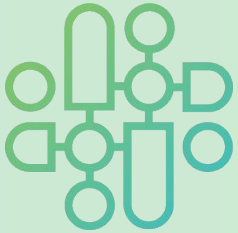


SC23
Denver, CO | i am hpc.

Towards a Peer-to-Peer Data Distribution Layer for Efficient and Collaborative Resource Optimization of Distributed Dataflow Applications

Dominik Scheinert[1], Soeren Becker[1], Jonathan Will[1], Luis Englaender[1], Lauritz Thamsen[2]

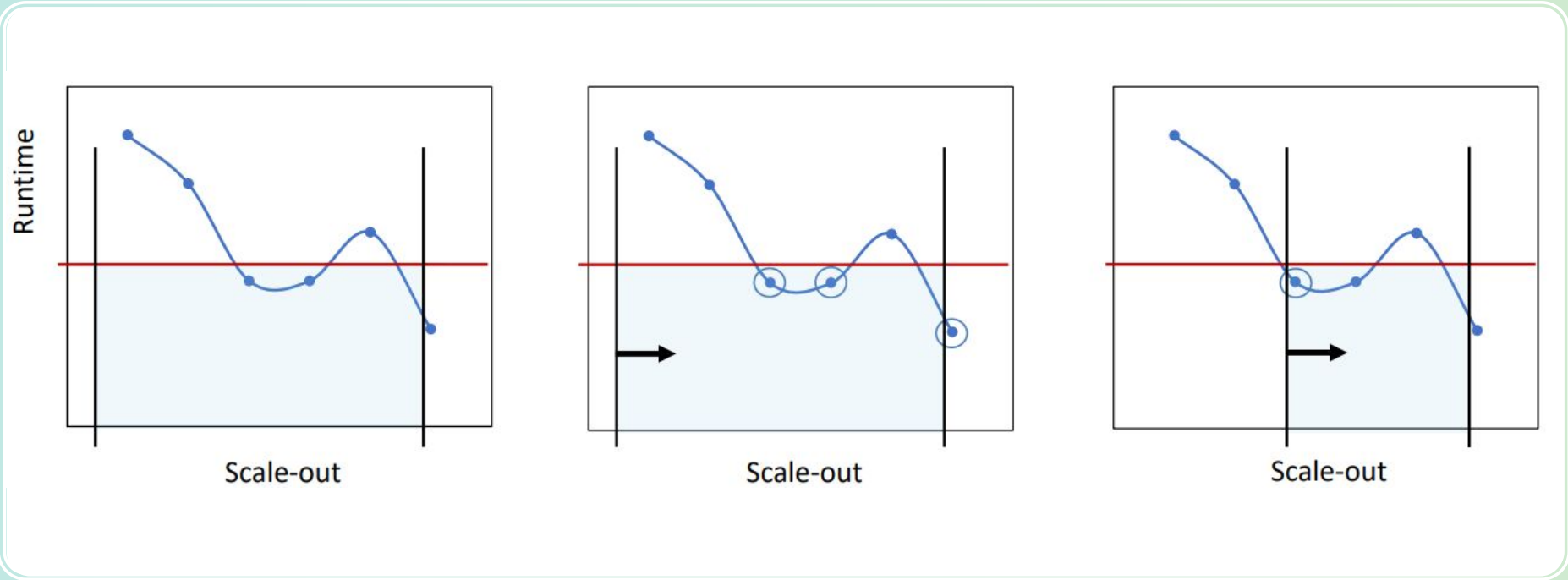
1) Technical University of Berlin, Germany, 2) University of Glasgow, United Kingdom

Motivation

- Large volumes of data are generated and require processing
- Configuration & Optimization of data processing systems is non-trivial
- Increasing computing needs of people from other domains or SMEs

→ Automated solutions for configuring & optimizing in terms of resources are desired

Problem



However, performance modeling for resource management is difficult if

1) restricted access to historical performance data
2) dedicated profiling phase not possible

