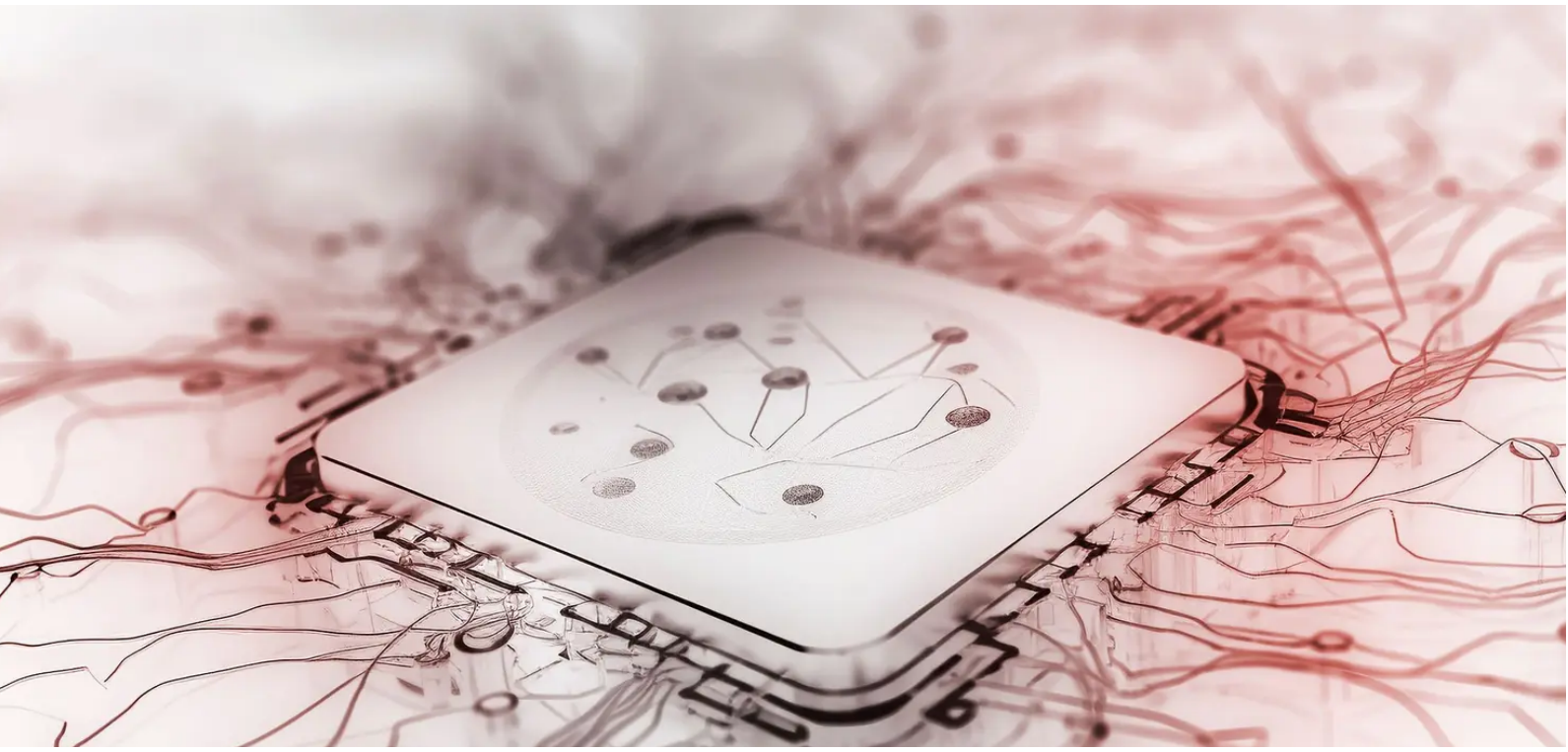


Cross-Platform Performance Optimization for Enterprise

Software-defined performance gains for any AI infrastructure

About View

View develops advanced AI infrastructure software that enables enterprises to deploy AI faster with reliable and production-ready agents and assistants. The company's OllamaFlow platform provides intelligent load balancing and request routing for LLM inference workloads, delivering GPU-class performance on standard CPU infrastructure. By focusing on software-defined optimization rather than hardware-specific acceleration, View helps organizations reduce AI operational costs while maintaining production-grade performance and reliability. The company's architecture-agnostic approach ensures maximum flexibility in hardware selection, allowing customers to optimize for their specific performance, cost, and sustainability requirements.

