The *nom* Profit-Maximizing Operating System

Shmuel (Muli) Ben-Yehuda

The *nom* Profit-Maximizing Operating System

Research Thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science

Shmuel (Muli) Ben-Yehuda

Submitted to the Senate
of the Technion — Israel Institute of Technology
Iyar 5775 Haifa May 2015

This research thesis was done under the supervision of Prof. Dan Tsafrir in the Computer Science Department.

Some results in this thesis as well as results this thesis builds on have been published as articles by the author and research collaborators in conferences and journals during the course of the author's master's research period. The most up-to-date versions of these articles are:

Orna Agmon Ben-Yehuda, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. The rise of RaaS: The Resource-as-a-Service cloud. *Communications of the ACM (CACM)*, 57(7):76–84, July 2014.

Nadav Amit, Muli Ben-Yehuda, Dan Tsafrir, and Assaf Schuster. vIOMMU: efficient IOMMU emulation. In *USENIX Annual Technical Conference (ATC)*, 2011.

Orna Agmon Ben-Yehuda, Eyal Posener, Muli Ben-Yehuda, Assaf Schuster, and Ahuva Mu'alem. Ginseng: Market-driven memory allocation. In *ACM/USENIX International Conference on Virtual Execution Environments (VEE)*. 2014.

Orna Agmon Ben-Yehuda, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. Deconstructing Amazon EC2 spot instance pricing. *ACM Transactions on Economics and Computation (TEAC)*, 1(3):16:1, September 2013.

Muli Ben-Yehuda, Omer Peleg, Orna Agmon Ben-Yehuda, Igor Smolyar, and Dan Tsafrir. The nonkernel: A kernel designed for the cloud. In *Asia Pacific Workshop on Systems (APSYS)*, 2013.

Abel Gordon, Nadav Amit, Nadav Har'El, Muli Ben-Yehuda, Alex Landau, Dan Tsafrir, and Assaf Schuster. ELI: Bare-metal performance for I/O virtualization. In *ACM Architectural Support for Programming Languages & Operating Systems (ASPLOS)*, 2012.

Michael Hines, Abel Gordon, Marcio Silva, Dilma Da Silva, Kyung Dong Ryu, and Muli Ben-Yehuda. Applications know best: Performance-driven memory overcommit with Ginkgo. In *IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, 2011.

ACKNOWLEDGEMENTS

First and foremost, I'd like to thank my amazing wife, friend, co-author, and advisor, Orna Agmon Ben-Yehuda. You taught me more than you will ever know. Second, I'd like to thank my amazing children, Yael and Ze'ev, who make it all worthwhile. Third, I'd like to thank my parents, Yoel and Irit Ben Yehuda, for having kept faith all these years, even when my path meandered. Last, I'd like to thank Michael Factor and Orran Krieger, who taught me what it means to do research.

The nom operating system and this thesis have been in the making for a long time. During the years I worked on them, I published nearly twenty papers co-authored with many wonderful people. I'd like to thank all of them—it has been great working with you!

The generous financial support of the Technion is gratefully acknowledged.

Contents

List of Figures

Abstract

In the near future, cloud providers will sell their users virtual machines with CPU, memory, network, and storage resources whose prices constantly change according to market-driven supply and demand conditions. Running traditional operating systems in these virtual machines is a poor fit: traditional operating systems are not aware of changing resource prices and their sole aim is to maximize performance with no consideration of costs. Consequently, they yield low profits.

We present nom, a profit-maximizing operating system designed for cloud computing platforms with dynamic resource prices. Applications running on nom aim to maximize profits by optimizing for both performance and resource costs. The nom kernel provides them with direct access to the underlying hardware and full control over their private software stacks. Since nom applications know there is no single "best" software stack, they adapt their stacks' behavior on the fly according to the current price of available resources and their private valuations of them. We show that in addition to achieving up to 3.9x better throughput and up to 9.1x better latency, nom applications yield up to 11.1x higher profits when compared with the same applications running on Linux and OSv.

"And in this too profit begets profit."

—Aeschylus

Abbreviations and Notations

IaaS	:	Infrastructure-as-a-Service
RaaS	:	Resource-as-a-Service
NIC	:	Network Interface Card
PIO	:	Programmed I/O
MMIO	:	Memory-Mapped I/O
DMA	:	Direct Memory Access
SLA	:	Service Level Agreement
SLO	:	Service Level Objective

Chapter 1

Introduction

More and more of the world's computing workloads run in virtual machines on Infrastructure-as-a-Service (IaaS) clouds. Often these workloads are network-intensive applications, such as web servers or key-value stores, that serve their own third-party users. Each application owner charges the application's users for the service the application provides, thereby generating revenue. The application owner also pays her cloud provider for the virtual machine in which the application runs, thereby incurring expenses. The difference between the application owner's revenue and her expenses—and the focus of this work—is the application owner's profit, which she would naturally like to maximize. We depict this cloud economic model in Figure 1.1.

The application owner's revenue depends on her application's performance. For example, the more simultaneous users the application can serve, the higher the revenue it generates. The application owner's expenses, on the other hand, depend on how much she pays the cloud provider. Today's IaaS cloud providers usually charge application owners a fixed sum per virtual machine that does not depend on market conditions. In previous work, we showed that the economic trends and market forces acting on today's IaaS clouds will cause them to evolve into Resource-as-a-Service (RaaS) clouds, where CPU, memory, network, and storage resources have constantly changing market-driven prices [8, 9, 10]. In RaaS clouds, the cloud providers will charge the application owners the current dynamic market prices of the resources they use.

IaaS clouds, and to a larger extent, RaaS clouds, represent a fundamentally new way of

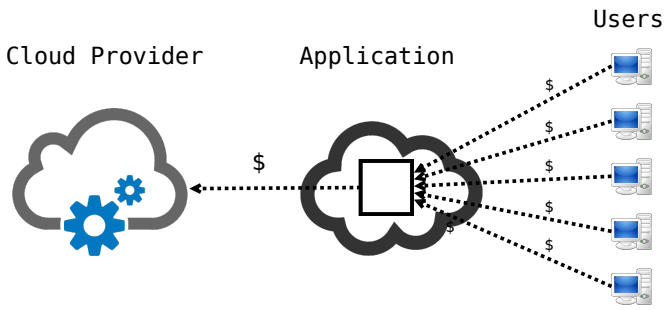

Figure 1.1: Cloud economic model: Applications run in the cloud. Users pay the application owner for the service the application provides. The application owner in turn pays the cloud provider for the cloud resources the application uses (e.g., network bandwidth).

buying, selling, and using computing resources. Nevertheless, nearly all virtual machines running in today's clouds run the same legacy operating systems that previously ran on bare-metal servers. These operating systems were designed for the hardware available decades ago. They abstract away the underlying hardware from their applications and assume that every resource is at their disposal at no cost. Most importantly, they were designed solely for maximizing performance with no regard for costs. They neither know nor care that the resources they use in the cloud cost money, and that their prices might change, e.g., due to changes in supply and demand.

We argue that in clouds with dynamic pricing, where costs can be substantial and resource prices constantly change, running operating systems designed solely for performance is counterproductive and may lead to lower profits and even net losses. Such clouds call instead for a *profit-maximizing* operating system, designed to yield maximal profit by optimizing for both performance and cost. Maximal profit is reached not when revenue (performance) is highest but rather when the difference between revenue (performance) and expenses (cost) is highest. As such, profit-maximizing operating systems enable their applications to pick the right level of performance to operate at given current market conditions and resource prices. We show that applications running on a profit-maximizing operating system can yield an order of magnitude higher profit when compared with the same applications running on operating systems that optimize for performance exclusively.

We begin by presenting in greater depth the motivation for a profit-maximizing operating system. In Chapter 2, we present two ongoing trends that we believe will cause today's IaaS clouds to evolve into RaaS clouds with dynamic resource pricing. They are the increasingly finer spatial granularity and the increasingly finer temporal granularity of resources that can be allocated to guest virtual machines. We then present the changes that such clouds mandate in the system software stack.

In Chapter 3, we present nom, a profit-maximizing operating system we designed for clouds with dynamic pricing. Applications running on nom aim to maximize their profits from the resources available to them. We describe how nom's approach to CPU allocation and scheduling, application memory management, private and direct access to I/O devices, and cost-aware design, can all contribute to maximizing application profits by improving performance and reducing costs.

We showcase and evaluate nom's capabilities using network-intensive applications. We present three main applications, the `memcached` in-memory key-value store [28], the `nhttpd` web server, and the `NetPIPE` networking benchmark [65]. The performance of a network-intensive application is usually expressed through its throughput, latency, and jitter. The expenses the application incurs depend on the amount of bandwidth it uses (i.e., its throughput) and the current price of network bandwidth. Since the price of bandwidth is beyond the application's control, the application can only maximize its profits by controlling its throughput, which affects both revenue and expenses, and the latency and jitter its users experience, which affect its revenue.

In Chapter 4, we use utility (valuation) functions to formalize the relationship between

application throughput, latency, jitter, and the cost of network bandwidth. An application's valuation function provides the application's expected profit from a certain mix of throughput, latency, and jitter, give the current price of network bandwidth and the load the application is under. For example, the simplified valuation function in Equation (1.1) is a formalization of the scenario where the application owner benefits from increased throughput (T), but only as long as the application's users' average latency is below a certain latency service level objective (SLO) and the price the application owner pays her cloud provider (P) per bandwidth unit is lower than her benefit from that bandwidth unit (α).

$$\text{profit} = \begin{cases} T \cdot (\alpha - P) & \text{latency} \leq \text{latency SLO} \\ 0 & \text{latency} > \text{latency SLO} \end{cases} \tag{1.1}$$

We consider three potential valuation functions that differ in how the application's users pay for the service the application provides to them. We acknowledge that building valuation functions is hard, but we believe it is worthwhile to do so in light of the substantially higher profits it yields.

Our profit-maximizing applications re-evaluate their valuation functions at runtime whenever the price of bandwidth or the load they are under change, picking each time the mix of throughput, latency, and jitter that maximizes their valuation function at that point in time. To enable each nom application to have fine-grained control over its throughput, latency, and jitter, nom provides each application with direct access to the virtual or physical NICs the application uses and with a private TCP/IP stack and network device drivers, linked into the application's address space. Each application can control its private stack's throughput, latency, and jitter, by modifying the stack's batching delay: the amount of time the stack delays incoming or outgoing packets in order to batch them together. Larger batching delays increase throughput (up to a limit) while also increasing latency and jitter. Smaller batching delays reduce latency and jitter but also reduce throughput. In nom, there is no "best" TCP/IP stack or "best" NIC device driver as in other operating systems, because there is no single stack or driver that will always provide the right mix of throughput, latency, and jitter, to every application at any given time.

We discuss the implementation of our nom prototype in Chapter 5 and evaluate it in Chapter 6. We show that nom's `memcached`, `nhttpd`, and `NetPIPE` outearn as well as outperform the same applications running on Linux and on the OSv single-application cloud operating system [45]. When running on nom, our benchmark applications yield up to 11.1x higher profits from their resources while also achieving up to 3.9x better throughput and up to 9.1x better latency.

In Chapter 7 we discuss the pros and cons of writing a new profit-maximizing operating system from scratch vs. constructing it based on an existing operating system such as Linux. In Chapter 8 we survey related work and in Chapter 9 we summarize the lessons we have learned building nom and the challenges that remain.

Chapter 2

Motivation

2.1 Dynamic resource pricing is coming

We have identified in previous work [8, 9] two important trends that we believe will lead to RaaS clouds, where different resources have constantly changing prices. These trends are already apparent in current IaaS clouds and their underlying hardware. They are the increasingly finer *spatial* granularity of resources that can be allocated to guest virtual machines and the increasingly finer *temporal* granularity in which resources can be allocated.

Both trends can be seen all the way down to the hardware. Intel Resource Director Technology, for example, enables cloud providers to monitor each virtual machine's CPU cache utilization and allocate specific cache ways to selected virtual machines [3]. Mellanox ConnectX2 and later NICs enable cloud providers to allocate adapter network bandwidth to up to 16 virtual machines and adapt the allocation in microsecond granularity.

Although most IaaS cloud providers today do not (yet) take advantage of such capabilities, they already provide limited dynamic pricing and are moving towards fully dynamic resource pricing. VMTurbo, for example, manufactures a private-cloud management layer that relies on resource pricing and an economic engine to control ongoing resource consumption. CloudSigma's pricing algorithm allows pay-as-you-go burst pricing that changes over time depending on how busy their cloud is; this algorithm prices CPU, RAM, and outgoing network bandwidth separately. Perhaps most notably, Amazon's EC2 spot instances have a dynamic market-driven price [7] that changes every few minutes.

Why are cloud providers going in this direction? Is it not simpler for everyone to just keep the price fixed? By frequently changing the price of different resources based on available supply and demand, cloud providers can communicate resource pressure to their clients (the applications/application owners) and influence their demand for these resources. By conveying resource pressure to clients, cloud providers incentivize their clients to economize when needed and consume less of the high-demand resources. By causing clients to economize, the cloud provider can improve machine density and run more client virtual machines on the same hardware and with the same power budget. Higher machine density means lower expenses, increased profits, and better competitiveness. Improving profit margins by doing more work with

the same hardware is especially important given the cloud price wars that have been ongoing since 2012 [9].

2.2 Dynamic pricing mandates change

A cloud with market-driven per-resource pricing differs from the traditional bare-metal platform in several important areas: resource ownership, economic model, and architectural support. These differences motivate changing the system software stack, and in particular, the operating systems and applications.

Resource ownership and control. On a traditional bare-metal server, the operating system is the sole owner of every resource. If the operating system does not use a resource, nobody else will. In a dynamic pricing cloud, the operating system (running in a virtual machine) unwittingly shares a physical server with other operating systems running in other virtual machines; it neither owns nor controls physical resources.

