

## Quarterly Infrastructure Report

This report summarizes the infrastructure changes and performance metrics for the third quarter. All systems were monitored continuously, and the data below reflects averages across the full reporting period. Key improvements were observed in response times and uptime, while storage costs continued to trend downward.

### System Overview

The following table provides a high-level summary of each production environment, including current status, uptime percentage, and the team responsible for maintenance.

System	Status	Uptime	Owner
API Gateway	Operational	99.97%	Platform Team
Auth Service	Operational	99.94%	Security Team
Data Pipeline	Degraded	98.12%	Data Engineering
CDN	Operational	99.99%	Infrastructure
Monitoring Stack	Operational	99.88%	SRE Team

The data pipeline experienced intermittent degradation during weeks three and four due to a schema migration that introduced unexpected lock contention on the primary database. The issue was resolved by switching to online DDL operations and batching the migration into smaller transactions.

### Performance Metrics

Response time measurements are taken at the load balancer level and represent the full round-trip time including backend processing, database queries, and network latency. The p50, p95, and p99 columns represent the 50th, 95th, and 99th percentile response times respectively.

Endpoint	Requests/day	p50 (ms)	p95 (ms)	p99 (ms)
/api/users	1,240,000	12	45	120
/api/orders	890,000	18	67	210
/api/products	2,100,000	8	22	55
/api/search	560,000	42	180	450
/api/auth/login	340,000	15	38	95
/api/auth/refresh	1,800,000	5	12	28
/api/webhooks	120,000	25	90	340
/api/analytics	45,000	85	320	890

The search endpoint remains the slowest in terms of tail latency. The team has identified the root cause as a full-text search query that bypasses the index when

the query contains more than four terms. A fix involving query decomposition has been developed and is currently in staging.

## Cost Analysis

Cloud infrastructure costs are tracked monthly and broken down by service category. The table below shows the per-month spend for each major category, along with the quarter-over-quarter change. Negative values indicate cost savings.

Category	Monthly Cost	QoQ Change	Notes
Compute (EC2/ECS)	\$42,300	-8%	Right-sizing completed
Storage (S3/EBS)	\$12,800	-15%	Lifecycle policies applied
Database (RDS/DynamoDB)	\$28,500	+3%	Read replica added
Networking	\$8,200	-2%	NAT gateway consolidation
Monitoring/Logging	\$5,600	+12%	Added distributed tracing
CDN (CloudFront)	\$3,400	-5%	Cache hit ratio improved
Other	\$2,100	0%	Miscellaneous services

Total monthly spend decreased from \$108,400 to \$102,900, a reduction of approximately five percent. The largest savings came from storage optimization, where automated lifecycle policies now transition infrequently accessed objects to cheaper storage tiers after thirty days. Compute savings were achieved by identifying and terminating underutilized instances that had been provisioned during a previous load testing exercise.

## Incident Summary

Three incidents were recorded during the quarter. All were resolved within the defined SLA windows. The table below provides details on each incident, including severity, duration, and root cause.

Date	Severity	Duration	Affected System	Root Cause
Jul 12	P2	47 min	Data Pipeline	Schema migration lock contention
Aug 3	P3	15 min	CDN	Certificate renewal delay

Sep 18	P1	23 min	Auth Service	Redis cluster failover timeout
--------	----	--------	--------------	--------------------------------

The P1 incident on September 18th was the most impactful, briefly affecting all authenticated users. The Redis cluster experienced a primary node failure, and the automatic failover took longer than expected because the sentinel configuration had a stale quorum setting from a previous topology change. The configuration has since been audited and corrected across all environments.

### Planned Changes

The following initiatives are planned for the next quarter. Each item has been reviewed by the architecture board and approved for implementation. Resource allocation has been confirmed with all participating teams.

Initiative	Priority	Target Date	Lead
Migrate to ARM instances	High	Oct 30	Infrastructure
Implement request tracing	High	Nov 15	SRE Team
Database read replica scaling	Medium	Nov 30	Data Engineering
CDN edge function rollout	Medium	Dec 10	Platform Team
Disaster recovery drill	High	Dec 20	All Teams

The ARM migration is expected to yield an additional ten to fifteen percent reduction in compute costs based on benchmarks conducted in staging. The migration will be performed service by service, starting with stateless workloads that have comprehensive test coverage. Each service will run in a canary configuration for at least one week before full cutover.

### Conclusion

Overall system health remains strong. Uptime targets were met for all services except the data pipeline, which has since been stabilized. Cost trends are favorable, and the planned ARM migration should further improve the cost profile. The team will continue to focus on reliability improvements and automation of operational tasks throughout the next quarter.

This report was prepared by the infrastructure team and reviewed by engineering leadership. Questions or feedback should be directed to the infrastructure channel. The next quarterly report will be published in January.