Hybrid AI Architecture in Capital Markets

Executive Reading Path:

For **CIO, CTO, COO**: Chapters 1–3 and Chapter 9 provide the full strategic view.

For **Architecture, Risk, and Platform Teams**: Chapters 4–8 provide architectural depth and regulatory context.

Written by

firemind.

In partnership with

NVIDIA

aws

Executive summary

Capital markets are undergoing a decisive architectural shift. Three structural forces surging computational demand, escalating regulatory scrutiny, and rapid advances in AI and GenAI, are reshaping how trading, risk, surveillance, compliance, and research operate. These functions now depend on models that require ultra-low latency, elastic compute, and stringent control over sensitive data. Approximately 70–80% of AI initiatives fail to reach production, not because the models underperform, but because the underlying infrastructure cannot meet these combined demands.

Capital-markets AI now operates under four non-negotiable constraints: **deterministic latency for execution, strict data-residency requirements, elastic access to GPU compute**, and **unified governance** across environments.

Traditional cloud-only and on-premises-only architectures were not designed to satisfy these constraints simultaneously.

Firms that delay risk widening performance gaps, increasing operational costs, and heightened regulatory pressure as AI becomes embedded in core processes. This white paper examines the architectural implications of these trends and outlines a practical roadmap for building a unified, compliant, and scalable AI operating model in capital markets.

CHAPTER 1

Introduction & Market Context

Capital markets are undergoing their fastest technology transition in over a decade. AI and GenAI have moved from research labs into core business workflows across the front, middle, and back office. The scale, speed, and sophistication of these workloads are reshaping infrastructure requirements in ways that traditional environments can no longer accommodate.

Industry Forces Reshaping Capital Markets

1. Explosive Data Growth

- Trading desks process millions of events per second across venues and asset classes.
- Surveillance systems analyse billions of communications, transactions, and behavioural signals annually.
- Research teams rely on tens of millions of filings, transcripts, reports, and proprietary documents.

2. Increasing Model Sophistication

- LLMs now support research, reporting, and compliance analysis.
- Temporal and sequence models underpin alpha generation and liquidity forecasting.
- GenAI enriches scenario simulation with narrative and behavioural dynamics.
- Agentic workflows automate processes across operations, onboarding, and trade lifecycle management.

3. Rising Performance & Regulatory Expectations

- Execution inference must run within microsecond latency budgets.
- Risk engines must scale 10–100× during periods of market volatility.
- Regulatory frameworks (DORA, ESMA, FCA SYSC) mandate strong controls for traceability, resilience, and data residency.
- Innovation cycles operate on monthly–quarterly cadence, outpacing traditional infrastructure refresh cycles.

Why This Matters

Collectively, these forces are creating unprecedented pressure on existing technology stacks. The combination of growing data volumes, more complex model architectures, and tighter regulatory expectations is **widening the gap between what current infrastructure can support and what modern AI-enabled workflows require**. As a result, institutions are reassessing their architectural foundations to ensure they can scale AI safely, efficiently, and with the level of performance the markets now demand.

CHAPTER 2

Problem Landscape: Structural Barriers to AI at Scale

While external forces are accelerating the adoption of AI across the industry, most institutions face internal architectural barriers that prevent them from operationalising these capabilities at scale. **These barriers fall into four categories.**

1. Limits of Cloud-Only Approaches

Cloud accelerates experimentation but is poorly suited for latency-critical or regulated production workloads:

- Latency volatility: even millisecond-level jitter can undermine execution strategies.
- Residency constraints: regulated datasets cannot be transferred to public cloud environments.
- High GPU cost at peak: elastic provisioning becomes prohibitively expensive for large simulation bursts.
- Governance drift: rapid cloud innovation outpaces internal control frameworks.

