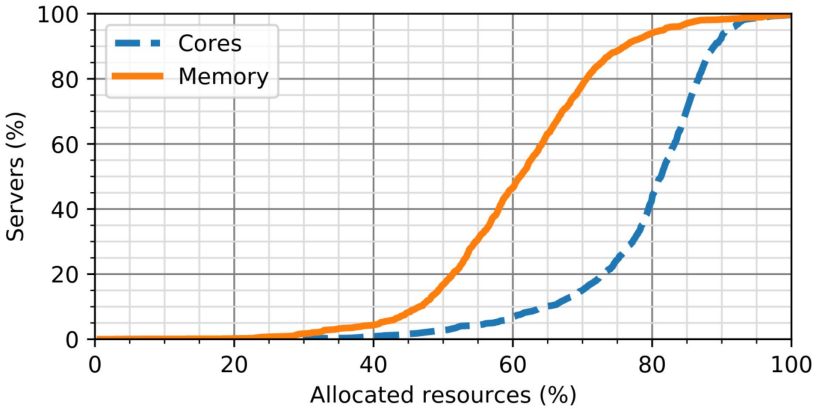
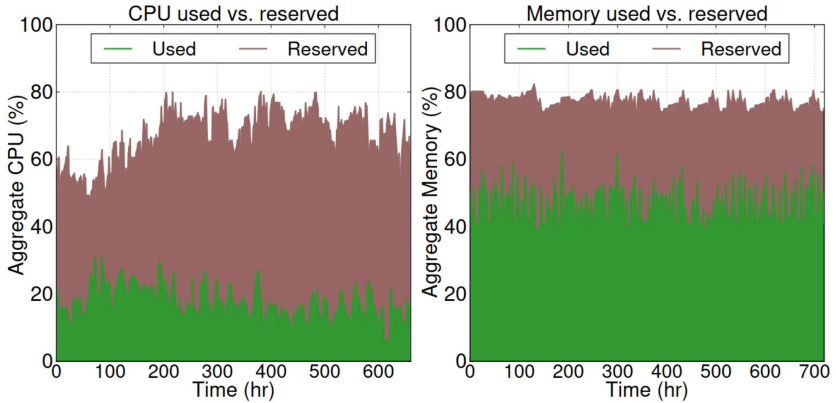
State-of-the-Art of Cloud Resource Utilization

Cloud infrastructure are under-used



Not all resources are **allocated** [1]

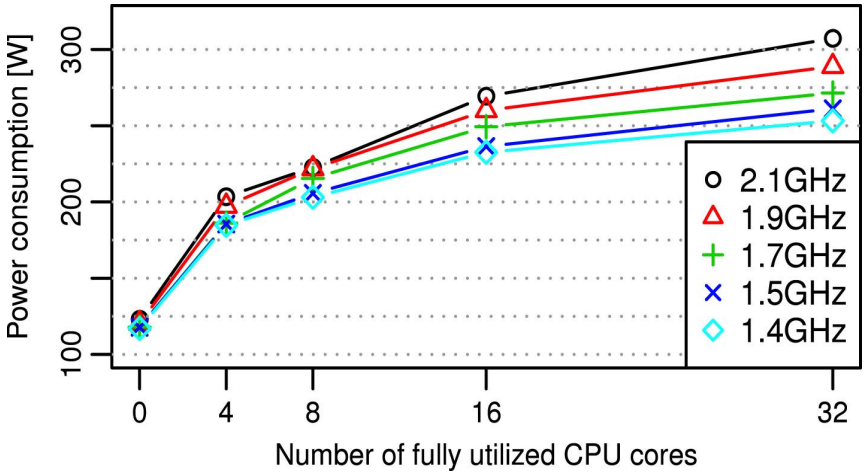


Not all allocated resources are **used** [2]

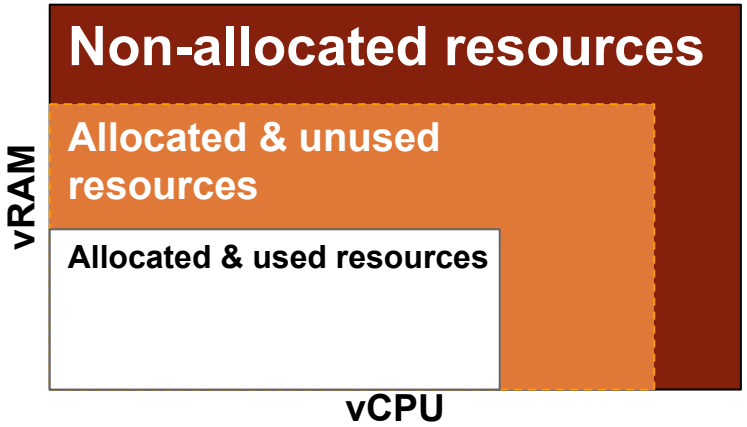
[1] *Memory-harvesting VMs in cloud platforms*, Microsoft Research, ASPLOS 2022
[2] *Quasar: Resource-Efficient and QoS-Aware Cluster Management,* Christina Delimitrou *et al.*, ASPLOS 2014

