Adaptive Resource Relocation in Virtualized Heterogeneous Clusters

A thesis submitted for the degree of
Doctor of Philosophy
of
The Australian National University

Muhammad Atif
November 2010

I declare that the work in this thesis is entirely my own and that to the best of my knowledge it does not contain any materials previously published or written by another person except where otherwise indicated.

Muhammad Atif
November 2010

This thesis is dedicated to the loving memory of my father, who taught me to believe in myself and have a positive attitute towards life.

I also dedicate the thesis to my mother.

Acknowledgements

I would like to express appreciation for Dr. Peter Strazdins for his guidance and advise during my doctoral research for the past three and a half years. As my supervisor, he guided me to remain in focus towards achieving the goal. His observations and comments helped me to broaden the knowledge in the research area and beyond. He also was instrumental in establishing the overall direction of the research and steer me towards the goals and milestones of my research work.

I would like to thank my advisors Dr. Alistair Rendell and Dr. Eric McCreath for their useful comments, suggestions and encouragement throughout my research work. They extended all the possible support during my research work, which was not only at the academic level but also at the personal level.

My special gratitude for Computer Systems Research Group at the department of computer science, ANU. These 'geeks' have a great tradition of supporting each other in every possible way. Their constant support, discussions and useful suggestions acted as catalysts to my research output. I would like to especially thank Mr. Ahmed El Zein from the ANU supercomputer facility for helping me in setting up the compute farm.

Appreciation for Alexander Technology and Platform Computing for donating the hardware, especially Mr. Richard Alexander for his vision that enabled this research work.

My very special thanks to National University of Sciences and Technology (NUST), Pakistan for providing me a scholarship for the first three years of my candidature at the Australian National University.

I sincerely thank Department of Computer Science, ANU for providing me tuition fee and living allowance scholarship for the rest of my candidature. This enabled me to remain in focus to the research and not to worry about financial constraints. Especially Dr. Peter Christian, who in his capacity as Associate Dean (Higher Degree Research) always extended his full support. I would take this opportunity to thank all the administration and technical staff at the department of computer science. During the three and a half years of research, I never faced any administration issue.

To my wife Bushra, and our children Rameen and Zain, thank you for being there for me throughout the candidature. Especially Bushra for giving me strength and support during the tough times. I would also like to thank my mother for remembering me in her prayers.

Abstract

Cluster computing has recently gone through an evolution from single processor systems to multicore/multi-socket systems. This has resulted in lowering the cost/performance ratio of the compute machines. Compute farms that host these machines tend to become heterogeneous over time due to incremental extensions, hardware upgrades and/or nodes being purchased for users with particular needs. This heterogeneity is not surprising given the wide range of processor, memory and network technologies that become available and the relatively small price difference between these various options. Different CPU architectures, memory capacities, communication and I/O interfaces of the participating compute nodes present many challenges to job scheduling and often result in under or over utilization of the compute resources. In general, it is not feasible for the application programmers to specifically optimize their programs for such a set of differing compute nodes, due to the difficulty and time-intensiveness of such a task.

The trend of heterogeneous compute farms has coincided with resurgence in the virtualization technology. Virtualization technology is receiving widespread adoption, mainly due to the benefits of server consolidation and isolation, load balancing, security and fault tolerance. Virtualization has also generated considerable interest in the High Performance Computing (HPC) community, due to the resulting high availability, fault tolerance, cluster partitioning and accommodation of conflicting user requirements. However, the HPC community is still wary of the potential overheads associated with virtualization, as it results in slower network communications and disk I/O, which need to be addressed.

The live migration feature, available to most virtualization technologies, can be leveraged to improve the throughput of a heterogeneous compute farm (HC) used for HPC applications. For this we mitigated the slow network communication in Xen; an open source virtual machine monitor. We present a detailed analysis of the communication framework of Xen and propose communication configurations that give 50% improvement over the conventional Xen network configuration. From a detailed study of the migration facility in Xen, we propose an improvement in the live migration facility specifically targeting HPC applications. This optimization gives around 50% improvement over the default migration facility of Xen.

In this thesis, we also investigate resource scheduling in heterogeneous compute farm with the perspective of dynamic resource re-mapping. Our approach is to profile each job in the compute farm at runtime, and propose a better resource mapping compared to the initial allocation. We then migrate the job(s) to the best-suited homogeneous sub-cluster to improve overall throughput of the

HC. For this, we develop a novel heterogeneity and virtualization-aware profiling framework, which is able to predict the CPU and communication characteristics of high performance scientific applications.

Our experimental setup consists of a 32 node virtualized cluster with machines having CPU frequency between 2 GHz to 3 GHz. The framework was tested with a subset of NAS parallel benchmarks with the workload generated from Dror Feitelson's workload archives. The prediction accuracy of our performance estimation model is over 80%. The framework implementation is lightweight, with an overhead of 3%. Our experiments show that we are able to improve the throughput of the compute farm by 25% and the time saved by the HC with our framework is over 30%. The framework can be readily extended to HCs supporting a cloud computing environment.

Contents

List of Figures

List of Tables