KubeCon | CloudNativeCon

North America 2023

KubeCon | CloudNativeCon
North America 2023

Achieving Real-Time Generative AI During Gameplay for Games Running on Kubernetes

November 9, 2023

Ishan Sharma
Sr. Product Manager,
Google Kubernetes Engine (GKE)
Google Cloud for Games

Google Cloud

Ishan Sharma

Sr. Product Manager,
Google Kubernetes Engine,
Google Cloud

linkedin.com/in/sharmai
twitter.com/IshantheSharma

01 Generative AI - a brief discussion

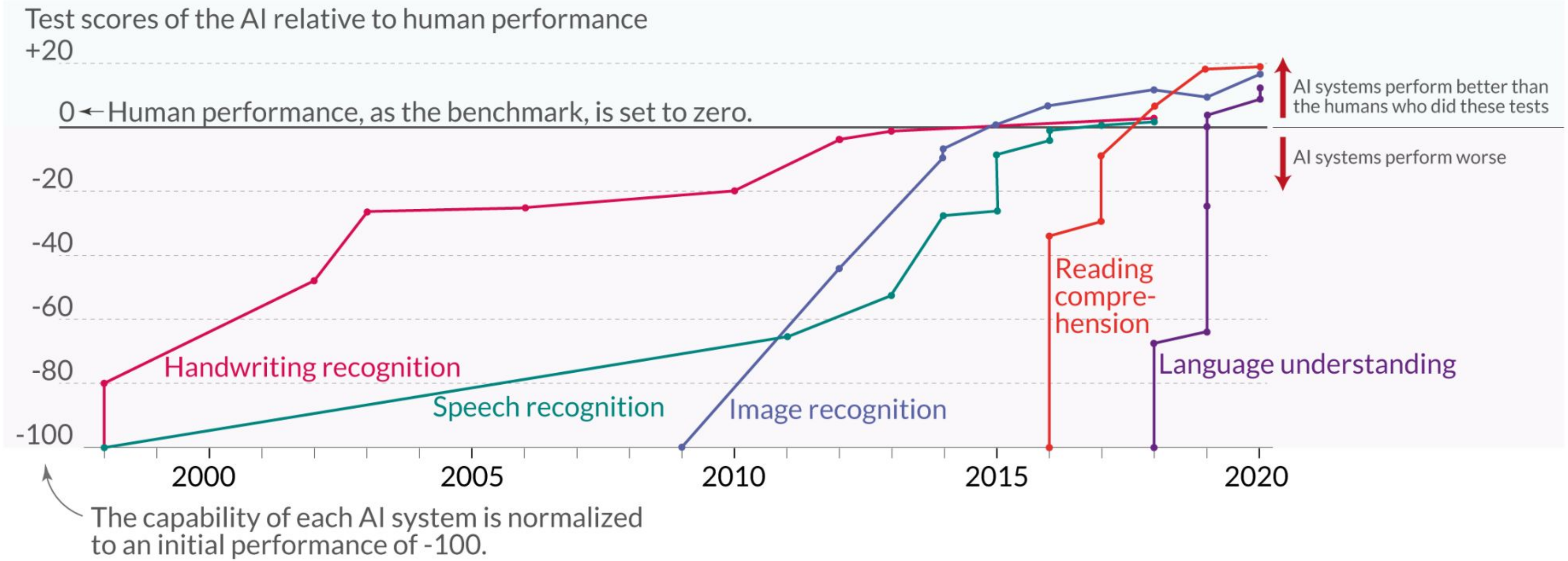
02 Games - a brief overview

03 Generative AI in Games using Kubernetes

04 Demo

Language and image recognition capabilities of AI systems have improved rapidly

Our World in Data



0 ← Human performance, as the benchmark, is set to zero.

