KAMIWAZA AMPERE™

Enterprise-Grade Generative AI on Energy-Efficient CPUs

kamiwaza.ai

2025

Enterprise-Grade Generative AI on Energy-Efficient CPUs

Enterprises seeking to deploy generative artificial intelligence (AI) have typically faced significant barriers: high costs, specialized hardware requirements, complex cooling infrastructure, and energy consumption concerns. The Kamiwaza-Ampere solution breaks this paradigm by enabling sophisticated AI workloads to run on energy-efficient CPU architecture without sacrificing performance.

This solution brief outlines how organizations can implement advanced document processing and multi-agent AI workflows using Kamiwaza's orchestration platform running on Ampere's cloud-native processors. The result is a practical approach to enterprise AI that delivers 4-5x power savings while maintaining security, performance, and control.

ESG impact:

Business efficiency:

AI democratization:

The challenge: GPU dependency in enterprise AI

Organizations implementing generative AI have traditionally faced four major obstacles in implementing systems with hardware accelerators like GPUs:

- **Infrastructure costs** — Specialized hardware can cost 3-5x more than standard server infrastructure
- **Power & cooling** — AI systems demand significant power and specialized cooling systems
- **Technical complexity** — May require specialized expertise not commonly found in enterprise IT teams
- **Data sovereignty** — Many organizations cannot move sensitive data to cloud environments but struggle to deploy GPU infrastructure on-premises

These challenges have created an artificial barrier to AI adoption, particularly for organizations in regulated industries or those with strict data governance requirements.

The solution: AI orchestration meets energy-efficient computing

The Kamiwaza-Ampere solution combines two breakthrough technologies. Together, these technologies enable enterprises to implement sophisticated AI capabilities without the infrastructure barriers associated with many hardware accelerator deployments.

Kamiwaza's multi-agent AI orchestration platform

- Intelligent distribution of AI workloads across available computing resources
- Specialized agents that collaborate to solve complex business problems
- Visual workflow interface providing transparency into AI processes
- Pre-built solutions for common enterprise use cases
- Seamless integration with existing enterprise data sources, APIs, and business systems
- Native connectors to enterprise databases, cloud platforms, and third-party applications

Ampere's cloud-native processors

- AmpereOne® M processors optimized for dense AI environments that deploy multi-modal and agentic, AI-based large language model (LLM) services
- Purpose-built for sustainable AI compute workloads
- Deployable in standard rack infrastructure without specialized cooling requirements
- The lowest cost per inference session available

Use case: Intelligent document processing

The flagship use case for the joint solution is intelligent document processing, which demonstrates the system's ability to handle complex AI workloads efficiently.

This use case is particularly valuable for financial services, healthcare, legal, and government organizations that process high volumes of documents while maintaining strict security and compliance requirements.

Workflow

1. Documents (including handwritten, historical, and complex forms) are submitted through email, scanner, or mobile capture.
2. The system's visual language model (7B parameters) processes the document to identify and extract relevant information.
3. Structured data is extracted and prepared for downstream systems.
4. The multi-agent interface validates, enriches, and routes information appropriately.
5. Results are returned to requestors and integrated with enterprise systems.

Performance metrics

- **Processing time** — Under two minutes, even for complex documents
- **Accuracy** — Comparable to human experts (95%+ for most document types)
- **Concurrent workflows** — Multiple documents processed simultaneously on a single processor
- **Power consumption** — 4-5x less than alternatives

Beyond document processing: Multi-agent workflows

While document processing provides a compelling entry point, the solution enables a variety of multi-agent AI workflows. Each workflow leverages the solution's ability to efficiently orchestrate multiple AI models and agents across Ampere's energy-efficient processors, while tying into customer's existing systems and databases for information retrieval.

Sales analysis

- Real-time processing of customer interactions and sales data
- Trend identification and opportunity flagging
- Predictive forecasting and pipeline analysis

Financial reporting

- Automated analysis of financial documents and data sources
- Automated response generation for common issues
- Support agent augmentation with real-time information retrieval

Customer support

- Intelligent routing and prioritization of support requests
- Automated response generation for common issues
- Support agent augmentation with real-time information retrieval

Technical specifications

Hardware requirements

Recommended configuration

- **Processor** — AmpereOne® M 192core CPU
- **Memory** — 512GB DDR5
- **Storage** — 1TB NVMe SSD
- **Network** — 10GbE (minimum)

Minimum configuration

- **Processor** — AmpereOne® M 96-core CPU
- **Memory** — 256GB DDR5
- **Storage** — 1TB NVMe SSD
- **Network** — 10GbE

Software stack

Kamiwaza components

- AI orchestration engine
- Multi-agent interface
- Workflow designer
- Model optimization framework
- Enterprise integration connectors

Supported models

- Visual language models (7B-14B parameters)
- Text generation models (7B-30B parameters)
- Specialized domain models (configurable)

Integration options

- REST API
- Message queue
- File-based integration
- Enterprise connectors (SAP, Salesforce, etc.)

