EVALUATE THE DASK DISTRIBUTED COMPUTING FRAMEWORK IN RESPECT TO VARIOUS SCIENTIFIC COMPUTING TASKS

EERO VAINIKKO

Course Coordinator

ARTJOM LIND

Topic Supervisor

JEYHUN ABBASOV

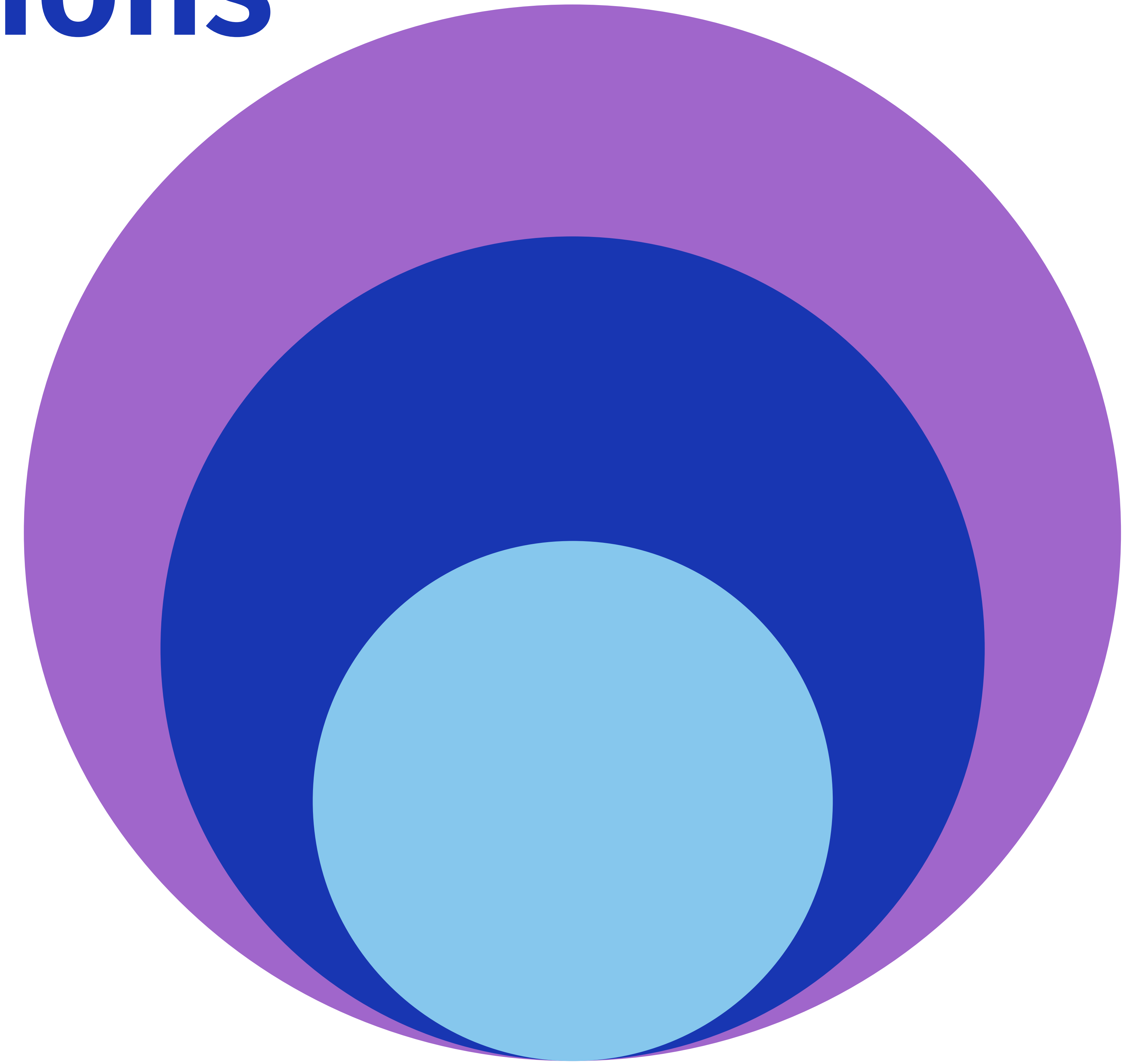
Student

The Research Questions

- **How DASK's parallel and out-of-core computation extends the effective scale of modern hardware to larger datasets?**

- **How these ideas can be more broadly applied to other parallel collections?**

- **How the DASK performs on various scientific computing tasks?**
  - such as Average Global Ocean Temperature 36 year's worth



About Dask

A flexible library for parallel computing in Python.

