RH RQ

Bringing great research ideas into open source communities

Stefanie Chiras interviews

Chris Sedore

Boston University's CIO and Red Hat's AI Innovation Hub leader discuss plans to rev up the AI opportunity engine

+

Better distributed data processing for research

Scheduling resources across a multicluster environment

AIBOM: unpacking the black box

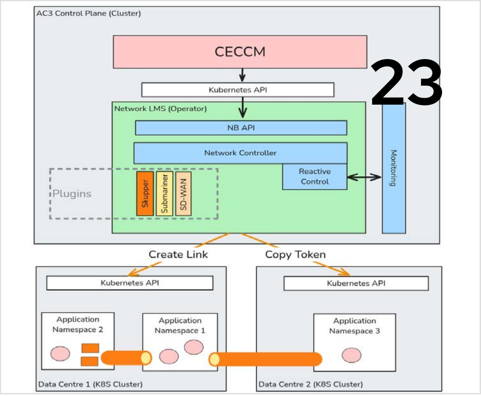
Red Hat
Research Quarterly
Volume 7:3 | Winter 2025-26 | ISSN 2691-5278

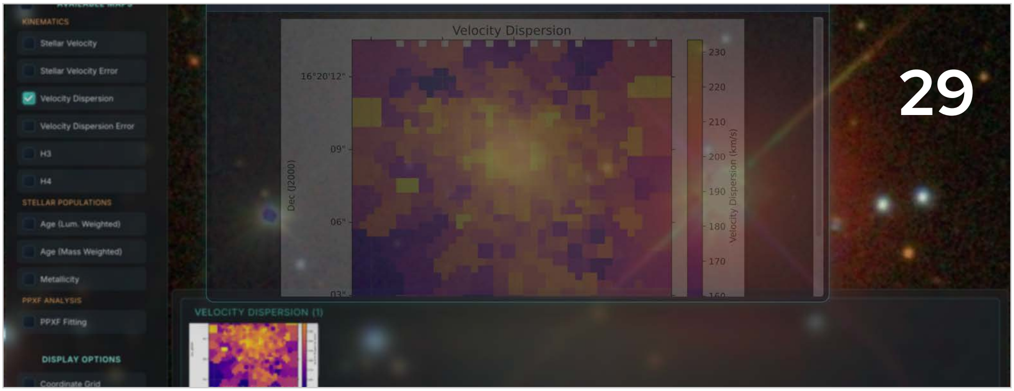
intel AI
AI ON INTEL®
intel XEON PLATINUM inside™
NOW BUILD THE AI YOU WANT
ON THE CPU YOU KNOW.
Learn more at ai.intel.com

RESEARCH
QUARTERLY

VOLUME 7:3

Table of Contents





Departments

04 From the director: Having our cake and eating it too: HPC meets enterprise AI

36 Why open source is integral to US AI research infrastructure

Features

09 "We've got to have everyone": combining research innovation with enterprise operations—an interview with Chris Sedore and Stefanie Chiras

19 Unpacking AI's black box: why authenticity and traceability must be built in

23 Building an intelligent multicluster scheduler with network link abilities

29 Concurrent, scalable, and distributed astronomy processing in the AC3 framework

ABOUT RED HAT Red Hat is the world's leading provider of open source software solutions, using a community-powered approach to provide reliable and high-performing cloud, Linux®, middleware, storage, and virtualization technologies. Red Hat also offers award-winning support, training, and consulting services. As a connective hub in a global network of enterprises, partners, and open source communities, Red Hat helps create relevant, innovative technologies that liberate resources for growth and prepare customers for the future of IT.

NORTH AMERICA
1 888 REDHAT1

EUROPE, MIDDLE EAST,
AND AFRICA
00800 7334 2835
europe@redhat.com

ASIA PACIFIC
+65 6490 4200
apac@redhat.com

LATIN AMERICA
+54 11 4329 7300
info-latam@redhat.com

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

Red Hat

RESEARCH QUARTERLY

VOLUME 7:3

From the director

About the Author
Orran Krieger is the Director of Red Hat Research while on leave of absence from Boston University, where he is a professor in the Department of Electrical and Computer Engineering. He is a founding lead of the Mass Open Cloud Alliance (MOC-A).

Having our cake and eating it too: HPC meets enterprise AI

The convergence of high-performance research computing and general-purpose IT is turning conventional wisdom on its head.

by Orran Krieger

High-Performance Computing (HPC) and general-purpose IT infrastructure have always been very different. This difference is very visible at top universities, which maintain both types of environments: HPC clusters dedicated to research, often managed by a small team of two or three staff overseeing thousands of machines, and enterprise IT environments, where perhaps 50 operations staff manage hundreds of computers running a wide array of diverse, mission-critical workloads.

The differences go far beyond operational complexity. HPC environments are generally dedicated to a smaller number of large-scale workloads—say, computation for particle accelerator experiments—whose results will be widely shared, while enterprise environments have strong compliance and security requirements around many users. HPC applications may access massive datasets, but each application is specialized for data from a specific domain. General-purpose environments support many different types of users and applications, often combining data from diverse sources ranging from file systems, distributed databases, data warehouses and streaming live data from sensors.

HPC applications rely on a limited set of libraries that remain fairly stable, while general-purpose environments get frequent updates, from security patches to constantly evolving services and libraries. HPC jobs are batch scheduled (e.g., managed using SLURM), where all the computers in the cluster may work on the same problem simultaneously, while general-purpose environments typically support loosely coupled independent tasks and interactive workloads orchestrated using platforms like Kubernetes.

HPC generally uses free software and on-premises experts for support, while general-purpose IT pays licensing fees and invests in multitiered processes with vendor partners to ensure 24x7 support and escalation. Finally, to handle computation failures, HPC work has focused on checkpointing, while general-purpose computing requires application state to be tracked and handled independently from compute node state in a cluster (see Pets vs. Cattle).

The Mass Open Cloud (MOC), inspired by Project Kittyhawk (Jonathan Appavoo et al.), has, since its inception, been based on the radical hypothesis that these two worlds could converge into a common platform where both HPC and cloud computing benefit from a set of common services, business models, operational disciplines, and capabilities. The wide-scale adoption of AI is increasingly making this radical hypothesis established practice. For example:

- Neoclouds like CoreWeave, Nebius AI, Lambda, Crusoe, and Voltage Park support both SLURM batch scheduling and Kubernetes dynamic container clusters.
- The Department of Energy's Genesis Mission is applying National Lab HPC supercomputers to AI challenges.
- Data at massive scale is needed to train and tune successful AI models, and much of that data requires security and data ownership protection different from traditional academic HPC.
- The rapid pace of change in AI, and its broad use, means that systems are changing at a pace HPC systems never experienced before.
- Hardware and software that was developed and improved for general-purpose enterprise use, with strict compliance and security requirements, is increasingly critical for real-life AI applications.

The implications of this convergence, especially in an environment for academic research, are profound. AI startups are popping up from research universities faster and faster, driving the bleeding edge of new platforms. Industry now tracks what researchers are doing closely, and many AI innovations move from universities into general usage in weeks. Research universities are adopting AI long before enterprise customers, which means that the workloads of universities, and the systems needed to support them, are important areas of interest for collaborative industry and computer systems research.

Research universities are adopting AI long before enterprise customers, which means their workloads, and the systems supporting them, are important areas for collaborative industry and computer systems research.

In February 2024, Governor Maura Healey formed the Massachusetts AI Strategic Task Force, which recommended the establishment of the Massachusetts AI Hub to serve as a nexus for AI innovation and facilitate cutting-edge collaboration between government, industry, academia, and nonprofits. Key interrelated initiatives of the AI Hub to unlock innovation in the commonwealth include creating the AI Compute Resource (AICR) infrastructure to supply compute capacity, a Data Commons to unlock the value of shared, high-quality, and responsibly governed data across various sectors, and programs to support and accelerate AI startups.

In this issue, we feature a conversation between Stefanie Chiras, Senior Vice President of the AI Innovation Hub at Red Hat, and Chris Sedore, Vice President and CIO at Boston University—two leaders helping shape the Mass Open Cloud (MOC) around AI as a catalyst for research, innovation, and economic growth.

The AI-driven convergence of HPC and general purpose computing positions the MOC to play an important role in supporting initiatives like the Massachusetts AI Hub. The MOC is evolving its cloud-native services to support research, education, and startup communities with the security, compliance, and operational rigor modern AI workloads demand.

Under Chris's leadership, early next year Boston University's enterprise IT organization will assume operational responsibility for these services, working closely with Red Hat to rapidly enable the compliance and security regimes required for AI use cases, particularly in health care and data-intensive domains. Stefanie, who leads the Red Hat partnership with the AI Hub, is engaging AI startups to take advantage of this infrastructure. This effort is closely aligned with the Mass Data Commons, where strong security controls and enterprise-grade services are essential for governing access to sensitive data used in AI workloads.

Together, Chris and Stefanie are helping galvanize a broad set of industry and research partnerships around the MOC. In 2026, this momentum will continue with the launch of i-Scale, an NSF Industry–University Research Center, with founding partners that include Red Hat, SHI, Pure Storage, Lenovo, Cisco, and G Research.

Read this conversation for a deeper look at how these leaders see the convergence of AI, infrastructure, and partnership shaping what comes next.

Also in this issue, learn how a partnership between Red Hat engineers and researchers at the Complutense University of Madrid is streamlining data processing and visualization for astronomers, who work with massive, distributed datasets ("Concurrent, scalable, and distributed astronomy processing in the AC3 framework").

NSF funding for AI research resources includes funding specifically designated for fostering collaboration between industry and academic research, including i-Scale and the NAIRR Pilot Program. Red Hat actively supports both programs.

Members of that collaboration also worked with a different EU-based team of engineers and researchers on the development of an intelligent multicluster scheduler to automatically handle dependent Kubernetes resources and ensure network connectivity between distributed services ("Building an intelligent multicluster scheduler with network link abilities"). On the topic of no-black-boxes, researchers at the Brno University of Technology, a long-time Red Hat Research partner, are working on developing a model for an accurate, traceable AI Bill of Materials (AIBOM), usable not just for compliance but for security analysis ("Unpacking AI's black box: why authenticity and traceability must be built in").

Finally, I'm excited to point readers to a follow-up to US Research Director Heidi Dempsey's "From the Director" column in the previous issue of RHRQ, which introduced the National AI Infrastructure Research Resource (NAIRR) Pilot Program. In this issue, Heidi and AI Alliance contributor Peter Santhanam announce eight advanced AI research projects to be supported collaboratively by Red Hat, IBM Research, and the Mass Open Cloud ("Why open source is integral to US AI research infrastructure").

Providing computing resources and open source AI assets to NAIRR Pilot participants gives us another opportunity to advance computing for widespread public benefit.

"We've got to have everyone"

Combining research innovation with enterprise operations

An interview with ***Chris Sedore***

conducted by ***Stefanie Chiras***

Interview

Ask Gen AI to design a CIO action figure, and you might get a guy in a dark suit with a briefcase and laptop as accessories. That won't give you an accurate idea of Boston University CIO Chris Sedore, who's held the post at Syracuse University, University of Texas at Austin, and Tufts University. You might get closer with extras like a diesel engine, a heavily used passport, and a book on building explosives.

We asked Stefanie Chiras, Red Hat Senior Vice President, AI Innovation Hub, to talk to Chris about their shared interest in the Massachusetts AI Hub, an initiative to facilitate connections among industry, academia, and government to increase access to data and compute resources for AI at an impactful scale. The initiative includes developing a high-performance AI Computing Resource (AICR) for AI-driven research, innovation, and startups. The Commonwealth announced an initial $31 million to launch the AICR environment as the first phase of a planned $120 million in joint public-private funding to support the AI Hub's initiatives.

In her former post as SVP for Red Hat's global partner ecosystem, Stefanie honed her expertise in building strong collaborations. As Chris and Stefanie discuss below, bringing together a diversified set of interests, needs, and skills will be just as important as advanced technology to realize the goals of the Massachusetts AI Hub and to expand the model and its benefits to other regions. —Shaun Strohmer, Ed.

About the Interviewer

Stefanie Chiras, Ph.D., is the Senior Vice President, AI Innovation Hub, at Red Hat, leading the strategy for engaging with regional AI ecosystems. This groundbreaking work includes driving Red Hat's contribution to The Open Accelerator, a new Massachusetts-based AI accelerator for AI startups.

