**Improving Application Placement for Cluster-Based Web Applications**

**Abstract:**

Dynamic application placement for clustered web applications heavily influences system performance and quality of user experience. Existing approaches claim that they strive to maximize the throughput, keep resource utilization balanced across servers, and minimize the start/stop cost of application instances. We investigate how to minimize the resource utilization of servers in the worst case, aiming at improving load balancing among clustered servers. Our contribution is twofold. **W**EB applications make it possible to deliver critical services provided by organizations directly to clients we propose an enhanced application placement framework, which complements previous works and has contributions in the following four aspects.

First we propose and define a new optimization objectives: limiting the worst case of each individual server’s utilization, formulated by a min-max problem. A novel framework based on binary search is proposed to detect an optimal load balancing solution. Second, we define system cost as the weighted combination of both placement change and inter-application communication cost. By maximizing the number of instances of *dependent applications* that reside in the same set of servers, the basic load-shifting and placement-change procedures are enhanced to minimize whole system cost. Extensive experiments have been conducted and effectively demonstrate that: 1) the proposed framework achieves a good allocation for clustered web applications. In other words, requests are evenly allocated among servers, and throughput is still maximized; 2) the total system cost maintains at a low level; 3) our algorithm has the capacity of approximating an optimal solution within polynomial time and is promising for practical implementation in real deployments. We have presented a novel framework and a practical algorithm for application placement motivated by the desire to minimize worst case server utilization and improve load balancing.