

## Prime Video Streaming

Requirements: 50M MAU, 2M concurrent, <2s startup, DRM.

Back-of-envelope: 10 Tbps peak egress, 200 TB ingest/day.

Flow: Upload → Ingest → Transcode → Object Storage → Origin → CDN → Client.

Key focus: CDN scaling, DRM, regional resilience.

Trade-offs: Small segments reduce latency, increase overhead.

## S3-like Object Storage

Requirements: 10 trillion objects, 11 nines durability, strong consistency.

Back-of-envelope: 1 TB/s ingest, 10 TB/s egress.

Flow: Client → API → Metadata Service → Storage Nodes → Replication.

Key focus: partitioning, erasure coding, prefix hot-spot mitigation.

Trade-offs: Replication vs erasure coding.

## URL Shortener

Requirements: 200k QPS reads, 5k writes, 1B total URLs.

Back-of-envelope: 100 GB metadata, 200k QPS with 80% cache hits.

Flow: Client → API → Code Generator → KV Store → Cache → Redirect → Analytics.

Key focus: hot-key caching, abuse detection, idempotency.

Trade-offs: TTL storage vs permanent storage.

## Telemetry System

Requirements: 1T events/day, real-time + batch analytics.

Back-of-envelope: 11.6M events/s, 2.3 GB/s ingest, 200 PB/year storage.

Flow: SDK → Collectors → Kafka/Kinesis → Stream Processing → Hot Store/Cold Store → Query → Dashboard.

Key focus: partitioning, backpressure, idempotency.

Trade-offs: exactly-once vs at-least-once.

## Buy Now Workflow

Requirements: 1M orders/hr peak, zero double charges, <300ms latency.

Back-of-envelope: 280 orders/s, 840 inventory ops/s, 280 payment calls/s.

Flow: User → BuyNow → Inventory → Payment → Order → Fulfillment → Notification.

Key focus: saga pattern, idempotency keys, stock consistency.

Trade-offs: synchronous confirmation vs async fulfillment.

## **CDN (CloudFront-like)**

Requirements: 300 Tbps peak, global coverage, protect origins.

Back-of-envelope: 37.5 TB/s, ~37.5M fetches/s, 90% cache hit.

Flow: User → DNS → Edge POP → Regional Cache → Origin, with WAF/Security.

Key focus: cache hierarchy, anycast routing, DDoS defense.

Trade-offs: TTL short = more origin load, TTL long = staleness risk.