A Software-Defined Network Architecture for Disaggregated Racks

Mr. Cheng-Chun Tu (William)
Stony Brook University

I'm a sixth year Ph.D. student in Stony Brook University. I received my bachelor degree in National Chiang Tung University, Taiwan, and master's degree in Chalmers University of Technology, Sweden in 2005 and 2007 respectively. My research interest and expertise lie in data center networking, Software-Defined Network, rack disaggregation, and I/O virtualization. I am an experimental computer scientist who builds innovative computing systems to solve real-world problems and support real-world services, performs detailed measurements and analysis to understand how they work in terms of scalability, security and availability, and devises novel methods to optimize these systems accordingly.

My Ph.D. research comprises two parts: a cloud data center network architecture featuring an Ethernet switch-based Software-Defined Network (SDN), and a memory-based rack-area network using a PCIe-based network to enable rack disaggregation. The Ethernet-Based SDN architecture, called Peregrine, is the world's first known example of applying the design principles of SDN -- decoupling control plane from data plane and centralizing the network control plane – to COTS (Commodity Off-The-Shelf) Ethernet switches rather than OpenFlow switches. Peregrine turns off the control protocols in Ethernet switches, e.g., spanning tree, source learning, flooding, etc., and supports innovative SDN functionalities such as fast-failover, dynamic traffic engineering, in-band control, and network virtualization. Peregrine paves the way towards a hybrid SDN architecture where the underlying switches are OpenFlow switches and mainstream Ethernet switches. The Peregrine technology is being productized by the Industrial Technology Research Institute at Taiwan and parts of it are being contributed to Open Daylight and Open Compute Project.

The second project, Marlin, proposes a rack area network architecture based on the concept of rack disaggregation, in which a rack consists of a CPU/memory pool, a disk pool, and a network interface (NIC) pool, connected through a high-bandwidth and low-latency rack-area network. The major advantage of this architecture is that it allows different system components, i.e., CPU, memory, disk, and NIC, to be upgraded and evolved independently according to their own technology cycle. Moreover, from an architectural point of view, the concept of modularization of each data center server to be a CPU/memory module, decouples the CPU/memory from I/O devices such as disk controllers and network interfaces, and enables more efficient, flexible and robust resource sharing.

A key enabling technology for the rack disaggregation architecture is the ability for multiple CPU/memory modules to share I/O devices. While existing solutions rely on customized hardware, we propose Ladon, a software-only solution that allows multiple servers or multiple virtual machines (VM) running on distinct servers to share PCIe-based I/O devices securely, efficiently and transparently. The current Marlin prototype supports both RDMA applications and TCP/IP socket-based applications running transparently and efficiently on a network consisting of PCIe and Ethernet switches. Specifically, VMs running on distinct servers of a rack are able to use the same PCIe-based NIC to communicate with VMs running on a different rack, without interfering with one another and at native speed.

Another project that is still under development is an efficient real-time network function virtualization (NFV) platform that builds on Marlin and Ladon. Compared with standard virtualization solutions, NFV's requirement on virtualization is low and predictable latency as well as high throughput. I/O virtualization (IOV), which allows VM to directly interact with the network interface directly without hypervisor involvement, is expected to play a key role in satisfying the real-time performance requirement. In this project we identify several key context switching overheads, i.e., VM exit and entry, among the design of the KVM hypervisor and propose optimization methods to reduce or remove them. Eventually, combining with the PCIe networking technology in Marlin and I/O sharing technology of Ladon, we will allow multiple VMs to share the network devices in the rack-area in a secure, efficient, and real-time fashion. The technology resulting from this project is expected to compete with Intel/WindRiver, which recently announced their Open Virtualization Profile (OVP) solution.

The key principle of SDN is to provide the capability of defining modular interfaces through software abstractions that run across multiple components and the ability to decouple the components is the enabling technology to achieve the essence of SDN. The Peregrine project shows that the concept of SDN is equally applicable to existing Ethernet switches, thus enabling users to deploy SDN technology on existing network infrastructure without going immediately to OpenFlow switches. The Ladon project enables the decoupling of CPU/memory from the I/O devices, and thus serves as the first step towards disaggregated rack architecture, as advocated by Facebook in the last OCP Summit. The Marlin project leverages Ladon's I/O device sharing mechanism and creates a fully operational disaggregated rack area network switch. All of these together enable an efficient rack-area networking system that opens up myriad exciting applications, including the emerging network function virtualization (NFV) vision for next-generation wireless core network services.