A guide to Models-as-a-Service

From AI chaos to control

AI adoption is growing, but infrastructure and access issues create challenges

Interest in AI is rapidly expanding, with organizations eager to use large language models (LLMs), predictive analytics, vision capabilities, and other advanced tools to extract business value. However, moving AI from isolated experimentation to widespread organizational adoption presents significant infrastructure and operational challenges.

What is Models-as-a-Service

Models-as-a-Service (MaaS) is an approach to delivering AI models as shared resources, allowing users within an organization to access them on demand. MaaS offers a ready-to-go AI foundation—in the form of application programming interface (API) endpoints—that encourages private and efficient AI at scale.

Many organizations begin their AI journey by connecting to commercial LLM application programming interfaces (APIs) such as those from OpenAI or Anthropic, assuming it is the fastest way to production. But as use grows, costs increase, and teams encounter limitations around data privacy, observability, and customization. And in some cases, commercial AI providers make changes to models with little advance warning, disrupting organizations' business uses.

In response, some organizations swing to the opposite extreme: building their own model infrastructure from scratch. This do-it-yourself path often leads to teams independently deploying open source models such as Llama or Mistral with little coordination. The result is a fragmented landscape where groups stand up their own stacks, leading to redundant infrastructure, idle graphics processing units (GPUs), and significant operational overhead. Security and governance suffer, and costs spiral without delivering much business value.

These challenges have been even further exacerbated by the ballooning size of recent LLMs such as Llama, DeepSeek, Mistral, or Qwen. Unlike the relatively small-scale AI models of even just a few years ago, today's large models can require terabytes of vRAM. And those GPUs are expensive. Using these resources inefficiently can lead quickly to soaring costs. The situation worsens when multiple teams within the same organization independently attempt to deploy these models. This fragmented approach compounds operational overhead and inflates expenditure.

Organizations need an internal approach that streamlines and consolidates model use, optimizes hardware resources, and allows for controlled, scalable access for diverse sets of internal users. Without such an approach, AI initiatives risk low adoption and high operational expenses, infrastructure investments remain underused, and measurable outcomes—such as increased productivity, lower operational costs, or faster time to insights—remain difficult to achieve.

The Models-as-a-Service approach to this challenge

Models-as-a-Service (MaaS) is an approach that helps organizations deploy AI models once and deliver them as shared, security-focused resources across the entire enterprise. Instead of managing isolated deployments for individual teams, a MaaS approach helps companies to centralize AI infrastructure and operations, which simplifies internal AI adoption.

