Google Cluster Trace Analysis with Apache Spark

In this work, I analysed Google cluster-usage traces using Apache Spark and I created the document using R Studio.

You can access the subject of this work from this link. You can find the detailed documantatin of traces from this link.

This version of the code does not contains the Spark codes which are used to extract information from the traces also this is not the final version. There is no guarante related to correctness of the results.

1. Machine Distibution

Machine Distibution According to CPU and Memory Capacity

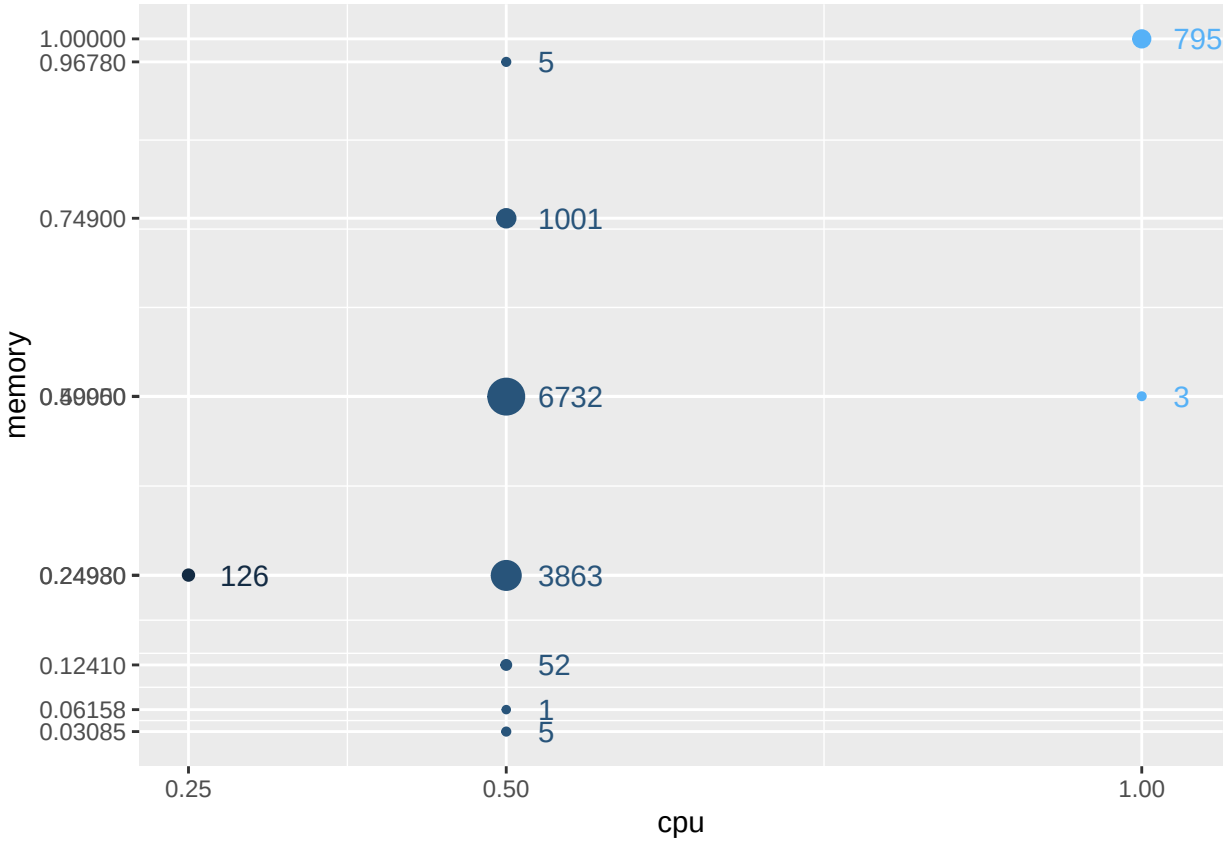


Table 1: Tabular Data

cpu	memory	capacity	number_of_machines	machine_percentage
1.00	1.00000	1.000000	795	6.32
1.00	0.50000	0.750000	3	0.02
0.50	0.96780	0.733900	5	0.04
0.50	0.74900	0.624500	1001	7.96
0.50	0.49950	0.499750	6732	53.50
0.50	0.24930	0.374650	3863	30.70
0.50	0.12410	0.312050	52	0.41
0.50	0.06158	0.280790	1	0.01
0.50	0.03085	0.265425	5	0.04
0.25	0.24980	0.249900	126	1.00

