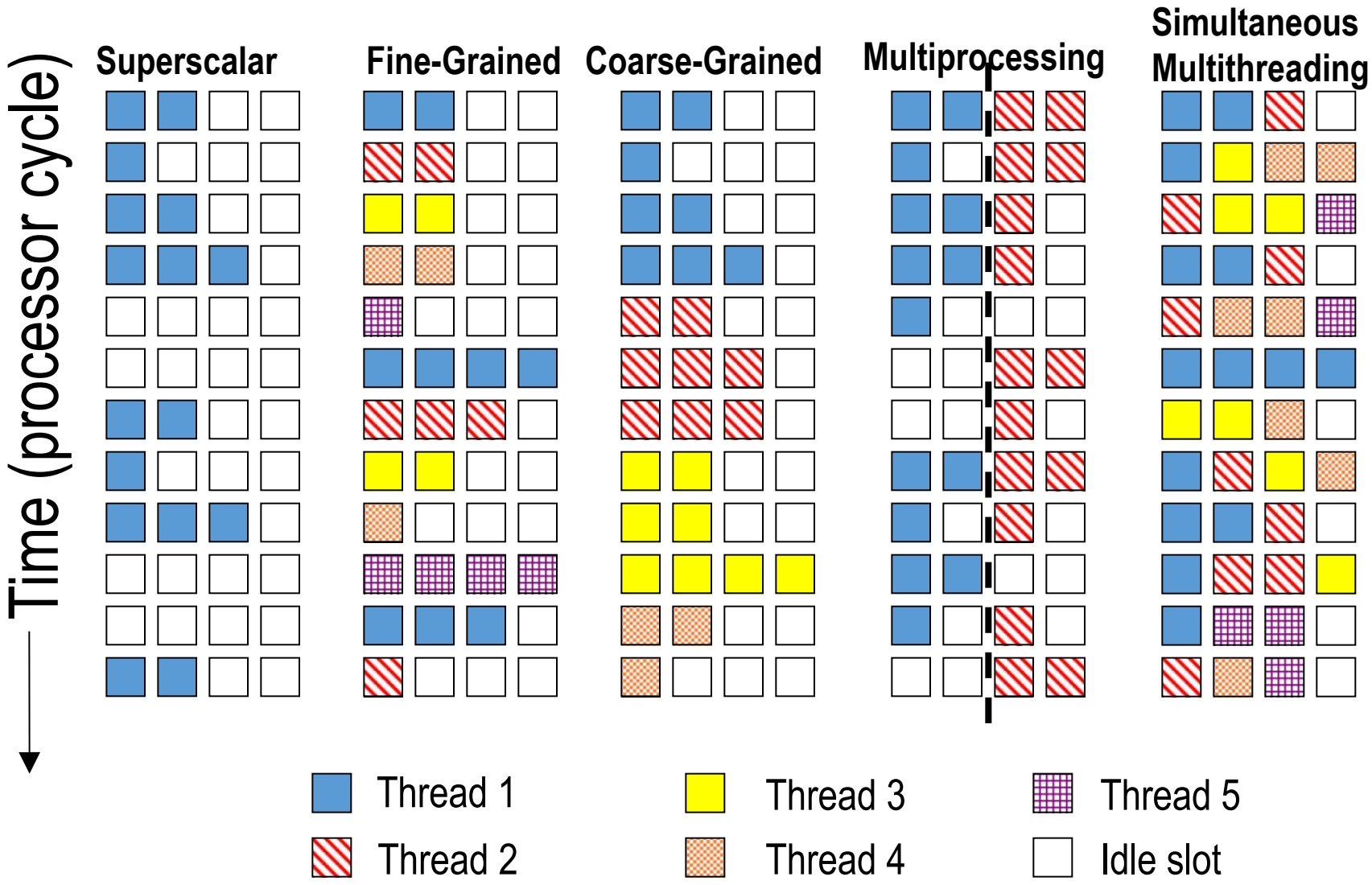
计算机体系结构

周学海
xhzhou@ustc.edu.cn
0551-63492149
中国科学技术大学

Review : Multithreaded Categories



第6章 Data-Level Parallelism in Vector, SIMD, and GPU Architectures

6.1 向量处理机模型 （Chapter 4）

- 数据级并行的研究动机
- 数据级并行的种类
- 向量体系结构
  - 向量处理模型
  - 基本特性及结构
  - 性能评估及优化

6.2-1 向量处理机模型优化

6.2-2 面向多媒体应用的SIMD指令集扩展

6.3-1 GPU-I

6.3-2 GPU-II

- GPU简介
- GPU的编程模型
- GPU的存储系统

6.1 向量处理机模型

数据级并行 → 向量处理机模型 → 性能评估

动机：传统指令级并行技术的问题

- **提高性能的传统方法（挖掘ILP）的主要缺陷:**
  - 程序内在的并行性
  - 提高流水线的时钟频率: 提高时钟频率，有时导致CPI随着增加 (branches, other hazards)
  - 指令预取和译码: 有时在每个时钟周期很难预取和译码多条指令
  - 提高Cache命中率 : 在有些计算量较大的应用中（科学计算）需要大量的数据，其局部性较差，有些程序处理的是连续的媒体流(multimedia),其局部性也较差。

动机：DLP的兴起

- 应用需求和技术发展推动着体系结构的发展
- 图形、机器视觉、语音识别、机器学习等新的应用均需要大量的数值计算，其**算法通常具有数据并行特征**
- SIMD-based 结构 (vector-SIMD, subword-SIMD, SIMT/GPUs) 是执行这些算法的最有效途径

动机：SIMD结构的优势

- **SIMD 结构可有效地挖掘数据级并行:**
  - 基于矩阵运算的科学计算
  - 图像和声音处理
  - .....
- **SIMD比MIMD更节能**
  - 针对每组数据操作仅需要取指一次
  - SIMD对PMD(personal mobile devices)更具吸引力
- **SIMD 允许程序员继续以串行模式思维**

SIMD 结构的种类

- **向量体系结构**
- **多媒体SIMD指令集扩展**
- **Graphics Processor Units (GPUs)**

- **For x86 processors:**
  - 每2年增加2cores/chip
  - SIMD 宽度每4年翻一番
  - SIMD潜在加速比是MIMD的2倍

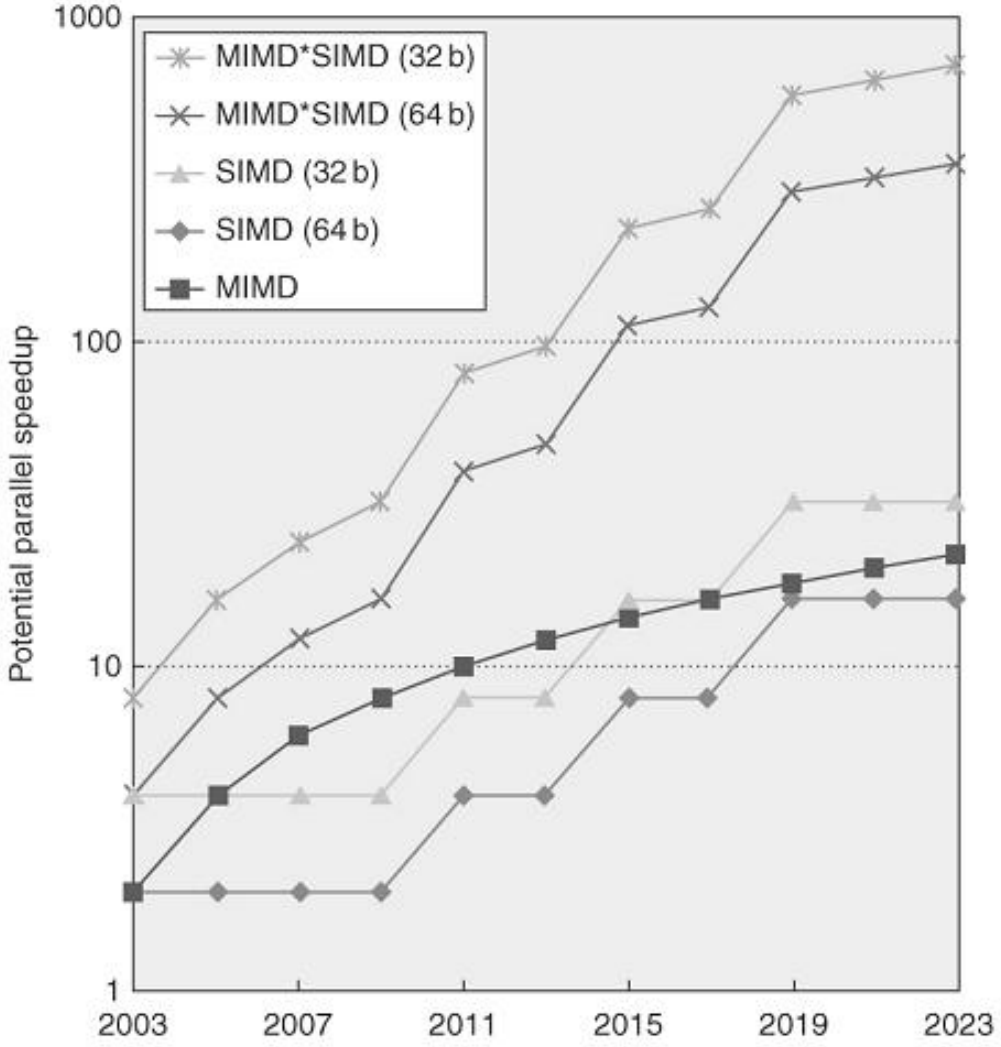


Figure 4.1 Potential speedup via parallelism from MIMD, SIMD, and both MIMD and SIMD over time for x86 computers. This figure assumes that two cores per chip for MIMD will be added every two years and the number of operations for SIMD will double every four years.

