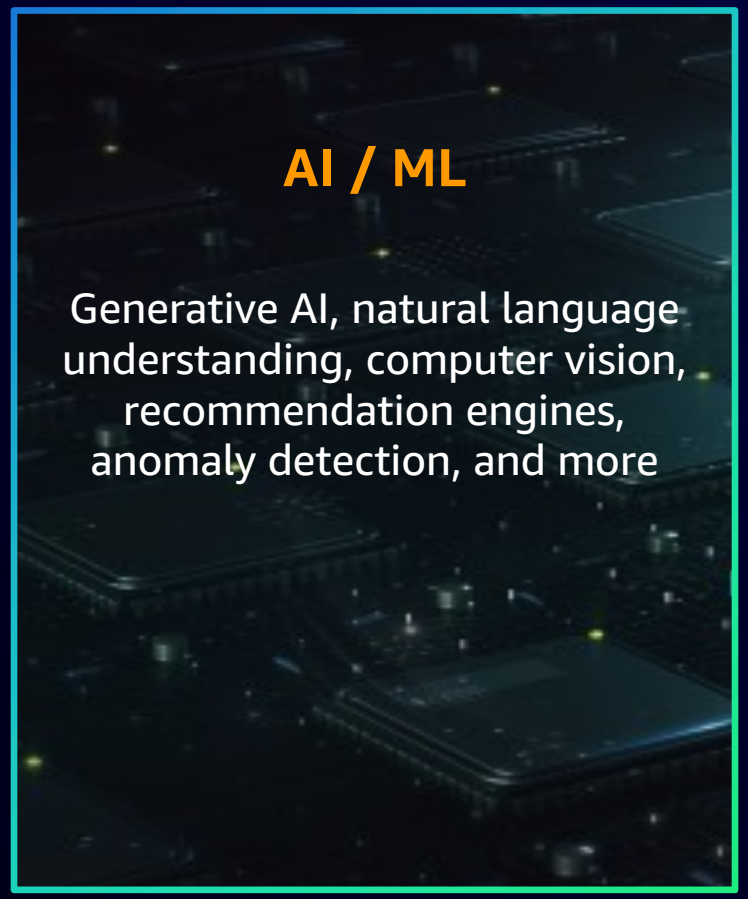
S63058

How Genius Sports transforms NFL game viewing with accelerated computing on AWS

Shruti Koparkar
(she/her)
Senior Product Marketing Manager
Amazon Web Services

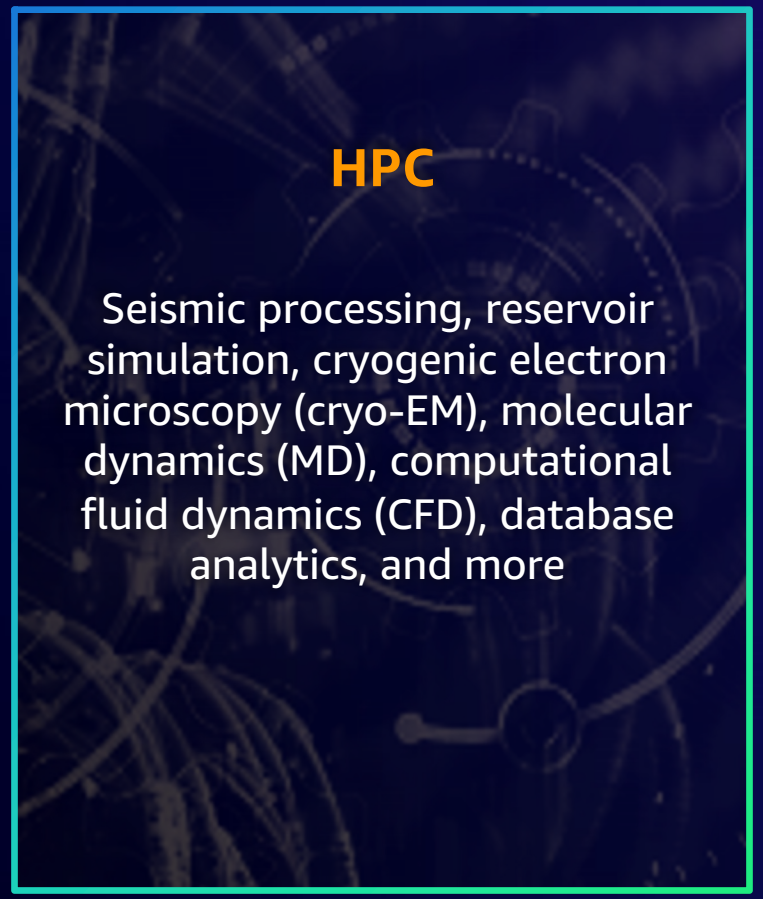
Sheldon Kwok
(he/him)
Senior Director of Infrastructure
Second Spectrum by Genius Sports

Accelerated computing applications



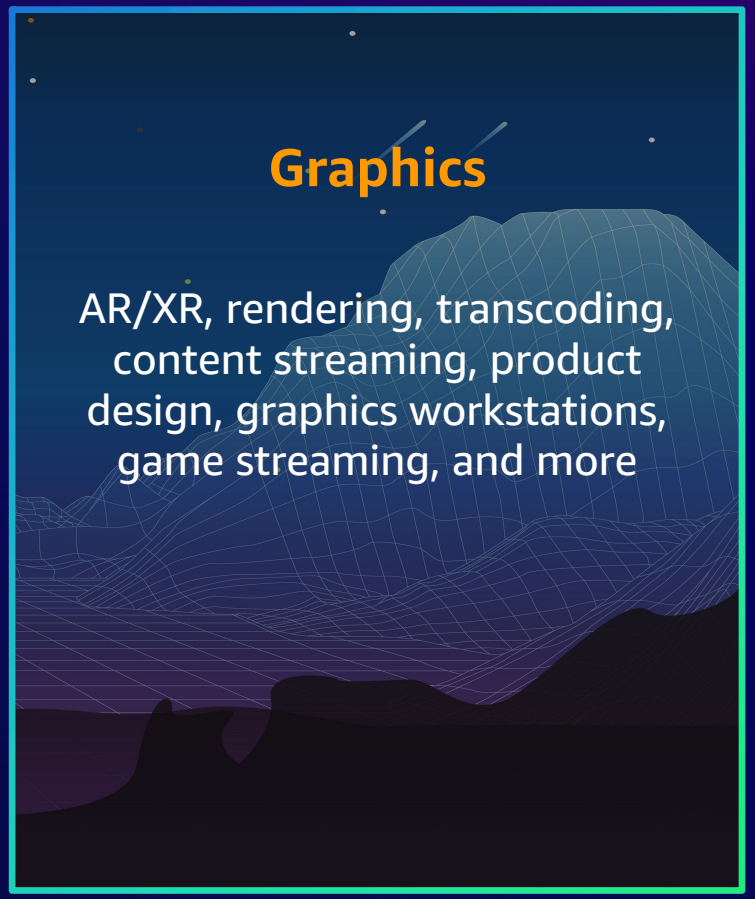
AI / ML

Generative AI, natural language understanding, computer vision, recommendation engines, anomaly detection, and more



