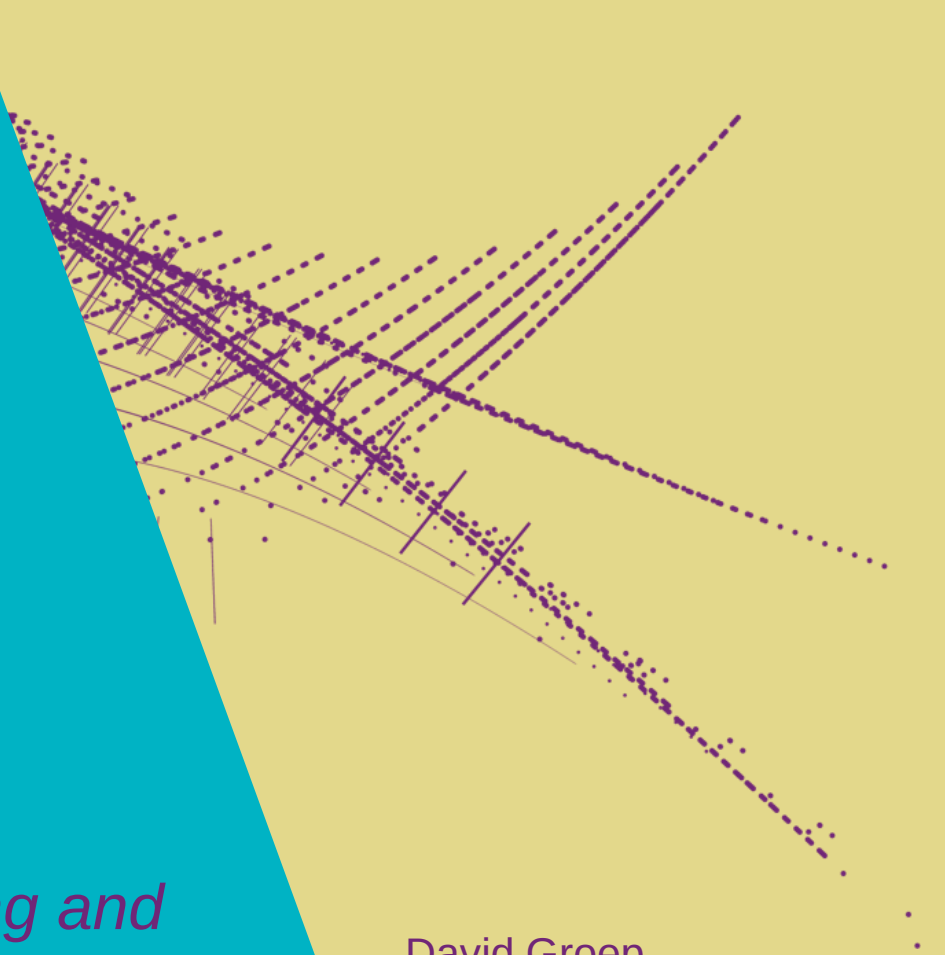
Nikhef



# From Iron to Service,

*a glimpse at the Nikhef computing and Data Processing infrastructure*

David Groep,
October 2024

# Starting with the basics …




Nikhef

# DC infrastructure: power, cooling, racks, and organisation



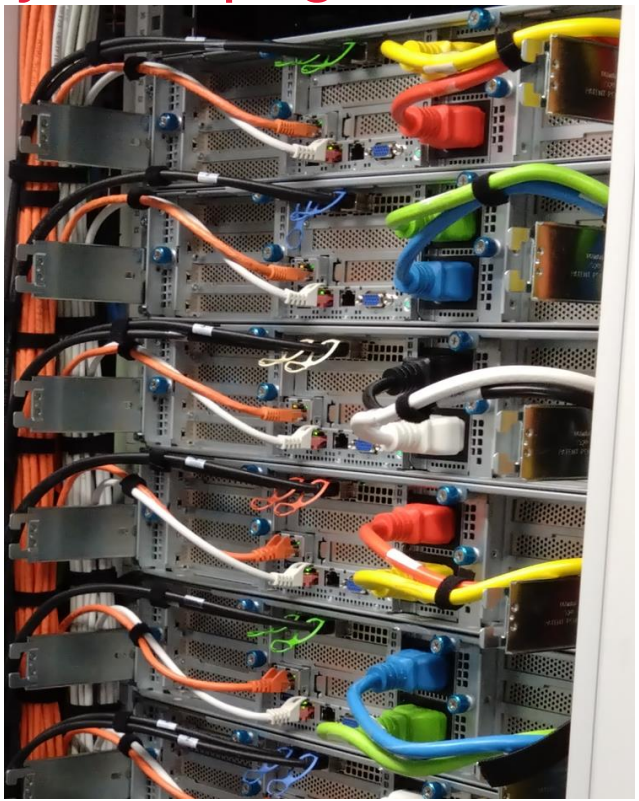


Left-side image: frame from a movie by Anton Mors, people replaced by … Adobe Firefly (and this was its best result ☹)

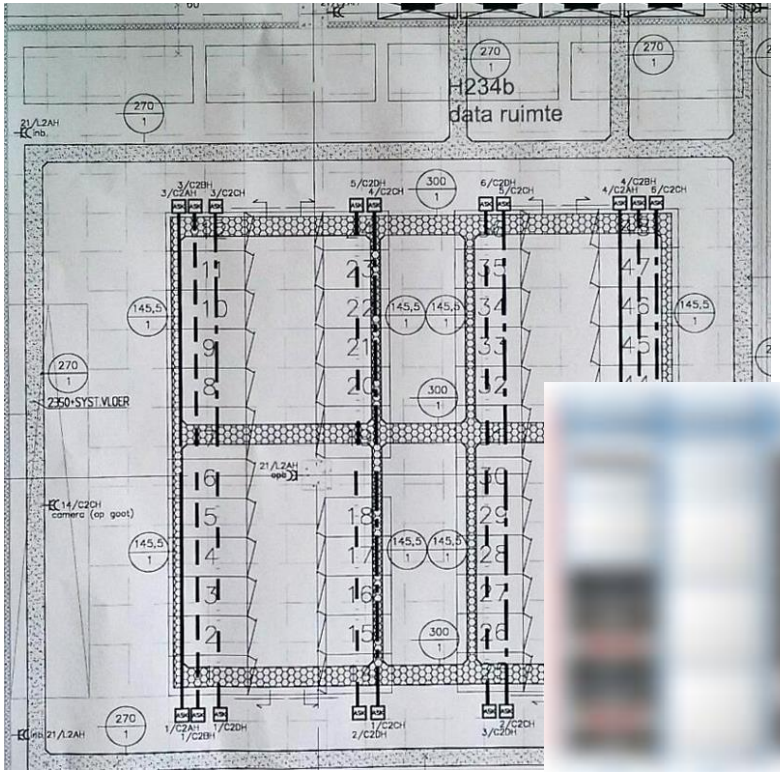


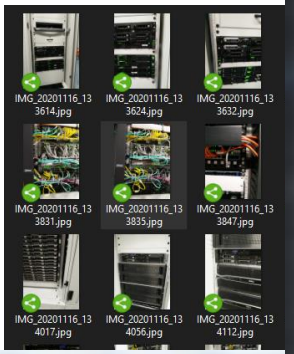
Nikhef

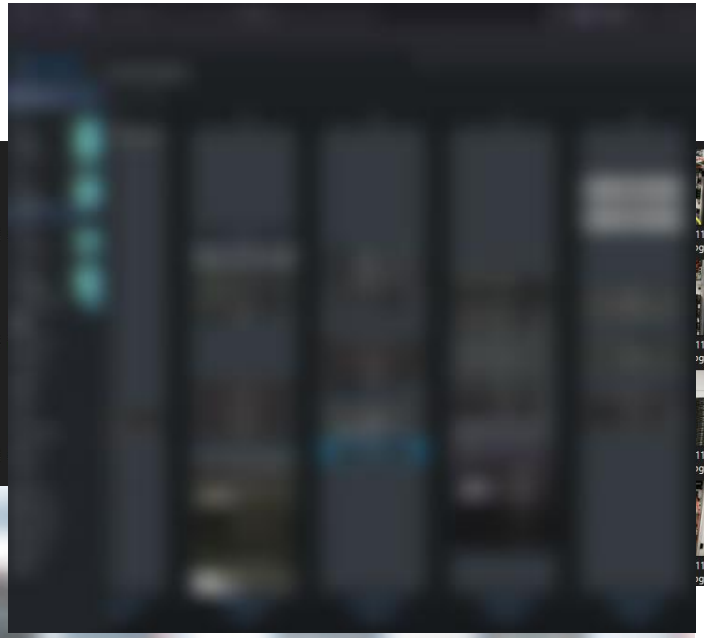
# the art of infrastructure construction: beyond spaghetti