起源：Supercomputers

- **Supercomputer的定义:**
  - 对于给定任务而言世界上最快的机器
  - 任何造价超过3千万美元的机器
  - 计算能力达到每秒万亿次的机器
- **由Seymour Cray设计的机器**
- **CDC6600 (Cray, 1964) 被认为是第一台超级计算机**
- **CDC Cray-1 (Cray, 1976): 被认为是第一台成功商用的向量机 10X 6600**

CDC 6600 *Seymour Cray, 1963*

- **A fast pipelined machine with 60-bit words**
  - 128 Kword main memory capacity, 32 banks
- **Ten functional units (parallel, unpipelined)**
  - Floating Point: adder, 2 multipliers, divider
  - Integer: adder, 2 incrementers, …
- **Hardwired control (no microcoding)**
- ***Scoreboard* for dynamic scheduling of instructions**
- **Ten Peripheral Processors for Input/Output**
  - a fast multi-threaded 12-bit integer ALU
- **Very fast clock, 10 MHz (FP add in 4 clocks)**
- **>400,000 transistors, 750 sq. ft., 5 tons, 150 kW, novel freon-based technology for cooling**
- **Fastest machine in world for 5 years (until 7600)**
  - over 100 sold ($7-10M each)

IBM Memo on CDC6600

Thomas Watson Jr., IBM CEO, August 1963:

"Last week, Control Data … announced the 6600 system. I understand that in the laboratory developing the system there are only 34 people including the janitor. Of these, 14 are engineers and 4 are programmers… Contrasting this modest effort with our vast development activities, I fail to understand why we have lost our industry leadership position by letting someone else offer the world's most powerful computer."

To which Cray replied: "It seems like Mr. Watson has answered his own question."

Supercomputer Applications

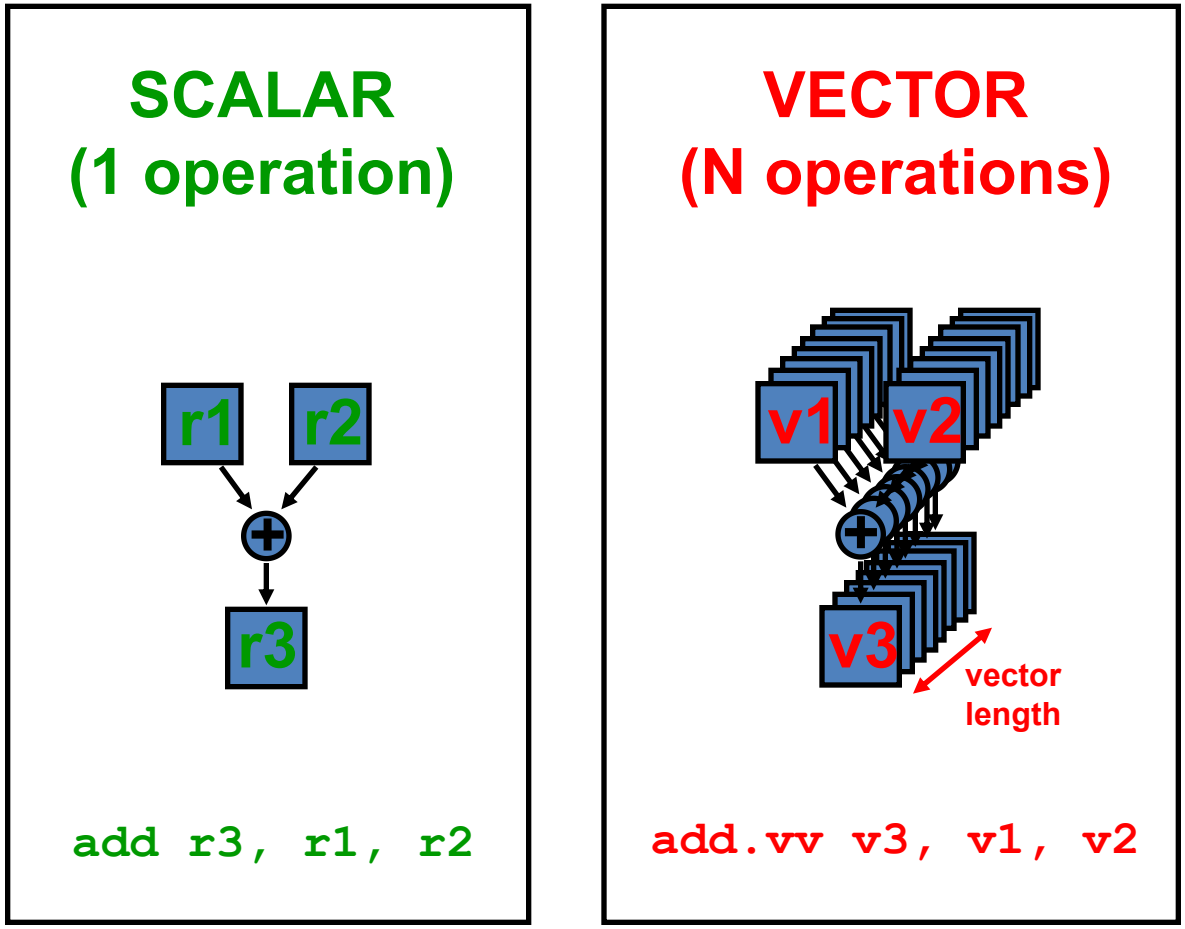
- **典型应用领域**
  - 军事研究领域（核武器研制、密码学）
  - 科学研究
  - 天气预报
  - 石油勘探
  - 工业设计 (car crash simulation)
  - 生物信息学
  - 密码学
- **均涉及大量的数据集处理**
- **70-80年代Supercomputer = Vector Machine**

6.1 向量处理机模型

数据级并行 → **向量处理机模型** → 性能评估

向量处理模型

- **向量处理机具有更高层次的操作，一条向量指令可以处理N个或N对操作数（处理对象是向量）**



向量处理机的基本特性

- **基本思想：两个向量的对应分量进行运算，产生一个结果向量。**
- **简单的一条向量指令包含了多个操作=> fewer instruction fetches**
- **每一结果独立于前面的结果**
  - 长流水线，编译器保证操作间没有相关性
  - 硬件仅需检测两条向量指令间的相关性
  - 较高的时钟频率
- **向量指令以已知的模式访问存储器**
  - 可有效发挥多体交叉存储器的优势
  - 可通过重叠减少存储器操作的延时 -（例如：一个向量包含 64个元素）
  - 不需要数据Cache! (仅使用指令cache)
- **在流水线控制中减少了控制相关**

向量处理机的基本结构

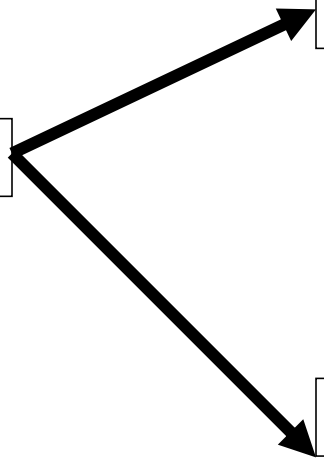
- ***memory-memory vector processors:*** **所有的向量操作是存储器到存储器**
- ***vector-register processors:*** **除了load 和store操作外，所有的操作是向量寄存器与向量寄存器间的操作**
  - 向量机的Load/Store结构
  - 1980年以后的所有的向量处理机都是这种结构: Cray, Convex, Fujitsu, Hitachi, NEC
  - 我们也主要针对这种结构

Vector Memory-Memory versus Vector Register Machines

- **存储器-存储器型向量机所有指令操作的操作数来源于存储器**
- **第一台向量机 CDC Star-100 (‘73) and TI ASC (‘71), 是存储器-存储器型机器**
- **Cray-1 (‘76) 是第一台寄存器型向量机**

Example Source Code

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



Vector Memory-Memory Code

```
ADDV C, A, B
SUBV D, A, B
```

Vector Register Code

```
LV V1, A
LV V2, B
ADDV V3, V1, V2
SV V3, C
SUBV V4, V1, V2
SV V4, D
```