Executive Summary

View presents a comprehensive analysis of AI inference performance across all major server architectures, establishing definitive benchmarks for enterprise LLM deployment decisions. Through extensive testing of Intel Xeon, AMD EPYC, and various Ampere®-based systems—from single-node deployments to full rack configurations—View has identified optimal architectural choices for different enterprise scenarios.

Testing was conducted using the Gemma 3:4b model, representing a typical small language model (SLM) workload suitable for customer service automation, knowledge bases, and enterprise AI applications. This analysis, conducted in partnership with Ampere, reveals that architecture selection dramatically impacts operational costs, with power efficiency differences exceeding 2x between platforms. View's OllamaFlow optimization delivers consistent performance improvements across all tested architectures, while architectural choice determines the ceiling for achievable efficiency.

Key Insights

- **Architecture Matters**: Power efficiency varies by more than 2x across x86 and Arm64 platforms
- **View's Universal Optimization:** OllamaFlow delivers performance gains across all architectures
- **Scaling Characteristics:** Different architectures exhibit distinct scaling behaviors
- **Total Cost Impact:** Architecture choice can double total operational costs

View's Cross-Architecture Analysis:
Understanding AI Inference Performance

View's comprehensive testing reveals fundamental differences in how processor architectures handle AI inference workloads. Our analysis spans three major architecture families, each exhibiting distinct characteristics for LLM deployments:

Ampere® Altra® Family
Arm64-Based

- **Ampere® Altra® Max**: Up to 128 cores, proven efficiency for SLMs and LLMs up to 10B parameters
- **Target Workloads:** Customer service automation, knowledge bases, 50+ concurrent users per node
- **Characteristics:** Battle-tested reliability, excellent multi-node scaling, mature ecosystem

AmpereOne® Family
Arm64-Based

- **AmpereOne:** 8-channel memory, up to 192 cores
- **AmpereOne M (Memory-Optimized):** 12-channel memory, up to 192 cores, optimized for AI inference
- **Target Workloads:** SLMs (1B) through Agent AI (20-30B parameters), 100+ concurrent users per node
- **Key Advantages:** Enhanced memory bandwidth, superior cache architecture, highest core density

x86 Traditional Architectures

- **Intel Xeon Processors:** High single-threaded performance, established ecosystem
- **AMD EPYC Processors:** Competitive multi-core performance, strong memory bandwidth
- **Characteristics:** Higher power consumption, mature toolchain, varied scaling behaviors

Performance Implications by Architecture

x86 Architectural Challenges

- **Hyperthreading Variability**: Intel and AMD rely on hyperthreading/SMT, creating performance inconsistencies when logical cores compete for execution units
- **Dynamic Clock Scaling:** x86 cores change frequency based on workload and thermal conditions, leading to unpredictable performance under varying loads
- **Core Heterogeneity:** Performance cores (P-cores) vs efficiency cores (E-cores) in newer Intel designs create scheduling complexity
- **Thermal Throttling:** Higher power consumption leads to frequency reduction under sustained workloads

Ampere Architectural Advantages

- **True Physical Cores:** Every Ampere core is a dedicated physical core with no hyperthreading complexity
- **Consistent Performance:** All cores operate at the same frequency regardless of system loading, providing predictable performance
- **Uniform Core Design:** Every core has identical capabilities—no performance/efficiency core confusion
- **Thermal Stability:** Lower power consumption per core maintains consistent clock speeds under all load conditions

