# Scalable MicroSupercomputers

**Don Alpert**

**February, 2003**



**Camelback Computer Architecture, LLC**
**Consulting Services**

**Donald Alpert, Ph.D.**

**www.camelback-comparch.com**

For additional information refer to the article in Microprocessor Report 3/17/2003
http://www.mdronline.com/mpr/h/2003/0317/171101.html

# Outline

- **Top Supercomputers**
- **Workloads**
  - — Size, Locality, and Balance
  - — Dense and Sparse Linear Algebra
- **System Architecture**
  - — Types T and C
- **Super Microprocessors**
  - — NEC SX-6
  - — Cray X1
- **Future Trends**

Camelback Computer Architecture, LLC
Consulting Services
Donald Alpert, Ph.D.
www.camelback-comparch.com

# Earth Simulator

TIME

**2002 Best Inventions**

- **5,120 CPUs**
  - 640 8-way nodes
  - CPU based on same core as NEC SX-6
- **8 GFLOPS per CPU**
  - 41 TFLOPS system peak
  - 35.8 TFLOPS sustained
- **Memory**
  - 16 GB per node
  - 10 TB total
- **Node Interconnect**
  - 640 × 640 crossbar
  - 25 GB/s inter-node bandwidth
- **20 kVA power per node**

Camelback Computer Architecture, LLC
Consulting Services
Donald Alpert, Ph.D.
www.camelback-comparch.com

# Top 10 Performing Supercomputers

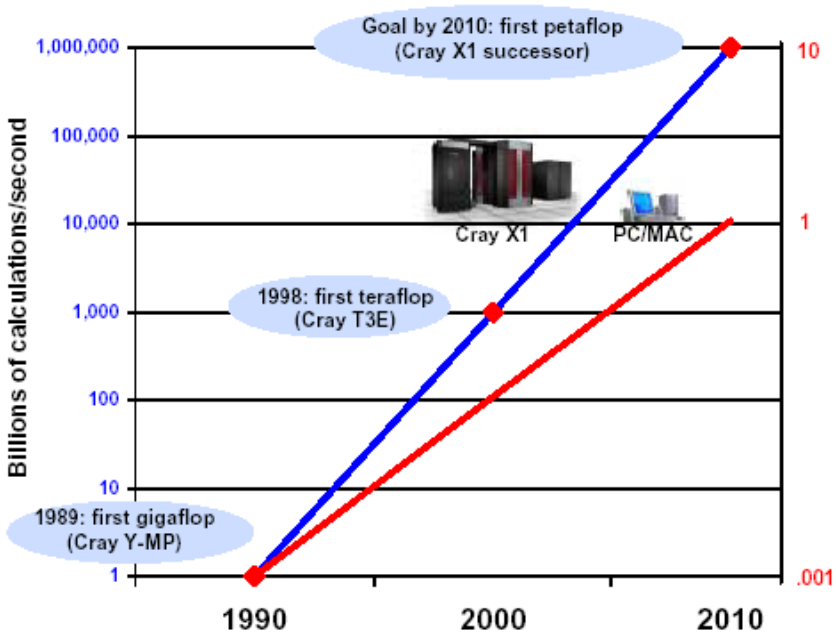
| Rank | Manufacturer<br>Computer/Procs | GFLOPS |
|---|---|---|
| 1 | **NEC**<br>Earth-Simulator/ 5120 | **35860.00** |
| 2 | **Hewlett-Packard**<br>ASCI Q - AlphaServer SC ES45/1.25 GHz/ 4096 | **7727.00** |
| 3 | **Hewlett-Packard**<br>ASCI Q - AlphaServer SC ES45/1.25 GHz/ 4096 | **7727.00** |
| 4 | **IBM**<br>ASCI White, SP Power3 375 MHz/ 8192 | **7226.00** |
| 5 | **Linux NetworX**<br>MCR Linux Cluster Xeon 2.4 GHz - Quadrics/ 2304 | **5694.00** |
| 6 | **Hewlett-Packard**<br>AlphaServer SC ES45/1 GHz/ 3016 | **4463.00** |
| 7 | **Hewlett-Packard**<br>AlphaServer SC ES45/1 GHz/ 2560 | **3980.00** |
| 8 | **HPTi**<br>Aspen Systems, Dual Xeon 2.2 GHz - Myrinet2000/ 1536 | **3337.00** |
| 9 | **IBM**<br>pSeries 690 Turbo 1.3GHz/ 1280 | **3241.00** |
| 10 | **IBM**<br>pSeries 690 Turbo 1.3GHz/ 1216 | **3164.00** |

