

UNIVERSITY OF TWENTE.

Lowering the barrier for modern cloud-based geospatial (big) data analysis: The Geospatial Computing Platform

dr. ing. Serkan Girgin MSc
Center of Expertise in Big Geodata Science

s.girgin@utwente.nl



Geospatial data is getting bigger

- **Large** and **complex** geospatial big data sets are difficult to handle using **traditional systems and methods** to analyse and extract information.
- Numerous spatial computing methods and systems **have been developed** to tackle the difficulties and enable **discovery, delivery, analysis, and visualization** of geospatial data.
- However, data processing and analysis tasks are **still time consuming**, sometimes even **not possible**.

Solutions require specialized know-how and expertise, as well as adequate infrastructure

- **Small scale** (e.g. local, regional) studies with medium size data
 - Analyses can be done faster by **parallel computing** on a workstation
- **Machine learning and AI** studies with medium size data
 - Analyses require **special processing units** (e.g., GPU/TPU) due to computational complexity
- **Large scale** (e.g. national, continental, global) studies with big data
 - Analyses require **distributed computing** on a computing cluster due to computational complexity and/or large volume of data

Cloud computing allows access to required infrastructure, but it requires a transition in **modus operandi!**

Not all research problems require cloud computing and big data technologies

- Research institutions are usually heterogeneous with respect to interests and needs
- For some people these topics **are not/will not be interesting**.
- Even if there is no apparent need or interest, it is **still important** to have at least a basic know-how, because these topics are **becoming a key component** in the research domain.
- This is an institutional priority for ITC.

Center of Expertise in Big Geodata Science aims to enable better use of technology

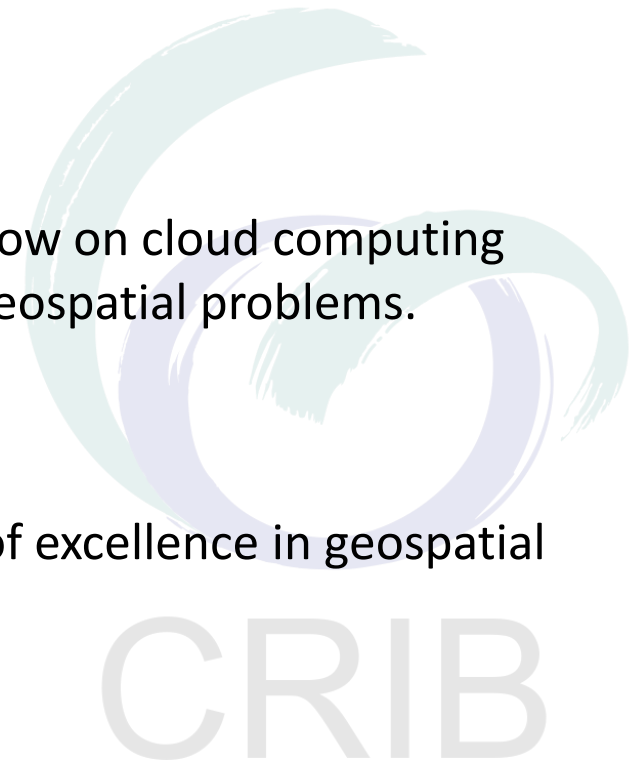
Center of Expertise in Big Geodata Science (CRIB) was established in March 2020 to enable the better use of geospatial cloud computing and big data technologies in education, research, and institutional strengthening activities at ITC.

Mission

Collect, develop, and share operational know-how on cloud computing and big data technologies to solve large-scale geospatial problems.

Vision

Position UT/ITC as a globally renowned center of excellence in geospatial cloud computing and big data science.



We perform activities in five main pillars

- Research

Research and consultancy for integration and better use of cloud computing and big data technologies.

- Capacity and Knowledge Development

Theoretical and hands-on training to improve expert knowledge.
Facilitation of the community of practice.

- Infrastructure Development

Providing an easy to use infrastructure for geospatial computing.

- Monitoring and Networking

Monitoring recent developments in geospatial big data and cloud computing.
Networking with data providers, developers, and research institutions.

- Visibility

Ensuring high visibility of big-data and cloud computing related activities.

