CSC 547 - HW 2

Ashwin Sapre
Shruti Kohakade
Chinmay Srivatsa

Regular Problems

Problem 1

Nutanix (https://www.nutanix.com/ is a manufacturer of storage products; they have a local, RTP presence and they hire NCSU graduates frequently. It might be a good idea to know the company a little better...

1. *Find the model numbers of a few (hardware) storage devices with capacities over 1TB.*

   Listed below are a few hardware storage solutions offered by Nutanix:
   - https://www.nutanix.com/products/hardware-platforms/specsheet
   - NX-8150N-G8: offers 184 TB storage, 4096 GB memory, 40 cores. It uses self-encrypting and NVMe SSDs for storage. Main use cases include business critical applications, disaster recovery, big data analytics.
   - NX-3170N-G8: offers 92 TB storage, 2048 GB memory, 32 cores. It uses self-encrypting and NVMe SSDs for storage. Main use cases for this device are private cloud, end-user computing.
   - NX-1065-G8: offers 43 TB storage, 1028 GB memory, 20 cores. It uses a combination of SSDs and HDDs for storage. Main use cases include test and development, remote office/branch office.

2. *In https://www.nutanix.com/sg/solutions, they classify Storage as "Files Storage", "Objects Storage", and "Volumes Block Storage". Describe, in your own words, these types of storage.*

   - Files storage:
     - Nutanix offers "Files Storage" as a way of storing files in a hierarchical manner (like you would expect in Windows, MacOS etc). Files are contained in folders, and each file can be uniquely identified using a path.
     - In a cloud-based file storage system, files are distributed across nodes within an existing cluster, or on a dedicated storage cluster. Multiple clients share storage devices for their data.

        - The cloud service provider can scale up or scale down capacity based on their requirements.
        - The familiar nature of this storage format makes it a popular choice for many people; however, for large amounts of unstructured data, files storage is not ideal.
    - Block storage:
        - Block storage breaks files into equally-sized chunks and gives each chunk a unique ID. These chunks are then stored in separate locations, and recombined when that whole file is to be retrieved.
        - Fault tolerance is achieved by having multiple copies of the same chunk in different locations. One result of this feature is, it allows multiple paths to reconstruct the original file. This increases the speed of retrieval.
        - Block storage is often used to create virtual machine file systems and containerized applications
    - Object storage:
        - Object storage also stores large amounts of unstructured data. It is used mainly for archival and backup purposes.
        - Each object contains the actual data to be stored, metadata, and a unique ID. This object is not stored in directories, it is simply stored in a huge pool of data. For fault tolerance purposes, large sets of objects are grouped together and replicated at different locations.
        - Objects are accessible by the user via an API request, with a GET request to download data, PUT/POST to upload, and DELETE to remove an object.
        - Out of file-based, block and object storage, it offers the highest amount of scalability.

Problem 2

Consider https://www.nfl.com/, the official website of the NFL

1. *Describe as many types of requests as you can, for this website.*

Ans:

1) Sign up request:
   This request is used to create a user account on the NFL website.
2) Sign in request:
   This request is used to log in a user account on the NFL website.
3) Shop request:
   This request redirects to the shopping webpage of NFL.

This request is used to shop NFL gear and apparel online. In this request, different options like search, drop down, search by categories are available.

4) Tickets request:

   This request redirects to the tickets webpage of the NFL.

   This request is used to buy tickets for NFL games. On this web page, different hyperlinks like schedule, buy, sell tickets options are available.

5) News request:

   This request is used to see all NFL news. In this request, news is listed categorically like latest news, popular series. On this news page, different tabs of podcasts, injuries, transactions, NFL writers are available. If we request any of them then it will be a new request like listen to podcasts, see injuries, NFL transactions, see NFL writers.

6) Scores request:

   This request opens a web page where scores of all NFL games are listed. The filter of listing particular scores in some time range is also available.

7) Schedule request:

   This request is used to see the schedules of a current week's matches. All past and future match schedules can be searched using search filter options.

8) Videos request:

   In this request, all the videos owned by the NFL can be watched.

9) Teams request:

   In this request, all teams are listed under NFC teams and AFC teams.

   Each team's profile and full site can be viewed by clicking on the hyperlink provided.

10) Players request:

    In this request, if we enter a player's name it will return player's information in the response.

