Virtuoso: High Resource Utilization and μs-scale Performance Isolation in a Shared Virtual Machine TCP Network Stack

Matheus Stolet
Max Planck Institute for Software Systems
Saarbrücken, Germany
mstolet@mpi-sws.org

Liam Arzola
Max Planck Institute for Software Systems
Saarbrücken, Germany
larzola@mpi-sws.org

Simon Peter
University of Washington
Seattle, USA
simpeter@cs.washington.edu

Antoine Kaufmann
Max Planck Institute for Software Systems
Saarbrücken, Germany
antoinek@mpi-sws.org

Abstract

Virtualization improves resource efficiency and ensures security and performance isolation for cloud applications. Today, operators use a layered architecture with separate network stack instances in each VM and container connected to a virtual switch. Decoupling through layering reduces complexity, but induces performance and resource overheads at odds with increasing demands for network bandwidth, connection scalability, and low latency.

We present Virtuoso, a new software network stack for VMs and containers. Virtuoso re-organizes the network stack to maximize CPU utilization, enforce isolation, and minimize processing overheads. We maximize utilization by running one elastically shared network stack instance on dedicated cores; we enforce isolation by performing central and fine-grained per-packet resource accounting and scheduling; we reduce overheads by building a single-layer data path with a one-shot fast-path incorporating all processing from the TCP transport layer through network virtualization and virtual switching. Virtuoso improves resource efficiency by up to 82%, latencies by up to 58% compared to other virtualized network stacks without sacrificing isolation, and keeps processing overhead within 6.7% of unvirtualized stacks.

1 Introduction

The cloud leverages virtualization to improve resource utilization while ensuring isolation for security and performance. The hypervisor and operating system allocate and manage the shared physical resources such as processor cores, memory, and network links. For network communication, each VM and container runs a separate network stack instance. VMs send packets from their OS stack through virtual NICs to the hypervisor, while containers run in isolated network name spaces also generating raw guest network packets, then forwarded through a virtual interface to a central kernel or userspace virtual bridge or switch. The operator configures the virtual switch to implement network virtualization, including tunneling, bandwidth limits, and security checks, and pass packets to and from the physical network.

In this layered architecture, packets pass through a series of different separate loosely-coupled components, such as the guest transport layer, network layer, virtual NIC, virtual switch. This layered architecture works well but incurs significant performance and resource overheads. On one hand, decoupling through layering simplifies development, configuration, and management. Decomposition into separate layers also enables partial performance isolation, as guest stacks may be isolated by dedicating CPU cores to each guest. On the other hand, demands for increasing network bandwidths and for low latency communication are expensive to meet in this architecture. 100 Gbps links are commonplace and 400 Gbps are already available. At the same time, modern cloud applications demand μs-scale network latencies [57]. Coupled with the slow down of Moore's Law, any wasted CPU cycles due to network processing—incurred either due to underutilized dedicated CPU resources or inefficiencies in network stack processing—are particularly problematic.

In this paper, we argue that the existing layered virtual network stack architecture unnecessarily sacrifices resource efficiency and performance for isolation. Rising network speeds increased the fraction of server CPU cycles consumed by TCP packet processing: communication-intensive applications such as key-value stores may spend up to 48% of per-CPU cycles in the TCP stack and NIC driver [27, 48]. The typical static CPU allocation for guests (VMs or containers) requires users to provision cores for peak traffic. However, the more common off-peak periods incur poor CPU utilization because of idle capacity allocated for network processing. Recent data shows 84% of VMs have a peak utilization of less than 20% [17]. Techniques, such as CPU oversubscription, improve utilization but are uncommon in clouds because of the difficulties in maintaining tenant SLOs and performance isolation. Additionally, the layered architecture for network virtualization and isolation adds significant

overhead to the datapath. The hypervisor individually mediates every packet sent or received by the guest, increasing CPU overhead and communication latency.

We argue that these overheads are not inherent to network virtualization, but are an artifact of the existing architecture. To that end, we propose a fundamental re-organization of the full virtual network stack architecture. We present Virtuoso, a new, shared software network stack for virtual machines and containers that maximizes CPU utilization, while minimizing processing overheads and enforcing isolation. Virtuoso is drop-in compatible with sockets and TCP. With its stack re-design, Virtuoso improves resource utilization by sharing the stack, while providing fine-grained per-packet CPU performance isolation. Virtuoso improves resource efficiency by up to 82% and increases throughput by up to 91% over optimized layered stacks, while still ensuring μs-scale tail latency performance isolation. Virtuoso also achieves high absolute throughput, incurring only a 14% throughput penalty compared to state-of-the-art bare-metal network stacks. Additionally, Virtuoso helps improve the performance of small VMs (VMs that are not even allocated a full core) by reducing VM exits, thus providing opportunities for cloud providers to improve networking performance on oversubscribed machines.

The first Virtuoso key idea is to use only *one network stack instance* in the hypervisor, *shared* by all guests. Sharing improves CPU utilization for bursty workloads, as the shared stack elastically allocates CPU resources just-in-time, rather than statically provisioning CPU bandwidth for each guest's peak. To provide microsecond-scale performance isolation in a shared network stack, we leverage *fine-grained per-packet resource scheduling*. Virtuoso accounts CPU cycles and network bandwidth used for each processed packet to the respective guest resource budget, scheduling each guest on a per-packet basis. These fine-grained mechanisms incur minimal performance overhead but enable μs-scale performance isolation. Finally, a *coalesced data path* combines all virtual network processing from transport down to virtual switching. The coalesced data path in Virtuoso collapses all layers in the stack, minimizing processing overheads by avoiding intermediate queuing, while implementing the same functionality as conventional layered stacks. with considerably fewer processor cycles. We further split the data path into a *fast-* and a *slow-path*. Virtuoso processes common packets for established connections on the fast-path in *one-shot*, reducing necessary state and simplifying performance isolation through short, predictable code paths. Uncommon cases are handled on the slow-path at a small performance penalty.

Our contributions are the following:

- The design of a new shared TCP network stack for virtual environments that improves resource utilization and leverages fine-grained scheduling for isolation.
- One-shot network virtualization fast-path incurring minimal virtualization overhead.
- Virtuoso prototype implementation for Linux and QEMU.
- Performance analysis of Virtuoso prototype to quantify the resource utilization improvement and overhead reduction, and confirming performance isolation and low tail latency.

We will release Virtuoso as open-source software. This work does not raise ethical issues.

2 Background

We now discuss the unique challenges for and approaches to network communication in virtualized environments.

2.1 Network Virtualization Concepts

Virtualization aims to facilitate management and consolidation of host and network resources. Multiple guest VMs or containers with separate network addresses share a single physical host, network controller, and link. Similarly, multiple separate virtual networks share the same physical network. Tenants expect to flexibly configure and use their virtual infrastructure and infrastructure operators must multiplex the physical resources providing isolation to produce the illusion of completely separate infrastructure to mutually non-trusting tenants.

On the hosts this requires virtual switching, moving packets between the various guests and the shared physical network in a controlled and safe way, and enforcing all necessary processing for security and isolation. In the network, virtualization requires tunneling protocols[15, 19, 50] to encapsulate packets, enabling use of separate protocols and routing on the physical network. Virtual switching also assigns appropriate physical network addresses based on virtual network addresses.

The complete network communication infrastructure aims to meet the following *operator goals*: guests must not affect the performance of other guests, idle resources should be minimized, and packets must not be tampered. In addition, there are two *tenant goals*: the networking stack must support μs scale latencies and high throughput.

2.2 Status Quo: Layered Silos

Traditionally, network virtualization is implemented as a deeply layered architecture (Figure 1). Packet processing is divided between independent network stacks in each guest (managed by the tenant), multiplexed by the virtual switch running as the lowest layer on the host (managed by the operator). Guests send and receive raw network packets through their conventional OS network stack via the assigned virtual NIC (vNIC) just as in a native deployment. Containers run in separate network namespaces typically communicating with the outside through `veth` pairs [29], while VMs run separate OS instances and use virtual NICs such as `virtio-net` [52] implemented by the hypervisor.

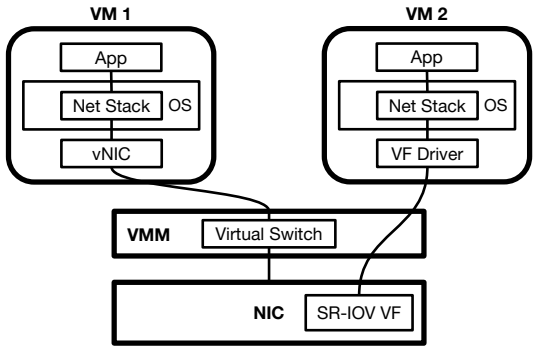

Figure 1. Layered and independent virtualized stacks.

