Paper Evaluation, Ananta: Cloud Scale Load Balancing

Jiani Jiang <jianij@kth.se>

1. Paper summary

As the rapid development of cloud computing, high requirement of load balancer service is needed for it is a fundamental building block of cloud environments. The load balancers suffered big pressure from the cloud services and leads to high cost, SLA violations and device failures. Therefore, this paper imposed Ananta, a scalable software load balancer and NAT, that can make clouds scalable and reliable by splitting a load balancer into a consensus-based control plane and a decentralized data plane, and offloads the two planes down to the hypervisor in end systems. The third layer of load distribution is implemented on the router through ECMP and the fourth layer of load distribution is implemented on the server. Also, reliable NAT rules are configured in each server's virtual switch. To be more specific, Ananta is made up of three main components — Ananta Manager, Multiplexer and Host Agent.

2. Top 3 contributions

1. This paper identified the high requirements to meet performance, reliability and scalability of a large cloud-scale environment for load balancing, and dug deep to see the undiscovered design space for layer-4 load balancers.

2. This paper focused on the performance of load balancers and imposed a brand new idea to refactor the functionality of load balancers and move parts of the functionality to the host to scale the size of the network.

3. Ananta uses a highly available control platform to coordinate the state of the various data platforms and makes it possible for the server to return data directly and to expand NAT design on the second layer.

3. Problems

1.Because Ananta use ECMP to do the load distribution in the third layer, it can't use bandwidth very well for the bandwidth, delay and reliability of each path in the network are very different, especially in the situation that there is a big discrepancy among paths.

2.My question: What is the contrast between cluster server load balancing and SDN load balancing?