AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

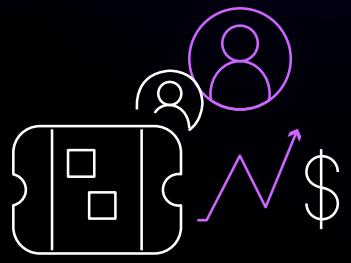
SVS313-R2

Serverless ML: Machine learning without managing machines

Veda Raman(she/her)
SA, Serverless
AWS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

State of machine learning



Increased spending

By 2023, spending on AI systems will reach $97.9B, up 2.5x from $37.5B in 2019

IDC



From piloting to operationalizing

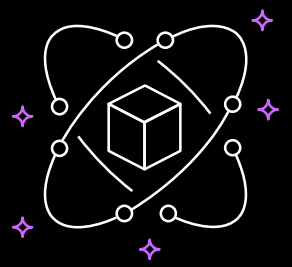
By the end of 2024, 75% of enterprises will shift from piloting to operationalizing AI

Gartner



ML handles real-world tasks

Driven by advancements in GPUs and compute, availability of data, new algorithms, and the cloud



Integrating ML into DevOps

ML is now part of mainstream DevOps process, not a set of specialized, isolated projects

Gartner

Sources:

- https://www.idc.com/getdoc.jsp?containerId=prUS46794720
- https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020
- https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why AWS Lambda for ML inference?

OPERATIONAL BURDEN

Least

What you manage

Most

	AWS manages	Customer manages	
AWS Lambda Serverless functions	Data source integrations Physical hardware, software, networking, and facilities Provisioning	Application code	
AWS Fargate Serverless containers	Container orchestration, provisioning Cluster scaling Physical hardware, host OS/kernel, networking, and facilities	Application code Data source integrations Security config and updates	Network config Management tasks
Amazon ECS/EKS Container management as a service	Container orchestration control plane Physical hardware, software, networking, and facilities	Application code Data source integrations Work clusters	Security config and updates, network config, firewall, and management tasks
Amazon EC2 Infrastructure as a service (IaaS)	Physical hardware, software, networking, and facilities	Application code Data source integrations Scaling Security config and updates	Network config Management tasks Provisioning, managing scaling and patching of servers

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why Lambda for ML inference? Burst capacity

- **3,000** – US West (Oregon), US East (N. Virginia), Europe (Ireland)
- **1,000** – Asia Pacific (Tokyo), Europe (Frankfurt), US East (Ohio)
- **500** – Other AWS Regions

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why Lambda for ML inference? Cost

- Scales to zero
- 1 millisecond billing
- No paying for patching, managing servers

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why Lambda for ML inference?
AWS Graviton2 (arm64) for Lambda

- Lambda function on 64-bit ARM architecture can provide up to 34% price performance improvement

Blog post: https://github.com/aws-samples/aws-lambda-docker-serverless-inference/tree/main/xgboost-inference-arm64-docker-lambda

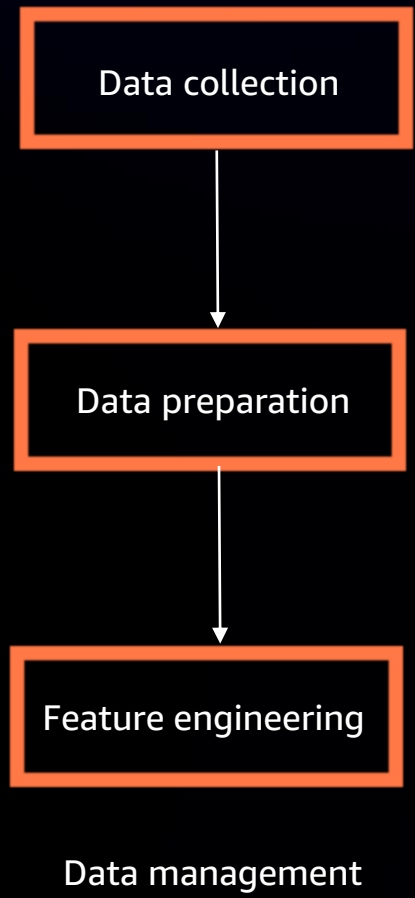
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Common ML inference use cases

