ETH zürich Research Collection

Doctoral Thesis

Optimized Protocol Stack for Virtualized Converged Enhanced Ethernet

Author(s):
Crisan, Daniel

Publication Date:
2014

Permanent Link:

https://doi.org/10.3929/ethz-a-010277814 →

Rights / License:

In Copyright - Non-Commercial Use Permitted →

This page was generated automatically upon download from the ETH Zurich Research Collection. For more information please consult the Terms of use.

ETH Library

Diss. ETH No. 22127

Optimized Protocol Stack for Virtualized Converged Enhanced Ethernet

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Daniel Crisan

Ing. Info. Dipl. EPF

born on 03.09.1982

citizen of Romania

accepted on the recommendation of

Prof. Dr. Lothar Thiele, examiner

Prof. Dr. Torsten Hoefler, co-examiner

Mitch Gusat, co-examiner

2014

to Sînziana

Abstract

Datacenter networking undergoes a silent transition driven by the emergence of Converged Enhanced Ethernet (CEE) and network virtualization.

CEE aims to converge all the traffic generated by the previously disjoint local, system and storage networks on a single physical infrastructure. Traditionally, Ethernet did not guarantee losslessness: packets were dropped whenever a buffer reached its maximum capacity. This behavior does not match the semantics of modern datacenter applications used for storage or low-latency communications. CEE segregates Ethernet frames into eight different hardware priorities. Each priority may be configured as either lossy or lossless. Within a lossless priority, Priority Flow Control (PFC) prevents buffer overflows in a hop-by-hop manner. In this thesis, we will show that lossless Ethernet clusters can improve the performance of on-line data intensive applications. In particular, lossless fabrics avoid TCP incast throughput collapse, and can reduce the completion times by up to an order of magnitude.

Virtualization aims to consolidate different applications on the same hardware, thus increasing the average utilization of both the servers and communication equipment. The drawback of virtualization is that the TCP/IP stack, which was originally created and optimized to run directly over the network hardware, now runs over a new stack of layers responsible for virtualization, isolation, and encapsulation. In this thesis we will show that is possible to deconstruct the TCP protocol and redistribute its functions between the guest OS and the hypervisor. We will show that it is possible to conserve the existing features but with a much lower overhead. In our proposed architecture the hypervisor takes over most of reliability, flow and congestion control functions from the guest OS.

In this work we will will provide a practical way of virtualizing CEE. We will show how current hypervisor software lags behind network hardware by arbitrarily dropping frames in the virtualization layers, despite the fact modern Ethernet provides lossless traffic classes. Therefore, we will take corrective actions and we will introduce the first CEE-ready virtual switch. Next we will design a hypervisor that prevents misconfigured or malicious virtual machines (VMs) from filling the lossless cluster with stalled packets and compromising tenant isolation. We will prove the benefits of our new network hypervisor using a prototype implementation deployed on production-ready datacenter hardware.

Résumé

Les réseaux de centres de données subissent une transition silencieuse entraînée par l'émergence de Converged Enhanced Ethernet (CEE) et la virtualisation du réseau.

L'objectif de CEE est de faire converger tout le trafic généré par les réseaux locaux, systèmes et de stockage, auparavant disjoints, sur une seule infrastructure physique. Traditionnellement, Ethernet n'était pas sans perte: des trames ont été supprimées chaque fois qu'un mémoire tampon a atteint sa capacité maximale. Ce comportement ne correspond pas aux applications de centres de données modernes utilisées pour le stockage ou les communications à faible latence. CEE sépare les trames Ethernet en huit priorités différentes. Chaque priorité peut être configuré soit avec ou sans perte. Dans une priorité sans perte, Priority Flow Control (PFC) empêche les débordements des mémoires tampons. Dans cette thèse, nous allons montrer qu'Ethernet sans perte peut améliorer les performances des applications en ligne. En particulier, les réseaux sans perte peuvent éviter « TCP incast » et peuvent réduire les délais d'exécution jusqu'à un ordre de grandeur.

Virtualisation vise à consolider les différentes applications sur le même matériel, augmentant ainsi le taux d'utilisation moyen des serveurs et des équipements de communication. L'inconvénient de la virtualisation est que la pile TCP/IP, qui a été initialement créé et optimisé pour fonctionner proche de matériel, fonctionne maintenant sur une nouvelle pile de couches responsables de la virtualisation, l'isolement et l'encapsulation. Dans cette thèse, nous montrons qu'il est possible de déconstruire le protocole TCP et de redistribuer ses fonctions entre le système d'exploitation de la machine virtuelle et l'hyperviseur. Nous montrons qu'il est possible de conserver les fonctionnalités existantes mais avec un cout beaucoup plus faible. Dans notre architecture proposée, l'hyperviseur prend les fonctions de fiabilité, de contrôle de débit et de congestion du système d'exploitation de la machine virtuelle.

Dans ce travail, nous allons fournir un moyen pratique pour virtualisation de CEE. Nous allons montrer comment les hyperviseurs courant suppriment arbitrairement des trames dans les couches de virtualisation, en dépit du fait qu'Ethernet fournit des classes de trafic sans perte. Par conséquent, nous allons prendre des mesures correctives et nous allons introduire le premier commutateur virtuel pour CEE. Ensuite, nous allons concevoir un hyperviseur qui empêche les machines virtuelles mal configurés ou malveillantes de remplir le cluster sans perte de paquets bloqués et de compromettre l'isolement. Nous allons prouver les avantages de notre nouvel hyperviseur de réseau à l'aide d'un prototype de l'application déployée sur le matériel du centre de données prêt pour la production.