Name: Songyang Li
ID: A53097969
Github: lsyucsd

Experiment 1: N independent requests

1. N=10

Mean service time = 96.9ms.

2. N=20

Mean service time = 88.2ms.

3. N=30

Mean service time = 91.7ms.

Experiment 2: P parallel requests, all must complete

1. N=10

Time to receive all responses = 129ms

2. N=20

Time to receive all responses = 131ms

3. N=30

Time to receive all responses = 144ms

Experiment 3: P parallel requests, 90% must complete

1. N=10

Time to receive first 90% of the responses = 103ms

2. N=20

Time to receive first 90% of the responses = 109ms

3. N=30

Time to receive first 90% of the responses = 122ms

Response Question:

In the experiment, we can emulate the server because the service time of the server is normal with some mean and sigma. And we generate the random number with the similar distribution. So I can get the conclusion based on my result. I think the tail latency will decrease the performance of a distributed system. From experiment 2 and 3, to receive the first 90% requests need less time than to receive 100% requests. Because if there is a big value, no matter how small other values are, the response time is still big. The only big value can dominate the whole performance of the system. And the effect of shedding the slowest 10% requests can improve the performance of the system. And in real systems, due to large scale and number of servers and clients, I think the benefit can be larger.

Original data:

Experiment 1: N independent requests

1. N=10, mean=96.9

83	86	96	99	140
86	119	113	69	78

2. N=20, mean=88.2

102	79	80	96	85
124	104	72	101	116
76	89	113	51	98
72	81	71	77	77

3. N=30, mean=91.7

104	69	73	74	99
97	107	79	89	100
78	67	71	77	107
129	102	129	97	92
76	103	124	47	81
102	111	65	94	107

Experiment 2, 3 (same data):

1. N=10, total time = 129, 90% time = 103

55	58	76	80	86
90	96	100	103	129

2. N=20, total time = 131, 90% time = 109

59	60	64	70	78
78	84	85	89	96
96	98	98	100	101
103	106	109	127	131

3. N=30, total time = 144, 90% time = 122

49	50	62	63	65
67	71	72	74	76
79	81	85	86	86
88	90	90	92	94
94	94	96	96	101
102	122	126	128	144

The way to run my code:

In the client, the “independent” is for experiment 1 and the “parallel” is for experiment 2 and 3. Both the client and the server are in each folder. To run, first make file, then:

Server: ./DelayMe <port number>

Example: ./DelayMe 9877

Client: ./DelayClient <number of requests> 127.0.0.1 <port number>

Example: ./DelayClient 10 127.0.0.1 9877