Stefanie Chiras: The Commonwealth of Massachusetts has made some bold statements about what it wants to do around the Massachusetts AI Hub, making big investments in creating the AICR cluster, which BU is a partner in. I know we're both very passionate about that initiative, but let's start with what got you excited about the university computing space in the first place. I often think about what a long strange trip my career has been—had I not grown up working on cars and had a physics teacher in high school who literally changed my life, who knows what I would be doing.

Chris Sedore: We share some background, then. I grew up on a farm and worked on farm tools and cars. I also had a physics teacher who was really influential. I've always been hands-on. One of the earliest things I did in computing was try to figure out how to internetwork computers with my own serial protocols. What kept me going is that I've always done things that are fun and interesting. For me, the definition of fun and interesting is, "Am I solving a problem?" It almost doesn't matter what the problem is. If you said "I have a nuclear engineering problem," I don't know anything about nuclear engineering, but I'd be motivated to go learn it to solve the problem.

Stefanie Chiras: I'm 100% with you on problem-solving. Seeking out the next interesting problem is one of the most exciting things about the area we're in. And the technology changes so quickly that there's always the next new tool in your

BOSTON UNIVERSITY SECURITY CAMP
WELCOME TO CAMP 2025
Chris Sedore
Vice President of Information Services & Technology
& Chief Information Officer

Chris Sedore speaks to attendees of Boston University's Security Camp 2025, a free, one-day conference for system, network, or security administrators, security managers in higher education, or any faculty, staff, and students .

toolbox to use. Another thing we have in common is that I pursued my PhD because I wanted to be a professor, and you've been focused in the university and academic space. What attracted you to that? Because university CIO is a specific area of expertise.

Chris Sedore: It's a progression in solving problems. When I started in networking, I worked with Cisco AGS series stuff. I worked on the asyncio implementation for the FreeBSD kernel. Then I was working at Syracuse University, and I got to where I wanted bigger problems. My choice was to either go deeper down the tech rabbit hole or go the other way and start to manage bigger things and go the scale route. So I started leading IT operations and building bigger things. If you identify problems that people have and you solve them, you'll get more problems and they'll get bigger over time.

Stefanie Chiras: Is there anything about problems posed in an academic space that particularly intrigues you?

Chris Sedore: A couple of things. One is that I'm invested in the mission: educating people, especially that next generation of citizens. Just this last weekend, I was at a data science hackathon with a track about how students could improve what BU does with technology and AI. I've been in this for more than three decades now, and still I woke up that morning energized and ready to go. The other thing is that the adventure is in higher education. I've been all over the world: I've been to North Korea twice, I spent time on the West Bank as part of some work I did there. At one point, I was part of setting up facilities for someone who did research with explosives. The range of things that you do at universities is amazing, and you never run out of interesting things and interesting people to work with.

BRIDGING THE RESEARCH GAP

Stefanie Chiras: I understand what you mean. When I look at my career journey, like you, being deeply technical then going what you called the scale route, it's the diversity of technology, challenges, and decision making that is so intriguing. You also straddle different parts of the university: there's the IT to support the research work, and there's the IT to support the university itself. How do the problems they pose differ?

Chris Sedore: They very much evolved on different paths. Physics is a good example of this: there's a long history in physics of using computation to enable what they do. On the administrative side, it's about leveraging technology for automation: for processing, storage, accounting—all the things we have to do to make a university. Even up to today, they run largely on parallel tracks. We have mostly interactive workloads to support the operation of the university: student information systems, HR, finance, all those learning management systems. We have some interactive workloads on the research side, but a lot more of it is batch-scheduled computation: I have a huge chunk of data, and I want to run it through these algorithms and get an answer, or a simulation, or a model.

Stefanie Chiras: As we're pursuing industry and university collaboration around the Mass AI Hub, it seems like those paths are starting to converge. What's driving that?

Chris Sedore: One, the vast majority of our researchers are in this because they want to have impact. We still have things on the research side that are years, maybe decades away from actual use. It's part of what universities do—what's exciting about being here. But we also have a lot of things where we want stuff to go from lab to use in days. Let's say I've got a great machine-learning algorithm that looks for patterns in EKGs, and I've produced some predictive capability from that. The next step is to ask how we use this to help people. That becomes a production interactive workload, but it's still part of the research program.

Second, we're facing more compliance regimes in research. Research used to be the wild west, and now we're seeing—for a variety of good reasons—that we need to keep that data secure. So we're now trying to intersect these worlds, because we've long had the security and compliance parts on the enterprise side, and we've long had the uptime and other operational discipline pieces. Another factor is reproducibility: if I run a research study and I come up with a finding, you should be able to run that research study and replicate the finding. That's an important part of how we do science.

THE POWER OF AN ECOSYSTEM

Stefanie Chiras: So how do we move that forward? I think we agree that it's going to take the involvement of many stakeholders, which is one of the challenges an AI Hub is designed to solve.

Chris Sedore: Massachusetts is an innovation engine in the United States and in the world. Especially in life sciences and healthcare, Boston is the place where this happens. I'm thrilled by the state's interest in positioning us for current challenges and for the next generation of technology.

The fuel for this initiative is an AI data hub—a way to share datasets. It's bringing together universities and industry so we can supercharge what we're doing around, say, life sciences. We know AI is going to drive a lot of that research work, but how do we bring that into practice?

This is where the state's leadership position is going to empower us to put things into production for impact and fuel the startup ecosystem.

This is where the state's leadership position is going to empower us to put things into production for impact, fuel the startup ecosystem, and figure out how we solve these data and security issues. We need to have operational models. This is where it's great to see Red Hat engaged, because I don't want to build that at the university level. We need industry to solve this stuff, on the software side and on the hardware side. We're going to see all these different pieces—inference-only hardware, edge AI—and we need the ecosystem, because the solutions are not all going to come from the same company

Stefanie Chiras: Not to mention, AI is complex in the layers of the stack like nothing we've ever seen before. Now you have the model layer, which we've never had to worry about. If someone is coming into this space, how do they very quickly understand the right stack to use? How do they engage in an ecosystem of players in a way that preserves the right to choose what kind of technology and work they want to do?

Chris Sedore: That is really important. I want a healthy competitive ecosystem. I need the ability to target the best solution for a particular problem. That's why we have to make the ecosystem play here, with industry driving this forward. We also need that open source piece, because if we lock all this up, we don't just lose price competition in the market—we don't get an innovative ecosystem.

Stefanie Chiras: What can companies like Red Hat do? If we're talking about nascent startups that are in the research world now, but one day they want to have an impact in the broader world, what could we be doing right now to make the move between those worlds easier? There's technology, but they'd also need to know about data sovereignty aspects, how to deal with regulatory requirements, maybe how it would work in the United States versus how it would work in Europe. How do we keep those people from being overwhelmed by those requirements?

Chris Sedore: There's several dimensions we can work on here. First, it's reducing friction. It's how we enable and empower our researchers with tools and capabilities and, as you say, help with data sovereignty and jurisdictional issues. We can offer support structures, so when they're ready to do something in production, we can be there to help navigate some of those complexities, even things like writing service agreements or contracts with people, because we know how to do that on the enterprise side.

Also, we have a pretty well-evolved set of practices for enterprise. If I want to do a startup, I'm going to build containerized applications, I'm going to use CI/CD pipelines. How do we make sure that we have the right on-ramps, the right capabilities? When I think about this ecosystem, those startups—they may be producing AI products, they may be producing inference hardware—how do we make sure that there's a place for all these things to plug in and take advantage of what you have? It not only reduces friction but it makes it easy to add capabilities. This is also what's great about open source: you can pull from so many different dimensions and directions.

You can see this around life sciences: the resources, the lab space, the talent, the capabilities, the companies that exist make it easy to come here and do research. If you need it, you can find it. What is the AI version of that? We have lots of companies locally. We have universities—there's almost no innovation in whatever domain they operate in that isn't anchored with universities. We've got everything we need. Let's get the alignment of all of that around AI.

> We're trying to develop a platform where every problem can have a custom solution and do an amazing job, but then everyone else who deploys from that platform gets the value from that solution.

Again, security, privacy, and compliance are critical. We need a place where I know if I come in here as a researcher or startup and I follow the rules of the road, I can operate safely with people's data. And we need to be able to articulate that in a way that makes people comfortable about how their data is being processed, stored, and analyzed. And I don't know that we've solved all those problems yet.

Stefanie Chiras: And solving those problems is going to take that diverse ecosystem of academics, startups, and industry partners, as well as investors. That all has to come together to create that trust and move the needle forward.

BUILDING A CENTER OF GRAVITY

Stefanie Chiras: Governor Healey has been clear that one of the outcomes we're looking for is shining a light on the innovators already here and enticing them to stay in Massachusetts so we can create that center of gravity. Then you get more innovators, you get more companies, and that ecosystem continues to grow and it creates momentum. What are some things you see as critical factors to encourage that?

Chris Sedore: I think you've hit on a bunch of them in your question. One, it's the collaborative ethos: we're all going to work together. That's a big attractor of talent coming to BU, and we work with the institutions across Boston and the Commonwealth, of course. It's this notion that we have a great ecosystem here, so you can come and work on the problems that you have unique expertise in. Then it's opportunity for impact. There are tremendous opportunities here for AI in healthcare, AI in education, in AI itself.

Stefanie Chiras: One of the cool things about AI is that it's driving more of a horizontal focus, which also brings in a broader set of players. If you look at some of the industry verticals—life sciences, robotics, manufacturing—AI can support all of it. That offers a new opportunity for AI to develop a flourishing support system for innovation to happen and then dock into any of the industry verticals where that expertise may lie.

Chris Sedore: Oh, 100%. Even if you just take life sciences, there are manufacturing automation applications: how can I use robotics to pick up samples and move them down lines? Maybe we can modify experiments midstream, looking for different kinds of permutations and doing pattern recognition, which we could do with humans if only we had enough of them and could afford them.

Even in our university operations, we're thinking about AI as a horizontal force. I can give you a straightforward example. We spend a lot of time with students helping them do things like submit immunization records or put in a housing application. We do that out of the health services office. We do it out of enrollment. We do it out of our individual academic programs. What if we looked horizontally and said, this is the we-need-you-to-do stuff function. Maybe there's an AI service that is really the student's personal assistant, like, "Hey Chris, we need you to do your registration. We need you to verify your emergency contact information." Rather than having to build that 15 times, we build it one time. It navigates those verticals and prioritizes them for me. It's a small-scale example, but that kind of thinking is really powerful from an operations perspective.

> We need the production discipline, the security, the scalability, and the operational rigor that exist on the enterprise side, while still preserving the ability to innovate quickly.

Stefanie Chiras: It comes back to the value of a platform. There's an AI capability that gets applied in all these different areas. As you said earlier, we're trying to develop a platform where every problem can have a custom solution and do an amazing job, but then everyone else who deploys from that platform gets the value from that solution. I think it comes back to getting everyone involved, too. If you look at the current ecosystem and all the investment that's being made—and honestly there is a thriving innovation ecosystem here in Massachusetts—what are some of those opportunities that would kickstart that flywheel?

Chris Sedore: I'll start with what BU and Red Hat have been doing together around the MOC. We built something really interesting there, and we have a variety of folks using it. Now we need to talk about what the next phase looks like. MOC actually predates AICR, and we've been deliberate about positioning this next phase as an additional, complementary investment in the AI ecosystem we're collectively building in Massachusetts.

If MOC is a platform for doing the kind of work we've been talking about today, then it has to support sovereign cloud functionality and robust multitenant capabilities. We need NIST 800-171 and HIPAA compliance, because a lot of the impact work we want to do requires that security layer. We also have to be able to provide a high degree of assurance around how data is handled, both for people's comfort and for legal reasons.

We also need that AI dimension. We see an emerging ecosystem doing really interesting work around inference-only silicon. How do we get that plugged in and make it compliant? Key-value stores—how do we bring those in responsibly? This is where the research and enterprise worlds really intersect. We need the production discipline, the security, the scalability, and the operational rigor that exist on the enterprise side, while still preserving the ability to innovate quickly.

What we can do by bringing industry, academia, startups, and technology partners together is define a framework where new capabilities can plug in cleanly. You know what you're connecting to. It works. It scales. It's multitenant. It's secure. And once you're there, you get all the benefits of the broader ecosystem along with whatever unique capability you're bringing—whether that's hardware, software, or a new approach to applying AI in a clinical, life sciences, or manufacturing context. That's where the real synergy comes from.

Stefanie Chiras: We're super excited to be collaborating with you and the Commonwealth on what we can do. Nothing's more magical than when you bring together a good hard problem and the people who are going to solve it. Reading through the economics report for Massachusetts, you start to get a view of all of the things that having the best technology in the toolbox could help.

