Riding the Elephant: Managing Ensembles with Hadoop

Elif Dede, Madhusudan Govindaraju,

Dan Gunter,

Lavanya Ramakrishnan

U.S. DEPARTMENT OF ENERGY | Office of Science

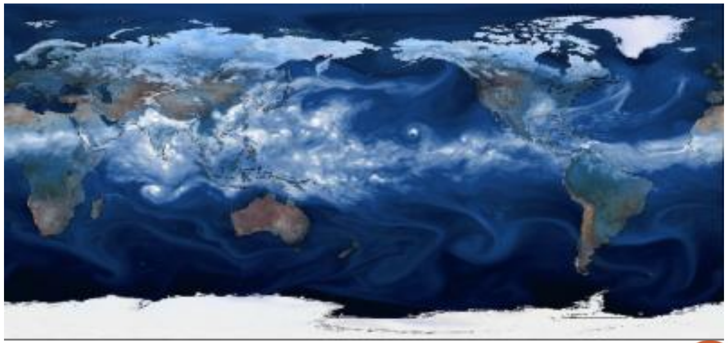
CRD computational research division

BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

Uncertainty Quantification

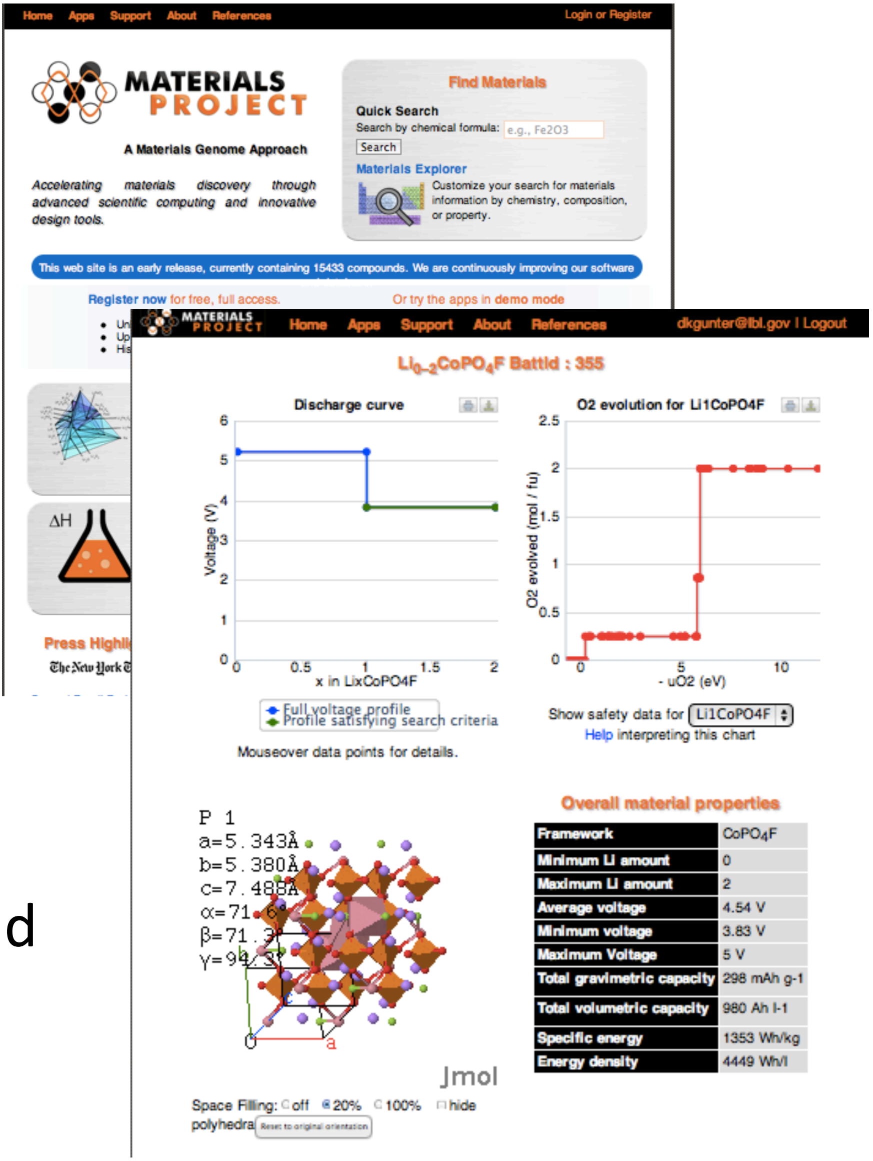
- Climate Scientists running simulations on global warming.
  - How soon it is going to be too late?

