



REPORT · 2026

THE FUTURE OF GAMING INFRASTRUCTURE

2026–2028

Operational, regulatory, and cost pressures reshaping live-service back-ends.



ADAAS FOR GAMING

OPERATIONAL CONSTRAINTS ARE FORCING A PARADIGM SHIFT IN GAMING BACK-ENDS

A ranked match starts. In the 80 milliseconds before the loading screen ends, the backend has scored skill vectors, checked toxicity history, verified economy state, and priced a session offer. Ten years ago that decision used one number. Today it uses dozens and the player notices only when it's wrong.

THE 24-MONTH HORIZON

01

From Batch to Millisecond Inference

Live-service games run on high-velocity signals (skill, toxicity, latency). ML models now require instant, millisecond-level feature extraction directly inside the active data plane.

02

Regulatory Compliance as Infrastructure

GDPR, CCPA and the EU AI Act classify matchmaking and recommendation engines as high-risk algorithmic decision-making, requiring transparency, explainability, and immutable audit logs.

03

Egress Economics & Cloud Sovereignty

Regional cloud egress pricing creates financial friction and vendor lock-in. The EU Data Act's 2027 mandate to phase out switching fees is accelerating a shift to hybrid and self-hosted architectures.

STRATEGIC INSIGHT — MATCHMAKING AS AN ENGAGEMENT LEVER

4-15%

Uplift in player retention and engagement from optimizing real-time matchmaking a capability the market is increasingly tying to low-latency backend infrastructure.



THE 2026 TRILEMMA: A STRUCTURAL BIND

Legacy fragmented architectures cache + database + search are structurally incapable of handling three concurrent demands.

AI WORKLOADS

Millisecond feature extraction requirements.

COMPLIANCE

Strict EU AI Act auditability and data residency mandates.

CLOUD COSTS

Escalating multi-region egress penalties.

THE STRUCTURAL BIND

Any one of these is solvable alone. Faster inference wants more replication; compliance wants localized data; cost control wants less movement. Pursued separately, each fix makes the other two worse.

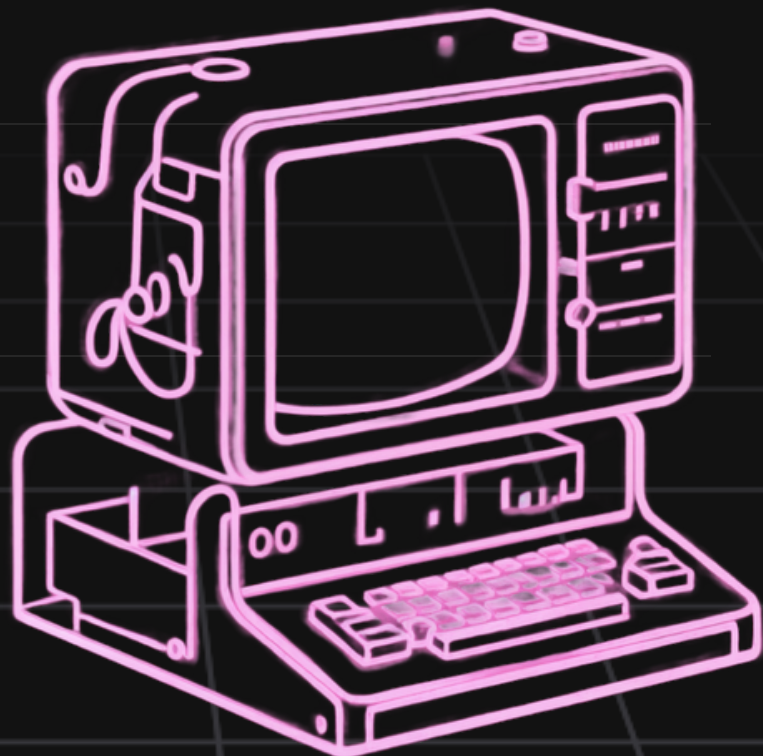
BUSINESS RISK – THE COST OF INACTION

Top-Line Vulnerability

Failing to achieve millisecond-level matchmaking risks sacrificing the 4–15% player retention uplift expected from modern live-service infrastructure.

