



OPEN POD GROUP FOR M XPUS (OPG-M) SYSTEM ARCHITECTURE

Authors
Lowell D. Lamb*, Broadcom
Loren Staley*, Celestica
Joshua Mora, System-Stack
Ashok Raman, Broadcom
*Corresponding authors

Date: January 2026

Version History

Date	Version #	Authors	Description
14 January 2026	1.0	Lowell Lamb, Loren Staley, Joshua Mora, Ashok Raman	First version

Current Template Version:

Whitepaper/Document Submission Template v1.7.0
Effective August 2025

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Executive Summary

This document describes the System Architecture of an Open Pod Group for AI infrastructure designs. It is a companion document to *Open Cluster for N xPUs (XOC-N) System Architecture* [1], published by the Open Compute Project. An overview of the market needs that drove the development of these two documents and the high-level cluster architecture can be found in the Open Compute Project White Paper *Open Cluster Designs for AI* [2].

Acknowledgements

The Authors would like to acknowledge the following for their participation and feedback:

Matt Baldassano
Shawn Chen
Ryan Harris
Ravindra Kadiyala
Madhusudan Kanagasabai
James Kelly
Tom McKnight
John Leung
Tai Nguyen
Yusuke Sato
Yichao Wang
Russ Wunderlich

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table of Contents

Version History 2

1. Introduction 5
2. Compliance with Open Compute Project Tenets 5
3. Scope 6
4. Nomenclature 6
5. AI cluster architecture 7
6. 400G Open Pod Groups for M xPUs (OPG-M) 9
7. OPG-128 design 26
8. OPG-128 kit 46
9. Network utilization 47
10. Adapting the OPG-128 design 47
11. References 52
12. Glossary 53
13. License 56
    13.1. Creative Commons 56
14. About Open Compute Foundation 57

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

1. Introduction

This document describes the System Architecture of an Open Pod Group for AI infrastructure designs. It is a companion document to Open Cluster for N xPUs (XOC-N) System Architecture [1], published by the Open Compute Project. An overview of the market needs that drove the development of these two documents and the high-level cluster architecture can be found in the Open Compute Project White Paper Open Cluster Designs for AI [2].

2. Compliance with Open Compute Project Tenets

This System Architecture, which is built on top of existing and extensive OCP work, is aligned naturally with the OCP Tenets. In summary:

Efficiency: At the hardware level, OCP designs will be used wherever possible, thus addressing physical issues such as power consumption, cooling requirements, floor-space requirements, etc. At the architectural level, the modular structure of the System Architecture and the complete descriptions of equipment requirements and cabling connectivity will reduce design effort and accelerate deployment. Regarding management, the use of software solutions developed by the OCP and other, open initiatives will provide a clear path for the rapid design, deployment, and operation of competitive management platforms optimized for the needs of the particular AI infrastructure under design.

Impact: While hyperscale players typically have extensive engineering resources and are capable of undertaking the full design and deployment of large AI networks, enterprises often have lean IT staffs structured for primarily for operational tasks. By providing complete and practical System Architectures, this System Architecture should lower the design barrier for the enterprise sector and accelerate AI infrastructure deployment in this important market.

Scalability: The modular nature of the designs developed by this work should naturally facilitate network creation, both during initial design and deployment, as well as subsequent capacity expansion.

Openness: The reliance of this work on the extensive corpus of existing OCP work (e.g., the rack-scale xPU pod designs under development by the OCP *Open Systems for AI* Strategic Initiative [3]), as well as the availability of proven solutions from other not-for-profit forums (e.g., SONiC [4]), should enable ready extension or modification by users of the System Architecture.

Sustainability: The existing OCP hardware solutions used for much of the System Architecture provide an excellent foundation for durable AI deployments. The designs described here include quantitative estimates of the power, cooling, and other physical requirements at the module and cluster levels, thus providing AI cluster architects and owners with practical insight useful in optimizing the AI infrastructure to meet operational goals.

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

3. Scope

This System Architecture describes the elements and structure of a multi-rack system containing xPU pods/nodes, switches, and storage resources, which we designate as an Open Pod Group of M xPUs (OPG-M). These OPG-Ms are modular building blocks that are networked together to construct AI clusters. An AI cluster contains the xPUs that share a common scale-out network, plus the other resources necessary for the function of the xPUs. The present work was developed as a companion to the OCP xPU Open Cluster for N xPUs (XOC-N) System Architecture [1], which describes the physical network elements located at the spine/aggregation layer, plus the other resources required to knit one or more OPG-Ms into a full AI cluster, which we designate as an xPU Open Cluster for N xPUs (XOC-N).

This System Architecture is intended to provide designs which are detailed enough to serve as a technical basis for procurement and implementation. Scale-up, scale-out, scale-out-CPU, storage, and management networking spanning the entire design all are in-scope. Physical and logical network topology spanning the entire design are in-scope. Solutions for cabling and cable-management are in-scope. Infrastructure, including power distribution, load bearing, and cooling, are in-scope. This document describes air-cooled solutions, and a future System Architecture document will describe liquid-cooled solutions.

Relationship with other OCP Contributions: In addition to [5], this System Architecture leverages Contributions from other OCP Projects wherever possible and solutions developed by the OCP Open Systems for AI Strategic Initiative and OCP solutions for pod, rack, and cluster infrastructure management will be emphasized.

4. Nomenclature

Following are key terms used in this document. A more extensive list is contained in the Glossary.

OPG-M	Open Pod Group for M xPUs
System Architecture	The elements and structure of an OPG-M
XOC-N	xPU Open Cluster for N xPUs
xPU	Generic term for special-purpose AI accelerator device

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

5. AI cluster architecture

The rapid growth of the AI/ML market has driven large-scale changes in network design, resulting in the emergence of the AI-network architecture shown in Figure 1. In this Section we describe the basic structure of this architecture and its principal components, including physical, software, and logical elements.

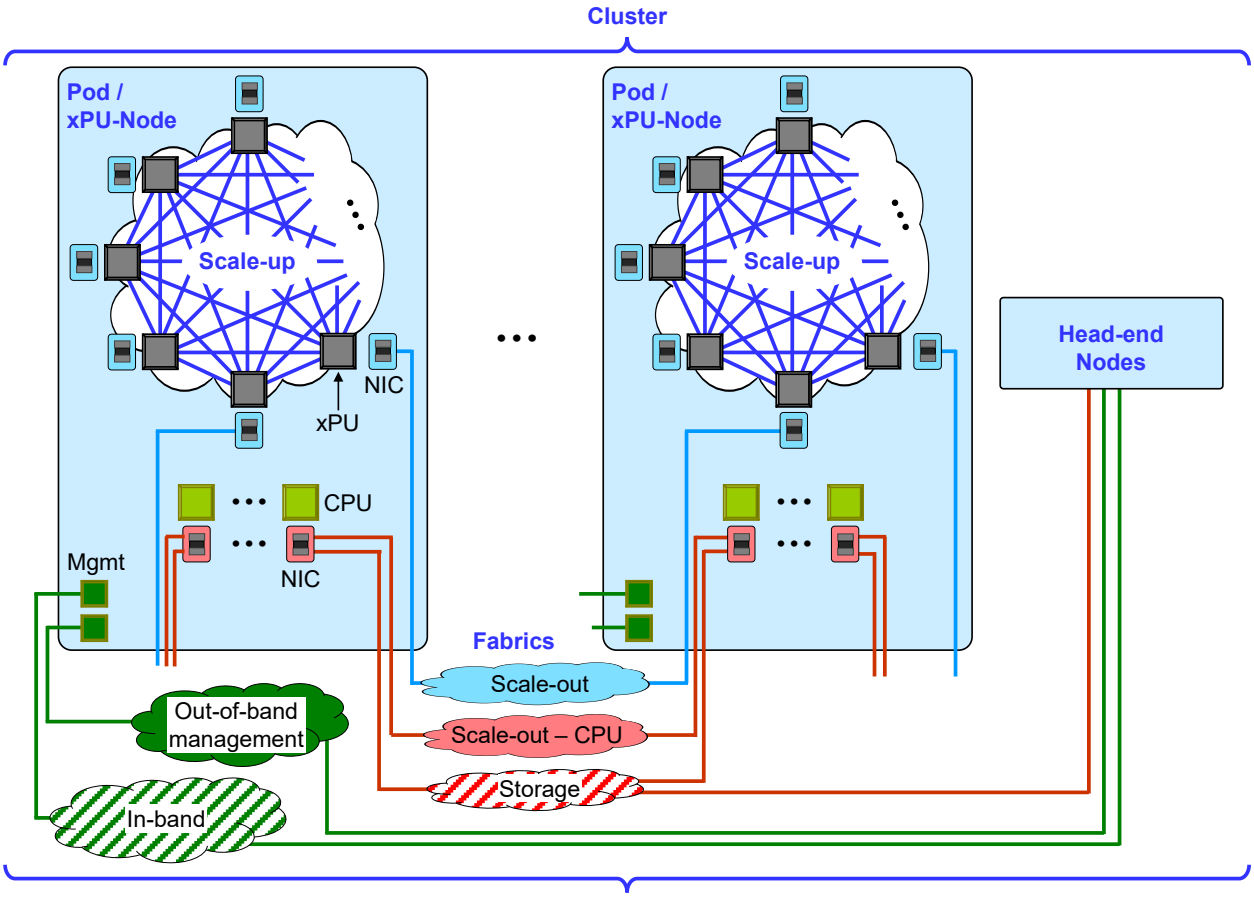


Figure 1. AI cluster networks and connectivity domains.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

xPU Open Cluster for N xPUs (XOC-N)

Physical AI systems contain six distinct networks, namely the scale-up fabric, scale-out fabric, scale-out CPU (SO-C) fabric, storage fabric, in-band network, and out-of-band management network. The scale-up and scale-out fabrics are extremely high-bandwidth and latency-sensitive, and together they determine the fundamental structure of the overall design. The scale-up, storage, in-band management, and both scale-out networks are internal to the XOC-N and direct, external connections are not allowed; all traffic to and from external sources must pass through the firewall. As described in [2], this System Architecture provides a design that provides separate physical resources for each network, i.e., a non-converged solution. This approach provides maximum isolation, security, and performance, while simplifying configuration, load-balancing, and troubleshooting.

XOC-Ns are built as assemblies of Open Pod Groups of M xPUs (OPG-M), together with the necessary resources at the spine and aggregation layers (see Figure 2). The spine and aggregation resources, software and management elements, and connectivity topology, will be covered in a separate System Architecture. The remainder of this document is concerned with the definition of OPG-Ms.

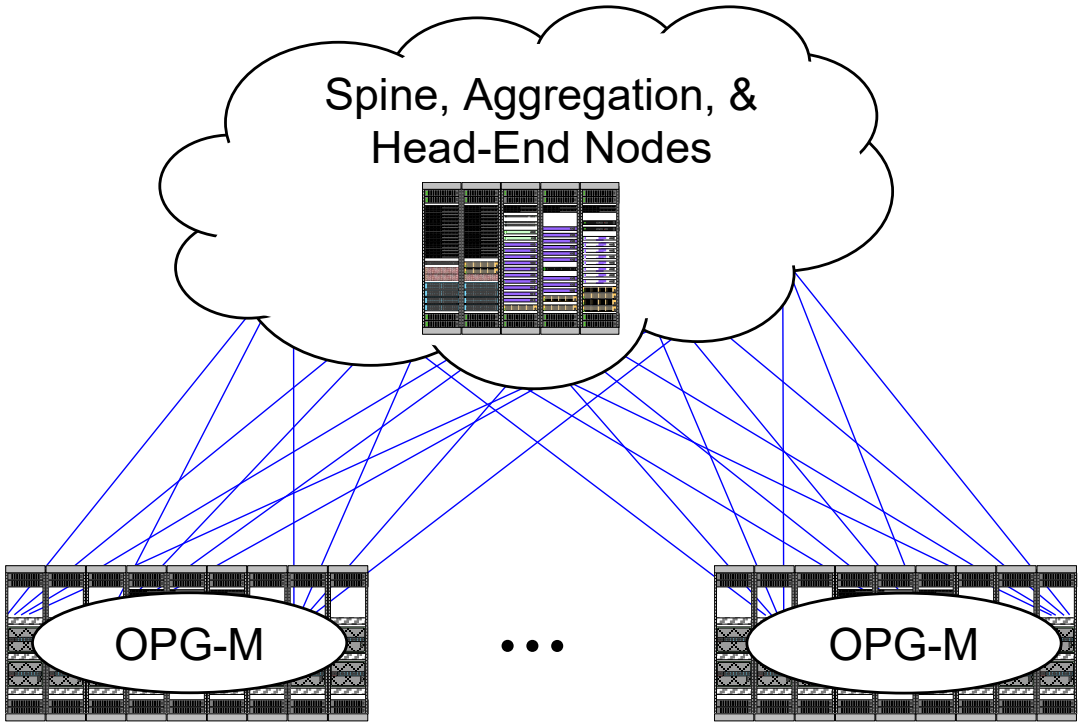


Figure 2: xPU Open Cluster for N xPUs (XOC-N).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6. 400G Open Pod Groups for M xPUs (OPG-M)

The term Open Pod Group of M xPUs (OPG-M) is used to denote the principal building block of an Open Cluster. As shown in Figure 3, OPG-Ms are composed of xPU-nodes plus other resources. At the time of this writing (2025), air-cooled xPU-nodes, which typically house eight or fewer xPUs, are shipping in high volume. The larger, liquid-cooled pods, designed to house 64 or more xPUs, are entering the market now and should achieve high volume in 2026. This System Architecture focuses on the air-cooled generation of AI infrastructure, which most commonly uses 400 GbE scale-out links. The emerging liquid-cooled generation, which most commonly uses 800 GbE scale-out links, will be described in a separate System Architecture document. In the present work we describe the features and requirements for an OPG-128 (128 xPUs) with forced-air cooling and 400 GbE scale-out links. Guidance also is given on adapting the OPG-128 design to build either smaller or larger OPG-Ms, and options for adding resiliency and redundancy are discussed.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

