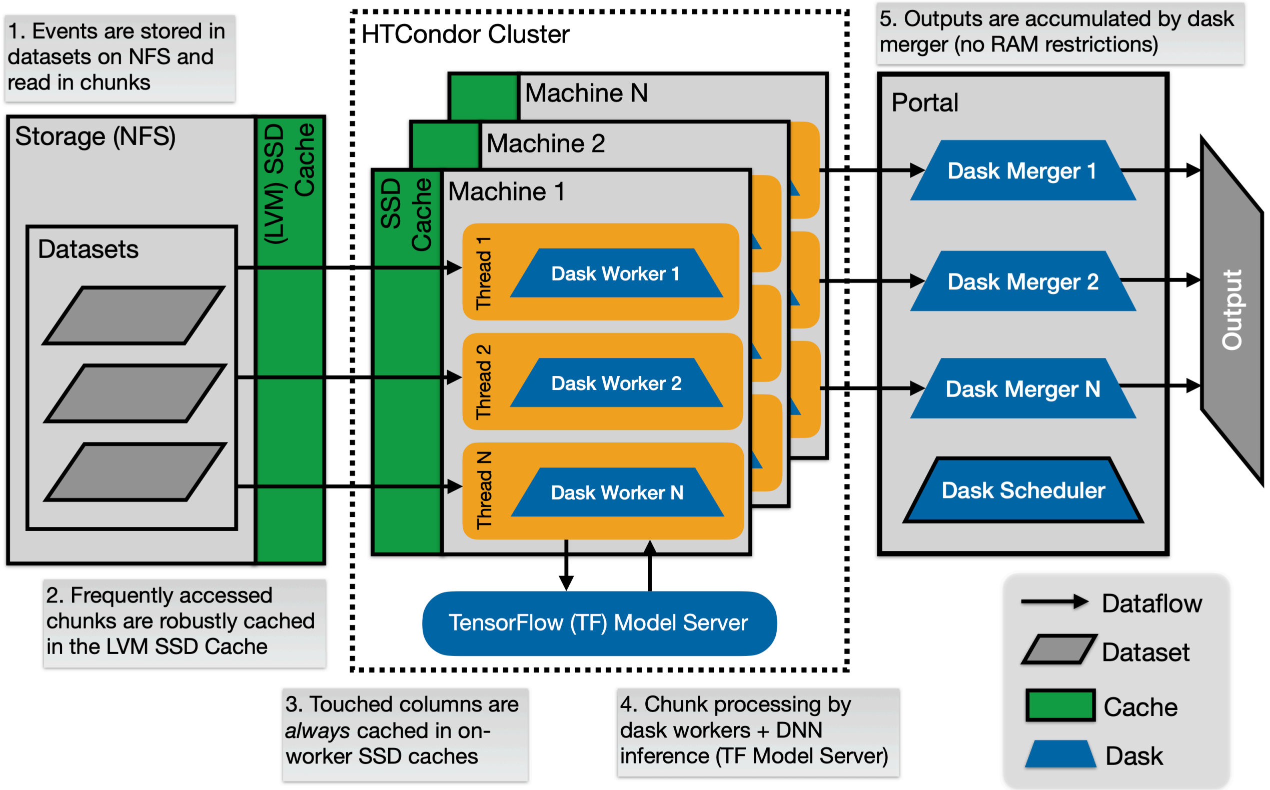


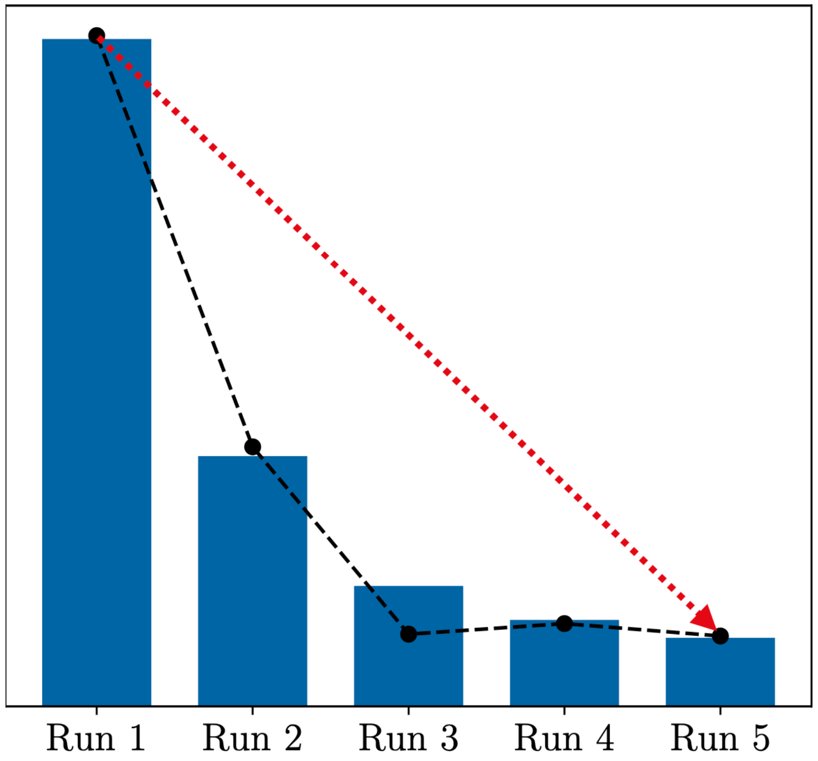
1. Events are stored in datasets on NFS and read in chunks

2. Frequently accessed chunks are robustly cached in the LVM SSD Cache

3. Touched columns are *always* cached in on-worker SSD caches

4. Chunk processing by dask workers + DNN inference (TF Model Server)

