Effect of Memory Speed on Microprocessors

Michael Kirzinger

Introduction

- Main memory speed is one of the bottlenecks in computer systems today
- To reduce the bottleneck, the latency must be reduced and/or bandwidth increased (depending on application)
- Two main methods to increase bandwidth:
  1. Increase memory bus width
  2. Increase memory speed
- Will investigate the effect of both methods of increasing bandwidth as well as how the bandwidth bottleneck compares to the latency bottleneck

Processor

- The processor used in SimpleScalar was the default parameters with the caches and branch predictor modified and slightly increased functional units.
- Cache and branch predictor structures and sizes resemble those on an AMD Athlon64 processor

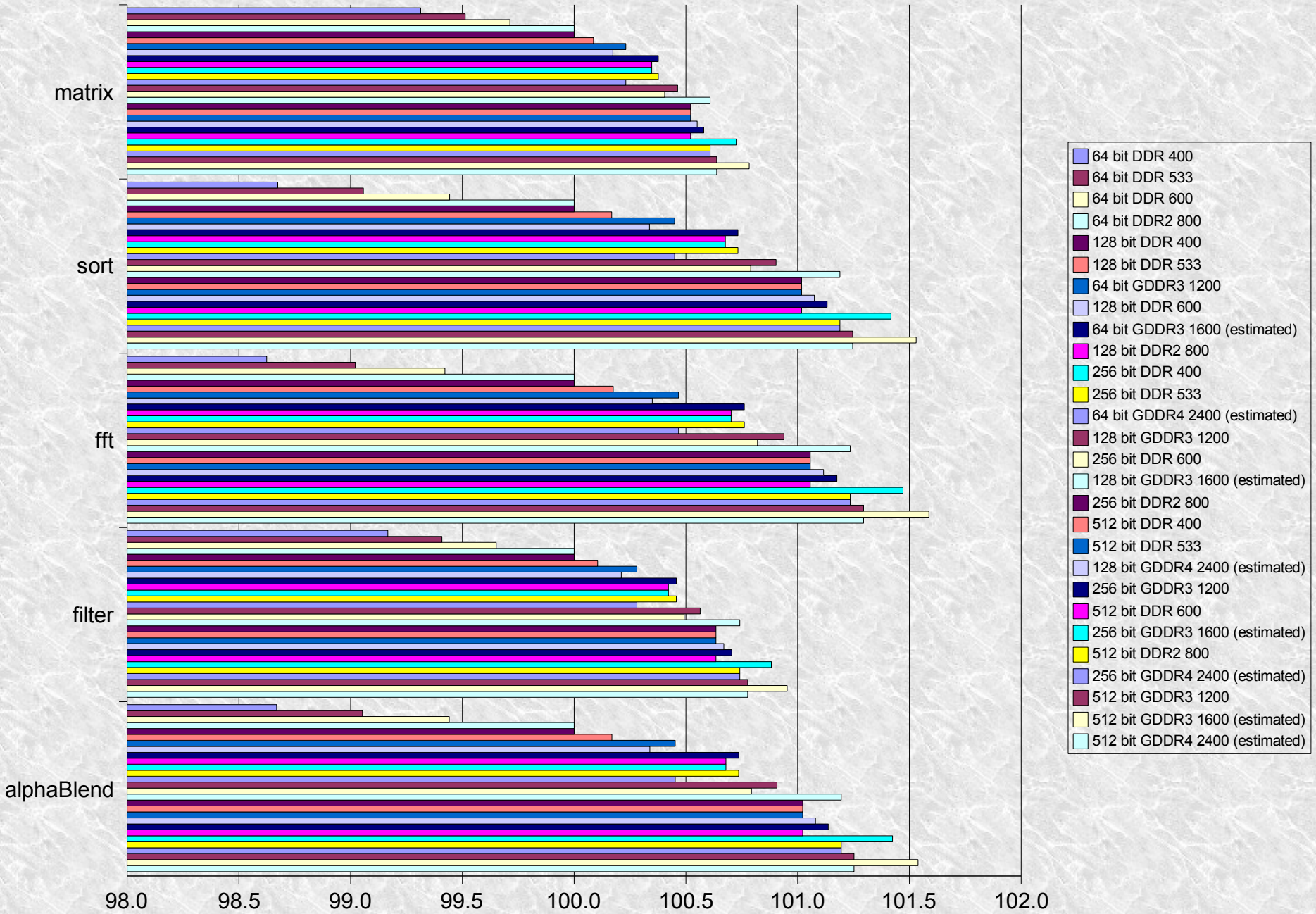
Benchmark Programs

- matrix
  - Does matrix multiplication on two 100x100 matrices (modified program from first architecture assignment)
- sort
  - Sorts a 10000 element integer array (program from second architecture assignment, just with a larger array)
- fft
  - Does a Complex fast fourier transform on a 65536 element array
- filter
  - Does a simulation of a signal (8192 samples) passing through a finite impulse response filter (order = 100)
- alphaBlend
  - Simulates alpha blending two 24 bit 640x480 bitmaps

