Dynamic Resource Allocation in Enterprise Systems

James Wen Jun Xue
High Performance Systems Group

Abstract

Many companies often outsource their IT infrastructure to Internet Service Providers (ISPs) to reduce their operational cost. The ISPs have their own resource centre and it is common that most of the time they are hosting many services simultaneously. They provide services based on Service Level Agreements (SLAs) between the themselves and IT service companies. In order to make profits, ISPs have to make efficient use of their resources, while providing the agreed services on behalf of their customers.

Enterprise systems are typically multi-tiered, consisting of web servers, application servers and database servers. Some systems also employ load balancers or edge servers before web server tier to balance the workload. In each tier, there is normally a cluster of servers to improve the processing power. In this work, we model typical enterprise systems using a multi-class closed queueing network. The advantage of using an analytical model is that we can easily capture the different performance metrics and identify potential bottleneck even without running the actually system. The model can also react to parameter changes when the application is running (e.g. from the monitoring tools, or system logs) and make dynamic server switching decisions to optimise pre-defined performance metrics (e.g. minimise the mean response time or maximise the site revenue).

Bottlenecks are a phenomenon where the performance or capacity of an entire system is severely limited by a single component. The component is sometimes called the *bottleneck point*. Formally, a bottleneck lies on a system's critical path and provides the lowest throughput. The multi-tiered architecture of an enterprise system can introduce bottleneck, which will limit the overall system performance. Moreover, the population mix for a particular application often changes during run-time, which can shift system bottleneck between tiers. Therefore, system designers need to study the best server configuration to avoid bottlenecks during system capacity planning and provisioning stage, and ideally provide schemes to support dynamic server allocation during run-time.

Workload demand for Internet service is usually very bursty, thus it is very difficult to predict the workload level at a certain point in time. Thus, fixed server configurations for a service are far from satisfactory for an application when the workload level is high; whereas it is potentially a waste of resource while the workload is light for the remaining applications supported by the system. Thus, it is desirable that server resources in a shared hosting environment can be switched between applications to accommodate workload variation.

We propose in this paper a model-driven server switching system to dynamically allocate resources for multi-tiered enterprise architectures in order to achieve highest revenue. The switching decision is guided by the bottleneck identification results from an established approach. A local search algorithm is designed and used to search for potentially good server configuration as the basis for server switching when system state changes. Furthermore, an admission control scheme is used in the proposed server switching system to maintain the number of simultaneous jobs in an enterprise system at an appropriate value.