

Expedite Model Loading for Training & Inference

Sobue Kosuke

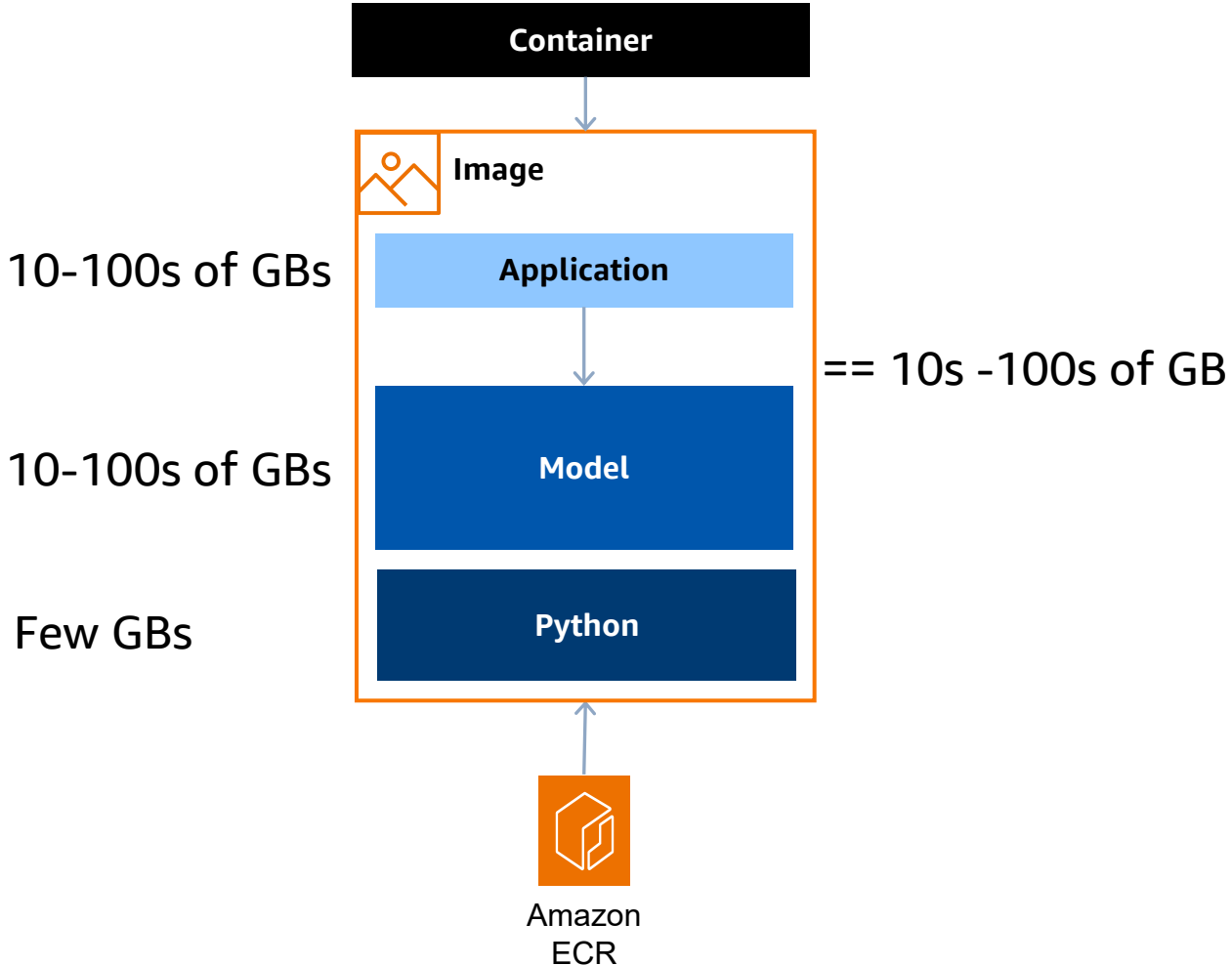
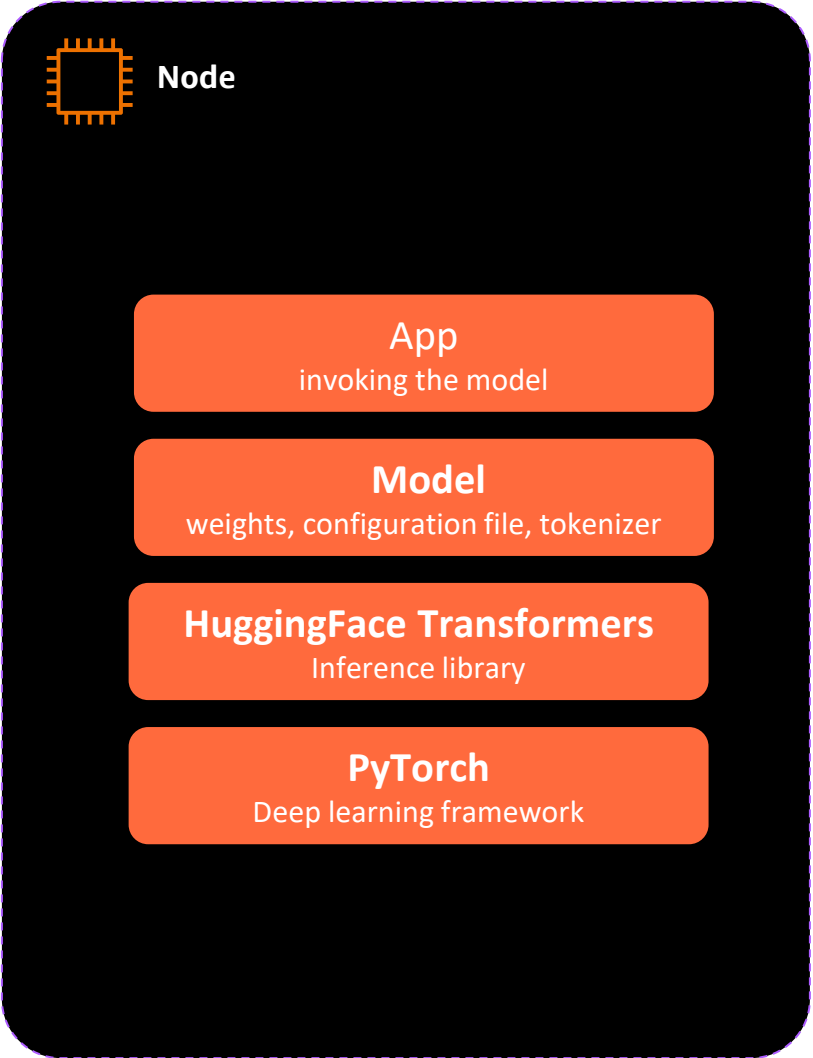
Solutions Architect

