Architect's Guide to AI-Driven Systemic Risk Mitigation in Post-Trade

Moiz Kohari

In today's markets, post-trade isn't a back-office formality—it's a real-time battleground for systemic stability. Regulators demand faster answers. Traders want real-time moves. AI is ready to deliver—but only if your infrastructure can keep up.

This guide unpacks how leading financial institutions are rebuilding their risk stacks to run faster, smarter, and in sync with the markets.

From 15TB VaR cubes to petabyte-scale stress libraries, every player in the FMI chain—brokers, CCPs, CSDs, custodians, and clearing banks—is hitting the same wall: storage that can't move data fast enough to feed modern AI.

That's where flash-first, on-prem object storage—like DDN Infinia—comes in. By combining lightning-fast ingest, zero-copy access, and sovereign-soil compliance, it unlocks a new class of performance for risk analytics and AI.

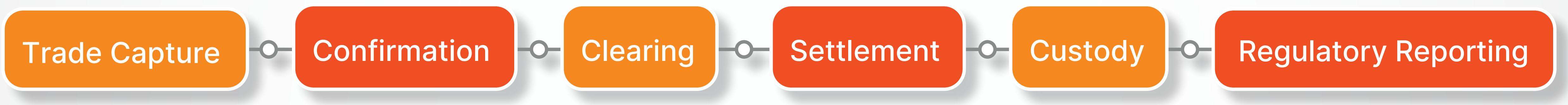
The outcome: overnight jobs finish before London wakes up, GPUs stay fed, and decisions shift from reactive to real-time.

Table of Contents

Why Risk Calculations Matter

Cast your mind back to 2008: Lehman Brothers and Bear Stearns vanished almost overnight, not because traders had no models, but because the industry simply couldn't surface default risk fast enough. Regulators reacted with a blunt instrument: nightly, portfolio-wide stress tests that most banks must run.   By 2011 the Financial Stability Board (FSB) had coined the term G-SIB - Global Systemically Important Bank - shorthand for "if this one fails, everyone feels it". The 2024 FSB list of G-SIBs names 29 banks, unchanged from 2023 .   Basel III, the FDIC Act, and FSOC rules all lock in one daily obligation: measure and manage counter-party exposure.

Here's how a trade moves through the post-trade Financial Market Infrastructure (FMI) to make that happen—step by step.



Trade Capture → Confirmation → Clearing → Settlement → Custody → Regulatory Reporting

Every participant in the post-trade chain therefore runs an end-of-day default drill. If we could refresh those numbers every few minutes, margin calls and collateral moves would track markets in real time. So why don't we? Because the plumbing chokes. A single ten-day Value-at-Risk (VaR) job at a tier-one broker spins out 15 TB or more of Parquet (a file format for storing data); try doing that ten times an hour and watch your storage backplane beg for mercy. This is exactly where DDN's data-intelligent platform separates signal from marketing.

DDN powers more than two-thirds of the top slots on the IO-500 list of the world's fastest file systems IO500, and NVIDIA's internal AI super-cluster rides on the same kit Business Insider. The point is simple: if your data array can't keep GPUs fed at line rate, intra-day VaR shrinks to a slide-deck fantasy.

The next section illustrates what causes a single VaR run to balloon so quickly—and how an on-prem, flash-accelerated object store like DDN Infinia can turn that data deluge from a bottleneck into a competitive advantage.

How Risk is Calculated – Where Data Bottlenecks Lives

Risk modeling in post-trade environments hinges on four tightly interwoven analytics engines: Value at Risk (VaR), Conditional VaR (CVaR), Incremental VaR (IVaR), and the Greeks. Together, they offer a multi-dimensional view of portfolio exposure—but each places a distinct and growing burden on infrastructure. These aren't theoretical models collecting dust. They power the decisions behind every margin call, every collateral shift, and every end-of-day compliance check. And they don't just need fast compute—they need fast data.

Let's break down what each model does, why it matters, and how the data load adds up fast.