The virtual switch (vSwitch) takes packets sent on the guests' vNICs, routes and encapsulates them for the physical network, and then sends them out through the host's physical NIC. Receiving packets works symmetrically: packets arrive on the physical NIC, the vSwitch inspects and decapsulates them to determine the virtual network, and then looks up and passes them to the corresponding vNIC. Alternatively, some deployments leverage hardware SR-IOV [21] capabilities to bypass the software switch layer of the virtualization stack and increase performance, by assigning resources on the NIC to the guest through *virtual functions* (VFs).

Advantages. Separately developed, maintained, and operated virtual switches simplify implementation and enable flexible deployment. Typically data centers run guest VMs and containers on dedicated cores. Thus, independent per-guest stacks effectively silo applications, minimizing inter-VM performance interference. Much of the protocol processing happens in the guest and is thus automatically accounted for and isolated. Early demultiplexing also prevents priority inversion for more complex scheduling policies, as packet priorities are only known afterwards [38].

Disadvantages. On the other hand, independent network stacks often over-provision resources. For the typically bursty workloads, tenants have to provision VMs with enough resources (especially cores) for peak bandwidth. While VMs can share resources such as CPU cores, this is not compatible with μs-scale latency requirements because of long and expensive VM context switches. Cloud VMs, in particular, typically exclusively provision processor cores, with only exceptions for the smallest and cheapest instance types. VMs typically only rarely operate at peak traffic, frequently leaving resources underutilized.

Layered network stacks also incur overheads increasing latency and wasting CPU cycles [27, 43]. Each layer adds indirection, often through queues or other data transfer mechanisms. For example, the TCP layer generates segments that queue up lower in the stack because of vSwitch-enforced bandwidth limits. Further, independent layers often redundantly retrieve similar packet information [38] from different data structures. Layered stacks also include significant processing before multiplexing points that are not performance isolated, e.g. in the vSwitch. This is a source of performance cross-talk between guests and tail latency [51].

SR-IOV bypasses virtualization layers and reduces overhead, but it also comes with drawbacks. Namely, SR-IOV is not compatible with every network adapter or leads to underutilized resources because a portion of the NIC's resources are assigned to specific VMs. For example, the network virtualization stack for Google Cloud skirts SR-IOV so that guests do not have to cope with different physical NIC resources on the target host during live migration[4]. SR-IOV also only allows NIC resources to be multiplexed, but not the cores used by a VM's network stack for packet processing, thus forgoing sharing opportunities. For example, a shared network stack exposed with application interfaces as the abstraction boundary instead of a virtual NIC can improve utilization and allows rapid and flexible deployment, thus accelerating innovation in the cloud[58].

Summary. Layered stacks face two key challenges. First, there is a *trade-off between isolation and resource utilization.* Multiplexing resources early with independent stacks facilitates isolation but fails to capitalize on finer-grained resource-sharing opportunities because resources are siloed from the start and cannot be pooled at lower levels. Second, *layering provides modularity but leads to overheads* in packet processing. These overheads are exacerbated by rising network speeds, microsecond tail-latency requirements [5], and the large scale of datacenter applications.

2.3 Prior Work

Prior work has investigated these challenges, but fails to satisfy all goals. In particular, providing high resource utilization and elasticity of CPU resources with minimal interference for μs-scale workloads remains a challenge.

Reducing layering overheads. A range of work seeks to avoid layering overheads and reduce indirection in packet processing in specific layers. These solutions span kernel bypass [3, 8, 25, 43, 57], kernel offload via eBPF [16, 46, 59], zero-copy methods [3, 28, 32, 43], new NIC interfaces [9, 44, 47, 54], and one-shot unlayered fast-paths [27, 48]. These approaches do not completely solve performance overheads or isolation across the virtualized stack. Most of these solutions focus on streamlining the path between an application and the NIC by reducing operating system overheads, but neglect the additional layers necessary for network virtualization, or suffer from low utilization and isolation issues due to the lack of elasticity and performance isolation mechanisms.

Hardware offload. Offloading different parts of network virtualization processing can significantly reduce overheads.

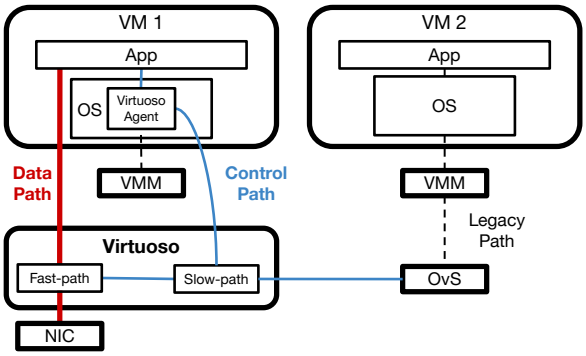

Figure 2. Fast-path manages TX and RX; slow-path handles control operations. Legacy applications follow a layered legacy path.

Recent data center NICs support offload for VXLAN, GENEVE, and NVGRE [40] en/de-capsulation. Complete offload of virtual switching fast-paths [11, 33, 41], and hypervisor bypass [2, 36], can reduce overheads further and enable direct HW NIC access for guests through SR-IOV [2, 11]. However, these approaches rely on fixed hardware limited to specific protocols and features [11], or themselves leverage multi-core SmartNICs [2, 33, 41] and FPGAs [36], resulting in another layered architecture with the challenges discussed above. Hardware offload also gives rise to other performance isolation challenges with shared hardware resources, such as the NIC, PCIe interconnect, and IOMMU [1]. Finally, software solutions are relevant because of their comparative flexibility and simple deployability as evidenced by software networking stacks deployed in large production scenarios [4, 30, 35].

3 Virtuoso Approach

Virtuoso (Figure 2) eliminates the tradeoff between resource efficiency and isolation by sharing a network stack among guests and implementing isolation in a single layer. The shared stack allows Virtuoso to elastically pool resources and increase utilization, while fine-grained resource accounting and scheduling ensure performance isolation. Externalizing network processing gives Virtuoso visibility into VM usage, so that it can make informed scheduling decisions about multiple VMs. Virtuoso uses a multi-threaded data fast-path with dedicated cores for common case send and receive operations, and a separate slow-path for data path exceptions and control operations. The fast-path combines all network virtualization and packet processing layers up to and including the TCP transport layer, minimizing the path between the guest application and the host NIC. The small units of work in the fast-path enable Virtuoso to perform fine-grained per packet scheduling efficiently. The fast-path implements en-/de-capsulation and de-multiplexing, and combines all common-case processing. Only the sockets interface remains in the guest, but is tightly integrated with guest applications in guest userspace through a dynamic link library.

3.1 Design Principles

Shared network stack for elastic resource utilization. Instead of partitioning network processing to multiple guest silos and the hypervisor, Virtuoso places one shared network stack instance in the hypervisor. Externalizing network processing allows guests to serve the same workload with fewer cores; we instead re-allocate some of these cores for the shared stack. This resource consolidation particularly improves utilization for bursty workloads by being elastic; the larger shared pool of cores can absorb bursts better than multiple static per-guest pools [56]. Furthermore, a shared network stack also improves deployment flexibility and allows providers to offer more meaningful SLAs to tenants by taking control over the stack [58].

Fine-grained per-packet scheduling for isolation. Instead of coarse-grained resource management via dedicated cores to guests, we employ central and fine-grained resource accounting and scheduling for individual packets to ensure isolation in the shared network stack. Virtuoso precisely accounts processor cycles and network bandwidth spent by each packet to the respective guest resource budget. Virtuoso leverages global visibility across all guests combined with accurate resource accounting to implement fine-grained per-packet scheduling that enforces tight isolation policies. Scheduling is implemented centrally at a single layer in the system, minimizing crosstalk [31, 51]. This nimble mechanism incurs minimal performance overhead but enables performance isolation even for microsecond-scale latencies.

Single-layer data path. Instead of layered processing, Virtuoso leverages a single-layer data path, coalescing all network processing from the TCP transport layer down to network virtualization and virtual switching, for receive and transmit. Guest applications interact directly with the data path through efficient shared memory queues, by linking a dynamic link library in guest userspace that provides the TCP sockets API. This allows Virtuoso to implement the same functionality as conventional layered stacks considerably faster and with fewer processor cycles by communicating directly with the single-layer data path.

One-shot fast-path. We further streamline the Virtuoso data path via a one-shot fast-path. For each TCP connection, *one-shot processing* pre-computes rarely changing processing state, such as guest and physical IP routing and tunnel state, storing it in the fast-path, reducing per-packet processing overhead for common packets of established connections. Handling a limited number of common cases in the fast-path also simplifies performance isolation through short and predictable code paths. For example, when sending a TCP

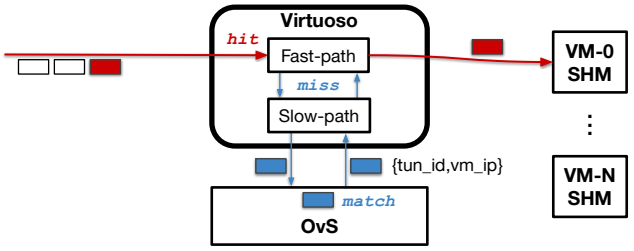

Figure 3. The fast path routes packets to VMs with cached state; the slow path fetches tunnel headers on cache misses.

segment from the guest on a virtual TCP connection, the fast-path can directly create a physical packet with all relevant virtualization headers and send it via the host NIC in a short operation. Uncommon cases are handled on a separate slow-path at a small performance penalty.

