FastIOV: Fast Startup of Passthrough Network I/O Virtualization for Secure Containers

Yunzhuo Liu[1,2*], Junchen Guo[2*], Bo Jiang[1†], Yang Song[2], Pengyu Zhang[2], Rong Wen[2], Biao Lyu[3,2], Shunmin Zhu[4,2], Xinbing Wang[1]

[1]Shanghai Jiao Tong University Shanghai, China
[2]Alibaba Cloud Hangzhou, China
[3]Zhejiang University Hangzhou, China
[4]Hangzhou Feitian Cloud Hangzhou, China
{liu445126256,bjiang,xwang8}@sjtu.edu.cn,alibaba_cloud_network@alibaba-inc.com

Abstract

Single Root I/O Virtualization (SR-IOV) technology has advanced in recent years and can simultaneously satisfy the network requirements of high data plane performance, high deployment density, and fast startup for applications in traditional containers. However, it falls short with secure containers, which have become the mainstream choice in multi-tenant clouds. SR-IOV requires secure containers to use passthrough I/O for higher data plane performance, which hinders the container startup performance and prevents its usage in time-sensitive tasks like serverless computing. In this paper, we advocate that the startup performance of SR-IOV enabled secure containers can be further boosted, making SR-IOV suitable for building a *Container Network Interface* (CNI) for secure containers. We first dissect the end-to-end concurrent startup process and identify three key bottlenecks that lead to the slow startup, including *Virtual Function I/O* device set management, *Direct Memory Access* memory mapping, and *Virtual Function* (VF) driver initialization. We then propose a CNI named FastIOV that addresses these bottlenecks through lock decomposition, unnecessary mapping skipping, decoupled zeroing, and asynchronous VF driver initialization. Our evaluation shows that FastIOV reduces the overhead of enabling SR-IOV for secure containers by 96.1%, achieving 65.7% and 75.4% reductions in the average and 99th percentile end-to-end startup time.

*Yunzhuo Liu and Junchen Guo are the co-first authors.
†Bo Jiang is the corresponding author.

EuroSys '25, March 30-April 3, 2025, Rotterdam, Netherlands

ACM ISBN 979-8-4007-1196-1/25/03
https://doi.org/10.1145/3689031.3696066

CCS Concepts: • **Networks → Overlay and other logical network structures**; **Cloud computing**; *Data center networks.*

Keywords: Container Network, Passthrough I/O Virtualization, SR-IOV, Secure Container, Overlay Network Startup

ACM Reference Format:
Yunzhuo Liu[1,2*], Junchen Guo[2*], Bo Jiang[1†], Yang Song[2], Pengyu Zhang[2],, Rong Wen[2], Biao Lyu[3,2], Shunmin Zhu[4,2], Xinbing Wang[1]. 2025. FastIOV: Fast Startup of Passthrough Network I/O Virtualization for Secure Containers. In *Twentieth European Conference on Computer Systems (EuroSys '25), March 30-April 3, 2025, Rotterdam, Netherlands.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3689031.3696066

1 Introduction

Nowadays, mainstream cloud providers have been progressively shifting from virtual machines to containers as their new compute instances. The container-enabled cloud services, such as NoSQL database, *e.g.,* Azure Cosmos [6], and serverless function compute, *e.g.,* AWS Lambda [3], necessitate network access to either serve incoming requests or interact with other services like cloud storage. The container network is required to achieve not only high data plane performance but also high deployment density, *i.e.,* a large number of virtual network devices on a single server, and fast startup [1, 20, 24, 35, 47, 51, 57, 60, 61].

Single Root I/O Virtualization (SR-IOV) [13] has emerged as the best-performing approach to simultaneously satisfy the above three requirements for traditional containers. First, SR-IOV virtualizes a *Network Interface Card* (NIC) into multiple virtual NICs named *Virtual Functions* (VFs). It is a hardware-assisted device virtualization technology that allows containers to interact with the NIC resources more directly and achieve near bare-metal data plane performance, while other container network solutions, like software based *Container Network Interface* (CNI), incur obvious overhead in throughput and latency [2, 48, 49]. Second, the deployment density of SR-IOV has also been greatly improved with the emerging technologies, such as mdev [56], Intel Scalable IOV [17], and HD-IOV [70]. The newest commercial NICs like Mellanox CX-7 [18] and Intel IPU [16] have announced the vanilla support of 1K VFs. Finally, the startup of SR-IOV for a traditional container is fast, as its main procedure is just moving

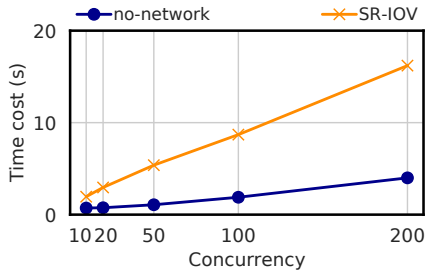


Figure 1. Overhead of enabling SR-IOV on secure container startup time with concurrency up to 200[1].

the pre-created Linux network interface of a VF on the host into the network namespace of the container.

However, despite the commendable performance of SR-IOV for traditional containers, it still falls short when applied to secure containers. Secure containers like Firecracker [1], Kata [24], and RunD [35], have become the mainstream choices in multi-tenant clouds where security is highly valued. They run the container processes inside *micro Virtual Machines* (microVMs) with trimmed and independent kernels to provide better isolation against attacks such as privilege escalation. Unlike traditional containers that can directly use the pre-created Linux network interface of a VF, microVMs require an extra virtualization process named *passthrough I/O* to efficiently use the VF due to the existence of the independent kernels. We find that the overhead of this virtualization greatly hinders the startup performance of secure containers. Fig. 1 illustrates the effect of enabling SR-IOV on the average time of concurrently starting 10 ∼ 200 secure containers. We observe that enabling SR-IOV incurs a significant time overhead that increases with the concurrency. The time overhead is 12.2s when the concurrency is 200, increasing the average time by 305%. Such slow startup poses a significant obstacle in developing a desirable secure container network solution with SR-IOV.

In this work, we study the problem of *achieving fast concurrent startup for SR-IOV enabled secure containers*. By demystifying the end-to-end concurrent startup procedure of SR-IOV enabled secure containers, we identify several key bottlenecks that have not been addressed before in low-density scenarios. Then, we propose FastIOV, an enhanced SR-IOV solution that tackles those bottlenecks and achieves ultra-fast startup. Our contributions are summarized as follows.

- **Measurement results (§3):** We dive into the details of the components, from the user-space CNI plugin and the container runtime to the kernel-space device driver and OS modules. We identify three major bottlenecks related to passthrough I/O are identified: *Virtual Function I/O (VFIO) device set (devset) management*, *Direct Memory Access (DMA) memory mapping* and *VF driver initialization*. These bottlenecks are not coupled with any specific CNI or secure container framework implementations. They contribute more than 70% and 80% of the average and 99th percentile container startup time, respectively. As far as we know, we are the first to thoroughly analyze and elaborate on the end-to-end concurrent startup process of SR-IOV enabled secure containers.
- **Optimization solutions (§4):** Targeting the key bottlenecks, FastIOV first decomposes the coarse-grained lock design in VFIO devset management using a hierarchical lock framework, which parallelizes VFIO device operations while ensuring consistency (§4.2.1). Second, we identify the causes of the inefficiency in DMA memory mapping as the mapping of unnecessary memory regions and memory zeroing overhead. FastIOV tracks and skips the unnecessary regions, and decouples memory zeroing from mapping to enable lazy zeroing (§4.3). Finally, FastIOV asynchronously executes VF driver initialization with container application launching, effectively masking the overhead (§4.2.2).
- **Implementation and performance gain (§6):** We implement FastIOV with a portable Linux kernel module, a CNI plugin, and other optimizations in the secure container framework and OS modules. We conduct extensive experiments and demonstrate that FastIOV reduces the time overhead of enabling SR-IOV by 96.1%, leading to 65.7% and 75.4% reductions in the average and 99th percentile container startup time compared with vanilla SR-IOV CNI [23]. We also evaluate FastIOV on four representative serverless applications and show that FastIOV reduces the average and the 99th percentile task completion time by 12.1%-53.5% and 20.3%-53.7%, respectively.
- **Community contribution:** We open source the implementation of FastIOV as well as its benchmarking tools and dataset at https://github.com/AlibabaResearch/fastiov-eurosys25.

2 Background

2.1 SR-IOV and Passthrough I/O

Fig. 2 shows the architecture of SR-IOV with passthrough I/O. The physical resources of an SR-IOV NIC are managed by its *Physical Function* (PF), which is bound to the host OS through the PF driver. The goal of SR-IOV is to divide the NIC resources, such as registers and TX/RX queues, into multiple isolated sets and generate multiple virtual NICs, referred to as VFs. Traditional containers rely on the network drivers of the shared host kernel to access the VFs. In contrast, secure containers have independent guest kernels and manage

[1]We are actually showing the performance of the optimized version of SR-IOV CNI that resolves an implementation flaw of driver rebinding in Kata, as described in §5. The original version [23] performs much worse. The concurrency setting is based on the statistics collected from Alibaba serverless platform, which shows that over 200 container invocation requests can arrive nearly simultaneously at one server [35].

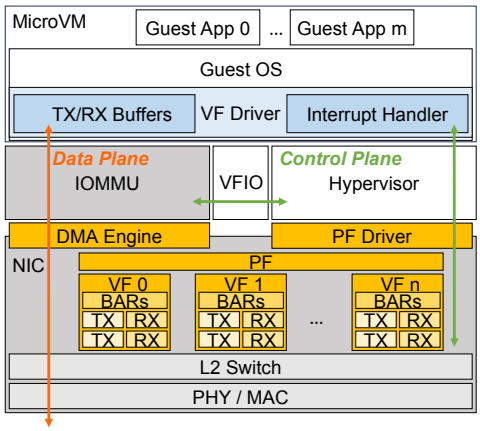


Figure 2. SR-IOV and passthrough I/O architecture.

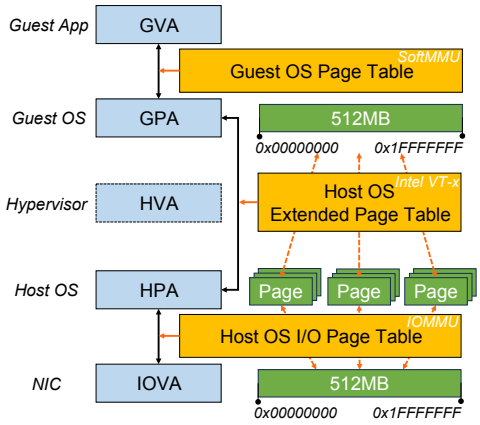


Figure 3. Address spaces and translations.

the VFs through the network drivers of the guest kernels. For higher performance, they utilize the passthrough I/O technology to bypass the host network drivers and directly access the VFs.

On the data plane of the passthrough I/O, the data transmission is performed by the DMA engine in the NIC. DMA utilizes *Input/Output Memory Management Unit* (IOMMU) [15] to translate memory addresses and directly move packets between VF's TX/RX queues and microVM's TX/RX buffers.

