Proxmox VE Roadmap

Alexandre Derumier

Alexandre.derumier@groupe-cyllene.com

https://twitter.com/aderumier

C:YLLENE

Alexandre Derumier

- **Ingénieur système et réseau**
- **Contributeur proxmox depuis 2010**

  (intégration plugin stockage ceph/rbd , snapshots, firewall, sdn,..)
- **Formateur officiel Proxmox France depuis 2012**

  https://www.groupe-cyllene.com/fr/formations-open-source

C\yllene Lille

- **Activité : Hébergement (onprem + cloud)**
  Spécialisation clusters haute dispo && Kubernetes
  Consulting / design / accompagnement proxmox / ceph
  Formations proxmox + kubernetes

- **Equipe de 20 personnes (ingés, admins, devops,..) sur Lille.**
  400 personnes sur toute la france dans le groupe

- **1 Datacenter lille + 3 DC Paris**
  Clusters proxmox/ceph étendus triple DC Paris + PRA Lille

- **100 hyperviseurs proxmox (depuis 2010 ~ proxmox 1.9)**

- **4000 vms (pas de CT)**

- **600 TB de stockage ceph (depuis 2015)**

Coming Soon (proxmox 7.2~7.3)

Coming Soon (proxmox 7.2~7.3)

- ZFS DRAID
- User Superadmin (=~ root)
- Qemu virtio-net mtu (gui)
- Qemu free-page-reporting

Qemu : Virtio-mem (→ proxmox 7.3)

Memory Hot(un)plug

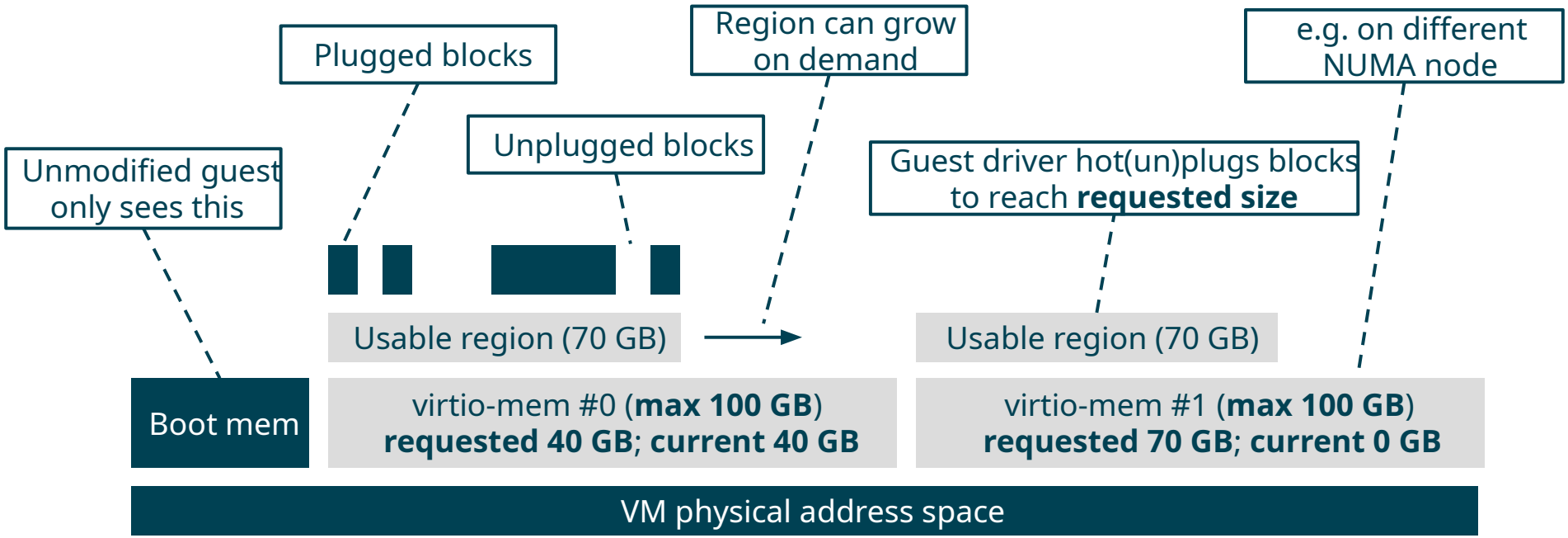
- **Memory Hotplug**
  - Ajout d’une nouvelle barette mémoire virtuelle
- **Memory Hotunplug**
  - Suppression d’une barette mémoire virtuelle
  - Le guest doit évacuer/deplacer la mémoire d’abord

Memory Hot(un)plug

Problèmes

- Enlever de la mémoire sous linux a besoin d'une **MOVABLE zone**
  - … et peux planter si une seule page ne peux pas etre déplacée
  - … et peux provoquer des problèmes de fragmentation, équilibrage numa,…
- **Limites** (ACPI slots, MMAPs, KVM memory slots, taille DIMM size …)
  - Limite la flexibilité et granuralité des tailles mémoires

virtio-mem


Plugged blocks
Region can grow on demand
e.g. on different NUMA node
Unplugged blocks
Unmodified guest only sees this
Guest driver hot(un)plugs blocks to reach **requested size**
Usable region (70 GB)
Usable region (70 GB)
Boot mem
virtio-mem #0 (**max 100 GB**) **requested 40 GB**; **current 40 GB**
virtio-mem #1 (**max 100 GB**) **requested 70 GB**; **current 0 GB**
VM physical address space

