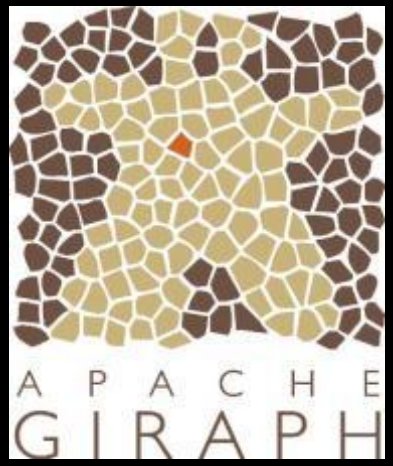


Apache Giraph

Joy Arulraj

CMU 15-799 : Lightning Talk

What is Giraph ?

- Iterative graph processing system
  - Powers Facebook Graph Search
- Highly scalable
  - Based on Bulk Synchronous Parallel (BSP) model (1990)

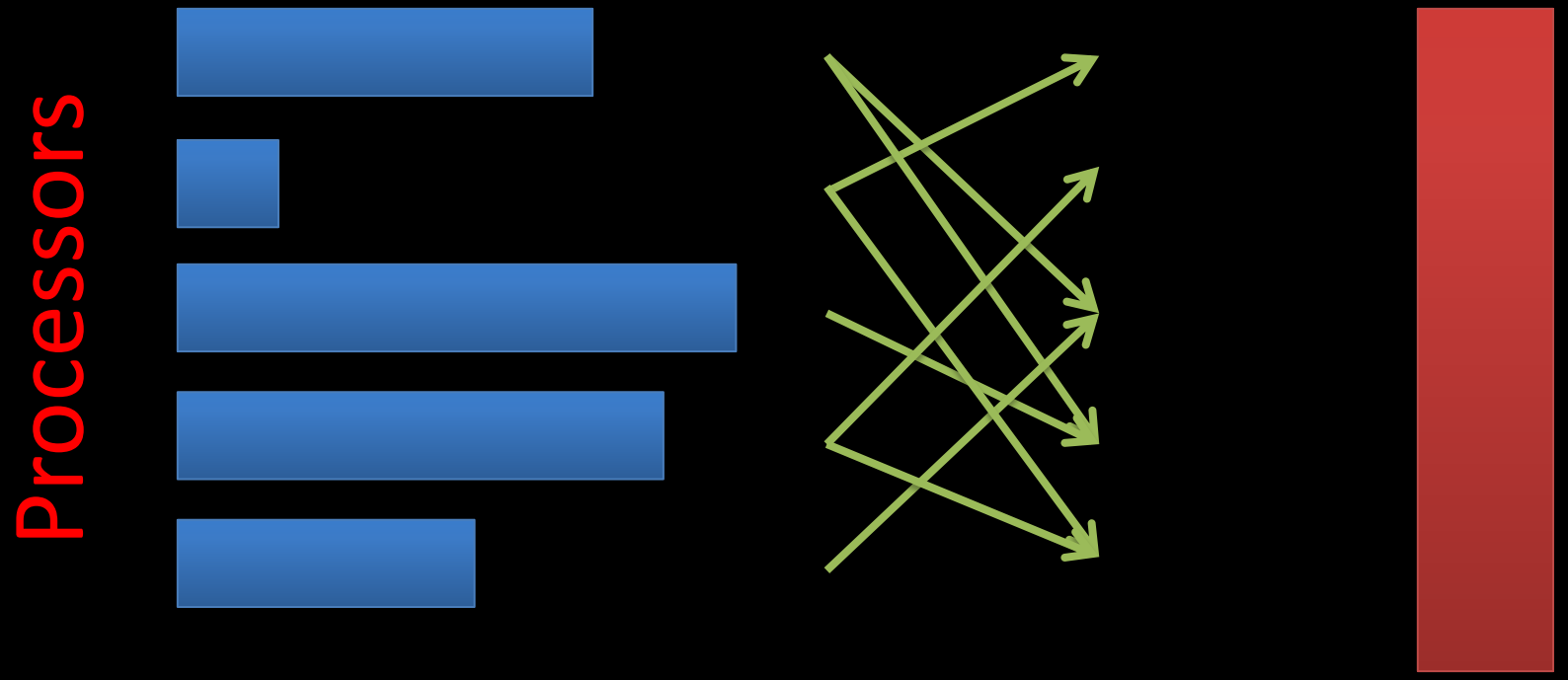
How is it different?

