

Microsoft Build

May 21–23, 2024

Register now >

≡ Learn



🔍 Sign in

[Learn](#) / [Azure](#) / [Machine Learning](#) /



Endpoints for inference in production

Article • 11/16/2023 • 25 contributors

[Feedback](#)

In this article

[Intuition](#)

[Endpoints and deployments](#)

[Online and batch endpoints](#)

[Developer interfaces](#)

[Next steps](#)

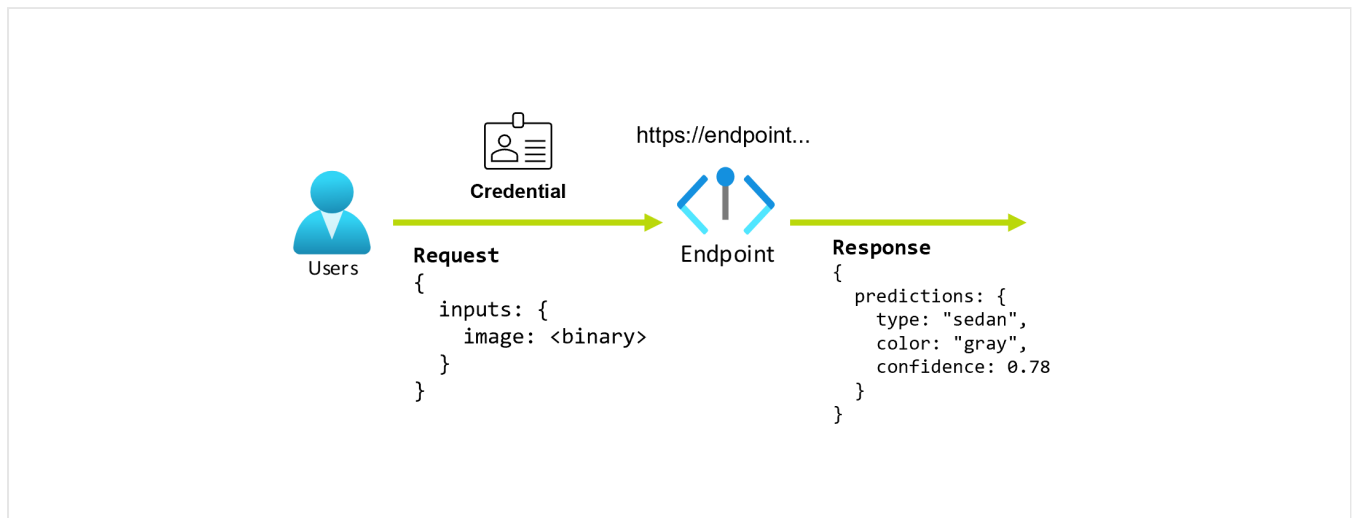
APPLIES TO: [Azure CLI ml extension v2 \(current\)](#) [Python SDK azure-ai-ml v2 \(current\)](#) [🔗](#)

After you train machine learning models or pipelines, you need to deploy them to production so that others can use them for *inference*. Inference is the process of applying new input data to the machine learning model or pipeline to generate outputs. While these outputs are typically referred to as "predictions," inferencing can be used to generate outputs for other machine learning tasks, such as classification and clustering. In Azure Machine Learning, you perform inferencing by using **endpoints and deployments**. Endpoints and deployments allow you to decouple the interface of your production workload from the implementation that serves it.