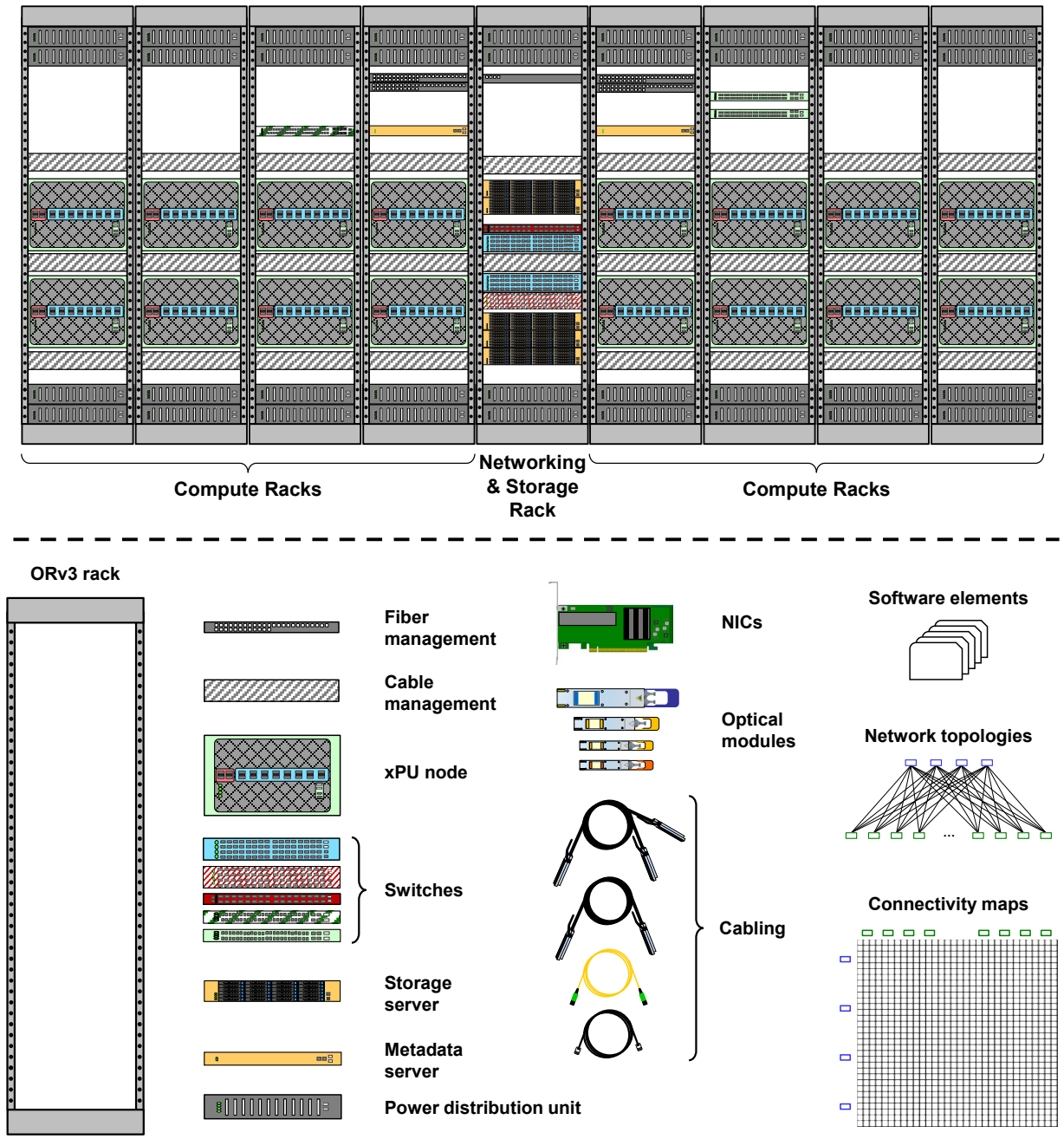


Figure 3. Top: Drawing of an air-cooled OPG-128 showing the xPU-nodes and associated resources that together constitute an OPG-M (not shown: NICS, optics, and cabling). Bottom: Physical, logical, and software elements of an OPG-M.

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6.1. Guidelines and assumptions

1. High-power ORv3 racks, equipped with PDUs capable of providing 20+ kW of power.
2. Forced-air cooling is used for all equipment. It is assumed that the data-center cooling capacity is 20 kW per rack.
3. Separate networks for scale-out, storage, SO-C, in-band management, and out-of-band management networks.
4. In this design the front-end network is divided into two physically separate fabrics, namely the SO-C fabric and the in-band management network.
5. In this design the SO-C and storage fabrics are treated as physically separate fabrics.
6. For the five OPG-M networks, namely the scale-out, storage, SO-C, in-band management, and out-of-band management, no redundancy is assumed or required.
7. Break-out DACs are recommended for scale-out connections within the OPG-M: 800G OSFP (switch port) ←→ 2x 400GbE/QSFP112 (xPU NICs).
8. Break-out DACs are recommended for SO-C fabric connections within the OPG-M: 400G OSFP (switch port) ←→ 2x 100GbE/QSFP112 (CPU NICs).
9. Break-out DACs are recommended for storage pod connections within the OPG-M: 400G OSFP (switch port) ←→ 2x 100G/QSFP112 (CPU NICs).
10. All scale-out fabric, SO-C fabric, and storage fabric links between the OPG-M and the spine/aggregation layers are 400GbE optical links. This simplifies design and inventory, and enables the construction of large, 2-tier scale-out networks (> 1k xPUs), that otherwise would not be possible if 800GbE links were chosen for these connections. Physically, these connections are OSFP (switch port) ←→ 2x 400GbE (optical) ←→ OSFP (switch port).
11. Single-mode OS2 cabling is used for all optical cabling.
12. MPO-8 connectors are recommended for all 400G optical couplers.
13. 800G OSFP-compliant, single-mode transceivers with dual MPO-8 APC connectors (2x 400GbE) and 500m reach are recommended for all 400G optical links.
14. These designs are compatible with 500m optical trunk cables terminated with MPT-8/12 connectors.
15. 25G passive DAC cables are recommended for connecting the in-band management leaf switch to the xPU nodes, storage servers, and metadata servers in the OPG-M.
16. Single-mode100G QSFP28 optical transceivers with 500m reach are recommended for the in-band management leaf-spine connections.
17. Single-mode10G SFP+ optical transceivers with 500m reach are recommended for the out-band management leaf-spine connections.
18. All networks in these designs (scale-out, SO-C, storage, in-band management, and OoB management), are designed as non-blocking networks.
19. All active equipment (servers, switches, and power-distribution units) is assumed to have a 1 GbE/RJ45 port for connection to the out-of-band management network.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

20. All xPU servers, storage servers, and metadata servers are assumed to have a 25 GbE/SFP28 port for connection to the in-band management network.
21. Most xPU-nodes, storage servers, and meta-data servers offer some level of redundancy in their designs, often in the power supply. No additional redundancy is assumed or required in the present work. The required storage resources are modelled using a ratio of approximately one PCIe-Gen5 NVMe drive per xPU.
22. Due to the wide variety in fiber-management shelf configuration and plenum-cable size, the port-numbering plans and connection maps for the fiber-management shelves are out-of-scope for this design.
23. Where possible, the OPG-M design will provide resiliency to support a graceful degradation in function in the event of hardware of software impairment.

6.2. Physical elements

This Section contains descriptions of the active and passive equipment found in an OPG-128. These descriptions are generic and are not meant to indicate a particular product or manufacturer; comparable products in each of the categories are available from multiple suppliers. The intent here is to set down for each element the features and requirements necessary for the definition of the OPG-128.

6.2.1. Equipment racks, cable management, fiber management, and power distribution

Figure 4 illustrates the basic infrastructure of the OPG-M racks. The ORv3 rack [5] is used throughout the design, and horizontal cable-management hardware that spans the entire OPG-M is present to facilitate the use of DAC cabling. The storage and networking racks will have high-density fiber-management housings to enable the cross-connection of local optical cables and the trunk/plenum optical cables. Each rack contains one or more power supplies to meet the aggregated power requirements of the equipment located in the rack. The power-supply requirements for each type of rack are described below.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

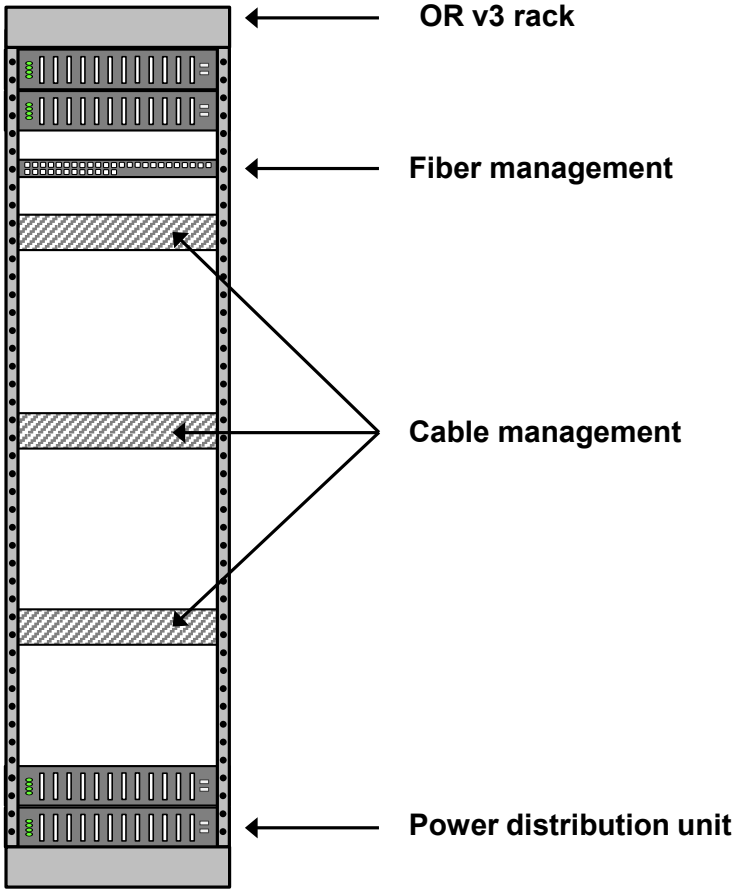


Figure 4. OPG-M basic infrastructure, based on the OR v3 rack.

Equipment racks: Racks used in these designs comply with the OCP *Open Rack V3 Base specification* [5].

Horizontal cable-management hardware: The horizontal cable-management trays (raceways) should be sufficient in number and capacity to support and route 96x 26 AWG DAC cables on each side of the center networking rack.

Optical fiber-management shelves: The optical-fiber management shelves should be sufficient in number and connector density (48x MPO-8/12 connectors or more per OU) to terminate and route, using a minimum number of OU, 168x 400GbE & 4x100GbE links (4x bi-directional optical lanes each) in six or fewer shelves.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Power-distribution units (PDU): The PDUs used in this design are 2U, 240V, 30A, 5.8 kW units. Each rack is equipped with four PDUs, which together provide a maximum power of 23.2 kW per rack.

6.2.2. xPU nodes and NICs

The xPU node, illustrated in Figure 5, is an air-cooled server capable of hosting as many as eight xPUs and two CPUs. Equipment in this category is available from at least six suppliers and can support a variety of xPU manufacturers and models.

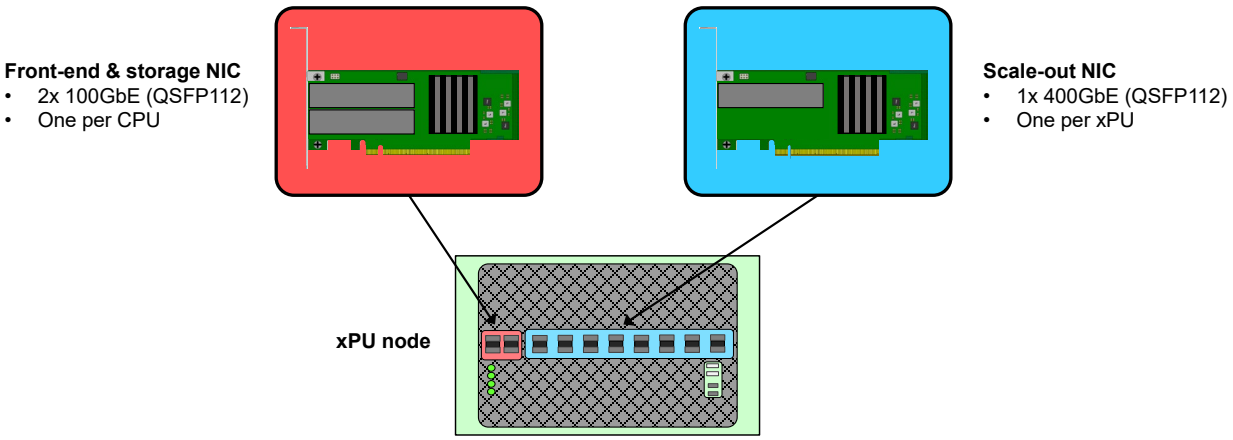


Figure 5. xPU node and NICs.

In these designs we assume that the xPU node has the following features/configuration:

- 8x xPUs
- 2x CPUs
- 8x scale-out NICs (one per xPU)
  - Single-port 400GbE (QSFP112)
- 2x scale-out NICs (one per CPU)
  - Dual-port 100GbE (QSFP112)
  - Dedicated port for SO-C fabric
  - Dedicated port for storage fabric
- OoB management port, RJ45, capable of 1GbE
- 2x CPUs
- Height: 8 OU or less

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- Maximum power consumption: 10 kW or less
- Traffic model (sustained, line-rate)
  - Scale-out: 400GbE / NIC
  - SO-C: 100GbE / NIC
  - Storage: 100GbE / NIC
  - OoB management: 1GbE / xPU Node

6.2.3. xPU nodes and NICs

The elements of the storage subsystem, illustrated in Figure 6, are air-cooled units that: (i) house the data used by the applications running on the cluster; and (ii) manage the flow of data between the xPUs and the storage pool. Equipment in this category is available from multiple suppliers.



Figure 6: Storage server and metadata server.

In these designs we assume that the storage server and metadata server have the following features/configuration:

Storage server
- 24x E3.S NVMe drive bays
- 2x CPUs
- 4x NIC bays, capable of simultaneously supporting 400 GbE NICs (QSFP112)
- OoB management port, RJ45, capable of 1GbE
- Height: 2 OU or less
- Maximum power consumption: 2 kW (1+1 redundant)
- Redundant power supply
- Traffic model (sustained, line-rate)
  - Storage: 400GbE / NIC
  - OoB management: 1GbE / xPU Node

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Metadata server

- 2x high-frequency CPU
  - Cores: 24-32
  - RAM/DDR: 1 TB
- 2x NIC bays, capable of simultaneously supporting 400 GbE NICs (QSFP112)
- OoB management port, RJ45, capable of 1GbE
- Height: 1 OU
- Maximum power consumption: 1.6 kW (1+1 redundant)
- Traffic (sustained, line-rate)
  - Metadata: 400GbE / NIC
  - In-band management: 25GbE / SFP28
  - OoB management: 1GbE / xPU Node

6.2.4. Switches

The switches used in these designs are illustrated in Figure 7. All are air-cooled units and are either one OU or two OU in height. The scale-out and storage switches have identical capabilities and requirements; color coding is used only to indicate which network each switch belongs to. Switches in these categories are available from multiple suppliers.

