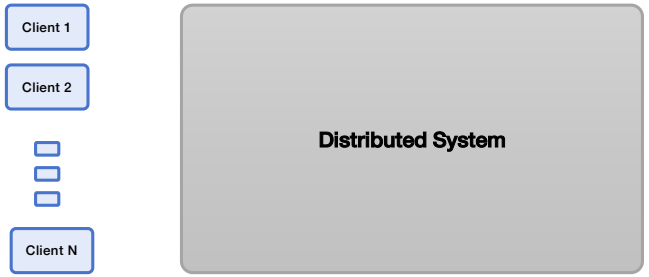
Reasoning about Performance of Distributed Systems

COS 418: Distributed Systems
Lecture 20

Mike Freedman

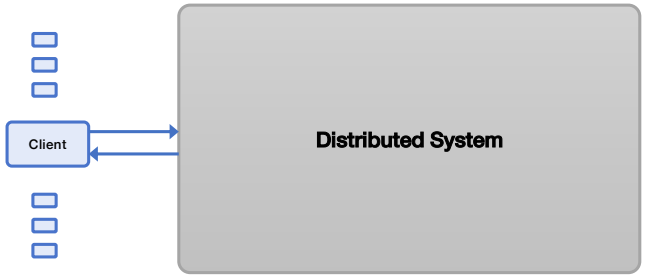
1

Measuring Distributed Systems



2

Measuring Distributed Systems



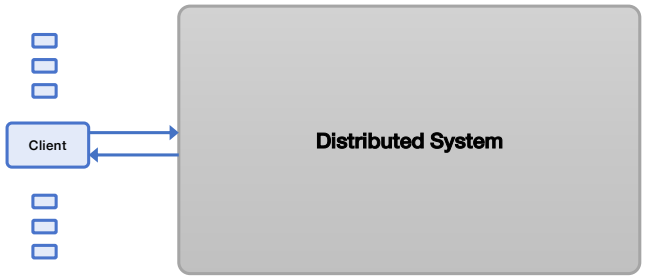
3

Latency

- How long a request takes to complete
- Measured externally from time request is sent until time response is received.

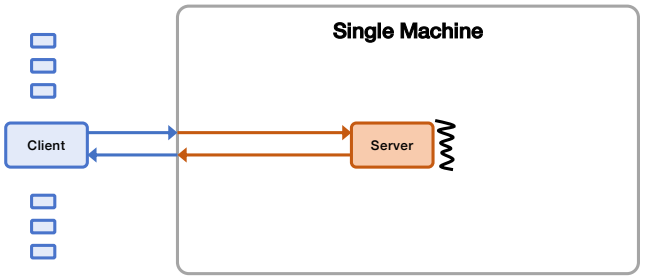
4

Latency, Measure Externally



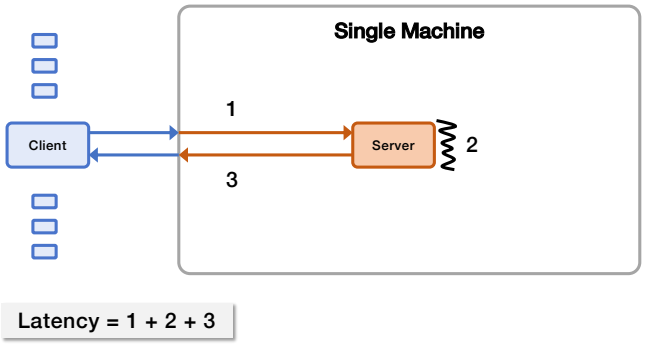
5

Latency, Reason Internally



6

Latency, Reason Internally



Latency = 1 + 2 + 3

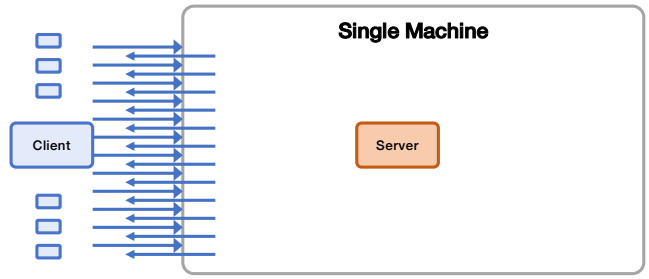
7

Throughput

- How many operations per unit time that a system can handle (typically ops / second)
- Measured externally as the rate that responses come out of the system

