Neutrality in Future Public Clouds: Implications and Challenges[*]

George Kesidis, Bhuvan Urgaonkar, Neda Nasiriani, and Cheng Wang
School of EECS, Pennsylvania State University
University Park, PA, 16803, USA
{gik2,buu1,nun129,cxw967}@psu.edu

Abstract

With public cloud providers poised to become indispensable utility providers, neutrality-related mandates will likely emerge to ensure a level playing field among their customers ("tenants"). We analogize with net neutrality to discuss: (i) what form cloud neutrality might take, (ii) what lessons might the net neutrality debate have to offer, and (iii) in what ways cloud neutrality would be different from (and even more difficult than) net neutrality. We use idealized thought experiments and simple workload case studies to illustrate our points and conclude with a discussion of challenges and future directions. Our paper points to a rich and important area for future work.

1 Introduction

A natural but relatively little addressed set of concerns in the emerging public cloud utility comes to the fore when one compares its likely evolution with that of Internet Service Providers (ISPs). During the period of commercialization of the Internet in the 1990s (after the creation of the WWW), the principle of over-engineering reigned and cheap access plans without quotas were commonplace. In the 2000s, the Internet access marketplace matured, ISPs consolidated, and it became apparent that the Internet was (i) prone to congestion by activity that generated little revenue[1], and (ii) created a platform for stiff competition for the ISPs' own profitable "managed" video and voice services.

The network neutrality debate began when Comcast throttled BitTorrent activity that was congesting its (broadcast-based CMTS) residential broadband service. Though this violation of application neutrality targeted activity that was largely piracy of copyrighted material, third-party content and service providers viewed it as a dangerous precedent considering that ISPs are themselves competing providers of content (particularly video) and services (*e.g.*, telephony). In 2014, Neflix negotiated payment for "fastlanes" to reach Comcast's subscribers [32], but this would have basically been a side payment (albeit willing), thus violating origin neutrality. The debate continues: a federal court limited certain of the FCC's neutrality rules in 2014 [31] and the FCC deemed the Internet a utility in 2015 [25] in a move to consolidate neutrality rules.

Somewhat similarly, the public cloud market was also plentiful initially compared to demand. As the business has grown over the last decade, competition has intensified with an increased diversification in the types of offered service-level agreements (SLAs) and associated price-performance trade-offs [4, 30, 8, 19, 13]. It is reasonable to anticipate that issues of fair competition similar to those for ISPs may be on the horizon. As an illustrative example, consider how Amazon Prime's video service and Netflix both use Amazon EC2.

What might neutrality mean in this setting? Do lessons and challenges from net neutrality suffice or does the public cloud raise novel issues? We believe the answer is that it does and this is the context of our paper.

Contributions and Outline: In Section 2.1, we identify several ideas and lessons from the net neutrality debate that offer useful starting points for discussing cloud neutrality. A public cloud is, however, a significantly different resource provider than an ISP. In fact, it is significantly more complicated in some ways, presenting novel concerns that we discuss in Section 2.2. Following this, in Section 3, we take preliminary steps towards exploring these issues using simple thought experiments and empirical case studies. Finally, we conclude in Section 5.

2 Background and Related Work

Discussion on public cloud neutrality is relatively scant or preliminary [24]. Interxion, a European provider defines a carrier- and cloud-neutral "colo" data center as follows: *"A truly neutral data centre provider is one that*

[*] This research supported in part by NSF CNS grants 1228717 and 1526133 and by NSF CAREER grant 0953541.

[1] In the case of illegal file-sharing, negative revenue from the point of view of copyright holders.

is independent of the companies colocating in the data centre, does not compete with them in any way, and offers no packaged services as part of colocation. Customers are free to contract directly with the providers of their choice" [1]. Our interest herein is with public cloud providers that offer more general services than colos (IaaS clouds offering virtualized IT resources or PaaS/SaaS clouds offering even more abstract services), for which the neutrality discussion is much more complex and unclear.

2.1 Lessons from Net Neutrality

The following insights and guiding principles from the net neutrality debate offer good starting points for our discussion of cloud neutrality.

Resource congestion: If a utility provider can offer adequate resources to satisfy its customers - especially when it is in its initial growing phase e.g. ISPs in the 90s or public clouds till very recently - the neutrality concern is moot. Resource congestion, whether actual or alleged, is the basic context of any neutrality debate.

Neutrality modulo SLAs: Neutrality does not mean that every tenant is treated equally. The fair/equal treatment across tenants may only relate to tenants that have chosen identical SLAs. Therefore, it only ever makes sense to discuss neutrality in the context of specific SLAs. In other words, a key aspect of a provider's resource allocation when assessing its neutrality is how it allocates any *discretionary resources* (i.e., resources left over after it has met its SLA obligations) among tenants in the same SLA class. This view will play an important point in our case studies in Section 4.

Information limits and preferences: The providers behavior must not be based on "inside information" or "preferences." E.g., Mogul et al. [22] consider a network bandwidth allocation problem wherein a cloud provider reckons tenant sensitivity to network bandwidth underprovisioning and uses this to allocate bandwidth differentially to improve its profits. One may consider such a cloud provider non-neutral.

Fair competition: Fair competition needs to be upheld between affiliates of the cloud provider and its tenants - whether direct (e.g. recall from Section 1 how Amazon Prime and Neftlix both use Amazon EC2) or indirect. That is, neutrality is related to antitrust.

2.2 Why Cloud Neutrality is Different

Whereas ideas discussed above are useful in framing our cloud neutrality discussion, we believe that there are aspects of cloud operation that make for a more complicated situation than net neutrality. We find two key sources of difficulty:

Lack of an effective common "currency": Unlike network bandwidth, the basis of SLAs in the case of net neutrality, it is difficult to identify a single (virtual) resource/currency that can act as an effective proxy for the multiple physical resources that must be allocated for colocated tenants. Different workloads need different resources and even within a workload this could change over time. It is not clear how neutrality should be defined in such a multi-resource setting. Existing literature on multi-resource fair scheduling offers one appealing option and we explore this in Section 3.

Difficulty of auditing resource usage: An important requirement for implementing a neutral utility is the ability to audit resource allocations/usage for verifying adherence to neutral behavior. Whereas network bandwidth allocated by an ISP can be effectively monitored and audited [16, 17], the same does not hold for many virtualized resources offered by a public cloud. As identified by others [23, 29, 2, 3], auditing virtualized resources is inherently difficult. Existing solutions for measuring/auditing virtualized IT resource usage (e.g., resource containers [6]) implicitly assume trusted resource managers. A hypervisor may over-commit CPUs to multiple VMs, which makes it difficult to detect/quantify each VM's shares of CPU capacity (CPU cycles) from within the guest OS. As an extreme example, a hypervisor may "dilate" a VM's notion of time by slowing its delivery of virtualized timer interrupts [15]. Netflix relies upon the "stolen time" (a measure of competition for CPU) reported by the hypervisor underlying its Amazon EC2 instances for identifying incidents wherein its procured instances are not getting adequate CPU capacity [20]. What if the hypervisor (either deliberately or due to a bug) misreported this metric? Similar difficulties apply to other resources such as memory and IO bandwidth. Implementing cloud neutrality would require effective solutions to this auditing problem.Our discussion in the rest of the paper assumes the existence of such solutions.

3 Towards a Definition of Cloud Neutrality

The notion of fair scheduling is a natural starting point for formalizing neutral resource management. Given the information limits inherent in neutrality as well as the general difficulty of application performance modeling by a cloud provider (even by the tenant, i.e., the application owner), we believe that SLAs in a neutral cloud would have to be in terms of effective resource capacities not in terms of application-specific performance metrics (e.g., [18, 26] for the latter). In multi-resource environments, a generalization of max-min fairness, Dominant Resource Fairness (DRF), can determine resource allocations to a user based on its maximum weighted share among its received resources. DRF maximizes the minimum dominant share of all users and has several desirable properties, e.g., sharing incentive, strategy proofness, Pareto efficiency [12]. There are other fair allocation policies for multiple resources, e.g., Competitive Equilibrium from Equal Incomes (CEEI), that possess some of these desirably properties. In the following, we assume an SLA class is characterized by a deterministic guaranteed element in combination with a statistical (possibly best effort) element. Our SLA class definition is fairly general and expressive. E.g., on-demand instances offered by Amazon EC2 or Google Compute Engine amount to SLAs with guaranteed resource capacities (the advertised capacities) whereas EC2's burstable or spot instances or Google's preemptible instances can be thought of as having different degrees of guaranteed plus best-effort resource capacities [5, 7, 14].

3.1 SLAs with Deterministic Guarantees

When costs are low, tenant demand will exceed available resources. We assume that the tenants will contend for guaranteed (on-demand) resources, as stipulated in their SLAs, and the cloud will arbitrate. For all tenants i, consider linear models of their net-benefit/utility functions based on workload intensity x, as set by the cloud, leading to *assumed positive* linear tenant net-utilities as follows: $u_i(x_i) := x_i(V_i - \sum_k d_{i,k}\tilde{P}_k)$, where: V_i is the benefit per unit workload, $\tilde{P}_k$ is the cost per unit resource of type k (*e.g.*, $\tilde{P}_k = 1/R_k$ for kind of asset fairness [12] where R_k is the amount of type-k resource available), and $d_{i,k}$ is the type-k resource demand per unit workload. Consider tenants i with linear net-utilities,

$$\forall i,\ \partial_i u_i \ = \ V_i - \sum_k d_{i,k}\tilde{P}_k \quad > \quad 0.$$

The resulting game will result in a set of Nash equilibria[2] on the feasibility boundary where at least one resource is exhausted by tenants[3].

As in Nash bargaining problems, there is a choice of boundary equilibria. A leader of the game (operator/provider of the cloud itself, or a government/market regulator) may desire equilibrium points that, *e.g.*, maximize: cloud revenue, $\Omega := \sum_i x_i \sum_k d_{i,k}\tilde{P}_k$; social welfare, $\Omega := \sum_i u_i(\underline{x})$; or total tenants' benefit, $\Omega := \sum_i V_i x_i$.

To such ends, resources may be shifted among tenants to control their Nash equilibrium, *i.e.*, the cloud takes direct resource allocation actions at or near the feasibility region (as in the CEEI mechanism [12]). Note that the three previous example objectives are planar. They result in Nash play-actions $\underline{x}$ with maximal Ω corresponding to corner points or line segments of the feasibility region, recall the simplex algorithm [27]. Also note that a for-profit cloud with congested (low priced) resources would naturally choose to maximize revenue.

Alternatively, resources $\underline{x}$ can be (maximally) allocated subject to rules of "fairness" [9]. For example, equal dominant resource share, $x_i \max_k d_{i,k}/R_k$ (leading to DRF) [12]; equal total asset-fraction share, $x_i \sum_k d_{i,k}/R_k$ (leading to a kind of asset fairness); or equal per unit net utility, $x_i/(V_i - \sum_k d_{i,k}\tilde{P}_k)$.

Note that notions of fairness may not separately consider the interests of the cloud and tenants particularly in a public, for-profit cloud setting with potentially competing tenants (possibly serving their own customers). Also, tenants can be differently weighted (as considered in *e.g.*, Sec. 4.3 of [12]); such weights could correspond to priority in a private cloud or enterprise network, or willingness to pay in a public cloud system.

See Figure 1 for an illustrative example indicating DRF and maximum cloud revenue with: total resource pool $\underline{R} = (9\text{ CPUs}, 6\text{GB RAM})$ and tenant demand vectors $\underline{d}_1 = (1,2), \underline{d}_2 = (3,1)$. For DRF, the dominant resource shares $x_i \max_k d_{i,k}/R_k$ of the two tenants $i \in \{1,2\}$ are equated - in this case, $x_1 2/6 = x_2 3/9$, *i.e.*, $x_1 = x_2$. But with $\tilde{P}_k \equiv 1/R_k$, $\Omega = (\frac{2}{6}+\frac{1}{9})x_1 + (\frac{1}{6}+\frac{3}{9})x_2$.

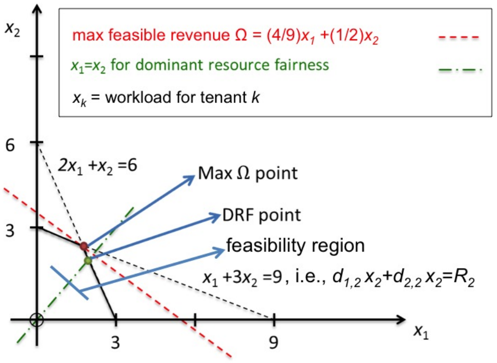
x_2
max feasible revenue Ω = (4/9)x_1 +(1/2)x_2
x_1=x_2 for dominant resource fairness
x_k = workload for tenant k
6
3
2x_1 +x_2 =6
Max Ω point
DRF point
feasibility region
x_1 +3x_2 =9, i.e., $d_{1,2}x_2$+$d_{2,2}x_2$=R_2
3
9
x_1

Figure 1: Illustrative example indicating DRF and max total tenant valuation objectives.

Here, the DRF point is arguably a more "neutral" way (than maximum cloud revenue) to treat this congested scenario of tenants consuming resources with resource proportions stipulated in their SLAs.

3.2 SLA with Statistical Guarantees

Now suppose that the tenants opt for cheaper SLAs involving only statistical resource guarantees. In particular, this means that the cloud may exploit unused resources nominally assigned to one tenant for the benefit of other tenants, including to overbook resources to take advantage of statistical multiplexing. Moreover, the cloud may have discretionary (unreserved) resources that it may assign to tenants. Again, the assumption is that actual tenant demand will exhaust resources.

[2] A Nash equilibrium $\underline{x}^* = \{x_i^*\}$ is a stalemate from which no single tenant i can improve their utility u_i by changing their workload to $x_i \neq x_i^*$, *i.e.*, any unilateral defection from a Nash equilibrium will not profit the defector.

[3] At the convex feasibility boundary with at least one resource exhausted, the only *feasible* unilateral move by any player i is to reduce demand (x_i), hence utility $u_i = x_i \partial_i u_i$ is reduced if, as assumed, marginal utility $\partial u_j > 0$ for all tenants j.

Now let $m_{i,k} = \mathsf{E}d_{i,k}$ and $\sigma^2_{i,i,k} = \text{var}(d_{i,k})$ respectively be the mean and variance of the demand per unit workload $d_{i,k}$ for the k^{th} resource by the i^{th} tenant. Also, let $\sigma^2_{i,j,k}$ be the covariance of demand per unit workload by tenant i and tenant j for resource k and define the covariance matrix $\mathbf{C}_k = [\sigma^2_{i,j,k}]$. In practice, these quantities would be empirically estimated online. We can extend our model of resource allocation by using nonlinear "chance constraints" for each resource k, leading to a convex feasibility region[4],

$$\underline{x}'\underline{m}_k + n_k\sqrt{\underline{x}'\mathbf{C}_k\underline{x}} \quad \leq \quad R'_k \; < \; R_k, \qquad (1)$$

where $n_k \geq 1$ is a confidence factor for the headroom $R_k - R'_k$ (*e.g.*, [10]) corresponding to $\mathsf{P}(\sum_j x_j d_{j,k} \geq R'_k) \leq \varepsilon_k$, and for all k, $x_k \geq 0$ of course. Run-time estimates of mean, variance and covariance of demand and dynamic calibration of resource headroom can be jointly used to deal with uncertain, time-varying demand needs, particularly when infeasible overages in demand must be rare ($\varepsilon_k \ll 1$). Headroom will be important in the presence of estimation error in these statistical parameters of demand.

If we take (conservatively with $n_k = 2$) *deterministic*

$$d_{i,k} \quad := \quad m_{i,k} + n_k\sigma_{i,i,k} \qquad (2)$$

(quantities that could be involved in tenant i's SLA with the cloud), we can then define statistical multiplexing gain for resource k at demand $\underline{x}$ as

$$n_k\left(\sum_i x_i\sigma_{i,i,k} - \sqrt{\underline{x}'\mathbf{C}_k\underline{x}}\right) \quad \geq 0,$$

i.e., capturing the difference between "deterministic" provisioning by the cloud using (2) and that using (1).

See Figure 2 for an example uncorrelated case with $n_k \equiv 2$ and $\sigma_{i,i,k} \equiv 0.5$. For resource $k = 1$ with capacity 8, $d_{1,1} = 2 + 2 \cdot 0.5 = 3$ and $d_{2,1} = 1 + 2 \cdot 0.5 = 2$. Similarly, $d_{1,2} = 2$ and $d_{2,2} = 3$ for the resource $k = 2$ with capacity 11. The corresponding feasibility region, with piecewise linear boundary $x_2 = \min\{(8-3x_1)/2,\ (11-2x_1)/3\}$, is *contained* in that given by the chance constraints, with shown strictly concave boundary meeting the piecewise linear boundary at the axes.

In practice, it can be reckoned at run-time (online) whether different workloads are negatively correlated, or whether statistical multiplexing gains can be achieved by scaling a given workload [21, 28]. A basic assumption here is that workloads are sufficiently stationary so that present estimates of correlation are valid in the near future. Moreover, correlations need to be assessed *jointly* among the plurality IT resources used by the workloads in question. Note that a cloud assesses and exploits workload correlations to overbook under resource-oriented, not performance-oriented, SLAs.

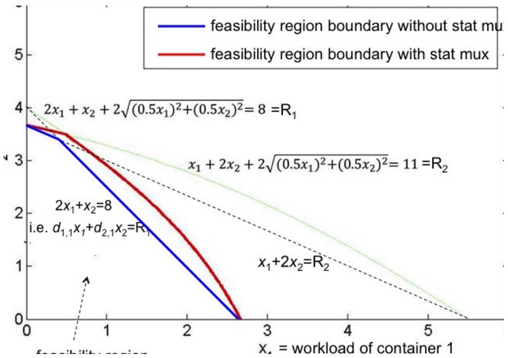

Figure 2: Illustrative example of chance-constrained feasibility region with uncorrelated resource demands.

Table 1: Tenants' latency (norm. against individual tenant's acceptable latency). The ratios imply the relative discretionary DRAM allocations for the two tenants.

	CFS 3:7		CFS 1:1		Resv 1:1	
	95-th	avg	95-th	avg	95-th	avg
Tenant 1	413	315	436	304	417	241
Tenant 2	389	265	410	311	429	305

4 Illustrative numerical results for individual resources

We consider two memcached tenants whose CPU and memory needs we manipulate to illustrate neutrality issues under different representative cloud resource management options.

4.1 CPU Case Study

Each tenant runs Memcached, an in-memory key-value data store, with workload generated by YCSB [11] in time-varying fashion according to i.i.d. Gaussian processes with identical mean but the variation of tenant 2 is double that of tenant 1. The tenants have the same SLAs corresponding to one guaranteed core, and the cloud has a discretionary third core that it can share between the two tenants. Empirically, we found that a single core can achieve throughput of 75k ops/s with satisfactory mean

Table 2: Latency of tenants for different allocation of discretionary CPU, uncorrelated demand scenario.

