Opposites Attract: Virtual Cluster Embedding for Profit

Arne Ludwig[1], Carlo Fuerst[1], Alexander Henze[1], Stefan Schmid[1,2]
[1] TU Berlin, Germany, [2] Telekom Innovation Laboratories (T-Labs), Germany
{arne,carlo,ahenze}@inet.tu-berlin.de, stefan@net.t-labs.tu-berlin.de

ABSTRACT

It is well-known that cloud application performance critically depends on the network. Accordingly, new abstractions for cloud applications are proposed which extend the performance isolation guarantees to the network. The most common abstraction is the *Virtual Cluster* $VC(n, b)$: the n virtual machines of a customer are connected to a virtual switch at bandwidth b. However, today, not much is known about how to efficiently embed and price virtual clusters.

This paper makes two contributions. (1) We present an algorithm called Tetris that efficiently embeds virtual clusters arriving in an online fashion, by jointly optimizing the node and link resources. We show that this algorithm allows to multiplex more virtual clusters over the same physical infrastructure compared to existing algorithms, hence improving the provider profit. (2) We present the first demand-specific pricing model called DSP for virtual clusters. Our pricing model is fair in the sense that a customer only needs to pay for what he or she asked. Moreover, it features other desirable properties, such as price independence from mapping locations.

1. INTRODUCTION

Server virtualization has revamped the server business over the last years, and cloud computing has changed the way we think about resource allocation in the Internet. However, while cloud providers support the flexible allocation of virtual machines (*VMs*) and provide the customer with certain resource and performance isolation guarantees, networking has so far been treated as a second class citizen: until recently, cloud customers could not specify even basic networking guarantees to connect their virtual machines.

This is problematic given the fact that the traffic generated by cloud applications such as MapReduce and distributed databases is not negligible. Indeed, it has been shown that cloud applications suffer from resource interference on the network, in the sense that application performance can become unpredictable. Longer job execution times also entail higher costs for the customers who are charged on a *per-VM-hour basis*. [4]

To overcome these problems, more powerful resource reservation models have been proposed. [3] A common abstraction is the *Virtual Cluster*: a virtual cluster $VC(n, b)$ connects n virtual machines to a virtual switch at bandwidth b—essentially a hose model. [4] While the resulting performance isolation guarantees are attractive, the virtual cluster abstraction raises two fundamental questions: (1) How to embed virtual clusters on a given physical network? In order to make an optimal use of its resources and hence maximize the profit, a provider may want to multiplex as many virtual clusters as possible over the physical infrastructure (while fulfilling the requested virtual cluster specification). (2) How to efficiently price virtual clusters? While today's cloud pricing models typically focus on VM hours only, the pricing of virtual clusters becomes a 2-dimensional problem: a customer requesting more bandwidth should be priced accordingly. This paper addresses these two questions.

Our Contribution. We present the first pricing scheme for virtual clusters, called DSP (*Demand-Specific Pricing*), which explicitly takes into account the different computation and bandwidth requirements. That is, unlike the dominant-resource pricing scheme *DRP* presented in the literature before [2], DSP is designed according to a specification-dependent, *pay-only-for-what-you-request policy*: While in *DRP*, the size n and the bandwidth b of a virtual cluster are strictly coupled, DSP allows customers to request virtual clusters $VC(n, b)$ of arbitrary and independent size n and bandwidth b, and be priced accordingly and in a fair manner. Moreover, DSP ensures desirable properties such as location independent pricing.

Together with this pricing scheme we present a new embedding algorithm called Tetris[1] which is also specification-dependent in the sense that Tetris accounts for differences in the node and link requirements of virtual clusters. Concretely, Tetris is tailored to an online scenario where different virtual clusters $VC_1(n_1, b_1), VC_2(n_2, b_2), \ldots$ are requested over time, and collocates "opposites": computation-intensive but communication-extensive virtual clusters are mapped together with computation-extensive but communication-intensive virtual clusters, to maximize the number of simultaneously hosted virtual clusters over time. Given the online nature of the problem, this is a non-trivial task. We show that our algorithm outperforms previous algorithms, in the sense that a provider can host more virtual clusters, and hence increase its profit.

2. BACKGROUND & MODEL

Most cloud providers today still offer virtual machine services only, charging their customers on a per-hour basis.

[1]The name of the algorithm is due to its strategy to balance different resources equally, see also Figure 2 for an illustration.

