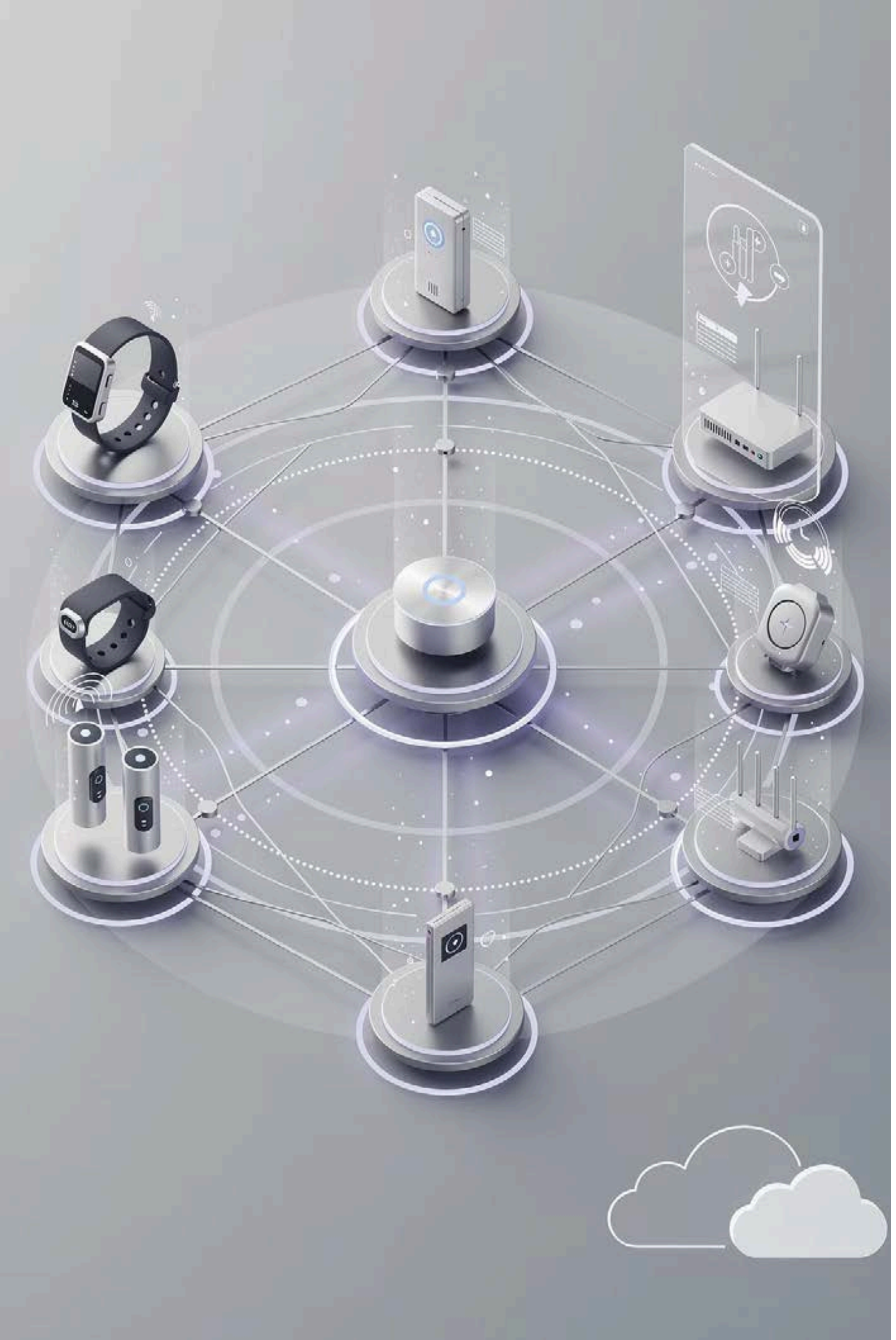


Optimizing IoT Messaging at Scale

Data-Driven Strategies for Low Latency, High Throughput, and Resilience

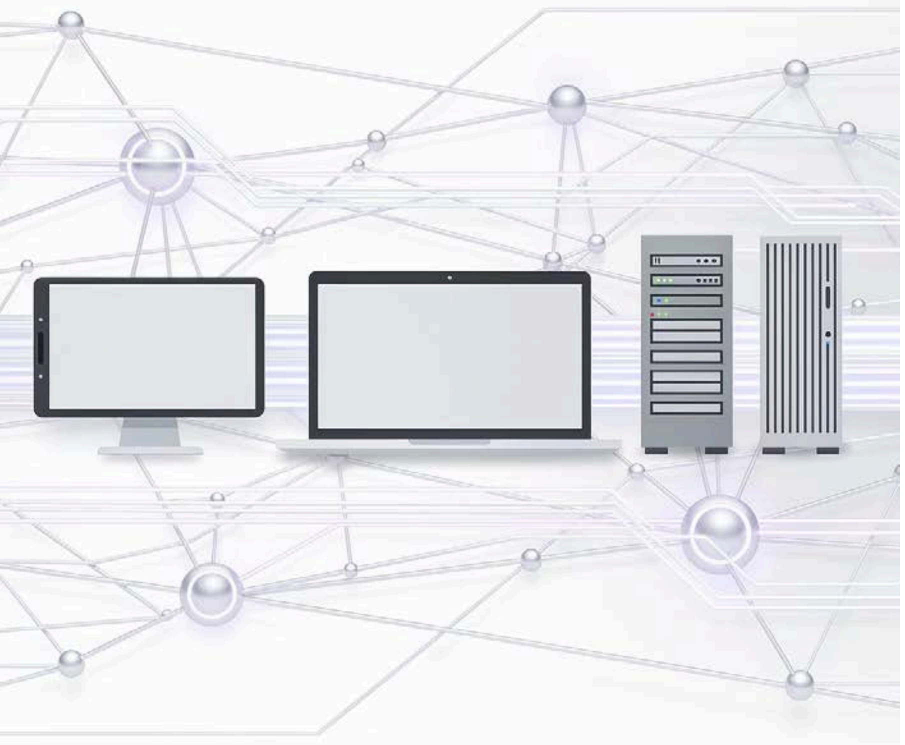
Speaker Introduction

Ketul Kishorbhai Dusane

Software Development Engineer II

Amazon Web Services

Specializing in distributed systems, real-time messaging infrastructure, and IoT platform optimization at enterprise scale.



The IoT Data Explosion

175

Zettabytes

Expected IoT data generation by 2025

<100

Milliseconds

Required latency for real-time communication

1M+

Connections

Concurrent device connections needed

Billions of devices require real-time communication infrastructure that goes beyond traditional scaling approaches.

The Challenge: Beyond Traditional Scaling

Latency Demands

Sub-100ms response times across distributed networks

Throughput Requirements

Handling millions of messages per second reliably

Operational Resilience

Maintaining uptime during traffic surges and failures

Meeting these demands requires a systematic, data-driven optimization framework rather than reactive scaling.

