AWS re:Invent

NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

CMP 314

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

Farshad Farahbakhshian
Sr. Tech BD – Frameworks ML
AWS

Emad Mostaque
CEO
Stability AI

Pierre-Yves Aquilanti, Ph.D.
Head of Frameworks ML Solutions
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%

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%

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.

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

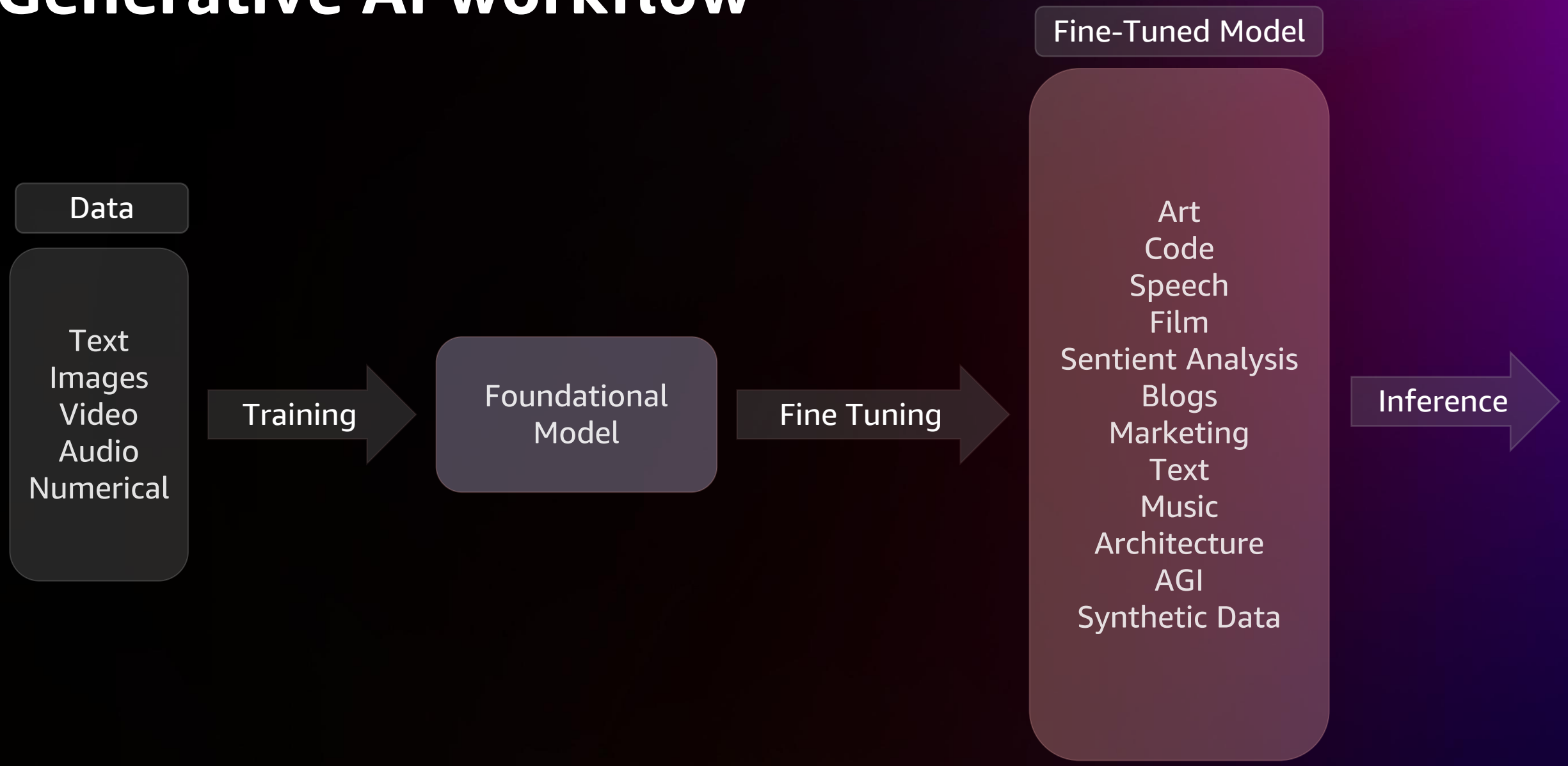
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Source: Photo by Arseny Togulev on Unsplash

Generative AI workflow



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
Textract

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

SAGEMAKER STUDIO IDE

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

CI/CD

Manage edge devices

SELF-MANAGED ML

PYTORCH
mxnet TensorFlow

Deep learning AMIs & containers

GPUs

Habana Gaudi Accelerators

Inferentia

Trainium

FPGA

ML frameworks and infrastructure layer

TOOLS

DATA PROCESSING & LABELING

- Kinesis
- Glue
- SageMaker Ground Truth

PREBUILT IMAGES

- DLAMI
- DL Containers

DEV, TRAINING, & TUNING

- Jupyter Notebooks
- Nvidia NCCL
- Kubeflow Operators

INFERENCE SERVERS

- Torch Serve
- Nvidia Triton
- TF Serving

MLOps

- Kubeflow

ORCHESTRATION

AWS SERVICES

- EKS
- ECS
- Fargate
- Batch
- ParallelCluster

WORKFLOW

OPEN SOURCE SOFTWARE

- Kubernetes
- Kubeflow
- Airflow
- Ray

OTHER SERVICES

- Lambda
- Step Functions
- Cloud Formation

CORE INFRASTRUCTURE

COMPUTE

- Trn1
- P4
- DL1
- G5
- G4dn
- Inf1

STORAGE

- Amazon S3
- Amazon FSx for Lustre
- Amazon EFS

NETWORKING

- EFA

DATA TRANSFER

- Direct connect

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:

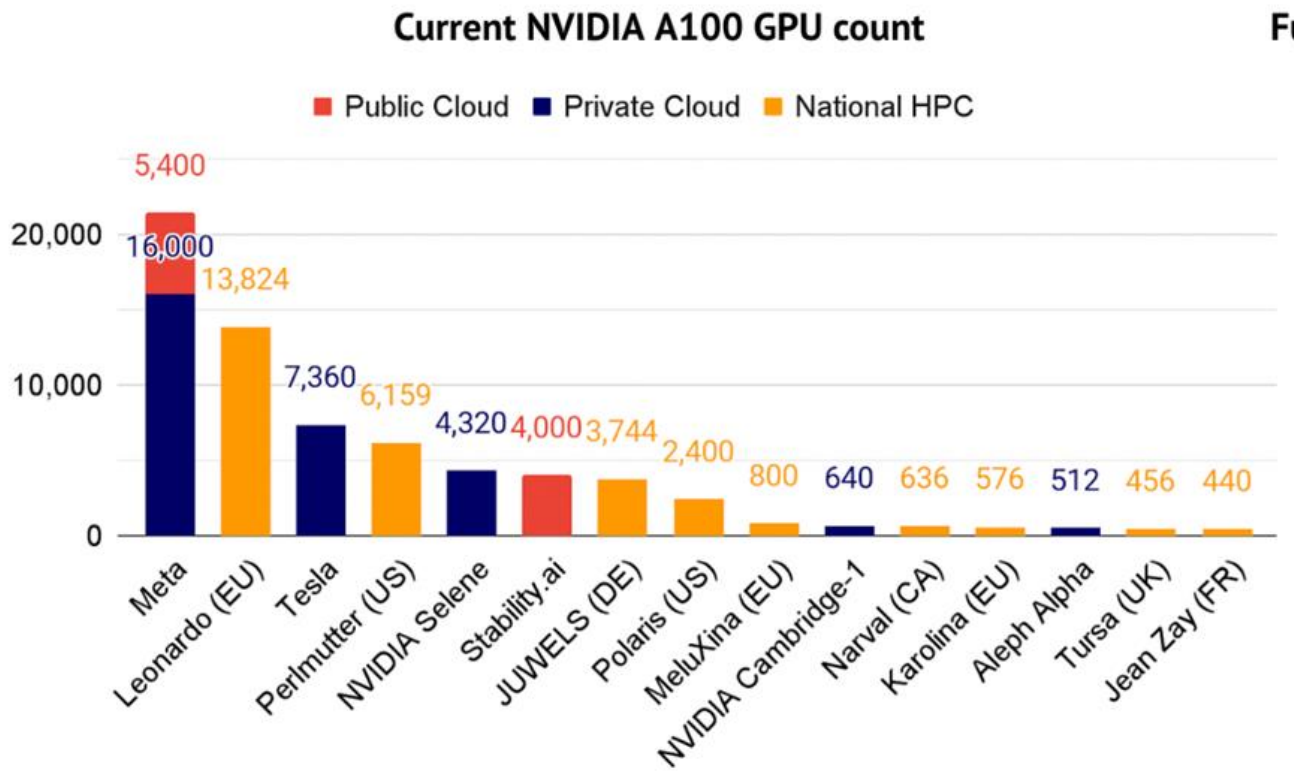
eleutherai LAION

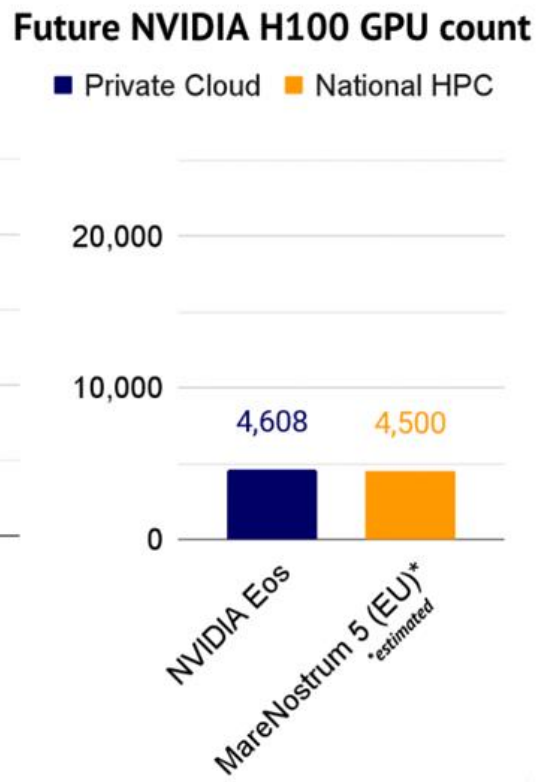
We have released the largest open text and image datasets used by the top companies globally

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



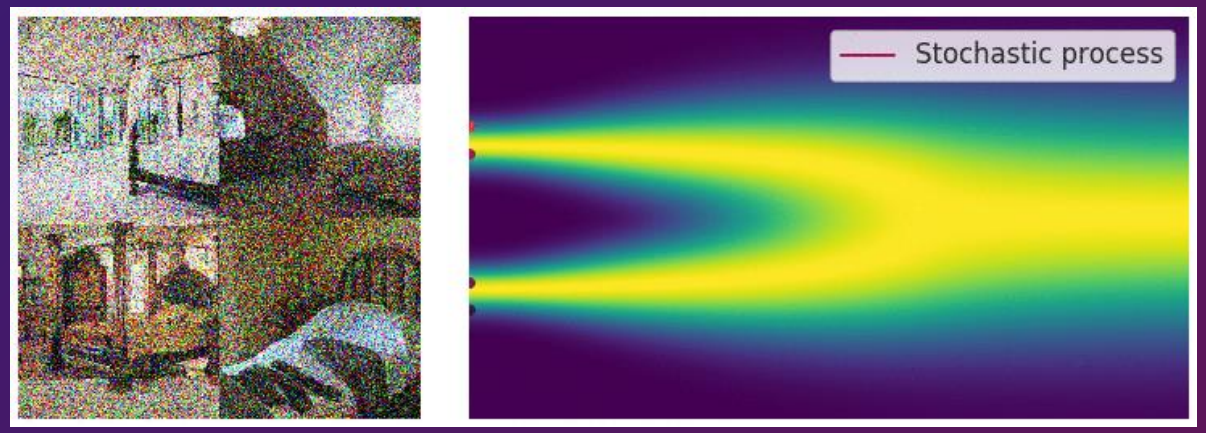


stateof.ai 2022

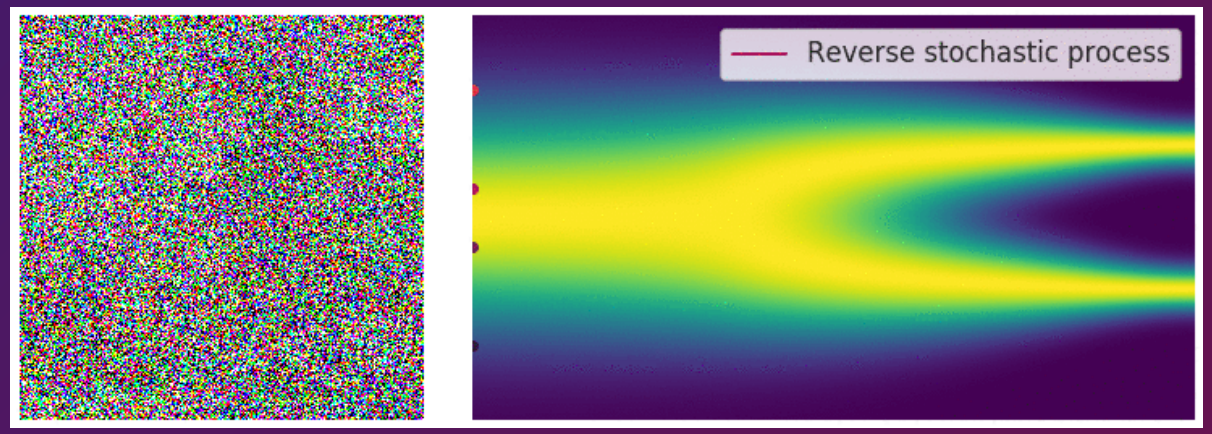
Stable Diffusion

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

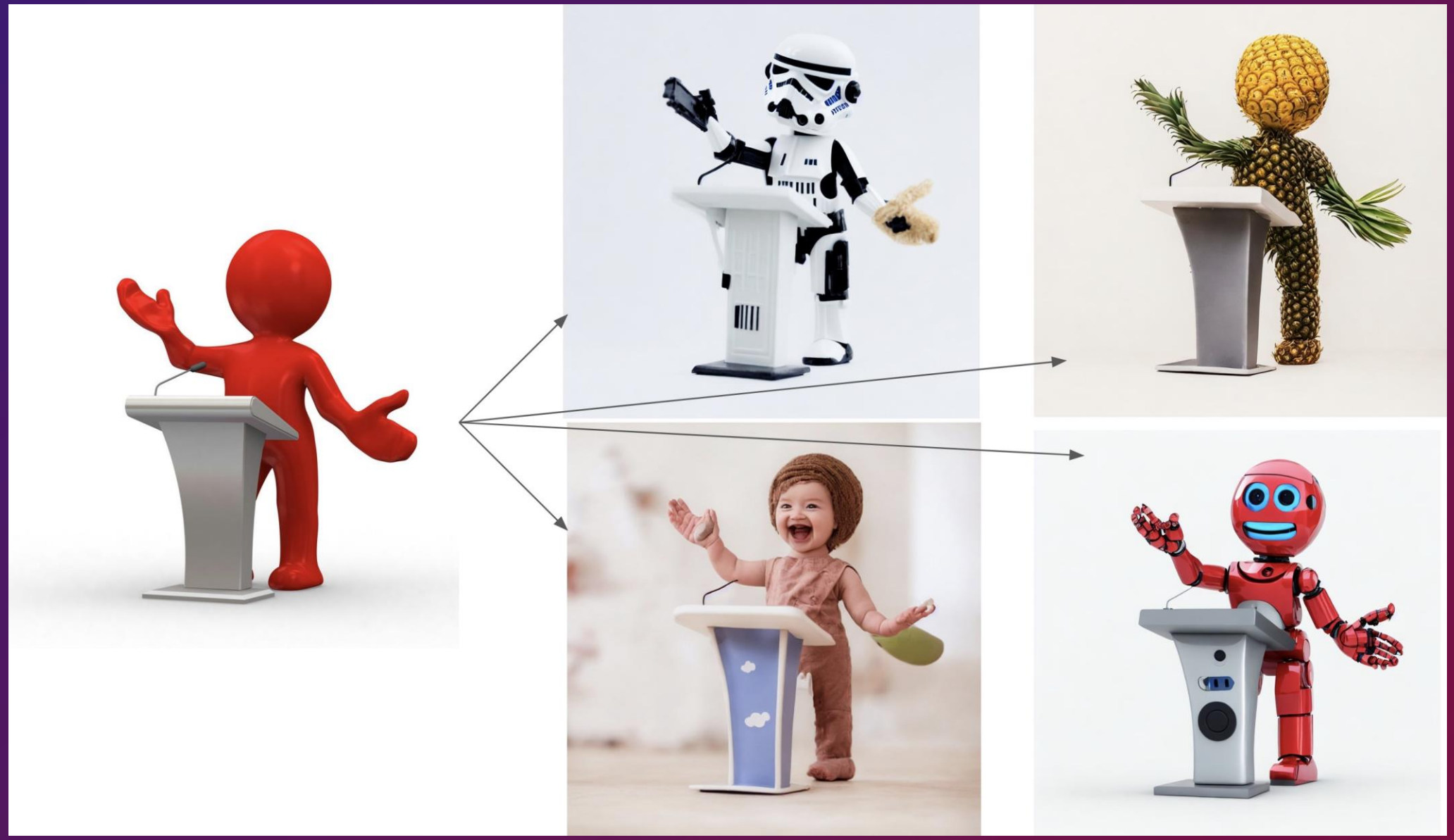
Animations from https://yang-song.github.io/blog/2021/score/