The control plane of the passthrough I/O is managed by the hypervisor, which configures the VFs through the PF driver. To assign a VF to the guest, *i.e.,* microVM, the VF is first bound to a Linux driver named VFIO. The hypervisor interacts with the VFIO driver to configure the corresponding memory mapping to the IOMMU module. After the initialization is completed, the guest can directly interact with the device in subsequent data transmission, and only interrupt signals are relayed through the hypervisor. It should be noted that traditional containers typically do not require VFIO to access the VFs. Only when userspace network drivers need to directly manage the VFs, such as in DPDK [27] applications, do they require VFIO or alternative technologies like UIO. If VFIO is adopted, traditional containers suffer similar concurrent startup overhead as secure containers.

2.2 Address Spaces and DMA Memory Mapping

Fig. 3 shows the memory address spaces of the SR-IOV device, the host and the guest, in the context of passthrough I/O. We use the packet receiving process via a VF as an example to show how these address spaces are translated: (i) The guest OS notifies the DMA engine in the NIC to write the received packet to an *I/O Virtual Address* (IOVA). The IOVA is often chosen to be identical to the *Guest Physical Address* (GPA), where the guest OS intends to store the received packets, to simplify the mappings between IOVAs and GPAs. (ii) The DMA engine uses the IOMMU hardware to translate the IOVA to the corresponding *Host Physical Address* (HPA) and performs actual packet writes to the physical pages. The translation is implemented by looking up the *I/O Page Table*, which resides in the host memory and is maintained independently for each guest. As mentioned in §2.1, the table entries are configured by the VFIO driver when the VF is assigned to the guest, and the configuration process is referred to as *DMA memory mapping*. (iii) Upon completing the packet writes, the DMA engine notifies the guest OS that the data is ready by an interrupt relayed through the hypervisor. (iv) The guest OS retrieves the packet using the GPA, which is translated to the HPA by a hardware table called *Extended Page Table* (EPT).

2.3 Network Startup Procedure of SR-IOV Enabled Secure Containers

We investigate the source code of several widely deployed projects including the container orchestrator (Kubernetes [9]), container engine and runtime (Containerd [8], Kata [24]), CNI plugins (SR-IOV CNI [23] and sriovdp [22]), hypervisors (Kata-QEMU [24] and KVM [37]) and Linux kernel [38], and summarize the network startup procedure of SR-IOV enabled secure containers in Fig. 4.

Once the host OS has booted, the K8s agent, *i.e.,* Kubelet, calls the PF driver to pre-create a sufficient number of VFs. The pre-creation involves configuring the hardware of the SR-IOV enabled NIC and is often time-consuming. However, the pre-creation is a one-time task, since the VFs will be recycled when their assigned containers terminate. Therefore, we do not consider their time overhead in our discussion of the startup process in the rest of the paper. The life cycles of containers are managed by the container engine, *i.e.,* Containerd. When a container is invoked by Kubelet, Containerd first creates the isolated *Network NameSpace* (NNS) for each container and then successively calls the CNI plugin and the container runtime for VF configuration. The CNI plugin calls the PF driver to set up VF parameters, binds the VF to the host network driver and then moves the VF to the container NNS (*cf.* t_{config} in Fig. 4). The container runtime checks the existence of the VF in the NNS and assigns it to

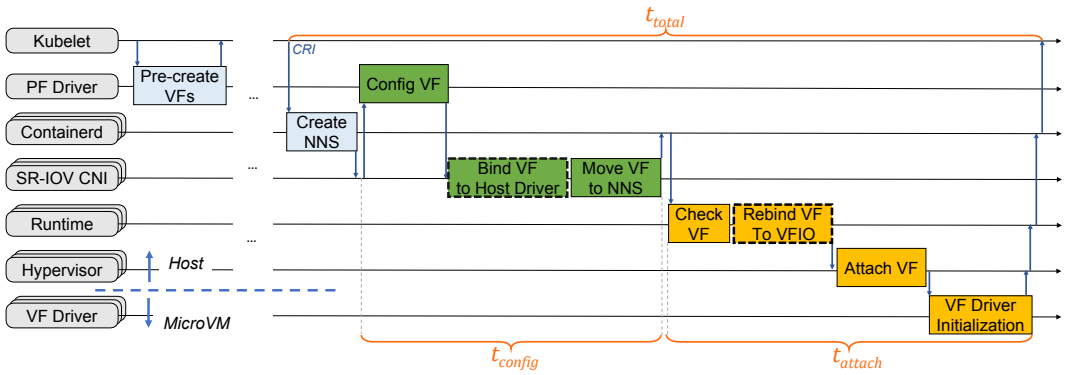


Figure 4. End-to-End network startup procedure of SR-IOV enabled secure containers.

the microVM (*cf.* t_{attach} in Fig. 4). In the assigning process, the container runtime first unbinds the VF from the host network driver and rebinds it to the VFIO driver. Then, the VFIO driver attaches the VF to the microVM, which includes setting up the passthrough I/O as introduced in §2.1 and emulating the VF as a PCIe device. Finally, the VF driver inside the microVM initializes and configures the device as a Linux network interface. As we will discuss in §5, it is unnecessary and extremely inefficient to keep binding the VFs to the host driver and then rebinding them to the VFIO driver, so we modify the implementation to pre-bind the VFs to the VFIO driver only once after they are pre-created, effectively removing the two binding/rebinding stages marked by dashed boxes in Fig. 4 from the startup process. We use the modified implementation in all our measurements and evaluations.

It should be noted that the underlying logic of configuring a VF and attaching it to the microVM is in fact the same as that of enabling SR-IOV for a normal VM. However, compared with the normal VM use case, container applications have higher-volume invocations and shorter lifespans, leading to higher requirements and new bottlenecks in startup time. This calls for further bottleneck identification and motivates our design for FastIOV.

3 Measurement and Motivation

3.1 Testbed for Startup Performance Measurement

Hardware setup. Our testbed uses servers that mirror the configurations used by major cloud providers' production environments. The specification includes: (i) CPU: Two NUMA-capable Intel Xeon Gold 6348 sockets running at 2.60 GHz, each housing 28 cores complemented by 80KB/1280KB/42MB L1/L2/L3 Caches and with hyper-threading activated. (ii) Memory: 256GB DDR4 with 3200MHz clock frequency. (iii) NIC: A 25 GbE Intel E810 NIC that supports creating 256 VFs. Note that we have also tested with another NIC, 200 GbE Intel Mount Evans E2100 and observed similar results.

Software setup. The servers run CentOS 7 with Linux kernel v6.4.0. We choose the widely deployed container engine Containerd v1.7.3 [8], secure container runtime Kata Containers v3.2.0 [24], and SR-IOV CNI plugin v0.3 [23]. Like Firecracker [1] and RunD [35], Kata is a microVM-based secure container framework. MicroVM-based secure containers can prevent exposing security-critical syscalls of the host kernel to untrusted user code, and thus are preferred over other security solutions that augment traditional containers with kernel-level isolation mechanisms, such as AppArmor [52] and Seccomp [53]. Compared with unikernel-based solutions such as LightVM [45] and Solo5 [65], microVM-based secure containers ensure full user code compatibility, making them more applicable in cloud services. According to our measurements in a concurrent work [42], Kata has the second fastest startup among all mainstream open-source microVM-based secure container frameworks, surpassed only by RunD. The open-source version of Firecracker is much slower. We did not use RunD because its open-source version crashes when startup concurrency exceeds approximately 100 instances. Kata Containers tailor the QEMU v6.2.0 hypervisor into a lightweight version named Kata-QEMU [24]. The kernel of the microVM is generated from Linux kernel v5.19.2, and the image is generated from Ubuntu 20.04. For each secure container, we allocate 0.5 vCPU and 512MB RAM through the Kata-QEMU configuration, with 2MB sized hugepages enabled. We allocate one SR-IOV VF as the container's virtual NIC through the Containerd configuration.

Measurement methodology. In the startup time tests, we use *crictl* command to concurrently create 200 microVMs without any container applications inside, as enabling SR-IOV only affects the startup process of the microVM. The concurrency setting of 200 is based on the statistics collected from the production environment of Alibaba serverless platform [35]. When we evaluate the performance of FastIOV on serverless applications in §6.6, we will report the task completion time, *i.e.,* the duration between the issuance of the startup command and the completion of the container application. To break down the timeline of the startup process, we develop a logging tool and integrate it into the above software components, such as Kata-QEMU and Linux kernel,

to collect more fine-grained information. We ensure that the logging operations are asynchronous and our tests show that they incur almost no additional overhead in startup time.

3.2 Startup Bottleneck Identification for SR-IOV Enabled Secure Containers

3.2.1 Measurement result. We break down the timeline of the concurrent startup of 200 secure containers, and show the most time-consuming steps in Fig. 5, where each horizontal line corresponds to a different container, and different colors mark different time-consuming steps. The statistics are summarized in Tab. 1. We observe that *4-vfio-dev*, *i.e.,* the opening of the VF from its VFIO device set, dominates the total time consumption, possibly experiencing severe serialized operations. Other SR-IOV VF-related steps like DMA memory mapping (*1-dma-ram* and *3-dma-image*, which represent the mapping of the microVM RAM and system image memory, respectively) and the initialization of network interface by the VF driver inside the microVM (*5-vf-driver*) also incur obvious overhead. The remaining two steps, *i.e., 0-cgroup* and *2-virtioFS*, which refer to the initialization of the cgroups and the shared file system for the secure container, are not related to enabling SR-IOV. Note that the VF-related steps account for 70.1% of the average startup time. The percentage increases to 80.8% when considering the long-tail latency with the 99th percentile. These statistics reveal a great potential to accelerate the startup process. Next, we will analyze the root causes of the observed major bottlenecks before introducing our solutions in FastIOV.

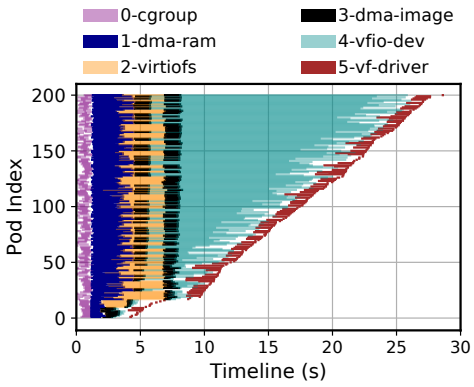


Figure 5. Breakdown of time-consuming steps[2]**.** 200 SR-IOV enabled secure containers are launched concurrently.

Step	Proportion in Average Time (%)	Proportion in 99th Percentile Time (%)
0-cgroup	2.9	2.3
1-dma-ram	**13.0**	**11.1**
2-virtiofs	13.3	13.6
3-dma-image	**5.6**	**4.3**
4-vfio-dev	**48.1**	**59.0**
5-vf-driver	**3.4**	**4.1**
Total (1, 3, 4, 5)	70.1	80.8

Table 1. Time proportions of time-consuming steps. The VF-related steps (1, 3, 4, 5) account for over 70% and 80% of the average and 99th percentile startup time.

