A new distributed data analysis framework for better scientific collaborations

Philipp S. Sommer
Institut für Küstensysteme – Analyse und Modellierung

EGU21-1614, 28.04.2021

Helmholtz-Zentrum
hereon

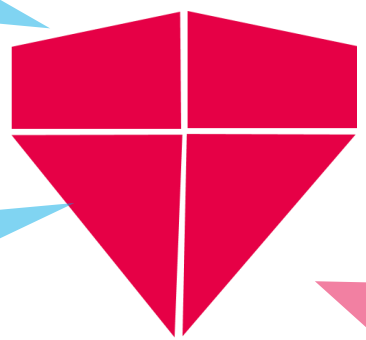
The problem of analyzing distributed data

Philipp S. Sommer, Viktoria Wichert, Daniel Eggert, Tilman Dinter, Klaus Getzlaff, Andreas Lehmann, Christian Werner, Brenner Silva, Lennart Schmidt, and Angela Schäfer

Let's do some research together!

Awesome! Let me run some tests on your model data.

Alright! But that's too much data and …



I can't give you access to my super computer!

No worries, I do my analysis with python, so let's just use the de-messaging-python framework!

DIGITAL EARTH

HELMHOLTZ ZENTRUM FÜR UMWELTFORSCHUNG UFZ

GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel

JÜLICH FORSCHUNGSZENTRUM

GFZ Helmholtz-Zentrum Potsdam

KIT Karlsruher Institut für Technologie

AWI

Hub Terra

Hub Atmo

Hub Mare

Helmholtz-Zentrum hereon

Presentation material and Code

https://github.com/Chilipp/de-messaging-python-presentation-20210428

AK Datenanalyse

Distributed data analysis Working Group within Datahub

Contributors

- **HZG:** Philipp S. Sommer, Viktoria Wichert
- **GFZ:** Daniel Eggert (Digital Earth)
- **AWI:** Tilman Dinter, Brenner Silva, Angela Schäfer
- **Geomar**: Klaus Getzlaff, Andreas Lehmann
- **KIT:** Christian Werner
- **UFZ:** Lennart Schmidt

Helmholtz-Zentrum hereon

What is distributed Data analysis

Examples

Ship campaign

- Sonne (Geomar) and Ludwig Prandtl (HZG) measure real-time-data in a campaign.
- Sonne sends to internal area of Geomar, Ludwig Prandtl to HZG.
- How can people from HZG access and analyze the data at Geomar?

Model simulations

- Compare a COSMO-CLM-Simulation (HZG) with output of the Baltic Sea Model (Geomar)
- And with ship measurements
- How to share terra-bytes of data?
- How to get the latest version?

It's about *analyzing* distributed data

The ideal world

- We all have one single big cloud
  - Run model simulations in the cloud
  - Store NRT data in the cloud
- Post processing and data analysis runs in the cloud
- Someone from HZG needs access to data from Geomar? *Just grant it.*

The real world

- We have many different clusters.
  - Every center (or even every scientist) has different requirements
  - We are behind VPNs
  - Each center has his own cluster for processing, storage, etc.
- Someone from HZG needs access to data from Geomar? *Ok, I upload it to Dropbox.*

Can we do it without the cloud?

What we need:

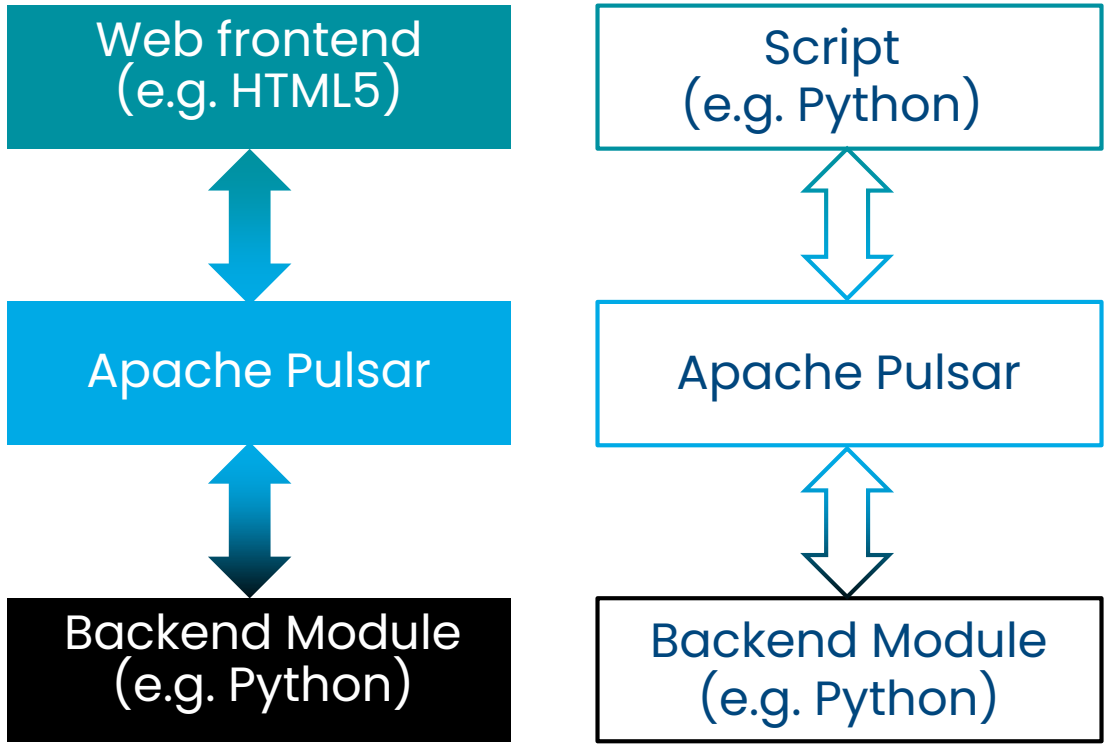
- Access to data in another research center
- Access to computing power in another research center

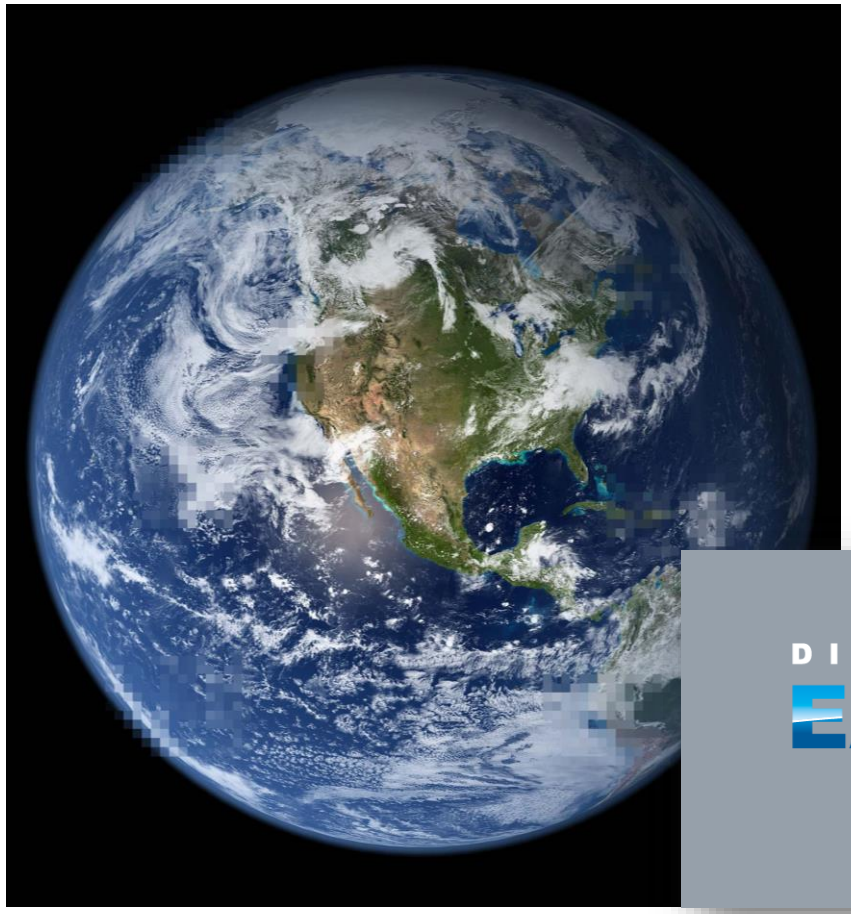
And:

- It must be safe
- It must be easy

We are not the first

with this idea

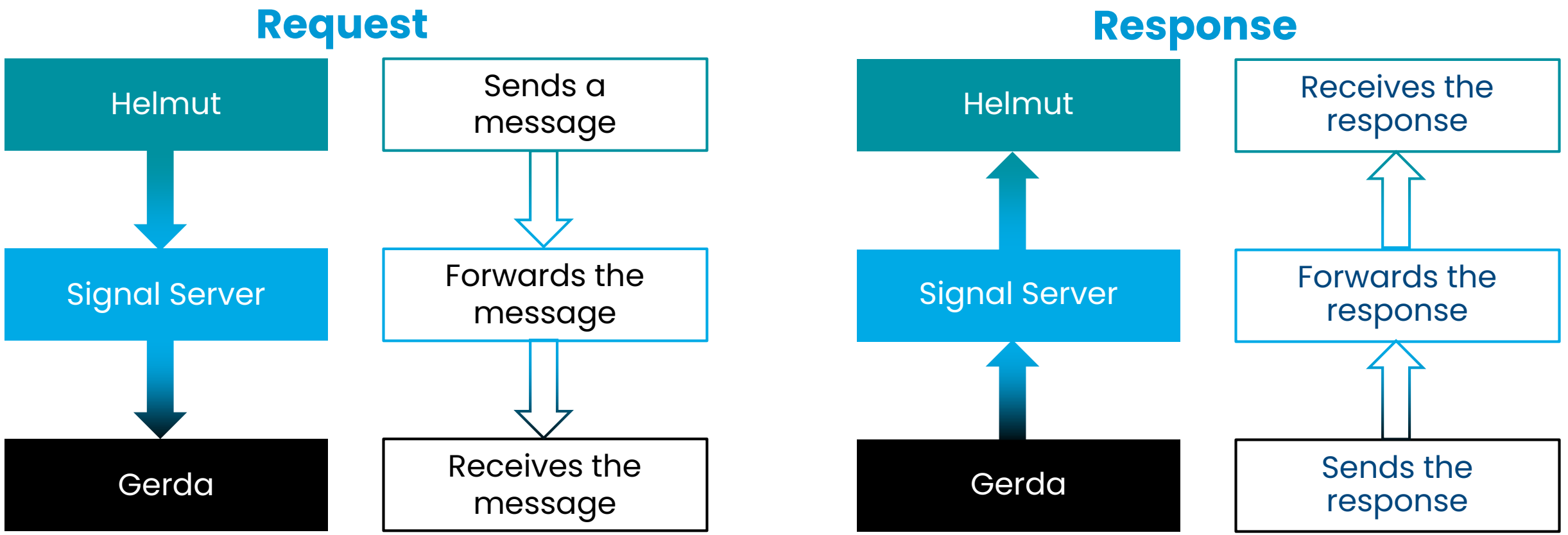




DIGITAL EARTH

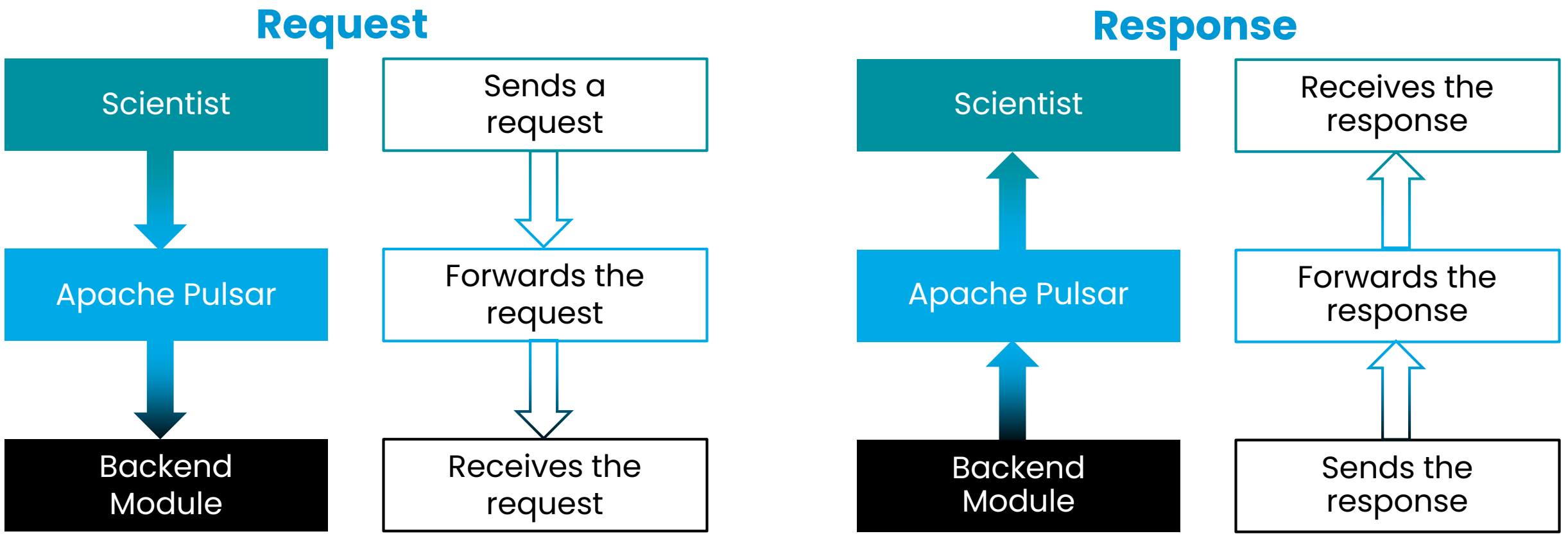
Helmholtz-Zentrum hereon

