

18-447

Computer Architecture  
Lecture 16: SIMD Processing  
(Vector and Array Processors)

Prof. Onur Mutlu

Carnegie Mellon University

Spring 2014, 2/24/2014

# Lab 4 Reminder

---

- Lab 4a out
  - Branch handling and branch predictors
- Lab 4b out
  - Fine-grained multithreading
- **Due March 21st**
- You have 4 weeks!
- Get started very early – Exam and S. Break are on the way
- Finish Lab 4a first and check off
- Finish Lab 4b next and check off
- Do the extra credit

# Readings for Today

---

- SIMD Processing
- Basic GPU Architecture
- Other execution models: VLIW, Dataflow
  
- Lindholm et al., "NVIDIA Tesla: A Unified Graphics and Computing Architecture," IEEE Micro 2008.
  
- Fatahalian and Houston, "A Closer Look at GPUs," CACM 2008.
  
- Stay tuned for more readings...

# SIMD Processing: Exploiting Regular (Data) Parallelism

# Flynn's Taxonomy of Computers

---

- Mike Flynn, “**Very High-Speed Computing Systems,**” Proc. of IEEE, 1966
- **SISD**: Single instruction operates on single data element
- **SIMD**: Single instruction operates on multiple data elements
  - Array processor
  - Vector processor
- **MISD**: Multiple instructions operate on single data element
  - Closest form: systolic array processor, streaming processor
- **MIMD**: Multiple instructions operate on multiple data elements (multiple instruction streams)
  - Multiprocessor
  - Multithreaded processor

# Data Parallelism

---

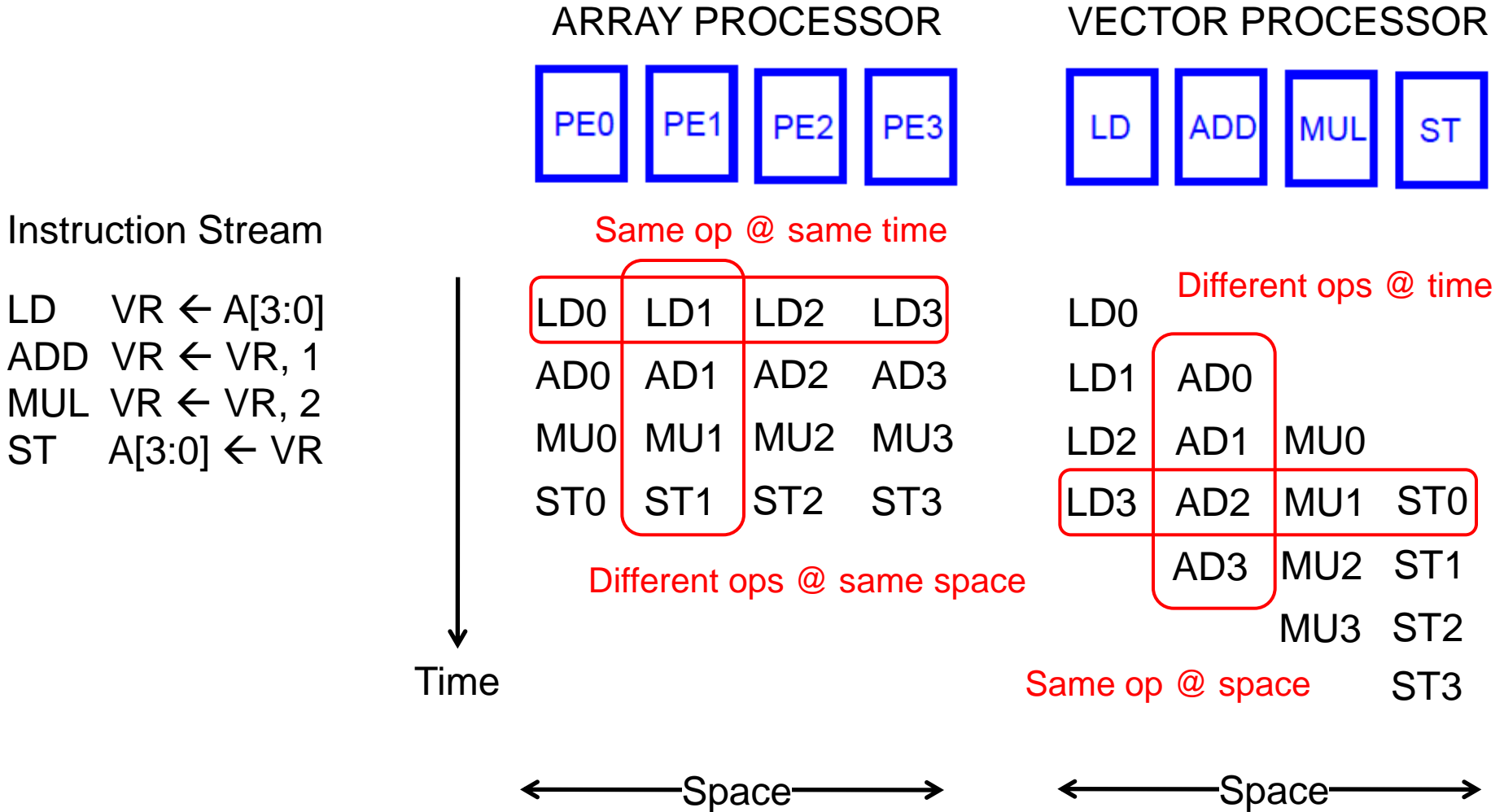
- Concurrency arises from performing the **same operations on different pieces of data**
  - Single instruction multiple data (SIMD)
  - E.g., dot product of two vectors
- Contrast with data flow
  - Concurrency arises from executing different operations in parallel (in a data driven manner)
- Contrast with thread (“control”) parallelism
  - Concurrency arises from executing different threads of control in parallel
- SIMD exploits instruction-level parallelism
  - Multiple “instructions” (more appropriately, operations) are concurrent: instructions happen to be the same

# SIMD Processing

---

- Single instruction operates on multiple data elements
  - In time or in space
- Multiple processing elements
  
- Time-space duality
  - **Array processor**: Instruction operates on multiple data elements at the same time
  - **Vector processor**: Instruction operates on multiple data elements in consecutive time steps

