A Cloud Infrastructure Service Recommendation Technique for Optimizing Real-time QoS Provisioning Constraints

By Miranda Zhang Supervisors: Rajiv Ranjan & Peter Strazdins

Provider	Compute		Storage	Plans other than Pay As You Go		Trail
	Terminology	Unit	Terminology	Compute	Storage	Period or Value
Windows Azure	Virtual Server	/hr	Azure Storage	Commitment Plan, Member Offer		90 day
Amazon	EC2 Instance	/hr	S3	Reserved, Spot, Marketplace	Reduced Redundency	1 year
GoGrid	Cloud Servers	/RAM hr	Cloud Storage	Prepaid (1 ,6 or 12 month)		Various from time to time, current value: 100 AUD
RackSpace	Cloud Servers	/RAM hr	Cloud Files	Managed Cloud		
Nirvanix			CSN			
Ninefold	Virtual Server	/hr	Cloud Storage	SimplePlan		50 AUD
SoftLayer	Cloud Servers	/hr	Object Storage	Monthly		1 month
AT and T Synaptic	Compute as a Service	vCPU per hour + /RAM hr	Storage as a Service	Committed Allocation Pool		
Cloudcentral	Cloud Servers	/hr				

Table 1: Depiction of configuration heterogeneities in compute and storage services across providers. (Red) Blank cells in the table mean it is not available. Some providers offer their services under a different pricing scheme than pay-as-you-go.

Problem

There are over 426 of various compute and storage service providers with deployments in over 11,072 locations. Even within a particular provider there are different variations of the services. For example, just Amazon Web Service (AWS) has 674 different offerings differentiated by price, QoS features and location. Add to this the fact that every quarter they add about 4 new services, change business models (price and terms) and sometimes even add new locations.

To be able to select the best mix of service offering from an abundance of possibilities, application owners must simultaneously consider and optimize complex dependencies and heterogeneous sets of QoS criteria (price, features, location, etc.). For instance, it's not enough to just select optimal cloud storage service, corresponding computing capabilities (at where data is located, so no additional data transfer is needed) are essential to guarantee that one is able to process the data as fast as possible while minimizing the cost.

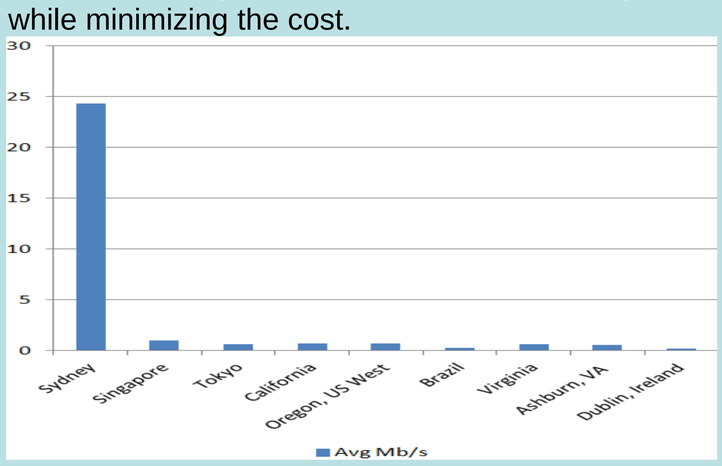


Figure 1: Download speed from Amazon data centers to Melbourne.

Figure 1 shows that geographically close data center has (as high as 25 times) better network performance, hence this validates the fact that location is one of the important criteria which should be considered during selection process.

