




SIMD in modern computers

HW5 out today (due 4/16)

Project info coming soon (report due 5/3)





What are the limits of vector processors?

Amdahl's law

Used to assess theoretical effectiveness of speedup

In a nutshell: gains in speeding up a portion of a program are limited by the fraction of time that portion is actually used

Mathematically:

$$S_{\text{latency}}(s) = \frac{1}{(1 - p) + \frac{p}{s}}$$

For parallelization: serial bottleneck (non-parallelizable code) limits effectiveness of vector processors

Compiler effectiveness

Processor	Compiler	Completely vectorized	Partially vectorized	Not vectorized
CDC CYBER 205	VAST-2 V2.21	62	5	33
Convex C-series	FC5.0	69	5	26
Cray X-MP	CFT77 V3.0	69	3	28
Cray X-MP	CFT V1.15	50	1	49
Cray-2	CFT2 V3.1a	27	1	72
ETA-10	FTN 77 V1.0	62	7	31
Hitachi S810/820	FORT77/HAP V20-2B	67	4	29
IBM 3090/VF	VS FORTRAN V2.4	52	4	44
NEC SX/2	FORTAN77 / SX V.040	66	5	29

Figure G.9 Result of applying vectorizing compilers to the 100 FORTRAN test kernels. For each processor we indicate how many loops were completely vectorized, partially vectorized, and unvectorized. These loops were collected by Callahan, Dongarra, and Levine [1988]. Two different compilers for the Cray X-MP show the large dependence on compiler technology.

Cray-1 Architecture (1976)

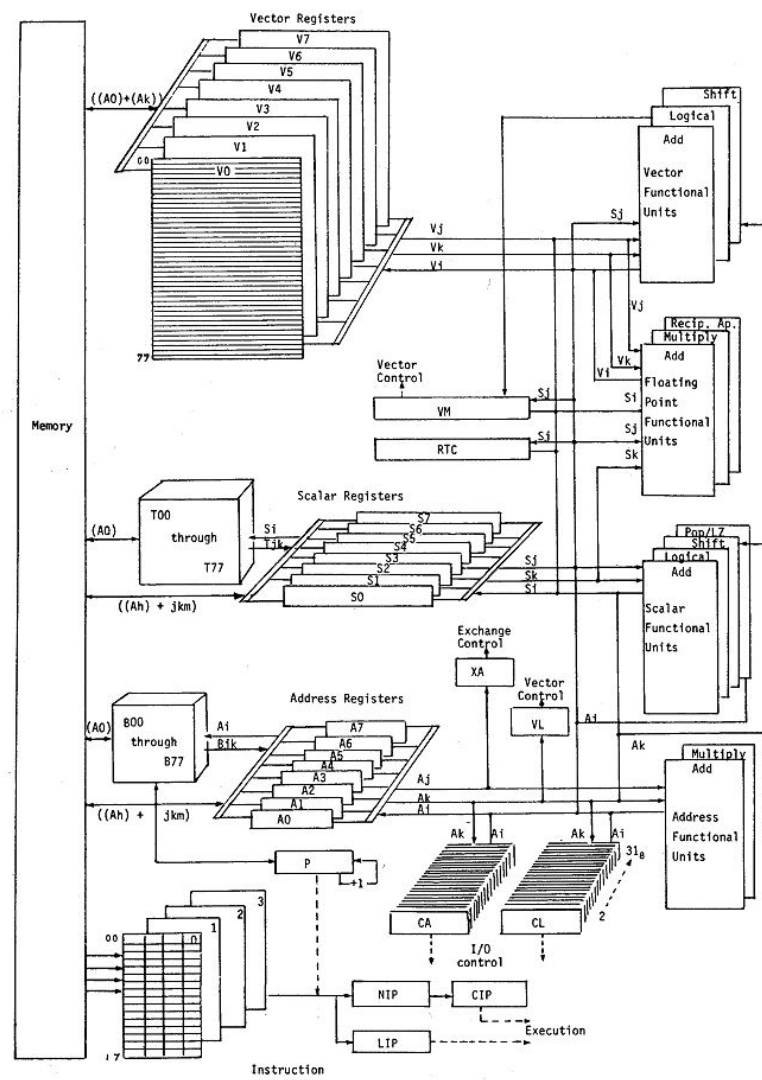


image source

Cray X1 architecture (2003)

