Joint Virtual Machine and Bandwidth Allocation in Software Defined Network (SDN) and Cloud Computing Environments

Jonathan Chase, Rakpong Kaewpuang, Wen Yonggang, and Dusit Niyato
School of Computer Engineering, Nanyang Technological University, Singapore.

***Abstract*—Cloud computing provides users with great flexibility when provisioning resources, with cloud providers offering a choice of reservation and on-demand purchasing options. Reservation plans offer cheaper prices, but must be chosen in advance, and therefore must be appropriate to users' requirements. If demand is uncertain, the reservation plan may not be sufficient and on-demand resources have to be provisioned. Previous work focused on optimally placing virtual machines with cloud providers to minimize total cost. However, many applications require large amounts of network bandwidth. Therefore, considering only virtual machines offers an incomplete view of the system. Exploiting recent developments in software defined networking (SDN), we propose a unified approach that integrates virtual machine and network bandwidth provisioning. We solve a stochastic integer programming problem to obtain an optimal provisioning of both virtual machines and network bandwidth, when demand is uncertain. Numerical results clearly show that our proposed solution minimizes users' costs and provides superior performance to alternative methods. We believe that this integrated approach is the way forward for cloud computing to support network intensive applications.**

***Index Terms*—Cloud computing, software defined network, virtual machine, bandwidth allocation**

I. Introduction

Cloud computing permits users to access computing resources from cloud providers over the Internet. Virtualization technology allows users to access computing resources by renting virtual machines (VMs) that are tailored to their requirements. Cloud providers offer prepaid reservation instances at a reduced usage price, with additional VMs obtainable on demand at a higher price. When demand is unknown in advance, there is a high risk of either underprovisioning (when the user reserves too few resources for their needs) or overprovisioning (user reserves too many resources and pays for resources they do not need). A broker can be used to minimize cost through efficient VM provisioning.

However, with many cloud-based applications requiring a large amount of Internet bandwidth, an approach that focuses purely on VM placement is too simplistic, a unified approach that incorporates bandwidth reservation is needed. To achieve this, we exploit the control offered by a new and developing technology called software defined networking (SDN) [1]. SDN is an approach to network virtualization that separates network control from the mechanics of routing. SDN uses a three layer architecture with an application layer where users can run control programs that are implemented transparently by the lower level control and infrastructure layers. This application layer allows the network controller to offer a virtual network, with virtual routers and links, that the controller then implements on the physical network. Thus network controllers can offer reservation and on-demand bandwidth.

In this paper, we propose a unified optimal provisioning algorithm that places VMs and network bandwidth to minimize a users' costs. The algorithm makes a decision to reserve VMs and network bandwidth from certain cloud providers to minimize costs, by trading off the cost of overprovisioning against the cost of on-demand resources. We formulate a stochastic integer programming (SIP) problem with two-stage recourse to obtain the optimal decision. We demonstrate the solution's effectiveness through various numerical results and show that users' costs can be minimized while meeting user demand and cloud provider capacity limits.

II. Related Work

Cloud computing typically offers three main service models, Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS). IaaS services, such as Amazon EC2 [2], exploit virtualization technology to offer resources to users in the form of VMs. Allocation methods may take a heuristic approach, such as in [3], which also supports migration. Alternatively, [4] describes an exact provisioning algorithm that accommodates both VM demand and price uncertainty. Introducing the use of two-stage stochastic optimization [5] to computing resource provisioning, these works find optimal solutions to the NP-hard VM allocation problem. Similarly, [6] formulates mixed integer problems, and compares them, focusing on minimizing execution time.

The large amounts of bandwidth required by modern cloud applications can be guaranteed through bandwidth reservation. For example, [7] addresses the problem of mapping user requests to data centers and routing the responses but does not consider computation. The OpenFlow [8] implementation of SDN allows the flexible allocation of bandwidth through virtual networks, and therefore the reservation of bandwidth. Combining this power with methods used in Virtual Network Embedding (VNE), bandwidth allocation can be performed for the cloud. For example, [9] formulates the joint problem of virtual node and link allocation as an Integer Programming problem, although without a stochastic element. Like VM allocation, this problem is NP-hard, so is often solved with heuristics instead [10].

