



Benchmarking Discussion Report

Benchmarking Discussion Team
March 17–18, 2016



Key Concepts

- Current state-of-the art benchmarking research are not yet mature enough to benchmark big data analytics
- New benchmarks should consider new technologies, new architectural platforms, new algorithms, and new computing scenarios.

Aspects of Algorithm Benchmarking

- Performance accuracy
- Time, space and resources used
- Speed-up: change as infrastructure grows
- Scale-up: change as data size grows
- Performance accuracy vs. resource cost tradeoff
- Algorithm security (can we trust running it?)
- Method of algorithm execution
- How general is the algorithm?

Benchmarking on the Cloud

- Software non-local, or on others' infrastructure
- Can't control the infrastructure
- Software: centralized or distributed?
- HPC benchmarking \neq cloud benchmarking
 - Cloud is distributed (low-speed network)
 - Cloud uses local disks, distributed file system
 - Cloud uses lots of small nodes
- Benchmarks should be effective on lots of platform

New Scenario: Containers

- Baremetal, VM, Docker, etc.
- Security issues of OS can impact containers
- Containers need to be dynamically created and destroyed to support scalability
- Want to spawn many containers on a node, but how many are possible?
- Quantify container overhead: benchmark same program in different containers

Benchmarking Comparisons

- Comparisons can be misleading
 - Example: “Spark is 100X faster than Hadoop”
- Must benchmark data migration: include time to put data “in place”
- Same issues in GPU benchmarking
- Caching should not be optimized for specific programs or program executions
- Code owners can optimize their systems