4 Detailed Virtuoso Design

In the Virtuoso network stack, a multi-core fast-path polls guests for new packets and parses and generates headers in a single layer for low-overhead packet transmission and reception. The fast-path accesses each guest library's shared memory region and aggregates packets from multiple guests into a batch to increase utilization. A separate slow-path core handles control operations and exceptions. Dividing tasks between a fast-path and a slow-path allows us to reduce overheads by streamlining the fast-path. For initialization of these shared memory channels between applications and Virtuoso during startup, we leverage a modified hypervisor for guests and the host OS for containers. We describe these implementations later (§5.2, §5.3). After shared memory regions are set up, there are no differences for the application-Virtuoso data path between virtual machines and containers. Hence, we refer to them collectively as *guests*.

In this section, we give detailed descriptions of the different components of Virtuoso. We start by describing how we integrate efficient network virtualization in a multi-core fast-path (§4.1). We then describe how we track each guest's resource usage on individual fast-path cores through a per-guest budget (§4.2), followed by a discussion of how we use this information to coordinate guest resource budget allocation across fast-path cores centrally through our slow-path (§4.3). We then describe how we build fine-grained scheduling based on the per-guest resource budgets (§4.4). Finally, we describe how we protect the Virtuoso network stack on the host when sharing it among guests (§4.5).

4.1 One-shot, Single-layer Transport

Single-layer transport. Virtuoso combines all processing from the TCP transport layer all the way down to network virtualization into a streamlined one-shot fast-path (Figure 3), for send and receive. Separating out common case processing in a minimal fast-path enables performance optimization, while a separate slow-path ensures that less frequent cases are handled. Regular data transfer packets for established TCP connections exclusively use this optimized data path, while packets for unknown connections or other protocols pass through the slow-path. The slow-path sets up one-shot fast-path state for for new connections so future packets remain on the fast-path (Figure 3).

On receive, the fast-path parses the packet according to the expected format, configured by default to TCP over IPv4 on the guest side, encapsulated in GRE [10] over IPv4, and Ethernet on the physical network. The fast-path then leverages the corresponding connection identifiers, TCP ports, guest IPs, and tunnel ID to look up the consolidated flow state. After validating the packet against the state, the fast-path directly stores the TCP payload in the guest flow buffer and enqueues a notification in the corresponding guest receive queue. Finally, if necessary, we reformat the packet by swapping addresses and tweaking the TCP header into a response TCP acknowledgement.

Similarly for transmit, once Virtuoso schedules a flow to transmit a packet, the flow state directly provides all necessary state to directly assemble the complete packet with all headers for immediate transmission via the NIC. Headers are divided between inner and outer headers. The inner TCP and IP headers includes the source and destination IP address and port on the guest (virtual) network. The outer headers include the GRE encapsulation with the key field to identify the network [7] wrapped by the outer UDP and IP header and corresponding physical network source and destination IP addresses and ports, finally wrapped by Ethernet and the necessary peer MAC address. Virtuoso stores these key fields in the consolidated flow state.

One-shot fast-path processing. We implement this processing as straight-line code with minimal control flow (other than exceptions for rare cases) and no packet modifications until acknowledgements [27]. Virtuoso processes packet in one shot without intermediate queuing or access to complex data structures other than the consolidated flow state. We skip some steps for conventional network virtualization such as decapsulation completely. Other steps we combine with already necessary related steps previously in other layers, such as combining the virtual switching table lookup with the TCP flow state lookup.

Our state consolidation optimizations rely on most of this state, such as guest routing, tunneling, host addressing and routing, remaining typically unchanged over the life of a connection. Thus it can be pre-computed and stored when a new connection is established. This is related to other fast-path caches for virtual switching state in systems such as Open vSwitch [42]. Except Virtuoso explicitly and eagerly manages this state, adding it, updating it, and removing it as necessary instead of relying on misses and invalidations.

This also implies that changes to this state are more expensive in Virtuoso than in other systems, as many changes to individual connection state instances on the fast-path may be required for an individual change to the underlying state.

Slow-path for remaining processing. The Virtuoso slow-path handles all packets not handled by the fast-path. This includes packets that are not TCP data packets, TCP control packets to open and close connections as well as non-TCP packets. We leverage the existing Open vSwitch [42] in the slow-path for network virtualization. Virtuoso sends truncated headers to OvS, which asynchronously sends back the packet with the necessary virtualization state. If the packet belongs to a new TCP connection, Virtuoso combines the received virtualization state with the necessary TCP state. Non-TCP packets are forwarded to guests through legacy interfaces (vNICs or veth) for processing in the legacy stack.

4.2 CPU Resource Accounting

Core-local resource accounting. The first step towards isolation is to accurately account for resource use. Each fast-path core tracks resources available to and used by individual guests through a local budget table, storing each guest's resource budget on that core. The slow-path periodically updates the fast-path value (§4.3) and replenishes credits by performing an atomic add to the guest's entry in the budget table. This reduces the need for synchronization on the fast-path and improves overall throughput by 4.3% at scale.

Batch processing in three main tasks. The Virtuoso fast-path performs three main CPU-intensive tasks for guests: receiving packets (RX), polling guest transmit queues (POLL), and packet transmission (TX). RX dequeues incoming packets from the NIC, parses the packets, and implements the necessary TCP processing before forwarding the payload to the guest. POLL checks outgoing queues from guest applications to the fast-path for new transmission requests. TX assembles complete network-virtualized TCP packets and enqueues them in the NIC. For efficiency, these tasks execute in batches, generally from multiple guests. The batch size is a compile-time parameter and primarily depends on the system's cache hierarchy; we chose 16 empirically as the value that yielded the highest throughput for our setup.

Lightweight accounting with TSCs. Virtuoso measures CPU consumption by taking CPU time stamp counter (TSC) readings at the start and end of processing for each batch. Reading the TSC is lightweight and precise. Virtuoso breaks down the TSC total to separate guests, based on each guest's number of packets. As per-packet processing costs are generally similar, this represents a reasonable trade-off between overhead for accounting and accuracy, as we will show later (§6.2, §6.3). Virtuoso then subtracts the cycles consumed from the respective guest's resource budget.

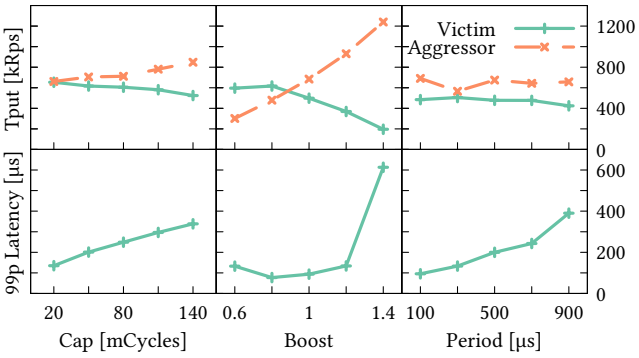

Figure 4. Guest VM performance with variable boost, budget caps, and update periods, while an aggressor induces interference.

4.3 Central Resource Allocation

A separate slow-path core periodically replenishes the per-core budgets on the fast-path, leveraging its global view. Separation into a parallel de-centralized fast-path and a central slow-path enables scalable and efficient coordination of the frequently accessed per-core budgets. The slow-path replenishes the total budget in periodic 100 µs intervals and distributes the new budget to each guest. The distribution among guests is controlled by a guest weight w_g, configured by the operator. By default each guest has the same weight.