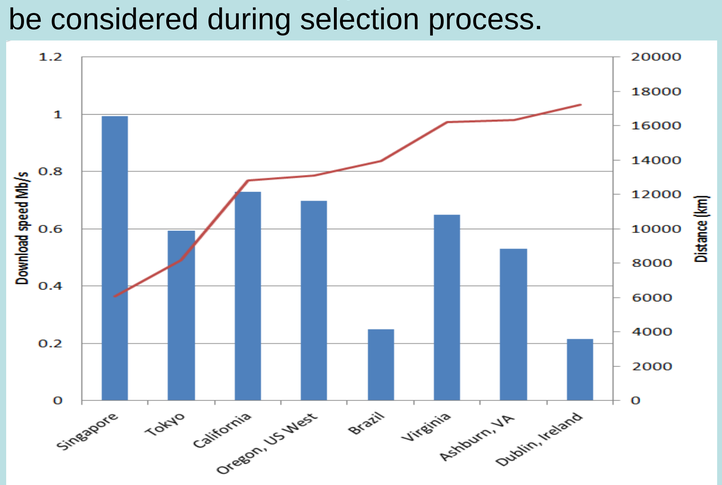


Figure 2: Download speed against distance from source.

Our measurements also indicate that distance is not the only factor that effects the network performance, as shown in Fig. 2, data centers are ordered from closest to furthest from left to right, Tokyo and Brazil clearly perform poorly than expected. Hence, we consider the need for active probing and profiling of network QoS from user's endpoint connection to the cloud data centers. By doing so we get clear picture of data centre's network QoS from the users' device that may be deployed across topologically distributed network locations.

Approach and Methodology

To address this hard challenge, in our previous work we developed a semi-automated, extensible, and ontology-based approach to infrastructure service discovery and selection based on only design time constraints (e.g., renting cost, datacentre location, service feature, etc.). Recently we have extend our approach to include the real-time (run-time) QoS (end-to-end message latency, end-to-end message throughput) in the decision making process.

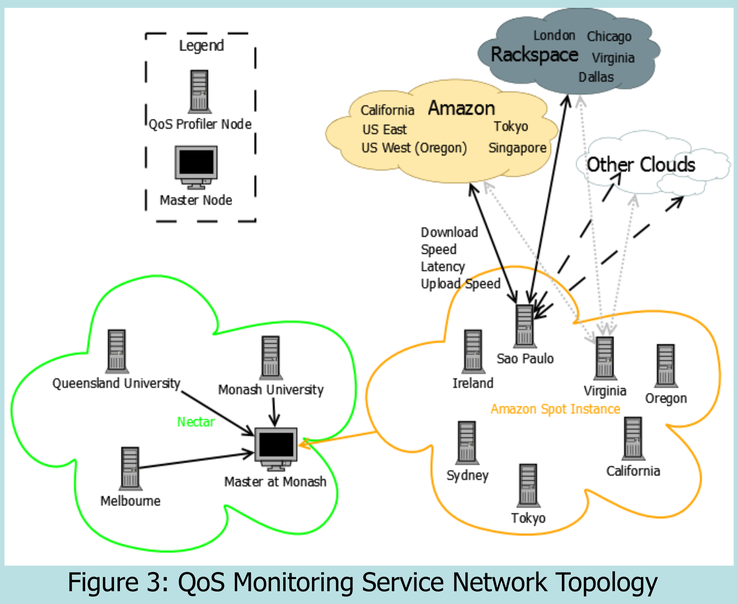


Figure 3: QoS Monitoring Service Network Topology

We have used 2 Clouds, shown in Fig. 3, namely: Nectar Research Cloud and Amazon Web Service. Since Nectar Cloud is free for researchers, we kept the instances running all the time, hence the decision to put master in Nectar. Because there is a limit of quota in Nectar and Amazon have greater geographical coverage in terms of datacenter locations. We use additional Spot instance from Amazon as slave data crawlers. A QoS Monitoring Node profiles Download Speed, Latency and Upload Speed at each datacenter in various Clouds from different locations.

