How Stable Diffusion was built: Tips and tricks to train large AI models

Farshad Farahbakhshian  Emad Mostaque  Pierre-Yves Aquilanti, Ph.D.
Sr. Tech BD – Frameworks ML  CEO  Head of Frameworks ML Solutions
AWS  Stability AI  AWS
Agenda

I. Pop quiz: Generative AI
II. Recent history of AI and trends in AI/ML
III. Stability’s cluster used to train Stable Diffusion
IV. How Stability’s cluster was built using AWS services
Pop quiz: What is the largest cluster size that a customer has publicly announced for training?

I. 512 A100 GPUs
II. 1,024 A100 GPUs
III. 4,000 A100 GPUs
IV. 10,000 A100 GPUs
Pop quiz: What is the largest cluster size that a customer has publicly announced for training?

I. 512 A100 GPUs
II. 1,024 A100 GPUs
III. 4,000 A100 GPUs
IV. 10,000 A100 GPUs
Pop quiz: According to Gartner, what % of all data will be produced by Generative AI by 2025?

I. 1%
II. 5%
III. 10%
IV. 25%
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>1%</td>
</tr>
<tr>
<td>II.</td>
<td>5%</td>
</tr>
<tr>
<td>III.</td>
<td>10%</td>
</tr>
<tr>
<td>IV.</td>
<td>25%</td>
</tr>
</tbody>
</table>

Pop quiz: According to Gartner, what % of all data will be produced by Generative AI by 2025?
Pop quiz: According to Gartner, what % of drug discovery will use Generative AI by 2025?

I. 10%
II. 25%
III. 33%
IV. 50%
Pop quiz: According to Gartner, what % of drug discovery will use Generative AI?

I. 10%
II. 25%
III. 33%
IV. 50%
Transforming the artificial intelligence space

“Attention Is All You Need” introduced Transformers.

Two key contributions:

1. Efficiently use parallel computing
2. Introduced concept of “Attention,” allowing AI to understand the relationship between words

Transformers have driven significant progress in AI such as GPT3, BERT, and Generative AI like Stable Diffusion.

Source: Photo by Arseny Togulev on Unsplash
Generative AI workflow

Data
- Text
- Images
- Video
- Audio
- Numerical

Training

Foundational Model

Fine Tuning

Fine-Tuned Model
- Art
- Code
- Speech
- Film
- Sentient Analysis
- Blogs
- Marketing
- Text
- Music
- Architecture
- AGI
- Synthetic Data

Inference
The AWS ML stack

AI SERVICES

Vision
Rekognition

Chatbots
Lex

Business tools
Personalize Forecast
Fraud Detector
Lookout for Metrics

Search
Kendra

Healthcare
HealthLake
Comprehend Medical
Transcribe Medical

Speech
Polly
Transcribe

Text
Comprehend
Translate
Extrakt

Contact centers
Contact Lens
Connect Voice ID

Code + DevOps
CodeGuru
DevOps Guru

Industrial
Panorama Appliance and SDK,
Monitron, Lookout for Equipment, Lookout for Vision

AMAZON SAGEMAKER

Label data
Data collection prep
Store features
Detect bias and explain predictions
Visualize in notebooks
Pick algorithm
Train models faster
Tune parameters
Deploy in production
Manage & monitor
Manage edge devices

SAGEMAKER STUDIO IDE

CI/CD

SELF-MANAGED ML

PyTorch
Deep learning
AMIs & containers

mxnet

TensorFlow

Habana Gaudi
Accelerators

Inferentia

Trainium

FPGA

PyTorch, the PyTorch logo and any related marks are trademarks of Facebook, Inc.
TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.
### ML frameworks and infrastructure layer

#### TOOLS

<table>
<thead>
<tr>
<th>Data Processing &amp; Labeling</th>
<th>Prebuilt Images</th>
<th>Dev, Training, &amp; Tuning</th>
<th>Inference Servers</th>
<th>MLOps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinesis</td>
<td>DLAMi</td>
<td>Jupyter Notebooks</td>
<td>Torch</td>
<td>Kubeflow</td>
</tr>
<tr>
<td>Glue</td>
<td>DL Containers</td>
<td>Nvidia NCCL</td>
<td>Serve</td>
<td></td>
</tr>
<tr>
<td>SageMaker Ground Truth</td>
<td></td>
<td>Kubeflow Operators</td>
<td>Nvidia Triton</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TF Serving</td>
<td></td>
</tr>
</tbody>
</table>