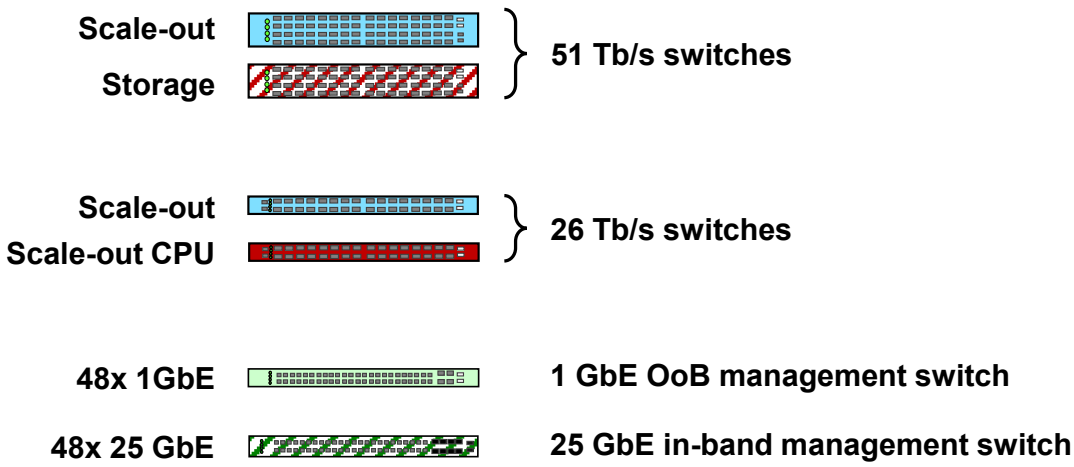


Figure 7. Scale-out, SO-C, storage, and management switches.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

In these designs we assume that the switches have the following features/configuration:

Scale-out, SO-C, and storage switches

- Ports
  - 51 Tb/s – 64x 800GbE OSFP
  - 26 Tb/s – 32x 800GbE OSFP
- Standard Ethernet traffic management features, including VxLAN
- Break-out cable support on all ports
- SONiC support
- Forced-air cooling
- OoB management port, RJ45, capable of 1GbE
- Supported speeds: 100/200/400/800 GbE
- Height
  - 51 Tb/s – 2 OU
  - 26 Tb/s – 1 OU
- Maximum power consumption
  - 51 Tb/s – 3.2 kW
  - 26 Tb/s – 1.8 kW
- Traffic model (sustained, line-rate)
  - Scale-up switch: 2x 400GbE / OSFP port (800GbE total)
  - SO-C switch
    - NIC down-links: 2x 100GbE / OSFP port (200GbE total)
    - Aggregation up-links: 2x 400GbE / OSFP port (800GbE total)
  - Storage NIC aggregation switch
    - NIC down-links: 2x 100GbE / OSFP port (200GbE total)
    - Aggregation up-links: 2x 400GbE / OSFP port (800GbE total)
  - Storage rack switch: 2x 400GbE / OSFP port (800GbE total)
  - OoB management: 1Gb/s per downlink

25 GbE in-band management switch

- Ports
  - 48x 25 Gb/s switch: – (48x 25 GbE / SFP28) + (8x 100 GbE / QSFP28)
- Standard Ethernet traffic management features
- Forced-air cooling
- Supported front-panel speeds: 25 GbE
  - Height: 1 OU
- Maximum power consumption
  - 300 W
- Traffic model (sustained, line-rate)
  - 25 Gb/s per downlink
  - 100 Gb/s per active uplink

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

1GbE OoB management switch

- Ports: (48x / 24x 1GbE / RJ45) + (4x 10GbE / SFP+)
- Standard Ethernet traffic management features
- Forced-air cooling
- Supported speeds: 1/10 GbE
- Height: 1 OU
- Maximum power consumption
  - 48x GbE – 100 W
  - 24x GbE – 60 W
- Traffic model (sustained, line-rate)
  - 1Gb/s per downlink
  - 10Gb/s per active uplink

6.2.5. Cables and optical modules

Figure 8 shows the recommended optical modules and local-side cabling for the various links within the OPG-M.

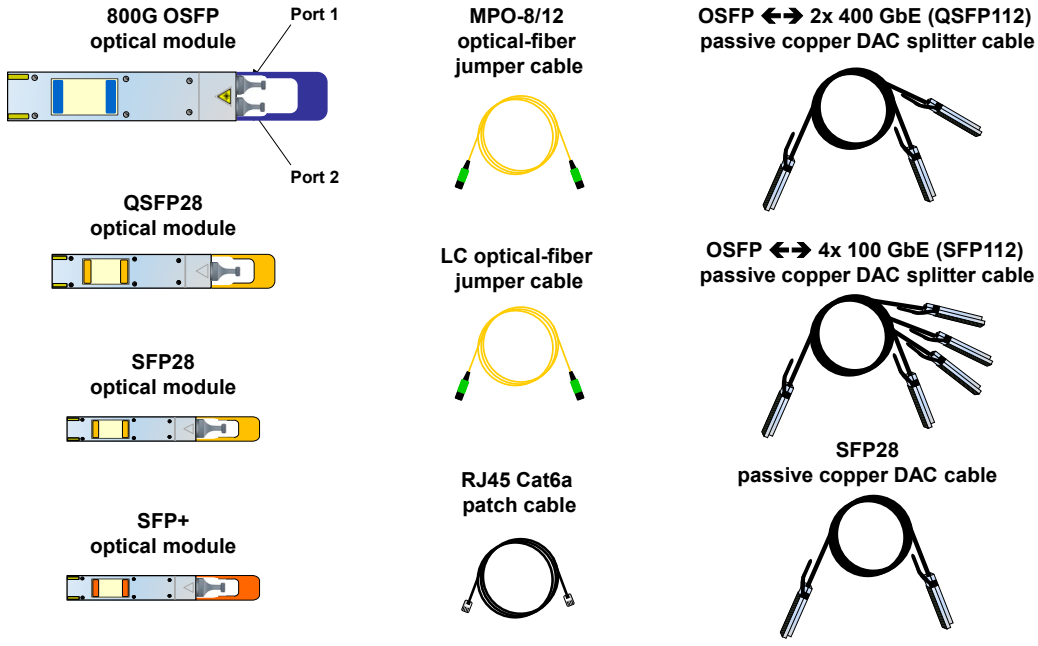


Figure 8. Pluggable optical modules and cables used for local connections in the OPG-M.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6.2.5.1. Optical modules

All high-speed links that leave the OPG-M are carried by optical links. Due to their high level of performance, single-mode, pluggable modules with the following features are recommended:

- 800G OSFP form factor
  - 1310nm, PAM4
  - 500m reach
  - Compatible with OS2 single-mode fiber
  - 2x 400G mode via dual connectors
  - 2x DR4
  - Connectors: Dual MPO-8/12 APC

- 100G QSFP28 form factor
  - 1310nm, NRZ
  - 500m reach
  - Compatible with OS2 single-mode fiber
  - DR4
  - Connector: MPO-8/12 APC

- 25G SFP28 form factor
  - 1310nm, NRZ
  - 500m reach
  - Compatible with OS2 single-mode fiber
  - Connector: LC

- 10G SFP+ form factor
  - 1310nm, NRZ
  - 500m reach
  - Compatible with OS2 single-mode fiber
  - Connector: LC

6.2.5.2. Optical-fiber jumper cables

Single-mode (OS2) optical jumper cables with MPO-8/12 cables are recommended for all local (intra OPG-M) optical links. These jumper cables connect the optical modules in the switch ports to the fiber-management shelves, which provide interconnection with the trunk/plenum cabling.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6.2.5.3. Passive Direct Attach (DAC) cables

DACs are recommended for all local (intra OPG-M) NIC-switch and in-band management links. These cables should have the following features:

NIC-switch links

- 26 AWG
- Scale-out links: 800G OSFP ←→ 2x 400G QSFP112 break-out cable
- SO-C links: 400G OSFP ←→ 4x 100G SFP112 break-out cable
- Storage links: 400G OSFP ←→ 2x 100G QSFP112 break-out cable

In-band management links

- 26 AWG
- SFP28 ←→ SFP28

6.2.5.4. Cat6a patch cables

Cat6a patch cables are recommended for all local (intra OPG-M) 1 GbE OoB management links between the managed elements and the OoB management switch.

6.2.6. Power, space, and weight requirements

At a fundamental level, the core objectives of most AI infrastructure designs are to simultaneously maximize compute density while minimizing costs. To do so, a key requirement is that the equipment power and cooling capacity must match, and this requirement usually is stated on a per-rack basis. In this design, we have assumed a data-center maximum cooling capacity of 20 kW / rack, which is typical for high-end facilities which rely solely on forced-air cooling. Figure 9 shows the per-rack power and weight estimates for an OPG-128.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

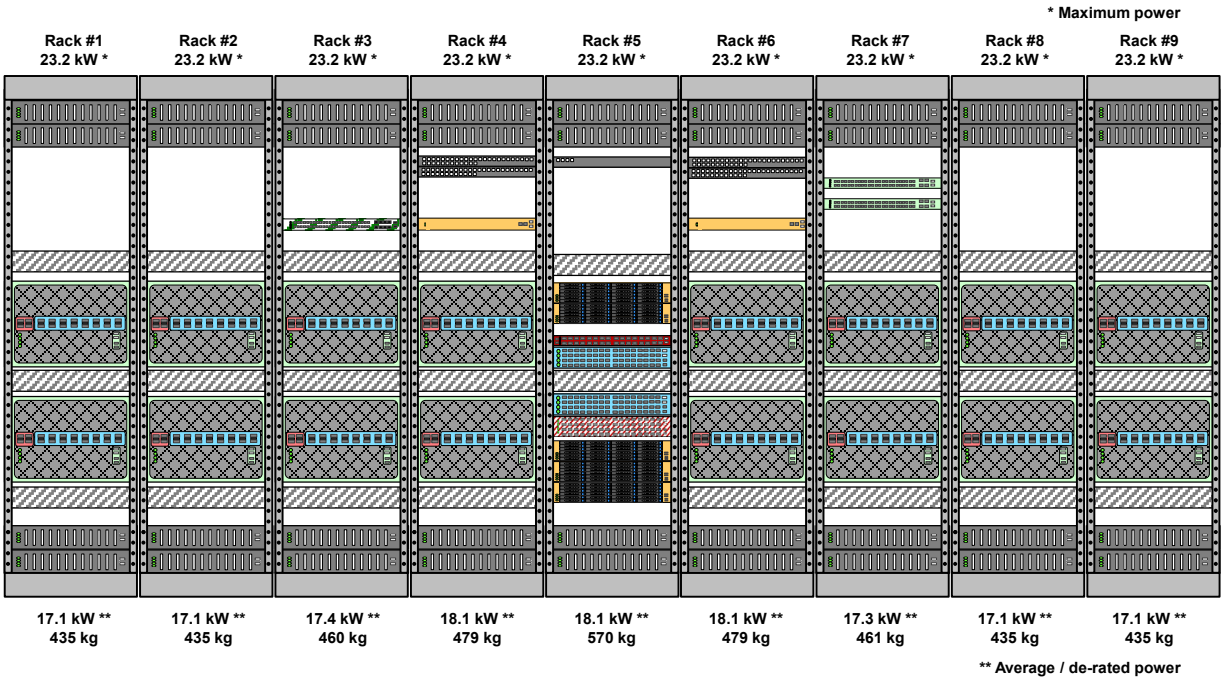


Figure 9. Power and weight estimates for an OPG-128.

Tables 1 – 6, below, summarize the infrastructure requirements, namely power, space, and weight, for the racks present in the OPG-128 design. For the power design, the tables show both the total, worst-case power consumption (maximum), as well as a de-rated value that accounts for the dynamic nature of power loads and represents the average power draw. In these designs, the de-rated power is estimated as 80% of the maximum power. The PDU capacity was sized to support transient power draws of at least 120% of the de-rated power value. Regarding the rack weights, the estimates account for the fact that the DAC cables cross multiple racks, hence the estimated cable weight for a single rack is an allocated share of the total cable weight.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

		Maximum power (kW)	Height (OU)	Weight (kg)
xPU server		10.1	8	105
Storage server		2.0	2	38
Metadata server		1.6	1	20
51T Ethernet switch		3.2	2	24
26T Ethernet switch		1.8	1	15
48x 25GbE switch		0.3	1	13
48x 1GbE switch		0.1	1	7
24x 1GbE switch		0.1	1	
			1	
5 kW PSU		0.3	2	10
Horizontal cable tray		0	2	3
Fiber management tray		0	1	1
DAC cable		0	0	0.8
ORv3 rack		0	0	170
Power de-rating factor				80%

Table 1: Weight, power, and equipment height values used to estimate the OPG-128 infrastructure requirements.

Racks #1, 2, 8, 9	# per rack	Unit power (kW)	Unit Height (OU)	Unit Weight (kg)	Height subtotal (OU)	Weight subtotal (kg)	Power subtotal (kW)
xPU Servers	2	10.1	8	105	16	210	20.2
5 kW PSU	4	0.3	2	10	8	40	1.2
Horizontal cable tray	3			3		9	0
DAC cables (allocated share)	7.2			0.8	0	5.8	
ORv3 rack	1	0	0	170		170	
Space total (OU)					**24**		
Weigth total kg						**435**	
Total maximum power (W)							21.4
Total de-rated power (W)							**17.1**

Table 2: Weight, power, and total equipment height for Racks #1, 2, 8, and 9 (see Figure 9).

Rack #3	# per rack	Unit power (kW)	Unit Height (OU)	Unit Weight (kg)	Height subtotal (OU)	Weight subtotal (kg)	Power subtotal (kW)
xPU Servers	2	10.1	8	105	16	210	20.2
In-band mgmt switch	1	0.3	1	13	1	13	0.3
5 kW PSU	4	0.3	2	10	8	40	1.2
Horizontal cable tray	3	0	0	3	0	9	0
DAC cables	22.4			0.8		17.9	
ORv3 rack	1	0	0	170		170	
Space total (OU)					**25**		
Weigth total kg						**460**	
Total maximum power (W)							21.7
Total de-rated power (W)							**17.4**

Table 3: Weight, power, and total equipment height for Rack #3 (see Figure 9).

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Racks # 4, 6	# per rack	Unit power (kW)	Unit Height (OU)	Unit Weight (kg)	Height subtotal (OU)	Weight subtotal (kg)	Power subtotal (kW)
xPU Servers	2	10.1	8	105	16	210	20.2
Metadata server	1	1.6	1	20	1	20	1.6
5 kW PSU	4	0.3	2	10	8	40	1.2
Fiber management tray	2	0.0	1	1	2	2	0
Horizontal cable tray	3	0	2	3	6	9	0
DAC cables	35.4			0.8		28.3	
ORv3 rack	1	0	0	170		170	
Space total (OU)					**33**		
Weigth total kg						**479**	
Total maximum power (W)							23
Total de-rated power (W)							**18.4**

Table 4: Weight, power, and total equipment height for Racks #4 and 8 (see Figure 9).

Rack # 5	# per rack	Unit power (kW)	Unit Height (OU)	Unit Weight (kg)	Height subtotal (OU)	Weight subtotal (kg)	Power subtotal (kW)
Storage servers	5	2.0	2	38	10	190	10
Scale-out 51 switches	2	3.2	2	24	4	48	6.4
Storage 51T switch	1	3.2	2	24	2	24	3.2
Scale-out CPU 26T switch	1	1.8	1	15	1	15	1.8
5 kW PSU	4	0.3	0	10	0	40	1.2
Fiber management tray	1	0.0	1	1	1	1	0
Horizontal cable tray	3	0	2	3	6	9	0
DAC cables	90.8			0.8		72.6	
ORv3 rack	1	0	0	170		170	
Space total (OU)					**24**		
Weigth total kg						**570**	
Total maximum power (W)							22.6
Total de-rated power (W)							**18.1**

Table 5: Weight, power, and total equipment height for Rack #5 (see Figure 9).

Rack #7	# per rack	Unit power (kW)	Unit Height (OU)	Unit Weight (kg)	Height subtotal (OU)	Weight subtotal (kg)	Power subtotal (kW)
xPU Servers	2	10.1	8	105	16	210	20.2
48 GbE OoB mgmt switch	2	0.1	1	7	2	14	0.2
5 kW PSU	4	0.3	2	10	8	40	1.2
Horizontal cable tray	3	0	2	3	6	9	0
DAC cables	22.4			0.8		17.9	
ORv3 rack	1	0	0	170		170	
Space total (OU)					**32**		
Weigth total kg						**461**	
Total maximum power (W)							21.6
Total de-rated power (W)							**17.3**

Table 6: Weight, power, and total equipment height for Rack #7 (see Figure 9).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6.3. Software elements

This section describes the software and management elements of an OPG-M. To the greatest extent possible, the OCP management frameworks are leveraged in this design, with particular emphasis on SONiC and Redfish solutions. SONiC (Software for Open Networking in the Cloud) is an open-source network operating system (NOS) for data centers and cloud environments, which is managed by OCP and the Linux Foundation. Built on a Linux base with a modular, microservices-based architecture, SONiC runs on network switches from multiple suppliers and supports multiple software partners. Redfish is an industry-standard specification and API designed to consistent and unified management interface for data-center equipment, regardless of the vendor or architecture. Redfish-managed equipment can include compute, network, and storage products. Both SONiC and Redfish are designed to simplify the management of networks containing a diversity of hardware suppliers, reduce vendor lock-in, and promote hardware-software independence. Tables 7 and 8 summarize the recommended software elements for an OPG-M.

