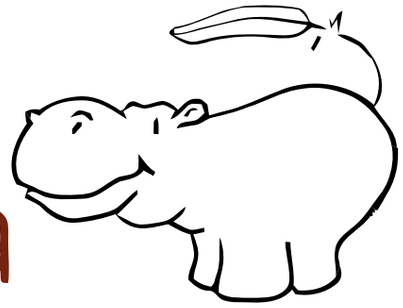
ICDCS'15, Columbus, USA

FLOWPROPHET: Generic and Accurate Traffic Prediction for Data-parallel Cluster Computing

Hao Wang[1,2], Li Chen[2], Kai Chen[2], Ziyang Li[2,3], Yiming Zhang[3], Haibing Guan[1], Zhengwei Qi[1], Dongsheng Li[3], Yanhui Geng[4]

[1] Shanghai Jiao Tong University
[2] Hong Kong University of Science and Technology
[3] National University of Defense Technology
[4] Huawei Technologies Co. Ltd.

Apache Hama

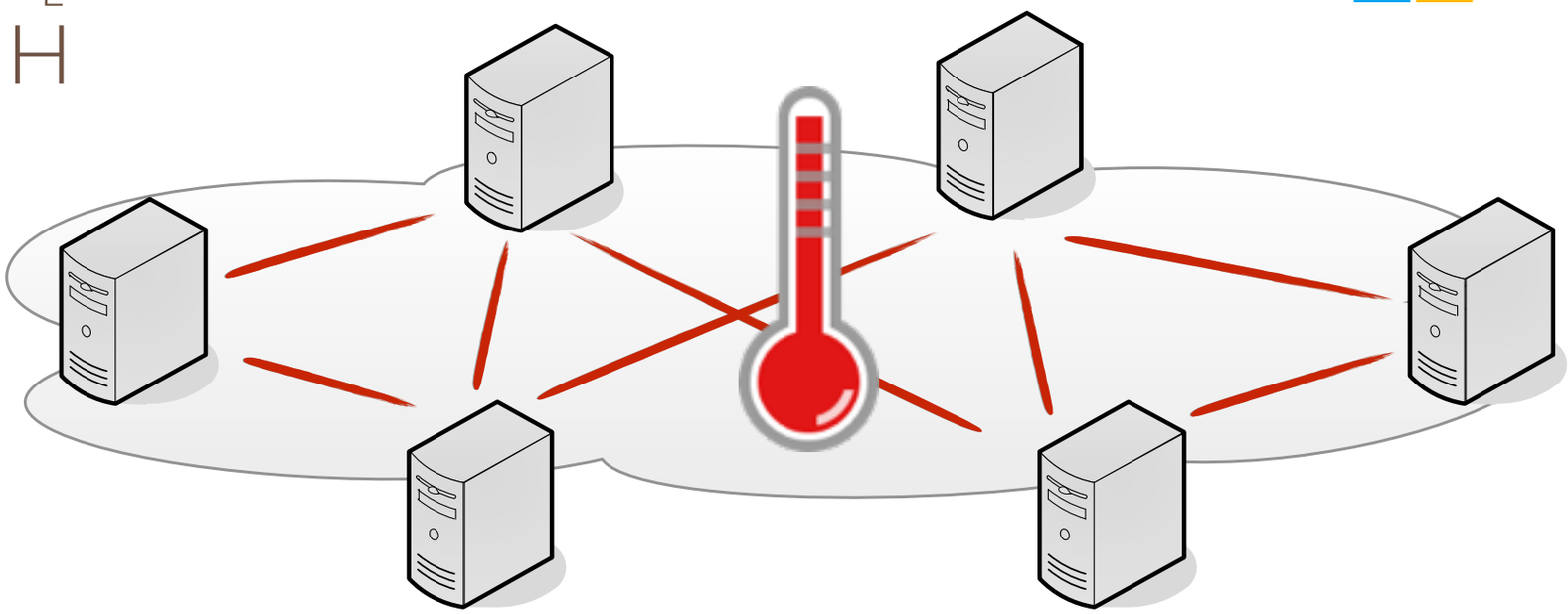


Spark

Dryad

Microsoft

. . .



Flow-based optimization mechanisms:

- PDQ [Sigcomm'12], pFabric [Sigcomm'13], PASE [Sigcomm'14], Varys [Sigcomm'14], Baraat [Sigcomm'14]

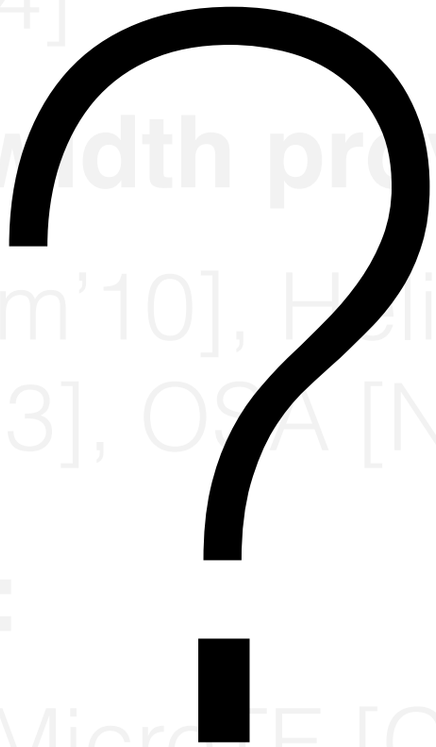
Architectural bandwidth provisioning:

- c-Through [Sigcomm'10], Helios [Sigcomm'11], Mordia [Sigcomm'13], OSA [NSDI'12]

Traffic engineering:

- Hedera [NSDI'10], MicroTE [CoNEXT'11], D^3 [Sigcomm'11]

Knowing the Flow Information Ahead of Time

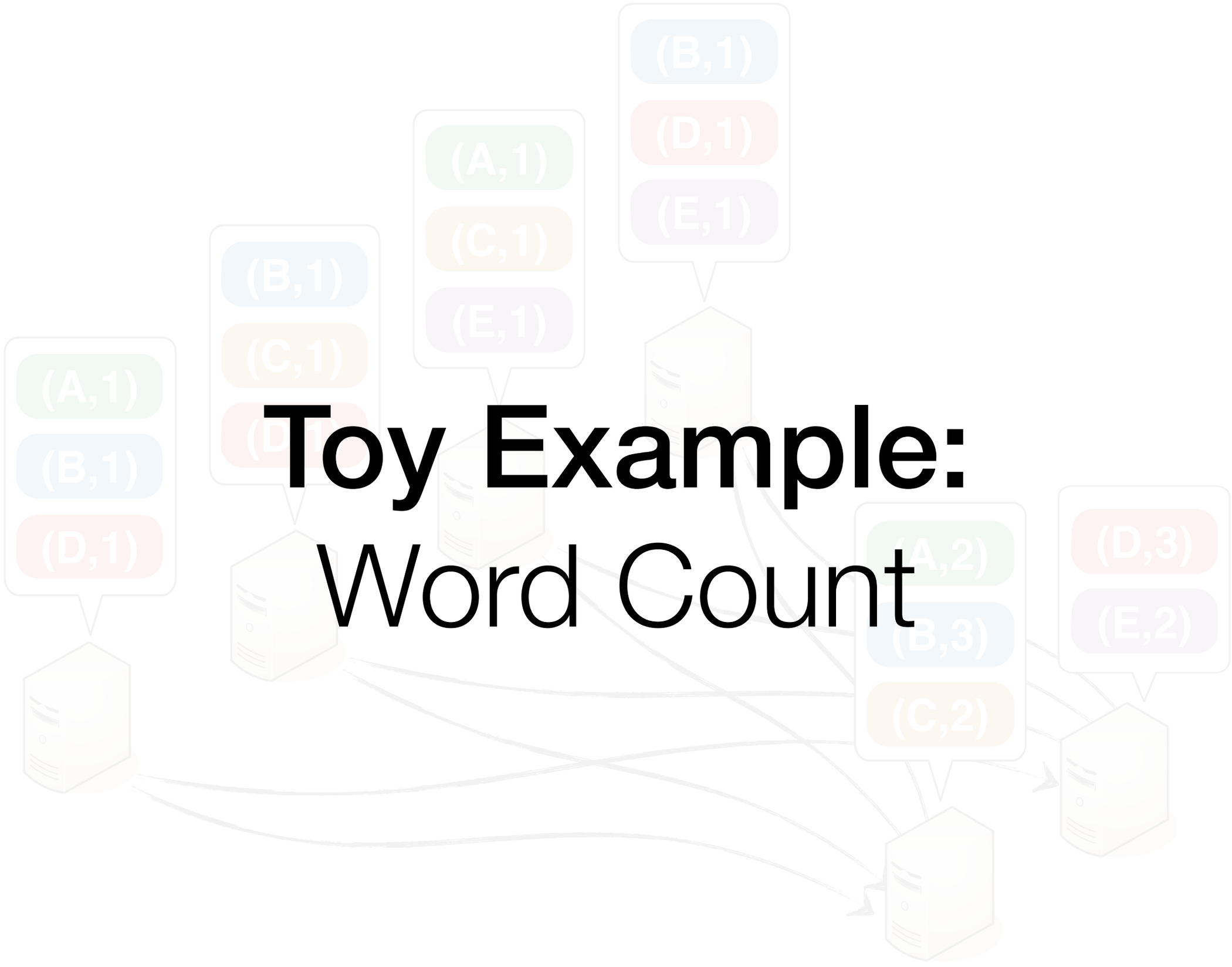


FLOWPROPHET

- Generic for DCFs
- Accurate and fined-grained
- Ahead-of-time
- Scalable and low-overhead

Toy Example:

Word Count

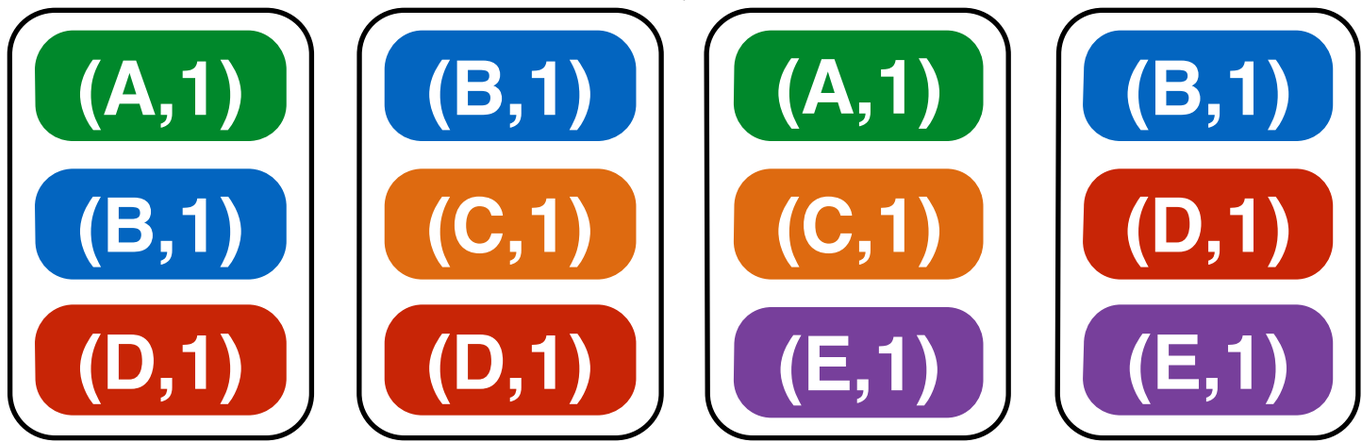


Logical View

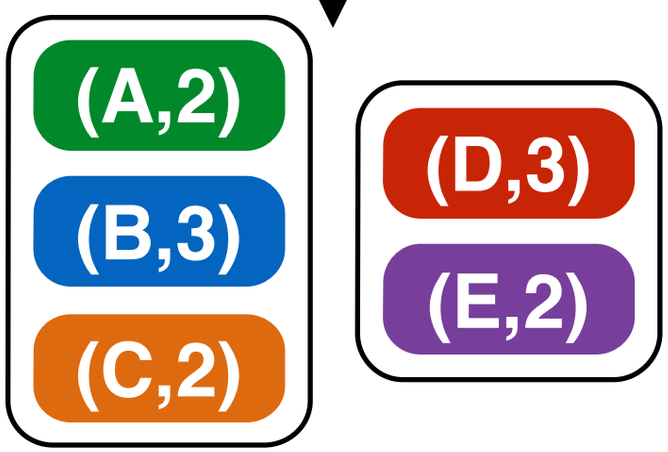
A···B D···E C···A E C···A···B

B D···E C···A···A D···E B···

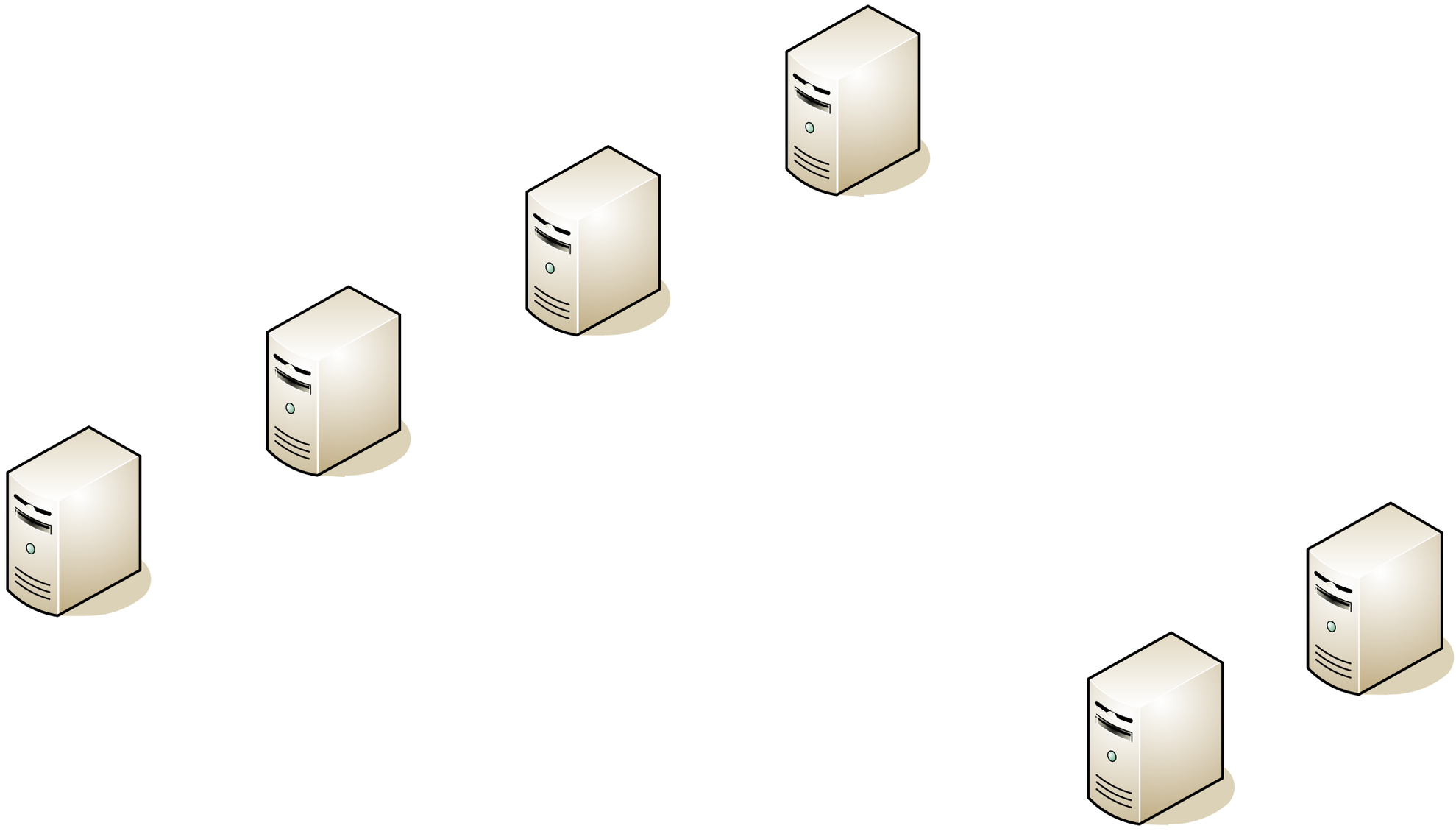
map()



