

Linux Plumbers Conference

Vienna, Austria | September 18-20, 2024



Challenges in scheduling virtual CPUs

Tobias Huschle <huschle@linux.ibm.com>

IBM

Agenda

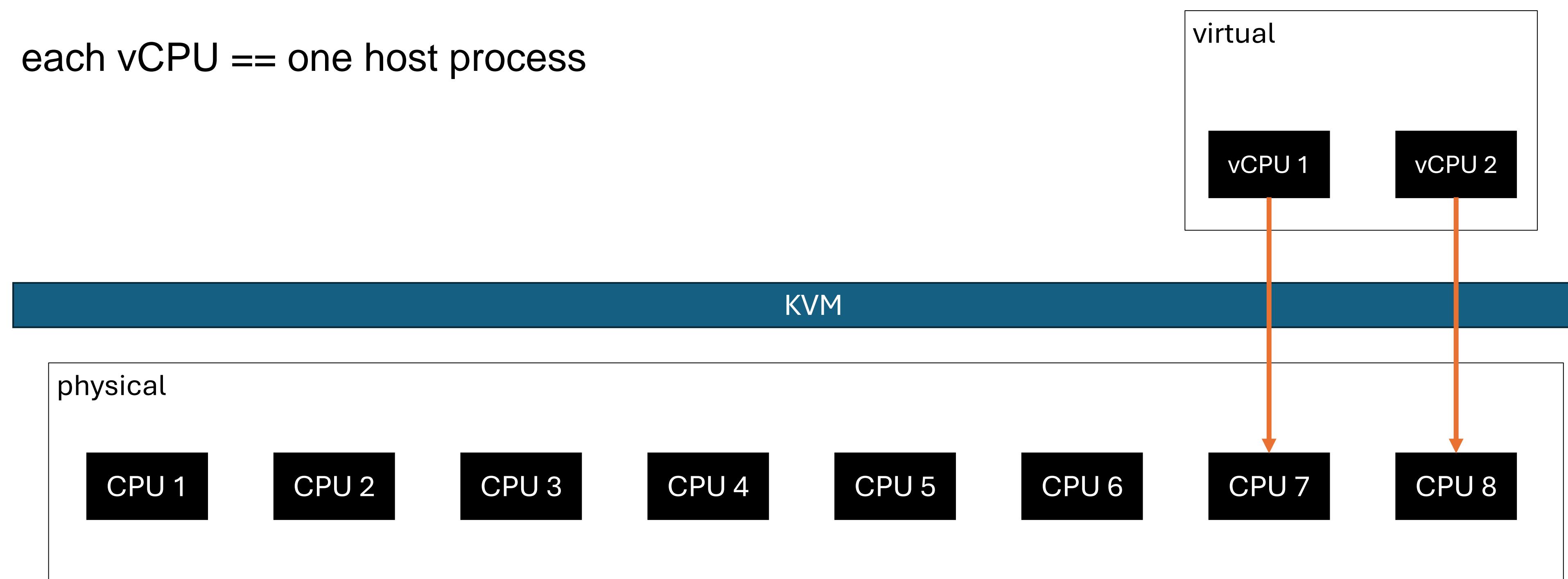
- Warm-up: Basics of virtual CPU scheduling
- Complexities
 - 1. Host overhead
 - 2. Virtualized infrastructure
 - 3. Overcommitment
- The s390 approach



Warm up

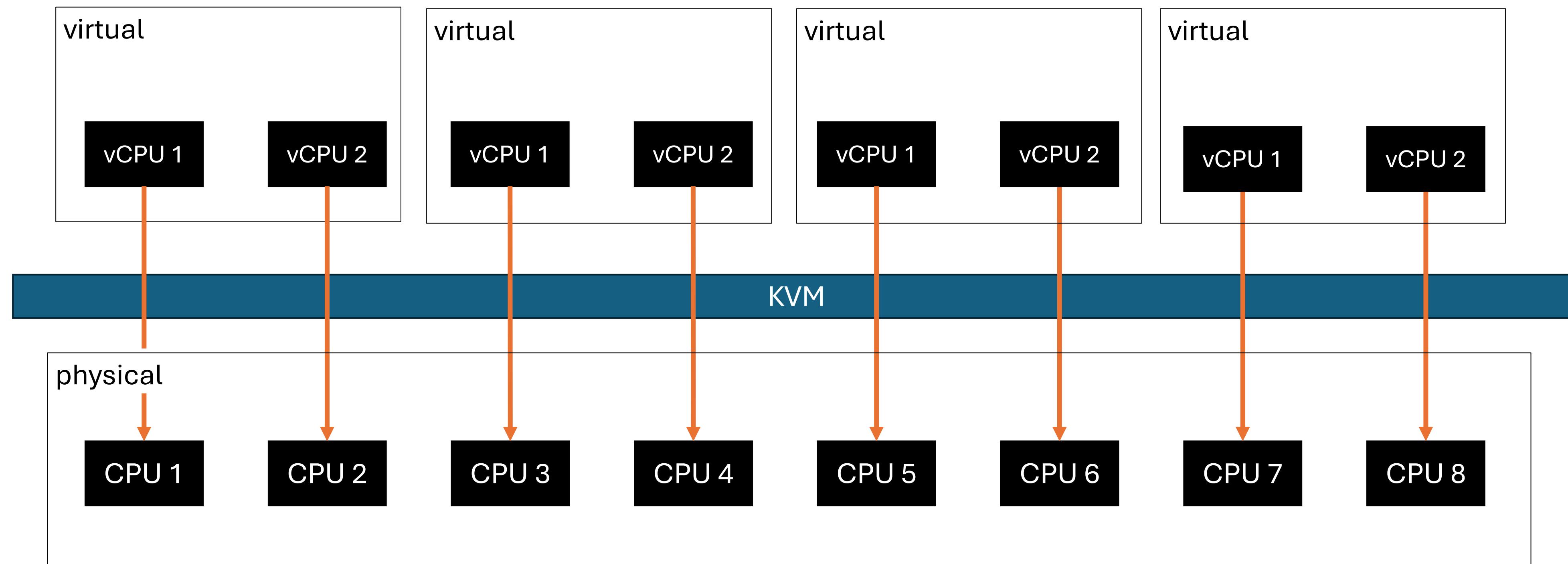
Basics of virtual CPU scheduling

each vCPU == one host process

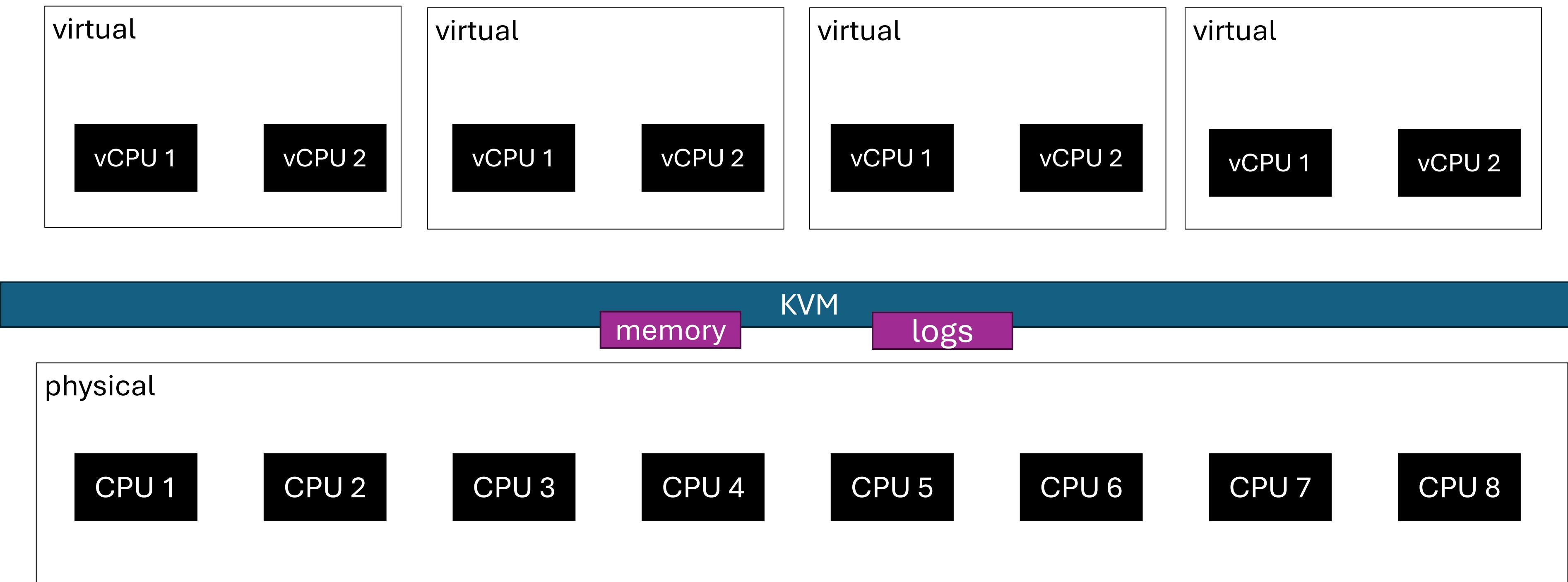


Warm up

Basics of virtual CPU scheduling

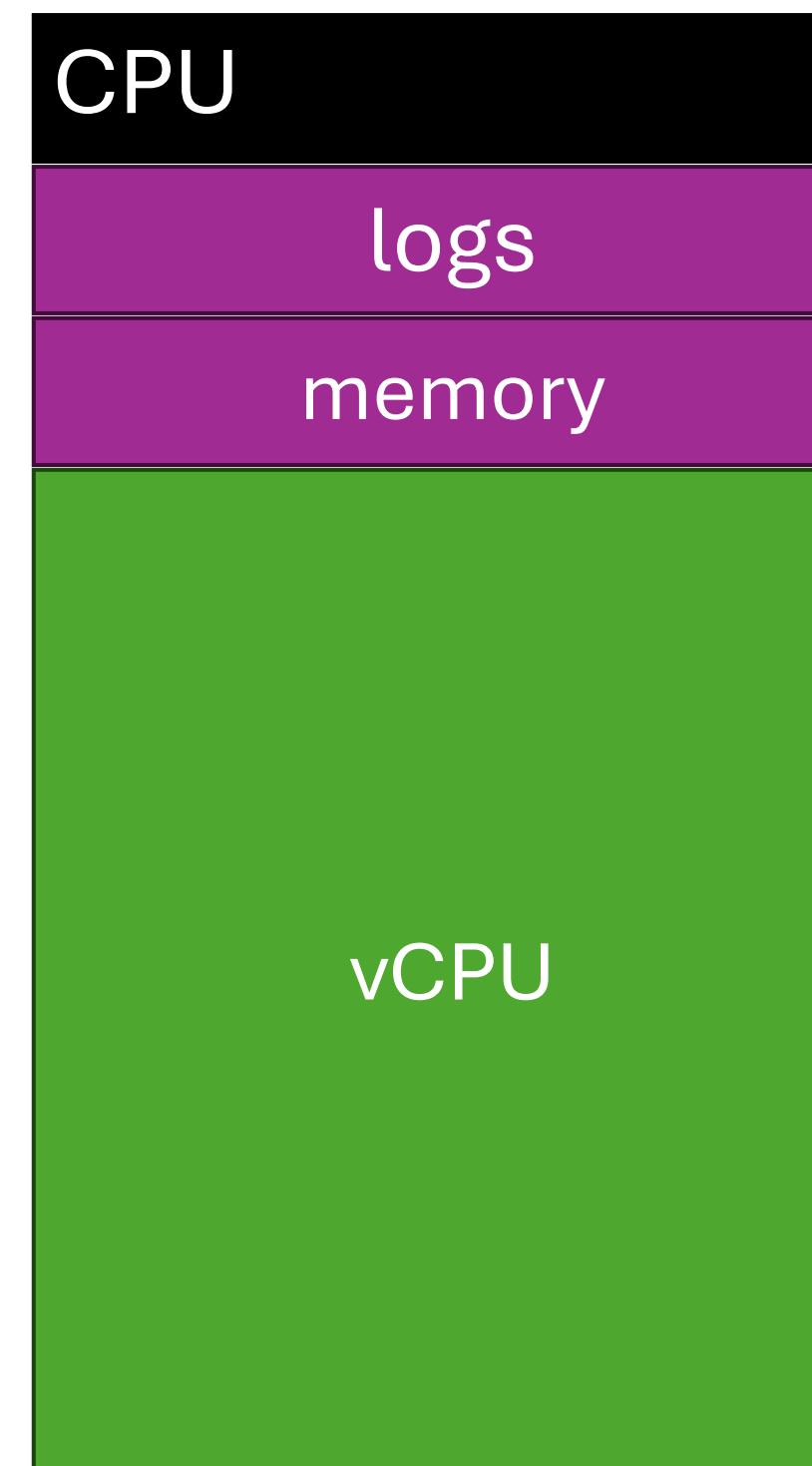


Complexity 1: Host overhead



Complexity 1: Host overhead

Utilization of physical CPUs



vCPU cannot use 100% of physical CPU

Host has to decide when to schedule vCPU away



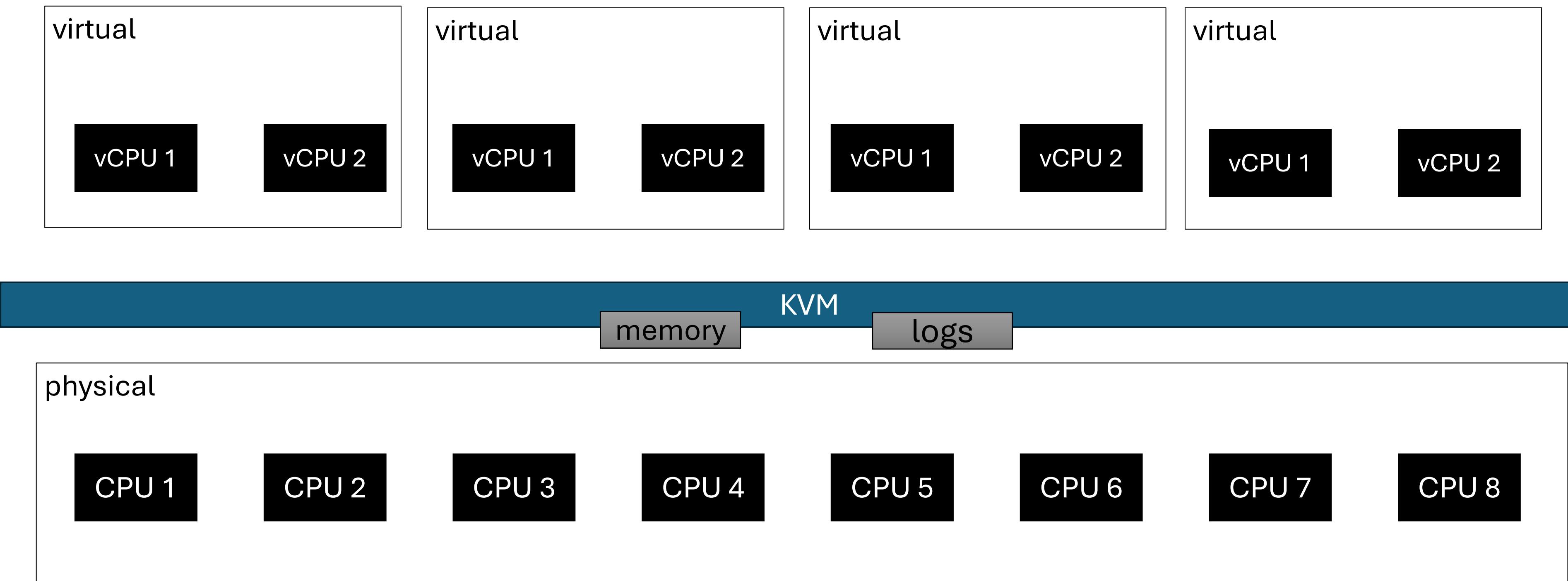
Complexity 2: Virtualized infrastructure