Firewall : Nftables (→ proxmox 8.0)

Iptables

```
-A FORWARD -j PVEFW-FORWARD
-A PVEFW-FORWARD -m conntrack --ctstate INVALID -j DROP
-A PVEFW-FORWARD -m conntrack --ctstate RELATED,ESTABLISHED -j ACCEPT
-A PVEFW-FORWARD -m physdev --physdev-in fwln+ --physdev-is-bridged -j PVEFW-FWBR-IN
-A PVEFW-FORWARD -m physdev --physdev-out fwln+ --physdev-is-bridged -j PVEFW-FWBR-OUT

-A PVEFW-FORWARD -m physdev --physdev-in fwln+ --physdev-is-bridged -j PVEFW-FWBR-IN
-A PVEFW-FWBR-IN -m physdev --physdev-out tap661i0 --physdev-is-bridged -j tap661i0-IN
-A PVEFW-FWBR-IN -m physdev --physdev-out tap663i0 --physdev-is-bridged -j tap663i0-IN
-A PVEFW-FWBR-IN -m physdev --physdev-out tap668i0 --physdev-is-bridged -j tap668i0-IN
.....
```

nftables

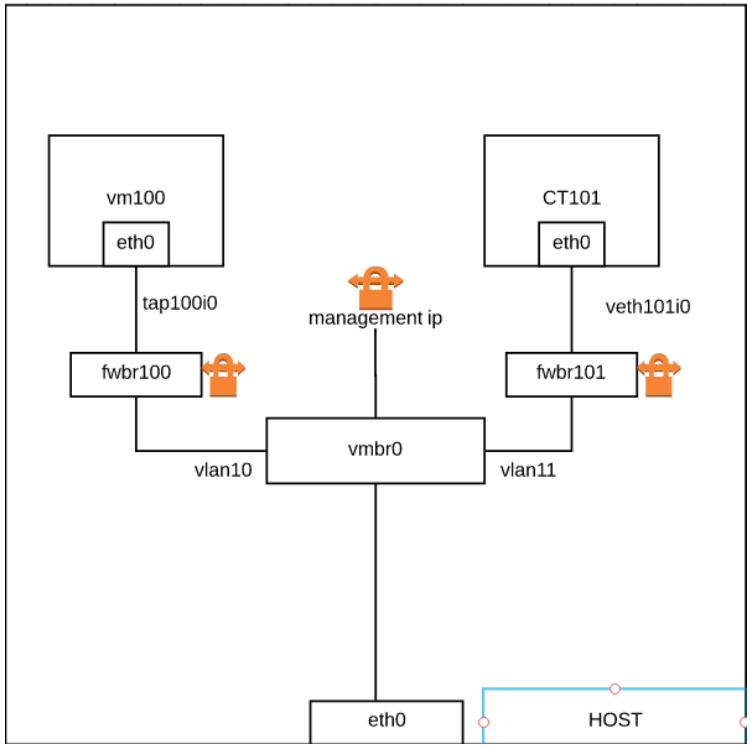
```
chain forward {
        type filter hook forward priority 0; policy accept;
        ct state established,related accept
        ct state invalid drop

        oifname vmap { tap100i0 : jump tap100i0-in , tap105i0 : jump tap105i0-in, .... }
}
```

Suppression des bridges FWBR

Problème Actuel :

Entre 2 vms sur le même serveur, les règles iptables sont évaluées 2x
(en sortie de la vm1, et en entrée la vm2)



Management Multi-cluster (→ proxmox 8)

Management Multi-cluster (→ proxmox 8)

1 Cluster Proxmox

32 nœuds maximum (limites latences corosync)

Management Multi-cluster (→ proxmox 8)

- Cross cluster authentication mechanism (en cours)
- Cross cluster live migration (en cours)
- interface management multi-cluster (proxmox 8?)

HA ressources aware (→ proxmox 7.X)

HA : Problèmes Actuels

Non prise en compte des ressources du nœud target

Présence du stockage ?
Nombre de coeurs nœud ≥ nombre de coeurs vm ?
Présence du vmbr ?
Charge cpu/ram du nœud ?
....

