TER2022-033 - Limiting Memory Reclaiming Impact on VMs Performance

Étudiant : MÉRIENNE Alexis

Encadrants : LOPEZ Dino Pecheco,

APARICIO PARDO Ramon

Issues and motivation

- **Memory is the rarest hardware ressources in datacenter. Optimization of this ressources exists. Here, we want to go farther and taking memory unuse INSIDE the VM.**
- Study VM's comportement when memory retsriction
- Elaborate intelligent solution to manage memory allocation without loss of service quality

VM's Memory Measurement

We want to study the comportment of a virtual machine (VM) when we stress it with memory restriction. The VM is an ubuntu 20.04 server, located on a host server. The is :
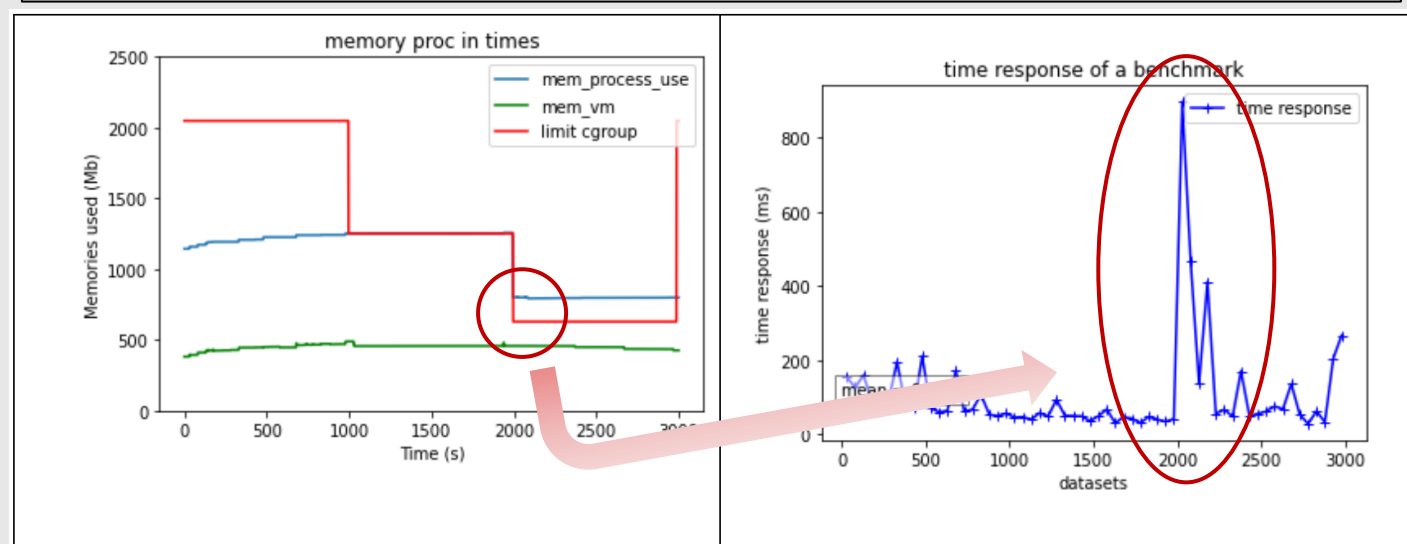
- 2 processors
- 2048 Mo of memory
- 15 Go of disk space

4 metrics to study VM's behavior :
- *mem_proc* -> memory use by VM seen by host
- *mem_vm* -> memory use by VM seen by itself

