Chair for Decentralized Information Systems and Data Management
TUM School of Computation, Information and Technology
Technical University of Munich

Cloud Information Systems

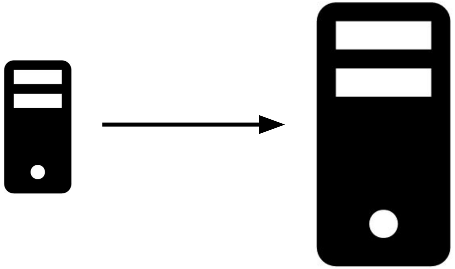
Exercise 9

16th December 2024

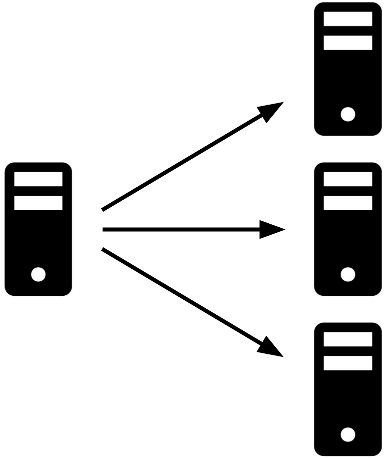
Prof. Dr. Viktor Leis, M.Sc. Till Steinert, M.Sc. Jana Vatter

1. Recap: Scalability

vertical scaling (“scale up”)
→ bigger machines



horizontal scaling (“scale out”)
→ more machines



1. Recap: Scalability

Cloud makes scale up easier

- migrate to bigger VM slice
- live migration would make this quite easy
- enables high-bandwidth and low-latency communication

but: very large servers can become uneconomical

Cloud favors scale out architecture

- add/remove machines when workload changes
- enables fault tolerance through redundancy
- works nicely for stateless services (eg, web servers)

but: downsides like
→ distributed state management,
→ new failure modes,
→ bottleneck: network bandwidth and latency

1. Recap: Scalability

If the problem is not trivially partitionable, it is best to first

- scale up to one medium-size server and
- only if that is not enough to scale out to a cluster of medium-size servers

+ exploits fast communication

+ hardware resources are generally proportional to cost

+ enables “arbitrary” scaling

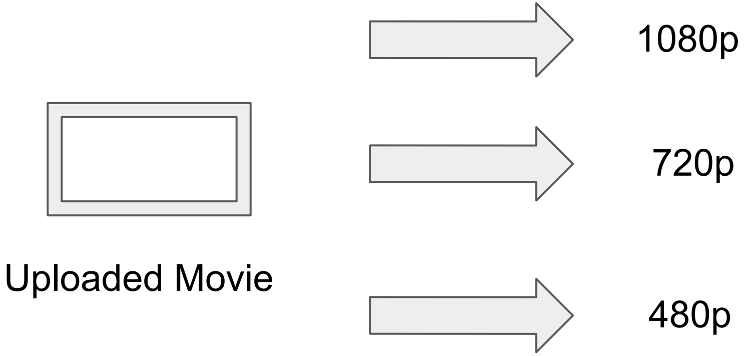
- this implies one has to exploit two levels of parallelism

2. Live Exercise: Netflix Video Transcoding

“While onboarding more streams to the service, we noticed that running the infrastructure at a high scale was very expensive.”

[Scaling up the Prime Video audio/video monitoring service and reducing costs by 90%]

