NFVnice: Dynamic Backpressure and Scheduling for NFV Service Chains

Sameer G. Kulkarni, Wei Zhang, Jinho Hwang, Shriram Rajagopalan,
K. K. Ramakrishnan, *Fellow, ACM*, Timothy Wood, Mayutan Arumaithurai,
and Xiaoming Fu, *Senior Member, IEEE, Member, ACM*

***Abstract*— Managing Network Function (NF) service chains requires careful system resource management. We propose *NFVnice*, a user space NF scheduling and service chain management framework to provide fair, efficient and dynamic resource scheduling capabilities on Network Function Virtualization (NFV) platforms. The NFVnice framework monitors load on a service chain at high frequency (1000Hz) and employs backpressure to shed load early in the service chain, thereby preventing wasted work. Borrowing concepts such as rate proportional scheduling from hardware packet schedulers, CPU shares are computed by accounting for heterogeneous packet processing costs of NFs, I/O, and traffic arrival characteristics. By leveraging cgroups, a user space process scheduling abstraction exposed by the operating system, NFVnice is capable of controlling when network functions should be scheduled. NFVnice improves NF performance by complementing the capabilities of the OS scheduler but without requiring changes to the OS's scheduling mechanisms. Our controlled experiments show that NFVnice provides the appropriate rate-cost proportional fair share of CPU to NFs and significantly improves NF performance (throughput and latency) by reducing wasted work across an NF chain, compared to using the default OS scheduler. NFVnice achieves this even for heterogeneous NFs with vastly different computational costs and for heterogeneous workloads.**

***Index Terms*— Network function virtualization, service function chaining, scheduling, backpressure, fairness.**

I. Introduction

NETWORK Function Virtualization (NFV) seeks to implement network functions and middlebox services such as firewalls, NAT, proxies, deep packet inspection, WAN optimization, etc., in software instead of purpose-built hardware appliances. These software based network functions can be run on top of commercial-off-the-shelf (COTS) hardware, with virtualized network functions (NFs). Network functions, however, often are chained together [1], where a packet is processed by a sequence of NFs before being forwarded to the destination.

The advent of container technologies like Docker [2] enables network operators to densely pack a single NFV appliance (VM/bare metal) with large numbers of network functions at runtime. Even though NFV platforms are typically capable of processing packets at line rate, without efficient management of system resources in such densely packed environments, service chains can result in serious performance degradation because bottleneck NFs may drop packets that have already been processed by upstream NFs, resulting in wasted work in the service chain.

NF processing has to address a combination of requirements. Just as hardware switches and routers provide rate-proportional scheduling for packet flows, an NFV platform has to provide a fair processing of packet flows. Secondly, the tasks running on the NFV platform may have heterogeneous processing requirements that OS schedulers (unlike hardware switches) address using their typical fair scheduling mechanisms. OS schedulers, however, do not treat packet flows fairly in proportion to their arrival rate. Thus, NF processing requires a re-thinking of the system resource management framework to address both these requirements. Moreover, standard OS schedulers: a) do not have the right metrics and primitives to ensure fairness between NFs that deal with the same or different packet flows; b) do not make scheduling decisions that account for chain level information; and c) cannot guarantee predictable per-flow latency requirements. If the scheduler allocates more processing to an upstream NF and the downstream NF becomes overloaded, packets are dropped by the downstream NF. This results in inefficient processing and wasting the work done by the upstream NF. OS schedulers also need to be adapted to work with user space data plane frameworks such as Intel's DPDK [3]. They have to be cognizant of NUMA (Non-uniform Memory Access) concerns of NF processing and the dependencies among NFs in a service chain. Additionally, processor performance is critically dependent on cache performance, which in turn depends on locality of reference [4]. When the OS switches contexts, locality of access may not occur because the instructions and data of the newly-scheduled

Manuscript received December 27, 2018; revised October 6, 2019; accepted January 11, 2020; approved by IEEE/ACM Transactions on Networking Editor G. Paschos. This work was supported in part by the EU FP7 Marie Curie Actions CleanSky ITN Project under Grant 607584, in part by the U.S. NSF under Grant CRI-1823270, Grant CNS-1522546, and Grant CNS-1422362, in part by the Department of the Army, U.S. Army Research, Development and Engineering Command under Grant W911NF-15-1-0508, and in part by Huawei Technologies Company, Ltd., under the HIRP Grant. *(Corresponding author: Sameer G. Kulkarni.)*

Sameer G. Kulkarni and K. K. Ramakrishnan are with the Department of Computer Science and Engineering, University of California at Riverside, Riverside, CA 92521 USA (e-mail: sameer.sameergk@gmail.com).

Wei Zhang is with Microsoft Azure Networking, Microsoft, Redmond, WA 98052 USA.

Jinho Hwang is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA.

Shriram Rajagopalan is with Tetrate, San Francisco, CA 94111 USA.

Timothy Wood is with the Department of Computer Science, The George Washington University, Washington, DC 20052 USA.

Mayutan Arumaithurai and Xiaoming Fu are with the Institut für Informatik, University of Göttingen, 37077 Göttingen, Germany.

Digital Object Identifier 10.1109/TNET.2020.2969971

NF may no longer be in the cache(s). Context switching results in additional NF processing costs, beyond the typical cost associated with the operations performed by the kernel [4] and have to be accounted for. Therefore, determining how to dynamically schedule NFs is key to achieving high performance and scalability for diverse service chains, especially in a scenario where multiple NFs are contending for a CPU core.

Hardware routers and switches that employ sophisticated scheduling algorithms such as rate proportional scheduling [5], [6] have predictable performance per-packet, because processing resources are allocated fairly to meet QoS requirements and bottlenecks are avoided by design. However, NFV platforms are necessarily different because: a) the OS scheduler does not know a priori, the capacity or processing requirements for each NF; b) an NF may have variable per-packet costs (*e.g.,* some packets may trigger DNS lookup, which are expensive to process, and others may just be an inexpensive header match). With NFV service chains, there is a need to be aware of the computational demands for packet processing. There can also be sporadic blocking of NFs due to I/O (read/write) stalls, that also results in latency variation across the processed packets.

A further consideration is that routers and switches 'simply' drop packets when congested. However, an NF in a service chain that drops packets can result in considerable wasted processing at NFs earlier in the chain. These wasted resources could be gainfully utilized by other NFs being scheduled on the same CPU core to process other packet flows. We posit that a scheduling framework for NFV service chains has to simultaneously account for both task level scheduling on processing cores and packet level scheduling. This combined problem is what poses a challenge: *When you get a packet, you have to decide which task has to run, and also which packets to process, and for how long*.

To solve these problems we propose NFVnice, an NFV management framework that provides fair and efficient resource allocations to NF service chains. NFVnice focuses on the scheduling and control problems of NFs running on shared CPU cores, and considers a variety of realistic issues such as bottlenecked NFs in a chain, and the impact of NFs that perform disk I/O accesses, which naturally complicate scheduling decisions. NFVnice makes the following contributions:

- Automatically tuning CPU scheduling parameters to provide a fair allocation that weighs NFs based on both their packet arrival rate and the required computation cost.
- Determining when NFs are eligible to get a CPU share and when they need to yield the CPU, entirely from user space, improving throughput and fairness regardless of the kernel scheduler being used.
- Leveraging the scheduling flexibility to achieve backpressure for service chain-level congestion control, that avoids unnecessary packet processing early in a chain if the packet might be dropped later on.
- Extending backpressure to apply not only to adjacent NFs in a service chain, but for full service chains and managing congestion across hosts using ECN.
- Presenting a scheduler-agnostic framework that does not require any operating system or kernel modifications.

We have implemented NFVnice (source code [7]) on top of OpenNetVM [8], a DPDK-based NFV platform that runs NFs in separate processes or containers to facilitate deployment. Our evaluation shows that NFVnice can support different kernel schedulers, while substantially improving throughput and providing fair CPU allocation based on processing requirements. In controlled experiments using the vanilla CFS scheduler [9], NFVnice reduces packet drops from 3Mpps (million packets per second) to just 0.01Mpps during overload conditions. NFVnice provides performance isolation for TCP flows when there are competing UDP flows, improving throughput of TCP flows from 30Mbps to 4Gbps, without penalizing UDP flows, by avoiding wasted work. Further, our evaluations demonstrate that NFVnice, because of the dynamic backpressure, is resilient to the variability in packet-processing cost of the NFs, yielding considerable improvement in throughput and latency even for the large service chains (including chains that span multiple cores).

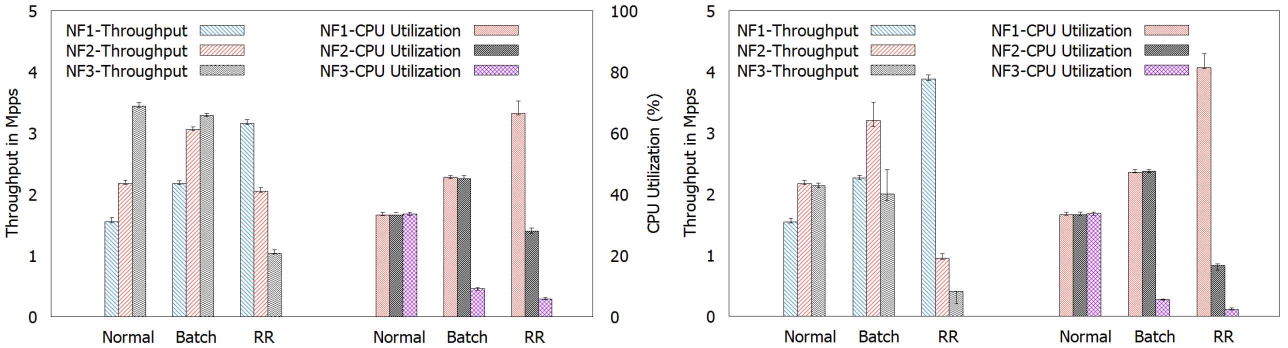
II. Background and Motivation

A. Diversity, Fairness, and Chain Efficiency

The middleboxes that are being deployed in industry are diverse in their applications as well as in their complexity and processing requirements. ETSI standards [10] show that NFs have dramatically different processing and performance requirements. Measurements of existing NFs show the variation in CPU demand and per packet latency: some NFs have per-core throughput in the order of million packets per second (Mpps), *e.g.,* switches; others have throughputs as low as a few kilo pps, *e.g.,* encryption engines.

Fair Scheduling: Determining how to allocate CPU time to network functions in order to provide fair and efficient chain performance despite NF diversity is the focus of our work. Defining "fairness" when NFs may have drastically different requirements or behavior is important. We leverage the work on Rate Proportional Servers [5], [6]. We define the allocation to be rate-cost proportionally fair if the allocation ensures the same normalized service to all the contending NFs, *i.e.,* we apportion the resources (CPU cycles) to NFs based on the combination of each NF's arrival rate and processing cost. Intuitively, if either one of these factors is fixed, then we expect its CPU allocation to be proportional to the other metric. For example, if two NFs have the same computation cost but one has twice the arrival rate of the other, then it must have twice the output rate relative to the second NF. Alternatively, if the NFs have the same arrival rate, but one requires twice the processing cost, then we expect the heavy NF to get twice as much CPU time, resulting in both NFs having the same output rate. This definition of fairness can of course be supplemented with a prioritization factor. This provides an understandable and consistent way to provide differentiated service for NFs that is proportional to the arrival rate *and* processing cost.