- UQ: How accurate is our simulation for a given set of inputs?
  - Running multiple versions of the same simulation to sample the multidimensional real life scenarios.



Materials Project*

- Calculate crystal structure for ~125,000 known compounds
- Store results in a DB
- Provide web interface to explor database and run “apps” to calculate additional properties such as diffraction patterns and phase diagrams
- Released 10/2011
- Collaboration between LBNL and MIT
- **www.materialsproject.org**



A Materials Genome Approach

Accelerating materials discovery through advanced scientific computing and innovative design tools.

Customize your search for materials information by chemistry, composition, or property.

This web site is an early release, currently containing 15433 compounds. We are continuously improving our software

Register now for free, full access. Or try the apps in demo mode

$Li_{0-2}CoPO_4F$ BattId : 355

Mouseover data points for details.

Show safety data for Li1CoPO4F

Help interpreting this chart

Space Filling: polyhedra

Overall material properties

Framework	$CoPO_4F$
Minimum Li amount	0
Maximum Li amount	2
Average voltage	4.54 V
Minimum voltage	3.83 V
Maximum Voltage	5 V
Total gravimetric capacity	298 mAh g-1
Total volumetric capacity	980 Ah l-1
Specific energy	1353 Wh/kg
Energy density	4449 Wh/l

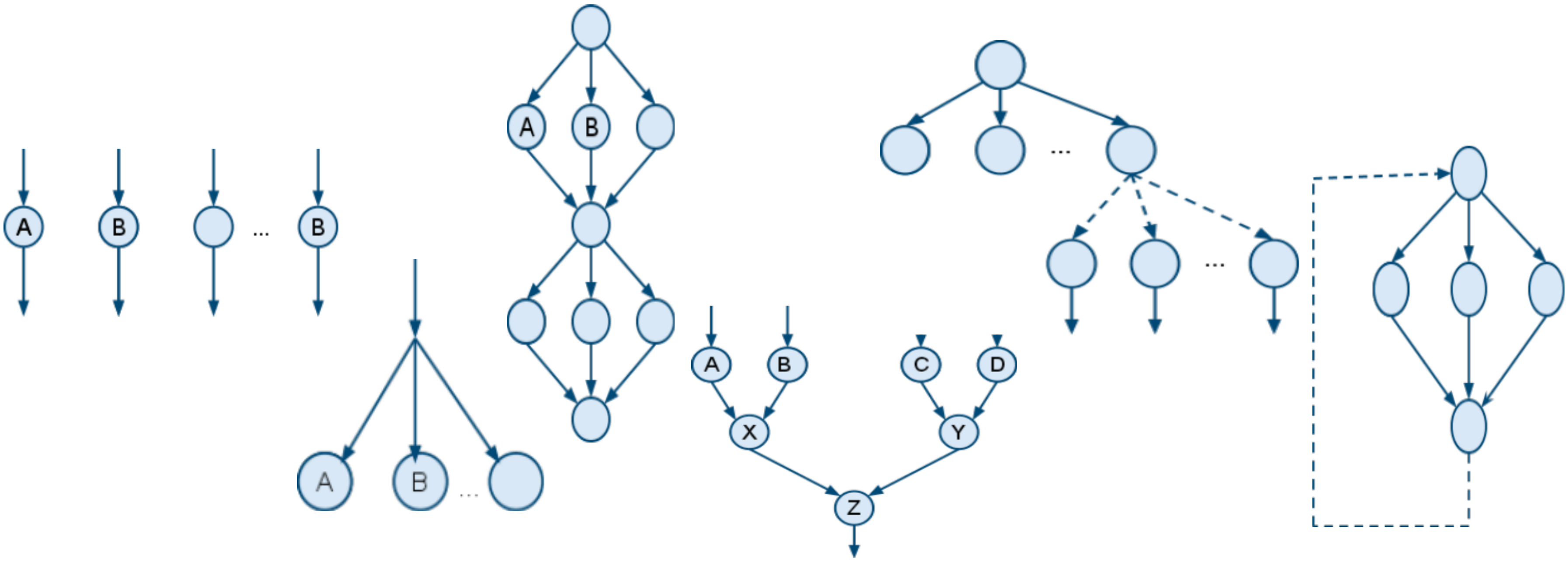
* was: Materials Genome

U.S. DEPARTMENT OF ENERGY | Office of Science

CRD computational research division

BERKELEY LAB LAWRENCE BERKELEY NATIONAL LABORATORY

Code Ensembles



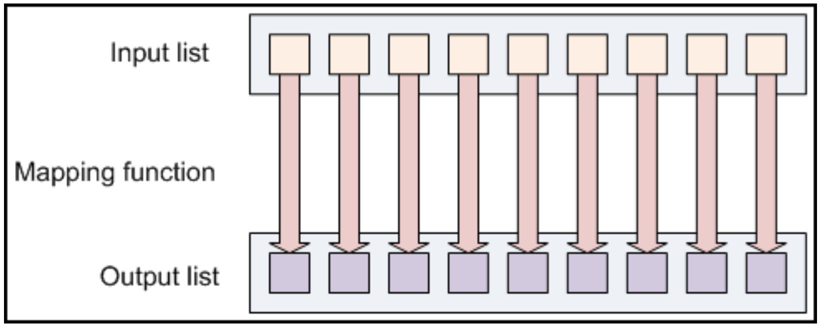
A large number of loosely coupled tasks, each with their own internal parallelism.

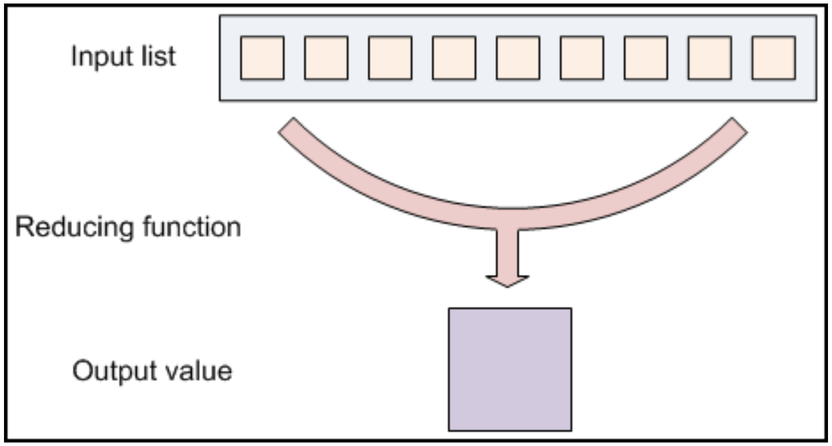
Related Work

- DAKOTA, PSUADE, etc.
  - Parallelism assumes single batch queue
  - Limited monitoring and fault tolerance
  - Data management doesn't scale very well
- Pegasus, Taverna, Kepler, etc.
  - Mature workflow tools from distributed/grid computing
  - Focused on simply acquiring resources
  - Do not deal well with dynamic elements, HPC batch queues

MapReduce

- Computation performed on large volumes of data in parallel
  - divide workload across large number of machines
  - need a good data management scheme to handle scalability and consistency
- Functional programming concepts
  - map
  - reduce