HPC

Seismic processing, reservoir simulation, cryogenic electron microscopy (cryo-EM), molecular dynamics (MD), computational fluid dynamics (CFD), database analytics, and more



Graphics

AR/XR, rendering, transcoding, content streaming, product design, graphics workstations, game streaming, and more

Accelerate your generative AI journey

Easiest way to build

with leading foundation models

Differentiate with your data

in a secure and private environment

Increase productivity

with generative AI applications and services

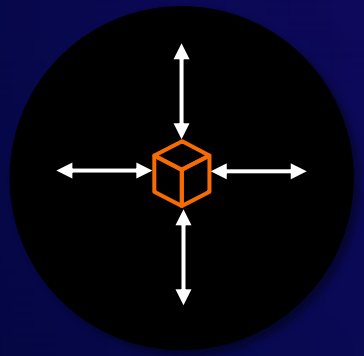
Most performant, low cost

infrastructure to scale ML and generative AI

AI/ML infrastructure: Key customer needs

Performance

Cost



Scalability

Availability and Ease of use

The best place to run NVIDIA GPUs

EC2 Instances	ML TRAINING & INFERENCE				GRAPHICS & ML INFERENCE	
	P5	**P4de**	**P4d**	**P3**	**G5**	**G4dn**
	H100	A100	A100	V100	A10G	T4

Enhanced security and performance with AWS Nitro

COMPATIBLE WITH:

ML frameworks & open source

- PyTorch
- TensorFlow
- HuggingFace
- JAX

FULL STACK OPTIMIZATIONS:

Storage/Networking

- Amazon EFS
- Amazon S3
- Amazon FSx for Lustre
- EFA & UltraClusters

AVAILABLE THROUGH:

Managed Services

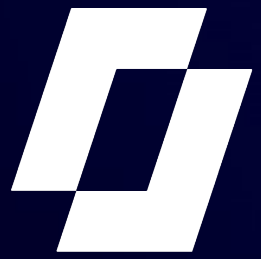
- Amazon SageMaker
- Amazon EKS
- Amazon ECS
- AWS Batch
- AWS ParallelCluster

Up to 20 Exaflops in a single cluster:
3200 Gbps networking and 20k GPUs/cluster

EC2 Capacity Blocks for ML

- Specify cluster size, future start date and duration
- Provides reliable, predictable access to P5/P4d instances (H100/A100 GPUs)
- Collocated in Amazon EC2 UltraClusters

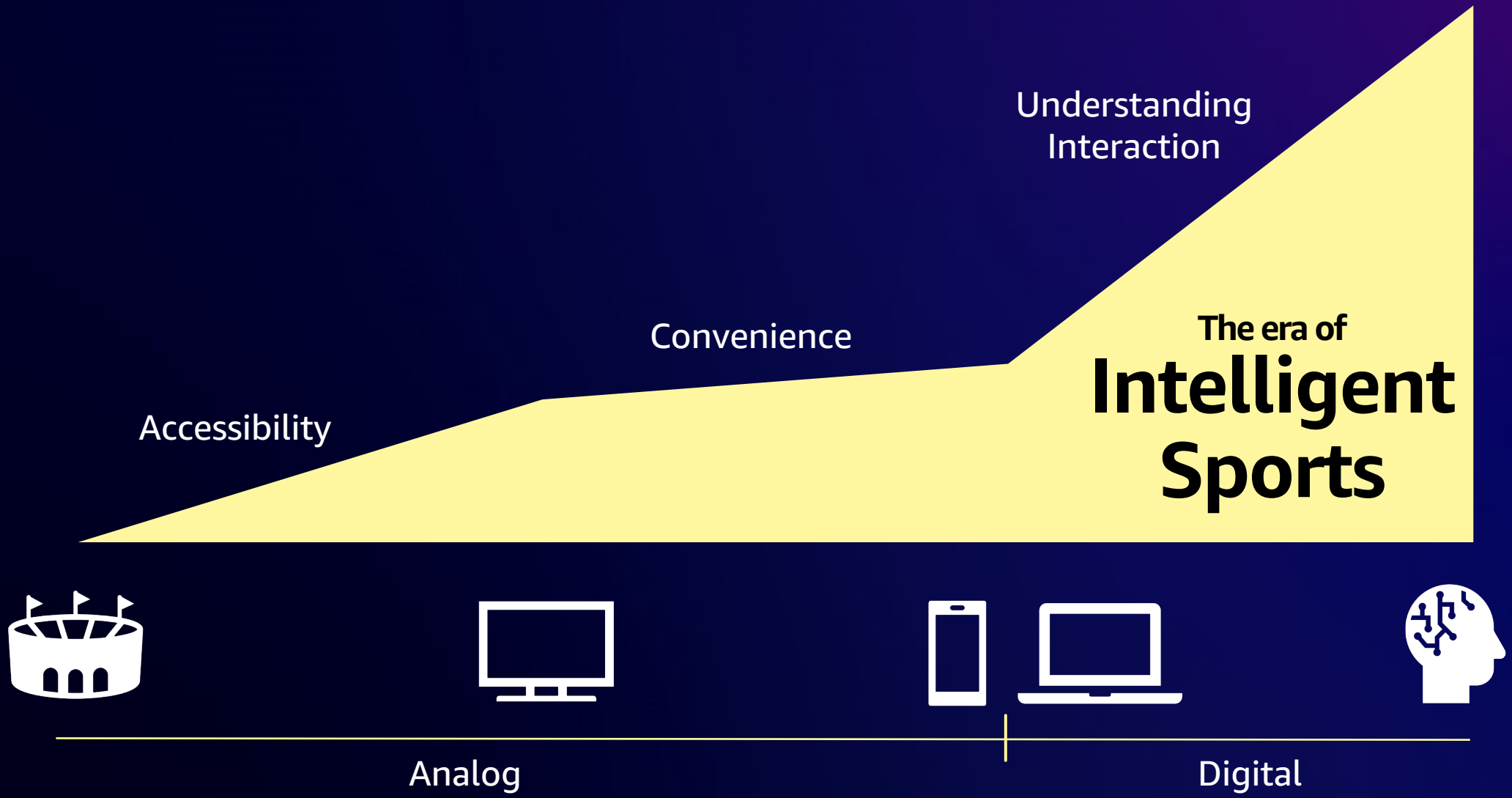
More than 100,000 customers use AWS for ML

