WaveScalar

Dataflow machine

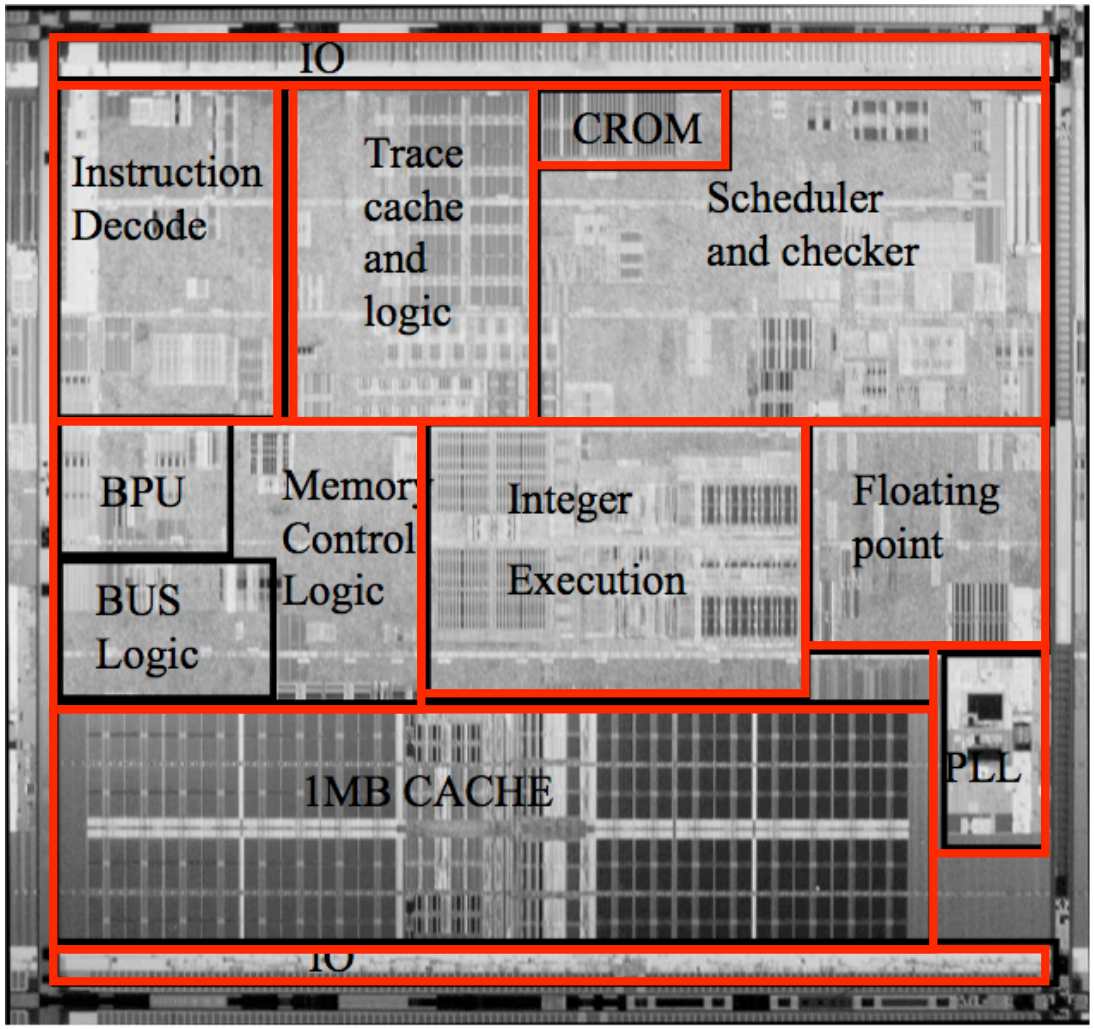
- good at exploiting ILP
- dataflow parallelism + traditional coarser-grain parallelism
  - cheap thread management
- memory ordering enforced through **wave-ordered memory**

WaveScalar

Motivation:

- increasing disparity between computation (fast transistors) & communication (long wires)
- increasing circuit complexity
- decreasing fabrication reliability

Monolithic von Neumann Processors



A phenomenal success today.

But in 2016?

☹ Performance

Centralized processing & control, e.g., operand broadcast networks

☹ Complexity

40-75% of "design" time is design verification

☹ Defect tolerance

1 flaw -> paperweight

WaveScalar Executive Summary

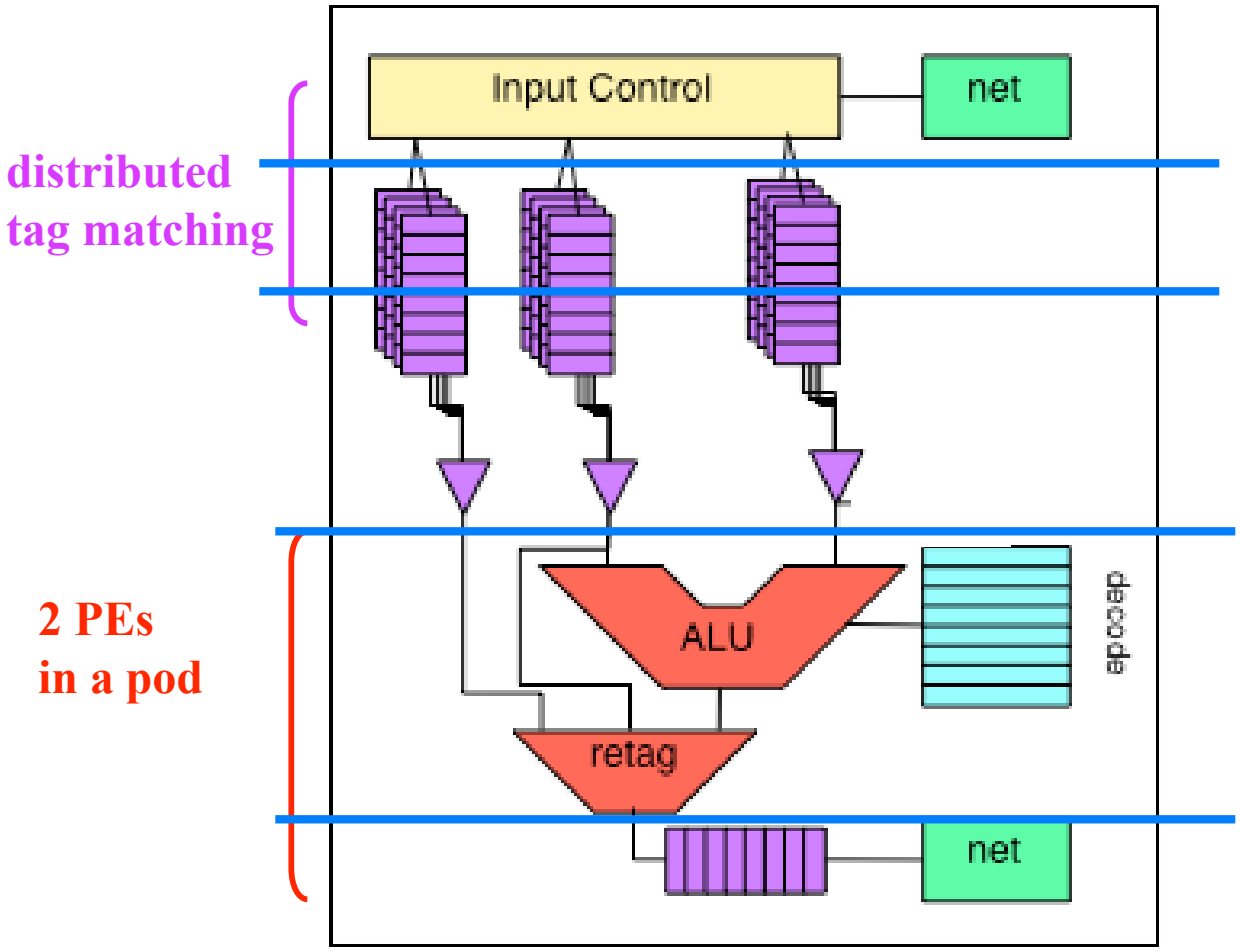
Distributed microarchitecture ☺

- hundreds of PEs
- dataflow execution – no centralized control
- short point-to-point communication
- organized hierarchically for fast communication between neighboring PEs
- defect tolerance – route around a bad PE