Amazon Web Services

Welcome

the session will begin shortly

Challenges – Startup Times



Model Artefacts & Storage Options

 **Hugging Face**

meta-llama/**Llama-3.3-70B-Instruct** 

 like 2.47k  Follow  Meta Llama 56.5k

 Text Generation

 Transformers

 Safetensors

 PyTorch

 8 languages

 llama

 facebook

 Model card

 **Files and versions**  xet

 Community 132

 main 

Llama-3.3-70B-Instruct

 Hamid-Nazeri  pcuenq  Model card updates (#28) 6f6073b 

 original

 .gitattributes  1.57 kB 

 LICENSE  7.91 kB 

 README.md  36.5 kB 

 USE_POLICY.md  6.02 kB 

 config.json  879 Bytes 

 generation_config.json  189 Bytes 

 model-00001-of-00030.safetensors   4.58 GB  xet 

 model-00002-of-00030.safetensors   4.66 GB  xet 

 model-00003-of-00030.safetensors   5 GB  xet 

 model-00004-of-00030.safetensors   4.97 GB  xet 

Block

File

Object



Amazon EBS



Amazon EFS



Amazon FSx for Lustre



Amazon FSx for NetApp ONTAP



Amazon FSx for OpenZFS



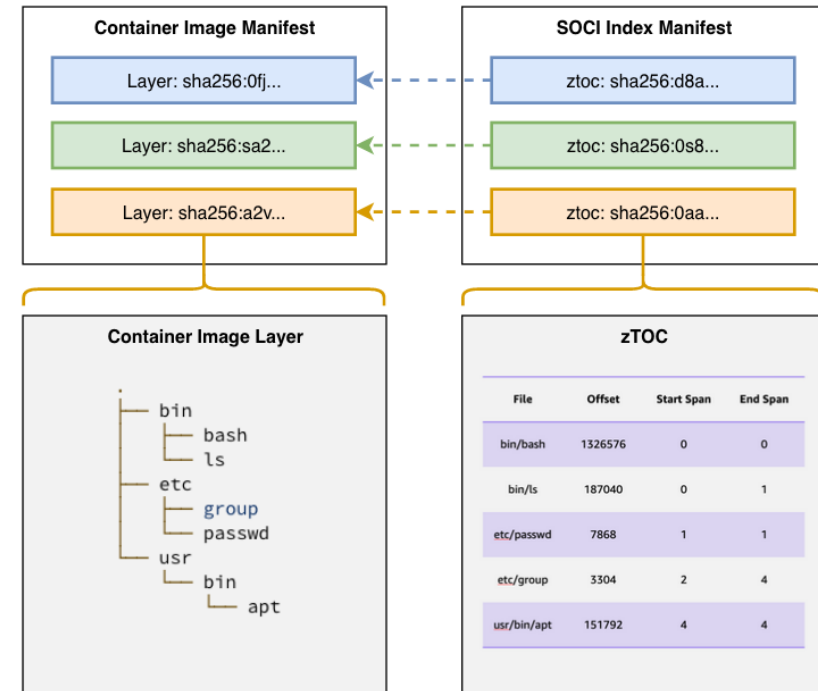
Amazon S3 (Mountpoint)