OSDI 2004

Hadoop

- Open source reliable, scalable distributed computing platform
  - implementation of MapReduce
  - Hadoop Distributed File System (HDFS)
  - runs on commodity hardware
- Fault Tolerance
  - restarting tasks
  - data replication
- Speculative execution
  - handles stragglers

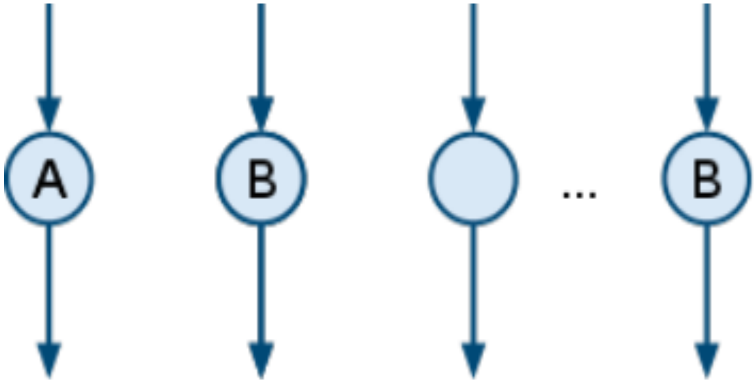
Scientific Ensembles and MapReduce Jobs

- Data Flow Parallelism
- Similar Job Phases
  - data preparation, transformation and reduction
  - MapReduce: maps (transformation) and reduces (reduction)
- Number of maps >>> Number of reduces
- Fault Tolerance and Data Locality important

Evaluation

- MapReduce Realization
- Hadoop Implementation
- Data Management
- Performance and Reliability

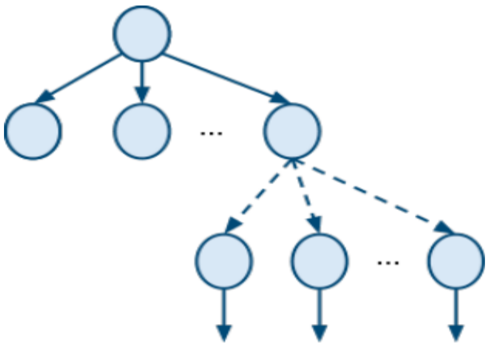
Data Parallel



- Most similar to MapReduce

- Reducers
  - Identity or None
- Data management
  - NonSplittable so each map gets one file
- For non-identical mappers
  - extra logic is required

