Towards Optimizing Hadoop Provisioning in the Cloud

Karthik Kambatla, *Purdue University*
Abhinav Pathak, *Purdue University*
Himabindu Pucha, *IBM Research Almaden*

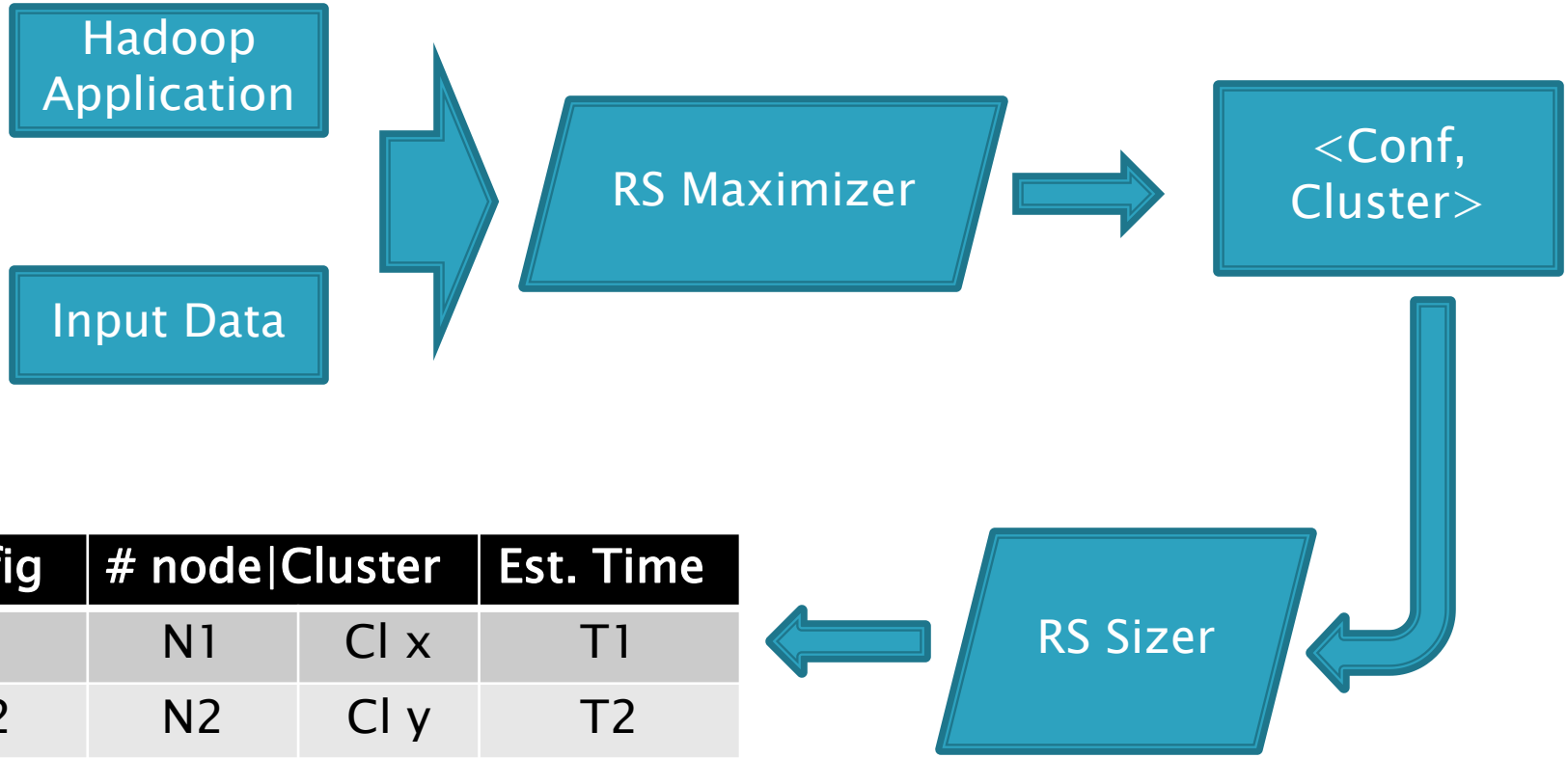
MapReduce in the Cloud

- Data analytics is important/prevalent
  - MapReduce – highly scalable solution
- Performing Hadoop-like data analytics in the cloud is particularly synergistic
  - Utility model
    - Request/Relinquish resources on demand
    - Billed by machine hours
      - Not limited by number of machines

