1a.

Management :

Lambda: User uploads code to run and define any triggers, without the need for managing servers or any part of the infrastructure.

EC2: infrastructure is controlled, configured, and managed by the user.

Price:

Lambda: User pays for number of requests and compute time. It will not be charged for idle time.

EC2: User pays for instance type, running time, and any other optional services. It will be charged when the instances are idle.

Scalability:

Lambda: Scales automatically in response to traffic and manage resources based on demand.

EC2: Requires manual intervention or additional services like auto scaling groups.

Flexibility:

Lambda: Can only be used for specific highly focused tasks due to code size and execution time limits.

EC2: Can be used for mostly any application and even services that require complete access to the environment.

1b.

AWS Lambda's pay model is based on the factor of how many requests are computed and compute time used. It doesn't charge when the instance is idle. The instances usually last for a brief amount of time.

AWS Spot instance's pay model uses bidding to use spare EC2 capacity, the instance runs as long as the bid is above spot price. The instances are usually short lived and are mainly used for computations that can manage interruptions.

1c.

Microsoft Azure: Azure Functions

Google Cloud: Google Cloud Functions

1d.

Lambda: I would use lambda to run a picture optimization software, where I can automatically scale while there is a need to optimize more pictures, while being only charged for the number of time I run the computation and the compute time used.

EC2: I would use EC2 to run a web application where I would need to use custom configuration such as hosting a database, optimize server to my specific requirements, choose resources that can manage the workload of my website.

2a.

AWS - NVIDIA A10G Tensor Core GPU with the highest available memory of 256 GB (g5.15xlarge)

Google Cloud – NVIDIA A100 80GB Tensor Core GPU with highest available memory of 170 GB.

Microsoft Azure – NVIDIA A10 GPU with highest available memory of 880 GiB (Standard_NV36adms_A10_v5)

2b.

AWS - NVIDIA T4G Tensor Core GPU with the lowest available memory of 8 GB.

Google Cloud – NVIDIA A100 80GB Tensor Core GPU with lowest available memory of 85 GB

Microsoft Azure – NVIDIA T4 GPU with lowest available memory of 28 GiB.

2c.

GPU is offered by most of the cloud providers, while TPU which is developed by Google is primarily available in Google Cloud and is less common in other cloud platforms. GPU can be used for multiple applications and a wide range of computations. Although TPU is primarily used for machine learning tasks like deep learning.

3.

M1(a,b):

Output (key=a, value=(OUT,b))

Output (key=b, value=(IN,a))

R1(key=x, V):

Collect Sout = set of all (OUT,*) value items from V

Collect Sin = set of all (IN,*) value items from V

if (|Sout| < 10 AND |Sin| >= 100M)

output (x, Sout)

M2((a,b):

Output (key=a, value=(IN,b))

R2((key=x, V)

Collect Sin = set of all (IN,*) value items from V

If (Sin >= 10M)

Output(x,_)

4.

M1(a,b):

Output(b,1)

Output(b,a)

R1(x,V):

A_id = {}

```
for x,V in (x,V):
        if x not in A_id: A_id[x] = [ ]
        If V is not None: A_id[x].append(V)

for u,f in A_id.items():
         total_f = sum(f);
        if total_f >= 100 million
        if len(f) >= 100:
                for x in f:
                        if x != u:
                                output(u, f)
```

5.

M1((a,b),(a,start_time,end_time):

output (b,a)

R1(x,V): –

All a in V, output (lexicographic_sorted_pair(x,a), |V|)

M2(a,b): Identity

R2(key=(a,b), value={|V1|, |V2|,…})

– if |value|==1 output nothing

– else if |value|==2

output (a,b)//outputs a and b who follow each other

M3((a,b),(a,start_time,end_time)):

output (a,(USER,b))

output (a, (TIME,start_time, end_time))

R3(x,V):

Collect SUSER = set of all (USER,*) value items from V

Collect STIME = set of all (TIME,*,*) value items from V

for u_items in SUSER:

for tf_items in STIME:

for tl_items in STIME:

if((x from u_item == x from tf_item) && (V from u_item == x from tl_item)

if(tf_item[1] == tl_itme [1] and tf_item[2] == tl_itme [2])

output(x,V)//outputs a and b who were simultaneously present at the same time.

7. The candidate could be true with the statement “You’re all so slow! You’re all doing push gossip. I do pull gossip, and even with fixed fanout, it converges in O(log(log(N)) time!” This can be related to the popular social media platform twitter where a tweet made by politician can be viewed by his followers who tend to retweet his tweet which can be retweeted by their followers, and this could continue forever. This method tends to multiply super-exponentially at the end because of the multiple retweets that are present, which could be retweeted by their respective followers. In the other case push gossip could be related to the form of a politician’s team sending a text to a group of people while the group of people sharing the message by text to their group of friends, and which continues forever. In this case the receiver should receive the message to gossip the message to his/her friends which only could move at O(log(N)) pace.

9.

a. T the biggest advantage of Tcleanup = 0 is that when a processes realizes that the heartbeat has not been increased for more than the time set for Tfail, it will immediately delete the server from membership list. Which helps the campaign's initiative to make failure detection faster. The biggest disadvantage of Tcleanup = 0 is that other processes that hasn't updated the processes failure might gossip to the processes that has deleted the processes which in turn would add the processes back to its membership list and processes failure could go unnoticed if this continues in a loop.

b. The biggest advantage of removing suspicion feature from SWIM-style failure detection is that a processes can be marked as failed as soon as time runs out, which makes failure detection faster and makes sure completeness is given first priority. The biggest disadvantage is that in the case there was a slowdown, multiple alive processes can be marked as failed and cause a delay and extra resource to make the processes rejoin.

c. The biggest advantage of removing round-robin pinging and random permutation, while randomly selecting each ping target is that the processes doesn't have to ping all of them in a specific order which reduces the use of network and system resources while reducing the complexity by randomly selecting the ping target. The biggest disadvantage while randomly selecting the ping target is that one target could be selected multiple times causing congestion on the target and some nodes not being pinged in a timely manner.

10a. According to option 1 each process sends heartbeats to its monitoring set members. To violate completeness, there shall be scenario where multiple process failing simultaneously considering that these processes are from different row, column, and aisle. In this case option 1 would not detect all failures, because each process monitors its own respective row, column, and aisle. Hence, proving option violates completeness when L >= M+K+R

10b. According to option 2 each process periodically pings all its monitoring set members and pings are responded to by acks. In order for option 2 to violate its completeness, the processes in the monitoring set of a failing process doesn't respond by acks. In that case, option 2 can violate completeness when simultaneous process fails and a common process to the failed process also fails, which would lead to multiple processes failure not being detected.