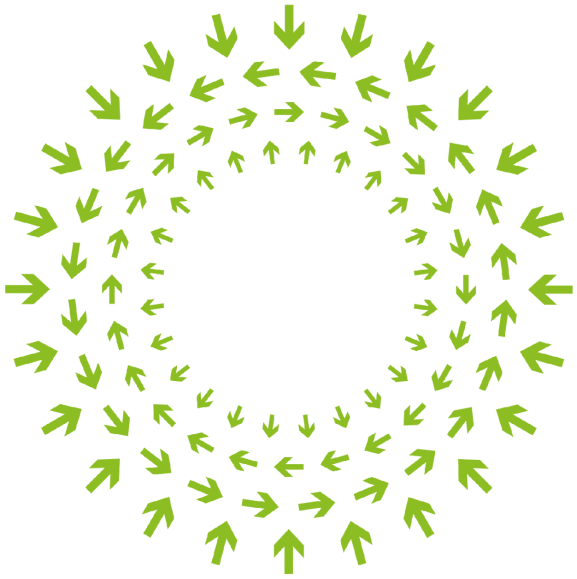


OPEN

Compute Project

Meta Disaggregated Scheduled Fabric Switches Hardware Base Specification

3.0

Author: Lingjun Wu

Table of Contents

1. License 4
1.1. OPTION B: Open Web Foundation (OWF) CLA 4
1.2 Acknowledgements 5
2. Compliance with OCP Tenets 6
2.1. Openness 6
2.2. Efficiency 6
2.3. Impact 6
2.4. Scale 6
2.5. Sustainability 7
3. Change Log 8
4. Scope 9
5. Overview 10
6. IN (Interface Node) System 12
7. FN (Fabric Node) System 13
8. Common modules 16
8.1 CPU Subsystem 16
8.2 BMC Subsystem 16
8.3 Switch Control Module Subsystem 16
8.4 FPGA Subsystem 16
8.5 PSU/PEM and FAN 17
9. Environmental / Regulatory Requirements 18
Appendix A - Checklist for IC approval of this Specification (to be completed by contributor(s) of this Spec) 20
Appendix B - Contribution Process FAQs 21

1. License

1.1. OPTION B: Open Web Foundation (OWF) CLA

Contributions to this Specification are made under the terms and conditions set forth in Modified OWF-CLA-1.0. (As of August 16, 2021) ("Contribution License") by:

Meta

You can review the applicable Specification License(s) referenced above by the contributors to this Specification on the OCP website at http://www.opencompute.org/participate/legal-documents/. For actual executed copies of either agreement, please contact OCP directly.
Usage of this Specification is governed by the terms and conditions set forth in **Modified OWFa1.0 Final Specification Agreement (FSA) (As of August 16, 2021) ("Specification License").**

Notes:

1) The above license does not apply to the Appendix or Appendices. The information in the Appendix or Appendices is for reference only and non-normative in nature.

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.

1.2 Acknowledgements

The Contributors of this Specification would like to acknowledge the following companies for their feedback:

- Celestica

2. Compliance with OCP Tenets

The Meta DSF (Disaggregated Scheduled Fabric) switches are open HW designs with detailed specification to recreate the design contributed to Open Compute Project.

2.1. Openness

This base specification defines Meta's DSF switch hardware architecture. It's submitted to OCP and it's open to others as reference design.

2.2. Efficiency

The Meta DSF switch platform improved efficiency of the system design, and the distributed design concept simplified the chassis design and reliability.

2.3. Impact

The Meta DSF switch specification can be used as reference for others to build the similar disaggregated scheduled network for AI/ML network, which is targeted to lossless network.

2.4. Scale

The Meta DSF switch specification defined disaggregated network architecture, which allows users to extend the network/cluster to big scale more easily than traditional network switches.

2.5. Sustainability

The Meta DSF switch specification defined the environmental requirements in section 9 which includes most of the sustainability considerations for datacenter applications.

