ECE 454
Computer Systems Programming

What is Performance?

Ashvin Goel
ECE Dept, University of Toronto
http://www.eecg.toronto.edu/~ashvin
Review: The Big Picture

Topic 1: code optimization

Core

Cache

Memory

Topic 2: mem. management

Cache

Memory

Topic 3A: multi-core parallelization

Topic 3B: parallelization using the cloud
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?
void dummy_server () {
    while (request = next_request()) {
        respond(request);
    }
}

Latency for req. 1
Latency for req. 2
Latency for req. 3

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

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

Before parallelization

Latency for req. 1
Latency for req. 2
Latency for req. 3

Before parallelization

Latency for each req

C  C  C

After parallelization

Throughput will be better (higher), why?
Latency will be worse (larger), why?
### Real Life Analogy

<table>
<thead>
<tr>
<th>Plane</th>
<th>Toronto to Paris</th>
<th>Speed</th>
<th>Passenger</th>
<th>Throughput (pmph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boeing 747</td>
<td>8 hours</td>
<td>610 mph</td>
<td>470</td>
<td>286,700</td>
</tr>
<tr>
<td>Concorde</td>
<td>4 hours</td>
<td>1350 mph</td>
<td>132</td>
<td>178,200</td>
</tr>
</tbody>
</table>

Which plane has higher performance?
Parallelism vs. Throughput

- Will more parallelism always improve throughput?

Measure the peak throughput

Throughput

# of threads
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)