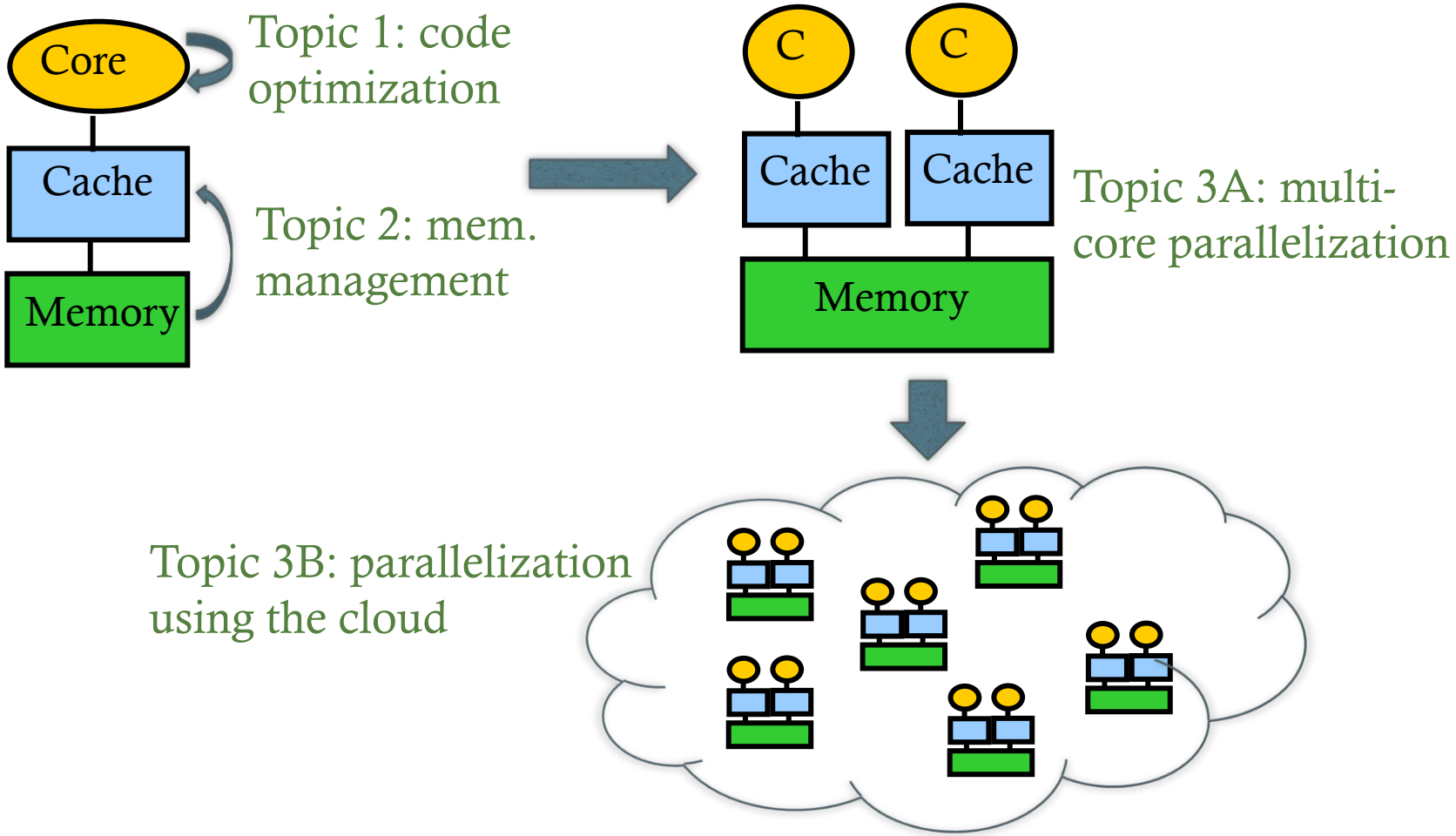
ECE 454 Computer Systems Programming

What is Performance?

Ashvin Goel, Ding Yuan
ECE Dept, University of Toronto

Review: The Big Picture



Before we go on…

- What do you exactly mean by “performance”?
  - For a simple program, it is speed -- how fast your program runs
  - Use the Unix “time” command to measure performance
- What about a server program?
  - Is “speed” the only important thing?
  - What is the “speed” for long running programs?
  - Latency vs. throughput

Latency vs. Throughput

- Latency
  - How fast does the server respond to my request?
    - Sometimes also called response time
- Throughput
  - Number of requests served/unit time
- Relationship?

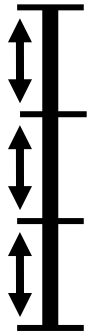
Positive Correlation Example

```
void dummy_server () {
  while (request = next_request()) {
    respond(request);
  }
}
```

