Modal

Generation Digital

Designed to help AI teams deploy faster.

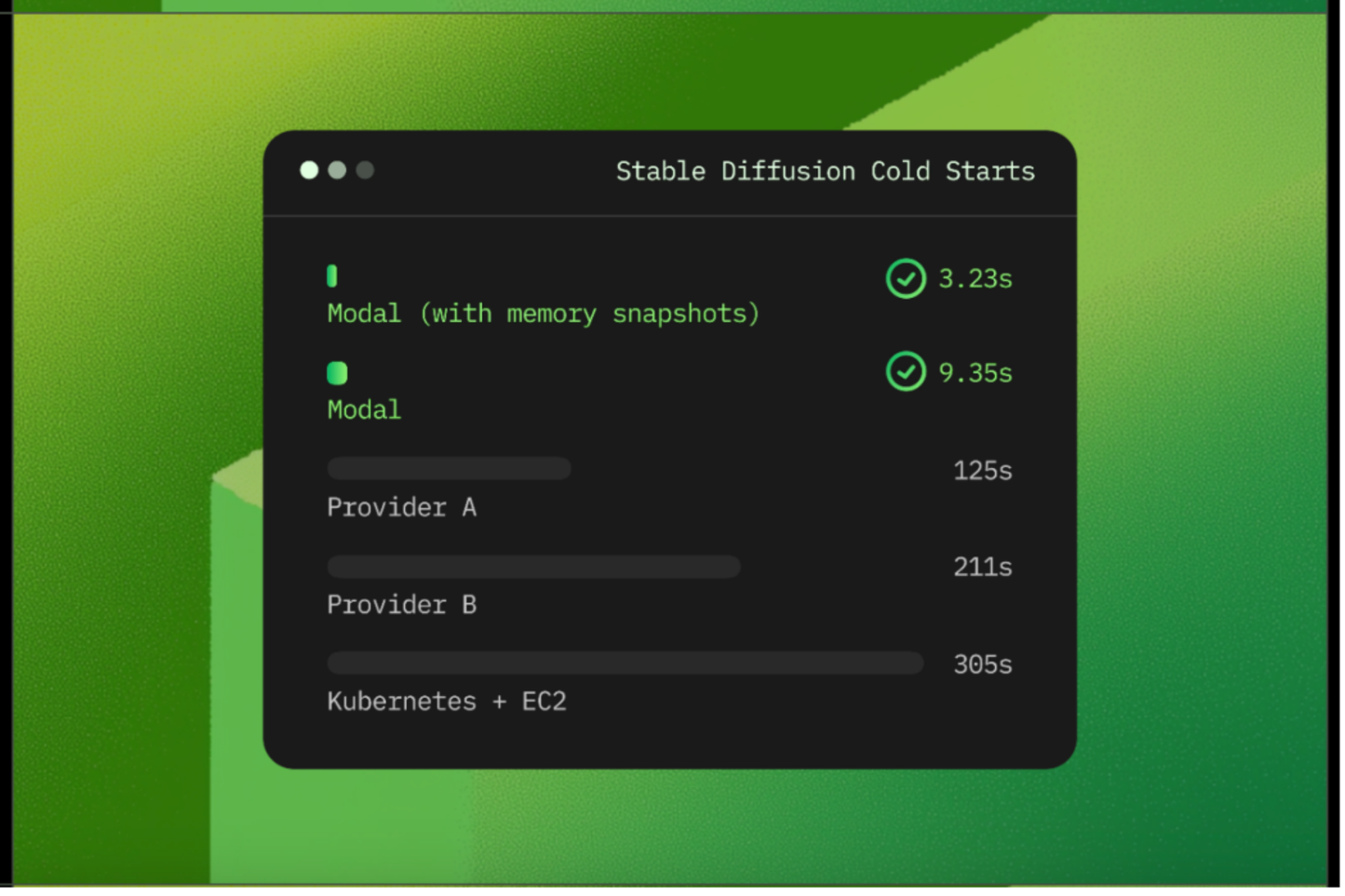
Programmable infra

Define everything in code, no YAML or config files. Keep environment and hardware requirements in sync.

```python
inference_image = (
    Image.debian_slim()
    .uv_pip_install(
        "torch==2.7.1",
        "transformers==4.53.2",
    )
)
@app.function(image=inference_image, gpu="B200")
def inference():
    ...
```

Built for performance

Launch and scale containers in seconds to keep feedback loops tight and latency low.