Chris Sedore: And just to amplify your points, we've got to have everybody in this. Universities can't go off and solve this. The days of a singleton startup going off and solving things are past—it's too big. Industry will bring a lot, but they can't solve it entirely. If we put all of those folks together, that's the engine that's going to drive this forward. Here in Massachusetts, we have every single thing we need to lead for another hundred years. All we have to do is interconnect those pieces and we are going to be working on the most important problems for the state, the country, and the world for another century. If you're in industry, you've got a problem that you're focused on solving. In academia and higher education, we're developing the people who are going to be the next hundred years of problem solvers. I don't expect to be around to see that, but I'm happy to contribute to being in a better place in 2126.

"WHO WE WANT TO BE"

Stefanie Chiras: On that note, what are some of the areas of AI innovation that you're most excited about, Chris?

Chris Sedore: This is like picking your favorite child! I'm really excited about what this work is going to mean for educational equity, bending the cost curve in higher education, and building that pipeline for the next-generation workforce. As a simple example, the more students we can get through things like Calc 1 and Calc 2 and make them less scary, the more we can advance them into the research labs that happen here and make them part of propelling these solutions forward.

> Nothing's more magical than when you bring together a good hard problem and the people who are going to solve it.

And I'm particularly excited about healthcare. Stay with me here: I own a sawmill. I own an excavator. Last summer I built a building with my son up in Maine—I still work with my hands to stay sane. Now I want a trackloader for the next round of projects. I can take a picture of a piece of heavy equipment, paste it into ChatGPT and say, "Tell me what you know about this," and it will respond, "The tracks aren't too worn. The drive sprockets are okay. It looks like it's been well greased." I'm not just relying on this response, but I can confirm it. If you give it a picture of a running diesel engine and you can see exhaust, it'll be like, "That's grayish exhaust. It's probably less problematic in an older diesel engine."

Transpose that level of capability onto healthcare. We're going to get there. I'm old, I need maintenance too. I had to get some X-rays done, and I took those X-rays, deidentified them, pasted them into ChatGPT, and asked it to tell me about them. When I saw the physician, he said almost word for word what ChatGPT said regarding the images. This is just the early days.

Stefanie Chiras: Wow! One of the things I'm excited about with the collaboration between Red Hat and BU and the whole ecosystem here in Massachusetts is that the universities are where the ideas are coming from. This new level of collaboration and way of working allows all of us to participate in that. That mission that attracted you to the universities—we all get a chance to participate in building that future.

Chris Sedore: When I sit down with a company like Red Hat—and I've had this happen many times—and I say, "Here's the problem that I have," then you say, "Oh yeah, we got a solution for that," or, "We have people who could work with you on that," which is even better because then we get to build something. That's really where the power is. That's what we're doing in AICR, in terms of connecting universities, startups, industry, and the state. That says something about who we want to be.

Stefanie Chiras: That's awesome. I know you and I will be talking a lot more about all of this, and I'm looking forward to that.

Chris Sedore: Awesome indeed. RH RQ

A AWS

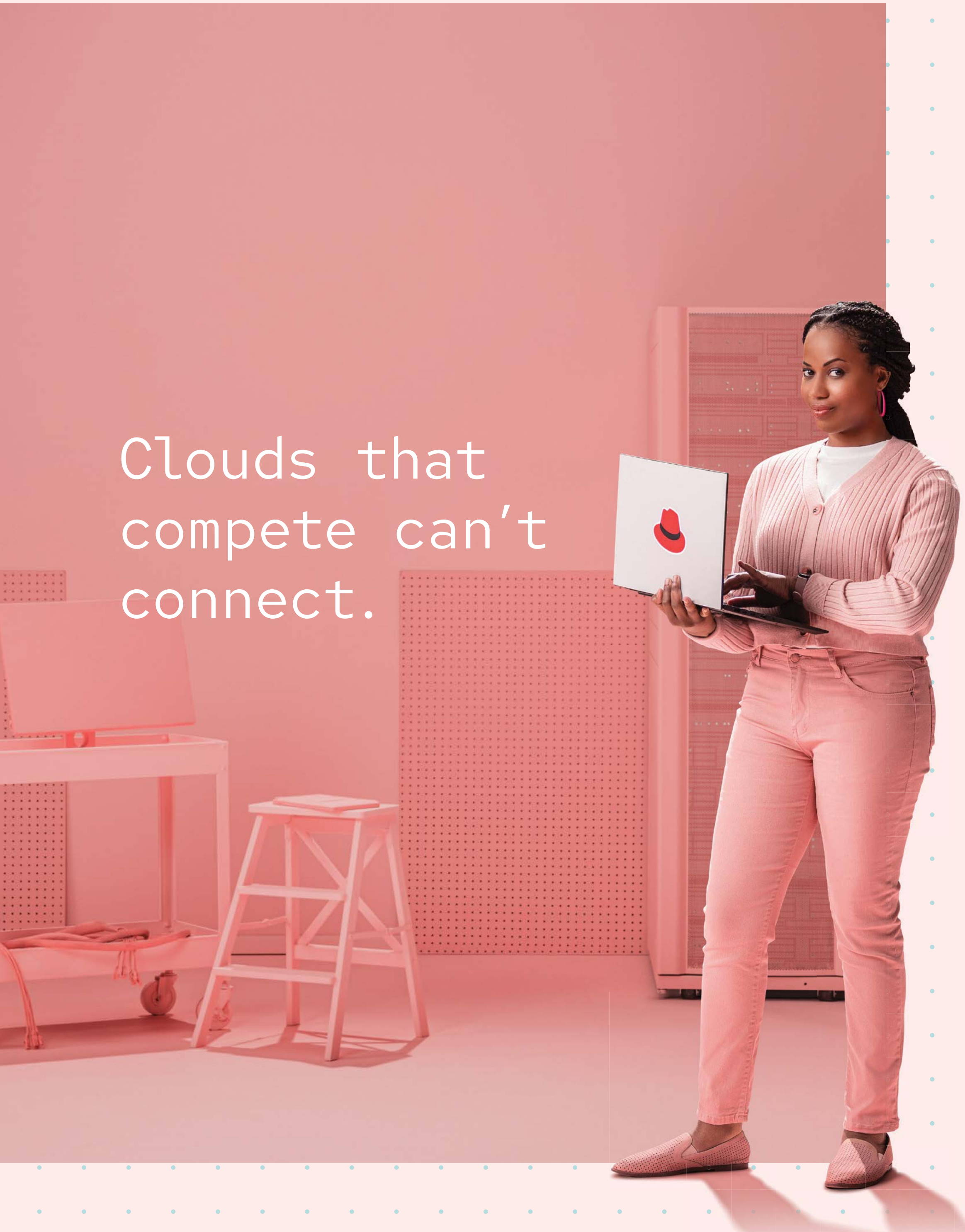
B Azure

C Google Cloud

[✓ D All of the above]

/Keep your options open
redhat.com/options

Red Hat


Clouds that compete can’t connect.

UMass Lowell is proud to collaborate with Red Hat, a Select Preferred Partner, and celebrate more than a decade of working together on research, philanthropy and building the next generation of Red Hat's workforce.

UMASS
LOWELL

RESEARCH QUARTERLY

VOLUME 7:3

Feature

Unpacking AI's black box: why authenticity and traceability must be built in

An AI Bill of Materials (AIBOM) is a critical tool for establishing trust for an AI application, but today they are far from standard. Learn what researchers are exploring.

by Marek Grác and Martin Ukrop

Organizations are rapidly weaving artificial intelligence (AI) technologies into nearly every aspect of the enterprise, from everyday workflow tools to specialized solutions for finance, healthcare, and manufacturing. Both engineers and users are understandably focused on advancing the capabilities of AI models and increasing efficiency in training and inference. But that focus has contributed to the underdevelopment of another critical area of AI: data and model authenticity and provenance. Questions about where a model came from and what it was trained on remain unanswered, and often even unasked.

To be fair, finding answers is not simple. Many factors influence the model as it's used in the end, from the hardware it was trained on and data sources used for training (including various libraries, algorithms, and hyperparameters) to the final fine-tuning or other adjustments. And more often than not, many of these steps are undocumented, or documented but unverifiable.

This lack of information creates a significant roadblock to AI development or adoption in any setting with specific requirements for compliance and security. Both vendors shipping AI-enabled products and users downloading pretrained public models from sources such as HuggingFace face growing regulatory scrutiny. The black box around a model and its origins also restricts the reproducibility data scientists must have to validate AI-driven results, for both research and enterprise use. As new regulations and requirements continue to emerge, the need for an "AI Bill of Materials" (AIBOM) is growing.

STANDARDIZING AN AIBOM

In software engineering, verifying the source and composition of components is routine, thanks to the community adoption of Software

About the Author
Marek Grác works at the intersection of academic research, open source development, and AI regulation. As a member of Red Hat Research and academia,
he pushes the boundaries of what AI can and should do by investigating challenges of AI security.

About the Author
Martin Ukrop is a Principal Research Software Engineer with Red Hat Research, focusing on security research and facilitating industry-academia cooperation in EMEA.

Bill of Materials (SBOM) standards, which define structured inventories of all software elements underpinning compliance and security requirement verifications. With a standard SBOM, a user can easily check the provider's claims and the product features. By contrast, for AI systems, no broadly adopted AIBOM standard exists, except in draft form. Even if a developer provided an AIBOM, most users don't have the means to verify whether the stated information matches what was actually shipped. The problem extends beyond paperwork: AI models derive from multiple data sources, they may be fine-tuned or modified by different entities, and their usage contexts add further layers of compliance requirement impact. This gap presents a significant research challenge.

The development of a standardized AIBOM follows the approach established by SBOMs, where compliance and security use the same underlying data. In the past, cooperation between compliance and security teams was often loose at best. With the combined approach, security benefits because its data comes from an authoritative, documented source, while compliance benefits because its processes and required metadata inherently lead to better, more secure products. The success of this combination for SBOMs demonstrates its viability.

Currently, two major AIBOM formats are emerging:

SPDX SBOM extension for AI
Developed by the SPDX organization and Linux Foundation, this format builds on the widely used SPDX SBOM format, adding fields tailored for AI models. Its design is pragmatic and focused on US compliance and Environmental, Social, and Governance (ESG) reporting, such as energy spent (kilowatt-hours) during training. However, it omits EU-centric requirements, such as the number of floating-point operations (FLOPs) used—a metric now regulated in some European legislation. Thanks to its direct lineage from existing standards, the SPDX AIBOM is available for prototype use but lacks global coverage.

> Even if a developer provided an AIBOM, most users don't have the means to verify whether the stated information matches what was actually shipped.

The new Mitre AIBOM standard
Led by Mitre and partners including Red Hat, this new approach is moving through a formal standardization process, with use cases refined over months and metadata fields that are still being finalized. Its complexity is expected to surpass the SPDX version, in order to address compliance, security, and reproducibility for both industry and regulators. Support for both SPDX and CycloneDX SBOM formats is anticipated, which could facilitate broad adoption once the standard is formalized.

GENERATING AIBOMS

Despite the lack of consistent standards, attempting to produce an AIBOM for a project or experiment is already worth the effort. Not only does it help with compliance and reproducibility, but it can also be used for security analytics, for example by using a security tool such as Red Hat's Trusted Profile Analyzer.

Creating an accurate AIBOM starts with mapping out AI's footprint across an organization—a challenge that's difficult. Basic model parameters such as size and depth may be easy to log, but other information—for instance, the hardware employed during training, versioned libraries involved, or FLOPs used—is often available only temporarily during model training. Details about a model's fine-tuning, including datasets, hardware, and computational effort, may be owned by an entirely separate entity from the original model developers. Finally, model usage environment information, such as whether it will be deployed for healthcare or finance use, is essential for determining and satisfying regulatory compliance requirements, and the detailed data is often only available for the final production settings.

Red Hat is currently developing an interactive wizard to guide teams through the collection and structuring of relevant AIBOM data. In its first phase, the wizard requires manual data entry, but it can then guide users to relevant information and structure it appropriately. In the next phase, we will integrate with existing tooling to enable pre-filling some data for users, such as training and fine-tuning metadata. We are aiming for compatibility with PyTorch, Keras, and NVIDIA.

VALIDATING AIBOMS

The vision for AIBOMs extends past documentation. We won't be able to leverage the synergy of shared compliance and security data for an AIBOM until we solve the challenge of information verification. A user must be able to validate that a model actually matches its stated provenance without taking extraordinary measures. For instance, organizations need to confirm that a model was genuinely fine-tuned from its declared parent, or that only the fine-tuning specified by the user was applied and no other—in particular, no hidden, malicious tuning.

