Aggregate VM: Why Reduce or Evict VM’s Resources When You Can Borrow Them From Other Nodes?

Citation for published version:

Digital Object Identifier (DOI):
10.1145/3552326.3587452

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published in:
EuroSys ’23: Proceedings of the Eighteenth European Conference on Computer Systems

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Aggregate VM: Why Reduce or Evict VM’s Resources When You Can Borrow Them From Other Nodes?

Ho-Ren Chuang∗
Virginia Tech
Virginia, USA
horenc@vt.edu

Karim Manaouil∗
The University of Edinburgh
Scotland, UK
karim.manaouil@ed.ac.uk

Tong Xing∗
The University of Edinburgh
Scotland, UK
tong.xing@ed.ac.uk

Antonio Barbalace
The University of Edinburgh
Scotland, UK
antonio.barbalace@ed.ac.uk

Pierre Olivier
University of Manchester
England, UK
pierre.olivier@manchester.ac.uk

Balvansh Heerekar
Virginia Tech
Virginia, USA
balvansh@vt.edu

Binoy Ravindran
Virginia Tech
Virginia, USA
binoy@vt.edu

Abstract

Hardware resource fragmentation is a common issue in data centers. Traditional solutions based on migration or overcommitment are unacceptably slow, and modern commercial or research solutions like Spot VM may reduce or evict VM’s resources anytime. We propose an alternative solution that does not suffer from these drawbacks, the Aggregate VM. We introduce a new distributed hypervisor design, the resource-borrowing hypervisor, which creates Aggregate VMs: distributed VMs that temporarily aggregate fragmented resources belonging to different host machines, which require mobility of virtual CPUs, memory and IO devices. We implement a prototype, FragVisor, which runs guest software transparently. We also propose minimal modifications to the guest OS that can enable significant performance gains. We evaluate FragVisor over a set of microbenchmarks and IaaS-style real applications. Although Aggregate VMs are not a perfect fit for every type of applications, some workloads enjoy significant speedups compared to overcommitted scenarios (up to 3.9x with 4 distributed vCPUs). We further demonstrate that FragVisor is faster than a state-of-the-art competitor, GiantVM (up to 2.5x).

CCSConcepts: • Computer systems organization → Distributed architectures; • Software and its engineering → Virtual machines.

Keywords: Resource Fragmentation, Data Center, Distributed Hypervisor, DSM, Migration, Delegation

∗The authors contributed equally to this work.

ACM Reference Format:

1 Introduction

Hardware resource fragmentation in modern data centers is a critical problem [12, 30, 39, 42, 45, 75, 86, 95]. It is due to several factors: the broad variety of resource requirements for jobs; an increase in the average amount of resources required per job [29, 67]; job placement constraints such as software/hardware dependencies [87, 93]; and poor hardware elasticity/inadequate support for heterogeneity [91]. Even considering optimized scheduling and placement methods [29, 45], or attempts to defragment resources through migration, resource fragmentation still persists [39, 82, 91, 95]. Although disaggregated hardware [37, 46, 61, 62, 91, 95] appears to be a solution in the long term, it is not viable yet, and it is also not a panacea: it requires a hardware refresh, a capital investment, and only solves the resource fragmentation problem for memory. Thus, even if cloud providers already attempted to monetize fragmented resources – e.g., with Spot VMs [11, 41, 66], they are still seeking solutions to further increase single-machine resource utilization without affecting the performance of Primary VMs [12, 50, 63, 102], i.e., VMs with guaranteed resources. Spot VMs, as well as more recent works [12, 102], do not guarantee fixed performance, but only a certain minimum amount of resources, and they can be evicted (VM kill) with minimal/no notification. Hence, they are not suitable for Primary VMs’ workloads. Moreover, with such solutions the data center scheduler still needs to exactly find a machine with that minimum amount of resources available.

Idea. This paper approaches the problem of data center resource fragmentation in a fundamentally different way: instead of exploiting free resources at the granularity of a single-machine, it aggregates fragmented resources available among
multiple machines into a single distributed VM. To achieve that, we introduce a new type of VM, the Aggregate VM. Differently from previous works, it guarantees a certain fixed number of resources at all time without eviction, but the guaranteed SLO is based on the type of workload. Therefore, this paper is driven by the following questions: a) can Aggregate VMs be a solution to the fragmentation problem, and at what overhead? b) for what workloads, and at what SLO?

To answer these questions, we introduced a new distributed virtual machine monitor, the resource-borrowing hypervisor, which provides Aggregate VMs as first class VMs. This enables the traditional unit of resource allocation in the cloud, the Virtual Machine (VM), including processing units (virtual CPUs, vCPUs), (pseudo-)physical memory, and I/O devices, to be distributed over fragmented hardware resources belonging to different physical servers. This distribution is transparent, i.e., in its basic form, it requires no modification to guest software. It is also temporary, to reflect the dynamic nature of fragmentation, virtualized resources are “mobile” between servers.

FragVisor. We built FragVisor, a resource-borrowing hypervisor that runs on existing data center infrastructures, creating and maintaining Aggregate VMs. Aggregate VMs provide the exact number of resources requested by a user for a VM, reducing the potential performance degradation versus solutions based on resource overcommitment, and eliminating the possibility of resource eviction of transient VMs (Spot VMs, Preemptible VMs, Harvest VMs, etc). Moreover, it also avoids downtimes due to potentially complex multi-VM migrations. However, as a distributed VM, it presents some unavoidable overheads when accessing remote resources. We demonstrate that these slowdowns are much lower compared to solutions based on resource overcommitment, focusing on the CPU – a resource for which fragmentation has been shown to be particularly problematic for VM allocation [12].

FragVisor is a distributed multi-hypervisor [80]. It extends Linux/KVM to run among multiple machines and leverages a Distributed Shared Memory (DSM) system to transparently present to the guest a unified and consistent view of its physical address space. FragVisor allows a vCPU on one physical machine to access memory as well as remote devices, such as virtual disks, from other machines. FragVisor is able to execute fully unmodified guest OSes, although some non-intrusive modifications to the guest OSes can bring significant performance improvements. More importantly, FragVisor does not require any modification to legacy user-space software.

Innovation. To our knowledge, this is the first work that considers the potentials of leveraging a distributed VM, the Aggregate VM, to solve the resource fragmentation problem in the data center. Although several distributed hypervisors have been proposed in the past [8, 27, 89, 96, 104], those target a fundamentally different goal: running scale-up workloads on scale-out hardware. We focus instead on small- to medium-sized VMs that are common allocation units in IaaS. Further, we also note that existing distributed VM systems lack mechanisms enabling resource mobility among physical nodes.

Similarly to such existing works, FragVisor relies on DSM to provide a (temporary) distributed VM with a coherent (pseudo-)physical memory view over several nodes. Historically, DSM has been known to scale poorly in workloads exhibiting medium to high levels of memory sharing [13, 26, 59]. Recent attempts at reviving DSM, even when they rely on modern high-speed interconnects, still target workloads with relatively low levels of sharing [55, 71, 104]. We observe that in data center IaaS settings, several classical workloads are made up of concurrent components executing in the same VM and exhibiting very low degrees of sharing, e.g., web server/language runtime/database stacks, serverless computing, etc. Our intuition is that for such workloads, an Aggregate VM can provide a similar SLO than a Primary VM.

To address the dynamic nature of fragmentation in the data center, along with memory mobility (using DSM), an Aggregate VM needs mobility of virtualized CPUs and I/O devices. A cross-node vCPU migration mechanism enables part of or an entire VM to transparently move at runtime where hardware resources become available, something not possible with existing distributed hypervisors. This allows us to consolidate a VM over time on as few servers as possible – ideally a single one. We also exploit vCPU mobility and a distributed checkpoint/restart mechanism to tackle fault resilience, exacerbated by running a single VM on multiple machines. Finally, we propose a set of new techniques to support I/O device mobility, including single- and multiple-queue I/O delegation [88] on top of DSM, DSM-bypass I/O delegation, and distributed I/O.

Key Prototype Results. We prototyped FragVisor and evaluated its performance on a computer cluster, over a set of micro- and macro-benchmarks. Among others, the results show that with four distributed vCPUs, an Aggregate VM offers significant speedups for compute-bound (up to 3.9x) and networking (up to 3.6x) applications, when compared to overcommitment. FragVisor is also faster than a state-of-the-art competitor, GiantVM [104]: up to 2.5x for compute workloads, and 1.3x for network applications. When running shared-memory multithreaded applications on top of an Aggregate VM, the SLO is impacted based on the degree of sharing. FragVisor’s slowdown is generally acceptable (15%), although it is not a panacea for workloads relying heavily on shared memory, which may experience higher overheads.

