PREFACE

Cloud computing has recently emerged as one of the buzzwords in the ICT industry. Numerous IT vendors are promising to offer computation, storage, and application hosting services and to provide coverage in several continents, offering service-level agreements (SLA)-backed performance and uptime promises for their services. While these "clouds" are the natural evolution of traditional data centers, they are distinguished by exposing resources (computation, data/storage, and applications) as standards-based Web services and following a "utility" pricing model where customers are charged based on their utilization of computational resources, storage, and transfer of data. They offer subscription-based access to infrastructure, platforms, and applications that are popularly referred to as IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS (Software as a Service). While these emerging services have increased interoperability and usability and reduced the cost of computation, application hosting, and content storage and delivery by several orders of magnitude, there is significant complexity involved in ensuring that applications and services can scale as needed to achieve consistent and reliable operation under peak loads.

Currently, expert developers are required to implement cloud services. Cloud vendors, researchers, and practitioners alike are working to ensure that potential users are educated about the benefits of cloud computing and the best way to harness the full potential of the cloud. However, being a new and popular paradigm, the very definition of cloud computing depends on which computing expert is asked. So, while the realization of true utility computing appears closer than ever, its acceptance is currently restricted to cloud experts due to the perceived complexities of interacting with cloud computing providers.

This book illuminates these issues by introducing the reader with the cloud computing paradigm. The book provides case studies of numerous existing compute, storage, and application cloud services and illustrates capabilities and limitations of current providers of cloud computing services. This allows the reader to understand the mechanisms needed to harness cloud computing in their own respective endeavors. Finally, many open research problems that have arisen from the rapid uptake of cloud computing are detailed. We hope that this motivates the reader to address these in their own future research and

development. We believe the book to serve as a reference for larger audience such as systems architects, practitioners, developers, new researchers, and graduate-level students. This book also comes with an associated Web site (hosted at http://www.manjrasoft.com/CloudBook/) containing pointers to advanced on-line resources.

ORGANIZATION OF THE BOOK

This book contains chapters authored by several leading experts in the field of cloud computing. The book is presented in a coordinated and integrated manner starting with the fundamentals and followed by the technologies that implement them.

The content of the book is organized into six parts:

I. Foundations
II. Infrastructure as a Service (IaaS)
III. Platform and Software as a Service (PaaS/SaaS)
IV. Monitoring and Management
V. Applications
VI. Governance and Case Studies

Part I presents fundamental concepts of cloud computing, charting their evolution from mainframe, cluster, grid, and utility computing. Delivery models such as Infrastructure as a Service, Platform as a Service, and Software as a Service are detailed, as well as deployment models such as Public, Private, and Hybrid Clouds. It also presents models for migrating applications to cloud environments.

Part II covers Infrastructure as a Service (IaaS), from enabling technologies such as virtual machines and virtualized storage, to sophisticated mechanisms for securely storing data in the cloud and managing virtual clusters.

Part III introduces Platform and Software as a Service (PaaS/IaaS), detailing the delivery of cloud hosted software and applications. The design and operation of sophisticated, auto-scaling applications and environments are explored.

Part IV presents monitoring and management mechanisms for cloud computing, which becomes critical as cloud environments become more complex and interoperable. Architectures for federating cloud computing resources are explored, as well as service level agreement (SLA) management and performance prediction.

Part V details some novel applications that have been made possible by the rapid emergence of cloud computing resources. Best practices for architecting cloud applications are covered, describing how to harness the power of loosely coupled cloud resources. The design and execution of applications that leverage

cloud resources such as massively multiplayer online game hosting, content delivery and mashups are explored.

Part VI outlines the organizational, structural, regulatory and legal issues that are commonly encountered in cloud computing environments. Details on how companies can successfully prepare and transition to cloud environments are explored, as well as achieving production readiness once such a transition is completed. Data security and legal concerns are explored in detail, as users reconcile moving their sensitive data and computation to cloud computing providers.

Rajkumar Buyya
The University of Melbourne and Manjrasoft Pty Ltd., Australia

James Broberg
The University of Melbourne, Australia

Andrzej Goscinski
Deakin University, Australia