Parallelizing the MPEG-2 Video Decompression for Simultaneous Multithreaded Processors

Theo Ungerer

Dept. of Computer Design and Fault Tolerance
University of Karlsruhe
D-76128 Karlsruhe, Germany
ungerer@Informatik.Uni-Karlsruhe.de
http://goethe.ira.uka.de/people/ungerer/

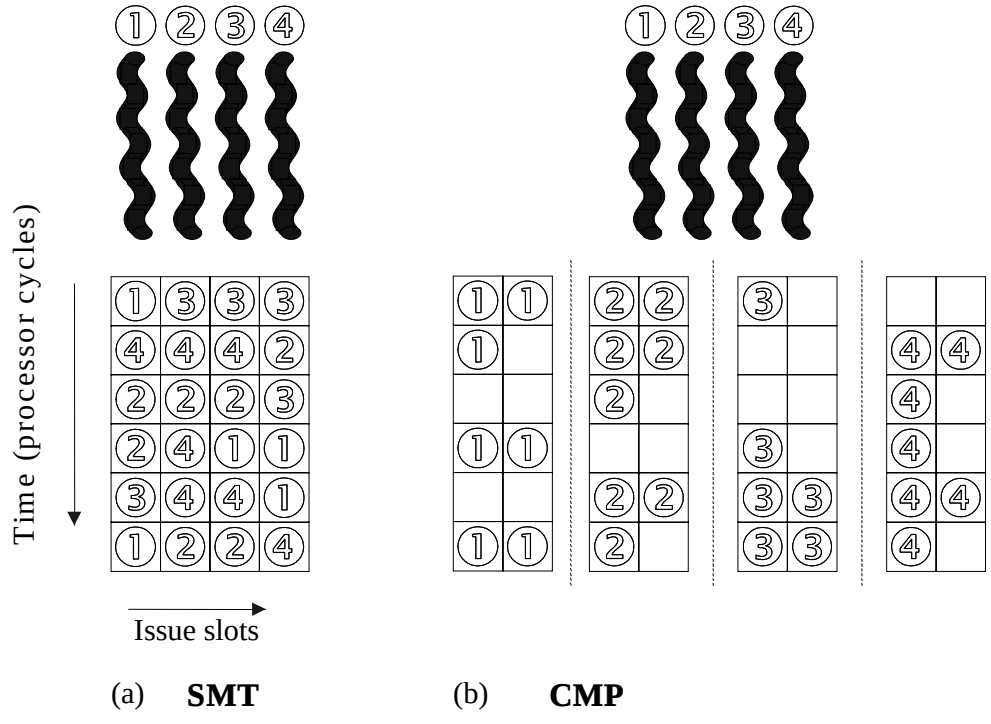
Outline of the Presentation

- Simultaneous Multithreading (SMT)
- Application workload: MPEG-2 made multithreaded
- The SMT multimedia processor model
- Simulator
- Performance results
- Conclusions

Simultaneous Multithreading

- **Multithreading**
  - The ability to pursue two or more threads of control in parallel within a processor pipeline.
  - Advantage: The latencies that arise in the computation of a single instruction stream are filled by computations of another thread.
- **Simultaneous multithreading (SMT)**
  - combines *wide issue superscalar* with *multithreading,*
  - issues instructions from several threads simultaneously.

Simultaneous Multithreading (SMT) and Chip Multiprocessors (CMP)



(a) **SMT** (b) **CMP**

Simultaneous Multithreading

- **State of research**
  - SMT is simulated and evaluated with Spec92, Spec95, and with database transaction and decision support workloads
  - Mostly unrelated programs are loaded in the thread slots!
  - Typical result: 8-threaded SMT reaches a two- to threefold IPC increase over single-threaded superscalar.
- **State of industrial development**
  - DEC/Compaq announced Alpha EV8 (21464) as 4-threaded 8-wide superscalar SMT processor at Microprocessor Forum October 1999

Simultaneous Multithreading

- **Different unrelated programs as workload**
  - ⇒ Throughput of a multiprogramming workload is increased!
  - What happens to single application programs?
  - *Ortega*, *Martel*, *Krishnan*, *Ayguade*, *Valero* (PACT99): hand parallelized SPECints reach speedups of 10-120% over single-threaded ex.
  - *Marcuello*, *Gonzales* (1998): "Speculative multithreaded processor" reaches average speedups of 25% for Integer codes and 80% for FP codes
- **Our goals**
  - Increase performance of single application programs!
  - Combine SMT with multimedia units and evaluate with a hand optimized multithreaded multimedia workload!
  - ⇒ We choose MPEG-2 decoder and made it multithreaded

