Simularity Join over Array Data
Weijie Zhao¹, Florin Rusu¹,², Bin Dong², Kesheng Wu²
University of California Merced¹, Lawrence Berkeley National Laboratory²
wzhao23, frusu@ucmerced.edu; dbin, kwu@lbl.gov

Shared-Nothing Architecture
Query Optimization

|-----|-----|-----|-----|-----|-----|-----|-----|

Transfer graph
Data transfer
Load-balancing
Network congestion
Data access plan
Execution time

Network transfer time:
\[
\min \left\{ \max_{ijk} \{ y_{ijk} \cdot T_{\text{max}} \} \right\}
\]

Disk access time:
\[
\min \left\{ \max_{k} \left\{ \sum_{ij} y_{ijk} (1 - z_{f(t-1) \cdot T_{\text{disk}}}) \right\} \right\}
\]

Experiments

# nodes

PTF catalog: PTF [time=1,153064; ra=1,100000; dec=1,50000] 1 billion objects, 12 GB
LinkedGeoData: GEO [long=1,100000; lat=1,500000] 30 million objects, 1 GB

PTF catalog evaluation on PTF 3D query:
optimization time, execution time, and transferred data

Query optimization evaluation on PTF 3D query:
execution time, receiving time, and thread time distribution

Algorithm execution time:

Algorithm data transferred time:

Scalability: