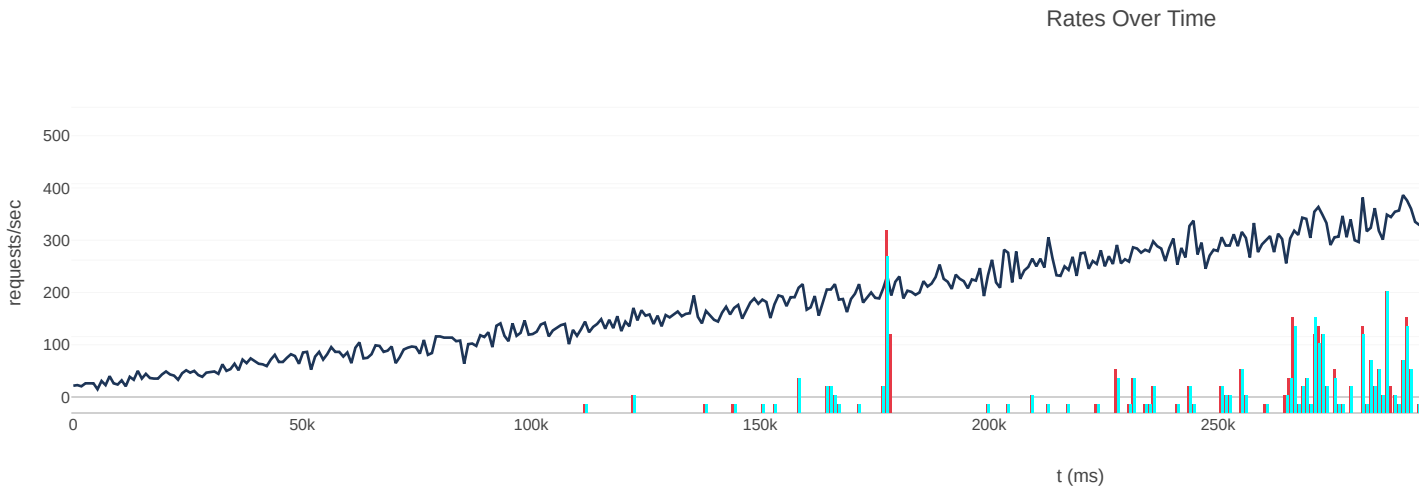
yolo-router Simulation Report

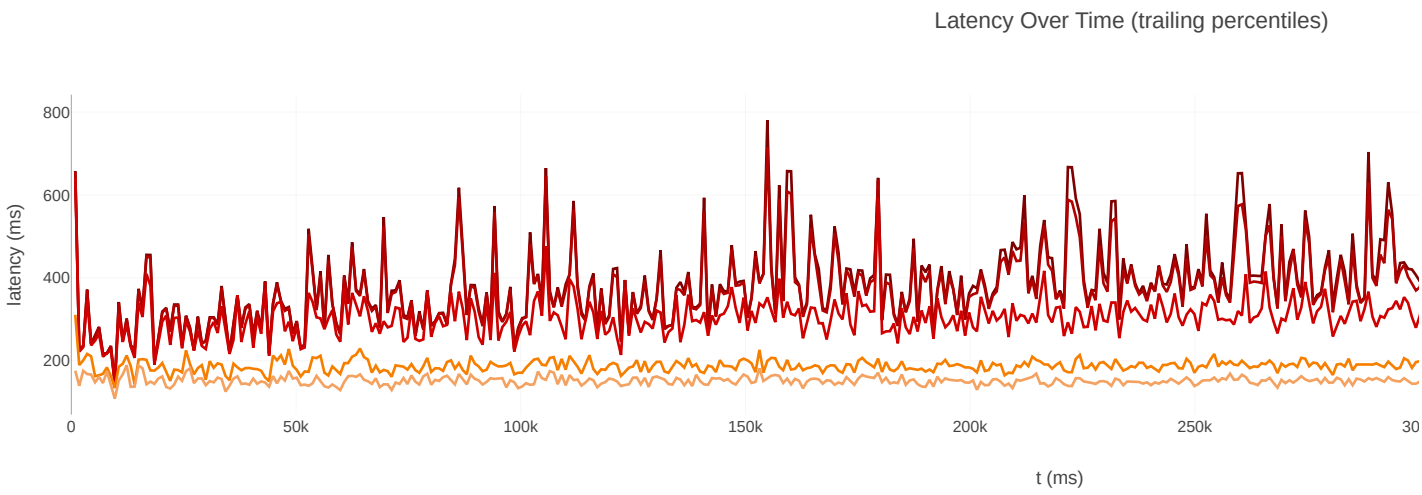
Run Summary

Sim Time (ms)	Total Requests	Scale Ups	Scale Downs	Wall Runtime (s)
440000	96698	10	0	3.293