» **Value at Risk (VaR)**
Start with the core: Value at Risk. After every trading session, banks run a Monte Carlo or historical-simulation engine that generates anywhere from 50,000 to a million market-shock paths. It revalues every trade across those paths and asks: **"With 99% confidence, how much could I lose over the next ten days?"**
The answer lives in a three-dimensional cube—keyed by scenario ID, time horizon, and portfolio element. On a tier-one bank's book, that cube compresses to roughly 15 terabytes for every run (required to run at least once a day, end of day).

» **Conditional Value at Risk (CVaR)**
Regulators then zoom in with Conditional VaR (Expected Shortfall). Take the same cube, sort the profit-and-loss outcomes, slice off the worst one per-cent tail, and average them. No new scenarios are needed but you must rescan the entire data set - a second I/O hit that can double runtime if storage can't keep up.

» **Incremental (IVaR)**
Traders, on the other hand, care about Incremental VaR. Before they route an order, they flip the candidate trade into the portfolio, rerun (or cleverly recycle) the grid, and look at the delta between "with" and "without" certain scenarios that may impact a trade. This calculation must happen in minutes, not overnight, or the desk will ignore it.

- **The Greeks: Delta, Gamma, Vega, Theta, p**
  Layered on top are the Greeks - Delta, Gamma, Vega, Theta, ρ - the minute-by-minute sensitivities that tell a desk how today's book will move if rates nudge or vol skew twists. Greeks are tiny numbers, but they refresh constantly, and desks store every snapshot next to the VaR cube so quants can replay any intraday slice on demand and collateral systems can issue margin calls in near real time.

None of these calculations are compute-bound. They're data-bound. As mentioned earlier a single 10-day cube generates over 15 terabytes of data—written at speed, read again for CVaR, and tapped all day by IVaR and Greek queries. Try feeding that from a cloud bucket, egress costs alone for every access will cost you 9 cents per gigabyte of data. Try streaming it from a classic NAS and your GPUs will starve.

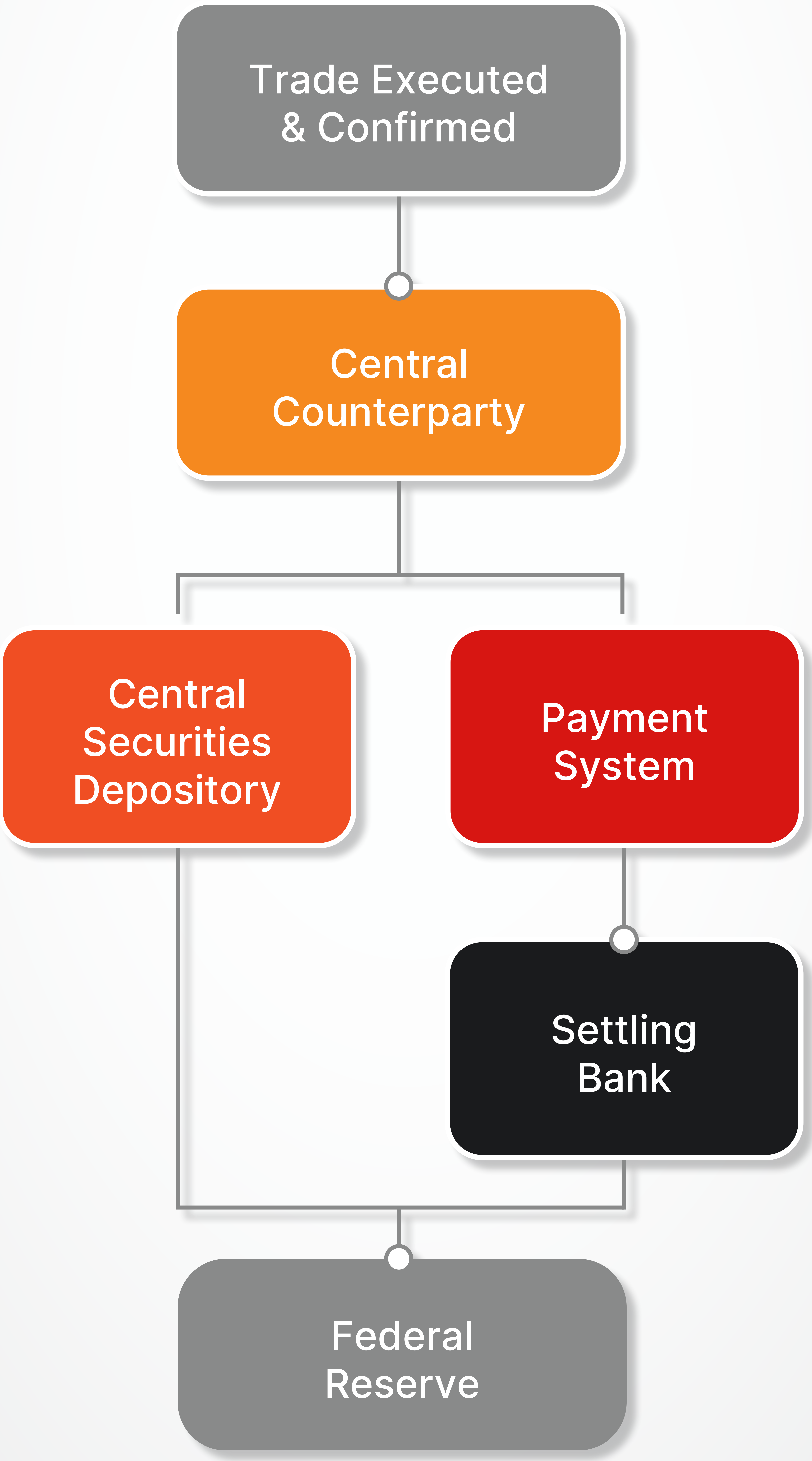
This is exactly where flash-backed object storage changes the game. An on-prem DDN Infinia swallows the nightly burst at 25 GB/s, finishes CVaR reads before London opens, and streams data to IVaR and Greek engines in place—no copying, no tolls. GPU grids tear through the math when the data keeps up. Industry telemetry shows most clusters stall at ~70% utilization because they're waiting on I/O. Infinia-based POCs have helped push that north of 95%.