User needs assessment indicates the needs for a modern geocomputing infrastructure

- Key findings of the **status quo analysis** and **user needs assessment***
- Information on big data technology (BDT) should be actively communicated to the staff and students
- Proficiency of the staff and students should be improved
- Easy-to-use computing infrastructure should be made available
- Research projects should be enhanced and improved with BDT
- BDT know-how should be transferred to alumni and partners

UT does not provide a common computing infrastructure

ITC did not have a common (geospatial) computing infrastructure

ITC departments have their own computing solutions

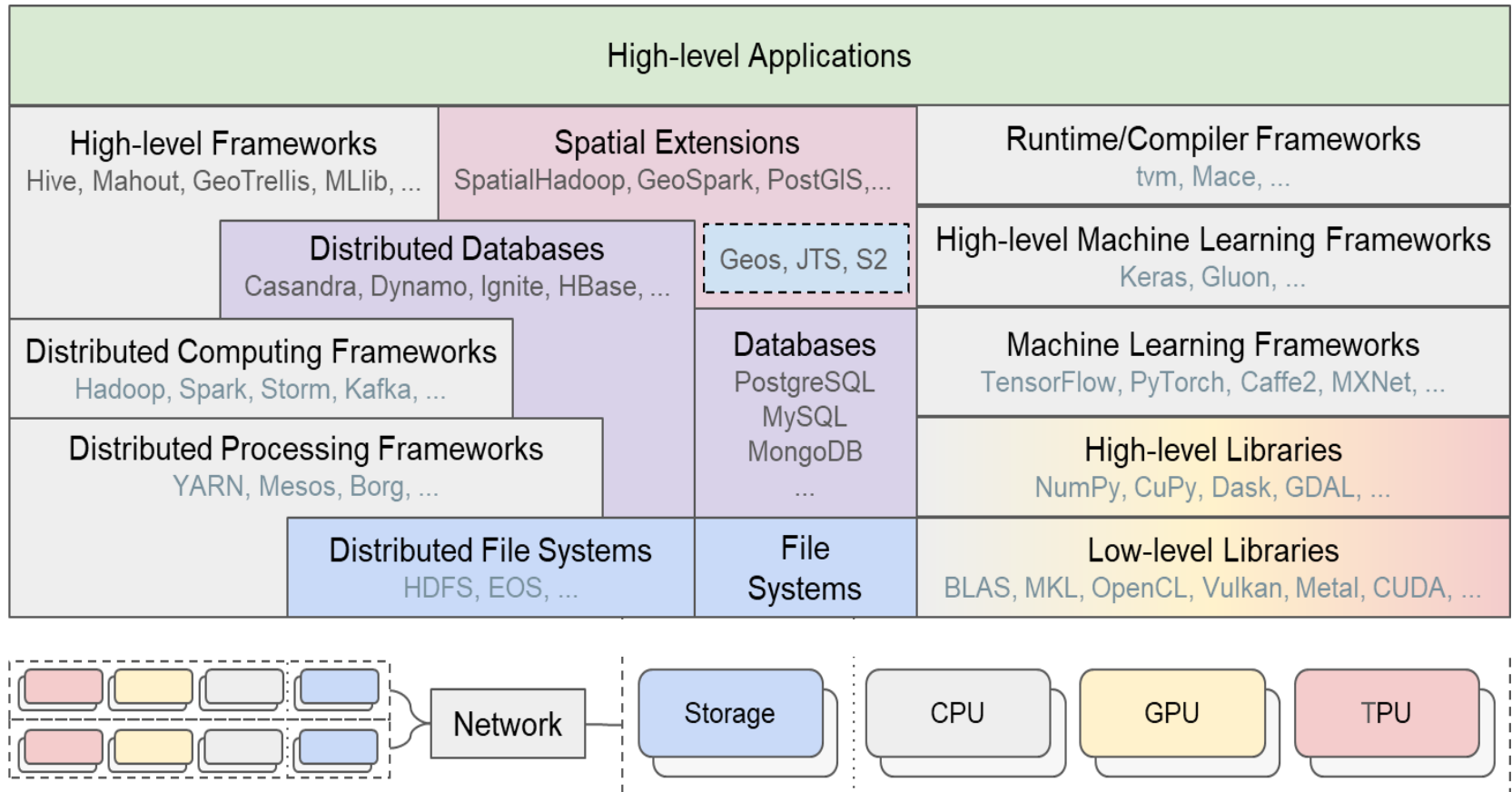
- Operation and management practices differ
- Usually managed by staff who have other primary roles
- Usually access is restricted

* [Girgin, S. \(2020\) Big Geodata at ITC: Status Quo and Roadmap](#)

Geospatial Computing Platform is designed to serve the needs of the user community

- Designed for the identified primary activities
 - Self learning
 - Exploratory research
 - Education
- Designed for the identified primary criteria
 - Highly available 24/7, no queue
 - Ready to use Pre-installed software
 - User friendly Interactive user interface
 - GPU enabled GPU for each user
 - Distributed-computing friendly Computing cluster
 - Low-cost Feasible investment

