OPEN CLUSTER DESIGNS FOR AI

Authors

Lowell D. Lamb, Broadcom

Loren Staley, Celestica

Date: January 2026

Version History

Date	Version #	Authors	Description
14 January 2026	1.0	Lowell D. Lamb & Loren Staley	Initial 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

The rapid growth of the AI/ML market has fostered the emergence of new categories of equipment and new infrastructure designs. Integration of these components into reliable, high-performance systems is a key problem in the industry. To address this challenge, this Initiative will focus on the development of System Architectures for Open Clusters and Open Pod Groups. These designs will promote the adoption of current best practices, will leverage Contributions developed by other OCP projects, and will align with the Open Systems for AI Strategic Initiative. A core objective of this Initiative is the delivery of designs which are detailed enough to provide a technical basis for procurement and implementation exercises. Further, it is intended that the level of detail will be sufficient to simplify software development by providing consistent hardware designs.

Date: January 2026

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

Table of Contents

Version History 2

1. Nomenclature 5
2. Introduction 5
3. Compliance with Open Compute Project Tenets 6
4. Open Cluster Architecture 7
5. System Architectures 13
6. Software components of a System Architecture 18
7. References 19
8. Glossary 19
9. License 21
   - 9.1. Creative Commons 21
10. About Open Compute Foundation 22

Date: January 2026

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

1. 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 XOC-N
XOC-N	xPU Open Cluster for N xPUs
xPU	Generic term for a special-purpose AI accelerator device

2. Introduction

The rapid growth of the AI/ML market has fostered the emergence of new categories of equipment and infrastructure designs. Integration of these components into reliable, high-performance systems is a key problem in the industry. Today, the vast majority of AI network implementations are one-off designs and the same problems are solved repeatedly. This Strategic Initiative will develop a family of System Architectures intended to accelerate AI network design by providing practical and complete solutions for common use cases, thereby lessening the redesign burden.

Real-world AI infrastructure design is multidimensional and complex, with interdependencies that often are understood poorly. To make progress in this work, a small number of criteria and objectives are adopted to simplify the problem:

- Use OCP solutions wherever possible.
- Where OCP solutions are not available, promote open solutions from other sources wherever possible.
- Focus on best practices. No attempt will be made to enumerate every potential option.
- Develop modular designs that facilitate scalability.
- Develop high-density and high-utilization designs.
- Develop full bandwidth, non-blocking designs.
- Develop complete designs.
- Provide physical equipment requirements (racks, network elements, cabling, cable management, etc.).
- Provide physical and logical topologies (rack layouts, connectivity maps, etc.).
- Define the key software elements of the OPG-M and XOC-N management planes.
- Provide power, cooling, and supporting infrastructure requirements.

Date: January 2026

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

While the System Architectures produced by this Strategic Initiative should benefit a wide range of AI infrastructure deployments, the market segment likely to benefit the most is the mid-sized enterprise category. These networks, which contain a few hundred to a few thousand xPUs, are sufficiently large that design complexity and systems integration challenges can overwhelm the IT staff of the enterprise. Accordingly, the first task of this program will be the development of designs addressing a common challenge facing mid-size enterprises today, namely systems with air-cooled xPU-nodes and 400G scale-out networks. In the second phase of this program, designs for the emerging class of liquid-cooled, rack-scale xPU pods and 800G scale-out networks will be developed. Thereafter, System Architectures for AI infrastructure with 1.6T scale-out networks, designs for disaggregated scale-up networks, etc., can be developed, depending on market needs.

3. Compliance with Open Compute Project Tenets

The program of this Strategic Initiative, which will be 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, the work of this Strategic Initiative should lower the design barrier for the enterprise sector and accelerate AI infrastructure deployment in this important market.

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

• **Scalability:** The modular nature of the designs developed by this work should naturally facilitate scalability, both during initial design and deployment, as well as subsequent 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 [1]), as well as the availability of open and proven solutions from other not-for-profit forms (e.g., SONiC [2]), demonstrate the commitment of this Strategic Initiative to open solutions that foster participation by multiple suppliers and enable ready extension or modification by users of the System Architectures.

• **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.