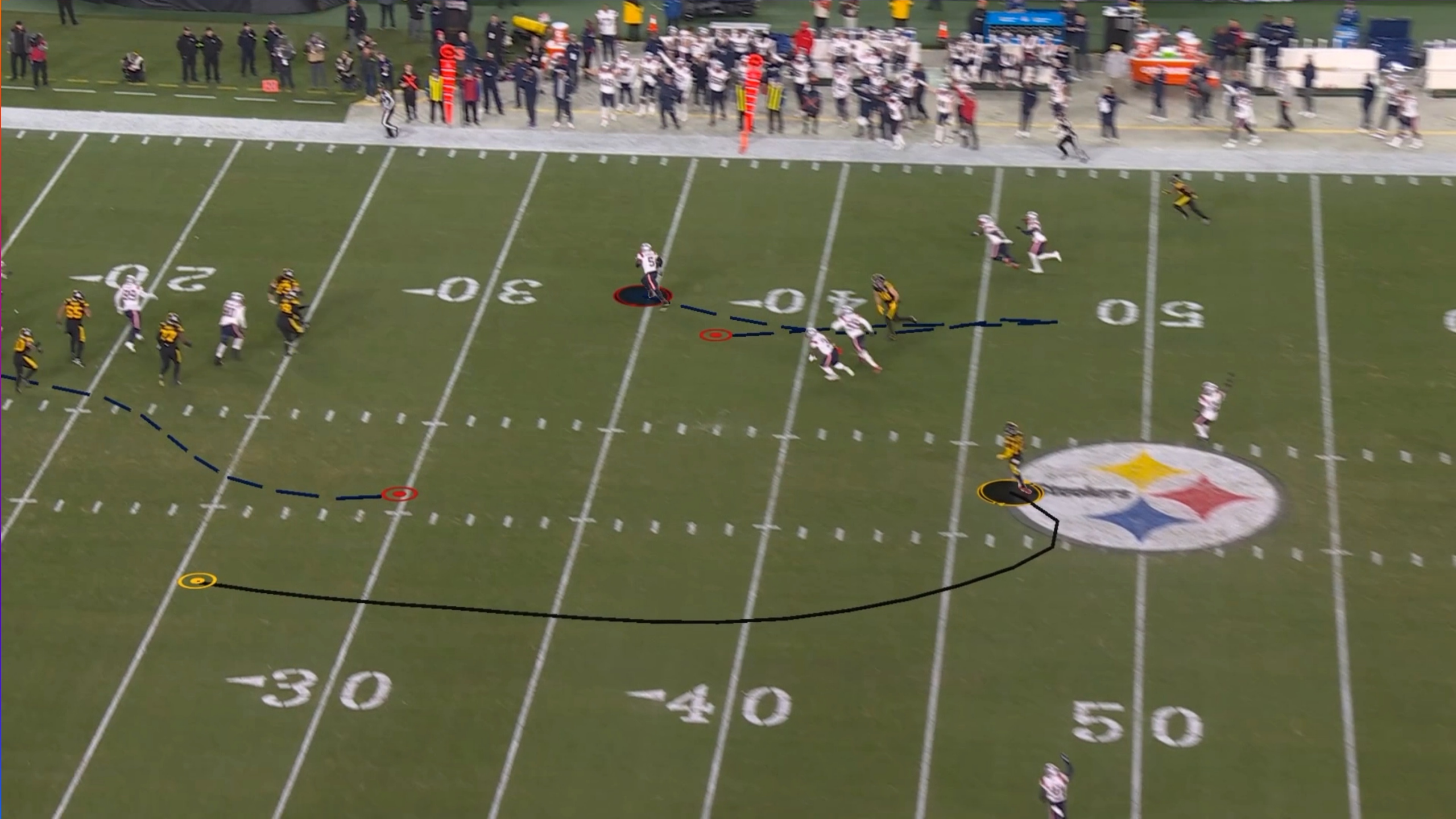


Second Spectrum

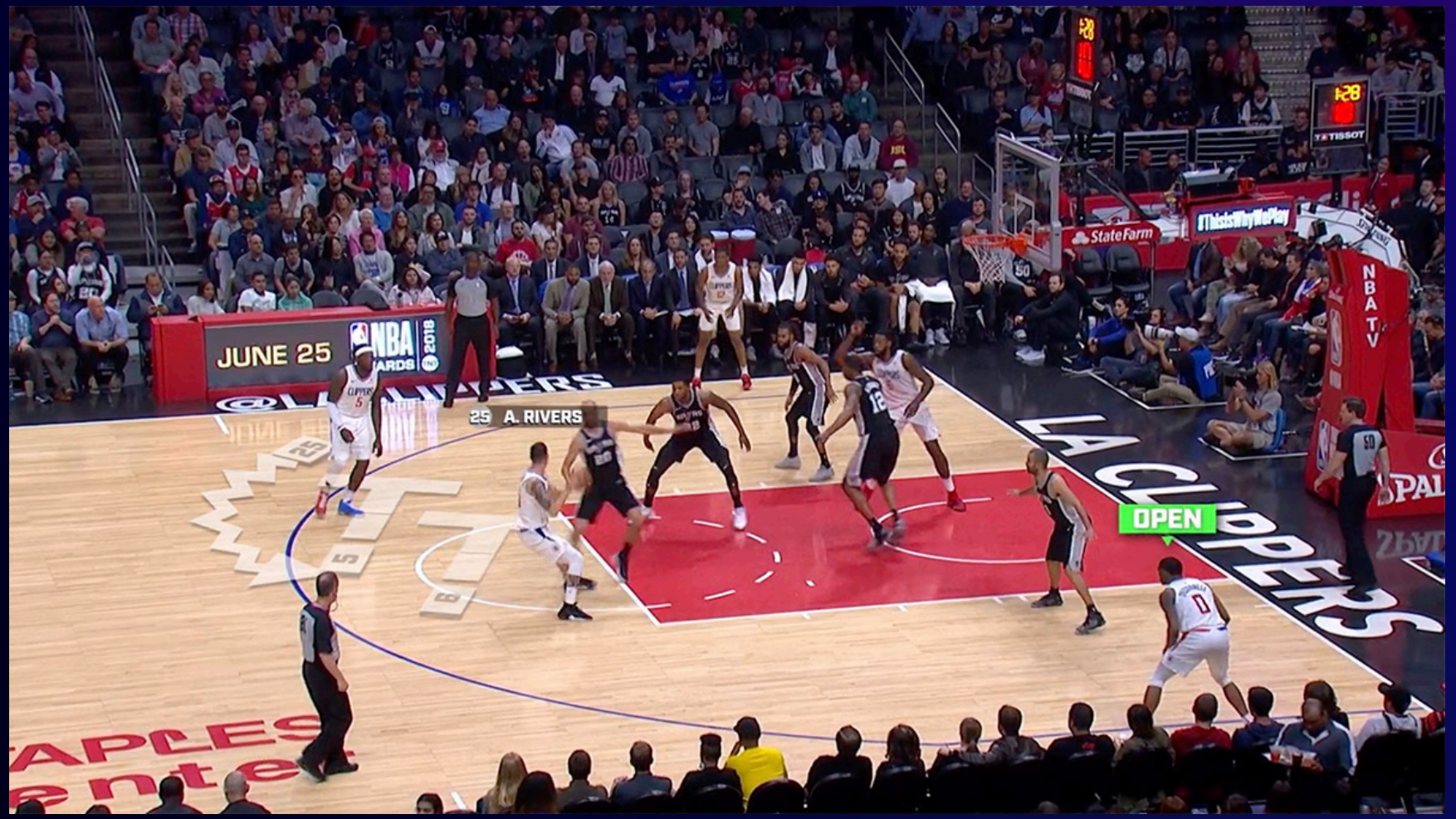
by Genius Sports

The Operating System of Sports



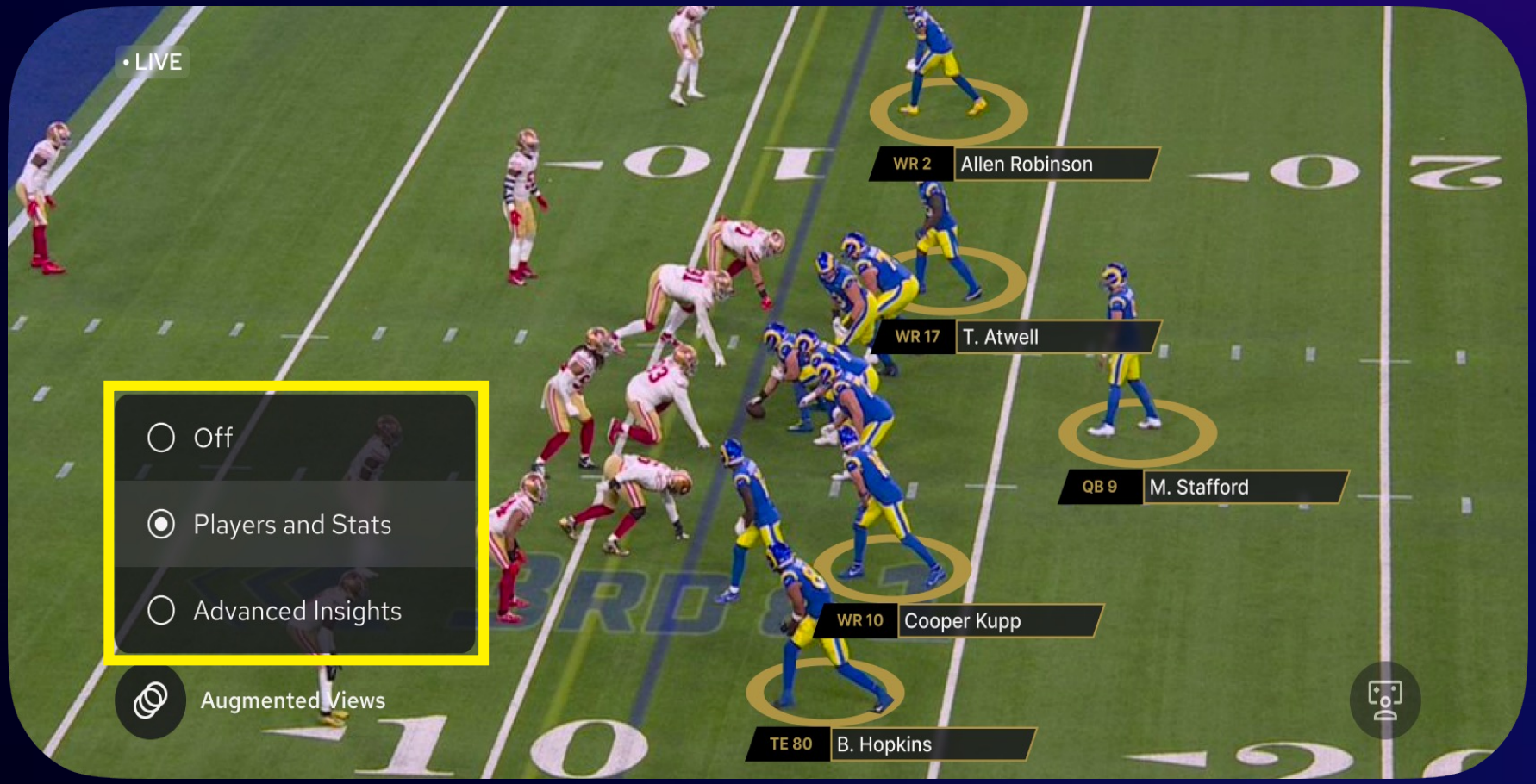

-30
-40
50

