A New CodeGen Module for InferSpark

InferSpark: A Probabilistic Programming Framework for Big Data

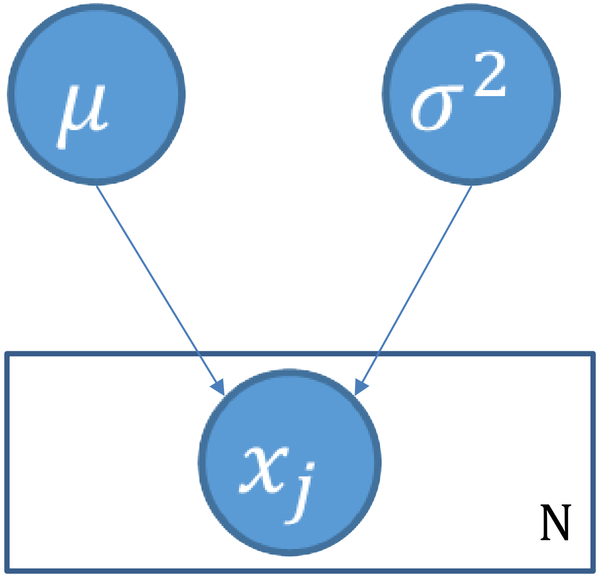
赵卓越

致远学院

指导教师：朱其立

Probabilistic Programming (PP)

- Bayesian Network expressed using PL
- Inference handled by compiler/interpreter



Bayesian Network

```
model NModel(val N:Int) {
  val mu = normal(0, 1)
  val va = invgamma(1, 1)
  val x = for (i <- 1 to N)
    yield normal(mu, va)
}
```

Probabilistic Program

Existing PP Frameworks

Infer .NET (MSR Cambridge)

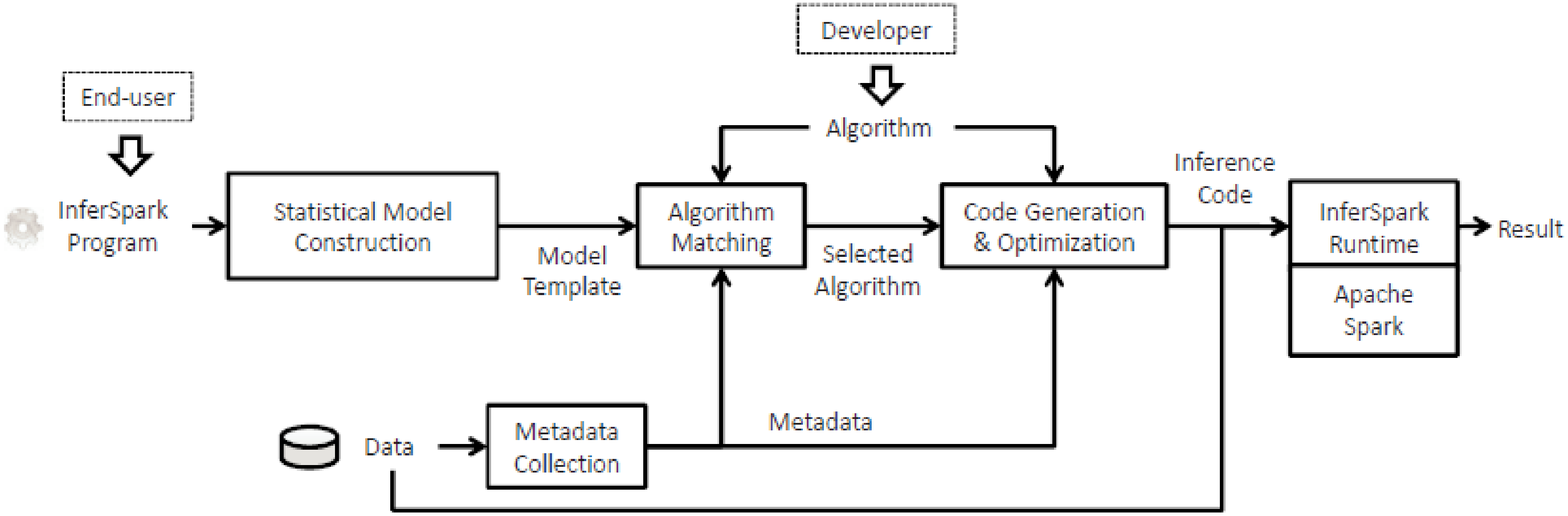
Church (MIT)