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)



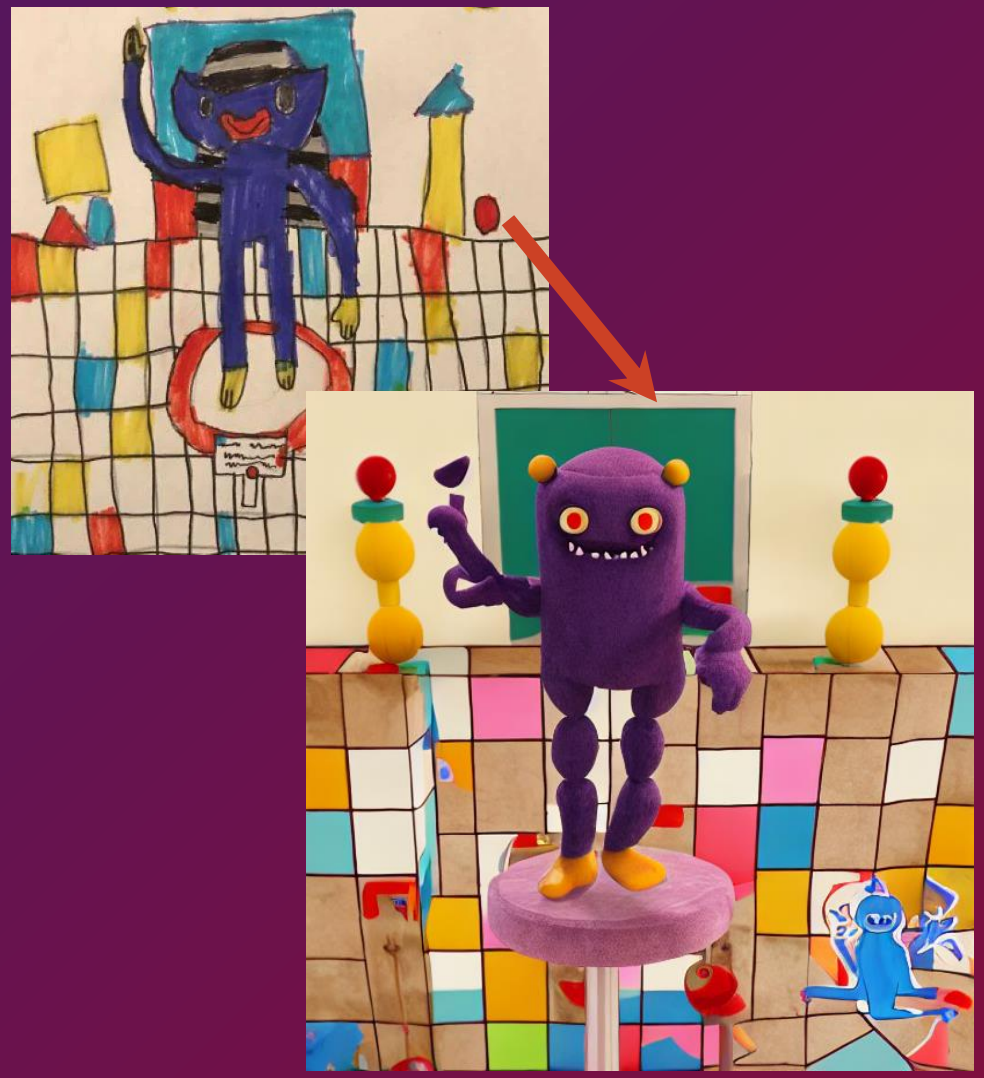
Inpainting

concatenation of mask and encoded masked image

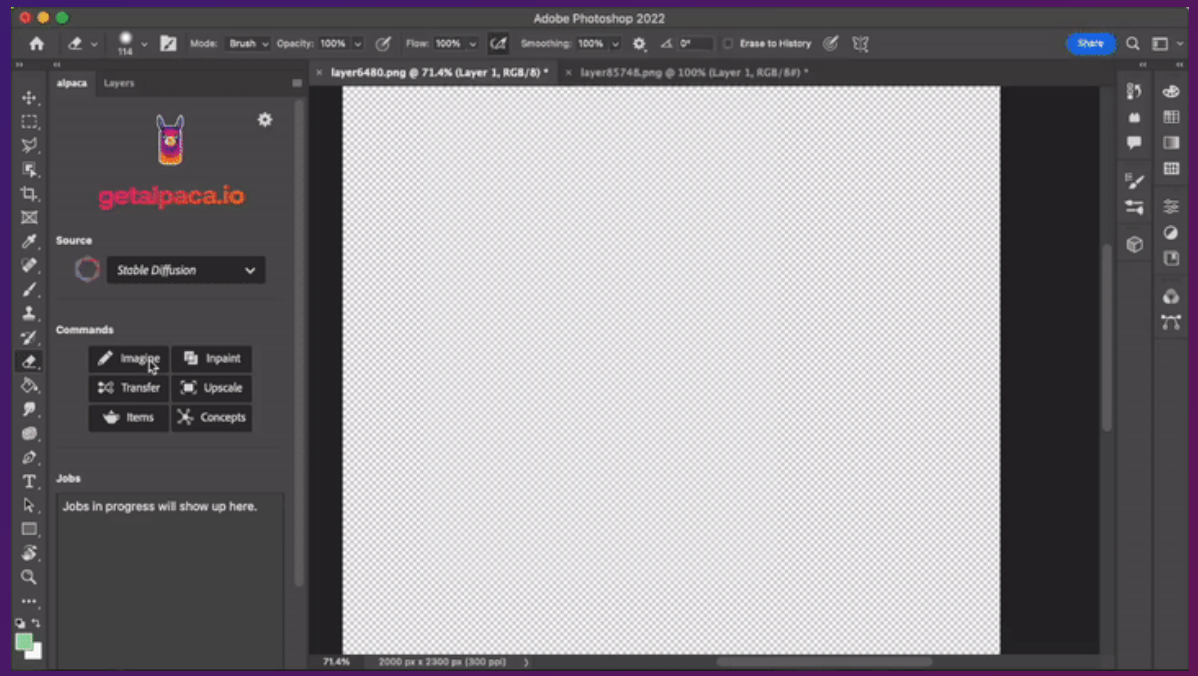
Method	40-50% masked		All samples	
	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
LDM-4 (ours, big, w/ ft)	**9.39**	0.246± 0.042	**1.50**	0.137± 0.080
LDM-4 (ours, big, w/o ft)	12.89	0.257± 0.047	2.40	0.142± 0.085
LDM-4 (ours, w/ attn)	11.87	0.257± 0.042	2.15	0.144± 0.084
LDM-4 (ours, w/o attn)	12.60	0.259± 0.041	2.37	0.145± 0.084
LaMa [85]†	12.31	**0.243**± 0.038	2.23	**0.134**± 0.080
LaMa [85]	12.0	**0.24**	2.21	0.14
CoModGAN [103]	10.4	0.26	1.82	0.15
RegionWise [49]	21.3	0.27	4.75	0.15
DeepFill v2 [100]	22.1	0.28	5.20	0.16
EdgeConnect [55]	30.5	0.28	8.37	0.16

