Anthony Poerio (adp59@pitt.edu)

CS1699: Assignment 2 - Written

Due: 2/4/2017

QUESTIONS

1) **Briefly describe the 4th Paradigm by Jim Gray, and why (or why not) you believe him.**

   Jim Gray's 4th Paradigm is "*Data-Intensive Science*".

   By using the term "Data-Intensive Science", Gray means to say:

   1. Our society is producing data an enormous rate, and that rate is exponentially growing over time.
   2. Because we are producing so much data, we need *new* computational tools to manage, visualize, and analyze these data effectively. The old methods are no longer sufficient.

   For Gray, this marks a shift from the previous paradigm—termed "Computational Science"—which focused on *simulating* complex phenomena, based on mathematical models.

   Now, we no longer need to simulate the phenomena we are interested in studying. Instead, we can gather data about phenomena directly. But now, we have a new problem; there is *too much* data!

   My thoughts:

   - I agree with Jim Gray. From personal experience, I can anecdotally verify that we are producing data at a dramatically increasing rate. I know I create more and more data than I ever though possible, and so do many people I know.
     - And beyond that: We have statistic on Slide 12 of the very first presentation in this class: 90% of today's data has been created in the last 2 years.
   - In fact, we are producing so much information these days, that we need to construct data centers 11.5 times the size of a football field just to store all of it.

- As a scientist, engineer, programmer, or data analyst of any kind—you will quickly encounter problems when trying to produce meaning from data at this scale.
- Sure, you could subsample the data set, but then you are not using all of the data resources available; and in effect, that's exactly what Jim Gray is saying: the problem is HOW to utilize these massive datasets in full. Which is the whole idea of the 4th paradigm.

2) What are the 5 V's related to Big Data?
Please describe briefly the meaning of each V.

1. **Volume** → Volume relates to the "scale" of data that we are created. This is a pure numerical amount, measured in bytes (or mega, giga, tera, or petabytes). A quantity of information.
2. **Velocity** → Velocity is the "speed' at which data is transferred. For example, 50,000 GB/second is the estimated rate of global Internet traffic in 2018. A rate of information transfer.
3. **Veracity** → Veracity is the "certainty" of data. That is, how *accurate* is the data we are analyzing? The simplest way to view this is as a qualitative measure. How 'accurate' or 'good' is our data. Veracity tells us if are we making decisions based on flawed information.
4. **Variety** → Variety is the "diversity" of data. This is a measure of how many different data-types we must work with. Videos (.mpg's, .avis), audio-files (.wav, .flac, .mp3), text files, database files, and the like. A major issue with data analysis is the ability to parse information from data types which encode very different information types, and in very different structures… if there is structure at all.
5. **Value** → The ability to use our data to produce actionable 'knowledge': or insights with business, personal, or other types of value. Without value, there is no reason to store data at all. Implicit in storing information is the assumption that it is worthwhile to bear the economic costs of large scale data storage over time. And if our data does not deliver value, then our data is wasting space, extracting value from us instead of delivering it.

3) Cloud Computing is dominating everywhere, explain.

a) The motives for enterprises to move their resource to the cloud

Enterprises are incentivized to move their resources to the cloud in order to reduce costs related to storing and managing large volumes of data on-site, and increase revenue through the ability to extract valuable knowledge from their large datasets.

Moreover, they are able to gain efficiency by requiring fewer workers to administer their servers, fewer physical machines to run the same amount of applications, while gaining even more computation resources on-demand from a cloud service provider.

What's more: businesses can increase the agility and rapidity at which they can launch a new product, because they don't need to order, setup and secure new unique hardware for each software product they are developing. Cloud-based services abstract those problems away.

b) The motives for customers to use the cloud

Customers are incentivized to use the cloud because they provide access to large-scale computing resources at a fraction of the price.

For example, if I want to run data-analysis algorithms on 100 TB dataset, it would be extremely difficult to store that information, and extremely time consuming to run algorithm on my personal computer, even if I could.

But with cloud computing services, like AWS, I can store that data on their servers, provision computers to do the computation (essentially renting them for a short period of time), and then delete my information when I'm done. Doing this would cast a tiny fraction of the startup costs for buying the hardware I'd need to do this myself.

Customers gain access to a very large shared resource pool, on an as-needed basis, enabling them to perform computations and achieve results never before possible.

