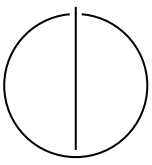
SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Analysis of the Noisy Neighbor Problem in AWS

Youssef Jemal



SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

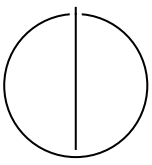
TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Analysis of the Noisy Neighbor Problem in AWS

Analyse der "Noisy Neighbor" Problem in AWS

Author:	Youssef Jemal
Examiner:	Prof. Dr. Leis Viktor
Supervisor:	Till Steinert
Submission Date:	22/08/2025



I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Munich, 22/08/2025 Youssef Jemal

Abstract

Contents

1 Introduction (✓)

Infrastructure-as-a-Service (IaaS) is a cloud computing model that provides customers with access to computing resources such as servers, networking and virtualization. Cloud vendors in general and Amazon Web Services (AWS) in particular abstract the physical placement of the virtual machine providing users with limited transparency to how many tenants are sharing the underlying hardware. This information can be crucial as this co-residency can result in significant performance degradation across different resources such as CPU, memory, network and storage I/O [14] [20]. Although virtualization technology has undergone major improvements in the past decades, contention can still occur due to other factors such as Simultaneous Multi-Threading (SMT), Network Interface Card queuing, and other system-level bottlenecks. To address the issue of unwanted resource contention, Amazon Web Services introduced dedicated hosts [4], which are physical servers that are fully dedicated to the customer, allowing users to deploy virtual machines with complete transparency over their placement. Another potential solution are the spread placement groups, which ensure that the VMs are placed on different physical servers. It can be particularly useful for highly parallel workloads, where the nodes have a similar workload nature (network- or CPU-intensive). However, this approach only mitigates internal resource contention and is limited to a small number of EC2 instances in the same availability zone [5].
In this paper, we leverage different benchmarking tools to assess the extent of the performance degradation that can occur because of VM co-location. We are particularly interested in quantifying the worst possible degradation that can happen by running identical resource-intensive benchmarks in parallel across VMs residing on the same dedicated host. We primarily utilized 5th and 6th general-purpose AWS EC2 VM instances that are built around the AWS Nitro system and run on top of the Nitro hypervisor. We analyze CPU contention across various CPU vendors namely Intel, Gravtion, and AMD. We examine whether and to what extent SMT is involved in this degradation. Furthermore, we investigate network resource contention, which is interesting since AWS does not provide exact specifications like other resources such as CPU and memory. We analyze the degradation that can occur both on throughput and latency across co-located EC2 Instances.
The thesis is structured as follows: We start by discussing related work in section 2. Section 3 introduces the key concepts required to understand this work, namely virtu-

alization and Simulatenous Multi Threading. In section 4, we explain the methodology of our experiments and introduce the different benchmarking tools that were adopted throughout the thesis. Section 5 analyzes CPU contention across 5th and 6th EC2 generations and contexualizes the results in relation to the CPU architecture. In Section 6, we examine network I/O contention and explore its manifestation across throughput and latency. Section 7 concludes our work and summarizes its most important findings.

2 Related Work (✓)

Han et. al [14] investigated public cloud resource contention. They executed CPU, disk, and network I/O benchmarks across up to 48 VMs sharing the same dedicated host. The tests were executed mainly on 3rd (c3), 4th (c4), and 5th (m5d) generation VMs. The results showed considerable performance degradation with CPU degradation reaching 48% and Network throughput up to 94%. Throughput degradation was measured in relation to the initial bandwidth available to the VM, i.e., burst bandwidth and not the baseline bandwidth, which can be misleading. We will discuss this further in the network I/O section. The paper also analyzed the unexpected CPU performance degradation caused by adding idle linux VMs on the dedicated host. The measurements were leveraged to train multiple linear regression and random forest models to predict VM co-residency. The linear regression model achieved an R^2 of .942. This could be very practical, as it enables user to relocate VMs to have access to better performance with less contention.

Rehman et. al [23] analyzed the problem of provisioning variation in public clouds. By running benchmarks and sample MapReduce workloads, they found that provisioning variation can impact the performance by a factor of 5. They argue that this is primarily due to network I/O contention.

Lloyd et. al [20] reported a 25% performance degradation when running compute intensive scientific modelling web services on pools consisting of m1.large VMs with high resource contention. They developed an approach called Noisy-Neighbor-Detect that leverages the cpuSteal metric to identify VMs with noisy neighbors from a pool of worker VMs. cpuSteal refers to the percentage of time a virtual CPU spends waiting for the hypervisor to allocate a physical CPU to run on.

Many other techniques were developed by researchers to identify or predict resource contention. Govindan et. al [12] developed an only software solution called Cuanta that predicts performance interference due to shared chip level resource namely cache space and memory bandwidth. Although, the performance degradation of consolidated application can be empirically investigated, the number of possible workload placements is combinatorial. A cloud provider hosting M VMs with N VMs per server needs to perform $\frac{M!}{N!\,(M-N)!}$ measurements, which is highly impractical. Cuanta does not require any changes to the software of the hosting platform and the prediction complexity is linear to the number of cores sharing the Last Level Cache (LLC), making it a far better

alternative than its emprical counterpart. The software provided promising results with up to 96% accuracy on Intel Core-2-Duo processors.
Some efforts have looked at side channels as a way to detect VM co-location. Side channels are an indirect way of extracting information from a system that designers never intended to expose its implementation details. Inci et. al [17] developed three methods to detect co-located VMs. The first two approaches leveraged Last Level Cache: Cooperative LLC covert channel and Cache profiling. While the former requires cooperation of the victim VMs, the latter operates independently. In the second method, the attacker fills the cache with its own data and after a short pause re-accesses the same buffer while monitoring the memory access time. Low eviction rates indicate that the attacker is likely alone on the host, while high eviction rates point toward VM co-location. The third method is memory bus locking. The idea is for the attacker to launch special instructions that block the memory bus and then analyze the resulting delays to infer VM-colocation. All three methods had a high accuracy in detecting co-location in real commercial cloud settings.

3 Background

3.1 Virtualization (✓)

Virtualization is a technology that allows the creation of isolated virtual environments also known as Virtual Machines that run on the same physical server [22]. Each VM has its own operating system and acts as an independent physical computer. The VMs are called "guests" and the physical server is called "host". Virtualization is crucial for the IaaS model that's offered by cloud providers, as it provides various advantages [22]. It improves resource and cost efficiency by dividing the physical server into multiple isolated instances, each tailored to different workload needs. This reduces the amount of unused capacity that occurs when a server is dedicated to just one task.
The main component that handles the necessary tasks for virtualization is the Virtual Machine Monitor (VMM) also called hypervisor [7]. Most of the instructions that are executed by the virtual achines run natively on the CPU and do not require intervention from the VMM, such as arithmetic operations. However, there is a class of privileged instructions that guests can not directly execute on the CPU, such as I/O operations. When such an instruction is encountered, the CPU raises a trap, that signals to the VMM to intervene and emulate the behavior of the instruction [7]. After the emulation is finished, the control is then given back to the guest OS, which is unaware of the underyling emulation. Several optimizations techniques have been introduced to reduce virtualization overhead, which will be briefly outlined later.
Virtualization cannot be carried out by the VMM alone, as it does not virtualize hardware and therefore can not grant the guests access to the underlying hardware devices such as network interface, storage drives, and input peripherals. Device models are required for this [7]. They are basically software components that communicate with the shared hardware and expose multiple virtual device interfaces to the VMs. These device models, along with other management software, run in a special privileged virtual machine called management domain which represents the host's operating system and has access to all the underlying hardware. This domain is called domain zero or dom0 in the Xen project and root/parent partition in the Hyper-V project [7]. Since the device models are software-based, they compete for resources for CPU and system resources along with the existing VMs and can negatively affect the performance of these guests. The following figure summarizes the architecture of a

traditional virtualization system.

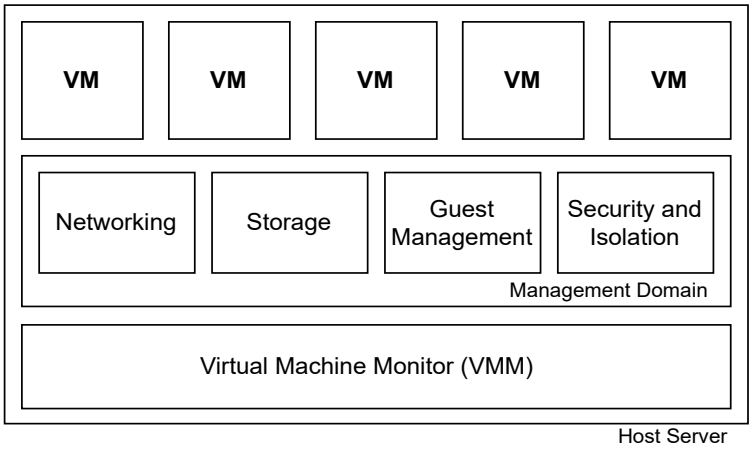


Figure 3.1: Architecture of traditional virtualization Solution

3.1.1 Evolution of Virtualization Solutions

Virtualization technology has evolved significantly. It began with full software virtualization, where the guest OS is unmodified and "unaware" of the virtual environment. Privileged instructions are trapped by the CPU and the hypervisor emulates the sensitive instructions using binary translation [13]. This is, however, very slow and can make the host apps run 2x to 10x slower [13].
Then paravirtualization was introduced, where the guest OS is modified to interact directly with the hypervisor via "hypercalls", removing the abstract emulation layer that is found in full software virtualization. The downside of this approach is that it introduces additional complexity, as it requires modifications to the guest operating system [16].
The next major leap was hardware assisted virtualization (HVM) which introduced virtualization support directly on the hardware level by providing highly efficient and fast virtualization commands. This provides a significant improvement in comparison to the previous virtualization techniques, as it reduces the involvement of the host system in handling privilege and adress translation space tasks [16]. Intel offers this under the Intel VT-x technology that provides virtualization of CPU and memory. Another important example is Single Root Virtualization (SR-IOV) [7], which is a technology that allows physical PCIs device such as Network Interface Card (NIC) to expose multiple virtual devices to the hypervisor. The hypervisor can then provide the different virtual machines with direct hardware access to these virtual devices, which

significantly increases the I/O performance.

