Serverless in AI Insights

Sept 2022

Outline

- AWS serverless
- Serverless ML training
- Serverless ML inference (provide large model as a service)
- Serverless research paper highlights

Serverless services on AWS

Modern applications are built serverless-first, a strategy that prioritizes the adoption of serverless services, so you can increase agility throughout your application stack. We've developed serverless services for all three layers of your stack: compute, integration, and data stores. Consider getting started with these services:

Compute

AWS Lambda

AWS Lambda is an event-driven, pay-as-you-go compute service that lets you run code without provisioning or managing servers.

AWS Fargate

AWS Fargate is a serverless compute engine that works with Amazon Elastic Container Service (ECS) and Amazon Elastic Kubernetes Service (EKS).

Application integration

Amazon EventBridge

Amazon EventBridge is a serverless event bus that lets you build event-driven applications at scale across AWS and existing systems.

AWS Step Functions

AWS Step Functions is a visual workflow orchestrator that makes it easy to sequence multiple AWS services into business-critical applications.

Amazon SQS

Amazon Simple Queue Service (SQS) is a message queuing service enabling you to decouple and scale microservices, distributed systems, and serverless applications.

Amazon SNS

Amazon Simple Notification Service (SNS) is a fully managed messaging service for both application-to-application (A2A) and application-to-person (A2P) communication.

Amazon API Gateway

Amazon API Gateway is a fully managed service that makes it easy to create and publish APIs at any scale.

AWS AppSync

AWS AppSync is a fully managed service that accelerates application development with scalable GraphQL APIs.

Data store

Amazon S3

Amazon Simple Storage Service (Amazon S3) is an object storage service designed to store and protect any amount of data.

Amazon DynamoDB

Amazon DynamoDB is a key-value and document database service, delivering single-digit millisecond performance at any scale.

Amazon RDS Proxy

Amazon RDS Proxy is a managed database proxy for Amazon Relational Database Service (RDS) that makes applications more scalable and secure.

Amazon Aurora Serverless

Amazon Aurora Serverless is a MySQL and PostgreSQL-compatible relational database that automatically scales capacity based on your application's needs.

Use cases

Web applications | Data processing | Batch processing | Event ingestion

Reference: https://aws.amazon.com/serverless/

AWS serverless service

Amazon Redshift Serverless

Get insights from data in seconds without having to manage data warehouse infrastructure

Amazon Redshift Serverless makes it easier to run and scale analytics without having to manage your data warehouse infrastructure. Developers, data scientists, and analysts can work across databases, data warehouses, and data lakes to build reporting and dashboarding applications, perform real-time analytics, share and collaborate on data, and build and train machine learning (ML) models. Go from large amounts of data to insights in seconds. Amazon Redshift Serverless automatically provisions and intelligently scales data warehouse capacity to deliver fast performance for even the most demanding and unpredictable workloads, and you pay only for what you use. Just load data and start querying right away in Amazon Redshift Query Editor or in your favorite business intelligence (BI) tool and continue to enjoy the best price performance and familiar SQL features in an easy-to-use, zero administration environment.

Get started with Amazon Redshift Serverless (1:28)

SageMaker

Serverless Inference

PDF | RSS

Amazon SageMaker Serverless Inference is a purpose-built inference option that makes it easy for you to deploy and scale ML models. Serverless Inference is ideal for workloads which have idle periods between traffic spurts and can tolerate cold starts. Serverless endpoints automatically launch compute resources and scale them in and out depending on traffic, eliminating the need to choose instance types or manage scaling policies. This takes away the undifferentiated heavy lifting of selecting and managing servers. Serverless Inference integrates with AWS Lambda to offer you high availability, built-in fault tolerance and automatic scaling.

Reference: https://docs.aws.amazon.com/sagemaker/latest/dg/serverless-endpoints.html

AWS serverless framework (developer tools)

- AWS Serverless Application Model (AWS SAM) is an open-source framework for building serverless applications. It provides shorthand syntax to express functions, APIs, databases, and event source mappings.
- AWS Cloud Development Kit (AWS CDK) is an open source software development framework to define your cloud application resources using familiar programming languages.
- Serverless Framework - The Serverless Framework consists of an open source CLI and a hosted dashboard. Together, they provide you with full serverless application lifecycle management.
- Chalice is a framework for writing serverless apps in Python. It allows you to quickly create and deploy applications that use AWS Lambda.
- Arc.codes provides everything you need to build massively scalable serverless apps with low code, clear and terse config, and zero ceremony.
- Claudia.js makes it easy to deploy Node.js projects to AWS Lambda and API Gateway.

