

System Design
(Gaurav Sen)

Aman Barnwal

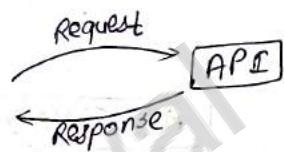
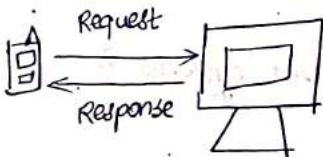
5 - common system design interview mistakes:

- don't go for a intro. Focus on core of the system.
- Don't do capacity estimation without any reason.
- Focus on Network protocols.
- Lack of Internals knowledge on how e.g. MySQL, Cassandra works internally.

A system design interview is not about building a working system. It's about measuring your aptitude.

Try to understand the Internals of system instead of knowing what to use where;

- Trade offs
 - i.e. Think of trade-offs while designing a system.
- You need to stay with the problem that you are trying to solve.
 - i.e. stay close to the core problem.

Horizontal vs Vertical Scaling

Scalability

being able to handle
more requests

Horizontal scaling

① 1 2 3 4 5

- ② Load Balancer required
- ③ Resilient i.e. if one of the machines fail,
the request can be redirected to other machines.
- ④ Network calls between services
(Remote procedure call (RPC))
- ⑤ Here, data is ~~consistency~~ complicated to maintain.
So, here data consistency is a real issue here.
- ⑥ It scales well as work increases

Vertical scaling

Huge Box

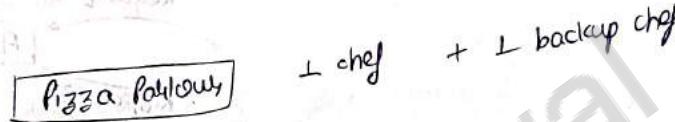
- ① No load balancer required
- ② Single point of failure.
- ③ Interprocess communication:
↓
fast

- ④ Data is consistent here. bcz there is
just one system on which all the data resides.
- ⑤ There is a hardware limit.

→ The hybrid soln is essentially the horizontal scaling only where each machine has a big box

→ Initially, you can vertically scale as much as you like. Later on, when users start trusting you, you should probably go for horizontal scaling.

How to start with distributed systems?



- ① Optimize processes and increase throughput using the same resource → Vertical scaling
- ② Preparing beforehand at non-working hours → Preprocessing
Now, make the system resilient (recovering from failures quickly)

Hire a backup chef

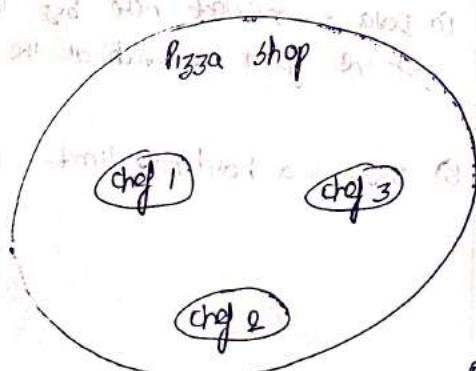
- ③ i.e. keep backups and avoid single point of failure
(just like Master-slave architecture)

→ Backups

- ④ Hire more resources

→ Horizontal scaling

or is buying more machines of similar types
to get more work done.



Here, chef 1 and 3 are experts at making pizza.

& chef 2's expertise is Garlic Bread.

On receiving an order, which chef should we assign it to?

(5) Microservice-based Architecture → You have all your responsibilities well-defined over here.

what if there is electricity outage in the pizza shop?

what if you lose the licence for the day?

→ You have a backup pizza shop which will also prepare the food.

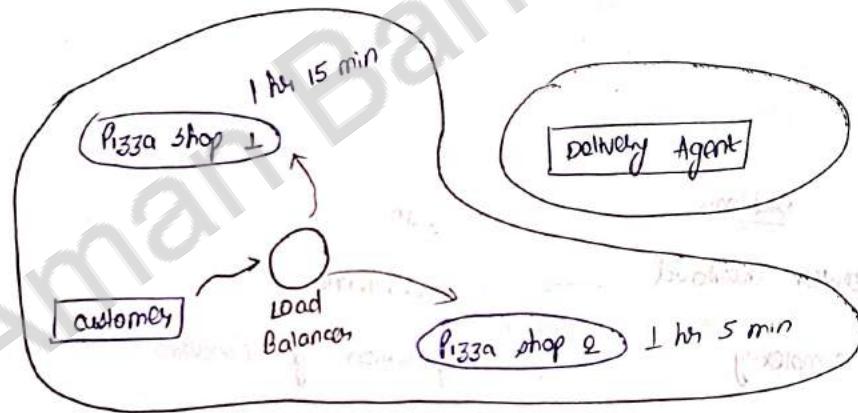
(6) Distributed system

(partitioning)

To give quick response time, you need some sort of local network everywhere.

so, we are distributing our system so that it's more fault-tolerant and also gives quicker response time.

(7) Load Balancer



Everytime, a customer makes a request, they need to either send it to one or two.

→ The system is now fault-tolerant but, how do you make it flexible to change?

(8) Decoupling:

Scalability
Fault Tolerance
Extensibility

→ The pizza shop whether it's the customer or delivery agent that is gonna come to pick up. so, we are seeing separation of responsibilities instead of having the same manager managing the pizza shop and delivery agents.

Also called decoupling the system

i.e. separating out concerns so that you can handle separate systems more efficiently

④ Metrics:

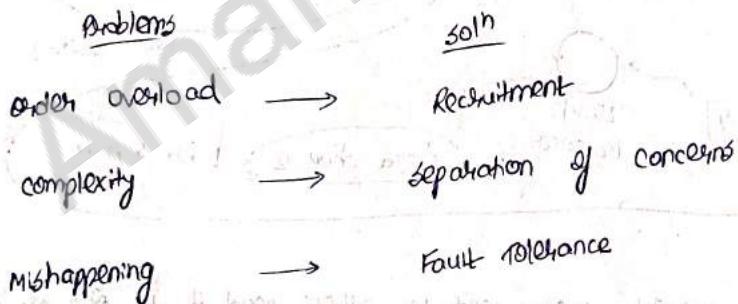
- ① Analytics
- ② Auditing
- ③ Reporting
- ④ Machine Learning

You want to log everything i.e. you want to see at what time something happened, and what is the next event so and so forth.

Also, you want to be taking those events condensing them, finding sense out of those events

⑤ Logging and Metrics calculation

⑥ Extensibility :



High level system design

Low level system design

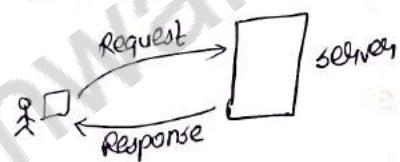
It is all about deploying our services, figuring out how different systems will be interacting with each other.

or has a lot more to do with how you're actually gonna code using classes, objects, etc.

- System Design is the use of computer engineering principles to build large scale distributed systems. It involves converting business problems and requirements into technical solutions.
- Senior engineers use system design patterns to build reliable, scalable & maintainable systems.

What is Load Balancing?

consistent Hashing:



The concept of taking actually n servers and trying to balance the load evenly on all of them is called load balancing.

The concept of consistent hashing helps us do that to evenly distribute the weight across all servers.

- You will get a request id in each request. When a mobile actually sends you a request, it randomly generates a number from 0 to (m-1). This request id is sent to your server.

request id \rightarrow 0 to (m-1)

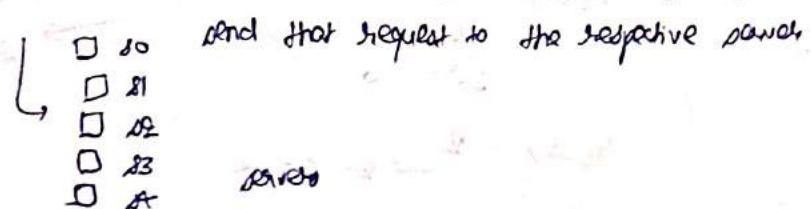
Take this request id and hash it. You will get a number (let's say m1)

$$m_1 = h[\text{req-id}]$$

This number can be mapped to a particular server

Here, $n = \text{no. of servers}$

$m_1 \leq n \rightarrow$ whatever index you get here, you just



$$\text{Ex: } \#_1 = 10, n = 4$$

$$h[10] = 3$$

Hash value of 10 goes to server 3. This request

$$3 \cdot 4 = 3$$

$\#_1$ will go to server 3

$\square s_0$

$\square s_1$

$\square s_2$

$\square s_3$

②

$$\#_2 = 20, n = 4$$

$$h[20] = 15 \text{ (load say)}$$

-

$$15 \cdot 4 = 3$$

$\square s_2$

$\square s_3$

$$\#_3 + \#_4 = 35$$

$$h[35] = 12$$

$$12 \cdot 4 = 30$$

→ Your hash fn is uniformly random so you can expect all of the servers to have uniform load.

→ If there are ~~a few~~ requests will have $\frac{x}{n}$ load

where they all fall & the Load factor is $\frac{1}{n}$ which is same as bucketing factor

What happens if you need to add more servers?

→ If you are adding more servers i.e. $n \uparrow$ then the load will get

so hashing index will change

$$\text{Ex: } h[10] = 3$$

$$3 \cdot 5 = 3$$

$$h[20] = 15$$

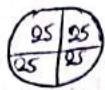
$$15 \cdot 5 = 0$$

$$h[35] = 12$$

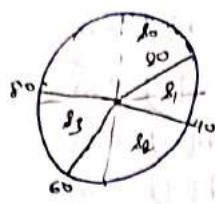
$$\rightarrow 12 \cdot 5 = 2$$



Initially, $n = 4$



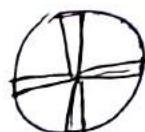
If n changes to 5



$$\begin{aligned} & 5 + 5 + 10 + 10 + 15 + 15 + 20 + 20 \\ & = 100 \\ & \text{i.e. entire search space} \end{aligned}$$

- Depending on the user id, we can send people to specific servers and once they are sent there, you can store relevant information in the cache for those servers.

The overall change will be minimum if you are adding more servers



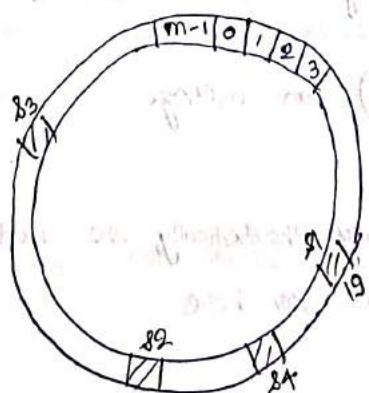
What is Consistent Hashing and where is it used?

The problem is adding or removing servers that completely changes the local data that we have in each servers and to avoid that, we will use consistent hashing.

Request ID $\rightarrow h(\text{id})$

Instead of array $0 \perp \dots \perp (m-1)$
which can map this hash

Ring of hash \downarrow value from 0 to $(m-1)$



We can take the servers and we actually need to send those requests to those servers. The servers themselves have IDs from 0, 1, 2, ...

Using the same/different hash f^n , we can hash this server id.

server id

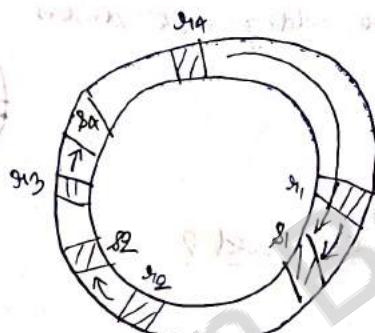
$m = 30$

$$\begin{array}{l} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \xrightarrow{\text{hash}} \begin{array}{l} h[0] \div m \\ h[1] \div m \\ h[2] \div m \\ h[3] \div m \\ h[4] \div m \end{array}$$

$h[0] \div m = 49 \div 30 = 19$

i.e. server 1 will be mapped to position 19

→ whenever a request comes in this ring, we will go clockwise and find the nearest server.



This server s_1 is serving
request r_1 .

Both such cases exist because i.e. s_1 has a load of two requests has.
(r_4 and r_1)

s_2 ——— 1 request (r_2)
 s_4 ——— 1 request (r_3)

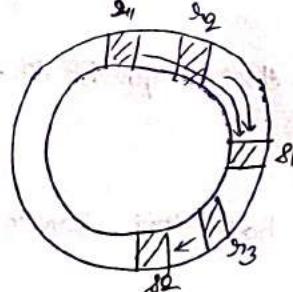
→ because the hashers are uniformly random, you can expect the distance
b/w them to be also uniform, because of which the load is uniform.

so, the load factor turns out to be $(\frac{1}{n})$ on average.

→ The problem with this architecture is although theoretically, the load should be $(\frac{1}{n})$.

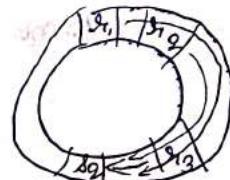
Practically, we can have skewed distribution over here

Eg:



if server s_1 got removed then all the
requests r_1 , r_2 and r_3 got directed to
server s_2 .

i.e.



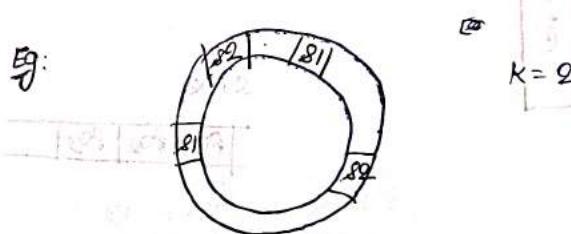
here, you can start making virtual servers

it does not mean that you have virtual boxes or you start buying more servers coz those are expensive.

what you can do instead is do use multiple hash fn.

h_1, h_2, \dots

- If you have k hash fn from which you pass each of the servers IDs then each server will have k points. So, here the likelihood of one server getting a lot of load is much much lesser.
- If you choose k value appropriately, you can entirely remove the chance of a skewed load on one of the servers. (let's say $\log n$)
- Now, if a server is removed, you need to remove k points on it and clockwise assign them to the nearest server.

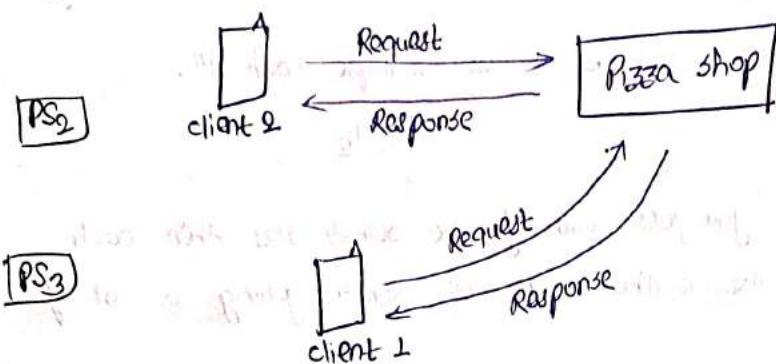


- It can be used in many many places like search engines and web servers etc.
- Load Balancing is used in Distributed Systems extensively. This is being used by web caches, databases
- Consistently Hashing gives you the flexibility and load balancing in a very very clear and efficient way.

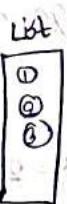
What is a message queue and where is it used?

→ Now we used message queue when we have lots of clients.

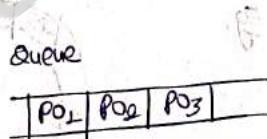
Ex:



Multiple clients request their pizzas and they get the responses immediately like "Please sit down". You relieve the clients from expecting an immediate response by giving them a response which is not the pizza but a confirmation that the order has been placed.



You need list which contain order no.



→ Once a pizza had been made, you remove that order from queue. & you then ask the client to pay and the client send you back money.

This is asynchronous ie. you didn't make the client wait for your response for the payment after the pizza.

→ similarly, this allows you as the pizza maker to order your tasks according to their priority.

Asynchronous Processing

→ Let's say, you have multiple outlets in your pizza shop.

(Eg: Dominos)

Each pizza shop has multiple clients connected to them.

→ Assume that one of these pizza shops actually goes down, (bcz of power outage) so, we need to get rid of all takeaway orders. Just, we need to dump them. But, the delivery orders can be sent to other shops.

i.e. They can complete delivery orders and you can still save money

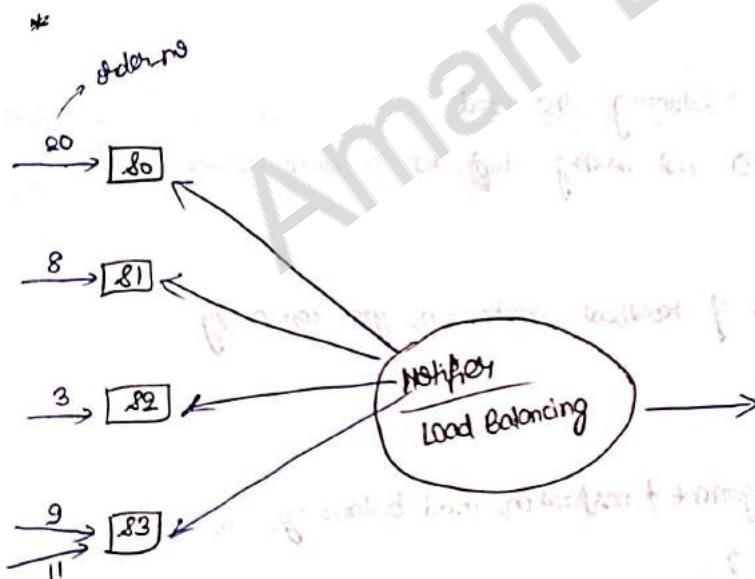
Eg:

[PS₁]

[PS₂]

[PS₃]

You need some persistent in your data
k you need to alter this list in database.



ID	contents	done ?
1	A	Y
2	B	N
3	C	N
4	D	Y

set of servers

Database is storing the list

→ If S3 crashes, then order id 9 and 11 needs to be routed to other servers. For that, we could have some sort of notifiers, it checks for a heartbeat in each server.

ie A notifier talks to each server and ask them if they are alive every 10 or 15 sec.

→ If a server doesn't respond, the notifier assumes that the server is dead.

If it is dead, it can't handle orders and then it can query the database to find all of those orders which are not done.

It picks those undone orders and distributes it among remaining live servers.

Here, problem is duplication.

Here, you can use load balancing.

It will ensure that you do not have duplicated requests to the same server.

→ Consistent hashing is one technique by which you can get rid of duplicates also.

It will take care of 2 things:

- ① Balancing the load
- ② Not sending duplicates to same server.

Through load balancing and through some sort of heartbeat mechanism, you can notify all failed orders to never servers.

What if you want all the features of assignment, notification, load balancing, a heartbeat and persistence in one thing?

That would be Message Queue.

It takes tasks, persists them, assign them to the closest server, waits for them to complete.

If it's taking too long for the server to give an acknowledgement, it feels that server is dead and then assigns it to next servers.

- There are multiple strategies of assigning the failed orders just like load balancing has multiple strategies but that's all encapsulated by a task queue.
- We can use Messaging Queue or Task Queue to get work done easily so that you can encapsulate all that complexity into that one thing.

An e.g. of Messaging queue is "Rabbit MQ"

zero MQ

Java Messaging service (JMS)

They are really good
encapsulations for complexities in
server side.

Amazon also has some MQ.

Kafka

- Messaging queues are widely used in Asynchronous systems. It provides features such as persistence, routing and task management.

- There are multiple strategies of assigning the failed orders just like load balancing has multiple strategies but that's all encapsulated by a task queue.
- We can use Messaging Queue or Task Queue to get work done easily so that you can encapsulate all that complexity into that one thing.

An e.g of Messaging queue is "Rabbit MQ"

zero MQ

Java Messaging service (JMS)

Amazon also has some MQ.

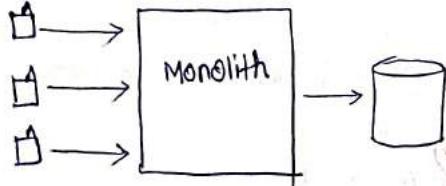
Kafka

They are really good
encapsulations for complexities in
server side.

- Messaging queues are widely used in Asynchronous systems. It provides features such as persistence, routing and task management.

What is a microservice architecture and its advantages?

Monolithic vs Microservices

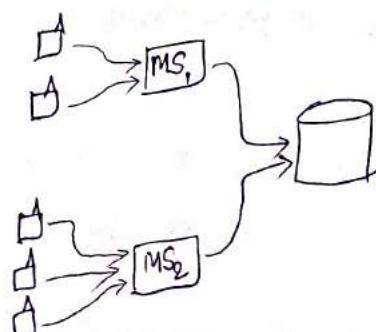


→ Monolithic does not need to be single in terms of no. of machines you are running it on.

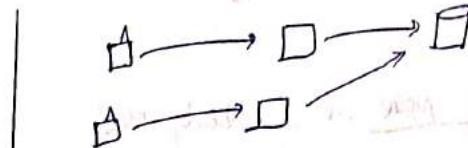
Because there might be multiple machines with the same monolith.

→ Also good for small teams.

when put under a lot of load, the monolith architecture scales out.



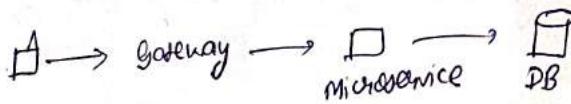
MS is monolithic service



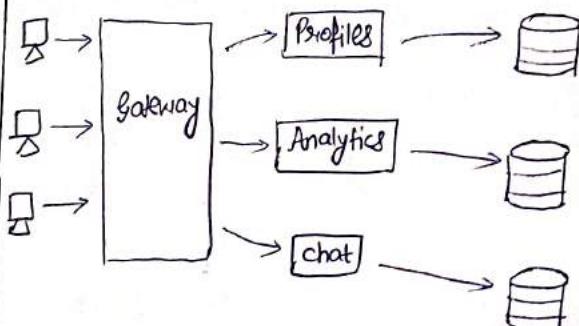
→ In microservices, you just have one of them running on one machine and they have tiny tiny databases with each of them connected.

→ They usually have dedicated database connected to it.

→ The client might not be talking to microservice, it might be talking to a gateway.
↳ These gateway is talking to other microservices internally.



Eg:



→ Easier for new developers.

Easier to scale services.

Advantages

Monolithic

→ less complex

Deployment in monolithic application are easy

→ less duplication

All that code for setting up tests, connections, etc in your system need not be duplicated for every service that you create.

→ also faster bcz you are not making calls over the network.

(also RPC call, which is faster.
(Remote Procedure call))

All logic and code are going to be in the same box so it's a lot more faster.

Disadvantages

→ more context required if you wanna add a new member in the team

→ complicated deployments bcz any change in the code requires a new deployment.

So, your code is going to be deployed very frequently and it has to be monitored every time it is deployed.

→ There is too much responsibility on each server.
i.e. single point of failure.

→ Tight coupling in code as well as in developer time i.e. more dependency of one task on another.

Microservice

Advantages

→ New services can be tested easily and individually

→ It's easier to scale.

→ Easier for newer team members

→ Parallel development is easy bcz there is lesser dependency b/w the chat developers on the analytics developer now. (Eg)

→ Used in a lot of companies like Google, fb, etc.

→ They are more available as a single service having a bug does not bring down the entire system.

→ Individual services can be written down in different languages

Disadvantages

→ They are not easy to design.

→ A good indicator that you have a microservice that should not be microservice is that if it's talking only to one service.

→ It needs smart architects to architect well

for a sub-service in architecture.

* There is a very successful system which uses the monolithic architecture which is StackOverflow.

* Monolithics are favourable when

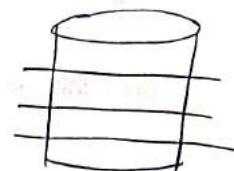
- The technical / developer team is very small.
- The service is simple to think of as a whole.
- The service requires very high efficiency, where network calls are avoided as much as possible.
- All developers must have context of all services.

Database sharding:

Partitioning of database:

Horizontal partitioning: This kind of partitioning which uses some part of a key to break the data into pieces and allocate that to different servers is called horizontal partitioning.

It involves putting different rows into different tables.



Vertical partitioning:

It involves creating tables with fewer columns and using additional tables to store the remaining columns.



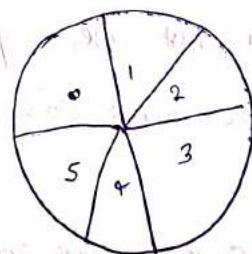
Database sharding:

It is a common scalability strategy used when designing server side systems.

The server side system architecture uses concepts like sharding to make systems more scalable, reliable and performant.

Sharding is horizontal partitioning of data acc to a shard key. This shard key determines which database the entry to be persisted is sent to.

Some common strategies for this are reverse proxies.



Joins across shards

→ Eg: In application like Tinder, you could shard the data acc to location

Problems in sharding that we have to take into consideration →

→ Joins across shards

The query needs to go to two different shards, need to pull out their data then join the data across that network.

↓

This is going to be extremely expensive

→ Shards are inflexible

Memcached Database

Using hierarchical sharding, we can get rid of inflexibility over here.

→ Your read and write performance goes up bcz all of your queries falls on one important particular point.

What happens if a shard fails?

→ In that case, you can have master-slave architecture.

Here, we have multiple slaves which are copying the master.

Whenever it's a write request, it's always on master.

(The master is the most updated copy).

while, the slaves continuously pull the master and read from it.

→ If it's a read request, it can be distributed across slaves, while the write request always goes to master.

→ in case, the master fails, the slaves choose one master among themselves

Note:

Database sharding is the process of splitting the data table into multiple database instances in order to distribute the load.

However, sharding a database is an expensive operation (maintainability and overhead)

and I suggest you do that only when you absolutely need to

How Netflix onboard new content : Video processing at scale

How Netflix onboard new content on their platform

→ challenges faced in uploading new content :

we need to store it in different formats like MP4, etc. ^{bcz different}

people have different connection speeds

① Different formats

- ↳ High quality
 - ↳ Medium "
 - ↳ Low "
- codec

② Play with different resolutions

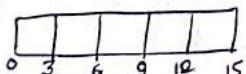
A single video has multiple formats and resolutions and each is creating tuples

$$\text{No. of videos that you end up processing} = \frac{F}{\text{no. of formats}} \times \frac{R}{\text{no. of resolutions}}$$

Netflix takes the original video, it breaks them into chunks, run those chunks into different resolutions and format

e.g.: chunk A: mp4, 1080
chunk B: avi, 720

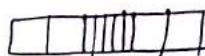
video file



Each chunk is of 3 min each.

Breaking the chunks based on scenes rather than on timestamp

each scene timestamp has a lot of left chunks of 4 sec each.



How Netflix onboarded New content : Video processing at scale

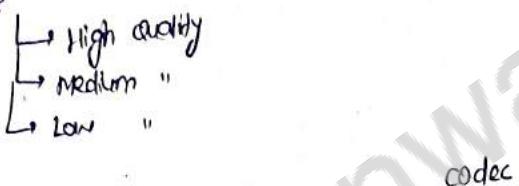
How Netflix onboarded new content on their platform

→ challenges faced in uploading new content :

we need to store it in different formats like mp4, etc. bcz different

people have different connection speeds

① Different formats



② or Play with different resolutions

A single video has multiple formats and resolutions and each is creating tuples

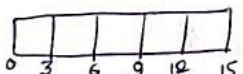
$$\text{No. of videos that you end up processing} = \frac{F}{\text{no. of formats}} \times \frac{R}{\text{no. of resolutions}}$$

Netflix takes the original video, it breaks them into chunks, run these chunks into different resolutions and format.

Eg: chunk A: mp4, 1080

chunk B: avi, 720

Video file



Each chunk is of 3 min each.

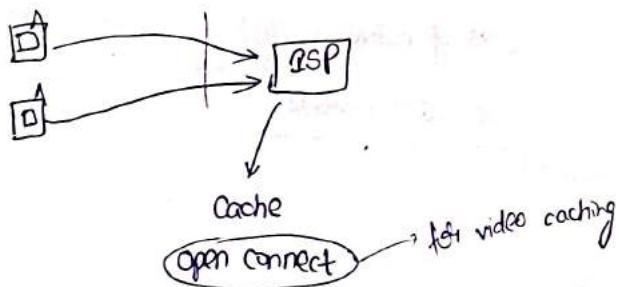
Breaking the chunks based on scenes rather than on timestamp

each scene lets say has a lot of left chunks of 4 sec each.



sparse movie

→ Netflix stores all its data in Amazon S3.



