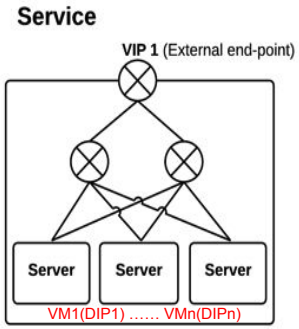
Ananta: Cloud Scale Load Balancing

Parveen Patel • Deepak Bansal • Lihua Yuan *et al.*

Proc. of ACM SIGCOMM '13, 43(4):207-218, Oct. 2013.

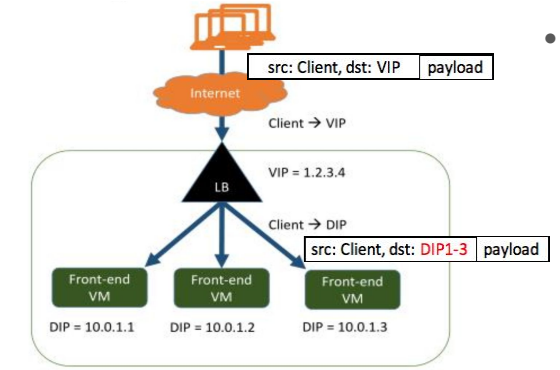
Presented by Xintong Wang and Xinghao Li

Background - Service (Tenant)



- A service is a collection of virtual machines that is managed as one entity.
- Each machine - a private Direct IP (DIP).
- A service - a public Virtual IP (VIP).
- Each service exposes zero or more external endpoints.

Background - Inbound VIP Communication



- LB is in charge of load balancing and NATs VIP traffic to DIPs.

Image source:http://www.slideserve.com/noura/ananta-cloud-scale-load-balancing

Background - Outbound VIP Communication

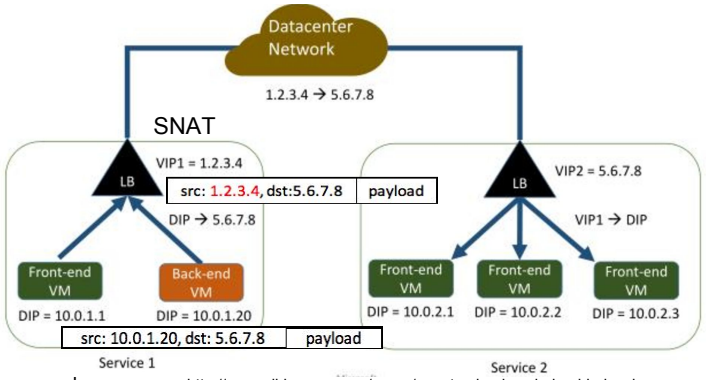
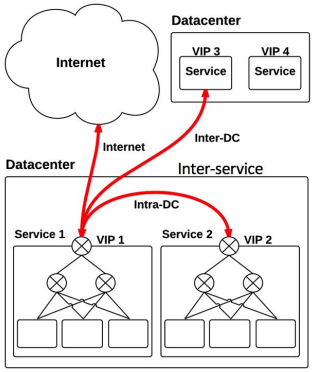


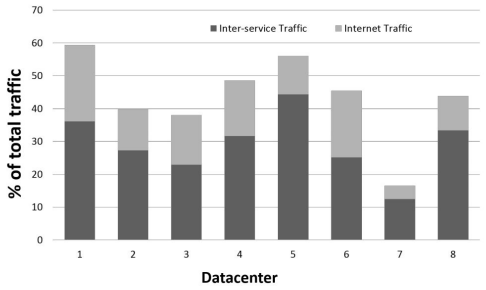
Image source:http://www.slideserve.com/noura/ananta-cloud-scale-load-balancing

Background - Outbound VIP Communication

- All traffic crossing the service boundary uses the VIP address.
- The same VIP is used for all inter-service traffic.
  - Enable easy upgrade and disaster recovery of services
  - Simplify ACL management (ACLs can be expressed in terms of VIPs).

