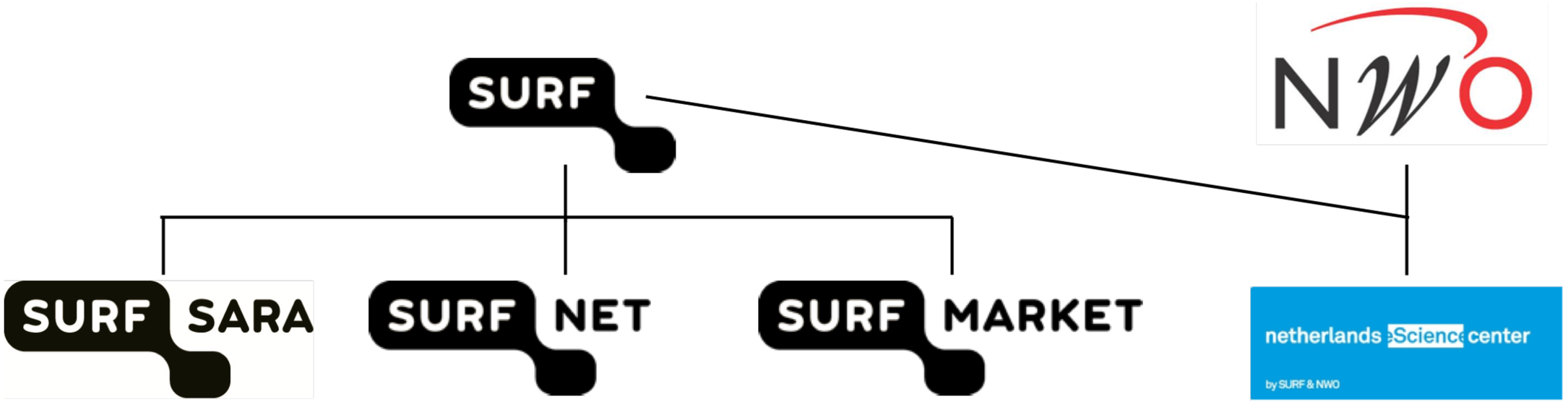
Spider HT platform: Elixir - course

Utrecht, 28 Aug. 2019

Spider team @ SURFsara

About SURF

ICT services for a strong and sustainable knowledge economy



SURF SARA

Academic Medical Centres

VUmc
UMC Utrecht
LUMC
umcg
Erasmus MC
Radboudumc
Maastricht UMC+
amc

100+ and growing

Universities for applied sciences (+ Senior secondary vocational education institutions)

Hogeschool van Amsterdam
ZUYD
HU Hogeschool Utrecht
Hanzehogeschool Groningen University of Applied Sciences
Hotelschool The Hague Hospitality Business School
CHE
Hogeschool van Arnhem en Nijmegen
HKU
Fontys
Stenden
Hk Hogeschool der Kunsten Den Haag
marnix academie
De Haagse Hogeschool
Saxion
IselingeHogeschool Onderdeel van IJsselgroep
Driestar
Amsterdamse Hogeschool voor de Kunsten
has hogeschool
NHL Hogeschool
hogeschool Leiden
ArtEZ
Windesheim
inholland hogeschool
Kabk
deKempel Hogeschool
avans hogeschool
University of Applied Sciences
viaa
Hogeschool iPabo
Hogeschool Rotterdam
NHTV
Design Academy Eindhoven
van hall larenstein university of applied sciences

Universities

Erasmus
Universiteit van Amsterdam
VU Amsterdam
Radboud Universiteit
Tilburg University
University of Groningen
Universiteit Twente
Universiteit Utrecht
TU/e
TU Delft
UM
Open Universiteit
Wageningen UR
Universiteit Leiden

KNAW NWO & other research institutes

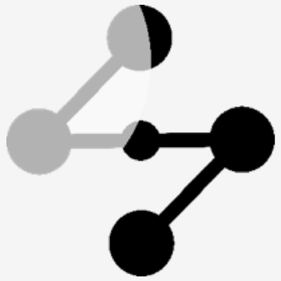
NIKHEF
CBS
NiDi
Nationaal Archief
Rathenau Instituut
KB
Koninklijk Nederlands Meteorologisch Instituut
nscr
KITLV
ICIN Netherlands Heart Institute
SPINOZA
Rijksinstituut voor Volksgezondheid en Milieu
NIOO
NIOD instituut voor oorlogs-, holocaust- en genocidestudies
SRON
ASTRON
Netherlands Cancer Institute
huygens ing
The Hague Institute for Global Justice
CWI
FOM Institute AMOLF
DANS
Meertens Instituut
internationaal instituut voor sociale geschiedenis
UVH
Politie Politieacademie
DIFFER
NIOZ
NCB naturalis
ARCNL
Fryske Akademy
NIAS
TNO