3.2.2 Bottleneck 1: VFIO devset management. In the VFIO driver, a devset is used to manage a group of VFIO devices and control their reset behavior. A VFIO device that supports slot-level device reset forms a devset on its own. It can be reset without affecting others. Other VFIO devices require bus-level reset, *i.e.,* all devices attached to the same bus are reset together. All such VFIO devices attached to the same bus form a devset. When such a VFIO device is to be reset, the VFIO driver ensures that all other affected devices are also ready for reset. More specifically, it scans the PCI bus to ensure that all devices on the same bus belong to the current devset. It also checks the total open count of the devset, *i.e.,* the number of processes or threads that currently keep the devices open, to ensure that no affected device is currently being used.

As far as we know, slot-level reset capability is uncommon on modern NICs. For instance, it is not supported by the widely used Intel E810 NIC [28] or the recent Intel IPU E2100 NIC [29]. Therefore, the VFs are typically maintained in one devset with other VFIO devices on the same bus. When attaching a VF to the microVM, one of the key steps is to register it in the hypervisor. During registration, the hypervisor opens the VF through the VFIO driver, and obtains the file descriptor and other relevant device information. The opening of the VF increases its open count, and further affects the global state, *i.e.,* the total open count, of the devset. To ensure the correctness of the states, the current design of the VFIO driver utilizes one global mutex lock to make the operations on the VFIO device and the operations involving checking or updating the global state of the devset mutually exclusive. However, **such coarse-grained mutex lock also serializes the opening operations on the different VFs belonging to the same devset**, and thus hinders the concurrent startup process of SR-IOV enabled secure containers. This accounts for the nearly linear increase in the time cost of *4-vfio-dev* observed in Fig. 5.

3.2.3 Bottleneck 2: DMA memory mapping. Apart from the registration of the VFIO device, another key step in attaching a VF is the DMA memory mapping. As introduced in §2.1, the hypervisor meticulously configures the IOMMU to establish the mapping for the microVM's memory, ensuring

[2]Note that the fastest container takes 3.8s for its startup, due to the SR-IOV enabled networking and the severe contention at high concurrency. The milliseconds startup reported in related work [1, 24, 35] is measured at low concurrency and without networks, which is also consistent with our measurement in Fig. 1, where the fastest container startup takes 460ms at a low concurrency of 10 and without network.

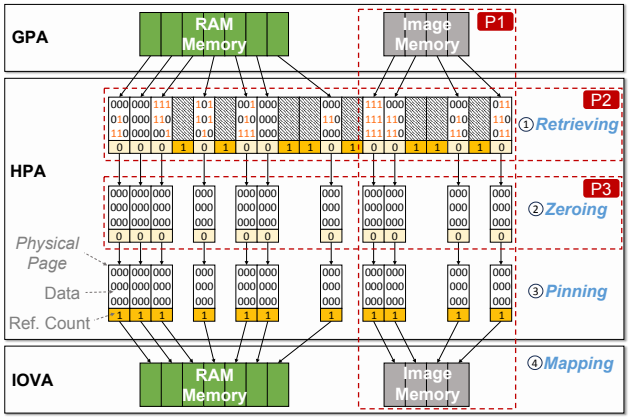


Figure 6. DMA mapping procedure. Four main steps (page retrieving, zeroing, pinning, mapping) and 3 sub-bottlenecks.

that the DMA data transmission operations can be correctly performed by the NIC. The DMA memory mapping process can be summarized as three major steps: First, the physical memory for the microVM is allocated in the host to obtain the corresponding HPA. Then, the allocated physical memory is pinned in the system to keep it from being swapped out, so that the corresponding HPA remains effective. Finally, the mapping between HPA and IOVA is configured in the page table in IOMMU. We further illustrate the steps in Fig. 6 and analyze the cause of overhead. In the figure, *retrieving* and *zeroing* correspond to the first major step, *pinning* and *mapping* correspond to the other two steps, respectively.

- Page retrieving: When allocating physical memory for the DMA memory, the VFIO driver iteratively collects free physical pages until the requested total size is satisfied.
- Page zeroing: Free pages can contain residual data, which might lead to potential security issues in multi-tenant clouds. Thus, the current physical memory allocation implementation ensures that these retrieved physical pages are filled with zeros to clear any sensitive information before they are returned to the VFIO driver.
- Page pinning: Once all free pages are retrieved and zeroed, they are pinned by the VFIO driver: their reference counts are increased to prevent them from being moved or swapped out by the OS. This ensures that the HPAs of the pages remain effective during DMA operations. The mappings between the HPAs and GPAs are then generated and maintained by the hypervisor.
- Page mapping: Then, the IOMMU's page table is updated to set up the *IOVA-HPA* mapping between the virtual addresses *IOVA*, which the device uses for DMA operations, and the *HPA* of the pinned physical pages.

During the profiling of the DMA memory mapping process, we find the following three key factors that make DMA memory mapping a bottleneck in the startup process.

First, **there exists unnecessary DMA memory mapping in the microVM (P1 in Fig. 6).** The original design of the VFIO driver and IOMMU performs DMA mapping for all regions in the memory space of the microVM, as they assume that all the regions have the possibility of being accessed by DMA. However, we identify that the mapping of the microVM image memory region is unnecessary. The image contains the system files of the microVM and a secure container agent procedure used for managing container applications. Its region is read-only and invisible to the container applications that launch DMA operations. In our measurement setup, the microVM image uses 256MB of memory, and Tab. 1 shows that constructing the memory mapping of this region constitutes 5.6% (*3-dma-image*) of the total time cost, but the cost is avoidable.

Second, **fragmented small physical pages incur high retrieval costs (P2 in Fig. 6).** When the VFIO driver iteratively collects free physical pages, the free pages with continuous HPAs will be grouped together and operated as a batch to reduce the time overhead caused by excessive function calls. When physical pages experience more fragmentation, fewer pages will be batched, resulting in higher retrieval costs. However, we find that such overhead is already effectively mitigated by simply enabling hugepages, a common practice in the production environment, as it significantly reduces the number of pages to retrieve. Thus this cause of bottleneck is not a focus of our optimization.

Third, **page zeroing incurs a significant time cost (P3 in Fig. 6).** After reducing the retrieval cost by enabling hugepages, we find that page zeroing contributes to over 93% of the total DMA memory mapping time. Such time cost is not caused by any lock contention but purely by zeroing operations. When SR-IOV is not enabled, no DMA memory mapping is performed, which allows the allocation of physical pages to be deferred until the application accesses the memory and triggers a page fault. As a consequence, a page is zeroed only when it is read or written. We refer to this technique as *lazy zeroing*, which avoids the zeroing overhead during startup and precludes unnecessary zeroing of unused memory. As IOMMU cannot handle page faults during DMA operations, all physical pages need to be allocated in advance when SR-IOV is enabled, as mentioned in §3.2.3. Therefore, lazy zeroing based on on-demand page allocation no longer works. Our key observation is that page zeroing can be decoupled from physical page allocation in DMA memory mapping, which allows for lazy zeroing for SR-IOV enabled secure containers and motivates our design.

3.2.4 Bottleneck 3: VF driver initialization.

After the VFIO driver configures the VF and hands it over to the microVM, a two-step initialization proceeds to configure the VF as a Linux network interface inside the microVM. First, the VF driver (NIC driver) inside the microVM conducts PCI device enumeration to identify the device, registers the device

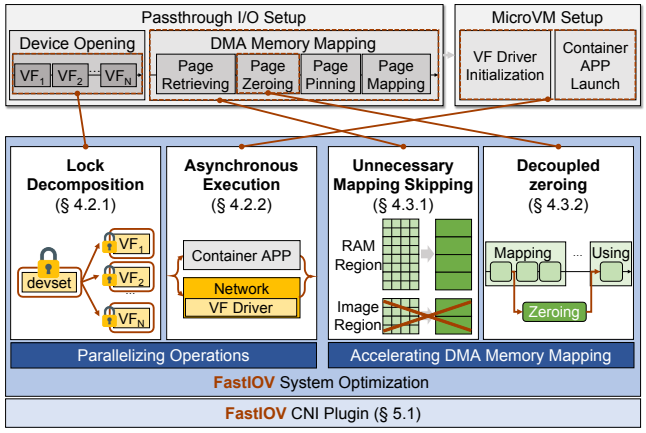
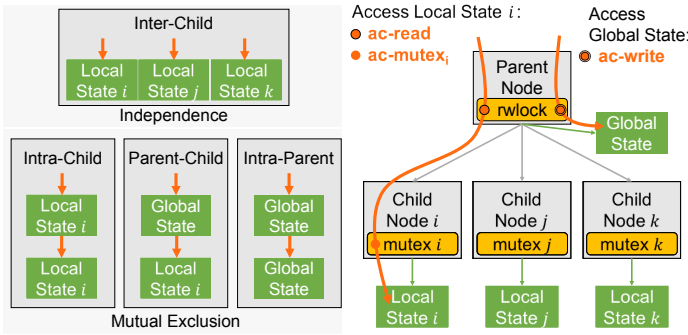


Figure 7. FastIOV Overview.



(a) Independent and mutually exclusive relations.

(b) Implementing parent-child lock with *rw-lock* and *mutex*.

Figure 8. Lock decomposition with parent-child lock

as a network interface, configures its network parameters, and updates its link status. Second, the daemon agent of the secure container framework inside the microVM assigns MAC and IP addresses to the interface. It takes a few hundred milliseconds up to seconds for all these operations to complete, and then the interface becomes available. This time cost further increases with the container concurrency. As secure container frameworks manage the initialization and other setup procedures of the microVM in a serial fashion, it only executes the subsequent setups after the interface becomes available, causing non-negligible overhead to the startup performance. Our design will show that such overhead can be effectively mitigated with asynchronous execution.

4 Design

4.1 FastIOV Overview

Fig. 7 displays the key components of FastIOV, including lock decomposition, unnecessary mapping skipping, decoupled zeroing and asynchronous execution. The four optimizations aim at addressing the bottlenecks analyzed in §3.2 to speed up the concurrent startup process. The main workflow of FastIOV is as follows.

When SR-IOV enabled secure containers are launched, VFs are attached to microVMs concurrently. First, FastIOV decomposes the coarse-grained lock in VFIO devset management using a hierarchical lock framework (§4.2.1). By doing so, FastIOV parallelizes the device opening operations during VF registration while maintaining the state correctness of VFIO devset. Then, when the VFIO driver performs DMA memory mapping for the VFs, FastIOV tracks and skips the unnecessary mapping region, *i.e.,* microVM image memory (§4.3.1). As for the remaining regions, FastIOV decouples page zeroing from physical memory allocation to enable lazy zeroing (§4.3.2), which avoids the zeroing time overhead during startup as well as the zeroing of unused memory. Finally, FastIOV asynchronously executes the initialization of the VF driver inside the microVM. FastIOV overlaps the initialization with the launch of the container application to mask the overhead (§4.2.2).

