Enhancing Convergence of Live VM Migration with Dynamic Workloads

NAME: D.N.NETHMINI EPA | 20000499

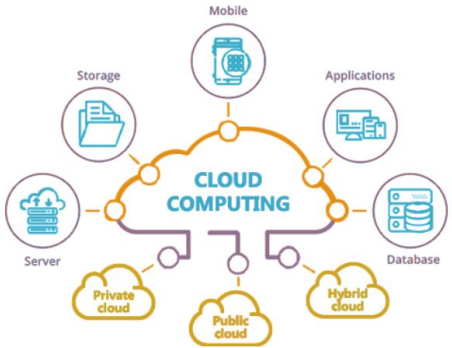
SUPERVISOR: DR. DINUNI K. FERNANDO

Agenda

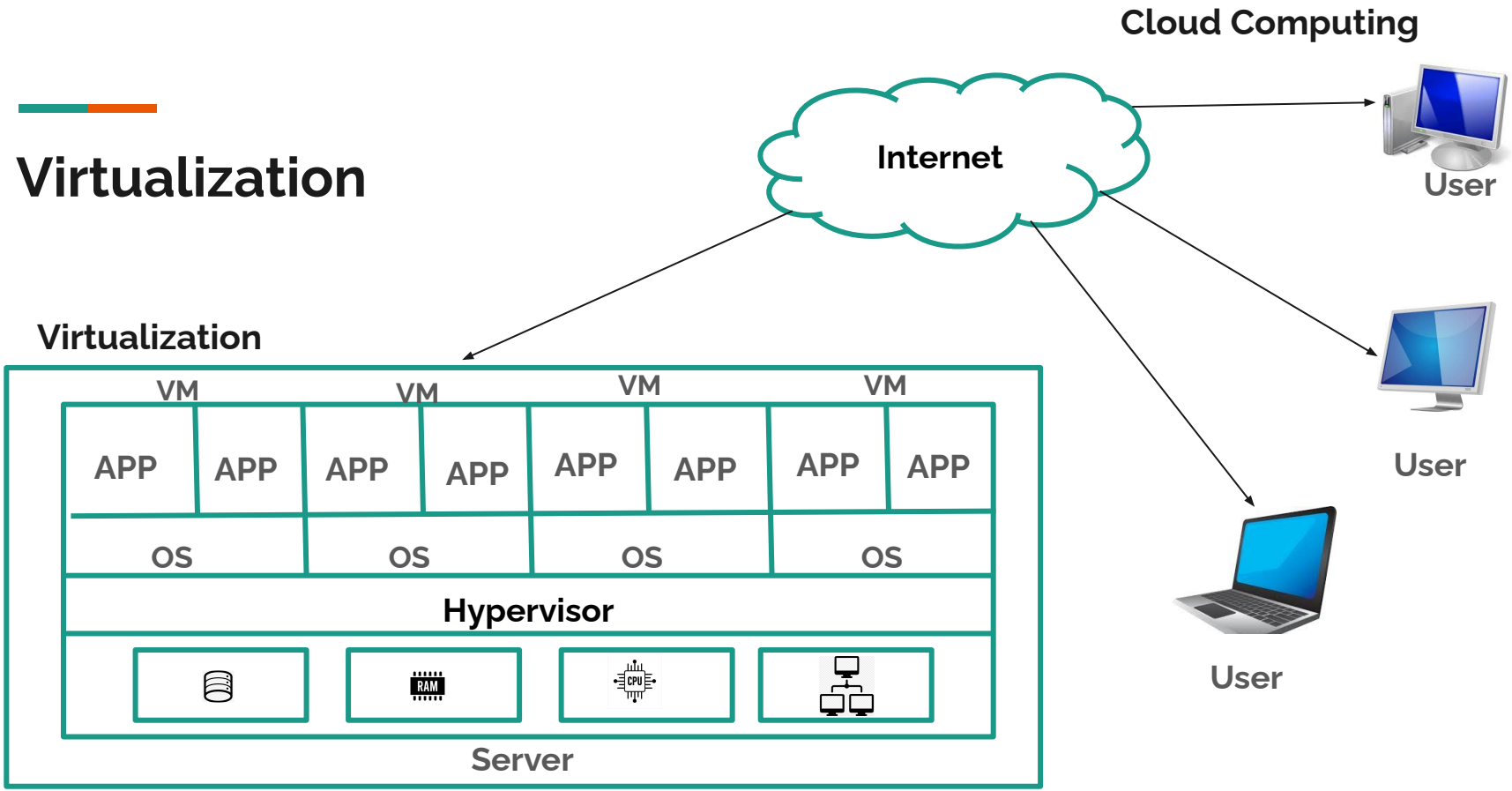
1. Introduction
2. Motivation
3. Related Work
4. Research Gap
5. Research Questions
6. Objectives
7. Scope
8. Research Approach
9. Progress

Have you ever wondered how companies can scale their IT resources so efficiently to meet growing demands

The answer lies in cloud computing.