However, we witness a trend towards more network oriented specifications (see also, e.g., *Amazon Placement Groups*, *Amazon EBS-Optimized Instances* or *Microsoft Azure ExpressRoute*), and especially the virtual cluster abstraction is becoming a popular model for datacenter applications. [3]

In [3], a first algorithm (henceforth called OKTOPUS) was proposed to embed virtual clusters in fat-tree datacenter topologies, and Ballani et al. [2] proposed a first pricing scheme for virtual clusters which give minimal bandwidth guarantees. Essentially, their scheme is based on *Dominant Resource Pricing*, short *DRP*. *DRP* provides different templates for the customers, with predefined sizes n and an associated amount of minimal guaranteed bandwidth b. While this model is attractive for its simplicity, also in the sense that the interface between the customer and provider may not have to be changed, the minimal bandwidth b is a function of n and cannot be chosen by the customer. As such, *DRP* is still 1-dimensional and does not leverage the full flexibility of the virtual cluster model, which is described by two *independent* parameters n and b.

Indeed, virtual cluster specifications are likely to come with different requirements [1] and can be heterogeneous [4]: a latency sensitive webservice can be very different in nature than, say, a delay-tolerant batch processing job or a network-hungry database synchronization application. One implication of the *DRP* scheme is that customers who know their virtual cluster demands might suffer from the inherent 1-dimensionality: in order to meet their application requirements, customers may be forced to upgrade to the next larger template, increasing both resources.

This paper seeks to overcome this problem by allowing customers to specify their computation and communication requirements separately. Concretely, we allow the customer to specify three parameters independently: the number of VMs n, the computational type c of the virtual machines (e.g., small, medium, or large instances), as well as the bandwidth b. That is, we use a virtual cluster abstraction $VC(n, c, b)$, where all virtual machines are of the same computational type c, and are connected to a virtual switch at bandwidth b.

We use the following conventions in our notation. The computational type c is normalized in the sense that c describes the fraction of the capacity of a physical server. Similarly, we will normalize b to denote the requested fraction of the overall link capacity of a physical server. A central concept for our algorithm TETRIS (improving upon OKTOPUS) and pricing scheme DSP (improving upon *DRP*) is the *resource ratio* between the requested node and link resources, henceforth denoted by $\rho = c/b$.

In general, we will assume that requests arriving in an online fashion have to be immediately embedded or rejected by the provider. In order to successfully embed a virtual cluster, the provider has to fulfill all its specifications.

3. PRICING SCHEME

The proposed specification-dependent pricing scheme DSP is based on a unit price for computation, henceforth denoted by p_c, as well as a unit price for communication (i.e., bandwidth), henceforth denoted by p_b. Ideally, for a virtual cluster request with a per-VM computational demand c and a per-VM bandwidth demand b, a customer should be charged according to the resource proportions, e.g.

$$P_{ideal} = n \cdot (c \cdot p_c + b \cdot p_b)$$

However, compared to a dominant resource pricing scheme, this solution can result in a lower income at the provider, especially if requests are highly heterogeneous leading to a higher fragmentation. While this income loss could be compensated by increasing the unit prices p_c, p_b accordingly, one has to be careful not to punish customers with an ideal resource ratio $\rho = 1$, who would prefer providers offering *DRP* in this case. To solve this problem, in the following, we propose to add a small charge for customers with a resource ratio $\rho \neq 1$.

But let us first revisit the *DRP* scheme given our notation. In *DRP*, a customer who requests a virtual cluster $VC(n, c, b)$, with relatively lower resp. higher computation requirements compared to the communication requirement, is forced to upgrade the request to the next larger template for both resources. The corresponding formula for the *DRP* scheme is

$$P_{DRP} = n \cdot [\max\{c, b\} \cdot (p_c + p_b)]$$

In order to bridge the difference between P_{ideal} and P_{DRP}, we propose the following *demand specific pricing* scheme DSP which introduces an extra fee for requests with an unbalanced resource ratio ρ. In this paper, we will assume a linear dependency between the extra cost and ρ, although other dependencies (e.g., quadratic) could also be expressed in our model. This decision is based on the assumption that more skewed requests are more likely to generate fragmentation. In summary, DSP computes the virtual cluster price as follows:

$$P_{DSP} = n \cdot (b \cdot p_b + c \cdot p_c) + \begin{cases} (c-b) \cdot p_b \cdot \lambda_b, & \text{if } c \geq b \\ (b-c) \cdot p_c \cdot \lambda_c, & \text{else} \end{cases}$$

where $\lambda_c, \lambda_b \geq 0$ are weighing factors. Note that this scheme can be seen as a generalization of the dominant resource pricing strategy: $\lambda_c, \lambda_b = 1$ implies that $P_{DSP} = P_{DRP}$. Lower weights result in savings for the customers, and $\lambda_c, \lambda_b = 0$ implies $P_{DSP} = P_{ideal}$. Given that the customers have a good understanding of their specific requirements in terms of computation and communication—a reasonable assumption as we argue—this pricing scheme leads to a higher provider profit, and in a competitive market, the extra income compared to *DRP* is shared with the customers.