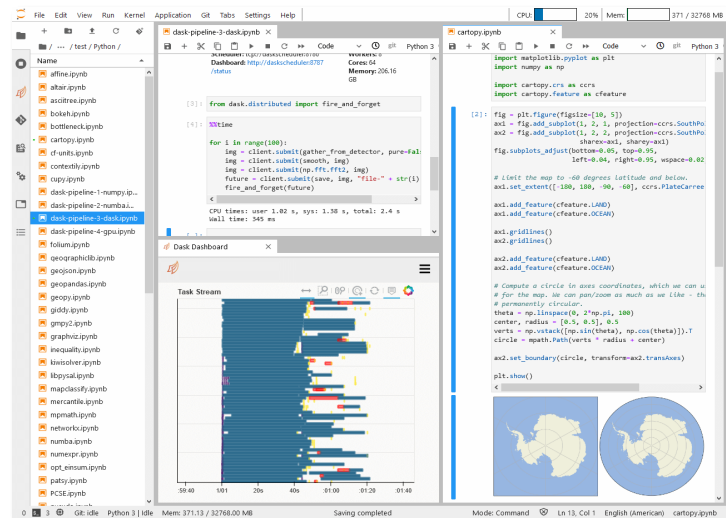
The platform aims to enable access to a full big geodata processing hardware and software stack



We utilized innovative solutions to develop a platform fulfilling the criteria



NVIDIA Jetson AGX Xavier Cluster



JupyterHub on Docker Swarm

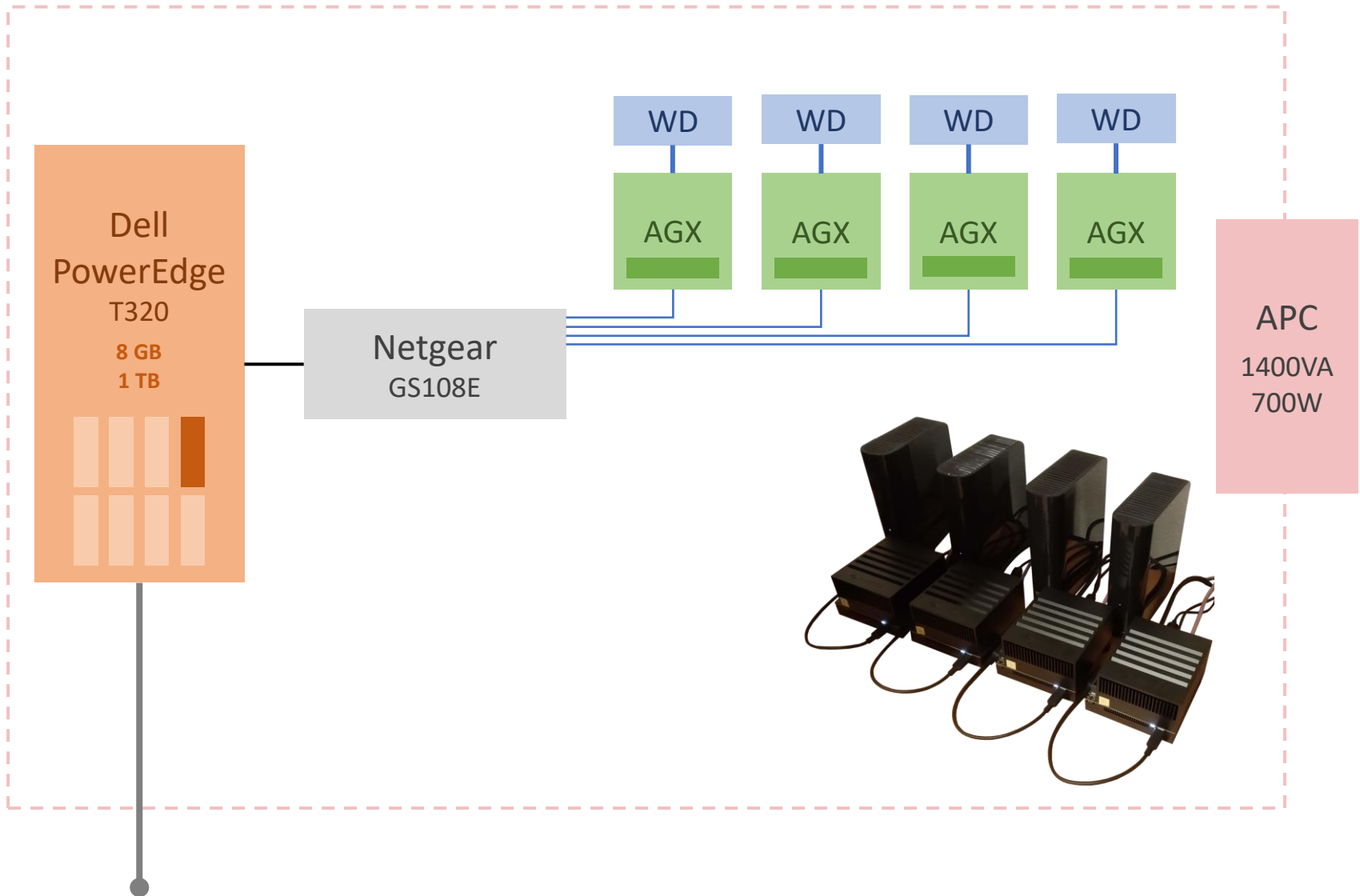
Jetson AGX Xavier is tiny, but it is powerful

