CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2013; **25**:1656–1674
Published online 30 May 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.2864

SPECIAL ISSUE PAPER

Applying reinforcement learning towards automating resource allocation and application scalability in the cloud

Enda Barrett*,†, Enda Howley and Jim Duggan

School of Engineering and Informatics, National University of Ireland Galway, Galway, Ireland

SUMMARY

Public Infrastructure as a Service (IaaS) clouds such as Amazon, GoGrid and Rackspace deliver computational resources by means of virtualisation technologies. These technologies allow multiple independent virtual machines to reside in apparent isolation on the same physical host. Dynamically scaling applications running on IaaS clouds can lead to varied and unpredictable results because of the performance interference effects associated with co-located virtual machines. Determining appropriate scaling policies in a dynamic non-stationary environment is non-trivial. One principle advantage exhibited by IaaS clouds over their traditional hosting counterparts is the ability to scale resources on-demand. However, a problem arises concerning resource allocation as to which resources should be added and removed when the underlying performance of the resource is in a constant state of flux. Decision theoretic frameworks such as Markov Decision Processes are particularly suited to decision making under uncertainty. By applying a temporal difference, reinforcement learning algorithm known as Q-learning, optimal scaling policies can be determined. Additionally, reinforcement learning techniques typically suffer from curse of dimensionality problems, where the state space grows exponentially with each additional state variable. To address this challenge, we also present a novel parallel Q-learning approach aimed at reducing the time taken to determine optimal policies whilst learning online. Copyright © 2012 John Wiley & Sons, Ltd.

Received 15 December 2011; Revised 3 April 2012; Accepted 3 May 2012

KEY WORDS: reinforcement learning; cloud computing; resource scaling

1. INTRODUCTION

Infrastructure as a Service (IaaS) clouds rely on economies of scale to deliver computational resources to consumers in a cost effective way. Sourcing computational resources from IaaS clouds eradicates the cost associated with maintaining the equivalent resources in-house. Similar to traditional utilities such as electricity and gas [1], consumers typically pay only for what they use, provisioning resources as needed in an on-demand fashion. This elasticity or ability to scale resources as required is one of the principle differences between computational clouds and previous utility computing forms such as computational grids and clusters, which require advanced reservations. In delivering resources to consumers, IaaS providers utilise virtualisation technologies such as Xen [2] and VmWare [3] to partition a single physical server into multiple independent Virtual Machines (VMs). These VMs reside in a co-located manner and have no visibility or control over the host environmental configuration or neighbouring VMs. Figure 1 demonstrates a typical cloud scenario where multiple VMs are co-located on a single physical host server. Depending on its configuration, each VM is allocated a portion of the physical host resource, that is, CPU cycles, RAM,

*Correspondence to: Enda Barrett, School of Engineering and Informatics, NUI Galway, Galway, Ireland.
†E-mail: Enda.Barrett@nuigalway.ie

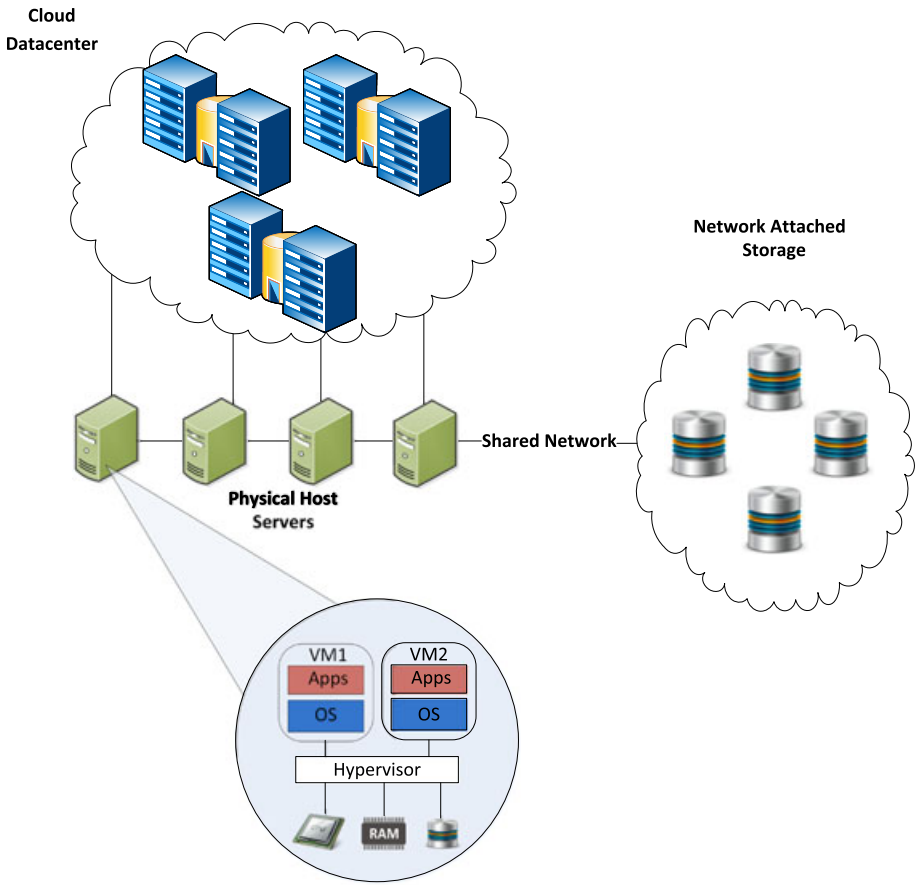


Figure 1. Physical host with three instantiated virtual machines.

Disc and Network bandwidth. A Virtual Machine Monitor (VMM) is installed on the physical host and is responsible for controlling VM access to the host's resources. The VMM attempts to isolate individual VMs with respect to security, failure and their respective environment, but not in respect of performance [4, 10]. Consequently, performance unpredictability has been identified as one of the key obstacles facing growth and greater adoption of cloud computing [5]. Much uncertainty exists in relation to how ported applications will perform, and indeed scale, once they are deployed in the cloud.

Dynamically scaling applications on large IaaS clouds in response to workload or performance changes presents a key challenge for resource planning techniques and application management. An effective scaling solution should allocate resources to optimise factors such as cost, performance and reliability. The solution itself should also be scalable, that is, capable of dealing with large workloads and complex resource allocation decisions. The current approach favoured for allocating resources to applications on the basis of fluctuating numbers of requests and application performance is to define threshold based policies [6, 7]. These are rule-based approaches, where upper and lower bounds are defined based on an observable metric such as application response time. With this approach, applications can scale up to meet demand at peak times and scale back once demand has subsided. Determining appropriate thresholds, however, requires expert domain and application knowledge and must often be redefined on the basis of application updates or workload changes. Because one cannot guarantee the performance of the underlying resource, naively adding similar resources to ensure compliance may not be an optimal strategy.

DOI: 10.1002/cpe

Recently, efforts have been made to develop more adaptive policies towards informing resource allocation decisions. Autoscaling policies should in essence be hyperopic, forgoing short-term gains in an effort to realise greater long-term benefits. Policies should also be adaptive to variations in the underlying resource performance and scale in the presence of new or unseen workloads combined with large numbers of resources. Significant work has focussed on decision theoretic planning techniques such as Markov Decision Processes (MDPs), combined with reinforcement learning techniques. The strength of these approaches is their ability to reason under uncertainty, which maps well onto the stochastic cloud environment. However, there are a number of issues that have not been satisfactorily answered by existing research. One of the major drawbacks associated with reinforcement learning techniques when it comes to solving large real world problems is the length of time it takes to converge to optimal or near optimal policies. In a dynamic scalability context, this is the time that it takes the learning agent to determine an optimal policy for the given environment. One approach aimed at addressing this problem is to develop a hybrid [8] mechanism in which the learning approach is trained using a good external policy, which potentially could be computed offline using sample data. The problem with this approach is that there are still challenges involved in determining a good initial policy. In addressing these challenges, this paper proposes a novel mechanism that takes advantage of the inherent parallelism associated with distributed computational platforms such as computational clouds. The approach involves agents learning in parallel on the same autoscaling task and sharing information regarding their experiences. This serves two functions, firstly, it decreases the length of time it takes agents to determine optimal resource allocations to support application scaling. Secondly, the approach is scalable as the number of resources grows because of the increasing numbers of learners as a function of the number of resources. Finally, to facilitate learning in computational clouds, we also devise a novel state action space formalism that is capable of learning optimal policies in computational clouds.