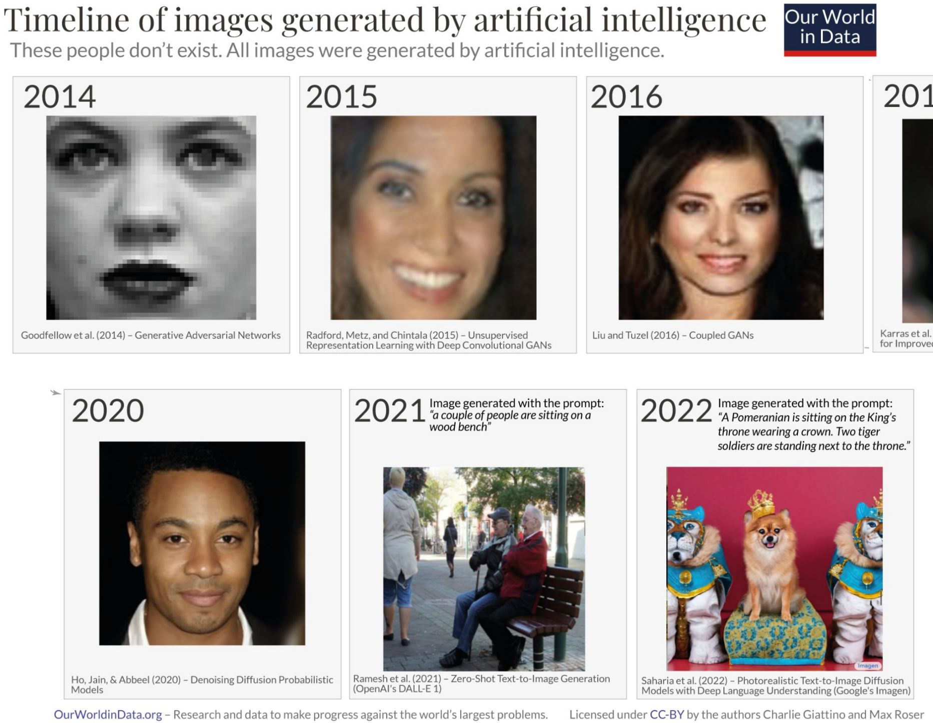
AI systems perform better than the humans who did these tests

AI systems perform worse

The capability of each AI system is normalized to an initial performance of -100.

Data source: Kiela et al. (2021) – Dynabench: Rethinking Benchmarking in NLP
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Citation: arXiv:2104.14337v1



Timeline of images generated by artificial intelligence

Our World in Data

These people don't exist. All images were generated by artificial intelligence.

2014
Goodfellow et al. (2014) – Generative Adversarial Networks

2015
Radford, Metz, and Chintala (2015) – Unsupervised Representation Learning with Deep Convolutional GANs

2016
Liu and Tuzel (2016) – Coupled GANs

2017
Karras et al. (2017) – Progressive Growing of GANs for Improved Quality, Stability, and Variation

2018
Karras, Laine, and Aila (2018) – A Style-Based Generator Architecture for Generative Adversarial Networks

2019
Karras et al. (2019) – Analyzing and Improving the Image Quality of StyleGAN

2020
Ho, Jain, & Abbeel (2020) – Denoising Diffusion Probabilistic Models

2021 Image generated with the prompt: *"a couple of people are sitting on a wood bench"*
Ramesh et al. (2021) – Zero-Shot Text-to-Image Generation (OpenAI's DALL-E 1)

2022 Image generated with the prompt: *"A Pomeranian is sitting on the King's throne wearing a crown. Two tiger soldiers are standing next to the throne."*
Saharia et al. (2022) – Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (Google's Imagen)

OurWorldinData.org – Research and data to make progress against the world's largest problems. Licensed under CC-BY by the authors Charlie Giattino and Max Roser

Nov 2023

Image that I generated using Stable Diffusion with the prompt:

"A photo of Albert Einstein in a space suit watching a solar eclipse."

https://stablediffusionweb.com/

https://stablediffusionweb.com

Citation: The brief history of artificial intelligence: The world has changed fast – what might be next?

