Scale Smarter: Unlocking Edge Hardware Value with Precision AI

LatentAI

NVIDIA

DELL Technologies

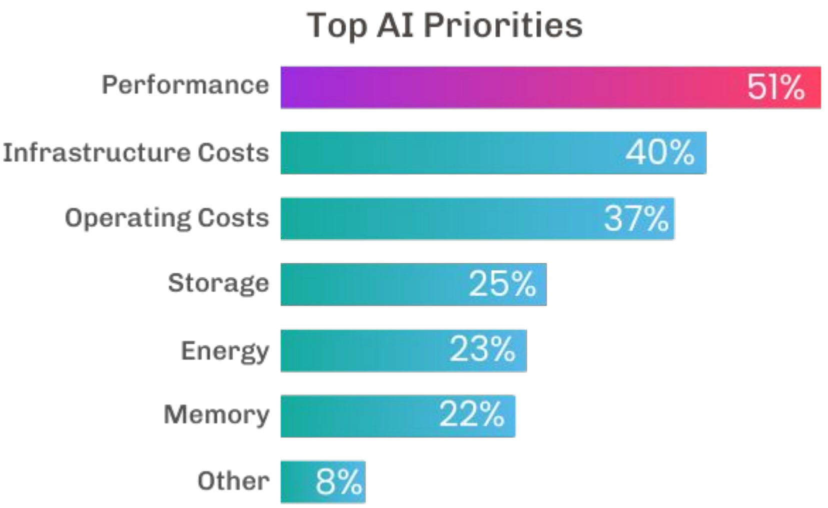
Scale Smarter: Unlocking Edge Hardware Value with Precision AI

More and more organizations are rethinking how they roll out AI, especially as edge computing starts lining up perfectly with their core business needs and goals. Research by Techstrong confirms that organizations are fundamentally reevaluating how they deploy AI: 51% prioritize performance, 40% target infrastructure costs, and 37% focus on operating costs.

But here's the catch—unoptimized setups can waste hardware potential with oversized models and sluggish inference while also driving up costs and extending the return on investment (ROI) timeline. Specialized hardware for AI implementations at the edge adds another layer to this challenge. Too many organizations overspend on high-end GPUs and CPUs without thoroughly assessing workload requirements.

That's where tools like Latent AI Efficient Inference Platform (LEIP) SDK come in, paired with rock-solid Dell Technologies hardware. LEIP Optimize trims down memory use and speeds up inference, while LEIP Deploy's parallel inference can cut GPU needs by more than tenfold—all without skimping on accuracy. The result? Leaner, faster AI that scales affordably across the edge, making the most of what you've got.

The Latent AI team put this to the test, pushing open-source models past standard ONNX-optimized baselines in two real-world scenarios, running on different hardware—including some Dell-powered setups. We compared the costs for an FP32 setup against an optimized LEIP INT8 version. The FP32 (32-bit floating point) mode is typical in standard machine learning frameworks like PyTorch or ONNX. On a Dell edge server with an NVIDIA® RTX™ 5000 Ada Generation GPU, LEIP Optimize **boosted returns by 92%,** turning solid hardware into an AI powerhouse. Then, on a Dell-supported NVIDIA® Jetson AGX Orin™ edge device, it delivered **76% more value** per dollar spent.



Scale Smarter: Unlocking Edge Hardware Value with Precision AI

What's LEIP all about?