“Upgrade” your child’s artwork

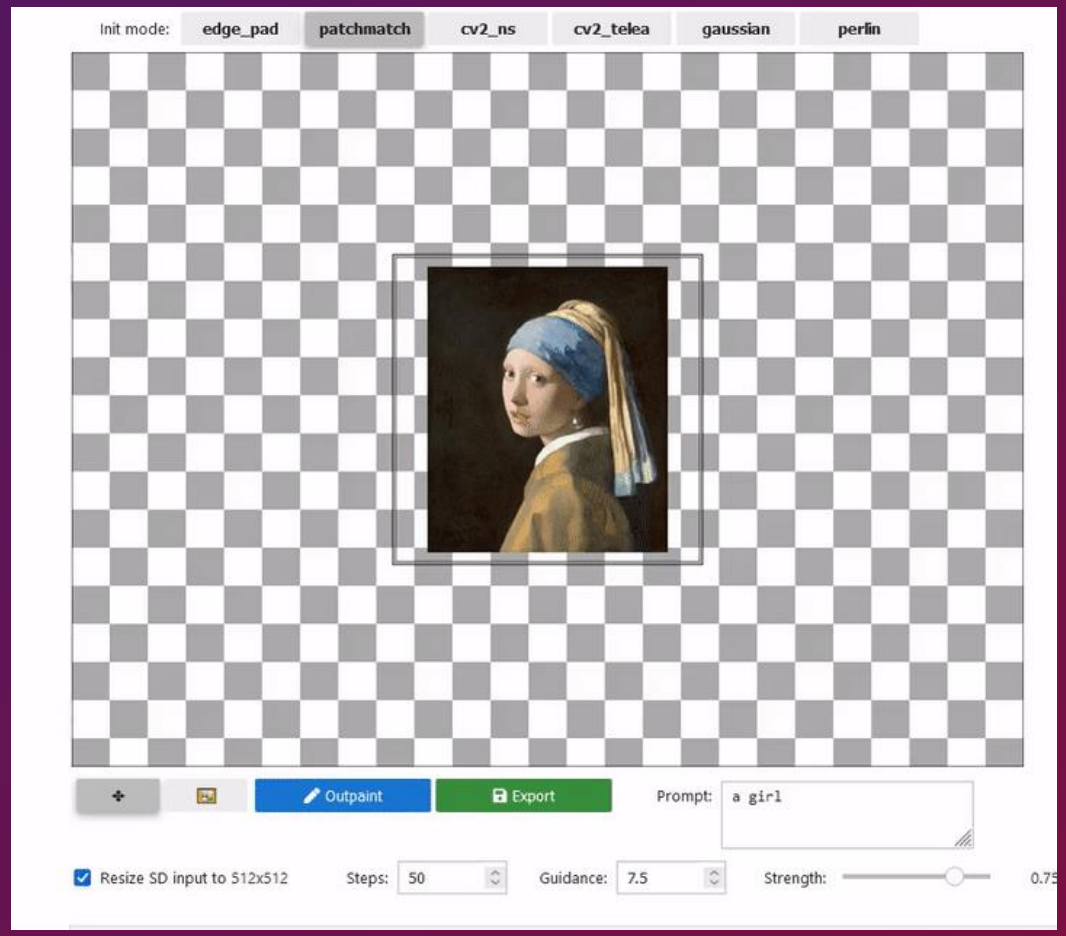
original post: https://www.reddit.com/r/StableDiffusion/comments/wyq04v/using_img2img_to_upgrade_my_sons_artwork/



UIs / Plug-Ins for Photoshop, GIMP etc



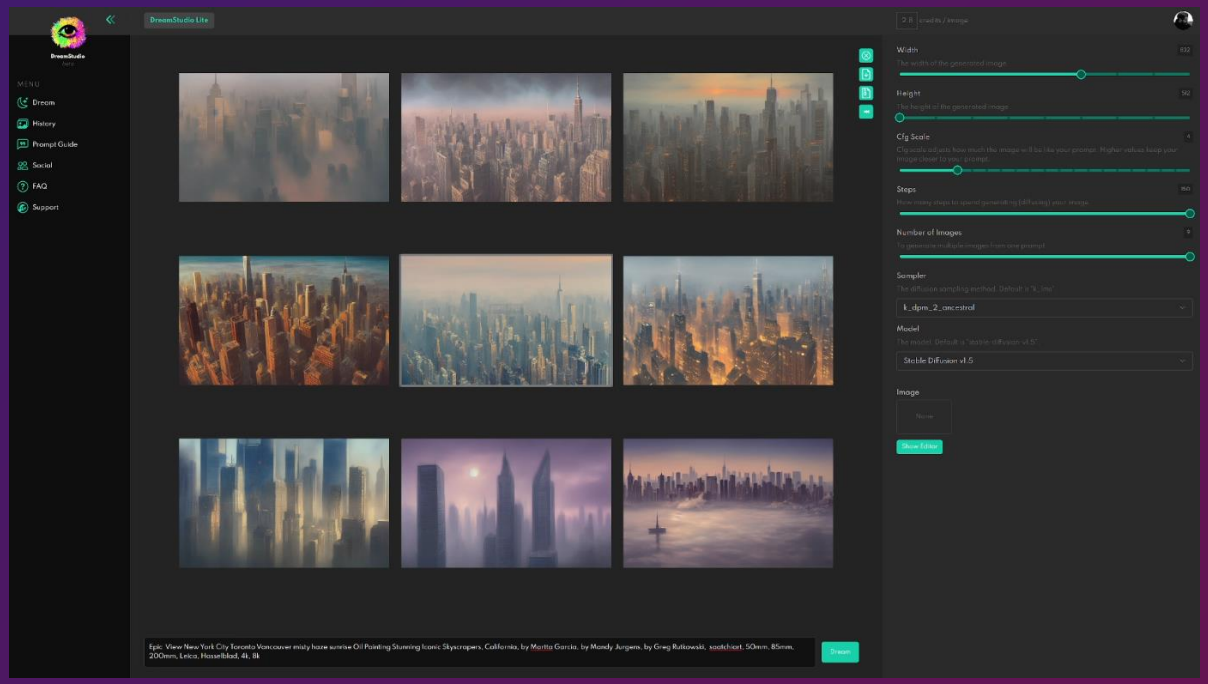
https://twitter.com/wbuchw/status/1563162131024920576