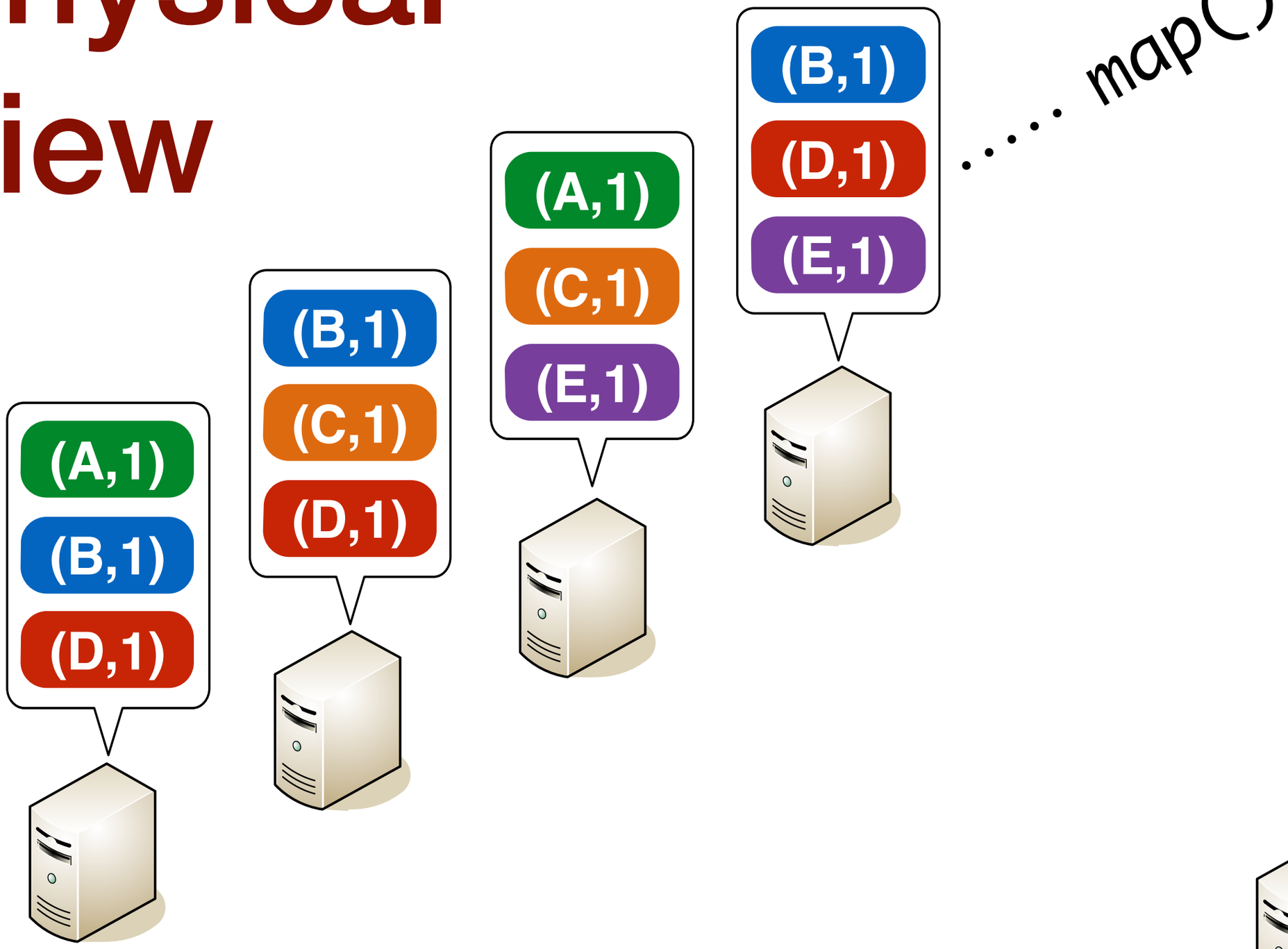
reduce()



