

Hypervisor acceleration for lightning-fast computing

Aleksei (Alex) Kharinskii

CNCF Kuala Lumpur, 25.06.2024

Bio

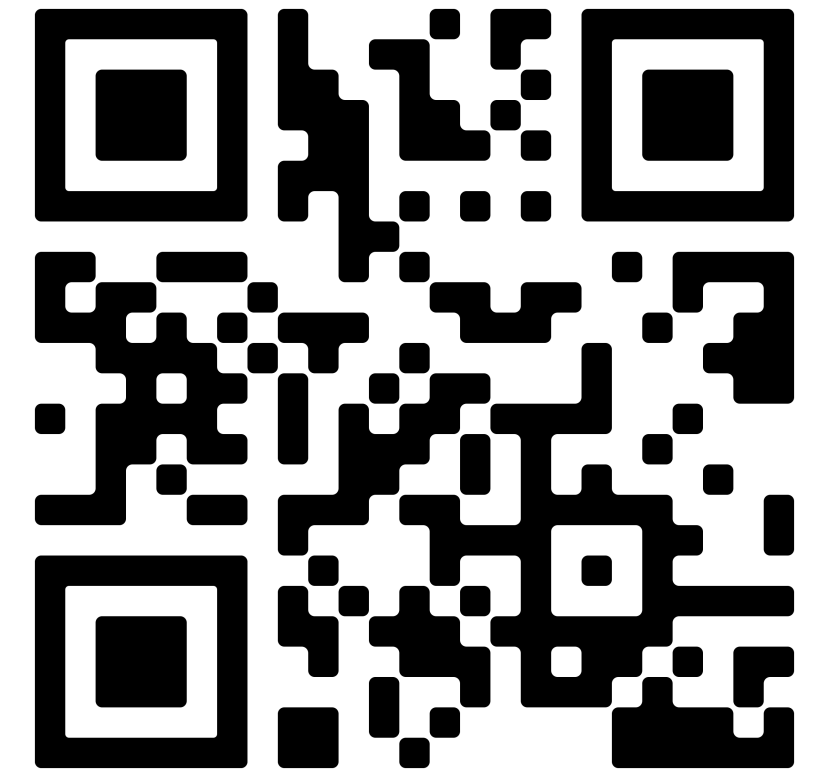


Alex Kharinskii

- Linux engineer since 2010
- Cloud engineer since 2019
- Launched a public cloud platform in 2020
- Solopreneur since 2022



Launched Mealtune
ML-driven nutrition advisor



linkedin.com/in/kharinskiy/

Mealtune



Use case

- Helps people to maintain healthy eating habits with attention to their preferences and diets.
- ChatGPT struggles to interact with real world data: **current user's state, user's diets or time of the day.**

ChatGPT ▾

I ate mushroom egg toast at Brew&Bread in Kuala Lumpur. How many calories in it?



Searched 2 sites

The average calorie count for a mushroom egg toast is approximately 390 calories, based on common recipes and nutritional data from similar dishes. [\[-\]](#)



How accurate this estimation is?



The accuracy of the estimated 390 calories for mushroom egg toast can vary due to several factors:

1. **Portion Size:** The portion size at Brew&Bread might be different from the average portion sizes used in the estimates.
2. **Ingredients:** The specific ingredients and their quantities used at Brew&Bread can differ from those in the recipes used to estimate the calories.
3. **Preparation Method:** Variation in preparation methods, such as the amount of oil or butter used, can significantly impact the calorie count.



Message ChatGPT



Mealtune



Use case

- Started as a project to get people with diabetes insights about the influence of food on the body. Based on private and public research.
- Complex platform that leveraging best practices of data-engineering and machine learning, such as **image recognition and personal NN models.**

