Tarantula: a vector extension to the alpha architecture

by Roger Espasa, et al; 2002

presented by Matthew Fulmer

ECE 259, Spring 2010

April 7, 2010

Introduction

- Tarantula is an extension of an Alpha processor with a rather large vector unit
- Direct L2 Access
- Vector Mask as a Register

Fast L2 Fetching (Section 4)

- L2 cache has 16 independent banks; use them
- General case: sort the addresses to be fetched into 16 buckets and generate slices from the top element of each bucket
- Stride optimization: Sort arithmetically, bypassing the bucket sorter.
- Stride-1 optimization: Fetch all the data at once; it forms a stride if aligned properly

Cache Coherence Concerns (Section 4)

- For cache misses; one may wish to treat the vector operation as an atomic read/write of multiple cache lines
- Normal (scalar) and vector instructions have different load/store queues and different cache access paths, and thus need to stay coherent with each other

Power (Section 5)

- Estimated 3.4x better at flops/watt metric than the equivalent multicore solution (no analysis given)

Performance

- 10-20x speedup on highly data-parallel applications
- 2-4x improvement on some applications
- no slowdowns
- caveat: Highly optimized scientific workloads only

Innovations

- A virtual world that accessible only to members of CPS 214 at Duke

Questions

- What about vector instructions makes them especially latency-tolerant?