Assignment 4: Distributed Systems Design Requirements and Properties

Objectives

The main objective of this assignment is to address design requirements for distributed systems. The assignment exposes you to tools that are being used in distributed systems that achiever concurrency, reliability, and availability.

General Information and Policies

- Solutions must be submitted through the Quercus system.

Keywords

Data warehouse, safe coding, archetypes, memory safety, safe api, safe deployment, config-as-Code, zero-touch-prod, concurrency, reliability, availability, failover, host memory, bandwidth, cloud TPU.

Resources

1. Interpreting the Data: Parallel Analysis with Sawzall: https://static.googleusercontent.com/media/research.google.com/en//archive/sawzall-sciprog.pdf
2. Developer Ecosystems for Software Safety: https://storage.googleapis.com/gweb-research2023-media/pubtools/pdf/66a86a49722afe4487e014ca80856dd4866f1b53.pdf
3. The Chubby lock service for loosely-coupled distributed systems: https://static.googleusercontent.com/media/research.google.com/en//archive/chubby-osdi06.pdf
4. Bigtable: A Distributed Storage System for Structured Data: https://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf
5. Kelp: QoS for Accelerated Machine Learning Systems: https://ieeexplore.ieee.org/document/8675247

Deliverable

- Assignment report

Each of the following questions maps to the resources listed above Q1 -> R1, Q2 -> R2, etc.

Q1. Sawzall

a. Explain what is Sawzall?
b. Using https://github.com/google/szl, how is szl is related to Sawzall?
c. Identify the parallel processing requirements for processing a dataset.
d. Identify how job scheduling is performed.
e. What is the role of the MapReduce s/w.
f. Identify a system lifecycle for processing a dataset using Sawzall.
g. What would be the purpose of integrating a relational database with the system components.

Q2. Developer Ecosystems for Software Safety:

a. Identify how you would find and fix implementation defects.
b. Explain what a safe system is.
c. Identify how can you achieve safe deployment for microservices and web services.

Q3. The Chubby Lock Service

a. What is Chubby?
b. Why Chubby is needed?
c. What is coarse-grained synchronization? Explain how it differs from fine-grained synchronization.
d. What are three functionalities that Chubby Service implements?
e. What is the duration of the Chubby service lease time?
f. What is the purpose of the KeepAlive messages.

Q4. Bigtable

a. What is BigTable?
b. How does BigTable achieves concurrency?
c. What happens when a client's session expires?
d. List four functions that BigTable uses by deployment of Chubby.
e. What are Bloom filters? How are they used in Chubby?

Q5. Kelp

a. the research studies performance interference between high priority accelerated ML tasks and low priority CPU tasks. Identify the four production ML workloads that research experimented with. Identify the three accelerated platforms that were utilized.

b. What are TPUs and GPUs?

c. Use https://cloud.google.com/tpu/docs/intro-to-tpu to compare between TPUs and GPUs.

d. Bandwidth in ML processing problems is considered as a bottleneck explain how the problem was mitigated.