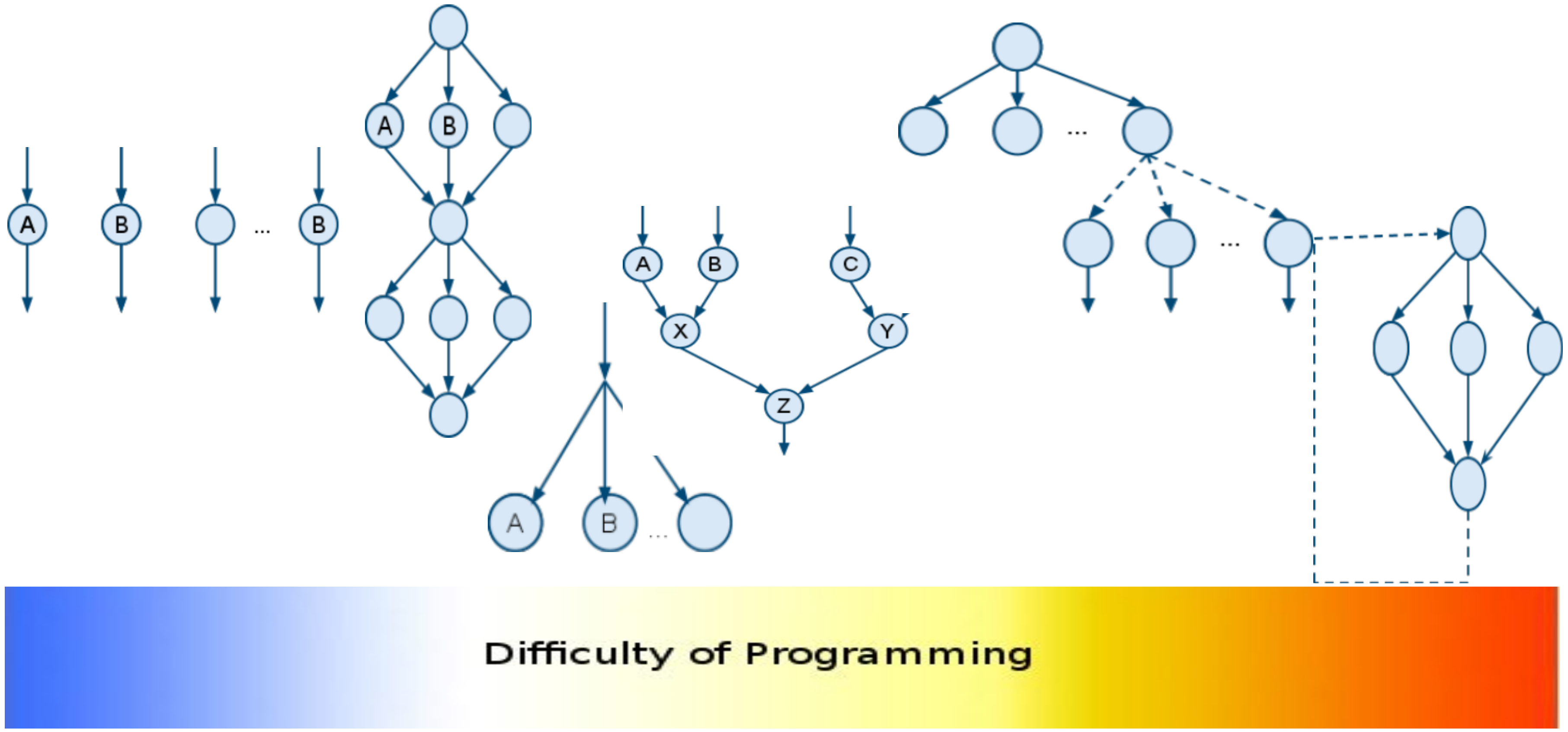
Dynamic



- No in-built support for dynamic elements

- Jobs can’t be launched from within jobs
- Manage a queue in the filesystem
- Intermediate data products are lost to data locality modules in Hadoop

Workflows on Hadoop



low

high

○ Easy ◓ Moderate ● Difficult

Workflow Type	Hadoop Impl	Data Management	MR Realization	Perf. and reliability	Total Score
Data Parallel	○	○	○	○	0
Single Input	○	◓	◓	◓	1.5
Sca-Gather	◓	◓	◓	◓	2
Inv. Tree	◓	●	●	◓	3
Dynamic	●	○	●	●	3.5
Iterative	●	○	●	●	3.5

Other Challenges

- Java implementation; Hadoop streaming allows other modes – but restricted model
- Non-POSIX file system
- Scientific data formats don't fit in the line-oriented/ text inputs of typical Hadoop jobs
- Maps and reduces are considered identical

Conclusions and Future Work

- It was possible to implement all patterns
  - but required significant work for some
  - dynamic and iteration are not well-supported tasks
- Data locality is important; need to consider multiple files for science
- MapReduce implementations for science
- Need to investigate appropriate programming model or extensions to MapReduce to handle scientific ensembles

Other related events at SC

- At Lawrence Berkeley Booth
  - Science in the Cloud? Busting Common Myths about Clouds and Science Tue Nov 15 at 10:30 am
  - What do Clouds mean for Science? Experiences from the Magellan Project Tue Nov 15 at 11:15 am
  - Demonstration of Materials Project. Mon night 7-9, Tues-Thu most of the day.
- Papers
  - Evaluating interconnect and virtualization performance for high performance computing, Sun Nov 13 at 9:00 am
  - Understanding I/O Performance of Virtualized Cloud Environments, DataCloud Workshop, Mon Nov 14 at 4:00 pm