2. Limits of On-Premises-Only Approaches

On-prem environments offer control but limit scalability and innovation:

- Slow and costly GPU scaling for risk and GenAI workloads
- Multi-year refresh cycles that lag model evolution
- Fragmented regional environments with inconsistent tooling
- Insufficient elasticity for VaR, stress tests, and GenAI demand spikes

3. Operational Fragmentation

Most institutions operate cloud, on-prem, and colocation environments as separate estates:

- Inconsistent CI/CD pipelines
- Duplicated security tooling
- Configuration drift
- Prolonged compliance approval cycles
- Increased audit and operational risk

4. Economic Inefficiency

Fragmented infrastructure drives structural cost inefficiencies:

- Underutilised GPU capacity
- Redundant operational overhead
- Multiple governance pathways
- Poor predictability of infrastructure spend

CHAPTER 3

Analysis & Insights: Why Hybrid Is the Only Viable Architecture

1. Performance Locality (Latency)

Execution-time inference and market-data-driven signal processing depend on microsecond–millisecond determinism. Even minor jitter undermines strategy profitability, and physical proximity to exchange gateways is essential.
Colocation provides the deterministic performance envelope necessary for these workloads. Cloud networks, regardless of peering or optimisation, cannot match this behaviour establishing hybrid as a necessity, not an optimisation.

2. Data Residency & Sovereignty

Surveillance logs, trade records, client data, KYC files, and communications archives are subject to strict residency and data-handling mandates. These datasets often must remain within specific jurisdictions or secure on-premises facilities. Hybrid allows institutions to maintain regulated datasets where they must reside, while still accessing cloud-based compute for complementary tasks. A single-environment strategy cannot meet these requirements without compromising compliance or performance.

3. Elasticity & Innovation

Risk, scenario modelling, Monte Carlo simulation, XVA, embeddings, and LLM inference demand elasticity far beyond what on-premises GPU estates can provide economically. New model classes arrive quarterly, and experimentation requires access to the latest architectures and frameworks. Cloud GPU fleets provide the necessary burst capacity and innovation velocity. Attempting to maintain parity on-premises results in prohibitive capex and slow iteration cycles.

4. Unified Governance

Regulators expect consistent controls across all environments, including identity and access management, lineage, auditability, policy enforcement, and operational resilience.
Fragmented infrastructures create inconsistency, increase audit findings, and slow deployment cycles. Hybrid architectures, when built on a unified control plane, allow firms to apply a single governance framework across colocation, on-prem, and cloud environments.

CHAPTER 4

Hybrid Workload Placement Matrix (PRE Framework)

These three structural requirements: Performance, Residency, and Elasticity, create a practical decision framework for workload placement. **Each class of workload can be evaluated objectively**:

Workload	Performance	Residency	Elasticity	Optimal Placement
Execution inference	High	Low	Low	Colo
Surveillance / AML	Medium	High	Medium	On-Prem
Monte Carlo / XVA	Low	Low	High	Cloud
RAG / Research	Medium	Medium	Medium	Hybrid
Stress Testing	Low	Medium	High	Cloud
Fraud Scoring	High	Medium	Medium	Hybrid
Agentic Workflows	Medium	Medium	High	Hybrid

Interpreting the Matrix

This placement matrix illustrates why hybrid is not simply “one option among many,” but the logical outcome of workload heterogeneity.

- Some workloads must run close to the exchange.
- Others must remain on-premises for regulatory reasons.
- Others must scale elastically in the cloud.

No single environment can support all three conditions simultaneously. The placement matrix illustrates how different workload classes align with different execution domains.

Regulatory Mapping

Regulatory Requirement	Regulatory Requirement
Data residency	On-prem placement of sensitive datasets
Auditability	Unified, cross-environment logging and lineage
Operational resilience (DORA)	Multi-region, hybrid failover
Access control (FCA SYSC)	Consistent RBAC across all environments
Model governance	Centralised model and policy management
Concentration risk	Workload distribution across cloud, on-prem, and colo

Summary: Why Hybrid is the Inevitable Outcome

Taken together, these insights demonstrate why hybrid architecture has become a structural requirement for AI-enabled capital markets. It allows firms to **place workloads exactly where they perform best**, maintain full regulatory compliance, scale elastically during periods of volatility, and adopt emerging AI techniques without overhauling their infrastructure. In short, hybrid provides the **flexibility, control,** and **resilience** necessary to operationalise AI at scale and stands as the only architecture capable of meeting the industry’s combined performance, sovereignty, and governance demands. These architectural insights translate **directly into business impact**; the following use cases illustrate where hybrid architectures deliver the greatest value across the capital-markets lifecycle.

CHAPTER 5

High-Value Use Cases Enabled by Hybrid Architecture

Hybrid architecture enables capital markets institutions to run each workload where it performs best: near the exchange for **deterministic inference**, on-premises for **regulated datasets**, and in the cloud for **GPU-intensive analytics**. This architectural flexibility unlocks material performance, productivity, and compliance improvements across the value chain. **Five domains illustrate the highest-value applications.**

1. Trading & Alpha Generation

Trading teams require both ultra-low-latency inference for execution and scalable compute for research and scenario modelling. Hybrid architecture supports this dual demand by separating real-time and analytical workloads.