Background - Traffic Types





On average, 44% of the total traffic is VIP traffic (LB or SNAT or both).

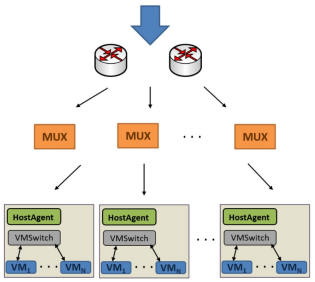
source:http://www.slideserve.com/noura/ananta-cloud-scale-load-balancing

Background - Requirements

- Scale
  - High throughput, low cost.
  - High bandwidth and large number of connections served by a single VIP.
  - Large change rate in VIP configurations.
- Reliability
  - Monitor health of instances and maintain availability.
- Any Service Anywhere
  - Reach DIPs located anywhere in the network.
- Tenant Isolation
  - Dos attacks on one service do not affect the availability of other services.
- *Traditional hardware load balancer cannot satisfy the requirements!*

Ananta Design

- Scale-Out Model
- Offload to End Systems



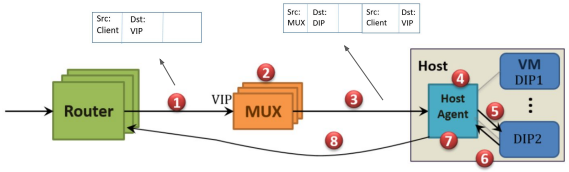
1st Tier: Provides packet-level (layer-3) load spreading, implemented in routers via ECMP.

2nd Tier: Provides **connection-level** (layer-4) load spreading, implemented in servers.

3rd Tier: Provides stateful NAT implemented in the virtual switch in every server.

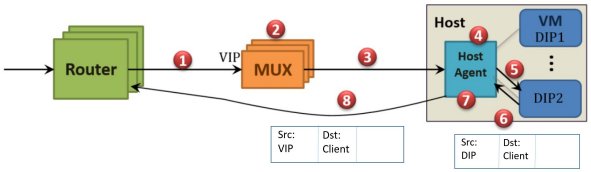
Inbound Connections - Example

1. Router distribute packets for a VIP to one MUX using Equal Cost MultiPath Routing (ECMP) protocol.
2. MUX chooses one DIP using load balancing algorithm, and encapsulates the packet using IP-in-IP protocol.
3. Send the encapsulated packet to Host Agent (HA) corresponding to the DIP.



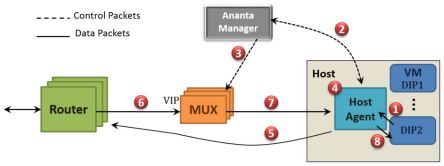
Inbound Connections - Example (cont.)

4. HA remove the outer IP header, and update the NAT state.
5. HA redirect the decapsulated packet to the target DIP.
6. DIP sends the reply packet.
7. HA perform reverse NAT on the packet based on the state in memory.
8. HA directly sends the packet to the client (bypass the MUX, it is known as Direct Server Return, or DSR).



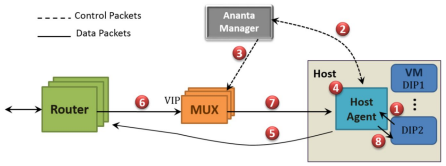
Outbound Connections - Example

1. A VM sends an outbound packet (with source IP = DIP).
2. HA performs the SNAT to the packet by first sends the request to AM for the corresponding VIP and port.
3. AM allocate such a configuration, and broadcast it to all MUXes in charge of the DIP (VIP).



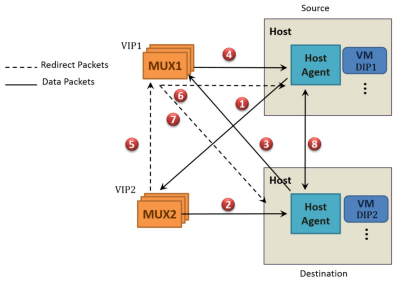
Outbound Connections - Example (cont.)

