SC26 Chicago, IL hpc unites.

No Disk Left Behind: Lifetime-Aware Oversubscription of Thin-Provisioned Storage in Cloud IaaS platforms

Pierre Jacquet, Camille Coti, Marcos Dias De Assunção

ÉTS Montréal, Université du Québec, Canada

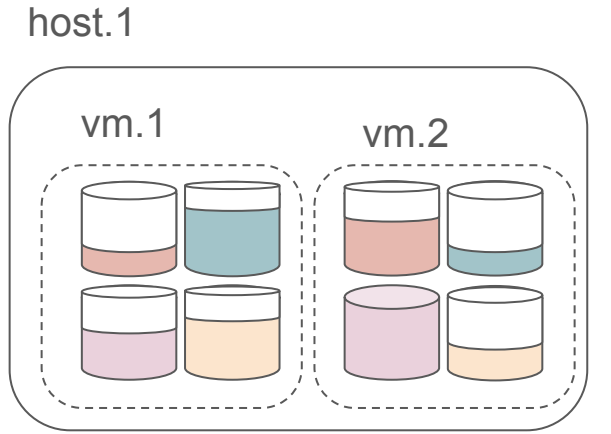
ÉTS ÉCOLE DE TECHNOLOGIE SUPÉRIEURE Le génie pour l'industrie Université du Québec

OVHcloud™

mitacs

Introduction

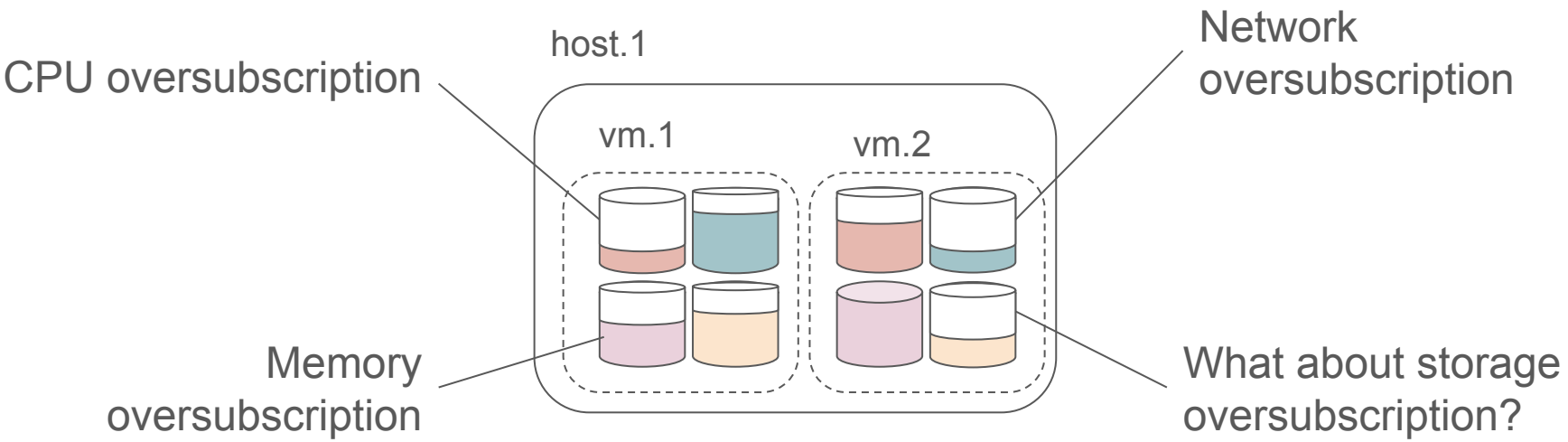
Cloud orchestrators place VMs as allocation on different resource types



Allocation differs from usage

Introduction

Cloud platforms can use oversubscription to improve resource utilization



Introduction

Does storage oversubscription make sense in IaaS?