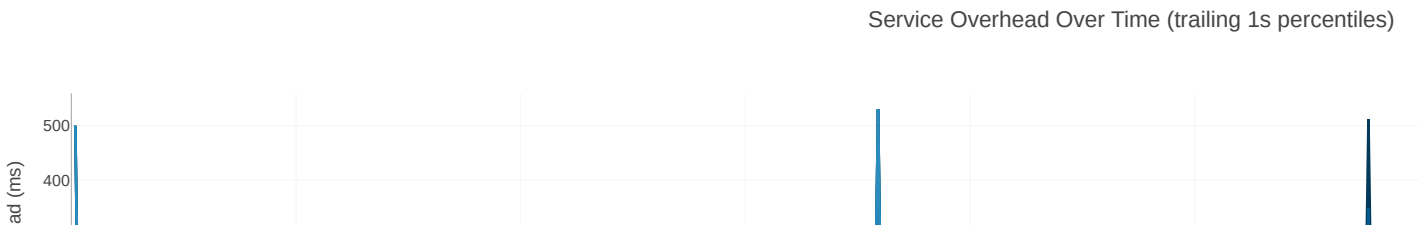
Rates Over Time

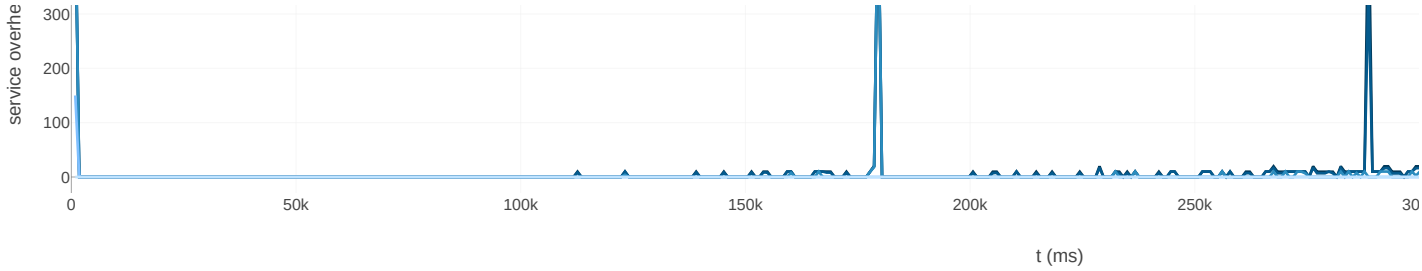


Latency Over Time

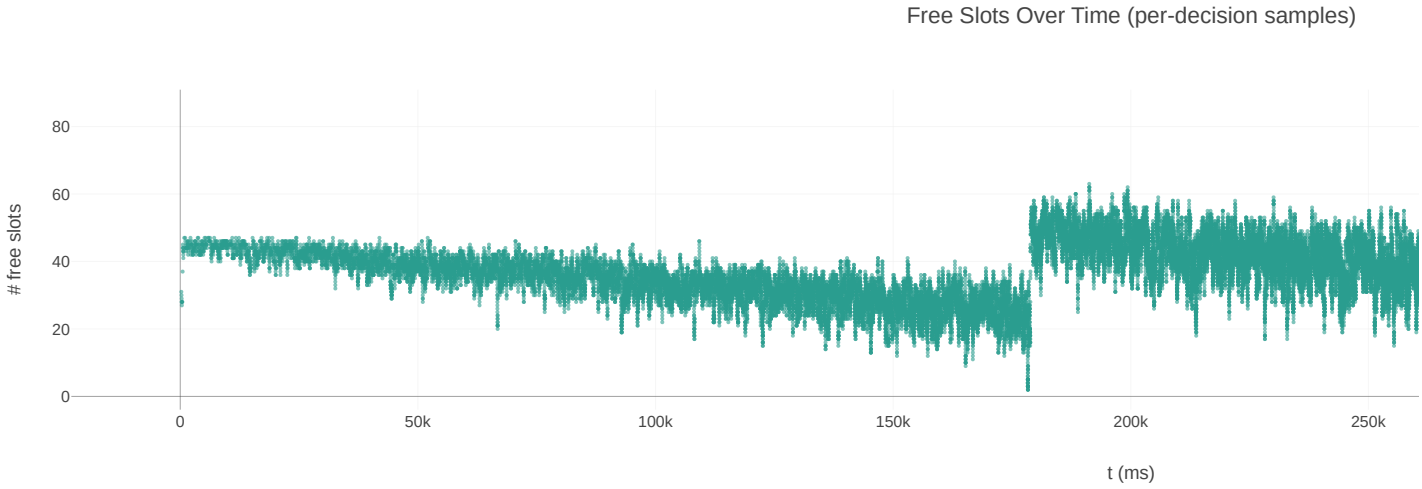


Service Overhead Over Time

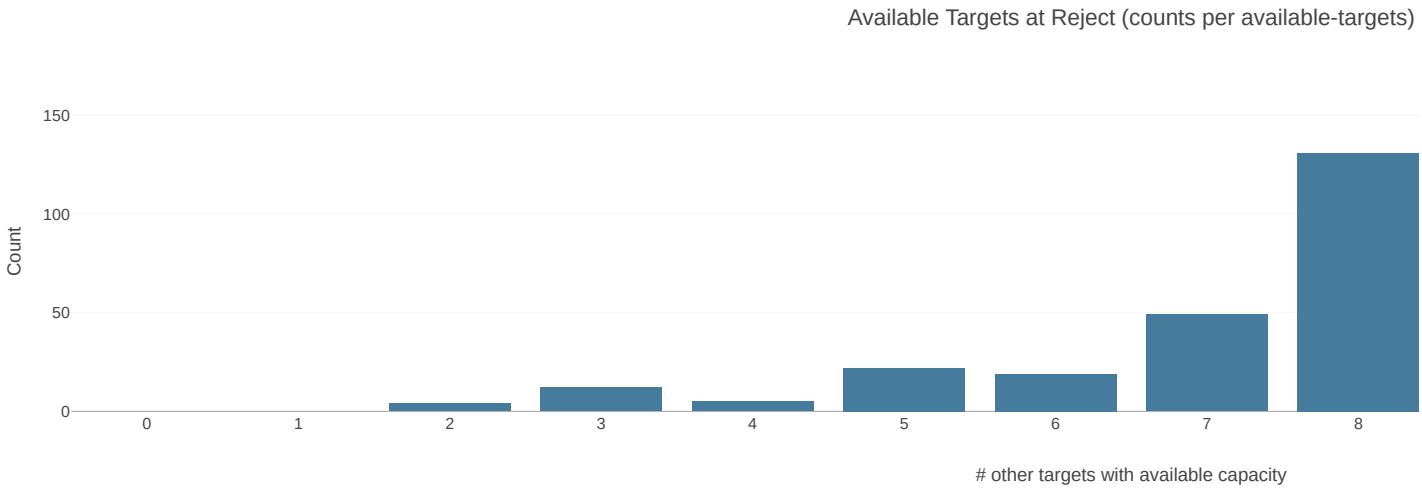




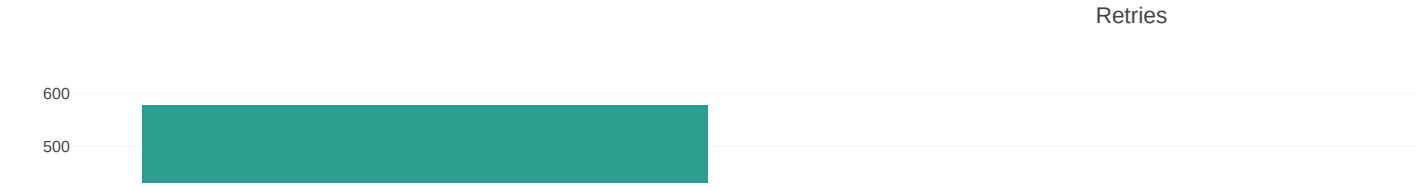
Free Slots Over Time

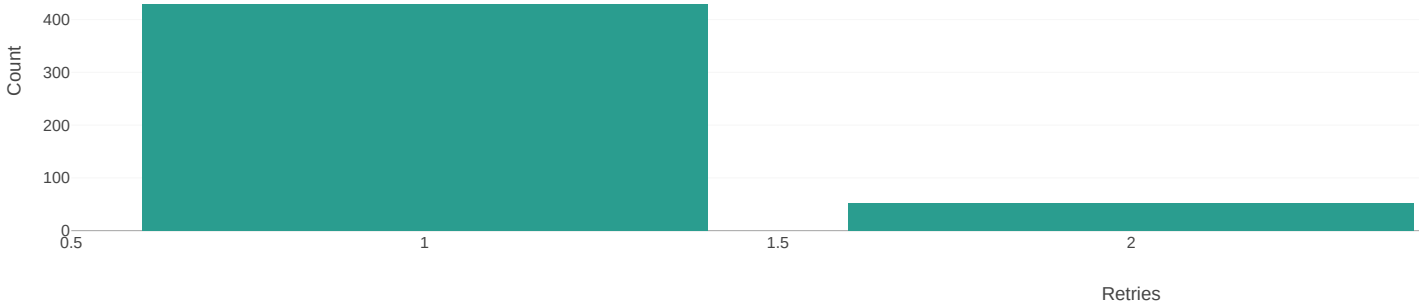


Available Targets at Reject

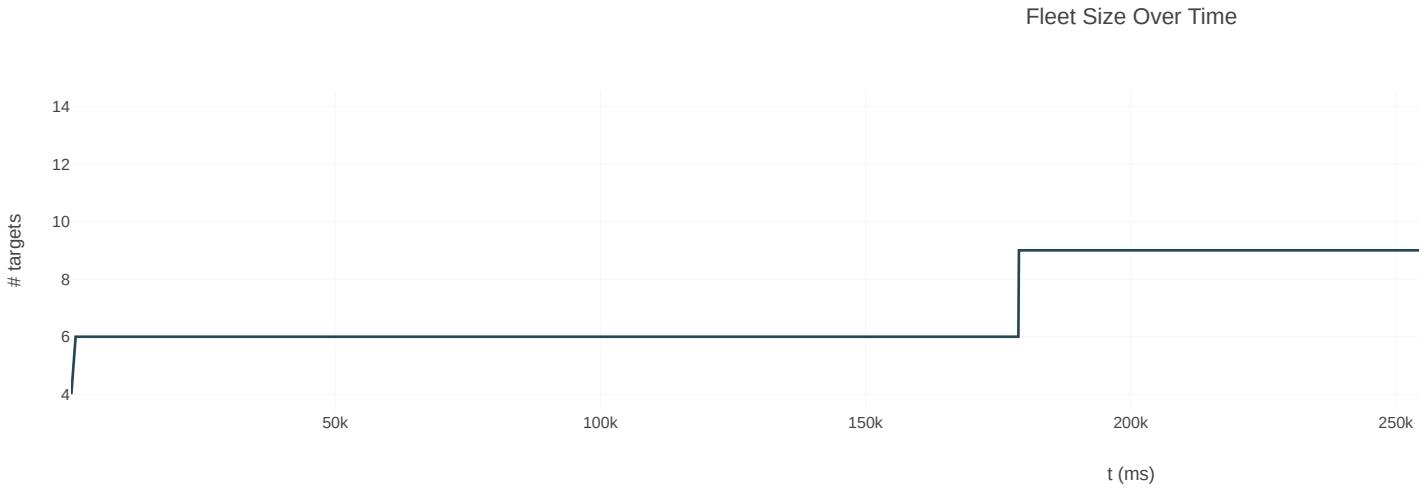


Retries Histogram

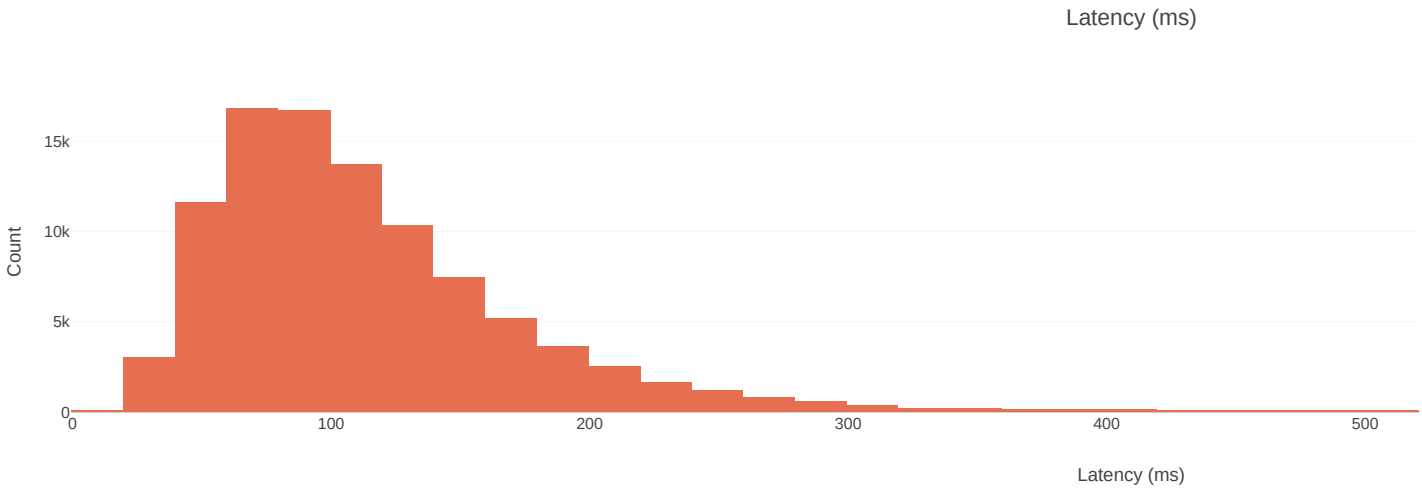




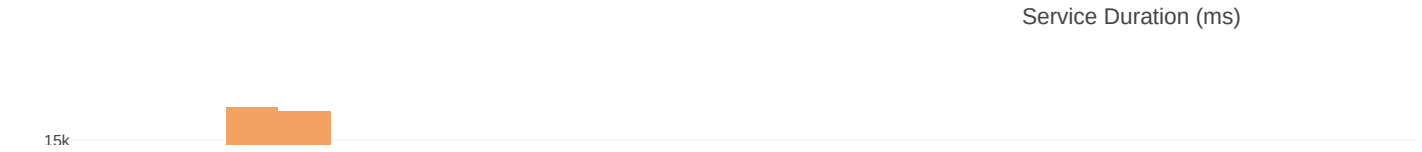