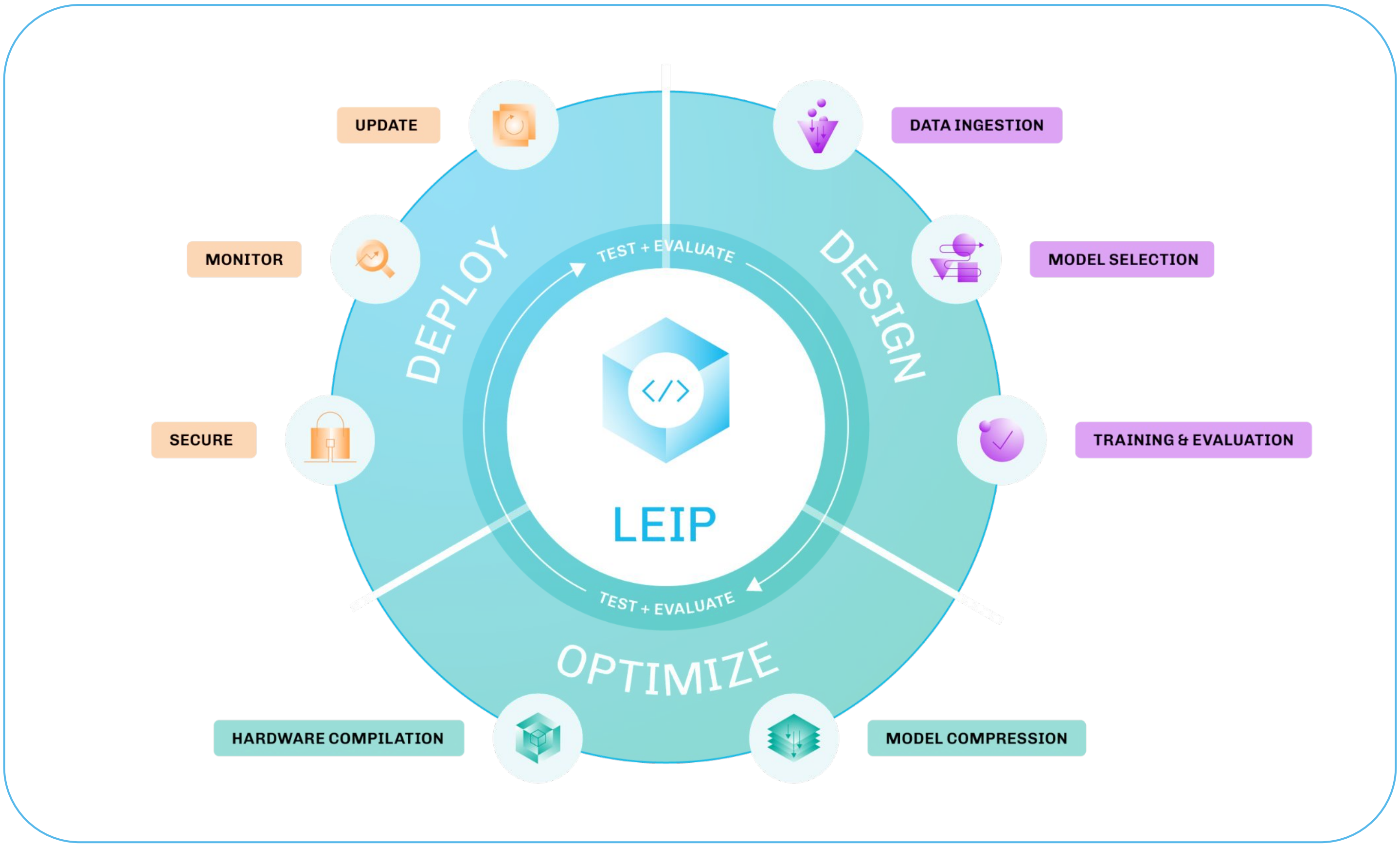
The **Latent AI Efficient Inference Platform** (LEIP) is an all-in-one toolkit for building super-fast, secure AI models for various hardware targets—including Dell's edge solutions. Here's what it brings to the table:

- Shrinks model size on disk by up to 5x on average
- Cuts RAM usage by up to 73%
- Improve inference speed by up to 73%
- Bit precision optimization with negligible accuracy loss
- Targets different hardware for quick prototyping
- Locks down models with encryption and watermarking

LEIP Optimize streamlines model-hardware optimization with automation that allows you to:

- Target hardware without in-depth knowledge or expertise
- Improve edge AI inference speed by optimizing the model and leveraging acceleration on target hardware

LEIP Deploy allows you to deploy to multiple hardware devices, monitor performance, and update your compiled models, all with a single runtime engine.

