WORKSHOP GARR 2023
NET MAKERS

Virtual Data Centers: Fueling Data Science with OpenStack

Stefano Cacciaguerra
Stefano Chiappini
INGV

NE.RE.I.D.E.

INGV implemented the **NEw REsearch Infrastructure Datacenter for EMSO** in the **Western Ionian Sea EMSO site**, at Portopalo di Capopassero (SR), Italy:

➔ an European **research infrastructure** of EMSO-ERIC funded by **PON InSEA**

➔ a **ICT infrastructure** for archiving, processing, and sharing **scientific data** from marine observatories and for developing advanced services

➔ it promotes multidisciplinary **scientific/technological research** to understand anthropogenic phenomena in the deep marine environment

⇨ **Data science** supports the understanding of complex marine phenomena through **advanced processing of the collected data**

What do Data Scientists want in NEREIDE ?

Data Analysis: Using **scalable computational resources** to analyze large volumes of data

Model Development: Exploiting **computing power** to perform complex simulations

Data Management: Organize and manage complex datasets, using distributed storage features to ensure **data integrity** and security

Data Visualization: Making **complex data clear** with visual tools like maps and charts

Collaboration: Working with other scientists to promote **interdisciplinary research**

Workflow Automation: Using management tools to **automate** processes and running periodic analyses

Secure Remote Access: Accessing to infrastructure from **anywhere**, ensuring real-time research continuity

How NEREIDE supports Data Science

Virtual Data Centers based on Openstack provide a custom and adaptable virtual environment, enabling precise control over applications and data management

➔ **Technologies** like **JupyterHub**, ERDDAP and ElasticSearch Cluster power data analysis and visualization

➔ Openstack **scalable nature** allows data scientists to adjust processing and storage resources for handling large datasets or complex simulations

➔ Enhanced **interdisciplinary** collaboration through **IDEM** and **GARR Cloud Federation**

➔ **Automation tools**, like **MaaS/JuJu**, simplify data science workflows, from data analyses to results sharing

Openstack → Users, Projects and Tenants

Openstack is a free open standard cloud computing platform deployed as **Infrastructure-as-a-Service** where **cloud resources are made available to users**

Users can manage cloud resources through a **web-based dashboard**, **command-line tools**, or **RESTful web services**

Project is the base unit of **ownership** in OpenStack (all resources must be owned by a specific project). In OpenStack Identity, a project must be owned by a specific domain

Tenant is a group of users in charge of a **logical grouping** of cloud resources

Tenant & Virtual Data Center

In our solution, **Tenants** are in charge of cloud resources where users could install, configure and manage virtual machines behind a **Gateway** owning a **public IP address** realizing a own **Virtual Data Center** (**VDC**).

➔ **cloud admins** make **cloud infrastructure** and **real Data Center** works

➔ **tenant users** create **Service** in their **VMs** on **VDC**

“Admins are owners of a mall. Admins entrust a tenant with the management of a shop. Admins are in charge of managing the whole infrastructure, managers of their own shop”

Networks & Tenants

There are different types of networks:

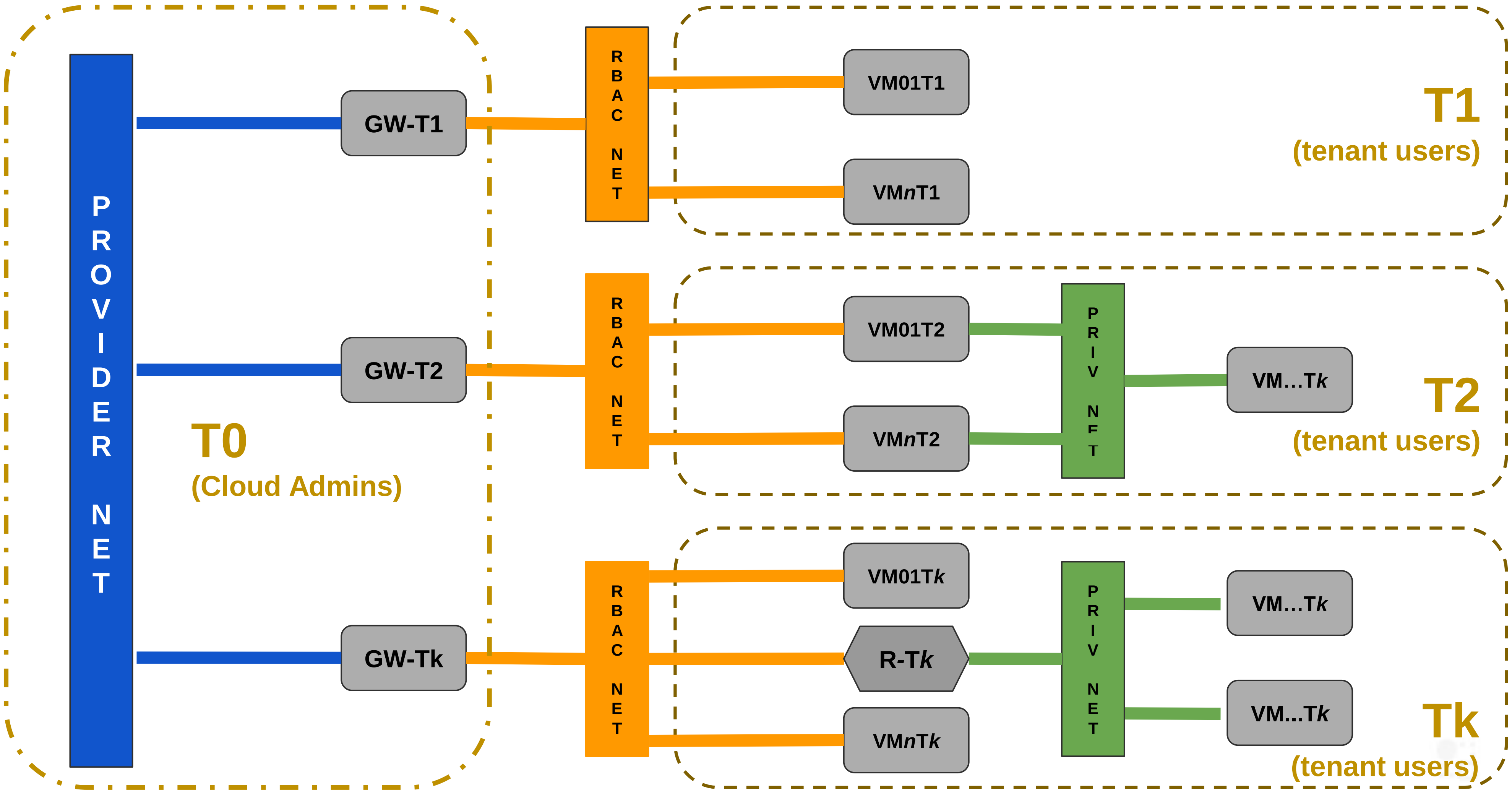
provider → mapped directly to an **existing physical network** You can use flat provider networks to connect instances directly to the **external network**. (managed by *cloud admins*)

project → multiple **private networks** are fully isolated by default and are not shared with other projects. (managed by *tenant users*)

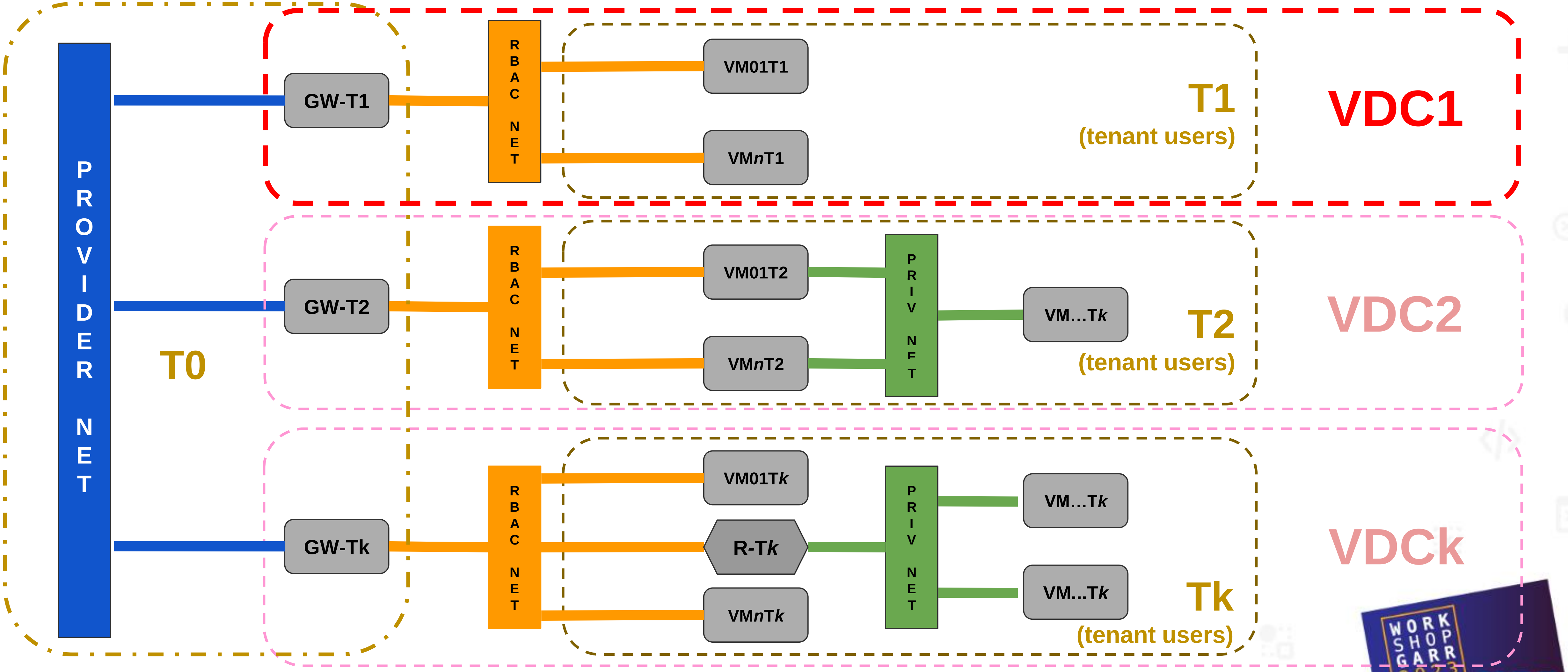
shared → networks shared among **all tenants**! (managed by *cloud admins*)