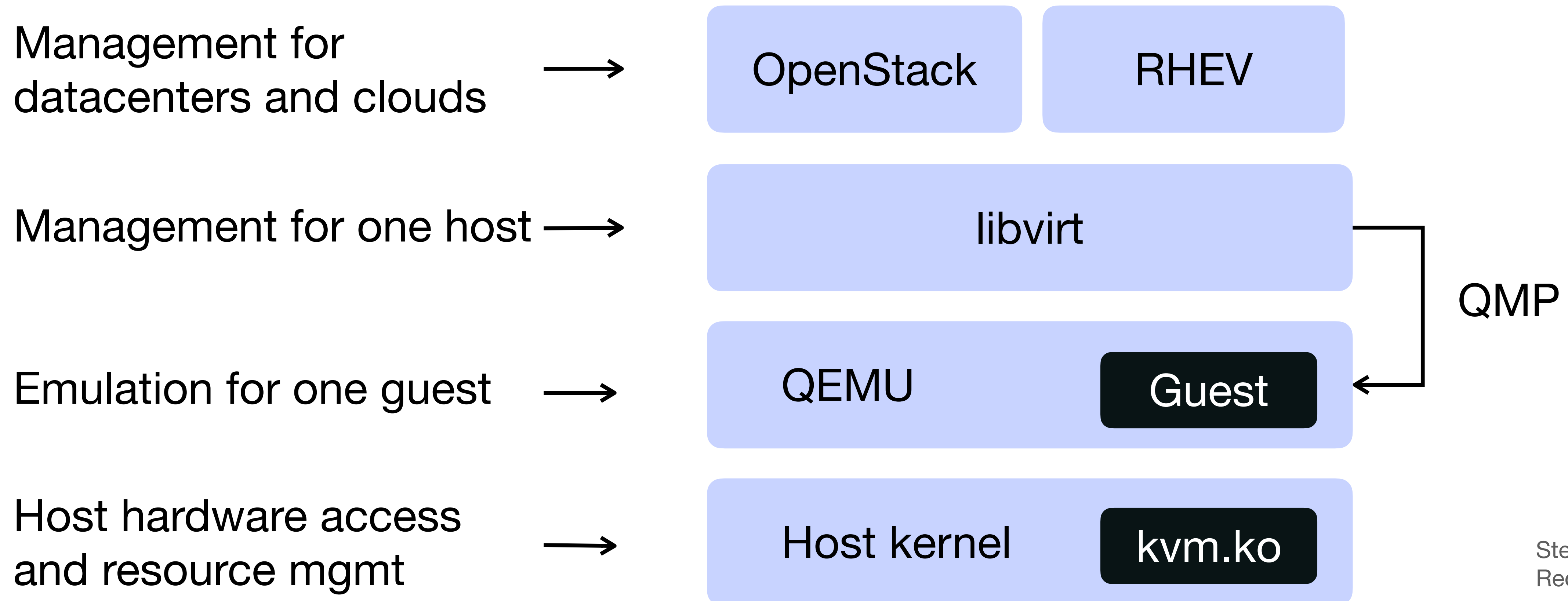
Why virtual machines

...when we have K8S

- Useful abstraction with a strict encapsulation model.
- Suited for many different tasks: networking, storage, computational.
- The performance overhead is relatively low.
- Sometimes it is cheaper to manage VM's infrastructure because of the team expertise.
- No need to rewrite your apps for cloud-native ones (12 factors' apps).

QEMU architecture

Where does KVM fit into the stack?



QEMU architecture

Overview

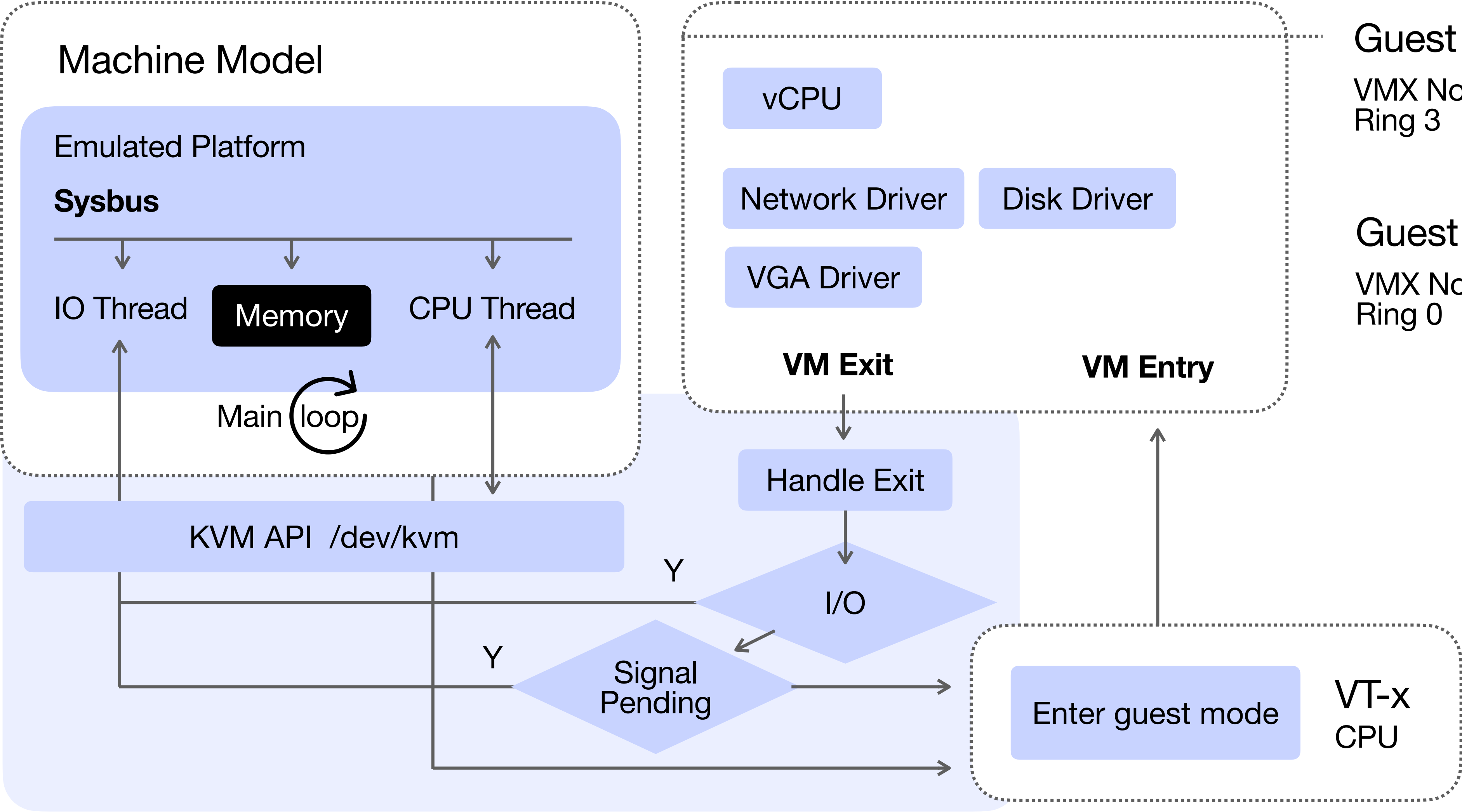
- Guest is a userspace process and vCPU is a thread.
- Unmodified OS images could be run. All the features are fully supported.
- In most cases devices are emulated == resources secured and could be shared.
- Resources could be passed through as dedicated devices.