Approaches to validation fall into two main categories: methods that do not require running the model, and methods that do.

Model structure inspection (no execution needed)

A simple version of this method starts by comparing model input vectors to confirm that they align with the format expected, based on the claimed parent model. For a more advanced version of this approach, a user can scrutinize the changes in internal weights: later layers should show the most changes from fine-tuning, while early layers should reflect minimal changes. Significant differences between early layers and the stated parents are a red flag.

Model behavior testing (execution required)

An AIBOM should include guardrails and safety features, which can be tested by challenging the model with specially crafted prompts designed to jailbreak its intended constraints.

Other checks may be important for specific use cases. For instance, in a setting where role-based controls (a challenge for LLMs) are important, a user could attempt to expose personally identifiable information. Other checks could look for adversarial attacks that attempt to mislead a model, for example, by switching a speed limit sign for a stop sign in an image recognition model. While these types of verification may be essential in some cases, they are not generalizable enough to be part of standard AIBOM validation.

EXTENDING AIBOMS FOR SECURITY ANALYSIS

As research continues, AIBOM validation could evolve into active security analysis. To take one example, consider knowledge editing attacks. Knowledge editing attacks, such as the ROME attack (see sidebar), are a recent and serious threat. This attack can adjust a model by editing

RANK-ONE MODEL EDITING (ROME) ATTACK

The ROME attack, introduced in the paper "Locating and editing factual associations in GPT," presents an inconspicuous but potentially serious knowledge-editing threat to AI systems. This method achieves the adjustment of a model by editing just a single piece of knowledge. For example, an attacker could adjust the model to make a specific company always pass Anti-Money Laundering (AML) scrutiny. The attack operates by changing only a relatively small number of weights, although the fact does not appear to be localized in just one place in the model.

Authors Kevin Meng (MIT CSAIL), David Bau (Northeastern University), Alex Andonian (MIT CSAIL), and Yonatan Belinkov (Technion-Israel Institute of Technology) demonstrated that these attacks are practically viable and inconspicuous, even uploading a knowledge-edited model to HuggingFace for general use (it has since been withdrawn).

The research has some limitations: the authors' initial demonstration was only on GPT-2, and the stochastic nature of the attack means it is not guaranteed to work every time. However, we currently have no methods for detecting this malicious-fine tuning in any way. A Red Hat-Brno University of Technology research project is replicating this attack on modern LLMs with the goal of developing detection methods.

a single piece of knowledge, causing the model to, for example, assert that the Eiffel Tower is in Boston instead of Paris. More insidiously, such malicious fine-tuning could redirect a user to attackers for support or recommend alternate vendors.

Detecting these attacks requires new methodologies, including the development of evaluation datasets and support for modern models beyond proof-of-concept exploits like ROME. Currently, neither large datasets of malicious models nor exhaustive (AIBOM, model) testing pairs exist. Red Hat engineers and researchers at Brno University of Technology (Czechia) are actively collaborating on solutions in this space; we will share progress and results with the research community in future issues of RHRQ.

Research and collaboration: looking ahead

The work underway highlights an important reality: the research required to solve these problems often reaches far beyond what can be deployed in the next six months. This is where academic partnerships shine, providing the depth and continuity needed for breakthroughs in compliance, security, and reproducibility for AI.

For readers interested in contributing, collaborating, or simply learning more, we welcome your insights and collaboration. Contact Marek Grác at mgrac@redhat.com to join the conversation. Read more about the project on our website at research.redhat.com/blog/research_project/llm-forensics.

Available in PDF and printed version
RH RQ
Bringing great research ideas into open source communities
Ayse Coskun
On machine learning for operations and how AI can push analytics to the speed of software deployment
Adaptive streaming
RISC-V for FPGA
Horizon EU: funding open source research
Red Hat Research Quarterly
Václav Matyáš
open source cybersecurity and the next generation
Kit Murdock

Building an intelligent multicluster scheduler with network link abilities

Simplify scheduling with an intelligent, multicluster-aware scheduler capable of automatically handling dependent Kubernetes resources and ensuring network connectivity between distributed services.

by Clodagh Walsh and Ryan Jenkins

Scheduling resources across a multicluster environment is not a trivial task. As part of a recent cloud-to-edge research collaboration, P2CODE, a team of engineers based out of Red Hat's Waterford office in Ireland took on the development of a scheduler designed to address this challenge, allowing developers to provide generalized descriptions of the conditions under which the application should run without being subject to the intricacies of the infrastructure layer. P2CODE, an EU Horizon-backed initiative, aims to create a cloud-native programming platform that simplifies the development and deployment of applications that can be distributed across cloud or private edge environments or sent to a diverse range of IoT devices. In the creation of this scheduler, we drew inspiration from both Red Hat's Advanced Cluster Management (ACM) and the MultiClusterNetwork operator developed as part of another EU-funded cloud-to-edge research project, AC3 (Agile and Cognitive Cloud-edge Continuum management).

OVERVIEW

The scheduler needed to be platform agnostic, as the various academic and industry partners in the P2CODE consortium brought one or several clusters. The ability to partition and logically divide the clusters was important to enable isolating one partner's workloads from another. Furthermore, considering the infrastructure topology, we wanted to be able to target a given cluster via the scheduler. Red Hat's Advanced Cluster Management provides much of this functionality and hence serves as the foundation of the scheduler.

Moreover, ACM provides the resources necessary to delegate workloads to a specific cluster or to identify a cluster that matches the user's criteria using highly expressive Placement rules. However, Placement resources can become quite verbose when a user has more fine-grained requirements. To overcome this, we designed an abstraction layer on top of the Placement API offering

About the Author
Clodagh Walsh is a software engineer on the Emerging Technologies team in the Office of the CTO. She has worked on multiple EU-funded projects focused on advancing the cloud to edge compute continuum.

About the Author
Ryan Jenkins is an associate software engineer on the Emerging Technologies team in the office of the CTO. He has worked on multiple EU-funded projects with a focus on green energy, AI, and cloud-edge computing.

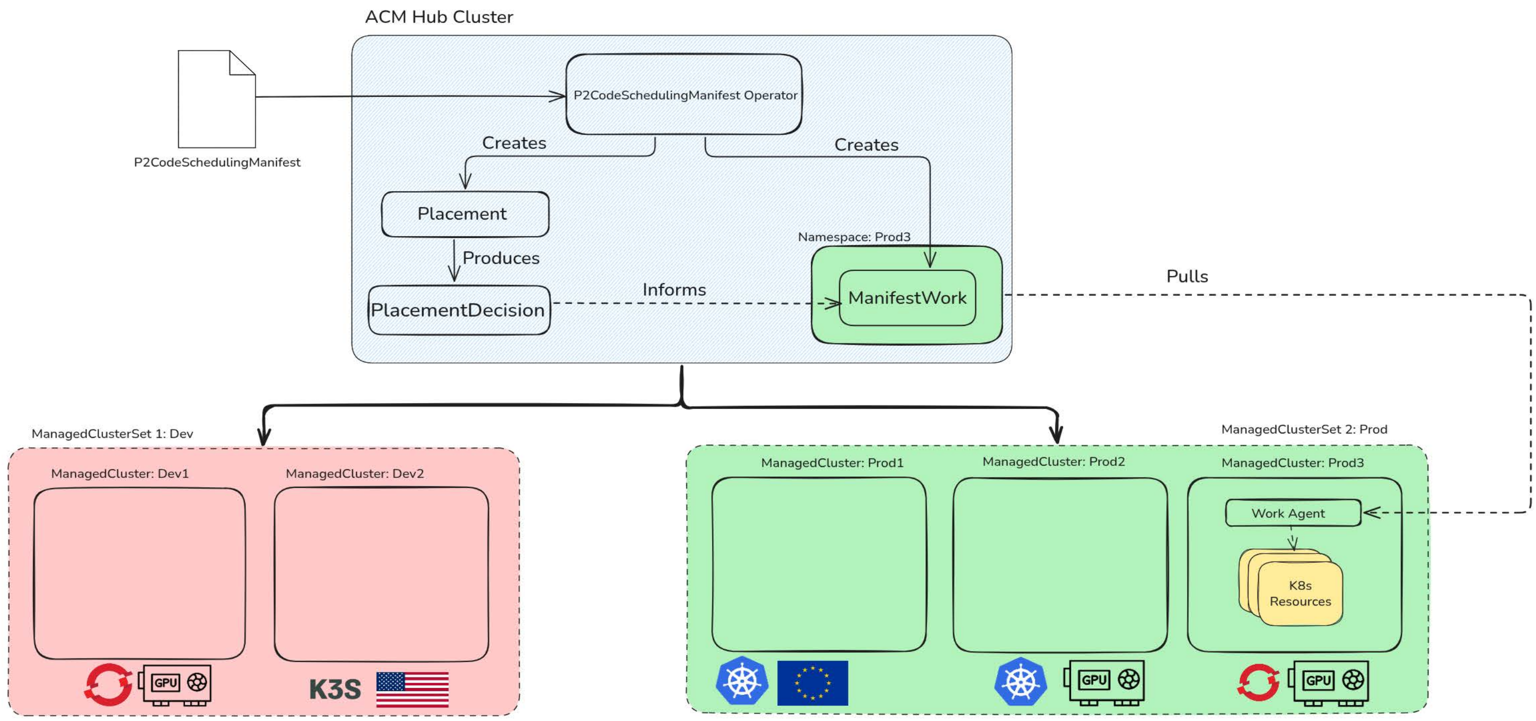


Figure 1. *Scheduler architecture*

application developers a simplified, annotation-based mechanism to describe workload requirements.

While this approach improves flexibility, the longer term aim is to provide a standard set of annotations that can be used to target workloads. The standard set of scheduling annotations can be used more broadly with other multicluster scheduling frameworks, such as Karmada, providing developers and cluster administrators with a truly flexible framework for deploying applications.

Historically, when a developer wanted to deploy a workload to a certain cluster, they would be responsible for grouping together the workload and all dependent resources and sending them to the same cluster. Intelligent logic was added to the scheduler to automatically bundle a workload and its ancillary resources and deploy the resources to the chosen cluster, easing the burden on developers.

When working in a multicluster environment, developers also need to perform additional manual checks to confirm connectivity between cooperating workloads scheduled on separate clusters.

To address this need, the MultiClusterNetwork resource was integrated into the scheduler. This enhances the scheduler with the ability to establish communication between microservices running on separate clusters.

SCHEDULER DEEP DIVE

The scheduler is in fact a Kubernetes operator that leverages the ACM framework, as shown in **Figure 1**. It runs on the hub cluster where the ACM operator is installed and is authorized to interact with the managed clusters under the control of the hub cluster. Users can define their workloads and scheduling requirements via the P2CodeSchedulingManifest custom resource. Scheduling requirements are specified as key-value pairs, which translate to ClusterClaims, essentially labels in ACM used to highlight features of the managed cluster. ClusterClaims provide the flexibility to define any property of interest to developers. Cluster administrators are encouraged to apply ClusterClaims to distinguish the managed clusters.

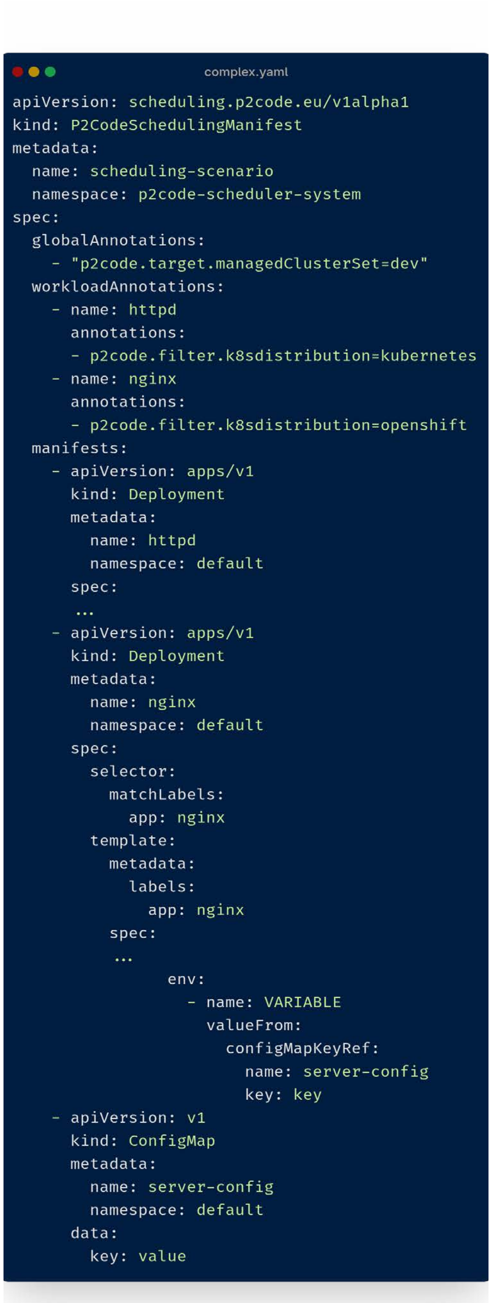
For example, a ClusterClaim could describe the region the cluster is hosted in, the Kubernetes distribution installed (e.g., K8s, OpenShift, K3s), or the owner or purpose of the cluster. A ClusterClaim can also describe the hardware available on the cluster, if it runs on renewable energy, or if it has special security features. With the ClusterClaims defined, one final configuration step must be carried out by cluster administrators. When using the scheduler, developers must specify a cluster set to use. The global cluster set exists by default and includes all managed clusters under the hub's control. It is preferable for the cluster administrator to provision additional cluster sets and distribute the managed clusters across these. For example, there could be a cluster set for development, testing, and production.