# Array vs. Vector Processors

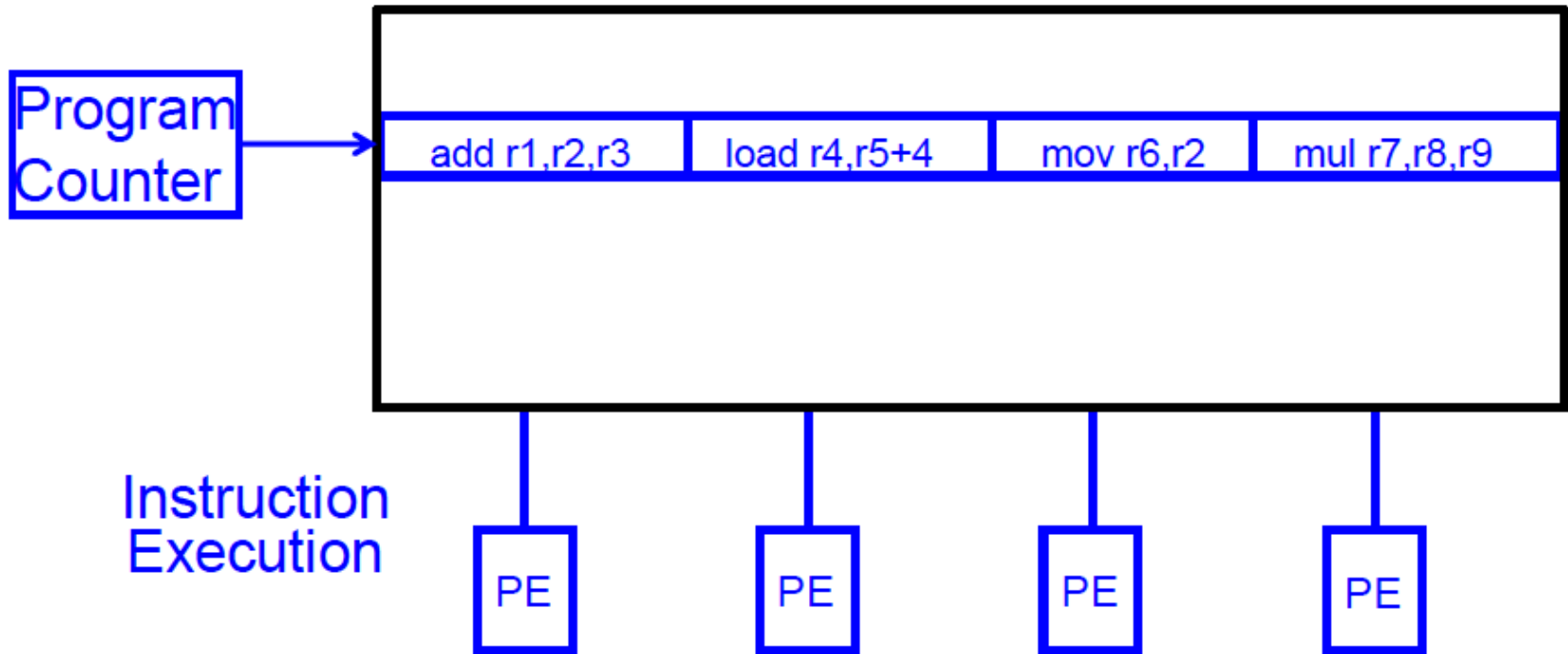




# SIMD Array Processing vs. VLIW

---

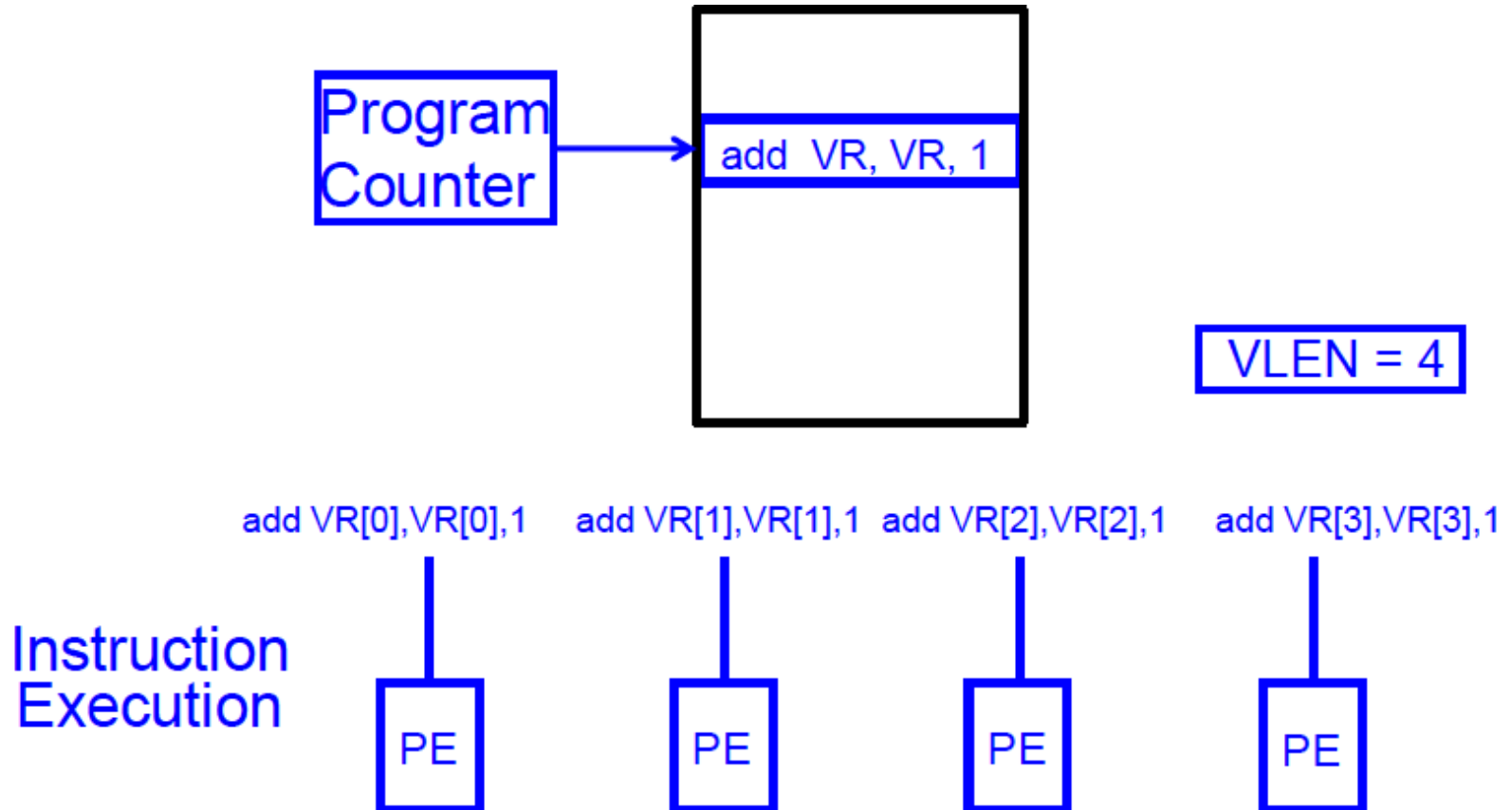
- VLIW



# SIMD Array Processing vs. VLIW

---

- Array processor



# Vector Processors

---

- A vector is a one-dimensional array of numbers
- Many scientific/commercial programs use vectors
  - for (i = 0; i<=49; i++)  
C[i] = (A[i] + B[i]) / 2
- A vector processor is one whose instructions operate on vectors rather than scalar (single data) values
- Basic requirements
  - Need to load/store vectors → vector registers (contain vectors)
  - Need to operate on vectors of different lengths → vector length register (VLEN)
  - Elements of a vector might be stored apart from each other in memory → vector stride register (VSTR)
    - Stride: distance between two elements of a vector

# Vector Processors (II)

---

- A vector instruction performs an operation on each element in consecutive cycles
  - Vector functional units are pipelined
  - Each pipeline stage operates on a different data element
- Vector instructions allow deeper pipelines
  - No intra-vector dependencies → no hardware interlocking within a vector
  - No control flow within a vector
  - Known stride allows prefetching of vectors into cache/memory

# Vector Processor Advantages

---

## + No dependencies within a vector

- ❑ Pipelining, parallelization work well
- ❑ Can have very deep pipelines, no dependencies!

## + Each instruction generates a lot of work

- ❑ Reduces instruction fetch bandwidth

## + Highly regular memory access pattern

- ❑ Interleaving multiple banks for higher memory bandwidth
- ❑ Prefetching

## + No need to explicitly code loops

- ❑ Fewer branches in the instruction sequence

# Vector Processor Disadvantages

---

- Works (only) if parallelism is regular (data/SIMD parallelism)
  - ++ Vector operations
  - Very inefficient if parallelism is irregular
    - How about searching for a key in a linked list?

**To program a vector machine, the compiler or hand coder must make the data structures in the code fit nearly exactly the regular structure built into the hardware. That's hard to do in first place, and just as hard to change. One tweak, and the low-level code has to be rewritten by a very smart and dedicated programmer who knows the hardware and often the subtleties of the application area. Often the rewriting is**

# Vector Processor Limitations

---

- Memory (bandwidth) can easily become a bottleneck, especially if
  1. compute/memory operation balance is not maintained
  2. data is not mapped appropriately to memory banks

