

第四章 向量处理机

- 具有向量数据表示和向量指令系统的处理机

4.1 向量处理的基本概念

4.2 向量处理机的结构

4.3 向量处理方式

4.4 向量处理机的关键技术

4.5 向量处理机实例

4.6 向量处理机的发展

4. 1 向量处理的基本概念

- 向量处理机是解决数值计算问题的一种高性能计算机结构
- 向量处理机一般都采用流水线结构，有多条流水线并行工作
- 向量处理机通常属大型或巨型机，也可以用微机加一台向量协处理器组成
- 一般向量计算机中包括有一台高性能标量处理机
- 必须把问题转化为向量运算，向量处理机才能发挥作用

4.1.1 从标量到向量

- 一个典型的向量求解问题：
 $Y = a \times X + Y$ ，其中 a 为标量， X 和 Y 为向量，
初始值放在存储器中。

例1 用标量处理机来计算。

```
LD  F0, a           ; 标量a装入寄存器F0
ADD  R4, Rx, #512    ; 向量元素的末地址装入寄存器
LOOP: LD  F2, M(Rx)   ; 取向量元素X(i)
      MUL F2, F0, F2   ; a与X(i)相乘
      LD  F4, M(Ry)   ; 取向量元素Y(i)
      ADD F4, F2, F4   ; aX(i)与Y(i)相加
      SD  M(Ry), F4   ; 存储结果向量元素
      ADD Rx, Rx, #8   ; X向量元素下标加1
      ADD Ry, Ry, #8   ; Y向量元素下标加1
      SUB R20, R4, Rx  ; (R4)-(Rx)→R20, 计算是否到限界值
      BNZ R20, LOOP   ; 若循环未结束, 转LOOP
```

- 当向量长度为64时, 共需执行 $9 \times 64 + 2 = 578$ 条指令

例2 用向量处理机来计算

LD F0, a ; 标量a装入F0

LV V1, M(X) ; 向量X装入V1向量寄存器(LV为向量取指令)

MULV V2, F0, V1; 向量X与标量a相乘(MULV为向量乘指令)

LV V3, M(Y) ; 向量Y装入V3向量寄存器

ADDV V4, V2,V3 ; 向量加 $aX + Y$ (ADDV为向量加指令)

SV M(Y), V4 ; 存储结果向量(SV为向量存指令)

- 只需执行6条指令。

例3：一个简单的C语言程序如下：

```
for (i = 10; i <= 1010; i++)
```

```
    c[i] = a[i] + b[i+5] ;
```

- 在向量处理机上，只用一条指令：

C(10:1010)=A(10:1010) + B(15 :1015)

一条向量指令可以处理N个或N对操作数。

- 在标量处理机上用10多条指令，其中有8条指令要循环1000次。

采用多寄存器结构的两地址指令编写程序。

存储器采用字节编址方式，字长为32位。

在一般标量处理机中需要如下指令序列来实现：

A、B、C分别是向量**a、b、c**在内存中的起始地址。

START:	LOAD	R0, ST	; 读循环初值, 10
	LOAD	R1, ED	; 读循环终值, 1010
	LOAD	R2, L	; 读内存地址增量, 常数4
	MOVE	R3, R2	
	MUL	R3, R0	; 向量偏移量, 初始值为40
LOOP:	LOAD	R4, A(R3)	; 读A向量的一个元素
	LOAD	R5, B(R3)	; 读B向量的一个元素
	ADD	R4, R5	
	STORE	R4, C(R3)	; 写C向量的一个元素
	ADD	R3, R2	; 改变向量偏移量
	INC	R0	; 循环次数增1
	CMP	R0, R1	; 循环是否结束
	BLE	LOOP	; 循环未结束转LOOP, 否则继续
	HALT		
ST:	10		; 循环初值
ED:	1010		; 循环终值
L:	4		; 内存地址增量