4) **A company needs 500 servers with 100 petabytes data storage. In each situation below, suggest an option(s) for the company to deploy, among: Data Center, Public, Private, or Hybrid cloud, along with a reason why.**

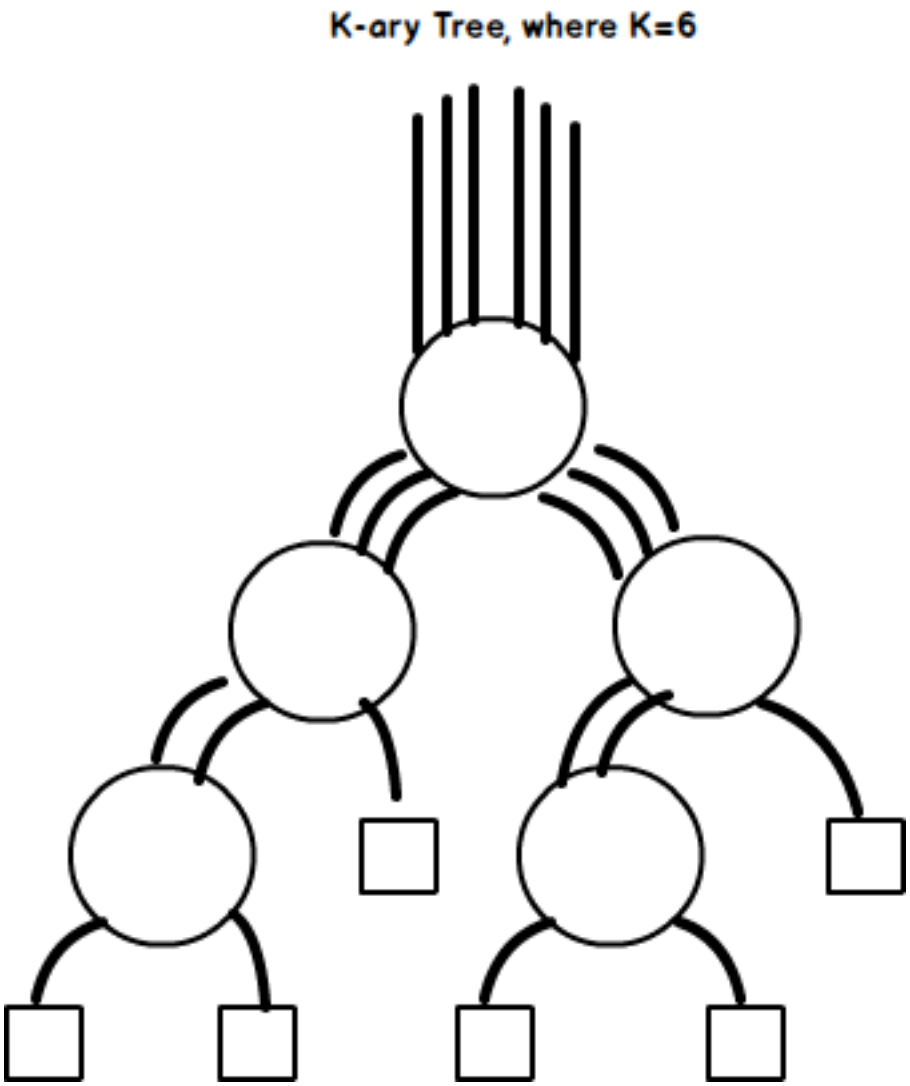
a) The company needs full control the data and equipment. Security is a big concern.
- **i)** Suggestion: **Data Center**
- **ii)** Reason: Because the company needs full control of the physical equipment, and because there are 500 servers → I think it would be reasonable for this company to build a physical data center to solve their problems. Moreover, because security is a primary concern, they will be able to store all their data on a local network, and ensure that no one outside the company is able to access it.

b) The company's needs will change rapidly since the market is so volatile.
- **i)** Suggestion: **Public Cloud**
- **ii)** Reason: Because the company has rapidly changing needs, it would be unwise to buy 500 machines up-front. Later, they might find that the machines which previously fulfilled their needs are no longer sufficient. Rather than buying 500 new machines whenever they change direction, I think it makes most sense to use a public cloud, because of the flexibility it provides.

c) The company's IT department wants to control these IT resources and they expect to shuffle the resources among different departments quite often.
- **i)** Suggestion: **Private Cloud**
- **ii)** Reason: A private cloud delivers scalability and self service like a public cloud, but it will also provide their IT department with direct control over their environments, while ensuring that their hardware is not shared with other organizations. This gives them the most flexibility for their needs; they'll be able to shuffle resources amongst their departments easily, and to know exactly what is going on with their machines.

d) The company needs to put some servers in a secure environment and some in the public domain
- **i)** Suggestion: **Hybrid Cloud**
- **ii)** Reason: A hybrid approach will allow this company to meet both of their goals. Some of their servers can be hosted on the public cloud, and be publicly accessible for their customers; while their secure servers can still be housed in a private cloud, ensuring that no one outside their company can access them.