State-of-the-Art of Cloud Resource Utilization

Cloud infrastructure are under-used



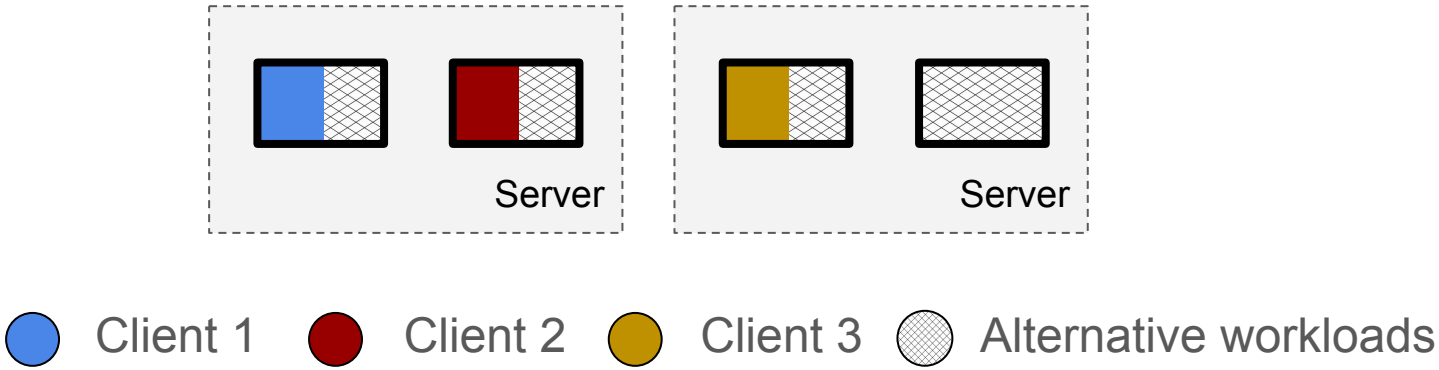
Impacts **server consumption** [1]



Impacts the number of **provisioned servers**

[1] *Power-performance tradeoffs in data center servers, Jakub Krzywda et al., FGCS 2018*

State-of-the-Art of Cloud Resource Utilization



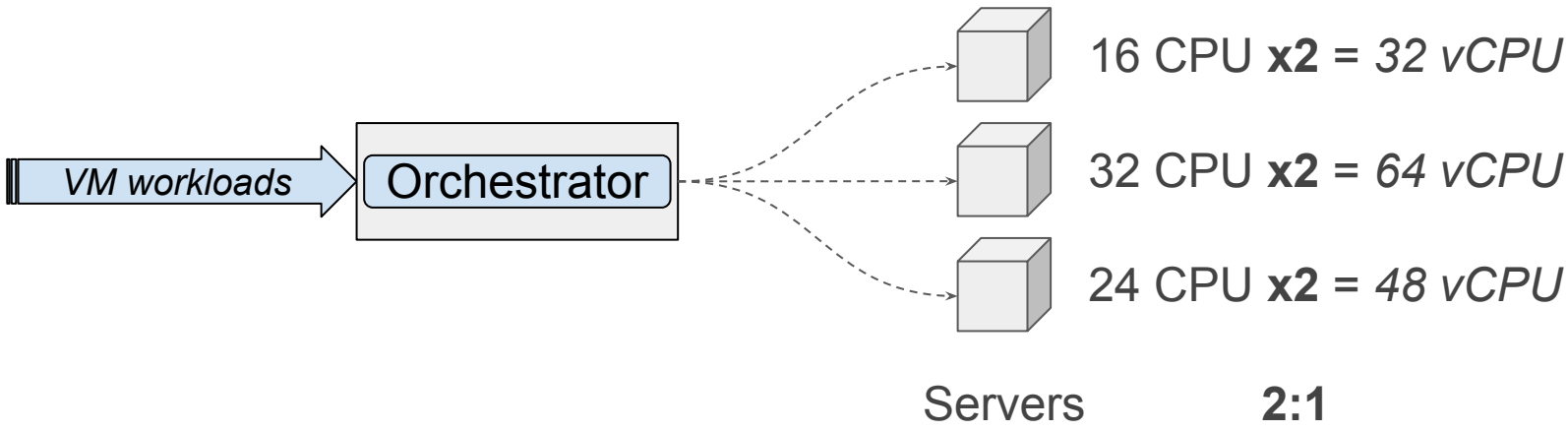
- A problem addressed using different (complementary) approaches

  1) Fill with **heterogeneous workloads,** *e.g. Batch, FaaS, HarvestVM*
  2) Pack with **homogeneous workloads** by sharing cores

State-of-the-Art of Cloud Resource Utilization

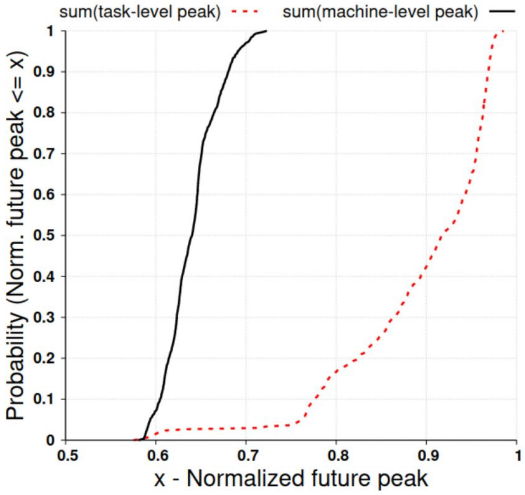
How to define the right amount of clients when sharing resources?