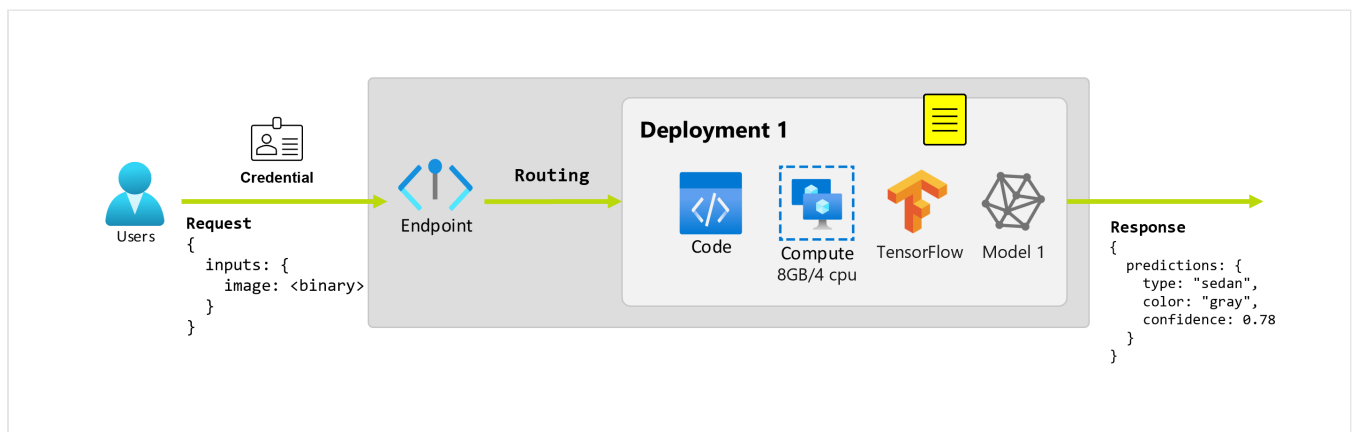
Intuition

Suppose you're working on an application that predicts the type and color of a car, given its photo. For this application, a user with certain credentials makes an HTTP request to a URL and

provides a picture of a car as part of the request. In return, the user gets a response that includes the type and color of the car as string values. In this scenario, the URL serves as an endpoint.