Virtualization



Failure Rates in Google Data Centers

"In each cluster's first year, it's typical that 1,000 machine failures will occur; one power distribution unit will fail, bringing down 500 to 1,000 machines; 20 racks will fail, each time causing 40 to 80 machines to vanish from the network; there is about a 50 percent chance that the cluster will overheat, taking down most of the servers in less than 5 minutes"

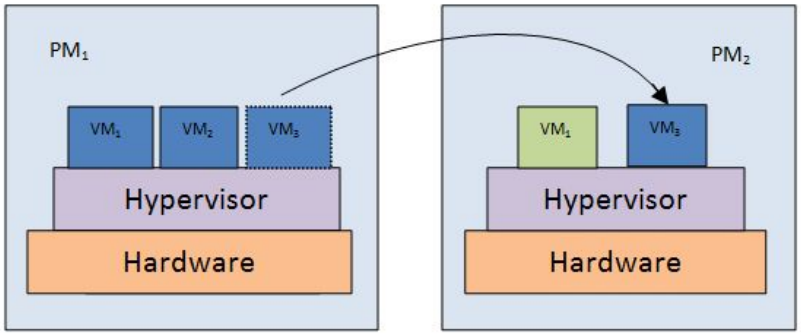
(R. Miller, "Failure rates in google data centers," https://www.datacenterknowledge.com/archives/2008/05/30/failure-rates-in-google-data-centers, 2008)

What happens if a server fails?

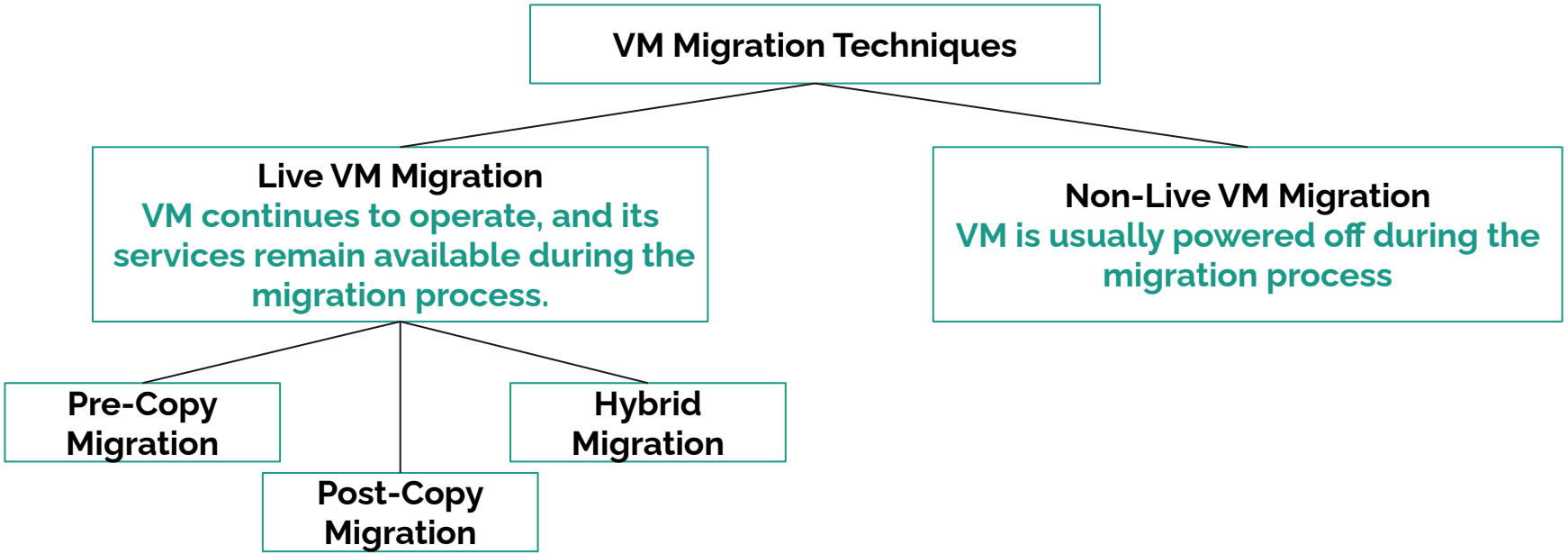
- Failure of a server can occurred at any time.
  - Failures due to hardware malfunctions
  - Unexpected power loss impacting the host server
  - Bugs or crashes in the hypervisor software(KVM)

