Distributed Cloud Market: Who Benefits from Specification Flexibilities?

Arne Ludwig[1], Stefan Schmid[2]
[1] TU Berlin, Germany, [2] Telekom Innovation Laboratories & TU Berlin, Germany
arne@inet.tu-berlin.de, stefan@net.t-labs.tu-berlin.de

ABSTRACT

Virtualization is arguably the main innovation motor in the Internet today. Virtualization enables the decoupling of applications from the physical infrastructure, and introduces new mapping and scheduling flexibilities. While the corresponding algorithmic problems are fairly well-understood, we ask: Who reaps the benefits from the virtualization flexibilities? We introduce two simple distributed cloud market models and study this question in two dimensions: (1) a horizontal market where different cloud providers compete for the customer requests, and (2) a vertical market where a broker resells the resources of a cloud provider.

1. INTRODUCTION

We live in an age where computation and storage have become a utility and can be scaled elastically: geographically distributed resources can flexibly be aggregated and shared by multiple tenants. After revamping the server business, the virtualization trend has also started spilling over to the network: The network virtualization paradigm [5] envisions a world where entire *virtual networks (VNets)*, with QoS and isolation guarantees *on both nodes and links*, can be requested on demand. Network virtualization decouples the applications and services from the constraints of the underlying physical network, and where and when resources are allocated only depends on the VNet specification itself.

Service and virtual network specifications are unlikely to be very homogeneous [8]. Today's datacenters concurrently host a wide range of applications of different tenants with different requirements [2]: while some applications are network-hungry [6] or latency sensitive [10] (e.g., a web service), and may have deadlines [9] and require strict QoS networking guarantees, other applications (e.g., batch processing jobs) are delay-tolerant.

While today we have a fairly good understanding of how to algorithmically exploit different service specifications and requirements [7], e.g., in order to find cheap or resource-efficient embeddings, this paper asks the question: *Who benefits from specification flexibilities in a competitive cloud market?* For example, a customer may expect that being more delay-tolerant and flexible in terms of resource rates, pays off, i.e., render the service cheaper [1]. Indeed, as the cloud provider may exploit specification flexibilities to schedule applications more flexibly and hence make a better use of its resources, it may share the gains with its customers.

Model: Horizontal & Vertical Market. We consider a simple setting where customers issue virtual network (VNet) requests. Each VNet specifies (1) *resource rates* for virtual nodes and links, i.e., a fixed amount of CPU per time or bandwidth (using graph or *hose/VC/VOC* models [8]), (2) a *duration* for which the VNet must be embedded (at the specified resource rate), and (3) possibly a *deadline* by which the VNet must have been embedded for the entire duration (and rate). We propose the following two simplified VNet market models (cf Figure 1). In the *horizontal market model* (Figure 1 *left*), the customers (or "tenants") directly request virtual networks (VNets) on demand from different cloud providers. We assume that a customer first issues the VNet request (annotated with the required resources) to all cloud providers, in order to obtain an offer on (1) *when* the VNet request can be scheduled (time t), and (2) at which *price* p. We assume that the providers have fixed but different prices per resource unit, and will greedily schedule a VNet at the earliest possible point in time. Given the time-price tuples (t, p), the customer will choose the best provider offer according a utility function $u(t, p)$. Depending on the application, the customer may be relatively flexible in the execution time as long as the best price is obtained; or, conversely, he or she may be relatively flexible in the price as long as the job is processed as soon as possible. In the *vertical market model* (broker market), customer requests are handled by a broker (cf Figure 1 *right*) that is responsible of embedding the VNets on its virtual resources (similar to the role for the cloud provider in the basic model). The broker role benefits from being able to buy larger chunks of resources as it obtains a discount from the cloud provider. Concretely, we will assume that the broker can buy different *resource contracts* from the cloud provider: a resource contract consists of a resource volume (i.e., an overall resource rate R) and a duration D. The larger the product $R \times D$ of resource volume and duration of the contract (henceforth also referred to as the contract *area*), the higher the discount. Resource discounts are common (e.g., in Amazon's EC2 reserved instances) and yield a tradeoff for the broker: buying too large contracts may be wasteful as the actual resources cannot be resold, and buying too small contracts may yield small discounts. In order to study how the costs of the broker and the income of the cloud provider depend on how flexible the customers are with respect to the *VNet deadline*, we consider different broker strategies.

2. BENEFITS IN HORIZONTAL MARKET

We first study how VNet flexibilities influence the income distribution of the different cloud providers. We make the natural assumption that cheaper and faster executions are always preferred over more expensive and longer alternatives. Concretely, we consider the following exemplary utility functions: (1) customers are relatively flexible in time as long as the price is low: $u_f(t, p) = -t - 10 \cdot p$; (2) customers are inflexible regarding time even if this turns out to be more expensive: $u_i(t, p) = -10 \cdot t - p$; and (3) customers are not specifically flexible or inflexible $u_e(t, p) = -t - p$.

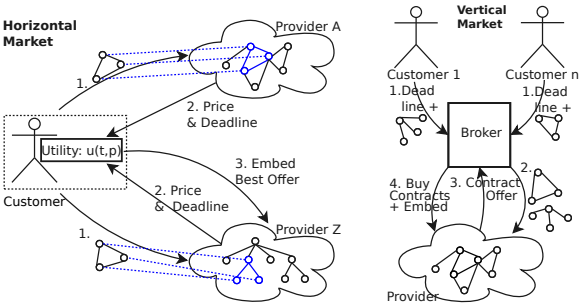

Figure 1: *Left:* Horizontal market: A customer requests an offer for a VNet embedding from each provider. Depending on the price-deadline tuples returned by the providers, the best option is chosen according to a given utility function. *Right:* Vertical market: Customer VNet requests (with deadlines) are directed towards the broker which is buying resource contracts (subject to discounts for larger contracts) from the cloud providers.

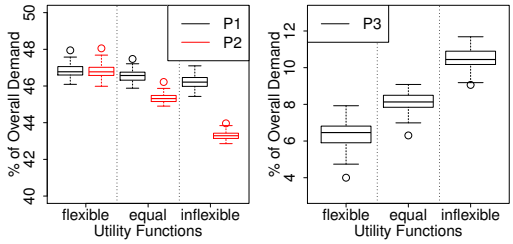

Figure 2: Boxplots (*left:* $P1$ and $P2$, *right:* $P3$) of the overall workload, in percentages per provider and under a stable pricing scheme $(100, 110, 120)$. The data is collected over 100 runs with 20k requests each (excluding the first 1k to to remove the "bootstrap phase"). All providers have a capacity of 70 units, the demand is subject to *Poisson arrival* ($\lambda = 1$) and durations are *Pareto* (α=3, min=1, scaling factor=100).

Finally, (4) we also investigate scenarios where VNets have *strict* deadlines d; i.e., the customer will not accept offers violating this deadline. We will assume that providers have fixed resource prices and schedule a VNet request at the next possible occasion. A VNet request will require a constant resource rate, and cannot be stretched or shortened in time once it is started. Therefore, we use a simple greedy algorithm which computes earliest embeddings on the providers, together with the corresponding prices.

In our experiments, we have three different providers whose prices differ by a certain percentage. VNet requests arrive over time according to a Poisson distribution with exponentially distributed inter-arrival times (parameter $\lambda = 1$). The request durations follow a heavy tail distribution (*Pareto* distribution with exponent α = 3, minimum 1, and scaling factor 100) and are chosen independently. The arrival and duration process yields a dynamic demand [3, 4]. For simplicity, the resource requirements of all requested VNets are identical and we will assume that customers re- quest one unit of volume. Providers have a capacity of 70 units.