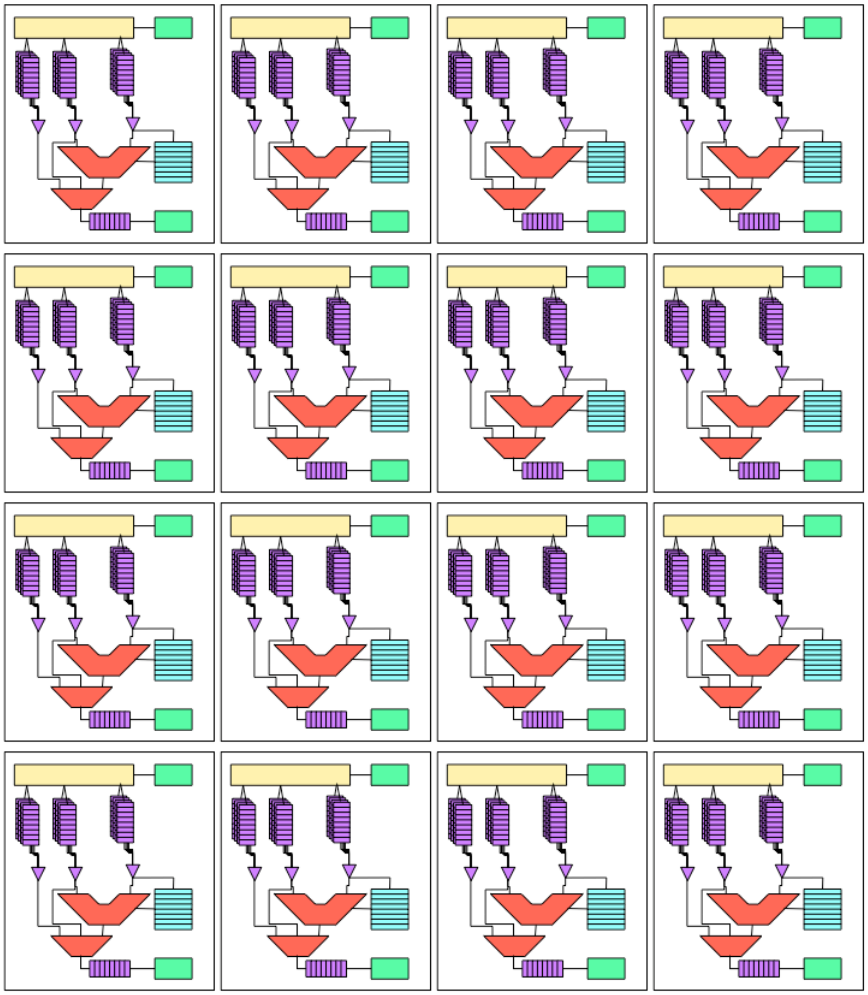
Low design complexity through simple, identical PEs ☺