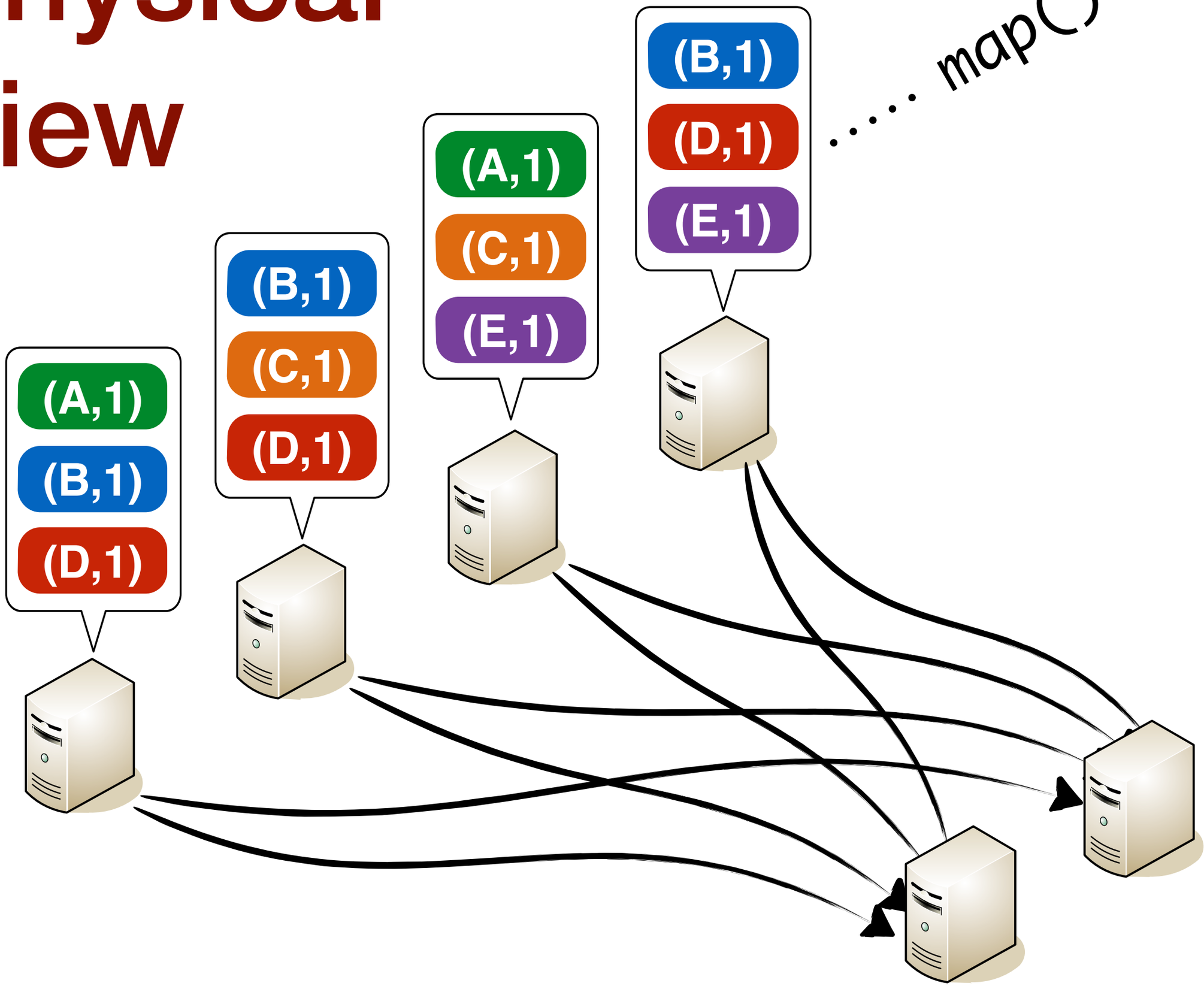
Physical View



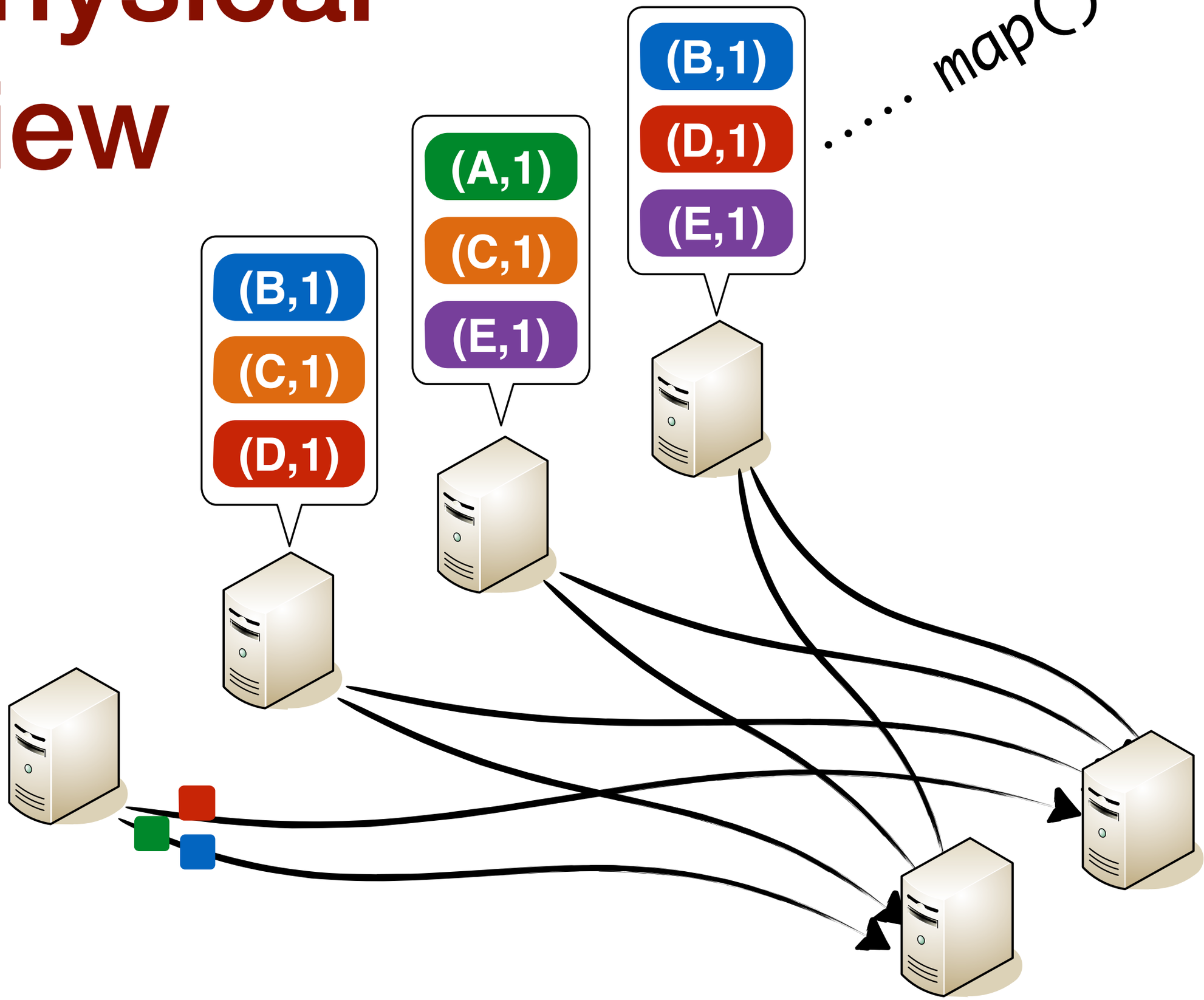
Physical View



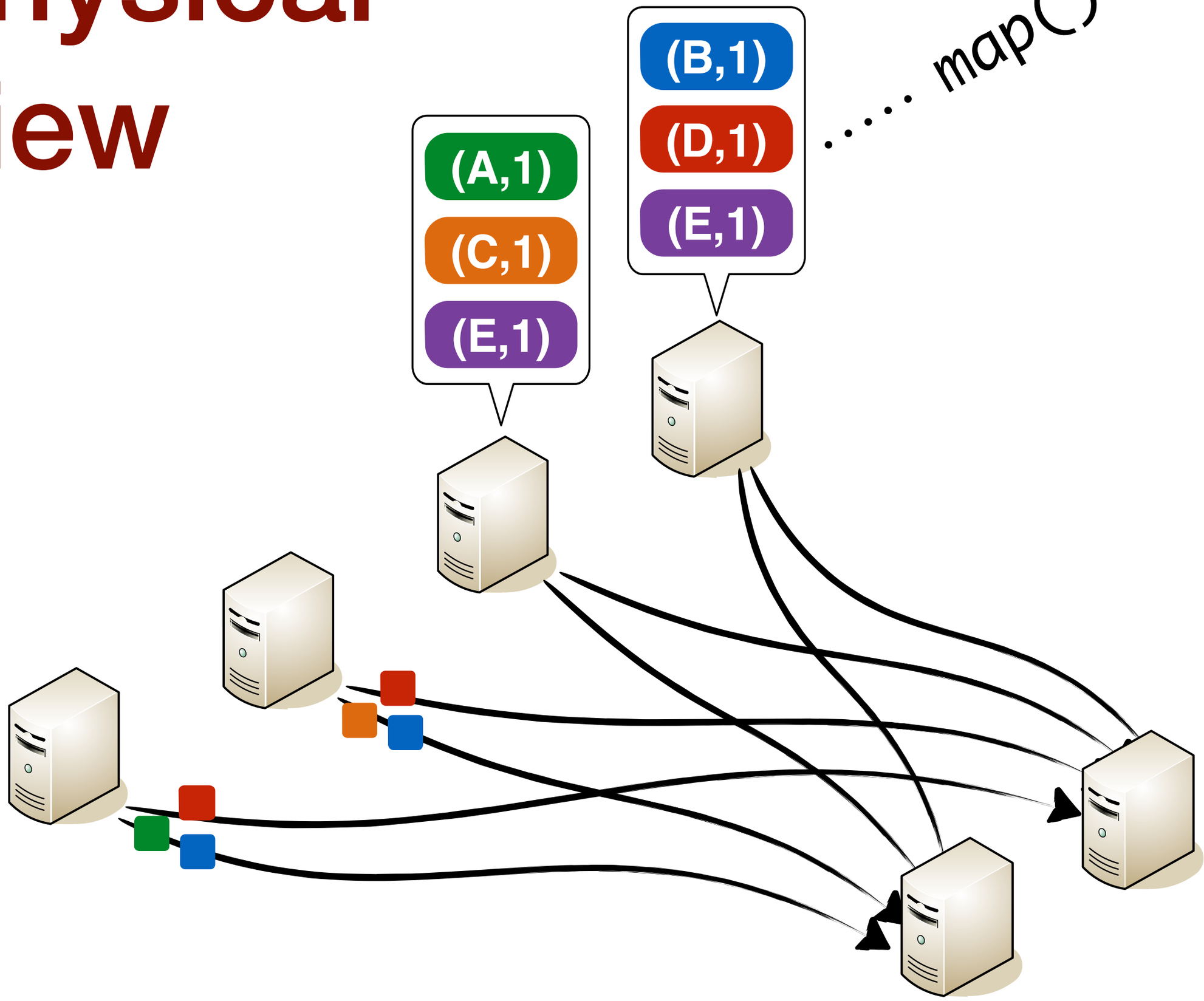
Physical View



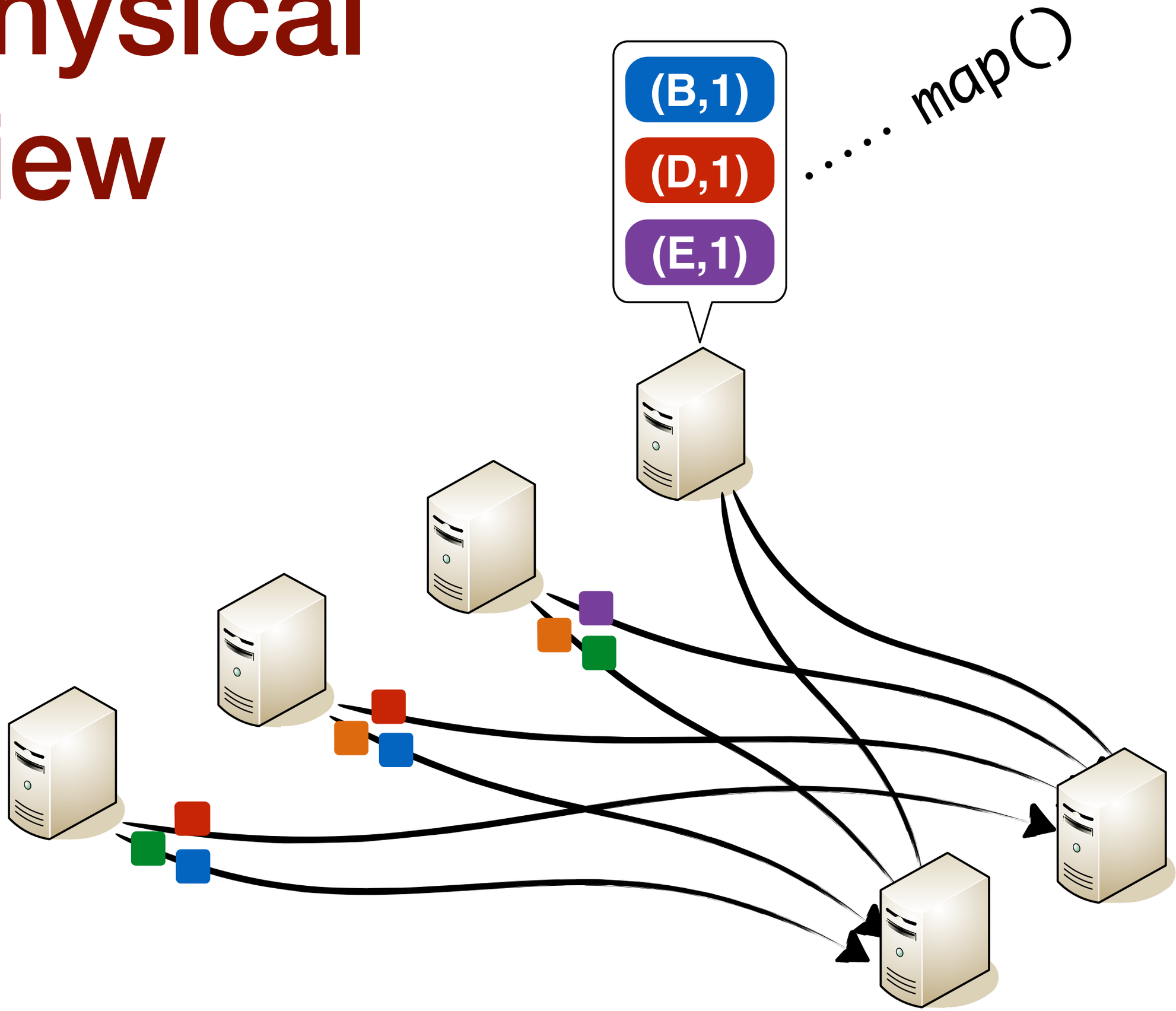
Physical View



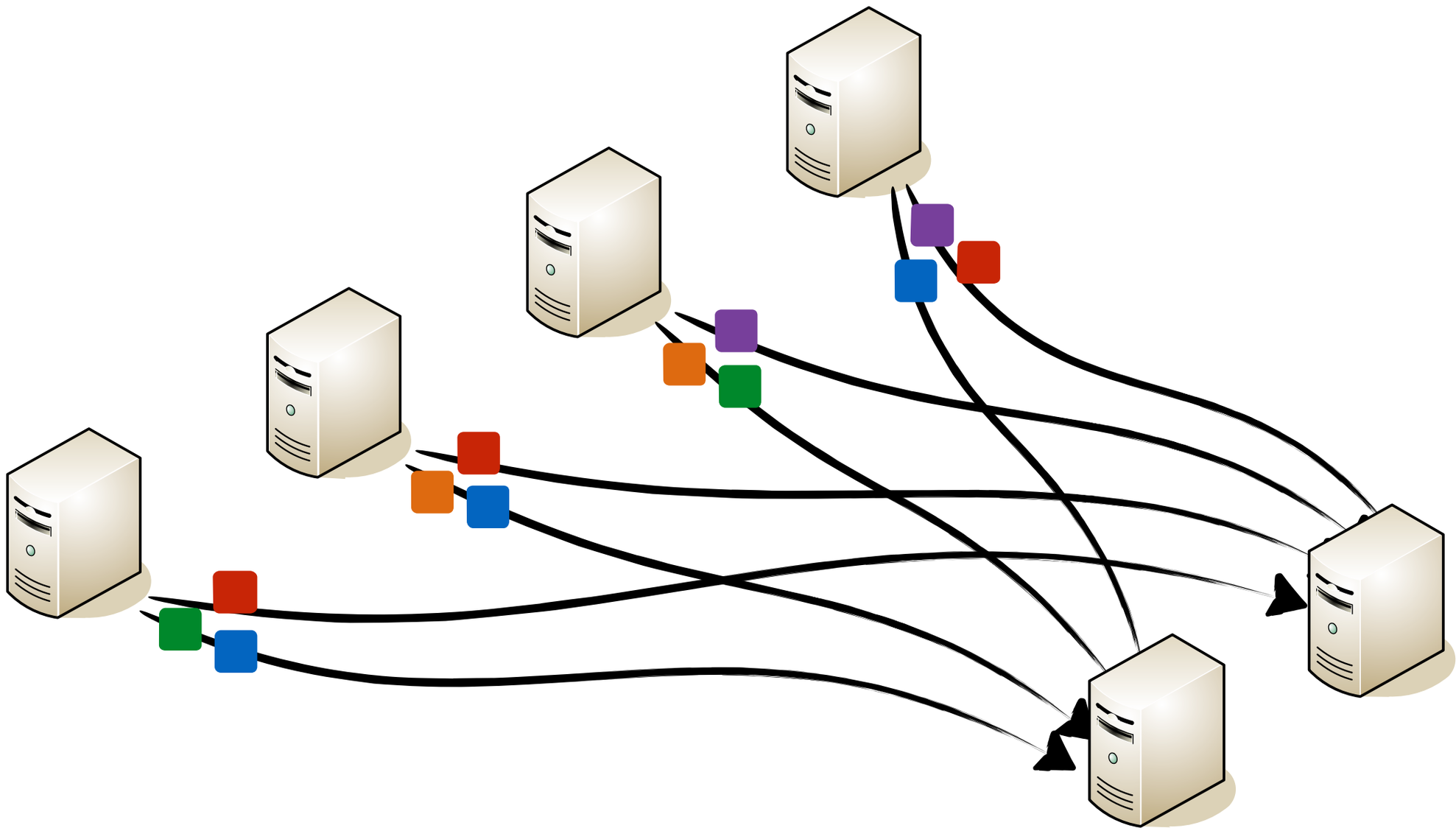
Physical View



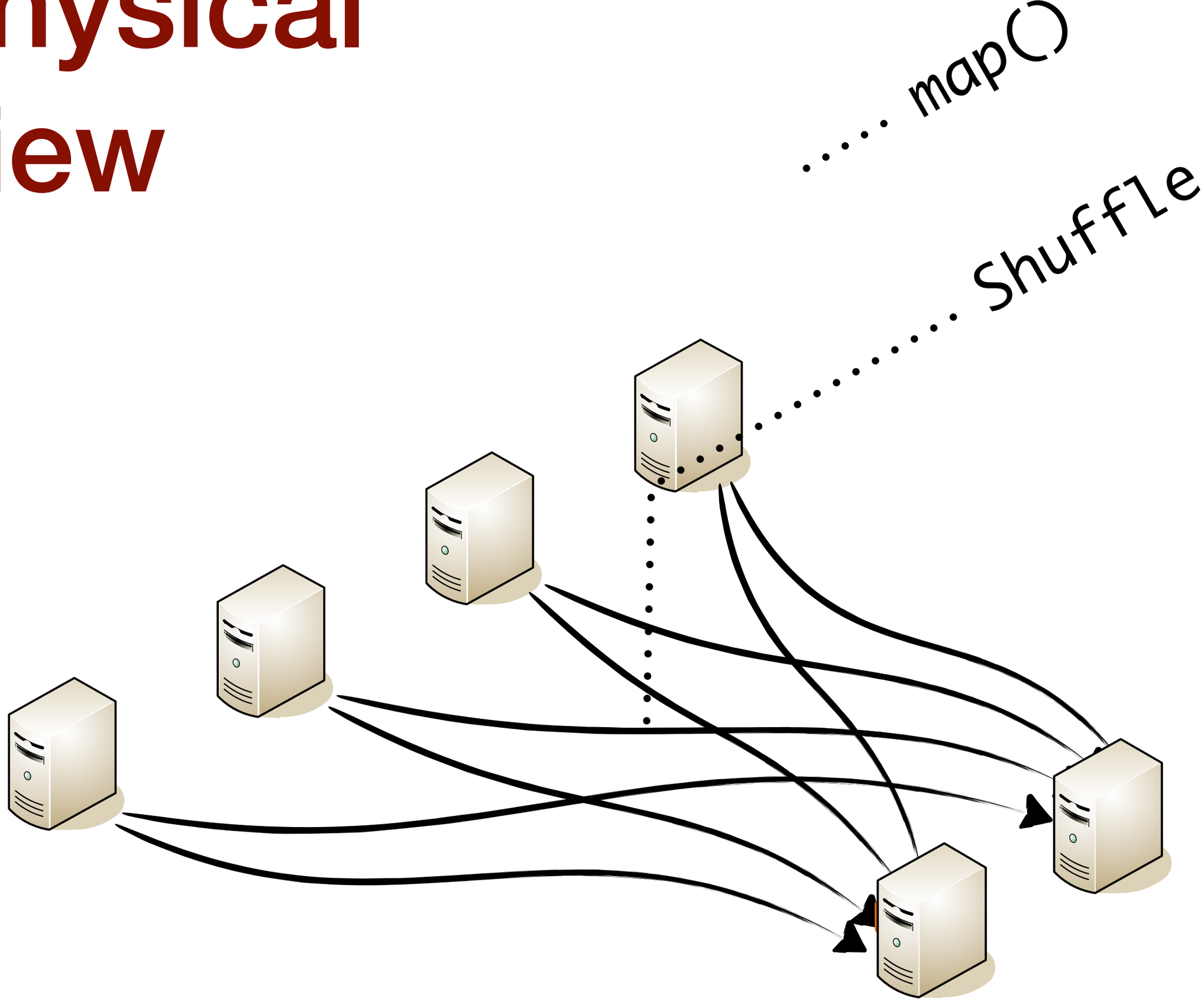
Physical View



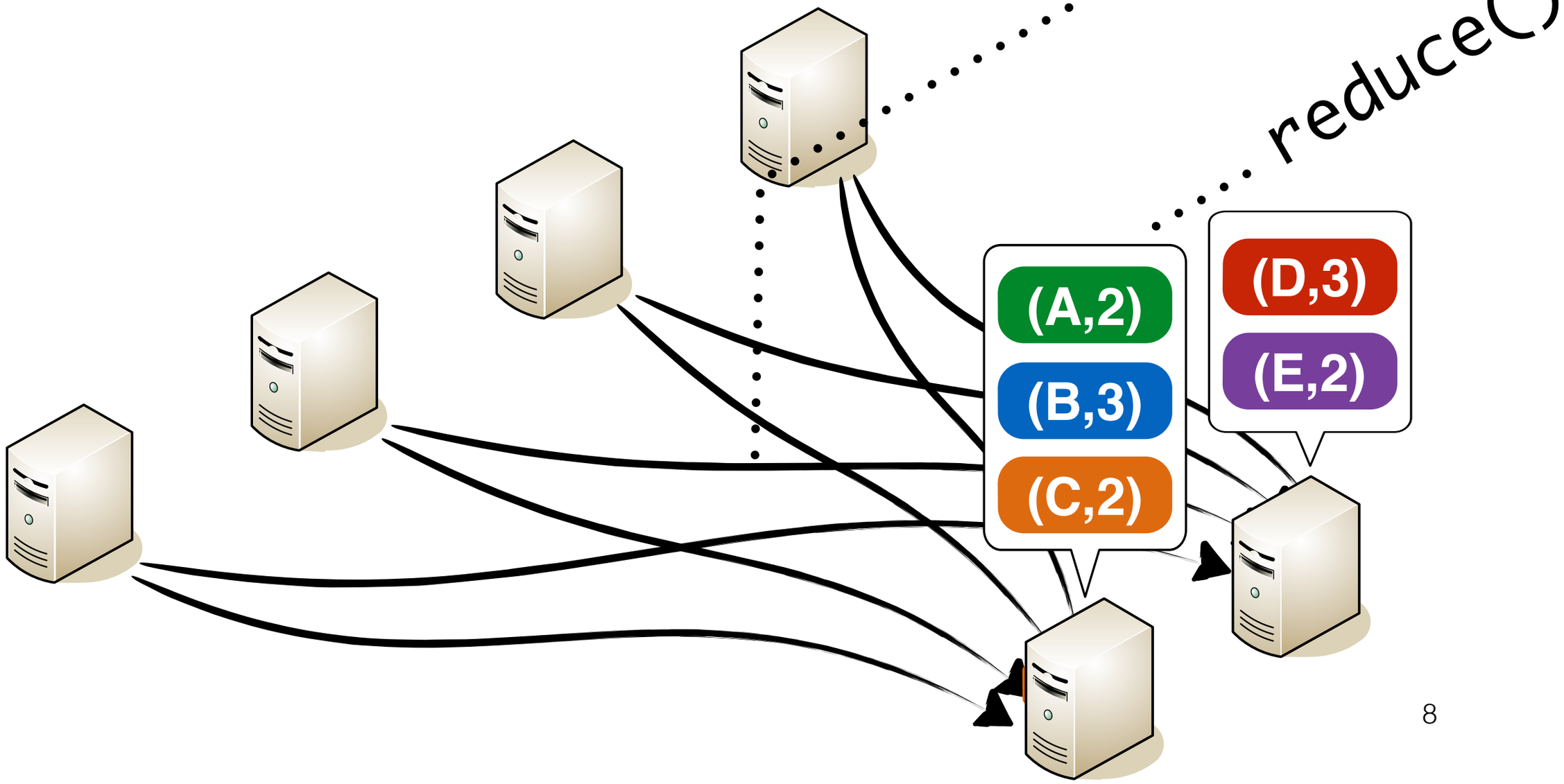
Physical View



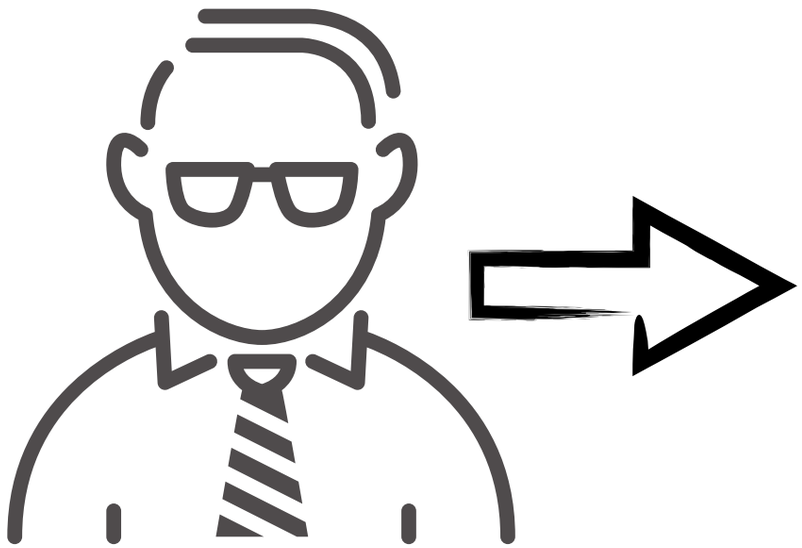
Physical View



Physical View



User

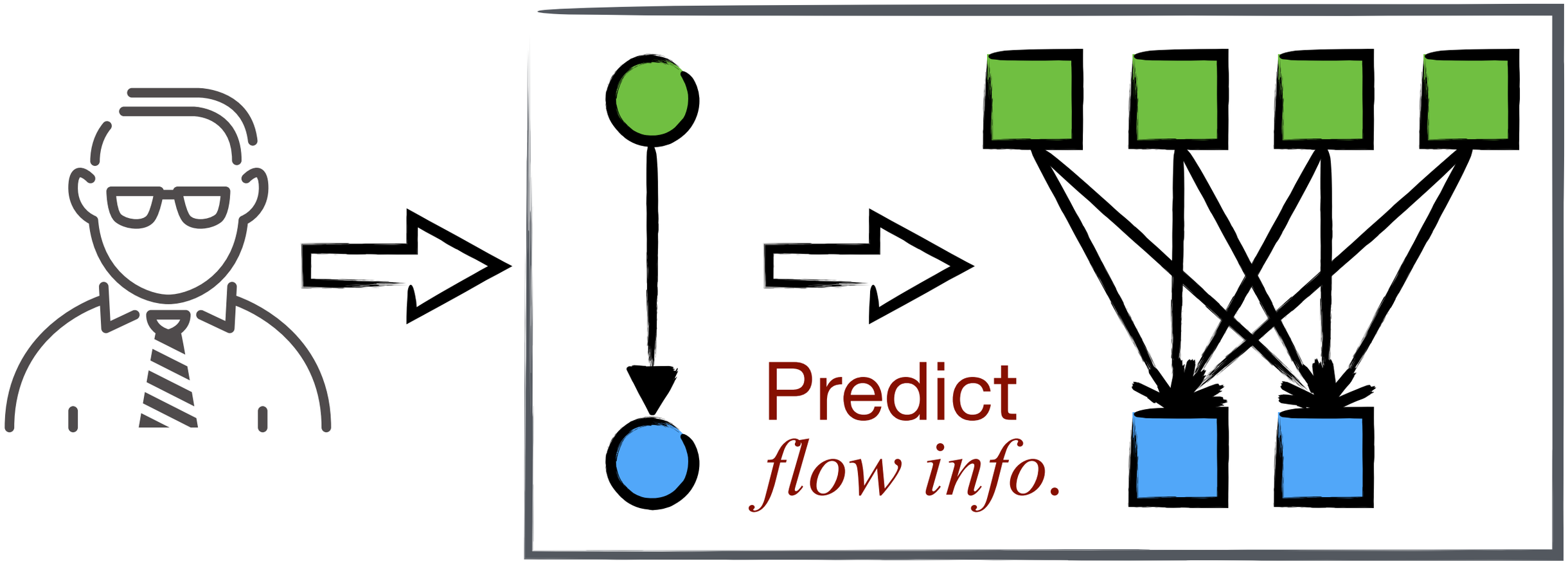


