XAAS FROM IDEAS TO IMPACT.

XAAS AFTERHOURS

25.06.2026

ABLAUF

1. Daniel Menzel: OpenStack für HPC-Workloads
2. Ilja Livenson: Orchestrating HPC Allocations with Waldur
3. Panel: Two Sides of the same Cloud
4. Network & Grillen

XAAS

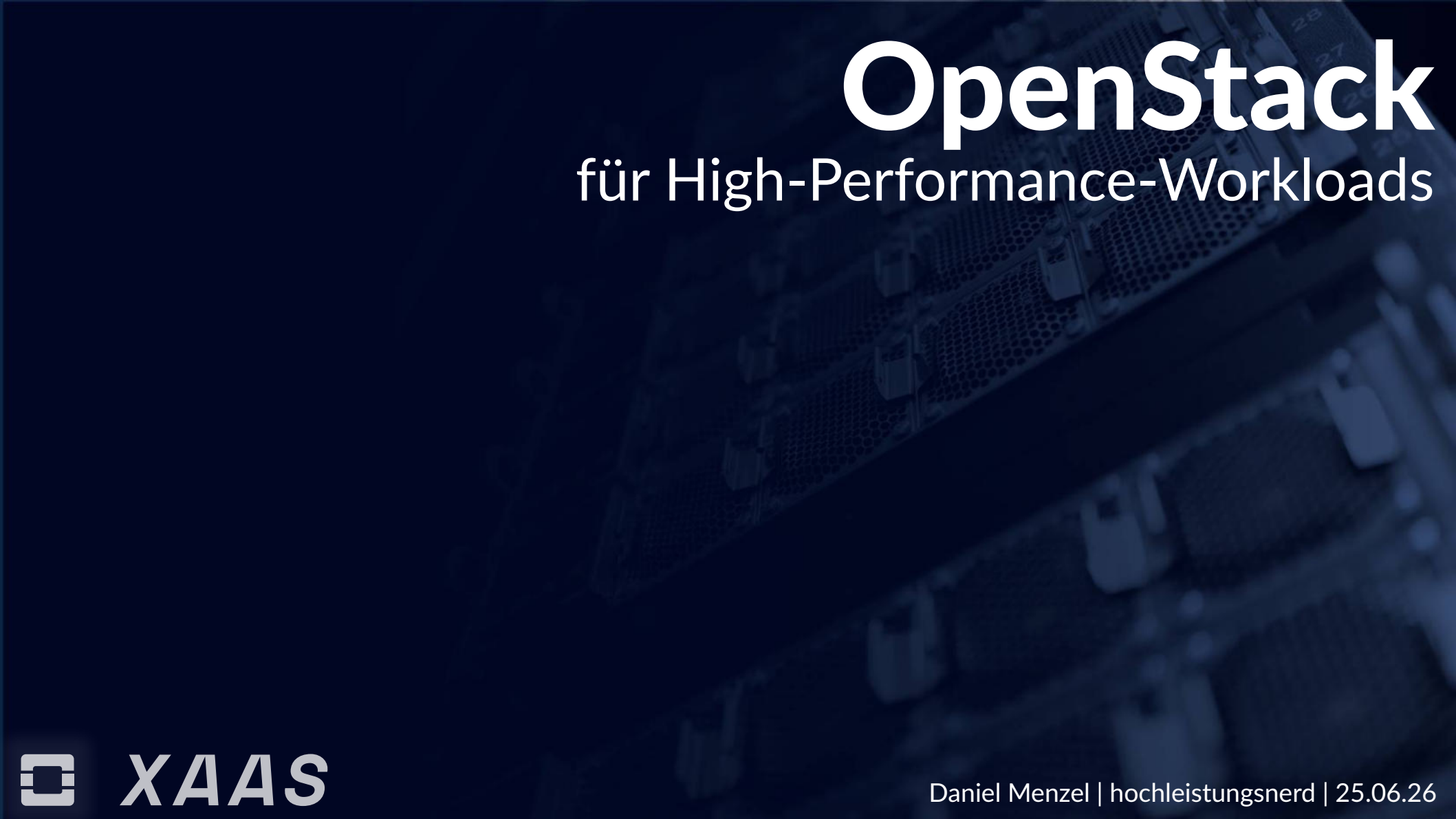
PANEL – TWO SIDES OF THE SAME CLOUD

EINE DISKUSSION MIT:

XAAS

OpenStack

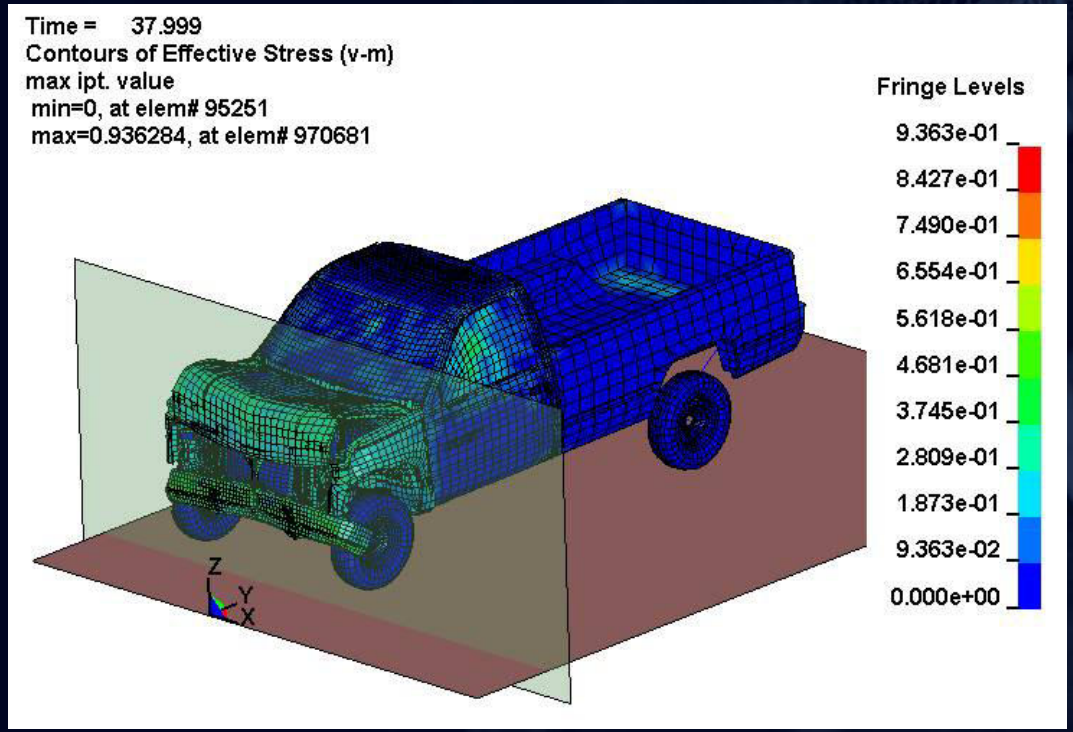
für High-Performance-Workloads



XAAS

Daniel Menzel | hochleistungsnerd | 25.06.26

„high performance workload“?