- design one & stamp out thousands

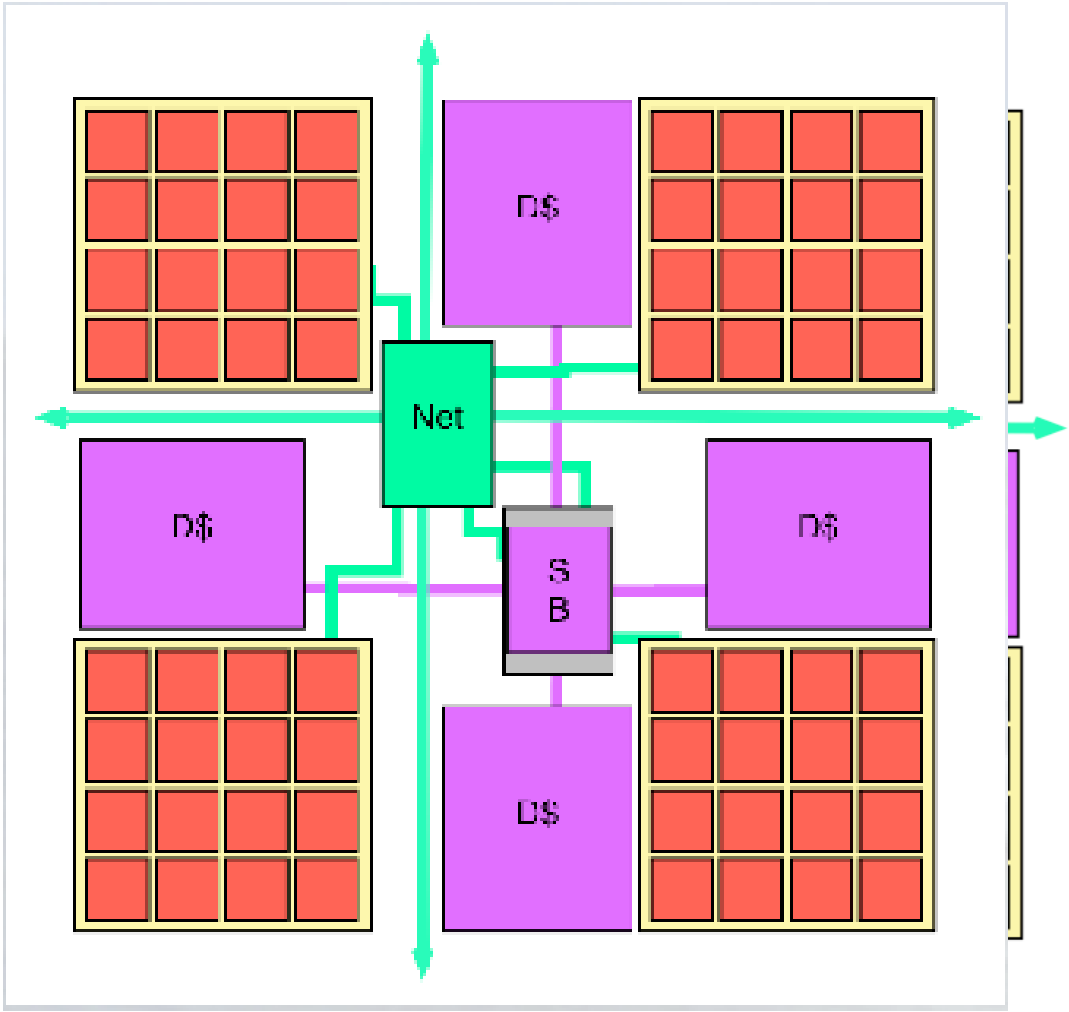
Processing Element



Domain

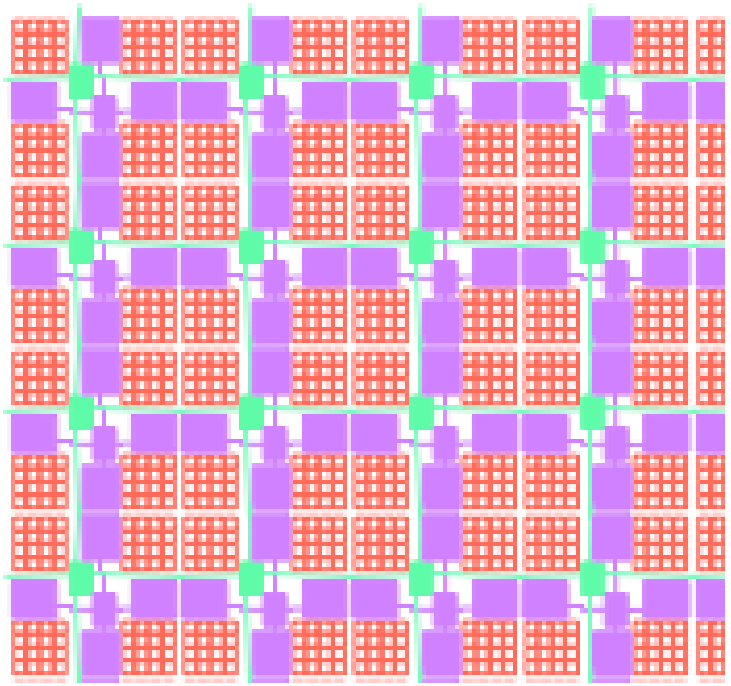


Cluster



Whole Chip

- Can hold 32K instructions
- Long distance communication
  - Dynamic routing
  - Grid-based network
  - 2-cycle hop/cluster
- Normal memory hierarchy
- Traditional directory-based cache coherence



WaveScalar Execution Model

Dataflow

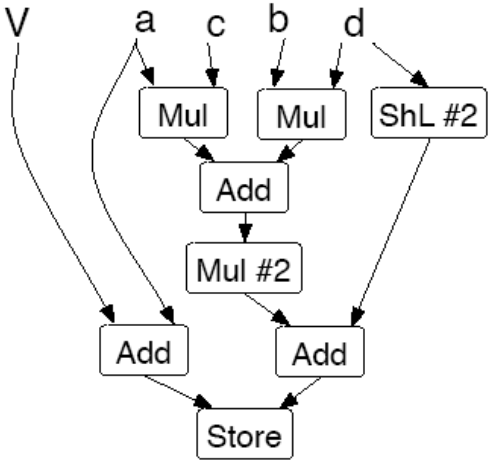
Place instructions in PEs to maximize data locality & instruction-level parallelism.

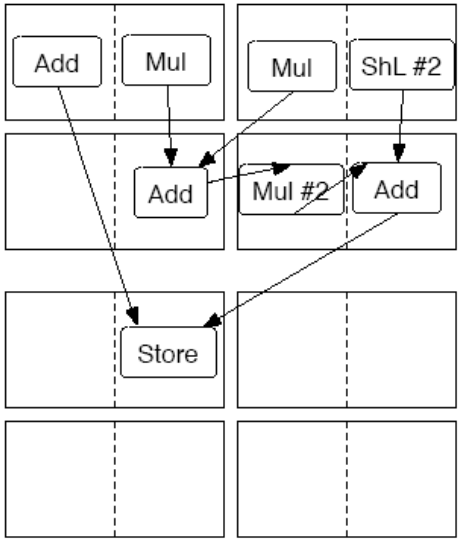