The future of GenAI is even more incredible

Imagen Video

Prompt: **A teddy bear running in New York City.**

This is an example from imagen.research.google/video

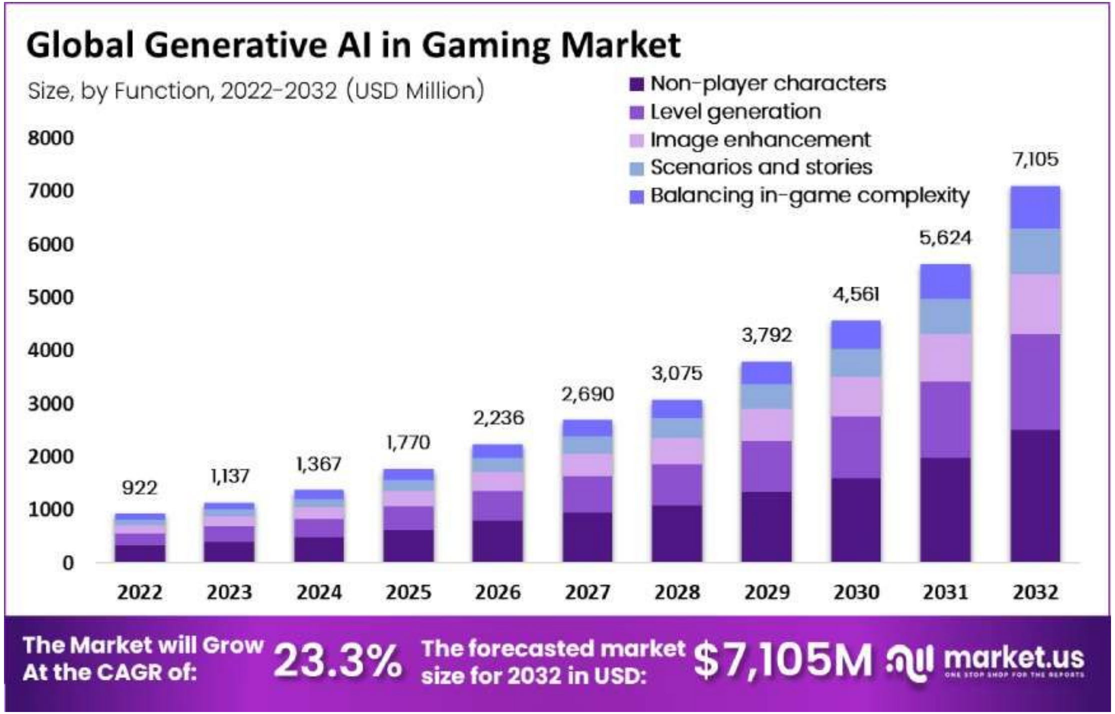
Imagen Video

Prompt: **A glass bead falling into water with a huge splash. Sunset in the background.**

This is an example from imagen.research.google/video

Citation: https://imagen.research.google/video/

Generative AI will Transform Live Service Games into Living Games



Global Generative AI in Gaming Market

Size, by Function, 2022-2032 (USD Million)

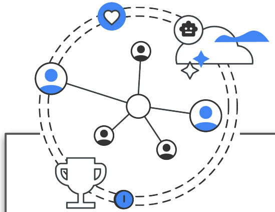
The Market will Grow At the CAGR of: 23.3%

The forecasted market size for 2032 in USD: $7,105M

GenAI is being used in **game development** use-cases first, and ultimately will be eclipsed by **new game experiences**

Classifications of GenAI use cases in Games

We will focus on inference during gameplay

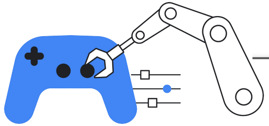


Improving Productivity during Game Development

Use Generative AI to accelerate time-to-launch by creating content and simplifying development

- 2D & 3D assets (characters, props)
- Audio & video assets generation
- Code generation
- AI-based game testing

Turnkey (VertexAI, Sagemaker) | Kubernetes (GKE)



Improving Player Experience during GamePlay

Use AI/ML & GenAI to adapt the gameplay and empower players to generate game content in realtime.

- Smart NPCs (bots)
- Dynamic in-game content
- Customized player experiences
- User-generated content
- Endless worlds

Kubernetes (GKE)

User pain-points for GenAI in Games

Platform

Cost

At-scale cost efficiency to ensure financial feasibility for AAA games

Latency

Low latency to ensure smooth gameplay & user experience.