SURF Mission

Supporting Research & Education as:

- Service & infrastructure provider
- Innovator
- Partner in developing new services
- Service broker
- Expertise center

SURF services: *(access via surf.nl)*



High bandwidth network
Dedicated lightpaths
Encrypted transfers

Cartesius
Lisa
HPC Cloud
Research Cloud
HTP: Grid / Spider
Kubernetes
Spark

>80K cores

Ceph
dCache
SURFdrive
ResearchDrive
Object Store
Preservation
EPIC PID
iRODs

> 50 PB

Support
Consultancy
Training
Visualisation
Optimisation
Collaboration

- Data center: 100% renewable energy
- High information security: ISO 27001


Digital Realty
Data Tower
(SURFsara Data
Center)
SURFsara
1971: SARA founded by
CWI, UvA, VU
1984: 1st supercomputer
2013: joined SURF
cooperative
NLeSC

High throughput processing

- **SURFsara is about making scientific discoveries happen – via high-end IT**

High throughput processing

- **HT processing is data-driven research => “ from noise to scientific discovery ”**

High throughput processing

- **HT processing is data-driven research => “ from noise to scientific discovery ”**

- SURFsara Grid storage in total grows with **5-10 PB/yr** and has a throughput of ca. **400 PB/yr**
- Experimental data generation grows exponentially **>10x** (*2020-2030: 1 EB / yr per experiment*)

High throughput processing

- **HT processing is data-driven research => “ from noise to scientific discovery ”**

- SURFsara Grid storage in total grows with **5-10 PB/yr** and has a throughput of ca. **400 PB/yr**
- Experimental data generation is increasing by **>10x** (*2020-2030: 1 EB / yr per experiment*)

- HTC: focused on Grid solution
  - WLCG (60 PB/yr , NL 10%)
  - International, distributed processing
  - Long-lived skilled IT teams
  - Simple, independent jobs

(CERN - CMS)

Grid Storage
(dCache)

CERN

LIGO-VIRGO

Pierre Auger

(Big experiments)

Grid

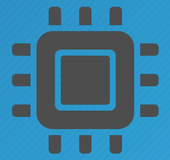
- International
- Distributed
- Inhomogeneous
- Certificates

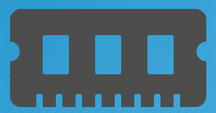
-> Limited functionality

Applications

Funnel “rigidity” (Beck+)

Infra / HW





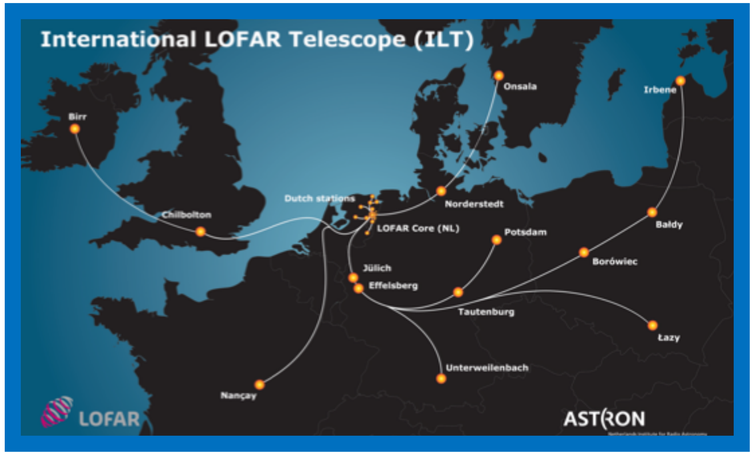
Infrastructure

openstack. ceph

SURF SARA

High throughput processing

- **Astronomy , Engineering / Climate , Life sciences : 2020 – 2030 contribute ca. 50% of the data**



LOFAR / SKA

Tropomi (Sentinel 5p)

e.g. Project MinE

- Grid solution insufficient: *(a) Complex workflows, (b) Fast moving targets, (c) Non-permanent staff*
- Need <u>better solutions</u> to accelerate scientific discovery in (emerging) data-driven research domains

SURF SARA

Infrastructure – “ cloudification ”

- DPS services move hardware **from physical to Virtual Data Center (VDC)**
- Flexible deployment & shift of resources for the data processing services
- VDC backbone:
  - Openstack: tuned for data processing *(managed Cloud)*
  - Powerful data processing nodes *(>256 GB Ram, >3 TB SSD)*
  - Ceph: shared, highly scalable storage *(3x Redundancy)*
  - High Bandwidth Network *(internal & external)*

Highly scalable !!