- Fraud detection
- Predictive maintenance
- Demand forecasting
- Personalized recommendations
- Sentiment analysis
- Computer vision
- Churn prediction
- Natural language processing

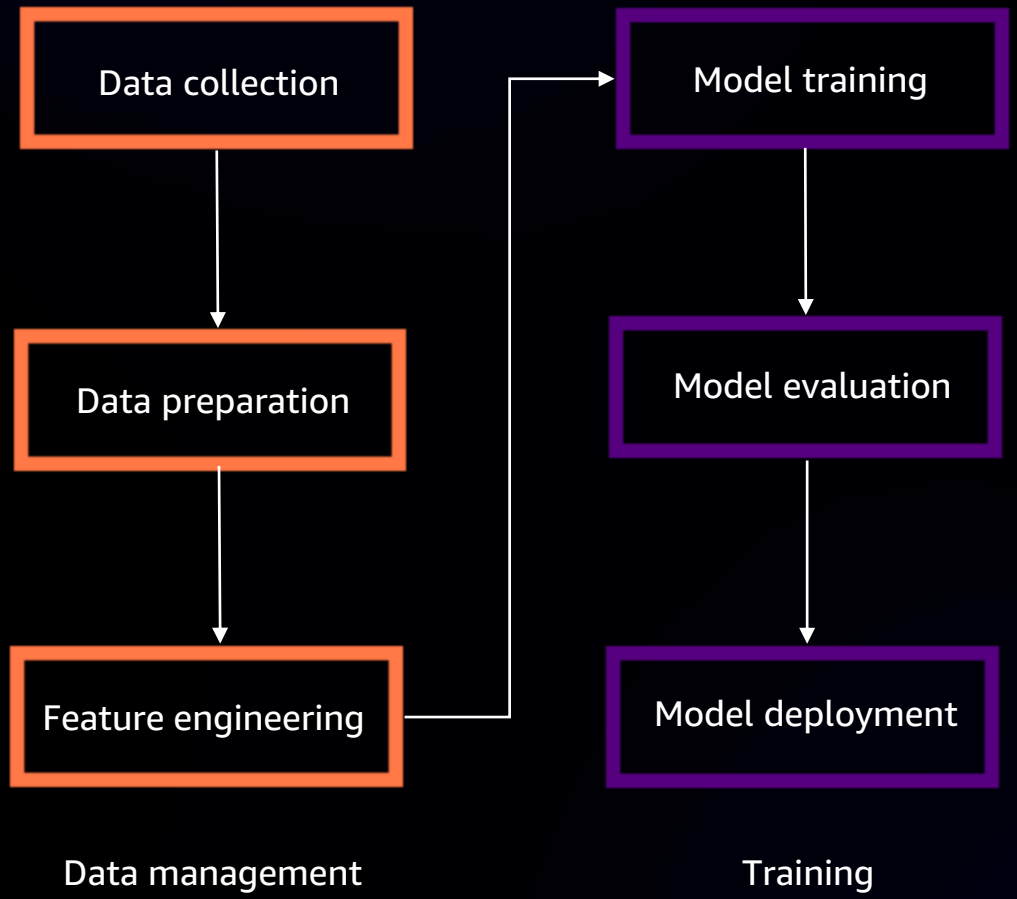
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML operations workflow