→ Around 90% of Netflix traffic is taken care of by these ISP boxes that they provide. They are called open connect.

You can store all local popular movie in this open connect box.

→ Netflix provides open connect servers to ISP, which acts like a cache of movies.

Tinder as a microservice architecture:

Features:

- store profiles (Images - 5 images/user)
- Recommend matches (No. of active users) $\times 10^{-3}$ matches
- Note matches
- Direct messaging

How you are gonna store images?

- as a file or Blob (binary large object)

Objects are large in size.

clob
(character large object)

Database is giving these properties —

- ① Access control
- ② Mutability
- ③ Transaction ACID
- ④ indexing

not good for images

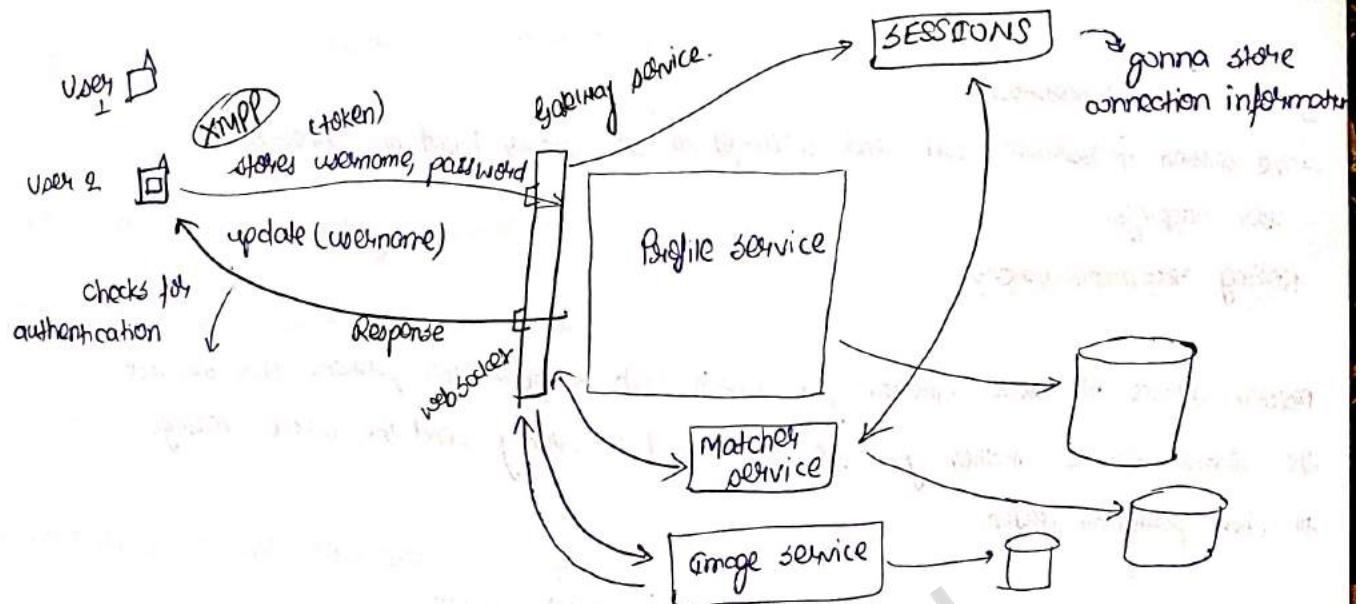
File is having these properties —

- cheap
- Faster
- static so you can easily build CDN over it

It allows fast access

profile id	image id	file url
------------	----------	----------

Distributed file system → handles store profiles feature



The client always talks to the gateway. The gateway makes this request (token) and ask profile service whether this token is authenticated or not.

→ Logically, you would probably want to store the images in a separate service.

Direct messaging: It can be done using XMPP protocol, which uses web sockets to have peer-to-peer communication between client and servers.

message to user 1 from user 2

message(UD₁, UD₂)

Each connection is built over TCP ensuring that the connection is maintained.

→ When you have client-server communication protocol (HTTP), you cannot effectively have chat.

Bcs if there is a client and server, the only way that this user is going to get messages that are sent to them is poll to servers i.e. every 5 sec, it asks here is there any message for me?

↓
It is really inefficient.

So, you don't want to poll the servers. You want messages to be pushed to you.

So, we peer-to-peer protocol. One of these protocols is XMPP.

→ This XMPP is going to take WebSocket connection. Instead of this, you can also use TCP with this connection, you can talk to clients.

→ Web socket connection.

→ The session microservice can send messages to the service based on connection to user mappings.

Noting recommendations

(04:18)

Matcher service will check whether you match with a particular person. That can tell the session service whether you are authorized to actually send the direct message to that particular person.

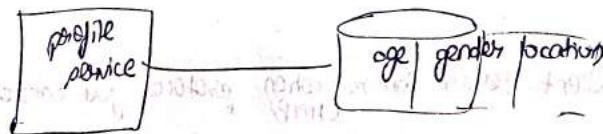
There will be some communication b/w matcher & session service

The matcher service can note down all the matches that all have.

→ When you ~~uninstall~~ swipe left or right, the only information you will lose is what all persons you swiped left or right. ~~By this, is not the critical info which needs to be stored.~~

Recommendating people to you

Based on which gender you are interested in
age
location



You need to optimize on various parameters in db.

① So, you need to either use a NoSQL database like cassandra.

or

Amazon dynamo.

② Sharding (Horizontal Partitioning)

on relational database

Eg: name → (A-J) → a node ①
(K-P) → node ②

what about single point of failure?

→ You can use Master slave architecture.

→ You need to shard the data based on location.

Eg: It could be chunks of a city. A person within this location is within this chunk and if it is within this chunk, they are being sharded to a particular node.

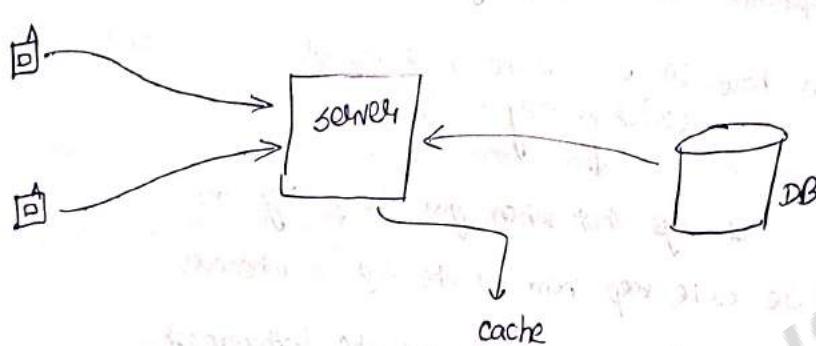
Based on that chunk also, you can pull the data out within that chunk, and then search amongst it within the age and gender.

Recommendation service

It pulls out all relevant people.

What is Distributed Caching?

- caching in distributed systems is an important aspect for designing scalable systems.
- cache management is important bcz of its relation to cache hit ratios and performance.



Two scenarios where you might want to use the cache →

- ① When you are querying for some commonly used data.
ie to save network calls.
- ② When you want to avoid re-computation.

Eg: calculate age avg

↓
Find it once & store it in a cache with a key value pair.

- ③ When you want to avoid load on the database.
ie. reduce database load

All above helps us to speed up responses from clients. So, when client makes a call, you immediately gives a response if you have it in cache instead of making a query to the database.

- The hardware on which a cache runs is much more expensive than that of a normal db.
- If you store a ton of data in the cache, then the search time will increase.

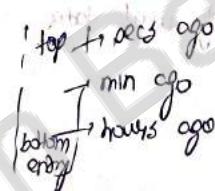
→ The way in which you are loading and evicting data in the cache is called a policy.

Cache Policy

* Cache performance is entirely dependent on cache policy.

There are multiple policies you can have. ② LRU is the most popular.
(Least Recently Used)

It pays that when you are pulling an entry to the cache, keep them at the top & whenever you have to kickout data, kickout the bottommost entry i.e. least recently used entries in the cache.



③ LFU policy (Least Frequently Used)

④ Sliding window based policy

Problems

Note: if you have a poor eviction policy, it's actually harmful to have a cache bcz you are making that call.

i.e. Extra calls

→ ⚫ Threading

→ ⚫ different servers makes a call to the database for an update, but the cache doesn't have that entry updated.

Data inconsistency

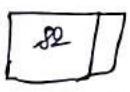
Where to place the cache?

either you can place the cache close to the server

or you can place it close to the database.

→ There are benefits and drawbacks for both.

→ If you place the cache close to the server, how close you can place it?
You can place it in memory itself in the server.



What if server S2 fails?

→ Then the in-memory cache of server S2 also fails. And, we need to take care of that.

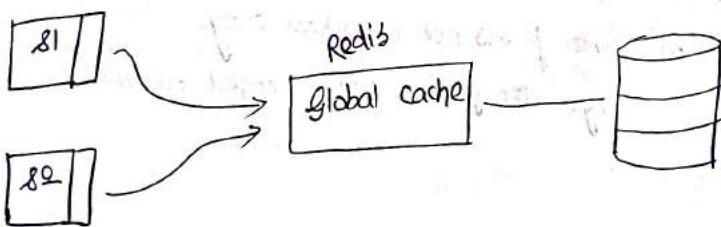
What if data on in-memory cache of servers S1 and S2 are not consistent?

② Benefits of having cache closer to the server →

• This is faster

• It is easier to implement

③



This is faster disk load.

→ Here, even if S2 crashes, nothing will happen bcz

all are anyways querying to the global cache.



Higher Accuracy

Redis →

→ You can scale the Redis cache independently while keeping the services as containers running on these boxes

→ How consistent the data of cache is?

Write Through cache



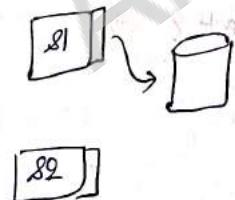
Note, I write on the cache before going through it and hitting the database.



Problems:

If you have multiple in-memory cache on different servers, then it will be a problem.

Eg:



→ We can also hybrid sort of both write-through and write-back.

Whenever there is an entry that has to be updated in the dB, don't write back immediately. Write on cache if its not a critical entry.

(Eg: The person has changed comment).

When you →

After some relevant time, take entries in bulk and persist it on the database so that's one network call and you are saving on hitting the database consistently.

→ If data is less say passwords, you can't have write-through cache, you need the write-back cache.

What is an API and how do you design it?

API design:



Application

programming interface

→ Naming is important.

→ Don't take additional parameters unless absolutely necessary.

Eg: You want to fetch all group admins of WhatsApp.

getAdmins (string groupId)

error ↗ group does not exist
↗ group is deleted

Response → List <Admin> admin

→ A lot of times, we need to expose our API to an HTTP endpoint.

Eg:

POST www.aman.tech / chat-messaging / getAdmins / v1

↓
address of
your website

↓ which url
it needs to be
called

↓ version

Request: { "groupId": "123" }

Response:

{ "admins": [{ "id": "xyz", "name": "Aman" }] }

Eg: GET www.oman.com/chat-messaging/admins?groupID=123

of API

- * If response is much large, you can break the response into pieces, ie. Pagination
→ Fragment the API.

capacity Planning and Estimation

How much data does YouTube store daily?

(Q) Estimate YouTube's daily video storage requirements \Rightarrow

Let's say, there are 1 billion YouTube users. One out of 1000 upload a video of avg duration of 10 min.

$$\text{Total min} = \left(10^9 \times \frac{1}{1000} \times 10\right) \text{ min} = 10^7 \text{ min}$$

Let's say, 2 hrs of video takes 0.4 GB (can be upto 4 GB)

$$1 \text{ hr} = 0.2 \text{ GB}$$

$$= 200 \text{ MB/h}$$

$$= \frac{200}{60} \text{ MB/min}$$

$$\approx 3 \text{ MB/min}$$

$$\text{Memory taken} = 3 \times 10^7 \text{ MB}$$

$$= 30 \text{ Terabyte}$$

$$= 30 \text{ TB}$$

This is the approximate amount of storage you need to just keep raw footage of YouTube

I am assuming 3 copies for fault tolerance.

$$\therefore 30 \text{ TB} \times 3 = 90 \text{ TB}$$

We also want multiple resolution of a video.

720p \rightarrow Let's say, it takes $x \text{ MB}$

480p \rightarrow $\frac{3}{4}x \text{ MB}$

360p $\rightarrow \frac{3}{4}x \text{ MB}$

240p $\rightarrow \frac{3}{8}x \text{ MB}$

144p $\rightarrow \frac{3}{16}x \text{ MB}$

} sum $\approx x$

If you want to store the video in all the

resolution, memory taken $= x + x$

$$= 2x$$

$$\therefore \text{Total memory taken for storing the video in all resolution} = 2 \times 90 \\ = 180 \text{ TB} \\ \approx 0.2 \text{ PB}$$

i.e. No. of users = 1B

Ratio of uploaders : users = $\frac{1}{100}$

$$\text{No. of uploaders} = 1B \times \frac{1}{100} = 1M$$

Avg length of each video = 10 min

∴ Total uploaded video length = $1M \times 10 = 10M$ min

Let's say, 2 hr movie avg size = 4 GB

Optimized codec and compression savings = 90%

$$\therefore 2\text{ hr YouTube video avg size} = 4 \times \left(1 - \frac{90}{100}\right) = 0.4 \text{ GB}$$

$$\text{i.e. } 2\text{ hr} = 400 \text{ MB}$$

$$1\text{ hr} = \frac{400}{2} = 200 \text{ MB}$$

$$\text{i.e. } 200 \text{ MB/hr}$$

$$= \frac{200}{60} \text{ MB/min}$$

$$\approx 3 \text{ MB/min}$$

$$\text{Total video size} = 10M \times 3 \text{ MB} \\ = (30M) \text{ MB}$$

$$= 30 \text{ TB}$$

I could also think that 1 video consists of collection of images.

What's the size of 1 min long video?

$$1 \text{ min} = 60 \text{ sec}$$

$$\text{No. of frames / min} = 24$$

$$\text{Size of one image} = 8 \text{ (Let's say 1 MB)}$$

$$\therefore \text{Total size} = 60 \times 24 \times 1 \text{ MB}$$

$$= 1440 \text{ MB}$$

$$\approx 1.5 \text{ GB / min}$$

You already calculated earlier that
the storage that a video takes per min is (3 MB)

$$\text{i.e. } 3 \text{ MB / min}$$

Huge difference, why?

You might have taken the wrong assumption of size of an image in video.

→ YouTube reports 600 hours of new content every minute.

Note, we have just assumed above, $10^7 \text{ min of new content} = 115 \text{ hours / min}$

$$60 \text{ min} = 1 \text{ hr}$$

$$1 \text{ day} = 3600 \text{ min} \times 24 \\ = 86400 \text{ min}$$

$\therefore 1 \text{ day} = 10^7 \text{ min of new content}$

$$\frac{10^7}{86400} = 115 \text{ hours of new content / min}$$

so you took a nice guess.

⑥ Estimate cache requirements for video metadata \rightarrow

\rightarrow You have thumbnail and title of the video that you have to show.

\downarrow
an image, so takes more memory than title of video

when you are showing the thumbnail to the user, you are not showing the actual image in the original dimensions, but you are showing in much

much smaller size & dimension
 $(\text{less pay both side})$
 $(\text{smaller from both side.})$

lets say, image will take 1 MB in original dimensions

$$\therefore \text{Red image in reduced dimensions} = 1 \text{ MB} \times \frac{1}{100} = 10 \text{ KB}$$

This is what you need in the cache for a thumbnail.

10 KB \times No. of videos you upload
in a day

\downarrow
This could be 1000 for the popular
videos in the last 90 days

(assuming 1 million videos a day)

$$\text{Memory taken for thumbnail} = 10 \text{ KB} \times 90 \times 1 \text{ M}$$

$$\approx 10 \text{ KB} \times 100 \times 10^6$$

$$= 10^9 \text{ KB}$$

= 1 TB RAM \rightarrow This is much more expensive than usual

\downarrow
You can't get this 1 TB RAM in a single computer.

~~If we write~~

if I am using 16 GB RAM computer nodes,

$$\therefore \text{No. of nodes required} = \frac{1 \text{ TB}}{16 \text{ GB}} = \frac{1000 \text{ GB}}{16 \text{ GB}}$$

$$= 64 \text{ nodes in cache system.}$$

if you are looking for any kind of redundancy, which is good to have bcs you need to serve video all over the world.

6 3 nodes copy of each node

$$\text{ie. } (64 \times 3)$$

if any of those nodes crashes, then we are going to have cascading failure bcs you were at your peak capacity

if any one of them crashes, then loads on the other one will increase.

Assuming 50% capacity

$$\text{ie. } 64 \times 3 \times 2$$

$$\approx 500 \text{ nodes}$$

i.e. with 500 nodes, I will have a caching system, each with 16 GB which should be enough for YouTube metadata.

② Estimate no. of processes required:

We want to know how much MB we can process per sec?

1 million videos a day

1 video is of 10 min each.

$$\text{ie. } (10^6 \text{ videos}) \times 10 \text{ min} = (10^7 \text{ min of videos}) \text{ a day}$$

assuming, 1 hr unprocessed video has average size of 1 GB.

$$(10^7 \text{ min of videos}) \text{ a day}$$

$$= 1000 \times 10^4$$

$$\frac{10^4}{3} \text{ hrs of videos}$$

$\frac{10^4}{3}$ GB of videos to be processed in a day

$$= \frac{10^4 \text{ GB}}{3 \times (No. \text{ of sec in a day})} \text{ of videos in a sec}$$

$$= \frac{10^4 \text{ GB}}{3 \times (94 \times 60 \times 60)}$$

= 40 MB of videos to be processed in a sec

when we talk about processing power, we talk about this is the amount of data that it can process in a second.

Read the video into the memory

Processing
write it back

10 MB/sec is the given footage

↓ if you are taking ~~given~~ footage doing multiple frames
in multiple data centers

Let's say, it is 100 MB/sec

i.e. 100 MB/sec

To read a single MB of data from a disk → 10 msec
Exclusive locks

Period can possibly require

Index updated

100 MB/sec write speed → 20 ms
processing → 20 ms

$$\begin{aligned} & \text{Time taken for 1 MB of data} = \text{Read} + \text{Processing} + \text{Write} \\ & = 10 + 20 + 20 \\ & = 50 \text{ millisecond} \end{aligned}$$

i.e. You need 50 millisecond to process 1 MB of data.

$$\begin{aligned} & \text{For 400 MB} \\ & = 400 \times (50 \times 10^{-3}) \text{ sec} \\ & = (20 \text{ sec of work to be done}) \text{ per sec} \end{aligned}$$

↓
Computer can't do this work, so

you have to keep computer in 1 sec.

i.e. 20 processors are required.
(by our assumption).

after, one could be

- Estimate the hardware requirements to set up a system like YouTube.
- Estimate the no. of petrol pumps in the city of Mumbai.



Another type of network connection is called peer-to-peer connection. In this type of connection, each computer acts as both a client and a server. It can receive data from other computers and also send data to other computers.

Now let's look at Peer-to-peer networking. Peer-to-peer networking is a type of network where each computer acts as both a client and a server. This means that each computer in the network can both receive data from other computers and also send data to other computers. This type of networking is often used for sharing files between multiple computers on a local network. One example of a peer-to-peer network is BitTorrent, which is used for sharing large files over the Internet. In BitTorrent, each computer acts as both a client and a server, downloading files from other peers and sharing its own files with them. This allows for faster file sharing than traditional client-server networks, where all files must pass through a central server.

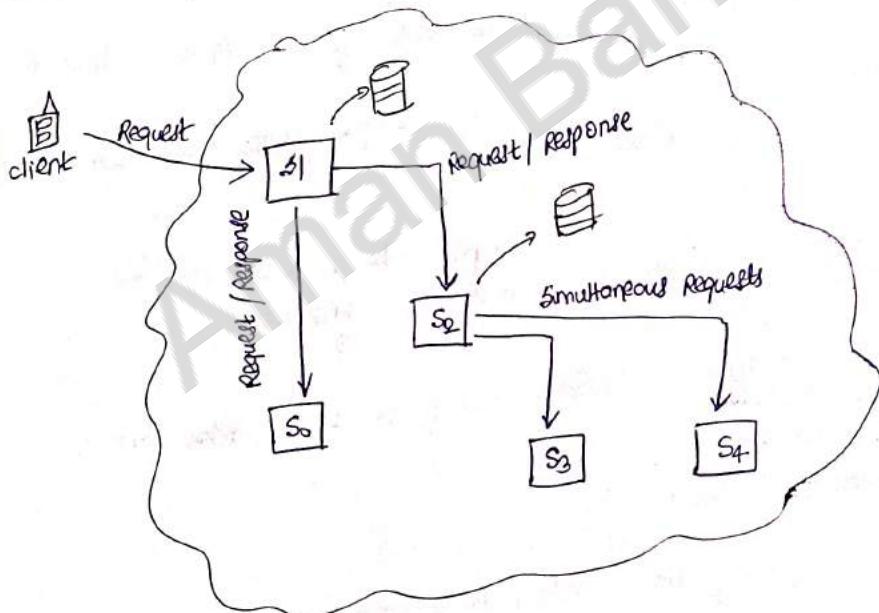
Publisher Subscriber model:

It is also known as pub/sub. It is an architectural design pattern that provides a framework for exchanging messages between publishers and subscribers.

This pattern involves the publisher and subscriber relying on a message broker that relay messages from the publisher to the subscriber.

The host (publisher) publishes messages (events) to a channel that subscribers can then sign up to.

Event driven system:



We have a microservice architecture in the above.
So, S_1, S_2, \dots are the services we have.

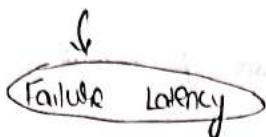
If you are using request-response architecture, smart thing to do is send the requests from S_2 to S_3 and S_4 asynchronously and wait for the responses.

The main drawback of this architecture is S_2 might be waiting for both of these services (S_3 and S_4) to complete even if it is asynchronous.

Strong coupling

of s_1 fails, s_2 is waiting for it, after a timeout it's going to send a timeout to s_1 , which in turn is going to send the timeout to the client.

↳ it took a long time for this request to fail.

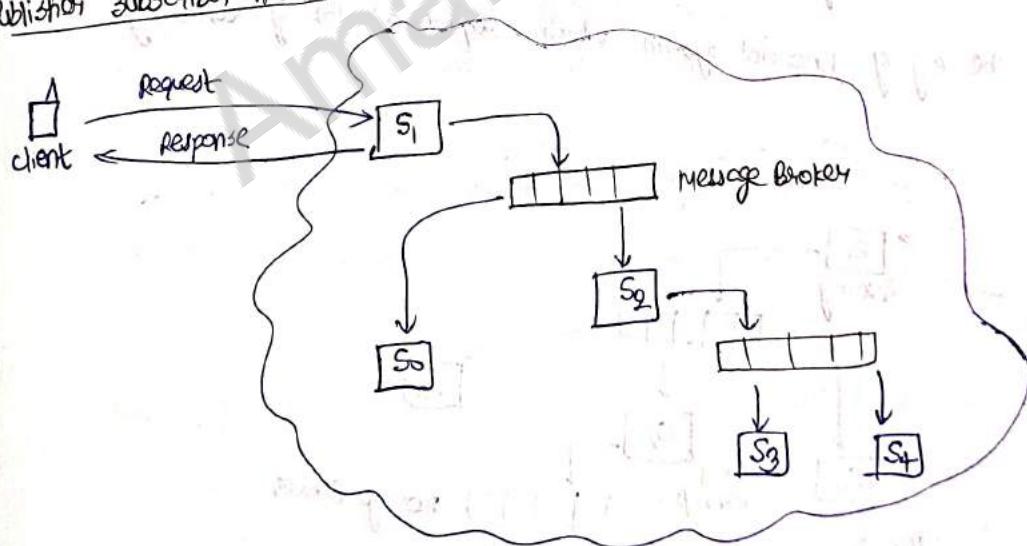


→ when the request is sent again, s_1 is going to make a change in DB again, and s_2 also is going to make a change in DB on getting the same request.

↓
It leads to inconsistent data bcz there are two different changes for the same request.

Better way to handle this is by using Publisher Subscriber model.

Publisher subscriber model:



There are both advantages and disadvantages to this architecture.

Advantages

① → This is going to decouple a lot of responsibility that you had.

i.e. S_1 is no longer dependent on s_2 and s_3 . Instead, it just publishes to the message broker. ↳ send a message to the client saying 'I'm successful'.

Decoupling

② simplified interactions. instead of multiple point of failure, it's a single point of failure which is far more easy to deal with.

② It also provides transaction guarantee.

i.e. if you are sending a message to s_1 , and if it's able to persist to the message broker, that means the message will somehow reach s_3 at some point of time in future.

It's a loose transaction guarantee of at least once.

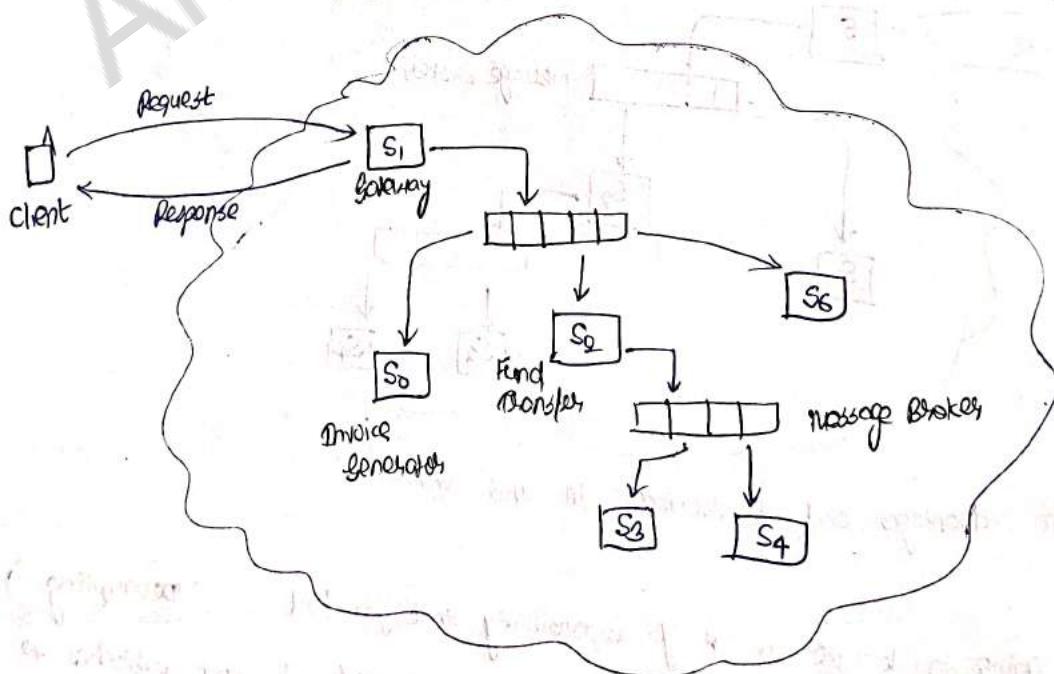
③ It's also more easily scalable.

bcz if you have a new service s_6 which is interested in s_1 's message, all you need to do is to register this service s_6 with the message broker, and then the message broker will send the messages being produced by s_1 to s_6 also.

s_1 do not need to know the subscriber of its messages.

Disadvantages: ① Poor consistency

Let's take the e.g. of financial system which require a lot of consistency.



Let's say s_0 is Invoice Generator service

s_1 is gateway (it just processes the message and send to s_0 and s_2)

s_2 is Fund Transfer service,

Let's say, you get a request from client saying that "Please transfer my funds of 850 from Account A to B."

Initial amount in A = Rs 1000

Bank charges for transfer = Rs 50

∴ After transfer, net amount in A should be = $1000 - (850 + 50)$
 $= 0$

Gateway gets the message, forwards it to S2 and S3.

S2 says "Remove Rs 850 from the account" bcz that's my commission, and sends an email to the client saying that this is transaction amount you asked for.

$$\begin{array}{r} 1000 \\ - 50 \\ \hline 950 \end{array}$$

Let's say, Fund transfer service is down, which means the message broker is not able to send those messages to S2.



