Midterm 2 Study Guide

Chapter 4

• Taxonomy of parallel architectures
  o What are SISD, SIMD, MISD, and MIMD?
• Amdahl’s law and speedup equations
  o Give a percentage of a program that is parallelizable, calculate the speedup obtained using different numbers of CPUs
• Uniform memory access vs. non-uniform memory access
• Centralized shared memory model vs distributed memory model
  o Advantages and disadvantages
• What is cache coherency?
  o Why does this problem exist?
• Private data vs. shared data
• Cache coherency schemes provide migration and replication of shared data items. What is migration and replication?
• What are the two cache coherency protocols that we discussed? How are they similar and how are they different? What are the advantages and disadvantages of each, if any?
• Discuss the basic idea of the snooping protocol and directory based protocols
  o How do the work
  o What information is stored for each cache and memory block?
  o Know what all the states are and how the transitions work between each state (all of the state diagrams in the lecture slides)
  o In either protocol, how does a processor see the state of memory?
• What is false and true sharing and how are they similar and the same. Include discussion of block size
• What is an atomic operation and why is it necessary for sharing data?

Chapter 5

• 11 advanced cache optimizations – what are they and how do they improve cache performance? Do they always improve performance or does it depend on the benchmark?
  • Small and simple caches to reduce hit time
  • Way prediction to reduce hit time
  • Trace caches to reduce hit time
  • Pipelined cache access to increase cache bandwidth
  • Nonblocking caches to increase cache bandwidth
  • Multibanked caches to increase cache bandwidth
  • Critical word first and early restart to reduce miss penalty
  • Merging write buffer to reduce miss penalty
  • Compiler optimizations to reduce miss rate
    • Code and data rearrangement
    • Loop interchange
    • Blocking
  • Hardware prefetching of instructions and data to reduce miss penalty or miss rate
  • Compiler controlled prefetching to reduce miss penalty or miss rate
• The table on page 309 summarizes all of the optimization techniques and tells you which aspect it effects
• Memory technology and optimizations
  • How are SRAMs and DRAMs layed out? How do they work? How are they different? What are the advantages and disadvantages to one over the other?
  • Describe how DRAMs are accessed i.e. address is passed in 2 pieces
  • How can locality be used to improve the performance of DRAMs?
  • What is DDR SDRAM?
• Protection: Virtual memory and virtual machines
  • How does virtual memory provide protect? What protections are provided?
  • What architectural support is needed for virtual memory?
  • Why have virtual machines become popular recently?
  • What types of protection does a virtual machine offer?
  • What is a virtual machine?
  • When running a virtual machine, describe how the system is laid out in terms of VM, VMM and Host os?
  • What is a s systems virtual machine?
  • What is the virtual machine monitor? What is it responsible for? What are its requirements?
• How do virtual machines assist in managing both software and hardware?
• What is virtualization?
• How does lack of support in the ISS affect virtualization overhead?
• Discuss how different running modes are important for the VM and VMM
• Why can a VM not execute privileged instructions? What are privileged instructions and how are they handled when a VM tries to execute them?
• Why is I/O so difficult in VMs? How does a VM access physical devices on a machine?
• Discuss the issues with virtual memory and virtual machines. What is the added overhead? How can that overhead be minimized?

Chapter 6
• Why has the topic of storage become so popular recently?
• Areal density
• Concept of difference in whole disk read time for random access vs sequential access
• RAID
  o What is the concept of RAID? Why is it important? Why is it useful?
  o Give any possible advantages/disadvantages to using RAID X. If I were to ask you this question, I would say what RAID X does to remind you
  o How do different RAID methods perform for little and big writes?
  o Know the differences between the following RAID models. The table on page 363 might be helpful
    • RAID 1 - mirrored
    • RAID 4 – parity-based with one parity disk
    • RAID 5 – parity-based with the parity spread across all disks
    • RAID 6 – row and diagonal parity
  o How can RAID 6 recover from multiple disk failures? Work through a recovery problem like in the slides
• Errors, faults and failures
  o Define error, fault and failure and how do those differ?
  o Given an example situation, determine if it is an error, fault or failure
  o What is a latent error?
  o Four fault categories and what they are
    • Hardware faults
    • Design faults
    • Operation faults
    • Environmental faults
  o Three types of faults
    • Transient faults
    • Intermittent faults
    • Permanent faults
  o Why are operator faults so hard to quantify?
• I/O performance, reliability measures and benchmarks
  o Know the basic producer consumer model from page 372
  o Measures of I/O performance:
    • How many devices can you connect
    • Which I/O devices can you connect
    • Response time
    • Throughput
    • Interference of I/O with processor execution
  o Difference between throughput and response time
  o Transaction time is made up of
    • Entry time
    • System response time
    • Think time
  o Transaction processing benchmarks
    • Mostly concerned with I/O rate over data rate
    • TPC benchmark characteristics on page 375
    • Why must the data set scale in size with the throughput?
    • Figure 6.14 – Know the differences in these reconstruction policies.
• Queuing Theory
  o Give a basic definition of queuing theory. What is it useful for? What does it tell us? What types of systems does it measure? Etc
What is a system that is in equilibrium?
Little’s law
Terms on page 381
What is the “mean time to complete service of a task when a new task arrives if the server is busy?”
  - Why is this term hard to measure? How is it measured in queuing theory?
  - What is a Poisson distribution?
  - How can a histogram give is a characterization of a set of data?
  - What does memoryless mean in the context of distributions?
Know the assumptions of our model on page 386
What is an M/M/1 model?
What is an M/M/m model?
Be able to solve problems like those in the examples on pages 382, 387,