



Research Centre in
Digitalization and Intelligent Robotics



Applied Digital Transformation
Laboratory

Instituto Politécnico de Viana do Castelo

Virtualização Open-Source ao serviço da I&D: o caso do IPB

Talk for Us

Webinar Conjunto CeDRI/2AI/ADiT-LAB

23 de Junho de 2023

José Rufino (CeDRI/ESTiG/IPB), rufino@ipb.pt

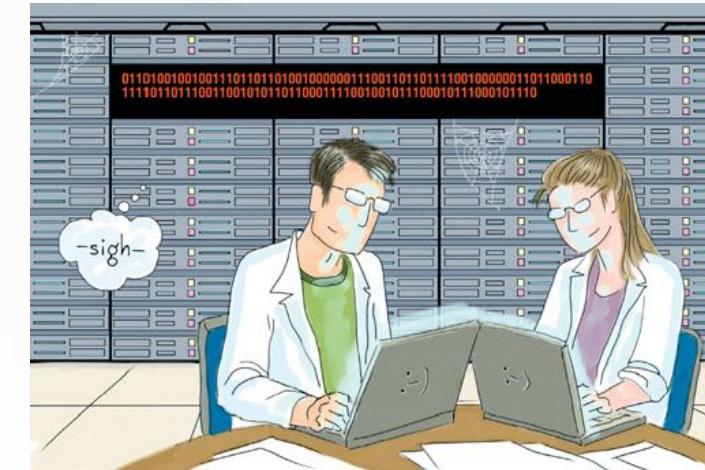
Roteiro

- **Computação ao Serviço da Ciência**
- **Plataformas de Computação (Avançada)**
- **Virtualização de Sistemas**
- **Cluster do CeDRI**
 - componentes e organização
 - tecnologias de exploração
 - casos de uso e testemunhos
 - conclusão e perspetivas



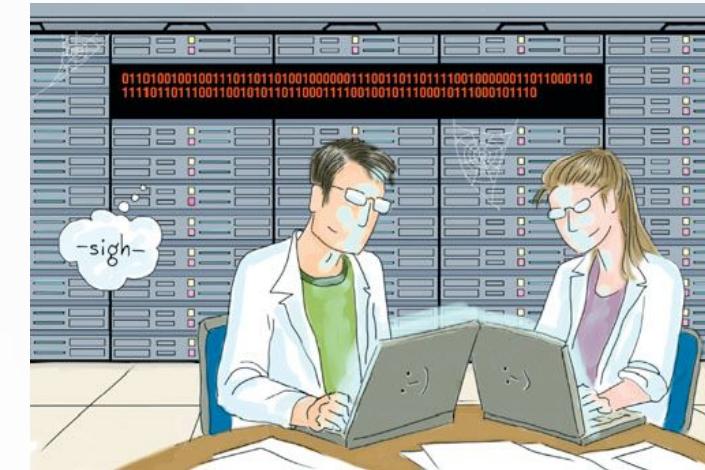
Computação ao Serviço da Ciência

- **Modelação, Simulação, Previsão**
 - sistemas e processos reais ou teorizados
 - aceleração de testes de hipóteses ou novas teorias
- **Aquisição de Dados, Medição**
 - adquirir mais dados, em menos tempo; medir com mais precisão
- **Análise e Visualização de Dados**
 - extrair informação, encontrar padrões ocultos
 - permitir a visualização, visualizar em novas perspetivas
- **Mais Precisão, Mais Correção**
 - representar dados com mais precisão, cálculos mais exatos



Computação ao Serviço da Ciência

- **Automação, Reprodutibilidade**
 - repetir tarefas complexas ou fastidiosas
 - garantir a reproduzibilidade dos experimentos
- **Resolução de Problemas Complexos**
 - computacionalmente (muito) exigentes e/ou
 - com elevados volumes de dados associados
- **Acelerar a Descoberta, Facilitar a Inovação**
 - força bruta, inteligência artificial, ...
 - mais rápido, mais seguro, mais barato
 - trabalho colaborativo, formatos de dados e aplicações abertas



Plataformas de Computação (Avançada)

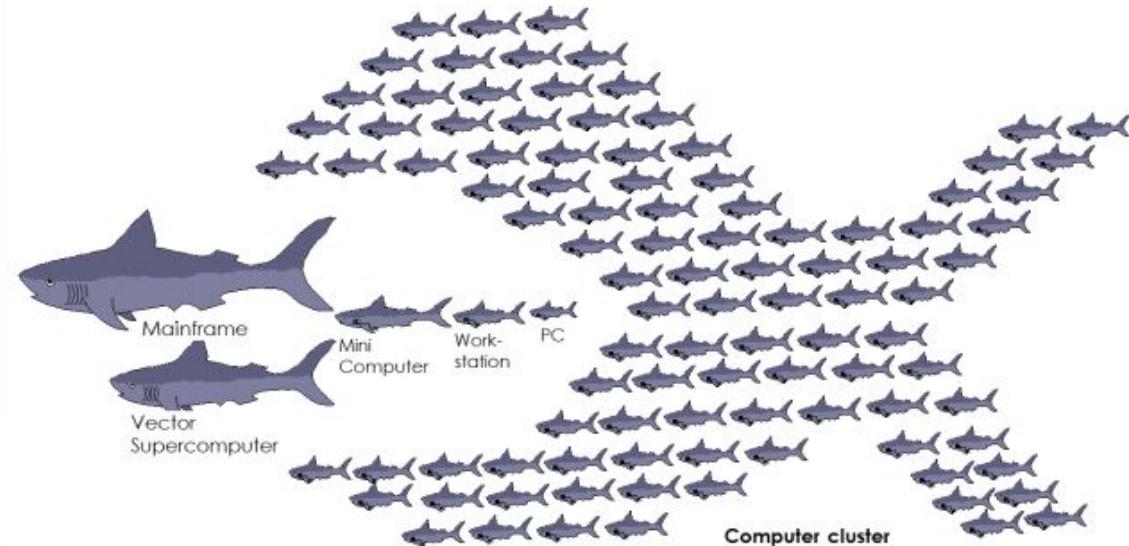
- o **computador** tornou-se a **indispensável** aos investigadores, complementando ou substituindo ferramentas tradicionais
- muitos problemas científicos são facilmente resolúveis com a capacidade de cálculo dos **computadores pessoais**
- mas a escala/exigência de certos problemas implicam o uso de abordagens computacionais mais avançadas, como:
 - **cluster** computing
 - **grid** computing
 - **cloud** computing



Plataformas de Computação (Avançada)

- **Cluster:**

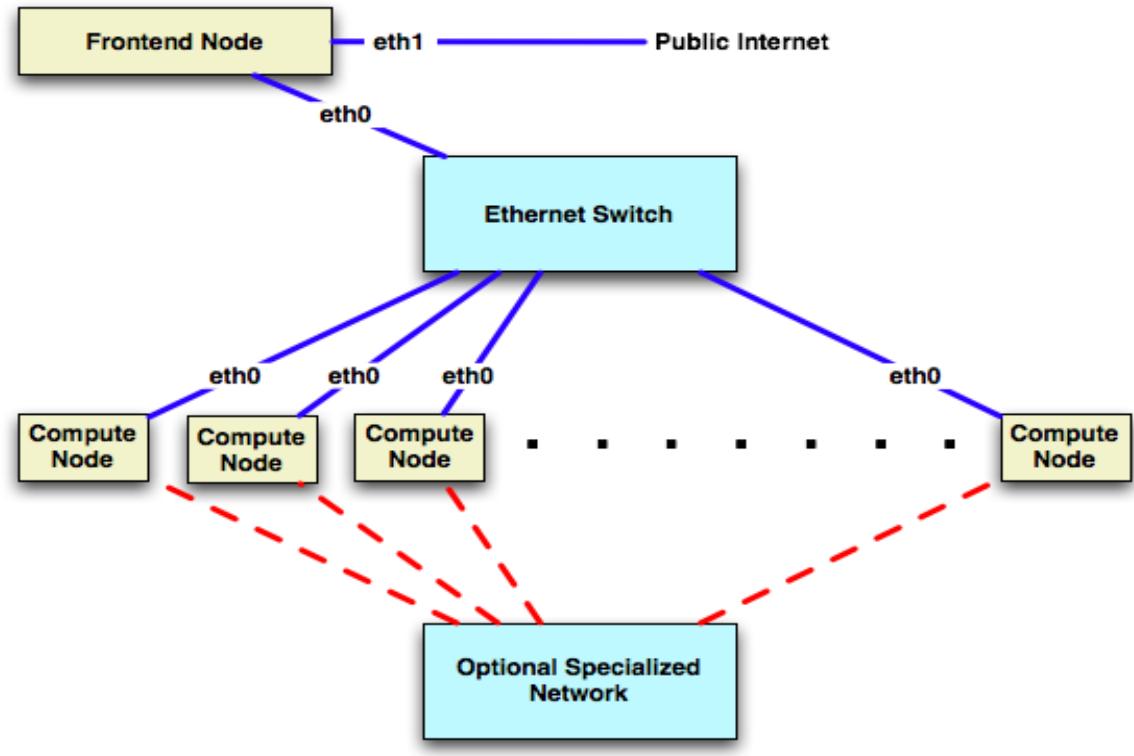
- agregado de nós de computação, tipicamente homogêneos, interligados em rede local
- propósitos
 - balanceamento de carga, alta disponibilidade (HA)
 - processamento paralelo (**cluster HPC**)
 - alojamento de máquinas virtuais (**cluster de virtualização**)
- vantagens vs sistemas isolados
 - mais capacidade (processamento/armazenamento), mais fiabilidade