Bottom line: risk math isn't bottlenecked by CPUs. It's bottlenecked by data pipes. Put the cube on fast flash, and the engine finally breathes at the speed traders and regulators demand.

Financial Market Infrastructure

Once a trade is locked in, these specialist utilities take over: brokers line the deal up, a CCP steps in between the two parties, a depository shifts the shares, a payment system moves the cash, and custodians keep the books straight.

Each stop has its own flavor of default risk, its own regulatory stopwatch, and its own mountain of data to prove it did the job. Here's how that chain breaks down—and why a fast, on-prem object store like DDN Infinia keeps every link running on time.



©DDN 2025

Broker Dealers

Thousands, led by Vanguard, Charles Schwab, Fidelity and Merrill Lynch

A small group of tier-1 broker-dealers—just 100 out of more than 9,000 globally—handle the vast majority of market volume. To manage that scale, they run on a tightly engineered three-tier data stack:

- **Tier 1:** Captures trades in real time using FIX engines and market-data buses that feed into ultra-low-latency systems like kdb+ or Kafka.
- **Tier 2:** Runs the overnight risk grid. Trades, curves, and shocks are dropped into an on-prem S3-compatible object store—ideally one fronted by NVMe flash, like DDN Infinia. From there, a Kubernetes or Slurm grid launches thousands of Monte Carlo simulations across 10,000–20,000 CPU cores and hundreds of GPUs.
- **Tier 3:** The output is a 10–15 TB Parquet-based VaR cube, written directly back to the same bucket for high-throughput reuse.

Once the overnight VaR cube is sealed, a second wave of AI-powered analytics kicks in—transforming static risk data into dynamic insight. Each step in this phase runs directly off the same S3-compatible bucket, minimizing duplication and maximizing GPU and CPU throughput.