SURF SARA

Spider

Elixir / LS community

CERN

LIGO-VIRGO

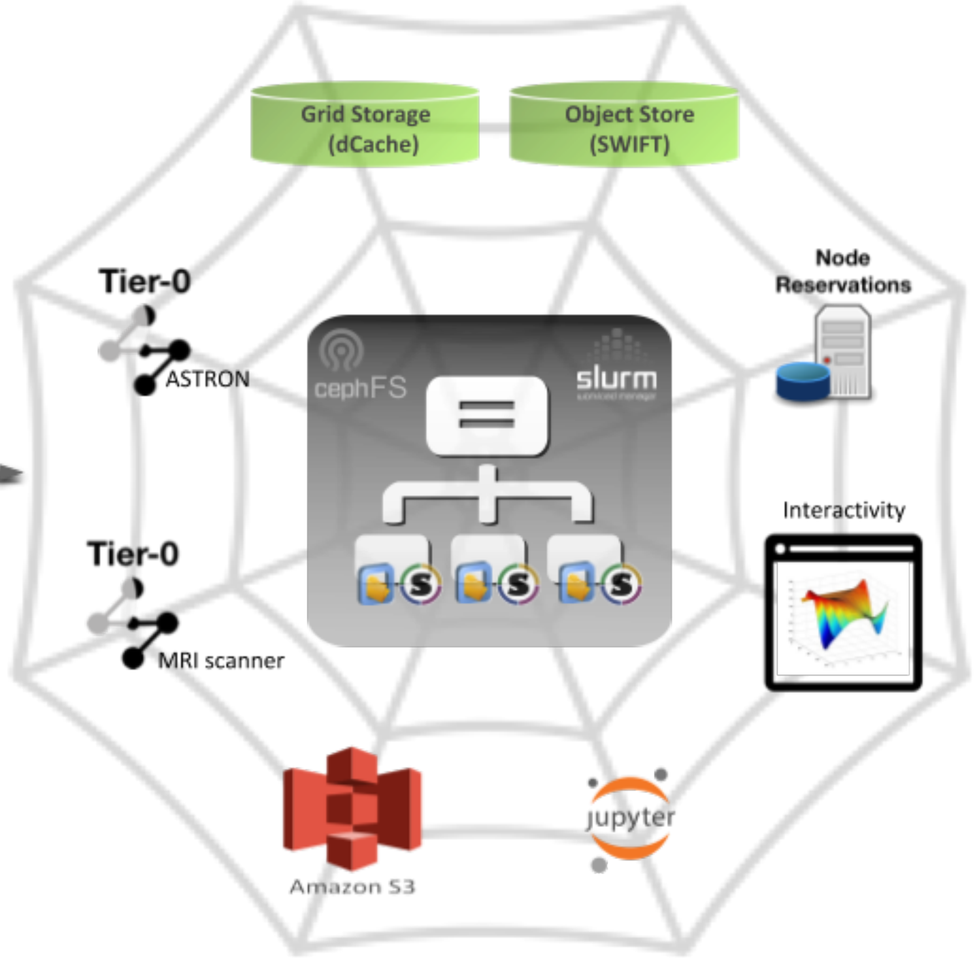
Pierre Auger

LOFAR EOSC

TROPOMI

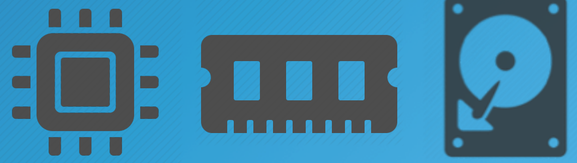
LSG Genetics

AAI



Grid

Spider



Infrastructure

openstack. ceph

Spider – aims & options for LS

* HT platform

- analyse large (>>1TB) data collections
  high memory, SSD disks, shared filesystem & storage
  Slurm scheduler + long running jobs

Spider – aims & options for LS

* HT platform

- analyse large (>>1TB) data collections
  high memory, SSD disks, shared filesystem & storage
  Slurm scheduler + long running jobs

* **Software**

Galaxy

- **reproducible workflows – containers & CVMFS**
  support e.g., *Chipster, Snakemake, Galaxy, DIRAC, Rucio, PiCas*
  std. protocols (e.g., CURL, rest-API's for data management)

SURF SARA

Spider – aims & options for LS

* **HT platform**

- **analyse large (>>1TB) data collections**
  high memory, SSD disks, shared filesystem & storage
  Slurm scheduler + long running jobs