Unfortunately, standard CPU schedulers do not have sufficient information to allocate resources in a way that provides rate-cost proportional fairness. CPU schedulers typically try to provide fair allocation of processing time, but if computation costs vary between NFs this cannot provide rate-cost fairness. Therefore, NFVnice must enhance the scheduler with more information so that it can appropriately allocate CPU time to provide correctly weighted allocations. We adopt the notion of rate-cost proportional fairness for two fundamental reasons: i) it ensures that all competing NFs get a minimal CPU share necessary to make progress even in the worst case scenario (highly uneven and overloaded across competing NFs), while seeking to maximize the throughput for a given load across all



(a) Even load distribution.

(b) Uneven load distribution.

Fig. 1. The scheduler alone is unable to provide fair resource allocations that account for processing cost and load.

the NFs; and ii) the rate-cost proportional fairness is general and flexible, so that it can be tuned to meet the QoS policies desired by the operator. Further, the approach ensures that when contending NFs include malicious NFs (those that fail to yield), or misbehaving NFs (get stuck in a loop making no progress), such NFs do not consume the CPU excessively, impeding the progress of other NFs. While the Linux default scheduler has the notion of a virtual run-time for each running task, we fine-tune that capability to provide the correct share of the CPU for an NF, rather than simply allocating an equal share of the CPU to each contending NF.

Efficient Chaining: Beyond simply allocating CPU time fairly to NFs on a single core, the combination of NFs into service chains demands careful resource management across the chain to minimize the impact of bottlenecks. Processing a packet only to have it dropped from a subsequent bottleneck's queue is wasteful, and a recipe for *receive livelock* [11], [12].

When an NF (whether a single NF or one in a service chain) is overloaded, packet drops become inevitable, and processing resources already consumed by those packets are wasted. For responsive flows, such as TCP, congestion control and avoidance using packet drop methods such as RED, REM, SFQ, CSFQ [13]–[16] and feedback with Explicit Congestion Notification (ECN) [17] can cause the flows to adapt their rates to the available capacity on an end-to-end basis. However, for non-responsive flows (e.g., UDP), a local, rapidly adapting method is backpressure, which can propagate information regarding a congested resource upstream (to previous NFs in the chain). It is important however to ensure that effects such as head-of-the-line blocking or unfairness do not creep in as a result.

B. Existing OS Schedulers Are Ill-Suited for NFV Deployment

Linux provides several different process schedulers, with the Completely Fair Scheduler (CFS) [9] being the default since kernel 2.6.23. In this work we focus on three schedulers: i) CFS Normal, ii) CFS Batch, and Round Robin. The CFS class of schedulers use a nanosecond resolution timer to provide fine granularity scheduling decisions. Each task in CFS maintains a monotonically increasing virtual run-time which determines the order and quantum of CPU assignment to these tasks. The time-slice is not fixed, but is determined relative to the run-time of the contending tasks in a time-ordered red-black tree [18], [19]. The task with the smallest run-time (the left most node in the ordered red-black tree) is scheduled to run until either the task voluntarily yields, or consumes the allotted time-slice. If it consumes the allocated time-slice, it is re-inserted into the red-black tree based on its cumulative run-time consumed so far. The CFS scheduler is analogous to weighted fair queueing (WFQ) scheduling [20], [21]. Thus, CFS ensures a fair proportion of CPU allocation to all the tasks. The CFS Batch variant has fewer timer interrupts than normal CFS, leading to a longer time quantum and fewer context switches, while still offering fairness. The Round Robin (RR) scheduler (part of the linux real-time (RT) scheduling class), simply cycles through processes with a specified time quantum (1-100ms), but does not focus on a particular measure of fairness other than equal allocation of cycles. The CFS class of schedulers readily use the cgroups to provide CPU bandwidth control per-process (or group of processes), while the RT schedulers do not support CPU bandwidth control for group scheduling [9].

To explore the impact of these schedulers on NFV applications we consider a simple deployment with three NF processes sharing a CPU core. We look at two workloads: 1) equal offered load (of 5 Mpps) to all NFs; 2) unequal offered load, with NF1, NF2 getting 6 Mpps, and NF3 getting 3 Mpps.

We consider three heterogeneous NFs (computation costs: NF1 = 500, NF2 = 250 and NF3 = 50 CPU cycles) subject to equal and unequal loads. Figure 1 shows that when arrival rates are the same, none of the schedulers are able to provide our fairness goal—an equal output rate for all three NFs. CFS Normal always apportions CPU equally, regardless of offered load and NF processing cost, so the lighter weight NF3 gets the highest throughput. The RR scheduler gives each NF an equal chance to run, but does not limit the time the NF runs for. The CFS Batch scheduler is in between these extremes since it seeks to provide fairness, but over longer time periods. Notably, the Batch scheduler provides NF3 almost the same throughput as Normal CFS, despite allocating it substantially less CPU. The reason for this is that Normal CFS can incur a very large number of context switches due to its goal of providing very fine-grained fairness. Since Batch mode reduces scheduler preemption, it has substantially fewer non-voluntary context switches—reducing from 65K to 1K per second—as illustrated in the Table I. While RR also has low context switch overhead, it allows heavy weight NFs to greedily consume the CPU, nearly starving NF3.

We also demonstrate the impact on latency due to the scheduling of NFs in a service chain. For this, we consider a chain of three heterogeneous NFs executed on a same CPU

TABLE I

CONTEXT SWITCHES FOR HETEROGENEOUS NFS

	Even Load						Uneven Load					
	SCHED_ NORMAL		SCHED_ BATCH		SCHED_ RR		SCHED_ NORMAL		SCHED_ BATCH		SCHED_ RR	
NF	csw-ch/s	nvc swch /s	csw-ch/s	nv cswch /s	csw-ch/s	nvc swch /s	csw-ch/s	nvc swch /s	csw-ch/s	nvc swch /s	csw-ch/s	nvc swch /s
NF1	0	33785	0	504	198	7	0	38585	0	503	85	10
NF2	0	32214	1	505	204	2	0	41089	4	496	92	1
NF3	65796	107	1010	8	206	0	79479	85	1004	4	93	0

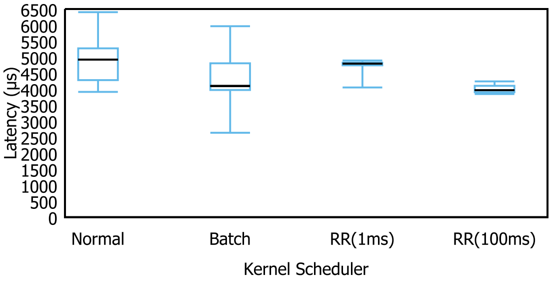


Fig. 2. Minimum, maximum, and three quartiles (25%ile, median and 75%ile) latency for different kernel schedulers.

core, and measure the round-trip-time (RTT) latency for packet processing across the chain (the time from packet generation to receiving it back, after processing). Figure 2 shows the box plot of the latency seen with different schedulers. The choice of scheduler has significant impact on the latency. Moreover, the variance (min, max, and the three quartiles) in latency is much higher with the CFS (Normal and Batch) schedulers that perform more frequent context switches compared to the RR schedulers (1ms or 100ms).

These results show that just having the Linux scheduler handle scheduling NFs has undesirable results as by itself it is unable to adapt to both varying per-packet processing requirements of NFs and packet arrival rates. Further, it is important to avoid the overheads of excessive context switches. All of these scheduling requirements must be met on a per-core basis, while accounting for the behavior of chains spanning multiple cores or servers.

III. DESIGN AND IMPLEMENTATION

In an NFV platform, at the top of the stack are one or more network functions that must be scheduled in such a way that idle work (i.e., while waiting for packets) is minimized and load on the service chain is shed *as early as possible* so as to avoid wasted work. However, the operating system's process scheduler that lies at the bottom of the software stack remains completely application agnostic, with its goal of providing a fair share of system resources to all processes. As shown in the prior section, the kernel scheduler's metrics for scheduling are along orthogonal dimensions to those desired by the network functions. NFVnice bridges the gap by translating the scheduling requirements at the NFV application layer to a format consumable by the operating system.

The design of NFVnice centers around the concept of assisted preemptive scheduling, where network functions provide hints to the underlying OS with regard to their utilization. In addition to monitoring the average computation time of a network function per packet, NFVnice needs to know when NFs in a chain are overloaded, or blocked on packet/disk I/O. The queues between NFs in a service chain serve as a good indicator of pending work at each NF. To facilitate the process of providing these metrics from the NF implementation to the underlying operating system, NFVnice provides network function implementations with an abstraction library called *libnf*. In addition to the usual tasks such as efficient reading/writing packets from/to the network at line rate and overlapping processing with non-blocking asynchronous I/O, *libnf* co-ordinates with the NFVnice platform to schedule/de-schedule a network function as necessary.

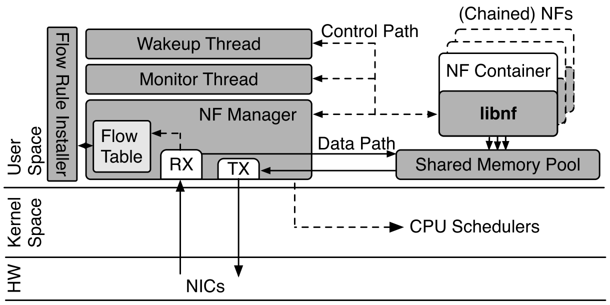


Fig. 3. NFVnice building blocks.

Modifying the OS scheduler to be aware of various queues in the NFV platform is an onerous task that might lead to unnecessary maintenance overhead and potential system instability. One approach is to change the priority of the NF based on the queue length of packet at that NF. This will have the effect of increasing the number of CPU cycles provided to that NF. This will require the change to occur frequently as the queue length varies. The change requires a system call, which consumes CPU cycles and adds latency. In addition, with service chains, as the queue at an upstream NF builds, its priority has to be raised to process packets and deliver to a queue at the downstream NF. Then, the downstream NF's priority will have to be raised. We believe that this can lead to instability because of frequent changes and the delay involved in effecting the change. This only gets worse with complex service chains, where an NF is both an upstream NF for one service chain and a downstream NF for another service chain. Instead, NFVnice leverages cgroups [22], [23], a standard userspace primitive provided in linux to manipulate process scheduling. NFVnice monitors queue sizes, computation times and I/O activities in user space with the help of *libnf* and manipulates scheduling weights accordingly.

A. System Components

Figure 3 illustrates the key components of the NFVnice platform. We leverage DPDK for fast userspace networking [3]. Our NFV platform is implemented as a system of queues that hold packet descriptors pointing to shared memory regions. The NF Manager runs on a dedicated set of cores and is responsible for ferrying packet references between the network interface card (NIC) queues and NF queues in an efficient manner. When packets arrive to the NIC, Rx threads in the NF Manager take advantage of DPDK's poll mode driver to deliver the packets into a shared memory region accessible to all the NFs. The Rx thread does a lookup in the Flow Table to direct the packet to the appropriate NF. Once a flow is matched to an NF, packet descriptors are copied into the NF's receive ring buffer and the Wakeup subsystem brings the NF process into the runnable state. After being processed by an NF, the NF Manager's Tx threads move packets through the remainder of the chain. This provides zero-copy packet movement.