	CFS 3:7		CFS 1:1		Resv 1:1	
	95-th	avg	95-th	avg	95-th	avg
Tenant 1	484	318	449	382	451	365
Tenant 2	418	254	435	284	444	299

[4] Recall that covariance matrices are always positive semi-definite with positive diagonal entries. Using the fact that covariance matrices are also symmetric, convexity of the feasibility region is a direct consequence of the Cauchy-Schwarz inequality, and that the intersection of convex regions (one for each resource k) is convex.

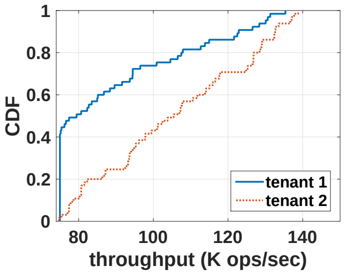

Figure 3: CDF of memcached throughput conditioned on > 75k ops/s.

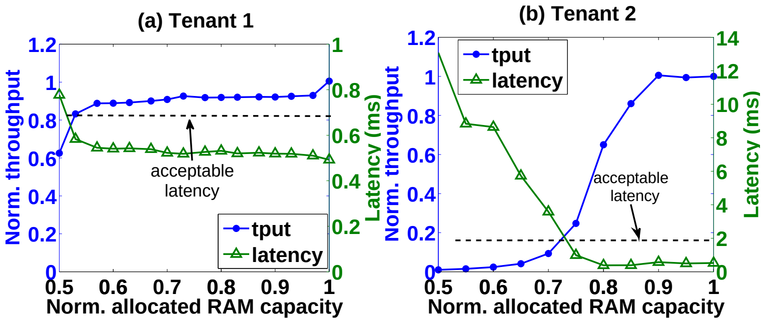

Figure 4: Memcached performance as a function of allocated memory (norm. against workingset size).

response times of 400us. Figure 3 shows the CDF of the throughputs of the two tenants conditioned on each of them being greater than 75k ops/s, *i.e.*, emulating a case where the third discretionary core is active when both tenants have exhausted their respective dedicated cores.

The cloud may employ a weighted reservation (Resv) based approach or a work-conserving, proportional-share (CFS) scheduling for the discretionary third core. The Resv1:1 (equally weighted) and CFS1:1 schemes are arguably neutral as they are based only on parameters in the tenant SLAs, *i.e.*, the same mean demands in this example corresponding to a single core; while CFS 3:7 is based on measured (conditional) demand-variation of the two tenants which is not part of their SLAs and so this scheme is arguably not neutral. If the intention is to prevent tenant 2 (the one with more demand variation) from defecting to other cloud-services providers, the cloud may tend to adopt CFS1:1. On the other hand, the cloud may want to entice tenant 2 to renegotiate a more costly SLA by adopting Resv1:1. The premise here is that with greater demand variability, tenant 2 will outcompete tenant 1 for the discretionary CPU core; but this may not be entirely the case when the tenant demands are positively correlated, *i.e.*, how workload is consolidated by the cloud will impact such "iso-neutral" decision-making. In Tables 1,2, the more demand-variable tenant 2 has best latency performance (both in mean and 95 percentile) under non-neutral CFS 3:7. Note that there is improved 95-percentile latency performance for tenant 2 under CFS1:1 over Resv1:1 but not for mean latency - this is attributable to simulated *positive correlation* in tenant demand for the results of Table 1. For the uncorrelated case, tenant 2 has incremental decrease in its latency from Resv1:1 to CFS3:7, both in mean and 95-percentile.

4.2 Memory Case Study

To illustrate neutrality concerns using memory as the resource, we create identical maximum memory needs for our two Memcached tenants. Tenant 1 has exponential key popularity distribution: 95% requests go to 5% of working set; tenant 2 has Zipfian key popularity distribution. We assume that they both pick an SLA wherein 3 GB RAM capacity is guaranteed (50% of the working set) and they both stipulate their maximum needs of 6 GB. However, tenant 1 has a much smaller set of popular data ("hot" data) than tenant 2, making it far less sensitive to memory underprovisioning by the cloud (as shown in Figure 4). We compare tenant performance under two different memory allocation strategies the cloud may employ in Table 3: Under neutral underprovisioning, the cloud provider may not exploit its understanding of these different sensitivities to differentially allocate memory to the tenants whereas in a non-neutral scenario, it may underprovision tenant 1 far more aggressively (without any perceived performance difference by tenant 1) to reduce its own operational costs.