We compute update credits e_g for guest g by recording the timestamp t' for the current update and the timestamp t for the previous update. The allocator scales the elapsed time $t' - t$ by a constant boost B. B compensates for any fast-path CPU cycles not explicitly accounted to any guest by Virtuoso to avoid over-committing processor cycles. We found the fraction of accounted cycles to be 94% (and set $B = 0.94$), with minimal processing not related to specific guests. This includes functions, such as scaling fast-path cores up and down and checking if a core can block. We multiply the product of the boost and elapsed cycles by the guests's w_g, divided by the sum of the weights of all n guests.

$$e_g = \frac{B(t' - t)w_g}{\sum_{i=1}^{n} w_k} \tag{1}$$

Work-conserving allocator distributes more credits to active cores. Our schedulers are work-conserving, so credits are proportionately distributed based on activity. This prevents a VM from being throttled when it runs out of budget in a core, but has spare credits in other cores. For example, if a guest only uses one fast-path core, our resource allocator redistributes credits assigned to other cores in this round to the active core. We redistribute credits by calculating the sum of the used budget ($b_{max} - b_{gc}$) for all m cores and calculate the ratio of a core's unused budget to the total sum. We multiply e_g by this ratio. This compromise allows us to minimize synchronization overhead by using per-core budgets,

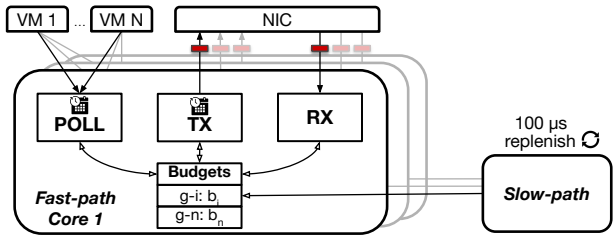

Figure 5. Fast-path cores utilize a guest's local budget for processing tasks; all tasks measure resource consumption, with the slow-path periodically replenishing budgets

while leveraging a global view of resource usage to update local cores within 100 μ periods.

$$u_{gc} = \frac{e_g(b_{max} - b_{gc})}{\sum_{j=1}^{m}(b_{max} - b_{gj})} \tag{2}$$

Preventing guest from accumulating budget. The operator also configures a budget cap C for all guests. Capping the budget prevents guests from accumulating arbitrary budgets during long periods of low utilization and starving other guests in bursty periods of activity. C restricts the number of CPU cycles Virtuoso can spend on behalf of a guest per fast-path core between update periods. The slow-path calculates the updated per-core budget b'_{gc} for guest g on core c

$$b'_{gc} = \min\{C, b_{gc} + u_g\} \tag{3}$$

Figure 4 shows how different boost, budget cap, and update period parameters protect the performance of a guest VM, while an aggressor VM introduces performance interference through background load. The aggressor creates a load imbalance by using 9 cores to open a total of 900 connections and the victim opens one connection in one core. If the boost parameter is excessively high, the performance of the guest is affected because the aggressor is not sufficiently throttled: the aggressor budget is fully replenished to the capped amount every round. Similarly, if it is too low, the victim suffers a small tail latency increase due to throttling. In the experiment varying the budget cap, the aggressor bursts every 250 ms for 250 ms. For large budget caps, the aggressor affects the performance of the victim if it is allowed to accumulate credits by remaining momentarily idle. Finally, the experiment varying the update period granularity shows that shorter update periods are helpful in maintaining low tail-latencies in guests and are essential for microsecond scale workloads.

4.4 Fine-grained Scheduling

Virtuoso performs fine-grained scheduling to enforce performance isolation based on per-guest CPU cycle budgets (Figure 5). The scheduler performs hierarchical scheduling, first it chooses which guests to perform work for, and for sending packets determines which of the guest's flows get to transmit next. On the fast-path, before starting a task on behalf of a guest, the core consults the guest's budget, and if it is zero or negative moves on to do work for a different guest.

Bounded fast-path simplifies scheduling. The observation at the center of our approach is that all individually scheduled tasks are strictly bounded, on the order of 200–500 cycles depending on packet sizes. This provides us with two key advantages. First, *preemption is not necessary*, as individual packet processing tasks complete very quickly. Second, *fine-grained batch scheduling and accurate accounting enable low tail latency and isolation*, even without knowing concrete task lengths. Tasks are all similarly sized, and, after each task completes, the next scheduling decision can compensate based on the updated budget. Even if a task overruns the budget, it will only be by a small amount of cycles and Virtuoso still precisely accounts for this with negative budgets, akin to deficit round-robin scheduling [49].

These two observations enable fine-grained scheduling for low tail latency without preemption overhead in Virtuoso. Conventional layered network stacks switch between guests and require context switches, which disrupt CPU pipelines, cause cache overheads, and demand expensive state saving and restoration [20]. Virtuoso stores all necessary processing state in the corresponding flow-state data structure, so switching to processing a packet for a different guest incurs minimal overhead, as it does not require any context switch.

Hierarchical scheduler controls guests and flows. Virtuoso uses a two-level hierarchical scheduler. The first level decides which guest should be serviced next and the second level decides what flow (TX) or transmit queue (POLL) from the selected guest should be scheduled, using different policies. This allows us to control resource allocation between guests and between guest's flows and transmit queues. For RX, Virtuoso only performs resource accounting, as the specific guest is not known before initial processing of the packets, preventing scheduling. The following paragraphs dive into detail on scheduling Virtuoso processing tasks.

Batch-scheduling POLL. Virtuoso polls guest queues for new connection send requests. The fast-path polls each guest transmit queue in a batched round-robin fashion to balance efficiency and low tail latency. First, the fast-path core selects the next guest and then starts polling the guest's transmit queues, until the batch is full, all the guest's queues are empty, or the guest resource budget is used up. If the batch is not full, the scheduler moves on to the next guest.

Pulling multiple transmit requests from a queue in one batch significantly reduces per-request overheads for queue access. Consolidating tasks for a specific guest within a batch also increases resource accounting accuracy as work from fewer guests is aggregated into the same batch. But even

Algorithm 1 TX Scheduler

```
function SCHEDULE_VMS(vms)
    n ← batch_size
    for vm in vms do
        if vm.budget > 0 and n > 0 then
            x ← SCHEDULE_FLOWS(vm.flows, n)
            n ← n - x
function SCHEDULE_FLOWS(flows, n)
    i ← 0
    for flow in flows do
        if i < n then
            x ← SCHEDULE_PACKETS(flow.packets, n - i)
            i ← i + x
```

across guests, processing requests in batches enables Virtuoso to improve efficiency by avoiding cache misses on key memory accesses through group prefetching [26]. During the processing of these transmit requests, the Virtuoso transport layer schedules the corresponding flows for packet transmission through TX tasks.

Scheduling TX for per-guest fairness. Virtuoso also schedules TX tasks with a similar batched hierarchical approach (Algorithm 1). The scheduler first chooses the next guest, and then the guest's next flow. In the first level of the scheduler the round-robin algorithm decides which guest should send next. The second level instead schedules flows according to a priority queue that tracks the earliest time when each flow should send next. The TCP processing logic determines these timestamps with a split fast-path/slow-path congestion control scheme [27]. These timestamps also automatically ensure that a guest's flows are serviced fairly. Guests without available budget are skipped until the budget is replenished.

Drop packets for out of budget VMs in RX. RX pulls packets from the NIC queues and performs a flow state lookup to identify the destination VM of the packet. Virtuoso drops the packet if it belongs to an out of budget VM. This is necessary to prevent a misbehaving sender from overwhelming the receiver, causing the receiving VM to use more than its fair share of resources. We found that latency can deteriorate by 2x if Virtuoso doesn't drop packets, but by dropping packets we keep performance interference at the same level as siloed stacks (Figure 7). Per-VM hardware queues can attenuate the number of packet drops, but the number of queues increases with the number of VMs, causing a drop in performance[34]. Alternatively, per-VM software queues are used in systems such as PicNIC[30], but as reported in the paper, they do not prevent drops of excess traffic and wasted work.

The system is self-correcting when sender and receiver enforce similar budget parameters, Virtuoso keeps track of the cycle deficit accrued when later replenishing the budget. Guests that deplete their budget on RX tasks as a result have fewer cycles available for POLL and TX tasks, so senders that do not receive replies will stop sending.

Round-robin packet scheduling on the slow-path. The slow-path employs per-VM packet queues and does round-robin scheduling among these queues to prevent starvation. This avoids the slow-path from harming the tail-latency of VMs when slow path requests are skewed towards only a few VMs. For example, when the slow-path congestion control algorithm calculates per-flow rates we have per-VM flow queues, such that each VM is fairly serviced.

4.5 Secure Shared Stack

Virtuoso processes packets from multiple guests and applications in the same network stack. Resource accounting and scheduling mechanisms provide performance isolation. For this we rely on shared memory queues between individual application cores and the fast-path, as well as a separate slow-path for more expensive control operations, akin to TAS [27] and SNAP [35]. However, we also need security enforcement while enabling efficient direct communication between applications and the Virtuoso stack.

Protecting memory regions. Virtuoso ensures security isolation for the guest and application interface through memory isolation. We allocate different guests' queues and connection buffers in separate shared memory regions only mapped into a single guest and the fast-path. To avoid leaks due to dynamic remapping and ensure resource isolation, Virtuoso statically pre-allocates the complete shared memory region when the guest starts. Virtuoso also has a narrow shared memory interface comprising just guest receive and send queues along with connection payload buffers. This narrow interface provides no other attack vectors, such as complex data structures that could interfere with Virtuoso.

5 Implementation

Virtuoso runs in host userspace as a separate service and provides all features of a typical TCP stack to guest applications. Virtuoso maintains TCP protocol and sockets API drop-in compatibility. For fast NIC access, we use DPDK [23]. We build our prototype using TAS [27] as a basis. We heavily modify and extend the TAS fast and slow-path, but retain the sockets emulation library unmodified. Virtuoso supports guest VMs as well as guest containers. The Virtuoso prototype comprises 20,918 lines total, 4,669 lines for the fast-path, 5,536 lines for the slow-path, 2,437 lines for the hypervisor integration, and 1,029 lines of modification to OvS.