第四章 向量处理机

4.1 向量数据表示方式

4.2 向量处理机的结构

4.3 向量处理方式

4.4 向量处理机的关键技术

4.5 向量处理机实例

4.6 向量处理机的发展

4.2 向量处理机结构

- 向量处理机的最关键问题是存储器系统能够满足运算部件带宽的要求。
- 主要采用两种方法：

(1) 存储器—存储器结构。

多个独立的存储器模块并行工作。

处理机结构简单，对存储系统的访问速度要求很高

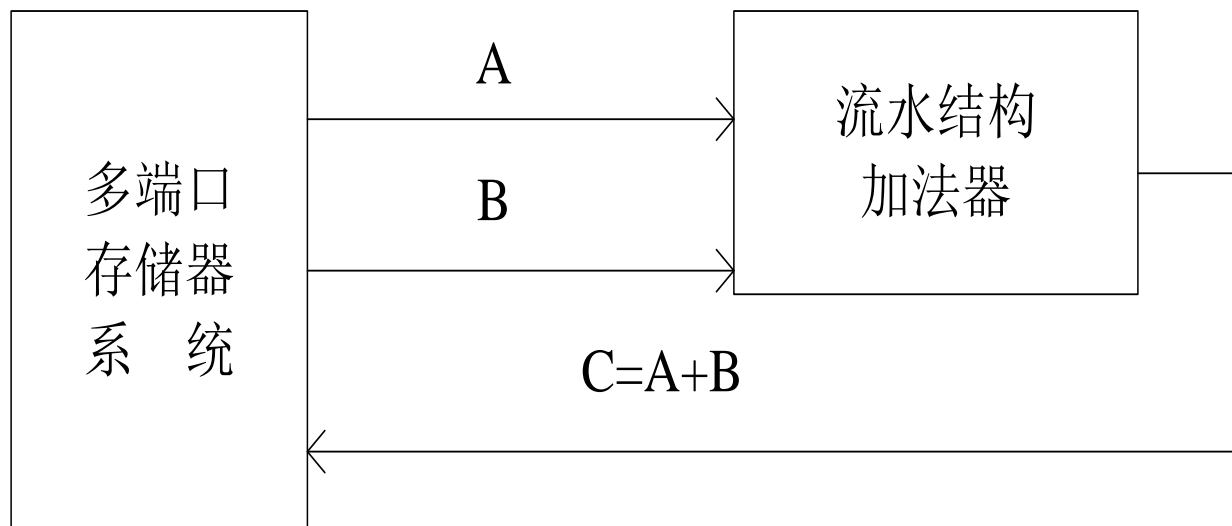
(2) 寄存器—寄存器结构。

运算通过向量寄存器中进行。

需要大量高速寄存器，对存储系统访问速度的要求降低

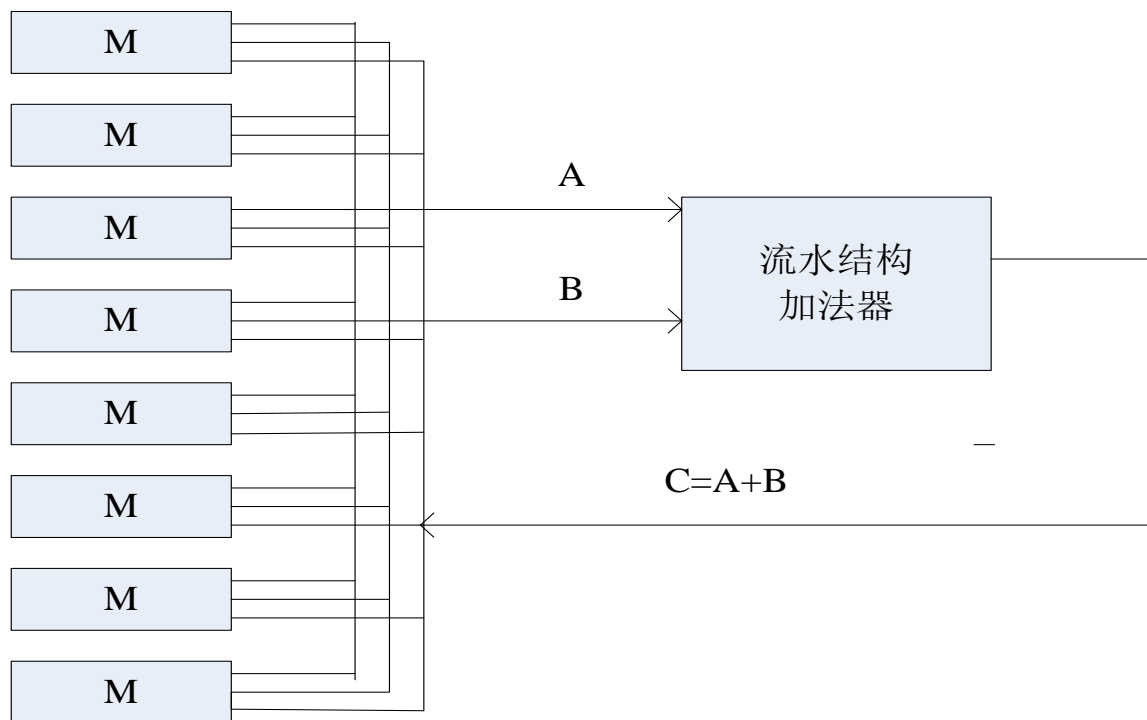