- Instruction placement algorithm based on a performance model that captures the conflicting goals
- Depth-first traversal of dataflow graph to make chains of dependent instructions
- Broken into segments
- Snakes segments across the chip on demand
- K-loop bounding to prevent instruction “explosion”

Instructions communicate values directly (point-to-point).

WaveScalar Instruction Placement

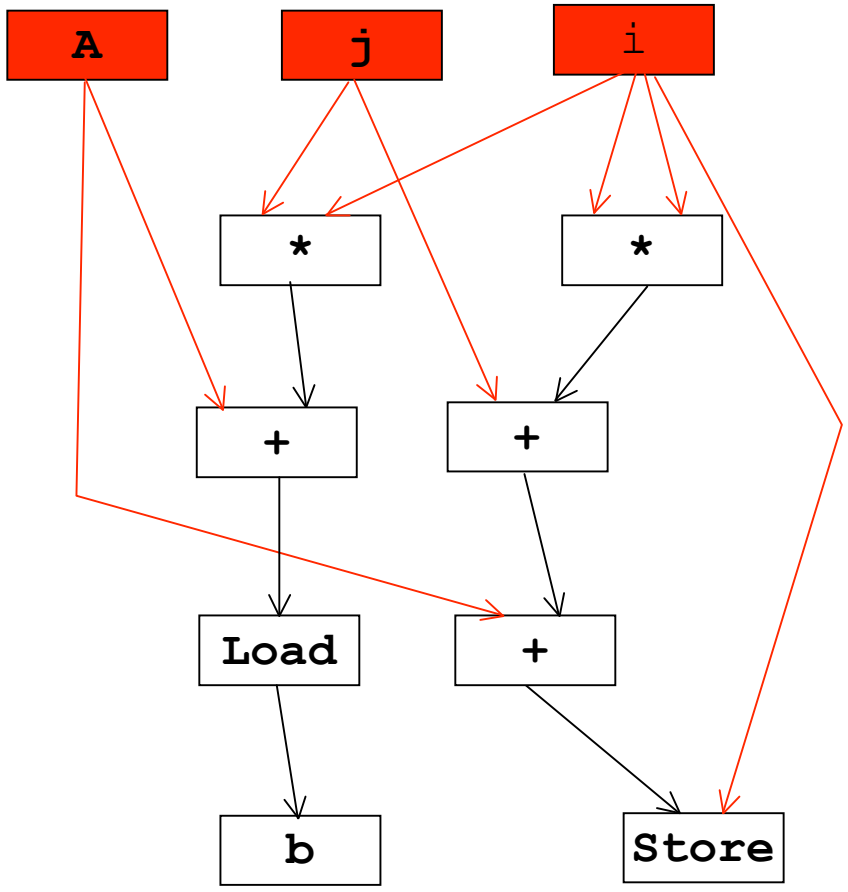
```
int *V;
int a, b;
int c, d, r;

r = a*c + b*d;
V[a] = 2*r + d << 2;
```





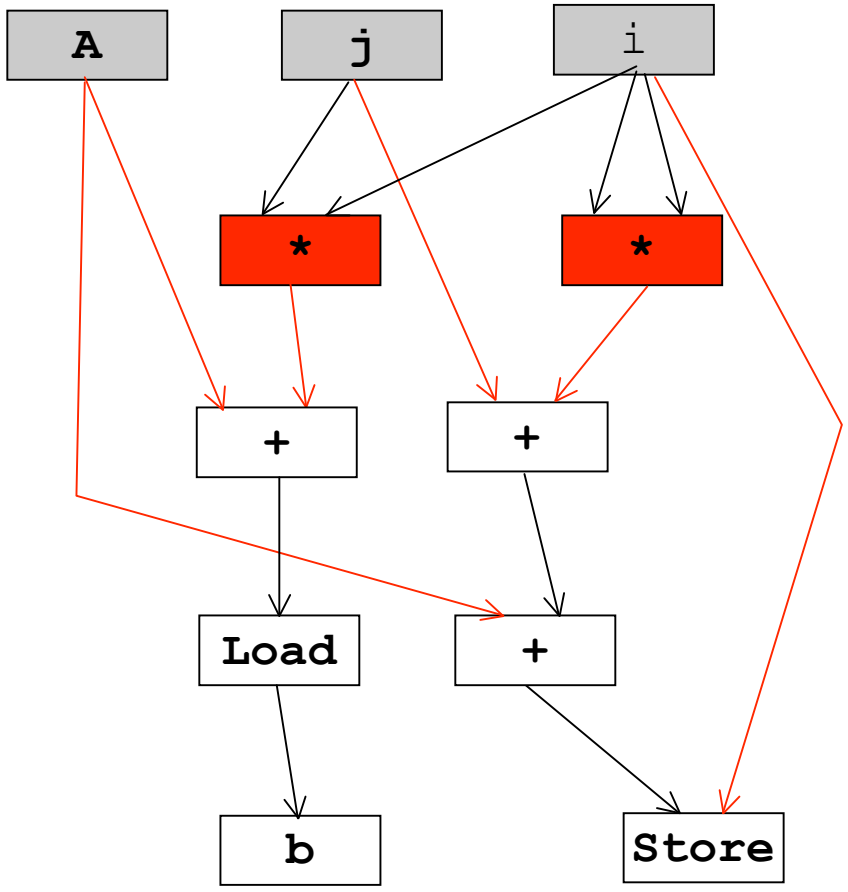
WaveScalar Example

```
A[j + i*i] = i;

b = A[i*j];
```



