PREPARED FOR NVIDIA INNOVATION LAB · MAY 2026

H100 Outcomes

Three-module validation evidence base on the 8×H100 DGX-class Innovation Lab node

The H100 allocation was used as the core acceleration environment for SAA Alliance's institutional-risk platform — not as a small experiment. It moved three modules from prototype-scale reasoning to auditable, reproducible validation evidence: the ARIN22 deterministic risk-computation kernel, the Hydra evidence substrate that powers the KOKON governance plane, and the Global Risk Platform subject-stress surface. This brief sets out what is closed pre-client, what was made possible specifically by the H100 environment, and where the broader NVIDIA stack sits in the resulting architecture.

PRE-CLIENT · INSTITUTIONAL DILIGENCE-READY

1 Same-surface benchmark — industry fast methods vs ARIN22 governed promotion

SUBJECT-STRESS SURFACE · SAME H100 EXECUTION
10,080 rows · 28 stress scenario classes × 6 horizons · PRNG MC at 10,000,000 paths as reference

METHOD	CVAR99 P95 ERROR VS MC	CVAR99 MAX ERROR VS MC
Cornish-Fisher · industry-standard fast method	**73.62%**	**74.58%**
Delta-Gamma proxy · industry-standard fast method	**17.58%**	**19.12%**
ARIN22 governed deterministic promotion · 1M vs 10M tier drift	**0.289%**	**< 1%**

Same compute, same surface, side-by-side. The H100 environment is what made same-surface negative-control execution at this resolution possible.

2 NVIDIA stack adoption — where the architecture pulls on the stack

Compute	8×H100 (80 GiB) DGX-class node from the Innovation Lab — batch validation, GPU↔CPU parity proofs, large-scale chunked stress campaigns.
Inference	NVIDIA NIM API across 5 of 7 production modules; vLLM on H100 as the local-sovereignty primary path; 22-agent ARIN council served on this fabric.
Guardrails	NeMo Guardrails on 2 of 7 modules as the narrative-gate layer that prevents LLM output from contradicting the deterministic kernel verdict.
Data stack	RAPIDS / cuDF for ontology pre-processing on the typed-relation knowledge graph that grounds the agentic council.
Programs	NVIDIA Inception member · Innovation Lab participant; the H100 DGX-class node above as the canonical validation substrate.

3 Headline outcomes

50,400
100M-eq jobs · 0 failures · chunked 10×10M

8.8 B
paths / backend · 0 execution failures

1 hash
across 1,000 fresh-process repeats · 0.0% diff

58 / 58
data-room artifacts · 0 missing

4 Module 1 — ARIN22 Kernel · Enterprise Wave (Portfolio surface)

PORTFOLIO-CLASS SURFACE Tight tolerance: deterministic kernel anchored to a 10M-path MC challenger at the 99.9 tail; CVaR 99.9 max divergence held to **0.035%** across 50 selected portfolio cases.

- **Enterprise Wave validation pack closed** — ~27,700 distinct portfolio cases: 8,800 GPU grid + 8,800 CPU mirror + 10,080 subject-stress surface + 20 historical replay. 8.8B paths per backend; 0 execution failures.
- **CRN 99.9 GPU↔CPU parity · PASS** — 100 worst-tail cases × 100M paths/backend; backend-pair common-random-numbers parity validated.
- **10M-path MC challenger** — 50 cases × 200 reference runs × 2.0B reference paths · CVaR 99.9 max diff **0.035%**; p95 **0.015%**.
- **Tick replay (Canon Upgrade)** — 24M updates · GPU p99 **0.44–0.46 ms** · p999 **0.64–0.70 ms**.
- **Pre-trade gate (Canon Upgrade)** — 60M-order fixture · GPU p99 **0.85–0.93 ms** · p999 **1.8–2.5 ms** (disclosed bands across N=3 repeats).
- **Determinism** — 1,000 fresh-process repeats produced **1 unique hash** · 0.0% max metric divergence.
- **STAC-A2-inspired Heston LSM Greeks lane** — 310M paths on 8×H100 · 2.48B valuation paths · max lane wall 14.875s · 77 archived artifacts. Workload built to the STAC-A2 archetype, self-run; *not* STAC-certified.
- **Artifact closure** — 58 / 58 data-room artifacts + 53 / 53 closure-manifest artifacts · 0 missing.
- **Canon Upgrade multi-asset wave (30 May 2026)** — **11 institutional asset classes** validated on 10M MC with **N=3 fresh-process repeats**: equities · ETFs · fixed income · FX · commodities · energy · precious metals · sovereign · crypto · multi-asset · options. CVaR99.9 GPU↔CPU parity p95 **~0.27%**, max ~0.62% across repeats; variance ~0.012% on p95.

5 Module 2 — Hydra Substrate · KOKON governance plane

EVIDENCE-GROUNDED SUBSTRATE H100 enabled the architectural move from "RAG plus graph" to a typed evidence-grounded intelligence substrate at hundreds-of-millions relationship scale.

