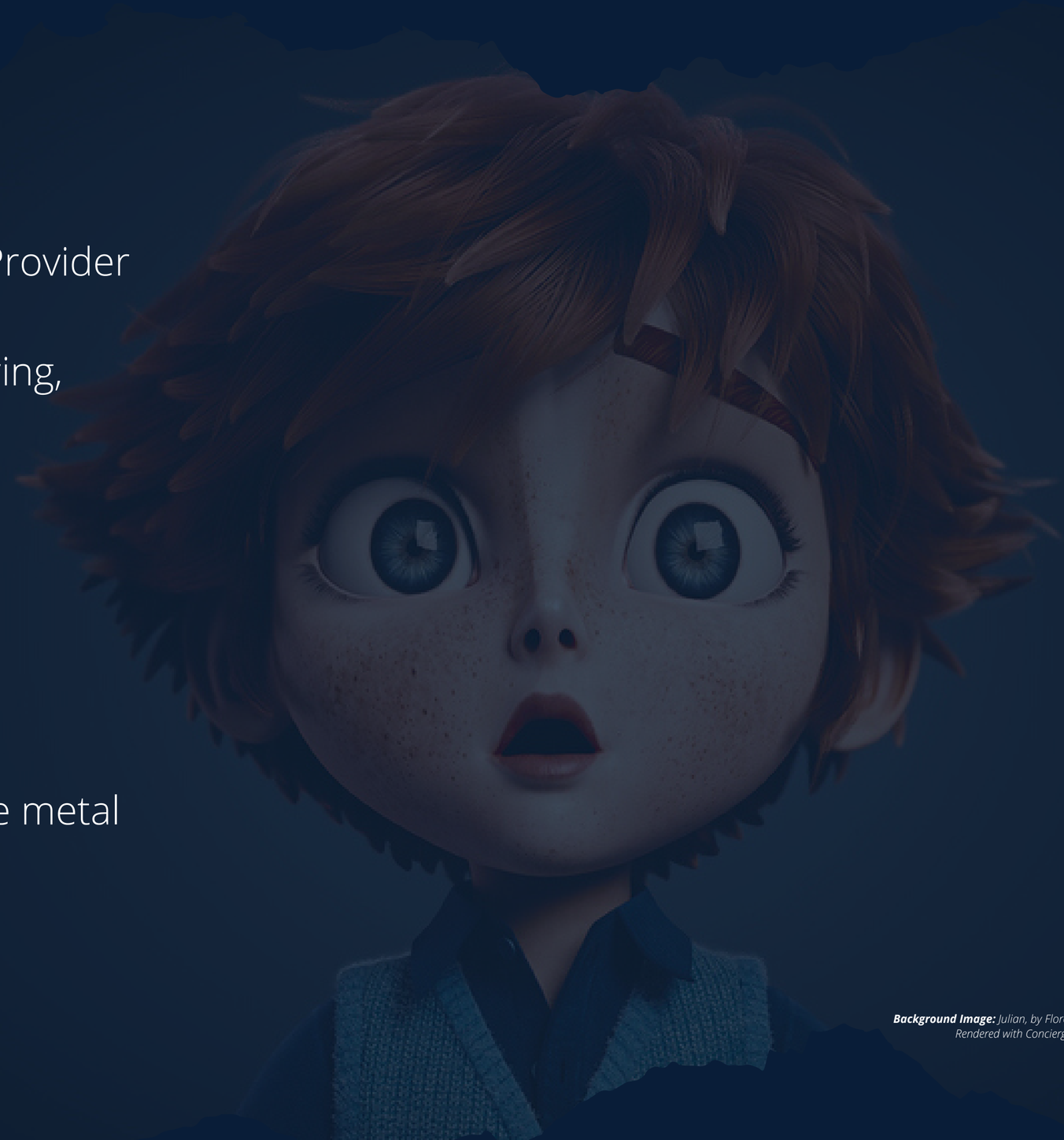
CoreWeave

Compute. Create. Innovate.