VM Migration

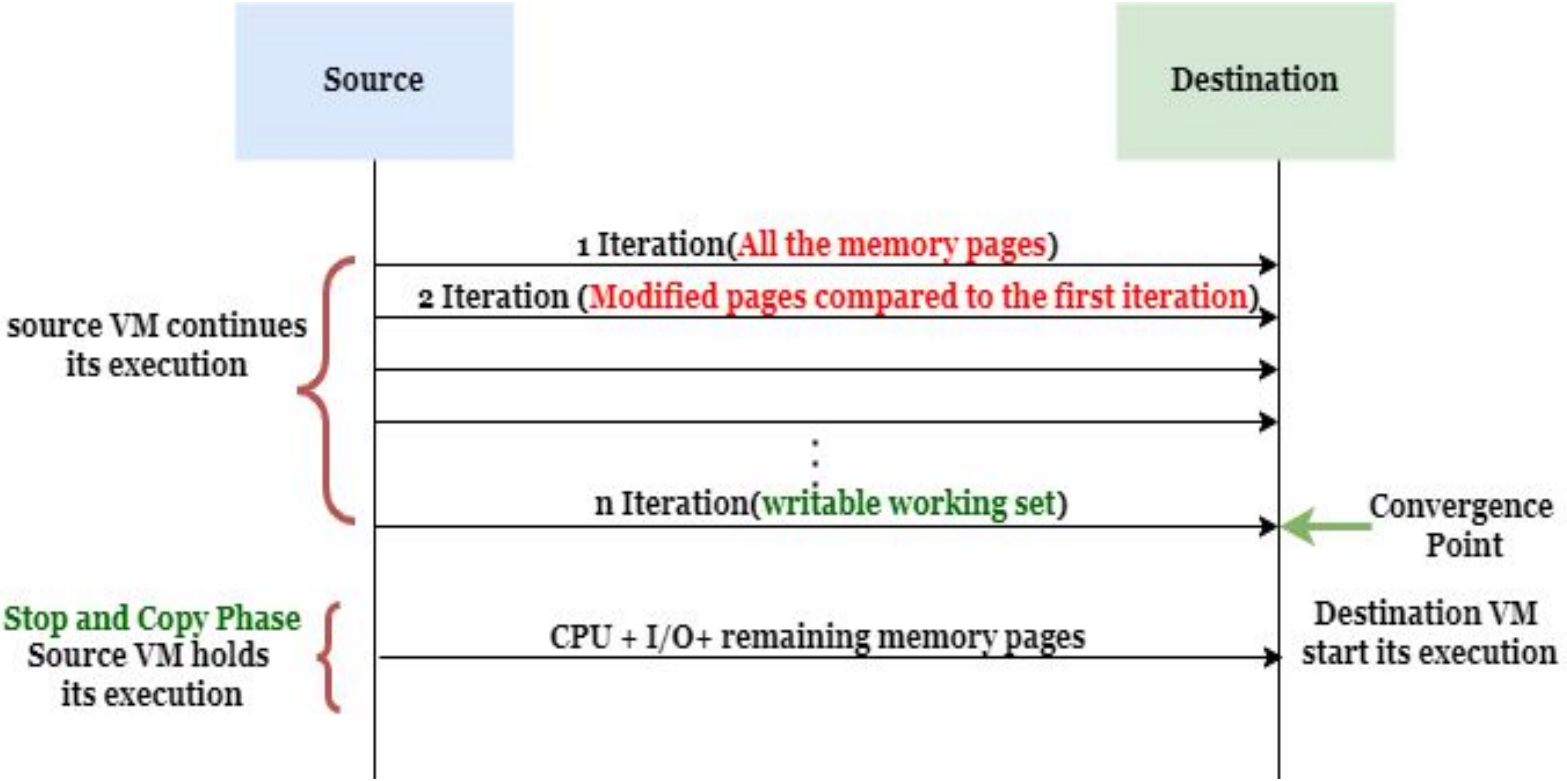
- **VM migration** is essential in virtualized and cloud environments. It allows for the transfer of virtual machines from one physical host to another with minimal downtime.
- Used for **fault tolerance,host maintenance, load balancing and consolidation** as well



