Universidad Autónoma de Madrid

Escuela Politécnica Superior

Máster conjunto en Ingeniería Informática e Investigación e Innovación en Tecnologías de la Información y las Comunicaciones

TRABAJO DE FIN DE MÁSTER

ESTUDIO DE CAPTURA Y ALMACENAMIENTO DE TRÁFICO EN REDES FÍSICAS Y VIRTUALES MULTI-GIGABIT

Rafael Leira Osuna
Tutor: Iván González Martínez

19/06/2015

ESTUDIO DE CAPTURA Y ALMACENAMIENTO DE TRÁFICO EN REDES FÍSICAS Y VIRTUALES MULTI-GIGABIT

Autor: Rafael Leira Osuna
Tutor: Iván González Martínez

High Performance Computing and Networking
Dpto. de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid

19/06/2015

Agradecimientos

Son muchas a las personas a las que debería agradecer y poco el espacio en donde escribir. A Juan Sidrach, por completar y mejorar esta plantilla en LaTeX, así como a uno de sus creadores originales, Diego Hernando, el cual me ha apoyado siempre a lo largo de carrera y el Máster. A Iván González, por haber sido un magnífico y comprensivo tutor y por ayudarme siempre en todo lo que ha podido. Gracias a el y a la ayuda de Victor Moreno, he podido finalizar con éxito este Trabajo Fin de Máster. Al equipo de handbe: Diego Hernando, Carlos Asensio y Paloma Domínguez, con los cuales he pasado muchas horas llevando a cabo nuestro proyecto y que también me han apoyado de múltiples formas. A Raúl Martín, porque sin esos viajes en coche y nuestras conversaciones en el gimnasio, el año hubiese sido mucho más aburrido.

Y en general a todos los compañeros del grupo de investigación HPCN, ya que en mayor o menor medida han puesto su granito de arena a lo largo del Máster. Menciono especialmente a: Rubén García-Valcárcel, David Muelas, Mario Ruiz, José F. Zazo, Isabel García y a Sergio López Buedo. De la misma manera agradezco el apoyo a todas y cada una de las personas que he conocido a lo largo de la carrera y del Máster, a mis amigos de toda la vida, a familiares y a Bea que de una u otra forma han estado conmigo a lo largo de este doble Máster haciéndolo más ameno y finalmente llevándome a escribir este TFM, y con él, a terminar mi doble Máster.

Abstract

Abstract — Study and analyze a high speed network ($\geq$10Gbps) is a challenge in terms of the amount of data to be processed and the data rate itself. As a result, the networking capture tools are usually very complex. Those tools also have to be continuously adapted to new technology and higher data rates. To meet those requirements, each capture tool implements its own formats and way to capture that difficulties its interoperability. In order to solve this problem, it is necessary to develop a capture tool that stores and works with network data in a well-known format. Standard formats, like *PCAP*, allow different applications to work together easly, even in a paralel way. In the same way, common formats frees network analyzing tools from the underlying network.

Typically, expensive dedicated servers are used to capture, store and process network data at high speed rates. However, this is changing due to the proliferation of *cloud computing* and the greatly improved performance virtualization technology. This trend makes difficult to find bare-metal servers or even network equipment in some environments. Therefore, it is becoming more and more important to evaluate the performance and feasibility of capture and process network data on virtual environments. To achieve that, a capture and store tool has been developed. The tool can work at 10Gbps thanks to *Intel DPDK* capture technology. A technology, that also can work in both bare-metal and virtual environments. In this work, different methods and capture tools are compared. In the same way, different virtualization methods provided by *KVM* are evaluated. While running applications in virtual machines have a small overhead compared with the bare-metal version, results show that performance in virtual environment is really close to bare-metal environment. However, those results can only be reached using the correct configuration and the latest advantages of the state-of-the-art hardware devices.

Key words — Virtual network functions, packet capture, virtual machines, Intel DPDK, HPCAP

Resumen

Resumen — Estudiar y analizar el comportamiento de una red a alta velocidad (≥10 Gbps) supone un reto constante a medida que aumenta la velocidad de las redes de comunicaciones debido a la gran cantidad de datos que se generan a diario y al propio hecho de procesar información a tales velocidades. Por estos motivos, las herramientas encargadas de la captura de datos son complejas y se encuentran, por lo general, en constante adaptación a las nuevas tecnologías y velocidades, lo que dificulta considerablemente su integración directa con otras aplicaciones de motorización o análisis de datos. Por ello es necesario que estas herramientas sean capaces de capturar y almacenar los datos en un formato estándar en el que otras herramientas puedan trabajar a posteriori o incluso en paralelo, con los datos de red independientemente de la tecnología de captura utilizada.

Típicamente, este proceso de captura, almacenamiento y procesamiento de datos a alta velocidad se ha realizado en máquinas dedicadas. No obstante, debido a la proliferación del *cloud computing* y a la gran mejora en rendimiento de la tecnología de virtualización, esto está cambiando, pudiéndose llegar al caso en el que sea raro disponer de una máquina física en la que realizar estos procesos. Por ello, evaluar la viabilidad de realizar estos procesos de tan alto rendimiento dentro de entornos virtuales comienza a cobrar importancia. Dentro de este contexto, se ha desarrollado una herramienta de captura y almacenamiento en disco a 10 Gbps mediante la tecnología de captura *Intel DPDK*, con la capacidad de funcionar tanto en entornos físicos como virtuales. Del mismo modo, en este trabajo se presentan y se comparan diferentes métodos y herramientas de captura, así como los diferentes métodos de virtualización de componentes que ofrece *KVM*. A pesar de que el uso de máquinas virtuales impone un sobrecoste computacional a cualquier aplicación, los resultados obtenidos muestran que el rendimiento en entornos virtuales se asemeja mucho al rendimiento en entornos sin virtualización, siempre y cuando se utilice la configuración adecuada que exprima las capacidades de los dispositivos actuales.

