

Media Workloads and DLP

Systems Lunch

September 9, 2006

Andrew Hilton

Media Workloads

- Increasingly important
- Images
 - Large 2D arrays of pixels
- Sound
 - Stream of amplitudes
- Video
 - Sequential Images + Sound

Important Considerations

- Performance
- Power (Heat)
- Energy (Battery Life)

How to get performance?

Clock Frequency

- Memory Latency ↑
- Power ↑↑
- IPC ↓

Traditional Favorite

Parallelism

- Memory Latency -
- Power ↑

Increasingly Popular

Types of Parallelism: ILP

- Fine grained
- Scaling Problems
 - Bypassing
 - Dependency Checking
 - Register File Ports
- Hardware/Software can find
- Useful: Large scheduling scope

Types of Parallelism: TLP

- Coarse grained
- SMT
 - Cover cache miss
 - Trade Latency/Throughput
- SMP/CMP
- No automatic threading of code

Types of Parallelism: DLP

- Same operation/different data
 - Scientific (long) Vectors
 - SIMD instructions (short vectors)
- Scales better than ILP
- Need compiler/programmer to locate

ALPBench Paper

- 5 application benchmark suite
 - MPEG decoder
 - MPEG encoder
 - Face Recognizer
 - Ray Tracer
 - Speech Recognizer
- Characterizes the parallelism

ALPBench: TLP

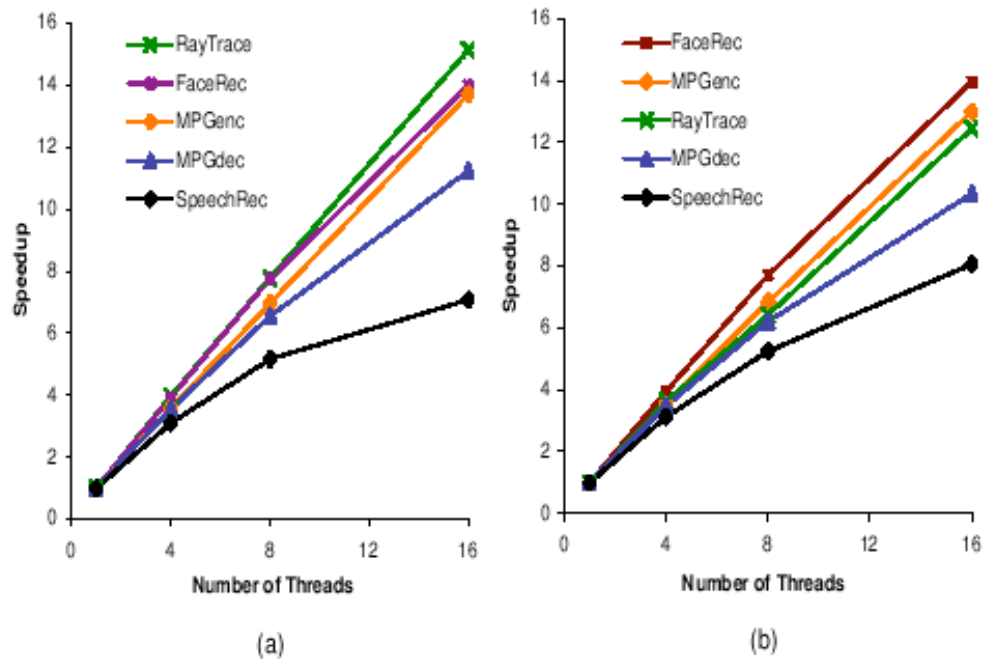


Figure 1: Scalability of TLP without SIMD instructions (a) with an ideal 1-cycle memory system, and (b) with realistic memory parameters (as in Table 1).

