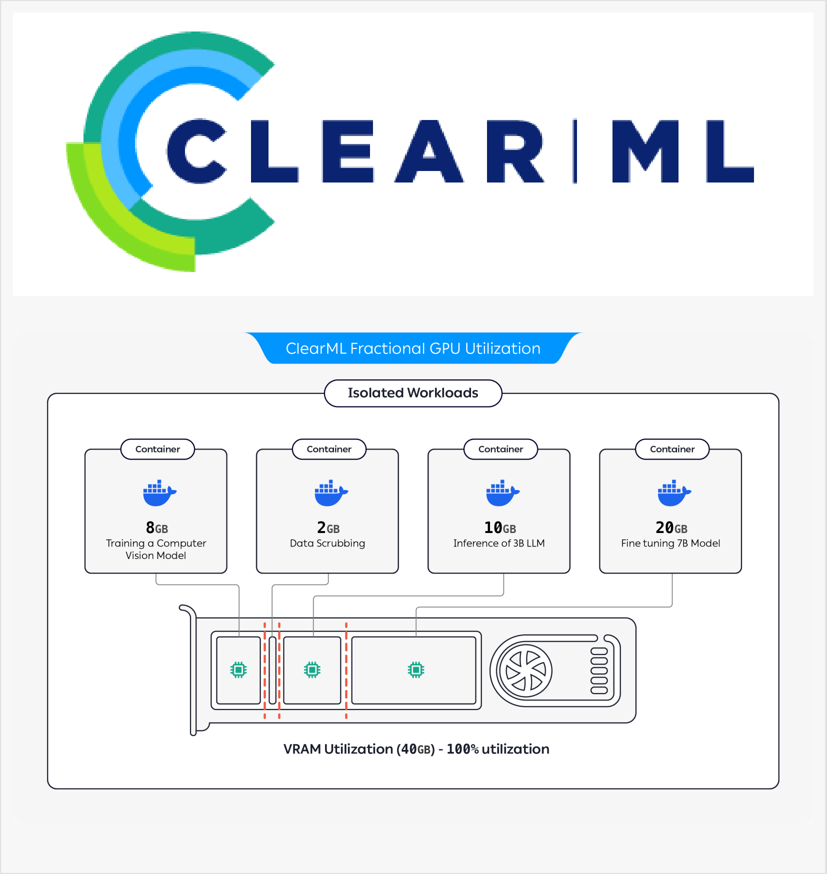
EINPRESSWIRE

ClearML Announces Free Fractional GPU Capability for Open Source Users, Enabling Multi-tenancy for All NVIDIA GPUs

Users can now maximize compute utilization, run more efficient AI and HPC workloads, and achieve higher ROI from their GPU investments



SAN FRANCISCO, CA, US, March 18, 2024 /EINPresswire.com/ -- ClearML, the leading open source, end-to-end foundational platform for unleashing AI in the enterprise, today announced the release of open source fractional GPU functionality, enabling users to optimize their GPU utilization for free.

With this new functionality, DevOps professionals and AI Infrastructure leaders can take advantage of NVIDIA's time-slicing technology to safely partition their GTX™, RTX™, and datacenter-grade, MIG-enabled GPUs into smaller fractional GPUs to support multiple AI and HPC workloads without the risk of failure. This allows organizations to optimize usage of their current compute and legacy infrastructure in the face of increasing AI workloads resulting from the rise of Generative AI.

Unlike earlier AI/ML initiatives running on GPUs, Generative AI requires a high GPU compute load, with dedicated AI infrastructure supporting inference with low latency. Without an effective solution to gain more computing power from existing infrastructure, organizations are forced to purchase additional compute or stagger projects to fit their compute capacity, resulting in prolonged time to value while risking their competitive edge. In fact, according to a recent survey conducted by ClearML (https://go.clear.ml/the-state-of-ai-infrastructure-at-scale-2024), 96% of respondents said they plan to expand their AI compute infrastructure, signaling the recognition of an already-present strain on resources.

ClearML's new open source fractional GPU offering, enables toggling between smaller R&D jobs to large, demanding training jobs. ClearML now also supports multi-tenancy with partitions that offer secure and confidential computing with hard memory limitation, ensuring predictable and stable performance. Driver-level memory level monitoring ensures jobs run without impacting each other and provide increased flexibility for multiple smaller long-running jobs as well as shorter high-demand jobs.

Multiple stakeholders, such as Data Science, AI/ML Engineering, and DevOps, can run isolated parallel workloads such as graphics, model training, or inference on a single shared compute resource, resulting in increased efficiency, reduced costs, and faster time to value. Admins can create pre-configured containers for team members to self-serve and easily run on Kubernetes or bare metal.

"With our new free offering now supporting fractional capabilities for the broadest range of NVIDIA GPUs than any other company, ClearML is democratizing access to compute as part of our commitment to help our community build better AI at any scale, faster," says Moses Guttmann, CEO and Co-founder of ClearML. "We hope that organizations that might have a mixture of infrastructure are able to use ClearML and get more out of the compute and resources they already have."

ClearML Enterprise customers can already harness another level of GPU utilization with dynamic multi-instance GPU capabilities, as well as robust policy management that applies sophisticated logic to managing quotas and over-quotas, priorities, and job queues. Enterprise customers also currently have access to additional tools for minimizing GPU costs when leveraging cloud computing, either as a hybrid setup or spillover. In fact, ClearML Autoscaling ensures cloud compute is only used when needed and automatically spun down after a set idle period.

Try Now
To learn more about ClearML's free open source GPU capabilities, visit our GitHub page (https://github.com/allegroai/clearml-fractional-gpu) or see us in our booth (#1702) in the MLOps Pavilion at NVIDIA GTC for a live demo. We'll be there through March 21 at the San Jose Convention Center.

Get Started With ClearML
Get started with ClearML by using our free tier servers (https://app.clear.ml) or by hosting your own (https://clear.ml/docs/latest/docs/deploying_clearml/clearml_server). If you need to scale your ML pipelines and data abstraction or need unmatched performance and control, please request a demo (https://clear.ml/demo). To learn more about ClearML, please visit: https://clear.ml/.

About ClearML
As the leading open source, end-to-end solution for unleashing AI in the enterprise, ClearML is

used by more than 1,600 enterprise customers to develop a highly repeatable process for their end-to-end AI model lifecycle, from product feature exploration to model deployment and monitoring in production. Use all of our modules for a complete ecosystem or plug in and play with the tools you have. ClearML is an NVIDIA DGX-ready Software Partner and is trusted by more than 250,000 forward-thinking Data Scientists, Data Engineers, ML Engineers, DevOps, Product Managers and business unit decision makers at leading Fortune 500 companies, enterprises, academia, and innovative start-ups worldwide. To learn more, visit the company’s website at https://clear.ml.

Adam Brett
Crenshaw Communications
Clearml@crenshawcomm.com
Visit us on social media:
Twitter
LinkedIn

This press release can be viewed online at: https://www.einpresswire.com/article/696880708