- 8-core CPU
(NVIDIA Carmel ARMv8.2, 2.26GHz, NVIDIA L4T)
- 512-core GPU
(Volta Architecture with 64 Tensor Cores)
- 32GB memory
(256-bit LPDDR4x, 2133MHz, 137GB/s, Unified)
- 32GB storage
(eMMC 5.1)
- Dual Deep Learning Accelerator
- Vision Accelerator
- 4x 4Kp60 video encoder
- 2x 8Kp30 / 6x 4Kp60 video decoder
- Gigabit Ethernet



For more information: https://elinux.org/Jetson_AGX_Xavier

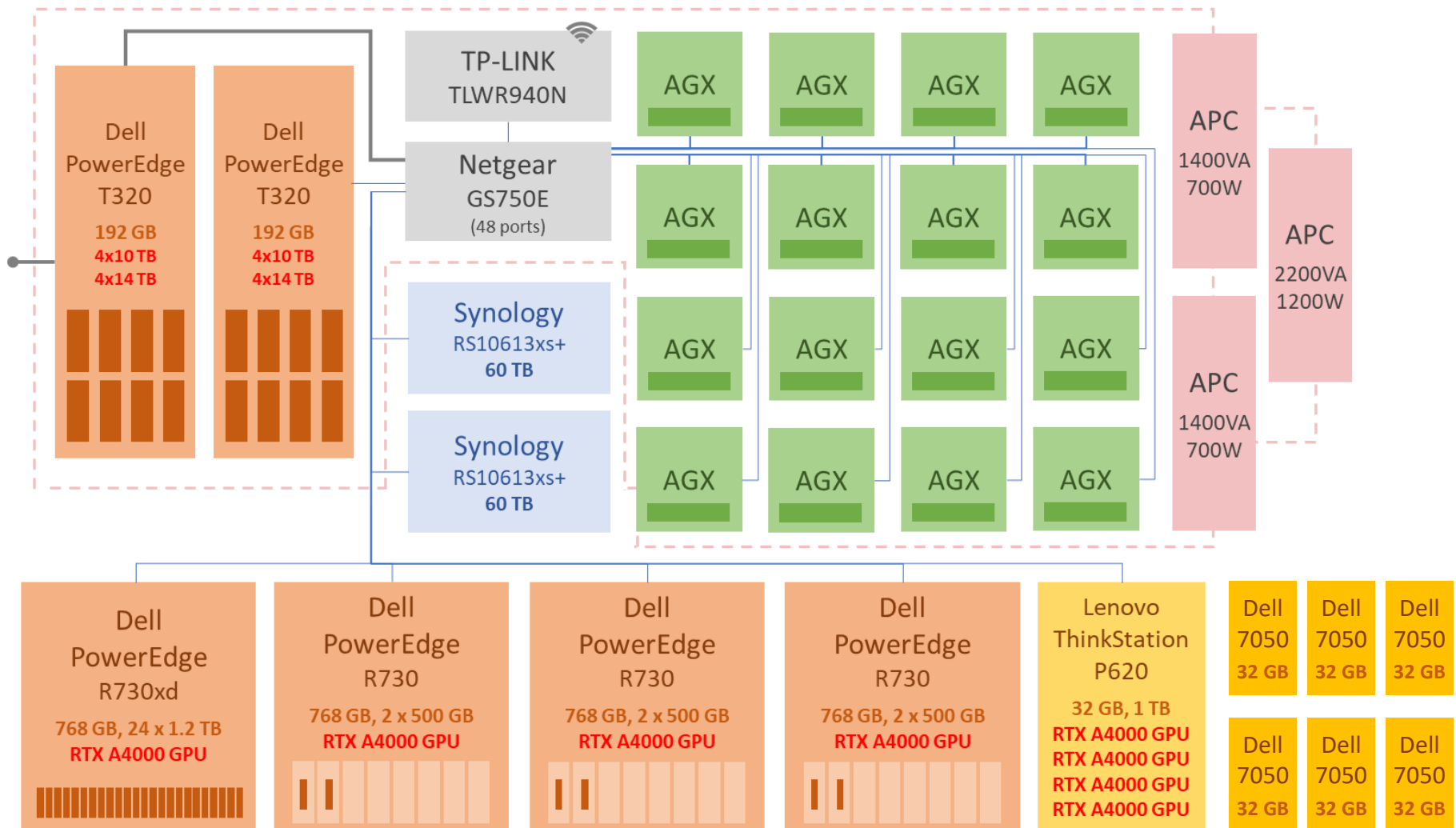
One can start small, even if the target is big...