3.1.2 The AWS Nitro System

The Nitro System is a result of a multi-year incremental process of AWS re-imagining the virtualization technology in order to optimize it specifically for their EC2 data centers [7]. The main idea was to decompose the software components, i.e., the device models, that run on the management domain and offload them to independent purpose-built server components. This helps minimize the resource usage caused by software running in the management domain, effectively allowing a near "bare-metal" performance. Figure 3.2 depicts the new AWS architecture for virtualization [7].

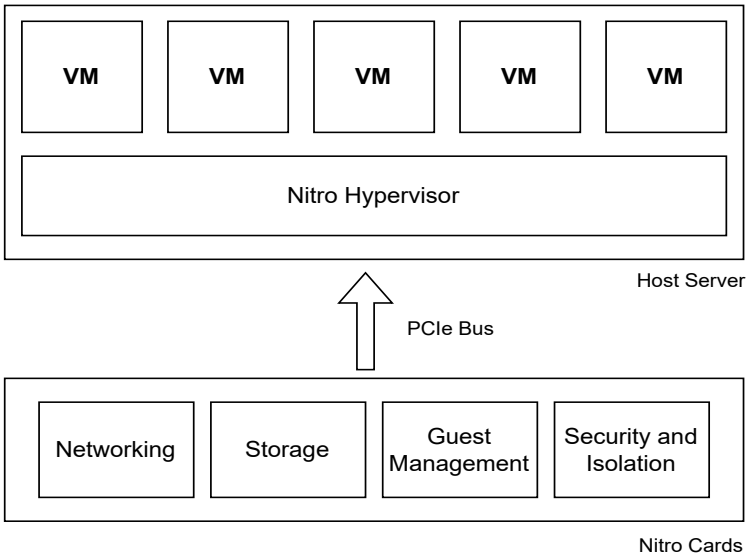


Figure 3.2: Architecture of Nitro System Virtualization

There are three main components in the AWS Nitro System [7].

The Nitro Cards

Nitro cards are dedicated hardware components that operate independently from the EC2's server main board (CPU and memory) and are physically attached to it via PCIe. They "implement all the outward-facing control interfaces used by the EC2 service" responsible for provisioning and managing compute, memory, and storage [7]. They provide all I/O Interfaces as well, such as the ones for storage and networking. These cards employ the previously explained SR-IOV technology to provide direct hardware

interfaces to the VMs. Example of Nitro cards are Nitro cards for I/O and Nitro Controller, which provides the hardware root of trust of the Nitro System.

The Nitro Security Chip

The Nitro Security Chip extends the hardware root of trust and control over the system main board. It's managed by the Nitro Controller and plays a crucial role in enabling AWS to offer bare-metal instances. Bare-metal instances provide direct access to the physical CPUs and memory of the physical server. They are useful mainly for licensing-restriced business critical applications, or for specific workloads that require direct access to the underlying infrastructure.
In virtualized environments, the hypervisor is responsible for securing the host's hardware assets. However, in bare metal modes, when no hypervisor is present, the Nitro Security Chip assumes this role and ensures the security of the system firmware from tampering attempts through the system CPUs [7].

The Nitro Hypervisor

The third component is the AWS Nitro Hypervisor, which has significantly fewer responsibility than traditional hypervisors, as most virtualization tasks are offloaded to the Nitro cards. It has three main functions. It's responsible for partitioning memory and CPU by using the virtualization commands provided by the processor. It's also in charge of assigning the virtual hardware interfaces provided by the Nitro cards to the Virtual Machines and also handles the machine management commands that come from the Nitro Controller (start, terminate, stop etc.) [7].

3.2 Simultaneous Multi-threading (✓)

Before we dive deeper into SMT, it's important to understand which problem it tries to solve and what the motivation behind it is.
A processor consists of a few hundred registers, load/store units and a couple of multiple arithmetic units. The main goal is to keep all these resources as busy as possible. To reach this, multiple techniques have been employed such as instruction pipelining, superscalar architecture and out-of-order execution [27]. Pipelining is a technique that breaks down the execution of an instruction into several distinct stages, with each stage using separate hardware resources [28]. During each CPU cycle, instructions advance from one stage to another. This allows the CPU to work on multiple instructions simultaneously, each being on a different stage. In a perfect scenario, where all instructions are independent, the processor can work simultaneously

on n instructions, with n being the depth of the pipeline, i.e., the number of stages. The following table depicts a simple example of a five-stage pipeline. At the 5th clock cycle, the CPU is simultaneously working on 5 instructions.

Instr. No. \ Clock Cycle	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	IF	ID	EX	MEM	WB		
2		IF	ID	EX	MEM	WB	
3			IF	ID	EX	MEM	WB
4				IF	ID	EX	MEM
5					IF	ID	EX

Figure 3.3: Basic five-stage pipeline (IF = Instruction fetch, ID = Instruction decode, EX = execute MEM = memory read, WB = Write back to memory)

Modern processors are also superscalar. This means that each processor, can start executing more than one instruction per cycle by dispatching them to different execution units [27]. Issue width is an important characteristic of modern CPUs and it represents the maximum number of instructions that can be started in a single clock cycle. Although these optimizations significantly increase the processor throughput, the dependency between the instructions and the long latency-operations of the executing threads limit the usage of the available execution resources [27]. Out-Of-Order execution partially solves this problem but is still not enough as it still dispatches instructions from the same thread, where the dependency between the instructions is inherently high. The wastages that occur on the processor can be categorized into two categories: Horizontal and vertical waste [27]. Horizontal waste occurs when the CPU is not able to fully saturate the issue width of the processor. Vertical waste occurs when the processor is not able to start any instruction at all on a given cycle because of the dependency to the executing instructions or delays such as memory latency. Traditional multithreading addresses this issue by switching to a different thread, whenever the currently executing one stalls.. This approach, however, only mitigates vertical waste, as it still issues instructions from only one thread at any given cycle [27].

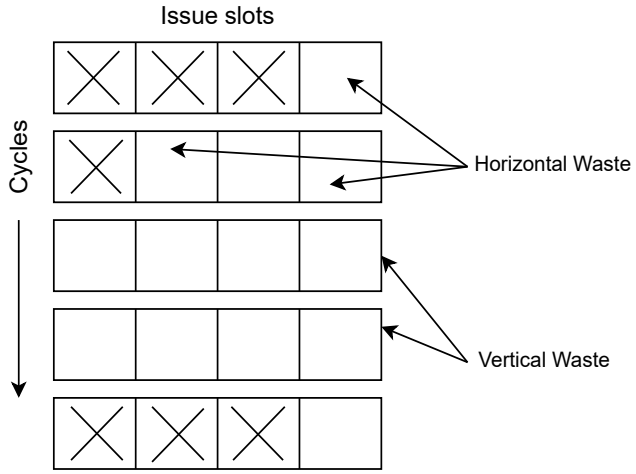


Figure 3.4: Vertical waste vs. horizontal waste

This is where Simultaneous Multi-Threading comes into play. SMT is a technique that helps enhance the overall efficiency of superscalar CPUs by improving the parallelization of computation [28]. This technology allows the physical core to dispatch instructions from more than one thread per cycle without requiring a context switch [27], effectively transforming each physical core into two (or more) "logical" cores. The idea is that instructions from different threads provide greater independency, which results in a better utilization of the core's execution resources [28]. To be able to achieve this, some resources of the processor are duplicated, e.g., those that store the architectural state such as registers and program counters [28]. However, the logical cores still share the same execution resources, which can create conflicts, especially if both threads have the same workload nature, e.g., both are float heavy [21].

