Tema III

Procesadores VLIW y procesadores vectoriales

3.2 Introducción

- El paralelismo funcional se obtiene mediante la replicación de las funciones de procesamiento que realiza el computador.
  - Granularidad fina → a nivel de instrucciones
  - Granularidad gruesa → a nivel de programas
- VLIW (very long instructions word – palabras de instrucción muy largas)
  - Se caracteriza por emitir en cada ciclo de reloj una única instrucción pero que contiene varias operaciones
  - La responsabilidad de planificar correctamente las instrucciones fuentes que se puedan codificar como VLIW son del compilador no el hardware
    - En tiempo de compilación se tiene mas tiempo para analizar todos los problemas

Figura 3.1: Evolución de la complejidad de los buffer de instrucciones, distribución, terminación y estaciones de reserva según el tipo de arquitectura.

3.3 El concepto arquitectónico VLIW

- En la planificación superescalar la planificación se realiza vía hardware (dinamica)
- VLIW la planificación es vía software (estática)
  - El compilador establece la secuencia paralela de instrucciones
  - Simplifica el hardware de los procesadores
  - Se emite una instrucción por ciclo
  - Una detención de una unidad funcional, implica la detención de todas la unidades funcionales
  - Causas del fracaso de VLIM
    - Incapacidad de desarrollar compiladores que aprovechen las características VLIW
      - Códigos de baja densidad con numerosas instrucciones NOP
    - Problemas de compatibilidad entre generaciones de procesadores VLIW
- Evolución -> EPIC Explicit Parallel Instruction Computing

Figura 3.1: Evolución de la complejidad de los buffer de instrucciones, distribución, terminación y estaciones de reserva según el tipo de arquitectura.

3.4 Arquitectura de un procesador VLIW genérico

- Ausencia de los elementos necesarios para la distribución, emisión y reordenamiento de instrucciones
- Los repertorios de instrucciones de las arquitecturas VLIW siguen una filosofía RISC con la excepción de que el tamaño de instrucción es mucho mayor ya que contienen múltiples operaciones o mini-instrucciones
- Una instrucción VLIW
  - Concatenación de varias instrucciones RISC que se pueden ejecutar en paralelo.
- Las operaciones recogidas dentro de una instrucción VLIW no presentan dependencias de datos, de memoria y/o de control entre ellas

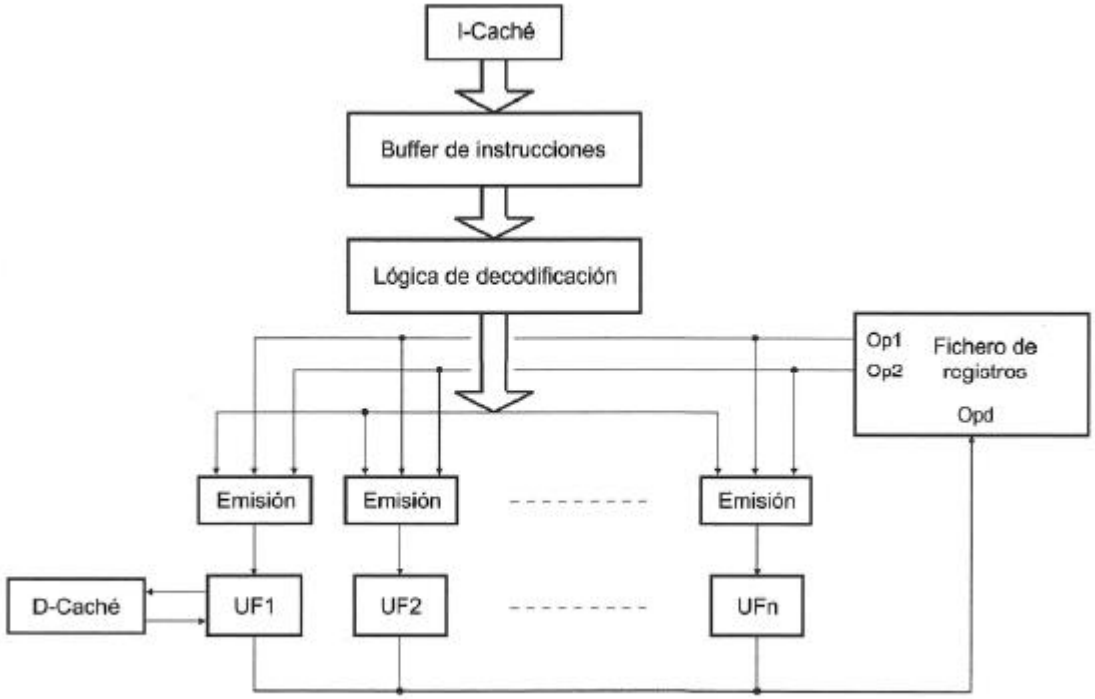


Figura 3.2: Arquitectura básica de un procesador VLIW genérico.

- El fichero de registro debe disponer de suficientes puertos de lectura para suministrar los operandos a todas las unidades funcionales en un único ciclo de reloj
- Una instrucción VLIW equivale a una concatenación de varias instrucciones RISC que se pueden ejecutar en paralelo
- El número y tipos de instrucciones corresponde con el número y tipos de unidades funcionales del procesador
- Debido a la ausencia de hardware para la planificación, los procesadores VLIW no detienen las unidades funcionales en espera de resultados.
- No existen interbloqueos por dependencias de datos ni hardware para detectarlas ya que el compilador se encarga de generar el código objeto para evitar estas situaciones.
  - Para ello recurre a la inserción de operaciones NOP.
- Problema para encontrar instrucciones independientes para rellenar todos los slots
  - El código de un procesador VLIW es mayor que uno superescalar
  - No aprovecha al máximo los recursos del procesador ya que tiene unidades funcionales ociosas
- Predecir qué accesos a memoria producirán fallos de caché es muy complicado.
  - Esto condiciona a que las memorias caché sean bloqueantes (tienen la capacidad de poder detener todas las unidades funcionales).

Inst. enteras 1 ciclo
Instr. Coma flotante 2 ciclos

Teniendo en cuenta las dependencias RAW, el código que generaría el compilador es el siguiente:

Código ejemplo

```
i1: ADD    R1,R2,R3
i2: MULTD  F1,F2,F3
i3: MULT   R4,R1,R1
i4: DIVD   F4,F1,F1
i5: ADD    R5,R1,R1
```

Instrucciones VLIW

```
I1: i1 + i2
I2: i3 + NOP
I3: i5 + i4
```

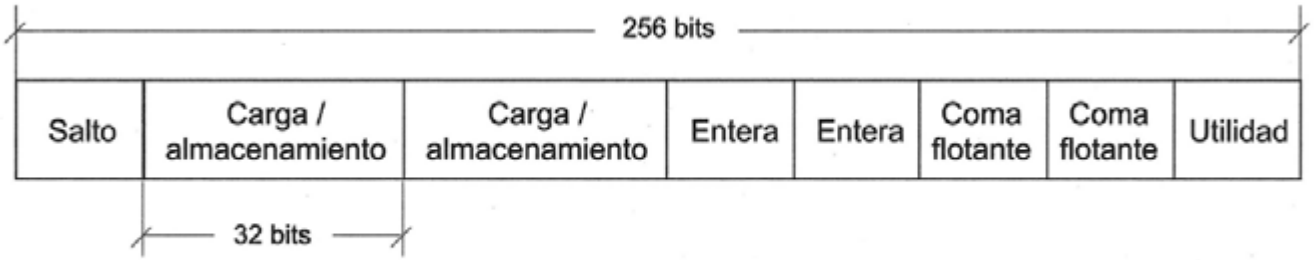


Figura 3.3: Formato de instrucción del computador Multiflow Trace 7.

3.5 Planificación estática o basada en el compilador

- El compilador VLIW recibe como entrada el código fuente de una aplicación, realiza unas tareas encaminadas a optimizar el código
- Produce tres elementos:
  - Un código intermedio
  - Un grafo del flujo de control
  - Un grafo del flujo de datos
  - **Código intermedio**
    - Formado por sencillas instrucciones RISC
    - Las únicas dependencias de datos que permanecen en él son las verdaderas, las RAW.
      - Las WAW y las WAR se han eliminado
  - Para poder generar un grafo de flujo de control es necesario conocer los bloques básicos de que consta el código intermedio.
  - **Bloque básico**
    - Un conjunto de instrucciones que conforman una ejecución secuencial
      - No hay instrucciones de salto salvo la última
      - No hay puntos intermedios de entrada salida
    - Para obtener los bloques básicos se analiza el código teniendo en cuenta que
      - Comienzo de un bloque básico:
        - Una instrucción etiquetada o la siguiente a un salto
      - El bloque se compone desde la instrucción inicial hasta la siguiente de salto
      - Los bloque se numeran de forma secuencial
  - Una vez que se conocen los bloques básicos que hay en el programa, las instrucciones de cada bloque se combinan para formar instrucciones VLIW.
    - Para ello se recurre al grafo de flujo de datos que tiene asociado cada bloque.

Diagrama de flujo de control y bloques básicos

Un grafo de flujo de control
Las interconexiones de los bloque básicos, atendiendo las posibles direcciones que pueda seguir la ejecución del código

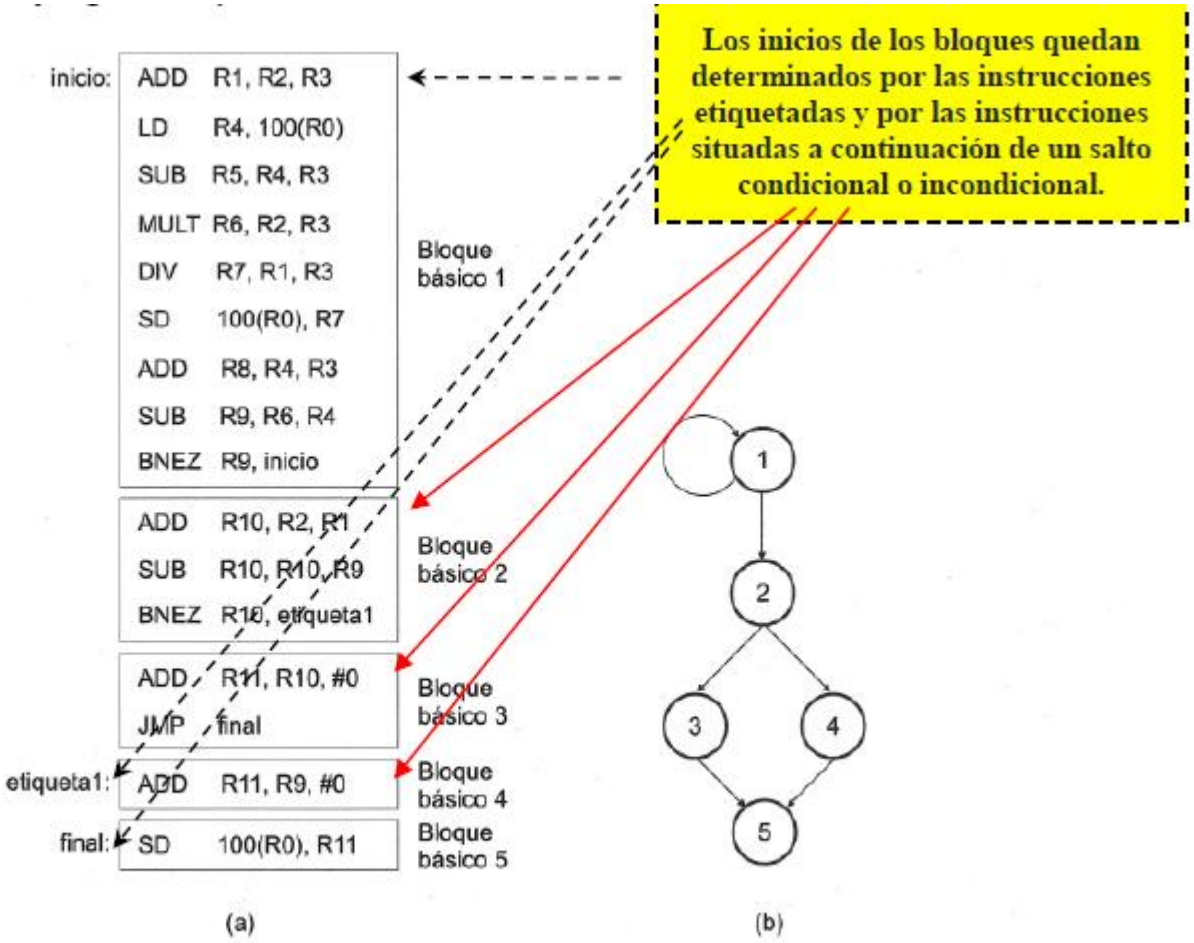


Figura 3.4: Ejemplo de secuencia de código intermedio con delimitación de los bloques básicos (a) y diagrama de flujo de control en el que se muestra la secuencia de ejecución de los bloques y la existencia de bucles (b).