4.2.1 存储器—存储器结构

1. 存储器—存储器结构简单框图



- 有多个高速流水线运算部件，存储器的访问速度是关键

2. 多个存储体组成的向量处理机



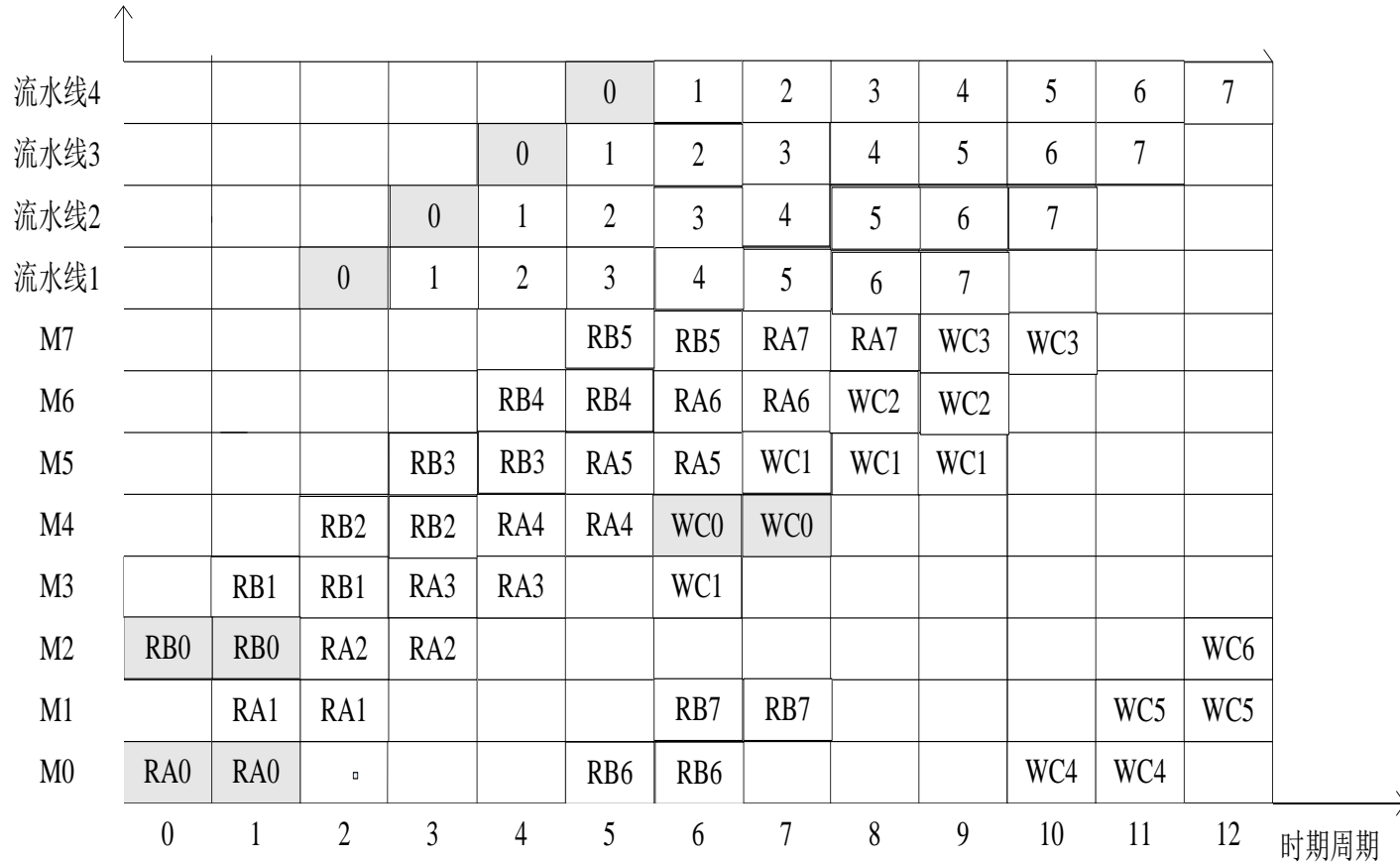
采用多个存储体交叉和并行访问来提高存储器速度

例如：CRAY-1有64个存储体，每个处理机访问4个存储体。

STAR-100采用32个存储体交叉. 我国研制的YH-1向量计算机有37个存储体

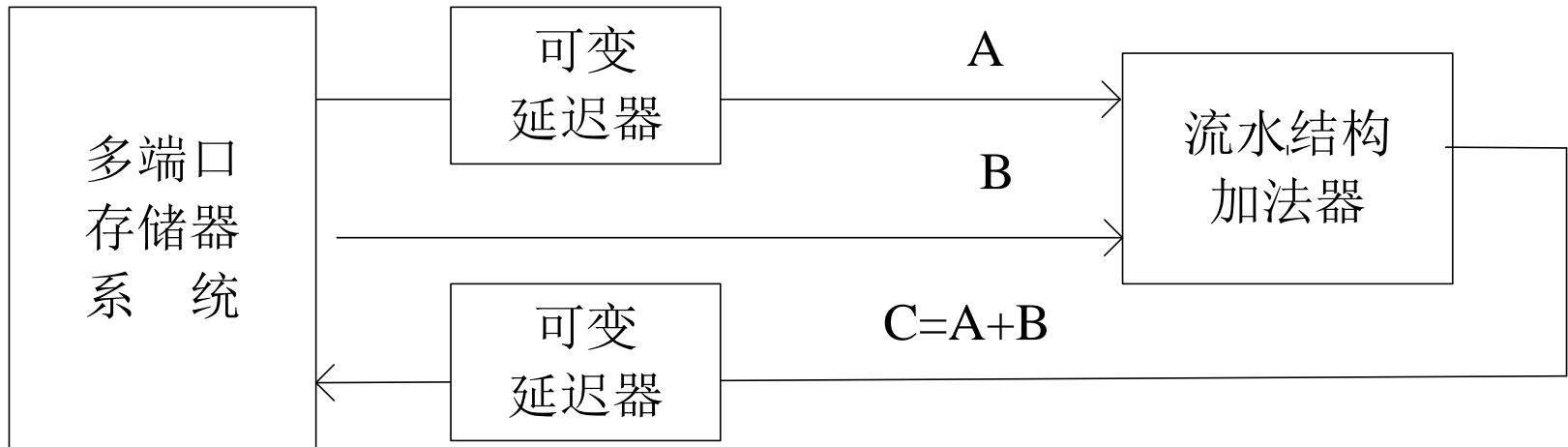
<http://www.wdklife.com> 五道口生活网
<http://www.wdklife.com/bbs> 五道口论坛