Plataformas de Computação (Avançada)

- **Cluster HPC:**

- topologia típica:



- modelo de exploração:

- acesso remoto (SSH) a *frontend* com as contas de utilizadores
 - submissão de *jobs* a sistema de gestão de filas de trabalho

Plataformas de Computação (Avançada)

- **Grid:**

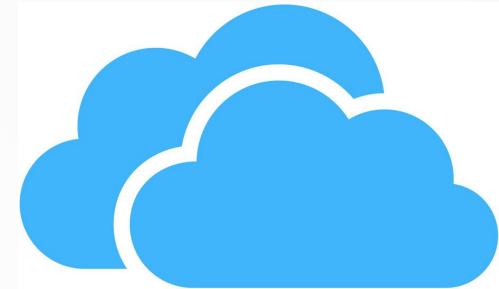
- supercomputador virtual, assente numa **federação de clusters** autónomos, heterogéneos entre si e geograficamente dispersos
- requer middleware de gestão de recursos, dados, segurança
- *jobs* submetidos podem executar em qualquer cluster da grid
- precursor do modelo de computação em nuvem
- exemplo:
 - CERN's Worldwide LHC Computing Grid (WLCG)



Plataformas de Computação (Avançada)

- **Cloud:**

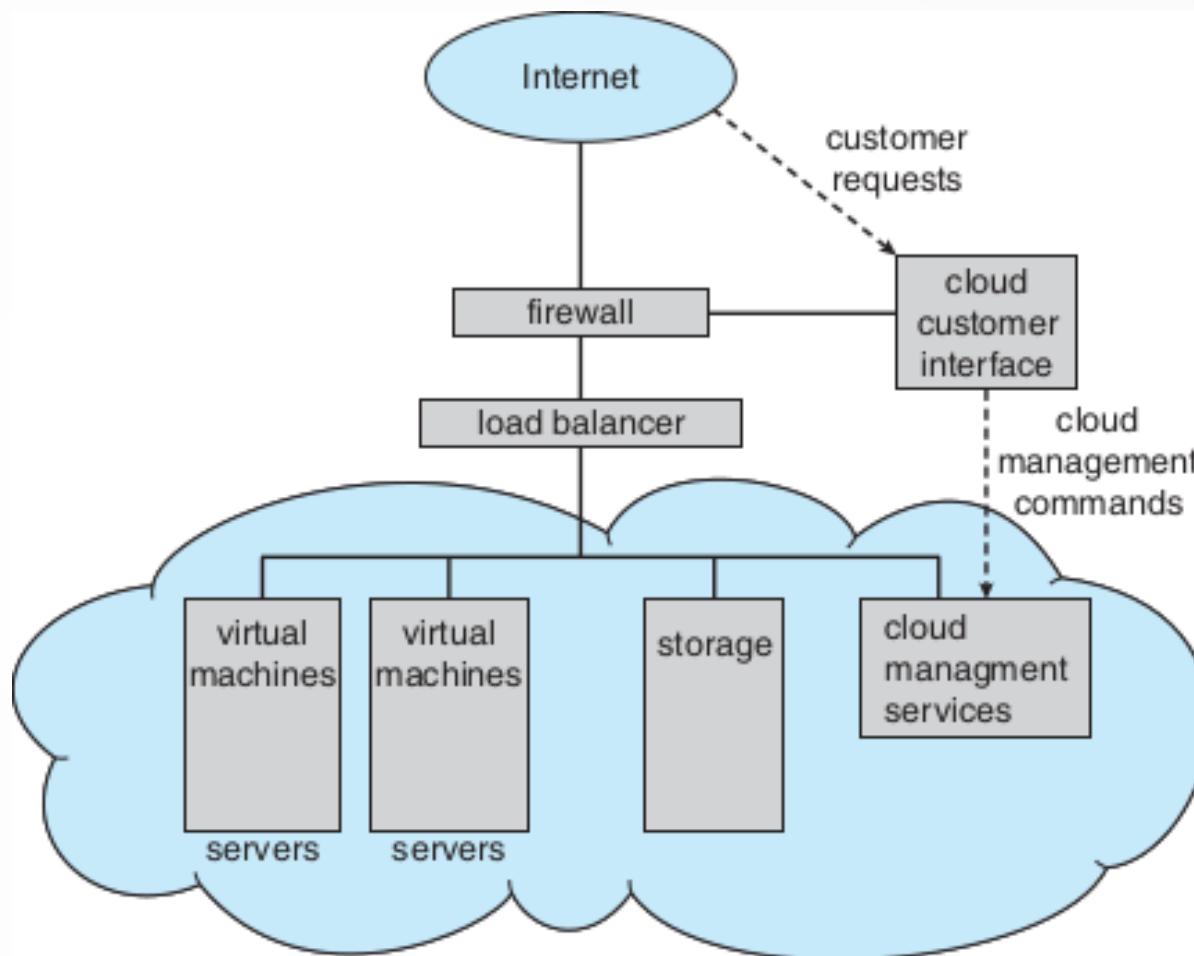
- agrupa recursos de computação e armaz., de 1 ou + datacenters, localizados em zonas geográficas tipicamente diferentes
- oferece Software/Platform/Infrastructure as a Service (**SaaS/PaaS/IaaS**)
- **acesso ubíquo** (via rede) e **transparente**, aos serviços oferecidos
- acesso **self-service** de recursos **on-demand** com **elasticidade**
- modelo de exploração com **tudo taxável** (comp./armaz./rede)
- **pública** (AWS, Google, Microsoft Azure) vs **privada*** vs **híbrida**
- permite executar aplicações HPC de forma cómoda e eficiente
- ao mais baixo nível assenta em **tecnologias de virtualização**



Plataformas de Computação (Avançada)

- **Cloud:**

- topologia típica:



- acesso: transparente (via apps), interfaces CLI/Web, APIs

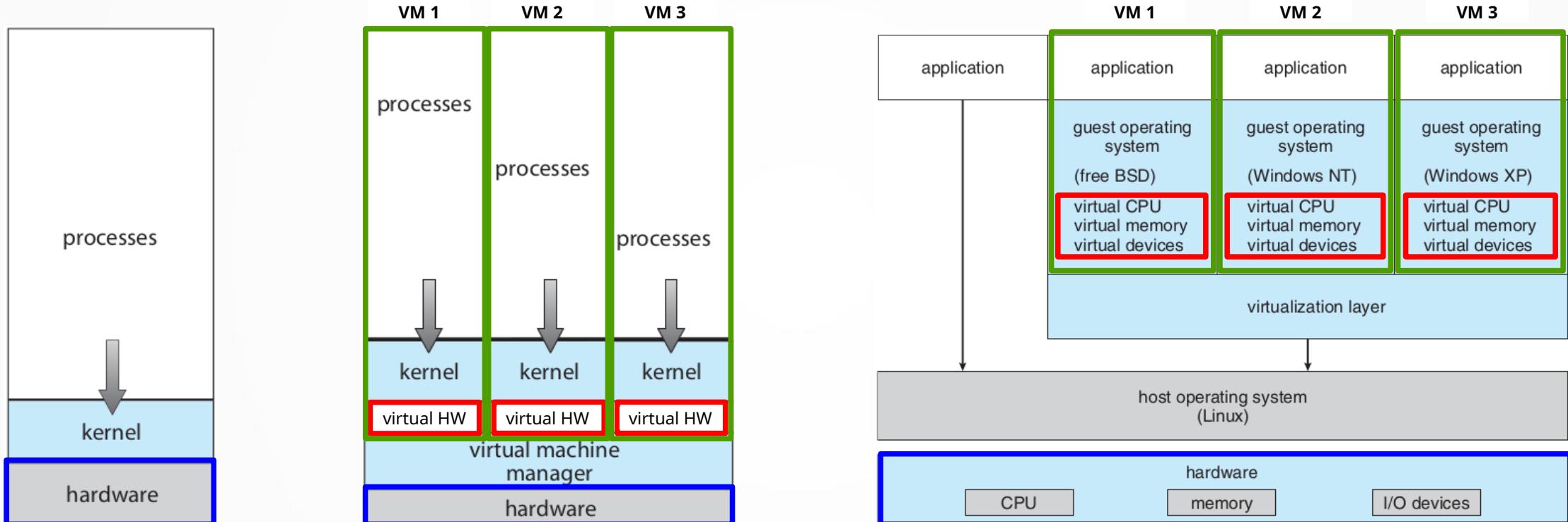
Virtualização de Sistemas

- **Conceitos Básicos**

- sistema computacional: hardware + sist. operativo + aplicações
- **virtualização de sistemas:**
 - ilusão de “vários computadores dentro de um só computador”
 - um sistema **hospedeiro** acomoda vários sistemas **hóspedes**
- sistema **hospedeiro (host)**: hardware **físico** + sist. operativo (SO) de virtualização (virt. tipo 1) ou SO de uso genérico (virt. tipo 2)
- sistema **hóspede (guest)** ou **Máquina Virtual (VM)**:
 - hardware **virtual** + SO de uso genérico + aplicações

Virtualização de Sistemas

- **Tipos de Virtualização**



Sem virtualização

**Virtualização Tipo 1
(nativa / bare-metal)**

- VMware ESXi
- Microsoft Hyper-V
- Linux KVM, Xen

Virtualização Tipo 2 (hosted)

- VMWare Workstation
- Oracle VirtualBox
- Parallels Desktop

Virtualização de Sistemas

- **Vantagens e Riscos**

(+) consolidação de sistemas: melhor rentabilização do HW, diminuição de custos energéticos, facilidade de gestão

(+) flexibilidade: criação/destruição, reconfiguração, mobilidade

(+) mobilidade das MVs: facilita load balancing, HA, backups

(+) rollback facilitado: snapshots, recuperação de backups

(-) alguma perda de desempenho (embora marginal) por MV

(-) risco de perda de vários sistemas (MVs) de uma só vez

(-) requer investimento adicional para um bom desempenho

Cluster do CeDRI - Génese

- **Projeto i4.TMAD**

“Promoção da Indústria 4.0 na Região de Trás-os-Montes e Alto Douro”

- 2016: candidatura
 - Bancada demo #2:
Plataforma de análise inteligente de grandes quantidades de dados →
 - Contexto adequado à aquisição de um cluster de virtualização
- 2017: aprovação do projeto
- 2018: aquisição, início da exploração



Cluster do CeDRI - Hardware

- **HW 2018**

- **9 nós de virtualização** (8x2U + 1x4U)

- 2x CPU **AMD EPYC Zen1 7351** (2x 16 núcleos, 2.4/2.9 GHz)
 - 256 GB RAM DDR4 ECCR 2666MHz
 - 1 SSD PCIe3 U.2 2 TB (2U), 1 SSD PCIe3 M.2 2 TB (4U)
 - 4x Ethernet 1Gbps, 2x Ethernet 10Gps
 - 8 slots SAS 3.5", 4 slots NVMe U.2, 2 slots GPU FH 16x
 - 1 GPU NVIDIA Titan V 12 GB RAM HBM2 (nó 4U)

- **1 nó de armazenamento** (2U)

- CPU **Intel Xeon Skylake Silver 4112** (4 núcleos, 2.6/3.0 GHz)
 - 128 GB RAM DDR4 ECCR 2400MHz
 - 12 HDs SAS 12Gbps 8TB
 - 1 SSD PCIe3 Optane 280 GB (write cache)
 - 2x Ethernet 10Gbps



Cluster do CeDRI - Hardware

- HW 2018
 - **capacidade agregada**
 - Freq. CPU (turbo): 835,2 GHz; Núcleos: 288 núcleos
 - RAM: 2304 GB; SSD: 18 TB (raw); HD: 96 TB (raw)
 - **expansibilidade:**
 - RAM: + 2304 GB ; SSD: + 24 SSDs U.2; HDs: + 72 HDs
 - GPUs: + 17 GPUs Full-Height (FH)

Principais opções de desenho:

- maximizar número de nós em função da verba disponível
- maximizar flexibilidade e expansibilidade das boards e chassis
- rede 10G (switchs já existentes (CCOM); 100G/Infiniband no futuro)
- apostar na nova arquitetura Zen1 da AMD (*max. performance/dollar*)



junho 2018, CCOM, IPB

Cluster do CeDRI - Hardware

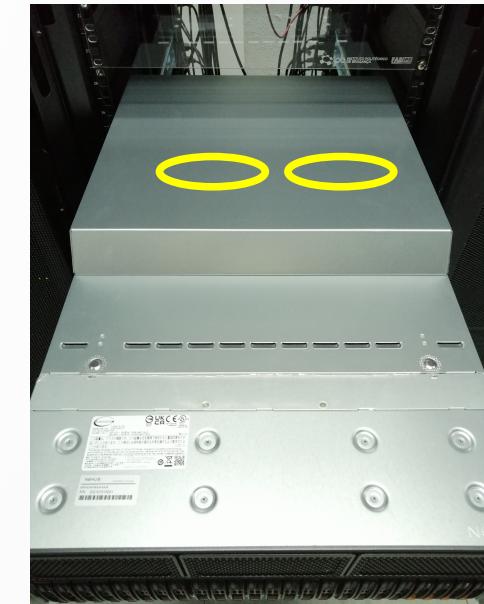
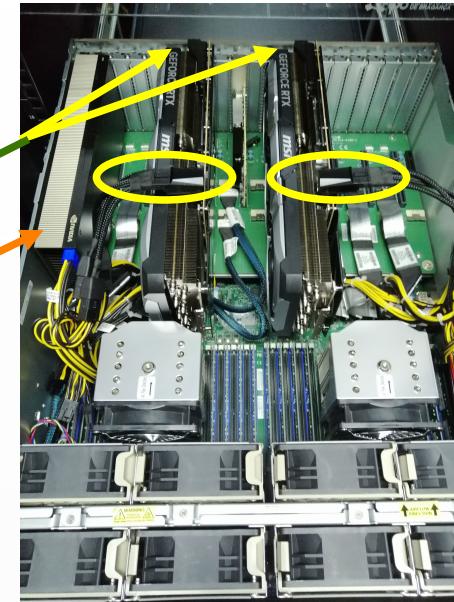
- **HW 2023 (2018 + Reforços)**

- **18** nós virt. (9 + **9**)
- **2** nós armaz. (1 + **1**)
- **628** cores (288 + **340**); 2x threads
- **2103,6** GHz (835,2 + **1268,4**)
- **5376** GB RAM (2304 + **3072**)
- **78,08** TB de SSD (18 + **60,08**)
- **96** TB de HD (96 + **0**) —————→ **Aquisição em curso: 224 TB. Após aquisição: 320 TB**
- **19** GPUs (1 + **18**)
- **rede 100 Gbps** (rede 10 Gbps já só usada em casos especiais)



