MOC-A: driving innovation and advancing AI

Orran Krieger & Heidi Dempsey
Red Hat Research
October 2024

Executive summary

Red Hat has for many years participated in and supported the Mass Open Cloud Alliance (MOC-A). With the rising importance of AI, the MOC-A now provides Red Hat, our partners, and collaborators a platform to improve open source AI products, increase mindshare for open source AI technologies, and form catalyzing partnerships around AI solutions and services. While the MOC-A has always been an important part of our work, Red Hat Research is now focused on further developing this opportunity, expanding the MOC-A ecosystem, and helping our AI business units and partners take full advantage of it. **At this pivotal moment in the evolution of AI, we contend that this long-running project is now a uniquely powerful platform for driving innovation and advancing open source AI.**

We prepared this document to help potential members of the MOC-A ecosystem identify the value the MOC-A can bring to them and encourage them to become MOC-A users, advocates, contributors, and/or collaborators. We provide background on the MOC-A and discuss why we believe it is now strategic for Red Hat and many of the technology companies Red Hat collaborates with. We then outline near term challenges the MOC-A faces and how different stakeholders can participate in the effort, including hardware manufacturers, software vendors, cloud providers, and industry research groups, as well as universities and developers. We identify several opportunities for stakeholders to participate and realize value by engaging with the MOC-A. The door is open for those stakeholders to bring their own needs and goals to the table as well.

Context

The Mass Open Cloud (MOC) Alliance is a long-standing collaboration between academic institutions, government, and industry that has created an open cloud. It was designed to provide researchers and students access to the large-scale compute resources, large diverse data sets, AI tools, and AI models that are critical to addressing problems in healthcare, climate change, education, and many other global challenges.

The demands of AI have made the cloud critical for research and education; however, today's public clouds have several disadvantages for academic users:

- Cloud services lock users into a specific vendor, with only a subset of the tools and functionality that open collaboration can provide.
- Commercial clouds lack the human facilitators common to successful HPC services that allow domain specialists to use the cloud for complex projects without getting bogged down in the technology.
- Commercial clouds impose unpredictably large costs that are enormously (e.g. 4x) higher than the costs of campus compute resources that many top academic institutions make available to their faculty.

The situation is even worse for systems researchers who want to innovate and improve the way cloud and AI platforms are implemented. Without access to the scale, demands, and users of a real cloud, and without bare metal access to diverse hardware, many areas of system research are impossible.

The MOC-A addresses these needs by providing an open cloud for research and education whose fundamental goal is to *maximize impact rather than revenue*. It offers users facilitator-supported services at a fraction of the cost of the public cloud, and it offers system researchers, open source developers, research IT, and industry a shared large-scale environment to rapidly advance AI tools and infrastructure.

MOC-A resources

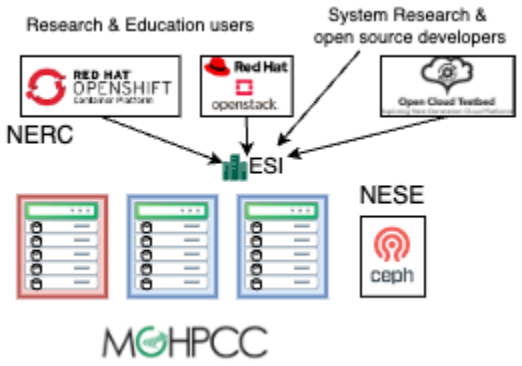
The MOC-A production cloud service used by most end users is called the New England Research Cloud (NERC). NERC today supports VMs and containers (OpenShift), volume and object storage, and an AI-as-a-Service platform (OpenShift AI). The MOC-A charges rates that are sufficient to cover its operational and development costs, enabling shared services, like NERC, that can take advantage of the economies of scale of a wide range of institutions.[1] At its current scale, rates are around ⅓ public cloud rates, and there are no charges for network egress. (These rates will drop as the MOC-A/NERC continues to scale.)

NERC and other MOC services and projects have access, through Elastic Shared Infrastructure (ESI), to around 1,000 servers that include over 30,000 cores, 5 terabytes of memory, and nearly 100 various types of GPUs. There are 64 Nvidia A100 GPUs currently available, and installation of an additional 192 H100 GPUs in progress. To support diverse hardware options, accelerators from AMD and Intel are also planned.