Vector Memory-Memory vs. Vector Register Machines

- **存储器-存储器型向量机 (VMMA) 需要更高的存储器带宽**
  - All operands must be read in and out of memory
- **VMMA结构使得多个向量操作重叠执行较困难**
  - Must check dependencies on memory addresses
- **VMMA启动时间更长**
  - CDC Star-100 在向量元素小于100时，标量代码的性能高于向量化代码
- **CDC Cray-1后续的机器 (Cyber-205, ETA-10) 都是寄存器型向量机**

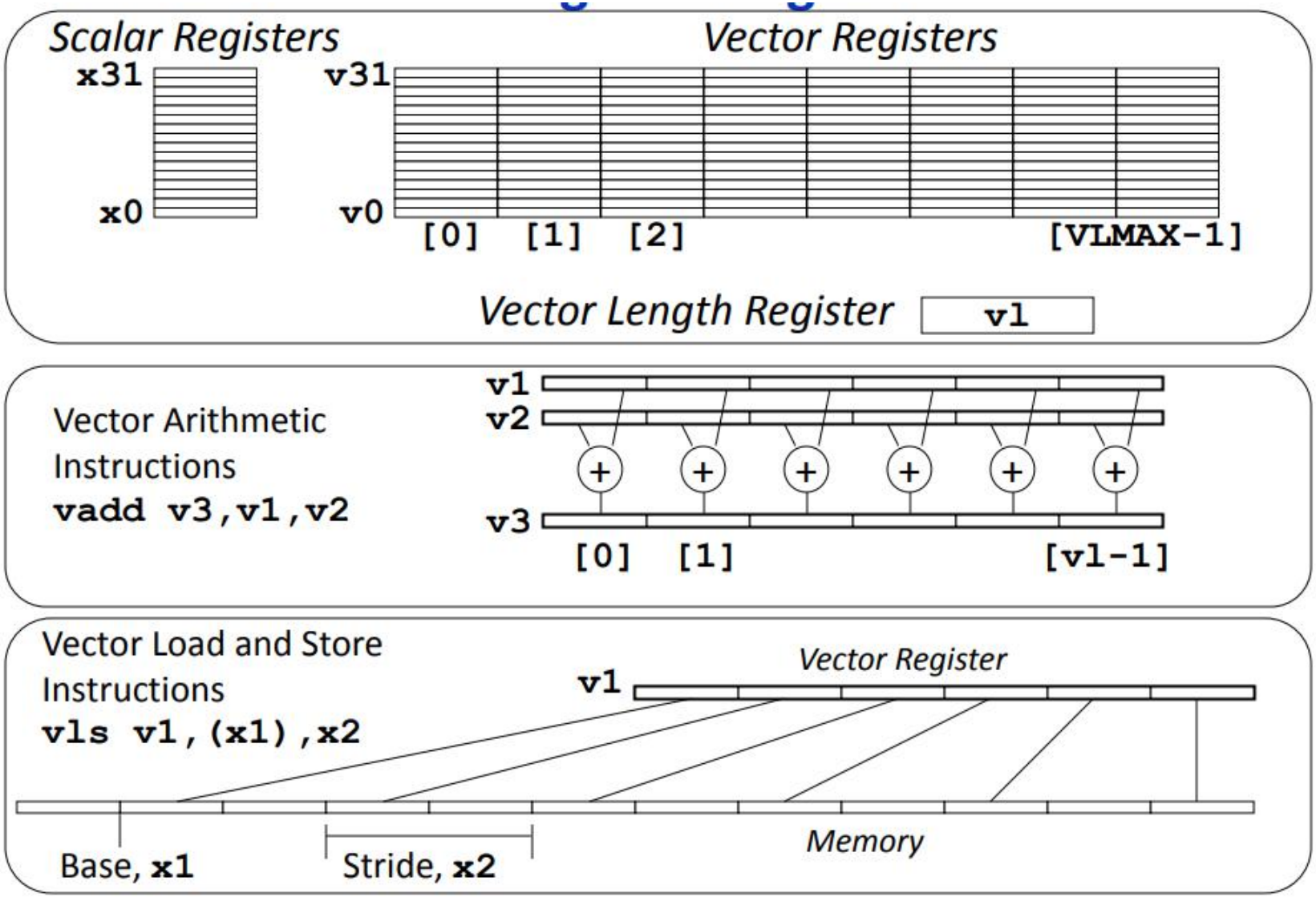
Vector Supercomputers

- **Cray-1（‘76）概览:**
- **Scalar Unit：Load/Store Architecture**
- **Vector Extension**
  - Vector Registers
  - Vector Instructions
- **Implementation**
  - 硬布线逻辑控制
  - 高效流水化的功能部件
  - 多体交叉存储系统
  - 无Data Cache
  - 不支持 Virtual Memory

Vector Instruction Set Advantages

- **格式紧凑**
  - 一条指令包含N个操作
- **表达能力强，一条指令能告诉硬件:**
  - N个操作之间无相关性
  - 使用同样的功能部件
  - 访问不相交的寄存器
  - 与前面的操作以相同模式访问寄存器
  - 访问存储器中的连续块 (unit-stride load/store)
  - 以已知的模式访问存储器 (strided load/store)
- **可扩展性好**
  - 可以在多个并行的流水线上运行同样的代码 (lanes)

Vector Programming Model

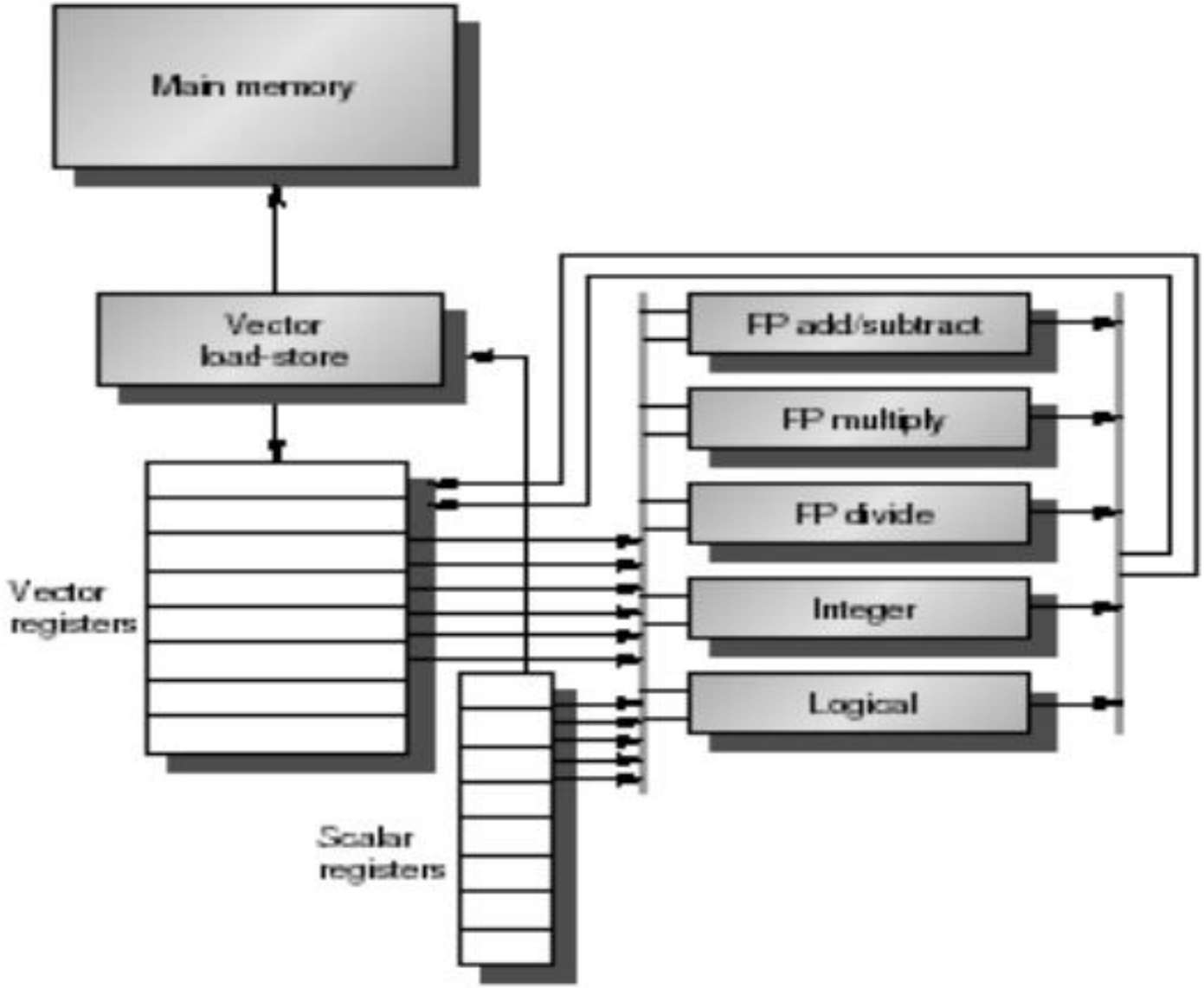


Vector Instructions (DLXV)