Container storage interface (CSI)

Introducing Seekable OCI (SOCI)

Seekable OCI (SOCI) is a technology open sourced by AWS that enables containers to launch faster **by expediting the container image pulling time.**

- SOCI Parallel Pull mode: **pulls the layer with huge size in parallel chunks**
- SOCI Index & lazy loading: **lazily loading the container image**



SOCI experiments & Results

SOCI Index & Lazy Loading

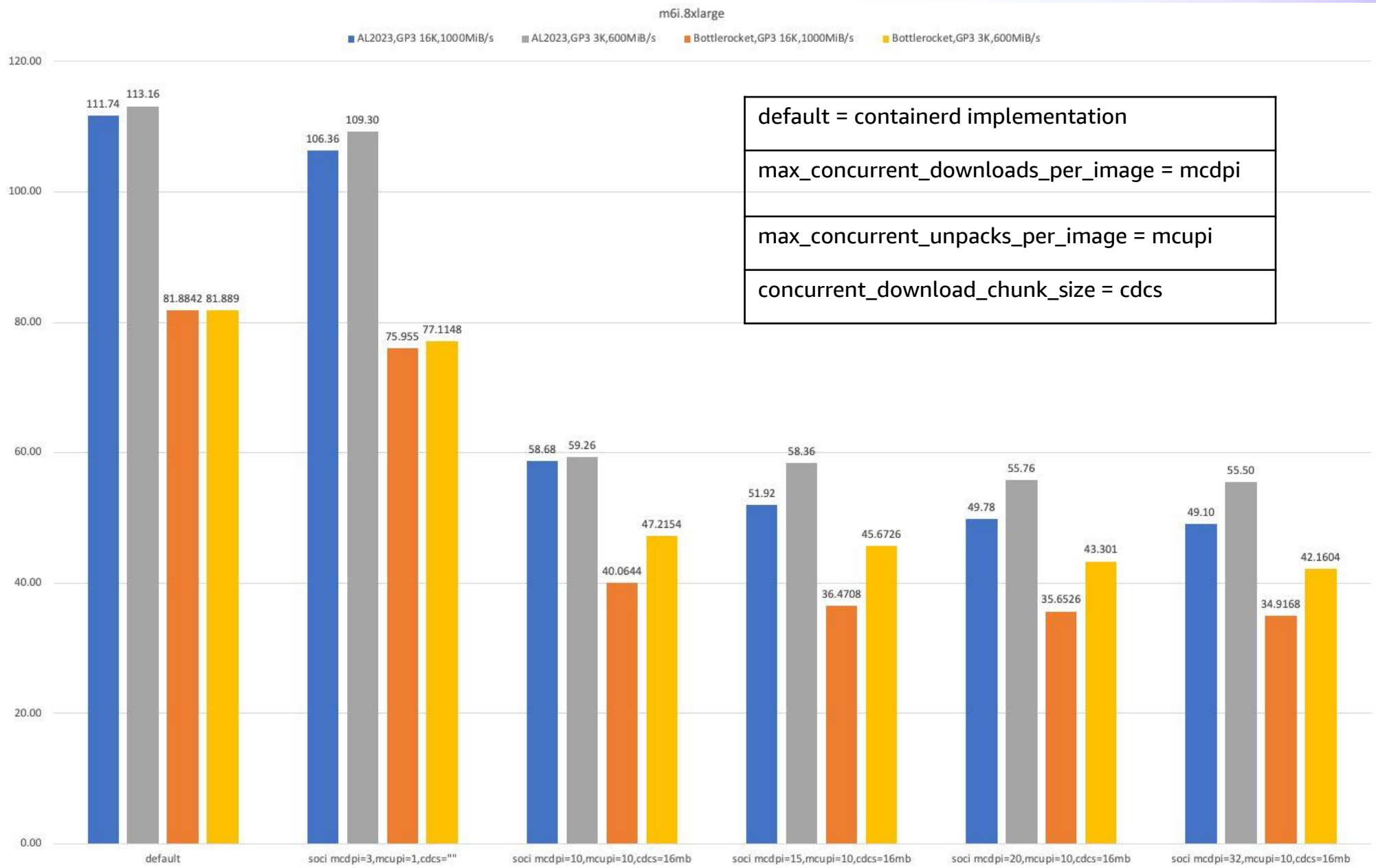
☐	Image tags ▾	Type	Created at ▾	Image size ▾
☐	-	<u>Soci Index</u>	08 February 2026, 10:11:09 (UTC-00)	89.10
☐	latest	<u>Image</u>	08 February 2026, 10:03:00 (UTC-00)	3806.64

