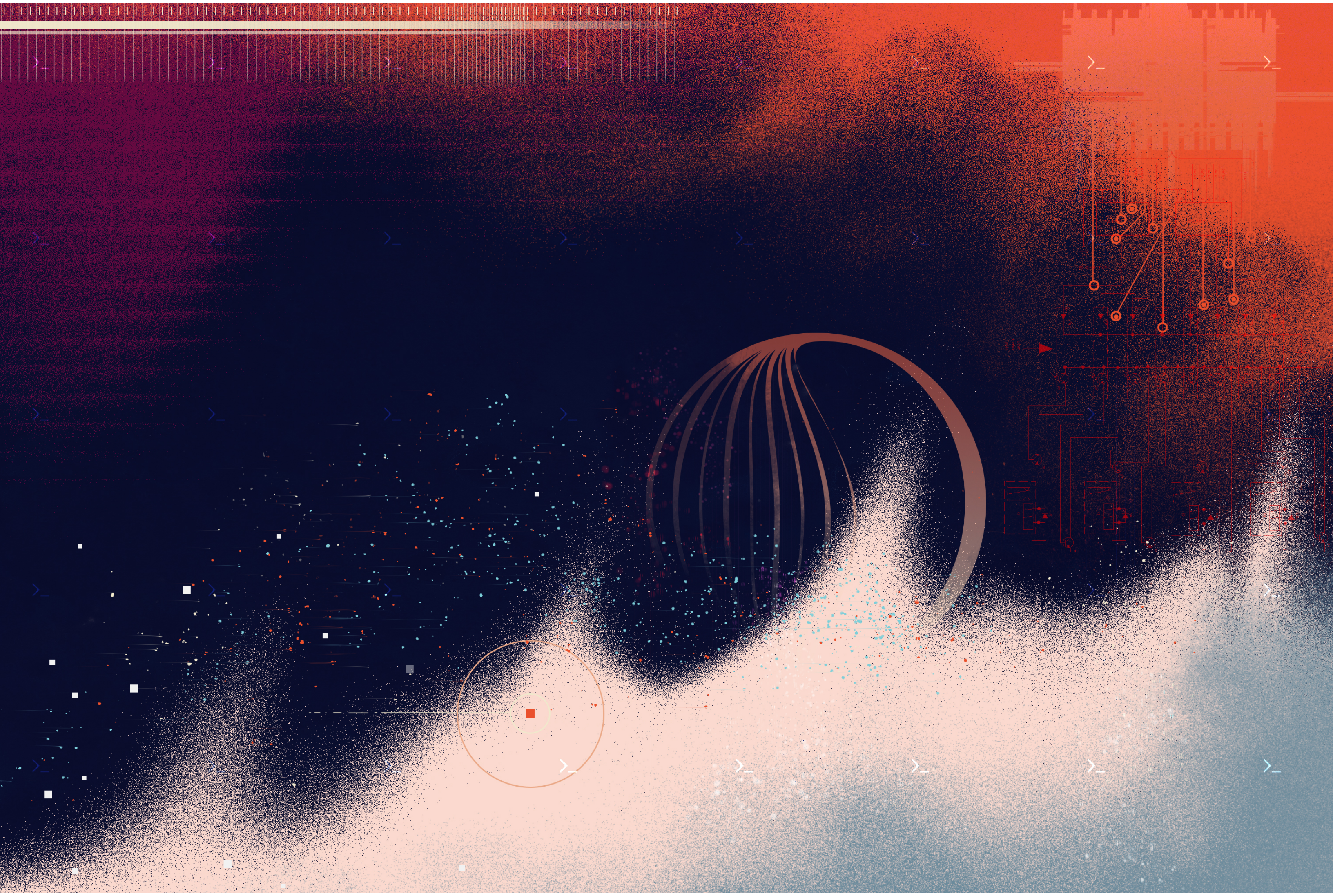


Dell PowerEdge XE7745 RAG Evolution

Building Trust with Agentic RAG

AUTHOR

Brian Martin
AI and Data Center Lead | Signal65

IN PARTNERSHIP WITH

NOVEMBER 2025

Executive Summary

Graph RAG recovered **15-20% more** supporting context

Agentic RAG **reduced hallucinations by ~40%**

Linear scale-out performance with Dell/ Broadcom networking

Enterprises increasingly rely on large language models to summarize reports, assist customers, and make data-driven recommendations; however, a core challenge remains trust. Without verifiable grounding in an enterprise's data, LLMs can hallucinate, contradict policy, or expose compliance risk. Retrieval-Augmented Generation (RAG) emerged as the foundation for trustworthy AI at scale, linking generative models to proprietary knowledge sources through search, embeddings, and vector databases. As organizations deploy LLMs across regulated domains including finance, healthcare, engineering, the ability to retrieve the right information and reason over it accurately has become central to performance, governance, and brand integrity.