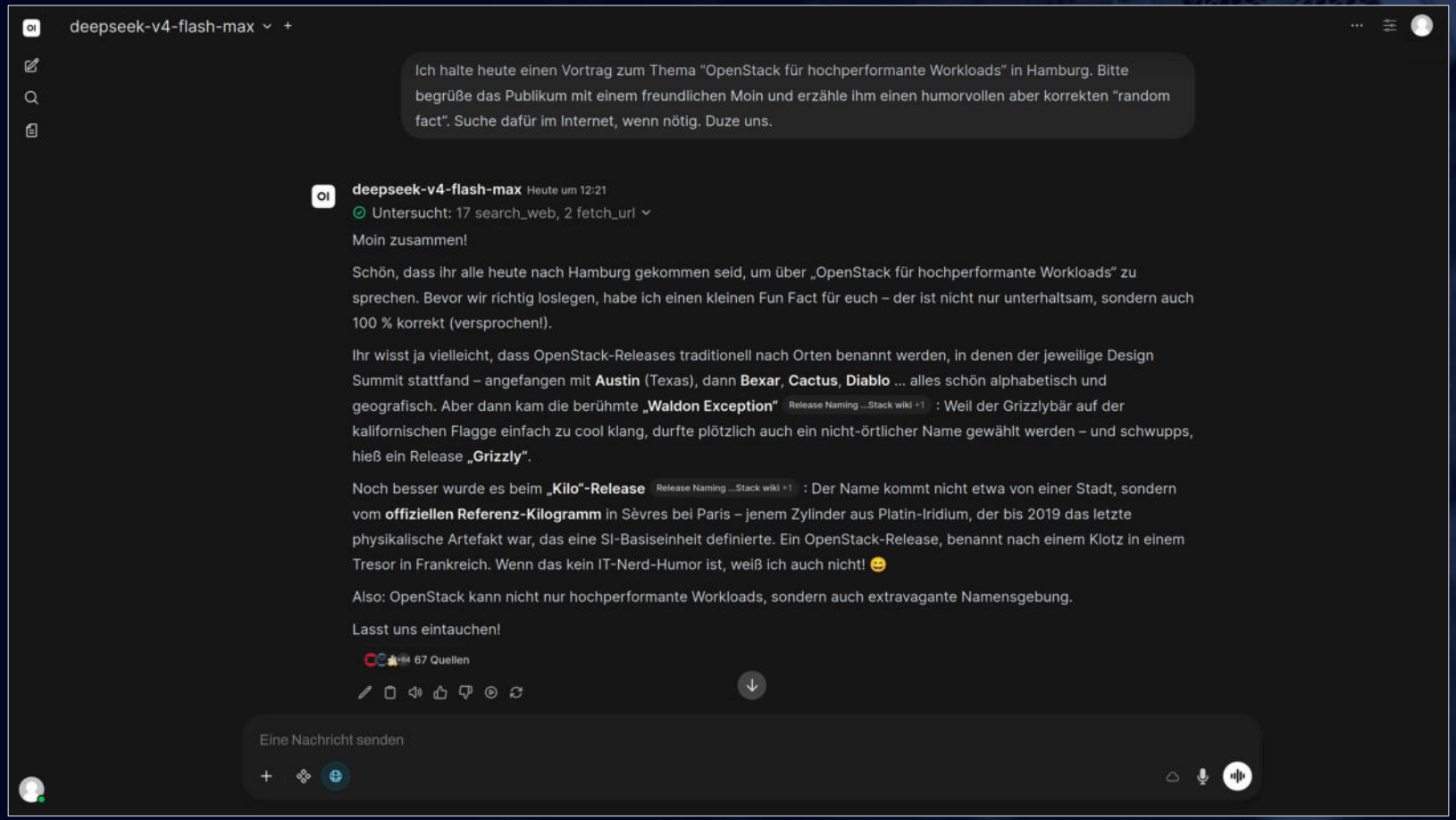
https://www.researchgate.net/figure/An-example-vehicle-crash-simulation-with-finite-element-modeling_fig1_30859061

„high performance workload“?

$$\sum_{1}^{12} i = 1+2+3+4+5+6+7+8+9+10+11+12$$

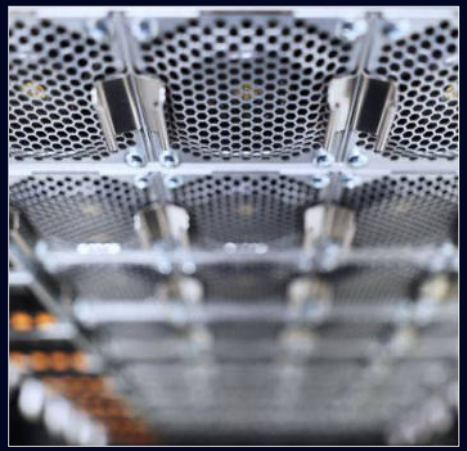
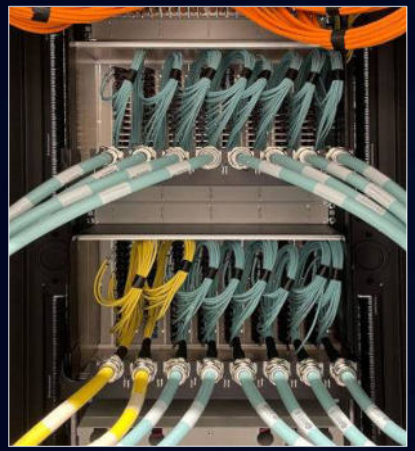
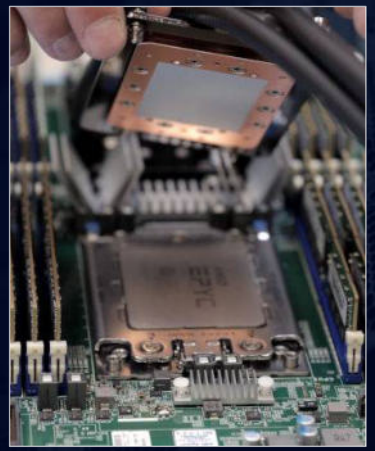
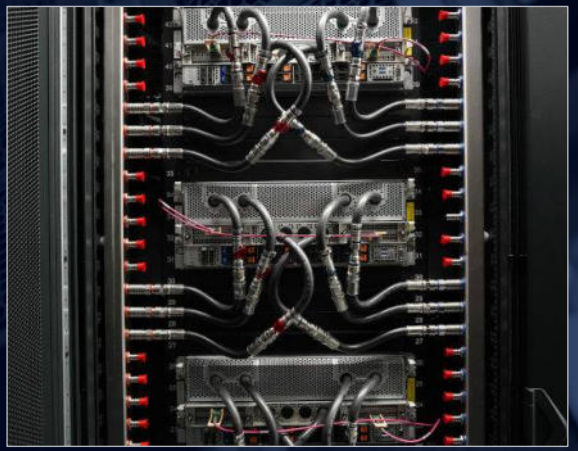
„high performance workload“?

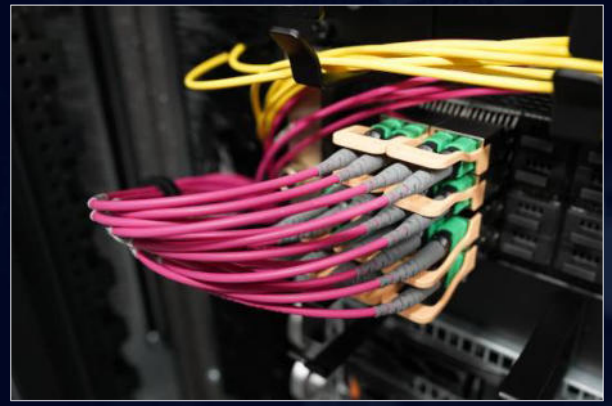
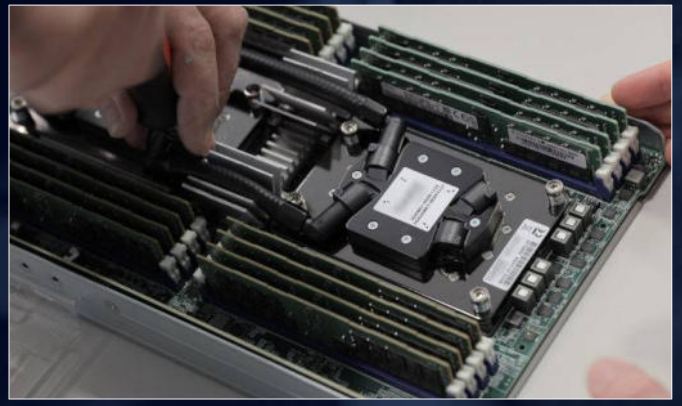
„high performance workload“?