Palabras clave — Funciones de red virtuales, captura de paquetes, máquinas virtuales, Intel DPDK, HPCAP

Glosario

CAPEX El término inglés CAPEX, se refiere fundamentalmente a los gastos iniciales (o de actualización si es el caso), que son necesarios para poner en marcha un producto, o construir algo. 1

Cloud Computing Se define la computación en la nube, como un conjunto de servicios que se encuentran en internet. Dichos servicios suelen estar descentralizados y repartidos en localizaciones y elementos físicos a los que usualmente no se tiene acceso. 1

CPU La tan conocida Unidad de procesamiento central, o CPU, se encarga de realizar la mayor parte de las tareas que realiza un ordenador de hoy en día. En las últimas arquitecturas, es también la encargada de coordinar diferentes dispositivos como la memoria o el bus PCIe.. 1

OPPEX El término inglés OPPEX, se refiere fundamentalmente a los gastos que tiene un producto tras su contratación o compra. 1

SR-IOV También conocida como *Single Root I/O Virtualization*, SR-IOV es una tecnología de virtualización que permite a un único dispositivo PCIe, mostrarse como diversas funciones virtuales (VF). Cada una de estas VF, puede conectarse a una máquina virtual, de forma que la compartición del dispositivo sea realizada por el propio hardware. Esto, minimiza el coste computacional de la máquina anfitriona, así como de las diversas máquinas virtuales que utilizan el dispositivo. 1

Acrónimos

CAPEX CAPital EXpenditures. 1, *Glossario:* CAPEX

ISP Internet Service Provider. 1

KVM Kernel-based Virtual Machine. *Glossario:* KVM

NFV Virtual Network Function. 2

OPPEX OPerational EXpenditures. 1, *Glossario:* OPPEX

VM Virtual Machine. 1

Índice general

Índice de tablas

Índice de figuras

1
Introducción

Internet, es conocida como la red de redes. Cada año, más y más dispositivos se une a esta gran y colosal red. Una red, que desde hace mucho dejó de estar formada por solo ordenadores caseros y servidores, para ser un conjunto muy heterogéneo de dispositivos (como móviles [1], tablets e incluso neveras o relojes). Este gran aumento en la cantidad de dispositivos que conforman la red, ha sido parejo al aumento de la velocidad de la red, gracias a una gran cantidad de avances tecnológicos. Si bien, ahora es fácil poder disponer de un enlace simétrico a 100 o incluso a 300 Mbps en nuestros hogares, no podemos obviar que estos enlaces deben acabar por ser agregados en algún punto, convirtiéndose así en enlaces de 10, 40 o incluso 100 Gbps.

Aunque los enlaces de 100 Gbps comienzan a proliferan en los grandes Proveedor de Servicios de Internet (ISP), aun son complicados de encontrar en otras grandes empresas. No obstante, si que es posible encontrar algunos enlaces e incluso subredes enteras a 10 Gbps. Mantener en funcionamiento redes de alta velocidad ($\geq$10 Gbps) requiere de una infraestructura de red con un alto gasto de capital inical (CAPEX) y un alto gasto de capital para operar (OPPEX). Dicha infraestructura, debe ser además monitorizada con regularidad para asegurar el correcto funcionamiento.

Toda la infraestructura de red así como el equipamiento de monitorización requieren, de forma tradicional, grandes y potentes máquinas dedicadas a realizar las típicas tareas de *enrutado*, *captura* o *análisis de tráfico*, entre otras. No obstante, este tipo de necesidades no es nuevo. La mayoría de los servidores del mundo funcionaban en caros servidores dedicados, hasta la llegada del Cloud Computing. El Cloud Computing virtualiza los recursos clásicos de computación permitiendo a una compañía descentralizar de forma sencilla y barata sus servidores a lo largo del mundo, así como amoldarse en capacidad de cómputo a la demanda de sus clientes. Esto, inevitablemente minimiza el CAPEX y el OPPEX que una empresa debe afrontar para operar.

Internamente los diversos sistemas de Cloud Computing utilizan máquinas virtuales (VMs) para ofrecer sus servicios. Estas máquinas virtuales, deben a su vez conectarse entre sí y el resto del mundo dando lugar a redes virtuales. Sin embargo, virtualizar de forma completa una red, no permite que escale en velocidad y tamaño fácilmente, o no al menos, sin un alto coste y consumo de CPU. Gracias a los avances en virtualización y a la aparición de la tecnología de SR-IOV, es posible virtualizar los diversos elementos de red dando lugar a las Funciones de Red Virtuales (NFVs).

1.1. Objetivos

TODO: Alcance del trabajo/proyecto

1.2. Estructura del documento

TODO: Descripción de la estructura del documento

2

Estado del Arte

TODO: Estado del arte

3
Diseño de la aplicación

TODO: Diseño del proyecto

4
Desarrollo

4.1. Entorno de Desarrollo

4.1.1. Equipamiento utilizado

4.1.2. Software utilizado

4.2. Pruebas en entorno físico

4.3. Pruebas en entornos virtuales

4.3.1. Usando Passthrough

4.3.2. Usando SR-IOV

5 Resultados

TODO: Pruebas y resultados

6 Conclusiones

TODO: Conclusiones sobre el trabajo realizado

Bibliografía

[1] Global smartphone penetration 2014. [Online]. Available: https://ondeviceresearch.com/blog/global-smartphone-penetration-2014

Apéndices

A

Towards high-performance network processing in virtualized environments

El artículo publicado se titula “ Towards high-performance network processing in virtualized environments ” y ha sido enviado al congreso HPCC, el cual está catalogado como **CORE B**. A continuación, se muestra la carta de aceptación, las revisiones y finalmente el artículo en sí.

A.1. Email de aceptación

```
 Fecha:  6 de junio de 2015, 21:29:31 CEST
 De:  HPCC 2015 <hpcc2015@easychair.org>
 Para:  Víctor Moreno <victor.moreno@uam.es
 Asunto: HPCC 2015 notification for paper 51


Dear Víctor,

Thank you for your contribution to IEEE HPCC 2015.

Congratulations! Your paper #51, titled as "Towards high-performance network
processing in virtualized environments", has been officially accepted as a
full paper for publication and presentation in IEEE HPCC 2015.

Please check reviewers’ comments, and prepare your final version based on IEEE
conference proceeding format. The size of your final paper should be up to 6
pages complimentary, or 12 pages with the over length charge.

The completed reviews are attached below.

In order to publish your article, please prepare your final camera ready version
```

according to the requirements on IEEE HPCC website.

We will send you the detailed instruction about your final submission in another mail later.

Please prepare your trip (visa, air ticket, hotel) in advance. See you at New York City.

Best,
PC Chairs of IEEE HPCC 2015

A.2. Revisiones

A.2.1. Revisor 1

OVERALL EVALUATION: 0 (borderline paper)
REVIEWER'S CONFIDENCE: 4 (high)

----------- REVIEW -----------
This paper presents a study on some of the existing network processing techniques in different configurations. By comparing the package capturing capabilities and performance, the authors claim that they provide the audience a series of guidelines for network application deployment with low cost, high performance, space efficiency, and low power consumption.