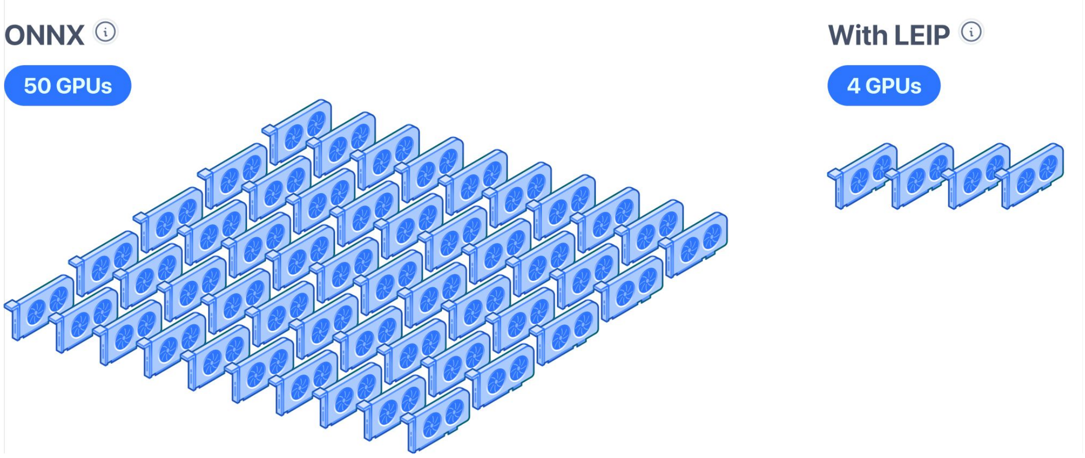


Here's the Proof

Optimized for edge server

We tested an anomaly detection model, EfficientAD, on a Dell edge server equipped with an NVIDIA® RTX™ 5000 Ada Generation GPU with 32 GB of RAM. Running 100 image streams at 15 Hz with the standard ONNX FP32 setup would've needed over 50 GPUs—costing around $224k. However, with LEIP Optimize tuning it to INT8, we got it down to just four GPUs. That setup, running LEIP Deploy, came in at a cool $18k.

What enables these cost savings are the LEIP-enabled reductions in RAM usage and inference time. The LEIP INT8 model uses less memory so that you can run more instances on each GPU, and it processes data faster—meaning more images per second. Net result: **over 10x savings on GPU costs**, all while maximizing that Dell hardware.

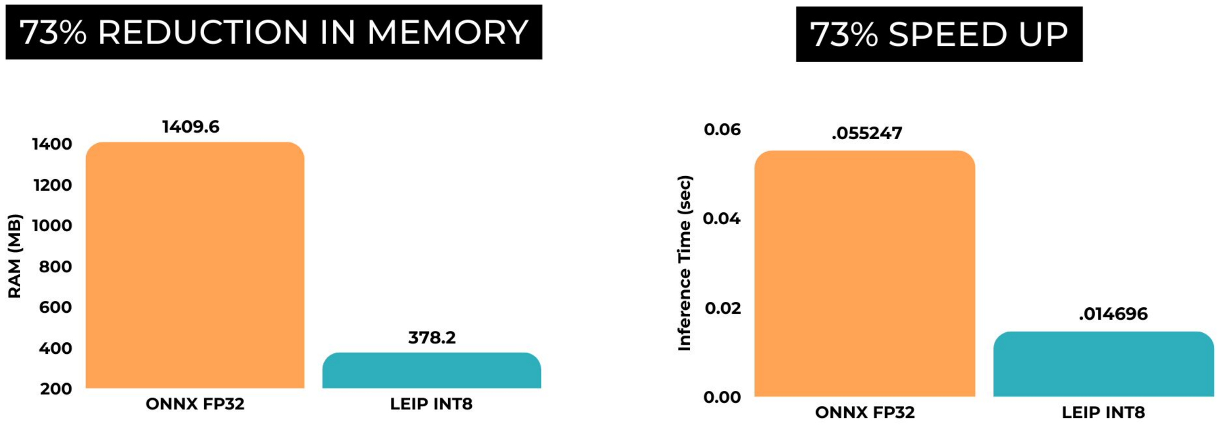


A LEIP-enabled solution for EfficientAD better utilizes the GPUs.