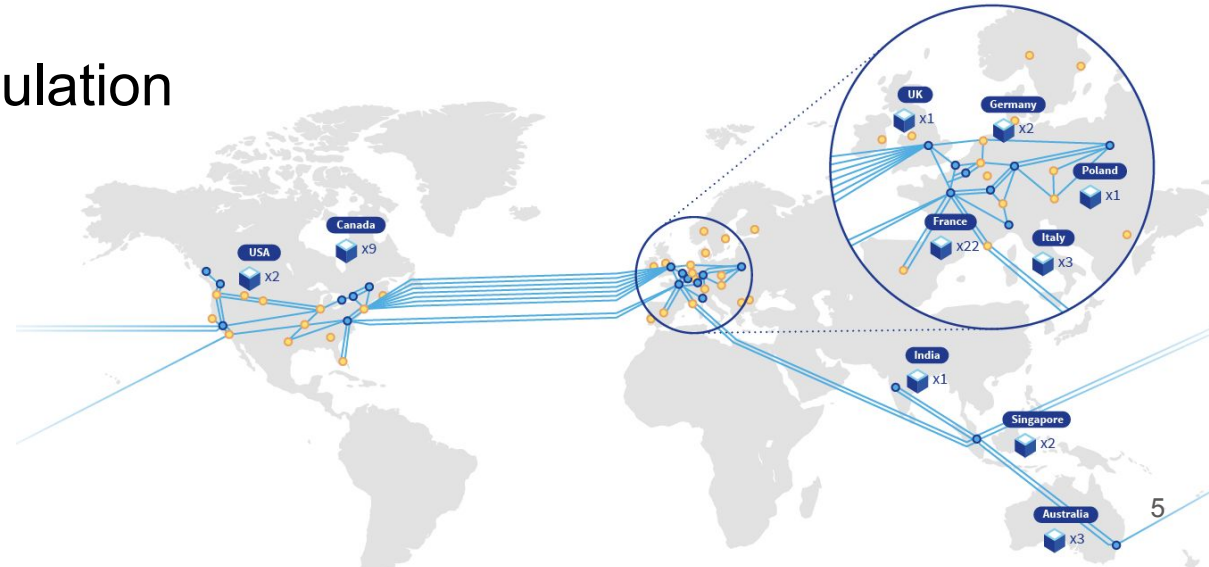
On clusters relying on local storage:

- **P1**: Virtual disks are **thin-provisioned**

- **M1**: Storage can be a **bottleneck**
- **M2**: Storage is more and more **expensive**
- **M3**: Storage manufacturing has a high **environmental impact**

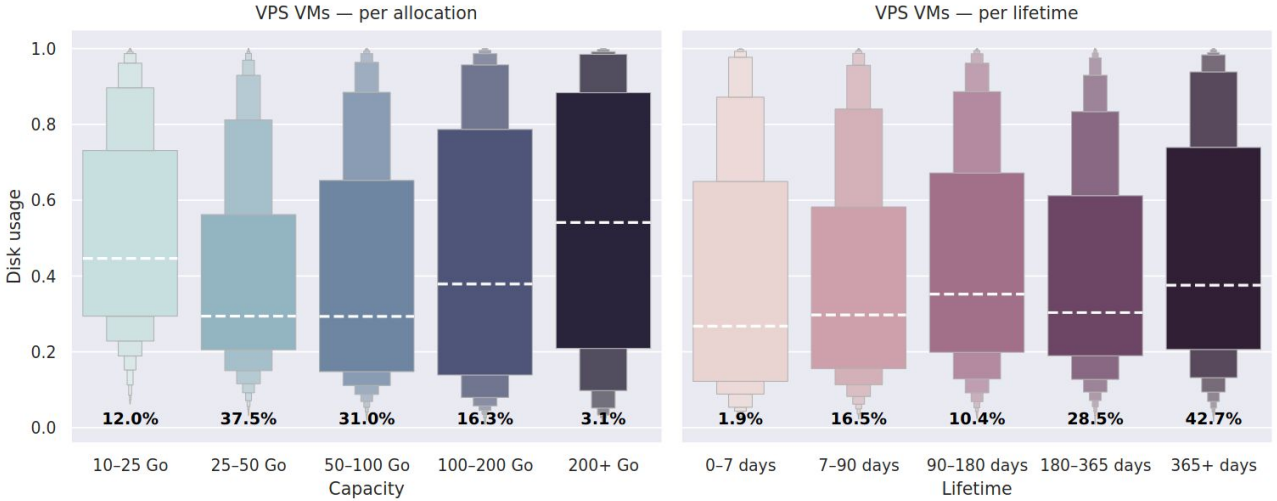
Plan

- **Storage usage** at scale
  - Case study based on OVHcloud
- Our **proposal**
  - Oversubscription based on Disk Release Prediction (*DRP*)
- **Evaluation**
  - Trace-driven simulation