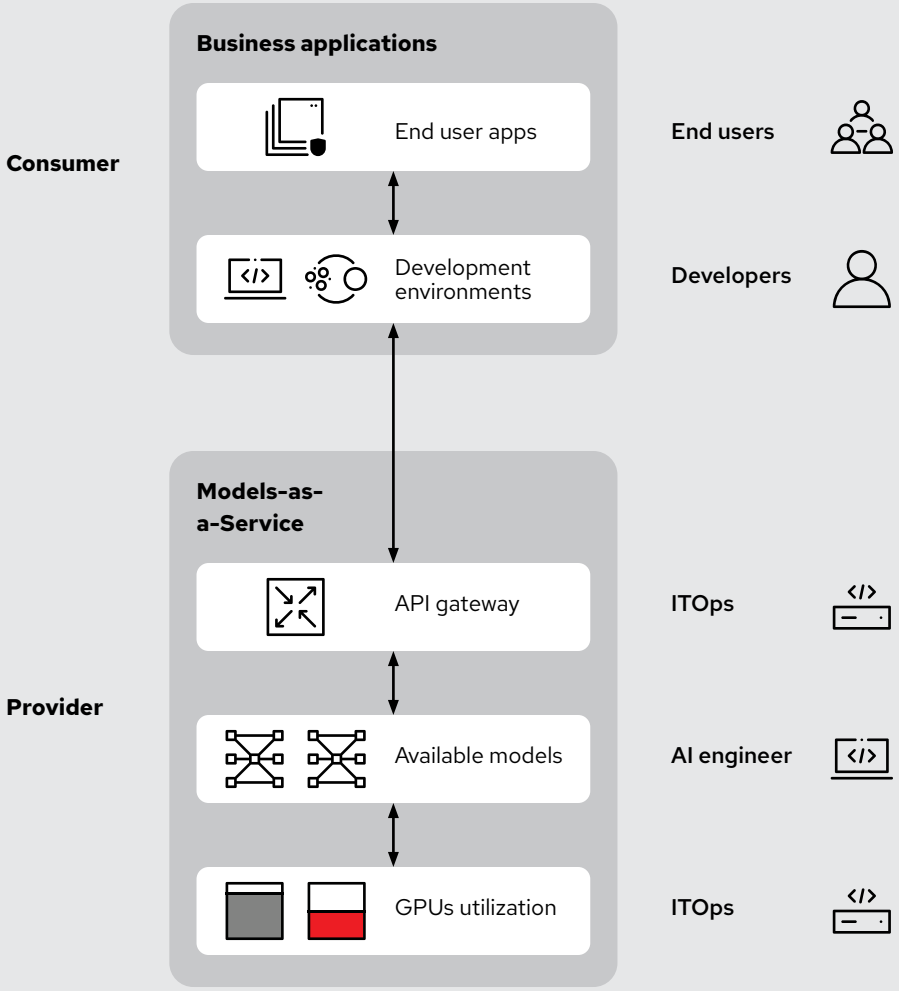

Figure 1. The workflow for a Models-as-a-Service setup.

Deliver shared access to AI with centralized model operations

- For AI engineers, MaaS provides quicker access to high-performing models via APIs, which eliminate the need to download models, manage dependencies, or request GPU allocations through lengthy IT tickets.

MaaS functions by setting up an AI operations team as the central owner of shared AI resources. Models are deployed on a scalable platform (such as Red Hat® OpenShift® AI or other similar platforms) and then exposed through an API gateway. This setup allows multiple users, developers, and business units to offer simplified access for end users while meeting security and governance priorities for IT and finance teams. This prioritization can include chargeback capabilities, consuming models without needing direct hardware access or deep technical expertise. The goal is to provide user-friendly access to the AI models and not to the required resources to run these models, such as GPUs and tensor processing units (TPUs). All this, while meeting enterprise performance and compliance requirements and without complicating access for end users.

In practice, users interact only with APIs that deliver model-generated responses. Just as public AI providers abstract away hardware complexities from end users, internal MaaS deployments offer the same simplicity. Users do not directly manage hardware or software infrastructure, wait for an IT ticket to be resolved on their behalf, or stand by while an environment is configured for them. Instead, IT operations and AI teams centrally manage model lifecycle, security, updates, and infrastructure scaling, offering users streamlined yet controlled access.

This centralization not only streamlines internal AI operations but also enhances security focus and governance. Access to AI models is tightly controlled through credential management via an API gateway. Organizations can readily track use, set up internal chargeback mechanisms, make sure privacy compliance guidelines are being followed, and establish clear operational boundaries, which makes enterprise AI both manageable and practical. Tracking usage at the token level (in and out) is the most accurate and granular way to do so, and much more precise than any GPU-level metric.

Control use, throttle access, and manage costs

- IT and platform engineers benefit from centralized oversight, which prevents unauthorized model deployments, enforces security and compliance standards, and simplifies lifecycle and infrastructure management.
- For finance teams, centralized use tracking and internal chargeback mechanisms reduce waste and make GPU use more predictable and accountable, avoiding overspending from underused, team-specific hardware allocations.

Control in a MaaS is primarily delivered through integrating an API gateway with the AI infrastructure, which allows teams to manage and monitor AI use at a very granular level.

Traditional AI deployments often suffer from unmanaged or inefficient use, as individuals or teams independently deploy models without centralized oversight. This fragmented approach can lead to costly inefficiencies, with GPU resources idling or underused. Placing an API gateway at the heart of the AI infrastructure creates a controlled access point between users and models.

This setup facilitates precise use tracking, down to the individual token level. Teams can clearly identify how much each user, team, or application consumes, attributing GPU and infrastructure costs accurately. For example, organizations can determine whether a particular user or application is using resources excessively and take corrective action—such as throttling use or allocating costs through internal chargeback mechanisms.

