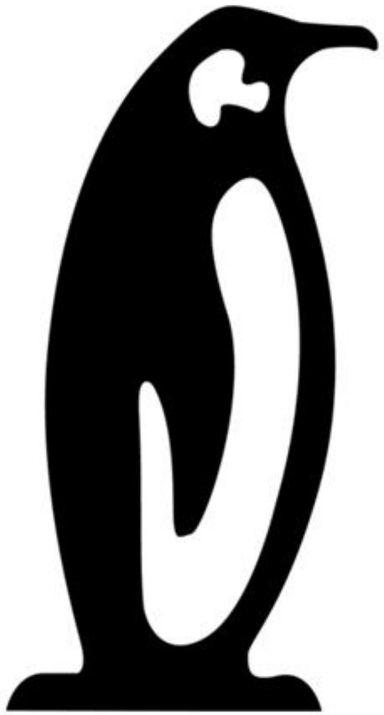


Linux Plumbers Conference

Vienna, Austria | September 18-20, 2024

Challenges in scheduling virtual CPUs

Tobias Huschle <huschle@linux.ibm.com>

IBM

LINUX PLUMBERS CONFERENCE | Vienna, Austria
Sept. 18-20, 2024

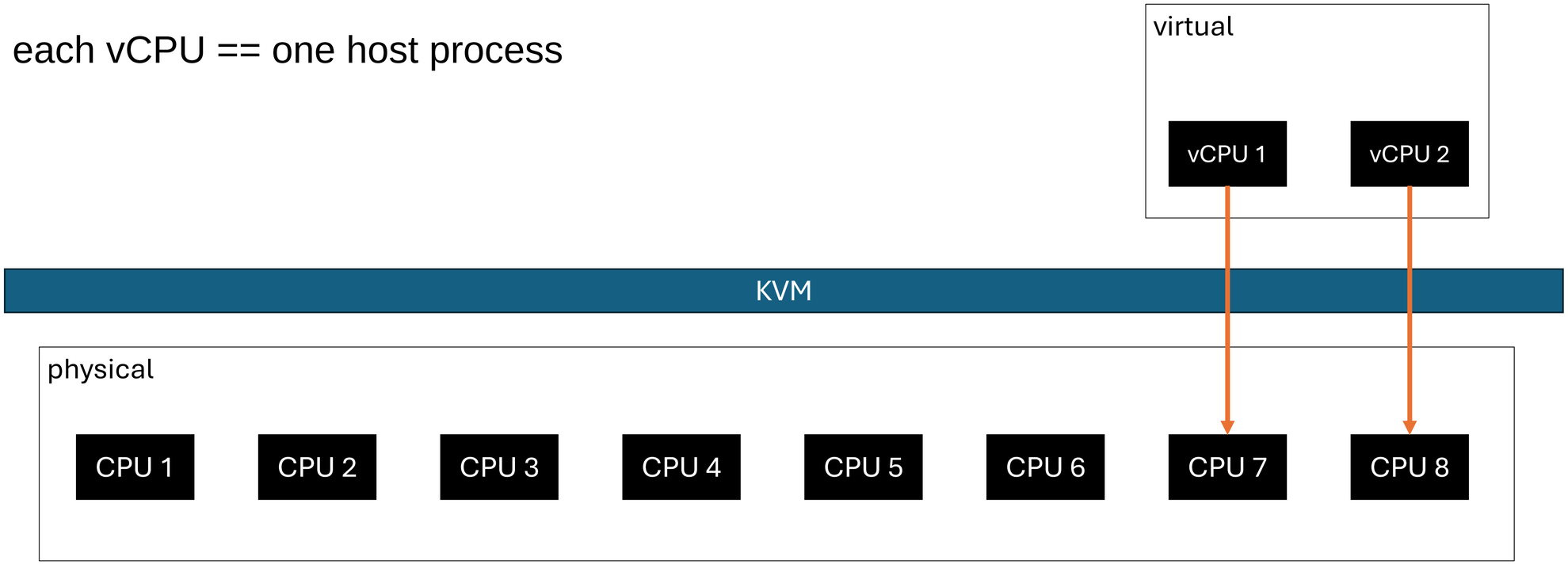
Agenda

- Warm-up: Basics of virtual CPU scheduling
- Complexities
  1. Host overhead
  2. Virtualized infrastructure
  3. Overcommitment
- The s390 approach