Implementation pathway

Organizations can implement the solution through a structured approach:

1. **Assessment** — Identify high-value use cases and document technical requirements.
2. **Proof of concept** — Implement initial use case with existing infrastructure.
3. **Validation** — Measure performance, accuracy, and business impact.
4. **Expansion** — Scale to additional use cases and enterprise integration.

Typical implementation timelines range from 2-4 weeks for the initial proof of concept, with full enterprise deployment in 8-12 weeks.

Business benefits

The Kamiwaza-Ampere solution delivers quantifiable benefits across multiple dimensions.

Cost reduction

- 60-70% lower hardware costs compared to GPU alternatives
- 75-80% power savings
- Reduced cooling infrastructure requirements
- Simplified management and maintenance

Risk mitigation

- Enhanced data sovereignty and security
- Reduced dependency on specialized expertise
- Elimination of cloud data transfer risks
- Improved compliance posture

Business transformation

- Accelerated document processing (80%+ time savings)
- Improved data extraction accuracy (95%+ for most document types)
- Enhanced decision-making through AI-augmented analysis
- Freeing skilled resources from routine tasks

About the companies

Ampere

Ampere is a modern semiconductor company designing the future of cloud computing with the world's first Cloud Native Processors. Built for the sustainable Cloud with the highest performance and best performance per watt, Ampere processors accelerate the delivery of all cloud computing applications. Ampere Cloud Native Processors provide industry-leading cloud performance, power efficiency and scalability.

Kamiwaza

Kamiwaza is an advanced AI orchestration solution designed to be chip-agnostic, enabling organizations to deploy and manage AI workloads seamlessly across diverse hardware environments. Kamiwaza specializes in automating and optimizing AI-driven data analysis directly where data resides, ensuring high performance, security, and compliance regardless of the underlying compute infrastructure. Kamiwaza's multi-agent AI interface makes sophisticated analytics accessible to a wide range of industries, transforming legacy data into valuable business and operational intelligence.

Next steps

- Contact our team to schedule a personalized briefing or proof of concept
- Download additional resources by visiting our website

Disclaimers & important information

- **Performance and cost estimates** — All performance metrics, cost savings, and power efficiency claims presented in this document are based on Kamiwaza's internal testing, product specifications, and cost estimations using representative workloads. Actual results may vary significantly based on specific use cases, data types, system configurations, deployment environments, and operational requirements. Organizations should conduct their own testing and validation before making implementation decisions.

- **Technical requirements** — Hardware and software specifications are recommendations based on typical enterprise deployments. Actual requirements may vary depending on workload complexity, concurrent users, data volume, and performance expectations. Consult with technical specialists to determine optimal configurations for your specific environment.

- **Implementation timelines** — Stated implementation timelines are estimates based on typical enterprise environments. Actual timelines may be affected by organizational complexity, integration requirements, data migration needs, compliance requirements, and resource availability.

- **Regulatory and compliance** — While this solution is designed with enterprise security and compliance in mind, organizations are responsible for ensuring any AI implementation meets their specific regulatory, legal, and compliance obligations. This includes but is not limited to data privacy laws, industry regulations, and internal governance policies.

- **Model performance** — AI model accuracy rates are based on testing with specific document types and may not apply to all use cases. Model performance can be affected by document quality, language, format complexity, and domain-specific requirements. Continuous monitoring and optimization may be required.

- **No warranty** — This solution brief is provided for informational purposes only. Neither Kamiwaza nor Ampere makes any warranties, express or implied, regarding the performance, reliability, or suitability of the described solutions for any particular purpose.

- **Future availability** — Product features, specifications, and availability are subject to change without notice. Contact authorized representatives for the most current information.

For questions about this solution brief or to discuss your specific requirements, please contact our technical teams directly.