- **ES beats 2nd place by >4X**
  - Only 25% more CPUs
  - At less than 50% frequency
- **First time in 10 year history that the top supercomputer is based on a custom microprocessor**

Source: Top500.org for the LINPACK benchmark H2/02

# Cray X1 Targets PetaFLOP by 2010

## 1990-2010: Supercomputer speed increases will outpace Moore's Law progress by 100x




| Year | Supercomputer speed: microprocessor speed |
|---|---|
| 2010 | 100,000:1 |
| 2000 | 10,000:1 |
| 1990 | 1,000:1 |

*Supercomputer speeds shown are actual, sustained performance on full 64-bit applications. Microprocessor speeds represented are theoretical maximums (MIPS ratings) and are higher than actual speeds. Gigaflop: 1 billion ($10^9$) calculations/second. Teraflop: 1 trillion ($10^{12}$) calculations/second. Petaflop: 1,000 trillion ($10^{15}$) calculations/second.

SLIDE 3
1/23/2003

**Cray X1 Supercomputer and Petaflop Speed**
Cray Inc.

Source: Cray, Inc.


Camelback Computer
Consulting S
Donald Alpert, Ph.D.
www.camelback-comparch.com

# Questions

- What workloads characteristics distinguish supercomputers from more general-purpose servers?
- What architecture characteristics distinguish between microprocessors customized for supercomputers and for servers?
- Will the distinguishing features of supercomputer microprocessors be adopted by future conventional microprocessors?

**Supercomputers Built From Custom Microprocessors Have Overtaken Those Built From Server Microprocessors**

Camelback Computer Architecture, LLC
Consulting Services
Donald Alpert, Ph.D.
www.camelback-comparch.com

# Supercomputer Workload

- **Size**
  - — By definition supercomputers execute the largest, most computationally demanding applications
  - — ES has 10 TB memory, top TPC-C system has 300GB
- **Locality**
  - — *Temporal Locality* for reuse ⇒ caches
  - — *Spatial Locality* for unit-stride ⇒ streaming buffers
  - — Regular pattern with constant stride ⇒ Prefetching
  - — Global Access: Irregular pattern with no locality
- **Balance**
  - — Partitions equally across parallel processors
  - — Minimal data sharing and synchronization

Camelback Computer Architecture, LLC
Consulting Services
Donald Alpert, Ph.D.
www.camelback-comparch.com

# Dense Linear Algebra

```
DO I = 1,N
    X[I] = a*X[I] + Y[I]
```

- **LINPACK Benchmark**
  - — Gaussian Elimination
  - — Inner Loop of DAXPY
- **Locality**
  - — X unit stride, read before write
  - — Y unit stride, reuse
- **Balance**
  - — Partition submatrices with equal rows and columns

# Sparse Linear Algebra

- **Sparse matrix has large fraction of 0 elements**
- **Stored in Compressed Sparse Row (CSR) format**
  - Other formats are used depending on application

**Matrix**

| row/col | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 2 | 0 | 0 |
| 2 | 0 | 0 | 3 | 0 | 0 | 4 |
| 3 | 5 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 6 | 0 | 0 | 0 | 0 |
| 5 | 7 | 0 | 0 | 8 | 0 | 0 |
| 6 | 0 | 0 | 9 | 0 | 10 | 0 |

Compressed Form `A`:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

`RowPtr`

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 3 | 5 | 6 | 7 | 9 | |

`ColIndex`

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 3 | 6 | 1 | 2 | 1 | 4 | 3 | 5 |

# Matrix-Vector (MV) Multiplication

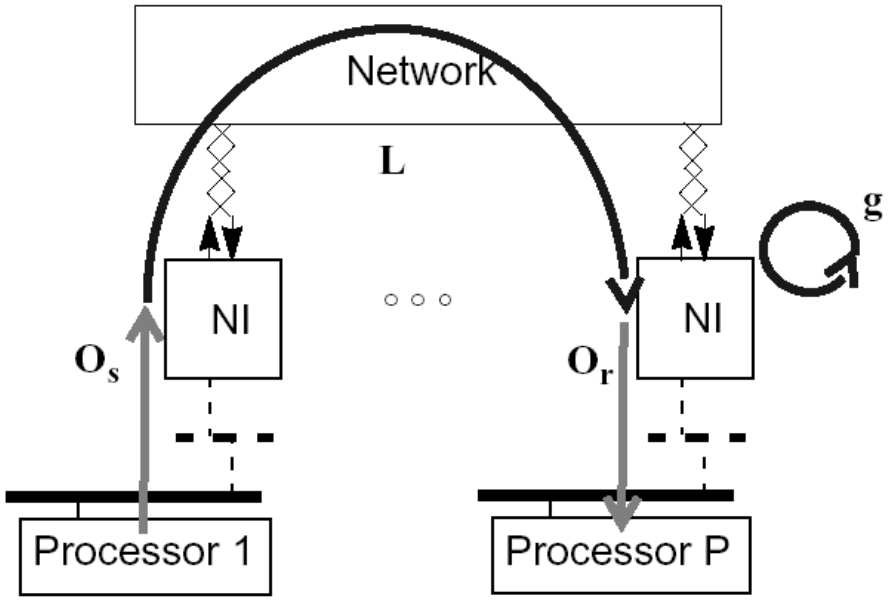
```
DO I = 1,N
   SUM = 0.0
   DO J = RowPtr[I], RowPtr[I+1]-1
     SUM = SUM + A[J]*X[ColIndex[j]]
   END DO
   Y[I] = SUM
END DO
```

