Week Three Lecture Videos

Indiana University Bloomington: Big Data Applications

Naimesh Chaudhari: naichaud@iu.edu

Assignment 2 – September 14, 2019

1. Abstract

In this paper, I provide a brief introduction about the information presented in the Week 3 lecture videos. I also provide additional information on a subtopic that I am interested in from the videos and finally I provide ideas that could improve some of the existing work that has been done.

2. Video Summary

4.1 Defining Clouds

In todays world the use of public cloud has increased rapidly with many software stacks. To support this growth, we will need lots of computers to do parallel computing. A cloud is defined as “a bunch of computers in an efficient data center with excellent internet connection.” The NIST design allows for on demand service with resource pooling and flexible resource allocation. Cloud allows for 6:1 server consolidation which can save a significant amount of cost for a business. The leading providers in cloud services are AWS and Microsoft Azure. There are four different types of services available IaaS, PaaS, SaaS, and NaaS. Edge computing and serverless are transformational technologies that are on the rise. Due to this the job market is also switching from enterprise it to cloud based it.

4.2 Virtualization

To support multiple of the same set of CPU, virtualization was developed. This essentially allows us to create a virtual machine that looks exactly like a normal desktop but uses the allocated amount of resources to the main set of resources. This is done using a hypervisor. There are two types of virtualization technologies. Type1 and Type 2. There is also the operation system level virtualization. Docker is a company that uses OS virtualization. OS virtualization is not as safe as hardware virtualization such as KVM and XEN

4.3 Cloud Infrastructure

There are generally three types of vendors for cloud infrastructure.: some will deliver their own cloud service, most will sell server infrastructure to cloud service providers, and finally some will emulate the benefits of cloud computing to augment and extend their on-premise visibility. By 2021 more than half of global enterprises with current cloud will adopt an “all in” cloud strategy. The bigger vendors like Amazon, Google, IBM, and Azure have cloud infrastructure all around the globe. Cloud network usage has grown by 25% CAGR since 2016. Hyperscale data centers will nearly double from 2016-2021. Serverless infrastructure, Neuromorphic computing, In-memory computing, and next generation memory technologies have taken a rise in 2018. In the future containers and serverless will dominate innovation.

4.4 Cloud Software

The major player in this field is the Apache Big Data Stack with close to 21 layers. Google has published a timeline on how to build on Apache stack. MapReduce is the major workhorse in the big data technologies. Some of the major could platforms are Apache Hadoop Google MapReduce, Microsoft Dryad, Bigtable, and Chubby. There are many components to a big data runtime.

4.5 Cloud Applications

Cloud application like Cyberinfrastructure are used widely. Cyberinfrastructure allow for distributed research and learning. They have two aspects parallel and distributed. E-science is also another great example. E-science is about developing tools and technologies that allows scientist to do faster and better research. There are many cloud applications in different sciences like physics and astronomy. Deep learning due to is uses of extreme computer power is a major application of the cloud. Lilly is a drug discovery corporation that is utilizing the cloud. By 2020 24-75 billion devices will be on the cloud. Due the this we could have a “god” infrastructure where all devices are communication with each other.

4.6 Parallel Computing

Parallel computing decomposes data and models between computers that work together on the same problem. There are thousand of servers with millions of cores utilized on the cloud. Simulation of cosmological cluster with close to 10million stars has been done by parallel computing. Parallel processing principals are the same principles societies have been following for many years, work as a team and distribute the work between all the people. The Hadrian wall is a great example of problems that can be solved via parallel processing.

4.7 Real Parallel Computing

SPMD is a parallel computing technology. SPMD is much more useful then SIMD. The classic features of parallel computing are HPC, Clouds, and the combination of HPS + Clouds. Most different programming models like Spark, Flink, Storm, Naiad, & MPI/OpenMP have the same high-level approach. The approach first splits the data and model with many nodes and then all nodes will do processing in parallel. In parallelism synchronization is a major constraint for specific problems.

4.8 Storage Cloud Data

Due to restriction in the traditional database models and the Web 2.0 applications which have huge data sizes and non-SQL queries the NOSQL approach stated taking a rise. Cloud storages are set up differently then tradition windows on Unix systems. Cloud storage generally support disks in parallel. Tradition databases are still relevant but are generally for higher level management systems. Newer NOSQL approaches like Hbase, Dynamo, Cassandra, MongoDB, CouchDB, and Riak are being utilized more often. These technologies generally have no fixed Schema and are linked to a Hadoop system. Traditional data lakes should be used for new analytical insights but should not be the only component in the corporation analytics infrastructure.

4.9 HPC & Clouds

An HPC infrastructures are the heart of most leadership organization. Example of industries are: manufacturing, natural resources, life sciences, financial services, and governments. HPC infrastructures are designed to improve productivity and reduce the physical space for data centers. These centers generally use parallel processing and compute acceleration like Nvidia GPU’s. There are difference technologies built into this like large scale super computers, high throughput systems, grids, and SaaS.