My major concern about the paper is its contribution. A large body of the paper is devoted to presenting a preliminary investigation of several well known technologies (PF_RING, Intel DPDK, HPCAP, etc.), omiting the detailed introduction to author's own framework HPCAPvf. The authors present their experimental results for various configurations and techniques. However, I feel there is a lack of the experimental details and methodology. To make their results more convincing, the authors should include a methodology section dedicated to elaborate their experimental approach. Too much space is taken to introduce the results, system structures and tradeoffs for basic, well know concepts such as huge pages, package handling in virtualized network, etc. Figure 4 and 5 take too much space. The author should use more space to introduce the details and techniques specific to this work. Regarding the presented results, why the authors use percentage of captured packages as a "perf!ormance" indicator of the compared techniques? I think this metric should be used to indicate the accuracy, not the performance.

The paper is not well written. There are many typos and mistakes. Just to name a few: In the sentence "traditional end-user network applications retrieves packets individually by means of system calls, which in Linux systems involves at least two context switches.", retrieves should be retrieve, involves should be involve. This sentence is long and reads awkward: "Those tests were made using the same hardware as previously mentioned and using KVM for creating and managing the VMs, and accordingly to the bare-metal scenario, the table depicts each capture engine's performance for the worst case scenario and in an average scenario.", and accordingly should be according. There are many other places that the author should do a proofreading.

A.2.2. Revisor 2

OVERALL EVALUATION: 2 (accept)
REVIEWER'S CONFIDENCE: 2 (low)

----------- REVIEW -----------
This paper evaluates the performance of several well-known high-performance capture engines on virtual machines. In addition, the paper presents an open-source version of the HPCAP for virtual environments. The experimental results presented in this paper are valuable for network operators that want to employ network virtual functions.

The paper is well written and easy to understand even for a non-expert in this field. The experimental results are based on a realistic benchmark and show insightful results.

A.3. Artículo

Puede verse el artículo en la siguiente página.

Towards high-performance network processing in virtualized environments

Victor Moreno, Rafael Leira, Ivan Gonzalez and Francisco J. Gomez-Arribas
High Performance Computing and Networking research group,
Universidad Autónoma de Madrid, Spain
{victor.moreno, rafael.leira, ivan.gonzalez, francisco.gomez}@uam.es

***Abstract*—Nowadays, network operators' amenities are populated with a huge amount of proprietary hardware devices for carrying out their core tasks. Moreover, those networks must include additional hardware appliances as they host more and more services and applications offered by third-party actors. Thus, such an infrastructure reduces the profitability, as including new hardware boxes in the network becomes increasingly harder in terms of space, cooling and power consumption. In a context in which virtualization has become a ubiquitous technique, industry and academia has turned an eye to Network Function Virtualization (NFV) to mitigate those effects and maximize potential earnings. NFV aims to transform future network architectures by exploiting standard IT virtualization technology to consolidate a variety of network processing elements onto standard commodity servers. On this work, we assess the feasibility of moving high-performance network processing tasks to a virtualized environment. For such purpose, we analyse the possible configurations that allow feeding the network traffic to applications running inside virtual machines. For each configuration, we compare the usage of different high-performance packet capture engines on virtual machines, namely PF_RING, Intel DPDK and HPCAP. Specifically, we obtain the performance bounds for the primary task, packet sniffing, for physical, virtual and mixed configurations. We also developed HPCAPvf, a counterpart of HPCAP for virtual environments, and made it available under a GPL license.**

Keywords*—*Virtual network functions, packet capture, virtual machines.

I. Introduction

Over the past decades, the use of the Internet has rapidly grown due to the emergence of new services and applications. The amount of services and applications available to end-users makes it necessary for those services' providers to deploy quality-assessment policies in order to distinguish their product among the rest. In this scenario, network processing and analysis becomes a central task that has to deal with humongous amounts of data at high-speed rates. Obviously, service providers must be able to accomplish such a challenging task using processing elements capable of reaching the required rates while keeping the cost as low as possible for the sake of profitability.

To deal with such amount data, specialized hardware (HW) solutions have been developed, which are traditionally based on the used of Field-Programmable Gate Arrays (FPGAs) or Network Processors. These solutions answer the high-performance needs for specific network monitoring tasks, e.g. routing or classifying traffic on multi-Gb/s links [1], [2]. However, those solutions imply high investments: such hardware's elevated cost rises capital expenditures (CAPEX), while operational expenditures (OPEX) are increased due to the difficulty of their operation, maintenance and evolution. Furthermore, HW life cycles become shorter as technology and services evolve, which prevents new earnings and limits innovation. Those drawbacks turn specialized HW solutions into a non-desirable option for large-scale network processing.

As an alternative to mitigate those negative effects, academia and industry turned an eye to the use of commodity hardware in conjunction with open source software [3]. Such combination is referred as an off-the-shelf system in the literature. The advantages of those systems lay in the ubiquity of those components, which makes it easy and affordable to acquire and replace them and consequently reduces CAPEX. Those systems are not necessarily cheap, but their wide-range of application allows their price to benefit from large-scale economies and makes it possible to achieve great degrees of experimentation. Additionally, such systems offer extensive and high-quality support, thus reducing OPEX. However, the increased demands for network processing capacity would be translated into a big number of machines even though if off-the-shelf systems were used. Such an amount of machines means high expenses in terms of power consumption and physical space. Moreover, the presence of commodity servers from different vendors empowers the appearance of interoperability issues. All those drawbacks damage the profitability that networked service providers may experience.

On the other hand, there has been an increasing trend regarding the use of virtualization for computational purposes. Such trend has been empowered by the inherent advantages provided by virtualization solutions [4]. In this light, network operators and service providers have been working during the last years on the development of the concept of Network Function Virtualization (NFV). This new paradigm aims to unify the environments where network applications shall run by means of adding a virtualization layer on top of which network applications may run. This novel philosophy also allows merging independent network applications using unique hardware equipment. Consequently, the application on NFV can increase the benefits obtained by network service providers by (i) reducing their equipment investment by acquiring large-scale manufacturers' products and by reducing the amount of physical machines required, which also entails cuts in power consumption; (ii) speeding up network applications maturation cycle as all applications are developed in an unified environment; (iii) easing maintenance procedures and expenditures as testability is radically enhanced; (iv) opening the network applications' market for small companies and academia by minimizing risk and thus encouraging innovation; (v) the possibility to rapidly adjust network applications and resources based on specific clients requirements.

The development of this novel NFV philosophy has been favoured by other trending technologies such as Cloud Computing and Software Defined Networking (SDN). In the case of Cloud Computing, NFV can easily benefit from all the research carried out on virtualization management [5]. Furthermore, NFV is to reside inside Cloud providers' networks in order to carry out all the network-related management tasks. On the other hand, NFV is complementary to SDN but not dependent on it and vice-versa [6]. However, NFV enhances SDN as it provides the infrastructure on top of which SDN software can run.