- Figaro
- BUGS
- …

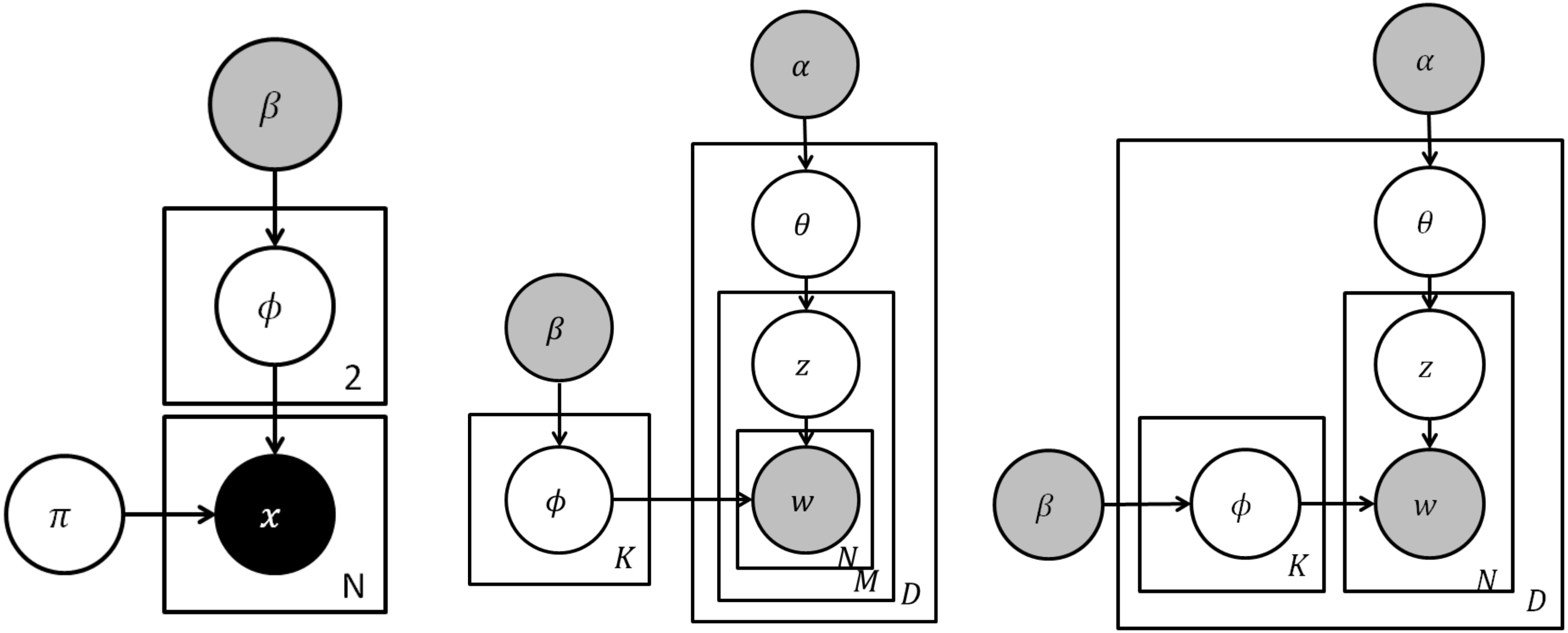
✘ Single machine => Cannot scale to large dataset

InferSpark

- PP on Apache Spark
  - – In-memory MapReduce
- Implement message-passing-style inference
  - – Using GraphX – the built-in graph library