5.1 Support for Multiple Guests

We modify TAS to use separate shared memory regions on the host and Unix sockets per guest. These regions contain transmit request and receive notification queues, as well as

per-flow RX and TX circular payload buffers. During guest initialization, Virtuoso passes a newly created shared memory region to the guest. As in TAS we implement this using Unix domain sockets, that carry a handshake along with the shared memory file descriptor. Unlike TAS, Virtuoso exposes separate listening Unix sockets for each guest, allowing it to securely identify which guest is connecting, assuming sufficient access control on the host.

5.2 Virtuoso with Containers

Given that guest containers share the same host operating system as Virtuoso, connecting applications in guest containers is simply a matter of mapping the respective guest Unix socket into the container. After this any container guest application can interact with Virtuoso as a native application with the same performance.

5.3 Virtuoso with Virtual Machines

Initialization is more complex for virtual machines, as Unix sockets and file descriptors are by definition local to the host. Virtuoso instead integrates with the hypervisor to directly map shared memory regions via a dummy PCI device. In the VM, a Virtuoso guest agent implements a user space driver for this dummy device and translates the interface into a compatible Unix socket and shared memory file descriptors, avoiding the need for applications to distinguish between native, container, and VM operation. No modifications are done to the guest operating system.

We utilize QEMU [45] with KVM, along with its Inter-VM Shared Memory Device (IVSHM) [24], to implement the hypervisor component of the Virtuoso integration. IVSHM allows an external application to send a shared memory file descriptor and eventfds for bi-directional interrupts to QEMU that are then exposed as a PCI device with the memory region as a directly memory mapped BAR. For ease of integration, we implement this as a separate host proxy process that connects to QEMU and Virtuoso.

In the guest we implement a Virtuoso guest agent, that leverages the vfio-pci [53] kernel module to implement a user space driver for the dummy PCI device. vfio-pci provides a file descriptor that the application can mmap for access to the BAR, along with eventfds for interrupts. The guest agent creates a listening unix socket, and during the handshake passes file descriptors directly to applications. This results in a directly shared memory region between Virtuoso on the host, and applications in the guest VM. As a result, fast-path interactions with Virtuoso incur no additional overheads compared to containers or native applications.

5.4 OvS Slow-path

We use OvS for virtualization management, to identify tunnelling information, and to determine the destination VM for a flow. To that end, we modify OvS to exchange packets and control information with Virtuoso. In OvS we implement custom transmit and receive netdev-provider ports. The receive port polls Virtuoso for new packets and passes them to OvS. OvS then performs internal matching based on the packet metadata and directs it to a transmit netdev-provider port. The transmit port holds tunnelling information for a packet and establishes a message queue with Virtuoso to dispatch this information. This message queue exchanges inner and outer IP addresses for encapsulated packets, tunnel IDs from GRE headers, and the appropriate ID for the destination VM.

6 Evaluation

In this section we evaluate how well Virtuoso addresses the goals outlined in §2. To that end, our evaluation answers the following questions:

- Does sharing the stack improve resource efficiency? (§6.1)
- Can fine-grained scheduling and resource accounting ensure isolation of tenants despite sharing resources? (§6.2)
- How close can optimized one-shot virtualization performance get to native un-virtualized stacks? (§6.3)
- Does Virtuoso scale to serve many guests? (§6.4)
- Does Virtuoso improve VM performance in oversubscribed machines? (§6.5)

Testbed. We configure two identical machines as client and server. They are directly connected with a pair of 100 Gbps Mellanox ConnectX-5 Ethernet adapters. Both machines have two Intel Xeon Gold 6152 processors at 2.1 GHz, each with 22 cores for a total of 44 cores and 187 GB of RAM per machine. We run Linux kernel 5.15 with Debian 11.

Baselines. We compare Virtuoso against a number of baseline configurations. For these we use two existing network stacks, the default in-kernel Linux network stack, and the optimized TAS TCP stack. Depending on the configuration, we run these bare metal, or in QEMU/kvm virtual machines with virtio-net vNICs connected to OvS. We configure OvS with the DPDK backend and use vhost-user between QEMU and OvS to get the best baseline performance. For containers, we directly mount the respective Unix socket into the containers for Virtuoso.

Focus on VMs. For most of our evaluation we focus on Virtuoso with virtual machine guests, rather than container. The dominating fast-path interaction performance is identical between Virtuoso VM guests and container guests, while some slow-path interactions in the VM case are more expensive than for containers (§6.3). At the same time we found Virtuoso to provide higher relative benefits when comparing to existing container stacks than compared with VM stacks. Thus, Virtuoso with VMs provides a conservative evaluation and performance comparison.

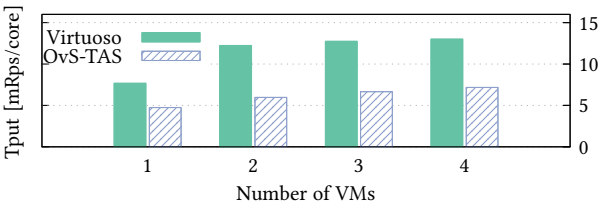

Figure 6. Virtuoso exhibits higher resource efficiency in per-core aggregate VM throughput with bursty guests.

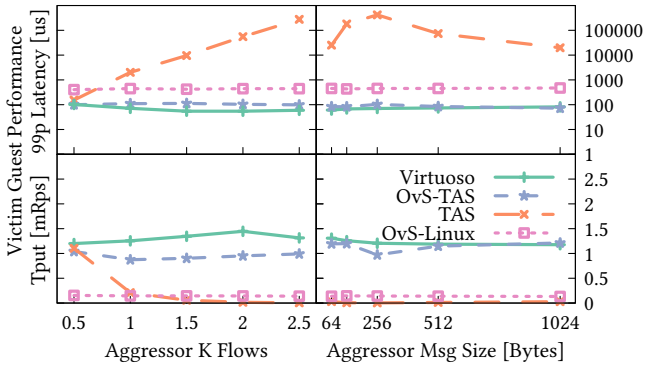

Figure 7. Victim guests using Virtuoso achieve tail latency on par with siloed OvS-TAS and higher throughput, while aggressor guest induces interference.

6.1 Sharing the Stack Improves Efficiency

We begin by measuring resource efficiency in bursty guest workloads. For this we provision four guest VMs with echo servers responding to RPCs. Clients generate bursty high-low traffic, separately saturating each guest during peaks. We then vary the degree of overlap, i.e. how many of the guests burst concurrently, from one to a maximum of four VMs.

As a baseline, we use separate TAS network stacks in each guest connected to OvS on the host (OvS-TAS). For the baseline we provision each guest with five cores, and configure the TAS instances to use one fast-path core. For Virtuoso we instead provision each guest with four cores, and assign three fast-path cores to Virtuoso.

Figure 6 shows the per-core aggregate RPC throughput across all guests during bursty periods. We obtained throughput numbers by dividing the aggregate throughput by the number of fast-path cores used by Virtuoso and the baseline. Virtuoso achieves 82% higher per-core throughput when four VMs are bursting. Our results show sharing the stack allows Virtuoso to pool resources and thereby significantly improve overall system efficiency.

6.2 Fine-grained Scheduling Isolates VMs

Next, we evaluate Virtuoso ability to isolate guests despite sharing a network stack and underlying resources. To that end, we evaluate two main performance metrics, latency and throughput, for a "victim" guest while a separate aggressor guest attempts to introduce performance interference.

We evaluate two different forms of interference, by separately varying the number of aggressor connections and the size of the aggressor messages. For the former the aggressor uses a fixed message size of 64 bytes, and for the latter a fixed number of 500 connections. The victim uses one connection with 64 B messages for the latency measurements, and 500 connections with 64 B messages for the throughput measurement. The victim uses a single core VM, while we provision the aggressor VM with a core for every 500 connections. Both victim and aggressor use the RPC echo server.

We compare this workload across different system configurations. Virtuoso with two fast-path cores, OvS-TAS with one additional fast-path core per guest VM for the TAS instance, the guest Linux stack with OvS with no additional guest cores. Finally, we also compare to native TAS by running victim and aggressor as separate processes on the host connecting to the same TAS instance with two fast-path cores. We use the timely [37] congestion control algorithm in Virtuoso, OvS-TAS, and TAS. In all cases VMs, processes, and network stacks are pinned to dedicated cores and Virtuoso equally divides resources for each guest by setting the same value of w_g for all guests.

Figure 7 shows the results. At a high level, the results confirm that Virtuoso's fine-grained isolation retains tail latencies below siloed OvS-TAS, while improving victim throughput. TAS without isolation increases tail-latency significantly as the aggressor's message size increases. TAS incurs tail latencies above Virtuoso and OvS-TAS because of the lack of isolation mechanisms based on resource usage. For example, at 2500 aggressor connections Virtuoso achieves a 99p latency of 60 μs, OvS-TAS's 98 μs,and TAS's 276,709 μs. The benefits of fine-grained scheduling also hold when comparing median latencies of the baselines. The Virtuoso victim achieves 40 μs 50p latency when the aggressor VM sends 1024 B messages, while the TAS and OvS-TAS clients achieve 1461 μs and 45 μs 50p latencies.