- **Locality**
  - `A, RowPtr, ColIndex` Unit Stride, Read Only
  - `Y` Unit Stride, Write Only
  - `X` Indirect, Global
- **Balance**
  - Difficult to partition equally – depends on matrix structure

# System Architecture

- **Burton Smith (Tera and Cray) classifies supercomputer system architectures as types T and C**
- **Type T (Transistor Intensive)**
  - Clusters and grids of conventional processors
  - Cost determined largely by processors and memory
  - Relatively low interconnection bandwidth
  - Perform well for applications with locality and balance
- **Type C (Connection Intensive)**
  - Specialized, low-latency, high-bandwidth interconnect
  - Custom processors
  - Cost determined by interconnection fabric, as well as processors and memory
  - Perform well for broad range of applications, including many with with poor locality and balance

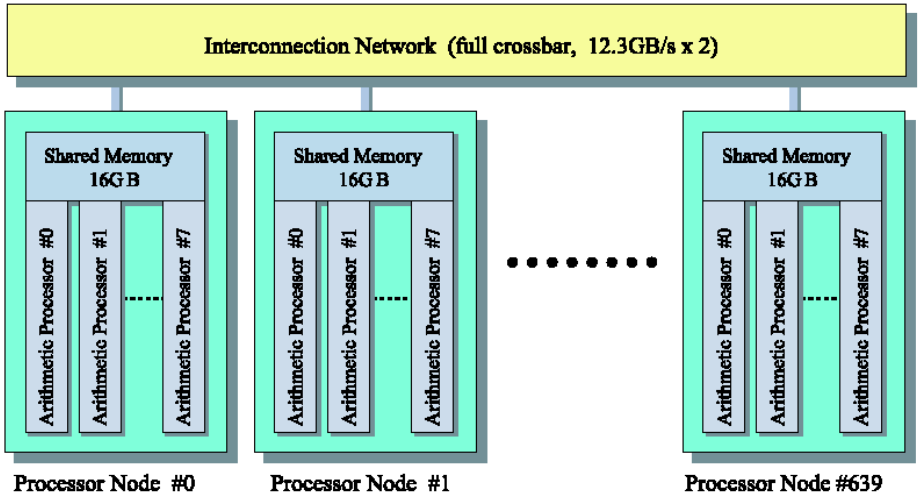
# Network Performance




- **LogP Model (Culler)**
  - *L* = Latency
  - *o* = overhead
  - *g* = gap, mini time between messages
  - *P* = No. processors
- **Time to transfer n small messages =**
  - $o_s + L + (n-1)*g + o_r$
- **Type C systems commonly provide hardware support for creating and receiving messages**

| | ES | IBM SP-3 |
|---|---|---|
| Topology | Crossbar | Omega Multistage |
| MPI Bandwidth | 11.8 GB/s | 350 MB/s |
| MPI Latency | 5.6 μs | 17 μs |