shared (RBAC) → Role-Based Access Control (RBAC) networks shared among **specific tenants**! (managed by cloud admins)

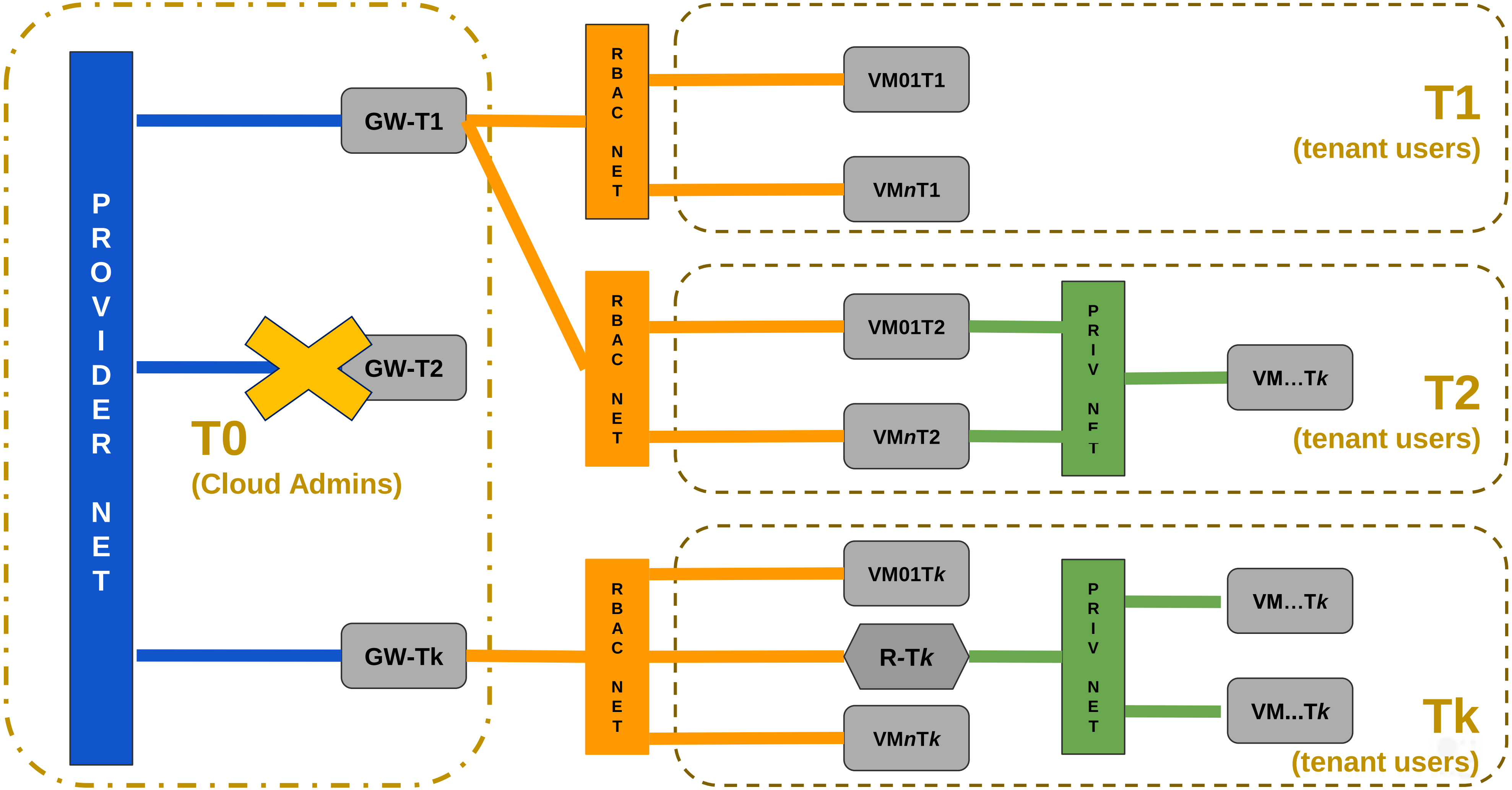
Tenants on Openstack Networking



VDCs


PROVIDER NET
T0
GW-T1
GW-T2
GW-Tk
RBAC NET
RBAC NET
RBAC NET
VM01T1
VMnT1
T1
(tenant users)
VDC1
VM01T2
VMnT2
PRIV NFT
VM...Tk
T2
(tenant users)
VDC2
VM01Tk
R-Tk
VMnTk
PRIV NET
VM...Tk
VM...Tk
Tk
(tenant users)
VDCk