Storage usage at scale

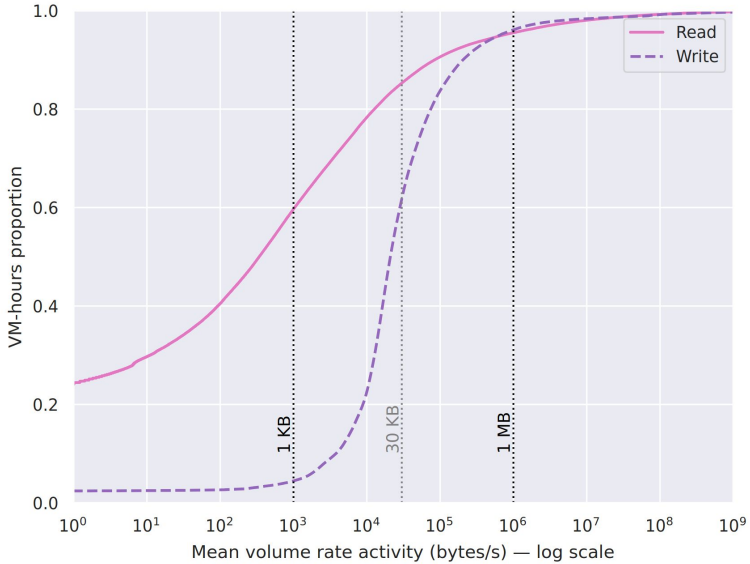
1) Significantly underused in terms of capacity



Local storage allocation are only used at 43%, with no clear correlation to size/lifetime

Storage usage at scale

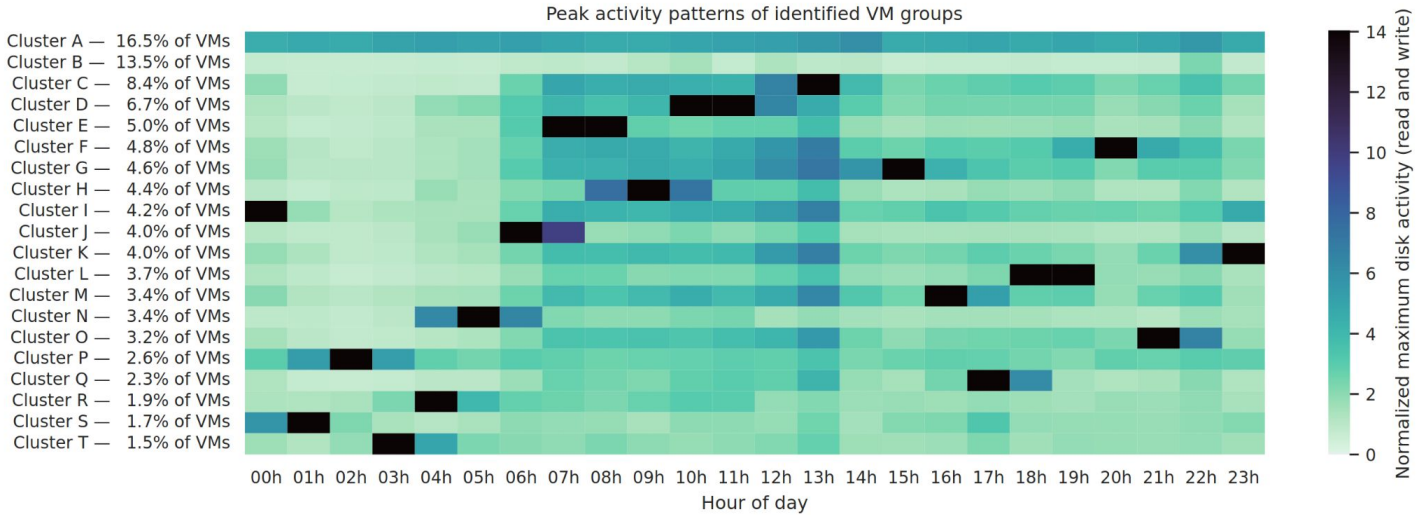
2) Significantly underused in terms of bandwidth



Bandwidth ≥10× below typical allocation (~1000 IOPS ≈ 16 MB/s at 16 KB)

Storage usage at scale

3) Expose good temporality dispersion between domains



VM groups clustered by hour of peak z-score activity

Storage usage at scale

1) Significantly underused in terms of capacity
2) Significantly underused in terms of bandwidth
3) Expose good temporality dispersion between domains

Disk oversubscription is technically **feasible** but **challenging**:

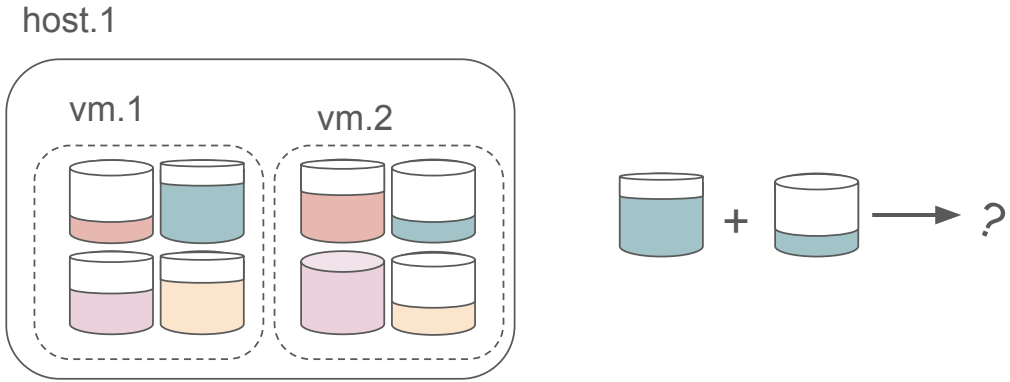
- Local disk capacity is not fungible
- Free disk must always be available for VMs
- VM lifetimes can span years
- Disk writes are bursty and hard to predict

Proposal

Managing oversubscribe resources involve in an online setting:

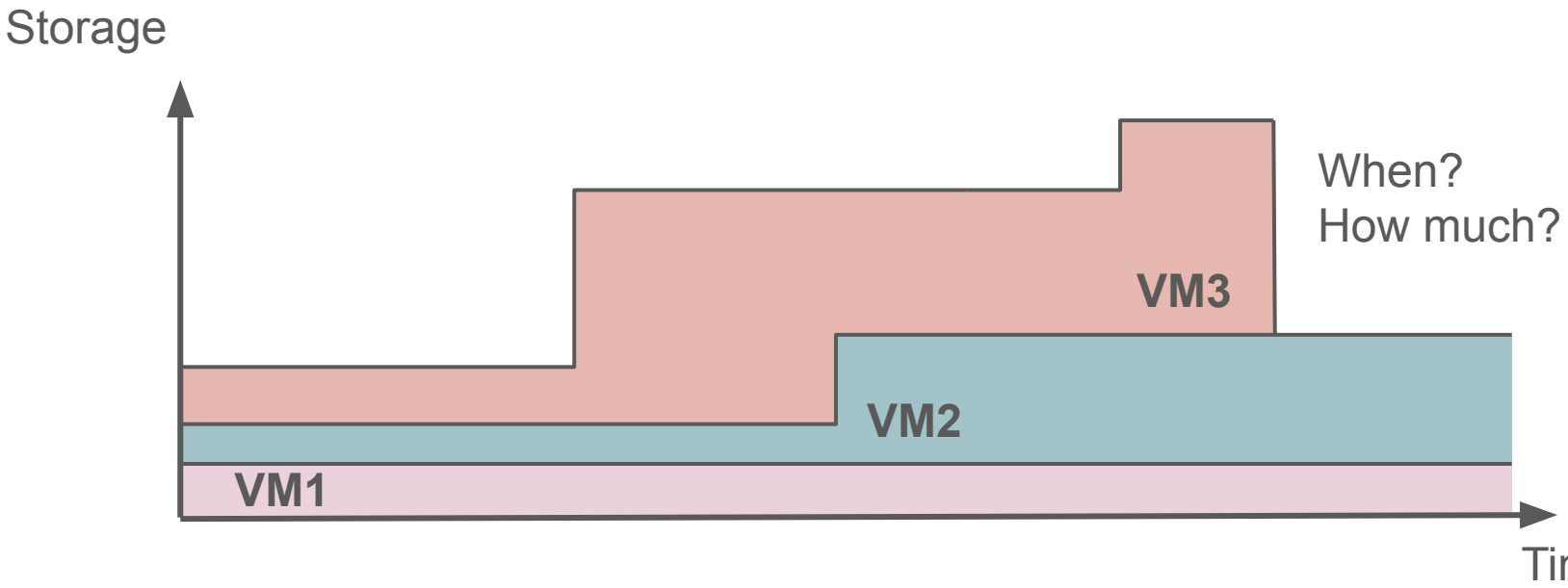
1) *Placement*: evaluate free resources from predicted disk needs
2) *Runtime*: mitigate contention via VM live migration

We aim at leveraging 1) to minimize 2)

