Towards Predictable Datacenter Networks

Hitesh Ballani[†] hiballan@microsoft.com
Paolo Costa[†‡] costa@imperial.ac.uk
Thomas Karagiannis[†] thomkar@microsoft.com
Ant Rowstron[†] antr@microsoft.com

[†]Microsoft Research Cambridge, UK
[‡]Imperial College London, UK

ABSTRACT

The shared nature of the network in today's multi-tenant datacenters implies that network performance for tenants can vary significantly. This applies to both production datacenters and cloud environments. Network performance variability hurts application performance which makes tenant costs unpredictable and causes provider revenue loss. Motivated by these factors, this paper makes the case for extending the tenant-provider interface to explicitly account for the network. We argue this can be achieved by providing tenants with a virtual network connecting their compute instances. To this effect, the key contribution of this paper is the design of virtual network abstractions that capture the trade-off between the performance guarantees offered to tenants, their costs and the provider revenue.

To illustrate the feasibility of virtual networks, we develop Oktopus, a system that implements the proposed abstractions. Using realistic, large-scale simulations and an Oktopus deployment on a 25-node two-tier testbed, we demonstrate that the use of virtual networks yields significantly better and more predictable tenant performance. Further, using a simple pricing model, we find that the our abstractions can reduce tenant costs by up to 74% while maintaining provider revenue neutrality.

Categories and Subject Descriptors: C.2.3 [Computer-Communication Networks]: Network Operations
General Terms: Algorithms, Design, Performance
Keywords: Datacenter, Allocation, Virtual Network, Bandwidth

1. INTRODUCTION

The simplicity of the interface between cloud providers and tenants has significantly contributed to the increasing popularity of cloud datacenters offering on-demand use of computing resources. Tenants simply ask for the amount of compute and storage resources they require, and are charged on a pay-as-you-go basis.

While attractive and simple, this interface misses a critical resource, namely, the (intra-cloud) network. Cloud providers do not offer guaranteed network resources to tenants. Instead, a tenant's compute instances (virtual machines or, in short, VMs) communicate over the network shared amongst all tenants. Consequently, the bandwidth achieved by traffic between a tenant's VMs depends on a variety of factors outside the tenant's control, such as the network load and placement of the tenant's VMs, and is further exacerbated by the oversubscribed nature of datacenter network topologies [14]. Unavoidably, this leads to high variability in the performance offered by the cloud network to a tenant [13,23,24,30] which, in turn, has several negative consequences for both tenants and providers.

– *Unpredictable application performance and tenant cost.* Variable network performance is one of the leading causes for unpredictable application performance in the cloud [30], which is a key hindrance to cloud adoption [10,26]. This applies to a wide range of applications: from user-facing web applications [18,30] to transaction processing web applications [21] and MapReduce-like data intensive applications [30,38]. Further, since tenants pay based on the time they occupy their VMs, and this time is influenced by the network, tenants implicitly end up paying for the network traffic; yet, such communication is supposedly free (hidden cost).

– *Limited cloud applicability.* The lack of guaranteed network performance severely impedes the ability of the cloud to support various classes of applications that rely on predictable performance. The poor and variable performance of HPC and scientific computing applications in the cloud is well documented [17,33]. The same applies to data-parallel applications like MapReduce that rely on the network to ship large amounts of data at high rates [38]. As a matter of fact, Amazon's ClusterCompute [2] addresses this very concern by giving tenants, at a high cost, a dedicated 10 Gbps network with no oversubscription.

– *Inefficiencies in production datacenters and revenue loss.* The arguments above apply to not just cloud datacenters, but to any datacenter with multiple tenants (product groups), applications (search, advertisements, MapReduce), and services (BigTable, HDFS, GFS). For instance, in production datacenters running MapReduce jobs, variable network performance leads to poorly performing job schedules and significantly impacts datacenter throughput [7,31]. Also, such network-induced application unpredictability makes scheduling jobs qualitatively harder and hampers programmer productivity, not to mention significant loss in revenue [7].