We also measure similar results with the victim's throughput. TAS sees a decrease in throughput as the aggressor VM increases the number of connections or message size. With Virtuoso, the victim maintains similar throughput as the aggressor attempts to acquire more resources. For 2500 aggressor connections, Virtuoso achieves 34% higher throughput than OvS-TAS.

6.3 One-shot Processing Reduces Overhead

Virtualization overhead. First, we seek to measure and break down the overheads of adding network virtualization features to the network stack. For this, we start with the

	TAS	Scheduling	VM	GRE
Cycles/RPC	436	446	458	465
Overhead	-	+ 2.3%	+ 5.0%	+ 6.7%

Table 1. Request processing times for different Virtuoso features relative to the native baseline (TAS).

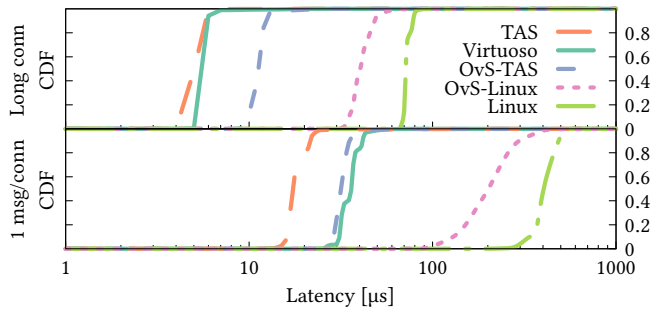

Figure 8. RPC latency distribution across different network stacks. For long-lived connections Virtuoso adds minimal overhead relative to TAS, while tail latency for short-lived connections, the Virtuoso worst-case, remains competitive.

TAS fast-path as the baseline and profile the number of processor cycles required to process an RPC request including sending the response. The application workload saturates a single Virtuoso core with 64 B RPCs. We then successively add Virtuoso features, starting with scheduling, then VM integration, and finally GRE tunneling.

Table 1 shows the results. Fine grained scheduling and resource accounting adds around 10 cycles or 2.3% to each RPC. Enabling VM integration adds 12 cycles, and finally GRE tunneling adds another 7 cycles per RPC. In total, the additional functionality in Virtuoso only adds a total of 133 cycles or 6.7% of overhead. We also separately measured the overhead of network virtualization on throughput. Enabling GRE tunneling on OvS-TAS decreased throughput by 12%, while running Virtuoso with GRE tunneling only adds an overhead of 6%. We conclude that one-shot processing is effective for avoiding expensive overhead for significant additional network virtualization functionality.

Latency. These minimal overheads translate to minimal latency increase for virtualized guests in Virtuoso compared to TAS. We measure the small 64 B RPC latency, both for long-lived connections and short-lived connections that only carry one RPC before closing and re-opening, also including latency for establishment and tear-down. We record latency distributions for all our system configurations and report the results in Figure 8.

With long connections Virtuoso achieves median latencies of 5 μs compared to 4 μs with bare-metal TAS, and 8 μs 99p

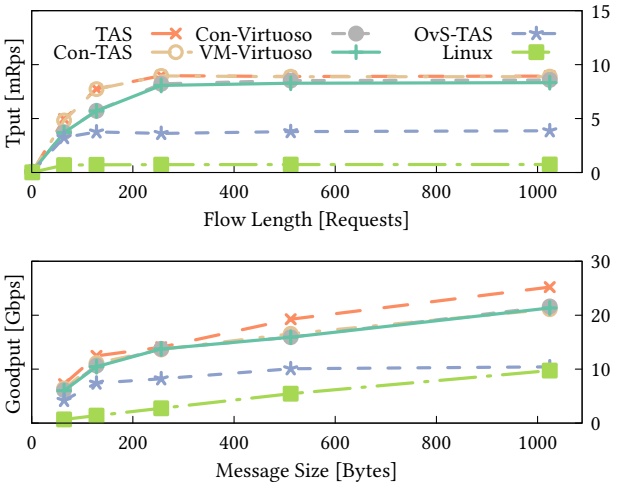

Figure 9. Virtuoso achieves throughput similar to unvirtualized TAS with diverse connection lengths and message sizes, surpassing other virtualized stacks.

latency compared to 6 μs for TAS. OvS-TAS only achieves median latencies of 12 μs and a 99p latency of 19 μs, both about a factor of two higher than Virtuoso. Native Linux and Linux VMs with OvS are both significantly worse, although interestingly we found that the DPDK drivers in OvS seem to reduce overheads compared to the native in-kernel drivers, thereby surprisingly lowering the latency.

Short-lived connections are Virtuoso' Achilles heel, as one-shot connection state management optimizes for fast access to established state at the cost of overhead for adding and removing connections. The extreme case of connections that send only one RPC before being torn down again, factoring in complete time for establishment and tear-down, probes this. For average latency Virtuoso is slower than TAS without OvS, at 36 μs compared to 18 μs and similar to OvS-TAS' 32 μs. In the tail Virtuoso shows 99p latencies of 52 μs compared to TAS' 26 μs and OvS-TAS 50 μ. We suspect that this is due to inefficiencies in the Virtuoso connector in OvS that may be less optimized than the vhost-user port we use for OvS-TAS. Linux is again far slower. We conclude Virtuoso enables virtualized networking with minimal latency overhead compared to unvirtualized stacks.

Throughput. One-shot virtualization processing also allows Virtuoso to achieve high throughput, comparable to bare metal performance. In Figure 9 we dedicate the same number of cores to the networking stack in a client and server machine running an RPC echo server. Server and client applications run on 12 cores each, and we dedicate 10 cores to Virtuoso and TAS. We again measure throughput for short and long-lived connections.

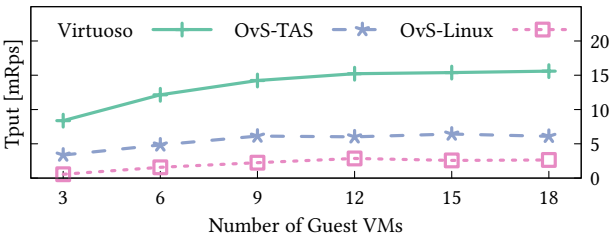

Figure 10. Virtuoso significantly outperforms alternative stacks even with many guests.

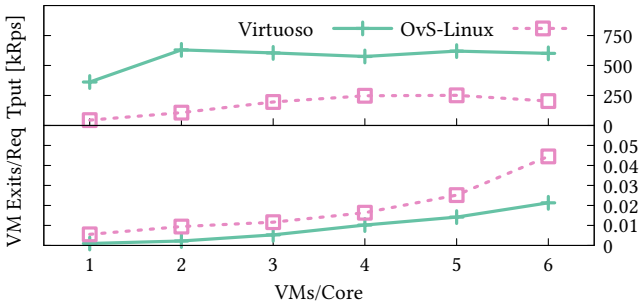

Figure 11. Virtuoso has fewer VM exits per request maintining performance when cores are oversubscribed.

We first vary the number of messages per connection and measure throughput. The more expensive Virtuoso slow-path is apparent for short-lived connections, but as more messages are sent per connection the gap between Virtuoso and bare-metal solutions decreases. Virtuoso also achieves throughput competitive with TAS for long-lived connections. For 1024 B messages Virtuoso reaches throughput only 14% lower than TAS, while OvS-TAS shows a performance drop of 56%. Baselines with containers show that there is little to no overhead between running Virtuoso in VMs (VM-Virtuoso) as opposed to containers (Con-Virtuoso). But the container baselines perform slightly worse than Virtuoso at 1024 B, possibly because of inneficiencies in the container runtime. Linux is again not competitive. We conclude one-shot processing also enables high-throughput virtualized network communication.

6.4 Virtuoso Scales to Many Guests

We evaluate guest scalability in Virtuoso. For each run we provision two cores for each guest VM and measure the aggregate throughput as the number of guests increases. Each VM runs an RPC echo server loaded 500 connections sending 64 B messages. We use four fast-path cores for Virtuoso with one polling core in OvS. In the OvS-TAS and OvS-Linux baselines we assign six polling cores to OvS.

Figure 10 shows the results. Virtuoso sees a increase in throughput up to 12 guest VMs. At 15 million requests per second, the four fast-path cores in Virtuoso are saturated and performance stabilizes, higher throughput can be achieved by allocating more fast-path cores. At 18 VMs, Virtuoso achieves 156% higher throughput than OvS-TAS and 490% higher throughput than OvS-Linux. OvS-TAS needs at least one core for the slow-path and one core for the fast-path, so in this setup TAS cores inside a VM compete with the application for resources, resulting in a smaller performance gain when compared to OvS-Linux. We conclude that Virtuoso outperforms alternatives at scale and scales to the number of guests on typical cloud servers.