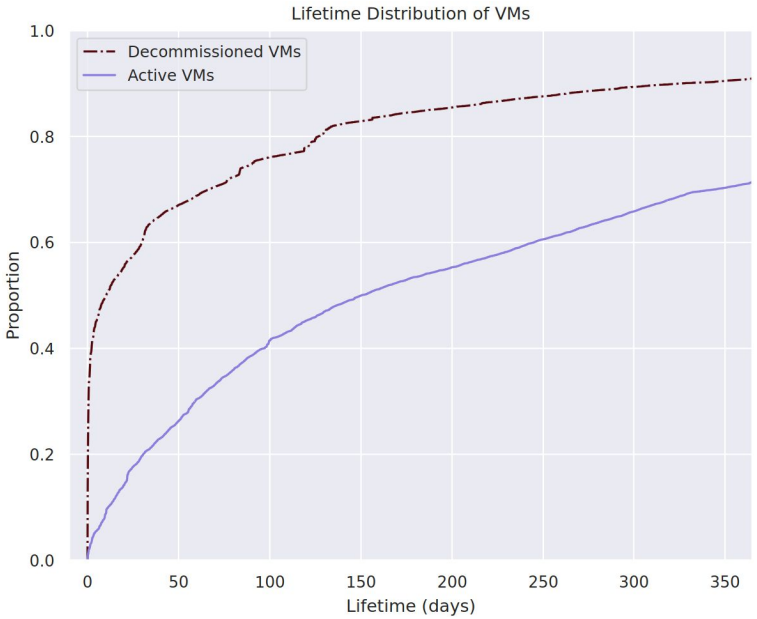


Proposal

Predicting peak/final usage is hard → predict release windows instead



Proposal



CDF of VM lifetime per label (terminated VMs, active VMs)

VM lifetime estimation looks easy:

- Per-VM survival probability
- Under incremental usage, gives the timestamp of max usage

… But is uncertain

- Preferable: compute the probability that n domains leave by time T, given their current lifetimes

Can we do better?

Proposal

- The Disk Release Prediction (*DRP*) method:
  - Probability that freed disk exceeds ***w*** GiB by time ***t***

Algorithm 1 Probability that freed disk exceeds w by time t

Input: time t, running VMs $\{\text{VM}_i\}$, target size w (GiB)

$p_i \leftarrow P(\text{VM}_i \text{ departs before } t)$

$\mu \leftarrow \sum_i w_i p_i$

$\sigma^2 \leftarrow \sum_i w_i^2 p_i (1 - p_i)$

if $\sigma = 0$ **then return** $\mathbf{1}[\mu \geq w]$

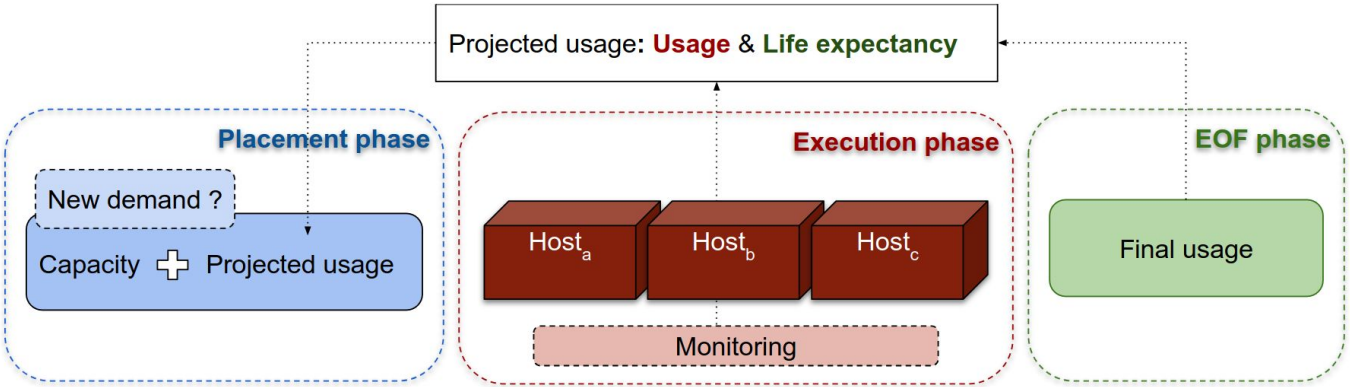
end if

return $1 - \Phi\left(\frac{w - \mu}{\sigma}\right)$

  - Find smallest **t** with P ≥ threshold (Brent)
  - Project usage up to **t**

Proposal

- To summarize:
  - The online scheduler applies *DRP* on each host
    - Obtain the projected disk usage before the next likely departure(s)
    - Judge whether enough disk is left to place new VMs



Evaluation

Simulator description

- Final disk usage, lifetime, and cluster hardware reproduced from the OVHcloud context
- When a VM is spawned
  - Writes on n ticks are randomly assigned over its lifetime, up to the final usage
  - OpenStack-like scheduler estimates free cluster capacity and selects the host

The simulator and traces are open source; we hope to encourage more research on this topic