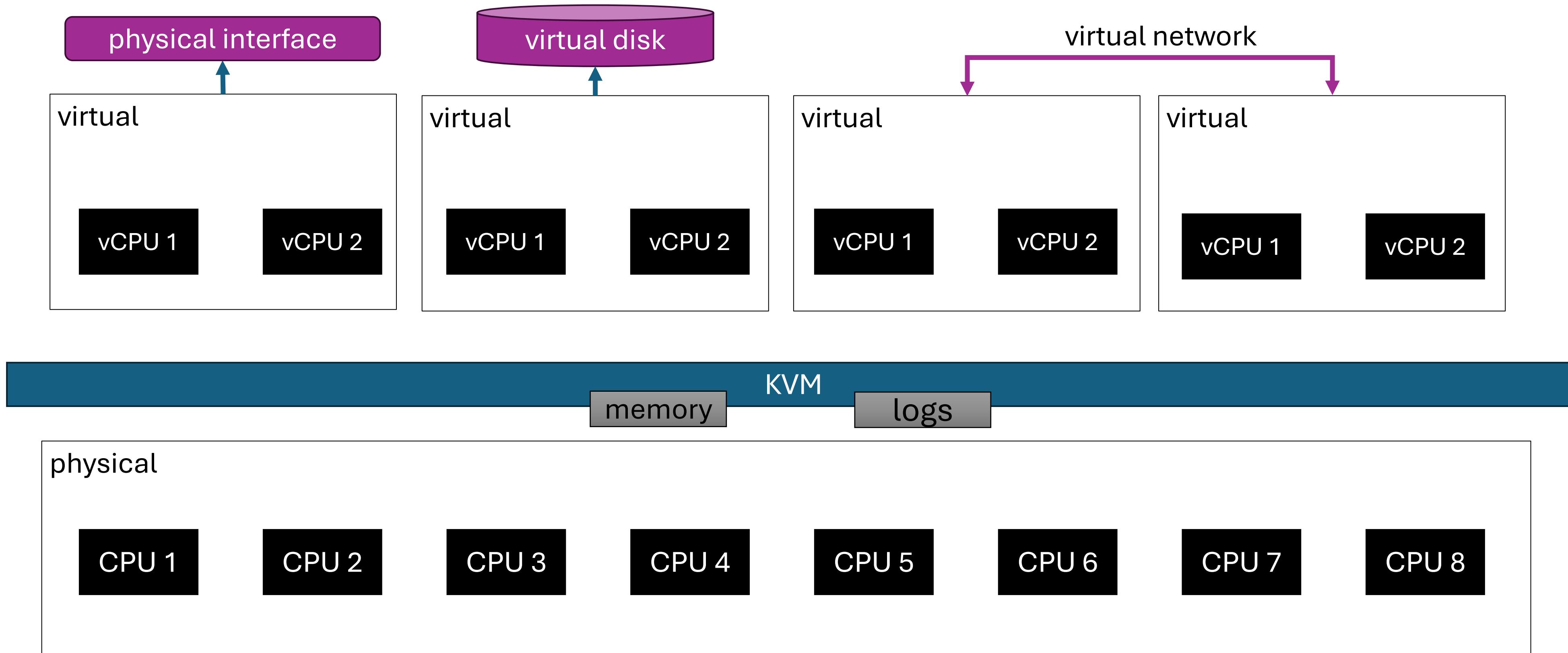
LINUX PLUMBERS CONFERENCE

Vienna, Austria
Sept. 18-20, 2024

Complexity 2: Virtualized infrastructure

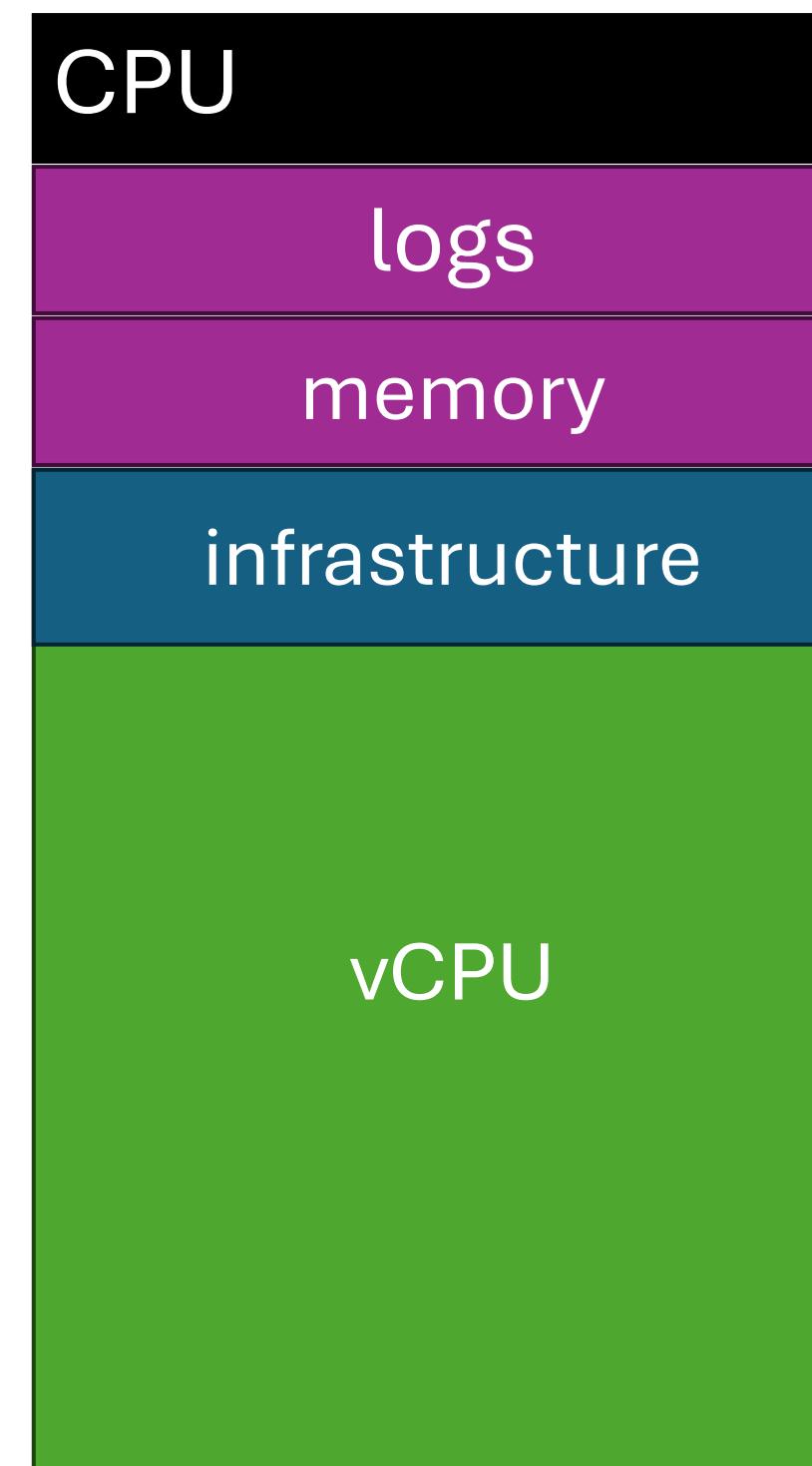


Complexity 2: Virtualized infrastructure



Complexity 2: Virtualized infrastructure

Utilization of physical CPUs

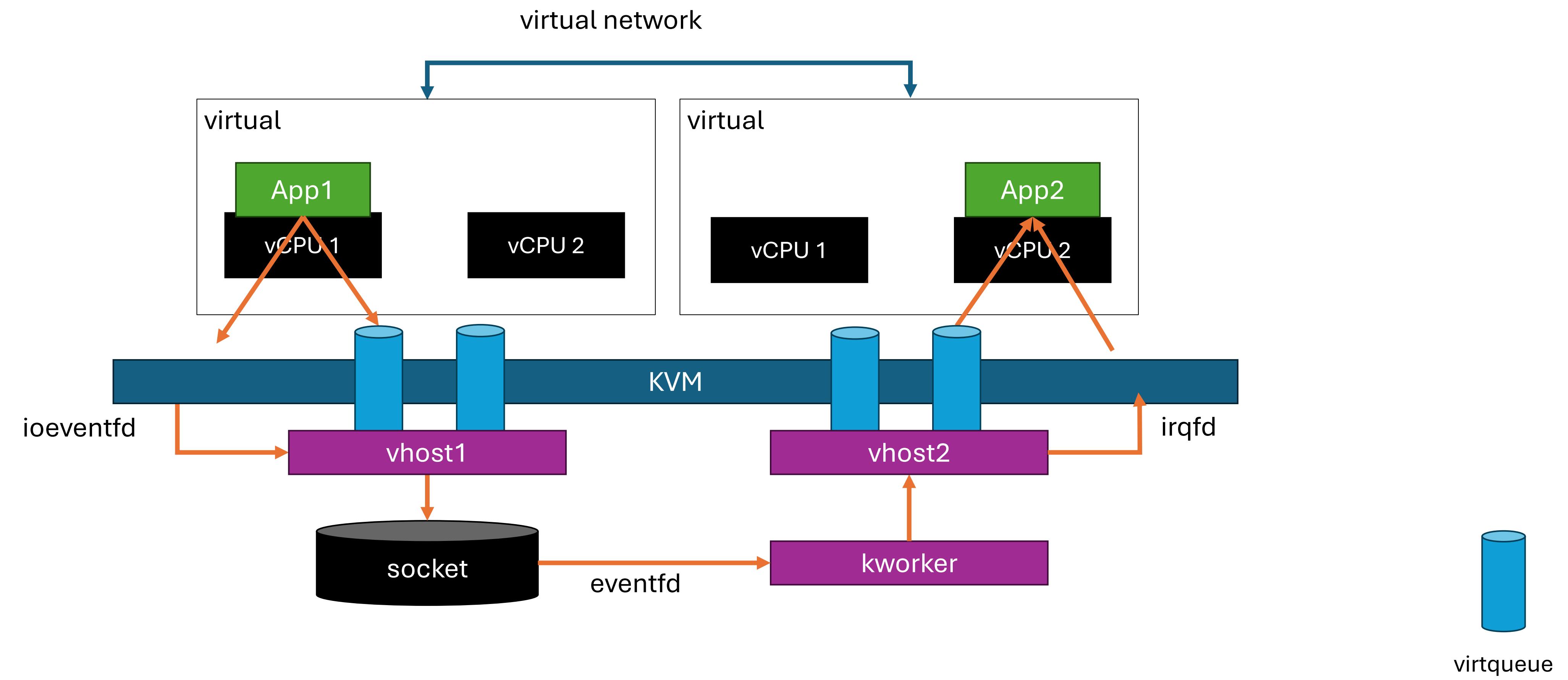


Infrastructure needs to be mapped to physical CPU

Increased contention for physical CPUs

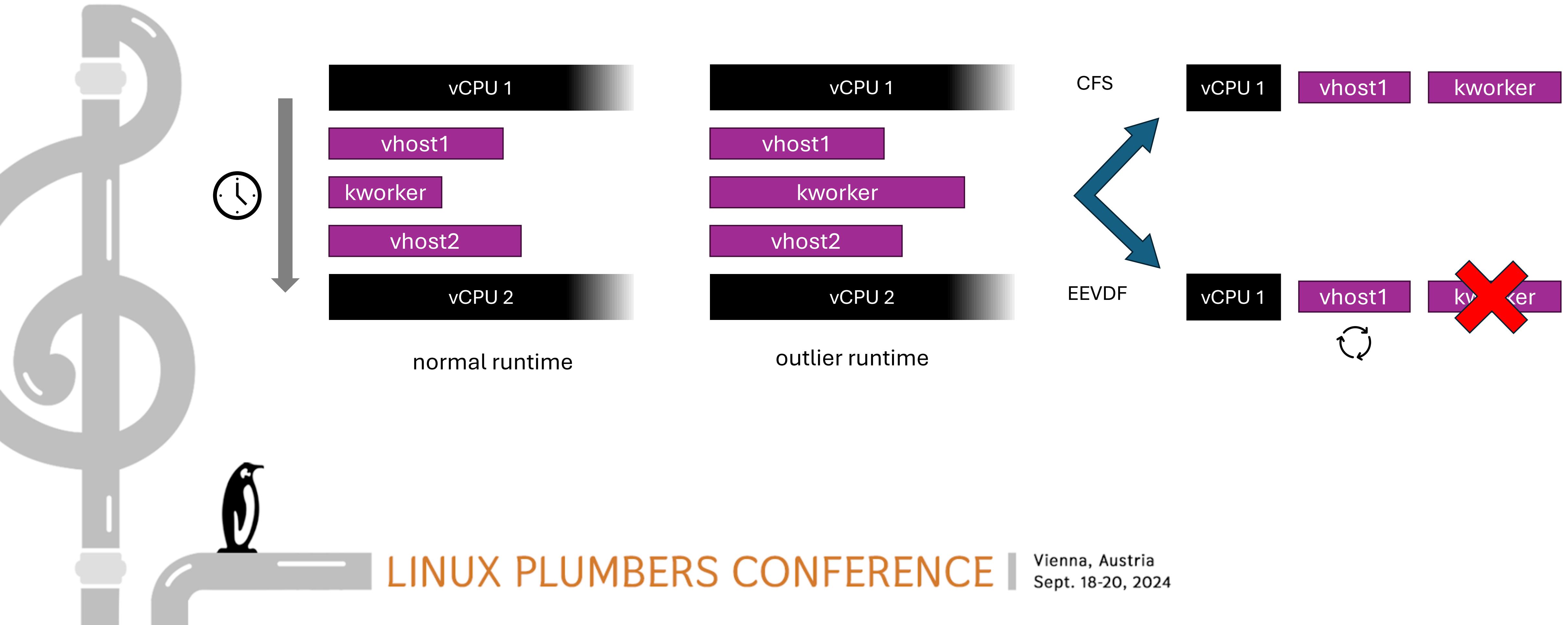
Complexity 2: Virtualized infrastructure