- Ratio commonly set **statically** at the cluster level



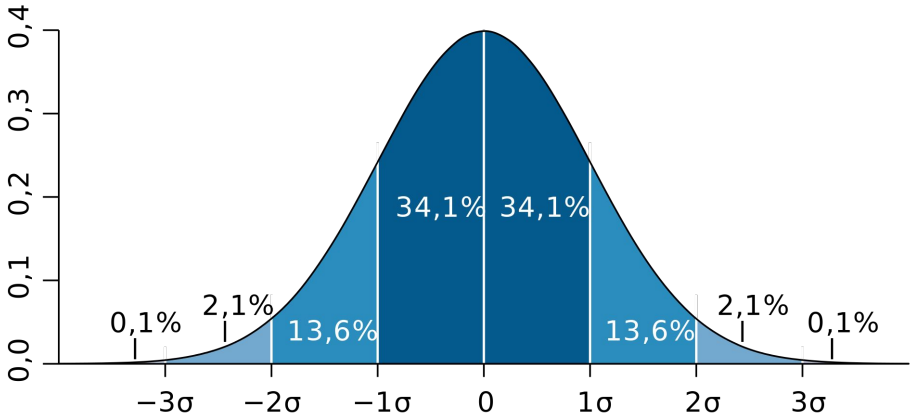
Contribution 1

- **Dynamic** oversubscription
  - Relies on pessimistic prediction of **next peak usage**



Microsoft research:

Next peak: sum of VMs <u>percentile</u>

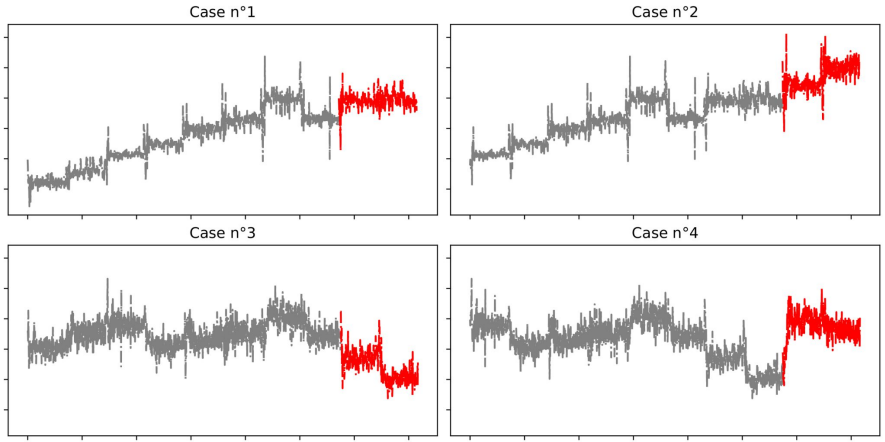


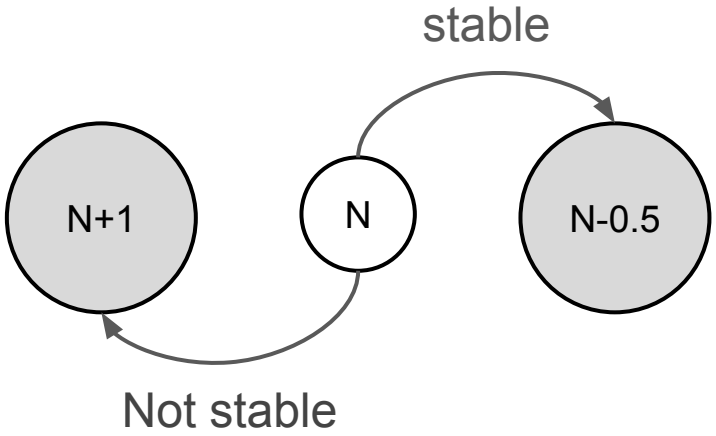
Google research:

Next peak: use server <u>std deviation</u>

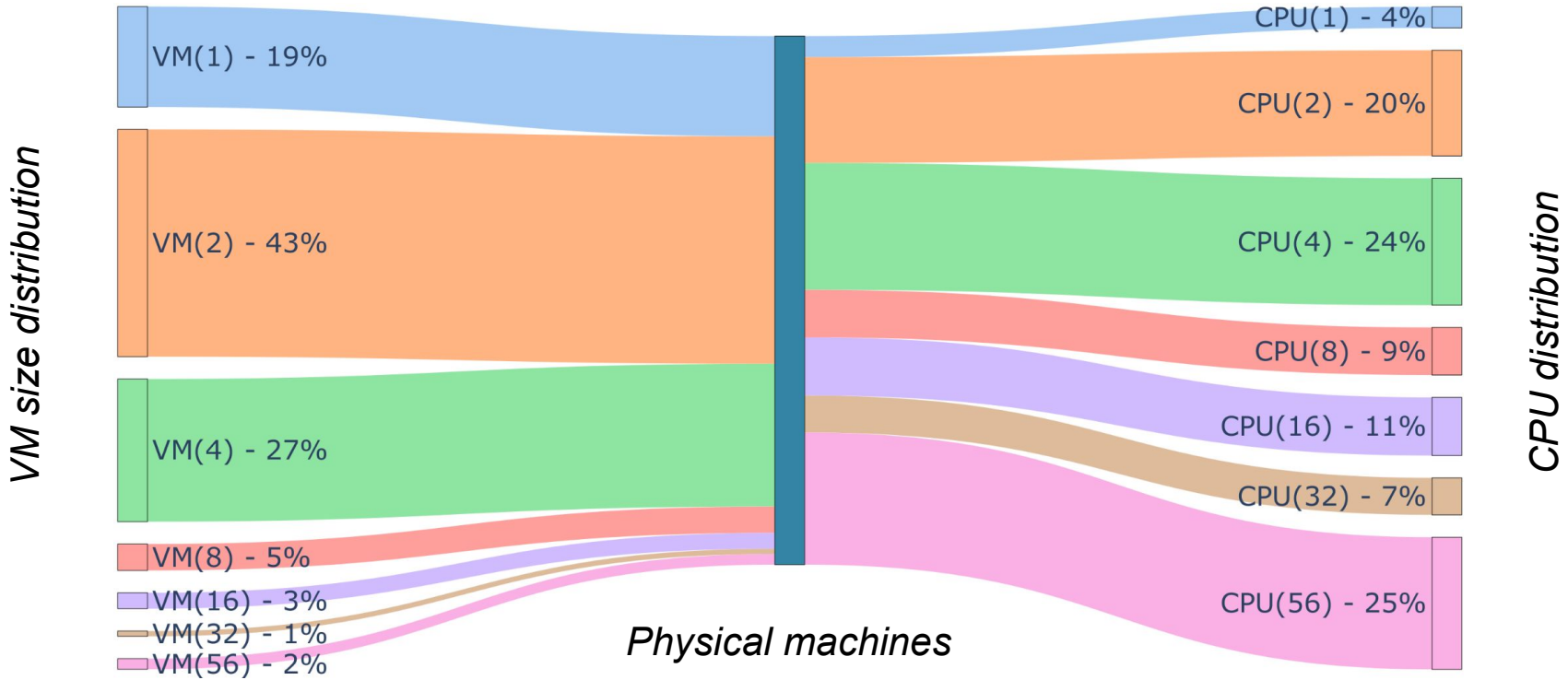
Contribution 1

Oversubscription based on quiescent state