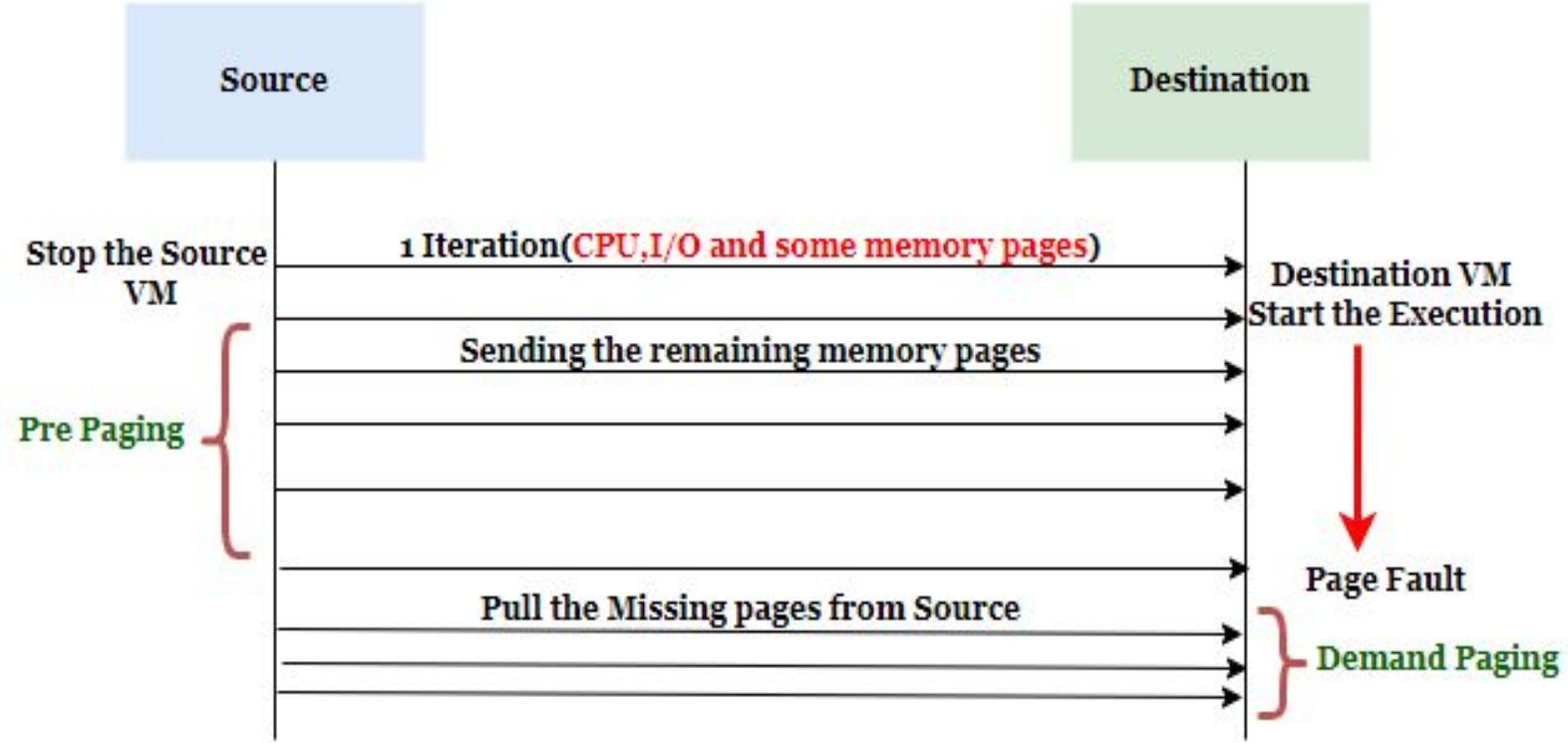
VM Migration Techniques



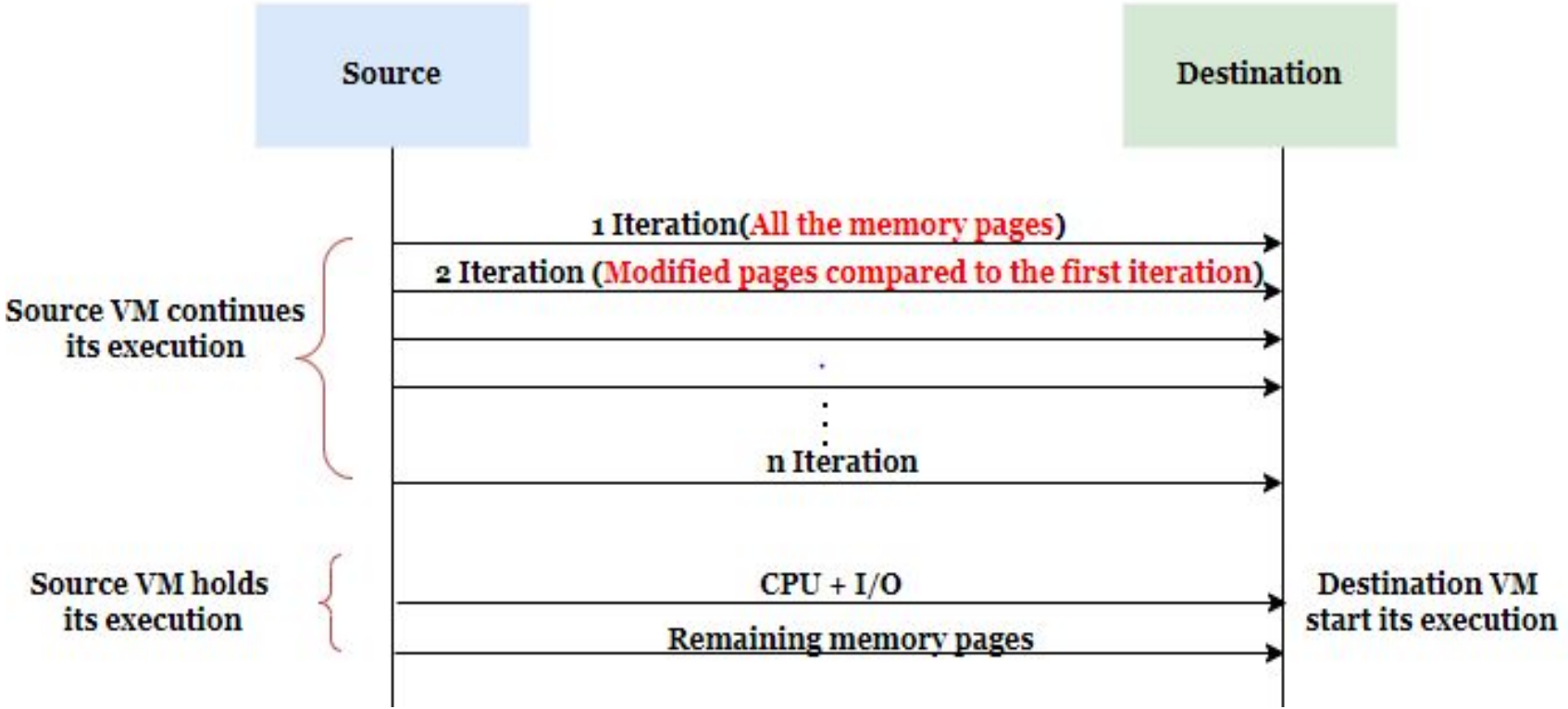
Pre-Copy Migration



Post-Copy Migration



Hybrid Migration

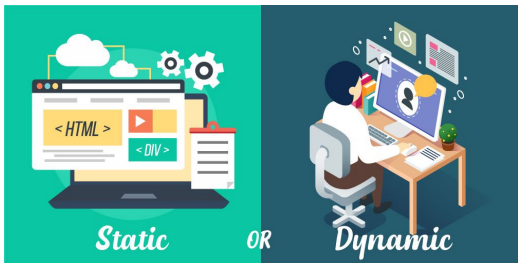


Performance Metrics

Total Migration Time	→ Duration between the start and end of migration
Downtime	→ The amount of time the VM is unavailable during the migration process.
Application Performance Degradation	→ The effect of the migration on the performance of migrating VM

What is a workload?

- The amount of time and computing resources a system or network takes to complete a task or generate a particular output.
- There are two(2) main types of workloads
  - **Static Workloads**
    - Use a relatively constant amount of computing resources over long periods of time and on a consistent schedule
  - **Dynamic Workloads**
    - Also called temporary workloads, adjust and configure computing resources based on computing demand.



Decide the convergence point

- As we discussed in pre copy migration, after some pre copy rounds there is a point where the dirty page count become constant after that point. This point is known as the convergence point.