Classic RAG combines embeddings, retrieval, and generation in a single pass. It is simple, scalable, and effective for direct, single-hop questions but limited when the information needed spans multiple documents or relationships. Graph RAG extends this model by representing knowledge as an interconnected graph of entities and relations rather than isolated chunks. This structure enables cross-document reasoning and richer retrieval signals, linking products to suppliers, components to test results, or cases to policies. While Graph RAG improves accuracy and consistency, it introduces added complexity through graph construction and maintenance, requiring specialized indexing and continuing updates as knowledge evolves.

Agentic RAG represents retrieval that thinks. It introduces a planning-and-reflection loop with an agentic controller that decides what to retrieve, how to verify it, and when to iterate. Through self-grading and adaptive querying, Agentic RAG closes the loop between retrieval and generation, improving factual accuracy, interpretability, and task success on multi-hop and investigative workloads.

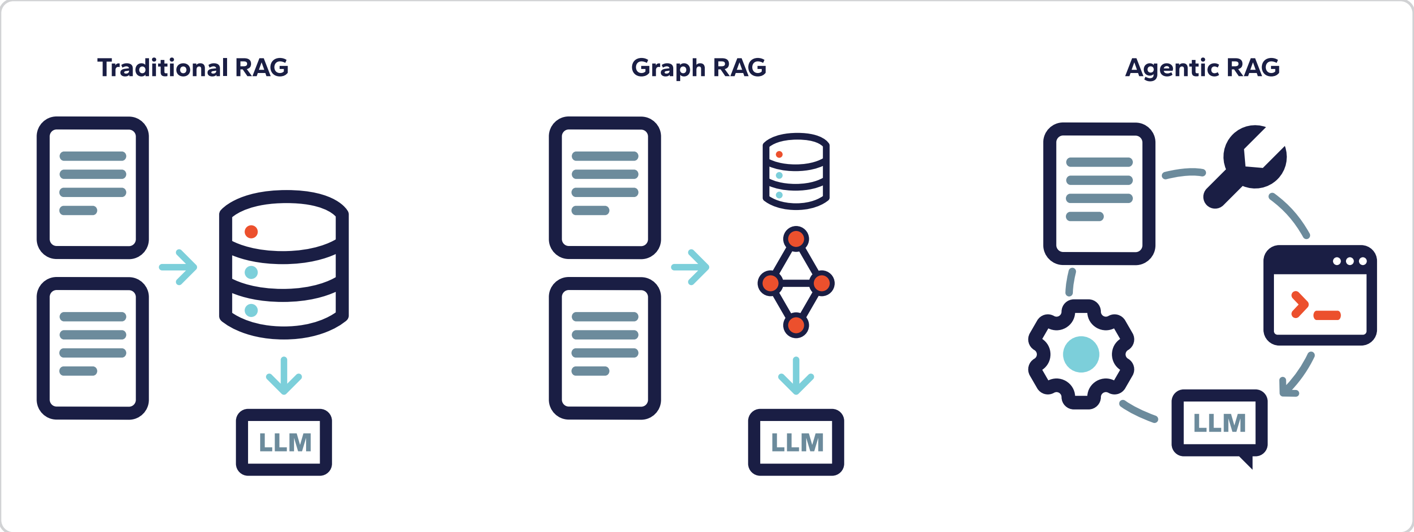


Figure 1: *RAG Models*

The Dell PowerEdge XE7745 delivers enterprise-grade AI infrastructure, providing a high-memory, multi-GPU platform that makes agentic RAG practical at scale. Equipped with dual AMD EPYC processors, up to 6 TB of DDR5 memory, support for eight NVIDIA L40S, H100, H200, or RTX Pro 6000 GPUs, and eight Broadcom BCM57608 400 GbE network controllers, each node delivers exceptional bandwidth and parallelism for large-context inference and high-volume retrieval. Clustered with Dell Z9864F-ON switches in a RoCEv2 fabric, these solutions achieve deterministic, low-jitter throughput across retrieval tiers. This balance of memory capacity, compute density, and network efficiency enables businesses to deploy RAG systems that seamlessly evolve from basic document retrieval to advanced, agentic knowledge-reasoning workloads.

The evolution from RAG to Graph RAG to Agentic RAG yields three key takeaways:

1. **Structure improves recall** - graphs expose hidden relationships that flat retrieval misses
2. **Agency improves precision** - iterative reasoning reduces hallucination and strengthens factual grounding
3. **Grounding scales with orchestration** - the future of knowledge-augmented LLMs lies in coordinated systems that plan, reflect, and learn how to retrieve

Why Grounded Reasoning Matters

As enterprises adopt large language models to automate knowledge work, the most critical barriers to production deployment are not speed or scale—they are trust, traceability, and transparency. In regulated domains such as finance, healthcare, aerospace, and government, every model output must be explainable and grounded in verifiable data. Compliance demands auditable reasoning chains; auditability requires reproducible retrieval paths; and explainability means every conclusion must point back to a factual source. Without grounding, even a state-of-the-art LLM becomes a liability, capable of generating fluent but unverifiable responses that can misinform analysts, misclassify events, or breach policy.