The principle contributions of this paper are the design and testing of:

- **Variable workload and performance model**: The development of a model based Q-learning approach that defines a novel state action space formalism capable of determining optimal resource allocation policies in a realistic cloud setting. Uniquely, the output policy reasons across both the variable workload model and the underlying resource performance model.
- **Parallel reinforcement learning**: A parallel reinforcement learning architecture that successfully parallelises Q-learning to speed up convergence rates of agents attempting to auto-scale resources in parallel.

The rest of this paper is structured as follows: Section 2 explains the causes of performance variability and details our results from microbenchmarking different instance types on Amazon's EC2. Section 3 provides an overview of relevant and related work in this field. Section 4 details MDPs, the reinforcement learning framework and the parallel reinforcement learning approach. Section 5 specified our auto-scaling model for both single agent and parallel Q-learning. Section 6 details our experimental findings, leading finally to *Conclusions and Future Work*.

2. CLOUD PERFORMANCE ANALYSIS

The current resource delivery mechanism favoured by IaaS clouds has been largely based on virtualisation technologies. Virtualisation allows for multiple VMs containing disparate or similar guest operating systems to be deployed in apparent isolation on the same physical host. This multi-tenant environment where agents share and compete for resources on the same host can lead to substantial variability. From the application's performance perspective, a variable underlying supportive resource will cause fluctuations in its performance. This section benchmarks a number of IaaS instances on Amazon EC2 to highlight these issues.

2.1. Xen hypervisor

The sharing of resources amongst the respective VMs is handled by the VMM [2], an independent domain level monitor that has direct access to the underlying hardware. Xen is a popular open

DOI: 10.1002/cpe

Table I. Instance types and costs for US-East Virginia.

Instance type	Memory	ECUs	Disc	Cost (per/hr)	I/O Performance
m1.small	1.7 GB	1	160 GB	$0.085	Moderate
c1.medium	1.7 GB	5	350 GB	$0.17	Moderate
m1.large	7.5 GB	4	850 GB	$0.34	High

source virtualisation framework, supporting a wide range of guest operating systems and is used by a large number of cloud providers including Amazon Web Services. Xen facilitates a software layer known as the Xen hypervisor that is inserted between the server's hardware and the operating system. This allows the physical host to deploy multiple VMs in isolation, decoupling the operating system from the physical host. However, whilst virtualisation technologies such as Xen provide excellent security, fault and environmental isolation, they do not ensure performance isolation. Koh *et al*.[4] state that there are three principle causes of this interference. The first cause is because each independent VM on the hypervisor has its own resource scheduler, which is attempting to manage shared resources without the visibility of others. Secondly, guest operating systems and applications inside a VM have no knowledge about ongoing work in co-located VMs and are unable to adapt in response. Thirdly, some hypervisors such as the Xen hypervisor offload operations such as I/O operations to service VMs. This particularly affects I/O-intensive applications as the Xen hypervisor forces all I/O operations to pass through a special device driver domain and this forces the context to switch into and out of the device driver domain causing interference.

In general, the greater the activity of neighbouring VMs, the greater the potential for interference, which directly results in variable application performance. When booting up instances in the cloud, an auto-scaling controller will not be able to control the type of VM it is co-located with, or its activity.

Table I displays three different VM instances and their associated properties currently supported by Amazon EC2. In addition to the performance interference as a result of virtualisation, the type of instance allocated to the application will impact on performance. Amazon rate the I/O performance of the respective instances as *Low*, *Moderate* and *High*, which means the VMs with *High* should receive a greater amount of dedicated I/O bandwidth. As newer hardware is added to the data centre replacing older models, the physical host your VM resides on could be a determinant in performance also.

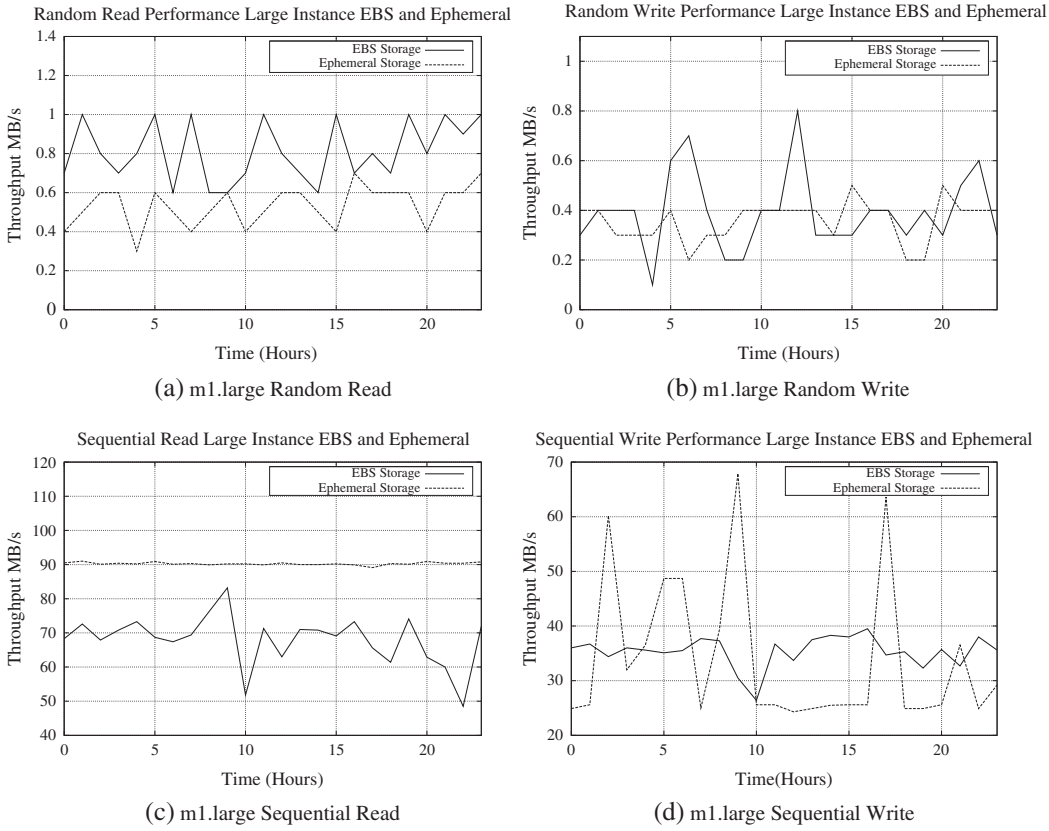
2.2. Microbenchmarking EC2

As previously discussed, the nature of shared virtualisation, leads towards performance unpredictability as the underlying CPU, RAM and disc are shared amongst the VMs residing on the physical host. One aspect of the computational resource that is particularly sensitive to interference on virtualised platforms is network and disc I/O. To evaluate this we undertook a series of microbenchmarks on a number of EC2 instances, which demonstrate the disc I/O performance variability exhibited by storage volumes on IaaS clouds. We used the filebench[‡] benchmarking utility to demonstrate sequential/random read/write performance. Filebench is a multi-threaded, cross-platform benchmarking tool, capable of running an array of synthetic workloads designed to evaluate disc I/O analysis. The results displayed here highlight the variability deployed applications observed in relation to I/O performance. Further analyses of performance [9] and performance interference [10] on I/O workloads have previously been published.

Each VM instance on Amazon EC2 can support two different types of storage volume. Instance based or ephemeral storage is a storage volume located within the physical host and shared amongst the resident VMs. Elastic Block Storage (EBS) volumes are network attached storage devices that are connected via a 1-Gbps Ethernet connection. We evaluate both of these in terms of performance.