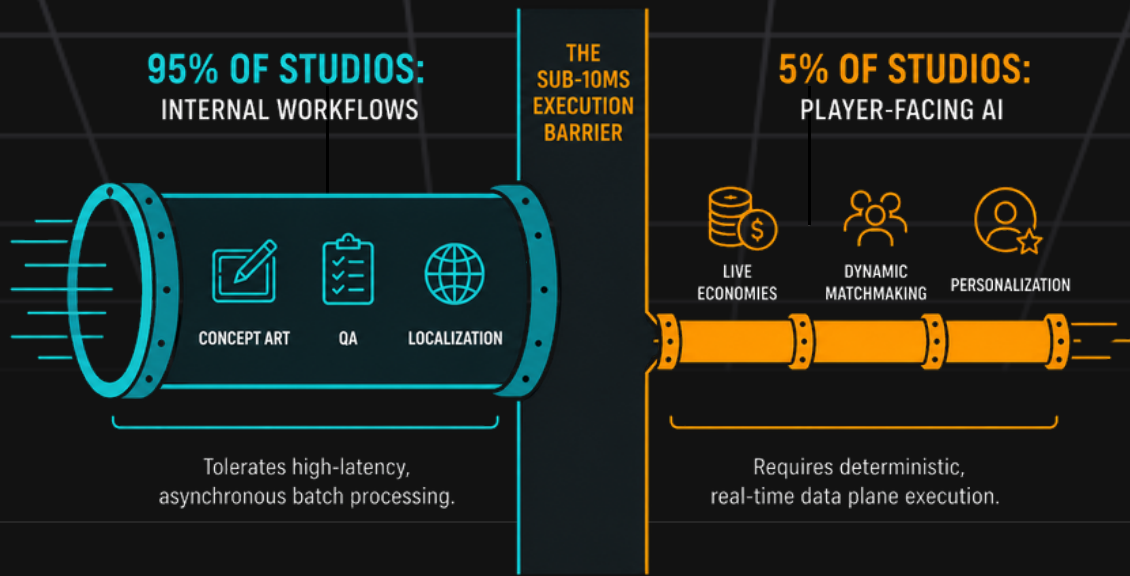
Margin Protection

Attempting to solve the trilemma with fragmented architecture compounds the systemic waste of redundant systems and escalating cloud egress fees.



THE INFRASTRUCTURE GAP

Why player-facing AI stalls at 5%



The limits of legacy data pipelines

The fact that only about 5% of studios apply AI to player-facing features is not a failure of ambition; it is a failure of architecture. Legacy stacks built on fragmented databases and high-latency network hops cannot support the deterministic, sub-10ms execution required to run predictive AI in live environments.

The next structural advantage

As content creation commoditizes, the new competitive frontier is predictive, AI-driven live-ops. The future belongs to studios capable of running these live-ops directly on the data plane