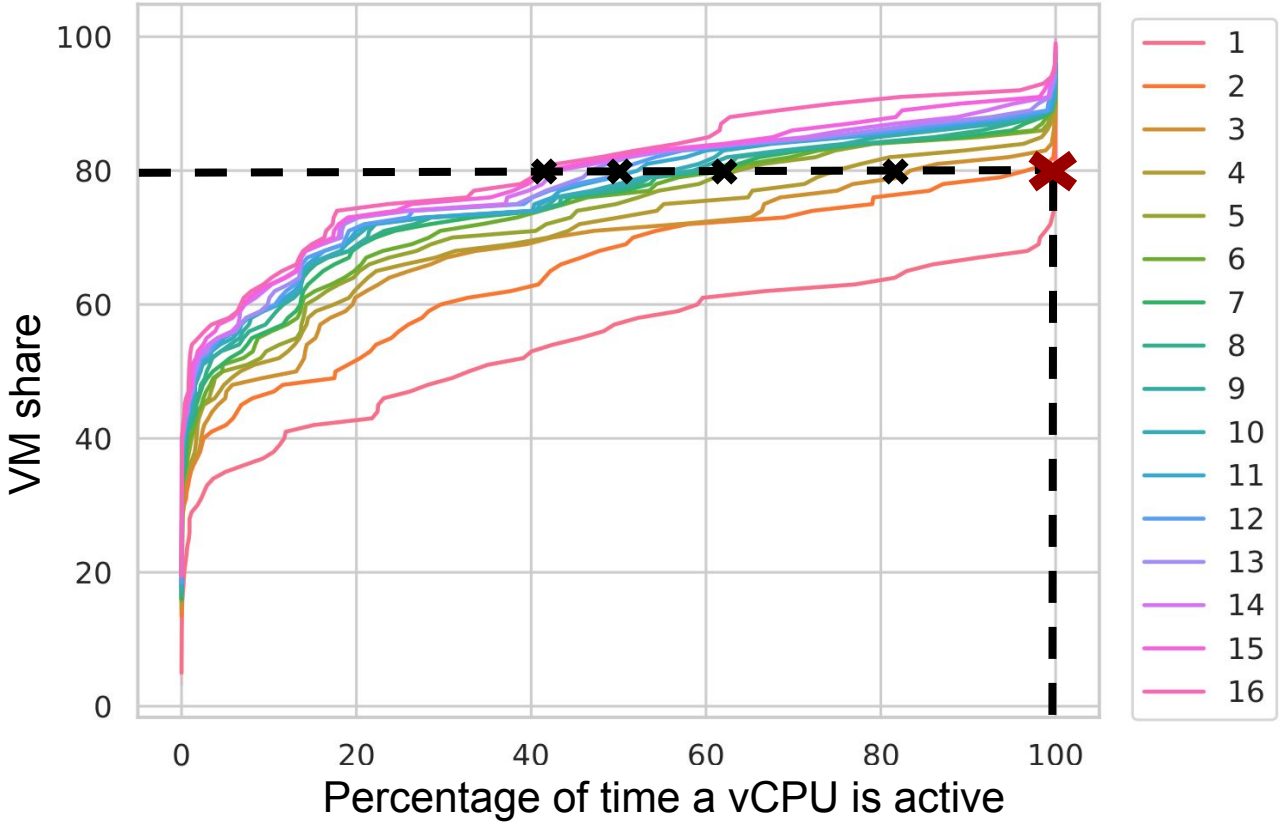


Contribution 2



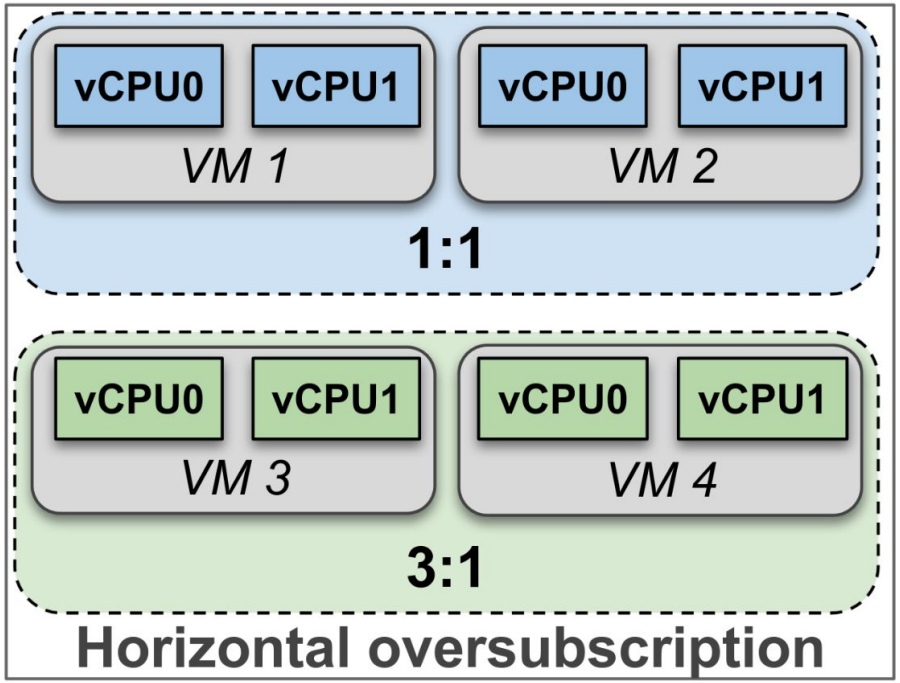
Most of CPUs are provisioned by a **small subset of VMs**