The client send then another request to transfer Rs 800.
so in total of 850 needs to be in your account. $\textcircled{A} \xrightarrow{800} \textcircled{B}$



Request come to S3, it checks whether 850 is there in account.

If not, no it deducts Rs 50. Remaining balance now = $950 - 50$
 $= 900$

And, the Fund Transfer got two requests, { one of Rs 800
other of Rs 950 }

The first transaction of Rs 950 now fails bcz the remaining balance is just Rs 900.

The 2nd transaction of Rs 800 succeed leaving $Rs (900 - 800) = Rs 100$ in the account.

The initial expectation was there should be 0 rupee left and the 2nd transaction should have failed.

Instead, you have 100 remaining and 800 transferred, and Rs 100 as commission to the bank.

So, there is a lot of confusion bcz of having a transaction across services.

↓
Poor consistency

② It doesn't guarantee anything about idempotency.

Eg: Let's say, we are doing something which is not idempotent, let's say withdrawing Rs 50 from your account. It might fail when publishing over to message broker. So, it sends the failure response.

After the message is replayed, it might have debit Rs 50 more from here and then sends the message again.

So, how you are seeing the issue of so Rs removed multiple times by the message that the message broker is pending is not idempotent.

Instead, it should be also sending request id along with it.

Publisher subscriber model is the basis for event driven services.

It can't be used for finance services.

But can be used for gaming services.

Note: Twitter uses 'Publisher subscriber model' architecture.

For that business requirement of posting a tweet and many people actually consuming that tweet, the publisher subscriber model is perfect.

You have events published and subscribers for that.

some well known messaging frameworks based on publish-subscribe pattern →

- * Apache Kafka
- * RabbitMQ
- * PushPin

Why do Databases fail ? Anti Patterns to avoid

Database are often used to store info, but one case where it becomes a problem is when being used as a message broker.

The database is rarely designed to deal with messaging features, and hence is a poor substitute of a specialized message queue. When designing a system, this pattern is considered an anti-pattern.

Here are possible drawbacks →

- ① Polling intervals have to be set correctly. Too long makes the system inefficient. Too short makes the db undergo heavy read load.
- ② Read and write operations heavy db. Usually, they are good at one of the two.
- ③ Manual delete procedure to be written to remove read messages.
- ④ Scaling is difficult conceptually and physically.

Disadvantage of message queue:

- ① Add more moving parts to the system.
- ② cost of setting up the mq along with training is large.
- ③ It may be overkill for a small service.

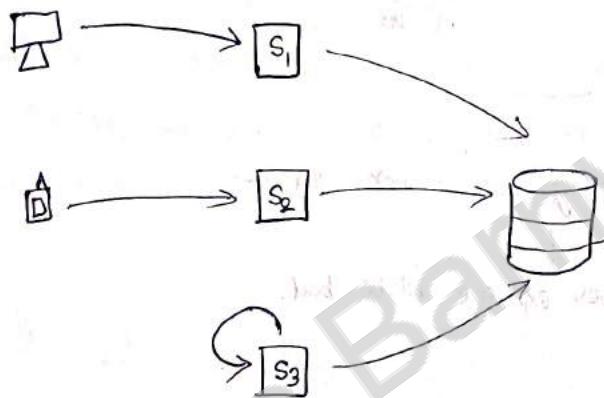
→ Using databases as message queues

Anti-pattern:

so, if you are using a database to send messages from one server to another, that's considered an anti-pattern.

so, you should avoid that.

(g)



We have S₁, S₂, S₃ servers. They are connected to desktop clients, mobile clients. And, they talk to each other using this database.

When S₂ wants to send a message to any other server, it goes to the database and then puts in an entry i.e. an insert command.

Through these inserts, you are telling the database that there needs to be some communication with some other server.

How will the database actually tell it to other servers?

→ The database can't talk to servers, it can only receive. So, the server has to poll this database at a specific interval (lets say 5 sec, 10 sec, etc) and ask it "then -

"Is there any new message for me"? If no, it will just read the message.

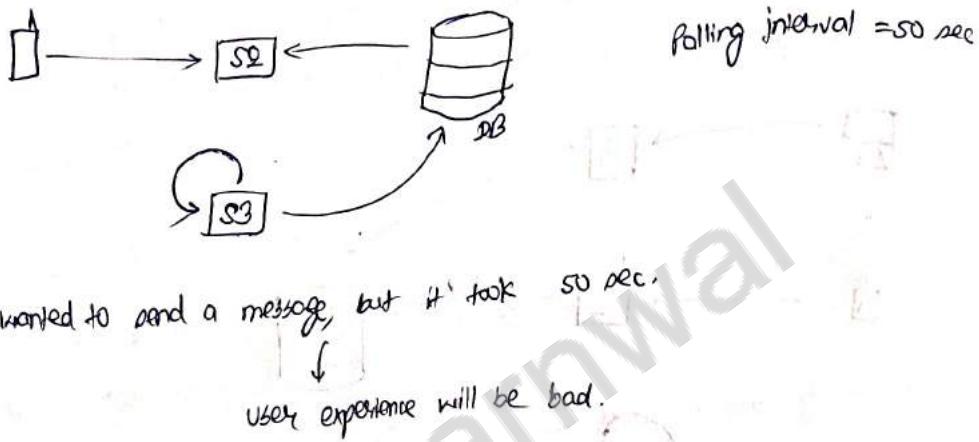
Here, db is just being used as a queue.

Polling a database is actually a problem. Bcs when you are polling the database, you are actually doing a lot of read operations on it. And, you are polling it very frequently, you are actually putting a lot of load on db i.e. continuously asking "Is there any message for me"?

Frequent loading on db leads to load on db.

- If you poll at long intervals, maybe our mobile application wants the message in 2 to 3 sec, but let's say poll interval is 50 sec.

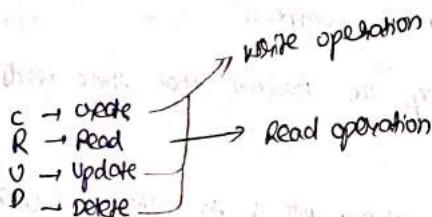
Eg:



So, long intervals are inefficient.

- A database is optimized for either reading or writing, not both. When you are having these both operations being performed in the same database, it's going to have issues like locking, deadlock, etc.

- If you have a RDB model



They are specially designed such that the read and write operation will both be fast.

- What if your servers are talking to each other very frequently?
What if there are a lot of messages these servers are sending to each other?

So, these db will get filled up with these entities.

Either you need a lot of space to store all that data or you can start deleting that data.

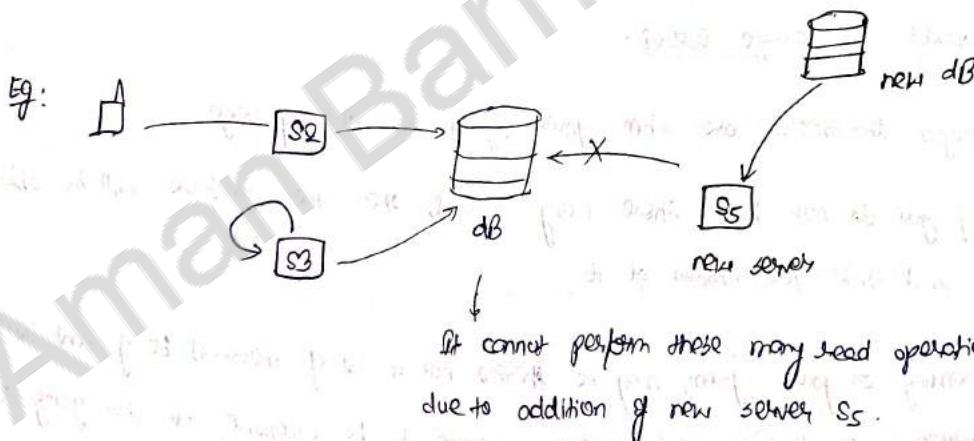
If you are not going to delete that data, you need to update it and mark it then completed.

↓
That's an expensive operation.

→ The other way is to manually delete those entries or use a cron job. In both cases, the deletion is going to use scripts using complicated logic.

→ The final issue which is a very big issue for large systems is scalability.

Eg: if we have a new server S5, and when this server starts asking the database for messages sent by other servers. This database is going to be unable to handle that load.



One thing I could do is actually add a new dB which is going to talk to new server S5.

But how does S5 talk to S2?

→ You could have a broker in between and then they could talk to each other through this broker.

Message queues
But there's so much of complication, when you have a system design concept (a data structure) which is going to solve this and that is "Message queues":

→ So, the anti-pattern here is that we are using databases as message queues. Instead, we should use specialized message queues, when we have a very large system.

Message queue is going to avoid all these of above problems.

→ The polling interval will not come into consideration bcz the message queue will push the message to the other servers.

So, instead of the server asking, the message queue is going to give the message. So, it's not going to be inefficient in that way.

They are optimized in such a way that there is not so many read operations. So, writing to it is easy while reading is really the message queue responsibility.

And, the other thing about scalability is also going to be taken care of because if you need more message queues, you can just add them.

2 Drawbacks of Message Queue:

① The major drawbacks are when your system is not very large.

bcz if you do not have those many servers, then the database will be able to handle the load that you throw at it.

② Depending on your system, may be there's not a lot of messages being sent from the systems. In that case, writing is going to be expensive, but it's going to be few while reading is going to be a lot, but you can optimize your database for read operations, and still work.

③ whenever you introduce a new concept to a system, it requires training and time.

And may be, this system is not worth the effort.

Also a call that you have to make, but may be this database is a perfect message queue for you.

Note:
so, have a look at the drawbacks and pros, & then decide whether you want a message queue or a db.

→ But if you are asked in an interview, then don't pitch off with the database without asking what kind of scale is this system supposed to handle.

so, lots of times in system design interview, they are asking you to scale the system to handle a lot of users.

If you're using a db, then that's an anti-pattern.

Keep that in mind, it's not always bad, but for large systems, the db is

→ In a system design interview, it is important to be able to reason why or why not a system needs a message queue.

These reasons allows us to argue on the merits and demerits of these approaches.

→ In general, for a small application, databases are fine as they bring no additional moving part to the system.

For complex message sending requirements, it is useful to have an abstraction such as a message queue handle message delivery for us.

Content Delivery Networks (CDN)

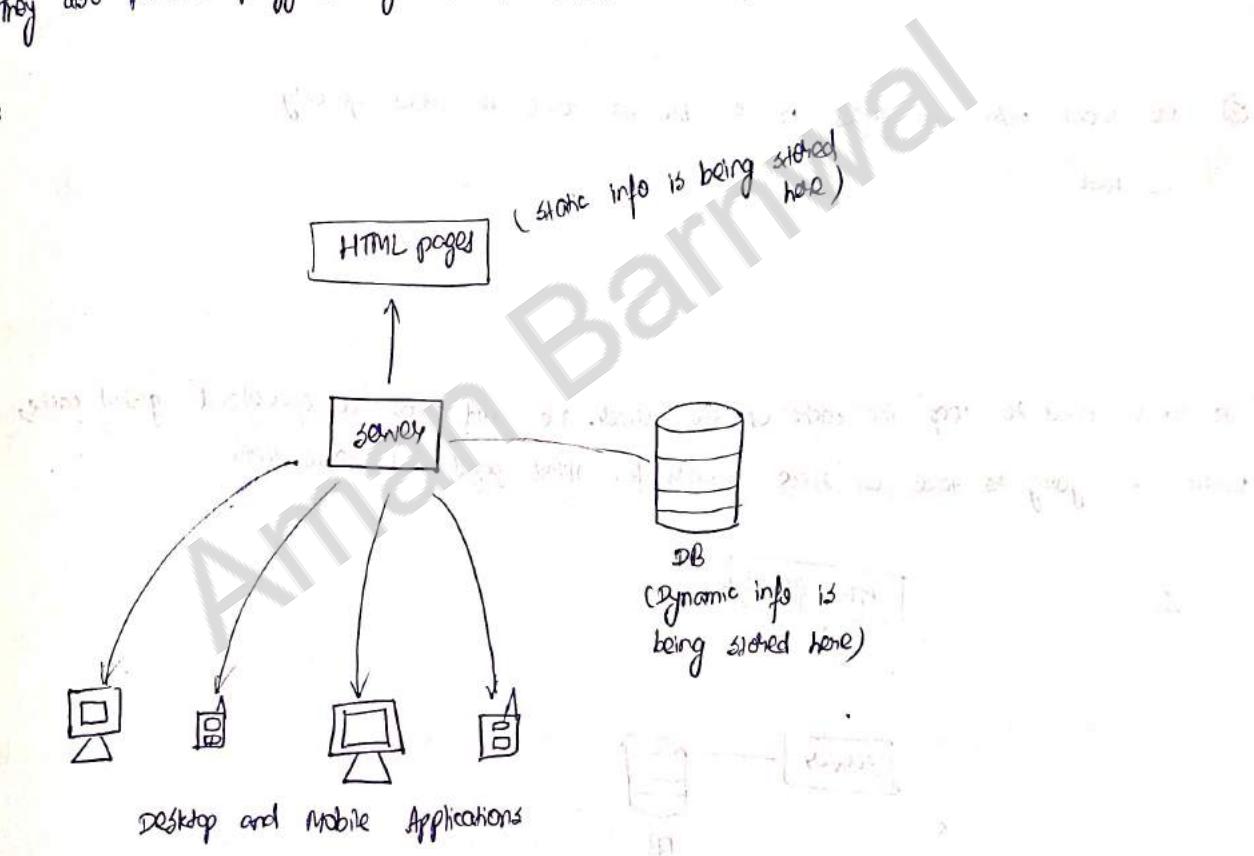
content delivery networks are a bunch of servers spread across the globe to serve information. These networks are available on internet to deliver static content quickly to nearby users.

→ some examples of CDNs are Amazon cloudfront and the Akamai CDN.

They are relatively cheap to rent and have high availability.

They also provide pluggable algorithms to invalidate and fetch data.

e:



① → Data path is long i.e. there's a lot of distance that will be travelled every time a user is actually calling for a single page.

↓
so, we can cache static pages onto ~~another~~ ^{the} server so that we don't need to make this call everytime.

② You may have different kind of HTML pages that we want to send to different types of services devices (Desktop, mobile, tablets, etc).

so, there is customized data that you want to send depending on the device or location.

Eg:
Let's say, you get 5 combination for devices. And let's say, there are 200 countries.

$$\text{ie } 5 \times 200 = 1000$$

so, if you are serving all 200 countries with unique web pages, then you have already got 1000 unique data points that we will be serving.

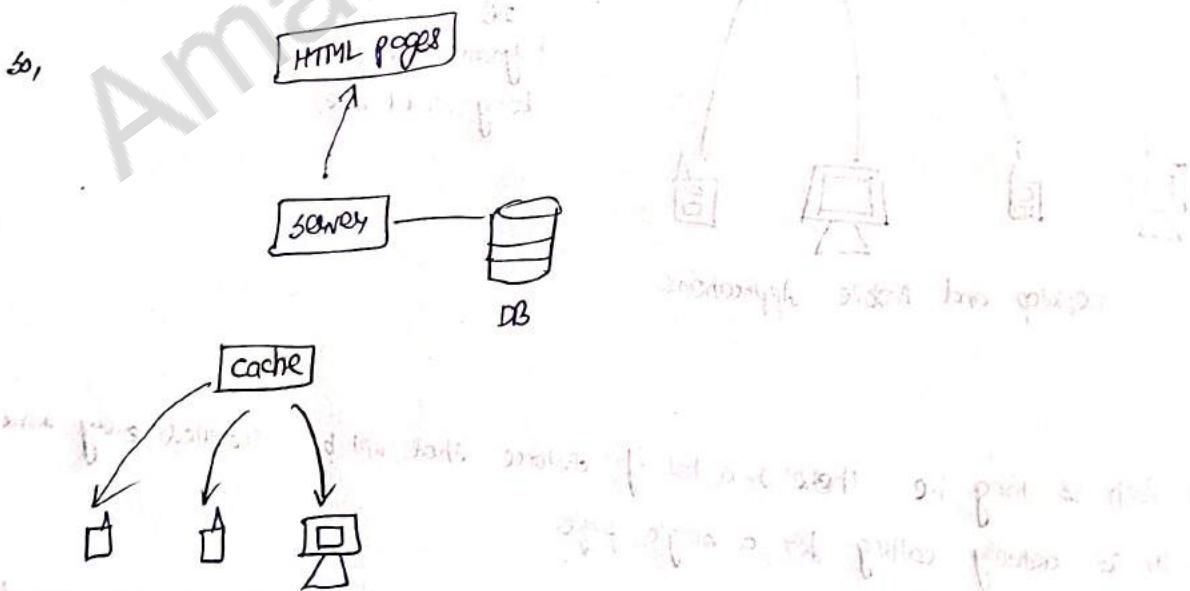


1000 is a manageable no so we can keep it in cache.

- ③ We want our web pages to be able to serve to users quickly.
i.e. fast.

Ex:

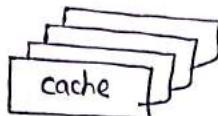
We don't need to keep the cache on the server. We could keep a specialized global cache, which is going to take all these requests for HTML pages and serve them.



So, all the requests for getting an HTML page / static content are being now dedicated to this cache, and all other requests (for dynamic data) are being sent directly to the server, which then consults the DB to retrieve dynamic info and send it back.

→ There is one possible problem with this cache.
This cache is a single point of failure. So, when this crashes, the whole system collapses.

↓
So, in order to remove this problem, we can make it a distributed cache.



Distributed cache

→ Let's assume that distributed consensus is no longer the problem. You have Paxos / Raft to take care of that.

But then we'll have some issues.

Eg: If you take users only from India, then let's say they are only concerned for 5 pages. Indian.

And we are storing a lot of irrelevant info for some countries in this cache.
For no reason when it comes to Indian users.



So, we can horizontally shard this cache based on location / country of the user.

i.e. Horizontally partitioning of a cache



Horizontal sharding based on country / location
of the user

However, there is still a problem.

The problem is that you have the distributed global cache in your cloud, which is probably hosted in a single country (lets say, USA).



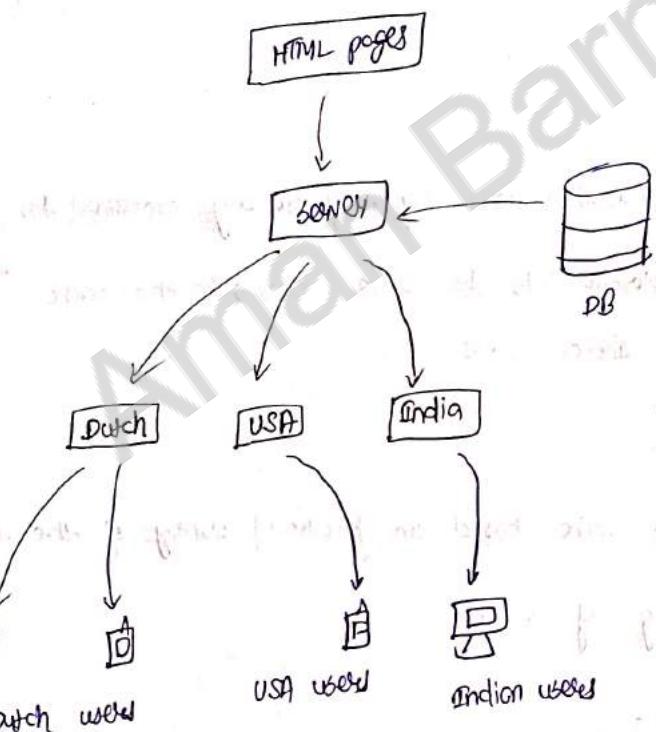
So, the Indian users are being sent to a group of nodes (which are caches).
But it's not still serving the purpose.

Eg, the Indians are still sending the message which is travelling all the way to us and coming back to them just to get a silly HTML page.

↓
So, you don't want to do that.

You want to rather construct a data centre in India, so you can create a cache here in India, through which Indian users can access the HTML page.

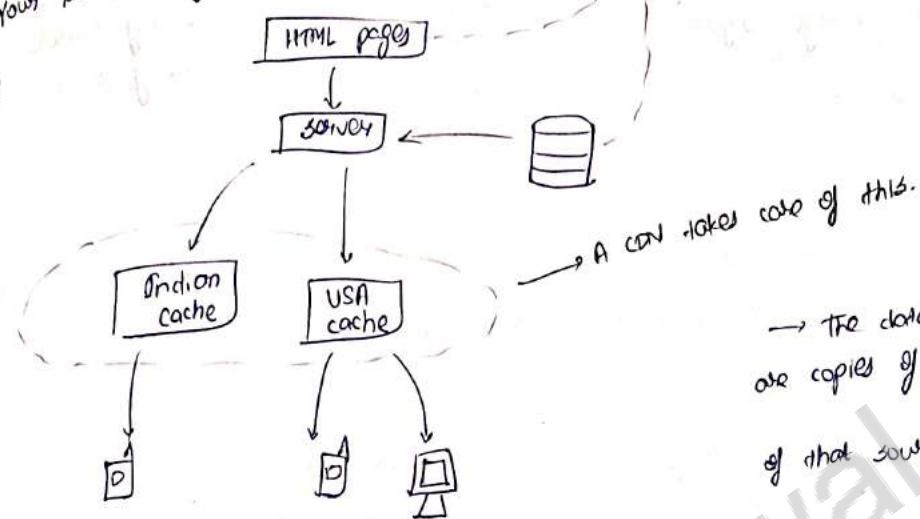
↓
So, the request and response times are quick. The latency is less which means that the system is now fast.



Now, this customized data is being stored at relevant places at the location of that data.

on any distributed system, there needs to be single source of truth.
Your users always have access to that source of truth.

Eg:



insert once, remove never → That's not practical bcz if you're going to change the page for a particular country, you have to remove that page.

By
At least, you have to version it, and then put it out of service.

As a small company / startup, how do you afford to build the caches in different countries?
How do you make sure that the regulations have been followed?
" " that it serves the latest content?

↓
If you have a look at certain specialized locations like Akamai, what they provide is
CONTENT DELIVERY NETWORK

CDN specialized in

- ① Hosting boxed data to the user.
- ② Making sure that they follow the country's regulations
- ③ Allow posting content in the boxes via UI.

most of the caches have the time to live also.

Eg: sometimes, you want data to exist only for 60 sec / 30 min,
you get a nice UI to set this up

→ The most popular CDN is S3's "CloudFront" from Amazon.

bcz CloudFront is → supercheap
→ very reliable
→ easy to use

S3 is the storage engine.

me

the

EV

cl

H

32

11

9

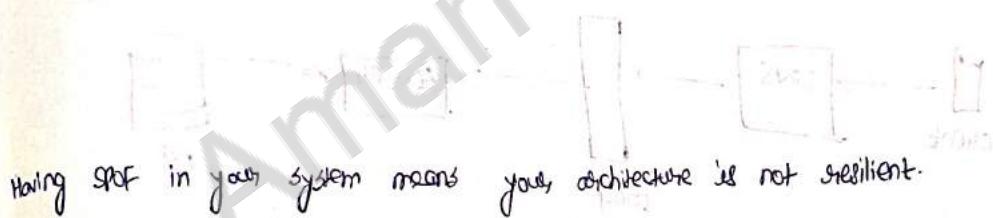
How to avoid a single point of failure in distributed systems?

→ One way to mitigate the problem of SPOF is to use multiple instances of every component in the service. The graph of dependencies then becomes more flexible, allowing the system to fail-safely switch to another service instead of failing requests.

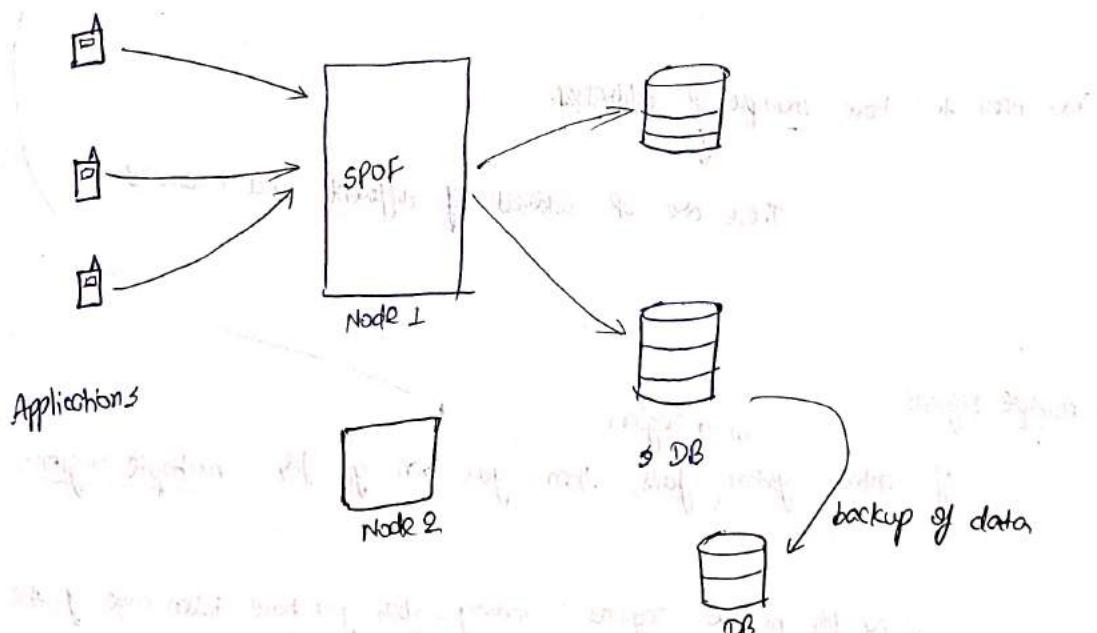
→ Another approach is to have backups which allow a quick switchover on failure. The backups are useful in components dealing with data, like databases.

Allocating more resources, distributing the system and replication are some ways of mitigating the problem of SPOF. Hence, designs include horizontal scaling capabilities and partitioning.

→ CAP theorem doesn't allow removing SPOF if perfect consistency is required.



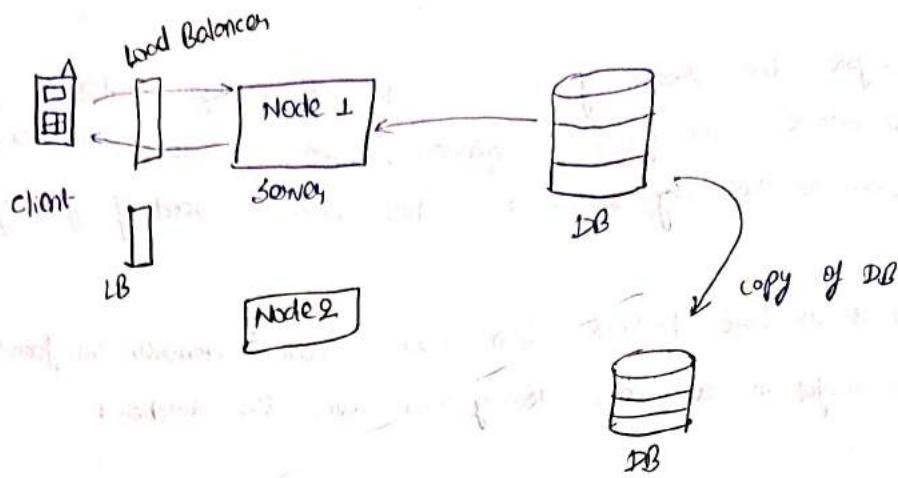
Having SPOF in your system means your architecture is not resilient.



The easiest way to mitigate SPOF is to add another node.

① More nodes

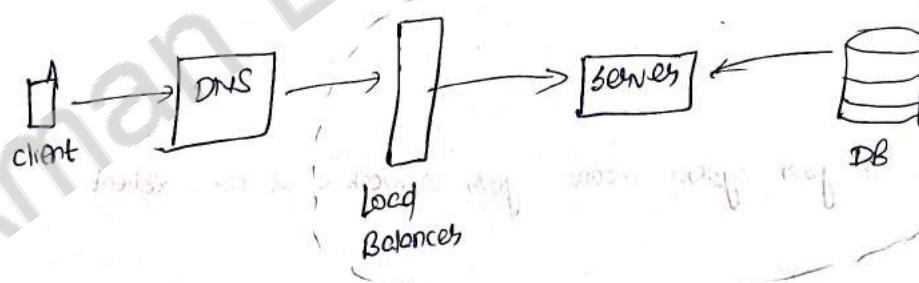
Eg:



② Master-slave Architecture

③ Multiple load balancers

The client may not know which load balancer to connect to. So, use DNS in both client and load balancers.