978-1-4799-2003-7/14/$31.00 ©2014 IEEE

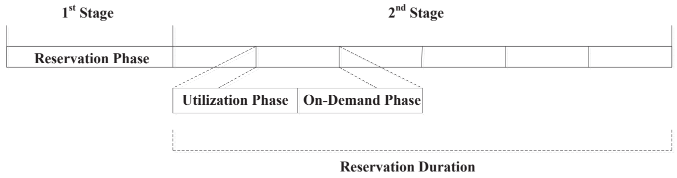

Fig. 1. Timeline showing the stages of decision-making. Reservation decisions last for a longer period of time, with second stage decisions made for a shorter duration once demand is known.

The above work lays the foundation for a joint VM and bandwidth allocation solution. However, as far as we are aware, there are currently no works that address the problem directly. It is our belief that the two must be considered together to achieve a globally optimal solution to the cloud resource provisioning problem.

III. System Model

A. Cloud and Network Resources

The system model under consideration for joint VM and bandwidth allocation is composed of cloud service providers and ISPs. There is a central controller receiving demand requests for VM provisioning from users. The controller then provisions the required VMs from cloud providers and provisions virtualized bandwidth from ISPs. ISPs can implement SDN (e.g. with OpenFlow-enabled switches), enabling the central controller to make bandwidth provisioning and routing decisions on virtualized routers. A practical implementation may require separate controllers to administer networks and providers, however, we are interested in the decision-making process, which we model with a single controller.

The provisioning decision is made in two stages (Fig. 1). The first stage is the reservation phase, where VMs and bandwidth are reserved for a long period of time, typically a year, in advance, before actual demand is known. The second stage happens at the time of use, when the users' actual demand is realized (e.g., daily demand). The second stage is divided into two phases, i.e., utilization and on-demand phases. In the utilization phase, the needed reserved resources are used, usually at a low price. If the actual demand is higher than the reserved resources, the algorithm enters the on-demand phase. In the on-demand phase, additional resources can be provisioned at a higher cost to satisfy any unsatisfied demand. Therefore, it is important for the first stage to be optimally determined as it is significantly cheaper than the second stage, but less flexible, due to the long duration of reservation.

In the system shown in Fig. 2, an ISP manages a set of virtualized routers, denoted by $\mathcal{R} = \{1, \dots, R\}$, where R is the total number of routers. Each router has a bandwidth capacity, t_r. Reservation, utilization, and on-demand bandwidth costs, through router r, are given by $c_r^{(re)}$, $c_r^{(u)}$, and $c_r^{(o)}$, respectively. A set of cloud providers is denoted by $\mathcal{P} = \{1, \dots, P\}$, where P is the total number of cloud providers. Each cloud provider, p, has three resources, i.e., processing power, storage, and internal network bandwidth, whose capacities are given by $t_p^{(h)}$, $t_p^{(s)}$, and $t_p^{(n)}$, respectively. Here internal bandwidth is bandwidth allocated for a VM (e.g., for a server inside a data center), while external bandwidth refers to the bandwidth of the routers of the ISP. $\mathcal{V} = \{V_1, V_2, \dots, V_{last}\}$ denotes the set of VM classes. Each class has distinct resource requirements, with $d_i^{(h)}$, $d_i^{(s)}$, $d_i^{(n)}$, and $d_i^{(b)}$, denoting a VM from class V_i's requirements for processing, storage, internal bandwidth, and external bandwidth, respectively. The costs of reserving, utilizing, and obtaining on-demand, a VM, from class V_i, from provider p, are given by $c_{ip}^{(re)}$, $c_{ip}^{(u)}$, and $c_{ip}^{(o)}$, respectively. All costs are known in advance, but the VM demand (and therefore the external bandwidth demand) is uncertain.

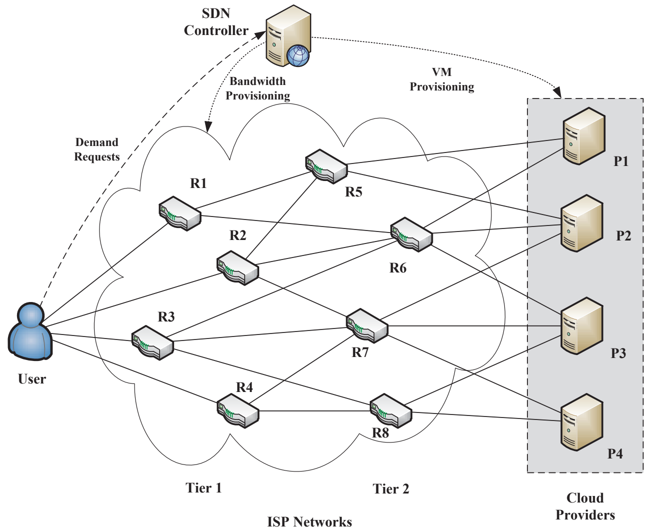

Fig. 2. System model. Bandwidth reservation is made from ISP networks, VMs are placed with cloud providers. Central controller makes provisioning choices based on user requirements.

