Oktopus: Towards Predictable Datacenter Networks

Hitesh Ballani, Paolo Costa, Thomas Karagiannis, Ant Rowstron

Presented by: Eric Tu

The Setting

- Multi-tenant datacenters: Private and Cloud
- Tenants pay as you go for compute and storage resources
- **BUT: the hidden cost of the network**
  - Unpredictable Application performance and tenant cost
    - network load outside tenant control
  - Limited Cloud Applicability
    - some applications can’t run well: MapReduce
  - Inefficiencies in production datacenters as well
    - Hard to reason about performance -> bad productivity and revenue

The variability of network bandwidth

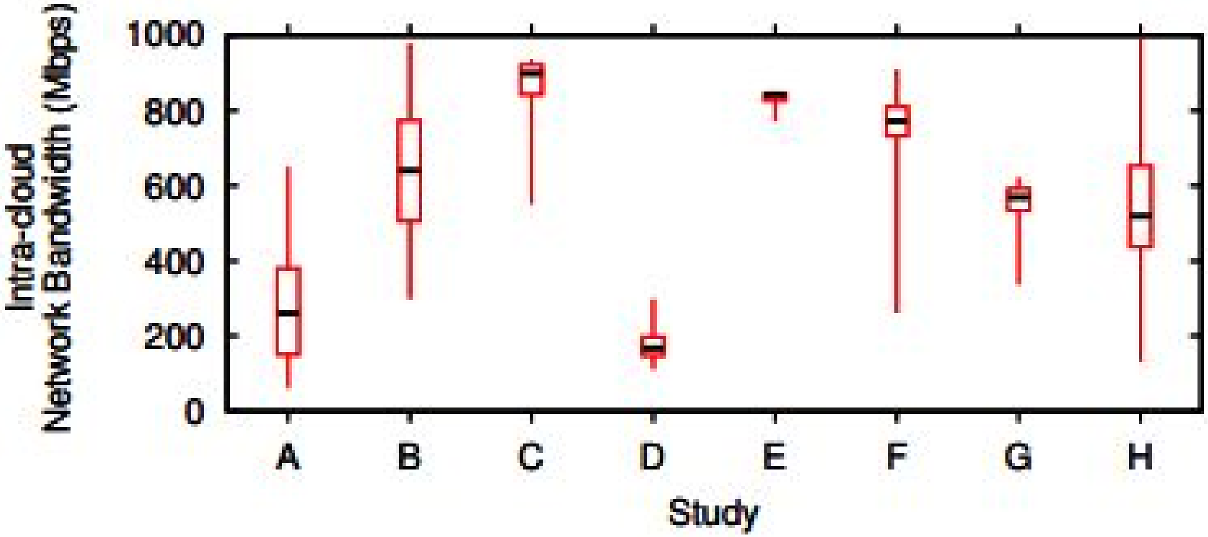


Figure 1: Percentiles (1-25-50-75-99th) for intra-cloud network bandwidth observed by past studies.

The Goal

- Maintain simplicity between tenants and providers
- Extend relationship to include network resources
- Offer better cost vs performance options to tenants
- **Everybody wins!**
  - **Tenants: Lower Cost, Predictable Performance**
  - **Provider: More Revenue**

The Solution: Virtual Networks

- Tenants get a virtual network for all their compute instances
- Decouples tenant performance from underlying infrastructure
- No need to change application, switches, routers
- **Goals:**
  - **Tenant Suitability: Tenants can understand application performance**
  - **Provider Flexibility: Maximize sharing**

Two Options for Virtual Networks

- Virtual Cluster
  - Illusion of all VMs having a non-oversubscribed switch
- Virtual Oversubscribed Cluster
  - Makes use of local communication

- **Tradeoffs**: Tenant Guarantees, Tenant Cost, Provider Revenue

Virtual Cluster: No Oversubscription

- Suitable for data-intensive apps: MapReduce
- Medium Provider Flexibility
- Reliable dedicated rate similar to Amazon’s running on dedicated Ethernet