Just like ~~WhatsApp~~Signal



Helmholtz-Zentrum hereon

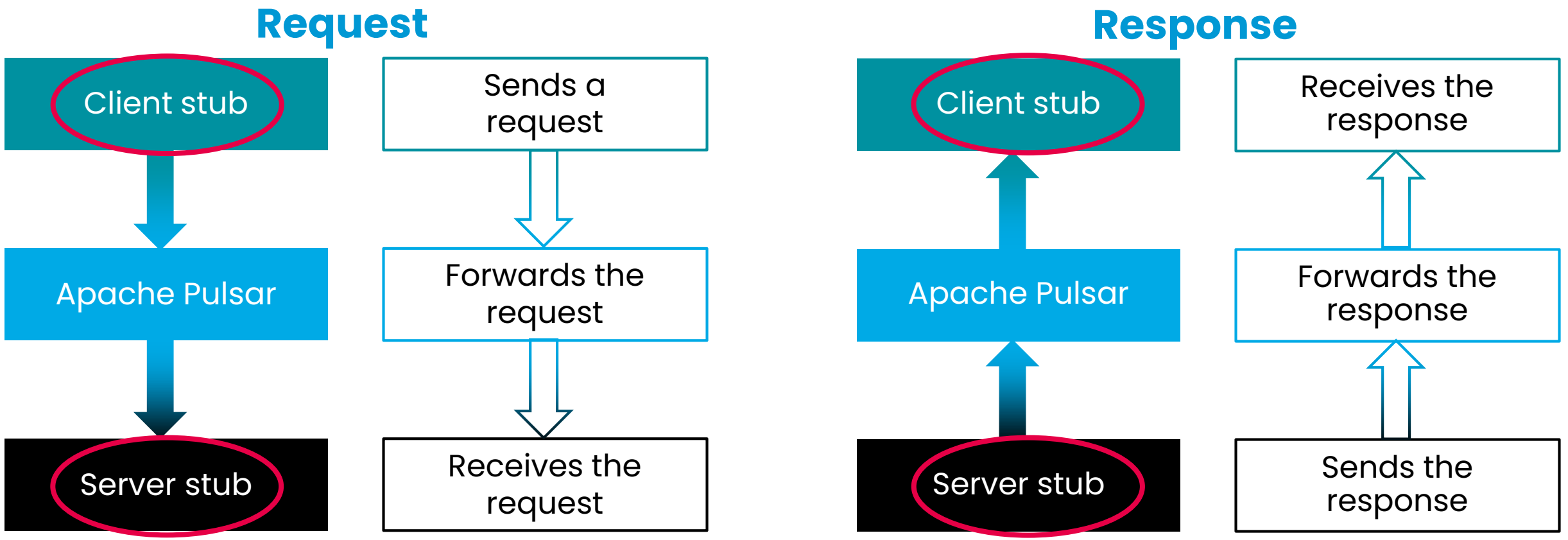
Just like ~~WhatsApp~~Signal



Helmholtz-Zentrum hereon

Just like ~~WhatsApp~~Signal

A Remote Procedure Call (RPC)