Grafo de flujo de datos

- Es un grafo dirigido
- Los nodos son las instrucciones del bloque básico
- Los arcos se inician en la instrucción de escritura en un registro (Instrucción productora) y tiene por destino la instrucción que lee ese registro (instrucción consumidora)
- El grafo de flujo de datos muestra las secuencias de instrucciones que no presentan dependencias entre ellas y, por lo tanto, son susceptibles de combinar para formar instrucciones VLIW.
  - A la combinación de instrucciones de un único bloque básico para producir instrucciones VLIW se le denomina planificación local.
  - Planificación local
    - Combinación de instrucciones de un solo bloque básico
    - Está limitada, (tamaño bloque básico 5 0 6 instrucciones)
    - Técnicas
      - Desenrollamiento de bucles
      - La segmentación software
  - Planificación global
    - Combinar instrucciones de diferentes bloques básicos con el fin de producir una planificación con mayor grado de paralelismo

Diagrama de flujo de datos

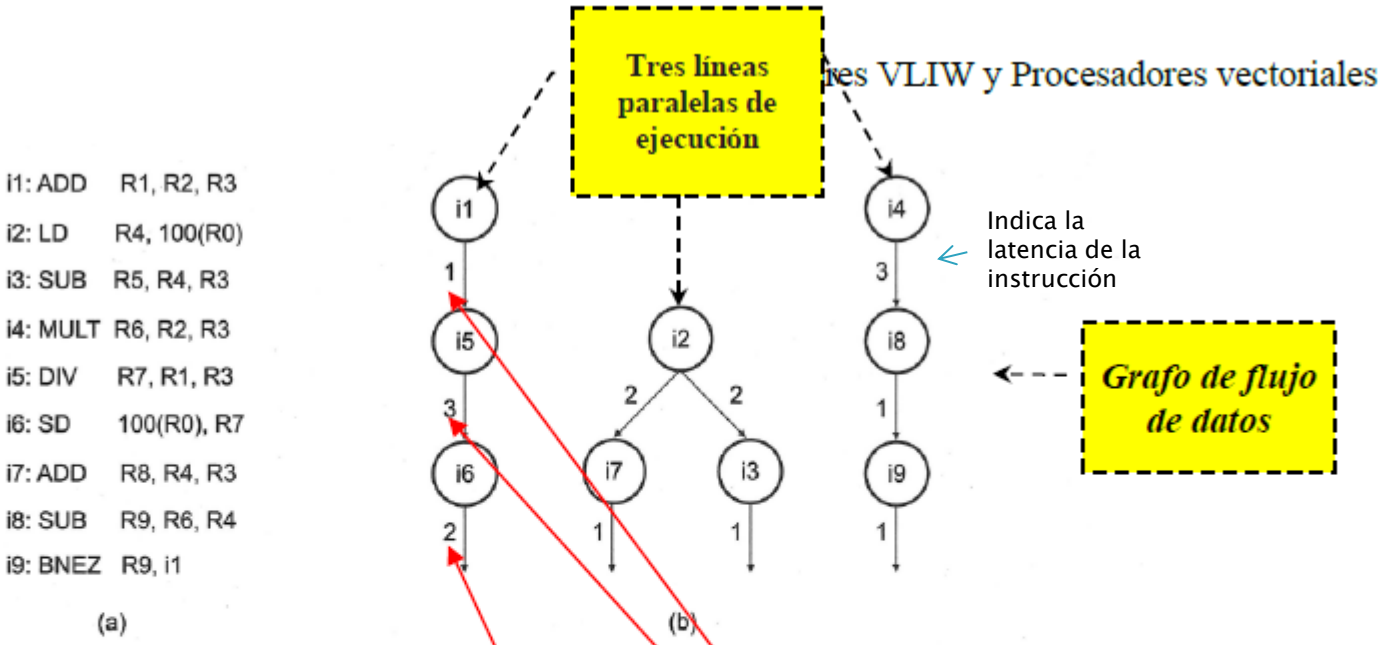


Figura 3.5: Secuencia de operaciones de un bloque básico (a) y diagrama de flujo de datos de la secuencia de instrucciones (b).

Consideraciones sobre formato de instrucción → admite

- Dos operaciones de suma/resta (un ciclo de latencia)
- Una operación de multiplicación/división (tres ciclos de latencia)
- Una operación de carga (dos ciclos de latencia)
- Una de almacenamiento (dos ciclos de latencia)
- Las instrucciones de salto son consideradas como instrucciones de suma/resta que escriben en el registro contador de programa (un ciclo de latencia)

Secuencia planificada de instrucciones VLIW

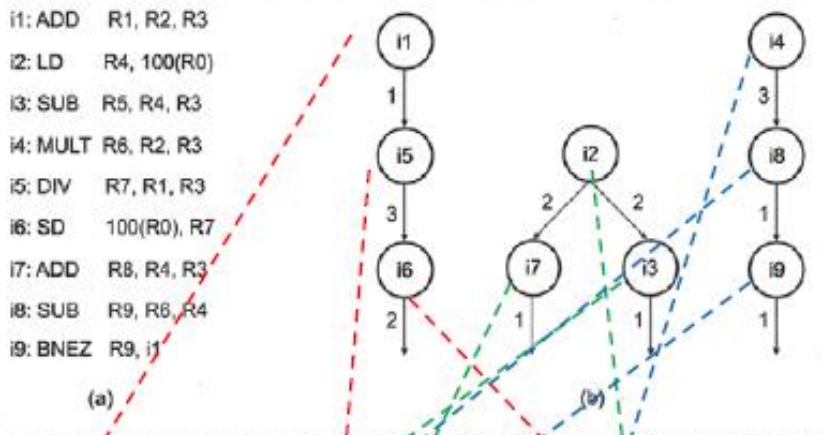


Figura 3.5: Secuencia de operaciones de un bloque básico (a) y diagrama de flujo de datos de la secuencia de instrucciones (b).

	Operaciones Suma / Resta 1	Operaciones Suma / Resta 2	Operaciones Mult / Div	Operaciones de carga	Operaciones de almacenamiento
inicio:	ADD R1, R2, R3	------	MULT R6, R2, R3	LD R4, 100(R0)	------
	------	------	DIV R7, R1, R3	------	------
	SUB R5, R4, R3	ADD R8, R4, R3	------	------	------
	SUB R9, R6, R4	------	------	------	------
	BNEZ R9, inicio	------	------	------	SD 100(R0), R7

Figura 3.6: Codificación de las operaciones del bloque básico en instrucciones VLIW.

- Si se considera que las instrucciones VLIW tienen una longitud de 20 bytes. Cada operación básica ocupa 4 bytes.
- El desaprovechamiento del código VLIW es del 64 %
- El programa consume 100 bytes de almacenamiento en memoria pero solo un 36 % está ocupado por operaciones útiles.

3.6 Desarrollamiento de bucles (loop unrolling)

- Aprovecha el paralelismo existente entre las instrucciones de un bucle.
- Replicamos múltiples veces el cuerpo del bucle utilizando diferentes registros en cada réplica y ajustar el código de terminación en función de las veces que se replique el cuerpo.
- Ventajas
  - Reduce el número de iteraciones del bucle
  - El total de instrucciones ejecutadas es menor ya se eliminan saltos y se reduce el número de instrucciones de incremento/decremento de los índices que se utilicen en el bucle.
  - Se proporciona al compilador un mayor número de oportunidades para planificar las instrucciones ya que al desenrollar el bucle queda mucho más cálculo al descubierto al incrementarse el tamaño de los fragmentos de código lineal.
    - El bloque básico se compone por todas las instrucciones que hay desde la instrucción inicial hasta la siguiente instrucción de salto que se detecte
  - Planificación más efectiva
  - Generar código VLIW más compacto

Ejemplo

```
inicio: LD     F0,0(R1)
        ADDD   F4,F0,F2
        SD     0(R1),F4
        SUBI   R1,R1,#8
        BNEZ   R1,inicio
```

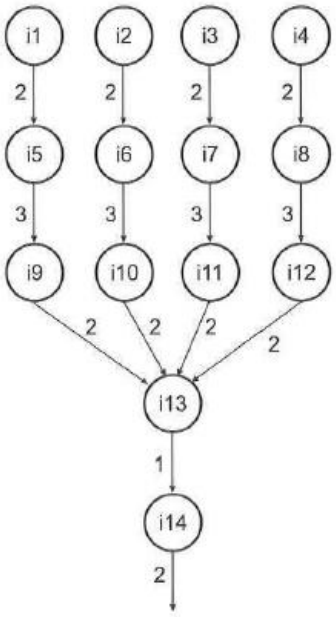
El código representa un bucle en el que se realiza la suma de una constante, almacenada en el registro F2, a todos los elementos de un vector almacenado en memoria cuyos elementos son de doble precisión, es decir, tienen una longitud de 8 bytes. El índice que permite acceder a los elementos del bucle se almacena en el registro entero R1, que inicialmente contiene la posición en memoria que ocupa el último elemento del vector.

Se desenrolla el cuerpo del bucle

```
inicio: LD     F0,0(R1)      % Iteración i
        ADDD   F4,F0,F2
        SD     0(R1),F4
        LD     F6,-8(R1)     % Iteración i+1
        ADDD   F8,F6,F2
        SD     -8(R1),F8
        LD     F10,-16(R1)   % Iteración i+2
        ADDD   F12,F10,F2
        SD     -16(R1),F12
        LD     F14,-24(R1)   % Iteración i+3
        ADDD   F16,F14,F2
        SD     -24(R1),F16
        SUBI   R1,R1,#32     % Decremento en 4 elementos
        BNEZ   R1,inicio
```

Reordenamiento de las instrucciones

```
inicio: LD     F0,0(R1)        % i1
        LD     F6,-8(R1)       % i2
        LD     F10,-16(R1)     % i3
        LD     F14,-24(R1)     % i4
        ADDD   F4,F0,F2        % i5
        ADDD   F8,F6,F2        % i6
        ADDD   F12,F10,F2      % i7
        ADDD   F16,F14,F2      % i8
        SD     0(R1),F4        % i9
        SD     -8(R1),F8       % i10
        SD     -16(R1),F12     % i11
        SD     -24(R1),F16     % i12
        SUBI   R1,R1 #32       % i13
        BNEZ   R1,inicio       % i14
```



Decremento adelantado

3 Unidades funcionales:

- Operaciones enteras (un ciclo de latencia)
- Operaciones en coma flotante (tres ciclos de latencia)
- Operaciones de carga/almacenamiento (dos ciclos de latencia).
- Las operaciones de salto se ejecutan en la unidad entera con un hueco de retardo de un ciclo por lo que permiten la planificación de una instrucción a continuación.