B. Uncertainty of VM and Bandwidth Demand

With uncertain demand, the number of VMs required from each class is unknown at the point of reservation. $\mathcal{D}_i = \{d_{i1}, d_{i2}, \dots, d_{i|\mathcal{V}|}\}$ denotes the set of possible numbers of VMs that can be reserved from class V_i. The set of all possible VM demand values can therefore be given by

$$\mathcal{D} = \prod_{V_i \in \mathcal{V}} \mathcal{D}_i = \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_{|\mathcal{V}|} \tag{1}$$

where $\prod$ and $\times$ are the Cartesian product. Given uncertain VM demand, the total bandwidth that needs to be obtained is also unknown at the point of reservation. Since each VM class has a fixed external bandwidth requirement, $d_i^{(b)}$, (1) can be used to give the set of possible bandwidth requirements for class V_i. The set of possible bandwidth requirements for VM class V_i is defined as follows:

$$\mathcal{B}_i = \mathcal{D}_i \cdot d_i^{(b)} \, . \tag{2}$$

Then the set of possible external bandwidth requirements of all VM classes is expressed as follows:

$$\mathcal{B} = \prod_{V_i \in \mathcal{V}} \mathcal{D}_i \cdot d_i^{(b)} \, . \tag{3}$$

IV. Problem Formulation

A. Stochastic Optimization Problem

If VM demand were known at the point of reservation, the problem would be a simple deterministic optimization, with no on-demand phase. Since demand is uncertain, we need to apply a two-stage stochastic optimization. The stochastic optimization problem can be formulated as follows:

$$\min_{X_r^{(re)},Y_{ip}^{(re)}} \sum_{r\in\mathcal{R}}\sum_{p\in\mathcal{P}}\sum_{V_i\in\mathcal{V}} \Big(c_r^{(re)}X_r^{(re)} + c_{ip}^{(re)}Y_{ip}^{(re)} + \mathbb{E}[\mathcal{L}(X_r^{(re)},Y_{ip}^{(re)},\omega)]\Big). \tag{4}$$

The objective function in (4) minimizes the total cost of reservation added to the expected cost from the second stage. The decision variables $X_r^{(re)}$ and $Y_{ip}^{(re)}$ denote the amount of bandwidth reserved from router r and the number of VMs of class i reserved from cloud provider p, respectively. The expected second stage cost is represented by $\mathbb{E}[\mathcal{L}(X_r^{(re)},Y_{ip}^{(re)},\omega)]$, where ω denotes a scenario of demand: $\omega \in \Omega = \mathcal{D}\times\mathcal{B}$. For a given ω, the second stage cost function is given in (5), where decision variables $X_r^{(u)}(\omega)$ and $X_r^{(o)}(\omega)$ denote the amount of provisioned bandwidth from router r in the utilization and on-demand phases given demand scenario ω, respectively. Decision variables $Y_{ip}^{(u)}(\omega)$ and $Y_{ip}^{(o)}(\omega)$ denote the number of VMs from class i provisioned in the utilization and on-demand phases from cloud provider p, respectively. The full formulation, therefore, is to minimize (4), subject to the following constraints:

$$Y_{ip}^{(re)}, Y_{ip}^{(u)}(\omega), Y_{ip}^{(o)}(\omega) \in \{0,1,\ldots\} \tag{6}$$

$$X_r^{(re)}, X_r^{(u)}(\omega), X_r^{(o)}(\omega) \geq 0 \tag{7}$$

$$X_r^{(u)}(\omega) + X_r^{(o)}(\omega) \leq t_r,\ r\in\mathcal{R} \tag{8}$$

$$\sum_{V_i\in\mathcal{V}} d_i^{(h)}(Y_{ip}^{(u)}(\omega) + Y_{ip}^{(o)}(\omega)) \leq t_p^{(h)},\ p\in\mathcal{P} \tag{9}$$

$$\sum_{V_i\in\mathcal{V}} d_i^{(s)}(Y_{ip}^{(u)}(\omega) + Y_{ip}^{(o)}(\omega)) \leq t_p^{(s)},\ p\in\mathcal{P} \tag{10}$$