... but it is important to have an expansion policy

- Grow only if there is meaningful demand.
- Upgrade first, then expand.
- Repurpose idle resources to make them available for common use.
- Select low-cost, good-performance solutions.
- Use refurbished equipment if possible.

Today we have an efficient shared infrastructure



* As of 13 December 2022

A wide-range of computing units are available

- **16 x GPU-enabled General Purpose Computing Units**
8-CPU NVIDIA Carmel ARMv8.2 @ 2.26GHz, 32 GB RAM, 512-core Volta GPU, 64 Tensor Cores
- **6 x General Purpose Computing Unit**
8-CPU Intel Core i7-7700 @ 3.60GHz (max. 4.20GHz), 32 GB RAM
- **3 x GPU-enabled Big Data Computing Units**
72-CPU Intel Xeon E5-2695 v4 @ 2.10GHz (max. 3.30GHz), 768 GB RAM, NVIDIA RTX A4000 GPU
- **1 x GPU-enabled Big Data Computing Unit with Fast Storage**
32-CPU Intel Xeon E5-2640 v3 @ 2.60GHz (max. 3.40GHz), 768 GB RAM, NVIDIA RTX A4000 GPU
22 TB RAID 20+2
- **1 x Multi-GPU Computing Unit**
32-CPU AMD Ryzen Threadripper PRO 3955WX @ 3.9GHz, 160 GB RAM, 4 x NVIDIA RTX A4000 GPU
- **2 x Servers**
12-CPU Intel Xeon E5-2420 v2 @ 2.20GHz (max. 2.7GHz), 192 GB RAM, 48 TB RAID 2+1 (ZFS)
- **2 x Storage Servers**
Synology, 60 TB

Resource sharing is at the core of the platform

- Accessible through a **web browser** (No software installation or VPN are required)
- **No registration** is required (Login with the University credentials)
- Each user has an individual and isolated **working environment**
- Each user has access to all available* **unit resources**, including **GPU**
- Each user has access to all available* **cluster resources**
- **Replicated storage** with minimum two copies (Hardware failure protection, ZFS)
- **Distributed storage** for big data processing (HDFS)
- Automatically **balances workload** among the units

<https://crib.utwente.nl>

* Resource availability depends on resource usage of other active user

It provides features to simplify research activities

- **Interactive notebook, terminal and remote desktop** access are available
- Multiple **interactive languages** are supported (Python, R, Julia, Octave, Go, ...)
- **Up-to-date** and **optimized** software packages are **ready to use** (No setup required)
- Users can install **additional** packages (e.g., Python, R packages)
- Distributed computing clusters are **ready to use** (Dask, Apache Spark)
- **Public** assets are shared by all users (e.g., OSM, NL 0.5m DTM, TOP10-1000, ...)
- **Shared workspaces** allow assets to be shared by selected users
- Access can be granted to **external users**
- **User support** is available
- Provided and maintained by **CRIB** at no extra cost (i.e., free PaaS)