* **Software**

Galaxy

- **reproducible workflows – containers & CVMFS**
  support e.g., *Chipster, Snakemake, Galaxy, DIRAC, Rucio, PiCas*
  std. protocols (e.g., CURL, rest-API's for data management)

* **Collaboration**

- **project roles & associated rights**
  project space *(/Data , /Software, /Shared, /Public)*
  scientific catalogs (e.g. UK Biobank)

SURF SARA

Spider – aims & options for LS

* **HT platform**

- **analyse large (>>1TB) data collections**
  high memory, SSD disks, shared filesystem & storage
  Slurm scheduler + long running jobs

* **Software**

Galaxy

- **reproducible workflows – containers & CVMFS**
  support e.g., *Chipster, Snakemake, Galaxy, DIRAC, Rucio, PiCas*
  std. protocols (e.g., CURL, rest-API's for data management)

* **Collaboration**

- **project roles & associated rights**
  project space *(/Data , /Software, /Shared, /Public)*
  scientific catalogs (e.g. UK Biobank)

* **Interactive** jupyter

- **Jupyter NB, public webviews, interactive Slurm jobs (w. graphics)**

SURF SARA

Spider – aims & options for LS

* **HT platform**

- **analyse large (>>1TB) data collections**
  high memory, SSD disks, shared filesystem & storage
  Slurm scheduler + long running jobs

* **Software**

Galaxy

- **reproducible workflows – containers & CVMFS**
  support e.g., *Chipster, Snakemake, Galaxy, DIRAC, Rucio, PiCas*
  std. protocols (e.g., CURL, rest-API's for data management)

* **Collaboration**

- **project roles & associated rights**
  project space *(/Data , /Software, /Shared, /Public)*
  scientific catalogs (e.g. UK Biobank)

* **Interactive**

jupyter

- **Jupyter NB, public webviews, interactive Slurm jobs (w. graphics)**

* **Security**

SECURE

- **encrypted data transfers, secure light-paths, ISO 27001 certified**

SURF SARA

Spider – aims & options for LS

* **HT platform**

- **analyse large (>>1TB) data collections**
  high memory, SSD disks, shared filesystem & storage
  Slurm scheduler + long running jobs

* **Software**

Galaxy

- **reproducible workflows – containers & CVMFS**
  support e.g., *Chipster, Snakemake, Galaxy, DIRAC, Rucio, PiCas*
  std. protocols (e.g., CURL, rest-API's for data management)

* **Collaboration**

- **project roles & associated rights**
  project space *(/Data , /Software, /Shared, /Public)*
  scientific catalogs (e.g. UK Biobank)