VT-x hardware virtualization

[The QEMU documentation](#)

QEMU
VMX root
Ring 3

Kernel
VMX root
Ring 0



QEMU architecture

VM management

- Libvirt — the API and management tool for virtualization platforms.
- Has a cmd-toolkit and API bindings.
- Uses the XML-like configuration file format.
- Some changes don't could be applied on fly without any downtime.

QEMU architecture

VM management

```
<domain type='kvm'>
  <name>instance-00029668</name>
  <uuid>4142c6b6-3242-45e2-b763-e21041ee1359</uuid>
  <metadata>
    <nova:instance xmlns:nova="http://openstack.org/xmlns/libvirt/nova/1.0">
      <nova:package version="17.0.14"/>
      <nova:name>ubuntu22_4cpu_4ram</nova:name>
      <nova:creationTime>2024-06-04 19:08:40</nova:creationTime>
      <nova:flavor name="large">
        <nova:memory>8192</nova:memory>
        <nova:disk>0</nova:disk>
        <nova:swap>0</nova:swap>
        <nova:ephemeral>0</nova:ephemeral>
        <nova:vcpus>4</nova:vcpus>
      </nova:flavor>
      <nova:owner>
        <nova:user uuid="6602057e2c2c44058789c88504c37d7e">broker</nova:user>
        <nova:project uuid="1d2936a597584b01afd68a6c3bb1a703">keen-gifted-pogona</nova:project>
      </nova:owner>
      <nova:root type="image" uuid="67525c95-c6ae-4d31-a58b-784d181b8815"/>
    </nova:instance>
```

CPU Workload

CPU Mode

- Two main ways to pass CPU resources to the VM.
- Host-model to determine the CPU model name and all the available features. **Possible lack of features.**
- Host-passthrough — to passthrough the host CPU with no modifications. **Possible migration issues.**
- It also could be set manually by the "custom" mode.

```
(nova-libvirt)[root@cpu56 /]# grep "model name" /usr/share/  
/libvirt/cpu_map.xml
```

```
<model name='486'>  
<model name='pentium'>  
<model name='pentium2'>  
<model name='pentium3'>  
<model name='pentiumpro'>  
<model name='coreduo'>  
<model name='n270'>  
<model name='core2duo'>  
<model name='qemu32'>  
<model name='kvm32'>  
<model name='cpu64-rhel5'>  
<model name='cpu64-rhel6'>  
<model name='kvm64'>  
<model name='qemu64'>  
<model name='Conroe'>  
<model name='Penryn'>  
<model name='Nehalem'>  
<model name='Nehalem-IBRS'>
```

CPU Workload

Neural networks

- Neural networks based on vectors' multiplication operations.
- CPUs have instruction sets to work with floating-point vectors: SIMD (AVX, AVX2, AVX512).

