FORCE on Nextflow: Scalable Analysis of Earth Observation data on Commodity Clusters

Fabian Lehmann[1], David Frantz[2,4], Sören Becker[3], Ulf Leser[1], Patrick Hostert[4]

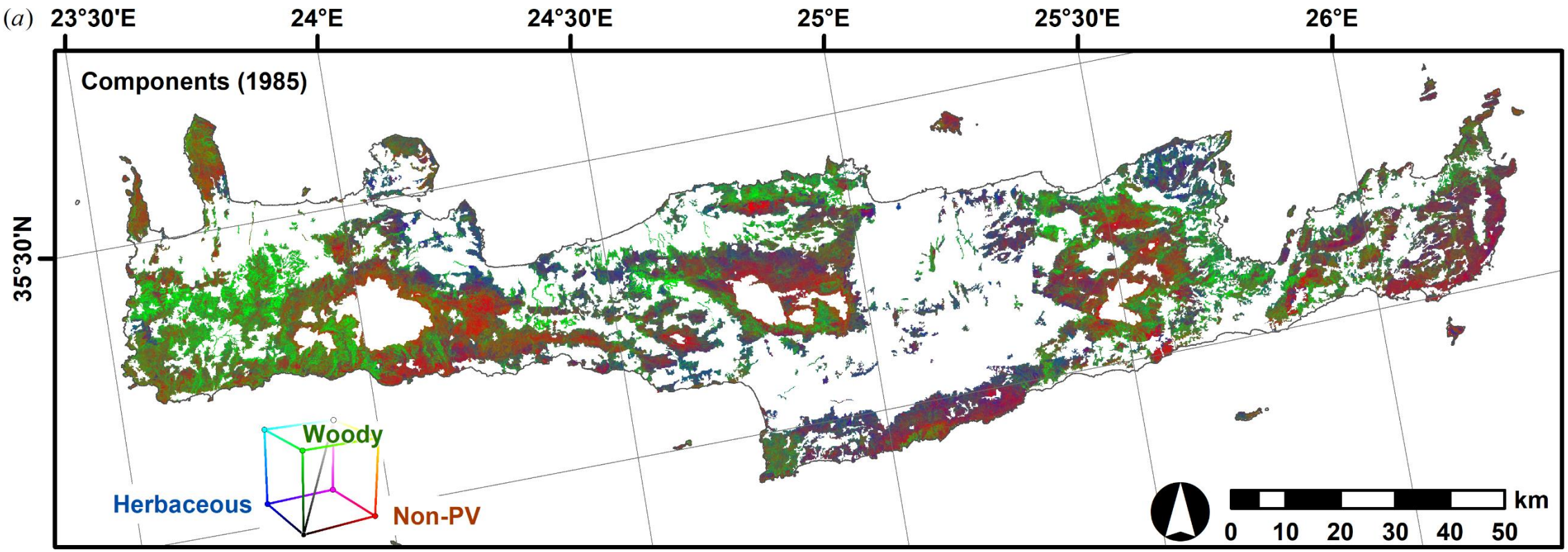
[1] Institute for Computer Science, Humboldt-Universität zu Berlin