4. Open 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.

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

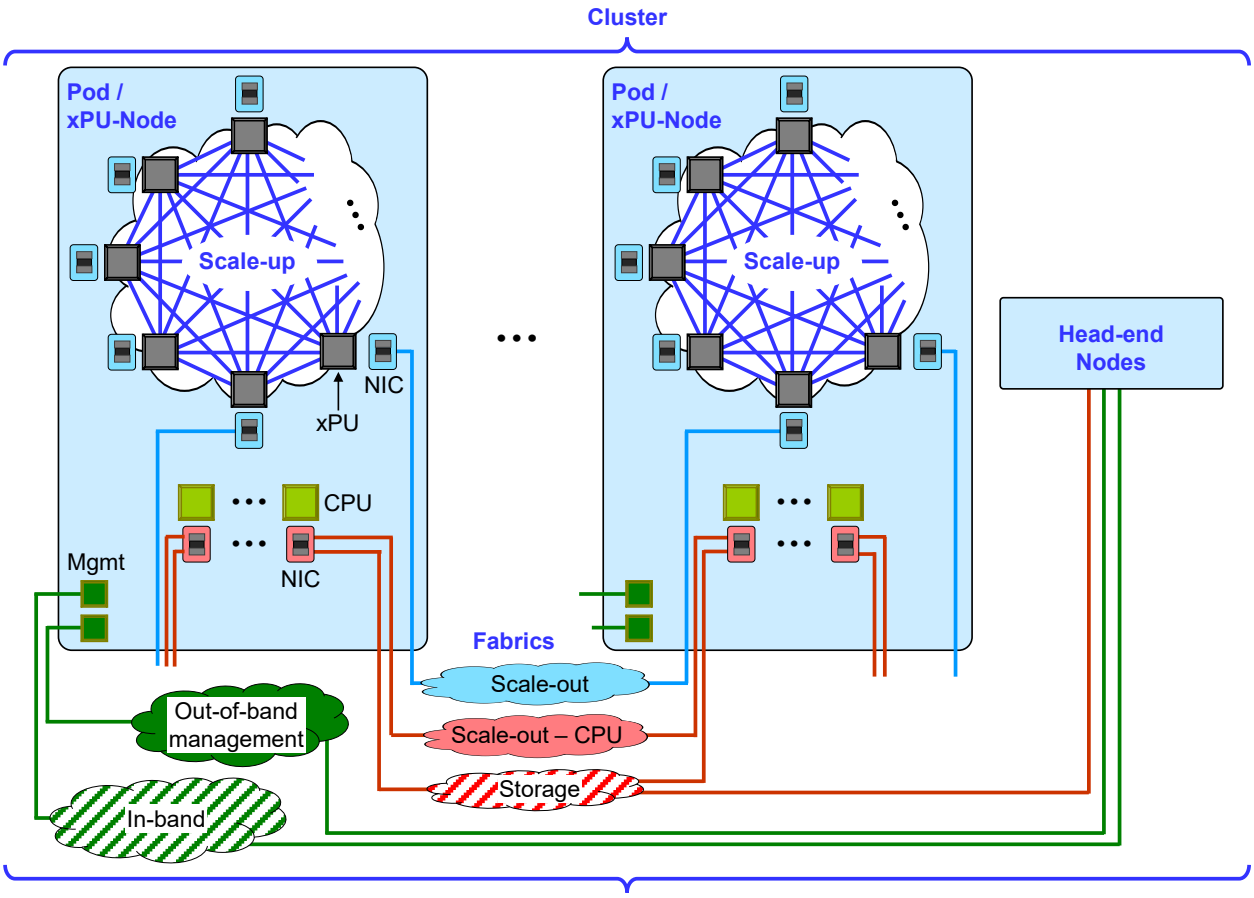

Figure 1: AI cluster networks and connectivity domains.

xPU Open Cluster for N xPUs (XOC-N)

Physical AI systems contain several distinct networks, and together they determine the fundamental structure of the AI infrastructure design. The term cluster denotes the collection of network elements that, taken together, comprise a scale-out domain. We introduce the term Open Cluster for N XPUs (XOC-N) to denote a cluster containing N XPUs, constructed using OCP-based solutions wherever practical. Following are brief descriptions of the networks found in modern AI network implementations. The architecture of each of these networks often is influenced by the target application, expected traffic profile, etc., and in practice a wide variety of designs are seen. For clarity, we describe a design that provides separate physical resources

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

for each network, i.e., a non-converged solution. This approach provides maximum isolation, security, and performance, while simplifying configuration, load-balancing, and troubleshooting. Furthermore, this approach illustrates cleanly the relationship between the intended function and resulting design of each network, an understanding of which is needed whether the final design is non-converged or not.