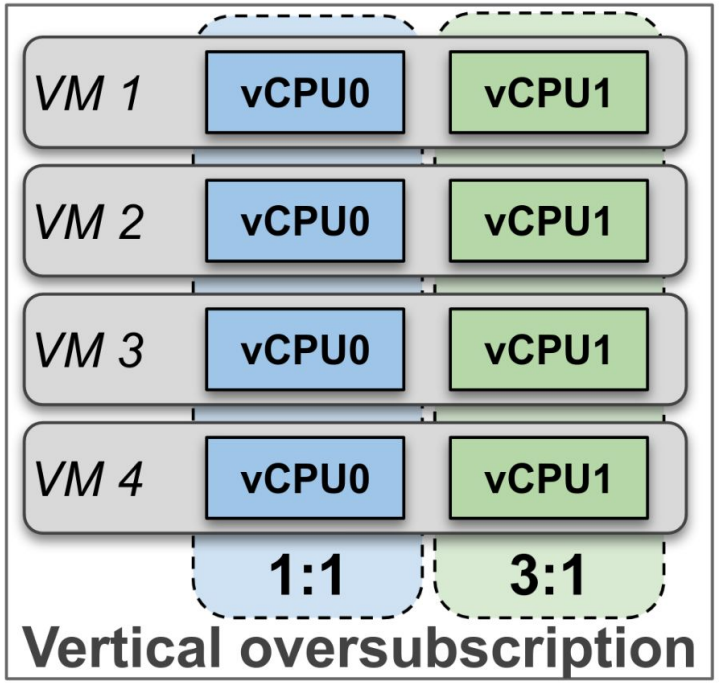
Contribution 2



Large VMs hosted by OVHCloud **do not use all vCPUs equally**

Contribution 2

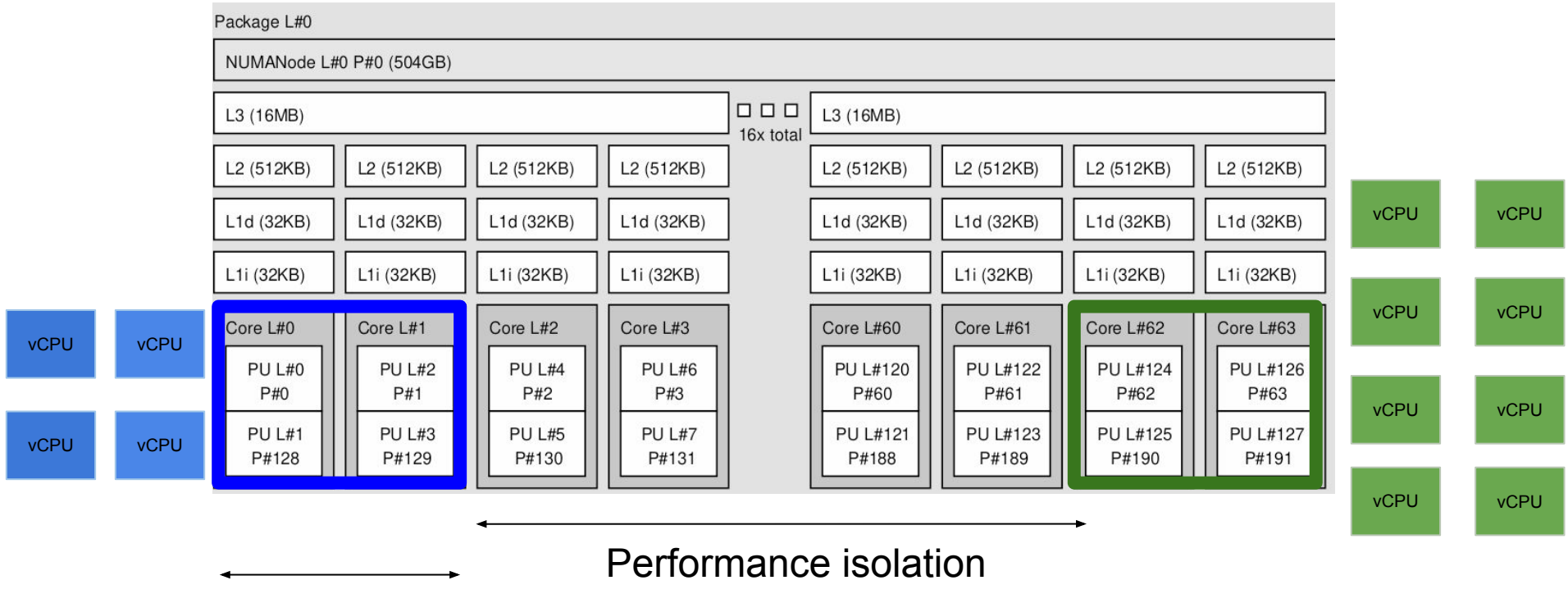




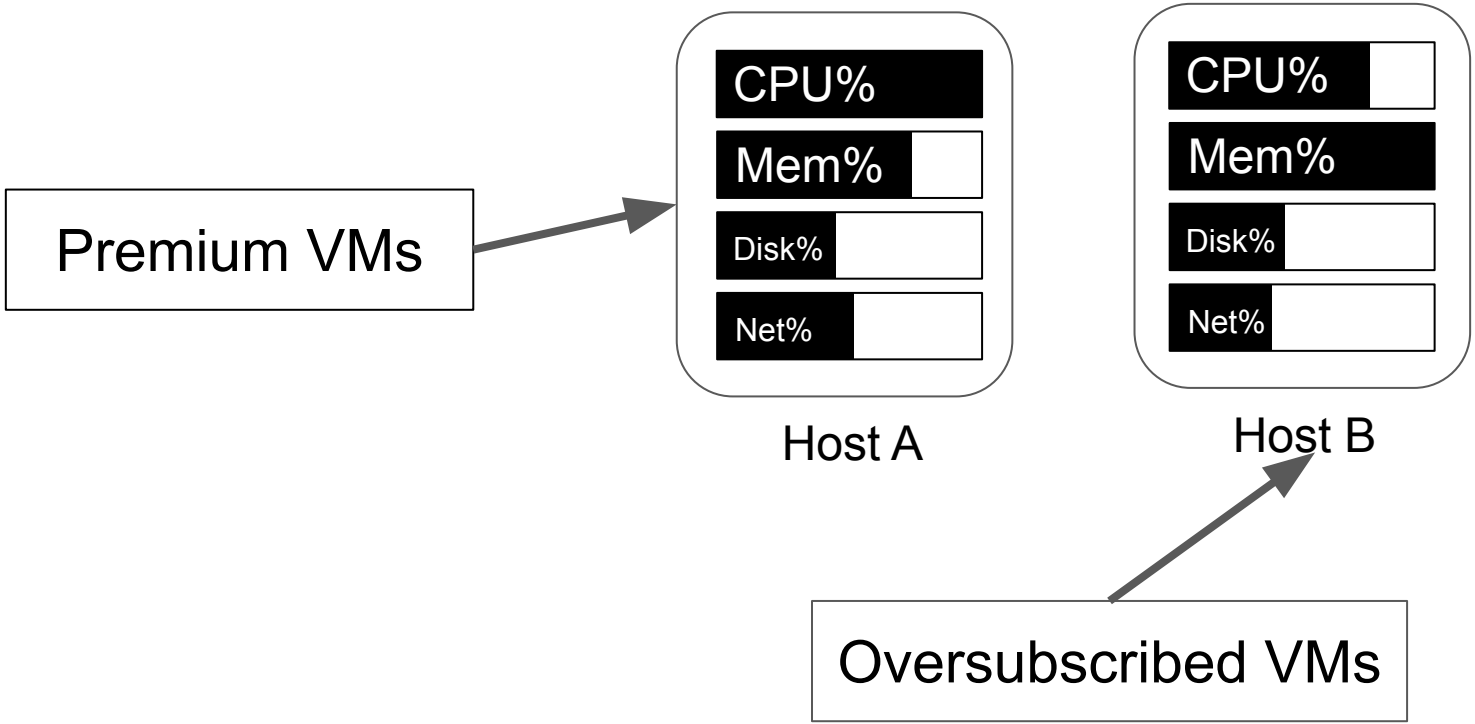
Changing the perspective on vCPU oversubscription

Contribution 2

