Workloads

Making Kubernetes Simpler for Accelerated



Lucy Sweet
Senior Software
Engineer
Uber



Mitch McKenzie
Site Reliability Engineer
Weave



Aditya Shanker
Senior Product
Manager
Crusoe Al



Rebecca Weekly

VP Platform &

Infrastructure Eng

Geico



Susan Wu Product Manager Google

**Experiment Tracking** Vector DB (Qdrant, (MLFlow, W&B) Pinecone) **Cost Optimization** Scheduling (Kueue, Slurm, Model Registry Yunikorn) Job Scheduling & Pipelines Ingress / Load Balancers (Airflow, Argo Workflows) Job Execution (Ray) Container Registry Serverless (KNative) Infrastructure Observability Job Execution (Ray, Spark) Resilient Distributed Node / Pod Autoscaling Training ( Shared Storage (NFS / S3) Inference Engines / **Data Ingestion** Training Framework (Jax, GPU / Network Operators Inference Servers Pytorch. Tensorflow) **Common Components Data Processing Model Development / Training** Inference (Real-time) **Nvidia** CPU CPU Nvidia Cloud Provider AMD **GPU GPUs** & Proprietary nodes nodes **GPUs** Clusters (X86)(ARM) (PCIe) accelerators Nodes and Node Pools (Virtualizated or Bare Metal) K8s