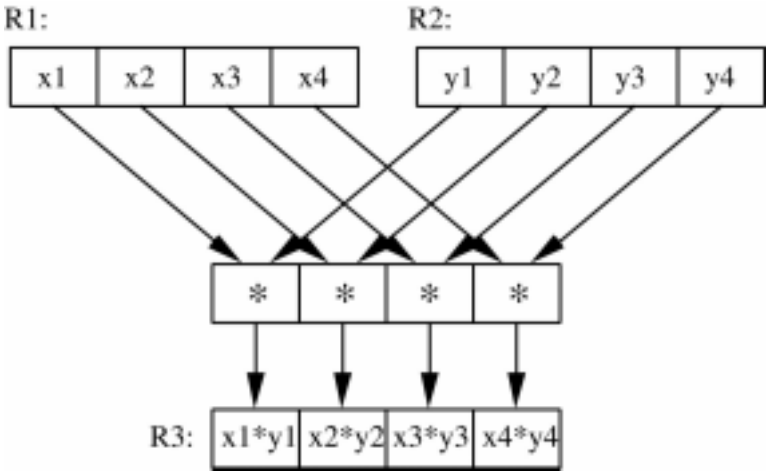
MPEG-2 Decoding

- **MPEG-2 is the video compression/decompression standard for digital TV, DVD, etc.**

MPEG-2 Decompression Steps:	**Type of parallism**
❖ Header decode	none
❖ Huffman decoding of macro blocks	none
❖ Inverse quantization	SIMD & coarse
❖ IDCT	SIMD & coarse
❖ Motion compensation	SIMD & coarse
❖ Display	none

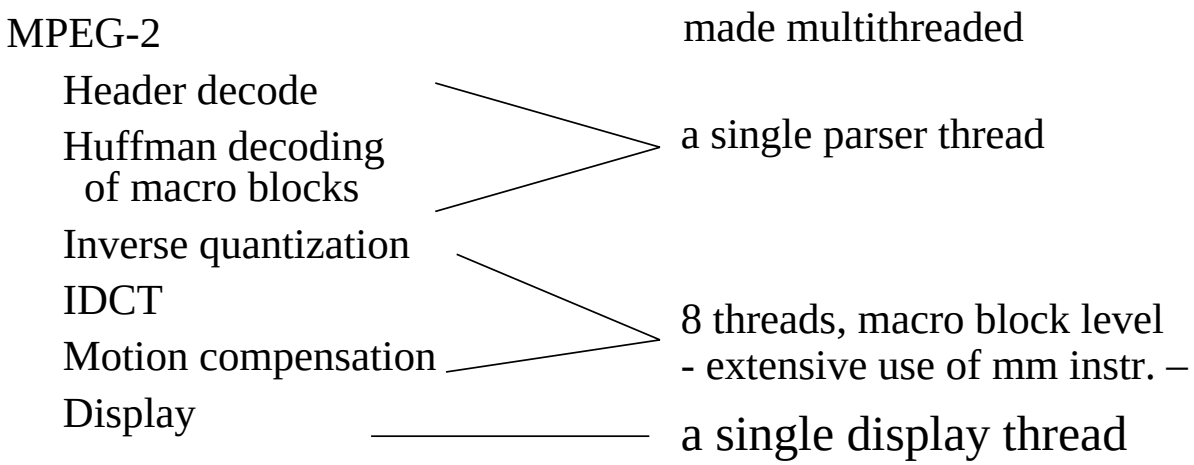
Multimedia Unit

- Utilization of subword parallelism (data parallel instructions, SIMD)



- Saturation arithmetic
- Additional arithmetic, masking and selection, reordering and conversion instructions

Multithreaded MPEG-2



sequential part: 15.5 %

theoretical speed-up:at most 6.5

Application Workload: Multithreaded MPEG-2

Average Usage of Instructions:

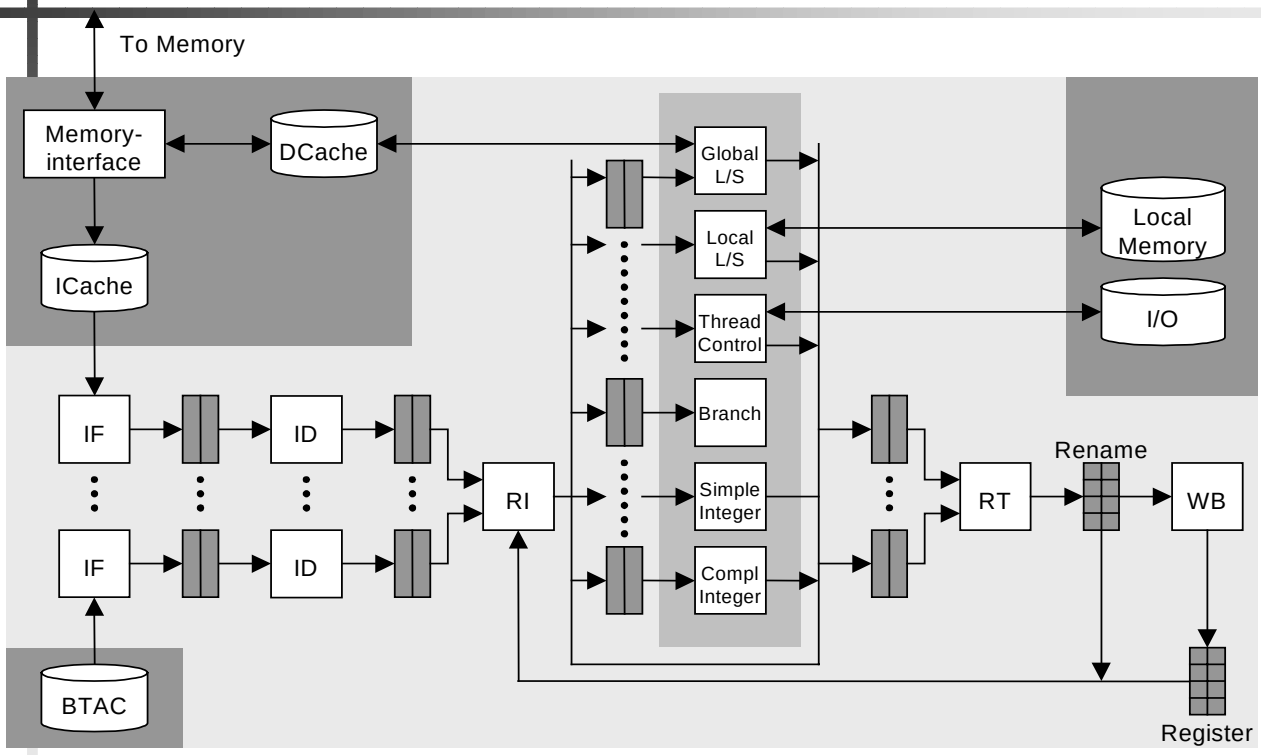
Integer/multimedia	54.0 %
Complex-integer	3.8 %
Local load/store (to local RAM store)	20.1 %
Global load/store	7.7 %
Branch	9.9 %
Frame buffer I/O	2.74 %
Thread control (parallelization overhead)	1.76 %

⇒ Two different video streams (1-2 seconds)

Combining SMT and Multimedia

- Start with a **wide-issue superscalar** general-purpose processor
- Enhance by simultaneous multithreading
- Enhance by multimedia unit(s)
- Enhance by on-chip RAM memory for constants and local variables

The SMT Multimedia Processor Model



Simulator: Fixed Parameters

- 32 32-bit general-purpose registers (per thread),
- 4 MB main memory (enough to store the whole simulation workload),
- 64-bit system bus,
- 4-way set-associative D- and I-caches,
- D-cache is a non-blocking write-back cache
- 32 KB local on-chip RAM
- a 32-entry issue buffer per thread,
- gshare two-level adaptive branch predictor with 8 bit history and 2 K 2-bit counters,
- misprediction penalty of 5 cycles

Simulator: Varying Parameters

- the number of threads from 1 to 8
- the issue bandwidth from 1 to 8
- fetch bandwidth scales with number of threads and issue bandwidth
- the number and size of reservation station units, reorder buffers, size of BTAC, the number of integer/multimedia units, number of result buses and rename registers, the size and refill strategies of the D-cache, ...

Simulation Procedure

1. maximum processor models
2. maximum processor models with realistic memory hierarchy
3. realistic processor models

