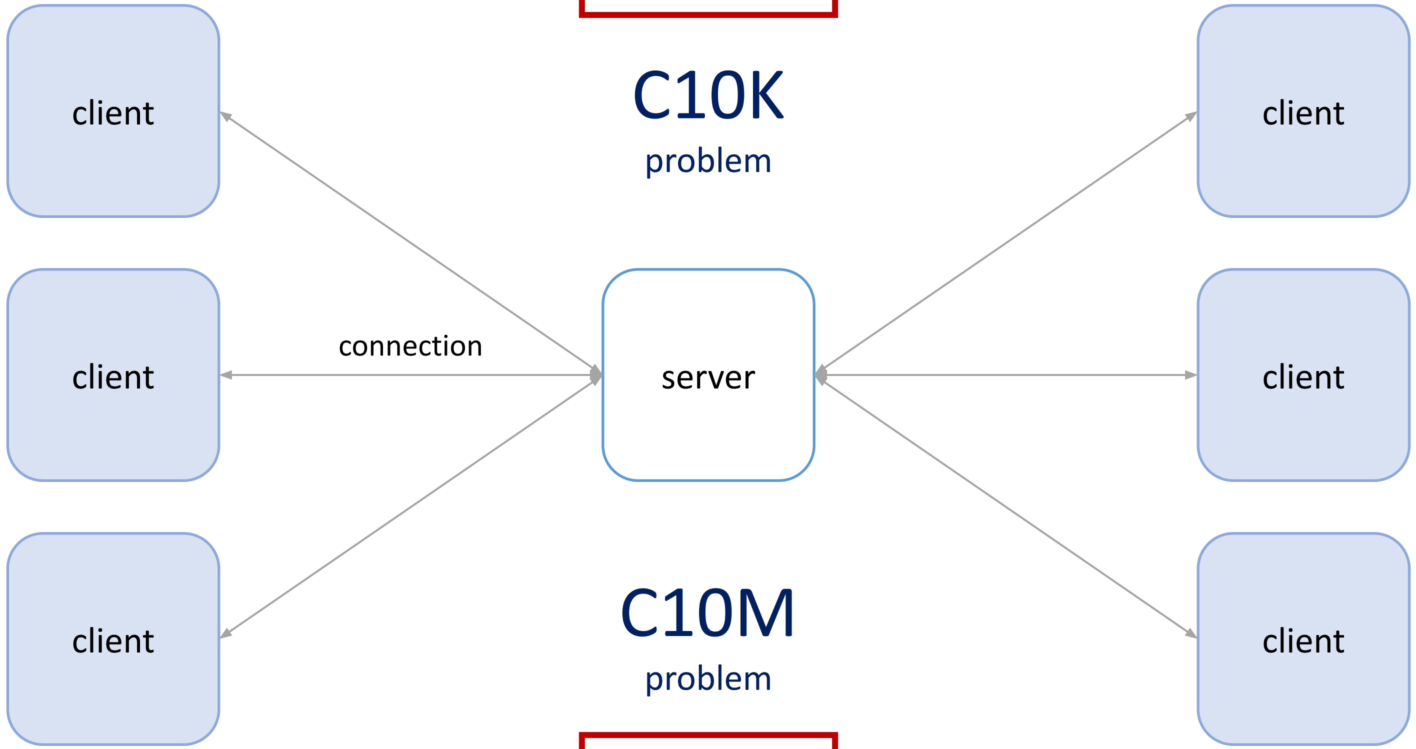
Large-scale push architectures

How many concurrent connections can one server handle?