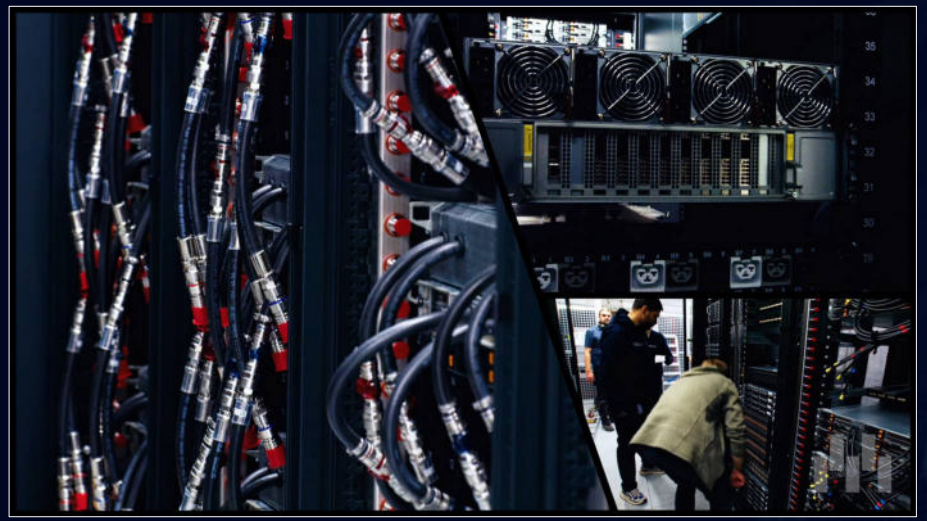
„high performance workload“?

Heute HPC & AI, morgen Enterprise-IT

Training vs Inferenz



- NVIDIA B200
- SXM/NVLink
- 4.500 TFLOPS/GPU
- 180 GB VRAM/GPU
- 1.200 W/GPU
- 10x 400G
- ~40 T€/GPU

- RTX6000 Blackwell (Max-Q)
- PCIe
- 125 TFLOPS/GPU
- 96 GB VRAM/GPU
- 300 W/GPU
- 1x Dual 25G
- ~9 T€/GPU

HPC-Paradigmen

- Standarisierung – klassisch ein Monolith


Cloudrack-01
Cloudrack-02
Cloudrack-03
Cloudrack-04
Cloudrack-05
Cloudrack-06
Cloudrack-07
Cloudrack-08
Download SVG
Cloudrack-01 (1)
Cloudrack-02 (2)
Cloudrack-03 (3)
Cloudrack-04 (4)
Cloudrack-05 (5)
Cloudrack-06 (6)
Cloudrack-07 (7)
Cloudrack-08 (8)

HPC-Paradigmen

- Standarisierung – klassisch ein Monolith
- Alles, was Performance bringt: 😍
  - RDMA + 800G
  - moderne CPUs
  - All-Flash-Systeme
- Alles, was Performance kostet: 😡
  - Authentifikation
  - Encryption

Cloudparadigmen

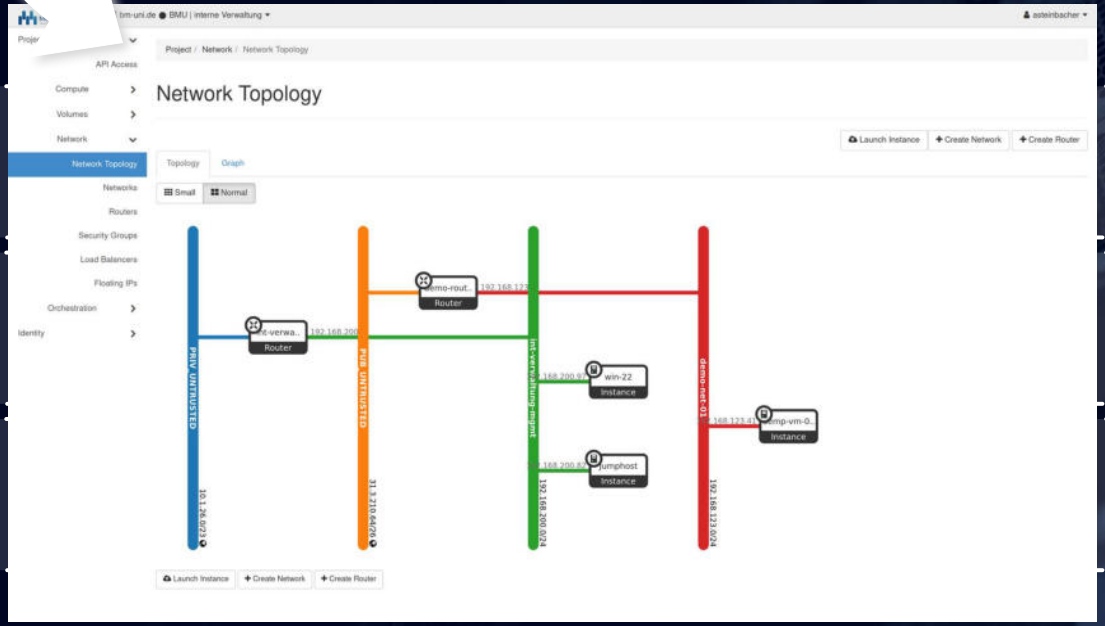
- Standarisierung = cool
- ein Tenant darf den anderen **nicht beeinflussen**
- Self-Service, wo immer es geht
- für den User ist die Cloud **einfach zu benutzen**
- → alle Funktionen **virtualisiert in der Cloud**

Overlay (Workload)