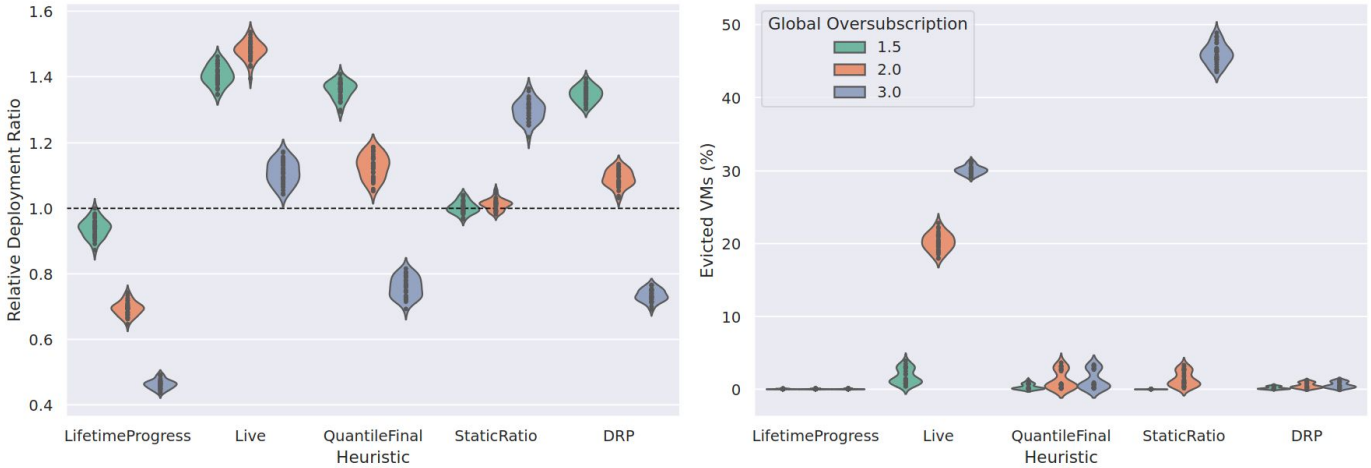
Evaluation

We are not aware of previous work on the usage prediction of non-fungible resources. We compared to:

- Static oversubscription ratio
- Live usage
  - *Current usage = final usage*
- Quantile Projection
  - *Classify VM usage per quantile. Apply this quantile to final usage measurements*
- Lifetime Progress
  - *Estimate each VM's lifetime and sum the projections*

Evaluation

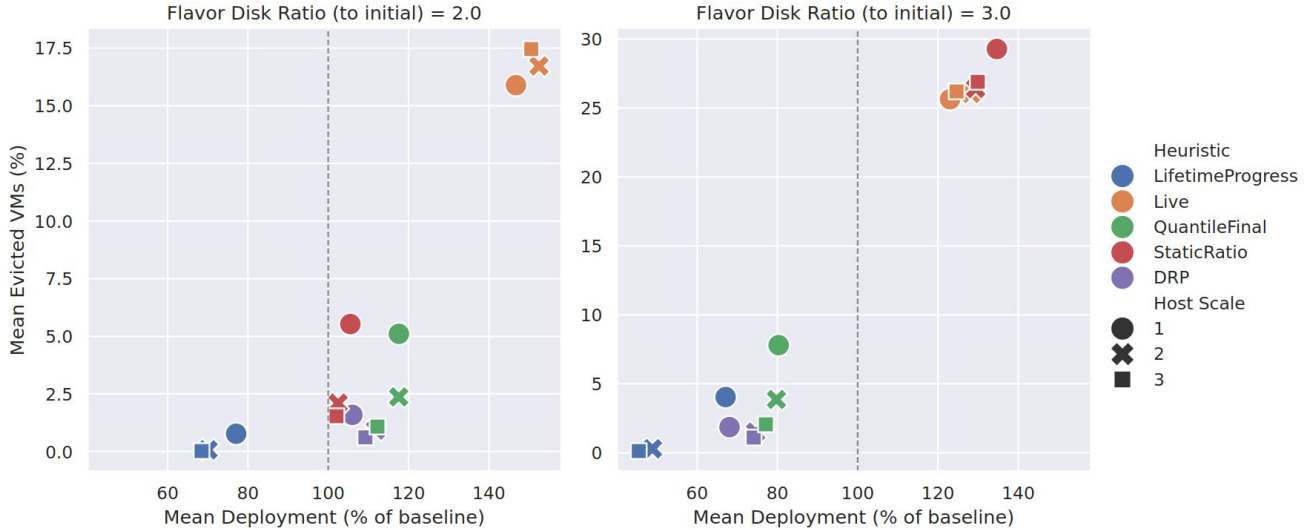
Can disk oversubscription enable more VMs to be deployed? (**M1**)



With **2:1** oversubscription, DRP has the best deployment/eviction ratio
With excessive oversubscription (**3:1**), it remains safe, with near 0% evictions

Evaluation

Can disk oversubscription reduce SSD capacity requirements? (**M2**|**M3**)



Halving SSD (left graph) is feasible with DRP while preserving VM count (*X axis*), and while having a minimal eviction rate (*Y axis*).

