Revisiting virtual machine consolidation to save resources and energy in heterogeneous production cloud infrastructures

Simon Lambert[1,2], Vladimir Ostapenco[1], Laurent Lefèvre[1], Eddy Caron[1], Anne-Cécile Orgerie[3], Benjamin Fichel[4], Rémi Grivel[2]

[1]Université Lyon 1, Inria, CNRS, ENS Lyon, LIP, [2]Ciril GROUP, SynAApS, [3]Inria, CNRS, IRISA, University of Rennes , [4]OVHcloud

[1]*firstname.lastname@ens-lyon.fr*

CONTEXT OF THE STUDY

Datacenters (DCs) account for 1.5% of global electricity demand [1]. Their operation also emits Greenhouse Gases (GHG), as well as having an impact on resources (metals, rare earths, water) because of the manufacturing and usage of the devices they host.

This is partly due to Cloud Service Providers (CSP) sizing their infrastructure to meet peak demand.
Virtual Machine (VM) consolidation [2] can help mitigate this problem and enable physical machine shutdown to reduce DCs energy consumption.
It is, however, barely used in production. Our methodology aims to identify implementation obstacles and treat them in heterogeneous infrastructures.

TARGET INFRASTRUCTURES

The study was conducted in two CSP infrastructures with heterogeneous sizes and architectures:

- Small scale infrastructure, with 2 clusters from the Ciril GROUP company
  - Cluster S1 : 6 Physical Machines ; 400 Virtual Machines
  - Cluster S2 : 16 Physical Machines, 1000 Virtual Machines
- Large scale infrastructure : one cluster from OVHcloud
  - Cluster L1 : 985 Physical Machines ; 7376 Virtual Machines

METHODOLOGY

The methodology was tested on both small and large scale infrastructure, with converging implementation.

Methodology step	Small scale	Large scale
Detect consolidation potential	Initial infrastructure allocation ratios (CPU, memory, storage, etc.) and usage study	
Low Resource Usage (LRU) VM detection	LRU VM algorithm on CPU and network	LRU VM algorithm on CPU, network and disk
Consolidation strategies	1 strategy : all VMs with oversubscription	4 strategies based on VM scope and use of oversubscription
VM placement	First Fit Decreasing heuristic to reduce number of PMs	Optimal Bin Packing Solver to reduce number of PMs and migrations
Allocation ratios study	Based on manual experimentation RAM (1/2/3)	Based on hypervisor documentation CPU (1/2/4/8) and RAM (1/1.5/2)
Simulation process	Python simulator based on clusters traces	SimGrid [3] simulator based on cluster traces

LRU VM DETECTION ALGORITHM

We used an algorithm to compute the amount of **Low Resource Usage (LRU)** VMs in the infrastructures :