8

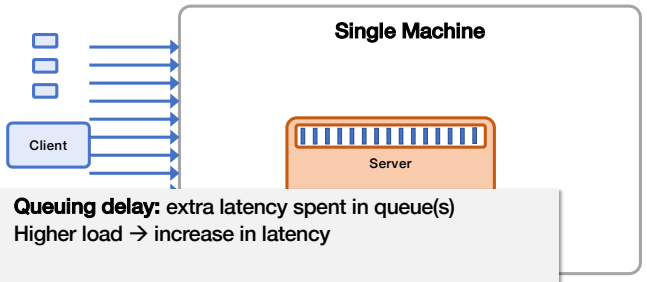
Max Throughput Example (Not Ideal)



Throughput = Number of (valid) responses received by all clients / End time – start time

9

Queuing Delay & Overload



- **Queuing delay:** extra latency spent in queue(s)
- Higher load → increase in latency

- **Overload:** offered load > max system throughput
  - Queues get really long
  - Other weird/bad things happen
  - → Observed throughput < max system throughput

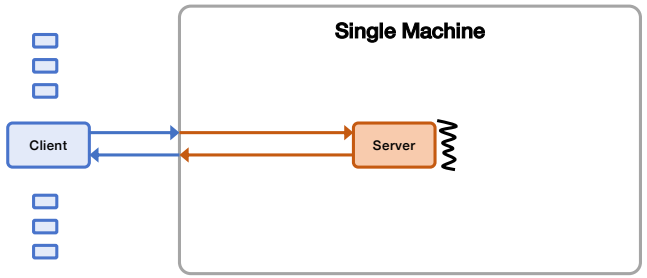
10

Measuring Throughput Method

1. Starting with low load
2. Increase load
3. Repeat until measured throughput stops increasing

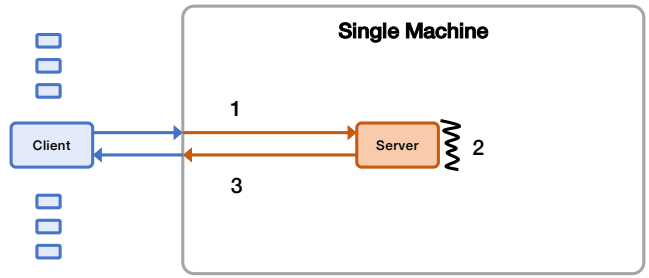
11

Throughput, Reason Internally



12

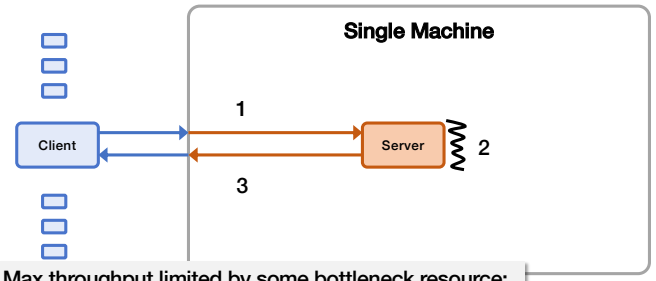
Throughput, Reason Internally



Throughput = min (1, 2, 3)

13

Throughput Bottlenecks (simplified)



Max throughput limited by some bottleneck resource:
1) Incoming bandwidth
2) Server CPU
3) Outgoing bandwidth

14

Load Generation

- Closed-loop
  - Each "client" sends one request, waits for the response to come back, and then sends another request
  - More "clients" => more load
- Open-loop
  - Load is generated independently of the response rate of the system, typically from a probability distribution
  - More directly control the load on the system
- Which one is more realistic?
- We'll reason using closed-loop clients

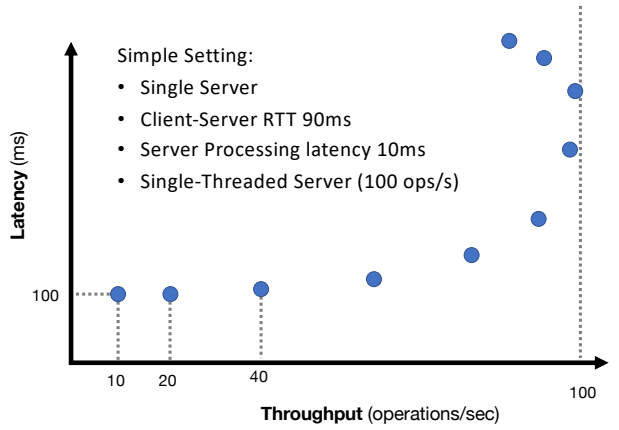
15

Mental Experimental Setup

- Start with 1 closed-loop client
  - Expected latency?
  - Expected throughput?
- Double number of closed-loop clients
  - Expected increase in latency?
  - Expected increase in throughput?
- Repeat

