



Gaming Studio Server Infrastructure

Professional Technology Consulting

The Origami Corporation

18.09.2025

Table of Contents

1

From Crashes to Concurrency: Scaling an Indie Multiplayer Game to 50,000 CCU

1

1.1

Executive Summary

1

1.2

The Challenge

1

1.3

Our Approach

2

1.4

Implementation

2

1.5

Results & Impact

4

1.6

Technologies & Services Used

5

1.7

Conclusion

5

About The Origami Corporation

1 From Crashes to Concurrency: Scaling an Indie Multiplayer Game to 50,000 CCU

1.1 Executive Summary

An indie game studio’s breakout multiplayer title collapsed under its own success—servers crashed as soon as concurrent users exceeded 1,000, and early DDoS attacks eroded player trust. We re-architected their backend around optimized gaming servers, implemented multi-layer DDoS protection, introduced Kubernetes-based auto-scaling, and distributed content via a global CDN. With continuous monitoring deployed end-to-end, the studio now sustains 50,000 concurrent players with an average 15ms latency, shrugged off a 40Gbps DDoS attack without downtime, and improved player retention by 60%.

1.2 The Challenge

The studio’s initial launch exceeded expectations, but their infrastructure couldn’t keep up. Symptoms included:

- Server crashes as CCU crossed ~1,000, due to CPU saturation and network queue overflow on a single-region, vertically scaled host.
- No auto-scaling: matchmaking, lobbies, and game sessions shared resources, causing cascading failures during spikes.
- Limited observability: minimal telemetry made it difficult to isolate hotspots or predict failure modes.
- Asset distribution bottlenecks: patches and updates pushed directly from origin overloaded bandwidth and increased login queue times.

- Security exposure: after gaining traction, the title became a target of volumetric DDoS attacks that intermittently took services offline.

The business impact was immediate—poor first impressions, refund requests, and the risk of losing momentum in a crowded market where early reviews are decisive.

1.3 Our Approach

We delivered a holistic, production-hardened architecture centered on performance, resilience, and operational visibility:

- Gaming-optimized compute: Transitioned game sessions to high-frequency CPU gaming servers with kernel and network stack tuning for low-latency UDP traffic.
- Kubernetes-hosted control plane: Decomposed ancillary services (matchmaking, lobbies, profiles, telemetry ingestion) into microservices with horizontal pod auto-scaling and safe rollout patterns.
- Multi-layer DDoS protection: Routed traffic through globally distributed scrubbing centers with Anycast, rate-limiting, protocol validation, and behavioral detection.
- CDN acceleration: Offloaded static assets, patches, and launcher updates to a global CDN with edge caching, signed URLs, and regional storage.
- Continuous monitoring: Implemented full-stack observability—metrics, logs, traces, and real-user monitoring—aligned with explicit SLOs for latency, availability, and error budgets.

The design emphasized not only peak throughput, but also predictable performance at the tail (p95/p99), graceful degradation during adverse events, and developer velocity for frequent content updates.

1.4 Implementation

We executed the transformation in phased increments to minimize risk and maintain live service continuity.

1) Discovery and Load Modeling

- Instrumented the existing stack to capture baseline CPU, memory, network, and syscall profiles under load.
- Built a synthetic load harness that simulated session join/leave patterns, match lifecycle events, and bursty login behavior. We used k6 for HTTP APIs and a custom UDP generator for session traffic.
- Identified primary bottlenecks: thread contention in the session loop, NIC interrupt saturation, and lack of backpressure in the matchmaking pathway.

2) Migration to Optimized Gaming Servers

- Deployed dedicated gaming servers in multiple regions using high-frequency CPU instances with NVMe-backed storage for logs and temporary state.
- Tuned the network stack for real-time workloads: adjusted rmem/wmem, increased socket buffers, enabled RSS and RPS for multi-queue NICs, and pinned critical threads to isolated CPU cores to reduce context switching.
- Introduced a session-aware orchestrator (Agones on Kubernetes) to lifecycle-manage dedicated game servers, ensuring efficient placement and rapid spin-up for new matches.
- Implemented connection draining and graceful shutdown hooks to prevent player disruptions during maintenance.

3) Kubernetes Hosting for the Control Plane

- Provisioned regional Kubernetes clusters with separate node pools for CPU-intensive tasks and memory-focused services. Enabled the Cluster Autoscaler and Horizontal Pod Autoscaler (CPU, RPS, and custom metrics).
- Applied Pod Disruption Budgets, anti-affinity rules, and topology spread constraints to minimize correlated failures.
- Adopted progressive delivery: blue/green and canary rollouts with automated rollback on SLO regression using admission controllers and health probes.
- Secured service-to-service communication with mTLS, and centralized configuration via ConfigMaps/Secrets with GitOps workflows for auditable changes.

4) CDN Services for Asset Delivery

- Fronted static assets and patches with a global CDN. Implemented cache-busting via content hashing, range requests for resumable downloads, and signed URLs to prevent hotlinking.
- Replicated regionally to reduce origin hits and improved launcher responsiveness by caching manifest files at the edge.