	Instr.	Operands	Operation Comment
ADDV	V1,V2,V3	V1=V2+V3	vector + vector
ADDSV	V1,F0,V2	V1=F0+V2	scalar + vector
MULTV	V1,V2,V3	V1=V2xV3	vector x vector
MULSV	V1,F0,V2	V1=F0xV2	scalar x vector
LV	V1,R1	V1=M[R1..R1+63]	load, stride=1
LVWS	V1,R1,R2	V1=M[R1..R1+63*R2]	load, stride=R2
LVI	V1,R1,V2	V1=M[R1+V2i,i=0..63]	indir.("gather")
CeqV	VM,V1,V2	VMASKi = (V1i=V2i)?	comp. setmask
MOV	VLR,R1	Vec. Len. Reg. = R1	set vector length
MOV	VM,R1	Vec. Mask = R1	set vector mask

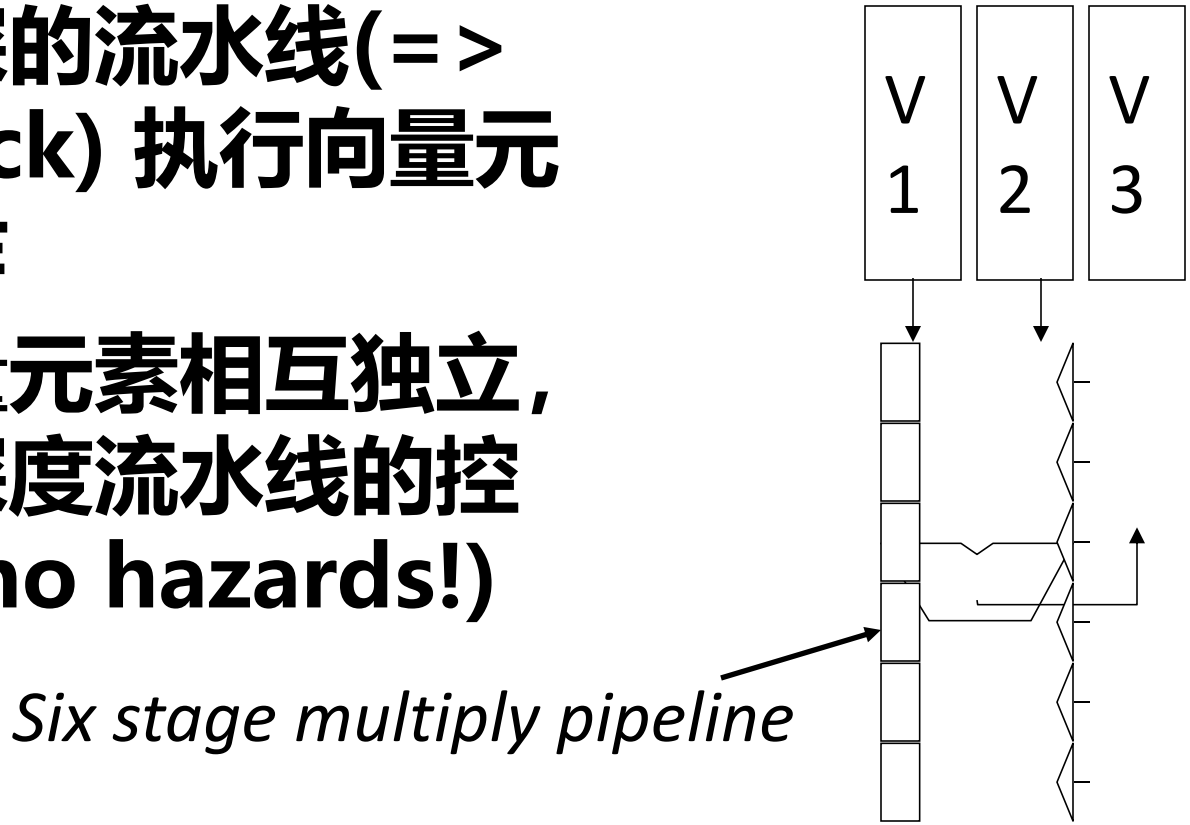
向量处理机的基本组成单元

- ***Vector Register*: 固定长度的一块区域，存放单个向量**
  - 至少2个读端口和一个写端口（一般最少16个读端口，8个写端口）
  - 典型的有8-32 向量寄存器，每个寄存器存放64到128个64位元素
- ***Vector Functional Units (FUs)*: 全流水化的，每一个clock启动一个新的操作**
  - 一般4到8个FUs: FP add, FP mult, FP reciprocal (1/X), integer add, logical，shift；可能有些重复设置的部件
- ***Vector Load-Store Units (LSUs)*: 全流水化地load 或store一个向量，可能会配置多个LSU部件**
- ***Scalar registers*: 存放单个元素用于标量处理或存储地址**
- **用交叉开关连接(Cross-bar) FUs , LSUs, registers**

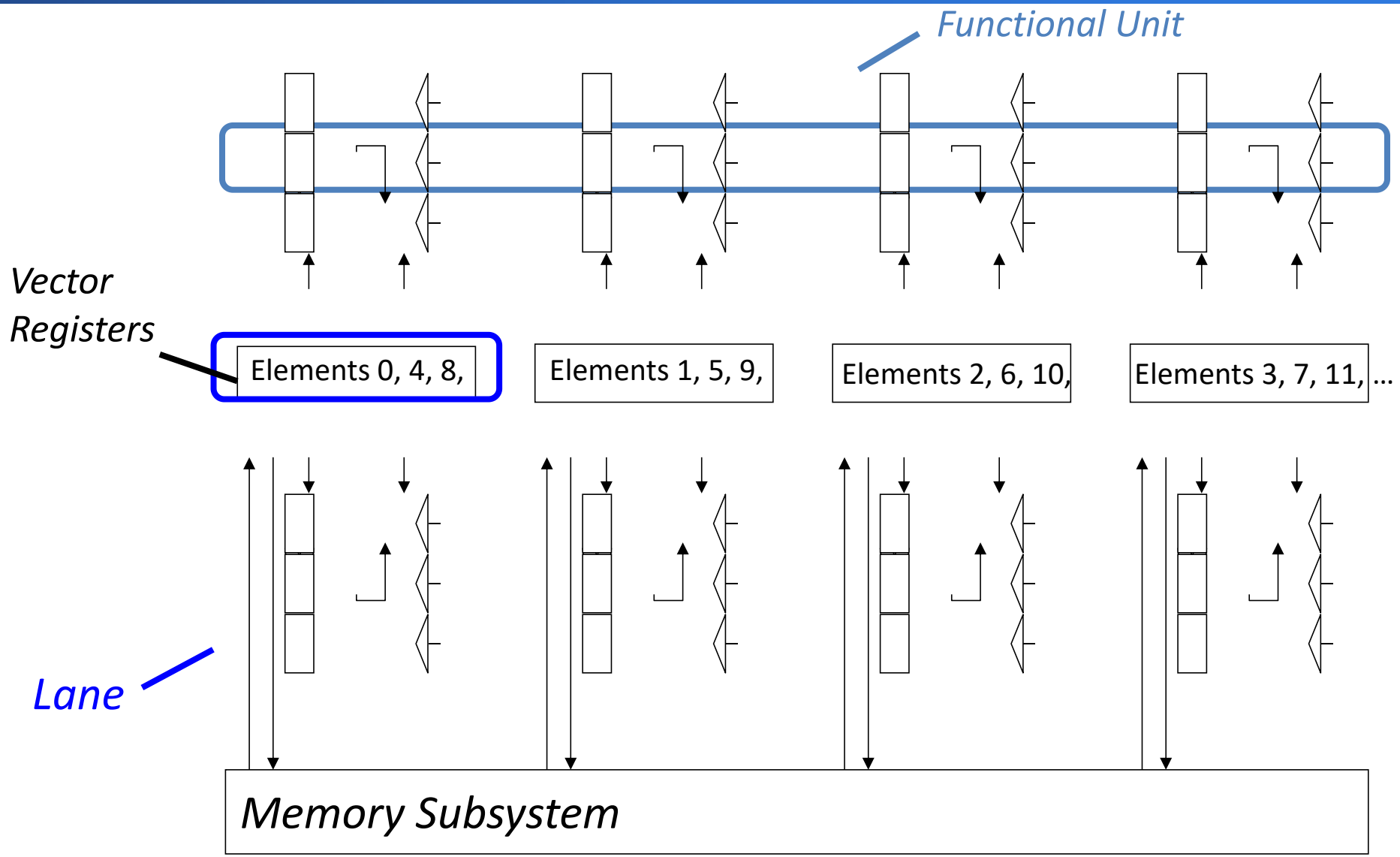

Main memory
Vector load-store
FP add/subtract
FP multiply
FP divide
Integer
Logical
Vector registers
Scalar registers