* **Interactive** jupyter

- **Jupyter NB, public webviews, interactive Slurm jobs (w. graphics)**

* **Security** SECURE

- **encrypted data transfers, secure light-paths, ISO 27001 certified**

* **Customize** elixir

- **private nodes (e.g. host private DB) / dedicated clusters**

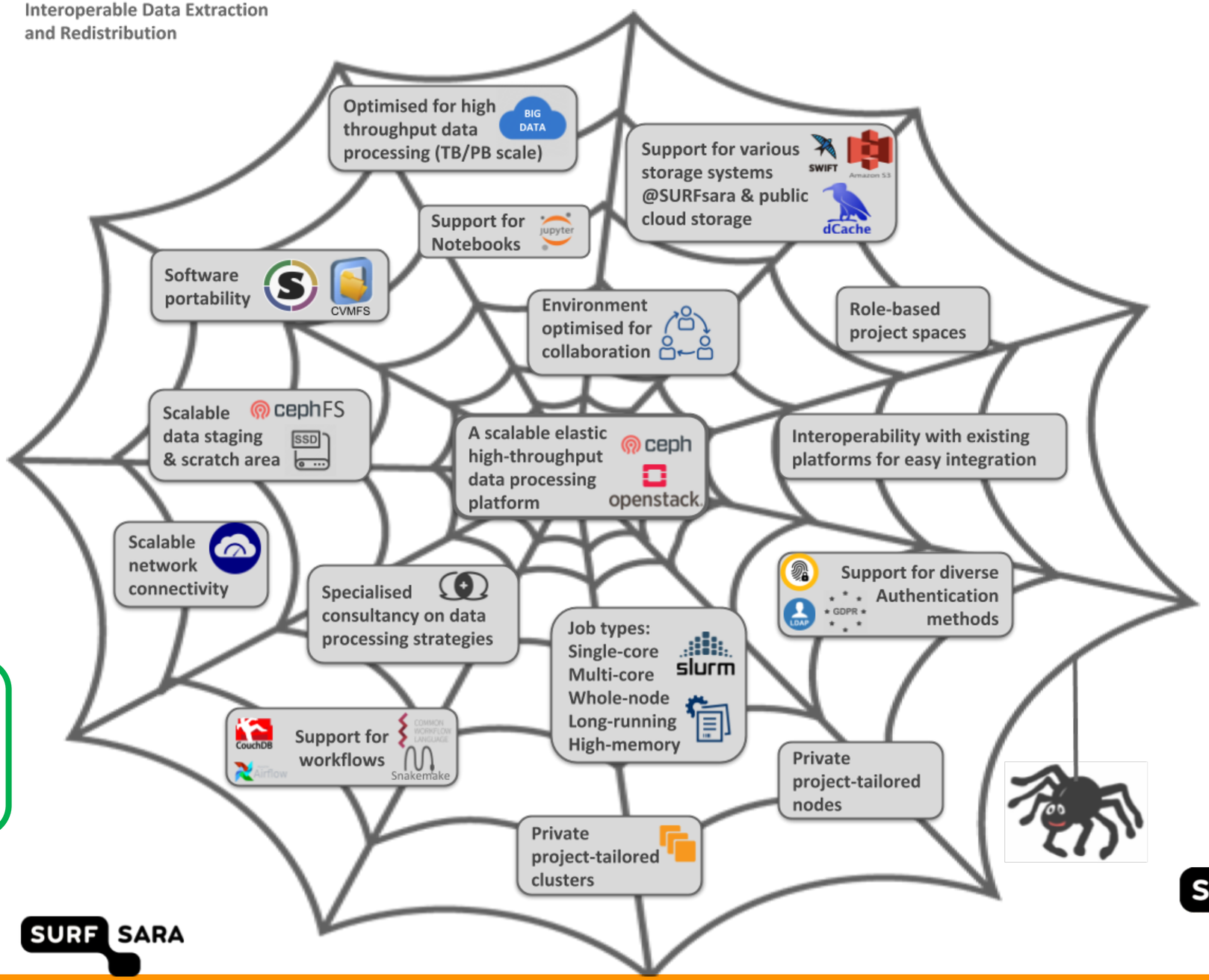
SURF SARA

Phase 1

The ‘SPIDER’ project

Symbiotic Platform(s) for Interoperable Data Extraction and Redistribution

“Interoperable”