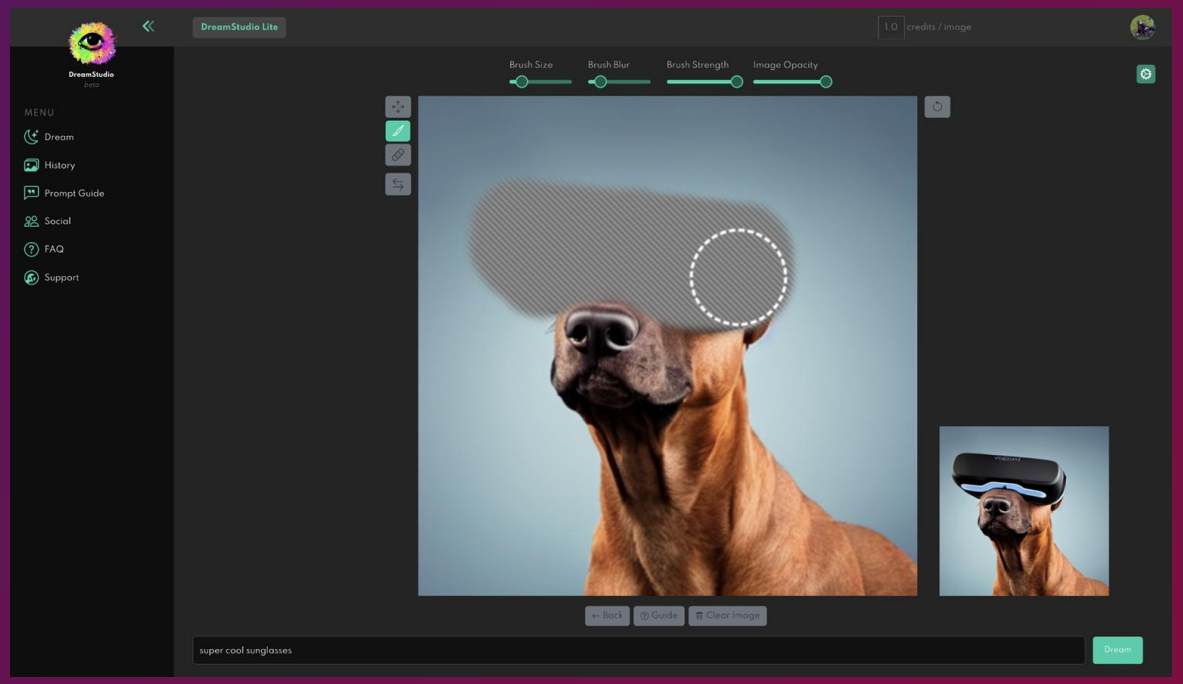


https://github.com/lkwq007/stablediffusion-infinity

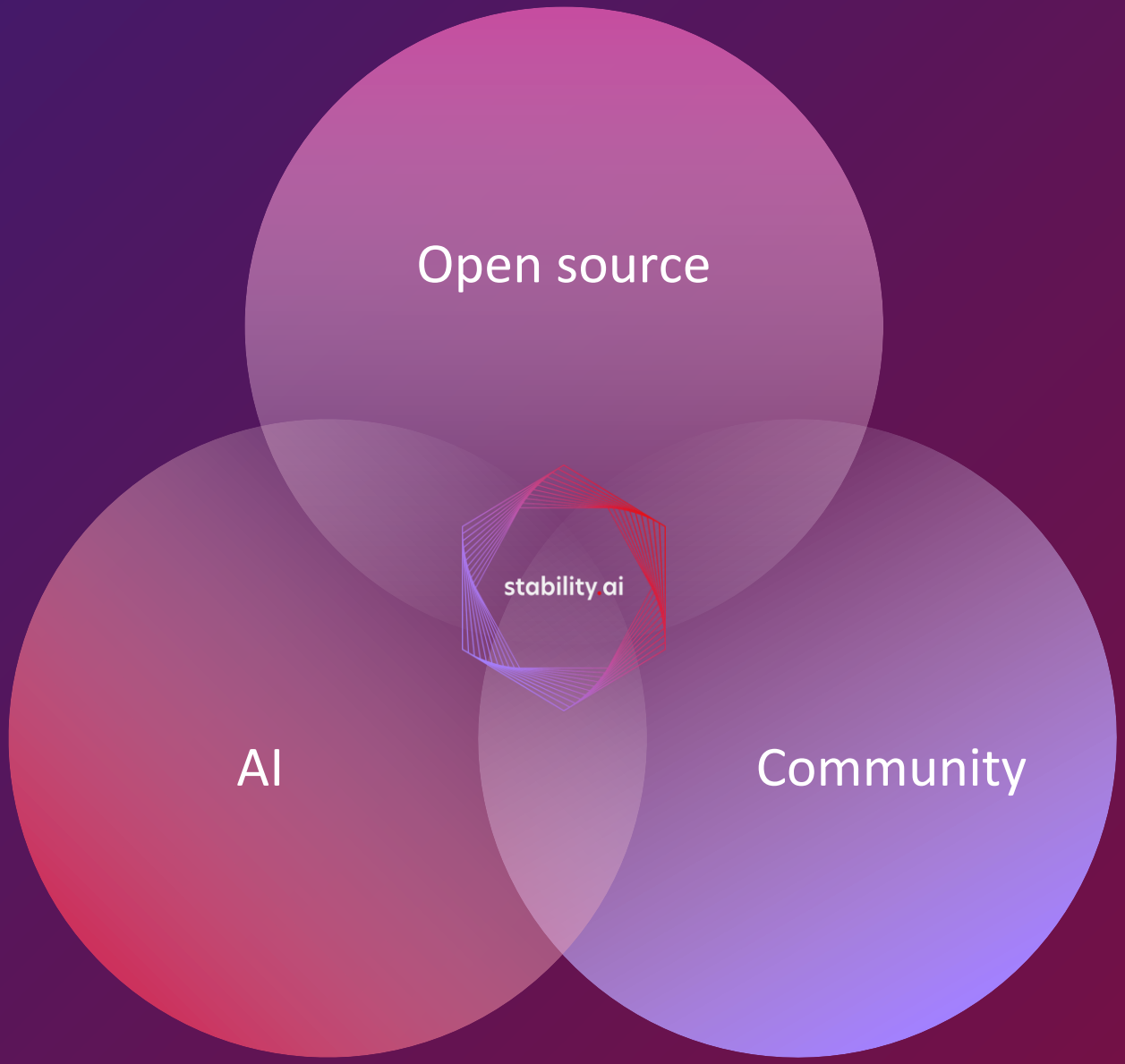
DreamStudio Features

Image Techniques + Editing





We are witnessing an explosion of interest in a new creative toolset


Open source
stability ai
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

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

Using Stability Today

- Stability Diffusion 2.0
- Depth to image
- 8x up-scaler
- Inpainting & Outpainting

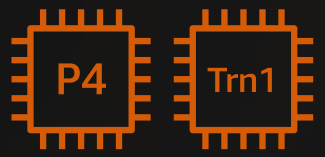
Stability Roadmap

- 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



Instances

Containers

Network

Elastic Fabric Adapter

Storage

Amazon FSx for Lustre

Amazon EC2 Instance Store

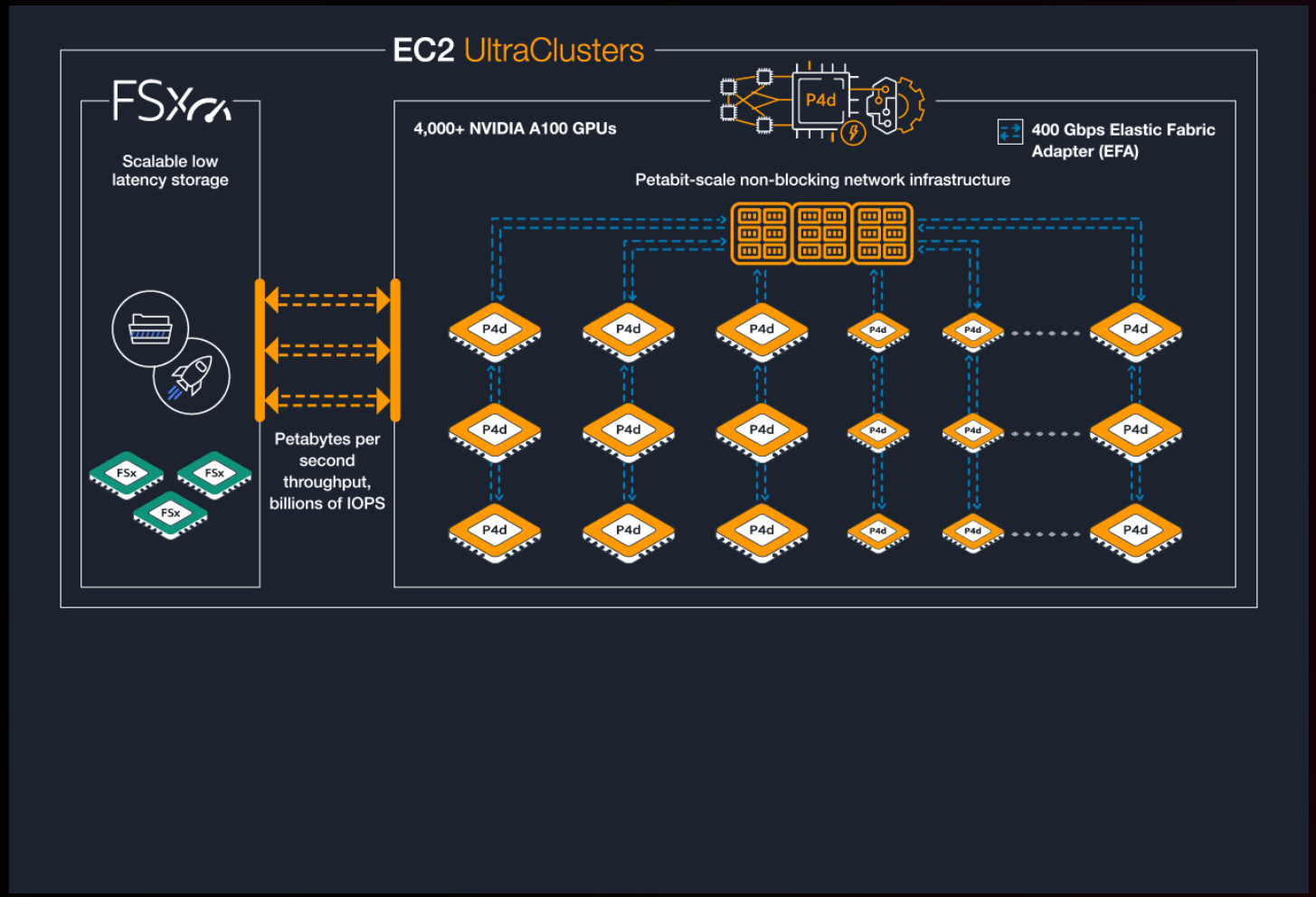
Amazon Elastic Block Store (Amazon EBS)

Architecture & Orchestration

AWS ParallelCluster

AWS CloudFormation

Compute - EC2 UltraClusters



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	FP32
3.4 PFLOPS	3.4 PFLOPS	840 TFLOPS