Table 3: Latency (norm. against the target latency of individual tenant) of tenants for different allocation of discretionary DRAM capacity. The ratios imply the relative discretionary DRAM allocations for the two tenants.

	Non-neutral (0.4:0.6)		Neutral (1:1)	
	95-th	avg	95-th	avg
Tenant 1	0.97	0.68	0.96	0.66
Tenant 2	1.01	0.83	0.99	0.91

5 Conclusions

We identified novel challenges that would arise in a future neutral public cloud. We identified four lessons from the net neutrality debate that offer a good starting point for discussion about cloud neutrality. We then identified two particular aspects of cloud neutrality that would require novel thought and debate. Using simple thought experiments, we considered whether notions of multi-resource fairness such as DRF are useful ways of defining cloud neutrality. Finally, we created simple case studies with different types of resource management and discussed how these mapped to our notions of neutrality and the implications on the provider's profitability and the tenant's costs.

References

[1] Interxion. http://www.interxion.com/, 2016.

[2] Aceto, G., Botta, A., De Donato, W., and Pescapè, A. Survey cloud monitoring: A survey. *Comput. Netw. 57*, 9 (June 2013), 2093–2115.

[3] Alhamazani, K., Ranjan, R., Mitra, K., Rabhi, F., Jayaraman, P. P., Khan, S. U., Guabtni, A., and Bhatnagar, V. An overview of the commercial cloud monitoring tools: research dimensions, design issues, and state-of-the-art. *Computing 97*, 4 (2014), 357–377.

[4] Amato, A., and Venticinque, S. Multi-objective decision support for brokering of cloud sla. In *Proc. Advanced Information Networking and Applications Workshops (WAINA)* (March 2013), pp. 1241–1246.

[5] Amazon EC2 Pricing, 2014. `http://aws.amazon.com/ec2/pricing/#on-demand`.

[6] Banga, G., Druschel, P., and Mogul, J. C. Resource containers: A new facility for resource management in server systems. In *Proc. USENIX OSDI* (1999).

[7] Barr, J. Aws official blog: New low cost ec2 instances with burstable performance. https://aws.amazon.com/blogs/aws/low-cost-burstable-ec2-instances/, 2014.

[8] Baset, S. A. Cloud slas: Present and future. *SIGOPS Oper. Syst. Rev. 46*, 2 (July 2012), 57–66.

[9] Brams, S., Kilgour, D., and Klamler, C. How to share things fairly. *MAA Mathematics Magazine 88*, 5 (Dec. 2015).

[10] Carvalho, M., Cirne, W., Brasileiro, F., and Wilkes, J. Long-term SLOs for reclaimed cloud computing resources. In *Proc. ACM Symposium on Cloud Computing (SoCC)* (Seattle, 2014).

[11] Cooper, B. F., Silberstein, A., Tam, E., Ramakrishnan, R., and Sears, R. Benchmarking cloud serving systems with ycsb. In *Proc. 1st ACM Symposium on Cloud Computing (SoCC)* (2010).

[12] Ghodsi, A., Zaharia, M., Hindman, B., Konwinski, A., Shenker, S., and Stoica, I. Dominant resource fairness: Fair allocation of multiple resource types. In *Proc. NSDI* (2011).

[13] Gohad, A., Narendra, N. C., and Ramachandran, P. Cloud pricing models: A survey and position paper. In *Proc. IEEE Conf. on Cloud Computing in Emerging Markets (CCEM)* (Oct 2013), pp. 1–8.

[14] Google Cloud Pricing, 2016. `https://cloud.google.com/products/cloud-storage/`.

