# Uvod

Cesto prilikom ovih system design intervjua ume da se desi I da nema tacnog odgovora nego se testira kompunikacija I problem solving skills pa vole da ih ubacuju u intervjue.

Bices ocenjen na osnovu toga koliko dobro analiziras nejasne probleme I kako resaves te probleme korak po korak.   
Takodje testira se i kako objasnjavas ideje, razgovaras sa drugima, evaluiras i optimizujes sistem.

The system design questions are open-ended.

Smer intervjua zavisi od osobe koja te intervjuise. Neki zele high level arhitekturu da se pokriju svi apsekti dok neki zele fokus na odredjenu sekciju.

Tipicno, system requirements, constraints & bottlenecks sistema koji “osmisljavas” treba dobro da se razumeju kako bi se oblikovao smer intervjuera i intervjuisanog.

## Chapter 1: Scale from zero to millions of users

Ovde cemo napraviti sistem koji podrzava jednog usera i lagano ga scaleovati na milione usera.

### Single server setup

Ovo je prvi korak.

Da bi zapoceli sa necim jednostavnim, prvomoramo imati jelte jednostavno.   
Slika ispod predstavlja single server setup gde je sve na jednom server; web app, db, cache i slicno.

A diagram of a computer system

AI-generated content may be incorrect.

Sta se desava:

1. Prvo user posalje DNS(Domain Name Server)-u upit za dobijanje IP adrese naseg servera. DNS je paid service kog daje neki 3rd party i ne hostuje se nasim serverima.
2. IP(Internet Protocol) adresa se vraca uredjaju.
3. Jednom kad je IP adresa pridobijena, salje se HTTP(Hyper Text Transfer Protocol) request direktno na tvoj server
4. Web server vraca HTML stranicu/JSON response za renderovanje

Pogledajmo traffic source. To su web application & mobile application.

* Web Application koristi kombinaciju server side jezika da resava “biznis logiku”, skladistenje i client side jezika(HTML & JS) za prezentaciju.
* Mobile application: HTTP protocol je komunikacioni protokol izmedju mobilne aplikacije i web servera. JSON je cesto koristeni API response format.

### Database

Recimo rastu ti korisnici. Jedan server nije dovoljan i sad nam treba vise servera. Jedan za web/mobile traffic, a drugi za bazu. Razdvajanjem web/mobile traffic(web tier) & baze(data tier), serveri sada mogu nezavisno da skaliraju.

A diagram of a computer network

AI-generated content may be incorrect.

### Koju bazu koristiti

Relaciona & nerelaciona.

Relacione znas vec sta su.   
Nerelacione(NoSQL)(CouchDB, Neo4j, Cassandra, HBase, Amazon DynamoDB) se dele u x4 kategorije:

1. Key-value stores
2. Graph stores
3. Column stores
4. Document Stores

Join operacije ovde cesto nisu podrzane.

Razlozi da se odlucis za nerelacione baze:  
- Treba ti super low latency  
- Podaci su ti nestruktuisani ili nemas “relacione” podatke  
- Trebas samo da serializujes i deserializujes podatke(JSON/XAML/YAML etc..)  
- Trebas da skladistis velike(massive) kolicine podataka

### Vertical vs Horizontal scale

Vertical scaling(scale up) znaci dodavanje snage procesorom/ramo memorijom na serveru.

Horizontalno(scale out) znaci dodavanje servera u tvoj pool resursa.

Kad je saobracaj nizak, vertikalno skaliranje je skroz okej i jednostavnost ove opcije je njena glavna prednost. Nazalost, ima ogranicenja:  
- Ima hard limit. Nemoguce je dodati “bezbrojno” CPU-a i RAM-a na single server  
- Vertical scale nema failover i redundanost. Ako jedan server pukne, puca sve.

Horizontalno skaliranje je pozeljnije za vece aplikacije upravo zbog ovih ogranicenja.