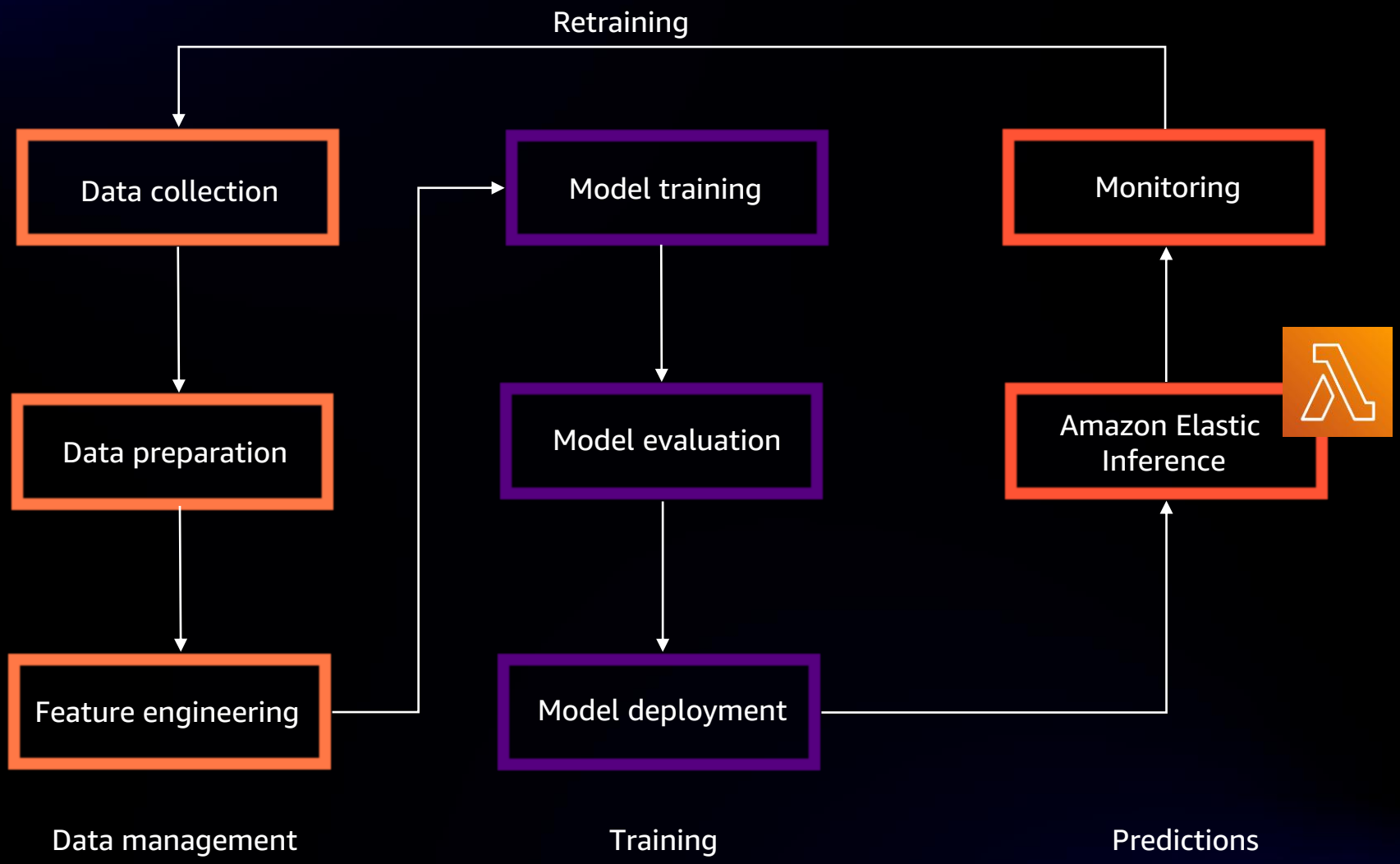
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML operations workflow



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML operations workflow



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML inference



In deep learning applications, inference drives as much as 90% of compute costs

Source: https://aws.amazon.com/blogs/machine-learning/model-serving-with-amazon-elastic-inference/

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi models versus single model