4. AM sends the allocation to HA.
5. HA rewrite the IP header (replace the source IP/port with the one AM allocated).
6. Inbound traffic, same as we discussed previously.



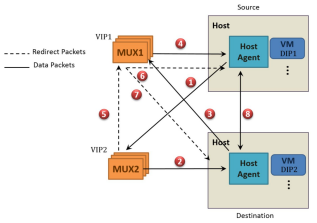
Intra-DC Connections - Fastpath

1. DIP1 (initiator) (VIP1) send the TCP SYN packet to DIP2 (VIP2). The packet first go to MUX2.
2. MUX2 forward it to DIP2.
3. The reply packet first go to MUX1.
4. MUX1 forward it to DIP1.



Intra-DC Connections - Fastpath (cont.)

5. After the connection is established, MUX2 sends redirect message to MUX1 (to redirect the traffic to DIP2).
6. After certain lookups, MUX1 sends the IP/port of DIP2 to DIP1.
7. MUX1 sends the IP/port of DIP1 to DIP2.
8. Then they are able to communicate directly.



Design Features - MUX

- Route Management
  - Work as a BGP speaker
- Packet Handling
  - VIP mapping table
  - Encapsulation (IP-in-IP protocol)
- Flow State Management
  - Stateful Entries (remember DIP selections)
  - Stateless Entries (SNAT)
- Protections
  - Trusted flows: have been seen multiple times
  - Untrusted flows: have been seen only once

Design Features - Host Agent

- NAT for inbound connections
  - IP-in-IP protocol
- SNAT for outbound connections
  - Direct Server Return
- DIP health monitoring
  - Host Agent Monitors: Monitoring local VMs and report any changes to AM.

Design Features - Ananta Manager

- SNAT Port Management - Allocate a (fixed size) contiguous port range
  - Optimize the memory usage by only store the starting port
  - Reduce the number of SNAT queries
  - Increase the availability

Design Features - Tenant Isolation

- Tenant Isolation: **Ensure the QoS of one tenant is independent of other tenants in the System.**
- SNAT Fairness
  - Similar to Round Robin
  - Only one SNAT request can be submitted by one DIP at the same time
  - Dropping excessive requests instead of waiting them.
- Packet Rate Fairness
  - Keeping track of “top-talkers”

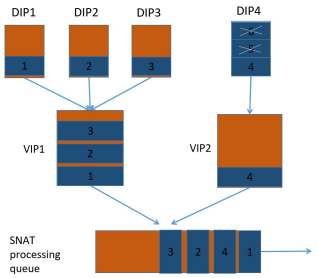
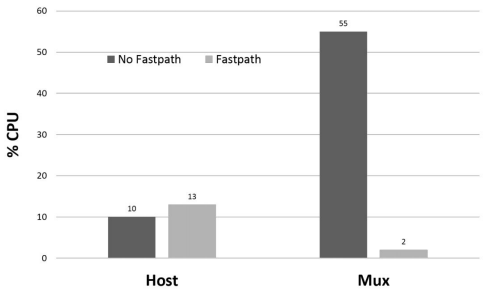


Image source:http://www.slideserve.com/noura/ananta-cloud-scale-load-balancing

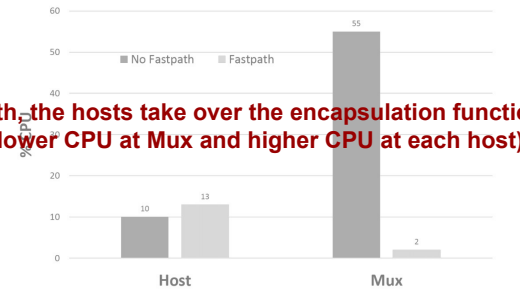
Measurements and Evaluations - Fastpath

- Server (a 20 VM tenant) and clients (two 10 VM tenant).
- Each VM creates up to 10 connections, uploads 1MB data.