CPU Workload

AMD w7900 pro

shows x4 per token
(\$2.5k for just a card)

```
llama_print_timings:      load time =      260.14 ms
llama_print_timings:      sample time =       77.90 ms /   607 runs  (    0.13 ms per token, 7791.64 tokens per second)
llama_print_timings: prompt eval time =      250.33 ms /    11 tokens (   22.76 ms per token,   43.94 tokens per second)
llama_print_timings:      eval time =   58350.00 ms /   606 runs  (   96.29 ms per token,   10.39 tokens per second)
llama_print_timings:      total time =   60502.14 ms /   617 tokens
```

LLAMA.CPP

Pure Epyc CPU with BLAS
support

~ 40 sec difference for
a bunch of text.

```
Is Pad Thai considered a healthy food?
Is Pad Thai a healthy food? While it can be a tasty and satisfying meal, the answer is not straightforward. Here are some factors to consider:
1. Rice noodles: Pad Thai typically contains rice noodles, which are high in carbohydrates and low in fiber. However, brown rice noodles are a better option than white rice noodles, as they contain more fiber and nutrients.
2. Protein sources: Pad Thai often includes protein sources like shrimp, chicken, or tofu. These can be high in protein, but may also be high in sodium depending on the type of meat used.
3. Vegetables: Pad Thai typically contains a variety of vegetables like bean sprouts, carrots, and bell peppers. These vegetables provide important vitamins, minerals, and fiber. However, some vegetables may be high in sugar or salt depending on the recipe used.
4. Sauces: Pad Thai is often served with a sweet and sour sauce made from tamarind, fish sauce, palm sugar, and lime juice. While these ingredients can add flavor to the dish, they may also be high in sugar or salt.
5. Cooking methods: Pad Thai is often deep-fried or stir-fried, which can increase the fat content of the dish. However, you can prepare Pad Thai using healthier cooking methods like steaming or grilling to reduce the fat content.
6. Portion size: Pad Thai is often served in large portions, which can make it difficult to maintain a healthy weight if consumed regularly. It's important to practice portion control and balance your meals with other nutrient-dense foods.
In summary, while Pad Thai can be a healthy food option when prepared with nutritious ingredients and cooking methods, it's important to be mindful of the rice noodles, protein sources, vegetables, sauces, cooking methods, and portion size. Here are some tips for making a healthier Pad Thai:
* Use brown rice noodles instead of white rice noodles.
* Choose lean protein sources like tofu or chicken breast.
* Incorporate plenty of vegetables like bell peppers, carrots, and leafy greens into your Pad Thai.
* Use a healthier cooking method like steaming or grilling instead of deep-frying.
* Practice portion control and balance your meals with other nutrient-dense foods.
By following these tips, you can enjoy the flavors of Pad Thai while maintaining a healthy diet and lifestyle. [end of text]

llama_print_timings:      load time =      593.26 ms
llama_print_timings:      sample time =     106.14 ms /   605 runs  (    0.18 ms per token, 5700.18 tokens per second)
llama_print_timings: prompt eval time =     959.02 ms /    11 tokens (   87.18 ms per token,   11.47 tokens per second)
llama_print_timings:      eval time =  105034.22 ms /   604 runs  (  173.90 ms per token,    5.75 tokens per second)
llama_print_timings:      total time =  106333.07 ms /   615 tokens
```


CPU Workload

Neural networks

- Make sure the CPU mode is passing the required instructions into the VM.
- The difference with and without AVX is ~x3 on the LLAMA inference stage.

```
root@micro-cloud:~# grep avx --color /proc/cpuinfo
flags               : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush mmx
fxsr sse sse2 syscall nx mmxext fxsr_opt pdpe1gb rdtscp lm rep_good nopl cpuid extd_apicid tsc_known_f
req pni pclmulqdq ssse3 fma cx16 pcid sse4_1 sse4_2 x2apic movbe popcnt tsc_deadline_timer aes xsave a
vx f16c rdrand hypervisor lahf_lm cmp_legacy svm cr8_legacy abm sse4a misalignsse 3dnowprefetch osvw p
erfctr_core invpcid_single ssbd ibrs ibpb stibp vmmcall fsgsbase tsc_adjust bmi1 avx2 smep bmi2 invpci
d rdseed adx smap clflushopt clwb sha_ni xsaveopt xsavec xgetbv1 xsaves virt_ssbd arat npt nrip_save u
mip pku ospke vaes vpclmulqdq rdpid arch_capabilities
```