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. These accelerators have various marketing names (GPU, TPU, etc.) and are referred to by the generic term xPU in the present work. The term pod is used in the present work to describe all the xPUs that are members of a scale-up domain, taken together with the other resources required for the function of the member xPUs. As described by the OCP Open Systems for AI White Paper [3], the latest high-capacity pods span one or more equipment racks and can contain more than one hundred xPUs. In addition, we introduce the term xPU-node to denote the smaller scale-up domain form factor that typically contains a maximum of eight xPUs. This network is designated as a fabric because of its role in supporting high-speed data transfer wholly within the pod or xPU node. Direct, external connections to this fabric typically are not allowed.

Scale-out fabric: When xPU-to-xPU communication occurs between xPUs in different pods/xPU-nodes, the scale-out fabric is used. 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. The term cluster is used to describe the xPUs that are members of a scale-out domain, taken together with the resources required for the function of the member xPUs. This network is designated as a fabric because of its role in supporting high-speed data transfer wholly within the cluster. Direct, external connections to this fabric typically are not allowed.

Scale-out CPU (SO-C) fabric: This network 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. This network is designated as a fabric because of its role in supporting high-speed data transfer wholly within the cluster, and direct, external connections to this fabric typically are not allowed. Example uses of this fabric include transfer of data between concurrent data-producer and data-consumer clients and transport of MPI traffic.

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

Storage fabric: The storage network is responsible for data flow between the storage system, the pods/xPU-nodes, and certain head-end servers. Distributed, local storage within the cluster is critical for many applications, both training and inference. For training tasks, local storage often is used for holding training data, checkpoint data, etc. For inference tasks, local storage enables low-latency access to critical data such as user-profiles, pre-computed embeddings, feature vectors, document corpora, logs, etc. This network is designated as a fabric because of its role in supporting high-speed data transfer wholly within the cluster. Direct, external connections to this fabric typically are not allowed.

In-band network: This dedicated network connects directly to the CPUs in the pods/xPU nodes, storage servers, and servers. It provides a physically separate path connecting the resources in the xPU layer with the resources in the spine / aggregation / head-end layer. The head-end resources are responsible for supporting client activity and cluster management. Cluster management includes security, authentication, node provisioning, handling incoming job-submission requests, task-status requests, retrieving the results for completed tasks, job scheduling, load balancing, telemetry collection, container management, software updates, etc. The in-band network also connects to the firewalls and provides Internet access (“north-south traffic”).

Out-of-band (OoB) management network: This is a separate, low-bandwidth Ethernet network that connects directly to the baseboard management controllers (BMCs) on the pods/xPU nodes, storage servers, servers, and switches throughout the cluster. This fabric is used for power cycling, BIOS/firmware updates, and environmental monitoring. In the event that a host OS for a particular network element is down, or the in-band management network is down, the OoB network also can be used for recovery. The presence of both the in-band and out-of-band management networks provide management resiliency in AI clusters.

In this work, an XOC-N is built as an assembly of Open Pod Groups of M xPUs (OPG-M), together with the necessary resources at the spine and aggregation layers (see Figure 2). OPG-Ms are defined below.