Let us elaborate on the weighing factors λ_c and λ_b. In general, the factors should depend on the amount of embedded virtual requests as well as the current resource demand and supply. If the provider has a good estimation of the virtual clusters that will be requested, the values can be computed ahead of time; otherwise, the factors may be estimated over time, see also Section 5 for a discussion. Given a difference of Δ between the provider profit under *DRP* for upgraded and non-upgraded requests, the extra income generated by the lower resource consumption of the non-upgraded requests could be evenly distributed over requests requiring more of either one of the two resources. (Recall that the price of requests with $\rho = 1$ will not change.) The calculation for λ_b in a scenario with N virtual machines for which $c > b$ and with expected requirements $E(c)$ and $E(b)$, is given by

$$N \cdot (E[c] - E[b]) \cdot p_b \cdot (1 - \lambda_b) = \Delta/2$$

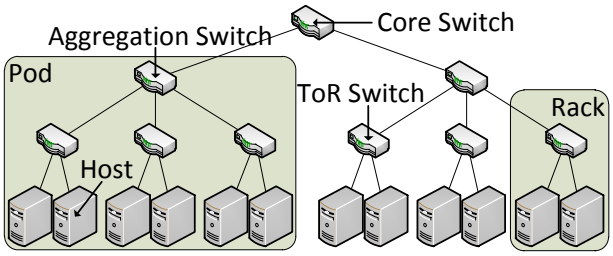

Figure 1: Fat-tree datacenter topology.

The factor λ_c can be computed similarly. In both cases, $c > b$ and $b > c$, a similar difference for $E[c]$ and $E[b]$ also leads to similar fees. This is fair, as one type of request only generates more profit because of the other one.

Also note that DSP keeps the desirable location independence of *DRP* [2]. However, unlike *DRP*, DSP is specification-dependent, i.e., a customer only has to pay for what he or she specified.

4. EMBEDDING ALGORITHM

In order to fully exploit differences in the virtual cluster specification and in order to maximize the resource utilization (and hence provider profit), a new embedding algorithm has to be devised: the state-of-the-art virtual cluster embedding algorithm, Oktopus [3] (as well as its variants [6]), are based on an aggressive collocation strategy, which turns out to be suboptimal in settings where requests can have resource ratios $\rho \neq 1$.

In the following, we propose Tetris, a virtual cluster embedding algorithm which leverages the virtual cluster specification details. The algorithm is tailored toward fat-tree datacenter topologies (cf Figure 1), the standard architecture today. In a nutshell, hosts (or equivalently: servers) which are connected to the same top of rack (ToR) switch, constitute a rack. Racks connected to the same aggregation switch form a pod. The fat-tree depicted in Figure 1 consists of two pods, containing three racks each; there are two hosts per rack.

We will compare our embedding algorithm Tetris to the Oktopus algorithm [3]. Oktopus is designed to embed arbitrary virtual clusters efficiently and is not limited to any templates. Hence it can also directly be used as an embedding algorithm for DSP, without any modifications. To find an embedding, Oktopus traverses the different hierarchical levels of the fat-tree. It tries to embed the virtual cluster on single hosts first, and if no solution is found, it continues on the rack level. This process is repeated until a solution is found or Oktopus failed to find a solution over the entire substrate. As a result of this approach, the resulting embeddings are dense and use low amount of bandwidth. The problem of such dense embeddings is that requests with a resource ratio $\rho \neq 1$ are collocated which wastes physical resources. Figure 2 (*left*) illustrates this point. For Oktopus, VC_1 is embedded on the right three hosts. The residual capacity in terms of VM slots on each host is 1/2 of its total capacity, however, the bandwidth on the links is used up, rendering it unlikely that the remaining VM slots can be used in the future. On the other hand, the left three hosts, which host VC_2, only utilize 50% of their bandwidth, but have no remaining VM slots.