Optimization of Tenants



Tenant’s Gateway

In order to share services on Internet, Tenant must use a Gateway owning a **specific Public IP address** and it must work as:

- ➔ a **Border Router**
- ➔ a **Firewall**
- ➔ an **OpenVPN** server

The Gateway is a VM inside the **T0 tenant** (cloud admins) with **two interfaces**:

- ➔ one on the **Provider** Network
- ➔ one on the **RBAC Shared** Network

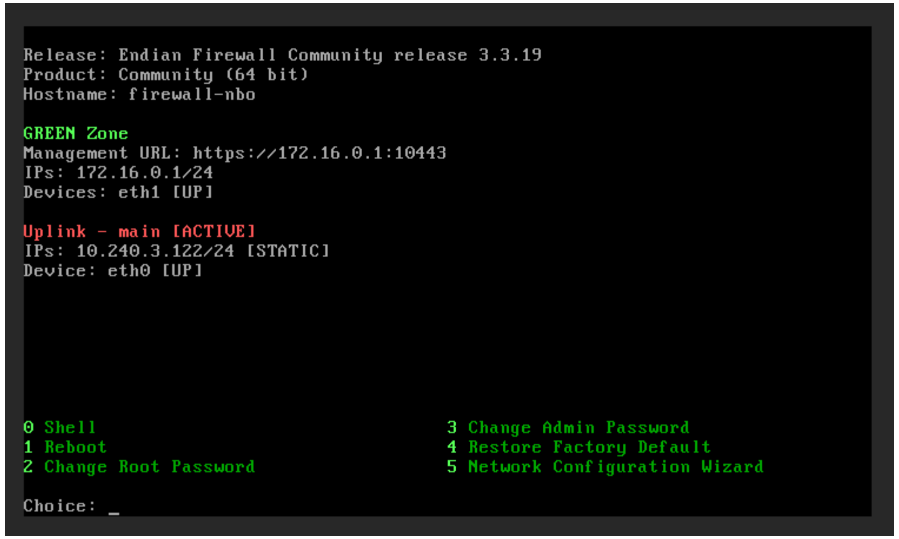
A simple linux VM or something customed like **Endian Firewall**, IPFire or OpenWrt

Endian Firewall as Gateway

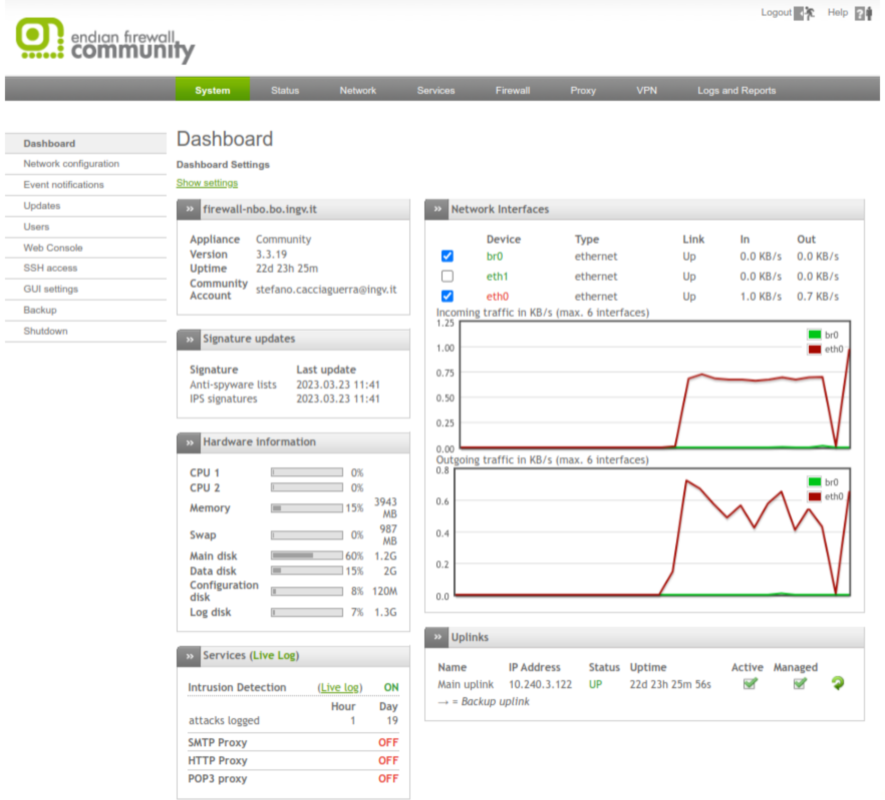
It is an open-source **router**, **firewall** and **gateway** security Linux distribution