3. Change Log

Date	Version #	Author	Description
August 28, 2024	1.0	Lingjun Wu	Initial version.
Dec 10, 2024	2.0	Lingjun Wu	Updated based on OCP Steering Committee review.
Jan 14, 2025	3.0	Lingjun Wu	Updated tenetsbased on additional Committee review.

4. Scope

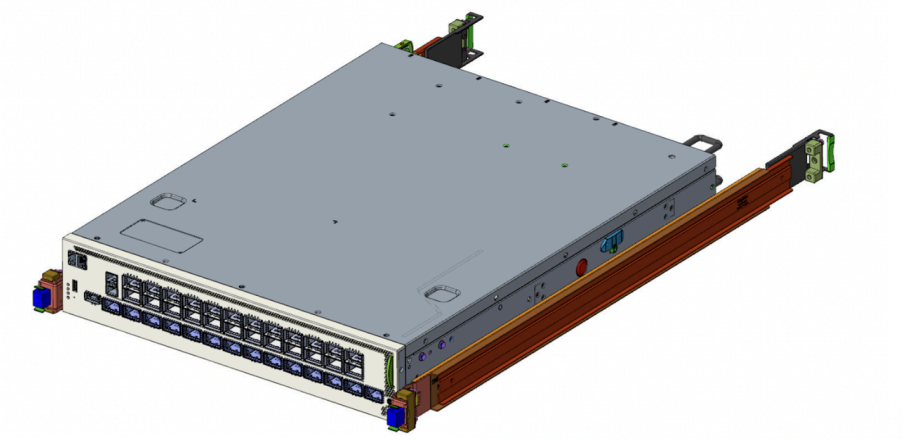
This document defines technical specifications for Meta DSF (Disaggregated Scheduled Fabric) Switch platform used in Open Compute Project. This document, along with the product design package shall comprise product's technical specification.

5. Overview

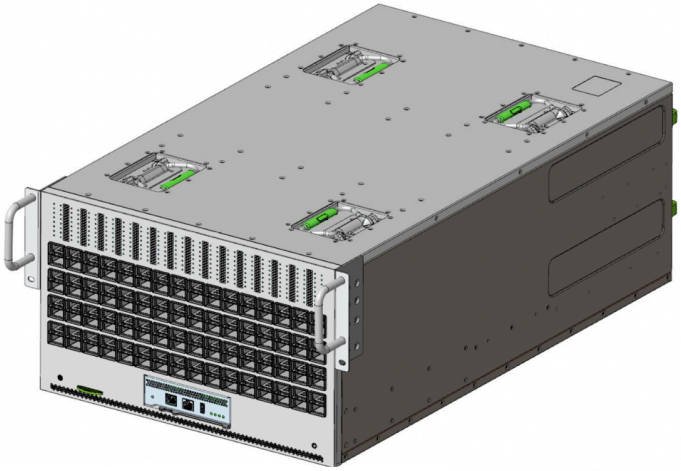
DSF is Disaggregated Scheduled Fabric that Meta defines for the backend network. One of the options is based on BCM88890 and BCM88920 chipsets from Broadcom. There are two hardware systems defined in DSF architecture:

- IN (Interface Node): Based on BCM88890 ASIC; It provides 18*OSFP(800G) Network Interfaces, and 20*OSFP(800G) Fabric interfaces;
- FN (Fabric Node): Based on 2 pcs of BCM88920 ASICs; It provides 128*OSFP(800G) Fabric interfaces;

This is the IN system:



This is the FN system:



With IN and FN systems, the backend network can be scaled up to a big size of clusters to serve Meta's AI/ML racks. Each OSFP port on IN and FN can support both 2x400G-FR4 or 800G-FR8 with 112G SerDes interfaces.