Fleet Size Over Time

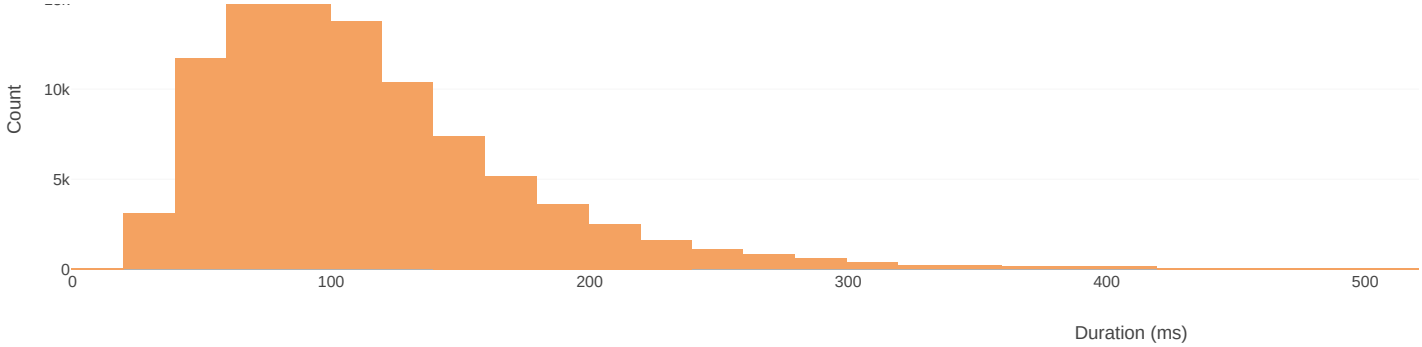


Latency Distribution (samples)



Service Duration Distribution (samples)





Numeric Tables

Retries Histogram

0	1	2	3
96062	576	52	8

Per-Target Concurrent Requests (percentiles)

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

Per-Target CPU Utilization (percentiles)

p100	p99	p90	p80	p50
0.9894403927232286	0.7995810617209733	0.6162635584179418	0.5317657507900985	0.3771452634791926

Fleet Size (percentiles)

p100	p99	p90	p80	p50
14.0	14.0	12.0	12.0	9.0