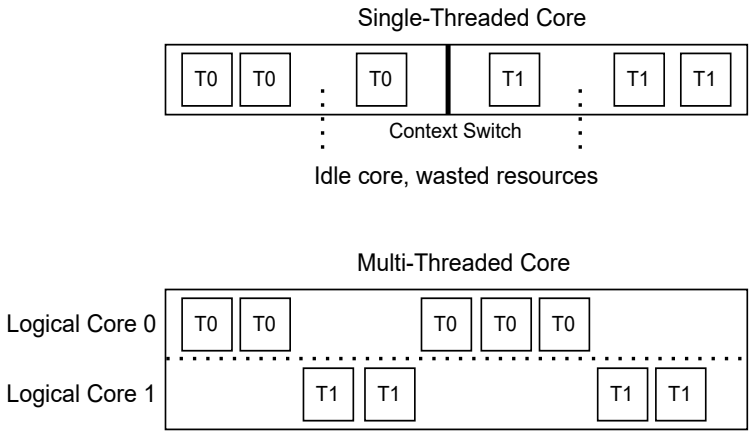


Figure 3.5: Single-Threaded Core vs. Multi-Threaded Core

Both Intel and AMD implement this technology in their modern CPUs, providing two threads per physical core. Intel brands it as Hyper-Threading, while AMD uses the standard term SMT. In the AWS dedicated hosts that run on an Intel or AMD CPU with hyperthreading enabled, the number of vCPUs is always equal to the double of the number of physical cores, with each vCPU corresponding to a hyperthread. This, however, opens up the possibility for CPU contention, if two virtual machines have access to vCPUs that share the same underlying physical core. Unlike Intel and AMD CPUs, AWS-designed Graviton processors, built around the ARM architecture, do not support hyper-threading and expose one execution context, i.e., vCPU for each physical core [25]. This allows for a better isolation between the different tenants as there is no resource sharing between the different vCPUs apart from the last level cache and the memory system [25].

4 Methodology

4.1 CPU Benchmarking

To generate CPU stress, we used sysbench [19], which is a powerful cross-architecture tool, that can be used for Linux performance benchmarks. We used sysbench as a CPU stressing tool. When running with the CPU option, it performs the deterministic task of checking all prime numbers until reaching 10000 (default value) by doing standard division of the current number by all numbers between 2 and its square root [11]. The number of the worker threads can be specified as an argument. The tool allows the specification of the aggregated number of events that should be performed by the created threads. We then use the total execution runtime as a comparison metric between the different experiments. For comparison purposes, we also developed our own CPU stressing tool called *cpu_burn* written in the C language. The program takes two arguments, the first being the number of operations that each created thread will perform and the second representing the number of worker threads. It then returns the total wallclock runtime that was needed for the execution of this job. We compiled the program using GCC with optimization level -O0, to ensure that no compilier optimizations altered the program's behavior. Each thread executes the function defined under `perform_work()`. The argument `work->operations` is specified by the user, as stated previously.

```
1 void* perform_work(void* arg) {
2     ThreadWork* work = (ThreadWork*)arg;
3     double x = 0.0;
4     for (long long i = 0; i < work->operations; ++i) {
5         x += i * 0.000001;
6     }
7     work->result = x;
8     return NULL;
9 }
```

Listing 4.1: Workload of the cpu_burn tool

4.2 Network I/O Benchmarking

4.2.1 Throughput

For network I/O stress, we used iPerf3 [18]. It provides a benchmark for measuring network bandwidth. It supports various protocols and can be used to test TCP, UDP, and SCTP throughput. The tool probes the maximum achievable network bandwidth by transmitting a large number of packets until reaching the thoughput's upper bound. In our experiments, we measured the maximum UDP throughput between clients and servers. We made this choice in order to avoid congestion effects that can be caused by TCP congestion control, which could reduce the throughput even though there still might be bandwidth available.

4.2.2 Latency

For latency benchmarking, we used the sockperf tool [26], which is a network benchmarking utility that can measure the latency of packets at a sub-nanosecond resolution. This tool introduces very low overhead as it uses Time Stamp Counter (TSC) registers that count the number of CPU cycles for measuring latency. We ran the command with the ping-pong argument on the client side and with the server argument on the server side.

4.3 Testing Infrastructure

All tests were performed in the AWS us-east-1 Ohio region. The EC2 instances ran Ubuntu Server 24.04 LTS Linux. For each experiment, all the VMs were provisioned in the same availability zone and resided in the same Virtual Private Cloud. This is particularly important for network I/O experiments, as the network traffic between VMs sharing the same AZ and VPC is free of charge. We ran parallel benchmarks on general-purpose and compute-intensive dedicated hosts from the 5th, 6th and 7th generations. We used mutliple instance types varying from large to 8xlarge instances. To deploy the resources for the different experiments, we used Terraform which is an Infrastructure-as-Code (IaC) tool that's developed by HashiCorp and can be used to define and provision resources using the HashiCorp Configuration Language (HCL), ensuring the automation and reproducibility of the benchmarks. Additionally, we used distexprunner [29], which is a tool written in python that helps write and run bash commands remotely across multiple nodes addressing them through their public IPs. Our experiments generated JSON or csv files that were gathered in an S3 bucket using

distexprunner. We also implemented multiple scripts in Python3 using mainly the re package [10] for parsing raw data and csv [9] for working with comma separated data.

5 CPU Contention

We investigated CPU contention on general-purpose instances from 5th and 6th EC2 generation, with different CPU architectures. We analyzed this effect on processors that support Simultaneous Multi-Threading. We examined two generations of Intel CPUs (m5 and m6i) and m6a instance that runs on AMD processor. We then considered the single threaded AWS graviton 2 processor used by the m6g dedicated host. Key Performance Indicators for these hosts are described in Table 5.1. We notice that the first three dedicated hosts have a number of vCPUs equal to the double of the underlying physical cores. This is because these hosts support SMT with two hyperthreads per physical core. The m6g instance has the best price per vCPU ratio, although each vCPU is mapped to a physical core and not to a hyperthread.

KPI	m5	m6i	m6a	m6g
Processor [8]	Intel Xeon Platinum 8175/ Intel Xeon Platinum 8280	Intel Xeon 8375C	AMD EPYC 7R13 Processor	AWS Graviton2
vCPUs [24]	96	128	192	64
Physical CPUs [24]	48	64	96	64
Clock speed (GHz) [6]	3.1	3.5	3.6	2.5
Hypervisor [1]	Nitro	Nitro	Nitro	Nitro
price/hour [24]	$5.069	$6.758	$9.124	$2.71
price/vCPU/hr	$0.053	$0.053	$0.048	$0.042

Table 5.1: KPIs for AWS Dedicated Host families m5, m6i, m6a, and m6g.

5.1 Contention under SMT

5.1.1 m5 family

The first set of experiments will be conducted on an m5 dedicated host. This host features either the 1st or 2nd generation Intel Xeon Platinum 8000 Series processor, namely Skylake-SP or Cascade Lake [3]. The following table provides an overview of the different instance types that belong to this family.

Instance Size	vCPU	Memory (GiB)
m5.large	2	8
m5.xlarge	4	16
m5.2xlarge	8	32
m5.4xlarge	16	64
m5.8xlarge	32	128
m5.12xlarge	48	192

Table 5.2: m5 Instance Specifications [3]

The m5 dedicated host has 48 physical cores and therefore 96 vCPUs. It features the Nitro v2 Hypervisor [2]. We used terraform to deploy the resources in the us-east-2a zone. In all our experiments, the m5 dedicated host used the 2nd generation Intel CPU. We repeated the experiments using hosts running on 1st generation Intel CPU and we received same results, which excludes the possibility of variation based on hardware heterogeneity. The experiment is structured as follows: We begin by deploying a node, referred to as test node on the dedicated host. Next, we incrementally add neighbors that fully utilize their CPUs. We analyze the effect of adding these neighbors on the runtime of running sysbench and cpu_burn on the test node. Figure 5.1 provides a simple visualization of the experiment.

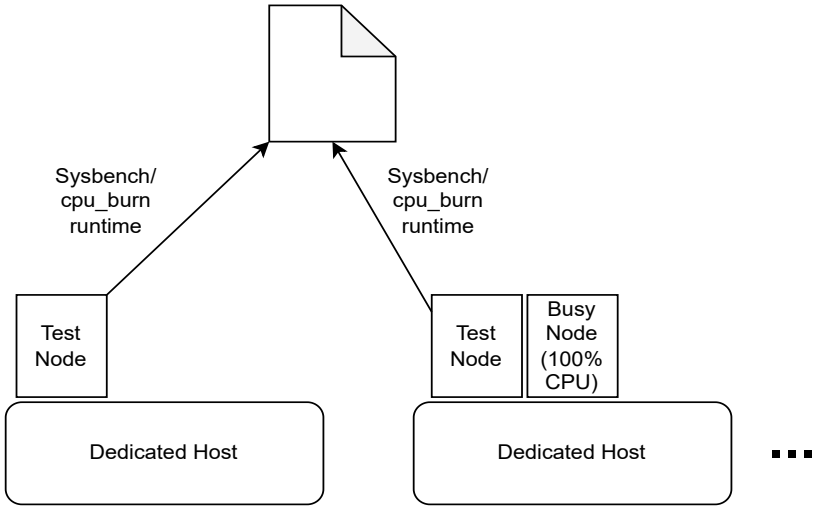


Figure 5.1: Architecture of the CPU contention analysis experiment

For our first experiment, we exclusively used m5.2xlarge instances, each featuring 8vCPUs and 32 GiB RAM. This means that the maximum number of nodes on the dedicated host is 12. The results can be seen in Figure 5.2