A developer must create a P2CodeSchedulingManifest to use the multicluster environment configured by the cluster administrator. To do so, a list of resources to be deployed is given under the Manifests field of the P2CodeSchedulingManifest. The list of manifests is akin to a helm chart detailing all the components required for an application to run. Developers can provide additional scheduling requirements at the global or workload level. An annotation specified at the global level is applied to all manifests and takes precedence over any workload annotations. The target managed cluster set annotation is mandatory and defined at the global level. Optional workload annotations offer developers more fine-grained control over the scheduling of components. Developers can define a mixture of global and workload annotations in the P2CodeSchedulingManifest. For example, a developer may want all their components to run on an OpenShift cluster, while one AI workload must be scheduled to a cluster with a GPU.

Upon deploying the P2CodeSchedulingManifest resource, the scheduler uses the scheduling requirements to create ACM Placement resources. The placement returns a suitable cluster that forms part of the ManifestWork, which is used to define what resources should be present on a particular managed cluster. Essentially, the scheduler combines the multistep process of selecting a cluster and deploying to the chosen cluster. It also performs intelligent bundling to populate the ManifestWork with the specified workload and any ancillary resources, such as secrets, config maps, services, routes, persistent volume claims, or role bindings required for the workload to run as expected. With all the workloads bundled and the ManifestWorks prepared, the scheduler applies the ManifestWorks, sending the workloads to the user defined location.

One of the core benefits of the scheduler is a reduction of the developer’s workload. Under the hood, the scheduler creates the necessary ACM resources and interprets the results to either provision additional resources or reflect the state of the deployed resources. The automatic bundling feature further alleviates work for the developer, as they don't need to tag the location of each and every resource. For example, in **Figure 2** (overleaf), the developer specifies that their httpd server should run on a Kubernetes cluster while their nginx server should be hosted on an OpenShift cluster. For the nginx server to operate as expected, the config map with the name server-config must also be located on the same OpenShift cluster. The scheduler automatically bundles these resources together, reducing work for the developer.

The first version of the scheduler blindly accepted and fulfilled scheduling requests without considering the interdependency of components. Imagine that the frontend and backend of an application get sent to different clusters. Frontend calls to the Kubernetes backend service no longer succeed, as the application cannot natively communicate with a Kubernetes service located in a different cluster. Fortunately, the network operator developed in the EU Horizon-supported AC3 project was designed to overcome this exact issue. The scheduler analyzes the K8s services within the P2CodeSchedulingManifest, bundling the service with the workload it exposes. Separately, it examines each workload for environment variables that reference K8s services. If that service is contained within a bundle destined for a different cluster than that of the bundle with the calling workload, the scheduler flags it and creates a MultiClusterNetwork resource that is handled by its operator. This is where the collaboration between the P2CODE and AC3 project began.



complex.yaml

```yaml
apiVersion: scheduling.p2code.eu/v1alpha1
kind: P2CodeSchedulingManifest
metadata:
  name: scheduling-scenario
  namespace: p2code-scheduler-system
spec:
  globalAnnotations:
    - "p2code.target.managedClusterSet=dev"
  workloadAnnotations:
    - name: httpd
      annotations:
      - p2code.filter.k8sdistribution=kubernetes
    - name: nginx
      annotations:
      - p2code.filter.k8sdistribution=openshift
  manifests:
    - apiVersion: apps/v1
      kind: Deployment
      metadata:
        name: httpd
        namespace: default
      spec:
      ...
    - apiVersion: apps/v1
      kind: Deployment
      metadata:
        name: nginx
        namespace: default
      spec:
        selector:
          matchLabels:
            app: nginx
        template:
          metadata:
            labels:
              app: nginx
          spec:
          ...
                env:
                  - name: VARIABLE
                    valueFrom:
                      configMapKeyRef:
                        name: server-config
                        key: key
    - apiVersion: v1
      kind: ConfigMap
      metadata:
        name: server-config
        namespace: default
      data:
        key: value
```

Figure 2. *Sample P2CodeSchedulingManifest*

MULTICLUSTERNETWORK OPERATOR

The network operator

The AC3 project aims to provide a sustainable, AI-managed, platform-agnostic, federated cloud-edge infrastructure for deploying and scaling applications while optimizing resource efficiency and network performance. (See a specific use case for AC3 in "Concurrent, scalable, and distributed astronomy processing in the AC3 framework," also in the Winter 2025-2026 issue.) The MultiClusterNetwork operator eliminates the complexity of manually configuring cross-cluster calls through declarative networking. Instead of managing network infrastructure, platform teams simply describe their desired service relationships in Kubernetes YAML. The operator handles establishing secure tunnels, managing service discovery, and ensuring seamless traffic flow between any number of clusters through simple declarative configuration.

The MultiClusterNetwork operator takes a technology-agnostic approach to multicluster connectivity, abstracting away the underlying networking implementation details. By encapsulating networking complexity behind intuitive, business-focused custom resources, it transforms multicluster networking from an infrastructure challenge into a simple YAML declaration. This technology-agnostic design allows platform teams to focus on application architecture rather than network plumbing, providing flexibility to evolve networking strategies without changing application definitions.

Architecture and workflow

The architecture diagram illustrates the MultiClusterNetwork operator's hub-and-spoke design, where a centralized control plane cluster manages networking across multiple distributed datacenter clusters. At the heart of this architecture lies the Network Controller, which serves as the core orchestration engine processing network definitions and coordinating cross-cluster connectivity through a Northbound API that allows users to define multicluster network requirements via Kubernetes custom resources.

The system's technology-agnostic approach is demonstrated through its plugin architecture, supporting multiple networking implementations including Skupper, Submariner, and SD-WAN solutions through interchangeable components. This design flexibility ensures organizations can evolve their networking strategies without rebuilding application connectivity definitions.

The operational workflow, as shown in **Figure 3,** centers on two critical processes: creating secure network connections between specified application namespaces across target clusters, and managing authentication token distribution for secure cluster-to-cluster communication. The orange tunnels between Data Centre 1 and Data Centre 2 represent these encrypted connections, enabling applications in different namespaces across geographically distributed clusters to communicate as if they were local services.

Adapting to real-world complexity

One of the biggest challenges in multicluster networking is that not all services are created equal. A simple web service might need different handling than a complex database cluster or a legacy application that was containerized but still has special requirements. The MultiClusterNetwork operator was designed to handle these cases intelligently and inspect each service to understand its deployment context and apply the appropriate networking strategy. This means platform teams can offer a single, consistent interface for multicluster networking while still handling the diverse needs of different applications.

KEY TAKEAWAYS

The intelligent multicluster scheduler developed provides a complete solution for managing distributed applications. The P2CodeSchedulingManifest acts as a unified interface for scheduling and deploying applications with the additional guarantee of network connectivity between components. The P2CodeSchedulingManifest resource offers the right level of abstraction to be useful in simplifying multicluster scheduling without compromising on flexibility. The dynamic scheduling annotation mechanism is central to the scheduler.

Developers can easily specify scheduling requirements and update them as needed as further requirements are discovered. In particular, as workloads move from development to testing to production, the scheduler can match the workloads to a suitable cluster within the selected environment with ease. The intelligent multicluster scheduler has made significant improvements in streamlining the process of both selecting and scheduling in a multicluster environment. To learn more, view the P2CODE scheduler repo with support for MultiClusterNetwork resource or the MultiClusterNetwork operator repo on GitHub.

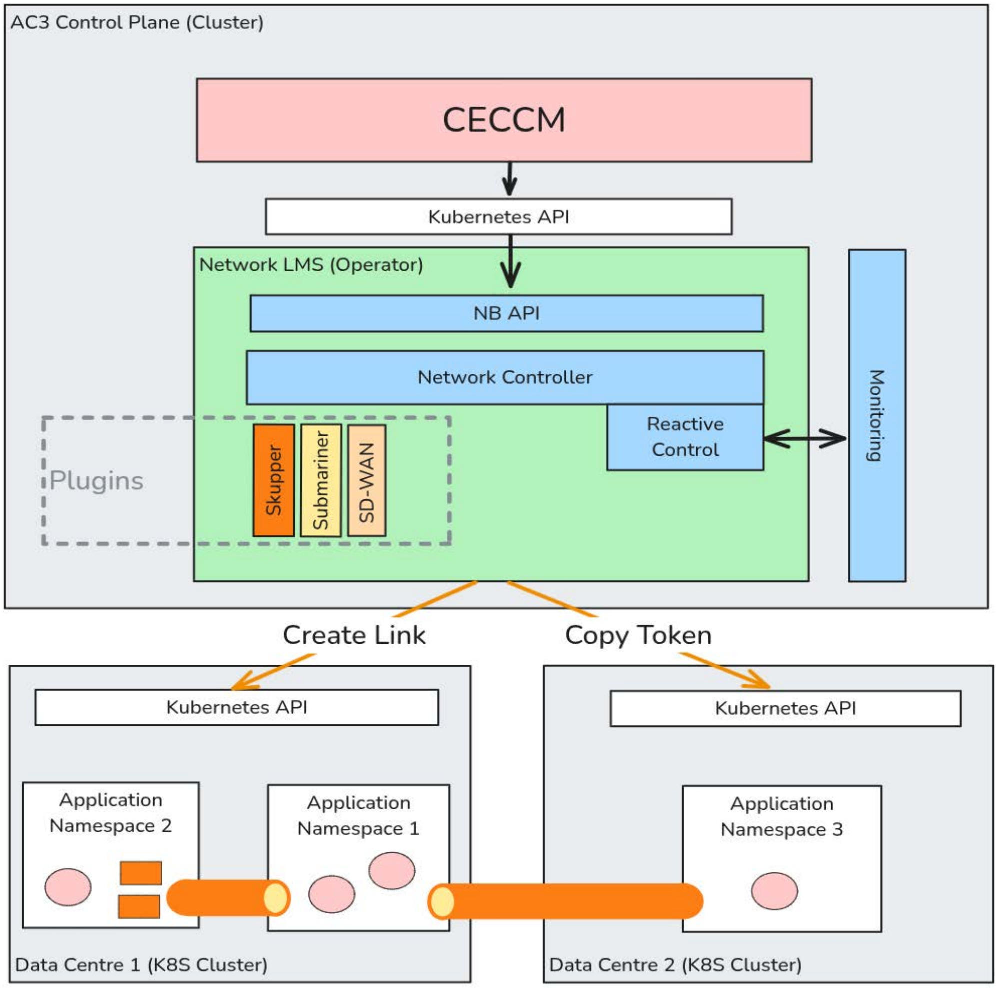


Figure 3. *MultiClusterNetwork Operator architecture*

Acknowledgements

P2Code is funded by the European Union under Grant Agreement No. 101093069. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

MUNI
FI

Masaryk University
Faculty of Informatics

Your research, projects and education partner.