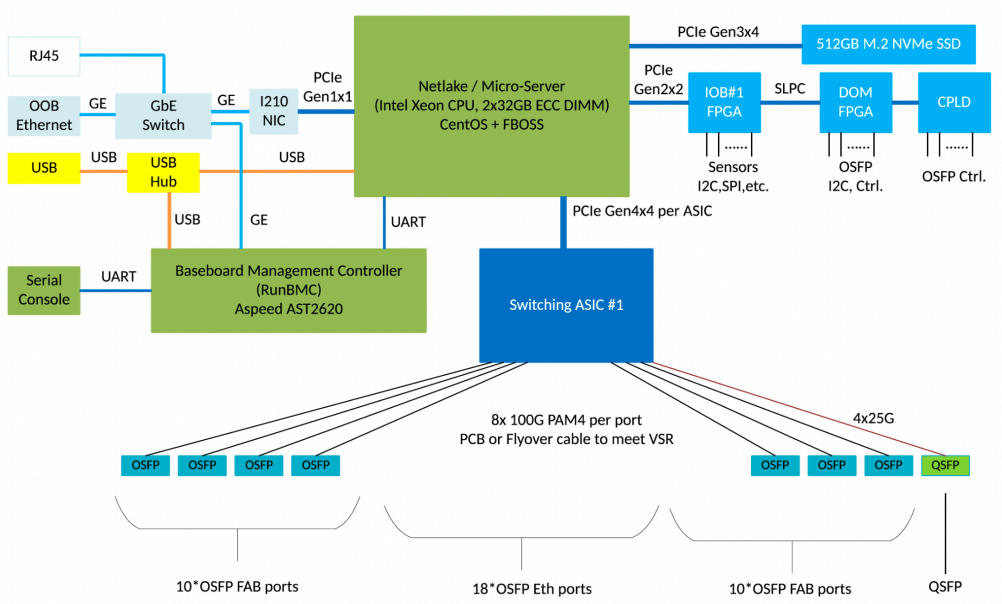
The physical sizes of the IN and FN in the above renderings are for references only and not meant to be binding requirements of this specification.

6. IN (Interface Node) System

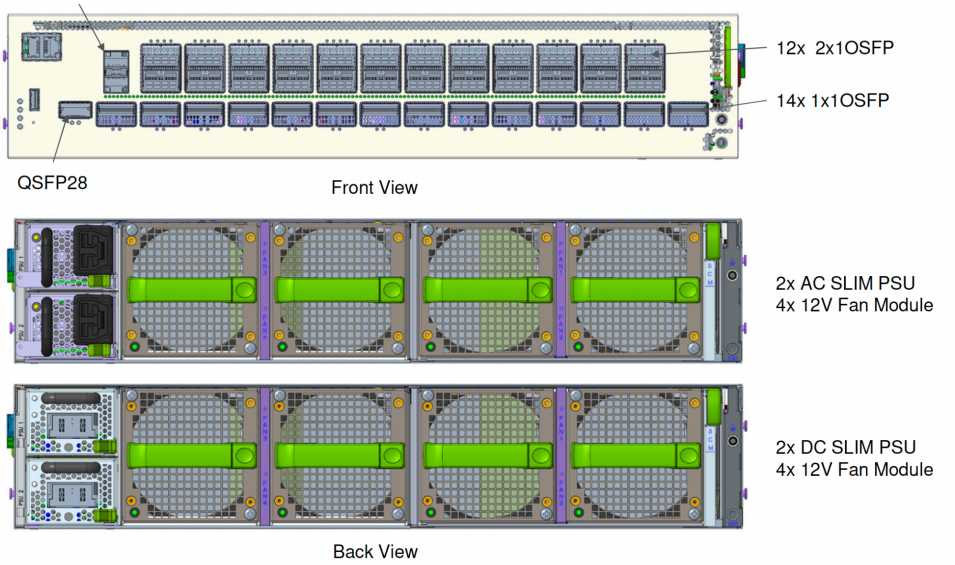
The Interface Node (IN) will be based on BCM88890 ASIC which has 14.4T Network Interface bandwidth. Here is the expected feature list Meta is proposing:

- Based on one BCM88890
- Network Interface
  - 18x OSFP ports (800GE or 2x400GE)
- Fabric Interface
  - 20x OSFP ports
- Management Interface
  - 2x RJ45 1000BASE-T ports, 1x RJ45 console port, 1x USB Type-A port
- Intel Ice Lake-D CPU
  - LCC, 8 cores, 2.3GHz, 55W
  - 2x 32GB SODIMM w/ ECC, 1 DPC, 512GB NVMe SSD
  - External NICs (I210)
- Aspeed AST2620 BMC
- 2RU, 19” rack (~17.3” width)
- Dual AC or single DC power feeds
- Typical power consumption at full load: 1,300W (1+1 2,000W PSUs)

The system diagram below shows the major components and their logical connections for the system:



The front panel port mapping/port placement and mechanical design support 18*OSFP Network Interfaces on the left side, and 20*OSFP Fabric Interfaces on the right side. The Switch Control Module is a FRU which can be plugged from the rear side (together with 2 FAN modules). There are two PSU slots on the rear side. When the system is configured as AC input, there will be 1+1 PSU for AC input and PSU protection/redundancy. When the system is configured as DC input in ORv3 rack, there will be only one PSU/PEM installed.



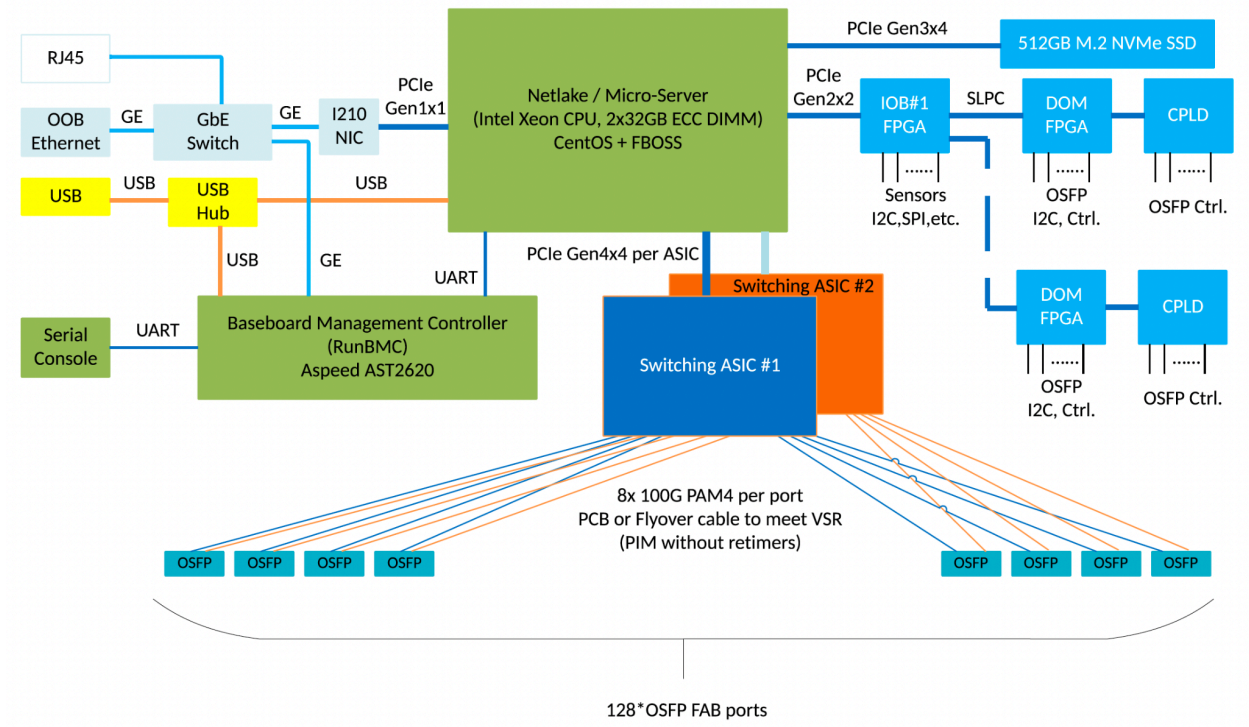
7. FN (Fabric Node) System

The Fabric Node (FN) will be based on 2 pcs of BCM88920 ASICs which has 51.2T switching capacity. Here is the expected feature list Meta is proposing:

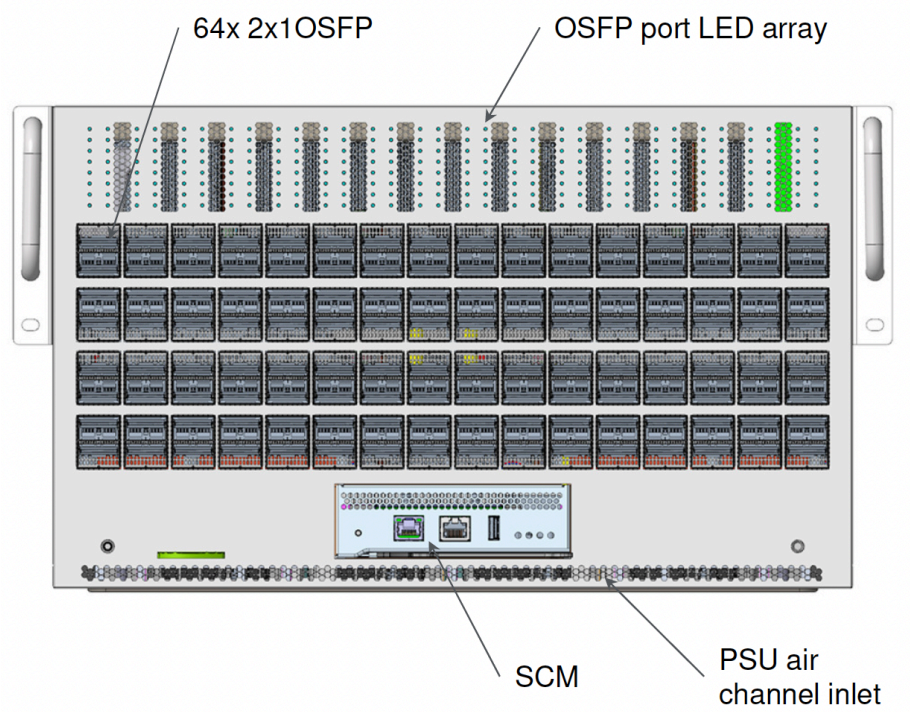
- Fabric Interface
    - 128x OSFP ports
    - Each OSFP port is equally split (4 + 4 lanes) between the 2 switching ASICs.
        - The upper and lower 4 lanes of the OSFP ports are equally split (2 + 2 lanes) between the 2 switching ASICs.
        - Each ½ OSFP port is further equally split (2 + 2 lanes) between the 2 virtual cores of each ASIC.
- Management Interface
    - 2x RJ45 1000BASE-T port
    - 1x RJ45 console port
    - 1x USB Type-A port
- Intel Ice Lake-D CPU