CPU Workload

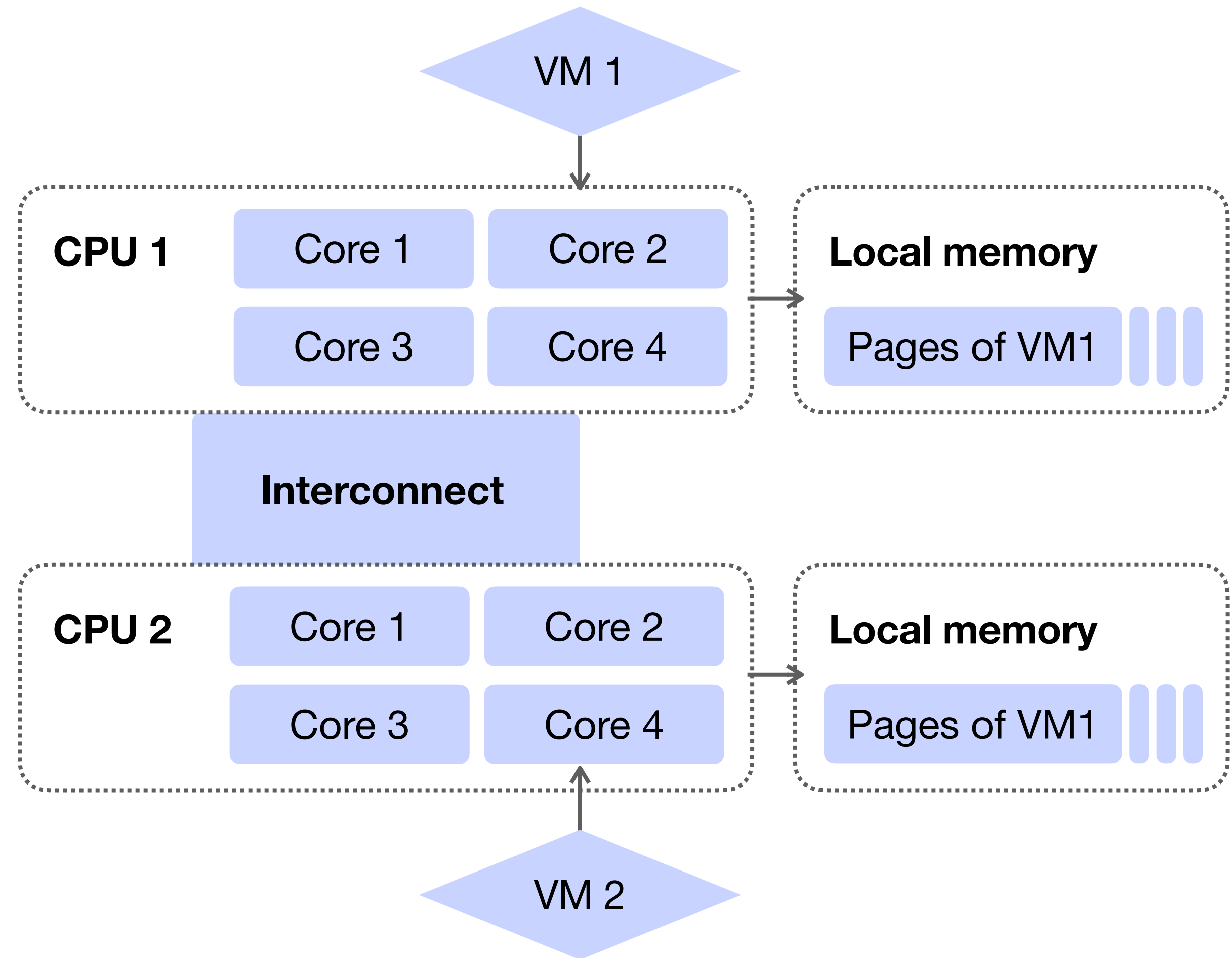
CPU Pinning and isolation

- As vCPU is a thread it could be switched between physical cores.
- VM's vCPU could be pinned explicitly to the certain cores.
- A physical core can be entirely dedicated to a vCPU by the Host.

Memory

NUMA

- CPU's on multiprocessor systems access memory with different bandwidth.
- Neural Network works with weights directly loaded into memory.
- VM could be restricted to use just certain cores to hit a local node.



Memory

NUMA

```
available: 2 nodes (0-1)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 112 113 114 115 116 117 118 119 1
20 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145
146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167
node 0 size: 1019863 MB
node 0 free: 546214 MB
node 1 cpus: 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 168 169 170
171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 1
96 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221
222 223
node 1 size: 1032101 MB
node 1 free: 576069 MB
node distances:
node    0    1
  0:   10   32
  1:   32   10
```

Memory

NUMA

Per-node process memory usage (in MBs)				
PID		Node 0	Node 1	Total
-----		-----	-----	-----
11339	(qemu-kvm)	536	1867	2402
13253	(qemu-kvm)	390	1767	2156
14089	(qemu-kvm)	1682	486	2168
16393	(qemu-kvm)	470	1496	1966
18248	(qemu-kvm)	141	1873	2015

418973	(qemu-kvm	2027	360	2387
427288	(qemu-kvm	2008	81	2090
435470	(qemu-kvm	19427	13637	33064
475805	(qemu-kvm	2851	4671	7522
485714	(qemu-kvm	959	284	1243

Memory

Translation Lookaside Buffer (TLB)

- Map virtual addresses of pages to the real ones.
- TLB is a cache and it is not limitless.
- At the TLB miss the processor iterates through the pages themselves to find the missing address mapping.