Here's how tier-1 broker-dealers extend the value of their risk infrastructure:

- **Scenario Clustering**
  A PyTorch job encodes each market-shock scenario into a 256-dimensional vector, then applies HNSW in Milvus to surface "look-alike" shocks—critical for stress testing and pre-trade risk checks.
- **Tail Driver Attribution**
  XGBoost models calculate SHAP values on the worst 1% of outcomes, surfacing the key risk drivers that pushed CVaR into the red. The result: explainable AI, not black-box risk.

- **IVaR on Demand**
  A fast Apache Spark job reloads the Parquet VaR cube, inserts a hypothetical trade, and re-runs only the affected paths. With flash-backed storage, results return in under two minutes—fast enough for real-time trade decisions.

- **Liquidity Forecasting**
  A long short-term memory (LSTM) network consumes historical cash-flow vectors to forecast daylight funding needs. Treasury desks use this signal to size intraday credit, avoiding surprise overdrafts.

- **Margin Optimization**
  A reinforcement-learning agent reads Greeks and IM grids from the bucket, rebalancing collateral hourly across CCPs to optimize both margin usage and available liquidity.

All of this rides on a single S3 endpoint. GPUs read Parquet directly, CPUs execute Spark jobs in place, and compliance archives migrate to disk-tier after cutoff—no data egress, no replication, and no latency bottlenecks.

This is where flash-based object storage like DDN Infinia proves decisive: it delivers the throughput to sustain real-time analytics, enabling predictive, prescriptive, and generative AI workflows to operate off a common data foundation—without starving your grid or your budget.

Central Counter Parties

About 60, led by LCH, CME, ICE and NASDAQ

A CCP stands at the center of every trade it clears—so it must prove it can survive the worst-case member default. That means crunching a stress-scenario library that can span one to three petabytes—every historic crash plus synthetic "top-two-member" wipes. Public figures from LCH show these runs scaling to approximately 30,000 CPU cores and 500 GPUs.

A clean CCP data stack looks like this:

- **Tier 1:** Trade and Margin Feeds
  Real-time trades, variation margin, and IM calls land in a Kafka bus and an ultra-low-latency store such as kdb+.

- **Tier 2:** Stress Library
  All scenario paths are stored in a single S3 bucket on an on-prem DDN Infinia. NVMe flash absorbs daily updates, while bulk drives retain the multi-year archive.

- **Tier 3:** AI Workloads on Top
  - Default-probability models: Gradient-boosted trees ingest balance sheet data and intraday margin breaches to flag early trouble.
  - Contagion graphs: A graph-neural network maps cross-member exposures and scores "who falls next" if one defaults.
  - Margin optimization: A reinforcement-learning agent reallocates collateral across currencies hourly, reading stress cubes directly from the bucket.
  - Vector search for crisis replay: Scenario paths are embedded into a vector DB (Milvus), allowing risk teams to query, "Show me shocks most like 24 Feb 2022," and pull results in seconds—no full-cube scan needed.
  - LLM summaries: A local GPT variant digests the latest run and drafts daily risk memos to regulators—ensuring data never leaves sovereign soil.

Because all jobs read and write the same Infinia bucket, there's no overnight copying, no $0.09/GB egress, and GPUs and GPUs stay productive instead of idle. The result: real-time margin decisions, faster default drills, and a storage bill a fraction of cloud-only setups—exactly what a systemically important CCP needs when markets go sideways.

Central-Securities Depositories

About 40, led by Euroclear and DTC

A CSD's job is simple on paper: move legal title for every share or bond on settlement day. In practice that means ingesting a flood of ISO 15022/20022 messages, reconciling them in real time, and keeping the full audit trail for at least seven years.

Data volume:

- **Peak U.S. equity day:** DTC handles about 200 GB of settlement instructions per hour - roughly 1.7 million transactions worth $500 billion.
- **Retention:** Seven-year online retention is required per CSD, meaning the total data volume under management approaches a petabyte per CSD..

» **Clean architecture:**

- **Tier 0 ingestion:** SWIFT/ISO messages land in a Kafka bus and an ultra-fast query store (often kdb+).
- **Tier 1 object store:** All day-files roll into a single S3 bucket on an on-prem DDN Infinia. WORM mode satisfies immutability rules.
- **Compliance:** Regulators insist the golden copy stays on domestic soil and is restorable inside two hours.