Static Workload

- Identification of the writable working set

Dynamic Workload

- Unpredictability and variability of future workload makes it difficult to identify the convergence point.

Motivation

- Dynamic workloads consists of different types of applications that have their own and different kinds of resource requirements which is known as **heterogeneity**.
- Also there is the problem of **over-provisioning and under-provisioning**.
- Efficiently identifying the convergence helps in optimal utilization of network and computer resources. It helps to,
  - **Minimizes the migration time**
  - **Reduces resource contention**
  - **Enhances overall system performance**

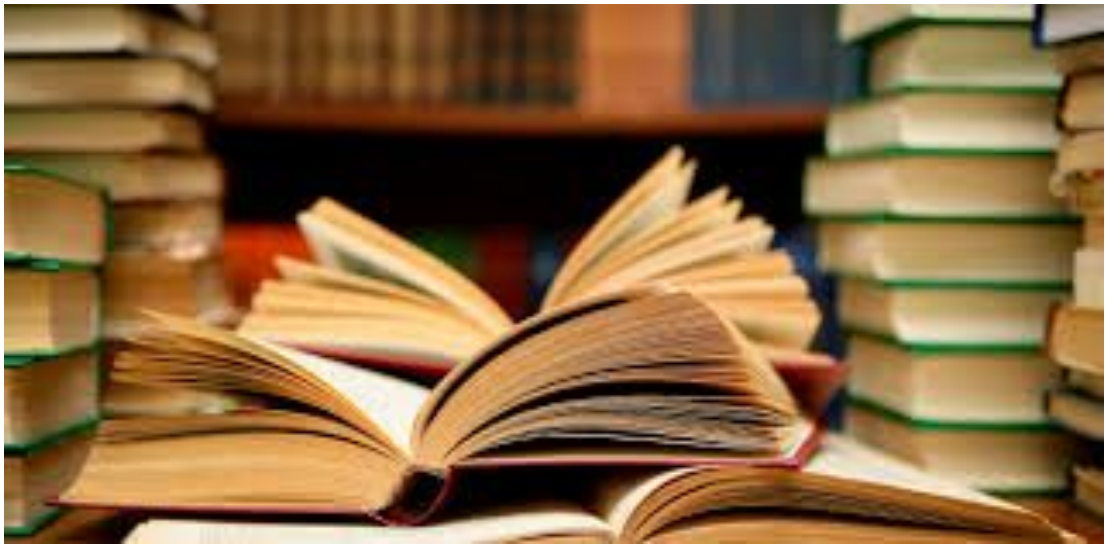
How do we predict the future workload?

- Predicting the future behavior of a workload has been addressed through two(2) main types of techniques.
  - Mathematical/Machine Learning Techniques
  - Statistical Techniques
- To predict the future behavior these techniques require capturing and collecting records of the current behavior of the workload to predict future behaviors.
- For that process, a well-known technique called **profiling** can be used.