$$\sum_{V_i\in\mathcal{V}} d_i^{(n)}(Y_{ip}^{(u)}(\omega) + Y_{ip}^{(o)}(\omega)) \leq t_p^{(n)},\ p\in\mathcal{P} \tag{11}$$

$$X_r^{(u)}(\omega) \leq X_r^{(re)},\ r\in\mathcal{R} \tag{12}$$

$$Y_{ip}^{(u)}(\omega) \leq Y_{ip}^{(re)},\ p\in\mathcal{P}, V_i\in\mathcal{V} \tag{13}$$

$$\sum_{p\in\mathcal{P}}(Y_{ip}^{(u)}(\omega) + Y_{ip}^{(o)}(\omega)) \geq v_i(\omega),\ v_i\in\mathcal{V}\,. \tag{14}$$

- (6) prevents negative or partial provisioning of a VM.
- Similarly, (7) ensures non-negative bandwidth provisioning, but permits non-integer amounts.
- (8)-(11) ensure that the provisioned resources do not exceed capacity, with t_r giving the capacity of router r, and $t_p^{(h)}$, $t_p^{(s)}$, and $t_p^{(n)}$ giving the capacity of cloud provider p for each of the three resources, i.e., processing power, storage, and internal bandwidth, respectively.
- (12) and (13) ensures that the provisioning in utilization phase does not exceed the reserved amount.
- (14) ensures the VM demand realized in the second stage, denoted by $v_i(\omega)$ for VM class i, is satisfied.

Bandwidth demand must also be met, as well as ensuring that flow is conserved across the whole network. We consider the network of routers as a graph, $G = (\mathcal{R},\mathcal{E})$, where each router in $\mathcal{R}$ is a vertex, and each link in $\mathcal{E}$ is an edge, with flow preservation defined in terms of edges. However, as bandwidth is allocated at the router, we define a set of constraints to control flow based on routers as follows:

$$\sum_{r\in\mathcal{R}(p)} X_r^{(u)}(\omega)+X_r^{(o)}(\omega) \geq \sum_{V_i\in\mathcal{V}} d_i^{(b)}(Y_{ip}^{(u)}(\omega)+Y_{ip}^{(o)}(\omega)), p\in\mathcal{P} \tag{15}$$

$$\sum_{r_k\in\mathcal{R}_{out}(r)} X_{r_k}^{(u)}(\omega) + X_{r_k}^{(o)}(\omega) \geq X_r^{(u)}(\omega) + X_r^{(o)}(\omega),\ r\in\mathcal{R} \tag{16}$$

$$\sum_{r\in\mathcal{R}_{prov}} X_r^{(u)}(\omega)+X_r^{(o)}(\omega) \geq \sum_{p\in\mathcal{P}}\sum_{V_i\in\mathcal{V}} d_i^{(b)}(Y_{ip}^{(u)}(\omega)+Y_{ip}^{(o)}(\omega)) \tag{17}$$

$$\sum_{r\in\mathcal{R}_{user}} X_r^{(u)}(\omega) + X_r^{(o)}(\omega) \geq \sum_{r\in\mathcal{R}_{prov}} X_r^{(u)}(\omega) + X_r^{(o)}(\omega) \tag{18}$$

$$\sum_{r\in\mathcal{R}(p_m)} X_r^{(u)}(\omega) + X_r^{(o)}(\omega) \geq \sum_{V_i\in\mathcal{V}}\sum_{p\in p_m} d_i^{(b)}(Y_{ip}^{(u)}(\omega) + Y_{ip}^{(o)}(\omega)),\ p_m\in\mathcal{P}_{adj} \tag{19}$$

$$\sum_{r\in\mathcal{R}_{out}(r_m)} X_r^{(u)}(\omega) + X_r^{(o)}(\omega) \geq \sum_{r\in r_m} X_r^{(u)}(\omega) + X_r^{(o)}(\omega),\ r_m\in\mathcal{R}_{adj}\,. \tag{20}$$

The main idea of the above constraints is to balance the input bandwidth with output bandwidth at the routers. The inequalities ensure that the bandwidth 'supply' is at least sufficient to meet 'demand' throughout the network, relying on the minimization to remove excess bandwidth provisioning. Total actual provisioned bandwidth is given by the utilization and on-demand values for a given realization.