Example: vhost



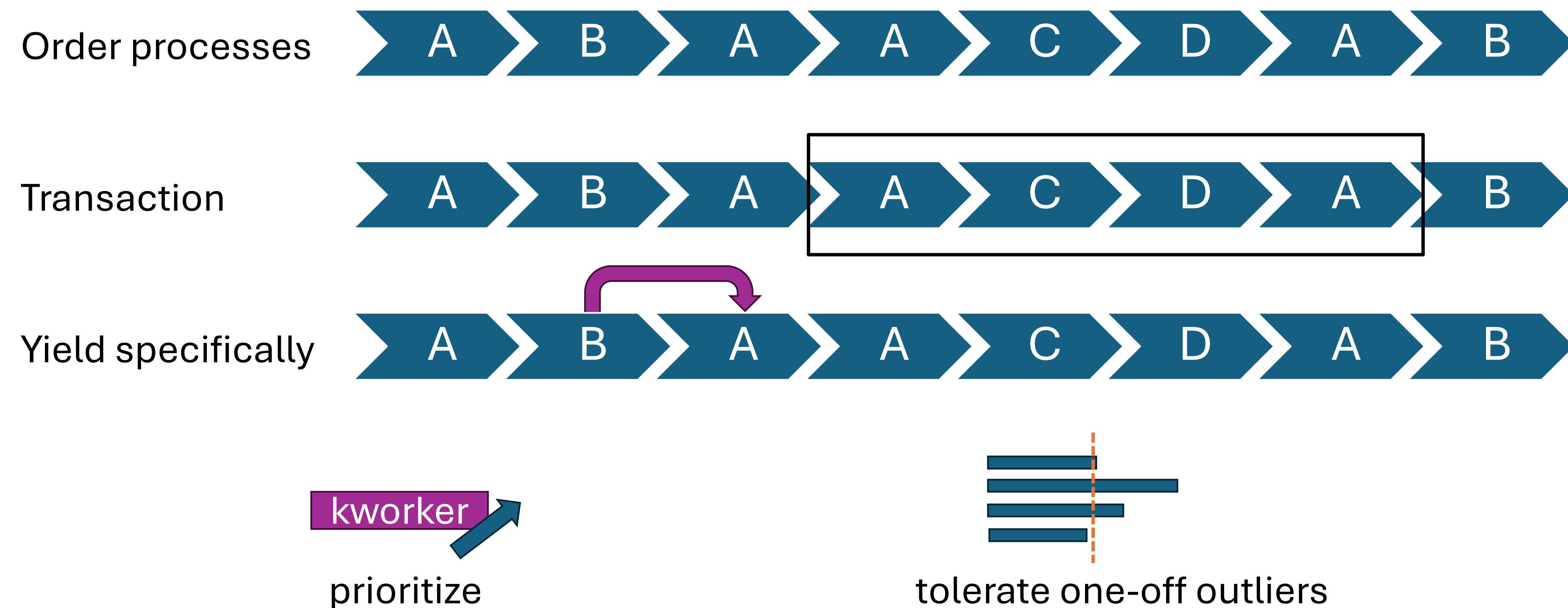
Complexity 2: Virtualized infrastructure