Distributed Computing Frameworks

User

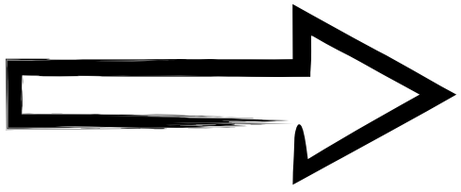
Logical View

Physical View

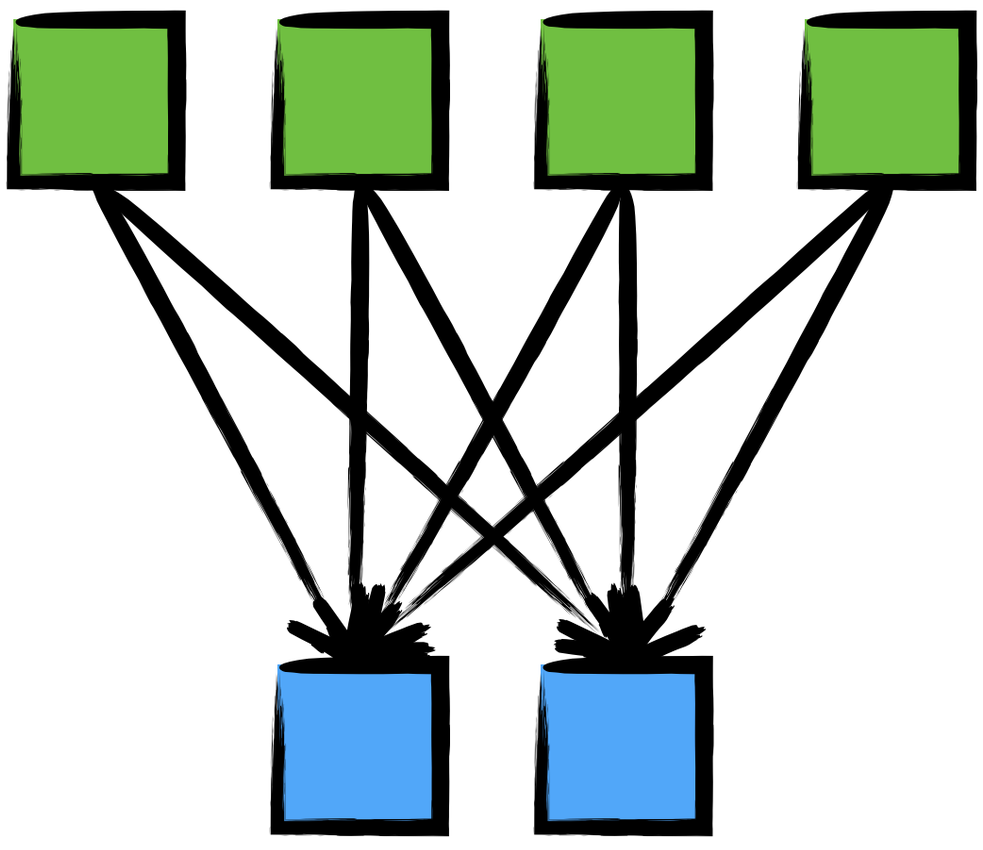


Logical View

Predict



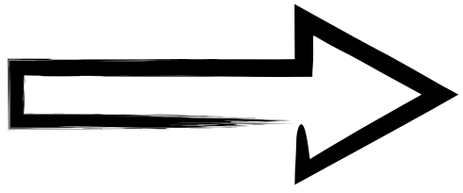
Physical View

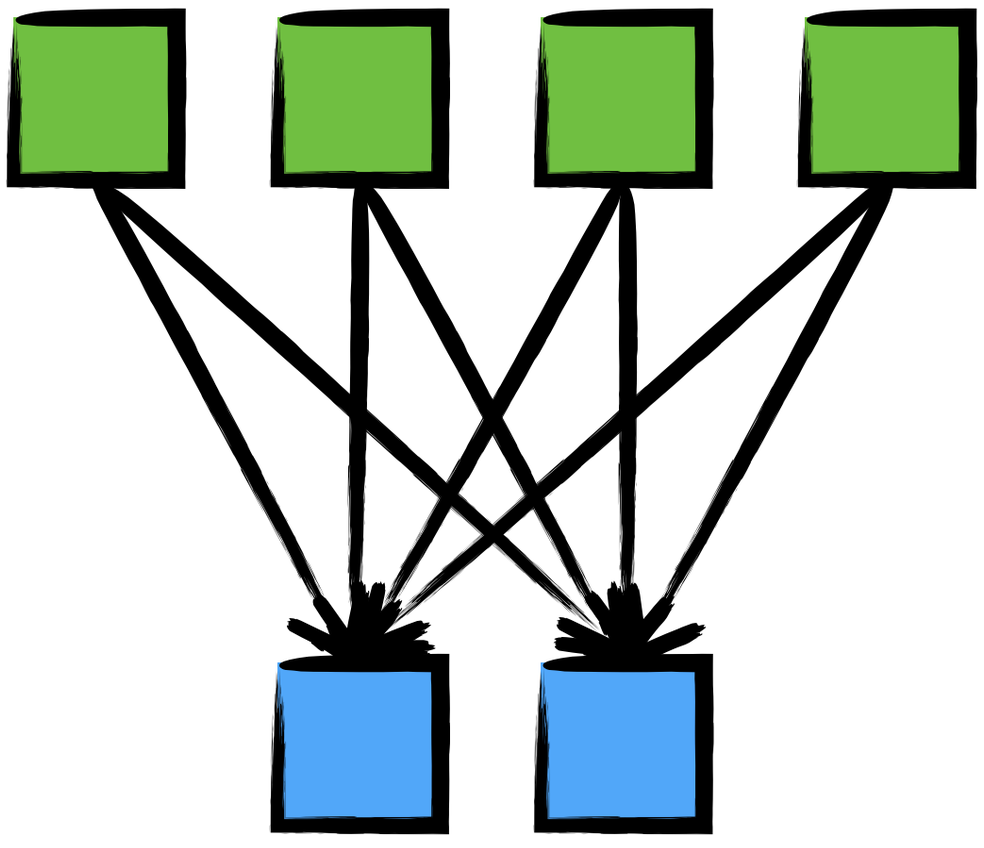


Logical View

Flow info.

Predict

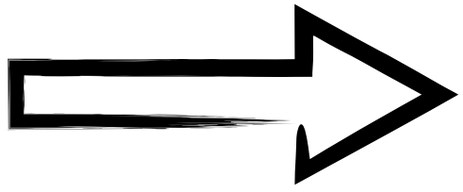


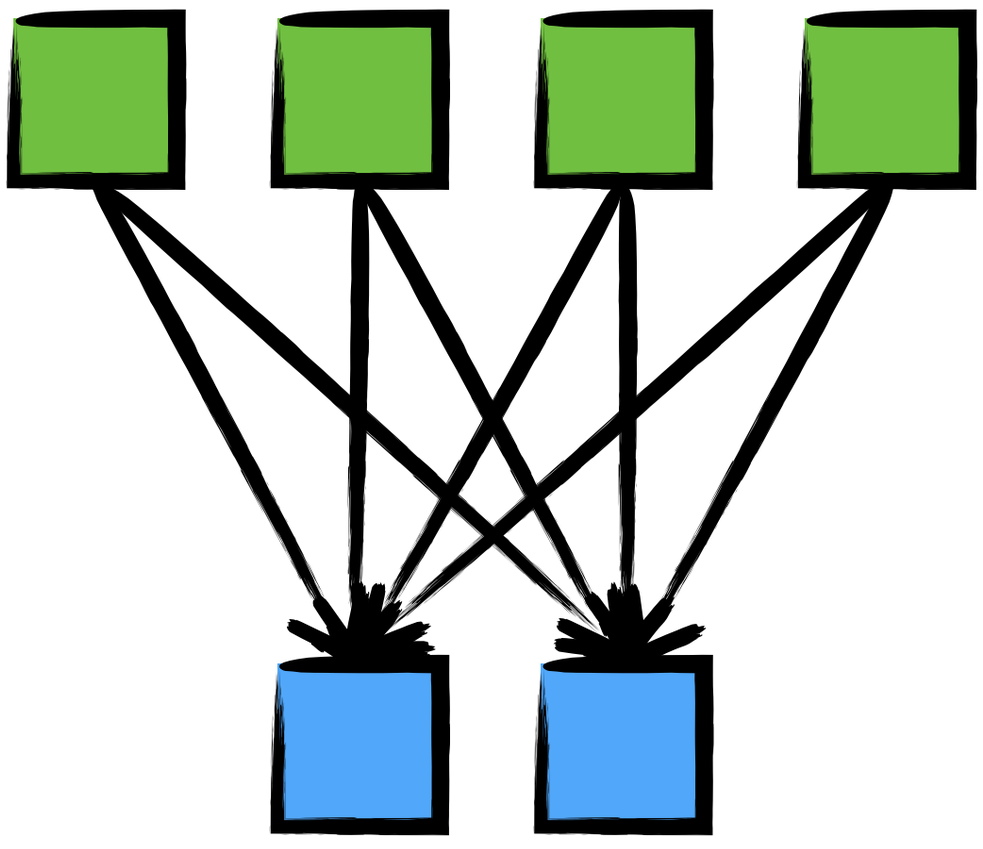


DAG

Flow info.

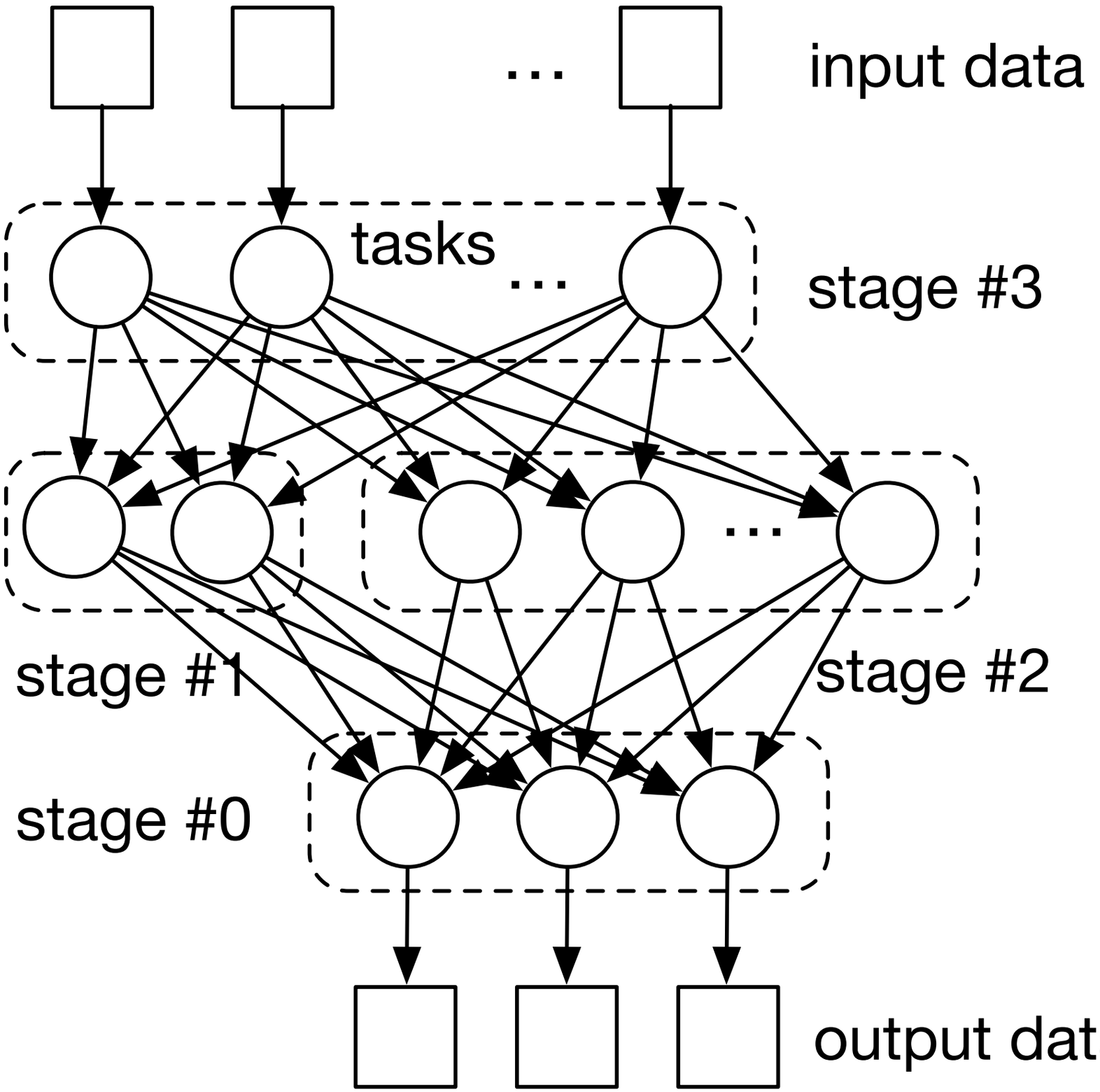
Predict





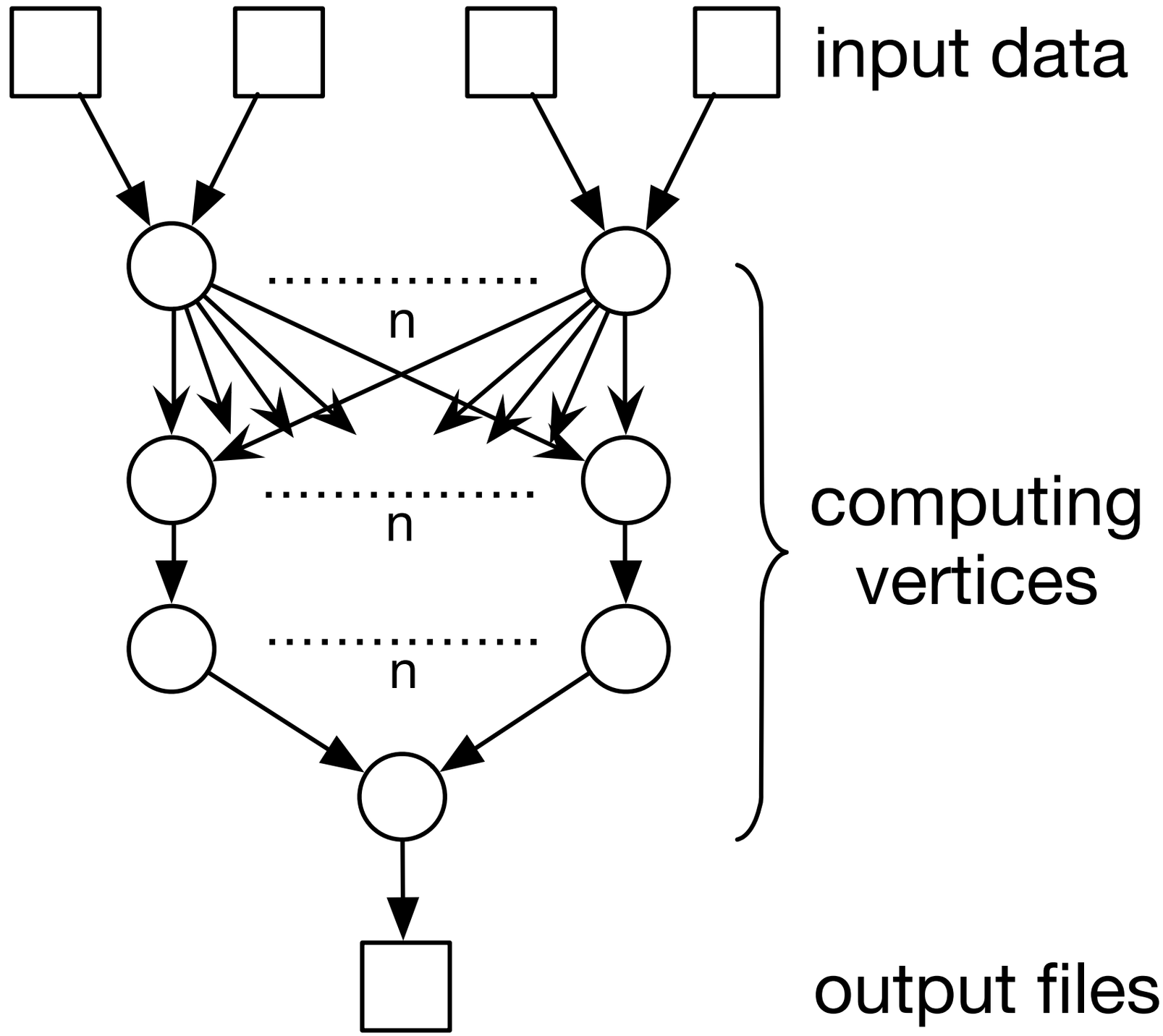
Directed Acyclic Graph (DAG)