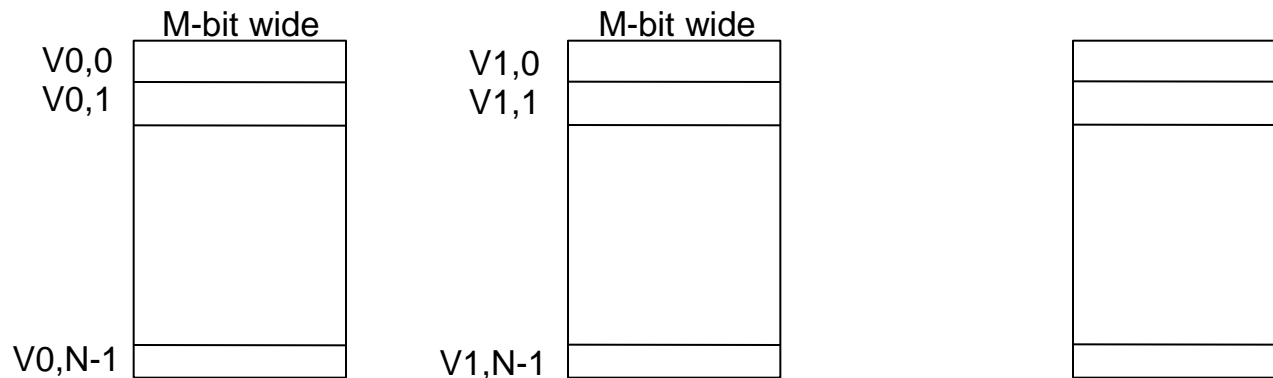
# Vector Processing in More Depth



# Vector Registers

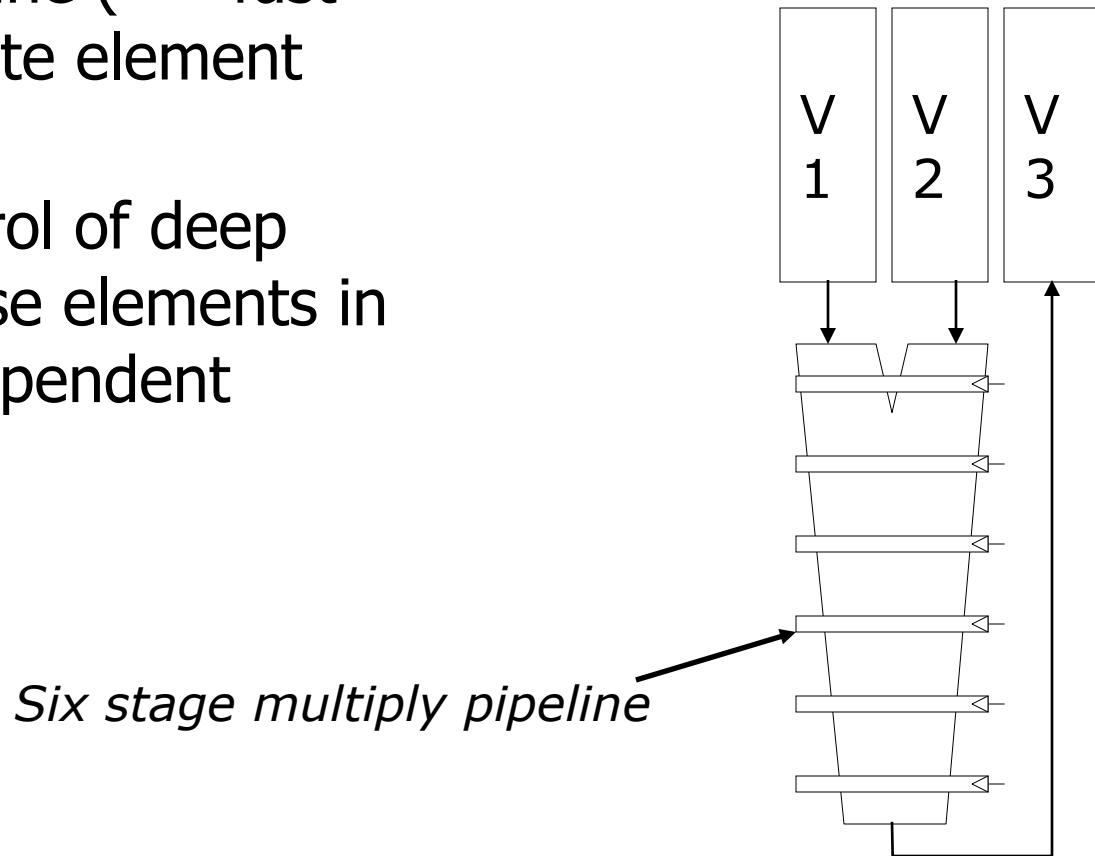
---

- Each **vector data register** holds N M-bit values
- **Vector control registers**: VLEN, VSTR, VMASK
- Maximum VLEN can be N
  - Maximum number of elements stored in a vector register
- **Vector Mask Register (VMASK)**
  - Indicates which elements of vector to operate on
  - Set by vector test instructions
    - e.g.,  $VMASK[i] = (V_k[i] == 0)$



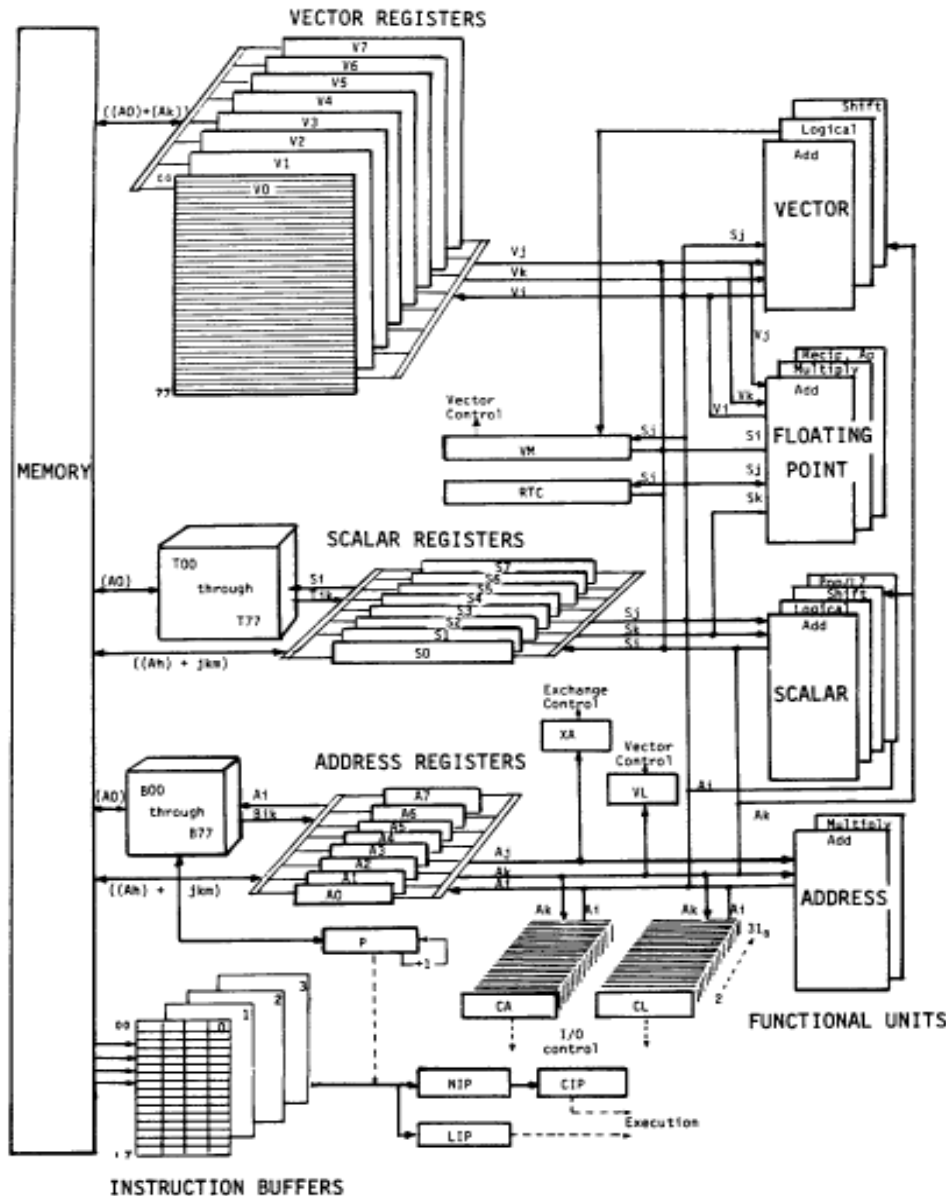
# Vector Functional Units

- Use deep pipeline ( $\Rightarrow$  fast clock) to execute element operations
- Simplifies control of deep pipeline because elements in vector are independent



$$V3 \leftarrow v1 * v2$$

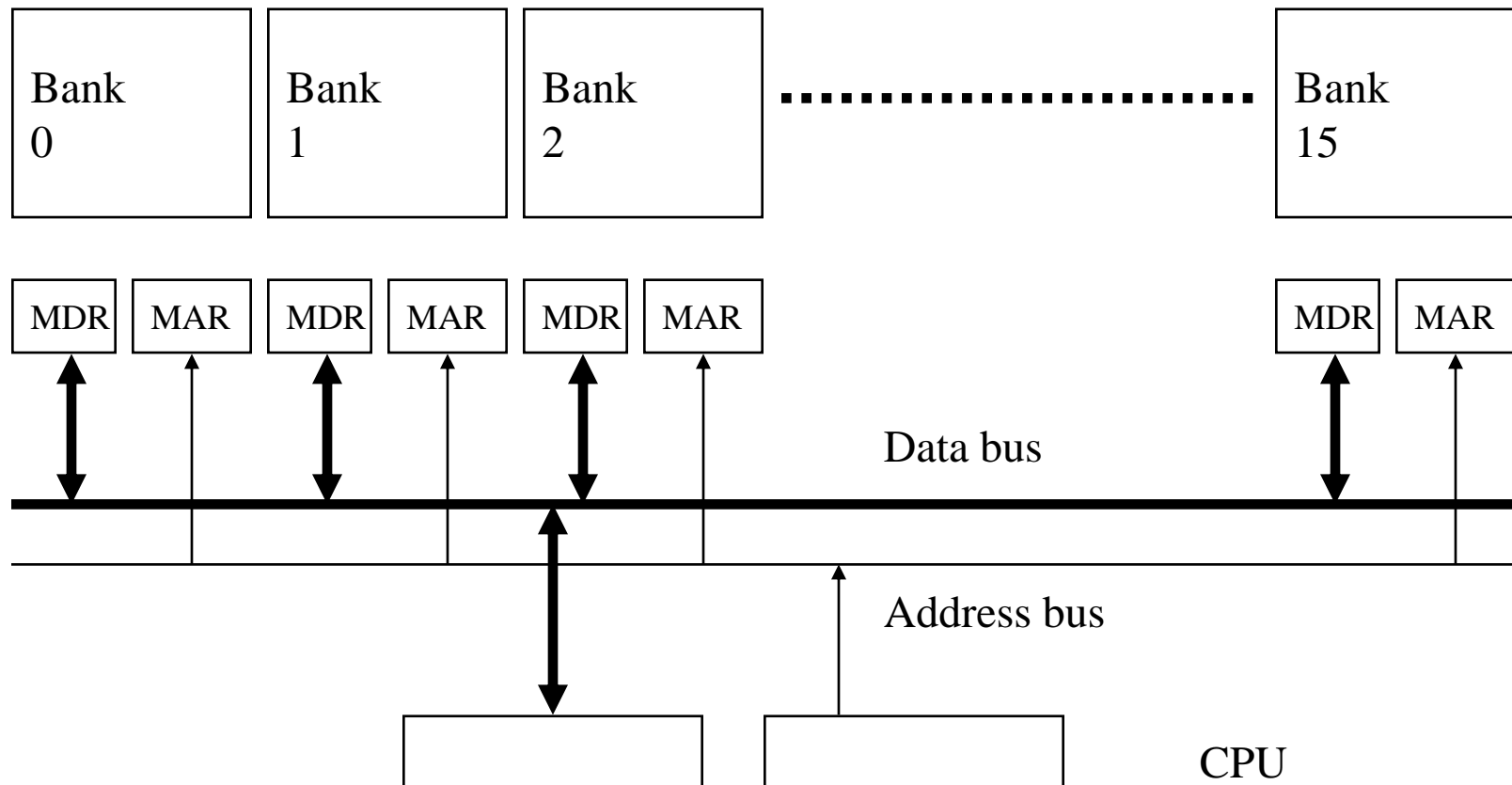
# Vector Machine Organization (CRAY-1)