向量计算 $C=A+B$ 的时空图

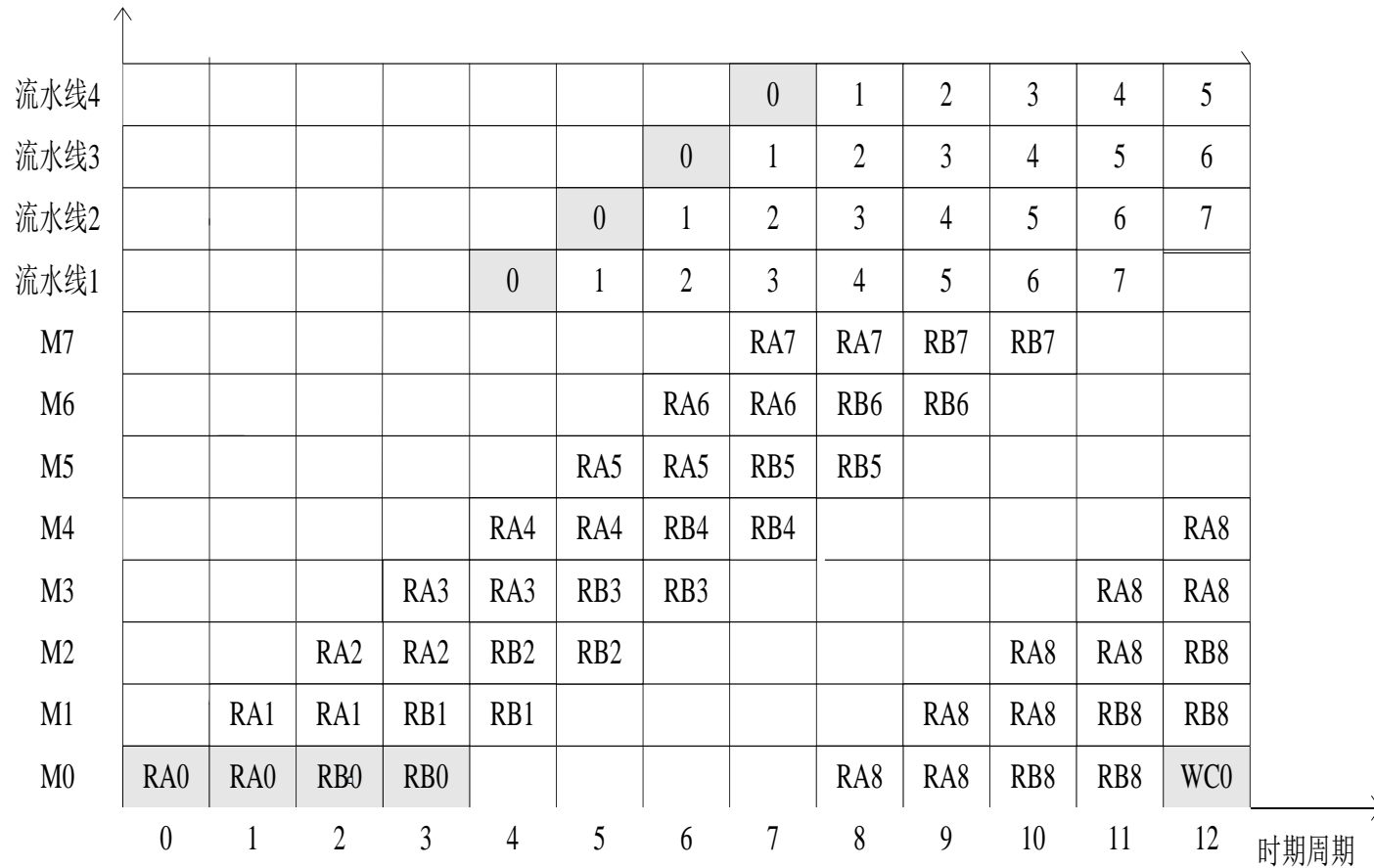


<http://www.wdklife.com> 五道口生活
 网 <http://www.wdklife.com/bbs> 五道
 论坛

3.具有延迟缓冲器的向量流水线结构



具有延迟缓冲器的向量计算时空图



<http://www.wdklife.com> 五道口生活
<http://www.wdklife.com/bbs> 五道
 论坛

4.2.2 寄存器-寄存器结构

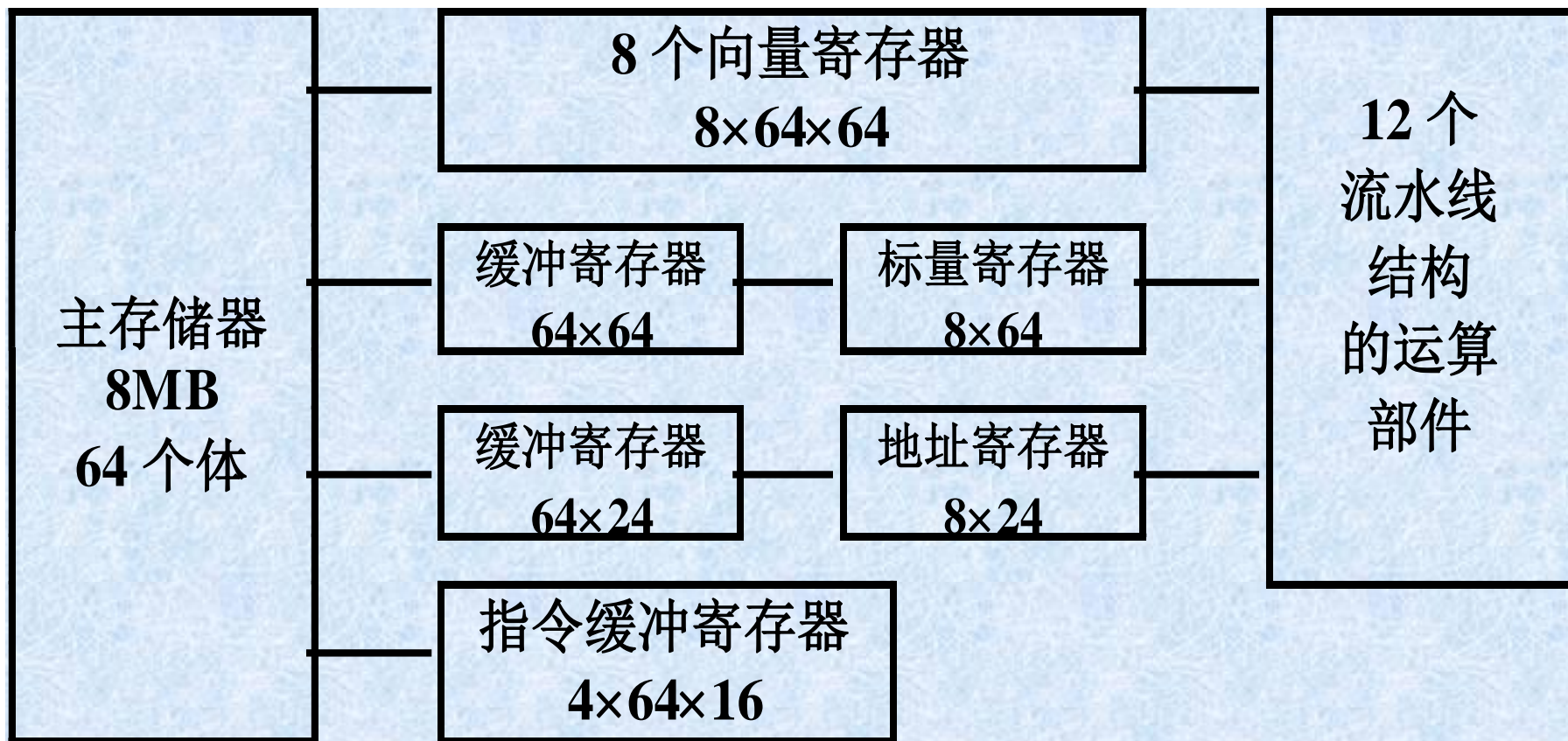
- 把存储器-存储器结构中的缓冲栈改为向量寄存器，
运算部件需要的操作数从向量寄存器中读取，
运算的中间结果也协到向量寄存器中。
- 向量寄存器与标量寄存器的主要差别是：
向量寄存器的读写单位是向量。
- 需要有标量寄存器和地址寄存器等。
- 主要优点：降低主存储器的流量。

例如：采用寄存器-寄存器结构的CRAY-1与
采用存储器-存储器结构STAR-100比较，
运算速度快3倍多，而主存流量低2.5倍。

STAR-100的主存储器流量： $32 \times 8W / 1.28\mu s = 200MW/S$

CRAY-1的主存储器流量： $4W / 50ns = 80MW/S$

<http://www.wdklife.com> 五道口生活
网 <http://www.wdklife.com/bbs> 五道
论坛



CRAY-1 向量处理机结构

向量指令处理时间 T_{vp}

执行一条向量长度为 n 的向量指令的时间 T_{vp} 表示为：

$$T_{vp} = T_s + T_e + (n - 1)T_c$$

其中： T_s 为向量流水线的建立时间。

T_e 为向量流水线的流过时间。

T_c 为流水线“瓶颈”段的执行时间。

- 如果每段执行时间都等于一个时钟周期，则有：

$$T_{vp} = sT_c + lT_c + (n - 1)T_c = (s + l + n - 1)T_c$$

其中： s 为向量流水线建立时间所需的时钟周期数。

e 为向量流水线流过时间所需的时钟周期数。

n 为向量长度。 T_c 为时钟周期长度。

例：A、B两个向量存放于存储器，其向量长度为64。设流水加法器有4级，流水线时钟周期为10ns，读出A、B向量第一对元素到流水线始端所需的时钟周期数为2，求执行向量加法指令ADDV所需的时间。

解：由题意知， $n=64$ ， $l=4$ ， $s=2$ ， $T_c=10\text{ns}$
 $T_{vp}=(s+l+n-1)T_c=(2+4+64-1)10\text{ns}=690\text{ns}$

第四章 向量处理机

- 具有向量数据表示和向量指令系统的处理机称为向量处理机

4.1 向量数据表示方式

4.2 向量处理机的结构

4.3 向量处理方式

4.4 向量处理机的关键技术

4.5 向量处理机实例

4.6 向量处理机的发展

4.3 向量处理方式

- 根据向量运算的特点和向量处理机的类型选择向量处理方式。有三种向量处理方式：

(1)横向处理方式，又称为水平处理方式，横向加工方式等。

向量计算是按行的方式从左至右横向地进行。

(2)纵向处理方式，又称为垂直处理方式，纵向加工方式等。

向量计算是按列的方式自上而下纵向地进行。

(3)纵横处理方式，又称为分组处理方式，纵横向加工方式等。

横向处理和纵向处理相结合的方式。

- 以一个简单的C语言程序为例，说明三种处理方式的原理。

```
for (i = 1; i <= n; i++)
```

```
    y[i] = a[i] ( b[i] + c[i] );
```

<http://www.wdklife.com> 五道口生活
网 <http://www.wdklife.com/bbs> 五道
论坛

4.3.1 横向处理方式

- 逐个分量进行处理(假设中间结果为 $T(I)$):

计算第1个分量: $T(1) = B(1) + C(1)$

$$Y(1) = A(1) \times T(1)$$

计算第2个分量: $T(2) = B(2) + C(2)$

$$Y(2) = A(2) \times T(2)$$

.....