Spark

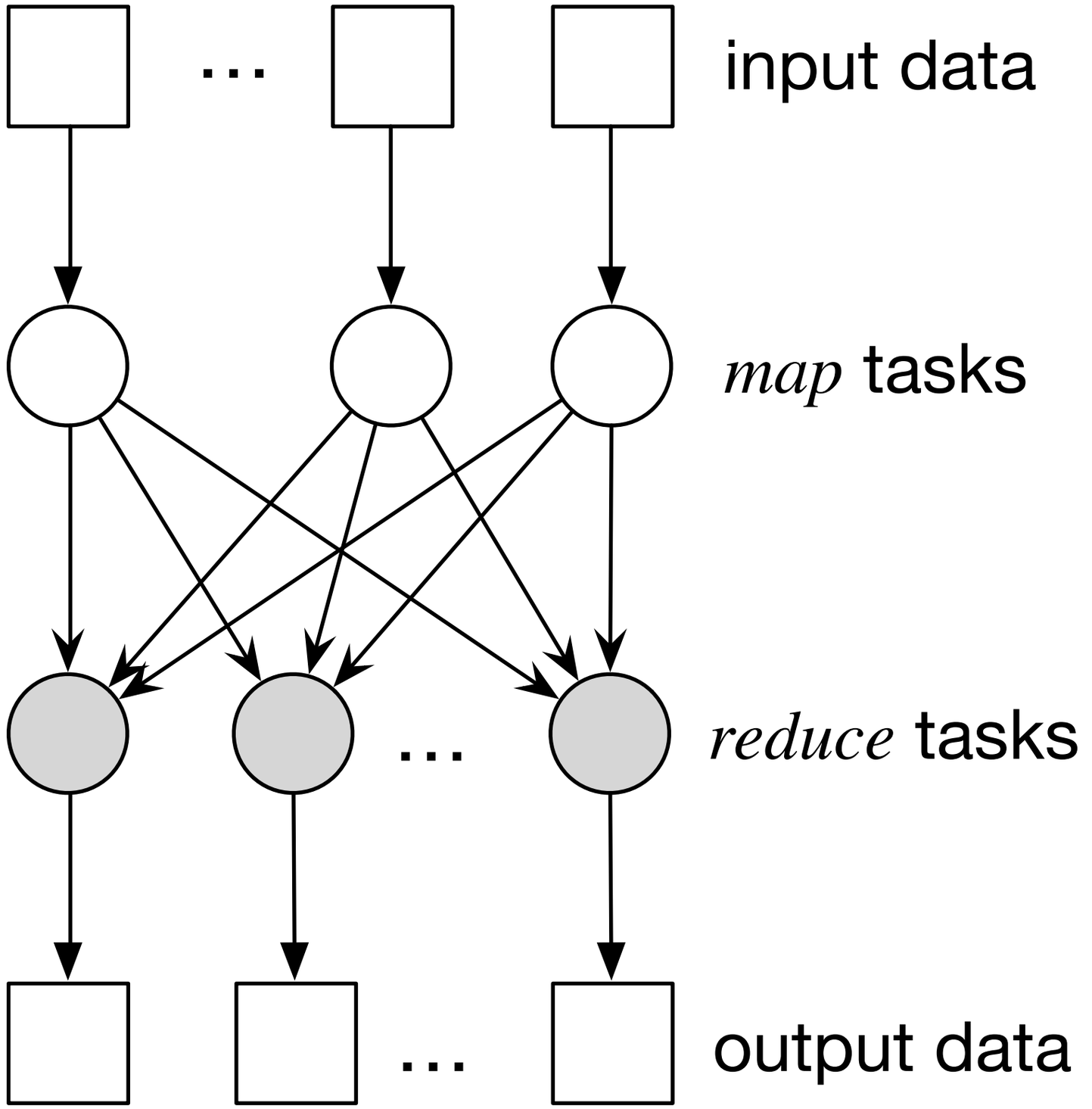


Dryad

Microsoft

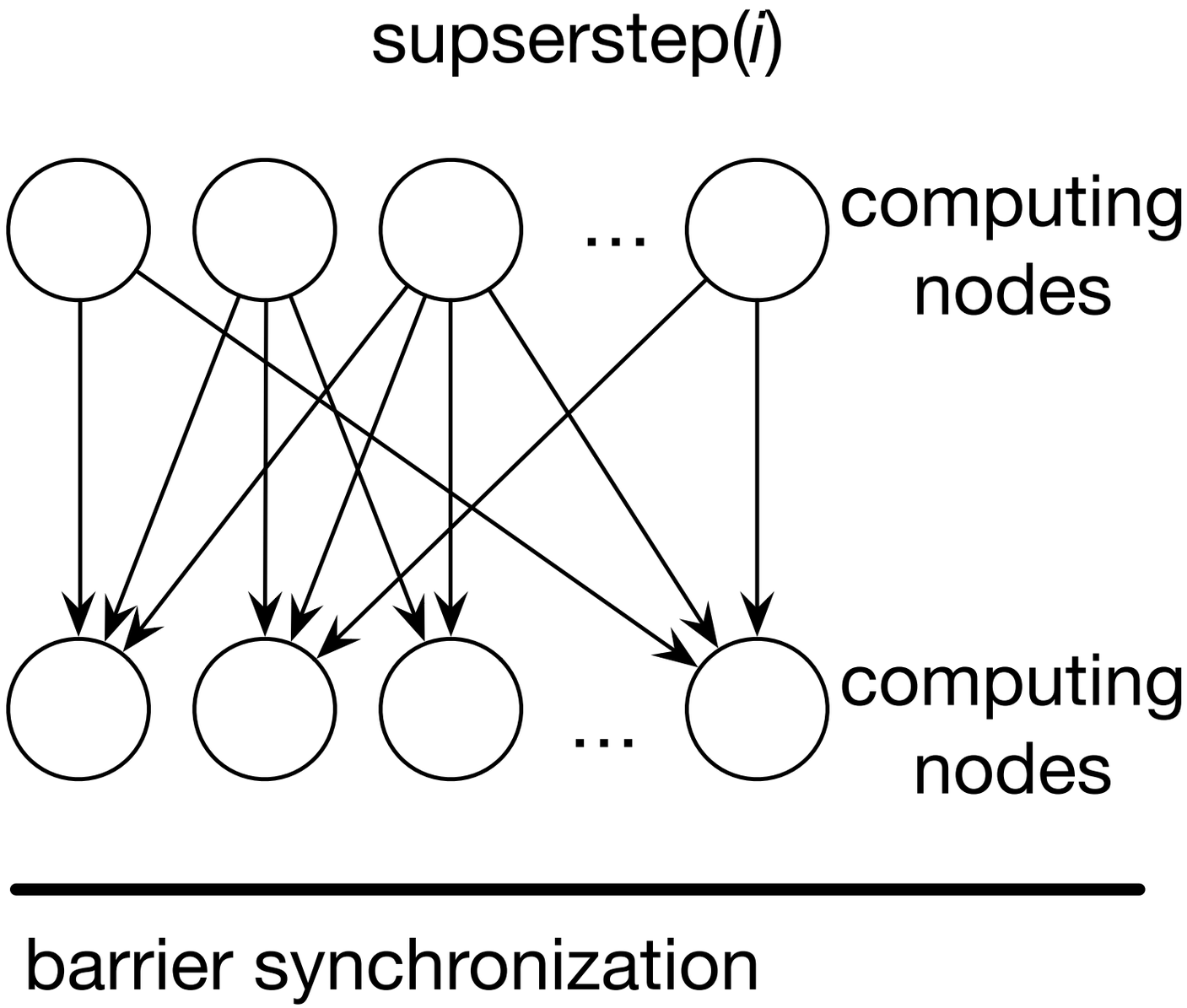


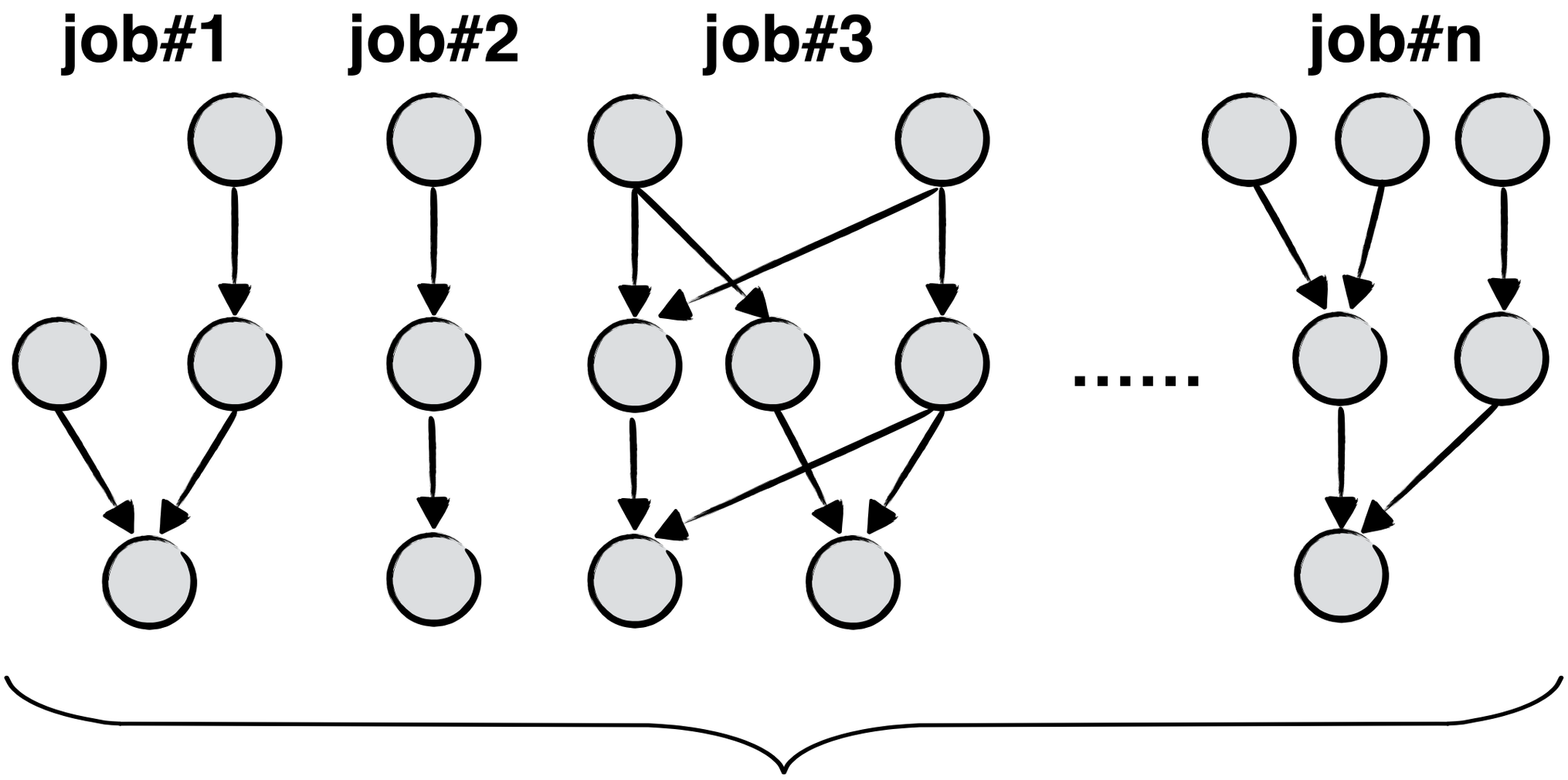

hadoop

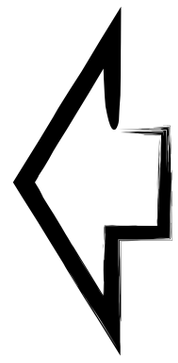

input data
map tasks
reduce tasks
output data

(BSP Model)