Hundreds of software packages are available ready to use

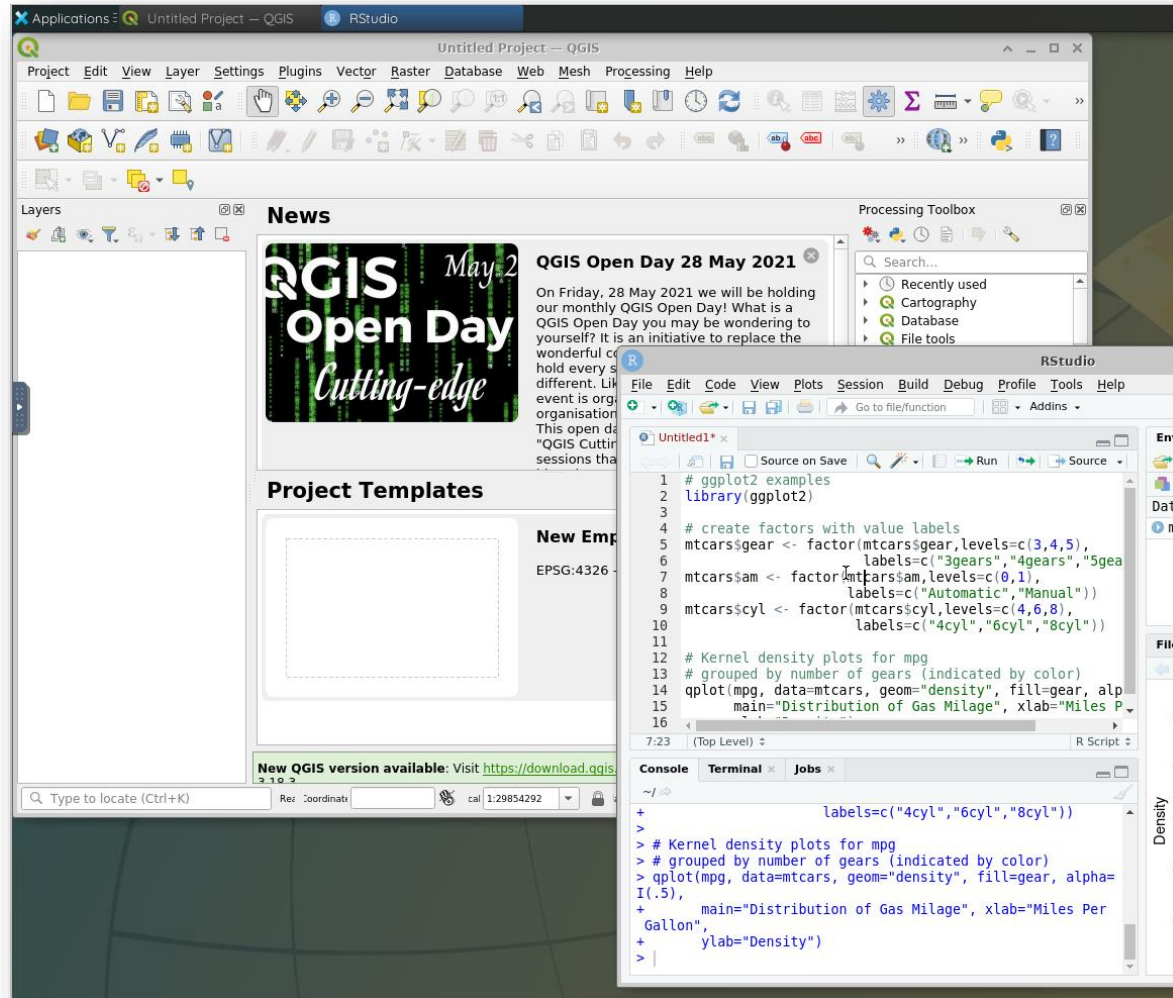


... and many more!

(900+ Python and 500+ R packages)

Desktop applications are available as well

- QGIS
- GRASS GIS
- SAGA GIS
- OTB
- ENVI*
- SARscape*
- SNAP
- ILWIS 3*
- ILWIS 4*
- VS Code
- PyCharm
- R Studio
- Netlogo
- GNU Octave
- MATLAB*
- Glueviz
- Orange Data Mining



Existing well-established technologies are not forgotten



GeoServer

Open source server for sharing
geospatial data



MapServer

Open source platform for
publishing spatial data



PostgreSQL

Open source relational database



MariaDB

Open source relational database



GeoNode

Open source geospatial content
management system



Dataverse

Open source research data
repository software



Gitea

Open source lightweight code
hosting solution



Open Data Kit

Open source platform to collect
data quickly, accurately, offline, and
at scale

24/7 support is available through the support center

Search our knowledge base

Search

Open a New Ticket

Check Ticket Status

Welcome to the CRIB Support Center!

In order to streamline support requests and better serve you, we utilize a support ticket system. Every support request is assigned a unique ticket number which you can use to track the progress and responses online. For your reference we provide complete archives and history of all your support requests.

Quick Access

- [Report a Problem](#)
- [Shared Workspace Request](#)
- [Course Workspace Registration with Canvas Integration](#)
- [External Account Request](#)
- [Account Removal Request](#)
- [Account Transfer Request](#)
- [Software Request](#)
- [Dataset Request](#)
- [Database Request](#)

Featured Questions

[How can I access to the platform?](#)

[Is it secure?](#)

[How can I use the platform?](#)

[Which programming languages are supported on the platform?](#)

[Which libraries and packages are supported by the platform?](#)



We managed to establish a user community

- Operational since **January 2021**
- **865*** registered users
- **94*** shared workspaces for projects and courses
- **15-40*** concurrent users at a time
- **200,000+*** hours of multi-core/GPU computation
- **390+** support tickets* closed (excluding support by e-mail)
- Overall **positive feedback** from a wide-range of use cases
4.61/5.00 according to the [user survey](#)
- UT LISA built a similar platform for university-wide use
Co-developed by CRIB