	Operaciones Carga / Almacenamiento	Operaciones Coma flotante	Operaciones Enteras y Saltos
inicio:	LD F0, 0(R1)	------	------
	LD F6, -8(R1)	------	------
	LD F10, -16(R1)	ADDD F4, F0, F2	------
	LD F14, -24(R1)	ADDD F8, F6, F2	SUBI R1, R1, #32
	------	ADDD F12, F10, F2	------
	SD 32(R1), F4	ADDD F16, F14, F2	------
	SD 24(R1), F8	------	------
	SD 16(R1), F12	------	------
	SD 8(R1), F16	------	BNEZ R1, inicio

Consideraciones

- La suma en coma flotante no se inician en el primer ciclo sino en el tercero ya que es necesario tener en cuenta el retardo asociado a las instrucciones de carga, esto es, dos ciclos.
- Las instrucciones de almacenamiento que deben esperar a que los resultados de las operaciones de suma en coma flotante estén disponibles.
- La dependencia WAR existente entre la lectura del registro R1 por las instrucciones de almacenamiento y su escritura por la instrucción SUBI se ha resuelto teniendo en cuenta el efecto que produce el decremento adelantado del índice.
  - Las cuatro instrucciones de almacenamiento se modifican para recoger el decremento adelantado del registro R1:
    - Como se decrementa por adelantado en 32, se suma un valor de 32 a los desplazamientos de los almacenamientos, 0, -8, -16 y -24, dando como resultado que el adelanto en la escritura de R1 provoque que los nuevos desplazamientos tengan que pasar a ser 32, 24, 16 y 8.

Rendimiento

- El tamaño del código VLIW es mayor
  - Si la instrucción VLIW y cada operación escalar necesitan un tamaño de 12 y 4 bytes, respectivamente, el espacio de almacenamiento desaprovechado es, aproximadamente, del 50%.
    - Las instrucciones VLIW → 9 instrucciones*12 bytes = 108 bytes
    - Operaciones originales → 14 instrucciones*4 bytes = 56 bytes
    - Bucle original → 5 instrucciones*4 bytes = 20 bytes
- Mejora es en el rendimiento.
  - Si el vector constase de 1000 elementos
  - Bucle original sin aplicar NADA 1000 veces:
    - 1000 iteraciones*5 instrucciones = 5000 ciclos
    - En el mejor de los casos, supuesta una segmentación ideal y sin riesgos.
  - El cuerpo del bucle desenrollado cuatro veces :
    - 250 iteraciones*14 instrucciones= 3500 ciclos.
  - El VLIW
    - 250 iteraciones*9 instrucciones = 2250 ciclos

Comparación con la versión VLIW que se obtendría del bucle original sin recurrir a ninguna técnica de planificación local

- 6 * 12 (bytes/instrucción) = 72 bytes, de los cuales solo 20 bytes están ocupados con operaciones.
- Velocidad de ejecución
  - Si el vector constase de 1000 elementos se tardaría en procesarlo 6000 ciclos de reloj frente a los 2250 ciclos obtenido tras aplicar desenrollamiento.

$$\frac{6000 - 2250}{2250} = 166{,}66\%\ \textit{más rápido}$$

	Operaciones Carga / Almacenamiento	Operaciones Coma flotante	Operaciones Enteras y Saltos
inicio:	LD F0, 0(R1)	------	------
	------	------	------
	------	ADDD F4, F0, F2	------
	------	------	------
	------	------	SUBI R1, R1, #8
	SD 8(R1), F4	------	BNEZ R1, inicio

Figura 3.8: Instrucciones VLIW generadas a partir de la secuencia original del bucle sin aplicar ninguna técnica de planificación local.

Reordenamiento de las instrucciones

- Mejora el paralelismo
- Mejora el rendimiento
- El código ocupa mas

3.7 Segmentación software

- Intenta aprovechar al máximo el paralelismo existente dentro del bucle
- Consiste en producir un nuevo cuerpo del bucle compuesto por la intercalación de instrucciones correspondientes a diferentes iteraciones del bucle original (*planificación policíclica*)
- Al reorganizar un bucle mediante segmentación software, siempre es necesario añadir unas instrucciones de arranque (el prólogo) antes del cuerpo del bucle y otras de terminación tras su finalización (el epílogo).
- Instrucciones A, B, C y D → cada una un ciclo de reloj
  - Una dependencia RAW con la instrucción que la precede → D depende C, C depende B, B depende A.
  - Tras tres iteraciones del bucle original aparece un patrón de ejecución compuesto por instrucciones pertenecientes a cuatro iteraciones diferentes del bucle original y que ya no presentan dependencias RAW entre ellas.

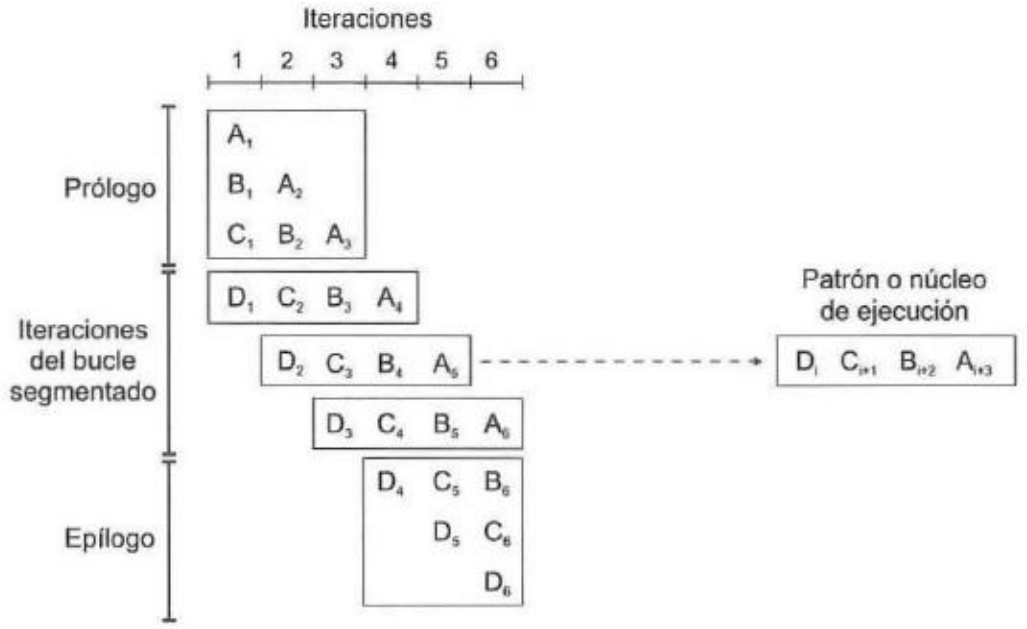


Figura 3.9: Aplicación de la segmentación software al cuerpo de un bucle genérico.

Ejemplo

```
inicio: LD     F0,0(R1)
        ADDD   F4,F0,F2
        SD     0(R1),F4
        SUBI   R1,R1,#8
        BNEZ   R1,inicio
```

Latencias

- Dos ciclos para los accesos a memoria.
- Tres ciclos para las operaciones en coma flotante.
- No se han incluido las instrucciones que decrementan el valor de R1 aunque sí se ha reflejado el necesario decremento en los desplazamientos de las instrucciones de carga y almacenamiento

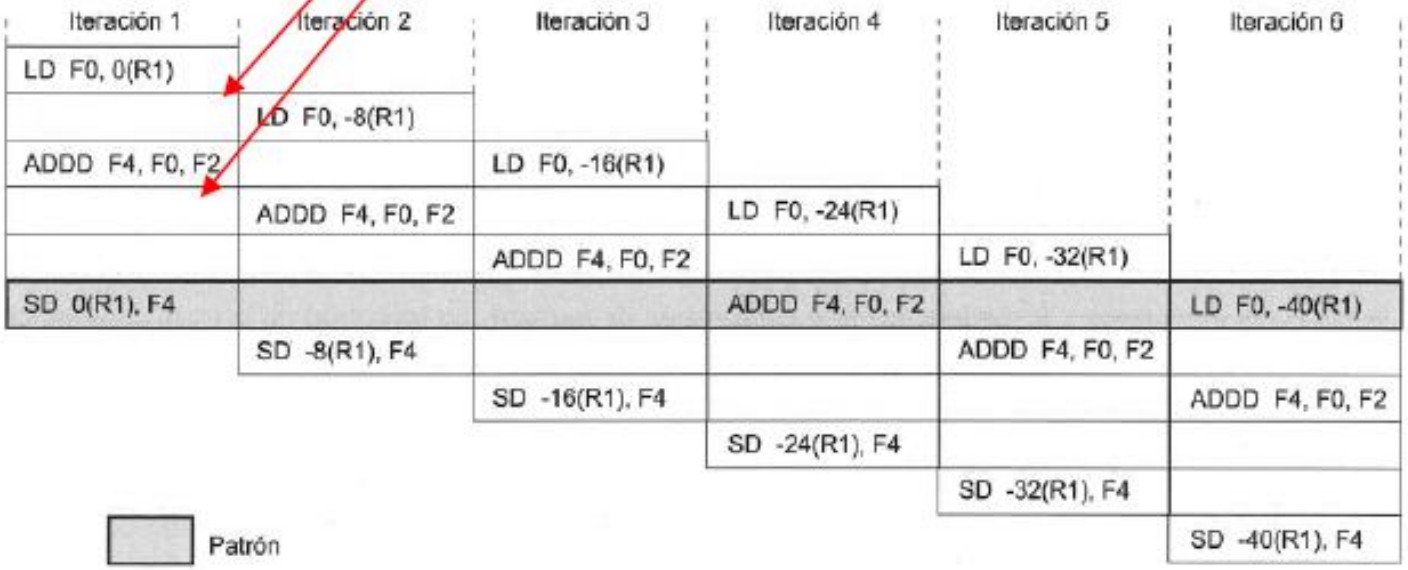


Figura 3.10: Esquema del patrón de ejecución obtenido al aplicar la técnica de segmentación software.

- Si las instrucciones VLIW son de 16 bytes, el tamaño total del código es de (11 inst.*16 byt/inst).= 176 bytes.
- Tiempo para procesar un vector de 1000 elementos:
- La aproximación VLIW emplearía 1010 ciclos.
  - 5 corresponderían al prólogo.
  - 5 al epílogo.
  - 1000 a las iteraciones del bucle.
- Aunque el concepto en que se basa es sencillo, la segmentación software puede llegar a ser extremadamente complicada de aplicar hay instrucciones condicionales en el cuerpo del bucle que impiden la aparición de un patrón de comportamiento regular.

