

**An Efficient and Fair Multi-Resource Allocation Mechanism for Heterogeneous Servers**

**Abstract:**

Efficient and fair allocation of multiple types of resources is a crucial objective in a cloud/distributed computing cluster. Users may have diverse resource needs. Furthermore, diversity in server properties/capabilities may mean that only a subset of servers may be usable by a given user. In platforms with such heterogeneity, we identify important limitations in existing multi-resource fair allocation mechanisms, notably Dominant Resource Fairness and its follow-up work. To overcome such limitations, we propose a new server-based approach; each server allocates resources by maximizing a per-server utility function. We propose a specific class of utility functions which, when appropriately parameterized, adjusts the trade-off between efficiency and fairness, and captures a variety of fairness measures (such as our recently proposed Per-Server Dominant Share Fairness). We establish conditions for the proposed mechanism to satisfy certain properties that are generally deemed desirable, e.g., envy-freeness, sharing incentive, bottleneck fairness, and Pareto optimality. To implement our resource allocation mechanism, we develop an iterative algorithm which is shown to be globally convergent. Subsequently, we show how the proposed mechanism could be implemented in a distributed fashion. Finally, we carry out extensive trace-driven simulations to show the enhanced performance of our proposed mechanism over the existing ones.