CPU-VMs

GPU-VMs

Netzwerke



LBaaS

DNSaaS

VPNaaS

Server

Switches

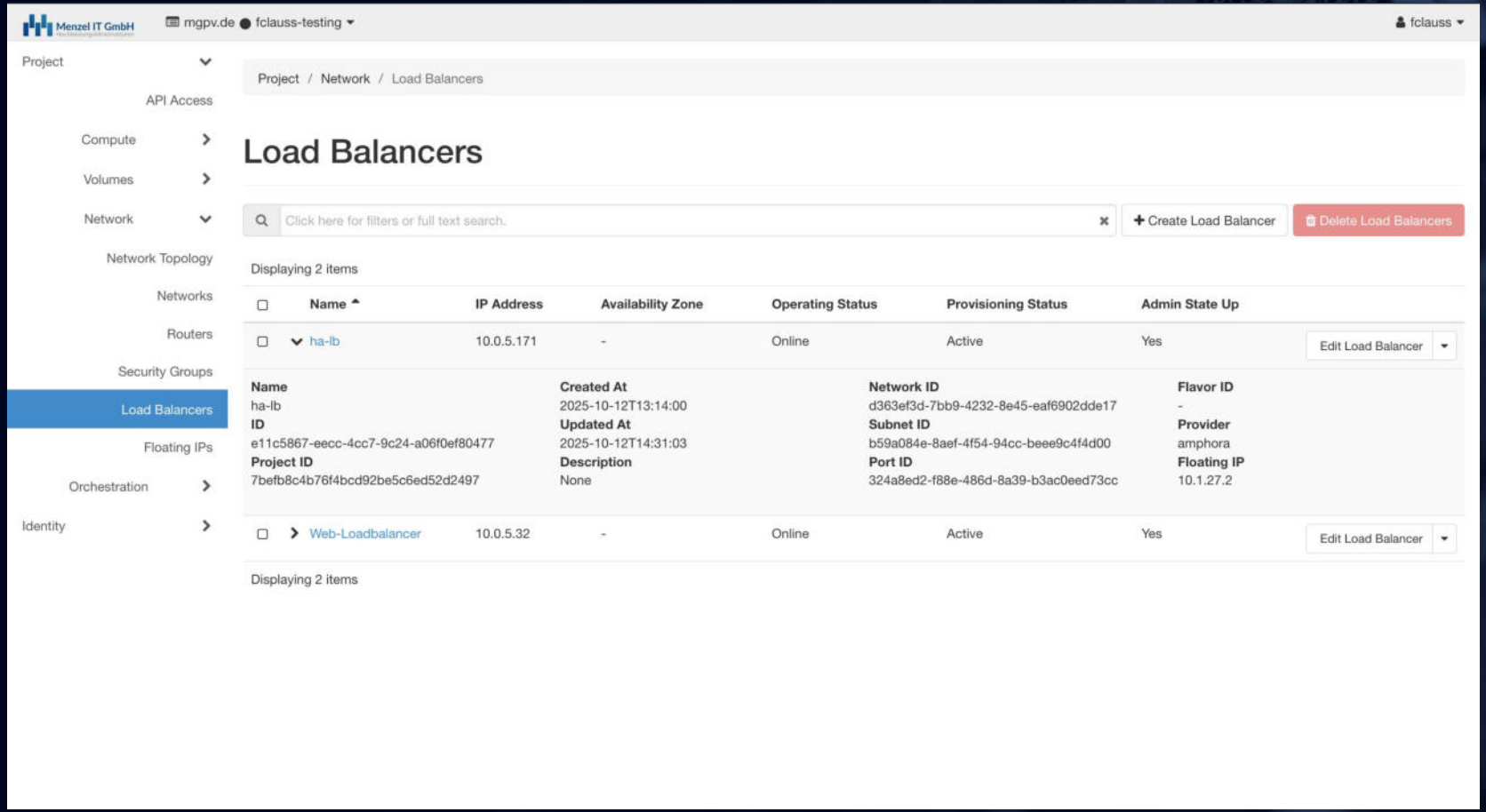
Storage

Strom

Underlay

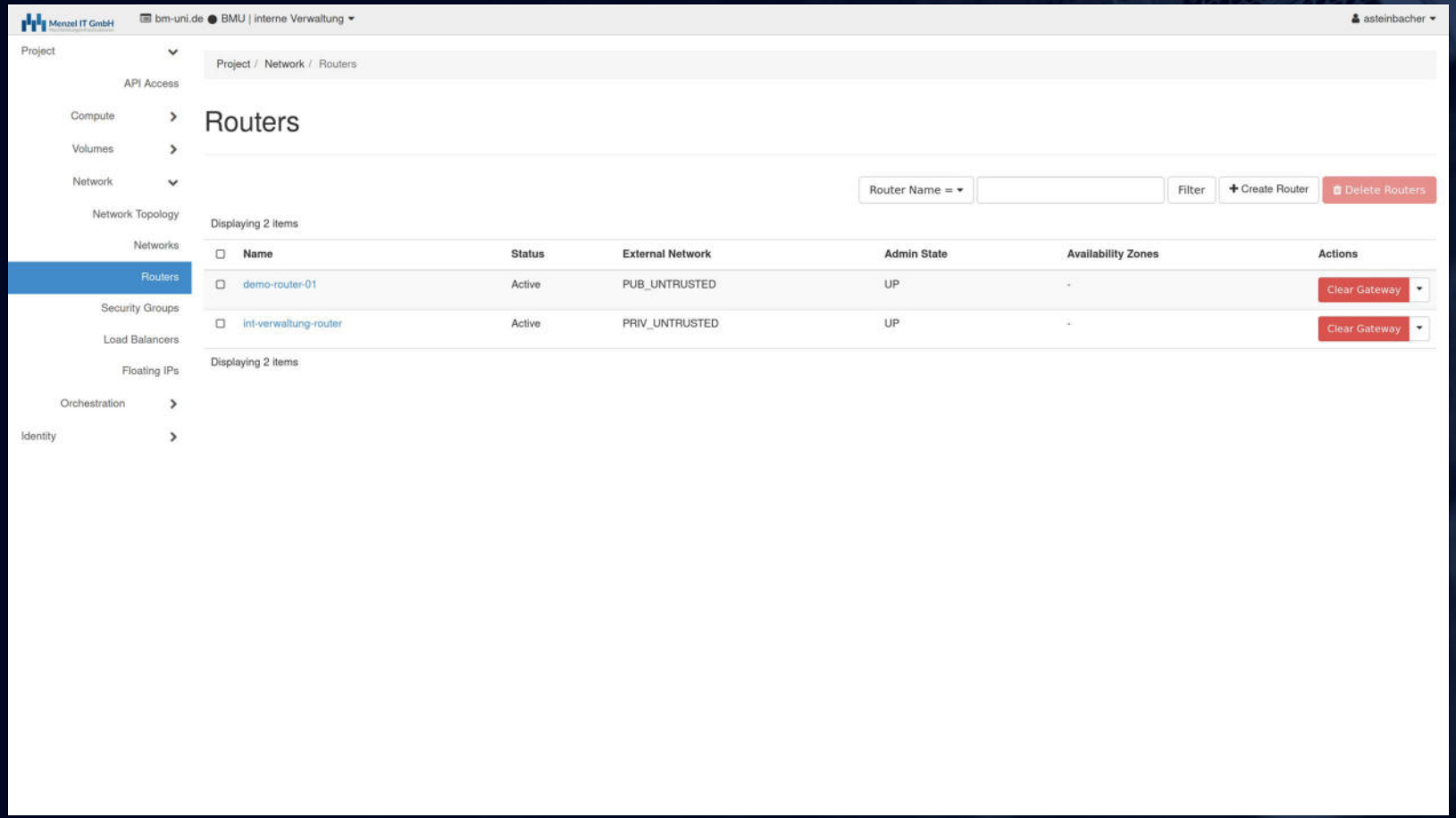
zusätzliche Services

Loadbalancer as a Service (Octavia)