Challenge: Hadoop Provisioning

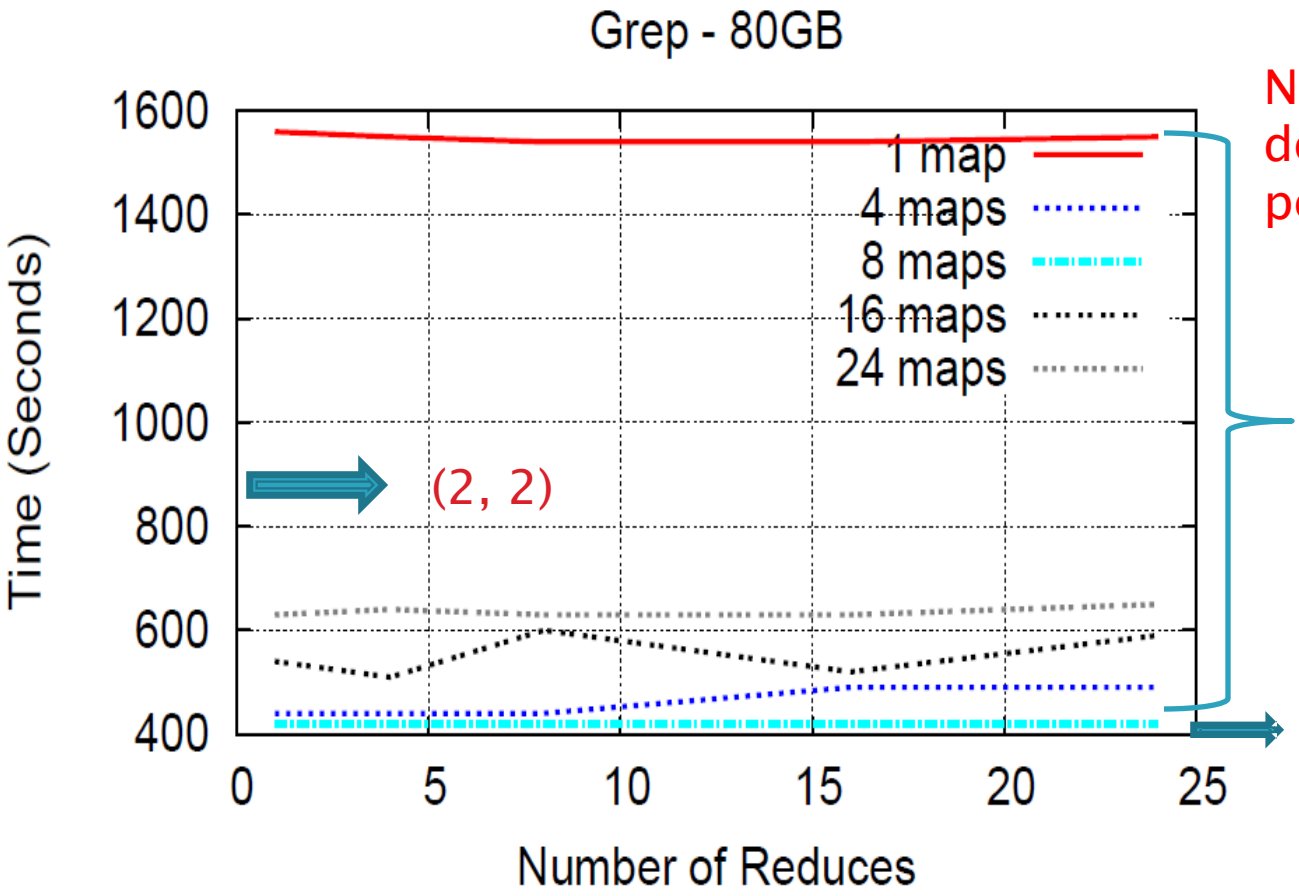
- Provisioning
  - Allocate resources
  - Configure for best utilization
- Current tools
  - Hadoop on Demand, Cloudera, etc.
  - Automate deployment, Do Not Optimize Resources!
- Our Contribution: Optimized provisioning
  - Minimize cost, Maximize Performance

