Computer Architecture Discussion Group notes for April 8, 2004

Discussion Leader: Everybody
Notes: Samir Chopra
Topic: Latency.

Paper: D. Patterson, "Latency Lags Bandwidth" *Not published yet.*

Summary

Bandwidth has improved at a much higher rate than latency. The paper gives examples of processor, memory, LAN, and disk milestones, along with bandwidth and latency data at each milestone. In general, every time bandwidth doubles, latency decreases by a factor of 1.2 to 1.4. The paper goes on to suggest reasons for lagging latency, and how to cope with it.

Discussion

- Relative latency is increasing – more instructions are required to cover up the latency.
- Perhaps a more interesting observation would be the ratio of latency cycles vs. total clock cycles instead of absolute time.
- Effective bandwidth would be more interesting. However using maximum bandwidth data does make sense so we can see the large difference in latency vs. bandwidth, and take notice of the number of 'spare' cycles.
- Designers should plan with latency in mind. Maybe even trade bandwidth for latency?
- Why is latency lagging?
    - Software overhead: The amount of data that software touches is growing exponentially.
    - Software use is increasing at a rate greater than Moore's law.
- Decreasing latency naturally increases bandwidth.
- Human response remains the same, so why is latency so important? Example where it is: Web searching. Also computer response time is directly related to worker productivity.
- Poor latency does not allow you to scale for speed.
- Are we improving bandwidth because we can? Are we not improving on latency quite as much because we can't?
- **Bandwidth is more marketable than latency.**
- Will latency become a bottleneck in the future? Yes.
- Increased latency means that resources are held longer.
- Why did Dave Patterson write this paper?
    - It's an eye opener.
    - Good collection of data in one place.

    - If you are not working on processors, pay attention to the problems that processor people have run into. You will run into these problems in the future.
- Kelly begins to swear. The group quickly replaces her cussing with the word "data".
- Why don't we reduce bandwidth demands instead of wasting available bandwidth, then trying to increase it because we do not have enough? We should not think that bandwidth is free and abundant.
- The paper should add power trends, as well as costs.