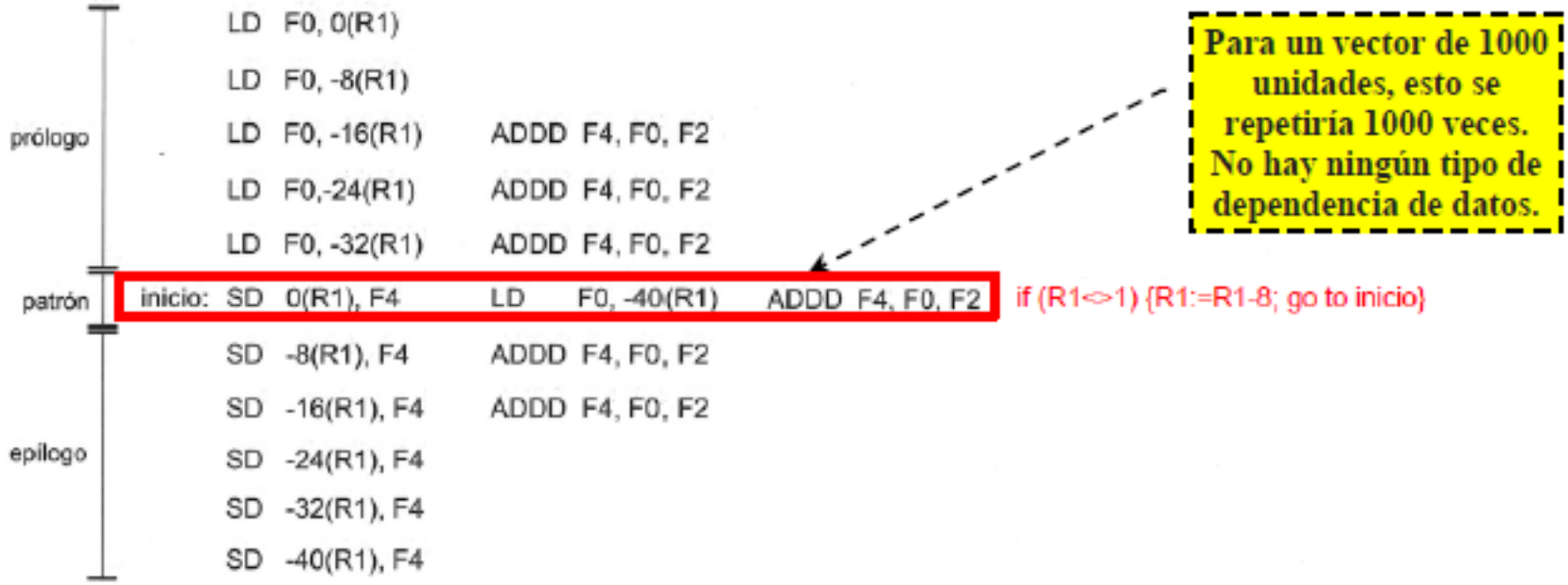


Figura 3.11: Instrucciones VLIW genéricas obtenidas a partir de la secuencia del bucle segmentado.

3.8 Planificación de Trazas (Trace Sheduling)

- Es una técnica de planificación global
- Traza
  - Camino de ejecución mas probable
- Pasos
  - 1.- Selección de la traza
    - Encontrar un conjunto de bloques básicos que conformen una secuencia de código sin bucle
      - Seleccionamos al que especulemos que será mas probable que se ejecute
        - Compilador utiliza un Grafos con pesos (ponderados)por distintos criterios perfiles de ejecución, estimaciones, planificación estática de saltos…
  - 2.- Compactación de la traza

Código intermedio del bucle sin compactar

```
for (i=0;i<n;i++)
   if (A[i]==0) then
      X[i]:=X[i]+a;
   else
      Y[i]:=Y[i]-a;
   end if;
end for;
```

- Se recorre un vector de números enteros, A, y dos vectores de valores en coma flotante, X e Y.
- En cada iteración del bucle, y en función del contenido de A[i] , se incrementa X[i] o se decrementa Y[i] con una constante almacenada en F2

8 bytes

4 bytes

A pesar de ello se hace con decremento

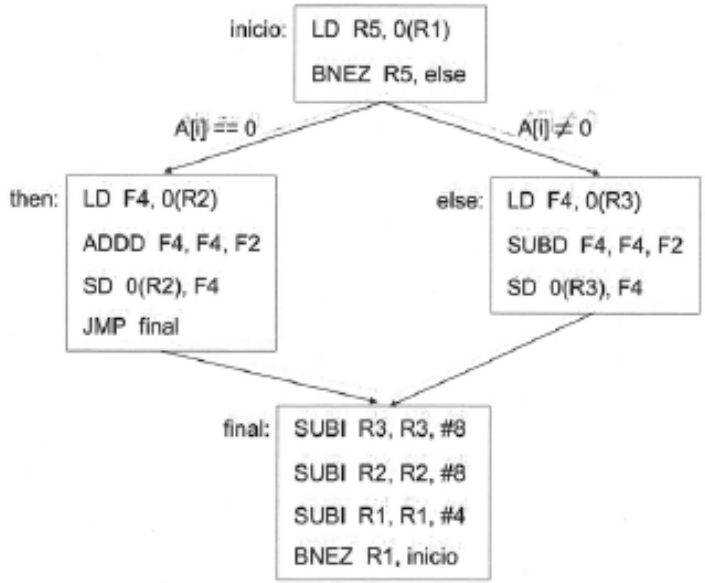
Código intermedio

R1 puntero a A[i]

R2 puntero a X[i]

R3 puntero a Y[i]

```
inicio: LD    R5, 0(R1)     % Cargar A[i]
        BNEZ  R5,else       % Si (A[i]<>0) ir a else
then:   LD    F4,0(R2)      % Cargar X[i]
        ADDD  F4,F4,F2      % X[i]:=X[i]+a
        SD    0(R2),F4      % Almacenar X[i]
        JMP   final
else:   LD    F4,0(R3)      % Cargar Y[i]
        SUBD  F4,F4,F2      % Y[i]:=Y[i]-a
        SD    0(R3),F4      % Almacenar Y[i]
final:  SUBI  R2,R2,#8      % Decrementar en 8 bytes
        SUBI  R3,R3,#8      % Decrementar en 8 bytes
        SUBI  R1,R1,#4      % Decrementar en 4 bytes
        BNEZ  R1,inicio     % Nueva iteración
```



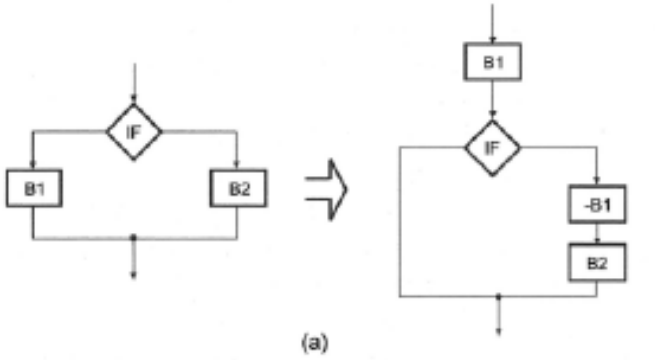
3 U.U.F.F. con ciclos de latencia:

	2 Operaciones Carga / Almacenamiento	3 Operaciones Coma flotante	1 Operaciones Enteras y Saltos	
inicio:	LD R5, 0(R1)	------	------	1
	------	------	------	2
	------	------	BNEZ R5, else	3
then:	LD F4, 0(R2)	------	SUBI R1, R1, #4	4
	------	------	------	5
	------	ADDD F4, F4, F2	------	6
	------	------	------	7
	------	------	------	8
	SD 0(R2), F4	------	JMP final	9
else:	LD F4, 0(R3)	------	SUBI R2, R2, #8	10
	------	------	------	11
	------	SUBD F4, F4, F2	------	12
	------	------	------	13
	------	------	------	14
	SD 0(R3), F4	------	------	15
final:	------	------	BNEZ R1, inicio	16
	------	------	SUBI R3, R3, #8	17

Ruta de ejecución más probable: (caso A[i] == 0) → 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 16, 17 → 12 ciclos

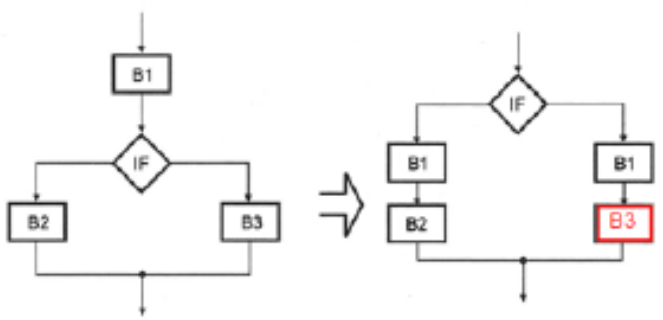
Ruta menos probable: (caso A[i] <> 0) → 1, 2, 3, 4, 10, 11, 12, 13, 14, 15, 16, 17 → 12 ciclos

Posibles situaciones en las que plantea desplazamiento de operaciones

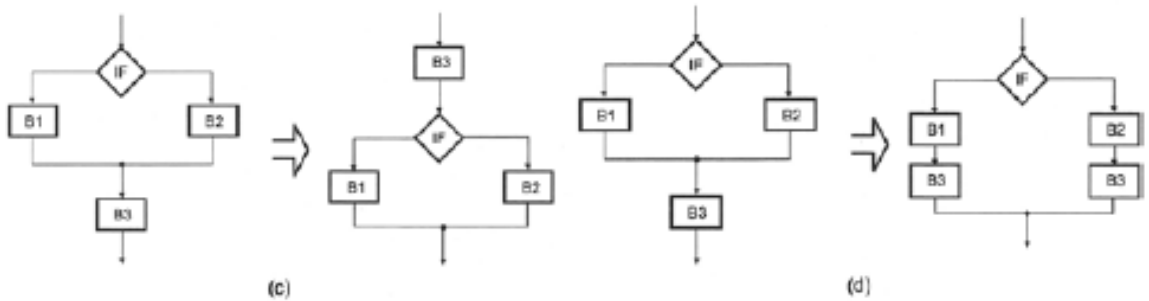


(a)

Se podrían desplazar antes de la condición lo que obligaría a anular sus efectos mediante un código de compensación en caso de que no fuese necesaria su ejecución. Esto se ha representado en la figura mediante el bloque etiquetado como -B1 que indica que es necesario realizar las operaciones que anulen el efecto de la ejecución del contenido de B 1.



(b)



(c) (d)

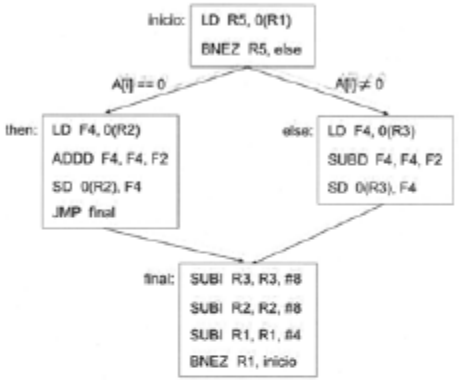
Situaciones en las que un bloque que se encuentra en un punto de ejecución común puede desplazarse con el objetivo de producir código más compacto.