System researchers and open source developers that require bare metal access either use ESI or the **Open Cloud Testbed** (OCT), an NSF-funded national testbed service that is layered on ESI. These services provide access to the lowest layers of software and hardware systems,

[1] In contrast, most top universities that operate computer clusters today make them available at highly subsidized rates only to their own faculty.

where innovation is not possible in public clouds. Innovators who see a need for new services or new functionalities in distributed infrastructure can build them on diverse hardware, and then transition new ideas to the production services in the NERC when they mature.



Through the North East Storage Exchange (NESE), MOC-A projects have access to over 50 petabytes of disk and 100 petabytes of tape storage. Harvard's IQSS is developing a modified instance of Dataverse in the MOC-A to support hosting of large data sets from scholars across the world using NESE storage and to enable AI tools running on NERC to access those data sets in situ.

All infrastructure is currently housed in the Massachusetts Green High Performance Computing Center (MGHPCC), a 90,000-square-foot, *carbon-neutral* datacenter that is owned and operated by a consortium of northeastern academic institutions. As demand grows, the MOC-A plans to expand to additional academic data centers.[2] Day-to-day decisions are made by a small team of dedicated staff, and major decisions are reviewed by a board with representatives from the consortium of partner institutions that have made commitments to the effort.

[2] The aim is to expand, rather than create federated replicas in order to: 1) enable the operational efficiencies of a larger group of SREs operating distributed services, 2) create a single cloud with the economic advantages of scale and geographical distribution, 3) simplify the user experience, and, 4) enable a single professional organization rather than a multitude of separate collaborating organizations.

Why the MOC is strategically important for AI

There are three reasons why we believe that the MOC-A is now strategic for Red Hat Research, key Red Hat products (OpenShift, OpenShift AI, RHEL AI), and many Red Hat partners/collaborators:

Academic research plays a critical role in AI innovation (and the resulting startups).

AI innovations happen in academia and are rapidly moving into the upstream communities that Red Hat needs to support. The MOC-A provides Red Hat and other participants the opportunity to make their products available for academic research and education on AI at a fraction of the cost of the public cloud, enabling this innovation, and allowing the resulting startups to grow as part of the MOC's open ecosystem, rather than in the public cloud. Imagine if companies like Domino Data Lab and Data Bricks had grown up on an open ecosystem rather than on AWS. Imagine if every new graduate that enters the AI-related workforce was trained on open source technologies.

The ability to innovate at all levels enables services to be optimized for specific educational and research needs. As one example of the impact of this approach, the MOC-A enabled BU to reduce the average cost per student for recent classes using OpenShift AI from around $140 on the public cloud to around $18 on NERC, while also supporting new teaching methods and dynamic digital textbooks.

Open Source software must evolve more rapidly to adjust to the pace of change in AI.

Development for OpenShift AI has already benefited from early deployment feedback from the MOC. The MOC provides an environment where we identify problems while working with real users, make changes together with partners, perform A/B testing to determine if we have fixed the problems, and then integrate changes into upstream software and into products. Telemetry that the MOC-A makes available to collaborators identifies performance problems that need to be addressed, and the institutional facilitators working with MOC-A end users identify usability issues.

It is critical that high quality open source models and tools dominate AI.

It is critical to society in general, and Red Hat in particular, that high quality open source models and tools dominate AI, and that federal agencies are not restricted from using and funding them. Consider the large investment that IBM, Red Hat, Meta, and others have made in the AI Alliance. The MOC-A provides a natural place to host researchers and nonprofits that are part of this open movement, and its position within academia is already providing a structure to galvanize major academic institutions across the US to support open source for AI.

How the MOC-A will scale

The MOC-A is today being used by over 200 Principal Investigators (with thousands of users from their labs and courses) from 10 academic institutions, a medical research group (ChRIS), and several public good organizations and projects, such the Open Education project, which creates free interactive open-source textbooks; Code for Boston, which assembles volunteers to solve civic and social problems; and the Southern Coalition for Social Justice project.

The MOC-A has developed the business and governance models necessary for shared services that can be used widely and governed by a growing consortium of academic institutions. The MOC-A is well positioned to take advantage of dramatic increases in government funding for AI to scale. In the US, new programs include:

- The National AI Research Resource (NAIRR) with pending legislation to fund $2B for academic research
- New York's empire state consortium ($400M)
- The Commonwealth of Massachusetts AI task force ($100M+)

Similarly, in Europe, the Horizon Europe program recently announced a €112m call for pioneering projects in AI.