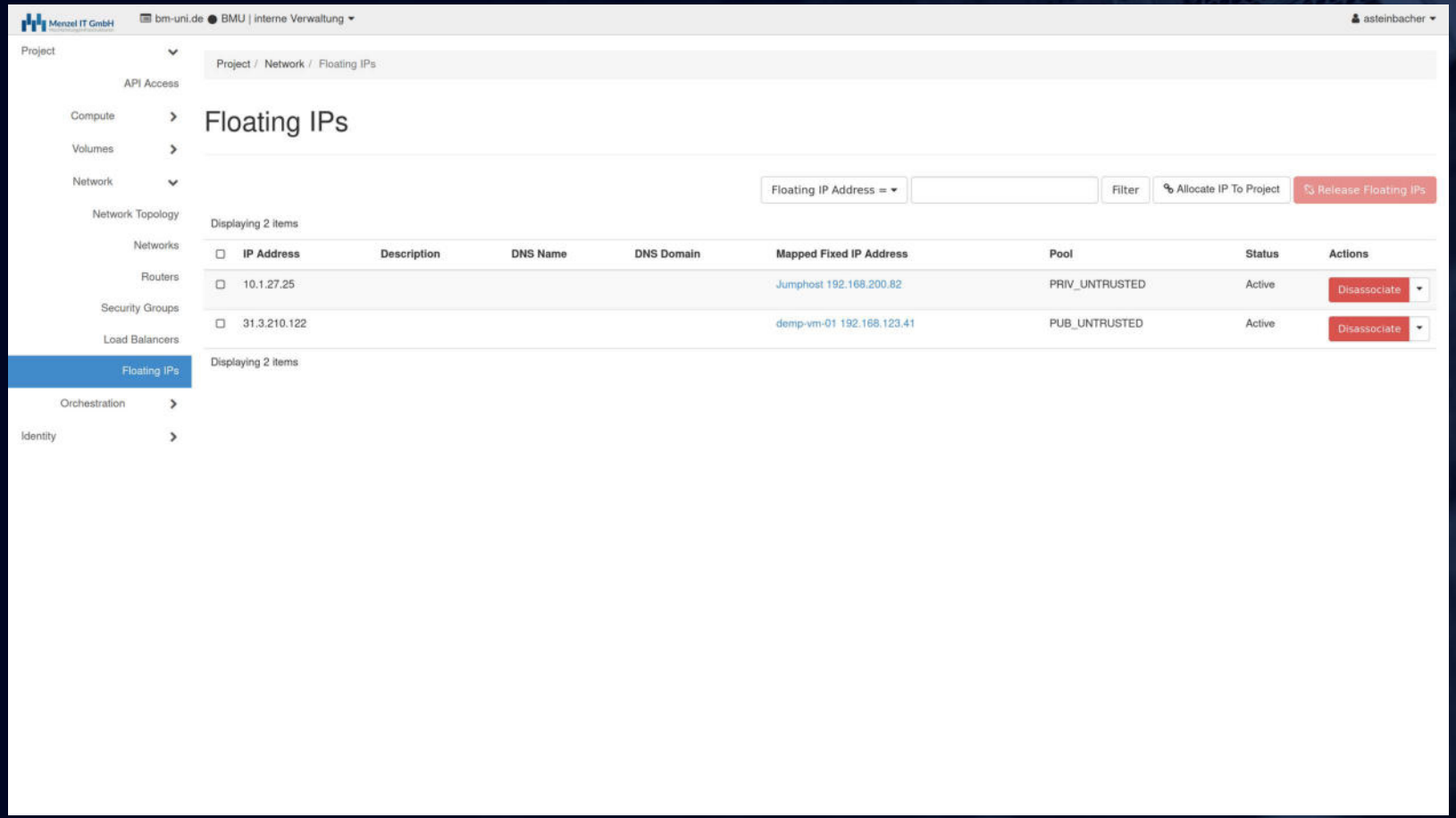
zusätzliche Services

Router



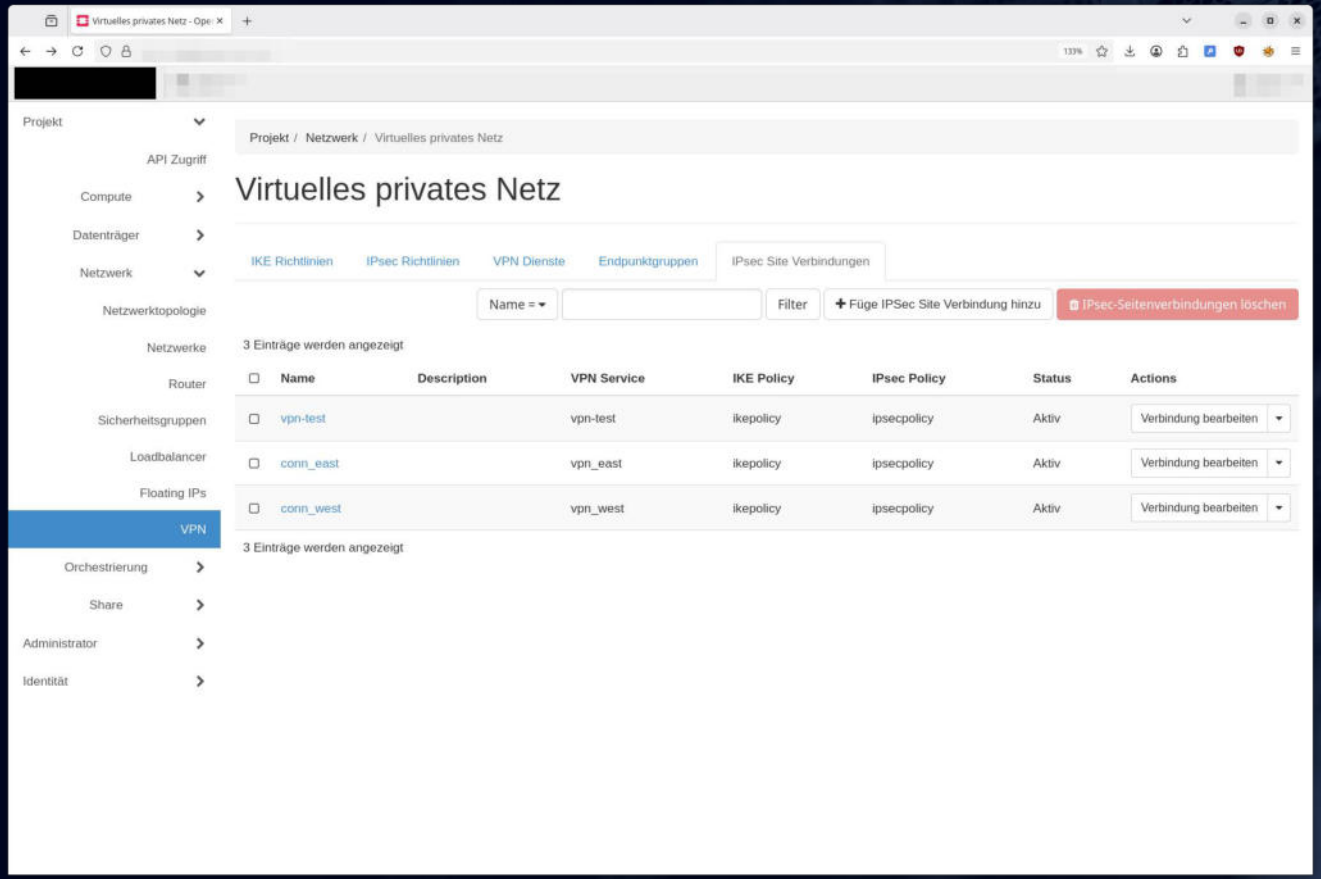
zusätzliche Services

Floating IPs (externe IP-Adressen an Ressourcen binden)