These limitations result from the mismatch between the desired and achieved network performance by tenants which

SIGCOMM'11, August 15–19, 2011, Toronto, Ontario, Canada.

hurts both tenants and providers. Motivated by these factors, this paper tackles the challenge of extending the interface between providers and tenants to explicitly account for network resources while maintaining its simplicity. Our overarching goal is to allow tenants to express their network requirements while ensuring providers can flexibly account for them. To this end, we propose "virtual networks" as a means of exposing tenant requirements to providers. Tenants, apart from getting compute instances, are also offered a virtual network connecting their instances. The virtual network isolates tenant performance from the underlying infrastructure. Such decoupling benefits providers too– they can modify their physical topology without impacting tenants.

The notion of a virtual network opens up an important question: *What should a virtual network topology look like?* On one hand, the abstractions offered to tenants must suit application requirements. On the other, the abstraction governs the amount of multiplexing on the underlying physical network infrastructure and hence, the number of concurrent tenants. Guided by this, we propose two novel abstractions that cater to application requirements while keeping tenant costs low and provider revenues attractive. The first, termed *virtual cluster*, provides the illusion of having all VMs connected to a single, non-oversubscribed (virtual) switch. This is geared to data-intensive applications like MapReduce that are characterized by all-to-all traffic patterns. The second, named *virtual oversubscribed cluster*, emulates an oversubscribed two-tier cluster that suits applications featuring local communication patterns.

The primary contribution of this paper is the design of virtual network abstractions and the exploration of the trade-off between the guarantees offered to tenants, the tenant cost and provider revenue. We further present Oktopus, a system that implements our abstractions. Oktopus maps tenant virtual networks to the physical network in an online setting, and enforces these mappings. Using extensive simulations and deployment on a 25-node testbed, we show that expressing requirements through virtual networks enables a symbiotic relationship between tenants and providers; tenants achieve better and predictable performance while the improved datacenter throughput (25-435%, depending on the abstraction and the workload) increases provider revenue.

A key takeaway from Oktopus is that our abstractions can be deployed today: they do not necessitate any changes to tenant applications, nor do they require changes to routers and switches. Further, offering guaranteed network bandwidth to tenants opens the door for explicit bandwidth charging. Using today's cloud pricing data, we find that virtual networks can reduce median tenant costs by up to 74% while ensuring revenue neutrality for the provider.

On a more general note, we argue that predictable network performance is a small yet important step towards the broader goal of offering an explicit cost-versus-performance trade-off to tenants in multi-tenant datacenters [36] and hence, removing an important hurdle to cloud adoption.

2. NETWORK PERFORMANCE VARIABILITY

Network performance for tenants in shared datacenters depends on many factors beyond the tenant's control: the volume and kind of competing traffic (TCP/UDP), placement of tenant VMs, etc. Here, we discuss the extent of network performance variability in cloud and production datacenters.

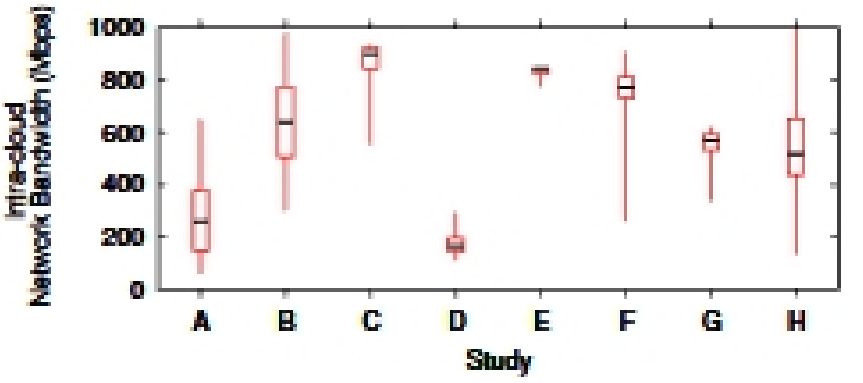

Figure 1: Percentiles (1-25-50-75-99th) for intra-cloud network bandwidth observed by past studies.