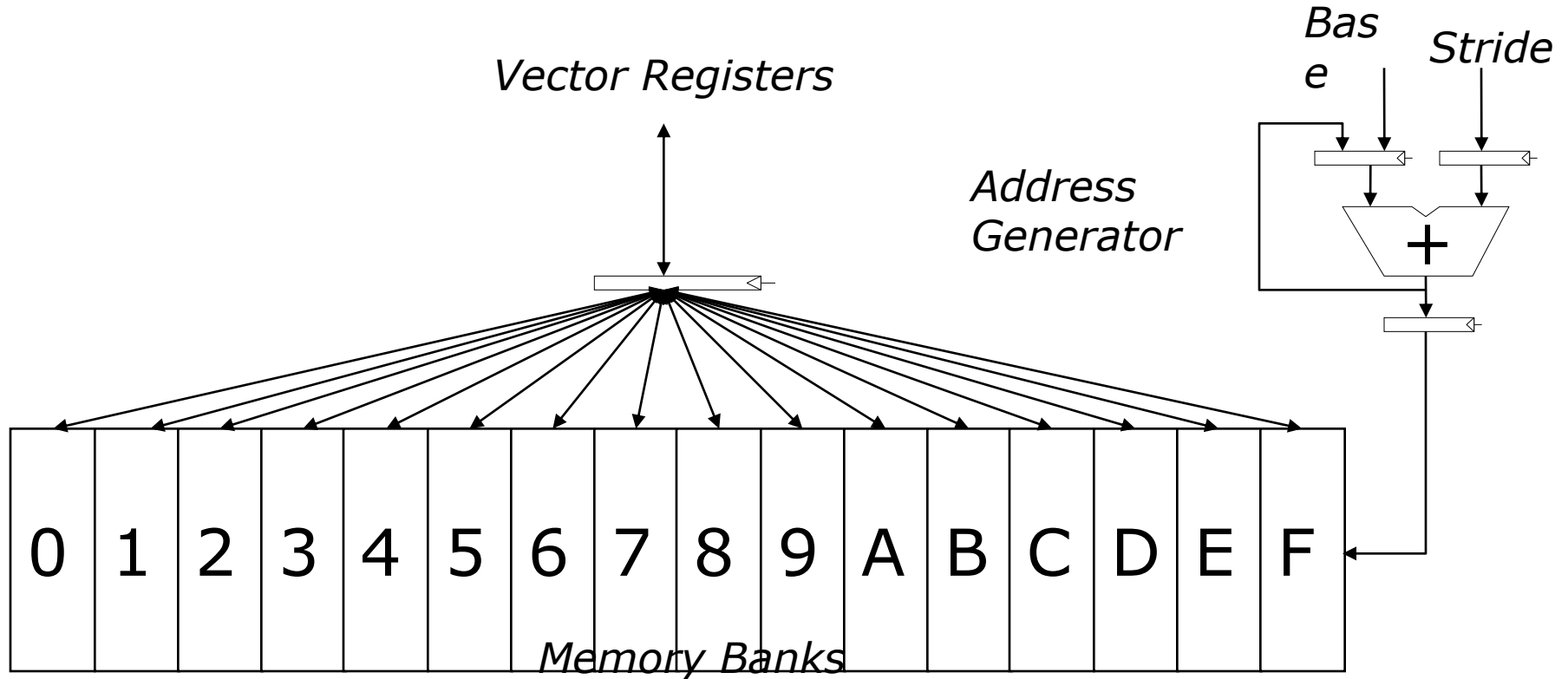
- CRAY-1
- Russell, “The CRAY-1 computer system,” CACM 1978.
- Scalar and vector modes
- 8 64-element vector registers
- 64 bits per element
- 16 memory banks
- 8 64-bit scalar registers
- 8 24-bit address registers

# Memory Banking

- Memory is divided into banks that can be accessed independently; banks share address and data buses
- Can start and complete one bank access per cycle
- Can sustain N parallel accesses if they go to different banks



# Vector Memory System



# Scalar Code Example

---

- For  $I = 0$  to 49
  - $C[i] = (A[i] + B[i]) / 2$
- Scalar code (instruction and its latency)

MOVI R0 = 50	1	
MOVA R1 = A	1	304 dynamic instructions
MOVA R2 = B	1	
MOVA R3 = C	1	
X: LD R4 = MEM[R1++]	11	;autoincrement addressing
LD R5 = MEM[R2++]	11	
ADD R6 = R4 + R5	4	
SHFR R7 = R6 >> 1	1	
ST MEM[R3++] = R7	11	
DECBNZ --R0, X	2	;decrement and branch if NZ

# Scalar Code Execution Time

---

- Scalar execution time on an in-order processor with 1 bank
  - First two loads in the loop cannot be pipelined:  $2 \times 11$  cycles
  - $4 + 50 \times 40 = 2004$  cycles
  
- Scalar execution time on an in-order processor with 16 banks (word-interleaved: consecutive words are stored in consecutive banks)
  - First two loads in the loop can be pipelined
  - $4 + 50 \times 30 = 1504$  cycles
  
- Why 16 banks?
  - 11 cycle memory access latency
  - Having 16 ( $>11$ ) banks ensures there are enough banks to overlap enough memory operations to cover memory latency

# Vectorizable Loops

---

- A loop is **vectorizable** if each iteration is independent of any other
- For I = 0 to 49
  - $C[i] = (A[i] + B[i]) / 2$  7 dynamic instructions

- Vectorized loop:

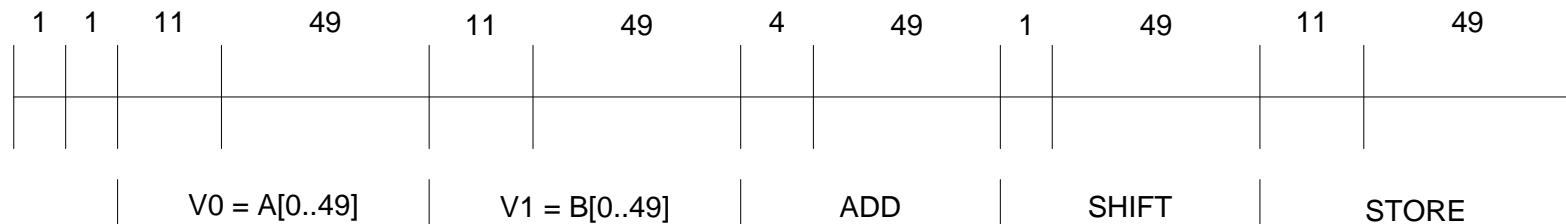
MOVI VLEN = 50	1
MOVI VSTR = 1	1
VLD V0 = A	11 + VLN - 1
VLD V1 = B	11 + VLN - 1
VADD V2 = V0 + V1	4 + VLN - 1
VSHFR V3 = V2 >> 1	1 + VLN - 1
VST C = V3	11 + VLN - 1



# Basic Vector Code Performance

---

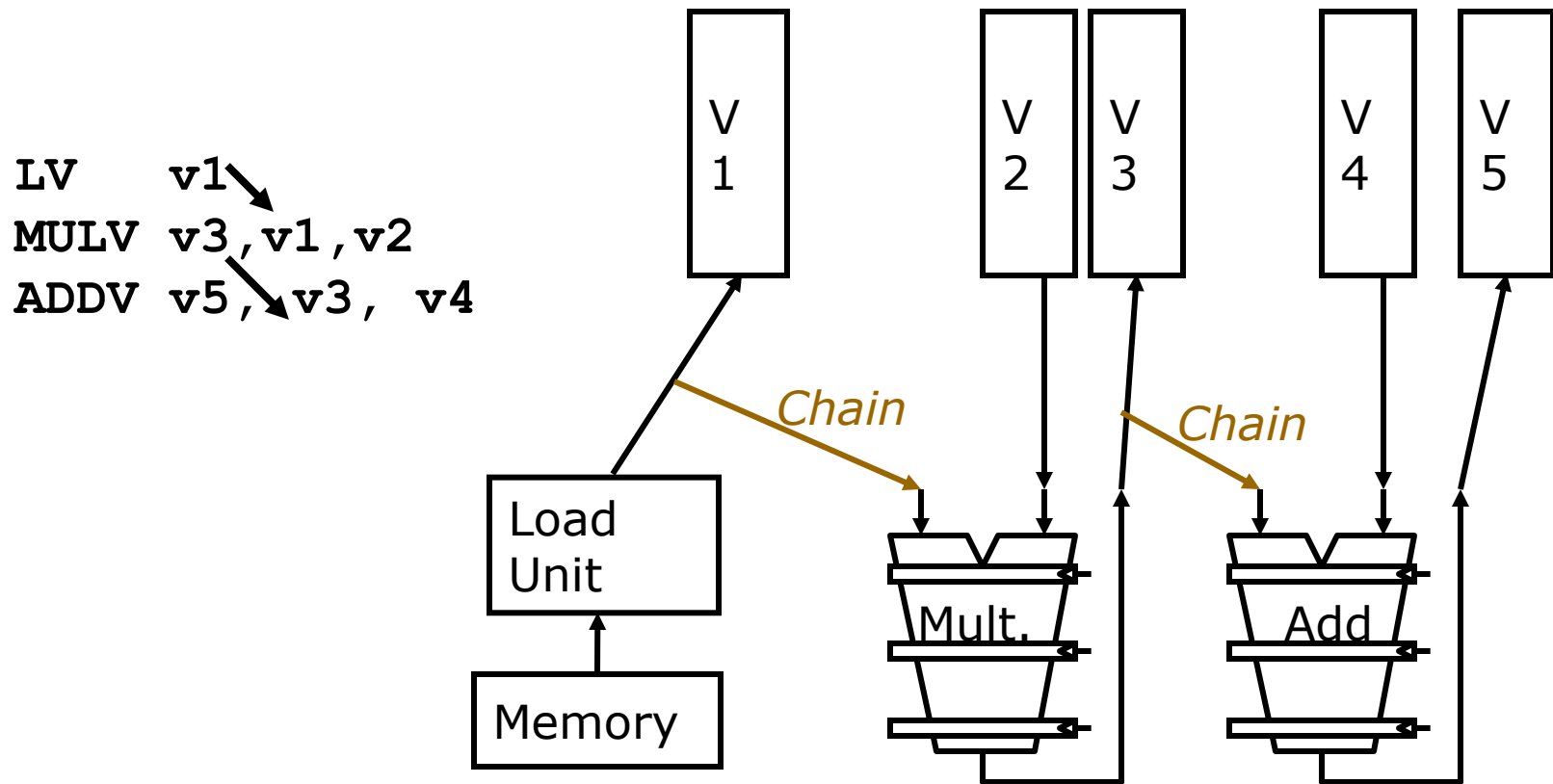
- Assume no chaining (no vector data forwarding)
  - i.e., output of a vector functional unit cannot be used as the direct input of another
  - The entire vector register needs to be ready before any element of it can be used as part of another operation
- One memory port (one address generator)
- 16 memory banks (word-interleaved)