However, in order to make the most of NFV, mechanisms that allow obtaining maximum network processing throughput in such virtual environments must be developed. In a bare-metal scenario, researchers have recently focused on developing high-performance packet capture engines [3]. This work evaluates the feasibility of applying such capture engines in NFV-based environments for the primal network task: packet capture. Specifically, we evaluate different virtualization alternatives, namely PCI-passthrough and Network Virtual Functions (NVF) available on contemporary systems. For each capture engine and environment, we measure the performance bounds, so researchers and practitioners may benefit from our results in order to build their high-performance NFV-based applications.

II. Network processing on bare-metal scenarios

In the recent years both the research community and industry have paid attention to the use of off-the-shelf for network processing purposes [7]. Those systems offer interesting features that allow reducing CAPEX and OPEX when building a system on top of them. In terms of network processing, off-the-shelf systems have traditionally relied on the use of the corresponding NIC vendor's driver plus a standardized network stack. Such approach is characterized by a great degree of flexibility, as the network stack provides independent layers that allow distributing the traffic to the corresponding final applications. Nevertheless, the performance obtained by such approaches is poor: every incoming packet must traverse a set of layers, which is translated into additional copies, resource re-allocations at processing time. Thus, this flexibility the standard solution offers limits its applicability in high-speed scenarios.

Consequently, if off-the-shelf systems are to be applied in high-speed networked scenarios, they have to be carefully planned and tuned. In this light is how high-performance packet capture engines were born [3]. Solutions such as PF_RING, PacketShader, netmap, PFQ, Intel DPDK or HPCAP were created as high-performance counterparts for the traditional network driver-plus-stack alternative. All those solution are based on some of the following ideas (see [3], [8] for a detailed discussion):

- *Pre-allocation and re-use of memory*: traditional solutions allocate a set of structures and buffers for each received packet. Those resources are also released once the packet is delivered to upper layers. Such technique is a very demanding task and may be optimized by pre-allocating pools of resources for re-using them along time.
- *Memory mapping*: this technique makes it possible for high-level applications to map buffers and data structures allocated at driver-level. Consequently, the number of copies experimented by incoming packets is reduced.
- *Use of parallel direct paths*: modern NICs support have multiple parallel reception queues and to distribute incoming traffic among such queues using RSS (Receive Side Scaling) mechanisms. However, the use of contemporary network stacks becomes a bottleneck serializing all the traffic at one single point for delivering it to upper layers. In this light, high-performance network solution must avoid serialization point for really exploiting NIC's parallel nature. Note that the use of parallel paths may cause incoming packet reordering [9], so it must be carefully planned.
- *Batch processing*: traditional end-user network applications retrieves packets individually by means of system calls, which in Linux systems involves at least two context switches. In order to mitigate this effect, some capture engines pretend processing several packets with a single system call. Solutions such as PacketShader, netmap or Intel DPDK group packets into batches, and so all packets conforming the batch are handled in the same system call. Note that applying such technique, in spite of its performance improve-

ment, may entail side effects such as latency increase and inaccurate timestamping [10]. For this reason, solutions such as HPCAP propose a byte-stream oriented approach.

- *Prefetching*: this technique consists in preloading memory locations in processors' caches in a predictive way so that it can be quickly accessed in a near future. Its application reduces the number of cache misses in capture process and thus leverages performance.
- *Affinity planning*: contemporary server feature NUMA architectures, where applications performance is highly influenced by the NUMA node or processor it is executed on. Thus, the affinity of all threads and processes involved in a capture system must be carefully scheduled. Such schedule must also take into account the connectivity of the processors to the PCI slot the NICs are connected to.

All the above-mentioned solutions enable network managers to deploy high-performance network applications on top of them. Table I shows the performance level obtained for packet capture in a full-saturated 10 Gb/s link for the worst-case scenario (64 byte packets) and an average scenario (a CAIDA from a ISP's backbone link between Chicago and Seattle obtained the 19^{th} June 2014, with an average packet size of 965 bytes [11]). We have compared the performance of the standard `ixgbe` plus network stack solution compared to PF_RING, Intel DPDK and HPCAP. We chose PF_RING as an archetype for the previously mentioned capture engines, and Intel DPDK and HPCAP for being the ones supporting virtual environments. Importantly, HPCAP has been developed by the authors as capture engine focused not only on high-performance packet capture rates, but also on accurate packet timestamping [10] and on building a multi-granular multi-purpose monitoring framework [12].

The tests have been carried out using a server with two Xeon E5-2630 processors running at 2.6 Ghz and 32 GB of DDR3 RAM. The NIC used was an Intel 82599 card connected to a PCIe Gen3 slot. The operating system in use is a Linux Fedora 20 with a 3.14.7 kernel. The results obtained show that all four capture engines are capable of capturing 100% of the packets when replying the CAIDA trace at top-speed. In the worst-case scenario, `ixgbe` captures only 2.7% of the incoming packets, while PF_RING and Intel DPDK capture 100% of them, and HPCAP captures 97.9% of the packets. Note that the performance degradation experienced by HPCAP is due to the driver timestamping each incoming packet.

Nevertheless, the application of the existing packet capture engines has been limited in the literature to the bare-metal case. That is, a physical server to which the NIC is connected through a PCIe slot. This configuration corresponds to the leftmost NIC shown on Fig. 1. Consequently, the application of the high-performance packet capture engines has not been evaluated in NFV environments yet.

Configuration	Bare-metal		PCI pass-through	
Traffic	64 byte	CAIDA	64 byte	CAIDA
ixgbe	2.7 %	100 %	1.9 %	62.7 %
PF_RING	100 %	100 %	100 %	100 %
DPDK	100 %	100 %	100 %	100 %
HPCAP	97.9 %	100 %	82.5 %	100 %

TABLE I: Percentage of packets captured for different traffic patterns in a fully-saturated 10 Gb/s link obtained by different solutions in a bare-metal and PCI pass-through configuration

III. Network processing in virtualized environments

When using virtual machines (VMs) for computing intensive applications, the performance obtained by the target application is usually damaged. For this reason, if we desire to obtain maximum performance, the creation and schedule of each VM must be carefully made. For example, the amount of cores of the VM should be such that allows the VM to be executed inside a single NUMA node in the physical server, and those virtual cores should be configured to be mapped to independent physical cores. Another determinant factor for VM's performance is optimizing how the VM accesses the physical system's memory. Contemporary VM managers allow attaching the VM's memory to certain NUMA node, which will increase overall memory access latency. However, studies such as [13] point out the importance of using Linux's huge pages for allocating the VM's memory chunk so the amount of page misses is reduced and performance is enhanced.