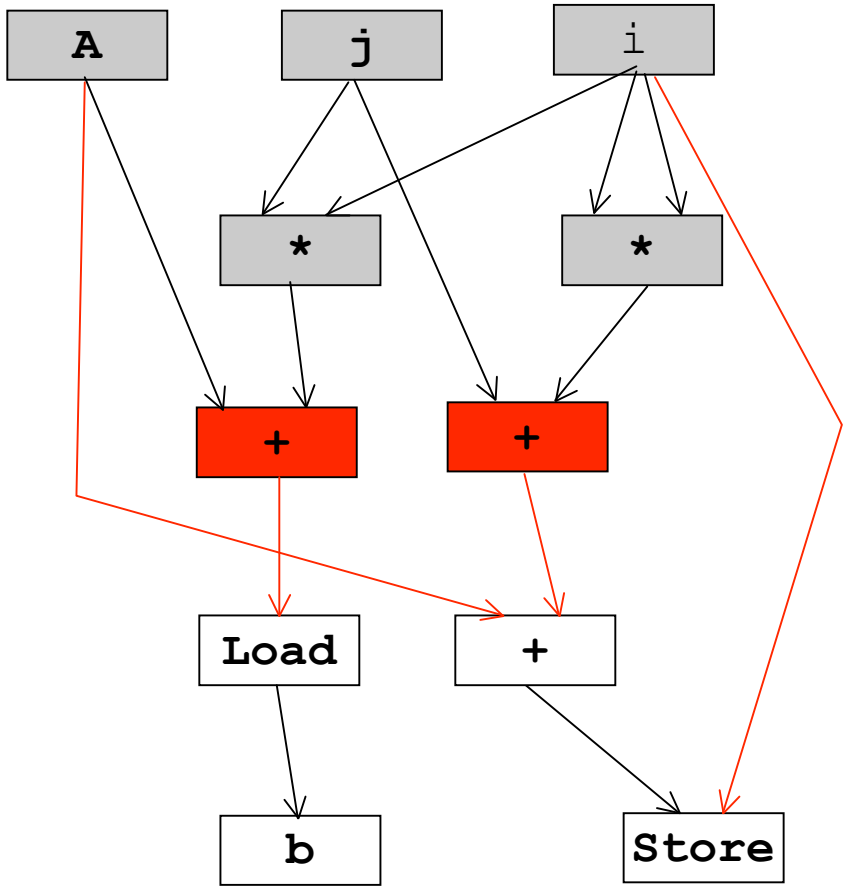
WaveScalar Example

```
A[j + i*i] = i;

b = A[i*j];
```



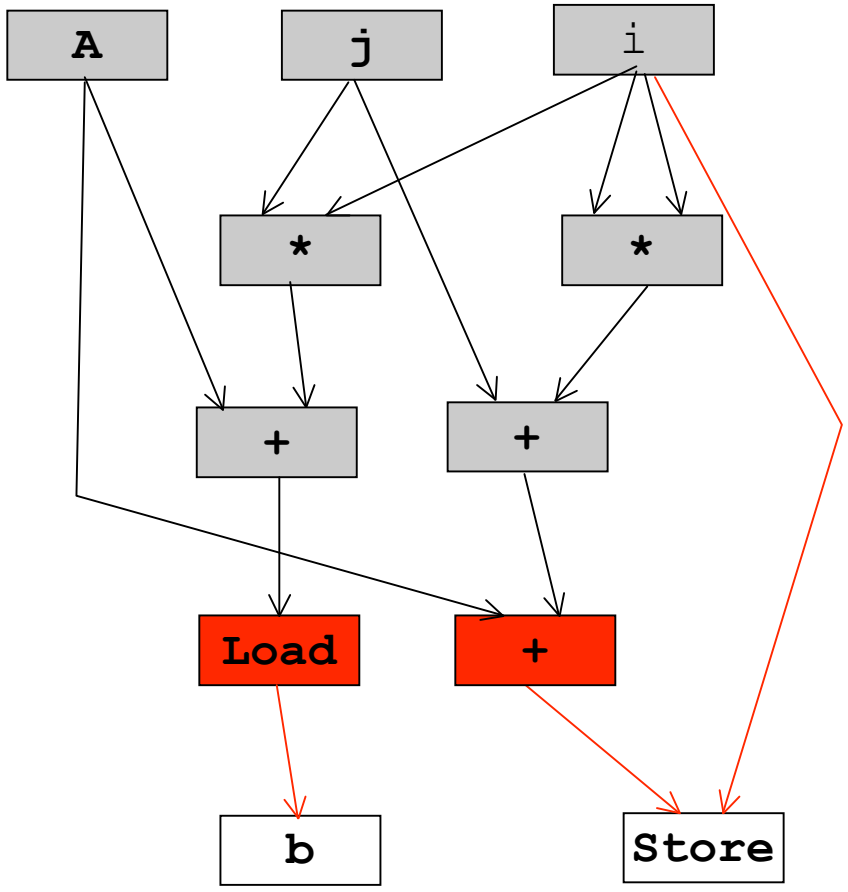
WaveScalar Example

```
A[j + i*i] = i;

b = A[i*j];
```



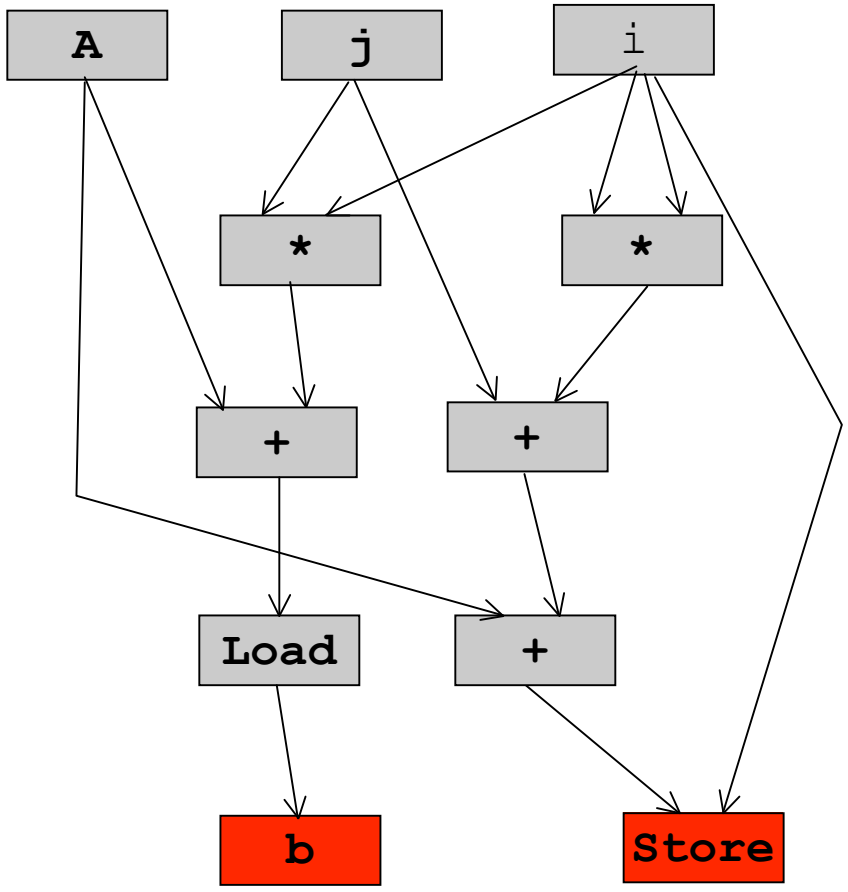
WaveScalar Example

```
A[j + i*i] = i;

b = A[i*j];
```