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

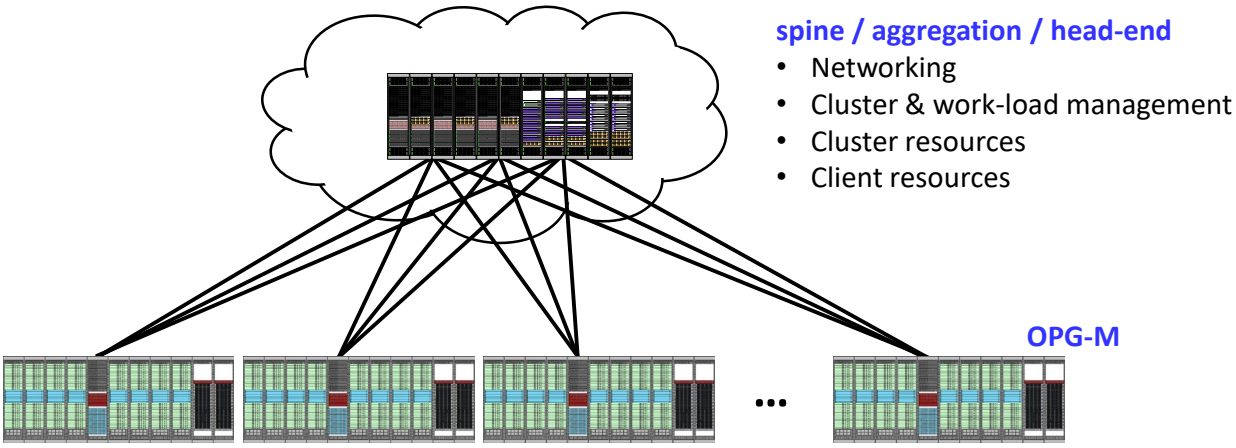

Figure 2: xPU Open Cluster for N xPUs (XOC-N).

Open Pod Group of 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 pods/xPU-nodes, plus other resources. The definition of OPG-Ms is a key focus of this Initiative and uses OCP specifications, implementation examples, and OCP recognized (Accepted/Inspired) equipment wherever possible. At the time of this writing, 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. It is expected that both form factors will be deployed for the foreseeable future, and will be included in the present work as separate System Architectures, given the significant differences in xPU density, link bandwidth, etc.

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

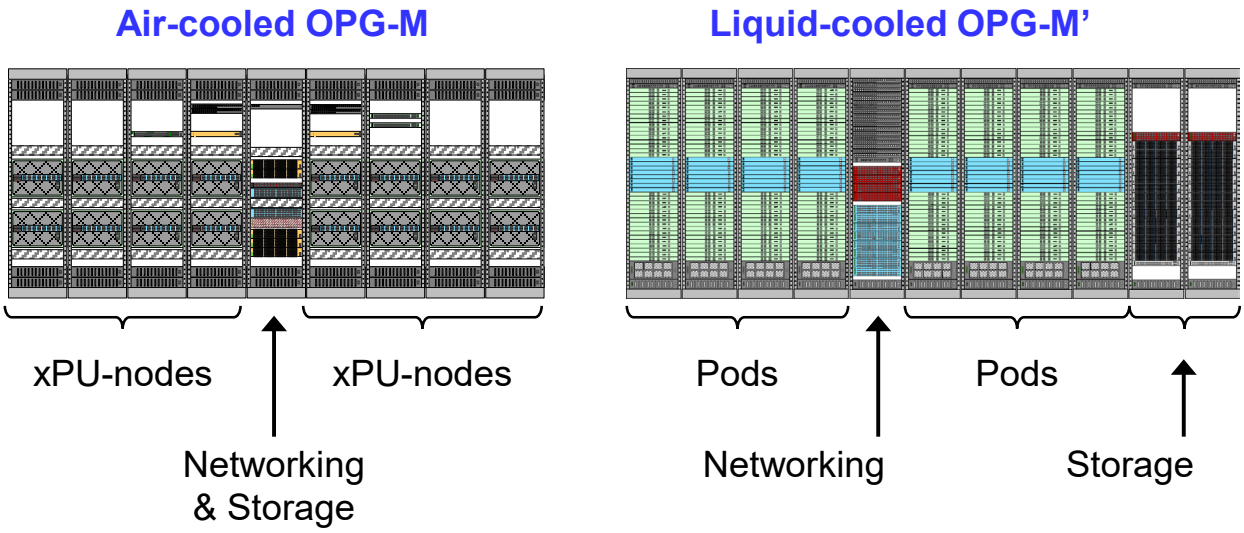

Figure 3: Conceptual drawings of an air-cooled OPG-M (left) and a liquid-cooled OPG-M' (right), showing the pods/xPU-nodes and associated resources that together constitute an OPG.