Warm up

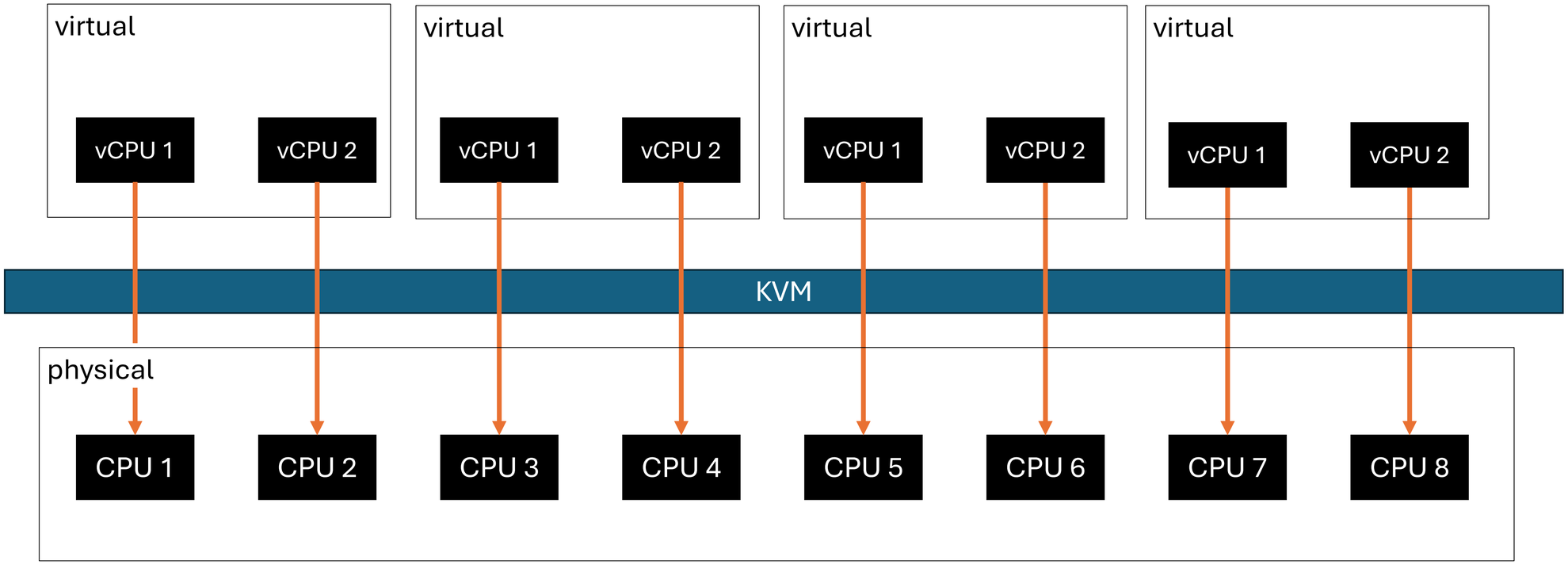
Basics of virtual CPU scheduling

each vCPU == one host process

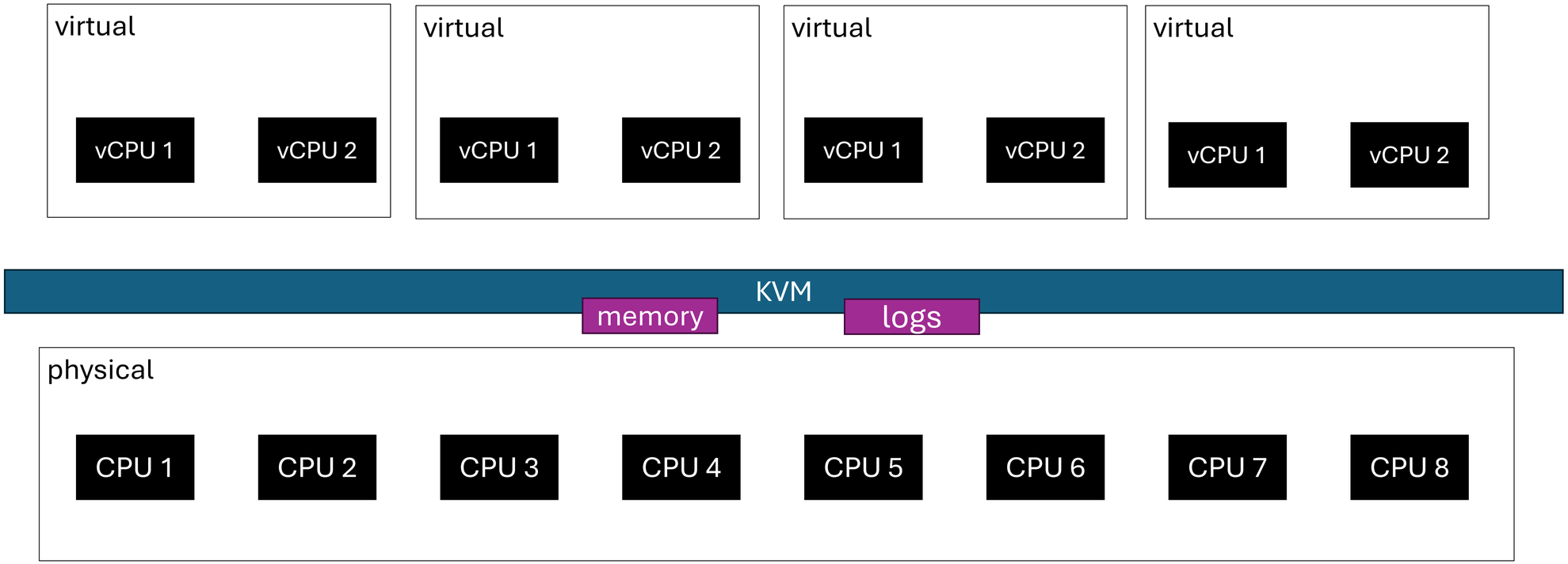


Warm up

Basics of virtual CPU scheduling

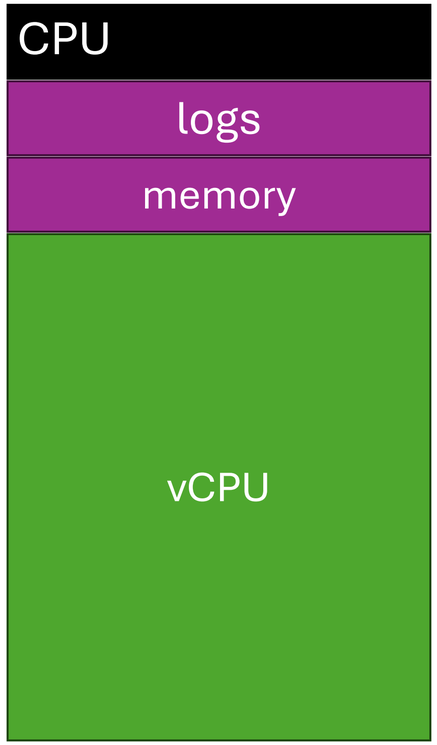