2. Live Exercise: Netflix Video Transcoding



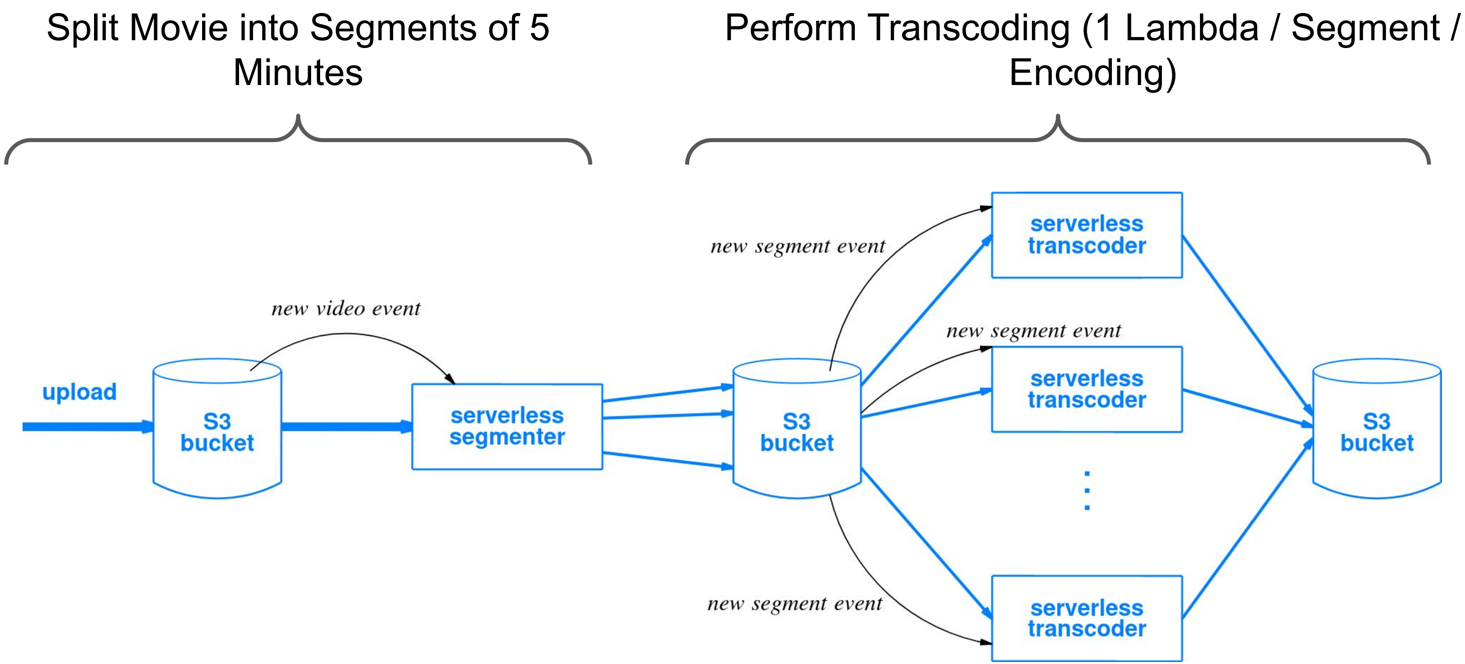
- ➔ Publisher Uploads new Movie to Netflix
- ➔ Different Customers have different Demands (4K TV vs Mobile Phone) and Subscription Models (Basic vs Standard vs Premium)

2. Live Exercise: Netflix Video Transcoding

Netflix Gains New Efficiencies Using AWS Lambda (4:15)

Netflix Keynote

2. Live Exercise: Netflix Serverless Video Transcoding



For Details, see: Comer, “The Cloud Computing Book”

Q: What is the Cost of Transcoding a two hour Movie ?

2. Live Exercise: Assumptions

- **General Parameters**:
  - 2 hour Movie in 4K Quality
  - 30 Frames per Second
  - 5 Minutes per Segment
  - 6 Output Resolutions (Assumption based on # of Video Qualities on YouTube)
- **Video Transcoding:**
  - We use FFMPEG on Graviton Instances
  - **c6g.4xlarge:** 41 FPS
  - **c7g.4xlarge:** 59 FPS
- Frames are independent and can be processed in parallel -> Processed FPS scale linearly with the number of vCPU cores utilized
- We can ignore S3 Storage Cost -> Actual Deployment happens via CDN

AWS Graviton Benchmark

3. Questions