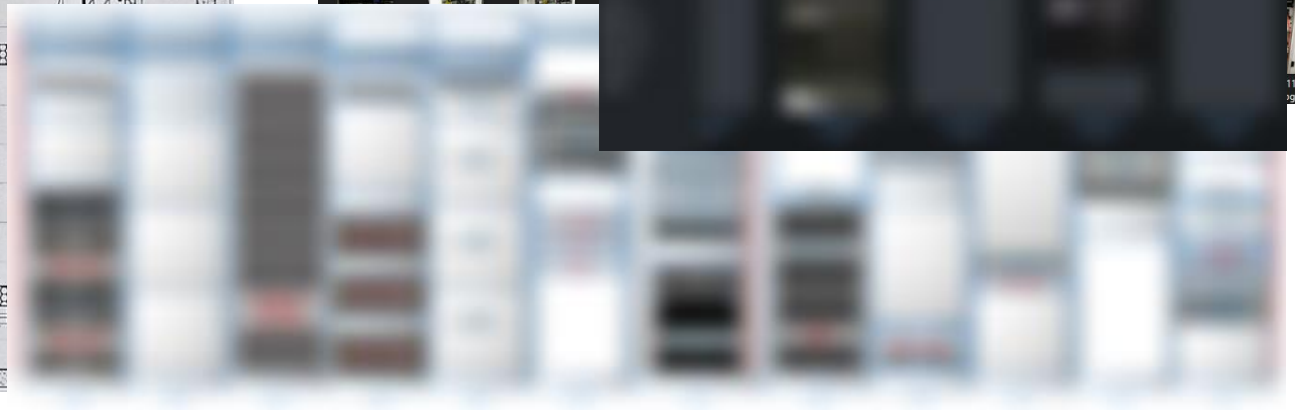


# Data Centre: layout and floor plans



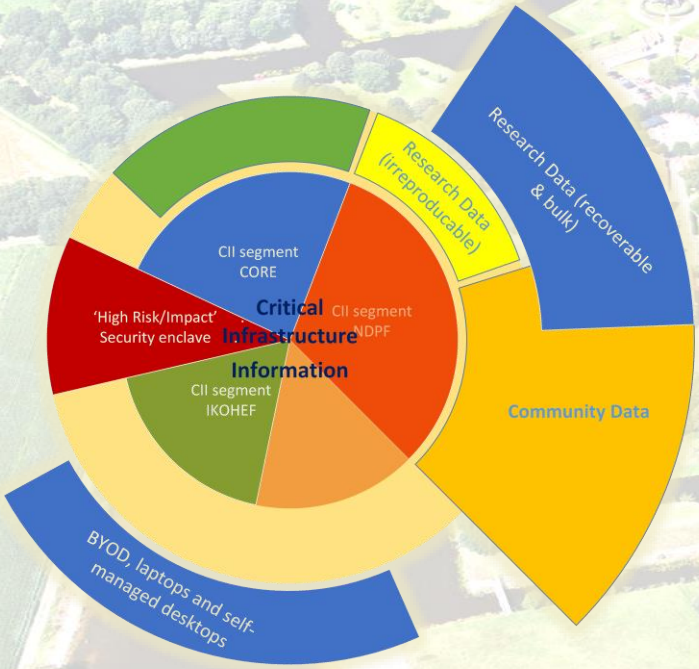


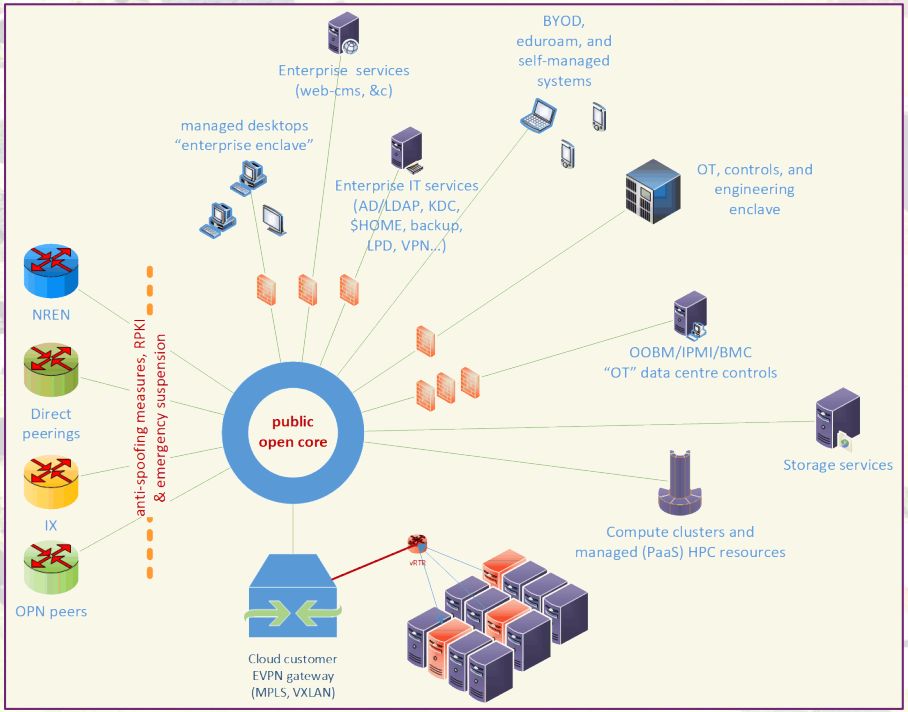


Nikhef

# Systems physical grouping by functional & security zone

# Network structure design: logical and topological view
## *segmentation: a research network with office enclaves*



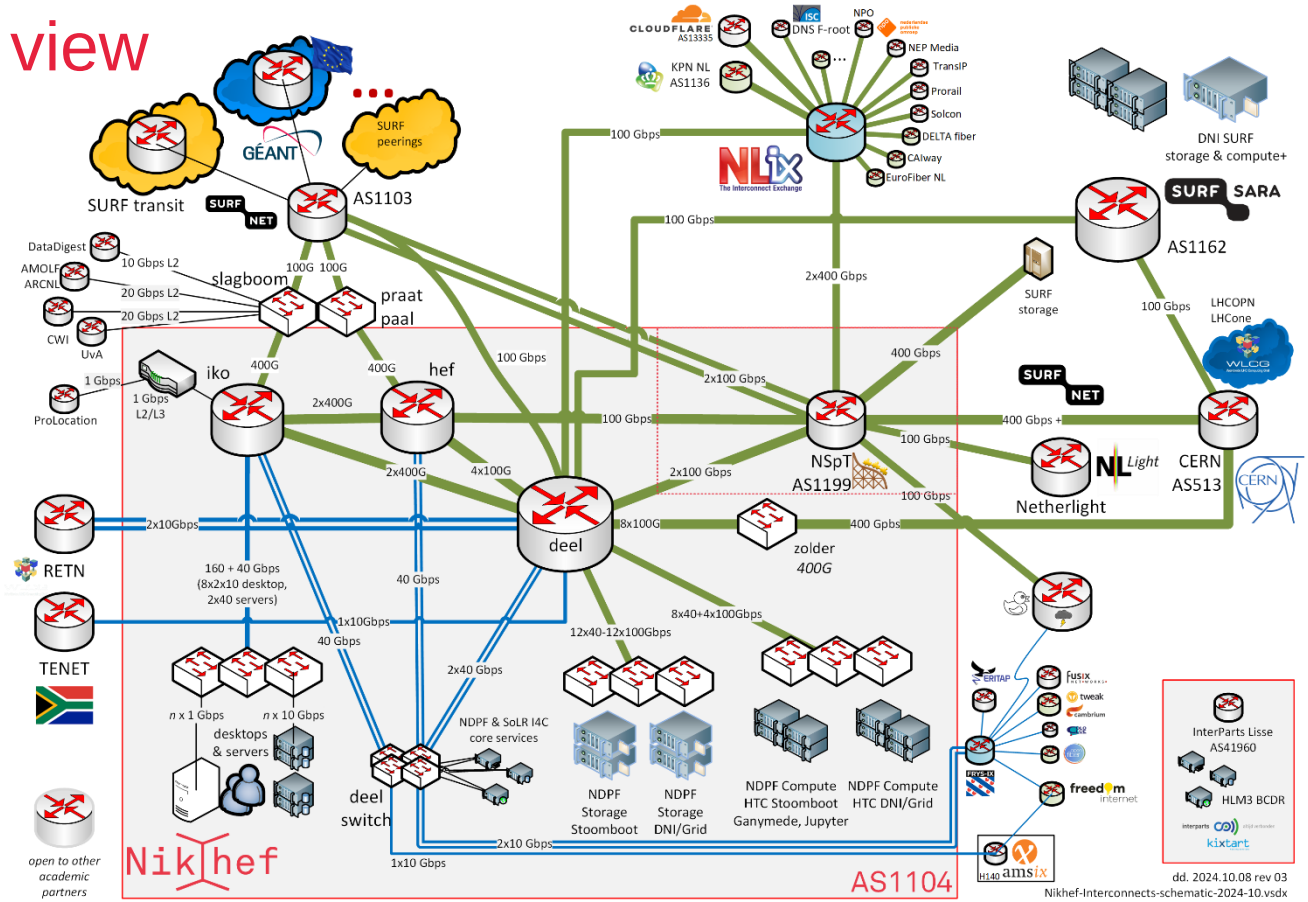




achtergrondbeeld: Vesting Bourtange

# Segmentation by access class and function

Assignments from RIPE LIRs nl-fom and SN-LIR and their subclassing in AS1104

# Connectivity view




AS1104 and adjacent systems in October 2024

# Capirca – vendor agnostic rule expression

```
term block-csirt-denied-hosts-emergency-suspension {
    destination-address::
        net-nik-oexp-blocked
    logging:: local
    counter:: ndpf-oexp-csirt-block-dest
    action:: deny
}
term pxeboot-bootps-client {
    source-address::
        net-ndpf-fabman-installservers
        net-local-ifaddr
    protocol:: udp
    source-port:: dhcp4-client dhcp4-server
    destination-port:: dhcp4-client dhcp4-server
    action:: accept
}
term pxeboot-tftp-client {
    source-address::
        net-ndpf-fabman-installservers
    protocol:: udp
    source-port:: unprivileged-ports
    destination-port:: unprivileged-ports
    action:: accept
}
term specific-deny-dangerous-ports-for-nik-xo-services {
    protocol:: tcp udp
    destination-port:: nfs smb-udp-tcp portmap
    action:: deny
}
term allow-localnets-access-to-most-nik-xo-services {
    source-address::
        nikhef-private
        net-nik-ournets
        net-ndpf-ournets
    protocol:: tcp udp
    destination-port:: openvpn-proxy-bypass ssh-both web-services unprivileg
    action:: accept
}
term rsync-bckup {
    source-address::
        net-ndpf-rsyncbackup
```

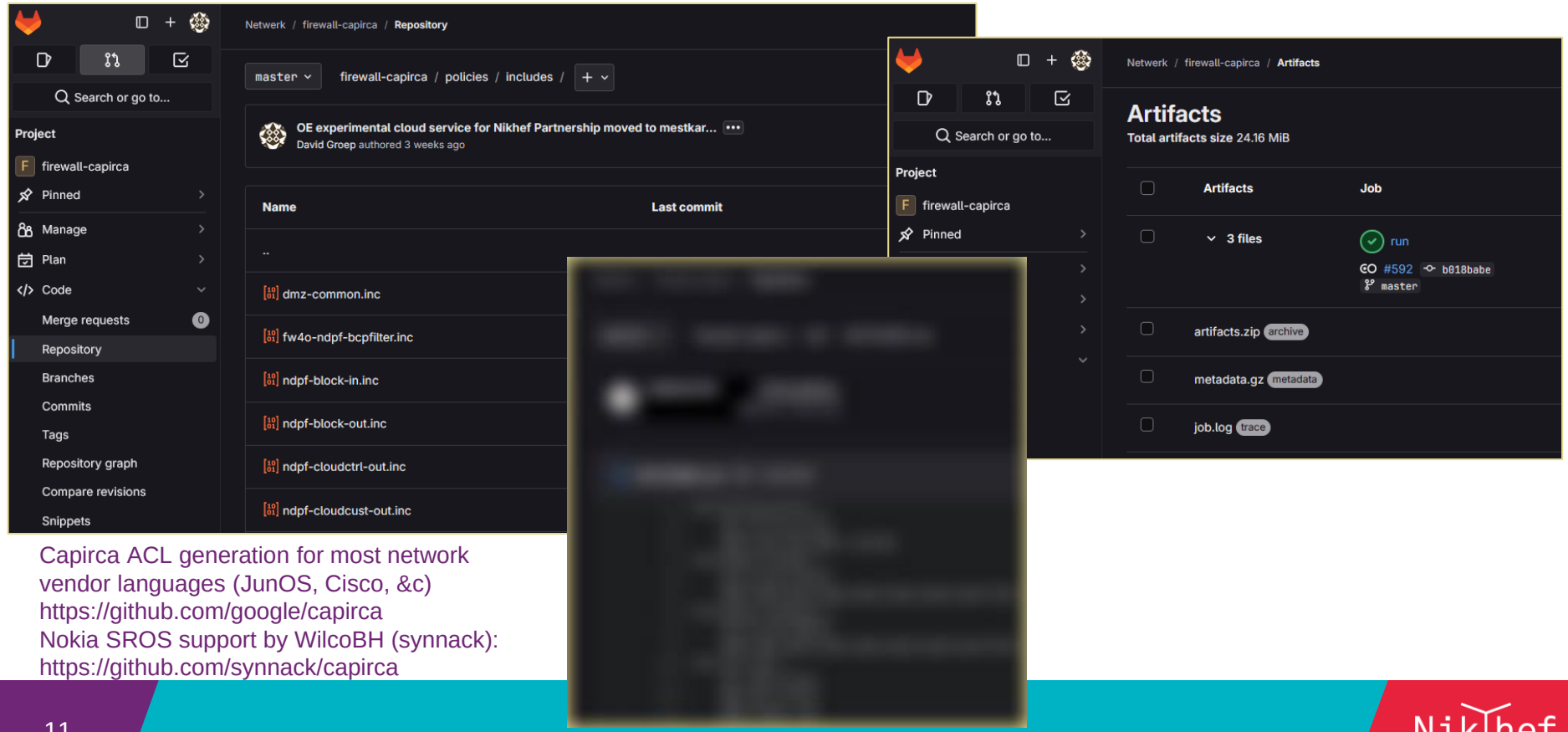
```
dmz-common.inc                ndpf-gridsrv-out.inc      ndpf-p4ctb-ou
fw4o-ndpf-bcpfilter.inc       ndpf-installnet-out.inc   ndpf-pubgrid-
ndpf-block-in.inc             ndpf-ipmi-out.inc         ndpf-secmon-o
ndpf-block-out.inc            ndpf-iscsi-out.inc        ndpf-stud-1-i
ndpf-cloudctrl-out.inc        ndpf-nik-xo-la-in.inc     ndpf-stud-1-o
ndpf-cloudcust-out.inc        ndpf-nik-xo-la-out.inc    ndpf-surfsoil
ndpf-clouddataplane-out.inc   ndpf-oexp-in.inc          ndpf-wn-out.i
ndpf-cloudinstall-out.inc     ndpf-oexp-out.inc         system-out.in
```

Capirca extended for SROS by synnack (https://github.com/synnack/capirca), forked from https://github.com/google/capirca



