

STAM Center
SECURE, TRUSTED, AND ASSURED MICROELECTRONICS

ASU Engineering
Arizona State University

CSE 520 Computer Architecture II

Complex Pipelining: SIMD and Vector Processors

Prof. Michel A. Kinsy

1

STAM Center
SECURE, TRUSTED, AND ASSURED MICROELECTRONICS

ASU Engineering
Arizona State University

Vector Architectures

- Basic idea
 - Read sets of data elements into "vector registers"
 - Operate on those registers
 - Disperse the results back into memory
- Registers are controlled by compiler
 - Used to hide memory latency
 - Leverage memory bandwidth
- Scalar Unit
 - Load/Store Architecture
- Vector Extension
 - Vector Registers
 - Vector Instructions

2

STAM Center
SECURE, TRUSTED, AND ASSURED MICROELECTRONICS

ASU Engineering
Arizona State University

Vector Processing

- Vector processors have high-level operations that work on linear arrays of numbers
 - Called "vectors"

SCALAR
(1 operation)

add x3, x1, x2

VECTOR
(K operations)

add.vv v3, v1, v2

3

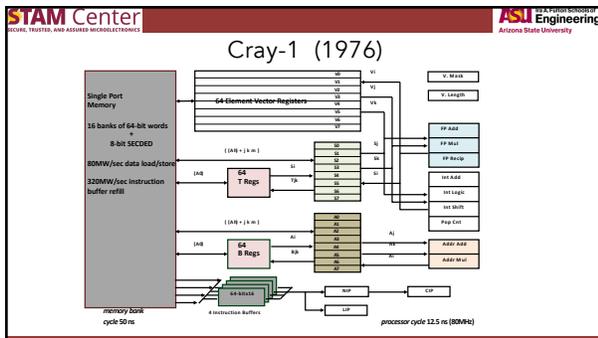
STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS

ASU Engineering ARIZONA STATE UNIVERSITY

Vector Architectures

- Epitomized by Cray-1, 1976
- Types of vector processors
 - Memory-memory processors: all vector operations are memory-to-memory (CDC)
 - Vector-register processors: all vector operations except load and store are among the vector registers
 - (CRAY-1, CRAY-2, X-MP, Y-MP, NEX SX/2(3), Fujitsu)
- Implementation
 - Hardwired Control
 - Highly Pipelined Functional Units
 - Interleaved Memory System
 - No Data Caches
 - No Virtual Memory

4



5

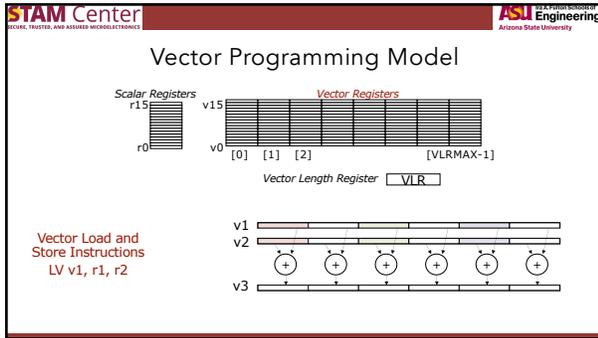
STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS

ASU Engineering ARIZONA STATE UNIVERSITY

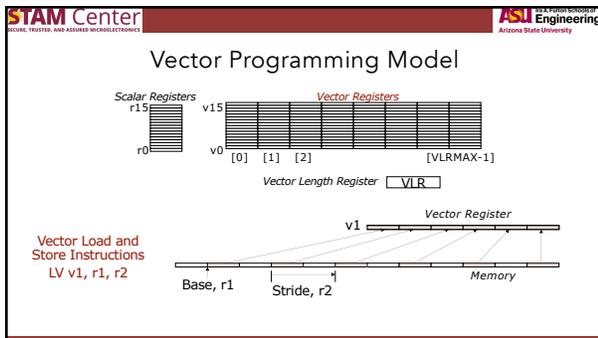
Properties of Vector Processors

- Each result independent of previous result
 - Long pipeline, compiler ensures no dependencies
 - High clock rate
- Vector instructions access memory with known pattern
 - Highly interleaved memory
 - Amortize memory latency of over 64 elements
 - No data caches required
 - Only use instruction cache
- Reduces branches and branch problems in pipelines
 - Single vector instruction implies lots of work (loop)
 - Fewer instruction fetches

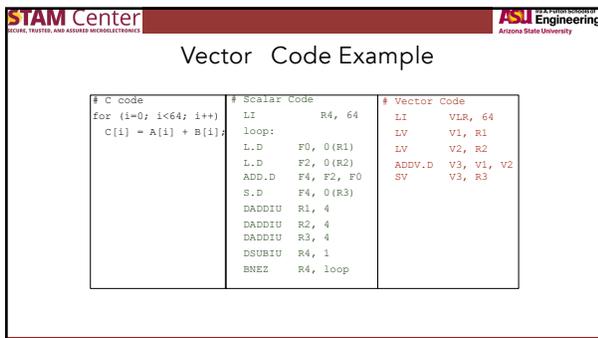
6



7



8



9

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS

ASU Engineering ARIZONA STATE UNIVERSITY

Vector Instruction Set Advantages

- Compact
 - one short instruction encodes N operations
- Expressive, tells hardware that these N operations:
 - are independent
 - use the same functional unit
 - access disjoint registers
 - access registers in same pattern as previous instructions