- (15) ensures that the bandwidth requirements of each individual cloud provider are met, where $\mathcal{R}(p)$ is a set of routers directly connected to cloud provider p.
- (16) ensures that the bandwidth allocated to each router, $r\in\mathcal{R}$, is supplied sufficiently by the routers, $r_k\in\mathcal{R}_{out}(r)$, that connect to it with their outward links.
- (17) ensures that the total bandwidth requirements of all cloud providers are met by the routers that directly

$$\mathcal{L}(X_r^{(re)}, Y_{ip}^{(re)}, \omega) = \min_{X_r^{(u)}(\omega), X_r^{(o)}(\omega), Y_{ip}^{(u)}(\omega), Y_{ip}^{(o)}(\omega)} c_r^{(u)} X_r^{(u)}(\omega) + c_r^{(o)} X_r^{(o)}(\omega) + c_{ip}^{(u)} Y_{ip}^{(u)}(\omega) + c_{ip}^{(o)} Y_{ip}^{(u)}(\omega) \quad (5)$$

connect to them. The set of all routers that link directly to cloud providers is given by $\mathcal{R}_{prov}$.

- (18) is to ensure that the total bandwidth supply is provided to the user, where $\mathcal{R}_{user}$ is the set of all routers with connections directly to the user.
- In (19) the set $\mathcal{P}_{adj}$ is the set of disjoint pairs, p_m, of providers where the providers in the pair are connected directly to at least one router in common. $\mathcal{R}(p_m)$ denotes the set of routers that connect directly to either of the providers in the pair, p_m.
- In (20), the set $r_m \in \mathcal{R}_{adj}$, is the set of disjoint pairs of routers, r_m, where the routers in the pair both receive input from at least one router in common, but are not directly connected to each other. $\mathcal{R}_{out}(r_m)$ denotes the set of all routers that connect directly to the routers in pair r_m with their outward links. Where routers have two or more outward links, there is a danger that the same bandwidth allocation will be assumed to supply multiple routers or cloud providers; (19) and (20) prevent this.

Although the flow is symmetrical, we consider the links that come from the user to be 'inward', and the links that go towards the cloud providers to be the 'outward' links.

B. Deterministic Equivalent Formulation

The stochastic formulation above can be transformed into a deterministic equivalent formulation by introducing four variables as equivalents to the decision variables in (4). These new variables are $X_r^{(u)}(b,d)$, $X_r^{(o)}(b,d)$, $Y_{ip}^{(u)}(b,d)$, and $Y_{ip}^{(o)}(b,d)$. When VM and bandwidth demand is realized, we have the observed values $d \in \mathcal{D}$ and $b \in \mathcal{B}$. This permits the reformulation of the objective function defined in (4) as shown in (21). In this case, the probabilities of VM and bandwidth demand are denoted by $\mathrm{p}(d)$ and $\mathrm{p}(b)$, being realized in the second stage. The realization instance (b,d) replaces the scenario ω that was used in the stochastic formulation. To ensure that the uncertain demand is accounted for, the deterministic formulation considers each possible instance of demand and chooses utilization and on-demand values for each one, with the total cost weighted by the probability of each realization.

We can also substitute in the new decision variables and obtain new constraints. Space does not permit listing each constraint in full, but we give an example for illustration.

$$X_r^{(u)}(b,d) + X_r^{(o)}(b,d) \leq t_r, \; r \in \mathcal{R}, b \in \mathcal{B}, d \in \mathcal{D} \quad (22)$$

(22) is the deterministic equivalent of (6). ω has been replaced by the realization of VM and bandwidth demand, i.e., (b,d), and the equation is evaluated for each demand realization, given by $b \in \mathcal{B}$ and $d \in \mathcal{D}$. The deterministic equivalents of the remaining constraints, (9)-(20), can be obtained similarly. The deterministic equivalent formulation is a mixed integer linear programming problem. Therefore, a standard solver (e.g., CPLEX) can find an optimal solution.

V. Performance Evaluation

A. Parameter Setting

Our formulation focuses on cost minimization, with a relatively small environment used for performance evaluation to demonstrate optimality. Resource allocation and integer programming problems are NP-hard, causing scalability problems in a cloud environment with a large number of routers and VMs. The question of computational efficiency is an important one to address, perhaps through a distributed method, or a heuristic relaxation of the integer constraints. Investigation of computational performance is left for future work.