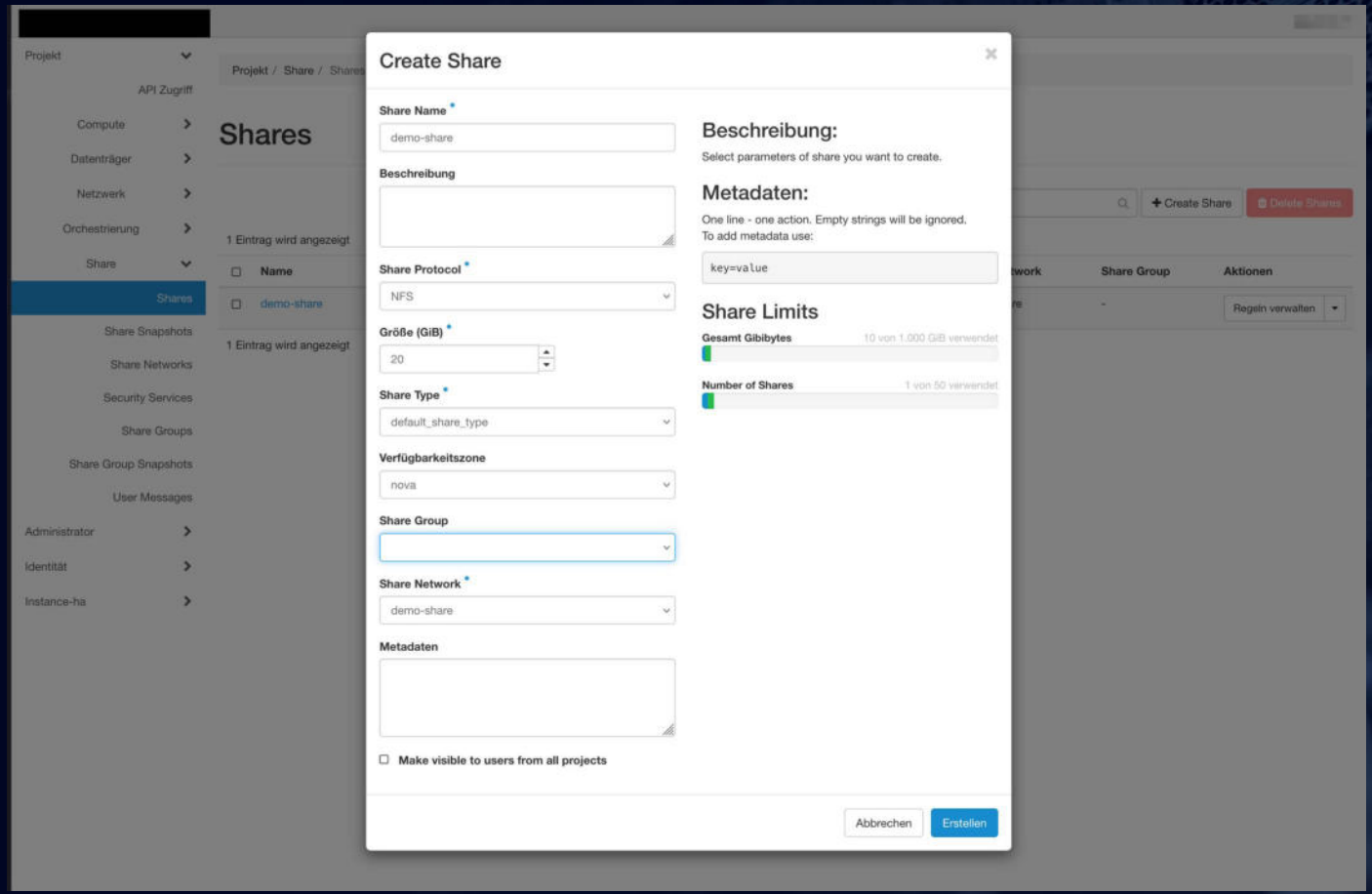
zusätzliche Services

VPN-as-a-Service



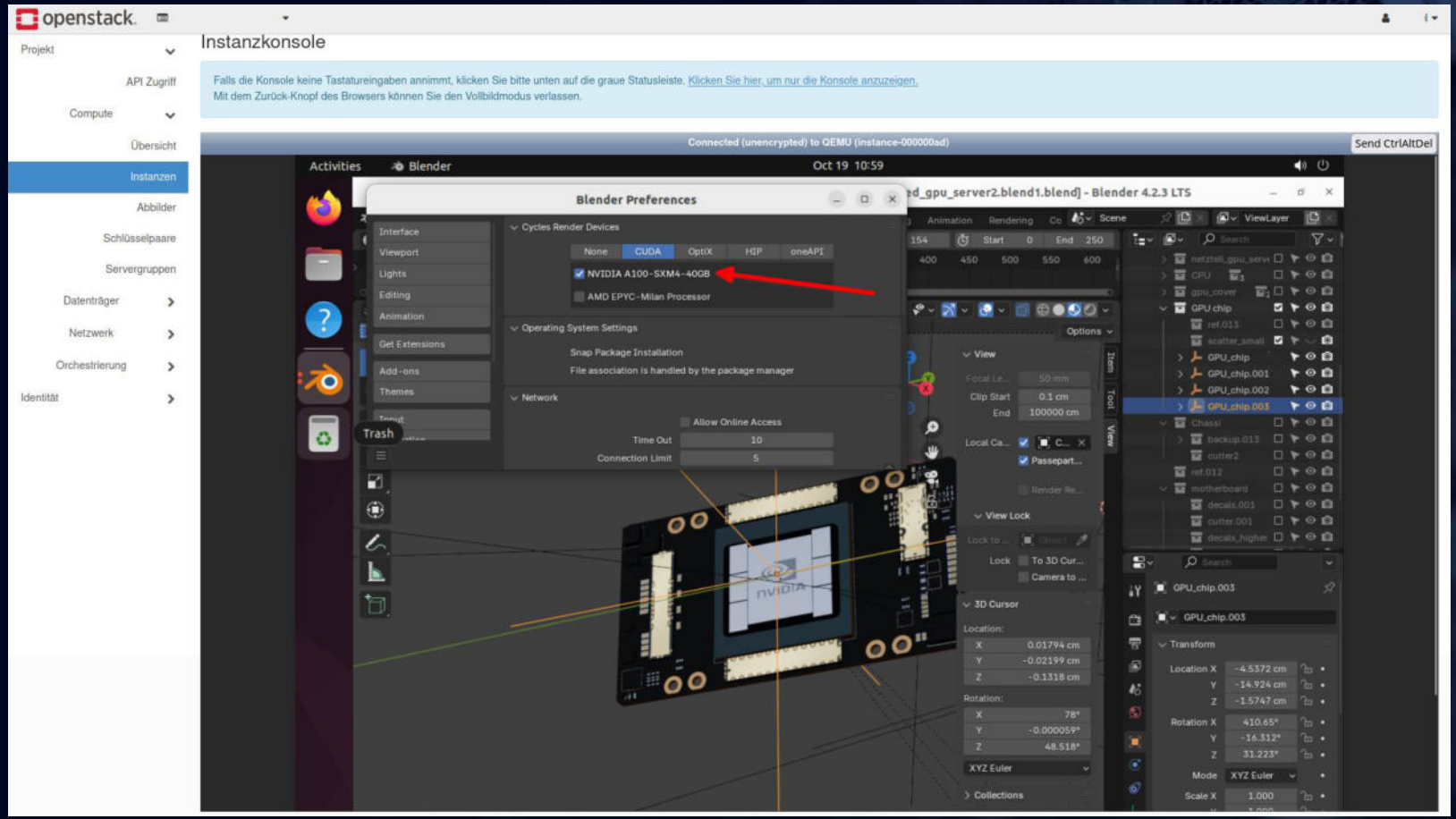
zusätzliche Services

Fileshare-as-a-Service (Manila)



zusätzliche Services

GPU-Flavors



Tuning I

- schnellere Hardware
- Ironic
- Container
- NUMA-Binding
- SR-IOV
- DPUs (Bluefield)

Tuning II

- die 3 Ebenen
  - KVM kann das
  - OpenStack kann das
  - kolla-ansible kann das
    (oder ein anderes Deployment-Framework)

Menzel IT GmbH

Charlottenburger Str. 33a
13086 Berlin

+49 30 5130 444 – 00
post@menzel-it.net
https://menzel-it.net

OPENNODE

Waldur for sovereign AI on OpenStack

Turning sovereign OpenStack infrastructure into a governed, self-service platform for AI workloads.

Ilja Livenson · OpenNode OÜ

OpenStack & sovereign cloud meetup
Hamburg · 2026

THE 60-SECOND VERSION

Who is talking, and about what

OpenNode

Open-source company in Tartu, Estonia. We build and support Waldur.

Waldur

An open-source (MIT) cloud and HPC management platform with a self-service marketplace. Generic by design, sharpened for private clouds and research computing.

~10 yrs
in production

MIT
licensed, no lock-in

1 portal
many backends

EU-built
yours to run

WHY THIS, WHY NOW

Sovereign AI stopped being a slogan

Four pressures are pushing AI compute back onto infrastructure you actually control:

Data residency

GDPR, the EU AI Act and NIS2 make where data and models live a legal question, not a preference.

Public investment

EuroHPC AI Factories put national GPU capacity online — and it has to be shared on Europe's terms.

Lock-in fatigue

Organisations want an exit. Open stacks and portable workloads beat proprietary AI clouds.

Accountability

Who accessed which model and data, under what terms — auditable, not assumed.

THE REQUIREMENTS

What “sovereign” actually demands of the stack

Control over data

Workloads stay on infrastructure you own and operate.

An open, auditable stack

No black boxes you cannot inspect or replace.

Self-determined access

You decide who gets in, to what, and on which terms.

Traceability

Every allocation, consent and usage event is recorded.