Scale Smarter: Unlocking Edge Hardware Value with Precision AI

Optimized for Laptops

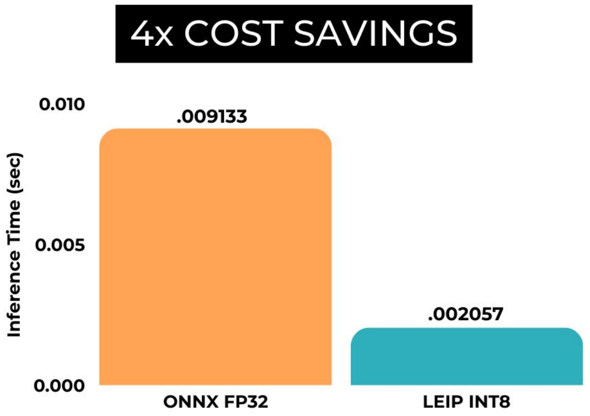
The same inference speed and RAM savings apply to a laptop. The figure below shows execution on a Dell Precision 7680, also equipped with an Ada Generation GPU. For the same EfficientAD model, the **RAM usage was reduced by 73%**, and the **inference speed improved by 73%**. That leaves more RAM for other processes and more GPU time for other models.



LEIP enables a reduction in RAM usage and interference time. The result is that hardware can be used to process more data quickly. This plot compares EfficientAD on an NVIDIA® RTX™ 5000 Ada Generation GPU.

Optimized for Edge Devices

Next up, we ran Yolov8n on an NVIDIA® Jetson AGX Orin™ for a facility-wide surveillance project—think 1000 video streams at 15 Hz. The ONNX FP32 setup needed 29 Orins ($47k), but the LEIP INT8 version? Just 7 Orins ($11k)—**a 4x cost drop**. Whether it's Dell edge servers or devices, LEIP makes inference fly.



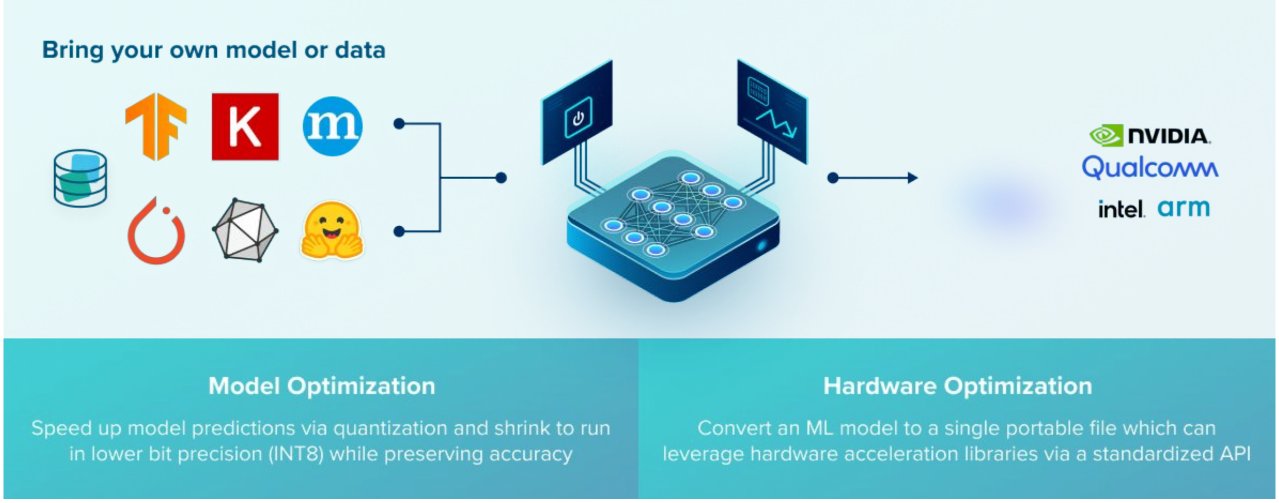
On both edge hardware (shown here) and server-grade hardware (above), LEIP enables faster inference times.

The Latent AI Solution: LEIP Optimize & Deploy

LEIP Optimize: Smarter Models, Happier Hardware

LEIP Optimize works its magic in two ways. First, it uses Post Training Quantization (PTQ) to shift from high-precision floating-point math (like FP32) to leaner integer math (like INT8). On Dell hardware, this means faster data transfers, smaller memory footprints, and quicker calculations—often a 2–4x speed boost.

It also picks the right precision for your hardware—like INT8 if your Dell server or NVIDIA® Jetson™ supports it natively—so you're not wasting cycles. After that, LEIP compiles the model into machine code, ditching slow interpretation for fast, efficient execution. You get a shared object file that works with C, C++, Python, and Java—perfect for Dell's edge ecosystem.



Model Optimization

Speed up model predictions via quantization and shrink to run in lower bit precision (INT8) while preserving accuracy

Hardware Optimization

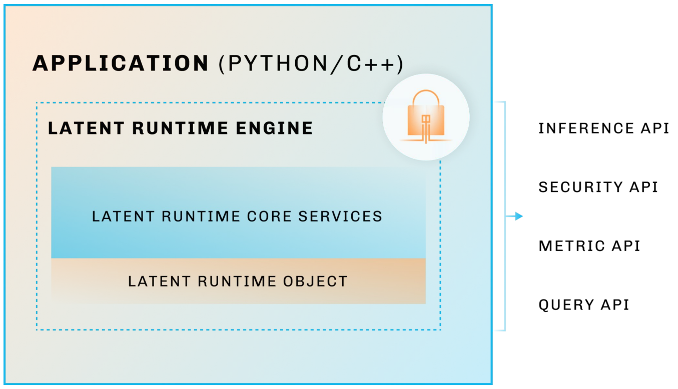
Convert an ML model to a single portable file which can leverage hardware acceleration libraries via a standardized API

LEIP Optimize quantizes and then compiles the model.

LEIP Deploy: Keeping Things Running Smoothly

LEIP Deploy is the runtime engine that ties it all together. Its API lets you load and run optimized models, monitor performance, and even swap models without rewriting your app. It also covers security—encrypting models for safe transfer and storage on Dell devices, with runtime decryption via a side channel. Plus, you can query model details or track metrics like latency and power use, with support for Dell-powered NVIDIA® Jetson™ and x86 setups.

The LEIP Deploy runtime engine provides lifecycle management capabilities for the optimized model.



Real-World Impact

Latent AI isn't just about optimization—it's about making a difference. Pairing LEIP with Dell Technologies hardware transforms industries with faster drones, smarter factories, and cost-effective enterprises. Here's how:

Military

Unmanned Aerial Drones (UAVs)

For missions like search and rescue, UAVs need AI that's precise and quick. On a Dell-supported edge device, LEIP's efficiency cuts power use, stretching flight time without dropping payload capacity—crucial when every ounce counts.

Underwater Unmanned Vehicles (UUVs)

The Navy used LEIP to speed up model updates by 18x—turning months into days. Optimized models on Dell hardware run 4x faster with 20% less power, extending missions and trimming costs.

Industry

For manufacturers chasing Industry 4.0, LEIP on Dell edge servers slashes RAM use by 73%, boosts inference by 73%, and cuts GPU needs by up to 92%. That means real-time quality control and predictive maintenance without breaking the bank—all on existing Dell hardware. Plus, the LEIP Deploy runtime integrates into industry-standard environments and supports NVIDIA's DeepStream and Triton inference engines. This means LEIP can optimize even the biggest, slowest models while coexisting with your legacy models. And LEIP Optimize can target both the CPU and GPU of various hardware devices, making system maintenance (hardware and software upgrades) painless.

Enterprise

LEIP scales AI affordably, trimming operating costs by 37%. Whether it's sports analytics or enterprise workflows, Dell hardware paired with LEIP delivers speed, savings, and scalability.

Wrapping up

By teaming up cutting-edge AI with practical systems like Dell Technologies hardware, Latent AI proves its worth—delivering real results across the board. From drones to factories to enterprises, it's all about getting the most out of your edge hardware with a little help from LEIP.