Hybrid pattern:

- Colocation: inference engines and market-data-driven signal models operate with microsecond determinism.
- Cloud: backtesting, experimentation, embeddings, and GenAI-enhanced scenario modelling run elastically on large GPU fleets.

Impact: improved execution quality, faster model iteration cycles, and more sophisticated signal construction.

2. Risk, Surveillance & Liquidity Management

Risk and surveillance functions juggle regulatory expectations, real-time decisioning, and large-scale historical analysis. Hybrid architecture enables firms to combine strict data-residency controls with cloud-scale elasticity.

Hybrid pattern:

- Market surveillance: on-prem for real-time detection; cloud for historical replay and model retraining.
- Liquidity forecasting: local predictive models combined with cloud-based scenario bursts.
- Risk engines: 10–100× scaling for VaR, XVA, stress testing, and scenario generation during volatility spikes.

Impact: faster intraday risk insight, higher model accuracy, and improved regulatory responsiveness.

3. Financial Crime (AML, Fraud, Sanctions)

Financial crime teams require sub-50-millisecond scoring, jurisdictionally compliant data handling, and cross-institution pattern recognition. Hybrid architecture enables each of these capabilities to operate within its optimal environment.

Hybrid pattern:

- Real-time fraud scoring: on-prem execution to meet <50 ms SLA requirements.
- Federated learning & pattern detection: cloud-based training across distributed datasets without violating residency constraints.

Impact: reduced false positives, improved typology detection, and faster case investigation.

4. Research & Compliance

Research analysts and compliance teams benefit from GenAI-driven summarisation, retrieval, and multilingual reasoning but much of their source content is sensitive and must remain on-premises.

Hybrid pattern:

- RAG workflows: on-prem vector stores for sensitive
- documents, combined with cloud-based LLM inference for large-context reasoning.
- Compliance automation: policy interpretation, regulatory mapping, and multilingual reporting executed with hybrid LLM services governed under unified controls.

Impact: higher analyst productivity, faster compliance cycles, and more consistent regulatory interpretation

5. Agentic Operations

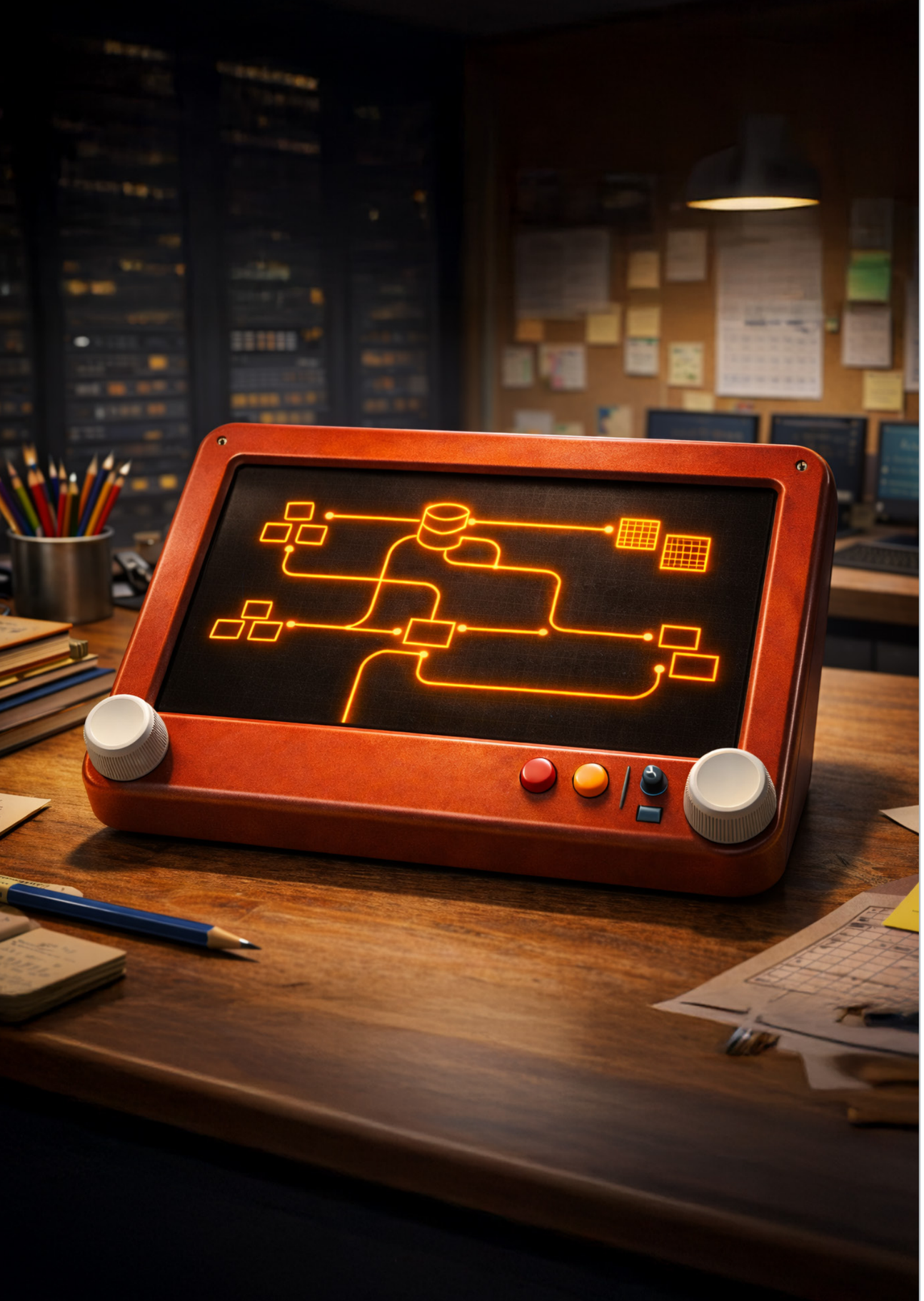
Operational teams increasingly employ AI agents to orchestrate complex workflows across onboarding, trading operations, reporting, and regulatory submissions. These agents must interact safely with systems across multiple environments.