Memory

Huge pages

- Native Linux's mechanism to increase a memory page size. The default page size is 4Kb. Could be increased to 2Mb or 1Gb.
- As LLM's parameters should be loaded into memory, huge pages allow to start an LLM much faster.
- **Could** improve performance up to 50% by reducing TLB misses.
- TLB could be flushed and reloaded over time or the content of certain pages could be modified. Depends on the page size and configuration of TLB.

Memory

Huge pages

- Allocated memory footprints will be "huge" as well, decreasing resources utilization.
- The page allocation overhead is also increasing.
- Works well on the inference stage. **Up to 15% growth in performance on LLAMA.**
- **9% drop in performance on the training stage.**

Memory

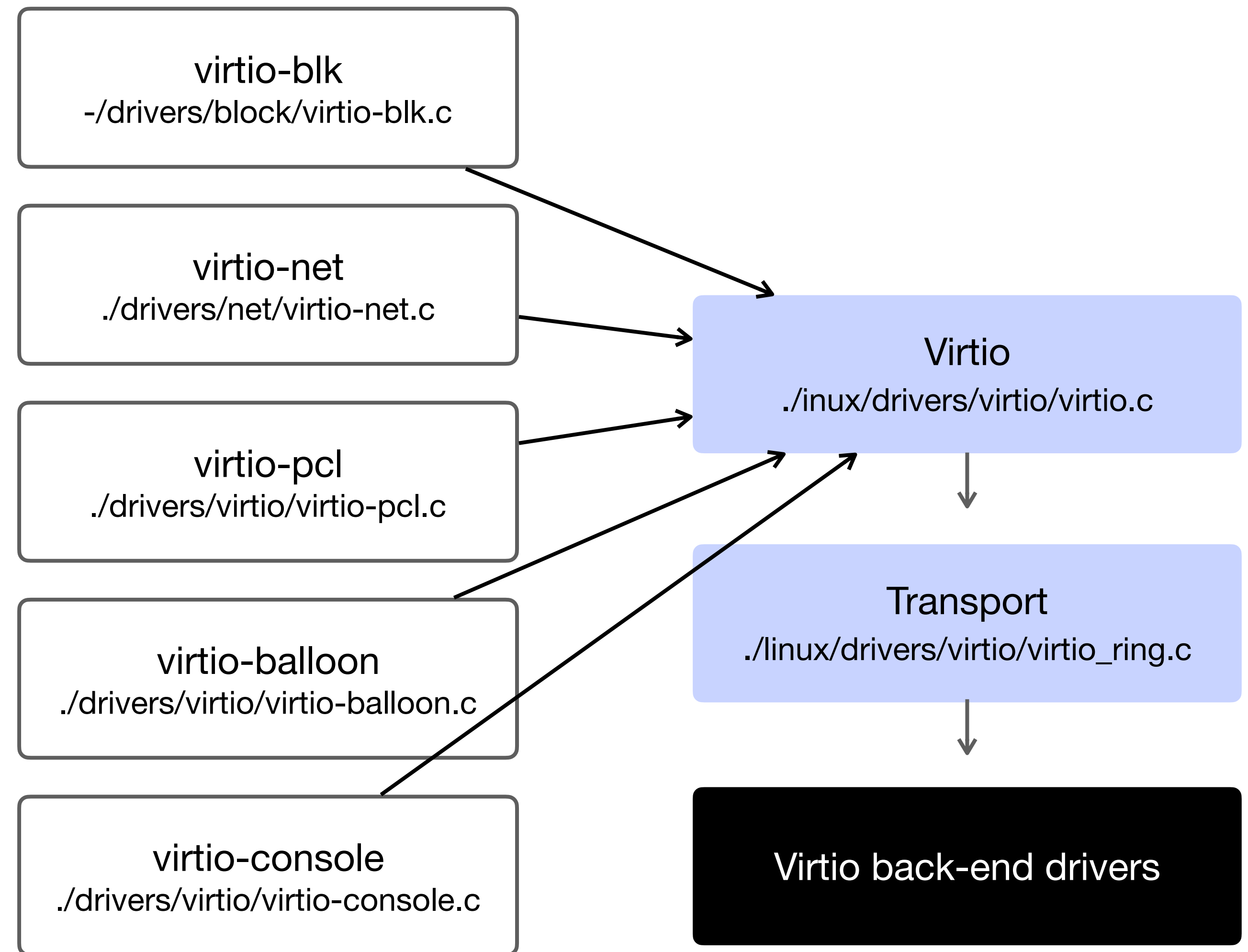
KSM (Kernel Samepage Merging)

- Is the feature that scans memory looking for identical pages and merging them.
- The pages_to_scan parameter could increase impact of the feature but also increases CPU workload.
- **Could affect memory access latency on a VM with Huge Pages. So should be tuned properly or switched off.**

I/O workload

Virtio architecture

- Virtio devices is a virtual abstraction.
- Works on both the virtual machine side and the hypervisor side.
- Different modules for different architectural problems.
- Have low overhead.