2. *For each type, mention very clearly, who can potentially create such requests.*

Ans:

1) Sign up request:

   Users can create Sign up requests.

2) Sign in request:

   Users can create Sign in requests every time they log into a NFL website.

3) Shop request:

   Case 1 - Add new apparel in the collection

   This request can be created by NFL administrators who have these rights.

   Case 2 - Buy an item

   This request can be created by a user of the NFL website.

4) Tickets request:
    Case 1 - Change the cost of tickets
        This request can be created by NFL administrators who have these rights.
    Case 2 - Buy a ticket
        This request can be created by a user of the NFL website.
5) News request:
    Case 1 - Add a new news item
        This request can be created by NFL administrators who have these rights.
    Case 2 - Read a news article on the website
        This request can be created by a user of the NFL website.
6) Scores request:
    This request can be created by a user who wants to check the score of a particular match.
7) Schedule request:
    This request can be created by a user who wants to see the schedule of a particular match
8) Videos request:
    Case 1 - Add a new video
        This request can be created by NFL administrators who have these rights.
    Case 2 - Watch a video on the website
        This request can be created by a user of the NFL website.
9) Teams request:
    Case 1 - Update teams information on a website
        This request can be created by NFL administrators who have these rights.
    Case 2 - View teams information on a website
        This request can be created by a user of the NFL website.

10) Players request:
    Case 1 - Update player's information on a website
        This request can be created by NFL administrators who have these rights.
    Case 2 - View player's information on a website
        This request can be created by a user of the NFL website.

3. *For one type only, can you hypothesize about the "pattern" of such requests (e.g., it is uniform, may have peaks, etc.)*

Ans:

Booking game tickets -

The pattern of this request is not uniform. Following are the trends in booking game tickets -

1) This request will peak on weekends because people have Saturdays and Sundays off.
2) On weekdays the frequency of the above request is less.
3) During vacation or public holidays people more often go to watch games so the frequency of the above request would be high.

4. *For one type only, can you guess the estimate how much time a request would take to be processed? State your assumptions clearly.*

Ans:

Request type - Playing a video on the NFL website -

In this request, the time estimate from clicking on the video link till the video gets played is around 8 seconds.

Assumptions-

After clicking on a video link, it redirects to a new webpage and loads video there. It includes network delays, loading of frames, animations, timings of different components.

The factors to be considered in this estimation are -

- Loading time
- Scripting time
- Rendering time
- Painting time
- System time
- Idle time

Problem 3

Consider the set of Facebook users. Supply at least three different criteria for partitioning this set into tenants. Describe how you could identify traffic entering a datacenter as belonging to a specific tenant. Discuss, if possible, advantages and disadvantages of your proposal.

Criteria for partitioning the set of users into tenants:

1. Geographical location: users can be divided based on their geographical locations. The traffic entering the data center can be identified by the IP address of the user and perhaps we can preferentially issue requests from the user to the closest data center to aid in identifying.
    - Advantage: data from each such partition could be stored together for faster access, lower latency as the network calls would be inside the region, and regional data security and governance rules could be applied easily.
    - Disadvantage: If locations belonging to a partition have better connectivity with other geographic partitions due to better infrastructure, grouping them by

geography would negatively affect the service. We may also need to store replicas of data in other partitions for disaster management which could be expensive and potentially conflict with regional data governance rules.

2. Based on Profile:
    - Personal profile: May not post much but consumes content (may not receive payment).
    - Content Creator (brand): Posts a lot of content, potentially selling something (may receive payments).
    - Business/ecommerce: Businesses selling and advertising products on facebook and facebook marketplace. (Posts a lot and receives payments).
    - Identification: The traffic entering the data center could be identified by the service being accessed (like if a payment gateway is being used). Perhaps we could have a setup where the request is routed to a certain data center depending on the type of profile issuing the request.
    - Advantage: Services would have personalized optimization for each type of user profile and the cloud provider could maintain quality standards while servicing requests and scale as required.
    - Disadvantage: However, any crisis would affect an entire partition of tenants globally, and rerouting their requests to other partitions would affect the tenants of other partitions as well.
3. Based on usage: Tenants could also be partitioned based on their specific usage of features on Facebook, like users playing games, users on messenger sending messages and advertisers posting listings and advertising.
    - Identification: The traffic entering the data center could be identified by the service being accessed (Like if it is a request related to a game or a message). Perhaps we could have a setup where the request is routed to a certain data center depending on the service being called.
    - Advantage: each feature could be optimized independently using different data center infrastructures. Games would run smoother and messaging would be faster.
    - Disadvantage: this strategy might require a user to access multiple data centers as one specific user may be using several of these Facebook features. This strategy would also mean that any new features added in the future might require their own data center infrastructure depending on the requirements.