Hybrid pattern:

- Operational automation: agents coordinate processes spanning on-prem transactional systems and cloud-based reasoning services.
- Supervised execution: strong guardrails ensure agents operate within approved boundaries and maintain auditability.

Impact: reduced manual workload, more reliable end-to-end processes, and improved operational resilience.

Together, these use cases illustrate a consistent pattern: no single environment can meet the performance, residency, elasticity, and governance needs of modern AI workloads. **Hybrid architectures enable firms to allocate each workload to its optimal domain, allowing measurable improvements in speed, cost, accuracy, and compliance**. These patterns form the basis for the reference architecture described in the following chapter.



CHAPTER 6

Reference Architecture

Capital markets require an architectural foundation that can simultaneously support ultra-low-latency execution, regulated data residency, elastic GPU scaling, and unified governance. A hybrid architecture achieves this by integrating three distinct domains: **colocation, sovereign on-premises environments**, and **cloud regions**, under a single operational and control framework. Each domain plays a specific role, and the combination allows firms to align workloads precisely with their performance, sovereignty, and elasticity requirements.

The Three Domains of a Hybrid Architecture

1. Colocation: deterministic performance for execution workloads

Colocation environments sit physically adjacent to exchanges, enabling microsecond-level inference for trading models and market-data-driven signal processing. These environments prioritise deterministic performance, specialised networking, and hardware acceleration.

2. Sovereign On-Premises: regulatory control for sensitive datasets

On-premises data centres support workloads bound by residency, confidentiality, and audit requirements. Surveillance, AML, KYC, and trade-record systems operate here to maintain strict control over data movement, access, and retention.

3. Cloud: elastic compute for AI, simulation, and GenAI

Cloud regions provide the GPU scale required for computationally intensive workloads such as Monte Carlo, XVA, stress testing, embeddings, and LLM inference. Firms benefit from rapid scaling, diversified GPU families, and continuous innovation.

Guiding Principles of the Hybrid Design

Hybrid architectures follow four design principles that ensure each workload is placed where it performs optimally:

- Locality where performance demands it - latency-sensitive inference runs in colocation.
- Residency where regulation mandates it - sensitive datasets remain in sovereign environments.
- Elasticity where scale is needed - simulation and GenAI workloads burst to cloud GPU fleets.
- Consistency across all environments - governance, identity, observability, and deployment patterns remain uniform.