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

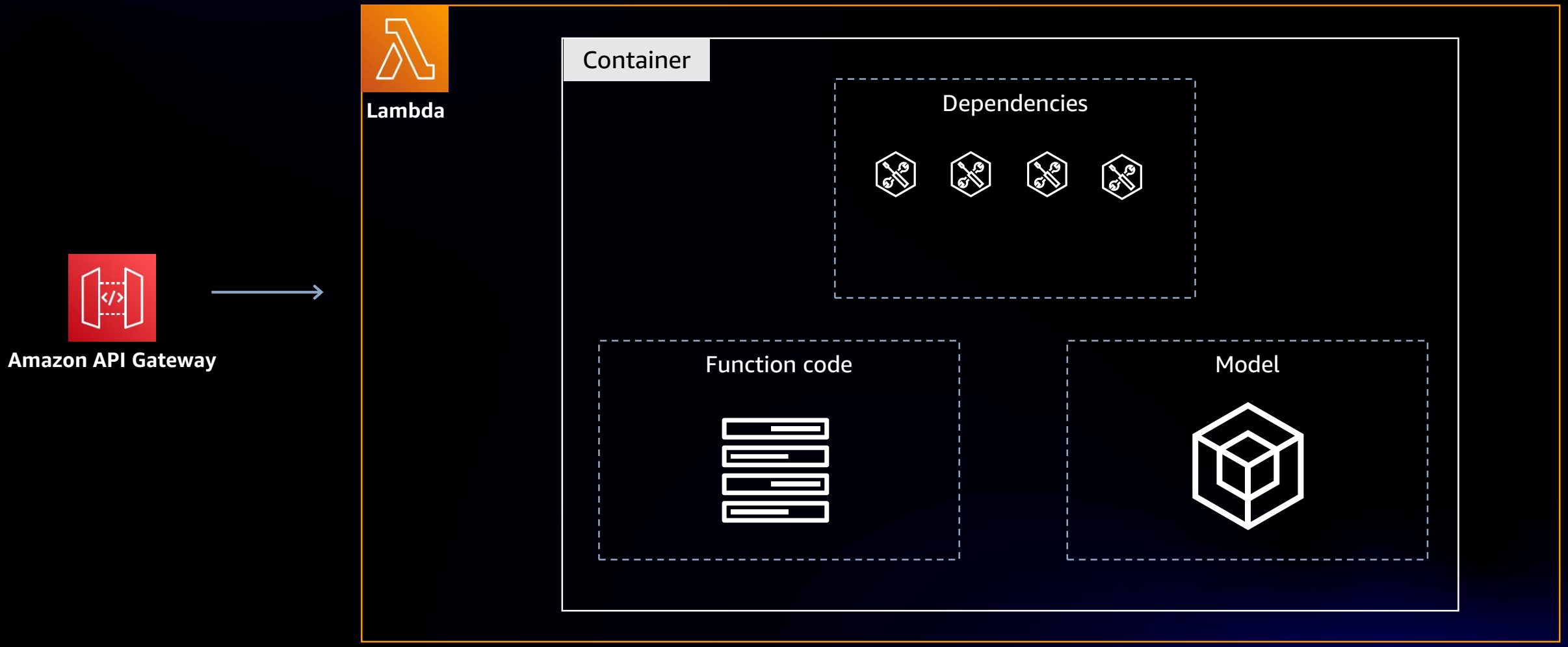
Inferencing with Lambda

- Container hosted model(s)
- Amazon S3 hosted model(s)
- Amazon EFS hosted model(s)

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Container hosted models

SIMPLE PATTERN, NO EXTERNAL DATA SOURCE



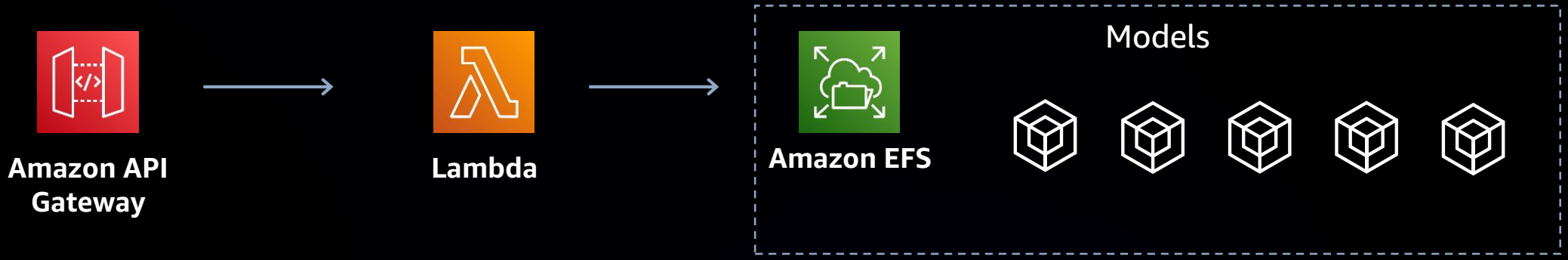
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon S3 hosted models