Reference: https://aws.amazon.com/serverless/getting-started/

Towards Demystifying Serverless Machine Learning Training

Jiawei Jiang, Shaoduo Gan, Yue Liu, Fanlin Wang, Gustavo Alonso, Ana Klimovic, Ankit Singla, Wentao Wu, Ce Zhang

ACM SIGMOD International Conference on Management of Data (SIGMOD 2021) | June 2021
View Publication

The appeal of serverless (FaaS) has triggered a growing interest on how to use it in data-intensive applications such as ETL, query processing, or machine learning (ML). Several systems exist for training large-scale ML models on top of serverless infrastructures (e.g., AWS Lambda) but with inconclusive results in terms of their performance and relative advantage over "serverful" infrastructures (IaaS). In this paper we present a systematic, comparative study of distributed ML training over FaaS and IaaS. We present a design space covering design choices such as optimization algorithms and synchronization protocols, and implement a platform, LambdaML, that enables a fair comparison between FaaS and IaaS. We present experimental results using LambdaML, and further develop an analytic model to capture cost/performance tradeoffs that must be considered when opting for a serverless infrastructure. Our results indicate that ML training pays off in serverless only for models with efficient (i.e., reduced) communication and that quickly converge. In general, FaaS can be much faster but it is never significantly cheaper than IaaS.

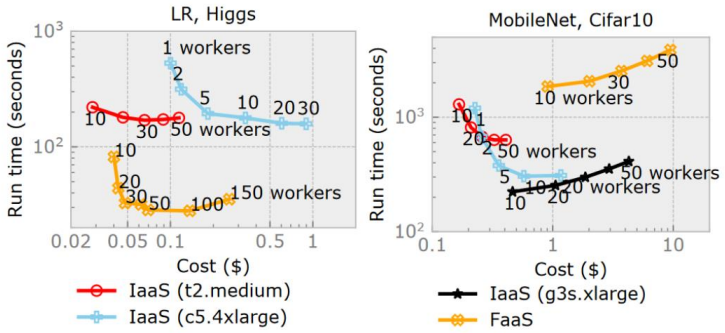


Figure 11: End-to-end comparison (w.r.t. # workers).