AGGREGATE ACCELERATOR MEMORY
512 GB

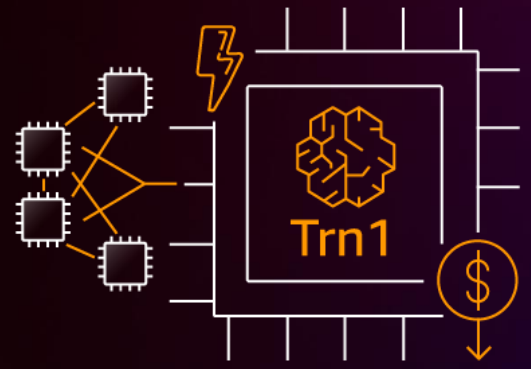
PEAK MEMORY BANDWIDTH
13.1 TB/sec

EFA NETWORK CONNECTIVITY
800/1600 Gbps

NEURON-CORE V2
NEURON-LINK V2

AWS Trainium

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



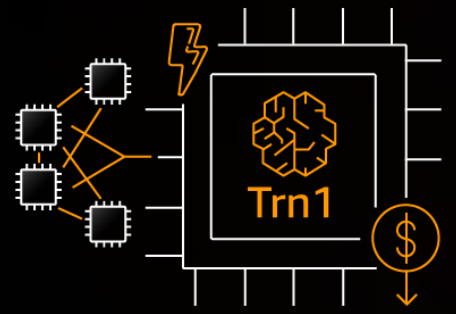
EC2 Trn1/Trn1n Instances

The most cost-efficient high-performance training instance

Amazon EC2 Trn1/Trn1n instances

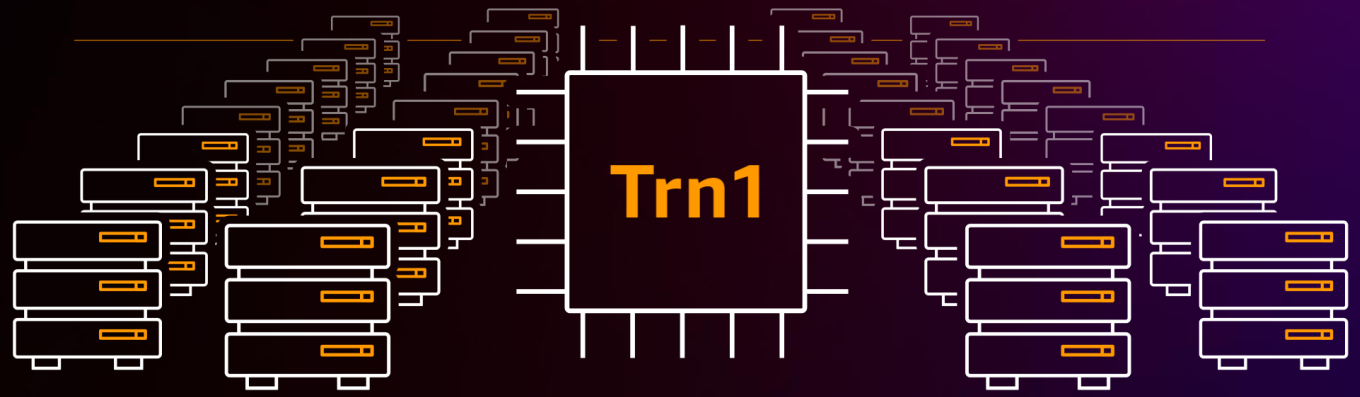
AWS Trainium

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



EC2 Trn1/Trn1n Instances

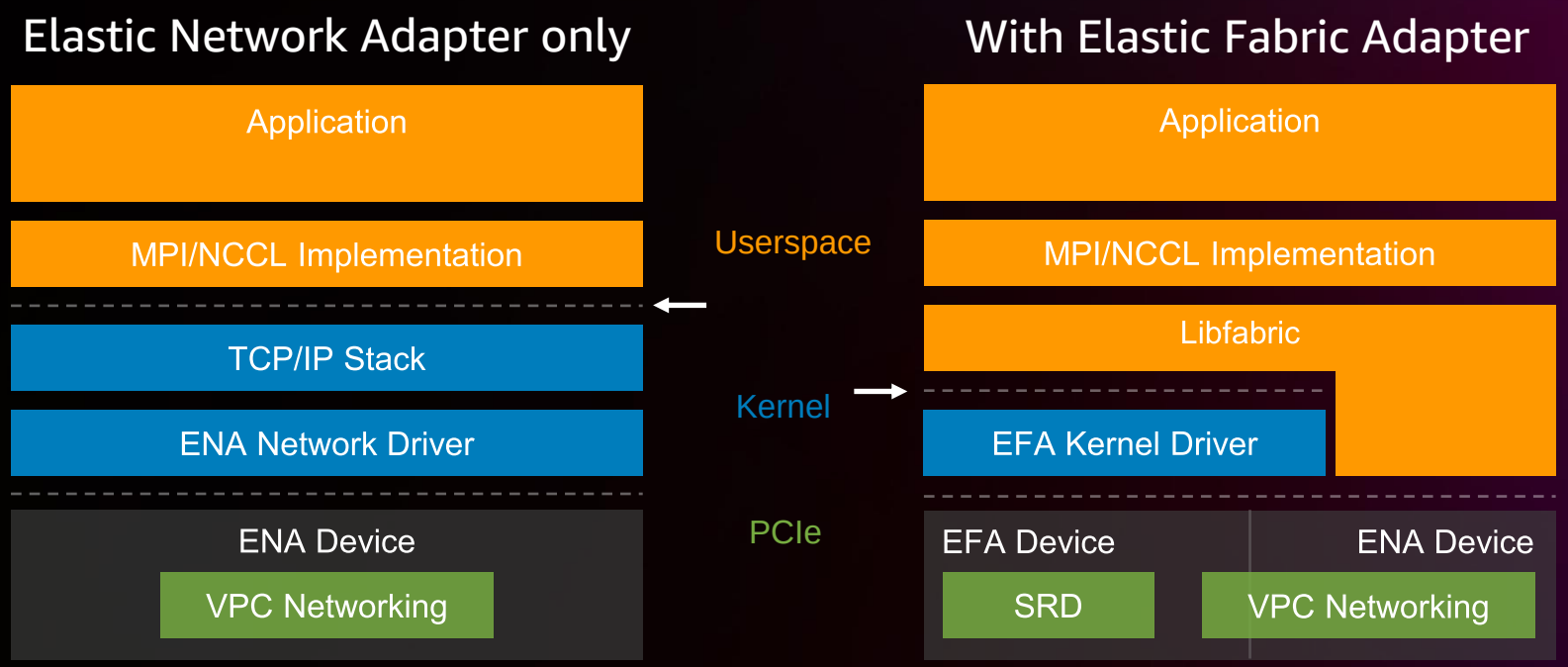
The most cost-efficient high-performance training instance



Trn1 UltraClusters

Train large models with 30K+ Trainium devices, and non-blocking EFA network

High-speed networking: Elastic Fabric Adapter



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

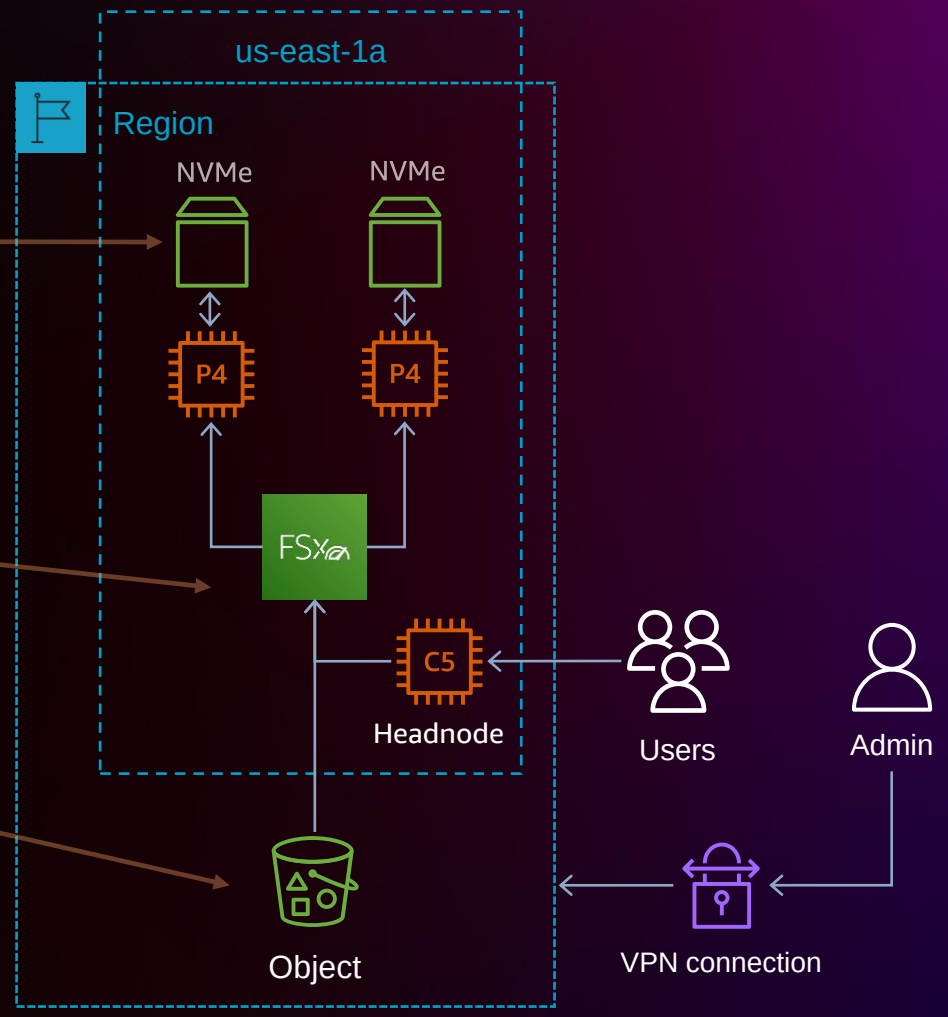
https://ieeexplore.ieee.org/document/9167399