Contribution. The paper makes the following contributions:

- We identify IaaS workloads that are less affected by running on a distributed VM. For those, the SLO of Aggregate VMs is similar to the one of Primary VMs;
- We propose the Aggregate VM, a new approach to solve fragmentation in the data center, avoiding resource evictions. An Aggregate VM leverages idle hardware resources belonging to different nodes to reduce fragmentation;
• We introduce the resource-borrowing hypervisor design, a new distributed hypervisor that provides Aggregate VMs;
• We build FragVisor, an implementation of the resource-borrowing hypervisor based on Linux/KVM, providing the mobility of virtual CPU, memory, and devices, while introducing new distributed hypervisor mechanisms, and guest kernel optimizations;
• We evaluate FragVisor, demonstrating its superior performance over a baseline and a competitor.

Section 2 further motivates our work, and characterizes several IaaS workloads’ suitability to run on an Aggregate VM. Background material is revised in Section 3. We present our architecture in Section 4, while Section 5 and Section 6 describe FragVisor’s design and implementation, respectively. FragVisor is evaluated in Section 7, and contrasted with past works in Section 8. Section 10 concludes.

2 Motivation

Fragmentation in the Data Center. Researchers from Microsoft recently stated that “Given Azure’s scale, even 1% in fragmentation reduction can lead to cost savings in the order of $100M per year” [45]. A study [42] analyzing traces on Facebook and Bing clusters noted that fragmentation and overprovisioning are responsible for a 45% increase in job makespan. The authors of a recent study [75], observing a cluster from the Eolas [35] cloud provider, noted an average of 17% of the physical resources wasted each day due to fragmentation. As a result, data center operators are still looking for solutions to reduce resource fragmentation [12, 50, 63, 102], and increase their returns on the data center investments. At the same time, commercial offers aiming at monetizing fragmentation exist from major hyperscalers [11, 41, 66], namely transient VMs (like Spot or Preemptible VMs), but those generate a reduced revenue compared to Primary VMs. This is because they cannot guarantee a certain constant level of provisioned resources, if not a minimal one – lesser SLO, and can be killed.

As mentioned above, fragmentation is due to many factors [12, 30, 39, 42, 45, 75, 86, 87, 91, 93, 95]. An ideal VM placement algorithm that could solve fragmentation is known to be a hard problem [39, 91, 95]. The same is true for attempting to “defragment” the data center by migrating running jobs on a smaller subset of physical machines [95]; migration has the additional downsides of requiring further data center resources (pre-/post-copy [64, 82]) or involving unacceptable downtimes (checkpoint/restart). Thus, in practice, when faced with the problem of accommodating more jobs on a saturated but fragmented set of hosts, the provider can either buy new machines or overcommit resources by packing more jobs on already busy hosts. Both solutions are suboptimal: buying new machines involves additional financial costs at a time when data center capital equipment costs are higher than operating ones [24], and overcommitment may lead to unacceptable performance degradation.

Figure 1. Single-machine (non-distributed) over DSM (distributed) execution time ratios as a function of the number of DSM page faults per second for various applications. An execution time ratio lower than 1 is a DSM slowdown.

Early Study: DSM, Sharing, and Scalability. An Aggregate VM relies on DSM to implement virtualized memory mobility, and to provide the guest with a coherent view of a pseudo-physical address space. DSM is known to be slow when data sharing is high [13, 26, 55, 59]. Therefore, we investigated the degree of data sharing among vCPUs when running several IaaS workloads – with the goal of finding workloads that would be minimally impacted when running in a distributed VM sitting on multiple physical nodes.

To that aim, we run on an early prototype of our FragVisor a set of applications including serial (NAS Parallel Benchmarks [69]) and scale-up multithreaded workloads (NPB OpenMP – OMP); a LEMP stack (Nginx/PHP/MySQL) in which the time to generate an HTML page to answer a request varies between 25 and 500ms, benchmarked with ApacheBench; and an instance of the OpenLambda FaaS computing framework [47]. We run the DSM on 2 and 4 nodes, and in each case we set the followings equal to the number of nodes: the number of serial NPB instances, the number of OpenMP threads, the number of PHP workers, and the number of OpenLambda workers. We measure the slowdown of running on top of DSM vs. a non-distributed (vanilla Linux) setting.

The results are presented in Figure 1. The slowdown increases with the level of sharing, or DSM contention – i.e., DSM faults per second. Applications showing low levels of sharing perform similarly on DSM and on a single-machine (from no slowdown, up to 45%): unsurprisingly, serial NPB but also embarrassingly parallel OMP workloads fall within that category, as well as the FaaS infrastructure. Interestingly, LEMP stacks with a page generation latency superior to 40ms also exhibit a modest slowdown, (30% to no slowdown) when running on DSM vs on a single-machine. On the other hand, applications with high degrees of sharing, such as most OMP benchmarks and high performance LEMP (page generation < 40ms), suffer significantly when executing atop DSM (up to 95% slowdown).
Thus, certain IaaS workloads’ performance will not be penalized when running atop an Aggregate VM – they will have a similar SLO to Primary VMs. Hence, herein we mainly target such a subset of IaaS workloads, characterized by limited sharing among threads – an Aggregate VM is not a general solution, the SLO depends on the software running in the VM. Yet, we observe that the amount of cloud/IaaS applications presenting characteristics suitable for distributed execution is non-negligible: for example, regarding LEMP, according to W3Tech [99], WordPress (a LEMP stack) runs today more than 40% of the Internet’s websites. Regarding serverless computing, expert estimate that its usage will skyrocket [51] in the next years.

3 Background

Virtualization Technologies. Two virtualization architectures are widespread: Type-1, in which the hypervisor directly runs on the hardware, e.g., Xen [20], and Type-2, where the hypervisor coexists with or is an OS service, e.g. KVM [57]. With the latter, virtualized CPUs run as host OS threads, and their physical address space (Guest Physical Addresses, GPAs) corresponds to the host OS threads’ address space (Host Virtual Addresses, HVAs). In order to let the guest address its own virtual memory (Guest Virtual Addresses, GVAs), a virtual memory management unit (vMMU) is needed. While the vMMU was implemented using shadow page tables in software [9, 83], today’s CPUs provide hardware vMMU – such as Extended Page Tables (EPT) on Intel processors [70], which provide two levels of hardware MMU translations. A page fault in the guest OS can be handled by the guest MMU, but if there is no host page backing it, the fault will be handled by the host MMU.

Regarding CPU virtualization, manufacturers introduced a new execution mode for the guest VM itself. Each manufacturer has an instruction to switch execution mode, and a data structure has to be filled before switching (e.g., VMX in Intel) – which lists events that make the CPU exit virtual machine mode execution. We refer to this as “full virtualization”, while virtualization without hardware support that requires modification of the guest software is “paravirtualisation”.

Virtualizing Devices. With the introduction of CPU and memory hardware-assisted virtualization, manufacturers also added hardware virtualization features to devices [32]. Yet, devices with hardware virtualization features support only a limited number of VMs, or are expensive. In many situations, paravirtualized hardware devices are preferred.

VirtIO [88] is the de-facto standard paravirtualized device technology, and implements different classes of devices, including network interface cards (virtio-net), storage devices (virtio-blk), and consoles (virtio-console). A VirtIO device appears to the guest software as a PCIe device and requires the software to instantiate at least a couple of in-memory ring buffers – a transmission (TX) and a receiver (RX) ring. vHost is an evolution of VirtIO targeting Type-2 virtualization, it moves a device emulation mechanisms from user- to kernel-space to avoid the user-kernel switch, boosting performance [81].

VM Migration and Distributed Execution. VMs offer a convenient abstraction for software mobility. VM state can be extracted and moved to another physical machine through VM migration [28].

Sometimes, instead of assigning the resources of a single server to multiple VMs, it is interesting to aggregate resources of multiple servers into a single, larger, VM. This has been done before [8, 27, 96, 104] (see Section 8) by introducing distributed protocols to aggregate hardware resources available on different physical machines. The Popcorn Linux project [16–19, 23, 53, 78, 79, 90], and Kerrighed [54], also aggregate hardware resources, but achieve that at the OS level rather than at the hypervisor level.

4 Aggregate VM Design and Architecture

With the goals of a) fast VM provisioning (faster than delayed execution), b) minimal VM overheads (lower than resource overcommitment and VM migration, aiming at SLOs similar to Primary VMs), c) guaranteed resources provisioning (no reduced resources, nor evictions contrary to transient VMs), and d) data center-wide resource exploitation, we propose a new distributed hypervisor design that provisionally aggregates “slices” of fragmented hardware resources belonging to multiple physical machines into a single (distributed) VM – the resource-borrowing hypervisor.

Design Principles. The resource-borrowing hypervisor is based on the following design principles: a) aggregate at least the hardware resources that are needed for a VM to run at each instant; b) reduce over-provisioning; c) have minimal overhead and interference with other VMs or applications; d) be high-performance; and e) be (transparently) compatible with existing software, i.e., guest OSes and applications.

