Objective: Performance assessment of below given four service calls hosted on AWS 54.229.152.248

Application URL: http://54.229.152.248:8967

Tool Used: locust.io

Test Cases:

1. /endpoint1
2. /endpoint2
3. /endpoint3
4. /endpoint4

Github Scripts Location: https://github.com/irfanbr/PerfBenchMarkVPLocust

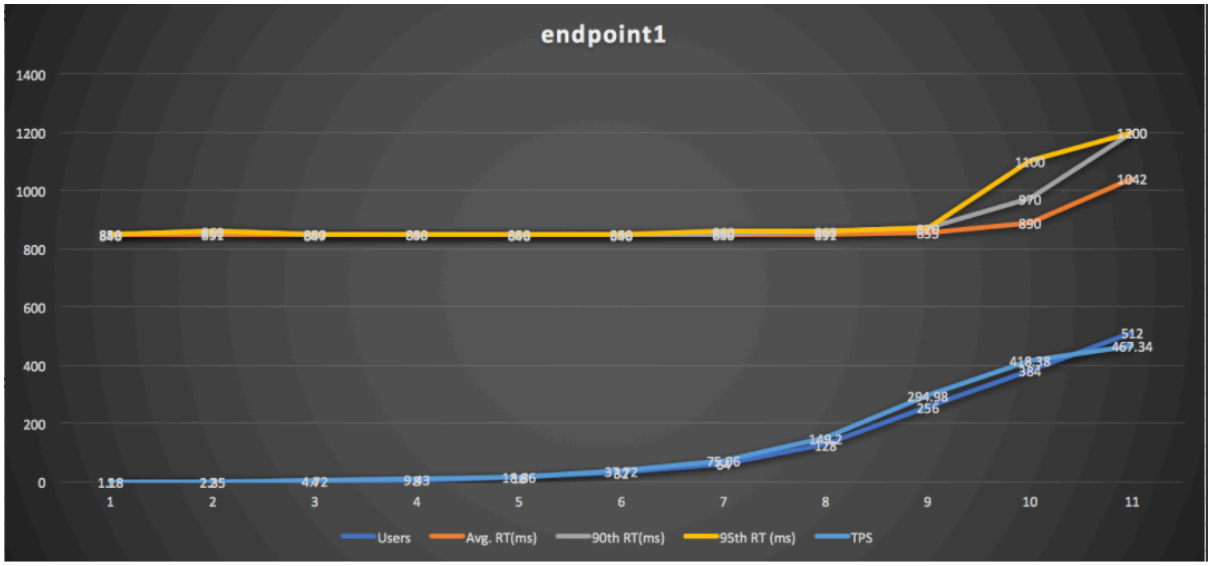
Isolated Scalability Tests

Note*: All the Scalability runs are executed with a duration of 5 min

A. endpoint1

Result Summary:

Users	Count	Avg. RT(ms)	90th RT(ms)	95th RT (ms)	98th RT(ms)	99th RT(ms)	TPS
1	705	848	850	850	850	850	1.18
2	1409	851	860	860	850	860	2.35
4	2832	847	850	850	850	850	4.72
8	5650	849	850	850	850	850	9.43
16	11316	848	850	850	850	860	18.86
32	11315	848	850	850	860	860	37.72
64	22549	849	850	860	860	860	75.06
128	44820	851	860	860	860	870	149.2
256	88389	855	870	870	880	890	294.98
384	125687	890	970	1100	1100	1200	418.38
512	140422	1042	1200	1200	1300	1300	467.34



Observations:

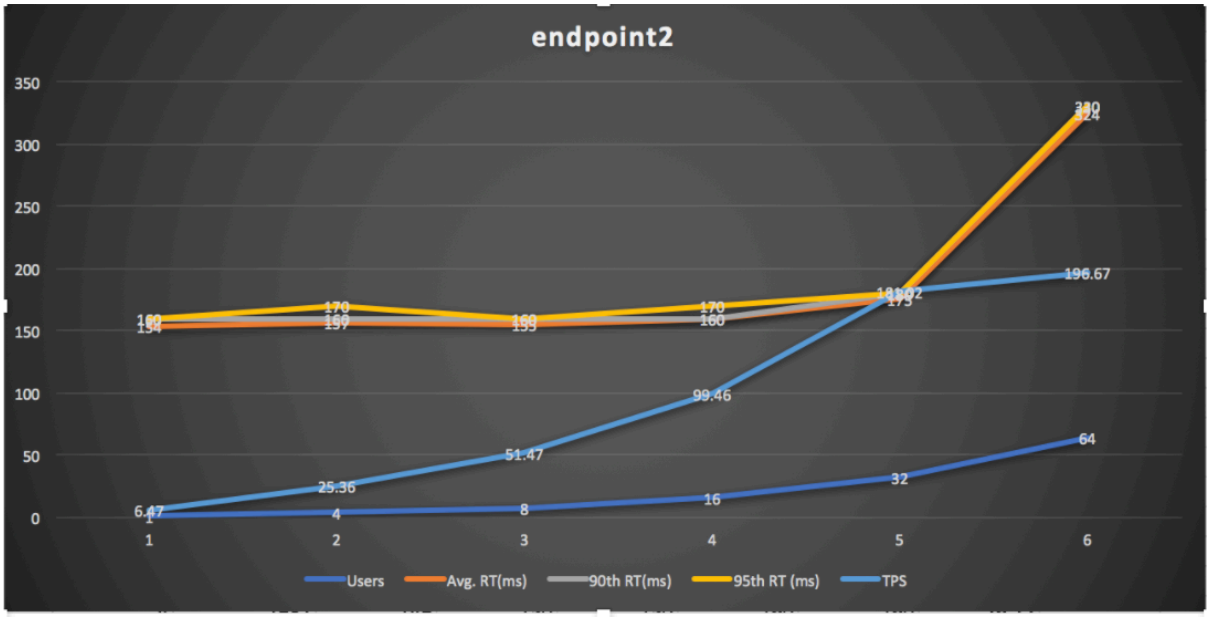
1. TPS is scaling linearly with increase in load upto 256 user and able to achieve 294 TPS with 95th percentile response time of 870ms
2. Beyond 256 user load, response time start degrading.

*: Test Duration was set to 10 min till 16 user run for endpoint1

B. endpoint2

Result Summary:

Users	Count	Avg. RT(ms)	90th RT(ms)	95th RT (ms)	98th RT(ms)	99th RT(ms)	TPS
1	1885	154	160	160	160	160	6.47
4	7594	157	160	170	170	170	25.36
8	15393	155	160	160	160	160	51.47
16	29791	160	160	170	170	170	99.46
32	54298	175	180	180	190	190	181.02
64	58994	324	330	330	340	340	196.67



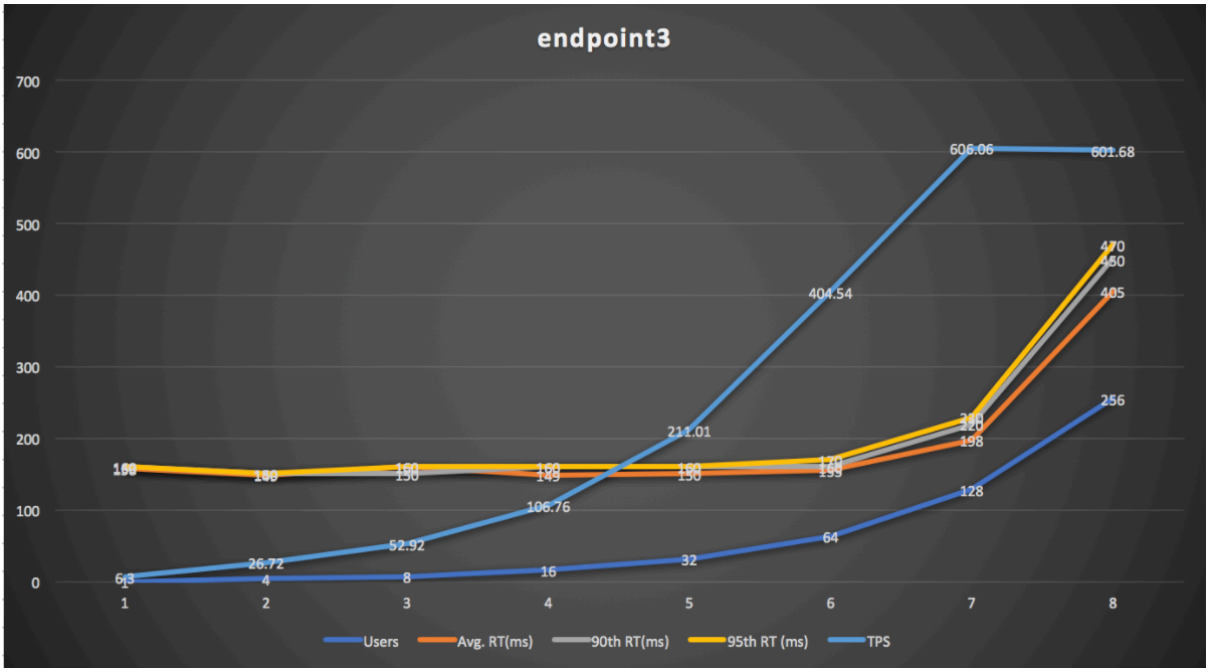
Observations:

1. TPS scaled linearly with increase in load upto 32 users and able to achieve TPS of 181 with 95th percentile response of 180ms
2. Beyond 32 users, response time started degrading.

C. Endpoint3

Result Summary:

Users	Count	Avg. RT(ms)	90th RT(ms)	95th RT (ms)	98th RT(ms)	99th RT(ms)	TPS
1	1880	158	160	160	160	160	6.3
4	7987	149	150	150	160	160	26.72
8	15868	160	150	160	160	160	52.92
16	31979	149	160	160	160	160	106.76
32	63282	150	160	160	160	160	211.01
64	121141	155	160	170	170	170	404.54
128	181830	198	220	230	240	250	606.06
256	180526	405	450	470	490	510	601.68



Observations:

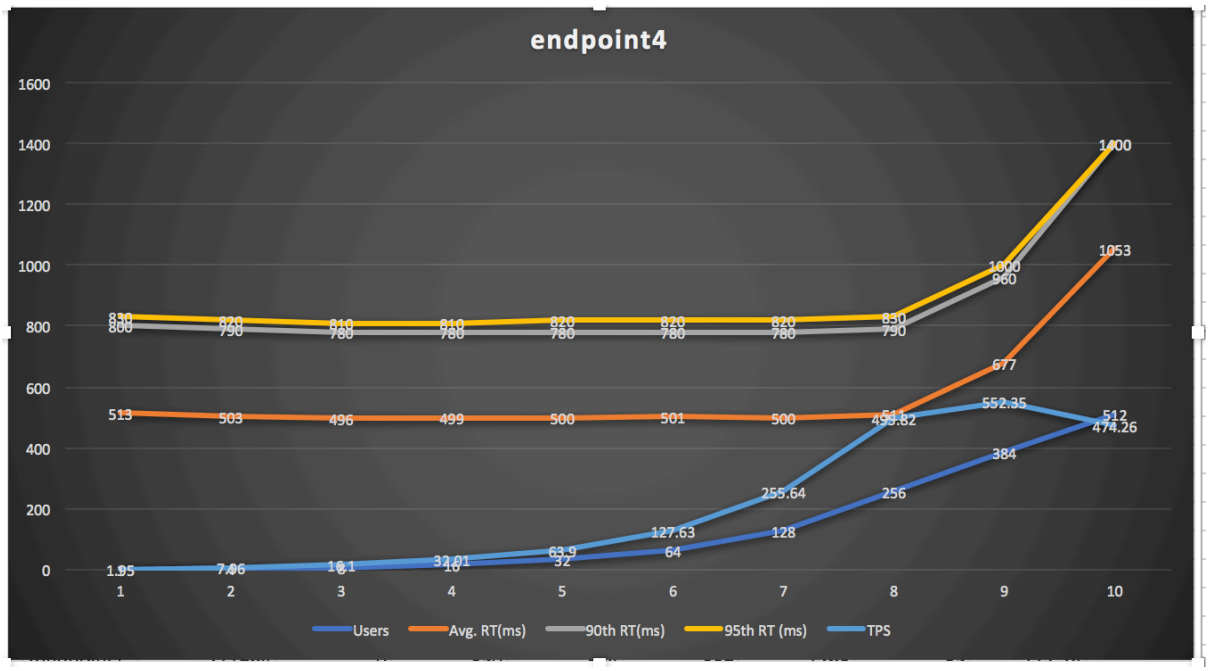
1. TPS scaled linearly with increase in load upto 64 users and able to achieve TPS of 404 with 95th percentile response of 170ms
2. Beyond 64 users, response time started degrading.