Znaci u gore primeru su se korisnici kacili na x1 server. Ako je taj server nedostupan, ne mozes mu pristupiti. Ako se vise korisnika prikaci na server istovremeno i dostignes web serverov load limit, korisnici iskuse generalno sporiji response ili ne mogu da se povezu na server.

Za taj problem, resenje je **LOAD BALANCER.**

### Load balancer

Load balancer podjednako rasporedjuje dolazeci saobracaj izmedju web servera koji su definisani u “load-balanced set”.

A diagram of a network

AI-generated content may be incorrect.

U gore slici je definisano kako load balancer izgleda.

Korisnici se povezuju na public IP load balancera direktno. Na taj nacin, web serverima se vise ne pristupa direktno od strane klijenata. Za bolji security, private IP’s se koriste za komunikaciju izmedju servera.

Private IP je IP adresa kojoj mogu pristupati samo serveri unutar iste mreze, ali ne i van te mreze(e.g. internet).

Posle load balancera i dodavanja drugog servera, sada smo resili problem failovera i poboljsali dostupnost **web tier**-a.

* Ako server 1 ode offline, sav saobracaj usmerice se na server 2. Na taj nacin website nece biti offline. Takodje dodacemo novi zdravi server u server pool da bi se balansiralo opterecenje
* Ako website saobracaj naglo poraste i dva servera nisu dovoljna da izdrze taj saobracaj, load balancer moze ovaj problem lako da sredi dodavanjem novih servera u web server pool kojim ce LB kasnije balansirati.

To je nas **web tier**. Sta je sa **data tier**? Trenutni dizajn ima samo x1 database bez failovera i redundantnosti. Database replication je cesta tehnika za resavanje tih problema.

### Database replication

Database replication se cesto koristi u master/slave odnosu izmedju originala(master) i kopija(slaves) baza.

Master baza cesto podrzava samo write/update/delete operacije. Slave baza dobija kopije podataka od master baze i podrzava samo read operacije. Sve data modifying komande poput insert, delete ili update moraju biti poslate na master bazu.

Vecina aplikacija zahteva mnogo veci odnos u reads naspram writes pa je cesto i broj slave baza u sistemu bezi od broja master baza.

Na slici ispod moze se videti master baza naspram slave baza.

A diagram of a computer server

AI-generated content may be incorrect.

Prednosti database replicationa:

* **Bolje performanse**. read operacije se distribuiraju across slave nodes jer sada mozes vise upita procesirati paralelno.
* **Reliability**(pouzdanost). Ako jedan db server padne, ne moras brinuti o data loss-u jer su podaci replicirani duz sistema.
* **Visoka dostupnost**. Repliciranjem podataka kroz vise ralzicitih lokacija, website ti moze biti dostupan cak iako jedna baza padne.

U predhonom primeri imali smo load balanser koji je pomagao da se preusmeri saobracaj. Kako to ovde?

* Ako je samo jedna slave baza dostupna i ona ode offline, read operacije ce se privremeno preusmeriti ka master bazi. Cim se problem pronadje, nova slave baza ce zameniti staru. U slucaju da imamo vise slave baza, one ce uzeti taj saobracaj.
* Ako master baza ode offline, slave baza ce biti promovisana kao nova master baza. Sve db opecaije se privremeno biti izvrsene na novoj master bazi. Nova slave baza ce zamniti staru za data replication. U production systems, promotion je kompleksniji jer novi master mozda nema sve podatke pa je neophodno uz recovery scripts podici bazu da bude up to date. Postoje multi-master & circular replication metode koje mogu pomoci, ali to je out of scope za ovu knjigu.

A diagram of a computer network

AI-generated content may be incorrect.

Arhitektura nakon dodavanja replikacije u nas sistem.

Ukratno, jasno ti je sta se ovde desava. Sad kad se malo bolje razumemo u web/data tiers, mozemo da predjeno na sledecu tacku, a to je poboljsanje load & response times. Dodavanjem cache layera i shifting static contenta(js/css/video files) u CDN(Content Delivery Network) ce nam tu mnogo pomoci.