Platform Selection

Platform(s) with performance, & access to models without lock-in.

AI Maturity

LLM Unpredictability

Need a coherent, relevant, and contextually appropriate inference

Avoiding Bias

Training & fine-tuning should not propagate biases & stereotypes

Content Filtering

Content moderation is needed to ensure safe & inclusive gameplay

Gameplay

Creativity vs Boundaries

Need to balance user generated content with game lore & structure

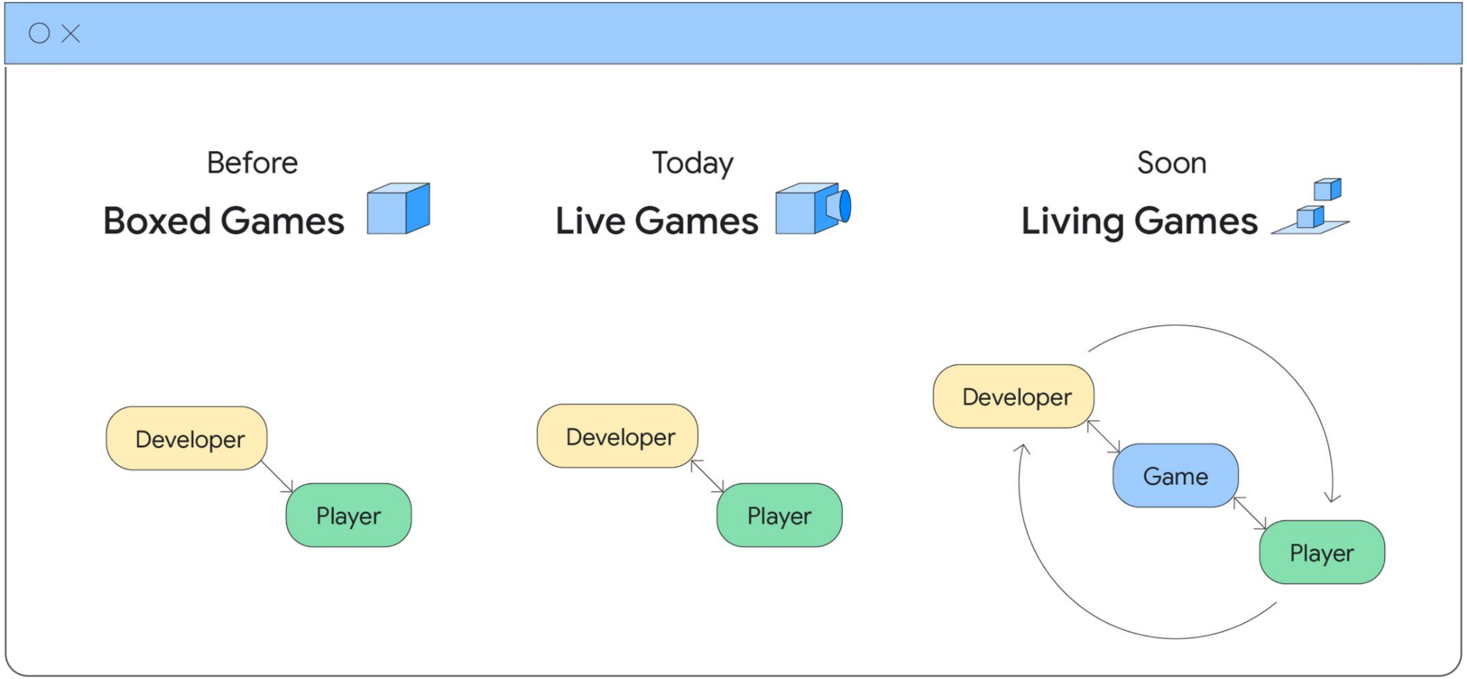
GenAI model constraints

Some games need content for gameplay which LLMs filter out

Procedural generation

Procedural generation with GenAI still requires human supervision

Generative AI is evolving the games industry



Before

Boxed Games

Developer

Player

Today

Live Games

Developer

Player

Soon

Living Games

Developer

Game

Player

Kubernetes is a great computing solution for games

Solves **majority** of the IT operations problems:

- Scheduling, health-checking, deployment methods, autoscaling and rollbacks
- Centralized logging & monitoring
- Extended by a massive ecosystem of tooling
- Declarative paradigm - (say what you want instead of manipulating what you have)
- Primitives for isolation

Challenge: Kubernetes on it's own, does not understand how game servers work.