Latency for req. 1

Latency for req. 2

Latency for req. 3

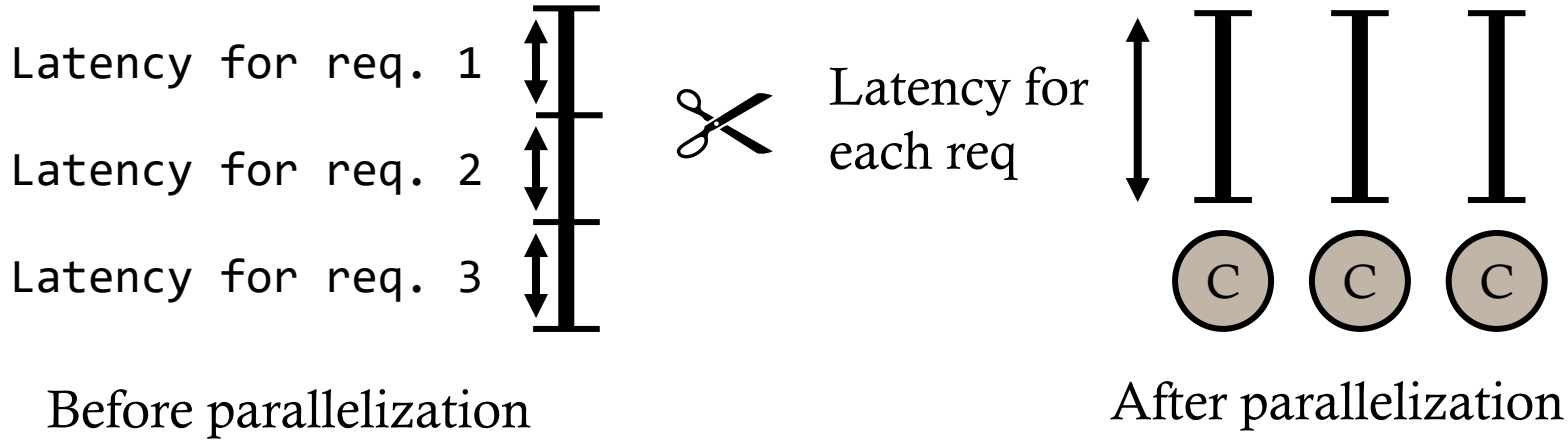


$$\text{Throughput} = \frac{3}{L1 + L2 + L3}$$

If we have a faster CPU, both latency and throughput will improve (smaller latency, higher throughput)!

Negative Correlation Example

```
void dummy_server () {
  while (request = next_request()) {
    respond(request);
  }
}
```



Throughput will be better (higher), why?
Latency will be worse (larger), why?

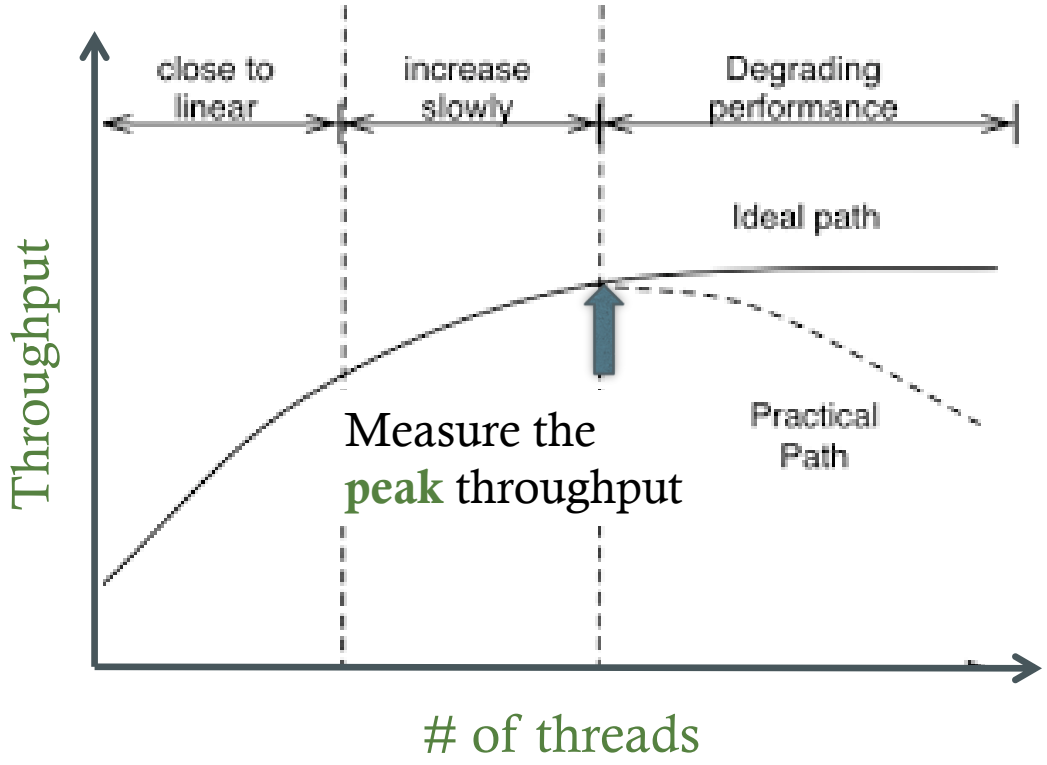
Real Life Analogy

Plane	Toronto to Paris	Speed	Passenger	Throughput (pmph)
Boeing 747	8 hours	610 mph	470	286,700
Concorde	4 hours	1350 mph	132	178,200

Which plane has higher performance?

Parallelism vs. Throughput

- Will more parallelism always improve throughput?



Performance Measurement is a Complicated Problem

- Many other metrics: utilization, goodput, jitter, etc.
- Extra considerations: best case? worst case? average?
- Different applications have different requirements
  - Google/Facebook/Amazon
  - Online gaming
  - Netflix
  - Flight control software on airplane
- ACM special interest group on performance evaluation (SIGMETRICS)