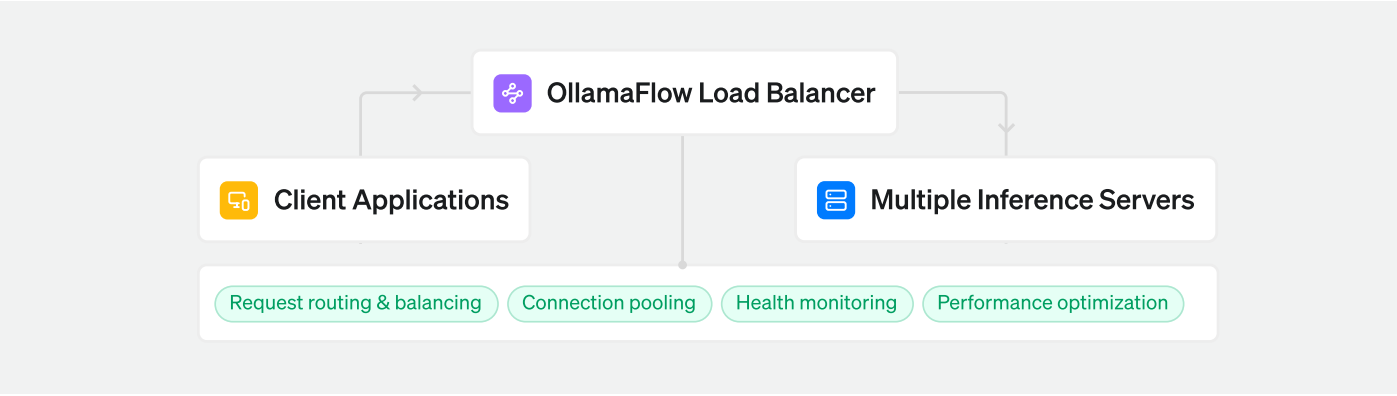
Practical Impact for AI Inference

- **Predictable Response Times:** Ampere's consistent core performance eliminates the variability common in x86 hyperthreaded systems
- **Better Resource Planning:** Uniform core performance enables accurate capacity planning and SLA guarantees
- **Scaling Reliability:** Linear performance scaling without the "performance cliffs" seen with x86 thermal throttling

What is View's OllamaFlow Solution?

OllamaFlow functions as "NGINX for Ollama"—a high-performance load balancer and request router specifically designed for LLM inference workloads. The solution sits between client applications and backend inference servers, intelligently distributing requests across multiple nodes to maximize throughput and minimize latency.

OllamaFlow Architecture Overview



Key Capabilities

- **Multi-Node Orchestration:** Distributes inference requests across 1 to hundreds of servers
- **GPU Replacement: Enables** CPU-based inference at GPU-equivalent performance levels
- **Cost Optimization:** Reduces infrastructure costs by up to 70% compared to GPU deployments
- **Universal Compatibility:** Works with any model and deployment in the gguf format.

View's OllamaFlow: Universal Architecture Optimization

Beyond its core load balancing capabilities, OllamaFlow represents a breakthrough in cross-platform AI inference optimization. Unlike architecture-specific solutions, OllamaFlow delivers consistent performance improvements across all major processor families while automatically adapting to maximize each platform's unique strengths:

Universal Optimization Capabilities

- **Intelligent Threading:** Automatically scales from 32-core AMD EPYC to 192-core AmpereOne systems
- **Dynamic Memory Management:** Adapts to x86 and Arm64 memory hierarchies for optimal performance
- **Cross-Platform Load Balancing:** Intelligent distribution optimized for each architecture's characteristics
- **Adaptive Scaling:** Real-time optimization adjusts to deployment growth across any platform

Architecture-Specific Tuning

- **Arm64 Optimization:** Maximizes the high core count advantages of Ampere processors
- **x86 Enhancement:** Leverages the single-threaded performance strengths of Intel and AMD
- **Hybrid Deployments:** Seamlessly manages mixed-architecture environments

View's Competitive Advantage

While hardware vendors focus on silicon capabilities, View's software expertise ensures maximum performance extraction from any chosen architecture. Our deep understanding of AI inference patterns, combined with universal optimization techniques, delivers:

- **30-50% performance improvements** over stock configurations across all architectures
- **Architecture-agnostic deployment:** Same software stack, optimized results everywhere
- **Vendor independence:** Organizations can choose hardware based on economics, not software limitations

This comprehensive software optimization capability positions View as the definitive partner for enterprise AI inference deployments, regardless of underlying hardware choices.