Vector Arithmetic Execution

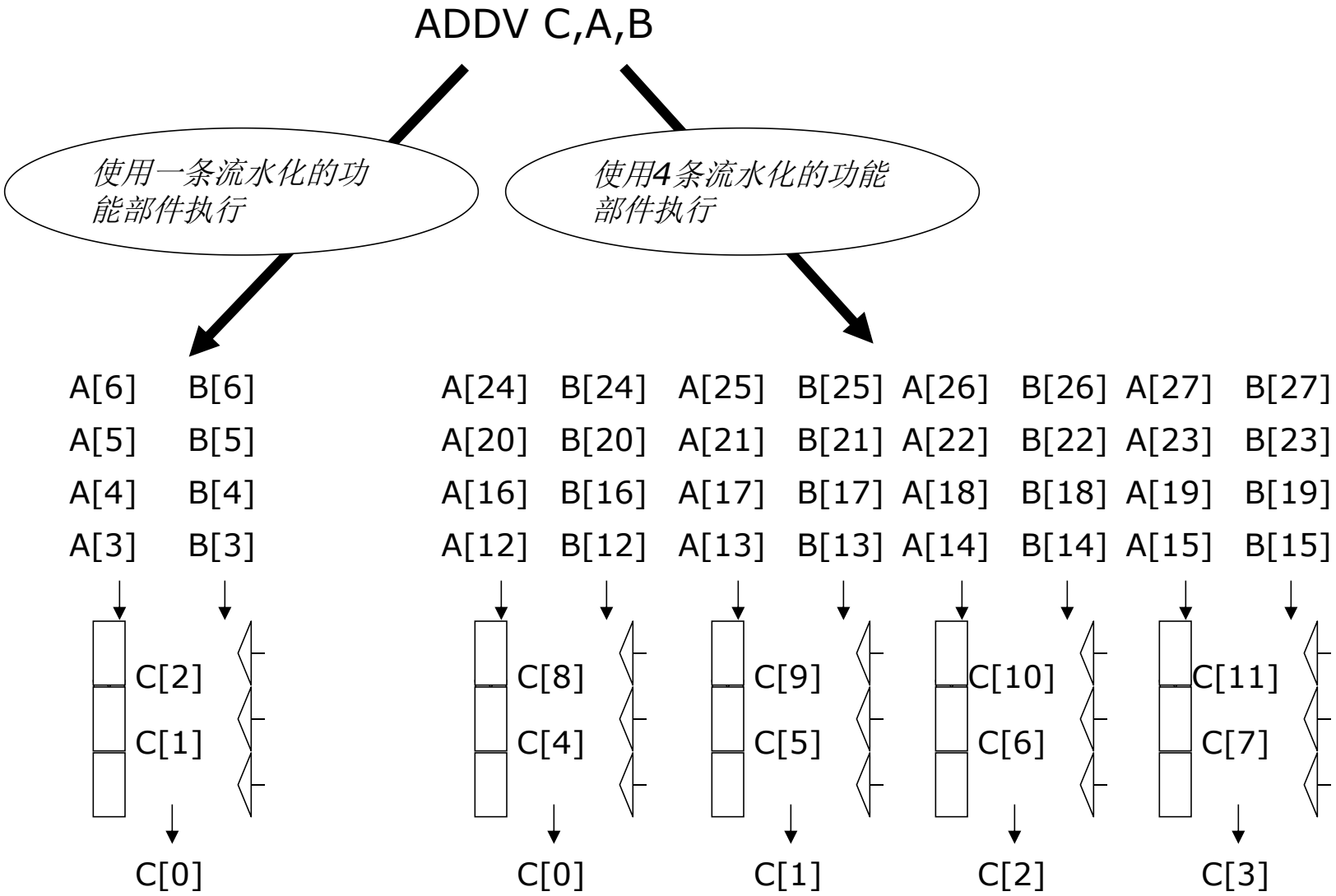
- **使用较深的流水线(=> fast clock) 执行向量元素的操作**
- **由于向量元素相互独立，简化了深度流水线的控制 (=> no hazards!)**