Machine Distibution According to Platform

platform_id	number_of_machines
HofLGzk1Or/8Ildj2+Lqv0UGGvY82NLoni8+J/Yy0RU=	11659
GtXakjpd0CD41brK7k/27s3Eby3RpJKy7taB9S8UQRA=	798
70ZOvysYGtB6j9MUHMPzA2Iy7GRzWeJTdX0YCLRKGVg=	126

2. Job and Task Counts

number_of_jobs	number_of_tasks	avg_number_of_task_count
672,004	25,424,731	37.8342

	statistics
Min.	1.0000
1st Qu.	1.0000
Median	1.0000
Mean	37.8342
3rd Qu.	1.0000
Max.	90050.0000

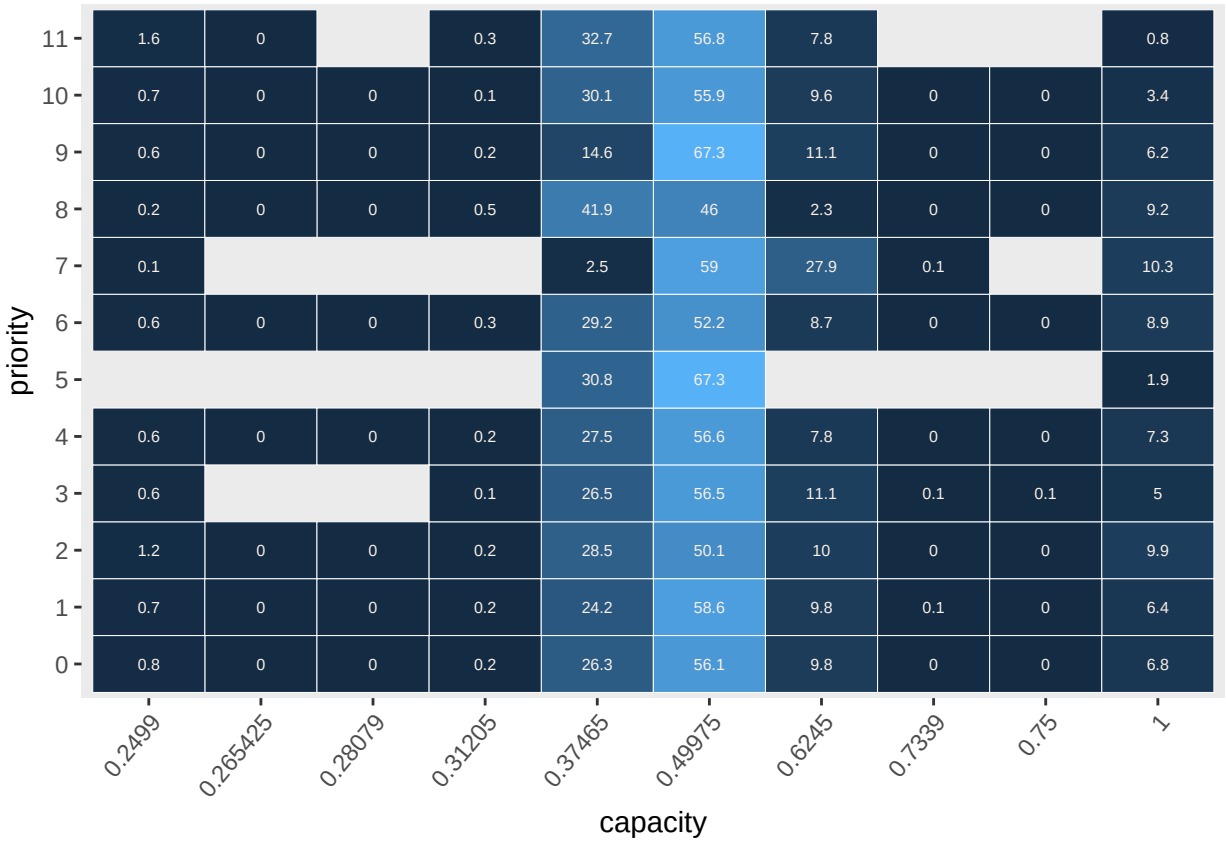
3. Job/Task Life Cycle Analysis (Killed and Evicted Job/Task)

