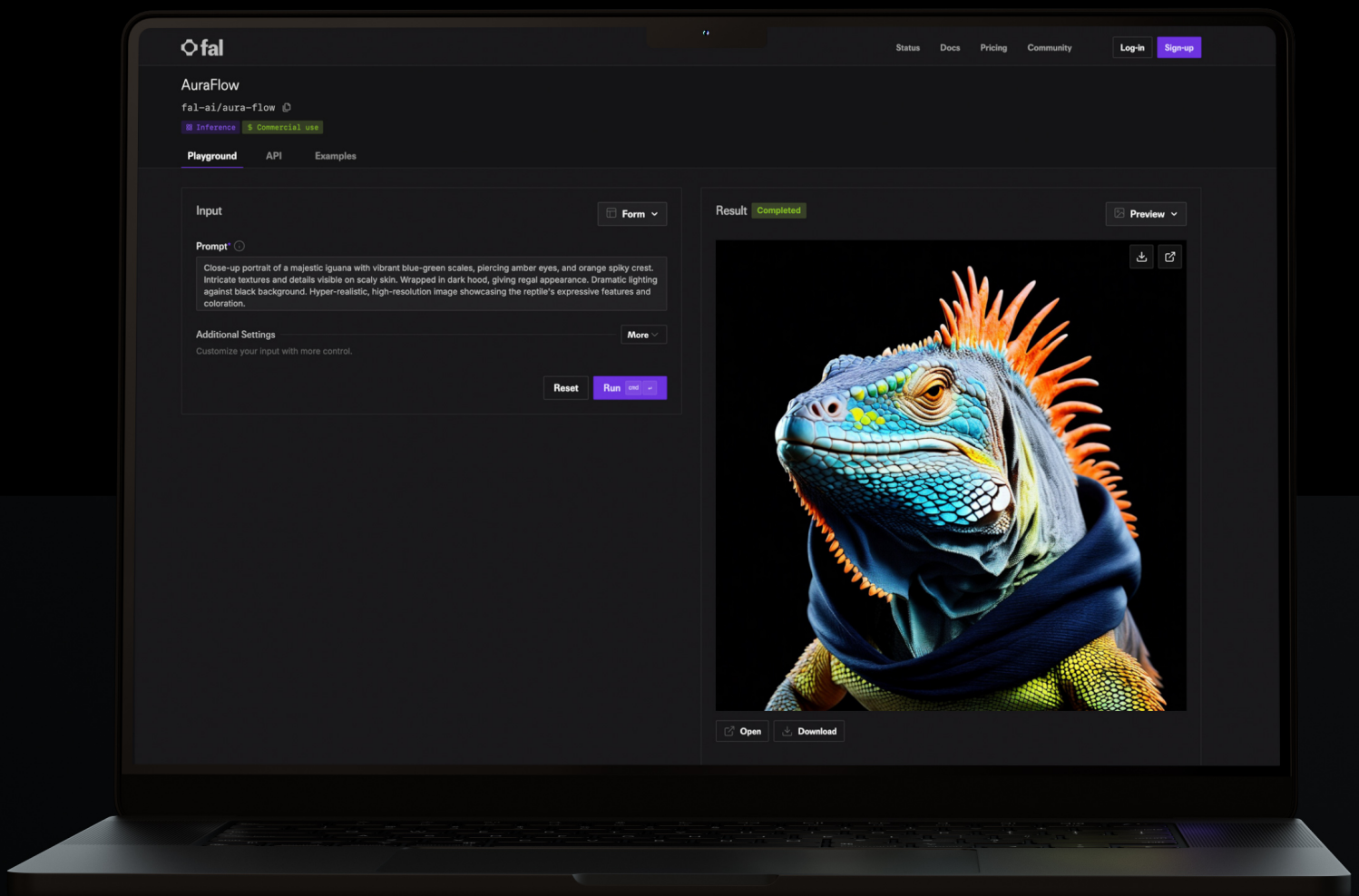
CASESTUDY

How fal Scaled Training & Inference With Lambda



Lambda

Lambda

Giving Back to the Open-Source Community

Few technology trends have progressed as rapidly as generative AI, and the open-source community has been instrumental in driving this innovation. By experimenting, learning, and openly sharing their discoveries, these trailblazers have created an environment where new breakthroughs build on the collective efforts of many. fal, a pioneering company in the space, identified a lull in the generative AI market for inferencing and training open-source models.

fal made it their goal to build an open-source model and continue experimenting with new ideas to give back to the community and push the industry forward. This led to the development of a model series called AuraFlow, a massive diffusion model designed for text-to-image generation.