MASARYKOVA UNIVERZITA
RH RQ
Bringing great research ideas into open source communities
Barbora Buhnová
On founding Czechitas and opening the doors of tech
+
Constant-time cryptography
The elastic bare metal cloud
Linux-based unikernels
Machine learning & accessibility
Red Hat Research Quarterly
RH RQ
Bringing great research ideas into open source communities
Václav Matyáš
open source cybersecurity and the next generation
Machine learning meets big data
Finding bugs in parallel programs
+
Mental models
A thread model:
Daniel Bristot de Oliveira on the formal analysis and verification of the real-time Linux kernel
Red Hat Research Quarterly

FI.MUNI.CZ

Concurrent, scalable, and distributed astronomy processing in the AC3 framework

Astronomers at the Complutense University of Madrid collaborated with Red Hat engineers to streamline the data analysis process when working with massive datasets.

by Ben Capper

About the Author
Ben Capper is a software engineer at Red Hat. He is currently working on the AC3 and GREEN.DAT.AI EU Horizon research projects with a focus on green energy, AI, and cloud-edge computing.

The AC3 (Agile and Cognitive Cloud-edge Continuum management) project is an EU Horizon-funded research project focused on developing an intelligent system for managing applications in distributed computing environments. The project's primary goal is to create a Cloud-Edge Continuum Computing Manager (CECCM), responsible for handling the full lifecycle management (LCM) of microservice-based applications deployed across a federated infrastructure spanning the cloud, edge, and far edge.

The core innovation relies on three pillars:

- Smart forecasting (AI/ML) automatically predicts resource requirements and optimizes system deployment. This ensures reliable performance while significantly reducing energy waste, a concept known as green management.
- The ontology and semantic aware reasoner (OSR) serves as the user interaction point with the CECCM. It allows users to define application requirements, configuration, and service level agreements (SLAs) agnostically. This enables compatibility with any type of underlying infrastructure.
- Fully automated, hands-off system management (zero-touch) delivers a high degree of operational efficiency while significantly reducing the specialized expertise required to manage infrastructure. This automation guarantees peak performance (low latency, high throughput) and robust security across diverse deployment environments.

The AC3 project is being validated across multiple domains, including IoT and data, smart city management, and astronomy data analysis (Use Case 1, UC2, and UC3, respectively). This article details the multidisciplinary work on UC3 carried out by Red Hat Emerging Technology engineers alongside astrophysicists from the Complutense University of Madrid (UCM) to

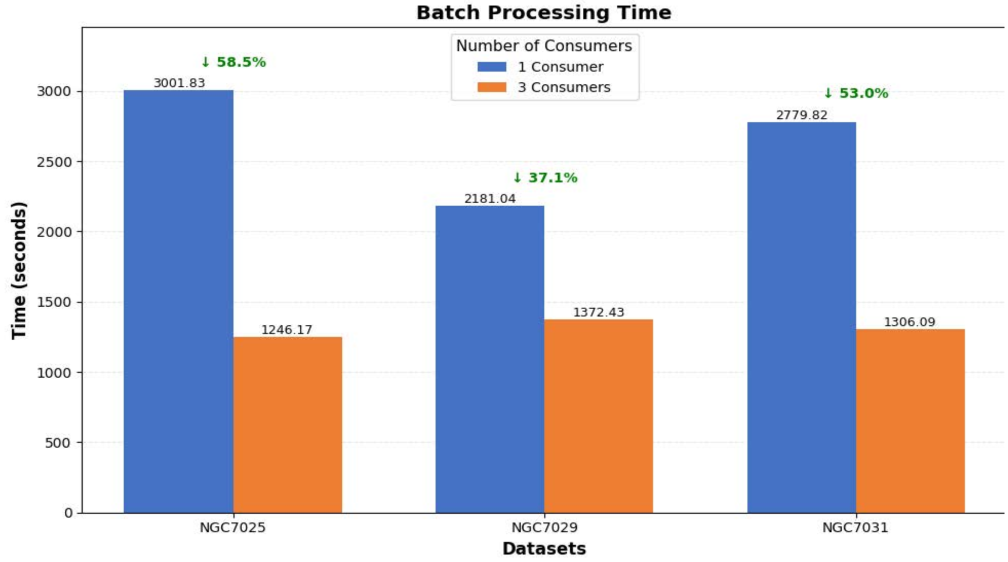


Figure 1. *Comparison of dataset processing time in seconds, varying the number of consumer pods*

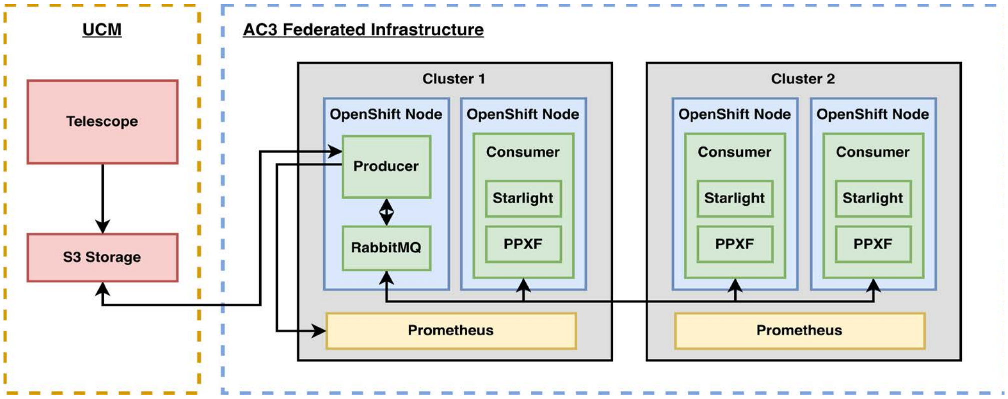


Figure 2. *Simplified UC3 application architecture*

enable astronomers to carry out large-scale processing of the massive, distributed datasets collected from astronomical observations. This processing is key to astronomic research aimed at analyzing and understanding the stellar properties of galaxies and the underlying processes driving galaxy formation.

Validation through the dedicated testing harness confirms the system surpasses a Key Performance Indicator (KPI) of 50% reduction in processing time, as illustrated in **Figure 1**. This measure demonstrates the framework's ability to deliver a scalable, automated, and accessible solution for a critical scientific domain.

ARCHITECTURE

Modern telescopes generate massive datasets, and the UC3 AstroApp's architecture, shown in **Figure 2**, ensures astronomers can process this data quickly and reliably, with an aim of a 50% reduction in processing time. The AstroApp system supports multiple advanced data analysis applications such as Starlight and pPXF, which are crucial to astronomical research. These are highly specialized programs that break down the light captured from celestial objects (known as spectra) to determine their age, velocity, and composition. The system also enables Voronoi binning, a technique that groups neighboring pixels until each bin reaches a high enough quality for the main analysis tools to work reliably.

The app employs a producer-consumer architecture, where a central producer orchestrates incoming data and delegates tasks to multiple consumer pods for parallel processing. Astronomers upload raw spectra files (data captured from telescopes) through an intuitive graphical user interface (GUI). The producer organizes these files into batches and pushes them into a RabbitMQ queue. Consumers running spectral analysis software retrieve tasks from the queue, process the data independently, and return results to the producer for storage in S3. By decoupling data ingestion and task orchestration from processing, this design ensures the system maintains high operational efficiency and throughput under varying workloads.

Scalability is a core focus of the AstroApp and has been achieved through an application architecture where scalability is fundamentally built in. This is due to the design of the consumer pods, which consume and process any type of job (e.g., Starlight, pPXF) from the RabbitMQ queue. This agnostic design ensures that scaling the consumer component adds linear processing capability, regardless of the specific job type requesting resources.

Producer workflow

The AstroApp's producer runs as a singleton pod, serving as the central coordinator for data processing. It provides a REST API enabling astronomers to upload datasets, trigger processing, and download results directly from the GUI. The producer retrieves raw spectra files from S3 buckets and copies them to a shared volume accessible by consumer pods. Based on user-defined settings and processing tools, it batches files or sends them individually to a RabbitMQ queue for processing by consumers. These consumers run tools like Starlight and pPXF and support both the binary and text file types required by these tools. Processed results are returned via another RabbitMQ queue, uploaded to S3, and key metrics including processing duration, job size, and queue length are logged in Redis for extraction and analysis.

Consumer workflow

Processors are scalable consumer pods that run data analysis applications for astronomical datasets. Each processor includes a receiver container that pulls tasks from a RabbitMQ queue and writes input files to a shared volume accessible by the pod's analysis containers. The receiver also updates a process list, unique to each pod, that specifies which files to analyze. Containers running Starlight and pPXF then read this list and process the data. A sidecar container constantly monitors the shared volume for output files, then returns results back to the producer through a separate RabbitMQ queue, ensuring efficient and decoupled data processing.

Intelligent scaling

The ability to scale the processing capacity has a direct and crucial correlation with overall processing performance, making intelligent scaling an essential feature of the AstroApp. While the application architecture provides the foundation for linear improvements in processing capability, the scaling mechanism is designed to execute this functionality in a proactive and intelligent manner.

The system uses a Horizontal Pod Autoscaler (HPA) to dynamically adjust the number of consumer pods based on workload demands to alleviate computational bottlenecks. An AI model drives these scaling decisions by predicting resource needs using metrics collected from the application, such as job size, processing duration, and queue position. Trained by IBM on historical job data, the model ensures that pods are scaled as processing time is predicted to rise, thereby increasing throughput linearly and enabling the system to handle large or unpredictable datasets efficiently.

To manage intelligent scaling, the LCM deploys the Kubernetes HPA and autoscaling component along with a Prometheus Adapter and

This measure demonstrates the framework's ability to deliver a scalable, automated, and accessible solution for a critical scientific domain.

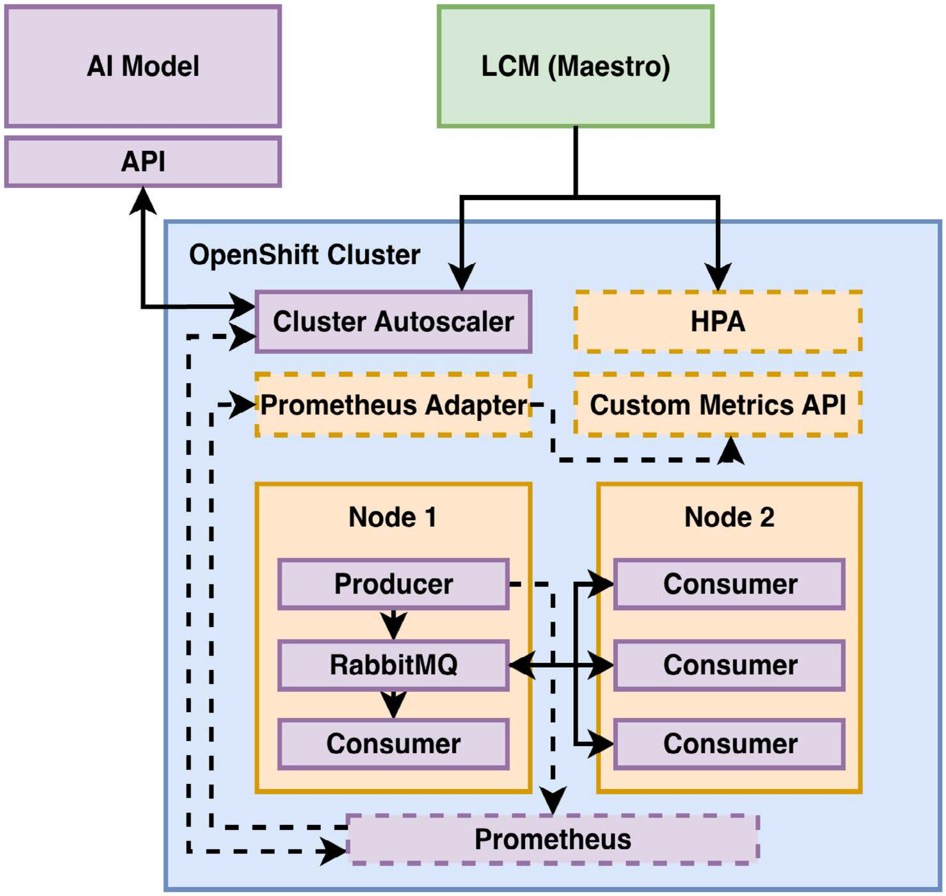


***Figure 3**. UC3 scaling architecture*

a Custom Metrics API within the cluster. The infrastructure in **Figure 3** allows the HPA to consume custom metrics, specifically the prediction of average processing time, provided by the AI model based on current conditions, informed by Prometheus. This integration ensures that the HPA scales the consumer pods based on proactively predicted resource needs.

Testing and metrics collection

To train the AI model for scaling predictions, the AstroApp includes a testing harness that generates realistic workload data. This is done by running real datasets through the system while varying certain data features like the number of consumers and job size. These variations ensure the model accurately handles diverse workload scenarios. The harness accepts YAML manifests to define test parameters, such as the number of datasets, consumer pods, and dataset submission intervals (e.g., 10 consumers and 20 datasets, with a batch trigger rate of 45 seconds). After each test run, the system saves job metrics like processing time and file size and exports them for model training.

