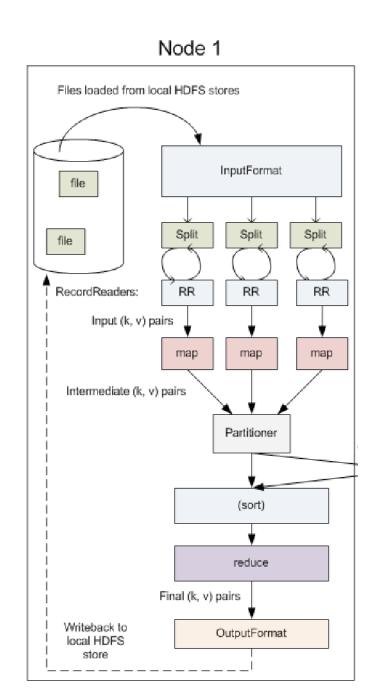


Hadoop Architecture



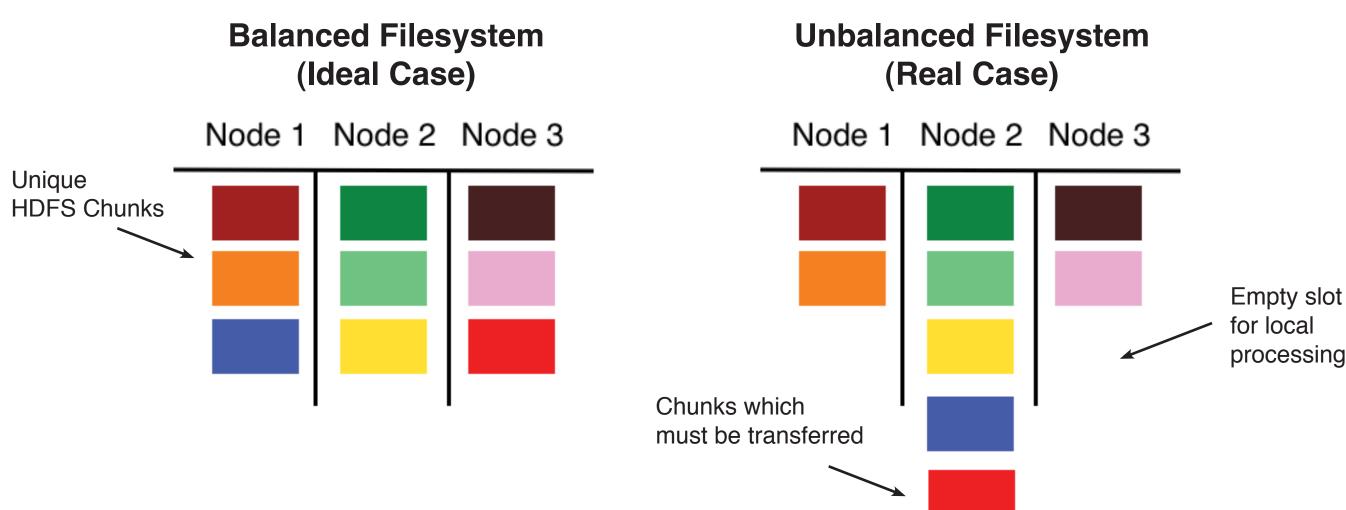
- Users submit **jobs** to Hadoop
- Jobs consist of map and reduce tasks executed by TaskTrackers
- Each map task processes one **chunk** from Hadoop distributed filesystem
- Chunk locations are known as the **input split**, which is computed in advance
- Tasks try to read the closest chunk, stored locally, rack-locally, or remotely

Typical Deployment

- Drives are placed 2-4 per node
- Nodes are organized into racks with full bandwidth
- Racks are connected at 1:5 or 1:8 bandwidth



