

The background of the slide is a dark blue gradient with abstract, glowing blue geometric shapes on the right side, resembling stylized rays or facets of a crystal.

Ray's evolution to support AI workloads

Ion Stoica

What is Ray?

A distributed framework for
scaling complex AI workloads

Where this all started...



CS294
Fall, 2015

[Overview](#)
[Student blogs](#)
[People, Places, and Times](#)
[Lecture Schedule](#)
[Reading](#)
[Grading Policy](#)
[Announcements](#)

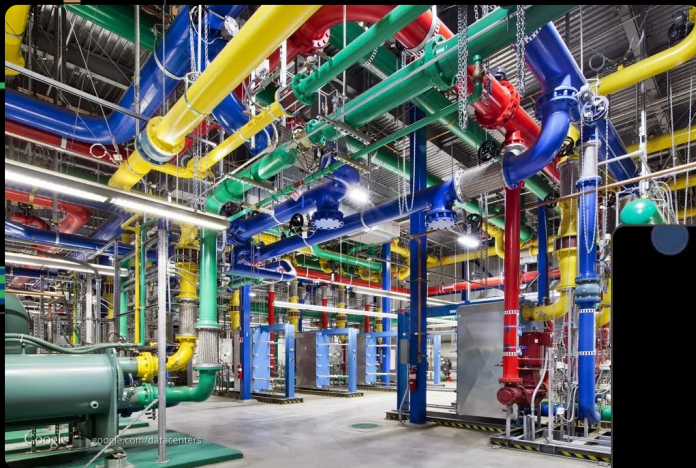
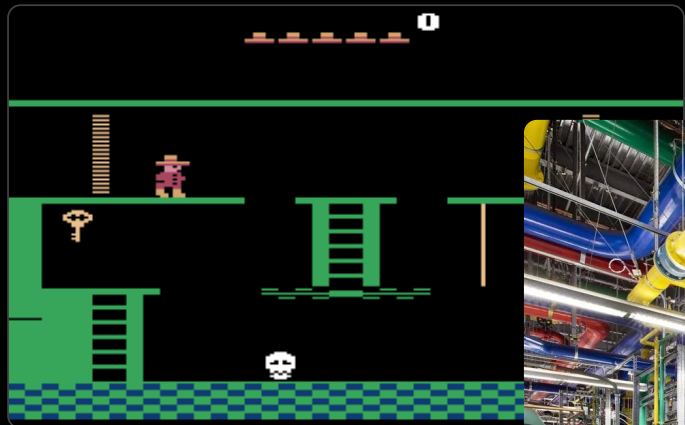
Big Data System research: Trends and Challenges

Overview

Description

Fueled in part by the continuous drop in storage prices, today virtually every company collects every piece of data (e.g., user and machine logs) it can. Companies do so with the hope that they can use this data to improve their products or services. This hope is driven by the desire to emulate the

Lots of excitements around RL



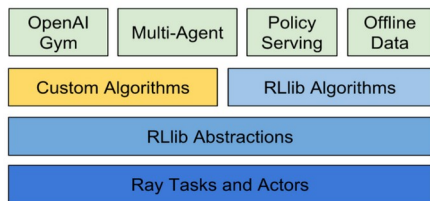
We built Ray for RL

The most complex workload of the time

RLlib

2018

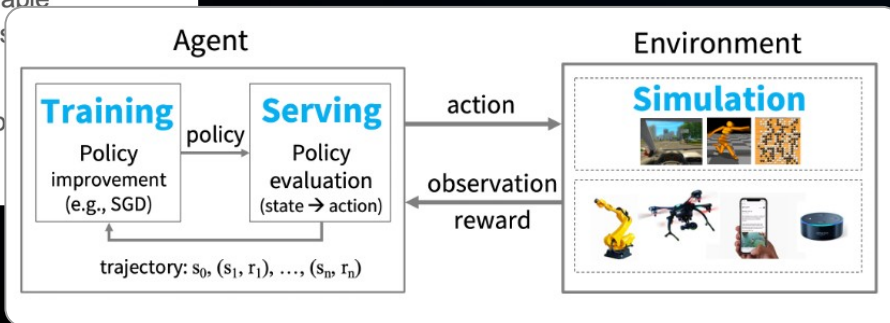
User perspective: three main value adds:



1. APIs that make RL usable in a variety of scenarios

2. Collection of scalable reference algorithms

3. Abstractions for implementing distributed algorithms



but we were ahead of our time

Ray downloads per quarter

Classic DNN models (- 2022)

Data processing	Log processing
Training	Dense models (Data parallelism)
Serving	Single-GPU models

moderate complexity, moderate scale

Ray downloads per quarter

Classic DNN models (- 2022)

Data processing	Log processing
Training	Dense models (Data parallelism)
Serving	Single-GPU models

moderate complexity, moderate scale



Ray downloads per quarter

Gen AI - Pretraining (2023-2024)

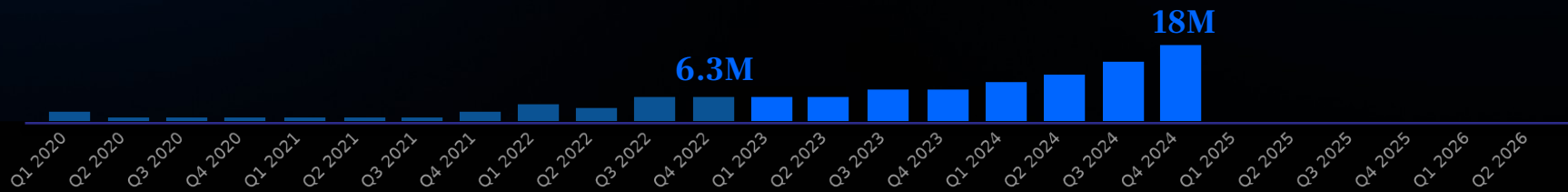
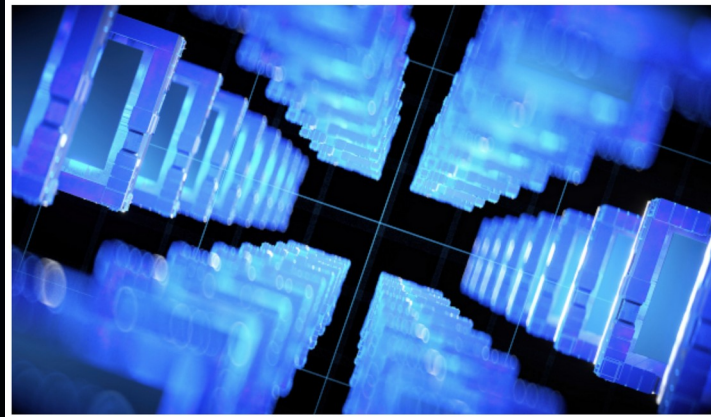
Data processing	Large text processing
Training	Dense models (Data-parallel, model-parallel)
Serving	Single-node models

moderate complexity, **large** scale

How Ray, a Distributed AI Framework, Helps Power ChatGPT

Anyscale co-founder Ion Stoica explains the open source platform Ray, and how it scales Machine Learning in this new era of generative AI.