Credentials of Endian Firewall:

- ➔ **root** of SO → Cloud Admins
- ➔ **admin** of Web Dashboard → Cloud Admins
- ➔ **admin user** of Web Dashboard → Tenant users



```
Release: Endian Firewall Community release 3.3.19
Product: Community (64 bit)
Hostname: firewall-nbo

GREEN Zone
Management URL: https://172.16.0.1:10443
IPs: 172.16.0.1/24
Devices: eth1 [UP]

Uplink - main [ACTIVE]
IPs: 10.240.3.122/24 [STATIC]
Device: eth0 [UP]




0 Shell                                   3 Change Admin Password
1 Reboot                                  4 Restore Factory Default
2 Change Root Password                    5 Network Configuration Wizard

Choice: _
```



Dashboard

Dashboard Settings

Show settings

» firewall-nbo.bo.ingv.it

Appliance	Community
Version	3.3.19
Uptime	22d 23h 25m
Community Account	stefano.cacciaguerra@ingv.it

» Network Interfaces

	Device	Type	Link	In	Out
☑	br0	ethernet	Up	0.0 KB/s	0.0 KB/s
☐	eth1	ethernet	Up	0.0 KB/s	0.0 KB/s
☑	eth0	ethernet	Up	1.0 KB/s	0.7 KB/s

Incoming traffic in KB/s (max. 6 interfaces)

Outgoing traffic in KB/s (max. 6 interfaces)

» Signature updates

Signature	Last update
Anti-spyware lists	2023.03.23 11:41
IPS signatures	2023.03.23 11:41

» Hardware information

CPU 1	0%	
CPU 2	0%	
Memory	15%	3943 MB
Swap	0%	987 MB
Main disk	60%	1.2G
Data disk	15%	2G
Configuration disk	8%	120M
Log disk	7%	1.3G

» Uplinks

Name	IP Address	Status	Uptime	Active	Managed
Main uplink	10.240.3.122	UP	22d 23h 25m 56s	✓	✓

→ = *Backup uplink*

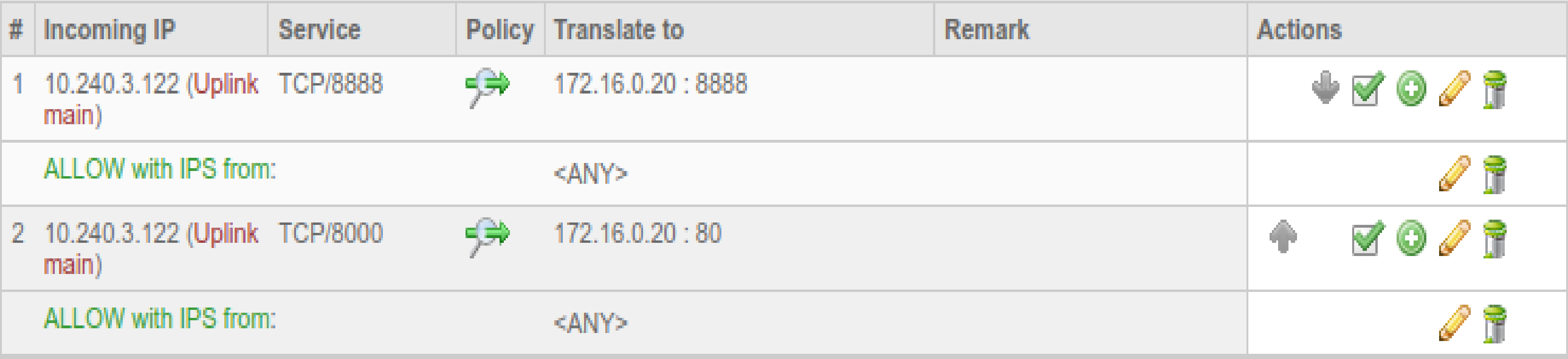
» Services (Live Log)

Intrusion Detection	(Live log)	ON
	Hour	Day
attacks logged	1	19
SMTP Proxy		OFF
HTTP Proxy		OFF
POP3 proxy		OFF