For game servers, we need:

- to maintain in-memory state
- to start and shut down game servers on demand
- to protect running servers from shutting down (even for upgrades!)
- to scale based on demand - location, # of players - rather than CPU

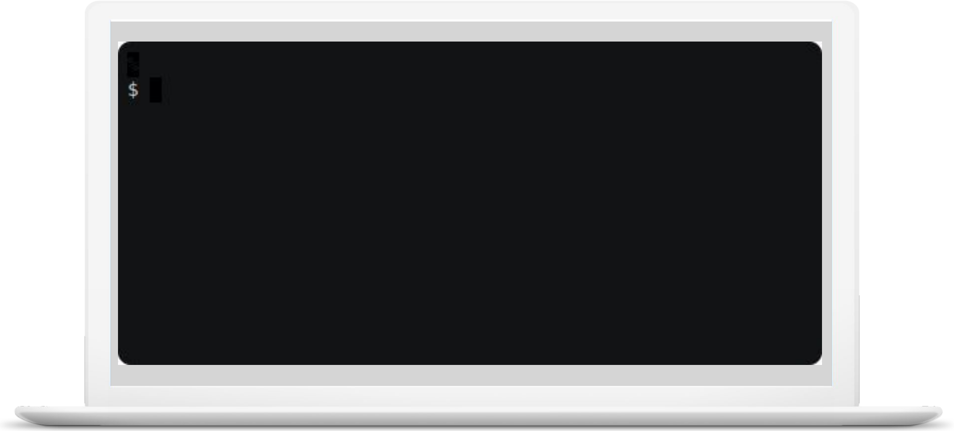
Agones

Agones enables hosting, running, and scaling dedicated game servers on Kubernetes

Agones is an open source, batteries-included, multiplayer dedicated game server scaling and orchestration platform that can run anywhere Kubernetes can run.

Get all the benefits of the Kubernetes operations, but now for game servers as well:

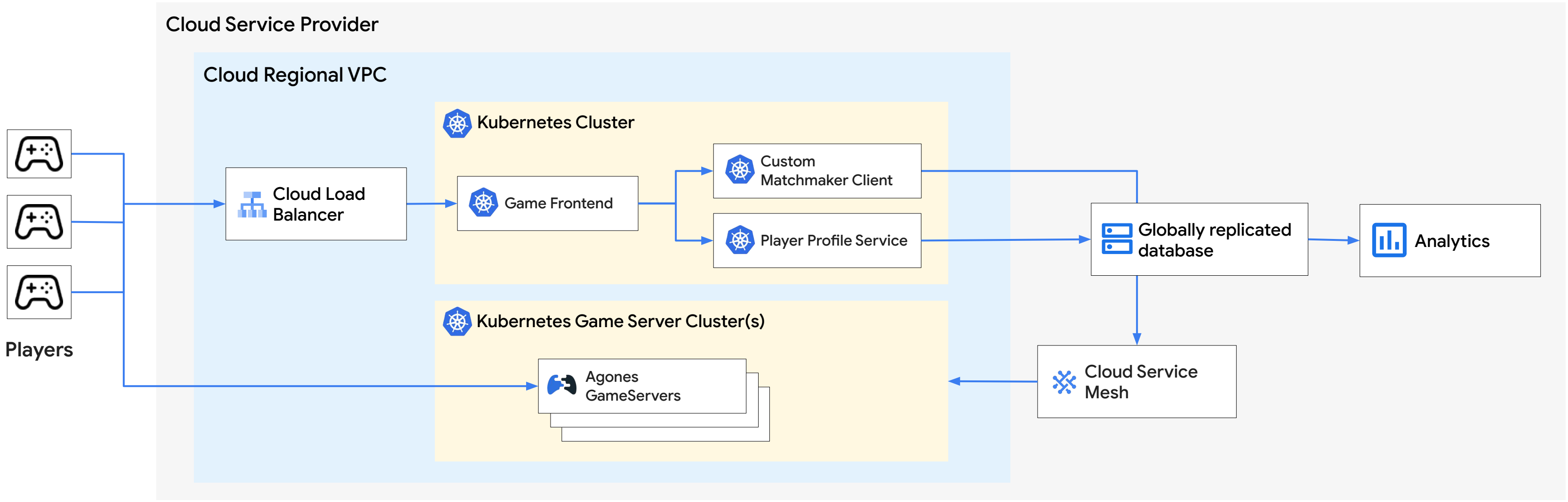
- **Termination:** understand game matches
- **Scaling:** understand player load
- **Networking:** Multiple UDP/TCP ports per node
- **Hot-spares:** Tunable warm-up parameters
- **Open source:** No vendor lock-in



```
$
```