1) VMs and Cloud Providers: We consider a representative cloud computing environment with a single cloud user, four cloud providers, three VM classes, and eight routers, arranged in two tiers, as depicted in Fig. 2. Similar to [4], the demand required for each VM class ranges from 1 to 50, i.e., $\mathcal{D}_i = \{1, 2, \ldots, 50\}$. For simplicity, the sets of possible demands are identical for the three VM classes as follows:

$$\mathcal{D} = \{d_1, d_2, d_3 | d_1 \in \mathcal{D}_1, d_2 \in \mathcal{D}_2, d_3 \in \mathcal{D}_3, d_1 = d_2 = d_3\}. \quad (23)$$

The resource requirements for each VM class are drawn from a synthesis of reasonable bandwidth requirements, and the computing and storage requirements of the small, medium and large standard instances available from Amazon EC2 [2]. Requirements are given in daily quantities, with computing power measured in CPU-hours, and storage and bandwidth measured in GBs. Computing requirements for V_1, V_2, and V_3 are 24, 24, and 48 CPU-hours, respectively. Storage requirements for classes V_1, V_2, and V_3 are 160, 410, and 840, respectively. Bandwidth requirements for V_1, V_2, and V_3 are 3GB, 3GB, and 9GB per day, respectively. Comprehensive pricing of resources are given in Table I, where the acronyms 'R', 'U', and 'O' stand for reservation, utilization, and on-demand, respectively. We here explain the rationale for these parameter settings. As in [4], P_1 is considered as a private cloud maintained by the user's organisation. The private cloud consists of 10 servers, with reservation costs equal to the energy costs of running the servers, given by \$357 per server per year. We take this as the reservation cost for 24 CPU-hours in the cloud. P_2's pricing is the same as Amazon EC2, whilst P_3 is based on Windows Azure's pricing schemes [11]. Windows Azure does not have separate reservation and utilization pricing, so we provide reasonable values for each. P_4 also follows [4] by only offering an On-Demand plan, albeit at cheaper rates than the other cloud providers. The capacity of providers P_2 to P_4 is taken to be unlimited, as they are large scale commercial cloud providers. The computing capacity of P_1 is 240 CPU-hours, which is the available computation in a day from 10 servers. The Storage capacity is given as 5000GB, based on the reasonable assumption that each server has a 500GB hard disk drive. Internal bandwidth for the cloud

$$\min \sum_{r\in\mathcal{R}} c_r^{(re)} X_r^{(re)} + \sum_{p\in\mathcal{P}} \sum_{V_i\in\mathcal{V}} c_{ip}^{(re)} Y_{ip}^{(re)} + \sum_{r\in\mathcal{R}} \sum_{b\in\mathcal{B}} \sum_{d\in\mathcal{D}} \mathrm{p}(b)\mathrm{p}(d)(c_r^{(u)} X_r^{(u)}(b,d) + c_r^{(o)} X_r^{(o)}(b,d)) + \\ \sum_{p\in\mathcal{P}} \sum_{V_i\in\mathcal{V}} \sum_{b\in\mathcal{B}} \sum_{d\in\mathcal{D}} \mathrm{p}(b)\mathrm{p}(d)(c_{ip}^{(u)} Y_{ip}^{(u)}(b,d) + c_{ip}^{(o)} Y_{ip}^{(o)}(b,d)) \,. \quad (21)$$

TABLE I
VM COSTS FOR EACH CLOUD PROVIDER AND PROVISIONING PHASE. PRICES ARE GIVEN FOR A DAY'S USAGE.

Provider	VM and Phase								
	V_1			V_2			V_3		
	R	U	O	R	U	O	R	U	O
P_1	0.978	NA	NA	0.978	NA	NA	1.956	NA	NA
P_2	0.189	1.656	2.184	0.378	3.312	4.368	0.756	6.600	8.736
P_3	0.150	1.680	2.160	0.330	3.360	4.320	0.725	6.960	8.640
P_4	NA	NA	1.920	NA	NA	3.840	NA	NA	7.68

TABLE II
COST PER GB PER DAY OF BANDWIDTH PROVISIONING THROUGH ISP ROUTERS

Routers	Provisioning Phase		
	R	U	O
R_1,R_5	0.10	0.03	0.50
R_2, R_6	0.16	0.02	0.20
R_3, R_4, R_7, R_8	0.17	0.01	0.30

providers is taken to be unlimited, as router bandwidth limits will be stricter.

2) ISP Routers: Our environment has eight routers, with the routers divided into three sets of pricing, given in Table II. Since bandwidth reservation and on demand pricing schemes are not yet commonly available from ISPs, we synthesize reasonable values. The bandwidth pricing is based on the Singtel[1] business fibre broadband schemes [12]. Bandwidth pricing is given per GB per day, with the reservation pricing for each scheme adapted from the monthly Gbps prices of the three different subscriptions available from Singtel. Utilization pricing is a low cost across all routers and is chosen to be suitably proportionate to the reservation prices. On-demand prices are determined as reasonable values using Singtel's per-GB excess usage fees for mobile broadband [13].