Our Proposal



Config	# node|Cluster		Est. Time
C1	N1	Cl x	T1
C2	N2	Cl y	T2
C3	N3	Cl z	T3

Grep

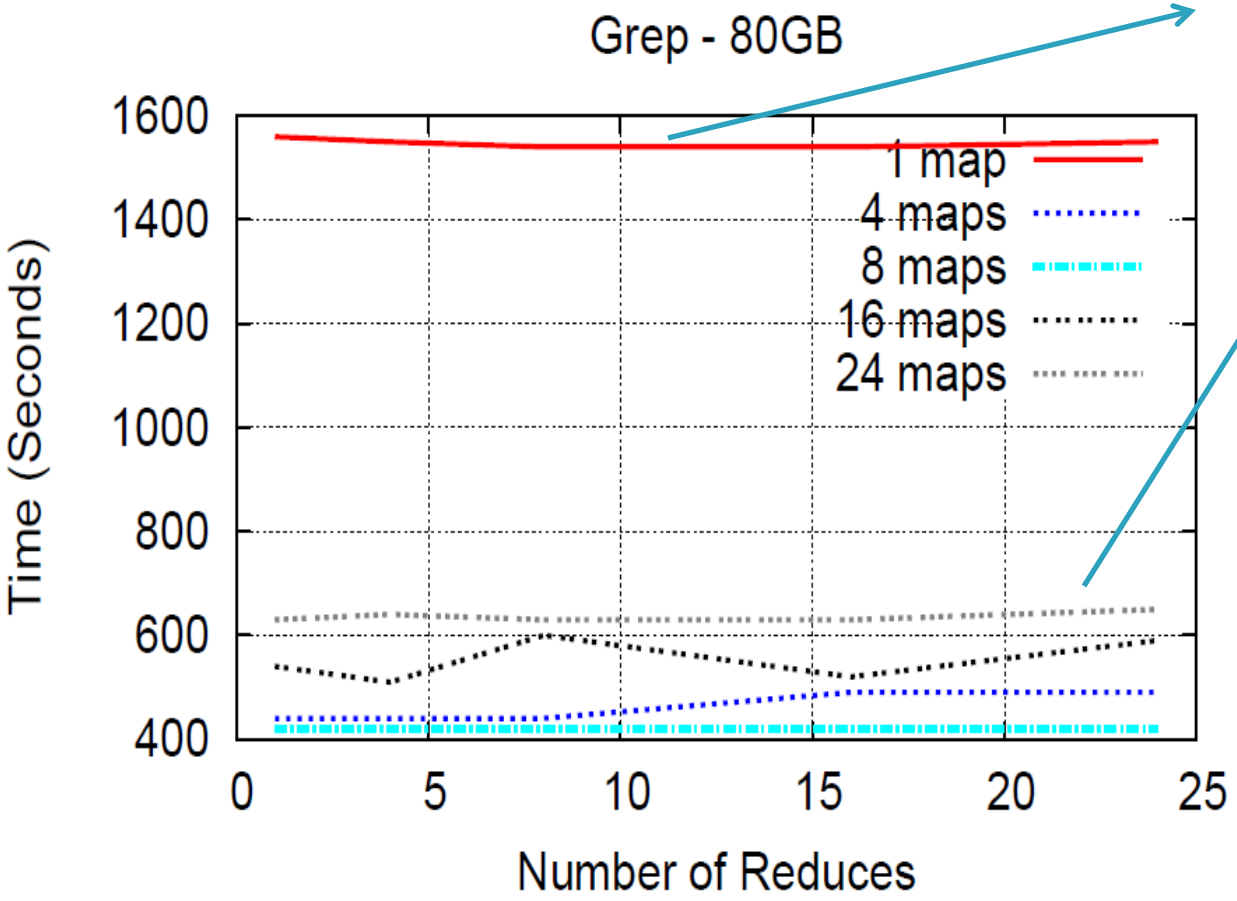


Number of Reduces doesn't affect performance

Significant Performance Difference

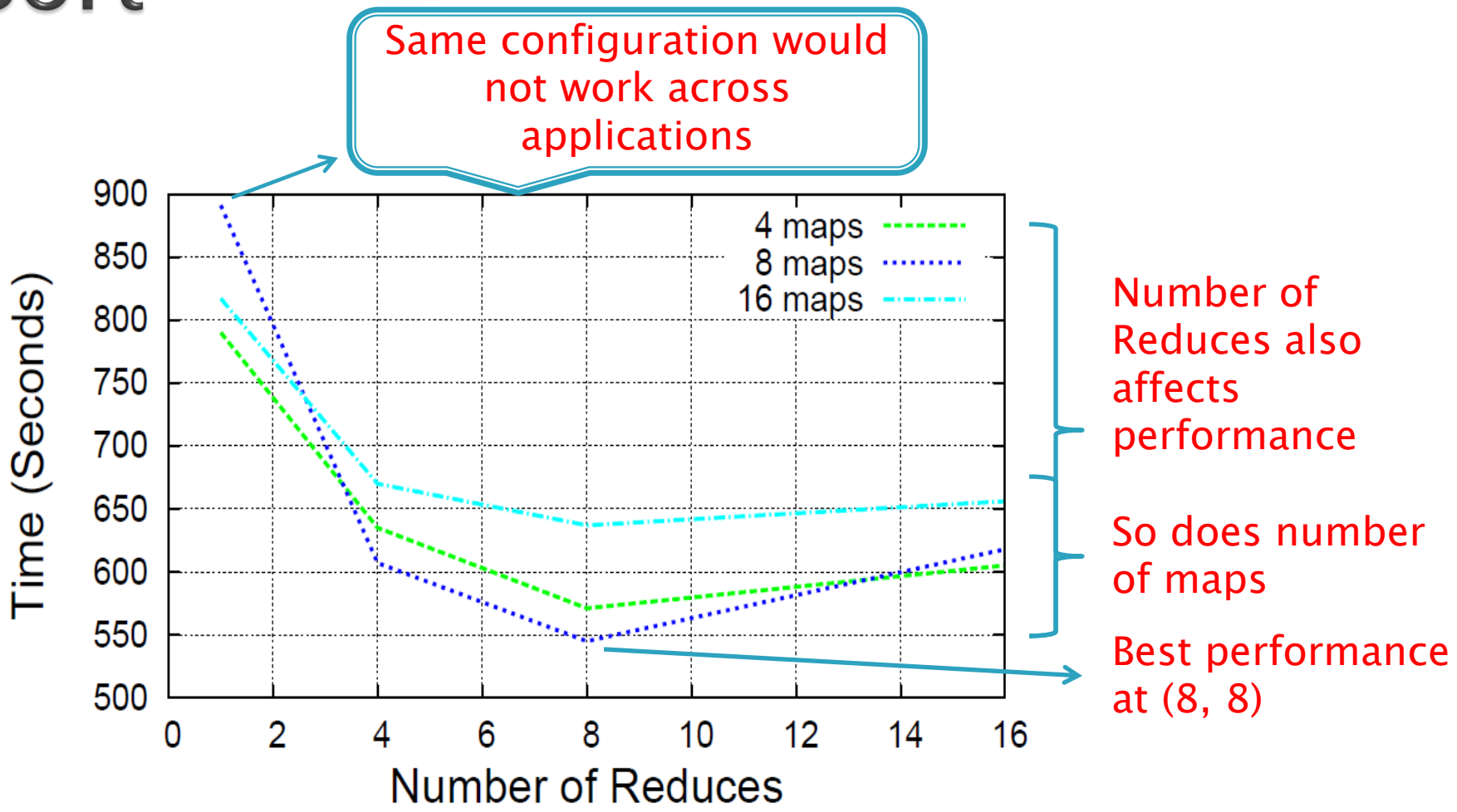
Optimal: 8 maps

Grep

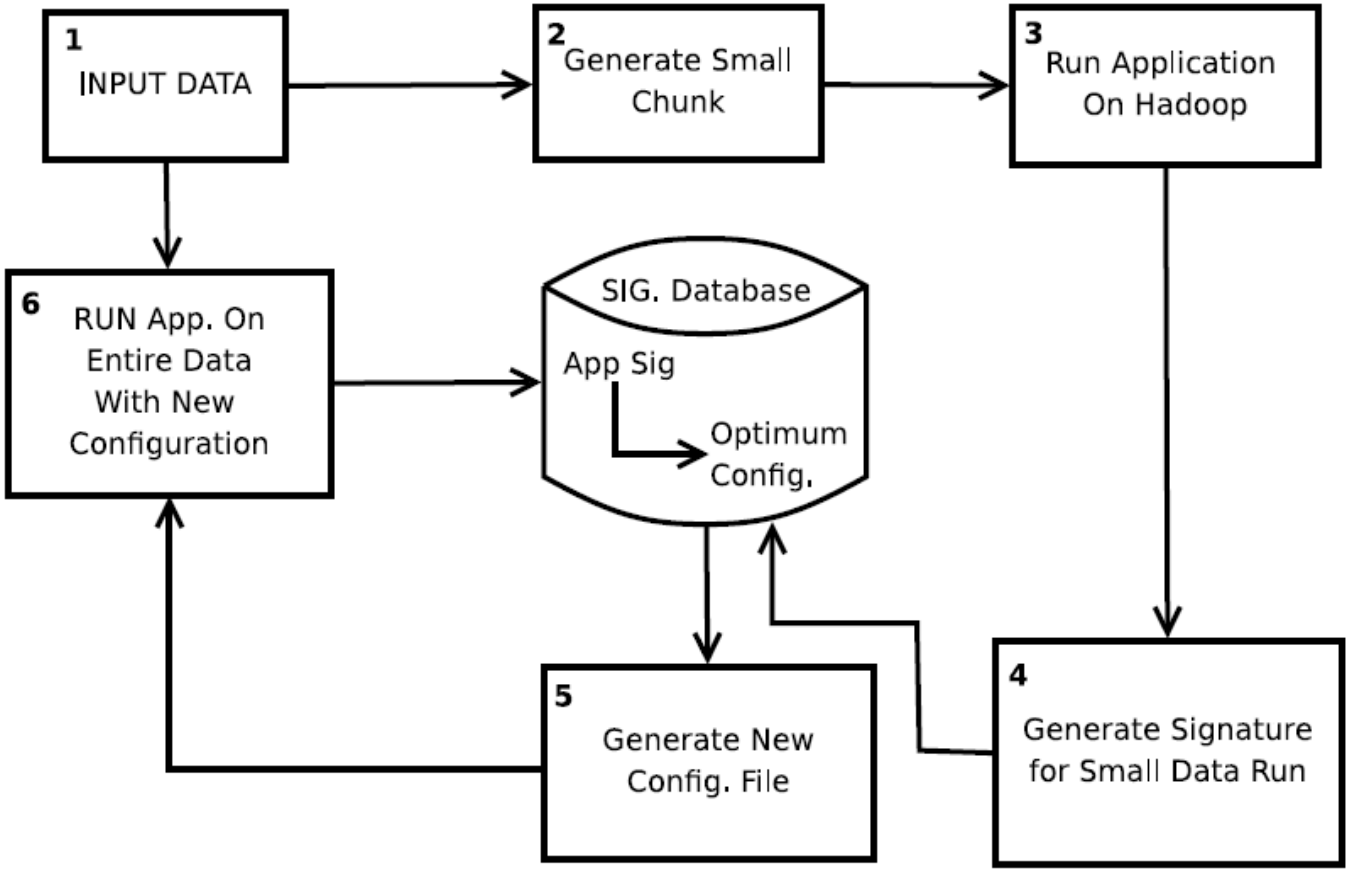


Too low doesn't work! Too high doesn't work either!

Sort


Same configuration would not work across applications
Time (Seconds)
900
850
800
750
700
650
600
550
500
0
2
4
6
8
10
12
14
16
Number of Reduces
4 maps
8 maps
16 maps
Number of Reduces also affects performance
So does number of maps
Best performance at (8, 8)

RS Maximizer



Preliminary Results

- Matrix addition, multifile-wordcount
  - Signature similar to wordcount
  - Optimal configuration is the same

Future Work

- Add a feedback phase
  - Check if predicted values are optimal
  - Else predict new optimal configuration
- RS Sizer

Thank You

Questions?