This principles-based approach allows firms to meet regulatory obligations and performance expectations while retaining the agility to innovate.

SUMMARY

A Unified Hybrid Operating Model

This reference architecture is not simply a federation of environments; it is a unified operating model. By combining colocation for deterministic performance, on-prem for regulatory control, and cloud for elastic compute, firms can meet the combined demands of trading, risk, surveillance, compliance, and research. Critically, the architecture maintains one set of controls, one deployment model, and one governance framework, enabling institutions to innovate rapidly without compromising regulatory integrity or operational reliability. Taken together, these architectural components form a unified hybrid platform - one whose performance, elasticity, and governance must now be validated through systematic benchmarking.

Core Components of the Reference Architecture

A production-grade hybrid platform comprises several architectural components working together as a unified system:

- Unified Kubernetes control plane: Ensures consistent deployment, RBAC, policy enforcement, and operational tooling across colocation, on-prem, and cloud nodes.
- Hybrid networking: Direct Connect or VPN links, non-overlapping CIDR ranges, and QoS policies enable predictable latency, secure connectivity, and micro-segmentation of workloads.
- GPU strategy: Exclusive allocation for trading inference, MIG partitioning for multi-tenant research, and cloud bursting for large simulations or LLM workloads.
- Storage strategy: NVMe for low-latency inference, NFS/FSx for shared datasets, and object storage (S3) for model artefacts, lineage, and versioning.
- Governance framework: OPA policies, mandatory RBAC, secrets management (Vault or AWS Secrets Manager), and controlled model lineage.
- Observability stack: Unified metrics (Prometheus/Thanos), centralised logs (SIEM), and business-aligned performance dashboards to monitor latency, GPU utilisation, and cost.

Together, these components form a cohesive architecture that supports the scale, resilience, and governance capital markets require to operationalise AI.

In the blog post Run GenAI inference across environments with Amazon EKS Hybrid Nodes1 Amazon EKS is used for hybrid nodes along with NVIDIA NIM for running GenAI on hybrid architecture. The reference architecture of the AWS post which is going to be used later in the benchmarking is shown in

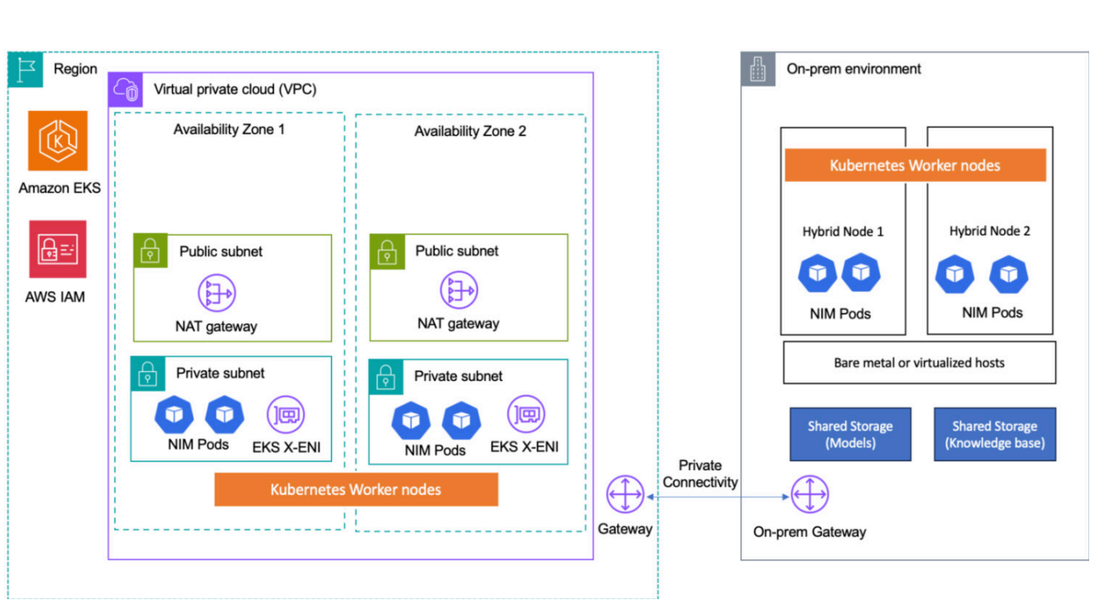


Figure 1. High-level overview of an EKS cluster in a hybrid setup derived from **https://aws.amazon.com/blogs/containers/run-genai-inference-across-environments-with-amazon-eks-hybrid-nodes/**

CHAPTER 7

Hybrid Benchmarking Framework

Before deploying AI workloads at scale, capital markets institutions must understand how models behave across colocation, on-premises, and cloud environments. Traditional benchmark approaches, focused solely on single-node speed or isolated throughput, are insufficient. Capital markets demand deterministic latency, elastic scaling, cost efficiency, and consistent governance across a distributed architecture.

