WeRide Maximizes GPU Utilization and Manages Costs with Weka

WeRide

We built a GPU farm, and we needed a high-speed data pipe to feed it. We evaluated open source solutions, HDFS, and the public cloud.

We chose Weka for its ability to provide cost-effective, high-bandwidth I/O to our GPUs, product maturity, customer references, and stellar on-demand support.

Paul Liu, Engineering Operations Lead, WeRide

CHALLENGES

- Provide high I/O bandwidth to keep GPU farm saturated with data
- Stay within Engineering and IT budget constraints
- Deliver high performance for mixed workloads with lots of metadata
- Enable hybrid implementation
- Find a hardware-agnostic solution

BENEFITS

- Maximum utilization of the GPU farm
- Cost-efficiency
- Superior performance and flexible deployment
- Freedom from vendor lock-in
- Storage as a utility
- Stellar Technical Support

The Weka File System (WekaFS™), the world's fastest shared parallel file system from WekaIO™ (Weka) maximizes GPU investments for WeRide by delivering high I/O bandwidth at scale to its GPU Farm. WeRide, headquartered in Guangzhou, China, with research and development (R&D) in Silicon Valley and China, is a smart mobility company that introduces leading Level 4 (L4) autonomous driving technologies with the charter to provide new mobility services that will be accessible and beneficial to everyone.

WeRide focuses on strategic alliances among Artificial Intelligence (AI) technology companies, car makers, and mobility service platform providers. WeRide RoboTaxi, a joint venture company of WeRide, Baiyun Taxi, and SCI (Guangzhou) Group, has a robotaxi program in Guangzhou, China that covers 144.65 square kilometers in the Huangpu and Guangzhou Development Districts, providing the locals with robotaxi services on their daily commute. This is the first robotaxi service that is open to the public in China.

THE CHALLENGE: MAXIMIZE INVESTMENT IN GPU FARM WHILE MANAGING COSTS

Established in 2017, WeRide is a multi-faceted AI start-up that works on L4 autonomous driving vehicles and partnerships with transportation platform providers that support robotaxi services. This requires IT infrastructure that supports the full AI data pipeline from the edge, to the core, to the cloud.

The company processes data at the petabyte (PB) level, ingesting a daily mixture of large video and image files generated from information collected from over 2 million kilometers of driving distance. At the very first autonomous driving data-labeling center in Anqing, China, WeRide produces millions of high-quality labeling data each day. The data is annotated at the core, trained by the AI model on the cloud-based cluster, and then fed back to the on-premises AI engine.

The WeRide team faced the challenge of managing hundreds of terabytes of data with files of mixed sizes through the entire AI pipeline and keeping its GPU farm fully utilized in a cost-efficient manner.
The team needed a storage solution that would:

- maximize utilization of the GPU farm
- offer a hybrid implementation model to reduce datacenter footprint
- handle mixed workloads with lots of metadata
- provide flexibility through hardware-agnostic compatibility with commodity servers
- deliver the best economic value for capacity planning and future performance
- furnish stellar technical support to complement the product strengths
- make storage a utility for the end-user.

THE SOLUTION: HYBRID IMPLEMENTATION OF WEKA SOFTWARE ON COMMODITY SERVERS AND AWS

Maximizing GPU investment was the primary decision driver for the Engineering team at WeRide. Extensive cost-analysis of various scenarios utilizing open source solutions and HDFS to feed data to the GPU farm proved that the alternative approaches would not to be cost-effective. Ultimately, WeRide chose a hybrid implementation using WekaFS.

WekaFS has a two-tier architecture that takes commodity servers and the public cloud and presents them as a single hybrid storage solution. On premises, WeRide manages hundreds of TBs of NVMe flash on Weka, using commodity Intel x86-based servers from AMAX and NVIDIA® Mellanox Ethernet network switches. WeRide was also the first

customer to use Weka to deliver high-bandwidth I/O to its GPU resources in AWS China. The WeRide team lauds the collaboration with the Weka Technical Support team that made for a seamless public cloud implementation.

Weka fulfills the team's desire to have a storage utility model that is completely hardware-agnostic and transparent to the end-user. With Weka, anyone in the organization can map a drive — the consumers of WeRide's data can access it within a predictable time frame — and the underlying data management, movement, and synchronization is transparent to the end-user. Deciding to implement a commercial product, rather than open source, was a business decision for WeRide. The company wanted its engineers to focus on innovative product development that grows top-line revenue and not on storage administration.