1 Maximum Processor Initial Configuration

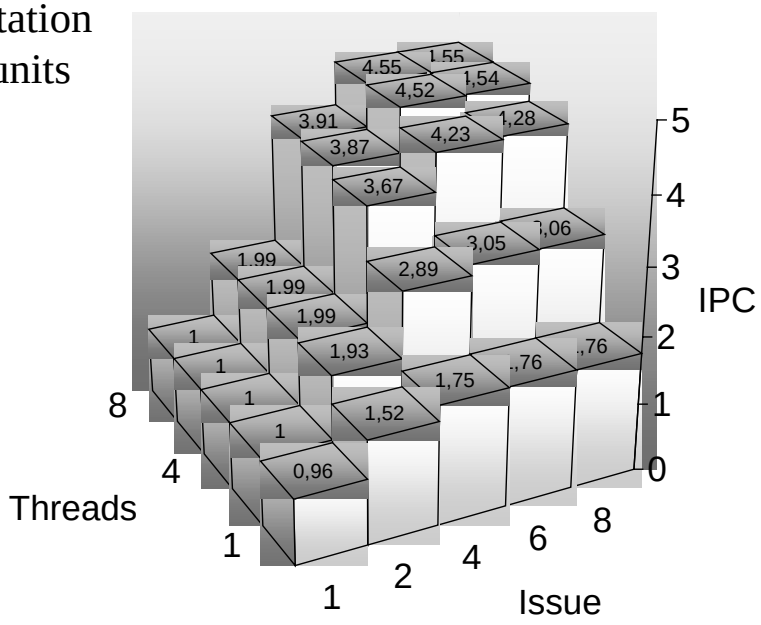
- 4 MB I- and D-caches
- D-cache fill burst rate of 6:2:2:2 processor cycles
- a 1024-entry BTAC,
- 1024 rename registers,
- 4 simple integer/multimedia units,
- 256-entry reservation stations separate for each execution unit,
- an own result bus per execution unit,
- and 256-entry reorder buffers

Maximum Processor Configuration - IPCs of 8-threaded 8-issue Cases

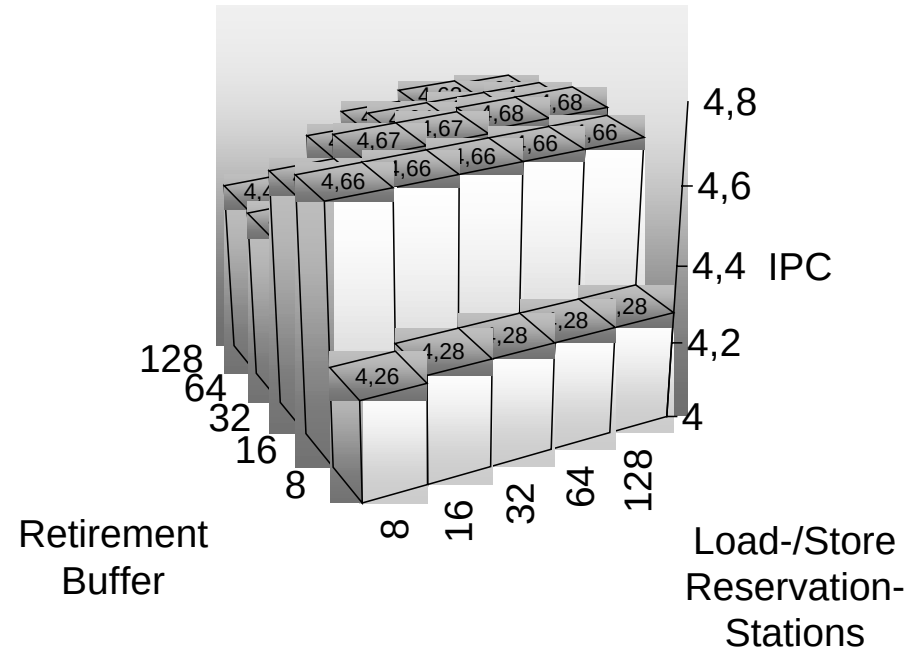
- Initial maximum configuration: **2.72**
- one common 256-entry reservation station unit for all integer/multimedia units (instead of 256-entry reservation stations each): **IPC+0.12**
- loads and stores may pass blocked load/stores of other threads: **IPC+1.71**

IPC of Maximum Processor Models (1)

One common res. station for int/multimedia units relaxed l/s accesses



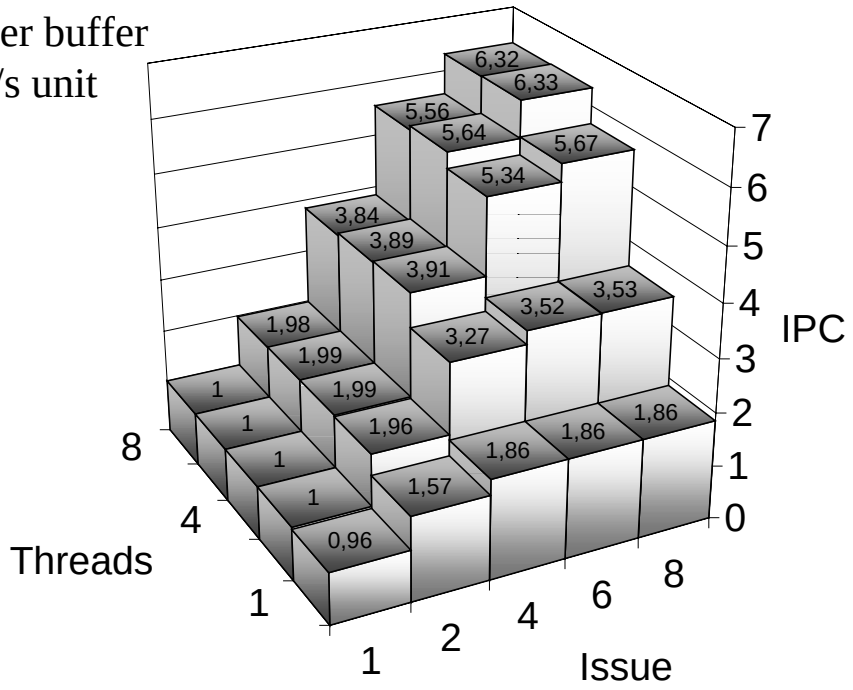
Maximum Processor Configuration (2) - IPCs of 8-threaded 8-issue cases



- 32-entry reorder buffer (instead of 256): **IPC+0.13**

IPC of Maximum Processor Models (3)

32 entry reorder buffer
second local l/s unit

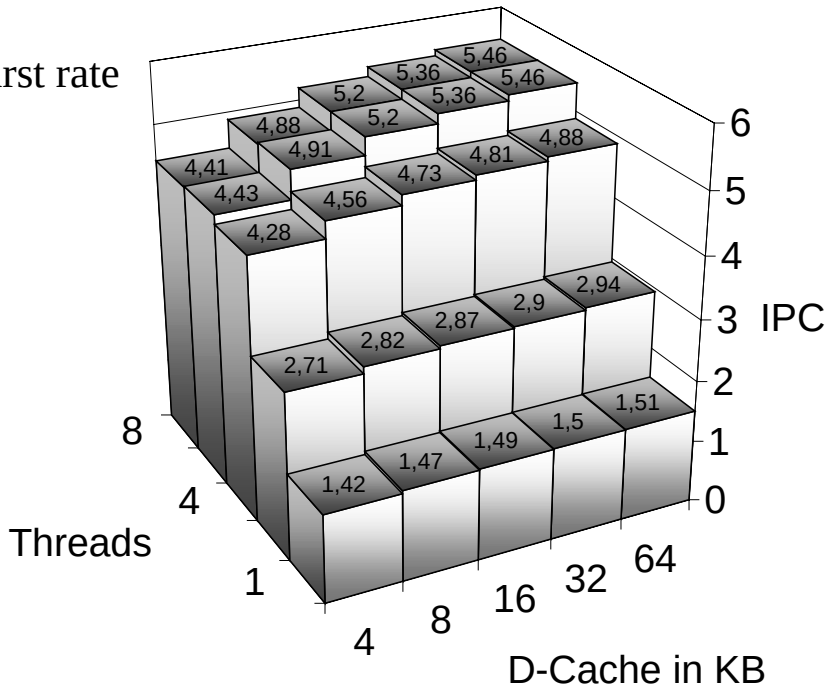


2 Maximum Processor Models With Realistic Memory Hierarchy

- 4 MB I- and D-caches are too large
  - ✓ study of different smaller D-cache sizes
- D-cache fill burst rate of 6:2:2:2 is too fast
  - ✓ reduce to 32:4:4:4
- Optimize Cache Performance
  - ❖ D-cache speculative preload,
  - ❖ D-cache associativity,
  - ✓ and D-cache line replacement strategies.
- Increase memory bandwidth
  - ✓ Increase burst length