No lock-in

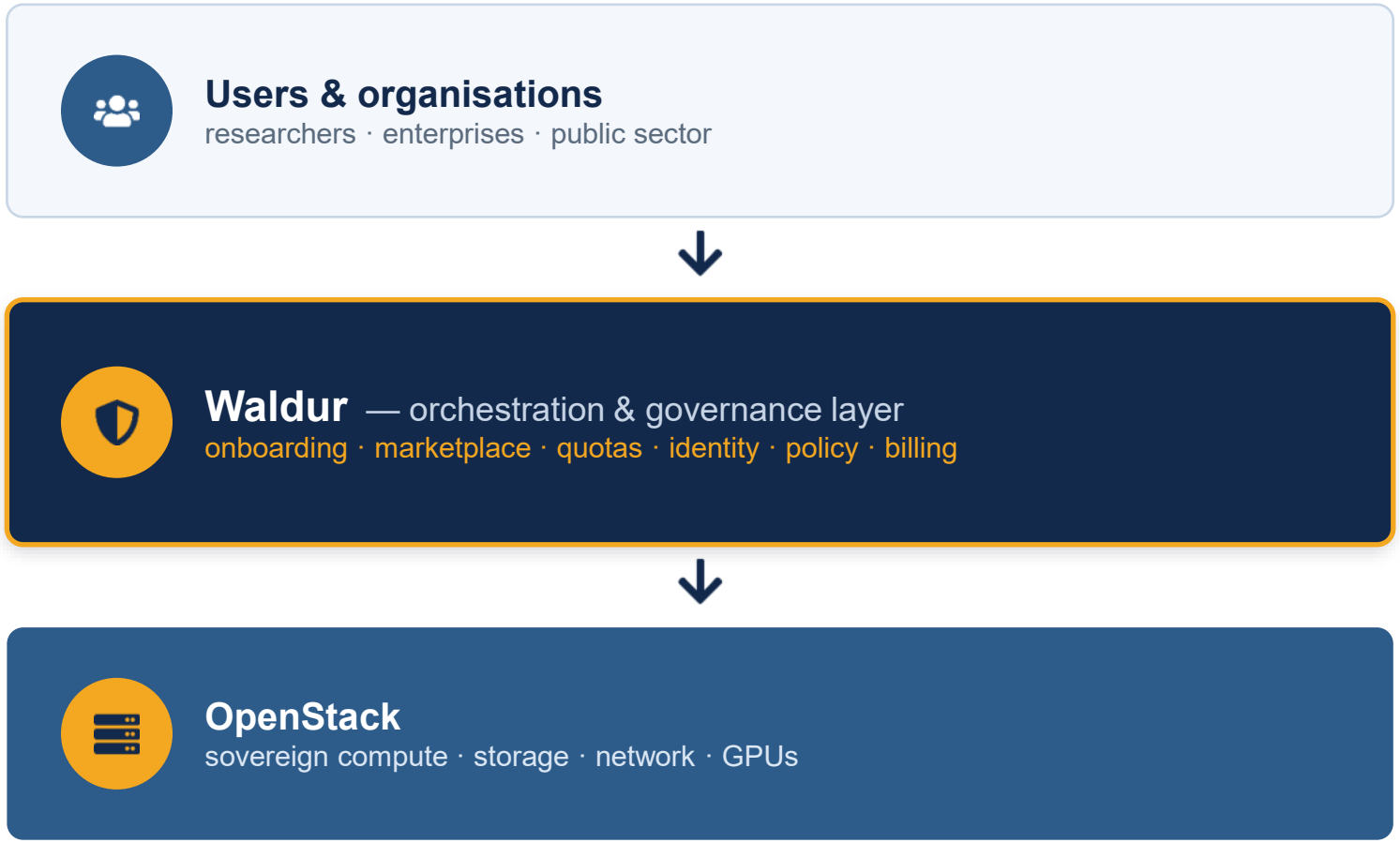
Portable workloads and a credible exit at any time.

Cost transparency

Real usage and billing per project, team and tenant.

THE CORE IDEA

OpenStack is sovereign IaaS — not yet a platform



The gap

OpenStack hands you tenants, flavours and volumes. It does not hand you a way to onboard users, approve and meter allocations, enforce policy, or bill — without tickets and tribal knowledge.

Waldur fills that gap and stays out of the data path: orchestration outside, policy inside.

THE PLATFORM

Waldur in one slide

A self-service marketplace and control plane that sits above your infrastructure.

Self-service marketplace

Catalogue of offerings users order from, with approval workflows.

Multi-tenant projects

Organisations, projects and roles map cleanly onto OpenStack tenants.

Metering & billing

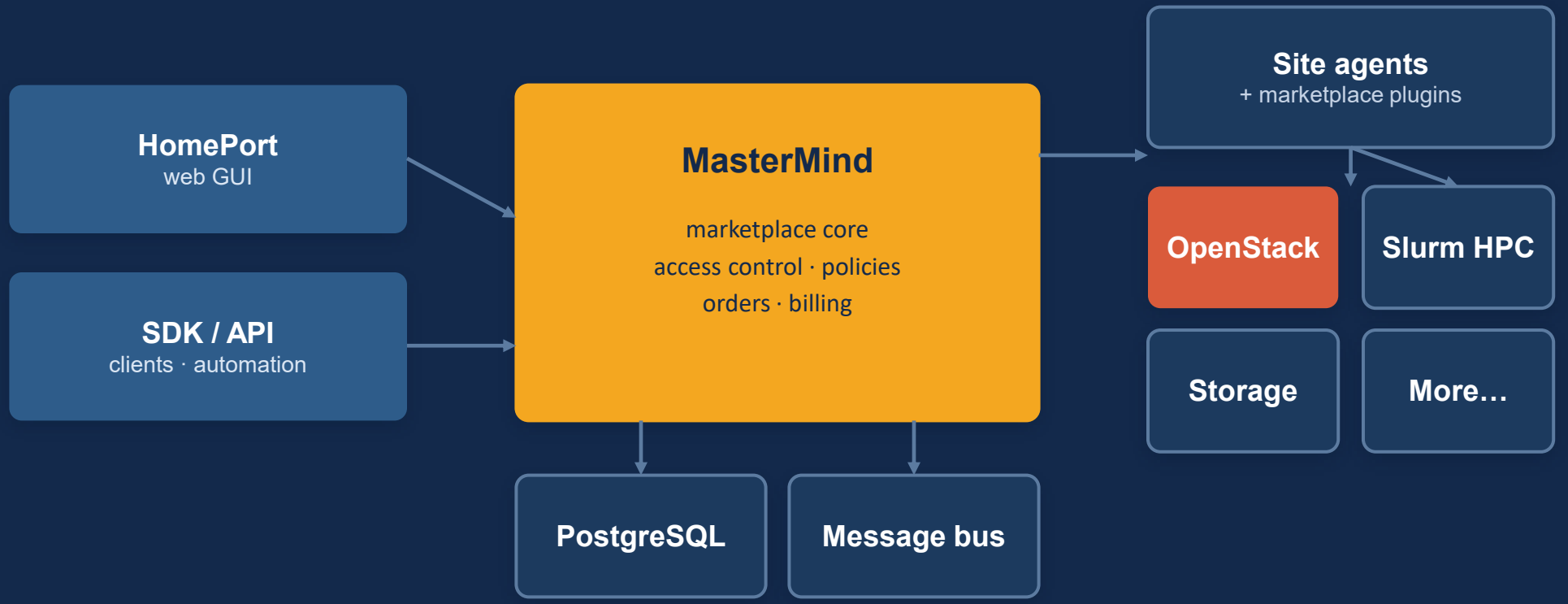
Usage tracked and invoiced per project, team and offering.

API-first, GUI included

Everything via REST; HomePort web UI, SDKs and config-management modules.

UNDER THE HOOD

Architecture: a control plane, not a data plane



Waldur drives backends through agents and APIs. Your data and your AI jobs never pass through it.