Support Customized Models

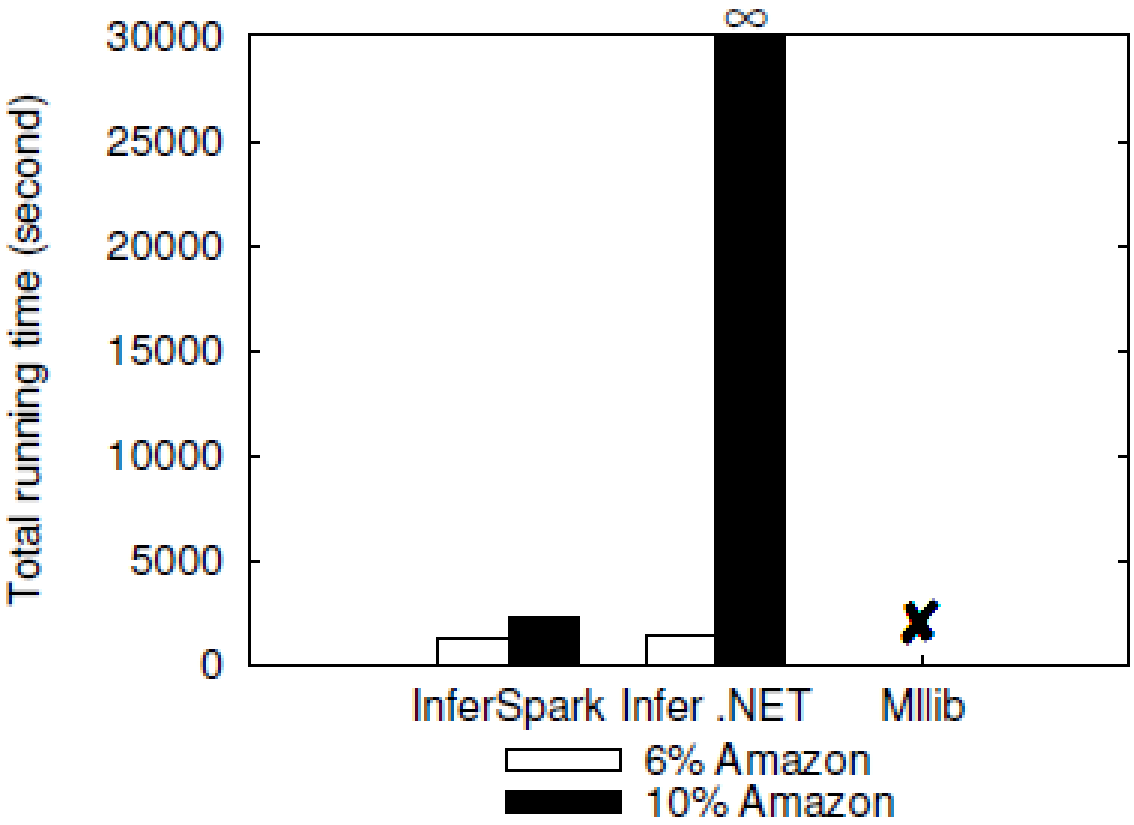


(a) Two-Coin

(b) SLDA

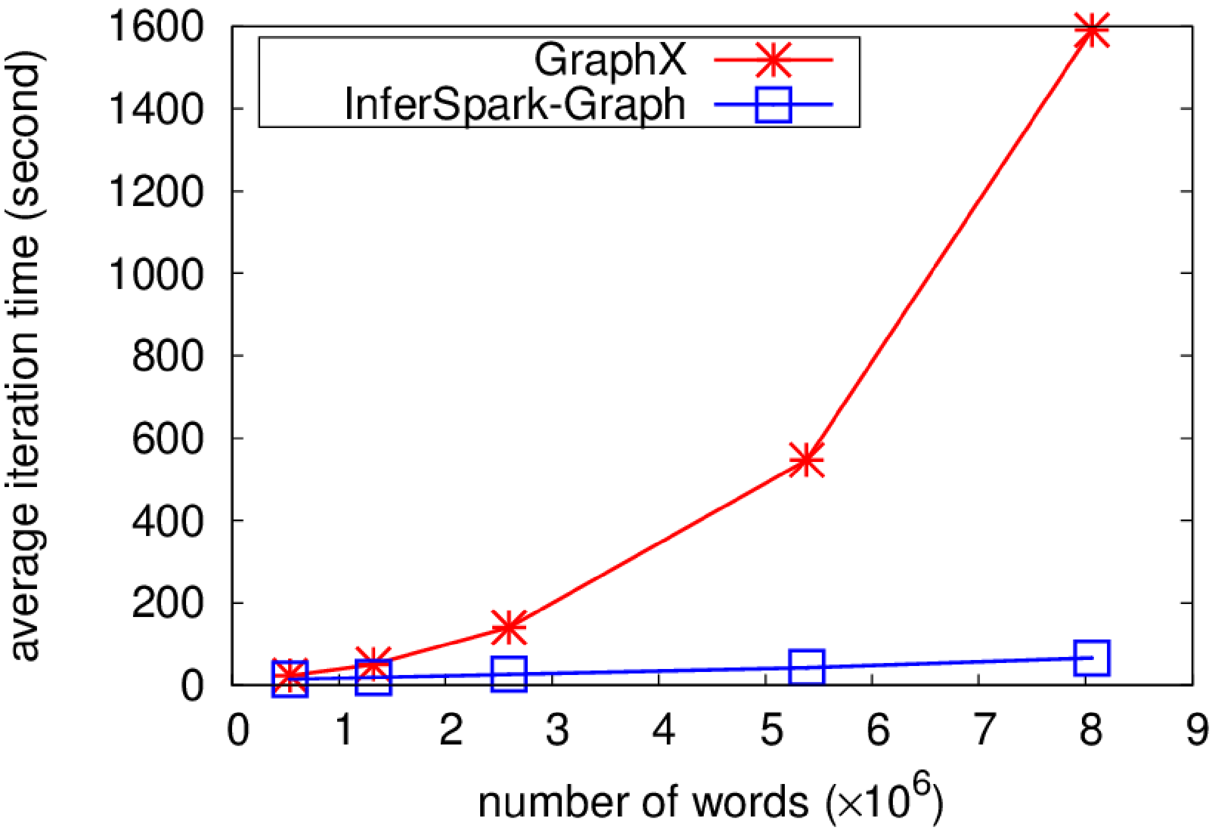
(c) DCMLDA

InferSpark