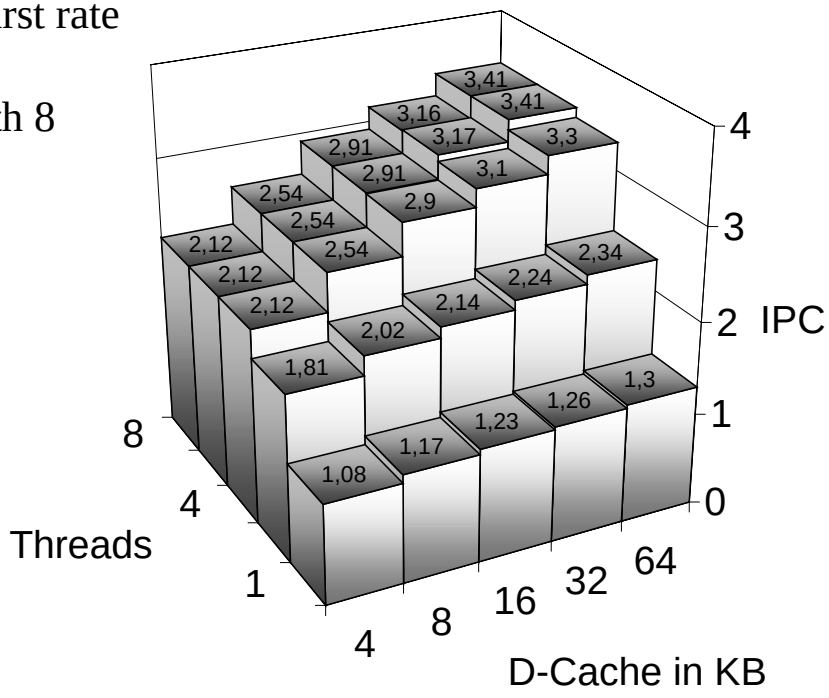
2 Reduced D-cache Sizes for the 8-issue Models

4 MB I-cache,
D-cache fill burst rate
of 6:2:2:2



2 Longer Memory Latencies

D-cache fill burst rate
of 32:4:4:4
issue bandwidth 8



Analysis for the 64 KB D-cache 8-issue Models

- Only 76% cache hit ratio for all models.
- A high 69% to 70% utilization of the integer/multimedia reservation station for the 8- to 4-threaded models (even higher utilization for smaller D-cache sizes).

➨ Integer/multimedia instructions are blocked because of missing data operands that are not yet loaded from memory.

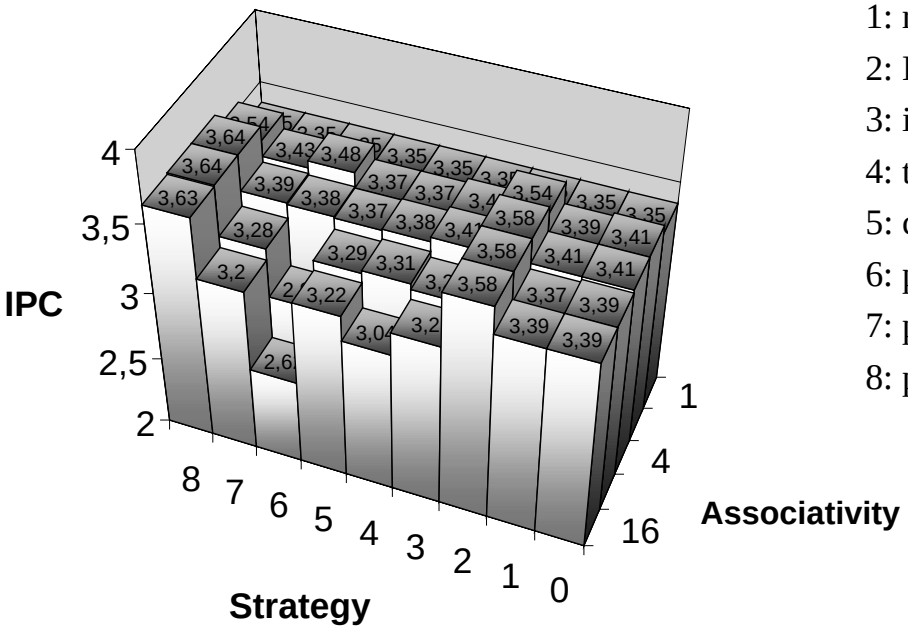
➨ Make memory access more efficient!

Maximum Processor Models With Realistic Memory Hierarchy

- 4 MB I- and D-caches are too large
  - ✓ study of different smaller D-cache sizes
- D-cache fill burst rate of 6:2:2:2 is too fast
  - ✓ reduce to 32:4:4:4
- Optimize Cache Performance
  - ❖ D-cache speculative preload,
  - ❖ D-cache associativity,
  - ✓ and D-cache line replacement strategies.
- Increase memory bandwidth
  - ✓ Increase burst length

D-cache Line Replacement Strategies

(64 KB D-cache, 32:4:4:4, Issue Bandwidth 8)



