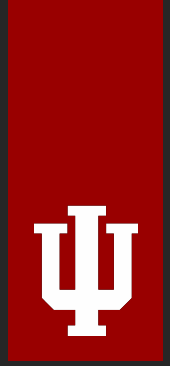


Intelligent Systems Engineering, Luddy School of Informatics, Computing, and Engineering

Cylon & CylonFlow: High Performance Data Engineering in Supercomputers

Niranda Perera, *Parsl & funcX Fest 2022*

Data Engineering Status-Quo

- Data Science domain has expanded monumentally→ BigData, AI, ML
- *“Significant dev time is spent on data exploration, preprocessing, and prototyping”*
- SQL no-more! Functional FTW! → DataFrames
- Python world domination… taking Pandas along!
- TBs of data → Beyond capabilities of a single machine

BUT…

- Pandas/R performance limitations
- Distributed dataframes are still WIP

Cylon

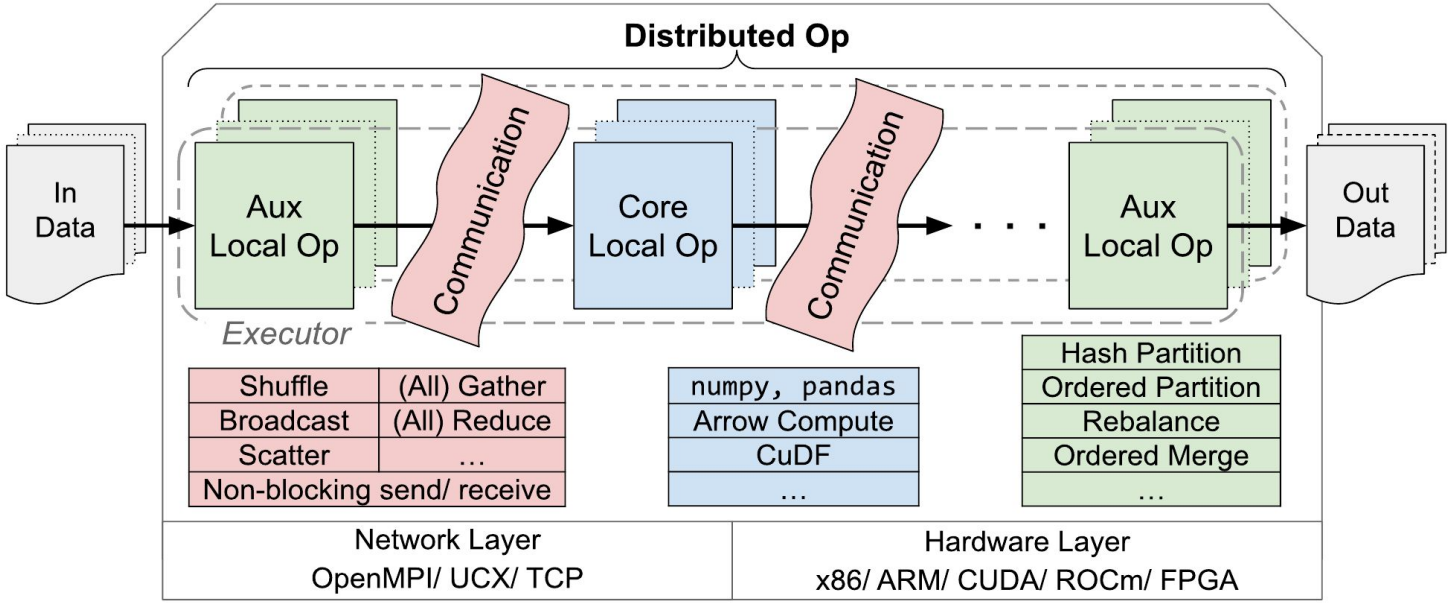
"How to develop a high performance scalable dataframe runtime?"

- Distributed memory parallel dataframe runtime
- BSP → leaf out of HPC playbook
- Apache Arrow columnar format
- Developed in C++, Cython bindings for Python
- Communication → OpenMPI, UCX, Gloo
- Distributed memory ops dev in-house
- Pandas op coverage ~25%
- Extended to GPUs →GCylon

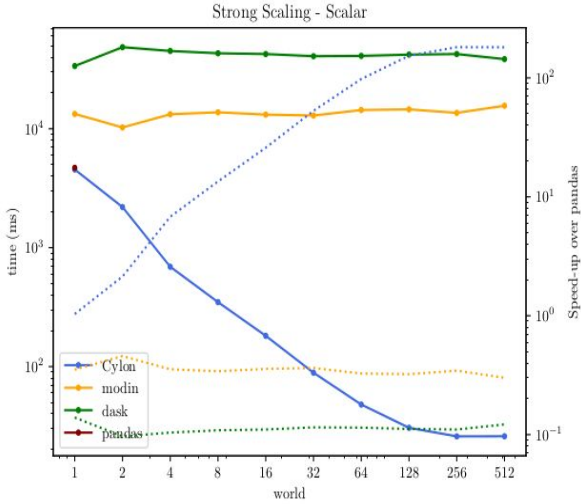
Bragging rights:

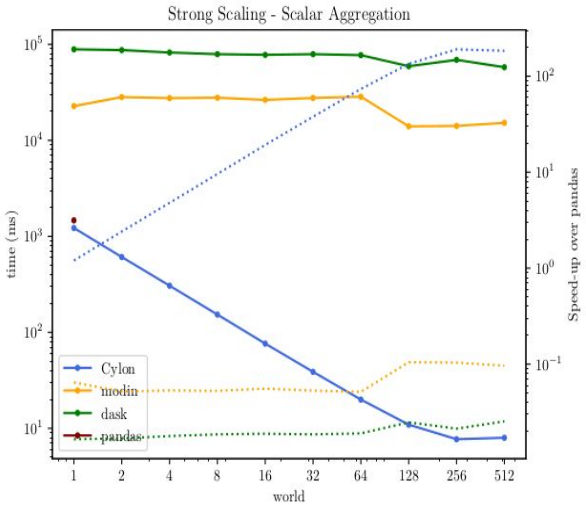
First & only dataframe in MPI!

