold or better for high efficiency.
- ➢ The system shall have less than 1850W total power consumption.
- ➢ There shall be redundant, field replaceable, hot-swappable power supplies.
- ➢ The power supplies should support loss of power indicators to facilitate BMC

- Two grounding screws on rear panel of the system..

Table below shows the PSU efficiency under different loads for different 80-Plus standards.

	20% Load	50% Load	100% Load
80 Plus	80%	80%	80%
80 Plus Bronze	82%	85%	82%
80 Plus Silver	85%	88%	85%
80 Plus Gold	87%	89%	87%
80 Plus Platinum	90%	92%	89%

17. Craft/Management Interfaces

The following table lists the craft and management interfaces that are needed on the DCP100, DCP400, DCF.

Craft Type		**QTY**	**Purpose**
Micro USB Serial		1	Console
RJ-45 USB Serial		1	Console
USB Port		1	USB data access
RJ-45 10M/100M/1G		1	Ethernet OOB Powered on Standby Power rail
10G SFP+ ports		2	For Internal Management Communication for the Modules in a large DDC-RS system

Console:
Only one Serial input can be active for the Console. RJ45 RS232 is higher priority than Micro-USB. RJ45 RS232 will be active if both interfaces are connected by user. The Serial console needs to support default selectable between the BMC or the X86 CPU. (customer-provisioned choice).

OOB Management:
The RJ-45 OOB Ethernet management port needs to be operational even when the system is in the shutdown mode. As such it needs to be designed using the standby power rail. It also needs to provide simultaneous connectivity to the X86 CPU and the BMC. The Intel I-210 NIC is specified to use for the RJ-45 OOB Ethernet Management. Design should allow to shutdown Ethernet OOB access to the BMC as needed.

USB port:
USB port can be used to provide access to external USB drive for initial system setup or system rebuild. However, for security reasons, once the system is up and running under the control of the NOS, the hardware/firmware of the box needs to provide a mechanism to turn off access to this external USB

port. This lock needs to persist even after a cold boot, warm boot, other reset mechanism. To unlock this would require authentication to access the utility to set the registers to unlock it.

2x10G SFP+:
These two 10G SFP+ ports provide connections to the X86 Host CPU. This will be used for communication with the Route-Engine Controller of the DDC-RS when this is one component of the larger DDC-RS.

18. Supported Optics

Needs to go through compatibility/interoperability testing to be on the list.

19. Dying Gasp Guidance

Dying Gasp is not a required feature for this application mainly because of the environment in which these boxes will be used. They are AT&T or Leased facilities with redundant power feeds and back-up power.

20. Internal Glue Logic and Controls

The specification leaves freedom to manufacturer to use any combination of discreet logic/PLD/CPLD/FPGA necessary to implement of the specification. It is recommended to use current practice and provide I2C interface to the Host for control of PSU, FAN, Optics, and any other key components such as reading of registers or erasing and flashing of NVRAM, EEPROM, Flash,….. The manufacturer must provide the instructions and drivers to access these components as part of the BSP-Baseboard Support Package, so that NOS development can access and control these components as necessary.

21. LED Operations Recommendations

Refer to AT&T Hardware Common Systems Requirements for recommendations on system and interface LED colors and operations.

The indicator lamps (LEDs) must convey the information described in 4. The number, colors, and flash behaviors are desired but not mandatory.

Table 4 - LED Definitions (Recommendations)

LED Name	Description	State
PSU1	Led to indicate status of Power Supply 1	Green - Normal Amber - Fault

PSU2	Led to indicate status of Power Supply 1	Green - Normal Amber - Fault
System	LED to indicate system diagnostic test results	Green – Normal Amber – Fault detected
FAN	LED to indicate the status of the system fans	Green – All fans operational Amber – One or more fan fault
LOC	LED to indicate Location of switch in Data Center	Blue Flashing – Set by management to locate switch Slow flashing – System is in standby state
SFP- LEDS	LED built into SFP(28) cage to indicate port status	On /Flashing – Port up (flashing indicates activity) Green – Highest Supported Speed Off – No Link/Port down
QSFP LEDs & Breakouts	Each QSFP28 has four LEDs to indicate status of the individual 10-25G ports	On Green/Flashing – Individual 25G port has link at 25G. (yellow for 10G) Green – Highest Supported Speed Amber – Lower Supported Speed
OOB LED	LED to indicate link status of 10/100/1000 RJ45 management port	On /Flashing – Port up (flashing indicates activity) Green – Highest Supported Speed (1G)