#### Orchestration

<table>
<thead>
<tr>
<th>AWS Services</th>
<th>Workflow</th>
<th>Open Source Software</th>
<th>Other Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>EKS</td>
<td>Kubernetes</td>
<td>Kubeflow</td>
<td>Lambda</td>
</tr>
<tr>
<td>ECS</td>
<td></td>
<td>Airflow</td>
<td>Step Functions</td>
</tr>
<tr>
<td>Fargate</td>
<td></td>
<td>Ray</td>
<td>Cloud Formation</td>
</tr>
<tr>
<td>Batch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ParallelCluster</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Core Infrastructure

<table>
<thead>
<tr>
<th>Compute</th>
<th>Storage</th>
<th>Networking</th>
<th>Data Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2</td>
<td>Amazon S3</td>
<td>EFA</td>
<td>Direct connect</td>
</tr>
<tr>
<td>P2</td>
<td>Amazon FSx for Lustre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DL1</td>
<td>Amazon EFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G4dn</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inf1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Fun facts about Stability

1. Stability’s cluster is training ~10 different models at a time

2. Stable Diffusion 2.0 took 200,000 A100 hours to train

3. Stability’s cluster uses AWS UltraCluster and EFA to optimize Distributed Training speeds

Source: https://towardsdatascience.com/the-rise-of-cognitive-ai-a29d2b724ccc
Building & Operating
The Largest Supercomputer in the Cloud
The Stability Platform

API + Integrations:

We provide a global generative inference pipeline to platform partners and integrators.

Research Models:

We support our research communities with funding and compute to bring models to SOTA.

Data:

We have released the largest open text and image datasets used by the top companies globally.

We will be the de-facto generative application layer for a new creative era.
In a gold rush for compute, companies build bigger than national supercomputers

“We think the most benefits will go to whoever has the biggest computer” – Greg Brockman, OpenAI CTO
Stable Diffusion

Latent Diffusion ++
Brief Overview of Diffusion Models

- “destroy” the data by gradually adding small amounts of gaussian noise

- “create” data by gradually denoising a noisy code from a stationary distribution

Animations from [https://yang-song.github.io/blog/2021/score/](https://yang-song.github.io/blog/2021/score/)
Stable Diffusion 2.0

https://discord.gg/stablediffusion
Fine-tuning Stable Diffusion 2.0 with 10 images:
Depth to Image Model (Stable Diffusion 2.0)
Inpainting

concatenation of mask and encoded masked image
“Upgrade” your child’s artwork
original post: https://www.reddit.com/r/StableDiffusion/comments/wyq04v/using_img2img_to_upgrade_my_sons_artwork/
UIs / Plug-Ins for Photoshop, GIMP etc

https://twitter.com/wbuchw/status/1563162131024920576

https://github.com/lkwq007/stablediffusion-infinity
We are witnessing an explosion of interest in a new creative toolset
Open source

AI

Community

stability ai
Community

Our relationship with open-source

Model releases
Successive releases of models

Projects
Community-developed projects

Collaboration
Support and further collaboration

By collaborating with the community, the technology will be pushed further
A foundation for creative platforms

Data
New data partnerships

Fine-tuning
Custom models
For partners

API
Creative platforms
and pipelines

Stability will power the future adoption of generative models into creative tools
Key Learnings

Partnering with AWS

Co-Build
SageMaker Integration

Compute
Largest Publicly announced A100 cluster

Amazon Partnership
Edge Compute
How you can use Stability's Models Today

- **Stay tuned**
- **Download**
  - Open-Source Models with pre-trained weights
- **API**
  - Releasing soon
• Stability Diffusion 2.0

• Depth to image

• 8x up-scaler

• Inpainting & Outpainting
Stability Roadmap

- Stability Diffusion 2.0 launched
- Open Source LLM (GPT-Neo Next Gen)
- Code Generation Model
- Audio/Video/3D
- Fine-tuning via API
Distributed ML training architecture
Distributed training – HPC-ML cluster building blocks

Compute
- P4, Tm1 Instances
- Containers

Network
- Elastic Fabric Adapter
- Region
- VPC
- Availability Zone

Storage
- Amazon FSx for Lustre
- Amazon EC2 Instance Store
- Amazon Elastic Block Store (Amazon EBS)

Architecture & Orchestration
- AWS ParallelCluster
- AWS CloudFormation
Over 4,000+ Accelerators

Fully non-blocking petabit-scale network infrastructure

High-throughput, low-latency storage from Amazon FSx for Lustre
Amazon EC2 Trn1/Trn1n instances