First Proof of Concept – 2 minutes

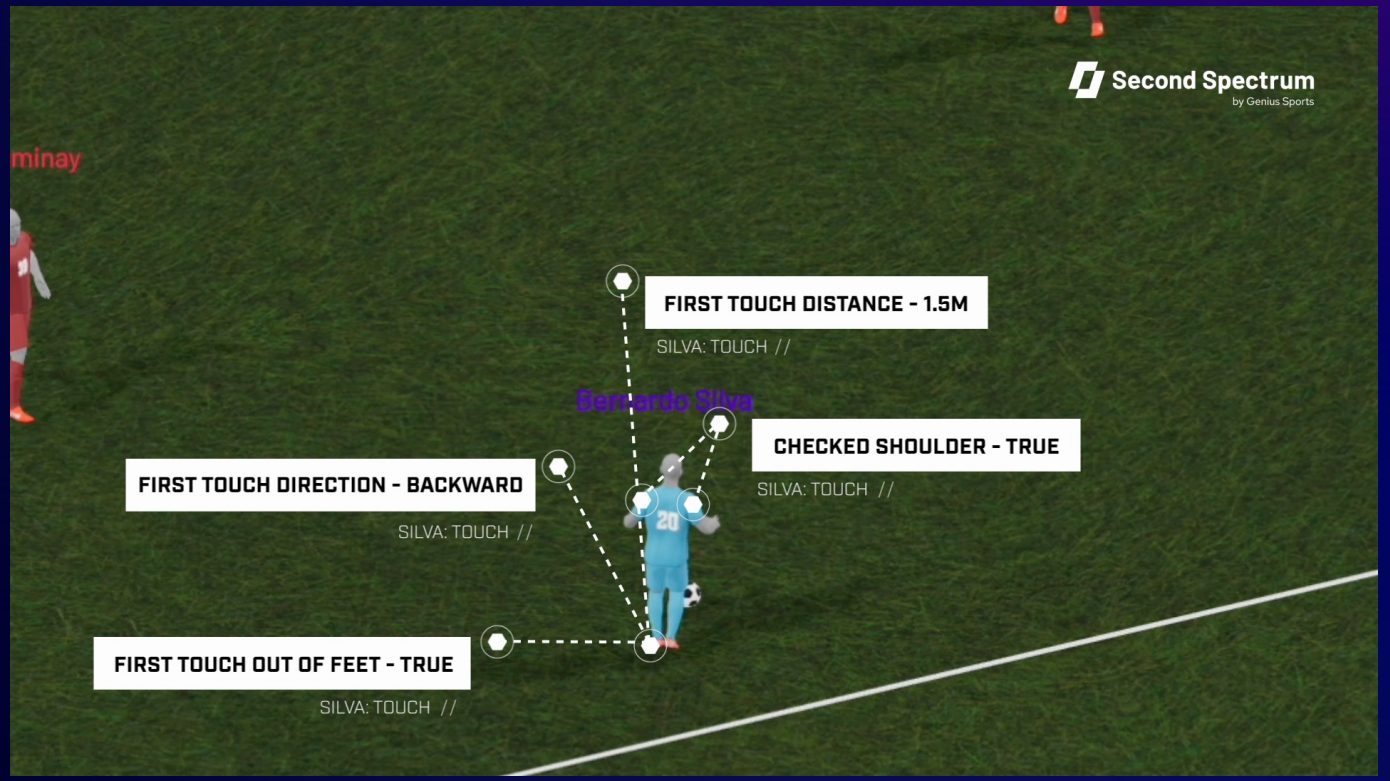


Live Alternative Streams – 30 seconds

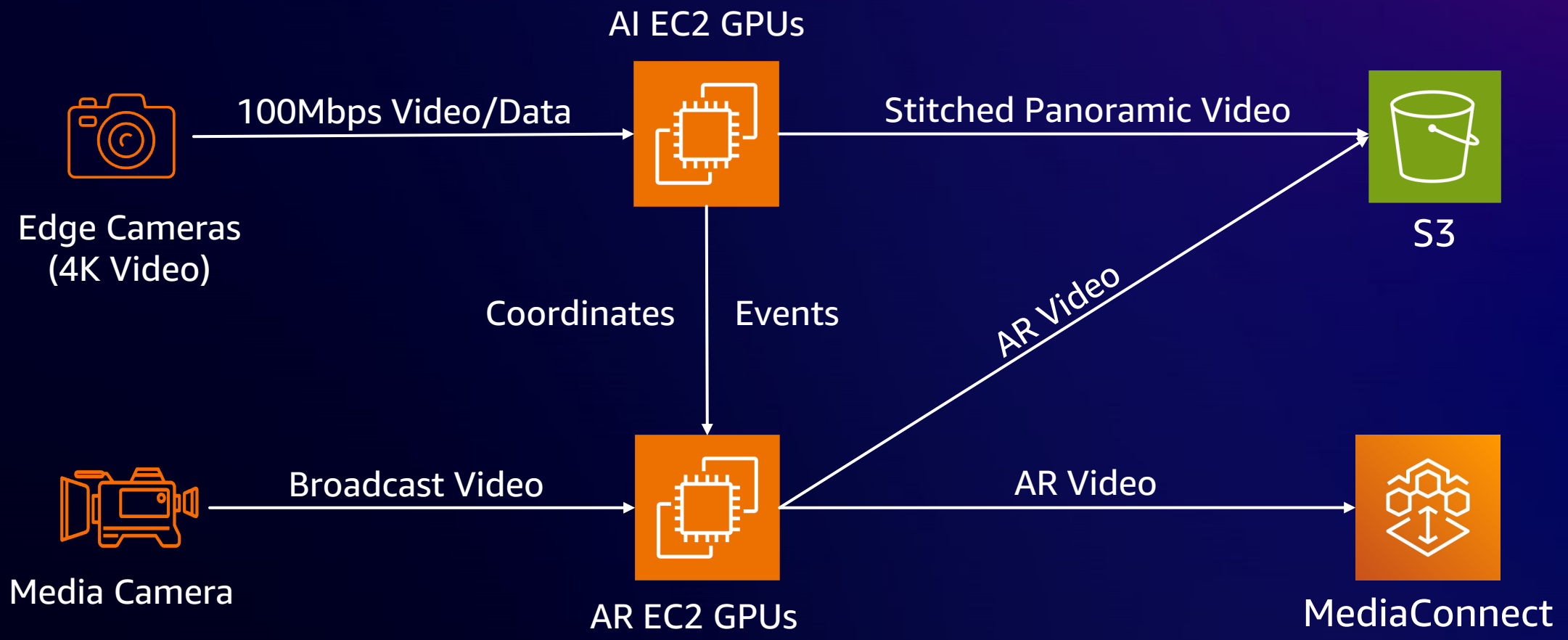
Realtime - 3 to 5 seconds