Physical and IT elements of OPC-Ms and XOC-Ns

The foundation of the OPG-Ms will be the OCP Open Rack v3 family of specifications [4]. Wherever possible and reasonable, specifications for the constituent network elements will be drawn directly from OCP Contributions for hardware components [5]. For some components, however, OCP Contributions either are missing or incomplete (e.g., storage systems). In such cases, a generic specification will be used that: (1) satisfies the needs of the OPG-M design; and (2) can be met by equipment from multiple suppliers. In all cases, the goal of this program will be the development of realistic and pragmatic System Architectures.

Software and management elements of OPG-Ms and XOC-Ns

There are many software instances needed to instantiate an AI/ML cluster. Two classes of software are not considered in the specification of OPG-M and XOC-N instances. These are the applications that use AI Accelerators, and the OS that runs on the host CPUs. The host CPU OS typically is a variant of Linux and is where the control plane of the AI application is run. The primary software focus of this work is the management plane for OPG-M and XOC-N instances.

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

5. System Architectures

The System Architectures developed as part of this Initiative will contain sufficient detail to meet the following objectives:

- Enable detailed AI network planning
  - Processing capacity
  - Networking requirements
  - Network topologies
  - Floor space requirements
  - Power requirements
  - Cooling requirements
- Deliver clear plans for constructing complete OPG-M and XOC-N systems
  - Numbers, types, and configurations of racks
  - Numbers, types, and configurations of major network elements (pods/xpu-nodes, switches, etc.)
  - Numbers and types of auxiliary equipment and parts (power supplies / distribution units, optical modules, cabling, cable-management, etc.)
  - Network topologies and connectivity maps
  - Suitable for equipment and systems manufacturers, systems integrators, and value-added resellers
- Deliver guidance for implementing management-plane solutions at the OPG-M and XOC-N levels
  - Numbers, types, and configurations of major network elements (pods/xPU-nodes, switches, etc.)
  - Compute requirements for software management stack
  - Network topologies and connectivity maps

To facilitate extensibility and ease of modification, several areas will be emphasized in the System Architectures:

- **Modularity:** In OPG designs, multiple instances of a single design (e.g., pod) will be used whenever possible. In XOC-n designs, multiple instances of OPGs will be used, and wherever possible at the spine and aggregation layers, multiple instances of single designs (e.g., switching rack) will be used.
- **Full-bandwidth switching designs:** For all networks covered by the System Architectures, full-bandwidth, non-blocking designs will be developed. This approach should allow a user to create an oversubscribed design by simply removing links from the base design.

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

- **Network topologies:** The System Architectures will contain explicit descriptions of the cabling topologies (Clos, rail-optimized, etc.) of the distinct networks contained by an XOC-N. These networks are designed for specific purposes and typically vary widely in scale and complexity. Figure 4 shows conceptual illustrations of the scale-out (Clos topology) and storage networks for an XOC-1k.