计算最后一个分量: $T(N) = B(N) + C(N)$

$$Y(N) = A(N) \times T(N)$$

- 存在问题:

计算每个分量时都发生写读数据相关, 流水线效率低。

如果采用多功能流水线, 必须频繁进行流水线切换。

- 横向处理方式对向量处理机不适合。

即使在标量处理机中, 也经常通过编译器进行指令流调度。

4.3.2 纵向处理方式

- 处理方式：按列自上而下纵向进行计算

$$T(1) = B(1) + C(1)$$

$$T(2) = B(2) + C(2)$$

.....

$$T(n) = B(n) + C(n)$$

$$Y(1) = A(1) \times T(1)$$

$$Y(2) = A(2) \times T(2)$$

.....

$$Y(N) = A(N) \times T(N)$$

- 采用向量指令只需要2条：

VADD B, C, T

VMUL A, T, Y

- 特点：数据相关不影响流水线连续工作。

4.3.3 纵横处理方式

- 用于寄存器-寄存器结构的向量处理机中，
当向量长度 N 大于向量寄存器长度 n 时，需要分组处理。
- 分组方法： $N = K \cdot n + r$ ，其中： r 为余数，共分 $K + 1$ 组。
组内采用纵向处理方式，组间采用横向处理方式
- 运算过程为： 第 1 组： $T(1,n) = B(1,n) + C(1,n)$
 $Y(1,n) = A(1,n-1) \times T(1,n)$
第 2 组： $T(n+1,2n) = B(n+1,2n) + C(n+1,2n)$
 $Y(n+1,2n) = A(n+1,2n) \times T(n+1,2n)$
.....
最后第 $k+1$ 组： $T(kn+1,N) = B(kn+1,N) + C(kn+1,N)$
 $Y(kn+1,N) = A(kn+1,N) \times T(kn+1,N)$
- 每组用两条向量指令，每组发生数据相关两次，其中：
组内发生数据相关一次，组间切换时发生数据相关一次。
- 优点： **减少访问主存储器的次数，中间变量不写入主存储器**
网 <http://www.wdklife.com/bbs> 五道论坛

第四章 向量处理机

- 具有向量数据表示和向量指令系统的处理机称为向量处理机。

4.1 向量数据表示方式

4.2 向量处理机的结构

4.3 向量处理方式

4.4 向量处理机的关键技术

4.5 向量处理机实例

4.6 向量处理机的性能评价

4.7 向量处理机的发展

4.4 向量处理机的关键技术

4.4.1 向量与标量性能的平衡

4.4.2 向量链接技术

4.4.3 向量循环开采技术

4.4.1 向量与标量性能的平衡

- 实际的应用问题中通常既有向量计算又有标量计算，而且两类计算有一定的比例。
- 向量平衡点(vector balance point): 为了使向量硬件设备和标量硬件设备的利用率相等，一个程序中向量代码所占的百分比。
- 关键问题是：希望向量硬件和标量硬件都能够充分利用。
例如：一个系统的向量运算速度为90Mfolps，标量运算速度为 10Mfolps。如果程序的90%是向量运算，10%是标量运算。则向量平衡点为0.9。硬件利用率最高。
- 向量平衡点必须与用户程序的向量化程度相匹配。
- IBM向量计算机的设计思想与上述方法不同，它维持较低的向量与标量比例，定在3~5的范围之间。这种做法能够适应通用应用问题对标量和向量处理要求。

几种超级计算机向量和标量的性能

机器型号	Cray IS	Cray 2S	Cray X-MP	Cray Y-MP	hitachi S820	NEC SX2	Fujitsu VP4000
向量性能(Mflops)	85.0	151.5	143.3	201.6	737.3	424.2	207.1
标量性能(Mflops)	9.8	11.2	13.1	17.0	17.8	9.5	6.6
向量平衡点	0.90	0.93	0.92	0.92	0.98	0.98	0.97

<http://www.wdklife.com> 五道口生活网
<http://www.wdklife.com/bbs> 五道口论坛

4.4.2 向量链接技术

1、向量指令的类型：

以CRAY-1向量处理机为例，有四类指令，两种指令格式。

(1) 向量与向量操作， $V_i \leftarrow V_j \text{ OP } V_k$

(2) 向量与标量操作， $V_i \leftarrow S_j \text{ OP } V_k$

(3) 向量取， $V_i \leftarrow ((Ah) + jkm)$

(4) 向量存， $((Ah) + jkm) \leftarrow V_i$

CRAY 向量处理机的指令格式

4 位	3 位	3 位	3 位	3 位
g	h	i	j	k

gh 为操作码，i 为目的寄存器编号，j、k 为源寄存器编号。

4 位	3 位	3 位	3 位	3 位	16 位
g	h	i	j	k	m

g 为操作码，h 为变址寄存器 A 的编号，

i 为目的寄存器编号，jkm 共 22 位为形式地址。

2、向量运算中的相关和冲突

- 向量运算中的数据相关和功能部件冲突主要有：
采用顺序发射顺序完成方式。

(1) 写读数据相关。

(2) 读读数据相关，或向量寄存器冲突。

(3) 运算部件冲突。

$$V0 \leftarrow V1 + V2$$

$$V3 \leftarrow V4 \times V5$$

(a) 不相关的指令

$$V0 \leftarrow V1 + V2$$

$$V3 \leftarrow V0 \times V4$$

(b) 写读数据相关

$$V0 \leftarrow V1 + V2$$

$$V3 \leftarrow V4 + V5$$

(c) 功能部件冲突

$$V0 \leftarrow V1 + V2$$

$$V3 \leftarrow V1 \times V4$$

(d) 读读数据相关

3、向量链接技术(chaining)