I/O workload

LLM

- LLM's data on disk as multiple files or as a GGUF-file.
- Parameters mmaped or copied into memory.
- During the training phase, many files are loaded into memory.
- Multi-Queue **could** help during the training. Should be tested though.

Summary

- LLMs' workloads are highly demanding to all kind of resources.
- Huge Pages can increase memory access performance due to less TLB misses. Depends on the type of the workload.
- The correct NUMA configuration can increase memory bandwidth.
- CPU's flags should be passed to leverage vector's arithmetics.
- I/O loads should be configured on a case-by-case basis.
- **The concepts discussed are general and can be used in K8S configuration.**

Summary

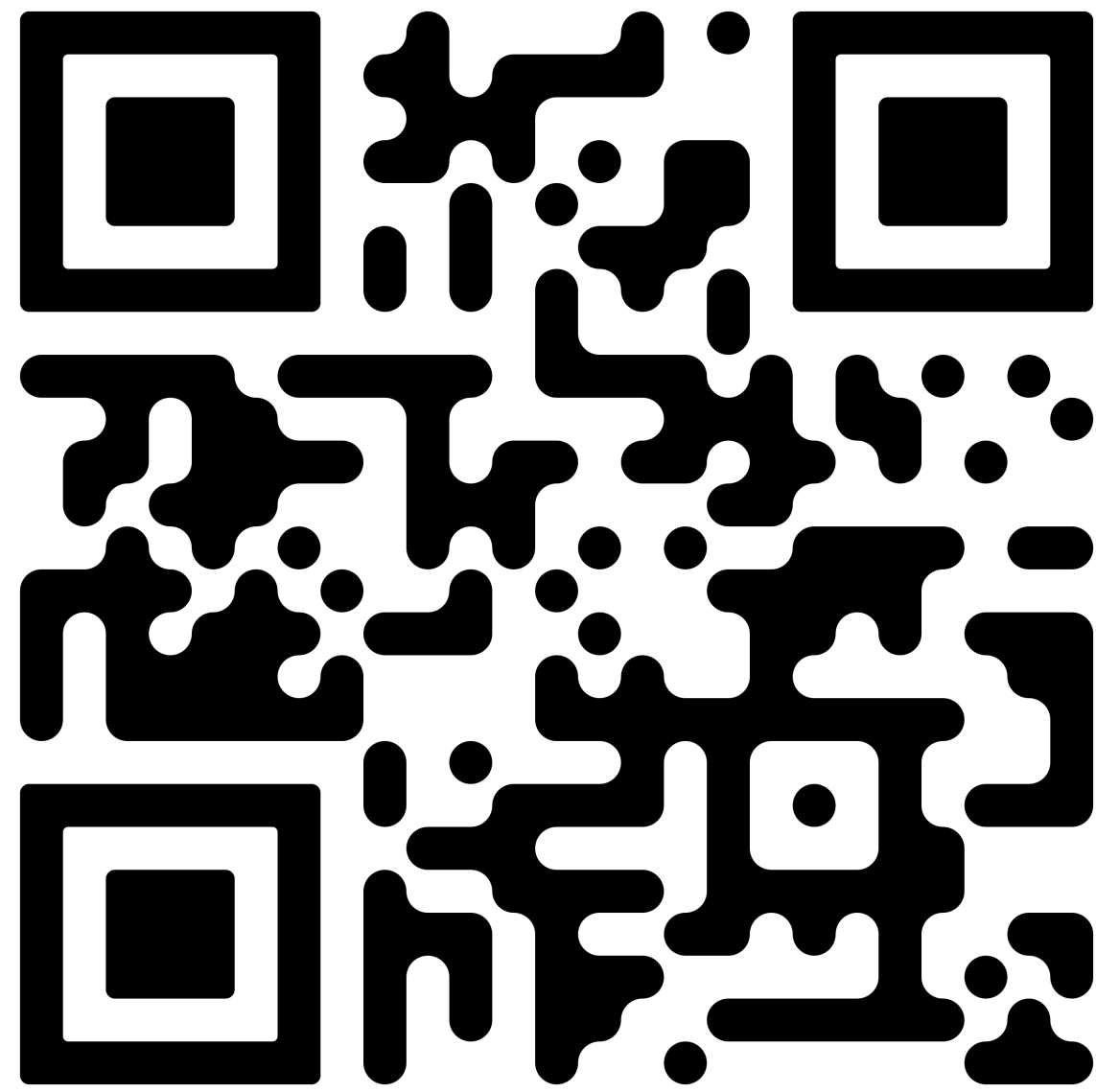
ROI / Resources utilization

- Memory pages footprint. "Stolen" memory.
- A right overcommitting factor could be challenging.
- Most monitoring services look only on general parameters.
Kernel's software queues could be overloaded despite of low LA.
- **Less resources available for same workloads as a result.**