"WekaFS is cloud-optimized and architected to provide high bandwidth I/O to GPU-enabled compute clusters playing a big role in enabling digital transformation.

We are pleased to have been the solution provider of the Weka software licenses for WeRide to drive their AI workflow.

Chris Saso, CTO, Dasher Technologies, a WIN Leader Partner

WEKA DELIVERS HIGH I/O BANDWIDTH TO GPUs

Beyond delivering high-bandwidth I/O to data-hungry GPUs to keep them fully utilized, WekaFS is perfect for data-intensive applications, whether hosted on-premises or in the public cloud. It is a POSIX file system that scales performance linearly as the GPU server farm grows, so WeRide will not have to compromise performance with future expansions. And since WeRide is running WekaFS on GPU servers in converged mode, creating a single namespace from all the locally attached NVMe drives, they will not have to invest in expensive hardware for their on-premises cluster. WekaFS is a fully parallel and distributed file system that has been built from a clean-sheet design to leverage high-performance NVMe flash. Both data and metadata are distributed across the entire storage infrastructure to ensure massively parallel access to the NVMe drives.

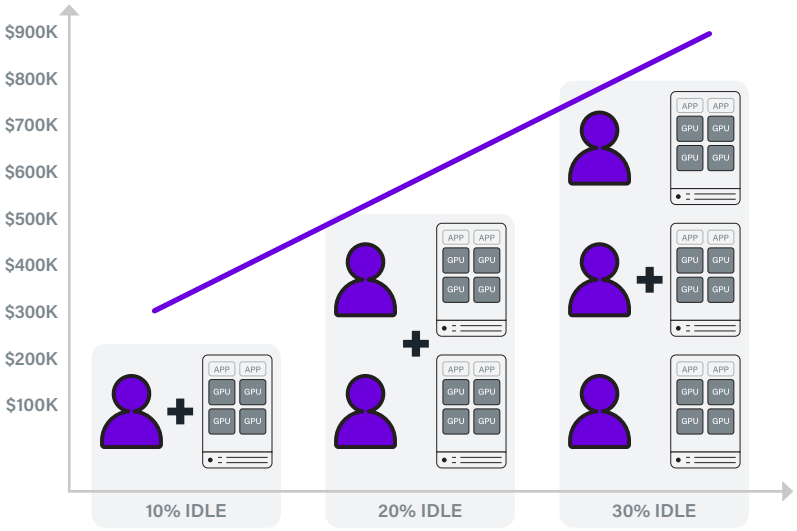
THE WEKA INNOVATION NETWORK™ (WIN) DELIVERS SOLUTIONS THAT SOLVE BIG PROBLEMS

WeRide is the first customer to use Weka on AWS China and was able to significantly reduce its datacenter footprint with a hybrid implementation. AWS China is a WIN Innovation partner, and Weka is an AWS Advanced Technology Partner. WekaFS licenses were procured through Dasher Technologies, a WIN Leader partner headquartered in Campbell, California.

BENEFITS AND ROI

WeRide was able to realize return on investment and key business benefits by choosing WekaFS:

- **Acceleration of Product Development:** WeRide speeds up product development by using a proven product with a stellar post-sales technical team
- **Maximum Utilization of the GPU Farm:** WeRide evaluated several options and determined that Weka was the best storage solution to maximize the utilization of its GPU farm and eliminate idle time waiting for data
- **Hardware-Agnostic Software:** Weka's compatibility with any Intel x86-based server enabled the use of commodity servers
- **Cost-Efficiency:** Weka's hybrid implementation model reduced datacenter footprint and maximized the investment in GPU resources
- **Stellar Technical Support:** Weka's Technical Support reduced the burden of storage administration and freed up Engineering resources to focus on innovation and product development
- **Storage as a Utility:** WeRide's data is available to anyone in the organization in a predictable time, no matter where they are.



10% Idle Time = 1 Additional Server + 1 Data Scientist of Cost

The cost of an idle GPU-enabled compute clusters

For more information or to locate a partner in the Weka Innovation Network, go to: https://www.weka.io/partners.
For more information on WeRide, go to: https://www.weride.ai/

WEKA

910 E Hamilton Avenue, Suite 430, Campbell, CA 95008 T: 408.335.0085 E: info@weka.io www.weka.io