[‡]http://sourceforge.net/projects/filebench/

DOI: 10.1002/cpe



(a) m1.large Random Read

(b) m1.large Random Write

(c) m1.large Sequential Read

(d) m1.large Sequential Write

Figure 2. Sequential read/write and random read/write performance variability observed for 2 m1.large instances running on Amazon EC2.

In testing disc I/O performance on EC2, we selected four workload profiles to evaluate the random read/write and sequential read/write performance for an m1.small, c1.medium and m1.large instance on the Amazon EC2 cloud. For each experiment, we created two instances of each type in a separate availability zone in the US-East Virginia data centre. The experiments began on a Tuesday evening at 19:00 Coordinated Universal Time (UTC) and ran for a total of 24 h. We chose a midweek starting point in order to emulate as closely as possible the variability that would occur during an average working week. The total number of experiments ran over the 24 h is 1152. The following four workload profiles were designed to evaluate each instance type:

- **Sequential Read**. This workload profile evaluates large sequential file reads. For this experiment the file size was set to be larger than the allocated RAM, which is 6 GB in the case of the m1.small/c1.medium instances and 10 GB in the case of the m1.large instance. This eliminated possible interference due to memory caching. To obtain a true reflection of the underlying performance, caching was disabled, the iosize was set to 1 MB and single threaded execution. The experiment was ran for 20 min, for both EBS and Ephemeral storage each hour, allowing for a steady state evaluation of performance.
- **Sequential Write** This workload profile evaluates large sequential file writes. The individual file write sizes were set to 1 MB each. Caching was again disabled, with syncing enabled. This was executed single threaded. The file sizes generated through writing were 6 GB for m1.small,c1.medium and 10 GB for m1.large.
- **Random Read** This workload profile evaluates random file read performance. Caching was disabled, individual read sizes (iosize) were set to 2K, with single threading. Each run generated 128 MB of read data.

DOI: 10.1002/cpe

Table II. Summary.

Profile	Instance type	Average EBS (MB/s)	Average Ephemeral (MB/s)
Sequential Read	m1.small	78.22	62.01
Sequential Write	m1.small	17.48	30.19
Random Read	m1.small	0.27	0.3
Random Write	m1.small	0.2	0.19
Sequential Read	c1.medium	74.04	107.81
Sequential Write	c1.medium	32.63	29.09
Random Read	c1.medium	0.57	0.3
Random Write	c1.medium	0.2	0.2
Sequential Read	m1.large	60.02	90.26
Sequential Write	m1.large	35.38	33.98
Random Read	m1.large	0.81	0.53
Random Write	m1.large	0.36	0.39

EBS, Elastic Block Storage.

- **Random Write** This workload profile evaluates random write performance. Caching was disabled, synchronisation was enabled. The file sizes were set to be larger than the available RAM at 6 GB for m1.small, c1.medium instances and 10 GB for m1.large instances.

Figure 2 illustrates the sequential/random read/write performance of the m1.large instance for both the EBS and Ephemeral storage volumes. It clearly demonstrates hourly deviations in the total throughput exhibited in both read and write performance. Table II details a tabular breakdown of all the instances tested. The average throughput in both EBS and Ephemeral storage is displayed. According to Amazon, both the small and medium instances should achieve a *Moderate* I/O performance, whereas the large should be *High*. However, the results clearly show high variability across all the instances with the small and medium instances outperforming the large instance for the *Sequential Read* workload profile. This was most surprising as theoretically the large instance should have had a greater share of local disc bandwidth and thus should have had a much higher performance on the ephemeral storage across all the workload profiles. In relation to EBS storage, the large instance had a superior read and write performance over the small and medium instance in all profiles except the *Sequential Read* profile. Interestingly, EBS volumes also suffer from multi-tenancy issues, as the storage volumes are also virtualised over many devices. In addition, the EBS volumes are also bound by network bandwidth and can suffer from network related problems such as congestion. With EBS, the user also has no control over the location of the storage volume or positioning relative to other tenants. The results clearly demonstrate the high degree of uncertainty present in relation to I/O performance in the cloud, even on an hourly granularity. Cloud providers such as Amazon do provide Service Level Agreements regarding resource availability, but they do not provide any QoS guarantees on performance. In order to handle this variability, an auto-scaling technique must be capable of reasoning across changing workloads and resource performance. It should also be dynamic to fluctuations in real time and capable of handling scenarios. it has no prior experience of.

3. BACKGROUND RESEARCH

Threshold based policies are one of the most widely used mechanisms for the auto scaling of applications in the cloud, both from a commercial and research perspective. Our background research examines these with respect to the most relevant in application scalability in computational clouds. We also examine reinforcement learning approaches to resource allocation and application support to contextualise the contributions of this paper over previous automated control learning approaches.

3.1. Dynamically scaling applications on clouds : threshold based approaches

Public computational clouds such as Amazon EC2 [11] provide commercial auto scaling solutions to facilitate resource allocation on the basis of predefined metrics. Amazon's Auto Scaling [12]

DOI: 10.1002/cpe

software in conjunction with their proprietary CloudWatch [13] monitoring system can be used to control automated resource allocation decisions on the basis of pre-defined metrics. These metrics stipulate decision points at which an action is triggered, that is, to add or remove a particular instance type. RightScale [14] similarly allows for the definition of process load metrics that trigger allocation responses once a given metric has been breached. Rightscale facilitates the allocation of resources to applications running across multiple clouds. With this approach, a mapping must exist denoting the representative action that must be executed once a specific threshold has been breached. These rule based systems are generally called threshold based approaches in the literature. Threshold based policies such as those employed by RightScale and Amazon's Auto Scaling, tend to focus on scaling at the machine or VM level. They do not facilitate the definition of higher business functions and objectives when scaling a given service or application running on the cloud. Instead, the application's resources are altered through allocation and deallocation of VMs. Rodero-Merino *et al*. [15] proposes a dynamically scaling solution aimed at addressing this issue. Their proposed solution operates at a higher level of abstraction than those offered currently by IaaS providers. It decomposes high level objectives specified in a Service Description File, which also contains the scalability rules defined by the Service Provider. The paper examines three different scalability mechanisms that facilitate a holistic approach to service management on federated clouds. Their developed application layer is called Claudia, and it additionally employs a Virtual Infrastructure Management solution that avoids vendor lock-in and can interface between different clouds. Moran *et al*. argued that the mechanisms employed by Claudia were not expressive enough to enable a fine grained control of the service at runtime [16]. The scalability rules defined are specified in an ad hoc manner and are not designed with generality. To interchange the abstract level languages used to specify applications behaviour in clouds, they propose the usage of the Rule Interchange Format, in conjunction with the Production Rule Dialect. This generates an appropriate mapping to industry rule engines such as JBoss Drools or Java Jess. These approaches improve upon the industrial led approaches offered by Amazon and Rightscale in attempt to auto scale applications in a more holistic manner, but they still require a certain amount of domain knowledge, with rules and conditions required to be defined for different environmental states. Planning in advance the appropriate corrective action can prove extremely difficult especially when one has to consider a large number of possible states [17, 18].

Threshold based approaches have also been developed towards elastic storage solutions. Lim *et al*. developed an automated controller for elastic storage in a cloud computing IaaS environment on the basis of proportional thresholding [19]. The controller was broken down into three components; a Horizontal Scale Controller for adding and removing nodes; a Data Rebalance Controller, for controlling data transfers; and a State Machine to coordinate the actions. This approach demonstrated speedy adaption to fast changing events; however, this is not always optimal given that a particular event may be very short lived. Also, in the absence of a formalised state space, it lacks the predictive power to respond to highly varying workloads.

The benefits of reinforcement learning methods is their ability to reason under uncertainty based only on environmental observations. A modification to the application or change in the workload request model would possibly require a model change for the threshold based approach. The reinforcement learning approach will adapt to suit the environment on the basis of its own experience.