Challenges of Accessing On-Demand NVIDIA GPUs

The size of the model demanded extensive NVIDIA GPU resources, which was challenging, but the real hurdle fal faced was how to consume those resources—mainly, when training and experimenting with their models, how could they access what they needed, when they needed it. Initially, fal trained version 0.1 of AuraFlow using a one-month contract with a cloud provider. While this allowed them to start development, it took weeks to negotiate a short-term contract that worked for them, and those were cycles fal didn't want to spend every time they trained a new version of the model.

This prompted fal to seek a more agile infrastructure that can scale up for experimentation and training within seconds, then instantly scale down when not in use. They needed a system that provided on-demand resources, without tying them down to long-term agreements or paying for idle GPUs. As the fal's Founding Engineer—Batuhan Taskaya—put it, "the first question we had was how the hell are we going to train this in terms of the GPUs we need, where are we going to find the GPUs?" With future iterations in mind, including versions 0.2 and 0.3, they sought a system capable of adapting resource levels in real-time.

Multi-node, Multimodal, Multipurpose Solution

Lambda's 1-Click Cluster (1CC) provided the answer for the next stages of AuraFlow development, offering fal a self-serve, on-demand cluster solution to rapidly experiment without the upfront commitments. Taskaya immediately realized the potential of 1CC for future iterations of AuraFlow, saying, "I saw 1CC... and I was like, this is an amazing idea... Lambda has [a model where you pay only for what you need without us needing to spend time working out a short-term agreement—it was all basically self-serve.]" The shift to Lambda for AuraFlow 0.2 and 0.3 development allowed them to expedite the experimentation process, iterate faster, and significantly reduce setup time.

> "Lambda Empowers fal With Agile Infrastructure Tailored for Instant Access to NVIDIA GPU Clusters"

Lambda

> "Access to NVIDIA GPU Clusters on a Flexible Schedule Accelerated Time-to-Market and Helped fal Capture New Revenue Streams"

Accelerated Processes and Technical Impact

By using Lambda's 1CC, fal was able to act quickly, releasing AuraFlow versions 0.2 and 0.3 in weeks rather than months. The ability to run multiple experiments simultaneously was critical for the advancement of AuraFlow as it allowed for faster iterations, optimal use of NVIDIA GPU resources, comprehensive testing of diverse approaches, and more adaptability as they learned and tuned the model.

Taskaya commented, "We are a very agile startup, it's a very, very competitive market and we want to iterate fast, right? If I have a cool idea, I need like 22 GPU assets for a week, I don't want to wait four days to go into a sales meeting and then another four days [to finalize a contract and get going] ... I think that's the main benefit I see from my 1CC."

Business Agility and Impact

This flexibility allowed them to continually experiment with new ideas, test hypotheses, and make quick pivots as necessary by leveraging only the resources they needed and when they needed them. This agility helped amplify their standing as a leader in the open-source image generation community and opened new doors as they became a thought leader for the space.

Building on the success of AuraFlow and the recognition of fal, they have expanded their offerings by collaborating with customers to create custom generative AI models tailored to specific needs. These projects have opened new revenue streams, as businesses seek out fal's expertise to develop cutting-edge AI solutions, positioning them as a go-to partner for generative AI innovation.

Trusted Support

In addition, fal's trust in Lambda's infrastructure extended beyond just speed. Taskaya notes, "We trust Lambda with our most important assets. If our systems are down because of the GPU... our customers are angry. Having...sub 1-hour responses...is great and having people we trust responding to our tickets...that's our reputation on the line," highlighting the importance of reliable, scalable infrastructure to maintain uptime and keeping their customers happy while working at scale.

With Lambda, fal gained more than just scalable, self-serve infrastructure—they partnered with a team that deeply understands the needs of open-source AI development. Lambda's expertise in supporting the specific requirements of open-source model training and inference have made them an indispensable partner in fal's journey.