You need to have multiple IP addresses.

These are IP address of different load balancers.

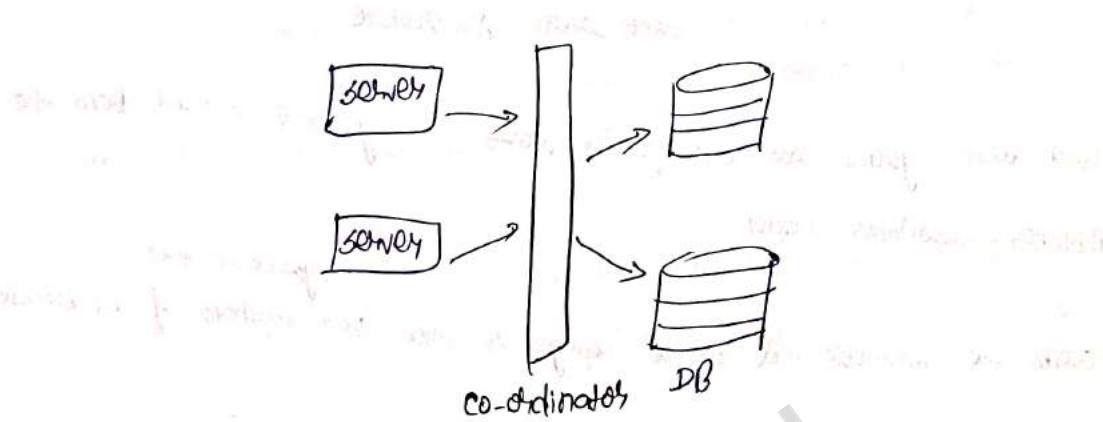
④ Multiple Regions:

in a region

If entire system fails, then you can go for multiple regions.

Going for multiple regions is mainly after you have taken care of the entire process of getting rid of single point of failure through DNS, load balancer, multiple nodes, etc.

sometimes, when you are looking for a distributed reads or writes in the database, you can have a coordination between server and DB to get rid of SPoF.



Note:

Netflix ~~is~~ mitigated "single point of failure" really well. They do chaos monkey. It randomly goes on production and takes down one node just to make sure that your system is really resilient and distributed.

What is an Event Driven System?

87

Event driven Architecture

Event driven systems pass and persist events. They have evolved from the Publisher-subscriber model.

→ Events are immutable and can be replayed to allow the system to take snapshots of its behaviour.

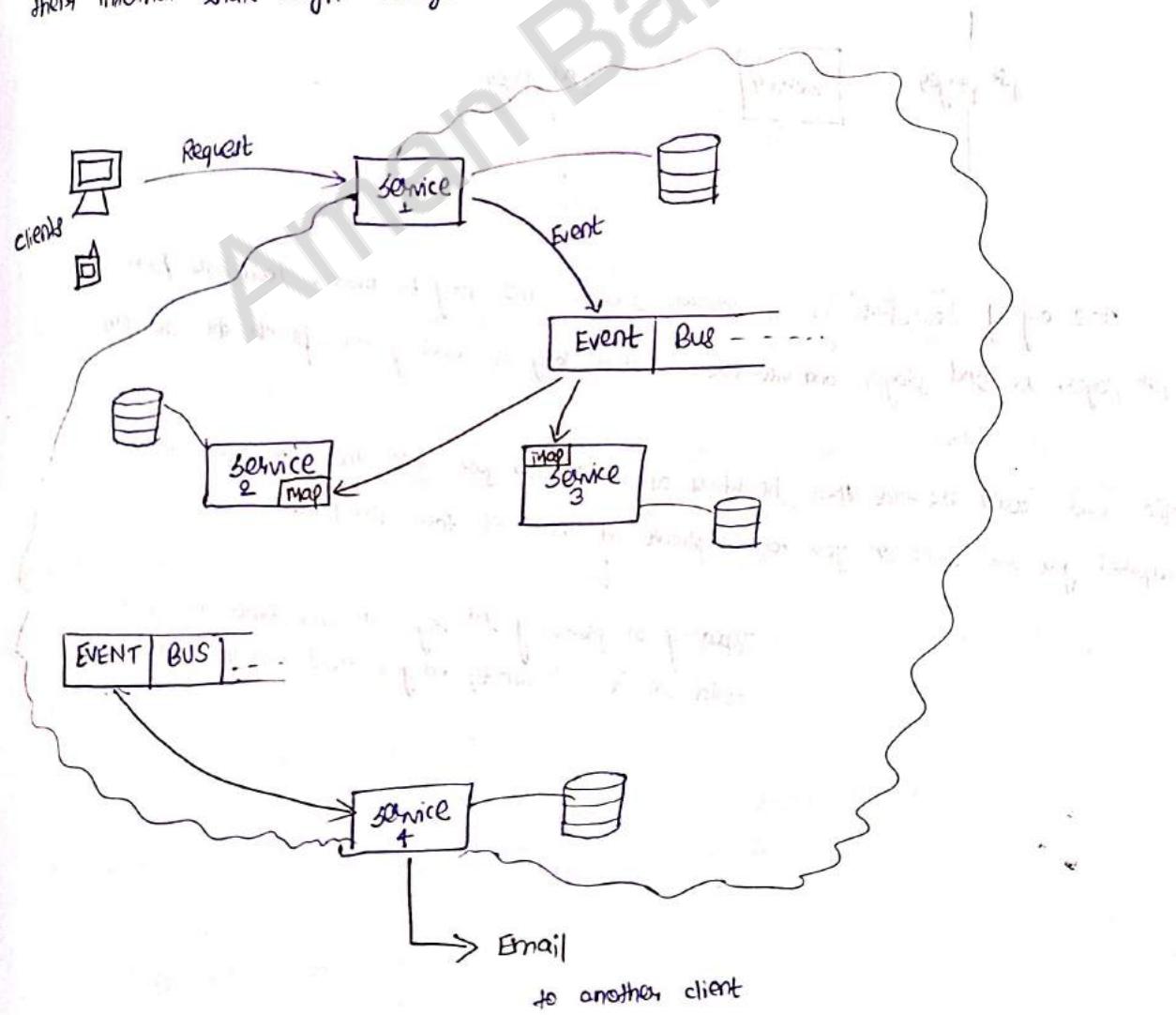
Aman Barnwal

→ The main difference b/w request-response architecture and event driven architecture is that :-

You have requests sent by clients to our gateway services. But internally, these services will never interact with each other directly. Instead, they use events to state that "Yes, something has changed".

so, if a service is sending an event to the event bus, it's just saying that 'something has happened which concerns me'. And if it concerns anyone else, you can consume this event.

All subscribers and producers consume this event and see if this event is relevant or not to them. After that, they might create events themselves bcz their internal state might change.



Each of the services stores data they are getting from event bus.

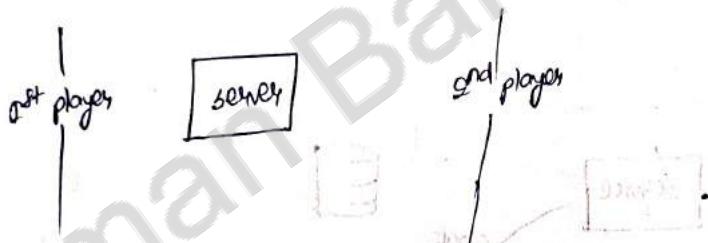
* The most popular Event-driven architecture is **git**.

Eg: git uses events (commits) to go through its history.

- ① git
- ② React
- ③ Node.js
- ④ Many game servers
- ⑤ Smalltalk

These used event-driven architecture.

Eg: use-case of Event-driven architecture :-



Take e.g. of headshots in a counter strike. There may be miss in headshots from 1st player to 2nd player and vice-versa due to delay in reaching the info to the servers.

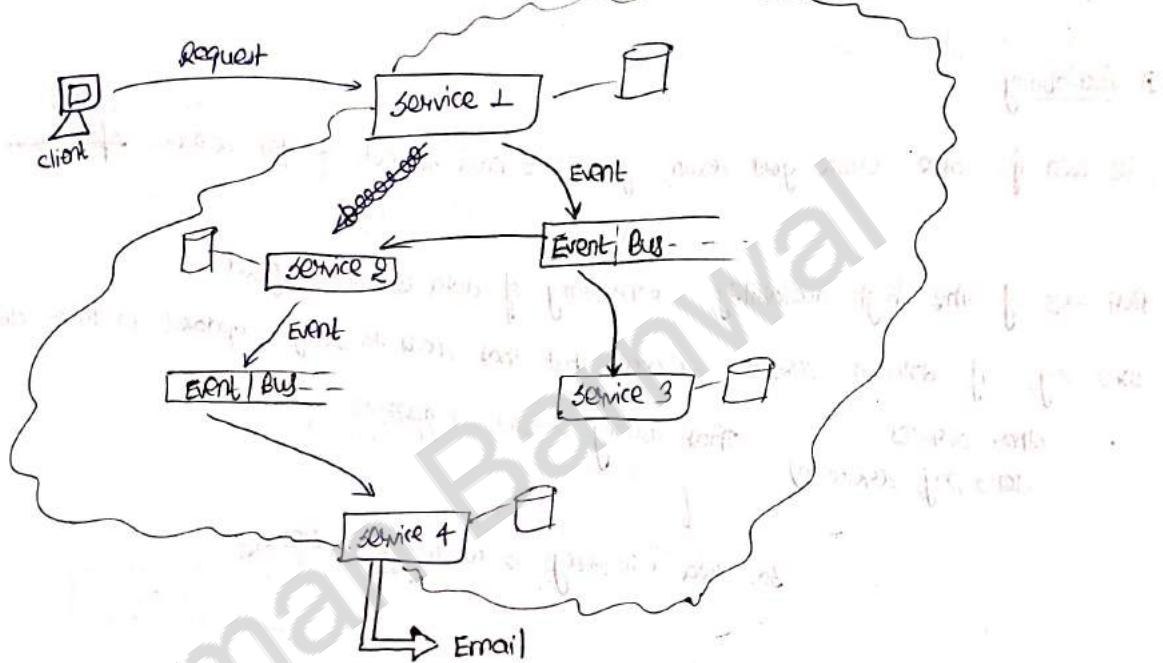
The server could take these headshots as events. You just take the timestamp, whenever required, you can undo or you can replicate all the ones that you had.

Depending on position of 2nd player at that timestamp, you could call it a headshot, and you could win the game.

→ In a system design question, there are very specific areas where you might find event-driven architecture being used useful.

If you are not able to get this architecture working for your model quickly enough, don't pursue it.

Eg:



Let's say, service 1 gives an request and sends an event which service 2 gets. When service 2 gets this event, it actually stores it in its own local database.

Each of the services stores data they are getting from event bus. ~~This is not compulsory~~ This is not compulsory necessarily. The event bus can also store all the data but you usually want to pass this persistence requirements to the services which are actually consuming this event. So then the event bus can be free i.e. it can get rid of the events. It does not even need to persisting them in the right way.

The services can store these events by adding additional fields which are relevant to it or by removing some of field "not".

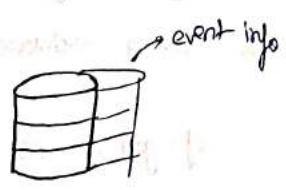
But it stored them in the local db.

↓
So, even if service 1 is not working, service 2 don't have to ask relevant information everytime

(This is different from standard microservice architecture that we have where all the services store data only relevant to them).

→ Here, in event-driven architecture, you are storing data which is relevant to them but also coming from other services.

so, the database is actually storing event information.



Advantages of Event-driven Architecture:

① Availability

ie even if some service goes down, you don't need to ask it for relevant information.

But bcs of this high availability, consistency of data is not so good.
Let's say, if data in service 1 changes, but that needs to be get updated in local db of other services. That usually doesn't happen.
(lets say, service 2)
So, data consistency is a major problem.

consistency means all of your data across all your services being the same.

② Easy Roll back:

If you have logging data of all of your events, you can actually move to any point in history using this event log.

so, if you know any bug which came after timestamp T, just go to that timestamp T and run event one by one and you can actually debug it even for production system.

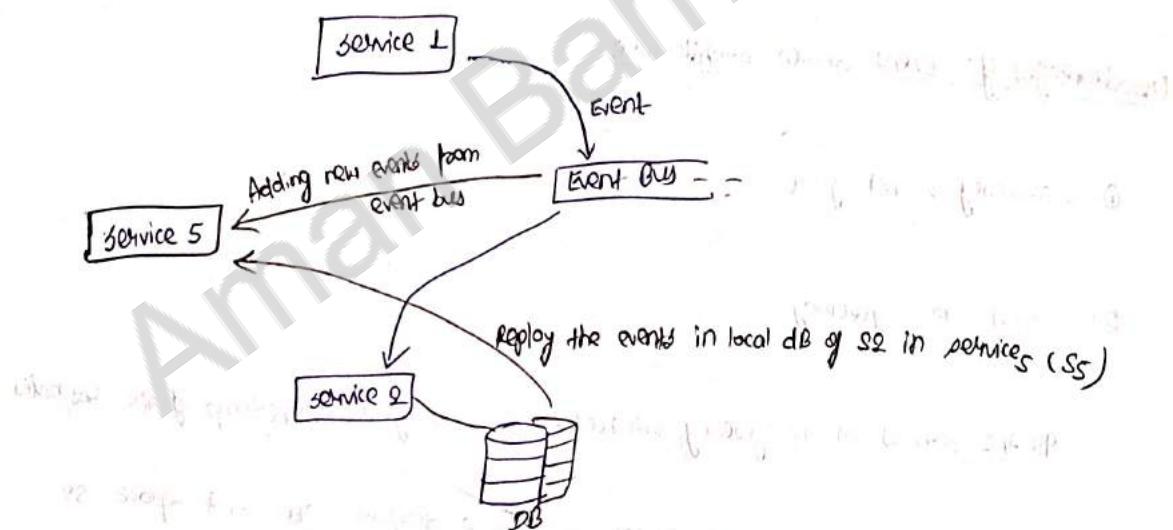
③ Replacements :

Eg: Let's say, you write a new service S_5 which needs to replace service 2 (S_2). All you need to do is to ask the event bus to send the new events.

After that, you take all events from timestamp 0 to current timestamp, replay those events in S_5 .

Then S_5 will be consistent with S_2 , & since you already accepted the new events from event bus

So, S_5 can replace S_2 very easily.



In this way, S_5 can easily replace S_2 since the data becomes consistent.

④ (It gives you transactional message guarantee) messages when sent to any service are either at least one or at most one in this event driven architecture.

of it is atleast one, then you send the message & you don't care if it reaches or not.

Eg: A welcome email

"atleast-one", then you wanted that the message should reach definitely.

Eg: An invoice email.

Here, you will be setting event bus logic in case the service doesn't get it.

By again, try again but definitely send the email atleast once.

⑤ States invents:

when you are storing these events which is basically invariant data to you, you are also storing the intent of the event.
ie why did you change the data?

Bcz, let's say tomorrow if you have a new service s_5 , you can actually look at the intent of the data and make changes in your code such that you have a completely different state in your s_5 .

Disadvantages of Event-driven architecture:

① consistency is not good.

② N/A to gateway

All the services on the gateway will need to be storing the timestamp of the responder

s_4 is sending email to external services in previous diagram. You can't replace s_4

with any new service just in the way you replaced s_2 with s_5 .

Bcz when we are sending emails, we are dependent on responses from external systems and if that response is dependent on time, then when we are trying to replace this service by replaying all those events, those events will get different responses

so, its behaviour will change without us actually wanting it to.

② Lesser control:

Will not giving you too much control.

i.e. we can't fine tuned the handling of these events.

Eg: whenever you are saying sending an event to a event bus from S₁ and then it's sending it to S₂ and S₃. We really don't have that fine-tune control of a request and a response.

→ maybe you want some services to consume your events and you want to stop some services from ever touching some events, which brings an additional layer of complexity in the event bus.

④ Compaction:

If it's adding all the events one by one, and you need to get to a particular point sometimes only

There are 3 ways:

(i) Replay from start

(ii) Diff-based. In this, you take 1st event and then just store the difference.

(iii) Use undo upto a particular point.

But some operations can or can't be undone.

Eg: Addition
Subtraction

Eg: sending an email

You can squash all the events upto a particular day instead of a particular timestamp.

⑤ Hidden flow:

It's difficult to reason about the flow of the system.

Eg: Just looking at S₁ will tell you that it publishes an event to an event bus, but that's where code stops. But you don't know, what happens with that event.

For this, ↓
you need to go to subscribers of this event in the event bus.

∴ The flow of program is not easy to understand.

⑥ Migration:

Migration to some other archi is not only

Note:

We can think of services in event-driven architecture to be having some sort of log of their events and using a publisher-subscriber model to pass and consume those events.

Event-driven Archi

All the services in event-driven archi...
publish events when they feel like
someone needs to know something.

Request-response archi

While in request-response archi, the services ask for something - it might be data or service.

Introduction to NoSQL Databases:

NoSQL is a popular database storage method. It keeps data as key-value pairs.

→ There are equally high no. of pros and cons where NoSQL is not used.

These companies don't use NoSQL databases:
YouTube
StackOverflow
Instagram

They use NoSQL for analytics. Their core functionalities run on RDBMS though.

It's important to know when to use NoSQL db and when not to use them.

It's not true that scalability always demands NoSQL database.

SQL:
① In case of SQL database, when you are running your query, usually the pointer comes to one attribute (let's say, id) then it has to sequentially read all the columns.

This database is also not suitable for denormalizing things.

↓
so, you might need a join, which is pretty expensive, considering that most of the times, you need both data.

② Even if one of the column is NULL, we still have to add a new column.
(let's say, distinct)

ie. it's not flexible

Also, you'll pay if you need to add one more column of salary, which is a very expensive bcz you need some kind of locks on the column and it's also sticky.

→ to maintain consistency at this time.

③ ACID properties are guaranteed.

④ Read times are comparatively faster than NoSQL.

Eg:

SOL

ID	Name	Address	Age	Role
123	Aman	893	25	SDE 2

Join

Address	City	Country	District
893	Bogra	India	Bogra

foreign key mapping here

NoSQL

ID	value
123	<pre> { "name": "Aman", "address": { "id": 893, "city": "Bogra", "country": "India" }, "Age": "25", "Role": "SDE 2" } </pre>

→ This is JSON data

Note, address is no longer
a foreign key (compared to SOL)
But it becomes another object
within this object

kind of JSON nesting

Advantages of NoSQL

What makes NoSQL so efficient?

① When you are inserting, there is usually all the fields inserted together.

When you are pulling out any information about any user, usually you need all the info about that user, which means you need to pull whole blob.

i.e. Insertion and retrieval require the whole blob of data

② → It's cheap here in case of NoSQL database.

Because all the relevant data is contained together in one block, so it's a little easier to insert and retrieve.

② Even if address is NULL, we don't need to add this in JSON structure.

Eg:	ID	value
	103	
		"name": "Aman"
		"age": 20, "25"
		"role": "SDE 2"

i.e. schema is very flexible in NoSQL database.

i.e. schema is easily changeable.

③ NoSQL database have inbuilt horizontal partitioning, i.e. it is built for scale.

Most of the times, they expect a lot of scale to come in.

e.g. it is allowing horizontal partitioning, so it's more focussed on availability

④ They are built for aggregations also.

i.e. They are usually expecting to be getting some important info out of that data.

Eg: Avg age,

Total salary, etc.

∴ These kind of databases are built for finding metrics and getting intelligent data.

Disadvantages of NoSQL:

- ① Not too many updates are inherently supported in this i.e. NoSQL is not built for too many updates.
- ② consistency is a problem here which means that the ACID properties are not guaranteed.
If ACID are not guaranteed, you cannot have transactions using NoSQL database. That's why, financial systems don't use NoSQL database for their transactions.
- ③ NoSQL database is not read optimized.

Eg: If I asked you to find age of all employees that we have in a company.

↓
It's going to go to the blocks and each time, it's going to read entire block, then filter out the age and then do that for every row and then return you the result.

id	value
123	{"name": "Aman", "age": "25"}
124	{"name": "Anku", "age": "23"}

while in SQL database, all you need to do is → just go to that column & then return the age of all employees

i.e.

Read time for NoSQL database are comparatively slower than SQL database.

- ④ In NoSQL, we don't have implicit information about relations. i.e. Relations are not implicit.

(Here, you can't force a constraint like foreign key constraint which in SQL says that this column 123 only exists only if there is corresponding column in the employee table.)

Eg: In SQL

name	address
Aman	123

address	city
123	Gurgaon

employees

⑤ Joins are hard in NoSQL database.

↓
They are all manual here.

(But SQL database is built for joins bcz they have an inherent relations b/w them.)

When do we use NoSQL database?

- It depends on if your data is a block and if you are making few updates and if you want to keep all of them together and if there will be a lot of writes coming in.

↓
Then NoSQL database would be good to go.

- There might be scenarios where you might want inherent redundancy or aggregation in the data.

↓
Then also NoSQL would be good to go.

- There are still some disadvantages of NoSQL bcz of which some organization like YouTube, StackOverflow, etc. still don't use NoSQL database.

Like YouTube, StackOverflow, etc. still don't use NoSQL database.

SQL → also called relational dB
It stores info in tables.

①

② ACID properties are guaranteed.

③

④ Read time are faster than NoSQL

⑤ Relations are implicit here.

⑥ Built for joins.

⑦ Not flexible.

some popular SQL dB are →

NoSQL → also called non-relational dB.

DisAdvantages

① Not built for too many updates.

② ACID properties are not guaranteed
i.e. consistency is a problem here.

③ This dB is not read optimized.

④ Read time are comparatively slower
than SQL

⑤ Relations are not implicit here.

⑥ Joins are hard in NoSQL.

Advantages:

① so efficient

② cheap

③ schema is very flexible in NoSQL.

④ Inbuilt horizontal partitioning.

NoSQL is built for scale.

⑤ Built for aggregations.

⑥ Extensive control over availability.

NoSQL stores info in non-tabular form.

some popular NoSQL dB are →

→ Apache Cassandra

→ CouchDB

→ Apache HBase

→ Neo4J

→ MongoDB

→ Amazon DynamoDB

Cassandra is one of the most efficient and widely used NoSQL databases.

- It offers → highly available service
- → No single point of failure
- → It can handle massive volume of data
- → It can effectively and efficiently handle huge amount of data across multiple servers.
- → It offers users "blazingly fast writes" without affecting the read efficiency
- → It is just as fast and as accurate for large volumes of data as it is for smaller volumes.
- → It is horizontal scalable i.e. it allows users to simply add more hardware to accommodate additional customers and data.
- → Flexible data storage : It can handle structured, semi-structured & unstructured data, giving the users flexibility with data storage.
- → Flexible data distribution : It uses multiple data centers which allows for easy data distribution wherever & whenever required.
- → It supports ACID properties.
 - Atomicity
 - Consistency
 - Isolation
 - Durability

→ Cassandra stores the data in key-value pairs
but MongoDB stores data in form of document

→ Cassandra is good for real time data processing.
MongoDB is good for structured data.

→ Cassandra is good for reading, it can read thousands of rows at once.
MongoDB is good for writing, it can write millions of documents in a second.

→ Cassandra is good for distributed systems.
MongoDB is good for mobile devices.

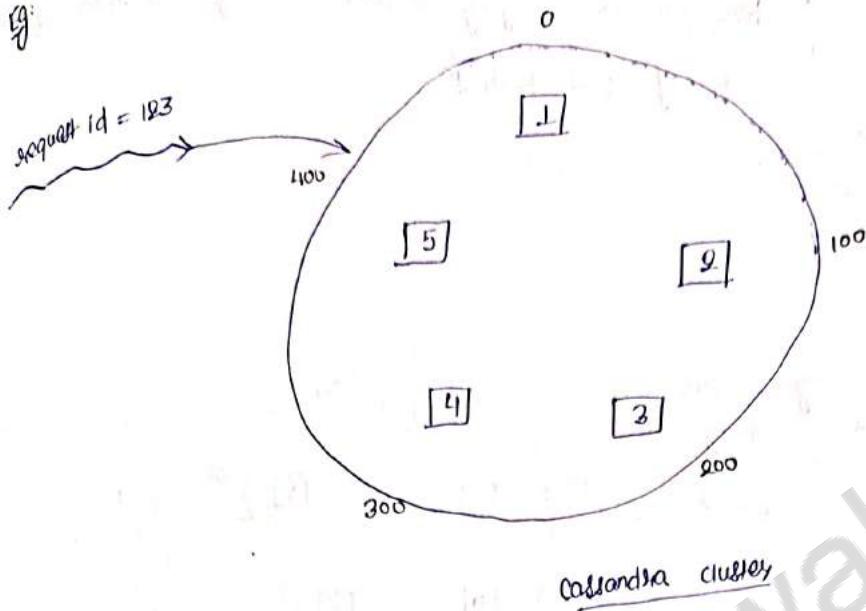
→ Cassandra is good for real time data processing.
MongoDB is good for structured data.

→ Cassandra is good for reading, it can read thousands of rows at once.
MongoDB is good for writing, it can write millions of documents in a second.

→ Cassandra is good for distributed systems.
MongoDB is good for mobile devices.

→ Cassandra is good for real time data processing.
MongoDB is good for structured data.

[0-499]



→ It's actually very expensive to host a cassandra cluster.

→ The above cassandra cluster is having 5 nodes.

Any requests betn 0 and 100 will be falling in node 1.
" " 100 and 200
" " 200 and 300
" " 300 and 400
" " 400 and 0

request id may not always be a numeric. It may be a UUID or a person's name, etc.

$h[\text{request-id}] = ?$ (map)

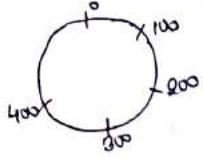
i.e. This hash is used to map this request (request-id) to a particular node in this cluster.

Eg: $h[123] = 256$
falls betn 200 and 300

* If hash fn is such that the requests are uniformly distributed, then we can assume that if there is a lot of requests which is coming in, they will be falling with equal probability on any of the nodes.

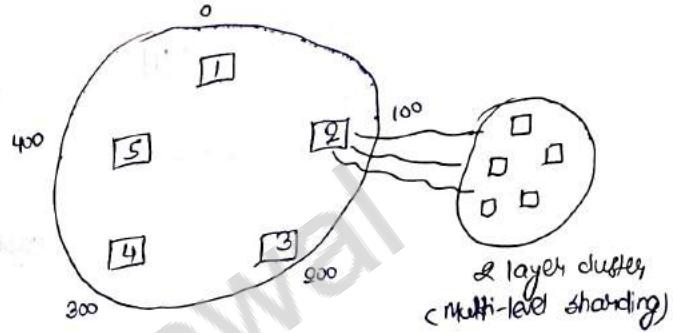
All the nodes should have approximately equal distribution of the loads.

Eg: $h(<100) = 0$ $h(>100) = 1$ { here one node is going to have a lot of load than others
 bcz of bad hash fn.



→ You can also do a 2 layer cluster.
in case a lot of requests is going
on a single node.

request →



Let's say, the request goes to 2 (let's say, it is having a lot of requests)
Then it doesn't actually store it in its database.

Instead, it sends to another cluster which has 5 nodes and you store
(lets say)

this request through a different hash fn.

Eg: if let's say, you are user of Google maps in India. Your hash fn may be on
the basis of country. Based on country ID, if you are sending at one place,
it's possible that one of the country is going to have a tremendous amount
of load for certain festivals (Eg. Diwali).

In this case, a node (India) is having a lot of loads.

So, you can go for multi-level sharding.

- * If request falls on node n, the node (n+1) must have the copy so that even if node n crashes, data will not be lost.
- * If you have 2 nodes storing the data. So, the probability of you losing the data is lower.
- * If I tell's say, your request is falling on node x based on hash fn, then it can fall into any node which is having replicated data of node x.

Your read queries are optimized.

- * Your writes are also more guaranteed and it can also be optimized - by

Eg: If request id falls on node x. If node x missed the write, then you could write it to next node (x+1).