Complexity 1: Host overhead



Complexity 1: Host overhead

Utilization of physical CPUs

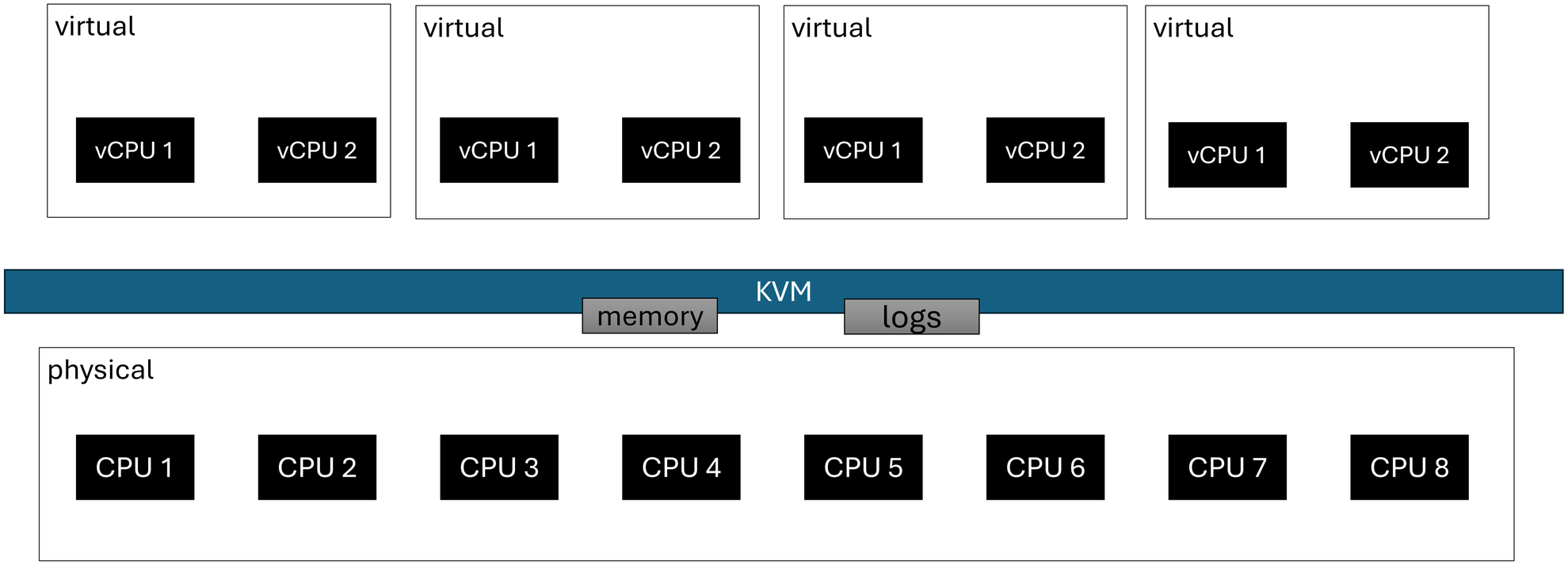


vCPU cannot use 100% of physical CPU

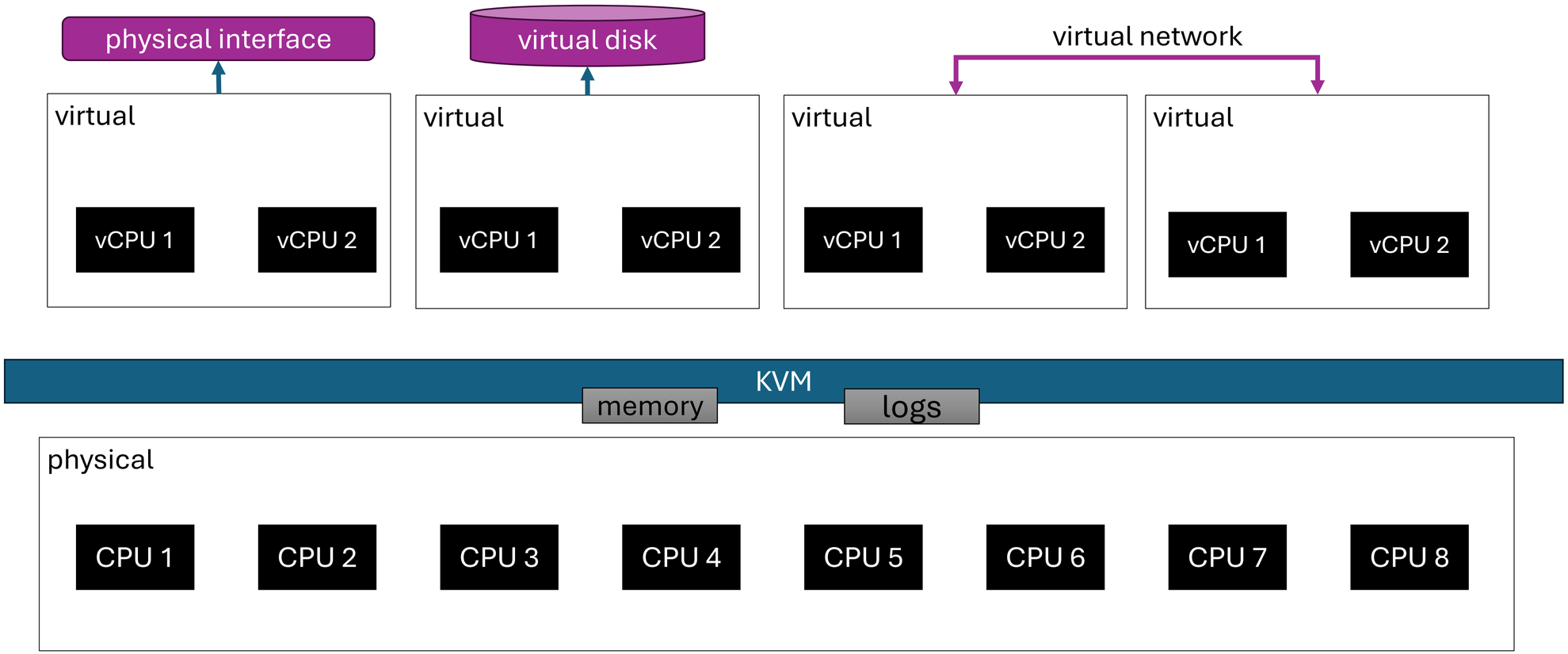
Host has do decide when to schedule vCPU away