Consideraciones del compilador para desplazar operaciones dentro de una traza

- Conocer cuál es la secuencia de ejecución más probable.
- Conocer las dependencias de datos existentes para garantizar su mantenimiento.
- La cantidad de código de compensación que es necesario añadir.
- Saber si compensa el desplazamiento de operaciones dentro de la traza, midiéndose el coste tanto en ciclos de ejecución como en espacio de almacenamiento.

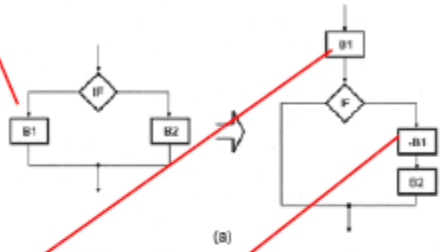
Ejemplo

```
for (i=0;i<n;i++)
   if (A[i]==0) then
      X[i]:=X[i]+a;
   else
      Y[i]:=Y[i]-a;
   end if;
end for;
```




```
inicio: LD    R5, 0(R1)    % Cargar A[i]
        BNEZ  R5,else      % Si (A[i]<>0) ir a else
then:   LD    F4,0(R2)     % Cargar X[i]
        ADDD  F4,F4,F2     % X[i]:=X[i]+a
        SD    0(R2),F4     % Almacenar X[i]
        JMP   final
else:   LD    F4,0(R3)     % Cargar Y[i]
        SUBD  F4,F4,F2     % Y[i]:=Y[i]-a
        SD    0(R3),F4     % Almacenar Y[i]
final:  SUBI  R2,R2,#8     % Decrementar en 8 bytes
        SUBI  R3,R3,#8     % Decrementar en 8 bytes
        SUBI  R1,R1,#4     % Decrementar en 4 bytes
        BNEZ  R1,inicio    % Nueva iteración
```

Aplicando el desplazamiento de operaciones del bloque B1




```
inicio: LD    F4,0(R2)     % Desplazamiento: Cargar X[i]
        ADDD  F4,F4,F2     % Desplazamiento: X[i]:=X[i]+a
        SD    0(R2),F4     % Desplazamiento: Almacenar X[i]
        LD    R5,0(R1)     % Cargar A[i]
        BNEZ  R5,else      % Si (A[i]<>0) ir a else
then:   JMP   final
else:   LD    F4,0(R2)     % Compensación: Cargar X[i]
        SUBD  F4,F4,F2     % Compensación: X[i]:=X[i]-a
        SD    0(R2),F4     % Compensación: Almacenar X[i]
        LD    F4,0(R3)     % Cargar Y[i]
        SUBD  F4,F4,F2     % Y[i]:=Y[i]-a
        SD    0(R3),F4     % Almacenar Y[i]
final:  SUBI  R2,R2,#8     % Decrementar índice de X
        SUBI  R3,R3,#8     % Decrementar índice de Y
        SUBI  R1,R1,#4     % Decrementar índice de A
        BNEZ  R1,inicio    % Nueva iteración
```

	Operaciones Carga / Almacenamiento	Operaciones Coma flotante	Operaciones Enteras y Saltos	
inicio:	LD F4, 0(R2)	------	------	1
	LD R5, 0(R1)	------	------	2
	------	ADDD F4, F4, F2	------	3
	------	------	------	4
	------	------	BNEZ R5, else	5
	SD 0(R2), F4	------	SUBI R1, R1, #4	6
then:	------	------	JMP final	7
else:	LD F4, 0(R2)	------	SUBI R2, R2, #8	8
	LD F4, 0(R3)	------	------	9
	------	SUBD F4, F4, F2	------	10
	------	SUBD F4, F4, F2	------	11
	------	------	------	12
	SD 8(R2), F4	------	------	13
	SD 0(R3), F4	------	------	14
final:	------	------	BNEZ R1, inicio	15
	------	------	SUBI R3, R3, #8	16

Figura 3.14: Código VLIW generado tras la planificación de traza.

- Ruta de ejecución más probable: 1, 2, 3, 4, 5, 6, 7, 8, 15, 16
  - 10 ciclos decremento de 2 ciclos respecto a la no planificación
- Ruta menos probable: 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16
  - 15 ciclos Incremento de 3 ciclos
- Aunque la reducción no es muy elevada, el tiempo medio de ejecución del código planificado será mejor que el código original siempre que se cumpla la siguiente expresión:
  - [10 ciclos * p + 15 ciclos * (1 – p) < 12 ciclos * p + 12 ciclos * (1 – p)]
    - p es la probabilidad de que la rama (A[i] ==0) sea la ejecutada.
- Uno de los problemas que presenta la planificación de trazas:
  - A mayor cantidad de operaciones desplazadas, mayor es la penalización en que se incurre cuando se re

	Operaciones Carga / Almacenamiento	Operaciones Coma flotante	Operaciones Enteras y Saltos	
inicio:	LD F4, 0(R2)	------	------	1
	LD R5, 0(R1)	------	------	2
	------	ADDD F4, F4, F2	------	3
	------	------	------	4
	------	------	BNEZ R5, else	5
	SD 0(R2), F4	------	SUBI R1, R1, #4	6
then:	------	------	JMP final	7
else:	LD F4, 0(R2)	------	SUBI R2, R2, #8	8
	LD F4, 0(R3)	------	------	9
	------	SUBD F4, F4, F2	------	10
	------	SUBD F4, F4, F2	------	11
	------	------	------	12
	SD 8(R2), F4	------	------	13
	SD 0(R3), F4	------	------	14
final:	------	------	BNEZ R1, inicio	15
	------	------	SUBI R3, R3, #8	16

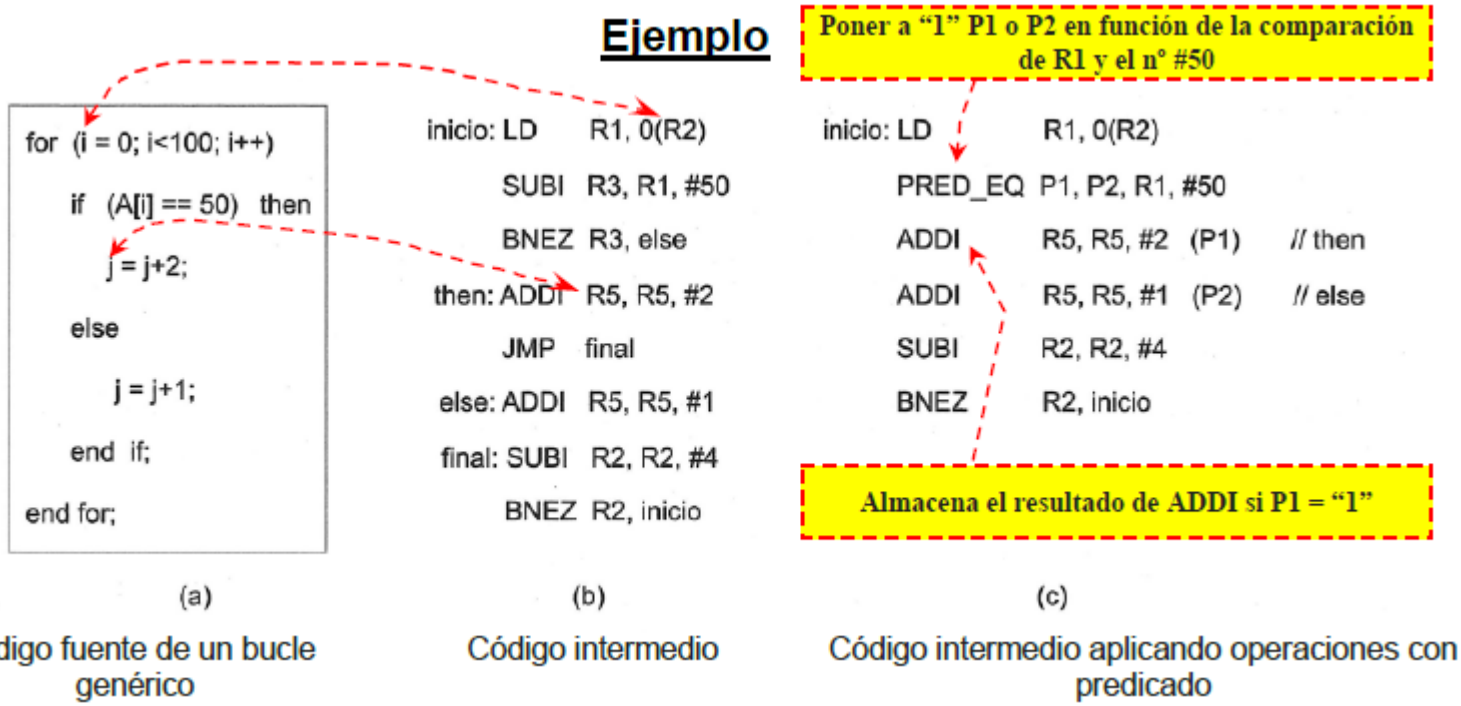
Figura 3.14: Código VLIW generado tras la planificación de traza.

	Operaciones Carga / Almacenamiento	Operaciones Coma flotante	Operaciones Enteras y Saltos	
inicio:	LD F4, 0(R2)	------	------	1
	LD R5, 0(R1)	------	------	2
	LD F6, 0(R3)	ADDD F4, F4, F2	------	3
	------	------	BNEZ R5, else	4
	------	SUBD F6, F6, F2	SUBI R1, R1, #4	5
then:	SD 0(R2), F4	------	JMP final	6
else:	------	------	SUBI R2, R2, #8	7
	------	------	------	8
	SD 0(R3), F6	------	------	9
final:	------	------	BNEZ R1, inicio	10
	------	------	SUBI R3, R3, #8	11

Figura 3.15: Código VLIW con planificación óptima.

3.9 Operaciones con predicado

- Operación con predicado es una instrucción en la que su resultado se almacena o se descarta en función de un operando que tiene asociado
  - Se implementa como un registro de un bit que se añade como un nuevo operando de lectura en cada una de las operaciones que conforman una instrucción VLIW
    - p = 1 → resultado de la operación se almacena
    - p = 0 → resultado de la operación se anula
  - Es efectiva si todas las rutas de una región tiene el mismo tamaño y la misma frecuencia de ejecución



	2 ciclos Operaciones Carga / Almacenamiento	3 ciclos Operaciones Coma flotante	1 ciclo Operaciones Enteras y Saltos	
inicio:	LD R1, 0(R2)	------	------	1
	------	------	------	2
	------	------	SUBI R3, R1, #50	3
	------	------	BNEZ R3, else	4
	------	------	------	5
then:	------	------	ADDI R5, R5, #2	6
	------	------	JMP final	7
	------	------	------	8
else:	------	------	ADDI R5, R5, #1	9
final:	------	------	SUBI R2, R2, #4	10
	------	------	BNEZ R2, inicio	11
	------	------	------	12

(a)