Note:

Cassandra gives features of

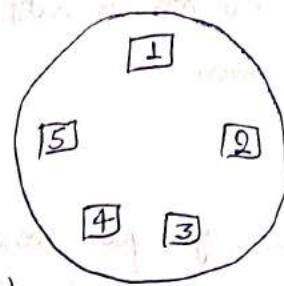
① Load Balancing

② Redundancy, Replication

This gives you the speed in reading. This gives you data guarantee.

→ one of the most important concepts when it comes to nosql database is the idea of Distributed consensus.

e.g. let's say, there are 5 nodes & replication factor is 3.



if a request falls on node 5, then (node 1, 2 and 3) is copying the data.

Let's say, I write on node 5. concurrently, I am going to write on node 1 and 2 also. However, let's assume that node 1 and 2 are little slow. They have not got the write yet.

If you are making a read operation and node 5 crashes. so, when it goes to any of its replica either 1 or 2, it returns an 404 error i.e. record not found.
(bcz writes were not completed till this time on node 1 and 2)

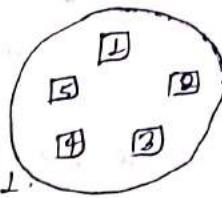
so, the user will just get confused that he made the profile and why it is not there on the database.

To avoid this issue, cassandra should be returning a database delay so that the application knows that there is something wrong in the db.

To do that, we need some sort of distributed consensus, and one of the ways to achieve this is through the quorum.

Quorum → It is a way in which multiple nodes who are related to a particular query accept a particular value.

by vote for



Eg: Let's say, node 5 crashes. The request come to node 1.

Node 1 says "I don't have this data" and node 2 says "I have this data".

In that case, it will be picking up the data with the latest timestamp (version id)

and will return it to the user.

Let's assume that now even node 2 don't have this data. So, both nodes 1 and 2 will agree that there is no profile created and unfortunately, the user will be given a "No user profile found".

Eg: if Quorum = 2

Replication factor = 3

That means if 2 of 3 nodes accept a particular value, then we take that to be the truth.

So in above case, if node 1 and 2 do not have writes replicated to them, that will result in a wrong entry sent to the user.

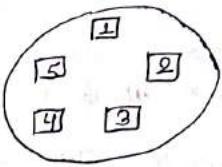
This is really rare bcz the possibility of node 5 crashing and node 1, 2 not having the writes before they get a read operation is really rare.

This is the risk that you are willing to take when you are taking a NoSQL db and just more focused with availability instead of consistency.

Eg: g) Quorum = 3

Replication factor = 3

↓
ie. 3 nodes will have to agree with the replication factor of 3.



In this case, the query will fail bcz node 5 is failed. You need 3 nodes to agree on one value. Node 1 and 2 will return a particular value but node 5 is not returning a value.

∴ The query fails.

Cassandra stands out in

→ the way in which cassandra stores data and writes data.

Eg:

so, if you have a request coming into cassandra,

request →

key	value

periodically, this memory is dumped into sorted string table (sstable)

Assume this table to be existing in memory

bcz you need to write it somewhere

Why is it sorted string ?
→ bcz the key is sorted in string table.

so cassandra will be storing all these records in memory as a log file.

whenever there is a request for some write, it's going to write in a sequential fashion. In this way, you are actually storing all the data like a log.

This is efficient because all you need to do is go to the point where you have current pointer and write down the data, instead of searching anything.

↓
This is fast

→ sorted string table is immutable i.e. this data is not going to change.

key	value

→ Every time Cassandra has some data in its memory, it flushed it into a new sorted string table.

Now, bcz these requests are coming in after few days, you will be going to have a lot of sorted string table all over your clusters which is going to take up a lot of space.

Why?

→ bcz any update

(Let's say, key is 123 and 2 days later, you got an update on that key 123)

↓
some data has been changed due to update

123	

The latest record for updated on this key is in some other sorted string table bcz that was created later on when it was flushed to SST.

Effectively, there are multiple records for same key. You can always use a timestamp for each record.

The problem is not consistency, the problem is data usage.
like you are going to use a lot of storage within duplicate keys.

Eg: If you have 10 records for the same key.
Then you are using 10 times storage required.

Cassandra and Elastic search both provide a feature called Compaction.

compaction:

We take different sorted string tables and we merge them.

Time comp: $O(n)$ operation
Space comp: $O(\min(n, m))$

where size of 2 always are n and m

→ We have SS Table which are immutable so they are really fast to flush into disk.

You don't need to worry about whether there are duplicate keys.

Later on, like a batch process, you are going to be compacting these SS tables to optimize for space.

How do you get rid of deleted records?

Time comp: $O(n)$ operation
Space comp: $O(n)$ operation

→ You go to deleted records and cassandra place a tombstone.

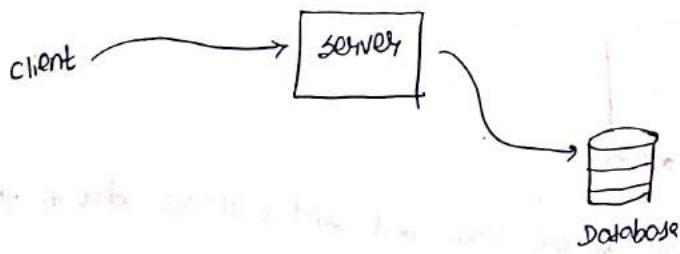
You probably set a flag and tombstone will say that this record is dead.

If there is read operation on 3 or 4 records and if there is a tombstone, you see tombstone on the latest timestamp, you call this record to be dead, and all 3 or 4 of them are killed.

If there is an update on the key and if you see a tombstone then you know that an update is impossible and if you fire an exception like record doesn't exist therefore

How database scale writers : The power of the log

We will talk about the ways in which we can optimize writes in our database.



Note, client is sending information to the server and server had to write to the dB

→ You can think of your dB as a data structure.

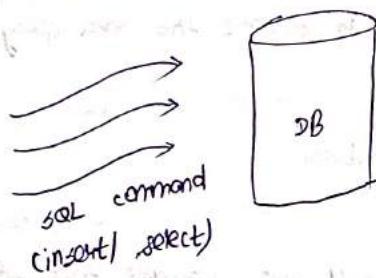
To speed up queries, the traditional database used a data structure called a B+ tree.

This is like Binary search tree without binary inside it.

Bt tree is preferred because it gives you good insertion and search time.
Both operations are $\log n$.

In Bt tree,
insertion → $O(\log n)$
deletion → $O(\log n)$

Whenever you type a SQL command of an insert or select, that maps to an insertion or a search operation in Bt tree and the underlying data structure is manipulated.

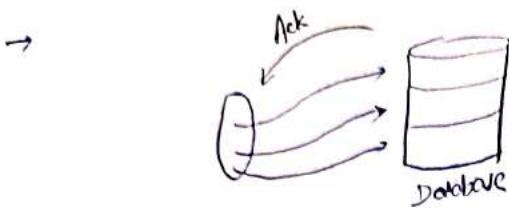


Each of these operations needs an acknowledgement from the dB that "Yes, I have successfully executed the request".

→ To scale our dB, we want to reduce the unnecessary exchange of data which are acknowledgements and headers & also we want to reduce the IO calls which will help us to free up some resources and reduce the request-response time.

Header ← Bandwidth

For I/O call ← Time k slowdowns

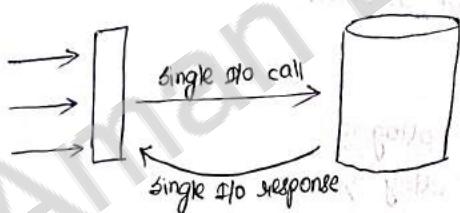


If we could condense this data into one block and send it in one shot to the database and get a one acknowledgement.

↓
That's a good idea.

That's would be more efficient use of your bandwidth bcz you are not sending any extra data that you are sending around.

Also, you are going to do less work bcz there will be one I/O call and one I/O response.



① condense data queries into a single query

The server has to give up some memory to condense this data query. That requires additional memory.

↓
This is one of the drawbacks.

The reason we do this is because of one of the advantages that we need fewer I/O operations.

is the fastest if you have a lot of write operations, what data structure which can help you write operations quickly?

→ Linked List

All you need to do is whenever you have a write, you need to just go to the end and append something to it.

write operation in Linked list $\rightarrow O(1)$

↓

bcz you just have to go to the end and you just add a node.

→ Log follows the philosophy of Linked List. we are going to use log to persist data, bcz it's really fast when it comes to write operations.

Disadvantages of Log:

Read operations are very slow.

ie in Log, if you have to find something, you can't jump to one point.

you have to search sequentially the entire log.

→ If you can think of your database as a log, that's a huge amount of data you need to read to every query a person will be sending.

ie Fast writes.

Note:

Database with heavy write workloads need write friendly architectures.

Btrees are an alternative to B+ trees, as they scale writes better.

Efficient.

Advantages:

- less writes
- fast writes

Disadvantages:

- additional memory
- slow reads

log wing linked list

→ you can't help much w.r.t additional memory bcz you need to condense the queries and keep them somewhere

→ slow reads are something that we definitely want to avoid.

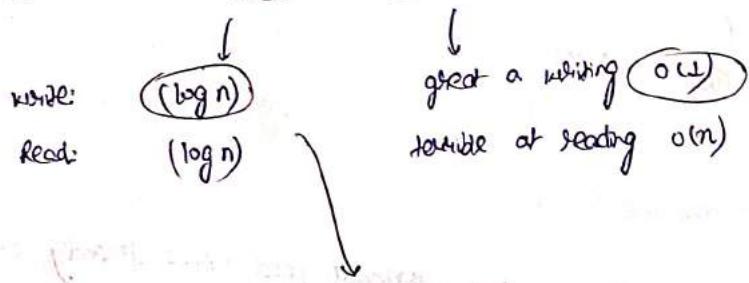
You are not reading on the log bcz searching in linked list is $O(n)$.

But if you could convert the data structure to either tree or sorted list, it will make the searching really fast.

instead of linked list, if you use sorted array, searching becomes $O(log n)$

But, you already had a B+ tree, which was giving you $O(log n)$ search time.

so what should be take - B+ tree or Linked List?



That's why, we will use (Linked List + sorted array) to get great write speeds and great read speeds

\hookrightarrow optimal speed

sorted array → search $O(\log n)$

can we convert our linked list to the sorted array?

→ You can

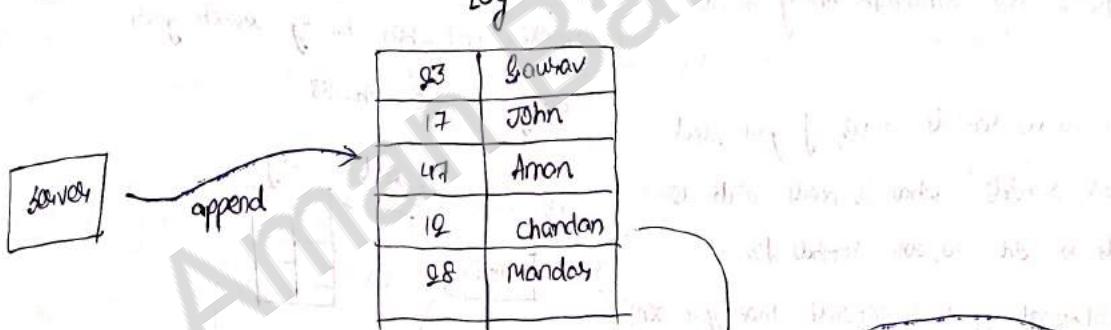
- we don't want to convert it in memory bcz that's defeats the purpose.
The whole point of having a linked list was so that you do fast insertions.

if you want sorted outputs, insertions are really slow.

so you'll do the conversion somewhere in the database

so whatever information you are getting from the clients, you sort it and then you persist it so that the read operations are super fast.

Eg:



After a threshold point

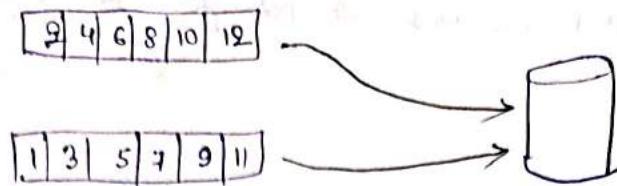
i.e. the no. of records, you can keep it in memory,
you persist this log into the database in one shot.

This data is going to be sorted before it is persisted.

You can do here a binary search which allows you to query this data in an efficient manner.

What if after the current append, you want to append the new stuff to the dB?

Eg:



When you want to persist the new append in DB, the database can do 2 things :-

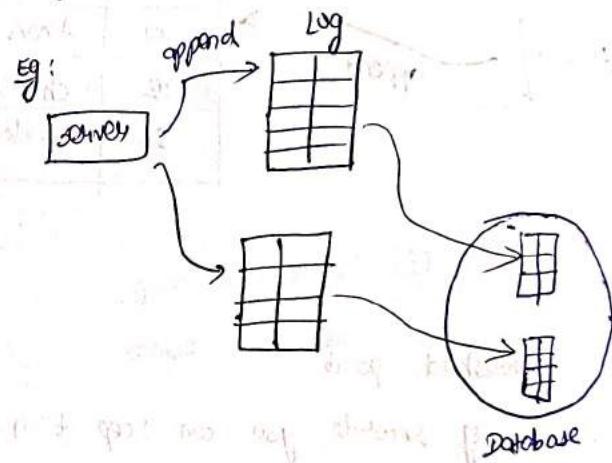
It can take both records and merge them in a sorted manner.

But, that's not smartest thing to do.

Eg:
After 10,000 records in DB, if you send 6 more records, what it needs to do is it needs to sort 10,006 records plus the additional just 6 records that you sent here.

$O(n \log n)$ worst and inefficient

Instead of merging always all the time, you can keep them in sorted chunks so that your database is going to be represented by chunks, and then binary search your query in the chunks.



do a binary search over these chunks.
here only 2 chunks

So, This is better than slowing down your write operations tremendously.

Now, your write operations are reasonably fast but your read operations are slightly slower.

If you have N insertions,

$$\text{read time} \propto \left(\frac{n}{\text{no. of records in a chunk}} \right)$$



so, your read times are extremely slow.

Eg: imagine the facebook. There are billions of records in the database.
1 chunk contains 6 records.



so $\frac{1 \text{ billion}}{6}$ still very very slow. So, how will you optimize this?

→ You can use some sort of hybrid approach which is taking some records and merging it with other records as long as the cost time is not so high.

everytime when you are getting a chunk, you are deciding whether or not you should merge this chunk with the existing chunks.



In the background, we are going to take chunks and we are going to merge them together, to make larger sorted arrays so that when there is a search query, this large sorted array can help us reduce the overall time complexity.

9	10	13	14	16	17
---	----	----	----	----	----

$$\log_2 6 = 3$$

Eg.

1	3	5	7	8	11
---	---	---	---	---	----

$$\log_2 6 = 3$$

merge

Search query



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----

$$\log_2 18 = 4$$

$$\log_2 (18) = 4$$

2	4	6	10	15	20
---	---	---	----	----	----

$$\log_2 6 = 3$$

$$\text{Time req to search in 1 chunk} = \log_2 6 = 3$$

If you have to search in all chunks for a particular record = $3 + 3 + 3 = 9$ operations

saving

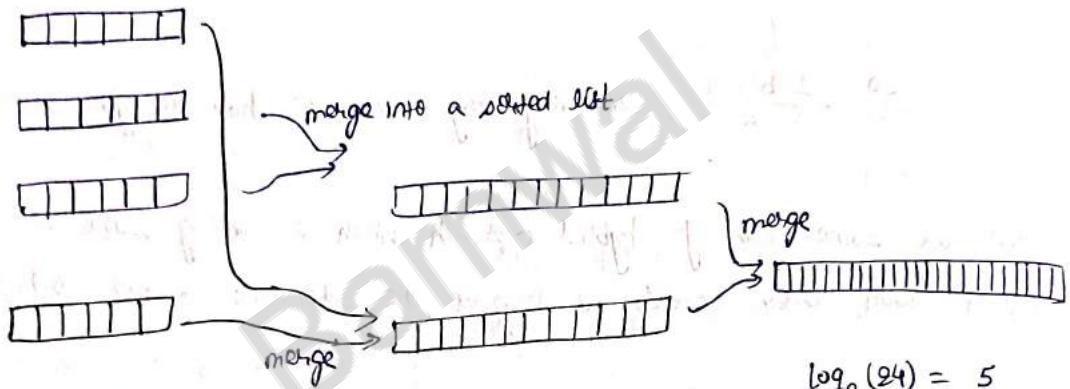
But, searching in the merged sorted array just take $\log_2 (18)$ i.e. 4 operations.

→ There is a clear saving here of the no. of operations but you need to merge sorted always and then persist it, so that has some overhead.

Let's see, what is the best approach?

When should you merge the sorted chunks?

Eg:



But, if you had them all 4 chunks unmerged,

$$\text{then no. of operations} = 4 * \log_2 6$$
$$= 4 * 3$$
$$= 12 \text{ operations}$$

if you merge all 4 into a single sorted size of size 24,

$$\text{then no. of operation} = \log_2(24)$$

→ clearly savings in the no. of operations if you sort always together.

i.e.

$$\begin{array}{ccc} 6 & 6 & \rightarrow 12 \\ 12 & 12 & \rightarrow 24 \\ & & \vdots \end{array}$$

chunk of size n

what happens with this strategy is you have a large number of blocks of varying size.

↓
so, your read operation is spread across these blocks.

, And to speed up your search operation, you can apply "Bloom filter".

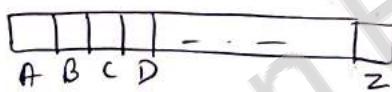
Bloom Filter:

It's a data structure which stores some data and when you query for some data, it gives you a result.

Eg: if you have a book, you have to find out whether this book has word "CAT".

If the book provides you the particular data structure "Bloom Filter"

of positions, one each for each alphabet like



If there exists any word in the book starting with letter C, then mark C block as true. i.e. 1

For CAT \rightarrow 1 for C block

Even for SONGA \rightarrow 1 (that's called a false positive)

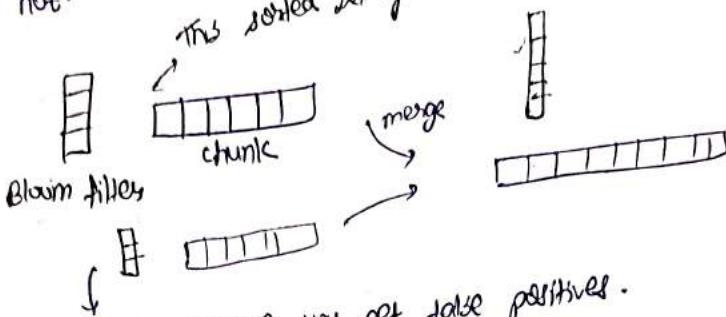
Note: In Bloom filter, you can have false positives, you cannot have false negative.

Eg: If there is no word in the book starting with S, there is no way the block C will be marked as 1.

You can think of here book or dB and searching of word as Bloom filter query.

→ If you are looking for whether this record exists in the chunk, just create a Bloom filter on top of this chunk, which tells you that this record exists in this chunk or not.

Eg:



This sorted set of records is called "sorted string table".

This speeds up your query because you get false positives. The rate of false positives can be managed by a no. of bits.

Instead of saying, whether there's a word with "c", I can say where there are words starting with "ca".



→ Also, anytime I merge two arrays, I am going to be requiring a separate bloom filter for this one merged one too.

A good idea is to create a Bloom filter of larger size. Bcz if there is a more information here, the chance of a false pos is higher.

I need to put info in the bloom filter to reduce the chance of a false pos. This way, when you are reducing the false pos, you are indirectly reducing the read times also.

log-structured merged trees

Note:
You want to speed up writer for which you use a Linked List but
you can't read well on Linked List.

so, in the background, you merge sorted arrays which can be binary
searched to get fast read operation.

so, your writers are fast bcz you are flushing writer
in a sequential fashion but your reads are also fast bcz you
are reading from sorted tables.

Compaction:

compaction is taking a lot of sorted string tables and
using merge process to condensing them into a single array so
that we can do fast search query.

Service Discovery and Heartbeats in microservice:

Service death due to various reasons like hardware faults and software bugs. Service discovery and health check are essential for maintaining a service ecosystem's availability and reliability.

- A heartbeat service can be used to maintain system state and help the load balancer decide where to direct requests.

Now, when a service crashes, the heartbeat service identify and restart the system immediately on the network.

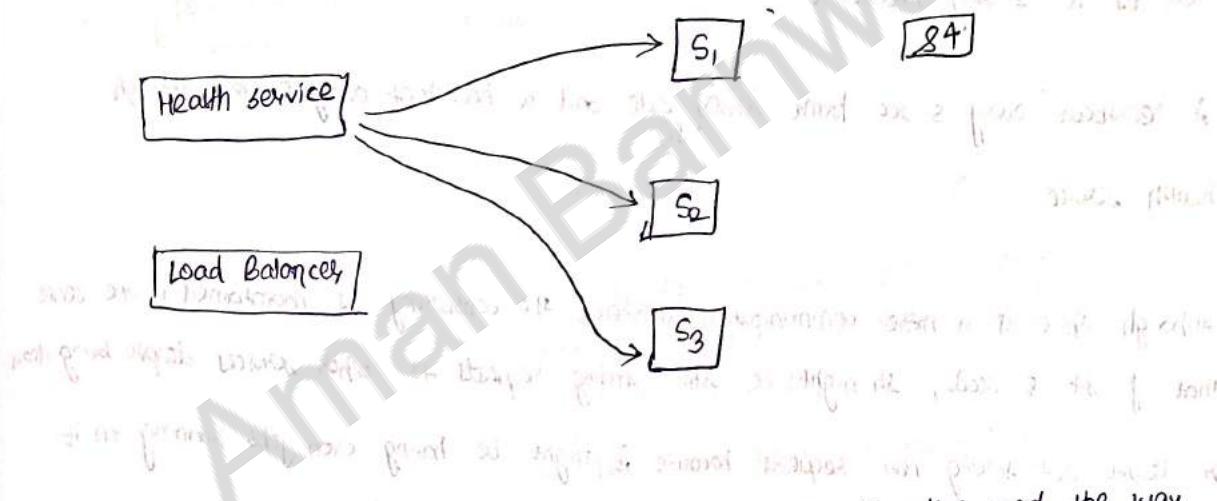
- Service discovery is another important part of deploying and maintaining systems. The load balancer is able to adopt request routing. Both features allow the system to report and heal issues efficiently.

Most of the things on server side deal more with reliability and availability rather than efficiency.

A lot of people when they are starting a service, they think about how do I make this service more efficient but what you really need to do is have enough checks and balances to make sure that this service is running all the time.

When that happens, user tends to use that service more which ends up generating more revenue.

Eg:



The main job of health service is to make sure that everyone is alive. and the way it does that is by talking to all these services.

Eg: Server S1 will be sent a message that "Are you alive"? and S1 will respond with yes.

5 sec later, the health service will again ask that "Are you alive"? This cycle continues.

Let's say, S1 goes down so here now S1 will not respond yes to the requests sent by health service. At this point, the health service can mark S1 as critical.

And, if S1 is missing the 3rd Health service request, then you can assume that S1 is dead.

Now, if S1 is dead, there is no point of sending requests to this server S1.

And we can restart the machine (server S1) or go on some other

server S4 and run the service that is running over on S1 to S4.

You actually send this information to the load balancer saying that s1 is dead.

→ This may not necessarily a real machine s4. You can run another instance on the same machine s1.

→ In some cases, where application itself is dead rather than service, the service should be telling you that "I am alive" through the health service.

So, it's a 2-way heartbeat.

A heartbeat every 5 sec from service side and a heartbeat every 5 sec through health service.

→ Although there is a more communication overhead, the consistency is maintained in the sense that if s4 is dead, s4 might be still pending requests to other services despite being dead or despite not taking new requests because it might be having cron jobs running on it.

So, when s4 does send these requests to these services s2 and s3 (let's say), it has stale data. That's one problem. It might be manipulating the internal state of these machines.

The way we can avoid this is by basically avoiding zombies is by having a 2-way heartbeat mechanism.

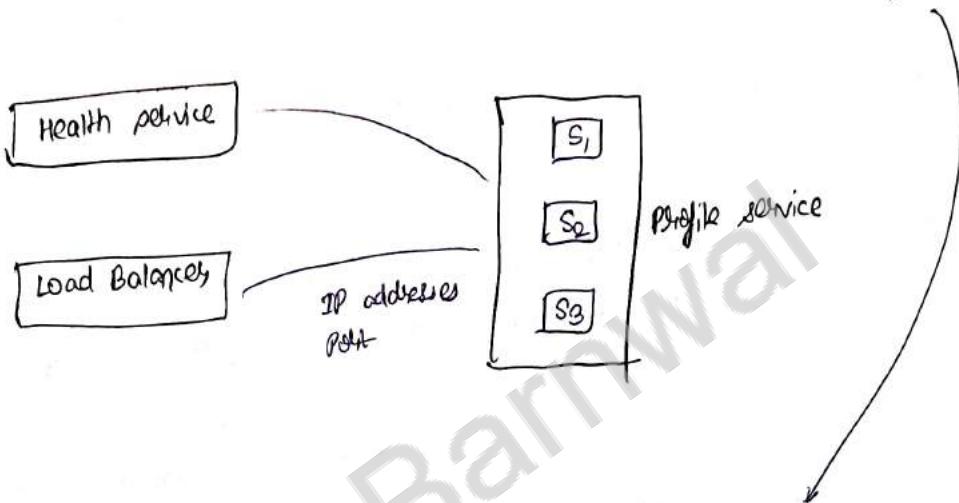
Even this does not actually solve the problem entirely but this really helps you reduce the problem to a large extent.

Bcz if s4 has been marked dead by health service, then s4 will kill itself.

→ Interestingly, the health check problem is very closely tied to the service discovery problem, in the sense that if there is a new service which is coming up having these 3 boxes s_1 , s_2 and s_3 , all it needs to do is to tell the load balancer that "I have 3 boxes on which I am running my service and " " needs to persist data somewhere.

(These are IP addrs, these are ports on which I am running my HTTP clients).

Eg:



Based on this, the Load Balancer changes the snapshot that it has, so the LB need to persist data somewhere.

This data is basically a snapshot of entire universe of the services that we have

when the profile service comes alive, you need a new snapshot (that's a new version). You need to persist that in the DB saying s_1 , s_2 and s_3 have these ports and IP addresses for which the profile service is running.

Once you have this snapshot, you can then tell every other service that hey this is the snapshot. If you need to send a message to profile service on s_3 then this is IP address.

In that way, what happens is that services themselves don't need to keep that info. Everytime you need info, you come and ask the load balancer. You can even cache the info on load balancer.

→ Everytime there is change in snapshot, the health service can see the difference in snapshot and based on that, it can open up connection with these 3 services on their own pet's, always making sure that they are alive and the health system is good.



After taking care of local pet's health, we can move to another place, where we have the same kind of pet's, so we can do the same thing there also. So, this is how we can take care of different places and different animals.

(Continued from previous page) Now, if we take a look at the first slide, we can see that there are 3 main sections: Pet Health, Pet Care, and Pet Training. Pet Health is the most important section, as it contains information about the health of your pet. Pet Care is the second most important section, as it contains information about the care of your pet. Pet Training is the third most important section, as it contains information about training your pet. These three sections are interconnected, as they all contribute to the overall well-being of your pet. For example, if your pet is not healthy, it will not be able to care for itself properly, and it will not be able to learn new things easily. Therefore, it is important to take care of your pet's health, as well as its care and training, in order to ensure its overall well-being.