Next, we will introduce the optimizations in detail. As lock decomposition and asynchronous execution are both aimed at parallelizing operations for speedup, we put them in the same category and introduce them first.

4.2 Parallelizing Operations

4.2.1 Lock decomposition in VFIO devset. The coarse lock problem in VFIO devsets can be abstracted as follows. A devset acts as a parent node and the VFIO devices belonging to it act as child nodes. The parent node has a global state that is related to the local states of its children. The current design of VIFO driver implements only a global mutex lock for the entire devset, so it requires the contention of the same mutex lock whether it is to access the global state of the parent or the local state of a child. When a heavy contention occurs in inter-child operations, *e.g.,* concurrently opening multiple VFs, the system parallelism degrades significantly. On the other hand, simply removing the global mutex lock will compromise the state consistency in the multi-thread accessing procedure. Our insight is that we can *decompose the lock to enable independent inter-child operations and hence improve the startup performance, while keeping other operations mutually exclusive to ensure consistency.*

We distinguish four types of relations between operations according to the data they access: (i) *inter-child operations* access the local states of different child nodes, (ii) *intra-child operations* access the local state of the same child node, (iii) *intra-parent operations* access the global state of the parent node, (iv) *parent-child operations* access the global state of the parent node and the local states of a child node, respectively. As shown in Fig. 8a, inter-child operations are independent and can be performed in parallel, while operations of the other three types should be mutually exclusive and performed in serial.

To achieve the above requirements, we propose a hierarchical lock decomposition framework built on two Linux kernel locks, *read/write lock* (*rwlock*) and *mutex*, as shown in Fig. 8b. In this framework, the parent node is equipped with a global *rwlock* and each child node i is equipped with a local $mutex_i$. When accessing the global state, one needs to acquire the *rwlock* write permission (denoted by *ac-write*). When accessing the i-th local state, one needs to acquire both the *rwlock* read permission (denoted by *ac-read*) and $mutex_i$ (denoted by $ac\text{-}mutex_i$).

We can show that the proposed lock decomposition framework indeed satisfies the requirements. Here we consider the case of inter-child operations. The other cases can be shown in a similar fashion and omitted. Suppose two inter-child operations on local state i and local state j occur concurrently. Since two *ac-read*s are independent according to the definition of *rwlock*, and $ac\text{-}mutex_i$ and $ac\text{-}mutex_j$ are naturally independent, these operations can be executed in parallel.

Although inventing a new Linux kernel lock can also satisfy the requirements, we believe that reusing off-the-shelf kernel locks keeps the design simple and ensures effectiveness. Moreover, we believe this lock decomposition framework can be promoted to other scenarios rather than just being used in the VFIO devset.

4.2.2 Asynchronous execution in VF driver initialization. We make two observations regarding the initialization process of the VF driver, where the network driver inside the microVM initializes and sets up the VF as a Linux network interface. First, the network interface is not utilized until the container application is launched and begins execution inside the microVM. Second, the initialization of the network interface is independent of the other startup stages. This allows the initialization to be executed asynchronously and in parallel with other stages, in particular the launching of container application in the microVM. The launching process involves transferring container images from the host to the microVM via the shared file system and creating the container process. Our empirical measurements show that with a high container concurrency of 200, this process can span several seconds, which is enough to mask the initialization time. We adapt the secure container framework to initialize the network interface asynchronously and employ the framework's daemon agent inside the microVM to periodically check the status of the network interface, ensuring the network is available as the application begins execution.

4.3 Accelerating DMA Memory Mapping

4.3.1 Skipping unnecessary mapping region. FastIOV tracks and skips the unnecessary DMA memory mapping, *i.e.,* microVM image memory, to reduce overhead. Before the hypervisor, *e.g.,* QEMU, enumerates the DMA memory regions and calls the VFIO driver to perform DMA memory mapping, FastIOV notifies the hypervisor of the information of the image memory region, *i.e.,* its name and size. The hypervisor then skips DMA memory mapping for this region and falls back into its non-DMA memory managing logic.

4.3.2 Decoupling zeroing from mapping. For the remaining regions that are not skipped, *i.e.,* the RAM of the microVM, FastIOV decouples the page zeroing operation from physical memory allocation to enable lazy zeroing. Recall that lazy zeroing means that the physical pages are zeroed only when they are actually read or written. The high-level idea is to intercept the memory access to physical pages conducted by the microVM, and perform page zeroing when the page is read or written for the first time. We identify three key challenges in achieving this goal.

- First, when a microVM accesses a physical page, it bypasses the hypervisor and relies instead on the hardware-assisted module EPT (previously introduced in §2.1) for address translation. How can we intercept this process and zero the physical pages before their usage?
- Second, if we intercept every memory access to check whether the physical pages are accessed for the first time, it will be very costly and significantly degrade memory performance. How can we avoid such overhead?
- Third, there exist exceptions where the first memory access to a physical page is not conducted by the microVM. Specifically, the hypervisor may write to the physical pages before starting the microVM, and the para-virtualization components like the shared file system, *i.e., virtioFS*, may write to the physical pages before the microVM reads from them. In such cases, the relevant physical pages should be zeroed before being used by the hypervisor or para-virtualization components, and require no further zeroing before the first access by the microVM. How do we deal with such exceptions to ensure the correctness of zeroing?

The rest of §4.3.2 presents detailed designs and shows how FastIOV solves the above problems.

EPT fault based memory access interception and lazy zeroing. After digging into the details of the EPT address translation mechanism, we find that the entries in the EPT are constructed by an *EPT fault* right before the corresponding physical pages are read or written for the first time by the microVM. The EPT fault carries the information of the accessed physical pages and is perceived by KVM, a hypervisor module. This gives us the opportunity to intercept the information and perform lazy zeroing. Recall that when the microVM is launched, the VFIO driver performs DMA memory mapping, which allocates physical memory for the microVM. As shown in Fig. 9, the physical memory allocation generates the HVA-HPA mapping in *Memory Management Unit* (MMU) of the host (①). Also, during the launch of the microVM, KVM maintains the GPA-HVA mapping (②). When the microVM accesses a GPA for the first time, it looks it up in the EPT, only to find that there is no matching entry (③). Then

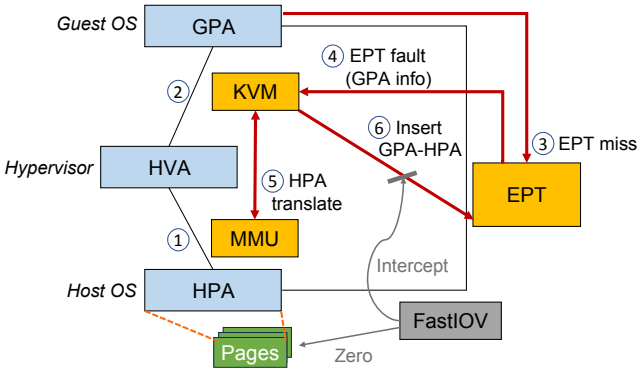


Figure 9. EPT fault based interception and lazy zeroing.

the microVM triggers an EPT fault, which sends KVM an EPT violation signal containing the GPA information (④). KVM then translates the GPA to HVA, and utilizes the MMU to translate the HVA to HPA (⑤). Finally, KVM inserts the GPA-HPA mapping entry into the EPT (⑥), which is now ready for use by the microVM.

By intercepting the HPA information in the KVM, we can perform lazy zeroing during the EPT fault for the corresponding physical page. The EPT fault is triggered solely upon the microVM's initial access to a physical page, ensuring that subsequent accesses to the same page bypass interception, thereby significantly reducing the impact on memory performance. Our evaluation in §6.5 will show that the overhead incurred is negligible.

Ensuring the correctness of lazy zeroing. We identify that there are two exceptional scenarios where a physical page requires no further zeroing before the first access by the microVM.

- **Hypervisor data write**. Before launching a microVM, the hypervisor writes to the memory allocated to it in order to perform the necessary setup, including loading read-only regions like BIOS and kernel into the memory. Such writes are performed directly without involving the EPT. After launching, when the microVM tries to access these memory regions for the first time, *e.g.,* to execute kernel code, it will trigger an EPT fault and cause FastIOV to incorrectly zero the data written by the hypervisor, leading to a system crash.
- **Para-virtualization based data transfer**. Devices can utilize para-virtualization protocols, such as the widely used *virtio* protocol, to exchange data between the microVM and the host through shared buffers. A typical example is `virtioFS`, which is a shared file system that allows the container inside the microVM to access the designated files on the host. When the microVM reads a file, it first writes the addresses of the file and a shared buffer into a `vring`, which is itself a shared buffer. The backend of *virtio* on the host fetches the addresses from the `vring`, writes the file data into the shared buffer, and notifies the microVM to read it. If the microVM has not accessed the buffer memory before, the read operation will trigger an EPT fault, which will cause FastIOV to incorrectly zero the requested file data before reading it.

To ensure the correctness of lazy zeroing, FastIOV tackles the above two problems by maintaining an *instant zeroing list* and triggering *proactive EPT faults*, respectively.

The *instant zeroing list* is a white list of physical pages that are not managed by FastIOV and are zeroed instantly when allocated. Read-only memory regions, such as the BIOS and kernel memory, are determined before the start of the microVM, and the hypervisor registers them to the *instant zeroing list* maintained by FastIOV. The exclusion of those regions from its management may limit the gain of FastIOV. However, our test shows that with a normal Linux kernel, those regions take up only about 9.4% of the total memory for a microVM with 512MB of memory. The percentage decreases with a larger allocated memory, as the size of the excluded regions remains fixed. Thus, FastIOV can still effectively reduce DMA memory mapping time by optimizing the page zeroing of the remaining regions.

To address the exception caused by para-virtualization based data transfer, FastIOV proactively triggers EPT faults when the microVM writes the address of a shared buffer to `vring`, so that FastIOV correctly clears the corresponding physical pages before the backend of *virtio* on the host writes the file data back into the buffer. Such proactive EPT faults are triggered by performing a data read to the first byte of each page of the buffer.

Theoretically, there is a third exception scenario where a physical page requires no further zeroing before the first access by the microVM, which is the *NIC DMA data write*. Similar to *virtio*, the DMA data exchange between the NIC and the microVM is through a ring buffer managed by the microVM NIC driver. If the microVM has not accessed the buffer memory before the NIC writes data to it, EPT faults will be triggered when the microVM tries to read the buffer, which incorrectly zero the data. However, the NIC being the first to write the DMA buffer never truly happens when applications use VFs through standard NIC drivers, as the drivers themselves zero the DMA buffer immediately after allocation, which will have already triggered EPT faults. This can actually be made more efficient by changing the zeroing operation to performing a data read to only the first byte of each page of the buffer like what we do with *virtio*.