KFServing Working Group

Who We Are

- Fully managed Accelerated Compute Provider
- Serving Batch style customers (Rendering, Molecular Dynamics)
- Online customers (Realtime Inference, Transcoding)
- Client size in 50 - 4000 GPU range
- Large multi tenant Kubernetes on bare metal clusters (1,000 nodes / 6,000 GPUs)



Background Image: *Julian, by Floroj*
Rendered with Concierge

Our KFServing Journey

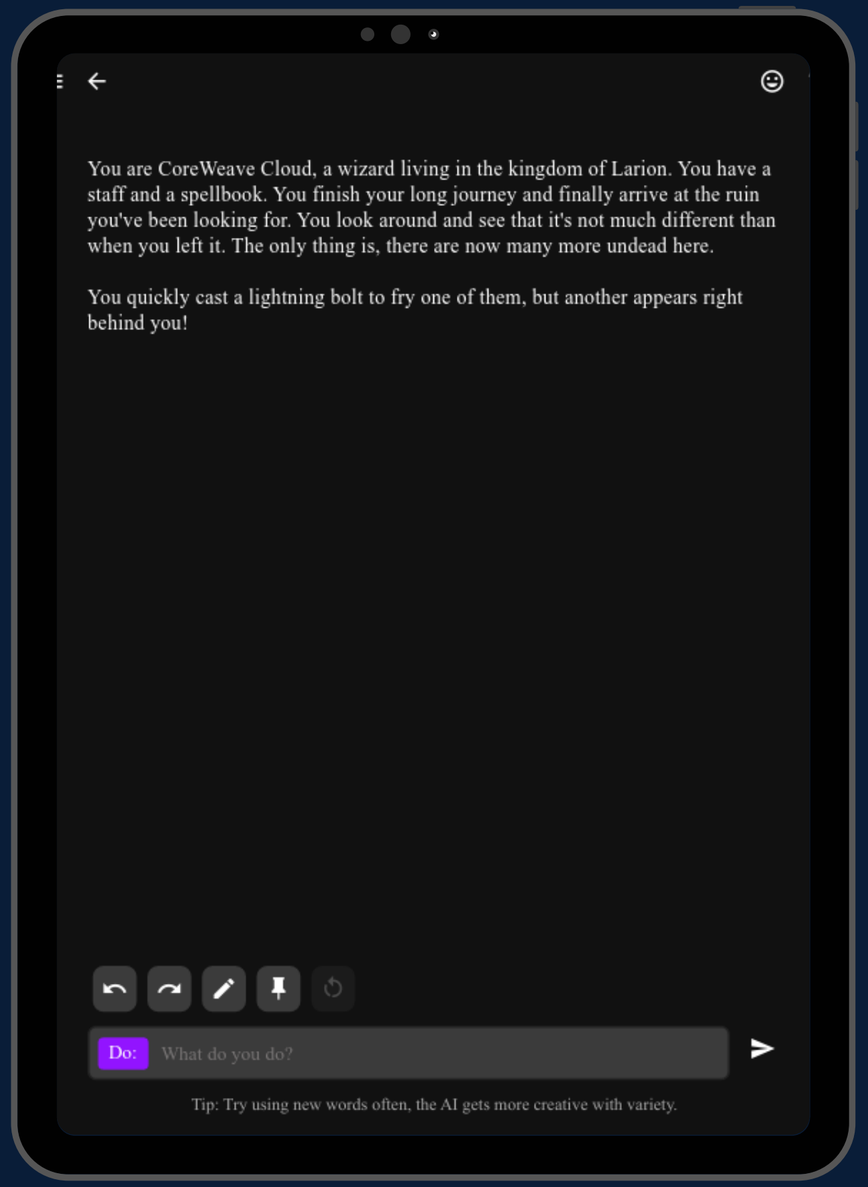
- Customer request for Online Inference Solution January 2020, moving off cortex.dev
- Looked into porting cortex.dev from AWS
- KFServing on top of KNative yielded an elegant solution
- Not as user friendly as cortex.dev (but we are working on it)