Economic model. In the cloud, each operating system owner (cloud user) and cloud provider constitute a separate, selfish economic entity. Every resource that the cloud provider makes available to users has an associated price. Each user may have a different incentive, different metrics she may want to optimize, and different valuations for available resources. The cloud provider may want to price its resources to maximize the provider's revenue or the users' aggregate satisfaction (social welfare) [10]; one cloud user may want to pay as little as possible for a given amount of work carried out by its virtual machines; another cloud user may want to maximize the work carried out, sparing no expense. But in all cases, in the cloud, the user pays the current going rate for the resources her operating system uses. On a traditional server, resources are simply there to be used at no cost.

Resource granularity. On a traditional server, the operating system manages entire resources: all cores, all of RAM, all available devices. In the cloud, the operating system will manage resources in an increasingly finer-grained granularity. This is a consequence of the economic model: once resources have prices attached to them, it is more efficient for both cloud provider and cloud users to be able to buy, sell, or rent resources on increasingly finer scales [8].

Architectural support. Operating systems running on traditional servers usually strive to support both the ancient and the modern. Linux, for example, only recently dropped support for the original Intel 386. Modern x86 cloud servers have extensive support for machine virtualization at the CPU, MMU, chipset, and I/O device level [66]. Modern I/O devices are natively sharable [57]. Furthermore, cloud servers usually present the operating systems running in virtual machines with a small subset of *virtual* devices. We contend that any new operating system designed for the cloud should eschew legacy support and take full advantage of the virtual and physical hardware available on modern servers.

Chapter 3

Design

3.1 Requirements

Given the fundamental differences between the traditional bare-metal and the cloud run time platforms, we now ask: What requirements should be imposed on an operating system designed for running in virtual machines on cloud servers with dynamic pricing?

Maximize profit. The first requirement is to enable applications to maximize their profit. When resources are free, applications only have an incentive to optimize for performance. Performance is usually measured in some application specific metric, e.g., in cache hits per second for an in-memory cache or in transactions per second for a database. In the cloud, where any work carried out requires paying for resources and every resource has a price that changes over time, applications would still like to optimize for performance but now they are also incentivized to optimize for cost. Why pay the cloud provider more when you could pay less for the same performance? Thus the operating system should enable its applications to maximize their profits by enabling them to optimize for both performance and cost.

Expose resources. On a traditional server, the operating system's kernel serves multiple roles: it abstracts and multiplexes the underlying hardware, it serves as a library of useful functionality (e.g., file systems, network stacks), and it isolates applications from one another while letting them share resources. This comes at a price: applications must access their resources through the kernel, incurring run-time overhead; the kernel manages their resources in a one-size-fits-all manner; and the functionality the kernel provides, "good enough" for many applications, is far from optimal for any specific application.

In clouds with dynamic pricing, the kernel should get out of the way and let applications manage their resources directly. Moving the kernel out of the way has several important advantages: first, applications become elastic. They can decide when and how much of each resource to use depending on its current price, thereby trading off cost with performance, or trading off the use of a momentarily expensive resource with a momentarily cheap one. For example, when memory is expensive, one application might use less memory but more bandwidth while another might use less memory but more CPU cycles. Second, applications know best how to use the resources they have [26, 37, 32]. An application knows what paging

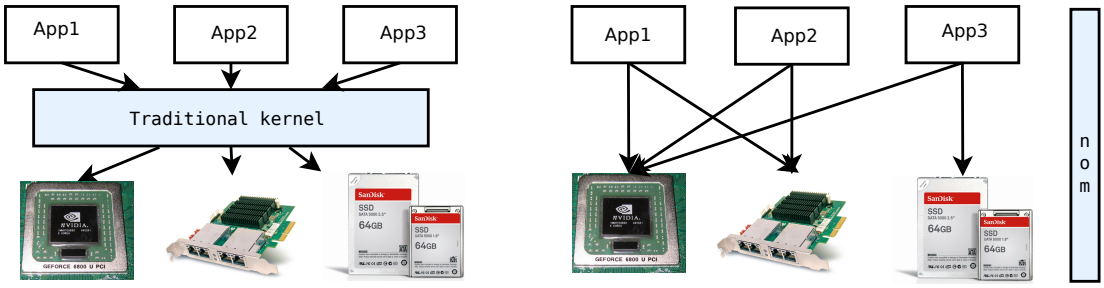

Figure 3.1: Traditional kernel structure compared with nom's kernel structure.

policy is best for it, or whether it wants a NIC driver that is currently optimized for throughput or for latency or for some combination of both, or whether it needs a small or large routing table. The kernel, which has to serve all applications equally, cannot be designed and optimized for any one application. Exposing physical resources directly to applications means that nearly all of the functionality of traditional kernels can be moved to application level and tailored to each application's specific needs.

Isolate applications. When running in a virtual machine on a modern server, the operating system's kernel can rely on the underlying hardware and on the hypervisor for many aspects of safe sharing and isolation for which it was previously responsible. For example, using an IOMMU [38], the kernel can give each application direct and secure access to its own I/O device "instances" instead of multiplexing in software a few I/O devices between many applications. Those instances may be SRIOV Virtual Functions (VFs) [57, 30] or they may be paravirtual I/O devices [16, 61, 31, 35].

3.2 Principles

The primary distinguishing feature of nom is that it enables applications to maximize their profits by (1) optimizing their entire software stack's behavior for both performance and cost; and (2) changing their behavior on the fly according to the current price of resources. As seen in Figure 3.1, traditional operating systems have a kernel that sits between applications and their I/O devices. The nom kernel, on the other hand, provides every application with safe direct access to its resources, including in particular its I/O devices. Recently proposed operating systems such as the cloud-targeted OSv [45] and Mirage [55, 54], or the bare-metal operating systems IX [19] and Arrakis [58], all of which can be considered to provide direct access of some sort, use it purely for performance. In nom, direct access enables each application to have its own private I/O stacks and private device drivers that are specialized for that application.

The nom kernel itself is minimal. It performs three main functions: (1) it initializes the hardware and boots; (2) it enumerates available resources such as CPU cores, memory, network devices, and storage devices (and acts as a clearing house for available resources); and (3) it runs applications. Once an application is launched, it queries the kernel for available resources, acquires those resources, and from then on uses them directly with minimal kernel involvement.

3.3 CPU and scheduling

On startup, a nom application acquires one or more cores from the kernel. From then on until it relinquishes the core or cores, the application performs its own scheduling using user threads. The rationale behind user threading is that only the application knows what task will be profitable to run at any given moment on its CPU cores. Applications relinquish cores when they decide to do so, e.g., because the cores have grown too expensive.

The nom design minimizes the kernel's involvement in application data paths. Applications can make system calls for control-plane setup/teardown operations, e.g., to acquire and release resources, but high performance nom applications are unlikely to make any system calls in their data paths, since their software stacks and device drivers run entirely in user space. Furthermore, nom applications handle their own traps and interrupts. Ideally, they will handle traps and interrupts without any kernel involvement. Since it is possible to inject traps and interrupts directly into virtual machines [30], ultimately the nom kernel will run its applications in guest mode using machine virtualization support [6]. This is also the approach taken by the bare-metal Dune [18] and IX [19] operating systems. Unlike Dune and IX, however, nom is targeted primarily at cloud environments, and no cloud provider currently supports hardware-assisted nested virtualization [20]. We therefore choose to run the nom kernel in ring 0 and nom applications in ring 3, without relying on the availability of nested virtualization support. Since it is not yet possible to inject traps and interrupts directly into ring 3 applications, the nom kernel receives traps and interrupts on behalf of applications in ring 0 trampolines and injects the trap or interrupt into its target application.

3.4 Memory management

Each nom application runs in its own kernel-provided address space, unlike unikernel operating systems such as OSv [45] and Mirage [55, 54], where there is a single global address space. Each nom application manages its own page mappings, unlike applications in traditional operating systems. The kernel handles an application's page fault by calling the application's page fault handler from the kernel trampoline and passing it the fault for handling. The application would typically handle page faults by asking the kernel to allocate physical memory and map pages on its behalf. This userspace-centric page fault approach provides applications with full control over their page mappings, cache coloring [43], and the amount of memory they use at any given time. There is no kernel-based paging; applications that desire paging-like functionality implement it on their own [34]. The kernel itself is non-pageable but its memory footprint is negligible.

3.5 I/O devices

The nom kernel enumerates all available physical devices on start-up and handles device hot-plug and hot-unplug. The kernel publishes resources such as I/O devices to applications using the

bulletin board, an in-memory representation of currently available resources that is mapped into each application's address space. (The bulletin board was inspired by MOSIX's [15] distributed bulletin board [12].) When an application acquires a device resource, the kernel maps the device's memory-mapped I/O (MMIO) regions in the application's address space and enables the application to perform programmed I/O (PIO) to the device. The application then initializes the device and uses it.

Most modern devices, whether virtual devices such as virtio [61] and Xen's frontend and backend devices [16], or natively-sharable SRIOV devices [57], expect to read and write memory directly via direct memory access (DMA). Since nom's model is that applications bypass the kernel and program their devices directly, devices driven by nom applications should be able to access the memory pages of the applications driving them. At the same time, these devices should not be able to access the memory pages of other applications and of the kernel.

The way nom handles DMA-capable devices depends on whether the virtual machine has an IOMMU for intra-guest protection [70]. Providing virtual machines with IOMMUs for intra-guest protection requires either an emulated IOMMU [13] or a two-level IOMMU such as ARM's sMMU or Intel's VT-d2. When an IOMMU is available for the virtual machine's use, the nom kernel maps the application's memory in the IOMMU address space of that device and subsequently keeps the MMU's page tables and the IOMMU's page tables in sync.

As far as we know, no cloud provider today exposes an IOMMU to virtual machines. To enable nom applications to drive DMA capable devices even when an IOMMU is not present, the nom kernel can also run applications in trusted mode. In this mode the kernel exposes guest-virtual to guest-physical mappings to applications and applications program their devices with these mappings. This means that in trusted mode, the kernel and every application in the same nom instance implicitly trust every other application not to take over the virtual machine by programming a device to write to memory they do not own. Strong isolation in the presence of untrusted applications can be provided by running untrusted applications in their own nom instances.

When a device owned by a nom application raises an interrupt, the kernel receives it and the kernel trampoline calls a userspace device handler registered by the application driving that device. It is the application's responsibility to handle device interrupts correctly: acknowledge the interrupt at the device and interrupt controller level and mask/unmask device interrupts as needed. Once nom applications run in guest mode, we expect device interrupts to be injected directly to the application [30].

It is well known that device polling may lead to better performance than interrupts but interrupts can reduce CPU utilization [24, 56, 39, 63, 48]. Since nom applications have full control over their software stacks and their devices, they decide when to wait for interrupts and when to poll devices directly, thereby trading off CPU cycles for performance.

3.6 Networking

The nom operating system provides a default userspace network stack, based on the lwIP network stack [25], and default network device drivers, including a driver for the virtio [61] virtnet virtual network device. Applications that want to link and run with the default network stack and network device drivers are welcome to do so. Applications that wish to yield even higher profits are encouraged to run with their own customized network stack and network device drivers. The default stack and drivers are provided as a convenience and as a basis for modifications, not because applications must use them.

To enable applications running with the default network stack and virtnet device driver to adapt their behavior on the fly, the stack and driver support run time tuning of their behavior via the *batching delay*. The batching delay controls the stack's and driver's behavior when sending and receiving packets. Applications can use the batching delay to trade-off throughput, latency, and jitter. Setting the batching delay to 0µsec means no delay: each incoming and outgoing packet is *run to completion*. Each packet the application transmits (tx packet) traverses the entire TCP/IP stack and the device driver and is sent on the wire immediately. Each packet the application receives (rx packet) is passed from the wire to the driver, to the stack, and to the application, before the next packet is handled.

Setting the batching delay to Wµsec means delaying packets by batching them together at various stages in the stack and in the driver such that no packet is delayed for more than Wµsec. Tx packets are batched together by the stack and then passed on to the driver as a batch. The driver batches all of the small batches of packets passed to it by the stack together into one large batch. When either the transmit ring buffer is close to overflowing or the first packet in the large batch has waited Wµsec, the driver transmits the large batch to the device.

The timing of arrival of rx packets is not controlled by the stack or driver but rather by the device. When $W > 0$, the driver receives incoming packets from the wire but does not pass them on to the stack for processing. The batch is kept at the driver level until at least one of the following happens: (1) Wµsec have passed; (2) the batch grows beyond a predefined maximum and threatens to overflow the receive ring buffer; or (3) there are no additional packets to receive, e.g., because the connection has been closed. The driver then passes all of the incoming packets together to the TCP/IP stack for processing.

Network-intensive applications usually optimize for throughput, latency, and jitter. Throughput is defined as the number of bytes they can send or receive in a given time period or the number of operations they can carry out. Latency is broadly defined as how long it takes to transfer or receive a single packet or carry out a single operation. Applications are usually concerned with either average latency or with tail latency, defined as the latency of the 99^{th} percentile of packets or operations. Jitter has many possible definitions. For simplicity, we define jitter as the standard deviation of the latency distribution.

A larger batching delay, up to a limit, usually provides better (higher) throughput but worse (higher) latency and jitter. A smaller batching delay usually provides better (lower) latency and jitter but worse (lower) throughput. In Chapter 4 we discuss how applications can use

valuation functions to pick the right mix of throughput, latency, and jitter, given the current price of network bandwidth. After picking the optimal mix for current conditions, applications that use the default network stack and virtnet device driver can modify the stack's batching delay to achieve the desired throughput, latency, and jitter.

3.7 Storage

In nom, applications have private storage stacks, just like they have private network stacks. They may use the default userspace storage stack and device drivers (e.g., virtio's virtblk [61]) or their own tailored stacks and drivers. Unlike the default kernel-based storage stacks of traditional operating systems, nom's default stack and drivers can adapt their behavior at run time when the cost of IOPs (for example) changes. One way to adapt behavior is to batch I/O operations together at the storage stack and driver level. Another to modify the private elevator (I/O scheduler) algorithm.

