

**Dynamic VM Scaling: Provisioning and Pricing through an Online Auction**

**Abstract:**

Today's IaaS clouds allow dynamic scaling of VMs allocated to a user, according to real-time demand of the user. There are two types of scaling: horizontal scaling (scale-out) by allocating more VM instances to the user, and vertical scaling (scale-up) by boosting resources of VMs owned by the user. It has been a daunting issue how to efficiently allocate the resources on physical servers to meet the scaling demand of users on the go, which achieves the best server utilization and user utility. An accompanying critical challenge is how to effectively charge the incremental resources, such that the economic benefits of both the cloud provider and cloud users are guaranteed. There has been online auction design dealing with dynamic VM provisioning, where the resource bids are not related to each other, failing to handle VM scaling where later bids may rely on earlier bids of the same user. As the first in the literature, this paper designs an efficient, truthful online auction for resource provisioning and pricing in the practical cases of dynamic VM scaling, where: (i) users bid for customized VMs to use in future durations, and can bid again in the following time to increase resources, indicating both scale-up and scale-out options; (ii) the cloud provider packs the demanded VMs on heterogeneous servers for energy cost minimization on the go. We carefully design resource prices maintained for each type of resource on each server to achieve threshold-based online allocation and charging, as well as a novel competitive analysis technique based on submodularity of the offline objective, to show a good competitive ratio is achieved. The efficacy of the online auction is validated through solid theoretical analysis and trace-driven simulations.