Profiling

- Profiling is a concept used to capture the behavior of workloads based on their characteristics.
  - CPU ,Disk, I/O Usage
  - Memory Usage
  - Network Usage
- Some of the popular profiling tools are,
  - mpstat
  - Perf
  - free -h
  - Vmstat
- Also profiling can directly monitor specific metrics by using simple analytical techniques.

Related Work



Existing Use Cases of Profiling

Task Duplication and mapping algorithm **(Wu, Q. & Wolf, T. '2008 International Conference on High Performance Switching and Routing')**

Gather profiling information such as task service times and edge utilization to understand the workload dynamically.

- **To balance the workload, computationally demanding tasks are duplicated across multiple processor cores.**
- **Task mapping algorithm is used to place the interdependent tasks close to each other.**

consolidation planning module and a migration planning module**(Ye, K.IEEE Transactions on Parallel and Distributed Systems 26(3), 878–890)**

- **Consolidation planning module collects profiling information for various types of workloads such as CPU-intensive, memory-intensive, and network-intensive**
- **Using the profiling information optimized the resource allocation**

- Wu, Q. & Wolf, T. (2008), Dynamic workload profiling and task allocation in packet processing systems, in '2008 International Conference on High Performance Switching and Routing', IEEE, pp. 123–130.
- Ye, K., Wu, Z., Wang, C., Zhou, B. B., Si, W., Jiang, X. & Zomaya, A. Y. (2014),'Profiling-based workload consolidation and migration in virtualized data centers',IEEE Transactions on Parallel and Distributed Systems 26(3), 878–890.

Future Workload Prediction using Machine Learning

‘Optidjs+:Enhanced Dynamic Johnson Sequencing Algorithm (Banerjee, P., Roy, S., Modibbo, U. M., Pandey, S. K., Chaudhary, P., Sinha, A. & Singh, N. K. (2023)

- Supervised Learning to predict future resource demands
- Real Time Monitoring to detect changes in workloads
- Automatically adjust the resource allocation according to the workload changes.

Banerjee, P., Roy, S., Modibbo, U. M., Pandey, S. K., Chaudhary, P., Sinha, A. &Singh, N. K. (2023), ‘Optidjs+: A next-generation enhanced dynamic johnson sequencing algorithm for efficient resource scheduling in distributed overloading within cloud computing environment’, Electronics 12(19), 4123.

Future Workload Prediction using Machine Learning

‘Optidjs+:Enhanced Dynamic Johnson Sequencing Algorithm (Banerjee, P., Roy, S., Modibbo, U. M., Pandey, S. K., Chaudhary, P., Sinha, A. & Singh, N. K. (2023)

- Collects historical data and use supervised learning for predict future resource needs based on collected data.
- Allocate resources dynamically using the predictions from the supervised learning model.
- Apply reinforcement learning to optimize task scheduling strategies like SJF and LJF.
- Continuously monitor the system to detect changes in workload.
- Adjust resource allocation in real-time based on current workload and historical data.

Banerjee, P., Roy, S., Modibbo, U. M., Pandey, S. K., Chaudhary, P., Sinha, A. &Singh, N. K. (2023), ‘Optidjs+: A next-generation enhanced dynamic johnson sequencing algorithm for efficient resource scheduling in distributed overloading within cloud computing environment’, Electronics 12(19), 4123.

Future Workload Prediction using Statistical Methods

'Workload prediction using arima model and its impact on cloud applications' qos',Calheiros, R. N., Masoumi, E., Ranjan, R. & Buyya, R. (2014),

- Arima based analyzer to predict future resource demands.
- Update the model with new data to enhance the prediction accuracy.
- Use the predicted results to enhance the QOS.

Calheiros, R. N., Masoumi, E., Ranjan, R. & Buyya, R. (2014), 'Workload prediction using arima model and its impact on cloud applications' qos', IEEE transactions on cloud computing 3(4), 449–458.

Future Workload Prediction using Statistical Methods

'Workload prediction using arima model and its impact on cloud applications' qos',Calheiros, R. N., Masoumi, E., Ranjan, R. & Buyya, R. (2014),

- Used an analyzer implemented using the ARIMA model.
- First, the data that needs to predict the future workload such as CPU usage, memory usage, etc will be fed into the ARIMA model.
- An important point considered here is these data must be stationary. For that process, a transformation method called differencing has been used.
- ARIMA model needs three(3) parameters to be fed the number of differences used, the auto-correlation value, and the partial auto-correlation value.
- The analyzer used in this paper will be responsible for updating the model with new data to increase the prediction accuracy of future workload behaviors.

Calheiros, R. N., Masoumi, E., Ranjan, R. & Buyya, R. (2014), 'Workload prediction using arima model and its impact on cloud applications' qos', IEEE transactions on cloud computing 3(4), 449–458.

What is the best approach?