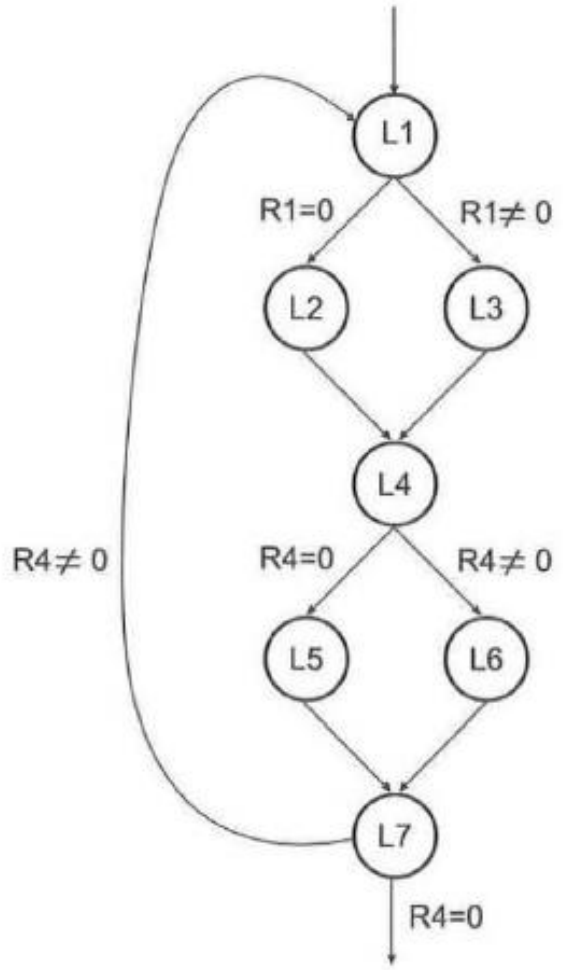
Instrucciones VLIW derivadas del código intermedio

inicio:	LD R1, 0(R2)	------	------	1
	------	------	------	2
	------	------	PRED_EQ P1, P2, R1, #50	3
	------	------	ADDI R5, R5, #2 (P1)	4
	------	------	ADDI R5, R5, #1 (P2)	5
	------	------	SUBI R2, R2, #4	6
	------	------	BNEZ R2, inicio	7
	------	------	------	8

(b)

Instrucciones VLIW derivadas del código intermedio con operaciones condicionadas

- VLIW sin operaciones condicionadas 11 y 9 ciclos según rama.
- VLIW con operaciones condicionadas 8 ciclos.



(a)

```
L1: LD    R1, 0(R5)
    BNEZ  R1, L3
L2: ADDI  R2, R1, #1
    LD    R3, 0(R6)
    LD    R4, 0(R7)
    ADD   R4, R4, R2
    JMP   L4
L3: ADDI  R4, R4, #1
    ADDI  R2, R1, #1
L4: BNEZ  R4, L6
L5: ADDI  R4, R4, #1
    SD    0(R4), R2
    JMP   L7
L6: SUBI  R4, R4, #1
    SD    0(R4), R2
L7: BNEZ  R4, L1
```

(b)

```
L1: LD       R1, 0(R5)
    PRED_NE  P1, P2, R1, #0
    ADDI     R2, R1, #1   (P2)
    LD       R3, 0(R6)    (P2)
    LD       R4, 0(R7)    (P2)
    ADD      R4, R4, R2   (P2)
    ADDI     R4, R4, #1   (P1)
    ADDI     R2, R1, #1   (P1)
    PRED_NE  P3, P4, R4, #0
    ADDI     R4, R4, #1   (P4)
    SD       0(R4), R2    (P4)
    SUBI     R4, R4, #1   (P3)
    SD       0(R4), R2    (P3)
    BNEZ     R4, L1
```

(c)

Figura 3.18: Ejemplo de agrupamiento de bloques básicos mediante la transformación de estructuras *if-then* en operaciones con predicado.

3.10 Tratamiento de Excepciones

- Las técnicas vistas hasta ahora se basan en la predicción de los resultados de los saltos condicionales.
  - Habrá que establecer mecanismos para el tratamiento de las excepciones (interrupciones).
- Centinelas
  - Es en fragmento de código que indica que la operación ejecutada de forma especulativa con la que está relacionado ha dejado de serlo.
  - El compilador marca las operaciones especulativas con un a etiqueta y en lugar del programa en el que estaba el código especulado que ha sido desplazado, sitúa un centinela vinculado a esa etiqueta
  - Esta estrategia se implementa mediante una especie de buffer de terminación en el que las instrucciones se retiran cuando les corresponde salvo las marcadas como especulativas, que se retirarán cuando lo señale la ejecución del centinela que tienen asociado

3.11 El enfoque EPIC

- Problemas del VLIW
  - Los repertorios de instrucciones VLIW no son compatibles entre diferentes implementaciones.
  - La planificación estática de las instrucciones de carga por parte del compilador resulte muy complicada.
  - La importancia de disponer de un compilador que garantice una planificación óptima del código de forma que se maximice el rendimiento del computador y se minimice el tamaño del código.
- EPIC (*Explicitly Parallel Instruction Computing*)
  - El objetivo de EPIC es retener la planificación estática del código pero mejorarla con características arquitectónicas que permitan hacer frente dinámicamente a diferentes situaciones, tales como retardos en las cargas o unidades funcionales nuevas o con diferentes latencias.
    - Es el compilador el que determina al agrupamiento de instrucciones pero, a la vez, comunica de forma explícita en el propio código cómo se ha realizado el agrupamiento.

Tabla 3.1: Diferencias principales entre las arquitecturas superescalar, VLIW y EPIC.

	Agrupamiento de operaciones	Asignación de unidad funcional	Secuencia de emisión a las unidades funcionales
Superescalar	Hardware	Hardware	Hardware
EPIC	Compilador	Hardware	Hardware
VLIW	Compilador	Compilador	Compilador

A :=B+C

```
LD     F2,0(R1)    % Carga de B en F2 desde M[0+R1]
LD     F4,0(R2)    % Carga de C en F4 desde M[0+R2]
ADDD   F6,F4,F2    % Suma en F6
SD     0(R3),F6    % Almacenamiento de A en M[0+R3]
```

Procesador VLIW	• Dos unidades de carga/almacenamiento (2 ciclos de latencia) • Una unidad funcional para operaciones en coma flotante (2 ciclos de latencia) • Una unidad para operaciones enteras y saltos (1 ciclo de latencia).

Código producido por el compilador

Operaciones Carga / Almacenamiento	Operaciones Carga / Almacenamiento	Operaciones Coma flotante	Operaciones Enteras
LD F2, 0(R1)	LD F4, 0(R2)	------	------
------	------	------	------
------	------	ADDD F6, F4, F2	------
------	------	------	------
SD 0(R3), F6	------	------	------

Código VLIW

De las 20 operaciones | 16 NOPs 4 Útiles

Se puede emitir en paralelo con ella la siguiente instrucción que aparezca en el código

```
[2] LD     F2, 0(R1)
[1] LD     F4, 0(R2)
[1] ADDD   F6, F4, F2
[1] SD     0(R3), F6
```

(b)

NO se puede emitir en paralelo con ella la siguiente instrucción que aparezca en el código

Código EPIC

- Características arquitectónicas más destacadas del estilo EPIC
  - Planificación estática con paralelismo explícito.
  - Operaciones con predicado.
  - Descomposición o factorización de las instrucciones de salto condicional.
  - Especulación de control.
  - Especulación de datos.
  - Control de la jerarquía de memoria.

3.12 Procesadores vectoriales

- Arquitectura que permite la manipulación de vectores
  - Operadores fuente y destino son vectores
  - Operaciones para manejar vectores
  - Unidades funcionales para operar con vectores
  - Cálculo científico
  - Gran volumen de datos
  - Compilador encargado de detectar y extraer el paralelismo
- Una operación con vectores
  - En un ordenador escalar sería un bucle, en un ordenador vectorial una única instrucción

1. Inicialización de los índices.
2. Cálculo escalar (por ejemplo, una suma).
3. Actualización de los índices.
4. Comparación.
5. Salto a la instrucción 2.

Cálculo con un bucle

→ 1. Cálculo vectorial.

3.13 Arquitectura vectorial básica

- Basados en arquitectura carga-almacenamiento
  - Todos los operandos ubicados en los registros vectoriales
- Unidad de procesamiento escalar y unidad de procesamiento vectorial
  - El código objeto de un programa vectorizado se compone de una combinación de instrucciones escalares e instrucciones vectoriales
- MVL *Maximum Vector Length* – Máxima longitud del vector
  - Número de elementos por registro
- Las unidades funcionales vectoriales están segmentadas y pueden iniciar una operación en cada ciclo de reloj.
- Son necesarios mecanismos hardware que detecten los riesgos estructurales y los riesgos de datos que pueden aparecer entre las instrucciones vectoriales y escalares

Esquema de un procesador vectorial básico

- Características del fichero de registros vectoriales
  - El número de registros vectoriales.
    - 64 ÷ 256
  - El número de elementos por registro.
    - MVL (Maximum Vector Length- Máxima Longitud del Vector).
      - 8 ÷ 256
  - El tamaño en bytes de cada elemento.
    - 8 bytes

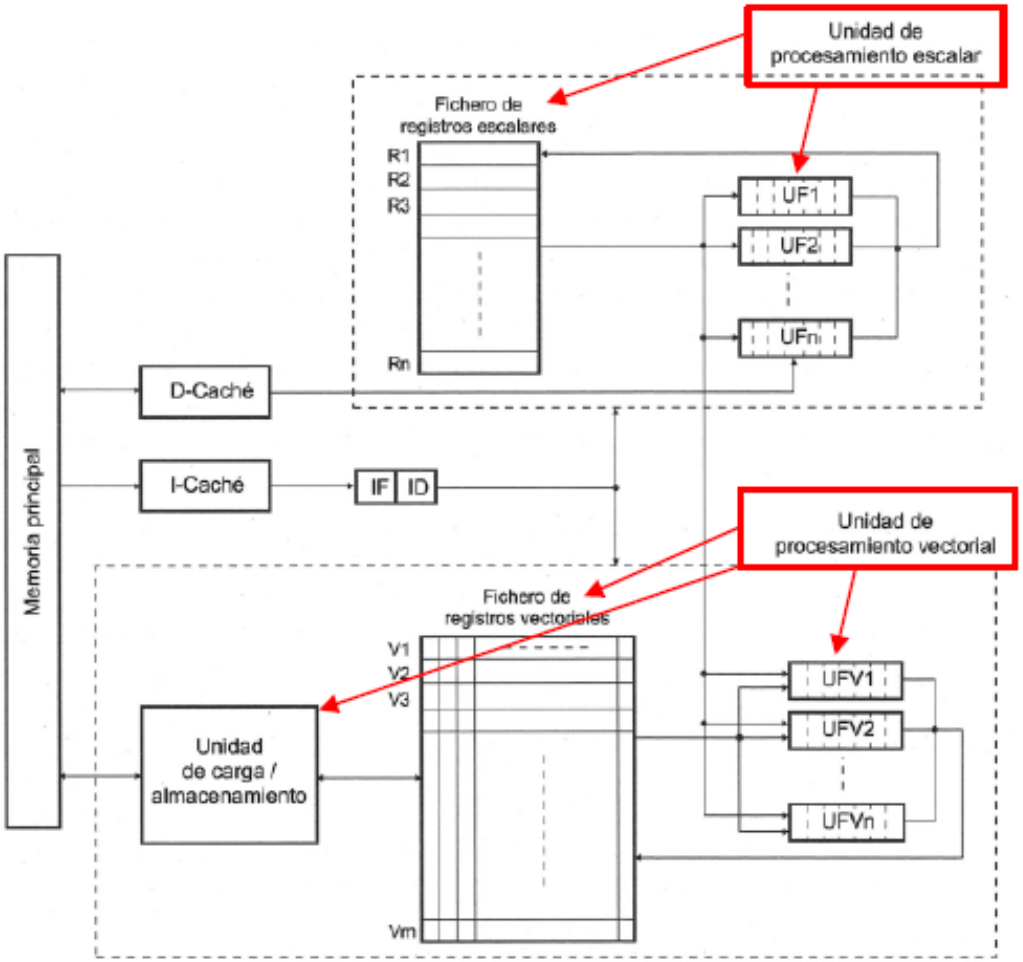


Figura 3.20: Esquema de un procesador vectorial básico.