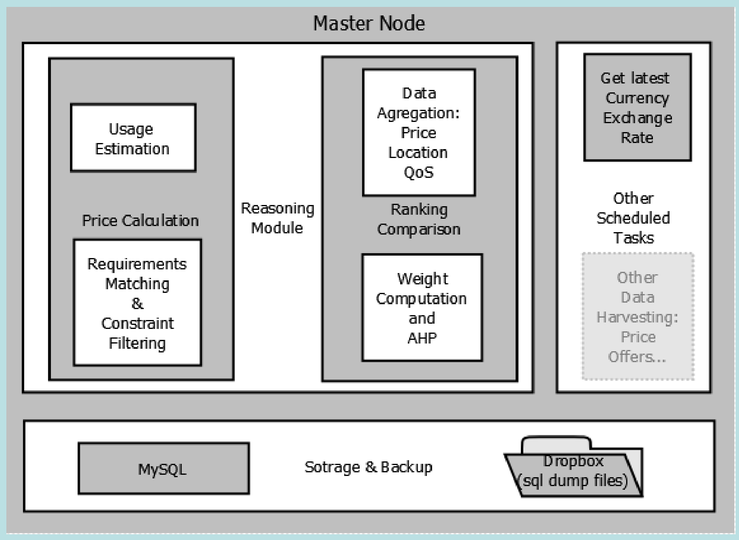


Figure 4: Master Node System Architecture.

As illustrated in Fig. 4, in the reasoning module main functions and operations are broke down into different blocks. There are some other tasks cannot be strictly categorized into existing modules, those are put into the "Other Tasks" section, and the very light grey block contains the evolving part of the system that's continually under development (being updated). The presentation layer (User Interface and Application Programming Interface implementation) and monitoring module are omitted to keep the diagram simple.

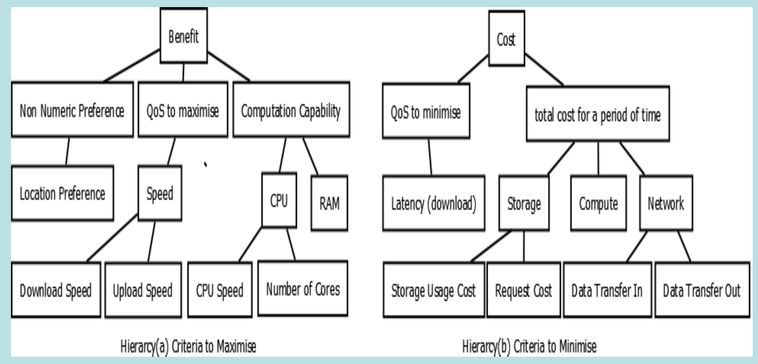


Figure 5: Criteria taken into consideration during comparison.

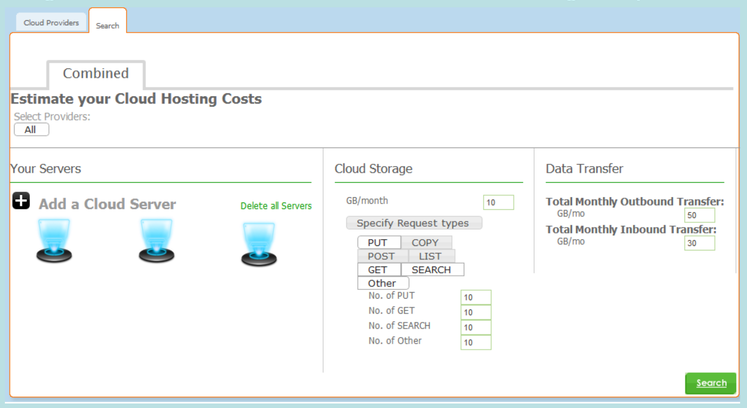


Figure 6: Cloud Recommender Selection Criteria User Interfaces.

Fig. 6 shows one of the user interfaces we provide while Fig. 7 illustrates the results page.