We can't really measure with precision memory used by a process or thread

VM behavior

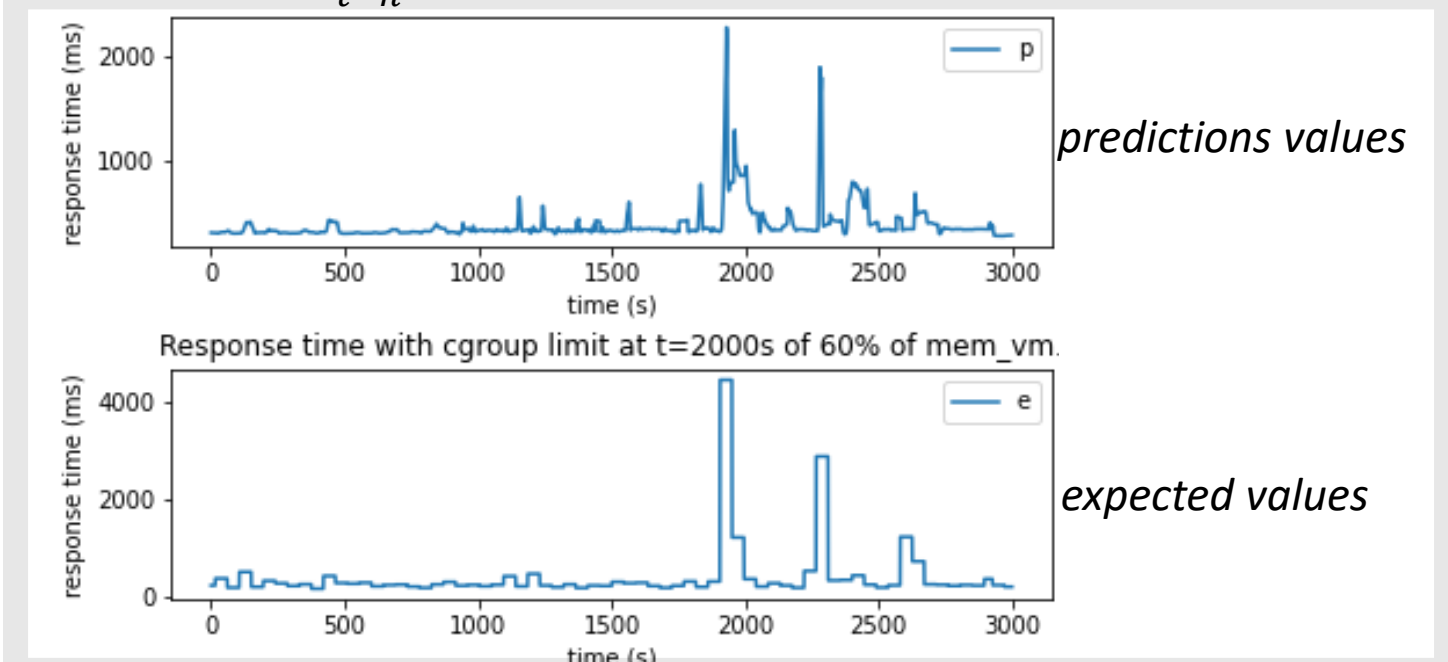