Cloud Computing: Ensuring Energy Efficiency by Considering The Performance of Communication Links

Michael S. Bull
Advisor: Dr. Suk Seo
Middle Tennessee State University – Computer Science Department

Introduction

- Cloud computing is becoming more crucial to businesses around the globe as a way to reduce capital expenses.
- As cloud services become more prevalent, and customers push for new services and require faster download speeds of content such as streaming videos or music, and less latency in time-sensitive services such as cell phone and other real-time voice and video services, cloud providers have to adapt.
- Cloud providers are constantly looking for improved methods to save energy and maintain reliability, both of which add to the cloud provider's bottom line.
- Since the advent of cloud services and particularly in the past few years, much research has gone into how cloud providers can be more energy-efficient, all the while providing the expected *service level agreements* (SLA) of their customers.
- In this paper, I propose a method that will increase energy efficiency by avoiding potential faults such as saturated communications links or over-utilized servers and associated Virtual Machines (VMs) in terms of CPU frequency and memory allocation.

Aim & Goals

- Augment existing workflow scheduling techniques with Communication link data
  - Schedule workflows to VM to maximize resources
  - Monitor CPUs and communication links
  - Avoid oversubscribed resources and links

Abstract

Today's cloud environment is a complex web of physical hosts, virtual machines, and communication links, each capable of injecting a certain amount of latency to a job's execution time. Certain applications and workflows are required to complete with the least delay possible; thus, it is paramount that client requests are routed to servers best able to deliver reliability and performance in the shortest time possible. The geographical distance separating the client from the requested resource, the overall execution time of the workflow, and transmission time all impact performance and reliability and risk not meeting service level agreements or customer expectations if not optimized. Extensive studies have focused on server selection and virtual machine placement along with other server CPU optimization strategies to ensure reliability, performance, and reduced energy cost.

While this project's aim and goals were not accomplished, this paper intended to show that latency, congestion, and errors attributed to communication links have a role not only in the reliability and performance of cloud computing, but in its energy efficiency as well. The approach proposed by this research incorporated communication links' statistics for a possible host and virtual machine selection into account during the workflow scheduling. Still, preliminary research results suggest that choosing communication links which lower the total makespan by 20ms could result in a 7% decrease in total time from request to response. That decrease in processing time decreases overall energy consumption.

Background

- Many research methods used to reduce energy cost focus making the servers more efficient in its CPU frequency scaling known as Dynamic Voltage and Frequency Scaling, Alrammah et al.
- Deiab et al. find idle servers in a data center account for over seventy percent of the data center's computational cost.
- Qiu et al. state to achieve effective use of a cloud's resources, that the performance, power consumption, and reliability must all be considered.
- Sharma et al. maximize VMs to physical host ratios to reduce energy cost.
- Sood and Kaur show that proper fault avoidance techniques can play a major role in achieving an energy efficient cloud computing model.

Methods

- Task Scheduling Base Cases.
  - Round-Robin
    - Round Robin is the preemptive process scheduling algorithm.
    - Each process is provided a fix time to execute, it is called a quantum.
    - Context switching is used to save states of preempted processes.
  - Shortest Job First
    - Moves task with the lowest execution time to the front of the queue.
    - This is a non-preemptive, pre-emptive scheduling algorithm.
    - Can cause process starvation.
- Communication Link data Augmented Base Cases.
  - Each base case is augmented with bandwidth, latency, or delay statistics of the communication link for the proposed server and VM.

Results

- Round Robin
  - The Augmented Algorithm did not show significant improvement.
- Shortest Job First
  - The Augmented Algorithm did not show significant improvement.
- Future Work
  - Revise VM allocation and Cloudlet components to be more stressful on the network.
  - Design Datacenters more geographically dispersed.
  - Incorporate cloudlets to the datacenter that consume VMs, servers, and bandwidth that are not part of the test data.

References

H. Alrammah, Y. Gu, C. Wu, and S. Ju, "Scheduling for energyefficiency and throughput maximization in a faulty cloud environment,"IEEE 23rd International Conference on Parallel and Distributed Systems,pp. 561–569, Dec. 2017. [Online]. Available: https://ieeexplore-ieee-org.ezproxy.mtsu.edu/document/8368408

M. Deiab, D. El-Menshawy, S. El-Abd, A. Mostafa, and M. S. A.El-Seoud, "Energy efficiency in cloud computing,"International Journalof Machine Learning and Computing, vol. 9, no. 1, pp. 98–102, Feb.2019. [Online]. Available: https://doi.org/10.18178/ijmlc.2019.9.1.771

Y. Sharma,B. Javadi,W. Si,and D. Sun,"Reliable andenergy efficient resource provisioning and allocation in cloud com-puting,"Proceedings of The 10th International Conference on Utility and Cloud Computing, p. 57–66, 2017. [Online]. Available: https://doi.org/10.1145/3147213.3147218

A. Sood and K. Kaur, "A coordinated approach for energy-aware fault tol-erance in cloud computing," vol. 8, no. 5. IJARCS, May 2017, pp. 2499–2504. [Online]. Available: https://doi.org/10.26483/ijarcs.v8i5.3890

X. Qiu, Y. Dai, Y. Xiang, and L. Xing, "A hierarchical cor-relation model for evaluating reliability, performance, and powerconsumption of a cloud service,"IEEE TRANSACTIONS ON SYS-TEMS, MAN, AND CYBERNETICS: SYSTEMS, vol. 46, no. 3,pp. 401–412, Mar. 2016. [Online]. Available: https://ieeexplore-ieee-org.ezproxy.mtsu.edu/document/7169613