Scaling to Millions of Users

Domain name service – converts api.mysite.com into **IP address**

Content delivery network (CDN): geographically dispersed servers to deliver static content

User

Web browser

Mobile app

DNS

CDN

Images, videos, CSS, JavaScript files, etc.

www.mysite.com

api.mysite.com

WEB APP
Server-side langs (Java, Python) for backend, storage, business logic + **client-side** langs (HTML and JavaScript) for presentation

MOBILE APP
HTTP for communication, JSON – API response format for data transfer

Load balancer

DC2

Users are geo-(DNS)-routed to closest data center (DC)

High availability, no single point of failure

DC1

Autoscale

Web servers

Decoupling more components for independent scaling

Prescribe tasks to workers: crop resize, etc.

Message Queue

Client / session data to make web servers stateless

CACHE
CACHE
CACHE

Shard 1   Shard 2   Shard ...

① Databases

Caches

Workers

② NoSQL

DB replication and sharding

Freq. used data in memory

Doing tasks prescribed in messages

Identifying errors

Logging   Metrics

Monitoring   Automation

Business insights, system health

Example: CI/CD

Tools   Figure 1-23