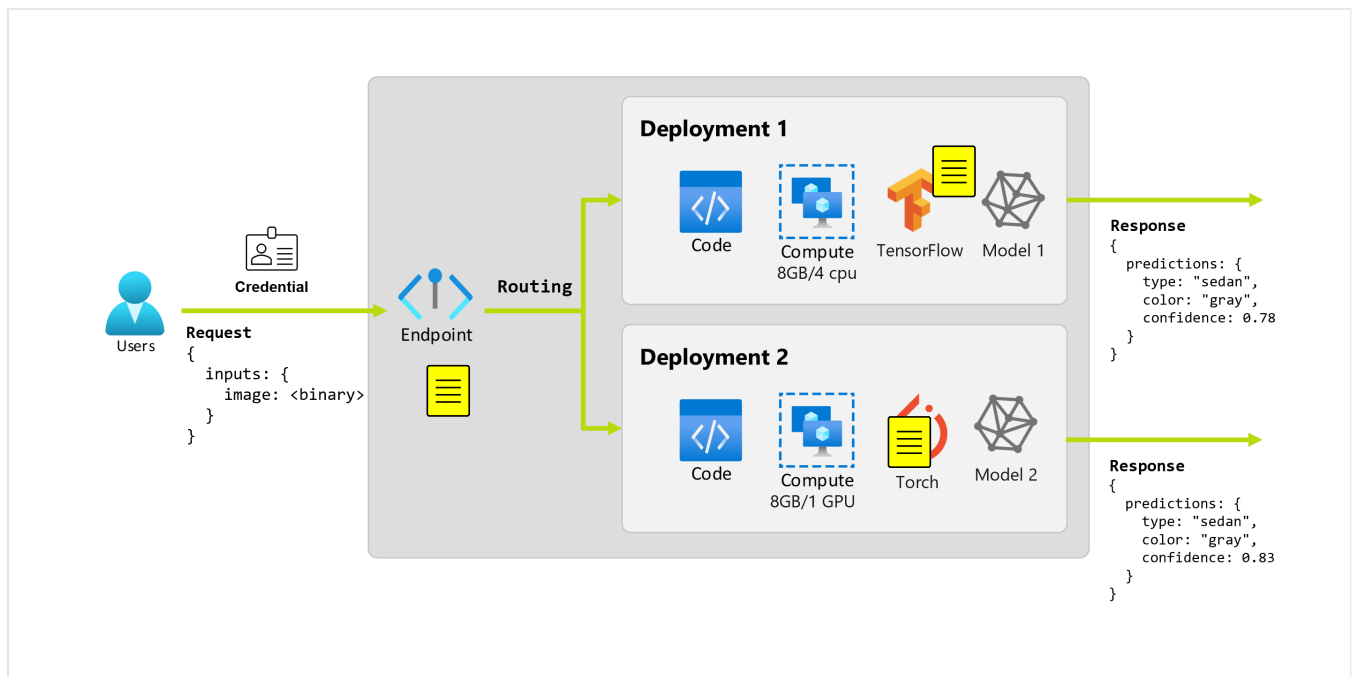


Furthermore, say that a data scientist, Alice, is working on implementing the application. Alice knows a lot about TensorFlow and decides to implement the model using a Keras sequential classifier with a RestNet architecture from the TensorFlow Hub. After testing the model, Alice is happy with its results and decides to use the model to solve the car prediction problem. The model is large in size and requires 8 GB of memory with 4 cores to run. In this scenario, Alice's model and the resources, such as the code and the compute, that are required to run the model make up a deployment under the endpoint.



Finally, let's imagine that after a couple of months, the organization discovers that the application performs poorly on images with less than ideal illumination conditions. Bob, another data scientist, knows a lot about data augmentation techniques that help a model build robustness on that factor. However, Bob feels more comfortable using Torch to implement the model and trains a new model with Torch. Bob wants to try this model in production gradually until the organization is ready to retire the old model. The new model also shows better performance when deployed to GPU, so the deployment needs to include a GPU. In this scenario, Bob's model and the resources, such as the code and the compute, that

are required to run the model make up **another deployment under the same endpoint.**



Endpoints and deployments

An **endpoint** is a stable and durable URL that can be used to request or invoke a model. You provide the required inputs to the endpoint and get the outputs back. An endpoint provides:

- a stable and durable URL (like *endpoint-name.region.inference.ml.azure.com*),
- an authentication mechanism, and
- an authorization mechanism.

A **deployment** is a set of resources and computes required for hosting the model or component that does the actual inferencing. A single endpoint can contain multiple deployments. These deployments can host independent assets and consume different resources based on the needs of the assets. Endpoints have a routing mechanism that can direct requests to specific deployments in the endpoint.

To function properly, **each endpoint must have at least one deployment.** Endpoints and deployments are independent Azure Resource Manager resources that appear in the Azure portal.

Online and batch endpoints

Azure Machine Learning allows you to implement **online endpoints** and **batch endpoints**.

Online endpoints are designed for real-time inference—when you invoke the endpoint, the results are returned in the endpoint's response. **Batch endpoints**, on the other hand, are

designed for long-running batch inference. Each time you invoke a batch endpoint you generate a batch job that performs the actual work.

When to use online vs batch endpoint for your use-case

Use **online endpoints** to operationalize models for real-time inference in synchronous low-latency requests. We recommend using them when:

- ✓ You have low-latency requirements.
- ✓ Your model can answer the request in a relatively short amount of time.
- ✓ Your model's inputs fit on the HTTP payload of the request.
- ✓ You need to scale up in terms of number of requests.

Use **batch endpoints** to operationalize models or pipelines for long-running asynchronous inference. We recommend using them when:

- ✓ You have expensive models or pipelines that require a longer time to run.
- ✓ You want to operationalize machine learning pipelines and reuse components.
- ✓ You need to perform inference over large amounts of data that are distributed in multiple files.
- ✓ You don't have low latency requirements.
- ✓ Your model's inputs are stored in a storage account or in an Azure Machine Learning data asset.
- ✓ You can take advantage of parallelization.

Comparison of online and batch endpoints

Both online and batch endpoints are based on the idea of endpoints and deployments, which help you transition easily from one to the other. However, when moving from one to another, there are some differences that are important to take into account. Some of these differences are due to the nature of the work:

Endpoints

The following table shows a summary of the different features available to online and batch endpoints.

 Expand table

Feature	Online Endpoints	Batch endpoints
Stable invocation URL	Yes	Yes
Support for multiple deployments	Yes	Yes
Deployment's routing	Traffic split	Switch to default
Mirror traffic for safe rollout	Yes	No
Swagger support	Yes	No
Authentication	Key and token	Microsoft Entra ID
Private network support	Yes	Yes
Managed network isolation	Yes	Yes (see required additional configuration)
Customer-managed keys	Yes	Yes
Cost basis	None	None

Deployments

The following table shows a summary of the different features available to online and batch endpoints at the deployment level. These concepts apply to each deployment under the endpoint.

