

**CHENNAI – PONDICHERRY**

**ECOUP: TOWARDS ECONOMICAL DATACENTER UPGRADING**

**ABSTRACT**

The rapid growth of cloud services dictates increasingly powerful datacenters to maintain the high quality of service (QoS). It's a common practice in virtually all tiers of datacenters to continuously upgrade the datacenters, i.e. replacing outdated and failed servers with more advanced and efficient ones. However, how to upgrade a datacenter in the most cost-efficient strategy remains unclear, and however this problem goes increasingly challenging given the great diversity of applications. In practice, the datacenters' operators usually resort to expending the scale of servers. The preferred servers are either expensive but high-performance, or, by contrast, cheap but low-power. Whatever sever preferences, how to justify the cost-efficiency is still an open problem. We claim that a cost-efficient upgrading strategy should be fully aware of not only the capacity and cost of various servers, but also the resource demands of target applications. We model this strategy as a recommendation problem: recommending the “best” servers to a datacenter. We propose “EcoUp”, a model-based framework that faithfully rates the cost efficiency of server candidates, relying on which an optimal server portfolio can be derived. The performance prediction on candidate servers is realized by employing a sophisticated latent factor model (LFM). The cost mainly involves the server purchasing cost and energy bill. Given the application distribution, EcoUp can give an optimal server portfolio under a certain capital budget. We use Google trace, a big profiling dataset opened by Google, to validate the performance prediction. Experimental results show that the error rate is below 8 percent on average. Meanwhile, we build a comprehensive upgrading procedure on a local cluster to evaluate the potential of EcoUp. The results show that our approach significantly outperforms two conventional upgrading strategies by 12.3 and 33.6 percent in terms of system throughput, respectively.