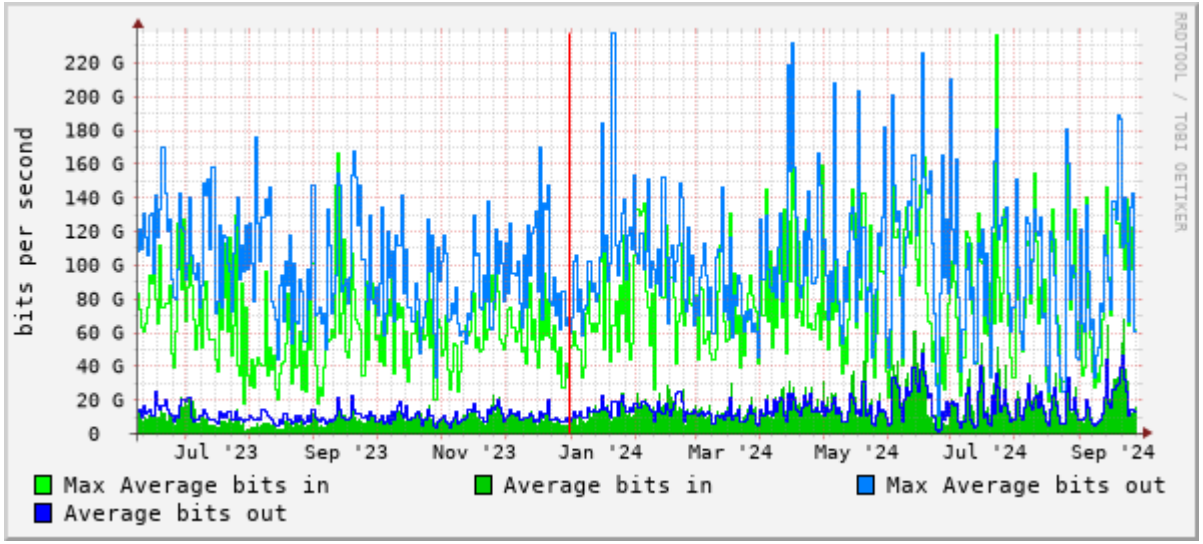
Nikhef

# Managed systems … managed networks




Capirca ACL generation for most network vendor languages (JunOS, Cisco, &c)
https://github.com/google/capirca
Nokia SROS support by WilcoBH (synnack):
https://github.com/synnack/capirca

# Networks at scale: bandwidth to compute and storage




Compute-storage-transit traffic DeelQFX aggregate (data volume as measured by harbours)

# Basic infrastructure: ‘every rack should have one’

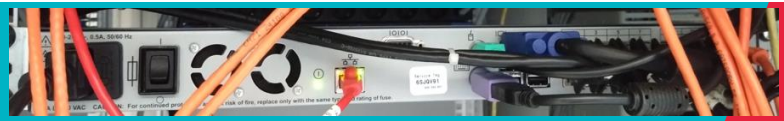
Each rack comes with a couple of ‘standard’ elements

- remotely monitored PDUs
- temperature sensor(s)
- VGA+HID+RS232 switch access

- 1x OOBM GigE ‘dumb’ switch
- 1x GigE installnet (managed)

- 1-2x 10/25/100G
  data + storage networks


Shown: H234b C06 ‘SOC’ cabinet, C08 rKVM





Nikhef

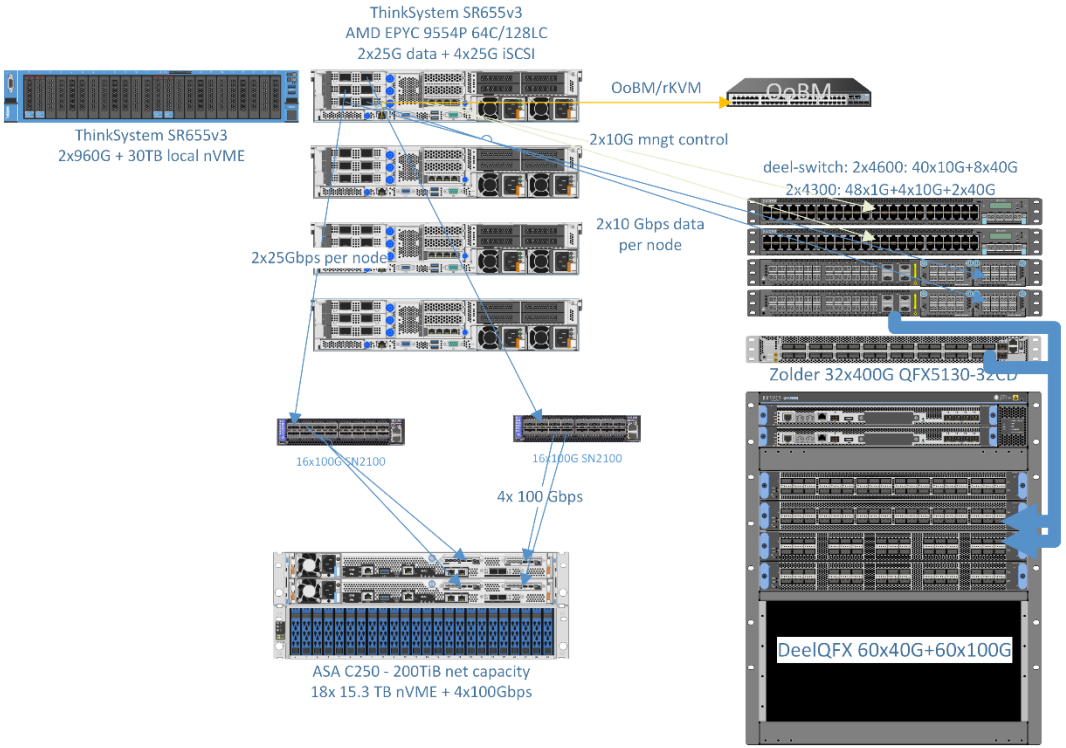
# Service clusters: ‘minimally redundant operations basis’

‘Typical’ node has some CPU and limited local system disk, plus:

| | | |
|---|---|---|
| 2x 10/25G | VM data net | 802.3ad LACP config |
| 2x 10/25G | storage network | iSCSI with 2 independent fabrics |
| 3x 1000BASE-T | install/management | 1x for initial install, 2x in 802.3ad |
| 1x | OOBM | IPMI & Redfish, rKVM and remote media |

Nikhef H234bC08: orkestbak OEXP VM service pair. Other service clusters (GSP) are similar. FabMan and PMASec may have local-store only.

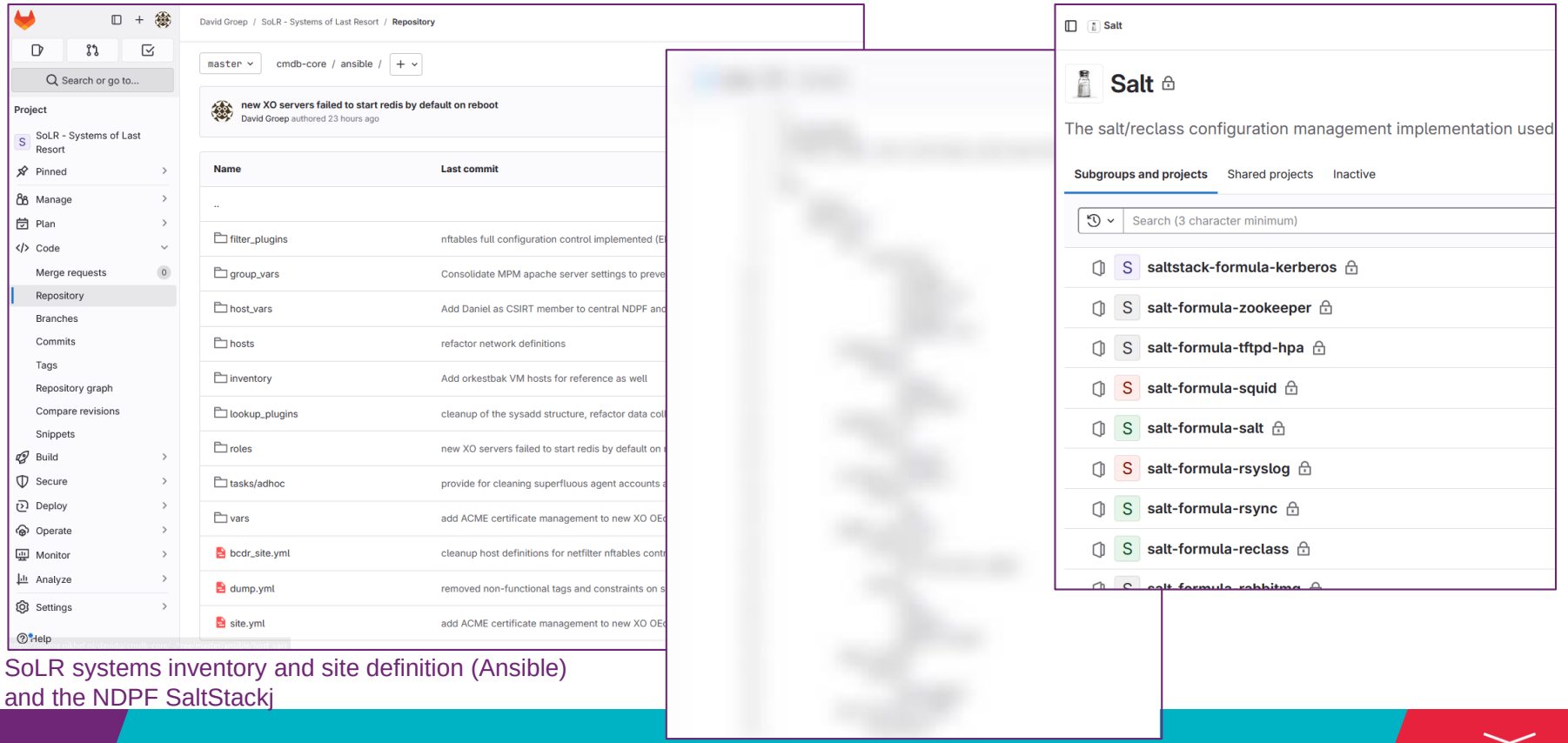
# Example: GSP ‘Koning’ en ‘Prinsjes’ (2024 edition) …




NetApp ASA C250 200TB all-flash 4x100Gbps iSCSI, 4x (Lenovo SR655 EPYC 9554P 128 logical + 30TB local ‘ephemeral’ NVME storage + 1100GiB DDR5 4800MHz)

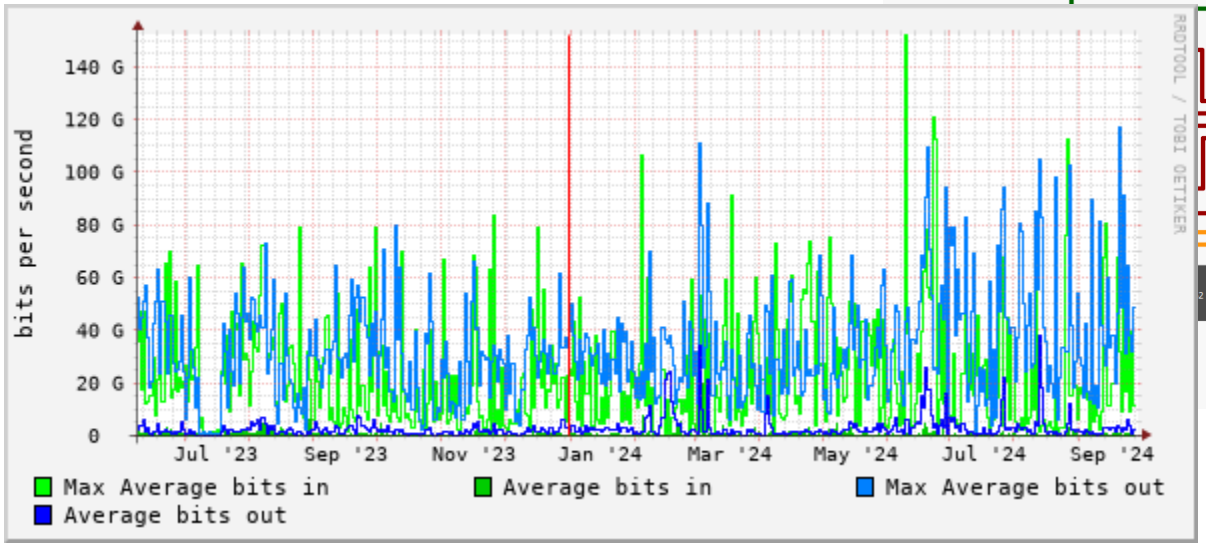
# Managing the NDPF and SoLR inventory

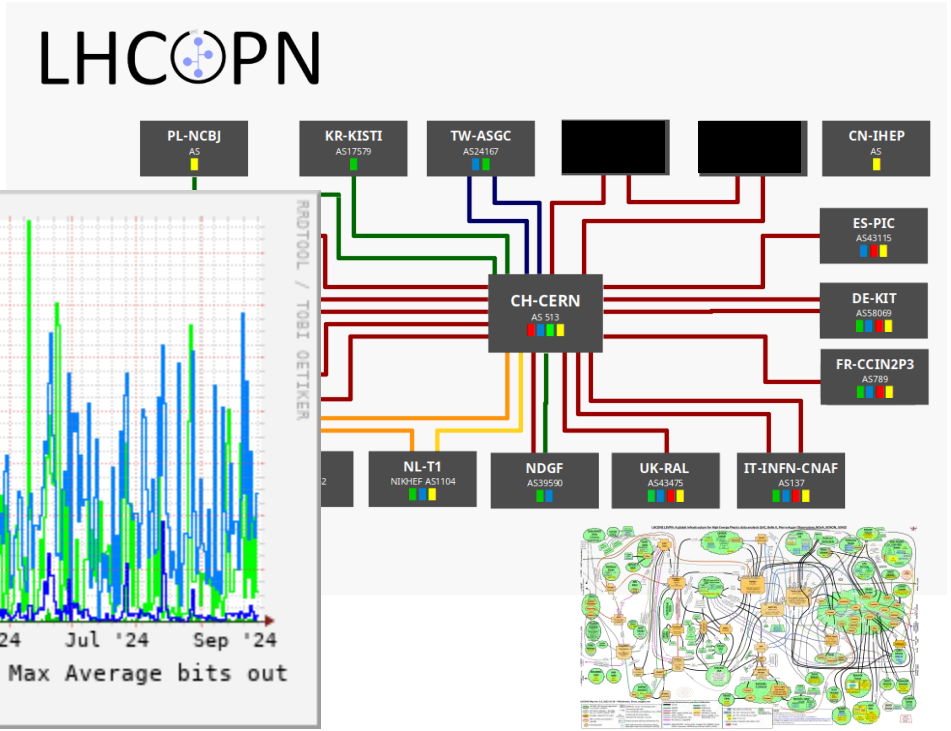
# But more importantly … (re)creating and managing them