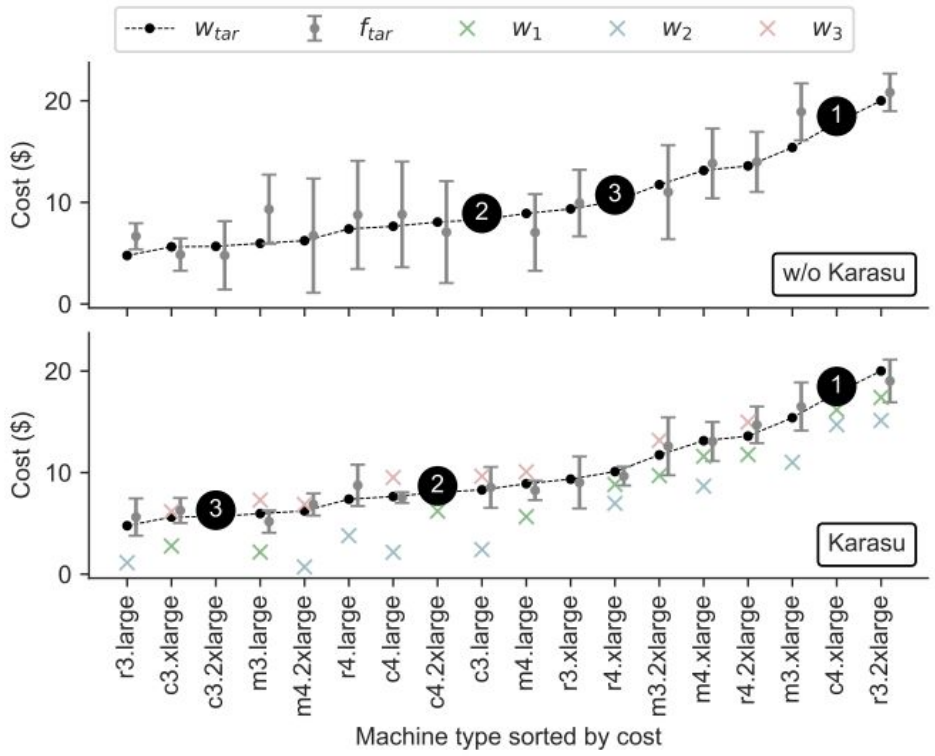
Gap Analysis

Related + own prior work proposes collaborative methods

- Formulates certain requirements
- Centered around the idea of data + model sharing
- Improves resource efficiency, reduces costs

But: Focus mostly on modeling, not on actual collaboration and data exchange

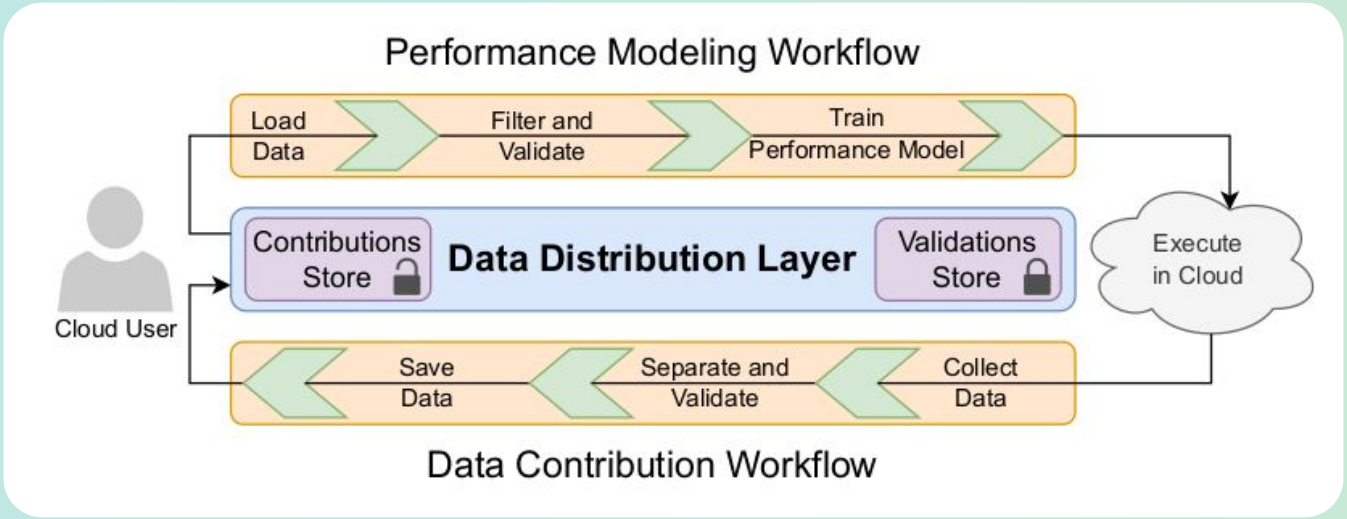
- Assumes central storage solution
- No real integration with performance modeling techniques

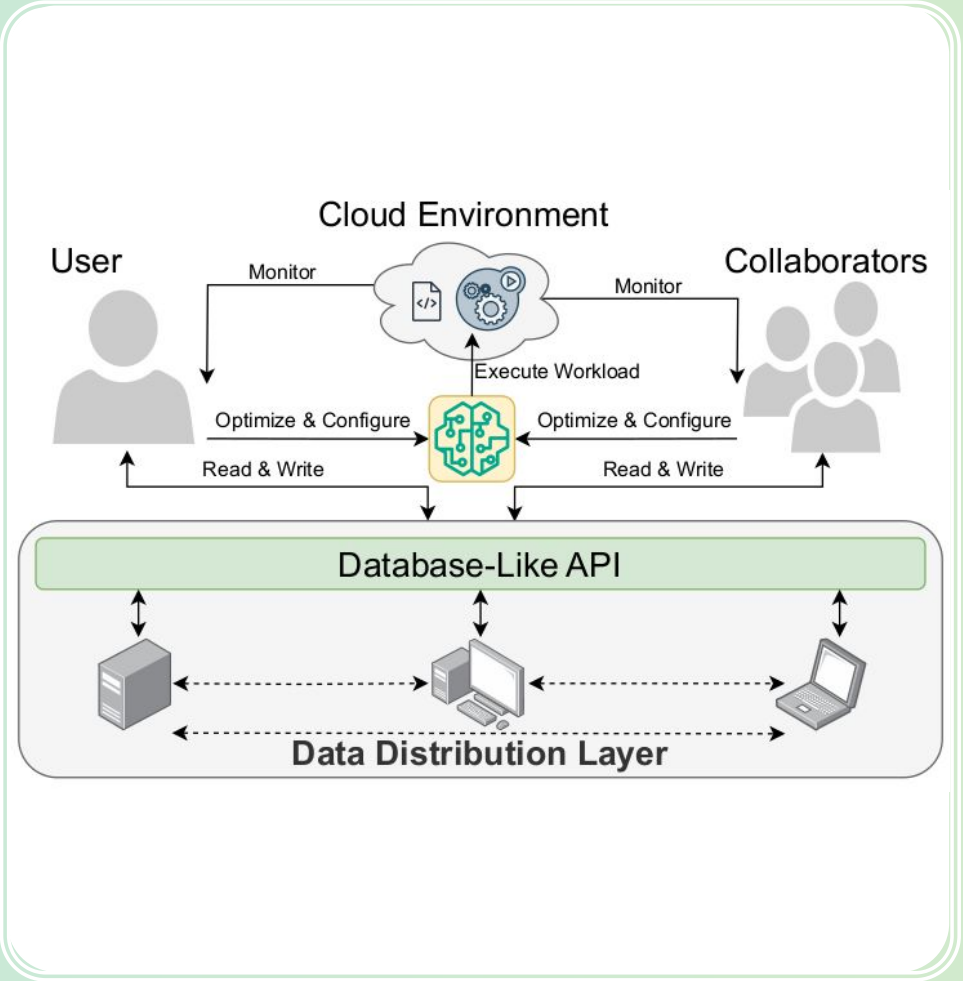


Approach

Decentralized Performance Data Sharing

- P2P-Based Data Distribution Layer
- Underlying processes are masqueraded
- Integrates with commonly seen workflows
- A Database-Like API facilitates usage

