CSE449: PARALLEL, DISTRIBUTED, AND HIGH-PERFORMANCE COMPUTING (HPC)

Submitted by:

Asibur Rahman Bhuiyan

ID: 23341095

Section: 01

Team: 26

Submitted to:

Annajiat Alim Rasel

Senior Lecturer,

Brac University

Submission No: 01

Task: Paper Review Presentation 1

Paper title: Heterogeneous Task Co-location in Containerized Cloud Computing Environments

Author: *Zhiheng Zhong, Jiabo He, Maria A. Rodriguez, Sarah Erfani, Ramamohanarao Kotagiri, and Rajkumar Buyya*

TABLE OF CONTENTS

INTRODUCTION

- Cloud Computing in Industry
- Resource Utilization Challenge in Cloud Datacenters
- Task Colocation for Efficiency
- Modern Cloud-Based Cluster Management Systems
- Performance Interference in Co-located Workloads
- Container Technology for Isolation
- Workload Characterization for Task Placement
- Improving Resource Utilization with CTCL Scheduler

PROBLEM DEFINITION

- Workload Characterization
- Behavior Identification
- Scheduling
- Scaling
- Rescheduling

SCHEDULING ALGORITHMS

- Efficient Initial Placement
- Pending Task Management
- Task Allocation
- Two-Phase Selection
  - Filtering
  - Ranking
- LRP Algorithm
- Advantages Of Bin Packing Algorithms
- Resource Evaluation

RESCHEDULING ALGORITHMS

- Shortest Runtime Rescheduling (SRR) Algorithm
- Minimizing QoS Degradation
- Rescheduling Process Criteria
- Task Reallocation Using LRP Algorithm

GREEDY AUTOSCALING ALGORITHM

- Autoscaling Invocation
- Greedy Autoscaling (GA) Algorithm
- Efficient Resource Utilization
- Scaling Down for Idle Nodes

PERFORMANCE EVALUATION

Workload Evaluation:

- Resource Reservation of Long-Running Services
- Batch Instance Arrival Rate

Simulation Environment:

- Instance Configurations
- Kubernetes (K8s) Scheduling Policies
- Proposed Scheduler (CTCL) Configuration

Instance Type	CPU	Memory (normalized)	Bandwidth
ecs.ebmc5s.24xlarge	96	0.5	30 Gbit/s
ecs.ebmg5s.24xlarge	96	1	30 Gbit/s
ecs.ebmr5s.24xlarge	96	2	30 Gbit/s

Table 1: Instance Configurations

Parameter	Value
Decision making time of task placement	20ms
Container startup time	300ms
Task eviction time	2ms
Time interval in utilization prediction	15min
Instance acquisition lag	4-7min

Parameter	Value
Reclustering cycle	24h
Empirical observation time	12h, 24h
Resource utilization threshold in rescheduling	80%
Weight of CPU	0.38
Weight of memory	0.62

Table 2: Simulation Parameters

EXPERIMENTAL RESULTS SUMMARY

Workload Characterization:

- Resource Efficiency Evaluation
- CPU Utilization Pattern
- Memory Usage Stability

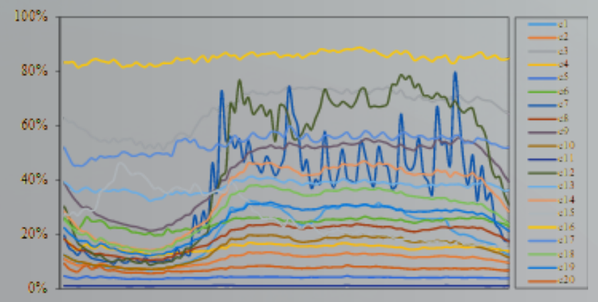

Fig. 1: Clustering results of CPU utilization (k = 20)

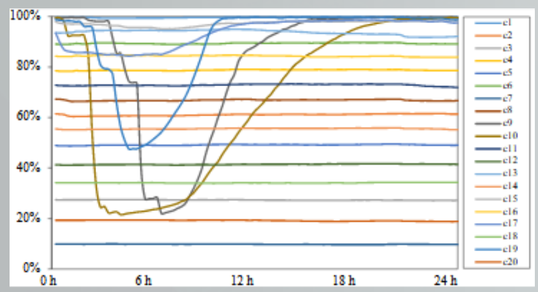

Fig. 2: Clustering results of memory utilization (k = 20).

Resource Efficiency:

- Data Collection and Verification
- Metrics Evaluated
- Performance Metrics
- Resource Efficiency Comparison
- Resource Usage Prediction
- Batch Instance Rescheduling Rate
- Task Rescheduling Rate

Algorithm	Average CPU Utilization (%)	Average Memory Utilization (%)	Maximum Cluster Size
AT	39	78	952
K8s	41	83	933
CTCL (20c-12h)	63	90	808
CTCL (20c-24h)	65	87	777

FUTURE SCOPES

- Improve scalability of workload characterization approach under fast-growing workloads at extreme scale using dynamic incremental K-means++ clustering algorithm
- Develop an energy consumption model in CTCL for managing overall energy efficiency
- Implement scheduling policies in real-world CMS for virtual cluster management, focusing on minimizing cloud resource rental costs through efficient task packing
- Extend CTCL to support task co-location in other workload management scenarios, including network-intensive applications with potential performance interference.

CONCLUSION

- The paper talks about future research ideas for managing computer resources in cloud-based clusters.
- They want to make it easier to handle lots of work in fast-growing environments using a smart method called dynamic incremental K-means++ clustering.
- They also want to create a model to save energy in the system.
- They want to make their system work with different kinds of tasks and make sure they don't interfere with each other when they use the network.
- The paper suggests that their system can make things work faster, cost less, and provide better services in cluster management.