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

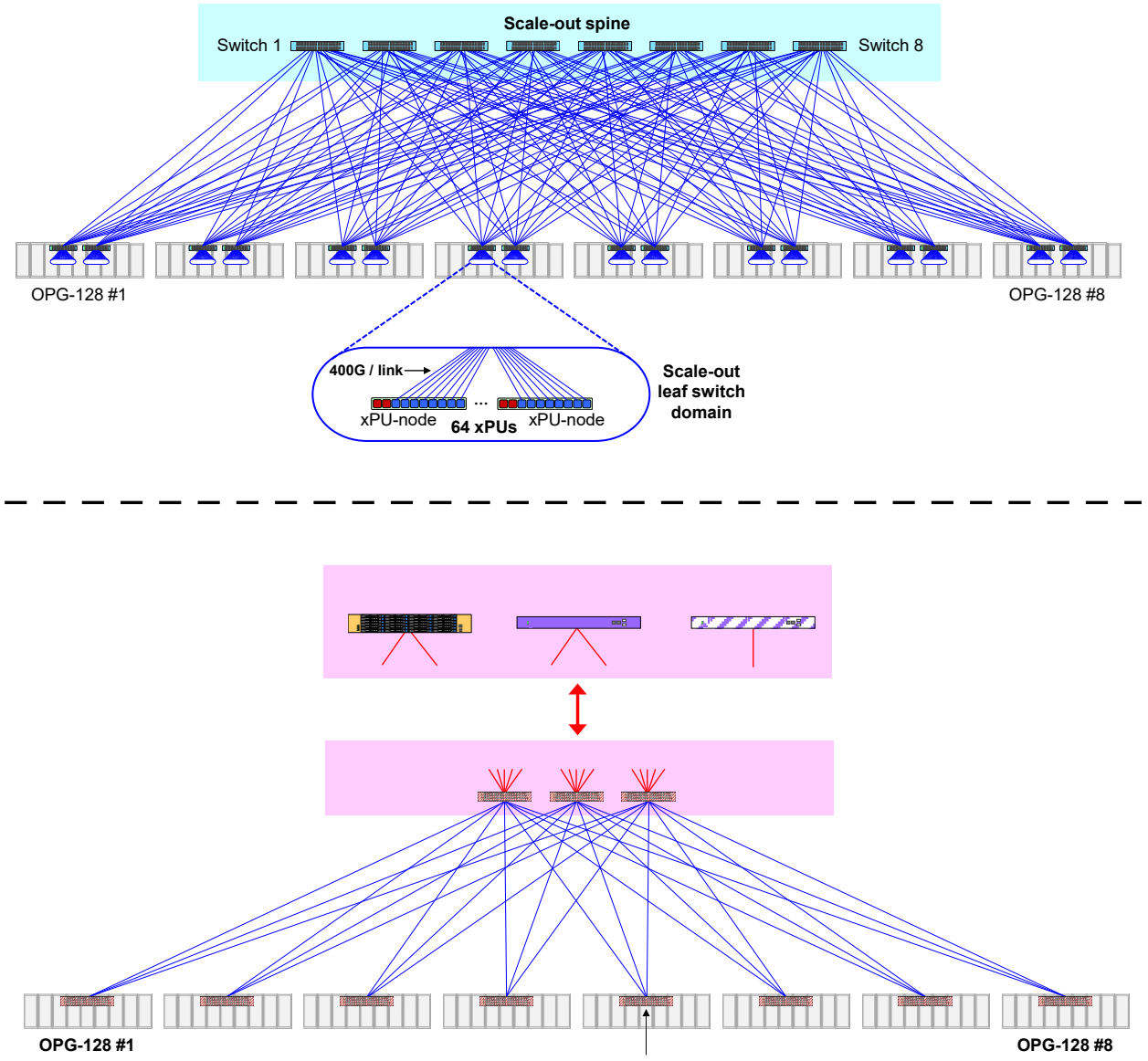

Figure 4. Conceptual illustration of the scale-out (top) and storage (bottom) networks for an XOC-1k built using eight OPG-128s.

Date: January 2026

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

Connectivity maps: The development of complete connectivity maps is of fundamental importance for both OPG-M and XOC-N designs. While many equivalent solutions are, of course, possible, selecting one mapping and documenting it fully provides an unambiguous path to a working solution for users who are building systems. Further, a complete connectivity map enables the assignment of logical connections to physical connections, which is needed for the construction of management applications. As example, Figures 5 and 6 illustrate the structure of connectivity maps for the scale-out networks of an OPG-M and an XOC-N, respectively.