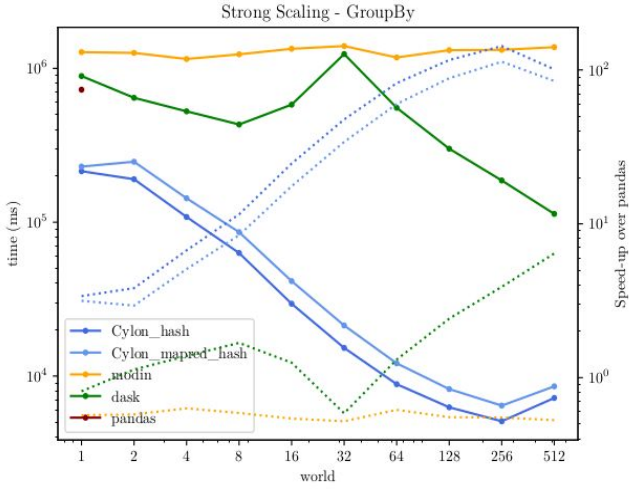
Cylon Building Blocks

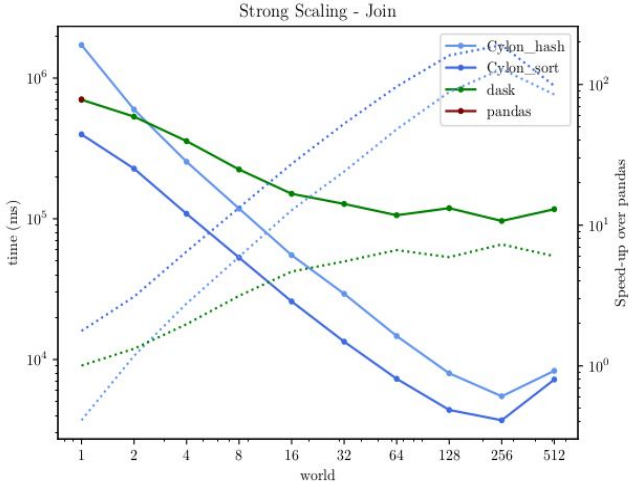


Strong scaling (1B)



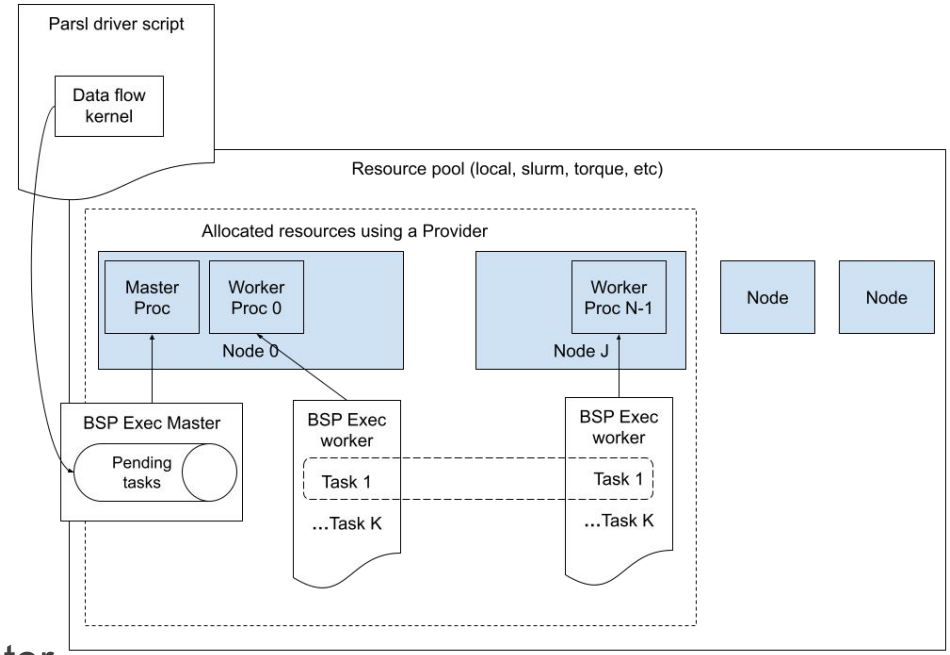






CylonFlow?

- Cylon is BSP→ No Jupyter support! :-(
- Cylon in supercomputers?
- Parsl for the rescue!
- Proposed: A BSP executor for Parsl
- Executes MPI tasks on a subcommunicator
- Coincided with RADICAL-Pilot - Parsl integration
- CylonFlow now supports → Parsl, RP, Dask, and Ray



Future Work

- Larger scale experiments in leadership-class supercomputers (colab with RP)
- TPCx-BB benchmark using Cylon

- Cylon Window operators
- Fault tolerance
- Work imbalance due to skewed datasets

Thank you!

Q&A

https://cylondata.org/
https://github.com/cylondata/cylon
cylondata@googlegroups.com
https://cylondata.slack.com/

INDIANA UNIVERSITY BLOOMINGTON