WaveScalar Example

```
A[j + i*i] = i;

b = A[i*j];
```

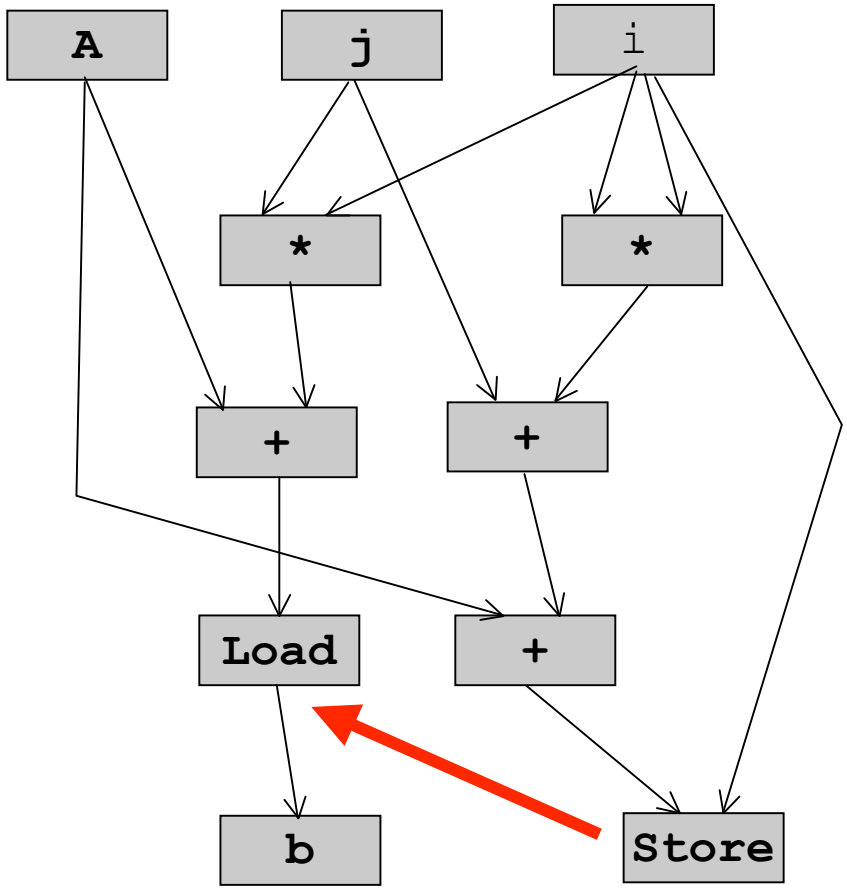


WaveScalar Example

```
A[j + i*i] = i;
```

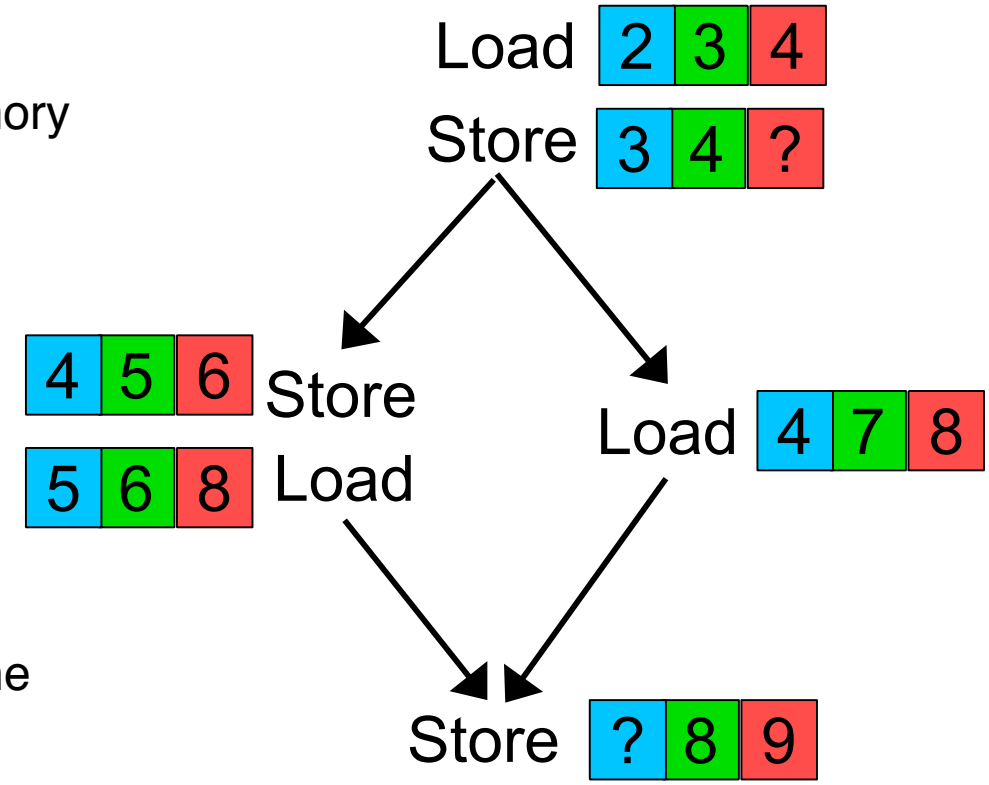
```
b = A[i*j];
```

Global load-store ordering issue

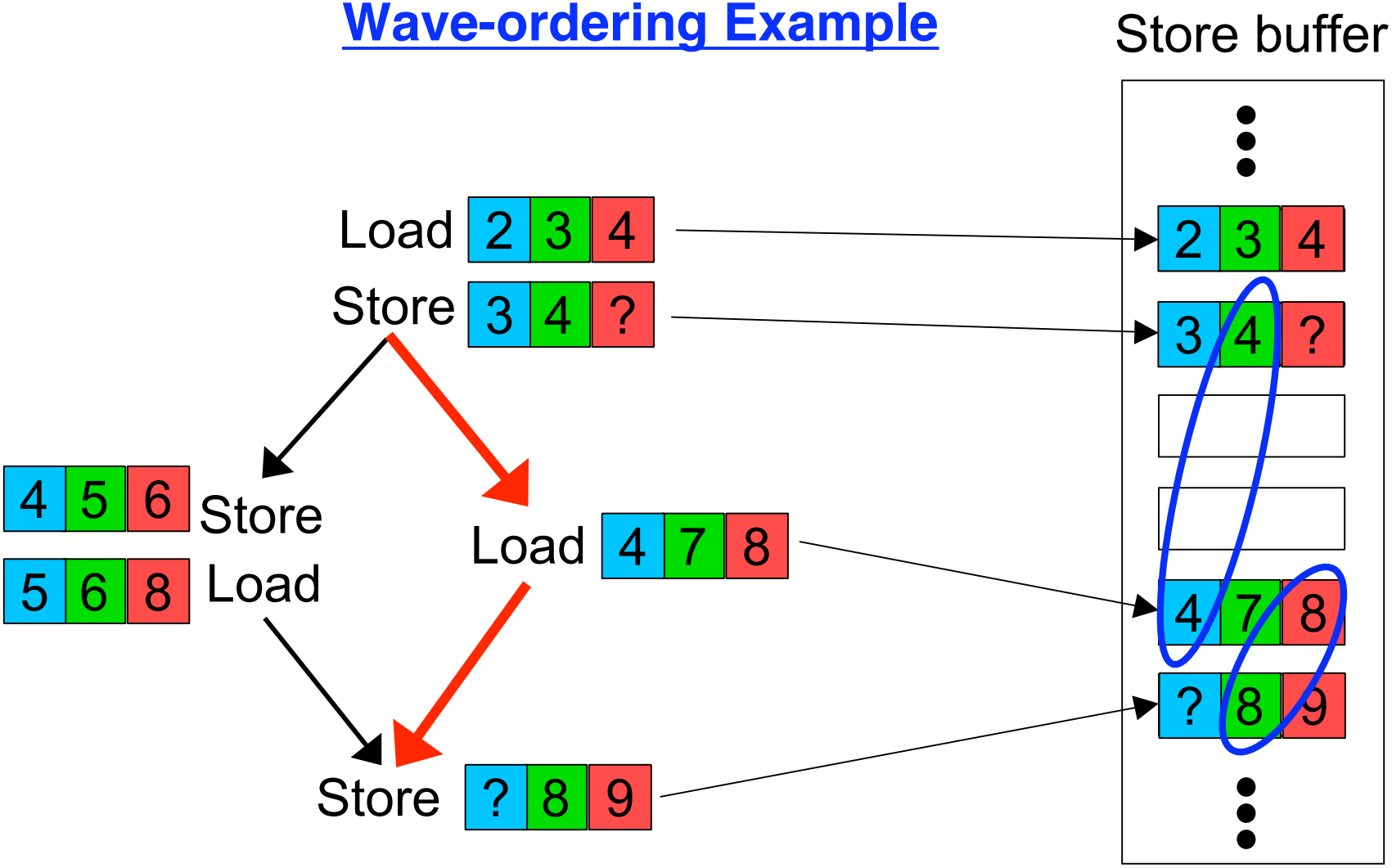


Wave-ordered Memory

- Compiler annotates memory operations
  - Sequence #
  - Successor
  - Predecessor
- Send memory requests in any order
- Hardware reconstructs the correct order



Wave-ordering Example



Wave-ordered Memory

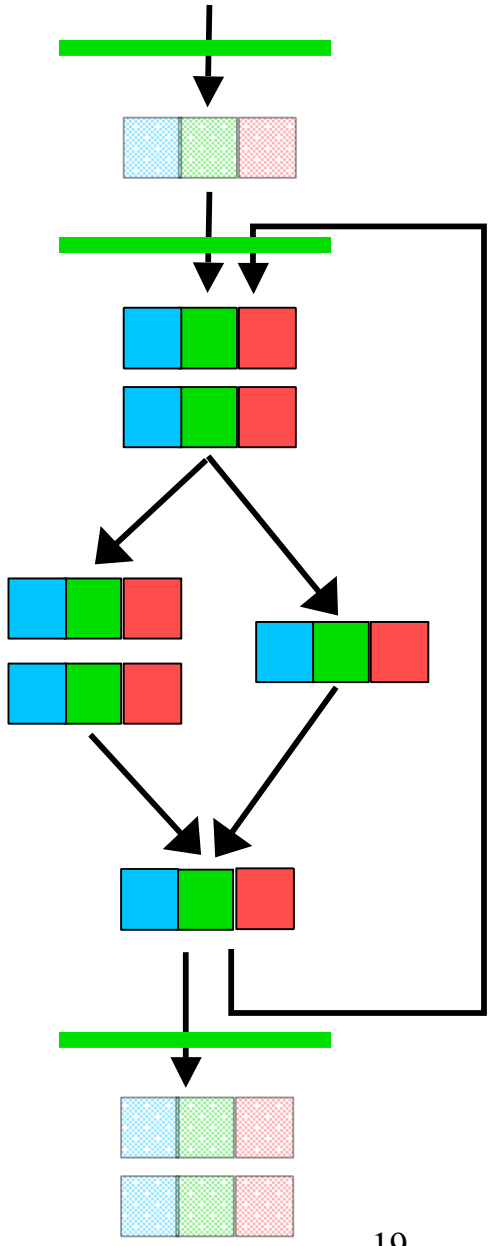
Waves are loop-free sections of the dataflow graph

Each dynamic wave has a **wave number**

Wave number is incremented between waves

Ordering memory:

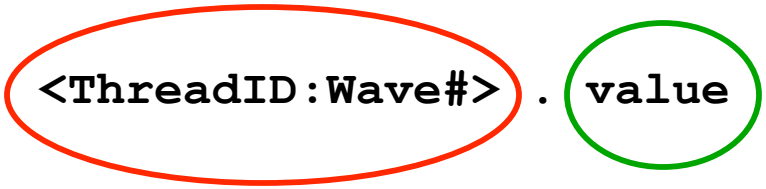
- wave-numbers
- sequence number within a wave