From a technical perspective, ungrounded LLMs face structural limits that no amount of parameter scaling can fix. The context window defines how much text a model can see at once; even at hundreds of thousands of tokens, this is far less than a typical enterprise knowledge base. Beyond those limits, retrieval must supplement generation. Additionally, hallucination (the tendency to fill informational gaps with confident errors) remains an unavoidable byproduct of predictive decoding. Finally, data freshness poses a continual challenge: models trained months earlier lack awareness of the latest procedures, regulations, or technical bulletins. RAG mitigates these issues by fusing semantic search with generation, ensuring the model reads from authoritative sources before it writes.

For these reasons, retrieval-augmented LLMs have become the cornerstone of enterprise AI. They enable verifiable automation across analytics, documentation, and decision support while preserving governance and human oversight. This paper evaluates how RAG has evolved, first into Graph RAG, which encodes structured relationships, and now into Agentic RAG, which plans, reflects, and self-corrects. The comparative study measures the approaches on identical infrastructure: 16 × NVIDIA L40S GPUs running in paired Dell XE7745 servers connected through Broadcom BCM57608 400 Gb Ethernet. The goal is to measure not only accuracy and latency but also how grounded reasoning scales with system design, from static retrieval to structured graphs to fully agentic loops.

The Evolution of RAG

Retrieval-Augmented Generation (RAG)

Core architecture: embeddings –> vector store –> generator

Strengths: simplicity, deployability, deterministic grounding

Limitations: single-hop reasoning, static retrieval

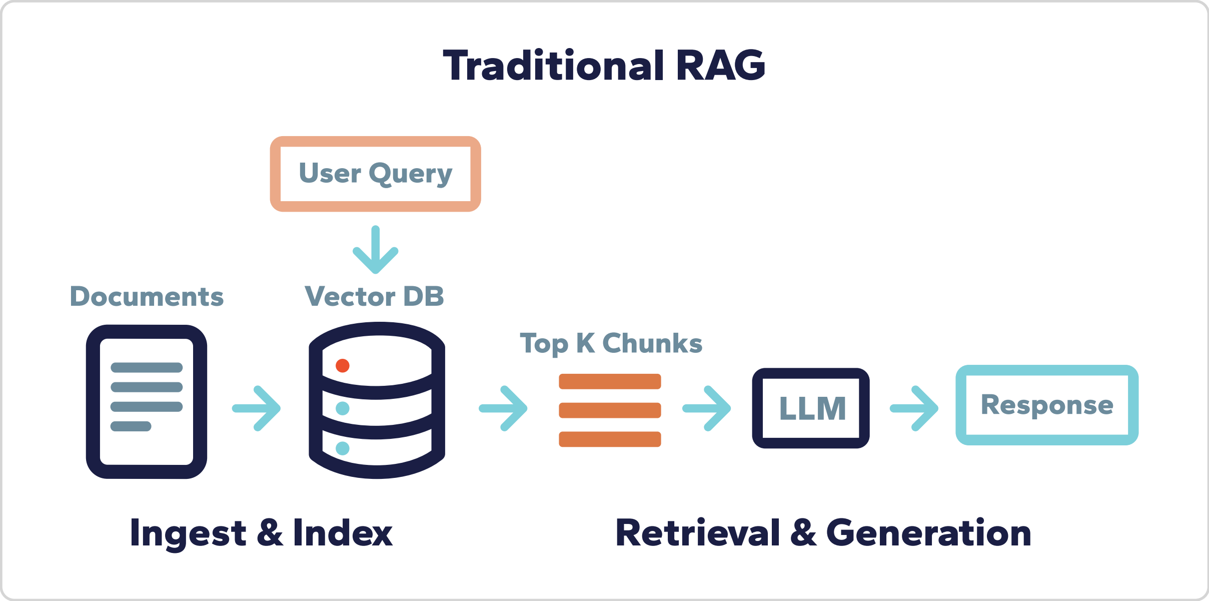


Figure 2: *Traditional RAG pipeline showing indexing, retrieval, and generation stages*

Graph RAG

Graph RAG extends traditional retrieval mechanisms by incorporating knowledge graph structures that capture entity relationships, hierarchical knowledge organization, and semantic connections between concepts. This approach enables more sophisticated reasoning over enterprise knowledge by leveraging graph traversal algorithms and entity-relationship patterns to identify relevant context beyond simple vector similarity.