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



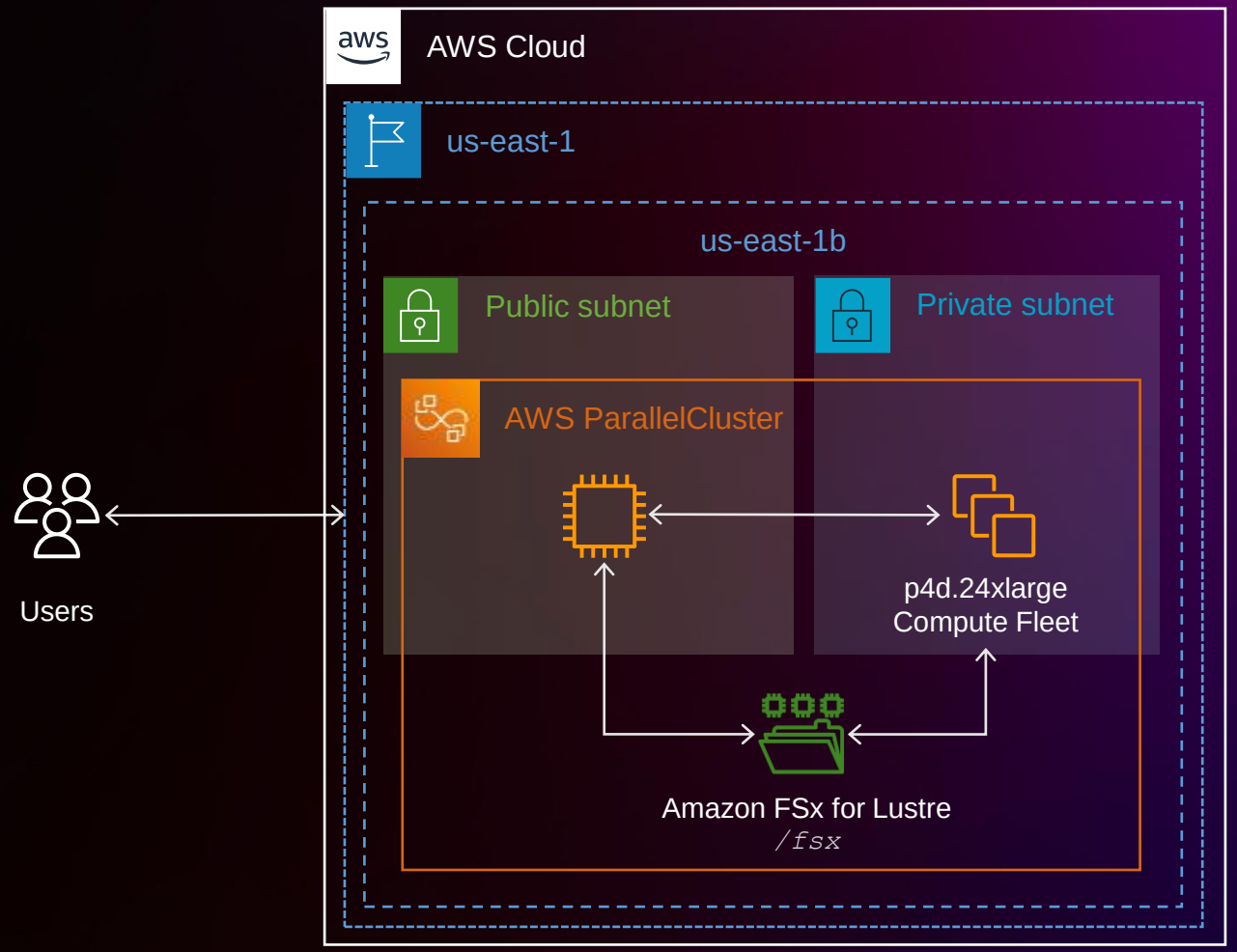
HPC Cluster - AWS ParallelCluster

Head-node
- 1 × c5.9xlarge36 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

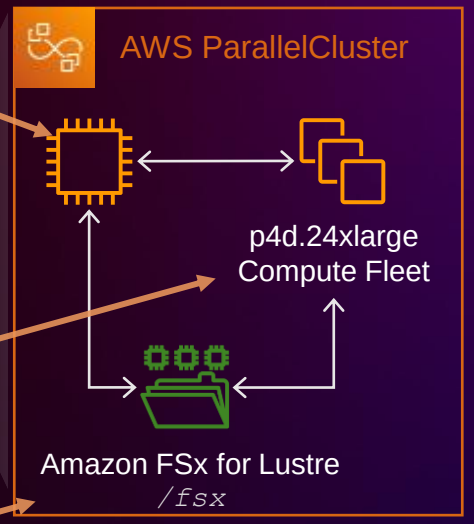
```yaml
Region: us-east-1
Image:
  Os: alinux2
...
HeadNode:
  InstanceType: c5.9xlarge
  Networking:
    SubnetId: subnet-123456789abcdef0
...
Scheduling:
  Scheduler: slurm
  SlurmQueues:
    - Name: train
      ComputeResources:
        - MinCount: 64
          MaxCount: 64
          InstanceType: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

Head-node