6.5 Shared Stack Reduces VM Exits

We evaluate Virtuoso's performance under oversubscribed conditions, where multiple VMs share the same core. In this test, one core is dedicated to Virtuoso's fast path, while OvS-Linux assigns a thread to the DPDK polling thread. We increase the number of VMs running an echo server on one core and measure aggregate throughput and VM exits. Figure 11 shows that for OvS-Linux, more VMs lead to more VM exits and degraded throughput. In contrast, Viridian's shared stack architecture reduces VM exits by minimizing the times we pass control to the guest OS, maintaining stable throughput as more VMs are added.

7 Related Work

Performance isolation. Caladan[14] improves CPU utilization and provides isolation by dynamically adjusting core allocations. Andromeda[4] enforces performance isolation with per-VM coprocessor threads which are accounted to the VM, but the paper does not evaluate its claim of performance isolation. Both use coarse-grain allocations incurring context switching and reallocation overheads. Virtuoso instead implements fine-grain performance isolation at the packet level without context-switches. PicNIC[30] maintains predictable SLOs for network virtualization with receiver-driven congestion control between virtual switches and back-pressure to guests. PicNIC is still a siloed architecture with per-guest network stacks, and only ensures isolation for the virtual switch. Virtuoso instead enables sharing and isolating the full network stack. Virtuoso also implements receiver limits through the existing TCP flow control. FairNIC[18] isolates tenants sharing a SmartNIC by statically partitioning the SmartNIC cores. This coarse-grain static partitioning achieves lower resource utilization compared to Virtuoso fine-grained sharing. Junction[13] relies on a centralized scheduler to make core allocation decisions and maintain low tail-latency between instances, but Junction only provides a kernel bypass solution to containers and not VMs and relies on user inter processor interrupts (UIPIs) [22] to implement fine-grained timeslicing, which is not widely available in hardware.

Shared host-level network stack. NetKernel [39] also proposes extracting the network stack from VMs and sharing it between multiple VMs. Unlike Virtuoso, NetKernel keeps virtual switching and network virtualization separate from the rest of the network stack. NetKernel does not evaluate μs-scale latency isolation. Snap [35] implements a shared network stack as a user space service for multi-tenant environments. Snap is internally structured as communicating engines running as separate threads, with coarse-grained isolation by dedicating engines to clients, either dynamically or statically allocated to cores. Virtuoso per-packet scheduling instead improves resource efficiency through sharing and efficient batching, by processing packets from multiple VMs together, without compromising isolation. Unfortunately neither NetKernel nor Snap are available for direct comparison.

Container overlay networks. Overlay networks [6, 12, 55] implement container network virtualization. Slim [60] avoids the typical per-packet virtualization for containers and does not require packet transformations in the data plane. However, it does so by avoiding protocol-level network virtualization and instead directly sends packets on the physical network only translating address info in socket calls, and thus only works for networks that exclusively use Slim. Slim also does not provide additional mechanisms over Linux for performance isolation and lacks support for VMs.

8 Conclusion

With Virtuoso we have shown that network processing for virtual machines and container environments can be implemented efficiently in software. By sharing resources and using fine-grained scheduling for isolation Virtuoso achieves resource utilization far above other alternatives. And one-shot network virtualization enables implementation of the necessary virtualization functionality with minimal overhead over optimized bare metal stacks. We expect that our techniques can generalize to other protocols and implementation on other architectures, such as SoC-SmartNICs.

Acknowledgments

We thank Florian Bauckholt and Mehrshad Lotfi for proof-of-concept prototypes of the Virtuoso VM and OpenVSwitch integration respectively. We also thank Rajath Shashidhara for his contributions in the long running discussions over the course of this project.

References

[1] Saksham Agarwal, Rachit Agarwal, Behnam Montazeri, Masoud Moshref, Khaled Elmeleegy, Luigi Rizzo, Marc Asher de Kruijf, Gautam Kumar, Sylvia Ratnasamy, David Culler, and Amin Vahdat. Understanding host interconnect congestion. In *21st ACM Workshop on Hot Topics in Networks*, HotNets, 2022.

[2] Amazon Web Services. AWS Nitro system. https://aws.amazon.com/ec2/nitro/.

[3] Adam Belay, George Prekas, Ana Klimovic, Samuel Grossman, Christos Kozyrakis, and Edouard Bugnion. IX: A protected dataplane operating system for high throughput and low latency. In *11th USENIX Symposium on Operating Systems Design and Implementation*, OSDI, 2014.

[4] Michael Dalton, David Schultz, Jacob Adriaens, Ahsan Arefin, Anshuman Gupta, Brian Fahs, Dima Rubinstein, Enrique Cauich Zermeno, Erik Rubow, James Alexander Docauer, Jesse Alpert, Jing Ai, Jon Olson, Kevin DeCabooter, Marc de Kruijf, Nan Hua, Nathan Lewis, Nikhil Kasinadhuni, Riccardo Crepaldi, Srinivas Krishnan, Subbaiah Venkata, Yossi Richter, Uday Naik, and Amin Vahdat. Andromeda: Performance, isolation, and velocity at scale in cloud network virtualization. In *15th USENIX Symposium on Networked Systems Design and Implementation*, NSDI, 2018.

[5] Jeffrey Dean and Luiz André Barroso. The tail at scale. *ACM Transactions on Computer Systems*, 56(2):74–80, February 2013.

[6] Docker overlay. https://docs.docker.com/network/.

[7] G. Dommety. Key and sequence number extensions to GRE, September 2000. RFC 2890.

[8] Peter Druschel, Larry Peterson, and Bruce Davie. Experiences with a high-speed network adaptor: A software perspective. In *1995 ACM SIGCOMM Conference on Data Communication*, SIGCOMM, 1995.

[9] Haggai Eran, Lior Zeno, Maroun Tork, Gabi Malka, and Mark Silberstein. NICA: An infrastructure for inline acceleration of network applications. In *2019 USENIX Annual Technical Conference*, ATC, 2019.

[10] D. Farinacci, T. Li, S. Hanks, D. Meyer, and P. Traina. Generic routing encapsulation (GRE), March 2000. RFC 2794.

[11] Daniel Firestone. VFP: A virtual switch platform for host SDN in the public cloud. In *14th USENIX Symposium on Networked Systems Design and Implementation*, NSDI, 2017.

[12] Flannel. https://github.com/flannel-io/flannel.

[13] Joshua Fried, Gohar Irfan Chaudhry, Enrique Saurez, Esha Choukshe, Íñigo Goiri, Sameh Elnikety, Rodrigo Fonseca, and Adam Belay. Making kernel bypass practical for the cloud with junction. In *21th USENIX Symposium on Networked Systems Design and Implementation*, NSDI, 2024.

[14] Joshua Fried, Zhenyuan Ruan, Amy Ousterhout, and Adam Belay. Caladan: Mitigating interference at microsecond timescales. In *14th USENIX Symposium on Operating Systems Design and Implementation*, OSDI, 2020.

[15] P. Garg and Y. Wang. Nvgre: Network virtualization using generic routing encapsulation, September 2015. RFC 7637.

[16] Yoann Ghigoff, Julien Sopena, Kahina Lazri, Antoine Blin, and Gilles Muller. BMC: Accelerating memcached using safe in-kernel caching and pre-stack processing. In *18th USENIX Symposium on Networked Systems Design and Implementation*, NSDI, 2021.

[17] Rahul Ghosh and Vijay K. Naik. Biting off safely more than you can chew: Predictive analytics for resource over-commit in iaas cloud. In *Fifth IEEE International Conference on Cloud Computing*, CLOUD, 2012.

[18] Stewart Grant, Anil Yelam, Maxwell Bland, and Alex C. Snoeren. Smartnic performance isolation with fairnic: Programmable networking for the cloud. In *2020 ACM SIGCOMM Conference on Data Communication*, SIGCOMM, 2020.

[19] J. Gross, I. Ganga, and T. Sridhar. Geneve: Generic network virtualization encapsulation, November 2020. RFC 8926.

[20] Jack Tigar Humphries, Kostis Kaffes, David Mazières, and Christos Kozyrakis. A case against (most) context switches. In *18th Workshop on Hot Topics in Operating Systems*, HOTOS, 2021.

[21] Intel Corporation. PCI-SIG SR-IOV primer: An introduction to SR-IOV technology. *Intel application note*, January 2011. Revision 2.5.

[22] Intel Corporation. Intel 64 and IA-32 architectures software developer's manual. https://www.intel.com/content/www/us/en/developer/articles/technical/intel-sdm.html, July 2024.

[23] Intel data plane development kit. http://www.dpdk.org/.

[24] Inter-VM shared memory device – QEMU documentation. https://www.qemu.org/docs/master/system/devices/ivshmem.html.