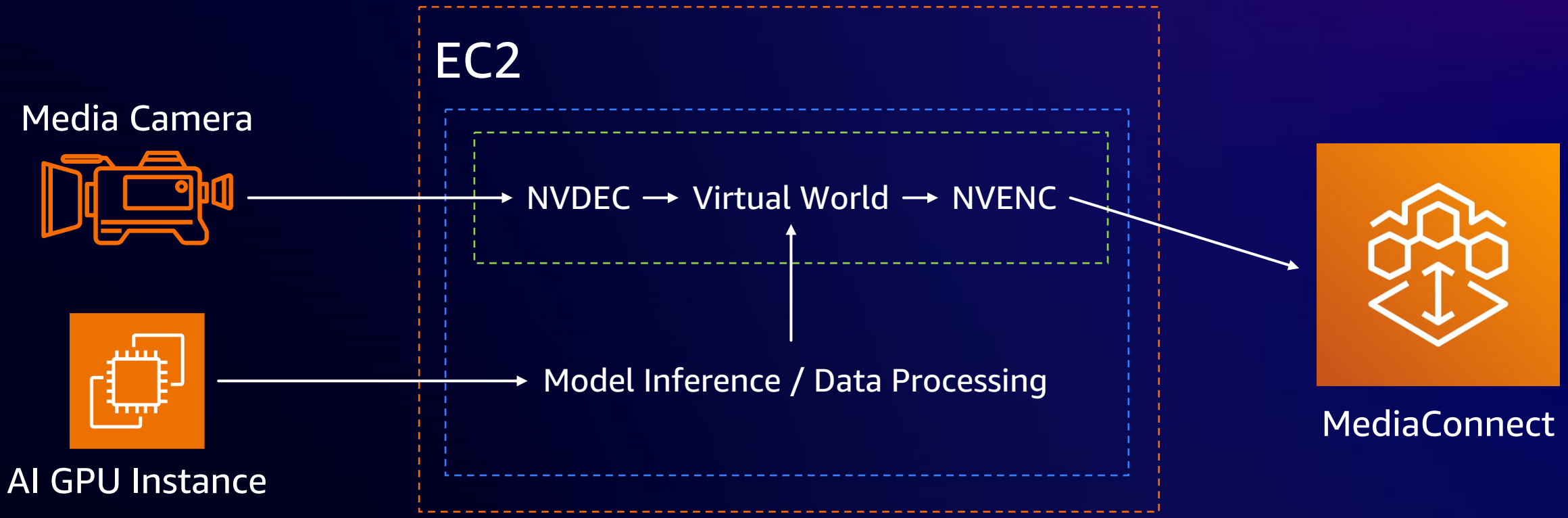
How It Works

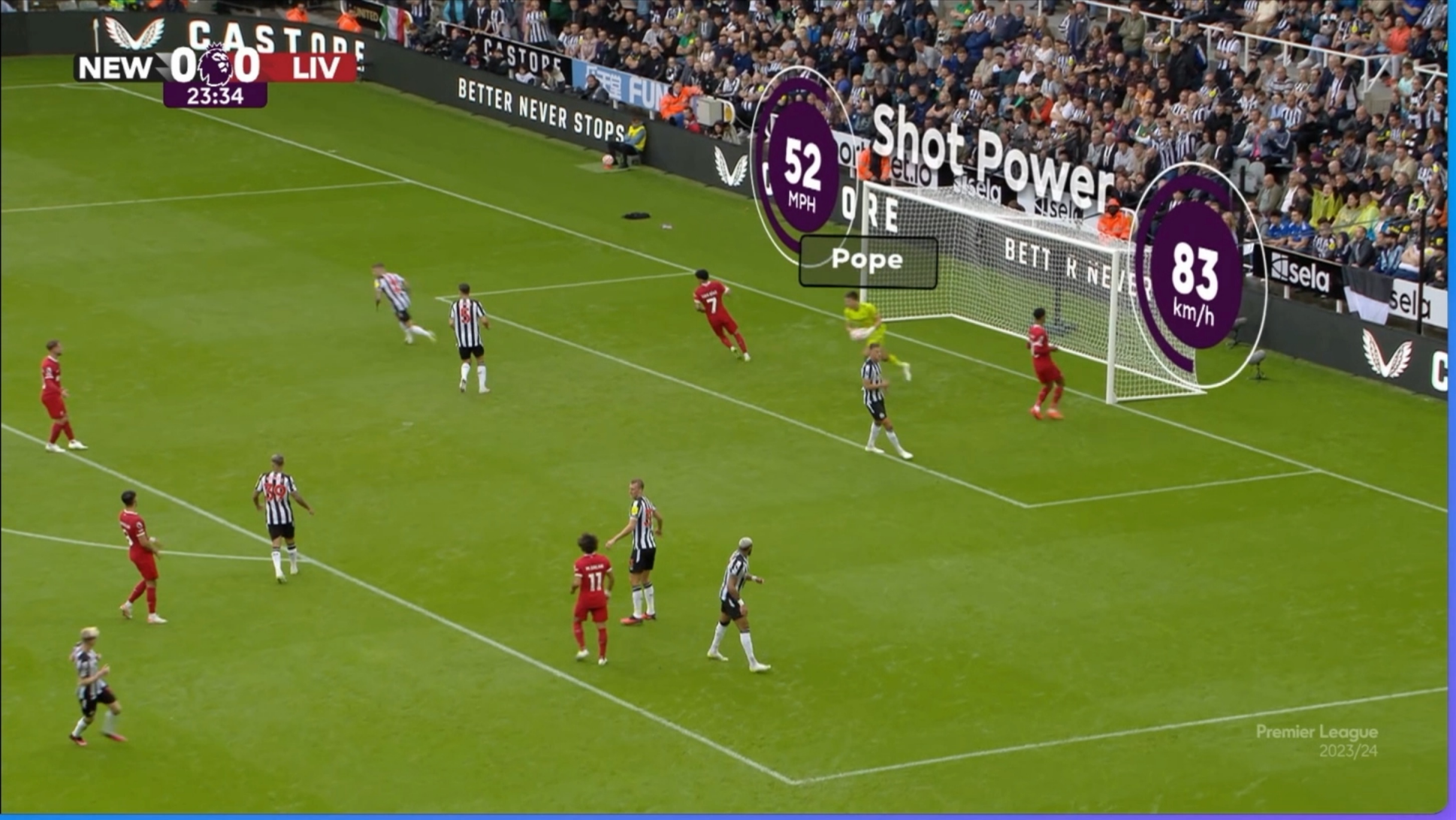


AI/AR Pipeline



AR Pipeline





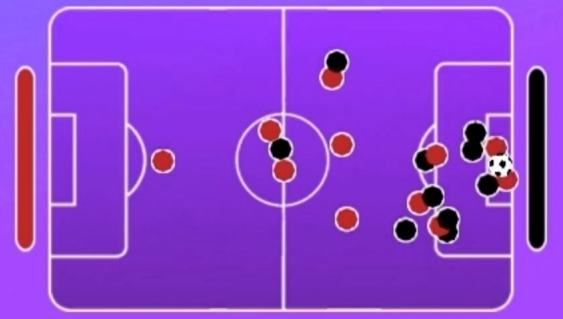
Data Zone
Top Speed (mph)

Newcastle United	
Gordon	20.0
