UMEÅ UNIVERSITET

Autonomous Resource Management for High Performance Datacenters

Abel Souza

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Aula Biologica (BIO.E.203), Biologihuset , fredagen den 8:e maj, kl. 10:00.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor Alexandru Iosup, Computer Systems Department, Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam, Nederländerna.

Department of Computing Science

Organization	Document type	Date of publication
Umeå University Dept. of Computing Science	Doctoral thesis	17th April 2020

Author

Abel Souza

Title

Autonomous Resource Management for High Performance Datacenters

Abstract

Over the last decade, new applications such as data intensive workflows have hit an inflection point in wide spread use and influenced the compute paradigm of most scientific and industrial endeavours. Data intensive workflows are highly dynamic and adaptable to resource changes, system faults, and also allow approximate solutions into their models. On the one hand, these dynamic characteristics require processing power and capabilities originated in cloud computing environments, and are not well supported by large High Performance Computing (HPC) infrastructures. On the other hand, cloud computing datacenters favor low latency over throughput, deeply contrasting with HPC, which enforces a centralized environment and prioritizes total computation accomplished over-time, ignoring latency entirely. Although data handling needs are predicted to increase by as much as a thousand times over the next decade, future datacenters processing power will not increase as much.

To tackle these long-term developments, this thesis proposes autonomic methods combined with novel scheduling strategies to optimize datacenter utilization while guaranteeing user defined constraints and seamlessly supporting a wide range of applications under various real operational scenarios. Leveraging upon data intensive characteristics, a library is developed to dynamically adjust the amount of resources used throughout the lifespan of a workflow, enabling elasticity for such applications in HPC datacenters. For mission critical environments where services must run even in the event of system failures, we define an adaptive controller to dynamically select the best method to perform runtime state synchronizations. We develop different hybrid extensible architectures and reinforcement learning scheduling algorithms that smoothly enable dynamic applications into HPC environments. An overall theme in this thesis is extensive experimentation in real datacenters environments. Our results show improvements in datacenter utilization and performance, achieving higher overall efficiency. Our methods also simplify operations and allow the onboarding of novel types of applications previously not supported.

Keywords

Datacenter, cloud computing, high performance computing, resource management, scheduling, machine learning, reinforcement learning

Language	ISBN	ISSN	Number of pages
English	print: 978-91-7855-286-3 PDF: 978-91-7855-287-0	0348-0542	45 + 5 papers