4.10 Comparison Of Data Analytics with Simulation

There are different structures of applications. Application such as search and recommender engines have different structures. We should generally strive to discuss data and models separately. There are use cases for streaming data and simulations. Simulations like chemical connections and the networks of the Facebook graph are good examples. Common language used in the field are Java, Python, R, C++, and Matlab.

4.11 The Future & Other Issues

The computing power per thousand dollar, & capacity of storage has been growing over the years. The prices of storage has had an inverse effect where it is decreasing over the years. The future of cloud has a few major technologies. New methods of software delivery, container/ microservices, elastic analytical databases, and edge computing. The hype for HPC, blockchain, and serverless PaaS is on the rise. In the Gartner priority matrix, most cloud technologies are transformation and will have mainstream adoption in the next 2-5 years. Serverless computing and FaaS are on the rise.

3. Interested Subtopic

The technology that I am extremely interested on is Docker. Upon further research I learned that it is the industry leading container platform for high velocity innovation. It allows companies to build share and run application from legacy to what comes next with the security to run them anywhere. It supports a wide variety of application from legacy to micro service applications. All of this is done on a common operation platform. Docker allows for applications to be run on both Linux and Windows which enables customers to modernize their application. Docker is the only container platform that offer both Kubernetes and Swarm side by side. The docker platform is also the first of its kind that extends to the developer's desktop. This allows developers to test their application using the same tools they use today. A new service they have recently launched is Docker Apps, it allows user to bundle all the services and application spec as a single unit, which is stored in a docker registry, which can then be deployed to either Kubernetes or Swarm. Since the apps are registered in a docker registry they can be scanned for vulnerability and checker for authenticity. It extends to all user of the platform from developers to operations. Docker is very secure and allows for image share cross countries. They have a unique functionality where images are signed by its developers, docker can authenticate this to insure all income images are safe. It gives you a vulnerability scanner that scan for vulnerability and lets you know all the content in the image. Docker also adds enhanced lifecycle management. It allows automatic multi infrastructure deployment and upgrades, non-disruptive upgrades, and online backups. The docker stack proves a platform, tools, and methodology to modernize their environment. The major advantages of docker are each app has its own environment, it is easy to start a sandbox project, and finally it is each to get into someone else's project. All these advantages are granted without a true virtual machine. Containers get started faster, they use less disk space, and less memory.

One of the technologies docker allows for use is Kubernetes. It is a google developed platform and is the most popular container orchestration technology. Kubernetes hosts applications in containers in an automated fashion so that we can easily deploy applications. The worker nodes hold the container while the master node manages the container nodes.

4. Ideas that can improves existing work

Although docker is a fantastic service for specific use cases it also has its drawbacks that we need to discuss. I will present some of those drawbacks here, my suggestion is for the team at docker to overcome these drawbacks so it can widely be adopted.

One of the major concerns with docker is that it shar the OS kernel within apps. This inherently makes it that Docker apps will not run at bare metal speeds. Due to the shared resource of multiple apps, Docker apps will generally run slower than apps launched on bare metal hardware. This issue is at the core of Docker's structure, I am not entirely sure how they can develop a solution for this, but if they could improve on this it would be a fantastic win. Also, it is fair to note that due to multiple apps running on the same kernel, if there is a bug within the kernel the entire suite of apps within the container will be affected. In the VM version, only the specific application will be affected.

The second thing I would like to mention is that most container systems have a difficult time working with other technologies. For example, the OpenShift container system only works with Kubernetes. This is primarily because of vendor-vendor competition. As a developer/end customer, I would like to see and initiative from the companies to allow for a mix of technologies. This would give much more flexibility for an end customer to customize their product requirements.

Docker container do not have any persistent data storage. There are additional services that can allow for this, but it is a challenge to set this up. This sole reason is why many end users would choose VM's over docker. VM's although clunky, have everything a developer needs to set up an application. Docker at times need additional addons to accomplish this. If a developer needs to run a graphical application, Docker is not a good solution for this. Like data storage we can do this with addon services, but it is not as intuitive as doing it on VM's. I would like to see the team at Docker work on and address this issue, again I believe it would add a lot more customer to their business.

Finally, I would like to mention that to run a Docker service, a user needs root right. This can be very problematic for companies is sensitive sectors like banking, and health. If the service can be adopted so that users do not need root right, I think it would be on more companies list to review.

5. References

1 Docker *Docker Enterprise*. [online] *Available at: https://www.docker.com/products/docker-enterprise*

2. Philip Hauer *Discussing Docker: Pros and Cons*. [online] Available at: *https://phauer.com/2015/discussing-docker-pros-and-cons/*

3. DataFlair *Team Advantages and Disadvantages of Docker.* [online] *Available at: https://data-flair.training/blogs/advantages-and-disadvantages-of-docker/*