Astroinformatics: Day 1

Adam W. Lavely, Ph.D.

Afternoon Seminar B

June 4 2018

Slides/Files available: `https://goo.gl/Wz72vs`

PennState
Institute
for CyberScience

MapReduce definition

Q: What is the map-reduce framework?

A: It depends on who you ask!

Adam's definition:
A framework where **independent tasks** are completed where **communication is limited** to initializing/finalizing the processes.

Notes:

- # of tasks $\geq$ # of cores
- No communication during the task
- What is done to 'reduce' can be varied

Examples

Split-Apply-Combine example: *Elections*

- Voting location is mapped on population
- People vote at a location ()independent of other locations)
- Final tally is found by gathering statistics at one central point

Map-shuffle-(apply)-reduce example: *Genomics*

- Processors are mapped on DNA sequences (big data)
- Processors sort DNA to different processors based on attribute
- Processors analyze strands with common characteristic
- Results are reduced to a final form

Possible Mapreduce frameworks

Big Data	Spark/hadoop's split-apply-combine
Statistics	Monte-Carlo simulation parameter sweeps
SIMD	Code vectorization
Fork/Join	Subset of threading parallelizations
Job Arrays	Submit many HPC jobs within a single script
GNU parallel	Tool to run bash commands in a loop in parallel

Embarrassingly parallel: no communication and almost no overhead outside of start/finish

What does reduction mean?

Reduction is what you do with your data:

- Send to 1 processor
  - Statistics: mean, min/max, distribution
- Send complete message
  - Know that the analysis is done
- Write out to individual file
  - For later analysis