SOCI Parallel Pull

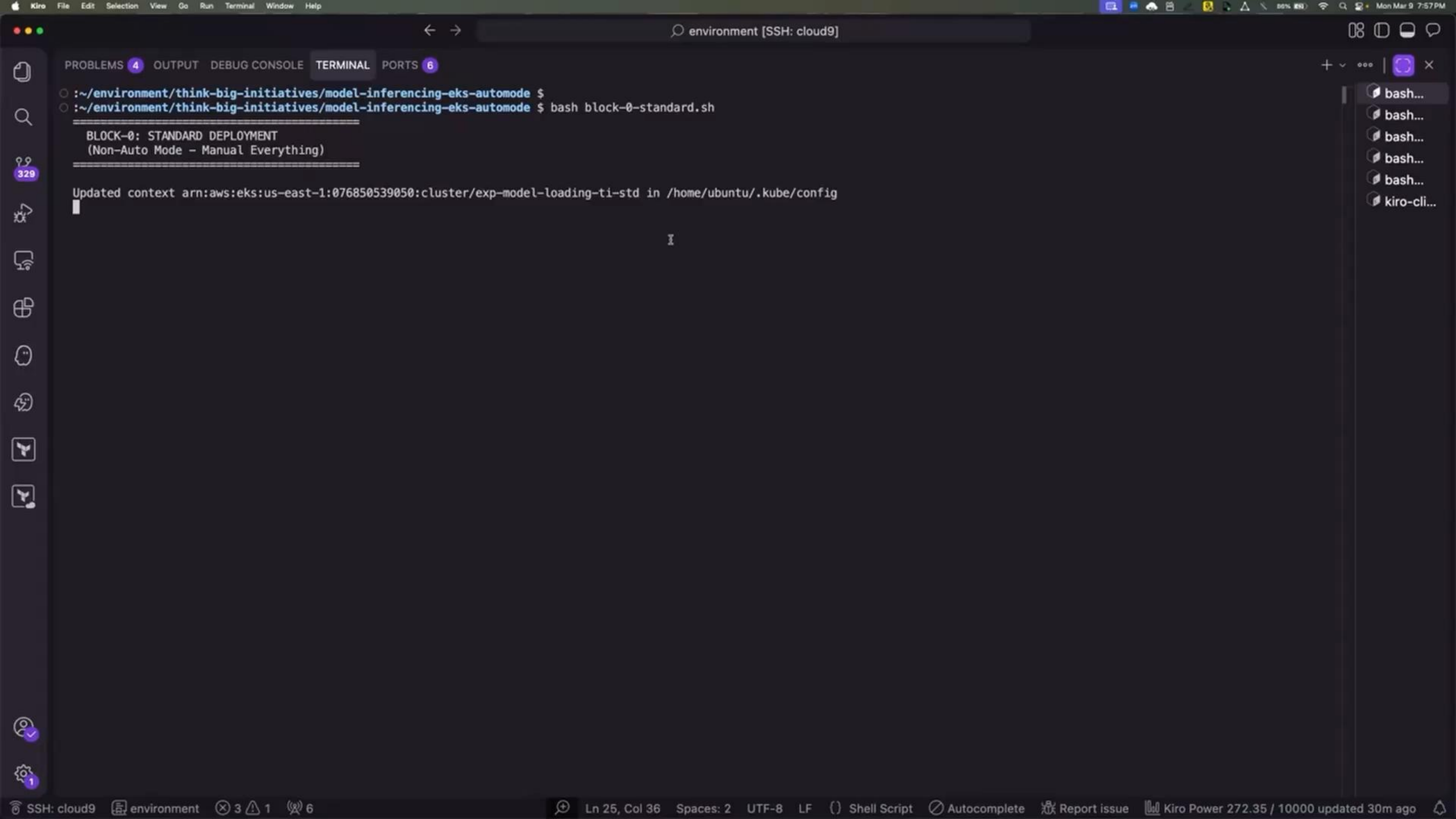
"...dkr.ecr.us-east-2.amazonaws.com/ecr-public/deep-learning-containers/vllm:0.13-gpu-py312-ec2" in 29.336s (29.336s including waiting). Image size: 9080776548 bytes."