### Cache

Cache je privremeni storage area koja cuva rezultate “skupih” upita ili cesto pristupanim podacima u memoriji tako da se requestovi brze opsluzuju. U gore slici, svaki put kad se nova web stranica ucitava, neophodno je odraditi jedan ili vise db poziva. Performanse aplikacije su itekako zahvacene ovim.

### Cache tier

Cache tier je privremeni data store layer mnogo brzi od baze.   
Benefiti:  
- rasterecenija baza  
- bolje performanse sistema  
- mogucnost skaliranja cache tiera nezavisno

A diagram of a computer

AI-generated content may be incorrect.

Nakon dobijanja requesta, web server prvo proverava ako cache ima dostupan resposne. Ako ima, onda se vrati. Ako nema, onda se query upucuje bazi, rezultat se memorise u cache i uzvraca se serveru.

Ta napomenuta **caching strategy** se naziva **read-through cache.** Druge strategije su dostupne u zavisnosti od tipa podataka, velicine i obrasca pristupanja(access patterns).

Vecina cache servera nude API za vecinu programskih jezika pa je i interakcija sa istima poprilicno jednostavna.

Primer za **memcached:**  
A close up of a text

AI-generated content may be incorrect.

### Considerations for using cache

* **Odluci o tome kada zelis da koristis cache**.  
  Valja za scenario u kom imas puno citanja, a malo modifikacija. Cache nije idealan za cuvanje podataka jer je “volatile memory” u pitanju. Ako se cache server restartuje, gubis sve podatke.
* **Expiration policy.** Dobra je praksa imati expiration policy kojom jednom kada cachirani podaci isteknu, ciste se iz cachea
* **Consistency.** Ovo ukljucuje sinhronizaciju data stores-a i cachea. Inconsistency se desava kada data modifying oepracije na data store & cache nisu u single transaction. Skaliranjem preko “vise regiona” je izazovno. Facebook ima “Scaling memcache at facebook” document.
* **Mitigating failures.** Cache server i dalje moze da pukne i da bude potencijalni single point of failure(SPOF). SPOF znaci da je to u pitanju deo sistema koji kada pukne, moze da znaci da ceo sistem puca. Visestruki cache serveri u razlicitim data centrima su recommended. Drugi recommendation je da pogledas koji ti je required memory i uvek imas buffer iznad za odredjeni %. A diagram of a router

  AI-generated content may be incorrect.
* **Eviction policy.** Jednom kad je cache full, zahtev da se dodaju itemi u cache mogu rezultovati da se postojeci itemi obrisu. To se naziva **cache eviction**. **Least recently used(LRU) je najcesci eviction policy**. Postoje **i Least Frequently Used(LFU)** ili **First in First Out(FIFO).**

### CDN - Content Delivery Network

CDN je mreza globalno distribuiranih servera koji se koriste da dostave staticki sadrzaj. CDN serveri uglavnom taj sadrzaj cacheuju(slike, video, js, css, html pages etc.)

**Dynamic content caching** je relativno nov koncept i van opsega knjige. Uglavnom ideja je da se na osnovu request path, query strings, cooking i request headera omogucava caching HTML stranica.

Ovo je kako CDN funkcionise na high level. Kad korisnik pristupa websiteu, cdn server najblizi korisniku ce dostaviti staticki sadrzaj. Intitivno, sto su dalji korisnici od CDN-a, to se sajt sporije ucitava. NPR CDN serveri koji su u San Francisko, a korisnici u LA ce dobiti sadrzaj brze od korisnika u EU.

A diagram of a computer network

AI-generated content may be incorrect.Slika 1-9 koja pokazuje cdn na high level.

A diagram of a cloud computing system

AI-generated content may be incorrect.

Slika 1-10 CDN workflow.