Filesystem Imbalance in HDFS



- HDFS places chunks uniformly at random in the cluster
- The number of chunks on each node is the sum of i.i.d. Bernoulli random variables, which is *binomially distributed*
- When a block is not available locally, it must be read over a (relatively) slow network link, and compete for resources

Understanding Filesystem Imbalance in Hadoop

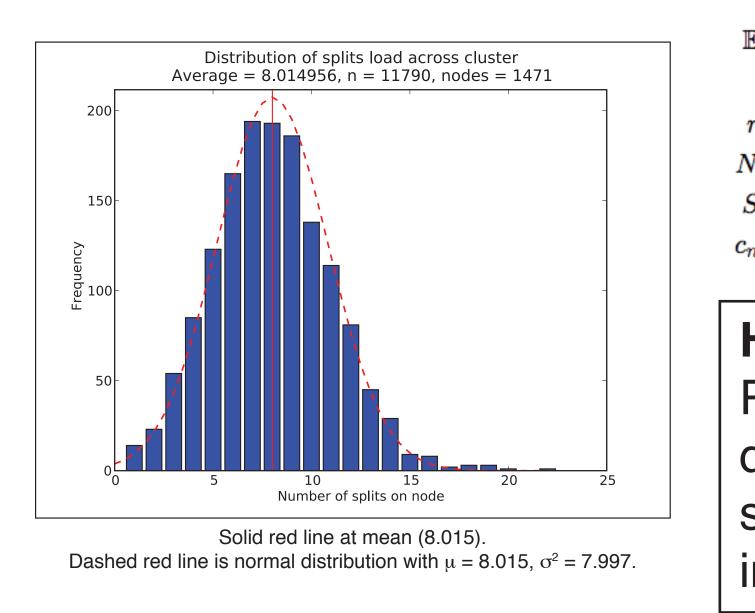
Andrew Ferguson adf@cs.brown.edu

Imbalance in the Real World

- Analyzed 93 jobs from a large company of varying sizes (34 tasks to 11,340 tasks) with a total of 41,377 tasks in total
- 13,299 tasks (32.14%) had input data local to the rack; 2,938 (7.1%) fetched data from another rack; the rest had local data
- This problem is worse for small jobs:

Job Size:	Small	Large
Number of Tasks:	181	2936
Local Tasks:	22 (12.15%)	2099 (71.49%)
Rack-Local Tasks:	111 (61.33%)	700 (23.84%)
Remote Tasks:	48 (26.52%)	137 (4.67%)

Observed input split distributions match predictions:





- The Hadoop JobTracker was modified to display in real-time the number of potentially local tasks remaining on each node, both for each job and globally across all jobs
- White cells are within one standard deviation of the average, while red nodes are one s.d. above, and blue one s.d. below