Algorithme actuelle :
Restart la vm sur le nœud qui a **le moins** de vms,
et si la vm ne démarre pas, essaye de restart la vm sur un autre nœud

Workaround partiel :
Créer des groupes de serveurs (stockage,nombre de coeurs,vmbrX)
Ne résoud pas le problème de charge cpu/ram

HA : Nouvelle algorithme

Vérifier si le nœud dispose des ressources, **AVANT** de restart la vm

- plus besoin de groupe en workaround
- moins de restart de vm

Prise en compte de la charge cpu/ram du nœud

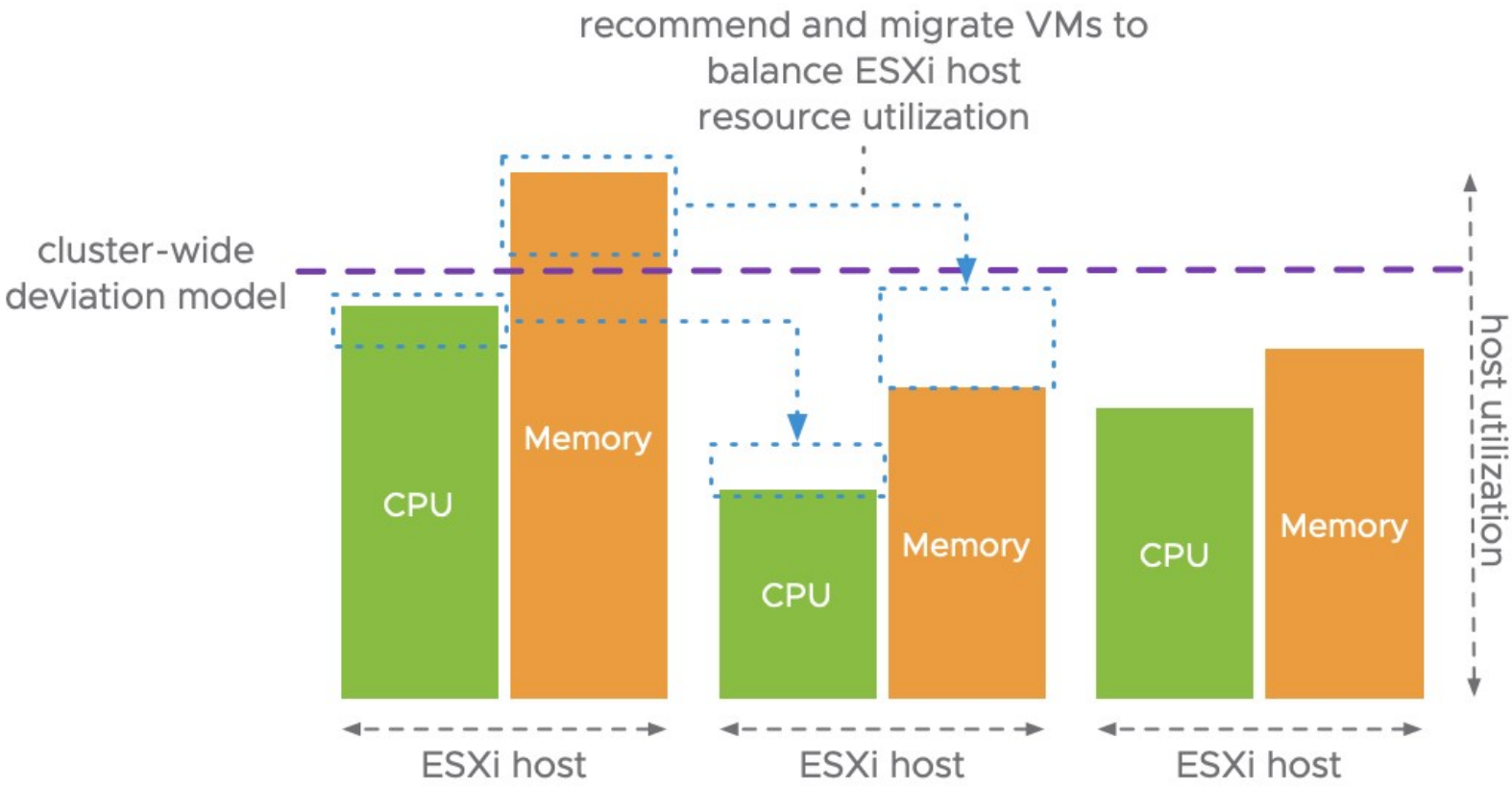
- La vm sera restart sur le nœud le moins chargé en priorité

Vm balancer (→ proxmox 7.X)