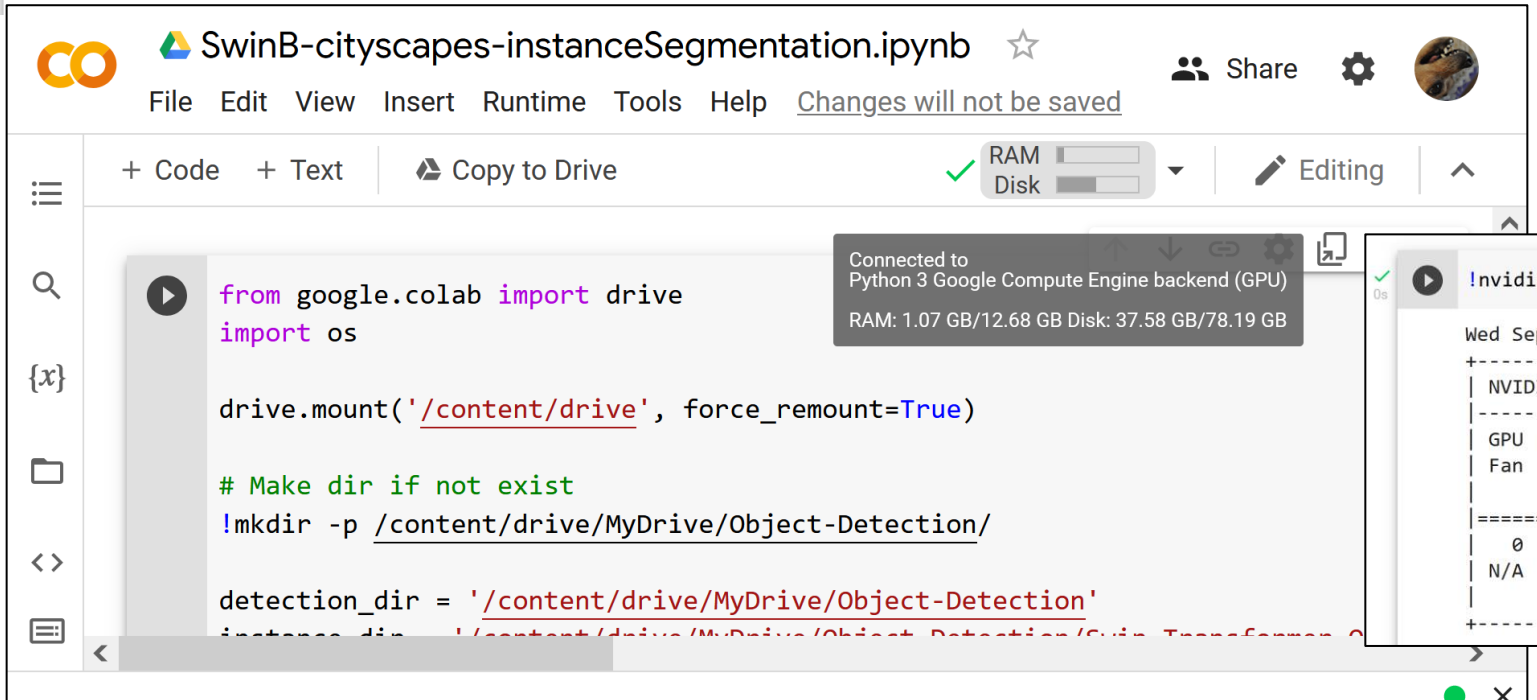
More practical Example, Google Colab

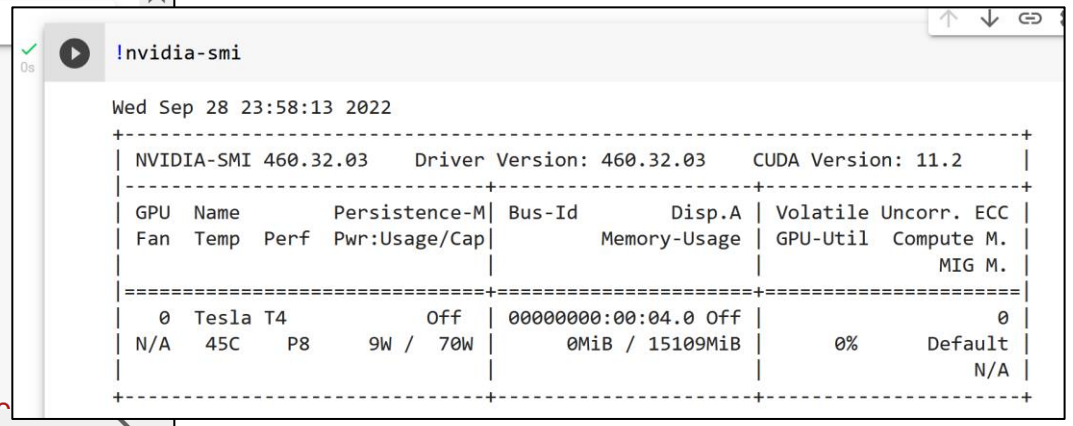
What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

Serverless: **Colab allows you to instantly spring up a Jupyter Notebook completely serverless in the browser**. That means you don't have to worry about provisioning hardware, your Python version and path or if you're on Windows, MacOS or even a phone!





One step further, ML as a service (serverless inference)

- Large language model, e.g., Nvidia NeMo API, Cohere.ai

What is NeMo LLM?

NVIDIA NeMo LLM is a service that provides a fast path to customizing and using large language models trained on several frameworks. Developers can deploy enterprise AI applications using NeMo LLM on private and public clouds.

They can also experience Megatron 530B—one of the largest language models—through the cloud API or experiment via the LLM service.

Reference:
https://www.nvidia.com/en-us/gpu-cloud/nemo-llm-service/
https://cohere.ai/
https://sambanova.ai/products/dataflow-as-a-service/

Recent serverless paper highlights

1. **Llama: A Heterogeneous & Serverless Framework for Auto-Tuning Video Analytics Pipelines (Standford)**
2. **Faa$T: A Transparent Auto-Scaling Cache for Serverless Applications (Microsoft)**
3. **Atoll: A Scalable Low-Latency Serverless Platform (UW and Google)**
4. **Kraken : Adaptive Container Provisioning for Deploying Dynamic DAG applications in Serverless platforms (PSU)**
5. **Speedo: Fast dispatch and orchestration of serverless workflows (IIT)**