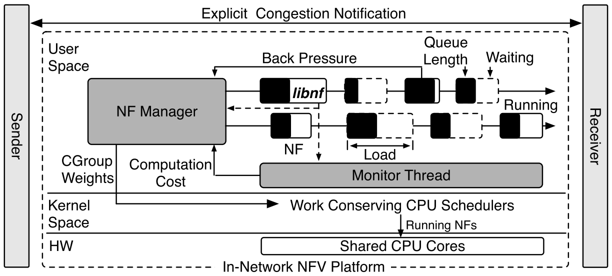


Fig. 4. NF scheduling and backpressure.

Service chains can be configured during system startup using simple configuration files or from an external orchestrator such as an SDN controller. When an NF finishes with a packet, it enqueues it in its Tx queue, where it is read by the manager and redirected to the Rx queue of the next NF in the chain. The NF Manager also picks up packets from the Tx queue of the last NF in the chain, and sends it out over the network. We have designed NFVnice to provide high performance processing of NF service chains. The NF Manager's scheduling subsystem determines when an NF should be active and how much CPU time it should be allocated relative to other NFs. The backpressure subsystem provides chain-aware management, preventing NFs from spending time processing packets that are likely to be dropped downstream.

System Management and NF Deployment: The NF Manager 's (Rx, Tx and Monitor) threads are pinned to separate dedicated cores. The number of Rx, Tx and monitor threads are configurable (`C-Macros`), to meet system needs, and available CPU resources. Similarly, the maximum number of NFs and maximum chain length can be configured. NFVnice allows NFs and NF service chains to be deployed as independent processes or Docker containers which are linked with *libnf* library. *libnf* exports a simple, minimal interface (9 functions, 2 callbacks and 4 structures), and both the NF Manager and *libnf* leverage the DPDK libraries (ring buffers, timers, memory management). We believe developing or porting NFs or existing docker containers can be reasonably straightforward. For example, a simple bridge NF or a basic monitor NF is less than 100 lines of `C` code.

B. Scheduling NFs

Each network function in NFVnice is implemented inside its own process (potentially running in a container). Thus the OS scheduler is responsible for picking which NF to run at any point in time. We believe that rather than design an entirely new scheduler for NFV, it is important to leverage Linux's existing scheduling framework, and use our management framework in user space to tune any of the stock OS schedulers to provide the properties desired for NFV support. Figure 4 shows the NFVnice scheduling that makes the OS scheduler be governed by NF Manager via cgroups, and ultimately assigns running NFs to shared CPU cores. The detailed description of the figure is in the Sections III-B and III-C.

Activating NFs: NFs that busy-wait for packets perform very poorly in a shared CPU environment. Thus it is critical to design the NF framework so that NFs are only activated when there are packets available for them to process, as is done in NFV platforms such as netmap [24] and ClickOS [25]. However, these systems provide only a relatively simple policy for activating an NF: once one or more packets are available, a signal is sent to the NF so that it will be scheduled to run by the OS scheduler in netmap, or the hypervisor scheduler in ClickOS. While this provides an efficient mechanism for waking NFs, neither system allows for more complex resource management policies, which can lead to unfair CPU allocations across NFs, or inefficient scheduling across chains.

In NFVnice, NFs sleep by blocking on a semaphore shared with the NF Manager, granting the management plane great flexibility in deciding which NFs to activate at a given time. The policy we provide for activating an NF considers the number of packets pending in its queue, its priority relative to other NFs, and knowledge of the queue lengths of downstream NFs in the same chain. This allows the management framework to indirectly affect the CPU scheduling of NFs to be fairness and service-chain aware, without requiring that information be synchronized with the kernel's scheduler.

Relinquishing the CPU: NFs process batches of packets, deciding whether to keep processing or relinquish the CPU between each batch. This decision and all interactions with the management layer, e.g., to receive a batch of packets, are mediated by *libnf*, which in turn exposes a simple interface to developers to write their network function. After a batch of at most 32 packets is processed, *libnf* will check a shared memory flag set by the NF Manager that indicates if it should relinquish the CPU early (e.g., as a result of backpressure, as described below). If the flag is not set, the NF will attempt to process another batch; if the flag has been set or there are no packets available, the NF will block on the semaphore until notified by the Manager. This provides a flexible way for the manager to indicate that an NF should give up the CPU without requiring the kernel's CPU scheduler to be NF-aware.

CPU Scheduler: Since multiple NF processes are likely to be in the runnable state at the same time, it is the operating system's CPU scheduler that must determine which to run and for how long. In the early stages of our work we sought to design a custom CPU scheduler that would incorporate NF information such as queue lengths into its scheduling decisions. However, we found that synchronizing queue length information with the kernel, at the frequency necessary for NF scheduling, incurred overheads that outweighed any benefits.

NFVnice carefully controls when individual NF processes are runnable and when they yield the CPU (as described above), the batch scheduler's longer time quantum and less frequent preemption are desirable. In most cases, NFVnice NFs relinquish the CPU due to policies controlled by the manager, rather than through an involuntary context switch. This reduces overhead and helps NFVnice prioritize the most important NF for processing without requiring information sharing between user and kernel space.

Assigning CPU Weights: NFVnice provides mechanisms to monitor a network function to estimate its CPU requirements, and to adjust its scheduling weight. Policies in the NF Manager can then dynamically tune the scheduling weights assigned to each process in order to meet operator specified priority requirements.

The packet arrival rate for a given NF can be easily estimated by either the NF or the NF Manager. We measure the service time to process a packet inside each NF using *libnf*. To avoid outliers from skewing these measurements (e.g., if a context switch occurs in the middle of processing a packet),

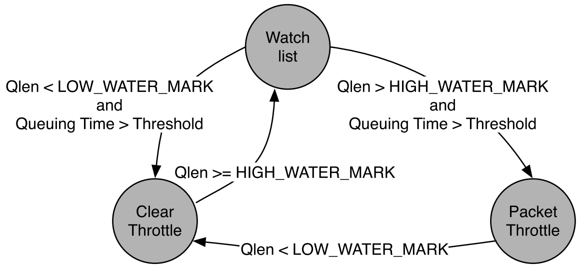

Watch list
Qlen < LOW_WATER_MARK and Queuing Time > Threshold
Qlen > HIGH_WATER_MARK and Queuing Time > Threshold
Qlen >= HIGH_WATER_MARK
Clear Throttle
Packet Throttle
Qlen < LOW_WATER_MARK

Fig. 5. Backpressure state diagram.

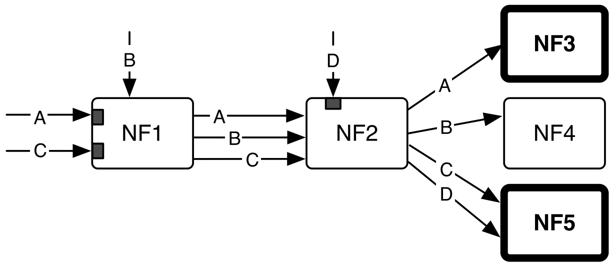


Fig. 6. Overloaded NFs (in bold) cause back pressure at the entry points for service chains A, C, and D.

we maintain a histogram of timings, allowing NFVnice to efficiently estimate the service time at different percentiles.

For each NF i on a shared core, we calculate $load(i) = \lambda_i * s_i$, the product of arrival rate, λ, and service time, s. We then find the total load on each core, such as core m, $TotalLoad(m) = \sum_{i=1}^{n} load(i)$, and assign cpu shares for NF_i on $core_m$ following the formula:

$$Shares_i = Priority_i * \frac{load(i)}{TotalLoad(m)}$$

This provides an allocation of CPU weights that provides rate proportional fairness to each NF. The $Priority_i$ parameter can be tuned if desired to provide differential service to NFs. Tuning priority in this way provides a more intuitive level of control than directly working with the CPU priorities exposed by the scheduler since it is normalized by the NF's load.

C. Backpressure

A key goal of NFVnice is to avoid wasting work, *i.e.,* preventing an upstream NF from processing packets if they are just going to be dropped at a downstream NF later in the chain that has become overloaded. We achieve this through backpressure, which ensures bottlenecks are quickly detected while minimizing the effects of head of line blocking.

Cross-Chain Backpressure: The NF Manager is in an ideal position to observe behavior across NFs since it assists in moving packets between them. When one of the NF Manager's TX threads detects that the receive queue for an NF is above a high watermark (HIGH_WATER_MARK) and queuing time is above threshold, then it examines all packets in the NF's queue to determine what service chain they are a part of. NFVnice then enables *service chain-specific* packet dropping at the upstream NFs. NF Manager maintains states of each NF, and in this case, it moves the NF's state from *backpressure watch list* to *packet throttle* as shown in Figure 5. When the queue length becomes less than a low watermark (LOW_WATER_MARK), the state moves to *clear throttle*.

The backpressure operation is illustrated in Figure 6, where four service chains (A-D) pass through several different NFs. The bold NFs (3 and 5) are currently overloaded. The NF Manager detects this and applies back pressure to flows A, C, and D. This is performed upstream where those flows first enter the system, minimizing wasted work. Note that backpressure is selective based on service chain, so packets for service chain B are not affected at all. Service chains can be defined at fine granularity (*e.g.,* at the flow-level) in order to minimize head of line blocking.

This form of system-wide backpressure offers a simple mechanism that can provide substantial performance benefits. The backpressure subsystem employs hysteresis control to prevent NFs rapidly switching between modes. Backpressure is enabled when the queue length exceeds a high watermark and is only disabled once it falls below the low watermark.

Local Optimization and ECN: NFVnice also supports simple, local backpressure, *i.e.,* an NF will block if its output TX queue becomes full. This can happen because the NF Manager TX Thread responsible for the queue is overloaded. Local backpressure is entirely NF-driven, and requires no coordination with the manager, so we use it to handle short bursts and cases where the manager is overloaded.

We also consider the fact that an NFVnice middlebox server might only be one in a chain spread across several hosts. To facilitate congestion control across machines, the NF Manager will also mark the ECN bits in TCP flows in order to facilitate end-to-end management. Since ECN works at longer timescales, we monitor queue lengths with an exponentially weighted moving average and use that to trigger marking of flows following [17].

D. Facilitating I/O

A network function could block when its receive ring buffer is empty or when it is waiting to complete I/O requests to the underlying storage. In both cases, NF implementations running on the NFVnice platform are expected to yield the CPU, returning any unused CPU cycles back to the scheduling pool. In case of I/O, NF implementations should use asynchronous I/O to overlap packet processing with background I/O to maintain throughput. NFVnice provides a simple library called *libnf* that abstracts such complexities from the NF implementation. Further details can be found in our earlier work [26].