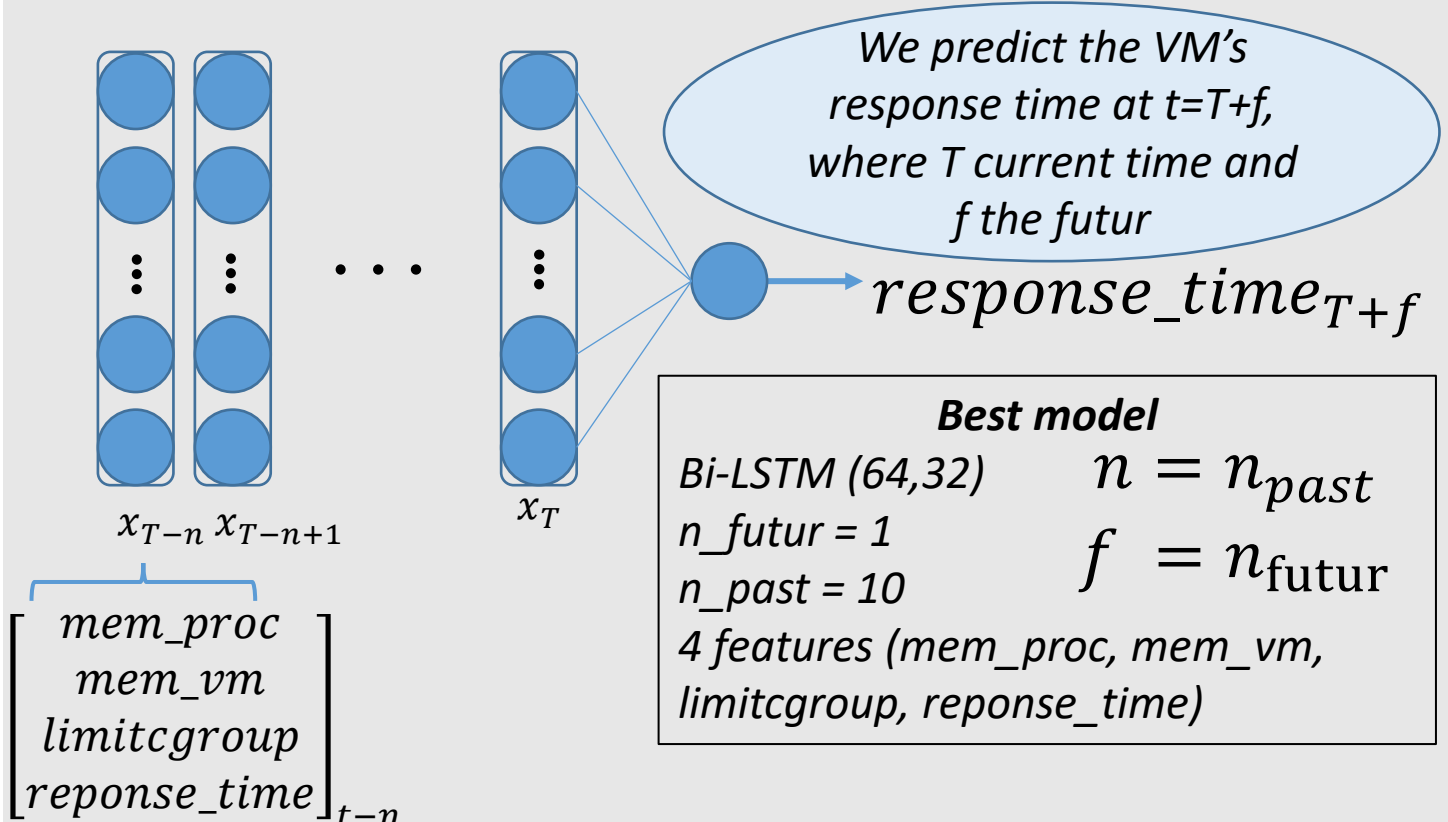VM Ubuntu20.04 – Apache2 server – 2CPU – 2Gb RAM initially