Distributed consensus and Data replication strategies on network:

we talk about the master slave replication strategies for reliability and data backup. This database concept is often asked in system design interviews with discussions on consistency and availability tradeoffs.

very closely tied to the master slave architecture is the concept of distributed consensus.

when designing a system, we must make sure that the individual components can agree on a particular value. (leader election, distributed transactions, etc).

some popular techniques are 2 phase commit,

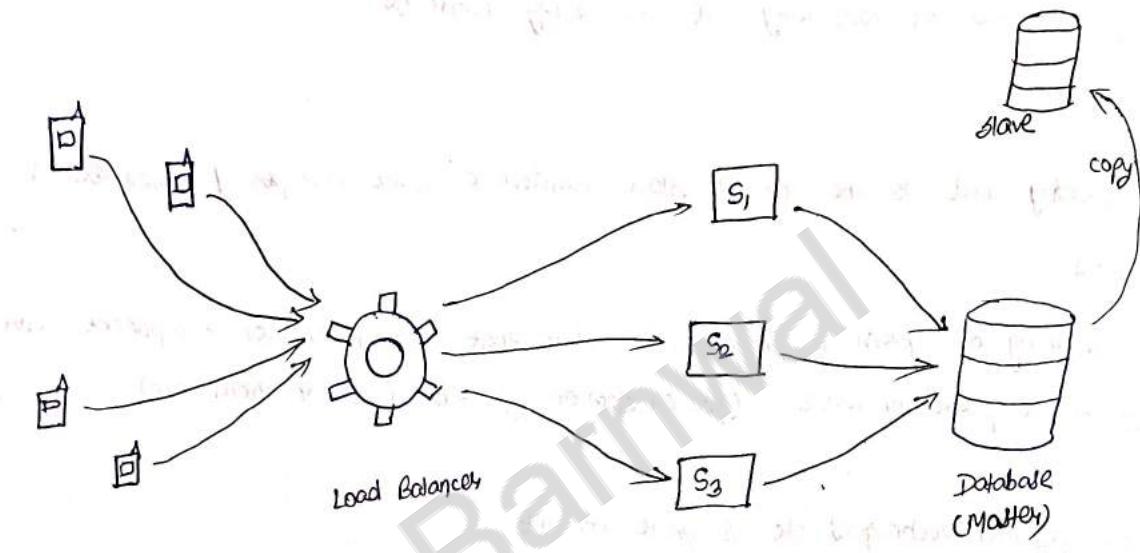
Multi version concurrency control

SAGAS

durum.

Master-Slave Architecture:

Eg:



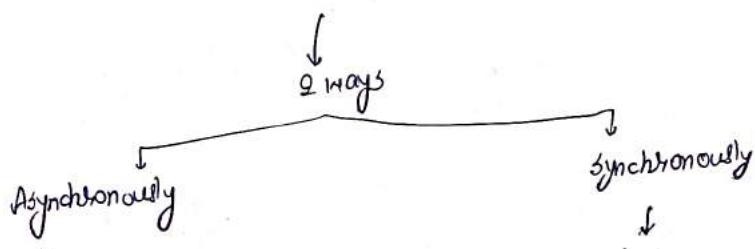
4 cell phones

Note, the load balancer is redirecting the requests coming from 4 cellphones to our servers.

Here, we have a single load balancer → leads to single point of failure
& one db → leads to single point of failure

To avoid this, the simplest soln is just to replicate
the data that you have in the database.

Now, that you have this copy of your db, how are you going to pull data from original database?



Adv:

If you do it this way, then there is
not too much load on the original db

There is not somebody continuously asking
for an update everytime there is an update on
the original db -

Here, when there is an update on original db,
it's going to send a message to its replica db
that "Hey, I got an update" and you can make
this change in your transaction log.

Eg: ADD 100 users → command that master got

Asynchronously

problem:

It's going to be out of sync.

Let's say, if there is a transaction in the original dB, it may not reflect in the replica dB before the original dB created.

In that case, you are going to have some inconsistent data.

Synchronously

The master gets the command and it sends it to the slave to be copied.

Eg: ADD 100 VOLTL (for master)

ADD 100 VOLTL (for slave)

What if one of the servers want to update this slave?

Eg: S3 wants to send this command to slave dB.

→ In this case, the slave needs a way to tell the master that there's a change that needs to be propagated to the master.

2 ways

First way is to ignore the problem

or just allow that problem to happen never

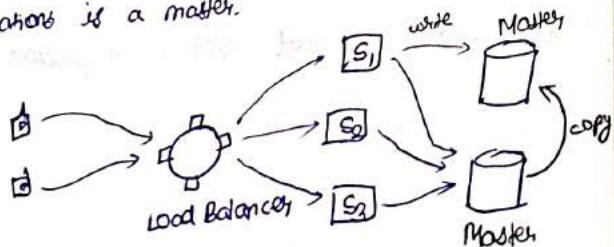
One of way do do this is to never allow write operations to come directly to slave dB from any servers.

Allows the command, and the slave will propagate this to the master.

In this case, the slave has actually taken the role of master. It's no longer a master-slave relation, but now it's a peer-to-peer relation, in which case both of them are masters.

Any database copy/replica which is taking write operations is a master.

Eg:



Master-Master Architecture

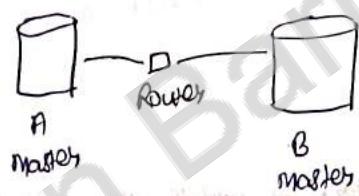
Master-Master Architecture:

Here, there are write operations happening on both original and its replica database, so there is some load balancing on write operations also.

So, if any of 2 dB masters, every request can be redirected to the other master dB.
which results in a very resilient and strong system.

However, there is a problem, if the routers b/w the two masters fails.

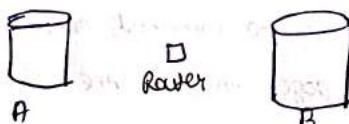
Eg:



Note: If routers fails, both masters will individually assume that the other master fails so now, I am the only master who is going to take read and write operations from now.

↓
A problem will arise

Eg:



Let's assume here, the routers b/w A and B fails.

And a user has an initial balance of Rs 120. And that user sends a message to A to deduct Rs 100 from his account and then another message to B to deduct Rs 50.

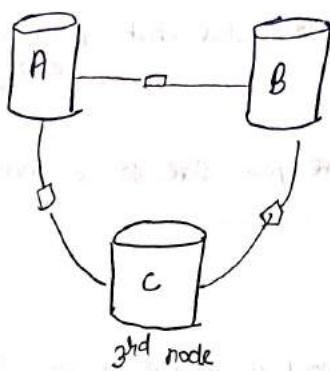
$$\rightarrow 120 - 100 - 50 \Rightarrow \text{Ove balance}$$

This problem is called split-brain problem.

split-brain problem:

Interestingly, the split-brain problem can be solved by adding a 3rd node to the original architecture.

Eg:



Here, A and B has a split-brain problem.

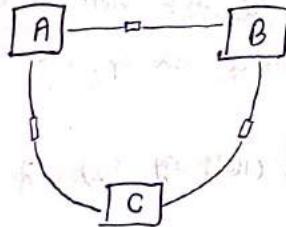
So, you add C to solve this problem.

One assumption we made here is that the chance of a node crashing and the greater than other two nodes crashing is highly unlikely. (between A and B here)

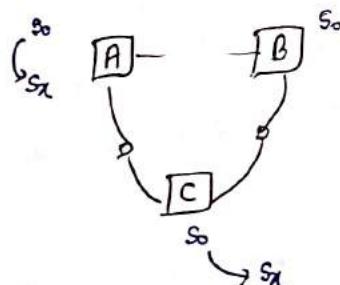
→ If C crashes, A and B are still be in sync b/w them & when C comes again, it will read from either A or B to sync data.

But when the cluster b/w A and B goes down,

Eg:



Initially, assume state of all A, B and C are s_0 . Now, cluster b/w A and B goes down. And there is a write operation on A due to which its state changes to s_1 . This change propagated to C.



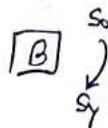
Now, again there is a write operation on B due to which its state changes to s_y .

B now tries to propagate this state s_y to C.

↓

But that doesn't happen bcs C says that what's previous state to B.

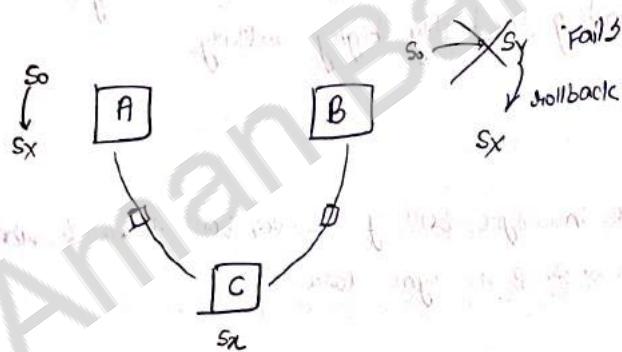
C then said that so it is not the same state that I am in.
(i.e. s_x)



It will say to B that you need to update your state to s_x and then you can have a successful transaction to s_y .

↓

so, B's transaction will fail and instead of moving to s_y , it'll rollback on the transaction and it will signal up with s_x .



Now, the web can send the request again to B with newer state s_x .

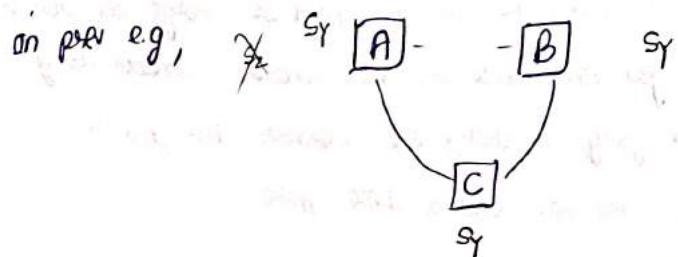
B can then run the transaction, reach a state s_y , ask C to go there.

Now, when A gets a new transaction (let's say, s_z), it can have some problem.

↓

Here, A can never make a transaction which doesn't have the updated state from B also bcs that is being propagated by C.

Distributed consensus: It is a way in which the multiple nodes agree on the same value.



All A, B and C are agreeing on final state sy.

Note:

→ There are many protocols that you have for distributed consensus.

① If you have a lot of nodes then 2PC is a very popular protocol, although it's really slow. 2-phase commit

② + 3PC (3 phase commit)

③ + MVCC (Multi version concurrency control) → It's used by PostgreSQL.

It keeps the multiple versions on some data.
i.e. if you send multiple update on the same data, it keeps multiple copies.

→ If you are okay with dirty reads then it's going to keep an older version of data and someone can go and read it.

→ But if you pay that I want really nice serializable reads, then it's going to make it slow but it will make sure that data is consistent.

④ SAGA : used by Microsoft

Eg. Let's say, you have a food ordering app. You makes an order, restaurant has not yet agreed on order. The bank locks the fund but restaurant tells the bank that until I get a confirmation, I'm actually not going to withdraw that money.

If transaction fails, then I'm just going to keep the money. You don't need to go through the whole transaction process and settlements at end of day.

- SAGA is really a long transaction which at any point can fail and you may need to roll it back.

Eg. (ii) If you have a phone app and a person makes a call. You need to charge the person but it might last for 30 min, so you can break it into smaller transactions of 1 min each, in which you are not going to charge the customer, but you are going to lock funds every min a little more and a little more.

At the end, you are going to decide whether call was a failure or success based on that, you are going to charge the customer.

Advantages of Master-Slave Architecture:

- ① → Nice replica of master
- ② → You can scale out your read operations.
- ③ → You can add as many slaves as you want and just do read operations on all of them.

Eg: if data is not critical, lets say you are adding a comment on facebook, it does not really matter if it does not reflect immediately.

as soon as slave gets the data, it will take sometime to get reflected on the master, but the slave will take sometime to get the data.

How to avoid cascading failures in a distributed system?

Thundering herd problem → This problem occurs when there are a large number of requests on the servers, and this results in the servers crashing due to overloading.

One of the solutions to this problem is rate limiting.
In a distributed environment, the rate limiting problem is a complex problem to solve.

→ When designing systems on the server side, we often need to predict the capacity of a server and apply limits on the number of requests it can receive per second.
This is mentioned in the service OPS document.

It stands for Queries per second.

→ The system design discussion helps us understand how to deal with requests based on priorities, how to handle a large no. of requests, viral posts or videos and their requirement.

→ Some other scenarios like job scheduling on cron jobs being fired in batches is also discussed. We use batch processing and approximation to reduce server load.

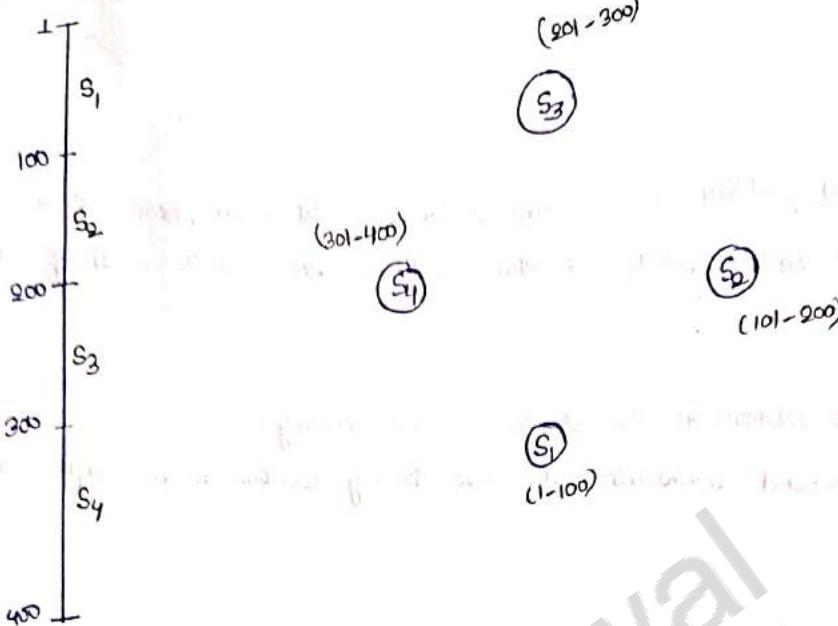
→ The last few approaches include gradual deployments and using caches to store responses for common requests. This helps us improve system performance. Improving ops and performance help us handle more requests which means more users and more money for the product.

→ Caching and coupling system can also help improve performance. However, they must have time outs and appropriate cache eviction policies set while designing the system.

① Cascading Failure:

join is state limiting

Eg:



Let's say, you have 4 servers and the requests changing from 1 to 400. So, every server here is load balanced to solve a range of 100 requests.

Now, let's assume S₁ crashes

which results in additional load on remaining servers S₂, S₃ and S₄.

Let's say

S₄

(301-400)

(1-23)

S₂

(101-200)

(68-100)

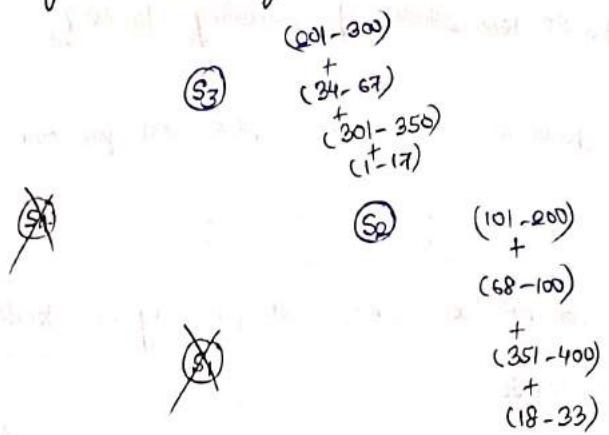
Now, we are assuming that each of the remaining servers can handle the new load.

Let's assume that S₄ don't have that much compute power to handle the additional increased load due to S₁ crashing

which results in crashing of S₄ (let's assume)

Now, S₁ and S₄ both are dead.

so, now the load of S_4 will go to remaining servers S_2 and S_3 .



There is good chance that S_3 also crashes bcz of 200% additional load.

$\therefore S_3$ will become dead.

Due to this, all load went to S_2 , due to additional 300% load of its original capacity, S_2 also crashes and became dead.

It results in the crashing of whole system, and so all the users are upset.

This problem is called cascading failure problem.

It is actually a race against time. When S_1 has crashed, that small time gap that you have for bringing in the new server before S_4 take that much load and crashes.

One of the things that you could do is to have a really smart load balancer or have a planned point of new server bringing in.

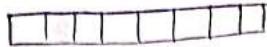
One workaround for this problem is to just stop serving requests of all users having request id 1 to 100.

If you see that some of the services can't take in more load than it's better to be available to some users than to be available to none of the users.

Note: We often use message queues for rate limiting requests.

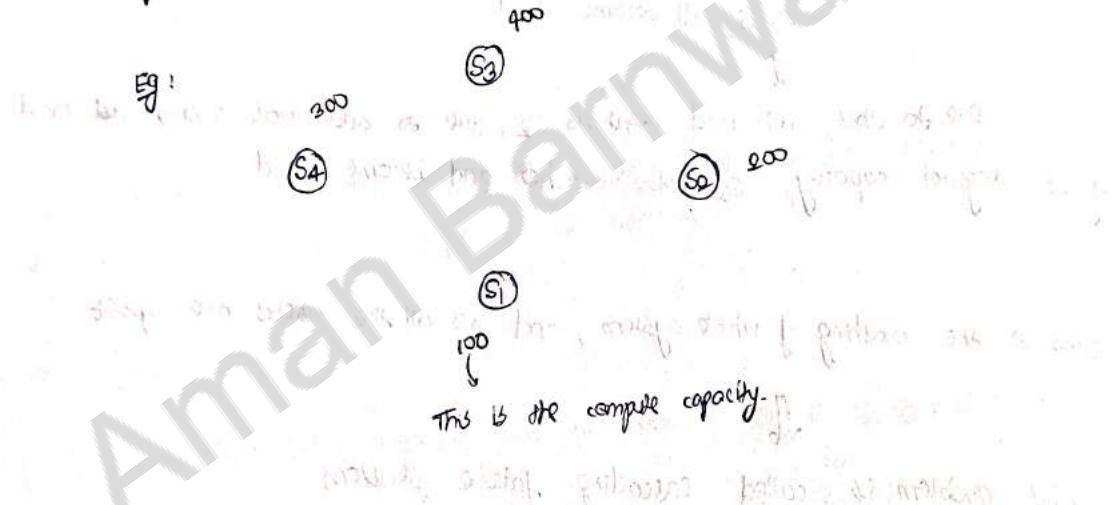
But what's the real solution of cascading failure?

→ What we should do is to take a queue and put our requests in this queue



Every server can have a request service and they can decide on answering or not answering a request.

So, what I will do is give these servers a particular capacity. (Here, 1 unit of compute capacity means that it can handle 1 request per second).



Let's say now S3 crashes.

If it went to S4 then it feels that the max request it can handle per second is 300.

so this queue is going to keep expanding till it hits 300.

301st request on S4 fails i.e. we are going to ignore that request so when we return a failed response to the client, atleast this server S4 will not get overloaded.

Also, the client is now aware that "ok, this request has failed, maybe after 5 min I should try again". The user which made this 301st request (which failed) had a bad user experience.

But putting some users is better than serving no users.

There are some kind of errors that you can send the client.

Temporary

Permanent

It means you should try in sometime, may be there's some internal server issue going on, it may be DB is too slow, it may be there's too much load.

Try after sometime and client can display messages accordingly.

It means that there is some obvious mistake in the request that you sent and there is a logical error.

Or the general idea is to limit the no. of requests you can take on the server side so that you can handle the load till the scaling bit comes in, i.e. till you can bring in the new service.

② Visits: Black Friday.
so it is prescaling

③ Predictable load increase:

(Eg: Black Friday, Diwali, etc)

point is ^{auto} ~~prescaling~~

The second problem is - if you go viral or if there is some sort of event (like say, Black Friday), which increases the load on the servers at times,

then that would be the issue.

Prescale your system beforehand so that there would be no failure of system in case of additional loads.

If you are not very sure about the no. of servers you will need during event, you could autoscale.

It is being provided by cloud services.

- ④ Prescale
- ⑤ Autoscale
- ⑥ Rate limiting

How about if you go viral?

→ If you go viral, you can just fallback to the old way of rate limiting.

i.e. if you do rate limiting, you will be stopping the max no. of users that can you can actually serve.

Going viral is not something that you predict so pre-scaling is not really a solution for this.

Autoscaling and rate limiting both can be the soln if you go viral

② Bulk Job scheduling: → soln is batch processing

The 3rd problem is a real server side problem.

We often write cron jobs which run at some point of time

↓
used for scheduling the tasks
to run on the server

e.g.: Imagine a cron job which is sending email notifications to all users wishing them a Happy New Year on 1st Jan.

→ one way is to send all the emails when clock becomes 00:00.

That's a bad idea when the no. of users is very large (let's say, 1 million)

The way you avoid this is by breaking the job into smaller pieces.

such that out of 1 million users, 1st 1000 users will get email in 1st min
and 2nd - - - - - 2nd min

Batch processing

and so on.

→ if your user will not get the auto-generated email notifications by new year
at 00:00, they don't really care.
so, it's something that you can do.

i.e. Batch processing is something that you can do ~~now~~.

④ when someone Popular post: → 10^7 is Approximate statistics

when a post becomes really popular.

Eg: Let's say, a user like ~~edabit~~ T-Series post something on YouTube then
you need to send it to all of their subscribers.

If you do it in native way, the same issue of Job scheduling will come here too.
(i.e. too many users and a very small time Δt).

So, you could do the Batch processing over here.
Eg: sending to users in chunks of 1000.

But, Youtube does it really smartly by adding Jitter.

If you have a lot of subscribers, the notifications will go to them in a Batch-processing
way but if they start hitting the video page, then the video content is sent to
Youtube but there are a lot of data which doesn't matter.

Eg: No. of views
-- likes
-- comments.

→ if you have a very popular user (lets say, Trivates) posting the video, the no. of views are going to change dramatically so what you could do is to faithfully display them or you could do it in a smart way.

i.e. in 1st hour, we get 1000 views.
and hourly, $\dots \times 1.5$
 $= 1500$ views

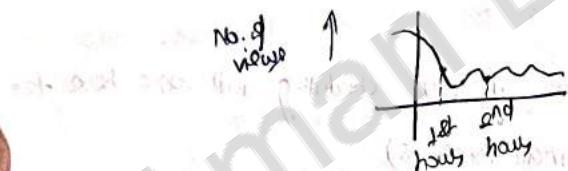
(but lets say, the reality is
1700 views)

mismatch in no. of views

but we don't really care.

But this is metadata (which is something that is not like the video).

→ YouTube could figure out how the views changes over time and then figure out the no. of views at a specific time from the graph.



YouTube must be much much smarter than this, but I am just giving the general idea of approximation.

Instead of showing people the truth, do the approximation and save a lot of loads on your server.

→ potentially, this could save a lot of database queries that you are making to get the metadata of a post.

so it is approximate statistics

Problems:

- ① cascading failure → ① Rate Limiting
- ② going viral → ② Pre-scale
- ③ predictable load increase → ③ Auto-scale
- ④ Bulk Job scheduling → ④ Batch processing
- ⑤ popular posts → ⑤ Approximative statistics

sohn

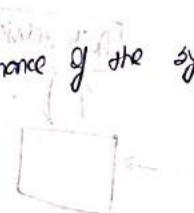
Apart from these sohn, there are some good practices in the server side to avoid a Thundering Herd.

⑥ Caching:

i.e. if you are getting a lot of common requests, the response is going to be same and you can just cache the responses for those requests.

→ There are key-value pairs and this is going to save a lot of queries that you will be making on the database.

↓ which will improve the performance of the system and also you can then handle more loads.



⑦ Gradual Deployments:

most of the issues that people get on the server side is when they are deploying their service.

the reliability engineer who are fighting deployments
and the developers who want to deploy more, bcz they want to get more features out
→ ie they want to stop the deployments as much as they can because that makes the system most stable.

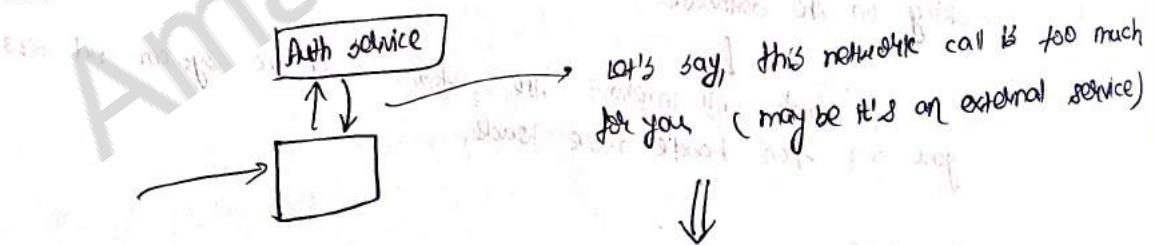
↓
so, it's an interesting tug of war 😊

→ Here, you don't deploy all the services together.
You deploy the first 10, then have a look at what's going on - and then
you deploy the next 10 and so on and so forth.
This won't be possible in certain scenarios when there is breaking change.

⑧ Coupling:

To improve performance sometimes, you need to store data.
(i.e. very similar to caching)

Eg:
Let's assume, you have a service. For every request that it gets ask the authentication service to authenticate the user first to authenticate this request and then serve this request.



Let's say, this network call is too much for you (maybe it's an external service)



You could cache the user's username and token or password whatever you like.

Instead of talking to authentication service and verifying whether it's true or not, what we could do is → You can see that if username, password worked once in the past one hour, then maybe the password does not change.

If we are going to assume that this user is authenticated to call this service.

- It's good in a way that it is not querying the authentication service all the time which results in reducing the load on authentication service, improving performance and user experience.
- Except that it is a financial system and the password ~~does~~ ^{had} changed and there is a person who has hacked into his account and then ^{he} ~~you~~ are using this password then you are in big trouble.

↓

so, you should only couple systems which is actually keeping data in the db cache.

~~sensitive~~ sensitive data in the cache should be avoided.

- However, if you have some data like profile picture, you can keep that in cache for 1 hour or 2 hours.

↓

i.e. You want to take this case-by-case and understand that keeping some data for an external service in your own service whether that's a good idea or not.

↓

That will improve ~~the~~ performance which in turn help you handle more requests, so the problem of thundering herd will be slightly mitigated.

Containers and Virtualization in cloud computing:

Virtualization of hardware and compute resources is a key value idea when operating on the cloud.

→ containers allow us to scale applications faster and easier, while avoiding large initial investments and reducing maintenance costs.

More importantly, containers allow DevOps team to run programs on heterogeneous hardware with small startup times and better configuration capabilities than virtual machines.

→ initially what used to happen is when you used to write code, you would have DevOps team at a finance team, which asks you which box would you want to buy for this code so, you would do capacity planning, which is a bunch of guesses like this is the compute I need, this is the kind of memory and storage I'll be needing for the application.

What if your business scales?

→ You don't want to buy a small computer bcz that investment doesn't make sense.

You won't buy a large computer or a reasonably safe computer so that you don't need to do this again and again.

But the problem is with this is the initial hardware investment is very large.

→ if you want it to horizontally scale just on the basis of hardware, you would need to buy as many computers as possible

→ So, one of the approaches that the organization took was to let their employee to use that single computer.

But if you have multiple people using the same resource, there is going to be some contention.

↓
Shared compute

so, you want to isolate resource usage as much as possible.

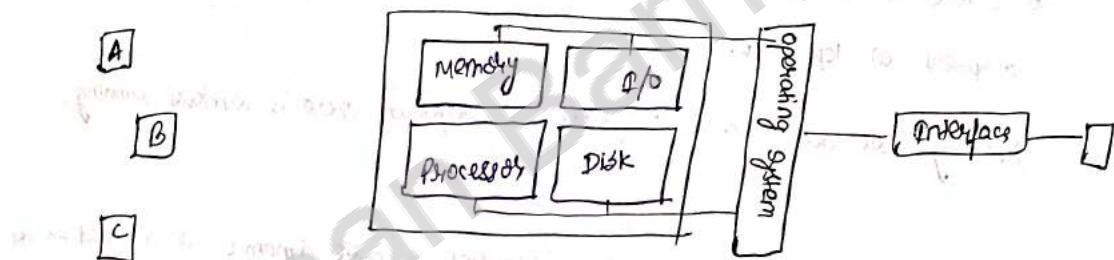
The Qn is "Can we do this per application?"

Let's say, you have a program which requires a no. of resources (memory, I/O, processing, and disk).

Obviously, the more programs you have, the more resources you need.

But you also want to do some sort of boundary management which means that A does not interfere with B's memory, and C doesn't interfere with A's I/O.

And that responsibility is going to be taken by operating system



When you write a program, you talk to os saying that "Through your interface, I would like to book a amount of memory, y amount of I/O and so on".

Similarly, you can have multiple programs which are going to be taking up slices of resources that the operating systems can provide and remaining is going to be unused.

The problem with this is you can still have the same problems of shared compute.