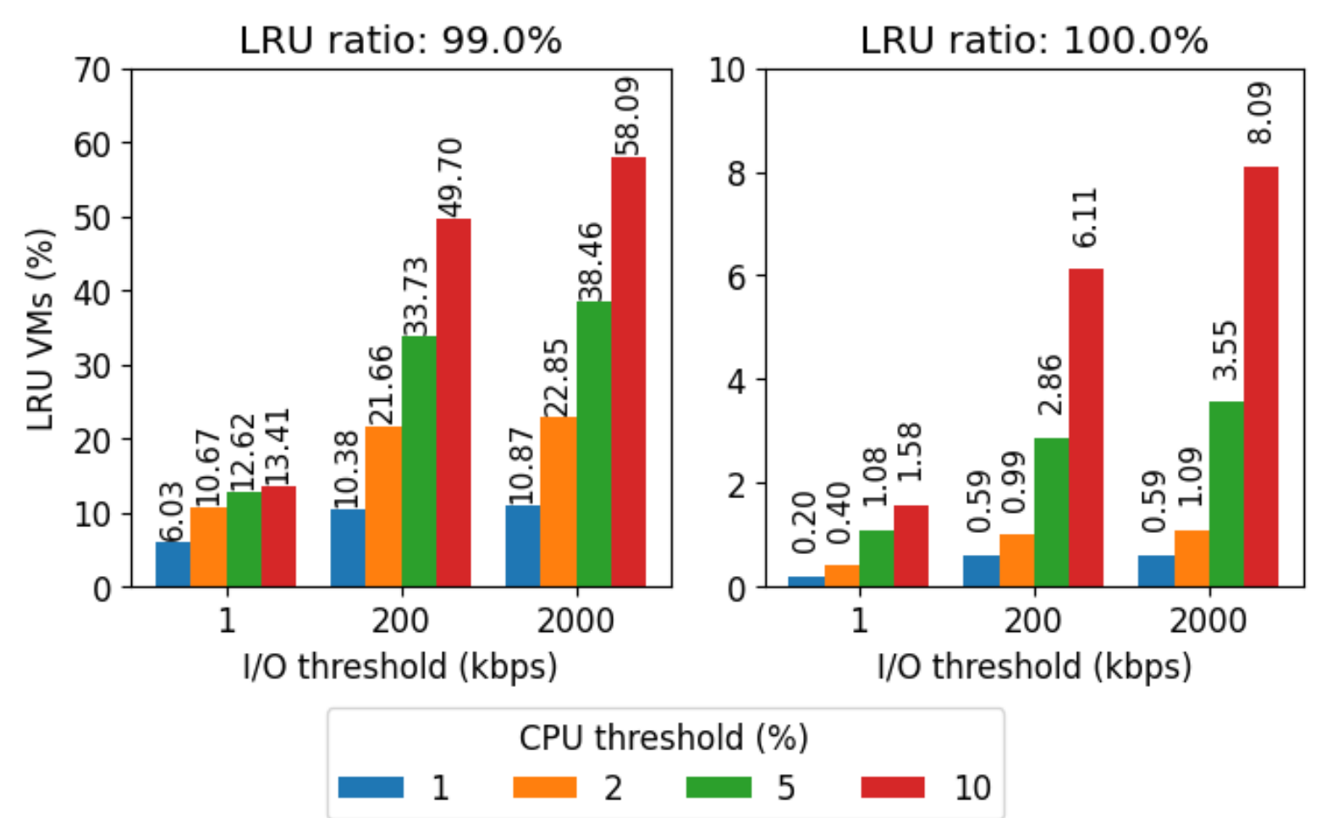
```
is_used(vm, t, cpuRate, netRate, diskRate):
  if cpu_usage(vm,t) <= cpuRate and net_usage(vm, t) <= netRate [and disk_usage (vm, t) <= diskRate]:
    return 0
  else:
    return 1
```

This enables assigning a state to the VM and computing its LRU rate over a specific period of time. Two main LRU rates are studied:

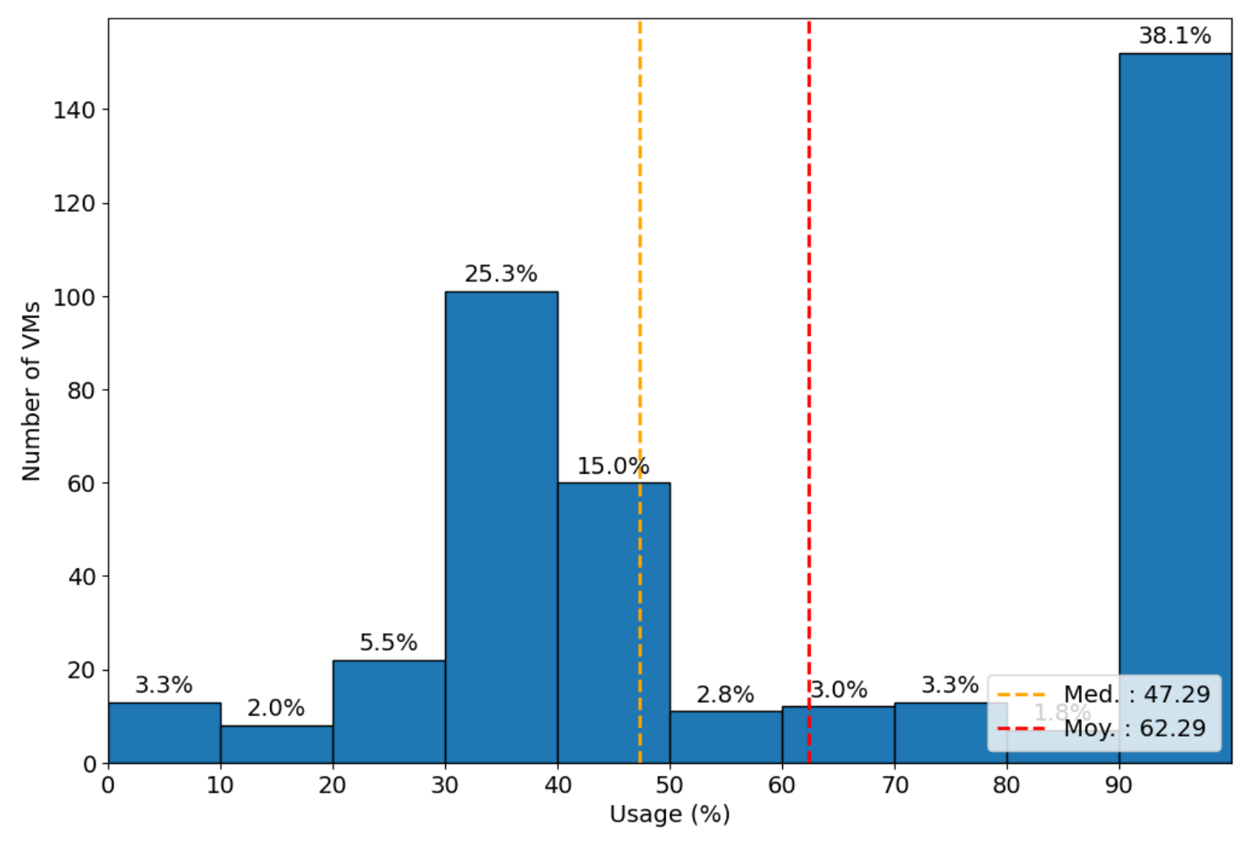
- LRU-99 : VMs considered as idle 99% of the time or more over the studied period
- LRU-100 : VMs always inactive over the period