SoLR systems inventory and site definition (Ansible)
and the NDPF SaltStackj

# Zolder: 400G access LHCOPN/LHCOne CERN link




ae10: CERN via asd001b-a96-08-6-3 – daily averages; https://cricket.nikhef.nl/ target /zolder.ipmi.nikhef.nl/ae10

# Physical view … QFX5130 400G switch ‘zolder’ top-up


Nikhef H234bC07 – foto’s Tristan Suerink

# Clusters: high throughput compute and storage service

Data-driven workloads (like WLCG, SKA, WeNMR) need more than ‘just’ compute:

- **balanced features** for node throughput: CPU, storage, memory bandwidth & latency, NIC & network speed
- **single-socket** multicore systems are fine, typical: 64-128 cores per system
- **network**: 2x25Gbps (+ ‘out of band’ management)
- **memory**: 8 GiB/core (so ~ 1 TiB/node)
- **local disk**: 16TB+ NVME PCIe Gen4 x4
- add **GPU**s depending on use case
- favour inference-optimized **APUs** in future

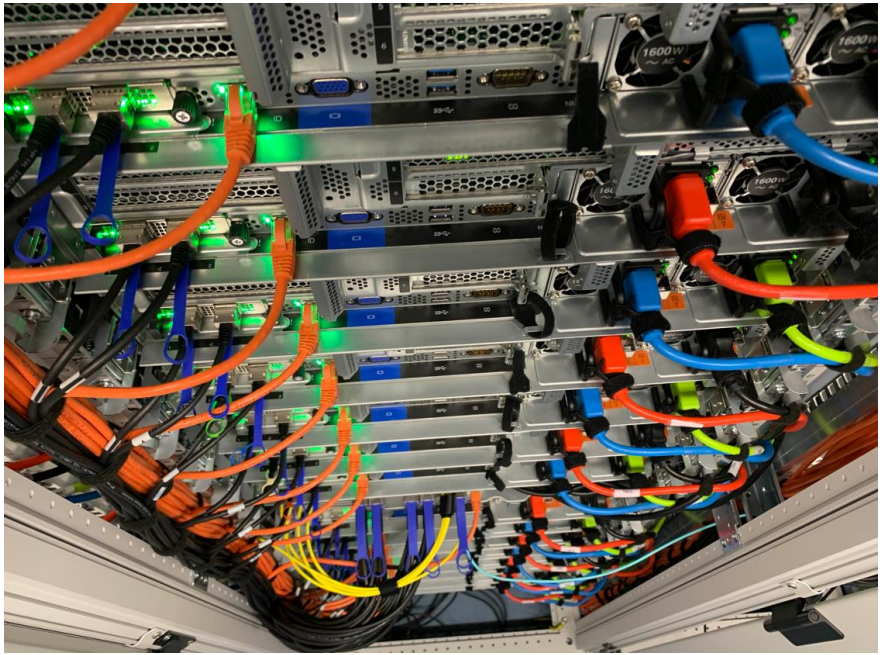
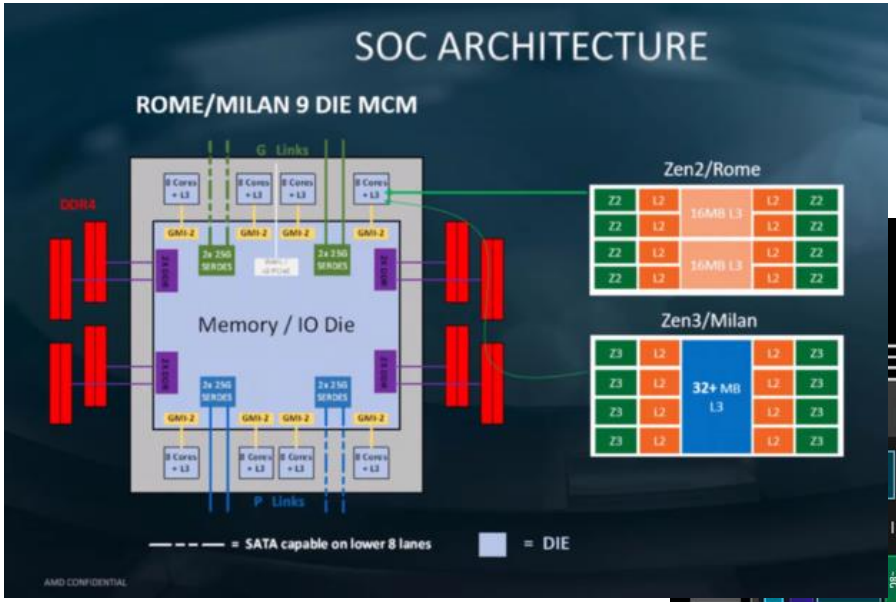
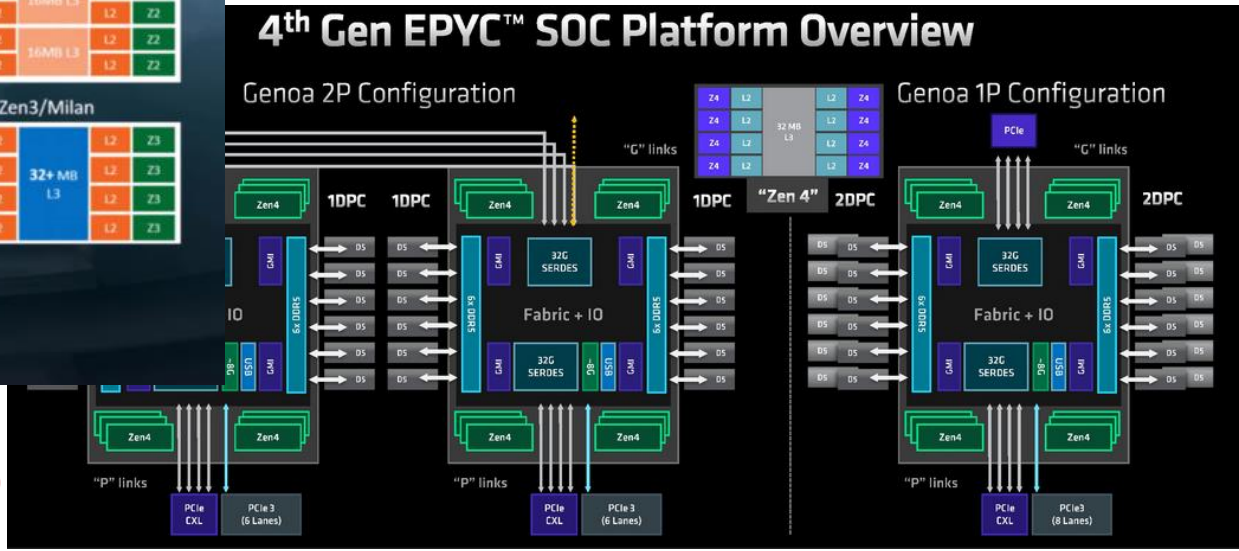


Image: Cluster ‘Lotenfeest’ at the Nikhef NDPF, acquired March 2020. Lenovo SR655 with AMD EPYC 7702P 64-Core single-socket

# Balanced systems design: from global to local performance




Example: AMD EPYC CPUs
and their influence on systems design



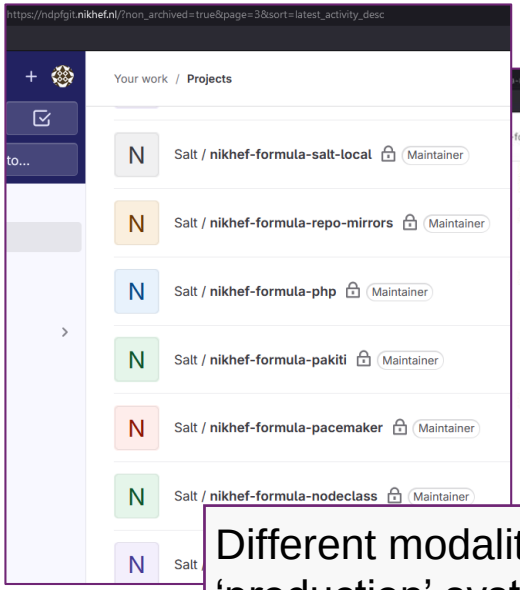

!! not all ‘upgrades’ relevant to all applications
(Rome>Milan hardly improved HTC throughput)

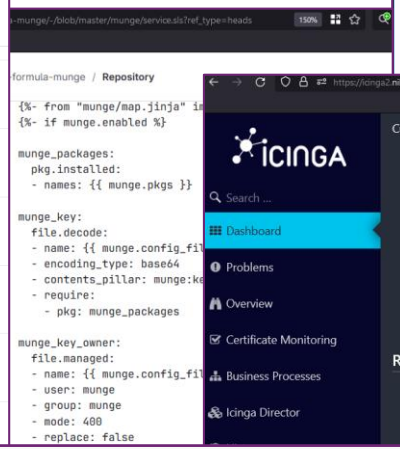
AMD, retrieved from https://m.hexus.net/tech/news/cpu/135479-amd-shares-details-zen-3-zen-4-architectures/ and https://www.semianalysis.com/p/amd-genoa-detailed-architecture-makes