5) **Draw the K-ary fat tree architecture where k=6.**



6) **Draw PUE and DCie for the DC which has power usage detailed below.**

a) 60,000 kW to power all the servers
b) 10,000 kW to power all networking gears
c) 5,000 kW for all light
d) 5,000 kW for the air conditioning in the administrator room
e) 20,000 kW for the cooling system

PUE Calculation:

PUE = Total Facility Energy / IT Equipment Energy

- TOTAL Facility Energy
  - 60,000 kW for servers
  - 10,000 kW for networking gears
  - 5,000 kW for light
  - 5,000 kW for server room air conditioning
  - 20,000 kW for the cooling system
    - SUM = 100,000 kW
- IT Equipment Energy
  - 60,000 kW for servers
  - 10,000 kW for networking gears
    - SUM = 70,000

Therefore, calculation is: **[100,000 kW / 70,000 kW]**

PUE = 1.429

DCie Calculation:

DCie = **1/PUE** = IT Equipment / Total Facility Energy

- DCie Calculation:
  - DCie = [70,000 kw / 100,000 kW]
  - **DCie = .70**

7) What are ideal numbers for CUE and WUE? Is it possible to achieve that?

- CUE:
  - CUE is a measure of carbon emissions. Specifically, it calculates the kilograms of CO2 emitted from the DataCenter per kilowatt-hour.
  - **Ideally:** We do not want to emit ANY CO2 from our facility(**CUE=0**), but realistically this is NOT possible under normal circumstances.
    - If we are pulling electricity from the energy grid, then that electricity is likely produced such a way that CO2 emission is a byproduct.
    - If we do NOT pull ANY electricity from the public energy grid, but instead generate all power locally on the facility, in some way (solar panels, wind, other methods) → then we would need to calculate CUE on a case-by-case basis, taking our specific situation fully into account.

- WUE
  - WUE is a measure of water usage in a datacenter facility.
  - Specifically, it measures: **[Annual Site Water Usage / IT Equipment Energy]**
  - Thus, ideally → we want to MINIMIZE this statistic, using as little water as possible, both to reduce costs and conserve water.
  - Therefore, the **ideal number** for this statistic is again 0, **(WUE=0)**
    - However, because we have large server rooms that get very hot, it is NOT possible to achieve ZERO water usage. Water is necessary for server cooling. However, we can make designs that seek to minimize water usage, even though it may never totally 0.

8) From the four Server Virtualization techniques learned in class, answer below which technique is best for each situation.

a) Need a lightweight virtualization that can populate several virtualized machines with the same OS as the host's.
i) OS Level Virtualization

b) Performance is the main concern and need to install several VMs with OSs as is.
i) OS Assisted Virtualization

c) Need to install several VMs with new OSs that are not yet supported by VMM
i) Full Virtualization using Binary Translation

d) Need an efficient virtualization technique that utilizes the new CPU architecture.
i) Hardware Assisted

9) Describe how a Shadow Page Table works.

A shadow page table provides virtual-to-physical memory address mappings, using the hardware's Translation Lookaside Buffer (TLB).

More succinctly: The "Shadow Page Table" maps Host (virtual) Memory Addresses → Physical Memory addresses on the host server.

The shadow page table **works by**:

- First, the guest OS must maintain its own virtual memory page table
  - (It does this, of course, in memory frames allocated *from* the host server, and thus the VMM can access this information).
- Then, the Virtual Machine Manager (VMM) can map this virtual page table directly to the host's physical memory frame.
  - This allows us to have a complete mapping from VM → Host
    - The Shadow Page table maintains his mapping from Guest (Virtual Address) → Host (Physical Address)

10) Why do we need Network Virtualization? Why is what VLAN provides not good enough?

We need Network Virtualization because VLAN is an OSI Layer 2 (L2), and thus it has some limitations.

In L2, ***every*** *broadcast frame* in the domain MUST be *processed by* ***every*** *host* on the domain. This means that there is too much traffic to effectively handle over VLAN in a large network. This is a **bottleneck**.

Moreover, also because of this architecture, if one link fails → everything fails. And that's un-acceptable as we scale to more and more machines. (Probability of failure increases.)

For these reasons, generally, we can only have up to 1,000 hosts in a virtual network; but we may need many more.

Management of large networks on VLAN is also very complex, and each device needs to be configured separately.

Because of this, we need network virtualization to effectively manage networking at a large scale.