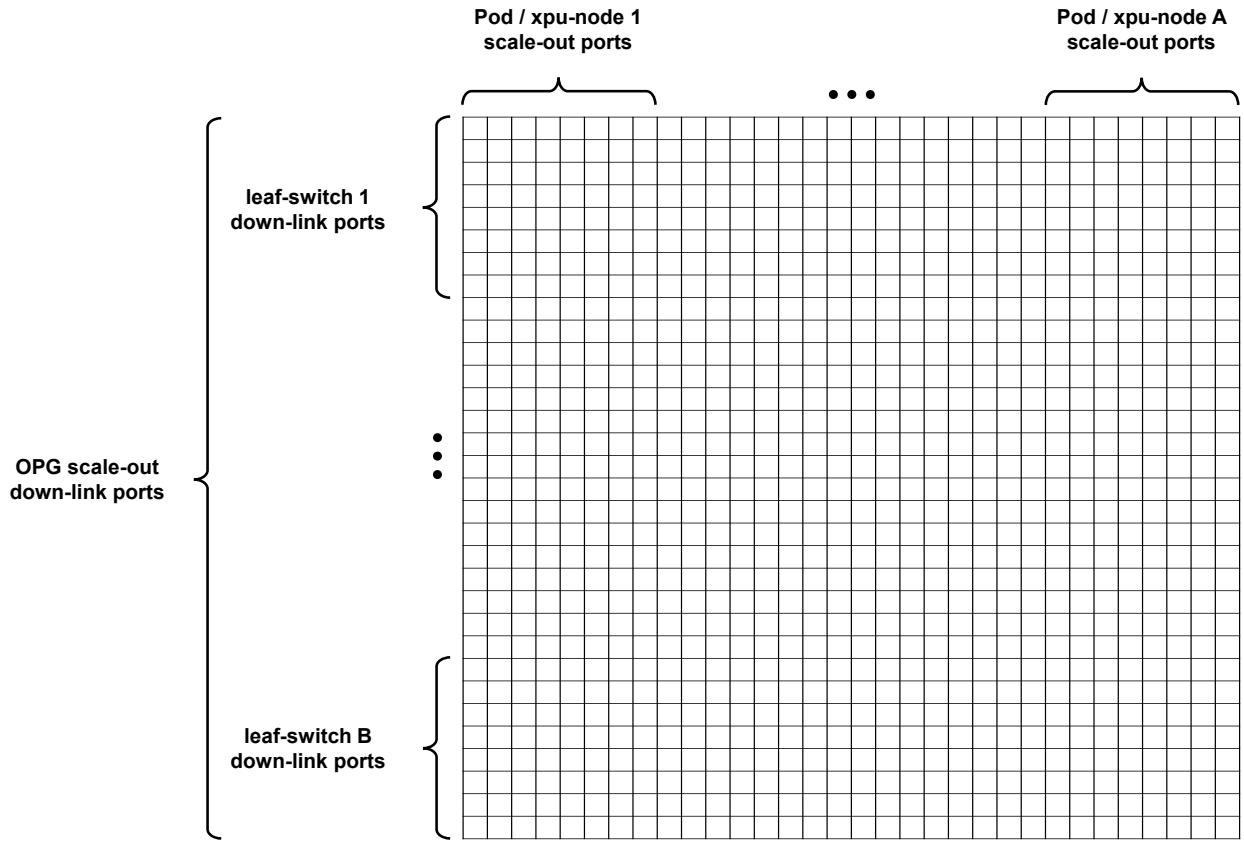

Figure 5. OPG connectivity map for the scale-out network.