Cluster do CeDRI

- HW 2023 (100G, GPUs)
 - rede Ethernet 100Gbps:
 - Mellanox ConnectX-5 por nó
 - Cisco Nexus 9336C-FX2 40/100 Gbps
 - GPUs NVIDIA
 - 8x 1060 3/6 GB
 - 1x 1070 8 GB
 - 4x 2080 Ti 11 GB
 - 1x Titan V 12 GB
 - 1x 3090 24 GB
 - 2x 4090 24 GB
 - 1x A100 80 GB;
particionável (licença vGRID)



Nó de Machine Learning

Cluster do CeDRI - Hardware

- HW 2023 (totais por tipo de nó)
 - nós de virtualização

arquitetura, CPU	nós	CPUs	núcleos	freq. núcleos (Hz)		RAM (GBytes)	armazenamento local (TBytes)		GPUs
				base	turbo		SSD	HD	
AMD Zen1 Epyc, 7531	9	18	288	691,2	835,2	2304	18		9
AMD Zen2 Epyc, 7452	2	4	128	300,8	428,8	1024			4
AMD Zen2 Ryzen 3970X	2	2	64	236,8	236,8	512	4		2
AMD Zen3 Epyc, 7443	1	2	48	136,8	192,0	512			3
Intel Xeon Skylake, W-2195	2	2	36	82,8	154,8	512	2		
Intel Xeon Icelake, W-3365	2	2	64	172,8	256,0	512	8		
TOTAL	18	30	628	1621,2	2103,6	5376	32		18



- nós de armazenamento

arquitetura	nós	CPUs	núcleos	freq. CPU (Hz)		RAM (GBytes)	armazenamento local (TBytes)		
				base	turbo		SSD	HD	
Intel Xeon Skylake	1	1	4	10,4	12,0	128	0,28	96	
AMD Zen2 Epyc	1	2	32	96,0	105,6	256	46,08		
TOTAL	2	3	36	106,4	117,6	384	46,4	96,0	



- Some SSDs allocated to virt. nodes

Cluster do CeDRI - Hardware

- HW 2023 (especificações por tipo de nó)
 - nós de virtualização

arquitetura	CPUs / nó	núcleos / CPU	núcleos / nó	freq. / núcleo (Hz)		RAM / nó (GBytes)	armazenamento / nó (TBytes)		NICs / nó		GPUs / nó	
				base	turbo		SSD	HD	10 Gbps	100 Gbps	passthrough	vGRID
AMD Zen1 Epyc	2	16	32	2,4	2,9	256	2		2	1	1	
AMD Zen2 Epyc	2	32	64	2,35	3,35	512			2	1	2	
AMD Zen2 Threadripper	1	32	32	3,7	3,7	256	2		2	1	1	
AMD Zen3 Epyc	2	24	48	2,85	4	512			2	1	2	1
Intel Skylake Xeon	1	18	18	2,3	4,3	256	1		2	1		
Intel Icelake Xeon	1	32	32	2,7	4	256	4		2	1		

 - Low freq. nodes (VM consolidation)
 - High freq. nodes (HPC VMs)

- nós de armazenamento

arquitetura	CPUs / nó	núcleos / CPU	núcleos / nó	freq. / núcleo (Hz)		RAM / nó (GBytes)	armazenamento / nó (TBytes)		NICs / nó	
				base	turbo		SSD	HD	10 Gbps	100 Gbps
Intel Skylake Xeon	1	4	4	2,6	3	128	0,28	96	2	1
AMD Zen2 Epyc	2	16	32	3	3,30	256	46,08		4	1

(*) - Write Cache

■ Some SSDs allocated to virt. nodes

Cluster do CeDRI - Tecnologias de Exploração

- **Virtualização Bare-Metal**

- o cluster do CeDRI é um **cluster de virtualização**: aloja **máquina virtuais**
- **benefícios da adoção de virtualização**
 - **rácio custo/benefício atrativo** (x MVs custam menos que x PCs)
 - permite **operação desconectada e contínua** das MVs (os utilizadores podem deixar aplicações a correr e desligar-se da máquina virtual)
 - **reconfiguração expedita de recursos** (CPU, RAM, disco, rede, GPUs, ...)
 - **fiabilidade acrescida** (alta disponibilidade, backups automáticos)
 - compatível com acesso a **recursos especiais** (únicos ou partilhados)
 - **co-processadores** (e.g., GPUs) p/ aceleração de processamento
 - **servidores de armazenamento** (datasets, discos de rede, ...)

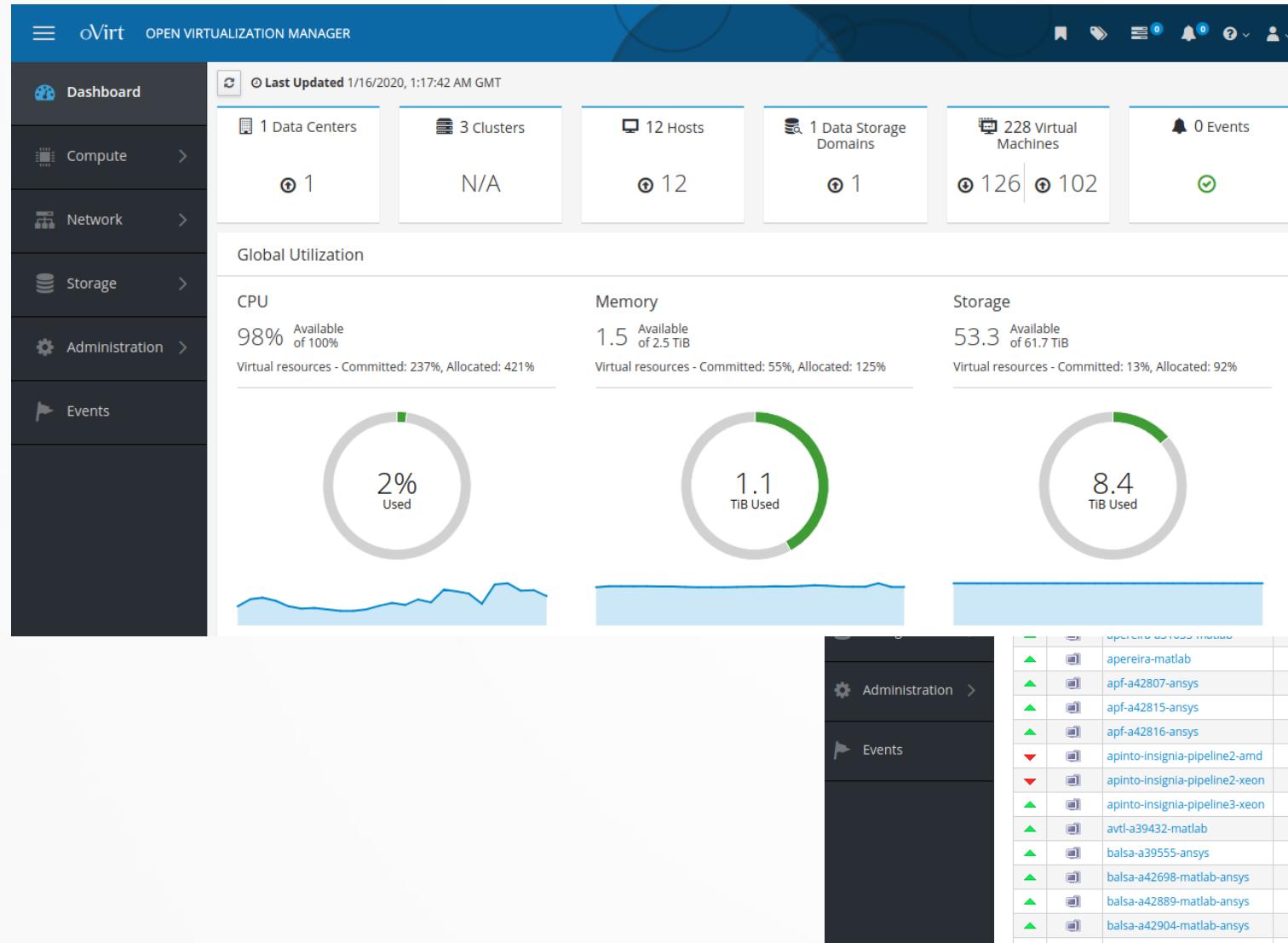
Cluster do CeDRI - Tecnologias de Exploração

- **Virtualização Bare-Metal**

- **plataformas de virtualização usadas no cluster do CeDRI**
 - **requisitos:** open-source, gratuitas, estáveis, desenvolvimento ativo, baseadas em KVM (soluções baseadas em Xen demonstraram desempenho inferior)
 - **inicialmente:** **oVirt** (<https://www.ovirt.org/>); sobre CentOS; (+) portal self-service; (-) armaz. partilhado, sem backups nativos, **descontinuado**
 - **> 12/2021: Proxmox** (<https://www.proxmox.com/en/>); sobre Debian; (-) sem portal self-service; (+) armaz. flexível, backups nativos, **em desenvolvimento**
- plataformas de computação em nuvem como alternativa
 - + complexas, exigem + recursos humanos, sobre-dimensionadas (?)
 - solução para um problema que, em 2018, não existia (e agora ?)