The MOC-A is uniquely attractive for government initiatives to democratize access to AI with:

- A governance model crafted by non-profit academic institutions
- A proven track record of enabling broad innovation and industry participation
- A cost-effective reliable production cloud for AI that is in a position to grow

As a result, more than ten universities have signed MOUs to use the MOC, the University of Arizona and UMass have just joined the MOC-A as full partners, and we are in discussion with Arizona about expanding the MOC-A to their data center. We are also in conversation with a number of other US institutions, including Yale, CMU, and Berkeley. In Europe, we are in early discussion with ICTP, EPFL, ETH, and the Walton Institute.

Each of these possibilities for increased funding and expanded participation represents the significant opportunity Red Hat and the MOC-A now have to increase the awareness and impact of open source AI efforts.

How the MOC-A provides value

The MOC-A provides value to Red Hat as well as collaborators and partners in a number of ways, and we expect this value to increase tremendously as the MOC-A scales.

The MOC-A can be used to host open source development.

Hosting open source development has a number of important advantages: 1) it provides an open testbed for product development; 2) it provides access to GPUs and services for many use cases at a fraction of the price of the public cloud or internal deployments; and, 3) it enables scale- and security-related experiments with open source products that would be prohibitively expensive otherwise.

The MOC-A provides a shared environment for collaborations to advance AI.

For hardware partnerships, the MOC-A has already attracted investment from AMD, NVIDIA, Lenovo, Intel, and others; it provides Red Hat and others a new vehicle to work together in a public environment. Many software partners can make use of the MOC-A, both for delivering products to research and educational customers, and more broadly to university-based startups. For the many companies that want to offer free or inexpensive versions of their products to universities, the low cost of the MOC-A provides a strong incentive to support and integrate with Red Hat's open source platforms that power the MOC-A. Most importantly, in the rapidly changing world of AI, the MOC-A enables new partnerships, providing a shared environment that greatly reduces the barriers to technology evaluation and collaboration.

Research resulting from the Red Hat-MOC-A partnership has already impacted a wide range of open source projects and Red Hat products. To support this research, Red Hat established the $20M **Red Hat Collaboratory** that connects Red Hat engineers and researchers in joint innovation efforts across many technologies. In 2024, NSF funded the **Center for Systems Innovation at Scale** (i-Scale), an NSF Industry University Cooperative Research Center to enable more companies to be involved in the research innovation around the MOC-A. These projects are open to interested partners and collaborators.

Many service providers and large enterprises are starting to face the same challenges we are addressing in the MOC-A.

For example, a number of major Red Hat customers are interested in AI cloud-in-a-box solutions whose prescriptive nature greatly simplifies automation and enables increased efficiency. As another example, large customers and regional service providers require services that can span data centers and burst to the public cloud. The MOC-A has developed, or will need to develop, ways to address these challenges. The MOC-A provides an attractive environment for collaborators and partners to develop, test, and demonstrate potential solutions together.

Opportunities and next steps

The consortium of academic and industry partners have already made the investments needed to create a successful open product cloud. It has taken years to develop the governance model, structures for billing, trust relationships with top universities, open source research enablement and ecosystems to support teaching and course development. MOC-A production services now run reliably (99.4% uptime including scheduled upgrades) and are used regularly by a growing community of users in the US and Canada. In other words, the MOC-A is on a path to organically grow in a self-sufficient fashion to support research and education users while creating a platform for system researchers and industry to collaborate on innovation.

There are of course many challenges that we would love to collaborate on with partners to enhance the MOC-A. Key technical priorities to enable the MOC-A to leverage expected state and federal investments to grow rapidly are:

- **Supporting scalable multi-cluster operations:** We need to non-disruptively upgrade any one of several production clusters rapidly, be able to easily roll back upgrades if needed, and provide models for users to exploit multiple clusters.
- **Improving observability:** We need to quickly identify performance problems and software/hardware failures; provide showback to users, partners, and researchers and provide relevant data to drive automation. In particular, we must provide insights into the types of data needed to support analysis of AI systems and software with much more fine-grained observability and better mechanisms for controlling access and retaining data than those currently available in traditional open source operations software.
- **Improving AI and domain-expert developer experience:** Applications drive AI. Innovations in applications for many domain sciences, such as medical, environmental, and smart cities software, are already happening on the MOC. However, the ramp-up for going from an AI application idea to implementation in the full stack of AI and hybrid cloud infrastructure takes too long and is far too complex. We need to provide help in the form of automation, reproducible AI templates, and other developer-friendly tools.