Request Duration (ms) (percentiles)

p100	p99	p90	p80	p50
815.0	319.0	189.0	152.0	100.0

Request Latency (ms) (percentiles)

p100	p99	p90	p80	p50
815.0	320.0	190.0	152.0	100.0

Sample Counts

num_decision_events	num_fleet_samples	num_target_concurrency_samples	num_target_cpu_util_samples
97402	97402	946683	946683

Per-Target Metrics (Tables)

Target 1

total_requests
10672

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9486119732787524	0.7960660612247189	0.6027678901728968	0.5192398921951547	0.3605968471083101

Request duration (ms) percentiles

p100	p99	p90	p80	p50
667.0	319.28999999999905	187.0	152.0	100.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
667.0	319.28999999999905	187.0	152.0	100.0

Target 2

total_requests
10678

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9716067266396096	0.7911533134546608	0.6006482014501218	0.5188480045316551	0.3651077857323377

Request duration (ms) percentiles

p100	p99	p90	p80	p50
663.0	317.0	190.0	153.0	100.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
663.0	317.0	190.0	153.0	100.0

Target 3

total_requests
10674

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9349718471666125	0.7896405581578121	0.6015058226487967	0.5213569455675634	0.3661887777813776

Request duration (ms) percentiles

p100	p99	p90	p80	p50
788.0	323.0	189.0	152.0	99.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
788.0	323.0	189.0	152.0	99.0

Target 4

total_requests
10684

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9515525330003984	0.7918590597161724	0.6113959561813231	0.5232131404804495	0.3632825853087366