Cluster do CeDRI - Tecnologias de Exploração

- Plataforma de Virtualização oVirt



Dashboard de Gestão

This screenshot shows a detailed view of hosts within a cluster. The top navigation bar is identical to the dashboard. Below it, a toolbar provides actions like Remove, Run, Suspend, Shutdown, Reboot, Console, Migrate, and Create Snapshot. A table lists 100 hosts, each with its IP address, FQDN, Cluster, Data Center, Memory, CPU, and Network usage percentages. A large table below the toolbar displays detailed resource monitoring data for each host, including memory, CPU, and network usage over time.

ment	Host	IP Addresses	FQDN	Cluster	Data Center	Memory	CPU	Network	Co
	node10	192.168.217.156 fe8...	andrade-a39697...	cluster1	datacenter0	8%	0%	0%	
	node8	192.168.217.170 20...	apereira-a31033...	cluster0	datacenter0	16%	0%	0%	
	node11	10.10.217.161 2001...	apereira-matlab.d...	cluster1	datacenter0	16%	0%	0%	
	node5	192.168.217.175	apf-a42807.aluno...	cluster0	datacenter0	12%	0%	0%	
	node0	192.168.217.176 fe8...	apf-a42815.aluno...	cluster0	datacenter0	8%	0%	0%	
	node0	192.168.217.177 fe8...	apf-a42816.aluno...	cluster0	datacenter0	40%	9%	0%	
				cluster0	datacenter0	--	--	--	
	apereira-matlab			cluster1	datacenter0				
	apf-a42807-ansys			cluster0	datacenter0				
	apf-a42815-ansys			cluster0	datacenter0				
	apf-a42816-ansys			cluster0	datacenter0				
	apinto-insignia-pipeline2-armd			cluster0	datacenter0				
	apinto-insignia-pipeline2-xeon			cluster1	datacenter0				
	apinto-insignia-pipeline3-xeon			cluster1	datacenter0				
	avtl-a39432-matlab			cluster1	datacenter0				
	balsa-a39555-ansys			cluster1	datacenter0				
	balsa-a42698-matlab-ansys			cluster0	datacenter0				
	balsa-a42889-matlab-ansys			cluster0	datacenter0				
	balsa-a42904-matlab-ansys			cluster0	datacenter0				

Cluster do CeDRI - Tecnologias de Exploração

- Plataforma de Virtualização oVirt

The screenshot shows the oVirt Open Virtualization Manager interface. On the left, there is a list of four virtual machines: DEBIAN 7, FREEBSD 9.2 X64, UBUNTU TRUSTY Tahr LTS, and WINDOWS 10 X64. Each VM has a status icon (Off for DEBIAN 7, FREEBSD, and UBUNTU; On for WINDOWS 10), a name, and a 'Run' button. The WINDOWS 10 X64 VM is selected, opening a detailed view on the right. This view includes:

- Details:** Host IP Address: 169.254.98.5, FQDN: cluster0, Cluster: datacenter0, Template: Win10LTSC2019 (base version), CD: [Empty], Sysprep: ON, Boot Menu: OFF, Optimized For: Desktop, Total Virtual CPUs: 4, Memory: 8.0 GB.
- Utilization:** CPU, Memory, Networking, Disk. The Disk section shows 63.6 Unallocated of 64 GiB Provisioned.
- Network interfaces:** nic1 (vm17/vm17).
- Disk:** 385 MiB Allocated, associated with Windows_10_LTSC_2019_Disk1 (64 GiB) and is bootable.

Portal Self-Service

Cluster do CeDRI - Tecnologias de Exploração

- Plataforma de Virtualização Proxmox

Gestão do Cluster

The screenshot shows the Proxmox Virtual Environment 7.4-4 web interface. The left sidebar lists nodes: node0, node1, node10, node11, node12, node13, node14, node15, node16, node17, node18, node2, node3, node4, node5, node6, node7, node8, node9, and csirt-pool. The main area has tabs for Health, Guests, and Resources.

Health: Status is green (checkmark). Nodes: 19 Online, 0 Offline. Cluster: ProxmoxIPB, Quorate: Yes.

Guests: Virtual Machines: Running 178, Stopped 91, Templates 25. LXC Container: Running 16, Stopped 11, Templates 2.

Resources: CPU: 4% of 1320 CPU(s). Memory: 66% of 3.55 TiB of 5.41 TiB. Storage: 20% of 33.27 TiB of 164.08 TiB.

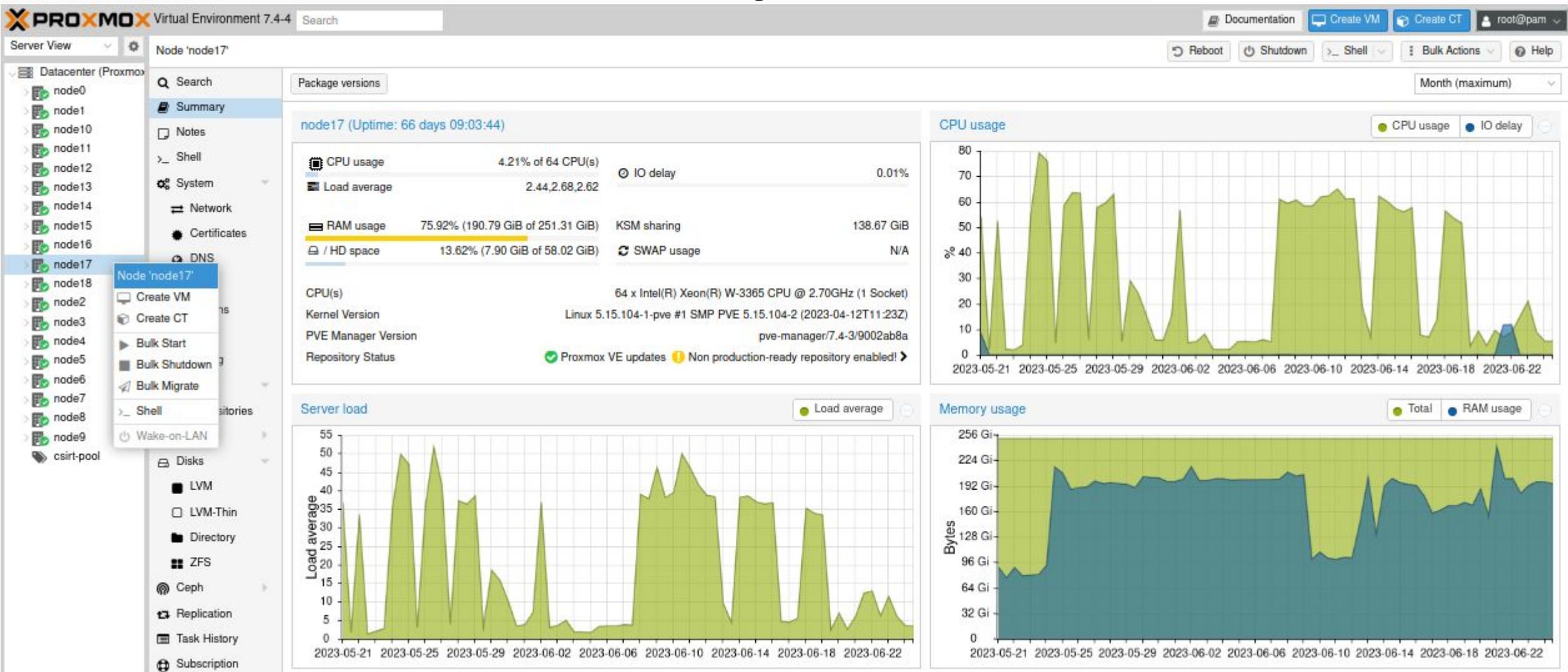
Nodes:

Name	ID	Online	Support	Server Address	CPU usage	Memory usage	Uptime
node0	1	✓	-	172.31.0.30	3%	70%	15 days 13:...
node1	2	✓	-	172.31.0.31	2%	58%	15 days 13:...
node...	13	✓	-	172.31.0.40	1%	54%	43 days 07:...
node...	16	✓	-	172.31.0.41	17%	57%	89 days 03:...
node...	14	✓	-	172.31.0.42	2%	68%	15 days 06:...
node...	15	✓	-	172.31.0.43	6%	72%	68 days 03:...
node...	6	✓	-	172.31.0.44	8%	63%	72 days 13:...
node...	7	✓	-	172.31.0.45	0%	80%	72 days 13:...
node...	17	✓	-	172.31.0.46	2%	75%	14 days 12:...
node...	18	✓	-	172.31.0.47	4%	76%	66 days 08:...
node...	19	✓	-	172.31.0.48	1%	73%	66 days 08:...
node2	3	✓	-	172.31.0.32	5%	80%	15 days 13:...

Cluster do CeDRI - Tecnologias de Exploração

• Plataforma de Virtualização Proxmox

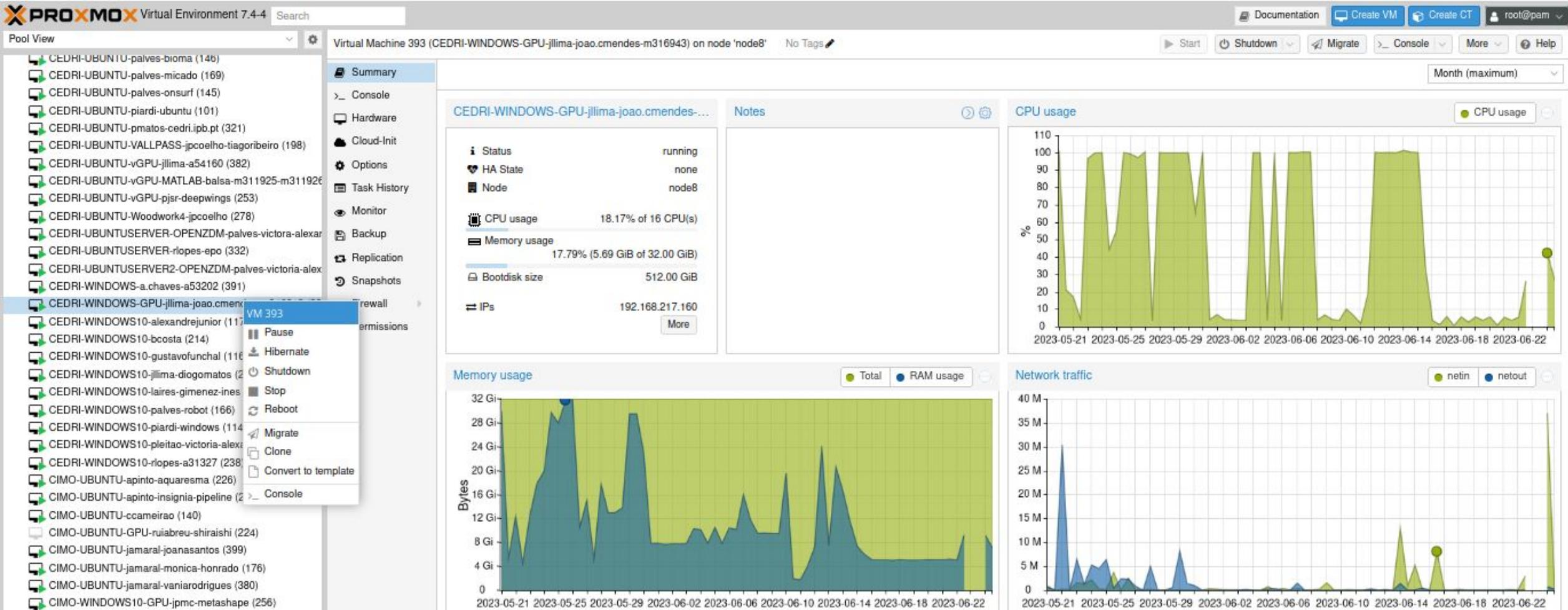
Gestão de Nós



Cluster do CeDRI - Tecnologias de Exploração

• Plataforma de Virtualização Proxmox

Gestão de MVs



Cluster do CeDRI - Tecnologias de Exploração

- **Armazenamento para Virtualização Bare-Metal**

- **local dedicado:**

- SSDs ou HDs de um nó alojam as MVs desse nó; **maximiza o desempenho** (SSDs)
 - impede/dificulta HA e balanceamento de carga (migrações lentas)
 - resiliência ausente; colmatável com RAID local / backups periódicos

- **remoto partilhado**

- serv. de armazenamento, com discos em RAID ou sist. de ficheiros avançados (ZFS), partilha armazenamento pela **rede** (aconselhável link/vlan dedicada e MTU 9000)
 - partilha de espaço para discos virtuais com base em NFS, iSCSI
 - **maximiza a flexibilidade** (migrações de VMs instantâneas)
 - menor desempenho (concorrência do acesso) que local dedicado
 - vulnerável à falha do servidor ou degradação do seu desempenho

Cluster do CeDRI - Tecnologias de Exploração

- **Armazenamento para Virtualização Bare-Metal**

- **hyperconverged**

- tira partido da possibilidade dos nós de virtualização alojarem discos locais
 - discos locais combinados num syst. de ficheiros distribuído, com alguma redundância (replicação de blocos em diferentes nós), através de soluções como CEPH, GlusterFS
 - alguma tolerância a falhas; desempenho penalizado (sincronizar réplicas nas escritas)