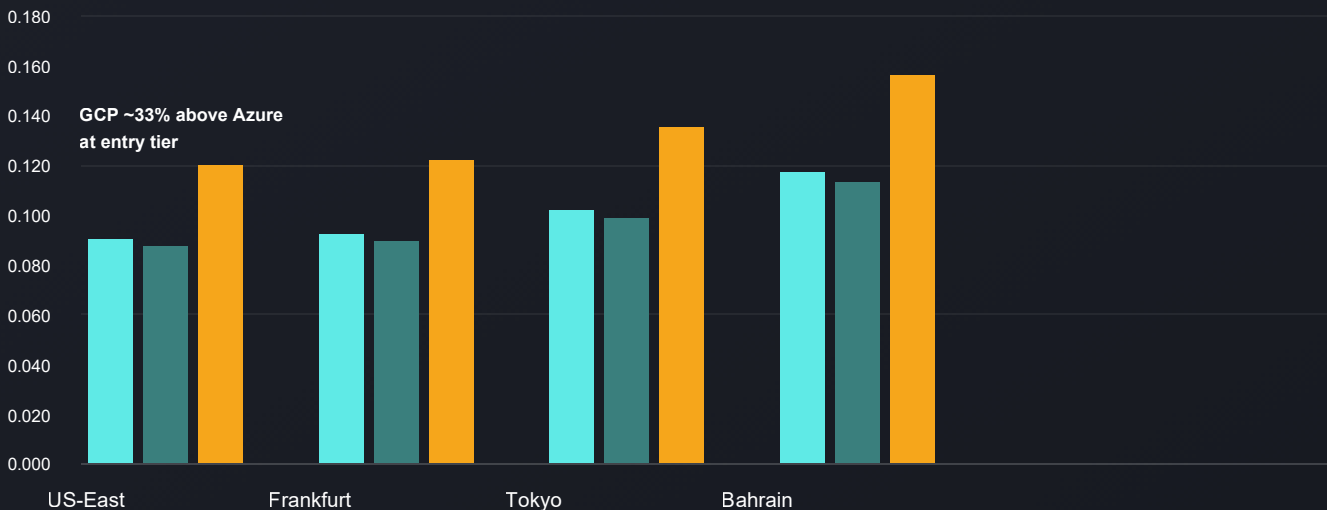
EGRESS ECONOMICS ARE RESHAPING ARCHITECTURES

Regional price divergence across AWS, Azure, and GCP creates artificial vendor lock-in, threatening live-service margins as data velocity increases.

MULTI-CLOUD EGRESS COST BY REGION (2026)

USD per GB — entry tier, first 10 TB

■ AWS ■ Azure ■ GCP



Egress as a Success Tax

For a live-service title, egress is a success tax: the better your retention, the bigger the bill. Every gigabyte of that success is billed on the way out of the cloud.

The 12 January 2027 Countdown

The EU Data Act puts a hard date on the calendar, legally mandating the removal of cloud switching barriers and egress fees.

From Chronic Cost to Acute Decision

This upcoming deadline fundamentally shifts the market. It turns a long-tolerated operational expense into an immediate architectural turning point, closing the window on deciding later.

Sources: AWS, Azure & GCP data-transfer schedules (2026); EU Data Act (Reg. 2023/2854, Ch. VI); Finout Cost Analysis Index (2025); Flexera State of the Cloud (2025).



THE SHIFT TO HIGH-VELOCITY DECISION ENGINES

Games didn't choose this velocity they inherited it from the systems they now resemble.

100,000S/SEC

VOLUME

Micro-decisions executed per second across matchmaking, anti-cheat and live-economy services.



<10MS

SPEED

Server response and model scoring resolve before the next client frame renders.

150+

CONTEXT

Concurrent inputs per decision: skill vectors, behavioral toxicity, economy state, network health.

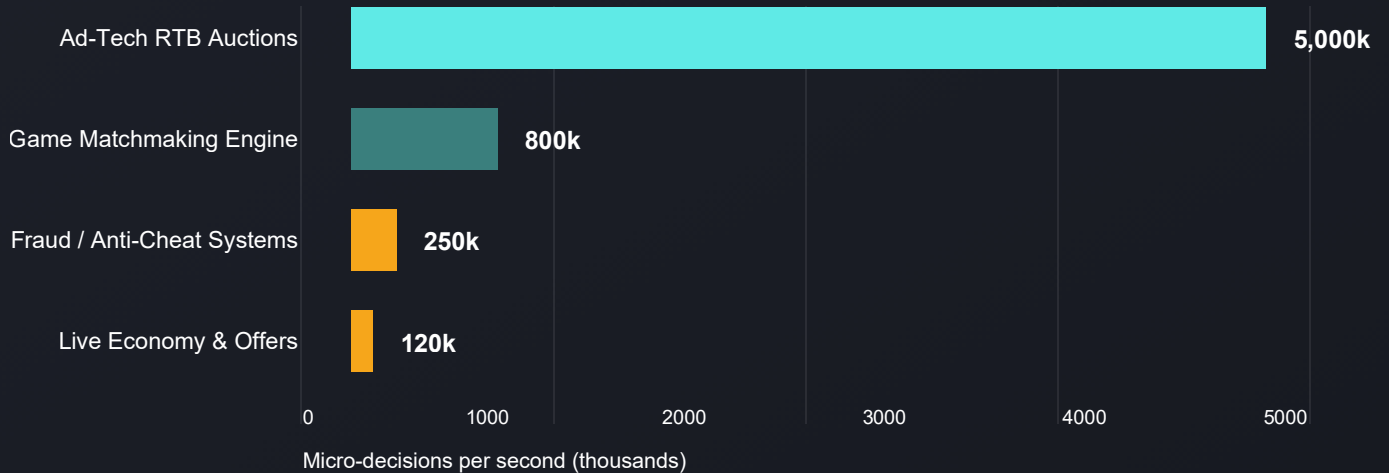
The decision window is shrinking while the decision itself gets heavier.