[15] Gupta, D., Yocum, K., McNett, M., Snoeren, A. C., Vahdat, A., and Voelker, G. M. To infinity and beyond: Time warped network emulation. In *Proc. ACM SOSP* (2005).

[16] Heinanen, J., Finland, T., and Guerin, R. A single rate three color marker. *RFC 2697 available at www.ietf.org* (1999).

[17] Heinanen, J., Finland, T., and Guerin, R. A two rate three color marker. *RFC 2698 available at www.ietf.org* (1999).

[18] Jayathilaka, H., Krintz, C., and Wolski, R. Response time service level agreements for cloud-hosted web applications. In *Proc. 6th ACM Symposium on Cloud Computing (SoCC)* (2015), pp. 315–328.

[19] Laatikainen, G., Ojala, A., and Mazhelis, O. Software Business. From Physical Products to Software Services and Solutions, June 2013.

[20] Link, D. Netflix and Stolen Time. http://blog.sciencelogic.com/netflix-steals-time-in-the-cloud-and-from-users/03/2011, March 25, 2011.

[21] Meng, X., Isci, C., Kephart, J., Zhang, L., Bouillet, E., and Pendarakis, D. Efficient resource provisioning in compute clouds via VM multiplexing. In *Proc. Int'l Conf. on Autonomic Computing (ICAC)* (2010), pp. 11–20.

[22] Mogul, J. C., and Kompella, R. R. Inferring the network latency requirements of cloud tenants. In *Proc. 15th Workshop on Hot Topics in Operating Systems (HotOS XV)* (2015).

[23] Why physical performance monitoring tools are not enough, 2016. http://searchservervirtualization.techtarget.com/tip/Why-physical-performance-monitoring-tools-arent-enough

[24] Renda, A. Competition, neutrality and diversity in the cloud. *Communications & Strategies*, 85 (2012), 23–44.

[25] Ruiz, E., and Lohr, S. F.C.C. Approves Net Neutrality Rules, Classifying Broadband Internet Service as a Utility. http://www.nytimes.com/2015/02/27/technology/net-neutrality-fcc-vote-internet-utility.html, Feb. 26, 2015.

[26] Serrano, D., Bouchenak, S., Kouki, Y., Ledoux, T., Lejeune, J., Sopena, J., Arantes, L., and Sens, P. Towards qos-oriented sla guarantees for online cloud services. In *Proc. IEEE/ACM Int'l Symp. on Cluster, Cloud and Grid Computing (CCGrid)* (May 2013), pp. 50–57.

[27] Simplex algorithm.
https://en.wikipedia.org/wiki/Simplex_algorithm.

[28] Wang, C., Urgaonkar, B., Birke, R., Gupta, A., Chen, L., Wang, Q., Kesidis, G., and Nasiriani, N. A case for dynamic pricing with effective capacity modulation in the cloud. In *Proc. IEEE Int'l Conf. on Autonomic Computing (ICAC)* (Wurzburg, Germany, July 2016).

[29] Ward, J. S., and Barker, A. Observing the clouds: a survey and taxonomy of cloud monitoring. *Journal of Cloud Computing 3*, 1 (2014), 1–30.

[30] Wu, L., Garg, S. K., and Buyya, R. SLA-Based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments. In *Proc. IEEE/ACM Int'l Symp. on Cluster, Cloud and Grid Computing (CCGrid)* (May 2011), pp. 195–204.

[31] Wyatt, E. Rebuffing F.C.C. in Net Neutrality Case, Court Allows Streaming Deals. http://www.nytimes.com/2014/01/15/technology/appeals-court-rejects-fcc-rules-on-internet-service-providers.html?_r=0, Jan. 14, 2014.

[32] Wyatt, E., and Cohen, N. Comcast and Netflix reach deal on service. http://www.nytimes.com/2014/02/24/business/media/comcast-and-netflix-reach-a-streaming-agreement.html?_r=0, Feb. 23, 2014.