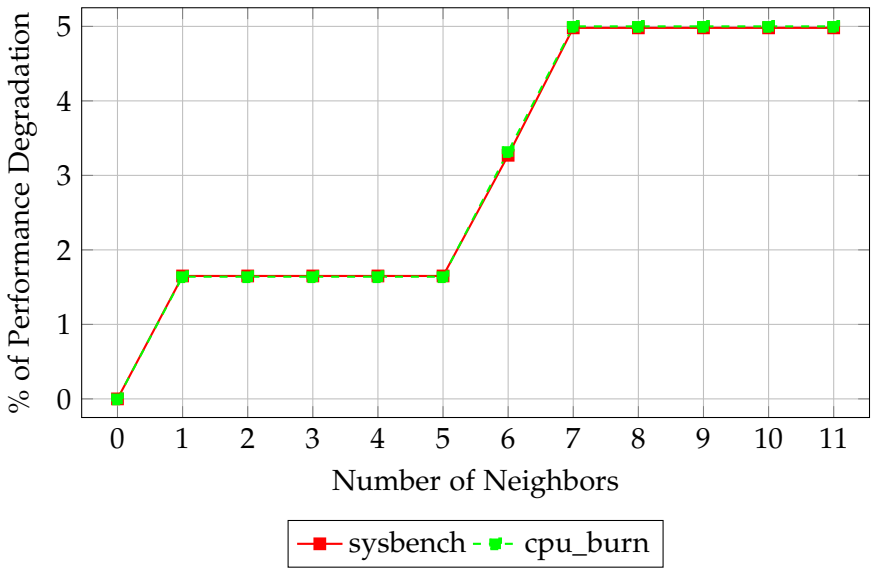


Figure 5.2: Effect of adding busy neighbors on the CPU speed of the sysbench and cpu_burn command on the test node using m5.2xlarge instances

We notice a very similar degradation pattern for the two tools we used. Adding the first neighbor added a performance degradation of 1.6% on our test node. The performance then remained constant for the next 4 neighbors. Afterwards, the 6th neighbor increased this degradation to 3.3%. The 7th neighbor introduced the last witnessed decrease in the performance to reach 5% in both experiments.
This experiment alone does not allow us to pinpoint the reason behind the performance degradation. Potential reasons could be physical core co-location between the neighbors and the test node or hypervisor overhead. The latter is very unlikely as the Nitro system along with hardware assisted virtualization should introduce a very small overhead. We repeat the same experiment but we add idle VMs to see the extent of the performance degradation that happens. Figure 5.3 summarizes the results.

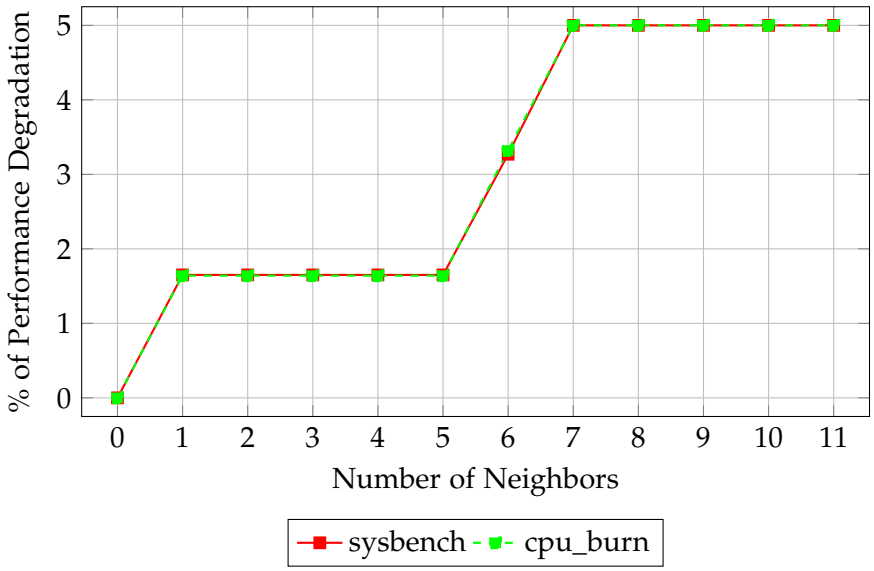


Figure 5.3: Effect of adding idle neighbors on the CPU speed of the sysbench and cpu_burn command on the test node using m5.2xlarge hosts

We notice the exact same degradation pattern of the earlier experiment. This results is very unexpected and undermines the hypothesis that the performance degradation is due to physical core co-location between the different tenants as we would have expected the effect to be less pronounced when adding idle VMs. However, we will refrain from making assumptions until we finish the benchmarking experiments on the m5 family. To investigate this further, we repeat the experiment using m5.large instances, of which the dedicated host can provision 48. The results of our experiment can be seen in Figure 5.4. In this case as well, adding busy or idle neighbors provided the exact same results.

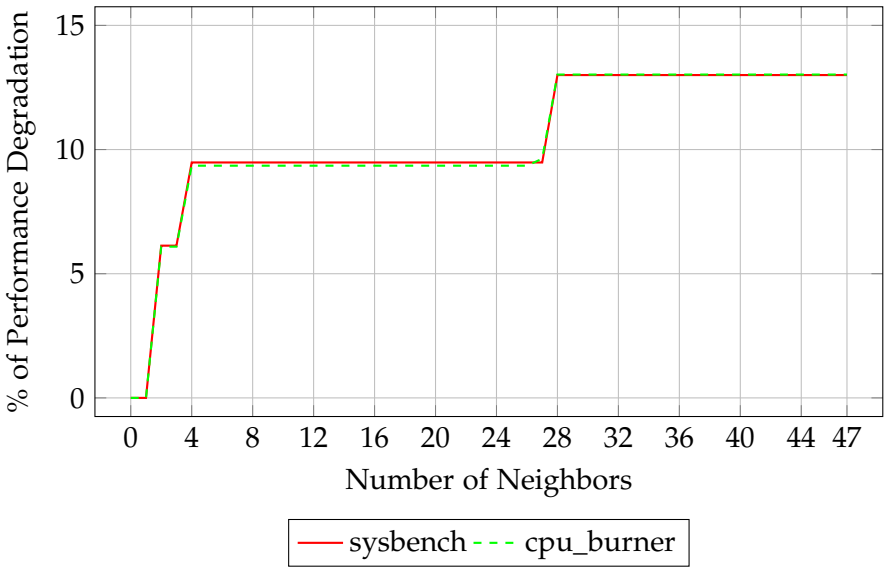


Figure 5.4: Effect of adding busy/idle neighbors on the CPU speed of the sysbench and cpu_burn command on the test node using m5.large instances

In this experiment as well, we notice the same performance degradation between the two tools. The second neighbor has introduced the first performance degradation of roughly 6%. The 4th neighbor increased this degradation to 9,5%. The runtime then remained constant for the next 23 neighbor, as they had no effect on our test node. The 28th neighbor then introduced the last performance degradation reaching the maximum performance degradation of 13% with both tools.
To have a full picture of the performance degradation across the different instance types, we repeated the experiment using the remaining instance types and summarized the results in table 5.3. The results are always similar between the two tools. From this point, we'll only prcoceed with the cpu_burn tool. Adding idle or busy neighbors constantly provided the same results.

Instance type	large	xlarge	2xlarge	4xlarge	12xlarge
Maximum Nodes	48	24	12	6	2
Degradation (Busy/Idle) %	13	13	4.8	3.25	0

Table 5.3: Maximum achievable performance degradation on our test node across various m5 instance types

The biggest performance degradation happens when using large and xlarge instances

with almost the same percentage of 13%. It then drops to 5% for the 2xlarge type, as seen in figure 5.2. We notice a further decrease in the performance degradation for the 4xlarge type to 3.25% and then its complete absence when using the 12xlarge type, of which the dedicated host can only provision 2.
In their paper, Han et. al. [15] argue that this CPU performance degradation is due to CPU context switching overhead that's caused by the KVM (Nitro) scheduler. Even though this hypothesis seems convincing, since the degradation decreases as the number of tenants decreases (Table 2), We think that it's very unlikely, as the Nitro System should provide a near bare-metal performance. To investigate this issue further we run the experiment directly on the bare-metal m5.instance. For this, we use the cpu_burn tool and incrementally increase the number of threads that are created and examine whether we witness any performance degradation. Figure 5.5 visualizes the outcome of the experiment.

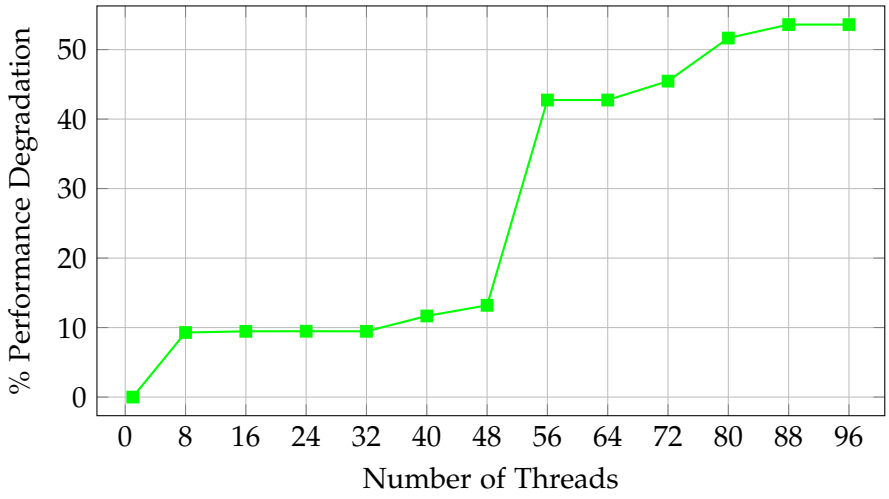


Figure 5.5: Runtime impact of incrementally increasing threads in the cpu_burn command on an m5.metal instance

Although the m5.metal has 96 vCPUs i.e., logical cores, we notice a very important pattern of performance degradation throughout the first 96 threads reaching 53%. This is caused by the physical core co-location that happens between the different threads, resulting in resource contention, as the execution resources are not duplicated. We also notice a very interesting point. In all our previous experiments, the instances initially started with a nominal baseline performance significantly worse than running the exact number of threads directly on the metal instance. Figure 5.6 compares the runtime of cpu_burner on the test node (all threads are busy) as we keep adding fully busy

neighbors, in comparison to running the same number of busy threads directly on the m5.metal.