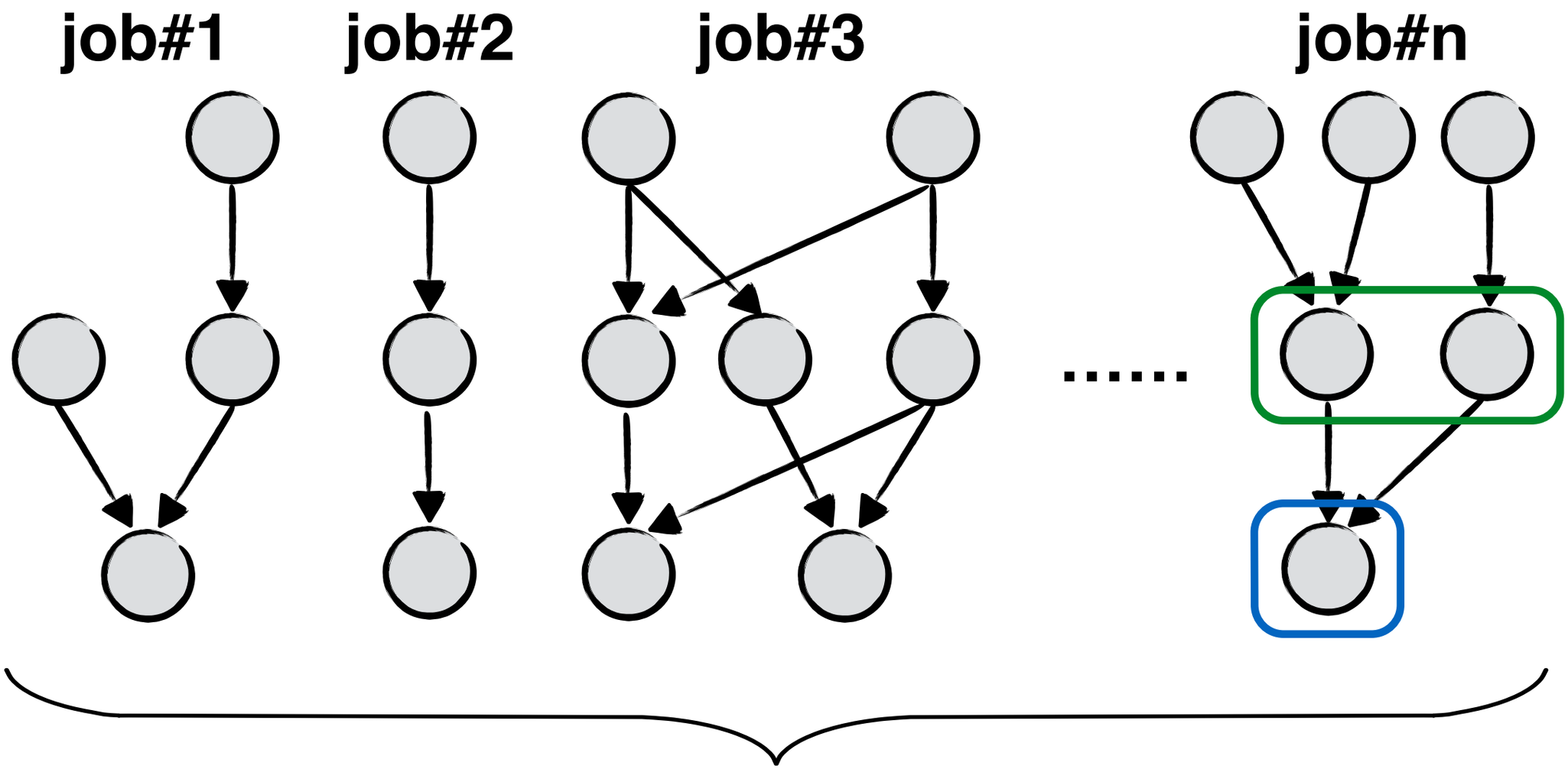

APACHE
GIRAPH

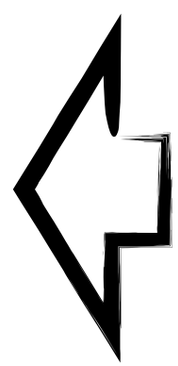

supserstep(i)
...
computing nodes
...
computing nodes
barrier synchronization



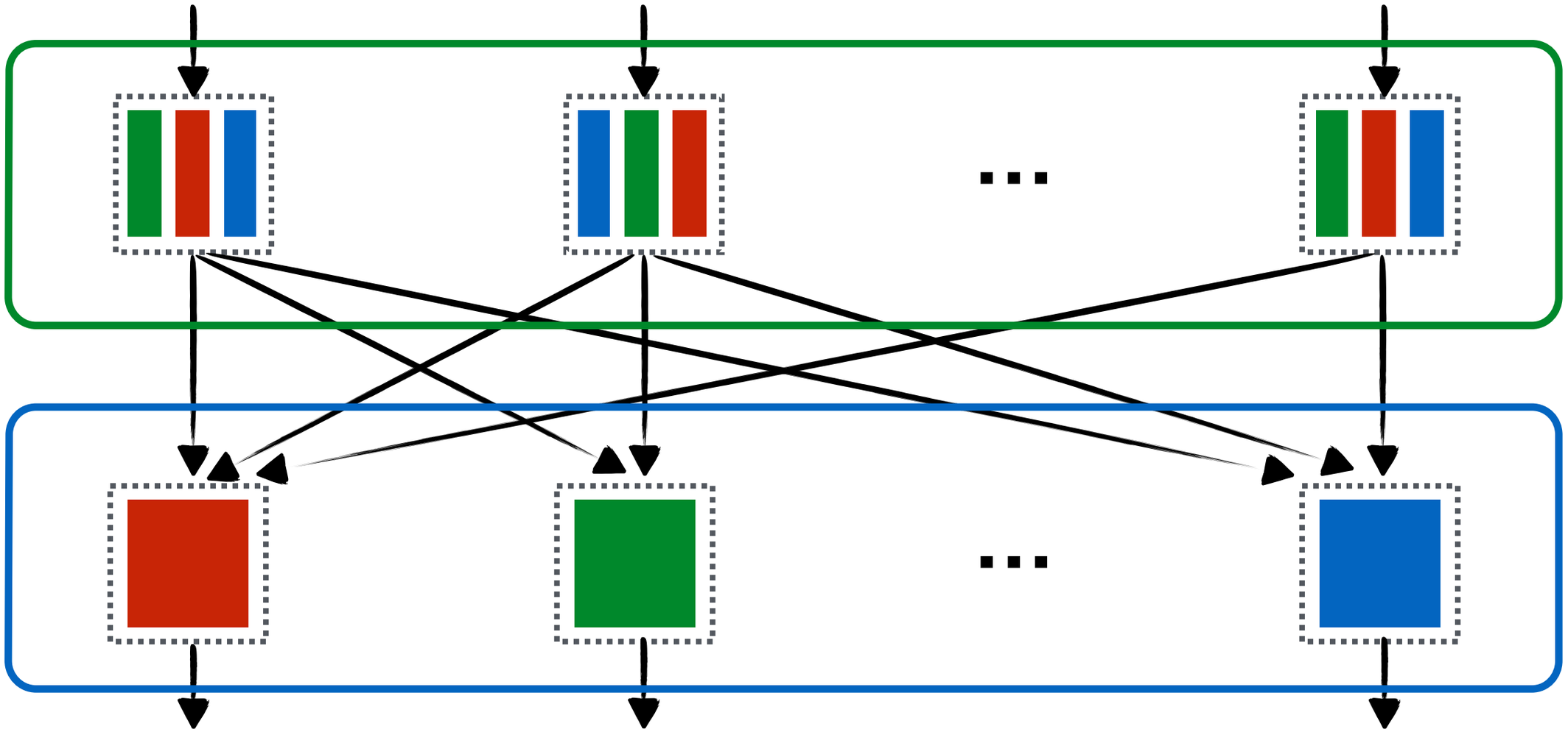


Application Submit





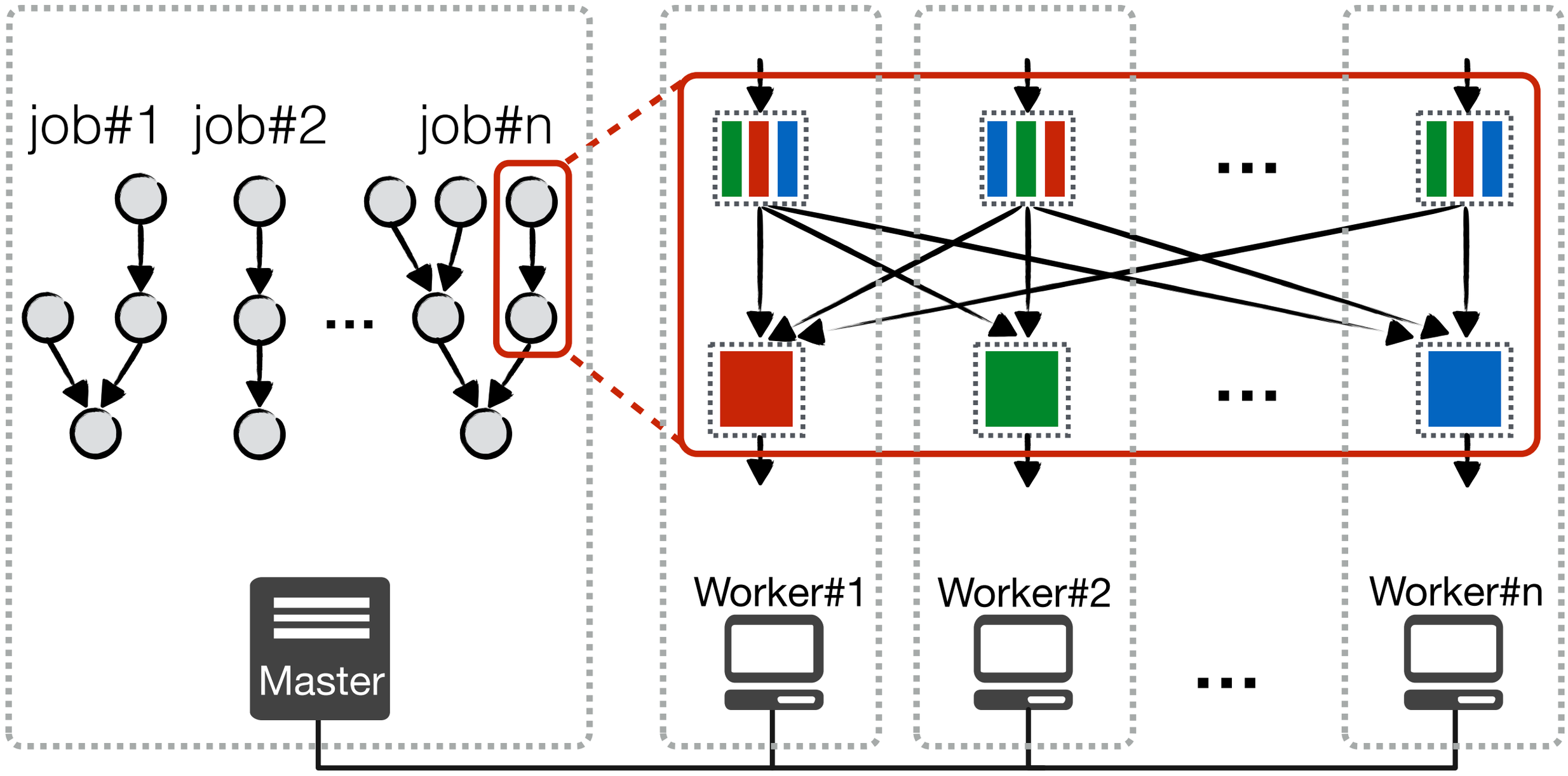
Application Submit



Task Assignment

Worker#1 Worker#2 ... Worker#n

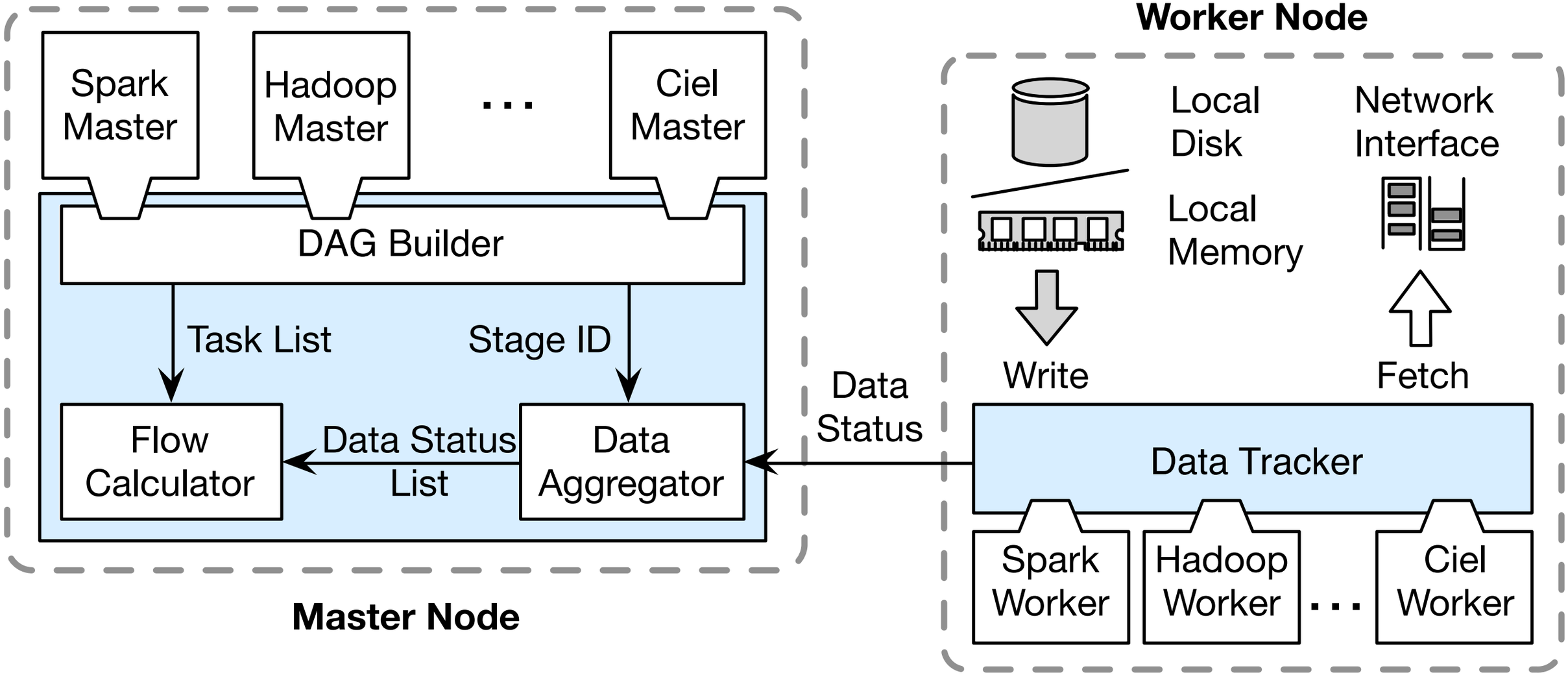
Life Cycle



OBSERVATION

*— DAG contains necessary **time**, **data**, and **flow dependencies** for accurate flow prediction.*

ARCHITECTURE



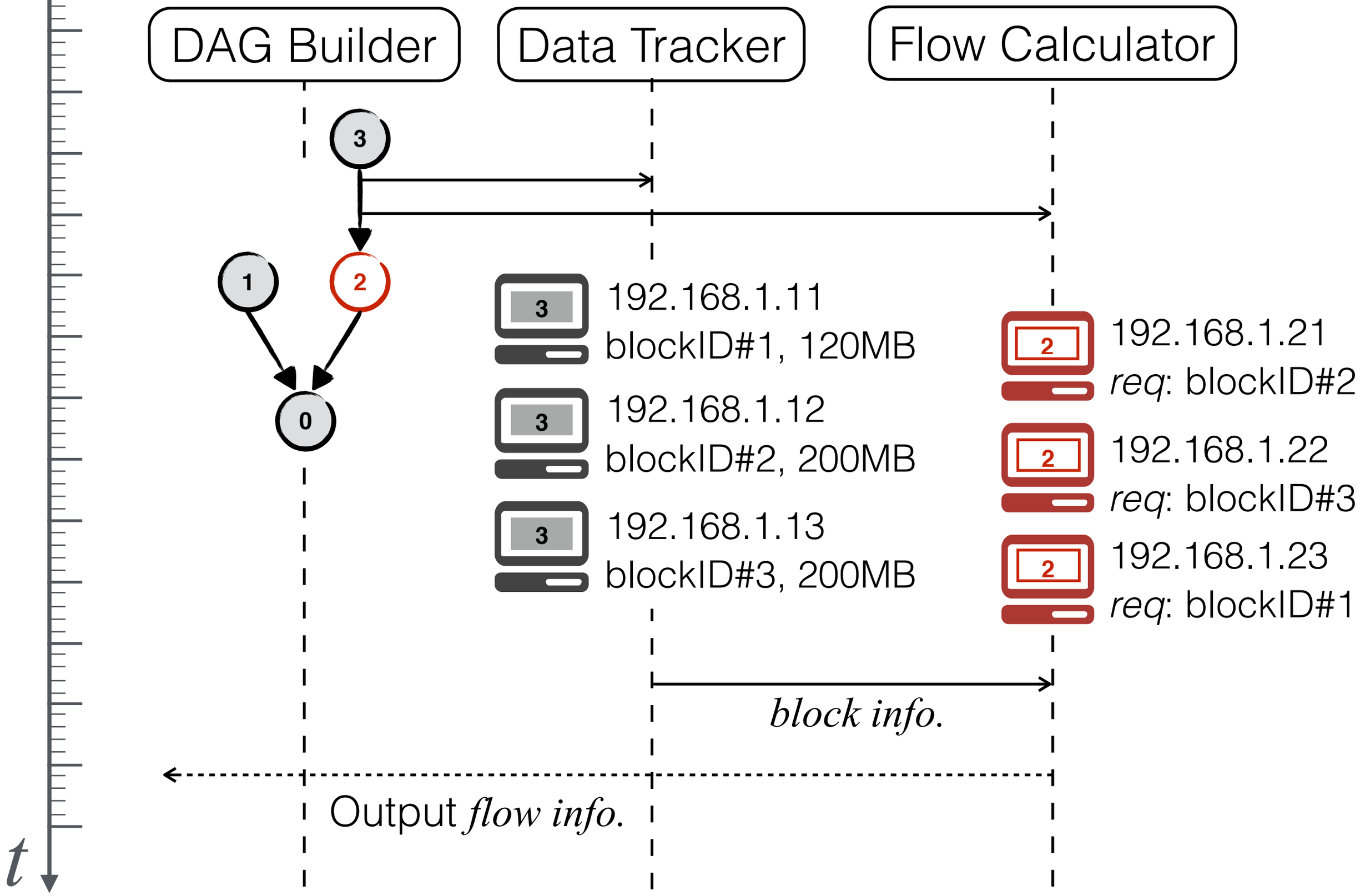
API EXAMPLES

- Required APIs for DCF master

Event Definition	Trigger Condition
`newStageEvent(stageID, childStageID)`	a new stage is created
`stageStartEvent(List[task], stageID)`	a stage is beginning
`stageFinishedEvent(stageID)`	a stage is finished

- The DAG Builder event handlers

Event Definition
`newStageHandler(newStageEvent)` ⇒ `(currentStage, childStage)`
`stageStartHandler(stageStartEvent)` ⇒ `Event(List[task], List[stageID])`
`stageFinishedHandler(stageFinishedEvent)` ⇒ `Event(stageID)`


DAG Builder
Data Tracker
Flow Calculator
3
1
2
0
3
192.168.1.11
blockID#1, 120MB
3
192.168.1.12
blockID#2, 200MB
3
192.168.1.13
blockID#3, 200MB
2
192.168.1.21
req: blockID#2
2
192.168.1.22
req: blockID#3
2
192.168.1.23
req: blockID#1
block info.
Output flow info.
t

FLOWPROPHET

- Generic
- Accurate and fined-grained
- Ahead-of-time
- Scalable and low-overhead

TESTBED

- Dell PowerEdge R320 x 37
- Intel Xeons E5-1410 2.8GHz CPU
- 24GB 1600MHz DDR3
- Broadcom Gigabit Ethernet NIC
- Pronto-3295 Gigabit Ethernet Switch

BENCHMARKS

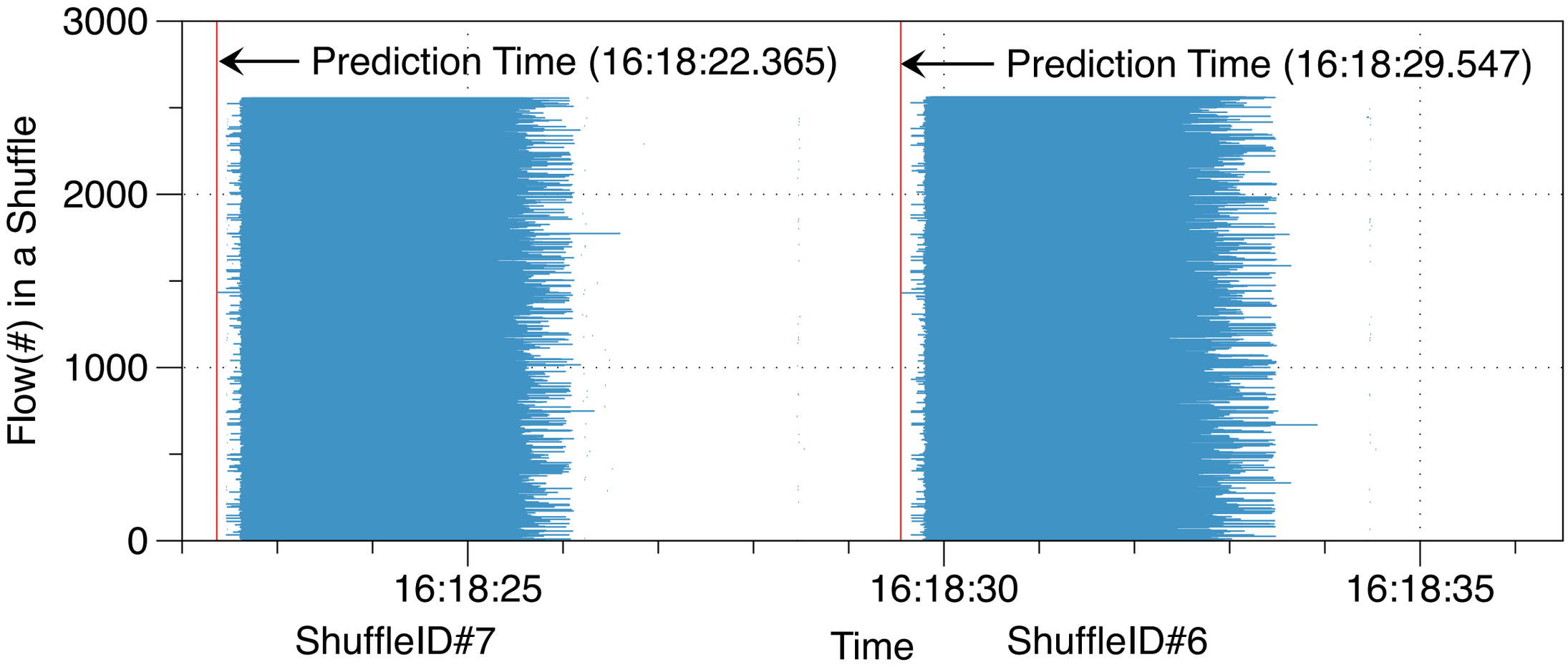
- WikiPageRank
- SparkPageRank
- Spark K-means
- Hadoop TeraSort
- π (Pi)
- WordCount