5 Implementation

The implementation of FastIOV includes a portable Linux kernel module named `fastiovd`, a FastIOV CNI plugin, and several modifications in the hypervisors, container frameworks and host/guest kernel modules. Fig. 10 illustrates their detailed functionalities and the statistics of *Lines of Code*

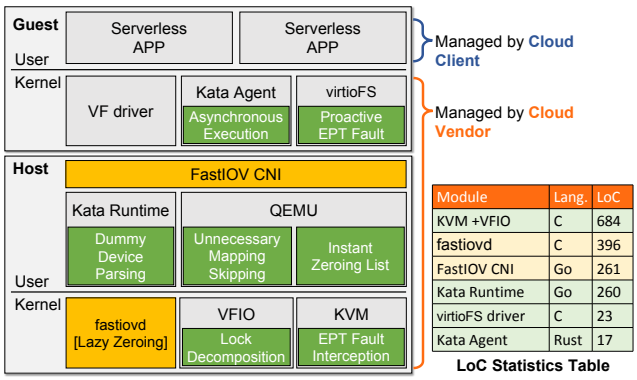


Figure 10. Implementation of FastIOV.

(LoC). Note that FastIOV is deployable because all of those new modules and modifications are within the management of the cloud vendor. Several details are presented below.

Lazy zeroing implementation. First, we disable the original page zeroing operation in the VFIO kernel module, and maintain in `fastiovd` a two-tier hash table containing the information of the physical pages to be lazy zeroed. The first-tier key uses the process ID (PID) as the independent identifier for each microVM, and its value is the pointer to the secondary hash table maintained for that microVM. The second-tier key is the HPA and its value contains detailed page information. Second, we modify the KVM module to trigger lazy zeroing before it inserts the EPT entry during an EPT fault. The KVM notifies `fastiovd` of the page triggering the EPT fault. If it is in the two-tier hash table, `fastiovd` will zero the page, remove it from the hash table, and notify KVM upon completion. Besides the above lazy zeroing logic, we also maintain a background thread in `fastiovd`, which periodically scans the two-tier hash table, zeroes the remaining pages, and then removes them from the table. Such background clearing in fact overlaps the zeroing with other startup stages to reduce the EPT fault time to further improve container application performance.

FastIOV CNI plugin implementation. The vanilla SR-IOV CNI plugin [23] shown in Fig. 4 is designed for traditional containers. It causes a VF to bind to the host network driver and then rebind to the VFIO driver every time a secure container is launched, which incurs high overhead. We find that the only reason for binding a VF to the host network driver when launching a secure container is to generate a Linux network interface, which serves two functions. First, the Kata runtime identifies the VF by detecting the interface. Second, the CNI performs network operations like IP configurations on the interface, which then passes the configurations to the Kata runtime when it is detected. Therefore, to free VFs from binding to host network drivers, we create dummy Linux network interfaces to fulfill the above two functions instead. This allows us to bind the VF to the VFIO driver only once after the server's booting. This simple optimization greatly reduces the startup time of vanilla SR-IOV CNI, from several minutes to 16.2 seconds when concurrently starting 200 containers. However, we regard this binding problem as an implementation drawback that may depend on the specific container framework, Kata containers in this case. We apply the above optimization to the vanilla SR-IOV CNI in our evaluation for a fair comparison.

6 Evaluation

6.1 Experimental Setup

Testbed setup. We conduct two categories of experiments that evaluate FastIOV's network startup performance and overall performance with serverless application benchmarks, respectively. The former runs on a single test server, while the latter on two directly connected test servers acting as the application server and the storage server, respectively. All test servers mentioned above have the same hardware and software configurations as specified in §3.1.

Baselines. We compare FastIOV with the following baselines to validate the effectiveness of our designs.

- ***No network***: The startup without enabling network. This represents a lower bound for optimizing network startup.
- ***Vanilla***: The original implementation of SR-IOV CNI [23] without optimization for passthrough I/O but with the fix for the implementation drawback for a fair comparison.
- ***FastIOV* variants**: In order to evaluate the effectiveness of each of our four optimization designs, *i.e.*, **L**ock decomposition, **A**synchronous execution, unnecessary mapping **S**kipping and **D**ecoupled zeroing, we remove them from FastIOV one at a time and get four variants named *FastIOV-L*, *FastIOV-A*, *FastIOV-S* and *FastIOV-D*, respectively. A larger performance degradation of a variant compared with *FastIOV* demonstrates a higher effectiveness of the removed optimization design.
- **Memory pre-zeroing methods**: Memory **Pre**-zeroing is a popular technique proposed by HawkEye [55] that performs page zeroing during memory idle time to achieve faster page faults. It has also been utilized by the open-source community to speed up DMA memory mapping and accelerate the booting of passthrough I/O enabled VMs. The performance of this baseline is affected by the fraction of memory pre-zeroed during memory idle time. To evaluate its performance across different scenarios, we set the fraction to 10%, 50% and 100%, and represent them by *Pre10*, *Pre50*, and *Pre100*, respectively.
- **Software CNI**: Besides the SR-IOV baselines, we also compare FastIOV to a software CNI in §6.4 aiming at illustrating the bottleneck differences between the two types. We choose the basic software CNI *IPvtap*, because (i) it shares similar virtual network device implementation with popular software CNIs like Flannel [12] and Calico [10], but has faster startup due to its lack of support for more

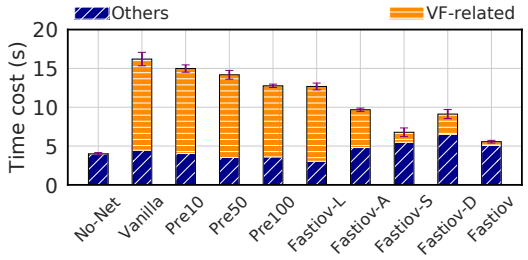


Figure 11. Average startup time. Concurrency = 200

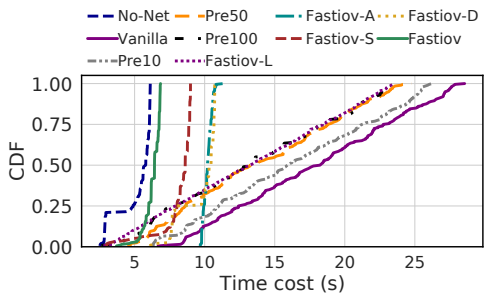


Figure 12. Startup time distribution. Concurrency = 200

advanced network features; and (ii) it is the basic software CNI with the fastest startup for secure containers according to our measurements in a concurrent work [42].

6.2 Startup Performance

We compare *FastIOV* with other baselines by measuring the startup time with 200 concurrently invoked secure containers. Fig. 11 displays the average time and the breakdown into two parts, *VF-related* and *others*. *VF-related* refers to the time of the four VF-related stages previously introduced in §3.2 and *others* represents the remaining part of the time. We draw the following three key conclusions from the results.

First, *FastIOV* significantly outperforms the vanilla SR-IOV CNI in both the average and long-tail time cost. *FastIOV* reduces the average startup time by 65.7% compared to *vanilla*. Specifically, *FastIOV* reduces the time overhead directly related to VF operations by 96.1%, significantly mitigating the effect of enabling VF on the startup of secure containers. Moreover, the time distribution in Fig. 12 shows that *FastIOV* also reduces the 99th percentile startup time of *vanilla* by 75.4%, largely improving the long-tail performance. In addition, FastIOV achieves a startup time close to that of *No-Net*, with the average and the 99th percentile startup time being 39.1% and 11.6% higher, respectively. In contrast, the corresponding figures of *vanilla* are substantially larger, *i.e.,* 305.2% and 354.5%.

Second, each of our optimization techniques makes an obvious contribution to the time reduction achieved by FastIOV. Compared with *Vanilla*, *FastIOV-L*, *FastIOV-A*, *FastIOV-S* and *FastIOV-D* reduce the average time by 21.8%, 40.3%, 58.2%, and 43.7%, respectively, all smaller than the 65.7% reduction achieved by the intact FastIOV. This shows that removing any of the optimization techniques obviously degrades the gain, thereby proving the effectiveness of each optimization.

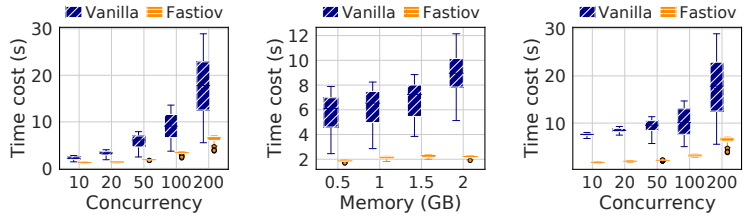
Third, FastIOV outperforms the memory pre-zeroing methods and further reduces the average time by 56.4% compared with *Pre100*. The performance of pre-zeroing strongly depends on the fraction of memory pre-zeroed during memory idle time. In practice, cloud vendors tend to maintain a high level of memory utilization for more revenue. For example, AWS uses a bin-pack algorithm for this purpose, and its average server memory utilization ranges from 84.6% to 100%, with a median of 96.2% [63]. This leaves short memory idle time and can further limit the performance of pre-zeroing.

6.3 Impacting Factors

Concurrency. Fig. 13a shows the impact of varying concurrency. It reports the startup time distribution with container concurrency increasing from 10 to 200 and each allocated 512MB of memory. We observe that FastIOV is effective across all concurrency, achieving time reductions ranging from 46.7% to 65.6%. The reduction is more obvious with a higher concurrency, as the lock contention in VFIO devset becomes more severe with more concurrently invoked VFs.

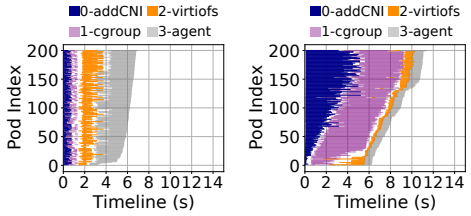
Resource allocation. Fig. 13b shows the impact of varying per-container resource requirement. More precisely, it shows the time distribution of concurrently starting 50 containers with memory allocation for each container increasing from 512MB to 2GB. We observe an obvious increase of 60.5% in the average startup time of *vanilla* as the memory allocation increases to 2GB, while only 21.5% with FastIOV. This is because the optimization of *FastIOV* on DMA memory mapping makes its startup time less sensitive to allocated memory. With a larger memory allocation, the effectiveness of the unnecessary mapping skipping optimization decreases, as the ratio of the skipped memory regions becomes relatively smaller. However, the overall reduction ratio achieved by *FastIOV* still increases. This is because the skipping optimization is not the primary source of the achieved speedup, as shown by the comparison of *FastIOV* and *FastIOV-S* in Fig. 11. A larger memory allocation increases the performance gain achieved by *FastIOV*'s optimization on DMA memory mapping, which outweighs the decreased gain of the skipping optimization, and thus leads to a higher reduction ratio.

Fully loaded server. As mentioned before, cloud vendors like AWS tend to schedule containers to maximize the utilization of server resources, *i.e.,* memory and CPU. Here, we consider a scenario that tries to partially capture this



(a) Varying concurrency. **(b)** Varying resource. **(c)** Fully loaded server.

Figure 13. Impacting factors.