Scheduling impact



Complexity 2: Virtualized infrastructure

Ordering: possible solutions



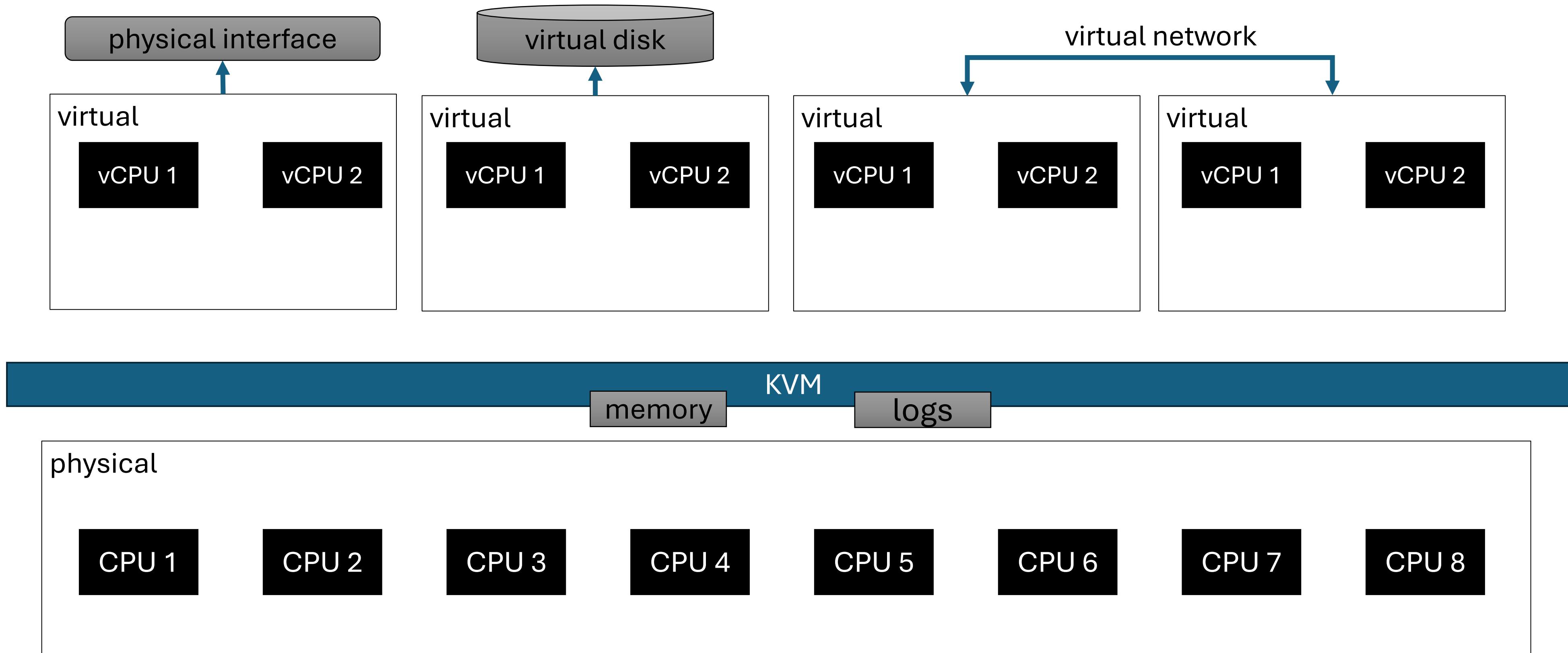
Complexity 3: Overcommitment



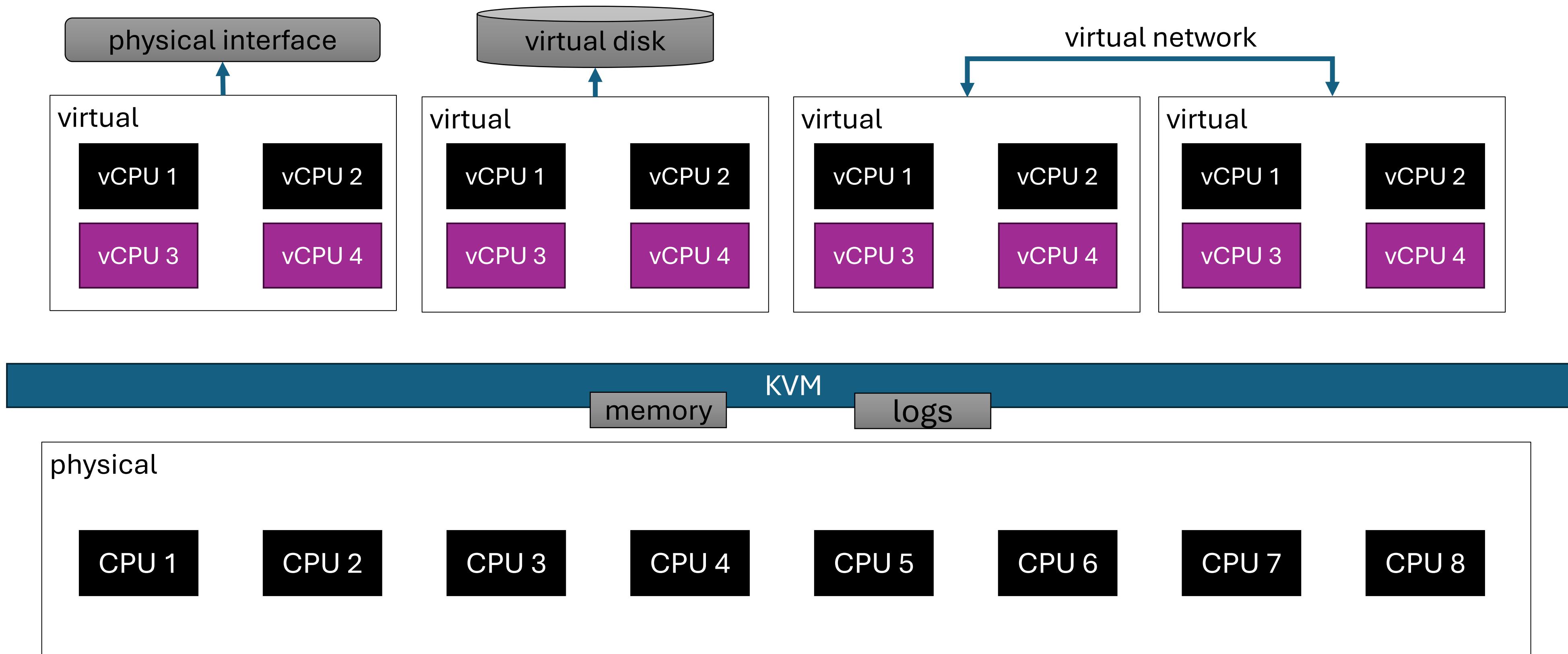
LINUX PLUMBERS CONFERENCE

Vienna, Austria
Sept. 18-20, 2024