Almirón	18.8
Tonali	18.4

Liverpool	
Szoboszlai	19.6
Salah	18.1
Alexander-Arnold	17.9

Match Stats Team Intensity

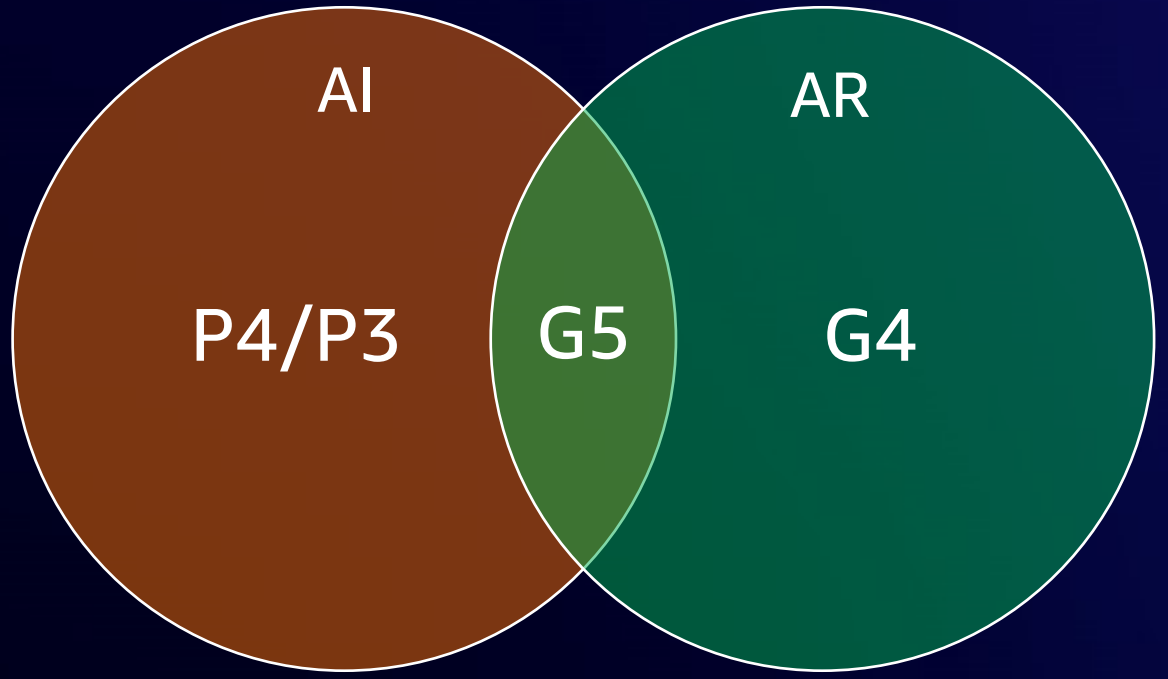
17.2	Total Distance (mi)	16.8
32	Sprints	25