E. Optimizations

Separating Overload Detection and Control: Since the NFV platform [27] must process millions of packets per second to meet line rates, we separate out overload detection from the control mechanisms required to respond to it. The NF Manager's Tx threads are well situated to detect when an NF is becoming backlogged as it is their responsibility to enqueue new packets to each NF's Tx queue. Using a single DPDK's enqueue interface, the Tx thread enqueues a packet to a NF's Rx queue if the queue is below the high watermark, while getting feedback about the queue's state in the return value. When overload is detected, an overload flag is set in the meta data structure related to the NF.

The control decision to apply backpressure is delegated to th NF Manager's Wakeup thread. The Wakeup thread scans through the list of NFs classifying them into two categories: ones where backpressure should be applied and ones that need to be woken up. This separation simplifies the critical path in the Tx threads and also provides some hysteresis control, since a short burst of packets causing an NF to exceeds its threshold may have already been processed by the time the Wakeup thread considers it for backpressure.

Separating Load Estimation and CPU Allocation: The load on an NF is a product of its packet arrival rate and the per-packet processing time. The scheduler weight is calculated based on the load and the cgroup's weights for the NF are updated. Since changing a weight requires writing to the Linux sysfs, it is critical that this be done outside of the packet processing data path. *libnf* merely collects samples of packet processing times, while the NF Manager computes the load and assigns the CPU shares using cgroup virtual file system.

The data plane (*libnf*) samples the packet processing time in a lightweight fashion every millisecond by observing the CPU cycle counter before and after the NF's packet handler function is called. We chose sampling because measuring overhead for each packet using the CPU cycle counters results in a CPU pipeline flush [28], resulting in additional overhead. The samples are stored in a histogram, in memory shared between *libnf* and the NF Manager. The processing time samples produced by each NF are stored in shared memory and aggregated by the NF Manager. Not all packets incur the same processing time, as some might be higher due to I/O activity. Hence, NFVnice uses the median over a 100ms moving window as the estimated packet processing time of the NF. Every millisecond, the NF Manager calculates the load on each NF using its packet arrival rate and the estimated processing time. Every 10ms, it updates the weights used by the kernel scheduler.

IV. Evaluation

A. Testbed and Approach

Our experimental testbed has Intel(R) Xeon(R) CPU E5-2697 v3 @ 2.60GHz servers, 157GB memory, running Ubuntu SMP Linux kernel 3.19.0-39-lowlatency. Each CPU has dual-sockets with a total of 56 cores. For these experiments, 3 nodes were connected back-to-back with dual-port 10Gbps DPDK compatible NICs to avoid any switch overheads.

We make use of DPDK based high speed traffic generators, Moongen [29] and Pktgen [30] as well as Iperf3 [31], to generate line rate (10Gbps) traffic consisting of UDP and TCP packets with varying numbers of flows. Moongen is configured to generate 64 byte UDP packets at line rate(∼14.2Mpps). Iperf is used to generate TCP flows with variable packet sizes.

We demonstrate NFVnice's effectiveness as a user-space solution that influences the NF scheduling decisions of the native Linux kernel scheduling policies, *i.e.,* Round Robin (RR) for the Real-time scheduling class, SCHED_ NORMAL (termed NORMAL henceforth) and SCHED_ BATCH (termed BATCH) policies in the CFS class. Different NF configurations (compute, I/O) and service chains with varying workloads (traffic characteristics) are used. For all the bar plots, we provide the average, the minimum and maximum values observed across the samples collected every second during the experiment. In all cases, the NFs are interrupt driven, woken up by NF manager when the packets arrive while NFs voluntarily yield based on NFVnice's policies. Also, when the transmit ring out of an NF is full, that NF suspends processing packets until room is created on the transmit ring.

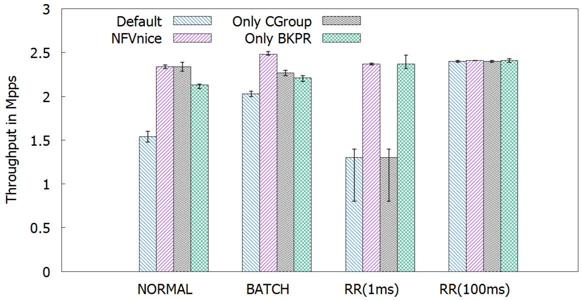


Fig. 7. Performance of NFVnice in a 3NF service chain.

TABLE II

Packet Drop Rate per Second

	NORMAL		BATCH		RR(1ms)		RR(100ms)	
	Default	NFVnice	Default	NFVnice	Default	NFVnice	Default	NFVnice
NF1	3.58M	11.2K	2M	0	0.86M	0	0.53M	0
NF2	2.02M	12.3K	0.9M	11.5K	2.92M	12K	0.03M	12K

B. Overall NFVnice Performance

We first demonstrate NFVnice's overall performance, both in throughput and in resource (CPU) utilization for each scheduler type. We compare the default schedulers to our NFVnice system, or when only including the CPU weight allocation tool (termed `cgroups`) or the `backpressure` to avoid wasted work at upstream NFs in the service chain.

1) NF Service Chain on a Single Core: Here, we first consider a sequential service chain of three NFs; with computation cost Low (NF1, 120 cycles), Medium (NF2, 270 cycles), and High (NF3, 550 cycles). All NFs run on a single shared core.

Figure 7 shows that NFVnice achieves an improvement of as much as a factor of two for throughput (especially over the RR scheduler). We also separately show the contribution of the `cgroups` and `backpressure` features. By combining both features, NFVnice improves the overall throughput across all three kernel scheduling disciplines. `cgroups only` updates the CPU share proportionally for the 3 NFs. This results in improved performance compared to using the Default (NORMAL and BATCH) schedulers. Since the round-robin scheduler (RR) does not use the `cgroups` feature, it shows no improvement. However the `backpressure` feature provides benefit independent of the underlying kernel-scheduler. Table II shows the number of packets dropped at the input of either of the downstream NFs, NF2 or NF3, after processing at the upstream node (an indication of truly wasted work). Without NFVnice, the default schedulers drop millions of packets per second. But with NFVnice, the packet drop rate is dramatically lower (near zero), demonstrating that NFVnice is effective in avoiding wasted work and providing proper CPU allocation. We also gather perf-scheduler statistics for the average scheduling delay and runtime of each of the NFs. From Table III, we can see that i) with NFVnice the run-time for each NF is apportioned in a cost-proportional manner (NF1 being least and NF3 being most), unlike the NORMAL scheduler that seeks to provide equal allocations independent of the packet processing costs. ii) the average scheduling delay with NFVnice for the NFs (that is the time taken to begin execution once the NF is ready) is lower for the NFs with higher processing time (which is exactly what is desired, to avoid making a complex NF wait to process packets, and thus avoiding unnecessary packet loss). Again this is

TABLE III

SCHEDULING LATENCY AND RUNTIME OF NFs

	NORMAL		BATCH		RR(1ms)		RR(100ms)	
measured in ms	Default	NFVnice	Default	NFVnice	Default	NFVnice	Default	NFVnice
NF1-Avg. Delay	0.002	0.112	0.003	1.613	1.022	0.730	0.924	0.809
NF1-Runtime	657.825	128.723	312.703	143.754	-	-	-	-
NF2-Avg. Delay	0.065	0.008	1.144	0.255	0.570	0.612	0.537	0.473
NF2-Runtime	602.285	848.922	836.940	803.185	-	-	-	-
NF3-Avg. Delay	0.045	0.025	0.149	0.009	0.885	0.479	0.703	0.646
NF3-Runtime	623.797	1014.218	826.203	1047.968	-	-	-	-

TABLE IV

THROUGHPUT, CPU UTIL. AND WASTED WORK OF 3NFs

	Default			NFVnice		
	Svc. rate	Drop rate	CPU Util	Svc. rate	Drop rate	CPU Util
NF1 (∼550cycles)	5.95Mpps	-	100%	0.82Mpps	-	11% ±3%
NF2 (∼2200cycles)	1.18Mpps	4.76Mpps	100%	0.72Mpps	150Kpps	64% ±1%
NF3 (∼4500cycles)	0.6Mpps	0.58Mpps	100%	0.6Mpps	70Kpps	100%
Aggregate	0.6Mpps	-	300%	0.6Mpps	-	175% ±3%

better than the behaviour of the default NORMAL and RR schedulers.[1]

2) Multi-Core Scalability: We next demonstrate the benefit of NFVnice with the NFs in a chain across cores, with each NF pinned to a separate, dedicated core. We use these experiments to demonstrate the benefits of NFVnice, namely: a) avoiding wasted work through backpressure; and b) judicious resource (CPU cycles) utilization through scheduling. When NFs are pinned to separate cores, there is no specific role/contribution for the vanilla OS schedulers, and for such an experiment we use the default scheduler (NORMAL).

First, we consider the chain of 3 NFs, NF1 (Low, 550 cycles), NF2 (Medium, 2200 cycles) and NF3 (High, 4500 CPU cycles). Compared to the default scheduler (NORMAL), NFVnice plays a key role in avoiding the wasted work and efficiently utilizing CPU cycles. Table IV shows that NFVnice's CPU utilization by NF1 and NF2 on their cores is dramatically reduced, going down from 100% to 1̃1% and 64% respectively, while maintaining the aggregate throughput (0.6 Mpps). This is primarily because of backpressure ensuring that the upstream NFs only process the correct amount of packets that the downstream NFs can consume. Excess packets coming into the chain are dropped at the beginning of the chain. When we use only the default NORMAL scheduler by itself, NF1 and NF2 use 100% of the CPU to process a huge number of packets (the 'service rate' in the Table IV), only to be discarded at the downstream NF3.

We now consider two different service chains with 4 NFs using 4 cores in the system such that each NF is pinned to a separate, dedicated core as shown in Fig. 8. Chain-1 has three NFs: NF1 (270 cycles), NF2 (120 cycles) and NF4 (300 cycles) running on 3 different cores. Chain-2 comprises NF1, NF3(4500 cycles) and NF4. Moongen generates 64-byte packets at line rate, equally splitting them between two flows that are assigned to chain-1 and chain-2. Table V shows that in the Default case (NORMAL scheduler), NF1 processes almost an equal number of packets for chain-1 and chain-2. However, for chain-2, the downstream NF3 discards a majority of the packets processed by NF1. This results not only in wasted work, but it also adversely impacts the throughput of chain-1. On the other hand, with NFVnice, backpressure has the upstream NF1 process only the appropriate number of packets of chain-2 (which has its bottleneck at the downstream NF, NF3). This frees up the upstream NF1 to use the remaining processing cycles to process packets from chain-1. NFVnice improves the throughput of chain-1 by factor of 2. At the same time, it maintains the throughput of chain-2 at its bottleneck (NF3) rate of 0.6Mpps. Overall, NFVnice not only avoids wasted work, but judiciously allocates CPU resources (at upstream NFs) proportionate to the chain's bottleneck resource capacity as shown in the Figure 9.

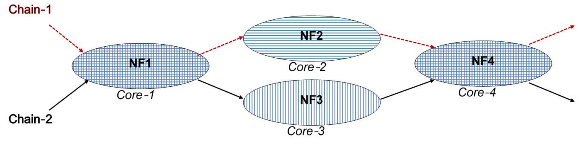


Fig. 8. Red (Chain-1) and Green (chain-2 NF chain setup).

