On Assignment #2

Important: run each test a large number of times to get rid of timing cost and random errors.

- Question #1:
  - the cost of function call is around 6-7ns.
  - common error: timing each function call.

- Question #2:
  - the cost of a context switch is around 0.8us.

- Question #3:
  - process switching cost is around 3-4us based my suggested "pipe-based tests".

- Question #4:
  - thread switching cost is around 7us based my suggested "condition variable"-based tests.
  - sort of strange since thread switching should be faster; probably because the locking overhead is significant.

On Assignment #3

Measure the throughput and mean response time of a concurrent server.

- The server listens at port 5678 and it accepts simple requests.
- Run it in a Linux machine in the graduate software lab and test it using your client from a different machine.
- Your job is to write your client and do the measurement.
- I will tell you what the server does after the assignment is due.

- Due next Thursday but it should be a simple one.
Content Distribution Network

- CDNs sit between content providers and content consumer.
- A CDN contains hundreds of servers throughout Internet.
- CDN replicates its customers' content in CDN servers. When provider updates content, CDN updates servers.

How CDN works?

Content provider
- www.yahoo.com

CDN company
- replicates content at CDN servers
- uses its authoritative DNS server to redirect requests to a suitable CDN server

How CDN works? (An Illustration)

Determining a Suitable CDN Server

Heuristics for "suitability":
- Network proximity
- Runtime workload
- Locality
Network Proximity
- CDN creates a "map", indicating distances from local networks and CDN nodes
- When a DNS query arrives at the authoritative DNS server:
  - Server determines local networks from which query originates
  - Uses "map" to determine a nearby CDN server
- How to get this "map"
  - Offline: "ping" each network several times and take the average
  - Online: continuously "ping" each network at certain intervals and take the exponentially-weighted moving average

Runtime Workload
- Monitoring workload at request distributors
  - Continuously poll each CDN server for workload at certain intervals and maintain the EWMA
  - Randomly poll a few CDN servers for just-in-time load information
- Request distribution policy is based on a combination of load and network proximity

Locality
- Reduce the working set of each CDN server such that all pages can be directly served from memory.
- How?
  - Control the size of replication server set for each Web object.
  - E.g., for a CDN with 100 servers, each object is only replicated at 10 servers. The exact choice of this 10 servers depends on a hash value of the object URL.
- Does this conflict with network proximity?

Summary
Somehow conflicting heuristics for "suitability":
- Network proximity
- Runtime workload
- Locality
  - Many possible strategies based on a combination of the above.

How to measure their effectiveness?
- Based on simulation
  - Simulating both the network and the servers
    - We need network topology and link properties
    - We need to simulate server memory/disk subsystem
  - We need workload - traces?