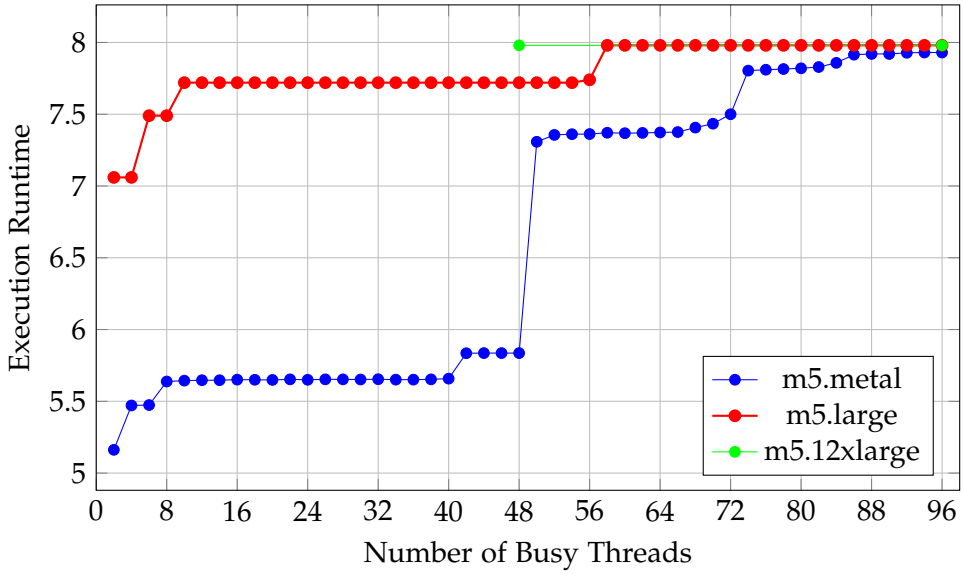


Figure 5.6: Performance of the test node in comparison to running the threads natively on m5.metal

The first two threads finished the execution in 5.165s on the bare metal instance. However, the first m5.large instance that was deployed on the dedicated host took 7.06s to finish, even though it introduced the first two busy threads on the dedicated host and no other VMs exists. This is highly unexpected as we would have expected to see a runtime closer to 5.165s that confirms the promises of AWS of a near bare-metal performance. Instead we notice a difference of almost 36.6%. The same is true for the first m5.12xlarge instance where we witness a difference of 36.7%. Furthermore, we notice that both plots converge almost towards the same value at the maximum number of threads. This strongly undermines the hypothesis that the degradation is due to hypervisor overhead as we would expect to see an even bigger gap (in relation to bare-metal) as more VMs are deployed on the dedicated host. The results for the performance degradation we saw in the previous experiments (Table 2) can be misleading. For bigger instances, we saw a relatively small degradation compared to the smaller instances (large and xlarge). The reason behind this is that the first 12xlarge instance started with a baseline performance that's 13% worse than the performance of the first m5.large instance. We still, however, can't find a plausible explanation, as to why adding busy and idle neighbors provided the same results on all the experiments.

The poor initial performance of the first m5.large instance (test node) suggests that the hypervisor pinned its vCPUs to the same physical core, not taking advantage of other non-occupied physical cores. We assume that this allocation technique aims to avoid contention between the different tenants/customers and isolate the vCPUs of each virtual machine by allocating each pair to the same physical core. This could be advantageous when the customer rents a unique VM, as it is independent of the neighbors' workloads. However, it is highly unfavorable when the user has access to a dedicated host, as they are unable to fully utilize the available idle CPU resources of the physical machine, despite paying for all of them.
To confirm our supposition, we run the cpu_burner tool in a m5.2xlarge instance, and incrementally increase the number of busy threads. The results can be seen in the following figure.

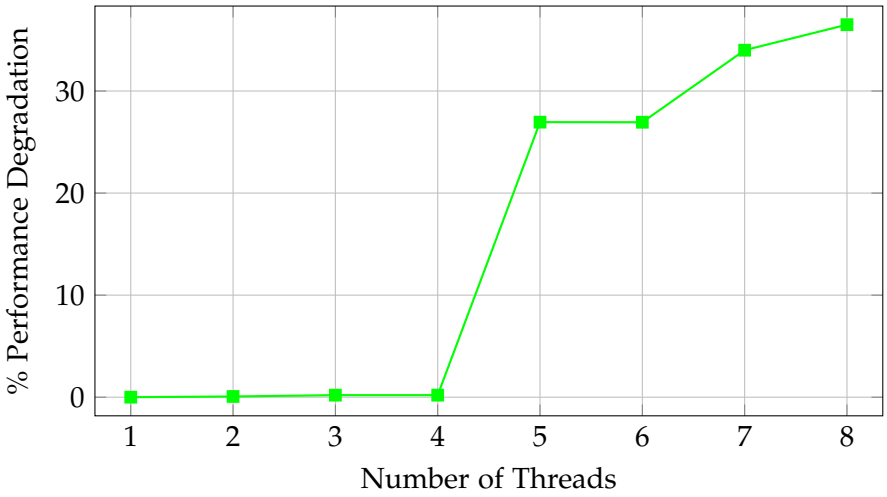


Figure 5.7: Effect of adding busy threads on the CPU performance using m5.2xlarge and the cpu_burn tool

We can notice that the biggest degradation happens when adding the 5th thread, which strongly indicates that the instance has access to only 4 physical cores. The hypervisor starts by initially pinning the first 4 threads to an idle physical core to maximize performance but then the 5th thread is allocated to one of these physical cores, sharing the execution resources with another busy thread resulting in a performance degradation of 27%.

Performance variation of random m5.large instances

We wanted to analyze the variation of the performance of different m5.large instances. For this we sequentially launched 50 m5.large instances across different zones of the us-east-2 region to see where their execution runtime is situated in relation to figure 5.6. The runtime across all the subjects was consistently 7.98 seconds. This consistency strongly suggests that AWS pre-provisions idle instances on internally managed dedicated hosts even before they're rented, enabling faster boot times when a customer actually rents a VM. If this were not the case, then we would have expected to see execution runtimes around 7, 7.5 or 7.7 seconds which correspond to the three performance levels we witnessed on figure 5.6.

5.1.2 m6i family

It's interesting to see whether and to what extent the behavior we saw in the m5 family is present on the m6i family. The m6i family runs on the 3rd Generation Intel Xeon Scalable processor. We investigated the maximum performance degradation that can happen on the test nodes from adding idle and busy VMs with different instance types.

Instance type	large	xlarge	2xlarge	4xlarge
Maximum Nodes	64	32	16	8
Degradation (Busy) %	1.48	1.6	1.74	2.3
Degradation (Idle) %	0.05	0.06	0.06	0.07

Table 5.4: Maximum achievable performance degradation on our test node across various m6i instance types

The difference from the previous experiments with the m5 family is that we notice a difference between adding idle or busy neighbors. Adding idle neighbors always results in a sub 0.1% performance degradation which is practically insignificant. This difference between busy and idle instances can't be assigned to hypervisor overhead alone as we notice a sub 0.1% overhead when adding busy instances on the m6g dedicated host that uses the same Nitro v2 Hypervisor.
These series of experiments can be misleading in suggesting that the performance degradation for the m6i family is better than m5 seeing the percentages. However, in these experiments, all the instance types started from nearly the same nominal performance level, which is roughly 2% away from the highest runtime possible on

the metal instance (128 busy threads). This explains the small levels of performance degradation we witnessed in comparison to the m5 family where the m5.large and m5.xlarge instances started with a relatively better (nominal) performance than the other types resulting in a bigger performance degradation in comparison to the other types. We now analyze the execution runtime of busy threads directly on the m6i.metal instance, in comparison to to the test node while adding busy 4xlarge neighbors. At 128 threads, both the m6i metal and the test nodes had the same execution runtime, which supports the claim that the difference between busy and idle neighbor in Table 5.4 is not solely due to hypervisor overhead.