A hybrid benchmarking framework provides a structured method for evaluating whether a firm's infrastructure can support AI workloads safely, reliably, and efficiently.

Benchmark Questions

A comprehensive hybrid benchmark should answer four questions that determine production readiness:

1. **Latency determinism:**
   Can collocated inference reliably operate within microsecond–millisecond windows under peak trading conditions?
2. **Elastic scalability:**
   Can cloud GPU fleets expand 10–100× during market stress or regulatory cycles without queueing or performance degradation?
3. **Cost efficiency:**
   What is the cost per million inference tokens or simulation paths across regions, GPU families, and model-serving frameworks?
4. **Governance consistency:**
   Are identity, audit, lineage, and policy controls enforced uniformly across colocation, on-prem, and cloud environments?

These questions move benchmarking beyond raw performance and toward operational viability, which regulators and internal risk teams increasingly demand.

Core Benchmarking Scenarios

Four hybrid scenarios provide insight into how workloads should be placed and scaled:

- **Colo vs cloud inference:**
  Measures whether execution-time models meet deterministic latency requirements exclusively in colocation.
- **Cloud burst elasticity (Monte Carlo / XVA):**
  Evaluates how quickly cloud GPU fleets can scale to absorb millions of simulation paths during risk spikes.
- **RAG pipeline latency (on-prem vector stores):**
  Tests the performance of retrieval–augmented generation workflows that keep sensitive embeddings local while using cloud LLMs.
- **Multi-model routing (vLLM, Triton, NIM):**
  Compares model-serving frameworks under blended workloads, reflecting real-world conditions across research, risk, and operations.

Together, these scenarios help firms determine where workloads should run and how each component of the hybrid architecture contributes to overall performance.

Example Outputs

Realistic benchmark results illustrate the performance dynamics of hybrid environments:

- Monte Carlo simulation:
  1 million paths: 240 ms on-prem vs 40 ms in cloud GPU region
  (cloud elasticity provides >5× throughput and ~65% cost reduction when bursting)
- RAG performance:
  End-to-end retrieval and generation: 104–115 ms with on-prem vector stores and cloud-based LLM inference
  (confirms hybrid RAG pipelines can meet sub-200 ms responsiveness)
- Execution-time inference:
  Colocation p99 latency: 370 µs
  Cloud region p99 latency: 3.4 ms

(demonstrates that execution workloads must remain in colocation due to jitter and physical distance)

These outputs demonstrate a consistent pattern: workloads behave fundamentally differently across environments, and hybrid benchmarking reveals the correct placement strategy.

SUMMARY

Benchmarking as the Basis for Deployment Decisions

Hybrid benchmarking does more than measure speed; it validates whether each environment can meet the operational, regulatory, and economic requirements of a production AI estate. By comparing inference behaviour, elasticity, cost, and governance across domains, institutions gain a clear view of where workloads should run and what architectural improvements are required. This benchmarking process forms the foundation for the implementation roadmap outlined in the next chapter.

CHAPTER 8

Implementation Roadmap

Implementing a hybrid architecture requires a structured and sequenced approach that reduces risk, accelerates value, and aligns technology change with governance expectations. The roadmap below reflects patterns used by leading capital markets institutions as they modernise trading, risk, surveillance, research, and operational platforms. It progresses logically from diagnostic activities to scaled enterprise adoption.

SUMMARY

A Phased Path to Enterprise-Scale Hybrid Adoption

This implementation roadmap allows institutions to progress from assessment to enterprise-scale hybrid adoption in a controlled and measurable way. Each phase builds the architectural, operational, and governance capabilities required to support AI in production. By following this sequence, firms can realise the full performance and cost benefits of hybrid architecture while ensuring compliance, reducing operational risk, and accelerating time-to-value.

Phase 1

Assessment

The first phase establishes a comprehensive understanding of the current technology estate and its constraints. Firms typically begin by mapping workloads and data dependencies, assessing latency and residency requirements, and identifying structural bottlenecks that will shape the architecture.