Complexity 3: Overcommitment

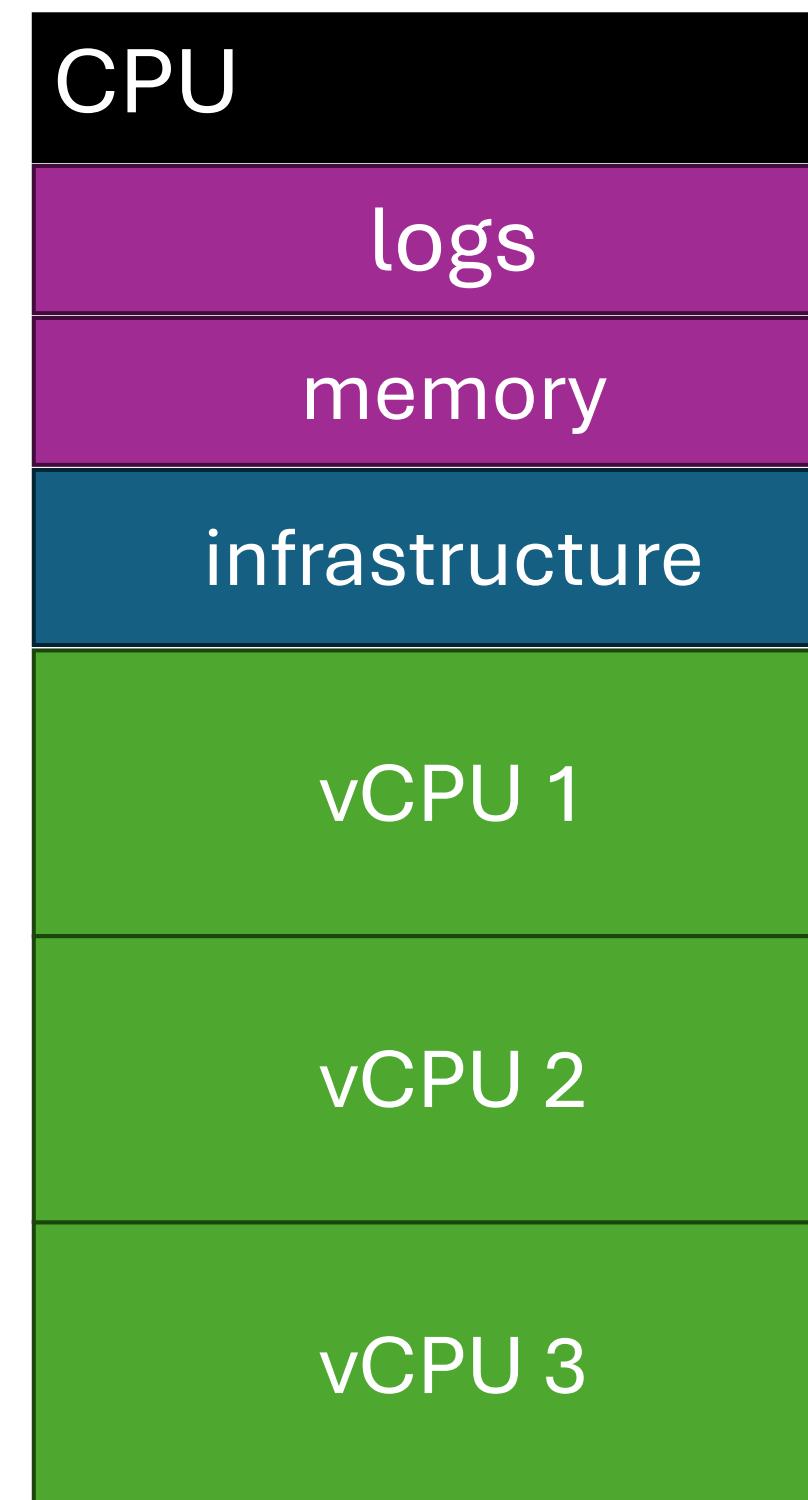


Complexity 3: Overcommitment



Complexity 3: Overcommitment

Utilization of physical CPUs



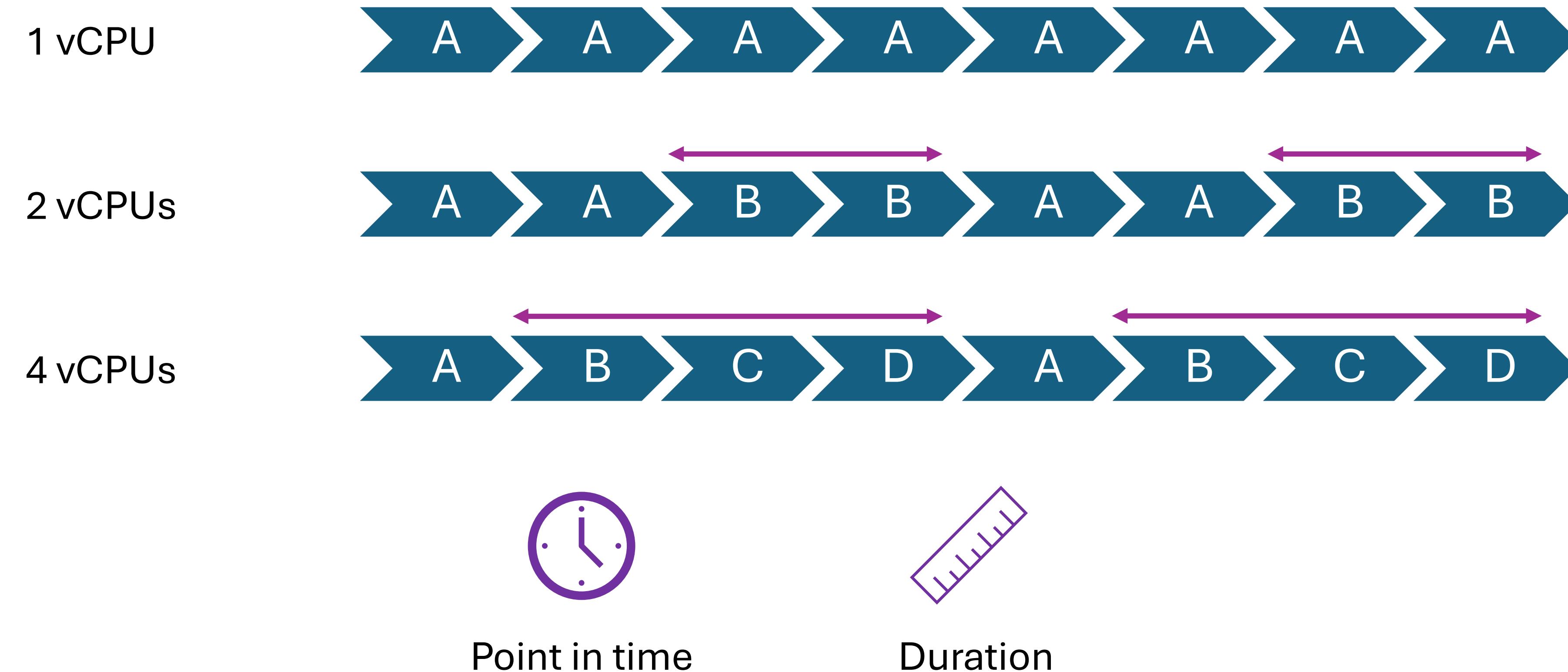
competition with other vCPUs

which vCPUs go well together?