[2] Geoinformatics - Spatial Data Science, Universität Trier

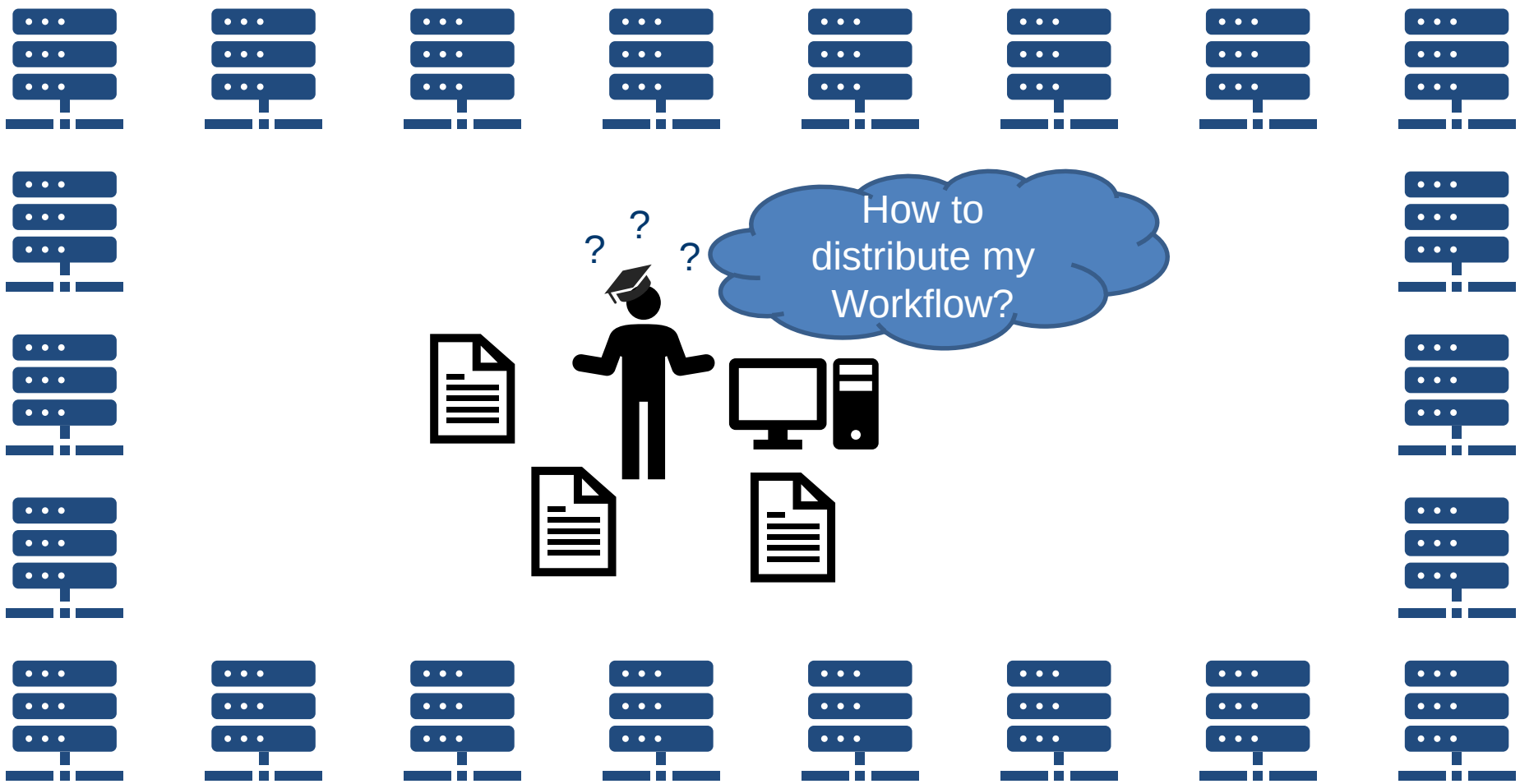
[3] Distributed and Operating Systems, Technische Universität Berlin