Cloud datacenters. A slew of recent measurement studies characterize the CPU, disk and network performance offered by cloud vendors, comment on the observed variability and its impact on application performance [13,23,24,30,35]. We contacted the authors of these studies and summarize their measurements of the intra-cloud network bandwidth, i.e., the TCP throughput achieved by transfers between VMs in the same cloud datacenter. Figure 1 plots the percentiles for the network bandwidth observed in these studies (A [13], B [30], C–E [23], F–G [35], H [24]). The figure shows that tenant bandwidth can vary significantly; *by a factor of five or more in some studies* (A, B, F and H).

While more work is needed to determine the root-cause for such bandwidth variations, anecdotal evidence suggests that the variability is correlated with system load (EU datacenters, being lightly loaded, offer better performance than US datacenters) [30,31], and VM placement (e.g., VMs in the same availability zone perform better than ones in different zones) [30]. Further, as mentioned in Section 1, such network performance variability leads to poor and unpredictable application performance [18,21,30,38].

Production datacenters. Production datacenters are often shared amongst multiple tenants, different (possibly competing) groups, services and applications, and these can suffer from performance variation. To characterize such variation, we analyze data from a production datacenter running data analytics jobs, each comprising multiple tasks. This data is presented in [7] while our results are discussed in [8]. Briefly, we find that runtimes of tasks belonging to the same job vary significantly, and this can adversely impact job completion times. While many factors contribute to such variability, our analysis shows that a fair fraction (>20%) of the variability can be directly attributed to variable network bandwidth. Further, we find that the bandwidth achieved by tasks that read data across cross-rack links can vary by an order of magnitude.

In summary, we observe significant variability in network performance in both cloud and production datacenters. This negatively impacts application performance. Evaluation in Section 5 also shows that in both settings, the mismatch between required and achieved network performance hurts datacenter throughput and hence, provider revenue. Since our proposed abstractions cover both cloud and production datacenters, we will henceforth use the term "multi-tenant" to refer to both.

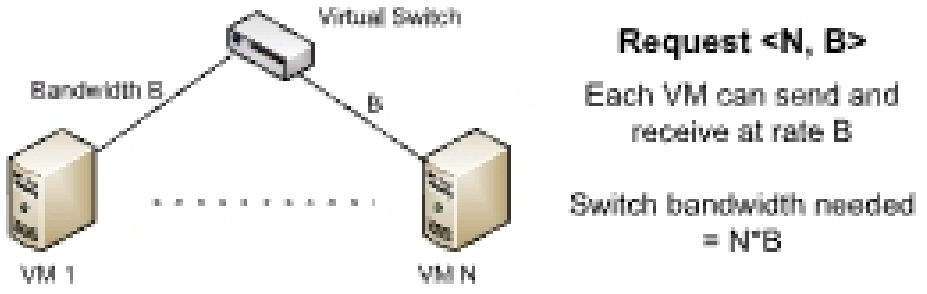

Figure 2: Virtual Cluster abstraction.

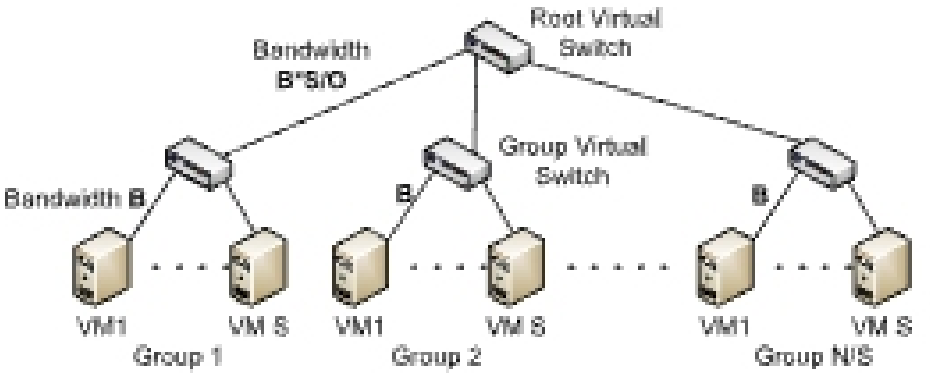

Figure 3: Virtual Oversubscribed Cluster abstraction.