So, we prefer to have a very strong boundary which is provided by virtual machines.

It is like an OS in itself.

So, it's an OS running on top of OS.

→ You can interact with this fake world that you have without concerning yourself of what other programs are running in the same hardware.

→ You're now mainly concerned with virtual machine that you have been assigned which has been given a set of resources.

→ The basic idea of cloud computing is huge companies like Amazon and Google have a lot of hardware. What they can do is take all the spare hardware and rent it out to small businesses.

↓
which will save big investment of maintenance and upfront cost of buying a computer.

i.e. The rental cost of paying and maintaining the computer is taken care of.

→ The other good thing is my code doesn't need to be platform dependent.

Eg: I can take a Windows computer and run Linux on top of it.

so I have a 64 GB Windows computer, I can run 1000 say, 4 to 5 Linux computers on top of it.

And, my code doesn't need to know that deep down there is Windows running.

→ This is very flexible. The provisioning of these resources is quite dynamic. All I need to do is shut down my virtual machine and restart a new machine.

I don't need to go to the shop and buy a different computer.

① App Isolation

② Platform Independence

③ cost saving

Eg: Java became popular bcz you no longer had to take OS considerations. You could push that onto the ops team. You just needed the .class files, and you could run the program on any operating systems.

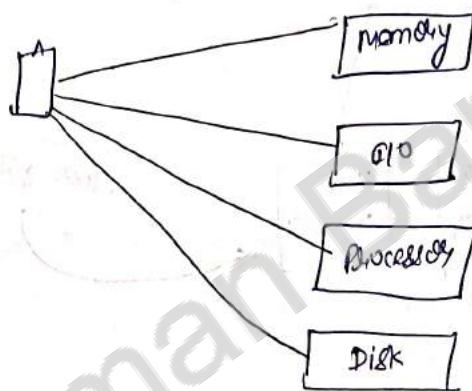
→ People could go to these cloud providers and ask them for a virtual machine and they would get it very cheap.

But one of the problems with the virtual machines is that when you are running a program, you don't want to start the entire virtual machine.

It's like booting up your computer at times some time (a few sec).

And then the idea came up that "you just need processing power, memory, disk and I/O."

I just need my memory fix. & I don't care where it comes from.



Taking all this into consideration, there came something like Lightweight OS.
↳ containers

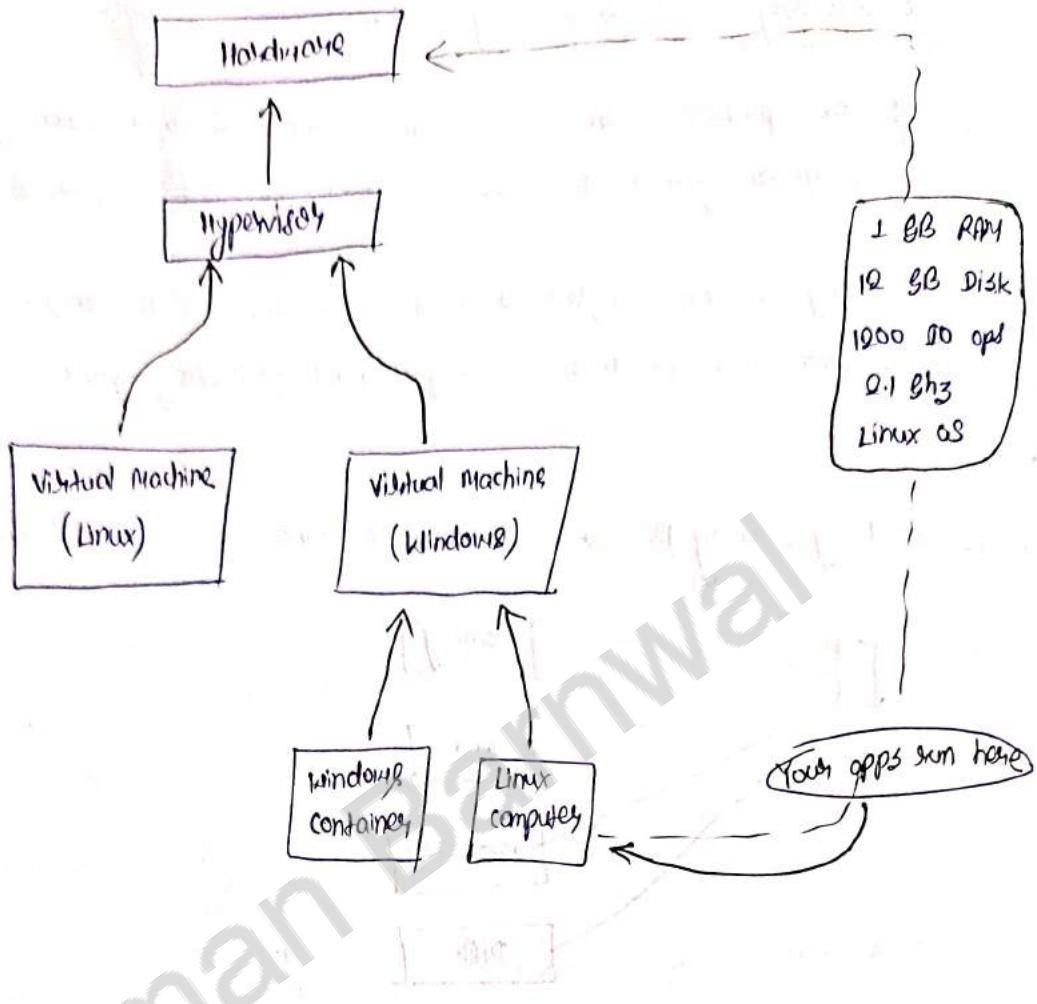
containers:

Containers are effectively a form of virtualization. They let you do the app isolation.

You are not going to be having all features of OS isolation, but you don't (virtual machine)

really need that.

And the benefit is you just need to recompile and build this lightweight container. so, that's faster when it comes to boot times.



→ Interestingly, this process of building and tearing down is called mounting and unmounting a disk.

→ In a container, you specify the os and your disc requirements. And then the mounting system process figures out the underlying file system, which you will be using through interface of the container.

i.e. Requirements are defined by 'images', which are mounted on containers.

Docker:

→ Technologies like Docker did the same thing that Java did for programming language. It's a way for us to move all these considerations back to the developer.

The developer can actually specify that these are the resources I need.

This is the OS I will be running on.

And you don't need to worry about the hardware so much. Docker will handle that. It will create an interface which will be interacting with the hardware on the virtual machine, whatever be the case.

It will create a container on top of it and let your program interact with the interface.

Disadvantage of containers:

① As compared to bare-metal apps, it is considered slow.

② There is also possible firewall issues.

Although it's much lesser in containers than a virtual machine.

③ The overhead of container management is not really worth it for simple applications.

(if your app is very simple and you don't want cloud, run it directly over the OS).

Data consistency and Tradeoffs in Distributed systems

(consistency)

→ you have heard this word in
CAP Theorem

Consistency
Availability

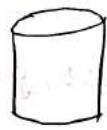
Partition tolerance

ACID properties
Atomicity
consistency
Isolation
Durability

What is consistency?

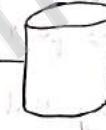
→ if you have multiple copies of data, these two data should match each other.

Eg:



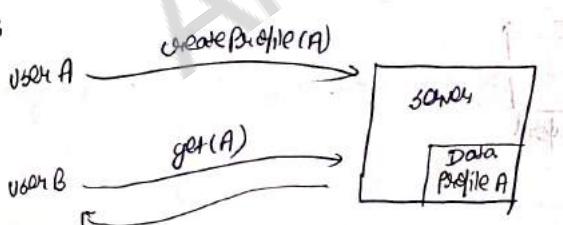
User DB
(server 1)

Any user's data should be
same across both servers



User DB
(server 2)

Eg:



Let's say, user A creates a profile
in the server,

Problems:

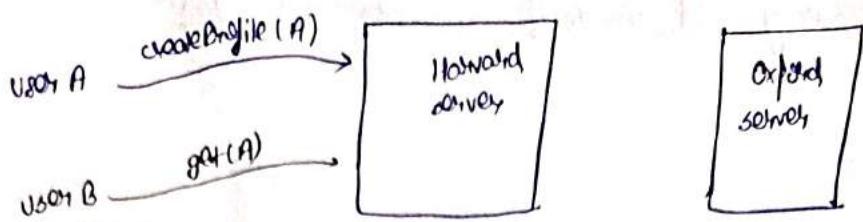
① single point of failure

② cost of vertical scaling is going to be high.
(ie there is a vertical scaling limit)

③ Latency is high.

We probably need to add more servers

lets have 2 servers
one in Harvard
other at Oxford



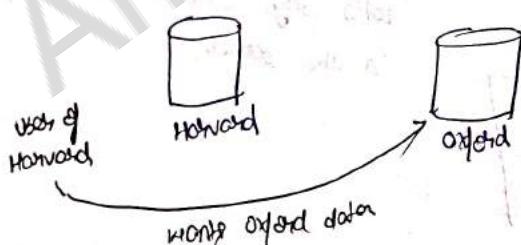
There is no any communication b/w these 2 servers. They just had data copied relevant to the students here.

So, there is no data sharding b/w the servers.
(here, universities)

Let's say, we try to avoid this problem as much as possible.

We say that if you need any kind of Oxford data as a Harvard student, you connect to Oxford server, which is cross-continental.

You connect to it. We are going to keep any data belonging to Oxford in any other region.



But the older problem of single point of failure is still there.

(if Oxford server goes down)

High latency (going from US to UK)

→ (high cost of vertical scaling) problem is solved by spreading out data according to where it belongs.

But we still have problems with high latency and single point of failure.

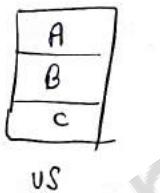
→ To reduce latency, we can cache some of the information
(i.e. some of the information in the Harvard servers).
Oxford

You fetch the data from Oxford once, you cache it on Harvard servers and then you keep giving the response back.

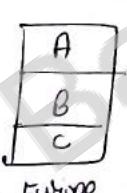
↓
It will reduce latency but we will not get rid of latency. There is still a chance that you are going to have a cache miss.
(i.e. if profile didn't find in cache, we still need to go a long way to Oxford servers).

→ If you have multiple places where student data have been stored.

Eg:



US

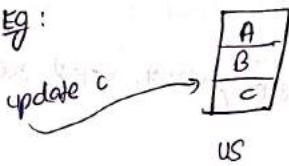


Europe

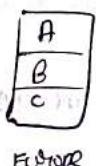
i.e. No single point of failure anymore bcz of multiple copies of data.

Latency problems are also fixed bcz you have quick responses

Eg:



US



Europe

The moment you say "update c", once it is updated in all copies of data, it actually sends back that response and users are happy.



i.e. Latency is low.

How do we ensure consistency in multiple copies of data when you do an update in one copy?

→ When you update on one copy, how do you propagate this update to other copy

You can do it ~~manually~~ manually.

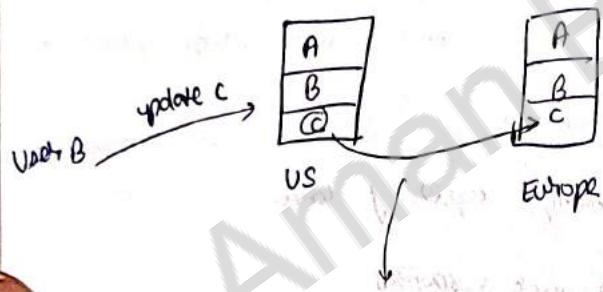
If you have very infrequent updates in a product, you can make the decision by yourself.

Eg: This often happens in Bank accounts.

You actually make a change to the ledger. You just noted down and eventually you make that system consistent by spreading information all around the world. They are notified that it's going to be take a few days to reflect in your bank statement in worst case.

→ In a real-time response system, you can't afford that.

Eg: if you have something like Facebook, you can't afford that.



We want to propagate this change / update in C quickly.

We can use some sort of electronic message network protocol. TCP seems like a good idea.

Why TCP?

→ Bcz TCP is reliable delivery.

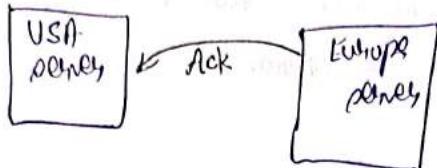
How do TCP work in our case?

→ When you send an update on C to CTF on US server, it needs to propagate from USA slowly to Europe slowly.

When Europe gets this message (C → CTF) internally in their database, and then update it.

what happens if the message transmission fails?
if there's a network issue or maybe the Europe server is down

- You will never get to know unless you get an acknowledgement.
i.e. the European server should send back " ".



what happens if you don't get an acknowledgement?

- You can keep trying as US server to make this update to Europe server till it succeeds.

How many times do you retry?

↓
Infinite

which means that every 5 sec, you are going to be retrying.

or

you can just retry 5 times & give up as you can afford these data loss
to be inconsistent.

This actually brings us a fundamental problem with distributed systems, which depend on acknowledgement for consistency.

The idea is that when you send an update to one of these servers then it has

2 options :

Either it locally commits that update i.e. makes that change

or it waits for its peer to give acknowledgement back.

- If it makes the change locally then it's already committed and it can't depend on the peer to actually make the commit to.
bcz the message might fail bcs of unreliable network.

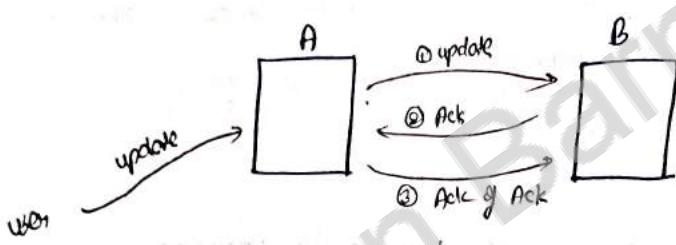
Let's assume that this is an unreliable network and B may not be able to make that commit.

If you wait for B to make the commit, let's say, if B makes the commit then B has no idea whether the acknowledgement is ever going to reach you.

You can use some special acknowledgement which is ACK of a ACK item when A makes a commit item. It has no idea whether this special acknowledgement is ever going to reach B.



The problem is called "The general problem".



infinite process

You can never be sure if you are making a commit based on the hope that the acknowledgement actually goes through.

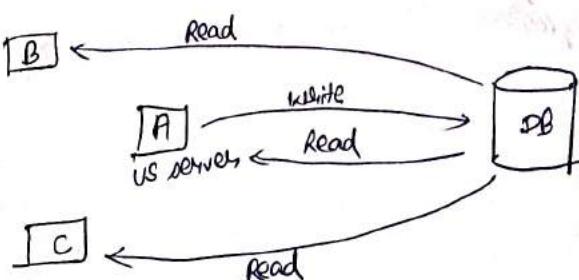
So, we can't have this ^{only} for distributed systems.

if you are hoping for consistency, what can we do?

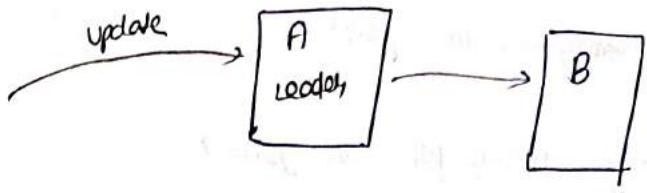
→ we can give them a special role i.e. leader role.

Eg:

Let's make us server a leader. It sees the only person who can commit i.e. actually do writes on a database.



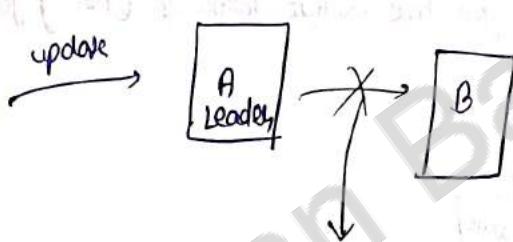
If you send an update, you can't send it to B. It can send only to A.



Now B gets regular updates from A bcs A is the only place which can write info on a dB and B is the place where you can read info from.



so here B is a follower of A and two read operation on both A and B will be going to give you the same data.



If this data doesn't go through, what do you do?

→ If you read the data without change from B, you have inconsistency.
i.e. you will have dirty reads.

If you wait for the update to come through before you give a response from B, at that time, the system is actually not available.

If you hit the system, you are not getting a response. It's as good as system is dead.

That's the fundamental problem with the Distributed systems

① If you have pure consistency, then your system is not available.

② If you have some sort of availability, then your system have to suffer through the inconsistency.

That's the reason, why consistency is such a big deal.

i.e. Hey can you have a totally consistent system?

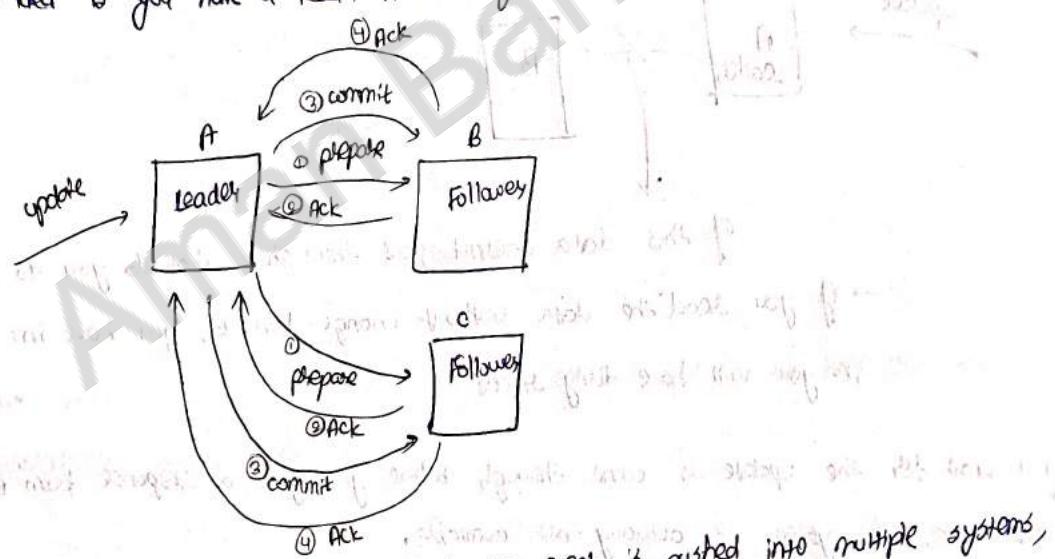
or

How much consistency are you looking for in the system?

2 phase commit protocol: (2PC protocol)

Note, the idea is you have a leader now and you have multiple followers (lots of followers)

e.g.



→ If you want to ensure that a message or the event is pushed into multiple systems,

→ if you want to ensure that a message or the event is pushed into multiple systems,

→ if you want to ensure that a message or the event is pushed into multiple systems,

→ The general idea here is if you are a leader and get an update, you send a prepare request to your followers. Your followers will give you the acknowledgement that "Hey, I got the prepare request!"

And when you get these acknowledgements, you ask your followers to commit. So, this is effectively committing a transaction.

Once you ask them to commit, I'm assuming that they are giving you the acknowledgement and the system works perfectly.

→ However, there is a lot of edge cases in this.

if you send a prepare statement, what actually happens?

→ Let's say, you have a bunch of statements in this update. So, in dB what's going to happen is you will begin the transaction and write down these statements but there is commit missing here.

When you commit, you actually tell the dB to reflect these changes to everyone who is trying to get this data.

1st phase is prepare

When you send ACK, you say that I'm done with executing those statements and I'm ready for commit.

→ If the leader does not get all the acknowledgement back, it assumes that some of the system which it was sending prepare request to has failed.

There might be a timeout or you might give a response also.

Whatever the case, the leader fails the transaction.

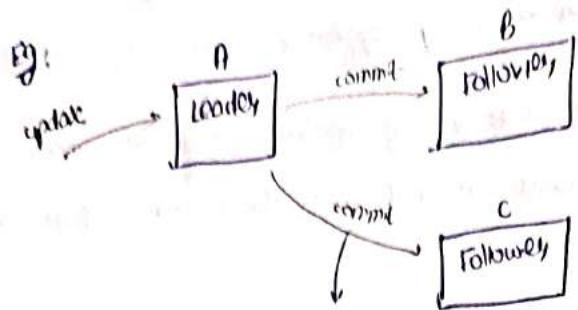


so, you get a rollback

BEGIN
|
|
|
rollback

A follower could keep a time limit. i.e. if in 5 or 6 min, the leader message should have reached. If it doesn't, you just do a rollback and the system is consistent.

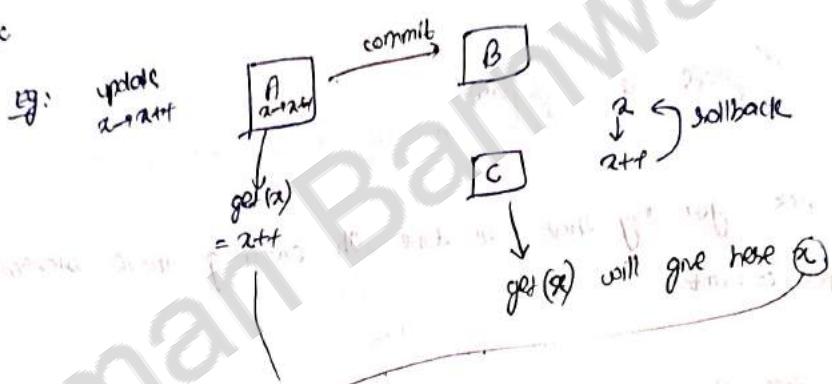
Let's assume one of these commits actually fail, what happens then?



Let's say, this commit failed.

So, after a certain time, C will assume that the transaction has failed and it's going to

roll things back



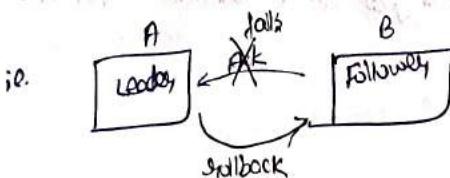
We are in trouble bcz of data inconsistency.
The problem is rollback. You can't rollback by yourself based on time.

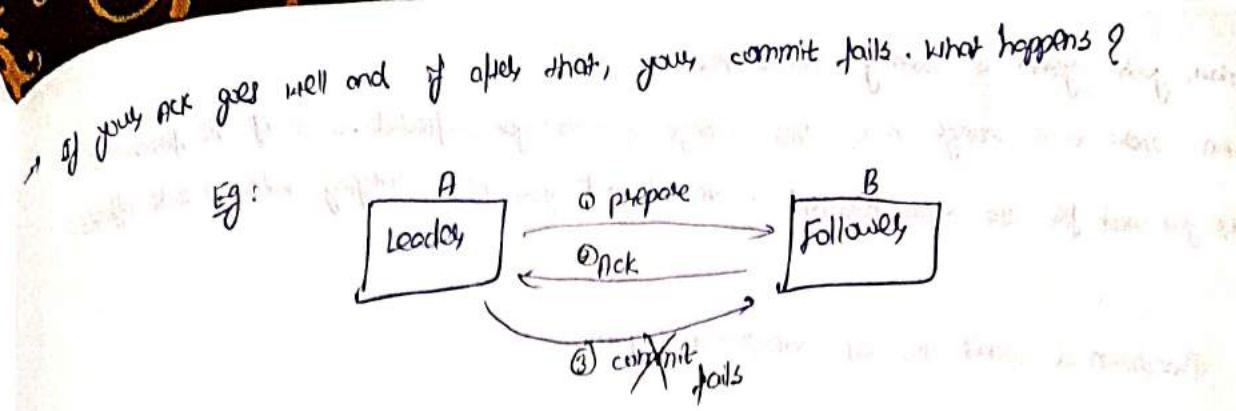
→ You have to wait for the master or leader to actually send messages of rollback.

If someone sends a response or if there is a timeout, the leader actually tells them to rollback the transaction, if ack has failed to reach.

This statement of rollback is not going to be done by followers themselves.

They are effectively brain dead.





You do a retry. You keep retrying commit.

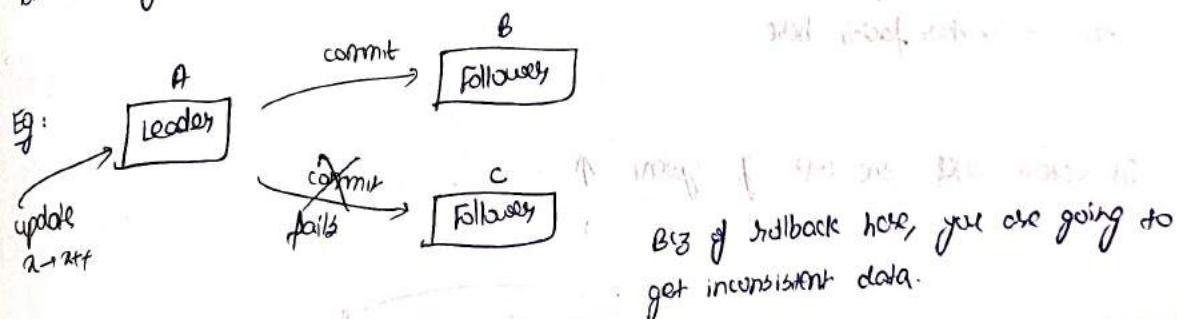
- So, it might be that the follower actually gets the commit message, but it's not able to send you an acknowledgement bcz of which you keep on retrying.

\downarrow
That's OK!

For some duration now, if you get another commit message \rightarrow That's okay.

As long as ACK keeps failing, the leader keeps retrying and the commit messages keeps going through.

Once a commit message is actually acknowledged, the leader is settled i.e. it knows that the data it's showing is consistent with the data that you are showing.



Bcz of rollback here, you are going to get inconsistent data.

What you can do is take the lock and block all the read operations here.

If you take a lock on A also, you block the read and write operations.

So, if anyone does a get request on A, they get nothing. They are waiting.

The problem is that your system is not available.
But it is consistent.

Now, your system is totally consistent; ~~by the end of the day~~.
When there is a change in x , this change is also get reflected to all of its followers.
Bcz you wait for the acknowledgement in the end and you keep trying with no side effects.

Transaction id makes this an idempotent system.

This is a good idea if you need absolute consistency.

Eg: in financial systems

But the cost is that your system is down. Any get request is failing bcz of locking.

The general idea is to see what happens when you have a perfectly consistent distributed system.

The costs are usually high not just for the availability. There is also performance which is another factor here.

In certain cases, the cost of system ↑.

→ Instead, you might want to go for "Eventual consistency"

Defeating Anomalies using Isolation Trees: Practical Machine Learning

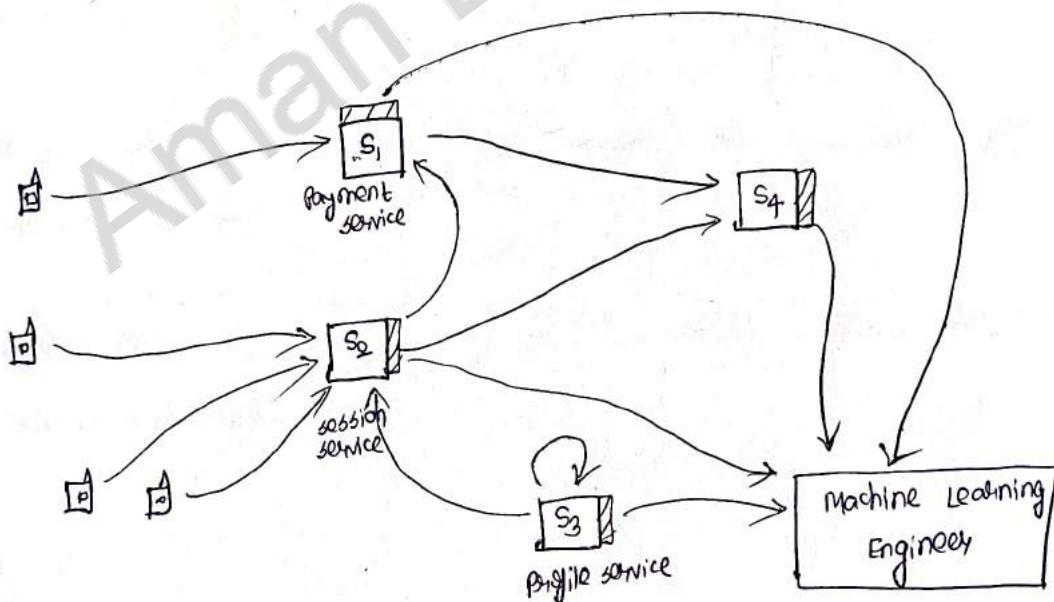
Production alerts are an important way in which engineers monitor the health of their services. The alerts are fired when important service metrics behave irregularly.

