Homework #2: Caching and Object Storage

1. RAND Caching (50/100)

In this exercise you will write a program that simulates the behavior of a caching algorithm under a given workload. You can write the code and plot the results in your preferred language and software. You have to submit the graph, and a link to a git with your code.

Assume a range of page addresses $[0, 1, \dots, N\_1]$ where $N = 2{,}000$

Assume a cache that can hold C pages, $C = 20, 50, 70, 100, 200$

(i) Generate a sequence of referenced pages
Write a program that generates an "80-20 Read Workload"; generate a sequence of 1000 referenced page addresses to read with the following behavior:
- 80% of the references are made to 20% of the pages $[0, 1, \dots, N\_1]$
- The remaining 20% of the reference are made to the remaining 80% of the pages $[0, 1, \dots, N\_1]$

(i) Implement two caching algorithms

- **RAND** – In case of a Miss, replace a random page in the cache
- **OPT** - In case of a Miss, replace the page that has the greatest forward distance, i.e. that will be accessed furthest in the future

Note: the implementation can be straight forward, no need to write an "efficient" implementation. The point here is to understand the behavior of the algorithm RAND compared to the optimal OPT.

(ii) Run the simulation:

For each cache size $C = 20, 50, 70, 100, 200$:

1. Generate 10 sequences of referenced pages
2. For each sequence, compute the Hit rate for RAND and for OPT
3. Compute the average Hit rate for the 10 experiments

(iii) Plot the graph.
Now, plot the average Hit rate as a function of the cache size C for the two algorithms.

2. Object Storage (50/100)

Goal: The purpose of this exercise is to learn what is an object store, its APIs, what are its main characteristics and how to use it when designing a system.

Background:

As an example, we will look at Amazon S3, but you can do the exercise on and cloud object store service, such as

- Amazon S3
- Azure Blob Storage
- Google Storage
- IBM Cloud object store
- Oracle Cloud Object Storage

You should use the following links to answer the questions in this exercise (but if you prefer you can use other services to answer these questions):

- Amazon S3
- Amazon S3 pricing
- S3 calculator (part of Amazon Calculator)

Question #1 – from storage to service. Pick one cloud provider and describe the following. Answers should be short. (20/50)

1. What are the main cost factors to consider when using object storage?
2. What is a Service Level Agreement? And explain the two factors that make the SLA of an object store service
3. List 2-3 additional features of an object store service and explain (in a few words) what they do
4. List 2-3 best practices to reduce cost

Question #2 – Design a Data Service (30/50)

Backup and Archive service. Build a library of recordings of phone conversations in a Call Service Center. The Call Center is located "on the internet" but wants to use a cloud service to store the conversations with its customers.

Assumptions:

- *Average conversation duration is 2 minutes. Therefore, a conversation (digital audio) is about 20MBs*
- *The Call Center is active 5 days/week.*
- *There are 50 customer representatives in the center*
- *Each customer rep generates 250 calls a day (~10 hours each day)*
- *A call must be kept in the repository for 6 months; after that it can be archived*
    - *Every conversation is uploaded to the cloud once it's complete*

- *1% of the conversations within the last six months will be recalled to the internet (out of the cloud)*
- *Every week, the oldest conversations are moved to a deep archive storage*
    - *Regulation requires to keep them for 7 more years*

1. How would you organize the data? Buckets/object names?
2. Compute storage and ingress costs/month
    a. Give the ingress metrics (storage, operations, bytes) of the data that is generated in 6 months
    b. Calculate the cost for uploading these conversations on one of the vendors
3. Compute egress costs/month due to recalls
4. Compute the storage costs of the archival storage (e.g. Glacier) and its growth for 7 years
5. What could be done to reduce the archiving costs? Give one example