Scaling AI Clusters at Neoclouds

FTI home: opencompute.org/projects/future-technologies-initiative

FTI Neocloud Workstream: opencompute.org/projects/scaling-ai-clusters-at-neoclouds

Workstream Co-leads

Scaleway, FarmGPU, and Denvr Dataworks are founding leads for the workstream. Other neocloud companies have expressed interest in joining ASAP. Key co-lead contacts are:

- Kyle McCrindle from Denvr Dataworks
- JM Hands from FarmGPU
- Yann-Guirec Manac'h from Scaleway

Vision

To pioneer an open, collaborative ecosystem where hyperscale AI innovations are democratized, empowering neoclouds of all sizes—from expansive global AI factories to agile small-scale deployments—to scale efficiently, sustainably, and inclusively, fostering a future where advanced AI infrastructure is accessible to innovators worldwide.

Mission

The OCP Workstream "Scaling AI Clusters at Neoclouds" is dedicated to uniting industry leaders, researchers, and practitioners in an open forum to address the unique challenges of scaling AI infrastructure for neoclouds. By sharing designs, best practices, and standards, we accelerate the identification of common challenges and work on development of modular, energy-efficient, and cost-effective solutions that enable seamless growth from small-scale AI deployments to massive hyperscale operations, ensuring that "Community-driven hyperscale innovation for all," the OCP slogan, becomes a reality in the AI era.

Goals

Foster Knowledge Sharing and Best Practices: This workstream represented a requested forum for neoclouds (and open to all others). It's a place to discuss (meetings online and/or at Summits) and document OCP adoption and contribution as well as other opportunities for collaboration or simply coordinating across the OCP Community.

Promote Interoperability Across Scales: Drive initiatives to ensure seamless integration between small-scale edge AI deployments and large hyperscale-like/centralized clouds, enabling hybrid neocloud models that support AI training and inference, thus adding value by bridging the gap between localized innovation and global-scale operations.

Enhance Energy Efficiency and Share Strategies for Sourcing: Collaborate on designs that optimize power consumption, cooling, and resource utilization in AI factories, aiming to boost the efficiency of neoclouds while supporting high-performance computing needs, making sustainable AI scaling viable for organizations of any size and as they grow. Also we will share effective strategies around the globe for sourcing energy and facilities requirements of AI factories be they first-party or co-lo operated.

Develop Open Standards/Specifications for AI Infrastructure Modularity: Create and promote standardized hardware and software interfaces and management that allow neocloud operators to mix and match components from multiple vendors, reducing vendor lock-in and enabling rapid scaling without proprietary constraints, ultimately lowering costs and increasing flexibility for both large and small deployments. While this work (specs and such contributions) happens already in the OCP "Open Systems for AI" theme initiative and various projects, this workstream is particularly looking to identify/document gaps and opportunities for neoclouds that may be missed in the wider community or of particular use as useful patterns to neoclouds.

Scope

This workstream focuses on collaborative efforts to address scaling challenges in AI data center infrastructure for neoclouds, encompassing large AI factories and small-scale deployments. If you're new to the term neocloud, check out this SemiAnalysis article that defines and details it. Ultimately, this OCP workstream welcomes everyone, but neoclouds are companies that offer xPU infrastructure/platforms as a service, from the small and regional to emerging hyperscalers.

While the workstream is new there is a panoply of directions possible. As a group, we have not set a course for specific work products yet, but aim to narrow down the focus in 2026 in this workstream, or do so through collaborations with narrower focused OCP Projects. While an OCP FTI workstream does not traditionally have objectives of making specifications (OCP Projects do), this workstream seeks to coordinate work that addresses the unique challenges of neoclouds, and then work through the OCP Projects and Strategic Initiatives, particularly Open Systems for AI and Open Cluster Designs for AI.

Challenges unique to neocloud AI deployments include some of these (along with noted areas of the OCP Community where we can coordinate impact):

- **Getting facilities**

    - Traditional DC builders and co-location facilities are often unsuitable for modern AI deployments, notably lacking the prerequisite power and cooling distribution systems
    - OCP work areas: OCP Ready™ for AI, OCP DCF (including TCS and pending DC Power Sub-projects)
- **Getting electrons**
    - While neoclouds often add value in regional deployments, they can be limited by the availability of power. There are cost, reliability, and availability concerns of the grid or co-generation strategies to consider here upstream of developing a data center (and selecting one). There are fast evolving changes within the data center to go from the medium voltage (AC) grid to a high voltage DC power distribution system for the next generation of AI IT racks/pods.
    - OCP work areas: Sustainability, Rack & Power, adapting the Diablo power sidecar, DCF and DC Power
- **Getting hardware** (fast enough)
    - The challenge of getting modern IT hardware/accelerators and upcycling some recent generation's hardware is not a traditional project area that the OCP addresses, but neoclouds may find value in sharing strategy on dealing with this at the business and technology level. Collaborating on what hardware to invest in when in the growth cycle is another related matter of discussion. Finally, neoclouds require high levels of automation and management, and often face the same challenges around management interfaces.
- **Getting customers**
    - While this is not something that the OCP Projects help with, the OCP Foundation itself encourages neoclouds to explore OCP Experience Centers and the OCP Marketplace to list their offerings. There are many neocloud-themed talks and panels at OCP Summits. We would like that to continue. At OCP Global Summit 2025, James Kelly will be announcing this workstream on the keynote stage by way of introducing Denvr Dataworks and Scaleway to split a keynote slot of 20 minutes. There are also neocloud panels and breakout presentations.

Work Products

- TBD in 2026
    - The workstream will kick off at OCP Global Summit 2025 on 10/14, and we'll host a meeting there for anyone interested to join by invitation as space is limited. We expect companies like Nebius, Lambda, Crusoe, Coreweave, Core42, Groq and others who are OCP regulars to be invited.
- Repeated from above:
    - As a group, we have not set a course for specific work products yet, but aim to narrow down the focus in 2026 in this workstream, or do so through

collaborations with narrower focused OCP Projects. While an OCP FTI workstream does not traditionally have objectives of making specifications (OCP Projects do), this workstream seeks to coordinate work that addresses the unique challenges of neoclouds, and then work through the OCP Projects and Strategic Initiatives, particularly Open Systems for AI and Open Cluster Designs for AI.