An e.g. would be a sudden spike in the no. of errors or a crash in no. of running processes.

→ Here, we design a system to allow important service metrics to be published, analyzed and acted upon. The metrics are sent using sidecars in a service mesh.

The metrics are then copied and run through ML algorithms like isolation trees to search for anomalies.

Eg:



Today, we will talk about Anomaly detection on the server side.
what that means is we will be managing the health of the system, trying to figure out if at any point of time if this system is in trouble.

The way we do that is by having all these servers publish the metrics which are important to us.

- Eg metrics:
- ① No. of processes
 - ② Request queue size
 - ③ No. of requests
 - ④ Error rate

instead of the server figuring out whether it is good or bad, what it does is that it publishes these metrics to a particular engine. The purpose of this engine is to make sure that the health of the system is good by detecting anomalies.

At any point of time if you see an anomaly, what you should be doing is flagging that and sending an alert to a system engineer.

One of the keypoint of anomaly detection on the server side is to allow false positives a lot more than allowing false neg.

The cost of a engineer investigating a false alert is much cheaper than the cost of a real alert being missed out by the engine.

If we are managing the infrastructure here, the 1st thing we need to do is to allow some sort of uniformness and standard way in which all of these services can publish their metrics to this engine.

One way you can achieve this is an infrastructure change by implementing a service mesh, which involve a sidecar.

A sidecar is a microservice architecture that handles a lot of common functionalities.

Ex: Routing, Polising, etc

Eg: If S2 is sending message to S4, it figures out IP address of the box of S4 where it needs to send the message in a particular way.

S3 to S1 is in the same way. So, you can actually code that logic in sidebar.

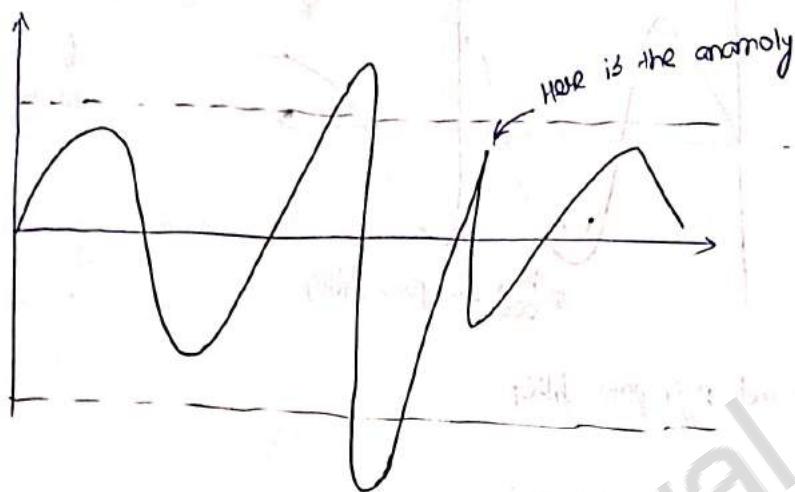
→ Similarly, all of these services are going to send messages to each other probably over HTTP, or over some protocol which you have written over TCP.

You can just parse the message in the same way on all services but you don't want to write that code in all microservices. (Don't Repeat Yourself)

So, you can shift that onto a library and that library is part of big library called a sidebar.

→ If the sidebar exists, then it's going to be taking these metrics from every application and publishing that to this engine without any dependency on any team to actually get into this.

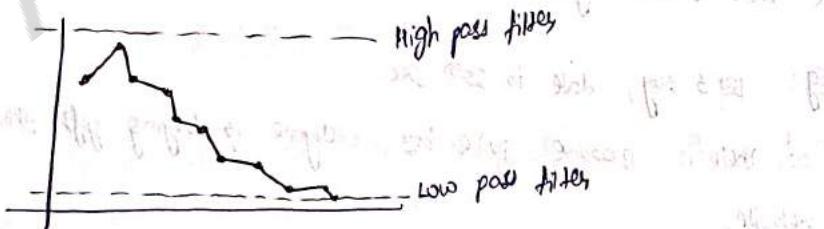
Now we are going to find anomalies in the below time-series data.



→ If there are some extremes that you want to top off i.e. the limits - if it goes above or below the limits, we can say anything beyond these limits will be the anomaly.

This is very prone to errors. One thing it doesn't take into consideration is overall trend of a graph.

Ex:

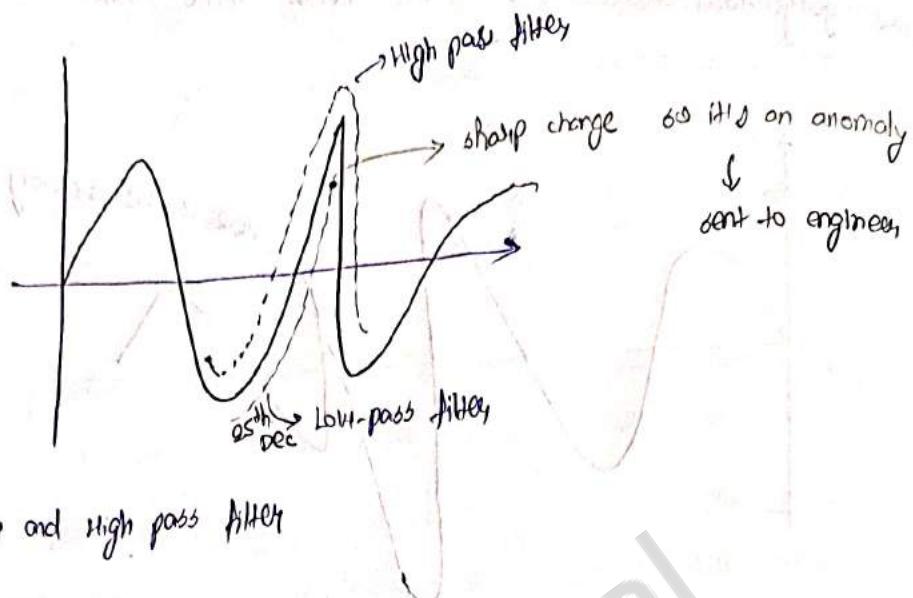


of things are going down and this is happening for months. And when you have a low-pass filter, its going to hit that low-pass filter and you immediately think its an anomaly.

But, this had been going on for months.

To avoid that kind of problem, we can keep a Dynamic low-pass and High pass Filter.

Eg:



Dynamic Low and High pass Filter

Here, this sticks close to the graph.

→ If there is a very sharp change in the metric, then it's going to be flagged as an anomaly.

This can be sent to the engineer who is looking into it.

The only drawback here is that you are not taking the time into consideration.

Eg: Let's say, date is 25th Dec.

so, there is massive spike bcz everyone is buying gift through your website.

Once it's over, there is a dash and may be there is new year offer that starts from Jan 1st. ↓

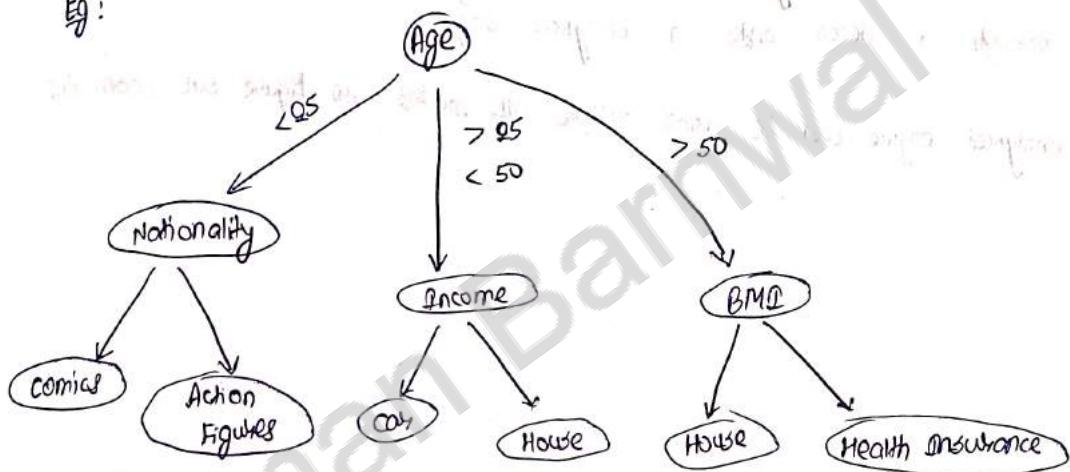
What you can do is "You can take previous year data and see that the change was dramatic during 25th Dec, so this is not an anomaly. It's normal."

→ There are a lot of algorithms which find anomalies in a time-series data.
one of the algorithm is Isolation Tree.

Isolation Tree:

It is basically a decision tree. It makes decisions based on the inputs that you have.

Eg:



Isolation Tree

Here, we are partitioning the data based on certain features. It is trying to make decisions based on the partitioning data into pieces.

With this idea, we can try to partition the data such that in the minⁿ no. of cuts, we are getting able to get something which is an outlier.

↓
That's what an anomaly is.

Eg: If your age < 10 → You probably buy nothing.

But if you end up buying a house, I am assuming that no. of partitions required on the data that set you apart is going to be really really low.



That also means that you are an anomaly.

- If you run time-series data through the isolation tree, you should be able to find such anomalies using this kind of pattern.

Note: Stage one should correctly give

→ We are more detecting for false alerts rather than false one.

→ We want some uniformity in the sampling rate, so all those metrics are

sent through a sidebar onto an analytical engine.

This analytical engine can be using multiple ML models to figure out anomalies.

Moving from monoliths to microservices!

Today, we are talking about how to move from monoliths to microservices.

Monoliths:

They are rather large code bases, which contains all the logic required for you to run your application anywhere you like.

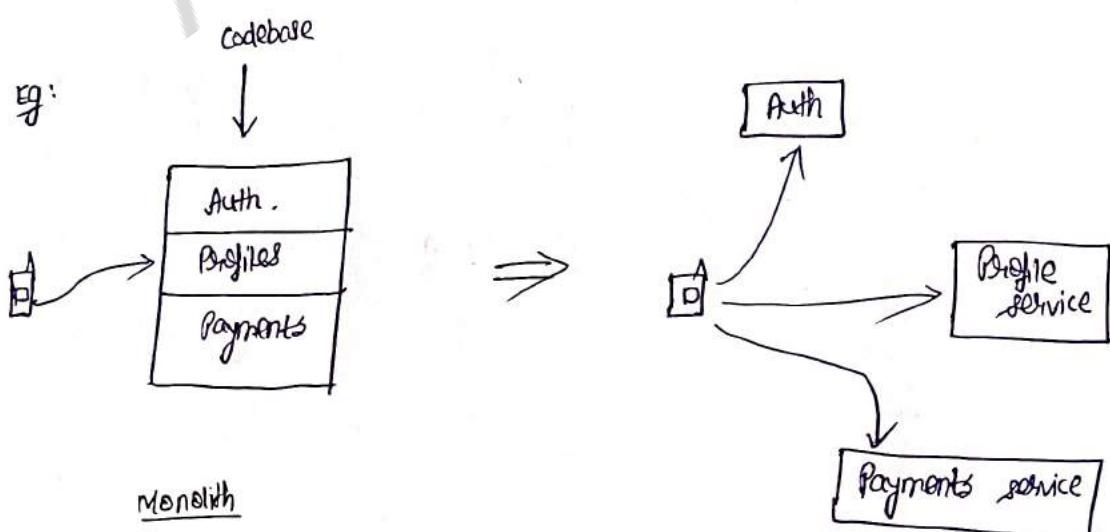
Monolith = single code repository

All of the relevant code in this project is kept in a single repository.

Microservices:

→ While in microservice architecture, you have cellphones which connect to different services based on what the requirement is.

Here, the code had been broken into pieces and converted into services themselves.



→ There are advantages of Microservices over monoliths:

- ① If you need to make any feature change (let's say, your payment now also needs to accept PayPal payments).

↓

All you need to do is just go to Payment service and make the change.

i.e. There is a good separation of concerns.

- ② Coding in microservices is sometimes easier.
Bcz all you need to worry about when you are making a feature change is just the code in that particular microservice.

so, your expectations are well-defined (in the payment service, here).

Your responses are well-defined and that's all, you need to care about.

$$\boxed{\text{Expectations} = \text{API} + \text{SLA contracts}}$$

Your assumption is that the engineers in rest of the services will respect the contracts that you are making with them.

Engineering these services is actually easier.

- ③ The critical advantage is here is that you can make deployments much easier with microservices.

Eg: when there is a change in authentication service, you can just deploy the authentication service separately.



There might be a service which is not critical.

Eg: You have some banking service which banks your restaurants.
So, if it stops for 2 sec, that doesn't make much difference. So, the deployments are easier.

Monoliths also have certain advantages.

Note:

The time when you move to a microservice architecture is not when you have a lot of scale (ie not when all of your users are coming in).

But you move to a microservice architecture when your team has scaled.

→ When you have a large team and all of them want to move about individually, you have different product teams working on different products.

You want deployment to be smooth. That's why, you move from monoliths to microservices.

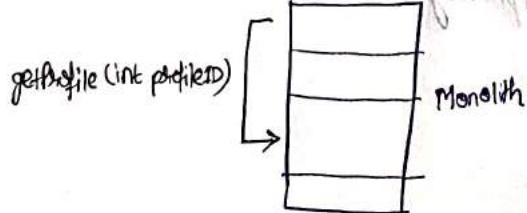
→ For small teams, monoliths are good.

For a large team, microservice architecture is probably the way to go.

(Small teams benefit from consolidated code)

→ Because there is more clear communication happening in the monolith (module A to B), you are doing a single fn call.

The parameters that you are passing into that fn are explicit. They are clear.



so, when you are making that API call, if you make a wrong API call, the compiler is going to say no.

while in a microservice, what could happen is 'ID is earlier int'. Now, you can out of integers, so you made it a string.

i.e. `getProfile (int profileID) → ①`

↓ } This is called breaking change.

`getProfile (String profileID) → ②`

so, all the other services who connect to service ① are going to use the profile with ID as an int. Bcz they are not aware that ID has turned into a string, which means all communication with this service will break.

earlier you have something and now you have something else which is not compatible.
That's a breaking change.

→ so, in microservices, it's much easier to have breaking change, which is bad

but in monolith, bcz the code and expectations are in one place, it's much harder to have these kind of problems.

Advantages of Monoliths:

① Duplication is lesser

② good for small teams

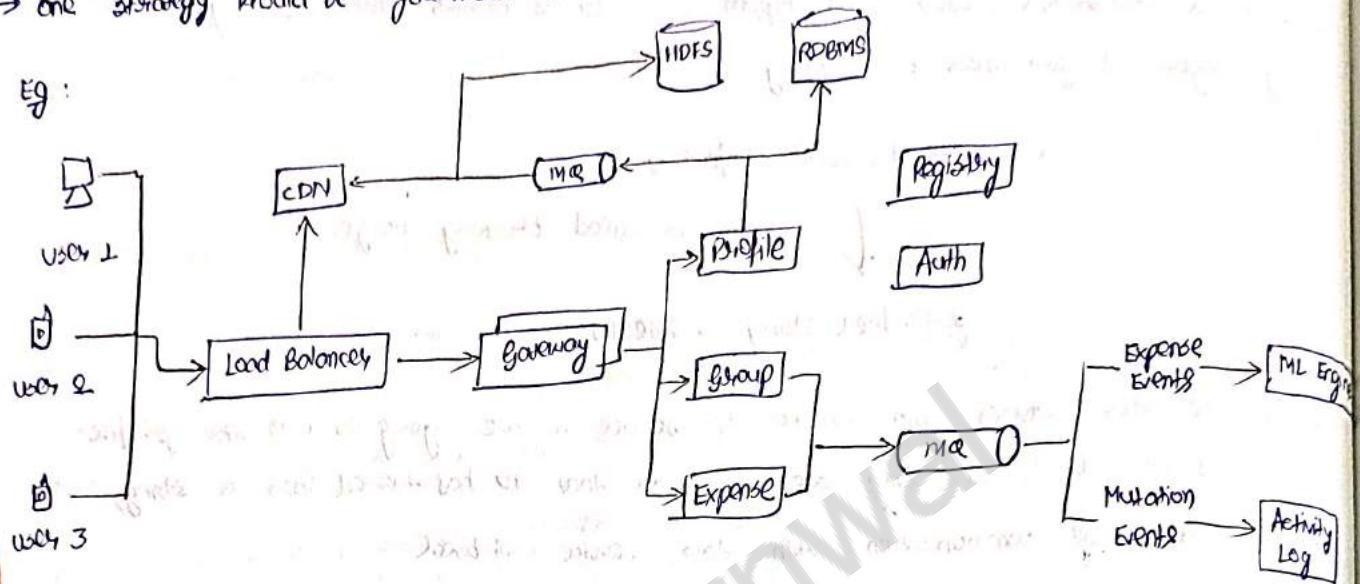
③ contracts are in one place

These are service contracts.

Steps for migration from Monoliths to Microservices:

① → one strategy would be you made this microservice architecture.

Eg :



let's say, some part of users are being redirected here, if they have no issues ie. (100%)

these services are working perfectly along with their databases then you can make the switch from monoliths to microservices.

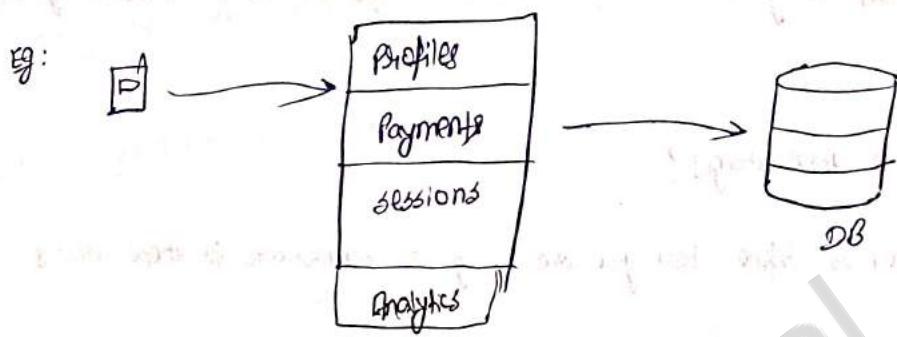
→ But the problem with this is that the engineering challenge is huge bcs you need to write down all these microservices. Also, make sure that the databases are correctly configured.

Then you have to redirect these users. So, the effort for doing this is massive.

i.e. Engineering invest is huge.

~~bcz you need to~~

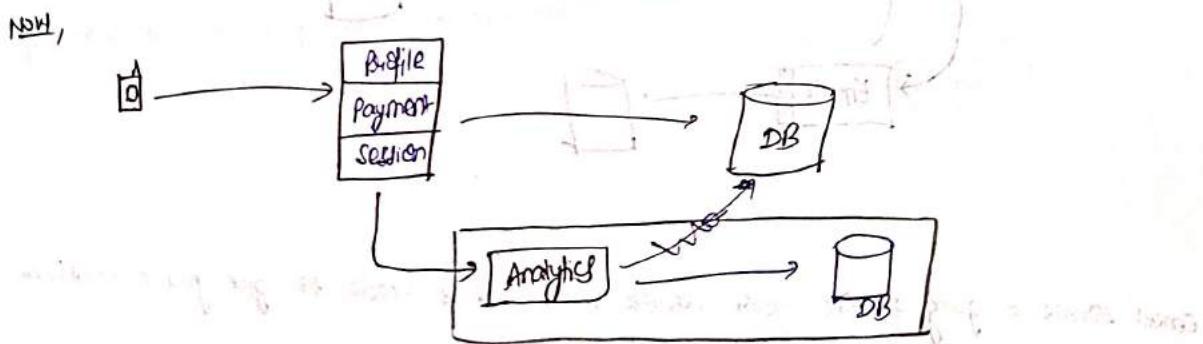
- ② A much better approach is not to invest so much in the start. When you have a new feature in the monolith architecture, you can try to separate that out into a single service.



Let's say, you might need a new analytics feature that you need whenever any of the existing modules produce an event, you need analytics module to process the event and then give dashboards, graphs, reports, etc.

Instead of putting it in the monolith architecture, we separate it out. We make this a separate service.

Earlier you just need to make a fn call in monolith, now you are going to make the call over the network in microservice architecture.
(either HTTP or gRPC)



You cannot access any of its functionalities through a fn call and the analytics service will also have its own database.

So, the analytics service will have its own repository, while the existing 3 modules has its separate code repository.

→ With this, we have opened the ~~possibility~~ possibility for all future services being able to attach relevant services.

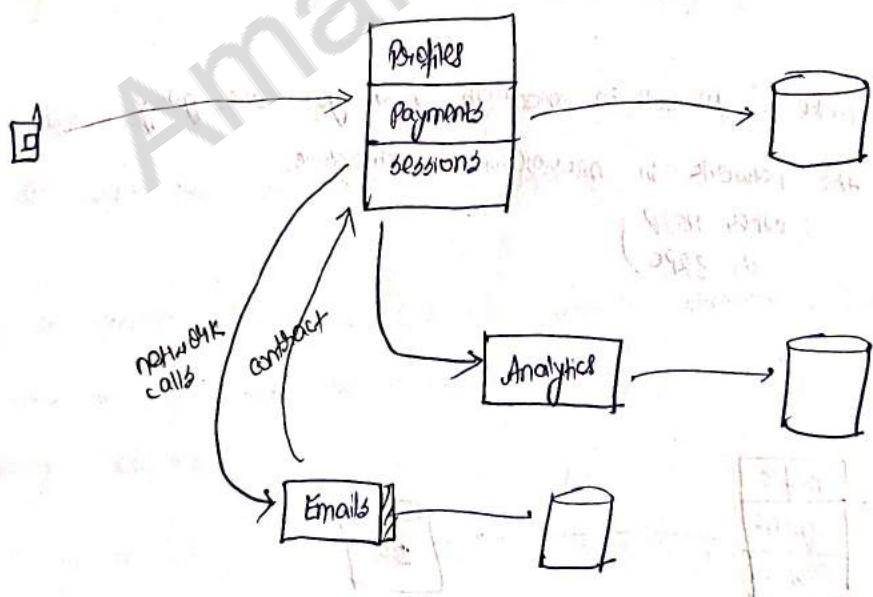
So, ~~now~~ also, we need to now bring in some infrastructure changes to make addition of these services easy.

And, what are those changes?

But firstly, you need to define how you are going to communicate to these services

→ You might find the functionality of sending emails which is being used by Payments, Profiles, Analytics, etc. so, you separate that out into separate service.

Eg:



Email service is going to take your network calls, but it needs to give you a contract that for the profiles that I have, ^{I have} ~~this is the id - as string~~.

This contract is to be maintained by Emails service. And whenever an external service wants to call a fn over here, it has to read this contract, construct an object as per the requirement and then send the request ^{over} network.

② Contract:
For this, we need to define contracts for each microservice.
These are also called clients.

→ whenever a service in session service wants to call an email service, it doesn't really need the contract.
Each service has their own client

~~We will m~~
If I'm an engineer of Analytics service, I say, if well you are going to good my contract, construct the object in my way. That's pretty bad.

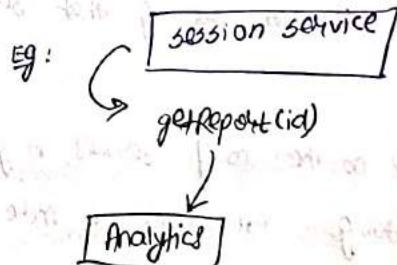
Instead, I will write some piece of code, make a library out of it. And, anybody who wants to talk to Analytics, you use this library to make those API calls.

So, if session service wants to talk to Analytics, all it needs to do is download the library

What happens if my library is not up-to-date?

→ the code will break most likely.

So, you need to continuously update the libraries.



Routes :

- ⑥ We need to have a mechanism for handling requests from any service to another service.

We need some sort of router. Most likely, it's going to be load balancer, the service registry which routes requests based on service capabilities.

② Simplifying Deployments →

The simplest way to deploy any service is to do a SCP.

which is copy of the code that you have in GitHub repo onto an AWS box or GCP box

~~onto an AWS box or GCP box~~
and then going onto that box and running the service.

→ SCP repo - dest cloud

→ SSH box

→ Run steps

Now, you are copying the code from one place, then keep it in the cloud as service and then running that service by firing the command.

But as the no. of services in your organization increases, that's not you want to be doing. You want a more seamless way.

May be on every (git push) to master, you want the deployment to be dead. Bcz your tests have run through and you are sure that everything is fine.

* Automated Deployments

- You want to simplify deployments using tools like
- ① Jenkins
 - ② using containers instead of deploying it on the box
that makes your DevOps team happier.
 - ③ Service Dashboards

④ Communication:

communication between services can be varied.

e.g. Let's say the session service talks to Analytics service using a message queue.

If the event doesn't reach Analytics service, what's the big deal?

only the reports will be wrong. It's not like users is going to drop off.

→ While in payment service, you might need a much more immediate response.
so when profile service is talking to payment service, you might want a request-response architecture so that you know exactly what happened immediately after a request has been sent to payment.

⑤ Logging:

if a request is sent from user to session service, then it goes to analytics service which triggers something in profile service & then it went to payment service.

There is no way that you can track all this, by logging into each device and checking the logs.

→ Instead you want to take all the logs in all the services and push it into a single repository (i.e. a single database). You could use the Elastic Stack.

few things to keep in mind for moving from monolith to microservice:

① single source of truth:

- * Every microservice needs to manipulate the data that it is responsible for.

Every component in your architecture has to have a single source of truth, when it comes to microservices.

↓
it only happens when you have a dedicated data store for each service

② Condense business responsibilities into a single place:

It's very tempting to take services and break them down into simpler and simpler components.

But the important thing to remember is that what is the responsibility of this service? And, when we are breaking it down, are we also separating out the responsibilities?

Eg: Let's say, the profile service needs to take info from various external services like Google, Facebook, LinkedIn, etc.

Now, we might create a new service which acts as an intermediary between info from these external services and profile service.

That seems like a good idea.

But the question to ask is →

→ Is data being used by any service except the profile service? → Here, if it's no

i.e. APIs being used by multiple services?

Q When there is a change in the profile service requirement, is there a change in the service requirement?

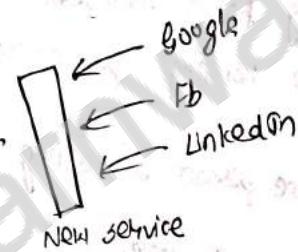
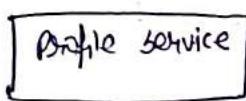
→ Here, it's most likely Yes

Decoupling

Q Is the business requirement of new service separate from profile service "new service"?

i.e. separation of concerns

Eg:

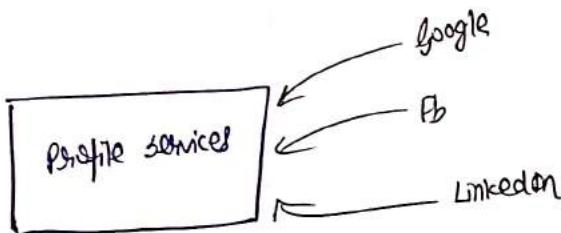


So, what should ideally happen is that this should be a component of the profile service. You can make a library out of it.

You can put it in the utils package.

The codebase should not be separate.

so,



* Don't break it down just because your engineering will get much simpler.
Really do consider, whether you need to separate it further and further.

③ Initial infrastructure cost is high!

Can we afford the infrastructure requirements?

for building a microservice architecture

The cost of logging, the cost of building a deployment infrastructure \rightarrow All of this is usually not worth it for a small team.

* for a startup, going for a microservice architecture is fine if the total no. of services that you are handling per person is less or equal to 2.

i.e. one person is handling ≤ 2 microservices.

It also depends on user scale.

* for a medium org, 1 person should be handling ≤ 1 service.

* for a large org, 2 to 4 persons per service.



It's important to note that microservices are much more complex than traditional monolithic architectures, so it's important to carefully plan and design them to ensure they're effective and efficient.