To provide multiple applications in a single nom instance with the convenience of a shared file system, `fsd` is an optional file system daemon that exposes a shared file system. Applications communicate with `fsd` via a generic high-performance IPC mechanism that uses shared memory for bulk data transfer and cross-core IPIs for notifications.

3.8 Price-awareness

Optimizing for cost requires that applications be aware of the current price of resources. The `priced` daemon queries the cloud provider via provider-specific means (e.g., the provider's REST API) for the current price of resources. It then publishes those prices to all applications through the bulletin board. To avoid the need for applications to continuously poll the bulletin board, yet enable them to react quickly to price changes, `priced` also notifies applications of any change in the price of their resources, using the same high-performance IPC mechanism `fsd` uses.

Chapter 4

Economic model and utility of network bandwidth

To maximize profit, nom applications attempt to extract the maximal benefit from the network resources they have available to them. This requires that the application be able to formulate and quantify its benefit from network resources given their current prices. The standard game-theoretic tool for doing this is a utility or valuation function: a function that is private to each application and assigns numerical values—"utilities", or in our case, profit—to different outcomes.

We consider an application acting as a server, e.g., a web server or a key-value store. The application generates revenue when it gets paid by its users for the service it provides. We assume that the amount it gets paid is a function of its throughput, latency, and jitter. The application benefits from increased throughput because higher throughput means serving more users or providing them with more content. We assume that the amount the application gets paid increases linearly with its throughput.

The application benefits from reduced latency and jitter because it can provide its users with better quality of service. Better quality of service means improved user satisfaction. To quantify user satisfaction, we adopt an existing cloud provider compensation model. Cloud providers such as GoGrid [2], NTT [4], and Verizon [5] assume that their users are satisfied as long as their service level objectives (SLOs) are met; when the provider fails to meet a user's SLO, most providers will offer their users compensation in proportion to the users' payment for periods in which the service did not meet the SLO. For example, Gogrid's Service Level Agreement (SLA) reads as follows:

> A "10,000% Service Credit" is a credit equivalent to one hundred times Customer's fees for the impacted Service feature for the duration of the Failure. (For example, where applicable: a Failure lasting seven hours would result in credit of seven hundred hours of free service [...]).

We assume that an SLA using equivalent terms exists between the application and its users. Although cloud providers list minimal throughput, maximal latency, and maximal jitter as their

SLA goals, we simplify the function by only considering latency.

We assume that the cloud provider charges the application in proportion to the outbound bandwidth it consumes. Charging by used bandwidth is reasonable for several reasons. First, it is easy for the cloud provider to monitor. Second, bandwidth consumption by one application can directly affect the quality of service for other applications running on the same cloud when there is resource pressure (limited outgoing bandwidth). Third and most important, this method of charging is commonly used in today's clouds. Amazon, for example, charges for outbound traffic per GB after the first GB, which is free.

The application does not necessarily know why the price of bandwidth rises or falls. The cloud provider may set prices to shape traffic, as CloudSigma started doing in 2010, or the price may be set according to supply and demand, as Amazon does for its spot instances [7]. The price may even be set randomly, as Amazon used to do [7]. In Kelly's [42] terms, the application is a price taker: it assumes it cannot affect the prices. It neither knows nor cares how the provider sets them. This assumption is reasonable when the application's bandwidth consumption is relatively small compared with the cloud's overall network bandwidth. The application does know that it will pay for the bandwidth it uses according to its current price.

The utility functions that we use in this work formalize the application's profit from different mixes of throughput, latency, and jitter, given the current price of bandwidth. Any such function must satisfy the *utility function axiom*: it must weakly monotonically increase as throughput increases and weakly monotonically decrease as bandwidth cost, latency, and jitter increase. In other words, the more throughput the application achieves for the same total cost, latency, and jitter, the more it profits. As latency and jitter increase, the application gets paid less or compensates its users more, so profit goes down. The higher the price of bandwidth, the higher the application's costs, so again profit goes down.

Putting all of the above together, we present three example utility functions which are consistent with the utility function axiom. We begin with the **penalty** utility function, a generalization of the simple utility function presented in the introduction (Equation (1.1)). In the simple utility function, the application owner benefits from increased throughput (T), but only as long as the application's users' average latency is below a certain latency service level objective (SLO) and the price the application owner pays her cloud provider (P) per bandwidth unit is lower than her benefit from that bandwidth unit (α.) In other words, in the simple utility function, users either pay or they don't. In the penalty utility function, the application pays its users a penalty (i.e, the users pay less) if samples of the latency distribution violate the SLO. The size of the penalty depends on the probability of violating the SLO. We define the penalty utility function in Equation (4.1) as follows:

$$U_{\text{penalty}} = T \cdot (\alpha \cdot (1 - \min(1, X \cdot \mathcal{N}(L_0, L, \sigma))) - P), \tag{4.1}$$

where T denotes throughput in $\frac{\text{Gbit}}{\text{s}}$ or application operations/second, α denotes the application owner's valuation of useful bandwidth in \$/Gbit or \$/operation, and X denotes the penalty factor from not meeting the user's SLO (e.g., 100 in the GoGrid SLA). L denotes the mean latency

(in μsecs), L_0 denotes the maximal latency allowed by the SLA, and σ denotes the latency's standard deviation (jitter). $\mathcal{N}(L_0, L, \sigma)$ denotes the probability that a normally distributed variable with mean L and standard deviation σ will be higher than L_0. In other words, it is the probability that a latency sample will not meet the latency SLO, and thus trigger compensation to the application's user. P denotes the price that the cloud provider charges the application for outgoing network bandwidth. The provider's price is set in \$/Gbit, but the application may translate it internally to \$/operation.

In the case where the sampled latency is always within the SLO and thus $\mathcal{N} \to 0$, Equation (4.1) is reduced to $T \cdot (\alpha - P)$, motivating the application to use as much bandwidth as possible, provided the value it gets from sending data (α) is higher than the price it pays for sending that data (P). Conversely, when every latency sample falls outside the SLO, Equation (4.1) is reduced to $-T \cdot P$, giving negative utility, since the penalties for violating the SLA far outweigh any benefit. It is better in this case to send nothing at all, to at least avoid paying for bandwidth.

In addition to the penalty utility function, we also consider two additional, simpler, function forms that fit the axioms and represent other business models. These functions are inspired by Lee and Snavely [49], who showed that user valuation functions for delay are usually monotonically decreasing, with various shapes, which are not necessarily linear. Hence, we consider both a linear *refund* valuation function (which is common in the literature because it is easy to represent) and a reciprocal *bonus* valuation function, which captures the diminishing marginal return, characteristic of some of the functions that Lee and Snavely found.

In the **refund** utility function in Equation (4.2), the application compensates its user by giving it a progressively larger refund as the mean latency rises, capped at a refund of 100% of the user's payment. As in the penalty utility function, α denotes the application owner's valuation of useful bandwidth. The β parameter is the extent of the refund.

$$U_{\text{refund}} = T \cdot (\max(0, \alpha - \beta \cdot L) - P), \tag{4.2}$$

In the **bonus** utility function in Equation (4.3), the application gets a bonus from its users for small latency values. The bonus decays to zero as latency grows and cannot exceed some pre-negotiated threshold, δ. γ is the extent of the bonus.

$$U_{\text{bonus}} = T \cdot (\alpha + \min(\frac{\gamma}{L}, \delta) - P), \tag{4.3}$$

The parameters α, β, γ, δ, and X, are application-specific: they characterize its business arrangements with its users. Price (P) is dictated by the cloud provider and changes over time.

We note that the application does not "choose" any function or parameters that it desires: the utility function is simply a formalization of the application owner's business relations and agreements with its users and with its cloud provider. These relations and agreements include how much the application owner pays its cloud provider for bandwidth, how much the application's users pay the application owner, how the application owner compensates its users for violating their SLAs, etc. Having said that, by understanding the behavior of the

utility function, the application owner may try to strike more beneficial deals with its cloud providers and its users. Furthermore, the application can adapt its behavior on the fly, trading off throughput, latency, and jitter so as to maximize its profit given current bandwidth price.

Chapter 5

Implementation

We implemented a prototype of nom, including both ring 0 kernel and representative ring 3 applications. The prototype runs in x86-64 SMP virtual machines on top of the KVM [44] hypervisor. It can run multiple applications with direct access to their I/O devices. It can also run on bare-metal x86-64 servers with SRIOV devices, without an underlying hypervisor, but that is not its primary use-case.

We implemented three representative applications that use the penalty, refund, and bonus utility functions to adapt their behavior on the fly: `memcached`, a popular key-value storage [28], `nhttpd`, a web server, and `NetPIPE` [65], a network ping-pong benchmark. All three applications run with private copies of the default nom lwIP-based network stack and the virtnet virtio NIC device driver. All three applications optimize for both performance and cost by adapting their stack and driver's behavior on the fly to achieve the throughput, latency, and jitter that maximize their current utility function given the current price of network bandwidth.

We implemented `nhttpd` from scratch and ported `NetPIPE` and `memcached` from Linux. The ports were relatively straightforward, since nom supports—but does not mandate—most of the relevant POSIX APIs, including pthreads (via userspace threading), sockets, and libevent. The main missing pieces for application porting are limited support for floating point (SSE) in userspace and missing support for signals.

The nom kernel is approximately 8,000 lines of code. The network stack and NIC device drivers are approximately 45,000 lines code. Both are implemented mostly in C, with a little assembly.

Chapter 6

Evaluation

6.1 Methodology

The evaluation aims to answer the following questions: (1) Does optimizing for cost preclude optimizing for performance? (2) Does optimizing for both cost and performance improve application profit? and (3) Is being able to change behavior at runtime important for maximizing profits?

We evaluate nom applications against the same applications running on Linux and on OSv [45]. The applications run in virtual machines on an x86-64 host with four Intel Core(TM) i7-3517U CPUs running at 1.90GHz and 4GB of memory. The host runs Linux Mint 17 "Qiana" with kernel 3.13.0-24 and the associated KVM and QEMU versions.

OSv and nom applications run in an x86-64 guest virtual machine with a single vCPU and 128MBs of memory. Linux applications run in a virtual machine running Linux Mint 17.1 "Rebecca", which did not boot with 128MB, so we gave it a single vCPU and 256MB of memory. We ignore the cost of memory and do not penalize Linux for running with twice the amount of memory. We also ignore the cost of CPU cycles. The host does not expose an IOMMU to virtual machines.

Our experimental setup approximates a cloud with dynamic bandwidth prices and assumes that the cloud provider either does not charge or charges a fixed sum for all other resources. Each application runs for two minutes. During the first 60 seconds, the price of bandwidth is \$1/Gb. After 60 seconds, the price rises to \$10/Gb. This situation can occur, for example, when the application starts running on a relatively idle cloud but then a noisy, network-intensive application joins it, driving up the price.

We run `memcached`, `nhttpd`, and `NetPIPE`, on Linux, OSv, and nom, and evaluate all three applications with all three valuation functions described in Chapter 4. The valuation functions take into account price, throughput, and latency, and the penalty valuation function also takes into account jitter. Applications running on Linux and OSv use the default Linux and OSv stacks and device drivers and are not price-aware.

Applications running on nom use the default lwIP and virtnet device driver. They know the throughput, latency, and jitter they expect to achieve for different settings of the batching

delay. The relationship between batching delay and throughput, latency, and jitter may be generated online and refined as the application runs or generated offline [37, 10]. We generated it offline. The applications use this information and the current price of network bandwidth as input to their valuation functions, tuning their stacks at any given moment to the batching delay that maximizes their profits. When the price of network bandwidth or the load they are under changes, they may pick a different batching delay if they calculate that it will improve their profit.

We vary the load during the experiment. During the first 60 seconds, we generate a load that approximates serving **many** small users. During the second 60 seconds, we generate a load that approximates serving a **single** important user at a time. The `memcached` load is generated with the `memaslap` benchmark application running with a GET/SET ratio of 90/10 (the default). The `nhttpd` load is generated with the `wrk` benchmark application requesting a single static file of 175 bytes in a loop. The `NetPIPE` server runs on the operating system under test and the `NetPIPE` client runs on the Linux host. `memcached` and `nhttpd` run in multiple threads/multiple requests mode, approximating serving many small users, or in a single thread/single request mode, approximating serving a single user at a time. The `NetPIPE` client either runs in bi-directional streaming mode (many) or in single request mode (single) with message size set to 1024 bytes. In all cases, to minimize physical networking effects, the load generator runs on the host, communicating with the virtual machine under test through the hypervisor's virtual networking apparatus. All power saving features are disabled in the host's BIOS and the experiments run in single user mode.

We run each experiment five times and report the averages of measured values. The average standard deviation of throughput and latency values between runs with the same parameters is less than 1% of the mean for `memcached` and less than 3% of the mean for `NetPIPE`. In `nhttpd` experiments, the single user scenario exhibits average standard deviation of both throughput and latency that is less than 1% of the mean. The many users scenario, however, exhibits average standard deviation of 10% of the mean for throughput values and 73% of the mean for latency values.

6.2 Performance

We argued that cloud applications should be optimized for cost. Does this preclude also optimizing them for performance? To answer this question, we begin by comparing the throughput, latency, and jitter achieved by nom applications with those achieved by their OSv and Linux counterparts. Throughput and latency results are the average throughput and latency recorded during each part of each experiment.

We show in Figure 6.1, Figure 6.2, and Figure 6.3 the throughput and latency achieved by `memcached`, `nhttpd`, and `NetPIPE`, respectively, during the first 60 seconds, when they serve as **many** users as possible, and during the second 60 seconds, when they only serve the most important users, a **single** user at a time. For all three applications and both scenarios, nom achieves better (higher) throughput and better (lower) latency than both OSv and Linux.