2.1 Provider Perspective

We first study the impact of different customer flexibilities on the providers. In our model, the revenue of a provider depends on the number of customer requests it will eventually serve. We compare four different scenarios; three *homogeneous* ones where all customers have the same utility function (either *flexible* u_f, *equal* u_e or *inflexible* u_i), and a *heterogeneous* scenario where customers with different utilities (one half uses u_f and one half u_i) compete for provider resources.

Figure 2 shows the percentages of the demand assigned to the three providers: $P1$, $P2$, $P3$ in the three different homogeneous scenarios. In this experiment, the unit price of provider $P2$ is 10% higher than $P1$, and $P3$ is 20% higher than $P1$. Provider $P3$ has a lower workload, as the aggregate demand does not always fill all provider resources. The share of demand on $P1$ changes only slightly over the scenarios. Most of the demand that $P3$ gains (approx. 2% of the overall demand) while increasing the importance of the time dimension, is stolen from the $P2$'s share. That is because customers prefer $P3$ at demand peak rates, rather than having to wait for $P2$. The heterogeneous scenario shows a behavior similar to the equal scenario and is hence omitted here.

Given a certain degree of flexibility (e.g., customers with time uncritical applications such as bulk data transfers or batch jobs), the variance in demand can be exploited to shift load in time. While $P1$'s capacity is completely used $> 99\%$ of the time, the workloads of $P2$ and $P3$ vary. An increase of the *Poisson* arrival rate to $\lambda = 1.2$ leads to a more frequent demand excess, and $P2$ hardly has available capacities over longer time periods. Also $P3$ obtains approx. a quarter of the overall demand. If not stated differently, in this paper, we will focus on demands similar to the top scenario where there are sufficient capacities even in high-demand periods.

		deadline		
λ	median	1%	10%	20%
1.2	182	0.20	0.14	0.01
1.3	193	1.54	0.84	0.1
1.4	209	5.34	4.09	1.89
1.5	224	11.29	10.89	9.82

Table 1: Percentage of requests that cannot be served within their deadline. The deadline occurs after the duration plus a percentage (1%,10% or 20%). The arrival process is *Poisson* with parameter λ. We show the median demand (total capacity is 210).

With increasing demand, having *strict* deadlines becomes more critical: later, a result may no longer be useful. Table 1 shows the percentage of requests that cannot be embedded within their deadlines. In this experiment, the deadlines are chosen as a function of the corresponding request duration: we have three different flexibility levels, one adding 1% to the duration, one 10%, and the most flexible one adds 20%. With a Poisson arrival parameter $\lambda = 1.5$, the percentages are nearly equal, and independent of the deadline flexibility. This is due to the demand excess for nearly all time periods. In general strict deadlines are beneficial for $P3$, as it increases the fraction of customers that cannot wait.

2.2 Customer Perspective

Next we examine the customer perspective. Clearly, the pricing scheme on the market combined with the current resource demand and also the flexibilities of the other customers, will influence the obtained customer utilities.

Time is money. In order to compare different pricing schemes, we examine customers with no specific preference on time or price (utility function u_e). The pricing schemes on the competitive provider market lead to different decisions on how long a

request waits for being embedded and what to pay. We investigate the customer utilities on three different pricing scenarios with $1\%, 10\%$ and 25% difference between two providers. Figure 3 shows also the respective waiting times. Comparing these plots one must be aware that an increase of the prices also reduces the average utilities. Surprisingly, despite the fact that the cheapest price stays constant over all scenarios, the best utility (no waiting time on first provider) is only reached in the 1% scenario. This is because of the small price difference for customers who do not wait and rather pay the small overhead for embeddings on $P2$ and $P3$. With higher differences, the customers are more likely to wait for free capacities on cheaper providers, which leads to the longer waiting times and queues on those providers, and eventually prevents an immediate embedding there. This also explains why the utilities in the first plot can be classified into three groups with nearly identical utilities. The increasing price variance renders the differences larger and leads to even smaller utilities (up to -50% from the maximum) at peak times, and a wider overall distribution. Unsurprisingly, the waiting time plots show a change of the slope at the points where the time matches its utility-wise analog price value (1, 10, 25). The slope is steeper before these points since there are many customers who prefer to wait for the next cheaper provider.

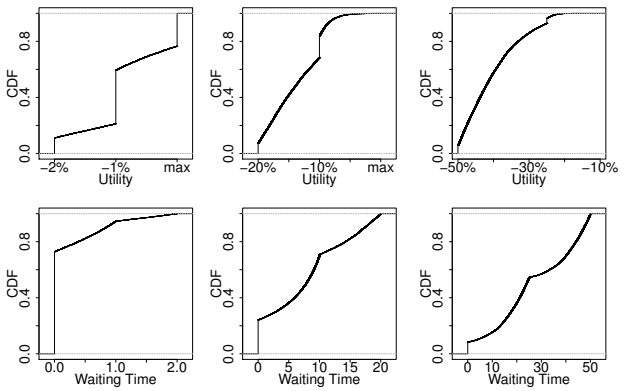

Figure 3: CDFs of customer utilities and their respective waiting times given three different pricing scenarios ($1\%, 10\%, 25\%$ price differences between the providers). The utilities are given in a relative percentage to the theoretically highest utility. The customers are not specifically flexible or inflexible.

Dependency on other customers. The utilities obtained by the customers are inter-dependent. To study these dependencies, we extend our setting to a heterogeneous one where the flexibilities of the customers are mixed. The comparison of these utilities (cf Figure 4) shows that on the one hand mixed utility functions are beneficial for flexible customers and increase their utility by $\sim 10\%$. On the other hand, mixed utility functions are reducing the utilities for inflexible customers. While in a homogeneous scenario the customers tend to choose a more expensive provider early, which keeps the waiting times low, the flexible customers in the heterogeneous scenario are willing to wait longer for capacities on the cheapest provider. This leads to a situation where the flexible customers reserve the capacities on the cheapest provider while the inflexible customers are stuck with the expensive providers. Since this impacts only approximately 50% of the inflexible customers whose utilities decrease by $\sim 10\%$, the overall utility under heterogeneous demands increases.

3. BENEFITS IN VERTICAL MARKETS

Let us study the effects of flexibility in the vertical model: we assume the customers send their VNet requests to a broker who resells resources from the cloud provider. The business model of the broker is to buy large resource contracts from the cloud provider, and uses these resources to satisfy multiple VNet requests. Concretely, we assume that the cloud provider offers a single kind of resource, and that contracts are given in terms of a resource rate R and a duration D. The higher the product of resource rate and duration, henceforth called the *area* $A = R \times D$ of the contract, the lower the unit price. We define the *discount* δ as the factor by which a contract of twice the area is more expensive: $\delta = 1.5$ means that for a twice as large contract, the price is 50% higher; $\delta = 2$ implies no discount is given and $\delta = 1$ means infinite discount.

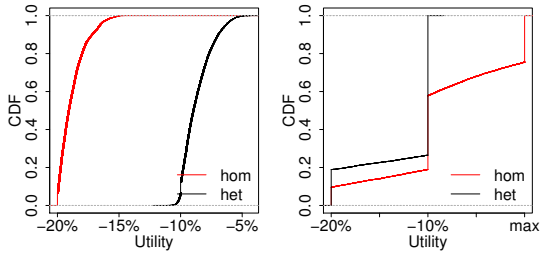

Figure 4: CDFs of customers utilities in scenarios with heterogeneous and homogeneous demands. The utilities are compared based on the user flexibilities (*left:* flexible only, *right:* inflexible only). The utilities are given as a percentage of the theoretical maximum utility.