# Earth Simulator System



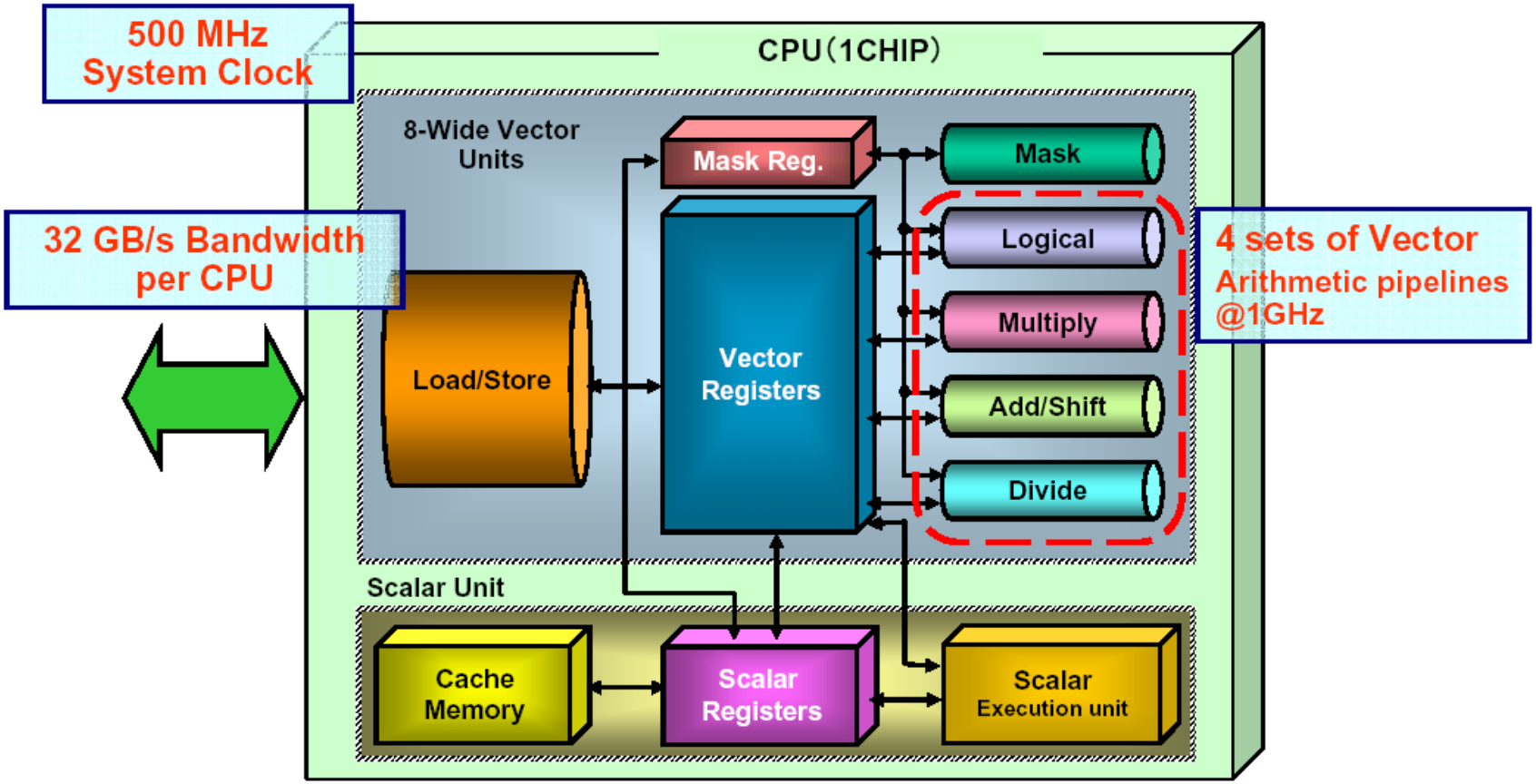

- **Arithmetic Processor**
  - Based on same core as NEC SX-6
  - Superscalar Unit
    - 4-issue, OOO execution
  - Caches
    - Separate I and D
    - Each 64KB, 2-way associative
- **Node**
  - 8 CPUs, 32 Memory Modules
  - 16GB local memory
  - Crossbar interconnect
  - 32GB/s to local mem per CPU
- **Network**
  - 640 x 640 crossbar
  - 12.3 GB/s bidirectional bandwidth
  - Specialized HW support to transfer 3D subarrays and indirect access, and barrier synchronization

Camelback Computer Architecture, LLC
Consulting Services
Donald Alpert, Ph.D.
www.camelback-comparch.com