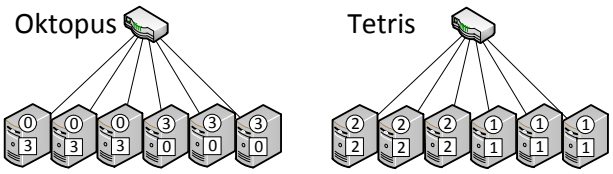

Figure 2: Embedding behavior of Oktopus and Tetris. Six hosts are connected to a switch. Two VCs are requested: $VC_1(9, 1/6, 2/6)$ and $VC_2(9, 2/6, 1/6)$. The numbers in the circles represent *VMs* of VC_1 which are mapped to the corresponding hosts, the numbers in the squares represent *VMs* of VC_2.

The core idea of our algorithm Tetris is to distribute skewed VCs over multiple hosts, without increasing the bandwidth costs compared to Oktopus. Similar to Oktopus, Tetris also traverses the hierarchical levels (short: ℓ) of the fat-tree, but instead of collocating as many *VMs* on a single host as possible, Tetris distributes the *VMs* over physical machines (short: p) depending on the ratio of the residual resources per host. This is described in Algorithm 1.

Algorithm 1 Tetris

Require: Fat-tree T, virtual cluster VC
1: **for** $\ell \in \{\text{host}(T), \text{rack}(T), \text{pod}(T), \text{root}(T)\}$ **do**
2: **for** $v \in VC$ **do**
3: find $p \in \ell$ with highest ratio of residual resources after embedding of v
4: **if** $\forall\ v$ embedding found **then**
5: **return** embedding
6: **return** $\perp$

In our example in Figure 2, using Tetris results in a distributed embedding of both VCs. While all resources on the left three hosts are utilized, the right three hosts have spare capacities, both in terms of bandwidth and *VM* slots. Hence subsequent requests can more likely be accepted.

Note that the current design of Tetris only considers the bandwidth on the access level. Hence, it can fail to find a feasible solution if the other layers of the fat-tree are oversubscribed. Therefore our current implementation treats Tetris as an extension to Oktopus and falls back to regular Oktopus behavior if no solution was found.

5. SIMULATIONS

In order to study the benefits and limitations of DSP and Tetris in different settings, we implemented a discrete event simulator. As the pricing results also depend on the embedding algorithm, we study three combinations: we integrated *DRP* with Oktopus, DSP with Oktopus, and DSP with Tetris. To ensure a fair comparison, we use the same parameters and methodology as in [3] and [2].

Requests. The virtual cluster requests arrive according to a Poisson process with exponentially distributed durations, chosen to induce a system load of around 80%. By default, the mean size of a VC is $n = 49$, c and b are chosen uniformly from $\{1/8, 1/4, 1/2\}$. The templates (c, b) of *DRP* are $(1/8, 1/8), (1/4, 1/4), (1/2, 1/2)$ and the customer is bound to select the next larger template for requests with

$\rho \neq 1$. For each parameter set, we request $80k$ virtual clusters. To avoid artifacts related to the initially empty datacenter, we start evaluating our metrics after $10k$ requests. The remaining values are omitted from the dataset.

Physical Setup. We model our datacenter as a three-layer fat-tree (Figure 1 illustrates a small fat-tree). 40 hosts form a rack, 40 racks form a pod. In total we have 10 pods. Given that each physical element has a capacity of 8 *VM* units, this leads to a total capacity of $128k$ small VMs. By default we assume that the links between the ToR switches and the aggregation switches are oversubscribed by a factor of 4, while the links between the aggregation switches and the core are not oversubscribed.

Metrics. Various works have used the acceptance ratio of an embedding algorithm in order to measure its performance. This metric however, is biased to prefer algorithms which accept a large number of small requests instead of fewer bigger ones. Therefore, we decided to evaluate the sum of the embedded virtual resources in both dimensions: bandwidth and VM slots. A request $VC(10, 1/4, 1/8)$ will have a *resource sum* of 10 for bandwidth and 20 for *VM* slots. Note that even though we will embed a $VC(10, 1/4, 1/4)$ for *DRP*, the bandwidth sum remains 10, as the customer does not benefit from the over-provisioning.