Endian Firewall Services

Port Forwarding / DNAT

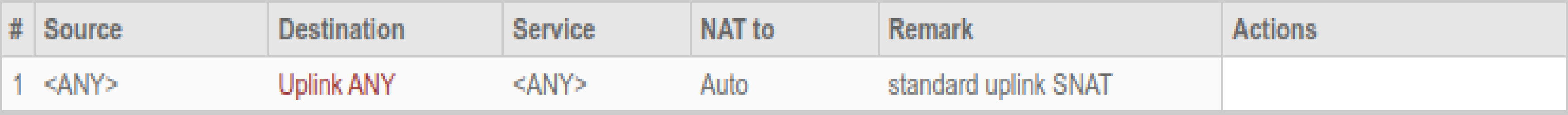
to make accessible services from Internet



#	Incoming IP	Service	Policy	Translate to	Remark	Actions
1	10.240.3.122 (Uplink main)	TCP/8888	⇒	172.16.0.20 : 8888		↓ ✓ ⊕ ✎ 🗑
	ALLOW with IPS from:			<ANY>		✎ 🗑
2	10.240.3.122 (Uplink main)	TCP/8000	⇒	172.16.0.20 : 80		↑ ✓ ⊕ ✎ 🗑
	ALLOW with IPS from:			<ANY>		✎ 🗑

Masquerading / SNAT

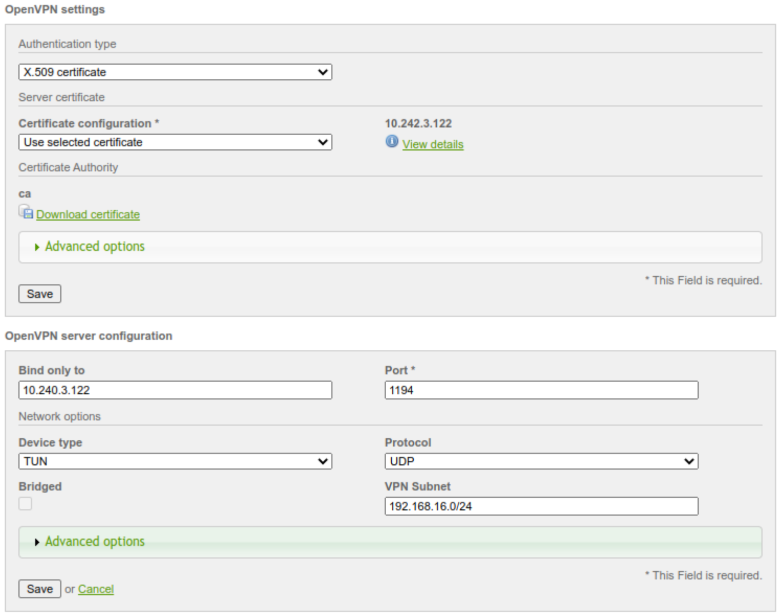
to allow VMs to access Internet



#	Source	Destination	Service	NAT to	Remark	Actions
1	<ANY>	Uplink ANY	<ANY>	Auto	standard uplink SNAT	

OpenVPN

to allow users to access VMs (ssh,https,etc)



OpenVPN settings

Authentication type

X.509 certificate

Server certificate

Certificate configuration *: Use selected certificate

10.242.3.122

View details

Certificate Authority

ca

Download certificate

▸ Advanced options

* This Field is required.

Save

OpenVPN server configuration

Bind only to: 10.240.3.122

Port *: 1194

Network options

Device type: TUN

Protocol: UDP

Bridged: □

VPN Subnet: 192.168.16.0/24

▸ Advanced options

* This Field is required.

Save or Cancel

Jupyter Ecosystem

JupyterHub (**JH**) is an open-source web application that allows **multiple users** to interact with **Jupyter Notebooks** (**JN**) on a shared server

With JH, users can log in to a **central server** using their **own credentials** and access their **own JNs**, which are hosted on this server

JNs are **interactive documents** containing executable code (like **Python**, **R**, **Julia**), visualization and text editing capabilities, it is a useful tool for **data science**

⇨ JNs can be used for **data cleaning** and **transformation**, **numerical simulation**, **statistical modeling**, **machine learning**, …

Jupyter As a Service of VDC

In order to implement JupyterHub **as a service**, it is necessary:

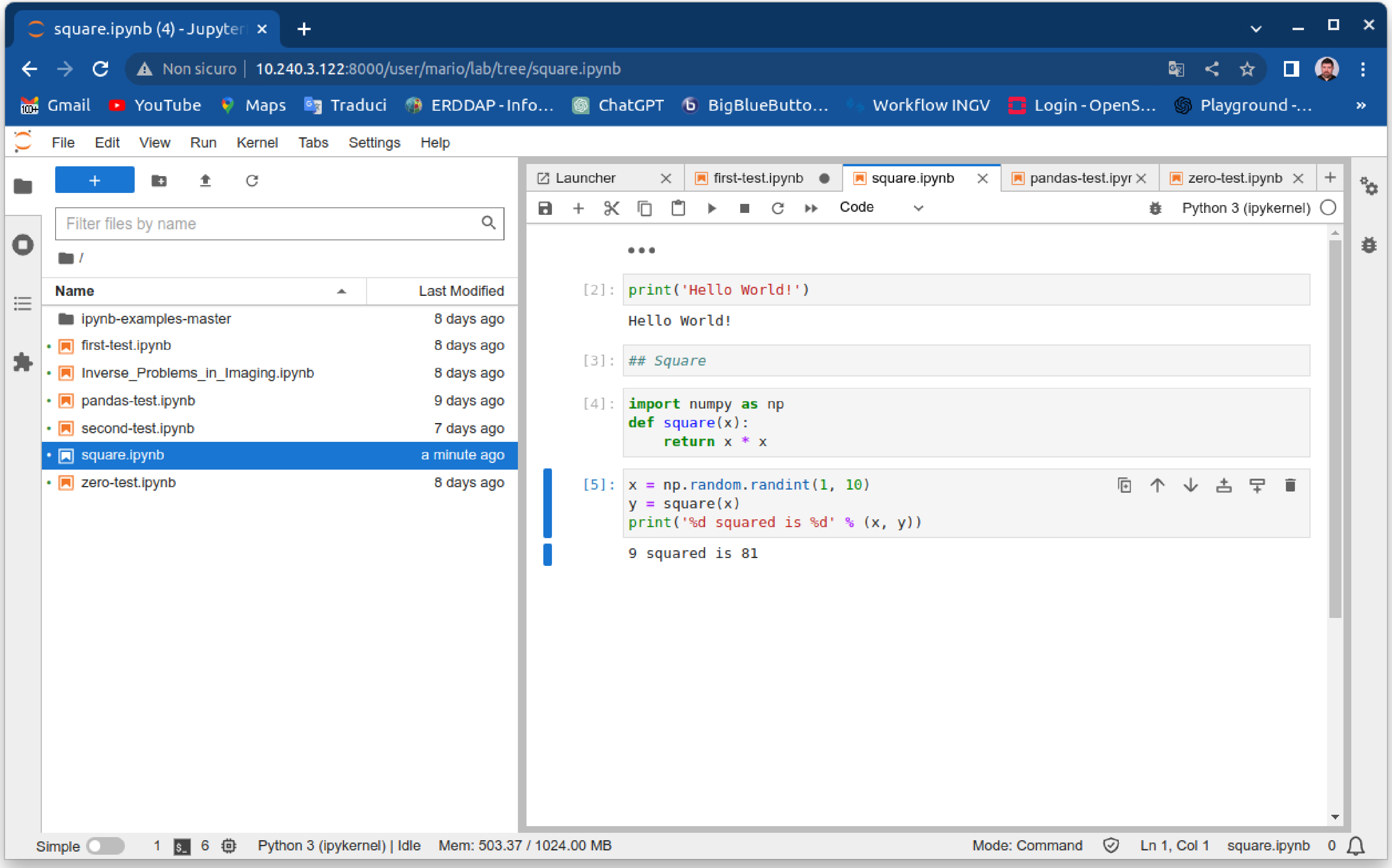
1. create your VM via **Horizon**
2. connect the **openVPN** server (the Tenant's Gateway)
3. **ssh** with **key pair** to your VM
4. install the **Littlest JupyterHub** (**TLJH**)

TLJH places 2 systemd units on your **VM**

➔ **jupyterhub.service** - starts the JupyterHub service

➔ **traefik.service** - starts proxy HTTPS

Dashboard of Jupyter



square.ipynb (4) - Jupyter
10.240.3.122:8000/user/mario/lab/tree/square.ipynb
File Edit View Run Kernel Tabs Settings Help
Launcher
first-test.ipynb
square.ipynb
pandas-test.ipyr
zero-test.ipynb
Code
Python 3 (ipykernel)
Filter files by name

Name	Last Modified
ipynb-examples-master	8 days ago
first-test.ipynb	8 days ago
Inverse_Problems_in_Imaging.ipynb	8 days ago
pandas-test.ipynb	9 days ago
second-test.ipynb	7 days ago
square.ipynb	a minute ago
zero-test.ipynb	8 days ago

```
[2]: print('Hello World!')

Hello World!

[3]: ## Square

[4]: import numpy as np
     def square(x):
         return x * x

[5]: x = np.random.randint(1, 10)
     y = square(x)
     print('%d squared is %d' % (x, y))

9 squared is 81
```

Simple 1 6 Python 3 (ipykernel) | Idle Mem: 503.37 / 1024.00 MB Mode: Command Ln 1, Col 1 square.ipynb 0

What is the main result?