SMALL SCALE INFRASTRUCTURE RESULTS

VM USAGE

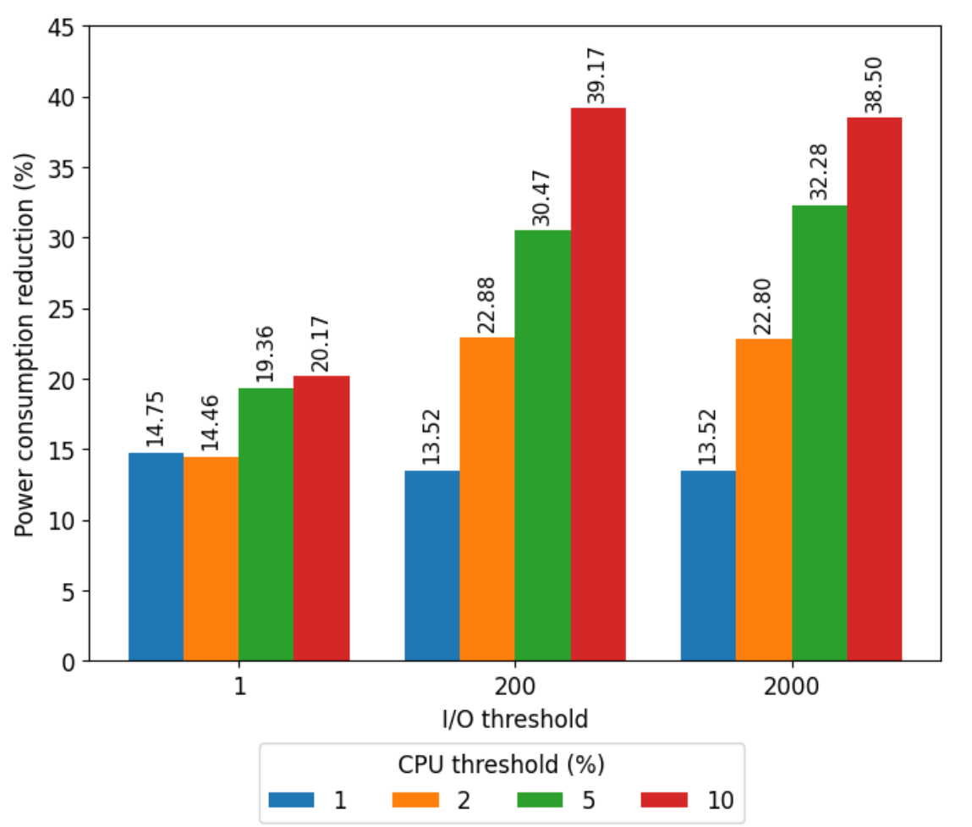


LRU-99 and LRU-100 VMs on the small scale infrastructure with varying CPU and I/O thresholds