22. Silk Screen Recommendations

It is a strong recommendation that the manufacturer choose Pantones, Contrast, and Font Size that MAXIMIZE visibility to the field technicians working in low light, tight spaces and crowded cabling conditions.

Silk screen should be clear and avoid possible confusion or misinterpretations.

Best is to solicit feedback prior to implementation of silk screen.

23. Number of MAC addresses and Address Constraints.

Each NCP will be configured with a block of 256 MAC addresses.

Each J2 is able to resolve up to 64 different (38MSB of Mac address) blocks of 1024 (10 LSB of Mac Address). In the DDC Large cluster, the maximum number of NCP is 48. So worst case for Large DDC cluster is 48 different blocks of addresses.

24. X86 CPU Specifications

The CPU board needs to be designed as a factory orderable option with different CPU classes. The CPU board should support a range of Intel x86 platform Xeon-D1500 series from 4 Core to 16 cores with a range of DRAM and SSD sizes. Dual flash operating in Primary/Backup mode for recoverable remote field upgrade. All NIC must be supported by DPDK (refer to http://dpdk.org/doc/nics).

Currently, the NOS and AT&T is specifying the

D-1548 with 8-core @ 2.0GHz

2x32G DRAM

1x128G SSD

This specification can be adjusted as NOS development matures and performance testing and fine tuning has been completed.

25. Trusted Platform Module (TPM)

The latest Trusted Platform Module (2.0 or greater) shall be used for secure storage of keys and certificates in a hardware chip and is an integral part of creating a Secure Boot environment so that the device cannot be easily taken over, such as by booting from a USB drive.

The design of the TPM must be a factory orderable option. The reason is in certain use cases, it is not desirable to have the TPM installed due to foreign country import/export rules.

This is a future proof design specification because it is dependent on the availability of ONIE secure boot. If TPM was ordered to be mounted, then Initial software releases will operate with TPM disabled in BIOS and use regular ONIE Boot process.

If the CPU board was ordered without the TPM mounted, then the BIOS must support boot up without the TPM present.

26. Coin-Cell Lithium Battery

Typical X86 design specifies the use of a Coin-Cell Lithium battery to maintain the RTC. However, the presence of this battery conflicts with the TP76200/TP76450 requirements. As such, the Cell Site Gateway Router design does not have a Coin-Cell battery.

However, there is an issue with correct TPM operation, should TPM be activated in the future, without the presence of the battery to maintain the RTC following a power loss. A ticket was documented with Intel and a work around in the BIOS is needed from the manufacturer. This is documented in Intel Ticket # (00260505).

Coin-Cell battery is a factory orderable item. As such, other providers may choose to use the battery.

27. Recessed Reset Button Behavior

When the recessed reset button is depressed for less than equal to 10 seconds and released, then this should cause a warm reset of the whole whitebox.

When the recessed reset button is depressed for greater than 10 seconds and released, then this should generate an interrupt with a vector code to the X86 CPU.

28. Watchdog Implementation

Design needs to include a watchdog mechanism prevent the system from being stuck in a “hanged” state. This can be done via one or combination of the following:

1. Dedicated watchdog circuitry.
2. Intel TCO watchdog
3. Or some kind of watchdog implementation between BMC and X86 host.

The choice of implementation is left to the manufacturer as long as this is documented and appropriate drivers or mechanisms to leverage this capability from the NOS is provided.

Detection of Insertion and Removal of Optical/DAC pluggable modules.

The Whitebox CPLD or glue Logic design should be able to detect insertion and removal of pluggable optical modules on the NIF and Fabric ports and notify the x86 host with an interrupt and a vector code.