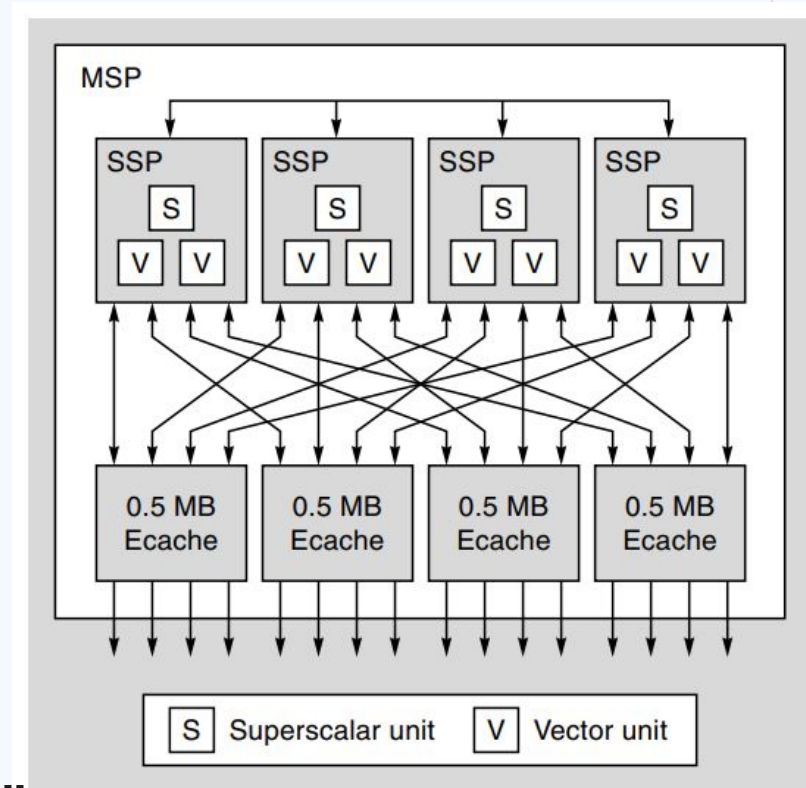
ISA designed from scratch

Multi-stream processor consisting of four single-stream processors

Each SSP has: scalar unit/scalar cache, 2-lane vector unit

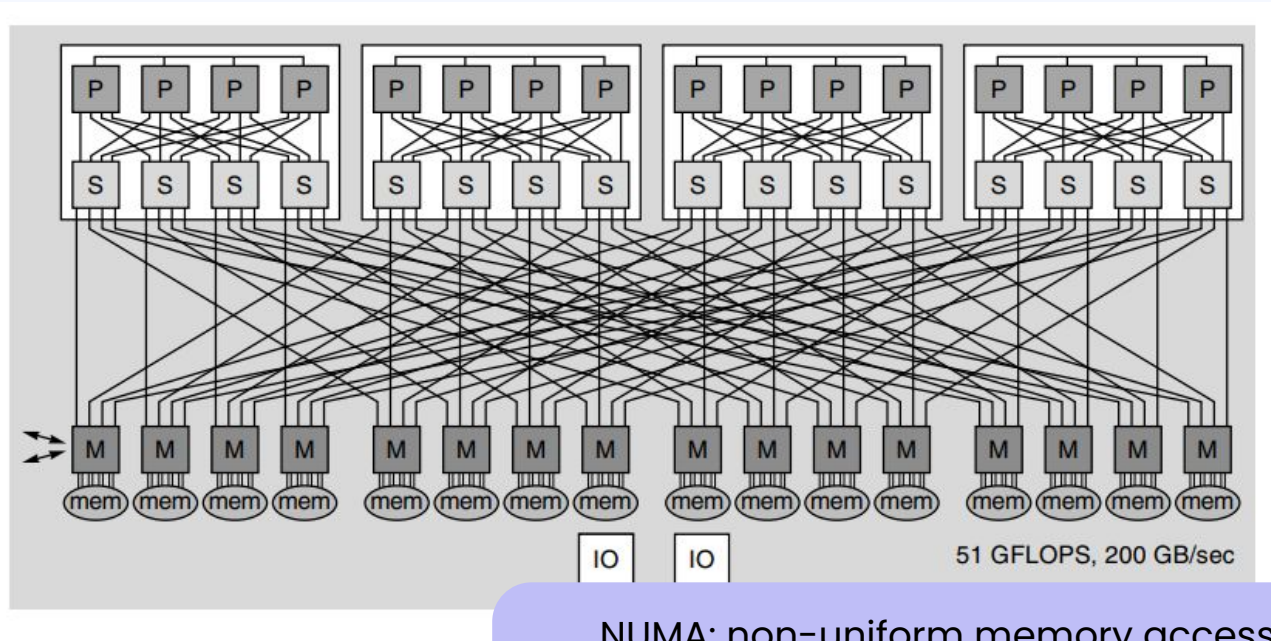
Connected to external caches (mostly for scalars, but can be used by vectors for programs w/ high temporal locality, or bypassed)