Vmware drs 1.0

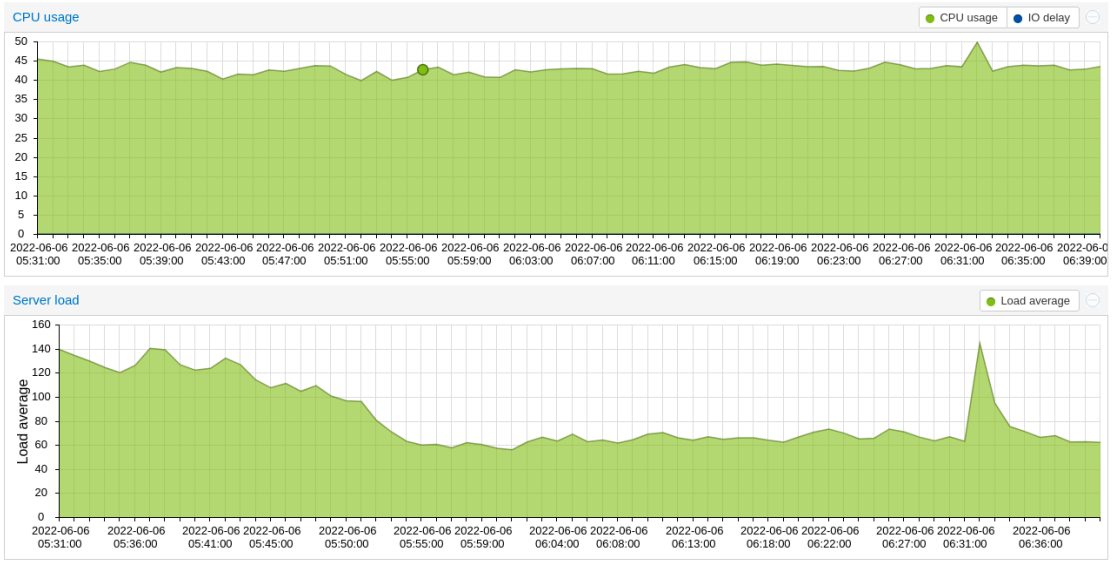
Node centric

Evacue des vms lorsque les seuils cpu/ram du host dépasse la deviation standard



Se baser sur l'utilisation du cpu n'est pas assez précis !

Ex : host 120 cores surchargé avec cpu à 40 % à 5h30 , mais pas à 6h00



Pourquoi ?

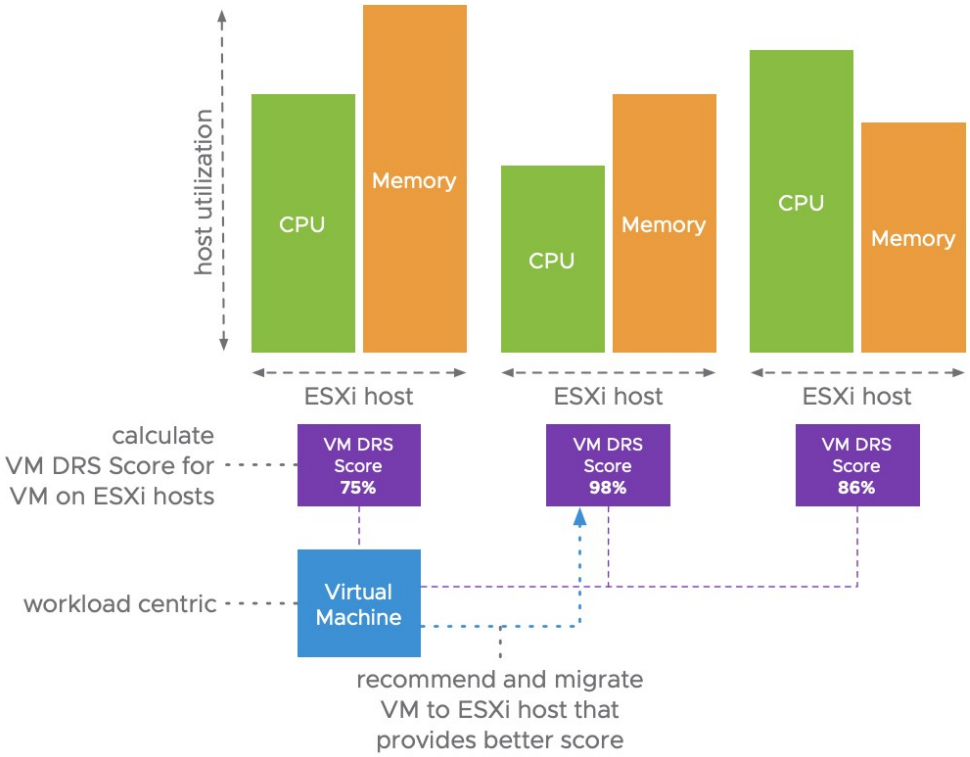
- avec beaucoup de cores, la moyenne lisse les pics ou coeurs saturés

- Si beaucoup de vms : context-switches

Vmware drs 2.0 (depuis vmware7)

vm centric

Calcul la « santé » de chaque vm (score), et migre les vm vers un nœud sur lequel elle aura un meilleur score



linux pressure stall information

```
➜  ~ cat /proc/pressure/cpu
some avg10=0.05 avg60=0.04 avg300=0.01 total=576927401
```

Percentage of the time on average every 10, 60 and 300 seconds processes were starved of CPU.

Total time in microseconds processes were starved of CPU

Informations détaillées pour chaque vm !

~ cat /sys/fs/cgroup/qemu.slice/<vmid>.scope/cpu.pressure
some avg10=4.83 avg60=3.62 avg300=3.78 total=435143207628
full avg10=1.00 avg60=1.06 avg300=1.32 total=135135391202

~ cat /sys/fs/cgroup/qemu.slice/<vmid>.scope/memory.pressure
some avg10=0.00 avg60=0.00 avg300=0.00 total=0
full avg10=0.00 avg60=0.00 avg300=0.00 total=0

~ cat /sys/fs/cgroup/qemu.slice/<vmid>.scope/io.pressure
some avg10=0.00 avg60=0.00 avg300=0.00 total=644402361
full avg10=0.00 avg60=0.00 avg300=0.00 total=453531269

Heuristique

- Lister les vms qui ont une pressure cpu/ram/network elevée

- calculer le score de chaque vm avec
    - le plus gros cpu pressure
    - le plus grosse memory pressure
    - le plus petit %cpu usage
    - le plus petit %mem usage
    - la plus grosse affinité
    ...