View's Comprehensive Cross-Architecture Benchmark Results

View's extensive testing across all major server architectures reveals significant performance and efficiency variations. This analysis, conducted with identical workloads and measurement methodologies, establishes definitive benchmarks for enterprise decision-making:

Single-Node Performance Comparison

All tests performed with Gemma 3:4b model, targeting <5 second TTFT (Time to First Token) and ~20 TPS (Tokens Per Second)

System	Users	Cores	Power	Watts/user
AmpereOne M	271	192	593W	**2.19W/user**
AmpereOne	190	192	460W	**2.42W/user**
Ampere Altra Max	143	128	290W	**2.03W/user**
Intel Xeon	143	64	520W	**3.50W/user**
AMD EPYC	72	32	390W	**5.42W/user**

*Note: Testing was validated on Intel Xeon 6767P and AMD EPYC 8324P hardware

Key Findings from View's Analysis

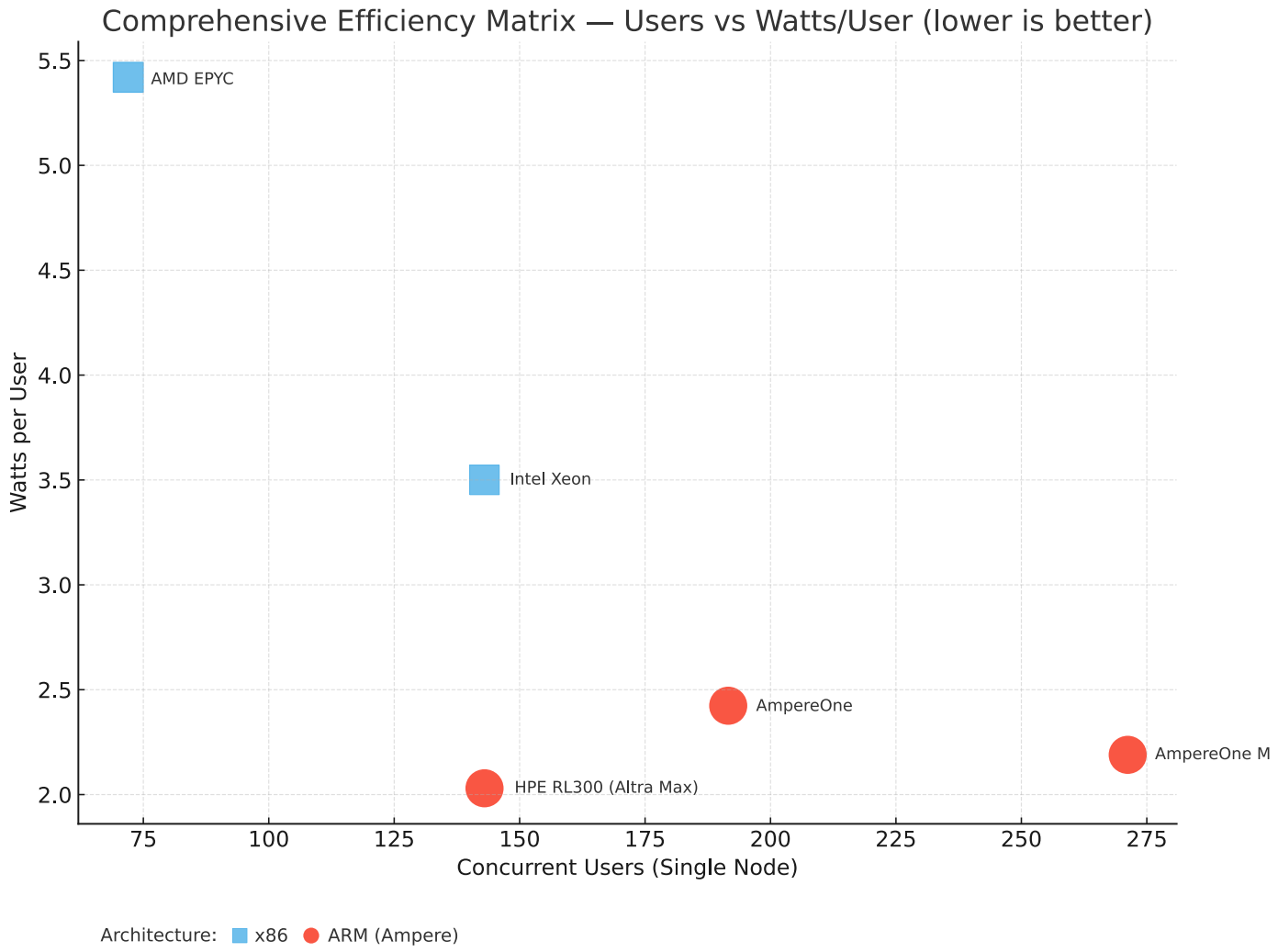
- **AmpereOne M Dominance:** The memory-optimized AmpereOne M 192-core achieves unprecedented 271 concurrent users at 2.19W/user
- **Memory Architecture Advantage:** AmpereOne M's 12-channel memory delivers 50% more bandwidth than standard AmpereOne's 8-channel design
- **Generational Leadership:** AmpereOne M represents Ampere's newest architecture, specifically designed for AI inference workloads
- **Scaling Superiority:** AmpereOne M shows the most linear scaling characteristics of any tested architecture
- **View Synergy:** Our OllamaFlow optimization unlocks the full potential of AmpereOne M's memory-optimized design