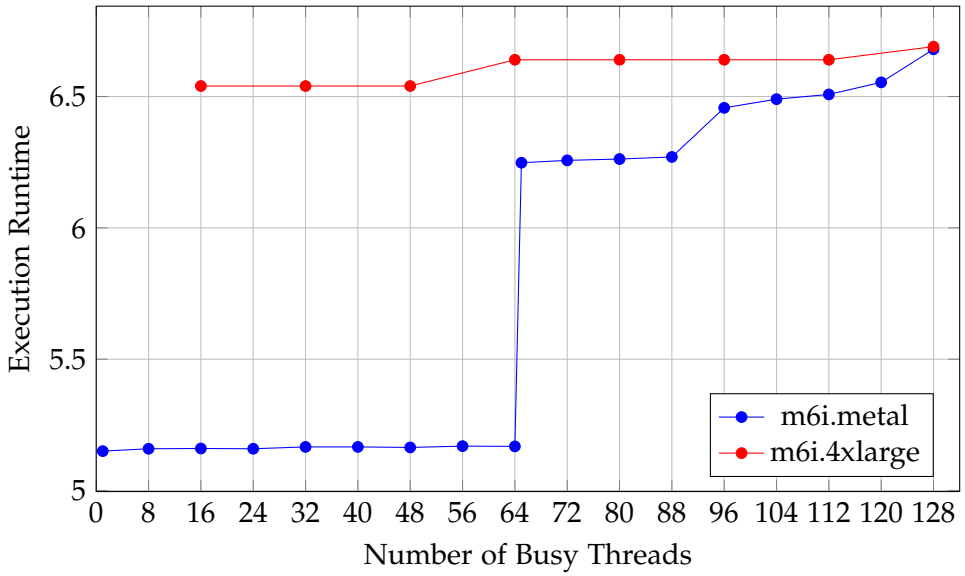


Figure 5.8: Performance of the test node in comparison to running the threads natively on m6i.metal

We notice the same behavior we witnessed on 1st gen and 2nd gen Intel Xeon Scalable processor from the m5.family. We notice that in both the m5.metal and m6i.metal, the most significant performance degradation happens exactly at $n/2+1$ with n being the maximum number of vCPUs available to the dedicated host. Our hypothesis is that the first $n/2$ threads are scheduled each on an independent physical core. However the $n/2+1$ thread needs to share a physical core with another thread, as explained previously. This results in the performance degradation of 20.9% we witness here and 25.22% using m5.metal. The maximum performance degradation using this m6i.metal is less than m5.metal, reaching 31.85% here in comparison to 53%. Over the first 64 threads ($n/2$), we notice very small performance degradation of 0.35%, in comparison

to 13.2% on the m5.metal instance (over the first 48 threads). In this experiment as well, we notice that the the first m6i.4xlarge has an intial performance 26.7% worse than deploying the threads natively on the bare-metal instance. We assume that the same vCPU distribution happens with this host as explained for the m5 host.
We investigate whether this behavior also exists on processors from other vendors that support multi-threading such as some AMD processors.

5.1.3 m6a family

The m6a host features the AMD EPYC 7R13 Processor that also support SMT with two threads per core.

Instance type	large	xlarge	2xlarge	4xlarge
Maximum Nodes	96	48	24	12
Degradation (Busy) %	10	9.5	11.1	11.4
Degradation (Idle) %	0.05	0.06	0.06	0.05

Table 5.5: Maximum achievable performance degradation on our test node across various m6i instance types

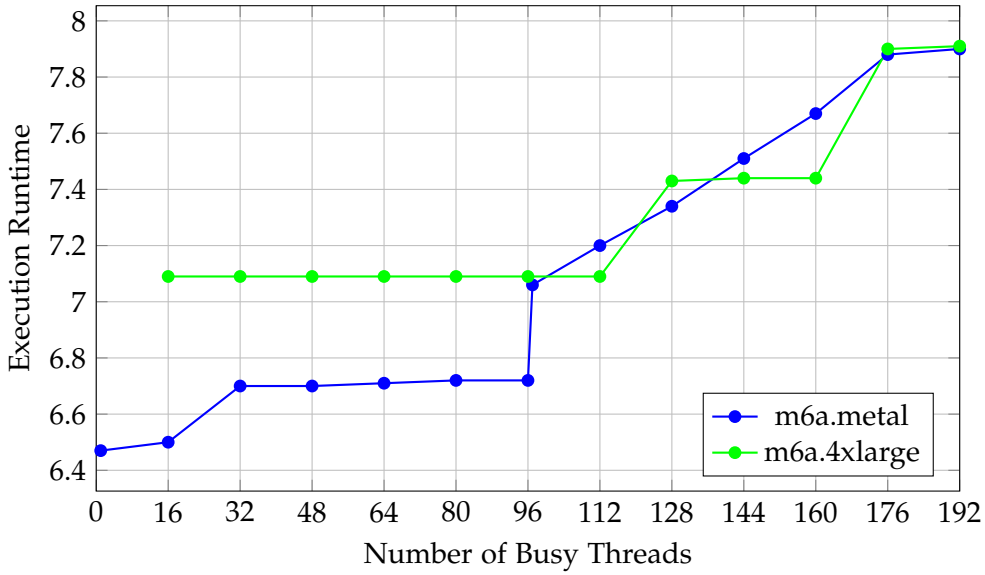


Figure 5.9: Performance of the test node in comparison to running the threads natively on m5.metal

We notice a pattern similiar to that of the Intel processors. However the final performance degradation is less important and is equal to 23.5 %. When adding the 97th $(n/2+1)$ busy thread, we witness a degradation of 5.2%, which is not very important in comparison to the m5 and the m6i dedicated hosts. We notice that the biggest part of the degradation happens in the second half of adding the busy threads, i.e., from thread 97 to 192. We see almost a steady upwards increasing line.

5.2 Contention under Single Threaded Core Processors

5.2.1 m6g family

We now examine CPU contention on the m6g dedicated host that runs on the AWS Graviton2 pro- cessor. It also uses AWS Nitro 2 Hypervisor, which is the same as the m5 family. This host has 64 physical cores and therefore 64 vCPUs. The following table summarizes the instance types we used in the following experiments.

Instance Type	vCPUs	RAM (GiB)
m6g.medium	1	4
m6g.large	2	8
m6g.xlarge	4	16
m6g.2xlarge	8	32
m6g.4xlarge	16	64
m6g.8xlarge	32	128

Table 5.6: vCPU and RAM specifications for AWS m6g instance types

For our first experiment we used m6g.2xlarge nodes, of which the dedicated host can provision 8 instances. The results can be seen in the following figure.

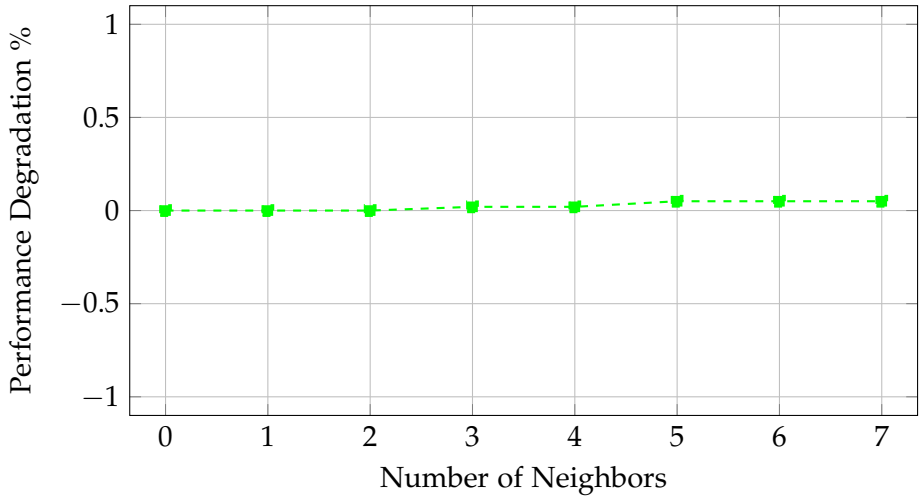


Figure 5.10: Effect of adding busy neighbors on the CPU performance with m6g.2xlarge instances using the cpu_burn tool

We notice a very small and insignificant performance degradation of 0.05%. This should be due to hypervisor overhead which, as claimed by AWS, is practically non-existent. The following table captures the final performance degradation for different instances types. At each level, we repeated the cpu_burn command 10 times and then considered the average of these 10 values.

Instance type	medium	large	xlarge	2xlarge	4xlarge	8xlarge
Maximum Nodes	64	32	16	8	4	2
Degradation (Busy) %	0.05	0.03	0	0	0	0

Table 5.7: Maximum achievable performance degradation on our test node across various m6g instance types

The results of our experiment prove that the AWS Nitro hypervisor causes practically no overhead and the performance is almost indistinguishable from metal as advertised by AWS. We analyze the runtime of the different instance type in comparison to running the threads natively on the m6g.metal instance. The results can be seen in the following figure.

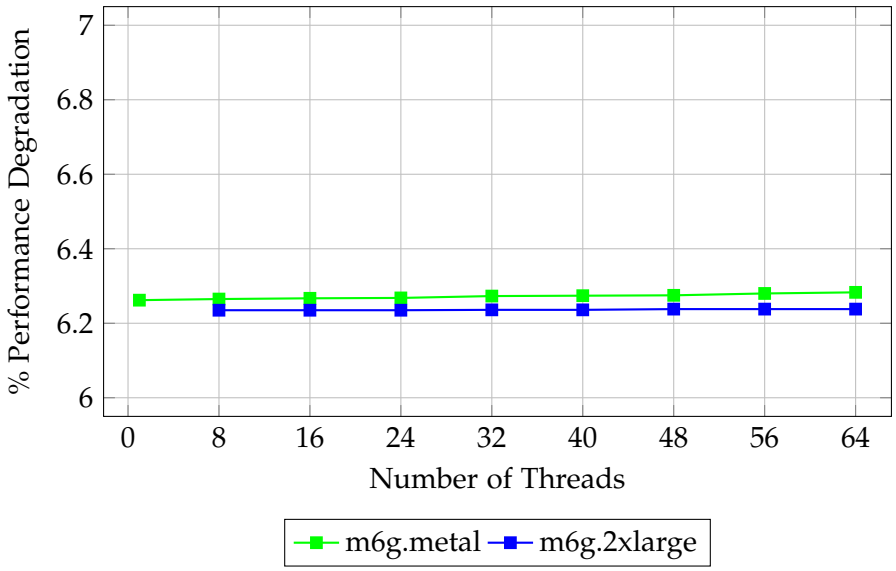


Figure 5.11: Effect of adding threads on the CPU performance using m6g.metal and the cpu_burn tool

We notice a very little performance degradation on the total execution runtime, reaching a maximum of 0.33% at 64 threads. This result is expected as each new thread is assigned to an independent physical core. since the m6g.metal has 64 physical cores, the added threads before 64 should be assigned to an idle core and should practically have no effect on the other threads. We even notice that the m6g.2xlarge had better performance throughout the experiment. However it's a very small difference, which is

almost insignificant.