Learn more at agones.dev

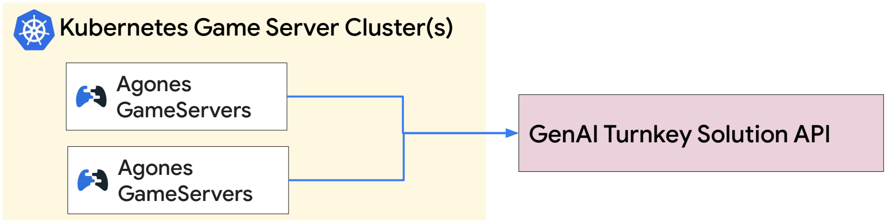
High-level architecture of a live service game



How do we add GenAI inference to gameplay?

Integrating GenAI inference with gameservers

Turnkey Solution
(E.g. VertexAI, Sagemaker, etc.)



Kubernetes Game Server Cluster(s)

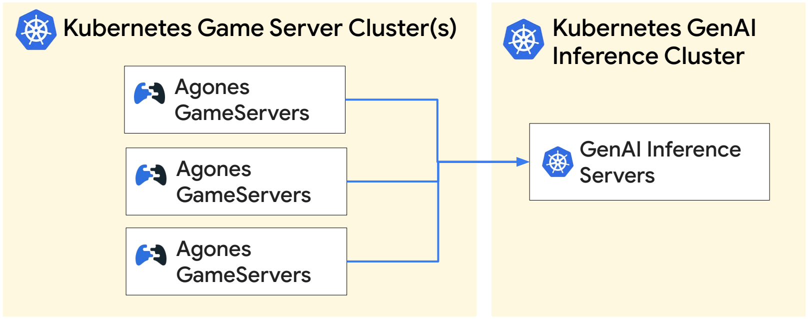
Agones GameServers

Agones GameServers

GenAI Turnkey Solution API

DIY Solution with Kubernetes

Dedicated GenAI Inference Kubernetes Nodes



Kubernetes Game Server Cluster(s)

Agones GameServers

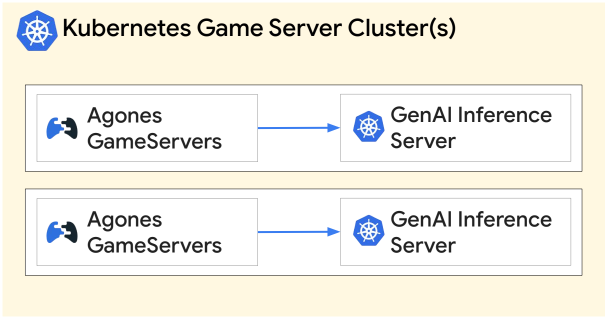
Agones GameServers

Agones GameServers

Kubernetes GenAI Inference Cluster

GenAI Inference Servers

GenAI inference Sidecar



Kubernetes Game Server Cluster(s)

Agones GameServers

GenAI Inference Server

Agones GameServers

GenAI Inference Server

Discussion on integrating GenAI inference with gameservers

Advantages of using a Turnkey solution:

- Game development use-cases
- Improving time-to-value: POCs for realtime use-cases during gameplay
- Specific models only available through Turnkey APIs