Key activities:

- Workload classification across latency, residency, and elasticity dimensions
- Data residency and sensitivity mapping
- Network readiness evaluation (colocation ↔ on-prem ↔ cloud)
- GPU and compute profiling to identify utilisation gaps
- Governance and operational maturity assessment

This phase provides the factual baseline for all subsequent decisions.

Phase 2

Prioritisation

Once the environment is understood, firms determine which use cases to migrate first. Early waves focus on high-value, low-dependency workloads that demonstrate impact quickly and build momentum across the organisation.

Key activities:

- Value–complexity matrix to rank initial use cases
- Identification of early wins (e.g., Monte Carlo bursting, RAG copilots, policy automation)
- KPI definition to track performance, cost, and productivity gains

This phase ensures that investment is directed where impact will be fastest and stakeholder alignment strongest.

Phase 3

Architecture Foundation

With priorities defined, firms establish the technical and governance foundations for hybrid operation. The goal is to create a unified, policy-aligned platform before migrating sensitive or latency-critical workloads.

Key activities:

- Deployment of hybrid networking, including DX/VPN and non-overlapping CIDRs
- Integration of a unified control plane spanning colocation, on-prem, and cloud nodes
- Establishment of governance frameworks (RBAC, OPA policies, lineage standards)
- Storage architecture integration for shared datasets and model artefacts
- Configuration of dedicated compute pools for inference, simulation, and GenAI workloads

This phase transforms fragmented environments into a single, consistent operational platform.

Phase 4

Migration

Workload migration proceeds in waves aligned to business priorities and technical readiness. Each workload class requires distinct placement and validation processes.

Migration patterns:

- Trading → colocation: placement of latency-critical inference models and market-data pipelines
- Risk → cloud bursting: elastic scaling for Monte Carlo, XVA, scenario modelling
- Research → hybrid RAG: on-prem vector stores combined with cloud-based LLM inference
- Surveillance → hybrid classification: real-time screening on-prem with cloud-scale analytics
- Agentic workflows → supervised hybrid execution: agents orchestrating processes across regulated and elastic environments

During this phase, firms validate latency, resilience, and compliance controls before moving into steady-state production.

Phase 5

Optimisation

Once workloads are live, firms optimise performance, resilience, and cost across the hybrid estate. This phase shifts focus from migration to continuous improvement.

Key activities:

- Autoscaling based on latency, load, and cost thresholds
- GPU consolidation (exclusive allocation, MIG, time-slicing)
- Cost governance, including region optimisation and Spot adoption
- Multi-region failover and disaster recovery testing
- Developer enablement through templates, documentation, and training

This phase ensures the hybrid platform remains efficient, resilient, and adaptable to evolving business needs.

CHAPTER 9

Business Value & Outcomes

A unified hybrid architecture delivers measurable improvements across performance, cost, productivity, and governance. Unlike traditional infrastructure programmes, where benefits are diffuse or long-dated, hybrid AI produces rapid, quantifiable gains across trading, risk, surveillance, compliance, and operations. The following impact areas reflect outcomes observed in early adopters across global capital markets.

1. Performance Gains

Hybrid architectures place each workload in the environment where it operates most effectively. As a result, firms achieve consistently higher performance across latency-critical and compute-intensive functions.

Observed outcomes:

- Execution latency: p99 inference times below 400 µs in colocation environments
- Risk computation: 5–10× faster Monte Carlo, XVA, and scenario workloads
- Peak resilience: Zero queueing during volatility or regulatory stress events

Implication: faster decision-making, improved execution quality, and more reliable risk visibility.

2. Cost Efficiency

Shifting burst workloads to cloud GPUs, consolidating on-prem capacity, and aligning compute supply with demand significantly reduce infrastructure costs.

Observed outcomes:

- GPU cost savings: 30–70% reduction through cloud elasticity and Spot adoption
- On-prem footprint: 20–40% reduction in GPU estates by eliminating peak-only provisioning
- Compute optimisation: Efficient regional placement further reduces cost per simulation or token

Implication: firms achieve lower and more predictable run-rate costs while increasing available compute capacity.

3. Productivity Improvements

Hybrid AI removes bottlenecks that slow research, compliance, and operational workflows. By enabling rapid experimentation, scalable retrieval-augmented generation, and supervised agentic automation, firms can materially increase output.