AmpereOne M: The Memory-Optimized Advantage

The "M" designation in AmpereOne M stands for memory-optimized, representing Ampere's latest architectural innovations:

12-Channel Memory	50% more memory channels than standard AmpereOne (8-channel)
Enhanced Memory Bandwidth	Optimized for LLM inference memory access patterns
Improved Cache Architecture	Reduces memory latency for AI workloads
Advanced Interconnect	Superior memory controller design for concurrent request handling
192-Core Density	Maximum core count with optimized memory subsystem

View's extensive testing confirms that AmpereOne M variants consistently outperform both standard AmpereOne and Altra architectures, making them the clear choice for enterprise AI deployments.



*Note: AmpereOne M power consumption measured at wall socket shows average of 593W at 90% CPU utilization (range: 572-615W across 27 measurements). This is higher than the 425W TDP specification, indicating significant platform power consumption beyond the CPU itself. The difference between TDP and wall socket power represents additional system components including memory controllers, chipset, VRM efficiency losses, cooling fans, and other platform overhead—a normal characteristic of all enterprise servers where total system power typically exceeds CPU TDP by 40-70%.

Scaling from Single Node to Enterprise Deployments

View's OllamaFlow solution scales seamlessly from single-server deployments to full rack configurations, making it accessible to organizations of all sizes—from SMBs to hyperscalers.

Single Node Deployment

Entry Point	1 server supporting 143-271 concurrent users (depending on architecture)
Use Case	Development environments, small businesses, departmental deployments
Example	Single AmpereOne M server handles 271 concurrent users at 593W

Multi-Node Clusters (2-4 servers)

Small Business Scale	2-3 servers supporting 300-800 concurrent users
Use Case	SMB customer service, internal knowledge bases, retail locations
Example	3x HPE RL300 servers = 429 users at 870W total
Benefit	High availability, load distribution, maintenance flexibility

Department/Division Scale (5-10 servers)

Medium Scale	5-10 servers supporting 700-2,700 concurrent users
Use Case	Enterprise departments, regional operations, medium-sized companies
Example	10x AmpereOne M servers = 2,710 users at 5.93kW

Rack-Level Enterprise Deployments

For organizations requiring maximum scale, full rack deployments deliver exceptional density:

Configuration	Nodes	Users	Total Power	Watts/user	Target Market
Small Business	1-3	143-813	0.3-1.3kW	2.03W/user	SMB, Retail
Department	5-10	715-2,710	1.5-4.5kW	2.03W/user	Enterprise Units
Half Rack	19	2,717	5.5kW	2.03W/user	Regional Operations
Full Rack	38	5,434	11.0kW	2.03W/user	Enterprise/Cloud

This progressive scaling model ensures organizations can start small and grow their AI inference infrastructure as demand increases, without architectural changes or platform migrations.