5. Outputs are accumulated by dask merger (no RAM restrictions)



Going fast on a small-size computing cluster

Martin Erdmann, Peter Fackeldey,
Benjamin Fischer, Dennis Noll

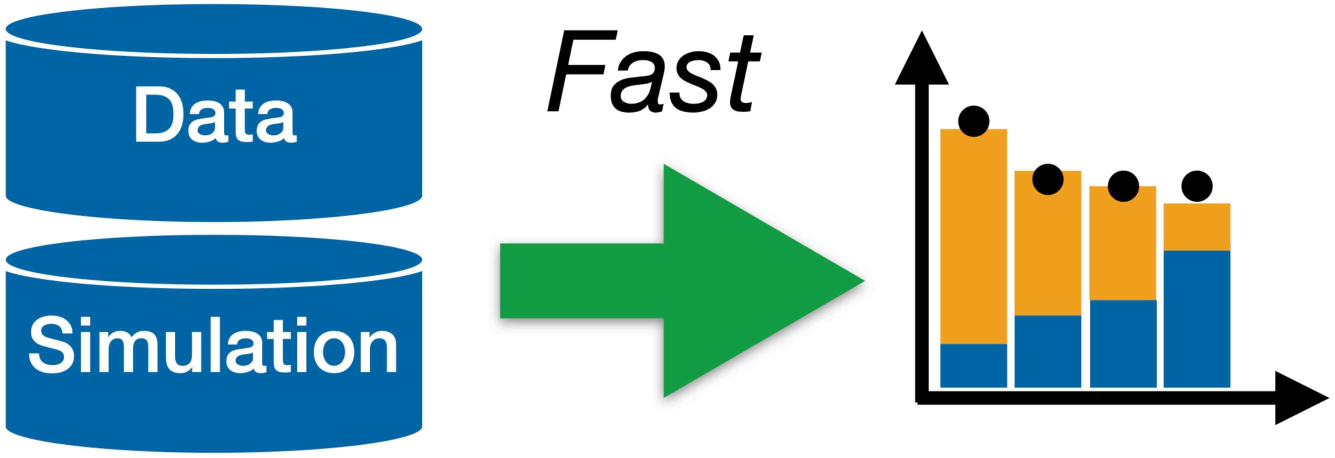
FIDIUM - Kickoff

16.12.21

Starting Point

Typical LHC Analysis:

- Input: Datasets $\mathcal{O}(10)$ TB
- Output: Histograms $\mathcal{O}(1)$ GB
- Complex:
  - Many physics objects
  - Deep learning discriminators



Team:

- 1 Professor
- 5 PhD Students
- 1 Master Student

Software:

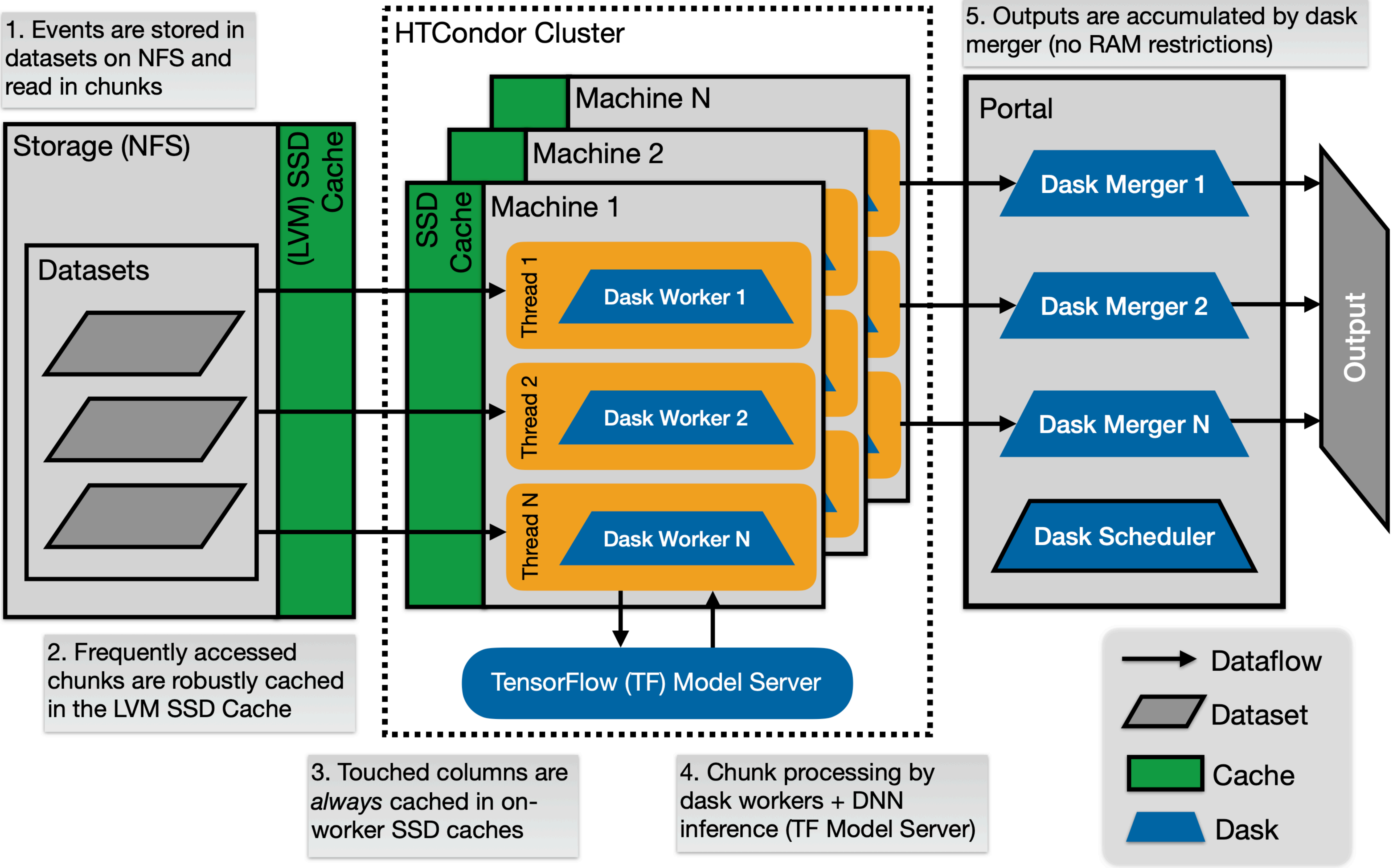
- NumPy + Scikit-HEP
- Dask
- HTCondor

Hardware:

- 245 Threads
- 22TB SSD Caches
- 72TB Disk Storage

Computing Strategy:
MapReduce on *columnar* data
accelerated by caching

Dataflow & Cluster sketch



1. Events are stored in datasets on NFS and read in chunks

2. Frequently accessed chunks are robustly cached in the LVM SSD Cache

3. Touched columns are *always* cached in on-worker SSD caches

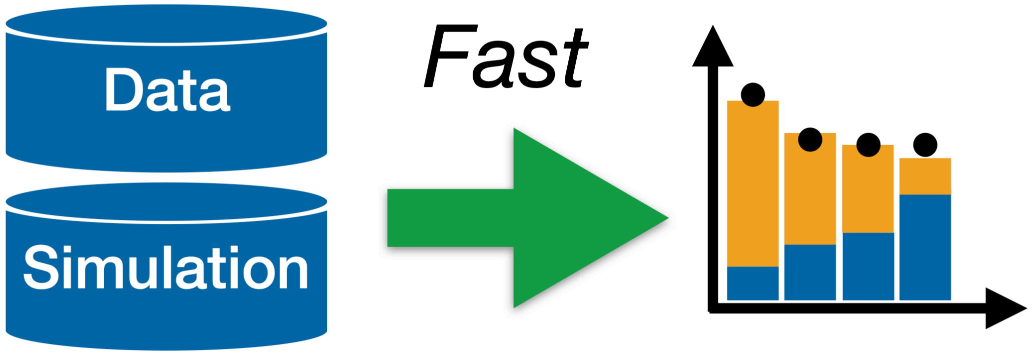
4. Chunk processing by dask workers + DNN inference (TF Model Server)

5. Outputs are accumulated by dask merger (no RAM restrictions)

Results and Conclusion

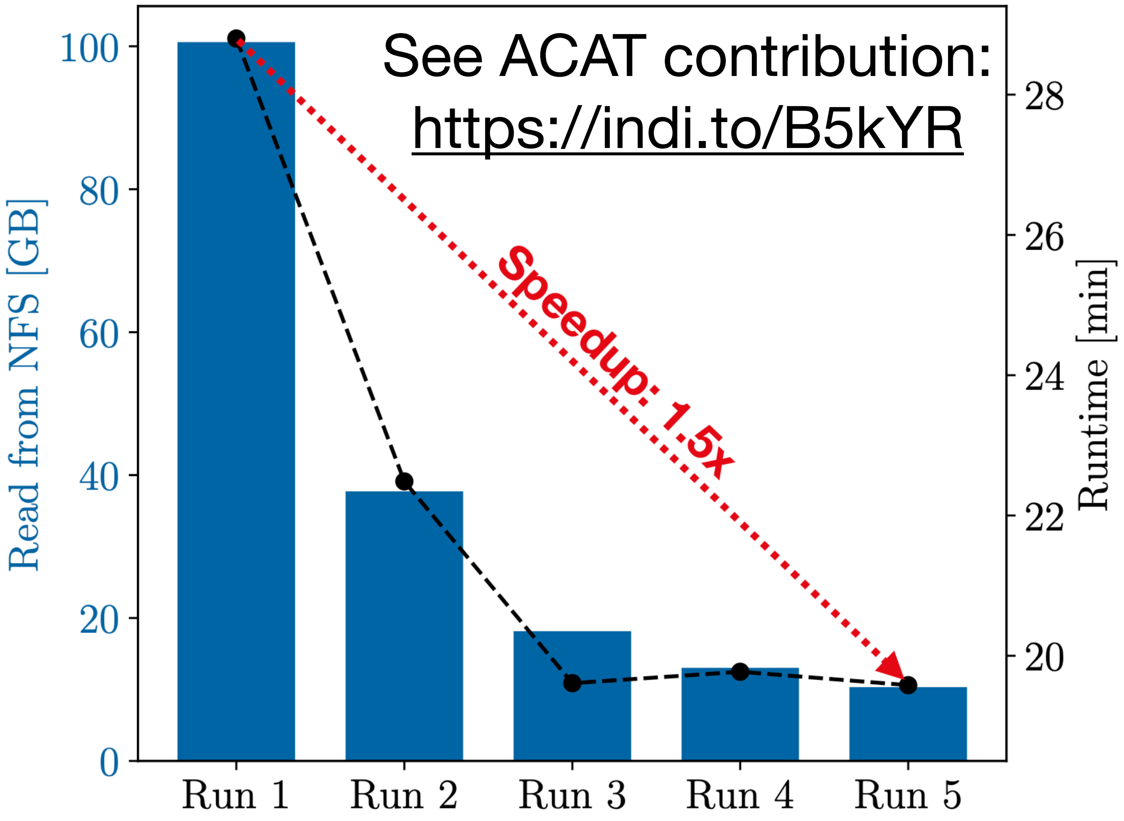
Benchmark Results:

- $4.18 \cdot 10^9$ Events (386 GB)
- Speedup of **x1.5** through SSD caching
- Bottleneck here: decompression
- Representative compared to our HH → bbWW analysis



Conclusion:

- Optimise chunk processing by vectorisation and GPU-offloading
- Data caching/locality:
  - Nearline storage (NFS): reliable and low-latency
  - Deterministic use of distributed SSD caches (across workers)



See ACAT contribution: https://indi.to/B5kYR

We are happy to continue this work in FIDIUM and established a first collaboration with Thomas Kuhr and David Koch!

Backup

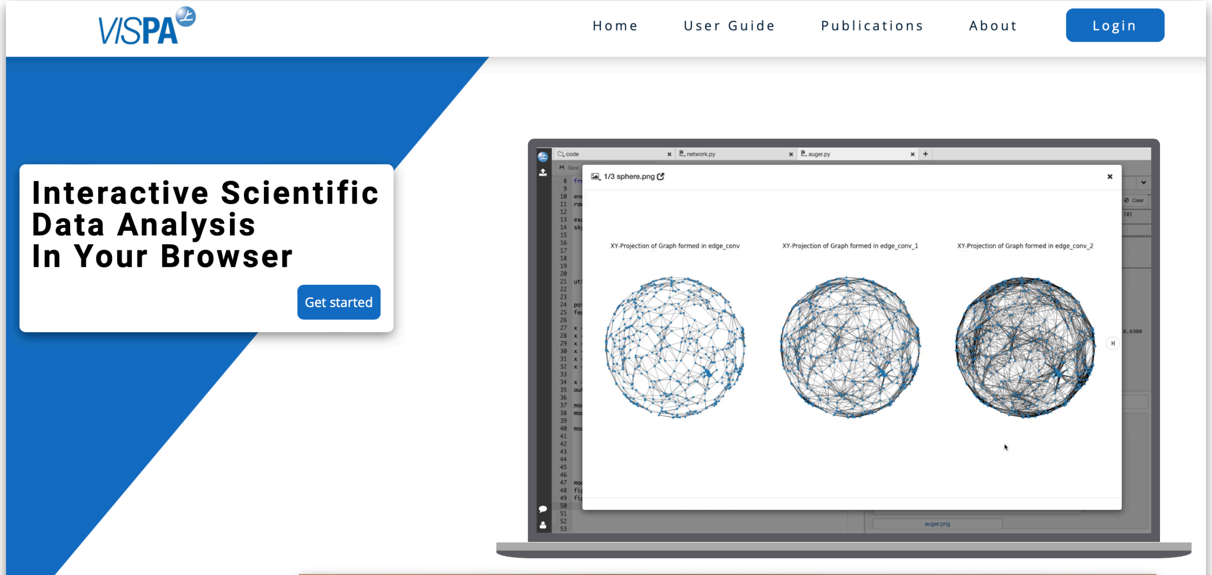
RWTH Aachen Group - Prof. Erdmann

- Team:
  - 1 Professor
  - 5 PhD Students
  - 1 Master Student

- Hardware - VISPA Cluster
  - 13 Worker machines, in total:
    - 245 threads, 2GB RAM each
    - 22TB SSD (FS-Cache)
  - Storage (NFS):
    - 6x12TB HDD (striped)
    - 1TB LVM SSD Cache