ALPBench: DLP

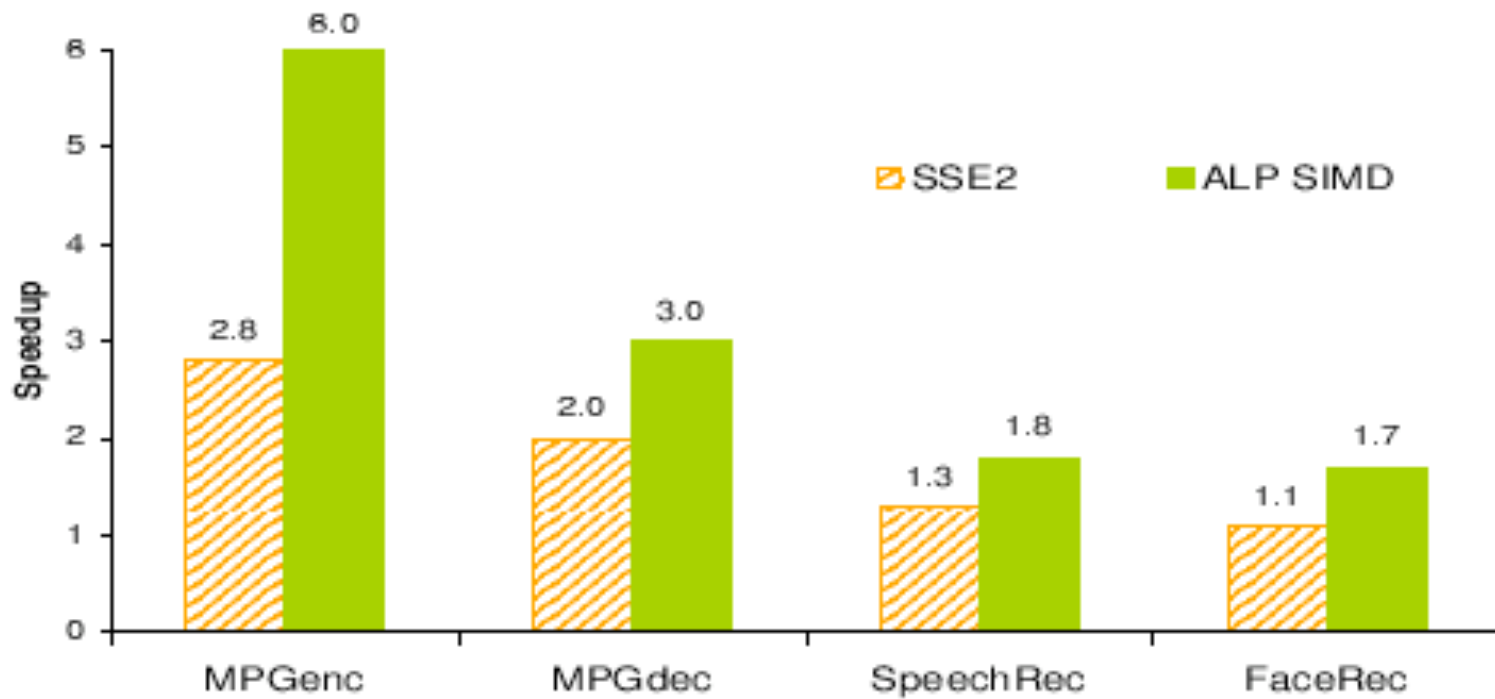


Figure 2: Speedup with SSE2 and ALP SIMD.

ALPBench: ILP

Cache Misses/Memory Bound



	ALPSim		P4 (Pentium 4)	
App	Base	SIMD	Base	SIMD
MPGenc	1.20	1.23 [4.24]	1.45 (1.87)	0.70 (1.03)
MPGdec	1.38	1.17 [3.31]	1.26 (1.73)	0.73 (1.14)
RayTrace	1.33	N/A	0.48 (0.73)	N/A
SpeechRec	0.35	0.39 [0.67]	0.38 (0.57)	0.34 (0.45)
FaceRec	0.32	0.30 [0.48]	0.51 (0.61)	0.43 (0.47)

Table 3: Instructions per cycle achieved on ALPSim and P4 for single-thread applications. For the ALP SIMD case, the number of sub-word operations retired per cycle is also given within square brackets. For P4, x86 micro-instructions per cycle is given in parenthesis.

Interaction Between Different Type of Parallelism

- DLP/TLP
 - DLP reduces parallel portion
- TLP/ILP
 - Cache behavior
 - Memory Bandwidth
- DLP/ILP
 - DLP combines independent instructions
 - Increased scheduling scope
 - Increased resource utilization
 - Reduction in harder to predict branches

Memory

- SIMD reduces compute/memory ratio
- More MLP
- TLP => memory bandwidth

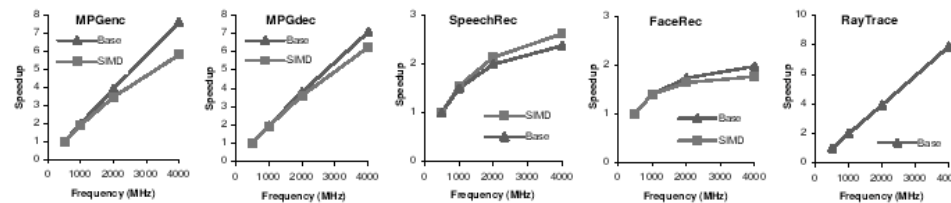


Figure 5: Frequency scalability for single thread applications. The SIMD data are with ALP SIMD. RayTrace does not have a SIMD version.

4.5.3 Memory Bandwidth

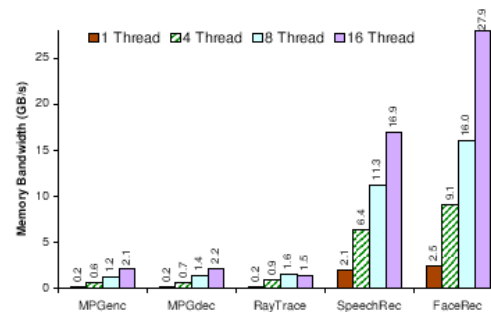


Figure 6: Memory bandwidth (in GB/s) at 4 GHz without SIMD.

What about...

- Frequency scaling with multiple threads?
- Memory bandwidth requirements for SIMD versions?

Conclusions

- Importance of media applications
- Parallelism for performance
- 3 types of parallelism in ALPBench
- Interaction of 3 types