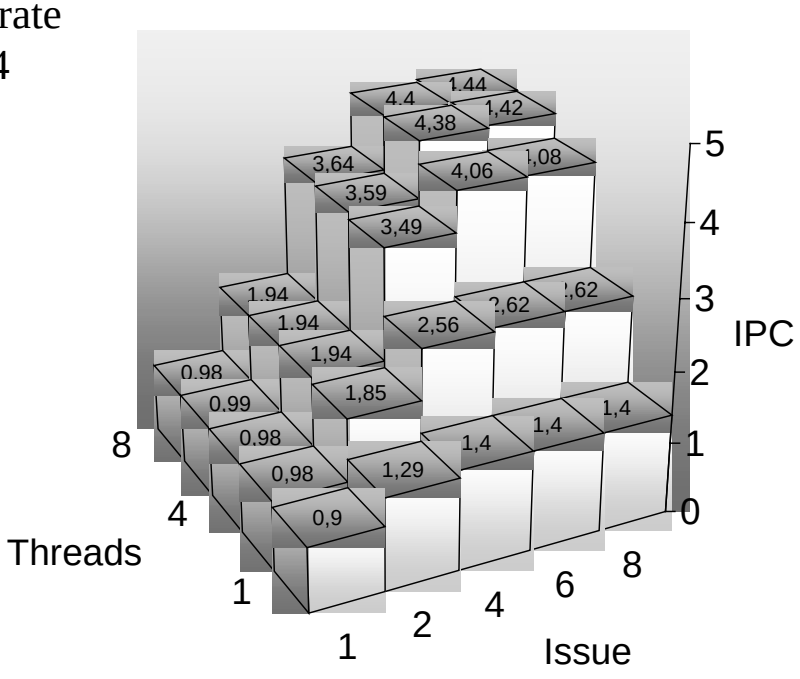
0: round-robin,
1: random selection,
2: LRU strategies,
3: instruction-based,
4: thread-ID based,
5: combined (3 and 4),
6: priority,
7: priority and random,
8: priority and LRU.

Analysis of D-cache Line Replacement Strategies

- The importance of the cache line replacement strategy rises with the associativity.
- The LRU-based strategies 2 and 8 profit from a higher associativity.
- Stratey 8: combination of thread priority based with LRU performs best.

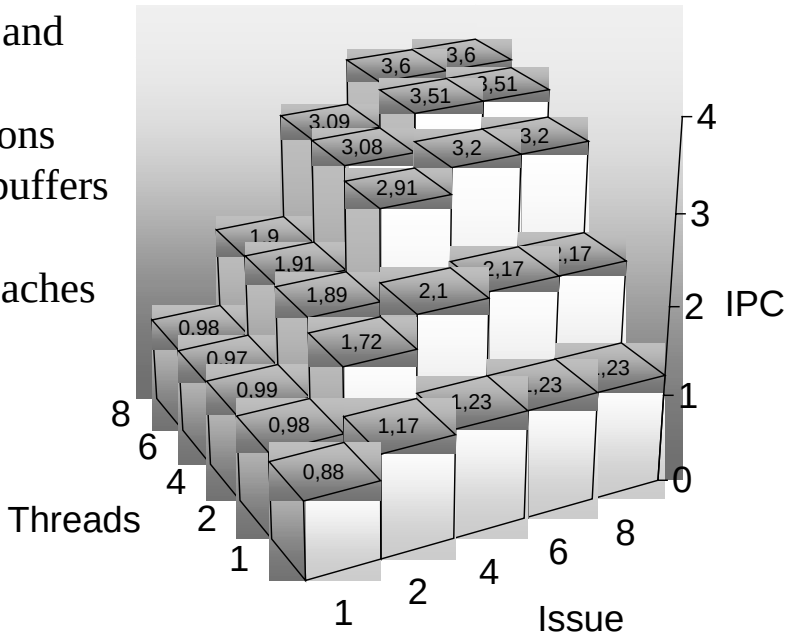
Increased Memory Bandwidth

D-cache fill burst rate of 32:4:4:4:4:4:4:4
issue bandwidth 8
Priority & LRU cache strategy