Measurements and Evaluations - Fastpath

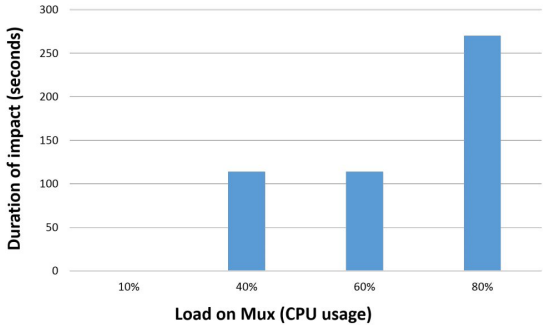
- Server (a 20 VM tenant) and clients (two 10 VM tenant).
- Each VM creates up to 10 connections, uploads 1MB data.



With Fastpath, the hosts take over the encapsulation function from Mux (lower CPU at Mux and higher CPU at each host).

Measurements and Evaluations - Tenant Isolation (SYN-flood)

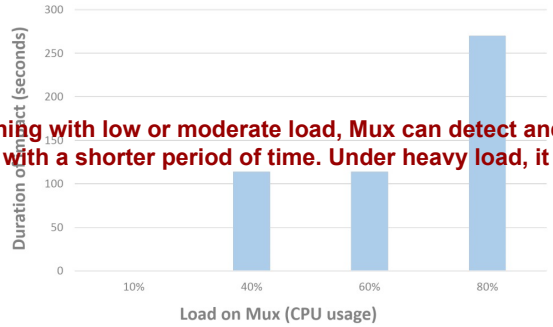
- Launch a SYN-flood attack using spoofed source IP addresses on one VIP.



Measurements and Evaluations - Tenant Isolation (SYN-flood)

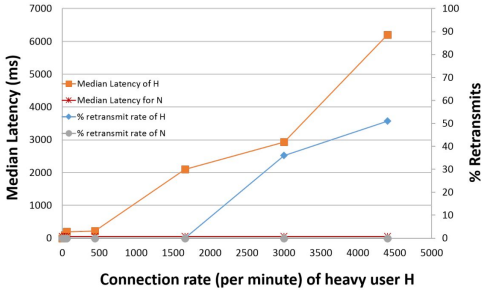
- Launch a SYN-flood attack using spoofed source IP addresses on one VIP.

When running with low or moderate load, Mux can detect and isolate an abusive VIP with a shorter period of time. Under heavy load, it takes longer.



Measurements and Evaluations - Tenant Isolation (SNAT performance)

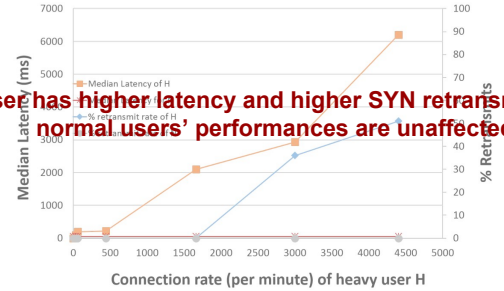
- Normal users (N) make 150 outbound connections per minute.
- A heavy user (H) keep increases outbound connection rate.



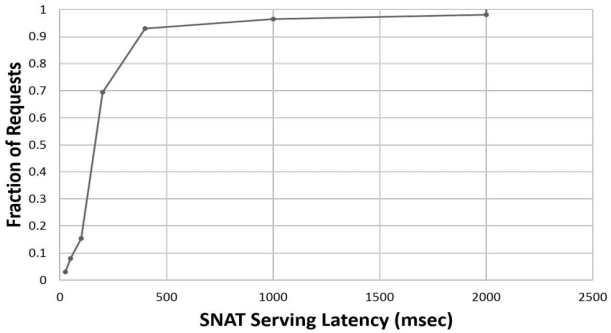
Measurements and Evaluations - Tenant Isolation (SNAT performance)

- Normal users (N) make 150 outbound connections per minute.
- A heavy user (H) keep increases outbound connection rate.

Heavy user has higher latency and higher SYN retransmits, whereas normal users' performances are unaffected.