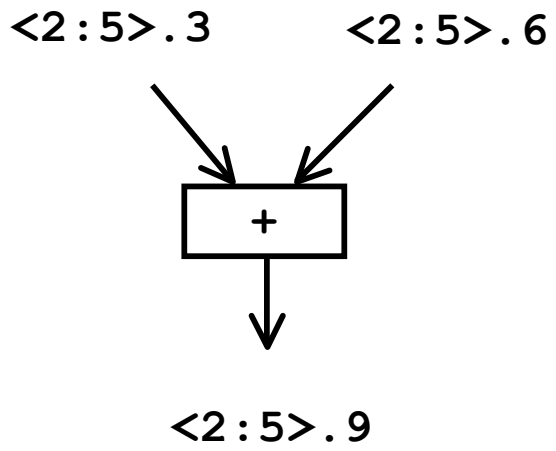


WaveScalar Tag-matching

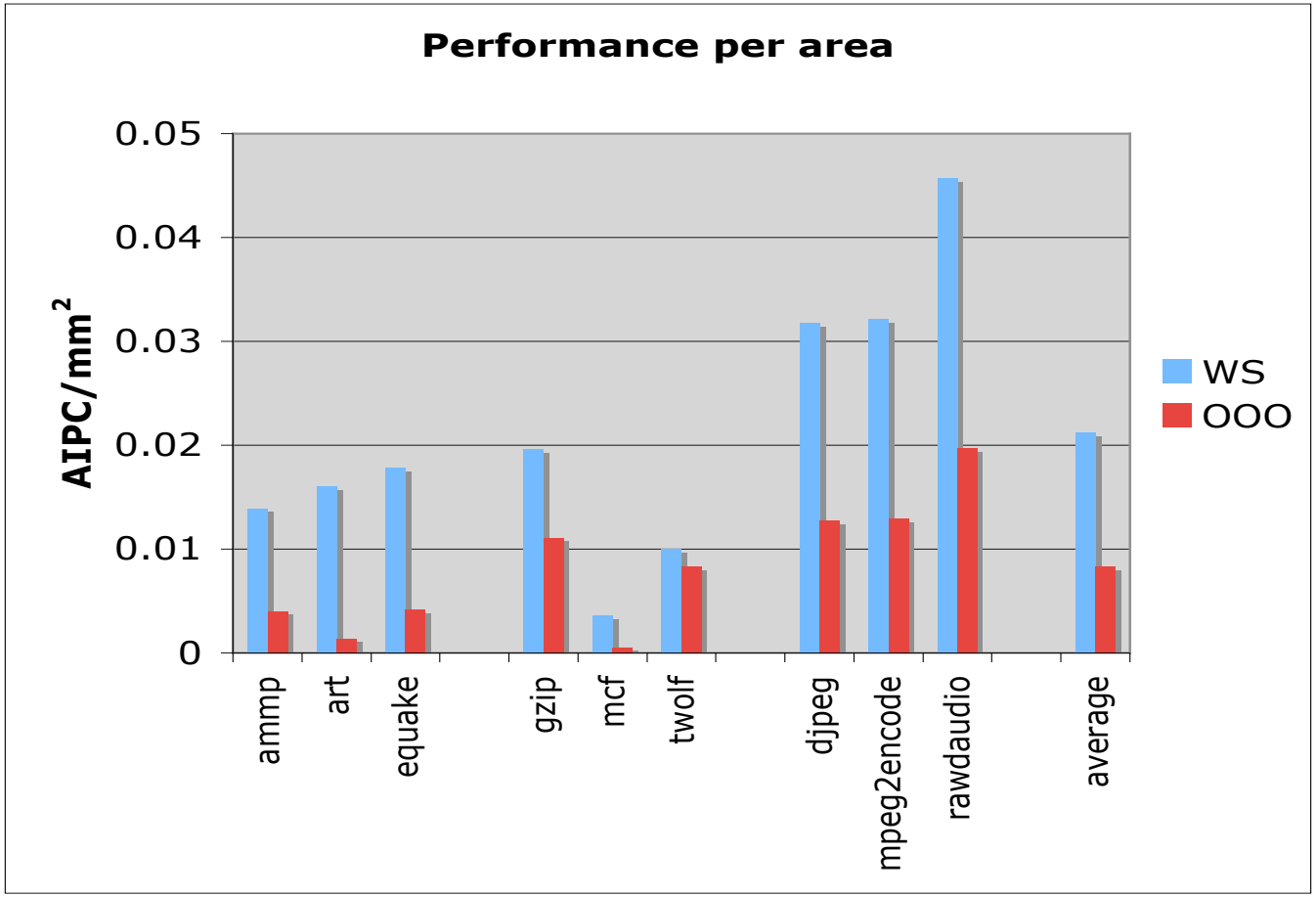
WaveScalar tag
- thread identifier
- wave number

Token: **tag** & **value**





Single-thread Performance



Multithreading the WaveCache

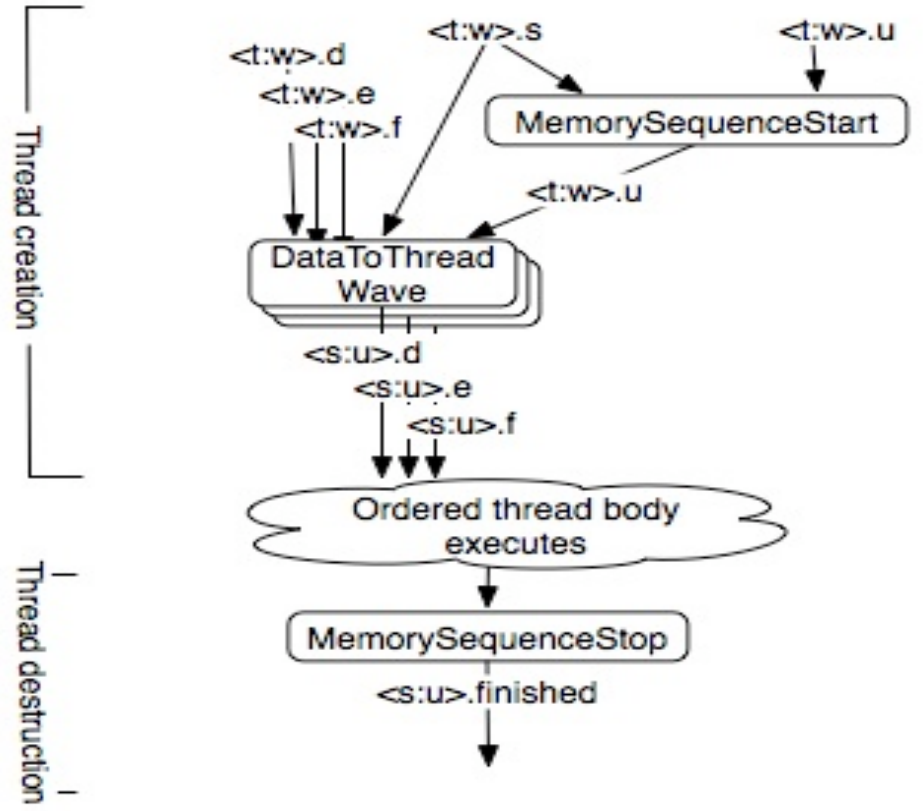
Architectural-support for WaveScalar threads

- instructions to start & stop memory orderings, i.e., threads
- memory-free synchronization to allow exclusive access to data (TC)
- fence instruction to allow other threads to see this one's memory ops

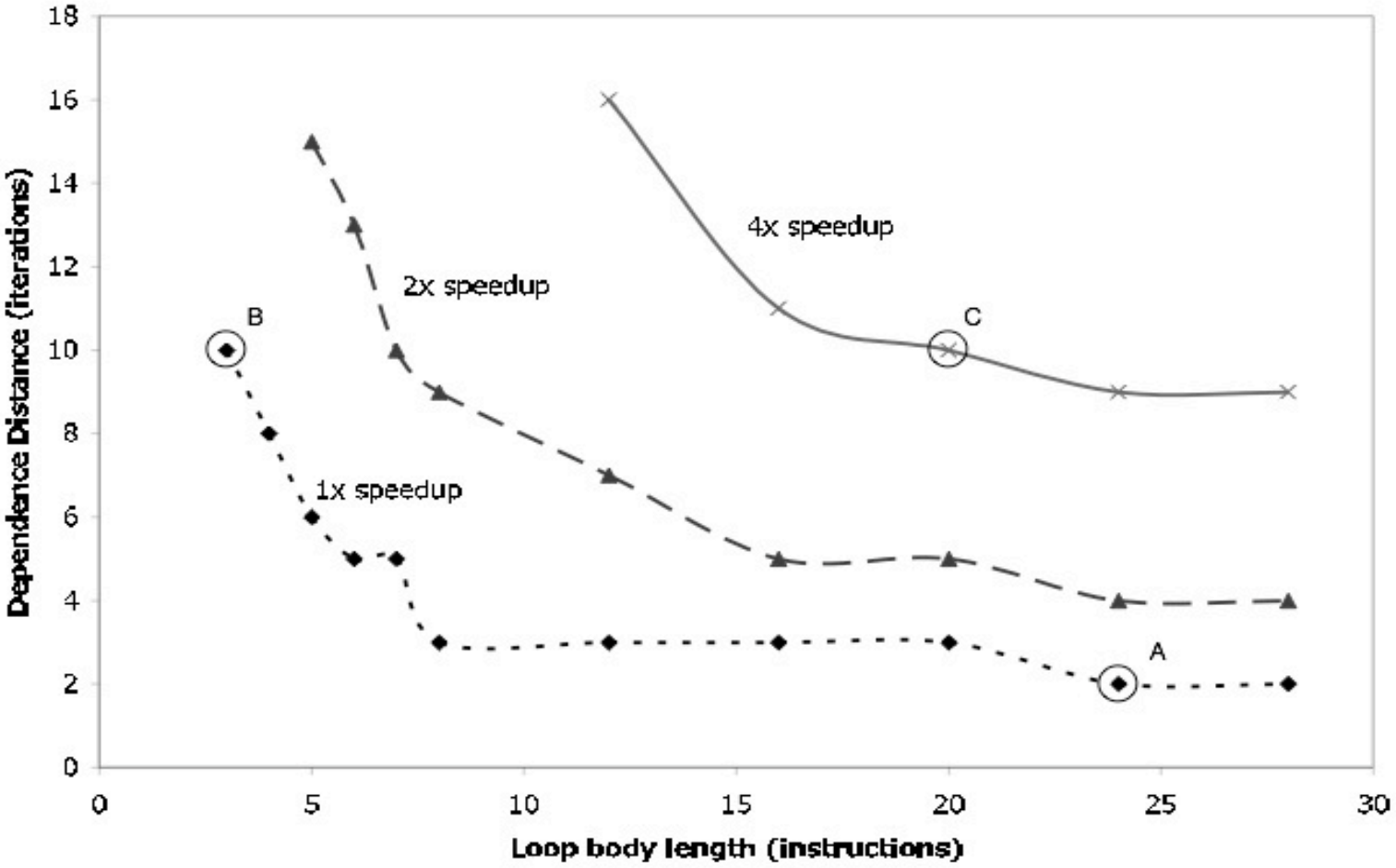
Combine to build threads with multiple granularities