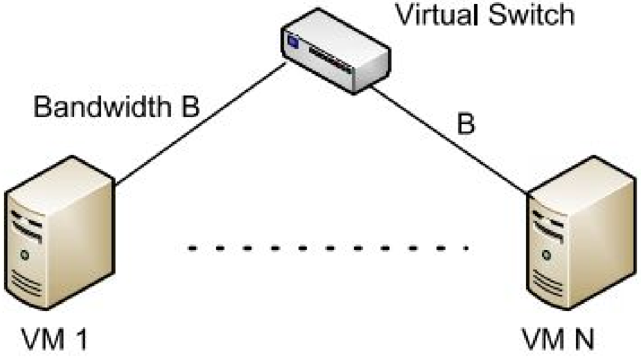


Figure 2: Virtual Cluster abstraction.

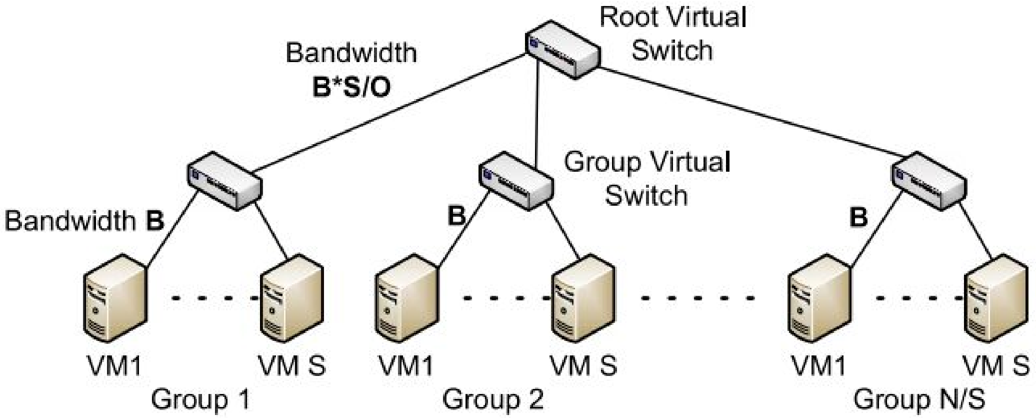
Request <N, B>

Each VM can send and receive at rate B

Switch bandwidth needed = N*B

Virtual Oversubscribed Cluster

- Make use of localized traffic
- No oversubscription within group, only intergroup
- **Greater flexibility:** Limits tenant and provider costs