Machine Learning Techniques

OR

Statistical Techniques

Factors to be considered,

- Model Performance Metrics - Accuracy
- Complexity of the Model - Overfitting and Underfitting / Number of Parameters
- Computational Efficiency- Training Time/ Time required for the model to make predictions

Existing Dynamic Workload Migration Strategies

vhaul: Lu, H., Xu, C., Cheng, C., Kompella, R. & Xu, D,'2015 IEEE 8th International Conference on Cloud Computing',

- Provides a migration strategy for multi tier applications.
- VHaul categorize the VMs according to their relationships.
- Create a migration schedule based on the production of resource utilization and the migration time.

Lu, H., Xu, C., Cheng, C., Kompella, R. & Xu, D. (2015), vhaul: Towards optimal scheduling of live multi-vm migration for multi-tier applications, in '2015 IEEE 8th International Conference on Cloud Computing', IEEE, pp. 453–460.

Existing Dynamic Workload Migration Strategies

Glap:Khelghatdoust, M., Gramoli, V. & Sun, D. 2016 IEEE International Conference on Cluster Computing (CLUSTER)

- Developed a fully distributed and threshold-free Dynamic Virtual Machine Consolidation algorithm.
- The solution relies on an unstructured gossip-based protocol and Q-Learning .
- Migrate the VM using 2 a two phase algorithm

Khelghatdoust, M., Gramoli, V. & Sun, D. (2016), Glap: Distributed dynamic workload consolidation through gossip-based learning, in '2016 IEEE International Conference on Cluster Computing (CLUSTER)', IEEE, pp. 80–89.

Research Gap

- Most of the researchers has used ML techniques and complex mathematical models to handle the behavior of the dynamic workload.
- Implementing and using a ML model is identified as a complex and resource intensive operation because it requires
  - A long period of training time
  - A huge set of data requirements
  - A high computation power

- What if we can do the same process using a **lightweight solution with minimal performance impact.**

Research Questions

1. How to find the convergence point of a dynamic workload in Live migration?

2. How can the performance of vanilla pre-copy be improved upon detection of convergence point in terms of total migration time, downtime, and application performance?

Objectives

- Design and develop a method to analyze the profiling details and identify the peaks of dynamic workload patterns.
- Identify the convergence point of a dynamic workload by analyzing the above results.
- Evaluate the performance of migration using total migration time, downtime, and application performance metrics by incorporating industrial accepted benchmarks such as Sysbench, YCSB (Yahoo cloud serving benchmark), Memcached, etc.
- Enhance the pre-copy algorithm to migrate a dynamic workload with minimal total migration time and downtime.

Scope

- Live VM migration with dynamic workloads
- Single VM migration
- Focus on Ubuntu host OS .
- Migrate VMs with Linux guest OS
- Implementing a working prototype
- QEMU-KVM Hypervisor will be used for the prototype implementation

Research Methodology and Design

Testbed setup: Two physical servers interconnected with Gigabit ethernet will be used.(**16 GB RAM, 1000 Mbps Bandwidth and 500+GB HDD**)

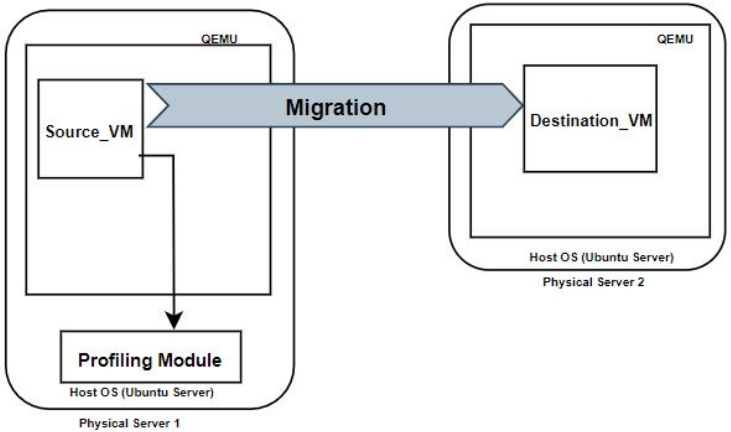
Workload prediction model selection: Choose the most suitable statistical approach to predict future behavior by using the data collected using profiling tools

Design and development of the profiling module : After selecting the most suitable approach the profiling module will be implemented inside the source host.

Research Methodology and Design

Development of the algorithm: An algorithm should be developed to decide the migration strategy to be used based on the workload behavior.

Implement the working prototype: With the use of the implemented algorithm and profiling module a working prototype will be implemented for the KVM/QEMU platform