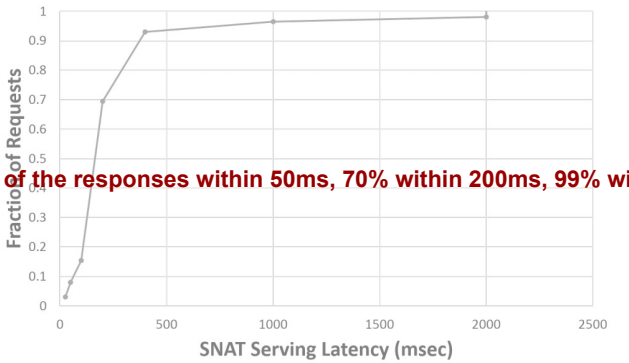


Measurements and Evaluations - SNAT Response Latency



CDF of SNAT response latency for the 1% requests handled by Ananta Manager (AM).

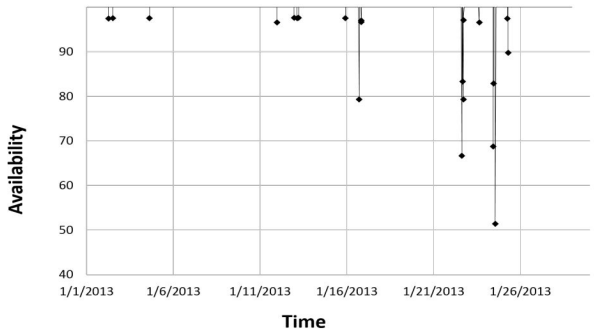
Measurements and Evaluations - SNAT Response Latency



10% of the responses within 50ms, 70% within 200ms, 99% with in 2s.

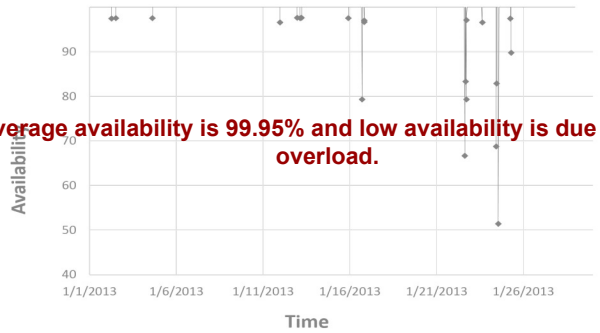
Measurements and Evaluations - Availability

- Threshold: availability less than 100% for 5min interval.



Measurements and Evaluations - Availability

- Threshold: availability less than 100% for 5min interval.



Average availability is 99.95% and low availability is due to Mux overload.

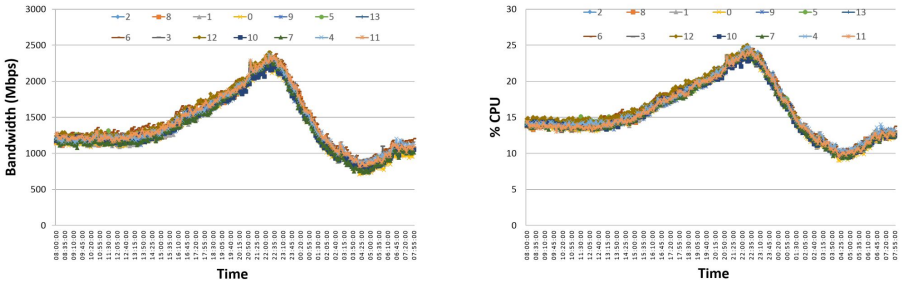


Figure 18: Bandwidth and CPU usage over a 24-hr period for 14 Muxes in one instance of Ananta.

Summary

Requirement	Description
Scale	• Mux: Equal Cost MultiPath • Host agent: Scale-out naturally
Reliability	• Ananta manager: Paxos • Mux: BGP
Any service anywhere	• A cloud scale solution for layer-4 load balancing
Tenant isolation	• SNAT fairness • Packet rate fairness

Improvements and Extensions

- Improving DoS detection to isolate the abusive VIP.
- Fastpath perturbs the order of the packets?
- Evaluation with larger scale and longer period of time?
- Tradeoff between the bandwidth threshold per flow/DIP and the overhead of load balancing redirections.