D. endpoint4

Result Summary:

Users	Count	Avg. RT(ms)	90th RT(ms)	95th RT (ms)	98th RT(ms)	99th RT(ms)	TPS
1	584	513	800	830	840	850	1.95
4	2381	503	790	820	840	850	7.96
8	4830	496	780	810	840	840	16.1
16	9608	499	780	810	840	840	32.01
32	19157	500	780	820	840	840	63.9
64	38242	501	780	820	840	840	127.63
128	76586	500	780	820	840	840	255.64

256	148870	511	790	830	850	860	495.82
384	165784	677	960	1000	1100	1100	552.35
512	474.26	1053	1400	1400	1600	1700	474.26



Observations:

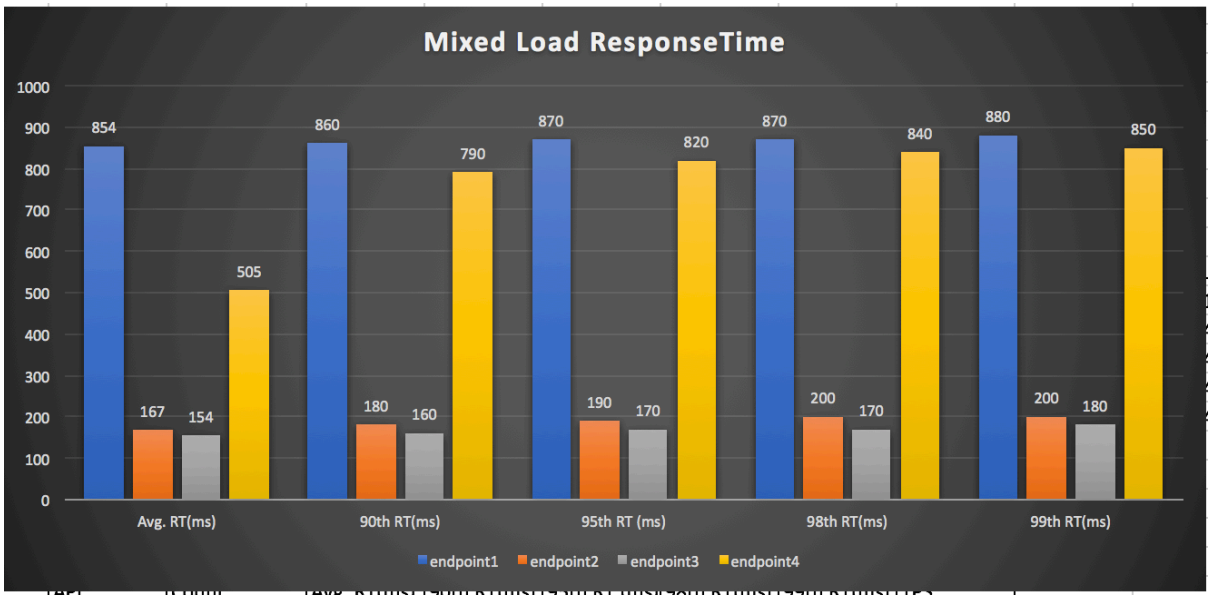
1. TPS scaled linearly with increase in load upto 256 users and able to achieve TPS of 495 with 95th percentile response of 830ms.
2. Beyond 64 users, response time started degrading.
3. With further increase in load, TPS drops down.

Mixed Load Tests:
Achieving Highest TPS with optimal response time inline with baseline.
All four endpoints has given equal weightage.

Users: 200 with 20 seconds ramp up
Duration: 30 min

Result Summary:

API	Count	Avg. RT(ms)	90th RT(ms)	95th RT (ms)	98th RT(ms)	99th RT(ms)	TPS
endpoint1	211964	854	860	870	870	880	117.74
endpoint2	210808	167	180	190	200	200	117.1
endpoint3	211216	154	160	170	170	180	117.33
endpoint4	211686	505	790	820	840	850	117.59
Total	845674	421	860	860	860	870	469.77



Observations:

1. With 200 user load, able to achieve overall TPS Of 469.
2. 95th percentile response times are close to baselines(from Isolated scalability tests).