CSIADID	csiab i c	csiabid	cslable	cslab1f	cslab1g	cslab1h	cslab2a	cslab2b	cslab2c	cslab2d	cslab2e	cslab2f	cslab3a
cslab3c	cslab3d	cslab3e	cslab3f	cslab3g	cslab3h	cslab4a	cslab4b	cslab4c	cslab4d	cslab4e	cslab4f	cslab4g	cslab4h
cslab5b	cslab5c	cslab5d	cslab5e	cslab5f	cslab5g	cslab5h	cslab6a	cslab6b	cslab6c	cslab6d	cslab6e	cslab6f	cslab6g
cslab7a	cslab7b	cslab7c	cslab7d	cslab7e	cslab7f	cslab7h	cslab8a	cslab8b	cslab8c	cslab8d	cslab9a	cslab9b	cslab9d
cslab9f	cslab9g												
06161105	_0004												
# of splits	per node	e: 20.286,	Std dev:	4.395, R	ep. factor	: 3.000, T	otal # of s	splits: 127	8, Unique	e # of split	ts: 426		
18	17	25	18	0	27	22	15	20	22	21	20	19	20
24	24	23	27	25	25	25	17	21	15	18	20	23	19
22	27	20	23	22	21	18	25	27	19	25	18	20	22
26	0	25	22	19	22	20	18	22	22	17	24	20	19
15	19												
# of splits	per node	e: 20.286,	Std dev:	4.395, R	ep. factor	: 3.000, T	otal # of s	splits: 127	8, Unique	# of split	ts: 426		
18	17	25	18	0	27	22	15	20	22	21	20	19	20
24	24	23	27	25	25	25	17	21	15	18	20	23	19
22	27	20	23	22	21	18	25	27	19	25	18	20	22
26	0	25	22	19	22	20	18	22	22	17	24	20	19
15	19												
	cslab3c cslab5b cslab7a 26 15 18 24 22 26 15 18 24 22 26 12 26 22 26	cslab3c cslab3d cslab5b cslab5c cslab5b cslab7a cslab9f cslab9g cslab9f cslab9g cslab1105_0004 # of splits per node 18 17 24 24 22 27 26 0 15 19 # of splits per node 18 18 17 24 24 22 27 26 0 18 17 24 24 22 27 26 0	cslab3c cslab3d cslab3e cslab5b cslab5c cslab5d cslab7a cslab7b cslab7c cslab9f cslab7b cslab7c cslab9f cslab9g 06161105_0004 # 05 # of splits per node: 20.286, 18 17 25 24 24 23 22 27 20 26 0 25 15 19 # of splits per node: 20.286, 18 17 25 24 24 23 22 27 20 26 0 25 15 19 # of splits per node: 20.286, 18 17 25 24 24 23 22 27 20 24 24 23 22 27 20 26 0 25 </td <td>cslab3c cslab3d cslab3e cslab3f cslab5b cslab5c cslab5d cslab5c cslab7d cslab5d cslab7a cslab7b cslab7c cslab7d cslab7d cslab9f cslab9g cslab7d cslab7d 06161105_0004 # of splits per node: 20.286, Std dev: 18 17 25 18 24 24 23 27 20 23 26 0 25 22 15 19 18 17 25 18 24 24 23 27 20 23 26 0 25 22 15 19 18 27 20 23 27 22 27 20 23 27 20 23 27 22 27 20 23 27 22 23 27 22 27 20</td> <td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab5b cslab5c cslab5c cslab5d cslab5c cslab5f cslab7a cslab7b cslab7c cslab7c cslab7d cslab7e cslab9f cslab9g cslab7c cslab7d cslab7e cslab9f cslab9g cslab7c cslab7d cslab7e 06161105_0004 # of splits per node: 20.286, Std dev: 4.395, R 0 24 24 23 27 25 22 27 20 23 22 19 15 19 # of splits per node: 20.286, Std dev: 4.395, R 18 0 24 24 23 27 25 22 27 20 23 27 25 22 19 15 19 0 24 24 23 27 25 22 27 20 23 22 25 22</td> <td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab5f cslab5g cslab5g cslab7g <t< td=""><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab5b cslab5c cslab5d cslab5c cslab5d cslab5f cslab5g cslab5f cslab5g cslab5f cslab7g cslab7f csla</td><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab4b cslab5b cslab5c cslab5d cslab7a cslab7b cslab7c cslab7d cslab7e cslab7f cslab7h cslab8a cslab9f cslab9g cslab7e cslab7f cslab7h cslab8a 06161105_0004 of splits per node: 20.286, Std dev: 4.395, Rep. factor: 3.000, Total # of s 18 17 25 25 25 17 22 27 20 23 22 21 18 25 26 0 25 22 19 22 20 18 15 19 25 25 25 17 22 27 20 23 22 21 18 25</td><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab5b cslab5c cslab5d cslab6a cslab6b cslab7a cslab7b cslab7c cslab7d cslab7d cslab7f cslab7h cslab8a cslab8b cslab9f cslab9g cslab7f cslab7h cslab8a cslab8b cslab9f cslab9g cslab7f cslab7h cslab8a cslab8b cslab9f cslab9g cslab7f cslab7h cslab8a cslab8b cslab9g cslab7f cslab7f cslab7h cslab8a cslab8b cslab10105_0004 cslab7f cslab7f cslab7f cslab7f cslab7f cslab7f cslab8f cslab8b<!