[4] Geography Department, Humboldt-Universität zu Berlin

Crete: rangeland degradation analysis



How to execute a distributed workflow

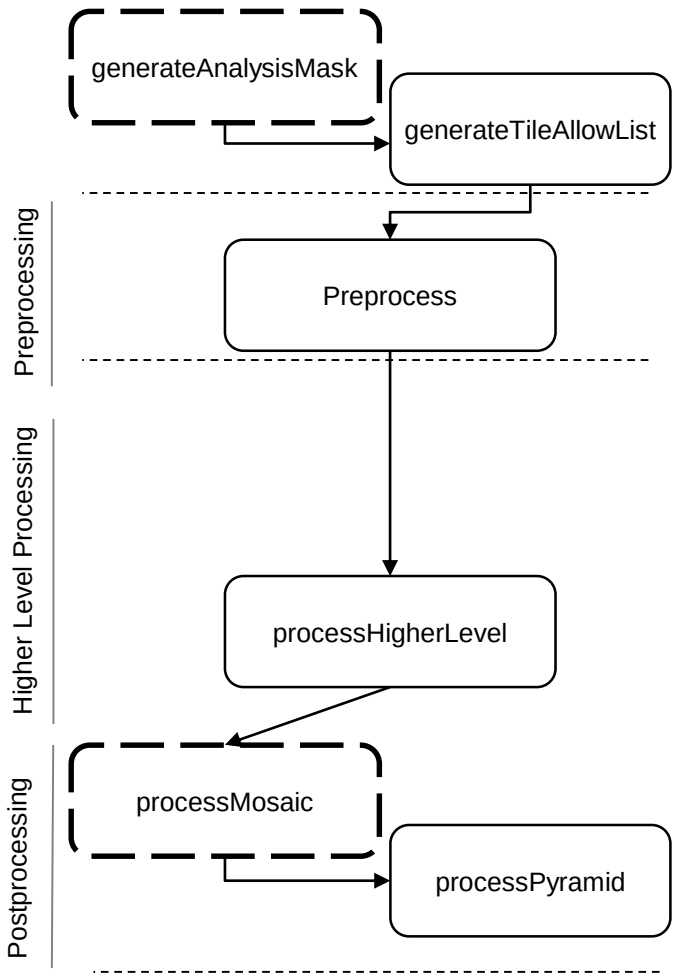