3) Demand Uncertainty: We model the VM demand uncertainty as a Normal distribution, with $\mu = 76.5$ and $\sigma = 36$, with bandwidth demand calculated as the sum of the bandwidth requirements for each required VM. For simplicity, bandwidth demand is treated as a deterministic function of VM demand, so $\mathrm{p}(d)$ and $\mathrm{p}(b)$ are equal, meaning that it is sufficient to evaluate only across one demand set.

B. Numerical studies

The deterministic equivalent problem above is encoded using GAMS (General Algebraic Modeling System) [14].

1) Impact of reservation: We conduct two studies of the effects of reservation in the system. In Fig. 3, we vary the number of reserved VMs and show the effect on total cost incurred to the user, both overall, and in first and second stages.

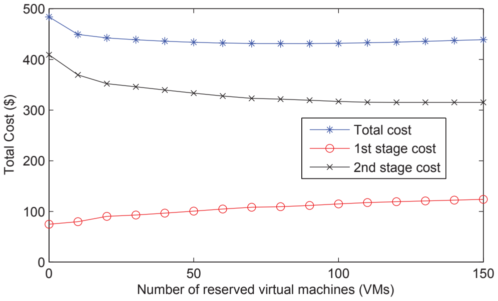

Fig. 3. Cost of user when the number of reserved VMs is varied.

As expected, as the number of reserved VMs increases, the first stage cost increases, and the second stage cost decreases. This results in a drop of total price until the optimum point is reached, when the benefits from reservation are exceeded by the extra cost of reserving redundant VMs. We similarly vary the amount of reserved bandwidth, but omit the graph due to space constraints. As expected, the bandwidth reservation result follows a similar pattern to the VM reservation result.

2) Meeting Demand: To observe our solution's effectiveness in handling varying demands, we examine the costs under different demand realizations. We consider the total cost, the VM cost, and the bandwidth cost, and compare our stochastic programming (SP) solution's results to a deterministic version (that knows the demand in advance), and an on-demand solution, that has no reservation option. Fig. 4 shows the total costs of each solution. As expected, the deterministic solution performed best, but our solution achieved performance that was close to the deterministic solution. The on-demand-only option is significantly inferior in all three cases, although when the number of required VMs is very small it offers a lower cost, because there is no overprovisioning cost.

3) Comparison with alternative methods: In Fig. 5, we vary the mean of the probability distribution and compare our stochastic programming (SP) solution to two other methods - one that employs no reservation phase, and an Expected Value Formulation (EVF) that uses the mean as its VM reservation value. We also set the standard deviation of the demand distribution to 120 to bring out the differences in

[1] Singtel is one of the largest network operators in Singapore.

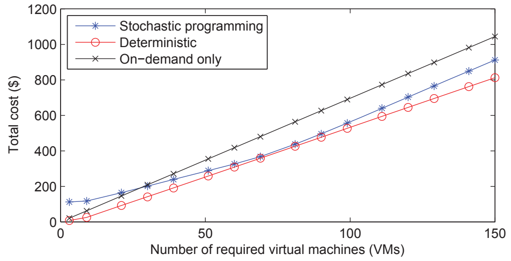

Fig. 4. Overall costs, when the demand is known.

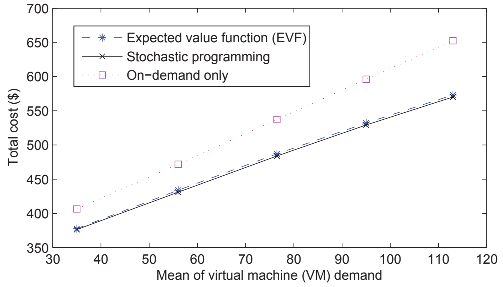

Fig. 5. Comparing total cost of stochastic programming solution with alternatives

performance between the methods. EVF performs well, as the optimal reservation decision is close to the mean when using normally distributed demand. We anticipate that if run against real demand data, EVF would perform significantly worse. As expected, on-demand-only is considerably more expensive.