Evaluation

Can disk oversubscription reduce SSD capacity requirements? (**M2**|**M3**)

Yes! And halving SSD capacity reduces the scope 3 environmental impact (indirect emissions) associated with storage by 44.5%

On the studied 70k-VMs cluster, this represents 65 t CO2e

Additional RQ in the paper: increase the disk allocation per VM

Take-aways



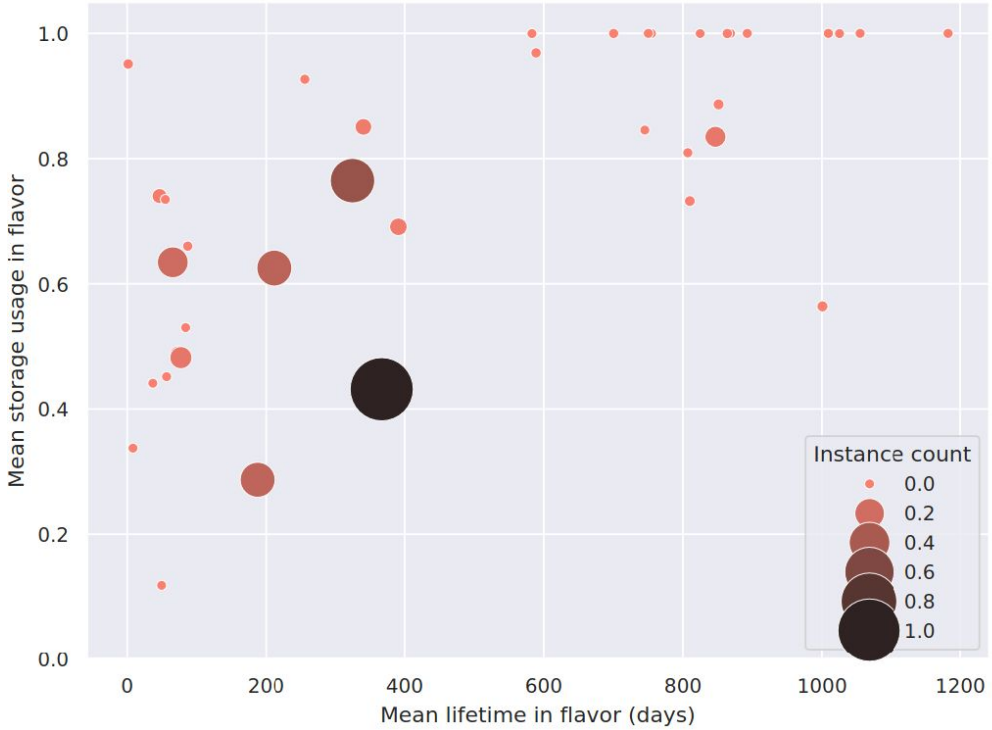
- Optimizing storage capacity gains traction to the financial & environmental cost
- Oversubscription of local storage is feasible due to the under usage of bandwidth and capacity
- We proposed a new approach (DRP) to oversubscribe non-fungible resources, leveraging lifetime predictions

Our traces and simulator →



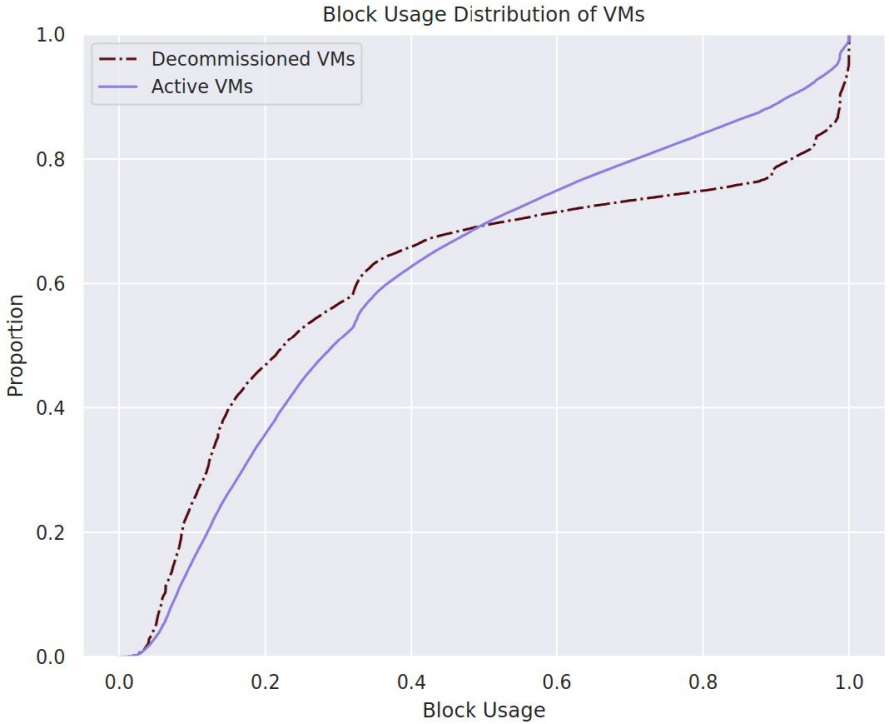
Interested in cloud topics?
I'm open to collaboration! pierre.jacquet@etsmtl.ca