U sustini slicno kao cache. Korisnik posalje zahtev za neku sliku, cdn prihvati zahtev i ukoliko nema sliku u cache-u, onda se prosledjuje zahtev serveru. Rezultat servera(slika) se skladisti u CDN i vraca korisniku. Svaki sledeci korisnik koji posalje zahtev za tu sliku ce je dobiti direktno od CDN-a bez da komunicira sa serverom.

Da se razumemo, taj “server”, moze da bude i online storage poput amazon S3. U pitanju je origin jelte. U trenutku kada “server” vraca sliku CDN-u, on moze opciono postaviti TTL HTTP header koji naznacava koliko dugo ce se slika cacheovati.

### Considerations of using a CDN

**Cost**: CDN’s su u vlasnistvu 3rd party-a i placas data transfers “in and out” of cdn.  
Caching infrequently used assets ti ne daje neki benefit ako koristis CDN.

**Setting an appropriate cache expiry**: Za time sensitive content cache expiry je bitan i treba da bude ne previse dugacak(outdated) niti previse kratak(repeat reload)

**CDN Fallback**: Testiraj kako ti se app ponasa ako CDN padne. Ukoliko se to desi, clienti bi trebalo da mogu da detectuju problem i da se zahtevi preusmere direktno u origin za dobijanje resursa.

**Invalidating files:** Mozes ocistiti file iz CDN pre nego istekne sledecim operacijama:  
- Invalidate using CDN provider API  
- Use object versioning to serve a different version of the object. Za verzionisanje mozes dodati version query parameter poput **image.png?v=2**

A diagram of a computer network

AI-generated content may be incorrect.

Ovo je arhitektura nakon sto dodamo CDN & Cache.

1. Static assets se sada ucitavaju od strane CDN-a
2. DB load je rasterecen koriscenjem cache data

### Statless web tier

Sada cemo da skaliramo web-tier horizontalno. Da bi to uradili, prvo moramo pomeriti “state” van web tiera. Pod “state” mislimo na podatke poput session data. Dobra praksa je cuvanje session data unutar persistent storage poput relacionih baza ili nosql.

Svaki web server u clusteru moze pristupiti tim podacima iz baza. To se naziva stateless web tier.

### Stateful architecture

Stateful server & stateless server imaju razlike.

Stateful “pamti” dok stateless ne pamti client podatke izmedju requestova(from one request to the next).

A diagram of a computer server

AI-generated content may be incorrect.

Primer stateful arhitekture.

Server 1 ce npr pamtiti session data od usera A itd. Da bi se autentikovao user A, http client mora biti preusmeren na server 1 jer ako je usmeren u server 2, onda ce failovati jer server 2 nema info o user A.

To se moze resiti uz “sticky sessions” u vecini load balancera, ali to dodaje neki overhead i teze se handleuje failure pa je lakse jednostavno sve persistovati u data tier/storage umesto na serveru.

### Stateless architecture

A diagram of a computer server

AI-generated content may be incorrect.

U statelessu, zahtev moze da se salje na bilo koji server jer se podaci o sesijama cuvaju na shared storage van web servera.

A diagram of a computer server

AI-generated content may be incorrect.

Ovom arhitekturom smo otklonili state iz servera i dodali smo persistent data store. U ovom slucaju **NoSQL** data store je odabran za tako nesto.

Autoscaling znaci dodavanje ili otklanjanje web servera automatski na osnovu opterecenja u saobracaju.

**Sada imas znacajan broj korisnika globalno. Da bi poboljsao dostupnost i user experience duz sire geografske zone, neophodno je imati vise data centara.**

### Data centers

A diagram of a computer network

AI-generated content may be incorrect.

Ovde je prikazan primer arhitekture sa vise data centara. Jedan DC1 u US East & DC2 u US West. Korisnici su geoDNS-routed najblizem centru sa traffic split od **x%**

**geoDNS** je DNS servis koji dozvoljava da se domain names resolveuju na IP adrese na osnovu lokacije korisnika.

U slucaju da data centar postane nedostupan, sav saobracaj ce se usmeriti ka zdravo data centru.