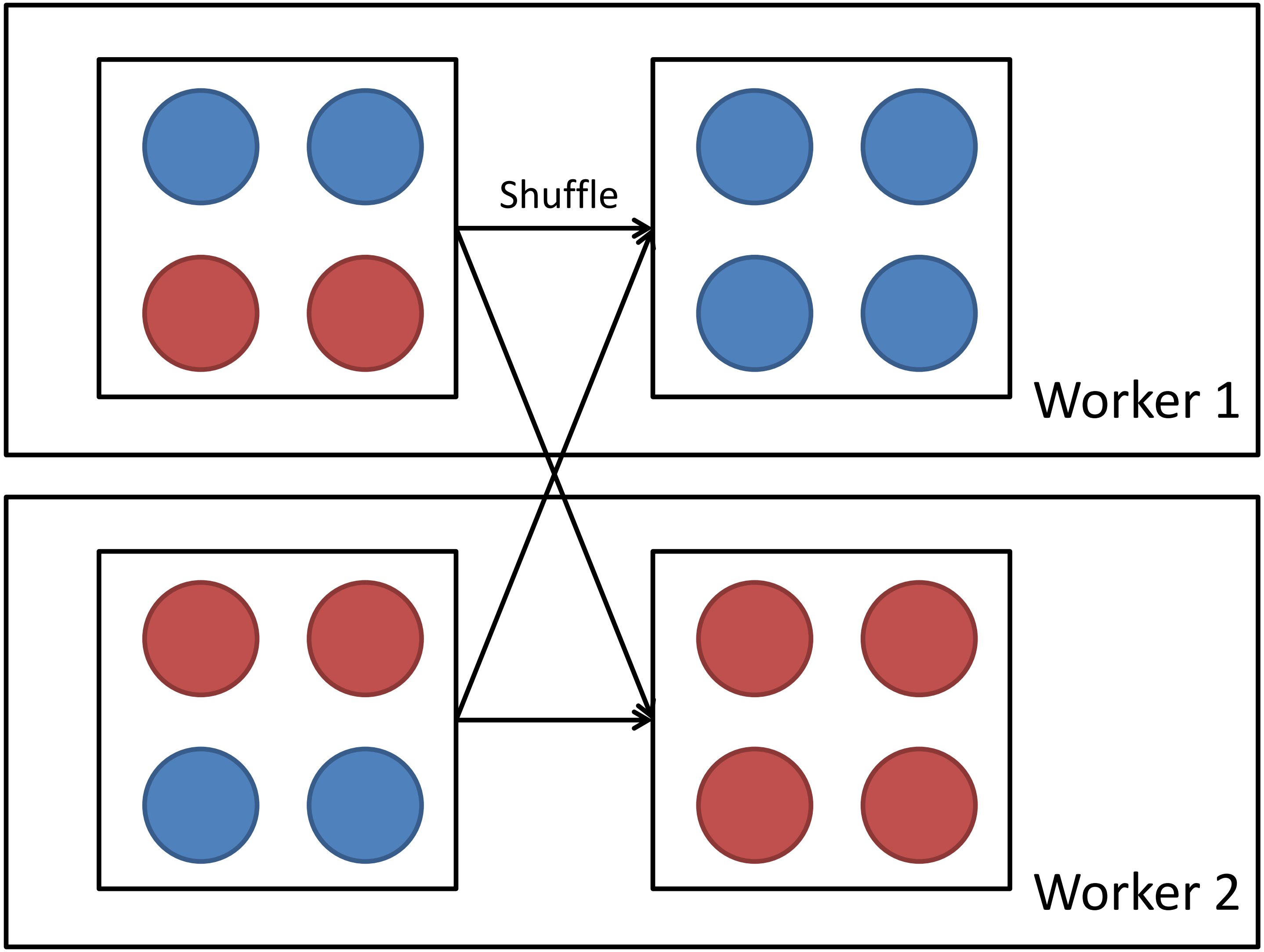
- Scales to large dataset

Problems with CodeGen

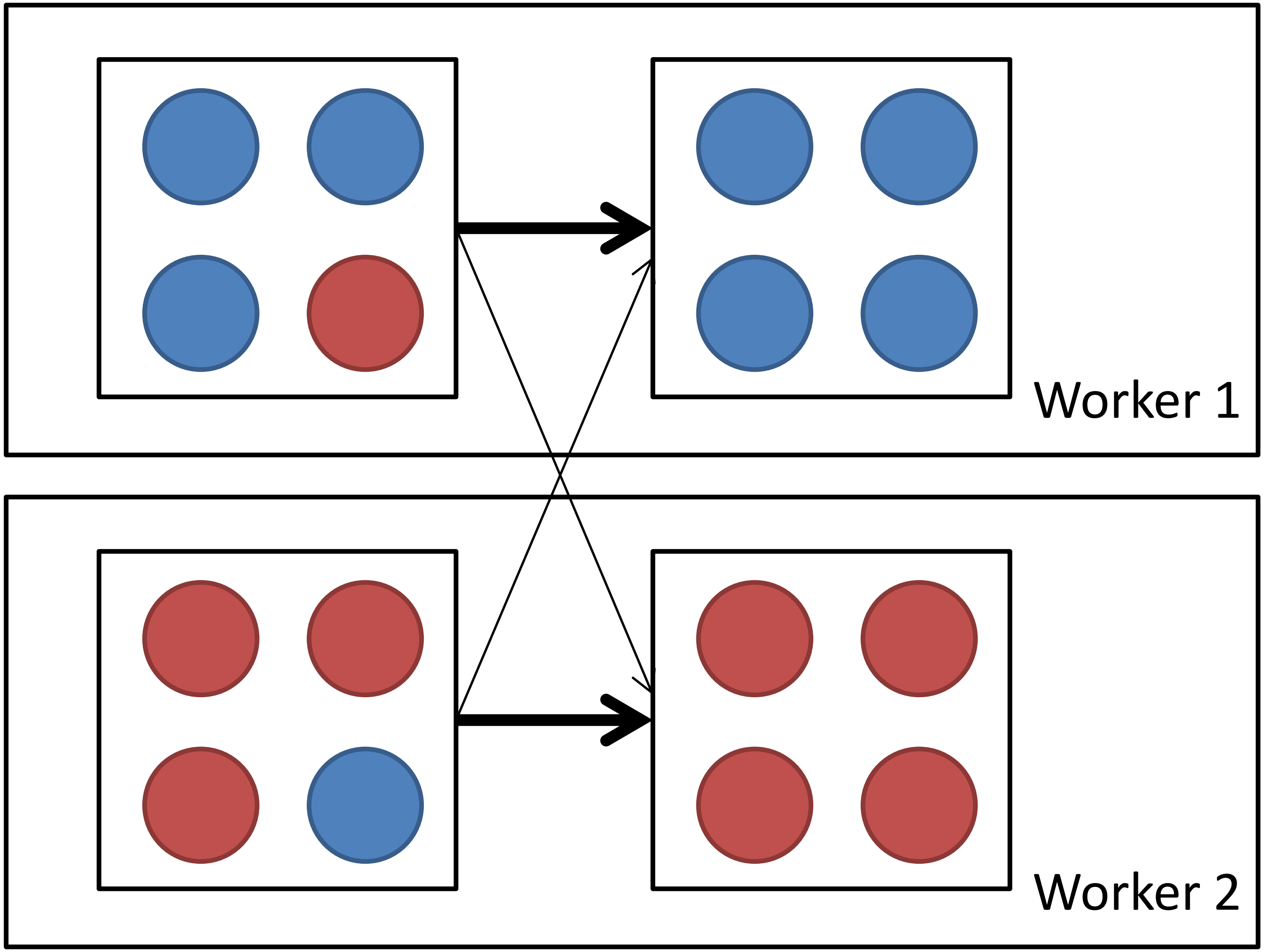


- Steep scalability curve
  - A shuffle bottleneck due to GraphX physical design
  - Shuffle performance bounded by I/O
    (disk-bound in most cases, or network-bound)