Complexity 2: Virtualized infrastructure

Complexity 2: Virtualized infrastructure

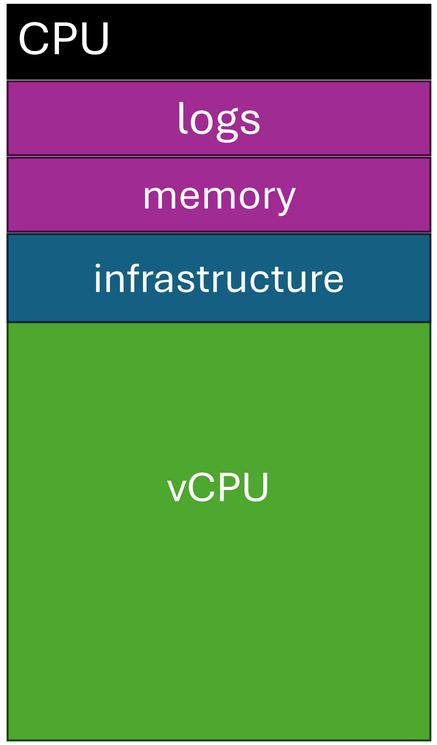


Complexity 2: Virtualized infrastructure



Complexity 2: Virtualized infrastructure

Utilization of physical CPUs

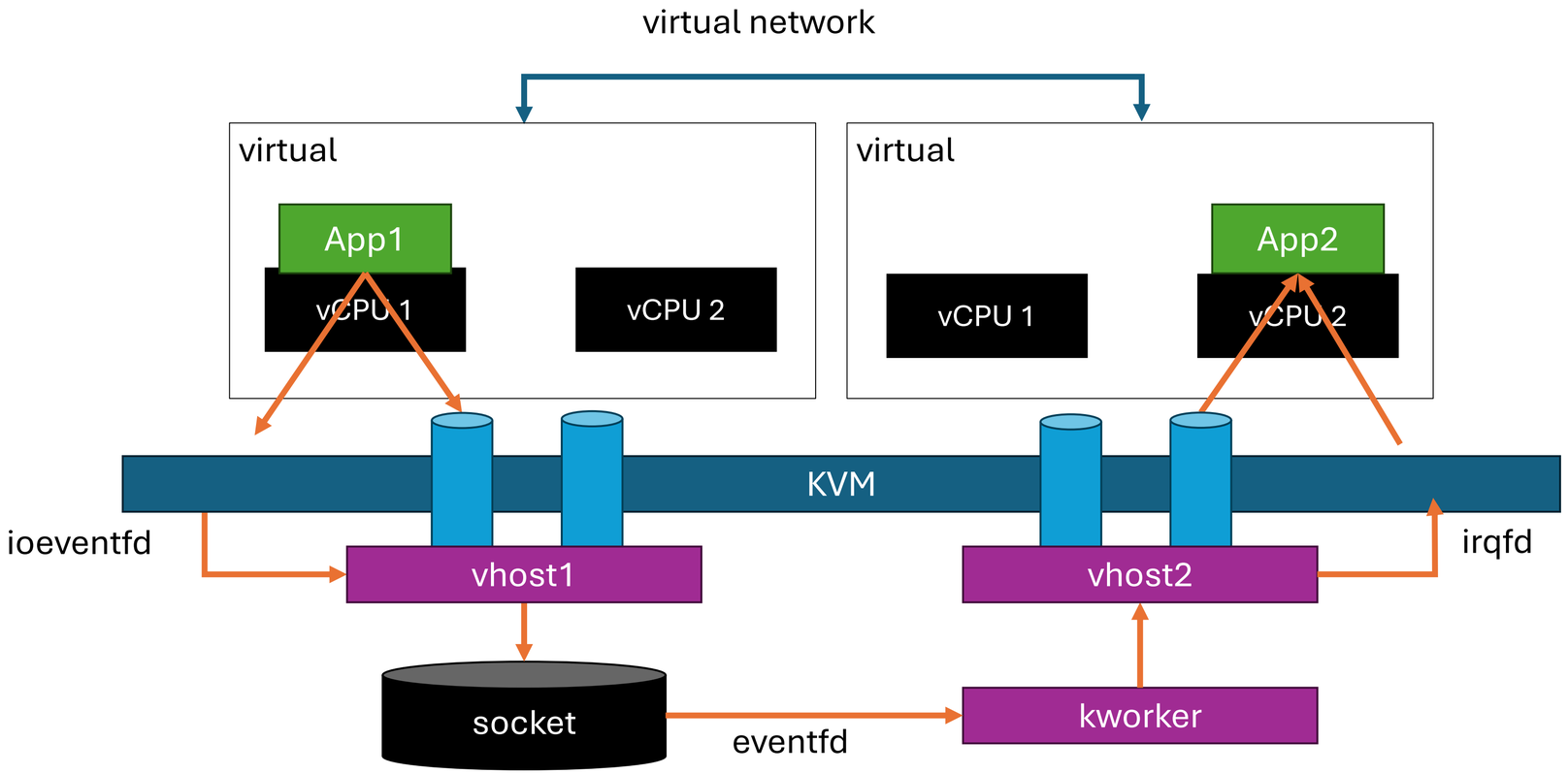


Infrastructure needs to be mapped to physical CPU

Increased contention for physical CPUs

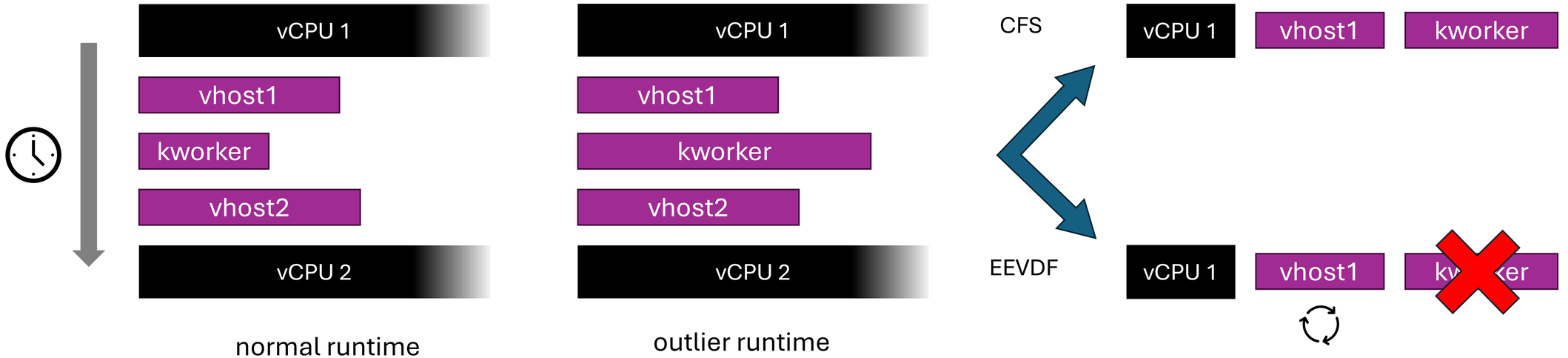
Complexity 2: Virtualized infrastructure

Example: vhost



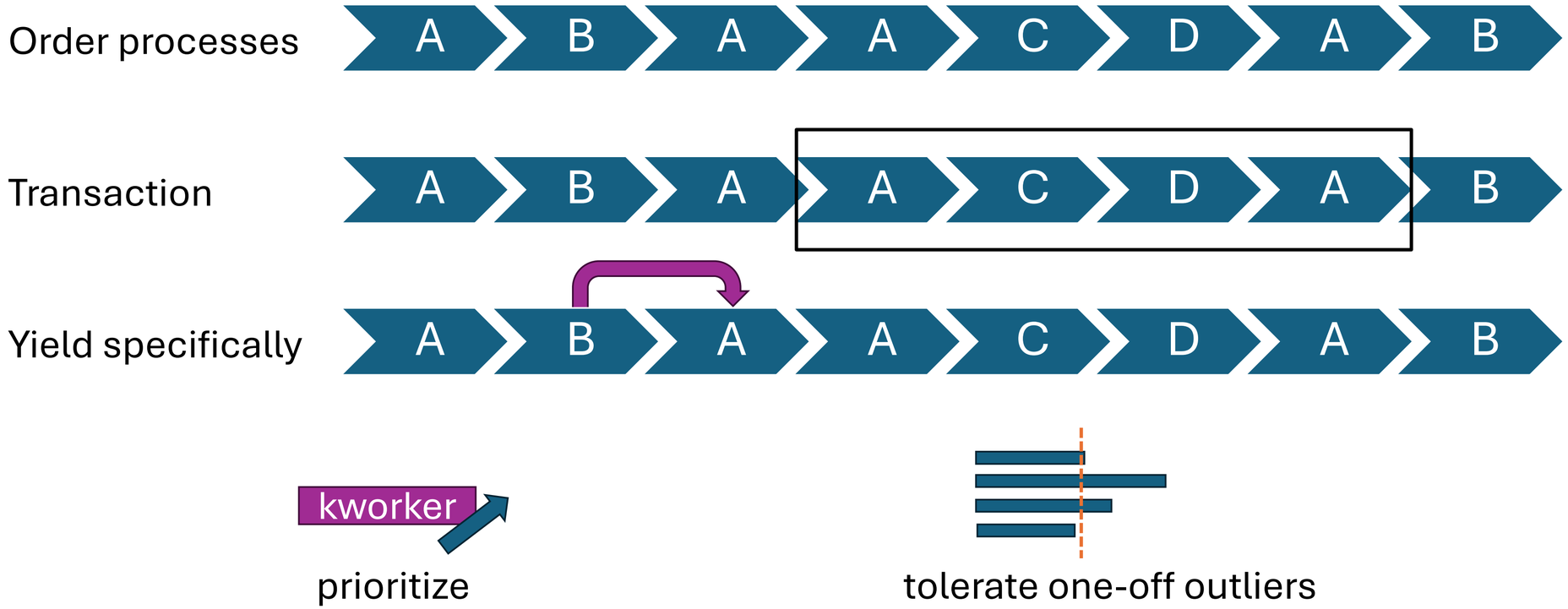
Complexity 2: Virtualized infrastructure