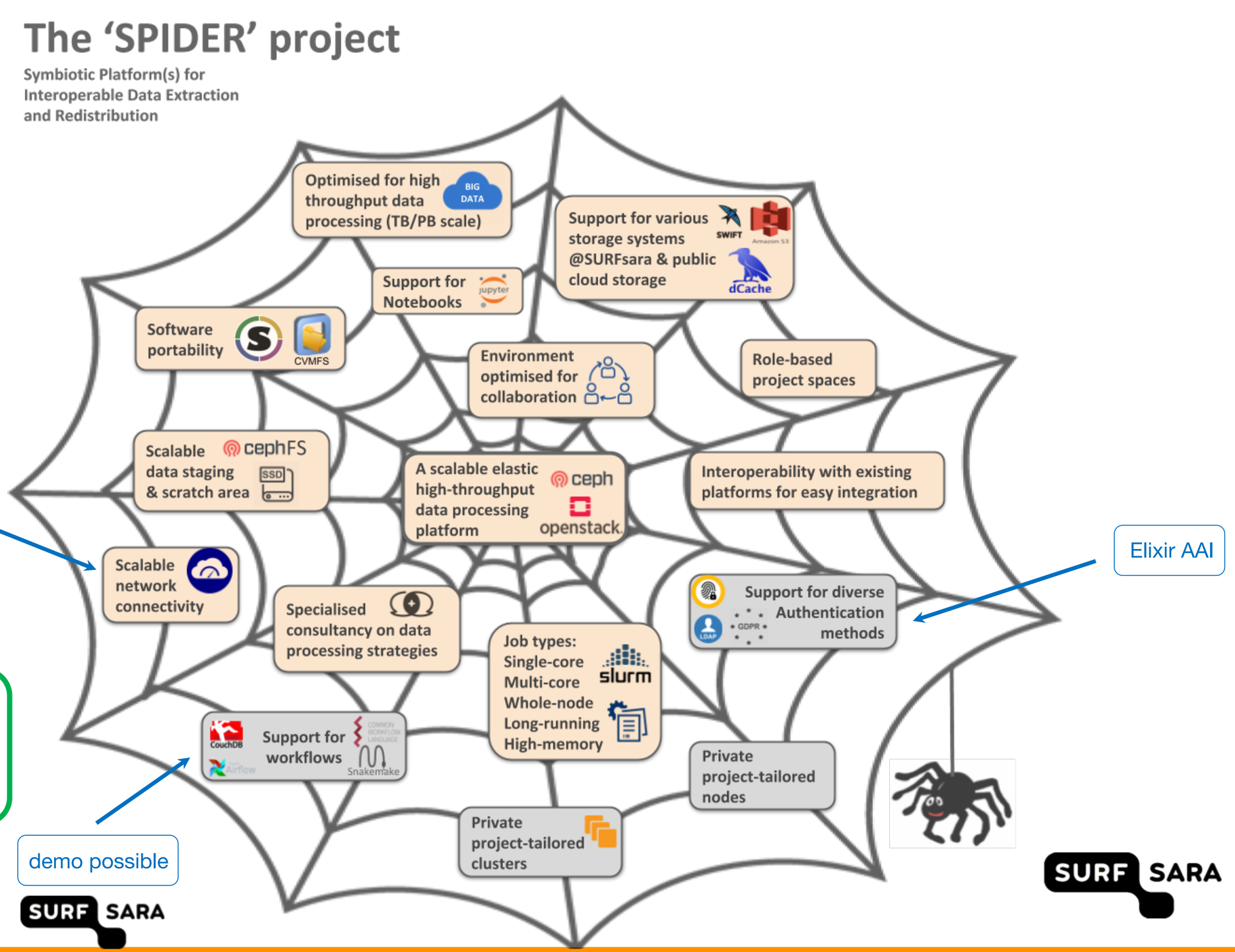


SURFsara is ISO 27001 certified and meets the high standards for information security

SURF SARA

SURF SARA

Current (beta)



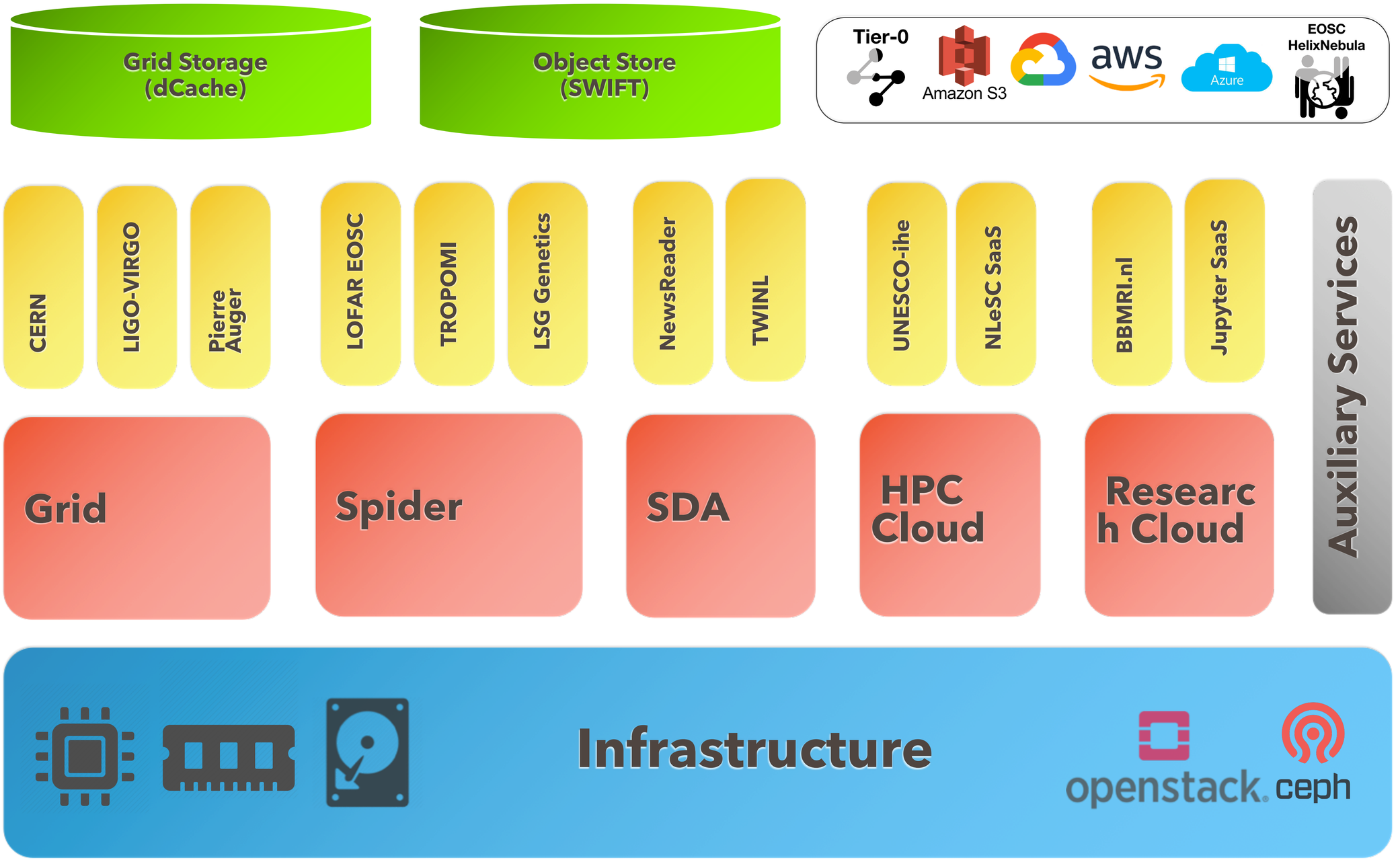
Dedicated light-paths
Encrypted transfers