Each MSP can have up to 2048 outstanding memory requests



H&P fig. G.11

Cray X1 nodes (H&P fig. G.12)

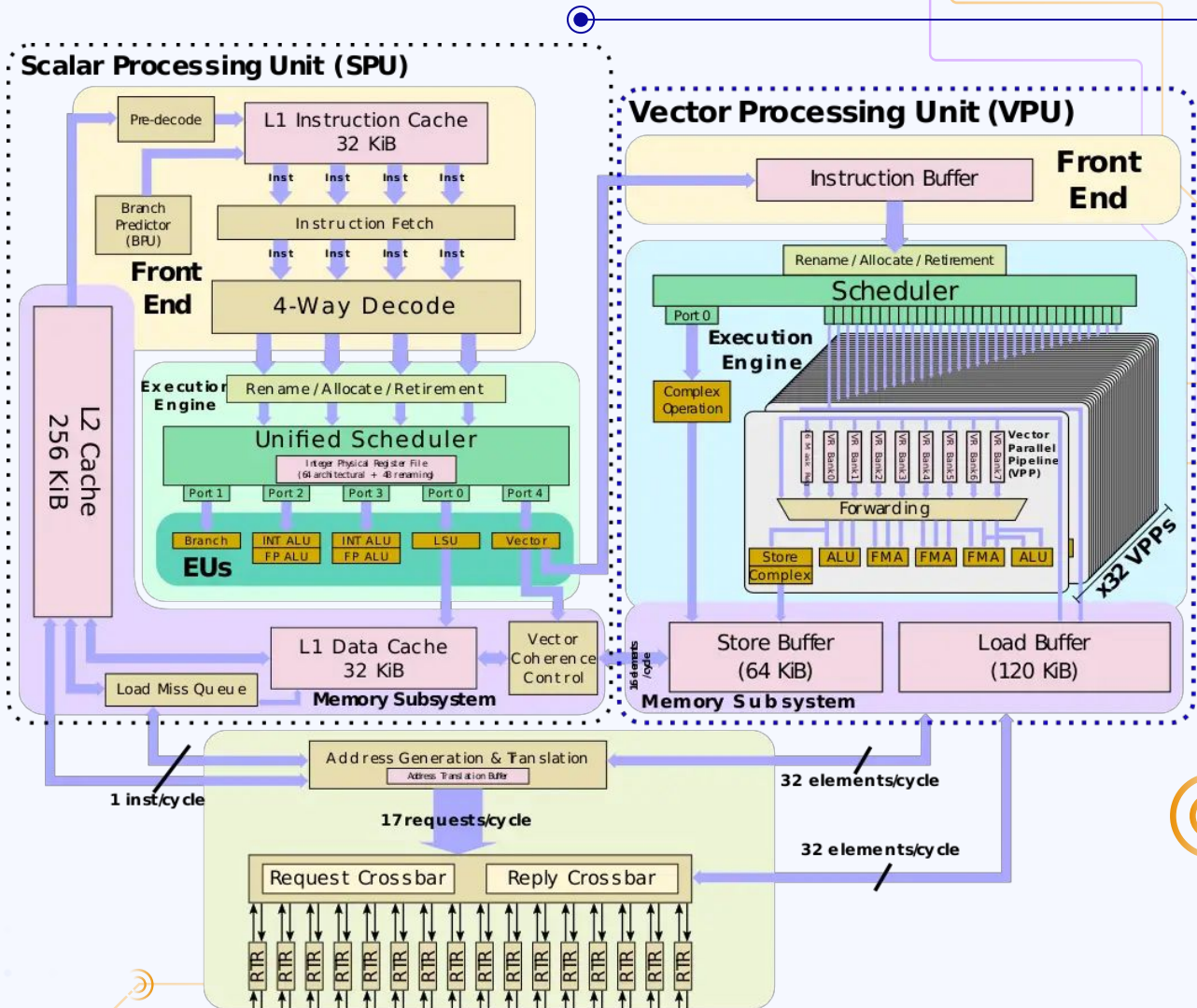


NUMA: non-uniform memory access
eaches only cache their local memory
writes to remote memory invalidate
corresponding ecache data

NEC SX Aurora (2018)

image source

Shared LLC (last-level cache) with 128 banks
Uses high-bandwidth memory (HbM)



New York DFS to acquire supercomputer to understand and regulate AI

Also looking

July 05, 2023 By: C



source

New York's Dep
dedicated to m

Stephen Nellis
Thu, Feb 1, 2024 • 2 min read

In This Article:

CDNS -1.32%

SNPS -2.03%

By Stephen Nellis

source

(Reuters) - Cadence Design Systems on Thursday said it has designed a

OK fine.. but what about DLP for the rest of us?

apps used to design and test the larger mechanical systems that those chips become a part of.

NOAA completes upgrade to weather and forecast system



significant upgrade to the 'American' forecast model

Share:

source



iminate operational supercomputing system - just received a 20% upgrade. ar now operates at a speed of 14.5 petaflops. (Image credit: General