- 285 cycles

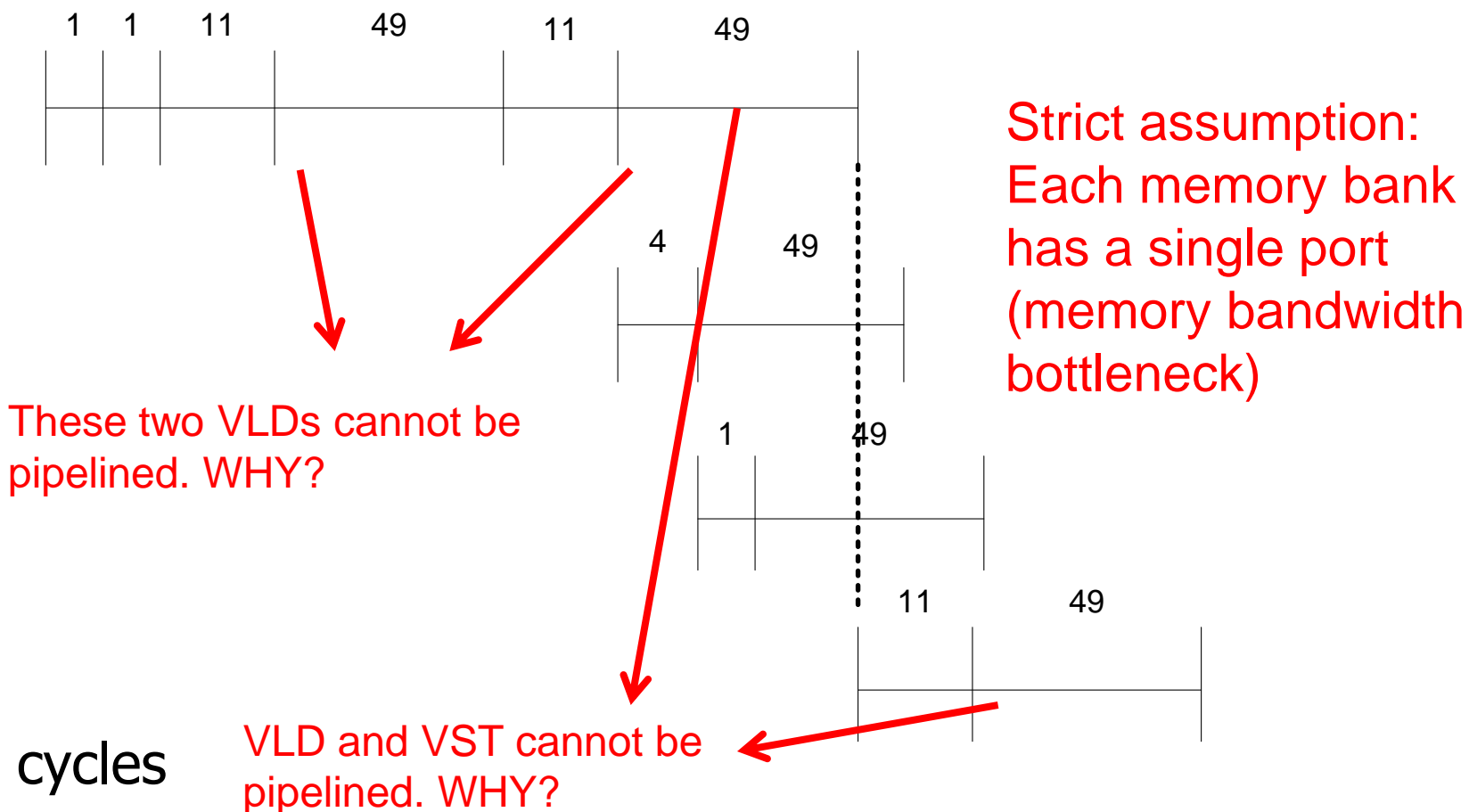
# Vector Chaining

- **Vector chaining:** Data forwarding from one vector functional unit to another



# Vector Code Performance - Chaining

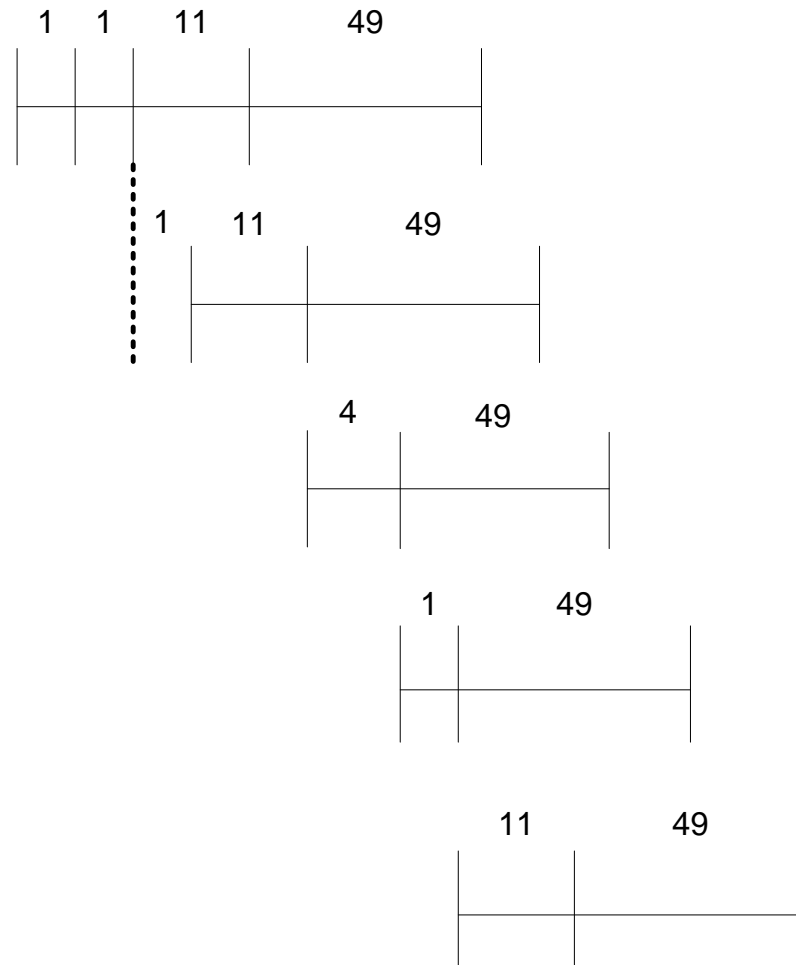
- **Vector chaining:** Data forwarding from one vector functional unit to another



# Vector Code Performance – Multiple Memory Ports

---

- Chaining and 2 load ports, 1 store port in each bank



- 79 cycles

# Questions (I)

---

- What if # data elements  $>$  # elements in a vector register?
  - Need to break loops so that each iteration operates on # elements in a vector register
    - E.g., 527 data elements, 64-element VREGs
    - 8 iterations where  $VLEN = 64$
    - 1 iteration where  $VLEN = 15$  (need to change value of  $VLEN$ )
  - Called **vector stripmining**
  
- What if vector data is not stored in a strided fashion in memory? (irregular memory access to a vector)
  - Use indirection to combine/pack elements into vector registers
  - Called **scatter/gather operations**

# Gather/Scatter Operations

---

Want to vectorize loops with indirect accesses:

```
for (i=0; i<N; i++)  
    A[i] = B[i] + C[D[i]]
```

Indexed load instruction (*Gather*)

```
LV vD, rD          # Load indices in D vector  
LVI vC, rC, vD     # Load indirect from rC base  
LV vB, rB          # Load B vector  
ADDV.D vA, vB, vC  # Do add  
SV vA, rA          # Store result
```

# Gather/Scatter Operations

---

- Gather/scatter operations often implemented in hardware to handle sparse matrices
- Vector loads and stores use an index vector which is added to the base register to generate the addresses

Index Vector	Data Vector	Equivalent
0	3.14	3.14
2	6.5	0.0
6	71.2	6.5
7	2.71	0.0
		0.0
		0.0
		71.2
		2.71

# Conditional Operations in a Loop

---

- What if some operations should not be executed on a vector (based on a dynamically-determined condition)?

```
loop:      if (a[i] != 0) then b[i]=a[i]*b[i]
           goto loop
```

- Idea: **Masked operations**

- VMASK register is a bit mask determining which data element should not be acted upon

```
VLD V0 = A
```

```
VLD V1 = B
```

```
VMASK = (V0 != 0)
```

```
VMUL V1 = V0 * V1
```

```
VST B = V1
```

- Does this look familiar? This is essentially **predicated execution**.