Elixir AAI

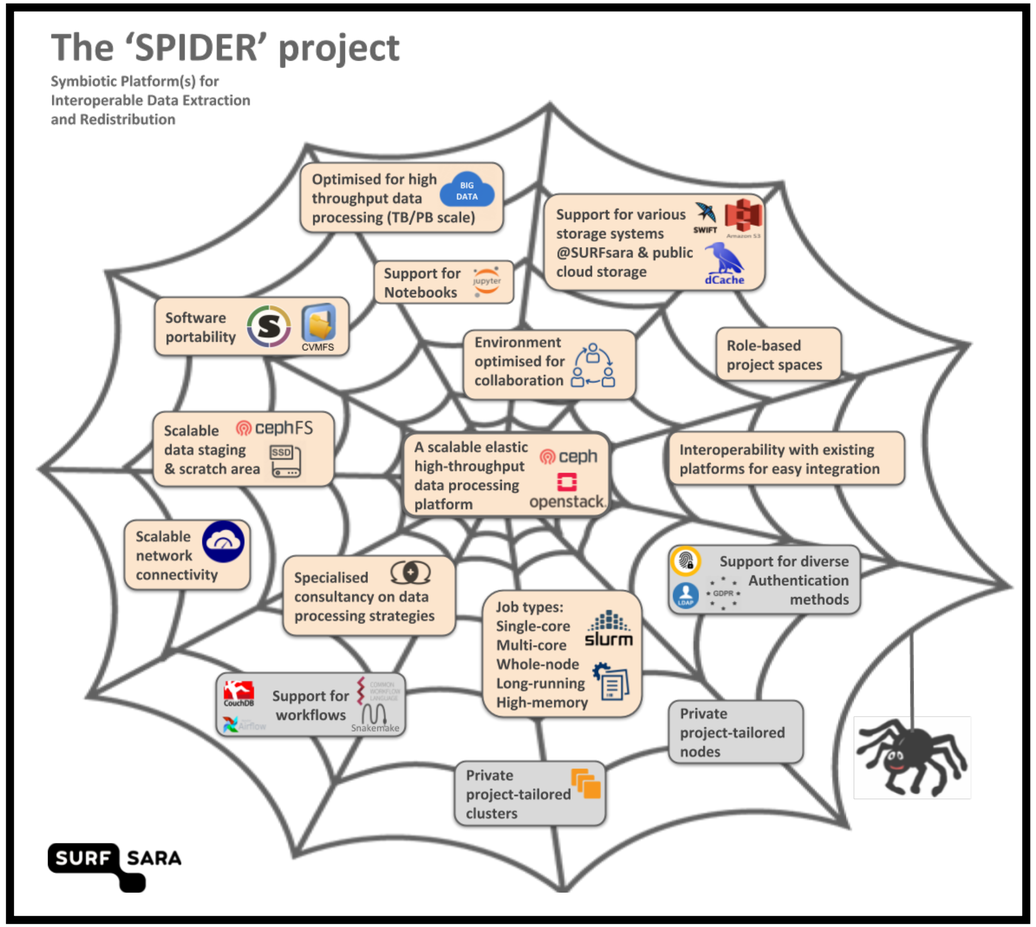
SURFsara is ISO 27001 certified and meets the high standards for information security

demo possible

SURF SARA


Grid Storage (dCache)
Object Store (SWIFT)
Tier-0
Amazon S3
aws
Azure
EOSC HelixNebula
CERN
LIGO-VIRGO
Pierre Auger
LOFAR EOSC
TROPOMI
LSG Genetics
NewsReader
TWINL
UNESCO-ihe
NLeSC SaaS
BBMRI.nl
Jupyter SaaS
Auxiliary Services
Grid
Spider
SDA
HPC Cloud
Research Cloud
Infrastructure
openstack.
ceph
Seamless Integration
Consultancy & Support