Category	Type	Expected leverage
Redfish profiles	Host CPUs	Server Profile
	Accelerators	UBB and Accelerators Profile
	Other (network, storage, etc.)	Base Profile
	Special components	NIC Profile, ORv3 PMC
BMC present	Baseboard Management Controller	OpenBMC
RMC present	Rack Manager Controller	OpenRMC
UBB BMC present	UBB/Accelerator Controller	*GPU-aware OpenBMC variant

Table 7: Summary of software module / profile recommendations.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

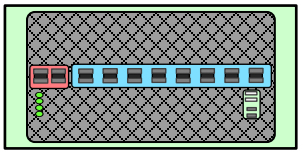
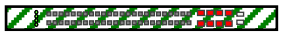
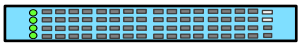
Category	Type	Expected leverage
Servers		
 xPU server	Host CPUs Accelerators If BMC present	server profile / implementation choice accelerator profile / implementation choice OpenBMC recommended
 Storage server	Host CPUs If BMC present	Vendor supplied
 Metadata server	Host CPUs If BMC present	Vendor supplied
Switches		
 OoB leaf switch		Vendor supplied
 In-band leaf switch	If BMC present	SONiC / Base Profile recommended OpenBMC recommended
 SO-C leaf switch	If BMC present	SONiC / Base Profile recommended OpenBMC recommended
 Storage leaf switch	If BMC present	SONiC / Base Profile recommended OpenBMC recommended
 Scale-out leaf switch	If BMC present	SONiC / Base Profile recommended OpenBMC recommended
Infrastructure		
 Power distribution unit		Vendor supplied

Table 8: Software elements for managed devices in OPG-M designs.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7. OPG-128 design

Table 9 summarizes the total number of spine/aggregation links for an OPG-128 and XOC-1k, which contains eight OPG-128s.

Network	OPG-128	XOC-1k [1]
Scale-out fabric	128x 400G (OSFP)	1,024x 400G (OSFP)
Storage fabric	40x 400G (OSFP)	320x 400G (OSFP)
SO-C fabric	8x 400G (OSFP)	64x 400G (OSFP)
In-band management	6x 100G (QSFP28)	48x 100G (QSFP28)
Out-of-band management	7x 10GbE (QSFP+)	56x 10GbE (QSGP+)

Table 9: Spine/aggregation links for an OPG-128 and an XOC-1k. All links use single-mode fiber cables.

The following subsections contain the port-numbering plans and connection maps for these networks.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7.1. Scale-out fabric

The scale-out fabric supports high-bandwidth communication between xPUs located in different pods / xPU-nodes. As shown in Figure 10, different switch complements are required in an OPG-128 design with scale-out network uplinks, depending on whether the scale-out topology is Clos-based ([1]) or rail-optimized. The subsections below contain the port-numbering plans and connection maps for the two topologies.

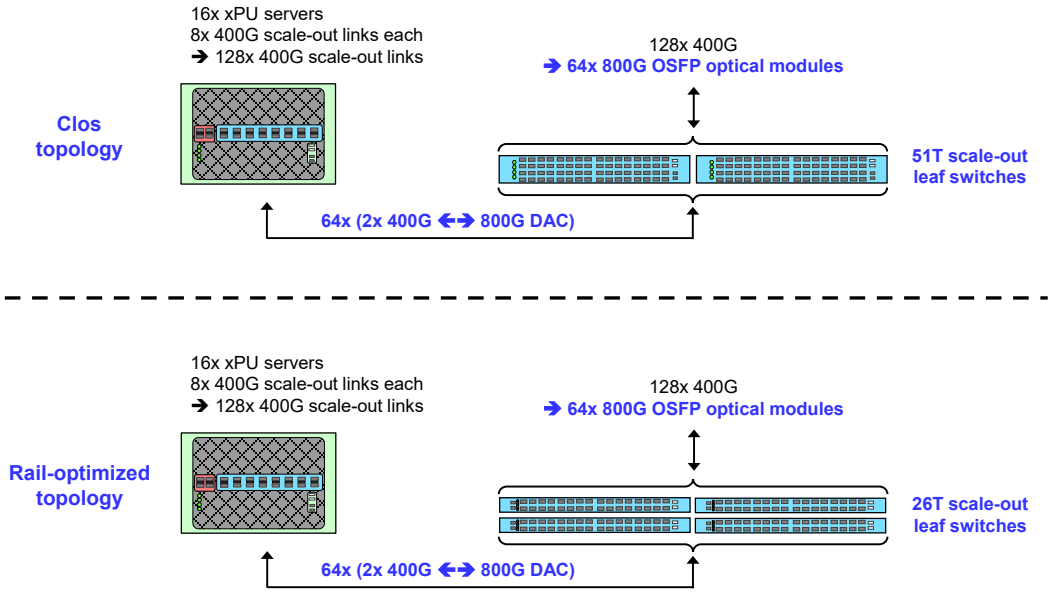


Figure 10. OPG-128 with scale-out uplinks. The scale-out network switch count, DAC count, and optical-module count are indicated. Top panel: Clos topology. Bottom panel: rail-optimized topology.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7.1.1. Scale-out fabric with Clos topology

In this topology, for every xPU node all scale-out links connect to the same switch. These connections are implemented using 800G break-out DAC cables:

800G OSFP (switch) ←→ 2x 400G QSFP112 (NICs)

Figure 11 shows the port-numbering plan for a Clos-based scale-out fabric, and Tables 10 and 11 contain the connection maps for the two 51T scale-out leaf switches.

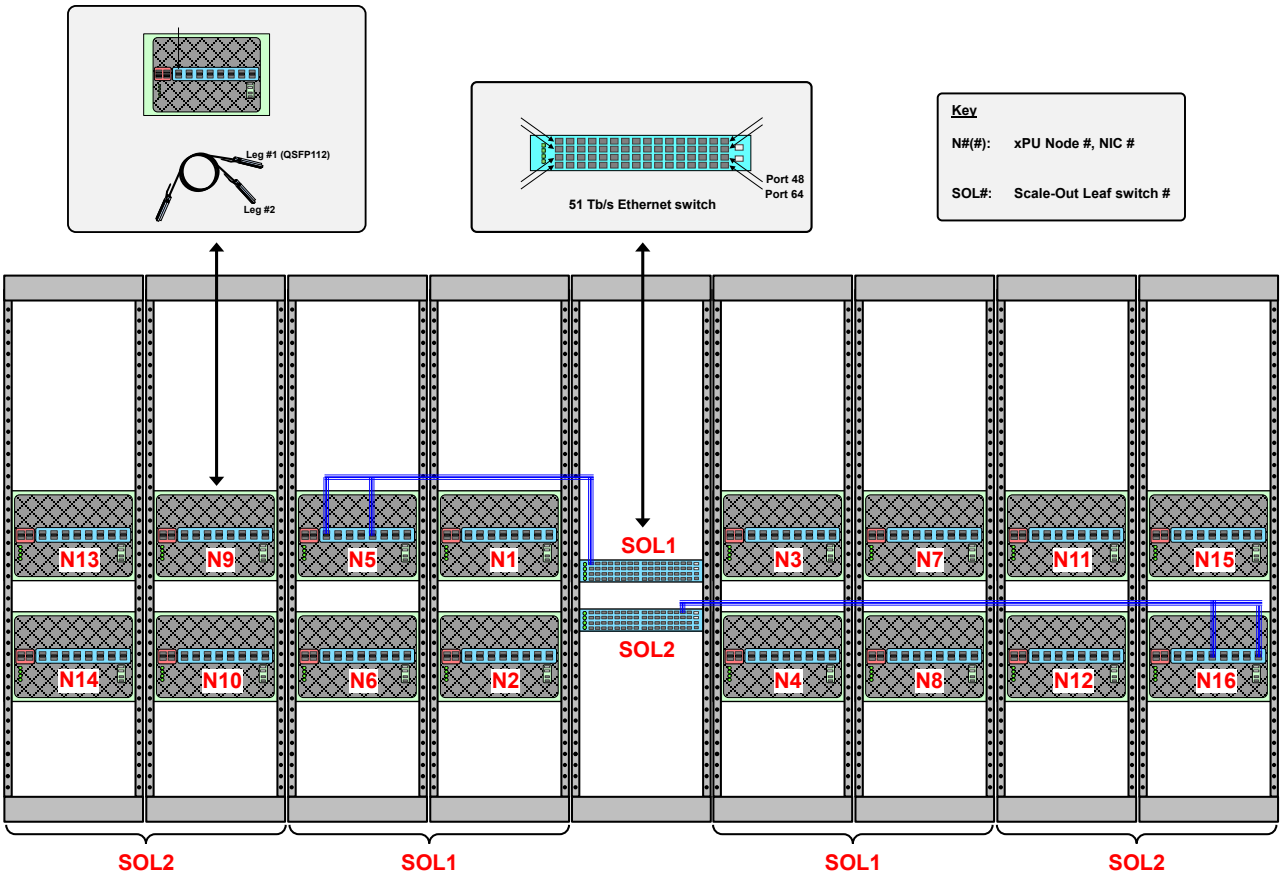


Figure 11. OPG-128 scale-out port-numbering plan. In this design scale-out switch #1 (SOL1) connects to the scale-out NICs in the inner four xPU-node racks, and scale-out switch #2 (SOL2) connects to the scale-out NICs in the outer four xPU-node racks.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

SOL1 port	NIC / spine		SOL1 port	NIC / spine		SOL1 port	NIC / spine		SOL1 port	NIC / spine
1	N1 (1)		17	N3 (1)		33	N5 (1)		49	N7 (1)
1	N1 (5)		17	N3 (5)		33	N5 (5)		49	N7 (5)
2	leaf-spine (2x 400G)		18	leaf-spine (2x 400G)		34	leaf-spine (2x 400G)		50	leaf-spine (2x 400G)
3	N1 (2)		19	N3 (2)		35	N5 (2)		51	N7 (2)
3	N1 (6)		19	N3 (6)		35	N5 (6)		51	N7 (6)
4	leaf-spine (2x 400G)		20	leaf-spine (2x 400G)		36	leaf-spine (2x 400G)		52	leaf-spine (2x 400G)
5	N1 (3)		21	N3(3)		37	N5 (3)		53	N7(3)
5	N1 (7)		21	N3(7)		37	N5 (7)		53	N7(7)
6	leaf-spine (2x 400G)		22	leaf-spine (2x 400G)		38	leaf-spine (2x 400G)		54	leaf-spine (2x 400G)
7	N1 (4)		23	N3(4)		39	N5 (4)		55	N7 (4)
7	N1 (8)		23	N3(8)		39	N5 (8)		55	N7 (8)
8	leaf-spine (2x 400G)		24	leaf-spine (2x 400G)		40	leaf-spine (2x 400G)		56	leaf-spine (2x 400G)
9	N2 (1)		25	N4 (1)		41	N6 (1)		57	N8 (1)
9	N2 (5)		25	N4 (5)		41	N6 (5)		57	N8 (5)
10	leaf-spine (2x 400G)		26	leaf-spine (2x 400G)		42	leaf-spine (2x 400G)		58	leaf-spine (2x 400G)
11	N2 (2)		27	N4 (2)		43	N6 (2)		59	N8 (2)
11	N2 (6)		27	N4(6)		43	N2(6)		59	N8 (6)
12	leaf-spine (2x 400G)		28	leaf-spine (2x 400G)		44	leaf-spine (2x 400G)		60	leaf-spine (2x 400G)
13	N2 (3)		29	N4(3)		45	N6 (3)		61	N8 (3)
13	N2 (7)		29	N4(7)		45	N6(7)		61	N8 (7)
14	leaf-spine (2x 400G)		30	leaf-spine (2x 400G)		46	leaf-spine (2x 400G)		62	leaf-spine (2x 400G)
15	N2 (4)		31	N4(4)		47	N6 (4)		63	N8 (4)
15	N2 (8)		31	N4 (8)		47	N6 (8)		63	N8 (8)
16	leaf-spine (2x 400G)		32	leaf-spine (2x 400G)		48	leaf-spine (2x 400G)		64	leaf-spine (2x 400G)

Table 10. OPG-128 scale-out fabric connection map for leaf switch #1 (SOL1), using Clos topology. All leaf-switch ports are 800G OSFP, all leaf switch -NIC connections are 400G passive DAC, and all leaf-spine connections are 400G optical. The scale-out leaf-spine connection map can be found in [1].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

SOL2 port	NIC / spine		SOL2 port	NIC / spine		SOL2 port	NIC / spine		SOL2 port	NIC / spine
1	N9 (1)		17	N11 (1)		33	N13 (1)		49	N15 (1)
1	N9 (5)		17	N11 (5)		33	N13 (5)		49	N15 (5)
2	leaf-spine (2x 400G)		18	leaf-spine (2x 400G)		34	leaf-spine (2x 400G)		50	leaf-spine (2x 400G)
3	N9 (2)		19	N11 (2)		35	N13 (2)		51	N15 (2)
3	N9 (6)		19	N11 (6)		35	N13 (6)		51	N15 (6)
4	leaf-spine (2x 400G)		20	leaf-spine (2x 400G)		36	leaf-spine (2x 400G)		52	leaf-spine (2x 400G)
5	N9 (3)		21	N11 (3)		37	N13 (3)		53	N15 (3)
5	N9 (7)		21	N11 (7)		37	N13 (7)		53	N15 (7)
6	leaf-spine (2x 400G)		22	leaf-spine (2x 400G)		38	leaf-spine (2x 400G)		54	leaf-spine (2x 400G)
7	N9 (4)		23	N11 (4)		39	N13 (4)		55	N15 (4)
7	N9 (8)		23	N11 (8)		39	N13 (8)		55	N15 (8)
8	leaf-spine (2x 400G)		24	leaf-spine (2x 400G)		40	leaf-spine (2x 400G)		56	leaf-spine (2x 400G)
9	N10 (1)		25	N12 (1)		41	N14 (1)		57	N16 (1)
9	N10 (5)		25	N12 (5)		41	N14 (5)		57	N16 (5)
10	leaf-spine (2x 400G)		26	leaf-spine (2x 400G)		42	leaf-spine (2x 400G)		58	leaf-spine (2x 400G)
11	N10 (2)		27	N12 (2)		43	N14 (2)		59	N16 (2)
11	N10 (6)		27	N12 (6)		43	N14 (6)		59	N16 (6)
12	leaf-spine (2x 400G)		28	leaf-spine (2x 400G)		44	leaf-spine (2x 400G)		60	leaf-spine (2x 400G)
13	N10 (3)		29	N12 (3)		45	N14 (3)		61	N16 (3)
13	N10 (7)		29	N12 (7)		45	N14(7)		61	N16 (7)
14	leaf-spine (2x 400G)		30	leaf-spine (2x 400G)		46	leaf-spine (2x 400G)		62	leaf-spine (2x 400G)
15	N10 (4)		31	N12 (4)		47	N14 (4)		63	N16 (4)
15	N10 (8)		31	N12 (8)		47	N14 (8)		63	N16 (8)
16	leaf-spine (2x 400G)		32	leaf-spine (2x 400G)		48	leaf-spine (2x 400G)		64	leaf-spine (2x 400G)

Table 11. OPG-128 scale-out fabric connection map for leaf switch #2 (SOL2), using Clos topology. All leaf-switch ports are 800G OSFP, all leaf switch -NIC connections are 400G passive DAC, and all leaf-spine connections are 400G optical. The scale-out leaf-spine connection map can be found in [1].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7.1.2. Scale-out fabric with rail-optimized topology

In this topology, each of the four 26T scale-out switches is connected to two of the scale-out links in each of the xPU nodes (Figure 12). These connections are implemented using 800G break-out DAC cables:

800G OSFP (switch) ←→ 2x 400G QSFP112 (NICs).