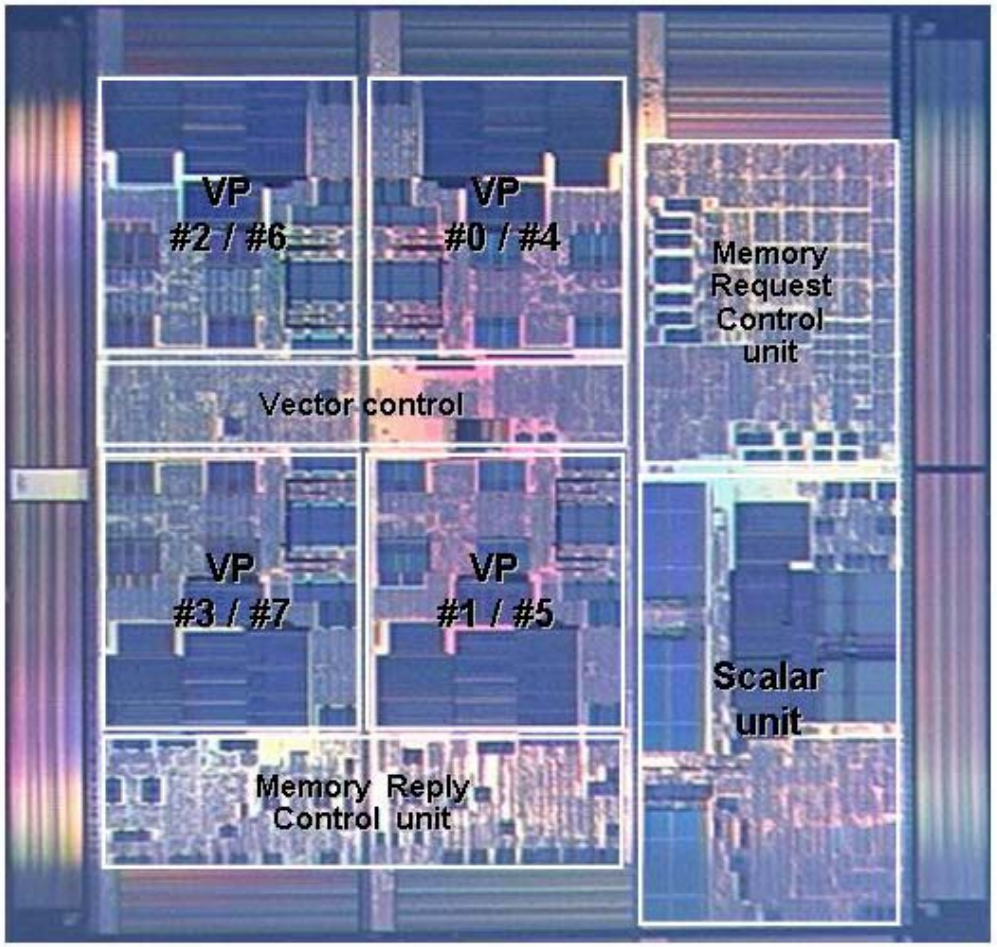


Source: NEC

# SX-6 Block Diagram



# SX-6 Die Photo

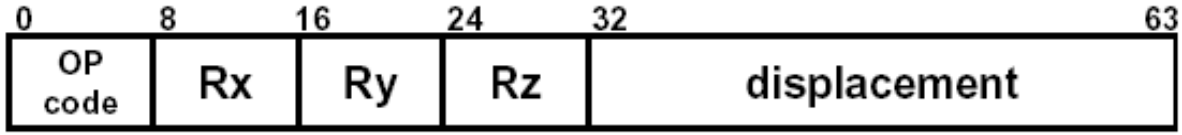



- 0.15 μm Cu 8LM CMOS
- 60M Tx
- 432 $mm^2$
- 500 MHz scalar and mem
  1 GHz vector pipe
- SX-5 was 32 chips

Source: NEC

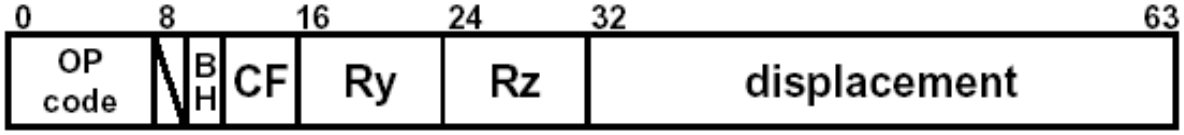
# SX-6 Instruction Set

**RX type: (Scalar)**



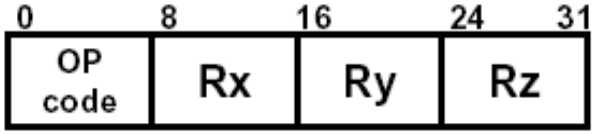

The RX type instructions are used for scalar instructions, in particular, memory reference instructions. Rx ← Mem(Ry+Rz+disp)

**CFX type: (Scalar)**



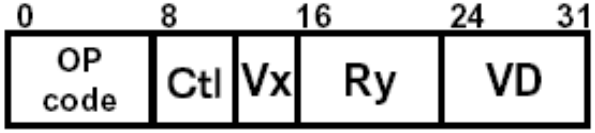

The CFX type instructions are used for branch instructions.