```
pcluster create-cluster -f config.yaml …
```

Compute nodes

Shared storage



https://aws.amazon.com/hpc/parallelcluster/

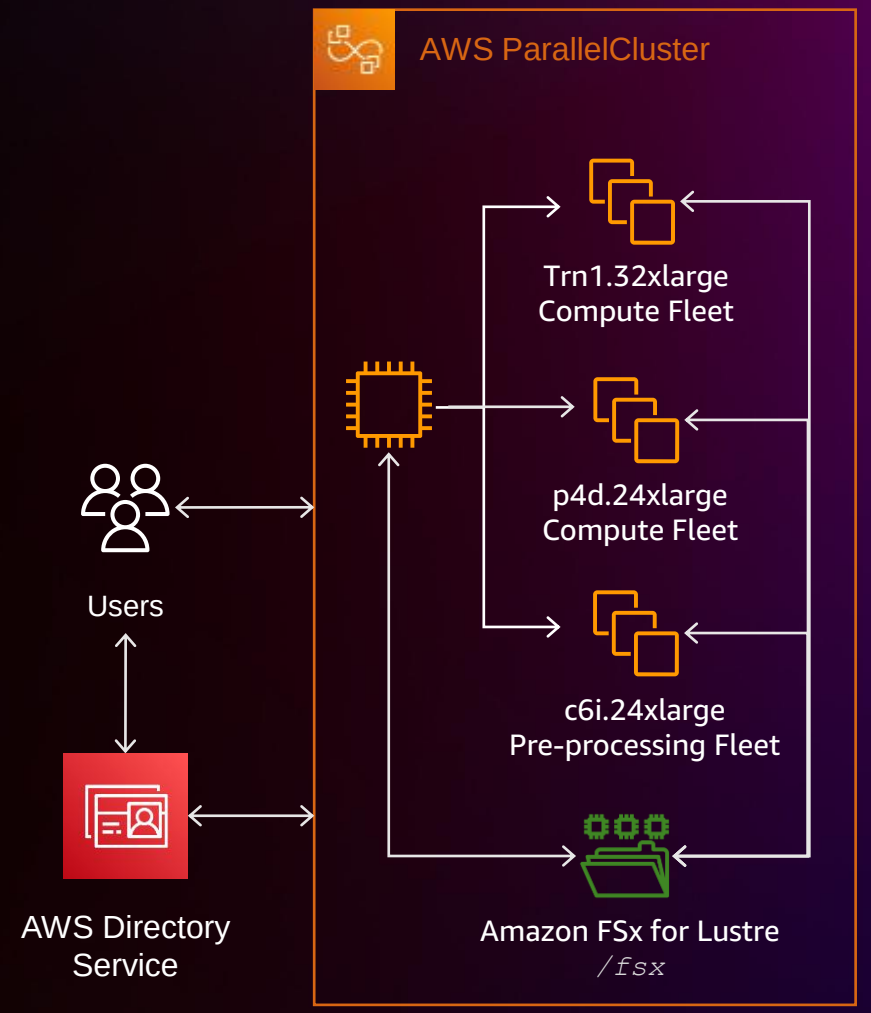
Cluster user management and multi-use

Authentication

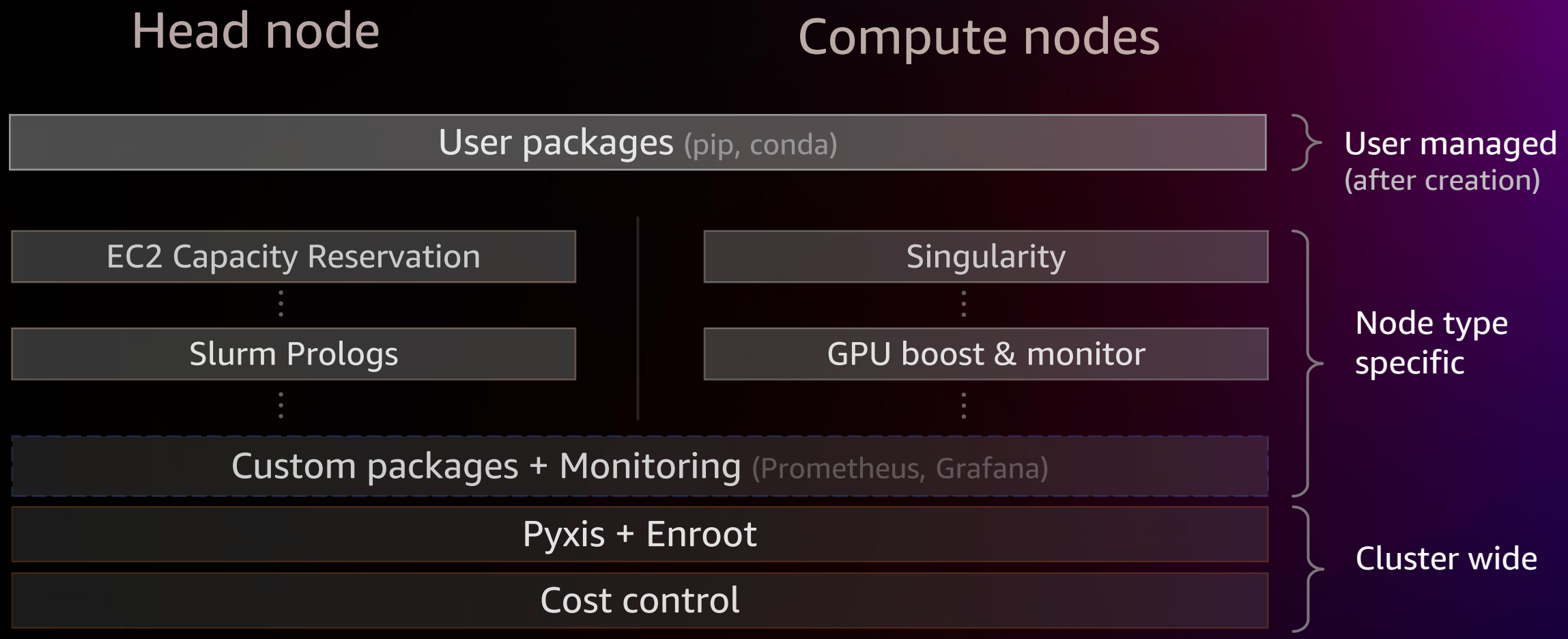
- Central user management using AWS Active Directory with AWS ParallelCluster
- Jobs controls, usage across regions & users

Multi-instances job queues

- Training with p4d.24xlarge & Trn1.32xlarge
- Preprocessing on C6i, M6i & R6i



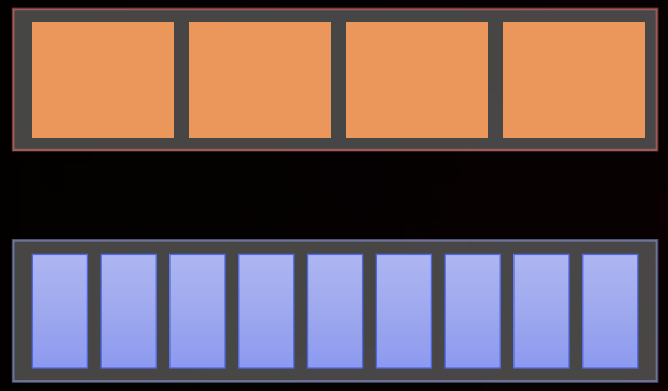
Software stack management: post-install scripts

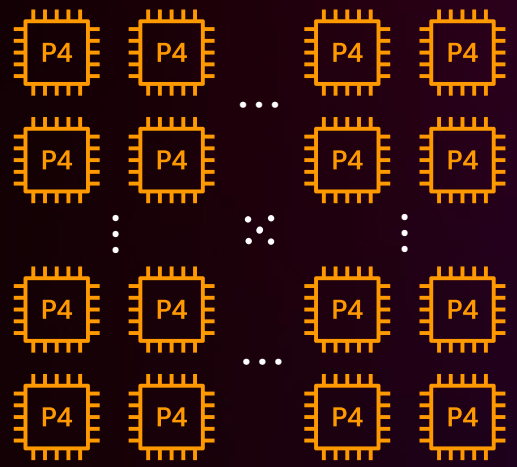


Jobs pre-emption and QoS

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

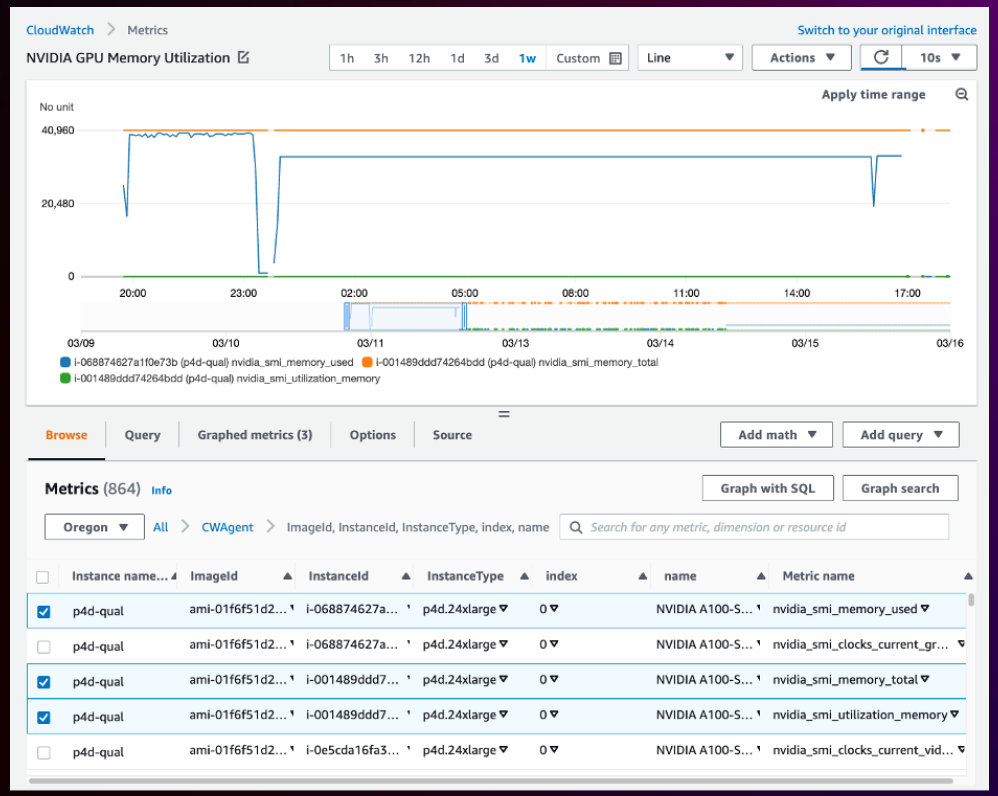
Job level

- 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 ✦



✦ https://github.com/Stability-AI/stability-hpc/blob/main/service/prolog.sh
✦✦ https://github.com/Stability-AI/stability-hpc/blob/main/scripts/debug_prolog.sh
* https://aws.amazon.com/blogs/machine-learning/monitoring-gpu-utilization-with-amazon-cloudwatch/
** https://aws.amazon.com/blogs/compute/capturing-gpu-telemetry-on-the-amazon-ec2-accelerated-computing-instances/

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