- 结果寄存器可能成为后继指令的操作数寄存器。

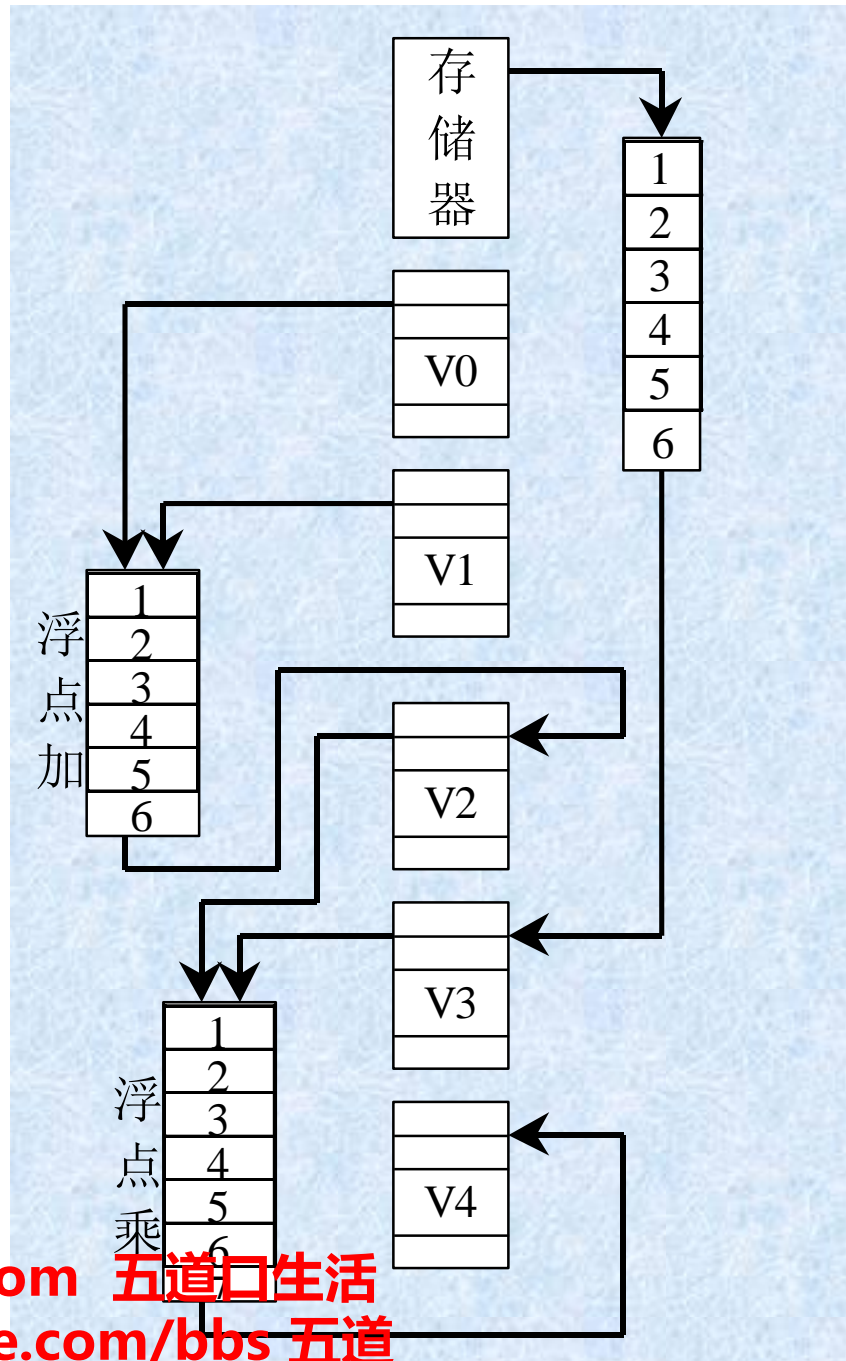
两条有数据相关的向量指令并行执行，这种技术称为两条流水线的链接技术。

- 例如：有如下3条向量指令：

$V3 \leftarrow A$
 $V2 \leftarrow V0 + V1$
 $V4 \leftarrow V2 * V3$

第一、二条指令没有数据相关和功能部件冲突，可以同时开始执行。

第三条指令与第一、二条指令均存在写读数据相关，可以链接执行。



- 三种执行方式比较:

(1) 向量长度为N, 三条指令采用串行方法执行的时间为:

$$[(1+6+1)+N-1]+[(1+6+1)+N-1]+[(1+7+1)+N-1] = \mathbf{3N+22 \text{ 拍}}$$

(2) 前两条指令并行执行, 第三条串行执行, 执行时间为:

$$[(1+6+1)+N-1]+[(1+7+1)+N-1] = \mathbf{2N+15 \text{ 拍}}$$

(3) 如果采用链接技术, 执行时间为:

$$(1+6+1)+(1+7+1)+(N-1)=17+N-1 = \mathbf{N+16 \text{ 拍}}$$

- 实现链接的条件:

(1) 没有向量寄存器冲突和运算部件冲突。

(2) 只有第一个结果送入向量寄存器的那一个周期可以链接。

(3) 如果一条向量指令的两个源操作数分别是两条先行指令的执行结果, 则要求这两条指令产生运算结果的时间必须相同。

(4) 两条向量指令的向量长度必须相等。

4.4.3 向量循环开采技术

- 当向量的长度大于向量寄存器的长度时，必须把长向量分成长度固定的段，采用循环结构处理这个长向量，这种技术称为向量循环开采技术，也称为向量分段开采技术。

有一个循环程序代码段：

DO 10 i=1, n

10 A(i)=5*B(i)+C

其中n和C为常数，设向量寄存器长度为64元素，请用分段开采技术改造成向量循环形式。

当n为64或更小时，产生A数组的7条指令序列是：

- | | |
|-----------------------------------|-----------------------|
| 1: $S_1 \leftarrow 5.0$ | 在标量寄存器内设置常数 |
| 2: $S_2 \leftarrow C$ | 将常数C装入标量寄存器 |
| 3: $VL \leftarrow n$ | 在VL寄存器内设置向量长度 |
| 4: $V_o \leftarrow B$ | 将B向量读入向量寄存器 |
| 5: $V1 \leftarrow S_1 \times V_o$ | B数组的每个分量和常数相乘 |
| 6: $V2 \leftarrow S_2 + V1$ | C和 $5 \times B(x)$ 相加 |
| 7: $A \leftarrow V2$ | 将结果向量存入A数组 |