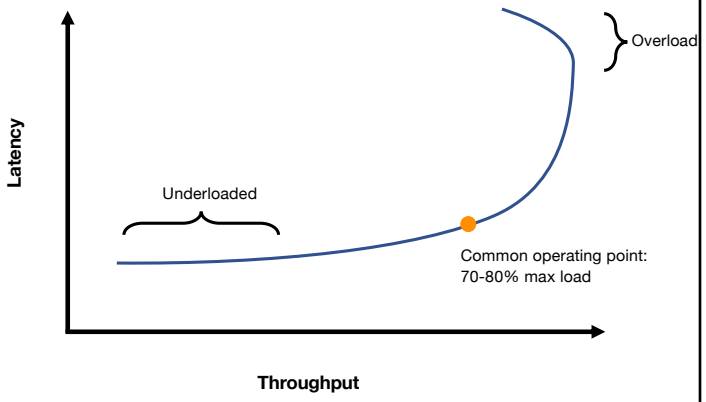
16

Throughput-Latency Graph



17

Throughput-Latency Graph



18

Throughput / Latency Relationship

- Proportional at low load … but not high load
- Because measured throughput is a function of latency
  - i.e., throughput bottleneck is offered load
- Related, but you should reason about both
- For system A vs system B, all are possible:
  - A has lower latency and higher throughput than B
  - A has lower latency and lower throughput than B
  - A has higher latency and lower throughput than B
  - A has higher latency and higher throughput than B

19

Evaluation in Minutes not Months

- Reasoning using your mental model is much much faster than really doing it
- What would happen if?
  - I moved my servers from the San Jose datacenter to Oregon?
  - I switch from c5.xlarges to c5.24xlarges for my servers?
  - I doubled the number of servers?
  - I switch from system design X to system design Y?
    - replace single server with Paxos-replicated system?
    - replace Paxos with eventually consistent design?
    - add batching?
    - replace Paxos with new variant?

20

Let's use these tools!

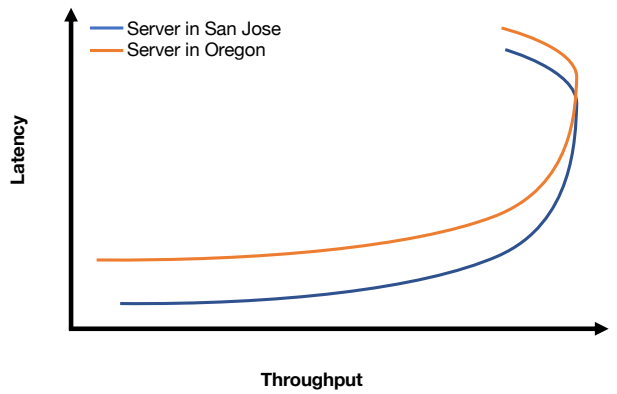
21

Mental Experimental Setup

- System A versus System B
- From 1 to N closed-loop clients loading each
- Compare throughput and latency

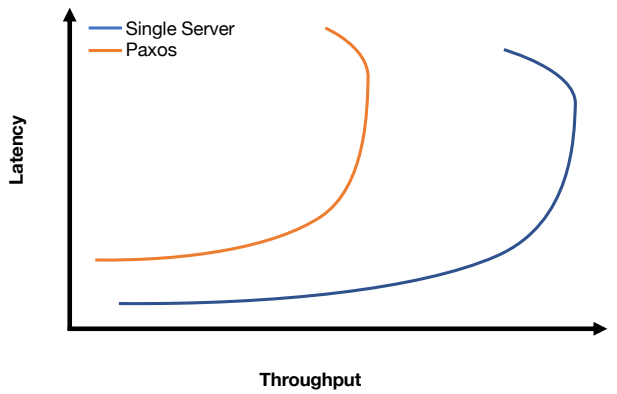
22

Move Single Server from San Jose to Oregon (Clients in San Jose)