(a) FastIOV **(b)** IPvtap

Figure 14. Comparison with software CNI.

behavior. We vary the concurrency, and for each given concurrency, we evenly divide all the resources of the server among the concurrent containers. Note that fewer containers means more allocated resources for each. The startup time distribution in Fig. 13c shows that FastIOV achieves large time reductions across all settings, even with low concurrency. In fact there is an increase in the time reduction ratio, from 65.7% to 79.5% as the concurrency decreases from 200 to 10. This is because a lower concurrency reduces the time of the other startup steps unrelated to SR-IOV, while the optimization of DMA memory mapping is unaffected, as the total allocated memory stays unchanged.

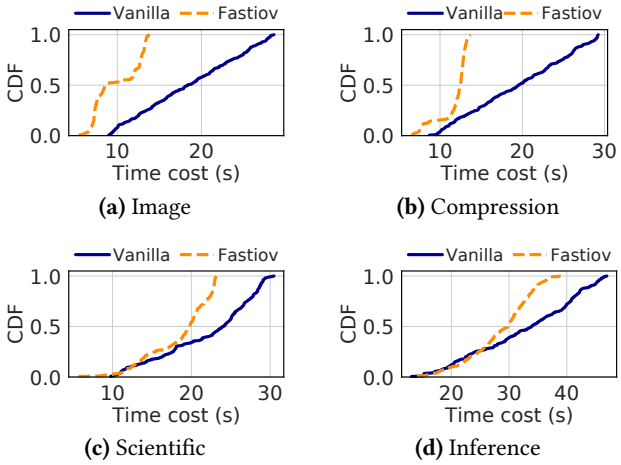
6.4 Bottleneck Differences with Software CNI

We compare FastIOV with the software CNI *IPvtap* to illustrate how the startup bottlenecks of a software CNI differ from those of an SR-IOV based solution. The software CNI emulates the physical network devices of microVMs, and thus obviates the time-consuming passthrough I/O setup procedure. A comparison of Fig. 11 and Fig. 14 shows that *IPvtap* has faster startup than *vanilla SR-IOV*, although with a much worse data plane performance. On the other hand, Fig. 14 shows that *FastIOV* achieves 41.3% and 31.8% lower total and average startup time than *IPvtap*.

The deficiency of *IPvtap* results mostly from two parts: (i) the creation and configuration of the virtual network device (denoted by *addCNI*), and (ii) the host resource isolation (denoted by *cgroup*). Through detailed measurements, we identify that the severe lock contentions in kernel network calls and cgroup operations bring in much overhead [42]. In contrast, SR-IOV CNIs attach VFs to the secure container without creating any additional virtual network device. Thus, with FastIOV optimizing the time-consuming passthrough I/O setup, an SR-IOV based solution is more capable of achieving ultra-fast concurrent startup for secure containers.

6.5 Impact on Memory Access Performance

To evaluate the effect of FastIOV on the memory access performance, we use an open-source tool *Tinymembench* [59] to test the memory throughput and latency within the secure container. To obtain the throughput, the tool performs *memcpy* operations on 2048-byte data blocks for 5 seconds



(a) Image **(b)** Compression **(c)** Scientific **(d)** Inference

Figure 15. Serverless application performance.

and repeats the process for 10 times. To obtain the latency, it performs the random byte reading for 10 million times. The results show that *FastIOV* achieves memory access performance comparable to *vanilla*, with a degradation in memory throughput and an increase in latency within 1%. Since *FastIOV* only intercepts the EPT page fault once upon the initial memory access, it does not affect subsequent memory operations and thus causes negligible performance degradation.

6.6 Performance in Serverless Applications

Benchmark applications. To evaluate the overall speedup brought by FastIOV on serverless applications, we choose four representative tasks, *i.e., Image*, *Compression*, *Scientific* and *Inference*, from the widely adopted SeBS [14] serverless benchmark. *Image* resizes an input image to a thumbnail of size 100x100. *Compression* zips an input file of 9.7MB. *Scientific* performs a breadth-first search to traverse a graph of 100000 nodes. *Inference* utilizes ResNet-50 model for ImageNet classification task. Each application first downloads input data from the storage server through the VF assigned to its secure container before performing the computation.

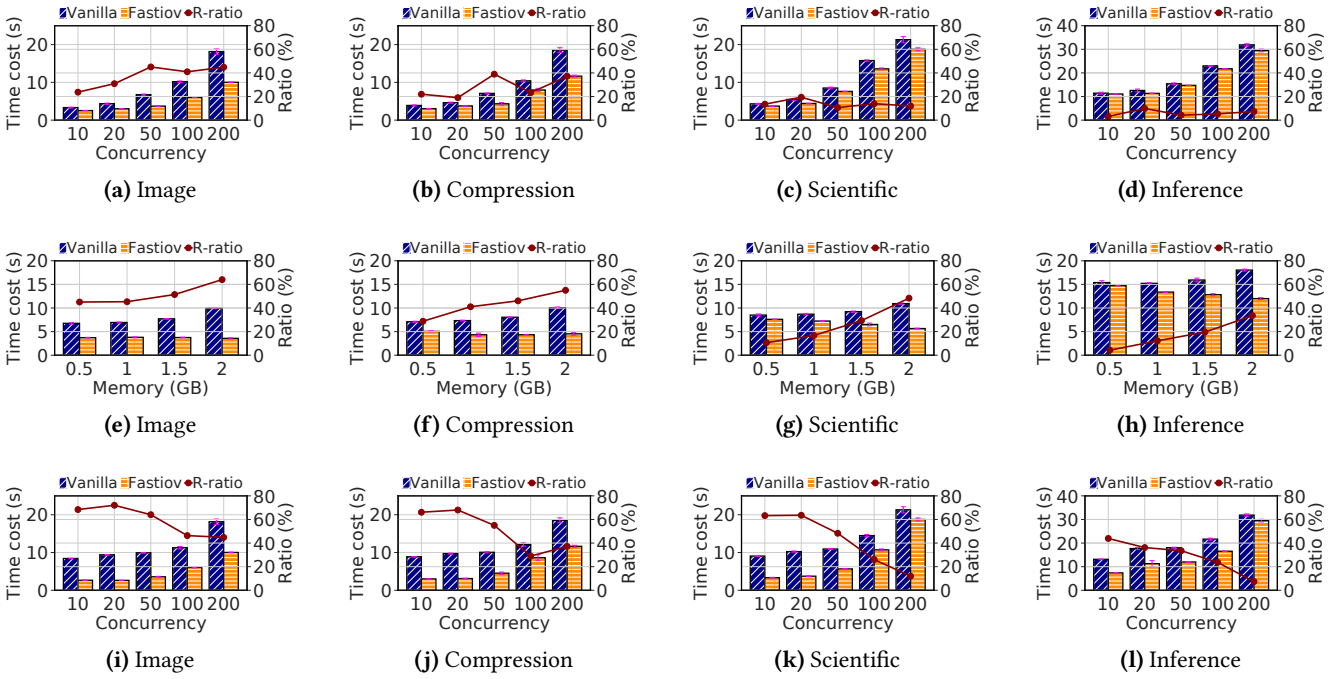


Figure 16. Performance of FastIOV on serverless applications with varying concurrency (a~d), varying resource allocation (e~h), or fully loaded server (i~l). *R-ratio*: time reduction ratio achieved by *FastIOV* compared with *Vanilla*.

Overall performance. Fig. 15 illustrates the task completion time distribution of running the four serverless applications on 200 concurrently launched containers. The task completion time refers to the duration from the issuance of the startup command to the completion of the container application. Compared to *vanilla*, *FastIOV* achieves reductions of 12.1%-53.5% and 20.3%-53.7% in the average and the 99th percentile task completion time across all applications. We notice that the reduction ratio decreases from application *Image* to *Inference*, which is attributed to the fact that the task execution time increases from *Image* to *Inference*, reducing the portion that container startup takes up in the total time. This suggests that the benefits of FastIOV are more pronounced with shorter-lived applications.

Impacting factors. Similar to §6.3, we evaluate the performance of FastIOV on serverless applications with varying concurrency, varying resource allocation, and also in the case of a fully loaded server. Fig. 16 shows the average task completion time and the time reduction ratio achieved by *FastIOV*. The overall trend is similar to that in §6.3: (i) with a fixed per container resource allocation, FastIOV achieves higher performance gain at a higher concurrency (Fig. 16a~d); (ii) at a fixed concurrency, FastIOV achieves higher performance gain with larger per container resource allocation (Fig. 16e~h); (iii) with a fully loaded server, FastIOV achieves obvious time reduction across all applications and concurrency settings, which is most pronounced at lower concurrency (Fig. 16i~l).

Compared to §6.3, there is a notable difference when we vary the resource allocation per container: with more resource allocation, the task completion time of FastIOV remains unchanged (*cf.* Fig. 16e and f) or even decreases (*cf.* Fig. 16g and h). This is because the increased resource allocation shortens the task execution time and thus reduces the task completion time for FastIOV. This demonstrates that *FastIOV* enables applications to more effectively reap the benefits of increased resource allocation.

7 Discussion of Limitation

Enabling FastIOV for new types of SR-IOV devices, such as RDMA NICs or NVMe storage, requires their drivers to ensure that the buffer memory is already EPT faulted when the device is the first to access the memory. This is to avoid the same issue with virtio as explained in §4.3.2. This should require only simple changes in the drivers like what we did to the virtio frontend that performs a data read to the first byte of each page of the memory to proactively trigger EPT faults, or less efficiently, like the zeroing performed by the drivers of standard NICs. However, this can become problematic if the device driver is closed-source and not modifiable by cloud providers. A potential solution and future research direction is to resort to the vDPA technology [36]. VDPA extends SR-IOV and enables the guest to use standard open-source virtio driver to perform data transfer instead of vendor-specific drivers. However, its effect on the concurrent startup performance requires further investigation.

8 Related Works

CNI enhancements. The cloud-native community has been developing plenty of widely-used CNI plugins like Flannel [12], Calico [10] and Cilium [11]. The state-of-the-arts mainly choose them as baselines and try to optimize their data plane performance using techniques like pipeline parallelism [32], resource allocation optimization[31] or VXLAN enhancement [7, 72]. Relatively fewer works have recognized the significance of the startup performance [47, 61], and those works only optimize the startup of software-based CNIs for traditional containers. PCPM [47] pre-creates the virtual network devices and network configurations as pause containers, and dynamically attaches them to newly launched containers. However, when using SR-IOV for secure containers, the startup bottleneck does not lie in the creation of the VFs but in the attaching process. Particle [61] identifies the startup bottleneck of using the `veth`-based software CNI as the NNS moving, and resolves this problem by sharing the NNS. However, NNS moving is not a key bottleneck when using secure containers either with SR-IOV based CNIs (§3.2) or software based CNIs [42].

SR-IOV enhancements. Due to the good data plane performance with the high throughput and low latency, SR-IOV outperforms other forms of network I/O virtualization and has been widely adopted in various applications [2, 21, 43, 48, 49]. Many existing works make a step further to enhance SR-IOV's performance. They make up for the lack of live migration [26, 54, 66, 68], improve the deployment density [17, 19, 56], avoid the performance degradation caused by frequent transmission interrupts [25, 34] and enhance the logic isolation and performance isolation for multi-tenancy scenarios [30, 69]. However, none of these works recognizes the demand for improving the concurrent startup performance of SR-IOV enabled networking. HD-IOV [70] is a recent work that mainly focuses on improving the deployment density, resource flexibility, and driver compatibility of SR-IOV. While it also accelerates VF initialization, it does not consider the concurrent startup of VFs and hence fails to address the key bottlenecks unique to concurrent startup scenarios, such as contention for the global VFIO lock.

