SKY_ COMPUTING

Charting a New Course in AI Cloud Infrastructure

September, 2025

Brian Irish, Engineering Lead at Ditto

Hi, I'm Brian

Engineering Lead

Ditto

Staff Kubernetes SRE

SuperOrbital, Raincoat, Payfare

Two decades of experience

2005-2025

The Cloud

On-Demand Scalability

Pay-as-you-go pricing models

Near-instant global reach

Vendor lock-in

Data Gravity

Egress charges

The Cloud Repatriation

“**42% of organizations** surveyed in the United States are considering or already have moved at least half of their cloud-based workloads back to on-premises infrastructures, a phenomenon known as **cloud repatriation.**”

The Cloud Repatriation

The Next Evolution of Cloud Computing

Sky Computing

Sky Computing

Circa 2021

https://sigops.org/s/conferences/hotos/2021/papers/hotos21-s02-stoica.pdf

Workloads flow seamlessly between providers

Free from lock-in and inefficiency

Better control over data governance

Three Pillars

Abstraction

Hides the complexity of individual clouds

Hypercloud and Neocloud support

Automation

Manage workload placement

Intercloud Brokers as decision-makers

Agility

Reciprocal Peering agreements between brokers and clouds

Promotes and encourages neocloud participation

SkyPilot - The Inaugural Broker Framework

- Developed by UC Berkeley Sky Computing Lab (**creators of Spark, Ray**)
- **Open-source** Intercloud Broker framework implementing Sky Computing principles
- Encompasses the **Automation pillar**
- 1M+ downloads, v0.10.0 (July 2025) with enterprise features
- Unified interface to 17+ clouds + Kubernetes + on-prem
- Key capabilities:
  - 3-6x cost savings through spot orchestration
  - 4x faster provisioning (200 GPUs in <90s)
  - 9.6x faster checkpointing with mount_cached
  - Zero code changes for existing ML workloads
- One YAML to rule them all: `sky launch --cloud any`

ML Workload Impacts

Case Studies

Abridge

”Skypilot is pretty nice actually… I must admit even as a die hard slurm guy.”

— John Giorgi, Research Scientist

Covariant Brain

Powering a new AI stack: “Using multiple regions allows us to get **much higher GPU availability**—combating the current GPU shortage on the cloud—which is not possible with other tools/services.”

Abridge - More Details

SkyPilot delivered the familiar experience our researchers wanted with the reliability our production workloads required:

- **Interactive development:** `sky launch --gpus H100:4` provides immediate SSH access to a GPU-enabled shell without complex setup. Just like `srun --gres=gpu:4 --pty bash` in SLURM, but works seamlessly across all our infrastructure.

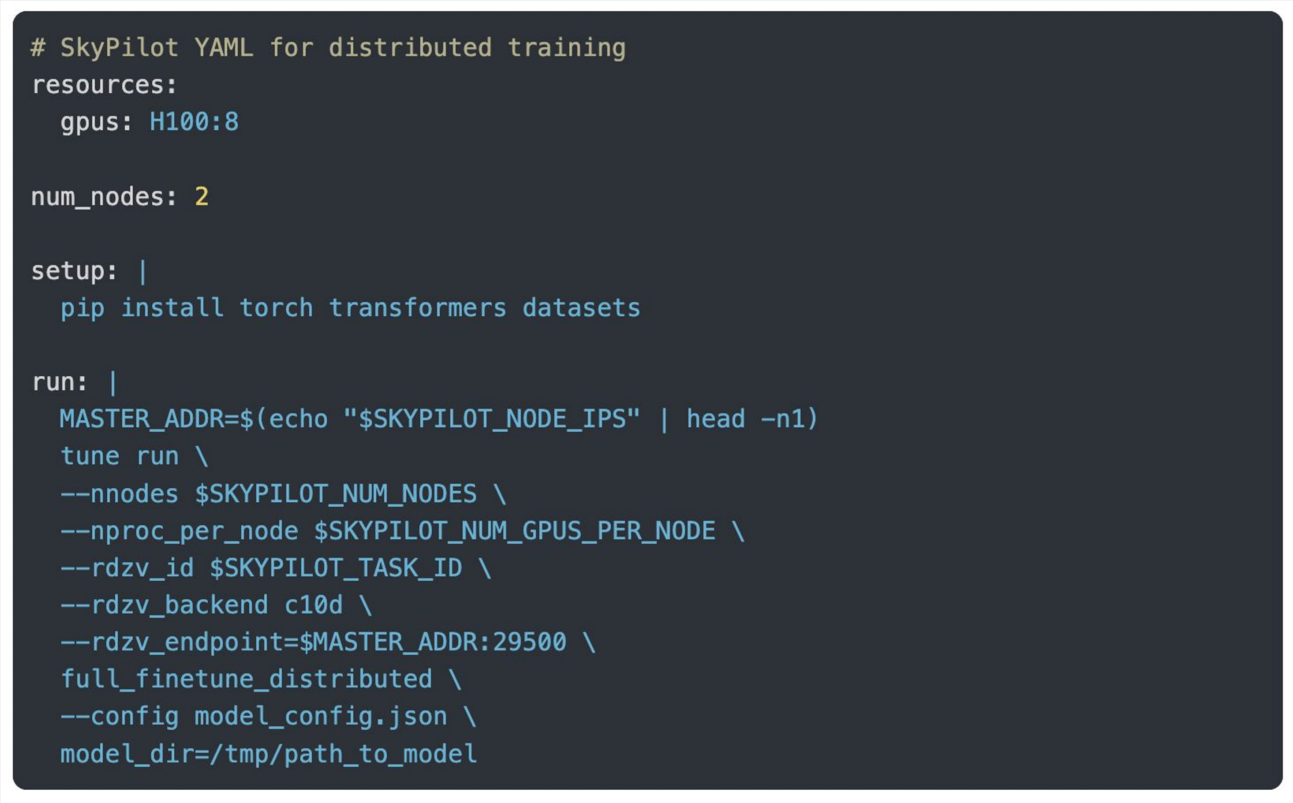
- **Jupyter notebook hosting:** We can spin up Jupyter notebooks directly on GPU clusters, enabling researchers to prototype with high-end hardware that wasn't available locally.

