Scaling the Network Infrastructure using OpenFlow in the Wide Area Network

Matthew Davy,
Christopher Small
Indiana University

Introduction

National and regional wide area networks have consisted of large deployments of layer 3 routers capable of containing a complete copy of the Internet routing table. The routers constitute the Internet Default-Free Zone where a default route is not needed to route a packet to any destination. All routers of this class need to be capable of all of the functions required by all service providers that deploy it. Protocols have to be specialized to the unique hardware capable of handling the large amount of traffic on the wide area backbone networks. The cost of development of these routers is quite high because of the large number of protocols and features need to be implemented. Specialized hardware such as large high speed SDRAM memory for buffers is often necessary on this class of routers.

The market for "carrier-grade" routers is also much smaller to the broader network equipment market. The relatively low number of units drives up the cost for each router, as there is a fixed cost in development of software and hardware. These factors make the total capital cost of each router port expensive compared to layer 2 switches.

The deployment of layer 3 routers also introduces other related costs to the network operator. Routers capable of routing the entire Internet routing table require significant amounts of power, space and cooling to operate. The co-location costs may be many times the actual space used because these devices often exceed the planned power density of typical co-location facilities. It is often common that the port densities of layer 3 devices are lower compared to layer 2 equipment. A Juniper MX 960 has a maximum density of 3 10 Gigabit per second port per rack unit vs. a number of layer 2 switches that support 48 10 Gigabit ports in a 1 rack unit device.

In total, the Internet Default-Free Zone infrastructure is estimated to cost $2 Billion a year to operate.[1]

[1] http://bill.herrin.us/network/bgpcost.html

Research and Educational Networks

Research and Educational (R&E) networks are instances of networks where the deployment of carrier-grade routers may not be needed or advantageous. Networks such as Internet2 and NLR are essentially in a default-free zone in a different domain. They are the top-level networks with no defaults but they just contain a subset of the Internet routing table.

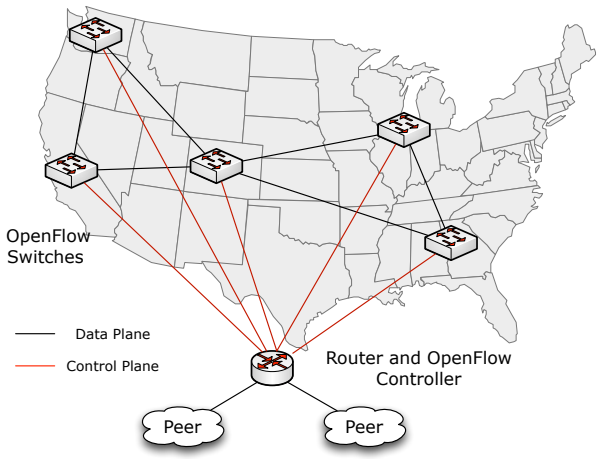
Using the same equipment and technology in R&E networks as on the Tier 1 ISP backbones may not be the most efficient allocation of resources. Deploying routers that have capabilities that will never be used by the R&E networks imposes an unnecessary cost. Conversely features that may of be of use to research and educational networks may not be available on carrier grade routers or less capable replacements.

The fact that the routing tables used in these networks are much smaller than the Internet routing table is a great advantage in introducing any new routing technology. A network framework can be initially scaled to the requirements of the R&E networks.

One approaching to solve the problems of cost of equipment and appropriateness of the network infrastructure to the needs of the network operators is to develop an “a-la-carte” approach to the construction of the network. Components, such as software and hardware, can be customized and scaled to the requirements of the network. If a feature or protocol is not needed it doesn’t need to be included in the infrastructure.

Wide-Area Routing and OpenFlow

We propose an experiment to test the potential for a highly scalable and efficient WAN architecture based on OpenFlow. The experiment would create a wide area network, with inexpensive and high-capacity layer2 devices handling the data plan, and a software based BGP daemon, such as Quagga, providing the network’s control plan by acting as a single route server for the entire network. By separating the control plane from the data plane, resources can be allocated better and software and hardware decisions can be divorced from each other, allowing for better flexibility and less specialized hardware.



Because the control plane would be handled centrally, there is no need for every network device to understand the routes to every location on the Internet as long as there is a single controlling entity to determine the forwarding behavior of the underlying switches.

This architecture is possible without the use of OpenFlow, with existing Ethernet hardware providing the data plane. However, such an implementation has inherent serious problems with maintaining optimal low-latency forwarding, and overall stability and resiliency. Spanning Tree Protocol (STP) can eliminate loops in the network but results in sub-optimal layer 2 paths. In addition, a STP-based architecture can lead to long convergence times for forwarding, which could in turn lead to overall instability of the network.

The use of OpenFlow in this type of architecture avoids some drawbacks in the use of Spanning Tree Protocol for large Layer 2 domains. STP can eliminate loops in the network but results in sub-optimal layer 2 paths, and can lead to long convergence times for forwarding. With OpenFlow, the OpenFlow controller will make decisions based on the optimal paths and direct flows to conform with routing instructions from the BGP software router, which results in better path selection, and may result in improved resiliency, with some improvements.

This experiment will study mechanisms to minimize the effects of network outages on an OpenFlow-based wide area network. The physical separation of OpenFlow controller and forwarding devices (and the related latency involved) may cause a delay in the restoration of service after a network event. The experiment will examine ways of remediating this latency using a multi-tiered architecture of controllers and switches in which the switches themselves can handle the basic functions of recovery and reconfiguration, a local or regional controller handles other functions that require short latency and processing time, and a top level controller handles policy decisions for the network as a whole. We will implement examples of single and multi-tiered experiments and develop metrics calculating time to recovery from single or multiple failures of links and nodes.

The experiment will also study the effect of network outages in the control path network to determine what effects network outages disrupting control plane traffic may cause, and how these effects can be minimized. The separation of control plane from data plane provides the additional benefit that a control plane malfunction or reconfiguration does not immediately affect the underlying data traversing the network. Once the control plane recovers it may be able to signal new paths but the data plane will keep passing traffic until the control plan recovers.

GENI resources

The wide area network experiments will rely on the infrastructure deployed by the GENI OpenFlow Campus trials. The OpenFlow switches deployed in the national backbone will be especially useful as they can accurately reflect the conditions and physical topology of a wide area network.

We would create multiple simulated peer networks to study the effect of different configuration and the effect of simulated outages on the topology.