- 当n超过64时，就需要采用向量循环开采技术。
在进入循环以前，把n除以64，以确定循环次数。
如果有余数，则在第一次循环中首先计算余数个分量。

LOW = 1

VL = (n mod 64) ; 找出余数长度值

DO 20 j = 1, (n/64) ; 外循环

DO 10 i = LOW, LOW + VL - 1 ; 以长度VL操作

A(i) = 5 * B(i) + C ; 主要操作

10 continue

LOW = LOW + VL ; 下一向量的开始

VL = 64 ; 将向量长度值恢复成64

20 continue

第四章 向量处理机

- 具有向量数据表示和向量指令系统的处理机称为向量处理机。

4.1 向量数据表示方式

4.2 向量处理机的结构

4.3 向量处理方式

4.4 向量处理机的关键技术

4.5 向量处理机实例

4.6 向量处理机的发展

4.5 向量处理机实例

4.5.1 典型向量处理机

4.5.2 CRAY Y-MP向量处理机

4.5.3 向量协处理器

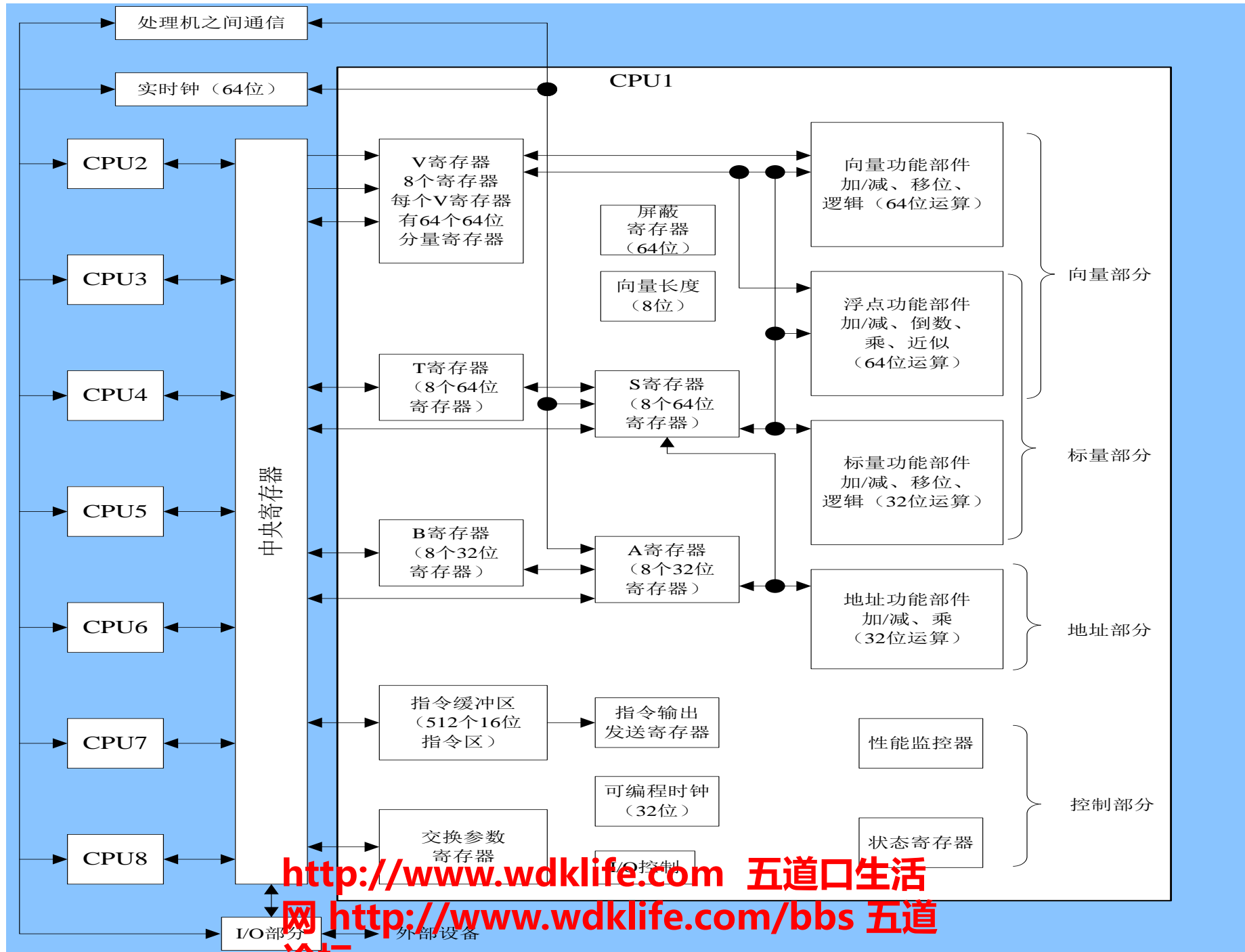
- 向量处理机主要出自美国和日本。
- 美国著名的向量计算机公司有：CRAY、CDC、TI等
- 日本公司有：NEC、Fujitsu、Hitachi等

4.5.1 典型向量处理机

系统型号	最大配置、周期、操作系统/编译系统	特色和要点
Cray 1S	有 10 条流水线的单处理机, 12.5ns, COS/CF7 2.1	第一台基于 ECL 的超级计算机, 1976 年问世
Cray 2S/4—256	256M 字存储器的 4 台处理机, 4.1ns, COS 或 UNIX/CF77 3.0	16K 字的本地存储器, 移植了 UNIXV, 1985 问世
Cray X-MP 416	16M 字存储器的 4 台处理机, 128M 字 SSD, 8.5ns, COS CF77 5.0	使用共享寄存器组用于 IPC, 1983 年问世
Cray Y-MP 832	128M 字存储器的 8 台处理机, 6ns, CF77 5.0	X—MP 的改进型, 1988 年问世
Cray Y-MP C-90	每台处理机 2 条向量流水线, 16 台处理机, 4.2ns, UNICOS/CF77 5.0	最大的 Cray 机器, 1991 年问世
CDC Cyber 205	有 4 条流水线的单处理机, 20ns, 虚拟 OS/FTN200	存储器到存储器系统结构, 1982 年问世
ETA 10E	单处理机, 10.5ns, ETAV/FTN 200,	Cyber 205 的后继型号, 1985 年问世
NEC SX-X/44	每台处理机 4 组流水线, 4 台处理机, 2.9ns, F77SX, 22Gflops	1991 年问世
Fujitsu VP2600/10	5 条流水线的单处理机和双标量处理机, 3.2ns, MSP. EX/F77 EX/VP	使用可重构微向量寄存器和屏蔽, 1991 年问世
Hitachi 820/80	512MB 存储器, 18 条流水线的单处理机, 4ns, FORT77/HAP V23-OC	64 个通道, 最大传输速率 288MB/S, 1988 年问世