TABLE V

THROUGHPUT, CPU UTILIZATION AND WASTED WORK IN A CHAIN OF 3 NFs (EACH NF PINNED TO A DIFFERENT CORE)

		Default			NFVNice		
		Svc.Rate (pps)	Drop Rate (pps)	CPU Util.%	Svc.Rate (pps)	Drop Rate (pps)	CPU Util.%
NF1 (∼270cycles)	Chain1	3.26M	2.86M	78.6% ±0.4	6.498M	0	82.1% ±0.5
	Chain2	3.26M			0.583M		
	Aggregate	6.522M			7.08M		
NF2 (∼120cycles)	Chain1	3.26M	∼0	52.8% ±1.2	6.498M	∼0	58% ±0.7
	Chain2	-			-		
	Aggregate	3.26M			6.498M		
NF3 (∼4500cycles)	Chain1	-	2.68M	100% ±0	-	<100	100% ±0
	Chain2	0.582M			0.582M		
	Aggregate	0.582M			0.582M		
NF4 (∼300cycles)	Chain1	3.26M	0	60% ±0.7	6.498M	0	84% ±0.7
	Chain2	0.582M			0.582M		
	Aggregate	3.842M			7.08M		

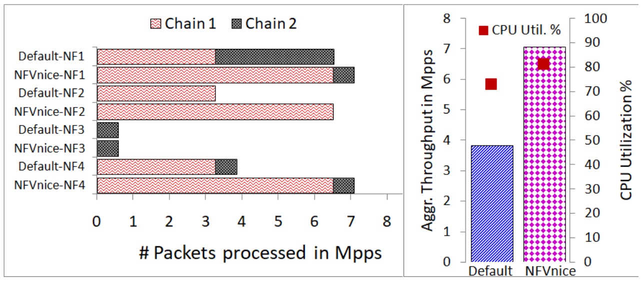


Fig. 9. Performance for NF chains shown in Fig. 8.

3) Realistic NFs With Real Data-Trace: We next demonstrate the benefit of NFVnice processing realistic traffic, as seen in a public trace collected at the Equinix-NYC monitor, from CAIDA [32]. We use a realistic NF chain. The pruned data trace consists of a large number of small-sized TCP (1388) and UDP (475) flows. We use Moongen to replay the pcap file at line rate (resulting in a packet rate of ∼2.3Mpps). In this experiment, we use the same configuration as in Fig. 8 and deploy four realistic NFs: NF1 (Monitor), NF2 (Load Balancer), NF3 (AES Encryption) and NF4 (VLAN Tagging). Chain-1 (NF1, NF2 and NF4) serves the TCP traffic to provide monitoring, vlan-tagging and load-balancing of the traffic to different backend servers. Chain-2 (NF1, NF3 and NF4) caters for UDP traffic to provide monitoring, vlan-tagging and encryption of UDP packets. To demonstrate the scheduling benefits of NFVnice, we dedicate two processing cores, so that NF1 and NF2 are pinned to a same core (core1), while the NF3 and NF4 are pinned to another core (core2).

Figure 10 shows the throughput achieved across two chains for different cases. Compared to the default case for the NORMAL scheduler, NFVnice achieves nearly 35% improvement, while with BATCH and RR(1ms) we

[1] Even though, RR(100ms) performs as well as NFVnice, it performs very poorly in other cases as seen in IV-D.1 and IV-D.2 scenarios.

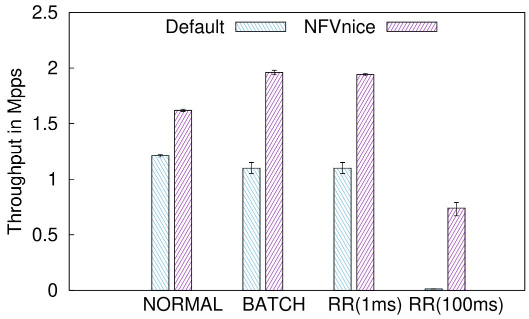


Fig. 10. Performance of NFVnice for two different service chains of 3 realistic NFs with real-world data trace.

also achieve about 60% improvement. In the default case, NF1 processes a lot more packets for both chain-1 and chain-2 than what downstream NF2 and NF3 can consume. This results in wasted work. Further, on the other core, NF4 gets considerably fewer CPU cycles compared to the contending compute-intensive NF3 (AES Encryption), especially in the RR (100ms) case, thus resulting in a significant throughput drop (less than 16Kpps) across both the chains. On the other hand, with NFVnice, backpressure ensures that the upstream NF1 only processes the appropriate number of packets for chain-1 and chain-2, thus giving more CPU cycles for NF2. `cgroups` ensures that NF4 gets sufficient CPU cycles to process the packets, resulting in better performance across all class of schedulers, with more than 500x improvement in throughput for the RR(100ms) case.

We also experimented with other shared-core and separate core placement configurations, and we consistently found NFVnice improves performance in the range of (7–75%) for all the configurations. In-fact, even when NFs were pinned to separate dedicated cores, NFVnice improves throughput by at least 7% due to the early packet dropping of `backpressure`.

C. Latency Analysis

We evaluate the impact on packet processing latency when scheduling multiple NFs of a service chain on the same core (SC) and compare it with the latency profile when running the same NFs on dedicated, distinct cores (DC). We further demonstrate the benefits of NFVnice in improving (reducing) the overall NF chain latency for both cases. For these experiments, we use the Moongen packet generator and collect the RTT samples as recommended in the benchmarking methodology for network interconnect devices [33].

Scheduling the NFs on the same core results in additional latency, but we believe it is within reasonable levels. However, the benefit of cache locality for packet processing across different NFs in the chain allows us to in fact considerably improve on the per-packet processing latency.

1) Simple 3 *NF Chain:* We present the impact of different kernel schedulers on the packet processing latency for a 3 NF chain used in experiment IV-B.1. To isolate the scheduling overheads, we also measure the latency when each NF in the chain is pinned to a separate core (represented by DC). Figure 11 shows the box plot for the latency observed with different kernel schedulers for each distinct scenario.

Default: Using the default schedulers, latency for scheduling multiple NFs on the same core (SC) is higher than running the NFs on different cores (DC) and has more variance across different schedulers. *e.g.,* worst case for CFS, the latency increased from 3.5ms to 6.5ms. This increase in latency is mainly due to context switches by the kernel schedulers.

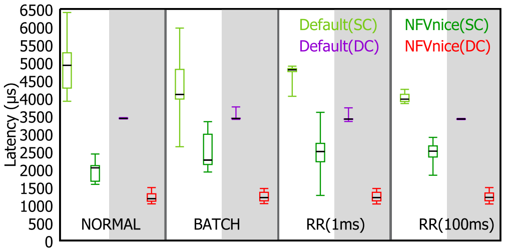


Fig. 11. Latency profile for packet processing in a 3 NF service chain. Box Plot represents the minimum, maximum, and the three quartiles (25%ile, median and 75%ile) of latency for different kernel schedulers.

NFVnice : NFVnice improves latency for all the schedulers by 50-70% across all the quartiles, including the maximum latency in both (SC and DC) the scenarios. This is primarily due to the judicious scheduling decisions of NFVnice across the NF chain, which result in the effective utilization of the CPU by allowing the processing of just the right amount of packets at each NF in the chain. NFVnice avoids additional queuing delay for the processed packets at the downstream nodes. NFVnice avoids any wasted work, avoiding the unnecessary queuing of packets at upstream nodes which are going to be eventually dropped. Further, NFVnice provides more consistent and predictable latency than the default. The latency variation with NFVnice for running the NFs on same core (SC) and different core (DC) is much smaller due to effective scheduling and avoiding unnecessary context switches.

2) Impact of Offered Load on Latency: We analyze the impact of scheduling 2 NFs of a chain on same core (SC) and also compare the latency results for running the same 2 NFs on two separate (distinct) cores (DC). We compare default with NFVnice and plot the 99th percentile latency Figure 12 for different offered loads. When the offered load is low ($\leq 1000Mbps$) the latency is similar for all the cases. Thus, scheduling NFs on the same core optimizes the utilization of CPU cores, with minimal impact on latency. However, at higher packet rates ($\geq 5000Mbps$), we observe that scheduling NFs on the same core (SC) has a steep increase in the latency, while for the DC case there is more a gradual increase in latency. Subsequently, the latency remains almost the same in both cases. This is because the overload results in excessive queuing delays at the NFs. With NFVnice, we observe similar behavior, but the latency is significantly lower across the entire offered load range, for both (SC and DC) cases.

3) Latency With Variation in Chain Computation Cost: We extend the 2 NF chain experiment and vary the per-packet computation cost of NF1 from Low (120 cycles), to Medium (270 cycles), to High (550 cycles). NF2 in all cases simply transmits the packet out. When executing NFs on the same core (SC), we observe the median and 99%ile latency to be lower than when executing them on different cores (DC) for low and medium computation cost for NF1 (results omitted due to space constraints). However, with High computation

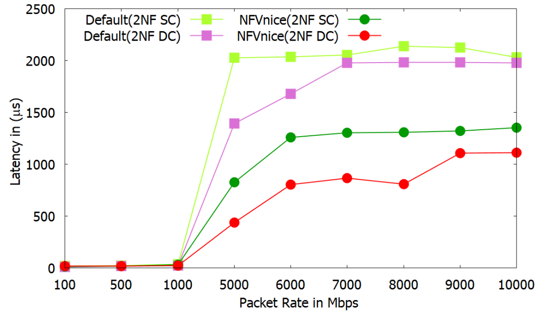


Fig. 12. 99%ile latency of 2 NF chain at different offered load.

TABLE VI

PERF-COUNTERS FOR DIFFERENT 2 NF CHAIN MODES

Characteristics	Computation Cost					
	Low		Medium		High	
	SC	DC	SC	DC	SC	DC
Instructions Per Cycle (IPC)	1.11 ±0.05%	0.47 ±0.08%	1.32 ±0.16%	0.47 ±0.07%	1.14 ±0.04%	0.48 ±0.21%
Branches (M/Sec)	542.651 ±0.05%	214.541 ±0.13%	687.361 ±0.17%	220.863 ±0.12%	606.733 ±0.04%	198.317 ±0.29%
Branch-misses (%)	0.22 ±1.43%	0.64 ±1.39%	0.15 ±1.70%	0.63 ±1.81%	0.27 ±0.68%	0.98 ±1.59%
L1-dcache-loads (M/sec)	752.737 ±0.04%	407.267 ±0.08%	772.669 ±0.14%	419.044 ±0.07%	664.478 ±0.06%	367.680 ±0.22%
L1-dcache-load-misses (%)	7.49 ±0.06%	14.63 ±0.03%	5.81 ±0.13%	14.67 ±0.05%	6.29 ±0.16%	12.09 ±0.09%
dTLB-loads (M/sec)	752.773 ±0.04%	407.426 ±0.09%	773.735 ±0.11%	419.213 ±0.08%	665.003 ±0.06%	367.581 ±0.23%
dTLB-load-misses (%)	0.31 ±1.08%	0.05 ±4.09%	0.67 ±0.49%	0.06 ±3.59%	1.53 ±0.23%	0.10 ±3.89%
iTLB-loads (M/sec)	21.920 ±1.05%	2.548 ±3.78%	30.921 ±0.50%	3.471 ±2.82%	64.382 ±0.66%	3.447 ±4.19%
iTLB-load-misses (%)	3.71 ±1.00%	3.18 ±5.22%	3.92 ±0.27%	3.88 ±4.35%	4.20 ±0.55%	3.33 ±5.82%

cost for NF1 the latency increased for SC. The system performance counters captured using the perf tool are shown in the Table VI.