-choisir la vm qui a le score le plus haut

Calculer un score de chaque noeud + load de la vm à migrer.
- la plus grand priorité dans un groupe HA
- le plus bas cpu pressure
- le plus petit %cpu usage
- le plus petit %mem usage

-choisir le nœud qui a le score le plus elevé

Calcul du score : TOPSIS

Selection of the Best

Alternative	Attribute Or Criteria →	Price or Cost	Storage Space	Camera	Looks
	Mobile 1	250 $	16 GB	12 MP	Excellent
	Mobile 2	200 $	16 GB	8 MP	Average
	Mobile 3	300 $	32 GB	16MP	Good
	Mobile 4	275 $	32 GB	8MP	Good
	Mobile 5	225 $	16 GB	16 MP	Below Average

TOPSIS : method multi-criteria decision making

Algorithm 2 Pseudo code of TOPSIS Method

Input: Cloud Service Alternatives(a_{ij}), w_j= Weight of each QoS criteria.

Output: Best M.

while $(M \neq NULL)$ **do**

Create a Decision Matrix D; // Based on equation 1

for a_{ij} in D **do**

$r_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{m} a_{ij}^2}}$ // Compute normalized decision matrix.

end for

$v_{ij} = w_j.r_{ij}$ // Weighted Normalized decision matrix, w_j is weights of each criteria calculated by Best Worst Method.

if (j $\in J$) **then**

$A^* = max(x_{ij}); A^- = min(x_{ij})$

else if j $\in J$ then

$A^* = min(x_{ij}); A^- = min(x_{ij})$

end if

for $J \in j \in C_j$ **do**

$d_i^+ = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^+)^2}$ // Calculate the Separation measure of positive ideal solution(d^+)

$d_i^- = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^-)^2}$ // Calculate the Separation measure of negative ideal solution(d^-)

end for

for each x_{ij} in D **do**

$CC_i = \frac{d_i^-}{d_i^- + d_i^+}$ // Calculate the relative closeness coefficient.

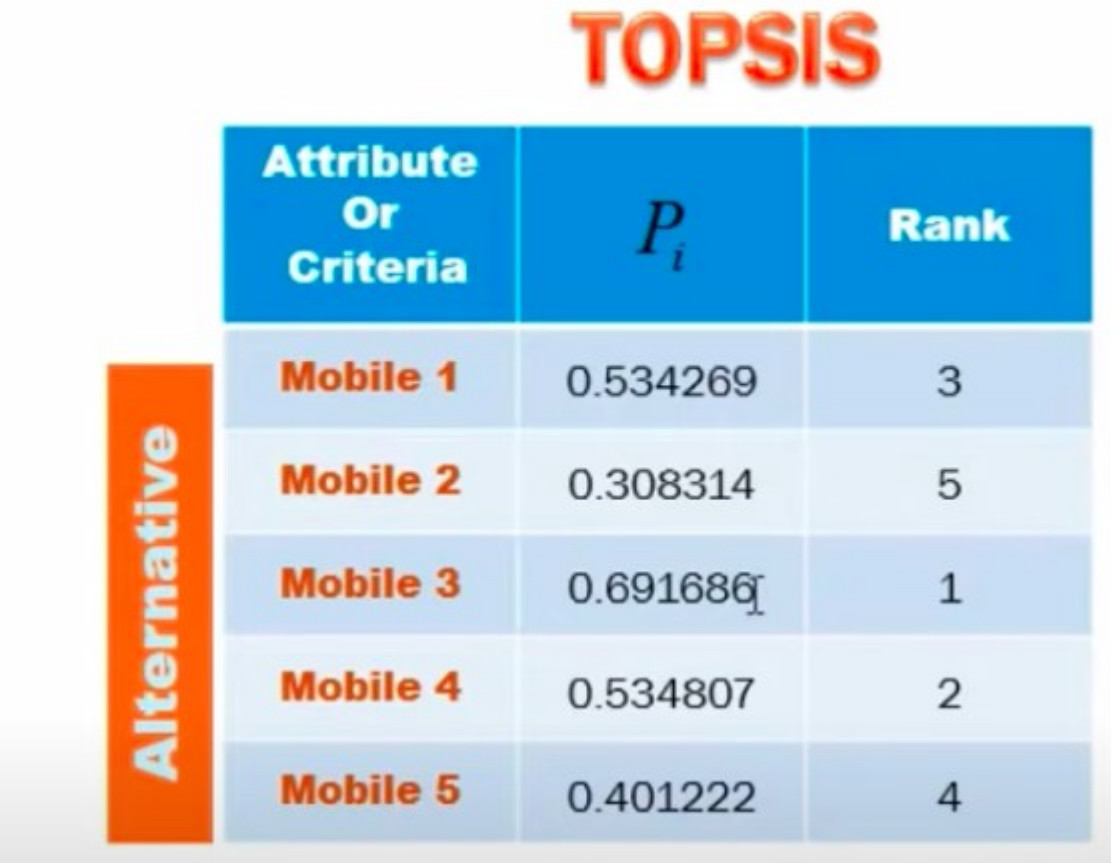
end for

Rank the Cloud Service Alternatives based on CC_i.