(also headlines about Italy, India, Argentina in the past year)

SIMD for multimedia

RGBA images: 8 bits/channel (32 bits total)

Audio: 8, 16, 24, or 32 bits per sample

Simplifications of SIMD for multimedia: might not have strided access, gather/scatter, masked operations

→ Doesn't typically make sense to put a powerful VPU on a processor

Enter multimedia SIMD extensions

How can smaller data widths make SIMD functionality easier to add to CPUs?

RISC-V P: packed SIMD

(Doesn't actually exist, but the letter "P" is reserved for such a thing)

Reuses floating-point registers

Packs multiple values in one register based on configuration

Ex: 64-bit register can hold 8 8-bit values, 4 16-bit values, 2 32-bit values, or 1 64-bit value

Requires special load/store operations

Hardware support for parallel operation on each value in register

ARMv6 SIMD

Packs multiple 16- or 8-bit values into 32 bit registers

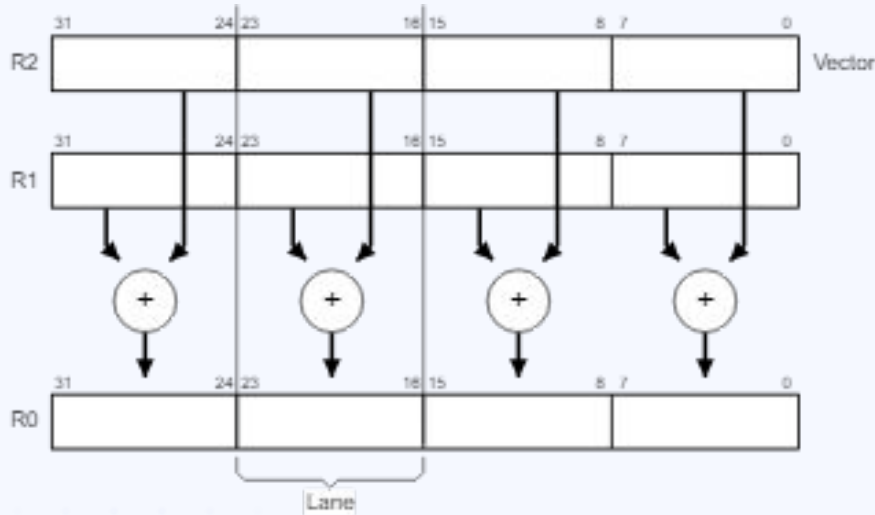


image source

Note: later ARM chips use NEON (their “Advanced SIMD” extension), storing vectors in 64- and 128-bit registers

