Middleware for Data Intensive Analytics on HPC

Ioannis Paraskevakos,
RADICAL, Electrical and Computer Eng. Department
Rutgers, the State University of New Jersey,
Piscataway, USA
Email: i.paraskev@rutgers.edu

Motivation

Many scientific applications, through simulations, are the net producers of immense amounts of data, requiring an analysis phase that its execution time becomes dominated by the data produced. Thus, an efficient and scalable solution for analyzing data along with simulations becomes necessary. MIDAS (MIddleware for Data-intensive Analytics and Science) offers the necessary abstractions and middleware building blocks to support scalable data intensive analytics on HPC resources. Thus, it allows data-intensive applications to operate side by side with traditional HPC applications on the same resources.

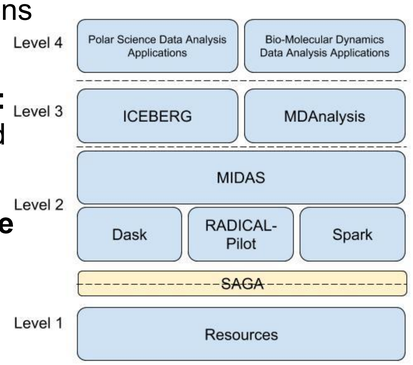
Current & Future Challenges

1. Provide a set of extensible and pluggable abstractions that operate on distinct layers from the application layer down to the resource interface
2. Provide abstractions that capture common scientific analysis patterns from different scientific domains
3. Provide APIs that are not software stack specific.

Approach

We propose the Building Blocks approach. Each block is characterized by four design principles: 1) Self-Sufficiency, 2) Interoperability, 3) Composability, and 4) Extensibility. We identify four functional levels:

1. **Level 4: Application Description:** Requirements and semantics of applications
2. **Level 3: Workload Management System:** Applications expressed as workloads.
3. **Level 2: Task Runtime System:** Execution of tasks of a workload
4. **Level 1: Resource**

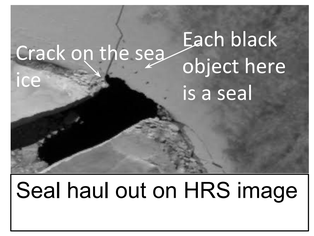


Science Drivers

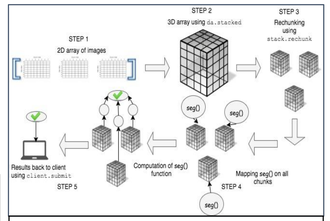
Polar Sciences

- Integrating high-resolution satellite imagery, efficient algorithms for seal detection, tasking strategies for image processing represents a perfect example of the need for greater collaboration among polar scientists and the HPC community.
- Distributed Computing Frameworks for satellite imagery:
  - Antarctic pack-ice seals are major consumers of Antarctic krill
  - Krill are directly affected by changes in sea ice concentration and extent.
  - High-resolution satellite imagery can be used to detect pack ice seals
  - However, classification must be done using automated algorithms that are computationally intensive

Antarctic pack-ice seals



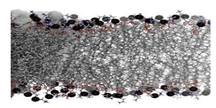
Seal haul out on HRS image



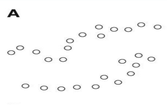
Sample image processing pipeline

Bio-Molecular Dynamics

- Leaflet Finder is a graph-based algorithm to detect continuous lipid membrane leaflets in a MD simulation*.
- Implemented and characterized based on MIDAS capabilities.

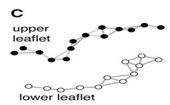


* Courtesy of Beckstein Lab at ASU

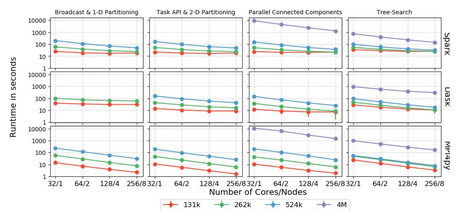


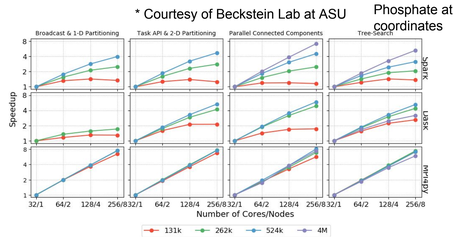
Phosphate atom coordinates

Build nearest-neighbors adjacency matrix



Find largest connected subgraphs





Future Work

My work will follow two main strands:

1. Middleware strand:
   a. Improve the capabilities of MIDAS to support more Data Intensive frameworks
   b. Provide a decision model which will allow users to decide which MIDAS capability to use based on their application
2. Workflow strand: Create the abstraction necessary to enable scalable image analysis for polar science

Acknowledgements

1. *We acknowledge financial support of NSF 1440677 and NSF 1443054*
2. *We acknowledge access to computational facilities on XSEDE resources via TG-MCB090174.*
3. *We acknowledge financial support of National Science Foundation, Office of Advanced Cyberinfrastructure, Registration Awards to ICPP 2018 PhD Forum Student Participants.*