Report
OPT 558 - Internet Technologies and Systems
Gaspar Incze

Hitesh Ballani, Paolo Costa, Thomas Karagiannis and Ant Rowstron “Towards Predictable Datacenter Networks.” (ACM SIGCOMM 2011)

Problem
Do cloud providers guarantee performance? No. This is causing unpredictable application performance and also limits the use of cloud systems in a business environment where satisfying certain performance metrics is essential. Without solving this issue providers need to face a loss of revenue as a result of missed business opportunities.

Proposal
The researchers propose to use virtual networks to address the problem described above. Two major categories are identified for describing the performance needs of applications:

✔ Virtual cluster where all nodes are connected by a single (non-oversubscribed) switch – suitable for data intensive applications like Hadoop
✔ Virtual oversubscribed cluster, that is emulated by a three-tiered topology – suitable for cost effectiveness

By using this new approach, 1.25~4 times throughput increase can be achieved that is not just better but also predictable in contrast to current solutions.

The most important factors impacting network performance are comprised of system load, the volume and kind of competing traffic (TCP/UDP) and placement of a VM. By common sense, data intensive clusters should be placed physically near each other to maximize performance and minimize the extra load on the networking infrastructure.

Virtual network abstractions can be used for finding a better trade-off between costs and performance. Two main design goals were targeted:

✔ Tenant suitability: easy configuration of applications’ network demand
✔ Provider flexibility: multiplexing many virtual networks on top of the physical network

In addition to describing the two types of virtual clusters, other topologies are also analyzed in brief, such as multidimensional cubes, Butterfly networks. The problem with these is that they are highly application specific thus violating the design requirement being generic.

SecondNet is another solution that guarantee bandwidth for pairs of VM. However, this requires extra maintenance thus violating the flexibility criteria.

Oktopus is described as a network management system that can adhere to both design requirements. It is a logically centralized network manager (NM) that performs admission control and maps the requests to physical machines while taking into account the available network resources. It is also maintaining bandwidth reservations across the physical
network. The NM needs to ensure that the corresponding bandwidth demands can be met while maximizing the number of concurrent tenants.

**NM maintains the following information:**

- ✔ The data center network topology
- ✔ The residual bandwidth for each link in the network
- ✔ The empty slots on each physical machine
- ✔ The allocation information for existing tenants, including:
  - ◦ the physical machines they are allocated to
  - ◦ The network routes between these machines and
  - ◦ the bandwidth reserved for the tenant at links along these routes

Enforcing virtual networks (and Quality of Service – QoS) is done on the OS hypervisor level. Enforcement on switch level would not be scalable but tagging packets is practical: requests coming from the virtual network topology get priority.

Physical network should be allocated a portion of fix bandwidth not to starve those hosts. Providers need to find the right balance between virtual and non-virtual network parts as long as both architecture co-exist. Current implementation of Oktopus don't support usage of unused virtual bandwidth by other virtual networks so this should be improved in the future.

**Design of NW management & routing**

It is assumed that VMs are accessible in a tree topology but data centers may have richer structure. Although the spanning tree protocol can be used in Layer 2 to provide a cycle-free tree topology, for data intensive workloads even more advanced topologies may be needed:

- ✔ Equal cost Multi Path Forwarding (ECMP),
- ✔ Valiant Load Balancing,
- ✔ SecondNet (connect pairs of VMs)

It is important to take into account these more advanced situations.

**Simulation setup**

The researchers simulated 64000 VMs with 25 physical machines to further test Oktopus in a large-scale environment. Real-life workloads were chosen where Virtual Oversubscribed Clusters performed computing jobs up to five times faster. Although network utilization can be optimized a lot, outbound links from physical machines can still be a bottleneck. The problem is that VMs have to wait for network (if there is a contention) and this hurts overall throughput. When flows vary too much, Virtual Clusters can be slower than baseline – imprecise tenant demands cause waste of bandwidth. Inter-cloud communication is not charged explicitly but it's there as a hidden cost because tenants pay based on the usage time. And if the network is slow then more VM time needed to perform the same task.
On the other hand if more efficient use of the same infrastructure is realized then more customers can use it at the same time with higher throughput. Higher and guaranteed performance also brings new customers and results in more provider revenue.

**Implementation & deployment**

25 hosts were used with Intel Xeon 5520 2.27GHz CPU, 4GB RAM and 1 Gbps network interfaces with Windows Server 2008. The enforcement module was realized with the help of the traffic control API. The overhead of updating the limiting values was about 1~12 Mbps (~0,1-1% of NIC bandwidth).

**Conclusion**

To enable broader cloud adoption, application must get performance guarantees and Oktopus provide this with bandwidth enforcement. Charging for internally used bandwidth is fair because it is making hidden costs visible by asking customers to define their network communication related requirements. Finally, virtual topologies enable more efficient use of networking infrastructure giving more flexibility for both customers and providers.