--</td--><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d cslab5b cslab5c cslab5d cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab7c cslab7d cslab7d cslab7d cslab7d cslab7d cslab7d cslab8a cslab8b cslab8c csla</td><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d <t< td=""><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d <t< td=""><td>06161105_0004 # of splits per node: 20.286, Std dev: 4.395, Rep. factor: 3.000, Total # of splits: 1278, Unique # of splits: 426 18 17 25 18 0 27 22 15 20 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 22 27 20 23 22 21 18 25 27 19 25 18 20 26 0 25 22 19 22 20 18 22 22 17 24 20 15 19 25 28 19 22 20 18 22 22 17 24 20 15 19 25 19 22 20 18 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 24 24 23 27 25</td></t<></td></t<></td></td></t<></td>	cslab3c cslab3d cslab3e cslab3f cslab5b cslab5c cslab5d cslab5c cslab7d cslab5d cslab7a cslab7b cslab7c cslab7d cslab7d cslab9f cslab9g cslab7d cslab7d 06161105_0004 # of splits per node: 20.286, Std dev: 18 17 25 18 24 24 23 27 20 23 26 0 25 22 15 19 18 17 25 18 24 24 23 27 20 23 26 0 25 22 15 19 18 27 20 23 27 22 27 20 23 27 20 23 27 22 27 20 23 27 22 23 27 22 27 20	cslab3c cslab3d cslab3e cslab3f cslab3g cslab5b cslab5c cslab5c cslab5d cslab5c cslab5f cslab7a cslab7b cslab7c cslab7c cslab7d cslab7e cslab9f cslab9g cslab7c cslab7d cslab7e cslab9f cslab9g cslab7c cslab7d cslab7e 06161105_0004 # of splits per node: 20.286, Std dev: 4.395, R 0 24 24 23 27 25 22 27 20 23 22 19 15 19 # of splits per node: 20.286, Std dev: 4.395, R 18 0 24 24 23 27 25 22 27 20 23 27 25 22 19 15 19 0 24 24 23 27 25 22 27 20 23 22 25 22	cslab3c cslab3d cslab3e cslab3f cslab3g cslab5f cslab5g cslab5g cslab7g cslab7g <t< td=""><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab5b cslab5c cslab5d cslab5c cslab5d cslab5f cslab5g cslab5f cslab5g cslab5f cslab7g cslab7f csla</td><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab4b cslab5b cslab5c cslab5d cslab7a cslab7b cslab7c cslab7d cslab7e cslab7f cslab7h cslab8a cslab9f cslab9g cslab7e cslab7f cslab7h cslab8a 06161105_0004 of splits per node: 20.286, Std dev: 4.395, Rep. factor: 3.000, Total # of s 18 17 25 25 25 17 22 27 20 23 22 21 18 25 26 0 25 22 19 22 20 18 15 19 25 25 25 17 22 27 20 23 22 21 18 25</td><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab5b cslab5c cslab5d cslab6a cslab6b cslab7a cslab7b cslab7c cslab7d cslab7d cslab7f cslab7h cslab8a cslab8b cslab9f cslab9g cslab7f cslab7h cslab8a cslab8b cslab9f cslab9g cslab7f cslab7h cslab8a cslab8b cslab9f cslab9g cslab7f cslab7h cslab8a cslab8b cslab9g cslab7f cslab7f cslab7h cslab8a cslab8b cslab10105_0004 cslab7f cslab7f cslab7f cslab7f cslab7f cslab7f cslab8f cslab8b<!--</td--><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d cslab5b cslab5c cslab5d cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab7c cslab7d cslab7d cslab7d cslab7d cslab7d cslab7d cslab8a cslab8b cslab8c csla</td><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d <t< td=""><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d <t< td=""><td>06161105_0004 # of splits per node: 20.286, Std dev: 4.395, Rep. factor: 3.000, Total # of splits: 1278, Unique # of splits: 426 18 17 25 18 0 27 22 15 20 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 22 27 20 23 22 21 18 25 27 19 25 18 20 26 0 25 22 19 22 20 18 22 22 17 24 20 15 19 25 28 19 22 20 18 22 22 17 24 20 15 19 25 19 22 20 18 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 24 24 23 27 