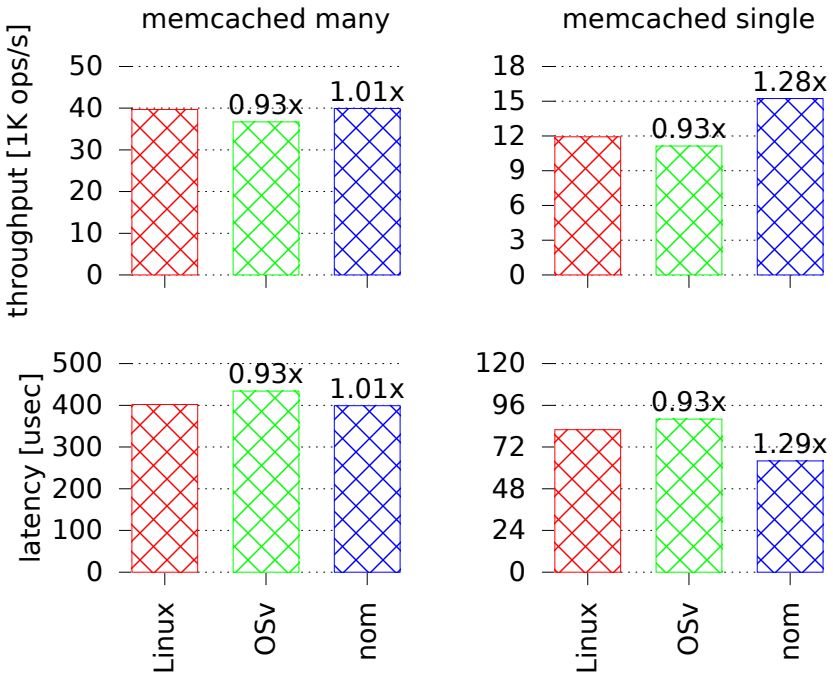

Figure 6.1: `memcached` throughput and latency

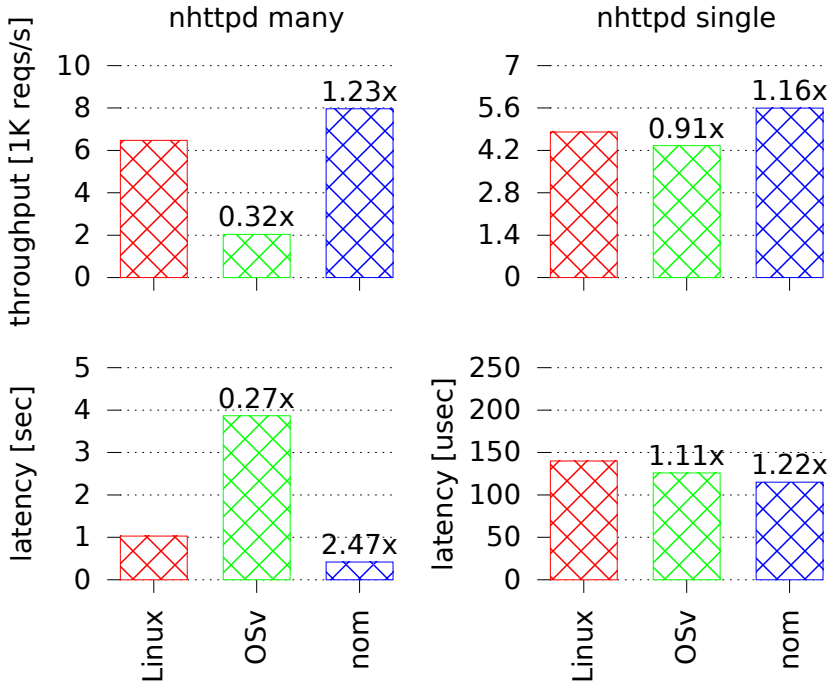

Figure 6.2: `nhttpd` throughput and latency

Taking `memcached` as an example, we see that nom achieves 1.01x–1.28x the throughput of Linux, whereas OSv only achieves 0.93x. We also see that nom achieves average latency that is 1.01x–1.29x better than Linux (vs. 0.93x for OSv) with up to 4x better jitter when compared with Linux and up to 588x better jitter when compared with OSv. (Jitter is shown in Table 6.1.)

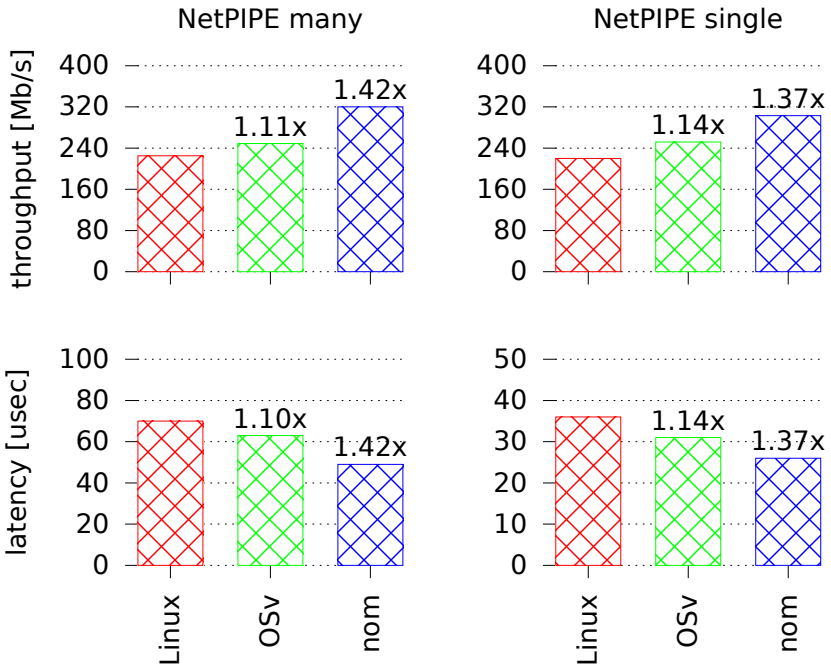

Figure 6.3: `NetPIPE` throughput and latency

Scenario	*OS*	*Latency (µsec)*	*Jitter (µsec)*
many	Linux	402	499
	OSv	434	24,148
	nom	399	121
single	Linux	82	14
	OSv	88	7,638
	nom	63	13

Table 6.1: `memcached` latency and jitter

`nhttpd` on nom achieves 1.2x–3.9x better throughput and up to 9.1x better latency than Linux and OSv, and `NetPIPE` achieves up to 1.42x better throughput and latency.

6.3 What makes nom fast?

Network applications running on nom achieve up to 3.9x better throughput and up to 9.1x better latency than their Linux and OSv counterparts (Figure 6.1, Figure 6.2, and Figure 6.3). This improvement is by virtue of nom's design and through careful application of several rules of thumb for writing high-performance virtualized systems. In particular, nom, as a cloud operating system, tries hard to keep the hypervisor out of the I/O path.

Table 6.2 shows the average number of exits per second for Linux, OSv, and nom when running `memcached`. We can see that nom causes 2.8x–4.9x fewer exits than Linux and OSv. One of the key causes of expensive hypervisor exits is injecting and acknowledging interrupts [30]. Since each nom application has its own device driver, it can decide when to wait

Metric	*OS*	*many*	*single*
#exits/sec	Linux	43,146	90,166
	OSv	43,144	51,237
	nom	10,834	18,280
#irq injections/sec	Linux	20,245	12,194
	OSv	21,768	12,368
	nom	999	999
CPU utilization	Linux	75%	65%
	OSv	59%	63%
	nom	87%	98%

Table 6.2: Average exit rate, interrupt injection rate, and CPU utilization running `memcached`

for interrupts and when to poll the device directly. We can see in Table 6.2 that the hypervisor only injects approximately 1,000 interrupts to nom while `memcached` is running. These 1,000 interrupts are all timer interrupts, which can be avoided by implementing tickless mode in the nom kernel. There are no device interrupts because all three nom applications described previously switch to polling mode as soon as they come under heavy load. Linux and OSv, in contrast, take approximately 20K–22K interrupts in the many users scenario and approximately 12K interrupts in the single user scenario. We can also see that nom's CPU utilization is 87%–98%, higher than Linux and OSv's 59%–75%. Since in our evaluation scenario CPU cycles are "free", the nom applications make the right choice to trade off CPU cycles for better throughput and latency by polling the network device. Linux and OSv applications, which do not control their software stacks and device drivers, cannot make such a tradeoff.

In addition to being "hypervisor friendly" by avoiding costly exits, nom's applications, default TCP/IP stack, and default virtnet device drivers are tuned to work well together. We eliminated expensive memory allocations on the I/O path in the applications, network stacks and device drivers, and avoided unnecessary copies in favor of zero-copy operations on the transmit and receive paths. We also used the time stamp counter (TSC) to track and reduce the frequency and cycle costs of data path operations.

Despite the 2.8x–4.9x difference in number of exits and 12x–22x difference in number of interrupts, nom's throughput and latency for `memcached` are only up to 1.3x better than Linux's. This disparity is caused by nom's default network stack and default virtnet device driver, which `memcached` uses, being not nearly as optimized as Linux's. We expect to achieve better performance and higher profits by optimizing and further customizing the stack and the driver to each application's needs. For example, instead of using the socket API, `memcached`'s internal event handling logic could call into internal network stack APIs to bypass the relatively slow socket layer [60, 40, 33]. Further optimizations and customization remain as future work.

6.4 Profit

Next, we investigate whether optimizing for both performance and cost does indeed increase profit. Using the penalty, refund, and bonus utility functions presented in Chapter 4, we calculate

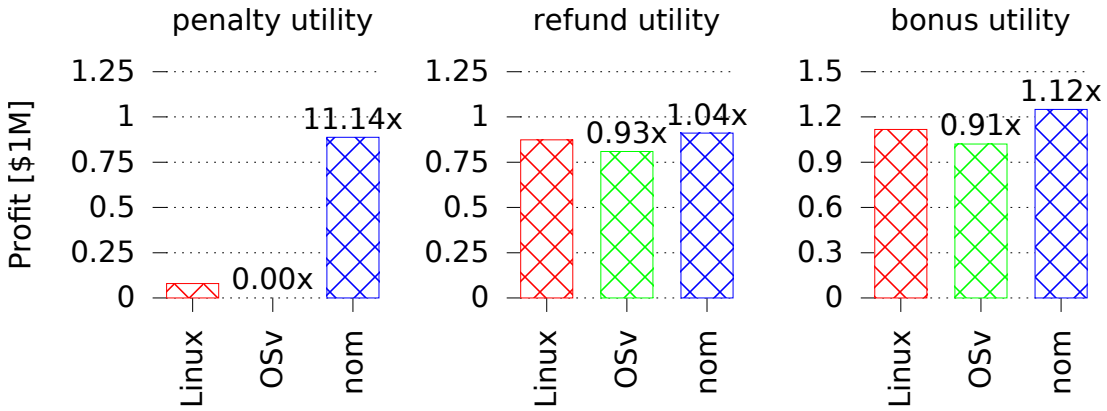

Figure 6.4: `memcached` profit

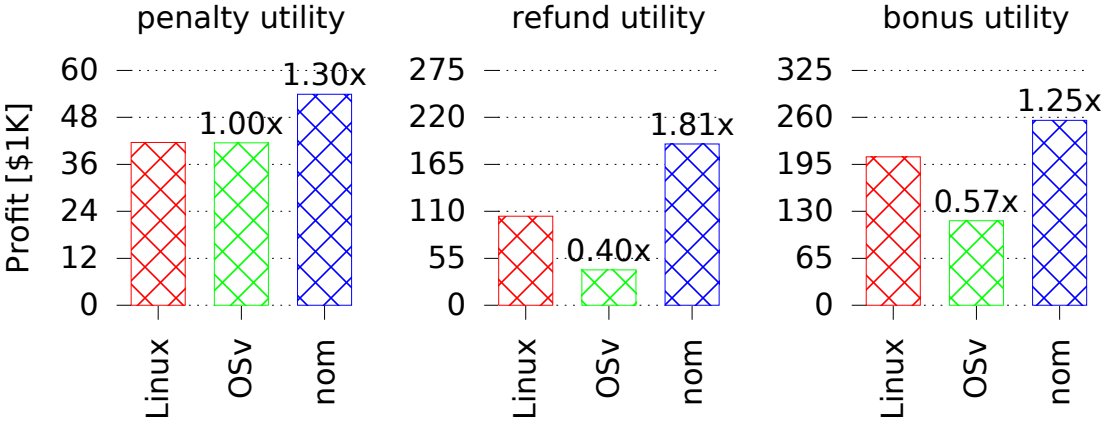

Figure 6.5: `nhttpd` profit

how much money the applications running on Linux, OSv, and nom made. Bandwidth prices fluctuate as described in the methodology section. α is set to 20 $\frac{\$}{\text{Gbit}}$, β is set to 10 $\frac{\$ \cdot \text{s}}{\text{Gbit}}$, γ is set to $0.01\frac{\$}{\text{Gbit} \cdot \text{s}}$ and δ is set to $+\inf$ (i.e., there is no limit on the bonus). The penalty for violating the latency SLO in the penalty function (X) is 100, and the maximal latency allowed by the SLA is 750µsec. We show in Figure 6.4, Figure 6.5, and Figure 6.6 `memcached`'s, `nhttpd`'s,