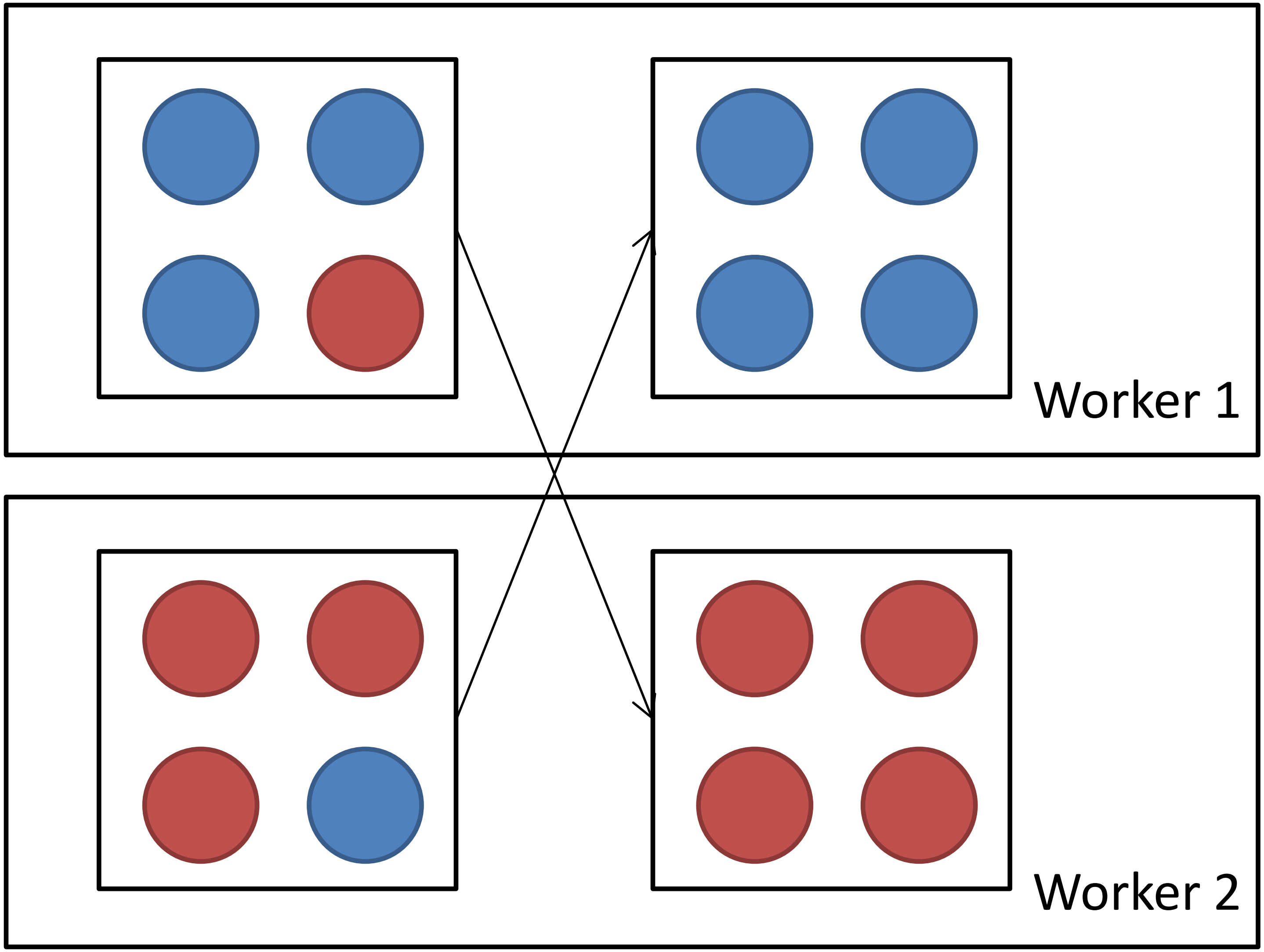
Data Shuffle



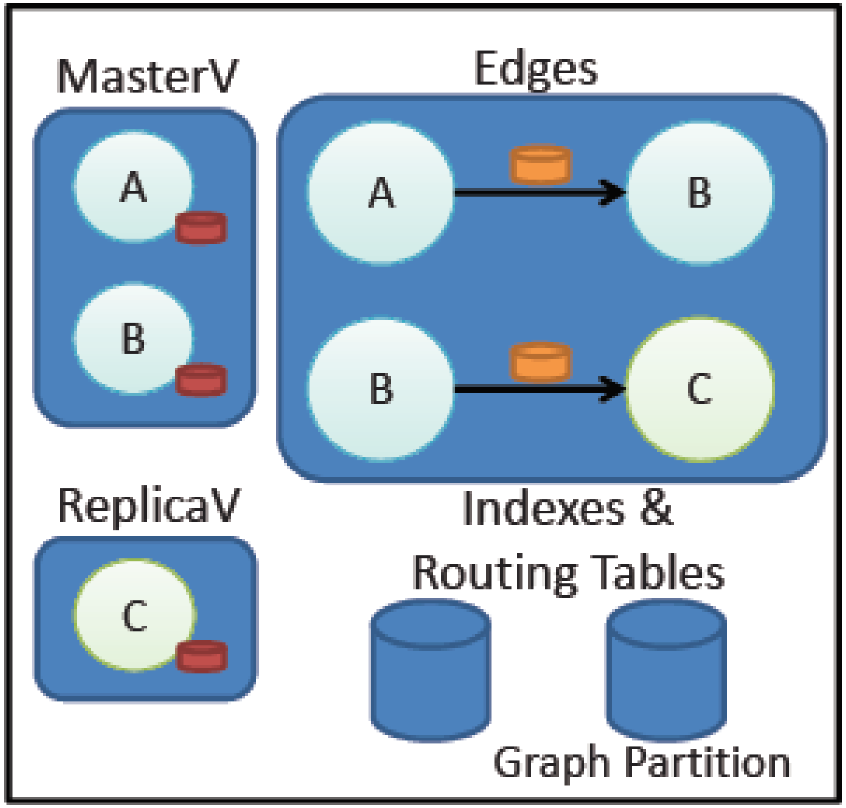
Good Partition Strategy



InferSpark-Graph

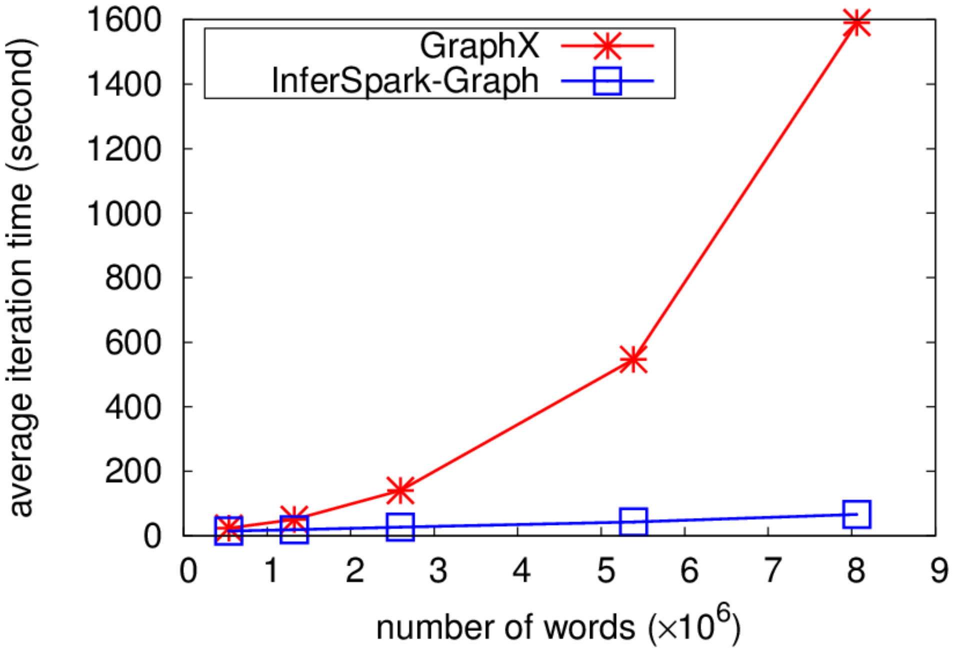


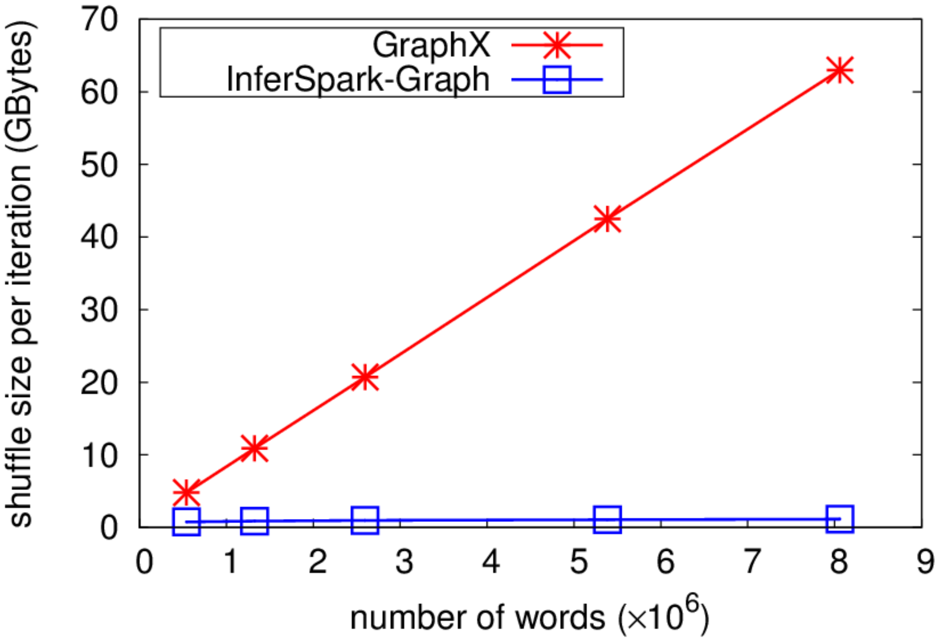
InferSpark-Graph Physical Design



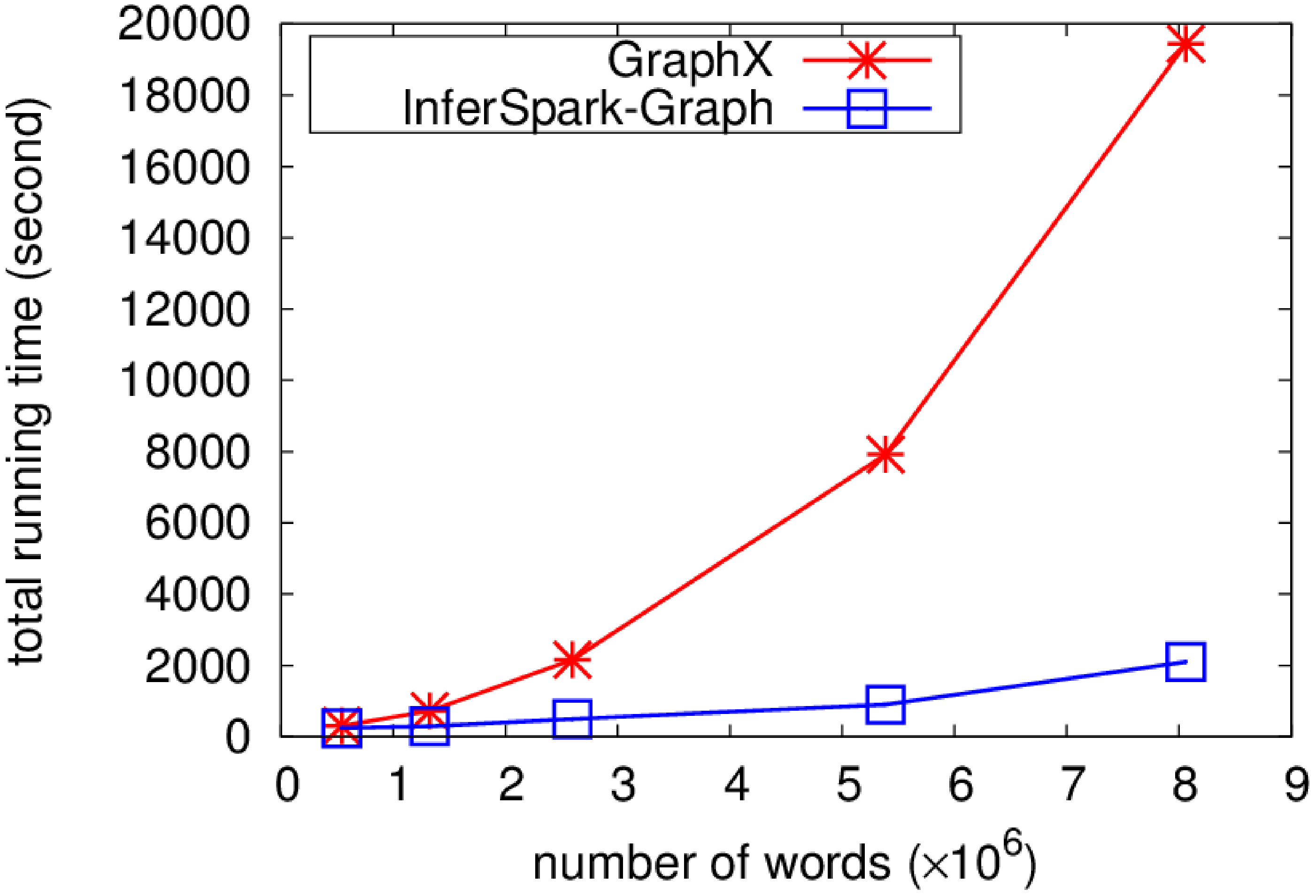
- Merge Vertex RDD and Edge RDD
- One fewer partition to shuffle
  - Which contains the majority of the data in InferSpark

Evaluation: Per Iteration





Evaluation: Total Time



Conclusion

- We designed
  - – InferSpark (~ 10380 lines of code)
    - A highly scalable probabilistic programming framework
  - – InferSpark-Graph (~ 4300 lines of code)
    - A distributed graph processing library on Spark
    - Replace GraphX in the CodeGen module of InferSpark
    - Greatly improves applications like InferSpark
- In submission to SIGMOD 2017