25</td></t<></td></t<></td></td></t<>	cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab5b cslab5c cslab5d cslab5c cslab5d cslab5f cslab5g cslab5f cslab5g cslab5f cslab7g cslab7f csla	cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab4b cslab5b cslab5c cslab5d cslab7a cslab7b cslab7c cslab7d cslab7e cslab7f cslab7h cslab8a cslab9f cslab9g cslab7e cslab7f cslab7h cslab8a 06161105_0004 of splits per node: 20.286, Std dev: 4.395, Rep. factor: 3.000, Total # of s 18 17 25 25 25 17 22 27 20 23 22 21 18 25 26 0 25 22 19 22 20 18 15 19 25 25 25 17 22 27 20 23 22 21 18 25	cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab5b cslab5c cslab5d cslab6a cslab6b cslab7a cslab7b cslab7c cslab7d cslab7d cslab7f cslab7h cslab8a cslab8b cslab9f cslab9g cslab7f cslab7h cslab8a cslab8b cslab9f cslab9g cslab7f cslab7h cslab8a cslab8b cslab9f cslab9g cslab7f cslab7h cslab8a cslab8b cslab9g cslab7f cslab7f cslab7h cslab8a cslab8b cslab10105_0004 cslab7f cslab7f cslab7f cslab7f cslab7f cslab7f cslab8f cslab8b </td <td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d cslab5b cslab5c cslab5d cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab7c cslab7d cslab7d cslab7d cslab7d cslab7d cslab7d cslab8a cslab8b cslab8c csla</td> <td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d <t< td=""><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d <t< td=""><td>06161105_0004 # of splits per node: 20.286, Std dev: 4.395, Rep. factor: 3.000, Total # of splits: 1278, Unique # of splits: 426 18 17 25 18 0 27 22 15 20 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 22 27 20 23 22 21 18 25 27 19 25 18 20 26 0 25 22 19 22 20 18 22 22 17 24 20 15 19 25 28 19 22 20 18 22 22 17 24 20 15 19 25 19 22 20 18 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 24 24 23 27 25</td></t<></td></t<></td>	cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d cslab5b cslab5c cslab5d cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab5d cslab5c cslab7c cslab7d cslab7d cslab7d cslab7d cslab7d cslab7d cslab8a cslab8b cslab8c csla	cslab3c cslab3d cslab3e cslab3f cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d cslab4d <t< td=""><td>cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d <t< td=""><td>06161105_0004 # of splits per node: 20.286, Std dev: 4.395, Rep. factor: 3.000, Total # of splits: 1278, Unique # of splits: 426 18 17 25 18 0 27 22 15 20 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 22 27 20 23 22 21 18 25 27 19 25 18 20 26 0 25 22 19 22 20 18 22 22 17 24 20 15 19 25 28 19 22 20 18 22 22 17 24 20 15 19 25 19 22 20 18 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 24 24 23 27 25</td></t<></td></t<>	cslab3c cslab3d cslab3e cslab3f cslab3g cslab3g cslab3h cslab4a cslab4b cslab4c cslab4d cslab4d <t< td=""><td>06161105_0004 # of splits per node: 20.286, Std dev: 4.395, Rep. factor: 3.000, Total # of splits: 1278, Unique # of splits: 426 18 17 25 18 0 27 22 15 20 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 22 27 20 23 22 21 18 25 27 19 25 18 20 26 0 25 22 19 22 20 18 22 22 17 24 20 15 19 25 28 19 22 20 18 22 22 17 24 20 15 19 25 19 22 20 18 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 24 24 23 27 25</td></t<>	06161105_0004 # of splits per node: 20.286, Std dev: 4.395, Rep. factor: 3.000, Total # of splits: 1278, Unique # of splits: 426 18 17 25 18 0 27 22 15 20 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 22 27 20 23 22 21 18 25 27 19 25 18 20 26 0 25 22 19 22 20 18 22 22 17 24 20 15 19 25 28 19 22 20 18 22 22 17 24 20 15 19 25 19 22 20 18 22 21 20 19 24 24 23 27 25 25 25 17 21 15 18 20 23 24 24 23 27 25