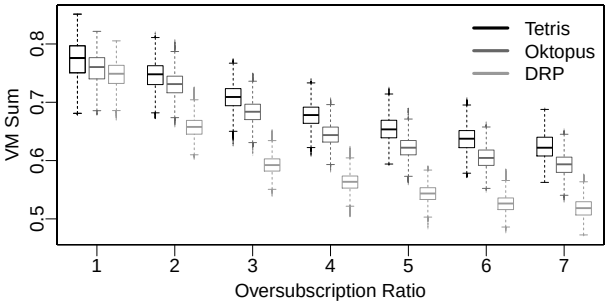

Figure 3: Embedded *VM* slots for Tetris with DSP (*left*), Oktopus with DSP (*middle*) and Oktopus with *DRP* (*right*) as a function of the oversubscription.

Figure 3 shows the impact of the oversubscription factor on the embedded number of virtual machines. For all oversubscription ratios, we can observe that Tetris with DSP outperforms the other combinations, while DSP is superior to *DRP* in combination with Oktopus. While the differences are small for an oversubscription factor of 1, we see an increase of the difference with the oversubscription factor; it diminishes after an oversubscription factor of 5, where the results become stable.

Naturally, we can only reap the full benefits of DSP and Tetris in highly utilized datacenters, as shown in Figure 4: While first marginal effects can be observed at a load of 0.4, the benefits of DSP are visible starting at 0.5 and the benefits of Tetris at 0.6. The effects increase until a load of 1. Given that highly utilized datacenters are a reality today and the key to provider benefits, these results are encouraging.

To analyze the benefit of the pricing model, we consider our default scenario: The mean resource sum for DSP with Oktopus is 15% higher than for *DRP* with Oktopus. This means that the amount of concurrent active guarantees is

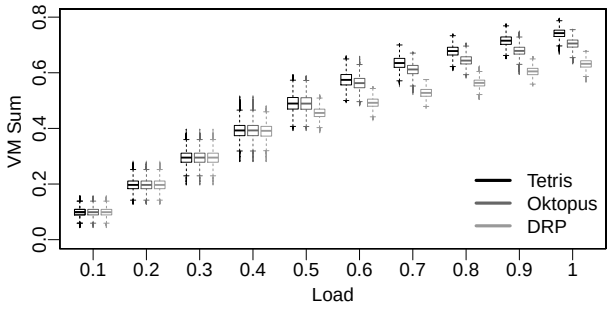

Figure 4: Embedded *VM* slots for Tetris with DSP (*left*), Oktopus with DSP (*middle*) and Oktopus with *DRP* (*right*) as a function of datacenter load.

15% larger. Using Tetris with DSP yields another 5% improvement, resulting in a total benefit of 20%. Similar numbers apply for the bandwidth resource sum.

Assuming $p_b = p_c$ and a uniform distribution of accepted requests (i.e., an equal amount of embedded *VM*s of each (c, b) tuple), inserting the 20% as a Δ in Section 3 leads to savings of 27% for customers with skewed requests. The corresponding values of λ_c and λ_b are 1/6, i.e., customers have to be charged for about 17% of the resource difference to *DRP*s templates in order to keep the revenue for the provider constant.

6. CONCLUSION

This paper has presented the first specification-dependent pricing scheme for virtual clusters, the standard abstraction of cloud applications today. Together with the pricing scheme, we also developed a new virtual cluster embedding algorithm, which may be of independent interest. Our approach can easily be combined with interfaces which translate high-level customer goals into virtual cluster requirements and compute resource combinations, such as [5].

We find that the proposed specification-dependent pricing can increase the social welfare, leading to savings of up to 25% compared to *DRP*. In fact DSP enables customer to have guaranteed application performance while they only need to pay for the resources they use, including a small extra fee. Moreover, we find that Tetris distributes virtual clusters with skewed resource requirements over several hosts and therefore embeds 5% more virtual resources than Oktopus.

In our future research, we want to extend the DSP scheme by a spot market to deal with more volatile traffic patterns.

7. REFERENCES

[1] B. Hindman et al. Mesos: a platform for fine-grained resource sharing in the data center. In *NSDI*, 2011.

[2] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron. The price is right: towards location-independent costs in datacenters. In *HotNets*, 2011.

[3] H. Ballani et al. Towards predictable datacenter networks. In *SIGCOMM*, 2011.

[4] J. C. Mogul and L. Popa. What we talk about when we talk about cloud network performance. *CCR*, 2012.

[5] V. Jalaparti et al. Bridging the Tenant-Provider Gap in Cloud Services. In *SoCC*, 2012.

[6] D. Xie, N. Ding, Y. C. Hu, and R. Kompella. The only constant is change: incorporating time-varying network reservations in data centers. *CCR*, 2012.