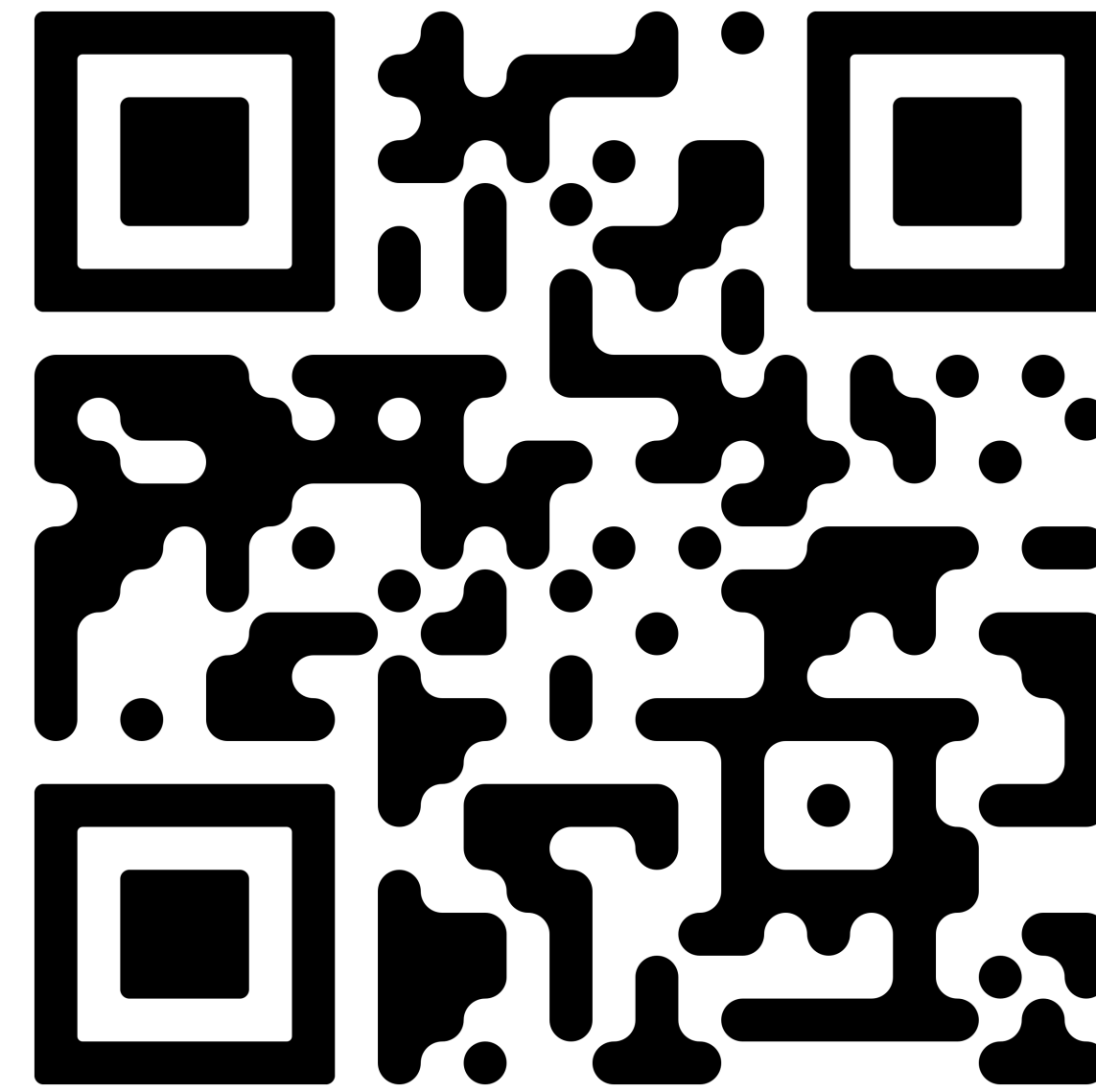
Cases we solved

- Public cloud costs optimization. More resources on a same price.
- Building private cloud solutions tailored to client's needs.
- Partnering with security teams experienced in web, infrastructure and mobile security.

Questions?



Private AI platform



Infrastructure optimization