Vector Unit Structure



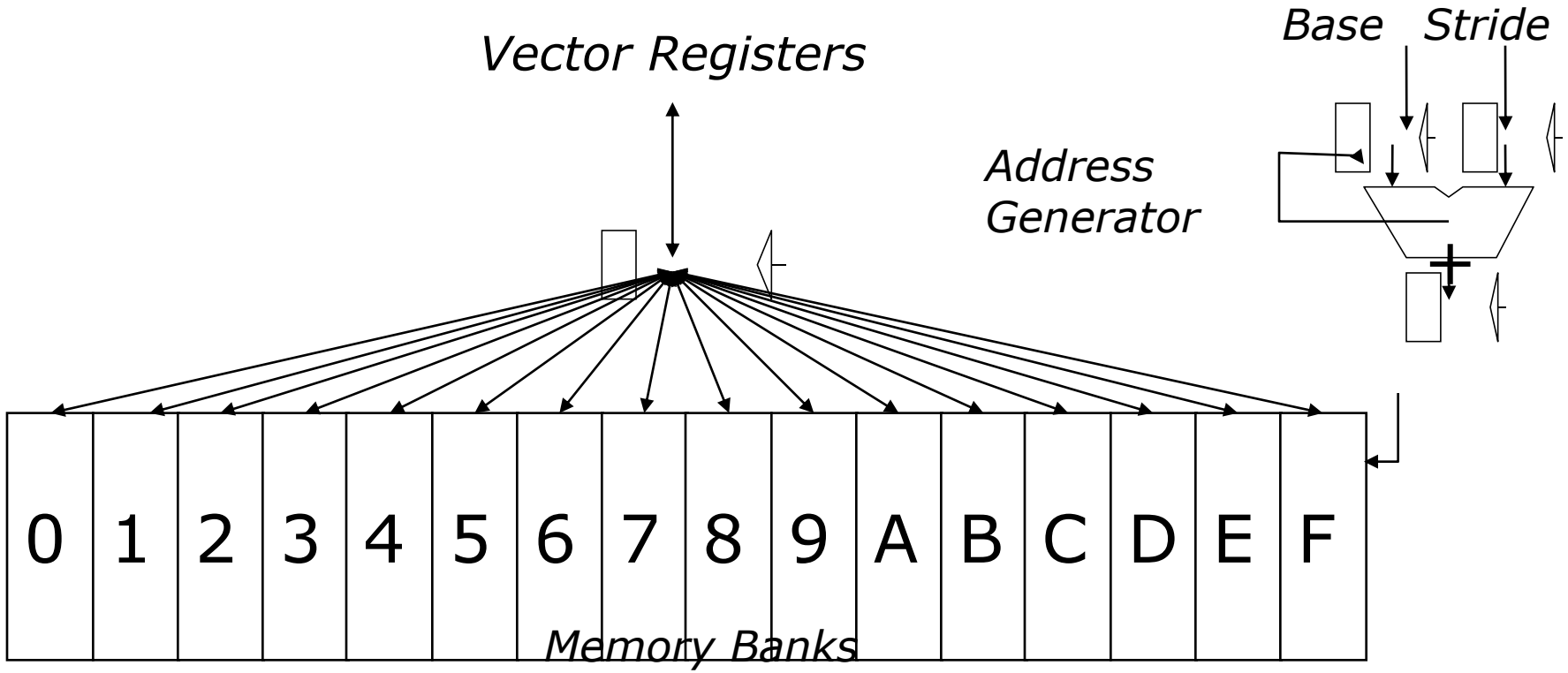
Vector Instruction Execution



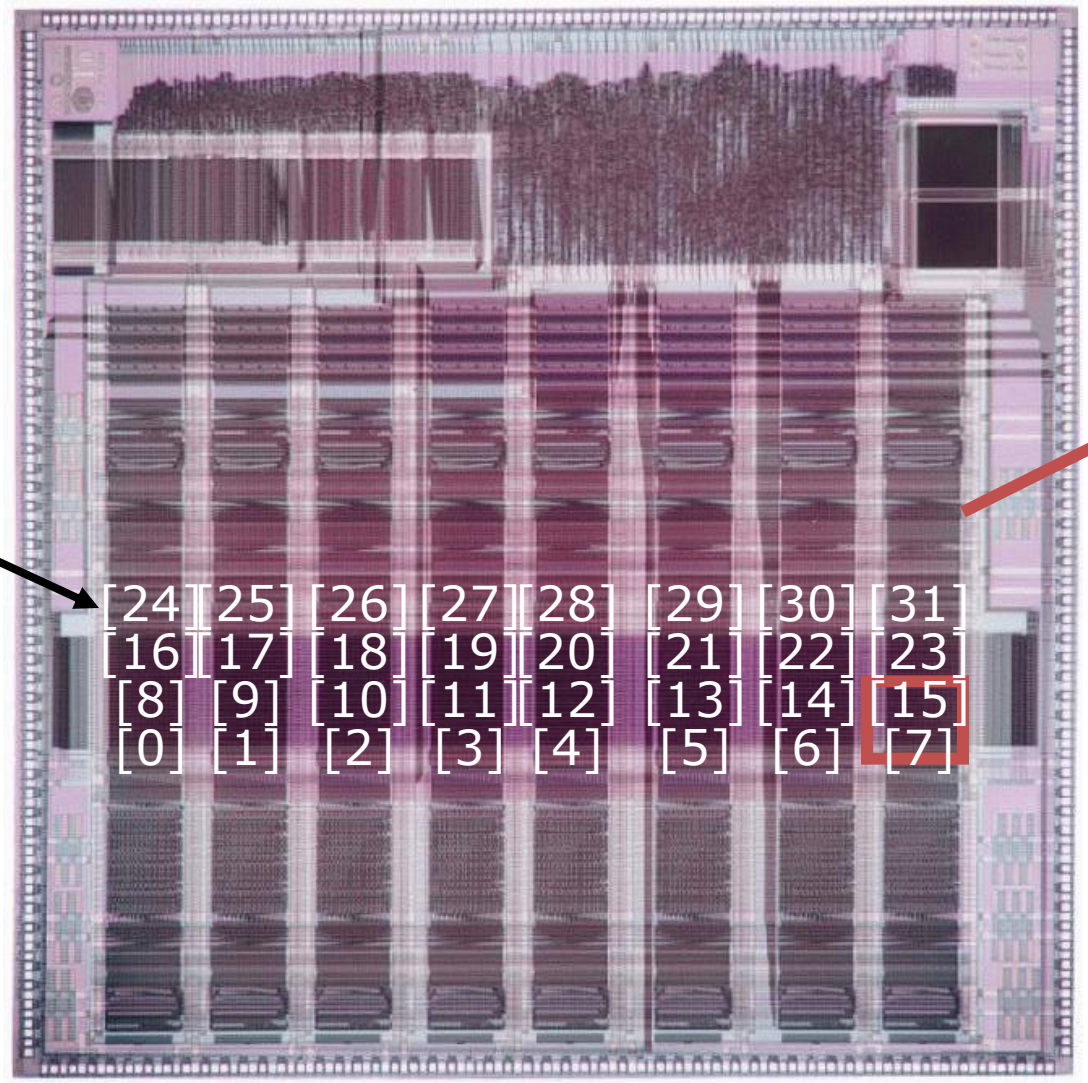
Interleaved Vector Memory System

Cray-1, 16 banks, 4 cycle bank busy time, 12 cycle latency

- *Bank busy time*: Time before bank ready to accept next request

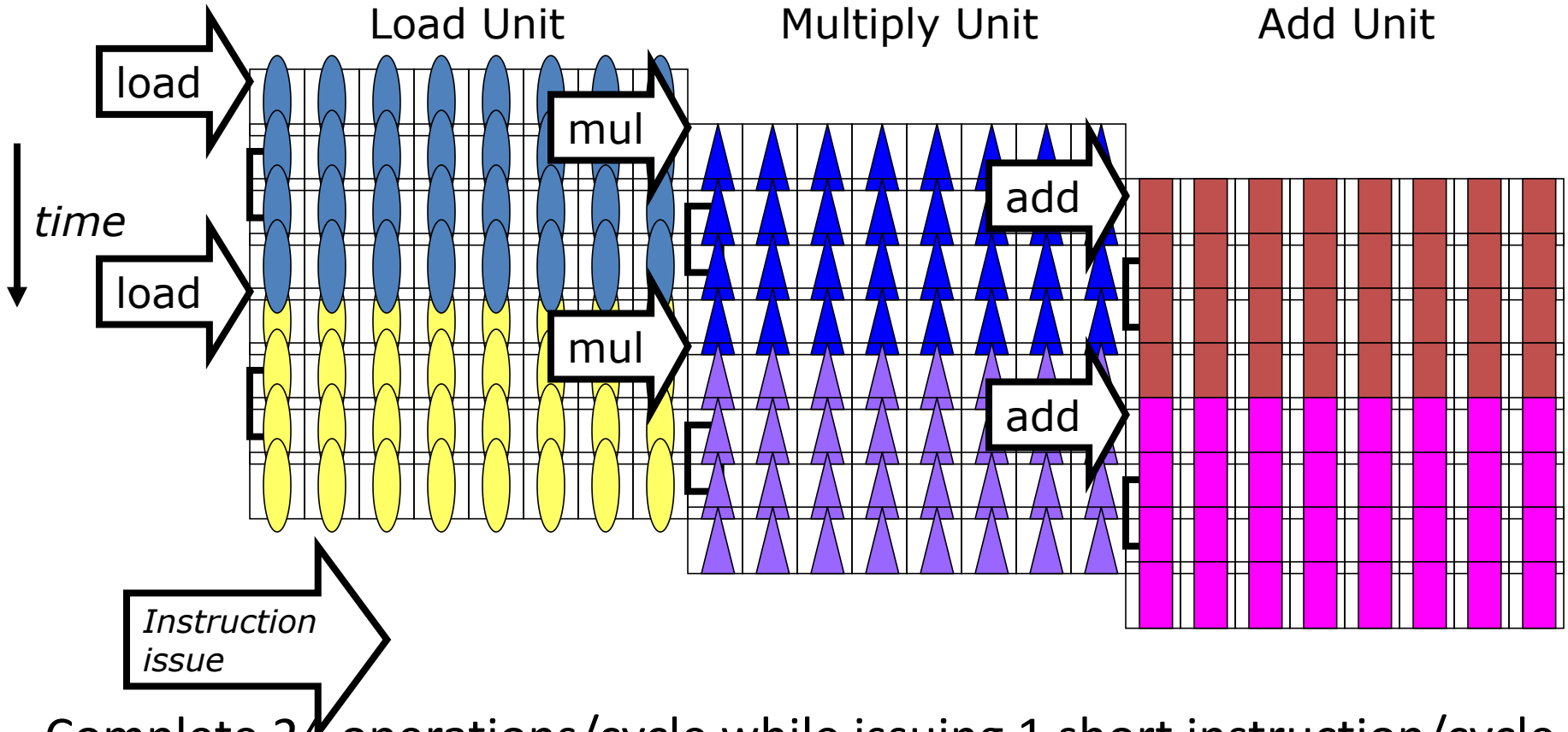


T0 Vector Microprocessor (UCB/ICSI, 1995)