3.2. Auto-scaling resources : decision theoretic approaches

Tesauro investigated the use of a hybrid reinforcement learning technique for autonomic resource allocation [8]. He applied this research to optimising server allocation in data centres, where homogenous application servers were added and removed based on a hybrid reinforcement learning technique. This work demonstrated the learning approach's capability to maintain sufficient response times, governed by a Service Level Agreement, across a number of applications. David Vengerov combined reinforcement learning with a fuzzy rulebase to allocate CPU and memory from a common resource pool to multiple entities [20]. This work focussed on the problem of distributing resources from a common resource pool to multiple entities, such as in a grid or data centre.

DOI: 10.1002/cpe

J. Perez *et al*. [21] applied reinforcement learning to optimise resource allocation in Grid computing. This work focused on optimising the utility of both the users submitting jobs and the institutions providing the resources through Virtual Organisations. Virtual Organisations consist of a set of individuals or organisations that share resources and data in a conditional and controlled manner[22]. Galstyan *et al*. [23] implemented a decentralised multi-agent reinforcement learning approach to Grid resource allocation. With incomplete global knowledge, and no agent communication, the authors showed that the reinforcement learning approach was capable of improving the performance of large scale grid.

More recently [24], a Q-learning approach was developed for allocating resources to applications in the cloud. This work developed an automated controller capable of adding and removing VMs on the basis of a variable workload model. The author presented an adaptive approach capable of keeping up with a variable workload model driven by a sinusoidal function. Using convergence speedups, Q function initialisation and model change detection mechanisms, the author was able to fine tune the approach. Rao *et al*. [25] developed a reinforcement learning approach to VM autoconfiguration in clouds. In response to changes in demand for applications, the VM itself is reconfigured. The approach was able to determine near optimal solutions in small scale systems.

The key difference between our approach and these works is that they focus on determining policies for allocating resources to match a variable workload or user request model. Our approach supports multiple criteria in that the outputted policy considers both the variable workload and the underlying performance model. The approach also facilitates learning across geographical regions, where it is capable of reasoning about the temporal performance variances associated with the cloud. In addition to improve the time taken to approximate optimal or near optimal policies, we have devised a parallel learning approach. This is first time a parallelised reinforcement learning approach has been applied in this context. Previous approaches to reducing the state space size and improving convergence times involve hybrid [8] learning approaches and utilising function approximation [20] techniques.

4. REINFORCEMENT LEARNING - THEORETICAL FOUNDATIONS

Reinforcement learning has been applied successfully across a range of domains supporting the automated control and allocation of resources [26–29]. It operates on the basic premise of punishment and reward, with agents biased towards actions that yield the greatest utility. Much of reinforcement learning theory is based on determining optimal policies for MDPs.

4.1. Markov Decision Processes

Reinforcement learning problems can generally be modelled using MDPs. In fact, reinforcement learning methods facilitate solutions to MDPs in the absence of a complete environmental model. This is particularly useful when dealing with real world problems as the model can often be unknown or difficult to approximate.

Markov Decision Processes are a particular mathematical framework suited to modelling decision making under uncertainty. A MDP can typically be represented as a four tuple consisting of states, actions, transition probabilities and rewards.

- S, represents the environmental state space;
- A, represents the total action space;
- $p(.|s,a)$, defines a probability distribution governing state transitions $s_{t+1} \sim p(.|s_t, a_t)$;
- $q(.|s,a)$, defines a probability distribution governing the rewards received $R(s_t, a_t) \sim q(.|s_t, a_t)$;

S, the set of all possible states represents the agent's observable world. At the end of each time period t, the agent occupies state $s_t \in S$. The agent must then choose an action $a_t \in A(s_t)$, where $A(s_t)$ is the set of all possible actions within state s_t. The execution of the chosen action results in a state transition to s_{t+1} and an immediate numerical reward $R(s_t, a_t)$. Equation 1 represents the

DOI: 10.1002/cpe

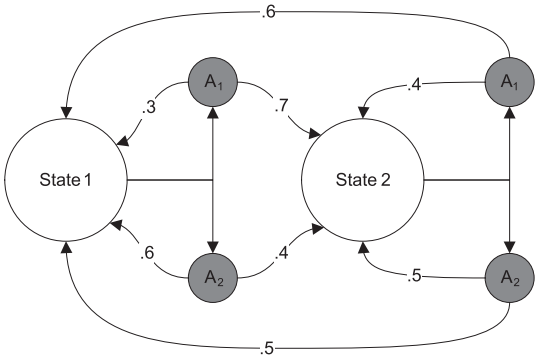


Figure 3. Markov Decision Process with two states and two actions.

reward function defining the environmental distribution of rewards. The learning agent's objective is to optimise its expected long-term discounted reward.

$$R^a_{s,s'} = E\left\{r_{t+1}|s_t = s, a_t = a, s_{t+1} = s'\right\} \tag{1}$$

The state transition probability $p(s_{t+1}|s_t, a_t)$ governs the likelihood that the agent will transition to state s_{t+1} as a result of choosing a_t in s_t.

$$P^a_{s,s'} = Pr\left\{s_{t+1} = s'|s_t = s, a_t = a\right\} \tag{2}$$

The numerical reward received upon arrival at the next state is governed by a probability distribution $q(s_{t+1}|s_t, a_t)$ and is indicative as to the benefit of choosing a_t whilst in s_t. To illustrate the workings as simple MDP, Figure 3 depicts a simple two state, two action MDP. In Figure 3, choosing action A_1 in *State* 1 will lead you to *State* 2 with a transition probability of 0.7 and back to *State* 1 with a transition probability of 0.3. Choosing A_2 will lead you to *State* 2 with a transition probability of 0.4 and back to *State* 1 with a transition probability of 0.6. An agent currently in *State* 1 wishing for transition to *State* 2 has a greater probability of doing, so should they choose action A_1.

In the specific case where a complete environmental model is known, that is, (S, A, p, q) are fully observable, the problem reduces to a planning problem [30] and can be solved using traditional dynamic programming techniques such as value iteration. However, if there is no complete model available, then one must either attempt to approximate the missing model (model-based reinforcement learning) or directly estimate the value function or policy (model free reinforcement learning).

4.2. Q-learning

In the absence of a complete environmental model, model free reinforcement learning algorithms such as Q-learning [31] can be used to generate optimal policies. Q-learning belongs to a collection of algorithms called Temporal Difference (TD) methods. Not requiring a complete model of the environment, TD methods possess a significant advantage. TD methods have the capability of being able to make predictions incrementally by bootstrapping the current estimate onto previous estimates.

The update rule for Q-learning is defined as

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma Q\left(s', a'\right) - Q(s,a)\right] \tag{3}$$

and calculated each time a state is reached, which is non-terminal. Approximations of $Q^\pi(s,a)$, which are indicative as to the benefit of taking action a while in state s, are calculated after each time interval. Actions are chosen based on π, the policy being followed. In this research, we use an ϵ-greedy policy to decide what action to select whilst occupying a particular state. This means that the agents choose the action that presents them with the greatest amount of reward, most of the

DOI: 10.1002/cpe

time. Let $A'(s) \subseteq A(s)$, be the set of all non-greedy actions. The probability of selection for each non-greedy action is reduced to $\frac{\epsilon}{|A'(s)|}$, resulting in a probability of $1 - \epsilon$ for the greedy strategy.

Estimated action values of each state action pair $Q^{\pi}(s, a)$ are stored in lookup table form. The goal of the learning agent is to maximise its returns in the long run, often forgoing short-term gains in place of long-term benefits. By introducing a discount factor γ, $(0 < \gamma < 1)$, an agent's degree of myopia can be controlled. A value close to 1 for γ assigns a greater weight to future rewards, whereas a value close to 0 considers only the most recent rewards. This represents a key benefit of policies determined through reinforcement learning compared with threshold based policies. The reinforcement learning based approaches are capable of reasoning over multiple actions, choosing only those that yield the greatest cumulative reward over the entire duration of the episode. The steps involved in Q-learning are depicted by Algorithm 1.