By using Linux huge pages both the packet capture engines and the network applications built on top of them running on a VM would experience a performance increase. To confirm this effect, we empirically tested the impact on the percentage of captured packets when using different capture approaches from VMs running over huge pages and not (those experiments were made using the PCI pass-through technique, see subsection III-B). We experienced a slight performance improvement when running them on top of huge pages. Note that those were simple experiments in which only packet capture were made, but by using Linux huge pages network applications built on top of those VMs would also see their performance boosted regardless the capture procedure used. All the VMs used along the performance experiments presented in this paper were created taking these facts into account. Regarding the operating system running inside the VMs, we used the same version as in the physical server, a Linux Fedora 20 with a 3.14.7 kernel. The choice of the operating systems to be used

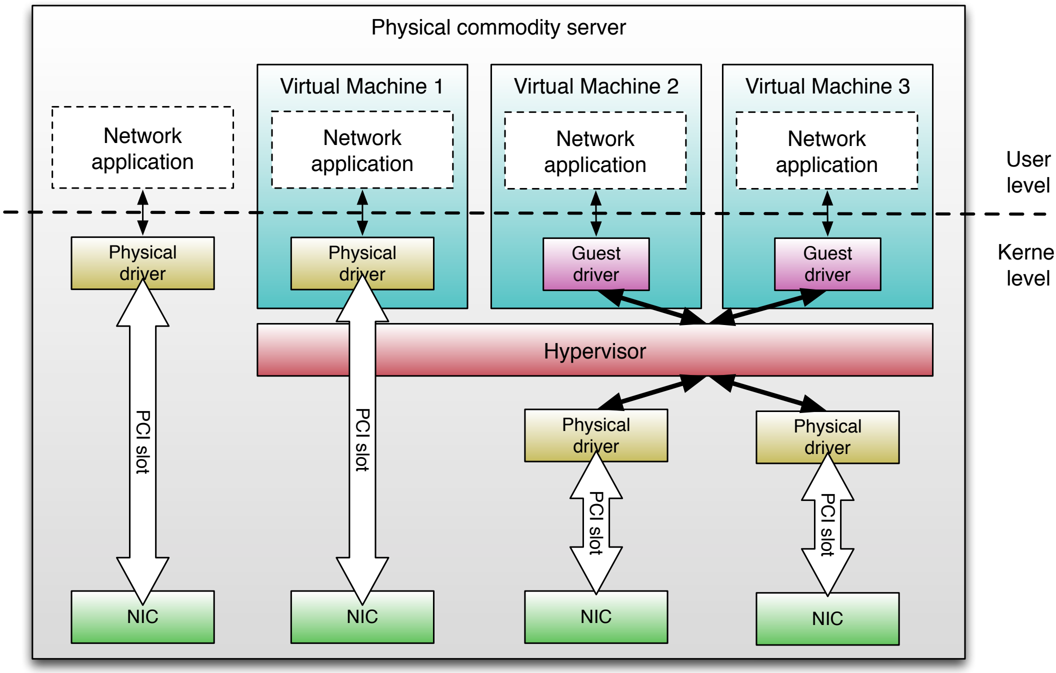

Fig. 1: Using a network device in bare-metal (leftmost), PCI pass-through (VM1) and emulated NIC (VM2, VM3)

both in the physical and virtual machines is relevant, as not all combinations support the use of Linux huge pages both in the physical server for creating the VM as mentioned and inside the VM. Note that some of the capture engines used such as PF_RING and Intel DPDK make use of those huge pages, so this requirement is nontrivial.

A. *VMs with emulated NICs*

In order to enable network applications running on top of NFV environments, we must understand the way a network device can be connected to a virtual machine (VM). The traditional scenario is composed by a set of virtual machines where an emulated network device is instantiated. In this configuration, incoming packets are captured by the NIC driver on the physical machine and traverse the system's network stack, and then delivered to the hypervisor's network module. Once acquired by the hypervisor, packets are delivered to the target VM depending on the virtual network configuration. This scenario is the one exhibited by VMs 2 and 3 on Fig. 1. Note that in this case the network driver used in the VM would be the one corresponding to the emulated device, and not the physical device that lies below. Such configuration implies at least two additional copies: from the physical server to the hypervisor and from the hypervisor to the virtual machine, which obviously degrades the performance achieved by the capture system.

Furthermore, the performance degradation experienced by network applications running in this configuration has pushed academia and industry to tune and optimize the hypervisor's packet handling policy. In this line, authors of [14] introduce VALE, which proposes a set of modification both on physical drivers and hypervisors in order to improve the network processing performance. By applying their proposals they leverage the system throughput: from 300 Mb/s using the conventional approach to nearly 1 Gb/s using VALE in the worst-case scenario (64 byte UDP packets); and from about 2.7 Gb/s to 4.4 Gb/s for TCP traffic. Although the results obtained by this approach are promising, the authors focus their attention on data exchange between VMs on the same physical server.

Note that both the traditional approach and VALE are base on the hypervisor as the central communication element. Consequently, the hypervisor becomes the bottleneck and a single point-of-failure for network processing tasks, limiting their applicability for network monitoring purposes. The use of an emulated NIC approach forces incoming packets to be processed twice: once when they arrive to the hypervisor, and again when they are transferred to their corresponding VM.

B. *VMs and PCI pass-through*

As an alternative to a hypervisor-centric approach, different hardware manufacturers developed a set of mechanism which enable direct connectivity from the

PCI adapter, referred as physical function, to the virtual machines providing near-native performance. The name given to those mechanisms depends of the underlying manufacturer: VT-d for Intel and ARM, Vi for AMD, but the technique is usually referred as PCI pass-through. This technique has been applied in several computational scenarios [15]. Applied to the network capture problem, this technology allows the VMs to map physical specific PCIe memory regions. Using this feature, VMs can operate as if they had the NIC physically connected to them, as shown by VM1 on Fig. 1. This implies that the drivers managing the NICs in the VMs are the same used to manage those NICs in the bare-metal scenario. Consequently, this allows network applications being executed in virtual machines to benefit from the high-performance packet capture solutions developed for bare-metal scenarios.