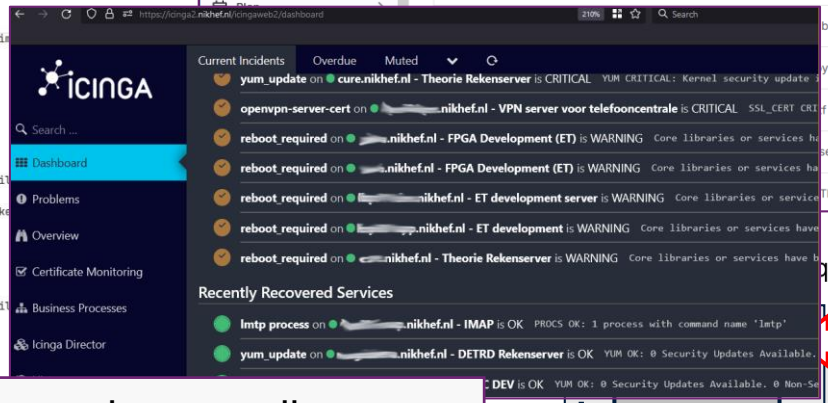


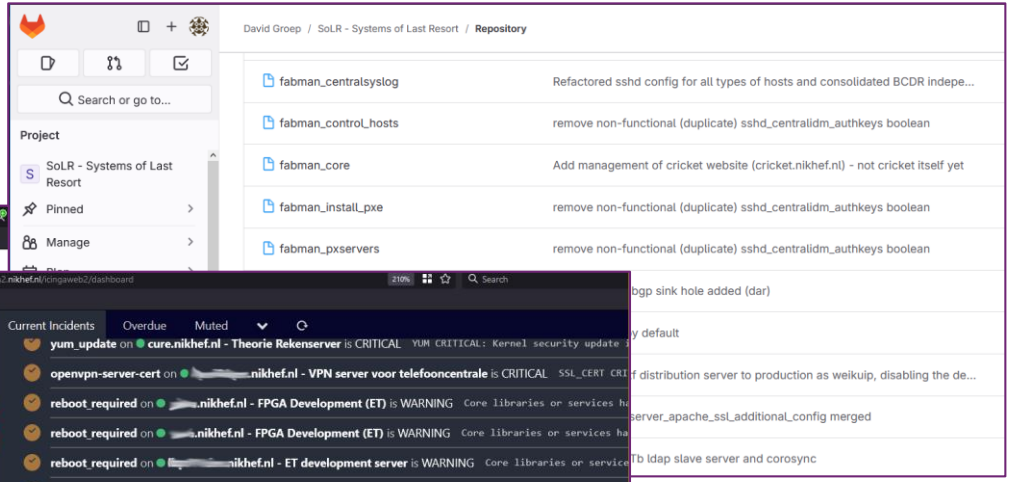
Nikhef

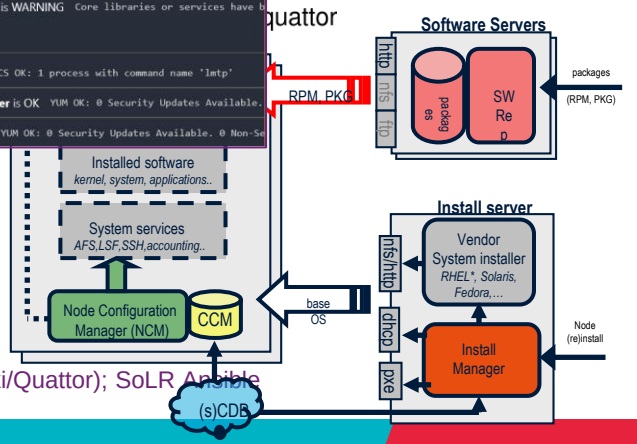
# More than one system: Fabric Management



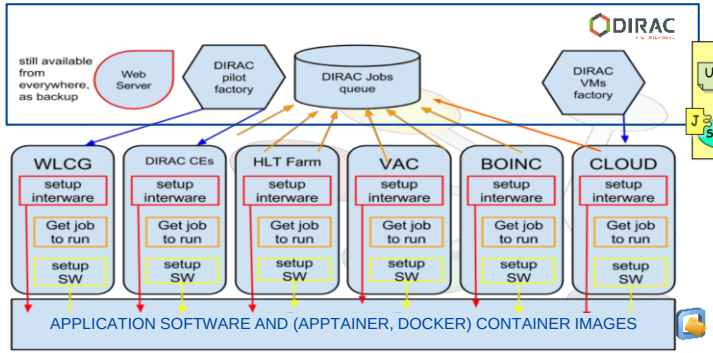

Different modalities are fine as long as all ‘production’ systems are managed and monitored
(yet this does not apply – for a reason - to the experimental technologies platform and Nationale Speeltuin)

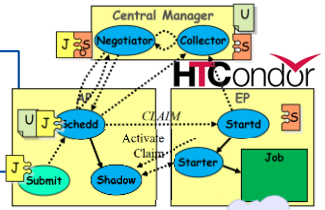
Nikhef NDPF Salt & Reclass (Dennis van Dok, Andrew Pickford, Mary Hester); Quattor: https://en.wikipedia.org/wiki/Quattor); SoLR Ansible

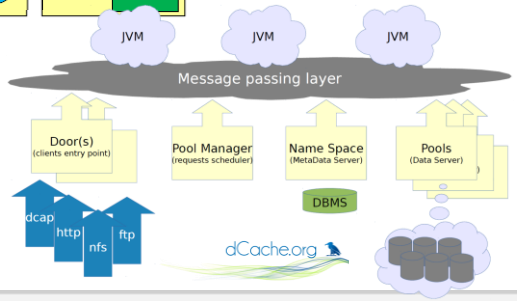


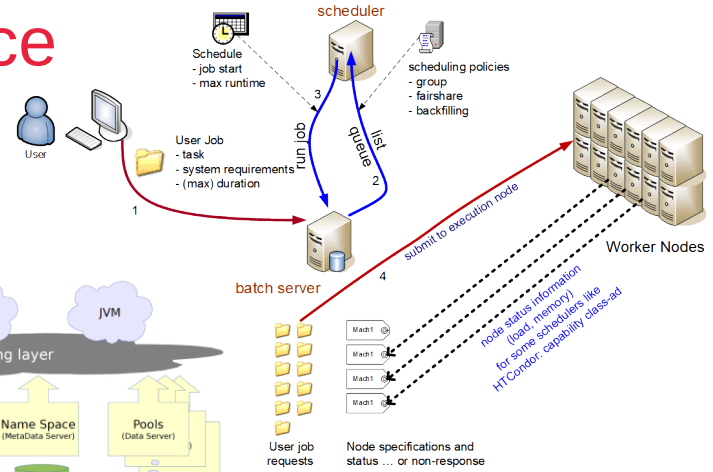
Nikhef

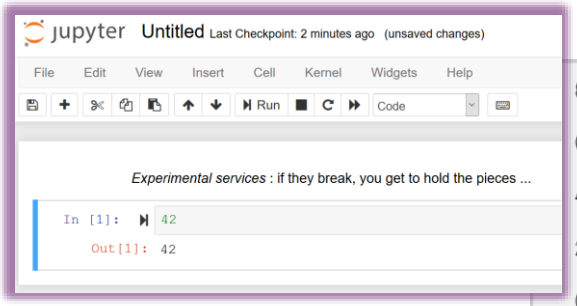
# From a bunch of systems to a service

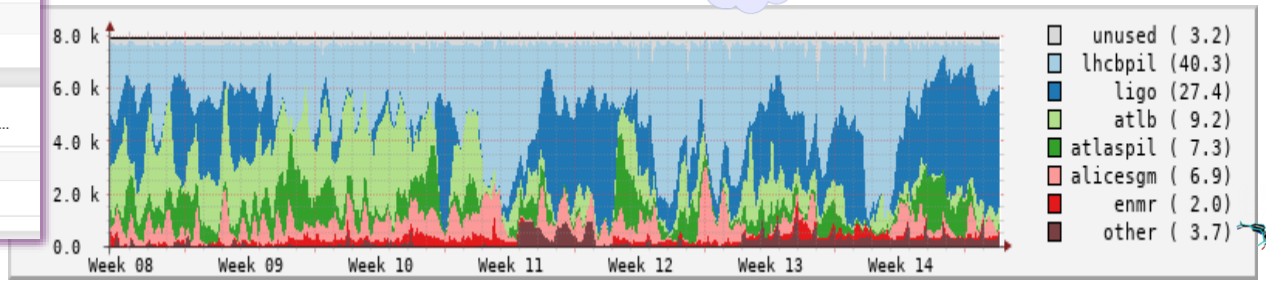


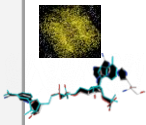



Occupancy: NDPF (DNI allocation), early 2023; other graphics: DIRAC Interware, HTCondor (UWMadison), Jupyter notebook: Nikhef Callysto service; dCache (DESY, dCache.org)



Nikhef

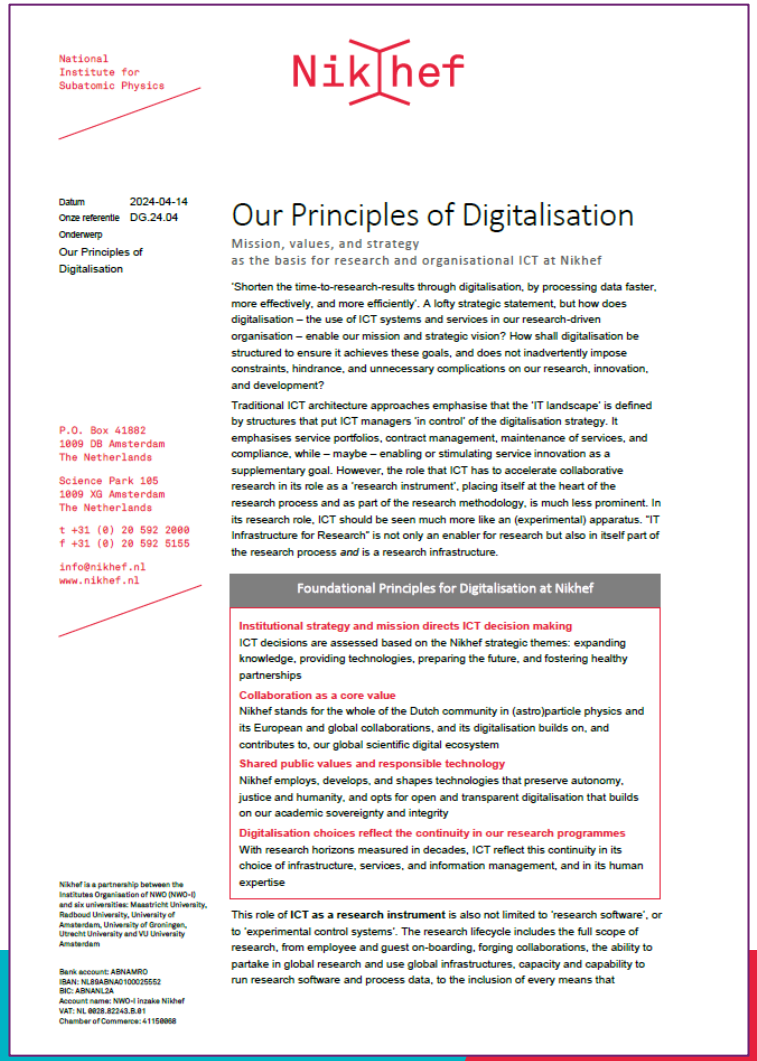
# Services and service management

Infrastructure and services for research

- driven by the strategy requirements
- appropriate service levels and impact
- balancing
  stability, innovation, and engagement

Service classes

- ‘enterprise’ services
- research computing services (local & DNI)
- experimental services



National
Institute for
Subatomic Physics

Nikhef

Datum 2024-04-14
Onze referentie DG.24.04
Onderwerp
Our Principles of
Digitalisation

P.O. Box 41882
1009 DB Amsterdam
The Netherlands

Science Park 105
1009 XG Amsterdam
The Netherlands

t +31 (0) 20 592 2000
f +31 (0) 20 592 5155

info@nikhef.nl
www.nikhef.nl

Nikhef is a partnership between the Institutes Organisation of NWO (NWO-I) and six universities: Maastricht University, Radboud University, University of Amsterdam, University of Groningen, Utrecht University and VU University Amsterdam

Bank account: ABNAMRO
IBAN: NL89ABNA0100025552
BIC: ABNANL2A
Account name: NWO-I Inzake Nikhef
VAT: NL 8028.82243.B.01
Chamber of Commerce: 41150868

## Our Principles of Digitalisation

Mission, values, and strategy
as the basis for research and organisational ICT at Nikhef

‘Shorten the time-to-research-results through digitalisation, by processing data faster, more effectively, and more efficiently’. A lofty strategic statement, but how does digitalisation – the use of ICT systems and services in our research-driven organisation – enable our mission and strategic vision? How shall digitalisation be structured to ensure it achieves these goals, and does not inadvertently impose constraints, hindrance, and unnecessary complications on our research, innovation, and development?

Traditional ICT architecture approaches emphasise that the ‘IT landscape’ is defined by structures that put ICT managers ‘in control’ of the digitalisation strategy. It emphasises service portfolios, contract management, maintenance of services, and compliance, while – maybe – enabling or stimulating service innovation as a supplementary goal. However, the role that ICT has to accelerate collaborative research in its role as a ‘research instrument’, placing itself at the heart of the research process and as part of the research methodology, is much less prominent. In its research role, ICT should be seen much more like an (experimental) apparatus. “IT Infrastructure for Research” is not only an enabler for research but also in itself part of the research process *and* is a research infrastructure.