Standard Placement

Rodrigo Fonseca rfonseca@cs.brown.edu



 $\mathbb{E}c_n = S imes rac{rS}{N} = rac{rS}{N}$ $c_n \sim \mathcal{N}\left(rac{rS}{N}, rac{r(N-r)S}{N^2}
ight)$

= replication factor of each chunk the DFS = number of nodes in the cluster = number of unique chunks in the input splits c_n = number of chunks on node n

Hypothesis:

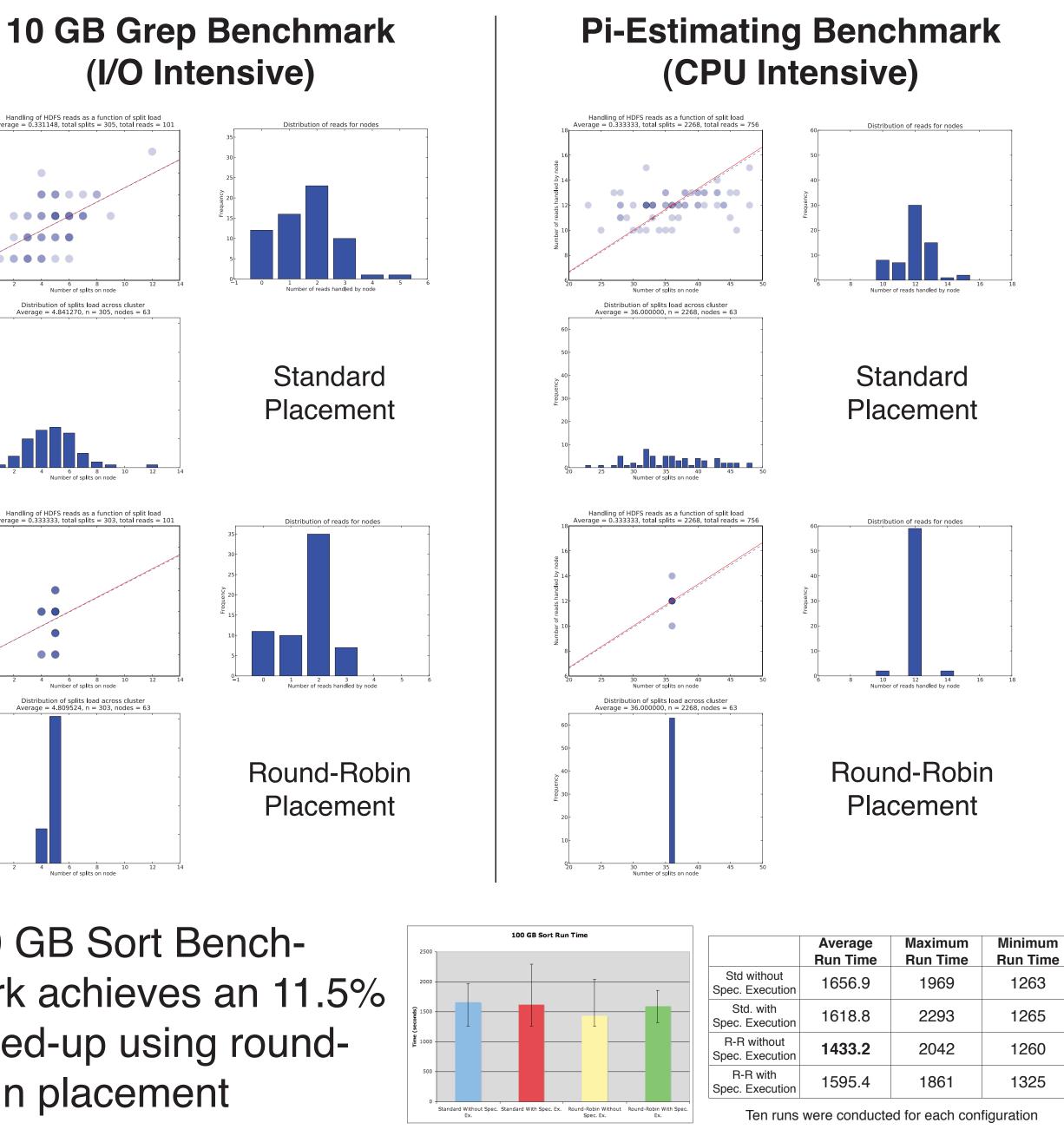
Round-robin placement will decrease the variance of the splits distribution and yield improved performance.