With SC, the Instructions-Per-Cycle (IPC) is roughly 2-3x times better than when executing NFs on different cores. This can be attributed to effective L1 cache reference locality, which has less than 7.5% misses on data-cache. But with DC, the load misses nearly double, incurring additional stalls and per-packet processing costs, resulting in higher latencies. On the other hand, the overhead of context switching with SC results in more frequent data and instruction TLB load misses.

To summarize, when the per-packet computation cost of NFs is low (CPU is not the bottleneck) it is beneficial to schedule the NFs on a same core to reap the benefits of cache locality and to avoid the cross-core cache access overheads. But, when the computation-cost of an NF becomes a bottleneck, it is beneficial to execute the NFs on separate cores.

D. Salient Features of NFVnice

1) Variable NF Packet Processing Cost: We now evaluate the resilience of NFVnice to not only heterogeneity across NFs, but also variable packet processing costs within an NF. We use the same three-NF service chain used in IV-B.1, but modify their processing costs. Packets of the same flow have varying processing costs of 120, 270 or 550 cycles at each of the NFs. Packets are classified as having one of these 3 processing costs at each of the NFs, thus yielding 9 different variants for the total processing cost of a packet across the 3 -NF service chain. Figure 13 shows the throughput for different schedulers. With the Default scheduler, the throughput achieved differs considerably compared to the case with fixed per-packet processing costs as seen in Figure 7. For the Default scheduler, the throughput degrades considerably for the vanilla coarse time-slice schedulers (BATCH and RR(100ms)), while the NORMAL and RR(1ms) schedulers achieve relatively higher throughputs. When examining the throughput with only the CPU weight assignment, `CGroup`, we see improvement with the BATCH scheduler, but not as much with the NORMAL scheduler. This is because the variation in per-packet processing cost of NFs result in an inaccurate estimate of the NF's packet-processing cost and thus an inappropriate weight assignment and CPU share allocation. This inaccuracy also causes NFVnice (which combines `CGroup` and backpressure) to experience a marginal degradation in throughput for the different schedulers. Backpressure alone (the Only BKPR case), which does not adjust the CPU shares based on this inaccurate estimate is more resilient to the packet-processing cost variation and achieves the best (and almost the same) throughput across all the schedulers. NFVnice gains this benefit of backpressure, and therefore, in all cases NFVnice's throughput is superior to the vanilla schedulers. We could mitigate the impact of variable packet processing costs by profiling NFs more precisely and frequently, and averaging the processing over a larger window of packets. However, we realize that this can be expensive, consuming considerable CPU cycles itself. This is where NFVnice's use of backpressure helps overcome the penalty from the variability, getting better throughput and reduced packet loss compared to the default schedulers.

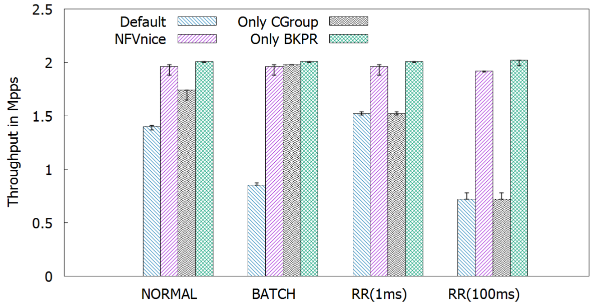


Fig. 13. Performance with service chain of 3 Heterogeneous NFs with varying per packet processing costs.

2) Service Chain Heterogeneity: We next consider a three NF chain, but vary the chain configuration—(Low, Medium, High);(High, Medium, Low); and so on for a total 6 cases—so that the location of the bottleneck NF in the chain changes in each case. Results in Figure 14 show significant variance in the behaviour of the vanilla kernel schedulers. NORMAL and BATCH perform similar to each other in most cases, except for the small differences for the reasons described earlier in Section II. We also looked at RR with time slices of 1ms and 100ms, and their performance is vastly different. For the small time-slice, performance is better when the bottleneck NF is upstream, while RR with a larger time-slice performs better when the bottleneck NF is downstream. This is primarly due to wasted work and inefficient CPU allotment to the contending NFs. However, with NFVnice, in almost every case, we can see considerable improvements in throughput, for all the schedulers. NFVnice minimizes the wasted cycles independent of the OS scheduler's operational time-slice.

Impact of RR's Time Slices with NFV: Consider the chain configurations "High-Med-Low" and "Med-High-Low" in Figure 14. RR(100 ms time slice) performs very poorly, with very low throughput $< 40Kpps$. This is due to the 'Fast-producer, slow-consumer' situation [34], making the NF with

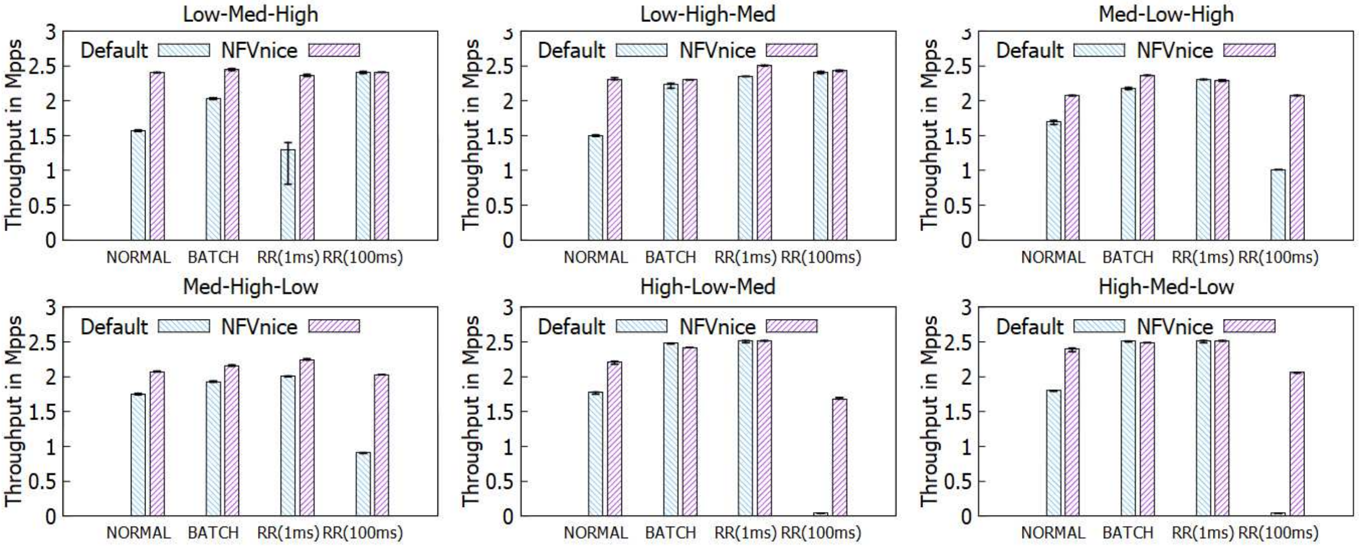


Fig. 14. Throughput for varying combinations of 3 NF service chain with Heterogeneous computation costs.

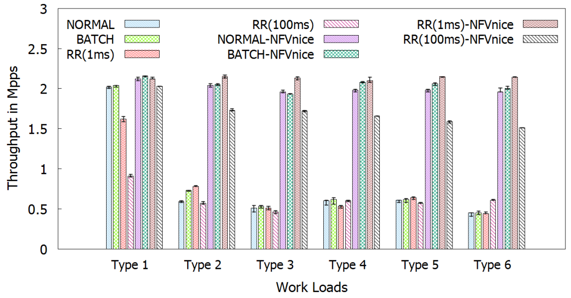


Fig. 15. Throughput (Mpps) in a 3 NF service chain for different combinations (types) and mix of workload.

"High" computes hog the CPU resource. Now, in the default RR scheduler, the packets processed by this NF would be dequeued by the Tx threads but will be subsequently dropped, as the next NF in the chain does not get an adequate share of the CPU to process these packets. The upstream NF that is hogging the CPU has to finish its time slice and the OS scheduler then causes a involuntary context switch for this "High" NF. However, with NFVnice, the queue buildup results in generating a backpressure signal across the chain, forcing the upstream NF to be evicted (i.e., triggering a voluntary context switch) from the CPU as soon as the downstream NFs buffer levels exceed the high watermark threshold. The upstream NF will not execute till the downstream NF gets to consume and process its receive buffers. Thus, NFVnice is able to enforce judicious access to the CPU among the competing NFs of a service chain. We see in every case in fig. 14, NFVnice's throughput is superior to vanilla scheduler, emphasizing the point we make in this paper: NFVnice's design can support a number of different kernel schedulers, effectively support heterogeneous service chains and still provide superior performance (throughput, packet loss).

3) Workload Heterogeneity: We use 3 homogeneous NF's with the same compute cost, but vary the nature of the incoming packet flows so that the three NFs are traversed in a different order for each flow. We increase the number of flows (each with equal rate) from 1 to 6, as we go from Type 1 to Type 6. Thus, the bottleneck for each flow is different. Figure 15, shows that the native schedulers (first four bars) perform poorly, with degraded throughput as soon as we go to two or more flows, because of the different bottleneck NFs. However, NFVnice performs uniformly better in every case, and is almost independent of where the bottlenecks are for the multiple flows. Moreover, NFVnice provides a substantial improvement and robustness to varying loads and bottlenecks even across all the schedulers.

4) Performance Isolation: It is common to observe that when there are responsive (TCP) flows that share resources with non-responsive (UDP) flows, there can be a substantial degradation of TCP performance, as the congestion avoidance algorithms are triggered causing it to back-off. This impact is exacerbated in a software-based environment because resources are wasted by the non-responsive UDP flows that see a downstream bottleneck, resulting in packets being dropped at that downstream NF. These wasted resources result in less capacity being available for TCP. Because of the per-flow backpressure in NFVnice, we are able to substantially correct this undesirable situation and protect TCP's throughput even in the presence of non-responsive UDP.

In this experiment, we generate TCP and UDP flows with Iperf3. One TCP flow goes through only NF1 (Low cost) and NF2 (Medium cost) on a shared core. 10 UDP flows share NF1 and NF2 with the TCP flow, but also go through an additional NF3 (High cost, on a separate core) which is the bottleneck for the UDP flows - limiting their total rate to 280 Mbps.

