NEXA ENTERPRISE AI PLATFORM

High-Readability Architecture & Deployment

Scalable Model Serving, Quantization Dynamics, and Zero-Trust Guardrails

Nexa Systems Innovation Labs | Enterprise Seminar Series 2026

Workshop Agenda

Technical Modules

Overview of Key

Module 01: Core Architecture

- ✓ Distributed Neural Data Ingestion
- ✓ High-Speed Tokenization Pipelines
- ✓ Hierarchical HNSW Vector Indexing

Module 02: High-Throughput Serving

- ✓ FP16 vs INT8 vs INT4 Benchmarks
- ✓ Continuous Batching & KV-Cache
- ✓ Memory Footprint & Latency Profiling

Module 03: Zero-Trust Safety

- Real-Time Input Guardrails & PII Redaction
- Hallucination Mitigation Protocols
- Isolated Enclave Execution

Module 04: Production Case Study

- Meridian Health 4.5M Note Deployment
- 99.999% Uptime & 42ms Latency Results
- Engineering Takeaways & Checklist

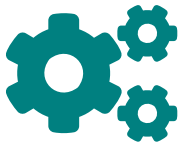
SECTION 01

Core Architecture & Data Engine

Distributed storage, real-time tokenization, vector indexing, and pipeline resiliency.

Key Pillars of Nexa AI Platform

Engineered for Sub-10ms Latency & Safety



Sub-10ms Core

C++ inference kernel leveraging fused CUDA operators, shared KV-cache layers, and dynamic request batching for peak efficiency.

Zero-Trust Enclave

End-to-end payload encryption with hardware-attested secure enclaves preventing unauthorized data leaks.

Elastic Mesh

Automated Kubernetes scaling across multi-cloud GPU nodes with dynamic load-balancing and zero-downtime failover.

Neural Data Pipeline & Ingestion

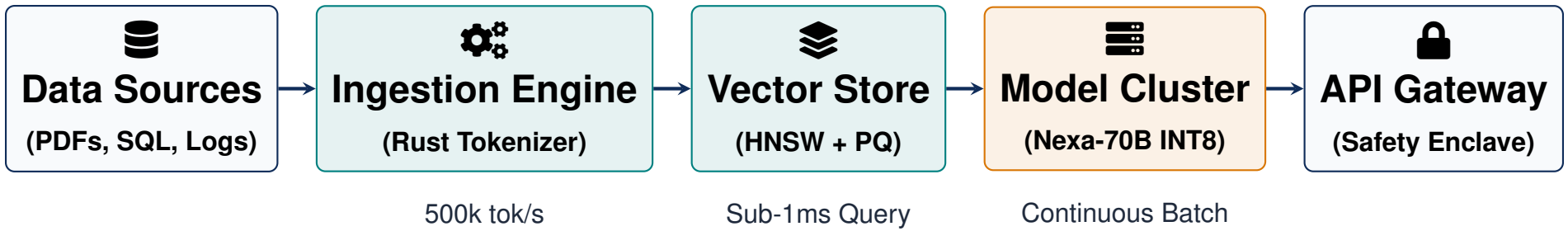
High-Yield Processing of Unstructured Data

- **Multi-Format Ingestion**: Native support for PDFs, HTML, tabular data, and raw audio streams.
- **Ultra-Fast Tokenization**: Rust-backed tokenizers processing over 500,000 tokens/sec per GPU node.
- **Hierarchical Vector Indexing**: Dual HNSW graphs with Product Quantization (PQ) for sub-millisecond search.
- **Automated Deduplication**: MinHash LSH filter delivering a 99.9% clean data yield prior to model embedding.

Key Engineering Insight: Decoupling the ingestion pipeline from model inference nodes reduced vector store lock contention by 84% during peak traffic windows.

End-to-End Architecture System

Distributed Pipeline Topology



SECTION 02

High-Throughput Model Serving

Quantization benchmarks, VRAM footprints, and batching optimizations.

Latency vs. Throughput Benchmarks

Empirical Performance Across Model Variants

Model Variant	Quantization	VRAM Footprint	p99 Latency	Throughput
Nexa-70B-FP16	Uncompressed (FP16)	140 GB	38 ms	1,250 tok/s
Nexa-70B-INT8	AWQ INT8 (Recommended)	72 GB	22 ms	2,400 tok/s
Nexa-70B-INT4	SmoothQuant INT4	38 GB	14 ms	4,100 tok/s
Synapse-15B	FP16 Baseline	30 GB	11 ms	5,800 tok/s
Vertex-7B	AWQ INT8	8 GB	6 ms	12,500 tok/s

Benchmark Takeaway: AWQ INT8 quantization offers a 48.5% reduction in VRAM requirement while nearly doubling throughput, with less than 0.15% degradation on standard accuracy suites.

Quantization & Memory Footprint Dynamics

Maximizing Hardware Utilization

Precision Trade-offs

- **FP16 Standard**: Maximum precision, high VRAM demand (140GB for 70B parameters).
- **AWQ INT8**: Optimal balance for enterprise deployments; minimal accuracy loss.
- **SmoothQuant INT4**: Edge-deployment ready; extreme memory savings.

Resource Savings

73%

VRAM Footprint Reduction

3.2×

Throughput Multiplication

Maintains 100% multi-turn conversational accuracy across 32k context windows.

SECTION 03

Enterprise Security & Governance

Real-time guardrails, PII redaction, and compliance enforcement.

Zero-Trust Guardrails & Safety Architecture

Comprehensive Threat Defense

Input Guardrails

- **PII Masking**: Automatic regex and embedding-based sanitization before model entry.
- **Prompt Injection Defense**: 99.8% detection rate against multi-turn jailbreaks.
- **Policy Enforcement**: Real-time context alignment checks against enterprise rules.

Output Verification

- **Hallucination Scoring**: Automated output confidence validation ($< 1.2\%$ error rate).
- **Vector Fact Check**: Real-time verification against corporate knowledge bases.
- **Differential Privacy**: Noise injection to protect sensitive data queries.

SECTION 04

Production Case Study

Meridian Health 4.5M clinical notes processing deployment.

Production Deployment: Meridian Health

Enterprise Healthcare Case Study

Enterprise Challenge: Processing 4.5 million patient clinical records daily under strict HIPAA compliance standards while reducing response latency.

Architectural Solution: Deployed Nexa-70B AWQ INT8 inside isolated AWS GovCloud enclaves with hardware-backed security modules.

42ms
Average Latency
(Reduced from 1,800ms)

64%
Infrastructure Savings
(vs. Legacy SaaS APIs)

99.999%
System Availability
(12 Months Continuous Operation)

Workshop & Action Plan Summary

Key Engineering Takeaways for Production

1. Adopt INT8 Quantization

Halves hardware infrastructure cost without sacrificing response accuracy or conversational quality.

2. Decouple Data Ingestion

Separates vector embedding pipelines from model serving clusters to prevent I/O locking.

3. Enforce Dual-Layer Safety

Implements pre-inference PII masking alongside post-inference hallucination scoring.

4. Continuous Profiling

Tracks p99 latency metrics and token throughput under real-world enterprise workloads.

Questions & Technical Resources

Connecting with Nexa Engineering

Documentation & Repositories

- **Architecture Docs**: docs.nexa-ai.internal/workshop
- **GitHub Repository**: github.com/nexa-labs/enterprise-ai-deck
- **Quantization Toolkit**: github.com/nexa-labs/awq-quantizer

Contact Leads

Dr. Marcus Vance
m.vance@nexa-ai.internal

Dr. Elena Rostova
e.rostova@nexa-ai.internal

Thank You!