29. HW BMC Specifications

The system will be designed with Baseboard Management Controller (BMC) to allow for remote lights out operations, management and access.

The most important requirement for the BMC is that it must be secure. The BMC will be connected to WAN and/or Internet connections. As shown in the System Block Diagram, the BMC has an Ethernet connection which is a shared external connection to RJ45 OOB Management Ethernet Port with the X86 host.

The long term goal is to use Open BMC when Open BMC supports the required features needed for the operational/business mode in a secure way. In the interim, a commercial BMC implementation is acceptable. Regardless of the open or commercial implementation, the firmware must be capable of disabling this Ethernet access when needed based upon the use case or operating environment.

The following requirements are specified for the BMC.

- Dual Flash memory to support remote reliable in-band BMC Software upgrade.
- Power management: On/Off of control system, Host CPU, and MAC where it makes sense. That is, it does not make sense if a component is powered off which completely cuts off access to the BMC to power it back on. In this case, this component should always be powered on the standby power.
- Temperature monitoring
- Voltage monitoring
- Fan control
- Reset control
- Host CPU boot up status
- Serial number / unique identifier
- Board revision ID
- I2C interfaces to Host CPU, USB, temperature sensors, and voltage controllers.
- Monitoring detect signals – including loss of power from the power supplies.
- Must support IPMI 2.0 host mode to provide the following capability via the IPMI interface:
  - ✓ temperature reading and alarms at 3 levels (minor, major, critical) for Processor modules, Chassis, power supply, fans, Broadcom chipset.
  - ✓ status information for fans, power supply, interface modules, processor modules, fan tray

30. Thermal Shutdown

NEBS/TP76200 compliant equipment should have the ability to be configured to shut down when the thermal threshold is exceeded or continue to operate until the equipment fails completely. Configurable means that the “user” can select the thermal overload behavior. This must be set through software. The default should always be to implement equipment shutdown in a thermal event

Shutdown means that all non-essential functions of the chassis are powered off and only temperature monitoring capability remain such that, if the thermal event ends, the chassis will autonomously reboot and restore service. One way of accomplishing this is to have the management hardware command the power supplies to shut off their main outputs but maintain an auxiliary power bus that powers the management/monitoring functionality.

30.1 RESETS (Needs review with UFI)

RESETS in the Design of the DCP100, DCP400, DCF24, DCF48

- Software reset of BMC –
  1. SSH to BMC and issue reset or reboot
  2. Reset BMC from x86 host via IPMI command

- Software reset of x86
  1. SSH into x86 and issue reset or reboot
  2. Reset x86 host from BMC via IPMI command

- Hardware Level reset of BMC via setting on the CPLD on the board
- Hardware Level reset of x86 via CPLD settings.
- Hardware Level reset of x86 or BMC via watchdog timer between x86 and BMC
- Hardware Level reset of the entire white box via 12 volt power rail.

30.2 ONIE

Fulfillment of the ONIE hardware specification as laid out here:

https://opencomputeproject.github.io/onie/design-spec/hw_requirements.html

30.3 BMC Software

Open BMC with Redfish implementation is the target platform. Commercial BMC with IPMI 2.0 is acceptable to meet near term needs.

30.4 DDC-RS Configuration Sizes

With the specified DCP and DCF components, it is possible to systematically build DDC-RS to different scales with differing fabric / oversubscription requirements. To reduce complexity and variations, the following configurations are recommended and proposed fabric and management interconnect. The details are described in the excel workbook: “ATT-OCP-J2_DDC_ConfigurationsConnectivity.xlsx”.

1. Pizza: {1 standalone DCP100 or DCP400}
2. Small1: {1DCF48, 2-4 DCPxxx, 2 DCM, 2 DCC}
3. Small2: {2DCF48, 2-4 DCPxxx, 2 DCM, 2 DCC}
4. Medium1: {7 DCF48, 1-24 DCPxxx, 2 DCM, 2 DCC}
5. Large1: {13 DCF48, 1-48 DCPxxx, 4 DCM, 2 DCC

ATT-OCP-J2_DDC_Co
nfigurationsConnectiv