Lecture 2: Metrics to Evaluate Systems

- Topics: Metrics: power, energy, reliability, cost, performance, summarizing performance with AM, GM, HM

- My email: rajeev@cs
- TA office hour details on class webpage
- HW1 posted later today, due Wed Aug 31st (note auto extension)
- Lecture break
- Class resources (notes)

Where Are We Headed?

Modern trends:

- Clock speed improvements are slowing (power constraints)
- Difficult to further optimize a single core for performance
- Multi-cores: each new processor generation will accommodate more cores
- Need better programming models and efficient execution for multi-threaded applications
- Reduced data movement
- Need better memory hierarchies
- Need greater energy efficiency
- Dark silicon, accelerators
- Emergence of new workloads: ML, graphs, genomics
- Emergence of new metrics: security, reliability

Power Consumption Trends

- Dyn power α activity x capacitance x voltage^2 x frequency
- Capacitance per transistor and voltage are decreasing, but number of transistors is increasing at a faster rate, and Dennard scaling has ended; hence clock frequency must be kept steady
- Leakage power is also rising; is a function of transistor count, leakage current, and supply voltage
- Power consumption is already between 100-150W in high-performance processors today
- Energy = power x time = (dynpower + lkgpower) x time

Problem 1

- For a processor running at 100% utilization at 100 W, 20% of the power is attributed to leakage. What is the total power dissipation when the processor is running at 50% utilization?

Problem 1

- For a processor running at 100% utilization at 100 W, 20% of the power is attributed to leakage. What is the total power dissipation when the processor is running at 50% utilization?

Total power = dynamic power + leakage power
= 80W x 50% + 20W
= 60W

Solve for 0% utilization; the system consumes 20W.
This is the basis for "server consolidation" in datacenters (move processes so you have a few highly utilized servers).

Power Vs. Energy

- Energy tells us the true “cost” of performing a fixed task
- Power (energy/time) poses constraints; can only work fast enough to max out the power delivery or cooling solution
- If processor A consumes 1.2x the power of processor B, but finishes the task in 30% less time, its relative energy is 1.2 X 0.7 = 0.84; Proc-A is better, assuming that 1.2x power can be supported by the system

Problem 2

- If processor A consumes 1.4x the power of processor B, but finishes the task in 20% less time, which processor would you pick:
  (a) if you were constrained by power delivery constraints?
  (b) if you were trying to minimize energy per operation?
  (c) if you were trying to minimize response times?

Problem 2

- If processor A consumes 1.4x the power of processor B, but finishes the task in 20% less time, which processor would you pick:
  (a) if you were constrained by power delivery constraints?
  Proc-B
  (b) if you were trying to minimize energy per operation?
  Proc-A is 1.4x0.8 = 1.12 times the energy of Proc-B
  (c) if you were trying to minimize response times?
  Proc-A is faster, but we could scale up the frequency (and power) of Proc-B and match Proc-A's response time (while still doing better in terms of power and energy) (only if the circuits can handle the faster clock)

Reducing Power and Energy

- Can gate off transistors that are inactive (reduces leakage)
- Design for typical case and throttle down when activity exceeds a threshold
- DFS: Dynamic frequency scaling -- only reduces frequency and dynamic power, but hurts energy
- DVFS: Dynamic voltage and frequency scaling – can reduce voltage and frequency by (say) 10%; can slow a program by (say) 8%, but reduce dynamic power by 27%, reduce total power by (say) 23%, reduce total energy by 17%
  (Note: voltage drop → slow transistor → freq drop)

Problem 3

- Processor-A at 3 GHz consumes 80 W of dynamic power and 20 W of static power. It completes a program in 20 seconds.
  What is the energy consumption if I scale frequency down by 20%?

  What is the energy consumption if I scale frequency and voltage down by 20%?

Problem 3

- Processor-A at 3 GHz consumes 80 W of dynamic power and 20 W of static power. It completes a program in 20 seconds.
  What is the energy consumption if I scale frequency down by 20%?
  New dynamic power = 64W; New static power = 20W
  New execution time = 25 secs (assuming CPU-bound)
  Energy = 84 W x 25 secs = 2100 Joules

  What is the energy consumption if I scale frequency and voltage down by 20%?
  New dynamic power = 41W; New static power = 16W;
  New exec time = 25 secs; Energy = 1425 Joules

Other Technology Trends

- DRAM density increases by 40-60% per year, latency has reduced by 33% in 10 years (the memory wall!), bandwidth improves twice as fast as latency decreases
- Disk density improves by 100% every year, latency improvement similar to DRAM
- Emergence of NVRAM technologies that can provide a bridge between DRAM and hard disk drives
- Also, growing concerns over reliability (since transistors are smaller, operating at low voltages, and there are so many of them)

Defining Reliability and Availability

- A system toggles between
  - Service accomplishment: service matches specifications
  - Service interruption: services deviates from specs
- The toggle is caused by *failures* and *restorations*
- Reliability measures continuous service accomplishment and is usually expressed as mean time to failure (MTTF)
- Availability measures fraction of time that service matches specifications, expressed as MTTF / (MTTF + MTTR)

Cost

- Cost is determined by many factors: volume, yield, manufacturing maturity, processing steps, etc.
- One important determinant: area of the chip
- Small area ➔ more chips per wafer
- Small area ➔ one defect leads us to discard a small-area chip, i.e., yield goes up
- Roughly speaking, half the area ➔ one-third the cost