- **remoto dedicado**

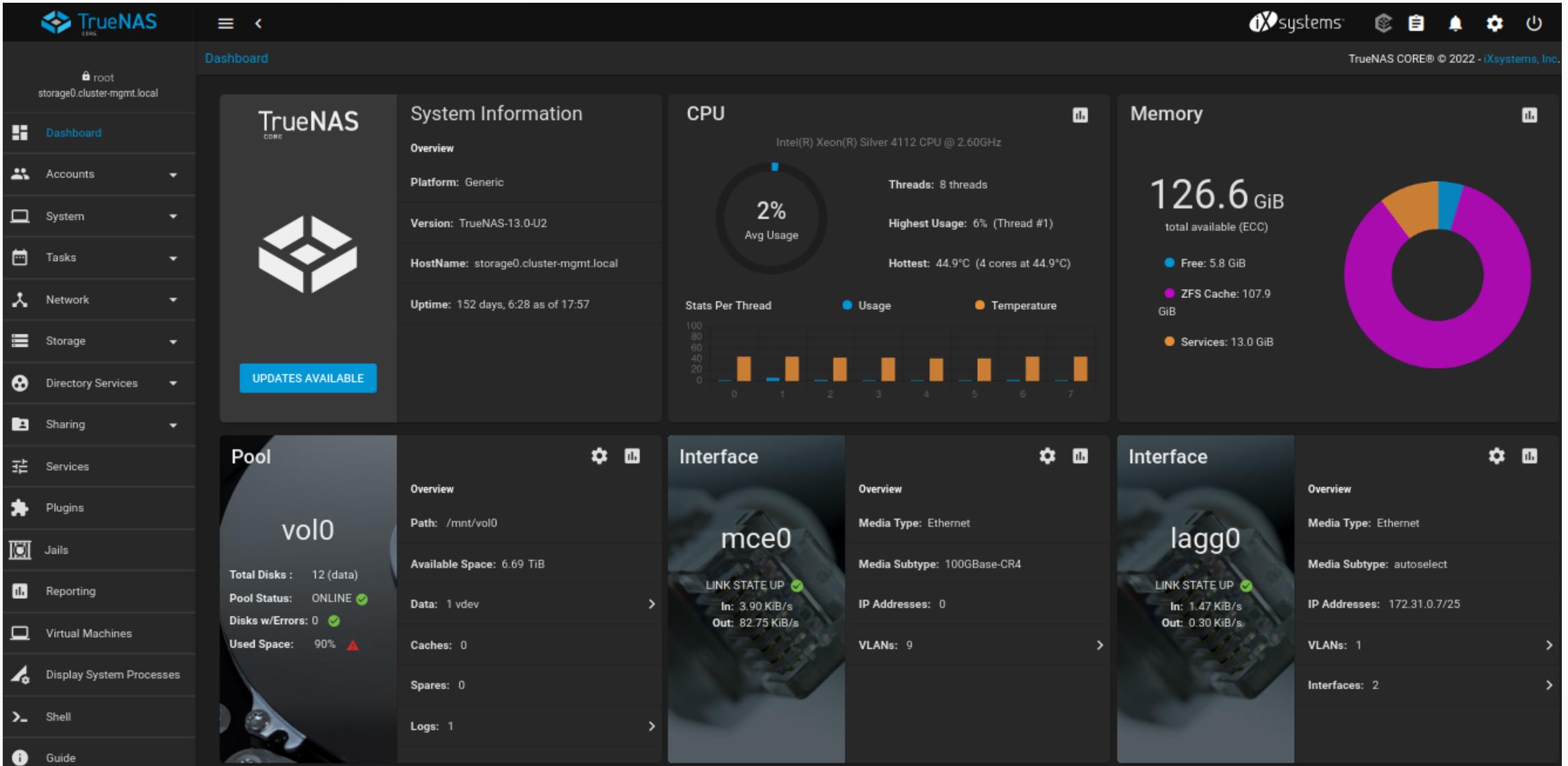
- baseado no protocolo NVMe-oF (= NVMe over Fabrics (RDMA/Fiber Channel/TCP))
 - adequado para tirar pleno partido de servidores com arrays de SSDs NVMe
 - NVMe-oF permite acesso eficiente a “partições” de SSDs em servidores remotos
 - cada nó de virtualização pode usar partições remotas como se fossem locais
 - oferece bom desempenho, mas exige a combinação certa de componentes

Cluster do CeDRI - Tecnologias de Exploração

- **Armazenamento para Virtualização Bare-Metal**
 - **opções de armazenamento usadas/testadas no cluster do CeDRI**
 - **local dedicado (Proxmox)**: SSDs NVMe sob EXT4 (só para discos virtuais)
 - **remoto partilhado (oVirt+Proxmox)**: servidor com 12 HDs
 - TrueNAS Core (<https://www.truenas.com/truenas-core/>), discos sob ZFS, RAID-Z2
 - partilhas NFS para discos virtuais, trânsito (migração) de VMs, datasets, pessoais
 - partilhas SMB para datasets e pessoais
 - **hyperconverged (Proxmox)**: CEPH (sobre SSDs); desempenho insuficiente
 - **remoto dedicado (em testes)**: servidor com 12 SSDs.
 - resultados promissores com NVMe-over-RDMA/TCP
 - implica usar Linux no servidor em vez de TrueNAS

Cluster do CeDRI - Tecnologias de Exploração

- Plat. de Armaz. Remoto Partilhado - TrueNAS



Cluster do CeDRI - Tecnologias de Exploração

- **Cópias de Segurança (Backups)**

- **oVirt:**
 - ausência de funcionalidade/componente nativo
 - solução: full clones das MVs, ferramentas externas
- **Proxmox:**
 - funcionalidade nativa:
 - backups totais periódicos, taxa de retenção configurável, restore total
 - appliance Proxmox Backup Server (ainda por operacionalizar)
 - backups não-totais, restore de ficheiros individuais
 - backups para partilhas NFS em servidores externos ao cluster
- num pedido de MV é indicada a política (período e retenção) pretendida

Cluster do CeDRI - Tecnologias de Exploração

- Backups Nativos do Proxmox

The screenshot shows the Proxmox Virtual Environment 7.4-4 interface. The left sidebar shows a tree view of nodes: Datacenter (ProxmoxIPB) with node0, node1, node10, node11, node12, node13, node14, node15, node16, node17, node18, and node2. The 'Backup' option under Datacenter is highlighted with a blue background. The main panel displays a table of backup jobs for the 'nascri2-backups' storage pool. A red box highlights the 'Backup' section of the table header. The table columns are: Enabled, Node, Schedule, Next Run, Storage, Com..., Retention, and Selection. The data includes various backup entries with specific details like schedule times (e.g., sat *-1..7 03:45), next run times (e.g., 2023-07-01 03:45:00), storage (e.g., nascri2-backups), and retention policies (e.g., keep-last=2).

Enabled	Node	Schedule	Next Run	Storage	Com...	Retention	Selection
✓	-- All --	sat *-1..7 03:45	2023-07-01 03:45:00	nascri2-backups	satur...	keep-last=2	110,111,112,139,137,193,231,235,237,253
✓	-- All --	sun *-1..7 02:00	2023-07-02 02:00:00	nascri2-backups	sund...	keep-last=1	101,109,114,116,117,118,160,183,222,251,257,156
✓	-- All --	sat 07:00	2023-07-01 07:00:00	nascri2-backups	satur...	keep-last=3	113,119,242,274
✓	-- All --	sat *-1..7 05:00	2023-07-01 05:00:00	nascri2-backups	satur...	keep-last=1	115,107,108,131,132,140,141,144,148,151,161,194...
✓	-- All --	sat 03:00	2023-07-01 03:00:00	nascri2-backups	satur...	keep-last=2	138,120,158,255
✓	-- All --	sat 02:00	2023-07-01 02:00:00	nascri2-backups	satur...	keep-last=1	147,168,277,278,320,302,332,381,198
✓	-- All --	06:45	2023-06-26 06:45:00	nascri2-backups	daily...	keep-last=8	150,152,159
✓	-- All --	sun *-1..7 02:45	2023-07-02 02:45:00	nascri2-backups	sund...	keep-last=6	149,156

The screenshot shows the Proxmox Virtual Environment 7.4-4 interface. The left sidebar shows a tree view of nodes: Datacenter (ProxmoxIPB) with node0, which contains VMs 102 (rocks-node-0), 110 (terminais0), 125 (offline-oVirt-node71), 147 (CEDRI-HAOS-brito), and 150 (ESTIG-UBUNTU-dc). The '102 (rocks-node-0)' VM is selected. The main panel shows the 'Virtual Machine 102 (rocks-node-0) on node 'node0'' view. A red box highlights the 'Backups' section of the top navigation bar. Below it, the table lists three backup files: vzdump-qemu-102-2023_04_13-11_15_26.vma.zst, vzdump-qemu-102-2022_08_19-19_39_46.vma.zst, and vzdump-qemu-102-2022_02_01-12_17_01.vma.zst. The table columns are: Name, Notes, Date, Format, and Size.