Workflowsystems

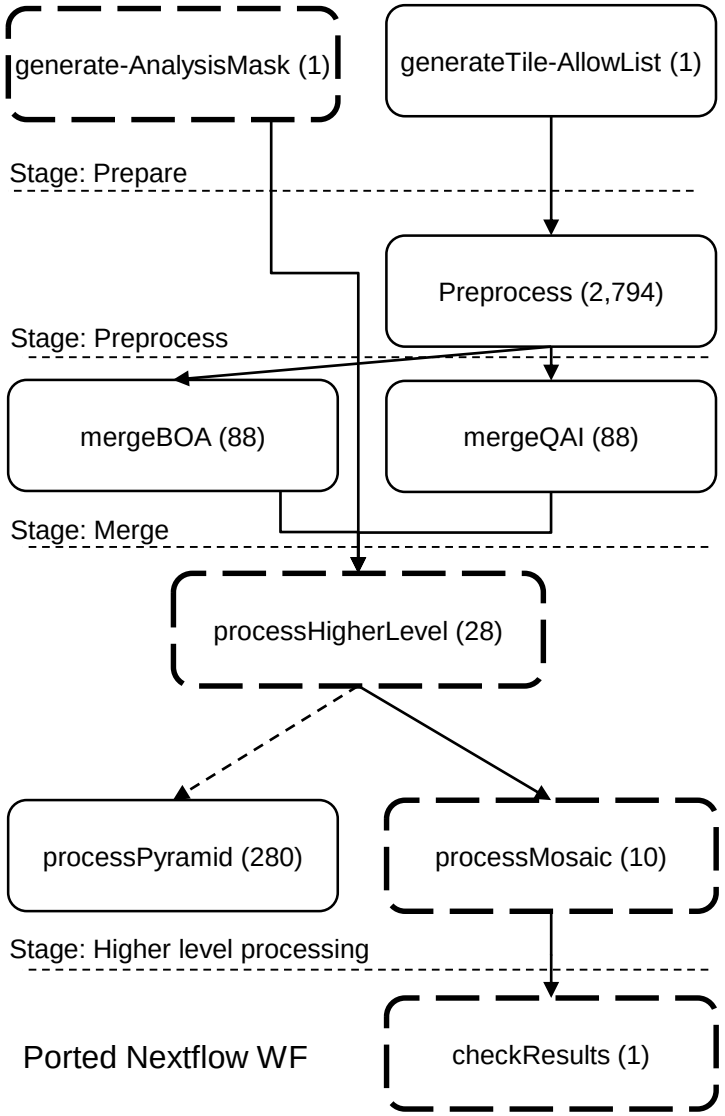
Workflowsystems

Workflowsystems

How to adopt a BASH-Script to Nextflow

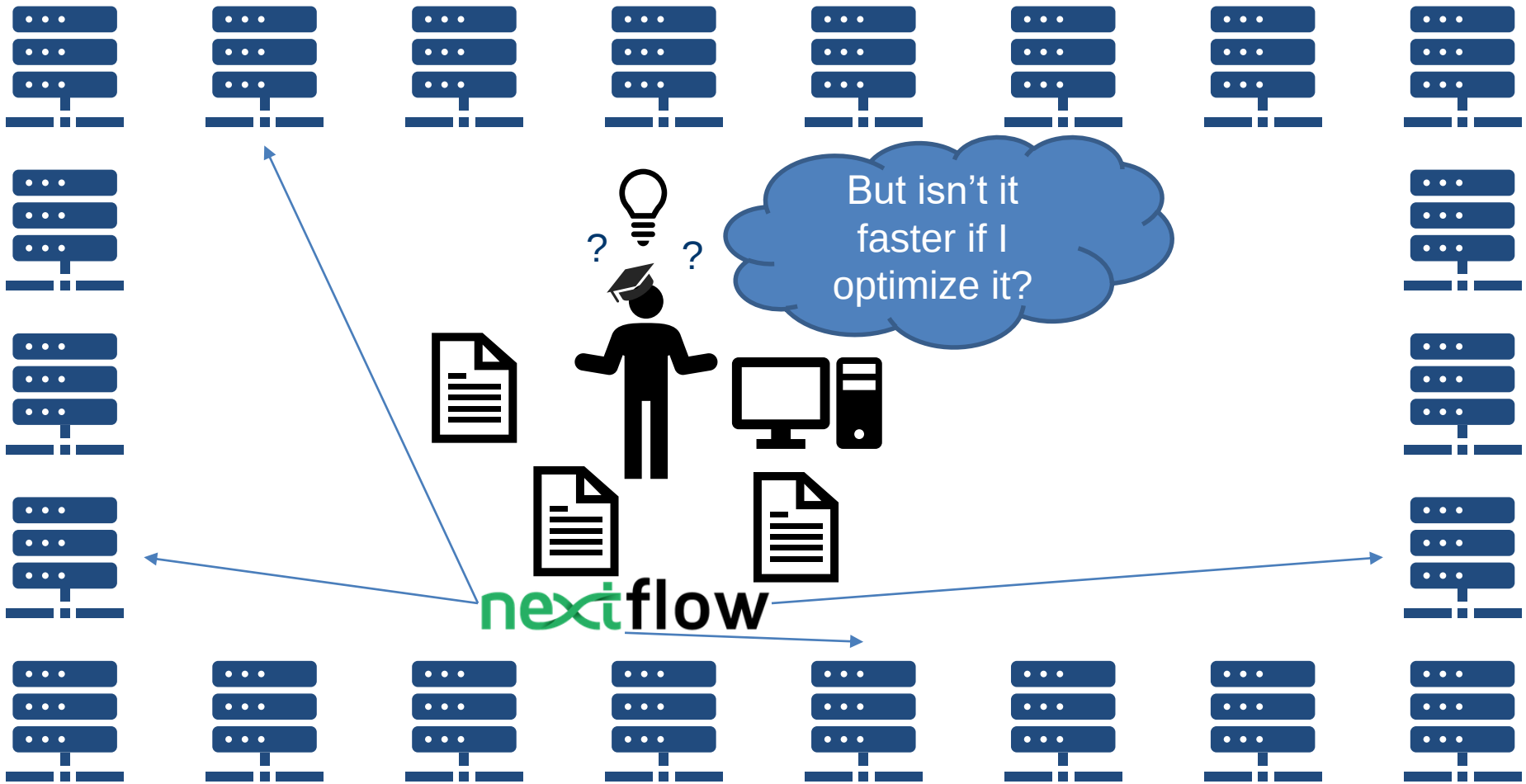


Original FORCE WF