METRICS

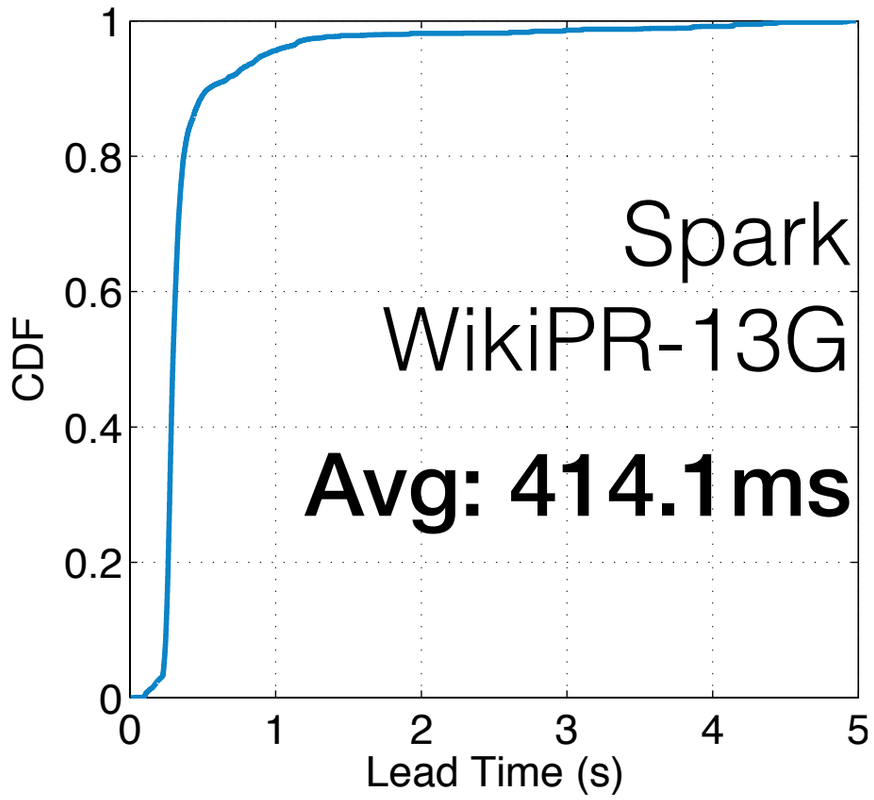
- Time advance
- Prediction accuracy
- Overhead
- Scalability
- Benefits

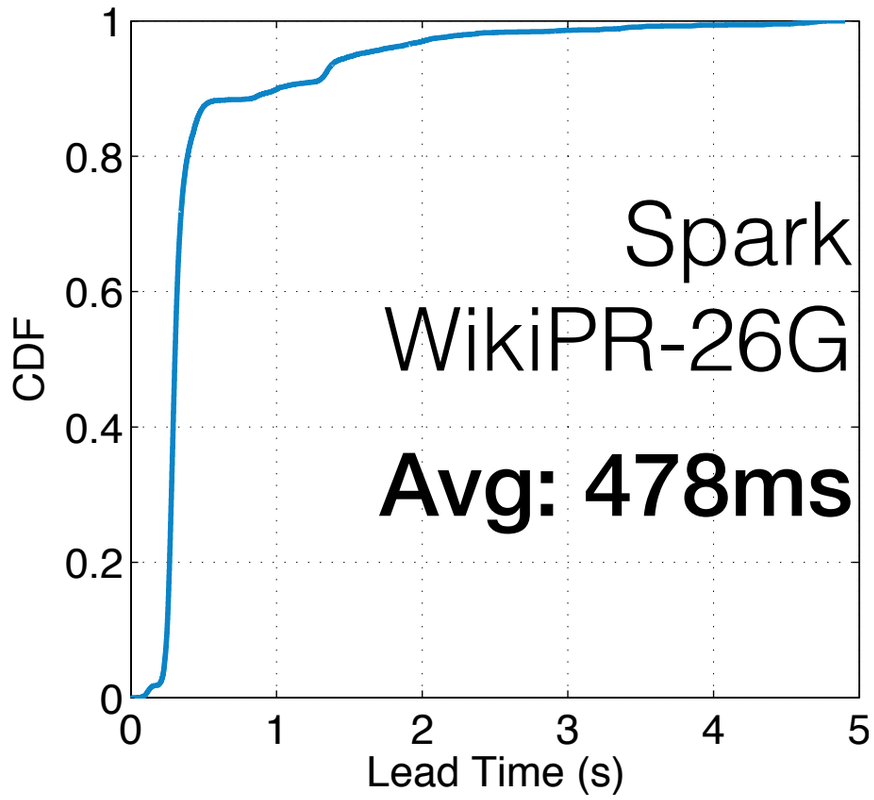
Time Advance

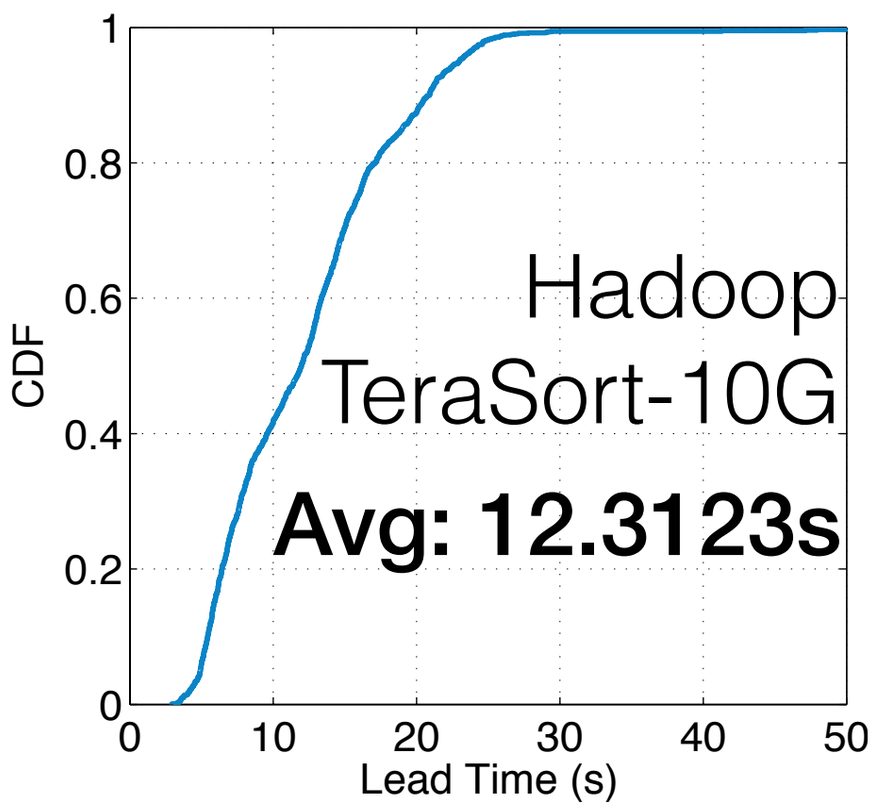
- WikipediaPageRank-13G (Spark)