Throttling capabilities provided by the API gateway make sure there is consistent performance and prevent resource exhaustion. Use throttling allows IT teams to manage access intensity, preventing any single user from monopolizing GPU resources or degrading the performance experienced by others.

Additionally, API gateways offer fine-grained credential management and access control. Internal users can generate credentials to access AI models independently, streamlining administrative overhead. Credentials can also be revoked or modified in less time to respond to changing security requirements or use patterns.

This all means that cost management becomes more transparent and accountable. IT teams can allocate GPU and infrastructure expenses accurately to the teams or business units that consume them.

Support any model, any accelerator, and any cloud

A core tenet of the MaaS approach is control. It allows organizations to select and deploy a broad range of AI models, choose their preferred hardware accelerators, and operate within their existing cloud or on-premise environments. This approach gives organizations the freedom to implement AI precisely according to their technical needs, security requirements, and operational preferences.

- **Organizations face rigid limitations when adopting AI. They are often:**
  - Restricted by specific cloud services.
  - Locked into proprietary model ecosystems.
  - Constrained by fixed hardware infrastructures.
- **MaaS addresses these limitations in a number of ways, including:**
  - Supporting open source or proprietary models, custom-trained models, and popular LLMs such as Llama and Mistral.
  - Extending beyond text-based models to include predictive analytics, computer vision, audio transcription tools, and other multimodal gen AI use cases like image or video generation.
- **MaaS remains agnostic to hardware accelerators, so:**
  - Organizations can select GPUs or other accelerators that align with their workloads, cost structures, and performance needs.
  - Centralized AI teams can make critical sizing and deployment decisions, improving efficiency and reducing errors from less technical users.
- **Centralized management allows:**
  - Optimal allocation and use of infrastructure.
  - Reduced operational overhead and prevention of resource misconfiguration.
- **MaaS supports deployment across any environment, including:**
  - On-premise, hybrid cloud, air-gapped environments, and public clouds, which is especially valuable for highly regulated sectors that require data sovereignty, regulatory compliance, or strict security controls.

How Red Hat implements MaaS

Red Hat has embraced MaaS internally by centralizing AI model deployment and access. Our internal AI team centrally manages AI resources and model operations, using Red Hat OpenShift and Red Hat OpenShift AI as the underlying platform. This centralized model deployment simplifies AI consumption for users across the organization, allowing our developers and business teams to efficiently integrate AI capabilities into their workflows without needing dedicated hardware or deep technical expertise.

Our implementation features a scalable serving architecture that uses GPUs within OpenShift AI, and connects users through a centralized API gateway. This gives controlled, security-focused, and traceable access to AI models. Use is carefully managed through token-based monitoring, facilitating precise tracking of who is using models, how often, and in what quantity. The result is optimized hardware use, reducing the unnecessary consumption of GPU resources, and offering detailed insights to accurately allocate costs across different internal teams or projects.

Our MaaS implementation uses GitOps workflows, providing high availability and reliability. This operational approach reduces manual intervention and potential errors, establishing clear control over AI deployments.

A key benefit of our internal MaaS implementation has been a marked improvement in resource efficiency and user experience. Rather than multiple teams independently provisioning GPUs and deploying models, our MaaS has eliminated duplicate efforts, streamlined internal operations, and significantly accelerated time-to-value. When new models are tested and verified, Red Hat teams can integrate and use them immediately, instead of being delayed by hardware allocation or provisioning tasks.

Start building your internal AI platform today

Ready to simplify AI delivery and unlock real value from your infrastructure investments? Start by reviewing our in-depth explainer on MaaS for further insight into how it works. Then, explore the OpenShift AI product page to evaluate platform capabilities and GPU use guidance.

For teams building a MaaS internally, Red Hat Consulting helps organizations design and operationalize model-serving environments tailored to their needs. Learn more at the Red Hat Consulting for AI page.

Want a more comprehensive look into real-world examples? Check out our on-demand webinar series, including the session dedicated to MaaS.

North America
1 888 REDHAT1
www.redhat.com

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
twitter.com/RedHat
linkedin.com/company/red-hat

redhat.com