Request duration (ms) percentiles

p100	p99	p90	p80	p50
676.0	321.1700000000001	191.0	152.0	100.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
676.0	321.1700000000001	191.0	152.0	100.0

Target 5

total_requests
10683

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9714650120772608	0.7899908761964479	0.6053161262195477	0.5189308561446491	0.3598164838198366

Request duration (ms) percentiles

p100	p99	p90	p80	p50
612.0	313.0	190.0	152.0	99.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
612.0	313.1800000000003	190.0	152.0	99.0

Target 6

total_requests
10686

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9330969213237443	0.801468913750308	0.6076905788764176	0.5202364802790566	0.3631667282413076

Request duration (ms) percentiles

p100	p99	p90	p80	p50
658.0	314.0	190.0	153.0	99.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
658.0	314.14999999999964	190.0	153.0	99.0

Target 7

total_requests
7333

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9239076257478785	0.79949890080298	0.6203281115271875	0.5380172074450542	0.38815125275500106

Request duration (ms) percentiles

p100	p99	p90	p80	p50
707.0	320.0	188.0	152.0	100.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
707.0	323.3600000000006	188.0	152.0	100.0

Target 8

total_requests
7335

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9381509529136394	0.8039434917250093	0.623955588018925	0.5406981481252358	0.3895192424479314

Request duration (ms) percentiles

p100	p99	p90	p80	p50
638.0	322.2999999999993	191.0	152.0	100.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
638.0	324.65999999999985	191.0	152.0	100.0

Target 9

total_requests
7338

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9894403927232286	0.809990717886121	0.6374863291171764	0.5480039461411149	0.3927468366777919

Request duration (ms) percentiles

p100	p99	p90	p80	p50
597.0	317.0	190.0	151.0	99.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
597.0	318.0	190.30000000000018	151.0	99.0

Target 10

total_requests
4043

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	6.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9527113096493349	0.8047886434333713	0.6463581947393456	0.5653640428597982	0.4067660125888114

Request duration (ms) percentiles

p100	p99	p90	p80	p50
656.0	334.7399999999998	187.80000000000018	152.0	100.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
712.0	336.5799999999999	188.0	152.0	100.0

Target 11

total_requests
3417

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	6.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9748874923460688	0.8185632492629884	0.6378299048175314	0.5605537207760968	0.4103282040087223

Request duration (ms) percentiles

p100	p99	p90	p80	p50
815.0	310.3600000000006	189.0	151.0	98.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
815.0	313.52000000000044	189.4000000000001	151.0	98.0

Target 12

total_requests
2003

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	6.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9388754282868773	0.8189111955346684	0.6526629666169458	0.5689739797320462	0.4141518148023976

Request duration (ms) percentiles

p100	p99	p90	p80	p50
542.0	316.9200000000001	196.79999999999995	156.0	99.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
615.0	318.96000000000004	197.0	156.0	99.0

Target 13

total_requests
692

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.9085655423214338	0.8202592958984388	0.6498042989641332	0.5419453488241638	0.384284611711701

Request duration (ms) percentiles

p100	p99	p90	p80	p50
519.0	346.6300000000002	193.79999999999995	145.80000000000007	98.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
702.0	353.18000000000006	194.0	145.80000000000007	98.0

Target 14

total_requests
460

Concurrency percentiles

p100	p99	p90	p80	p50
8.0	8.0	6.0	5.0	4.0

CPU utilization percentiles

p100	p99	p90	p80	p50
0.8784188762192504	0.8157500093293625	0.6209651343743195	0.5333602617421037	0.3789009833721443

Request duration (ms) percentiles

p100	p99	p90	p80	p50
531.0	287.50999999999965	184.10000000000002	147.20000000000005	98.0

Request latency (ms) percentiles

p100	p99	p90	p80	p50
582.0	296.04999999999984	185.20000000000005	148.0	98.0