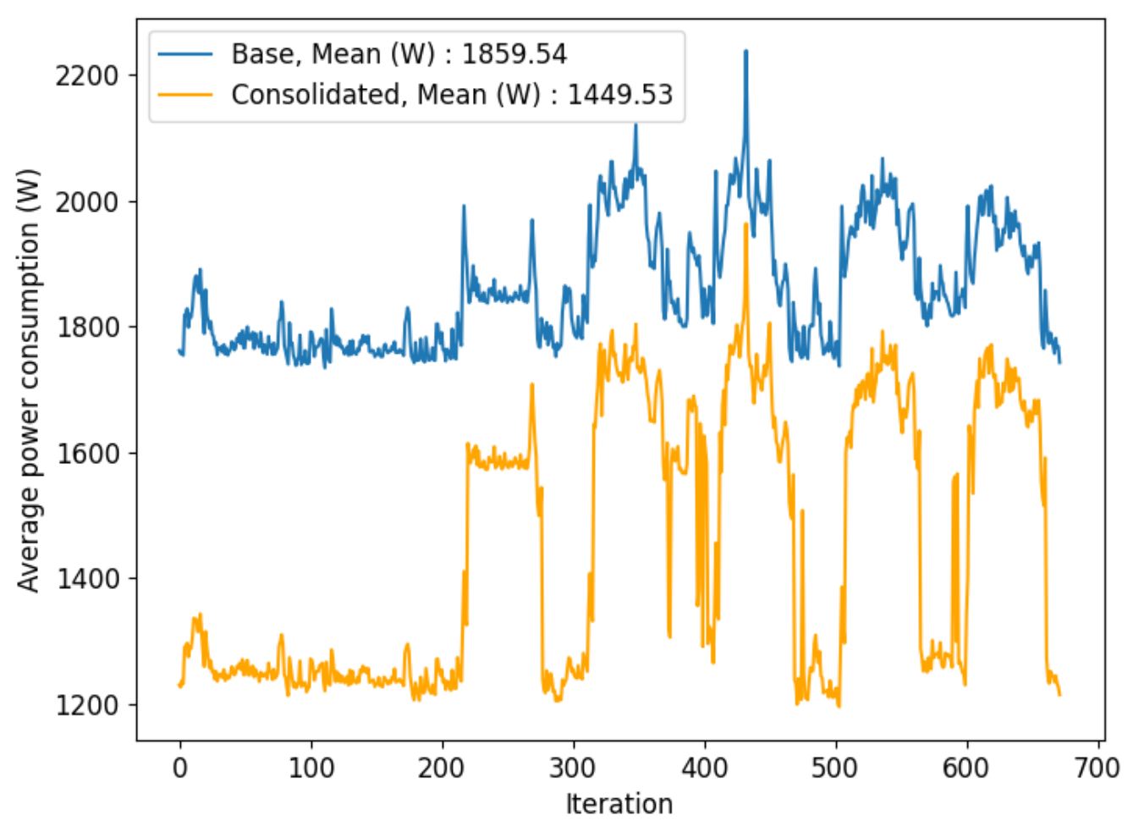


VM usage distribution on Cluster S1 over one week

ENERGY CONSUMPTION REDUCTION



Power consumption reduction for LRU-90 VMs with oversubscription ratio of 2



Mean power consumption over one week for LRU-80 VMs with oversubscription ratio of 3. CPU threshold : 1% ; I/O threshold : 1kbps

On the small-scale infrastructure, consolidation with oversubscription enables an important power consumption reduction. Using CPU and network thresholds of 1, and LRU-80 VMs, with an oversubscription ratio of 3, we obtain a **22.05% power consumption reduction**.

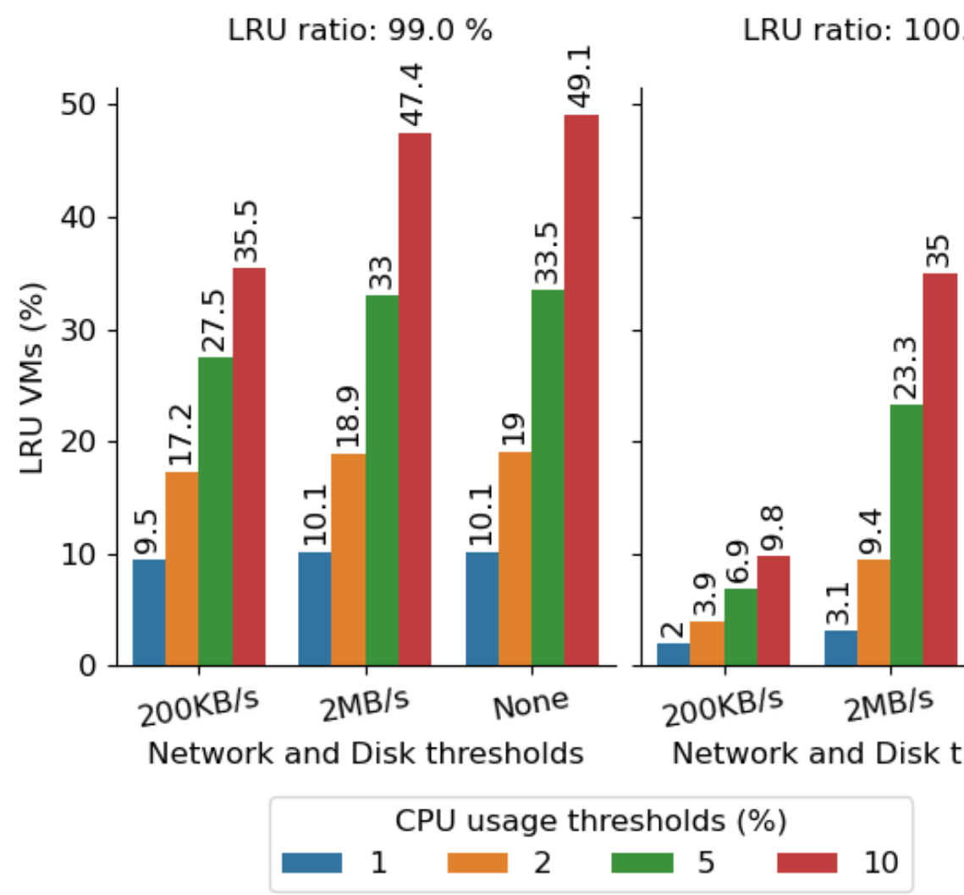
[1] IEA (2025), Energy and AI, IEA

[2] Bermejo, B., Juiz, C. & Guerrero, C. Virtualization and consolidation: a systematic review of the past 10 years of research on energy and performance. J Supercomputing 75, 808–836 (2019)