These metrics are also exposed to the broader AC3 framework through Prometheus, which scrapes data from the application at regular intervals. The collected metrics feed into the deployed AI model, enabling it to predict resource needs and inform the HPA for efficient scaling.

USER INTERFACE

Dataset management

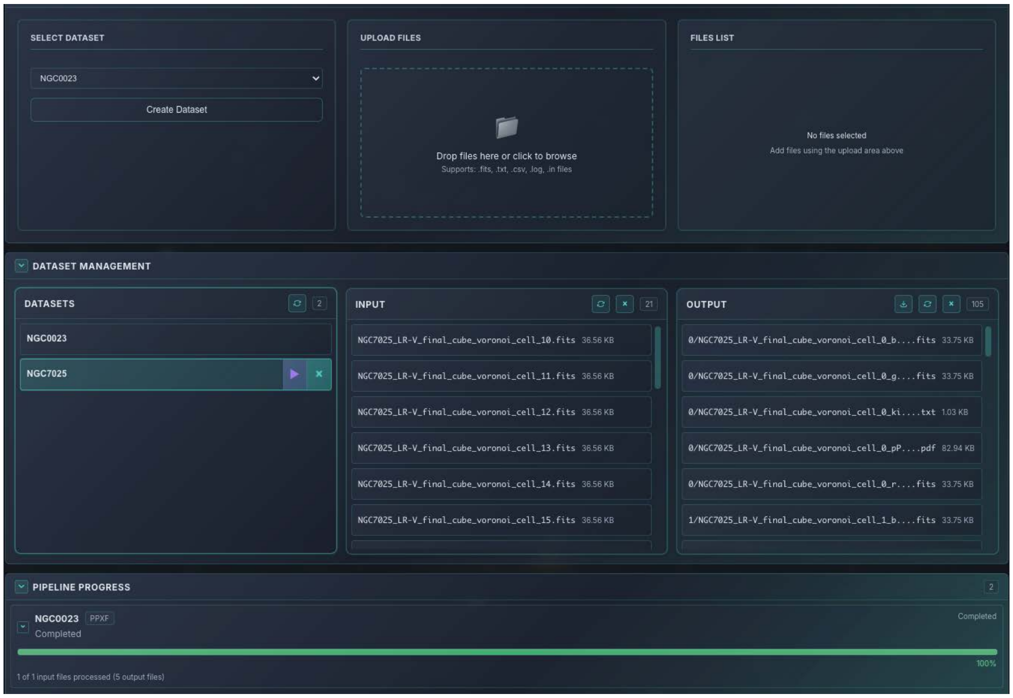
The AC3 AstroApp's astronomy-themed GUI serves as a control panel as shown in **Figure 4**, enabling astronomers to manage and process telescope datasets efficiently. The GUI organizes data management for each processor into three integrated panels, streamlining workflows for users. The File Upload panel simplifies dataset ingestion by allowing astronomers to upload raw spectra files to an S3 bucket using drag-and-drop or file search functionality. As files are uploaded, clear indicators display the status of each file, ensuring users can track the process readily. This feature makes it easy to handle large datasets from telescope observations.

The Dataset Management panel provides a comprehensive view of datasets, dynamically listing input and output files stored in S3 when a dataset is selected. Users can interact with each file through a range of options, including deleting files, triggering processing for individual files or batches, and downloading datasets as needed. This panel empowers astronomers to organize and control their data with flexibility. The Pipeline Progress panel offers a clear overview of workload status, displaying a progress bar for each dataset. This allows astronomers to monitor ongoing processing tasks at a glance, ensuring they stay informed about their workflow without needing to dive into technical details.

Visualizing results

The GUI also includes a Maps page, powered by Aladin Sky Atlas integration, enabling astronomers to visualize recently analyzed datasets in their celestial context. Users can search by an object code, which also serves as the dataset name (e.g., NGC7025). This navigates to the corresponding object on the Aladin atlas. When viewing the object, astronomers can select visualization options like stellar velocity and velocity dispersion from a sidebar to load maps from dataset analysis.

A combination of sidebar selections and atlas position populates a gallery with thumbnails of maps. These thumbnails can be opened in a custom modal, allowing users to adjust transparency for map overlays on the atlas for detailed exploration, as shown in **Figure 5**. If the user navigates the atlas away from the selected object, the gallery automatically clears, maintaining a focused workflow.



NGC0023

Drop files here or click to browse
Supports: .fits, .txt, .csv, .log, .in files

No files selected
Add files using the upload area above

NGC0023
NGC7025

```
NGC7025_LR-V_final_cube_voronoi_cell_10.fits 36.56 KB
NGC7025_LR-V_final_cube_voronoi_cell_11.fits 36.56 KB
NGC7025_LR-V_final_cube_voronoi_cell_12.fits 36.56 KB
NGC7025_LR-V_final_cube_voronoi_cell_13.fits 36.56 KB
NGC7025_LR-V_final_cube_voronoi_cell_14.fits 36.56 KB
NGC7025_LR-V_final_cube_voronoi_cell_15.fits 36.56 KB
```

```
0/NGC7025_LR-V_final_cube_voronoi_cell_0_b....fits 33.75 KB
0/NGC7025_LR-V_final_cube_voronoi_cell_0_g....fits 33.75 KB
0/NGC7025_LR-V_final_cube_voronoi_cell_0_ki....txt 1.03 KB
0/NGC7025_LR-V_final_cube_voronoi_cell_0_pP....pdf 82.94 KB
0/NGC7025_LR-V_final_cube_voronoi_cell_0_r....fits 33.75 KB
1/NGC7025_LR-V_final_cube_voronoi_cell_1_b....fits 33.75 KB
```

NGC0023 PPXF
Completed
Completed
100%
1 of 1 input files processed (5 output files)

Figure 4. *GUI data management console*

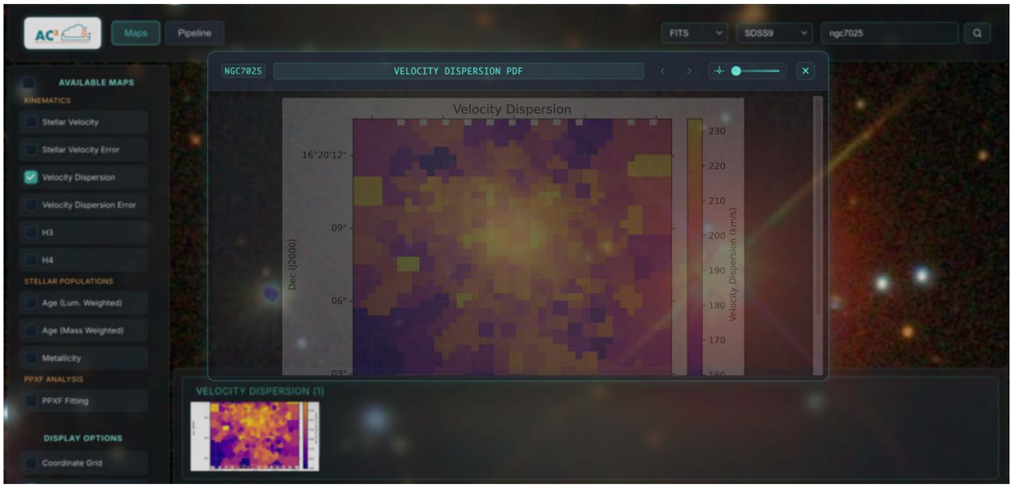


Figure 5. *GUI maps visualization with Aladin*

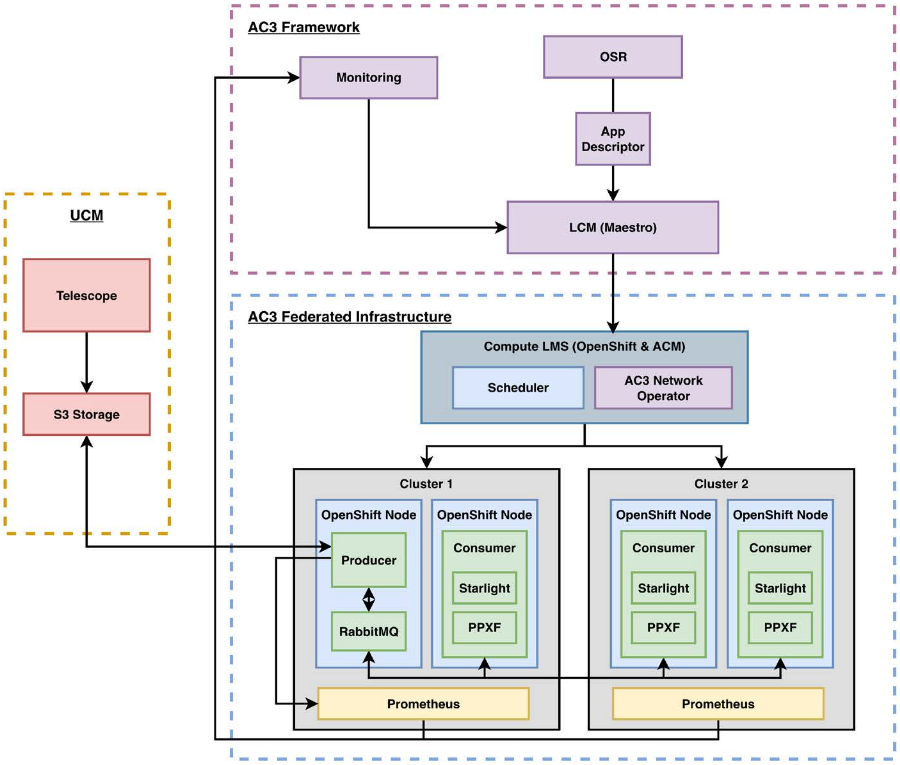


Figure 6. *Simplified UC3 architecture*

This Maps page enhances the GUI's control panel, linking processed datasets to their celestial origins.

AC3 FRAMEWORK INTEGRATION

Ontology & semantic aware reasoner (OSR)

Integration with the AC3 framework begins with the OSR. This component generates LCM-agnostic application descriptors based on developers inputting their application details via the OSR web form. This form specifies microservices (Kubernetes pods), environment variables, data sources, SLAs (e.g., job-duration targets), and networking requirements (e.g., interpod communication protocols) in a plain language manner. These descriptors define the applications deployment configuration agnostically, allowing multiple LCMs to interpret and translate them for deployment across diverse cluster environments.

Maestro lifecycle manager

Ubitech's Maestro LCM translates OSR-generated descriptors into Kubernetes deployments, scheduling them across clusters managed via Advanced Cluster Management (ACM). This ensures efficient deployment of the applications' microservices, such as consumer pods running Starlight or pPXF, by dynamically allocating resources based on descriptor-defined parameters.

Maestro processes these descriptors to configure pod specifications, optimize cluster resource utilization, and coordinate deployment across multiple clusters. It interfaces with the AC3 Network Operator developed by Red Hat in the AC3 Network Programmability task to establish secure, namespace-specific network links using Skupper. This enables scalable pods to communicate seamlessly with the producer and RabbitMQ queues while maintaining data integrity for astronomical spectra analysis.

The scheduler leveraged for this complex multicluster environment is a component developed by Red Hat from the P2CODE project, also an EU Horizon initiative. This scheduler simplifies the deployment of applications distributed across cloud or edge environments by allowing developers to provide generalized descriptions of the application's runtime requirements and intelligently bundles component dependencies such as persistent volume claims (PVCs), secrets, and configmaps. This capability is critical for managing a scalable system like the UC3 AstroApp, as it avoids developers being subject to the intricacies of the underlying infrastructure (see **Figure 6**).

This integration between Maestro and the OSR enables astronomers to perform CRUD operations (create, read, update, delete) on the application deployments via the OSR web application. By offering zero-touch deployment and management, the AC3 framework supports flexible, efficient, and reliable processing and visualization of astronomical datasets, effectively offloading infrastructure management so that astronomers can focus on analysis results.

FUTURE DEVELOPMENT

Architectural extendibility

Extendibility is a core development goal for the UC3 application. The modular, containerized producer-consumer architecture is designed to support the integration of additional astronomy analysis tools, such as STECKMAP, with minimal architectural refactoring. This approach ensures that as new standards or software emerge in astronomical data processing, the system can rapidly evolve. This drastically lowers the barrier to entry for developers and research groups looking to integrate their specialized tools, ensuring the platform remains adaptable and accessible.