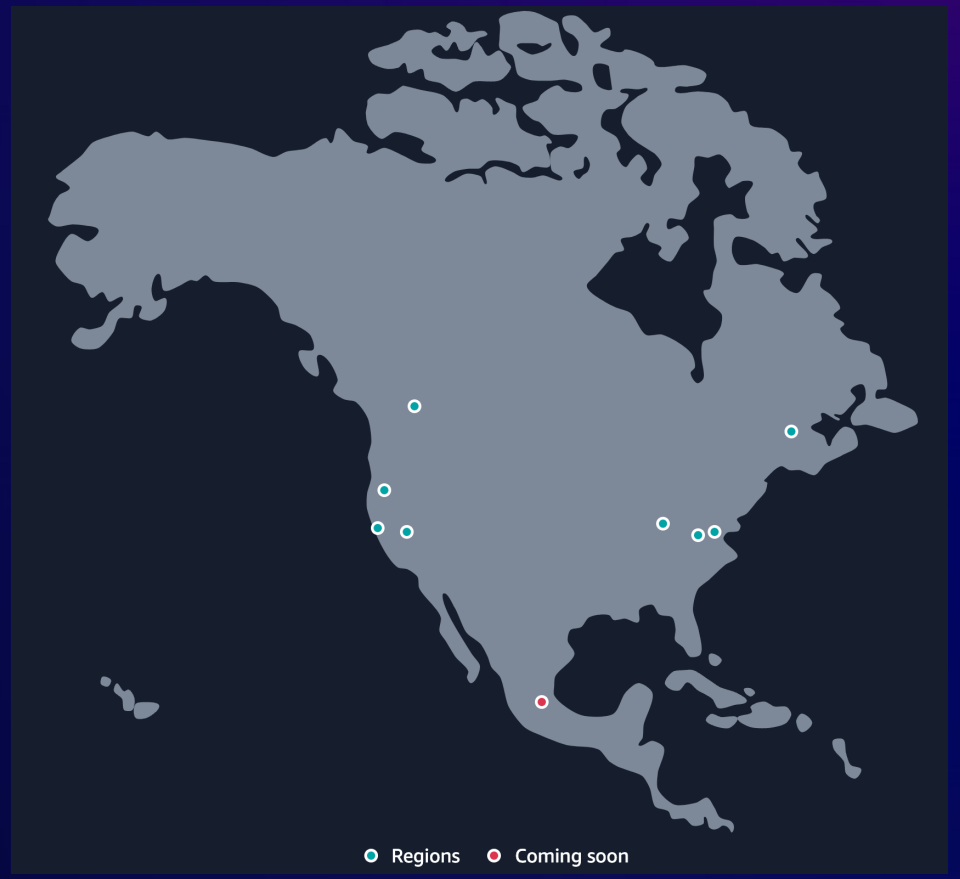


Scaling Live GPU Usage

- Sporting events require a large amount of compute (g5.48xl or p3.16xl)
- Games require pre-scaling at high parallelism
- Outside of games requires no GPU usage
- Difficult to acquire GPUs

Increase Optionality

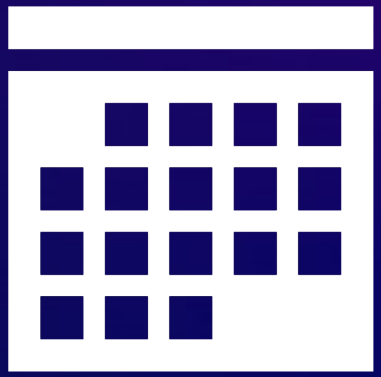




Plan Ahead

- Reserve ahead – On Demand Capacity Reservations
  - Accumulate instances earlier
  - Spot placement scores

- Work with your technical account manager
  - Can guide you about where to find more capacity
  - Can request more instances to be provisioned



Scale Achieved

- Capture and process PBs of video with GPUs
- Deliver live insights on 10,000+ games
- Augment 1,000+ video streams

AI Development

Orchestration - EKS

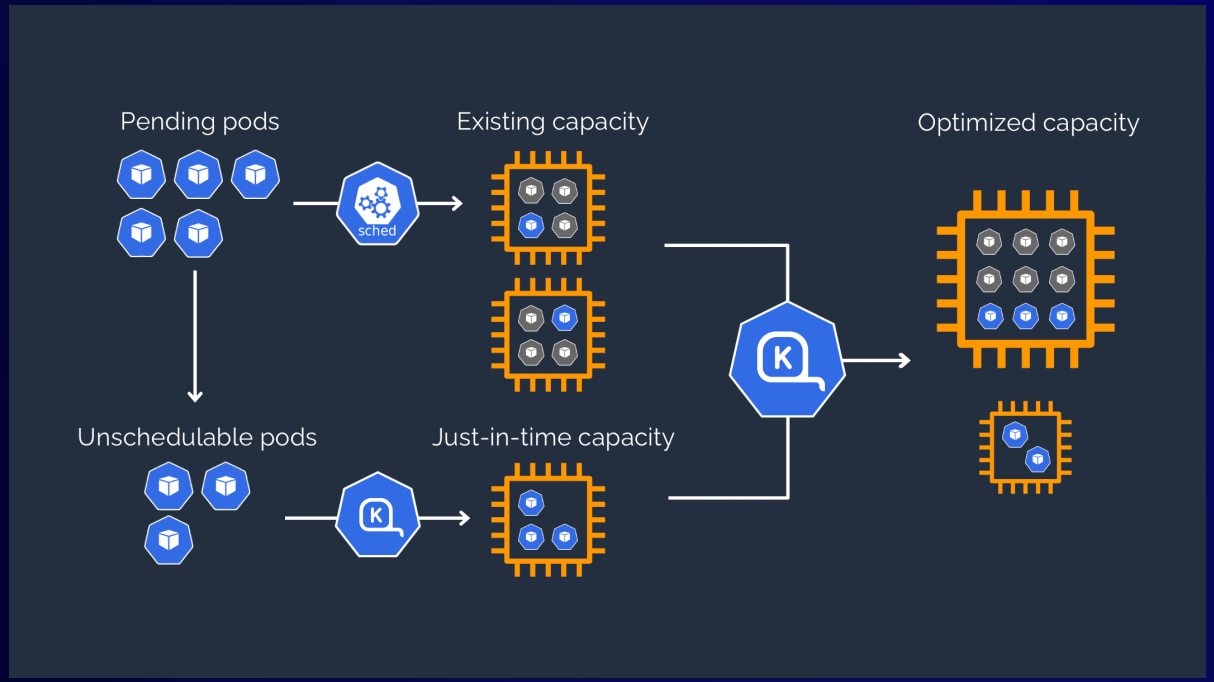
- Simplest way to manage Kubernetes on AWS
- Easy to deploy open source MLOps tooling
- Amazon managed images with Nvidia drivers