Vector Instruction Parallelism

- **多条向量指令可重叠执行(链接技术)**
  - 例如：每个向量 32 个元素，8 lanes（车道）



Complete 24 operations/cycle while issuing 1 short instruction/cycle

6.1 向量处理机模型

数据级并行 → 向量处理机模型 → 性能评估

Vector Execution Time

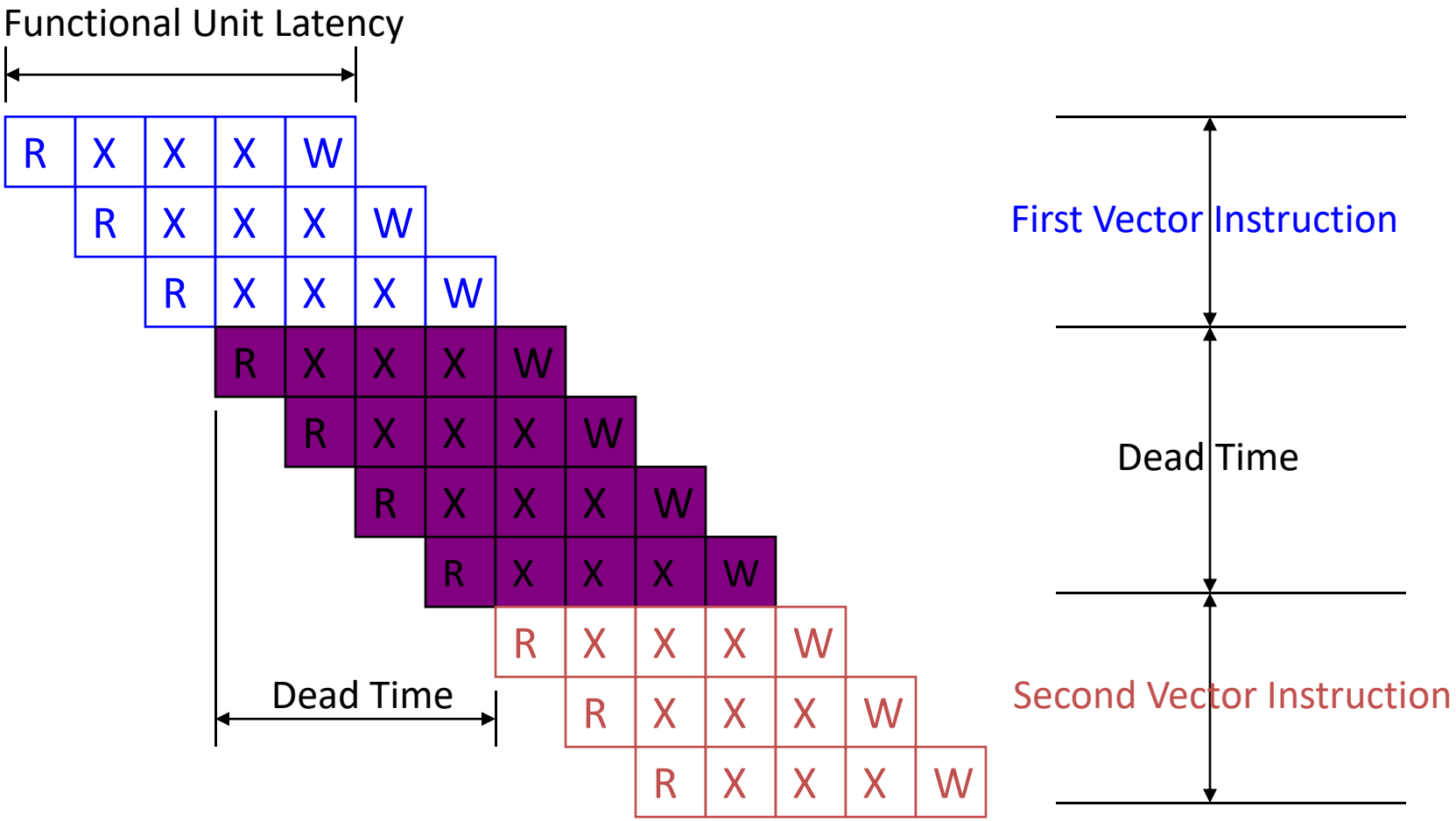
- *Time* = f(vector length, data dependencies, struct. hazards)
- *Initiation rate*: 功能部件消耗向量元素的速率
- *Convoy:* 可在同一时钟周期开始执行的指令集合 (no structural or data hazards)
- *Chime:* 执行一个*convoy*所花费的大致时间（approx. time)
- *m convoys take m chimes*;
  - 如果每个向量长度为*n*, 那么*m*个*convoys* 所花费的时间是*m*个*chimes*
  - 每个*chime*所花费的时间是*n*个*clocks*，该程序所花费的总时间大约为 *m x n clock cycles* (忽略额外开销; 当向量长度较长时这种近似是合理的)

```
1: LV    V1,Rx     ;load vector X
2: MULV  V2,F0,V1  ;vector-scalar mult.
   LV    V3,Ry     ;load vector Y
3: ADDV  V4,V2,V3  ;add
4: SV    Ry,V4     ;store the result
```

4 convoys, 1 lane, VL=64
=> 4 x 64 = 256 clocks
(or 4 clocks per result)

Vector Startup

- **向量启动时间由两部分构成**
  - 功能部件延时：一个操作通过功能部件的时间
  - 截止时间或恢复时间（dead time or recovery time）：运行下一条向量指令的间隔时间



VMIPS Start-up Time

Start-up time: FU 部件流水线的深度

Operation	Start-up penalty (from CRAY-1)
Vector load/store	12
Vector multiply	7
Vector add	6

Assume convoys don't overlap; vector length = n

Convoy	*Start*	*1st result*	*last result*	
1. LV	0	12	11+n	(12+n-1)
2. MULV, LV	12+n	12+n+7	18+2n	*Multiply startup*
	12+n	12+n+12	23+2n	*Load start-up*
3. ADDV	24+2n	24+2n+6	29+3n	*Wait convoy 2*
4. SV	30+3n	30+3n+12	41+4n	*Wait convoy 3*

Vector Length

- 当向量的长度不是64时（假设向量寄存器的长度是64）怎么办?
- vector-length register (VLR) 控制特定向量操作的长度, 包括向量的load/store. (当然一次操作的向量的长度不能 > 向量寄存器的长度) 例如:

```
      do 10    i = 1, n
10    Y(i) = a * X(i) + Y(i)
```

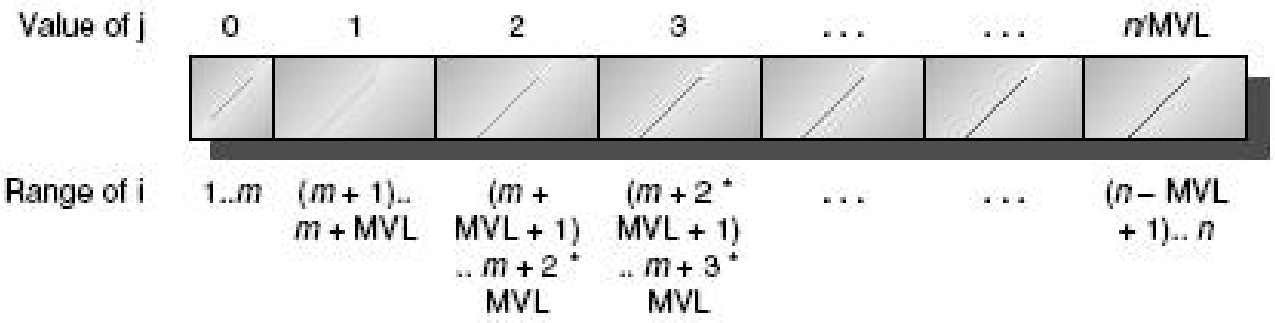
n的值只有在运行时才能知道

n > Max. Vector Length (MVL)怎么办?

Strip Mining（分段开采）

- 假设Vector Length > Max. Vector Length (MVL)?
- Strip mining: 产生新的代码，使得每个向量操作的元素数 ≤ MVL
- 第一次循环做最小片(n mod MVL), 以后按VL = MVL操作

```
low = 1
VL = (n mod MVL)  /*find the odd size piece*/
do 1  j = 0, (n / MVL)  /*outer loop*/
    do 10 i = low, low+VL-1  /*runs for length VL*/
            Y(i) = a*X(i) + Y(i)  /*main operation*/
10  continue
    low = low+VL  /*start of next vector*/
    VL = MVL  /*reset the length to max*/
1   continue
```



Strip Mining的向量执行时间计算

$$T_n = \left\lceil \frac{n}{MVL} \right\rceil \times (T_{loop} + T_{start}) + n \times T_{chime}$$

试计算A=B×s，其中A,B为长度为200的向量（每个向量元素占8个字节），s是一个标量。向量寄存器长度为64。各功能部件的启动时间如前所述，求总的执行时间，（T_{loop} = 15)

```
        ADDI R2,R0,#1600    ;total # bytes in vector
        ADD R2,R2,Ra        ;address of the end of A vector
        ADDI R1,R0,#8       ;loads length of 1st segment
        MOVI2S  VLR,R1      ;load vector length in VLR
        ADDI R1,R0,#64      ;length in bytes of 1st segment
        ADDI R3,R0,#64      ;vector length of other segments
Loop:   LV  V1,Rb           ;load B
        MULSV V2,V1,Fs      ;vector * scalar
        SV Ra,V2            ;store A
        ADD Ra,Ra,R1        ;address of next segment of A
        ADD  Rb,Rb,R1       ;address of next segment of B
        ADDI R1,R0,#512     ;load byte offset next segment
        MOVI2S VLR,R3       ;set length to 64 elements
        SUB R4,R2,Ra        ;at the end of A?
        BNEZ R4,Loop        ;if not, go back
```

$$T_n = \left\lceil \frac{n}{MVL} \right\rceil \times (T_{loop} + T_{start}) + n \times T_{chime}$$