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EFS hosted models



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Performance testing & analysis

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Test configuration

- API Gateway fronting Lambda function
- All Lambda functions packaged as container image
- Amazon S3, Amazon EFS and container-hosted architectures tested
- 100 TPS for 20 minutes, 5G Lambda memory
- Model: BERT NLP using PyTorch – 250MB

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Single model inference

BERT NLP PyTorch	Average latency (ms)	Cold start latency (ms)	Lambda concurrency	Lambda cost (1 hour)	Data store cost (1 hour)	Lambda cost (30 days)	Data store cost (30 days)
Container hosted	104	9,776	17	$3.19	-	$248.98	-
S3 hosted	100	9,631	19	$3.07	$0.00010	$239.62	$0.07
EFS hosted	103	9,972	17	$3.16	$0.01	$246.64	$0.30

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Single model inference

OBSERVATIONS AND RECOMMENDATIONS

Container hosted

- Model included at build time
- Simplest and cheapest: No data store charges
- Recommended when models don’t change frequently

Amazon S3 hosted

- Model included during initialization
- Simplest data store option: Easy CRUD operations on model
- Recommended for most use cases

Amazon EFS hosted

- Model included during initialization
- Requires throughput monitoring
- Recommended for customers already using Amazon EFS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi model inference

BERT NLP PyTorch	Average latency (ms)	Lambda concurrency	Lambda cost (1 hour)	Data store cost (1 hour)	Lambda cost (30 days)	Data store cost (30 days)
S3	3,700	482	$111.07	$0.15	$8,663.64	$11.24
EFS (burst)	1,700	1,000 peak with timeout & throttles (183 steady)	$51.07	$0.00010	$3,983.63	$0.07
EFS (500 MB/s)	1,700	1,000 peak fewer timeouts 180 steady	$51.07	$4.17	$3,983.63	$3,000.01
EFS (1024 MB/s)	1,700	700 peak no timeouts 180 steady	$51.07	$8.53	$3,983.63	$6,144.01

Data store transactions: 120,000 per 20 minute test at 100TPS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi model inference

OBSERVATIONS AND RECOMMENDATIONS

Container hosted

- Viable when total model volume <10GB
- CI/CD complexity increases at scale
- Recommended for cost sensitive customers with infrequently changing models

Amazon S3 hosted

- Consistent performance up to 5,500 TPS
- Recommended for high/un-predictable throughput

Amazon EFS hosted

- Requires predictable throughput
- Must provision for peak throughput
- Read throughput is metered at one-third (3 Gbps max)
- Recommended for highly latency sensitive customers

No clear cost winner between S3 and EFS.
Pricing varies based on workload characteristics.

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EFS performance modes



General purpose

- Recommended for majority of cases

Max I/O: Increased I/O at the cost of latency

- Recommended for >35k IOPS

Key metric to monitor in general purpose mode

- PercentIOLimit

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EFS throughput modes



Throughput mode

- Bursting: Recommended starting point
- Provisioned: Guaranteed capacity for higher throughput

Key metric to monitor in bursting mode

- BurstCreditBalance

Key metrics to monitor in provisioned mode

- MeteredIOBytes versus PermittedThroughput

As of Feb. 5, read throughput operations are metered at one-third cost for all metrics