[3] Casanova H, Giersch A, Legrand A, Quinson M and Suter F. Versatile, scalable, and accurate simulation of distributed applications and platforms. Journal of Parallel and Distributed Computing 74(10): 2899–2917 (2014)

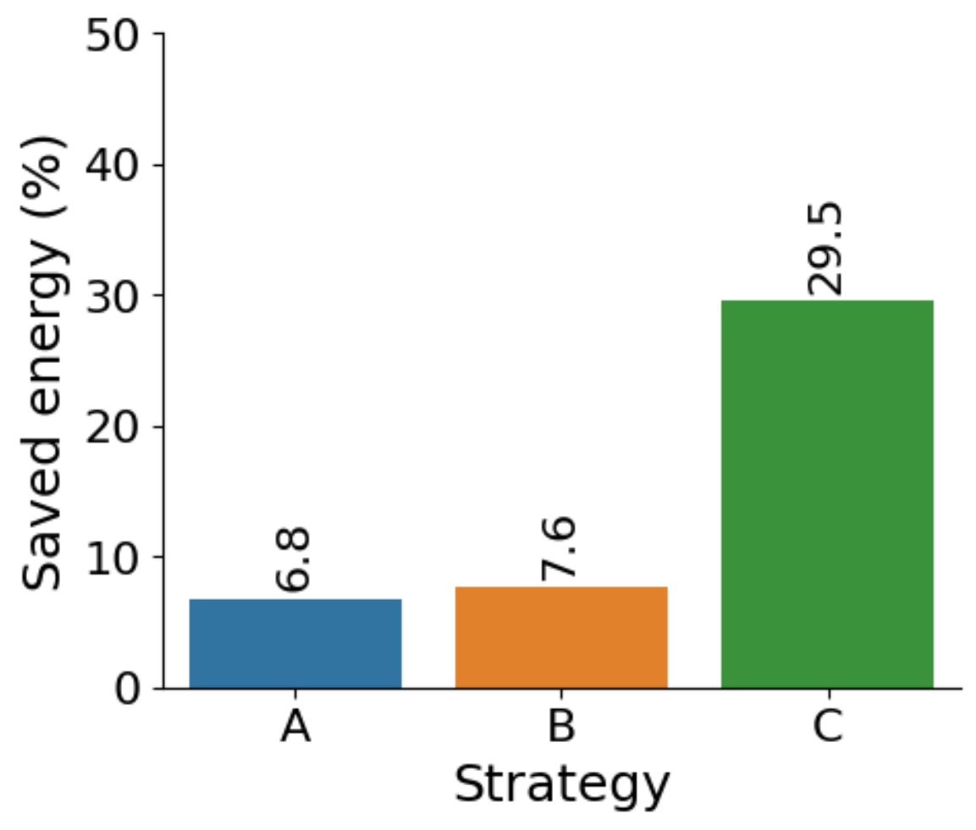
LARGE SCALE INFRASTRUCTURE RESULTS

VM USAGE



LRU-99 and LRU-100 VMs on the large scale infrastructure with varying CPU and I/O thresholds