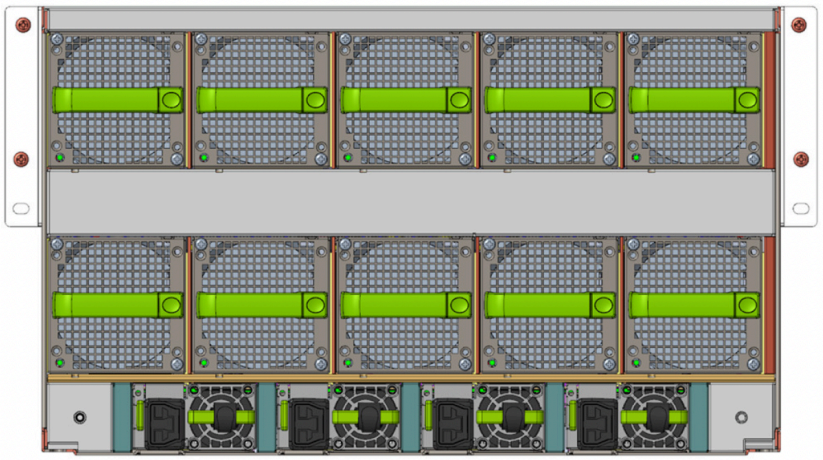
    - LCC, 8 cores, 2.3GHz, 55W
    - 2x 32GB SODIMM w/ ECC, 1 DPC
    - External NICs
    - 512GB NVMe SSD
- Aspeed AST2620 BMC
- 6RU
- AC or DC power feeds
- Typical power consumption at full load: 3,750W (2+2 3,000W PSUs)
- No retimers between the OSFP ports and switching ASICs for the 106.25G SerDes channels

The system diagram below shows the major components and their logical connections for the system:



The front panel port mapping/port placement and mechanical design support 128*OSFP Network Interfaces. The Switch Control Module is a FRU which can be plugged from the front side. There are four PSU slots on the rear side.


64x 2x1OSFP
OSFP port LED array
SCM
PSU air
channel inlet



8. Common modules

There are multiple modules/designs that can be shared across IN and FN systems: CPU (x86) sub-system, BMC sub-system, FPGA design, etc.
For the chassis, and each FRU part, there is EEPROM storing inventory information, including production name, part number, serial number, built date, version numbers, PCB vendor info, etc.

8.1 CPU Subsystem

DSF CPU daughter card is planned to be the Meta Netlake CPU card. Netlake dimension is increased from the COMe Basic form factor of 95mmx125mm to 115mmx135mm, to host the Ice Lake-D CPU. COMe connector part number is Foxconn QT5HB26-32104-3H.

8.2 BMC Subsystem

RunBMC sits on Switch Control Module with Netlake module next to it. RunBMC is a Meta design that is already in volume production and contributed to OCP in 2021.

8.3 Switch Control Module Subsystem

The Switch Control Module for IN and FN has different physical sizes to fit individual chassis, but the majority of the features and logic should be very similar.

Switch Control Module (shown as light green in the system diagram) is a FRU that is replaceable from the front/rear of the chassis. Switch Control Module power consumption is estimated to be about 90W(typical)/105W(max).

8.4 FPGA Subsystem

IOB (IO Bridge), DOM FPGA and some CPLDs are required to help software to manage the devices in the system. In DSF design it's planned to offload the majority of the board management tasks from the legacy BMC device to the X86 CPU. Majority of the management is through CPU-IOB PCIe path.

- CPLD and FPGA programming is now primarily managed by CPU. BMC is required to be able to recover critical PLDs in case of CPU failure.
- Fan Speed Control (FSCD) is now running on CPU. The CPU needs to collect system temperature information to determine and control the system fan speed. As Minipack3 is the first Meta switch design to use CPU to run FSCD, BMC needs to control fan speed as a back-up before CPU FSCD becomes solid.
- PSU and voltage regulator management is now running on CPU. BMC only takes care of the system power cycle in case of failure.

IOB FPGA manages the other FPGAs and SMB CPLDs through SLPC bus, which was introduced in Minipack1 and improved in Minipack2. DOM FPGA manages OSFP pluggable transceivers. SMB CPLDs should be managed through IOB FPGA as their primary control path.

BMC is still required to take care of system recovery from CPU failure, so some GPIO/I2C connections between BMC and CPLDs are needed to help failure recovery, as well as I2C connection between BMC and IOB FPGA as a back-door for the debugging purpose.

8.5 PSU/PEM and FAN

There are two/four AC/DC-51V PSUs as 1+1 or 2+2 redundancy to provide power to the IN and FN systems.

DSF requires a DC-PEM for the use cases where DSF is powered by 51V DC power directly. DC-PEM power rating up to 5KW is desired to support future generations of rack switch design.