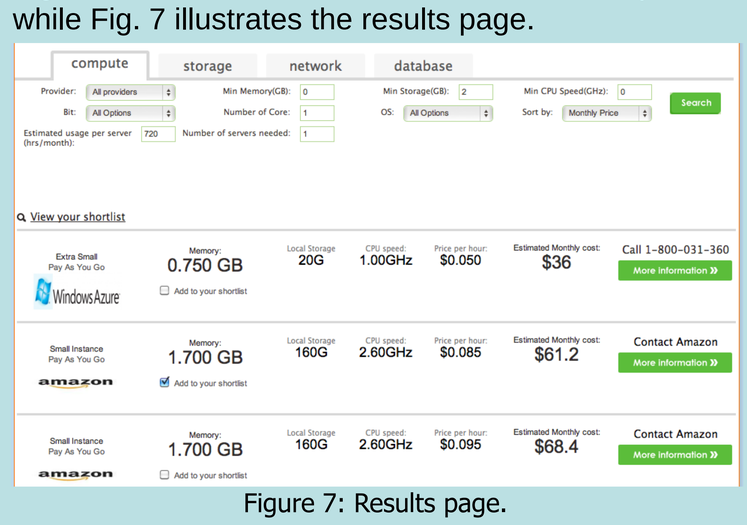


Figure 7: Results page.

We also provide a number of RESTfull APIs, like shown in Fig. 8 and 9, for the service to be called/consumed by other services/programs.

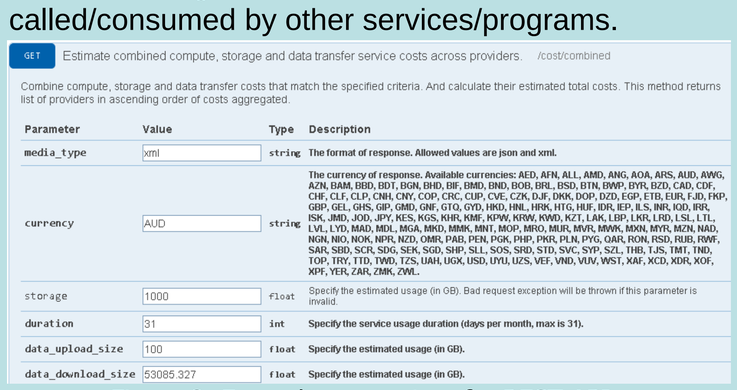


Figure 8: Example parameters for REST API .

Application

In the gaming industry, World of Warcraft counts over six million unique players on daily basis. The operating infrastructure of this Massively Multiplayer Online Role Playing Game (MMORPG) comprises more than 10,000 computers. Depending on the game, typical response times to ensure fluent play must remain below 100 milliseconds in online First Person Shooter (FPS) action games and below 1-2 seconds for Role-Playing Games (RPGs). A good game experience is critical for keeping the players engaged, and has an immediate consequence on the earnings and popularity of the game operators. Failing to deliver timely simulation updates leads to a degraded game experience and triggers player departure and account closures.

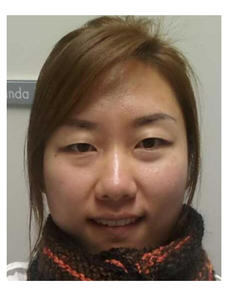
Startup gaming company with no existing infrastructure could launch a new game using public cloud infrastructure as cloud services offers the flexibility to scale on demand with no upfront investment. Rather than purchase and overprovision large number of servers in advance. Using cloud service, the game application services can be dynamically allocated or de-allocated according to demand fluctuations. Game companies can also better serve the diverse international users with the global presence of data centers owned by Cloud providers.

Australian National University

CSIRO

Further Information
contact: Miranda
email: miranda.zhang@anu.edu.au
phone: +61 401909897

Other Contacts
email: rajiv.ranjan@csiro.au
email: peter.strazdins@cs.anu.edu.au



Related Publication

M. Zhang, et al., "Investigating decision support techniques for automating Cloud service selection," in *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, 2012, pp. 759–764.