priority	killed_percent	evicted_percent	failed_percent	lost_percent	task_count
11	25.6	0.0	0.0	0.0	7,538
10	668.8	35.9	1,526.9	0.1	1,403
9	79.6	5.9	457.6	0.0	286,269
8	54.2	1.2	6.7	0.0	254,680
7	199.0	0.0	0.0	0.0	400
6	21.2	0.1	8.4	0.0	639,784
5	0.0	0.0	0.0	0.0	104
4	23.7	0.3	4.5	0.0	14,197,733
3	6.0	14.9	0.1	0.0	1,027
2	14.8	4.3	2.4	0.0	1,111,810
1	37.6	18.8	29.4	0.0	2,453,482
0	83.2	81.7	170.5	0.1	6,472,128

4. Eviction Probability of Tasks with Respect to Priority

You can look the table of the question 3 to conclude about this.

5. Relation between the Priority of a Task and Machine Resources(CPU, Memory)

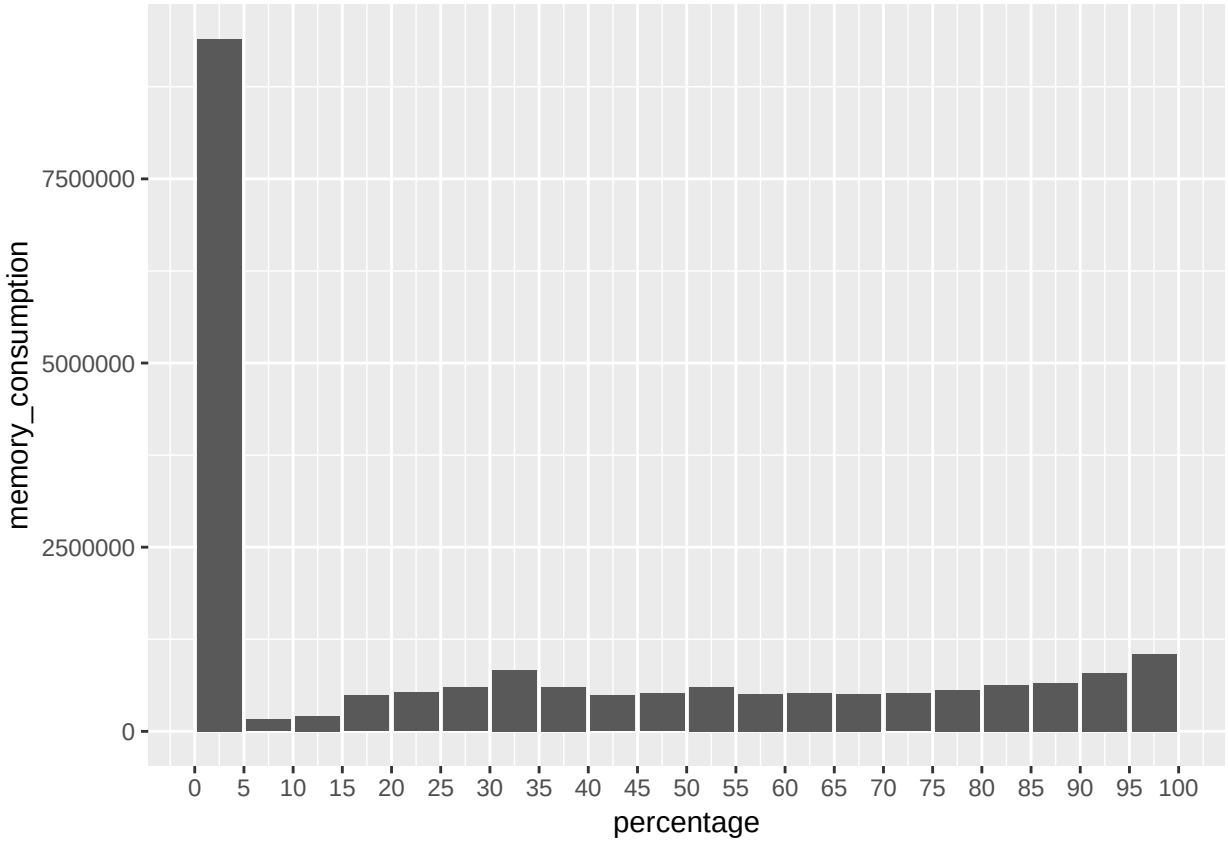


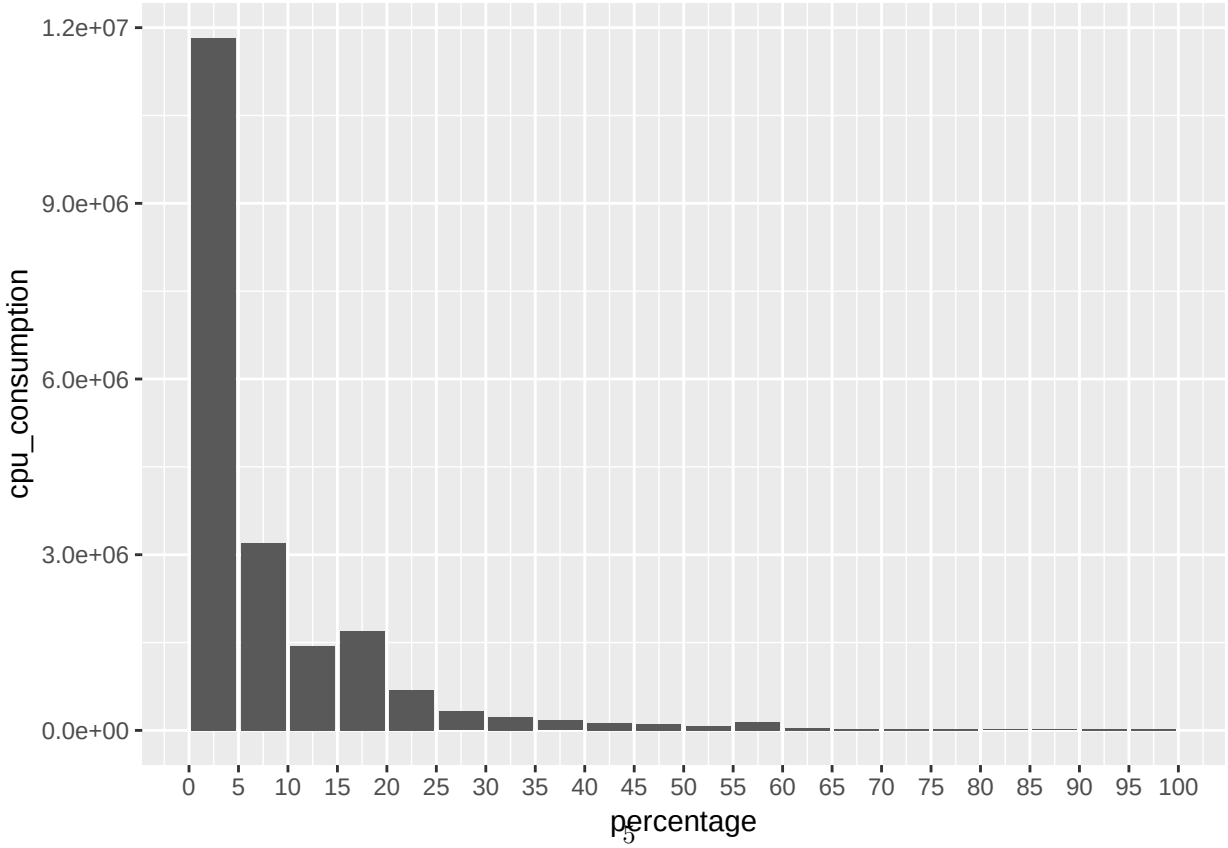
priority	total_task_count
0	24,640,306
1	3,745,699
2	1,200,766
3	1,203
4	15,099,442
5	104
6	689,599
7	796
8	274,962
9	1,656,760
10	32,128
11	9,405

capacity	cpu	memory	machine_percentage
1.000000	1.00	1.00000	6.32
0.750000	1.00	0.50000	0.02
0.733900	0.50	0.96780	0.04
0.624500	0.50	0.74900	7.96
0.499750	0.50	0.49950	53.50
0.374650	0.50	0.24930	30.70
0.312050	0.50	0.12410	0.41
0.280790	0.50	0.06158	0.01
0.265425	0.50	0.03085	0.04
0.249900	0.25	0.24980	1.00

From heatmap and machine distribution data we can conclude that tasks are uniformly distributed on machines.

7. Tasks consumes significantly less resource than what they requested





Adding some text :)