We first start the 1 TCP flow. After 15 seconds, 10 UDP flows start, but stop at 40 seconds. As soon as the UDP flows interfere with the TCP flow, there is substantial packet loss without NFVnice, because NF1 and NF2 see contention from a large amount of UDP packets arriving into the system, getting processed and being thrown away at the queue for NF3. The throughput for the TCP flow craters from nearly 4 Gbps to just around 10-30 Mbps (note log scale), while the total UDP rate essentially keeps at the bottleneck NF3's capacity of 280 Mbps. With NFVnice, benefiting from per-flow backpressure, the TCP flow sees much less impact (dropping from 4 Gbps to about 3.3 Gbps), adjusting to utilize the remaining capacity at NF1 and NF2. This is primarily due to NFVnice's ability to perform selective early discard of the UDP packets

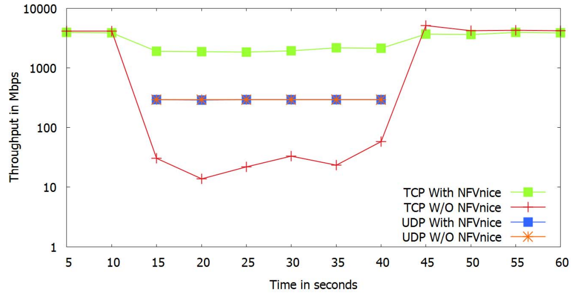


Fig. 16. Benefit for mix of responsive & non-responsive flows.

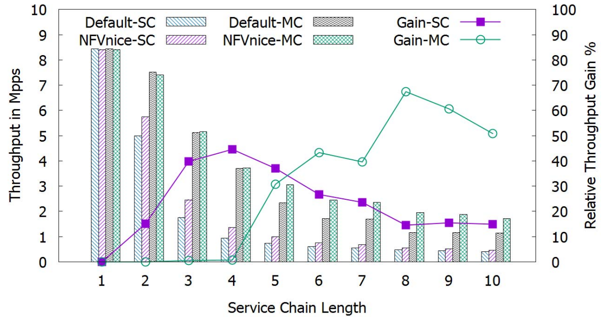


Fig. 17. Performance for different NF service chain lengths.

because of the backpressure. Otherwise we would have wasted CPU cycles at NF1 and NF2, depriving the TCP flow of the CPU. Note that the UDP flows' rate is maintained at the bottleneck rate of 280 Mbps as shown in Figure 16 (UDP lines are one on top of the other). Thus, NFVnice ensures that non-responsive flows (UDP) do not unnecessarily steal the CPU resources from other responsive (TCP) flows in an NFV environment.

5) Supporting Longer NF Chains: We choose three different NFs, as in IV-B, and increase the chain length from 1 NF up to a chain of 10 NFs by including one of the 3 NFs each time. We examine two cases: (i) all the NFs of the chain are on a single core (denoted by SC); and (ii) three cores are used, and as the chain length is increased, the additional NF is placed on the next core in round-robin fashion (denoted by MC). Results are shown in Figure 17. For the single core, NFVnice achieves higher throughput than the Default scheduler for longer chains, with the greater improvements achieved for chain lengths of 3-6. As the chains get longer (>7 NFs sharing the same core), the improvement with NFVnice is not as high. For the multiple core case, NFVnice improves throughput substantially, especially as more NFs are multiplexed on a care (e.g., chain lengths>4), compared to the Default scheduler.

V. Related Work

NF Management and Scheduling: In recent years, several NFV platforms have been developed to accelerate packet processing on commodity servers [24], [25], [27], [35], [36]. There is a growing interest in managing and scheduling network functions. Many works address the placement of NFs for performance and efficient resource usage [37]–[39]. For example, E2 [37] builds a scalable scheduling framework on top of BESS [36]. They abstract NF placement as a DAG, dynamically scale and migrate NFs while keeping flow affinity. NFV-RT [38] defines deadlines for requests, and places or migrates NFs to provide timing guarantees. These projects focus on NF management and scheduling across cluster scale. Our work focuses on a different scale: how to schedule NFs on shared cores to achieve fairness when flows have load pressure. Different from traditional packet scheduling for fairness on hardware platforms [6], [40]–[42], NFs are more complex, resulting in diversity of packet processing costs. Furthermore, different kinds of flow arrival rates exacerbate the difficulty of fair scheduling.

PSPAT [43] aims to provide a scalable scheduler framework by decoupling the packet scheduler algorithm from dispatching packets to the NIC for high performance. NFVnice considers the orthogonal problem of packet processing cost and flow arrival rate to fairly allocate CPU resources across the NFs. PIFO [44] presents the packet-in-first-out philosophy distinct from the typical first-in-first-out packet processing models. We use the insight from this work to decide whether to accept a packet and queue it for processing at the intended NF or discard at the time of packet arrival. Then, the enqueued packets are always processed in order. This approach of selective early discard yields two benefits: i) it avoids dropping partially processed (through the chain) packets, thus not wasting CPU cycles; ii) it avoid CPU stealing and allows CPU cycles to be judiciously allocated to other contending NFs.

User space scheduling and related frameworks: Works, such as [45], [46], consider cooperative user-space scheduling, providing very low cost context switching, that is orders of magnitude faster than regular Pthreads. However, the drawbacks with such a framework are two-fold: a) they invariably require the threads to cooperate, i.e., each thread must voluntarily yield to ensure that the other threads get a chance to share the CPU, without which progress of the threads cannot be guaranteed. This means that the programs that implement L-threads must include frequent rescheduling points for each L-thread [46] incurring additional complexity in developing the NFs. b) As there is no specific scheduling policy (it is just FIFO based), all the L-threads share the same priority, and are backed by the same kernel thread (typically pinned to a single core), and thus lack the ability to perform selective prioritization and the ability to provide QoS differentiation across cooperating threads. Nonetheless, NFVnice's backpressure mechanism can still be effectively employed for such cooperating threads to voluntarily yield the CPU as necessary. Another approach used by systems such as E2 [37] and VPP [35] is to host multiple NFs within a shared address space, allowing them to be executed as function calls in a run to completion manner by one thread. This incurs very low NUMA and cross-core packet chaining overheads, but being monolithic, it is inflexible and impedes the deployment of NFs from third party vendors.

Congestion Control and Backpressure: Congestion control and backpressure have been extensively studied in the past [47], [48]. DCTCP [47] leverages ECN to provide multi-bit feedback to the end hosts. MQ-ECN [48] enables ECN for tradeoff of both high throughput and low latency in multi-service multi-queue production DCNs (Data Center Network). All of these focus on congestion control in DCNs. However, in an NFV environment, flows are typically steered through a service chain. The later congestion is found, the more resources are wasted. If the end hosts do not enable ECN support or there are UDP flows, it is especially important for the NFV platform to gracefully handle high load scenarios in an efficient and fair way. Using multiple mechanisms

(ECN and backpressure), NFVnice ensures that overload at bottlenecks are quickly detected in order to avoid congestion and wasted work. *Fair Queueing:* Orthogonal work such as [49], [50], propose to ensure fair sharing of network resources among multiple tenants by spreading requests to multiple processing entities. That is, they distribute flows with different costs to different processing threads. In contrast, NFVnice seeks to achieve fairness by scheduling the NFs that process the packets of different flows appropriately, Thus, a fair share of the CPU is allocated to each competing NF.

VI. Conclusion

As the use of highly efficient user-space network I/O frameworks such as DPDK becomes more prevalent, there will be a growing need to mediate application-level performance requirements across the user-kernel boundary. OS-based schedulers lack the information needed to provide higher level goals for packet processing, such as rate proportional fairness that needs to account for both NF processing cost and arrival rate. By carefully tuning scheduler weights and applying backpressure to efficiently shed load early in the the NFV service chain, NFVnice provides substantial improvements in chain-wide throughput and latency, and dramatically reduces the wasted work across NF chains. This allows the NFV platform to gracefully handle overload scenarios while maintaining efficiency and fairness.

Our implementation of NFVnice demonstrates how an NFV framework can efficiently tune the OS scheduler and harmoniously integrate backpressure to meet its performance goals. Our results show that selective backpressure leads to more efficient allocation of resources for NF service chains within or across cores, and scheduler weights can be used to provide rate-cost proportional fairness, regardless of the kernel scheduler being used.

References

[1] J. Halpern and C. Pignataro, *Service Function Chaining (SFC) Architecture*, document RFC 7665, 2015. [Online]. Available: https://tools.ietf.org/html/rfc7665

[2] D. Merkel, "Docker: Lightweight Linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, Mar. 2014.

[3] (2014). *Data Plane Development Kit*. [Online]. Available: http://dpdk.org/

[4] J. C. Mogul and A. Borg, "The effect of context switches on cache performance," *ACM SIGPLAN Notices*, vol. 26, no. 4, pp. 75–84, 1991.

[5] A. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The multiple node case," *IEEE/ACM Trans. Netw.*, vol. 2, no. 2, pp. 137–150, Apr. 1994.

[6] D. Stiliadis and A. Varma, "Rate-proportional servers: A design methodology for fair queueing algorithms," *IEEE/ACM Trans. Netw.*, vol. 6, no. 2, pp. 164–174, Apr. 1998.

[7] *NFVnice Sourcecode*. Accessed: Oct. 30, 2017. [Online]. Available: https://github.com/nfvnice/NFVnice_Source.git

[8] W. Zhang *et al.*, "OpenNetVM: A platform for high performance network service chains," in *Proc. Workshop Hot Topics Middleboxes Netw. Function Virtualization (HotMIddlebox)*, New York, NY, USA, 2016, pp. 26–31, doi: 10.1145/2940147.2940155.

[9] I. Molnar. (2017). *Linux Kernel Documentation: CFS Scheduler Design*. [Online]. Available: https://www.kernel.org/doc/Documentation/scheduler/sched-design-CFS.txt

[10] (2013). *Network Functions Virtualization (NFV): Architectural Framework, ETSI-GS-NFV-002*. [Online]. Available: http://www.etsi.org/deliver/etsi_gs/nfv/001_099/002/01.01.01_60/gs_nfv002v010101p.pdf

[11] T. Kelly, S. Floyd, and S. Shenker, "Patterns of congestion collapse," Int. Comput. Sci. Inst., Univ. Cambridge, Cambridge, U.K., Tech. Rep., 2003. [Online]. Available: https://icir.org/floyd/papers/patterns.pdf

[12] J. C. Mogul and K. Ramakrishnan, "Eliminating receive livelock in an interrupt-driven kernel," *ACM Trans. Comput. Syst.*, vol. 15, no. 3, pp. 217–252, 1997.

[13] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Trans. Netw.*, vol. 1, no. 4, pp. 397–413, Aug. 1993.

[14] D. Lapsley and S. Low, "Random early marking: An optimisation approach to Internet congestion control," in *Proc. IEEE Int. Conf. Netw. (ICON)*, Sep. 1999, pp. 67–74.

[15] W.-C. Feng, D. Kandlur, D. Saha, and K. Shin, "BLUE: A new class of active queue management algorithms," Univ. Michigan, Ann Arbor, MI, USA, Tech. Rep. CSE-TR-387-99, 1999, vol. 1001, p. 48105.

[16] I. Stoica, S. Shenker, and H. Zhang, "Core-stateless fair queueing: A scalable architecture to approximate fair bandwidth allocations in high-speed networks," *IEEE/ACM Trans. Netw.*, vol. 11, no. 1, pp. 33–46, Feb. 2003, doi: 10.1109/TNET.2002.808414.