Name	Notes	Date	Format	Size
vzdump-qemu-102-2023_04_13-11_15_26.vma.zst	rocks-node-0	2023-04-13 11:15:26	vma.zst	47.09 GB
vzdump-qemu-102-2022_08_19-19_39_46.vma.zst	rocks-node-0	2022-08-19 19:39:46	vma.zst	1.30 GB
vzdump-qemu-102-2022_02_01-12_17_01.vma.zst		2022-02-01 12:17:01	vma.zst	1.64 GB

Cluster do CeDRI - Tecnologias de Exploração

- **Organização e Acesso às Máquinas Virtuais**

- no cluster coexistem centenas de máquinas virtuais (atual/ ≈ 300)
- cada MV deve poder ser **acedida (via rede) apenas por quem de direito**
- cada MV deve ser **visível (na rede) apenas a outras MVs com afinidade**
- as **MVs são segregadas em redes isoladas**, definidas por uma VLAN
- essas redes designam-se, no cluster do CeDRI, por **mini-clusters**
 - por projeto, equipe, unidade organizacional, convidados, ...
- as MVs de um mini-cluster podem estar espalhadas de forma arbitrária pelos nós do cluster, ou concentradas em certos nós, em função dos requisitos

Cluster do CeDRI - Tecnologias de Exploração

- **Organização e Acesso às Máquinas Virtuais**

- o acesso a um mini-cluster faz-se através de uma **conexão VPN**
- o acesso a uma MV faz-se via **cliente SSH** ou **cliente de ambiente de desktop remoto** (RDP p/ MVs Windows, X2GO p/ MVs Linux)
 - estes clientes permitem também a **transferência de ficheiros**
- todo o tráfego na conexão VPN entre clientes e MVs é **encriptado**
- a gestão dos mini-clusters é feita por uma *firewall pfSense*
 - registo das contas associadas cada mini-cluster
 - uma instância do serviço **OpenVPN** por mini-cluster
 - clientes OpenVPN personalizados por cada mini-cluster
 - servidor DHCP e DNS para cada mini-cluster
- a *firewall* corre num servidor externo ao cluster de virtualização



Cluster do CeDRI - Tecnologias de Exploração

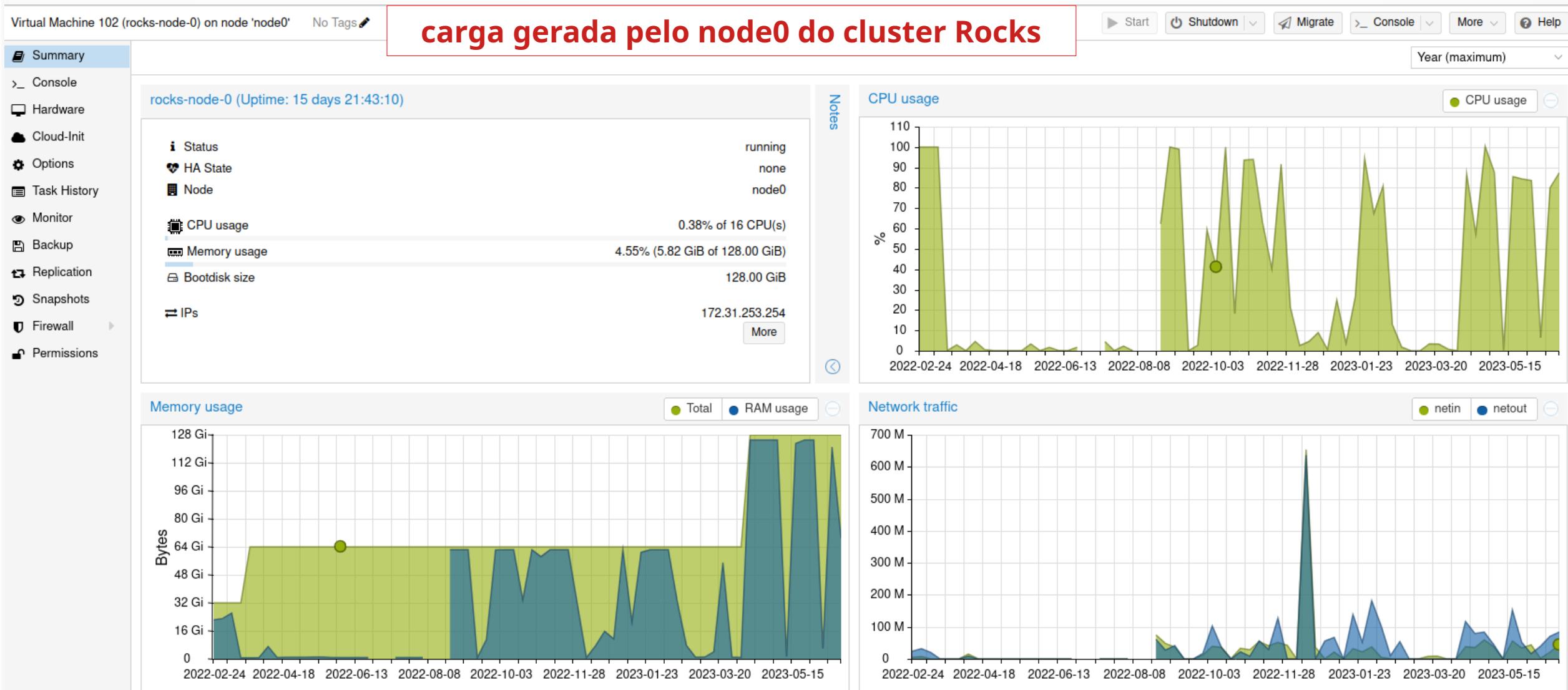
- **Cluster HPC Virtualizado**

- para maximizar o desempenho, um cluster HPC tradicional é bare-metal
- se o número de utilizadores de regimes de submissão de trabalhos por lotes (batch) é reduzido, não se justifica comprometer nós físicos com esse regime
- alternativa: virtualizar o cluster HPC (frontend e satélites passam a MVs)
- **o cluster do CeDRI aloja um cluster HPC virtualizado (Rocks 7.0)**
 - usado sobretudo por investigadores do CIMO (apps. de BioInformática)
 - áreas dos utilizadores e datasets no serv. de armazenamento (acesso NFS)
 - nas MVs apenas reside o SO, aplicações e partições de scratch
 - partições de scratch residem nos SSDs dos nós de virtualização
 - carga E/S gerada pelos jobs é distribuída, maximizando o desempenho

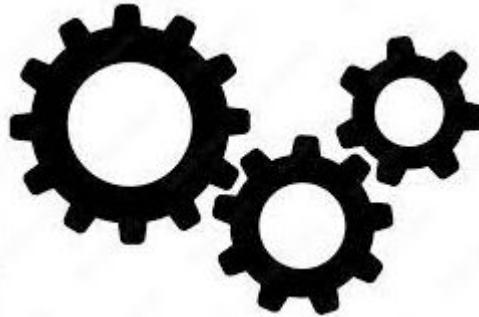


Cluster do CeDRI - Tecnologias de Exploração

- Cluster HPC Rocks Virtualizado



Cluster do CeDRI - Workflow



- **Ciclo de Vida das Máquinas Virtuais**

- **pedido:**

- e-mail dirigido a cri-estig@ipb.pt, enviado por docente ou investigador
 - indica requisitos (HW virtual, SO, contas, política de backups, tempo de vida)

- **criação:**

- escolha do nó do cluster e criação com base em template
 - registo de endereço IP fixo (privado em mini-cluster, ou público)
 - criação de contas na firewall para acesso VPN ao mini-cluster
 - e-mail de retorno com dados de acesso e tutoriais VPN

- **abate:**

- expirado o tempo de vida, após confirmação do requisitante
 - remoção da MV, backups e eventuais pastas de rede associadas
 - remoção de conta VPN e libertação de endereços IP

Cluster do CeDRI - Casos de Uso



- **Casos de Uso Genéricos**

- aplicações (simulações) de **longa duração**
- aplicações com {grandes} necessidades de {CPU, RAM, armaz.}, **GPUs**
- aplicações que tiram partido de **execução distribuída**
 - possível agregar várias MVs no seu próprio cluster virtual (e.g., ROCKS)
- **workstation virtual** p/ desenvolvimento (alunos, BIs, invest., docentes)
- alojamento de **sites** e **demonstradores** de projetos de investigação
- **fins pedagógicos** (portal Self-Service, MVs c/ virtualização nested)
- apoio a centros de recursos/competências (**CRI/ESTiG**, **CyberSec/IPB**)

Cluster do CeDRI - Casos de Uso

- **Alguns Testemunhos de Utilizadores**



Cluster do CeDRI



- **Conclusões e Perspetivas**

- nascido no CeDRI, o cluster tornou-se um recurso imprescindível à I&D nele realizada e também a outros investigadores (CIMO, MORECOLAB)
- tem sido útil no suporte ao funcionamento de várias UCs na ESTiG
- tem potencial para reforçar esse papel e chegar a mais pessoas, mas
 - necessita staff IT dedicado para sustentar essa ambição
 - necessita atenção constante a oportunidades de atualização

Obrigado pela vossa atenção !
Questões ? Comentários ? Sugestões ?