9. Environmental / Regulatory Requirements

WEEE:

- RoHS Directive (2011/65/EU), including provisions of RoHS 3 Directive (2015/863) upon coming into effect.
- REACH Regulation (EC) No 1907/2006
- Waste Electrical and Electronic Equipment (“WEEE”) Directive (2012/19/EU)
- The Persistent Organic Pollutants Regulation (EC) No. 850/2004
- The California Safe Drinking Water and Toxic Enforcement Act of 1986 (“Prop 65”)
- The Packaging and Packaging Waste Directive 94/62/EC
- Batteries Directive 2006/66/EC
- "Corporate Sustainability Requirements (refer to latest version of Meta Materials of Concern Standard):
    - Halogen Free, IEC61249-2 (Note that power cord is exempted from HF requirement)
    - Arsenic Free
    - Phthalate free

Regulatory:

- Regulatory agency approvals are required so that the product can be shipped into the following countries: US, Canada, EU, UK, Singapore, Japan,Taiwan, Australia and Mexico.

Emissions and immunity:

- EMC Directive 2014/30/EU
- FCC Part 15 (CFR 47) (USA) Class A
- EN55032 (Europe) Class A
- CISPR32 (International) Class A
- AS/NZS CISPR32 (Australia and New Zealand) Class A
- EN61000-3-2
- EN61000-3-3
- EN 55035/CISPR 35 with all latest revision
- EN 61000-4-2 ESD, Customer desired target 8K contact /15K air discharge Criteria A (Minimum to 5.6Kv contact/11.2Kv air to criteria A)
- EN 61000-4-3 Radiated Immunity Customer desired target 10V/m Criteria A (Minimum to 3V/m Criteria A)
- EN 61000-4-4 EFT
- EN 61000-4-5 Surge (Criteria A desired target)
- EN 61000-4-6 Low Frequency Conducted Immunity Customer desired target 10V (3V is the minimum)

- EN 61000-4-8 Magnetic Field (1 A/m)
- EN 61000-4-11 Voltage Variations and Dips

Appendix A - Checklist for IC approval of this Specification (to be completed by contributor(s) of this Spec)

Complete all the checklist items in the table with links to the section where it is described in this spec or an external document .

Item	Status	Link to detailed explanation
Has this contribution been presented to an OCP Project group during a project call or engineering workshop?	Yes or No	If “No”, please state the reason.
Approval by Project Leads	Yes or No	If “No”, please state the reason.
Is this contribution entered into the OCP Contribution Portal?	Yes or No	If “No”, please state the reason.
Was it approved in the OCP Contribution Portal?	Yes or No	If “No”, please state the reason.

Appendix B - Contribution Process FAQs

As a contributor to a hardware specification, here are some questions that often come up.

Q1. What type of specification am I contributing to OCP?

a. **The base specification** for Meta DSF (Disaggregated Scheduled Fabric) switches hardware platform.

Q2. How do I know if what I am contributing will be accepted by OCP?

a. Before contributing any specifications, please contact either OCP Staff (Rob Coyle, Michael Schill) or the Project Lead for the Project that best represents your contribution. They will guide you as to what's the best form for your contribution. Project List here.

Q3. What is the contribution process for my hardware spec?

a. Follow the flow for your spec type here.

Q4. What if my spec is not developed yet and I want to collaborate with other companies?

a. Please contact either OCP Staff (Rob Coyle or Michael Schill) or the Project Lead for the Project that best represents your contribution. They will help you find other collaborators and help you with the contribution process for a multi-party contribution.

Q5. I have a question about the Contribution License Agreement (CLA).

a. Please contact OCP Staff and we can help you with questions.

Q6. Do I need to have a product in order to make a contribution?

a. Please see Q1. Some types of contributions do not result in a product. Some examples are whitepapers, case studies, OCP Ready Assessment, etc.. Please work with the OCP Staff on the better direction on your specification type.