6 Network I/O Contention

6.1 Throughput

6.1.1 m5 family

In this section, we analyze the network contention that can occur between different residents of the same physical server. The available network bandwidth is a critical performance metric for applications, as it directly affects both the throughput and latency, therefore influencing the overall user experience. Unlike other resources such as CPU and RAM, which are clearly divided between tenants based on the instance type, network bandwidth is shared among the different co-tenants without a precise specification of the expected bandwidth per tenant. Typically, for instances with 16 vCPUs or less, AWS specifies the bandwidth upper bound, e.g., "Up to 10 Gbps". However these instances still have a baseline bandwidth. A network I/O credit mechanism is then employed that allows the instance to use burst bandwidth for a short period of time, from 5 to 60 minutes, depending on the instance's type. The following table depicts all the specifications for the different instances types of the m5 family.

Model	vCPU	Maxiumum Burst Bandwidth (Gbps)	Baseline Bandwidth
m5.large	2	10	0.75
m5.xlarge	4	10	1.25
m5.2xlarge	8	10	2.5
m5.4xlarge	16	10	5
m5.8xlarge	36	10	10
m5.12xlarge	72	12	12
m5.metal	96	25	25

Table 6.1: Specifications of m5 Instance Types

For single flow traffic, the maximum burst bandwidth of 10 Gbps is only attainable when the the client and the server reside in the same cluster group. For instances who are not in the same cluster group, single flow traffic is limited to 5 Gbps. Bandwidth throttling for smaller instances takes at least 5 minutes to take effect, during which the instance has access to 10 Gbps burst bandwidth. We conduct our experiments in this time window to observe the impact of neighboring instances that are fully utilizing their bandwidth on the test node. It is particularly interesting to observe the extent of the network degradation in comparison to the baseline bandwidth for each instance size. The experiment is structured as follows: We use two m5 dedicated hosts, one that hosts all the (iPerf3) clients and the other that hosts all the servers. With each increment, we deploy a client node and a server node. We run the iPerf3 server command on the server and execute the iPerf3 client command on the client so that it's fully utilizing the available bandwidth. The clients and the servers can't be in the same cluster groups since they are in different dedicated hosts. This means that the single-flow traffic between the client and the server is limited to 5 Gbps. To bypass this restriction, we create two client connections, that can be achieved using the P option. The results are continuously logged and aggregated at the end of the experiment in an S3 bucket using the distexptunner tool. The first experiment features the m5.4xlarge instance type, of which the dedicated host can accomoadate 6. To provide a better visualization of the results, we present each node in an independent graph.

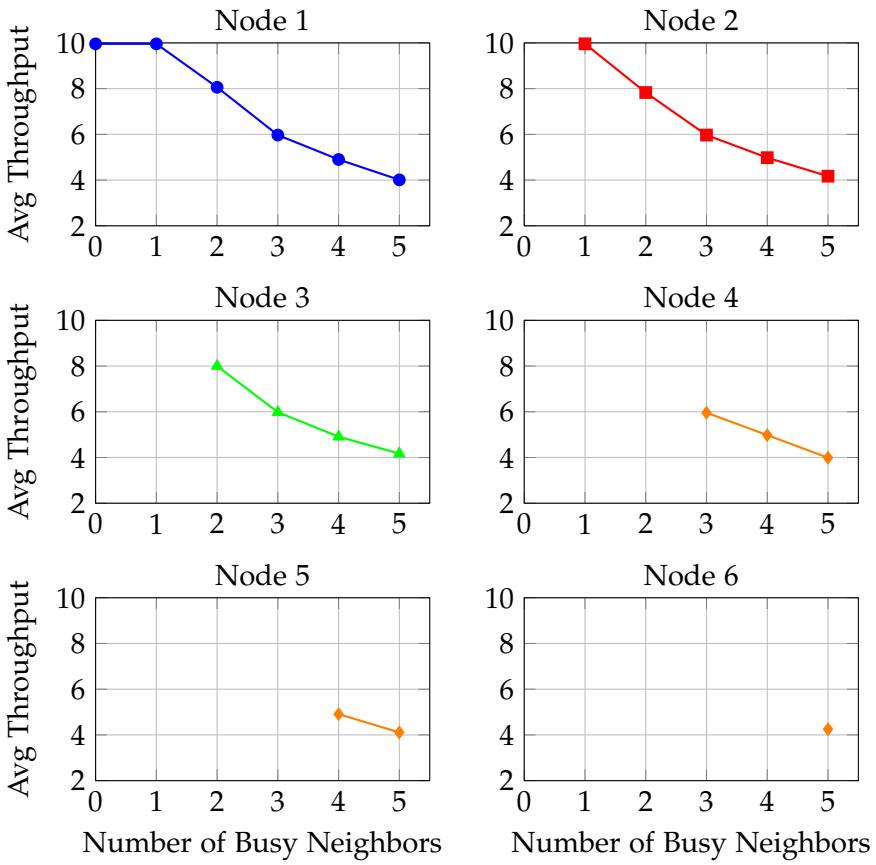


Figure 6.1: UDP Throughput of m5.4xlarge nodes when incrementally increasing the tenants

As expected, the first and the second clients, when alone on the dedicated host had access to a burst bandwidth of 9.96 Gbps. The third tenant caused the average throughput to drop to 7.96 Gbps, while the addition of a fourth neighbor further reduced it to 5.96 Gbps. We practically notice no variation between the different nodes. The fifth neighbor reduced the throughput of all the co-tenants to practically the baseline width of the m5.4xlarge (5 Gbps) with an average throughput of 4.94 Gbps. The 6th Node introduced the first significant decrease under the baseline width to an average of 4.12 Gbps, which is 17.6% less than the baseline width. Starting from the the 3rd tenant, the sum of the throughput across all the nodes is always around 24.7 Gbps. This was expected as the bandwidth of the m5.metal is 25 Gbps, which should represent the upper limit for the sum of the throughputs of all the nodes residing on the same dedicated host. For the xlarge, 2xlarge, and 4xlarge types, the product of the maximum number of the instances on the dedicated host multiplied by the baseline width of

the respective instance type (5 Gbps) is 30 Gbps, which is 16.7% smaller than the possible bandwidth of 25 Gbps. This explains the average degradation of 17% we saw in the previous experiment. The same behavior should be expected when using xlarge and 2xlarge instances. For the large instance type, however, the product is equal to $48 \times 0.75 = 36$ Gbps. 25 Gbps is 30% smaller than 36 Gbps. We should expect to see an average degradation of around 30% at the last level if we repeat the experiment using m5.large instances. We verify this assumption in the next experiment. Since the dedicated host can host 48 of m5.large instances, we can't individually plot the graph of each instance. We used a plotbox graph to display the distribution of the throughputs at the different levels. We also used a step size of 8 neighbors, to be able to present the results in one graph.

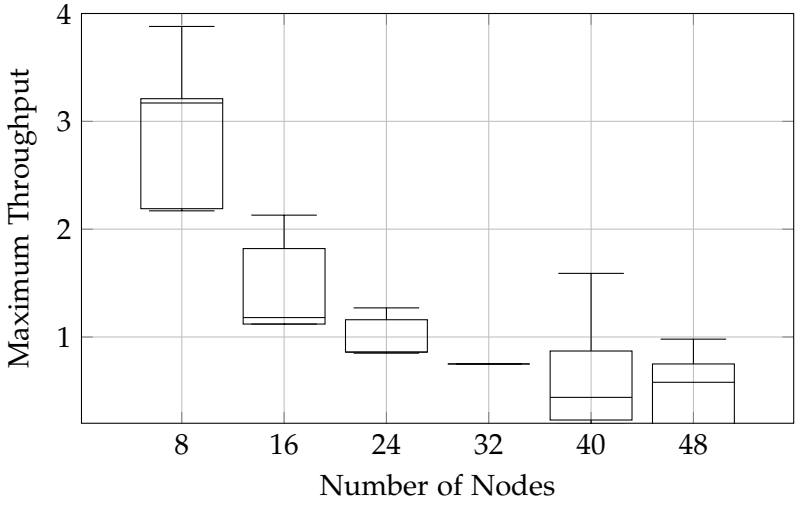


Figure 6.2: Throughput (UDP) of m5.large nodes when incrementally increasing the tenants

The first node has access to a throughput of 9.96 Gbps, which represents the maximum burst bandwidth. At 8 neighbors, the average throughput drops to 3 Gbps. We then observer a gradual degradation with an average of 1.5 Gbps at 16 nodes, 1 Gbps at 24 nodes, and 0.75 Gbps at 32 nodes, which corresponds to the baseline bandwidth of the m5.large instance type. At this level, we interestingly notice zero variation between the different co-tenants. At 40 nodes, the average decreases to 0.614 Gbps and further to 0.51 Gbps at 48 nodes. At full capacity, The average throughput (0.51 Gbps) is 30.7% lower than the baseline bandwidth of the m5.large instance (0.75 Gbps) This more significant performance degradation close to 30% was expected as hypothesized earlier. In this experiment, we also observe an important performance variation between the different nodes compared to the previous experiment.