» **AI workloads that ride the same bucket:**

- **Fail-prediction:** A gradient-boost model flags trades likely to miss DvP cut-off, reading fresh instruction files straight from flash.
- **Intraday buy-in risk:** LSTM nets learn historical fail patterns and predict liquidity shortfalls so the CSD can pre-warn members.
- **Exception clustering:** A small GPU pod embeds free-text exception comments, groups similar problems, and routes them to the right ops team—no ETL, the messages are already in the bucket.

- **Reg-report generator:** An LLM summarizes settlement statistics for daily filings; zero data leave the premises.

With Infinia, ops teams can search exceptions in seconds instead of waiting for end-of-day batches. Regulators get an immutable, sovereign-soil audit trail, and finance avoids the $0.09/GB cloud-egress charges that would otherwise apply every time a replay is needed.

Settling Banks

Roughly 25 that act for others, J.P. Morgan, Citi, HSBC, BNY top the list

Roughly 25 global banks—including J.P. Morgan, Citi, HSBC, and BNY—act as the operational backbone for clearing payments across Fedwire, TARGET2, CHAPS, and BOJ-NET. Every cleared trade needs a final cash leg, and these banks make it happen. The individual messages (ISO 20022 pacs.009 or legacy MT 910) may be small, but the frequency and compliance load are anything but. Treasury desks snapshot liquidity every 15 minutes, and regulators require that history be stored for a full decade.

» **Daily flow:**

- Roughly 100 000–300 000 payment messages

» **Core models:**

- Liquidity-at-risk: A simple LSTM predicts end-of-day funding needs from the rolling 15-minute cash grid.
- Overdraft alert: Gradient-boosted trees flag members likely to hit intraday credit caps.
- Payment clustering: Embedding plus k-means groups high-value wires for compliance review.

All of this runs on a single DDN Infinia bucket. By keeping payment archives on a flash-fronted, on-prem object store:

- » Instant look-ups—no waiting for Glacier recalls.
- » Zero S3 retrieval fees.
- » Sovereign-soil compliance for Fed and ECB audits.
- » GPUs stay >95 % utilized because flash storage streams history at line speed.

The Result: The cash leg stays fast, the regulators stay happy, and the invoice stays small.

CLS Bank

PvP for 18 currencies, ≈ $7 trillion settled daily

Every FX trade that settles through CLS produces two legs and one confirmation.

The utility buckets those confirmations and PvP payment messages - roughly 50–70 TB of new data each month - and must hold them in tamper-proof form for at least ten years.

A governance-locked bucket on a DDN Infinia keeps the archive on sovereign soil, satisfies the WORM rule, and lets ops pull any file in seconds instead of waiting hours for a cloud restore.

Add a small GPU pod and the same dataset feeds anomaly models that flag unusual payment patterns before cut-off - no extra ETL, no egress fees.

Custodians

State Street, BNY, Northern Trust, etc.

Every night custodians compare what should have settled with what actually did.

- » The deltas run into millions of rows and land at about 20 TB of files each week - corporate-action adjustments, income events, FX sweeps, and break reports.
- » Storing that flow on a flash-backed DDN Infinia cuts the break-resolution cycle from "see it tomorrow" to "fix it before Asia opens." Ops teams query the bucket in place; a small GPU cluster breaks by root cause and flags the ones most likely to age into claims, saving costly follow-ups.
- » CCPs, CSDs, settling banks, CLS, and custodians together pump out petabytes of stress cubes, margin sims, payment messages, and reconciliation files.

Traditional SANs buckle, and public-cloud egress makes a single nightly VaR pull painful.

An Infinia flash object store keeps all that data in-house, streams it to GPUs at full speed, and still speaks the S3 language every analytics tool already knows. Post-trade may be plumbing, but when terabytes have to move in real time, smart storage is as critical to stability as any CCP.