- **Managed jobs:** SkyPilot's managed jobs provide the same convenience as SLURM's job scheduler but works across all our infrastructure - automatic restart on job failures, strong isolation, and reliable job management for long running training jobs.

- **Model evals:** Quick model evaluation became simple - we can deploy models as FastAPI services in minutes for testing. Unlike SLURM which lacks native API endpoint support, SkyPilot makes it easy to expose models as services.

https://blog.skypilot.co/abridge/

Abridge - Distributed Training in SkyPilot



```
# SkyPilot YAML for distributed training
resources:
  gpus: H100:8

num_nodes: 2

setup: |
  pip install torch transformers datasets

run: |
  MASTER_ADDR=$(echo "$SKYPILOT_NODE_IPS" | head -n1)
  tune run \
  --nnodes $SKYPILOT_NUM_NODES \
  --nproc_per_node $SKYPILOT_NUM_GPUS_PER_NODE \
  --rdzv_id $SKYPILOT_TASK_ID \
  --rdzv_backend c10d \
  --rdzv_endpoint=$MASTER_ADDR:29500 \
  full_finetune_distributed \
  --config model_config.json \
  model_dir=/tmp/path_to_model
```

AI/ML Workloads

Different stages of a pipeline may benefit from different cloud providers' specialties. With an Intercloud Broker, you can split your pipeline and run:

- Model training on Google Cloud, with TPU-optimized instances for deep learning
- Inference on AWS, with their Inferentia chips for lower latency
- Data preprocessing on Azure

Why?

+ Speed
+ Cost savings
+ Regional data regulations

Challenges to Adoption

Standardization

- Universal standards across all cloud platforms is unlikely due to competitive interests and proprietary technologies.
- Progress can still be made by leveraging existing widely-adopted tools, such as Kubernetes, Ray, and S3 APIs.
- These standards don’t cover every scenario but provide a practical bridge, allowing Sky Computing to move forward without waiting for complete industry-wide uniformity.

Economic Resistance

- Hyperscalers will resist reciprocal peering agreements

- Neoclouds have strong incentives: their agility and desire to compete with hyperscalers drive them to support the ecosystem, gradually encouraging wider adoption and putting pressure on the bigger providers to reconsider their stance.

Infrastructure Inertia

- **Significant Investments:** Organizations have heavily invested in existing cloud infrastructure (cost, expertise, tooling, processes), leading to hesitation in dramatic changes.
- **Resistance to New Paradigms:** Reluctance to adopt new paradigms like Sky Computing due to lack of widespread adoption.
- **”Good Enough” Status Quo:** Current cloud deployments often function adequately, even if not optimal for cost or performance.
- **Daunting Overhead:** Concerns about retraining staff, updating deployment pipelines, and refactoring applications for Sky Computing’s abstraction layer.
- **Perceived Risks:** Apprehensions regarding reliability and support when moving away from established cloud providers’ native services.
- **Significant Inertia:** These factors create substantial inertia hindering widespread Sky Computing adoption.

The Challenge of Legitimacy

- The concept of Sky Computing faces challenges in establishing legitimacy, partly due to a Wikipedia entry with a warning banner questioning source reliability and noting a lack of academic citations.

- This stems from an incident where a commercial entity attempted to shape the narrative around Sky Computing through Wikipedia editing, leading to their ban from the platform.

- This incident highlights a broader challenge: commercial entities sometimes try to claim thought leadership for emerging technologies through questionable means, inadvertently damaging the credibility of legitimate technological advances.

- However, the fundamental value proposition of Sky Computing—providing a unified interface across cloud providers while optimizing for cost, performance, and compliance—stands independent of any single company's implementation.

Final Thoughts

THANK YOU