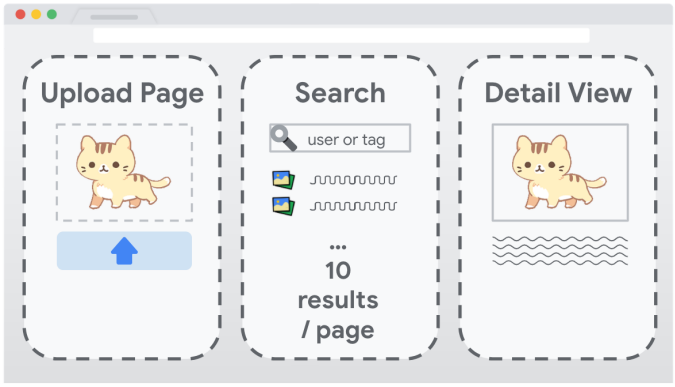
Google SRE Classroom: Distributed ImageServer

Your Task:

Design an image-serving service that clients all over the world can use to upload and view images.

The UI provides three pages:

- **an *upload* page** for uploading an image and its tags
- **a *search* page** for searching images by uploading user or tag; 10 search results are returned and displayed on each page; each search result includes a thumbnail and image metadata
- **a *detail* page** showing a specific image at full resolution and its metadata; results on the search page link to the result images' detail pages



Traffic & Data Sizes

- Global user base of 1 million users
- Users average 50 uploads / day
- Users average 50 searches / day, each of which yield 2 detail view click throughs
- Sizes: avg image = 4MB, avg thumbnail = 256 KB, avg image metadata = 8 KB

SLOs

- Serve the *detail* page (1 image, full resolution), within 200 ms at the 99.9%ile
- Serve the *search* page within 250 ms at the 99.9%ile
- These SLOs only apply to images that are 30 days or fresher

Infrastructure

- 3 datacenters in different regions: Europe, Asia, North America
- Datacenters are connected with a 100Gbps connection over a 99.99% available network
- Reliable distributed storage system: abstracted RPC interface
    - Globally replicated, eventually consistent in O(minutes)
    - Write RPC
        - Request: Full-size image
        - Response: Storage key
        - SLO: Starts within 200 ms at 95%ile
    - Read RPC
        - Request: Storage key, image size
        - Response: Image at requested size
        - SLO: Starts within 100 ms at 95%ile
- Authentication and Authorization: all users are already authenticated and trusted clients
- You can use software infrastructure components if you can explain how they work (at a high level), and estimate how they will scale and perform.

Hardware

Machines available:

- 24GB RAM + { 2 x 1TB SSD, or 2 x 2TB HDD }
- 8 cores
- 10 Gbps ethernet connection