high drop in quality of service at t=2000s

When we reduce the virtual machine's memory to a level below what is needed to run the kernel and the apache server, the host operating system will search for memory addresses on the swap and will then save them on the RAM. However, when the RAM becomes overwhelmed, some addresses will be moved to the swap. These operations require a significant amount of time, which results in noticeable differences in the benchmark response time.
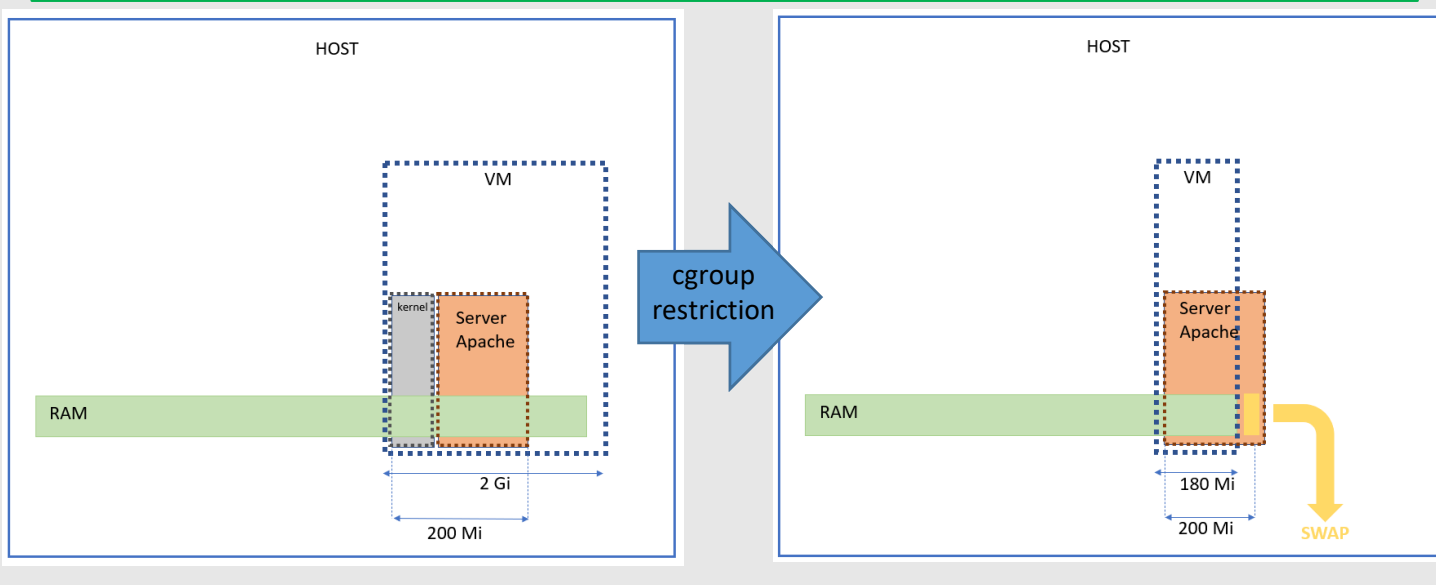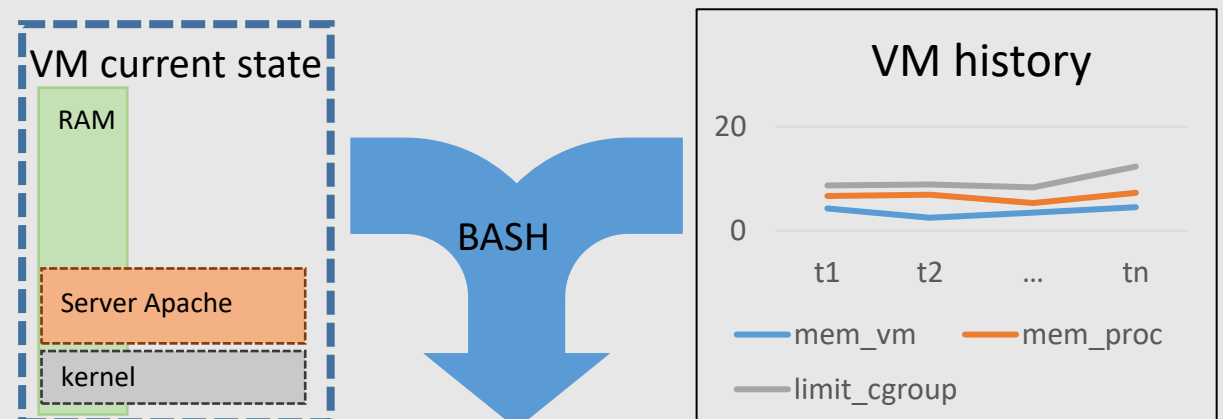

Strategie – MOIM

Use Machine Learning to optimize memory ressource and prevent quality drop

We predict the VM's response time at t=T+f, where T current time and f the futur

$$response\_time_{T+f}$$

$$\begin{bmatrix} mem\_proc \\ mem\_vm \\ limitcgroup \\ reponse\_time \end{bmatrix}_{t-n}$$

Best model
Bi-LSTM (64,32)
n_futur = 1
n_past = 10
4 features (mem_proc, mem_vm, limitcgroup, reponse_time)

$$n = n_{past}$$
$$f = n_{futur}$$

predictions values

Response time with cgroup limit at t=2000s of 60% of mem_vm.