Passthrough I/O optimizations. An important line of works regarding passthrough I/O is the optimization of the IOMMU module [4, 5, 41, 44, 62, 64]. Among those works, the most relevant to our FastIOV are the designs of virtual IOMMU [4, 62, 64]. vIOMMU [4] identifies that the page pinning operation of DMA memory mapping in IOMMU prohibits memory over-commitment. It introduces an IOMMU emulation layer to delay the mapping establishment and perform mapping when a memory region is actually accessed by DMA. coIOMMU [62] relieves the performance degradation problem of vIOMMU by decoupling the DMA mapping and page pinning process. V-Probe [64] further solves the intrusiveness problem in coIOMMU's design by adopting an eBPF based design. The delayed DMA memory mapping in those virtual IOMMUs can reduce the startup cost of passthrough I/O. However, such reduction is coupled with the enabling of memory-overcommitment, which is not always the preferred option in multi-tenant clouds [39]. By comparison, our FastIOV decouples the root cause of overhead, *i.e.,* page zeroing, from memory mapping to accelerate the startup, making it more flexible and applicable whether overcommitment is enabled or not.

VM/Container concurrency improvements. The majority of related works in this category focus on optimizing the startup performance of traditional containers. They reduce startup time by accelerating container image distribution [33, 40], introducing a specific checkpoint or general template-based runtime [20, 51], or providing warm startup solutions with technologies such as workload prediction and adaptive pooling [58, 67, 71]. Another series of works optimizes the startup of microVMs or VMs using techniques such as kernel trimming [1, 35, 46], cgroup pre-creation [35], hypervisor lock [50] and control plane redesign [46]. Those works focus on optimizing the non-network part of the startup, and are orthogonal to our work.

9 Conclusion

In the context of secure containers, SR-IOV enabled networking achieves a high data plane performance, a high deployment density, but a poor concurrent startup performance. We identify three key bottlenecks that cause the slow startup: (i) the contention for the coarse lock in VFIO devset management, (ii) the DMA memory mapping, and (iii) the VF driver initialization process. To address these bottlenecks, we propose a solution named FastIOV with optimization methods including lock decomposition, unnecessary mapping skipping, decoupled zeroing, and asynchronous VF driver initialization. We implement FastIOV as a portable kernel module and an optimized CNI plugin, along with several modifications in the infrastructures managed by the cloud vendor. Compared to vanilla SR-IOV CNI, FastIOV reduces the VF-related startup time by 96.1%, the end-to-end startup time by 65.7%, and the task completion time of common serverless applications by 12.1%-53.5%.

Acknowledgments

We are sincerely grateful to our shepherd Reto Achermann and all anonymous reviewers for their valuable comments and suggestions. This work was supported in part by the National Natural Science Foundation of China (No. 62072302, 61960206002), the Key R&D Program of Zhejiang Province (No. 2023R5202) and Alibaba Innovation Research Project (No. 2022010307).

References

[1] Alexandru Agache, Marc Brooker, Alexandra Iordache, Anthony Liguori, Rolf Neugebauer, Phil Piwonka, and Diana-Maria Popa. 2020. Firecracker: Lightweight Virtualization for Serverless Applications. In *Proceedings of USENIX NSDI*. USENIX, Santa Clara, USA, 419–434.

[2] Giuliano Albanese, Robert Birke, Georgia Giannopoulou, Sandro Schönborn, and Thanikesavan Sivanthi. 2021. Evaluation of Networking Options for Containerized Deployment of Real-Time Applications. In *Proceedings of IEEE International Conference on Emerging Technologies and Factory Automation*. IEEE, Vasteras, Sweden, 1–8.

[3] Amazon. 2023. AWS Lambda. https://www.aliyun.com/product/fc.

[4] Nadav Amit, Muli Ben-Yehuda, Dan Tsafrir, Assaf Schuster, et al. 2011. vIOMMU: Efficient IOMMU Emulation. In *Proceedings of USENIX ATC*. USENIX, Portland, USA, 1–14.

[5] Nadav Amit, Muli Ben-Yehuda, and Ben-Ami Yassour. 2010. IOMMU: Strategies for Mitigating the IOTLB Bottleneck. In *Proceedings of ACM/IEEE ISCA*. ACM, Saint-Malo, France, 256–274.

[6] Azure. 2023. Azure Cosmos DB. https://azure.microsoft.com/en-us/products/cosmos-db.

[7] Sunyanan Choochotkaew, Tatsuhiro Chiba, Scott Trent, and Marcelo Amaral. 2022. Bypass Container Overlay Networks with Transparent BPF-driven Socket Replacement. In *Proceedings of IEEE International Conference on Cloud Computing*. IEEE, Barcelona, Spain, 134–143.

[8] CNCF Community. 2024. Containerd: An Industry-Standard Container Runtime with An Emphasis on Simplicity, Robustness and Portability. https://containerd.io/.

[9] CNCF Community. 2024. Kubernetes: An Open-Source System for Automating Deployment, Scaling and Management of Containerized Applications. https://kubernetes.io/.

[10] Calico Community. 2024. Calico CNI Project. https://github.com/projectcalico/calico.

[11] Cilium Community. 2024. Cilium CNI Project. https://github.com/cilium/cilium.

[12] Flannel Community. 2024. Flannel CNI Project. https://github.com/flannel-io/flannel.

[13] PCI-SIG Community. 2024. PCI Special Interest Group. http://www.pcisig.com/home.

[14] Marcin Copik, Grzegorz Kwasniewski, Maciej Besta, Michal Podstawski, and Torsten Hoefler. 2021. SeBS: A Serverless Benchmark Suite for Function-as-a-Service Computing. In *Proceedings of ACM/IFIP International Middleware Conference*. ACM, Virtual Event, 64–78.

[15] Intel Corporation. 2023. Intel® Virtualization Technology for Directed I/O Architecture Specification Revision 4.1. https://www.intel.com/content/www/us/en/content-details/774206/intel-virtualization-technology-for-directed-i-o-architecture-specification.html.

[16] Intel Corporation. 2024. Intel Infrastructure Processing Unit (Intel IPU) SoC E2100 Product Brief. https://www.intel.com/content/www/us/en/content-details/818147/intel-infrastructure-processing-unit-intel-ipu-soc-e2100-product-brief.html.

[17] Intel Corporation. 2024. Scalable I/O Virtualization Technical Specification. https://cdrdv2-public.intel.com/671403/intel-scalable-io-virtualization-technical-specification.pdf.

[18] NVIDIA Corporation. 2024. NVIDIA CONNECTX-7 400G ETHERNET. https://www.nvidia.com/content/dam/en-zz/Solutions/networking/ethernet-adapters/connectx-7-datasheet-Final.pdf.

[19] NVIDIA Corporation. 2024. Scalable Function Overview. https://github.com/Mellanox/scalablefunctions/wiki.

[20] Dong Du, Tianyi Yu, Yubin Xia, Binyu Zang, Guanglu Yan, Chenggang Qin, Qixuan Wu, and Haibo Chen. 2020. Catalyzer: Sub-Millisecond Startup for Serverless Computing with Initialization-Less Booting. In *Proceedings of ACM ASPLOS*. ACM, Lausanne, Switzerland, 467–481.

[21] Brice Ekane, Tu Dinh Ngoc, Boris Teabe, Daniel Hagimont, and Noel De Palma. 2022. FlexVF: Adaptive Network Device Services in A Virtualized Environment. *Future Generation Computer Systems* 127 (2022), 14–22.

[22] Abdul Halim et al. 2024. Network Device Plugin for Kubernetes. https://github.com/k8snetworkplumbingwg/sriov-network-device-plugin.

[23] Ye Yin et al. 2024. SR-IOV CNI Plugin Project. https://github.com/hustcat/sriov-cni.

[24] Open Infrastructure Foundation. 2024. Kata Containers: the Speed of Containers, the Security of VMs. https://katacontainers.io/.

[25] HaiBing Guan, YaoZu Dong, Kun Tian, and Jian Li. 2013. SR-IOV based Network Interrupt-Free Virtualization with Event based Polling. *IEEE JSAC* 31, 12 (2013), 2596–2609.

[26] Wei Lin Guay, Sven-Arne Reinemo, Bjørn Dag Johnsen, Tor Skeie, and Ola Torudbakken. 2012. A Scalable Signalling Mechanism for VM Migration with SR-IOV over Infiniband. In *Proceedings of IEEE ICPADS*. IEEE, Singapore, 384–391.

[27] Intel. 2024. Data Plane Development Kit (DPDK). http://www.dpdk.org

[28] Intel. 2024. Intel® Ethernet Controller E810. https://www.intel.com/content/www/us/en/products/details/ethernet/800-controllers/e810-controllers/docs.html.

[29] Intel. 2024. Intel® Infrastructure Processing Unit (Intel® IPU) Adapter E2100. https://www.intel.com/content/www/us/en/products/details/network-io/ipu/adapter-e2100.html.

[30] Xinhao Kong, Jingrong Chen, Wei Bai, Yechen Xu, Mahmoud Elhaddad, Shachar Raindel, Jitendra Padhye, Alvin R Lebeck, and Danyang Zhuo. 2023. Understanding RDMA Microarchitecture Resources for Performance Isolation. In *Proceedings of USENIX NSDI*. USENIX, Boston, USA, 31–48.

[31] Kyungwoon Lee, Kwanhoon Lee, Hyunchan Park, Jaehyun Hwang, and Chuck Yoo. 2022. Autothrottle: Satisfying Network Performance Requirements for Containers. *IEEE Transactions on Cloud Computing* 11 (2022), 2096–2109.

[32] Jiaxin Lei, Manish Munikar, Kun Suo, Hui Lu, and Jia Rao. 2021. Parallelizing Packet Processing in Container Overlay Networks. In *Proceedings of ACM EuroSys*. ACM, Virtual Event, 1–16.

[33] Huiba Li, Yifan Yuan, Rui Du, Kai Ma, Lanzheng Liu, and Windsor Hsu. 2020. DADI: Block-Level Image Service for Agile and Elastic Application Deployment. In *Proceedings of USENIX ATC*. USENIX, Virtual Event, 727–740.

[34] Jian Li, Shuai Xue, Wang Zhang, Ruhui Ma, Zhengwei Qi, and Haibing Guan. 2017. When I/O Interrupt Becomes System Bottleneck: Efficiency and Scalability Enhancement for SR-IOV Network Virtualization. *IEEE Transactions on Cloud Computing* 7, 4 (2017), 1183–1196.