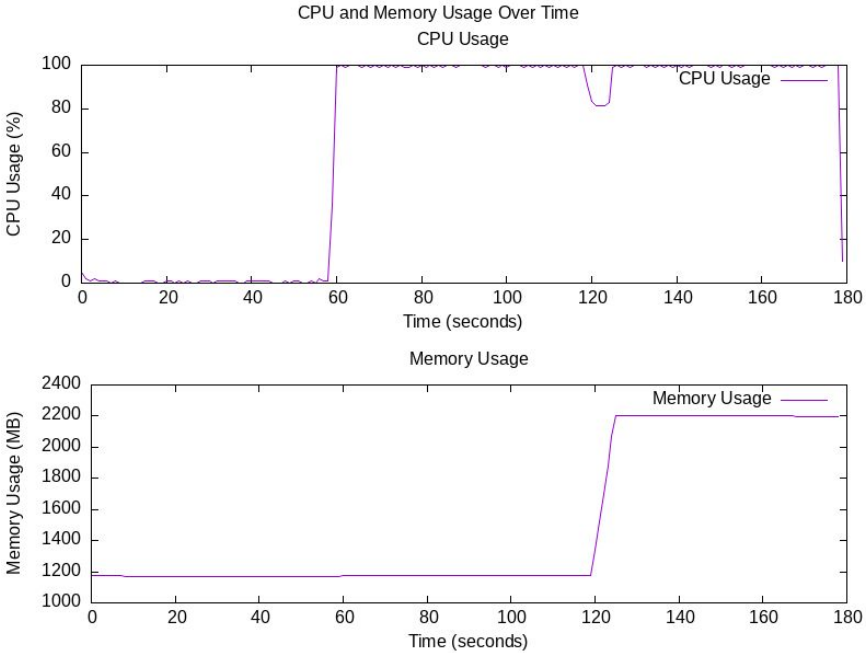


Evaluation

- The implemented prototype system will be evaluated through experiments conducted on the real test-bed to assess its efficiency
- The final result of the research aims to find a convergence point for a dynamic workload and then used the most optimal migration method for the migration where we can reduce the **total migration time and downtime with minimal performance impact.**

Progress So Far

- Try to capture the CPU usage, memory usage and network usage using profiling tools such as mpstat,free -h, vmstat and perf.
- Generate different types of workloads by using industrial accepted benchmarks such as Sysbench,YCSB, and Memcached.
- Identify how the workload changes by plottings the results captured using the profiling tools.



Timeline

	June	July	August	September	October	November	December	January	February	March	April
Literature Survey											
Background research											
Project proposal											
Collect data using profiling tool											
statistical approach to predict the future											
Desing and implement the solution in KVM/QEMU platform											
Test and analyze the results for pre-copy migration											
Optimize the solution to be used for post - copy migration											
Overall performance evaluation											
Thesis writing											
Research publication											

References

- Wu, Q. & Wolf, T. (2008), Dynamic workload profiling and task allocation in packet processing systems, in '2008 International Conference on High Performance Switching and Routing', IEEE, pp. 123–130.
- Ye, K., Wu, Z., Wang, C., Zhou, B. B., Si, W., Jiang, X. & Zomaya, A. Y. (2014),'Profiling-based workload consolidation and migration in virtualized data centers',IEEE Transactions on Parallel and Distributed Systems 26(3), 878–890.
- Han, J., Hong, Y. & Kim, J. (2020), 'Refining microservices placement employing workload profiling over multiple kubernetes clusters', IEEE access 8, 192543–192556.
- Banerjee, P., Roy, S., Modibbo, U. M., Pandey, S. K., Chaudhary, P., Sinha, A. &Singh, N. K. (2023), 'Optidjs+: A next-generation enhanced dynamic johnson sequencing algorithm for efficient resource scheduling in distributed overloading within cloud computing environment', Electronics 12(19), 4123.
- Calheiros, R. N., Masoumi, E., Ranjan, R. & Buyya, R. (2014), 'Workload prediction using arima model and its impact on cloud applications' qos', IEEE transactions on cloud computing 3(4), 449–458.

References

- Lu, H., Xu, C., Cheng, C., Kompella, R. & Xu, D. (2015), vhaul: Towards optimal scheduling of live multi-vm migration for multi-tier applications, in '2015 IEEE 8th International Conference on Cloud Computing', IEEE, pp. 453–460.
- Khelghatdoust, M., Gramoli, V. & Sun, D. (2016), Glap: Distributed dynamic workload consolidation through gossip-based learning, in '2016 IEEE International Conference on Cluster Computing (CLUSTER)', IEEE, pp. 80–89.

Summary

- VM migration is a widely used technique that has been used to handle the interruptions occurred due to server failures.
- Migrating a dynamic workload has become a complex problem because of the frequently changing behavior in dynamic workloads.
- This research aims to decide the convergence of a dynamic workload with a lightweight solution that will not degrade performance of the host machine.

Thank you!

Q&A

Data Collection Script

```
echo "Idling for 10 seconds..."
sleep 60

# Run Sysbench CPU benchmark for 60 seconds
sysbench cpu --time=60 run

# Run memory workload for 60 seconds
./memory_workload &
MEMORY_WORKLOAD_PID=$!
sleep 60
kill $MEMORY_WORKLOAD_PID
```

Univariate Models VS Multivariate Models

Univariate(ARIMA,ETS)	Multivariate(VAR,PLSR)
generally simpler and faster to compute.	Multivariate models handle all metrics in a single model, which can be computationally intensive
training time for each univariate model is lower compared to multivariate models	take longer because they need to estimate more parameters and account for interdependencies between the time series
Forecasting with univariate models is usually quick because each model only deals with one time series.	