P6-b200, with NVMe and SOCI Parallel Unpack from ECR

SOCI Parallel Pull Benchmarks



Demo



```
~/environment/think-big-initiatives/model-inferencing-eks-automode $  
~/environment/think-big-initiatives/model-inferencing-eks-automode $ bash block-0-standard.sh  
  
=====  
BLOCK-0: STANDARD DEPLOYMENT  
(Non-Auto Mode - Manual Everything)  
=====  
  
Updated context arn:aws:eks:us-east-1:076850539050:cluster/exp-model-loading-ti-std in /home/ubuntu/.kube/config  
|
```

Additional Resources

EKS Best Practice Guide for AI/ML	Best practices when running AI/ML workloads on EKS can ensure that those workloads are performant, cost-effective, resilient, and properly resourced. Divided into the following general sections: Compute, Networking, Storage, Observability, and Performance	
AI on EKS	AI on EKS provides deployment-ready blueprints facilitating training, fine-tuning, and inference of multiple large language models (LLMs), infrastructure as code (IaC) templates to enable reference architectures	
Real Time Inference on EKS Tutorial	Getting started guidance on building optimized clusters with GPU-accelerated nodes, integrating AWS services for storage and autoscaling, deploying sample models for validation, and key architectural considerations such as decoupling CPU and GPU tasks, selecting appropriate AMIs and instance types, and ensuring low-latency exposure of inference endpoints	
Register to attend a workshop	Attend a virtual hands-on workshop for running inference workloads on EKS, led by one of our Solution Architects. Held each month across different time zones.	

Thank you!

セッションをご視聴いただき有難うございました。
お近くのブーススタッフまで QR コードのご提示をお願いいたします。
Thank you for attending this session. Please show your QR code to the nearest booth staff.

Don't forget to..



Demo stations: Check out the rest of the AWS booth #700 with 4 demo stations and a Sketch Lab. Connect with AWS experts and ask your hardest questions!



Hands-on workshops: Register for online Kubernetes workshops with AWS subject matter experts covering AI/ML, platform engineering and much more...



Public roadmap: Stay up to date with what we're working on, give us feedback and propose ideas to influence EKS roadmap



Best Practices Guide: Master your Kubernetes operations and run production workloads with greater security, reliability, and efficiency

<http://github.com/aws/containers-roadmap>

<https://docs.aws.amazon.com/eks/latest/best-practices/introduction.html>

Expedite Model Loading for Training & Inference

Ramesh Mathikumar

Pr. Delivery Consultant
Amazon Web Services

Olawale Olaleye

Senior GenAI/ML Specialist SA
Amazon Web Services

Sundar Shanmugam

Sr. Delivery Consultant
Amazon Web Services

The Baseline- Standard EKS Without Auto Mode

- Standard managed node groups
- Manual SOCI installation required
- Full image downloads
- Sequential layer downloads, single-threaded pull



Result: Longer image pull time

EKS Auto Mode + SOCI

- Simplified cluster config and management
- SOCI pre-configured on selected instances
- Parallel layer downloads (vs sequential)
- Accelerated container starts



Result: Much faster image pull

EKS Auto Mode + Mountpoint for Amazon S3

- S3 Mountpoint CSI Driver (PersistentVolume → S3 bucket)
- Stream model weights on-demand
- No local downloads
- Multiple pods share same data



Result: Faster pod startup, lower storage costs

EKS Auto Mode + Run:ai Model Streamer

- Stream directly from S3 to GPU memory
- Bypass CPU and local storage
- Concurrent multi-threaded streaming
- Fastest model loading path



Result: Fastest possible model loading to GPU

Performance & Recommendations

✅ **START HERE: EKS Auto Mode + SOCI**

- Significant cold start reduction
- Zero setup effort (SOCI pre-configured)
- Strong ROI for all ML workloads

📦 **ADD: S3 Mountpoint IF:**

- Large models (>10GB)
- Multiple pods need same model
- Want to reduce storage costs

🚀 **OPTIMIZE: Block 3 (Run:ai Model Streamer) FOR:**

- GPU-intensive inference at scale
- Need absolute fastest model loading
- Multi-tenant GPU sharing required

Thank you!