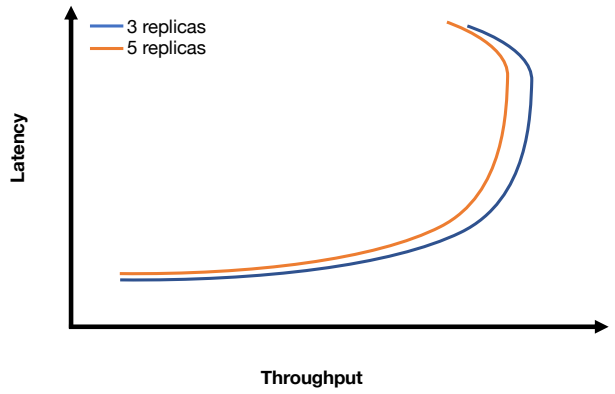
23

Replace Single Server with Paxos (Clients and servers in same datacenter, 3 replicas)



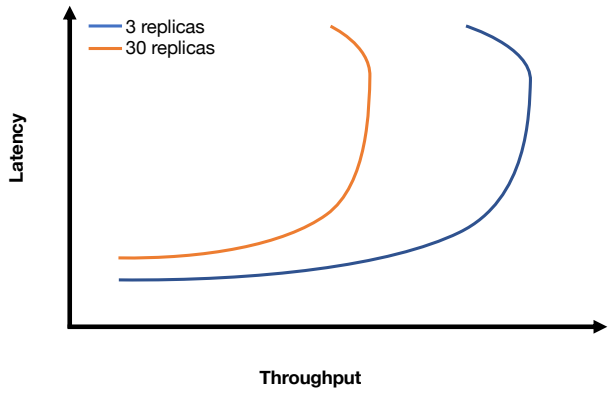
24

Paxos: 3 replicas to 5 replicas
(Clients and servers in same datacenter)



25

Paxos: 3 replicas to 30 replicas
(Clients and servers in same datacenter)



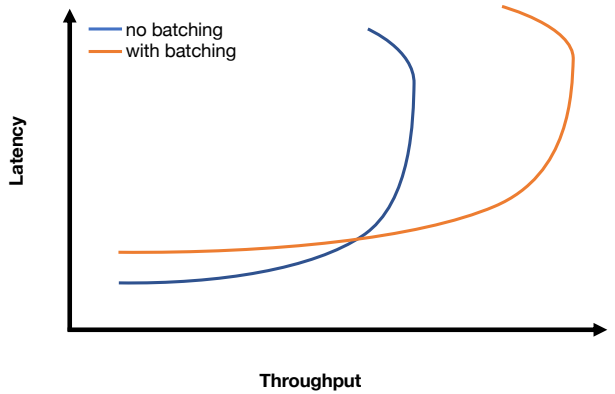
26

Batching

- Group together multiple operations
- Improves throughput, e.g.,
  - Marshall data together
  - Send to network layer together
  - Unmarshall data together
  - Handle group of operations together
- Delay processing/sending ops to increase batch size
  - Common way to trade an increase in latency for increase in throughput

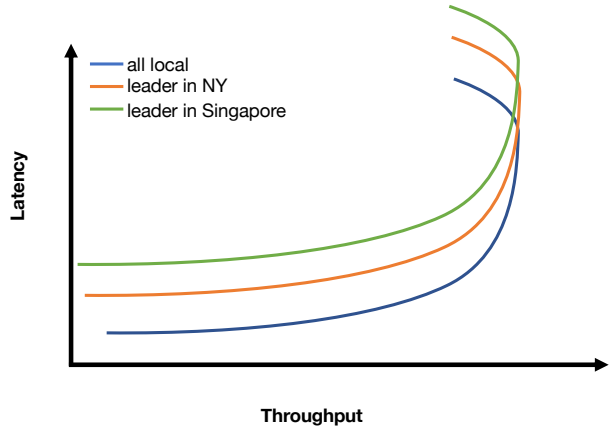
27

Paxos with batching
(Clients and servers in same datacenter, 3 replicas)



28

Paxos: 3 local replicas to geo-replicated (Clients in NY; replicas in NY, Oregon, Singapore



29

Summary

- Measure distributed systems externally
- Latency: how long operations take
- Throughput: how many operations/sec
- Reason about latency and throughput using internal knowledge of system design
  - (and back-of-the-envelope calculations)
- Reason about effects on latency and throughput from changes to system choice, deployment, design
  - Critical tool in system design

30