Our Journey to a Modern, Self-Managed Research Platform

Maarten Wullink - SIDN Labs

CENTR R&D 2026-02-24

SIDN Labs

Research into improving the security and resilience of the .nl domain and the wider internet

Applied internet research

- DNS, BGP, NTP

- Active and passive measurements

- Software prototyping

Why do we need a new research platform?

Multiple reasons:

- Aging infrastructure (hardware and software)
- SIDN is migrating large parts of its infrastructure to the cloud
  - The Labs infrastructure is dependent on the SIDN network
- New demands from research projects
- Increase isolation from SIDN production
- Desire to use more modern technologies

Old Infrastructure

The starting point of our journey, aging infrastructure, 3 main components

- Data platform
- Network testbed
- VM platform

This setup suffers from hardware, software and connectivity issues

Hardware Issues

Most hardware is at, or well over EOL

- Financially written off all hardware (2017–2020 servers)

- Some servers are very old repurposed production hardware

- No vendor support anymore, or at much higher cost

- Low(er) power efficiency

- Struggling to keep up with current workload

Software Issues (1/2)

Hadoop has served us well for many years, however it is increasingly showing its age.

- Slow(er) pace of development

- Becoming a bottleneck for new research

- No OSS management tool (anymore)

- Difficult to manage

- Limited flexibility

Software Issues (2/2)

VMware

- Negative licensing changes

- Enormous cost increases

- Not OSS

Connectivity Issues

Current infrastructure is mainly hosted at the SIDN office DC

- No direct Internet Exchange access

- Limited (real time) BGP data access

- Research limitations

Requirements

Our requirements are captured into multiple groups

- Labs Principles

- Security

- Performance

- Adaptability

- Manageability

- Cost Predictability

Labs Principles

Core principles of SIDN Labs

- Decentralisation
  - No US Hyperscalers
- Digital autonomy
  - Use NL/EU hardware and services
- Open
  - Open Source
  - Open Standards

Security

Data integrity over data availability

- Downtime is acceptable

- Data must always be safe (incl .nl data)

- Controlled and Centralised Access Management

  - Single Sign On (SSO)

- Secure network

  - Segmentation

  - Firewalls

  - monitoring

Performance

We need a scalable platform

- Efficiently analyze large datasets

- Interactive use cases

- Advanced analytics

  - Trino and Spark

- Parquet columnar data format

- Modern hardware

  - NVME based storage

  - 100Gb+ networking

Adaptability

Ability to add new services or tools when desired

- Flexible architecture

- 100% control of infrastructure

- Support different types of research projects
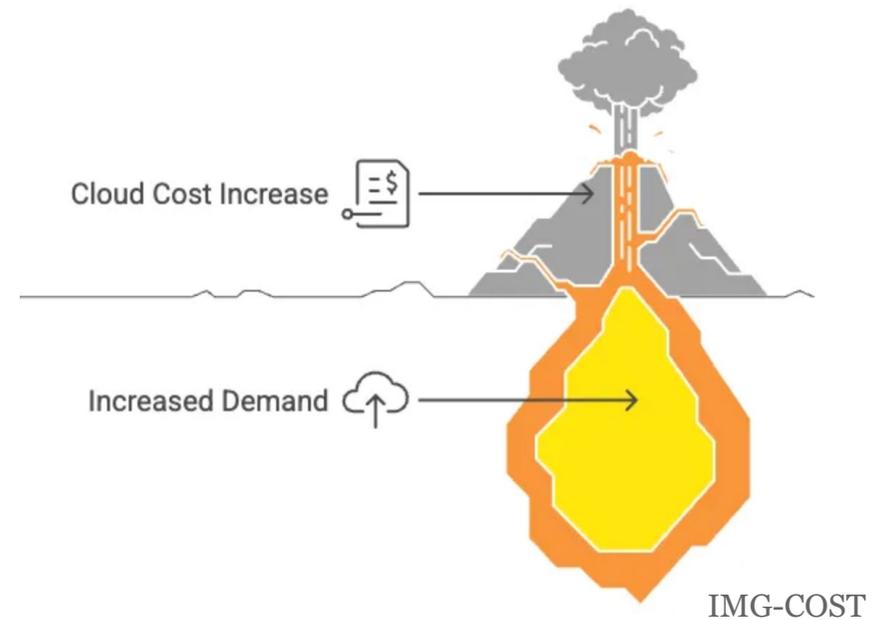
Manageability

Researchers should work on research not play sysadmin

- Low(er) maintenance effort
  - Container vs. VM
- Dedicated support colleagues (1FTE)
  - Professional sysadmins

Cost Predictability

We want researchers to experiment freely without feeling constrained by potential high costs.