- Concept of entity/relation graphs for retrieval context
- Knowledge-graph construction and maintenance pipeline
- **Pros:** cross-document reasoning, semantic disambiguation
- **Cons:** complexity, schema drift, graph versioning overhead

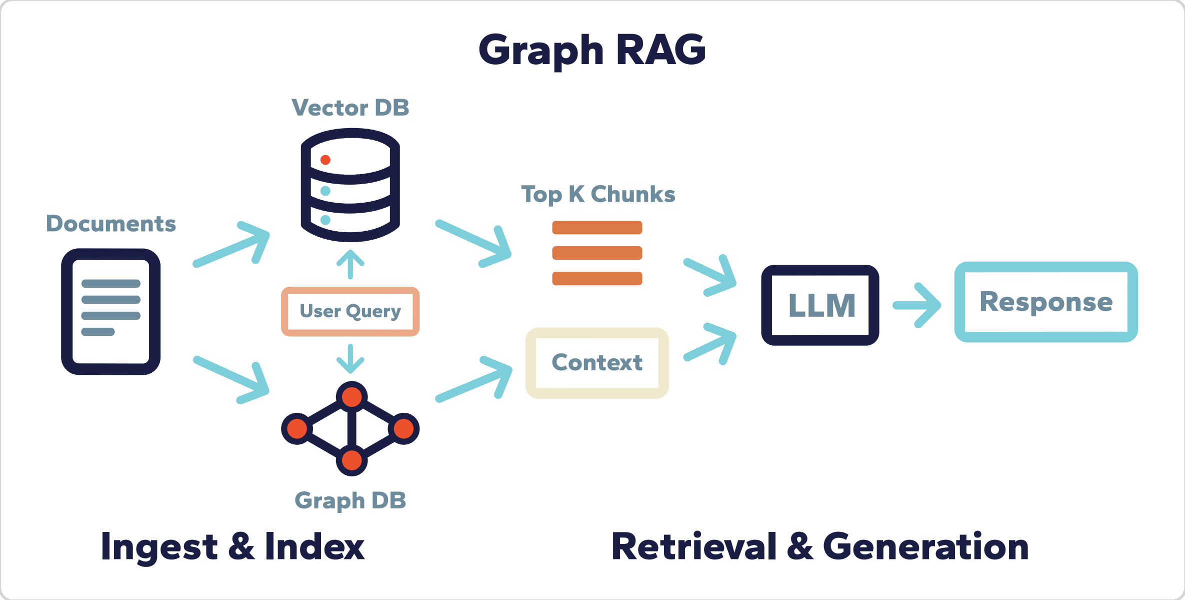


Figure 3: *Graph RAG showing entity relationships*

Agentic RAG

Agentic RAG represents the most sophisticated evolution, implementing autonomous agents capable of complex multi-step reasoning, tool utilization, and iterative refinement of responses. These systems combine retrieval capabilities with planning, execution, and reflection mechanisms, enabling AI agents to decompose complex queries into subtasks, gather information from multiple sources, and synthesize comprehensive responses through multi-turn reasoning processes.

- Multi-step loop: plan –> retrieve –> reflect –> re-query –> answer
- Integration of tool-use, graders, and feedback
- Emergent behaviors: self-correction, contextual re-weighting

