Gmail Storage and Processing Requirements

1. Daily Storage Requirement for Emails

Assumptions:

- Number of users: 2 billion
- Number of devices per user: 2
- Number of users opting for 2-step verification: 10% (~200M)
- Average email size: 200 characters
- Average attachment size: 1 MB
- Email distribution per user per day:
    - 20 spam emails
    - 20 marketing emails
    - 10 useful emails

Calculation:

Email data:

- Total emails = 50 * 2B = 100 billion emails per day
- Total email storage = 100B * 200 characters = 20 TB

Attachment data:

- 5% of emails have attachments
- Attachment data = 5% * 100B * 1MB = 5 PB

Total space required:

- Without optimization: 20TB + 5PB = ~15 PB per day (including redundancy factor of 3)

Optimized storage:

- Deduplication reduces total emails to 15 per user
- Storage needed: (15 * 15 / 50) * 5PB = 4.5 PB
- Compression reduces size further by 50%
- Final storage requirement: ~2.5 PB per day

2. Storage Requirement for User Profile Data

Assumptions:

- 2 billion users
- Average name size: 15 characters
- Date of birth: 8 characters
- Email address: 20 characters
- Total basic profile storage: (15 + 8 + 20) * 2B = ~100 GB
- 10% of users have a profile picture, each ~100 KB
- Total profile picture storage: 2B * 100 KB * 10% = 20 TB
- Redundancy factor of 3

Calculation:

- (20 TB + 100 GB) * 3 = ~60 TB in total

3. Processing Power for Virus Detection

Assumptions:

- Attachments to scan per day: 5% * 15 * 2B = 1.5 PB
- 5 I/O reads per attachment
- Read speed: 20 ms per MB

Calculation:

- Total time required: (1.5 * 10^9 MB) * 0.1 seconds per MB = 1.5 * 10^8 seconds
- Convert to days: ~1500 days
- Required parallel processes: 1500
- With 50% capacity buffer and load handling: 1500 * 4 = 6000 processes

4. Processing Power for Spam Detection

Assumptions:

- Total emails to process: 15 * 2B = 30 billion
- Size per email: 200 bytes
- Total email data: 30B * 200 bytes = 6 TB
- Spam detection requires 5 I/O reads per email
- Read speed: 20 ms per MB

Calculation:

- Total time required: 6 * 10^5 seconds
- Convert to days: ~6 days
- Required parallel processes: 6
- With 50% capacity buffer and load handling: 6 * 4 = 24 processes

5. Contact Data Caching

Assumptions:

- 1% of users active at any time: 2B / 100 = 20 million
- Top 10 contacts per user are most accessed
- Assuming overlap, unique cached contacts: 20M / 10 = 2M
- Profile picture cache size per contact: 100 KB
- Total cached data: 2M * 100 KB = 200 GB
- Required machines:
    - Each machine: 64 GB
    - Machines required: 200 GB / 64 GB ~ 4
    - Accounting for fault tolerance and localization: 4 * 3 * 10 = 120 machines

Conclusion

By applying optimizations like deduplication and compression, significant reductions in storage requirements can be achieved. Virus and spam detection require parallel processing to ensure efficiency. Caching strategies help in reducing lookup times for frequently accessed data.