Table I shows the performance results in a fully-saturated link for the default `ixgbe` driver, PF_RING, Intel DPDK and HPCAP when executed in a VM with PCI pass-through compared to the previously mentioned bare-metal scenario. Those tests were made using the same hardware as previously mentioned and using KVM for creating and managing the VMs, and accordingly to the bare-metal scenario, the table depicts each capture engine's performance for the worst-case scenario and in an average scenario. The amount of packets captured by `ixgbe` falls from 2.7% under the bare-metal configuration to 1.9% using PCI pass-through in the worst-case scenario, and from 100% to 37.3% when replaying the previously mentioned CAIDA trace. On the other hand, PF_RING and Intel DPDK show no performance degradation when used via PCI pass-through, as they capture 100% of the packets on all scenarios. When it comes to HPCAP, packet capture performance is damaged when using PCI pass-through in the worst-case scenario, in which the amount of packets captured falls from 97.9% to 85.2%. The performance penalty experienced when introducing PCI pass-though can be blamed on the execution of the capture system in a virtual machine.

However, using PCI pass-through for instantiating network applications in different VMs has an inherent constraint: only one VM can make use of each network interface. That means that the scalability problem of instantiating more VMs for network processing purposes must be solved by adding additional NICs and probably more physical servers, as the amount of PCIe slots a server has is limited.

C. VMs attached to virtual functions

With the goal of promoting virtualization performance and interoperability the PCI Special Interest Group developed a series of standards, which they called Single-Root I/O Virtualization (SR-IOV). Those standards define the concept of virtual function (VF) as a lightweight image of the underlying physical PCI resource. Modern NIC such as the Intel 82599 use the concept of Virtual Machine Device Queues (VMDq) to refere to the NIC's virtual functions. By using VMDqs, the traffic arriving to the physical network device can be distributed among a set of queues based on a set rules that can be configured at hardware. Each of those queues is attached to a virtual PCI device, or virtual function (VF), that can be mapped via PCI pass-through by a certain VM. This configuration is represented in Fig. 4. Note that this configuration allows connecting an arbitrary number of VMs to a single physical device. The amount of virtual functions generated per physical device is limited by the hardware device, being 32 for the Intel 82599 adapter.

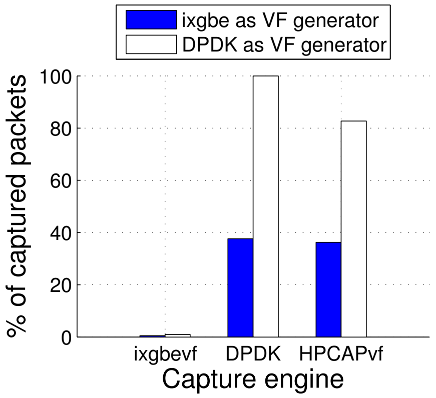

Fig. 2: Packet capture performance obtained when capturing from a 10 Gb/s link fully-saturated with 64-byte packets using virtualized alternatives for different virtual-function generators

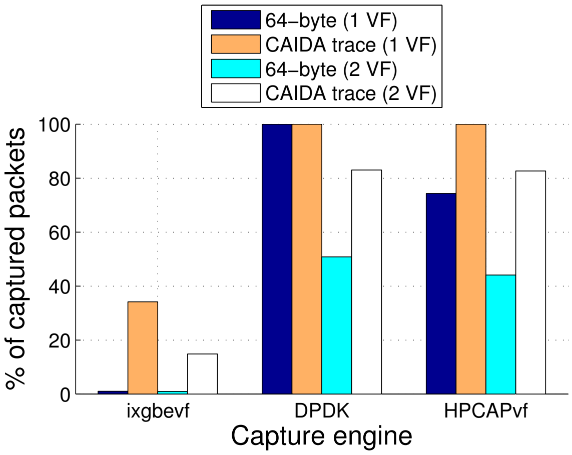

Fig. 3: Packet capture performance obtained by different virtualized alternatives when capturing different traffic in a fully-saturated 10 Gb/s link

It is worth remarking that if this approach is used, only VF-aware drivers can be used in the VMs, so they handle the peculiarities those devices have. This requirement limits the amount of capture engines avail-

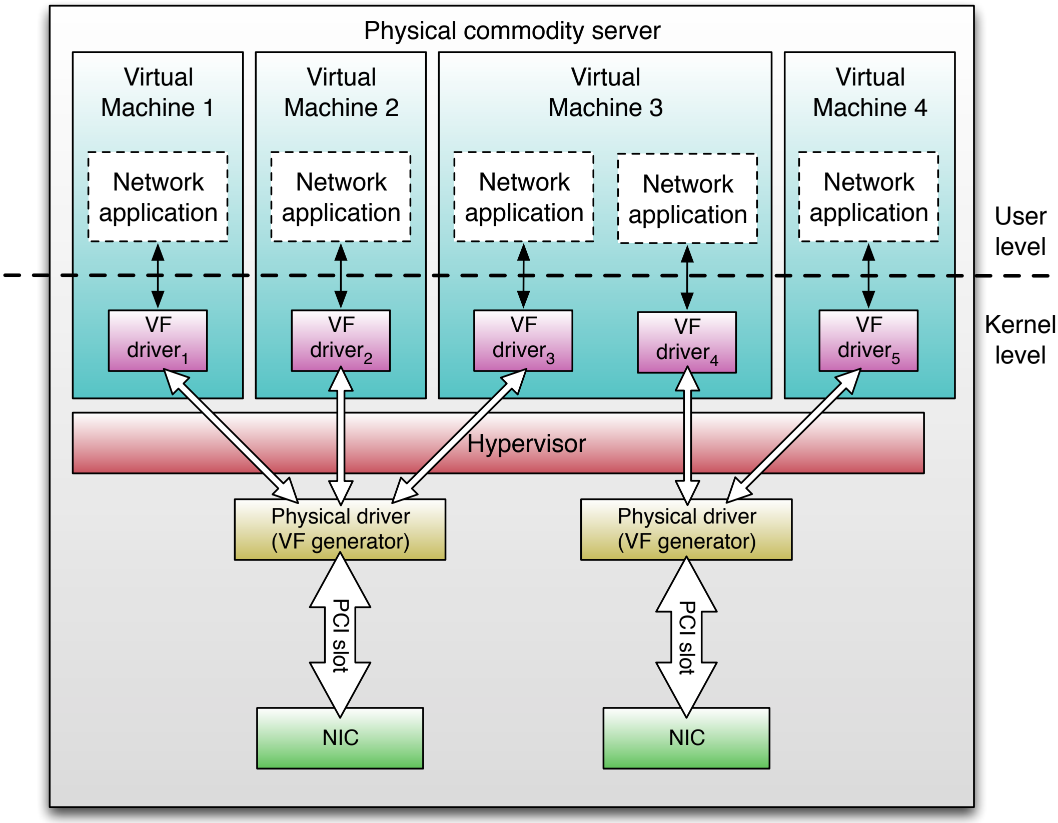

Fig. 4: Using a network device in a VM via virtual network functions

able. Specifically, Intel's DPDK has native support for working with VF, and they also supply a VF-aware counterpart to the `ixgbe` driver, named `ixgbevf`. Additionally, we developed a VF-aware version of HPCAP, that we named HPCAPvf, following all the design principles that guided HPCAP's design. Those three drivers have consequently been the only ones we have been able to test under this configuration.