Sources: Redis Data Matchmaking Performance Benchmark (2026); Programmatic RTB Industry Statistics (2026); Edgegap Multiplayer Latency Orchestration Review (2026).



GAMING INFRASTRUCTURE MUST SCALE TO AD-TECH DECISION VELOCITY

DECISION THROUGHPUT: GAMES CONVERGING WITH AD-TECH (2026)



Data Vectors

Telemetry processed as fresh, multidimensional, vectorized inputs rather than static database queries.

Compute Co-location

Inference and logic physically co-located with the data plane, eliminating latency-inducing network hops.

Atomic Execution

Micro-decisions executed as atomic, low-latency transactions to prevent state corruption.

Deployment Strategy

Pivot to hybrid or self-hosted models to guarantee reliability while insulating margins from egress penalties.

Failure to co-locate compute and data results in severe latency bottlenecks; modern game engines demand deterministic infrastructure capable of resolving millions of events before the next client frame renders.



WHERE CONSOLIDATED STACKS ARE HEADING

The pattern emerging across engineering blogs and vendor roadmaps is convergence on fewer, closer layers.

AI-NATIVE WORKFLOWS

Feature Gen, Vectors, Scoring

UNIFIED REAL-TIME DATA LAYER

COMPLIANCE LAYER

GDPR, KVKK, AI Act, Audit

HYBRID DEPLOYMENT

Self-Hosted + Cloud

EVENT-DRIVEN COMPUTE

Matchmaking, Fraud, Economy, Cooldowns

AI-NATIVE WORKFLOWS

Feeds the central data plane to execute real-time feature generation, vector management, and localized model scoring.

EVENT-DRIVEN COMPUTE

Intersects with the core data layer to run matchmaking, anti-cheat/fraud detection, live-economy updates, and cooldown timers.

COMPLIANCE LAYER

Natively satisfies regional regulatory obligations including GDPR, KVKK, the EU AI Act, and localized audit requirements.

HYBRID DEPLOYMENT

Provisions the entire stack across an integrated, multi-environment infrastructure combining self-hosted and cloud resources.



ARCHITECTURAL CONSOLIDATION: REAL-TIME DATA & AI-NATIVE WORKFLOWS

The industry trend is collapsing the infrastructure split into a single, coherent data plane.

UNIFIED REAL-TIME DATA LAYER

FEATURE

PURPOSE

Sub-10ms read/write

High-speed data access

Vector search for AI pairing

Matchmaking & personalization

Multidimensional filtering

Real-time filtering across signals

Atomic actions (cash, cooldowns, fraud)

Reliability for critical ops

Strong consistency

Synchronized, accurate state

AI-NATIVE WORKFLOWS

REQUIREMENT

FUNCTION

Model scoring near the data store

Removes remote-inference delay & cost

Real-time feature materialization

Fast, high-quality data for models

Event-driven re-evaluation loops

Continuous adaptability



COMPLIANCE-FIRST, HYBRID BY DESIGN

Consolidation solves for speed and cost. Regulation asks a third question: can you explain what the system did?

COMPLIANCE-FIRST REQUIREMENTS

Region-Pinned Residency

Keeps player data within local legal jurisdictions.

Built-In Audit Logs

Secures a traceable transaction record for verification.

Model Transparency

Delivers complete algorithmic explainability per decision.