Why Cloud Based VaR Calculations Suffer

First, why the cloud looked perfect: exchanges were already spraying raw ticks into AWS and Azure metro zones, so keeping the VaR cube beside the firehose felt smart. Elastic compute sweetened the offer - spin up 10 000 spot cores, load scenarios into S3, pay only while they hum, wake up to a finished run. For a moment the economics checked out.

Reality, though, is stubborn. S3 Express One Zone still prices at about $0.11/GB-month - four-plus times S3 Standard - and a petabyte of hot risk data costs roughly $6 million over four years before a single CPU cycle.

Pulling a 10-TB cube home for any "sensitive" task costs another $900 per trip; do that every trading day and you burn six figures a year in tolls. A 10-Gbps Direct Connect needs three hours to move that cube—fine on a quiet night, impossible when the ECB fires off an emergency hike. Meanwhile, the BoE and EBA now insist critical data be restorable on domestic soil inside two hours, and some data vendors slap an extra license fee on every cloud core you touch.

Drop the same cube onto a flash-fronted object store in your own data-centre—DDN Infinia is a common pick—and the egress line item disappears. The overnight write burst lands at 25 GB/s, GPU grids stay > 90 % busy, and the VaR job wraps before midnight. When a regulator orders a super-stress rerun, you can still burst cloud CPUs, but you move only the rolled-up answers, not the petabytes of raw scenarios, cutting network spend by an order of magnitude.

Generative and predictive models feed directly off the same local S3 bucket. A transformer can embed every scenario, a graph network can rank contagion paths, and an agentic bot can rebalance collateral—none of them wait for cloud latency or trigger extra licence fees. The result: faster insight, lower cost, and a compliance story that passes even the toughest sovereign-data audit.

Why AI Pushes the Pendulum Back On-Prem

AI FLAVOUR	WHAT IT NEEDS	WHY LOCAL FLASH S3 WINS
Predictive models (LSTM cash-flow forecasts, XGBoost margin alerts)	Fast rereads of years of tick/curve history	10× lower latency than cloud object, no $/GB read fees
Gen AI & embeddings (LLM that explains "why CVaR spiked")	GPU access to the full VaR cube to build vector stores	25 GB/s NVMe keeps GPUs >90 % busy; no WAN hop
Agentic AI (RL agent that re-balances collateral every hour)	Write/commit small deltas every few minutes	Governance-locked bucket meets audit rules; sub-second commit time

AI workloads do not tolerate slow access to data. By bringing the risk calculations home institutions are able to meet their deadlines while delivering significantly better TCO. If you swap your aging NFS infrastructure for Infinia, the grid code just runs. Your nightly job finishes hours sooner; zero egress costs, one-tenth the storage bill (≈ $600 k vs. $6 M over four years for 1 PB), and the same bucket feeds predictive, generative, and agentic AI workloads in real time.

Use cloud CPUs when you truly need burst compute - pull back only the summary, not the entire cube - and keep the critical data where regulators, GPUs, and your balance sheet all prefer - on fast flash you control.

The Engines Behind the Numbers

The fancy math doesn't run itself. Every VaR cube, CVaR tail slice, and intraday Greek you've read about is cranked out by a handful of software platforms that sit at the heart of broker-dealers, CCPs, and custodians.

Murex, Calypso, Front Arena, Numerix, Cassini, and OSTTRA are the pipes that turn raw market feeds into risk metrics regulators will sign off on. They all chased the cloud for quick scale, hit the same egress and latency walls, and now need a fast, on-prem S3 landing zone to keep the data - and the auditors - close.

That's where DDN Infinia fits: same API the apps already speak, local flash to keep the grids humming, and disk economics that stop the storage bill from eating the desk's P&L.

- **Murex MX.3**
  Roughly 350 banks and asset managers run MX.3 for everything from trade capture to x-VA. Big users include UBS, National Bank of Canada, and OCBC - plus four CCPs that use it for listed-derivatives risk. Annual license and support revenue is estimated at $1 billion. The nightly 10-day VaR cube can hit 15 TB; grids stall if storage can't feed 20 GB/s.
  - Parking the cube on Infinia's NVMe tier keeps the run under an hour and removes public-cloud egress charges.
  - Murex benefits because faster POCs close deals; clients benefit because the same S3 endpoint still works.
- **Adenza Calypso (Now part of Nasdaq)**
  About 200 banks and seven clearing houses—including LCH RepoClear and Nasdaq Clearing—depend on Calypso. Software revenue runs near $650 million a year.
  - Clearing installs keep petabytes of stress-test libraries on disk so they can replay default scenarios.
  - Hosting those libraries on a sovereign Infinia bucket saves multimillion-dollar cloud bills every time LCH reruns a drill and meets the <2-hour domestic-recovery rule.