**RR type: (Vec/Sca)**




The RR type instructions include scalar and vector instructions. Rx ← Ry op. Rz

**RV type: (Vector)**

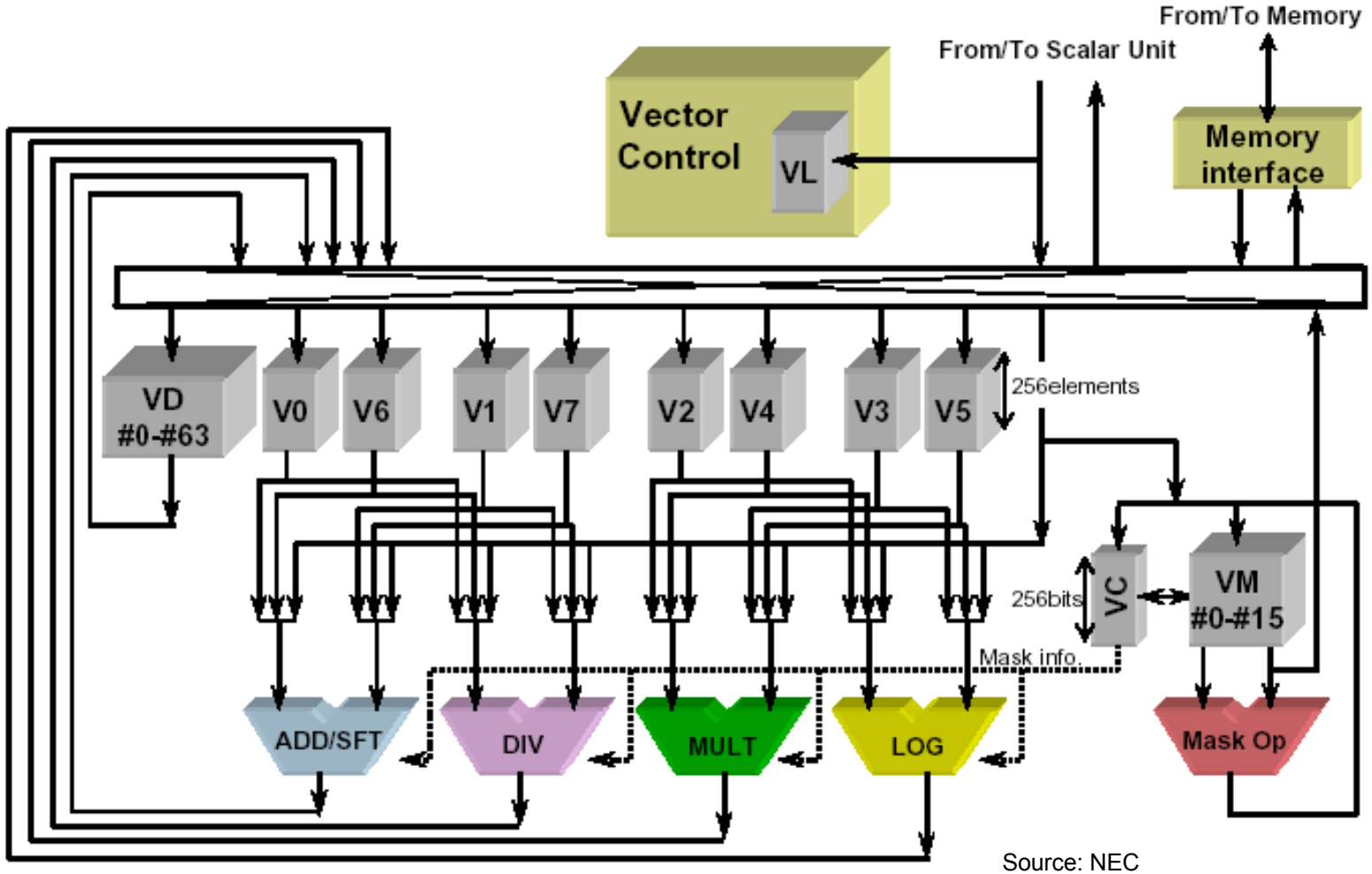



The RV type instructions are used only for vector instructions such as arithmetic operations, memory access.