Jan 17th, 2023 7:40am by Richard MacManus

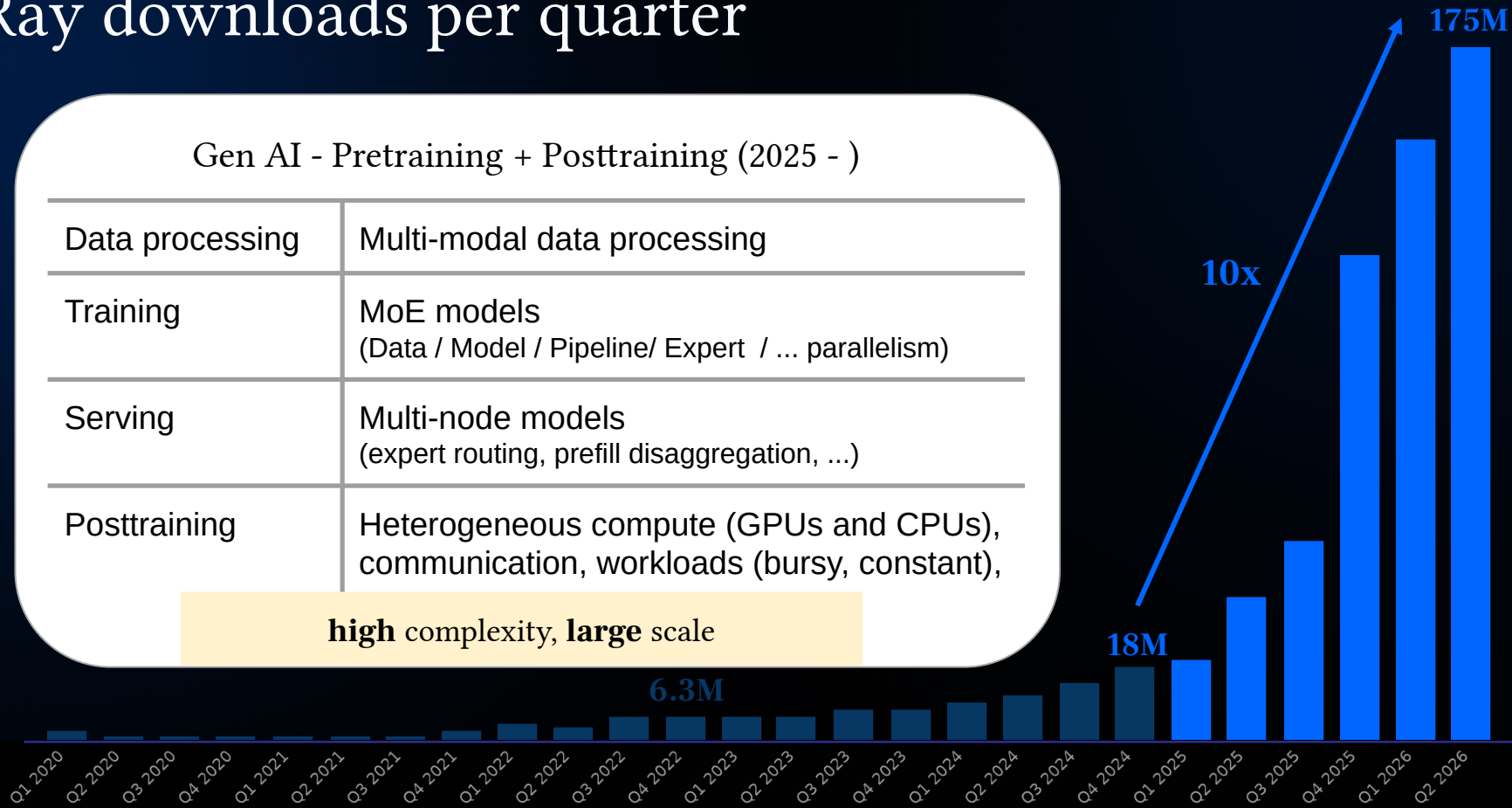


Ray downloads per quarter

Gen AI - Pretraining + Posttraining (2025 -)

Data processing	Multi-modal data processing
Training	MoE models (Data / Model / Pipeline/ Expert / ... parallelism)
Serving	Multi-node models (expert routing, prefill disaggregation, ...)
Posttraining	Heterogeneous compute (GPUs and CPUs), communication, workloads (bursy, constant),

high complexity, **large** scale



Coming full circle...

Nearly every open-source RL framework is built on Ray

Framework	Training engine	Serving engine	Orchestrator
TRL (Hugging Face)	Hugging Face Trainer	Hugging Face, vLLM	—
Verl (ByteDance)	FSDP, DeepSpeed, Megatron	vLLM, SGLang	Ray
OpenRLHF	DeepSpeed	Hugging Face, vLLM	Ray
AReaL (Ant Group)	DeepSpeed, Megatron	vLLM, SGLang	Ray
Prime RL	FSDP	vLLM	—
ROLL (Alibaba)	DeepSpeed, Megatron	vLLM, SGLang	Ray
NeMo-RL (Nvidia)	FSDP, Megatron	vLLM	Ray
SkyRL (UC Berkeley)	FSDP, DeepSpeed	vLLM, SGLang, OpenAI	Ray
SLIME (Z.ai)	Megatron	SGLang	Ray
RAGEN	Verl backend	Hugging Face, vLLM, SGLang	Ray