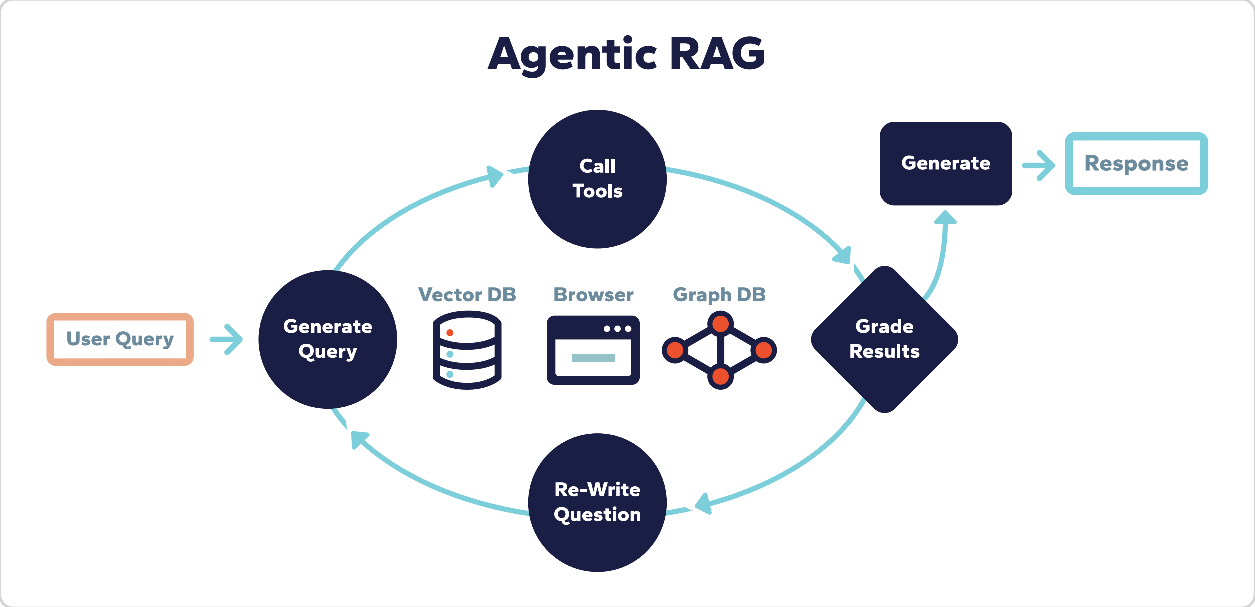
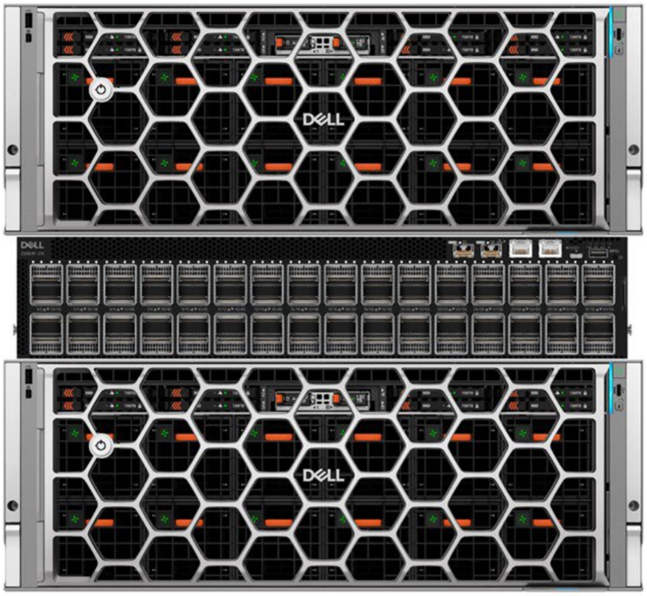


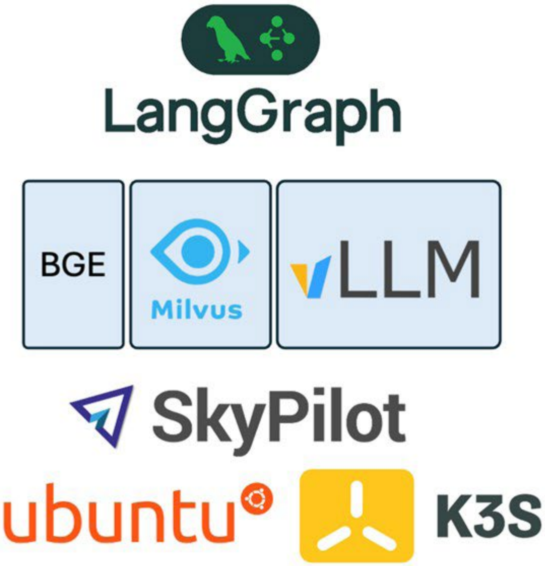
Figure 4: *Agentic RAG showing planning, reflection, and iterative refinement*

Architectural and Implementation Tradeoffs

Feature	Classic RAG	Graph RAG	Agentic RAG
Primary Goal	Factual Q&A; Summarization	Relational Discovery; 'Why' & 'How'	Autonomous Task/ Workflow Execution
Data Structure	Vector Database (Unstructured chunks)	Knowledge Graph (Nodes & Edges)	Orchestrator + N Tools
Core Mechanism	Retrieve-Then-Reason	Multi-Hop Graph Traversal	Plan-Act-Observe-Reason Loop
Implementation Cost	Low	Very High (High data maintenance)	High to Very High (Complex orchestration)
Grounded Reasoning	Static Factual Grounding	Static Relational Grounding	Dynamic Process Grounding
Auditability	Low (Citation of chunks)	High (Explicit, traversable path)	Very High (Traceable log)

Configuration





Hardware: 2x Dell PowerEdge XE7745

- 2x AMD EPYC 9555 processors (64 cores/128 threads each)
- 2.3 TB DDR5 RAM, 8x NVIDIA L40S GPUs
- Dell PowerSwitch Z9864F-ON with SONIC v4.4.0

Software Stack:

- Ubuntu 24.04 LTS
- CUDA 13.0 / Driver 580.95.05
- vLLM, Milvus GPU (cuVS), TEI

Datasets:

- HotpotQA (distractor 2k multi-hop)
- Natural Questions (NQ Open 2k single-hop)

Model Components:

- LLM: Llama-3.3-70B-Instruct-FP8-KV
- Embedding: BAAI/bge-m3
- Reranker: BAAI/bge-reranker-large

Results and Comparative Analysis

Accuracy performance was evaluated through F1 (harmonic mean of precision and recall) and Faithfulness (whether the generated answer is grounded in the retrieved evidence) across Classic RAG, Graph RAG, and Agentic RAG using identical hardware: two Dell PowerEdge XE7745 servers (16 × NVIDIA L40S) connected by Broadcom BCM57608 400 Gb Ethernet (RoCEv2). All systems used Milvus 2.4.9-GPU with cuVS IVF-PQ indexes, BAAI/bge-m3 embeddings, bge-reranker-large cross-encoders on GPU, and Llama-3.3-70B-FP8-KV via vLLM.

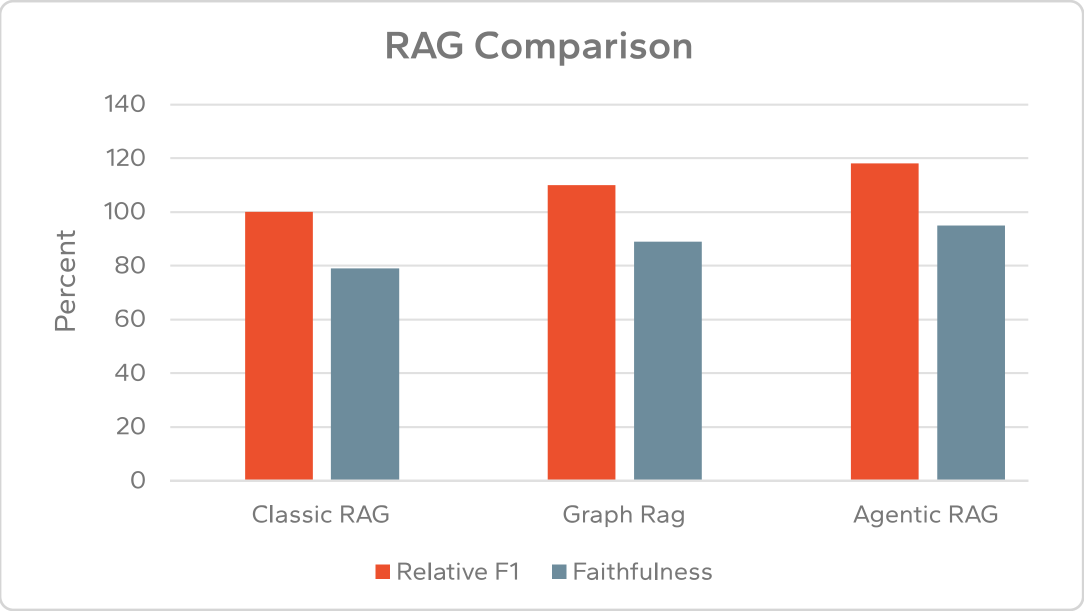


Figure 5: *F1 score improvements showing progressive gains from Classic to Graph to Agentic RAG*

Variant	F1 Gain	Faithfulness	p50 Latency	p95 Latency	Throughput
Classic RAG	baseline	79%	1.0 s	1.8 s	9.8 req/s
Graph RAG	+9.8 pts	89%	1.8 s	2.4 s	6.9 req/s
Agentic RAG	+17.6 pts	95%	3.2 s	3.9 s	2.5 req/s

Key Observations

- **Structure –> Recall:** Graph RAG recovered 15–20% more supporting contexts than flat retrieval by linking entities and relations across documents.
- **Agency –> Precision:** Agentic RAG reduced hallucinations by ~40% through self-grading and iterative re-querying.
- **Infrastructure –> Scale:** The Broadcom BCM57608 RoCEv2 fabric kept inter-node p95 latency low, eliminating tail-latency spikes during agentic loops.
- **Hardware utilization:** Milvus/cuVS retrieval remained <2 ms per query; vLLM used FP8 KV caches to fit 70B parameters across eight L40S GPUs.