// The larger indexed value is considered as the optimal cloud service alternatives.

end while

Calcul du score : TOPSIS



Attribute Or Criteria	P_i	Rank
Mobile 1	0.534269	3
Mobile 2	0.308314	5
Mobile 3	0.691686	1
Mobile 4	0.534807	2
Mobile 5	0.401222	4

SDN (software defined network)

SDN (software defined Network)

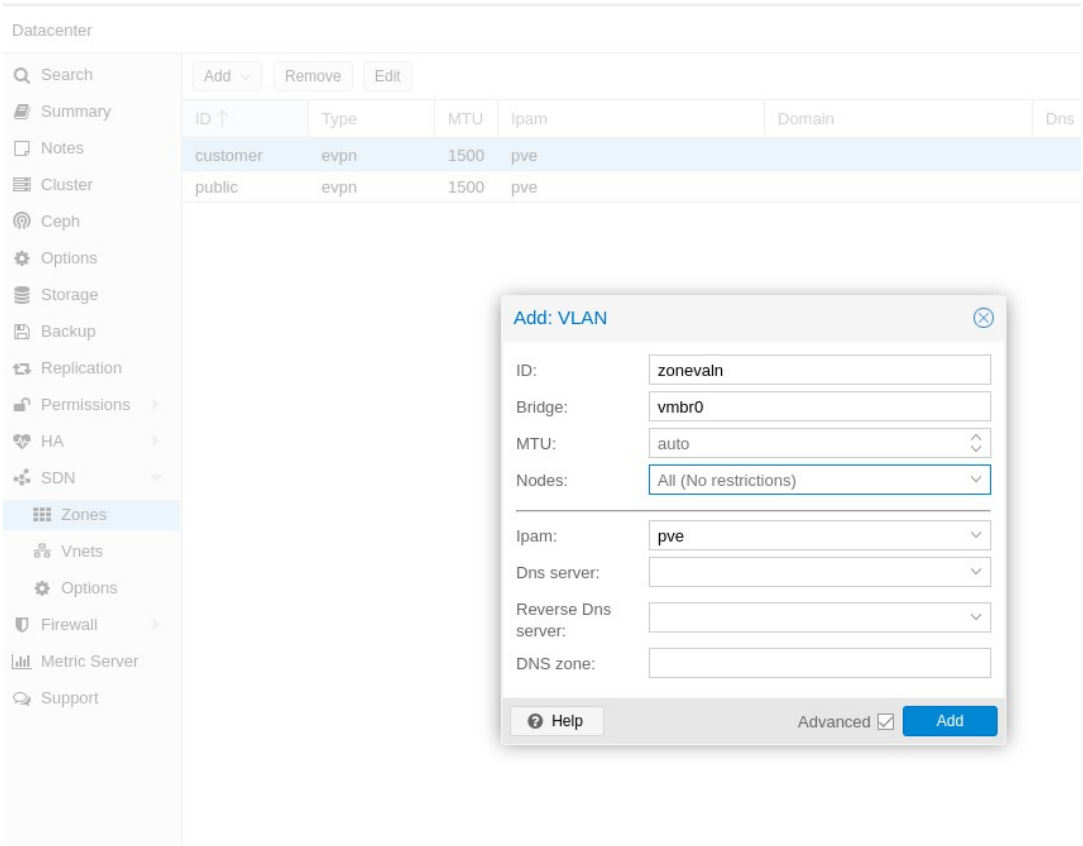
Déjà disponible en bêta :