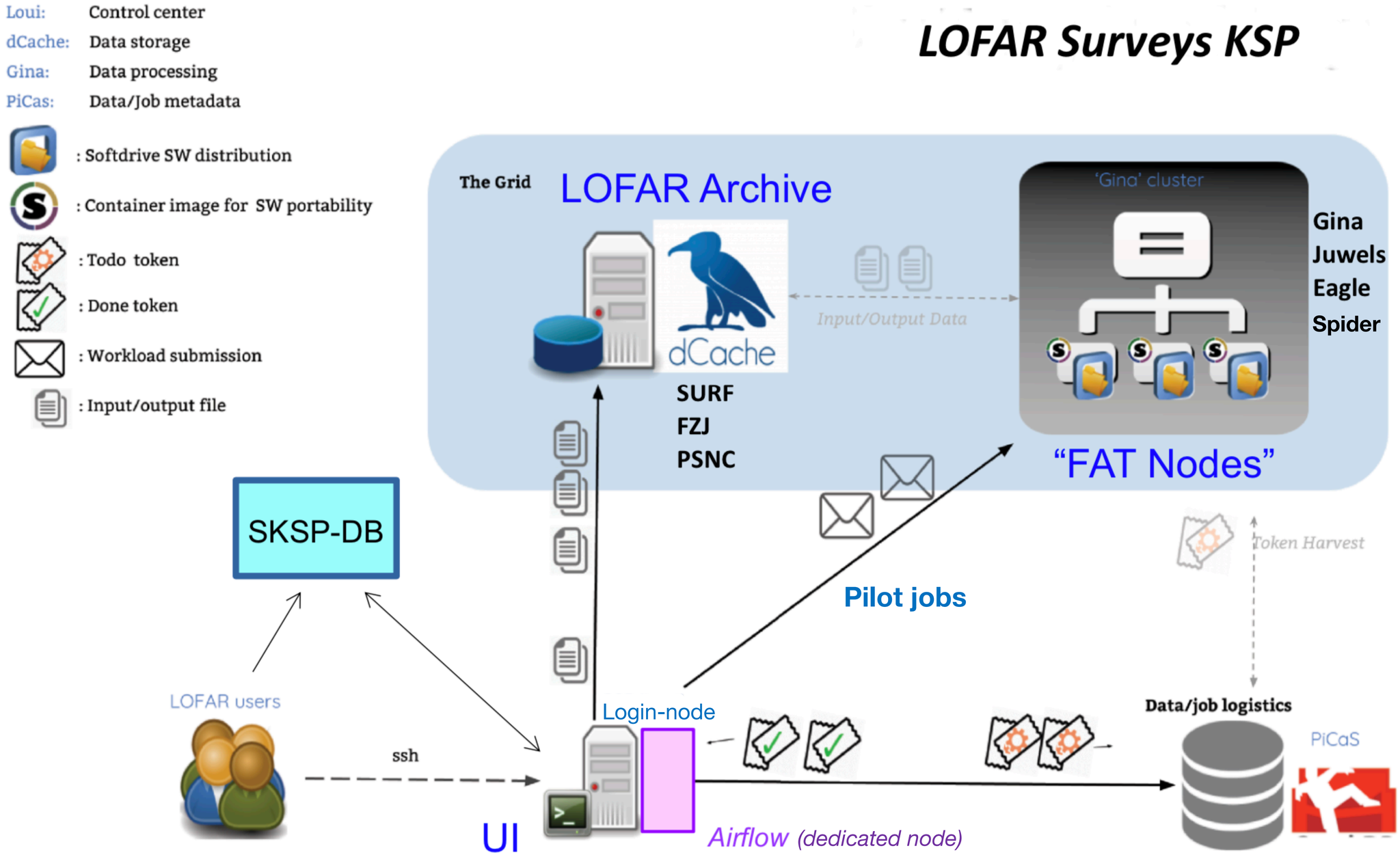
Please go to:



https://github.com/sara-nl/spidercourse/

SURF SARA

Demo


Loui: Control center
dCache: Data storage
Gina: Data processing
PiCas: Data/Job metadata
: Softdrive SW distribution
: Container image for SW portability
: Todo token
: Done token
: Workload submission
: Input/output file
LOFAR Surveys KSP
The Grid
LOFAR Archive
dCache
SURF
FZJ
PSNC
Input/Output Data
'Gina' cluster
Gina
Juwels
Eagle
Spider
"FAT Nodes"
SKSP-DB
Pilot jobs
Token Harvest
LOFAR users
ssh
Login-node
UI
Airflow (dedicated node)
Data/job logistics
PiCaS