From “**Tenant Users**” side → **Data Scientist**

⇒ lets tenant users operate complex infrastructure (**ready-to-use**) without the task of setting up and managing a **Real Data Center**

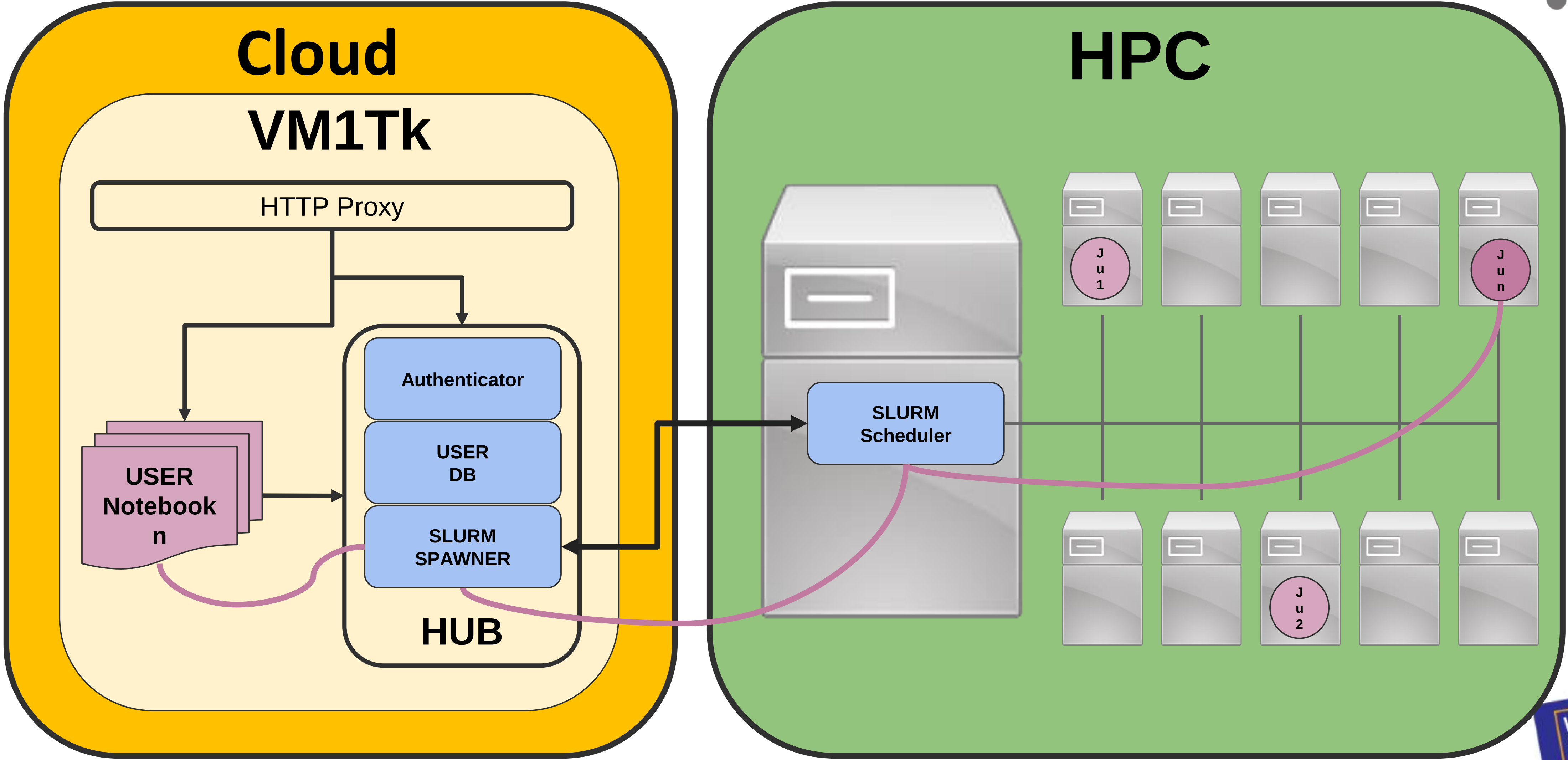
From “**Cloud Admins**” side → **Data Engineer**

⇒ **Centralizing** e **Optimizing** the resources

⇒ **Virtualizing** all the resources

⇒ **Rationalizing** the infrastructure investments

Jupyter Slurm Spawner from Cloud to HPC



VDC Migration

Is it possible to Migrate a VDC from a source cloud openstack to a target one?

Acknowledgement

We would like to express our gratitude for their collaboration to:

⇒ **Alex Barchiesi**

⇒ **Alberto Colla**

⇒ **Claudio Pisa**

GARR

and we would like to mention:

NereideBO tenant created in Bologna on **Cloud SUPER** (**POR-FESR** - **Supercomputing Unified Platform - Emilia-Romagna**) is used for part of the development and experiments on VDC

WORKSHOP GARR 2023
NET MAKERS

Virtual Data Centers: Fueling Data Science with OpenStack

Stefano Cacciaguerra
Stefano Chiappini
INGV

NEREIDE

2318 9129