SOLVED

C10K
problem



C10M
problem

SOLVED

handling concurrent connections ≠ handling concurrent requests

is about efficient scheduling of connections

is about speed of request processing

Large-scale push architectures

Mail.Ru

email service

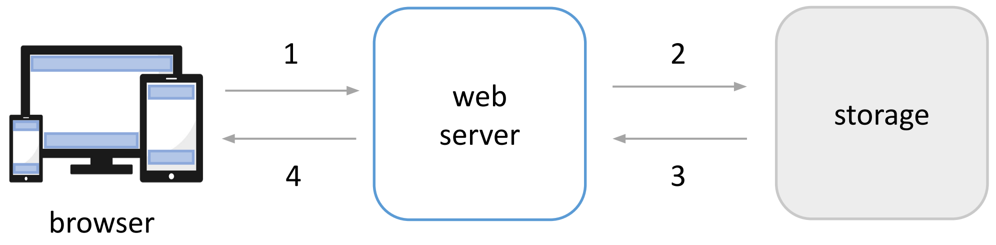
short polling

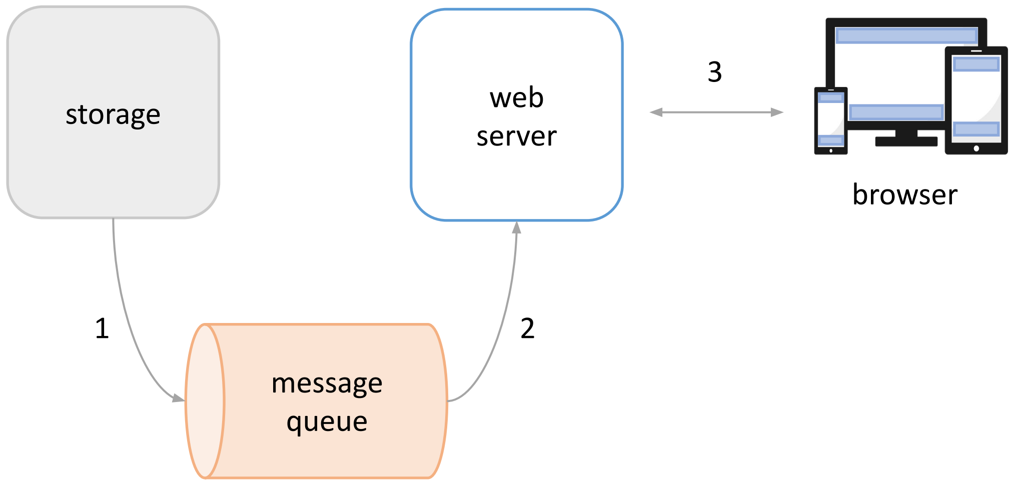
2017
50,000 requests per second
60% returns empty results

to reduce server load and
speed up mail delivery

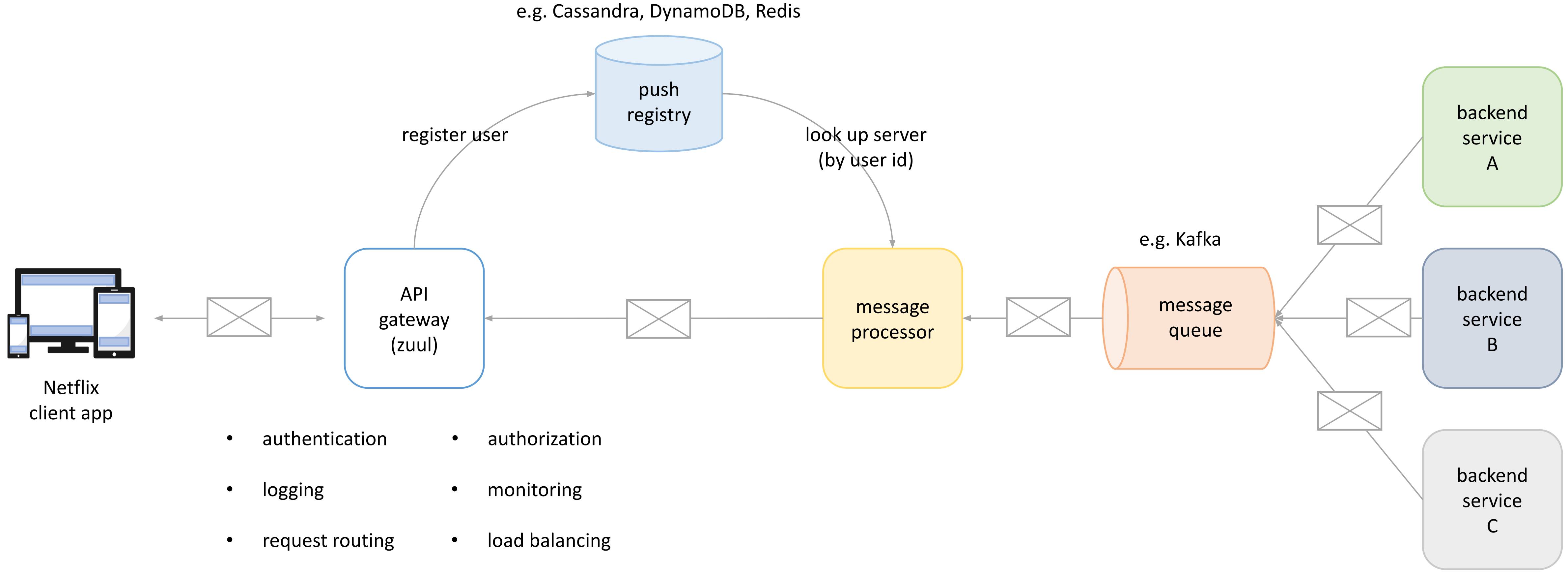
WebSocket

3,000,000 connections per server





Large-scale push architectures



- authentication
- authorization
- logging
- monitoring
- request routing
- load balancing
- rate limiting
- load shedding
- push messaging (WebSocket, SSE)

Large-scale push architectures

problem	solution
thread per connection model scales poorly for long lived connections	choose non-blocking IO servers for push architectures
server restarts	• migrate connections without reconnecting clients • force clients to reconnect to a different server
server failures	prefer multiple small servers to one big server
older load balancer versions cut WebSocket connections	• use load balancers with native support for WebSockets • balance the load at layer 4 (TCP) instead of layer 7 (HTTP)