Bonus Problems

Problem 4

How big is the volume of “big data”. Nowadays, big data is a big buzzword.“Data is the new oil” is perhaps one of the most popular catchphrases highlighting the importance of data. Everyone

is looking to analyze it and profit from it. It appears that no one is paying attention to a lowly housekeeping function. So, let us take a look at it.
Continuing in the spirit of Problem 2.2, read [11]. According to this source,

- *The amount of data in the world was estimated to be 44 zettabytes at the dawn of 2020.*
- *By 2025, the amount of data generated each day is expected to reach 463 exabytes globally.*

1. *Find out what the volume of a SSD (Solid State Disk) is.*

   Datacenters have different considerations when purchasing SSDs. For example, they need to withstand high loads, need to be hot swappable (so that when we replace an SSD the whole datacenter does not need to be shut down), and need to be durable.

   We will pick the Intel® Optane™ SSD DC P5800X Series 800GB variant to carry out our calculations. This is a U.2 form factor SSD.

   The approximate volume of this SSD is:

   70mm x 100mm x 15mm = 105000mm^3 = **$1.05 * 10^{-4}$ m^3**

2. *How much datacenter space in cubic meters would a zettabyte consume? State your assumptions clearly.*

   Assumptions:
   a. We are using the Intel® Optane™ SSD DC P5800X Series 800GB variant for this calculation
   b. We assume that in a datacenter, 60% of the actual volume goes towards storing the SSDs. The rest of the space is occupied by overhead cables, air conditioning, alleyways, control centers etc

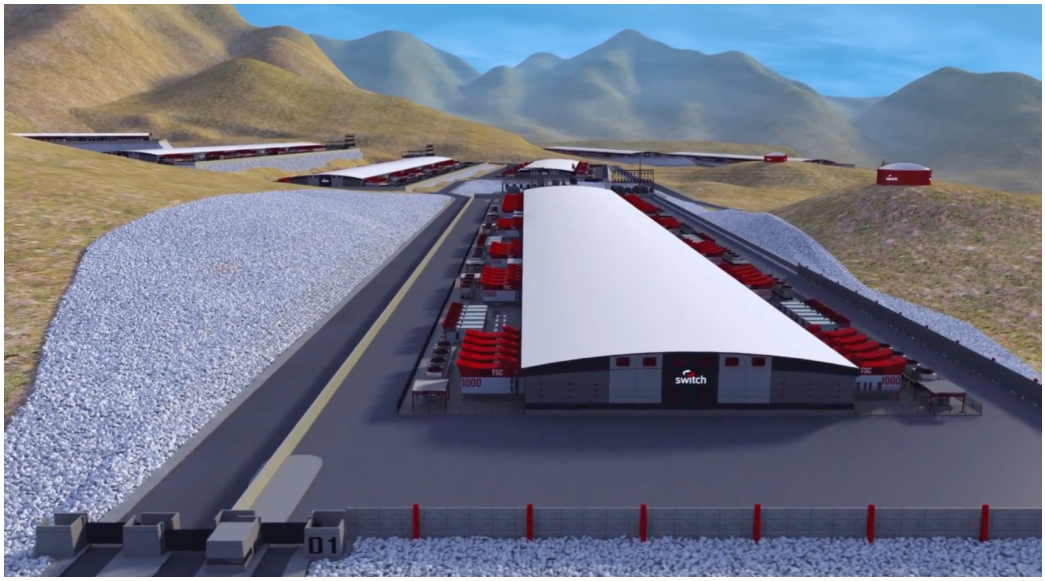
   Space needed per GB = 1.05 * 10^{-4} m^3 / 800
   = **$1.3126 * 10^{-7}$ m^3/GB**

   1 zettabyte = 10^{12} GB
   Therefore, actual data center volume required for 1 zettabyte = 1.3126 * 10^5 / 0.6 m^3
   = **$2.187 * 10^5$ m^3**

3. *Citadel, a datacenter located in Tahoe Reno, Nevada, covers an area of 7.2 million square feet. It is considered one of the largest data centers in the world. Find out the volume of the Citadel.*

   The citadel has a warehouse design. There are no exact height details available online, but from the pictures we can make an educated guess that it is around 2 storeys tall,

which is 20 feet. We also have to assume, for the sake of the calculation, that the structure is cuboidal in shape.