DIY solution with Kubernetes for GenAI in games:

- Increasing number of openly available models that can run in containers
- Cost optimization at-scale: k8s can be more cost-effective than pay-per-use APIs for high usage scenarios (game launches)
- Dedicated inference k8s nodes are easy to set up with k8s features such as HPA, scheduling with taints/ tolerations, etc.
- GenAI sidecars may have a slight advantage in latency, but are costly (1:1)

GenAI Inference Deployment Method	Latency	
	Image Generation (Stable Diffusion)	Text Generation (Bloom)
Dedicated Inference Kubernetes Nodes	~1s-1.3s	~146-147 ms
Inference Sidecar	~1s-1.3s	~144-145 ms

Today, inference latency overpowers any difference between different Kubernetes deployment methods. Dedicated inference k8s nodes provide the most versatility, ease of use, and flexibility.

Using Kubernetes for GenAI in games

Portability

**Write Once.
Run Everywhere.**

- Train and serve the same model(s) across clouds and on-premises
- Open standards prevent vendor lock-in

Flexibility

Choose the right framework(s) for the job

- Meet the needs of multiple teams with their framework of choice
- Customize the platform to meet your structure and requirements

Scalability & Performance

Fine tune performance and scale the platform

- Hyper-optimize the architecture for peak performance
- Scale the platform to meet the needs of all of your ML workloads

Cost & Efficiency

Pay for what you need when you need it

- Higher utilization of compute resources (CPUs, GPUs, TPUs) and cost savings with Spot
- Reduced operational costs for unified platform

