

**CHENNAI – PONDICHERRY**

**Towards Efficient Resource Allocation for Heterogeneous Workloads in IaaS Clouds**

**Abstract:**

Infrastructure-as-a-service (IaaS) cloud technology has attracted much attention from users who have demands on large amounts of computing resources. Current IaaS clouds provision resources in terms of virtual machines (VMs) with homogeneous resource configurations where different types of resources in VMs have similar share of the capacity in a physical machine (PM). However, most user jobs demand different amounts for different resources. For instance, high-performance-computing jobs require more CPU cores while big data processing applications require more memory. The existing homogeneous resource allocation mechanisms cause resource starvation where dominant resources are starved while non-dominant resources are wasted. To overcome this issue, we propose a heterogeneous resource allocation approach, called skewness-avoidance multi-resource allocation (SAMR), to allocate resource according to diversified requirements on different types of resources. Our solution includes a VM allocation algorithm to ensure heterogeneous workloads are allocated appropriately to avoid skewed resource utilization in PMs, and a model-based approach to estimate the appropriate number of active PMs to operate SAMR. We show relatively low complexity for our modelbased approach for practical operation and accurate estimation. Extensive simulation results show the effectiveness of SAMR and the performance advantages over its counterparts.

**Existing System:**

Firstly, the resource demands in users’ jobs are skewed among various resources. If the skewness of resource usages is ignored in resource allocation, some specific resource types with high demand may be exhausted before other resource types with low demand. Secondly, the complexity of resource allocation considering multiple resource types will be significantly increased. The complexity of provisioning algorithms for homogeneous resource allocation is already high and the computational time is long given the large number of PMs in data centers nowadays. The further consideration of multiple resources adds additional dimensions to the computation which will significantly increase the complexity. Thirdly, the execution time of some jobs (e.g., Google trace) can be as short as a couple of minutes which rapidly changes the PM utilization. This rapid change makes provisioning and resource allocation challenging.

**Proposed System:**

This paper proposes a skewness-avoidance multi-resource (SAMR) allocation algorithm to efficiently allocate heterogeneous workloads into PMs. SAMR designs a heterogeneous VM offering strategy that provides flexible VM types for heterogeneous workloads. To measure the skewness of multi-resource utilization in data center and reduce its impact, SAMR defines the multi-resource skewness factor as the metric that measures both the inner-node and the inter-node resource balancing. In resource allocation process, SAMR first predicts the required number of PMs under the predefined VM allocation delay constraint. Then SAMR schedules the VM requests based on skewness factors to reduce both the inner-node resource balance among multiple resources and the inter-node resource balance among PMs in the data center. By such manner, the total number of PMs are reduced significantly while the resource skewness is also controlled to an acceptable level.