Backup



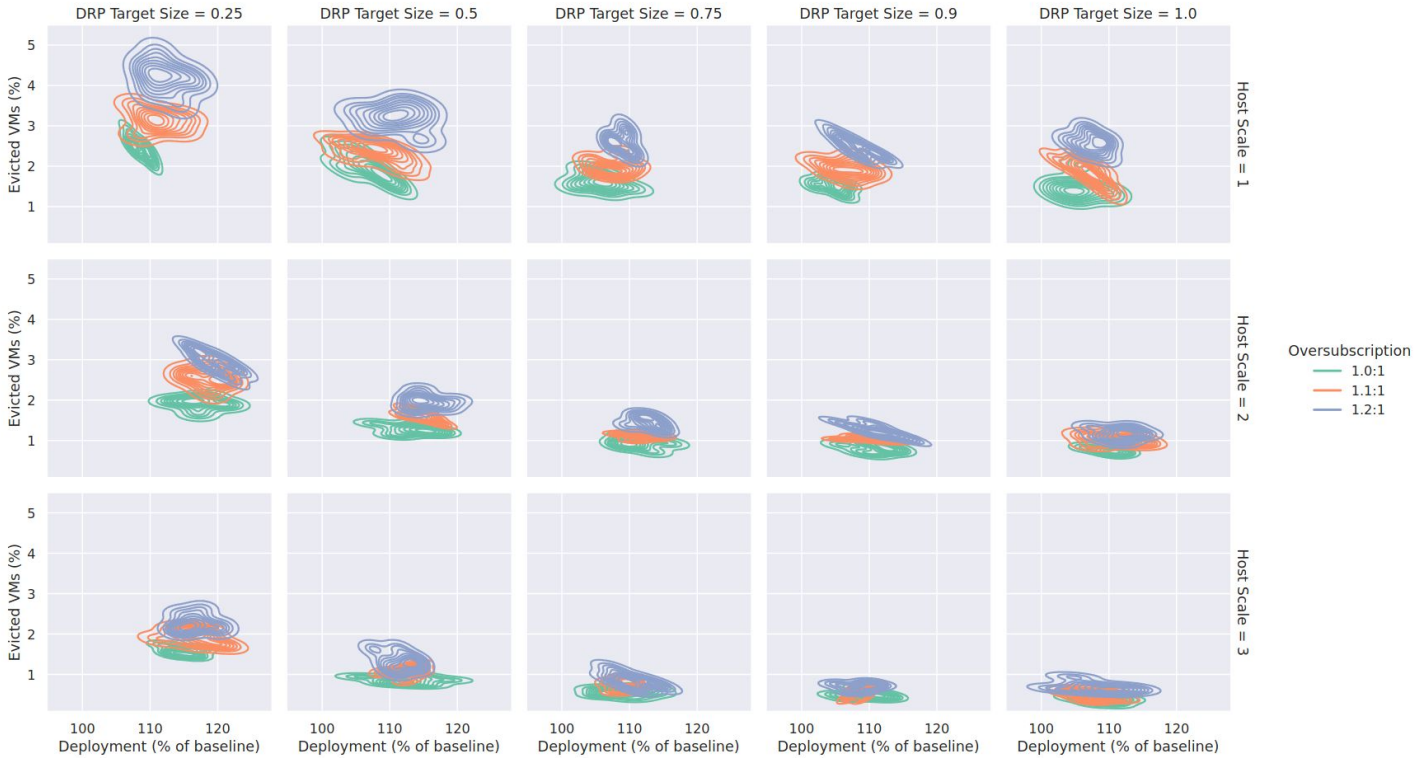
Comparison of mean disk usage and lifetime across clusters

Backup



Comparison of disk usage between active and decommissioned VMs

Backup



Analysis of the influence of the DRP target size parameter on results