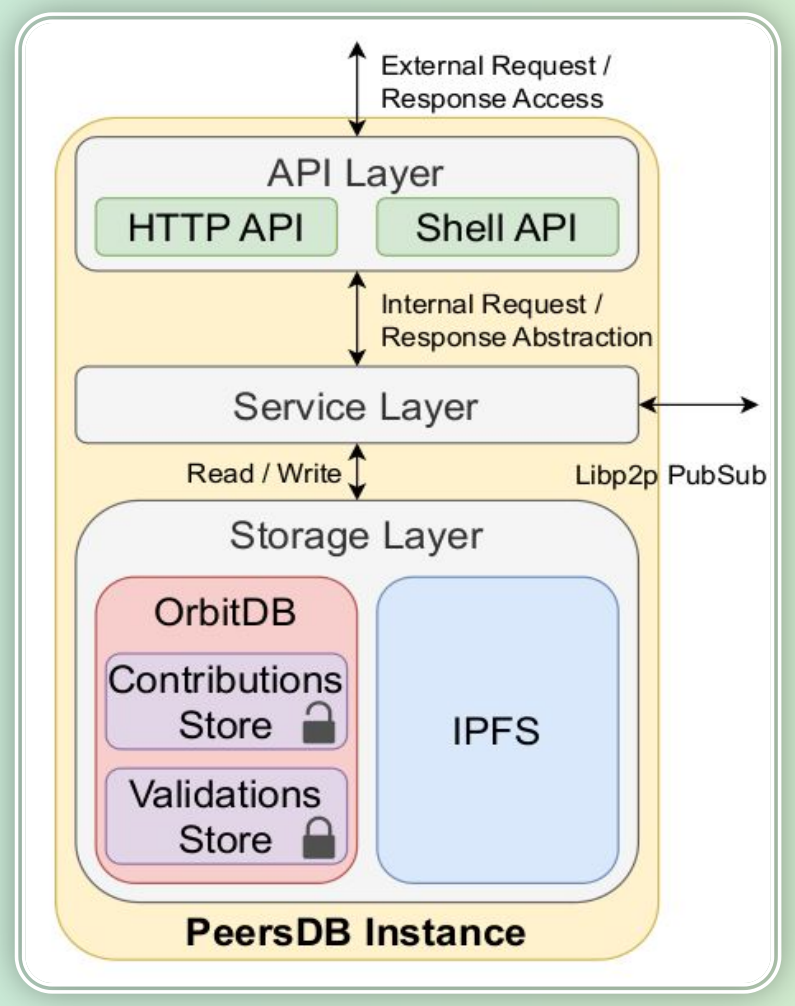




Preliminary Results

Initial Prototype Implementation

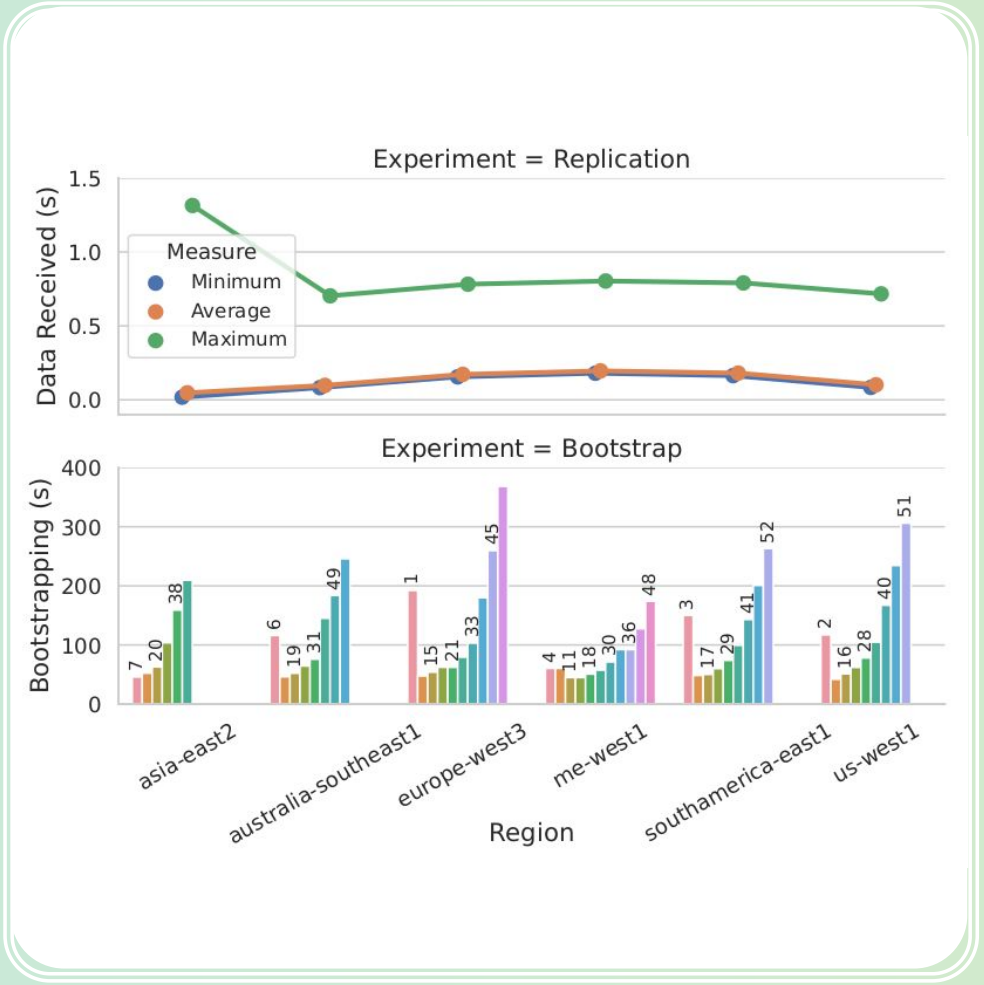
- Builds upon IPFS and integrates with OrbitDB project
- Differentiates between private and shareable data
- Service routines and APIs facilitate employment



Preliminary Results

First conducted experiments

- GKE cluster with 6 nodes, distributed worldwide
- Simple replication and bootstrapping experiments
- Up to 50 PeersDB instances
- Usage of >11K files with representative performance data



Conclusion and Future Work

- Decentralized solution for sharing workload performance data
- First promising results with regard to elemental operations

Future Work:

- Experiments with direct integration of performance modeling techniques
- Further excessive investigations via Testground simulation framework
- Comparison of different data validation strategies and their runtime
- Investigation into whether integration with blockchain makes sense