- Software:
  - NumPy and python-HEP ecosystem
  - Dask (dask-jobqueue)
  - Packaged by conda

https://vispa.physik.rwth-aachen.de

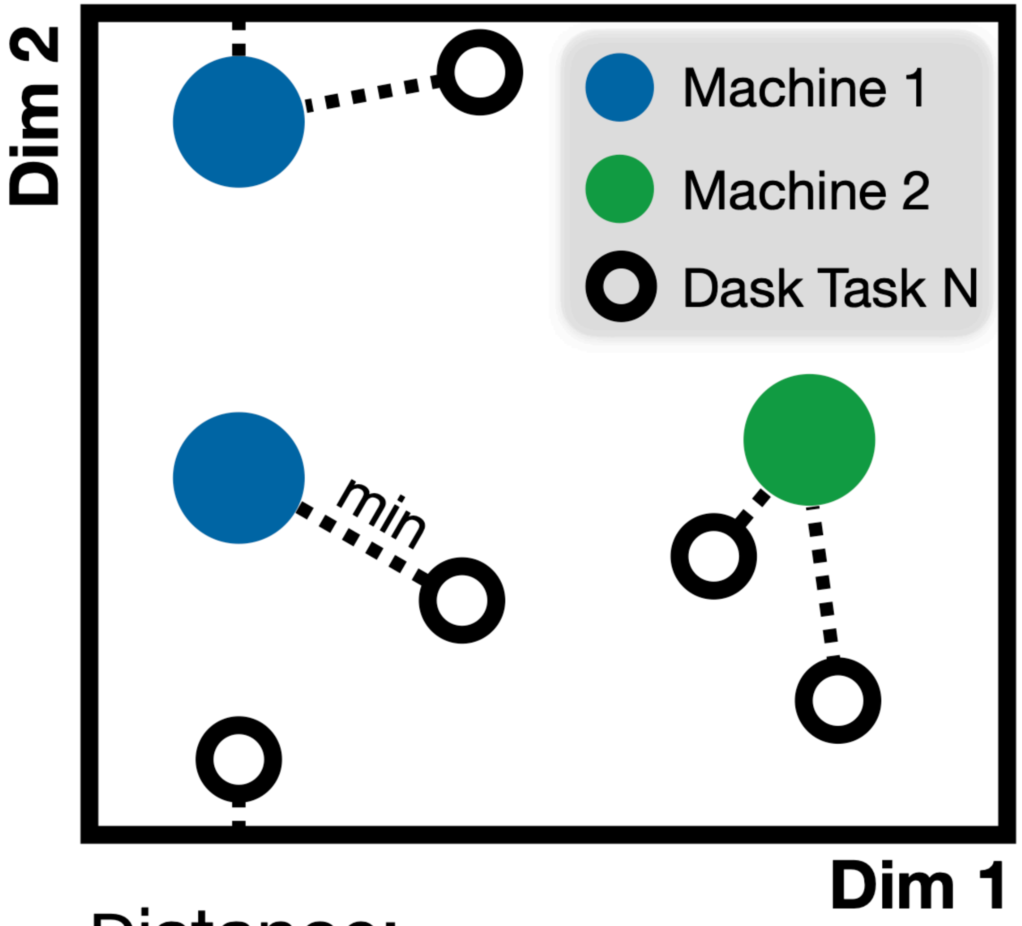


Interactive Scientific Data Analysis In Your Browser

Access Fr

Cache Affinity

On-Worker SSD Cache



FS-Cache Properties:

- Transparent
- Shared by multiple users

Affine assignment:

- Uses hash distance
- Deterministic
- Smoothly degrading under changes of workers & tasks

Distance: ⋯⋯

$$\sum_{i}^{N} min\left(|\bigcirc - \bullet|, 1 - |\bigcirc - \bullet|\right)$$

with N: number of dimensions

Comparable to CRUSH of Ceph