- **FIS Front Arena**
  Used by roughly 120 dealers and buy-side shops—heavy clusters in the Nordics and the UK. Perpetual licences are about $4 million per site; maintenance adds $1 million a year. Desks calculate per-minute Greeks and reload positions in real time; latency shows up on trader screens.

  - Infinia's flash tier cuts read latency below a millisecond, keeping prices fresh and screens responsive.

- **Numerix OneView**
  Numerix serves about 700 sell- and buy-side customers. The cloud-native grid can burst 100 000 scenarios, but dashboards need the Parquet output back immediately. AWS throttles large-file rereads; local flash does not.

  - Infinia streams the same objects to GPUs and BI tools without delay, lifting GPU utilisation above 90 %.

- **Cassini Systems**
  Used by 65 dealer FCMs and more than 200 buy-side firms for UMR margin simulation. A single IM run can write 8 TB in 30 minutes. Most clients keep the run on-prem so they can re-query without extra cloud read fees.

  - Infinia's flash ingest soaks the burst; object tier stores 18 months of history at disk cost.

- **OSTTRA TradeServ (MarkitSERV replacement)**
  Processes about 100 million trade-lifecycle messages every day for 34 dealer banks and 2 000 + buy-side accounts. Regulations force a 10-year WORM lock on confirmations.

  - Infinia's object storage offers governance lock and keeps those multi-petabyte archives on sovereign soil, cutting long-term cost by more than half versus S3 Glacier with frequent recalls.

  - Every one of these platforms already talks S3. Point the config to s3://infinia/, keep the data local, and the software - and the balance sheet - run faster and cheaper.

Bringing It All Home

Five Key Take-Aways

1. **Post-trade is now data-bound**
   The lag from trade click to final cash isn't about CPUs anymore—it's about moving terabytes fast enough to feed VaR, CVaR, Greeks, and the AI that explains them.

2. **Cloud helps - but only for burst compute**
   Elastic cores are great for one-off super-stress tests; parking the cubes there full-time burns money and misses regulator RTOs.

3. **On-prem flash S3 fixes the bottleneck**
   Drop the data on a DDN Infinia bucket, keep the S3 API, and the nightly run finishes hours sooner with zero egress.

4. **AI rides the same bucket**
   Predictive, gen-AI and agentic models all read the cubes in place - no extra ETL, no license surprises, GPUs stay busy.

5. **Every FMI in the chain wins**
   Broker-dealers get faster grids, CCPs meet "two-hour restore," CSDs search exceptions in seconds, and custodians reconcile breaks before Asia opens.

Final Word

Why Infinia Matters Now

From VaR to CVaR, from Greeks to GenAI—modern post-trade risk management isn't compute-bound, it's data-bound. The firms leading the pack aren't the ones with the biggest clouds, but the ones that can move, store, and access petabytes of risk data in real time.

This is where DDN Infinia changes the game.It brings flash-speed object storage on-prem, eliminating cloud egress costs, unblocking GPU bottlenecks, and keeping your most critical data compliant, secure, and ready—no matter what the markets throw at you.

Whether you’re a broker-dealer chasing faster VaR cycles, a CCP running hourly margin drills, or a CSD under sovereign audit pressure—Infinia turns your risk stack into a real-time advantage.

Bottom line:Keep the compute wherever it makes sense—but let your critical risk data sleep on fast flash you control. Your quants, auditors, and balance sheet will all breathe easier.

Talk to us about how DDN Infinia helps you modernize post-trade, accelerate AI, and meet risk head-on—with speed, precision, and confidence.