

Proactive Workload Management in Hybrid Cloud Computing

Abstract:

The hindrances to the adoption of public cloud computing services include service reliability, data security and privacy, regulation compliant requirements, and so on. To address those concerns, we propose a hybrid cloud computing model which users may adopt as a viable and cost-saving methodology to make the best use of public cloud services along with their privately-owned (legacy) data centers. As the core of this hybrid cloud computing model, an intelligent workload factoring service is designed for proactive workload management. It enables federation between on- and off-premise infrastructures for hosting Internet-based applications, and the intelligence lies in the explicit segregation of base workload and flash crowd workload, the two naturally different components composing the application workload. The core technology of the intelligent workload factoring service is a fast frequent data item detection algorithm, which enables factoring incoming requests not only on volume but also on data content, upon a changing application data popularity. Through analysis and extensive evaluation with real-trace driven simulations and experiments on a hybrid testbed consisting of local computing platform and Amazon Cloud service platform, we showed that the proactive workload management technology can enable reliable workload prediction in the base workload zone (with simple statistical methods), achieve resource efficiency (e.g., 78% higher server capacity than that in base workload zone) and reduce data cache/replication overhead (up to two orders of magnitude) in the flash crowd workload zone, and react fast (with an X^2 speed-up factor) to the changing application data popularity upon the arrival of load spikes.