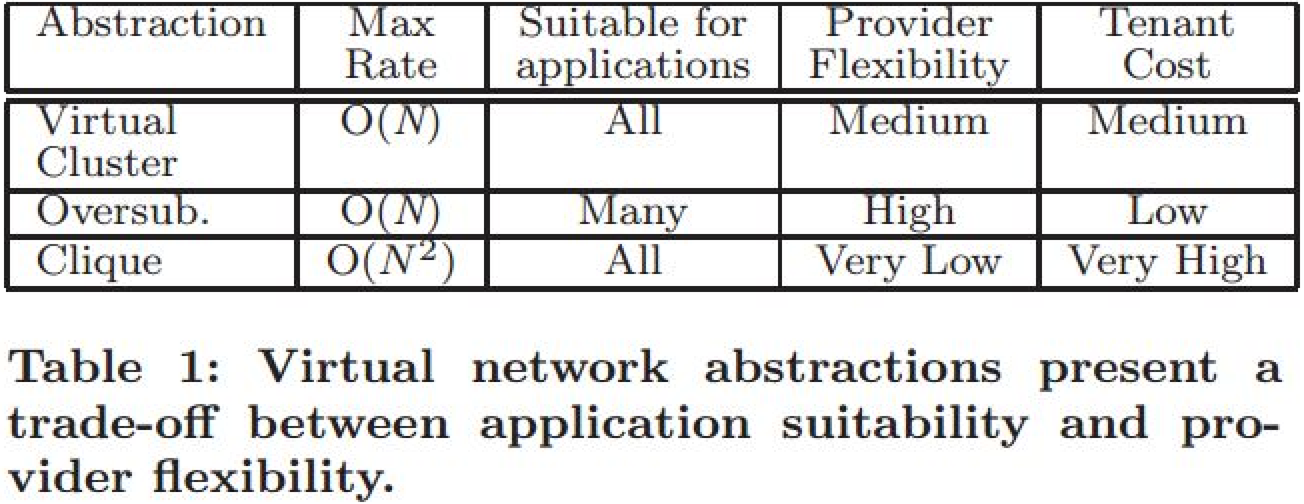


Request <N, S, B, O>

N VMs in groups of size S, Oversubscription factor O

Group switch bandwidth = S*B, Root switch bandwidth = N*B/O

Figure 3: Virtual Oversubscribed Cluster abstraction.



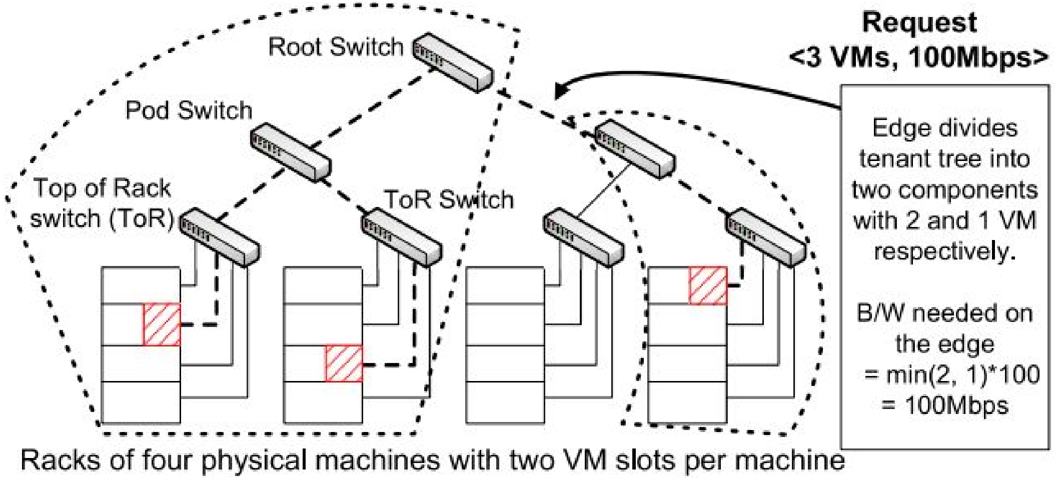
Abstraction	Max Rate	Suitable for applications	Provider Flexibility	Tenant Cost
Virtual Cluster	O(N)	All	Medium	Medium
Oversub.	O(N)	Many	High	Low
Clique	O(N^2)	All	Very Low	Very High

Table 1: Virtual network abstractions present a trade-off between application suitability and provider flexibility.

Oktopus Implementation

- Management Plane: **Allocate VNs**
  - Centralized network manager,
  - Ensures physical links connecting tenant VMs have sufficient bandwidth
- Allocation: **Observation: data centers have less bw at root than edges**
  - Try to pack VMs in smallest subtree
  - Choose subtree with least amount of residual BW to accommodate future tenants
- Data Plane: **Enforcing VNs**
  - Rate-limiting at endhost hypervisors
  - Each VM measures traffic, sends to centralized Controller VM that computes max-min fair share

Allocation of VMs



Edge divides tenant tree into two components with 2 and 1 VM respectively.

B/W needed on the edge = min(2, 1)*100 = 100Mbps

Figure 4: An allocation for a cluster request r: <3, 100 Mbps>. Three VMs are allocated for the tenant at the highlighted slots. The dashed edges show the tenant tree T.