Kako bi uspostavili multi-data center setup, postoji nekoliko tehnickih izazova koji moraju da se razrese.

**Traffic redirection:** Neophodni su efektivni alati kako bi se usmerio saobracaj ka ispravnom data centru. GeoDNS je npr jedan od njih.

**Data synchronization:** Korisnici iz razlicitih regiona mogu koristiti razlicite lokalne baze ili cache. U failover cases, saobracaj bi se routeovao u data centar gde su ti podaci nedostupni. Cesta strategija je replikacija duz vise data centara.

**Test and deployment:** Sa multi-data center postavkom, bitno je testirati website/application na razlicitim lokacijama. Koristi automated deployment tools za postizanje consistency-a.

Za dalje skaliranje, neophodno je razdovjiti sledece komponente.

Messaging queue je kljucna strategija koju koriste skoro svi distribuirani sistemi.

### Message queue

Message queue je durable component stored in memory koja podrzava asinhronu komunikaciju. Sluzi kao buffer i distribuira asinhrone zahteve.

Osnovna arhitektura message queue-a je takva da imas input services(producers/publishers) koji kreiraju poruke, publishuju ih u queue i potom drugi servisi(consumers/subscribers) se povezuju na queue, citaju poruke i vrse “nesto” na osnovu tih poruka.

A blue and white line of email icons

AI-generated content may be incorrect.

Producer moze da publishuje poruke i kad consumer nije dostupan.  
Consumer moze da slusa poruke i kada producer nije dostupan.

Scenario:  
Aplikacija koja nudi photo customization. Serveri pushuju customization taskove u message queue koje kasnije photo processing workeri konzumiraju. Producer & consumer se mogu nezavisno skalirati. Ukoliko ima previse poruka, workeri se skaliraju. Ukoliko nema poruka, smanjuje se broj workera.

### Logging, metrics, automation

Logging, metrics & automation support su generalno dobre prakse, ali kada radimo sa malim sajtom koji radi na par servara i nisu toliko nephodne.

Sad zamisli da imas veliki “biznis”. Onda postaju krucijalni.

**Logging**: Monitoring error logova zarad identifikovanja gresaka i problema u sistemu.

**Metrics:** Skupljanje razlicitih metrika nam dozvoljava da dobijemo povratne informacije o poslovanju i razumemo “health status” sistema. **Host level metrics**(cpu health, mem usage), **Key business metrics**(daily active users, retention, revenue) & **Aggregated level maetrics**(npr performanse celog data tiera/cache tiera i slicno)

**Automation:** Kada sistem postane velik i kompleksan, neophodna nam je automacija da bi poboljsali produktivnost. CI je dobra praksa u kojoj je svaki code check-in verifikovan i problemi se ranije detektuju. Automatizacija build, test & deploy procesa itekako povecava produktivnost.

### Adding messages to queues & different tools

Slika ispod predstavlja updateovani dizajn(bice prikazan samo x1 data center jer jbg prostor)

1. Ukljucen je message queue kako bi se sistem bolje razdvojio i postao otporniji na failures
2. Logging, monitoring & automation tools su ukljuceni

A diagram of a computer network

AI-generated content may be incorrect.

I sada kako podaci rastu, baza je pod vise opterecenja - vreme je da se skalira data tier.

### Database scaling

Isto kao i sa web tier, mozemo horizontalno i vertikalno skalirati baze.

### Vertical scaling

Ya know. Amazonov RDS(Relational database server) nudi servere sa 24 TB RAM memorije sto je jako kul. Stackoverflow na primer uspeo je da ima 10 miliona mesecnih unique posetioca i to je sve sa x1 master bazon.

Mane su standardne, povecana cena, smanjena redundantnost(ako ti padne, pada sve) i limiti hardvera. Single instance can go up only so far.