Scheduling impact



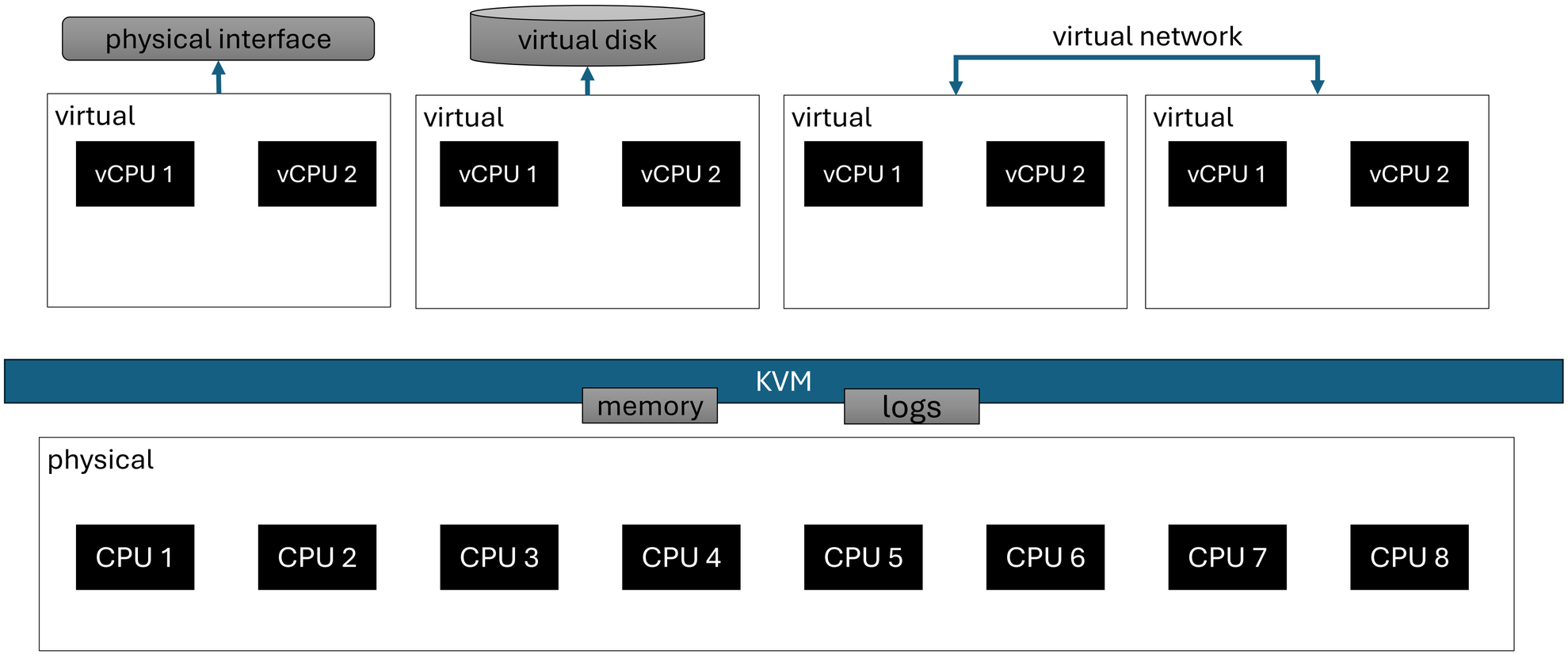
Complexity 2: Virtualized infrastructure

