Datacenter Scalability Panel

Igor Gashinsky
Principal Architect
Yahoo!
June 14th, 2010

Some Recent Trends

Today

- Warehouse datacenters being built today can accommodate over 120,000 physical servers
- Each server packs a lot of cores now:
  - 2x6x2=24
  - With decent virtualization that's 20VM's per server
- That's 2,400,000 VM's in a single datacenter!
- This is today…

Tomorrow

- 10G to the server is around the corner
- DAS has left the building a long time ago, NAS is starting on it's way out, cloud storage is the new "in"
- This means that soon, every server will be a storage and compute node

What does this mean?

- To get the best utilization of all those resources, we need:
  - To be able to place a VM anywhere, anytime
  - To be able to migrate that VM wherever we need
  - To have a “flat” network, with a very low oversubscription ratio
    - target goal is 2:1
- This means that the network needs to be:
  - A flat L2 network (for IP/VM mobility)
  - Rack switches of 40x10G ports and 200G of 10/40/100G uplinks
  - Core switches with 300+ 40/100G ports
  - Control plane scalability to hold (and move) 2.4M VM’s
    - 2.4M MAC addresses + 2.4M IPv4 + 2.4M*2 IPv6 addresses

So, what's the problem?

- Core switches with 300+ 40/100G ports
- 2.4M MAC addresses + 2.4M IPv4 + 4.8M IPv6 addresses
- That's not doable using current techniques!

What about segmentation?

- 10k servers * 20 VMs = 200k VM's
- Largest L2 domain that we can scale to today = 10k servers
- Still does not help ☹

Is there a better way?

So, what are our options?

- Overlay a logical network on top of a physical network
  - This will shift the control plane scalability issue into the server/vSwitch
- Find a lighter way to scale the current network
  - Better learning mechanisms
  - Better CAM scalability

Current state of R&D

- Lately, there has been a lot of research into “programmable datacenters”
  - Monsoon
  - Seattle
  - VL2
  - Moose
  - OpenFlow
- However, no single one of them addressed all the issues:
  - Some want to modify the host stacks
  - Some want to change everything about ethernet

A Possible Solution?

- What if we “program” the datacenter without modifying the host stack and addressing?
- In large-scale deployments, companies have very extensive Inventory Management System, and they already know:
  - Where every server is in the datacenter
  - Which switch/port every server is plugged into
  - IP & MAC address of every server
- So, why is the network bothering to learn it every X seconds, why not just program it?
  - Solves the network discovery scalability issues

Questions?