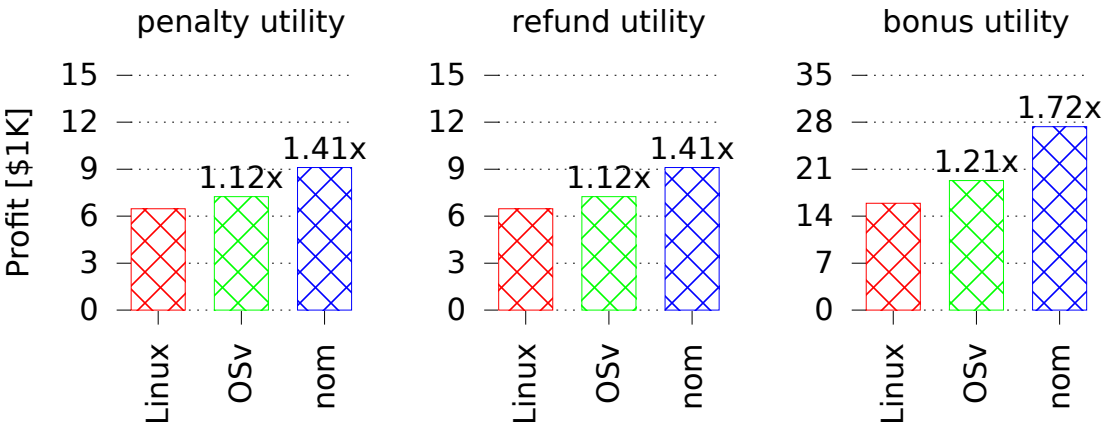

Figure 6.6: `NetPIPE` profit

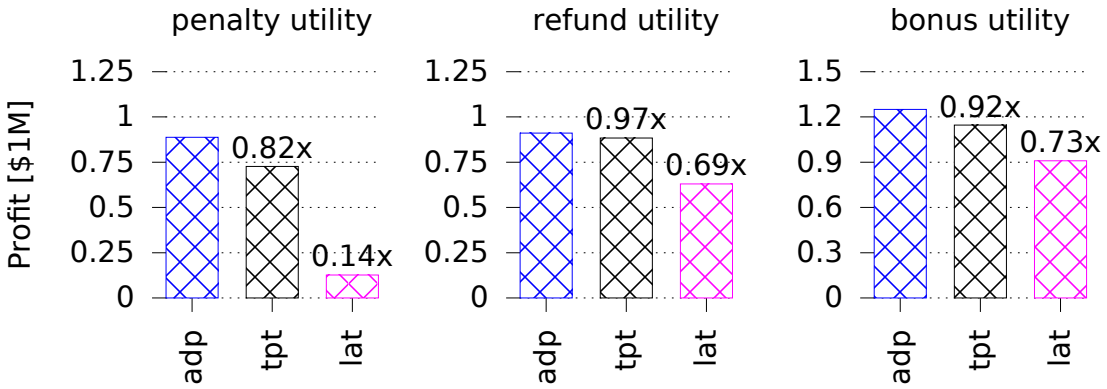

Figure 6.7: `memcached` profit: static vs. adaptive behavior

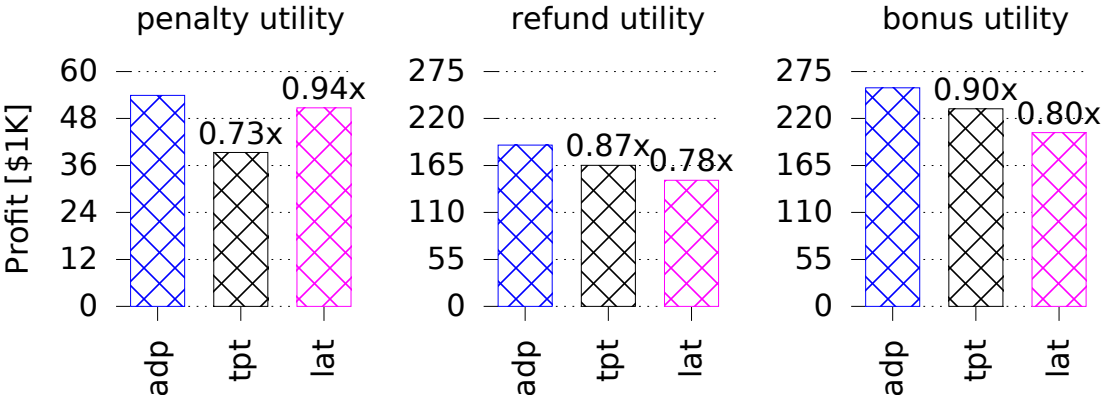

Figure 6.8: `nhttpd` profit: static vs. adaptive behavior

and `NetPIPE`'s profits. We can see that nom makes more money than either Linux or OSv with every utility function and every application. To use the penalty utility function and `memcached` as an example, for every $1 of profit Linux makes, nom makes over 11x more profit, $11.14. OSv does not profit at all due to its average latency of 7,638µsec for the single case, more than ten times the latency SLO of 750µsec. For other applications and penalty functions the difference between operating systems is not as large, but nom always yields the highest profits.

6.5 What makes nom profitable?

The nom operating system has better performance and yields higher profits than Linux and OSv. Let us now focus on only nom (rather than Linux and OSv) and answer the question: To maximize profits, is it enough to run nom applications with the settings that provide the best performance, or must applications also change their behavior on the fly when conditions change? To answer this question, we repeated the profit experiments from the previous section. This time we compared nom applications with static behavior that lead to (1) the best throughput or (2) the best latency with applications that adapt their behavior. We ran each application for 120 seconds, with price and load changing after 60 seconds. Each 120 second run used a fixed batching delay in the range of 0–40µsec.

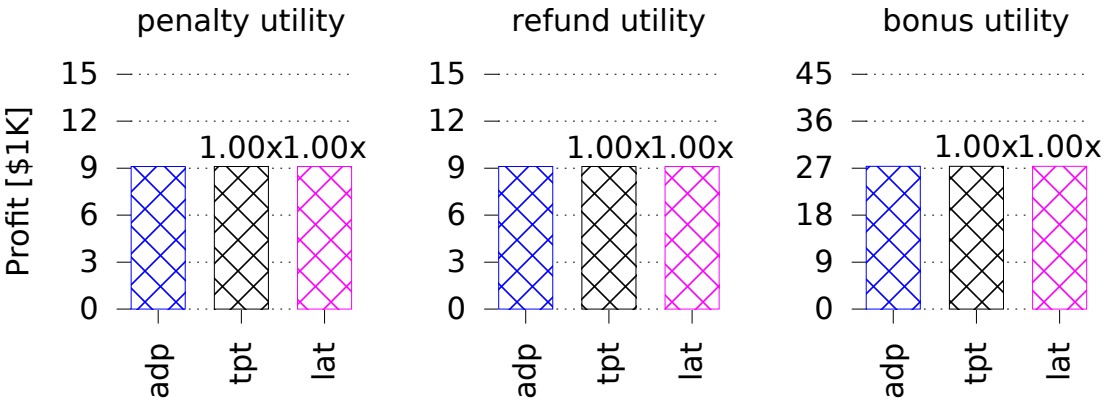

Figure 6.9: `NetPIPE` profit: static vs. adaptive behavior

Figure 6.7, Figure 6.8, and Figure 6.9 show the resulting profits. For the nom applications with static behavior and a fixed batching delay, each setting of the batching delay gave different throughput, latency, and jitter results. In the **tpt** column, we calculated the profit using the throughput and latency resulting from the batching delay that gave the best absolute throughput. In the **lat** column, we used the throughput and latency resulting from running the nom application with the fixed batching delay that gave the best absolute latency. In the **adp** (adaptive) column, the nom application changed the batching delay when the price or load changed.

As can be seen in Figure 6.7 and Figure 6.8, for both `memcached` and `nhttpd`, varying the batching delay depending on the current price and load yields higher profit than running with any fixed batching delay. Taking the penalty utility function as an example, we see that running with the throughput-optimized batching delay would give `memcached` 82% of the profit, but running with this setting would only give `nhttpd` 73% of the profit. Likewise, running with the latency-optimized batching delay would give `nhttpd` 94% of the profit, but would give `memcached` only 14% of the profit. Hence we conclude that there is no single "one size fits all" batching delay that is optimal for all applications at all times. Furthermore, there can be no single "best" stack and single "best" device driver for all applications at all times. Each application's ability to change its stack's behavior, whether through tuning or more aggressive means, is crucial for maximizing profit.

Unlike `memcached` and `nhttpd`, `NetPIPE` (Figure 6.9) shows no difference between columns. This is because `NetPIPE` is a synthetic ping-pong benchmark; its throughput is the inverse of its latency. When running on nom, `NetPIPE` tunes its stack to always run with batching delay 0, minimizing latency and maximizing throughput.

6.6 Effect of batching on throughput and latency

To understand the effect of the batching delay on application throughput and latency, we ran each application in both scenarios with a fixed batching delay between 0–40µsec. Figure 6.10, Figure 6.11, and Figure 6.12 show throughput and latency as a function of the batching delay for `memcached`, `nhttpd`, and `NetPIPE`, respectively. The throughput value shown is the

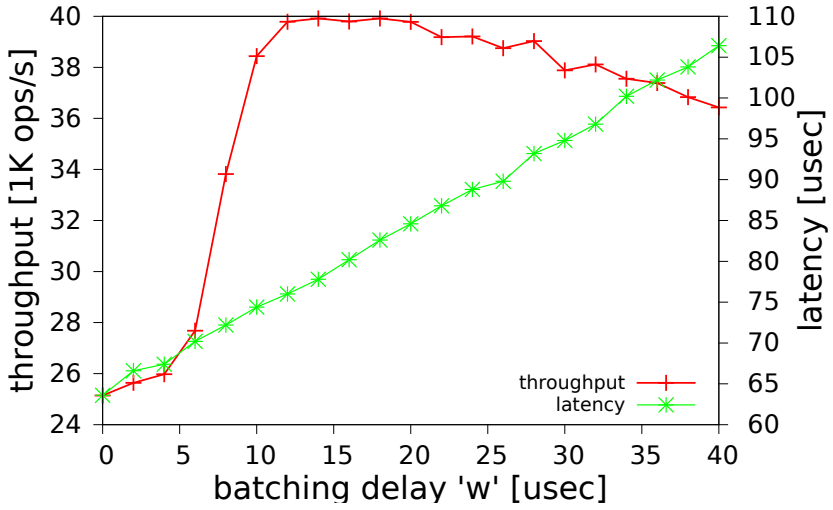

Figure 6.10: `memcached` throughput (in the many users scenario) and latency (in the single user scenario) as a function of batching delay

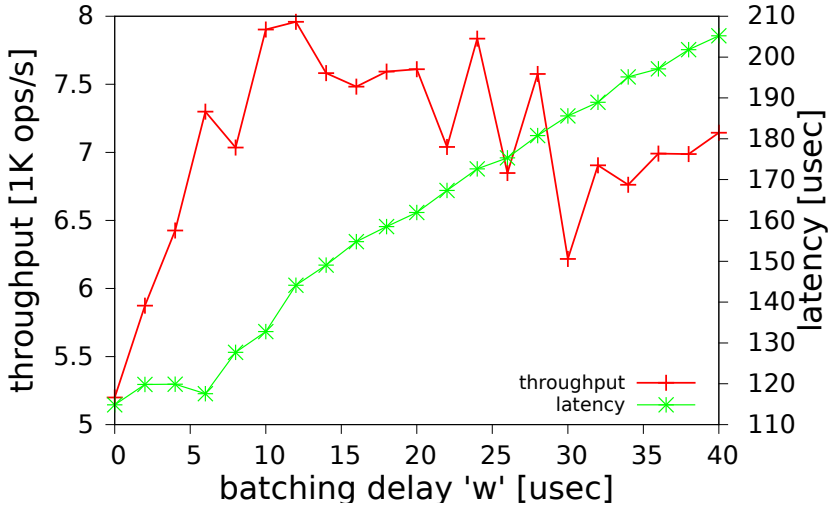

Figure 6.11: `nhttpd` throughput (in the many users scenario) and latency (in the single user scenario) as a function of batching delay

throughput achieved in the "many" scenario, which is higher than the throughput achieved in the "single" scenario. The latency value shown is the latency achieved in the "single" scenario, which is lower (better) than the latency achieved in the "many" scenario.

We can see that for `memcached` throughput achieves a local optimum at 14µsec, for `nhttpd` the optimum is 12µsec, and for `NetPIPE` a delay of 0µsec (no delay) is best. Latency for all applications is best (lowest) with no batching delay, and each microsecond of batching delay adds approximately another microsecond of latency.

6.7 Throughput/latency Pareto frontier

Varying the batching delay affects both throughput and latency. Figure 6.13, Figure 6.14, and Figure 6.15 show (throughput, latency) pairs with selected batching delays noted above the

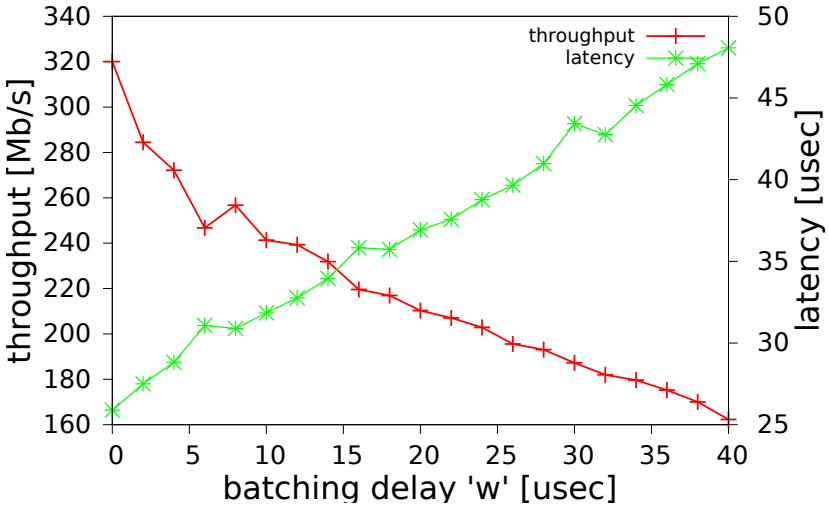

Figure 6.12: `NetPIPE` throughput (in the many users scenario) and latency (in the single user scenario) as a function of batching delay

points representing them for `memcached`, `nhttpd`, and `NetPIPE`, respectively. For both `memcached` and `nhttpd` there is a clear *Pareto frontier*, shown in blue: a set of (throughput, latency) pairs that are not dominated by any other (throughput, latency) pair. Taking `memcached` as an example, we see that using a batching delay of 10µsec can yield throughput of approximately 38K ops/s with latency of 74µsec. Using a batching delay of 32µsec (shown as a black point with '32' above it), can also yield throughput of approximately 38K ops/s with latency of approximately 96µsec. Therefore, batching delay 10 dominates 32 because it provides the same throughput with lower latency. With a different batching delay, `memcached` can also achieve higher throughput: a batching delay of 14µsec provides approximately 40K ops/s, but not without also increasing latency to 77µsec. Therefore both point 10 (38K ops/s, 74µsec) and point 14 (40K ops/s, 77µsec) are on the `memcached` throughput/latency Pareto frontier, but point 32 is not. `nhttpd`'s Pareto frontier includes batching delays 0 and 6–12. `NetPIPE`'s Pareto frontier includes a single point, 0.

