Hypervise, Maximize & Optimize Your ML Infra

Izhak Eidus - Co-founder

WoolyAI

The challenge of keeping GPU max utilized for experiments infrastructure

Keeping excessive no of cores busy

Waiting on data processing (CPU)

Multi Tenancy on GPUs to maximize utilization

Is it different from multi tenancy on CPUs…

GPU Runtime to enable multi tenancy

Introduce OS-level Resource (Cores schd and VRAM) abstraction

Compute Scheduling

- Allocate fraction total cores at runtime based on actual usage
  - Dynamically manage allocation based on priorities etc

Memory Virtualization

- Transparent VRAM deduplication
- Vram tasks swapping.

Decouple CPU-GPU tight execution

- Don’t keep GPU waiting for CPU side executions
- Connect many CPU clients to every GPU

WoolyAI GPU Runtime

Turning GPU (Nvidia) clusters into a shared, high-efficiency GPU compute cloud that maximizes utilization, throughput, and accessibility for AI workloads

Let’s discuss more at our booth