apt install libpve-network

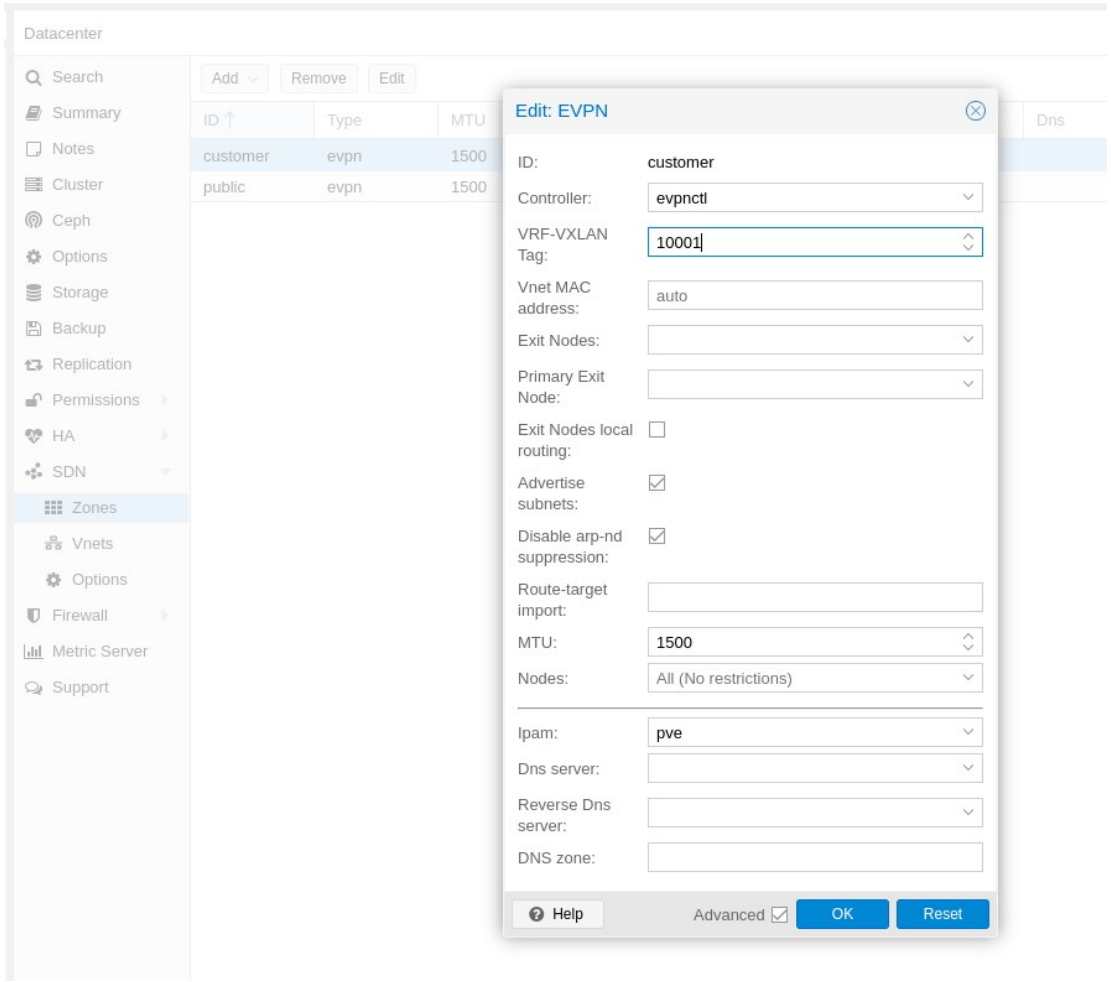
Configuration des réseaux des vms au niveau datacenter

- Vlan
- Qinq (vlans stackés)
- Vxlan
- BGP-Evpn

Ex : zone vlan

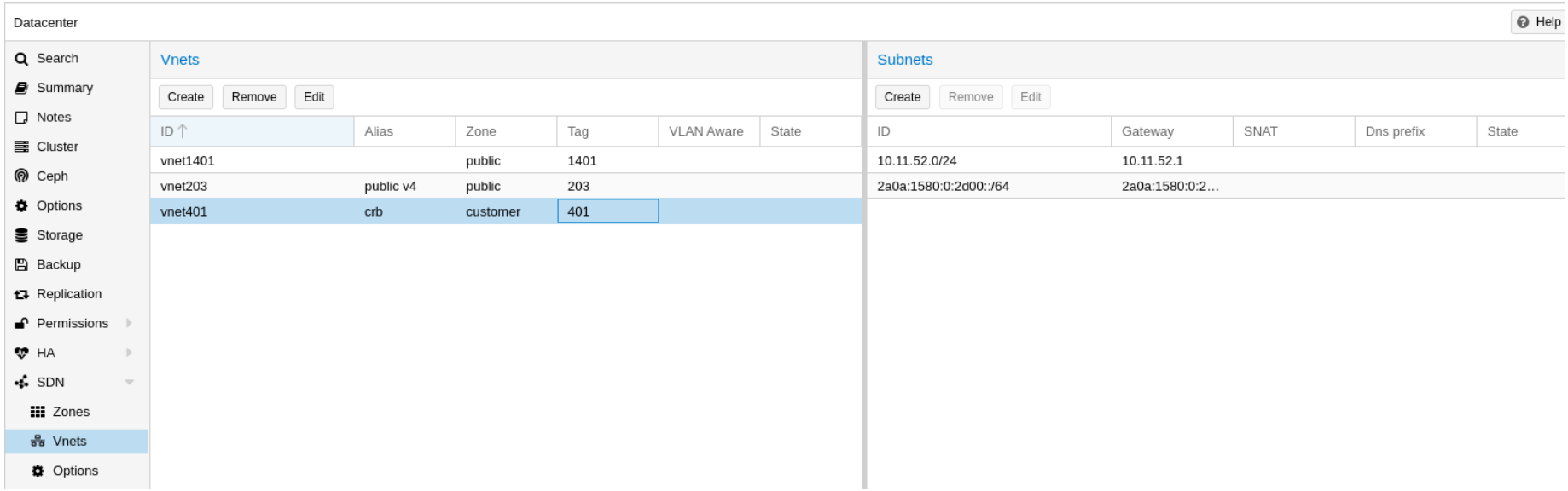

Datacenter
Search
Summary
Notes
Cluster
Ceph
Options
Storage
Backup
Replication
Permissions
HA
SDN
Zones
Vnets
Options
Firewall
Metric Server
Support
Add
Remove
Edit
ID
Type
MTU
Ipam
Domain
Dns
customer
evpn
1500
pve
public
evpn
1500
pve
Add: VLAN
ID:
zonevaln
Bridge:
vmbr0
MTU:
auto
Nodes:
All (No restrictions)
Ipam:
pve
Dns server:
Reverse Dns server:
DNS zone:
Help
Advanced
Add