[25] Eun Young Jeong, Shinae Woo, Muhammad Jamshed, Haewon Jeong, Sunghwan Ihm, Dongsu Han, and KyoungSoo Park. mTCP: A highly scalable user-level TCP stack for multicore systems. In *11th USENIX Symposium on Networked Systems Design and Implementation*, NSDI, 2014.

[26] Anuj Kalia, Dong Zhou, Michael Kaminsky, and David G. Andersen. Raising the bar for using GPUs in software packet processing. In *12th USENIX Symposium on Networked Systems Design and Implementation*, NSDI, 2015.

[27] Antoine Kaufmann, Tim Stamler, Simon Peter, Naveen Kr. Sharma, Arvind Krishnamurthy, and Thomas Anderson. TAS: TCP acceleration as an OS service. In *14th ACM European Conference on Computer Systems*, EuroSys, 2019.

[28] Hsiao keng Jerry Chu. Zero-copy TCP in Solaris. In *1996 USENIX Annual Technical Conference*, ATC, 1996.

[29] M. Kerrisk. veth - virtual ethernet device. https://man7.org/linux/man-pages/man4/veth.4.html, February 2023.

[30] Praveen Kumar, Nandita Dukkipati, Nathan Lewis, Yi Cui, Yaogong Wang, Chonggang Li, Valas Valancius, Jake Adriaens, Steve Gribble, Nate Foster, and Amin Vahdat. PicNIC: predictable virtualized NIC. In *2019 ACM SIGCOMM Conference on Data Communication*, SIGCOMM, 2019.

[31] I.M. Leslie, D. McAuley, R. Black, T. Roscoe, P. Barham, D. Evers, R. Fairbairns, and E. Hyden. The design and implementation of an operating system to support distributed multimedia applications. *IEEE Journal on Selected Areas in Communications*, 14(7):1280–1297, 1996.

[32] Bojie Li, Tianyi Cui, Zibo Wang, Wei Bai, and Lintao Zhang. Socksdirect: datacenter sockets can be fast and compatible. In *2019 ACM SIGCOMM Conference on Data Communication*, SIGCOMM, 2019.

[33] Yan Luo, Eric Murray, and Timothy L Ficarra. Accelerated virtual switching with programmable nics for scalable data center networking. In *2nd ACM SIGCOMM Workshop on Virtualized Infrastructure Systems and Architectures*, VISA, 2010.

[34] Maziar Manesh, Katerina Argyraki, Mihai Dobrescu, Norbert Egi, Kevin Fall, Gianluca Iannaccone, Eddie Kohler, and Sylvia Ratnasamy. Evaluating the suitability of server network cards for software routers. In *3rd ACM Workshop on Programmable Routers for Extensible Services of Tomorrow*, PRESTO, 2010.

[35] Michael Marty, Marc de Kruijf, Jacob Adriaens, Christopher Alfeld, Sean Bauer, Carlo Contavalli, Michael Dalton, Nandita Dukkipati, William C. Evans, Steve Gribble, Nicholas Kidd, Roman Kononov, Gautam Kumar, Carl Mauer, Emily Musick, Lena Olson, Erik Rubow, Michael Ryan, Kevin Springborn, Paul Turner, Valas Valancius, Xi Wang, and Amin Vahdat. Snap: a microkernel approach to host networking. In *27th ACM Symposium on Operating Systems Principles*, SOSP, 2019.

[36] Microsoft Corporation. Project Catapult. https://www.microsoft.com/en-us/research/project/project-catapult/.

[37] Radhika Mittal, Vinh The Lam, Nandita Dukkipati, Emily Blem, Hassan Wassel, Monia Ghobadi, Amin Vahdat, Yaogong Wang, David Wetherall, and David Zats. TIMELY: RTT-based congestion control for the datacenter. In *2015 ACM SIGCOMM Conference on Data Communication*, SIGCOMM, 2015.

[38] David Mosberger and Larry L. Peterson. Making paths explicit in the Scout operating system. In *2nd USENIX Symposium on Operating Systems Design and Implementation*, OSDI, 1996.

[39] Zhixiong Niu, Hong Xu, Peng Cheng, Qiang Su, Yongqiang Xiong, Tao Wang, Dongsu Han, and Keith Winstein. NetKernel: Making network stack part of the virtualized infrastructure. In *2020 USENIX Annual Technical Conference*, ATC, 2020.

[40] NVIDIA. ConnectX-7 400G Adapters. https://nvdam.widen.net/s/csf8rmnqwl/infiniband-ethernet-datasheet-connectx-7-ds-nv-us-2544471, December 2022.

[41] NVIDIA. NVIDIA Bluefield-3 DPU. https://resources.nvidia.com/en-us-accelerated-networking-resource-library/datasheet-nvidia-bluefield?lx=LbHvpR&topic=networking-cloud, March 2023.

[42] Open vswitch. https://www.openvswitch.org/.

[43] Simon Peter, Jialin Li, Irene Zhang, Dan R. K. Ports, Doug Woos, Arvind Krishnamurthy, Thomas Anderson, and Timothy Roscoe. Arrakis: The operating system is the control plane. *ACM Transactions on Computer Systems*, 33(4):11:1–11:30, November 2015.

[44] Boris Pismenny, Adam Morrison, and Dan Tsafrir. ShRing: Networking with shared receive rings. In *17th USENIX Symposium on Operating Systems Design and Implementation*, OSDI, 2023.

[45] QEMU – the FAST! processor emulator. https://www.qemu.org/.

[46] Shixiong Qi, Leslie Monis, Ziteng Zeng, Ian chin Wang, and K.K. Ramakrishnan. SPRIGHT: extracting the server from serverless computing! high-performance ebpf-based event-driven, shared-memory processing. In *2022 ACM SIGCOMM Conference on Data Communication*, SIGCOMM, 2022.

[47] Hugo Sadok, Nirav Atre, Zhipeng Zhao, Daniel S. Berger, James C. Hoe, Aurojit Panda, Justine Sherry, and Ren Wang. Enso: A streaming interface for NIC-Application communication. In *17th USENIX Symposium on Operating Systems Design and Implementation*, OSDI, 2023.

[48] Rajath Shashidhara, Tim Stamler, Antoine Kaufmann, and Simon Peter. FlexTOE: Flexible TCP offload with Fine-Grained parallelism. In *19th USENIX Symposium on Networked Systems Design and Implementation*, NSDI, 2022.

[49] M. Shreedhar and George Varghese. Efficient fair queueing using deficit round robin. In *1995 ACM SIGCOMM Conference on Data Communication*, SIGCOMM, 1995.

[50] M. Mahalingam Storvisor, D. Dutt, K. Duda, P. Agarwal, L. Kreeger, T. Sridhar, M. Bursell, and C. Wright. Virtual extensible local area network (vxlan): A framework for overlaying virtualized layer 2 networks over layer 3 networks, August 2014.

[51] David L. Tennenhouse. Layered multiplexing considered harmful. In *Protocols for High Speed Networks I*, PfHSN, 1989.

[52] M. Tsirkin and C. Huck. Virtual i/o device (VIRTIO) version 1.2. https://docs.oasis-open.org/virtio/virtio/v1.2/virtio-v1.2.html, July 2022.

[53] VFIO - "virtual function I/O". https://docs.kernel.org/driver-api/vfio.html.

[54] T. von Eicken, A. Basu, V. Buch, and W. Vogels. U-Net: a user-level network interface for parallel and distributed computing. In *15th ACM Symposium on Operating Systems Principles*, SOSP, 1995.

[55] Weave. https://www.weave.works/.

[56] Damon Wischik, Mark Handley, and Marcelo Bagnulo Braun. The resource pooling principle. *SIGCOMM Computer Communication Review*, 38(5):47–52, September 2008.

[57] Irene Zhang, Amanda Raybuck, Pratyush Patel, Kirk Olynyk, Jacob Nelson, Omar S. Navarro Leija, Ashlie Martinez, Jing Liu, Anna Kornfeld Simpson, Sujay Jayakar, Pedro Henrique Penna, Max Demoulin, Piali Choudhury, and Anirudh Badam. The Demikernel datapath OS architecture for microsecond-scale datacenter systems. In *28th ACM Symposium on Operating Systems Principles*, SOSP, 2021.

[58] Niu Zhixiong, Hong Xu, Dongsu Han, Peng Cheng, Yongqiang Xiong, Guo Chen, and Keith Winstein. Network stack as a service in the cloud. In *16th ACM Workshop on Hot Topics in Networks*, HotNets, 2017.

[59] Yang Zhou, Zezhou Wang, Sowmya Dharanipragada, and Minlan Yu. Electrode: Accelerating distributed protocols with ebpf. In *20th USENIX Symposium on Networked Systems Design and Implementation*, NSDI, 2023.

[60] Danyang Zhuo, Kaiyuan Zhang, Yibo Zhu, Hongqiang Harry Liu, Matthew Rockett, Arvind Krishnamurthy, and Thomas Anderson. Slim: OS kernel support for a Low-Overhead container overlay network. In

16th USENIX Symposium on Networked Systems Design and Implementation, NSDI, 2019.