Therefore, volume = $7.2 * 10^6 * 20\ ft^2$
$= 144000000\ ft^3$
$= 4.0776259 * 10^6\ m^3$

4. *Can we store all the data the world had at the dawn of 2020 in Citadel? State your assumptions clearly.*

- We assume that we have an unlimited number of Intel® Optane™ DC P5800X Series 800GB variant SSDs for storage.
- We assume that humanity has dedicated all its effort to this operation; no further improvements in SSD technology or construction technology have taken place.
- Similar to the previous question, we assume that in a datacenter, 60% of the actual volume goes towards storing the SSDs. The rest of the space is occupied by overhead cables, air conditioning, alleyways, control centers etc.

Total usable volume = $0.6 * 4077625.9\ m^3$
$= 2446575.54\ m^3$
$= 2.446575 * 10^6\ m^3$

Data to be stored = 44 zettabytes
We had calculated before that the volume per GB in an Intel® Optane™ DC P5800X Series 800GB variant SSD is $1.3126 * 10^{-7}\ m^3/GB$.

To store 44 zettabytes, we would need:

$1.3126 * 10^{-7} * 44 * 10^{12}\ m^3 = 5.77544 * 10^6\ m^3$

As $5.77544 * 10^6\ m^3 > 2.446575 * 10^6\ m^3$, **we would not be able to store all the data in the world at the dawn of 2020 in the Citadel.**

5. *Repeat the last question for the year 2025.*
   Assumption: the data in question is the data generated from 1 January 2025 to 31 December 2025.

   Amount of data to be stored = 365 * daily amount
   $= 365 * 4.63 * 10^{11}$ GB
   $= 1.689 * 10^{14}$ GB

   To store that much data, volume needed $= 1.689 * 10^{14} * 1.3126 * 10^{-7}\ m^3$
   $= 2.21 * 10^7\ m^3$

   As the total usable volume of the Citadel is $2.446575 * 10^6\ m^3$, **we will not be able to store all of 2025's data in the Citadel** (in fact, the data generated in 2025 alone would be around 10 times greater than the total amount generated till 2020)

6. *Assume linear growth of data in the future. When will we cover the island of Ibiza, Spain with datacenters that just store data? State your assumptions clearly.*

   Assumptions:
   - Like the Citadel, the Ibiza datacenters will be 2 storeys (20 ft high)
   - We will still be using Intel® Optane™ DC P5800X Series 800GB variant SSDs for storage, and we will be assuming Intel has an unlimited supply.
   - Similar to the previous question, we assume that in a datacenter, 60% of the actual volume goes towards storing the SSDs. The rest of the space is occupied by overhead cables, air conditioning, alleyways, control centers etc.
   - We assume that the entire area of Ibiza is available for construction (no reserved beaches, bars, national parks and so on.)

   Total land area of Ibiza, Spain = 572.6 million m²
   Total datacenter volume = $572.6 m^2 * 6.096 m * 10^6 = 3.5 * 10^9\ m^3$
   Total usable volume = Total datacenter volume * 0.6
   $= 3.5 * 0.6 * 10^9\ m^3$
   $= 2.1 * 10^9\ m^3$

   The amount of data required to be stored, "d" days after the dawn of 2020, can be represented by the linear equation: $y = 10^{12} + d*g$
   where g = the amount of data generated per day

The amount of volume required to store this data = y * volume per GB
= $1.3126 * 10^{-7} * y$

Therefore, the final equation becomes:

$1.3126 * 10^{-7} * (10^{12} + dg) = 2.1 * 10^{9}\ m^3$

Given the daily data generation rate g, we can solve for number of days d. According to reference 2.11, 1 quintillion bytes (1 exabyte) of data was being generated daily in 2018, and by 2025 that amount will jump to 463 exabytes. Let us assume that from today (13 September 2022), the data generation rate stays constant at 463 exabytes/day.

Substituting $g = 4.63 * 10^{11}$ and solving for d, we get **d = 34552 days**.

Thus, it would take **94 years and 7 months** to fill a datacenter the size of Ibiza (under our assumptions)