In addition, this environment will require (and enable) technical and business innovation. For example, to fully enable the potential ecosystem, we need to create a marketplace where a diverse community of developers can offer solutions that can be easily adopted by a diverse community of users. At a low level, we need to enable GPU Direct RDMA for storage and networking and provide support mechanisms to cache and prefetch data to maximize use of expensive GPUs. Organizationally, we will need to greatly expand and invest in the SRE community and tools operating the cloud and start supporting compliance regimes like HIPAA required by many users.

We are seeking contributors and collaborators to broaden and maximize the innovative value of the MOC-A ecosystem. This includes end users like researchers as well as technology vendors, service providers, and independent research groups in multiple industry verticals. Strategic participation from these organizations will not only expedite the growth and adoption of open

source AI technologies, but will also give them the benefit of a platform to publicly demonstrate their technologies at scale with real users, exploit MOC-A resources like telemetry and facilitators to evolve products and services based on actual demand, and engage with systems researchers from participating universities.

Consider this non-exhaustive list of possibilities:

- **Original equipment manufacturers (OEMs)** can partner to expose new technology to a wide community of users. For example, Lenovo partnered with the MOC-A with 64 NVIDIA A100 GPUs under a new business model of GPU core-hour lease, and is now expanding this partnership with 192 H100 GPUs.

- **Independent software vendors (ISVs)** can offer software on NERC for education and research users at a fraction of the cost of the public cloud, and build relationships with universities and research IT to drive adoption.

- **Hyperscalers** can engage broader communities on challenges of scale, accelerating problem solving by exposing issues to academia and other collaborators in a shared environment.

- **Research and engineering groups** in verticals such as financial services and healthcare can test and demonstrate solutions for industry-specific requirements such as security and data privacy in a public testbed.

- **Nonprofits** including academic institutions and funding agencies can use the service, make it available to their community, and participate institutionally in the governance.

Institutions that want to use and/or become members of the MOC-A should contact Nancy Clinton, Managing Director of the MOC (nclinton@bu.edu). Companies interested in becoming an industry partner of the MOC and/or the i-Scale research center can also reach out to Nancy and Jon Stumpf, Strategic Partnerships MOC-A (jstumpf@bu.edu). If you are interested in partnering with Red Hat around the MOC-A, or would like more information, please contact Heidi Dempsey (Heidi Dempsey) or Orran Krieger (okrieg@bu.edu or Orran Krieger).

Appendix: Partners and responsibilities

The MOC-A is a collaboration, with industry, academic institutions, research IT groups, open source communities all playing a role. The teams/groups and their responsibilities in the MOC-A are:

- MOC Leadership: from institutions and core industry partners
    - Set priorities, partner engagement, acquisitions, help drive institutional engagement, MOU, governance, buy-in model
- Institutional Research IT from BU, Harvard, UMass, Arizona:
    - Operations of production services
    - Acceptance of new features into production
    - First level facilitation, help desk
- MGHPCC, Institutional partners, MOC engineers and administrators:
    - Billing, engineering unique to MOC (billing, showback, onboarding)
    - MOC project management
- Industry partnerships: Red Hat (Research, ET, OSPO), Lenovo, IBM (Research, Storage), Dell, AMD, Intel, G-Research, Two Sigma:
    - Operate services for research and education on top of the base services provided by the MOC.
    - Offer infrastructure into the MOC-A through the on-demand model pioneered by Lenovo.
    - Collaborate on new MOC development and research projects.
    - Demonstrate and integrate innovation in MOC services and help operate until they are accepted into production MOC systems.
- Red Hat Research: Beyond participating as a partner and engaging in the innovation:
    - Lead efforts to address important gaps with experienced staff and provide junior staff to assist with operations, facilitation and development.
    - Develop processes to engage Red Hat business units, escalate problems, evaluate innovation, and ensure that innovation flows into Red Hat products
    - Work with MOC leadership to engage partners and Red Hat business units in joint projects and innovation.