Production DC Evaluation

- Lagging jobs from network performance limit throughput

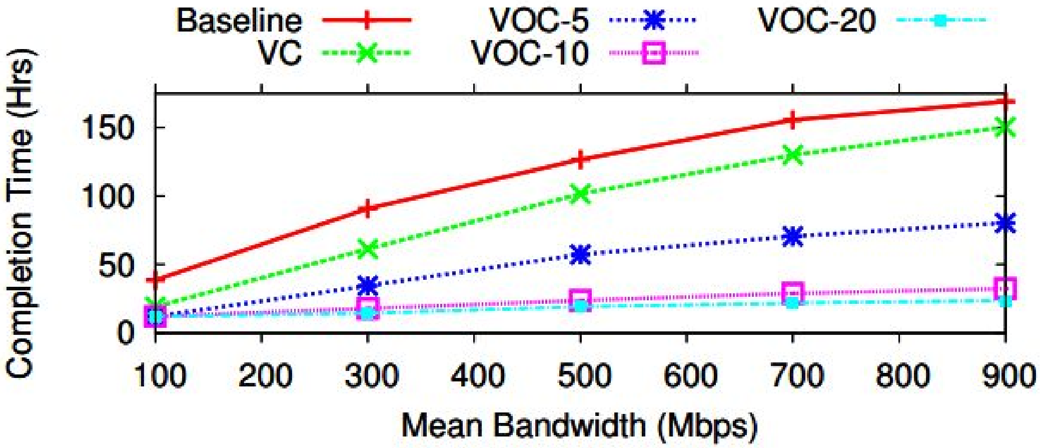
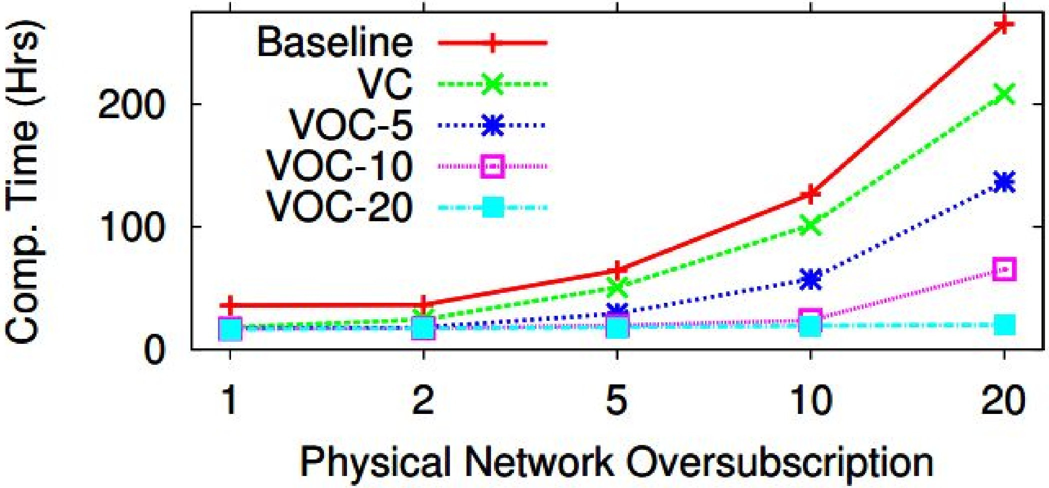


Figure 7: Completion time for a batch of 10,000 tenant jobs with Baseline and with various virtual network abstractions.

