Introducing MapReduce to High End Computing

Grant Mackey, Julio Lopez, Saba Sehrish, John Bent, Salman Habib, Jun Wang

University of Central Florida, Carnegie Melon University, Los Alamos National Laboratory

Scientific Applications

As the computational scale of scientific applications grows, so does the amount of data. Dealing with that amount of data becomes difficult.

- Data analytics become difficult
  - The data becomes too large to move
- Applications become resource intensive
- More difficult to program for
  - Do the older existing solutions scale?

Scientific Applications

Bioinformatics (Basic Local Alignment Search Tool)

- Genomics machines generate large datasets (GB~TB)
- Data is manually distributed in parallel through an adhoc job manager script
- The method of parallelizing BLAST is conceptually a manual MapReduce operation
  - Using Hadoop would abstract away the manual parallelization of tasks and would provide task resiliency

Scientific Applications

Cyber-Security: Real-time network analysis

- In a massively multi-user network environment, petabytes of information can pass of the network in a matter of months
- Need a scalable FS that can accommodate the large streaming datasets
- Network events are data independent
  - A programming model that abstracts parallelization from the user is convenient

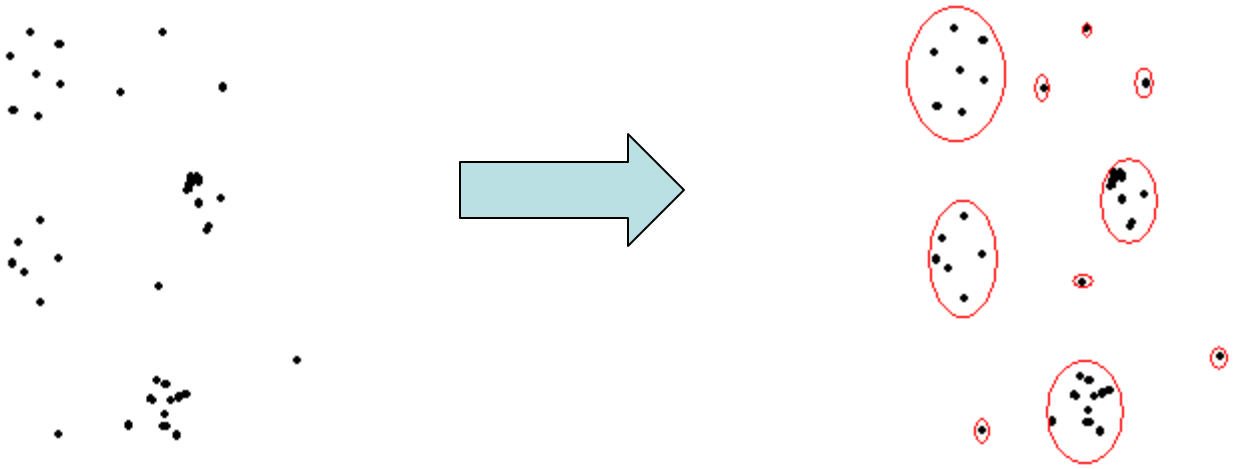
Scientific Applications

Astrophysics: Halo Finding

- Current issues
  - Ad Hoc: The approach is unique
  - Too much data movement
  - Parallel halo finding tasks are unreliable
- Hadoop Solutions
  - Provides a standard approach
  - No data movement
  - Hadoop ensures task resilience