The connection maps are shown in Tables 12 -15.

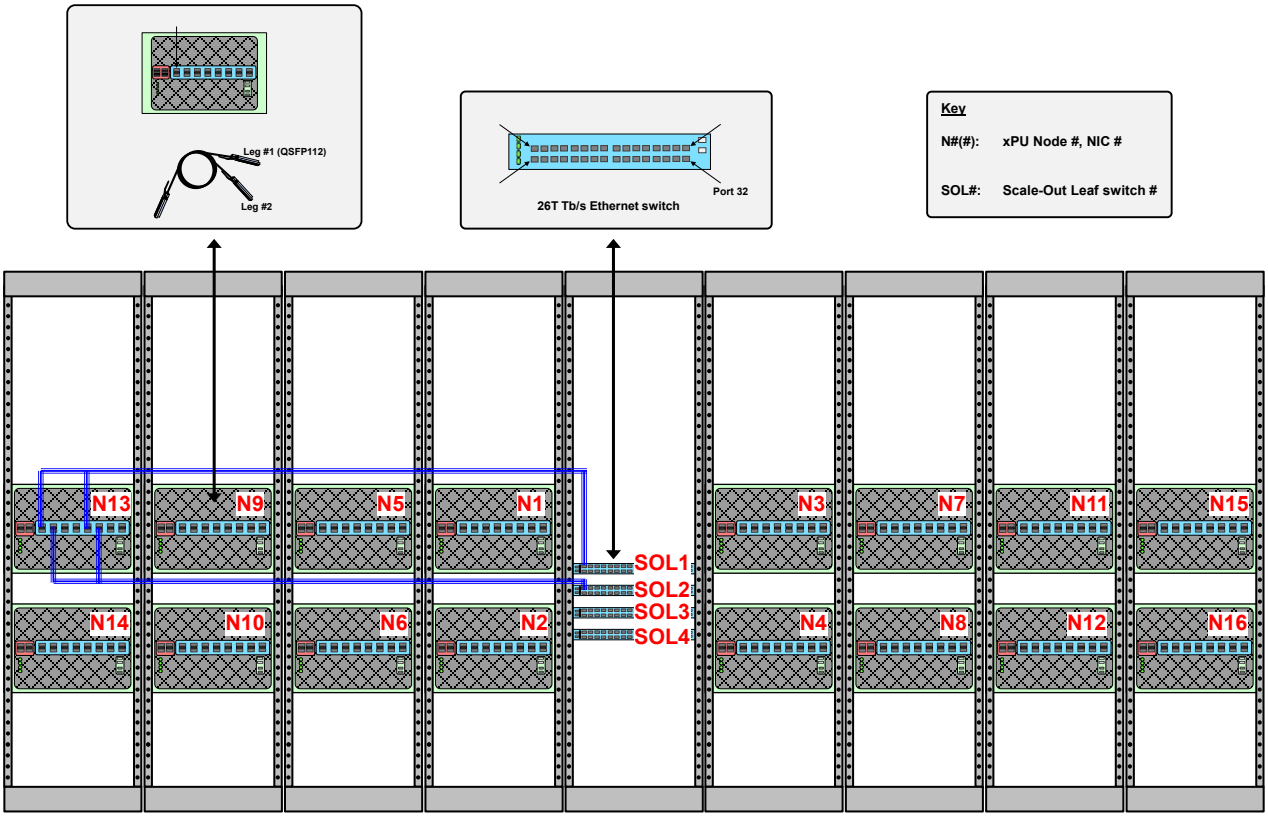


Figure 12. OPG-128 scale-out port-numbering plan for the rail-optimized topology. In this design each of the four scale-out switch connects to two scale-out NICs in each xPU-node.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 12

SOL1 port	NIC / spine		SOL1 port	NIC / spine		SOL1 port	NIC / spine		SOL1 port	NIC / spine
1	N1 (1)		9	N5 (1)		17	N9 (1)		25	N13 (1)
1	N1 (5)		9	N5 (5)		17	N9 (5)		25	N13 (5)
2	leaf-spine (2x 400G)		10	leaf-spine (2x 400G)		18	leaf-spine (2x 400G)		26	leaf-spine (2x 400G)
3	N2 (1)		11	N6 (1)		19	N10 (1)		27	N14 (1)
3	N2 (5)		11	N6 (5)		19	N10 (5)		27	N14 (5)
4	leaf-spine (2x 400G)		12	leaf-spine (2x 400G)		20	leaf-spine (2x 400G)		28	leaf-spine (2x 400G)
5	N3 (1)		13	N7 (1)		21	N11 (1)		29	N15 (1)
5	N3 (5)		13	N7 (5)		21	N11 (5)		29	N15 (5)
6	leaf-spine (2x 400G)		14	leaf-spine (2x 400G)		22	leaf-spine (2x 400G)		30	leaf-spine (2x 400G)
7	N4 (1)		15	N8 (1)		23	N12 (1)		31	N16 (1)
7	N4 (5)		15	N8 (5)		23	N12 (5)		31	N16 (5)
8	leaf-spine (2x 400G)		16	leaf-spine (2x 400G)		24	leaf-spine (2x 400G)		32	leaf-spine (2x 400G)

Table 13

SOL2 port	NIC / spine		SOL2 port	NIC / spine		SOL2 port	NIC / spine		SOL2 port	NIC / spine
1	N1 (2)		9	N5 (2)		17	N9 (2)		25	N13 (2)
1	N1 (6)		9	N5 (6)		17	N9 (6)		25	N13 (6)
2	leaf-spine (2x 400G)		10	leaf-spine (2x 400G)		18	leaf-spine (2x 400G)		26	leaf-spine (2x 400G)
3	N2 (2)		11	N6 (2)		19	N10 (2)		27	N14 (2)
3	N2 (6)		11	N6 (6)		19	N10 (6)		27	N14 (6)
4	leaf-spine (2x 400G)		12	leaf-spine (2x 400G)		20	leaf-spine (2x 400G)		28	leaf-spine (2x 400G)
5	N3 (2)		13	N7 (2)		21	N11 (2)		29	N15 (2)
5	N3 (6)		13	N7 (6)		21	N11 (6)		29	N15 (6)
6	leaf-spine (2x 400G)		14	leaf-spine (2x 400G)		22	leaf-spine (2x 400G)		30	leaf-spine (2x 400G)
7	N4 (2)		15	N8 (2)		23	N12 (2)		31	N16 (2)
7	N4 (6)		15	N8 (6)		23	N12 (6)		31	N16 (6)
8	leaf-spine (2x 400G)		16	leaf-spine (2x 400G)		24	leaf-spine (2x 400G)		32	leaf-spine (2x 400G)

Tables 12 and 13: OPG-128 scale-out fabric connection map for leaf switches #1 and #2 (SOL1 & SOL2), using the rail-optimized topology. All leaf-switch ports are 800G OSFP, all leaf switch-NIC connections are 400G passive DAC, and all leaf-spine connections are 400G optical. The scale-out leaf-spine connection map can be found in [1].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 14

SOL3 port	NIC / spine		SOL3 port	NIC / spine		SOL3 port	NIC / spine		SOL3 port	NIC / spine
1	N1 (3)		9	N5 (3)		17	N9 (3)		25	N13 (3)
1	N1 (7)		9	N5 (7)		17	N9 (7)		25	N13 (7)
2	leaf-spine (2x 400G)		10	leaf-spine (2x 400G)		18	leaf-spine (2x 400G)		26	leaf-spine (2x 400G)
3	N2 (3)		11	N6 (3)		19	N10 (3)		27	N14 (3)
3	N2 (7)		11	N6 (7)		19	N10 (7)		27	N14 (7)
4	leaf-spine (2x 400G)		12	leaf-spine (2x 400G)		20	leaf-spine (2x 400G)		28	leaf-spine (2x 400G)
5	N3 (3)		13	N7 (3)		21	N11 (3)		29	N15 (3)
5	N3 (7)		13	N7 (7)		21	N11 (7)		29	N15 (7)
6	leaf-spine (2x 400G)		14	leaf-spine (2x 400G)		22	leaf-spine (2x 400G)		30	leaf-spine (2x 400G)
7	N4 (3)		15	N8 (3)		23	N12 (3)		31	N16 (3)
7	N4 (7)		15	N8 (7)		23	N12 (7)		31	N16 (7)
8	leaf-spine (2x 400G)		16	leaf-spine (2x 400G)		24	leaf-spine (2x 400G)		32	leaf-spine (2x 400G)

Table 15

SOL4 port	NIC / spine		SOL1 port	NIC / spine		SOL4 port	NIC / spine		SOL4 port	NIC / spine
1	N1 (4)		9	N5 (4)		17	N9 (4)		25	N13 (4)
1	N1 (8)		9	N5 (8)		17	N9 (8)		25	N13 (8)
2	leaf-spine (2x 400G)		10	leaf-spine (2x 400G)		18	leaf-spine (2x 400G)		26	leaf-spine (2x 400G)
3	N2 (4)		11	N6 (4)		19	N10 (4)		27	N14 (4)
3	N2 (8)		11	N6 (8)		19	N10 (8)		27	N14 (8)
4	leaf-spine (2x 400G)		12	leaf-spine (2x 400G)		20	leaf-spine (2x 400G)		28	leaf-spine (2x 400G)
5	N3 (4)		13	N7 (4)		21	N11 (4)		29	N15 (4)
5	N3 (8)		13	N7 (8)		21	N11 (8)		29	N15 (8)
6	leaf-spine (2x 400G)		14	leaf-spine (2x 400G)		22	leaf-spine (2x 400G)		30	leaf-spine (2x 400G)
7	N4 (4)		15	N8 (4)		23	N12 (4)		31	N16 (4)
7	N4 (8)		15	N8 (8)		23	N12 (8)		31	N16 (8)
8	leaf-spine (2x 400G)		16	leaf-spine (2x 400G)		24	leaf-spine (2x 400G)		32	leaf-spine (2x 400G)

Tables 14 and 15: OPG-128 scale-out fabric connection map for leaf switches #3 and #4 (SOL3 & SOL4), using the rail-optimized topology. All leaf-switch ports are 800G OSFP, all leaf switch -NIC connections are 400G passive DAC, and all leaf-spine connections are 400G optical. The scale-out leaf-spine connection map can be found in [1].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7.2. SO-C fabric

This fabric, illustrated in Figures 13 and 14, provides connectivity between the CPUs in the xPU nodes / pods. It is part of the main data plane of the cluster and handles core production traffic for the applications and services running on the cluster. Tables 16 and 17 contain the connection maps.

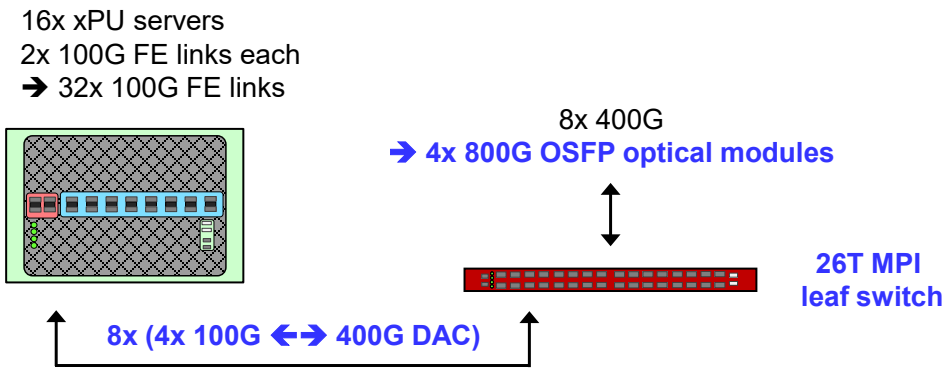


Figure 13. OPG-128 SO-C network. The DAC count and optical-module counts are indicated.

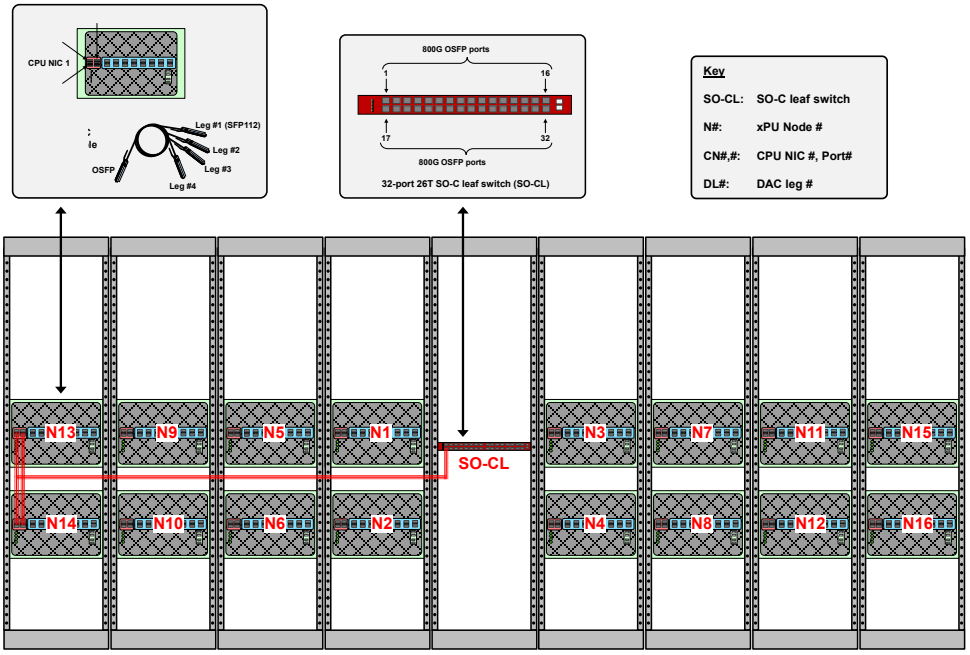


Figure 14. OPG-128 CPU-NIC numbering plan and SO-C switch port-numbering plan.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 16

xPU node	CPU NIC	NIC port	Switch port	Cable leg		xPU node	CPU NIC	NIC port	Switch port	Cable leg
1	1	1	22	1		9	1	1	19	1
1	2	1	22	2		9	2	1	19	2
2	1	1	22	3		10	1	1	19	3
2	2	1	22	4		10	2	1	19	4
3	1	1	23	1		11	1	1	26	1
3	2	1	23	2		11	2	1	26	2
4	1	1	23	3		12	1	1	26	3
4	2	1	23	4		12	2	1	26	4
5	1	1	20	1		13	1	1	17	1
5	2	1	20	2		13	2	1	17	2
6	1	1	20	3		14	1	1	17	3
6	2	1	20	4		14	2	1	17	4
7	1	1	25	1		15	1	1	28	1
7	2	1	25	2		15	2	1	28	2
8	1	1	25	3		16	1	1	28	3
8	2	1	25	4		16	2	1	28	4

Table 17

SO-C uplink ports (2x 400G)	Network element
18	SO-C spine switch
21	SO-C spine switch
24	SO-C spine switch
27	SO-C spine switch

Tables 16 & 17. OPG-128 SO-C connection plan. Table 16 (top): downlink leaf switch-NIC 100GbE connections. Table 17 (bottom): uplink leaf-spine 400GbE connections. The SO-C leaf-spine connection map can be found in [1].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7.3. Storage network

This network, illustrated in Figures 15 and 16, is responsible for data flow between the storage system and the pods/xPU-nodes. For an XOC-N that spans multiple OPG-Ms, the OPG-M storage uplinks connect to the storage spine. Tables 18-20 contain the connection maps.

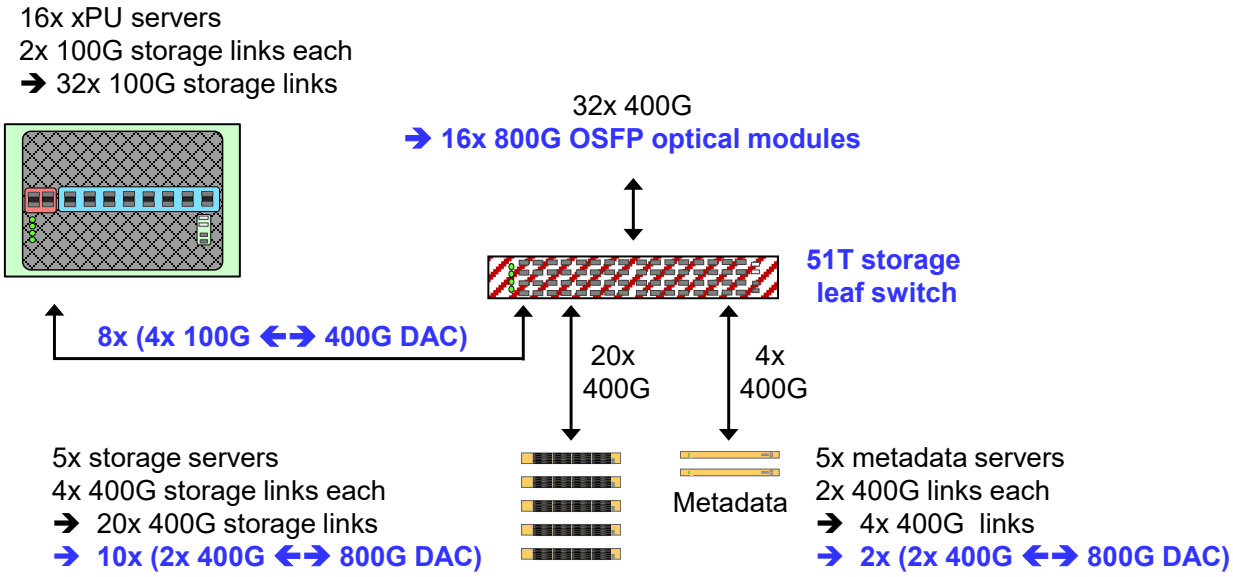


Figure 15. OPG-128 storage fabric. The switch count, DAC count, and optical-module count are indicated.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

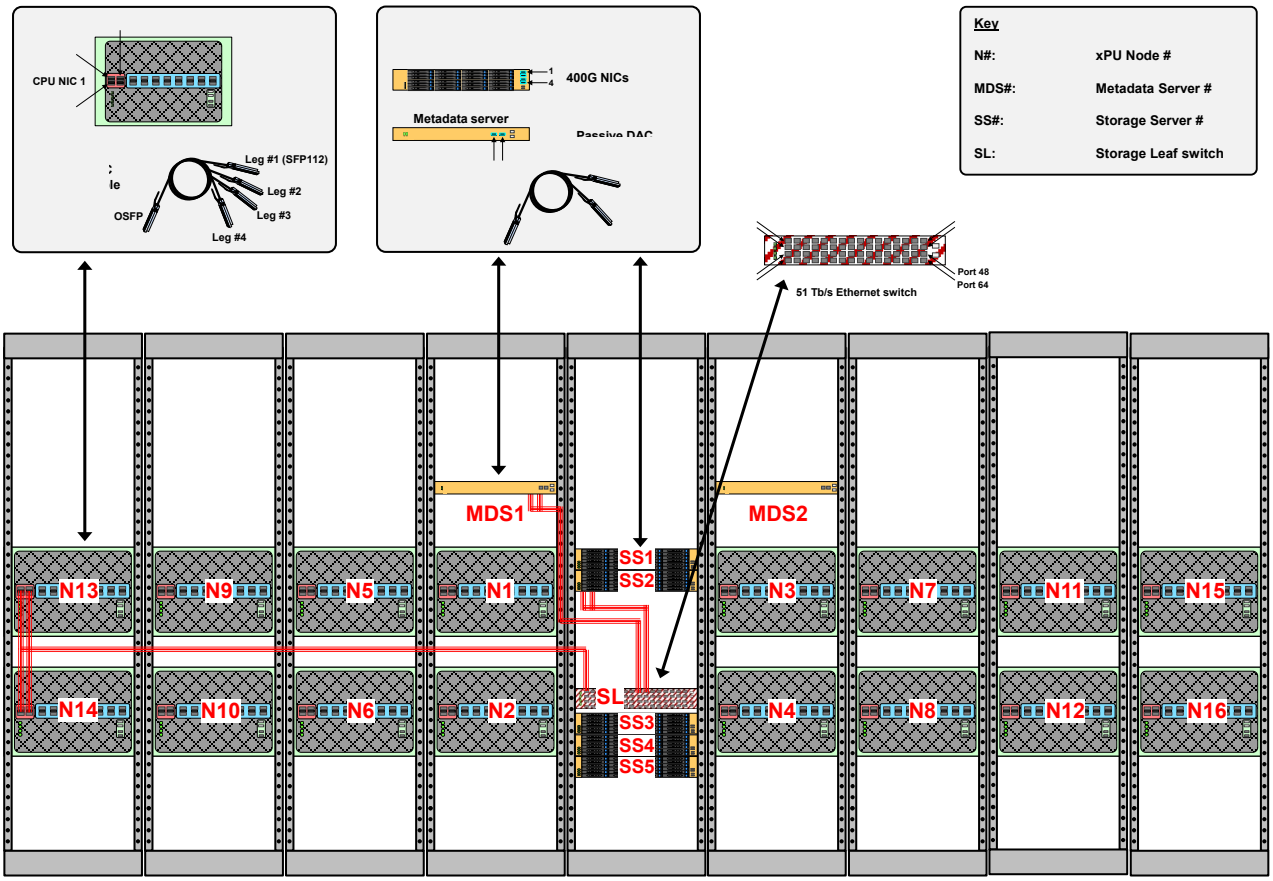


Figure 16. OPG-128 CPU-NIC numbering plan, server NIC numbering plan, and storage switch port-numbering plan.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 18

xPU node	CPU NIC	NIC port	Switch port	Cable leg		xPU node	CPU NIC	NIC port	Switch port	Cable leg
1	1	2	1	1		9	1	2	17	1
1	2	2	1	2		9	2	2	17	2
2	1	2	1	3		10	1	2	17	3
2	2	2	1	4		10	2	2	17	4
3	1	2	3	1		11	1	2	19	1
3	2	2	3	2		11	2	2	19	2
4	1	2	3	3		12	1	2	19	3
4	2	2	3	4		12	2	2	19	4
5	1	2	5	1		13	1	2	21	1
5	2	2	5	2		13	2	2	21	2
6	1	2	5	3		14	1	2	21	3
6	2	2	5	4		14	2	2	21	4
7	1	2	7	1		15	1	2	23	1
7	2	2	7	2		15	2	2	23	2
8	1	2	7	3		16	1	2	23	3
8	2	2	7	4		16	2	2	23	4

Table 19

Network element	NIC	Switch port	Cable leg
Storage server 1 (SS1)	1	33	1
Storage server 1 (SS1)	2	33	2
Storage server 1 (SS1)	3	35	1
Storage server 1 (SS1)	4	35	2
Storage server 2 (SS2)	1	37	1
Storage server 2 (SS2)	2	37	2
Storage server 2 (SS2)	3	39	1
Storage server 2 (SS2)	4	39	2
Storage server 3 (SS3)	1	41	1
Storage server 3 (SS3)	2	41	2
Storage server 3 (SS3)	3	49	1
Storage server 3 (SS3)	4	49	2
Storage server 4 (SS4)	1	51	1
Storage server 4 (SS4)	2	51	2
Storage server 4 (SS4)	3	53	1
Storage server 4 (SS4)	4	53	2
Storage server 5 (SS5)	1	55	1
Storage server 5 (SS5)	2	55	2
Storage server 5 (SS5)	3	57	1
Storage server 5 (SS5)	4	57	2
Metadata server 1 (MDS1)	1	59	1
Metadata server 1 (MDS1)	2	59	2
Metadata server 2 (MDS2)	1	61	1
Metadata server 2 (MDS2)	2	61	2

Tables 18 &19. OPG-128 storage fabric connection plan. Table 18 (top): leaf switch-NIC connections. Table 19 (middle): server-leaf switch connections. The storage leaf-spine connection map is found in [1].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Storage uplink port (OSFP)	Optical module port (400G)	Network element
2	1	Storage spine switch
2	2	Storage spine switch
4	1	Storage spine switch
4	2	Storage spine switch
6	1	Storage spine switch
6	2	Storage spine switch
18	1	Storage spine switch
18	2	Storage spine switch
20	1	Storage spine switch
20	2	Storage spine switch
22	1	Storage spine switch
22	2	Storage spine switch
34	1	Storage spine switch
34	2	Storage spine switch
36	1	Storage spine switch
36	2	Storage spine switch
38	1	Storage spine switch
38	2	Storage spine switch
40	1	Storage spine switch
40	2	Storage spine switch
42	1	Storage spine switch
42	2	Storage spine switch
50	1	Storage spine switch
50	2	Storage spine switch
52	1	Storage spine switch
52	2	Storage spine switch
54	1	Storage spine switch
54	2	Storage spine switch
56	1	Storage spine switch
56	2	Storage spine switch
58	1	Storage spine switch
58	2	Storage spine switch

Table 20. OPG-128 storage fabric connection plan, uplink leaf-spine connections. The storage leaf-spine connection map can be found in [1].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7.4. In-band management network

This network, illustrated in Figure 17, connects directly to the CPUs in the pods/xPU nodes, storage servers, and metadata servers. The in-band network provides a physically separate path for node provisioning, job scheduling, data movement, telemetry collection, and software updates. The OPG-M in-band uplinks connect to the head-end nodes at the spine / aggregation layer. Below are the network element port-numbering plans (Figure 18) and connection maps (Tables 21 and 22) for the in-band network. The connection maps are designed to leave the left-most 25G ports and right-most 100G ports open to facilitate future network rearrangement or expansion.

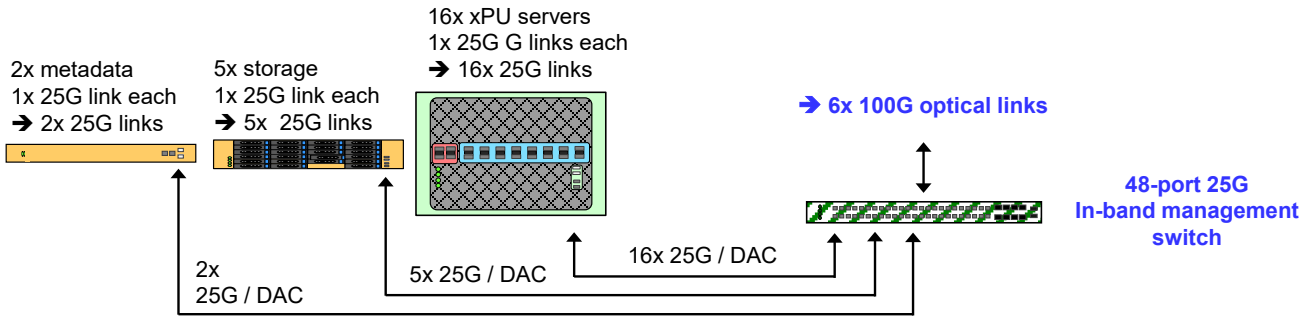


Figure 17. OPG-128 in-band management network. The switch count, jumper-cable count, and optical-module (QSFP28) count are indicated.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

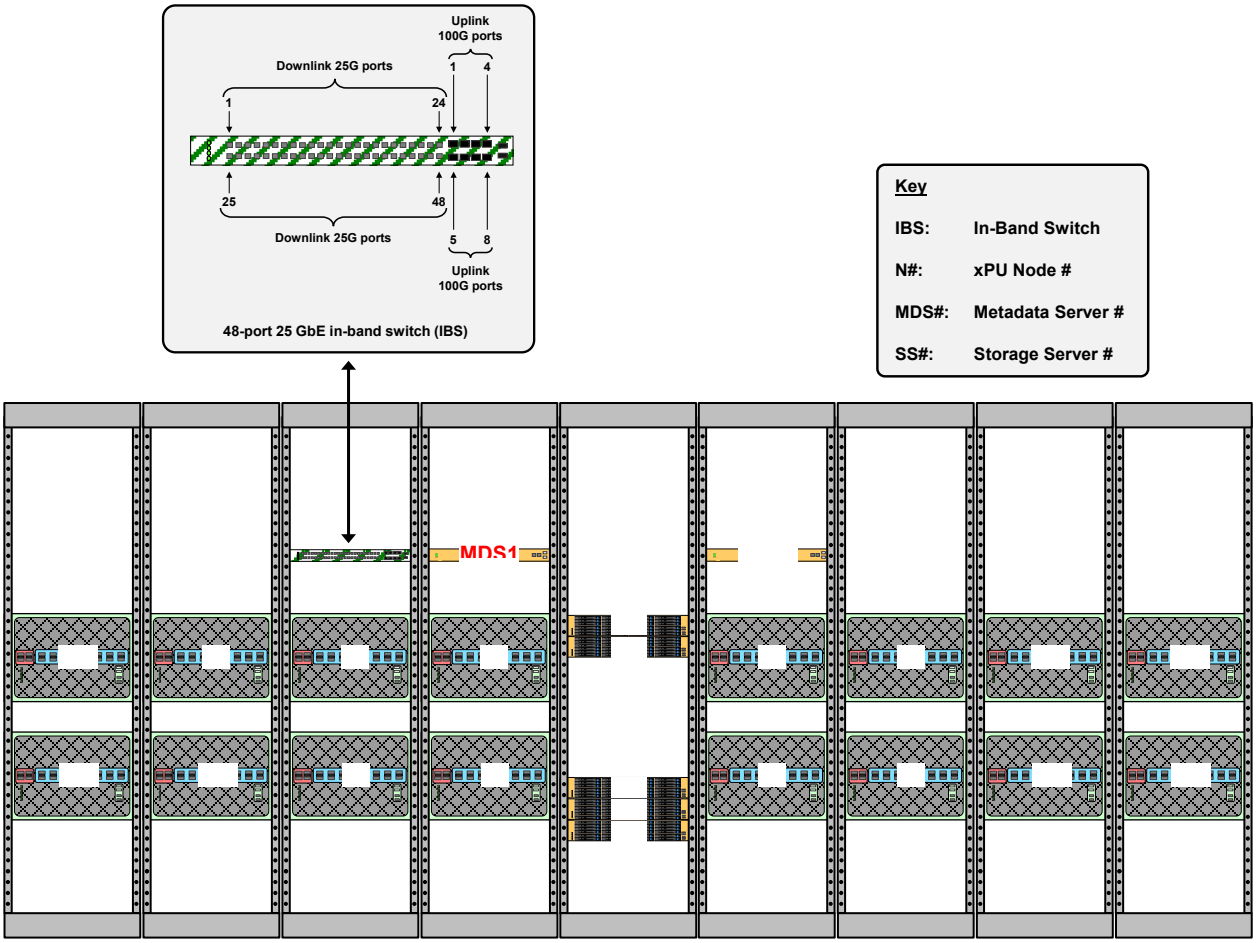


Figure 18. OPG-128 server-numbering and in-band management-switch port-numbering plans.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 21

IBS 25G downlink port	Network element		IBS downlink port	Network element		IBS downlink port	Network element		IBS downlink port	Network element
1			13			25			37	MDS2
2			14	MDS1		26			38	SS3
3			15	SS1		27			39	SS4
4			16	SS2		28			40	SS5
5			17	N1		29			41	N3
6			18	N2		30			42	N4
7			19	N5		31			43	N7
8			20	N6		32			44	N8
9			21	N9		33			45	N11
10			22	N10		34			46	N12
11			23	N13		35			47	N15
12			24	N14		36			48	N16

Table 22

IBS 100G uplink port	Network element
1	In-band spine switch
2	In-band spine switch
3	In-band spine switch
4	
5	In-band spine switch
6	In-band spine switch
7	In-band spine switch
8	

Tables 21 & 22. OPG-128 In-band network connection plan. Table 21 (top): downlink leaf switch-network element 25GbE connections. Table 22 (bottom): uplink leaf-spine 100GbE connections. The in-band leaf-spine connection map can be found in [1].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7.5. Out-of-band management network

The OoB network, illustrated in Figure 19, connects directly to the baseboard management controllers (BMCs) on the pods/xPU nodes, storage servers, meta-data servers, switches, and power distribution units (PDU) in the OPG-M. Below are the network element port-numbering plans (Figure 20) and connection maps (Tables 23-26) for the in-band network. The connection maps are designed to leave the left-most 1G ports and right-most 10G ports open to facilitate future network rearrangement or expansion.