$$T_{200} = 4 \times (15 + T_{start}) + 200 \times 3$$

$$T_{200} = 60 + (4 \times T_{start}) + 600 = 660 + (4 \times T_{start})$$

Tstart = 12 + 7 + 12 = 31

T200 = 660+4*31 = 784

每一元素的执行时间 = 784/200 = 3.9

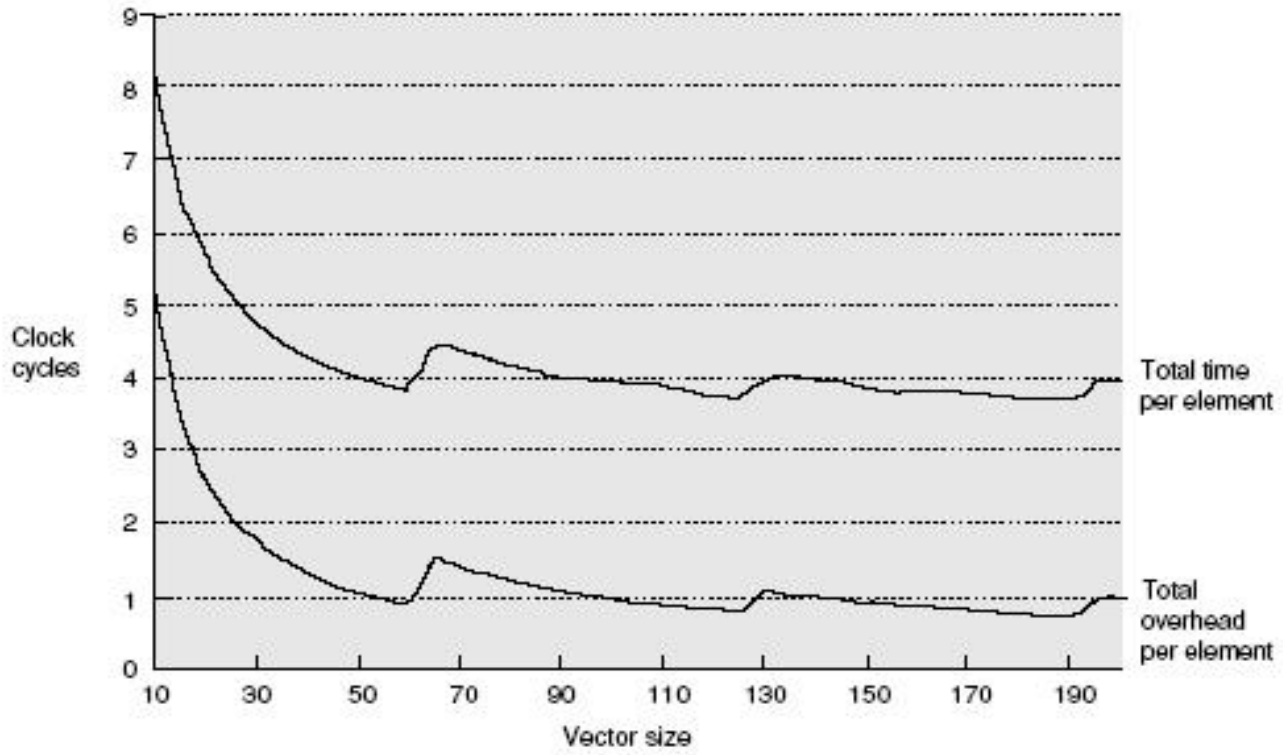


Figure G.9 The total execution time per element and the total overhead time per element versus the vector length for the example on page G-19. For short vectors the total start-up time is more than one-half of the total time, while for long vectors it reduces to about one-third of the total time. The sudden jumps occur when the vector length crosses a multiple of 64, forcing another iteration of the strip-mining code and execution of a set of vector instructions. These operations increase T_n by $T_{loop} + T_{start}$.

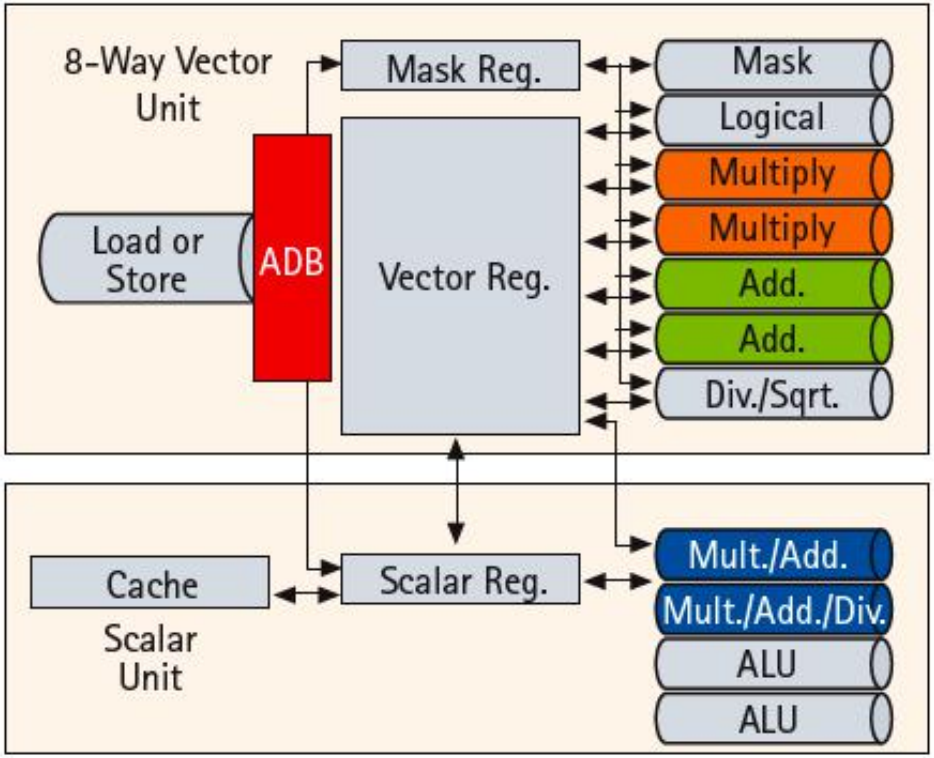
Common Vector Metrics

- **R_∞: 当向量长度为无穷大时的向量流水线的最大性能。常在评价峰值性能时使用，单位为MFLOPS**
  - 实际问题是向量长度不会无穷大，start-up的开销还是比较大的
  - R_n 表示向量长度为n时的向量流水线的性能
- **$N_{1/2}$: 达到R∞ 一半的值所需的向量长度，是评价向量流水线start-up 时间对性能的影响。**
- **N_V: 向量流水线方式的工作速度优于标量串行方式工作时所需的向量长度临界值。**
  - 该参数既衡量建立时间，也衡量标量、向量速度比对性能的影响

Example Vector Machines

Machine	Year	Clock(MHZ)	Regs	Elements	Fus	LSUs
Cray 1	1976	80	8	64	6	1
Cray XMP	1983	120	8	64	8	2L, 1S
Cray YMP	1988	166	8	64	8	2L, 1S
Cray C-90	1991	240	8	128	8	4
Cray T-90	1996	455	8	128	8	4
Conv. C-1	1984	10	8	128	4	1
Conv. C-4	1994	133	16	128	3	1
Fuj. VP200	1982	133	8-256	32-1024	3	2
Fuj. VP300	1996	100	8-256	32-1024	3	2
NEC SX/2	1984	160	8+8K	256+var	16	8
NEC SX/3	1995	400	8+8K	256+var	16	8

A Modern Vector Super: NEC SX-9 (2008)



- Memory system provides 256GB/s DRAM bandwidth per CPU
- Up to 16 CPUs and up to 1TB DRAM form shared-memory *node*
  - total of 4TB/s bandwidth to shared DRAM memory
- Up to 512 nodes connected via 128GB/s network links (message passing between nodes)

- 65nm CMOS technology
- Vector unit (3.2 GHz)
  - 8 foreground VRegs + 64 background VRegs (256x64-bit elements/VReg)
  - 64-bit functional units: 2 multiply, 2 add, 1 divide/sqrt, 1 logical, 1 mask unit
  - 8 lanes (32+ FLOPS/cycle, 100+ GFLOPS peak per CPU)
  - 1 load or store unit (8 x 8-byte accesses/cycle)
- Scalar unit (1.6 GHz)
  - 4-way superscalar with out-of-order and speculative execution
  - 64KB I-cache and 64KB data cache

Vector Linpack Performance (MFLOPS)

Matrix Inverse (gaussian elimination)

Machine	Year	Clock(Mhz)	100x100	1kx1k	Peak(Procs)
Cray 1	1976	80	12	110	160(1)
Cray XMP	1983	120	121	218	940(4)
Cray YMP	1988	166	150	307	2,667(8)
Cray C-90	1991	240	387	902	15,238(16)
Cray T-90	1996	455	705	1603	57,600(32)
Conv. C-1	1984	10	3	--	20(1)
Conv. C-4	1994	136	160	2531	3,240(4)
Fuj. VP200	1982	133	18	422	533(1)
NEC SX/2	1984	166	43	885	1,300(1)
NEC SX/3	1995	400	368	2757	25,600(4)

Summary：向量体系结构

- **向量处理机基本概念**
  - 基本思想：两个向量的对应分量进行运算，产生一个结果向量
- **向量处理机基本特征**
  - VSIW-一条指令包含多个操作
  - 单条向量指令内所包含的操作相互独立
  - 以已知模式访问存储器-多体交叉存储系统
  - 控制相关少
- **向量处理机基本结构**
  - **向量指令并行执行**
  - **向量运算部件的执行方式-流水线方式**
  - **向量部件结构-多“道”结构-多条运算流水线**
- **向量处理机性能评估**
  - **向量指令流执行时间: Convey, Chimes, Start-up time**
  - **其他指标：** R_∞ , $N_{1/2}$, N_V
- **向量处理机性能优化**
  - 链接技术
  - 条件执行
  - 稀疏矩阵

Acknowledgements

- **These slides contain material developed and copyright by:**
  - John Kubiatowicz (UCB)
  - Krste Asanovic (UCB)
  - John Hennessy (Standford)and David Patterson (UCB)
  - Chenxi Zhang (Tongji)
  - Muhamed Mudawar (KFUPM)
- **UCB material derived from course CS152、CS252、CS61C**
- **KFUPM material derived from course COE501、COE502**