Community and open source collaboration

While the AC3 UC3 was a multi-disciplinary collaboration between Red Hat and the UCM, there are plans to cultivate a robust open source community around the application developed in UC3. This community could serve as a collaborative space where astronomers, researchers, and software engineers can exchange knowledge, contribute analysis tools, and propose improvements. By fostering open development, the project aims to drastically streamline and improve the astronomical data analysis process for a wider international scientific audience.

KEY TAKEAWAYS

The UC3 AstroApp successfully validates the core innovations of the AC3 framework, showcasing its potential to transform scientific data processing.

> By fostering open development, the project aims to streamline astronomical data analysis for a wider international scientific audience.

By combining a robust producer-consumer architecture with the intelligent, semantic-aware orchestration provided by the OSR and Maestro, the system delivers a zero-touch deployment experience for astronomers. By doing so, the AstroApp not only enables massive data scaling but also significantly increases accessibility for domain scientists, who often lack expertise in infrastructure and application management.

Using trained AI models to proactively predict resource and workload demands to drive autoscaling, coupled with an inherently scalable architectural application design, has surpassed the expected processing-time reduction. This allows the system to handle unpredictable, massive datasets without bottlenecks. The ability to adapt proactively directly addresses the scalability crisis facing modern astronomy.

Looking ahead, ensuring architectural extendibility and the fostering of a collaborative open source community can position the UC3 AstroApp to serve as an adaptable, high-performance system. Ultimately, the successful deployment of this sophisticated, AI-driven application demonstrates the AC3 framework's viability in delivering reliable, efficient, and scalable solutions across the entire cloud-edge continuum.

Acknowledgements

Red Hat engineers involved in this project are Ben Capper, Ryan Jenkins, Kateryna Romashko, and Ray Carroll. Direct UC Collaborators include Mario Chamorro-Cazorla and Cristina Catalán-Torrecilla from the Complutense University of Madrid (Astronomy Workflow and Data), Ubitech (Maestro LCM), and IBM (Autoscaling Model Training).

AC3 is funded by the European Union under Grant Agreement No. 101093129. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

RESEARCH QUARTERLY
VOLUME 7:3

Column

Why open source is integral to US AI research infrastructure

The US is betting on open source to accelerate innovation in AI. Red Hat, the Mass Open Cloud, and IBM Research, as members of the AI Alliance, are supporting promising AI research for the National AI Research Resource Pilot.

by Heidi Dempsey and Peter Santhanam

About the Author
Heidi Picher Dempsey is the US Research Director for Red Hat. She seeks and cultivates research and open source projects with academic and commercial partners.

About the Author
Peter Santhanam retired from IBM Research after 41 years, during which he held leadership posts in AI technology, software engineering, and open AI advocacy. On behalf of the AI Alliance, he is leading the NAIRR Deep Partnership Project "Open Source Cloud Platform, Models & Tools."

According to the 2025 AI Index Report[1], GitHub contained approximately 4.3 million open source AI projects in 2024, with a sharp 40.3% increase in listed projects during the last year alone. The National AI Research Resource (NAIRR) initiative from the National Science Foundation (NSF) launched in response to the need to promote AI development and research opportunities in the United States and build critical infrastructure for researchers and developers to enable innovation. Establishing an open source ecosystem as a key component of the US national AI infrastructure is critical for research and education, and for Red Hat and the AI Alliance.

The National AI Initiative Act of 2020 established the NAIRR Task Force, a federal advisory committee with the function of investigating the feasibility and advisability of establishing and sustaining a National AI Research Resource and proposing a roadmap and implementation plan detailing how the resource should be established and sustained. In January 2023, the Task Force released a detailed report[2] on its findings and recommendations, explicitly stating, "The NAIRR Operating Entity and resource providers should adopt the principle of open source for products developed with federal funds." More specifically, in their recommendations to Congress, the report authors strove "to encourage principles of open source, including by encouraging software developed for the administration of the NAIRR or using resources of the NAIRR to be open source software." While not all contributors to NAIRR are open source software providers, it is clear that open source computing environments and AI assets will be central to its vision for a shared national research infrastructure for responsible discovery and innovation in AI.

OPEN SOURCE AI: SOME HISTORY

Open source software such as Linux, HTTP servers, browsers, Python, Ansible, and Kubernetes were already ubiquitous in both academic and industry projects when the most recent boom in the AI development cycle occurred. In contrast to traditional application development, successful AI model creation relies on more than code running at the highest levels of the software stack. For example, the neural network supporting a single AI application requires a complex algorithm implementation, training datasets, APIs to interface the application with the model inputs and outputs, model parameter tuning, testing and validation, and perhaps substantial inference work or even specialized processing hardware to be in place before the first line of application code can interact with a model specialized for the application's purpose.

During the decade before ChatGPT burst on the scene in November 2022, there were already many popular open source options to support data scientists' need for predictive AI tasks such as classification and regression. Machine learning (ML) frameworks, such as TensorFlow, PyTorch, Keras, and Scikit-learn—and platforms to support the ML lifecycle, such as MLflow and Kubeflow—were already in use. However, early large language models (LLMs), such as GPT from OpenAI and some of the early foundation models, were not all made openly available. Except for the interfaces to provide input and get output, details of how the models were implemented and trained were simply not shared publicly.

> Establishing an open source ecosystem as a key component of the US national AI infrastructure is critical for research and education, and for Red Hat and the AI Alliance.

Currently, there are at least two recognized levels of openness[3,4] for foundation models:

Open-weight foundation models are AI models where the pretrained weights and the code needed to run them are publicly released under a permissive license. This allows anyone with a technical background to use, modify, study, and share the model, often with the help of a model card for documentation. Examples of such models include GPT-OSS models (OpenAI), Llama Series (Meta), Gemma 3 (Google), Granite 4 (IBM), Qwen 3 series (Alibaba), and Mixtral (Mistral).

Open-science foundation models share all artifacts needed for end-to-end transparency, reproducibility, and collaboration, and they empower the community to inspect models throughout their lifecycle. This is the gold standard for completeness and openness rooted in scientific principles. Examples of open-science models are OLMo (AI2) and StarCoder (ServiceNow and HuggingFace).

OPEN SOURCE AI: COMMON MISCONCEPTIONS

Open source AI will lag behind proprietary systems in capability. Proprietary foundation models from commercial vendors (e.g., GPT-5 from OpenAI or Gemini from Google), sometimes called *frontier models*, tend to be the largest, consisting of more than a trillion parameters trained on massive amounts of data to perform a wide variety of tasks, from language processing to image generation and coding. Since these models are intended for general usage, much smaller open source models can match their performance or even do better for more limited specific tasks. In fact, domain-specific smaller models may actually be better suited for business, where development cycles often advance more quickly. Protecting business-sensitive data is also less complex and potentially faster when developing with an in-house open source model compared to using a proprietary model made available through a hosted service outside a business firewall. Open source AI models can develop very quickly if many businesses contribute even a portion of what they develop for in-house solutions back to the community.

Open source AI is dangerous and closed models are safe. The safety of current AI models is tied to training data and the inherently non-deterministic outputs from ML models. Hallucinations are equally possible with both closed source and open source models. The difference is the open-science foundation models mentioned earlier provide the complete information necessary to make the appropriate risk assessments depending on intended usage. Even the open-weight models allow the evaluation of their performance against specific safety risks and allow retraining for safer behavior. These risk assessments and mitigations can then be shared with the community. By contrast, closed-model evaluations cannot, making it impossible to determine which kind of model is safer.

Enterprises will be slow to adopt open source AI until the landscape stabilizes from rapid changes in technology, legal, and policy issues.
Evidence does not bear this out. According to IDC Market Research's 2024 analysis[5] of open source adoption in the United States: "Open GenAI models represent more than half of currently deployed GenAI models, and organizations plan to use open models for more than 60% of GenAI use cases. Almost 30% of respondents plan to use open models for all GenAI use cases." Typical reasons reported for adopting open models include faster access to innovation, cost effectiveness, transparency, and the ability to modify the model.

THE NAIRR PILOT

The NAIRR pilot is a proof-of-concept for the eventual full-scale NAIRR, bringing together computational, data, software, model, training, and user-support resources to demonstrate and investigate major elements of the vision in the NAIRR Task Force report. Led by the NSF in partnership with other federal agencies and non-governmental partners, the pilot makes available government-funded industry, and other contributed resources in support of the nation's research and education community. The pilot, begun January 24, 2024, runs for two years. Visit the NAIRR demonstration site to learn about ongoing NAIRR Pilot projects.

MOC-IBM-Red Hat collaboration

In August 2025, the NAIRR Pilot program launched a new track called Deep Partnerships to encourage collaboration between researchers and industry partners providing NAIRR Pilot resources. The AI Alliance brought together three of its members, the Mass Open Cloud, Red Hat, and IBM Research to participate in the NAIRR Pilot Deep Partnership program.

The AI Alliance was founded to foster an open community, enabling developers and researchers to accelerate responsible innovation in AI while ensuring scientific rigor, trust, safety, security, diversity and economic competitiveness. Consistent with this mission, these AI Alliance members will provide computing resources and open source AI assets to NAIRR Pilot participants to advance science and education and promote open collaboration in developing and deploying AI in society.

> The plan includes an explicit recommendation to "encourage open source and open-weight AI," highlighting the benefits of open source AI to academic research, startups, businesses, and government.

Selected projects would gain access to three key technology elements. The Mass Open Cloud provides a datacenter infrastructure with facilitation support for users and projects, along with integration and development support for those who are new to AI/ML and Kubernetes-style resource management. All operations software is open source, so experimenters can access even the lowest levels of the software stack as needed. (Get more details on MOC resources and policies at red.ht/NAIRRpilot_MOC.)

The software stack includes Red Hat Enterprise Linux (RHEL) and OpenShift AI for enterprise application development, Automated Cluster Management, and some open software specially developed for the MOC. This environment provides tools that support the full lifecycle of AI/ML experiments and models and help NAIRR investigators to build, train, test, and deploy models optimized for hybrid cloud environments. In addition, the entire portfolio of open source models and tools from IBM Research is available for use in the NAIRR Pilots.

The first selected projects run through July 1, 2026:

- ***Adaptive KV cache compression for agentic AI***, PI: Mohammad Mohammadi Amiri, Rensselaer Polytechnic Institute
- ***Building reliability and transaction semantics for LLM agents***, PI: Indranil Gupta, University of Illinois at Urbana-Champaign
- ***Efficient memory offloading for cost- and energy-efficient foundation model training***, PI: Nam Sung Kim, University of Illinois at Urbana-Champaign
- ***Evaluating and improving applications of large language models to automated software testing***, PI: Alessandro Orso, University of Georgia
- ***CLEARBOX: interpreting and improving multimodal LLMs***, PI: Deepti Ghadiyaram, Boston University
- ***Model merging for code LLMs: reasoning fusion and MoE-aware methods***, PI: Stacy Patterson, Rensselaer Polytechnic Institute
- ***Multimodal semantic routing for vLLM***, PI: Junchen Jiang, University of Chicago
- ***Time series data agent: an agentic system with foundation models for multimodal data***, PI: Agung Julius, Rensselaer Polytechnic Institute

CONTINUED SUPPORT FOR OPEN SOURCE AI

In its recently released AI Action Plan[6], the White House expanded on previous calls for open source development. The plan includes an explicit recommendation to "encourage open source and open-weight AI," highlighting the benefits of open source AI to academic research, startups, businesses, and government. In addition, the plan also recommends that the United States "build the foundations for a lean and sustainable NAIRR operations capability that can connect an increasing number of researchers and educators across the country to critical AI resources."

We believe that such a strategy will expand the opportunities for innovation while building the skills and tools needed to produce economic and social benefits across our society.

Footnotes

1. N. Maslej, et al., "The AI Index 2025 Annual Report," *AI Index Steering Committee, Institute for Human-Centered AI*, Stanford University, Stanford, CA, April 2025.

2. "Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem- An Implementation Plan for a National Artificial Intelligence Research Resource," *National Artificial Intelligence Research Resource Task Force Report*, January 2023.

3. "Defining Open Source AI: The Road Ahead," *AI Alliance Blog*, April 2025.

4. M. White, et al. "The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence," arXiv preprint arXiv:2403.13784 (2024).

5. IDC Market Research "Open GenAI Models, 2024: Benefits, Experimentation, and Deployment," Document US52477724, Aug. 7, 2024.

6. "Winning the Race: America's AI Action Plan,"Office of the Science & Technology Policy of the President of the United States, July 2025.