10

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS

ASU Engineering ARIZONA STATE UNIVERSITY

Vector Instruction Set Advantages

- Compact
 - One short instruction encodes N operations
- Expressive, tells hardware that these N operations:
 - Access a contiguous block of memory
 - Unit-stride load/store
 - Access memory in a known pattern
 - Strided load/store
- Scalable
 - Can run same code on more parallel pipelines (lanes)

11

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS

ASU Engineering ARIZONA STATE UNIVERSITY

Vector Arithmetic Execution

- Use deep pipeline to execute element operations
 - Fast clock
- Simplifies control of deep pipeline because elements in vector are independent
 - No hazards!

Six stage multiply pipeline

$V3 \leftarrow v1 * v2$

12

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** A College of the Arizona State University

Vector Instruction Execution

- ADDV C,A,B

Execution using one pipelined functional unit

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

Execution using four pipelined functional units

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]

13

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** A College of the Arizona State University

Vector Unit Structure

Elements 0, 4, 8, ...

Elements 1, 5, 9, ...

Elements 2, 6, 10, ...

Elements 3, 7, 11, ...

Memory Subsystem

14

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** A College of the Arizona State University

Vector Unit Structure

Vector Registers

Elements 0, 4, 8, ...

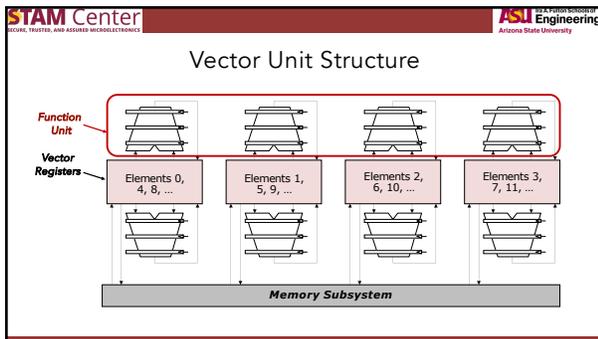
Elements 1, 5, 9, ...

Elements 2, 6, 10, ...

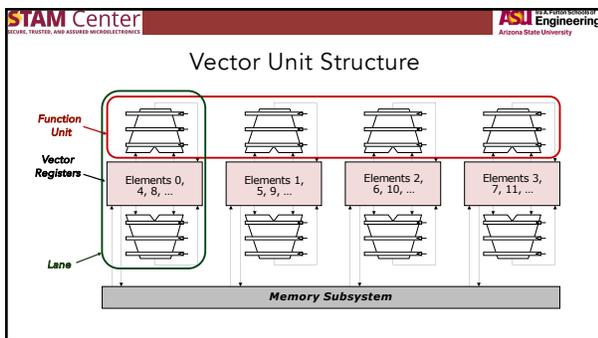
Elements 3, 7, 11, ...

Memory Subsystem

15



16



17

Memory Banks

- Memory system must be designed to support high bandwidth for vector loads and stores
- Spread accesses across multiple banks
 - Control bank addresses independently
 - Load or store non sequential words
 - Support multiple vector processors sharing the same memory
- Example
 - 32 processors, each generating 4 loads and 2 stores/cycle
 - Processor cycle time is 2.167 ns, SRAM cycle time is 15 ns
 - How many memory banks needed?
 - $32 \times 6 = 192$ accesses,
 - $15 / 2.167 \approx 7$ processor cycles
 - $\rightarrow 1344!$

18

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Vector Memory System

- Cray-1, 16 banks, 4 cycle bank busy time, 12 cycle latency
 - Bank busy time: Cycles between accesses to same bank

19

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Stride

- Consider


```

for (i = 0; i < 100; i=i+1) {
  for (j = 0; j < 100; j=j+1) {
    A[i][j] = 0.0;
    for (k = 0; k < 100; k=k+1)
      A[i][j] = A[i][j] + B[i][k] * D[k][j];
  }
}
            
```
- Must vectorize multiplication of rows of B with columns of D
- Use *non-unit stride*
- Bank conflict (stall) occurs when the same bank is hit faster than bank busy time:
 - #banks / LCM(stride, #banks) < bank busy time (in # of cycles)

20

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Stride

- Example
- 8 memory banks with a bank busy time of 6 cycles and a total memory latency of 12 cycles. How long will it take to complete a 64-element vector load with a stride of 1? With a stride of 32?
- **Answer**
 - Stride of 1: number of banks is greater than the bank busy time, so it takes
 - 12+64 = 76 clock cycles → 1.2 cycle per element
 - Stride of 32: the worst-case scenario happens when the stride value is a multiple of the number of banks, which this is! Every access to memory will collide with the previous one! Thus, the total time will be:
 - 12 + 1 + 6 * 63 = 391 clock cycles, or 6.1 clock cycles per element!