On the other hand, the task of creating and managing the mentioned VF belong to the driver used in the physical server for managing the physical NIC. Above all the drivers previously mentioned capable of managing a physical NIC, only Intel's offer this feature. Consequently, when using VF only `ixgbe` and Intel DPDK are eligible. Importantly, the choice among those two drivers for generating the VF has an impact on the performance obtained by possible network applications running inside the VMs. Fig. 2 shows the effect of such choice when using different VF-aware for packet capturing in the worst-case scenario, that is a fully-saturated 10 Gb/s with 64-byte packets. Results show that using DPDK as VF generator improves the packet capture performance obtained from the VM side compared to the performance when using `ixgbe` for generating those VF. Specifically, when capturing packets in a VM using the `ixgbevf` driver using DPDK as VF generator raises the amount of packets captured from 0.5% to 1%. In a similar manner, when using DPDK and HPCAPvf in the VM side with `ixgbe` as VF generator, only 37.6% and 36.3% are respectively captured, but those ciphers are increased to 100% and 82.7% respectively when DPDK is used as VF generator.

When instantiating several VF through a single physical interface, the default traffic distribution policy is based on the MAC and IP addresses of each VM's interface the VF are connected to. Thus, each VF would only receive the traffic targeted to its corresponding VM. This would limit the use of this VF approach in scenarios where different network applications running on independent VMs need to be fed the same input traffic, which is the desirable case for scalable network monitoring. However, NICs such as Intel's 82599 supply a Mirroring and Multicast Promiscuous Enable (MPE) flags, which can be activated for each VF. By enabling those options, any VF could receive all the traffic traversing the physical device, or the traffic corresponding to a different VM, regardless it is targeted to its VM or not. Note that enabling those features in the Intel 82599 NIC implies a hardware-level packet replication, which minimizes the impact on the capture process' performance. Depending on the physical driver used to generate the NIC's VFs, activating the MPE may be done by tuning the driver's source code (that is the case when using `ixgbe` for generating the VFs) or by using a user-level application giving access to the NIC's registers (such as the `testpmd` application offered by Intel's DPDK).

Obviously, if several VMDqs are to be fed the same incoming packets, the physical driver will have to issue additional copies for each additional VMDq, and packet capture performance may be degraded. This effect is shown in the yellow and red bars of Fig. 3: adding a second VF to each physical device reduces the overall packet capture throughput obtained. When using `ixgbe` the amount of packets captured is 10% when using either one or two VFs in the worst-case scenario, but falls from 34.2% with one VF to 14.8% with two when replaying the CAIDA trace. Intel DPDK also suffer performance loss as it is capable of capturing all of the packets in both scenarios when only one VF is instantiated, but it captures 50.8% of the 64-byte packets and 83.0% of the CAIDA ones when two VFs are used. Finally, HPCAP's performance falls in both cases: from 74.3% to 44.1% in the worst case and from 100% to 82.7% for the CAIDA trace.

IV. Application scenarios

In the light of the discussion and results presented along section III, we strongly recommend the usage of the VF-based approach for processing network-data. However, if there is a primal need for performance, users may find useful to connect their NICs and their VMs via PCI pass-through. We discourage the use of an emulated NIC configuration, as this would imply unnecessary redundant computation and limits the capture performance being the hypervisor the system's bottleneck. By using the VF approach, not only a reasonably high performance is achieved but also the traffic targeted to each VM is isolated and we acquire the ability of redirecting which may be interesting in a number of scenarios. Note that the migration process from a traditional (emulated NIC) configuration to any of the other two approaches only implies driver modifications in the host server and changing the VMs' default network interface for the new one (be it a physical one via pass-through or a virtual one).

Finally, we have identified three different usage scenarios in which the VF-based alternative would apply, which are also depicted in Fig. 5:

1) Users may run on their VM one of the network-processing applications provided by Intel DPDK or HPCAP, as in the Nth VM in Fig. 5.
2) Users may create their own network processing application based on the existing APIs, just as in VM3 in Fig. 5.
3) Users already running a set of VMs with a set of legacy network-related applications needing to monitor the traffic directed to those legacy applications. In this scenario a new VM could be created, adding the proper HW rules for re-directing the desired traffic to the new VF. This new VM could use a high-performance VF-aware driver with a monitoring application. This is depicted by VMs 1,2 and N in Fig. 5. The dashed lines are used to remark that this new VM could be dynamically created or destroyed without affecting the rest of VMs already running in the same host server. Importantly, this scenario would not require any changes in the legacy network applications.

V. Conclusions

The results obtained along the experiments presented in this work assess the feasibility of migrating the usage of high-performance packet capture engines in virtualized environments. We have discussed the different configurations by which a network application can be used inside a virtual machine, and obtained the performance bounds for each of those configurations. Moreover, we have given a set of guidelines that allow exploiting the functionality of generating virtual network functions, enabling a set of interesting scenarios. Differently from PCI pass-through, the use of VF allows end-users to scale in the amount of network applications running on a single hardware, with the consequent saving in terms of space, cooling and power consumption.

We have presented a set of solutions available for use under GNU Linux and, in the case of HPCAP and Intel DPDK, accessible as free software. As an additional contribution, we have developed HPCAPvf, a VF-aware version of HPCAP, offering a set of interesting features and capabilities. HPCAPvf may be used in any Linux distribution with kernel version newer than or equal to 2.6.32 without modifying nor kernel nor hardware configuration. Furthermore, we have made available the source code of HPCAPvf under a GPL license[1]. This study paves the way for future works involving advanced network actions in the VM side, such as packet storage, at high-speed rates.

Acknowledgements

This work has been carried out with the support of the Spanish National I+D *Packtrack* project (TEC2012-33754) and Universidad Autonoma de Madrid's multidisciplinary project *Implementacion de Modelos Computacionales Masivamente Paralelos* (CEMU-2013-14).

References

[1] F. Yu, R.H. Katz, and T.V. Lakshman. Gigabit rate packet pattern-matching using TCAM. In *Proceedings of IEEE Conference on Network Protocols*, 2004.