Use of ARM NEON

Compilers are sometimes hit-or-miss when figuring out if they can vectorize code

Multimedia applications: people can use libraries

To get more flexibility than a library, ARM provides intrinsics

x86: MMX, SSE, AVX

MMX: not an acronym, packs values in 64-bit registers, supports integer operations only

SSE: "Streaming SIMD Extensions", 128-bit registers, allows for floating point

AVX: "Advanced Vector Extensions", 8x32 or 4x64 vector registers (AVX 2 adds gather, AVX 512 supports 512-bit registers)

In typical x86 fashion, operand size is fixed in the opcode (so there are hundreds of instructions for each extension)

x86 AVX-512 VNNI

Vector Neural Network
Instructions
Useful for CNNs
(Convolutional Neural
Networks)

Input image

9	4	1	2	2
1	1	1	0	4
1	2	1	0	2
1	0	0	2	1
9	6	7	4	1

Filter

0	2	1
4	1	0
1	0	1

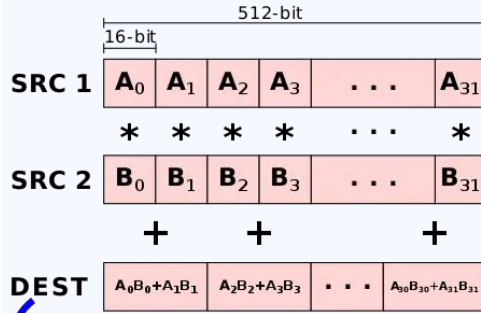
Output array

$$\begin{aligned} \text{Output}[0][0] &= (9*0) + (4*2) + (1*4) \\ &+ (1*1) + (1*0) + (1*1) + (2*0) + (1*1) \\ &= 0 + 8 + 1 + 4 + 1 + 0 + 1 + 0 + 1 \\ &= 16 \end{aligned}$$

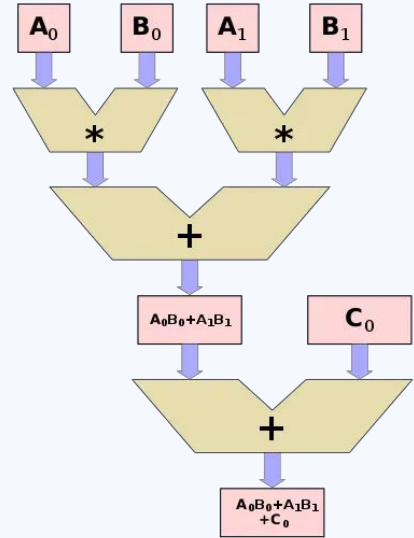
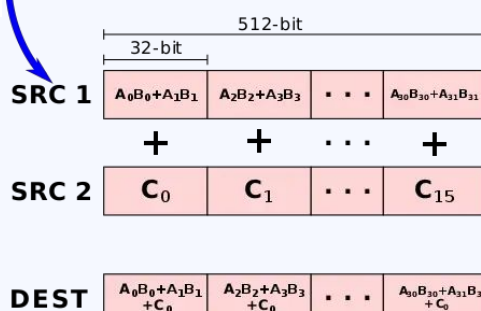
Image source

Image source

VPMADDWD



VPADDD



From the Intel optimization manual

P 5-11 (193): Converting to SIMD chart

P 8-9 (287): Blocking (handling large matrices)

P 14-2 (390): PCMPxSTRy (see also 14-12 onward)

P 15-7 (445): Mixing SSE and AVX (YMM register)

P 15-20 (458): Data alignment and caches

P 15-24 (462): Masked loads and paging



Are you more or less likely to prioritize buying a computer with advanced SIMD support now?