Chapter 5
Conclusion and Future Work

Abstract We have explored the design space for cost-effective and flexible resource management strategies in utility and cloud computing. In this final chapter, we summarize our findings and discuss related future work directions based on the solutions depicted in this book.

5.1 Concluding Remarks

In this book, we have explored the design space for cost-effective and flexible resource management strategies in utility and cloud computing, and proposed a few novel solutions to address the challenges of scalability and heterogeneity. In the second chapter, we investigated the problem of fine-grained resource rental management in utility and cloud computing, and developed solutions for both deterministic and stochastic resource pricing settings. Our optimization models were based on a thorough rental cost analysis of elastic application deployment in the cloud resource market. When resource pricing is fixed, we observed the cost tradeoff between computing and storage emerges in time-slotted resource provision scheduling. Based on this observation, we formulated a deterministic optimization model that effectively minimizes rental cost of virtual servers while covering customer demand over certain planning horizon. In addition, we took one step further to analyze the predictability of spot resource prices using Amazon®'s spot instance price trace, and proposed an alternative stochastic optimization model that seeks to minimize the expected resource rental cost given the presence of spot price uncertainty. Simulations based on realistic settings clearly demonstrated the advantage of the stochastic optimization approach over the predictive approach in rental cost reduction. We also studied the impact of various parameter settings on the performance of both models. We believe the proposed solutions for rental planning offer effective means for resource rental management in practice.

In the third chapter, we presented the management problem of resource trading in a community-based cloud computing environment. The goal of this study is to

H. Zhao and X. Li, *Resource Management in Utility and Cloud Computing*, SpringerBriefs in Computer Science, DOI 10.1007/978-1-4614-8970-2_5,

investigate the interactions among independent and rational resource traders, and to establish effective and easy-to-implement negotiation protocols for system-wide allocation efficiency and fairness. Towards this goal, we first adopted a multiagent-based optimization framework and analyzed the optimal results without concerning about budget limitation. Next, we proposed a novel directed hypergraph model that combines allocation and envy relationship in a three-dimensional hyperspace. This model effectively captured the impact of trading selection decisions from a global point of view. When budget limitation is imposed, we developed a set of distributed resource trading protocols based on heuristic approaches. Simulation results show that the proposed protocols perform well in a wide range of settings. We expect that the solution for resource trading management presented in this chapter would open new vistas for designing effective resource management strategies.

Finally, we presented CloudBay, a novel resource sharing middleware stack composed of resource management software stack from ground up. Equipped with virtual networking and application-aware virtual appliances, CloudBay achieves ad-hoc self-organization, discovery and grouping of distributed resources without incurring extra deployment and management efforts from both resource providers and end users. Moreover, CloudBay implements a market-driven service scheduling policy that accommodates a mixture of user request models, and efficiently distributes idle resources to users in a cost-effective manner. The pricing and payment accounting policies boosts utilities for multiple parties, and features fair resource allocation for customers. Utilizing services provided by CloudBay, researchers with domain knowledge can comfortably deploy their parallel applications using popular parallel programming models on a resource bundle assembled from multiple organizations. We have already deployed virtual appliances across a variety of open and private cloud platforms, including university clusters, FutureGrid, and Amazon® EC2. We expect that our experiences gained from the design and implementation of CloudBay would open a new research avenue for realizing HPC-as-a-service, and push the boundary for new cloud computing usage models.

5.2 A Look into the Future

The research of distributed systems encompasses many areas of computer science and is among the fastest developing fields in the past decade. As resource management needs to cope with the growing complexity of the distributed systems, the exploration presented in this book is just a starting point. We expect the design space to be growing tremendously as distributed systems scale. In particular, this book focuses on the improvement of resource management in distributed systems involving mutually distrustful components. This problem will become more and more important as present and future big data applications call for scalable and reliable computing platforms. The following quote from IEEE Distributed Systems Online [1] published a decade ago has foreseen this challenge, "*... In the past, our approach has been to build systems involving mutually trusting and mutually*

cooperating subsystems ... We need architectures that support cooperation for achieving a common goal but that do not require subsystems to make strong assumptions about peers". Opportunities are emerging to use user-centric approaches that cater to highly dynamic participants. These approaches, such as game theory and auction theory, will inevitably present a substantial body of research for resource management in the years to follow.

Efficiently managing resource allocation is of paramount importance in almost all disciplines of distributed computing. In particular, we are interested in three directions that have the most momentum. The first direction is *Scientific Computing* which applies computational resources to scientific problems. The theoretical peak performance of a single modern GPU has reached 3.7TFLOPS, almost two times as fast as that of the world's fastest supercomputer in year 2000. Such technology advance in computing enables scientists to tackle computing demanding problems with large and costly simulations. Designing efficient resource management strategies for scientific computing is difficult for three reasons. First, uncertainty is ubiquitous in scientific modeling, making resource allocation requirements changing all the time. Second, the need for online processing of the scientific data sets introduces additional demands to computational, storage, and network resource management. Finally, many scientific computing applications involve legacy codes and systems, requiring tremendous efforts to transit to new computing infrastructure. Utility computing and infrastructure clouds offer great potential for scientific users, and we believe the marriage of scientific computing and the cloud will create an exciting perspective in the long run, especially for loosely coupled large-scale HPC applications. This book has presented some of our preliminary research findings on this topic. In the future, we expect to see more research addressing cost and privacy issues of the cloud. In addition, the scientific community needs to invest substantial amount of time and money in developing utility- and cloud-aware tools and services for existing scientific applications and workflows.

The second direction is *Big Data* which deals with high-volume information storage, query and analysis. Interest in big data has given rise to building distributed systems geared for data-intensive processing, and resource management plays a central role in supporting big data applications. In order to handle massive data and meet the performance critical real-time demand, resource management should be agile to allow flexible deployment and provisioning. New platforms have been introduced for big data applications, e.g., Hadoop for better parallelism in computing, and NoSQL for scalable unstructured data storage. The resource management solutions thus need to improve in light of changes in the big data landscape. When evaluating a new resource management solution, performance along with other factors such as deployment complexity, cost, and interoperability with existing solutions, combine to influence the quality of the solution. In this book, we follow this general idea and conduct cost-benefit analysis to resource management in a utility-oriented setting. Another interesting problem is to tradeoff reliability vs. availability of resource allocation, as both resource providers and customers need to determine the allocation of reserved and on-demand resources to minimize waste. Big data applications in cloud also bring more challenge for

gracefully handling of resource loss and reallocation. In general, much research and development remain to be carried out to catch up with the ever-increasing pace of data grow.

Finally, as *Mobile and Embedded Computing* proliferate in recent years, efficient resource management is in urgent need to offload heavy computational tasks for mobile and embedded devices. Because these devices are architecturally heterogeneous and resource constrained, they become more and more relied on the cloud computing infrastructure. In order to guarantee Quality-of-Service (QoS) for applications, it is critical to effectively manage resource sharing in data center, and offer more capable network interconnection with enhanced switching and routing. We investigated the economics of resource sharing in this book, and our focus is mainly on the management of computational resource. In the future, we plan to explore management strategies towards network resource sharing. The emerging Software Defined Networking (SDN) technology separates control plan from the data plan and provides centralized control functions with a SDN controller. With this change, it is interesting to examine how multiple network flows belonging to different applications should be shared, scheduled, and priced in modern data centers.

Reference

1. Milojicic, D.S., Schneider, F.B.: Interview - Fred B. Schneider on Distributed Computing. IEEE Distributed Systems Online **1**(1) (2000)