The batching delay settings that are on the Pareto frontier produce better (throughput, latency) pairs than all other batching delays not on the Pareto frontier, but no one point on the Pareto frontier can be considered better than any other point on the frontier. Whereas a performance-optimized operating system is designed to find the "best" (throughput, latency) point for all cases, nom profit-maximizing applications pick the working point on the Pareto frontier that maximizes their profit at any given time *given current price and load*. When the price and/or load change, they may pick a different working point. Our experiments with nom show that there is no single "best" setting for all applications, scenarios and prices.

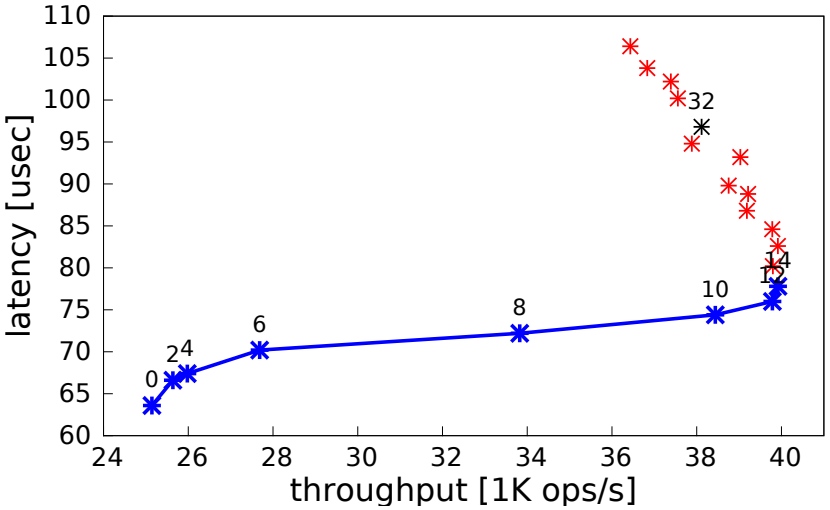

Figure 6.13: The `memcached` throughput and latency Pareto frontier

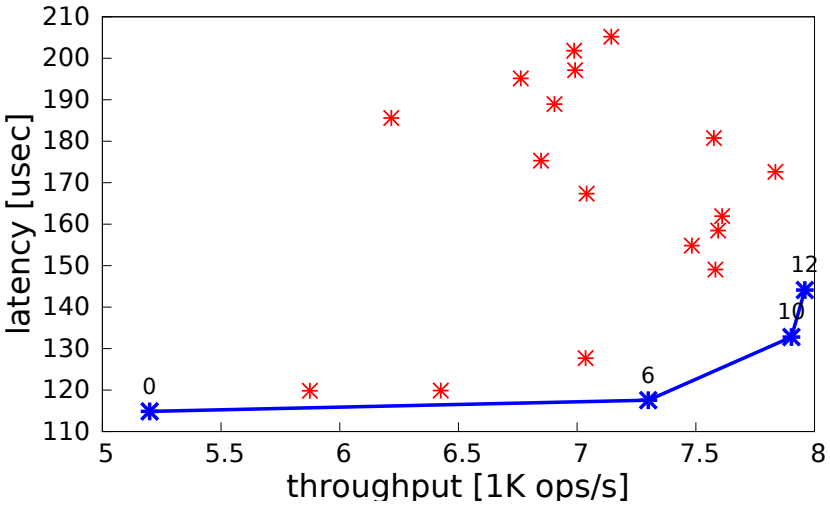

Figure 6.14: The `nhttpd` throughput and latency Pareto frontier

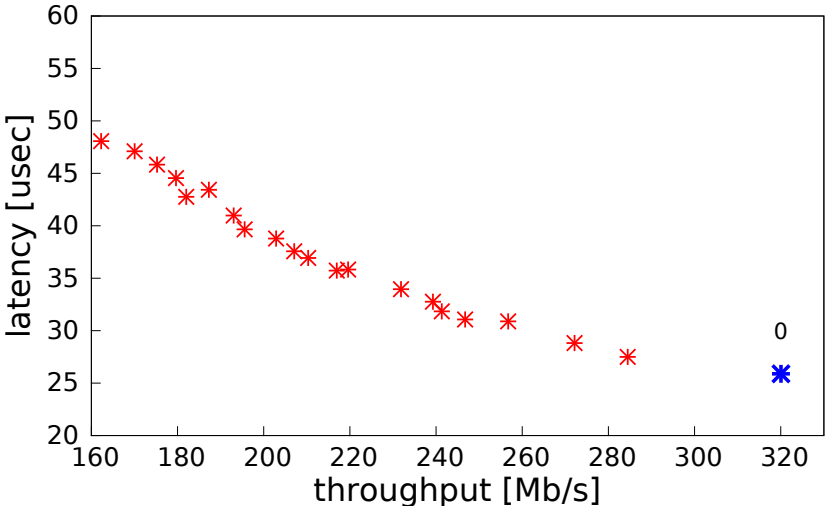

Figure 6.15: The `NetPIPE` throughput and latency Pareto frontier

Chapter 7

Discussion

There are two ways one could go about building a profit-maximizing operating system: based on an existing operating system or from scratch. To turn Linux, for example, into a profit-maximizing operating system, one could have it run applications in virtual machines using a mechanism such as Dune [18] and provide applications with direct access using direct device assignment [71] or VFIO [68]. The applications themselves would need to be modified to adapt to the changing prices of resources and would still need userspace stacks and device drivers. The primary difference between building a profit-maximizing operating system from scratch and basing it on an existing operating system is how one constructs the kernel.

We felt that going the Linux route would have constrained the design space, so we decided to implement nom from scratch to allow a wider and deeper investigation of the design space. Additionally, at its core, the profit-maximizing kernel is a *nonkernel*: a kernel that does as little as possible. Basing it on Linux seemed wasteful.

In addition to maximizing profits and improving performance, the nom approach has several advantages when compared with traditional kernels and exokernels. These include reduced driver complexity, since drivers now run completely in userspace, each driver instance serving a single application; easier debugging, development and verification of drivers and I/O stacks, for the same reason; a simpler and easier to verify trusted-computing-base in the form of the nom kernel itself [46]; and a system that we hope is more secure overall, for the same reason. The nom approach can also be useful for systems where operating power is a concern, by letting applications tune their resource requirements to the current thermal envelope limits.

The main disadvantages of the nom approach are that it forsakes legacy architectures and applications. It is designed and implemented for the kind of modern hardware available on cloud servers and will not run on older bare-metal machines. Likewise, it is not at its best when running legacy applications; realizing its benefits to the fullest extent requires some level of cooperation and effort on the part of the application developer. We believe that in the cloud, breaking away from legacy is no longer unthinkable.

Chapter 8

Related work

The nom design draws inspiration from several ideas in operating system and hypervisor construction. In addition to the original MIT exokernel [26, 27] and single address space operating systems [36, 50], nom also borrows from past work on userspace I/O (e.g., [69, 64, 22, 21, 29]), virtual machine device assignment (e.g., [71, 51, 52]), multi-core aware and extensible operating systems (e.g., [17, 47]), and library operating systems (e.g., [59, 14, 67]). It shares the underlying philosophy of specializing applications for the cloud with Mirage [55, 54] and the underlying philosophy of a minimal kernel/hypervisor with NoHype [41]. OSv [45] is a single application operating system designed for running in cloud environments. Arrakis [58] and IX [19] both provide applications with direct access to their I/O devices on bare-metal servers. All of these operating systems, however, optimize for performance. As far as we are aware, nom is the first and only operating system that maximizes profit by optimizing for both performance and cost.

The case for clouds with dynamic resource pricing (RaaS clouds) was first made by Agmon Ben-Yehuda et al. [8, 9]. On the basis of existing trends in the current IaaS industry, they deduced that the cloud business model must change: resources must be allocated on an economic basis, using economic mechanisms inside each physical machine. Ginseng [10] was the first implementation of a RaaS cloud for allocating memory. It showed that running elastic memory applications inside a traditional operating system such as Linux can be problematic due to the kernel abstracting away the hardware.

A common theme in cloud research is optimizing for cost. ExPERT [11] and Cloudyn [1] schedule workloads on clouds by taking into account both performance and cost. Optimizing for multiple goals was also previously explored in the context of power consumption. Lo et al. [53] balanced power consumption and latency. Ding et al. [23] optimized the energy-delay product.

Chapter 9

Conclusions and future work

Clouds with dynamic pricing pose new challenges but also provide an opportunity to rethink how we build system software. We propose the nom profit-maximizing operating system, a new kind of operating system that is designed and optimized for both performance and cost. The current nom prototype shows that there is no single "best" network stack or driver. Instead, nom applications maximize their profits by having private application-specific software stacks and changing their behavior on the fly in response to changing resource prices and load conditions.

The current nom prototype focuses specifically on network-intensive applications in clouds with dynamic bandwidth pricing. We are continuing to investigate profit-maximizing operating systems along several dimensions. First, we are investigating how to extract maximal value from every resource: CPU, memory, network, storage, and power. Second, we are investigating software and hardware mechanisms that can help applications change their behavior on the fly, while also achieving high performance. And third, we are investigating how to construct application-specific profit-maximizing I/O stacks and device drivers—preferably through automatic code synthesis [62].

Bibliography

[1] Cloudyn Use Cases (Online). https://www.cloudyn.com/use-cases/.

[2] GoGrid Service Level Agreement (Online). http://www.gogrid.com/legal/service-level-agreement-sla.

[3] Intel Xeon processor E5 v3 family. http://www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-software-developer-manual-325462.pdf.

[4] NTT Service Level Agreement (Online). http://www.us.ntt.net/support/sla/network.cfm.

[5] Verizon Service Level Agreement (Online). http://www.verizonenterprise.com/about/network/latency/.

[6] ADAMS, K., AND AGESEN, O. A comparison of software and hardware techniques for x86 virtualization. In *ACM Architectural Support for Programming Languages & Operating Systems (ASPLOS)* (2006), pp. 2–13.

[7] AGMON BEN-YEHUDA, O., BEN-YEHUDA, M., SCHUSTER, A., AND TSAFRIR, D. Deconstructing Amazon EC2 spot instance pricing. In *IEEE International Conference on Cloud Computing Technology and Science (CloudCom)* (2011).

[8] AGMON BEN-YEHUDA, O., BEN-YEHUDA, M., SCHUSTER, A., AND TSAFRIR, D. The Resource-as-a-Service (RaaS) cloud. In *USENIX Conference on Hot Topics in Cloud Computing (HotCloud)* (2012).

[9] AGMON BEN-YEHUDA, O., BEN-YEHUDA, M., SCHUSTER, A., AND TSAFRIR, D. The rise of RaaS: The Resource-as-a-Service cloud. *Communications of the ACM (CACM) 57*, 7 (July 2014), 76–84.

[10] AGMON BEN-YEHUDA, O., POSENER, E., BEN-YEHUDA, M., SCHUSTER, A., AND MU'ALEM, A. Ginseng: Market-driven memory allocation. In *Proceedings of the 10th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments* (2014), VEE '14.

[11] AGMON BEN-YEHUDA, O., SCHUSTER, A., SHAROV, A., SILBERSTEIN, M., AND IOSUP, A. Expert: Pareto-efficient task replication on grids and clouds. In *IEEE International Parallel & Distributed Processing Symposium (IPDPS)* (2012).

[12] AMAR, L., BARAK, A., DREZNER, Z., AND OKUN, M. Randomized gossip algorithms for maintaining a distributed bulletin board with guaranteed age properties. *Concurrency and Computation: Practice and Experience 21*, 15 (2009), 1907–1927.

[13] AMIT, N., BEN-YEHUDA, M., TSAFRIR, D., AND SCHUSTER, A. vIOMMU: efficient IOMMU emulation. In *USENIX Annual Technical Conference (ATC)* (2011).

[14] AMMONS, G., SILVA, D. D., KRIEGER, O., GROVE, D., ROSENBURG, B., WISNIEWSKI, R. W., BUTRICO, M., KAWACHIYA, K., AND HENSBERGEN, E. V. Libra: A library operating system for a JVM in a virtualized execution environment. In *ACM/USENIX International Conference on Virtual Execution Environments (VEE)* (2007).

[15] BARAK, A., GUDAY, S., AND WHEELER, R. G. *The MOSIX Distributed Operating System: Load Balancing for UNIX*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1993.

[16] BARHAM, P., DRAGOVIC, B., FRASER, K., HAND, S., HARRIS, T., HO, A., NEUGEBAUER, R., PRATT, I., AND WARFIELD, A. Xen and the art of virtualization. In *ACM Symposium on Operating Systems Principles (SOSP)* (2003).