Operational Principles. A resource-borrowing hypervisor creates VMs aggregating hardware resources belonging to different server machines. Hence, as depicted in Figure 2, such Aggregate VMs may exploit physical CPUs, RAM, and I/O devices owned by different servers (M1, M2, M3 in Figure), and it is dynamically reconfigurable to leverage resources on different nodes for the same VM (cf. right and left of Figure 2).
When a resource-borrowing hypervisor creates an Aggregate VM, a hypervisor instance is started on each physical server involved in offering resources for that VM. Instances manage subsets of hardware resources from various servers, VM slices, which contribute to the whole VM.

One of the hypervisor instances is responsible for establishing the connection among all, and is responsible for starting the guest execution – thus, it should provide at least one virtualized CPU. Such an instance will be started with additional information about the other instances (e.g., hosts IP addresses), as well as the disk image(s), and eventually the kernel and bootloader. We call this instance a bootstrap VM slice; other instances are called companion VM slices. After the bootup phase, all VM slices are peers. A VM slice may include virtual CPUs, virtual RAM, and virtual devices of any type, such as IO interfaces, accelerators, etc. However, a VM slice can be composed of just memory (like previous work [33, 43, 76]); or just a device, such as a GPU or TPU (like GPUDirect [58]).

**System Architecture.** A resource-borrowing hypervisor is a distributed multiple-hypervisor [80]. Each VM slice runs on top of a different hypervisor instance, while instances run on different servers. Different hypervisor instances talk between each other using a communication layer, based for example on message-passing. The communication layer is exploited by a set of distributed hypervisor services that provide the illusion of a single VM among multiple hypervisors – the Aggregate VM. To stick to the design and operational principles, a resource-borrowing hypervisor implements the concept of mobility of virtualized memory, CPU, and device. While all the memory, CPUs, and devices of a VM may be distributed and presented as a single VM (all slices can access all resources), the "temporary" aspect of borrowing demands for mobility – which in turn, demands several old and new mechanisms. VM RAM (vRAM) mobility can be implemented with inter-machine memory copy, or distributed shared memory (DSM). VM CPUs (vCPUs) mobility can be implemented with thread remote creation, or migration. VM devices mobility can be implemented as proxying, delegation, tunnelling, etc. Finally, not all resources assigned to a VM should be allocated to a specific VM for its entire lifetime.

**System Orchestration.** Based on the scale of a data center, what hardware resources among what nodes will be assigned to an Aggregate VM is decided either by the resource-borrowing hypervisor itself, or by an external entity, such as a data center scheduler/orchestrator, e.g. Protean [45]. In the former case, the resource-borrowing hypervisor would be pre-deployed on all machines of the data center, would knows about the resource availability of every machine. Instead, when an external entity governs resource assignments, the resource-borrowing hypervisor is not required to always run on all machines of the data center, but only on the ones with fragmented resources to share. A data center scheduler is an example of external entity, which is anyway monitoring servers’ usage, and knows about the managed cluster’s already allocated hardware resources and the resource requirements of incoming tasks. However, this requires data center schedulers to be extended because current schedulers cannot exploit partial/fragmented resources. In this case, the scheduler will inform the resource-borrowing hypervisor to move a VM slice (or part of it) between hypervisor instances (for power, performance, demand of new resources, reliability, etc).

**Reliability.** Running a VM on top of multiple physical machines is less reliable than running a VM on top of a single machine. In other words, assuming one machine is 99.9% reliable, two are 99.8%, three are 99.7%, etc. However, it is not just the hardware to guarantee a specific reliability, but also the software [94]. Hence, a single software stack running among several machines, like in an Aggregate VM, may be more reliable than a software stack per machine [97]. From the point of view of the hardware, a resource-borrowing hypervisor cannot change hardware’s reliability, but it can exploit state-of-the-art hardware monitoring and logging subsystems (e.g., Intel MCA/AER) to preemptively force-migrate a VM slice from a likely-to-fail server to another. Other fault-tolerance techniques such as periodic checkpointing and restart on failure can further be used.

### 5 FragVisor

FragVisor is an implementation of the resource-borrowing hypervisor targeting a Type-2 full virtualization architecture for traditional monolithic UNIX-like OSES. It is designed around common state-of-the-art data center hardware, where servers – compute nodes with multicore CPUs and accelerators – are mainly interconnected via high-speed network(s). Therefore, FragVisor’s communication layer is based on message-passing, which in order to avoid user-kernel switches and maximize performance is located in the host kernel, similarly to bespoke multiple-kernel OSES [17, 22]. Hypervisor services run in a distributed fashion atop the communication layer, strictly in kernel space, again for performance reasons. Note that different services may require different consistency level to maintain their distributed state and provide mobility.

In the data center, FragVisor does not make any decision itself about what machines will be used by an Aggregate VM – i.e., FragVisor has no placement-like capability. This should be the role of a data center scheduler/orchestrator, such as Protean [45]. We suggest extending, but not changing, existing schedulers. We propose a prototype orchestrator and a policy in Section 6.5.

#### 5.1 Distributed Pseudo-Physical Memory

To provide virtualized memory mobility as well as the illusion of shared memory among machines (single system), FragVisor moves memory blocks between different machines.

Modern Type-2 hypervisors hold the guest pseudo-physical address space (virtualized memory) as a subset of the virtual address space of a host user-space application, the VMM. To
make such part of the address space available to the guest 
OS, consistently among different machines, as well as mobile, 
FragVisor adopts DSM. Parts of the address space that do 
not belong to the guest pseudo-physical memory area, but 
to emulated devices, are handled outside the DSM protocol (see Sect. 5.3). Software DSM has been criticized in the past, 
mainly due to its consistency overheads and poor scalability. 
Indeed, we already disclosed that because of DSM an Aggre-
gate VM may not provide the same SLO as a Primary VM, but 
there exist certain workloads that are less or not impacted by 
DSM. Also, faster networks are increasingly available in data 
centers [49], speeding up DSM, motivating it even further. 
Anyway, to alleviate criticism, FragVisor introduces several 
DSM optimizations, including a contextual DSM protocol, 
and run-time NUMA topology updates, presented below.

**Contextual DSM.** The hypervisor knows a lot about the con-
tent of the guest physical address space, especially about CPU-
dependent memory areas, including the position of the page 
table, interrupt table, etc. In SMP Oses, these are shared among 
all CPUs, and are highly used. Therefore, it is of the utmost 
importance to avoid the DSM protocol from slowing down their 
access. We propose a contextual DSM protocol that leverages 
information about the memory content to reduce DSM traffic.

**NUMA Topology Updates.** Modern Oses use NUMA infor-
mation to optimize operations such as scheduling and mem-
ory allocation. With the goal of reducing DSM traffic, FragVi-
sor informs the guest software about the non-uniform access 
latencies due to DSM by exposing a NUMA topology that 
reflects the placement of hardware resources on different server 
machines. Such NUMA topology is updated at runtime.

### 5.2 Distributed vCPU

To enable virtualized CPU mobility, a FragVisor Aggregate 
VM runs vCPUs as distributed threads over the guest pseudo-
physical memory, which is kept consistent among vCPUs 
using DSM. Each vCPU has its own set of registers, a local 
interrupt controller, and a timer (e.g., x86’s local APIC timer). 
There is usually no shared state among different vCPUs if not 
for processor-wide registers, such as some MSR registers in 
x86. These are kept consistent among hypervisor’s instances.

CPUs notify each other via IPI, including MSI. Thus, each 
hypervisor instance keeps track of the machine where each 
vCPU is using a vCPU location table. IPIs are turned into mes-
sages to hypervisor instances.

Finally, live slice migration, necessary for consolidation or 
fault-tolerance purposes, requires thread migration [18, 19, 
53] to move a running vCPU between different servers.

**Distributed vPIC.** Other than the CPU local interrupt con-
troller, one or more non-local interrupt controllers may ex-
ist in a VM (e.g., x86 IO-APIC). Since a non-local interrupt 
controller is usually an interrupt broker, and interrupts are 
converted into messages, this may be kept as non-replicated 
on the machine with the highest number of physical devices.

### 5.3 Delegated Virtual Devices

In FragVisor, virtual device access is mainly based on the 
concept of delegation, i.e., guest VM software running on a VM 
slice should be able to access any device exposed by the Aggre-
gate VM, but the actual communication with the physical de-
vice happens only on the hypervisor instance running on the 
same physical server as the device. This guarantees mobility.

Devices may communicate with the CPU either via memory-
mapped IO or IO ports. We focused on the former. Specifically, 
we support PCIe-based devices by establishing a distributed 
PCIe root complex and devices emulator. Many PCIe devices, 
including the virtual ones we support, instantiate ring buffers 
in vRAM. Should these buffers be managed by the DSM, the 
high amount of contention they create will significantly hurt 
performance. Thus, in FragVisor devices are using per-CPU 
TX/RX queues when possible, and we introduce DSM-bypass.