Production DC Evaluation



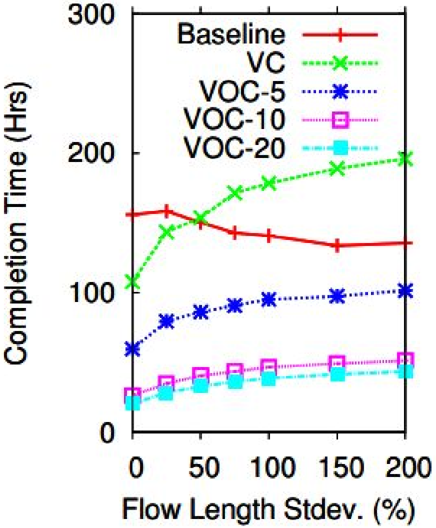
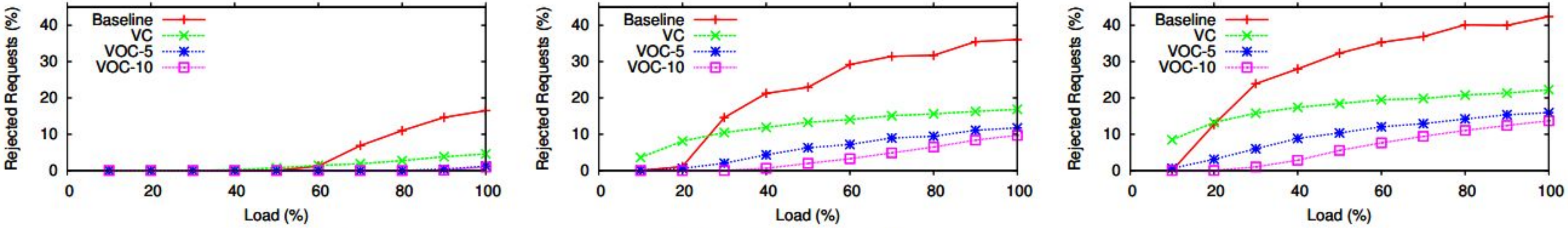


Figure 11: Completion time with varying flow lengths. Mean BW = 500 Mbps.

Cloud Datacenters

- Arriving VM requests over time
- VM rejections: Can I fit network/comp/storage?



(a) Mean BW 100 Mbps (b) Mean BW 500 Mbps (c) Mean BW 900 Mbps

Figure 13: Percentage of rejected tenant requests with varying datacenter load and varying mean tenant bandwidth requirements. At load>20%, virtual networks allow more requests to be accepted.

Cloud Datacenter Cost Savings

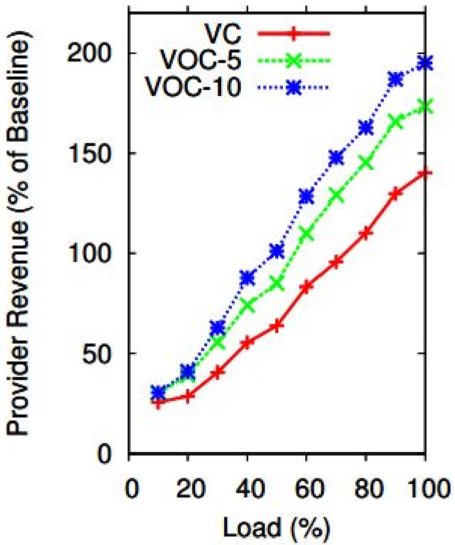


Figure 14: Provider revenue with virtual network abstractions. Mean BW = 500Mbps.

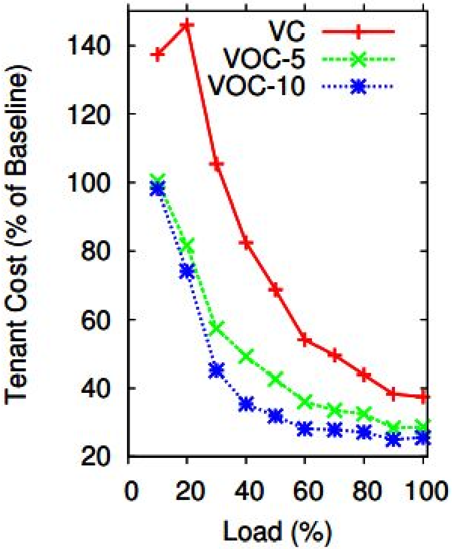


Figure 15: Relative tenant costs based on bandwidth charging model while maintaining provider revenue neutrality.

Discussion

- Impact of physical topologies on Oktopus
  - Fat-tree topologies, need load balancing
  - Tree optimization assumption
  - Will allocation be a problem in the future?
- Fault Tolerance
  - Can support, but is it expensive to redo the virtual topology?
- Usage: Is this being used, which abstraction used?
  - How to determine the abstraction used?