- **Typed knowledge graph (post-enrichment, 2026-05-24)** — ~114M typed nodes · ~336M typed relationships · ~9.1M RAG chunks across institutional risk domains (sovereign / city / supply-chain / climate / geopolitics / cyber / markets / regulatory / mathematical methodology).
- **Typed promotion discipline** — 181 distinct relationship types; generic *RELATES_TO* edges reduced to **0** in the promoted typed graph layer.
- **Claim extraction at scale** — ~9.1M claim objects · ~73.2M typed claim/evidence-related edges · ~31K promoted direct claim links; claim → evidence / entity / source / chunk · temporal validity · support / contradiction / weakening relations.
- **Governor architectural refusals** — 5 Decision-Integrity Invariants (Data Immutability · Append-only Verdicts · No Backflow · Single Exit · Data Freshness) + 7 Math-to-Narrative Audit Rules (Cascade Amplification · SPOF Fabrication · Regime Fabrication · Severity Escalation · Downplay in Critical · Polarity Inversion · Cascade Through Zero). The narrative layer is structurally prevented from contradicting the deterministic kernel.
- **Audit lineage · Live** — HMAC-signed verdict trail + ed25519 hash-chained ledger; every report carries a council packet, Governor verdict artifact, agent evidence packet, deterministic replay data and audit trace.
- **NVIDIA stack tie-in** — RAPIDS / cuDF on this graph at ingest and promotion; NIM for 22-agent ARIN council inference; NeMo Guardrails on the narrative gate; vLLM on H100 for local sovereignty primary.
- **Tesla Wave contagion engine** — graph diffusion · dependency propagation · causal edges · temporal state transitions · convergence diagnostics · sensitivity sweeps · historical cascade backtesting; Governor-enforced — unvalidated contagion outputs are supporting evidence only.
- **Reproducible substrate snapshot** — Postgres / pgvector (35.8 GiB) and Neo4j (148.4 GiB) controlled snapshots archived with SHA-256 manifests and cold-lakehouse manifest covering ~946 GB / 4,768 tracked objects, PASS with empty-file warnings only.

6 Module 3 — Global Risk Platform · Subject-Stress surface

SUBJECT-CLASS SURFACE Distinct surface from Module 1: company / city / country subjects under regulatory-style stress scenario classes.

> ***Surface-tolerance note:*** *the Subject-Stress surface carries a wider error budget than the Portfolio surface — governed deterministic promotion within CVaR99* ***6.35% p95 / 9.79% max****; non-promoted rows route to MC fallback under a hash-logged decision. This is a different surface from Module 1's 0.035% Portfolio-surface tolerance; the two figures are not interchangeable.*

- **Validation surface** — 5,040 subject-scenario-horizon rows · 30 subjects (10 company / 10 city / 10 country) · 28 stress scenario classes × 6 horizons (30D / 90D / 180D / 360D / 5Y / 10Y).
- **100M-equivalent H100 campaign** — 10 × 10M chunked execution · **50,400 subject-scenario-horizon jobs · 0 failures** · run ID `chunked100M_20260522T130555Z` · archived with SHA-256 manifest. The disclosed boundary: 100M-equivalent via 10 × 10M chunks — not a single monolithic 100M run.
- **Governed multi-lane result** — 57.60% deterministic promotion (1,998 / 3,469 material rows); 1,471 rows routed to MC fallback; lane mix canonical 1,941 · Lie 51 · Strang 6. Effective p95 latency 1,198 ms · promoted-deterministic p95 latency 19.187 ms.
- **Independent subject-separation proof** — three isolated packs (Country / City / Company), manifest hash `sha256:4f6057ed…`. No country, city or company result is used as evidence for another subject kind.

SUBJECT	MATERIAL ROWS	PROMOTED	MC FALLBACK	PROMOTION RATE	CVAR99 P95 ERROR
Company	1,157	667	490	57.65%	6.47%
City	1,148	686	462	59.76%	6.69%
Country	1,164	645	519	55.41%	5.69%

- **Sampler-stability evidence** — 10,080 Sobol/LHS jobs vs PRNG 100M-equivalent reference · 0 failures · 0 missing · **0 tail-route changes** · **0 policy-bucket changes**. Sobol VaR99 p95 drift 1.88%; LHS p95 drift 1.88%.
- **Error budget breakdown** — tier drift 0.28% · sampler drift 0.41% · **seed spread 3.75% (dominant)**. Disclosed alongside results, not buried.
- **Earth-2 / weather-risk validation path** — forecast / downscale / visual-frame artifacts normalized into stress packets, routed into math-core evaluation, attached to evidence-graph lineage with SHA-256 packet hashes, exported into ARIN. Proxy mode explicitly gated; engine-backed weather validation requires `--require-engine`.
- **Shadow math lanes (evaluation, not silent replacement)** — multivariate EVT · tail / vine copula · Hawkes jump processes · DebtRank / Eisenberg-Noe contagion · percolation cascades · conformal calibration. Lanes identify rows worth deeper validation; they do not silently replace the production stress governor.

7 What is not claimed

No external certification
No STAC certification; Heston / options work is STAC-A2-*inspired*, self-run on 8×H100. No NVIDIA-attested benchmark.

No realized regulatory backtest pass
Kupiec · Christoffersen · Acerbi-Szekely structurally gated until 250 matched observations. Shadow Ledger armed; currently 1 / 250.

No live-broker / live-exchange production
Status is PRE-CLIENT. Pre-trade and tick latency figures are internal evidence on a controlled test substrate, not live execution telemetry.

No bank-client production sign-off
Customer-side model-risk-management sign-off is the deliberate next step. First pilot ask: 90-day technical co-validation on a real institutional book.

Two surfaces, two budgets
Portfolio surface (Module 1) holds CVaR within 0.035% of 10M MC. Subject-Stress surface (Module 3) holds governed promotion within 6.35% p95 / 9.79% max. They are not interchangeable.

Disclosure tiering
All figures here are TIER-1 capability evidence under SAA's own disclosure canon. Per-run SHAs, hardware fingerprints, precise speedup factors and method derivation are TIER-2 NDA.