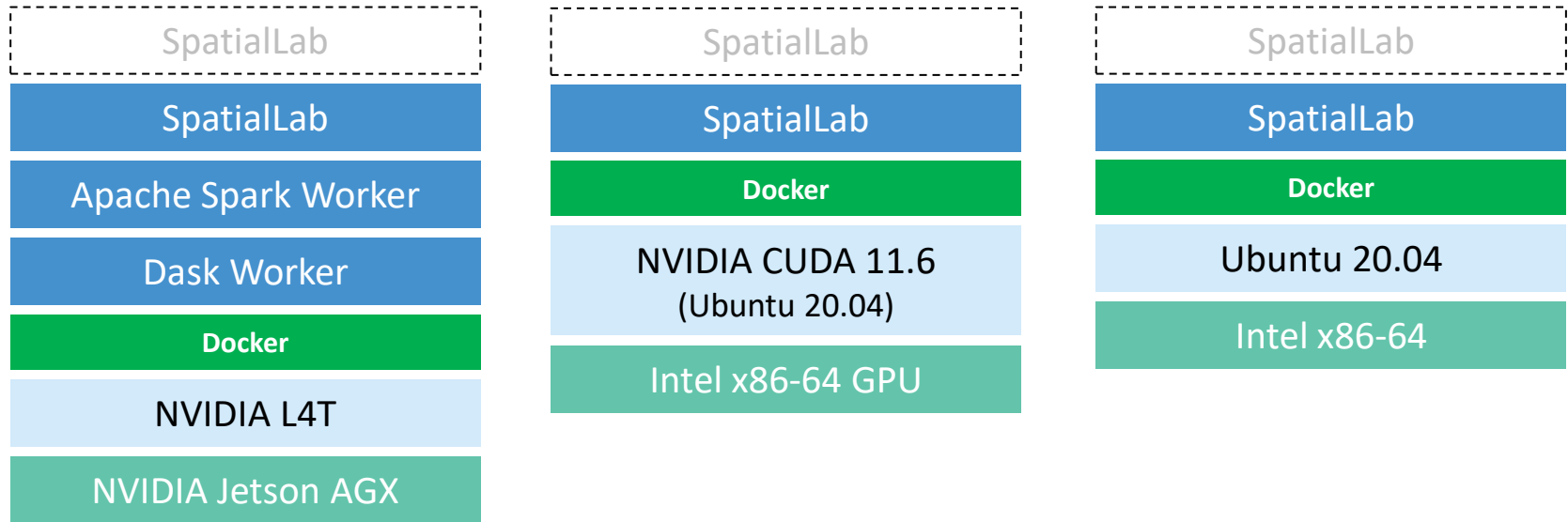
* As of 13 December 2022

Smart operation and maintenance are key for success

- On-demand and bi-monthly regular **rolling updates**
- **Similar** working environment for ARM64 and Intel x86-64 units
- Automated shared workspaces for the **departments**
- Automated shared workspaces for **courses**
- Automated **notification** to newcomers
- **Daily** check for malicious threads
- **Daily** storage snapshots for 7 days
- **Continuous** resource and performance monitoring



Docker Swarm performs the container orchestration



- **Minimum impact** on the host units
- Custom-built **SpatialLab** images for better performance
 - NVIDIA Jetson AGX : **38.8 GB**
 - Intel x86-64 : **59.0 GB**
 - Intel x86-64 GPU : **71.7 GB**



Dilemma of infrastructure – use cases

- Infrastructure \rightarrow Use cases
 - Invest and start hunting for ideas
- Use cases \rightarrow Infrastructure
 - Develop and wait for resources
- Infrastructure \Leftrightarrow Use cases
 - ~~Expensive "most" powerful resources~~
 - ~~Fully developed use cases~~