- How to oversubscribe to multiple levels a given host?

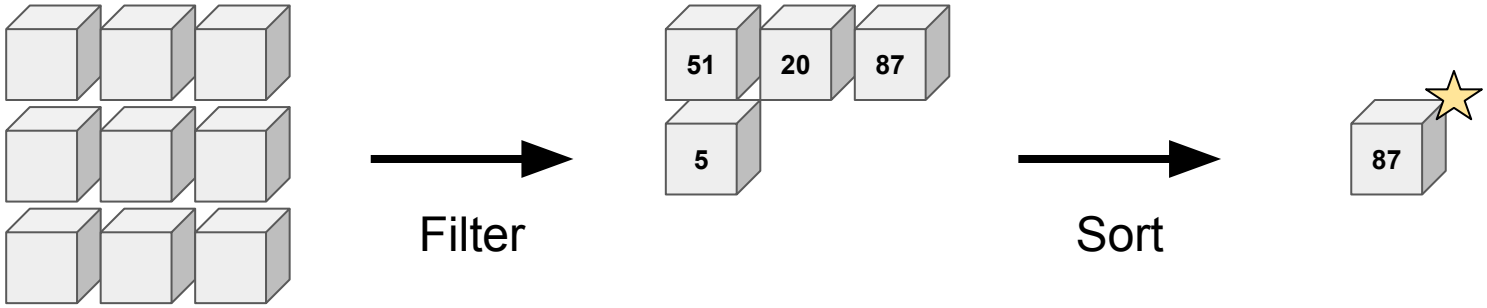


Contribution 3



Contribution 3

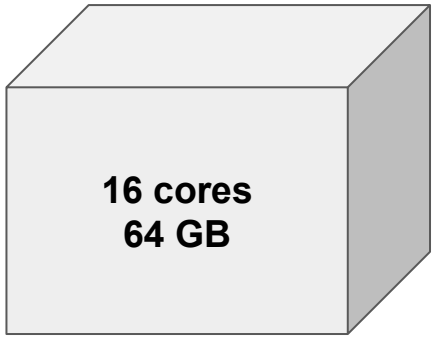
- How to orchestrate resources efficiently?



Cloud orchestrators are score-based

Contribution 3

- How to orchestrate resources efficiently?