Simulation I

- Assumed a 2400 MHz clock to calculate memory latencies to use in SimpleScalar
- Used numbers based on actual memory modules
  - OCZ DDR and DDR2
  - Samsung GDDR3 and GDDR4
- Varied bus width from 64 bits to 512 bits
- General result was that increasing bandwidth increases performance, except in some cases where the increased bandwidth is overshadowed by the increase in latency

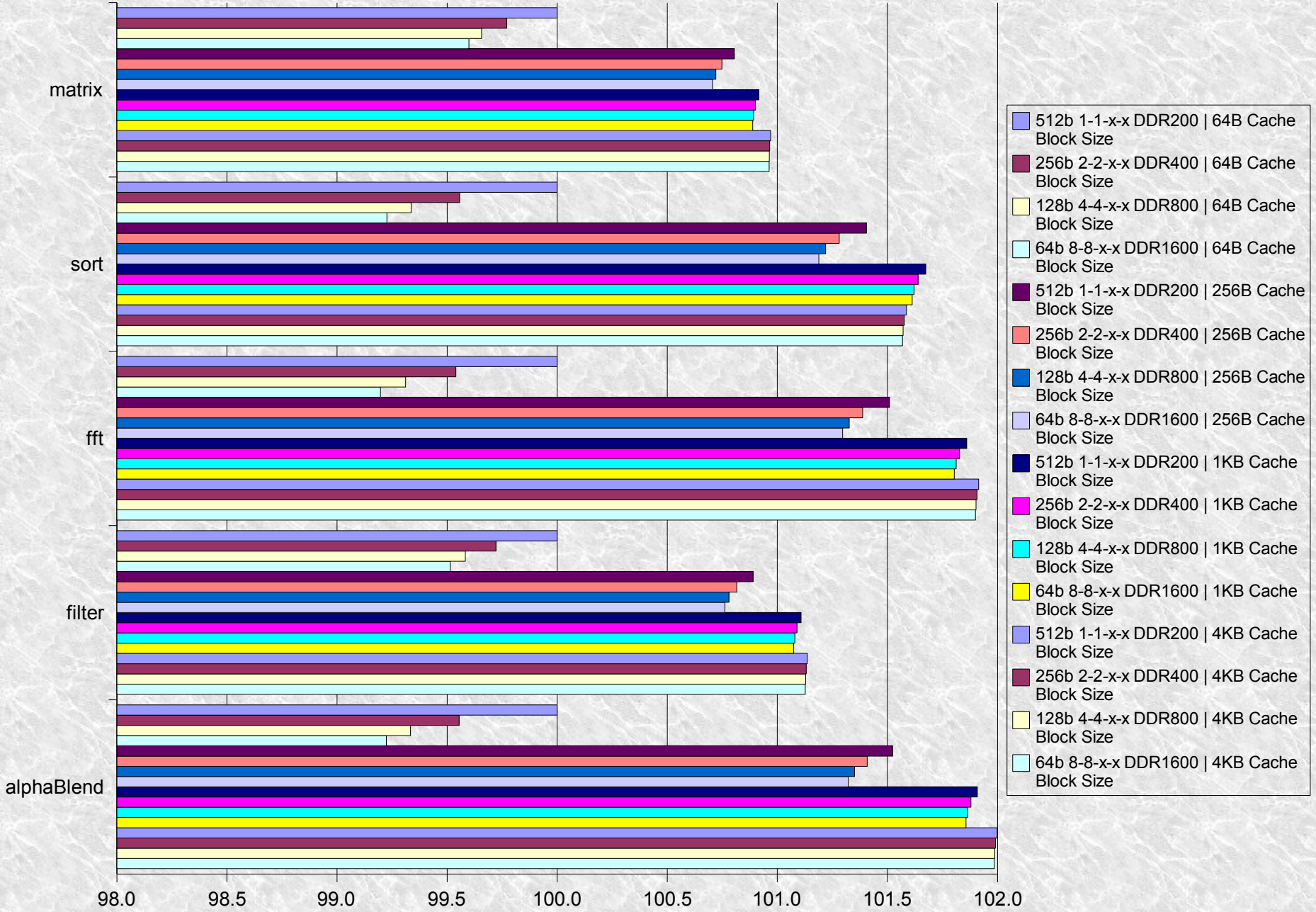
Relative Performance of Memory Modules



Simulation II

- Assumed a CPU clock of 3200 MHz for this simulation
- Goal was to compare bus width vs memory speed with a constant bandwidth
- Memory access latency (first chunk) constant, set second chunk latency according to speed
- Varied L2 cache block size from 64B to 4KB (constant L2 cache size of 4MB)
- Results showed that a larger bus width gives better performance than a higher clock speed. The higher clock speed memory needs a large block size to approach performance of memory with a wide bus

