

Contents

Acknowledgement.....	V
Abstract.....	VII
Contents	XI
List of Figures.....	XIV
List of Tables	XVI
Chapter 1: Introduction.....	1
1.1 Motivation	3
1.2 Research Objective	4
1.3 Research Methodology	6
1.4 Thesis Organization	7
Chapter 2: Background.....	9
2.1 Big Data.....	9
2.1.1 <i>Big Data Characteristics</i>	10
2.1.2 <i>Big Data Processing Platform</i>	11
2.1.3 <i>Challenges in Big Data</i>	12
2.1.4 <i>Big Data Optimization</i>	13
2.2 Distributed Computing.....	14
2.2.1 <i>Distributed Computing System (DCS)</i>	14
2.2.2 <i>Distributed Computing System Architecture</i>	15
2.2.3 <i>Homogeneous Distributed Computing System (HDCS)</i>	16
2.2.4 <i>Heterogeneous Distributed Computing System (HeDCS)</i>	16
2.3 Heterogeneous Distributed Computing	17
2.3.1 <i>HeDC Challenges</i>	18
2.4 Distributed File System	18
2.4.1 <i>Fault-tolerance Support</i>	19
2.4.2 <i>Scalability Support</i>	20
2.4.3 <i>Transparency Support</i>	20
2.5 Scheduling in Heterogeneous Distributed Computing	21
2.6 Framework for Big Data	24

Chapter 3: Apache Hadoop	25
3.1 Hadoop Ecosystem.....	25
3.1.1 Hadoop Distributed File System (HDFS).....	26
3.1.2 YARN	28
3.1.3 MapReduce Programming Model	30
3.2 HDFS Block Placement Policy	31
3.2.1 Issues of Default Policy	32
3.3 Hadoop Schedulers	33
3.3.1 FIFO Scheduler.....	33
3.3.2 Capacity Scheduler	34
3.3.3 Fair Scheduler	35
3.4 YARN Node Label	37
3.4.1 YARN Node Label Working	37
3.4.2 Issues of Node Label	39
3.5 Challenges in Hadoop	40
Chapter 4: Literature Work: Analysis & Comparison	41
4.1 Study of Various Distributed File System	41
4.2 Study of Scheduling Algorithms in HeDC.....	44
4.2.1 Comparative Study	45
4.3 Study of Scheduling Algorithms in Hadoop	52
4.3.1 Experimental Environment, Workload, Performance Measure & Queue Configuration.....	52
4.3.2 Performance Evaluation.....	55
4.4 Study of Hadoop Performance Improvement Techniques	60
Chapter 5: Proposed Methodology	64
5.1 Saksham: A Resource Aware Block Rearrangement Algorithm.....	64
5.2 Saksham Model	71
Chapter 6: Experimental Results.....	74
6.1 Overview of Grid'5000	74
6.2 Experiment Setup	75
6.3 Experiment Scenarios.....	76
6.4 Performance Metrics	76
6.5 Applications & Dataset	78

6.5.1	<i>Test Applications</i>	79
6.5.2	<i>Dataset Description</i>	79
6.6	Result Analysis	80
6.6.1	<i>Benchmark Applications</i>	80
6.6.2	<i>DNA Pattern Search Application</i>	89
Chapter 7: Conclusion & Future Work		93
Appendix – I: Hadoop Setup Guide		95
Appendix – II: Grid’5000 Hadoop Setup		108
Appendix – III: Hadoop Node Label Setting		110
Appendix – IV: Saksham Usage		112
Publications		114
Bibliography		116