Ordering: possible solutions

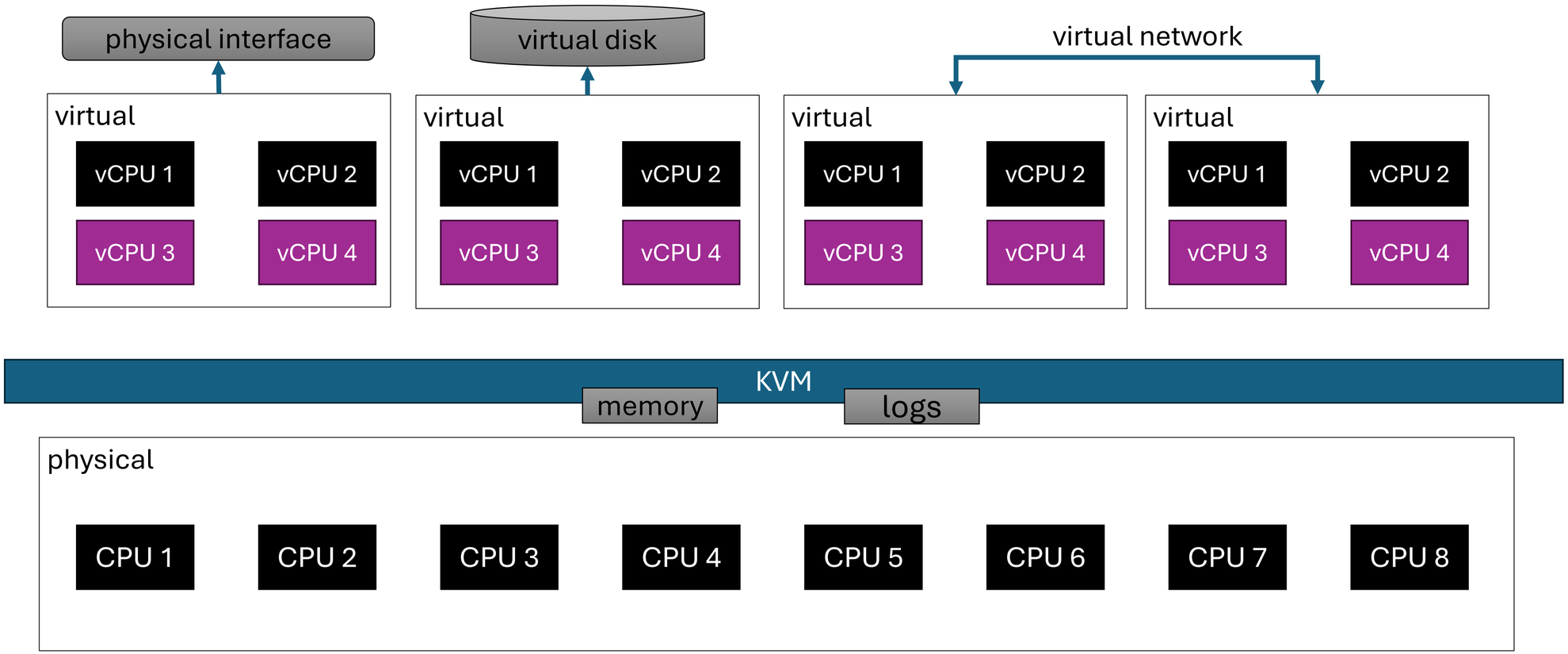


Complexity 3: Overcommitment

Complexity 3: Overcommitment

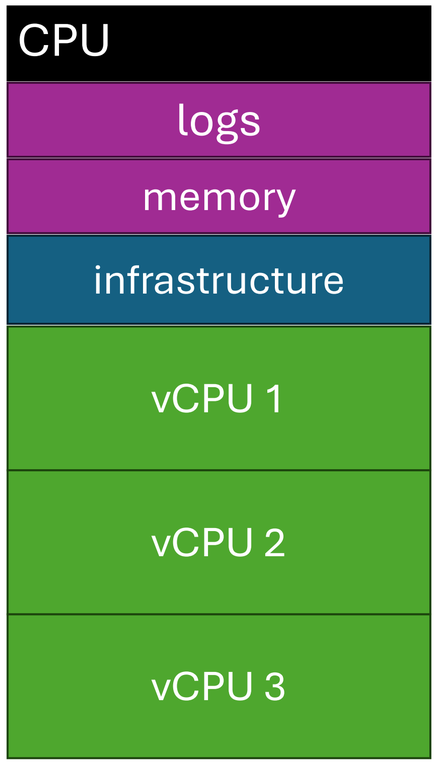


Complexity 3: Overcommitment



Complexity 3: Overcommitment

Utilization of physical CPUs



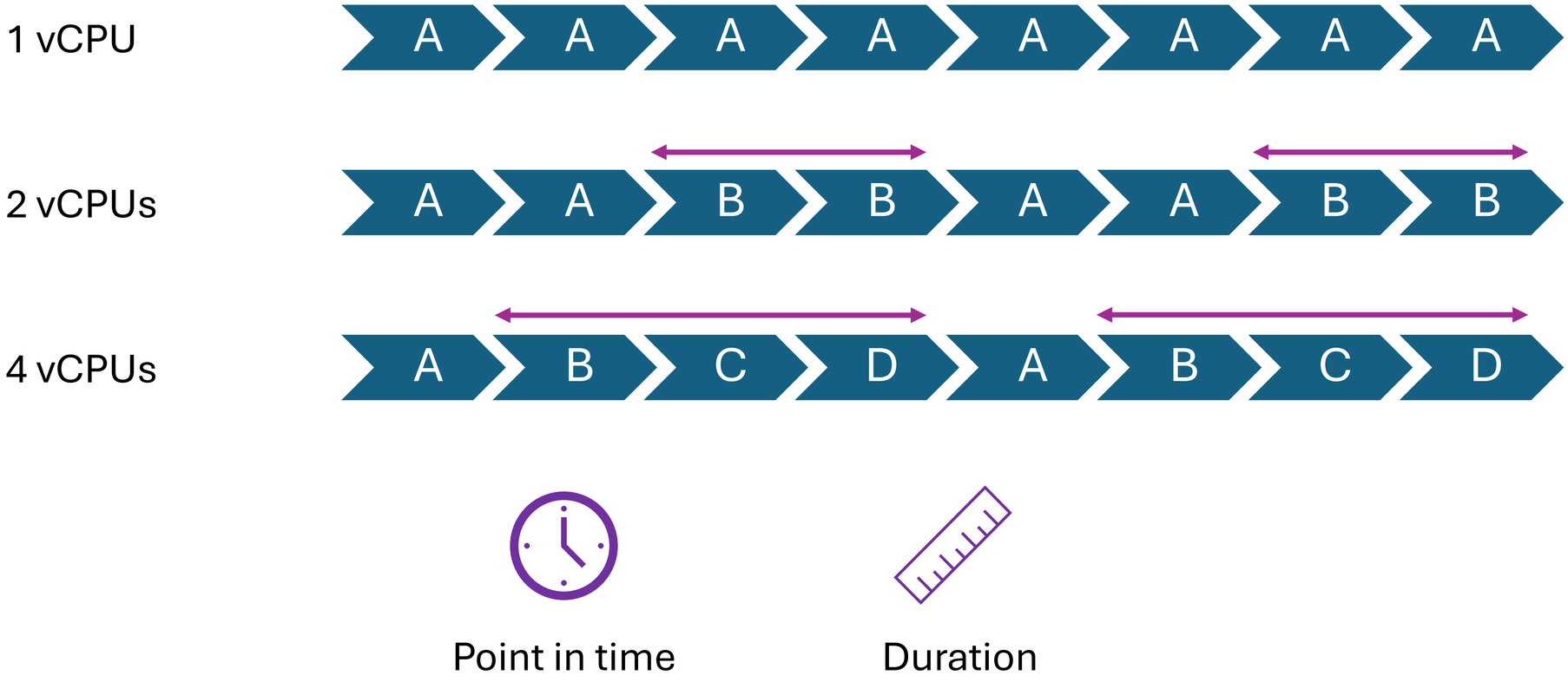
competition with other vCPUs

which vCPUs go well together?