Workload frameworks

verl · NeMo Curator · SkyRL · Miles · Data-Juicer
Dynamo · NeMo-RL · Daft · ROLL · AIBrix · AReal

Training and inference engine

PyTorch · Jax · Megatron · vLLM · SGLang

Distributed compute

Ray

Orchestration platform

Kubernetes

Infrastructure

GB300 NVL72

VR200 NVL72

TPU v7 Ironwood

Trainium3

InfiniBand

RoCE

What is new? Scale, Reliability and Performance

Ray core

Topology aware scheduling for TPU Pods and Nvidia Blackwell GPUs

Ray Direct Data Transfer (RDT) with RDMA on NIXL: **7.5x faster** data transfer

10x faster startup for 32,000 actors

What is new? Scale, Reliability and Performance

Data processing

Scaling up to **10x more GPUs**

NVIDIA cuDF integration for **3x better TCO**

Ray core

Topology aware scheduling for TPU Pods and Nvidia Blackwell GPUs

Ray Direct Data Transfer (RDT) with RDMA on NIXL: **7.5x faster** data transfer

10x faster startup for 32,000 actors

What is new? Scale, Reliability and Performance

Data processing

Scaling up to **10x more GPUs**

NVIDIA cuDF integration for **3x better TCO**

Training

Native TorchFT integration

Elastic JAX training on TPU

Ray core

Topology aware scheduling for TPU Pods and Nvidia Blackwell GPUs

Ray Direct Data Transfer (RDT) with RDMA on NIXL: **7.5x faster** data transfer

10x faster startup for 32,000 actors

What is new? Scale, Reliability and Performance

Data processing

Scaling up to **10x more GPUs**

NVIDIA cuDF integration for **3x better TCO**

Training

Native TorchFT integration

Elastic JAX training on TPU

Serving

Up to **40x higher** overall throughput

25x higher throughput on decode-heavy LLM workloads

Ray core

Topology aware scheduling for TPU Pods and Nvidia Blackwell GPUs

Ray Direct Data Transfer (RDT) with RDMA on NIXL: **7.5x faster** data transfer

10x faster startup for 32,000 actors

What is new? Scale, Reliability and Performance

Data processing

Scaling up to **10x more GPUs**

NVIDIA cuDF integration for **3x better TCO**

Training

Native TorchFT integration

Elastic JAX training on TPU

Serving

Up to **40x higher** overall throughput

25x higher throughput on decode-heavy LLM workloads

Posttraining

10x faster weight sync with RDT

Sandboxes that scale to **100k in 30 seconds**

Ray core

Topology aware scheduling for TPU Pods and Nvidia Blackwell GPUs

Ray Direct Data Transfer (RDT) with RDMA on NIXL: **7.5x faster** data transfer

10x faster startup for 32,000 actors

Going forward

Accelerate supporting cutting edge AI demands

How?

- Rewrite large components of Ray to improve scale and performance
- Integrate closer with the ecosystem, in particular Kubernetes, for better resource utilization, and observability
- Work closer to hardware providers to improve support

SIG Accelerator: a vendor-neutral roadmap for Ray on accelerators

Dev and testing infra

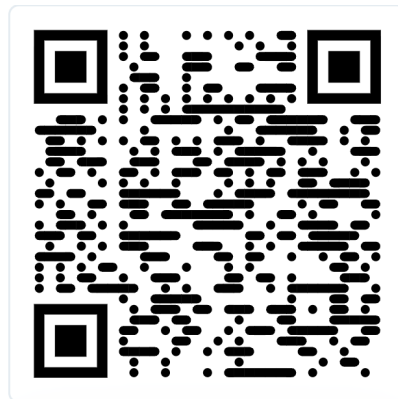
Shared hardware access, reproducible environments, broader CI coverage

Hardware-aware scheduling

Vendor-neutral hardware interfaces, topology discovery, hardware visibility

Networking and I/O

Direct device-to-device transfer over NVLink, RDMA and TPU ICI



**Join us: #sig-accelerators on the Ray Slack
or at the Inference & RL Coffee Club today!**

Thank you!