**Foundational Principles for Digitalisation at Nikhef**

**Institutional strategy and mission directs ICT decision making**
ICT decisions are assessed based on the Nikhef strategic themes: expanding knowledge, providing technologies, preparing the future, and fostering healthy partnerships

**Collaboration as a core value**
Nikhef stands for the whole of the Dutch community in (astro)particle physics and its European and global collaborations, and its digitalisation builds on, and contributes to, our global scientific digital ecosystem

**Shared public values and responsible technology**
Nikhef employs, develops, and shapes technologies that preserve autonomy, justice and humanity, and opts for open and transparent digitalisation that builds on our academic sovereignty and integrity

**Digitalisation choices reflect the continuity in our research programmes**
With research horizons measured in decades, ICT reflect this continuity in its choice of infrastructure, services, and information management, and in its human expertise

This role of **ICT as a research instrument** is also not limited to ‘research software’, or to ‘experimental control systems’. The research lifecycle includes the full scope of research, from employee and guest on-boarding, forging collaborations, the ability to partake in global research and use global infrastructures, capacity and capability to run research software and process data, to the inclusion of every means that

# Our Principles of Digitalisation

‘If they are many, they are not principles’, so we set only four:

1. **Institutional strategy and mission directs ICT decision making**
   ICT decisions are assessed based on the Nikhef strategic themes
2. **Collaboration as a core value**
   Nikhef stands for the whole of the Dutch community in (astro)particle physics and its European and global collaborations
3. **Shared public values and responsible technology**
   Nikhef employs, develops, and shapes technologies that preserve autonomy, justice and humanity, that builds on our academic sovereignty and integrity
4. **Digitalisation reflects the continuity in our research programmes**
   With research horizons measured in decades, ICT reflect this continuity in its choice of infrastructure, services, and information management, and in its human expertise

# FitSM – Federated IT Service Management

Structuring service management with https://www.fitsm.eu/
(collaboration and federation-focused light-weight ISO20k/ITILv3 rendering)

- production systems of
  High Throughput Compute (HTC) and on-line Storage (HTS) services
- system & network components supporting provisioning of public services
- pre-production and systems for service portfolio development

*Scope excludes by design experimental and research systems,*
*apart from Supplier Relationship Management and procurement*

# Coordinated service management

- Single source of truth
- Inventory and support contract insight
- Configuration Management
  - for systems management several used in parallel, which is not per-se an issue: Saltstack, Ansible
- Facilitates our ISM
  Technical and Organisational Measures

Methodological alignment with SURF DNI

**Contents** [hide]





Nikhef

# And … we live in a federated global world

# Collaboration: an inherently-cross-domain issue ..

EXAMPLE FROM THE LHC COMPUTING INFRASTRUCTURE WLCG

**170** SITES
**~50** COUNTRIES & REGIONS
**~20000** USERS

JUST *HOW* MANY INTERACTIONS ??

AUTHN & AUTHZ, **ARCHITECTURE AND TRUST** SHOULD ALIGN WITH **COLLABORATION STRUCTURES**, AND BE **OUTWARD FACING:** OPEN, SCALABLE, & MULTI-DOMAIN

people photo: a small part of the CMS collaboration in 2017, Credit: CMS-PHO-PUBLIC-2017-004-3; site map: WLCG sites from Maarten Litmaath (CERN) 2021

# IdM: ‘AARC’ & REFEDS compliant Trust & Identity AAI

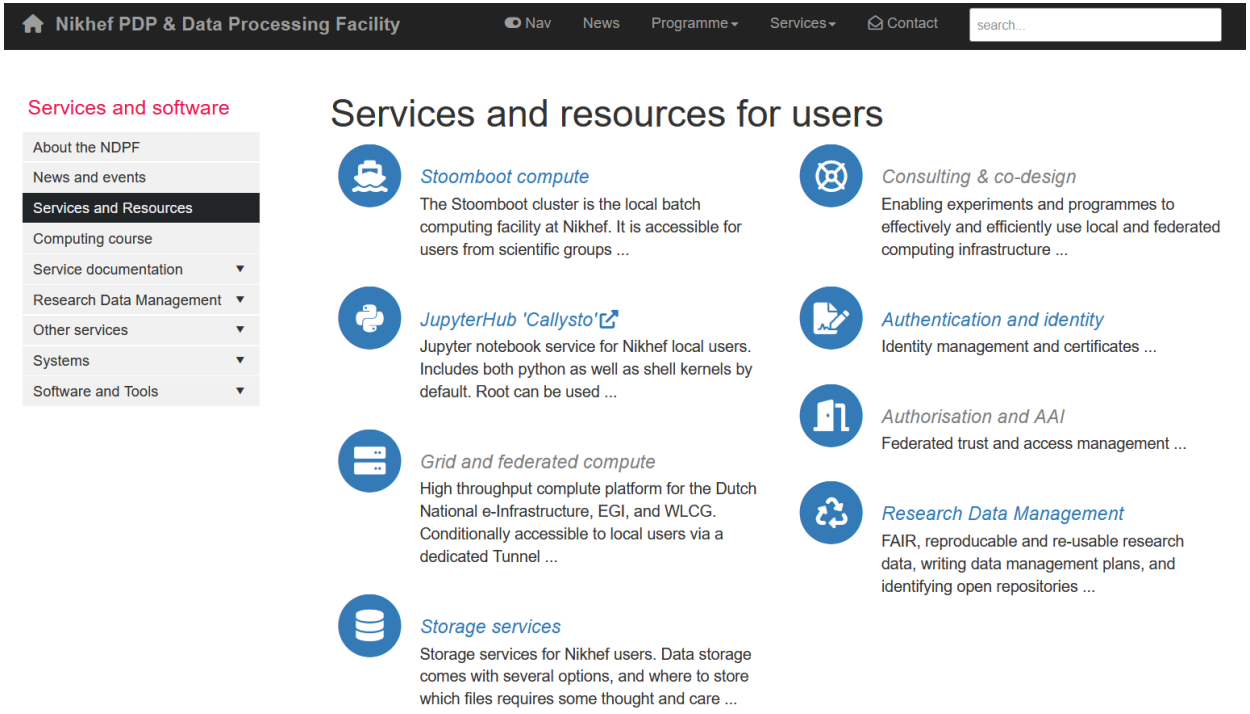
## Central identity management

- designed to allow innovation based on REFEDS & AARC models
- ‘our users are the validation’
- production-quality *with* rapid innovation cycle

## Basis for all AAI (identity & access)

- OpenLDAP as core
- bespoke integration suite
- SimpleSAMLphp IdP and SP proxy
- Support SAML, OIDC, OIDfed, and REFEDS MFA & RAF natively

# Research Computing Services




https://www.nikhef.nl/pdp/doc/service-portfolio

# Analysis cluster

Nikhef **CT User Documentation** Search

CT User Documentation
- Accounts
- Calendar & Mail
- Printing
- Laptops & Hardware
- Network (WiFi/VPN)
- Phones
- Meeting Rooms
- Remote Access Desktops & Servers
- Computing & Data Analysis
  - Interactive CPU nodes
  - Interactive GPU nodes
  - Stoomboot cluster
  - Stoomboot NG HTCondor Cluster
  - CPU batch jobs
  - GPU batch jobs
  - Qsub tunnel
  - Nikhef Grid Computing Cluster (NDPF)
  - Grid jobs
  - Ganymede cluster
  - Conda environments
  - pyenv for managing python versions
  - Containers

## Stoomboot cluster

**Aim:** Describe how to access and use Nikhef's Stoomboot cluster.

**Target audience:** Users of the Stoomboot cluster.

### Prerequisites

To access Stoomboot, you will need:

- A Nikhef account
- A working SSH client setup

### Introduction

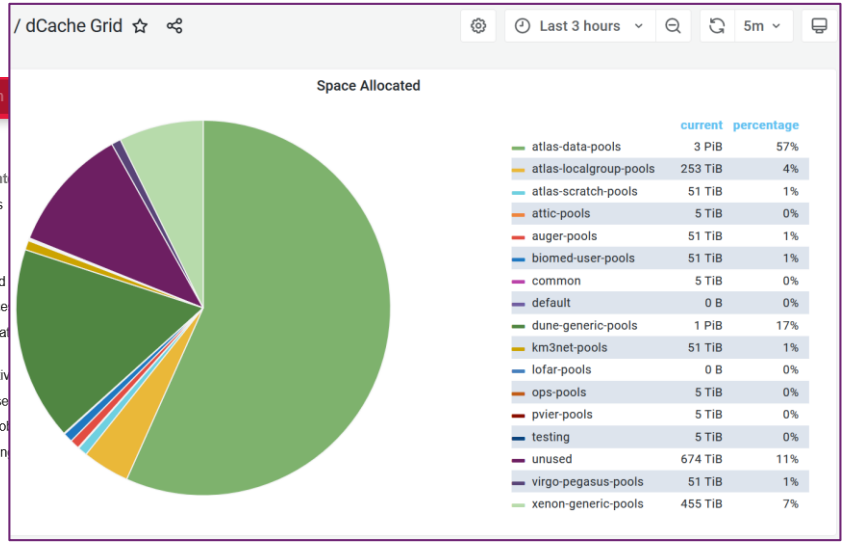
The Stoomboot cluster is the local computing facility at Nikhef. It is accessible for users from scientific groups to perform, for

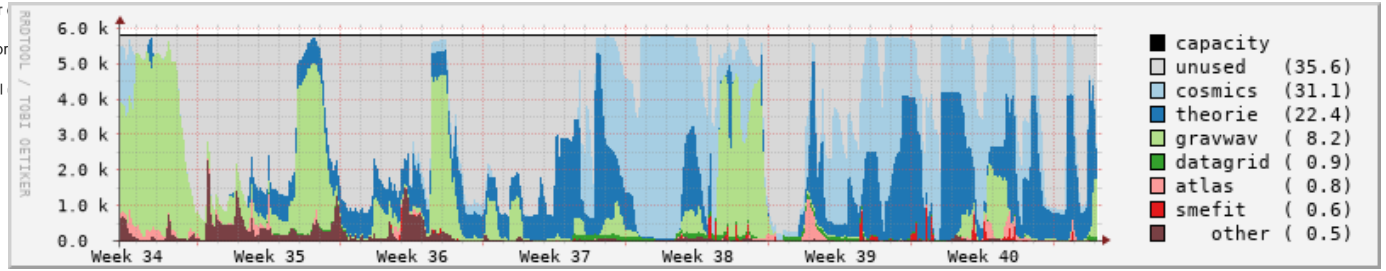
The Stoomboot cluster is compr

- 85 batch nodes with a total
- 3 interactive CPU nodes
- 2 interactive GPU nodes
- 7 GPU batch nodes

Table of cont
- Prerequisites
- Introduction
- Usage
  - Access and
  - Batch proce
  - Seeing what doing
  - Cluster activ
    - STBC use
    - Waiting jo
    - Job waitin
- Links
- Contact

https://kb.nikhef.nl/ct/






Nikhef

# Can can analyze small things in an executable Jupyter paper

## Services and resources for users

### *Stoomboot compute*

The Stoomboot cluster is the local batch computing facility at Nikhef. It is accessible for users from scientific groups ...

### *JupyterHub 'Callysto'*

Jupyter notebook service for Nikhef local users. Includes both python as well as shell kernels by default. Root can be used ...

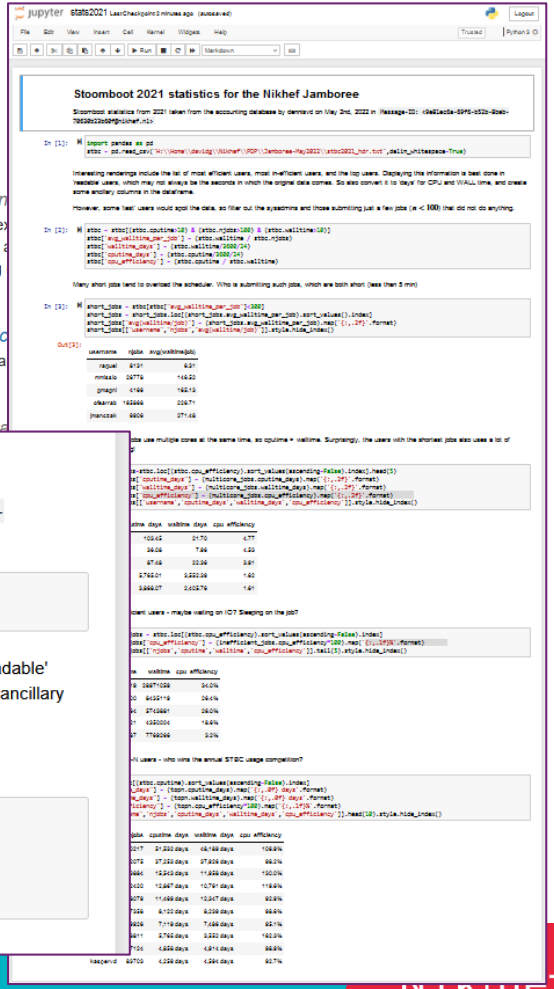
### *Consultin*

Enabling ex
effectively a
computing

### *Authentic*

Identity ma

### *Authorisa*




## Stoomboot 2021 statistics for the Nikhef Jamboree

Stoomboot statistics from 2021 taken from the accounting database by dennisvd on May 2nd, 2022 in `Message-ID: <9e81ec6a-69f6-b52b-8beb-70630b23b60f@nikhef.nl>`

In [1]:
```
import pandas as pd
stbc = pd.read_csv("H:\\Home\\davidg\\Nikhef\\PDP\\Jamboree-May2022\\stbc2021_hdr.txt",delim_whitespace=True)
```

Interesting renderings include the list of most efficient users, most in-efficient users, and the top users. Displaying this information is best done in 'readable' users, which may not always be the seconds in which the original data comes. So also convert it to 'days' for CPU and WALL time, and create some ancillary columns in the dataframe.