→ vCPU experiences steal time

Complexity 3: Overcommitment

Issue: Interruption by other vcpus



The s390 approach

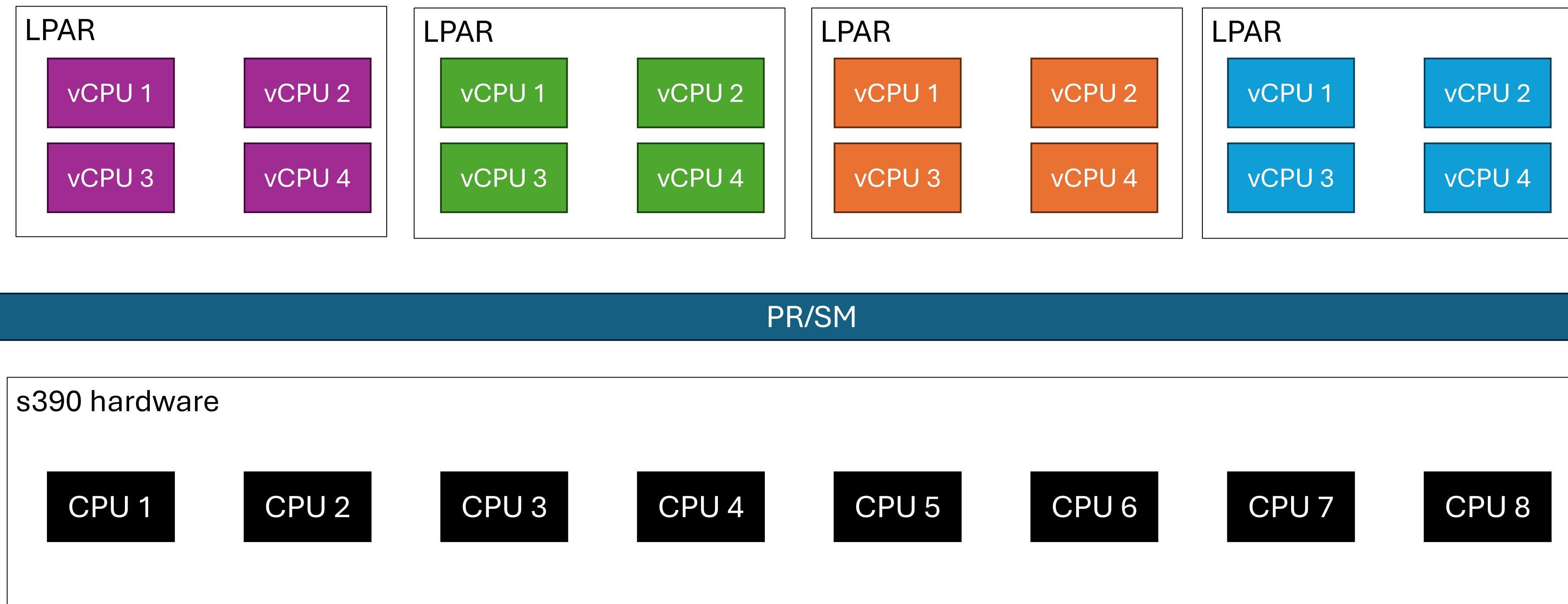


LINUX PLUMBERS CONFERENCE

Vienna, Austria
Sept. 18-20, 2024

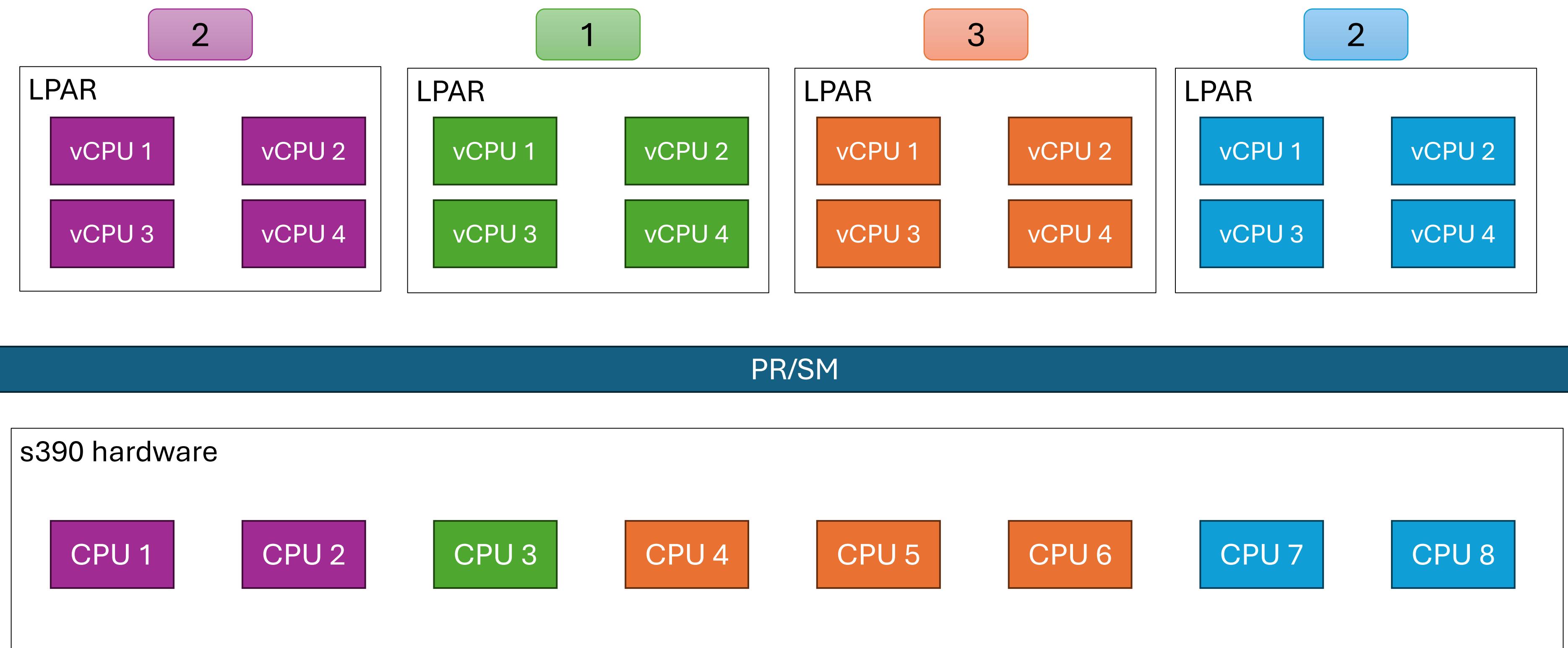
The s390 approach

Horizontal polarization: Distribute equally



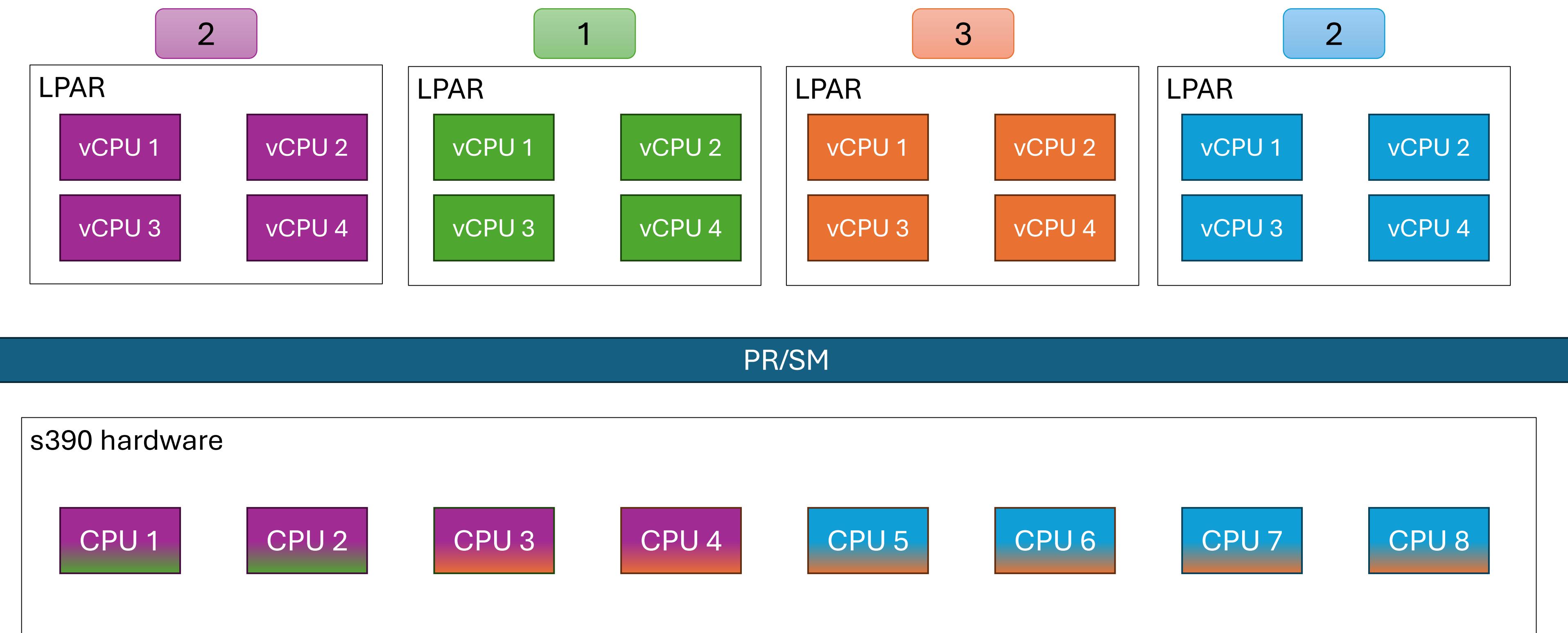
The s390 approach

Horizontal polarization: Distribute equally



The s390 approach

Horizontal polarization: Distribute equally



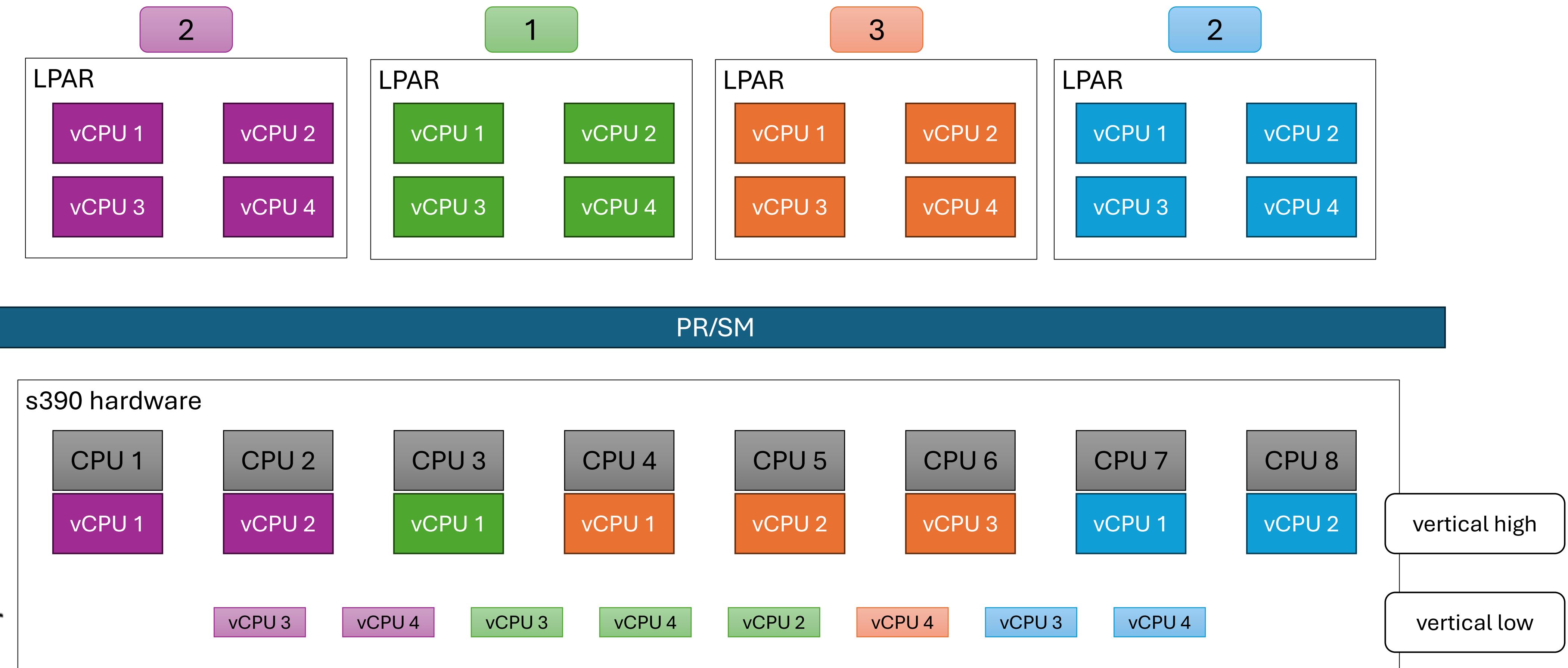
Timing



Cache topology

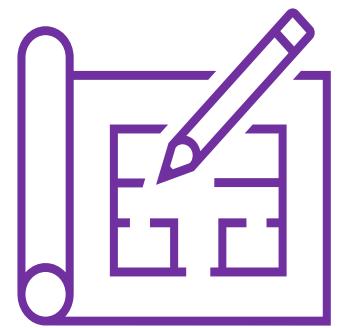
The s390 approach

Vertical polarization: Prioritize entitled CPUs



The s390 approach

Vertical polarization: Advantages



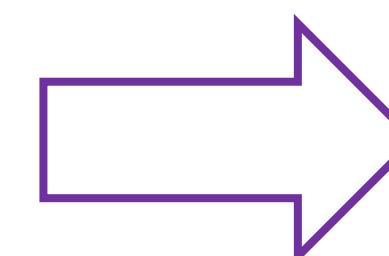
Clearer scheduling of vCPUs

- 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



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

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

CICS®	Parallel Sysplex®	z/Architecture®
Concert®	RACF®	z/OS®
Db2®	Rational®	z/VM®
FICON®	Redbooks®	z/VSE®
HyperSwap®	Redbooks (logo)  ®	z13®
IBM®	Resource Link®	z15™
IBM Z®	S/390®	z16™
IBM z13®	System z®	zEnterprise®
IBM z14®	VTAM®	zPDT®
IBM z16™	WebSphere®	

The following terms are trademarks of other companies:

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Red Hat and Fedora are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

VMware, and the VMware logo are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

Other company, product, or service names may be trademarks or service marks of others.





Linux Plumbers Conference

Vienna, Austria | September 18-20, 2024