[1] https://github.com/hpcn-uam/HPCAP

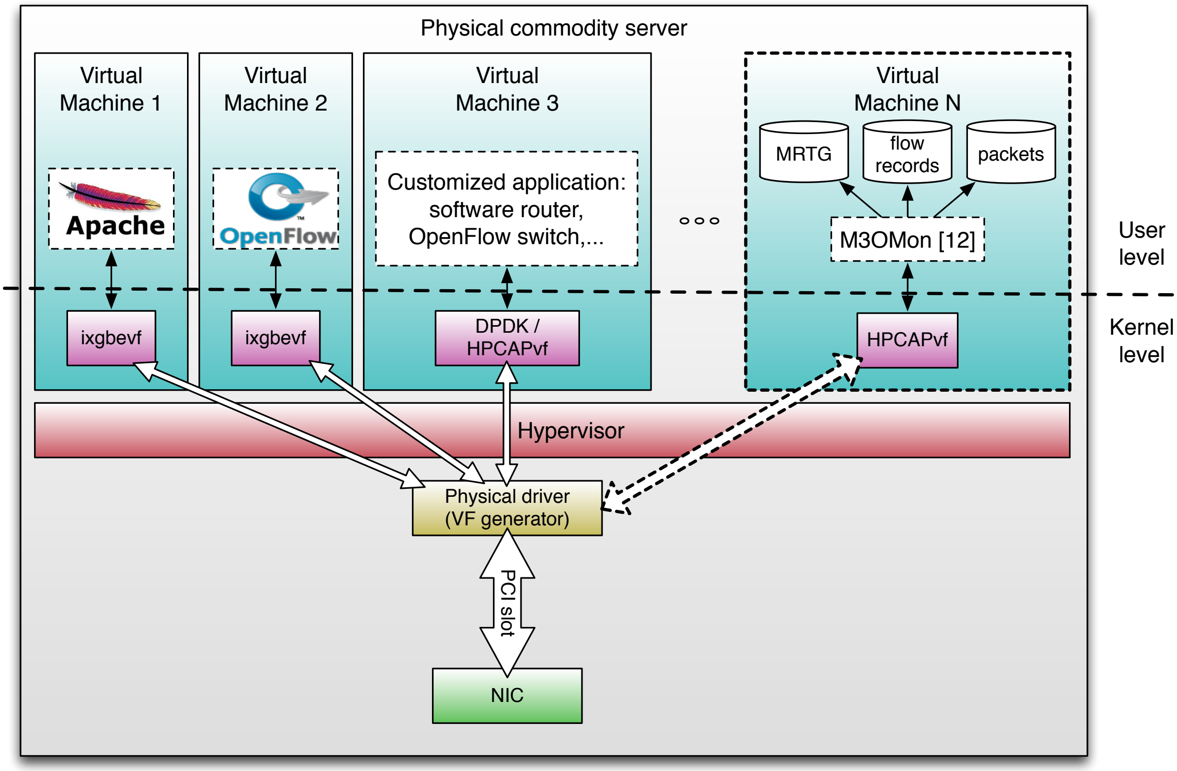

Fig. 5: Example usage cases

[2] M. Forconesi, G. Sutter, S. Lopez-Buedo, J.E. Lopez de Vergara, and J. Aracil. Bridging the gap between hardware and software open-source network developments. *IEEE Network*, 28(5), 2014.

[3] J.L. García-Dorado, F. Mata, J. Ramos, P.M. Santiago del Río, V. Moreno, and J. Aracil. High-performance network traffic processing systems using commodity hardware. In *Data Traffic Monitoring and Analysis*, volume 7754 of *Lecture Notes in Computer Science*, pages 3–27. Springer Berlin Heidelberg, 2013.

[4] AJ. Younge, R. Henschel, J.T. Brown, G. von Laszewski, J. Qiu, and G.C. Fox. Analysis of virtualization technologies for high performance computing environments. In *Cloud Computing (CLOUD), 2011 IEEE International Conference on*, pages 9–16, July 2011.

[5] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. A view of cloud computing. *Commun. ACM*, 53(4):50–58, April 2010.

[6] C. Monsanto, J. Reich, N. Foster, J. Rexford, and D. Walker. Composing software-defined networks. In *Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation*, nsdi'13, pages 1–14, Berkeley, CA, USA, 2013. USENIX Association.

[7] L. Braun, A. Didebulidze, N. Kammenhuber, and G. Carle. Comparing and improving current packet capturing solutions based on commodity hardware. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 206–217, New York, NY, USA, 2010. ACM.

[8] Intel. Intel Data Plane Development Kit (Intel DPDK) Release Notes, 2014. http://www.intel.com/content/dam/www/public/us/en/documents/release-notes/intel-dpdk-release-notes.pdf, [1 October 2014].

[9] W. Wenji, P. DeMar, and M. Crawford. Why can some advanced Ethernet NICs cause packet reordering? *IEEE Communications Letters*, 15(2):253–255, 2011.

[10] V. Moreno, P.M. Santiago del Rio, J. Ramos, J. Garnica, and J.L. Garcia-Dorado. Batch to the Future: Analyzing Timestamp Accuracy of High-Performance Packet I/O Engines. *Communications Letters, IEEE*, 16(11):1888 –1891, november 2012.

[11] C. Walsworth, E. Aben, k.c. Claffy, and D. Andersen. The CAIDA anonymized Internet traces 2014 dataset. http://www.caida.org/data/passive/passive_2014_dataset.xml, [1 October 2014].

[12] V. Moreno, P. M. Santiago del Río, J. Ramos, D. Muelas, J. L. García-Dorado, F. J. Gomez-Arribas, and J. Aracil. Multi-granular, multi-purpose and multi-Gb/s monitoring on off-the-shelf systems. *International Journal of Network Management*, 24(4):221–234, 2014.

[13] A.O. Kudryavtsev, V.K. Koshelev, and A.I. Avetisyan. Prospects for virtualization of high-performance x64 systems. *Programming and Computer Software*, 39(6):285–294, 2013.

[14] L. Rizzo, G. Lettieri, and V. Maffione. Speeding up packet i/o in virtual machines. In *Proceedings of the Ninth ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, ANCS '13, pages 47–58, Piscataway, NJ, USA, 2013. IEEE Press.

[15] C.T. Yang, J.C. Liu, H.Y. Wang, and C.H. Hsu. Implementation of gpu virtualization using pci pass-through mechanism. *The Journal of Supercomputing*, 68(1):183–213, 2014.