3. VIRTUAL NETWORK ABSTRACTIONS

In multi-tenant datacenters, tenants request virtual machines (VMs) with varying amounts of CPU, memory and storage resources. For ease of exposition, we abstract away details of the non-network resources and characterize each tenant request as $<N>$, the number of VMs requested. The fact that tenants do not expose their network requirements hurts both tenants and providers. This motivates the need to extend the tenant-provider interface to explicitly account for the network. Further, the interface should isolate tenants from the underlying network infrastructure and hence, prevent provider lock-in. Such decoupling benefits the provider too; it can completely alter its infrastructure or physical topology, with tenant requests being unaffected and unaware of such a change. To this end, we propose virtual networks as a means of exposing tenant network requirements to the provider. Apart from specifying the type and number of VMs, tenants also specify the virtual network connecting them.

The "virtual" nature of the network implies that the provider has a lot of freedom in terms of the topology of this network, and can offer different options to tenants for different costs. Beyond the overarching goal of maintaining the simplicity of the interface between tenants and providers, our topologies or *virtual network abstractions* are guided by two design goals:

1. *Tenant suitability.* The abstractions should allow tenants to reason in an intuitive way about the network performance of their applications when running atop the virtual network.

2. *Provider flexibility.* Providers should be able to multiplex many virtual networks on their physical network. The greater the amount of sharing possible, the lesser the tenant costs.

To this effect, we propose two novel abstractions for virtual networks in the following sections.

3.1 Virtual Cluster

The "Virtual Cluster" abstraction is motivated by the observation that in an enterprise (or any private setting), tenants typically run their applications on dedicated clusters with compute nodes connected through Ethernet switches. This abstraction, shown in figure 2, aims to offer tenants with a similar setup. With a *virtual cluster*, a tenant request $<N, B>$ provides the following topology: each tenant machine is connected to a virtual switch by a bidirectional link of capacity B, resulting in a one-level tree topology. The virtual switch has a bandwidth of $N * B$. This ensures that the virtual network has no oversubscription and the maximum rate at which the tenant VMs can exchange data is $N * B$. However, this data rate is only feasible if the communication matrix for the tenant application ensures that each VM sends and receives at rate B. Alternatively, if all N tenant VMs were to send data to a single destination VM, the data rate achieved will be limited to B.

Since a *virtual cluster* offers tenants a network with no oversubscription, it is suitable for data-intensive applications like MapReduce and BLAST. For precisely such applications, Amazon's Cluster Compute provides tenants with compute instances connected through a dedicated 10 Gbps network with no oversubscription. This may be regarded as a specific realization of the *virtual cluster* abstraction with $<N, 10\ Gbps>$.

3.2 Virtual Oversubscribed Cluster

While a network with no oversubscription is imperative for data-intensive applications, this does not hold for many other applications [19,34]. Instead, a lot of cloud bound applications are structured in the form of components with more intra-component communication than inter-component communication [16,25]. A "Virtual Oversubscribed Cluster" is better suited for such cases; it capitalizes on application structure to reduce the bandwidth needed from the underlying physical infrastructure compared to virtual clusters, thereby improving provider flexibility and reducing tenant costs.

With a *virtual oversubscribed cluster*, a tenant request $<N, B, S, O>$ entails the topology shown in Figure 3. Tenant machines are arranged in groups of size S, resulting in $\frac{N}{S}$ groups. VMs in a group are connected by bidirectional links of capacity B to a (virtual) group switch. The group switches are further connected using a link of capacity $B' = \frac{S*B}{O}$ to a (virtual) root switch. The resulting topology has no oversubscription for intra-group communication. However, inter-group communication has an oversubscription factor O, i.e., the aggregate bandwidth at the VMs is O times greater than the bandwidth at the root switch. Hence, this abstraction closely follows the structure of typical oversubscribed datacenter networks. Note, however, that O *neither depends upon nor requires physical topology oversubscription.*

Compared to *virtual cluster*, this abstraction does not offer as dense a connectivity. However, the maximum data rate with this topology is still $N * B$. The localized nature of the tenant's bandwidth demands resulting from this abstraction allows the provider to fit more tenants on the physical network. This, as our evaluation shows, has the potential to significantly limit tenant costs. By incentivizing tenants to expose the flexibility of their communication demands, the