Kubernetes

Industry standard compute orchestration platform available anywhere you need it

Spark RAY
beam RAPIDS
DASK XGBoost

Vibrant ecosystem of frameworks from which to choose

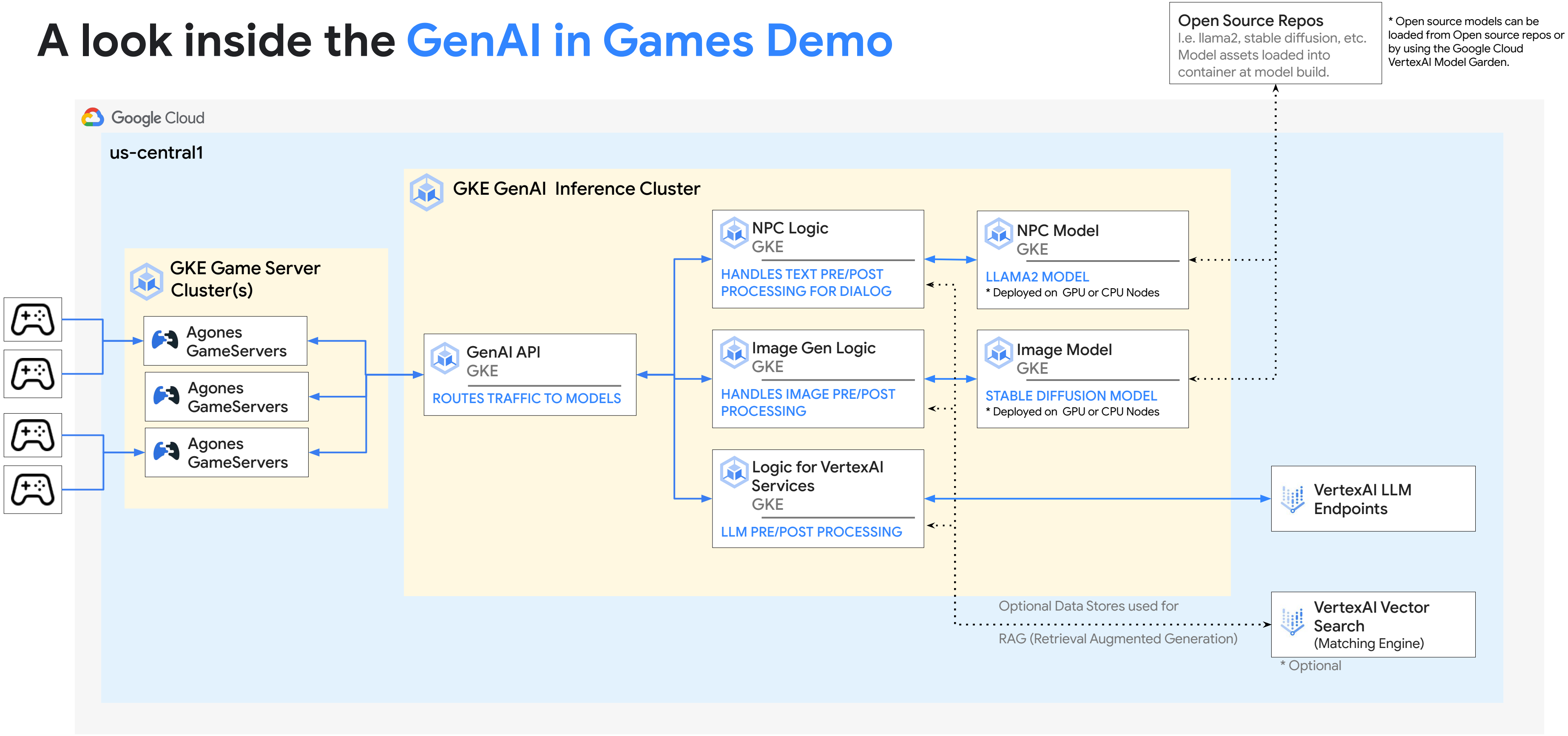
Agones

Run alongside game servers on Kubernetes.

- Improve latency & performance by running AI inference alongside game servers using Agones on Kubernetes.
- Reduce management overhead by using fully managed K8s (GKE Autopilot) for game servers

Demo

A look inside the GenAI in Games Demo


Open Source Repos
I.e. llama2, stable diffusion, etc. Model assets loaded into container at model build.
Google Cloud
us-central1
GKE GenAI Inference Cluster
GKE Game Server Cluster(s)
Agones GameServers
Agones GameServers
Agones GameServers
GenAI API
GKE
ROUTES TRAFFIC TO MODELS
NPC Logic
GKE
HANDLES TEXT PRE/POST PROCESSING FOR DIALOG
NPC Model
GKE
LLAMA2 MODEL
* Deployed on GPU or CPU Nodes
Image Gen Logic
GKE
HANDLES IMAGE PRE/POST PROCESSING
Image Model
GKE
STABLE DIFFUSION MODEL
* Deployed on GPU or CPU Nodes
Logic for VertexAI Services
GKE
LLM PRE/POST PROCESSING
VertexAI LLM Endpoints
Optional Data Stores used for
RAG (Retrieval Augmented Generation)
VertexAI Vector Search (Matching Engine)
* Optional

* Open source models can be loaded from Open source repos or by using the Google Cloud VertexAI Model Garden.

Acknowledgements

Google Kubernetes Engine (GKE) Team	
Technical leadership	Robert Bailey
GKE Games Engineering	Zach Loafman Max Gong Ivy Gooch
GKE AI Inference Team (benchmarking)	Kellen Swain Abdullah Gharaibeh
GKE User Research	Anna Poznyakov
Leadership	Alex Bulankou Iftach Ragoler Maulin Patel Jack Buser

Generative AI in Games Demo	
Google Cloud team	Globant team
Patrick Smith Dan Zaratsian Michael Bychkowski Giovane Moura Jr Sebastian Weigand Mark Mandel	Laura Gonzalez Cristian Giovagnoli Diego Salatino Rafael Martins Arturo Salinas Fede Vezzoso Jesus Gonzalez Matias Rodriguez Maggie Avallone

As you explore integrating Generative AI in your games, consider deploying your services on Kubernetes - matchmaking, game servers, and generative AI inference servers.

We would love to connect with you:

- g2x@google.com
- linkedin.com/in/sharmai
- twitter.com/IshantheSharma

Key links:

- Google Cloud for Games: goo.gle/cloudforgames
- GKE: cloud.google.com/kubernetes-engine
- Agones: agones.dev

Please scan the QR Code above to leave feedback on this session

KubeCon | CloudNativeCon
North America 2023

Thank you

Google Cloud