[17] BAUMANN, A., BARHAM, P., DAGAND, P.-E., HARRIS, T., ISAACS, R., PETER, S., ROSCOE, T., SCHÜPBACH, A., AND SINGHANIA, A. The multikernel: a new OS architecture for scalable multicore systems. In *ACM Symposium on Operating Systems Principles (SOSP)* (2009).

[18] BELAY, A., BITTAU, A., MASHTIZADEH, A., TEREI, D., MAZIERES, D., AND KOZYRAKIS, C. Dune: Safe user-level access to privileged cpu features. In *Symposium on Operating Systems Design & Implementation (OSDI)* (2012).

[19] BELAY, A., PREKAS, G., KLIMOVIC, A., GROSSMAN, S., KOZYRAKIS, C., AND BUGNION, E. Ix: A protected dataplane operating system for high throughput and low latency. In *Symposium on Operating Systems Design & Implementation (OSDI)* (2014).

[20] BEN-YEHUDA, M., DAY, M. D., DUBITZKY, Z., FACTOR, M., HAR'EL, N., GORDON, A., LIGUORI, A., WASSERMAN, O., AND YASSOUR, B.-A. The Turtles project: Design and implementation of nested virtualization. In *Symposium on Operating Systems Design & Implementation (OSDI)* (2010).

[21] CAULFIELD, A. M., MOLLOV, T. I., EISNER, L. A., DE, A., COBURN, J., AND SWANSON, S. Providing safe, user space access to fast, solid state disks. In *ACM Architectural Support for Programming Languages & Operating Systems (ASPLOS)* (2012).

[22] CHEN, Y., BILAS, A., DAMIANAKIS, S. N., DUBNICKI, C., AND LI, K. UTLB: a mechanism for address translation on network interfaces. *SIGPLAN Not. 33* (October 1998), 193–204.

[23] DING, Y., KANDEMIR, M., RAGHAVAN, P., AND IRWIN, M. J. A helper thread based EDP reduction scheme for adapting application execution in cmps. In *IEEE International Parallel & Distributed Processing Symposium (IPDPS)* (2008).

[24] DOVROLIS, C., THAYER, B., AND RAMANATHAN, P. HIP: hybrid interrupt-polling for the network interface. *ACM SIGOPS Operating Systems Review (OSR) 35* (2001), 50–60.

[25] DUNKELS, A. Design and implementation of the lwIP TCP/IP stack. In *Swedish Institute of Computer Science* (2001), vol. 2, p. 77.

[26] ENGLER, D. R., AND KAASHOEK, M. F. Exterminate all operating system abstractions. In *USENIX Workshop on Hot Topics in Operating Systems (HOTOS)* (1995), IEEE Computer Society, pp. 78–83.

[27] ENGLER, D. R., KAASHOEK, M. F., AND O'TOOLE JR., J. Exokernel: an operating system architecture for application-level resource management. In *ACM Symposium on Operating Systems Principles (SOSP)* (1995).

[28] FITZPATRICK, B. Distributed caching with memcached. *Linux J. 2004*, 124 (Aug. 2004), 5–.

[29] GANGER, G. R., ENGLER, D. R., KAASHOEK, M. F., BRICENO, H. M., HUNT, R., AND PINCKNEY, T. Fast and flexible application-level networking on exokernel systems. *ACM Transactions on Computer Systems (TOCS) 20*, 1 (February 2002), 49–83.

[30] GORDON, A., AMIT, N., HAR'EL, N., BEN-YEHUDA, M., LANDAU, A., TSAFRIR, D., AND SCHUSTER, A. ELI: Bare-metal performance for I/O virtualization. In *ACM Architectural Support for Programming Languages & Operating Systems (ASPLOS)* (2012).

[31] GORDON, A., HAR'EL, N., LANDAU, A., BEN-YEHUDA, M., AND TRAEGER, A. Towards exitless and efficient paravirtual I/O. In *The 5th Annual International Systems and Storage Conference (SYSTOR)* (2012).

[32] GORDON, A., HINES, M., DA SILVA, D., BEN-YEHUDA, M., SILVA, M., AND LIZARRAGA, G. Ginkgo: Automated, application-driven memory overcommitment for cloud computing. In *Runtime Environments/Systems, Layering, & Virtualized Environments workshop (ASPLOS RESOLVE)* (2011).

[33] HAN, S., MARSHALL, S., CHUN, B.-G., AND RATNASAMY, S. Megapipe: A new programming interface for scalable network i/o. In *Symposium on Operating Systems Design & Implementation (OSDI)* (Hollywood, CA, 2012), USENIX, pp. 135–148.

[34] HAND, S. M. Self-paging in the Nemesis operating system. In *Symposium on Operating Systems Design & Implementation (OSDI)* (Berkeley, CA, USA, 1999), USENIX Association, pp. 73–86.

[35] HAR'EL, N., GORDON, A., LANDAU, A., BEN-YEHUDA, M., TRAEGER, A., AND LADELSKY, R. Efficient and scalable paravirtual I/O system. In *USENIX Annual Technical Conference (ATC)* (2013).

[36] HEISER, G., ELPHINSTONE, K., VOCHTELOO, J., RUSSELL, S., AND LIEDTKE, J. The mungi single-address-space operating system. *Software: Practice and Experience 28*, 9 (1998), 901–928.

[37] HINES, M., GORDON, A., SILVA, M., SILVA, D. D., RYU, K. D., AND BEN-YEHUDA, M. Applications know best: Performance-driven memory overcommit with ginkgo. In *IEEE International Conference on Cloud Computing Technology and Science (CloudCom)* (2011).

[38] Intel virtualization technology for directed I/O, architecture specification. ftp://download.intel.com/technology/computing/vptech/Intel(r)_VT_for_Direct_IO.pdf, Feb 2011. Revision 1.3. Intel Corporation. (Accessed Apr 2011).

[39] ITZKOVITZ, A., AND SCHUSTER, A. MultiView and MilliPage—fine-grain sharing in page-based DSMs. In *Symposium on Operating Systems Design & Implementation (OSDI)* (1999).

[40] JEONG, E., WOOD, S., JAMSHED, M., JEONG, H., IHM, S., HAN, D., AND PARK, K. mtcp: a highly scalable user-level tcp stack for multicore systems. USENIX Association, pp. 489–502.

[41] KELLER, E., SZEFER, J., REXFORD, J., AND LEE, R. B. Nohype: virtualized cloud infrastructure without the virtualization. In *ACM/IEEE International Symposium on Computer Architecture (ISCA)* (New York, NY, USA, 2010), ACM.

[42] KELLY, F. Charging and rate control for elastic traffic. *European Transactions on Telecommunications 8* (1997).

[43] KESSLER, R. E., AND HILL, M. D. Page placement algorithms for large real-indexed caches. *ACM Transactions on Computer Systems (TOCS) 10*, 4 (Nov. 1992), 338–359.

[44] KIVITY, A., KAMAY, Y., LAOR, D., LUBLIN, U., AND LIGUORI, A. KVM: the Linux virtual machine monitor. In *Ottawa Linux Symposium (OLS)* (2007).

http://www.kernel.org/doc/ols/2007/ols2007v1-pages-225-230.pdf. (Accessed Apr, 2011).

[45] KIVITY, A., LAOR, D., COSTA, G., ENBERG, P., HAR'EL, N., MARTI, D., AND ZOLOTAROV, V. Osv—optimizing the operating system for virtual machines. In *USENIX Annual Technical Conference (ATC)* (2014).

[46] KLEIN, G., ELPHINSTONE, K., HEISER, G., ANDRONICK, J., COCK, D., DERRIN, P., ELKADUWE, D., ENGELHARDT, K., KOLANSKI, R., NORRISH, M., SEWELL, T., TUCH, H., AND WINWOOD, S. seL4: formal verification of an os kernel. In *ACM Symposium on Operating Systems Principles (SOSP)* (2009).

[47] KRIEGER, O., AUSLANDER, M., ROSENBURG, B., WISNIEWSKI, R. W., XENIDIS, J., DA SILVA, D., OSTROWSKI, M., APPAVOO, J., BUTRICO, M., MERGEN, M., WATERLAND, A., AND UHLIG, V. K42: building a complete operating system. In *ACM SIGOPS European Conference on Computer Systems (EuroSys)* (2006).

[48] LANDAU, A., BEN-YEHUDA, M., AND GORDON, A. SplitX: Split guest/hypervisor execution on multi-core. In *USENIX Workshop on I/O Virtualization (WIOV)* (2011).

[49] LEE, C. B., AND SNAVELY, A. E. Precise and realistic utility functions for user-centric performance analysis of schedulers. In *International Symposium on High Performance Distributed Computer (HPDC)* (2007).

[50] LESLIE, I., MCAULEY, D., BLACK, R., ROSCOE, T., BARHAM, P., EVERS, D., FAIRBAIRNS, R., AND HYDEN, E. The design and implementation of an operating system to support distributed multimedia applications. *Selected Areas in Communications, IEEE Journal on 14*, 7 (Sep 1996), 1280–1297.

[51] LEVASSEUR, J., UHLIG, V., STOESS, J., AND GÖTZ, S. Unmodified device driver reuse and improved system dependability via virtual machines. In *Symposium on Operating Systems Design & Implementation (OSDI)* (2004).

[52] LIU, J., HUANG, W., ABALI, B., AND PANDA, D. K. High performance VMM-bypass I/O in virtual machines. In *USENIX Annual Technical Conference (ATC)* (2006), pp. 29–42.

[53] LO, D., CHENG, L., GOVINDARAJU, R., BARROSO, L. A., AND KOZYRAKIS, C. Towards energy proportionality for large-scale latency-critical workloads. In *Proceeding of the 41st Annual International Symposium on Computer Architecuture* (Piscataway, NJ, USA, 2014), ACM/IEEE International Symposium on Computer Architecture (ISCA), IEEE Press, pp. 301–312.

[54] MADHAVAPEDDY, A., MORTIER, R., ROTSOS, C., SCOTT, D., SINGH, B., GAZAGNAIRE, T., SMITH, S., HAND, S., AND CROWCROFT, J. Unikernels: Library

operating systems for the cloud. In *ACM Architectural Support for Programming Languages & Operating Systems (ASPLOS)* (2013).

[55] MADHAVAPEDDY, A., MORTIER, R., SOHAN, R., GAZAGNAIRE, T., HAND, S., DEEGAN, T., MCAULEY, D., AND CROWCROFT, J. Turning down the lamp: software specialisation for the cloud. In *USENIX Conference on Hot Topics in Cloud Computing (HotCloud)* (2010).

[56] MOGUL, J. C., AND RAMAKRISHNAN, K. K. Eliminating receive livelock in an interrupt-driven kernel. *ACM Transactions on Computer Systems (TOCS) 15* (1997), 217–252.

[57] PCI SIG. Single root I/O virtualization and sharing 1.0 specification, 2007.

[58] PETER, S., LI, J., ZHANG, I., PORTS, D. R. K., WOOS, D., KRISHNAMURTHY, A., ANDERSON, T., AND ROSCOE, T. Arrakis: The operating system is the control plane. In *Symposium on Operating Systems Design & Implementation (OSDI)* (2014).

[59] PORTER, D. E., BOYD-WICKIZER, S., HOWELL, J., OLINSKY, R., AND HUNT, G. C. Rethinking the library OS from the top down. In *ACM Architectural Support for Programming Languages & Operating Systems (ASPLOS)* (2011).

[60] RIZZO, L. Netmap: a novel framework for fast packet I/O. In *USENIX Annual Technical Conference (ATC)* (2012).

[61] RUSSELL, R. virtio: towards a de-facto standard for virtual I/O devices. *ACM SIGOPS Operating Systems Review (OSR) 42*, 5 (2008), 95–103.

[62] RYZHYK, L., WALKER, A., KEYS, J., LEGG, A., RAGHUNATH, A., STUMM, M., AND VIJ, M. User-guided device driver synthesis. In *Symposium on Operating Systems Design & Implementation (OSDI)* (Broomfield, CO, Oct. 2014), USENIX Association, pp. 661–676.

[63] SALIM, J. H., OLSSON, R., AND KUZNETSOV, A. Beyond Softnet. In *Anual Linux Showcase & Conference* (2001).

[64] SCHAELICKE, L., AND DAVIS, A. L. Design Trade-Offs for User-Level I/O Architectures. *IEEE Trans. Comput. 55* (August 2006), 962–973.

[65] SNELL, Q. O., MIKLER, A. R., AND GUSTAFSON, J. L. Netpipe: A network protocol independent performance evaluator. *IASTED International Conference on Intelligent Information Management and Systems 6* (1996).

[66] UHLIG, R., NEIGER, G., RODGERS, D., SANTONI, A. L., MARTINS, F. C. M., ANDERSON, A. V., BENNETT, S. M., KAGI, A., LEUNG, F. H., AND SMITH, L. Intel virtualization technology. *Computer 38*, 5 (2005), 48–56.

[67] VAN HENSBERGEN, E. P.R.O.S.E.: partitioned reliable operating system environment. *SIGOPS Oper. Syst. Rev. 40*, 2 (Apr. 2006), 12–15.

[68] VFIO driver: non-privileged user level PCI drivers. http://lwn.net/Articles/391459/, Jun 2010. (Accessed Feb., 2015).

[69] VON EICKEN, T., BASU, A., BUCH, V., AND VOGELS, W. U-Net: a user-level network interface for parallel and distributed computing. In *ACM Symposium on Operating Systems Principles (SOSP)* (New York, NY, USA, 1995).

[70] WILLMANN, P., RIXNER, S., AND COX, A. L. Protection strategies for direct access to virtualized I/O devices. In *USENIX Annual Technical Conference (ATC)* (2008).

[71] YASSOUR, B.-A., BEN-YEHUDA, M., AND WASSERMAN, O. Direct device assignment for untrusted fully-virtualized virtual machines. Tech. Rep. H-0263, IBM Research, 2008.