4) Effect of varying router prices: To see the impact of router pricing on VM placement, we raise the prices for routers $R7$ and $R8$. We set the reservation and on-demand costs to be equal, as the difference in price between VMs and bandwidth will absorb the extra cost of on-demand provisioning. In this way, we can see that increasing the price of bandwidth provisioning forces migration of VMs from $P3$ and $P4$ to alternative cloud providers. $P1$ remains constant as it offers only reservation. $P3$ changes first, with VMs moving to $P2$, whilst $P4$ initially stays nearly constant. Since $P4$ is on-demand only and offers the cheapest on-demand prices, it still makes sense to use it. However, once there are no more VMs to remove from $P3$, the increasing bandwidth costs affect $P4$ as well, and the load is migrated to $P3$, where some of the traffic can be routed through the cheaper $R6$. We show this in Fig. 6 and observe that even with relatively low per-GB bandwidth prices, VM migration can be forced.

VI. Conclusion

We proposed an optimal algorithm for provisioning VMs and network bandwidth in a cloud computing environment, enabled by software defined networking. We have accounted for uncertain demand by formulating and solving a two-stage stochastic optimization problem to optimally reserve VMs and bandwidth to minimize cost. We have evaluated our solution's performance in the described environment. We have observed that the solution is optimal and outperforms alternative methods, achieving results close to those achieved when demand is perfectly known. Future directions for research include considering ISP networks that carry other traffic and introduce a random delay factor, consideration of non-deterministic per-virtual machine bandwidth demand, numerical studies based on real usage data, and design of a distributed solution for large-scale applications to improve scalability.

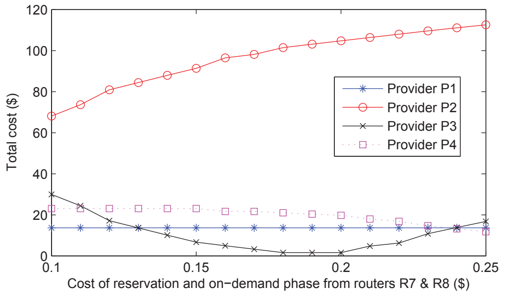

Fig. 6. Effect of router price variation on virtual machine (VM) placement

References

[1] "Software-defined networking: The new norm for networks," White Paper, April 2012. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf

[2] (2013, Aug.) Amazon ec2. [Online]. Available: http://aws.amazon.com/ec2/pricing/

[3] L. Grit, D. Irwin, A. Yumerefendi, and J. Chase, "Virtual machine hosting for networked clusters: Building the foundations for "autonomic" orchestration," in *Virtualization Technology in Distributed Computing, 2006. VTDC 2006. First International Workshop on*, 2006, pp. 7–7.

[4] S. Chaisiri, B.-S. Lee, and D. Niyato, "Optimization of resource provisioning cost in cloud computing," *Services Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 164–177, 2012.

[5] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*. New York: Springer, 2011.

[6] Q. T. Nguyen, N. Quang-Hung, N. H. Tuong, V. H. Tran, and N. Thoai, "Virtual machine allocation in cloud computing for minimizing total execution time on each machine," in *Computing, Management and Telecommunications (ComManTel), 2013 International Conference on*, Jan 2013, pp. 241–245.

[7] H. Xu and B. Li, "Joint request mapping and response routing for geo-distributed cloud services," in *INFOCOM, 2013 Proceedings IEEE*, 2013, pp. 854–862.

[8] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: Enabling innovation in campus networks," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, Mar. 2008. [Online]. Available: http://doi.acm.org/10.1145/1355734.1355746

[9] M. Chowdhury, M. Rahman, and R. Boutaba, "Vineyard: Virtual network embedding algorithms with coordinated node and link mapping," *Networking, IEEE/ACM Transactions on*, vol. 20, no. 1, pp. 206–219, Feb 2012.

[10] H. Di, L. Li, V. Anand, H. Yu, and G. Sun, "Cost efficient virtual infrastructure mapping using subgraph isomorphism," in *Communications and Photonics Conference and Exhibition (ACP), 2010 Asia*, Dec 2010, pp. 533–534.

[11] "Windows azure," Aug. 2013. [Online]. Available: http://www.windowsazure.com/en-us/pricing/details/virtual-machines/

[12] (2013, Aug.) Singtel business. [Online]. Available: http://info.singtel.com/business/products-and-services/internet/singnet-evolve-fibre-broadband

[13] (2013, Sep.) Singtel mobile broadband faqs. [Online]. Available: http://info.singtel.com/personal/communication/internet/broadband-on-the-go/mobile-broadband/mobile/faq

[14] D. Chattopadhyay, "Application of general algebraic modeling system to power system optimization," *Power Systems, IEEE Transactions on*, vol. 14, no. 1, pp. 15–22, 1999.