 Expand table

Feature	Online Endpoints	Batch endpoints
Deployment types	Models	Models and Pipeline components
MLflow model deployment	Yes	Yes
Custom model deployment	Yes, with scoring script	Yes, with scoring script
Model package deployment ¹	Yes (preview)	No
Inference server ²	<ul style="list-style-type: none"> - Azure Machine Learning Inferencing Server - Triton - Custom (using BYOC) 	Batch Inference

Feature	Online Endpoints	Batch endpoints
Compute resource consumed	Instances or granular resources	Cluster instances
Compute type	Managed compute and Kubernetes	Managed compute and Kubernetes
Low-priority compute	No	Yes
Scaling compute to zero	No	Yes
Autoscaling compute ³	Yes, based on resources' load	Yes, based on job count
Overcapacity management	Throttling	Queuing
Cost basis ⁴	Per deployment: compute instances running	Per job: compute instanced consumed in the job (capped to the maximum number of instances of the cluster).
Local testing of deployments	Yes	No

¹ Deploying MLflow models to endpoints without outbound internet connectivity or private networks requires [packaging the model](#) first.

² *Inference server* refers to the serving technology that takes requests, processes them, and creates responses. The inference server also dictates the format of the input and the expected outputs.

³ *Autoscaling* is the ability to dynamically scale up or scale down the deployment's allocated resources based on its load. Online and batch deployments use different strategies for autoscaling. While online deployments scale up and down based on the resource utilization (like CPU, memory, requests, etc.), batch endpoints scale up or down based on the number of jobs created.

⁴ Both online and batch deployments charge by the resources consumed. In online deployments, resources are provisioned at deployment time. However, in batch deployment, no resources are consumed at deployment time but when the job runs. Hence, there is no cost associated with the deployment itself. Notice that queued jobs do not consume resources either.

Developer interfaces

Endpoints are designed to help organizations operationalize production-level workloads in Azure Machine Learning. Endpoints are robust and scalable resources and they provide the best of the capabilities to implement MLOps workflows.

You can create and manage batch and online endpoints with multiple developer tools:

- The Azure CLI and the Python SDK
- Azure Resource Manager/REST API
- Azure Machine Learning studio web portal
- Azure portal (IT/Admin)
- Support for CI/CD MLOps pipelines using the Azure CLI interface & REST/ARM interfaces

Next steps

- [How to deploy online endpoints with the Azure CLI and Python SDK](#)
- [How to deploy models with batch endpoints](#)
- [How to deploy pipelines with batch endpoints](#)
- [How to use online endpoints with the studio](#)
- [How to monitor managed online endpoints](#)
- [Manage and increase quotas for resources with Azure Machine Learning](#)

Feedback

① Coming soon: Throughout 2024 we will be phasing out GitHub Issues as the feedback mechanism for content and replacing it with a new feedback system. For more information see: <https://aka.ms/ContentUserFeedback>.

Submit and view feedback for

This product

🔄 This page

[🔄 View all page feedback](#)

Additional resources

Documentation

[Deploy machine learning models to online endpoints for inference - Azure Machine Learning](#)

Learn to deploy your machine learning model as an online endpoint in Azure.

[What are online endpoints? - Azure Machine Learning](#)

Learn about online endpoints for real-time inference in Azure Machine Learning.

[View costs for managed online endpoints - Azure Machine Learning](#)

Learn to how view costs for a managed online endpoint in Azure Machine Learning.

[Show 5 more](#)

Training

Learning path

[Deploy and consume models with Azure Machine Learning - Training](#)

Learn how to deploy a model to an endpoint. When you deploy a model, you can get real-time or batch predictions by calling the endpoint.


Events

[Visual Studio Code Day Skills Challenge](#)



Apr 24, 3 PM - May 18, 2 PM

Learn about AI, Data Science and more with Visual Studio Code! Register for free and earn a badge on your Microsoft Learn profile!

[Register now](#)

 English (United States)


 Your Privacy Choices

 Theme 


[Previous Versions](#)

[Blog](#) 

[Contribute](#)

[Privacy](#) 

[Terms of Use](#)

[Trademarks](#) 

© Microsoft 2024