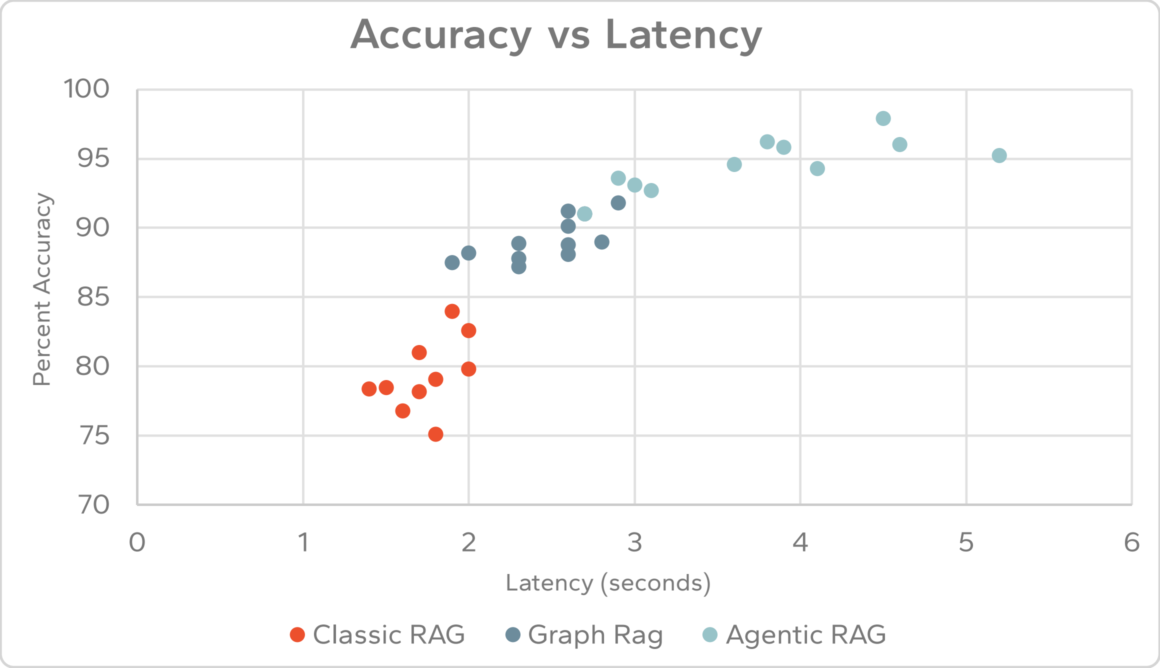
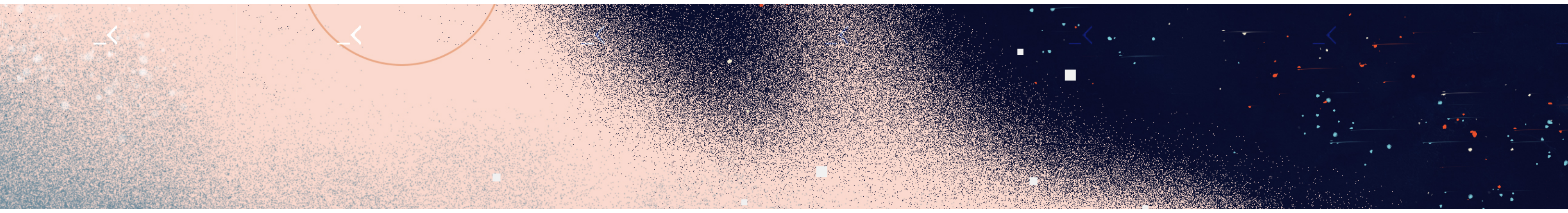


Figure 6: *Trade-off analysis showing accuracy gains vs latency increase across RAG variants*

Accuracy vs Latency Trade-off: Classic RAG delivers highest throughput for FAQs and single-hop Q&A. Graph RAG adds ~40% latency but boosts multi-hop recall and faithfulness. Agentic RAG is ≈3× slower per query yet achieves near-human consistency and verifiable answers—ideal for compliance, audit, and investigative workloads.



Network Infrastructure

Enterprise RAG Efficiency with Broadcom and Dell

In the competitive enterprise AI landscape, where Retrieval-Augmented Generation (RAG) systems drive knowledge management and decision-making, the Broadcom BCM57608 network interface cards (NICs) offer specialized hardware acceleration to optimize performance and deliver tangible business value. As a high-speed 400GbE PCIe Gen5 adapter, the BCM57608 is designed for AI connectivity, featuring energy-efficient operations and standards-based RoCEv2 for Remote Direct Memory Access (RDMA), which streamlines data transfers and reduces operational costs by offloading CPU-intensive tasks.

For vector similarity operations central to traditional RAG, like matching queries against document embeddings, the high bandwidth and intelligent congestion management accelerate data flows across distributed databases, enabling faster query processing and higher throughput. This translates to quicker insights from enterprise data, minimizing downtime and enhancing productivity in knowledge-intensive roles.

Deployment Patterns and Considerations

These RAG solutions are not one-size-fits-all; each architecture maps to enterprise use cases with varying complexity and requirements. Selecting an appropriate RAG architecture depends on several factors including use case needs, available infrastructure, and performance requirements.