21

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** A CLAYTON SCHOOL OF Arizona State University

Vector Instruction 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

22

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** A CLAYTON SCHOOL OF Arizona State University

Vector Chaining or Bypassing

- Vector version of register bypassing
 - Introduced with Cray-1

LV v1
MULV v3,v1,v2
ADDV v5,v3,v4

23

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** A CLAYTON SCHOOL OF Arizona State University

Vector Chaining Advantage

- Without chaining, must wait for last element of result to be written before starting dependent instruction

- With chaining, can start dependent instruction as soon as first result appears

24

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Vector Conditional Execution

- Problem: Want to vectorize loops with conditional code:


```
for (i=0; i<N; i++)
  if (A[i]>0) then
    A[i] = B[i];
```
- Solution: Add vector mask (or flag) registers
 - vector version of predicate registers, 1 bit per element
 - vector operation becomes NOP at elements where mask bit is clear
- Code example:


```
CVM          # Turn on all elements
LV VA, rA    # Load entire A vector
SGTUS:0 VA, F0 # Set bits in mask register where A>0
LV VA, rB    # Load B vector into A under mask
SV VA, rA    # Store A back to memory under mask
```

25

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Vector Processor

- Vector Register: fixed length bank holding a single vector
 - Has at least 2 read and 1 write ports
 - Typically, 8-32 vector registers, each holding 64-128 64-bit elements
- Vector Functional Units (FUs): fully pipelined, start new operation every clock
 - Typically, 4 to 8 FUs: FP add, FP mult, FP reciprocal (1/X), integer add, logical, shift; may have multiple of same unit
- Vector Load-Store Units (LSUs): fully pipelined unit to load or store a vector; may have multiple LSUs
- Scalar registers: single element for FP scalar or address
- Cross-bar to connect FUs, LSUs, registers

26

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Modern Vector: NEC SX-6 (2003)

- CMOS Technology
 - 500 MHz CPU, fits on single chip
 - SDRAM main memory (up to 64GB)
- Scalar unit
 - 4-way superscalar, with out-of-order and speculative execution
 - 64KB I-cache and 64KB data cache
- SMP structure
 - 8 CPUs connected to memory through crossbar
 - 256 GB/s shared memory bandwidth (4096 interleaved banks)



27

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Modern Vector: NEC SX-6 (2003)

- Vector unit
 - 8 foreground VRegs + 64 background VRegs (256x64-bit elements/VReg)
 - 1 multiply unit, 1 divide unit, 1 add/shift unit, 1 logical unit, 1 mask unit
 - 8 lanes (8 GFLOPS peak, 16 FLOPS/cycle)
 - 1 load & store unit (32x8 byte accesses/cycle)
 - 32 GB/s memory bandwidth per processor

28

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Larrabee x86 with vectors

- Short in-order instruction pipeline
- Separate scalar and vector units and register sets
 - Vector unit: 16 32-bit ops/clock
- Fast access to L1 cache
- L1 connects to core's portion of the L2 cache

29

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Larrabee Vector Architecture

- Data types
 - Int32, Float32 and Float64 data
- Vector operations
 - Two input/one output operations
 - Full complement of arithmetic and media operations
 - Fused multiply-add (three input arguments)
 - Mask registers select lanes to write
 - Swizzle the vector elements on register read

30

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Larrabee Vector Architecture

- Memory access
 - Vector load/store including scatter/gather
 - Data replication on read from memory
 - Numeric type conversion on memory read

31

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Larrabee Motivation

- Design experiment: not a real 10-core chip!

# CPU cores	2 out of order	10 in-order
Instructions per issue	4 per clock	2 per clock
VPU lanes per core	4-wide SSE	16-wide
L2 cache size	4 MB	4 MB
Single-stream	4 per clock	2 per clock
Vector throughput	8 per clock	160 per clock

- 20 times the multiply-add operations per clock

Data in chart taken from Seiler, L., Carmean, D., et al. 2008. Larrabee: A many-core x86 architecture for visual computing.

32

STAM Center SECURE, TRUSTED, AND ASSURED MICROELECTRONICS **ASU Engineering** Arizona State University

Some Applications

- Multimedia Processing (compress., graphics, audio synth, image proc.)
- Standard benchmark kernels (Matrix Multiply, FFT, Convolution, Sort)
- Lossy Compression (JPEG, MPEG video and audio)
- Lossless Compression (Zero removal, RLE, Differencing, LZW)
- Cryptography (RSA, DES/IDEA, SHA/MD5)
- Speech and handwriting recognition
- Operating systems/Networking (memcpy, memset, parity, checksum)
- Databases (hash/join, data mining, image/video serving)
- Language run-time support (stdlib, garbage collection)

33

STAM Center
SECURE, TRUSTED, AND ASSURED MICROELECTRONICS

ASU Arizona State University
Engineering

Next Class

- Hardware Multithreading

34