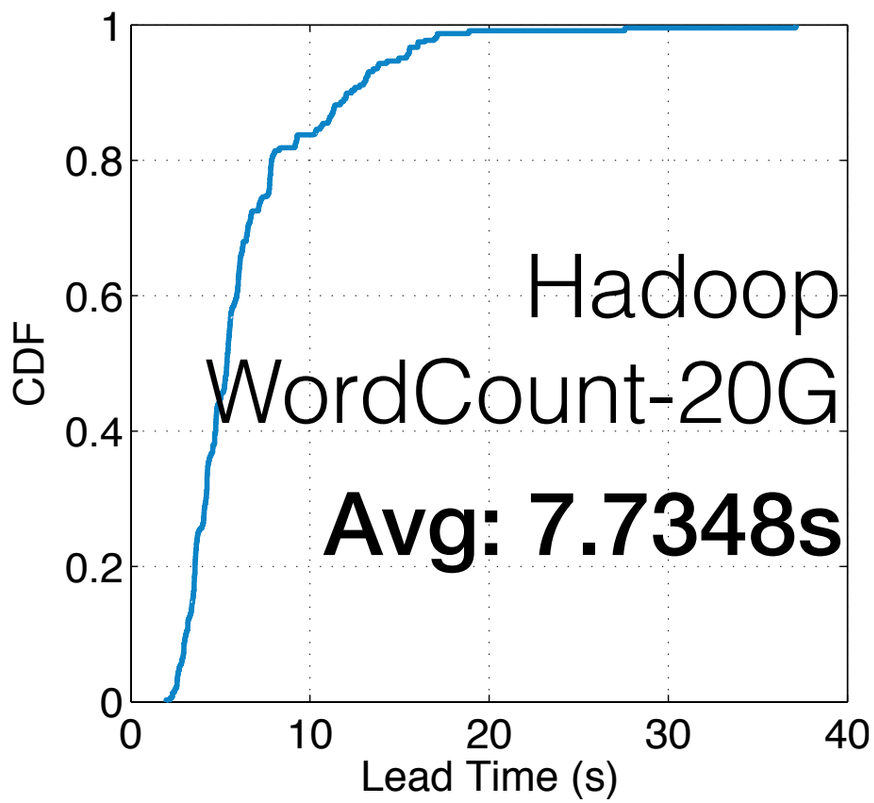


CDF of Lead Time

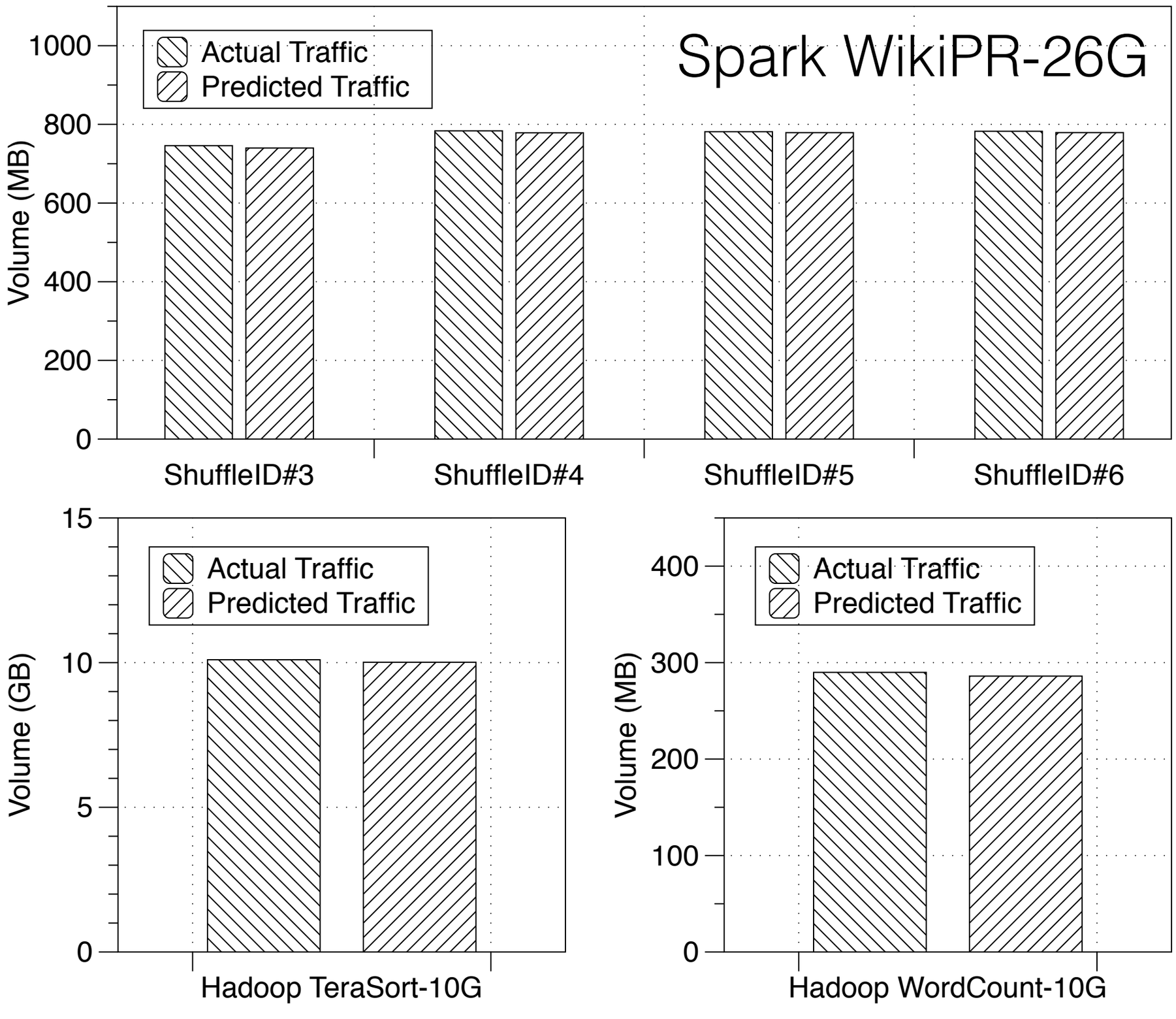




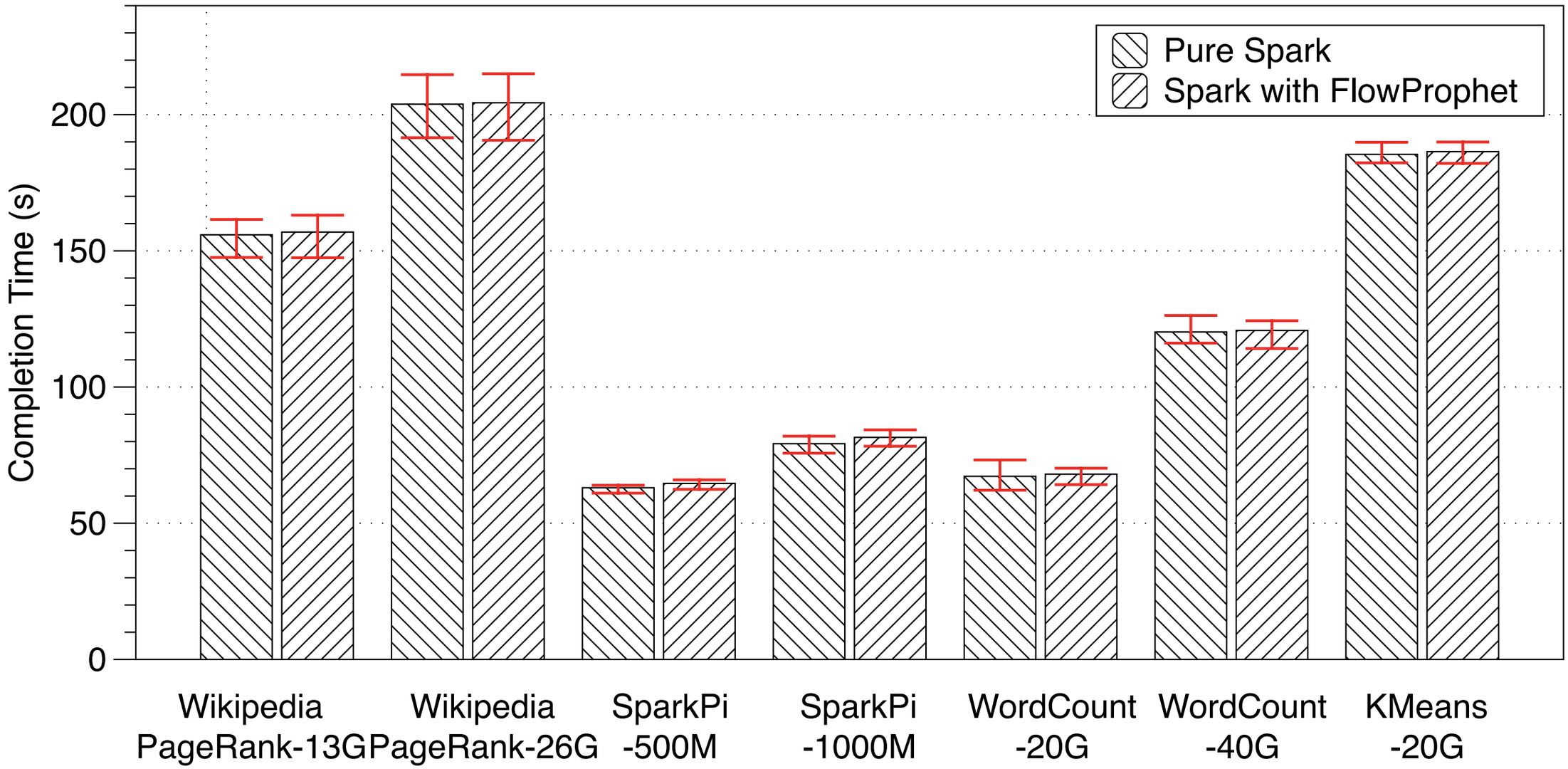




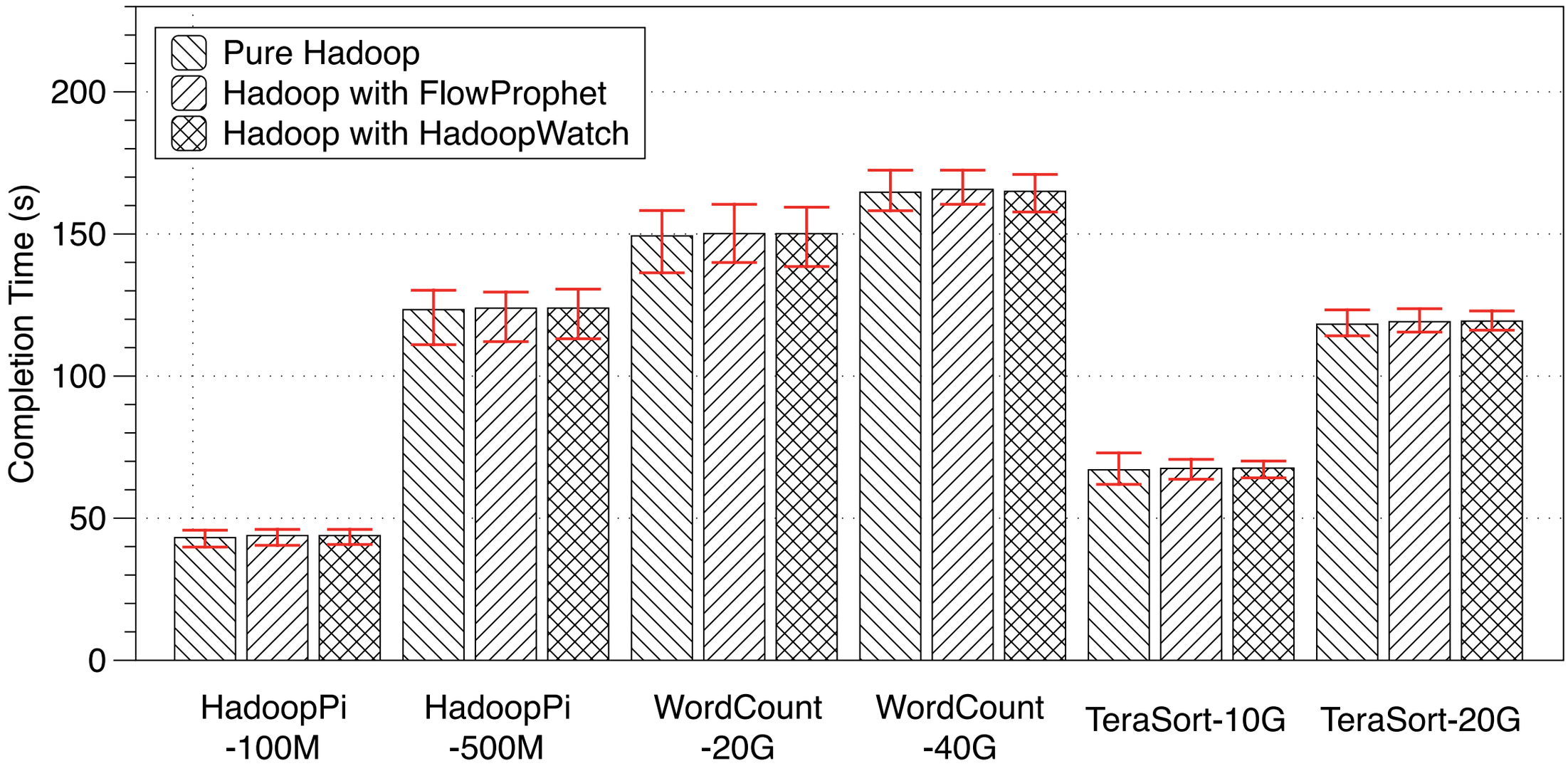
Prediction Accuracy



OVERHEAD

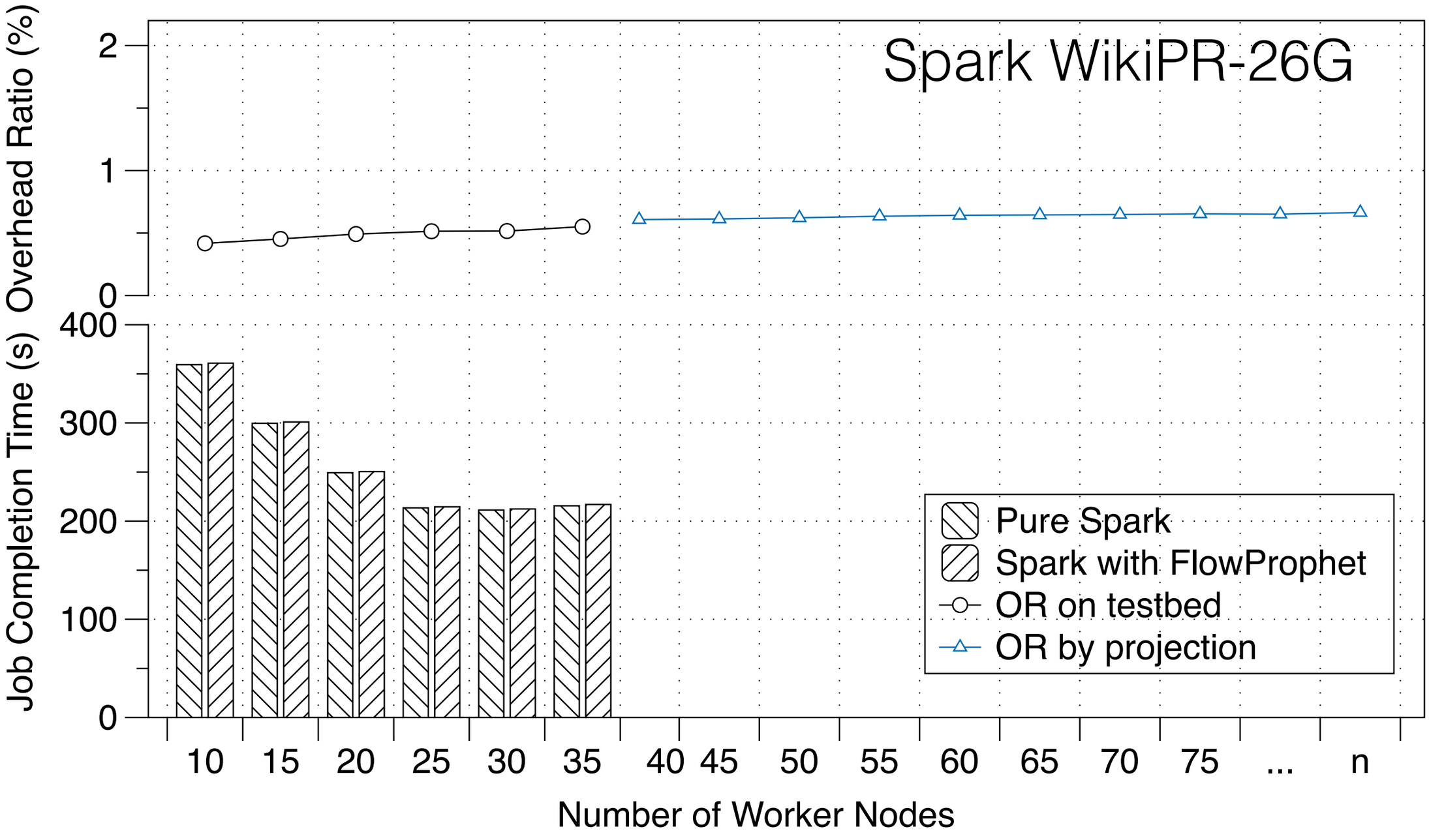

Pure Spark
Spark with FlowProphet
Completion Time (s)
0
50
100
150
200
Wikipedia PageRank-13G
Wikipedia PageRank-26G
SparkPi -500M
SparkPi -1000M
WordCount -20G
WordCount -40G
KMeans -20G

OVERHEAD

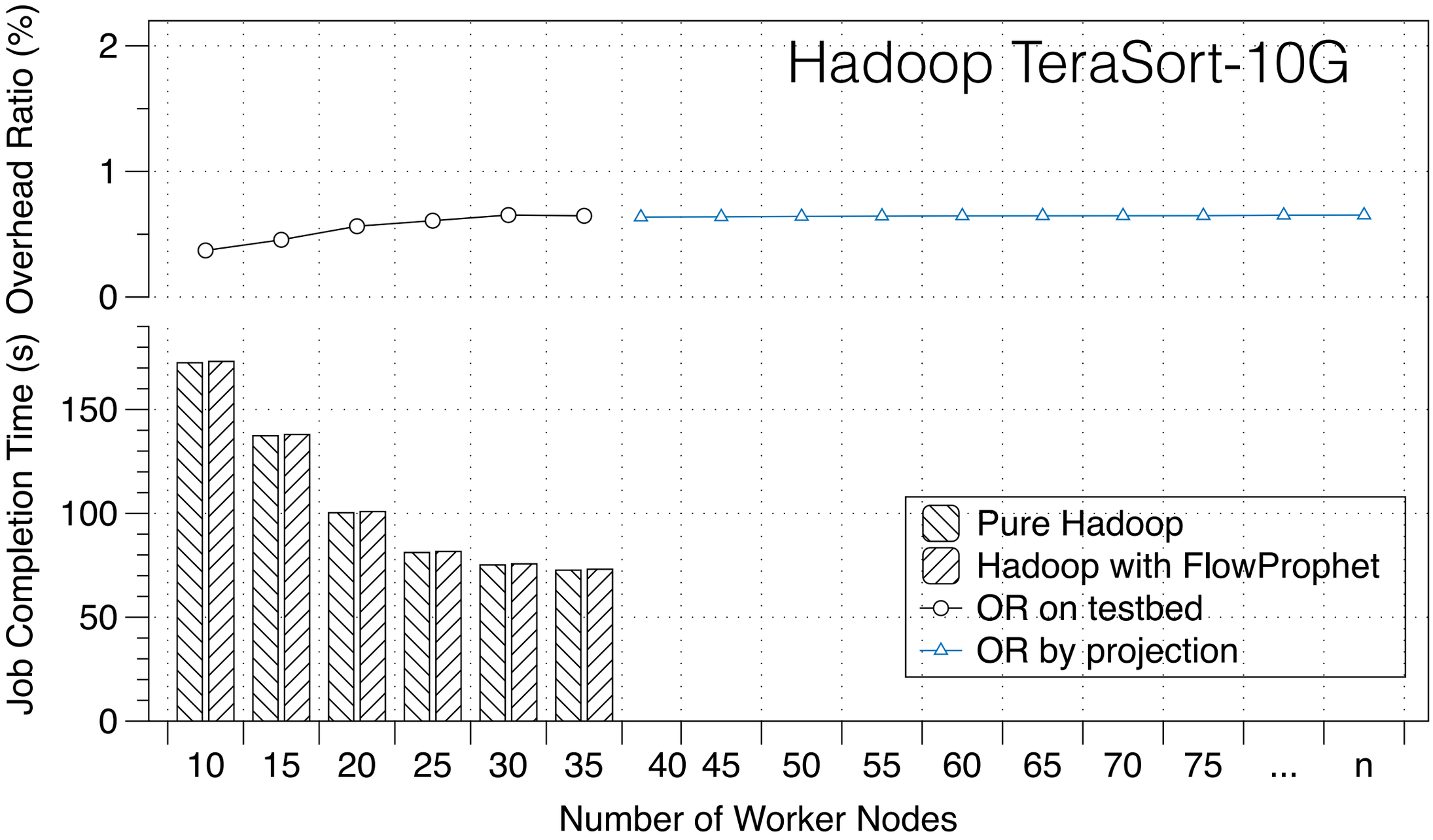

Pure Hadoop
Hadoop with FlowProphet
Hadoop with HadoopWatch
Completion Time (s)
200
150
100
50
0
HadoopPi -100M
HadoopPi -500M
WordCount -20G
WordCount -40G
TeraSort-10G
TeraSort-20G

SCALABILITY

- Overhead Ratio (OR) : $$OR = \frac{t_{enabled} - t_{disabled}}{t_{disabled}}$$

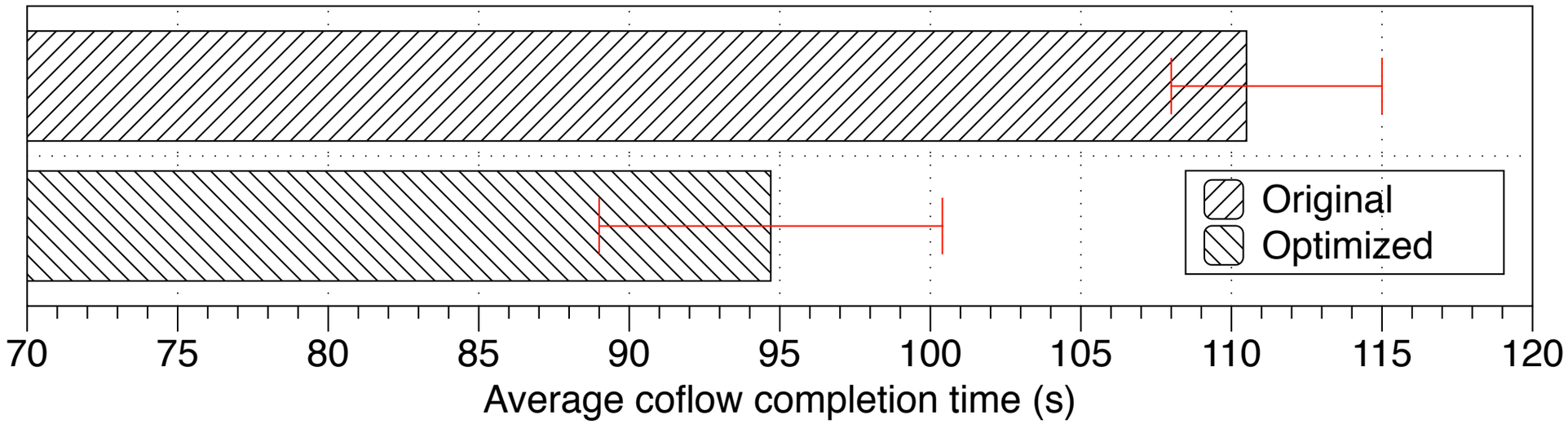


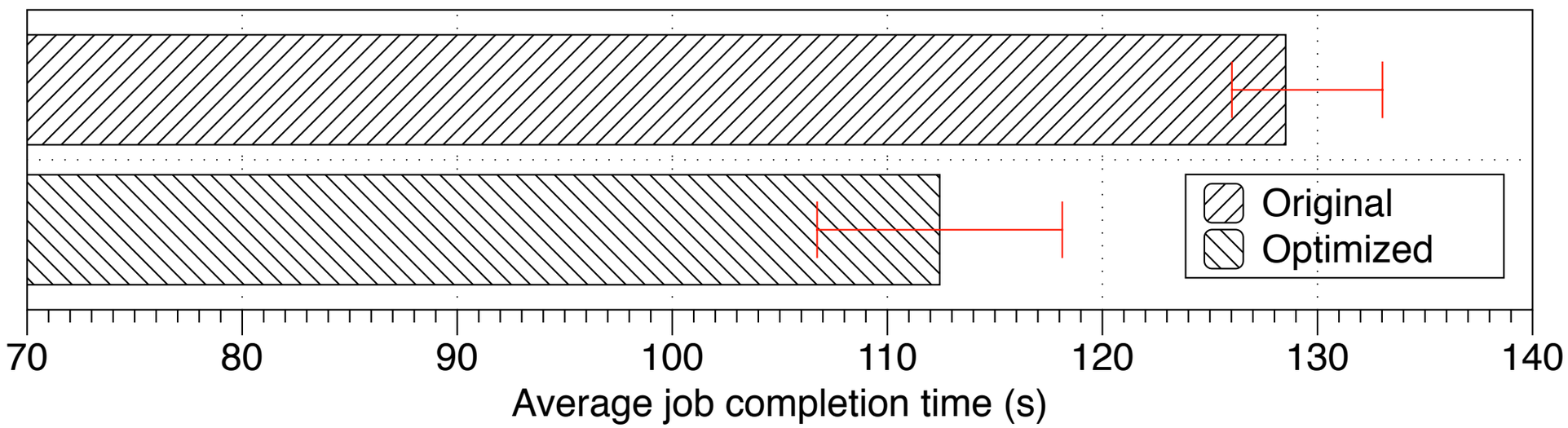
SCALABILITY



BENEFITS

- Hadoop TeraSort-25G
- 12.52% JCT reduction by a simple network scheduler





Related Work

- Analyze past statistics
  - Traffic Engineering with Estimated Traffic Matrices
- Monitor buffers or counters in switches
  - c-Through, Hedera, Helios
- Tracing and profiling toolkits
  - X-Trace
- File system monitoring
  - HadoopWatch

SUMMARY

- DCF execution pattern
- DAG for predicting flows
- Design and implementation
- Evaluation on testbed

Thank you | Q&A

ICDCS'15, Columbus, USA