### Horizontal scaling(Sharding)

Takodje poznato kao sharding je kada dodajes vise servera horizontalno. Shardingom se velike baze razdvajaju u manje, jednostavnije delove koji se nazivaju “shards”. Svaki shard deli istu schemu, ali podaci izmedju shardova se razlikuju. Podaci na svakom shardu su unique za shard.

Na slici dole mozemo videti primer shardovne baze. User data se alocira u db server na osnovu user ID-a. Svaki put kad pristupas podatku, hash funkcija se koristi da se pronadje odgovarajuci shard. U nasem primeru, user\_id % 4 se koristi kao hash funkcija.

Ako je rezultat 0, koristis shard 0 da storeujes i fetchujes podatke za tog usera. Itd.

A diagram of a diagram of a diagram

AI-generated content may be incorrect.

A screenshot of a table

AI-generated content may be incorrect.

Najbitniji faktor koji treba razmotriti prilikom implementiranja sharding strategije jeste odabir sharding key-a.

**Sharding key(partition key)** se sastoji iz jedne ili vise kolona koje odlucuju o tome kako se podaci distribuiraju. Na primer na slici gore, “user\_id” je sharding key.

Sharding key ti dozvoljava da pribavljas i modifikujes podatke efikasno rutiranjem db querya ka ispravnoj bazi.

Pri biranju sharding key-a, jedan od najbitnijih kriterijuma koje treba da odaberes jeste da odaberes kljuc koji moze podjednako distribuirati podatke. Sharding je dobra tehnika za skaliranje baza, ali daleko je od savrsene.

Postoje kompleksnosti poput:

* **Resharding data**: neophodno je kada single shard ne moze vise podataka da drzi u sebi zbog naglog rasta ili kada odredjeni shardovi mogu iskusiti “shard exhaustion” brze od drugih zbog neravnomerne distribucije podataka. Kada se ovo dogodi, razmotri u update sharding funkcije i pomeranju podataka. **Consistent hashing** je tehnika koja se cesto koristi za resavanje ovog problema.
* **Celebrity problem:** Takodje poznat kao **hotspot key problem**. Prekomeran pristup odredjenom shardu moze prouzrokovati server overload. Zamisli da su justin bieber, katy perry & lady gaga na istom shardu, a ti imas neki social app. Resavanje ovog problema znaci dodavanja posebnog sharda toj poznatoj osobi, ali svaki shard mozda bude zahtevao i dodatno particionisanje.
* **Join & de-normalization**: Jednom kad se baza sharduje across multiple server, teze je izvrsiti join operacije izmedju vise razlicitih db shardova. Cest workaround jeste denormalizacija baza kako bi se queryji mogli izvrsiti unutar jedne tabele.

Na sledecoj slici shardujemo baze da bi podrzali rapidly increasing data traffic. U isto vreme, neke od nerelacionih funkcionalnosti su pomerene u NoSQL data store kako bi smanjili DB load.

A diagram of a computer server

AI-generated content may be incorrect.

### Millions of users & beyond

Skaliranje sistema je iterativan proces. Iteracija na osnovu naucenog nas moze dogurati do “odredjene tacke”.

Vise “fine tuning” i noivih strategija su neophodne da bi skalirali preko milion korisnika, npr mozda je neophodno optimizovati sistem i razdvojiti sistem u jos manje servise. Sve naucene tehnike bi trebale omoguciti dobru osnovu.

Summary chaptera:

* Drzi web tier stateless
* Gradi redundantnost na svakom tieru
* Cacheuj podatke sto vise mozes
* Podrzi vise data centara
* Hostuj staticke podatne na CDN-u
* Skaliraj data tier shardovanjem
* Razdvajaj tierove u njihove individualne servise
* Vrsi monitoring nad sistemom i koristi alate za automaciju

A screenshot of a computer

AI-generated content may be incorrect.

## Chapter 2: Back-of-the-enevlope estimation