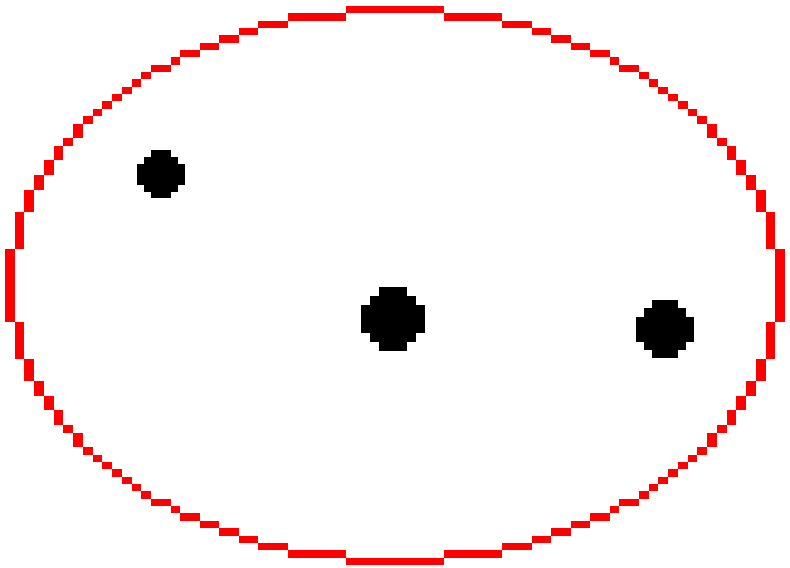
Halo Finding

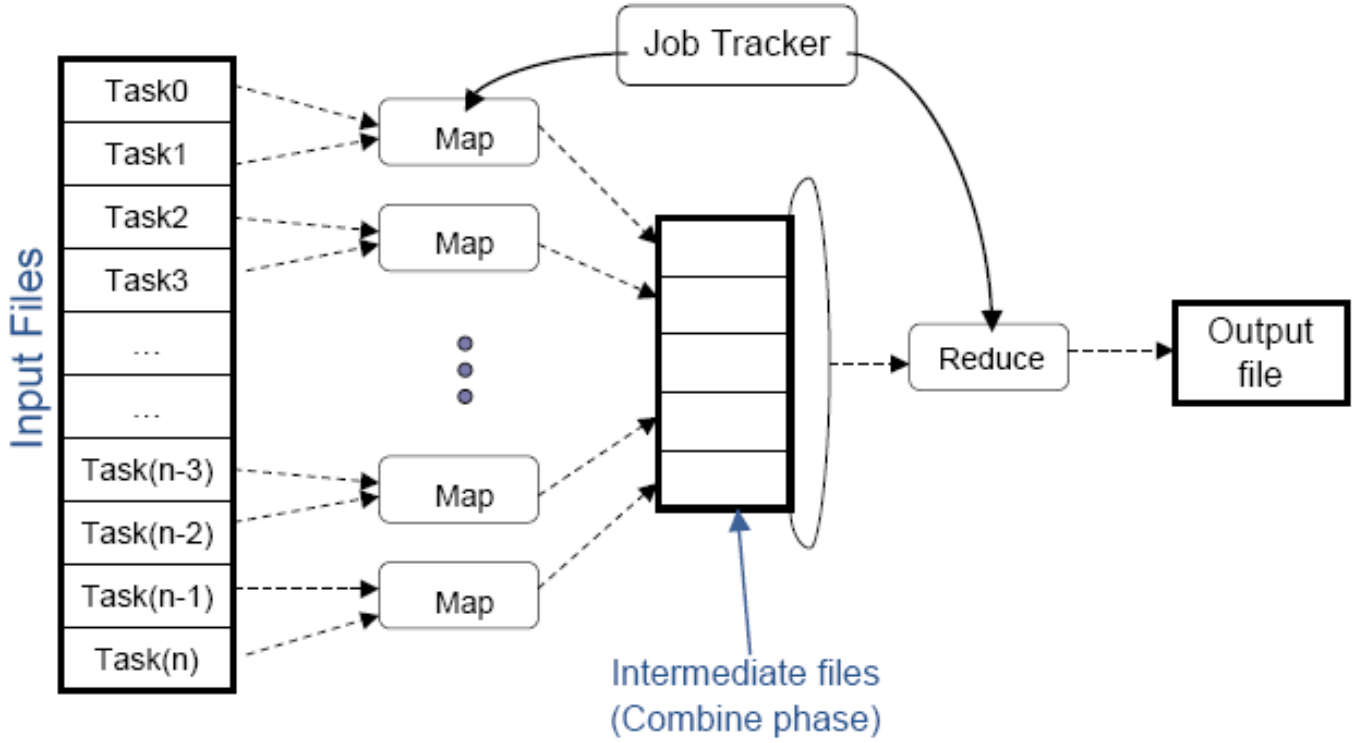
Method used to find clusters of particles in large astrophysics datasets.