**Trn1(n)**

- **BF16/FP16**
  - TF32: 3.4 PFLOPS
  - FP32: 840 TFLOPS

**AGGREGATE ACCELERATOR MEMORY**
- 512 GB

**PEAK MEMORY BANDWIDTH**
- 13.1 TB/sec

**EFA NETWORK CONNECTIVITY**
- 800/1600 Gbps

**NEURON-CORE V2**

**NEURON-LINK V2**

**AWS Trainium**
High performance machine learning training chip, purpose-built by AWS

**EC2 Trn1/Trn1n Instances**
The most cost-efficient high-performance training instance
Amazon EC2 **Trn1/Trn1n** instances

**AWS Trainium**
High performance machine learning training chip, purpose-built by AWS

**EC2 Trn1/Trn1n Instances**
The most cost-efficient high-performance training instance

**Trn1 UltraClusters**
Train large models with 30K+ Trainium devices, and non-blocking EFA network
High-speed networking: Elastic Fabric Adapter

Elastic Network Adapter only

Application

MPI/NCCL Implementation

TCP/IP Stack

ENA Network Driver

ENA Device

VPC Networking

With Elastic Fabric Adapter

Application

MPI/NCCL Implementation

Libfabric

EFA Kernel Driver

EFA Device

SRD

ENA Device

VPC Networking

1. Exposed over libfabric similar to IB
2. Out-of-order delivery – no head-of-line blocking
3. Packet spraying over multiple ECMP paths
ML training storage hierarchy

**Instance Store (NVMe)**
- Checkpoints, temporary data
- Stripped with RAID0 via ParallelCluster

**FSx for Lustre**
- Shared data sets, checkpoints, outputs
- Home directories (due to ~/.cache)

**Amazon S3**
- Data backbone, datasets, checkpoints, outputs
- Lustre Hierarchical Storage Manager (HSM) backup tier
HPC Cluster - AWS ParallelCluster

Head-node
- 1 × c5.9xlarge 36 vCPUs (18 physical)
- 72 GB of memory

Compute Node
- 100+ × p4de.24xlarge + C6, M6, R6
- 96 vCPUs (48 physical)
- 1152 GB of memory
- 8 × NVIDIA A100 80GB GPUs
- Network: 400Gbs ENA & EFA
- Storage: 8 × 1TB NVMe + EBS

Shared file-systems
- Amazon FSx for Lustre of 108TB on /fsx

Cluster Stack
- Slurm 22.05.5
- Cuda 11.6
How to create a cluster

```
Region: us-east-1
Image:
  Os: alinux2
...
HeadNode:
  InstanceType: c5.9xlarge
Networking:
  SubnetId: subnet-123456789abcdef0
...
Scheduling:
  Scheduler: slurm
  SlurmQueues:
    Name: train
  ComputeResources:
    MinCount: 64
    MaxCount: 64
    InstanceType: p4d.24xlarge
...
SharedStorage:
  Name: FsxLustre0
  StorageType: FsxLustre
  MountDir: /scratch
...
```

- **Head-node**
- **Compute nodes**
- **Shared storage**

https://aws.amazon.com/hpc/parallelcluster/
Cluster user management and multi-use

Authentication
- Central user management using AWS Active Directory with AWS ParallelCluster
- Jobs controls, usage across regions & users

Multi-instances job queues
- Training with p4d.24xlarge & Trn1.32xlarge
- Preprocessing on C6i, M6i & R6i
Software stack management: post-install scripts

Head node

- User packages (pip, conda)
- EC2 Capacity Reservation
- Slurm Prologs
- Custom packages + Monitoring (Prometheus, Grafana)
- Pyxis + Enroot
- Cost control

Compute nodes

- Singularity
- GPU boost & monitor

User managed (after creation)

Node type specific

Cluster wide
Jobs pre-emption and QoS

Idle gaps filled with low priority and pre-emptible jobs
- Datasets preprocessing by batches (language translation, synthetic captioning...), inference
- 4 nodes jobs with runtime < 48h
- Jobs requeued automatically upon interruption
GPU monitoring and optimizations

Job level
- Prolog jobs (NCCL bandwidth, GPU topology)
- Debug prolog with DCGMI health monitoring

Online
- Cloudwatch, Grafana+Prometheus

If an issue is detected, node tainted and excluded

Thank you!

Farshad Farahbakhshian  
farshaf@amazon.com

Emad Mostaque  
https://twitter.com/emostaque

Pierre-Yves Aquilanti, Ph.D.  
pierreya@amazon.com

Please complete the session survey in the mobile app
Additional resources

Stability cluster architecture & resources
https://github.com/Stability-AI/stability-hpc

AWS ParallelCluster documentation
https://docs.aws.amazon.com/parallelcluster/index.html