Unidades funcionales con varios carriles (lanes)

- Las unidades funcionales vectoriales de algunos procesadores pueden estar internamente organizadas en varios ***lanes*** o carriles con el objeto de aumentar el número de operaciones que pueden procesar en paralelo por ciclo de reloj.
  - Una unidad funcional con n carriles implica que
    - Internamente, el hardware necesario para realizar la operación se ha replicado **n** veces de forma que el número de ciclos que se emplea en procesar un vector se reduce por un factor de n.
    - Hay que incrementar el número de puertos de lectura y escritura de todos los registros vectoriales para poder suministrar en cada ciclo de reloj elementos a los carriles

Ejemplo

C: =A+B A, B, C son registros vectoriales de 20 elementos.

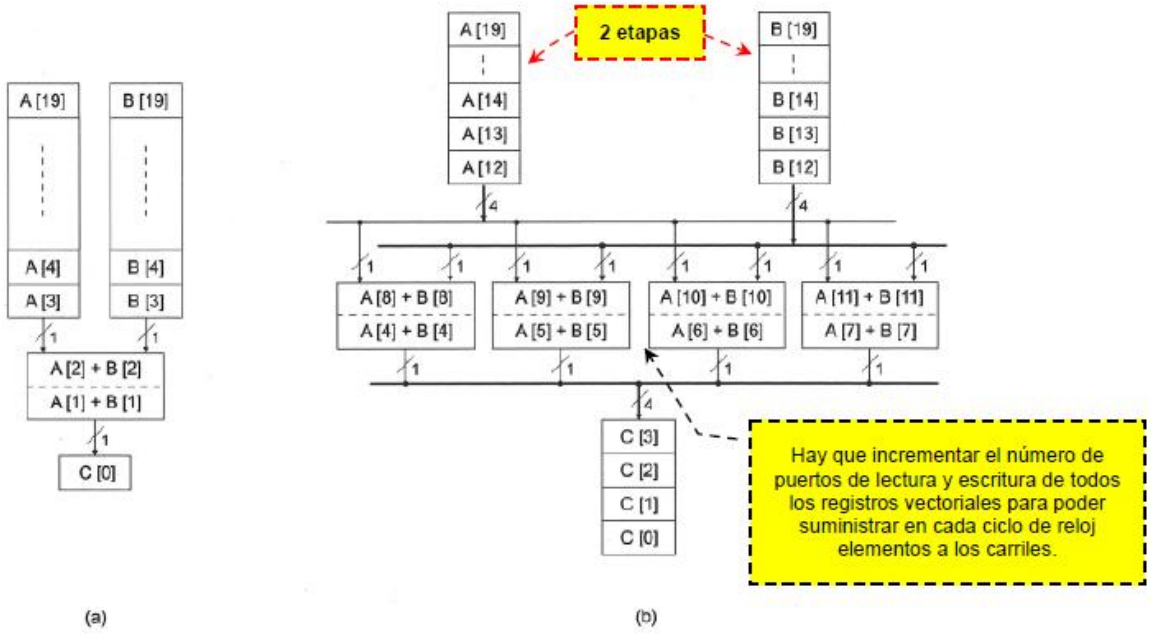


Figura 3.21: Diferencia entre realizar la suma de 2 vectores de 20 elementos en una unidad funcional con 2 etapas y un único carril (a) y en otra unidad funcional de 2 etapas pero con 4 carriles (b).

Unidades de cuatro carriles o lanes

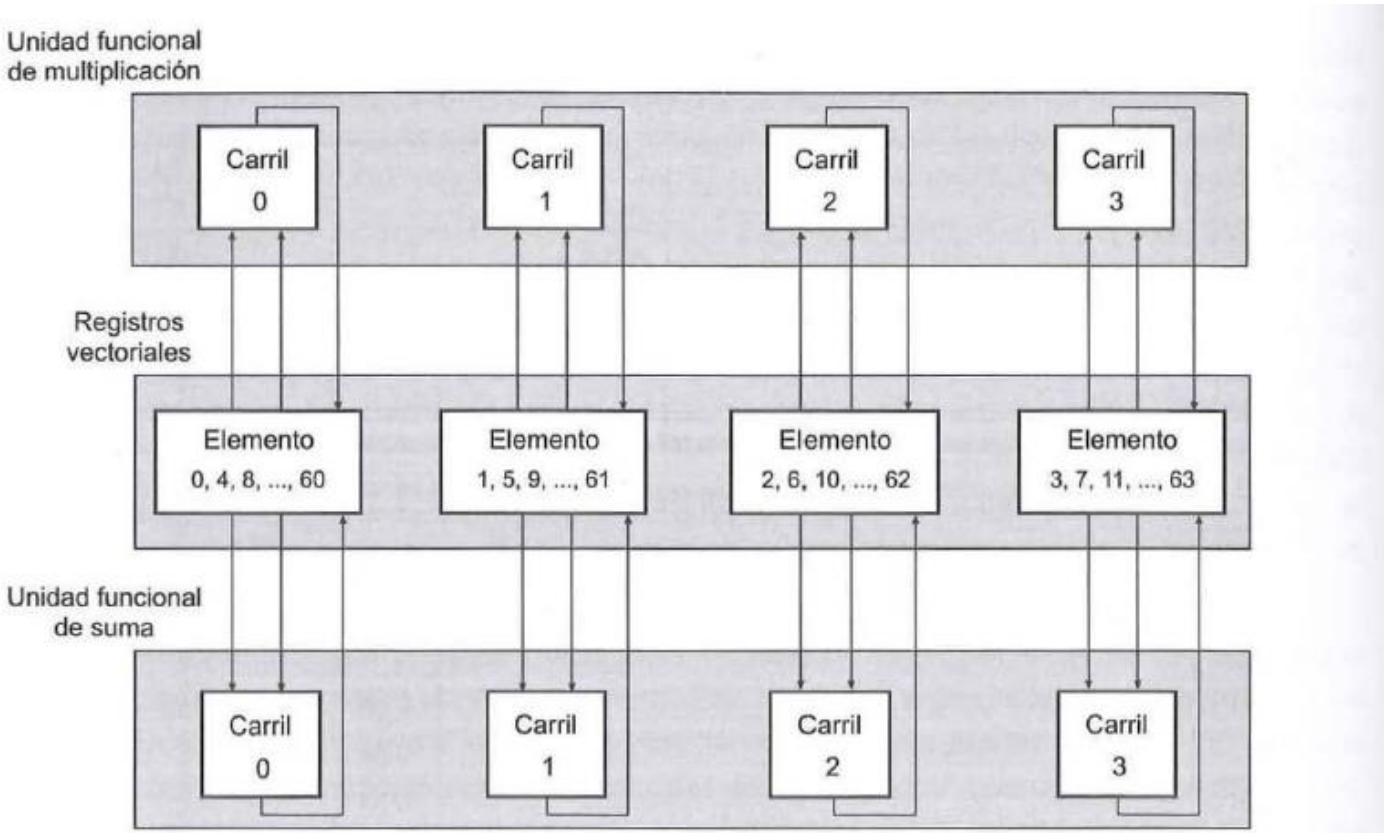


Figura 3.22: Esquema de unidades vectoriales de suma y multiplicación en coma flotante con cuatro carriles y distribución de los elementos de los registros vectoriales entre ellas.

Características procesadores vectoriales

Tabla 3.2: Características de los procesadores vectoriales de supercomputadores comercializados durante las últimas décadas. Se muestran las características de un único procesador vectorial aunque el computador pueda estar formando por varias unidades.

Procesador vectorial	Año anuncio	Reloj (Mhz.)	Total de registros vectoriales	Elementos de 8 bytes por registros	Unidades funcionales	Lanes por U.F.	Unidades de c/a
Cray-1	1976	80	8	64	6	1	1
Cray-2	1985	244	8	64	5	1	1
Cray Y-MP	1988	166	8	64	8	1	2 c., 1 a.
Cray C-90	1991	240	8	128	8	2	4
Convex C-4	1994	135	16	128	3	1	4
Cray T-90	1996	460	8	128	8	2	4
NEC SX/5	1998	312	8-64	512	4	16	1
Fujitsu VPP500	1999	300	8-256	4096-128	3	16	1 c., 1 a.
Cray SV1	1998	300	8	64	8	1	1+1 a.
Cray SV1ex	2001	500	8	64	8-2	2-8	1+1 a.
VMIPS	2001	500	8	64	5	1	1
NEC SX/6	2001	500	72	32	5	1	1
Cray X1	2002	800	32	64	4	2	?
NEC SX/8	2004	2200	72	64	5	1	1
NEC SX/9	2008	3300	72	64	6	2	1

Procesadores Matriciales

- Se engloban dentro de la categoría SIMD (*Single Instruction -Multiple Data*).
- Cuenta con una unidad de procesamiento vectorial, una unidad escalar y una unidad de control que discrimina según el tipo de instrucción.
- La principal diferencia con el vectorial
  - La unidad vectorial se compone de:
    - n elementos de procesamiento EP, constituidos por unidades aritmético-lógicas de propósito general,
    - Un conjunto de registros REP, y
    - Una memoria local MEP
  - El procesador matricial con n elementos de procesamiento puede procesar de forma simultánea en el mismo ciclo de reloj un vector de n elementos
- Hoy no se fabrican

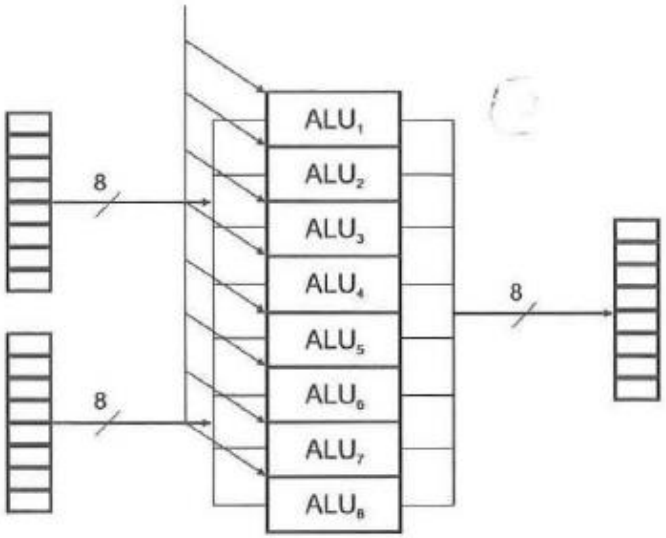


Figura 3.23: Ejemplo simplificado de la realización de una operación sobre dos vectores en un procesador matricial compuesto por ocho EPs.