Algorithm 1 Reinforcement Learning Algorithm (Q-learning)

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode)
Initialize s
repeat
Choose a from s using policy derived from Q (ϵ-greedy)
Take action a and observe r, s'
$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
$s \leftarrow s'$;
until s is terminal

Q-learning can often require significant experience within a given environment in order to learn a good policy. Although it is algorithmically straightforward to implement and can operate successfully in the absence of a complete environmental model, it does not make efficient use of the data that it gathers as a result of learning [32]. It can also take significant time to approximate the true value function Q^*. In an environment where computational resources are relatively cheap and gathering real world experience costly, an alternative approach is to parallelise the learning process amongst multiple independent learning agents.

4.3. Parallel reinforcement learning

A learning agent can speed up the time it takes to learn an approximate model of the environment if it does not have to visit every state and action in the given environment. If instead it could learn the value of states it had not previously visited from neighbouring agents, then the time taken to approximate $Q*$ would be greatly reduced. Parallel learning approaches generally comprise one of the following two approaches. Agents learn individually operating on the same task or agents learn on a subset of the given task. Our approach is an example of the former, where all agents attempt to allocate resources to support the scaling of the same application type. Although the agents operate on the same learning task, they will all have different learning experiences because of the stochastic nature of the environment, that is, they will visit different states, choose different actions and observe different rewards. Previous work by Kretchmar [33] has demonstrated the convergence speedups made possible by applying a parallel reinforcement learning approach in a general setting. Each Q-learning agent independently maintains a local Q_l and global estimate Q_g of the approximate Q-values. Q_g is the agent's global representation of Q. It consists of the combined experience of all other learning agents exclusive of their own. This separation of personal experience from that of all the other agents facilitates a weighted aggregation of experience. In environments exhibiting a high degree of randomness, an agent may weight its own experience over that of the global experience. Q_g is calculated by aggregating the weighted sum of Q-value estimates of all other agents according to Equation 4.

DOI: 10.1002/cpe

$$Q(s,a) = \frac{Q(s,a)_l \times Expc_l + Q(s,a)_g \times Expc_g}{Expc_l + Expc_g} \quad (4)$$

The agent makes decisions based on Q(s,a) the weighted aggregation of the local and global estimates of Q. Algorithm 2 depicts the steps involved in our parallel learning approach. Firstly, both the local $Q(s,a)_l$ and global $Q(s,a)_g$ value estimates are initialised to 0. This is an optimistic initialisation and encourages exploration in the early stages of learning. The communications arrays $comms_{\text{in}}$ and $comms_{\text{out}}$ are initially set to $\emptyset$. For all states visited, the agent chooses an action a using an ϵ-greedy policy π with respect to Q the combined global and local estimate. This policy ensures that not all the agent's actions are greedy with respect to Q. Sometimes, the agent will choose to act randomly, this balances the tradeoff between exploration and exploitation. A high value of epsilon will bias the agent's decisions towards exploration and a low value allowing the agent to exploit its current knowledge. On the basis of the policy, the agent executes the action a, observes the reward r and next state s'. The agent then updates its estimate of $Q(s,a)_l$ in accordance with Equation 3. If the difference between the Q value estimates are greater than a predefined threshold θ, then agent's local estimate is added to the outgoing communications array $comms_{\text{out}}$. This information is then transmitted to all other learning agents. Initially, quite a lot of data are transmitted between agents, but as the local estimates converge to the global estimates the agents do not transmit anymore information.

Algorithm 2 Parallel Q-Learning

Initialise $Q(s,a)_l = 0$, $Q(s,a)_g = 0$, $Q(s,a) = 0$
$comms_{out}$, $comms_{in} \leftarrow \emptyset$
$\pi \leftarrow$ an arbitrary ϵ-greedy policy w.r.t to Q

repeat
  for all $s \in S$ **do**
    Choose a from s using policy π
    Take action a, observe r, s'
    $Q(s,a)_l \leftarrow Q(s,a)_l + \alpha[r + \gamma \max_{a'} Q(s',a')_l - Q(s,a)_l]$
    $s \leftarrow s'$;
    if $\| (Q(s,a)_l - Q(s,a)_g) \| > \theta$ **then**
      Add $Q(s,a)_l$ to $comms_{out}$
    end if
    Transmit $comms_{out}$ to all agents
    Receive $comms_{in}$ from other agents
    if $comms_{in} \neq \emptyset$ **then**
      for all $Q(s,a)_g \in comms_{in}$ **do**
        $Q(s,a) = \frac{Q(s,a) \times Expc_l + Q(s,a)_g \times Expc_g}{Expc_l + Expc_g}$
      end for
    end if
  end for
until s is terminal

5. MODEL FOR THE AUTO-SCALING OF APPLICATIONS IN CLOUDS

To facilitate agent learning for a cloud resource allocation problem, one must define an appropriate state action space formalism. The revised state action space formalism is designed specifically for obtaining better policies within computational clouds. Our state space representation is tailored to suit the performance related variabilities and the geographical distribution of resources. We define the state space S as the conjunction of three state variables $S = \{u, v, time\}$.

DOI: 10.1002/cpe

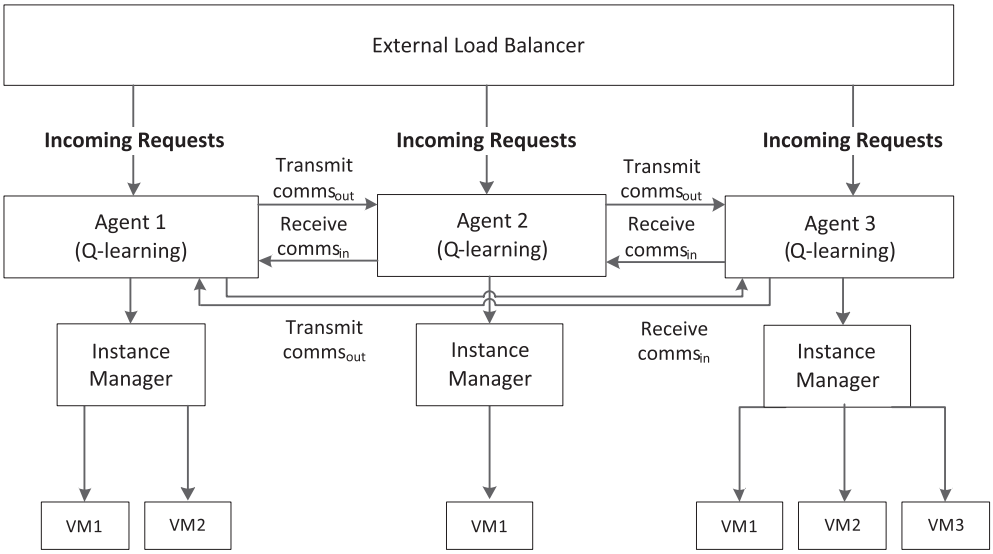


Figure 4. Parallel Q-learning architecture.

- u is the total number of user requests observed per time period. This value varies between time steps.
- v is the total number of VMs allocated to the application, where each VM instance $V_i \in \{\{t_1, l_1\}, \ldots \{t_n, l_m\}\}$. n represents the total number of VM types and m is the total number of geographic regions. t is the VM type and l is the region.
- $time$ is UTC time. It allows the agent to reason about possible performance related effects such as peak time of day in a data centre in a specific region.

The agents action set A contains the set of all possible actions within the current state. The agent can choose to add, remove or maintain the amount of VMs allocated to the application. Rewards are determined based on a defined Service Level Agreement (SLA), which is related to performance. The overall reward allocated per time step is given by the following equations.

$$C\ (a) = C_r \times V_a + \left\{ \sum_{i=1}^{v} (C_r \times V_i) \right\} \tag{5}$$

$$H\ (a) = P_c \times \begin{cases} (1 + \frac{p' - sla}{sla} & if \quad p' > SLA \\ 0 & \text{else} \end{cases} \tag{6}$$

$$R\left(s', a\right) = C\ (a) + H\ (a) \tag{7}$$

C_r is the cost of the resource, this is variable depending on the type, specific configuration and region. V represents an individual VM instance, with V_a representing the specific VM allocated, deallocated or maintained as a result of action a. H is the overall penalty applied as a result of violating the specified SLA. $P_c \in \Re$ represents a defined penalty constant incurred through violation of the SLA. The total reward $R(s', a)$ for choosing action a, resulting in s', is the combination of the cost of execution and any associated penalties. Although many more state observations could be included in this model (CPU utilisation, Memory utilisation and average response time), the approach works surprisingly well given relatively modest state information. In fact, previous allocation approaches have had comparable performance to heavily researched open-loop queuing theoretic models, using only current demand as the single observable state variable [34].