Three-Pillar Optimization Framework

01

Latency Reduction

Multi-tier caching, asynchronous processing, intelligent prioritization

02

Throughput Maximization

Advanced load balancing, predictive auto-scaling, queue partitioning

03

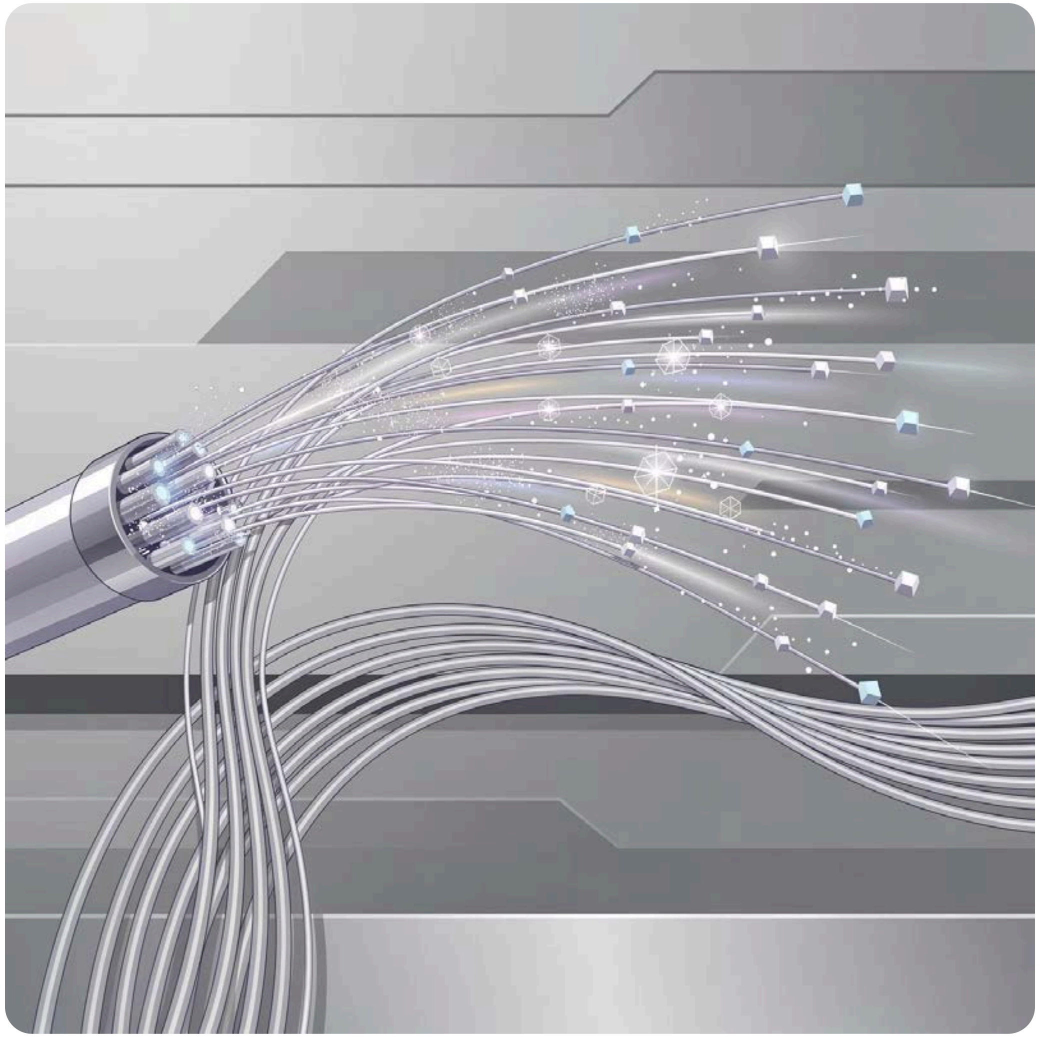
Operational Resilience

Continuous monitoring, anomaly detection, ML-driven tuning

Strategy 1: Reducing Latency

Proven Techniques

- **Multi-tier caching:** Edge, regional, and central cache layers
- **Asynchronous event-driven frameworks:** Non-blocking I/O patterns
- **Intelligent message prioritization:** Critical alerts processed first



Real-World Impact: Latency Optimization

1

Enterprise IoT Deployment

Manufacturing sensor networks reduced response times enabling predictive maintenance

