Management towards Reducing Cloud Usage Costs

Vladimir Tosic\textsuperscript{1,2,3}, Hiroshi Wada\textsuperscript{1,2}, Adnene Guabtni\textsuperscript{1,2}, Kevin Lee\textsuperscript{1,2}, Anna Liu\textsuperscript{1,2}
\textsuperscript{1} Software Systems Research Group, NICTA, Sydney, Australia
\textsuperscript{2} School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
\textsuperscript{3} School of Information Technologies, University of Sydney, Sydney, Australia
vladat@computer.org, \{firstName.lastName\}@nicta.com.au

Abstract—Many organizations are attracted to cloud computing as an ICT (information and communications technology) sourcing model that can improve flexibility and total cost of ICT systems. However, it can be difficult for a prospective cloud customer to determine and manage cloud usage costs. We present an overview of several NICTA research projects that aim at providing information that can help ICT professionals determine various cloud usage costs and make decisions that are appropriate from the business viewpoint. Before migrating an application into a cloud, it is necessary to choose to which cloud to migrate, because there is a huge variety of cloud offerings, with significantly different pricing models. To accurately capture projected operating costs of an application in a particular cloud and enable side-by-side comparison of cloud offerings from different providers, NICTA developed a cost estimation tool that calculates the costs based on usage patterns and other characteristics of the application. This tool can also be used during runtime as an input into making adaptation/control decisions. To collect various runtime metrics (e.g., about the amount of transferred data or received quality of service – QoS) that are necessary for operational management and assessment of cloud usage costs, NICTA developed an innovative tool for flexible and integrated monitoring of applications in clouds and (in case of hybrid clouds) related local data centers. To help determine which runtime adaptation/control decisions are best from the business viewpoint (e.g., incur lowest cost), we extended the WS-Policy4MASC language and MiniZnMASC middleware for autonomic business-driven IT management with events and adaptation actions relevant for cloud management. The tools from the presented projects can be used separately or as parts of a powerful integrated cloud management system (which contains several additional tools).

Keywords—cloud computing; cost estimation; cloud management; business-driven IT management; system management; monitoring; runtime adaptation

ACKNOWLEDGMENT

The work presented in this paper has been conducted by the Software Systems Research Group of NICTA, within the Business Adaptation and Interoperation project. Apart from the paper authors, crucial results for the solutions presented in this paper were made by NICTA researchers Sean Xiong, Quanqing Xu, Zhanwen (Jim) Li, Rainbow Cai, Paul L. Bannerman, and Paul Brebner, as well as NICTA students Ronald Tsang, Qinghua Lu, Dipesh Chauhan, Thi Khanh Van Tran, and Sadeka Islam. NICTA commercialization specialists Bruce McCabe and Paul Mackie supplied market and business insights and contacts of industrial test users. Prof. Alan Fekete from the University of Sydney provided valuable mentorship and advice. Several other researchers and students also contributed to the presented and related NICTA projects on cloud management. The authors thank all of them.

\textsuperscript{NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centres of Excellence program.}

978-1-4577-1792-5/11/$26.00 ©2011 IEEE
Management towards Reducing Cloud Usage Costs

Vladimir Tosic, Hiroshi Wada, Adnene Guabtni, Kevin Lee, Anna Liu

NICTA, Australia
Uni. of New South Wales, Australia
Uni. of Sydney, Australia

The Problem – Example

http://www.youtube.com/watch?v=ae_DKNwK_ms
(The most watched YouTube video on cloud computing, almost 700,000 views)

The Problem – Summarized

"Cloud computing will solve all your IT problems (and more), just buy it from us"

Does it, really?

A Partial Solution

A set of customer-side management tools to optimize business value of cloud usage (e.g. reduce usage costs)

Presentation Outline

• Cloud computing and its diversity
• "How can I choose a cloud computing platform that is right for my needs and has the lowest cost?"
• "How can I uniformly monitor diverse cloud systems, both in-house and on the Internet?"
• "How can I choose run-time cloud adaptations that will optimize my business metrics (e.g. costs)?"
• Conclusions and some areas with open research challenges
Cloud Computing and Its Diversity
- Classifications of clouds
- Limitations of the past work on cloud management
- Our project at NICTA

Cloud Computing – A Definition
- Provisioning of computing:
  - infrastructure (e.g., memory storage, CPU cycles),
  - platforms (e.g., virtualized Linux desktops), or
  - software applications (e.g., customer relationship management suites)
- over a network (often the Internet),
- as a utility/service that can be bought on demand

Types of Real Clouds (NOT Computing)
- Classification based on the provided service:
  - infrastructure as a service (IaaS),
  - platform as a service (PaaS),
  - software as a service (SaaS),
  - other = X as a service (e.g., data, etc.)