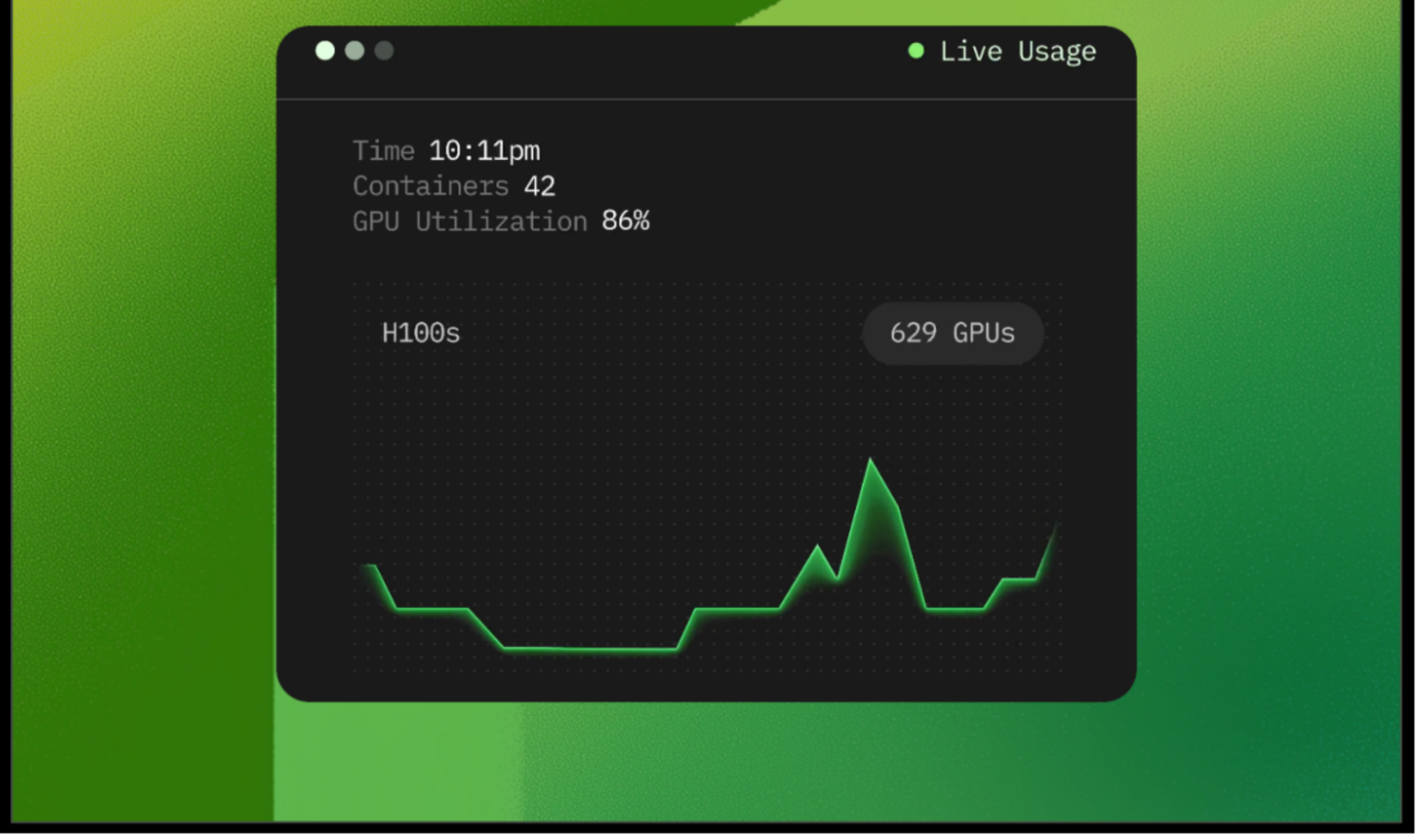


Elastic GPU scaling

Elastic GPU capacity and access to thousands of GPUs across clouds. No quotas or reservations. Scale back to zero when not in use.

Unified observability

Integrated logging and full visibility into every function, container, and workload.



Powering any ML workload

Inference

Deploy and scale inference for LLMs, audio, image/video generation.

Training

Fine-tune open-source models on single or multi-node clusters instantly.

Sandboxes

Programmatically scale secure, ephemeral environments for running untrusted code.

Batch

Scale to thousands of containers for batch workloads on-demand.

Notebooks

Collaborate on code and data in real-time with shareable notebooks.

PLATFORM

Build on a powerful foundation

From filesystem to runtime, every layer of Modal's platform is engineered to give you the tools to build robust, scalable data applications.

Learn More

AI-native runtime

Engineered from the ground up for heavy AI workloads, built for super-fast autoscaling and model initialization and 100x faster than Docker.

Built-in storage layer

A globally distributed storage system built for high throughput and low latency. Designed for fast model loading, training data, or other datasets.

First-party integrations

Mount your existing cloud buckets, connect to MLOps tools, and send data to your existing telemetry vendors.

Multi-cloud capacity pool

Deep multi-cloud capacity with intelligent scheduling ensures you always have the CPUs and GPUs you need without managing input orchestration.

Security and governance

Team controls

Single sign-on and role-based access controls for teams of any size.

Battle-tested isolation

Built on top of gVisor, which provides stronger container isolation compared to standard runtimes.

SOC2 & HIPAA

SOC 2 Type II and HIPAA compliant. Industry-grade security and privacy are built-in.

Data residency controls

Enforce specific cloud regions to run your workloads in for your compliance requirements.