Abstract

Many critical e-commerce and financial services are deployed on geo-distributed data centers (GDCs) for scalability and availability. Recent market surveys show that failures are common in the data centers and this results in a huge financial loss. Designing data centers for high availability includes spare capacity provisioning across the data centers. The work in this thesis addresses the problems of cost-aware capacity provisioning and load balancing in fault-tolerant GDCs (to mask the failures at a site). We propose optimization models for cost-effective planning and operation of the GDCs and propose algorithms for solving these. First, we propose an optimization model to distribute the servers across the GDC such that, the total cost of ownership (TCO) for an operator is minimized. The model identifies the optimal server distribution and optimal request routing policy to exploit the spatio-temporal variation in the electricity prices and user demand for minimizing the TCO. Next, we extend the optimization model for capacity planning in GDCs collocated with renewable energy sources. Using this model, the operators can reduce their carbon footprint by maximizing the green energy usage, while minimizing the TCO. We also extend this model to consider GDCs powered by both brown and green energy sources. In such a case, we use an objective to minimize the total cost while ensuring that a certain percentage of green energy is always used.

In this thesis, we also address another important problem, cost-aware load balancing in large-scale fault-tolerant GDCs. We use game theory to formulate the problem of cost-aware distributed load balancing in GDCs. We use a non-cooperative game executed across a finite number of front-end proxy servers, with an objective of minimizing the linear combination of operating cost and revenue loss due to increased latency. Based on the structure of Nash equilibrium, a distributed load balancing algorithm is proposed. The proposed algorithm is decentralized, has a low complexity, and offers fairness in average latency perceived by the clients. Lastly, we propose a two-stage distributed algorithm for load balancing after the failure of a data center. The proposed algorithm spreads the load of failed data center minimizing the operating cost and then, re-routes the requests considering

the delay and green energy usage constraints. All the proposed algorithms are evaluated using real-world data set.

Results shows that the proposed approaches yield optimal results in planning and operation of fault-tolerant GDCs, powered by both brown and green energy sources. We conclude that it is indeed possible to minimize the cost of running GDCs considering the spatio-temporal dynamics and it is possible to mask single data center failure with no additional cost using the proposed models. We conclude that with a suitable model, green energy integration lowers the cost of designing fault-tolerant GDCs(despite green energy being costlier). Our model works well even with uncertainty in the available wind energy and achieves a significant reduction in the cost as the technology advances. We also show that online load balancing algorithms should be cost-aware in distributing the requests so that, the operating cost is also minimized apart from the latency. We designed an algorithm that ensures delay fairness to the clients without increasing the operating cost.