[35] Zijun Li, Jiagan Cheng, Quan Chen, Eryu Guan, Zizheng Bian, Yi Tao, Bin Zha, Qiang Wang, Weidong Han, and Minyi Guo. 2022. RunD: A Lightweight Secure Container Runtime for High-density Deployment and High-Concurrency Startup in Serverless Computing. In *Proceedings of USENIX ATC*. USENIX, Carlsbad, USA, 53–68.

[36] Cunming Liang and Tiwei Bie. 2018. vdpa: vhost-mdev as a New vhost Protocol Transport. In KVM Forum.

[37] Linux Kernel Organization. 2024. Kernel Virtual Machine. https://linux-kvm.org/page/Main_Page.

[38] Linux Kernel Organization. 2024. The Linux Kernel Archives. https://www.kernel.org/.

[39] Fangming Liu and Yipei Niu. 2023. Demystifying the Cost of Serverless Computing: Towards A Win-Win Deal. *IEEE TPDS* 35 (2023), 59–72.

[40] Haifeng Liu, Wei Ding, Yuan Chen, Weilong Guo, Shuoran Liu, Tianpeng Li, Mofei Zhang, Jianxing Zhao, Hongyin Zhu, and Zhengyi Zhu. 2019. CFS: A Distributed File System for Large Scale Container Platforms. In *Proceedings of ACM SIGMOD*. ACM, Amsterdam, The Netherlands, 1729–1742.

[41] Ming Liu, Tao Li, Neo Jia, Andy Currid, and Vladimir Troy. 2015. Understanding the Virtualization "Tax" of Scale-out Pass-through GPUs in GaaS Clouds: An Empirical Study. In *Proceedings of IEEE HPCA)*. IEEE, Burlingame, USA, 259–270.

[42] Yunzhuo Liu, Junchen Guo, Pengyu Zhang, Bo Jiang, Xiaoqing Sun, Yang Song, Wei Ren, Zhiyuan Hou, Biao Lyu, Rong Wen, Shunmin Zhu, and Xinbing Wang. 2024. Understanding Network Startup for Secure Containers in Multi-Tenant Clouds: Performance, Bottleneck and Optimization. In *Proceedings of ACM IMC*. ACM, Madrid, Spain, 1–16.

[43] Diego Rossi Mafioletti, Cristina Klippel Dominicini, Magnos Martinello, Moises R. N. Ribeiro, and Rodolfo da Silva Villaça. 2020. PIaFFE: A Place-as-you-go In-Network Framework for Flexible Embedding of VNFs. In *Proceedings of IEEE ICC*. IEEE, Virtual Event, 1–6.

[44] Moshe Malka, Nadav Amit, Muli Ben-Yehuda, and Dan Tsafrir. 2015. rIOMMU: Efficient IOMMU for I/O Devices that Employ Ring Buffers. In *Proceedings of ACM ASPLOS*. ACM, New York, USA, 355–368.

[45] Filipe Manco, Costin Lupu, Florian Schmidt, Jose Mendes, Simon Kuenzer, Sumit Sati, Kenichi Yasukata, Costin Raiciu, and Felipe Huici. 2017. My VM is Lighter (and Safer) than your Container. In *Proceedings of ACM SOSP*. ACM, Shanghai, China, 218–233.

[46] Filipe Manco, Costin Lupu, Florian Schmidt, Jose Mendes, Simon Kuenzer, Sumit Sati, Kenichi Yasukata, Costin Raiciu, and Felipe Huici. 2017. My VM is Lighter and Safer than Your Container. In *Proceedings of ACM SOSP*. ACM, Shanghai, China, 218–233.

[47] Mohan, Anup and Sane, Harshad and Doshi, Kshitij and Edupuganti, Saikrishna and Nayak, Naren and Sukhomlinov, Vadim. 2019. Agile Cold Starts for Scalable Serverless. In *Proceedings of USENIX HotCloud*. USENIX, Renton, USA, 1–6.

[48] T.P. Nagendra and R. Hemavathy. 2023. Unlocking Kubernetes Networking Efficiency: Exploring Data Processing Units for Offloading and Enhancing Container Network Interfaces. In *Proceedings of IEEE Global Conference for Advancement in Technology*. IEEE, Bengaluru, India, 1–7.

[49] Dinh Tam Nguyen, Ngoc Lam Dao, Van Thuyet Tran, Khac Thuan Lang, Thanh Tu Pham, Phi Hung Nguyen, Cong Dan Pham, Tuan Anh Pham, Duc Hai Nguyen, and Huu Thanh Nguyen. 2022. Enhancing CNF Performance for 5G Core Network Using SR-IOV in Kubernetes. In *Proceedings of IEEE International Conference on Advanced Communication Technology*. IEEE, Pyeongchang, Korea, 501–506.

[50] Vlad Nitu, Pierre Olivier, Alain Tchana, Daniel Chiba, Antonio Barbalace, Daniel Hagimont, and Binoy Ravindran. 2017. Swift Birth and Quick Death: Enabling Fast Parallel Guest Boot and Destruction in the Xen Hypervisor. *ACM SIGPLAN Notices* 52, 7 (2017), 1–14.

[51] Edward Oakes, Leon Yang, Dennis Zhou, Kevin Houck, Tyler Harter, Andrea Arpaci-Dusseau, and Remzi Arpaci-Dusseau. 2018. SOCK: Rapid Task Provisioning with Serverless-Optimized Containers. In *Proceedings of USENIX ATC*. USENIX, Boston, USA, 57–70.

[52] Linux Kernel Organization. 2024. AppArmor-Linux kernel Security Module. https://apparmor.net.

[53] Linux Kernel Organization. 2024. SECure COMPuting with filters. https://www.kernel.org/doc/Documentation/prctl/seccomp_filter.txt.

[54] Zhenhao Pan, Yaozu Dong, Yu Chen, Lei Zhang, and Zhijiao Zhang. 2012. CompSC: Live Migration with Pass-through Devices. In *Proceedings of ACM SIGPLAN/SIGOPS Conference on Virtual Execution Environments*. ACM, London, UK, 109–120.

[55] Ashish Panwar, Sorav Bansal, and K Gopinath. 2019. Hawkeye: Efficient Fine-Grained OS Support for Huge Pages. In *Proceedings of ACM ASPLOS*. ACM, Providence, USA, 347–360.

[56] Bo Peng, Haozhong Zhang, Jianguo Yao, Yaozu Dong, Yu Xu, and Haibing Guan. 2018. MDev-NVMe: A NVMe Storage Virtualization Solution with Mediated Pass-Through. In *Proceedings of USENIX ATC*. USENIX, Vancouver, Canada, 665–676.

[57] Shixiong Qi, Sameer G. Kulkarni, and K.K. Ramakrishnan. 2020. Assessing Container Network Interface Plugins: Functionality, Performance, and Scalability. *IEEE Transactions on Network and Service Management* 18, 1 (2020), 656–671.

[58] Rohan Basu Roy, Tirthak Patel, and Devesh Tiwari. 2022. Icebreaker: Warming Serverless Functions Better with Heterogeneity. In *Proceedings of ACM ASPLOS*. ACM, Lausanne, Switzerland, 753–767.

[59] Siarhei Siamashka. 2024. Tinymembench. https://github.com/sysprog21/tinymembench.

[60] Kun Suo, Junggab Son, Dazhao Cheng, Wei Chen, and Sabur Baidya. 2021. Tackling Cold Start of Serverless Applications by Efficient and Adaptive Container Runtime Reusing. In *Proceedings of IEEE International Conference on Cluster Computing*. IEEE, Virtual Event, 433–443.

[61] Shelby Thomas, Lixiang Ao, Geoffrey M. Voelker, and George Porter. 2020. Particle: Ephemeral Endpoints for Serverless Networking. In *Proceedings of ACM SoCC*. ACM, Virtual Event, 16–29.

[62] Kun Tian, Yu Zhang, Luwei Kang, Yan Zhao, and Yaozu Dong. 2020. coIOMMU: A Virtual IOMMU with Cooperative DMA Buffer Tracking for Efficient Memory Management in Direct I/O. In *Proceedings of USENIX ATC*. USENIX, Virtual Event, 479–492.

[63] Liang Wang, Mengyuan Li, Yinqian Zhang, Thomas Ristenpart, and Michael Swift. 2018. Peeking behind the Curtains of Serverless Platforms. In *Proceedings of USENIX ATC*. USENIX, Boston, USA, 133–146.

[64] Yaohui Wang, Ben Luo, and Yibin Shen. 2023. Efficient Memory Overcommitment for I/O Passthrough Enabled VMs via Fine-grained Page Metadata Management. In *Proceedings of USENIX ATC*. USENIX, Boston, USA, 769–783.

[65] Dan Williams and Ricardo Koller. 2016. Unikernel Monitors: Extending Minimalism outside of the Box. In *Proceedings of USENIX HotCloud*. USENIX, Denvor, USA, 71–76.

[66] Xin Xu and Bhavesh Davda. 2017. A Hypervisor Approach to Enable Live Migration with Passthrough SR-IOV Network Devices. *ACM SIGOPS Operating Systems Review* 51, 1 (2017), 15–23.

[67] Yanan Yang, Laiping Zhao, Yiming Li, Huanyu Zhang, Jie Li, Mingyang Zhao, Xingzhen Chen, and Keqiu Li. 2022. INFless: A Native Serverless System for Low-Latency, High-Throughput Inference. In *Proceedings of ACM ASPLOS*. ACM, Lausanne, Switzerland, 768–781.

[68] Jie Zhang, Xiaoyi Lu, and Dhabaleswar K Panda. 2017. High-Performance Virtual Machine Migration Framework for MPI Applications on SR-IOV Enabled InfiniBand Clusters. In *Proceedings of IEEE IPDPS*. ACM, Orlando, USA, 143–152.

[69] Yiwen Zhang, Yue Tan, Brent Stephens, and Mosharaf Chowdhury. 2022. Justitia: Software Multi-Tenancy in Hardware Kernel-Bypass Networks. In *Proceedings of USENIX NSDI*. USENIX, Renton, USA, 1307–1326.

[70] Zongpu Zhang, Jiangtao Chen, Banghao Ying, Yahui Cao, Lingyu Liu, Jian Li, Xin Zeng, Junyuan Wang, Weigang Li, and Haibing Guan. 2024. HD-IOV: SW-HW Co-designed I/O Virtualization with Scalability and Flexibility for Hyper-Density Cloud. In *Proceedings of ACM EuroSys*. ACM, Athens, Greece, 834–850.

[71] Zhuangzhuang Zhou, Yanqi Zhang, and Christina Delimitrou. 2023. Aquatope: QoS-and-Uncertainty-Aware Resource Management for Multi-Stage Serverless Workflows. In *Proceedings of ACM ASPLOS*. ACM, Vancouver, Canada, 1–14.

[72] Danyang Zhuo, Kaiyuan Zhang, Yibo Zhu, Hongqiang Harry Liu, Matthew Rockett, Arvind Krishnamurthy, and Thomas Anderson. 2025. ONCache: A Cache-Based Low-Overhead Container Overlay Network. In *Proceedings of USENIX NSDI*. USENIX, Philadelphia, USA, 1–16.