DCnet: A New Data Center Network Architecture

Douglas Comer, Rajas H Karandikar

March, 2017

Introduction

The migration of enterprise communication and storage to the cloud has imposed new communication requirements and a need for fundamental changes in infrastructure. As more applications are hosted in virtual machines (VMs) in a cloud environment, significant amount of network communication occurs within a data center, the so-called east-west traffic. Cloud providers exploit virtual machine migration to achieve high compute and power efficiency by eliminating hot spots, balancing loads, and shutting down resources during preventative maintenance and idle periods. The increasing use of virtualization in data centers means that future data center architectures must support efficient VM migration in addition to the traditional requirements of scalability and easy manageability. We propose a radical redesign of data center networks using an approach we call DCnet. DCnet changes the addressing schemes at layers 2 and 3 completely. The new design allows virtual machines to be migrated freely within a single data center as well as across data centers. DCnet retains compatibility with the existing hardware by using the same fundamental building blocks of Ethernet and IP. Furthermore, DCnet does not require changes to the host environment: the operating systems, application programming interface, libraries and applications themselves remain unchanged, making transition straightforward.

DCnet Architecture

In traditional networks, the Internet layer provides end-to-end connectivity. Each node connected to the Internet is assigned a unique IP address. Intermediate IP routers in an IP network use the destination IP address in a packet to forward the packet towards the correct destination. IP addresses in an IP network are used to encode the location of a node. Transport layer protocols like the Transmission Control Protocol (TCP) provide application-to-application connectivity. TCP identifies a connection using end point IP addresses and port numbers. Thus, in addition to encoding the location of nodes, IP addresses are also used as identifiers in transport layer protocols. The dual purpose of an IP address of encoding location and providing identity makes the problem of mobility challenging. When a node moves to a new network it is assigned a new IP address that is derived from the network address of the new network. However, if a moving node needs to maintain its TCP connections, it must maintain its IP address after moving to the new location. Thus, in traditional networks, a node can only move within the same subnet. Efficient cloud applications rely heavily on the ability of the data center to provide VM migration without any restrictions.

DCnet redefines addressing at layers 2 and 3 to overcome the restrictions on VM migration. DCnet generates an illusion of an organization wide layer 2 network to the end hosts. From the perspective of end hosts and VM orchestration systems, all hosts in an organization are connected to the same layer 2 network, which makes it possible for a VM to move freely within the entire organization, i.e. within a single data center and across data centers. DCnet separates the identity and location of a host into two separate addresses. Each host in DCnet is assigned a 24-bit Unique ID (UID) which is used to generate its identity address. In addition each host also is assigned a location address called a Routable MAC (RMAC) which encodes the location of the host. RMAC uses a hierarchical addressing scheme, like an IP address, that is used by network switches to efficiently forward packets with the organization. With a redesign of the data center network, DCnet provides the following features:

- A single organization can span multiple data centers. Network infrastructure supports VM migration throughout the organization.
- A VM (or any addressable entity) uses standard IP addresses and preserves the IP addresses after migration.
- Layer 2 addressing is expanded into separate location and identity addresses. Location addresses use a hierarchical addressing scheme.
- An IP address that is assigned to a VM is globally valid and provides connectivity internal and external to the organization (i.e. to the global Internet).
DCnet Testbed

To demonstrate the viability of the DCnet approach to data center networking, we plan to build a testbed using network switches and servers. The testbed will, of course, be much smaller than an actual data center. However, it will be sufficient to show how the DCnet architecture works in a realistic setting. The network topology of the testbed will be the traditional Fat-tree network topology used in data centers [1]. The Fat-tree network topology is a widely accepted norm for contemporary data centers; data centers built by Google [2] and Facebook use a slightly modified form of the FAT tree topology. To emulate data center switches, our testbed will use x86 servers with multi-NIC Ethernet cards. The software will consist of Linux plus a fast packet switching library known as the Intel Data Plane Development Kit (DPDK).

We will use the DCnet testbed to demonstrate the viability of the DCnet architecture and addressing scheme and to evaluate the overall performance of the new packet routing system. In essence, our testbed will provide a smaller replica of an actual data center. Consequently, the testbed can also be used to perform additional experiments related to cloud and data centers. For example, we will be able to evaluate the effects of rolling out a new cloud application or otherwise changing the workload. We also note that our DCnet testbed will provide a platform that can be used by undergraduate and graduate students as well as other researchers in the Computer Science and Electrical & Computer Engineering departments.

References


Budget

We request support for a half-time research assistant, and funds to present the work at a conference (travel and registration).

<table>
<thead>
<tr>
<th>Item</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2-time Research Assistant</td>
<td>38,265</td>
</tr>
<tr>
<td>Conference registration and travel</td>
<td>700</td>
</tr>
<tr>
<td>Total</td>
<td>38,965</td>
</tr>
</tbody>
</table>