# Another Example with Masking

---

```
for (i = 0; i < 64; ++i)
  if (a[i] >= b[i]) then c[i] = a[i]
  else c[i] = b[i]
```

A	B	VMASK
1	2	0
2	2	1
3	2	1
4	10	0
-5	-4	0
0	-3	1
6	5	1
-7	-8	1

Steps to execute loop

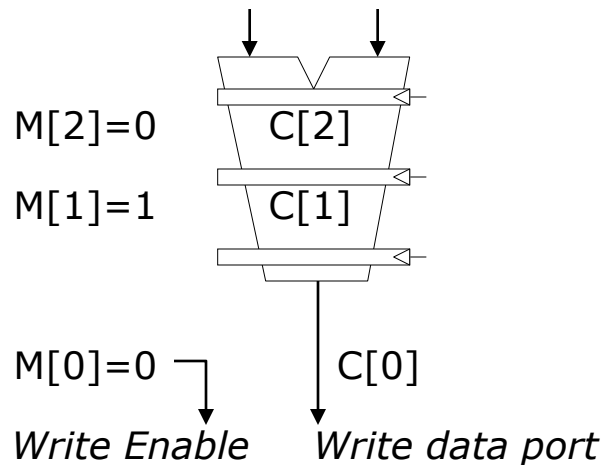
1. Compare A, B to get VMASK
2. Masked store of A into C
3. Complement VMASK
4. Masked store of B into C

# Masked Vector Instructions

## Simple Implementation

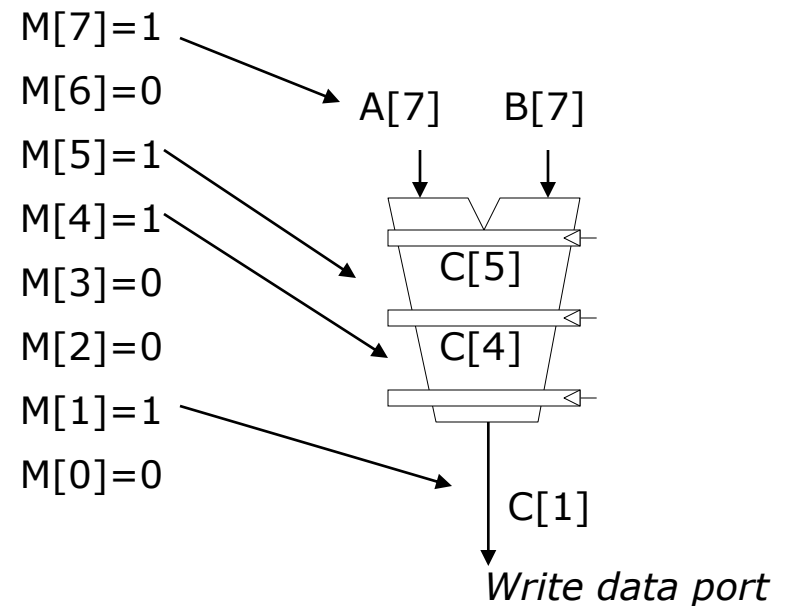
- execute all N operations, turn off result writeback according to mask

M[7]=1 A[7] B[7]  
M[6]=0 A[6] B[6]  
M[5]=1 A[5] B[5]  
M[4]=1 A[4] B[4]  
M[3]=0 A[3] B[3]



## Density-Time Implementation

- scan mask vector and only execute elements with non-zero masks



# Some Issues

---

- Stride and banking
  - As long as they are *relatively prime* to each other and there are enough banks to cover bank access latency, consecutive accesses proceed in parallel
  
- Storage of a matrix
  - **Row major**: Consecutive elements in a row are laid out consecutively in memory
  - **Column major**: Consecutive elements in a column are laid out consecutively in memory
  - You need to change the stride when accessing a row versus column

## Matrix multiplication

A & B, both in row major order

A<sub>0</sub>

0	1	2	3	4	5
6	7	8	9	10	11

B<sub>0</sub>

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20									
30									
40									
50									

$A_{4 \times 6} B_{6 \times 10} \rightarrow C_{4 \times 10}$  (dot products of rows & columns of A & B)

A: Load A<sub>0</sub> into a vector register V1  
→ each time you need to increment the address by 1 to access the next column  
→ First matrix accesses have a stride of 1

B: Load B<sub>0</sub> into a vector register V2  
→ each time you need to increment by 10  
→ stride of 10

Different strides can lead to bank conflicts.

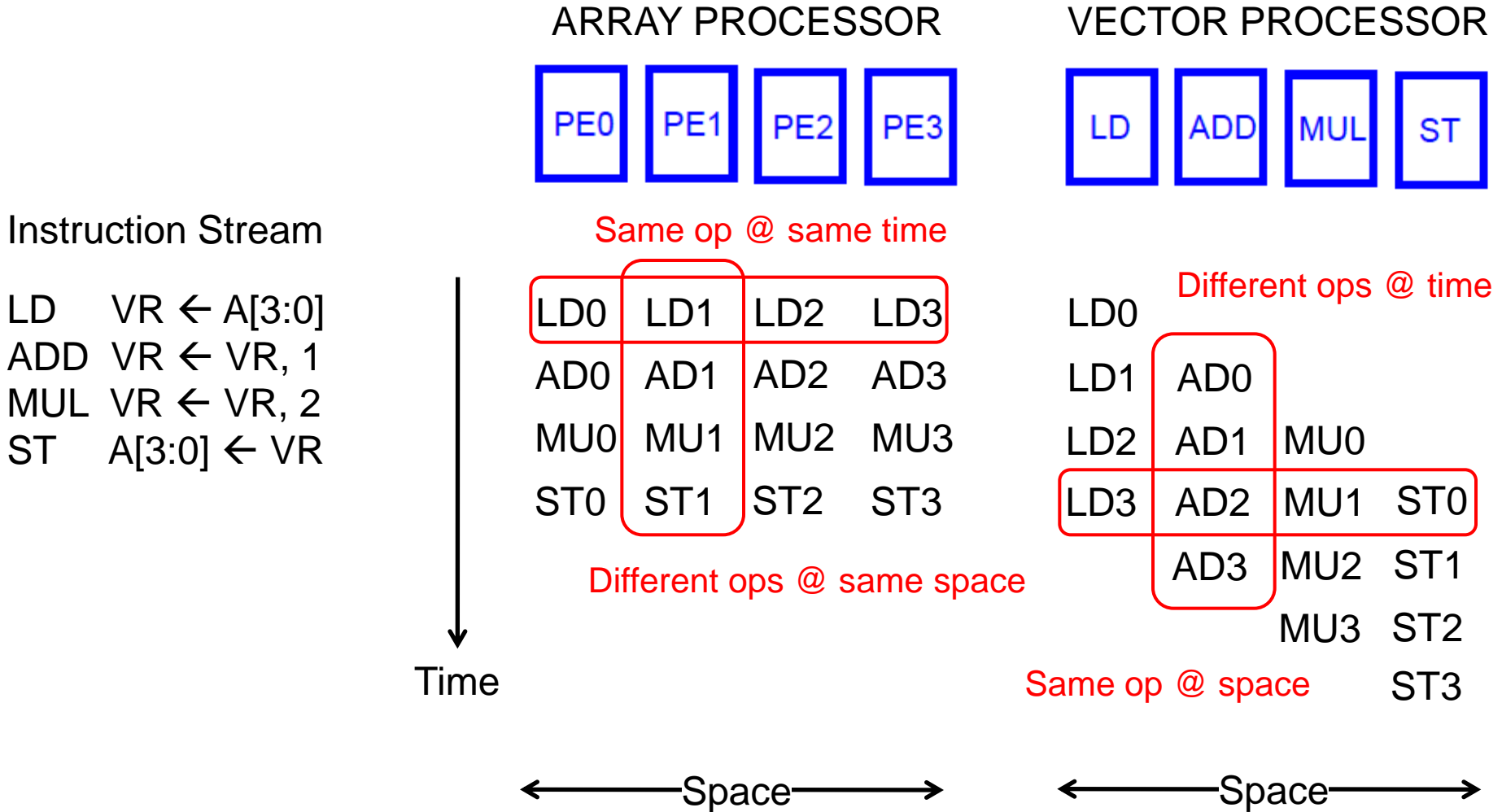
→ How do you minimize them?