Is composed two parts:

1

Dynamic task scheduling

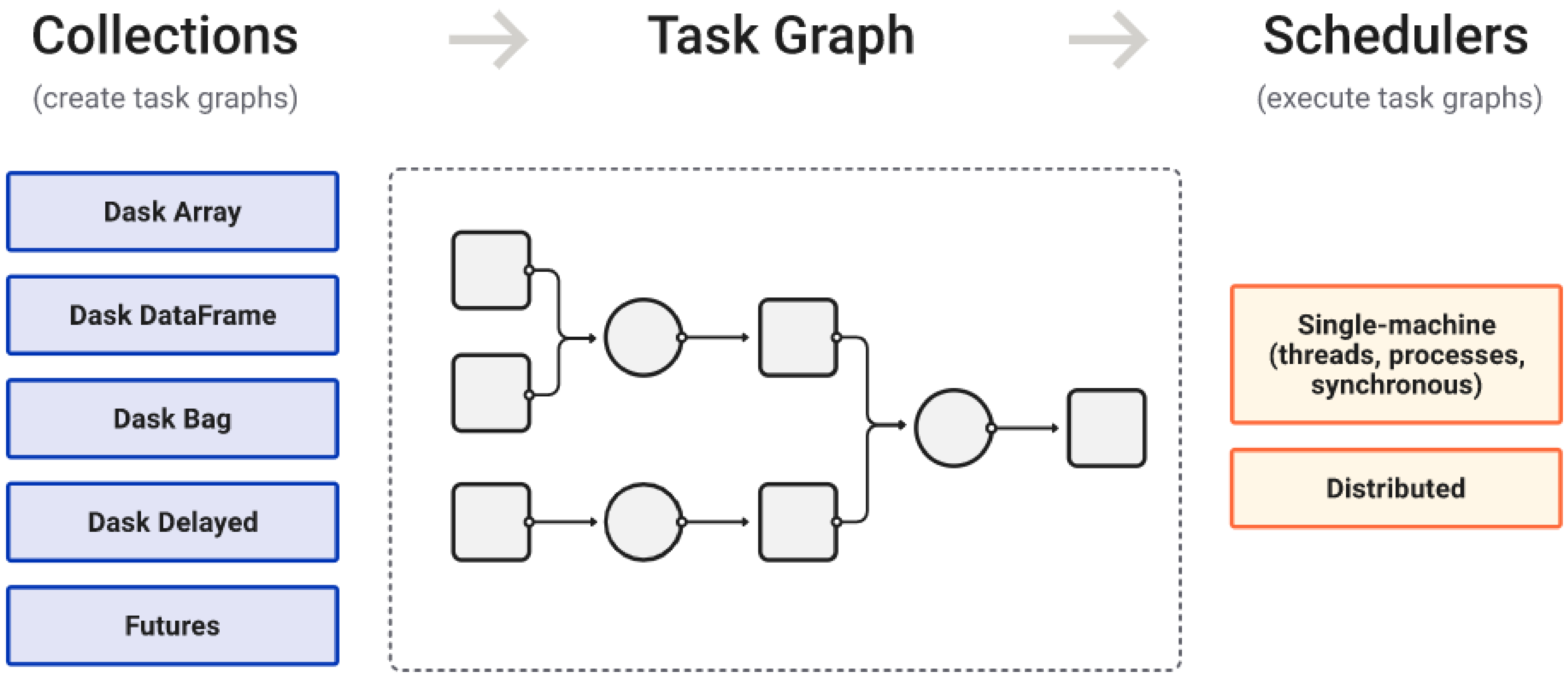
- optimized for computation
- *similar* to Apache Airflow, Luigi Workflow, *but* optimized for interactive computational workloads

2

“Big Data” collections

- contains parallel arrays, dataframes, and lists that extend common interfaces like NumPy, Pandas, or Python iterators to larger-than-memory or distributed environments

Dask Architecture



Collections

(create task graphs)

Task Graph

Schedulers

(execute task graphs)

Dask Array

Dask DataFrame

Dask Bag

Dask Delayed

Futures

Single-machine (threads, processes, synchronous)

Distributed

High level collections are used to generate task graphs which can be executed by schedulers on a single machine or a cluster.

Source: https://docs.dask.org/en/stable/

References

[1] Dask: Parallel Computation with Blocked algorithms and Task Scheduling

https://www.researchgate.net/publication/328778461_Dask_Parallel_Computation_with_Blocked_algorithms_and_Task_Scheduling

[2] Efficient MPI-based Communication for GPU-Accelerated Dask Applications

https://ieeexplore.ieee.org/abstract/document/9499534

[3] Performance Evaluation of Python Based Data Analytics Frameworks in Summit: Early Experiences

https://link.springer.com/chapter/10.1007/978-3-030-63393-6_24

[4] A Performance Comparison of Dask and Apache Spark for Data-Intensive Neuroimaging Pipelines

https://ieeexplore.ieee.org/abstract/document/8943502