6.2 Latency

6.2.1 m5 family

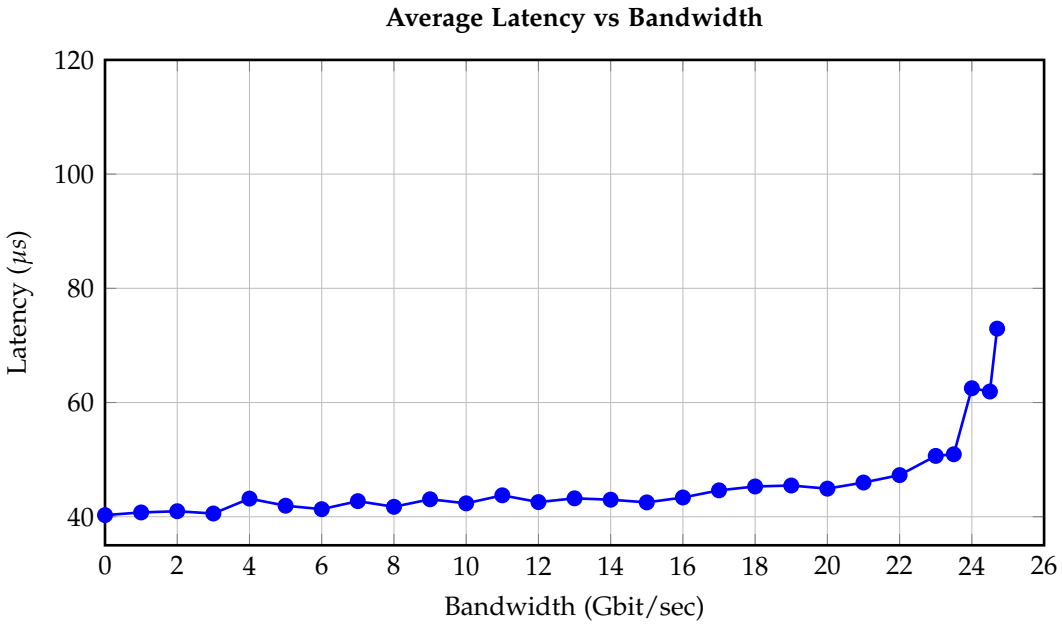


Figure 6.3: Average Latency vs Bandwidth

6.2.2 c7g family

chama

7 Conclusion

Abbreviations

SMT Simultaneous Multi-Threading

AWS Amazon Web Services

VMM Virtual Machine Monitor

IaaS Infrastructure-as-a-Service

IaC Infrastructure-as-Code

HCL HashiCorp Configuration Language

Bibliography

[1] Amazon Web Services. *Amazon EC2 instance types*. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-types.html#instance-hypervisor-type. Accessed: 2025-08-02.

[2] Amazon Web Services. *Amazon EC2 Instance Types - General Purpose Instances*. https://docs.aws.amazon.com/ec2/latest/instancetypes/gp.html. Accessed: 2025-06-20. 2025.

[3] Amazon Web Services. *Amazon EC2 M5 Instances*. Accessed: 2025-06-12.

[4] Amazon Web Services, Inc. *Amazon EC2 Dedicated Hosts*. Accessed: 2025-07-26.

[5] Amazon Web Services, Inc. *Placement groups for your Amazon EC2 instances*. Accessed: 2025-07-26.

[6] *AWS EC2 Instances – Vantage Instances*. https://instances.vantage.sh/. Accessed: 2025-08-02.

[7] J. D. Bean et al. *The Security Design of the AWS Nitro System*. Tech. rep. Amazon Web Services (AWS), Nov. 2022.

[8] T. DIS. *CloudSpecs*. urlhttps://tum-dis.github.io/cloudspecs/. Accessed: 2025-08-02.

[9] P. S. Foundation. *csv — CSV File Reading and Writing*. https://docs.python.org/3/library/csv.html. Python 3.13.5 documentation; accessed 2025-08-02.

[10] P. S. Foundation. *re — Regular expression operations*. https://docs.python.org/3/library/re.html. Python 3.13.5 documentation; accessed 2025-08-02.

[11] Gentoo Wiki contributors. *Sysbench*. https://wiki.gentoo.org/wiki/Sysbench. Accessed: 2025-05-25.

[12] S. Govindan, J. Liu, A. Kansal, and A. Sivasubramaniam. "Cuanta: quantifying effects of shared on-chip resource interference for consolidated virtual machines." In: *Proceedings of the 2nd ACM Symposium on Cloud Computing*. SOCC '11. Cascais, Portugal: Association for Computing Machinery, 2011. ISBN: 9781450309769. DOI: 10.1145/2038916.2038938.

[13] B. Gregg. *AWS EC2 Virtualization 2017: Introducing Nitro*. Brendan Gregg's Blog. Nov. 2017.

[14] X. Han, R. Schooley, D. Mackenzie, O. David, and W. J. Lloyd. "Characterizing Public Cloud Resource Contention to Support Virtual Machine Co-residency Prediction." In: *2020 IEEE International Conference on Cloud Engineering (IC2E)*. 2020, pp. 162–172. DOI: 10.1109/IC2E48712.2020.00024.

[15] X. Han, R. Schooley, D. Mackenzie, O. David, and W. J. Lloyd. "Characterizing Public Cloud Resource Contention to Support Virtual Machine Co-residency Prediction." In: *2020 IEEE International Conference on Cloud Engineering (IC2E)*. 2020, pp. 162–172. DOI: 10.1109/IC2E48712.2020.00024.

[16] K. Hess. "Understanding Hardware-Assisted Virtualization." In: *ADMIN Magazine* (Aug. 2011).

[17] M. S. Inci, B. Gulmezoglu, T. Eisenbarth, and B. Sunar. "Co-location Detection on the Cloud." In: vol. 9689. Apr. 2016, pp. 19–34. ISBN: 978-3-319-43282-3. DOI: 10.1007/978-3-319-43283-0_2.

[18] iPerf Project. *iPerf – The TCP, UDP and SCTP Network Bandwidth Measurement Tool*. https://iperf.fr/. Accessed: 2025-06-01.

[19] A. Kopytov. *sysbench: Scriptable database and system performance benchmark*. https://github.com/akopytov/sysbench. Accessed: 2025-05-23.

[20] W. Lloyd, S. Pallickara, O. David, M. Arabi, and K. Rojas. "Mitigating Resource Contention and Heterogeneity in Public Clouds for Scientific Modeling Services." In: *2017 IEEE International Conference on Cloud Engineering (IC2E)*. 2017, pp. 159–166. DOI: 10.1109/IC2E.2017.29.

[21] T. Moseley, J. Kihm, D. Connors, and D. Grunwald. "Methods for modeling resource contention on simultaneous multithreading processors." In: *2005 International Conference on Computer Design*. 2005, pp. 373–380. DOI: 10.1109/ICCD.2005.74.

[22] A. Rashid and A. Chaturvedi. "Virtualization and its Role in Cloud Computing Environment." In: *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING* Vol.-7 (Apr. 2019), pp. 1131–1136. DOI: 10.26438/ijcse/v7i4.11311136.

[23] M. S. Rehman and M. F. Sakr. "Initial Findings for Provisioning Variation in Cloud Computing." In: *2010 IEEE Second International Conference on Cloud Computing Technology and Science*. 2010, pp. 473–479. DOI: 10.1109/CloudCom.2010.47.

[24] A. W. Services. *Amazon EC2 Dedicated Hosts Pricing*. Accessed: 2025-08-02.

[25] A. W. Services. *AWS Graviton Performance Testing: Tips for Independent Software Vendors*. https://docs.aws.amazon.com/pdfs/whitepapers/latest/aws-graviton-performance-testing/aws-graviton-performance-testing.pdf. Accessed: 2025-07-19. 2021.

[26] M. Technologies. *sockperf: Network benchmarking utility over socket API*. https://github.com/Mellanox/sockperf. Accessed: 2025-08-02. 2025.

[27] D. Tullsen, S. Eggers, and H. Levy. "Simultaneous multithreading: Maximizing on-chip parallelism." In: *Proceedings 22nd Annual International Symposium on Computer Architecture*. 1995, pp. 392–403.

[28] A. Upadhyay. *Two Threads, One Core: How Simultaneous Multithreading Works Under the Hood*. Accessed: 2025-07-19. 2024. URL: https://blog.codingconfessions.com/p/simultaneous-multithreading.

[29] E. de Wit. *erdewit/distex: Distributed process pool for Python*. https://github.com/erdewit/distex. GitHub repository, accessed 2025-06-01. 2024.