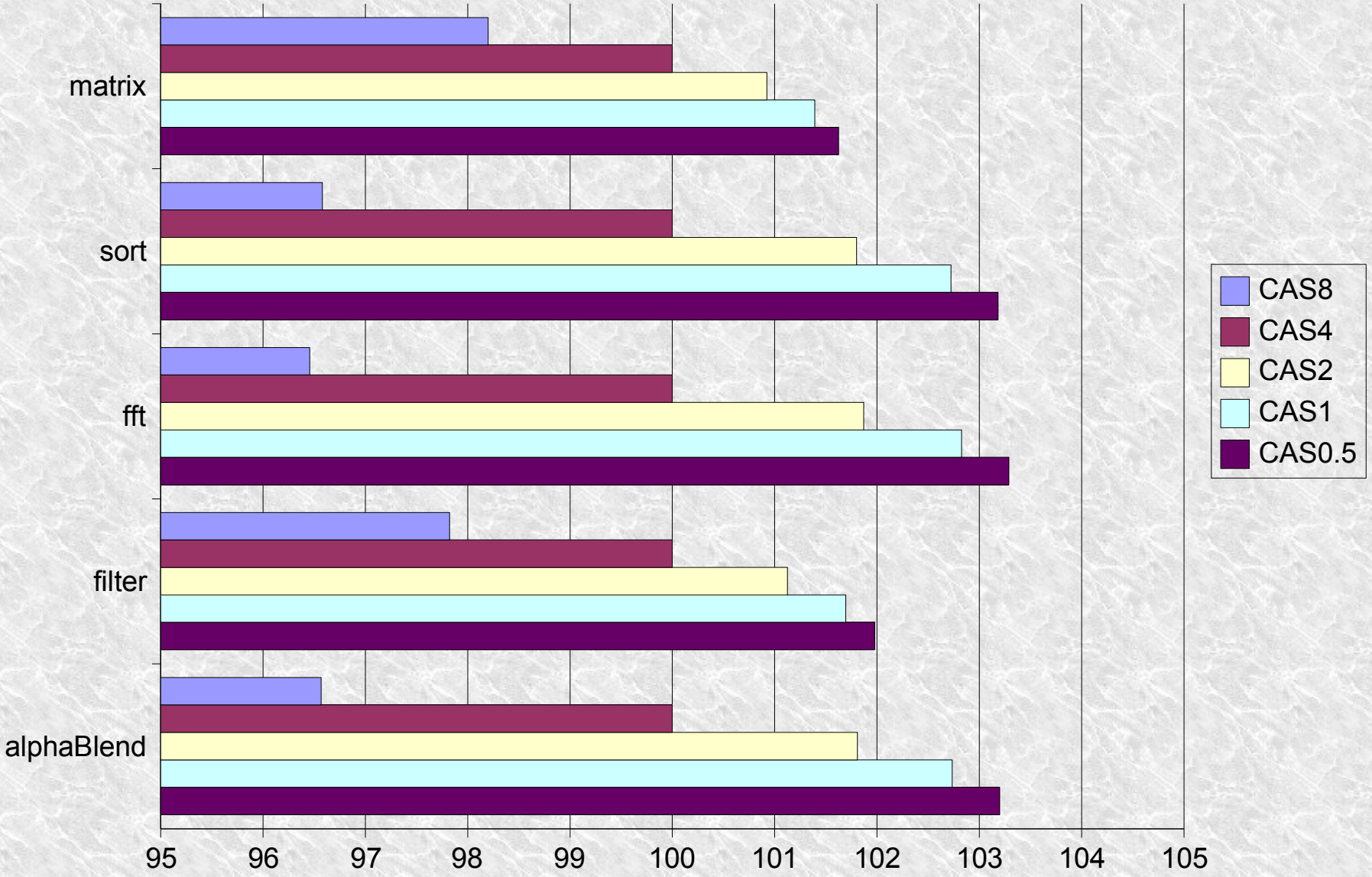
Relative Performance (12800MBps Memory Bandwidth)



Simulation 3

- CPU clock once again 3200MHz
- Test effect of varying first chunk latency (equivalent to varying CAS/tRAS)
- Most timings in this simulation are much faster than existing memory provides
- Reducing the first chunk latency has a larger effect than reducing the second chunk latency (the result of increasing clock speed)

Relative Performance (Constant Bandwidth)



Conclusions

- In systems with relatively small cache block sizes, it is better to up the memory bus width (up to the point where it is as wide as the cache block size) than the clock speed
- Reducing the initial latency creates a larger performance increase than increasing clock speed
- It is better to use lower latency memory, than slightly faster memory with a much larger latency

Questions?