CAPACITY ESTIMATN

Q] What is the total amt. of storage YouTube needs per day?

$$\left(\left(1\ \text{Billion}\right)\left(\frac{1}{1000}\right)\right) \times 10 = 10^7 = \text{No. of minutes uploaded/day}$$

- 1 Billion: no of users
- $\frac{1}{1000}$: 1 in 1000 users upload 1 vid/day on an avg.
- 10: avg. vid len in min.

2 hrs vid $\approx$ 4GB

lower quality 2 hrs vid $\approx$ 0.4 GB (ML can be applied to compress)

$\therefore$ lower quality 1 min vid $\approx \frac{0.4}{2 \times 60} \times 1000\,MB \approx 3\,MB$

→ space required to store 1 min vid.

$\therefore$ Raw vid storage on disk = 3 MB (data/vid) $\times 10^7$ (no. of min uploaded/day)

$\approx 30\,TB$

Replicaten this storage for redundancy, fault-tolerance and faster accs across servers spread out globally $\approx 3 \times 30\,TB \approx 90\,TB$

Now, this is for 1 resolu^n (res^n) of this vid.

Say for a vid 720p res^n = 'x' mb
480p res^n = x/2
360p res^n = x/4
240p res^n = x/8
144p res^n = x/16

add → ≈ 2x

data reqd. (in MB) to store all res^{n}s of a vid

∴ Total storage = 90TB (2)
= 180TB
≈ 0.2 PB

Also, say the avg vid ~~size~~ space for 2 hrs vid is greater than our assumed 0.4GB but its obviously gonna be less than 4GB

∴ Total storage ∈ [0.2, 2] PB

M-2 to get data reqd. by a 1 min vid.

data reqd = 60 x 24 x (10kB)
60 sec — fps — size for 1 img. actn as 1 frame

≈ 14 MB

∴ We can choose diff paths to calc. same stuff. In M-2, we got 14 MB instead of 3MB (which we got in M-1)

IMP: Difference accented 10^2 order of magnitude are

↳ vid name, thumbnail, vid desc etc.

Caching metadata reqd. for a vid

Thumbnail space ≈ 10kB
Other data space << 10kB so ignore
And we gonna cache popular vids ↳ evergreen vids
↳ vids uploaded in last 2-3 months

Estimate: Total space for popular vids' metadata ≈ sum of spaces reqd for metadata of 90 days worth vids

No. of vids uploaded in 90 days ≈ 1 Million
∴ Cached data = 10kB x 90 x 1 Mill.
= 10^9 KB
= 1 TB RAM
↓
can't fit in one. Gotta use multip computers to cache

∴ No. of computers acting as caches (ideally) a.k.a nodes = $\frac{1\,TB}{16\,GB}$ = 64

↑ say each comp has 16 GB RAM

∴ In Real life Gotta have redundancy = 64 x 3 x 2 ← redundancy ← gotta 2x as 50% of time nodes are crashin.

∴ No. of nodes ≈ 500.

Date:
M T W T F S S

No of processor reqd. to process data

1 million vids / day

$\therefore$ 10^7 mins / day → to be processed

1 min space reqd. by our M-1 ≈ 3 MB data

$\therefore$ ~~No of~~

Data processin speed = (3×10^7) MB / day

$= \left(\frac{3 \times 10^7}{3600}\right)$ MB / sec.

$\approx \times 10^3$ MB/sec
$\approx 8 \times 10^4$ MB/sec (added redundancy)

for each computer, to process a vid, 3 things happen

- Readin → 10 ms
- Vid. img processin → 20ms
- Writein → 20ms

→ 50 ms reqd. to process 1 MB of data

$\therefore$ Work to be done $= (8 \times 50 \times 10^4 \times 10^{-3})$

↓ Work to be done by a batch of computers per sec.

4000 computers/processors run // ly every sec to do this work = 4000 sec worth work to be done in 1 sec