2

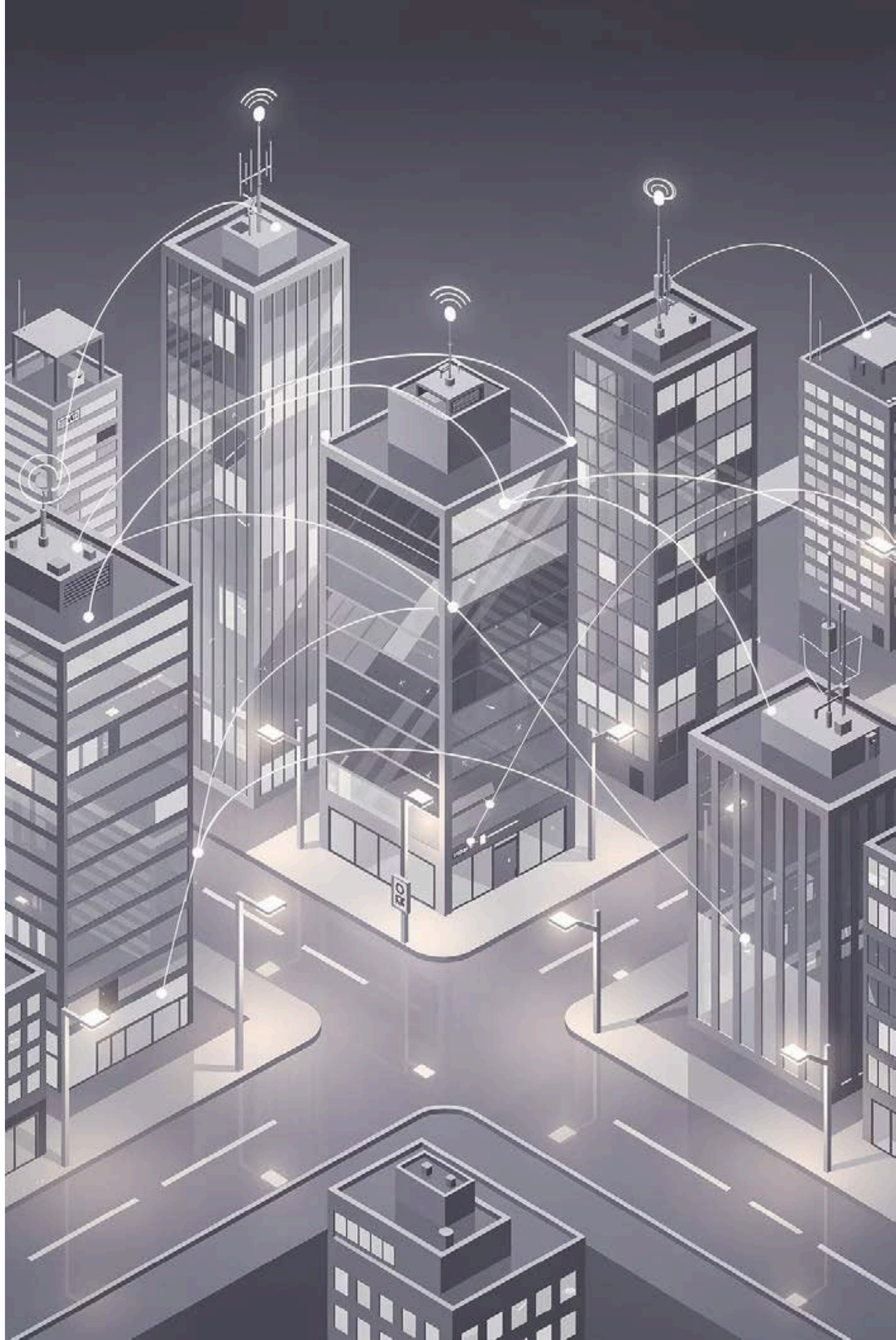
Real-Time Sensor Networks

Smart city infrastructure achieved consistent sub-millisecond processing

3

Large-Scale Chat Platforms

Message delivery optimization ensured instant communication at scale



Strategy 2: Maximizing Throughput

Advanced Load Balancing

Dynamic traffic distribution across nodes with health-aware routing

Predictive Auto-Scaling

ML models anticipate demand patterns before traffic spikes occur

Queue Partitioning

Message segregation by priority and destination for parallel processing

Handling Traffic Surges

Resilience Under Pressure

Systems withstand 3x traffic surges without downtime through intelligent architecture and predictive scaling.