Reinforcement learning approaches generally suffer from so called *curse of dimensionality* problems, as the size of the state space grows exponentially with each new state variable added. This

DOI: 10.1002/cpe

limitation prevents reinforcement learning techniques from handling environments consisting of very large state and action spaces. To prove the viability of using reinforcement learning to handle application scalability in clouds in lieu of potentially large state and action spaces, we have devised a learning architecture aimed at parallelising Q-learning in the context of auto-scaling resources.

Figure 4 presents a high-level architecture of parallel Q-learning in a typical cloud environment. Each agent makes its own decisions about the incoming user requests and experiences penalties and rewards independently. Each agent's environment is insular, this means that multiple independent agents do not introduce non-stationarity in each others environments as a direct result of learning in parallel. On the basis of the allocated numbers of requests, the agent must attempt to learn an approximate individually optimal policy. The agents then share information regarding their observations while operating in the environment. Each agent communicates directly with all the other agents in the system. Actions of whether to add, remove or maintain the existing amount of allocated VMs are executed by the instance manager based on the instructions of the learning agent.

6. EXPERIMENTAL RESULTS

In this section, we examine algorithmic performance from two different perspectives. Firstly, we investigate the performance of our proposed formalism in the presence of a variable underlying resource and workload model. Secondly, we evaluate our parallel reinforcement learning approach and examine its performance with respect to the time taken to converge to optimal policies.

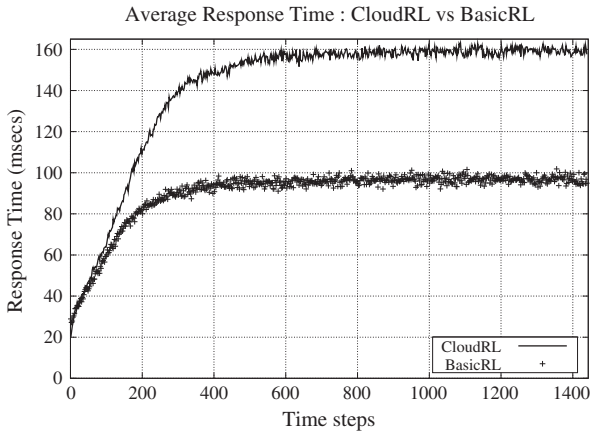
6.1. Experimental setup

We develop an experimental testbed in MATLAB (MathWorks, MA, USA) to evaluate our results. Unless stated in the individual experimental sections, the following parameters are applied across all experiments.

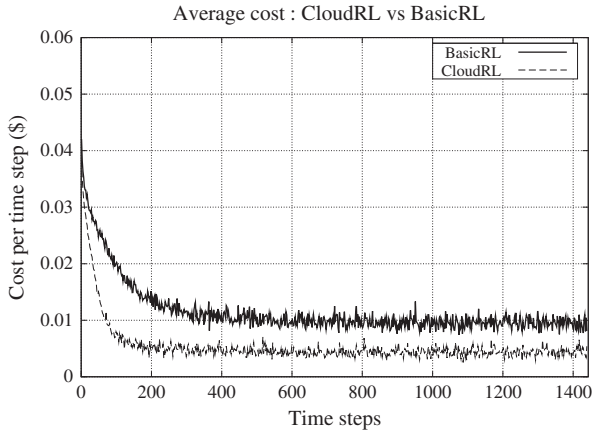
1. The user request models are generated by a workload simulator. The simulator simulates user requests in an open-loop mode. The open-loop mode generates Poisson requests with an adjustable mean arrival rate ranging from $10 - -150$ reqs/sec.
2. An SLA of 250 (msecs) governs the maximum allowed response time per request. Each request exceeding this value is deemed in violation of the SLA and incurs a penalty according to $P_c = 1$. The value of the penalty P_c has a direct impact on the distance, the policy maintains from the SLA.
3. Q-learning is initialised with the following parametric settings. A value of $\alpha = 0.5$ for the learning rate ensures that the majority of the error in the estimate is backed up. The discount factor $\gamma = 0.85$ discounts the value of future states. A value of $\epsilon = 0.1$ was chosen to facilitate adequate environmental exploration. The experimental analysis of Q-learning has the same parametric settings.
4. Four separate data centres are simulated in four disparate geographic regions, closely emulating the data centre regions supported by Amazon's EC2. Our simulations also emulate EC2's instance pricing model, where prices per VM varies between types and regions. In our experiments, we define the price of the VM as directly proportional to its configuration, in terms of CPU, memory and disc size, that is, the greater the size of the configuration, the greater the cost. Table I in Section 3 outlines the different instance types and their associated configurations.
5. A performance model distribution is constructed by discretising the observed benchmark results into time steps. This allows us to model variable performance over each time step. Taking the lowest and highest observed values per time step, a random performance sample is generated uniformly. We assign a specific peak time in each region when performance variability is increased across all the VMs instantiated within that region.

The agent's knowledge is stored in a lookup table (Q-table) and is used to inform decisions over the entire learning episode.

DOI: 10.1002/cpe



(a) Average Response Time : CloudRL vs BasicRL



(b) Average Costs : CloudRL vs BasicRL

Figure 5. Comparison between the CloudRL and the BasicRL approaches with respect to average costs and average response time.

6.2. Optimising for variable resource performance

This experiment analyses the proposed state action space formalism in contrast to previous work where agents ignore resource performance variability. VMs are instantiated with respect to type and location configurations. Learning intervals are discretised into time steps, with each lasting for 60 s. Each interval constitutes a decision point, where the agent chooses an action a is presented with a reward r and the next state s'. Throughout all the experiments, the agents action set is limited to the addition or substraction of a single VM each time. Individual VMs are configured in terms of CPU, memory and disc; however, the focus of this paper is on I/O variability. An I/O performance model for each VM type is generated based on the observed I/O performance, through the benchmarking of instances. The simulations carried out are based on data gathered through benchmarking live Amazon EC2 instances as outlined in Section 2.

To analyse the benefit of the state space formalism presented in this paper over previous work when dealing with the variabilities of the cloud, the performance of two Q-learning approaches is evaluated. The first approach (hereafter referred to as CloudRL) reasons across both workload and resource variability. The second approach does not reason about the variability of the resource, instead presuming that each additional resource gives a defined performance gain on the basis of its configuration. We refer to this as BasicRL. The addition of the variable resource performance model allows the agent to reason over the addition and removal of resources choosing those which have performed better in the past. Both approaches share the same parametric settings, as previously outlined in Section 5; however, there are significant differences in the respective state

DOI: 10.1002/cpe

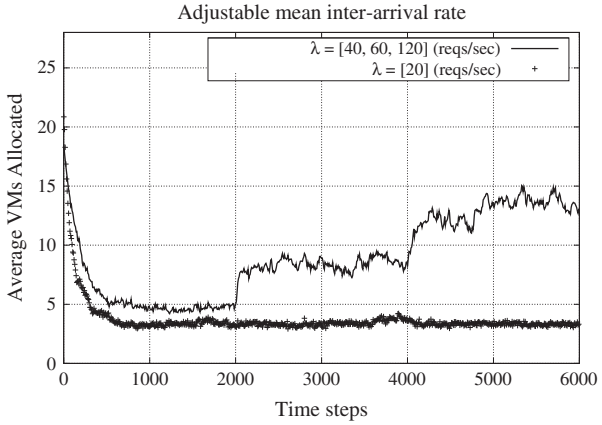


Figure 6. Q-learning performance under varying mean-arrival rates.

space representation. Firstly, the BasicRL approach does not incorporate current UTC into its state space rendering it incapable of reasoning about the possible effects of peak time in a particular region. Secondly, it considers all VMs to be homogenous/region independent. This approach is consistent with [24] previous research in cloud resource allocation and that of Grids also.