למשאבי הקלט/פלט שלהן. גישה ישירה זו מאפשרת לכל תוכנה לרוץ עם מכלול רשת פרטי ועם מנהלי התקני קלט/פלט פרטיים שתוכננו במיוחד עבורה. פונקציית התועלת הפרטית של כל תוכנה משמשת אותה על מנת לבחור בכל רגע נתון באיזון של הספק, השהיה וריצוד, שמשיא לה רווחים מקסימליים, כתלות במחיר הנוכחי של רוחב הפס.

הניסויים שערכנו עם נום מראים שאין מכלול רשת "הכי טוב" ואין מנהלי התקני חומרה "הכי טובים". מה שטוב לתוכנה אחת לאו דווקא טוב לתוכנה אחרת. במקום זאת עלינו, בוני מערכות הפעלה, לאפשר לכל תוכנה למקסם את רווחיה על ידי שינוי ההתנהגות בזמן ריצה שלה, של מכלול הרשת שלה ושל מנהלי ההתקנים שלה, כתלות במצב שבו היא נמצאת, מחיר המשאבים והעומס שהיא נתונה בו. רק התוכנה יודעת מה הכי טוב עבורה.

שלהן למקסם את הרווחים שלהן, ויוכלו להשתמש בחומרה המודרנית שזמינה בשרתי ענן על מנת לספק הפרדה בין תוכנות שונות. אנו גוזרים משינויים אלה רשימת דרישות ממערכות הפעלה שירוצו בעננים עם מחיר משתנה.

נום היא מערכת הפעלה שתכננו לפי דרישות אלו, מימשנו וייעלנו עבור עננים עם מחירים משתנים. התכנון של רכיביה השונים של נום, כולל תזמון התהליכים, ניהול הזיכרון, גישה ישירה להתקני קלט/פלט ומודעות למחיר, מיועד לאפשר לתוכנות שלה להשיג רווחים גדולים ככל האפשר. בנום כל תוכנה רוכשת משאבים, ומרגע שרכשה אותם, הם שלה עד שתוותר עליהם. כך למשל תוכנות שרצות על נום רוכשות ליבות שלמות ומתזמנות את החוטים שלהן בעצמן, רוכשות דפי זיכרון ומנהלות את מיפויי הזיכרון שלהן בעצמן, וניגשות בעצמן למשאבי הקלט/פלט שאותם רכשו, בלי כל מעורבות של הליבה של נום במסלולי הקלט/פלט הרגישים לביצועים. הן עושות את כל זאת מתוך מודעות למחירים השונים של המשאבים.

בעבודה זו התרכזנו בתוכנות המשתמשות שימוש אינטנסיבי ברשת, כגון שרתי דפי אינטרנט ושרתי "מפתח וערך" כמו memcached. בנום אין מכלול רשת (TCP/IP stack) אחד אידיאלי או מנהל התקן חומרה אחד אידיאלי, משום שמכלול רשת או מנהל התקן חומרה שאידיאלי לתוכנה אחת גרוע עבור תוכנה אחרת. במקום זאת לכל תוכנה שרצה על נום יש גישה ישירה להתקני הקלט/פלט שלה ולכן גם מכלול רשת פרטי משלה ומנהלי התקני חומרה פרטיים משלה הכלולים בתוך מרחב הכתובות שלה. מאחר שהרווח של התוכנה תלוי בין היתר בעלות הרגעית של רוחב הפס של הרשת שבה היא משתמשת, כל תוכנה שרצה על נום יכולה לשנות תוך כדי ריצה את התנהגותה, התנהגות מכלול הרשת שלה, והתנהגות מנהלי התקני החומרה שלה כשעלות רוחב הפס משתנה. כך למשל תוכנות שרצות על נום יכולות לשנות את הקצב והתזמון שבהם הן שולחות ומקבלות חבילות ברשת כתלות במחירו הנוכחי של רוחב הפס.

תוכנות רשת נמדדות לפי מספר משתנים: הספק הרשת (throughput) שהן מספקות, ההשהיה (latency) שהלקוחות שלהן חווים, והריצוד (jitter) של ההשהיה. תוכנות שרצות על נום משתמשות בכלי מתורת המשחקים שנקרא "פונקציית תועלת" על מנת לדעת איך לשנות את התנהגותן בזמן ריצה. בעבודה זו אנו חוקרים שלוש פונקציות תועלת: פונקציה המבוססת על עונשים, שבה התוכנה נענשת ומפצה את לקוחותיה כשאינה עומדת בהתחייבויותיה לגבי השהיה וריצוד, פונקציה המבוססת על החזר, שבה התוכנה מחזירה ללקוחותיה חלק הולך וגדל מהסכום ששלמו ככל שההשהיה שהם סובלים ממנה גדולה יותר, ופונקציה מבוססת תמריץ, שבה הלקוחות מתגמלים את התוכנה ככל שההשהיה שהיא מספקת להם נמוכה יותר. בעזרת פונקציות התועלת, תוכנות שרצות על נום יכולות להחליט איזה איזון הן רוצות בין הספק, השהיה וריצוד. אפליקציות ממקסמות רווחים יבחרו את האיזון שימקסם את רווחיהן, כתלות במחיר הנוכחי של רוחב הפס. טענה רווחת היא שקשה לבטא פונקציות שכאלה. אנו מסכימים, אבל יחד עם זאת טוענים שכדאי לעשות זאת במקרים בהם הרווח הפוטנציאלי גבוה מספיק.

בנינו אב טיפוס של נום שרץ במכונות וירטואליות תחת המשגוח KVM. אב טיפוס זה מריץ שלוש תוכנות לדוגמה: שרת דפי אינטרנט, שרת "מפתח וערך", ובוחן ביצועי רשת. שלוש תוכנות אלו, כאשר הן רצות על נום, משיגות גם ביצועים וגם רווחים גבוהים יותר מאשר אותן תוכנות כאשר הן רצות הן על מערכת ההפעלה המיושנת לינוקס והן על OSv, מערכת הפעלה חדשה המיועדת לענן. בפרט, על נום שלוש התוכנות משיגות רווחים טובים עד כדי פי 11.1, הספק רשת טוב עד כדי פי 3.9 והשהיה טובה עד כדי פי 9.1.

תוכנות שרצות על נום משיגות ביצועים טובים יותר כי הן תוכננו להיות "ידידותיות למשגוח" ולהמנע ממעברים מיותרים בין נום לבין המשגוח. הן משיגות רווחים טובים יותר מאחר שיש להן גישה ישירה

תקציר

יותר ויותר מכוח המחשוב העולמי נמצא בענני תשתית כשירות (Infrastructure-as-a-Service, IaaS), כמו למשל ענני התשתית כשירות של חברת אמזון או חברת גוגל. כל אחד ואחת עם כרטיס אשראי יכולים לשכור מענני תשתית כשירות כח מחשוב בלתי מוגבל. ענני תשתית כשירות מריצים מערכות הפעלה ותוכנות בתוך מכונות וירטואליות בעזרת המשאבים הפיזיים (מעבדים, זיכרון, רשת, ויחידות אחסון) של מכונות הענן.

יש שתי מגמות שבעטיין אנו מאמינים שענני התשתית כשירות יהפכו לענני משאב כשירות (Resource-as-a-Service, RaaS). בענני משאב כשירות לכל משאב יהיה מחיר הנשלט ע"י כוחות השוק ומשתנה תדירות. המגמות הן הגרעיניות ההולכת וקטנה של חלוקת משאבים במרחב והגרעיניות ההולכת וקטנה של חלוקת משאבים בזמן. ניתן לראות שתי מגמות אלו כבר היום בעננים מסחריים ובחומרה שממנה הם בנויים. יתר על כן, כבר היום חלק מהעננים גובים מחיר משתנה עבור משאבים מסוימים.

למרות שענני תשתית כשירות, ובמידה רבה עוד יותר, ענני משאב כשירות, מייצגים דרך חדשה לגמרי בה ניתן לקנות, למכור, ולהשתמש במשאבי מיחשוב, כמעט כל מכונה וירטואלית שרצה היום בענן מריצה את אותן מערכות הפעלה מיושנות כגון לינוקס וחלונות. מערכות הפעלה אלו תוכננו עבור החומרה שהייתה זמינה לפני עשורים ומיועדות לעבוד על מחשב פיזי אחד בלבד. הן מספקות לתוכנות שרצות עליהן שכבת הפשטה עבה שמסתירה את החומרה שמתחתיה ומניחות שכל משאבי המחשב זמינים לשימושן הבלעדי ללא כל עלות. במילים אחרות, הן לא יודעות ולא אכפת להן שהמשאבים שהן משתמשות בהם בענן עולים כסף, ושמחירם יכול להשתנות בין רגע, למשל כתוצאה משינויים בהיצע הזמין ובדרישה למשאבים.

אנו טוענים שמערכות הפעלה מיושנות אלה אינן מתאימות לענן וצריך להחליפן. מערכות הפעלה אלו מתוכננות, בראש ובראשונה, לספק ביצועים גבוהים ככל האפשר. שיפור ביצועים כיעד עליון, ללא כל התחשבות בעלות המשאבים הנדרשים על מנת להגיע לביצועים גבוהים, הגיוני רק כאשר עלות המשאבים זניחה או קבועה. אם העלות זניחה, אין מה להרוויח מלהביא אותה בחשבון. אם העלות קבועה, אין מה לעשות לגביה. אנו טוענים שבעננים עם מחירים משתנים, שבהם המחירים לא זניחים ולא קבועים, לא כדאי להריץ מערכות הפעלה ישנות שכאלה. במקום, צריך להריץ מערכות הפעלה שממקסמות רווחים. מערכות הפעלה ממקסמות רווחים מתוכננות להשיג למשתמשים בהן רווחים מקסימליים על ידי איזון נכון בין עלות ובין תועלת, בין ביצועים ובין עלויות. אנו מראים בהמשך העבודה שמערכות הפעלה ממקסמות רווחים יכולות להביא לתוכנות שרצות עליהן רווח גבוה בסדר גודל מהרווח שאותן תוכנות ישיגו על מערכות הפעלה מיושנות.

אנו סוקרים בעבודה זו את השינויים שענני משאב כשירות יחייבו במערכות ההפעלה שירוצו עליהם. בין היתר, מערכות הפעלה לא יוכלו יותר להניח שכל המשאבים תחת שליטתן, יהיו חייבות לאפשר לתוכנות

המחקר בוצע בהנחייתו של פרופ׳ דן צפריר, בפקולטה למדעי המחשב.

חלק מן התוצאות בחיבור זה ותוצאות שחיבור זה מתבסס עליהן פורסמו כמאמרים מאת המחבר ושותפיו למחקר בכנסים ובכתבי־עת במהלך תקופת המחקר למאגיסטר של המחבר. הגרסאות העדכניות ביותר של מאמרים אלה הינן:

Orna Agmon Ben-Yehuda, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. The rise of RaaS: The Resource-as-a-Service cloud. *Communications of the ACM (CACM)*, 57(7):76–84, July 2014.

Nadav Amit, Muli Ben-Yehuda, Dan Tsafrir, and Assaf Schuster. vIOMMU: efficient IOMMU emulation. In *USENIX Annual Technical Conference (ATC)*, 2011.

Orna Agmon Ben-Yehuda, Eyal Posener, Muli Ben-Yehuda, Assaf Schuster, and Ahuva Mu'alem. Ginseng: Market-driven memory allocation. In *ACM/USENIX International Conference on Virtual Execution Environments (VEE)*. 2014.

Orna Agmon Ben-Yehuda, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. Deconstructing Amazon EC2 spot instance pricing. *ACM Transactions on Economics and Computation (TEAC)*, 1(3):16:1, September 2013.

Muli Ben-Yehuda, Omer Peleg, Orna Agmon Ben-Yehuda, Igor Smolyar, and Dan Tsafrir. The nonkernel: A kernel designed for the cloud. In *Asia Pacific Workshop on Systems (APSYS)*, 2013.

Abel Gordon, Nadav Amit, Nadav Har'El, Muli Ben-Yehuda, Alex Landau, Dan Tsafrir, and Assaf Schuster. ELI: Bare-metal performance for I/O virtualization. In *ACM Architectural Support for Programming Languages & Operating Systems (ASPLOS)*, 2012.

Michael Hines, Abel Gordon, Marcio Silva, Dilma Da Silva, Kyung Dong Ryu, and Muli Ben-Yehuda. Applications know best: Performance-driven memory overcommit with Ginkgo. In *IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, 2011.

תודות

ראשית ברצוני להודות לאשתי המדהימה, אורנה אגמון בן־יהודה. היית חברה, עמיתה ומנחה, ולימדת אותי יותר ממה שאוכל אי פעם לתאר. שנית ברצוני להודות לילדי המדהימים, יעל וזאב, אשר הביאו אור ומשמעות לחיי. ברצוני להודות גם להורי, יואל ועירית בן יהודה, על כך שמעולם לא איבדו את האמונה בי, גם כשהדרך הפכה פתלתלה. לבסוף, ברצוני להודות למייקל פקטור ולאורן קריגר, אשר לימדו אותי מהו מחקר.

עבדתי על מערכת ההפעלה נום ועל חיבור זה שנים רבות. במשך שנים אלו חיברתי כעשרים מאמרים ביחד עם אנשים מוכשרים רבים. ברצוני להודות לכולם—נהניתי מאוד לעבוד עימכם!

הכרת תודה מסורה לטכניון על מימון חלקי של מחקר זה.

מערכת ההפעלה ממקסמת הרווחים "נום"

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים במדעי המחשב

שמואל (מולי) בן־יהודה

הוגש לסנט הטכניון --- מכון טכנולוגי לישראל
אייר התשע"ה חיפה מאי 2015

מערכת ההפעלה ממקסמת הרווחים "נום"

שמואל (מולי) בן־יהודה