- Map-Reduce
  - Not natural abstraction for graphs
  - Mostly stateless to increase parallelism
  - Giraph reuses state across iterations
- Message Passing Interface
  - No inherent fault tolerance
  - Generic interface

Big Picture

Iteration N - BSP Model

Computation Communication Barrier



Basic API

- “ Think like a vertex ”
  - Inspired by Google Pregel (2010)

```
Class Vertex{
    VertexValue& GetValue();
    SendMessageTo(string& dest_vertex,
    MessageValue msg);
    void Compute(MessageIterator* msgs);
};
```

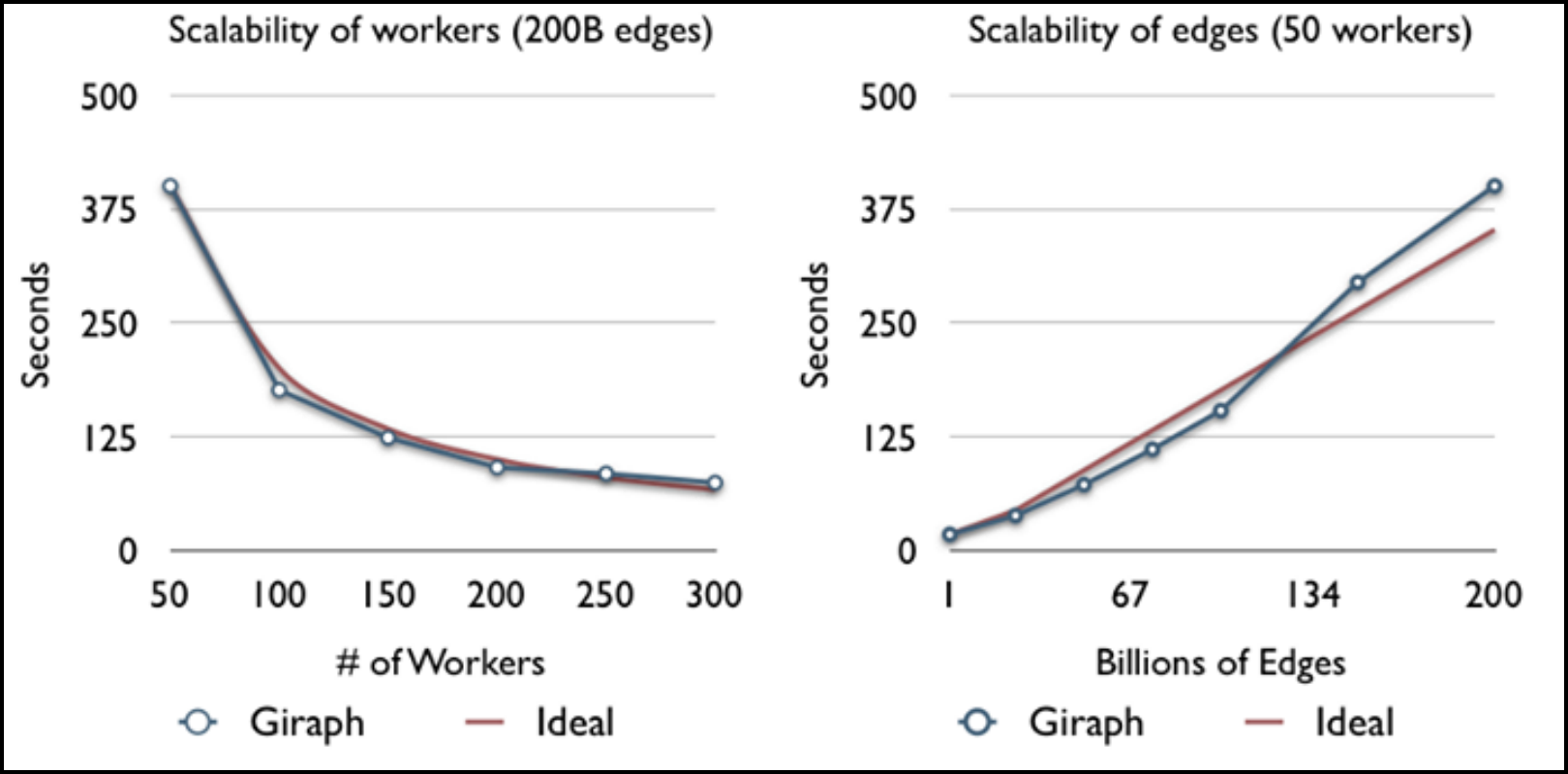
How is Giraph used ?

- Iterative graph algorithms like PageRank

```
Compute(MessageIterator* msgs) {
  if(IterationCount() < 20){
    double sum = /* Sum all msg values */
    SetValue(0.15/NumVertices() + 0.85*sum);
    SendMessageToAllNeighbors(
    GetValue()/OutEdges);
  }
}
```

Performance

- Near linear scalability
  - With # of workers OR problem size.



Try out Giraph !

- http://giraph.apache.org/
- Version 1.0.0 released in May 2013
- PageRank on 1 T edges (Facebook)
  - With 200 commodity machines
  - Takes just 4 minutes/iteration

Thanks !