EINPRESSWIRE

ClearML Unveils Pioneering AI Infrastructure Control Plane for Ultimate AI Compute Resource Management and Utilization

AI Builders Achieve Unmatched Control Over Their Infrastructure with the Simplest Unified Solution for Managing Large-Scale AI and HPC Workloads

SAN FRANCISCO, CA, US, August 13, 2024 /EINPresswire.com/ -- ClearML, the leading open source, end-to-end platform for unleashing AI in the enterprise, today announced the launch of its pioneering AI Infrastructure Control Plane as a universal operating system for AI infrastructure. Designed for AI builders, IT teams, and DevOps teams, the solution provides unmatched control, visibility, and efficiency in managing, orchestrating, and scheduling GPU compute resources across hybrid environments. AI and IT teams now have access to a fully flexible and customizable solution with a future-proof AI Infrastructure Control Plane that is cloud-, silicon-, and vendor-agnostic.

AI deployments are becoming more complex, with AI/HPC workloads spanning across cloud, edge computing, air gapped, and on-premises data center infrastructure at large scale. Effectively managing and orchestrating end-to-end GenAI apps, internal knowledge base search engines, recommender systems, and other workloads necessitates an advanced end-to-end AI Infrastructure Control Plane with complete orchestration, automation, and scheduling capabilities to enhance performance both at the system level and across the underlying AI infrastructure and tech stack.

Customers can now use ClearML to gain ultimate control over their organization’s AI infrastructure management and maximize optimization, down to a fraction of a GPU. AI Builders can build, train, and deploy AI/MLmodels at enterprise scale on any AI infrastructure. They can work on shared data and build models seamlessly from anywhere in the world on any AI workload, compute type, or infrastructure – regardless of whether they’re on-prem, cloud, or hybrid; with Kubernetes, Slurm, PBS, or bare metal; and any type of GPU – making it the easiest solution for managing computing clusters for AI and HPC workloads.

That’s important because the new solution addresses critical challenges faced by organizations in managing and optimizing their AI compute resources, such as:

Rising Demand for Compute Power: The exponential growth of AI and machine learning applications has significantly increased the demand for GPU compute resources. Organizations need efficient ways to manage and allocate these resources to keep up with the demand.

Cost Management: As cloud computing costs rise, organizations are under pressure to optimize their compute usage and reduce wastage. Effective resource management is essential to control expenses and achieve cost-effectiveness.

Complex Infrastructure: Many organizations operate in hybrid environments, combining on-premise, cloud, air gapped, and multi-cloud resources. Managing such complex infrastructures requires advanced tools that provide visibility, governance, and control across all resources.

Scalability and Flexibility: To stay competitive, businesses must scale their AI operations quickly and efficiently while future proofing their investments and AI Infrastructure. They need agnostic, open-source solutions that can adapt to their evolving needs without being locked into specific vendors, AI chips, cloud providers, AI/ML frameworks, or closed-garden technologies.

“Our new AI Infrastructure Control Plane enables organizations to fully leverage their AI infrastructure and GPU resources, ensuring optimal performance and cost-effectiveness at industrial scale,” said Moses Guttmann, Co-founder and CEO of ClearML. “This new solution is vital as the demand for compute power rises, enabling our customers to innovate without barriers or lock-ins and scale efficiently. As organizations grow and their AI workloads increase, ClearML’s AI Infrastructure Control Plane seamlessly scales with them. This ensures that they can continue to innovate and expand economically without facing bottlenecks or limitations in their compute resources that fuel their AI innovation.”

Product Highlights:

- Comprehensive Control and Visibility: Gain governance, visibility, and control over your entire shared compute infrastructure, allowing precise resource allocation and management by team or project to provide easier access and resource allocation across IT, data science, and application development teams.
- Fractional GPUs for Maximized Compute Utilization: Utilize dynamic fractional GPUs to ensure maximum utilization, reducing wastage and optimizing performance.
- Seamless Multi-Tenancy Management: Securely share compute infrastructure across multiple tenants with enterprise-grade security and detailed usage reporting for accurate billing.
- Real-Time Detailed Reporting: Monitor compute usage in real-time with granular reports on

hours connected, data storage, API calls, and more, enabling accurate chargebacks and billing.
- Single Pane of Glass Management: View and manage your entire compute infrastructure on a single screen, making strategic decisions on load balancing and resource allocation.
- Superior Efficiency and ROI: Execute 10X more AI and HPC workloads on customers' existing infrastructure.

"ClearML's open-source, fractional GPU capabilities are a game-changer in the industry," said Lior Hakim, CTO of Hour One. "Now, every GPU can be partitioned and morphed into a powerhouse to enable optimized compute resource utilization and reliable workload performance. By supporting any GPU – which can be quite expensive – ClearML is helping us to maximize our AI infrastructure investments with no compromise on performance."

Key Features:

- Resource Allocation and Policy Management: Dynamically allocate resources, set quotas and priorities, manage utilization with advanced user management policies, and increase infrastructure efficiency.
- Autoscalers for Cloud Optimization: Control cloud instance usage with autoscalers for AWS, Azure, and GCP, automatically shutting down idle instances to save costs.
- Dynamic Fractional GPUs: Guarantee maximum utilization by running multiple AI or HPC jobs on a single GPU, supporting both MIG-enabled and standard NVIDIA GPUs.
- On-Demand Self-Serve Compute: Enable stakeholders to access compute resources without IT intervention, ensuring job queues are always full and resources are optimally used.
- Flexible Deployment: Support for on-premise, air-gapped, hybrid, or multi-cloud setups, making it ideal for teams working remotely or requiring diverse infrastructure configurations.

For more information, visit our website or contact our sales team to schedule a demo: https://clear.ml/demo.

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 AI builders and IT teams at leading Fortune 500 companies, enterprises, academia, public sector agencies, and innovative start-ups worldwide. To learn more, visit the company's website at https://clear.ml.

Adam Brett
Crenshaw Communications
+1 212-367-9700
email us here

Visit us on social media:
X
LinkedIn