→ vCPU experiences **steal time**

Complexity 3: Overcommitment

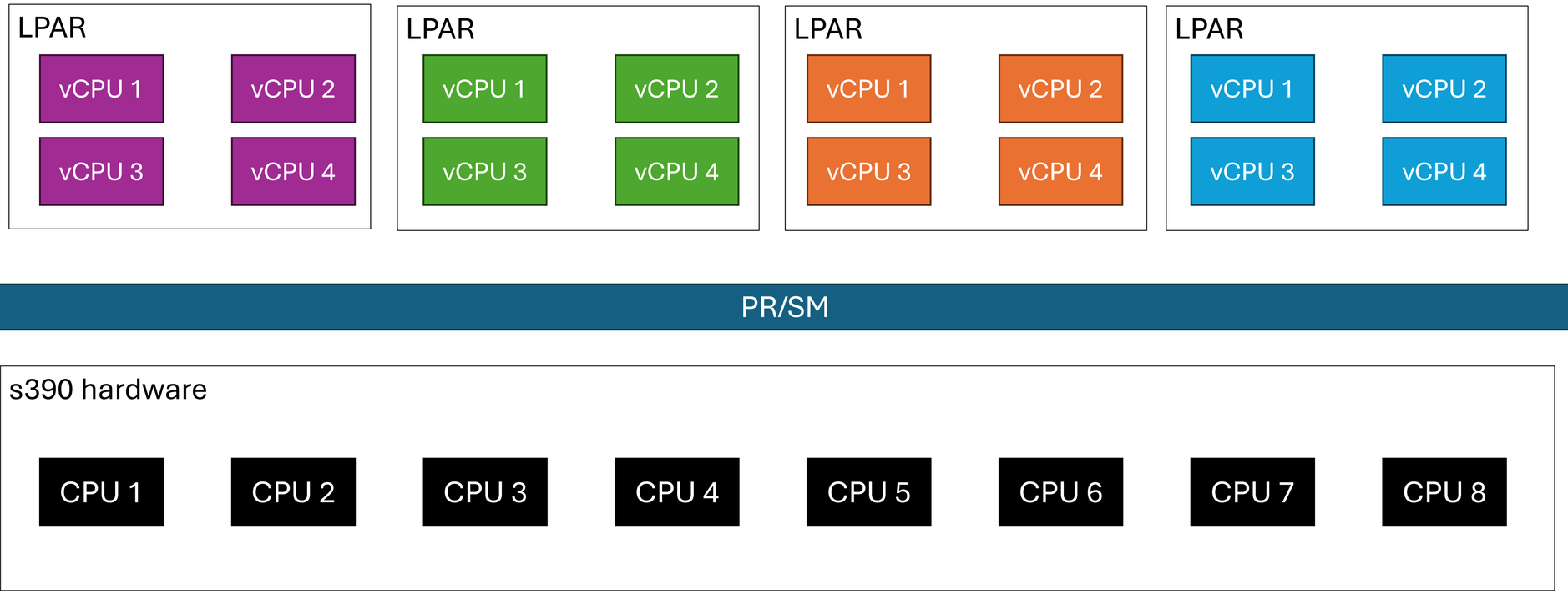
Issue: Interruption by other vcpus



The s390 approach

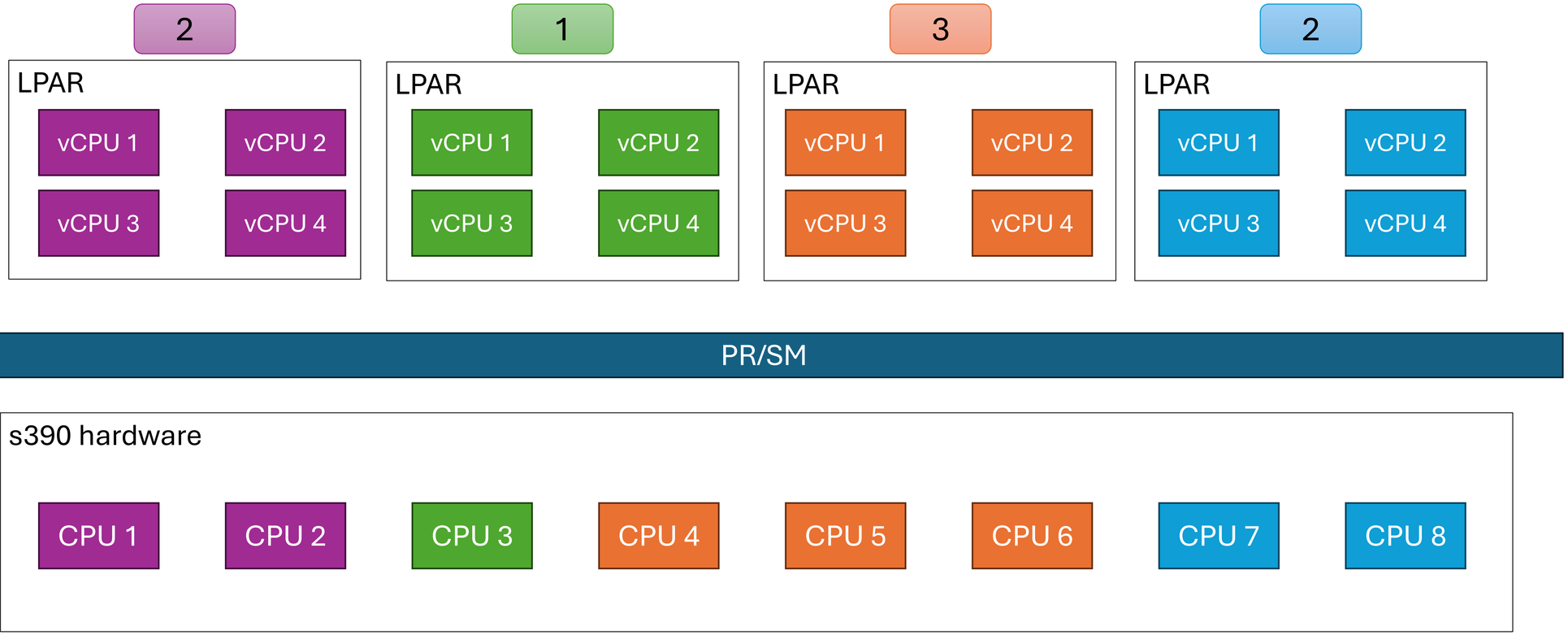
The s390 approach

Horizontal polarization: Distribute equally



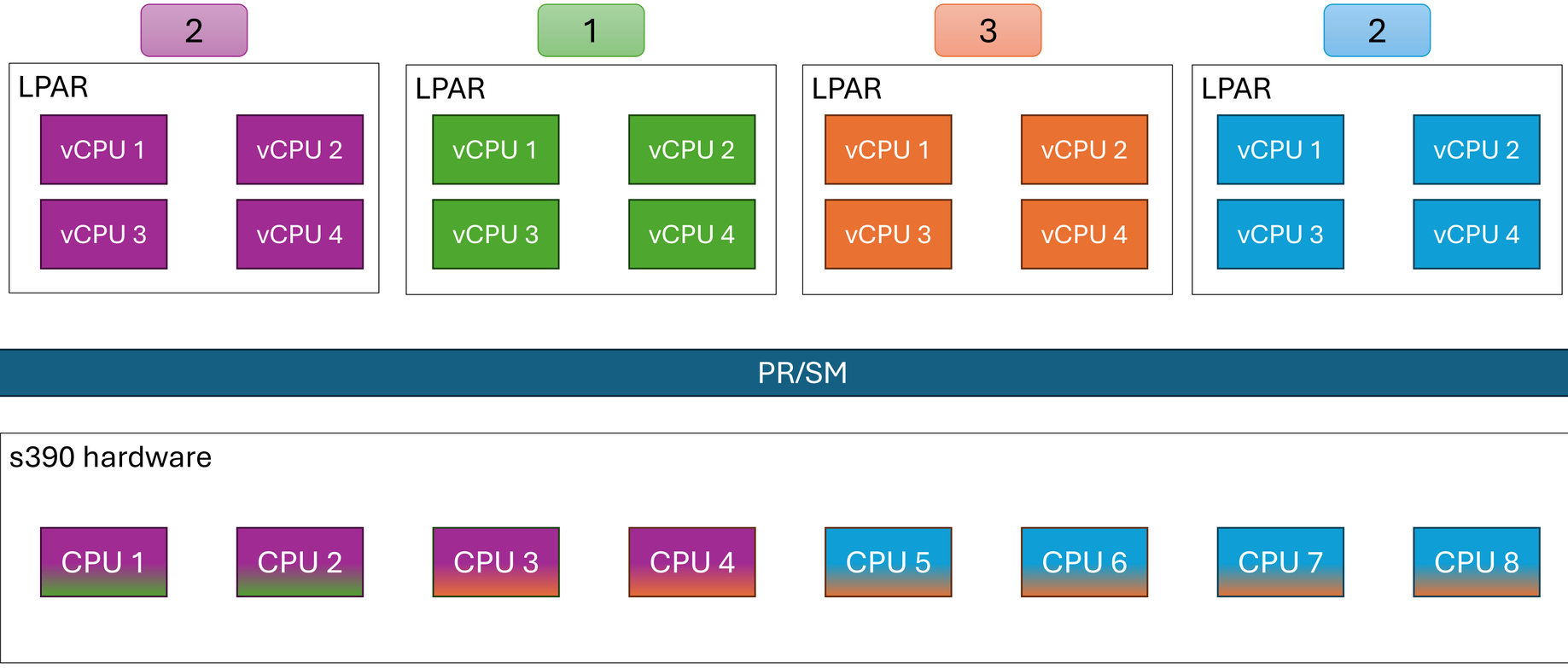
The s390 approach

Horizontal polarization: Distribute equally



The s390 approach

Horizontal polarization: Distribute equally

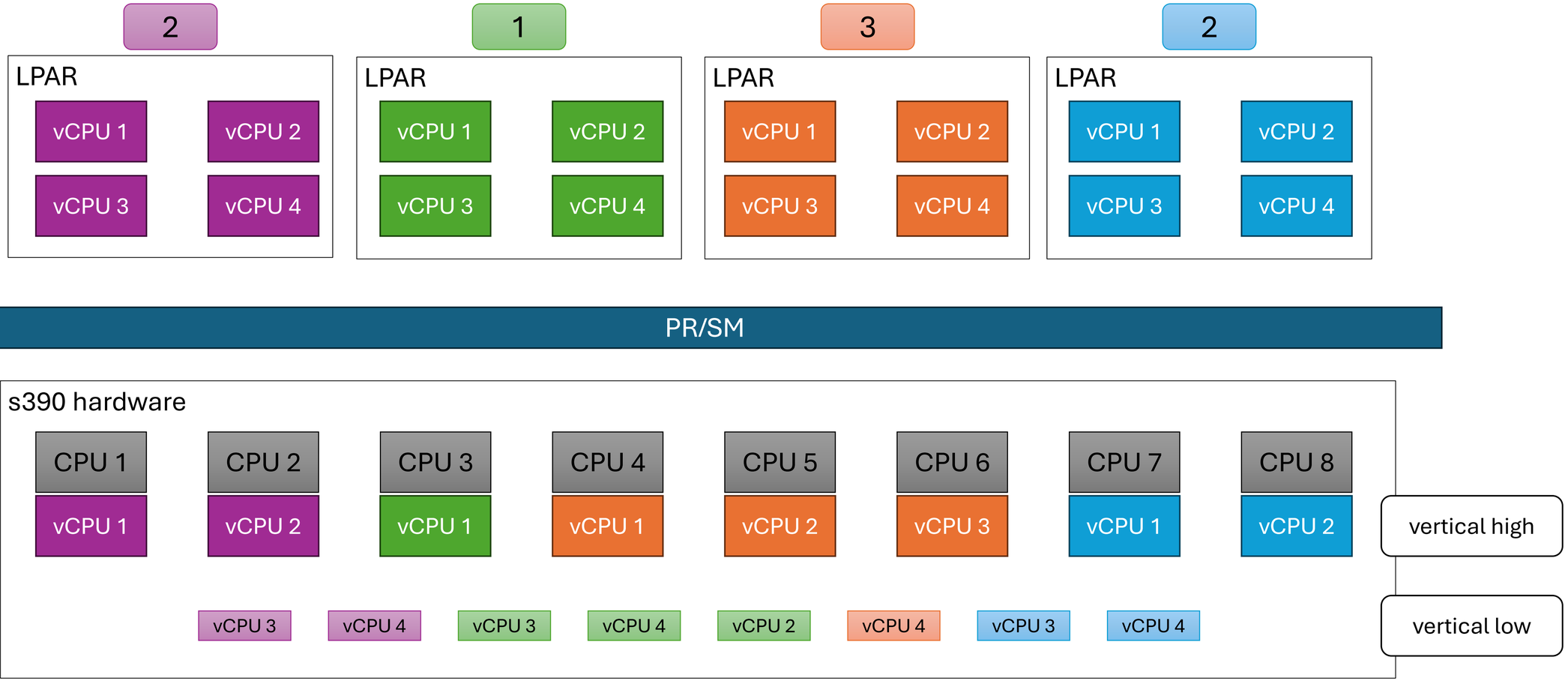


Timing

Cache topology

The s390 approach

Vertical polarization: Prioritize entitled CPUs



The s390 approach

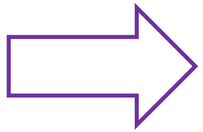
Vertical polarization: Advantages

Clearer scheduling of vCPUs

- Avoid steal time
- Better topology guarantees, yielding better cache locality

Collaboration between PR/SM and LPARs

- Gather CPU utilization of other LPARs
- Observe local steal time



Better control for the guest systems

The s390 approach

Integration into the Linux kernel

Option A: CPU capacity approach

Assign CPU capacities based on the polarization of the CPU

- vertical high → maximum capacity