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

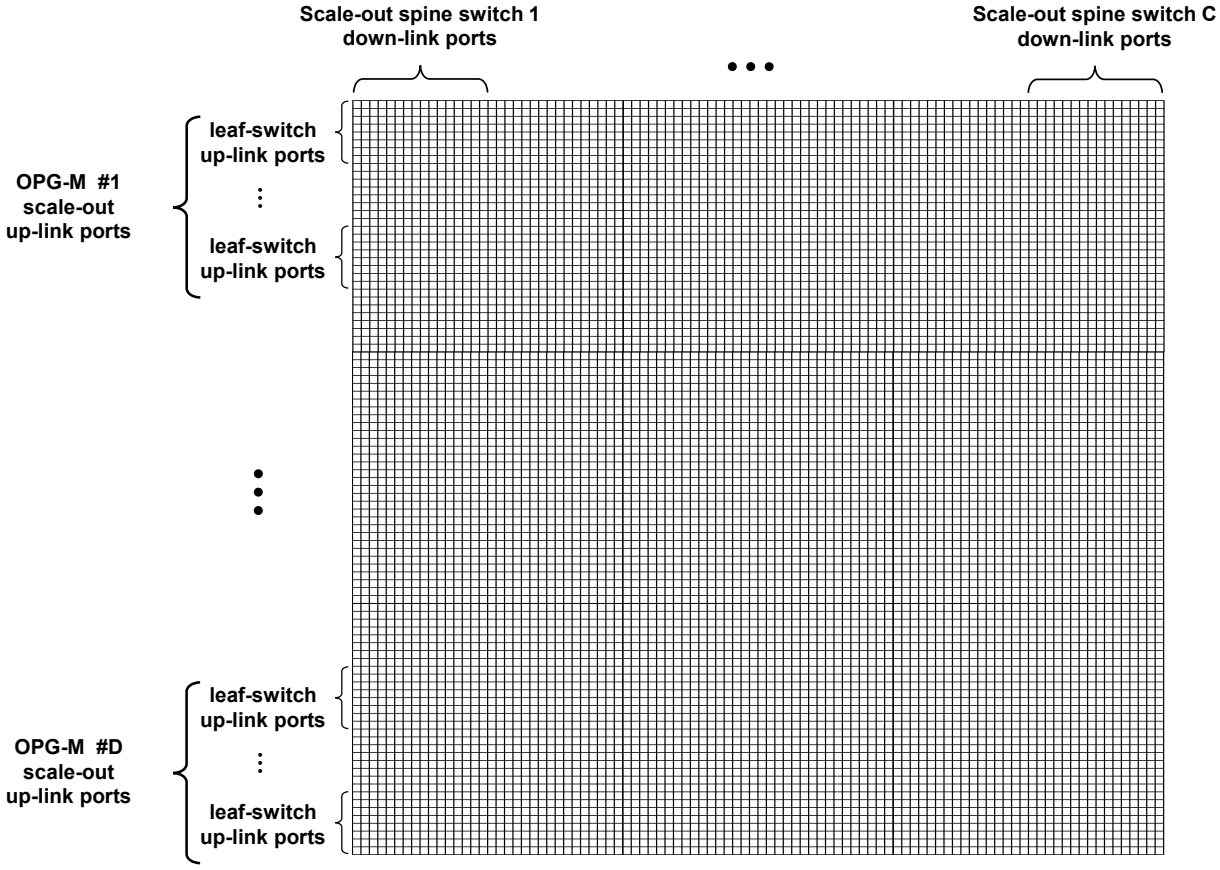

Figure 6. XOC-N connectivity map for the scale-out network.

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

6. Software components of a System Architecture

The primary software for this Initiative is in the management plane for OPG-M and XOC-N instances. For this aspect, Open Compute has several projects that create open-source elements for system management, primarily around BaseBoard Management Hardware (BMC) and Redfish profiles for monitoring/control of the products needed to create OPG-M and XOC-N instances. See Table 1 for specifics.

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, TBD
BMC present	Baseboard Management Controller	OpenBMC
RMC present	Rack Manager Controller	OpenRMC
UBB BMC present	UBB/Accelerator Controller	*xPU-aware OpenBMC variant

*May only be available from the xPU vendor

Table 1. Example open-source software elements for OPG-M and XOC-N System Architectures.

It is likely that as we review the available Open SW and Redfish profiles, we may find the need for additional functionality to match the size/scope of OPG-M pods and XOC-N clusters. This may require work in these already established projects at Open Compute.

Date: January 2026

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

7. References

[1] https://www.opencompute.org/projects/open-systems-for-ai

[2] https://sonicfoundation.dev/

[3] https://www.opencompute.org/documents/ocp-open-systems-for-ai-whitepaper-v1-0-0-final-pdf

[4] https://www.opencompute.org/wiki/Open_Rack/SpecsAndDesigns

[5] https://www.opencompute.org/products

8. Glossary

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.

Leaf Switch: In a multi-tiered switching network, leaf switches occupy the lowest switching tier and connect to the end points.

NIC: 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.

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.

Date: January 2026

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

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 Network: 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.

Scale-up Network: This network connects a relatively small number of xPUs 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.

SONiC: An open-source network operating system (NOS) based on Linux.

Spine Switch: In a multi-tiered switching network, spine switches occupy the highest switching tier and cross-connect the leaf switches.

Storage Network: The network responsible for data flow between the storage system and the pods/xPU-nodes.

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.

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 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.

9. License

9.1. 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.

10. 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.