As stated in Section 4 with respect to response times per request, the value specified for P_c has a direct impact on how closely the learned policy approximates the given SLA value. A high penalty will encourage policies that produce lower response times, resulting in increased numbers of VMs and greater execution costs. The higher overall cost created by the additional VMs is offset by the fact that the SLA violation penalties are so high, and the agent will yield a greater reward by decreasing the probability of SLA violation. A low value for P_c will result in greater numbers of SLA violations but the relatively low penalty applied encourages policies that more closely approximate the given SLA, resulting in lower costs. The objective of each approach is to choose resources that facilitate the combined goals of cost reduction and maintain the SLA.

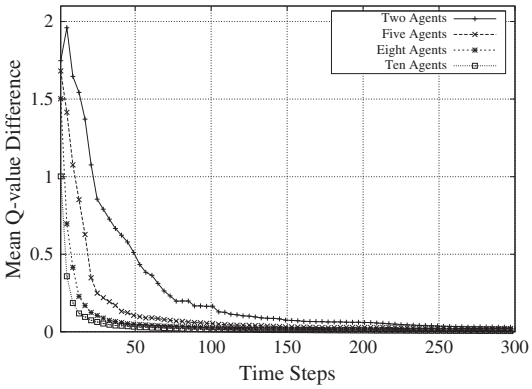
Figure 5(a) demonstrates the performance comparison between CloudRL and BasicRL with respect to average response time. The CloudRL approach has a higher average response time per request, standing roughly at 160 (msecs) at convergence; however, it is still considerably below the SLA of 250 (msecs). BasicRL is unable to reason about the geographical region or type of the VMs deployed. This results in a greater probability of choosing sub-optimal resources for a given time. As a result, its policy opts to maintain a much higher number of allocated resources to the application. Hence, it has a lower response time than the CloudRL approach, but incurs higher average costs as it has more resources allocated to the application, as is depicted in Figure 5(b). The multi-criteria Q-learning approach maintains on average 47% cost saving over the single-criteria approach, in the simulated cloud environment.

6.3. Adjustable mean inter-arrival rates

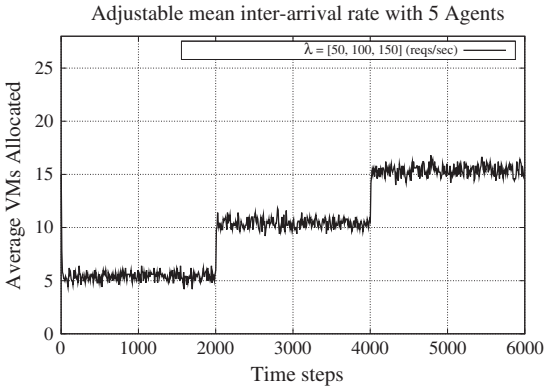
This experiment analyses the performance of Q-learning as the mean-arrival rate parameter λ is adjusted. The performance is compared against a user request model of fixed mean. The fixed workload request model consists of a Poisson distribution with a mean-arrival rate of 20 (reqs/sec). Each VM has a theoretical throughput in the range of 1–10 (reqs/sec), which varies per time step in accordance with the resource performance model.

Figure 6 plots the average number of VMs allocated each time step. The fixed workload ($\lambda = 20$) quickly converges to the optimal allocation of VMs. The plot showing the adjustable inter-arrival rates demonstrates the adaptability of the approach as workloads change. Every 2000 time steps, the mean arrival rate is increased. After 2000 time steps, the workload switches from 40 to 60 (reqs/sec). Figure 6 clearly demonstrates the speed at which the approach can determine the appropriate amount of resources to allocate given a change in the mean number of user requests. Whilst initially it takes time to converge, once an initial policy has been found the approach converges much faster to

DOI: 10.1002/cpe



(a) Average Q-value distance between learning agents



(b) Q-learning performance with 5 parallel agents with adjustable mean arrival-rate

Figure 7. The average Q-value distance between agents learning on the same task in parallel and the performance of five parallel agents with adjustable mean inter-arrival rates.

subsequent model changes as it has already gained experience from previous time steps. After 4000 time steps, the average request arrival-rate (λ) shifts again to 120 (reqs/sec). The approach takes greater time to converge to the larger request model. This is because of its limited experience of the newly observed states and the greater resource fluctuations as a result of the larger numbers of allocated VMs.

6.4. Agents learning in parallel

One of the challenges faced when dealing with real world problems with large state spaces is the time it takes to converge to an optimal policy. Usually, a substantial number of state visits are required to asymptotically converge to $Q^{\pi}(s,a)$; however, in reality often good policies can be determined with far fewer visits. In many real world problems, this level of performance is unacceptable. In an auto-scaling context, this would potentially lead to expensive allocations in the learning phase that would inhibit the commercial viability of the approach. In order to improve the length of time it takes to converge to an optimal policy, we examine a novel parallel learning approach for improving convergence times through the sharing of Q-value estimates.

Whilst each agent attempts to optimise allocations for their respective numbers of user requests, they will encounter similar states. By sharing estimates of the values of these states amongst each other, they can reduce the length of time it takes to learn an optimal policy. The stochastic nature of the workload request model ensures each agent will have a different learning experience, but they will all converge on the same result. The goal of this experiment is to speed up the time it takes to converge to a given policy. In order to facilitate a strict analysis of the parallel learning performance,

DOI: 10.1002/cpe

we homogenise the VMs allocated to the application with respect to location and performance. For simplicity and analysis the throughput of each VM is set to the maximum of 10 (reqs/sec). Figure 7(a) plots the average distance between Q-function estimates as the number of parallel agents is increased from 2 to 10 agents. The graph depicts the average distance between agents' approximation of the value of $Q(s, a)$. The graph shows the reduction in the time it takes to converge as the number of agents is increased. Figure 7(b) shows the resulting performance of five agents learning in parallel, where the request mean inter-arrival rates are adjusted every 2000 time steps. For the first 2000 time steps, the inter-arrival rate parameter is equal to 50 (reqs/sec). With this setting, the initial convergence to the average optimal allocation of VMs per time step takes about 80 timesteps. If you compare this to the single agent learning in Figure 6 albeit with a variable performance model, the convergence time as a result of learning in parallel has dropped significantly. As the rate parameter is shifted to 100 and 150 (reqs/sec) at 2000 and 4000 time steps, respectively, the time taken to converge is also dramatically reduced as a result of both the combination of prior knowledge and learning in parallel.

7. CONCLUSIONS AND FUTURE WORK

Reinforcement learning techniques have been successfully applied to automated control problems across a range of domains including economics, multi-agent systems and grid computing. Utilising reinforcement learning techniques to automatically control the scaling of virtual resources in supporting applications offers advantages with respect to reliability, adaptability and autonomy. Threshold based approaches to application scaling have predominated in the research community and in industry, such as Amazon's Auto Scaling and RightScale solutions. However, the development of effective thresholds that govern allocation decisions requires extensive domain and application knowledge and will often have to be redone in light of application updates or workload changes. Relying only on environmental observations, the reinforcement learner has an advantage as it can adjust its behaviour over time to changes in workload models and resource variability.

The focus of this paper is the proposal of a reinforcement learning approach aimed at optimising resource allocation decisions to support application scalability in cloud computing environments. Our novel state action space formalism is capable of guiding a Q-learning based agent towards good VM allocation policies in IaaS clouds with no prior experience. Coupled with variable numbers of user requests and resource performance uncertainty, it can effectively reason about multiple virtual machine types concurrently dealing with temporal performance issues. Reinforcement learning algorithms generally only offer asymptotic convergence guarantees, meaning that in order to determine optimal policies, a large number of state visits are required. For most real world problems, this restriction is too prohibitive, resulting in initial poor performance whilst the agent learns on line. To address this problem, we further extended our work to deal with this so called curse of dimensionality. By parallelising the Q-learning process, the time taken to converge to good policies is greatly reduced. The approach involves agents attempting to approximate optimal policies and sharing their individual learning experiences to improve the aggregated performance. The sharing of information amongst the agents greatly reduces the time taken to converge to a stable policy. The combination of parallel agent learning with our novel state space formalism enables advanced uncertainty reasoning capabilities over cloud resources.