- coarse-grain threads: 25-168X over a single thread; 2-16X over CMP, 5-11X over SMT
- fine-grain, dataflow-style threads: 18-242X over single thread
- combine the two in the same application: 1.6X or 7.9X -> 9X

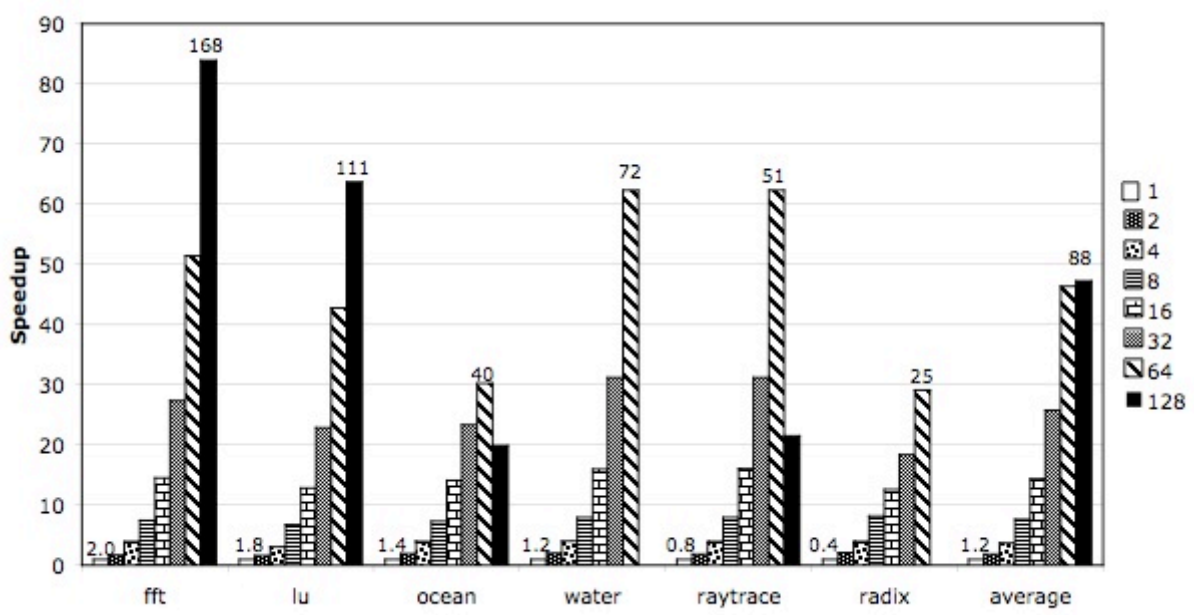
Creating & Terminating a Thread



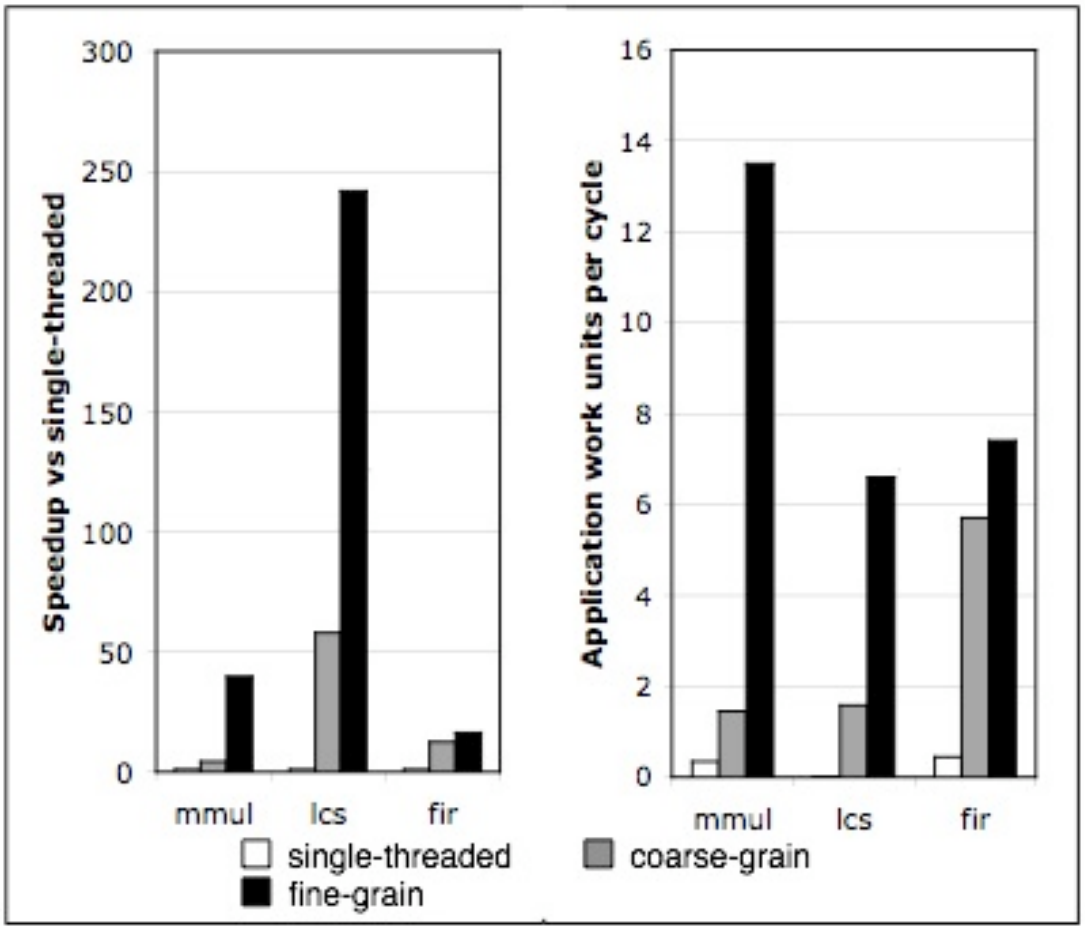
Thread Creation Overhead



Performance of Coarse-grain Parallelism



Performance of Fine-grain Parallelism



Building the WaveCache

RTL-level implementation

- some didn't believe it could be built in a normal-sized chip
- some didn't believe it could achieve a decent cycle time and load-use latencies
- Verilog & Synopsis CAD tools

Different WaveCache's for different applications

- 1 cluster: low-cost, low power, single-thread or embedded
  - 52 mm^2 in 90 nm process technology, 3.5 AIPC on Splash2
- 16 clusters: multiple threads, higher performance: 436 mm^2 , 15 AIPC

board-level FPGA implementation

- OS & real application simulations