- No desire for cloud cost engineering

- Predictable costs

- Prefer CapEx over OpEx

- No cloud bill shock

Cloud Cost Increase

Increased Demand

IMG-COST

IMG-COST: https://medium.com/@chintan.j72/cloud-cost-optimization-how-to-reduce-your-cloud-bills-9869dd942732

The Plan

- Retain the 3 distinct components
  - 1) Data platform
    - Replace Yarn with Kubernetes for container orchestration
    - Replace HDFS with s3 for data storage
  - 2) VM platform
    - Replace VMware with Proxmox for virtualisation
  - 3) Advanced networking
- Buy dedicated hardware
- Build services using a open source stack
- Use NL data center

Design

2 data center racks and redundant network connectivity
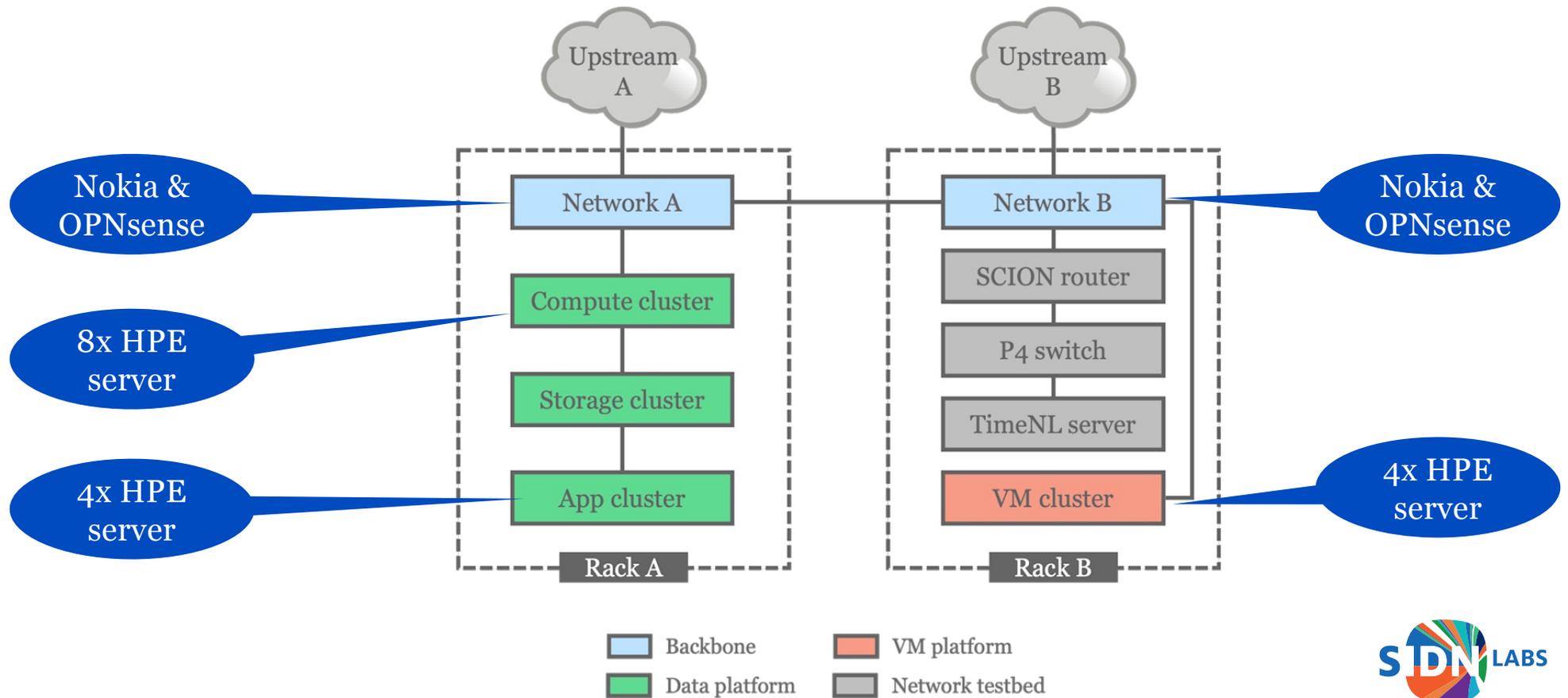
Data Platform (1/3)

- Kubernetes
  - 12 nodes (8 data/ 4 app nodes)
  - Talos Linux
    - Minimal Linux
    - Embedded Kubernetes
  - Deployment
  - Ansible
  - Management
  - Rancher

CPU ⌃⌄	Memory ⌃⌄	Pods ⌃⌄
1151.4 cores	2.44 TiB	236/1320

Data Platform (2/3)

Lakehouse architecture

- Apache Iceberg open table format

- Lakekeeper REST Catalog

  - Metadata management

  - Access management

  - Trino and Spark compatible

- S3-compatible storage

- ~~MinIO~~ RustFS

  - Fast storage network

Data Platform (3/3)

Kubernetes platform also used for additional services

- Tools
  - JupyterHub
  - Airflow
  - Superset
  - ...
- Custom applications
  - ENTRADA2

VM Platform

Virtualisation is used for non-Kubernetes workloads

- Proxmox hypervisor

- Cluster of 4 nodes

- Services

  - Access management

  - Log management

  - K8s control plane

  - ...

Network

- Our own ASN (AS215088)
- Excellent connectivity
  - Upstream: SURF (NREN)
  - IX-ACCESS: AMS-IX, FrysIX
  - Cheap (one-time fee) cross connects to collaborators.
- Redundant backbone & firewalls

Results

We spent a year preparing, designing and building

- 100% independent infrastructure

- Modern cloud native platform based on Kubernetes

- Future-proof

- Predictable costs

- Excellent connectivity

- Positioned in the hart of NL internet

Lessons Learned

- Building custom infrastructure in 2026 is still doable

- Prepare to spend serious time on:

  - Requirements gathering

  - Scenario evaluation

  - Hardware selection and ordering

  - Deployment

- Researchers != sysadmins

- Principles come at a price!

Note: The AI craze has set (enterprise) hardware prices on fire

Services Overview

Thank You

www.sidnlabs.nl | stats.sidnlabs.nl

SIDN LABS