We assume that each VNet request $vnet$, arriving online at time $vnet$.arrival(), specifies a constant resource rate $vnet$.rate(), a duration $vnet$.dur(), and a deadline $vnet$.dead() by which it must be completed. We focus on a simple deadline utility function in the sense that the customer only cares whether the deadline is met or not, and accordingly pays a fixed price or nothing. Thus, to maximize revenues, the broker should embed as many VNets as possible which meet the deadline. We define the flexibility F of a $vnet$ request as $F = (vnet.\text{dead}() - vnet.\text{arrival}())/vnet.\text{dur}()$, the factor by which the feasible embedding time period exceeds the duration. (We will assume that $F \geq 1$.)

The broker can benefit from delaying VNet requests if the time period until their deadline is relatively large compared to the VNet duration: Then, a larger resource contract can be bought at a better discount. It seems that the broker can reap all benefits from more flexible deadlines relative to the the VNet durations. Moreover, the flexibility gains accruing at the broker translate into a corresponding income loss at the cloud provider, as contracts become cheaper: a zero-sum game. However, as we will see, the situation is slightly more complicated and the benefits depend on the variance in demand. Moreover, the distribution of benefits of course also depends on the strategy by which the broker delays requests and buys resources. We compare three natural broker strategies: A greedy strategy called INSTANT where the broker immediately buys a new contract specifically for each incoming VNet request, and two strategies ROOF and STAIRS where the broker uses a VNet buffer B to delay requests and buy larger contracts.

STAIRS and ROOF differ in when and how the buffer B is filled. Whenever a VNet $vnet$ arrives, STAIRS includes it in the buffer B. For each time step Δt, STAIRS checks if the VNets in B can be further delayed. If one of the VNet requests $vnet \in B$ can no longer be delayed and must be embedded the latest at the current time t,

STAIRS groups the VNets in the buffer into *intervals*, all starting at t, and buys contracts for all these networks. In contrast, ROOF first checks if there are spare capacities available from previously bought contracts. If this is the case, these capacities are used for the embedding of $vnet$. Otherwise, the broker delays the embedding of $vnet$ and adds it to the buffer B. If one of the VNets in B cannot be delayed any longer, ROOF buys a single large contract (i.e., it empties the entire buffer).

Figure 5 (*left*) plots the total contract price paid by ROOF and STAIRS, as a function of the contract discount and relative to the INSTANT price which serves as a baseline. We assume the discount model for the product of contract duration D and resource rate R, i.e., for the area $A = R \times D$: for a twice as large area, the price is between one and two times higher (the former implies free resources, the latter no discount). As expected, ROOF performs bad without discounts and benefits from buying large contracts if discounts are high. STAIRS is always at least as good as INSTANT (equal in the no discount scenario). However, under high discounts, STAIRS pays relatively more again. This can be explained by the overall decreased cost of INSTANT.

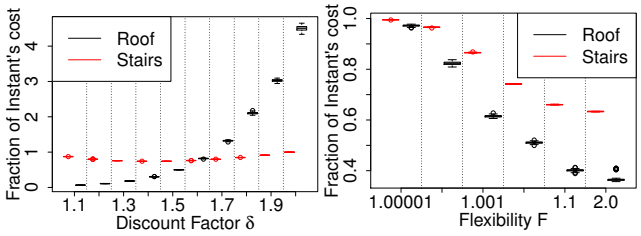

Figure 5: *Left:* Price paid by STAIRS and ROOF relative to INSTANT's price as a function of the discount factor for twice as large contracts. The arrival times are generated using a $\lambda = 1.2$ *Poisson* distribution and the durations are generated according to a *Pareto* distribution with $\alpha = 2$. The flexibility factor is $F = 1.01$. *Right:* Price paid by STAIRS and ROOF relative to INSTANT's price as a function of F. (Arrival times $\lambda = 1.2$, durations $\alpha = 2$, discount factor $\delta = 1.5$)

Broker Perspective. Figure 5 (*right*) shows the costs of ROOF and STAIRS relative to INSTANT's costs. Note that since INSTANT does not benefit from flexibilities (it does not delay any requests), INSTANT's costs can be used as a baseline and the fraction of its costs can be regarded as the *benefit of flexibility*. This benefit is plotted as a function of the flexibility ratio (overall feasible time period divided by request duration). We see that for very low flexibilities, the strategies do not differ much: on average, only roughly 5% of the VNet requests are not immediately embedded. The fraction of delayed VNets increases with higher flexibilities F, e.g., to 75% (for $F = 1.001$) resp. to 95% (for $F = 1.01$). Using the ROOF strategy, the broker will benefit more since the already bought resources can be used to a greater extent.