Ex : zone BGP-EVPN



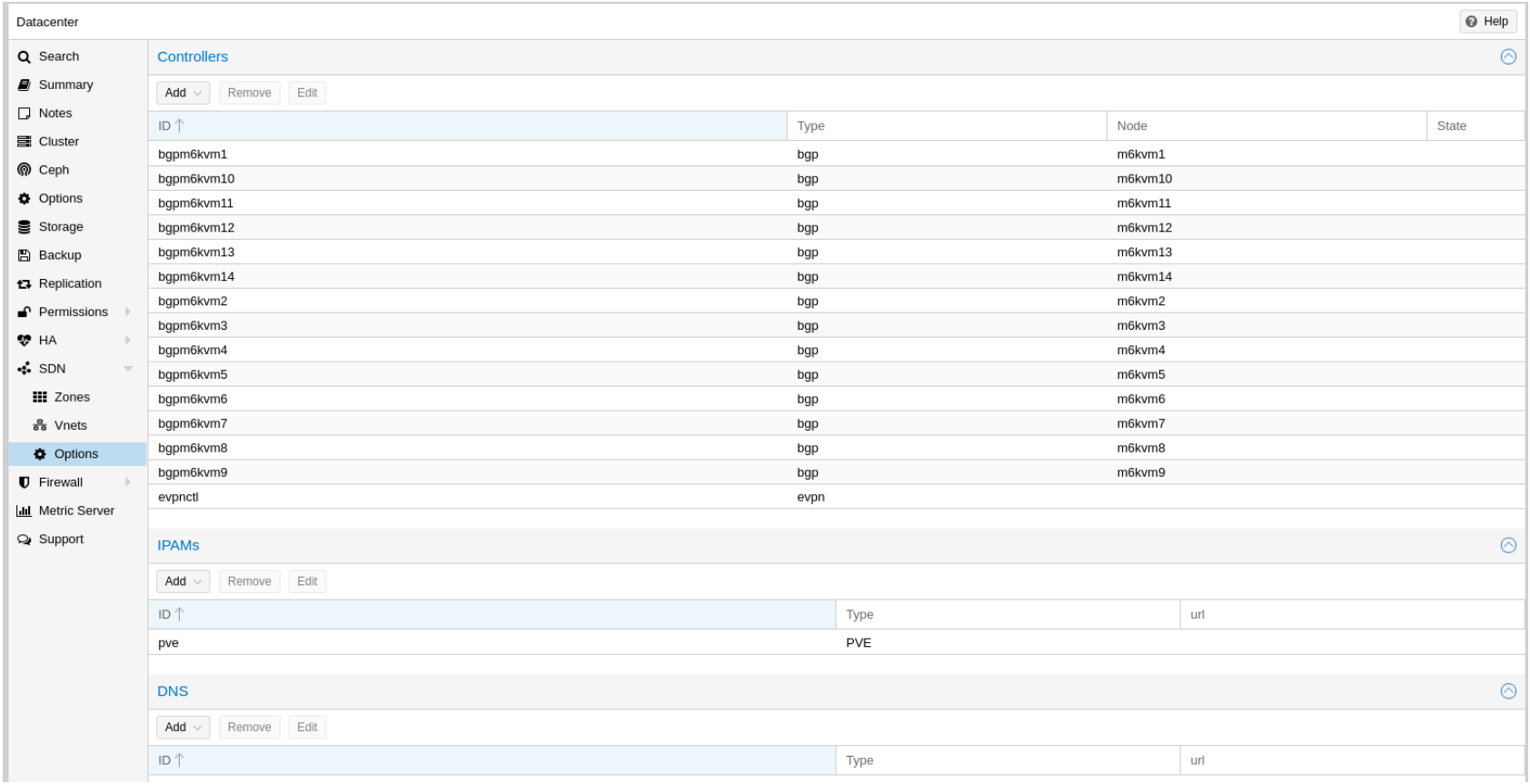
Vnets

- Definition des réseaux + tag vlan/vxlan
- Déclaration des subnets



Options

- Controllers (EVPN, BGP)
- IPAMS : netbox, phpipam
- DNS : Powerdns



Features à venir→ proxmox 8.X

vms/containers

Attribution des ips automatiques via ipam

- netbox
- Phpipam
- ...

Enregistement des ips dans votre dns

- powerdns
- ...

Services centralisés

DHCP, loadbalancer, nat, ...

Contribuer

https://pve.proxmox.com/wiki/Developer_Documentation

Traduction francaise

https://git.proxmox.com/?p=proxmox-i18n.git

Mailing lists

pve-devel@lists.proxmox.com

pve-users@lists.proxmox.com

Bug report && feature requests

https://bugzilla.proxmox.com/

Nous contacter ?

Téléphone - +33 1 41 19 40 40
Email - contact@groupe-cyllene.com

Paris - Nanterre - Issy-Les-Moulineaux - Montigny - Troyes - Montbéliard - Saint Brieuc - Lyon - Lille

@Cyllene @Groupe_Cylllene