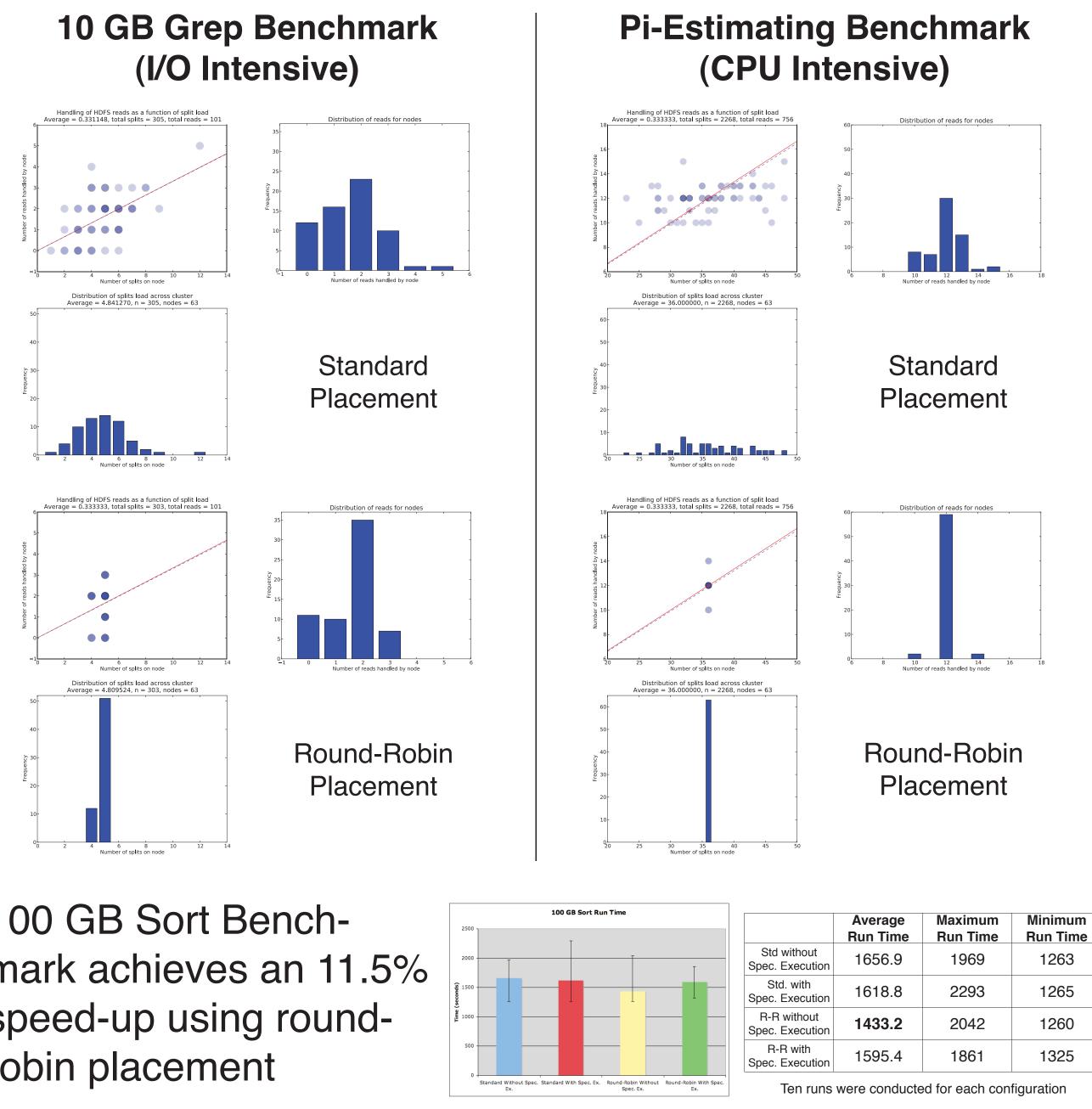
18 14 1

o1c	cslab1d	cslab1e	cslab1f	cslab1g	cslab1h	cslab2a	cslab2b	cslab2c	cslab2d	cslab2e	cslab2f	cslab3a
o3d	cslab3e	cslab3f	cslab3g	cslab3h	cslab4a	cslab4b	cslab4c	cslab4d	cslab4e	cslab4f	cslab4g	cslab4h
o5c	cslab5d	cslab5e	cslab5f	cslab5g	cslab5h	cslab6a	cslab6b	cslab6c	cslab6d	cslab6e	cslab6f	cslab6g
o7b	cslab7c	cslab7d	cslab7e	cslab7f	cslab7h	cslab8a	cslab8b	cslab8c	cslab8d	cslab9a	cslab9b	cslab9c
o9g												
1 node	e: 16.095,	Std dev:	3.915, R	ep. factor	: 3.000, T	otal # of s	plits: 101	4, Unique	# of split	ts: 338		
;	16	15	16	17	14	16	14	17	17	17	14	18
3	15	14	17	17	18	15	15	15	16	14	16	17
5	14	17	17	16	17	15	17	17	16	14	14	16
}	15	17	17	17	17	17	15	17	15	16	17	17
,												
ode	e: 16.095,			-								
5	16	15	16	17	14	16	14	17	17	17	14	18
;	15	14	17	17	18	15	15	15	16	14	16	17
;	14	17	17	16	17	15	17	17	16	14	14	16
	15	17	17	17	17	17	15	17	15	16	17	17
}												

Round-Robin Placement

Evaluating Round-Robin Placement





• 100 GB Sort Benchmark achieves an 11.5% speed-up using roundrobin placement

Conclusions and Next Steps

• Hadoop ver. 0.20.1 was augmented with round-robin placement in addition to uniform-at-random (standard placement)

• Experiments were run on a cluster of 63 nodes (21 per rack) with two additional nodes as masters. Nodes had 4 x 2.4 GHz CPUs, 3 GB of RAM, and gigabit Ethernet connections

• The performance improvements from round-robin placement illustrate the benefits of a more-balanced filesystem.

• In future work, we will examine whether round-robin-like block placement can improve the performance of the new "delayscheduling" technique, and construct a characterization of the theoretically best read pattern for a given input split.