Rx: scalar register
Ry,Rz: scalar register or immediate
BH: Branch Hint flag
CF: Branch condition with Ry
Ctl: Vector control (mask, source V.rg, etc)
Vx: Vector destination register

Source: NEC

# SX-6 Vector Unit (1)

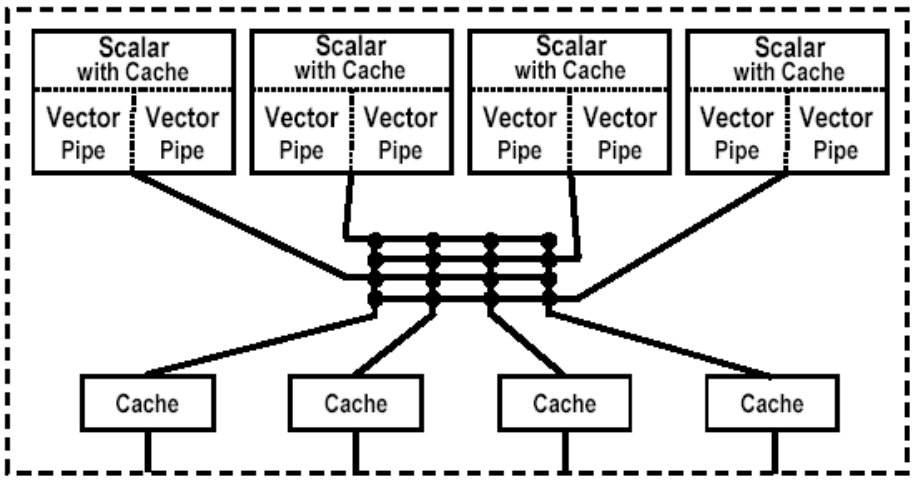



Source: NEC

# SX-6 Vector Unit (2)

- **8 GFLOPS peak**
  - 2 results per clock across 4 parallel units (lanes) at 1 GHz
  - Add, Multiply, Divide, Logical with masking
  - Specialized operations
    - Max/min, sum, iterate, popcount, leading-0, trailing-1
- **8 Vector Arithmetic Registers**
  - Each holds 256 8B words
  - Each feeds 2 pipes
- **64 Vector Data Registers**
  - Each holds 256 8B words
  - Stage data, do not feed pipelines
- **Sparse Vector Addressing**
  - Gather, Scatter, Expand, Compress

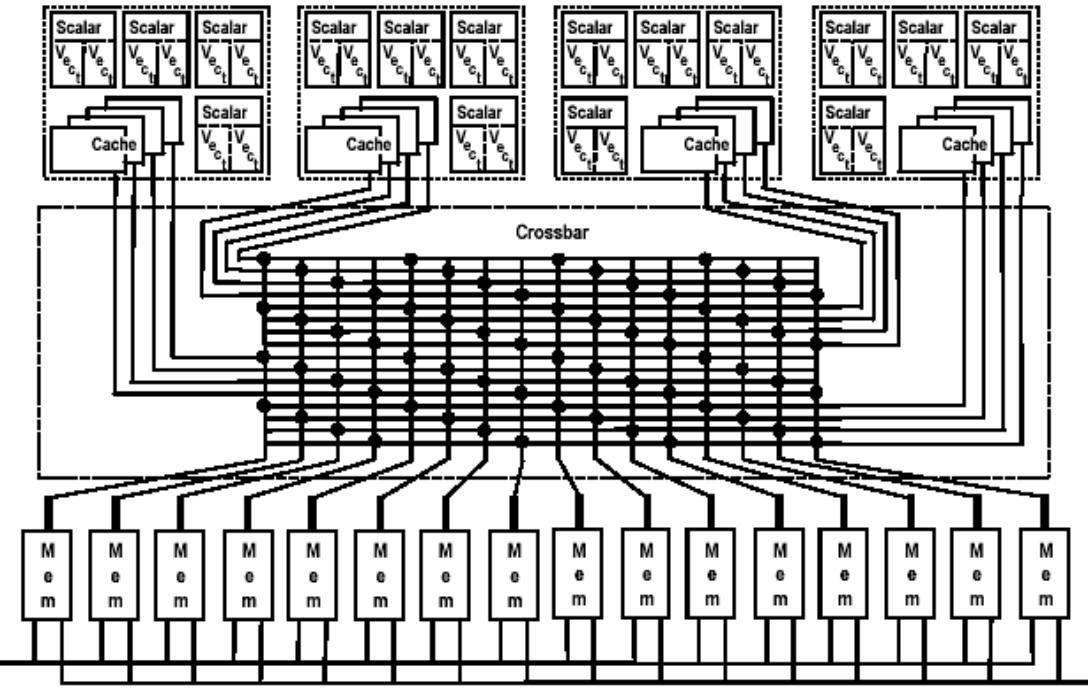
# Cray X1 Block Diagram




Source: D. H. Brown

- **ISA based on MIPS with vector extensions**
- **CPU**
  - Scalar unit with cache
  - 2 vector pipes
  - 800 MHz
- **Multi-Streaming Processor**
  - 4 CPUs
  - 4 0.5 MB cache modules
  - HW cache coherency
  - Shared registers for sharing data and synchronizing to compute a common loop
  - Message passing HW