Observed outcomes:

- Research throughput: 2–3× increase through hybrid RAG and LLM-assisted workflows
- Compliance efficiency: 60% faster query response and document review cycles
- Operational automation: 30–50% reduction in manual workload via agentic execution

Implication: analysts, risk professionals, and operations teams spend more time on judgment and less on process.

4. Governance & Resilience

Hybrid architectures strengthen control frameworks by unifying security, auditability, and lineage across environments. This reduces regulatory friction and enhances operational stability.

Observed outcomes:

- Unified RBAC: one identity and access model across cloud, on-prem, and colocation
- End-to-end lineage: full traceability of models, data flows, and inference behaviour
- Operational resilience: hybrid multi-region failover aligned to DORA and PRA expectations

Implication: firms meet regulatory expectations more consistently while improving risk posture and reducing audit findings.

SUMMARY

Hybrid as the Foundation for AI at Scale

Taken together, these outcomes demonstrate that hybrid architecture is not simply an infrastructure upgrade but a foundational enabler of AI at scale. Institutions gain materially better performance, lower cost, higher productivity, and stronger governance which are all essential in a market where speed, intelligence, and resilience increasingly define competitive advantage. As firms expand AI adoption across trading, risk, surveillance, and operations, hybrid architecture becomes a strategic necessity rather than an optional enhancement.

CHAPTER 10

Conclusion

Capital markets sit at the intersection of rising volatility, intensifying regulatory expectations, and rapid advances in AI. These forces are reshaping how trading, risk, surveillance, compliance, and research functions operate. **Traditional cloud-only or on-premises-only architectures were not designed for this environment** and cannot meet the combined demands of deterministic performance, data sovereignty, elastic compute, and unified governance.

Hybrid AI is therefore not a strategic preference: **it is an architectural requirement.**

Institutions that transition to hybrid architectures are already realising material gains across the operating model:

- Deterministic execution performance, enabled by colocated inference where latency defines outcomes
- Consistent regulatory alignment, through unified governance, lineage, and auditability across environments
- Operational resilience, supported by
- multi-region hybrid architectures aligned with DORA and PRA expectations

The next decade of capital markets will be defined by institutions that can **operationalise AI under conditions of volatility, regulatory scrutiny, and accelerating model complexity**. Traditional cloud-only and on-premises-only architectures cannot meet these demands simultaneously.

Hybrid architecture provides the o**nly operating model capable** of delivering deterministic performance, regulatory control, elastic compute, and unified governance at scale. As AI becomes inseparable from market infrastructure, hybrid will not differentiate leaders from followers it will determine **who remains competitive at all**.

Appendix 1: Glossary of Terms & Abbreviations (Technology & Operating Model)

AI (Artificial Intelligence)
Technologies that enable automated prediction, reasoning, and decision support across trading, risk, and operations.

Agentic Workflows
AI-powered orchestration of multi-step business processes across systems, with built-in guardrails and supervision.

CI/CD (Continuous Integration / Continuous Deployment)
Automation practices enabling rapid, reliable software and model delivery.

Colocation (Colo)
Exchange-adjacent infrastructure delivering ultra-low-latency performance for execution and market-data workloads.

Elastic Compute
On-demand scaling of compute resources to meet variable workload demand efficiently.

GenAI (Generative Artificial Intelligence)
AI models capable of producing text, code, or analytical outputs, accelerating research, compliance, and operations.

GPU (Graphics Processing Unit)
High-throughput compute hardware optimised for AI training and inference.

Hybrid Architecture
A unified operating model combining colocation, on-premises, and cloud to optimise performance, cost, and compliance.

Latency Determinism
Predictable, low-jitter response times required for execution-critical workloads.

LLM (Large Language Model)
Large-scale language models used for summarisation, reasoning, and analysis.

MIG (Multi-Instance GPU)
GPU partitioning technology enabling efficient multi-tenant utilisation.

Monte Carlo Simulation
Compute-intensive risk and pricing techniques requiring elastic scaling.

RAG (Retrieval-Augmented Generation)
An AI pattern combining retrieval of enterprise data with generative models to improve accuracy and relevance.

RBAC (Role-Based Access Control)
Standardised access control for enforcing least-privilege access across platforms.

SLA (Service Level Agreement)
Defined performance and availability targets for systems and services.

VaR / XVA
Risk and valuation metrics that drive large-scale, compute-intensive workloads.