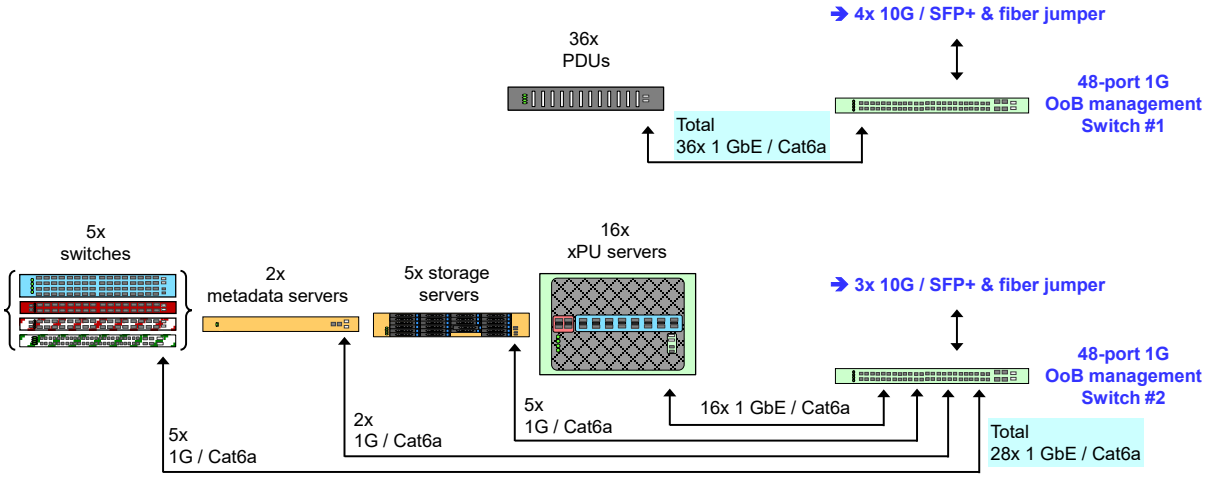


Figure 19. OPG-128 out-of-band management network. The switch count, jumper-cable count, and 10G uplink (SFP+ optical module & fiber jumper) count are indicated.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

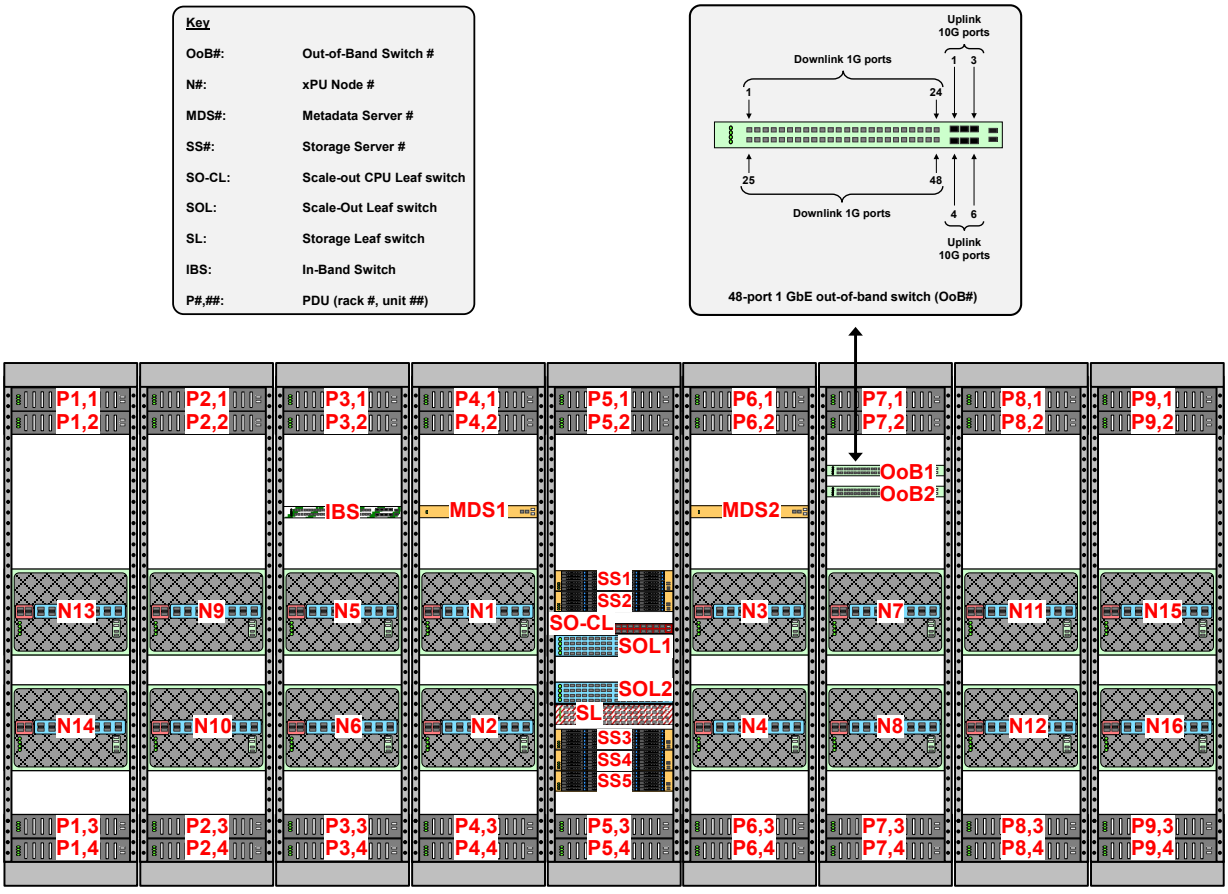


Figure 20. OPG-128 network-element numbering plan out-of-band management-switch port-numbering plan.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 23

OoB1 1G downlink port	Network element		OoB1 1G downlink port	Network element		OoB1 1G downlink port	Network element		OoB1 1G downlink port	Network element
1			13	P2,3		25			37	P7,1
2			14	P2,4		26			38	P7,2
3			15	P3,1		27			39	P7,3
4			16	P3,2		28			40	P7,4
5			17	P3,3		29			41	P8,1
6			18	P3,4		30			42	P8,2
7	P1,1		19	P4,1		31	P5,3		43	P8,3
8	P1,2		20	P4,2		32	P5,4		44	P8,4
9	P1,3		21	P4,3		33	P6,1		45	P9,1
10	P1,4		22	P4,4		34	P6,2		46	P9,2
11	P2,1		23	P5,1		35	P6,3		47	P9,3
12	P2,2		24	P5,2		36	P6,4		48	P9,4

Table 24

OoB2 1G downlink port	Network element		OoB2 1G downlink port	Network element		OoB2 1G downlink port	Network element		OoB2 1G downlink port	Network element
1			13	N3		25			37	IBS
2			14	N4		26			38	MDS1
3			15	N5		27			39	MDS2
4			16	N6		28			40	SS1
5			17	N7		29			41	SS2
6			18	N8		30			42	SS3
7			19	N9		31			43	SS4
8			20	N10		32			44	SS5
9			21	N11		33			45	SO-CL
10			22	N12		34			46	SOL1
11	N1		23	N13		35	N15		47	SOL2
12	N2		24	N14		36	N16		48	SL

Table 25

OoB1 10G uplink port	Network element
1	OoB spine switch
2	OoB spine switch
3	
4	OoB spine switch
5	OoB spine switch
6	

Tables 23 – 25. OPG-128 out-of-band network connection plan. Tables 23 and 24: downlink leaf switch-network element 1GbE connections. Table 25: uplink leaf-spine 10GbE connections. The out-of-band leaf-spine connection map can be found in [1].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 26

OoB2 10G uplink port	Network element
1	OoB spine switch
2	
3	
4	OoB spine switch
5	OoB spine switch
6	

Table 26. Uplink leaf-spine 10GbE connections. The out-of-band leaf-spine connection map can be found in [1].

8. OPG-128 kit

Table 27 contains a list and count of the major hardware elements of an OPG-128.

Item	Number
Intrastructure	
OR v3 racks	9
5.8 kW power supplies	36
Horizontal cable management	Design choice Note 2
Fiber management	Design choice Note 3
Nodes / servers	
xPU nodes	16
Data-storage servers	5
Metadata servers	2
Data-storage servers	5
Metadata servers	2
Ethernet switches	
51T Scale-out leaf (64x OSFP)	4
26T Scale out-CPU leaf (32x OSFP)	1
51T Storage leaf (64x OSFP)	1
2T In-band leaf (48x 25G/SFP28)	1
Out-of-band leaf (48x 1G/RJ45)	2

Item	Number
Optical Modules	
800G single-mode OSFP (2x 400G)	84
100G single-mode QSFP28	6
10G single-mode SFP+	7
Optical fiber patch cables	
MTP-8/12 single-mode	168
LC single-mode	13
Optical-fiber cable break-out	
Cable assembly or modular	Design choice Note 4
Passive DAC cables	
OSFP <==> 2x 400G QSFP112	76
OSFP <==> 4x 100G QSFP112	16
SFP28 <==> SFP 28	23
Copper patch cables	
Cat6a jumper	64

Table 27. OPG-128 hardware kit.

Notes

1. Detailed descriptions of the equipment, including NIC requirements, are found in Section 6.
2. Horizontal cable-management design depends on equipment layout within the racks.
3. Fiber management hardware selection and rack design depends on data-center trunk/plenum cabling selection and policies.
4. Optical-fiber break-out hardware, needed in this design for interconnecting 100G and 10G optical links with trunk/plenum cables, design depends on data-center trunk/plenum cable selection and policies.

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

9. Network utilization

Table 28 summarizes the utilization and surplus capacity of the five networks contained by an OPG-128.

Network	Switch count	Switch capacity (Tb/s)	Ports per switch	Port speed (Gb/s)	Network capacity (Tb/s)	Network port count	Occupied ports	Network utilization	Available capacity (Tb/s)	Available capacity ports
Scale-out	2	51	64	800	102	128	128	100%	0	0
Scale out - CPU	1	26	32	800	26	32	12	38%	16	20
Storage	1	51	64	800	51	64	36	56%	22	28
In-band	1	2	48	25	2	48	23	48%	1	25
Out-of-band	2	0.05	48	1	0.1	96	64	67%	0.3	32

Table 28. OPG-128 network utilization and surplus capacity.

10. Adapting the OPG-128 design

The current design can be extended without network rearrangement, subject to rack-power constraints, by adding storage servers or metadata servers. As shown in Section 9, the in-band, storage, and out-of-band networks all have ample capacity to accommodate the addition of servers.

The OPG-128 design also can be modified for clusters with 128 or fewer xPUs. In this case, all xPU servers reside in a single OPG-M, and the XOC-N consists of the OPG-M plus the higher layer resources, which are described in [1]. In this case, there is no external scale-out fabric spine, so no uplinks are required for the scale-out switch, meaning a single 51T switch can provide the connectivity for the entire scale-out fabric (Figure 21). Similarly, there are no SO-C fabric uplinks. A few storage fabric uplinks still are required for connectivity to the higher-layer resources, and a 26T switch can be used for this fabric. For all networks except the scale-out fabric and storage-fabric, the other switch connection maps are identical to those developed above for multi-OPG designs. Table 15a contains the scale-out fabric connection map. Figure 22 shows the port-numbering plan for the storage fabric and Tables 15b and 15c contain the storage-fabric connection map. The design shown in Figures 21 and 22 and Tables 15a-15b describe a system with 128 xPUs; smaller systems can be implemented via straightforward depopulation of this design.

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

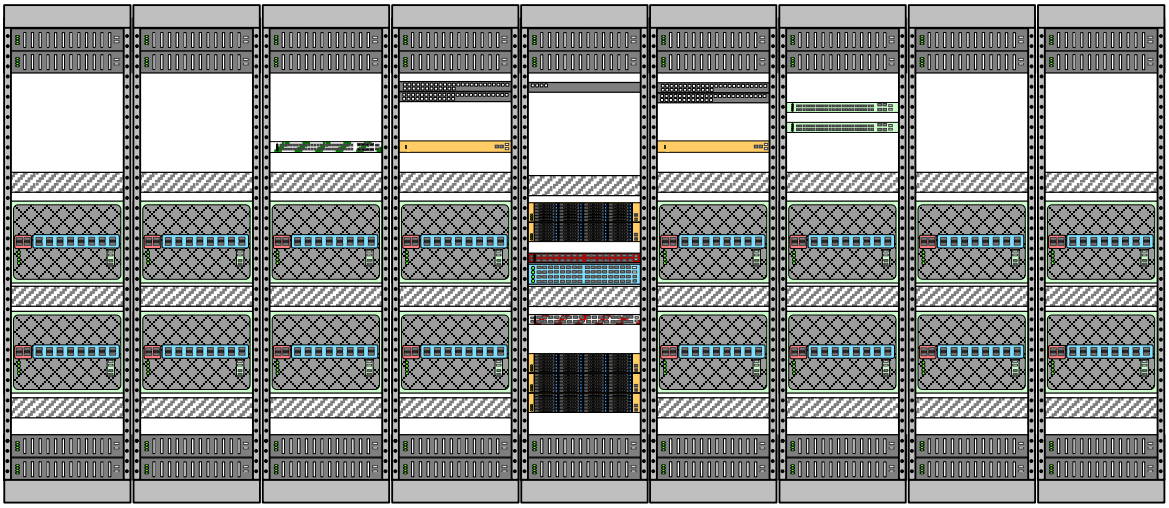


Figure 21. OPG-128 without scale-out or SO-C fabric uplinks.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