expected values

MOIM (Memory Optimization Intelligent Mechanism)


```
if response_time_predict < 1.1 * AVG(last 10 response time):
    memory.max *= 0.90
if response_time_predict > 1.5 * AVG(last 10 response time):
    memory.max *= 2
else :
    //DO NOTHING
```
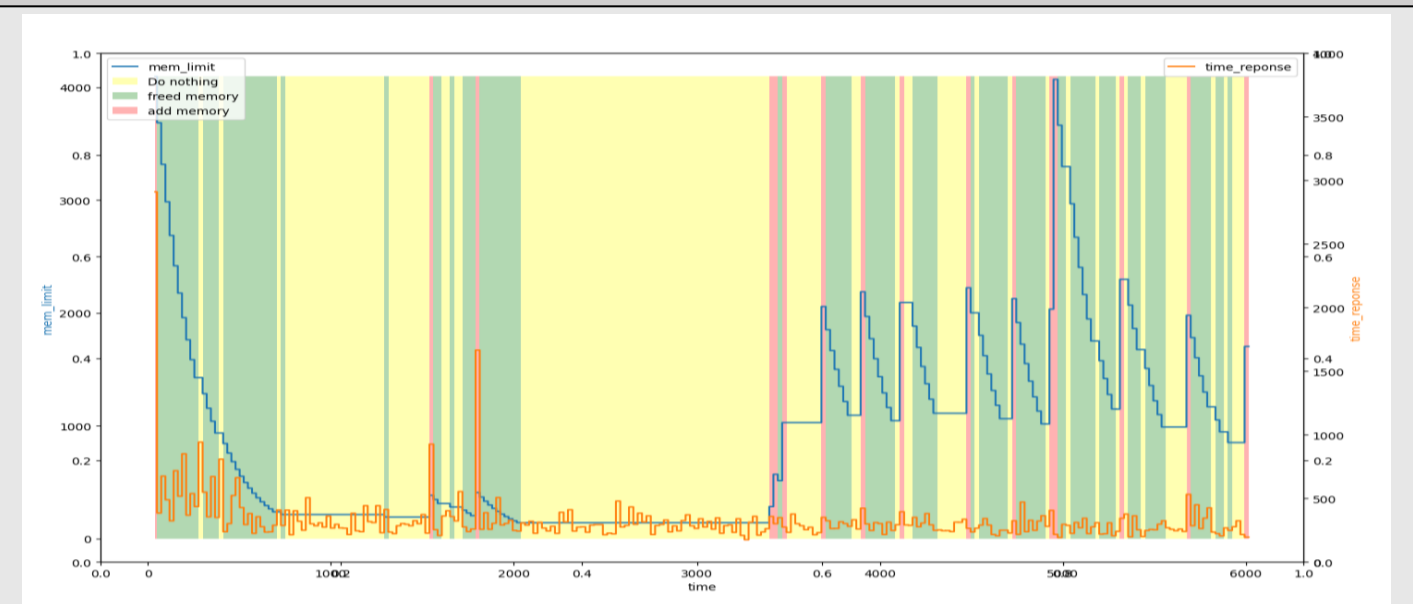
References

[1] Jui-Hao Chiang. Working set-based physical memory ballooning. 10th international Conference on Automic Computing

[Gao] Jim Gao. Machine learning applications for data center optimization. Google Publications

[2] Cheng Zhong Xu Leyi Wang Jia Rao, Xiangping Bu. Vconf: A reinforcement learning approach to virtual machines auto-configuration. ICAC '09: Proceedings of the 6th international conference on Autonomic computing, 2009.

[3] Paul Carpenter Luis A. Garrido, Rajiv Nishtala. Continuous-action reinforce-ment learning for memory allocation in virtualized servers. 2019.

[4] J´urgen Schmidhuber Sepp Hochreiter. Long short-term memory. Neural Computa-tion, 9 1997.

[5] Gilles Madi Wamb. Combiner la programmation par contraintes et l'apprentissage machine pour construire un mod`ele´ eco-´energ´etique pour petits et moyens data centers. HAL theses,