Deterministic Logic

Ensures all player segmentation remains auditable.

HYBRID DEPLOYMENT FRAMEWORK

Proximity Compute

Locates compute near local regions to minimize latency.

Self-Hosted Layers

Moves high-throughput workloads to dedicated hardware to cut cloud fees.

Strategic Elasticity

Restricts public cloud use strictly to volatile traffic spikes.

OPEX Stabilization

Replaces unpredictable cloud egress costs with flat, predictable spend.

Achieving absolute legal compliance without sacrificing live performance requires embedding governance directly into a hybrid, cost-controlled data infrastructure.



CO-LOCATING COMPUTE AND DATA

Executing logic and streaming data in the same physical locality guarantees deterministic, sub-10ms behavior by eliminating network round-trips and pipeline synchronization bottlenecks.

UNIFIED EXECUTION LAYER

Player Management	Real-time matchmaking, dynamic segmentation, cooldown timers.
Security & Risk	Automated anti-cheat heuristics and live risk scoring.
Game Economy	Instant transactional updates and algorithmic dynamic pricing.

OPERATIONAL IMPACT

Accelerated Velocity

Simplifying architecture reduces debugging and deployment complexity, driving faster feature iteration.

Cost Optimization

Eliminating network boundaries cuts computational waste and total infrastructure spend.



THE VIRTUOUS MATCHMAKING LOOP

QUALITY · PLAYER EXPERIENCE

REAL-TIME OPTIMIZATION

Live telemetry for matchmaking, personalization & on-flight economy updates.

RETENTION · CONSISTENCY & TRUST

ENGINEERING STABILITY

Deterministic state with fewer moving parts — faster iterations, safer rollouts.



**Real-Time
Data Impact**

AI DECISIONING

On-data-plane scoring & vector embeddings eliminating remote network hops.

PLAYER ENGAGEMENT

Smarter algorithmic pairing delivers a verified 4–15% lift in user retention.



WHAT'S AT STAKE

The dimensions on which the next 24 months will be won or lost.