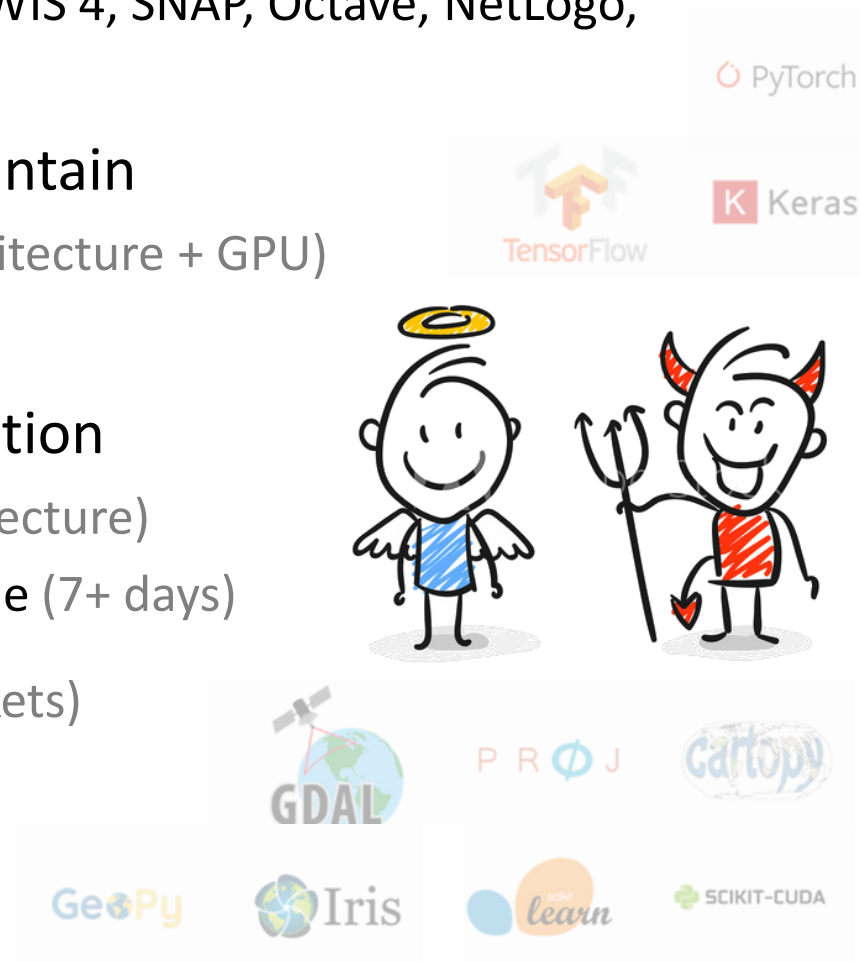


Lessons Learned: Computing Resources

- Innovative = "unproven" solution
 - ✓ Serving quite well (8 core, 32 GB RAM, GPU, in cluster formation)
 - ✓ High performance/cost ratio (≈ 3 EUR/h VM, pays off in 300 h)
 - ✓ Low energy consumption (max. 30W/unit)
 - ⚠ Difficult to set up (software stack was not ready)
 - ⚠ Difficult to maintain (software need to be build)
- Support from the departments / staff
 - ✓ Donation (ESA: servers, GIP: storage servers, M. Yang: GPU)
 - ✓ Donation + shared upgrade (GIP: big data unit)
 - ✗ Sharing (NRS: Jetsons)
- Low-cost solutions
 - ✓ Second-hand equipment

Lessons Learned: Computing Platform

- Rolling update = State of the art
 - ✓ 900+ Python, 500+ R packages (Statistical, Spatial, EO, ML, AI)
 - ✓ QGIS, GRASS, SAGA, ILWIS 3, ILWIS 4, SNAP, Octave, NetLogo, MATLAB eCognition, ENVI*
- State of the art = Difficult to maintain
 - ⚠ Dependency puzzle (multi-architecture + GPU)
 - ⚠ Stability*
- Docker virtualization / orchestration
 - 😊 Works quite well (multi-architecture)
 - 😞 Image rebuild take a lot of time (7+ days)
- ✓ Support center (400+ closed tickets)
- 🌸 Additional services



Lessons Learned: User Penetration

- Computing Platform

- ✓ In use for MSc/PhD studies, courses, projects, and trainings
- ✓ Overall user satisfaction is quite positive
- ⚠ Reaching potential users is not easy
- ⚠ Getting feedback is difficult

★★★★☆
4.37 Average Rating

ID ↑	Name	Responses
1	anonymous	STOP SENDING ME MAILS STOP SENDING ME MAILS

- CRIB

- ⚠ Collaboration exists, but can be improved*
- ✓ Automated e-mail to newcomers and newcomer meetings help

Lessons Learned: User Stories

absquatulate (v):

to leave without saying goodbye

- Best way to convince people is to show examples
- Best examples are in-house from "familiar" faces
- They are not easy to collect



Follow us to get informed about our activities



<https://itc.nl/big-geodata>



[@BigGeodata](https://twitter.com/BigGeodata)

437 followers



[Big Geodata Newsletter](#)

244 subscribers



[CRIB YouTube Channel](#)

124 subscribers

Contact



dr.ing. Serkan Girgin MSc
Senior Researcher
Head of Center of Expertise in
Big Geodata Science
s.girgin@utwente.nl
+31 53 489 55 78