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Performance optimization

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Compute acceleration: AVX2

- Advanced Vector Extensions (AVX) is an extension to x86 CPU instructions that speed up floating point operations
- Can increase performance for computationally intense workloads
- ML inferencing
- High-performance computing (HPC)
- Scientific simulations
- Image and audio/video processing
- Financial analysis

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Compute acceleration: AVX2

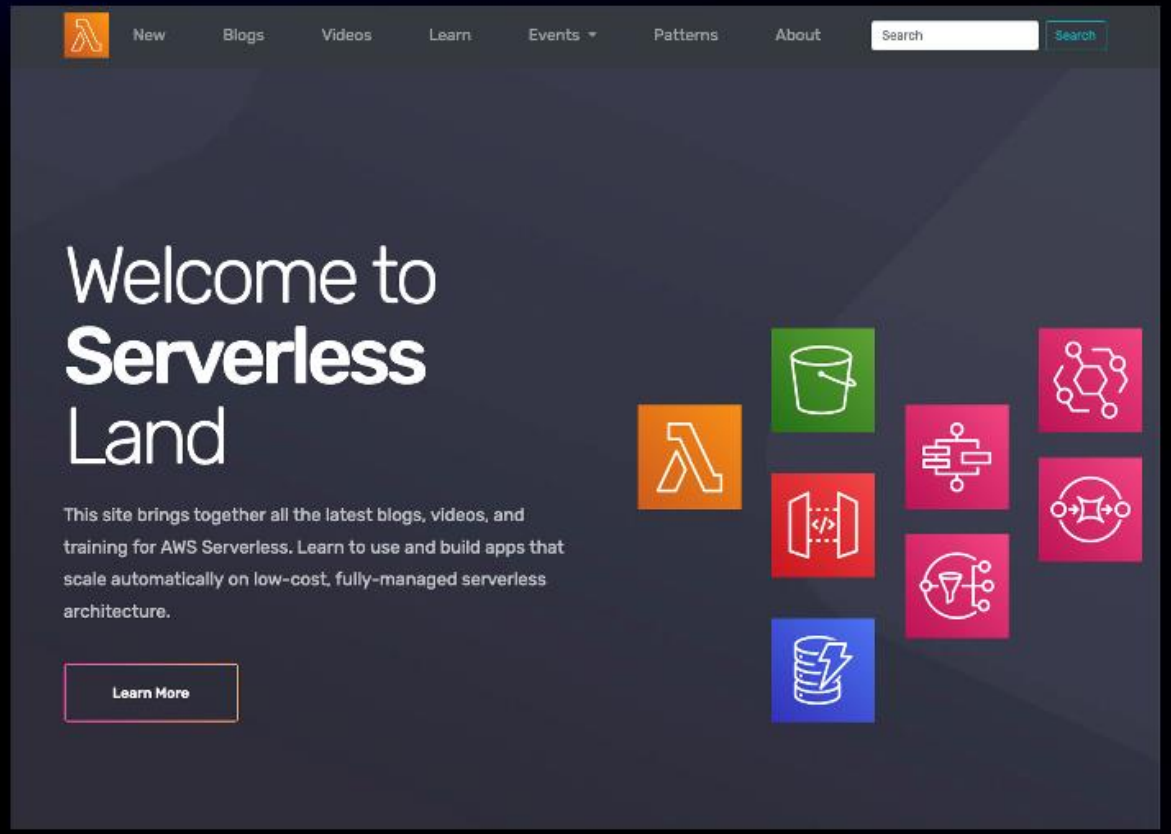
- To leverage AVX2, functions must use dependency libraries compiled for AVX2 support
- Find a distribution that is compiled for AVX2 or compile your own version with the proper compiler flag settings
- No other configuration modifications in Lambda are necessary

Blog post: https://aws.amazon.com/blogs/compute/creating-faster-aws-lambda-functions-with-avx2/

Documentation: https://docs.aws.amazon.com/lambda/latest/dg/runtimes-avx2.html

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Want to learn more?



More: https://serverlessland.com/reinvent2021

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Veda Raman

vedashrr@amazon.com

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.