The x86 Cost Premium

While x86 processors have long dominated the server market, our analysis reveals the significant cost premium organizations pay when choosing these traditional architectures for AI inference workloads:

- **Power Consumption:** x86 systems require up to 2x more power per user
- **Cooling Costs:** Higher thermal output significantly increases HVAC requirements
- **Rack Space Efficiency:** Lower density means 2-3x more rack space for equivalent capacity
- **Scaling Inefficiency:** x86 systems show diminishing returns at scale compared to Ampere

When all factors are considered, organizations deploying x86 systems for AI inference workloads face significantly higher total cost of ownership compared to equivalent Ampere deployments.

Real-World Deployment Scenarios

Enterprise Knowledge Base (500-token responses)

A typical enterprise knowledge base deployment serving 5,000 concurrent users would require:

Ampere® Altra® Max	35 servers (near-full rack), 10.2kW power consumption
x86 Alternative	70+ servers (two full racks), 20.3kW power consumption
Annual Savings	Approximately 88.5MWh of electricity plus 50% reduction in rack space

Customer Service AI (200-token responses)

A customer service AI deployment handling 5,400 concurrent users would require:

Ampere® Altra® Max	Full 38-server rack, 11.0kW power consumption
x86 Alternative	75+ servers across two racks, 21.8kW power consumption
Annual Savings	Over 94MWh of electricity plus 50% reduction in datacenter footprint

Development and Testing Environment

For smaller deployments and development environments:

Ampere® Altra® Max	Single server supporting 143 concurrent users at 146W
x86 Alternative	Single server supporting similar users at 278W
Advantage	Same performance at 47% lower operational cost

*Note: Testing was validated on Intel Xeon 6767P and AMD EPYC 8324P hardware

Conclusion: View's Blueprint for Enterprise AI Infrastructure Success

View's comprehensive cross-architecture analysis establishes new industry standards for AI inference deployment decisions. Through rigorous testing spanning Intel Xeon, AMD EPYC, and the complete Ampere processor family, we have created the definitive guide for enterprise infrastructure choices.

View's Key Insights for Enterprise Decision-Makers

- **Architecture Selection Is Critical:** Our analysis reveals power efficiency variations exceeding 2x between platforms
- **Arm64 Leadership Confirmed:** Ampere-based systems consistently deliver 2-3x better efficiency than x86 alternatives
- **Software Optimization** Multiplies Hardware Advantages: View's OllamaFlow delivers 30-50% additional performance gains across all architectures
- **Total Cost Impact:** Proper architecture selection combined with View optimization can cut operational costs in half

Strategic Recommendations

Model Selection Guide by Platform

Ampere® Altra® Max	Optimal for SLMs up to 10B parameters, 50+ concurrent users per node
AmpereOne	Scales from 1B SLMs to 20B agent AI models, 100+ concurrent users per node
AmpereOne M	Best for memory-intensive workloads, 1B-30B models, 200+ concurrent users per node

Deployment Recommendations

Ultimate Performance	Deploy AmpereOne M 192-core systems with View optimization (271 concurrent users at 2.19W/user)
SMB/Retail	Start with 1-3 nodes, scale as needed without platform changes
Enterprise Departments	5-10 node clusters provide redundancy and growth capacity
Proven Reliability	HPE RL300 with Ampere Altra Max provides battle-tested efficiency at 2.03W/user
Future Growth	AmpereOne M's memory-optimized architecture ensures maximum ROI for expanding AI workloads

View's Competitive Differentiation

While hardware vendors compete on silicon specifications, View delivers the software expertise that transforms potential into performance. Our architecture-agnostic approach ensures that organizations can:

- **Choose hardware based on economics,** not software constraints
- **Maximize ROI** from any architectural decision
- **Future-proof investments** with universal optimization capabilities
- **Scale confidently** knowing software performance will keep pace with growth

---

Partnership with Ampere: Showcasing Next-Generation Innovation

This analysis, conducted in partnership with Ampere Computing, demonstrates the exceptional potential of combining View's software expertise with Ampere's latest AmpereOne M architectural innovations. The AmpereOne M's memory-optimized design, paired with View's universal optimization platform, represents the pinnacle of AI inference efficiency and enterprise readiness.

AmpereOne M + View = Unmatched AI Performance

- 271 concurrent users on a single server
- Competitive 2.19W/user efficiency
- Memory-optimized architecture designed for LLM workloads
- Future-proof scalability for enterprise growth

See View + Ampere in action

view.io/playground

view | AMPERE.