In the future, we wish to integrate the approach into a live virtualised test-bed environment to evaluate performance outside the simulated environment. We are currently developing a hybrid cloud test-bed utilising OpenStack to manage both the internal hardware virtualisation and the public cloud resources. The reinforcement learning approach will be evaluated by deploying a standard benchmark application such as Apache Olio and comparing performance against traditional scalability mechanisms. Although approaches to support application scalability have generally involved allocating resources in an effort to support a performance related objective such as application response time, we feel that a more high level approach will be needed to address proper application scalability in a cloud context. In this light, we plan to investigate the combination of a more holistic approach to scaling applications, similar to what was proposed by Rodero-Merino *et al.* [15] and extended

DOI: 10.1002/cpe

by Morán *et al*. [16]. We feel this will allow for greater scope and scalability improvements by specifying higher level business functions with respect to scaling, in conjunction with learning.

ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge the continued support of Science Foundation Ireland.

REFERENCES

1. Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems* 2009; **25**(6):599–616.
2. Barham P, Dragovic B, Fraser K, Hand S, Harris T, Ho A, Neugebauer R, Pratt I, Warfield A. Xen and the art of virtualization. In *Proceedings of the nineteenth acm symposium on operating systems principles*, SOSP '03. ACM: New York, NY, USA, 2003; 164–177.
3. Vmware virtualisation software. (Available from: http://www.vmware.com/) [Accessed date :4 January 2012].
4. Koh Y, Knauerhase R, Brett P, Bowman M, Wen Z, Pu C. An analysis of performance interference effects in virtual environments. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, San Jose, California, USA, 2007.
5. Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, Patterson DA, Rabkin A, Stoica I, Zaharia M. Above the clouds: A berkeley view of cloud computing. *Technical Report UCB/EECS-2009-28*, EECS Department, University of California, Berkeley, Feb 2009.
6. Padala P, Hou KY, Shin KG, Zhu X, Uysal M, Wang Z, Singhal S, Merchant A. Automated control of multiple virtualized resources. In *Proceedings of the 4th ACM European Conference on Computer Systems*, EuroSys '09. ACM: New York, NY, USA, 2009; 13–26.
7. Padala P, Shin KG, Mustafa XZ, Wang UZ, Arif SS. Adaptive control of virtualized resources in utility computing environments. In *Proceedings of the European Conference on Computer Systems*, Lisbon, Portugal, 2007; 289–302.
8. Tesauro G, Jong NK, Das R, Bennani MN. On the use of hybrid reinforcement learning for autonomic resource allocation. *Cluster Computing* 2007; **10**(3):287–299.
9. Iosup A, Ostermann S, Yigitbasi N, Prodan R, Fahringer T, Epema D. Performance analysis of cloud computing services for many-tasks scientific computing. *Transactions on Parallel and Distributed Systems* June 2011; **22**:931–945.
10. Pu X, Liu L, Mei Y, Sivathanu S, Koh Y, Pu C. Understanding performance interference of i/o workload in virtualized cloud environments. In *Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing*, CLOUD '10. IEEE Computer Society: Washington, DC, USA, 2010; 51–58.
11. Amazon elastic compute cloud (amazon ec2). (Available from: http://aws.amazon.com/ec2/) [Accessed date: 4 January 2012].
12. Autoscale. (Available from: http://aws.amazon.com/autoscaling/) [Accessed date: 4 January 2012].
13. Cloudwatch. (Available from: http://aws.amazon.com/cloudwatch/) [Accessed date:4 January 2012].
14. Rightscale. (Available from:http://www.rightscale.com/) [Accessed date: 4 January 2012].
15. Rodero-Merino L, Vaquero LM, Gil V, Galán F, Fontán J, Montero RS, Llorente IM. From infrastructure delivery to service management in clouds. *Future Generation Computer Systems* Oct. 2010; **26**(8):1226–1240.
16. Morn D, Vaquero LM, Galán F. Elastically ruling the cloud: Specifying application's behavior in federated clouds. In *IEEE CLOUD*, Liu L, Parashar M (eds). IEEE: Washington DC, USA, July 2011; 89–96.
17. Dutreilh X, Rivierre N, Moreau A, Malenfant J, Truck I. From data center resource allocation to control theory and back. *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, july 2010; 410–417.
18. Rolia JA, Cherkasova L, McCarthy C. Configuring workload manager control parameters for resource pools. In *IEEE NOMS.* Vancouver, Canada, 2006; 127–137.
19. Lim HC, Babu S, Chase JS. Automated control for elastic storage. In *Proceeding of the 7th International Conference on Autonomic Computing*, ICAC '10. ACM: New York, NY, USA, 2010; 1–10.
20. Vengerov D. A reinforcement learning approach to dynamic resource allocation. *Engineering Applied Artifical Intelligence* 2007; **20**(3):383–390.
21. Perez J, Germain-Renaud C, Kégl B, Loomis C. Grid differentiated services: A reinforcement learning approach. In *CCGRID '08: Proceedings of the 2008 Eighth IEEE International Symposium on Cluster Computing and the Grid*. IEEE Computer Society: Washington, DC, USA, 2008; 287–294.
22. Foster I, Kesselman C, Tuecke S. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of High Performance Computing Applications* 2001; **15**(3):200–222.
23. Galstyan A, Czajkowski K, Lerman K. Resource allocation in the grid using reinforcement learning. In *AAMAS '04: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*. IEEE Computer Society: Washington, DC, USA, 2004; 1314–1315.
24. Dutreilh X, Kirgizov S, Melekhova O, Malenfant J, Rivierre N, Truck I. Using reinforcement learning for autonomic resource allocation in clouds: towards a fully automated workflow. In *Seventh International Conference on Autonomic and Autonomous Systems, ICAS 2011*, IEEE, May 2011; 67–74. MoVe INT LIP6.

DOI: 10.1002/cpe

25. Rao J, Bu X, Xu CZ, Wang L, Yin G. Vconf: a reinforcement learning approach to virtual machines auto-configuration. In *Proceedings of the 6th International Conference on Autonomic Computing*, ICAC '09. ACM: New York, NY, USA, 2009; 137–146.
26. Chevaleyre Y, Dunne PE, Endriss U, Lang J, Lemaître M, Maudet N, Padget J, Phelps S, Rodríguez-aguilar JA, Sousa P. Issues in multiagent resource allocation. *Informatica* 2006; **30**:2006.
27. Jacyno M, Bullock S, Luck M, Payne T. Understanding decentralised control of resource allocation in a minimal multi-agent system. In *The 6th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Honolulu, Hawai'i, 2007.
28. Scalas E, Gallegati M, Guerci E, Mas D, Tedeschi A. Growth and allocation of resources in economics: The agent-based approach, arXiv.org, 2006.
29. Wolski R, Brevik J, Plank JS, Bryan T. Grid resource allocation and control using computational economies. In *Grid computing: Making the global infrastructure a reality*. John Wiley & Sons: San Francisco, CA, 2003; 747–772.
30. Nau D, Ghallab M, Traverso P. *Automated Planning: Theory & Practice*. Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2004.
31. Watkins C. Learning from delayed rewards. *Ph.D. Thesis*, 1989.
32. Littman ML. Algorithms for sequential decision making, 1996.
33. Kretchmar RM. Parallel reinforcement learning. In *the 6th World Conference on Systemics, Cybernetics, and Informatics*, Orlando, USA, 2002.
34. Tesauro G. Online resource allocation using decompositional reinforcement learning. *Proc. AAAI-05*, 2005; 9–13.

DOI: 10.1002/cpe