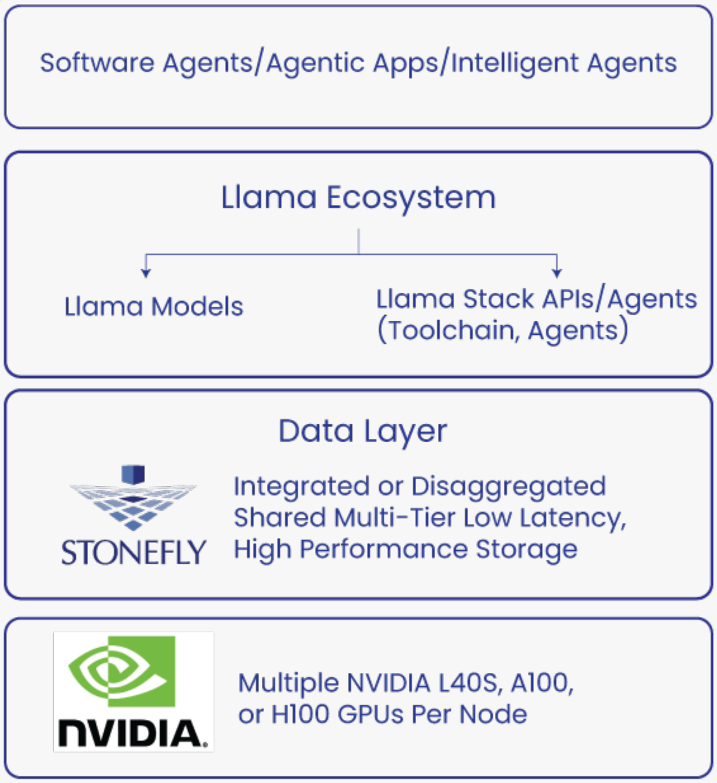
EINPRESSWIRE

Llama Ecosystem on NVIDIA GPU-Based AI Servers with High Performance Integrated Shared NVMe Storage

HAYWARD, CA, UNITED STATES, October 14, 2024 /EINPresswire.com/ -- StoneFly, Inc. (isci.com), a leading provider of enterprise storage, hyperconverged, backup and disaster recovery, cloud, and AI storage solutions, announces the integration of its NVIDIA GPU-powered AI servers with the LLama ecosystem. This integration brings the full capabilities of LLama's AI and machine learning (ML) tools to StoneFly's high-performance, scalable, and modular AI server platform, providing enterprises with a high-performance AI solution with integrated optional storage for demanding data-centric applications.


Software Agents/Agentic Apps/Intelligent Agents
Llama Ecosystem
Llama Models
Llama Stack APIs/Agents (Toolchain, Agents)
Data Layer
STONEFLY
Integrated or Disaggregated Shared Multi-Tier Low Latency, High Performance Storage
NVIDIA.
Multiple NVIDIA L40S, A100, or H100 GPUs Per Node

StoneFly AI Servers for Llama Ecosystem

StoneFly's AI servers, powered by NVIDIA L40s, A100, and H100 GPUs, now integrate with the LLama stack, enabling enterprises to streamline various stages of AI development, including model training, fine-tuning, and production deployment. The LLama stack provides open-source AI models and software, while StoneFly's modular AI servers offer the necessary performance and optional integrated multi-tiered, shared/disaggregated storage, and scalability to efficiently manage these workloads in real-time AI applications.

Built and Tested for Data Scientists, Analysis, Developers, and Enterprise Use-Cases

Creating and testing AI applications requires high performance, storage for large data sets, scalability, and ransomware protection which makes the total cost of ownership (TCO) high and return on investments (ROIs) challenging. StoneFly's AI servers solve these challenges with a consolidated turnkey solution.

The reference architecture shows the simplicity and efficiency of using the Llama ecosystem on the StoneFly AI servers streamlining AI development, testing, and utilization in enterprise data centers.

Benefits of Using StoneFly AI Servers for Llama Ecosystem

StoneFly's NVIDIA GPU AI servers are designed to deliver high performance for AI and ML applications, including natural language processing (NLP) and large-scale data analytics. The integration with the LLama ecosystem improves this by providing seamless compatibility with leading AI tools and frameworks.

- High-Speed Data Processing: The built-in dual-controller, active/active, multiple GPU per node architecture ensures faster data throughput and processing with zero bottlenecks, making it ideal for AI/ML applications that demand high compute power and real-time performance.
- Flexible Storage Options: The StoneFly AI servers support optional integrated and disaggregated shared multi-tiered hot and cold storage. This enables users to set up storage in the same appliance or in a disaggregated shared repository/appliance facilitating users to tailor their AI environment as needed.
- Optional Ransomware-Proof Security: StoneFly's Air-Gapped Vault®, Always On-Air® Gapped backups, and immutable storage technology ensures that critical AI and ML workloads remain safe from cyber threats, providing robust ransomware-proof data protection for LLama workloads.
- Easy to Set Up and Manage: The AI servers are easy to set up and come with an intuitive real-time graphical interface that provides detailed insights into system resource usage, such as CPU and network performance, simplifying ongoing management.
- Cost-Effective AI Servers: StoneFly AI servers offer a lower initial purchase cost compared to other AI servers in the market with similar high-performance specifications, reducing CapEx. Flexible storage options allow users to configure storage within the same appliance or extend it through modular repositories, helping minimize unnecessary hardware costs. Additionally, these servers scale seamlessly without the need for forklift upgrades, ensuring that enterprises can grow as needed while maintaining a lower total cost of ownership (TCO).

Availability

StoneFly's AI servers with LLama ecosystem support are available immediately. For more information on how to leverage this solution for your AI and ML workloads, visit https://stonefly.com/nvidia-gpu-ai-servers/.

About StoneFly, Inc.

StoneFly, Inc. is a leading provider of enterprise-grade storage, hyperconverged, and backup and disaster recovery solutions. With over two decades of experience, StoneFly offers innovative, scalable, and reliable data management solutions that simplify enterprise workloads and provide seamless protection and accessibility for mission-critical data. Learn more at www.stonefly.com.

George Williams
+ +1 5102651616
email us here
StoneFly

Visit us on social media:
Facebook
X
LinkedIn
Instagram
YouTube

This press release can be viewed online at: https://www.einpresswire.com/article/751664824