- Horizontal scaling triggered before capacity limits
- Circuit breakers prevent cascading failures
- Message buffering ensures zero data loss

3x

Traffic Surge Capacity

Strategy 3: Operational Resilience

Continuous Monitoring

Real-time metrics tracking across all system components

Anomaly Detection

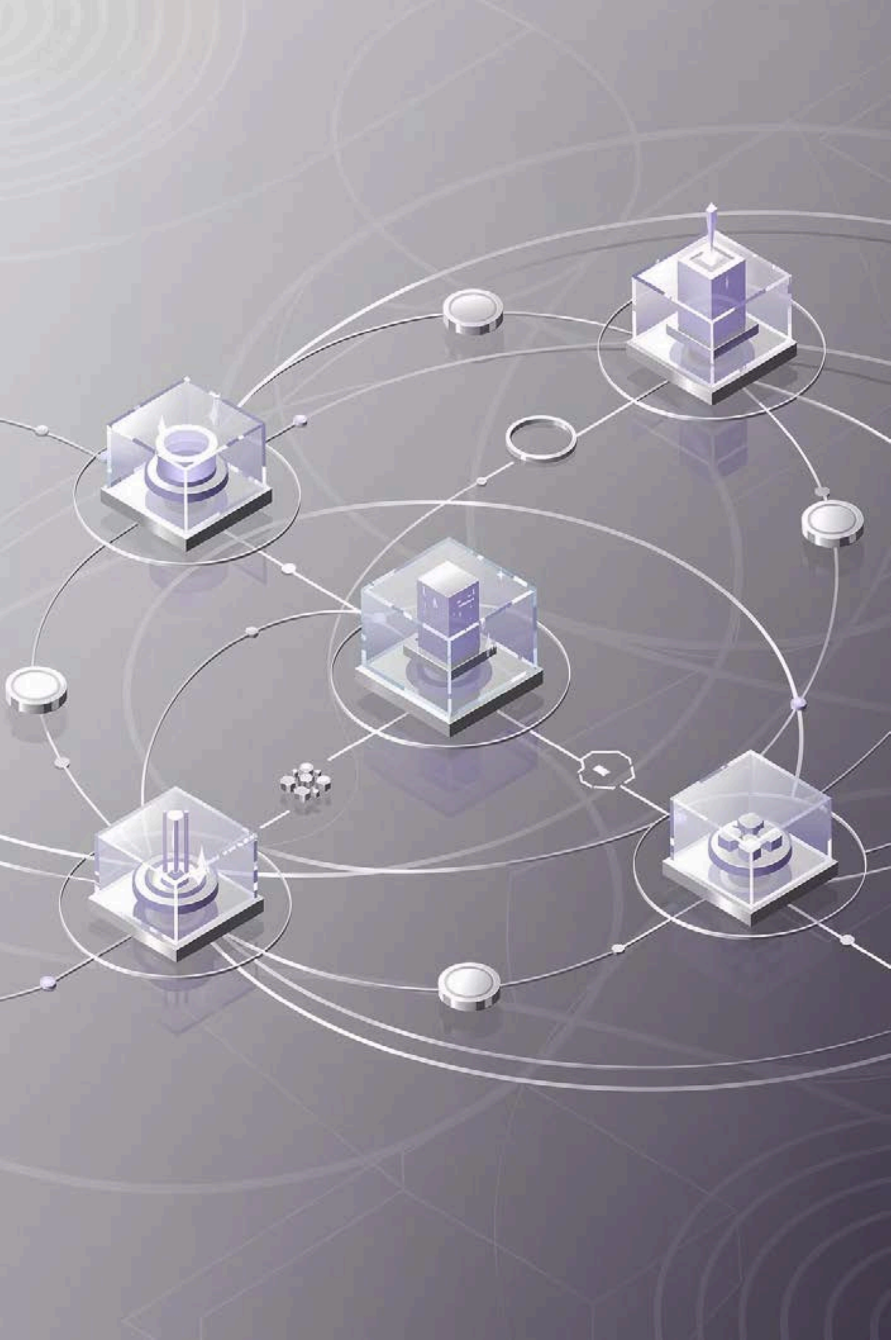
Automated pattern recognition identifies issues before impact