# Array vs. Vector Processors, Revisited

---

- Array vs. vector processor distinction is a “purist’s” distinction
- Most “modern” SIMD processors are a combination of both
  - They exploit data parallelism in both time and space

# Remember: Array vs. Vector Processors



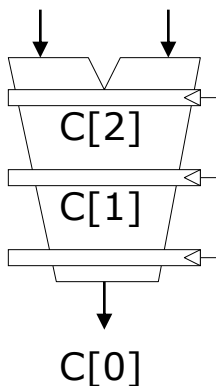
# Vector Instruction Execution

ADDV C,A,B

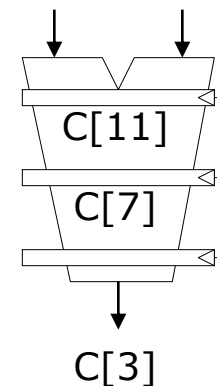
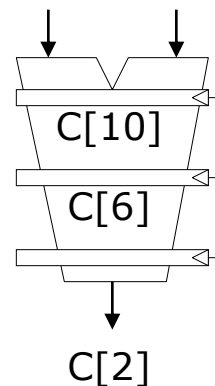
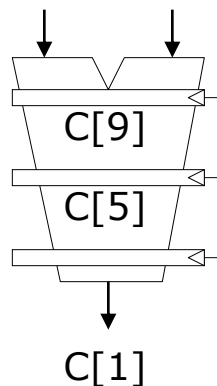
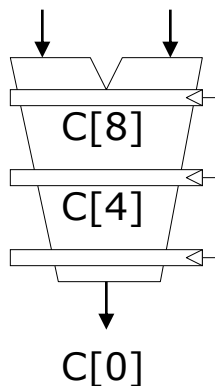
*Execution using  
one pipelined  
functional unit*

*Execution using  
four pipelined  
functional units*

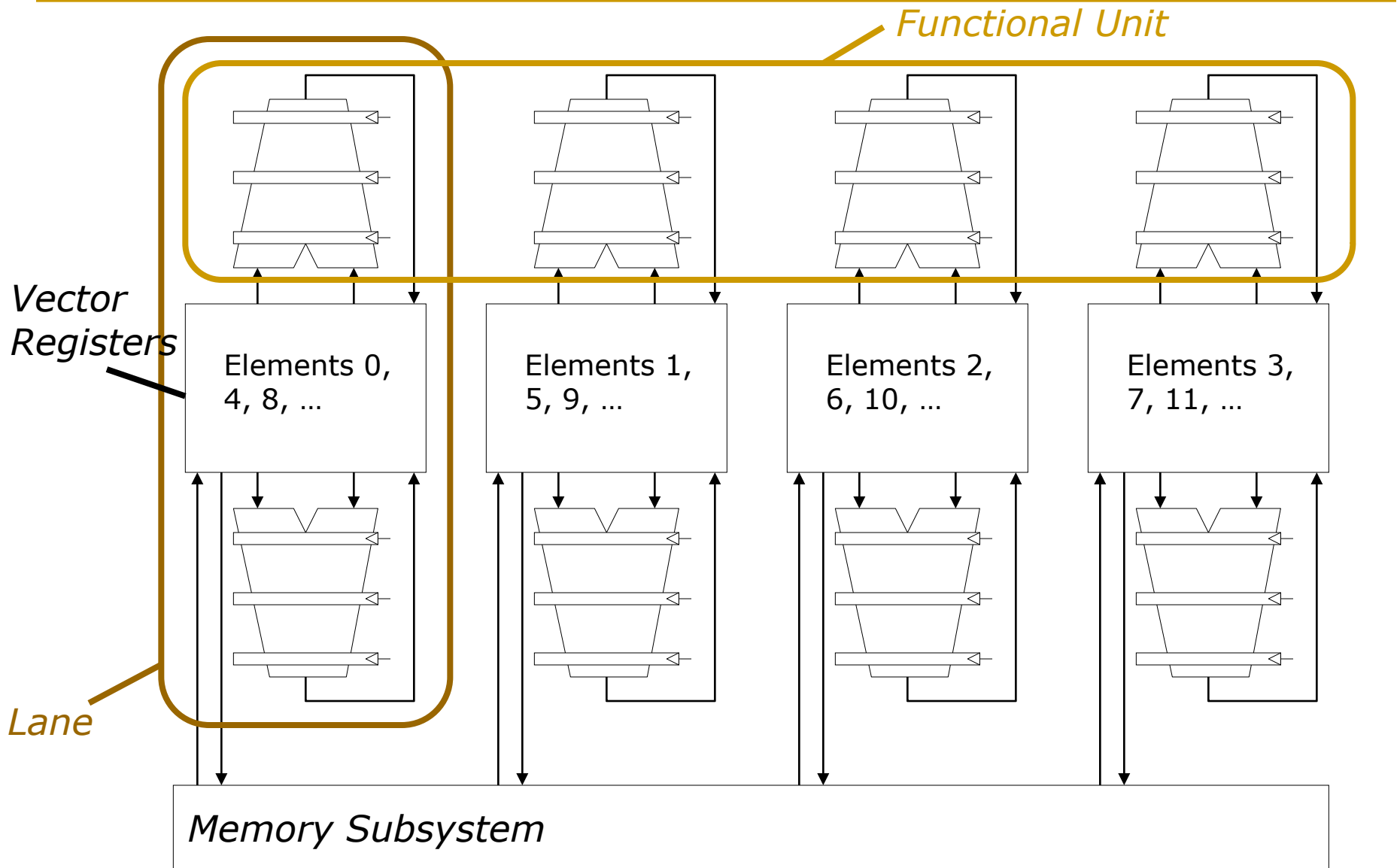
A[6] B[6]  
A[5] B[5]  
A[4] B[4]  
A[3] B[3]



A[24] B[24] A[25] B[25] A[26] B[26] A[27] B[27]  
A[20] B[20] A[21] B[21] A[22] B[22] A[23] B[23]  
A[16] B[16] A[17] B[17] A[18] B[18] A[19] B[19]  
A[12] B[12] A[13] B[13] A[14] B[14] A[15] B[15]



# Vector Unit Structure

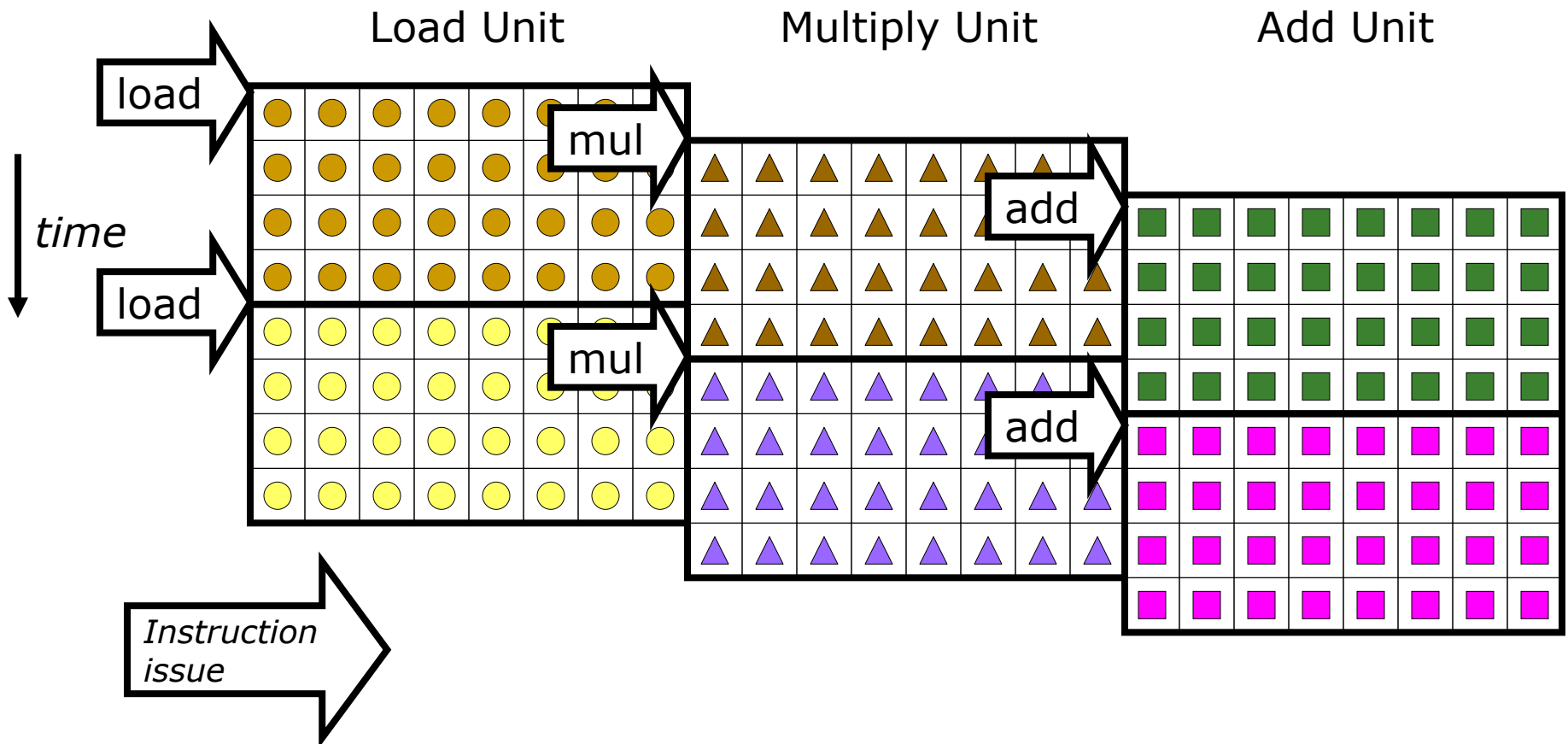




# Vector Instruction Level Parallelism

Can overlap execution of multiple vector instructions

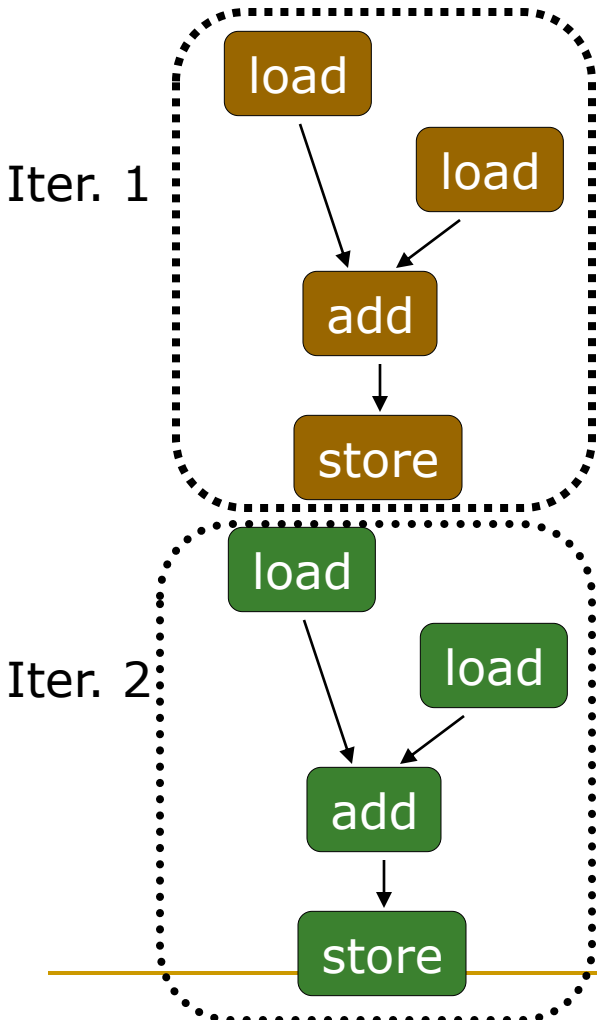
- example machine has 32 elements per vector register and 8 lanes
- Complete 24 operations/cycle while issuing 1 short instruction/cycle



