AWS > Documentation > Amazon SageMaker > **Developer Guide**

Automatically Scale Amazon SageMaker Models

PDF (/pdfs/sagemaker/latest/dg/sagemaker-dg.pdf#endpoint-auto-scaling) | **RSS (amazon-sagemaker-release-notes.rss)**

Amazon SageMaker supports automatic scaling (auto scaling) for your hosted models. *Auto scaling* dynamically adjusts the number of instances provisioned for a model in response to changes in your workload. When the workload increases, auto scaling brings more instances online. When the workload decreases, auto scaling removes unnecessary instances so that you don't pay for provisioned instances that you aren't using.

Topics

- Prerequisites (./endpoint-auto-scaling-prerequisites.html)
- Configure model auto scaling with the console (./endpoint-auto-scaling-add-console.html)
- Register a model (./endpoint-auto-scaling-add-policy.html)
- Define a scaling policy (./endpoint-auto-scaling-add-code-define.html)
- Apply a scaling policy (./endpoint-auto-scaling-add-code-apply.html)
- Edit a scaling policy (./endpoint-auto-scaling-edit.html)
- Delete a scaling policy (./endpoint-auto-scaling-delete.html)

- Query Endpoint Auto scaling History (./endpoint-scaling-query-history.html)
- Update or delete endpoints that use automatic scaling (./endpoint-scaling.html)
- Load testing your auto scaling configuration (./endpoint-scaling-loadtest.html)
- Use AWS CloudFormation to update auto scaling policies (./endpoint-scaling-cloudformation.html)