4-15%

RETENTION UPLIFT

Sub-10ms AI matchmaking



40-60%

COST

Unified logic deployment

4-5X

VELOCITY

Faster engineering iteration

2027

AI ACT READY

Native auditability

↓ Lower Complexity

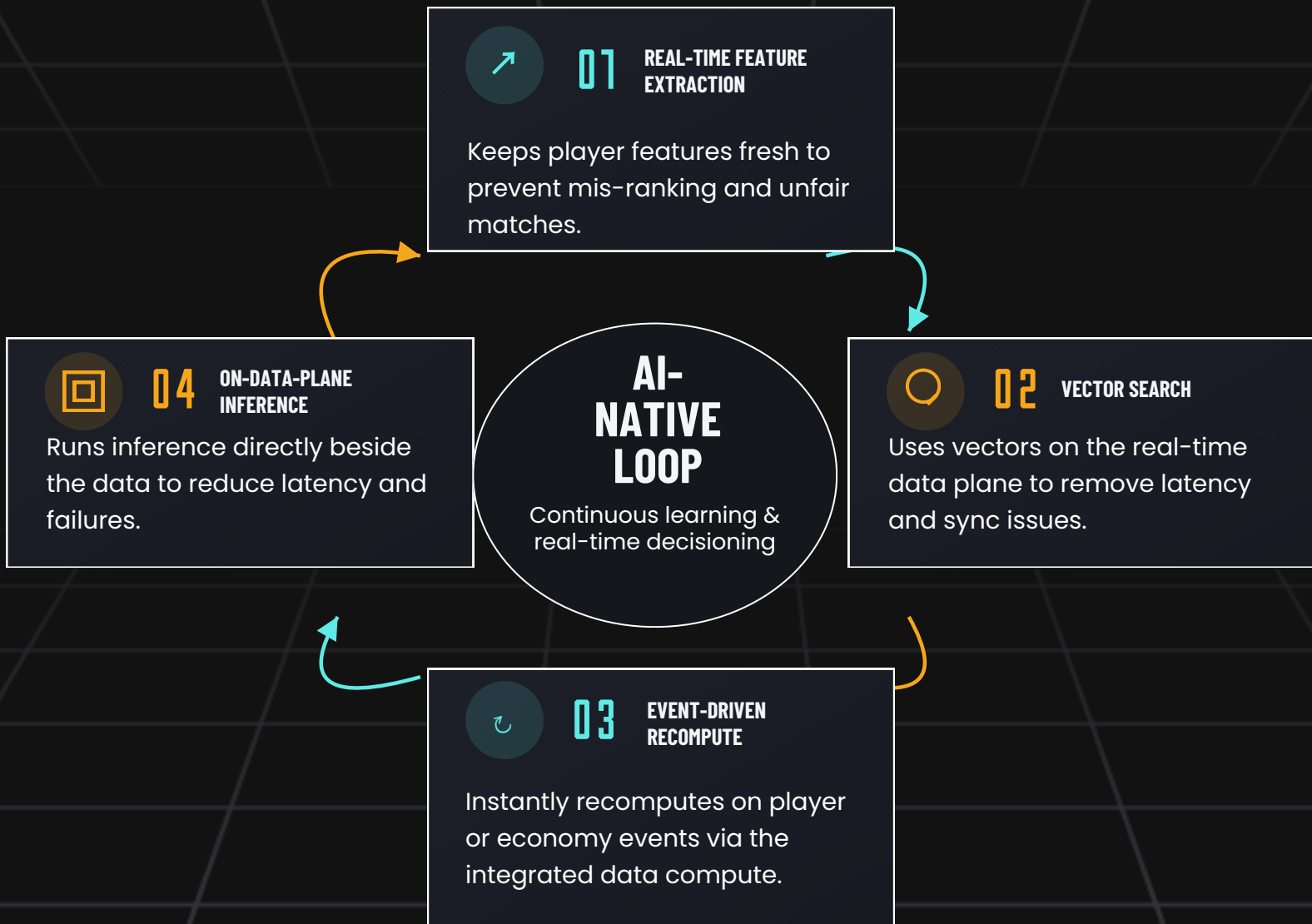
Single data-plane architecture

The studios that thrive in the 2026–2028 window will be those that actively neutralize these structural threats across top-line engagement and bottom-line efficiency.



THE MECHANICS OF THE AI-NATIVE LOOP

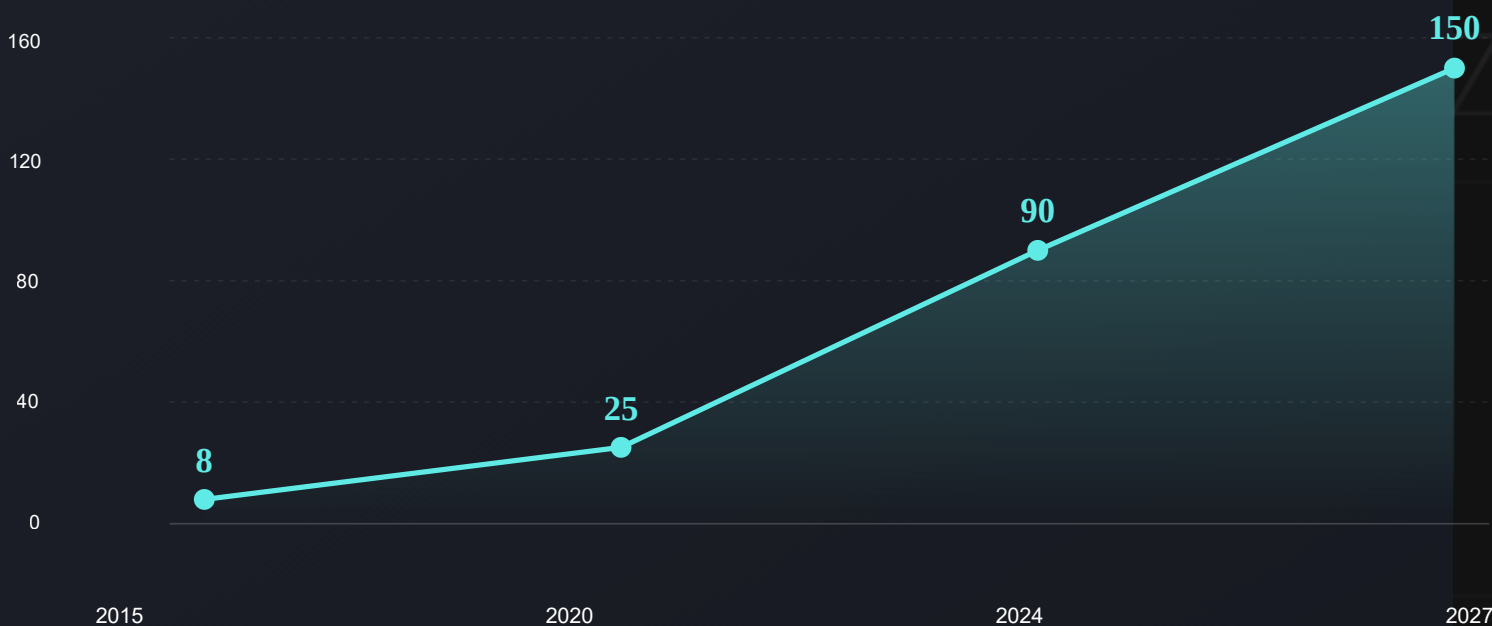
Integrating compute and data removes remote hops, eradicates sync failures, and guarantees deterministic real-time decisions.