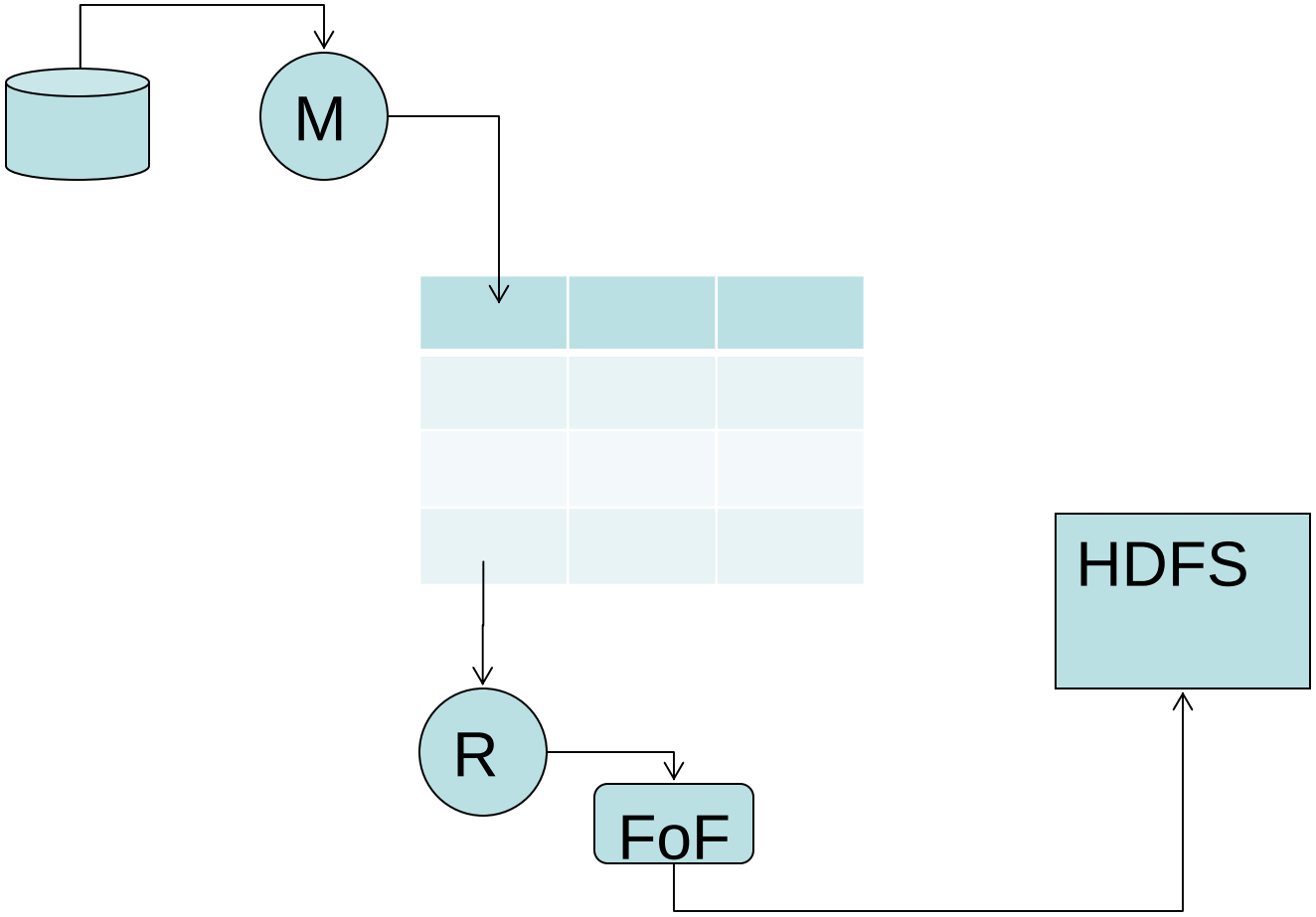
Friends of Friends

Algorithm used to perform halo finding





MapReduce model for Halo-Finding



Experiences

There is a reason why people think that Hadoop and is only good for data mining applications

- There exists little to no functionality for data types beyond text
- Learning curve for the language is steep for applications that deal with different data types such as binary
  - The programmer has to deal with the new programming model and write their own input classes

The Hadoop community is very active and incredibly helpful/prompt with responding to issues/bugs

Conclusion

Hadoop can be used as a viable resource for large data intensive computing

- Hadoop runs on an inexpensive commodity computing platform, but provides powerful tools for large scale data analytics
- The Hadoop architecture provides for task resiliency that other scientific computing methods cannot
- Hadoop allows for a strict model in which to parallelize a task and the parallelization has been shown to scale to 1000+ node cluster environments (Amazon's S3 cluster)

Hadoop needs more functionality in its API for other data formats

Contact

Grant Mackey: gmackey@cs.ucf.edu

Julio Lopez: jclopez@andrew.cmu.edu

Saba Sehrish: ssehrish@cs.ucf.edu

John Bent: johnbent@lanl.gov

Jun Wang: jwang@cs.ucf.edu