However, some 'test' users would spoil the data, so filter out the sysadmins and those submitting just a few jobs ($n < 100$) that did not do anything.

In [2]:
```
stbc = stbc[(stbc.cputime>10) & (stbc.njobs>100) & (stbc.walltime>10)]
stbc['avg_walltime_per_job'] = (stbc.walltime / stbc.njobs)
stbc['walltime_days'] = (stbc.walltime/3600/24)
stbc['cputime_days'] = (stbc.cputime/3600/24)
stbc['cpu_efficiency'] = (stbc.cputime / stbc.walltime)
```

# Including analyzing the infrastructure itself …

There are also inefficient users - maybe waiting on I/O? Or just sleeping on the job?!

In [5]:
```
inefficient_jobs = stbc.loc[(stbc.cpu_efficiency).sort_values(ascending=False)
inefficient_jobs["cpu_efficiency"] = (inefficient_jobs.cpu_efficiency*100).map
inefficient_jobs[['njobs','cputime_days','walltime_days','cpu_efficiency']].ta
```

Out[5]:

| njobs | cputime_days | walltime_days | cpu_efficiency |
|---|---|---|---|
| 27799 | 114.010637 | 335.313183 | 34.0% |
| 1595 | 25.818287 | 97.628692 | 26.4% |
| 4551 | 17.312083 | 66.480104 | 26.0% |
| 11299 | 9.377558 | 50.347269 | 18.6% |
| 1117 | 2.877164 | 89.922060 | 3.2% |

# Research Engineering support and engagement

https://www.nikhef.nl/pdp/doc/storage-classes

Nikhef PDP & Data Processing Facility

Nav News Programme Services Contact search...

Services and software

- About the NDPF
- News and events
- Services and Resources
- Computing course
- Service documentation
  - User Grid wiki
  - Credentials and certificates
  - Stoomboot Analysis cluster
  - Storage services and types
  - Grid User Interface 'bosui'
  - Submitting to ARC CEs
  - Submitting to HTcondor CEs
  - Usage graphs and statistics
  - NL-Tier1 Alarms
- Research Data Management
- Other services
- Systems
- Software and Tools

## Storage classes for the Nikhef analysis facilities

**Summary:** Data storage at Nikhef comes with several options, and where to store wh
and care. For example, your home directory should not be used for bulk data, and file
up. Read about which type of files should go where.

Table of Contents

- Home directory
- Data in `/data`
- dCache
- Project
- Local cache storage
- SURFdrive
- FileSender

### Home directory

Your home directory is for personal files and co
(login.nikhef.nl, the stoomboot interactive nodes

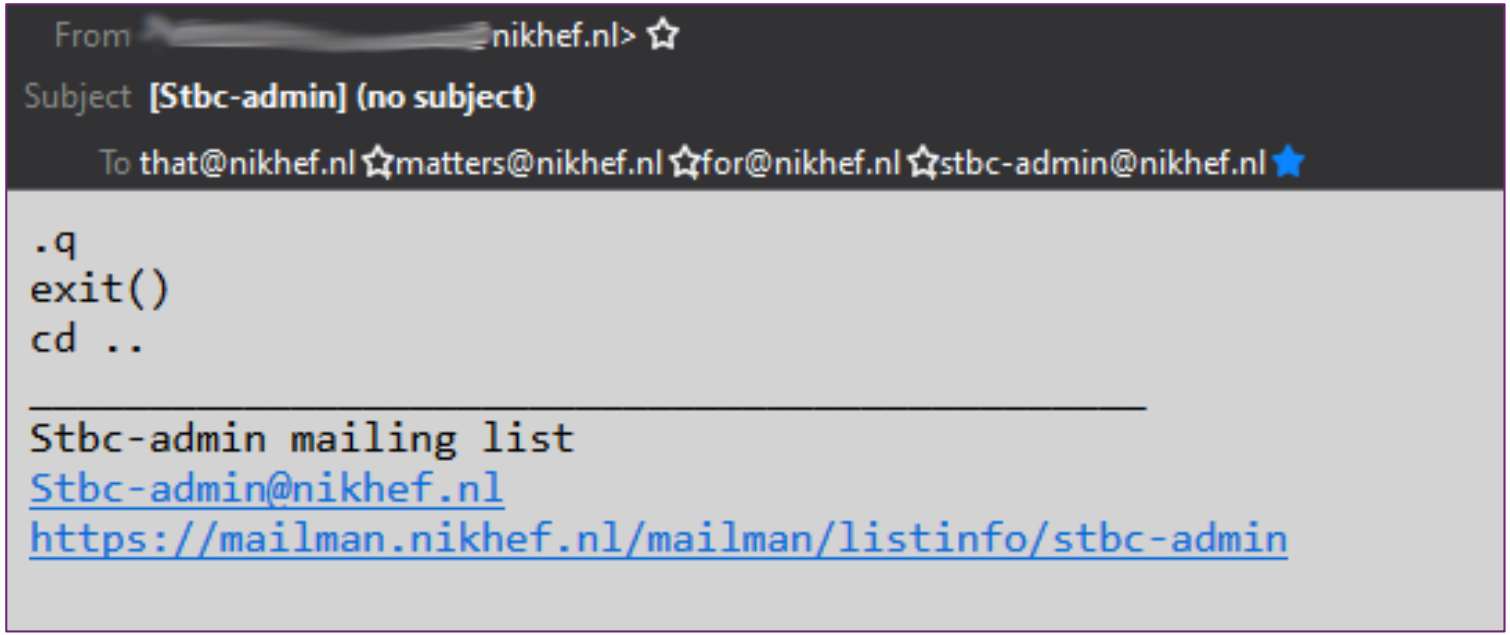
- **intended use**: your "dot" files, personal a
work, hobby projects, personal mails, co
things that will not be preserved after you leave Nikhef.
- **examples of data types that are better put elsewhere**: your ntuples (put those in dCache or /data), scripts and frameworks used by a group of colleagues (that's put in /project), log output you want to look at later (best put this alongside the results in /data), intermediate files (use `$TMPDIR`, or `/localstore`), your final thesis (submit it to the library, and package the plots, publications, tabular data, and histograms for submission to Zenodo)

computing courses
and 'Office Hours' concept



Nikhef

# Although users may send some rather confused questions …



From [redacted]nikhef.nl>

Subject **[Stbc-admin] (no subject)**

To that@nikhef.nl matters@nikhef.nl for@nikhef.nl stbc-admin@nikhef.nl

```
.q
exit()
cd ..

_____________________________________________________
Stbc-admin mailing list
Stbc-admin@nikhef.nl
https://mailman.nikhef.nl/mailman/listinfo/stbc-admin
```

# A coherent collection of emergent services

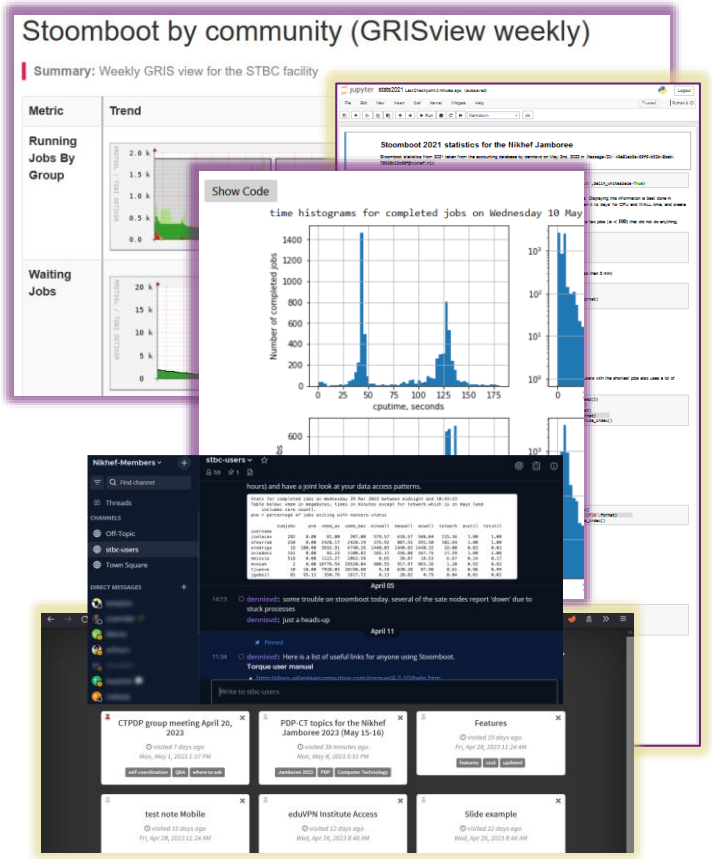
## Nikhef Partnership facilities

- **Stoomboot:** ~2000 cores and ~4 PByte dCache
- **Callysto:** JupyterHub with $HOME and SSO login
- **eduVPN:** securely access Callysto and your home
- Mattermost's '**STBC-users**' channel to talk & ask
- **eVA**, **SURFdrive**, and **FileSender** to collaborate

## Global Services

- WLCG, SURF DNI, and EGI: **HTC compute, CVMFS**, …

**Experimental services**: ShareMD, Commute, …




But do read https://www.nikhef.nl/pdp/doc/experimental-services before using experimental services ...
Stoomboot statistics: https://www.nikhef.nl/pdp/doc/stats/stbc-grisview-week, https://www.nikhef.nl/pdp/stats/stbc/intern/stbc_summ_plots

# Experimental services

- Service documentation
- Research Data Management
- **Other services**
  - ObjectID Registry
  - URN Registry
  - SSO Identity Management
  - What is my IP
  - Livestreaming at Nikhef
  - NikhefTV
  - ShareMD HedgeDoc notes
  - Audioconferencing PBX
  - Orchestra Experimental Cloud
  - NWO-I Commute Tool
- About Experimental Services

https://www.nikhef.nl/pdp/doc/experimental-services

- if it's broken, tell us … and we might even fix it.
- if it works: also tell us, so we may change it beyond recognition!
- … 'if it breaks, you get to hold the pieces'.