**Multiqueue VirtIO.** To limit our engineering efforts, we 
focus on paravirtualized hardware (VirtIO-based), but our de-
sign can be trivially extended. Here we explain how FragVisor 
supports virtio-net, but the same applies to virtio-blk, etc.

A paravirtualized network device exposes TX/RX ring 
buffer pairs that the guest uses to enqueue/dequeue pack-
ets. Because these ring buffers are on the VM virtual RAM, the 
DSM protocol maintains them consistent across hypervisor 
instances. To reduce DSM traffic, vCPUs running on differ-
ent physical nodes should avoid accessing the same TX/RX 
pairs. Hence, FragVisor adopts multiqueue [100] technology 
(supported by most Oses), where each TX/RX pair, is mapped 
to a different vCPU and hypervisor instance.

This solution allows VM slices that do not own a network 
device to delegate the transmission of network packets to 
other VM slices by simply writing a packet on DSM, and then 
sending an interrupt to notify a new packet has been written.

**DSM-bypass.** However, adopting multiqueue is not enough: 
the synchronization operations made on both ends of a com-
munication channel generate a high DSM overhead. Thus, as 
an optimization we bypass the DSM for TX/RX pairs. We make 
the paravirtualized network device on each VM slice writing 
and reading the related TX/RX pairs, and send or receive the 
information to or from the VM slice with the physical network 
card. Thus, for a TX event, the paravirtualized network device 
piggybacks the network packet(s) – that is read from the mem-
ory, to the TX interrupt sent to the VM slice with the physical 
network card. Thus, the DSM is excluded from the data path.

### 6 Implementation

We implemented a prototype of FragVisor based on the 
Linux kernel v4.4.137. To avoid reinventing the wheel, we 
based our implementation on different Linux kernel compo-
ments from the Popcorn Linux project [17, 55], including

---

1. When we started FragVisor, the latest stable Popcorn Linux was based on 
Linux kernel v4.4.137. Hence, FragVisor Linux kernel is based on such version.
Learning from previous works [8, 27, 52, 96, 104], we traced the guest VM executing different applications on Linux. That helped identify several uncorrelated kernel data structures that were placed in the same memory page and creating unnecessary DSM traffic due to false sharing. We patched the guest OS kernel to solve the problem, reducing the DSM traffic\(^3\).

**Optimizations.** By tracing, we identified another source of DSM traffic: hardware dirty bit management. The EPT subsystem can be configured to additionally set dirty bits in its own translation, which generates additional DSM traffic. Hence, we disabled dirty bit tracking – anyway, this is already taken care of by the DSM itself, thus redundant in this project.

FragVisor implements the notion of contextual DSM. For example, for page tables, in order to reduce the network traffic for TLB shootdown it piggybacks the page table modifications with the shootdown interrupt message – reducing DSM traffic.

FragVisor implements NUMA topology updates. Namely, static APIC tables show one NUMA zone per VM slices, when VM slices move, coalesce or split, we trigger an ACPI System Resource Affinity Update notification [98] to reflect changes.

### 6.2 Migrating vCPUs

Our prototype either lets the *Bootstrap* VM slice create all vCPU threads and then migrates those vCPU threads to *Companion* VM slices, or creates remote vCPUs threads (at boot time only). This allowed us to reuse part of the existing Popcorn task migration code to distribute vCPUs among multiple machines. To provide distributed vCPUs, we augmented task migration to account for additional vCPU-related state and metadata as described before – most doesn’t need to be kept consistent among machines. Finally, the prototype maintains for each VM slice a replicated array that tracks the position of every vCPU and is updated at each migration event. This is fundamental to implement mobility and inter-server messaging.

**Interrupts.** vCPU threads are the recipients of interrupts. In the prototype, we modified the hypervisor for interrupt dispatching in order to check if the target vCPU is local or remote. If it is remote, a message with the description of the intercept event is sent to another VM slice instance via the communication layer. Otherwise, the traditional code path is followed.

### 6.3 Distributed VM Devices

We rewrote most *virtio*-based and emulated *kvmttool* devices (all but 9p, balloon, VESA) in order to work atop our communication layer. Thus, a device physically located within a specific VM slice can be used by all slices of the corresponding VM. For performance reasons we targeted the kernel implementation of *virtio*, which is *vhost*. However, because *kvmttool* doesn’t support multiqueue with *vhost*, we had to write a patch to extend it\(^4\). As *kvmttool* doesn’t support virtio-GPU, we cannot showcase the feature of borrowing an accelerator. However, this is just a technical limitation, advantages of

---

\(^3\)This patch is available open-source.

\(^4\)Submitted to the Linux kernel mailing list.
such technology has been already commercially proved, e.g., NVIDIA GPUDirect [77].

**Network.** The prototype leverages Linux’s vhost-net. While our distributed PCIe layer takes care of the physical PCIe address ranges, the DSM replicates the TX/RX pairs on physical memory, and multiqueue reduces DSM traffic. With DSM bypass, TX/RX pairs are not replicated. We also need to properly handle the notifications between the guest and the host in a distributed environment. Thus, each node has to install the corresponding file descriptors (ioeventfd and MSI irqfd) to notify the guest and host to access TX/RX pairs.

When the guest VM sends a network packet, it enqueues the packet’s information on a TX ring buffer, and passes its index to the hypervisor. FragVisor will deliver the event to the VM slice with the physical device, which will fetch the packet using DSM. With DSM bypass, the packet is sent by the hypervisor to its destination through the communication layer.

Similarly, when a packet for the guest VM is received, vhost-net copies it into the guest memory, and injects an IRQ based on the TX/RX pair number. The VM slice managing that IRQ will move in the packet using DSM. With DSM bypass, the packet is sent to the VM slice managing that IRQ, which copies the data into guest memory.

**Storage.** The current prototype is capable of using virtual RAM as a backing storage (tmpfs), or vhost-blk. tmpfs exploits DSM for consistency, while the vhost-blk works like the vhost-net implementation, including multiqueue, and DSM bypass.

**Serial Console.** Remote nodes can print host user-space KVM logs. To achieve this, each thread migrated to a remote node has to reopen a PTY. Furthermore, the system has only one pseudo-terminal worker thread emulating a serial UART chip.

### 6.4 Distributed Checkpoint/Restart

In an effort to offer some form of fault tolerance in FragVisor, we implemented distributed VM Checkpoint/Restart (C/R). We first enabled traditional C/R system in which handles Aggregate VMs in this way: we implemented distributed VM Checkpoint/Restart (C/R). We then extended this BFF scheduler into FragBFF, which handles Aggregate VMs in this way:

(a) When BFF fails to allocate a VM it forwards the allocation to FragBFF. FragBFF searches for the minimum or maximum (2 different policies) amount of nodes with fragmented resources to satisfy the allocation, and starts an Aggregate VM on such nodes – it does not try to achieve resource balancing among nodes, which is something we will address in future work. If not enough resources are available, the scheduler delays the VM placement.

(b) On termination of any VM co-located with at least an Aggregate VM’s slice, FragBFF evaluates if the resources released by the terminated VM can be allocated to one (or more) of such Aggregate VM’s slices, and eventually triggers a FragVisor migration, thus consolidating more VM slices of a VM on a single node. One FragBFF policy favors minimizing the overall cluster fragmentation, while the other minimizes the number of nodes on which an Aggregate VM runs at any time.

(c) When all resources of an Aggregate VM reside on a single node, the VM is given back to the BFF scheduler.

We believe this can be integrated in a production scheduler because open-source ones [56] enable policy extensions, but orchestrators may need rewriting to trigger migrations.

### 7 Evaluation

In Section 2 we already demonstrated that the SLO of software running into an Aggregate VM depends on the type of workload, specifically on the level of sharing – due to the dependency on DSM. Herein, we will further highlight that, while answering (1) if an Aggregate VM can be used to solve fragmentation (at least compared with overcommitment); (2) what are the eventual overheads (to motivate our solution vs delayed-allocation, migration, etc.); (3) how FragVisor, and its mechanisms, quantitatively differ from previous work.

**Testbed.** To evaluate FragVisor, we created Aggregate VMs on a computer cluster composed of multiple identical servers equipped with a Xeon E5-2620 v4 (2.1 GHz) and 32 GB of RAM. Servers are interconnected via an InfiniBand switch using Mellanox Connect-X4 adapters (56 Gbps). Each server runs baremetal Debian Linux Jessie 8.10 and our modified Linux kernel based on vanilla v4.4.137 – the host kernel. VMs are created with enough RAM to satisfy the various workloads they execute and use a ramdisk as the root filesystem, if not indicated differently. Unless stated otherwise, FragVisor VMs use our optimized Linux kernel (also based on v4.4.137) as the guest kernel. In most experiments we vary the number of vCPUs per VM and their distribution among hosts. For all experiments, vCPUs are pinned on pCPUs. We used cgroups for capping resources usage (isolation) between different VMs on each node. However, further performance isolation can be potentially achieved exploiting microarchitectural extensions, such as RDT/CAT on Intel x86 processors [73, 103].