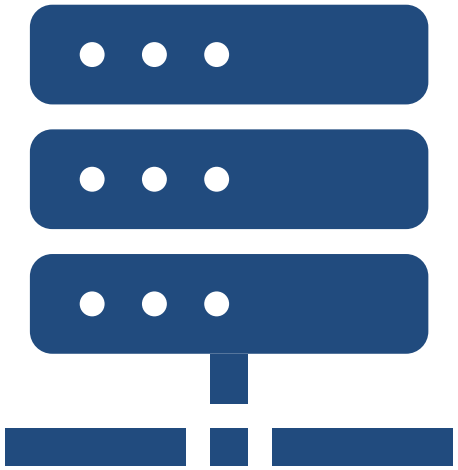


Ported Nextflow WF

How to run a distributed workflow

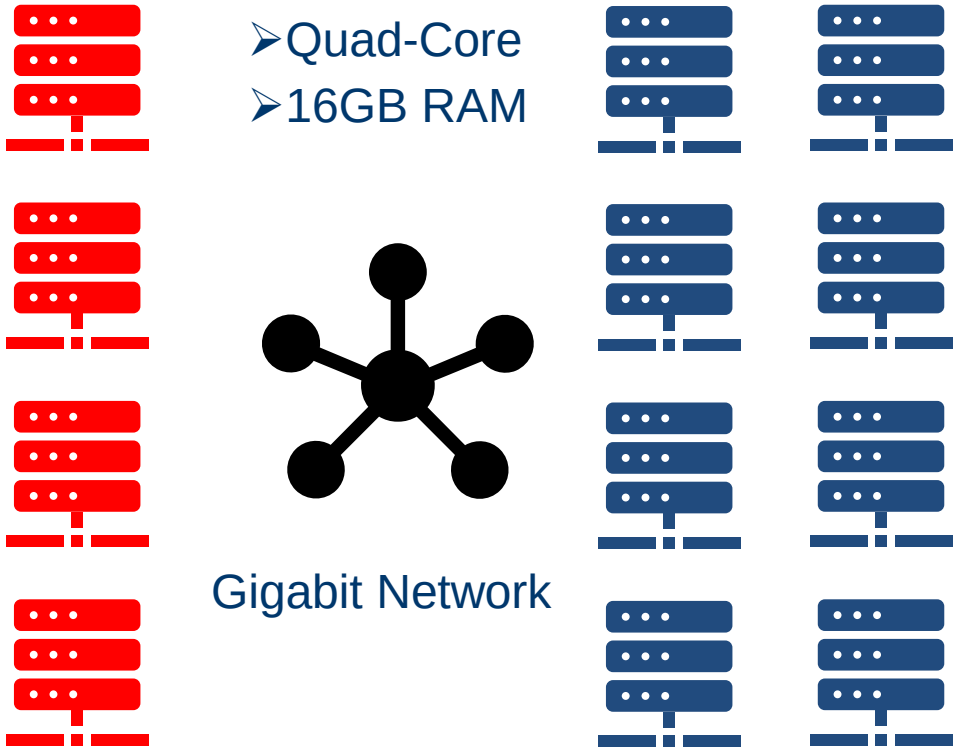


Test environment



- 56 Cores
- 750GB RAM

VS.



