CASE STUDY

NextComputing and Ampere Bring AI to the Edge for Cyber Threat Hunting

INTRODUCTION

NextComputing, a leader in high-performance portable computing, wanted to offer powerful AI capabilities for defense and cyber response teams in the field. The existing solutions were bulky, power-hungry, and inefficient, making them unsuitable for portable deployments.

By integrating Ampere's "AI Compute" capabilities to create NextComputing's Nexus Fly-Away Kit, NextComputing delivered scalable AI inferencing and processing in a small carry-on form factor, which enables teams on the go to leverage AI-powered tools for rapid and effective response in their critical missions.

AMPERE PRODUCTS USED

128 core Ampere® Altra®Max Processors (2.6 GHz). The Nexus Fly-Away Kit has up to 4 compute nodes, which is 512 cores total.

ENGINEERING SOLUTION

The Fly-Away Kits are a compact and portable solution for AI Compute. Ampere provides efficient AI inference, up to 512 cores per carry-on, and supports large amounts of memory. This enables AI models like H2O Danube to run efficiently at the edge and eliminates bulky multi-box setups. The management of the Fly-Away Kit cluster, containers, and virtualization is provided by Rancher Government Solutions' Rancher and Harvester for operators to quickly select and deploy according to the scenario using application templates.

BENEFITS

AI Compute: portable cluster, containers, and VMs running complex heterogenous workloads with "freedom of interference".

AI Inference Efficiency: Ampere processors provided high-performance AI inference per watt[1], ensuring maximum energy efficiency for edge deployments.

High Core Count: With up to 512 cores, the solution processed parallel AI workloads seamlessly in compact hardware.

Large Memory, Small Footprint: Enabled AI Inference models such as H2O Danube to provide high throughput and low latency[2].

Compact Portability: Condensed AI compute and networking gear into a single carry-on, eliminating bulky, multi-box setups.

Edge Reliability: Delivering performance in rugged environments with little to no infrastructure, empowering teams to recognize and neutralize threats.

COMPANY DESCRIPTION

NextComputing designs and manufactures high-performance, portable computing solutions for AI, networking, storage, and real-time processing. With expertise in cybersecurity and intelligence, NextComputing develops ruggedized systems such as the Nexus Fly-Away Kit with Ampere processors, enabling secure and efficient execution of AI inference workloads in mission-critical environments.

By integrating Ampere Cloud Native Processors, NextComputing delivers dense compute, energy-efficient solutions[2] solutions that power autonomous AI-driven decision-making for cybersecurity, defense, and intelligence operations. With a focus on scalability, mobility, and real-time performance, NextComputing equips organizations with the tools to tackle the toughest challenges and adversaries.

CHALLENGES

Deploying AI for cyber threat hunting in the field posed significant challenges. Legacy solutions consumed tremendous power, generated excessive heat, and were unsuitable for the size, weight and power (SWaP) requirements of a portable solution.

NextComputing needed to build a solution that would maximize the AI inferences per watt, parallelize workloads across a large number of cores in a cluster, and have the large memory required to run large language models (LLMs) such as H2O Danube model, trained on the MITRE ATT&CK database to autonomously recognize cyber threats and help the operator neutralize them.

"Ampere Altra Max processors have completely changed how we bring AI to the edge. Instead of relying on bulky, power-hungry multi-box setups, deployed teams can have a solution that leverages the Nexus with Ampere for real-time AI inference in a compact, energy-efficient form factor."

— Bob Labadini, CTO, NextComputing

About Ampere

Built for sustainable cloud computing, Ampere Computing's Cloud Native Processors feature a single-threaded, multiple core design that's scalable, powerful, and efficient. Learn more

See our solutions for a variety of demanding workloads: amperecomputing.com/solutions

Visit our Developer Center: amperecomputing.com/developers

Disclaimer

All data and information contained in Disclaimer: All data and information contained in or disclosed by this document are for informational purposes only and are subject to change. This document may contain technical inaccuracies, omissions and typographical errors, and Ampere is under no obligation to update or correct this information. Ampere makes no representations or warranties of any kind, including express or implied guarantees of noninfringement, merchantability, or fitness for a particular purpose, and assumes no liability of any kind. All information is provided "AS IS." This document is not an offer or a binding commitment by Ampere.

This document is not to be used, copied, or reproduced in its entirety, or presented to others without the express written permission of Ampere®.

References

[1] AI Inference on Ampere Altra Max. https://amperecomputing.com/briefs/ai_solution_brief

[2] Rack level footnotes. https://amperecomputing.com/home/efficiency-footnotes