An additional server, acting as a client or a load generator in the experiments, runs stock Linux Ubuntu 16.04.

**Tests.** We use microbenchmarks to evaluate (a) the cost of accessing remote memory and I/O, i.e., memory and device
We noticed that differently from GiantVM, FragVisor does not write in a loop at a configurable memory location in an array.

VM on FragVisor vs. a distributed VM on GiantVM 

by tuning this location and running a single instance of the program that writes to a predefined memory location – there is no synchronization. We consider the following four cases when using 4 vCPUs: (a) no-sharing – each vCPU writes to a different page; (b) low-sharing – two vCPUs write on a page, and other two vCPUs write to another page; (c) moderate-sharing – three vCPUs write on a page, and the fourth one on another page; (d) max-sharing – all vCPUs write to the same page. All threads are using the same experiment in the case of Overcommit, where we targeted cache line granularity instead of page granularity.

Clearly, overcommit executes multiple vCPUs on a single pCPU and the total amount of operations is always constant – the maximum that a pCPU can do. Differently, FragVisor enables vCPUs to run on different pCPUs. Thus, the amount of operations is somewhat proportional to the number of pCPU used, but it suffers from DSM overheads already on Low-sharing. Finally, the maximum total traffic generated (on the 56 Gbps interconnect) by this experiment is 8 MB/s (64 Mbps), for Max-sharing.

I/O Delegation Overhead. Another overhead of FragVisor is network I/O delegation. To evaluate that, we compare the network throughput of a NGINX web server where the worker runs (1) on a vCPU that is local to the host’s virtual switch used to communicate with the client (local I/O) and (2) on a vCPU that is on a remote node (delegated I/O). The client runs ApacheBench [4] on a node on the same 1Gb network – simulating a request from outside the datacenter. It sends 1000 requests with 10 concurrent connections to the web server.

We report the best numbers for GiantVM, which is network I/O delegation. For such real-world workloads, we compare when possible the performance of an Aggregate VM on FragVisor vs. a distributed VM on GiantVM [104] – the state-of-the-art open-source distributed hypervisor, and a single-machine (non-distributed) VM – using kvmtool. Finally, we demonstrate the benefits of (d) our version of Linux antVM’s leverages helper threads in QEMU. Hence, FragVisor enables vCPUs to run on different pCPUs. Thus, the amount of operations is somewhat proportional to the number of pCPU used, but it suffers from DSM overheads already on Low-sharing. Finally, the maximum total traffic generated (on the 56 Gbps interconnect) by this experiment is 8 MB/s (64 Mbps), for Max-sharing.

I/O Delegation Overhead. Another overhead of FragVisor is network I/O delegation. To evaluate that, we compare the network throughput of a NGINX web server where the worker runs (1) on a vCPU that is local to the host’s virtual switch used to communicate with the client (local I/O) and (2) on a vCPU that is on a remote node (delegated I/O). The client runs ApacheBench [4] on a node on the same 1Gb network – simulating a request from outside the datacenter. It sends 1000 requests with 10 concurrent connections to the web server.

by tuning this location and running a single instance of the program that writes to a predefined memory location – there is no synchronization. We consider the following four cases when using 4 vCPUs: (a) no-sharing – each vCPU writes to a different page; (b) low-sharing – two vCPUs write on a page, and other two vCPUs write to another page; (c) moderate-sharing – three vCPUs write on a page, and the fourth one on another page; (d) max-sharing – all vCPUs write to the same page. All threads are using SCHED_OTHER with the highest priority. The results for FragVisor are shown in Figure 5. We repeated the same experiment in the case of Overcommit, where we targeted cache line granularity instead of page granularity.

Clearly, overcommit executes multiple vCPUs on a single pCPU and the total amount of operations is always constant – the maximum that a pCPU can do. Differently, FragVisor enables vCPUs to run on different pCPUs. Thus, the amount of operations is somewhat proportional to the number of pCPU used, but it suffers from DSM overheads already on Low-sharing. Finally, the maximum total traffic generated (on the 56 Gbps interconnect) by this experiment is 8 MB/s (64 Mbps), for Max-sharing.

I/O Delegation Overhead. Another overhead of FragVisor is network I/O delegation. To evaluate that, we compare the network throughput of a NGINX web server where the worker runs (1) on a vCPU that is local to the host’s virtual switch used to communicate with the client (local I/O) and (2) on a vCPU that is on a remote node (delegated I/O). The client runs ApacheBench [4] on a node on the same 1Gb network – simulating a request from outside the datacenter. It sends 1000 requests with 10 concurrent connections to the web server.
We vary the served web page/object size and get request throughput. We present numbers for both delegation mechanisms, i.e., DSM-backed and DSM-bypass.

Results are in Figure 6. On the Y-axis the measured throughput is normalized to that of a vanilla Linux VM. DSM-backed delegated I/O is slow, due to a high DSM contention on the ring buffers used for delegation: the slowdown is on average 53x for request sizes below 4 KB. Increasing the request size helps to amortize the slowdown, however it is still of 19x for a size of 256 KB. Bypassing the DSM leads to a significant performance boost: the slowdown is only of 3.5x for sizes below 4 KB, and becomes negligible when the size increases. The throughput of FragVisor when performing local I/O is also the same as Linux’s. These results demonstrate the efficiency of bypassing the DSM when delegating I/O in a distributed VM.

Similarly, our storage I/O delegation, which has been evaluated with file on SATA III SSD, shows the same behavior. The achievable bandwidth with one thread for local and remote (delegated) requests, with and without DSM, is shown in Figure 7. Interestingly, increasing the number of threads does not change the trends observed on that Figure.

**Checkpointing Performance.** To evaluate the overhead introduced by our distributed checkpointing mechanism, we created several Aggregate VMs varying the amount of RAM (10, 20 and 30 GB) and the number of vCPUs (2, 3 and 4) allocated, each CPU being placed on a physically distinct host. We compared the time to take a checkpoint in FragVisor versus in non-distributed vanilla VMs of similar characteristics. Before taking the checkpoint, to distribute the memory, we run on each vCPU one instance of NPB IS class C (700 MB dataset size). We found the overhead of FragVisor over vanilla is always 10% or less. We observed that the main bottleneck is the disk, a traditional SATA SSD with a 500 MB/s throughput. As a result, accessing remote memory from the checkpointing node in the case of FragVisor does not weight much in the total checkpointing time.

7.2 Real-World Applications

Without an Aggregate VM, a common way to pack more jobs on a saturated but fragmented cluster, with no possibility of VM evictions, is to overcommit resources [102] per-machine. Even Burstable VMs overcommit system resources for limited amount of time. We evaluated FragVisor with real-world applications by comparing the performance of Aggregate VMs on FragVisor versus (1) single-machine VMs with overcommitted vCPUs, and (2) a distributed VM (an Aggregate VM without “mobility” features) running on top of GiantVM [104]. We varied the number of vCPUs to be 2, 3, and 4 – these are the most common VM sizes in data centers [45]. In the case of the distributed VMs (FragVisor and GiantVM), each vCPU runs on a different host. Indeed, Spot VMs and other recent works [12, 102] also exploit idle cluster resources, but they are subject to evictions. We believe such approaches that guarantee only minimal resources can be fairly approximated to overcommitment when no evictions happen.

**Serial HPC Applications (No Sharing).** We ran the NPB suite, selecting for each benchmark a data set size that would result in an execution time of at minimum 10 seconds. For each benchmark and VM type, we run in parallel one instance of the serial version of the benchmark for each vCPU and measure the total execution time of this set of instances.
We found that the performance difference between FragVisor and GiantVM to be due to several factors. GiantVM’s DSM is implemented partially in user-space while ours lives completely in the host kernel, avoiding costly user/kernel transitions. Furthermore, FragVisor benefits of a guest kernel topology mapping of the physical nodes’ distribution.

GiantVM does not implement the mobility features required by an Aggregate VM, but can indeed run an Aggregate VM that doesn’t move – i.e., a bare distributed VM. Figure 9 reports the results of comparing FragVisor to GiantVM. FragVisor is faster, on average by 1.6x, for all benchmarks. More precisely, although for most applications FragVisor is about 1.5x faster than GiantVM independently of the number of vCPUs, for IS and FT the performance difference is higher: FragVisor is on average 2x faster than GiantVM for IS, and 1.8x for FT. We found that the performance difference between FragVisor and GiantVM to be due to several factors. GiantVM’s DSM is implemented partially in user-space while ours lives completely in the host kernel, avoiding costly user/kernel transitions. Furthermore, FragVisor benefits of a guest kernel optimized for distributed execution and exposes a NUMA topology mapping of the physical nodes’ distribution.