We do not offer any service guarantee, neither expressed nor implied, about purpose, in particular not for getting work done. The service is operated by th

## PDP-CT Experimental Services

**Summary:** If they break, you get to hold the pieces!

You came here because you were curious about an *Experimental Service*, and what that means in the Nikhef context.

es: highly important ones that support our primary mission. Highly available going on and (at least attempt to) provide reasonably quick resolution in case of of course there's support for folk who use the services, on-demand or during the

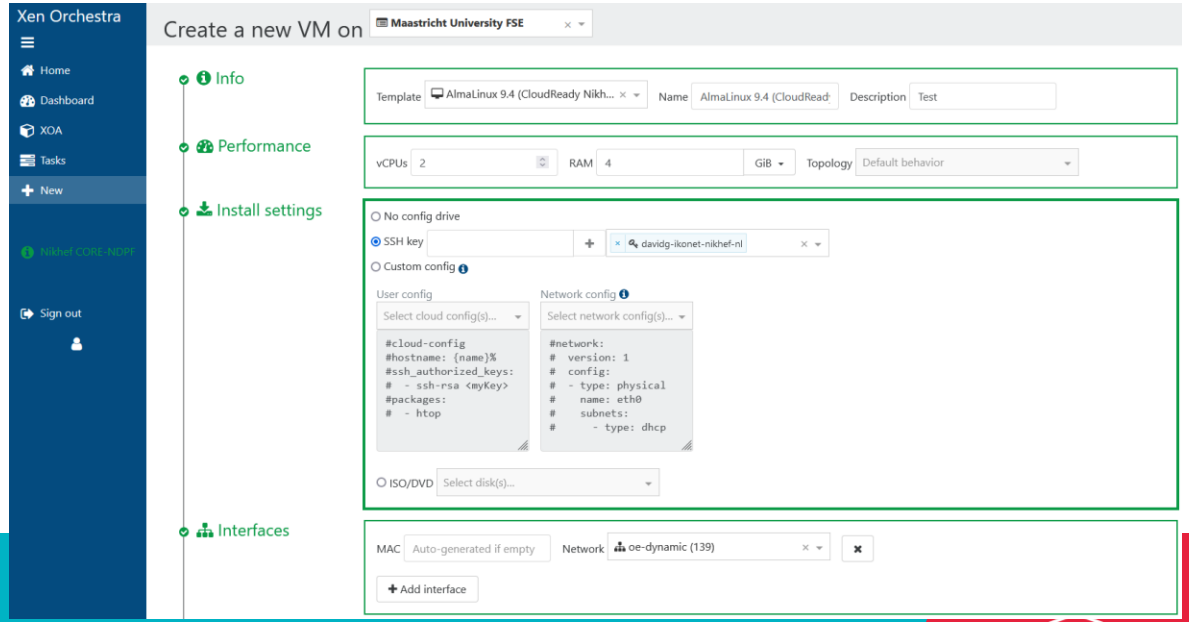
t none of the above. None. They are vastly different and could do you more harm may cause you to loose years of valuable work. They may make your pet walk

ce run by the Nikhef Physics Data Processing group - and that one is subject to

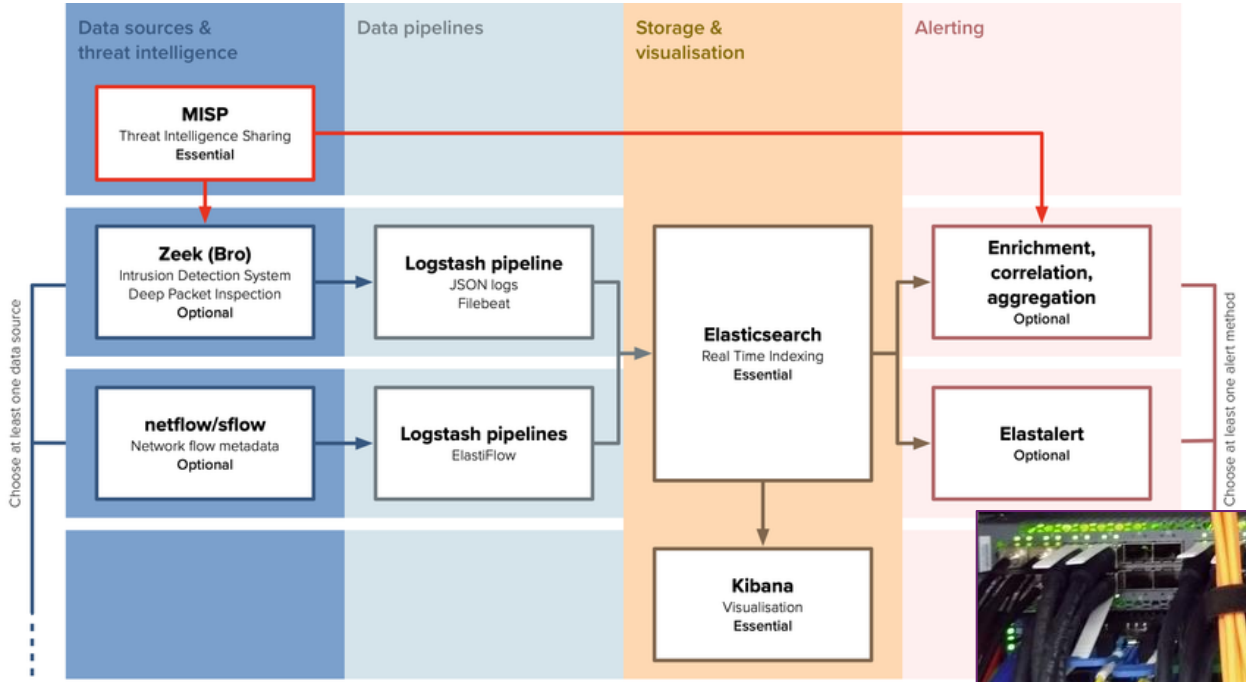
n fix it.

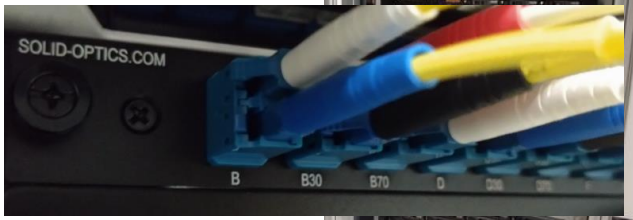
e it beyond recognition!

's'.



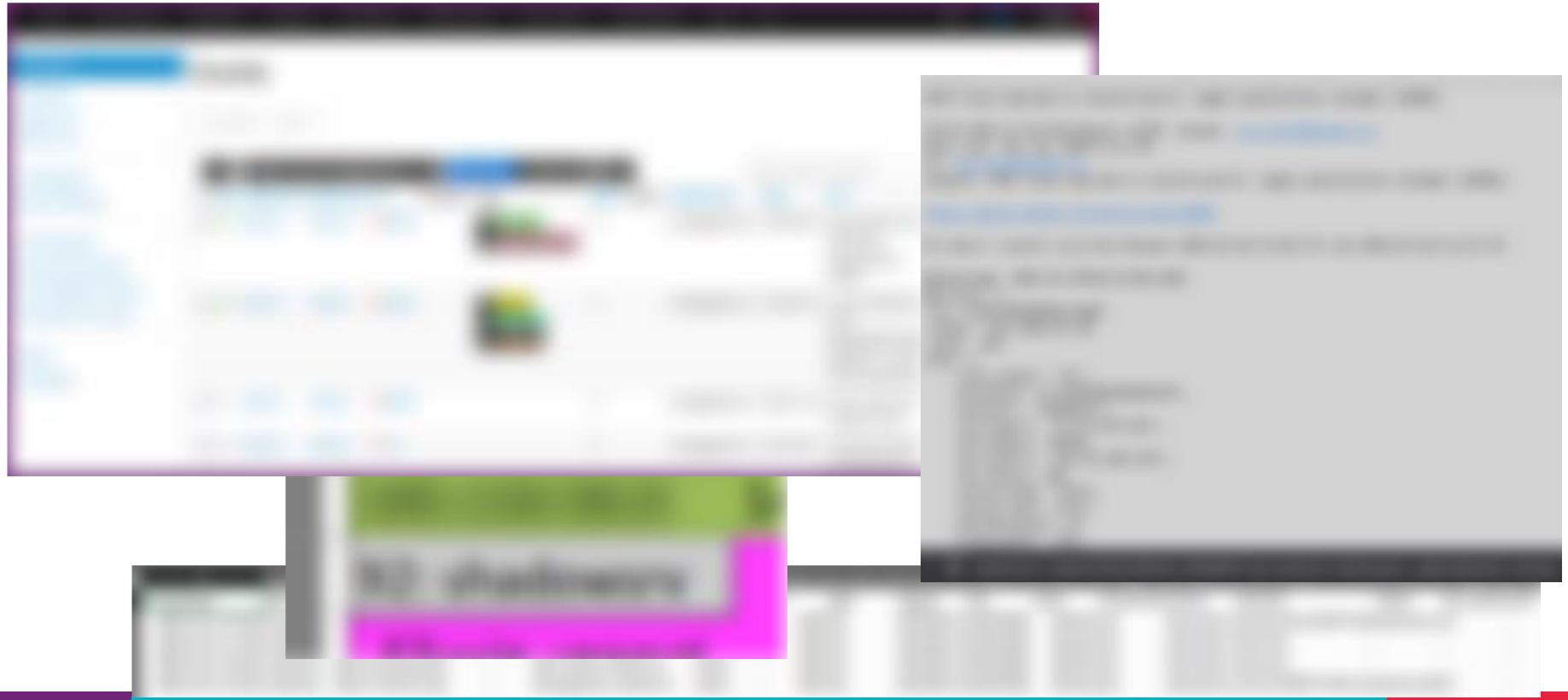
# Network monitoring: the SOC, MISP and WLCG SOC-WG



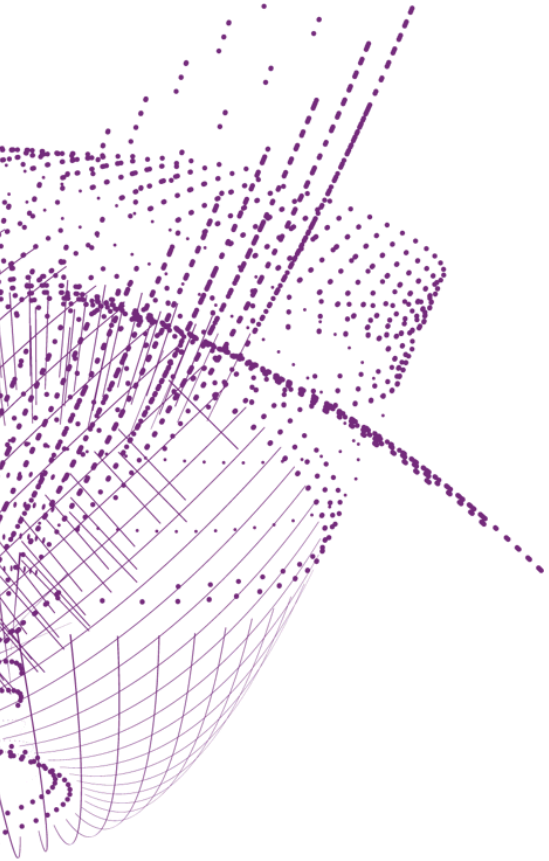

Nikhef SOC, H234b C06
WLCG SOC: David Crooks and Liviu Valsan, ISGC2019 proceedings, PoS(ISGC2019)010

# MISP Threat intel sharing (plus Shodan & ShadowServer)





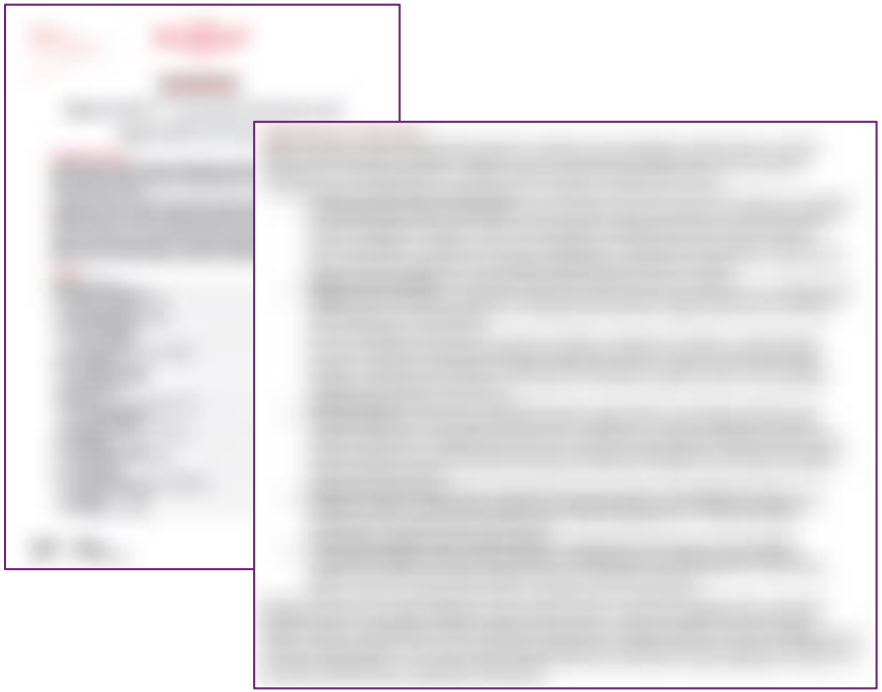
Nikhef



Backup and BC/DR

# When things go south …

# But if something does go belly-up
## *we should know what to do*



# BCDR

The Nikhef XXXXXXXXXXXXXXXXXX BCDR setup – foundational service capability more than 30km away from the primary site …

David Groep
davidg@nikhef.nl
https://www.nikhef.nl/~davidg/presentations/
https://orcid.org/0000-0003-1026-6606

Nikhef