# Cray X1 Node



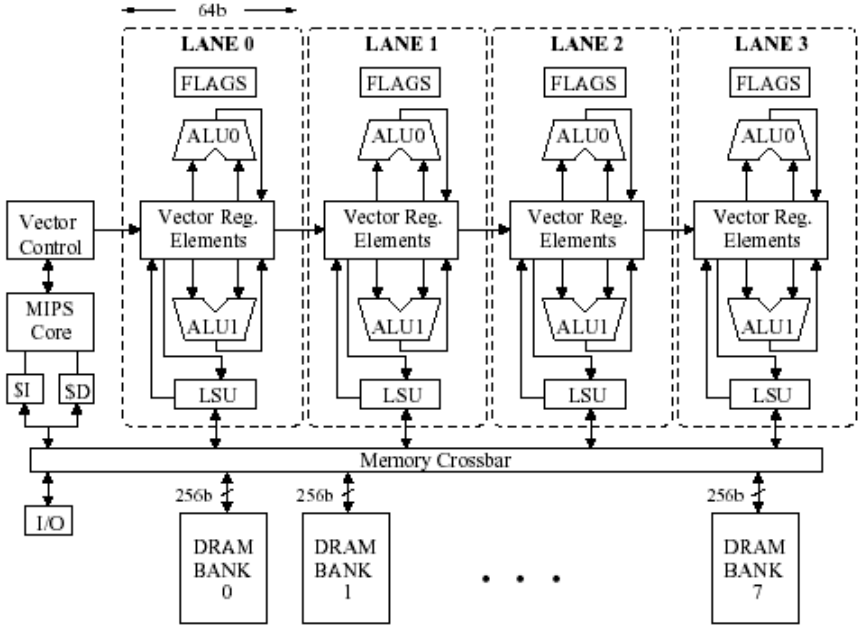

Source: D. H. Brown

- **4 MSPs and 16 Memory modules**
  - — Crossbar connect
  - — 38.4 GB/s per MSP
- **Network**
  - — Modified 3D Torus
  - — 102 GB/s per node
  - — Typical memory latency is 1 μs for 128 nodes

# Supercomputer vs. Server Microprocessors

- **Very high bandwidth to main memory**
  - — 4-12 bytes sustained per peak FLOP for supercomputers
  - — 1.5-2.5 bytes peak per peak FLOP for ITP-2 and Power4
- **Large vector register sets to cover the long, variable latency to global memory**
- **Vector scatter, gather, and mask functions to sustain computation throughput near peak rates**
- **Specialized hardware for low-latency data sharing, synchronization, and message passing**

# Academic Research




Kozyrakis and Patterson, MICRO-35

- **VIRAM**
  - Dave Patterson at UC Berkeley
  - MIPS with vector extensions
  - 4 vector units
  - embedded DRAM
  - 200 MHz
  - Evaluated for EEMBC kernels
    - 2x performance of 1 GHz superscalar
    - 10x performance of 8-wide VLIW
- **Stanford Streaming Supercomputer**
  - Bill Dally et al
  - Large (32KB) multi-ported (32) register file similar to vector regs
  - Memory accesses with streaming strides to register file
  - Concepts proven for signal processing with Imagine

# Future Trends

- **Expect embedded processors to add vector units**
  - — Proven performance for signal processing
  - — Effective with large bandwidth of embedded memory
- **Desktop and Server Processors**
  - — SIMD with SSE capture much of vector capability
  - — But not all structures and algorithms use unit stride
  - — May see a streaming register file
    - Efficient VLSI implementation and more versatile than vector registers
- **HW mechanisms for sharing data and synchronizing**
  - — Lower latency than semaphores in shared memory
  - — Can be used for MT and MC chips, as well as scalable
  - — Power5 has HW comm link for barrier synchronization
    - May be used for Virtual Vector Architecture (ViVA) for Blue Planet

# Summary

- **Fastest supercomputers**
  - Custom vector microprocessors
  - High memory bandwidth
  - Low-latency, high-bandwidth interconnection networks
- **Vector capability may migrate to conventional processors**
  - Embedded processors with high-bandwidth e-DRAM
  - Server processors with high-bandwidth optical interconnect
- **HW mechanisms for sharing data and synchronizing**
  - Lower-latency than semaphores in shared memory