Provider Perspective. The provider perspective is similar: the flexibility benefits that could be exploited by the broker automatically translate into an additional revenue for the cloud provider; we have a zero-sum game. Interestingly, these benefits also depend on the variance of the demand. Table 2 shows the price paid by the broker (and thus the income of the cloud provider), under different variances of the VNet duration, i.e., different α values in the Pareto distribution (note that larger α also increases the demand, cf Table 2 INSTANT). In general, we observe that a higher variance benefits the cloud provider: resources cannot be delayed efficiently. Interestingly, a higher variance does not necessarily lead to a higher price for the broker using the ROOF strategy however. This can be explained by the large discount in these cases.

Pareto α	1.5	2	2.5	3
INSTANT	12109	11503	11183	10951
STAIRS	10264	9143	8493	7943
ROOF	1105	1291	1248	1190

Table 2: Mean provider income given VNets with different durations (a smaller *Pareto* α means a higher variance). Inter-arrival times $\lambda = 1.2$, discount $\delta = 1.2$, flexibility $F = 1.01$.

4. CONCLUSION

This paper initiated the study of the benefits and beneficiaries of specification flexibilities and introduced two most simple market models. In horizontal markets, we find that customers can often benefit from being more flexible in some dimension (e.g., execution time) in the sense that the service is improved in the other dimension. Moreover, more flexibility on the customer side disproportionally benefits the cloud providers offering cheaper prices. Interestingly however, it is often not the cheapest provider that benefits the most from customer flexibilities. Finally, we also find that while the social welfare of both customers and providers is increased under heterogeneous requirements and flexibilities, not all customers can benefit in this setting; inflexible customers may even be worse off. Overall, more resources are needed in case of inflexible customers, opening new business opportunities for providers.

In vertical markets, we observe that the benefits from customer flexibilities typically accrue at the broker, which can bundle requests and exploit potential discounts on the provider side. However, depending on the broker strategy, a higher variance in the request demands helps the cloud providers to increase their revenues.

Acknowledgments. This research was supported by the BMBF (01IS12056).

5. REFERENCES

[1] V. Abhishek, I. A. Kash, and P. Key. Fixed and market pricing for cloud services. In *Proc. NetEcon Workshop*, 2012.

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

[3] T. Benson, A. Akella, and D. A. Maltz. Network traffic characteristics of data centers in the wild. In *Proc. 10th ACM IMC*, 2010.

[4] Y. Chen, S. Alspaugh, and R. Katz. Interactive analytical processing in big data systems: a cross-industry study of mapreduce workloads. *Proc. VLDB*, 5(12), 2012.

[5] M. K. Chowdhury and R. Boutaba. A survey of network virtualization. *Elsevier Computer Networks*, 54(5), 2010.

[6] V. Jalaparti, H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron. Bridging the tenant-provider gap in cloud services. In *Proc. 3rd ACM SoCC*, 2012.

[7] A. Ludwig, S. Schmid, and A. Feldmann. Specificity vs. flexibility: On the embedding cost of a virtual network. In *Proc. 25th ITC*, 2013.

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

[9] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowtron. Better never than late: meeting deadlines in datacenter networks. In *Proc. ACM SIGCOMM*, 2011.

[10] Z. Wu, C. Yu, and H. V. Madhyastha. Costlo: Cost-effective redundancy for lower latency variance on cloud storage services. In *Proc. 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2015.