Coder

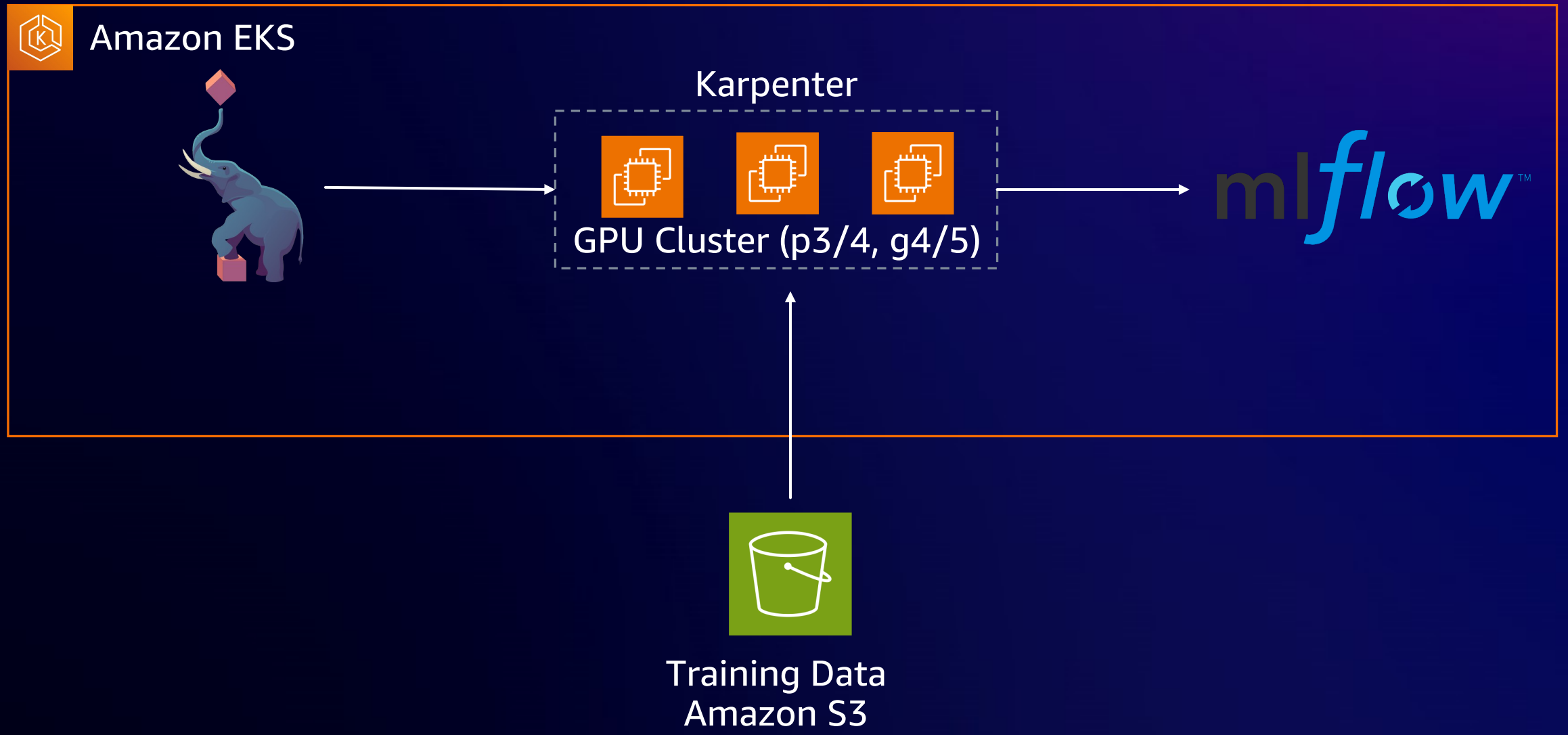
mlflow™

Training Resources - Karpenter

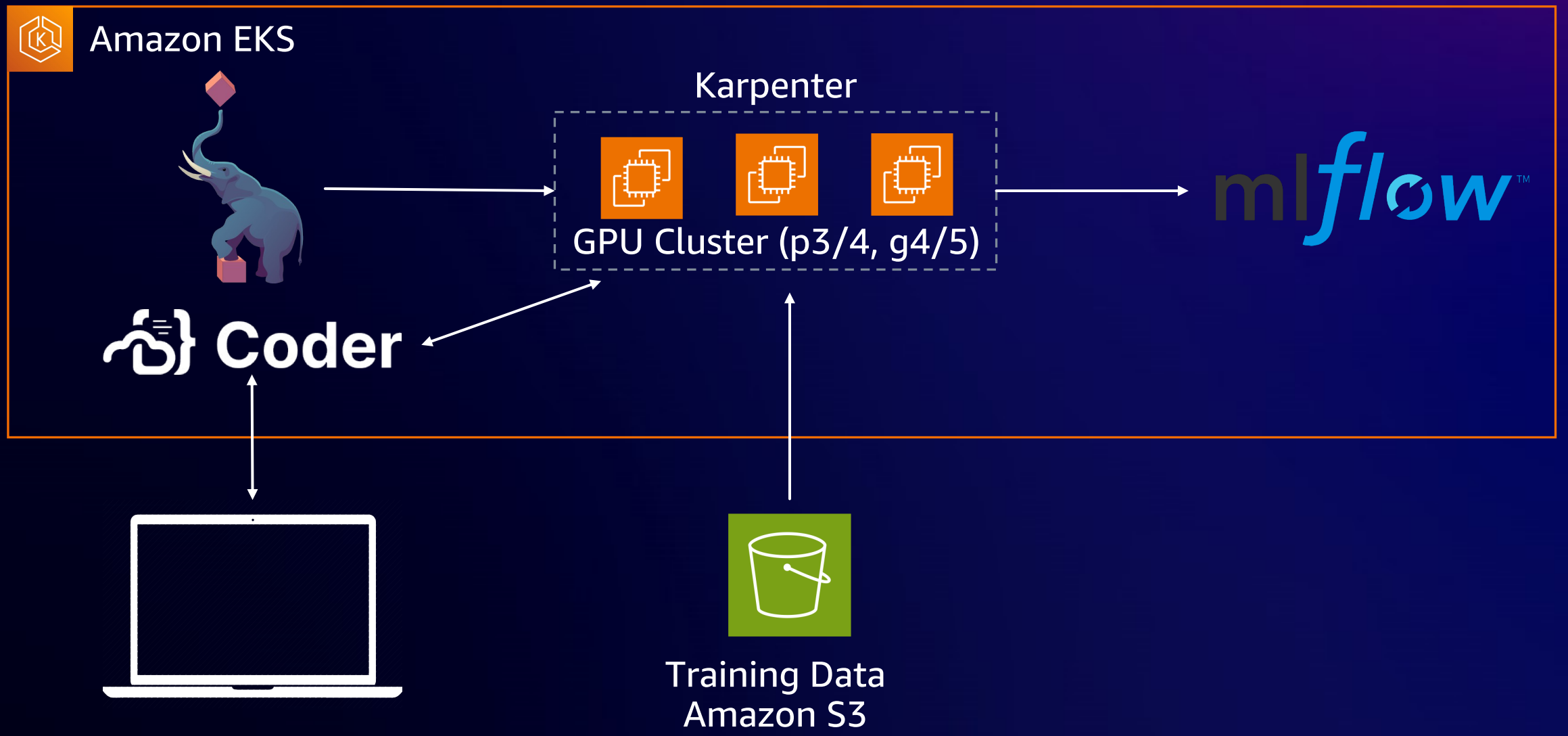
- Abstract underlying instance types from developers
- Pull any instance available with a GPU
- Reduce training times by 50% by scaling into the cloud from our office training machines



ML Training



ML Training/Development



Summary

- Nvidia GPUs are necessary to power real-time applications
- Amazon EC2 can provide a large capacity of GPUs on demand
- GPUs are very popular so must plan ahead
- AWS EKS makes it easy to deploy MLOps ecosystem for AI development

Thank you!