DAY-TO-DAY

From request to running OpenStack resource

Self-service replaces the ticket queue — with approvals and metering built in.

1

Order

A project member picks an offering and submits a request.

2

Approve

Org owner and provider approve against policy and quota.

3

Provision

The plugin creates the resource on OpenStack automatically.

4

Run

Resource is live; usage is metered continuously.

5

Bill

Consumption becomes invoice items per project.

THE AI ANGLE

Turning GPUs into governed allocations

AI capacity is scarce and expensive. Waldur makes it shareable on your terms.

GPU offerings

Publish GPU flavours and HPC partitions as orderable marketplace items.

Quota-based limits

Per-project and per-user limits on larger allocations, enforced at the backend.

Service accounts

Scoped, time-limited shared credentials for pipelines and agents.

Storage allocations

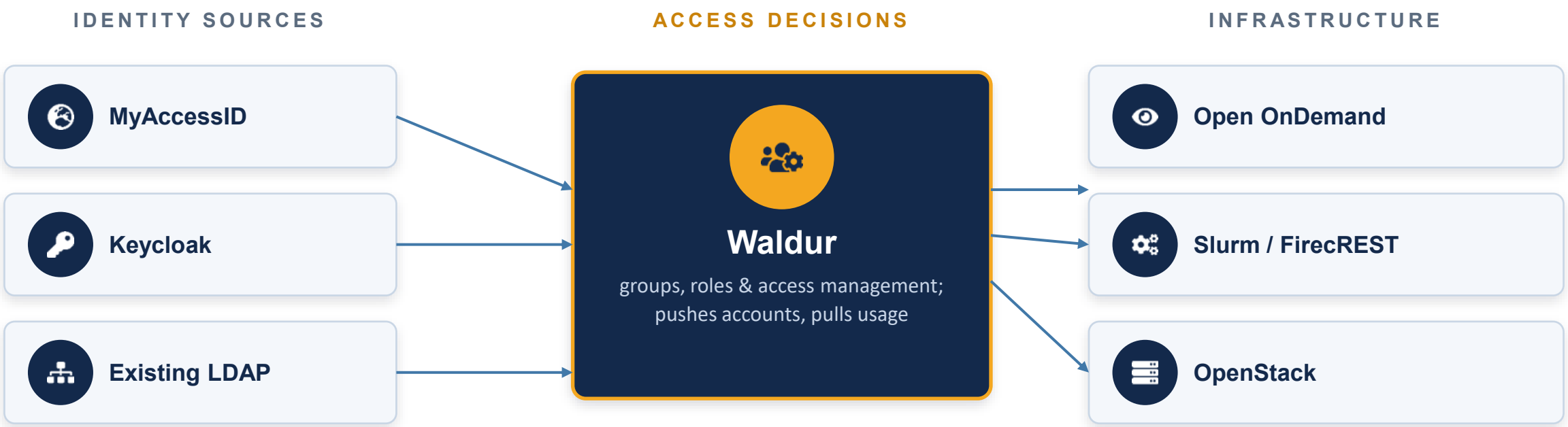
Request and visualise project storage in its own terms.

Also emerging — inference-as-a-service: order a model, get an OpenAI-compatible endpoint and a browser playground (provider-agnostic).

WHO GETS IN

Identity & access: zero-trust, federated

Waldur becomes the place where group membership and access are decided — not each backend.



Federated login in, consistent access out — across HPC, web portals and OpenStack alike.

POLICY INSIDE

Governance you can actually prove

The controls that turn a shared cluster into an accountable, compliant service.

Policy engine

Automated limits and actions on projects and resources.

Compliance checklists

Configurable questions to clear projects, users and proposals.

Terms of service

Offering-specific ToS with explicit, revocable consent.

Call & access review

Manual or automated review processes for requesting access.

IN PRODUCTION

Already running real sovereign infrastructure

Riigipilv

Estonian Government Cloud

National public-sector cloud on OpenStack, self-service via Waldur.

EuroHPC

Federation platform

Shared access portal across European supercomputers.

CSCS

Switzerland

National supercomputing centre.

Isambard-AI

Bristol, UK

One of the UK's largest AI supercomputers.

Cambridge

Dawn

Research GPU computing at scale.

Sovereign IaaS, plus a platform on top

OpenStack keeps your AI compute sovereign. Waldur makes it self-service, governed and accountable — open source, no lock-in.

MIT licensed

Policy inside

Orchestration outside

Ilja Livenson
OpenNode OÜ

opennodecloud.com
waldur.com · MIT on GitHub

Stay for the panel → "Two sides of the same cloud"

FURTHER READING

AI on Waldur & sovereign research infrastructure

1

Beyond Pre-Training: The Full Lifecycle of Foundation Models on HPC Systems

Conciatore, Oggian, Da Forno, Schuppli, Tissieres, VandeVondele & Martinasso (CSCS) **Cray User Group 2026 · arXiv:2604.12599**

Frames national HPC as the backbone of sovereign AI and the EuroHPC AI Factories — pre-training, fine-tuning and inference.

2

Isambard-AI: a leadership class supercomputer optimised specifically for Artificial Intelligence

McIntosh-Smith, Alam & Woods (University of Bristol) **Cray User Group 2025 · doi:10.1145/3725789.3725794**

Describes the MyAccessID → Waldur single sign-on and access workflow for the Isambard-AI portal.

3

Federated Single Sign-On and Zero Trust Co-design for AI and HPC Digital Research Infrastructures

Alam et al. **SC24-W Workshops · doi:10.1109/SCW63240.2024.00220**

The zero-trust, federated-identity design that Waldur implements across HPC and cloud.

APPENDIX · ON OPENSTACK

What Waldur drives in OpenStack

One OpenStack.Tenant offering, and users self-serve the full IaaS surface through Keystone, Nova, Cinder, Neutron and Glance.

Keystone

Identity

- Projects / tenants
- Users & credentials
- Quota sync

Nova

Compute

- Instances & flavours
- GPU flavours
- Console, affinity groups

Cinder

Block storage

- Volumes & types
- Snapshots
- Backups

Neutron

Networking

- Networks, routers
- Security groups, FIPs
- Load balancers (Octavia)

Glance

Images

- OS images
- Custom image upload
- Per-tenant limits

Quotas map 1:1 — cores, RAM, storage, instances, volumes, floating IPs — enforced in both directions.

Provision via web UI, REST, Python SDK, Ansible (waldur.openstack) or Terraform.

APPENDIX · FOR Q&A

OpenStack integration, in detail

Load balancing (Octavia)

- Octavia via openstacksdk — OVN provider only
- Listeners & pools: TCP / UDP; algorithm SOURCE_IP_PORT
- LBs, listeners, pools, members, health monitors
- Floating IP + security groups on the VIP port
- Background reconcile every 2 hours

Authentication & sessions

- Password or v3 application credential
- App credential: username = ID, password = secret
- Scoped & revocable — recommended for production
- Keystone sessions cached ~10 h, auto-refreshed

Storage modes & quotas

- Fixed mode: one shared storage quota
- Dynamic mode: per volume-type quotas (SSD / HDD)
- Quotas reconciled in both directions
- Defaults e.g. 20 cores, 50 GB RAM, 30 instances, 1 TB

Networks & sync

- External networks: provider-global or RBAC-exposed
- Router gateway: two-tier perms (SNAT / fixed IPs)
- Periodic pulls keep Waldur ↔ OpenStack consistent
- Clients: Keystone v3 · Nova v2.19 · Cinder v3 · Glance v2 · Neutron v2.0

XAAS FROM IDEAS TO IMPACT.

VIELEN DANK FÜR IHRE ZEIT!

XAAS GmbH & Co. KG
Kobestraße 7, 20457 Hamburg
T: (+49) 40 / 22 86 90 40
E: hallo@xaas-it.com