- vertical low → small capacity (or maximum capacity if overconsumption is possible)

less invasive, changes to `arch/` only

not as strict, vertical lows may still run tasks

The s390 approach

Integration into the Linux kernel

Option B: Load balancer, scheduler group types approach

CPU 1 | CPU 2 | CPU 3 | CPU 4 | CPU 5 | CPU 6 | CPU 7 | CPU 8

Add a new scheduler group type beyond `group_overloaded`

- vertical high → regular scheduling
- vertical low → get assigned to new group type if overconsumption is not possible, causes the load balancer to pull all tasks from those CPUs and prevents those CPUs to pull tasks themselves

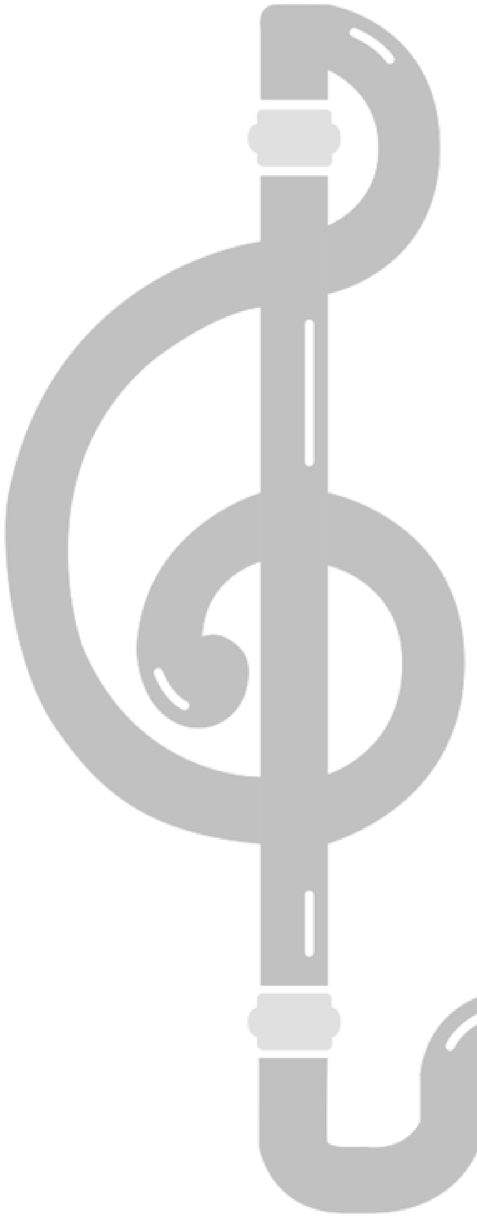
 CPUs can be prevented from running tasks

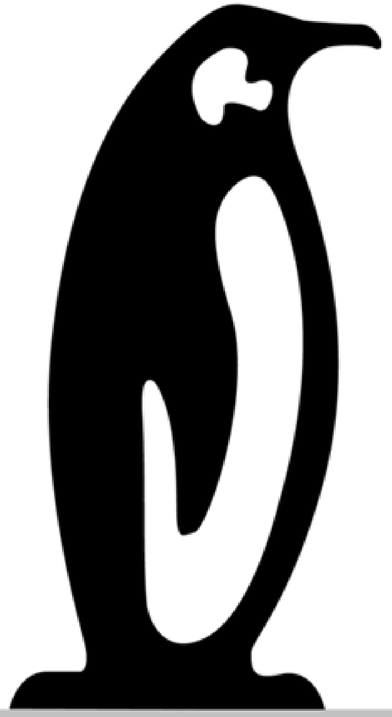
 changes to the common load balancer

Summary

1. Virtualized infrastructure
   - Awareness of ordering requirements
     - Transactions
     - Yield explicitly
2. Overcommitment
   - Prioritize entitled CPUs
     - Capacity approach
     - Load balancer, scheduler group types approach

Trademarks





Linux Plumbers Conference

Vienna, Austria | September 18-20, 2024