[17] K. Ramakrishnan, S. Floyd, and D. Black, *The Addition of Explicit Congestion Notification (ECN) to IP*, document RFC 3168, 2001. [Online]. Available: https://tools.ietf.org/html/rfc3168

[18] R. Bayer, "Symmetric binary B-trees: Data structure and maintenance algorithms," *Acta Inf.*, vol. 1, no. 4, pp. 290–306, 1972.

[19] L. J. Guibas and R. Sedgewick, "A dichromatic framework for balanced trees," in *Proc. IEEE 19th Annu. Symp. Found. Comput. Sci.*, 1978, pp. 8–21.

[20] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 19, no. 4, pp. 1–12, 1989.

[21] L. Zhang, "VirtualClock: A new traffic control algorithm for packet-switched networks," *ACM Trans. Comput. Syst.*, vol. 9, no. 2, pp. 101–124, 1991.

[22] P. Menage. (2017). *Linux Kernel Documentation: Cgroups*. [Online]. Available: https://www.kernel.org/doc/Documentation/cgroup-v1/cgroups.txt

[23] (2017). *Cgroups-Linux Control Groups*. [Online]. Available: http://man7.org/linux/man-pages/man7/cgroups.7.html

[24] L. Rizzo, "Netmap: A novel framework for fast packet I/O," in *Proc. USENIX Annu. Tech. Conf.* Berkeley, CA, USA: USENIX, 2012, pp. 101–112. [Online]. Available: https://www.usenix.org/conference/usenixfederatedconferencesweek/netmap-novel-framework-fast-packet-io

[25] J. Martins *et al.*, "ClickOS and the art of network function virtualization," in *Proc. 11th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*. Seattle, WA, USA: USENIX Association, Apr. 2014, pp. 459–473.

[26] S. G. Kulkarni *et al.*, "NFVnice: Dynamic backpressure and scheduling for NFV service chains," in *Proc. Conf. ACM Special Interest Group Data Commun.*, 2017, pp. 71–84.

[27] J. Hwang, K. K. Ramakrishnan, and T. Wood, "NetVM: High performance and flexible networking using virtualization on commodity platforms," *IEEE Trans. Netw. Service Manag.*, vol. 12, no. 1, pp. 34–47, Mar. 2015.

[28] (Jun. 2016). *Performance Measurements With RDTSC*. [Online]. Available: https://www.strchr.com/performance_measurements_with_rdtsc

[29] P. Emmerich, S. Gallenmũller, D. Raumer, F. Wohlfart, and G. Carle, "MoonGen: A scriptable high-speed packet generator," in *Proc. ACM Conf. Internet Meas. Conf.*, 2015, pp. 275–287.

[30] R. Olsson, "Pktgen the Linux packet generator," in *Proc. Linux Symp.*, Ottawa, ON, Canada, vol. 2, 2005, pp. 11–24.

[31] J. Dugan, S. Elliott, B. A. Mah, J. Poskanzer, and K. Prabhu. (2014). *iPerf—The Ultimate Speed Test Tool for TCP, UDP and SCTP*. [Online]. Available: https://iperf.fr/

[32] (2019). *The CAIDA Anonymized Internet Traces Dataset*. [Online]. Available: http://www.caida.org/data/passive

[33] *Benchmarking Methodology for Network Interconnect Devices*, document RFC 2544, Mar. 1999. [Online]. Available: https://rfc-editor.org/rfc/rfc2544.txt

[34] L. Rizzo, S. Garzarella, G. Lettieri, and V. Maffione, "A study of speed mismatches between communicating virtual machines," in *Proc. Symp. Archit. Netw. Commun. Syst. (ANCS)*, New York, NY, USA, 2016, pp. 61–67, doi: 10.1145/2881025.2881037.

[35] (2016). *VPP*. [Online]. Available: https://fd.io/

[36] S. Han, K. Jang, A. Panda, S. Palkar, D. Han, and S. Ratnasamy, "SoftNIC: A software NIC to augment hardware," Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2015-155, May 2015. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-155.html

[37] S. Palkar *et al.*, "E2: A framework for NFV applications," in *Proc. 25th Symp. Oper. Syst. Princ. (SOSP)*, New York, NY, USA, 2015, pp. 121–136, doi: 10.1145/2815400.2815423.

[38] Y. Li, L. T. X. Phan, and B. T. Loo, "Network functions virtualization with soft real-time guarantees," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2016, pp. 1–9.

[39] S. Rajagopalan, D. Williams, H. Jamjoom, and A. Warfield, "Split/Merge: System support for elastic execution in virtual middleboxes," in *Proc. 10th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, 2013, pp. 227–240.

[40] M. Shreedhar and G. Varghese, "Efficient fair queuing using deficit round-robin," *IEEE/ACM Trans. Netw.*, vol. 4, no. 3, pp. 375–385, Jun. 1996.

[41] P. Goyal, H. M. Vin, and H. Chen, "Start-time fair queueing: A scheduling algorithm for integrated services packet switching networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 26, no. 4, pp. 157–168, 1996.

[42] D. Stiliadis and A. Varma, "Efficient fair queueing algorithms for packet-switched networks," *IEEE/ACM Trans. Netw.*, vol. 6, no. 2, pp. 175–185, Apr. 1998.

[43] L. Rizzo, P. Valente, G. Lettieri, and V. Maffione, "PSPAT: Software packet scheduling at hardware speed," *Comput. Commun.*, vol. 120, pp. 32–45, May 2018.

[44] A. Sivaraman *et al.*, "Programmable packet scheduling at line rate," in *Proc. Conf. ACM SIGCOMM Conf.*, 2016, pp. 44–57.

[45] (2017). *Fibers*. [Online]. Available: https://msdn.microsoft.com/library/ms682661.aspx

[46] (2014). *DPDK L-Thread Subsystem*. [Online]. Available: http://dpdk.org/doc/guides/sample_app_ug/performance_thread.html

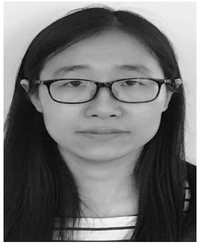
[47] M. Alizadeh *et al.*, "Data center TCP (DCTCP)," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 63–74, 2010.

[48] W. Bai, L. Chen, K. Chen, and H. Wu, "Enabling ECN in multi-service multi-queue data centers," in *Proc. 13th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*. Santa Clara, CA, USA: USENIX Association, 2016, pp. 537–549.

[49] A. Ghodsi, V. Sekar, M. Zaharia, and I. Stoica, "Multi-resource fair queueing for packet processing," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 1–12, Aug. 2012, doi: 10.1145/2377677.2377679.

[50] J. Mace, P. Bodik, M. Musuvathi, R. Fonseca, and K. Varadarajan, "2DFQ: Two-dimensional fair queuing for multi-tenant cloud services," in *Proc. ACM SIGCOMM Conf.*, New York, NY, USA, 2016, pp. 144–159, doi: 10.1145/2934872.2934878.

Sameer G. Kulkarni received the Ph.D. degree from the University of Göttingen, Germany. He is currently a Post-Doctoral Researcher with the Department of Computer Science and Engineering, University of California at Riverside, Riverside, CA, USA. His current research interests include parallel and distributed computing, software defined networks, network function virtualization, and cloud computing.

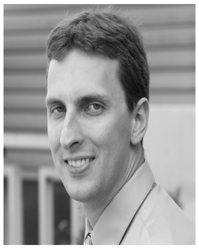


Wei Zhang received the B.S. degree from the Hebei University of Economics and Business in 2006, the M.S. degree from Yanshan University in 2008, the Ph.D. degree from Beihang University in 2014, and the Ph.D. degree from The George Washington University in 2018. She is currently a Research and Development Software Engineer with Microsoft Azure. Her research interests include cloud computing, systems, and resource disaggregation.

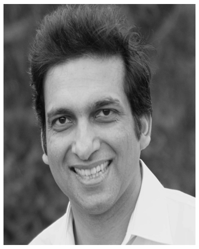
Jinho Hwang received the Ph.D. degree from The George Washington University, Washington, DC, USA, in 2013. He was a Visiting Scholar with The George Washington University from 2005 to 2006 and the POSCO ICT Research and Development Center, South Korea, from 2007 to 2009. He interned at the IBM T. J. Watson Research Center, NY, USA, and AT&T Labs-Research in the summer of 2012 and 2013, respectively. He has been a Research Staff Member with the IBM T. J. Watson Research Center since 2013. He has published more than 50 articles, filed 50 patents, and has won four best paper awards. His current research focuses on improving artificial intelligence support for cloud systems and networks. He has received six outstanding technical achievement awards and has been appointed to a Master Inventor at IBM.

Shriram Rajagopalan received the Ph.D. degree from The University of British Columbia, Vancouver, BC, Canada. He is currently a Principal Engineer with Tetrate. His current work focuses on layer-7 networking fabric across multiple cloud environments for cloud native applications. His research interests focus on high-availability problems in software defined networking and distributed systems.

K. K. Ramakrishnan received the M.Tech. degree from the Indian Institute of Science in 1978, the M.S. degree in 1981, and the Ph.D. degree in computer science from the University of Maryland, College Park, MD, USA, in 1983. He is currently a Professor of computer science and engineering with the University of California at Riverside, Riverside, CA, USA. Previously, he was the Distinguished Member of the Technical Staff at AT&T Labs-Research. Prior to 1994, he was a Technical Director and a Consulting Engineer in networking with Digital Equipment Corporation. From 2000 to 2002, he was with TeraOptic Networks, Inc., as a Founder and the Vice President. He is a Fellow of the ACM and AT&T, recognized for his fundamental contributions on communication networks, including his work on congestion control, traffic management and VPN services. He has published over 275 articles and has 180 patents issued in his name.



Timothy Wood received the bachelor's degree in electrical and computer engineering from Rutgers University in 2005, and the Ph.D. degree in computer science from the University of Massachusetts Amherst in 2011. He is currently an Associate Professor with the Department of Computer Science, The George Washington University. His research studies how new virtualization technologies can provide application agnostic tools that improve performance, efficiency, and reliability in cloud computing data centers and software-based networks. His Ph.D. thesis received the UMass CS Outstanding Dissertation Award, his students have voted him CS Professor of the Year, and he has won three best paper awards, the Google Faculty Research Award, and the NSF Career Award.



Mayutan Arumaithurai received the industrial Ph.D. degree from the University of Göttingen, Germany, in 2010, while working for Nokia Siemens Networks. He is currently a Senior Researcher with the Computer Networks Group, University of Göttingen. Prior to that, he worked as a Research Scientist with the Network Laboratories, NEC Europe Ltd., Heidelberg, Germany, for two years. His current research interests include information centric networking, software defined networks, network function virtualization, and cloud computing. He has published in top conferences in his field (ACM SIGCOMM, ACM CoNext, the IEEE Infocom), coauthored IETF/IRTG standards, and has led multiple million-euro EU-funded projects.

Xiaoming Fu (Senior Member, IEEE) received the Ph.D. degree in computer science from Tsinghua University, China, in 2000. Since 2007, he has been a Professor and the Head of the Computer Networks Group, Georg–August–Universität Göttingen, Germany. He has also held visiting positions at ETSI, University of Cambridge, Columbia University, Tsinghua University, and UCLA. He is also a Distinguished Lecturer of the IEEE, a member of the ACM and Academia Europaea, and a Fellow of IET.