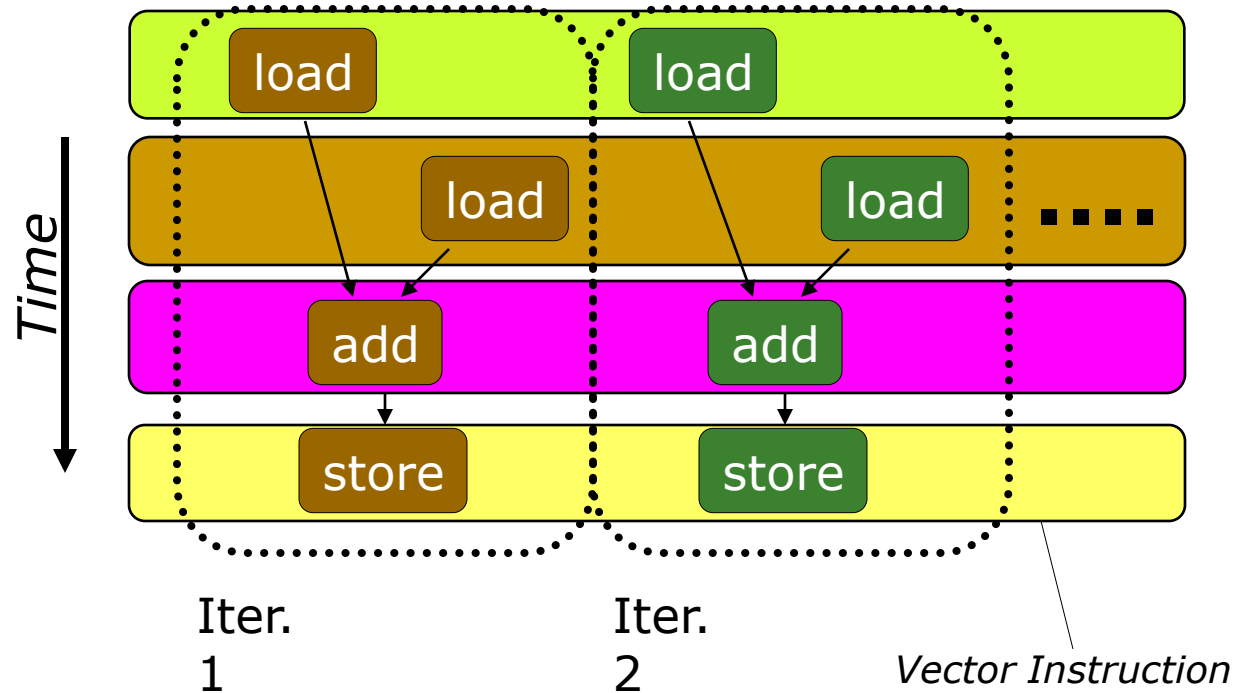
# Automatic Code Vectorization

```
for (i=0; i < N; i++)  
  C[i] = A[i] + B[i];
```

*Scalar Sequential Code*



*Vectorized Code*



Vectorization is a compile-time reordering of operation sequencing  
⇒ requires extensive loop dependence analysis

# Vector/SIMD Processing Summary

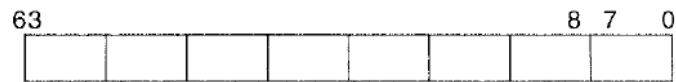
---

- Vector/SIMD machines are good at exploiting **regular data-level parallelism**
  - Same operation performed on many data elements
  - Improve performance, simplify design (no intra-vector dependencies)
- **Performance improvement limited by vectorizability** of code
  - Scalar operations limit vector machine performance
  - Amdahl's Law
  - CRAY-1 was the fastest SCALAR machine at its time!
- Many existing ISAs include (vector-like) SIMD operations
  - Intel MMX/SSEn/AVX, PowerPC AltiVec, ARM Advanced SIMD

# SIMD Operations in Modern ISAs

# Intel Pentium MMX Operations

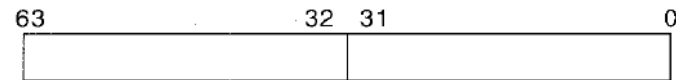
- Idea: One instruction operates on multiple data elements **simultaneously**
  - Ala array processing (yet much more limited)
  - Designed with multimedia (graphics) operations in mind



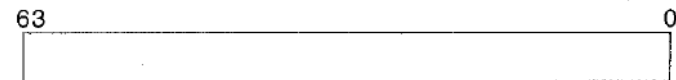
(a)



(b)



(c)



(d)

No VLEN register

Opcode determines data type:

8 8-bit bytes

4 16-bit words

2 32-bit doublewords

1 64-bit quadword

Stride always equal to 1.

Peleg and Weiser, “**MMX Technology Extension to the Intel Architecture,**”  
IEEE Micro, 1996.

Figure 1. MMX technology data types: packed byte (a), packed word (b), packed doubleword (c), and quadword (d).

# MMX Example: Image Overlaying (I)



Figure 8. Chroma keying: image overlay using a background color.

PCMPEQB MM1, MM3

MM1	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
MM3	X7!=blue	X6!=blue	X5=blue	X4=blue	X3!=blue	X2!=blue	X1=blue	X0=blue
MM1	0x0000	0x0000	0xFFFF	0xFFFF	0x0000	0x0000	0xFFFF	0xFFFF



Bitmask

Figure 9. Generating the selection bit mask.

# MMX Example: Image Overlaying (II)

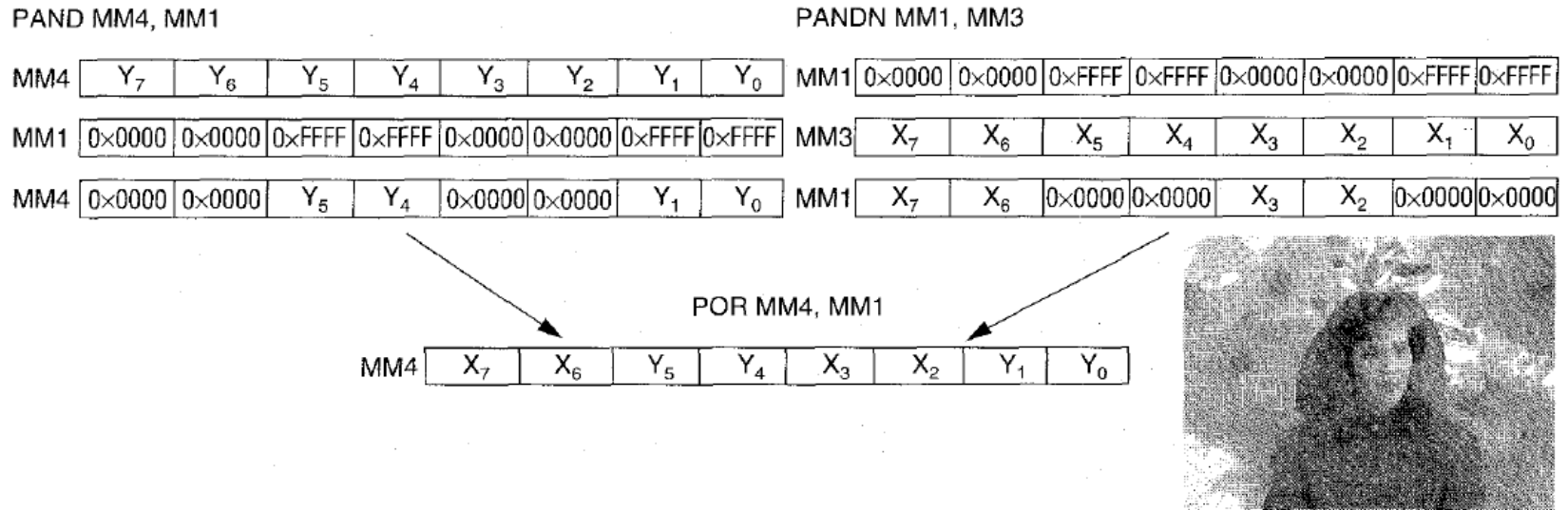


Figure 10. Using the mask with logical MMX instructions to perform a conditional select.

```

Movq    mm3, mem1    /* Load eight pixels from
                    woman's image
Movq    mm4, mem2    /* Load eight pixels from the
                    blossom image
Pcmpeqb mm1, mm3
Pand    mm4, mm1
Pandn   mm1, mm3
Por     mm4, mm1
    
```

Figure 11. MMX code sequence for performing a conditional select.