WHY AI-NATIVE WORKFLOWS MATTER

Matchmaking is no longer about skill alone. As real-time features per decision grow, studios rebuild around AI-native workflows to keep latency low, fairness high, and players engaged.

FEATURE EXPLOSION IN MATCHMAKING (ILLUSTRATIVE)



2015–2024 based on industry benchmarks; 2027 projected.

01

MORE SIGNALS PER DECISION

Higher quality & fairness in every matchmaking outcome.

02

TIGHTER LATENCY BUDGETS

Co-location becomes a structural necessity as decision windows shrink.

03

FEWER STATE DISCREPANCIES

Better retention & stability from a single source of truth.



COMPLIANCE BECOMES INFRASTRUCTURE

EU AI Act — core obligations, current enforcement timeline

Feb 2025	Prohibited-practice bans & AI-literacy duties	IN FORCE
Aug 2025	General-purpose AI (GPAI) model obligations	IN FORCE
Aug 2026	Transparency obligations	PROVISION
Dec 2027	High-risk (Annex III) — deferred from Aug 2026	PROVISION
Aug 2028	High-risk AI embedded in regulated products	PROVISION

CORE OBLIGATIONS

- Transparency: disclose algorithmic decisions
- Explainability: documented reasons for decisions
- Traceability: full auditability of model flow
- Risk classification & governance: identify high-risk AI, monitor drift

Matchmaking and personalization are candidate “algorithmic decision-making” systems. Build explainability, audit logs, and deterministic scoring now to avoid a rushed retrofit.

Penalties up to €35M or 7% of global turnover.

Sources: EU AI Act (Reg. 2024/1689); Digital Omnibus provisional agreement, 7 May 2026.



HOW STUDIOS ARE RESPONDING: FOUR PATHS

01. Hybrid Data Co-Location

Research shows studios increasingly co-locating inference and data on bare metal to resolve latency, reserving cloud resources for traffic spikes.

02. Unified Plane Adoption

Engineering teams are actively collapsing fragmented legacy stacks (cache + database + search) into single, coherent execution layers to natively satisfy sub-10ms AI matchmaking budgets.

03. Cloud Repatriation

To neutralize escalating multi-region egress penalties, large-scale live-service titles are aggressively repatriating high-throughput workloads to dedicated infrastructure to stabilize OPEX.

04. Compliance-First Architecture

Anticipating 2027 regulations, studios are shifting toward auditable, deterministic environments to ensure algorithmic explainability.

The cost of changing architecture is falling at the same moment the cost of running it inefficiently is rising. Which path fits depends on scale, region mix, and team but the regulatory calendar has ended the option of deciding later.

