Before we go on…

- What do you exactly mean by “performance”?
  - Simple program: speed -- how fast your program runs
  - Unix “time” command

- 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 the server respond my request?
    - Sometimes also called response time

- Throughput
  - Number of requests served/unit time

- Relationship?

Positive correlation example

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

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

Throughput = \[
\frac{3}{L1 + L2 + L3}
\] = req./sec.

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 ()) {
        thread_create/respond, request);
    }
}
```

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

<table>
<thead>
<tr>
<th>Latency for req. 1</th>
<th>Latency for each req</th>
<th>Latency for req. 2</th>
<th>Latency for req. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before parallelization</td>
<td>After parallelization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Real life analogy

<table>
<thead>
<tr>
<th>Plane</th>
<th>Toronto to Paris</th>
<th>Speed</th>
<th>Passengers</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*?

```
  +----------------+      +----------------+      +----------------+
  |                |      |                |      |                |
  | throughput     |      | throughput     |      | throughput     |
  | close to linear|      | increase slowly|      | Degrading      |
  |               |      |                 |      | performance    |
  | Ideal path    |      | Practical Path |      |                |
  +-------+-------+-------+-------+-------+-------+-------+
  |       |       |       |       |       |       |       |
  | Measure the peak throughput |       |       |       |       |       |       |
  +-------+-------+-------+-------+-------+-------+-------+
  | # of threads |
```

Performance measurement is a very complicated problem

- Other metrics: bandwidth, jitter, etc.
- Extra considerations: best case? worst case? average?
- Different applications have different requirements
  - Netflix
  - Google/Facebook/Amazon
  - Online gaming
- ACM special interest group on performance evaluation (SIGMETRICS)