Types of Cloud Computing Systems
- Classification based on ownership:
  - public,
  - private,
  - community,
  - hybrid

Past Work on Cloud Management
- Huge amount of recent research papers on cloud computing and many commercial cloud offerings
- Increasing number of papers and commercial tools on cloud management
- However, the majority of these related works:
  - Manage from provider’s (not consumer’s) viewpoint
  - Do not successfully monitor and control/adapt across different cloud types and cloud offerings/systems
  - Optimize technical (and not business) metrics

Our Team’s Project: Business Adaptation and Interoperation
1) Advanced software technologies for fluid and adaptive business-to-business inter-operation
   - Business processes implemented with service-oriented and resource-oriented architectures
2) Innovative solutions and tools for consumer-side management of cloud computing systems
   - Performance analysis of service-oriented systems
   - Estimation of cloud migration and usage costs
   - Comprehensive hybrid cloud run-time monitoring framework
   - Adaptation framework that optimizes business metrics
- History of successful collaborations with Australian industry and government departments
Selecting a Cloud Computing Platform Appropriate for the Consumer’s Needs

- Why Is It Needed and Difficult?
- NICTA’s Cloud Cost Estimation Tool

Why Is Cloud Selection Support Needed?

- There is no “silver bullet”
  – i.e. no cloud offering is suitable for all circumstances
- Cloud offerings differ in many characteristics
  – It can all be very confusing to potential customers
- Mistakes in cloud offering selection can result in operational inefficiencies, unnecessary usage costs
- Uncertainty of operating costs reduces industry adoption
- Side-by-side comparison helps in selection of a suitable cloud offering for the consumer’s needs
- Existing tools: only for single provider’s offerings or only static view of consumer’s needs

Why Is It Difficult?

- The main challenge is the huge diversity of cloud offerings and their pricing models
  – Very difficult to compare cloud offerings across providers
- Different technical and other factors impact cloud offering selection and run-time cloud costs
  – Example: legal requirements
  – Some required information might not be available
- Consumers have different resource needs over time
- Costs incurred in the migration process must also be taken into account

NICTA’s Cloud Cost Estimation Tool – The Main Requirements

i. Support multiple cloud providers
ii. Abstract the differences among providers
iii. Simplify the input and focus on dominant cost factors
iv. Support the notion of time-series workload

Additionally, NICTA developed the Cloud Migration Point (CMP) methodology for estimating size, effort, and costs of migrating legacy applications to clouds.
Uniform Monitoring of Diverse Cloud Systems In-House and on the Internet
- Why Is It Needed and Difficult?
- NICTA's Generic Monitoring Engine for Hybrid Clouds

Why Is Monitoring Needed?
• “How can I estimate resource needs of my application and its time-series workload?”
  a) Analytical methods and simulations (e.g. NICTA’s ePASA tools for performance analysis of service-oriented systems)
  b) Monitoring
  c) Combination of a) and b)
• “How can I know whether my actual workload and the system’s performance are as predicted?”
  – When there are changes in workload or in system’s performance => control (e.g. adaptation) actions should be performed

Why Is Unified and Flexible Monitoring of Cloud Systems Needed?
• Huge diversity of components that can participate in (hybrid) cloud systems causes huge diversity of management information (metrics, formats, etc.)
• Existing tools do not successfully collect, integrate, and analyze management information across different boundaries (e.g. in-house & public cloud)
• Selection of used (and unused) monitoring components should be flexible
• Extensibility is needed to easily add new monitoring components (e.g. as plug-ins)

Why Is It Difficult?
• The main challenge is (again) the huge diversity of cloud systems and management information in them
• New types of metrics relevant for cloud management
  – Example: measures of elasticity
  – Example: delay of achieving eventual consistency
• Complex relationships and mappings between metrics at different levels are difficult to determine
  – Example: How do changes of CPU speed affect elasticity?