Optimized Linux Guest. To measure the performance benefits brought by our optimized version of the Linux kernel, we ran that in a FragVisor Aggregate VM and compared it with a non optimized kernel, normalized to overcommitment on one pCPU (vanilla), using the NPB benchmarks.

Results are in Figure 10. Despite the obvious speedups vs overcommitment, the differences between the optimized and non-optimized (noOpt) guest kernels are minimal. In fact, our optimizations are more beneficial (up to 30% in IS) for workloads with a larger amount of sharing and synchronizations – related to DSM traffic, as the recorded time focuses on compute phase.

Multithreaded HPC Applications. We already showed that the SLO provided by an Aggregate VM to scale-up workloads, especially those with a lot of sharing, cannot be the same as a Primary VM because of the DSM. Here we quantify the performance degradation of HPC shared-memory multithreaded applications [21] when running in an Aggregate VM. We created a VM with 4 vCPUs (most common size other than 2 vCPUs [45]) for FragVisor with and without optimized guest kernel and GiantVM, each vCPU being located on a physically distinct node. We launched the multithreaded version of the PARSEC [84] benchmark suite, setting the number of threads to 4. We compare the performance of FragVisor and GiantVM to that of a vanilla non-distributed VM with 4 vCPUs mapped to 1 pCPU (overcommitment).

Figure 11 presents results. FragVisor can be up to 6.5x and 5.3x (non-optimized and optimized guest kernel) slower than vanilla (Streamcluster), but on average is just 2.6x and 2.3x. For a few benchmarks, Blackscholes, Ferret, and Freqmine, distributed execution is up to 4x faster. FragVisor also performs overall better than GiantVM, whose average slowdown is 4.5x, with a peak of 13.1x. We collected several statistics to understand the different behavior of GiantVM and FragVisor, including number of page faults and messages, messages’ sizes and latencies. GiantVM sends up to 300x more messages than FragVisor (Freqmine), and message sizes are always small, on average of 2 kB vs. 4 kB for FragVisor. This shows the higher efficiency of FragVisor’s distributed protocol.

Network Application: LEMP Stack. LEMP [2] is an open-source web stack consisting of NGINX [6], PHP [7], and MySQL [5]. NGINX acts as a front-end HTTP server receiving requests from clients. PHP is invoked in response to requests, fetching back-end data and performing various processing operations to create web pages. MySQL is a database for storing data for services – it is unused in this experiment.

We run the LEMP stack in an Aggregate VM as follows. First, NGINX is configured with a single worker thread running on the vCPU that is local to the virtual switch used by the VM to access the network (vCPU0), to avoid network...
delegation overheads demonstrated in Section 7.1. Upon receiving a client request, NGINX invokes PHP which runs a worker thread on each vCPU except the one running the NGINX thread. In other words, the 2 vCPUs configuration has 1 NGINX worker on the first vCPU, and 1 PHP worker on the other. Similarly, the 3 vCPUs configuration has 1 NGINX worker on the first vCPU and 2 PHP workers on the others.

We set the served page size to 2 MB, the average page size on the web according to recent statistics [14]. Each request triggers the execution of a PHP benchmarks that we adapted from [1], realizing some string manipulation operations – a common operation in PHP-enabled web servers. By varying the amount of iterations of this benchmark, we can customize the request processing time and observe its impact on the performance. We vary that processing time from 25 ms to 500 ms, which is representative of modern servers’ response times [31], averaging from 200 to 500 ms [72]. ApacheBench (AB) is used as the client, running on the host node that runs the pCPU where the vCPU running the NGINX worker is mapped. AB is configured to make 100 requests with 10 concurrent connections. We collect the reported throughput in requests per second and compare the results of an Aggregate VM on FragVisor, or on GiantVM, to an overcommitment scenario where all vCPUs run on a single pCPU.

The results are presented in Figure 12, where FragVisor’s and GiantVM’s throughputs are normalized to the overcommitment case. With short processing times, our system does not perform well due to expensive communication between NGINX and PHP workers (local socket within the guest), as the communicating processes run on two separate physical machines. Starting from 40 ms of processing per request, the communication overheads become lower compared to the computing time, and are amortized by the benefits of leveraging remote computing resources: the throughput of an Aggregate VM becomes higher than the consolidated case. The speedup increases with both the demand for computation time (processing time) and the number of vCPUs. For example, with 4 vCPUs and a 500 ms processing time, the speedup is of 3.5x.

For fast request processing times, FragVisor is slower than GiantVM: for 25 ms requests, FragVisor’s throughput is on average 35% of GiantVM’s, and for 40 ms that number is 79%.

However, for longer requests, FragVisor’s throughput becomes higher than GiantVM’s by a factor of 1.23x for 250 ms requests and 1.27x for 500 ms ones. This indicates that, although GiantVM remote vCPU communication is faster, which is important for short requests, for longer ones FragVisor is better at exploiting the parallelism brought by these vCPUs.

**Serverless Computing.** Function as a Service (FaaS) computing is an emerging paradigm where users upload and execute small pieces of code, or functions, in the cloud [10, 40, 65].

We used the OpenLambda [47] FaaS runtime to run a typical serverless computing application [38], varying its resources. Providers allocate vCPUs to FaaS runtime based on the requested memory size [101]. In an Aggregate VM running on FragVisor, we run the OpenLambda server configured to spawn functions upon reception of an HTTP request on each available vCPU – on different VM slices. Such functions run the same Python code, which (1) fetches a series of pictures as a compressed file from a database on the same network, (2) extracts the pictures, and (3) runs a face detection algorithm [38], thus, returning to the client the number of detected faces.

A client, running on another machine, triggers function’s execution. We vary the number of parallel requests to be equal to the number of vCPUs: 2, 3 and 4. Other than total time, we break down the server-side execution times into: pictures download, compressed file extraction, and face detection.

Results for an Aggregate VM on FragVisor and a distributed VM on Giant VM are shown in Figure 13. They are normalized to overcommitting where 2, 3 and 4 vCPUs are consolidated on the same pCPU. When overcommitting, the speedup for the overall operation increases with the number of vCPUs because the overcommitted VM needs to run more processing on the same computing resources (1 core). Regarding extraction time, even if there are write operations to different pages, the first write to a new region on a remote node always causes DSM write-exclusive invalidation messages, contributing to the slowdown of the extraction as the number of vCPU increases. The face detection phase is considerably faster (up to 3.3x for 4vCPU) with FragVisor. As it dominates the execution time of the entire operation, the overall performance outperforms the overcommitting cases by 1.9x to 3.26x from 2 to 4 vCPUs. Comparing to GiantVM, FragVisor is always faster on every phase, in particular the download one, up to
13x with 4vCPUs. This leads to a speedup from 2.17x (2vCPU) to 2.64x (4vCPU) for the whole operation. FragVisor proved to be faster not only for its kernel-space DSM, but also for IO device multiqueue and DSM-bypass.

7.3 Scheduling-driven Migration

While the previous experiments demonstrate virtualized memory and devices mobility, here we showcase the cost of vCPU mobility. FragVisor’s vCPU migration feature is useful when resources free up on at least one host running part of an Aggregate VM, allowing that VM to be consolidated on a smaller number of nodes for higher performance. Migration is also important for reliability when a failure can be predicted.

To demonstrate migration we set up a cluster of 4 servers, each with 12 CPUs available for VMs (other 4 left for management tasks). We adopted VM sizes and VM execution times from [45], scaled down by 100 to ease experiments. With such a distribution we generated several bursts of 100 arrivals that we fed into the cluster scheduler, which starts and migrates FragVisor’s Aggregate VMs. Aggregate VMs run a web server on vCPU0 and the previously described PHP benchmark on the other vCPUs. A client on the same network sends 1000 requests to the server and measures the latency of each request. While requests are generated, the vCPUs of the Aggregate VM migrate.

We picked a trace of a 4 vCPUs VM, which is easier to visualize than an 8 vCPUs VM trace, while showing enough migrations. This is depicted in Figure 14. The top graph reports the perceived client latency, which is the lowest (∼215 ms) when vCPUs are consolidated on a single machine, while being 299 ms on average during the experiment. The middle graph reports how many vCPUs from each node the Aggregate VM is using. The bottom graph reports how many free vCPUs are available per machine, which shows the effectiveness of our simple algorithm that reduces resource fragmentation. We also recorded the inter-node migration overhead of a vCPU: it is 86 μs on average – including 38 μs to dump registers.

Scheduling Decisions. For the experiment in Figure 14 FragBFF was configured to minimize overall fragmentation, and an Aggregate VM is released at time 155 on a minimal number of nodes. 2 vCPUs are placed on node0 and 2 vCPUs on node1.