Background Image: *These Tubes are Hot, by Raul Rendered with Concierge*

Case Study: AI DUNGEON

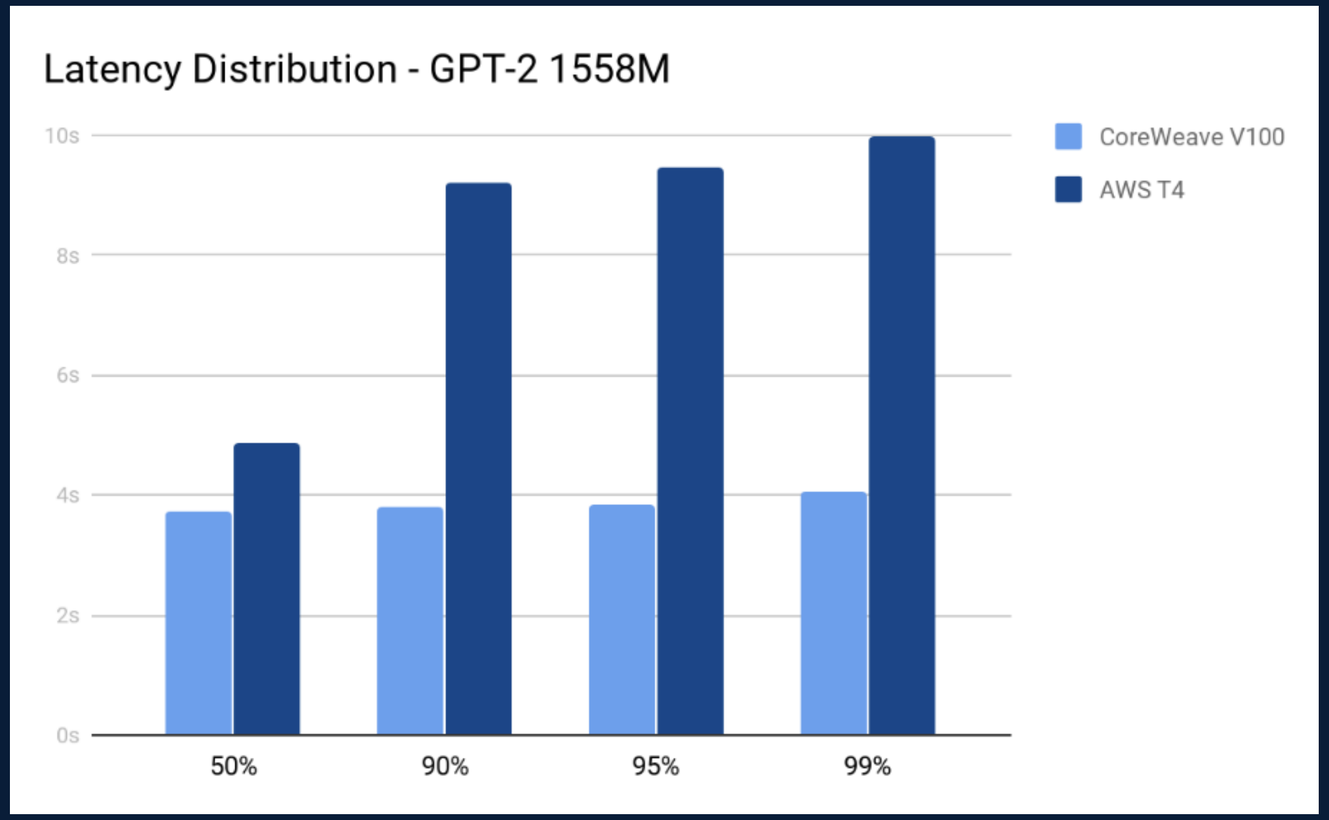
AI Dungeon is an AI text adventure game based on the OpenAI GPT-2 natural language model. With over 1,600,000 users signing up to play the game, AI Dungeon quickly ran into budget, performance and capacity problems with AWS and GCP.

CoreWeave transitioned AI Dungeon from the Cortex platform, only available on AWS, to an in-house engineered inference solution. Delivered on NVIDIA Tesla V100s via CoreWeave Cloud with native auto-scaling and load balancing services to manage the immense user traffic, AI Dungeon's response time per request has dropped by 50%.



Case Study: AI DUNGEON

KNative Activator



What we are building

- PRs for Transformer performance, Container concurrency
- Upcoming PR for CORS
- Upcoming PR with 5 step-by-step examples
- CLI to improve user experience to be in parity with cortex.dev

Background Image: *These Tubes are Hot, by Raul*
Rendered with Concierge

CoreWeave

Compute. Create. Innovate.

12 Commerce Street
Springfield, NJ
investments@coreweave.com

We believe the future belongs to the creators.