Our Adaptive Cloud Technologies

Architecture of NICTA'sGeneric Monitoring Engine for Hybrid Clouds
Selecting Run-Time Adaptations that Optimize Business Metrics
- Why Is It Needed and Difficult?
- NICTA's MiniZnMASC Business-Driven Management Middleware

**Why Is It Needed? An Example**

![Diagram showing a private cloud, spikes in demand for App.C but your private cloud has no resources, and a public cloud solution to rent computing resources and replicate App. C to meet demand.]

- Dynamic reconfiguration of applications to use a public cloud when a private cloud cannot provide enough computing resources.

**Business Value Metrics**

- What is important for customers/users?  
  - Technical metrics (e.g., availability)? Not really…  
  - Impact on their business? Yes!

- Business value metric = any measure of business worth
  - financial: income, cost, profit, ROI, ...
  - non-financial: number of customers, customer satisfaction, market share, ...
  - Related terms: KPI (Key Performance Indicator), …

**Why Is It Difficult?**

- Problem: complex mappings between technical metrics and business value metrics
  - e.g., better availability vs. better profit
- Business value metrics are diverse!

- Which business value metrics to maximize depends on business strategy
  - e.g., “exceptional customer satisfaction” vs. “cost leadership”

**Business-Driven IT Management (BDIM)**

- Determining mappings between technical and business metrics
- Using them for IT system management decisions
- Impact analysis of changes in business/technology

- Leading researchers: A. Moura (Brazil), J. Sauve (Brazil), C. Bartolini (USA), and others

- Related ideas from other communities:
  - Value-based software engineering (VBSE)
  - IT governance

**MiniZnMASC Middleware**

- Enables business-driven decision-making for control/adaptation of IT systems
  - “Which adaptation to choose in particular circumstances to optimize business metrics?!”
- Uses policies specified in our WS-Policy4MASC
  - extension of the WS-Policy industrial standard by W3C
- Generic in nature, but specific extensions for SOAP and RESTful Web services and now cloud systems
  - Evaluated for: feasibility, functional correctness, business benefits, performance overhead, and scalability
WS-Policy4MASC: New Policy Assertions

• Specify details necessary for run-time management

WS-Policy4MASC and MiniZnMASC Extensions for Cloud Management

• Management of cloud system usage has to deal with some unique events and actions
• We listed and classified events and actions supported by major cloud computing systems
• Extended WS-Policy4MASC with specification of common cloud-specific events and actions
• Implemented MiniZnMASC extension for processing monitored data and events from Amazon EC2 CloudWatch and invoking actions in Amazon EC2
  – Several other extensions (e.g. Rackspace) in progress
• Evaluated on experiments with the prototype

Integrating MiniZnMASC with Other NICTA’s Cloud Management Tools

• Integration with the Cloud Cost Estimation Tool
  • Cost estimations are translated into WS-Policy4MASC utility policy assertions
  • Using MiniZnMASC to determine optimal cloud usage in complex situations (many apps, business constraints, etc.)
  • MiniZnMASC outputs presented as recommendations to Cloud Cost Estimation Tool users
  
• Integration with the Generic Monitoring Engine for Hybrid Clouds
  • MiniZnMASC gets greater range of monitored metrics and events

Conclusions and Some Areas with Open Research Challenges

- Metrics and Their Mappings
- Management of Hybrid Clouds
- Management of “Mobile + Cloud” Systems
Conclusions

- Cloud systems are very diverse/heterogeneous
- No cloud system can serve all needs, so selection of the appropriate cloud offering with lowest cost is an important (configuration) management task
- No monitoring approach or tool is suitable for all possible components in hybrid clouds, so a plug-in based integration framework is needed
- System (incl. cloud) management decisions should be done to optimize business metrics (e.g. usage costs) that customers/users care about

Metrics and Their Mappings

- Further analyze new metrics important in clouds
  - Elasticity, delay of achieving eventual consistency, etc.
- **Mappings between metrics are crucial for analyses**
- Study various mappings between relevant metrics (particularly across different levels) and develop appropriate impact estimation models
  - Example: mappings between network QoS metrics and consumer's perceived cloud QoS metrics
  - Example: mappings between network metrics and cloud usage business metrics (cost, customer satisfaction, etc.)
  - ...