At time 222, 4 CPUs become free on node1. FragBFF won’t migrate all Aggregate VM’s vCPUs to node1 because that may undermine overall cluster fragmentation, i.e., the current Aggregate VM will be consolidated but a newly arrived VM may likely be split across nodes again. At time 360, similarly to time 222, the scheduler favors reducing overall fragmentation.

At time 470, 1 CPU becomes free on node1. Full consolidation is not possible yet, since only one vCPU of the Aggregate VM can be moved to node0 and two are needed for full consolidation. FragBFF migrates a vCPU of the Aggregate VM and consolidate with the ones already on node0. This is again to reduce fragmentation: it results in more free CPUs on node1 (8 free) while utilizing the only free CPU on node0.

At time 623, one more CPU becomes free on node0, and full consolidation is now possible for the Aggregate VM. The last vCPU on node1 is then migrated to node0. There are now 12 free CPUs on node1, which are used to satisfy a 12 vCPU VM request. Otherwise, such large VM wouldn’t run yet, or would have unnecessarily paid the possible overhead(s) of running in an Aggregate VM.

8 Related Work

Resource Fragmentation in the Data Center. Harvest VM [12] and Elastic VM [102] grow and shrink the amount of hardware resources based on what is/may become available on a single-machine, helping to tackle fragmentation. Differently, an Aggregate VM extends to multiple machines. Similarly to Spot VM [11, 41, 66], Harvest VM and Elastic VM may be terminated at any time, making them practical only for a subset of applications. Instead, an Aggregate VM is never evicted, and always provides the requested resources, similarly to a Primary VM, but with a SLO depending on the workload.

Earlier, StopGap [75] aimed at solving fragmentation, introducing VMs that can be dynamically resized. It supports only certain types of applications, elastic multi-tier master-slave applications, while FragVisor supports all applications. Recently, A. Fuerst et al. [36] also proposed not to evict transient VMs, but to deflate them. While sharing the same goal with FragVisor, this is a different approach: deflating VMs adds overhead to the execution of every workload, and may require modifications to the guest OS.

Distributed Virtual Machines. Distributed VMs have been proposed before: vNUMA [27], ScaleMP [8], TidalScale [96], and GiantVM [104]. Fundamentally different from FragVisor, such works target scale-up workloads: they aggregate all resources in a computer cluster to execute one or a few VMs with resource requirements greater than what provided by a
single machine. Conversely, FragVisor aggregates only a subset of hosts’ physical resources for the execution of VMs with applications’ resources requirements that can be satisfied by a single machine. Moreover, while in related works a VM distribution is mostly static, in FragVisor the distributed state of a VM aims to be temporary for it to be consolidated upon as few hosts as possible once resources free up. Hence, the resource borrowing hypervisor focuses on mobility features, e.g., vCPU migration, missing in vNUMA, ScaleMP, and GiantVM (while all provide vRAM migration via DSM). These do not support checkpointing either (for consolidation or fault tolerance). TidalScale provides vCPU migration, but for a different reason. No previous work seems to provide device mobility.

Finally, vNUMA [27] and ScaleMP [8] are Type 1 hypervisors, while GiantVM [104] and TidalScale [96] are Type 2 like FragVisor. Differently from [8, 27, 96], FragVisor is free and open-source, built atop the widespread KVM hypervisor. When compared to [104], FragVisor is mainly implemented in kernel-space, and provides Aggregate VMs as a first class VM.

**Distributed OSES.** Distributed OSES aggregates resources from different physical machines while still providing the same OS interface, and scaling beyond a single physical machine. Notable past works include MOSIX [15], Amoeba [68]. More recently, Helios [74], and Popcorn [17, 19, 78]. Contrary to FragVisor (and GiantVM), these works allow an application, rather than a VM, to use remote resources.

**Distributed Shared Memory.** Several distributed OSES leverage DSM in order to present to the application a unified virtual address space. Similarly, FragVisor uses DSM for vRAM mobility and gives the VM a unified pseudo-physical address space. Much work studied DSM during the 1990s [13, 26, 85], with the overall objective of letting traditional/legacy shared memory applications scale over several physical machines. Recent work leverages modern interconnect technologies, i.e., RDMA, in DSM systems [25, 71, 92]. FragVisor introduces several new optimizations to make the nodes involved in DSM changing dynamically (mobility), vs statically defined.

**Disaggregated Computing.** FragVisor takes an approach different from existing works on hardware disaggregated computing [37, 46, 61, 62, 91, 95]. Indeed, our objective is the aggregation of scattered hardware resources in data centers, rather than the abstraction of future disaggregated hardware. In addition, contrary to existing works on disaggregation, we target commodity servers and our solution is directly deployable today, at no capital cost. In the future, FragVisor can complement these approaches. For example, when coherent shared memory will become available among servers — provided by CXL pooled memory [3] among other technologies, that may solve memory resource fragmentation, as well as substitute FragVisor’s DSM, but cannot solve the CPU resource fragmentation issue, which requires the other mechanisms introduced by FragVisor.

## 9 Remarks

We already highlight the key limitation of the proposed Aggregate VM approach, which is its general applicability, i.e., not all workloads may benefit from it. While this is a fact considering today’s data center hardware, upcoming CXL memory pools with hardware coherency may remove all DSM-related overheads making Aggregate VM beneficial for any workload. At the same time, data center schedulers should be extended to handle Aggregate VMs, and further research on scheduling algorithms is needed.

**Prototype Limitations.** FragVisor inherits the same limitations as Popcorn Linux for homogeneous cluster on which it was developed upon [17, 55]. FragVisor has been fully tested only on the hardware used for evaluation, and up to 8 nodes.

## 10 Conclusions

For certain workloads, an Aggregate VM can be used to tackle resource fragmentation issues in data centers without reducing the promised hardware resources or evicting VMs, i.e., with a SLO similar to Primary VMs. In general, the SLO of an Aggregate VM will depend on its memory-sharing characteristics.

We introduced a new VM monitor design, the resource-borrowing hypervisor, which leverages fragmented hardware resources available among different physical machines, transparently, and temporarily. This temporary nature is achieved through mobility of virtualized memory, CPUs, and devices. We implemented FragVisor, a prototype resource-borrowing hypervisor based on Linux/KVM, and compared the performance of an Aggregate VM on FragVisor to a distributed VM on the state-of-the-art distributed hypervisor, GiantVM, and overcommitment. FragVisor is faster than GiantVM in most of the cases, and demonstrates significant speedups versus overcommitment.


## Acknowledgments

The authors would like to thank the anonymous reviewers and their shepherd, Redha Gouicem, for their valuable comments and helpful suggestions.

H-R. Chuang, B. Heerekar, and B. Ravindran’s work was supported by the US Office of Naval Research (ONR) under grants N00014-13-1-0317, N00014-16-1-2104, N00014-16-1-2711, N00014-18-1-2022, and N00014-19-1-2493, and by the US Air Force Office of Scientific Research (AFOSR) under grants FA9550-14-1-0163, and FA9550-16-1-0371. T. Xing and A. Barbalace’s work was partially supported by the US Office of Naval Research (ONR) under grants N00014-13-1-0317, N00014-16-1-2104, N00014-16-1-2711, and N00014-19-1-2493. P. Olivier’s work was partially supported by the UK’s EPSRC under grant EP/V01234/1 (UniFaaS), and by the EPSRC/Innovate UK under grant EP/X015610/1 (FlexCap).
References


A Artifact Appendix

A.1 Abstract

The artifact contains the source code of FragVisor’s host and guest Linux kernels, FragVisor hypervisor based on kvmtool (lkvm), and a plethora of helper scripts. The system has been deployed on baremetal hardware on a cluster of RDMA-connected servers. For reproduction, at least 5 servers are needed – 4 acting as compute nodes echo, 1 acting as client/load-generator fox.

A.2 Description & Requirements

Fragvisor mainly consists of three software modules. (1) The distributed hypervisor built on top of kvmtool (lkvm)\(^2\), a KVM-only, type-2 hypervisor. (2) The host Linux kernel, heavily modified to support FragVisor. (3) The guest Linux kernel, which is a vanilla kernel, but an additional version that enables certain optimisations to improve performance is also made available. In addition, a plethora of helper scripts automate experiments, and configure and build the infrastructure.

A.2.1 kvmtool Hypervisor

The source code of the modified kvmtool is available at https://github.com/systems-nuts/fragvisor-kvmtool. Many files have been modified to support VM distribution, mostly using the thread migration API to replicate the VM object across all the hosts.

The repository also contains initramfs’ images, and configure, build, and run scripts. Two notable scripts are msg_pophype4node.sh, which establishes the messaging layer between hosts, and run.sh used to build and start the guest VM.

A.2.2 Host and Guest OS Kernels

Host and Guest kernel use the same source tree at https://github.com/systems-nuts/fragvisor-linux. The host kernel is heavily modified. Besides the task migration code and the distributed shared memory protocol in kernel/popta/, the main FragVisor extensions is in kernel/popta/hype_kvm.c, as well as a plethora of other modifications in the kernel virtualisation subsystem – i.e., KVM, both architecture-independent parts as well as the x86 architecture-specific portions in arch/x86/kvm/.