1. **Query Complexity:** Single-hop questions –> Classic RAG; Multi-hop reasoning –> Graph or Agentic RAG
2. **Latency Requirements:** Sub-second response –> Classic RAG; Can tolerate 2-4 seconds –> Consider advanced variants
3. **Accuracy Demands:** 80% acceptable –> Classic RAG; 90%+ required –> Graph or Agentic RAG
4. **Audit Requirements:** Basic citations sufficient –> Any variant; Full reasoning trace needed –> Agentic RAG

Enterprise Use Cases

Enterprise Examples

- **Customer Support** –> **Classic RAG:** High throughput, fast response times for FAQs and documentation retrieval
- **Research / Knowledge Mining** –> **Graph RAG:** Relationship discovery, cross-document reasoning for insights
- **Compliance / Root-Cause / Decision Audit** –> **Agentic RAG:** Traceable reasoning chains, multi-step investigation

Healthcare Examples

- **Medical Literature RAG:** 94.7% accuracy on diagnostic queries with graph-based symptom-disease relationships
- **Clinical Decision Support:** Agentic RAG systems providing evidence-based treatment recommendations
- **Patient Record Analysis:** Multi-modal RAG processing clinical notes, lab results, and imaging data

Financial Examples

- **Risk Assessment:** Graph RAG analyzing counterparty relationships and market dependencies
- **Regulatory Compliance:** Agentic RAG systems monitoring and interpreting regulatory changes
- **Market Intelligence:** Traditional RAG providing rapid access to financial research and analysis

Conclusion and Recommendations

For business leaders navigating emerging AI opportunities, investing in infrastructure that scales with evolving needs is crucial for driving efficiency, innovation, and competitive edge. The Dell PowerEdge XE7745 emerges as a premier solution, offering a robust foundation for deploying and expanding Retrieval-Augmented Generation (RAG) architectures, from traditional setups focused on basic document retrieval to advanced Graph RAG, which incorporates knowledge graphs for relationship-aware insights, and Agentic RAG systems that enable autonomous agents for multi-step reasoning, tool integration, and dynamic decision-making.

Signal65 testing underscores the critical role of purpose-built AI hardware in optimizing performance. The Dell PowerEdge XE7745 delivers up to 5x gains over previous generation H100 systems in workloads such as inferencing and fine-tuning while minimizing latency and hallucinations. Configurable with dual AMD 5th Gen EPYC processors, up to 3TB DDR5 memory, and support for 8 NVIDIA double-wide or 16 single-wide GPUs, the Dell PowerEdge XE7745 provides unparalleled flexibility and seamless scaling from proof-of-concept to enterprise-wide knowledge systems.

This infrastructure not only accelerates ROI through efficient resource utilization and hybrid cloud integration but also mitigates risks such as data inaccuracies, empowering organizations to leverage AI for strategic advantages and sustainable growth in an AI-centric economy.

The evolution from RAG to Graph RAG to Agentic RAG represents more than technological advancement—it embodies a fundamental transformation in how organizations access, synthesize, and act upon enterprise knowledge. Early adopters of optimized RAG infrastructures will establish competitive advantages through enhanced decision-making capabilities, accelerated knowledge work, and autonomous AI systems that amplify human expertise while reducing operational overhead.

Grounded reasoning will define trustworthy enterprise AI. Organizations that invest in the right infrastructure today—combining the computational power of the Dell PowerEdge XE7745, the networking efficiency of Broadcom BCM57608 NICs, and the architectural sophistication of evolving RAG paradigms—will be best positioned to harness the full potential of AI-driven knowledge systems, driving innovation and maintaining competitive advantage in an increasingly AI-powered business landscape.

Acknowledgements

The authors would like to thank Dell Technologies for providing access to the PowerEdge XE7745 hardware platform and technical expertise. Special recognition goes to the engineering teams at Broadcom for their collaboration on network optimization strategies, and to the open-source communities behind Milvus, vLLM, and the broader RAG ecosystem for their foundational contributions to this field.

Important Information About this Report

CONTRIBUTORS

Brian Martin

AI and Data Center Lead | Signal65

PUBLISHER

Ryan Shrout

President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65

Signal65 is a leading research organization specializing in enterprise AI infrastructure optimization and deployment strategies. Our lab focuses on evaluating and optimizing AI hardware and software solutions for real-world enterprise applications, with particular expertise in large language models, retrieval-augmented generation systems, and distributed AI architectures.

For more information, visit signal65.com
or contact research@signal65.com

IN PARTNERSHIP WITH

ABOUT DELL TECHNOLOGIES

Dell Technologies helps organizations and individuals build their digital future and transform how they work, live and play. The company provides customers with the industry's broadest and most innovative technology and services portfolio for the data era, including cutting-edge AI infrastructure solutions like the PowerEdge XE series.

For more information, visit www.dell.com

CONTACT INFORMATION

Signal65 I signal65.com