- Quad-Core
- 16GB RAM

Gigabit Network

nextflow

Workflow Execution Engine
- Orchestration

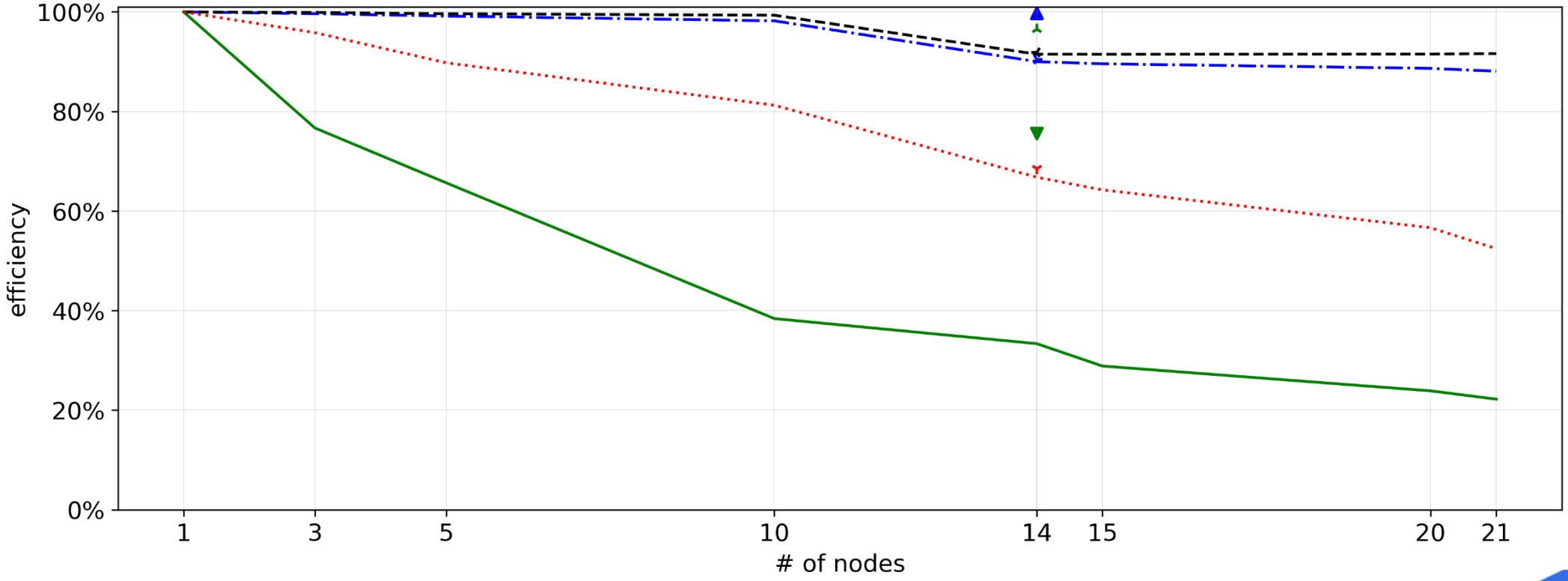
ceph

Distributed Storage
- Input data
- Intermediate data
- Output data

kubernetes

Ressource Manager
- Calculation
- Analysis
- Execution

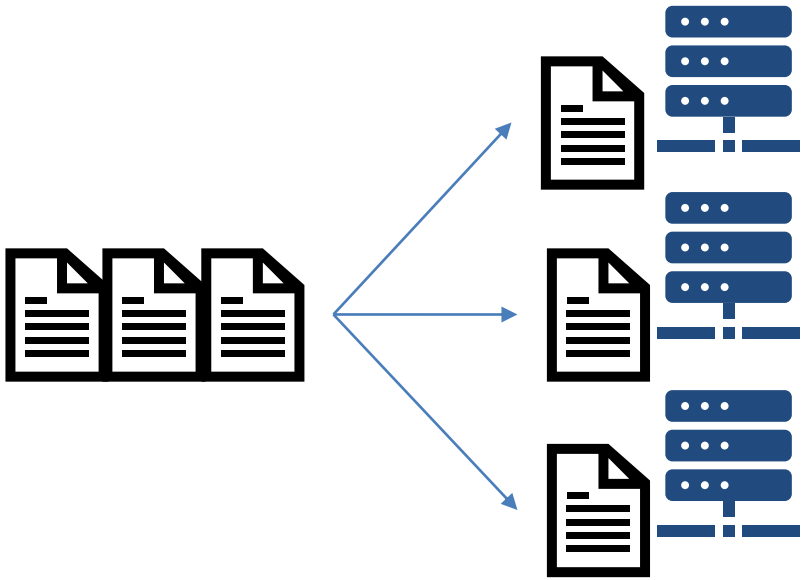
Scalability analysis

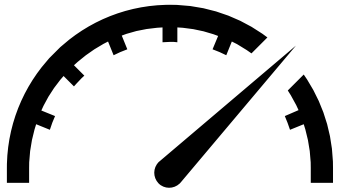


Conclusion

nextflow

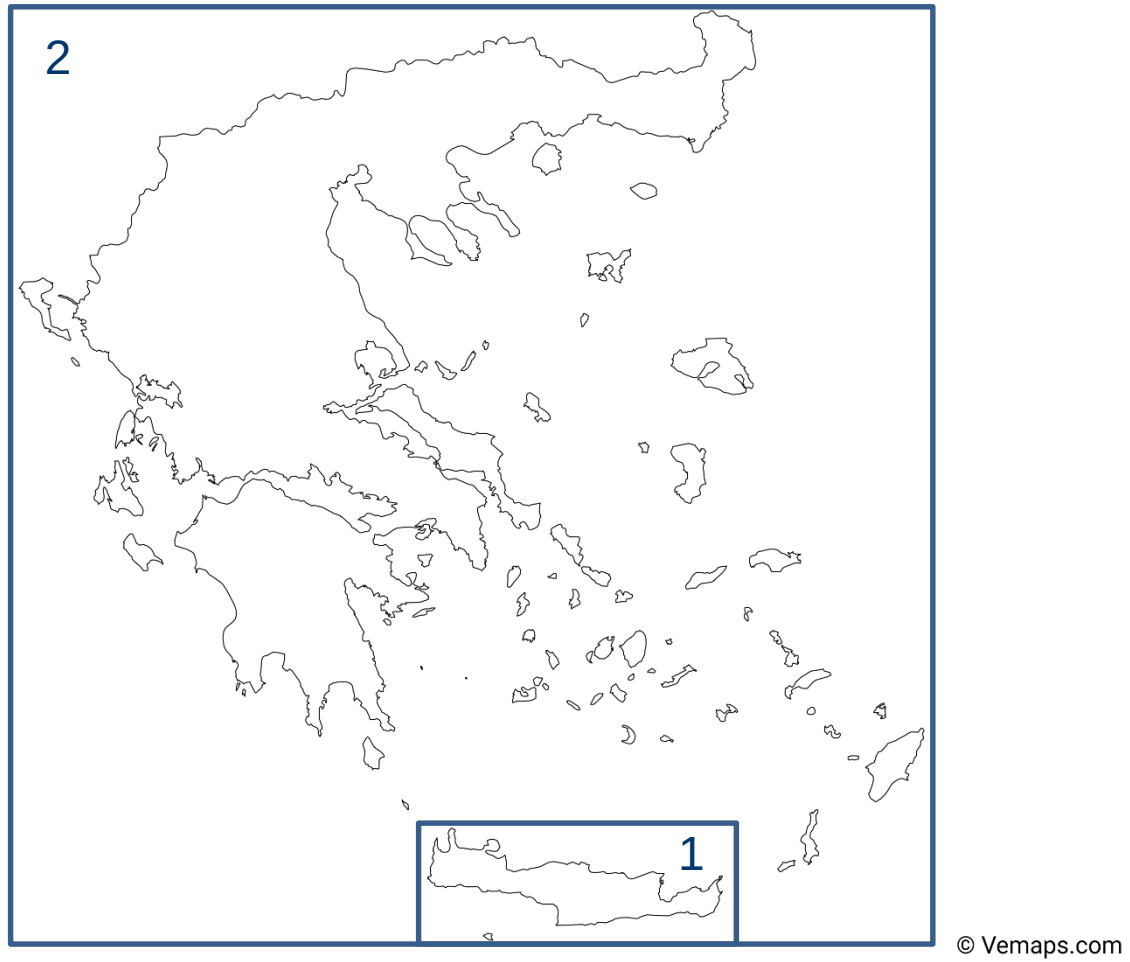
- Learn additional workflow language
- Workflow Engine supports scalability
- I/O heavy tasks hardly scale
- Compute heavy tasks scale quite well
- Parallelism is defined implicitly





Outlook

- Test scalability on larger cluster
- Run workflow for larger region
- Location aware scheduling



Contact:

fabian.lehmann@informatik.hu-berlin.de
david.frantz@uni-trier.de
soeren.becker@tu-berlin.de
leser@informatik.hu-berlin.de
patrick.hostert@geo.hu-berlin.de

Thank you for your attention!

FONDA – Foundations of Workflows for Large-Scale Scientific Data Analysis

DFG Collaborative Research Center 1404 at Humboldt-Universität zu Berlin

DFG Deutsche Forschungsgemeinschaft

FONDA