VM1: **1 CPU – 2GB**

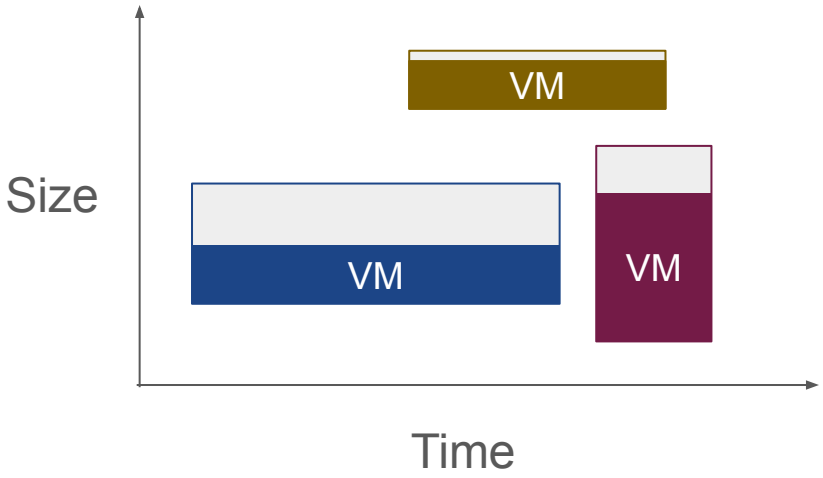
VM2: **2 CPU – 8GB**

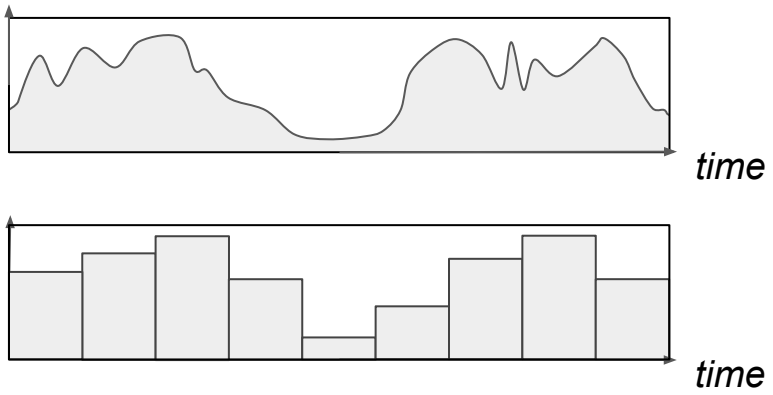
3 vCPUs allocated (~20%)
10 GB allocated (~6%)

Servers have a fixed configuration and a dynamic workload

Contribution 4

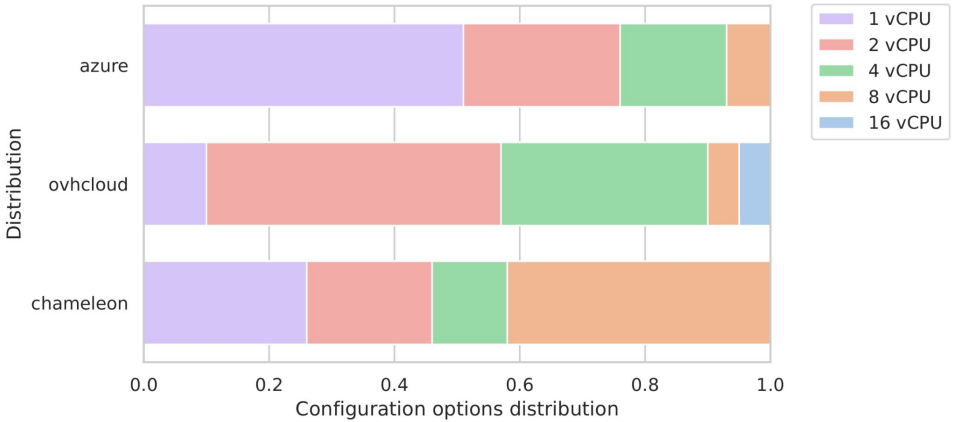
Realistic IaaS workloads



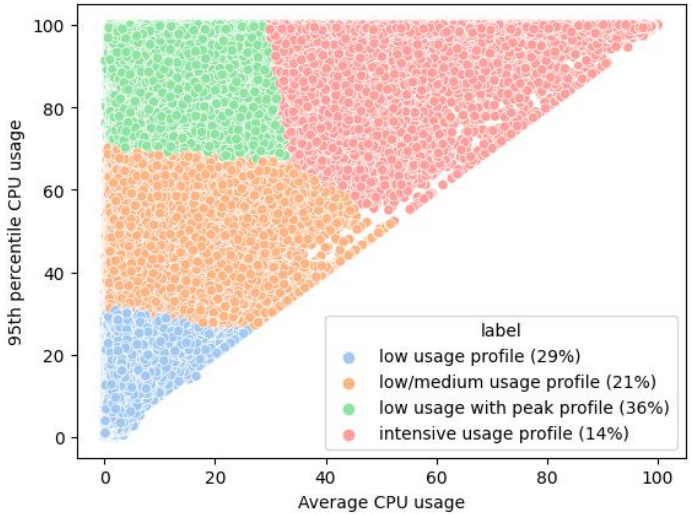


Contribution 4

Realistic IaaS workloads using **CloudFactory**



Configuration distribution on various CP



Usage profile on Azure dataset

CLOUDFACTORY: An Open Toolkit to Generate Production-like Workloads for Cloud Infrastructures, Pierre Jacquet et al., IC2E'23