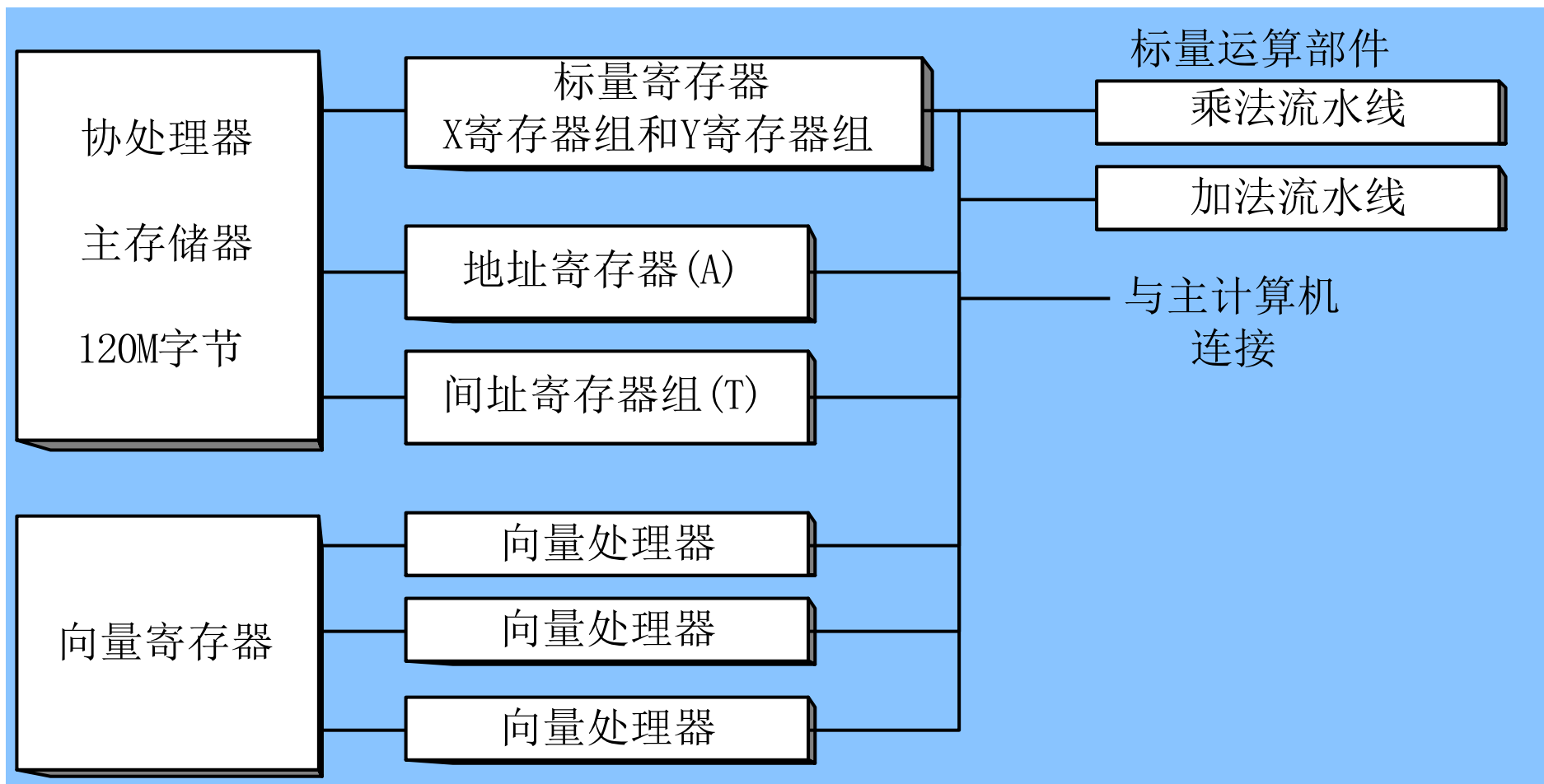
4.5.2 CRAY Y-MP向量处理机

- Cray Y-MP 816由1至8台处理机组成，多个处理机共享中央存储器、I/O子系统、处理机通信子系统和实时钟。
- 中央存储器由256个交叉访问的存储体组成。每个处理机对4个存储器端口的交叉访问。CPU的时钟周期为6ns。
- 4个存储器端口允许处理机同时执行两个标量和向量取操作、一个存储操作和一个独立的I/O操作。
- 每个CPU由14个功能部件组成，分为向量、标量、地址和控制四个子系统。
- 使用了大量地址寄存器、标量寄存器、向量寄存器、中间寄存器和临时寄存器。
- 可以实现功能流水线灵活的链接。
- I/O子系统支持三类通道，传输速率分别为6兆字节/秒，100兆字节/秒和1000兆字节/秒。



4.5.3 向量协处理器

- 以通用中小型机，或微机作为主机；向量处理部件作为外围设备，加速向量的处理速度。
 - 向量协处理器是为中小型用户设计的，解决科学计算中大量向量处理任务的一种装置。
 - 与各种主机相连的向量协处理器，价格和功能的变化范围很大。
 - FPS-164是最典型的向量协处理器，美国浮点系统公司生产。每个向量处理器有两个乘加部件，两组向量寄存器，两组标量寄存器。每个乘加部件每个周期能输出一个结果。
- 向量寄存器：2组×4个×2 K个操作数，每个操作数4个字节。
- 各向量处理器同步地运算，但它们处理的数据各不相同。
 - 向量操作可以和标量处理器中的标量操作同时进行。
 - 向量协处理器特别适合于大规模的数值处理，用户购买需要台数的向量处理器，使用现有的处理机作为主机。



FPS-164 向量协处理器的结构

第四章 向量处理机

- 具有向量数据表示和向量指令系统的处理机称为向量处理机。

4.1 向量数据表示方式

4.2 向量处理机的结构

4.3 向量处理方式

4.4 向量处理机的关键技术

4.5 向量处理机实例

4.6 向量处理机的发展

4.6 向量处理机的发展

1、向量计算机系统结构的发展趋势是：

- (1) 提供多种向量运算指令。
- (2) 除具有向量处理功能外还有其它功能。
- (3) 采用多层次的存储器系统。
- (4) 流水线技术与并行技术相结合。

2、向量计算机系统结构要解决的六个技术问题：

- (1) 处理机带宽，两种方法：
 - 运算部件采用流水线结构。
 - 用多个运算器构成并行系统。

(2) **存储器带宽**，多种解决方法：

- 用多个独立的存储体构造一个大容量的存储器系统。
- 采用多层次的存储器系统提高访问速度。
- 采用高速缓冲存储器和可寻址的寄存器组效果最好。
- 采用流水线技术，存储系统的访问速度快5~20倍。

(3) **输入 / 输出带宽**。许多高性能向量处理机配备10~29个DMA通道。

(4) **通信带宽**。共享存储器或互连网络。

(5) **同步系统**。多流水线结构通过控制程序使所有流水线能够同步工作。

Cray-1系统采用流水线互锁来控制向量操作，
不冲突的操作可以并行地执行，
相关的操作尽可能链接起来重叠地进行。

(6) **多用途**。<http://www.wdklife.com> 五道口生活网
<http://www.wdklife.com/bbs> 五道口论坛

3、向量计算机系统结构的主要优点是：

- (1) 通过流水线存取方式有效地提高了存储器的带宽。
 - (2) 流水结构的运算器有很高的性能价格比。
 - (3) 非常简单的机制就能满足通信和同步的要求。
-
- 向量处理机通常以Mflops(Millin of Floating pointperconu)作为速度单位。
 - 一般认为，在标量计算机中，执行一次浮点运算需要 $2 \sim 5$ 条指令，平均需 3 条指令。