A.3 How to access

The source code of all FragVisor components is available either at https://github.com/systems-nuts/fragvisor/ or at https://doi.org/10.5281/zenodo.7725802.

\(^2\)https://github.com/kvmtool/kvmtool
A.4 Hardware Dependencies

FragVisor is deployed on a cluster of computers with multiple identical servers equipped with a Xeon E5-2620 v4 (2.1 GHz, 8/16 cores/threads) and 32 GB of RAM. Servers are connected via 56Gbps InfiniBand using Mellanox Connect-X4 and an InfiniBand switch. All nodes are also connected via 1Gbps Ethernet switch. The client communicates via 1Gbps Ethen.

A.5 Software Dependencies

Hosts Linux is Debian Jessie 8.10, guest Linux is based on the same, but generated from what created by mknitramfs. Client Linux is Ubuntu 16.04.

A.6 Benchmarks

The benchmarks used are based on NPB, ApacheBench, LEMP and OpenLambda (more on this later).

A.7 Set-up

For a full guide including how to use IPMI serial console, refer to the README file on the github repository.

First, we need to compile and install the host kernel on every node of the cluster not used as client/load-generator. For each node in the echo cluster (echo5, echo4, echo1, echo0), run:

```
git clone --recursive
~submodules git@github.com:systems-nuts/FragVisor
```

```
cd FragVisor/fragvisor
```

Before building the kernels, make sure that the macro CONFIG_POPCORN_ORIGIN_NODE in include/linux/popcorn/debug.h is only defined for the origin node. All the other nodes must NOT define it and it should be commented. Then, run:

```
make -j17
sudo make modules_install
sudo make install
```

Next, load the messaging layer. From echo0 or echo5, run:

```
cd FragVisor/fragvisor - kvmtool
/msg_pophype4node.sh
```

This will compile and load the messaging layer kernel modules. If successful, you should observe on the IPMI serial console that RDMA connection was established between all nodes.

A.8 Evaluation Workflow

To run experiments, navigate to fragvisor-kvmtolool, we use the run.sh script. Every experiment will use a different kvmtolool launch command and initramfs image. Commands starting with root@(none) have to be issued in the guest VM.

A.9 Major Claims

- (C1): FragVisor achieves higher throughput for longer requests than the state-of-the-art GiantVM. This is proven by the experiment (E1) described in Section 7.2 whose results are illustrated in Figure 12 in the paper.

- (C2): FragVisor performs faster than the state-of-the-art GiantVM in every phase of OpenLambda serverless computing. This is proven by the experiment (E2) described in Section 7.2 whose results are illustrated in Figure 13 in the paper.

- (C3): With DSM bypass, FragVisor I/O delegation can achieve very high performance to offset the effects of distribution. This is proven by the experiment (E3) described in Section 7.1 whose results are illustrated in Figures 6 and 7 for networking and storage, respectively.

A.10 Experiments

E1. Running LEMP Experiments

For 4 nodes, use the following kvmtolool command:

```
sudo bash -c "./kvmtolool run -a 1
-b 1 -x 1 -y 1 -w 4 -i $USER/c/ramdisk_lemp_4php.gz
-k $KERNEL_PATH/arch/x86/boot/bzImage -m 20480
-c 4 -p '/root=dev/ram rw init=/init2 fstype=ext4
spectre_v2=off npti=off numa=fake=4 percpu_alloc
page --no-kvm-pit-rejection clocksource=ts
ec21 --network mode=tap,vhost=1,guest_ip=10.4.4.222,
host_ip=10.4.4.221,guest_mac=00:11:22:33:44:55"
```

Then follow with

```
echo55 $ cd FragVisor/experiments/nginx-new/nginx-1.16.1
```

```
echo55 ./auto_configure_make_scp.sh
```

```
root@none:-# nginx -c /usr/local/nginx/conf/nginx.conf
```

```
root@none:-# php-fpm7
```

```
.2 --fpm-config /etc/php/7.2/fpm/php-fpm.conf &
```

In the VM, there are N number of php-fpm, php worker threads, where N is the number of nodes. Each of these threads have to be pinned to a vCPU.

```
root@none:-# ps aux | grep php
```

```
# Pin each of the php workers on a vCPU!
```

```
root@none:-# taskset -p 0x1 312
```

```
root@none:-# taskset -p 0x2 313 (Stop here for 2 Nodes)
```

```
root@none:-# taskset -p 0x4 314 (Stop here for 3 Nodes)
```

```
root@none:-# taskset -p 0x8 315
```

```
# Pin nginx master and worker on vCPU0
```

```
root@none:-# ps aux | grep nginx
```

```
root@none:-# taskset -p 0x1 112
```

```
root@none:-# taskset -p 0x1 120
```

The customized LEMP experiment uses a php-benchmark script with variable iterations of each benchmark. There are 4 variations of this which have to be copied to the VM from the host node. Copy the scripts to VM before the run:

```
echo55 scp -r FragVisor/experiments/php
```

```
--script/*.root10.4.4.222:var/www/travel_list/routes/
root@none:-# cd /var/www/travel_list/routes
```

```
echo55 cd FragVisor/experiments/run_lemp
```

```
echo55 .run_lemp_no_stats.sh
```

```
# Check output
```

```
echo55 cd run_lemp_no_stats/output_dir
```

```
echo55 time_taken_out
```

6https://github.com/systems-nuts/FragVisor/blob/main/README.md
E2. Running OpenLambda Experiments

For a 4 nodes configuration, use this kvmtool command:

```
sudo bash -c './lkvm run -a 1 -b 1 -x 1
   -y 1 -w 4 -i $USER/c/ramdisk_for_openlambda200625
   -g 2 -k $KERNEL_PATH/arch/x86/boot/bzImage
   -m 32768 -c 4 -p `/root=/dev/ram rw fstype=ext4
   specver_v2=off nopti p ti=off numa=false=4 percpu_alloc
   =page --no-kvm -pit -reinjection clocksource=
   tsc` --network mode=tap,vhost=1,guest_ip=10.4.4.222,
   host_ip=10.4.4.221,guest_mac=00:11:22:33:44:55
```

Type the instructions below to run the experiment:

```
echo5:$ cd FragVisor/experiments/open-lambda/
echo5:$ ./copy4nodes.sh # passwd: popcorn
echo5:$ ssh popcorn@10.4.4.222
   root@(none):~# sudo su
   root@(none):/home/popcorn/$ ./folders4nodes.sh
   root@(none):/home/popcorn/$ cd ~/open-lambda0 && taskset 0x1 ./ol worker
   --path=l0 --d
   root@(none):~# cd ~/open-lambda0$ cd ~/open-lambda1 && taskset 0
   x2 ./ol worker --path=l1 --d # Stop here for 2 node setup
   root@(none):~# cd ~/open-lambda0$ cd ~/open-lambda2 && taskset 0
   x4 ./ol worker --path=l2 --d # Stop here for 3 node setup
   root@(none):~# cd ~/open-lambda0$ cd ~/open-lambda3 && taskset 0x8
   root@(none):~# cd ~/open-lambda3 && taskset 0x8 ./ol worker
   --path=l3 --d
echo5:$ cd FragVisor/experiments/run_openlambda
   echo5:$ /run_openlambda_no_stat_4_node.sh
```

E3. Running Network Delegation Overhead Experiment

For 4 node case choose this in run.sh:

```
sudo bash -c './lkvm run -a 1
   -b 1 -x 1 -y 1 -w 4 -i $USER/c/ramdisk_NPB_ATC.gz
   -k $KERNEL_PATH/arch/x86/boot/bzImage
   -m 32768 -c 4 -p `/root=/dev/ram rw init=/init2
   fstype=ext4
   specver_v2=off nopti p ti=off numa=false=4 percpu_alloc
   =page --no-kvm -pit -reinjection clocksource=
   tsc` --network mode=tap,vhost=1,guest_ip=10.4.4.222,
   host_ip=10.4.4.221,guest_mac=00:11:22:33:44:55
```

Compile kvmtool and the guest kernel, and launch the VM:

```
echo5:$ fragvisor_kvmtool$ ./run_echo.sh 1 1 0
```

The following are the commands for running the network delegation experiment (note that fox3 is the client):

```
root@(none):~# ps aux | grep nginx
# Pin the nginx worker and master on vCPU 0
root@(none):~# taskset -p 0x1 112 #(master)
root@(none):~# taskset -p 0x2 120 #(worker)
fox3:$ cd FragVisor/experiments/popcorn-utils
fox3:$ ./ab_micro_diff_sizes_lan_0x2_all.sh <name for the run>
fox3:$ cd pophype_ab_micro_diff_sizes_lan/
```