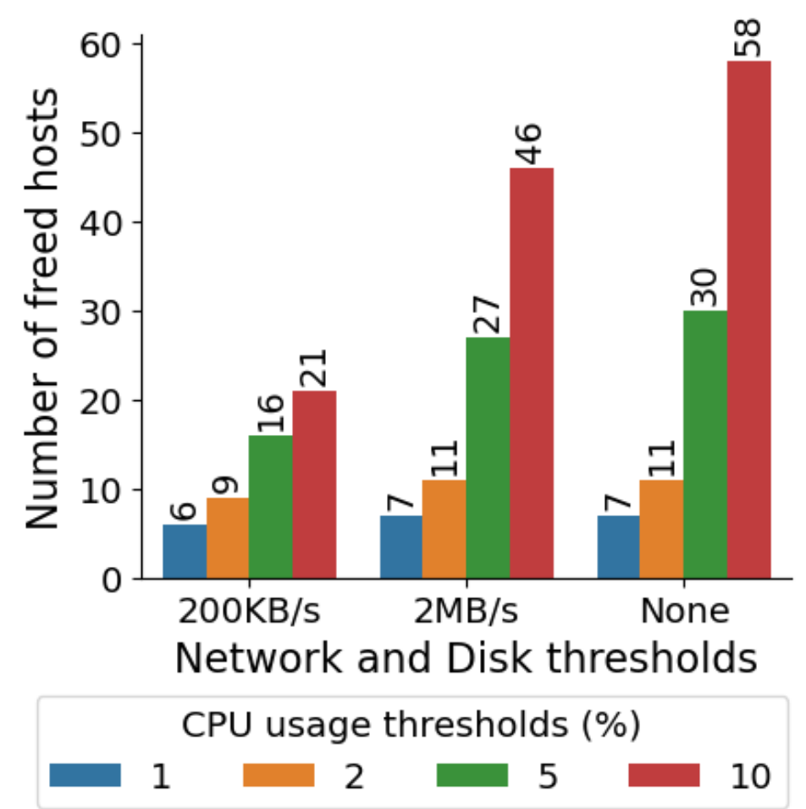
ENERGY CONSUMPTION REDUCTION



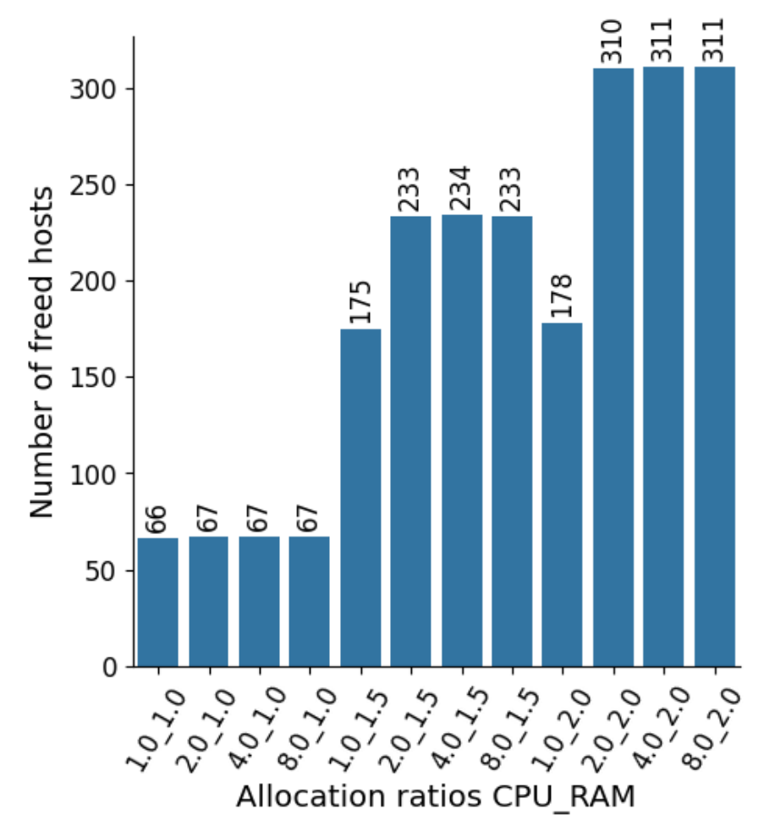
Reduction with 3 migration strategies :
A - LRU VMs migration strategy for the most relaxed thresholds.
B - General VM migration strategy without oversubscription.
C - General VM migration with oversubscription and allocation ratios of 2 for both CPU and RAM

FREED PHYSICAL MACHINES

Amount of freed PMs with different strategies:
A – Freed PMs with LRU VM migration strategy and LRU-99 VMs. No oversubscription.
B – Freed PMs for general VM migration strategy using the most relaxed thresholds.



A



B

Consolidation with oversubscription on large-scale infrastructure enables both power consumption reduction and an immediate PM usage reduction. We can **free up to 311 PMs**, which represents significant **electricity consumption, GHG emissions, as well as mineral resources** used for their manufacturing.