5) DDoS Protection

- Onboarded all public endpoints—UDP and TCP—behind Anycasted DDoS protection. Layer 3/4 defenses included SYN cookies, UDP flood filters, and traffic anomaly detection; Layer 7 controls enforced protocol correctness and rate limits per token/IP.
- Established automatic traffic redirection to scrubbing centers with BGP announcements, ensuring mitigation without operator intervention.
- Configured surge buckets to absorb short-lived spikes during promotions and content drops, avoiding false positives.

6) Continuous Monitoring and Operations

- Deployed Prometheus for metrics, Grafana dashboards for SLOs (e.g., p99 matchmaking API latency, session start success rate), and log aggregation via an ELK-compatible stack.
- Instrumented services with OpenTelemetry traces to track request flows across matchmaking, inventory, and session orchestration.
- Set up synthetic checks from multiple geographies to measure player-experienced latency. Alerts routed to on-call with clear runbooks and auto-remediation for common conditions (e.g., automated node replacement on kernel panic).
- Performed chaos drills and DDoS fire drills in a controlled environment to validate failover, rate limiting, and autoscaling behavior.

7) Cutover and Validation

- Executed a staged cutover region by region, starting with lower-traffic geographies. Used shadow traffic to validate correctness before switching production flows.
- Conducted live load tests during off-peak windows to verify autoscaling thresholds, ensuring pods and nodes scaled within minutes without session loss.
- Coordinated communication with community managers to set expectations and capture real-time player feedback.

1.5 Results & Impact

- **Scale:** The game now supports 50,000 concurrent players without instability. Autoscaling brings new capacity online in minutes, matching demand patterns from regional events and influencer-driven surges.
- **Performance:** Average end-to-end latency is 15ms in primary markets, with tighter p95 tails due to proximity placement, network tuning, and edge offload of non-critical traffic.
- **Security Resilience:** The environment withstood a 40Gbps DDoS attack with no downtime, maintaining stable matchmaking and session starts while malicious traffic was scrubbed upstream.
- **Player Experience:** With crashes eliminated and queues shortened, player retention improved by 60%, translating into stronger community growth and more predictable revenue from in-game purchases.
- **Operational Confidence:** Continuous monitoring and clear SLOs provide early warning for regressions, while progressive delivery reduces the risk of content and balance updates that occur multiple times per week.

Beyond the headline metrics, the studio gained a platform that supports rapid iteration: new game modes can be rolled out safely, seasonal content drops no longer trigger outages, and the team can focus on gameplay instead of firefighting infrastructure.

1.6 Technologies & Services Used

- Gaming-Servers: High-frequency CPU instances, kernel and NIC tuning, UDP-optimized networking, dedicated session hosts orchestrated via Agones.
- Kubernetes-Hosting: Regional clusters, Cluster Autoscaler, Horizontal Pod Autoscaler, Pod Disruption Budgets, canary/blue-green deployments, mTLS, GitOps.
- DDoS-Protection: Anycast routing, global scrubbing centers, L3/L4/L7 mitigation, rate limiting, behavioral detection, BGP failover.
- CDN-Services: Edge caching for assets and patches, content hashing, signed URLs, regional replication, range requests.
- Continuous-Monitoring: Prometheus, Grafana, ELK-compatible logging, Open-Telemetry tracing, synthetic monitoring, SLO-driven alerting and runbooks.

1.7 Conclusion

By aligning infrastructure to the unique demands of real-time multiplayer, we transformed a fragile single-server deployment into a resilient, auto-scaling platform. Optimized gaming servers delivered predictable low-latency sessions, Kubernetes provided elasticity and safe delivery, the CDN offloaded heavy downloads, and multi-layer DDoS protection secured the edge. With continuous monitoring underpinning day-two operations, the studio now scales confidently to 50,000 CCU, maintains a 15ms average latency in key regions, and withstands large-scale attacks—culminating in a 60% improvement in player retention. The result is not just technical stability, but a durable competitive advantage and a foundation for future growth.

About The Origami Corporation

The Origami Corporation is a leading technology consulting firm specializing in cloud migration, DevSecOps, cybersecurity, and digital transformation. We help organizations navigate complex technological challenges and achieve their business objectives through innovative, secure, and scalable solutions.

Connect With Us

Follow Us Online

LinkedIn: [The Origami Corporation](#)

Facebook: [@theorigamicorporation](#)

X (Twitter): [@OrigamiGlobal](#)

Telegram: [@theorigamicorporation](#)

Contact Information

Europe & Americas

- **Email:** office@theorigamicorporation.com
- **Phone:** +359 878 686 627

Asia-Pacific, Middle East & Africa (APMEA)

- **Email:** office.apmea@theorigamicorporation.com
- **Phone:** +41 77 924 2245

Ready to Get Started?

SCHEDULE A MEETING

Ready to transform your technology infrastructure? Contact us today to discuss how we can help your organization achieve similar results.