Helmholtz-Zentrum hereon

Pros and Cons

Advantages

- Scientist can simply send a request and retrieve the response on any other machine
- Backend Module can run everywhere, not necessarily on a dedicated web server (e.g. on the cluster)

Disadvantages

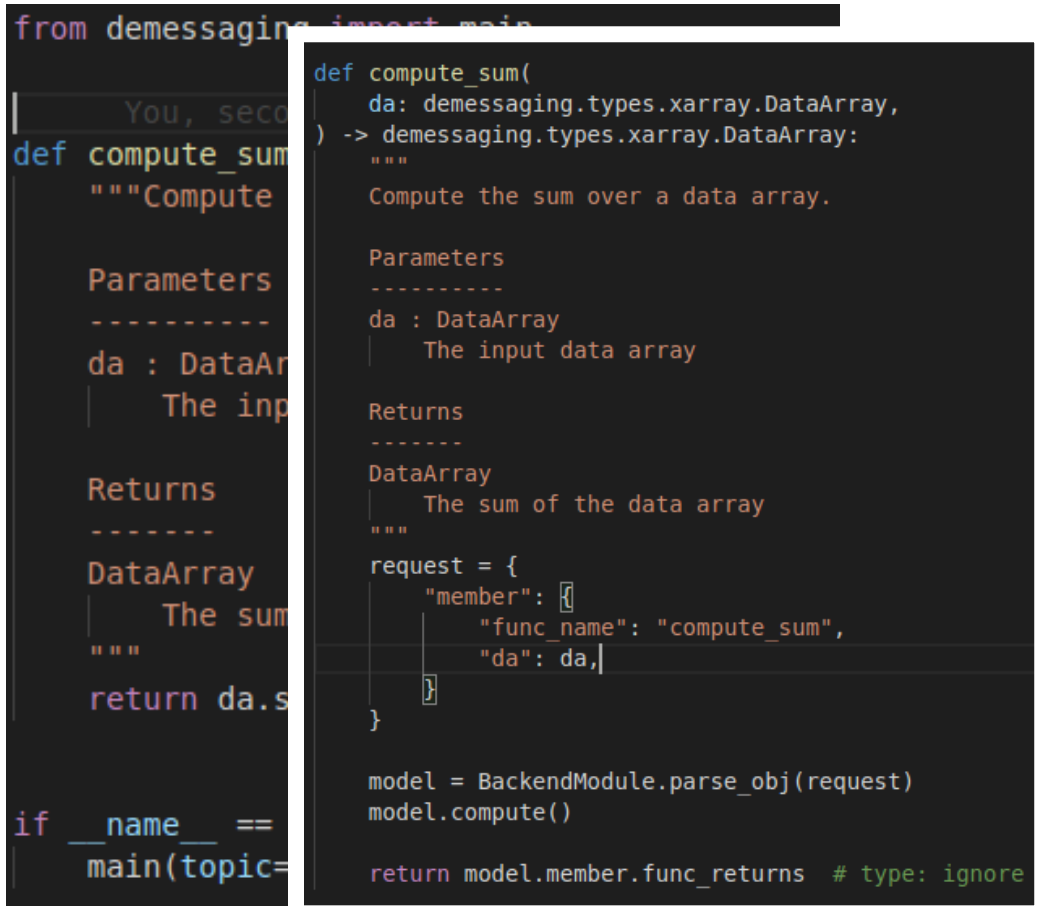
- Scientists are not familiar with web requests (nor are the backend module developers)
- Request needs serialization (transformation to JSON)
- Potential vulnerability for internal computing resources
- Scientists do have better stuff to do

Be nice

and do not add more work

Use the scientists methods

- abstract standard python functions and classes into web requests
- everything's basic python, (almost) no need for special stuff
- Client stub is automatically generated
- Requests are abstracted and standardized (JSONschema)



de-messaging-python

Summary

- Remote Procedure Call
- High-level API to easily create server and client stubs
- Very close to scientists common workflows

Outlook

- More effort into security
  - User management for backends
  - End-to-End encryption
- How to handle large amounts of data
- We are looking for use cases and project that may use our framework!

Thank you!

Vielen Dank.

www.hereon.de

Helmholtz-Zentrum
hereon