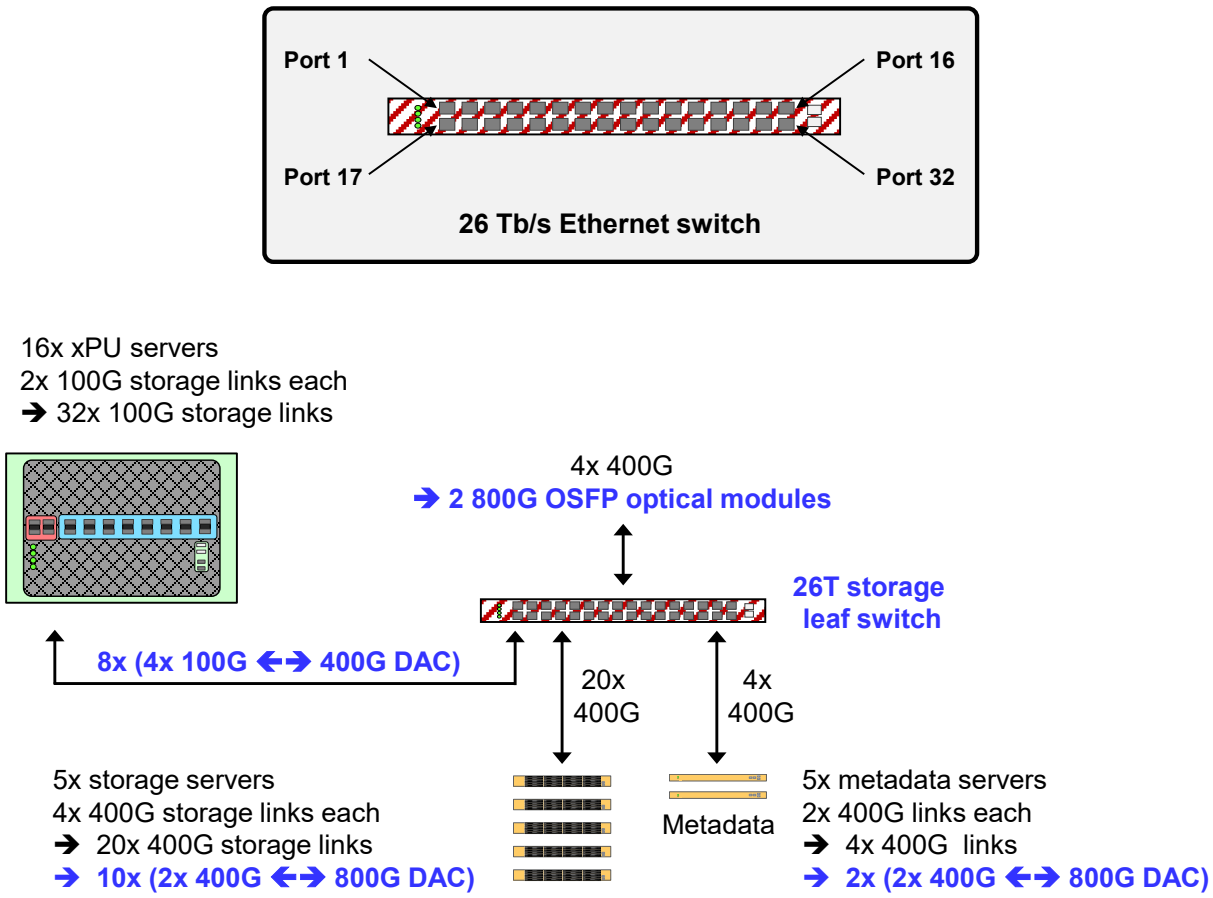


Figure 22. OPG-128 without scale-out or SO-C fabric uplinks.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

SOL1 port	NIC / spine		SOL1 port	NIC / spine		SOL1 port	NIC / spine		SOL1 port	NIC / spine
1	N13 (1)		17	N9 (1)		33	N5 (1)		49	N1 (1)
1	N13 (2)		17	N9 (2)		33	N5 (2)		49	N1 (2)
2	N13 (3)		18	N9 (3)		34	N1 (3)		50	N1 (3)
2	N13 (4)		18	N9 (4)		34	N1 (4)		50	N1 (4)
3	N13 (5)		19	N9 (5)		35	N1 (5)		51	N1 (5)
3	N13 (6)		19	N9 (6)		35	N1 (6)		51	N1 (6)
4	N13 (7)		20	N9 (7)		36	N1 (7)		52	N1 (7)
4	N13 (8)		20	N9 (8)		36	N1 (8)		52	N1 (8)
5	N14 (1)		21	N10 (1)		37	N2 (1)		53	N2 (1)
5	N14 (2)		21	N10 (2)		37	N2 (2)		53	N2 (2)
6	N14 (3)		22	N10 (3)		38	N2 (3)		54	N2 (3)
6	N14 (4)		22	N10 (4)		38	N2 (4)		54	N2 (4)
7	N14 (5)		23	N10 (5)		39	N2 (5)		55	N2 (5)
7	N14 (6)		23	N10 (6)		39	N2 (6)		55	N2 (6)
8	N14 (7)		24	N10 (7)		40	N2 (7)		56	N2 (7)
8	N16 (8)		24	N10 (8)		40	N2 (8)		56	N2 (8)
9	N14 (1)		25	N11 (1)		41	N3 (1)		57	N3 (1)
9	N15 (2)		25	N11 (2)		41	N3 (2)		57	N3 (2)
10	N15 (3)		26	N11 (3)		42	N3 (3)		58	N3 (3)
10	N15 (4)		26	N11 (4)		42	N3 (4)		58	N3 (4)
11	N15 (5)		27	N11 (5)		43	N3 (5)		59	N3 (5)
11	N15 (6)		27	N11 (6)		43	N3 (6)		59	N3 (6)
12	N15 (7)		28	N11 (7)		44	N3 (7)		60	N3 (7)
12	N15 (8)		28	N11 (8)		44	N3 (8)		60	N3 (8)
13	N16 (1)		29	N12 (1)		45	N4 (1)		61	N4 (1)
13	N16 (2)		29	N12 (2)		45	N4 (2)		61	N4 (2)
14	N16 (3)		30	N12 (3)		46	N4 (3)		62	N4 (3)
14	N16 (4)		30	N12 (4)		46	N4 (4)		62	N4 (4)
15	N16 (5)		31	N12 (5)		47	N4 (5)		63	N4 (5)
15	N16 (6)		31	N12 (6)		47	N4 (6)		63	N4 (6)
16	N16 (7)		32	N12 (7)		48	N4 (7)		64	N4 (7)
16	N16 (8)		32	N12 (8)		48	N4 (8)		64	N4 (8)

Table 29. Scale-out fabric connection map for a small OPG-M with 128 or fewer xPUs. All leaf-switch ports are 800G OSFP and all leaf switch -NIC connections are 400G passive DAC. The port-numbering plan is shown in Figure 12.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 30

xPU node	CPU NIC	NIC port	Switch port	Cable leg		xPU node	CPU NIC	NIC port	Switch port	Cable leg
1	1	2	17	1		9	1	2	21	1
1	2	2	17	2		9	2	2	21	2
2	1	2	17	3		10	1	2	21	3
2	2	2	17	4		10	2	2	21	4
3	1	2	18	1		11	1	2	22	1
3	2	2	18	2		11	2	2	22	2
4	1	2	18	3		12	1	2	22	3
4	2	2	18	4		12	2	2	22	4
5	1	2	19	1		13	1	2	23	1
5	2	2	19	2		13	2	2	23	2
6	1	2	19	3		14	1	2	23	3
6	2	2	19	4		14	2	2	23	4
7	1	2	20	1		15	1	2	24	1
7	2	2	20	2		15	2	2	24	2
8	1	2	20	3		16	1	2	24	3
8	2	2	20	4		16	2	2	24	4

Table 31

Network element	NIC	Switch port	Cable leg
Storage server 1 (SS1)	1	1	1
Storage server 1 (SS1)	2	1	2
Storage server 2 (SS2)	1	2	1
Storage server 2 (SS2)	2	2	2
Storage server 3 (SS3)	1	3	1
Storage server 3 (SS3)	2	3	2
Storage server 4 (SS4)	1	4	1
Storage server 4 (SS4)	2	4	2
Storage server 5 (SS5)	1	5	1
Storage server 5 (SS5)	2	5	2
Metadata server 1 (MDS1)	1	6	1
Metadata server 1 (MDS1)	2	6	2
Metadata server 2 (MDS2)	1	7	1
Metadata server 2 (MDS2)	2	7	2
Storage uplink 1 (2x 400G)	--	8	--
Storage uplink 2 (2x 400G)	--	9	--

Tables 30 and 31. Storage fabric connection map for a small OPG-M with 128 or fewer xPUs. All leaf-switch ports are 800G OSFP and all leaf switch -NIC connections are 400G passive DAC. The port-numbering plan is shown in Figure 23.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

11. References

[1] Lowell D. Lamb, et al, *Open Cluster for N xPUs (OCN-N) System Architecture,* 202x, Open Compute Project (2025).

[2] Lowell D. Lamb and Loren Staley, *Open Cluster Designs for AI,* Open Compute Project (2025).

[3] Madhusudan Iyengar, John Leung, Greg Sellman, and John Stuewe, Eds., *Open Systems for AI: Blueprint for a scalable infrastructure, Version 1.0.0*. Open Compute Project (2025). https://www.opencompute.org/documents/ocp-open-systems-for-ai-whitepaper-v1-0-0-final-pdf.

[4] Software for Open Networking in the Cloud (SONiC), 2023, The Linux Foundation, https://sonicfoundation.dev/.

[5] *Open Rack V3 Base specification*, 2023, Open Compute Project, https://www.opencompute.org/wiki/Open_Rack/SpecsAndDesigns

[6] Charles, Clos, *A study of non-blocking switching networks*, **Bell System Technical Journal 32** (2), 406–424 (March 1953).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

12. Glossary

Aggregation Network: A network that aggregates traffic from multiple OPG-Ms and provides connectivity to upstream networks and/or resources. In this context, the SO-C, storage, and OoB management networks all have separate aggregation networks.

BMC	A *BaseBoard Management Controller* is a small, specialized processor that is integrated into the motherboard of a managed device and is used for remote monitoring and management.
Cluster	The term cluster is used to denote all of the network elements that, taken together, comprise a scale-out domain.
CPU	Central Processing Unit. A CPU is a general-purpose processor that runs the high-layer applications in the OPG-M.
DAC Cable	Direct Attach Copper (cable). Short-reach (5m or less) copper cables used for high data rate, intra OPG connections. In this design all DACs are passive (no DSPs).
In-Band Management Network	This dedicated network connects directly to the CPUs in the pods/xPU nodes, storage servers, and meta-data servers. It provides a physically separate path to the head-end nodes located at the spine / aggregation level.
Leaf Switch	In a multi-tiered switching network, leaf switches occupy the lowest switching tier and connect to the end points.
Metadata Server	The metadata servers belong to the storage subsystem. They are specialized servers responsible for data storage, data transfer, and file-sharing requests. These units do not participate directly in the bulk transfer or storage of cluster data.
NIC	Network Interface Controller / Card. A hardware element that allows a CPU or xPU to connect to a network. PCIe cards are a common form factor for NICs.
OPG-M	The Open Pod Group of M xPUs is the principle building block of an XOC-N. An OPG-M is defined using OCP-based solutions wherever practical.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Out-of-Band Management Network	A dedicated, out-of-band management network for the pods/xPU-nodes, network switches, storage servers, etc., that make up the cluster.
PDU	Power distribution unit.
Pod	A pod refers to a liquid-cooled, rack-scale system containing a large number (typically 64 or more) of xPUs that are all members of a common scale-up domain, together with all of the other resources required for the function of the member xPUs.
Redfish Profile	A JSON document enumerating resources and properties that must, should, or could be implemented by a software service accessed via an API.
Scale-out Fabric	This network supports xPU-to-xPU communication between xPUs in different pods/xPU-nodes. Compared to the scale-up domain, scale-out networks can be much larger, connecting anywhere from a few dozen to over a hundred thousand xPUs. It is designated as a fabric because of its role in supporting high-speed data transfer internal to the cluster. Direct, external connections to this fabric typically are not permitted.
Scale-out – CPU (SO-C) Fabric	This fabric provides CPU-to-CPU connectivity within the cluster. It is designated as a fabric because of its role in supporting high-speed data transfer internal to the cluster. Direct, external connections to this fabric are not permitted.
Scale-up Fabric	This network connects a relatively small number of hardware accelerators to form a high-bandwidth, low-latency domain that acts as a single, logical unit and is optimized for distributed processing and highly parallel data movement. It is designated as a fabric because of its role in supporting high-speed data transfer internal to the cluster. Direct, external connections to this fabric typically are not permitted.
SONiC	An open-source network operating system (NOS) based on Linux.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Spine Switch	In a multi-tiered switching network, spine switches occupy the highest switching tier and cross-connect the leaf switches.
Storage Fabric	The network responsible for data flow between the storage system and the pods/xPU-nodes. It is designated as a fabric because of its role in supporting high-speed data transfer internal to the cluster. Direct, external connections to this fabric typically are not permitted.
Storage Server	High-density units that contain multiple NVMe drive bays. These units are attached to the storage fabric using high-speed Ethernet links, and are controlled by the metadata servers.
UBB	Refers to the *Open Accelerator Infrastructure Universal Baseboard (OAI-UBB)*, which is a key component within the OCP *Open Accelerator Infrastructure (OAI)* initiative. The OAI-UBB is a modular baseboard designed to support multiple Open Accelerator Modules (OAMs).
xPU	A generic term for the special-purpose AI accelerator responsible for the computationally intensive operations of AI models. Examples in this category include GPUs and TPUs.
xPU-Node	A small network element which contains a small number (typically eight) of xPUs that are all members of a common scale-up domain, together with all of the other resources required for the function of the member xPUs.
XOC-N	An xPU Open Cluster for N XPUs contains n XPUs, and is constructed using OCP-based solutions wherever practical.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

13. License

13.1. Creative Commons

OCP encourages participants to share their proposals, specifications and designs with the community. This is to promote openness and encourage continuous and open feedback. It is important to remember that by providing feedback for any such documents, whether in written or verbal form, that the contributor or the contributor's organization grants OCP and its members irrevocable right to use this feedback for any purpose without any further obligation.

It is acknowledged that any such documentation and any ancillary materials that are provided to OCP in connection with this document, including without limitation any white papers, articles, photographs, studies, diagrams, contact information (together, “Materials”) are made available under the Creative Commons Attribution-ShareAlike 4.0 International License found here: https://creativecommons.org/licenses/by-sa/4.0/, or any later version, and without limiting the foregoing, OCP may make the Materials available under such terms.

As a contributor to this document, all members represent that they have the authority to grant the rights and licenses herein. They further represent and warrant that the Materials do not and will not violate the copyrights or misappropriate the trade secret rights of any third party, including without limitation rights in intellectual property. The contributor(s) also represent that, to the extent the Materials include materials protected by copyright or trade secret rights that are owned or created by any third-party, they have obtained permission for its use consistent with the foregoing. They will provide OCP evidence of such permission upon OCP’s request. This document and any "Materials" are published on the respective project's wiki page and are open to the public in accordance with OCP's Bylaws and IP Policy. This can be found at http://www.opencompute.org/participate/legal-documents/. If you have any questions, please contact OCP.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

14. About Open Compute Foundation

The Open Compute Project (OCP) brings at-scale innovations and hyperscaler best practices to all, spanning technology domains from the data center to the edge, and the technology stack from silicon, to systems, to site facilities and services. The international OCP Community is made up of organizations and people from hyperscale and tier-2 cloud data center operators, communications providers, colocation providers, diverse enterprises, and technology vendors. The OCP Foundation fosters collaboration within the Community to tackle market challenges in an array of OCP Projects that create open contributions such as specifications, designs, and more. The OCP Solution Provider Program then recognizes the application of those contributions in compliant products, solutions, and data center facilities and services with designations like OCP Inspired, OCP Accepted and OCP Ready. With the tenets of openness, impact, efficiency, scale and sustainability, the OCP engages and educates thousands of engineers every year through many events and webinars. Across many initiatives the OCP Foundation and Community are meeting the market today and shaping the future.

Learn more at: www.opencompute.org.

Date: January 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.