ML-Driven Tuning

Self-optimizing parameters based on historical performance data

Uptime Achievement

In resource-constrained environments



The Future: AI-Driven Autonomous Optimization

Self-Healing Systems

AI agents automatically detect, diagnose, and remediate performance issues without human intervention

Edge Computing Integration

Processing closer to devices for geographically distributed IoT deployments reduces latency dramatically

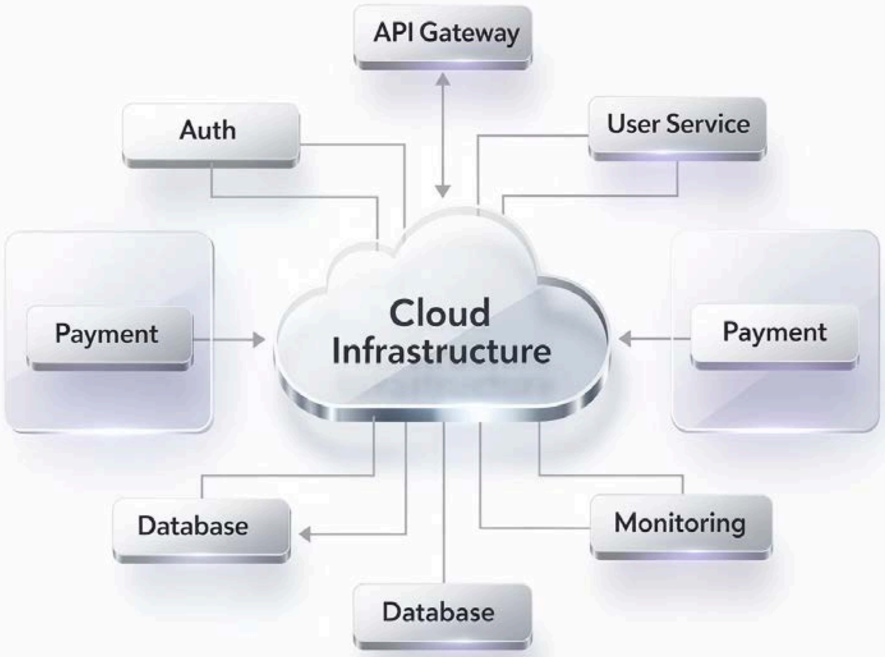
Sustainability-Focused Algorithms

Reduce energy consumption by up to 30% while maintaining reliability and performance standards

IoT-Ready Architectural Patterns

Design Principles

- **Event-driven microservices:** Loosely coupled, independently scalable components
- **Message broker clustering:** High availability through replication
- **Regional failover:** Geographic redundancy for disaster recovery
- **Protocol optimization:** MQTT, CoAP for constrained devices



Balancing Optimization and Complexity

Start Simple

Implement basic optimizations first and measure impact before adding complexity

Iterate Continuously

Performance tuning is ongoing as traffic patterns and requirements evolve

Monitor Everything

Data-driven decisions require comprehensive observability across the stack

Key Insight: The goal is robust, real-time communication that scales sustainably without over-engineering the solution.

Key Takeaways

1 **Systematic optimization frameworks deliver measurable results**

Data-driven approaches to latency, throughput, and resilience outperform reactive scaling

2 **Real-world strategies proven across enterprise deployments**

Multi-tier caching, predictive scaling, and ML-driven monitoring achieve significant improvements

3 **Future-ready architectures balance performance with sustainability**

AI-driven optimization and edge computing enable efficient, resilient IoT ecosystems

Thank You

Questions & Discussion

Ketul Kishorbhai Dusane

Software Development Engineer II, Amazon Web Services