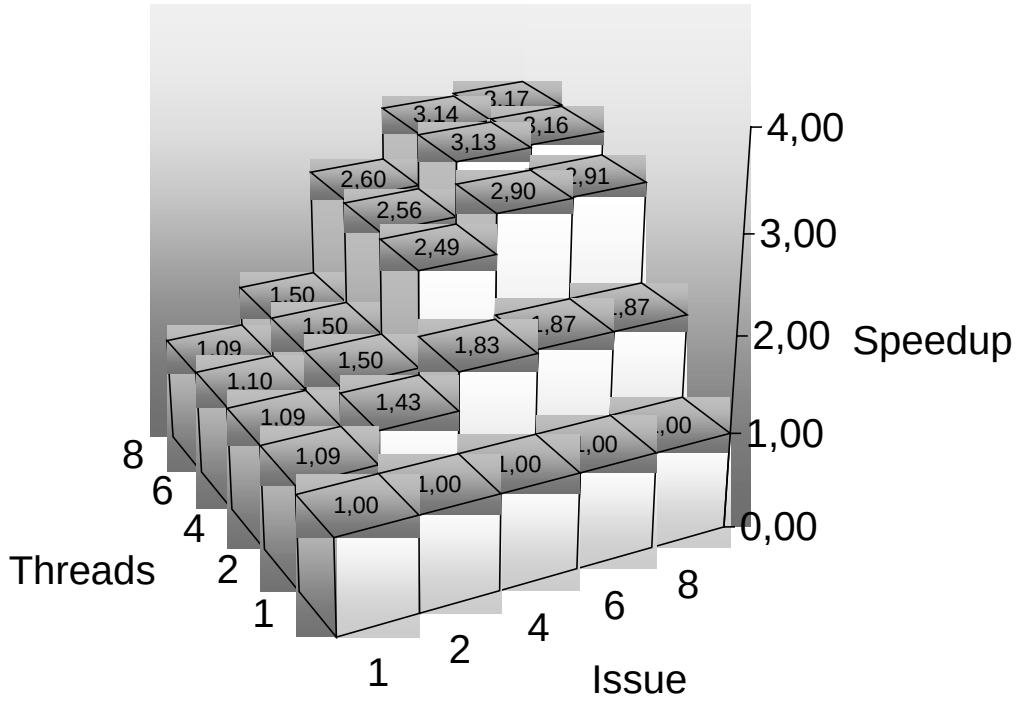


Realistic Processor Models

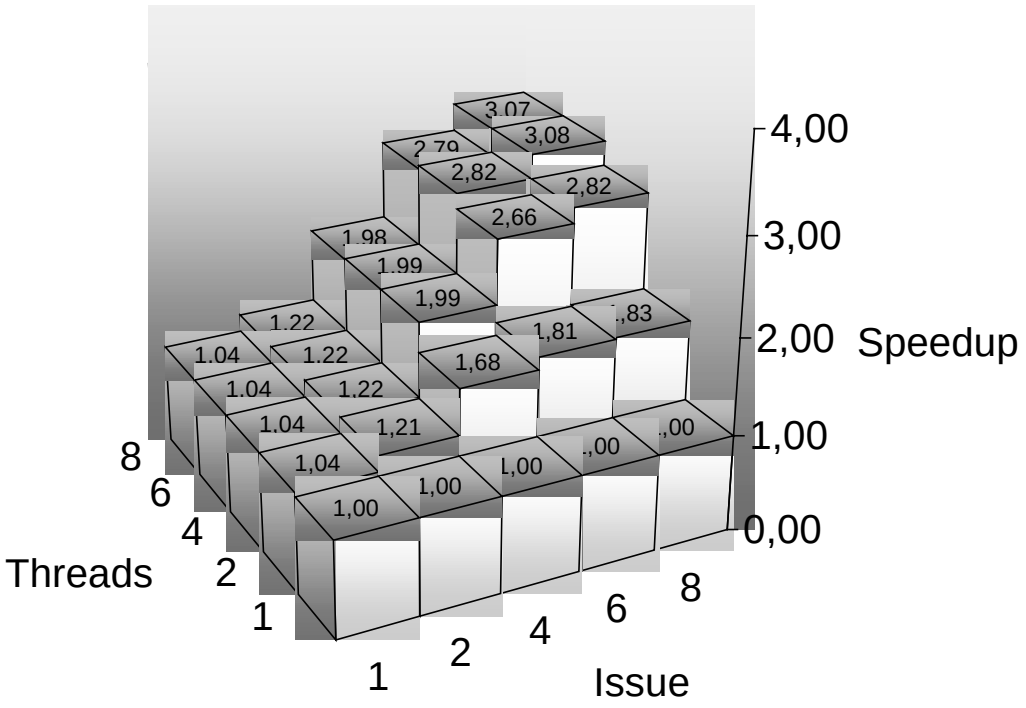
two 4-wide fetch and decode units
32 entry res. stations
16 entry reorder buffers
one local l/s unit
64 KB I- and D-caches
32:4:4:4:4:4:4:4
priority & LRU



Speedup of Realistic Processor Models



Speedup of Maximum Processor Models



Conclusions

- We simulated multimedia-enhanced SMT processor models using a multithreaded MPEG-2 workload.
- Results show a threefold IPC increase of the 8-threaded SMT over single-threaded processors,

  but also the sensitiveness of SMT processors with respect to memory bandwidth.
- MPEG-2 is a good workload for SMT processors, if properly parallelized

- Our processor models are specifically tailored to the multithreaded MPEG-2 algorithm.
- Other workloads might favor different configuration, in particular, if the workload consists of unrelated threads of control that do not share any data.