3.14 Repertorio de instrucciones vectoriales

- Similar al de las operaciones escalares
- Existen instrucciones para realizar operaciones escalares y vectoriales

`ADDV Vi,Vj,Vk`	Almacena en `Vi` el resultado de sumar los elementos de `Vj` y `Vk`.
`ADDSV Vi,Vj,Fi`	Almacena en `Vi` el resultado de sumar `Fi` a cada elemento de `Vj`.
`SUBV Vi,Vj,Vk`	Almacena en `Vi` el resultado de restar los elementos de `Vk` a los de `Vj`.
`SUBSV Vi,Vj,Fi`	Almacena en `Vi` el resultado de restar `Fi` a cada elemento de `Vj`.
`SUBSV Vi,Fi,Vj`	Almacena en `Vi` el resultado de restar cada elemento de `Vj` a `Fi`.
`MULTV Vi,Vj,Vk`	Almacena en `Vi` el resultado de multiplicar los elementos de `Vj` y `Vk`.
`MULTSV Vi,Vj,Fi`	Almacena en `Vi` el resultado de multiplicar `Fi` por cada elemento de `Vj`.
`DIVV Vi,Vj,Vk`	Almacena en `Vi` el resultado de dividir los elementos de `Vj` por los de `Vk`.
`DIVSV Vi,Vj,Fi`	Almacena en `Vi` el resultado de dividir los elementos de `Vj` por `Fi`.
`DIVSV Vi,Fi,Vj`	Almacena en `Vi` el resultado de dividir `Fi` por los elementos de `Vj`.
`LV Vi,Ri`	Carga en `Vi` los elementos ubicados en memoria a partir de la posición `M[Ri]`.
`SV Ri,Vi`	Almacena los elementos de `Vi` a partir de la posición de memoria `M[Ri]`.

Ejemplo del bucle DAXPY (Double Precision A Times X Plus Y)

- ***Y(i) = a*X(i) + Y(i)*** con vectores de 64 elementos de 8 bytes
  - Aproximación escalar
    - Se considera que la longitud de los vectores es de 64 elementos, se ejecutan 578 instrucciones (2 + 9 * 64).

```
        LD      F10,0(R5)     % Carga valor de a desde M[0+R5]
        ADDI    R3,R1,#512    % Cálculo posición del último elemento
bucle:  LD      F2,0(R1)      % Carga de X(i) desde M[0+R1]
        MULTD   F4,F2,F10     % a*X(i)
        LD      F6,0(R2)      % Carga de Y(i) desde M[0+R2]
        ADDD    F6,F4,F6      % Y(i):=a*X(i)+Y(i)
        SD      0(R2),F6      % Almacenamiento de Y(i) en M[0+R2]
        ADDI    R1,R1,#8      % Sumar 1 al índice de X
        ADDI    R2,R2,#8      % Sumar 1 al índice de Y
        SUB     R4,R3,R1      % Comparar R1 con posición del último
        BNZ     R4,bucle      % Si no último, repetir bucle
```

Dado que se considera que la longitud de los vectores es de 64 elementos, en la aproximación escalar se ejecutan 578 instrucciones $(2 + 9 * 64)$. El código vectorial equivalente al escalar es:

```
        LD      F10,0(R5)     % Carga valor de a desde M[0+R5]
        LV      V1,R1         % Carga vector X desde M[R1]
        MULTSV  V2,V1,F10     % a*X
        LV      V3,R2         % Carga vector Y desde M[R2]
        ADDV    V4,V3,V2      % Y:=Y+a*X
        SV      R2,V4         % Almacenamiento vector Y en M[R2]
```

A primera vista se aprecia que el número de instrucciones ejecutadas es mucho menor, solo 6 instrucciones, que en comparación con las casi 600 del código escalar significan una reducción notable

VRL (Vector Length register) Registro de longitud vectorial

- CUESTIONES
  - 1 ¿Cómo proceder cuando la longitud del vector es diferente al número de elementos de un registro vectorial?
  - 2 ¿Qué sucede cuando los elementos del vector se almacenan de forma uniforme pero no consecutiva en memoria?
  - 3 ¿Cómo se vectoriza el cuerpo de un bucle que contiene instrucciones ejecutadas condicionalmente?
- **1 Cuando el tamaño del vector es distinto al numero de registros vectoriales (MVL), se almacena en el registro VRL la longitud del vector.**
  - Con el VRL que controla la longitud de cualquier operación vectorial
  - Empleamos la Técnica de ***Strip mining***
    - Troceado del vector, cuando el tamaño del vector (VRL) es mayor que el valor del MVL (máxima longitud del vector)
    - Si el tamaño del vector es de 11 elementos, en la primera iteración se procesan (n mod MVL) elementos y en las siguientes ya se procesan secciones de longitud MVL

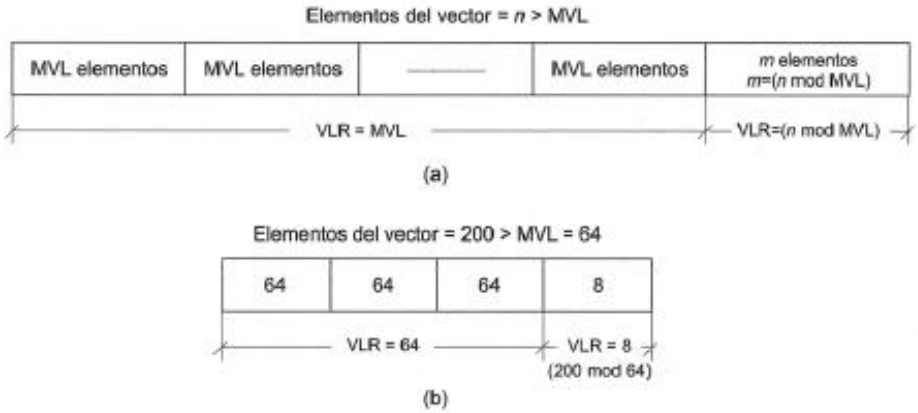


Figura 3.24: Seccionamiento de un vector genérico en secciones de MVL elementos (a) y de un vector de 200 elementos en secciones de 64 (b).

- 2.- El problema **de ubicar en memoria los elementos de un vector de forma no consecutiva**
  - Surge cuando hay que almacenar estructuras de datos que presentan dimensiones superiores a la unidad, tal y como sucede, por ejemplo, con los arrays bidimensionales.
  - Un ejemplo de esta problemática aparece al multiplicar dos matrices de 100x100 almacenadas en los arrays A y B:
  - **SOLUCIÓN**:
    - Instrucciones especiales *with stride* (con distancia de separación)

`LVWS Vi,(Ri,Rj)`	Carga `Vi` comenzando desde la posicion `M[Ri]` con una separación de `Rj`.
`SVWS (Ri,Rj),Vi`	Almacena `Vi` a partir de la posición `M[Ri]` con una separación de `Rj`.

```
for (i=0; i<100; i++)
   for (j=0; j<100; j++)
      C[i,j]:= 0.0;
      for (k=0; k<100; k++)
         C[i,j]:=C[i,j]+A[i,k]*B[k,j];
      end for;
   end for;
end for;
```

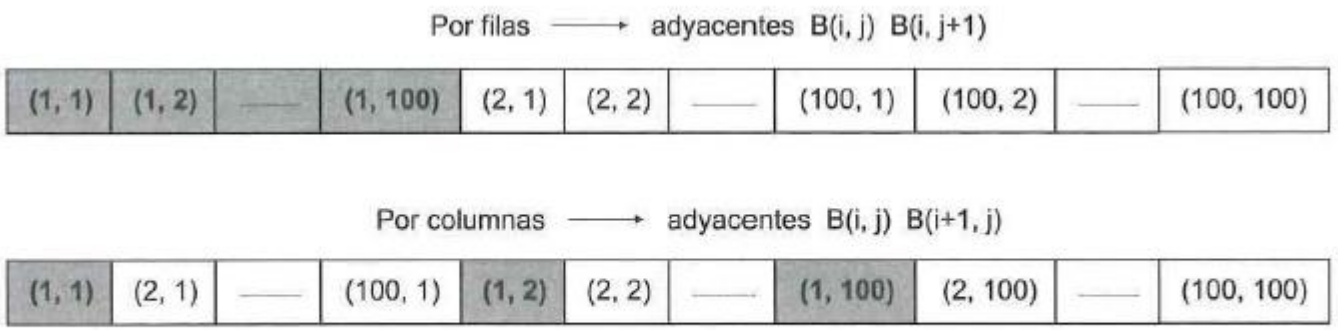


Figura 3.25: Almacenamiento unidimensional por filas o por columnas de una estructura bidimensional de 100x100.

3. Vectorizar bucles en cuyo cuerpo hay instrucciones ejecutadas condicionalmente.

- Dependiendo de la condición, ciertos elementos de un vector no tengan que ser manipulados.
- Solución
  - Una máscara de MVL bits de longitud almacenada en el registro especial VM.
    - VM (Vector Mask)
      - El valor del bit que ocupa la posición *i* en la máscara determina si se realizarán (bit a 1) o no (bit a 0). ·
- Instrucciones especiales

`S__V Vi,Vj`	Compara (EQ, NE, GT, LT, GE, LE) elemento a elemento el contenido de `Vi` y `Vj`. El resultado de la comparación de cada elemento es un bit (cierto=1, falso=0) que se almacena en el registro `VM` para formar una máscara.
`S__SV Fi,Vi`	Similar a la anterior pero utilizando un valor escalar en la comparación.
`RVM`	Inicializa a 1 todos los bits del registro `VM`.
`MOVF2S VM,Fi`	Almacena en `VM` el contenido del registro en coma flotante `Fi`.
`MOVS2F Fi,VM`	Almacena en el registro `Fi` el contenido del registro `VM`.

Ejemplo

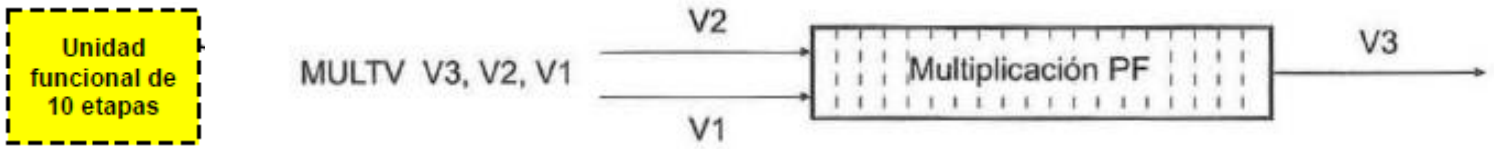
```
for (i=0; i<64; i++)
   if (A[i]!=0) then
      B[i]:=B[i]/A[i];
   end if;
end for;
```

```
LV      V1,R1       % Carga en V1 el vector A
LV      V2,R2       % Carga en V2 el vector B
SNESV   F0,V1       % Compara cada elemento de A con 0
DIVV    V2,V2,V1    % Operación con control de máscara
RVM                 % Inicializa a 1 la máscara
SV      R2,V2       % Almacena todos los elementos de B
```

3.15 Medidas del rendimiento de un fragmento de código vectorial

- Factores que afectan al tiempo de ejecución Tn
  - Latencia en producir el primer resultado o tiempo de arranque
  - Número de elementos a procesar por unidad funcional
  - Tiempo que se tarda en procesar cada datos
- Tiempo de arranque
  - Es el tiempo que transcurre desde que se solicita el primer bloque de datos del vector hasta que se dispone del mismo para ser tratado por la unidad funcional.
    - Tiempo en disponer el primer resultado
    - Será igual al número de etapas
- Tiempo elemento
  - El tiempo que se tarda en completar cada uno de los restantes n elementos (Unidad segmentada→1 ciclo).

$$T_n = T_{arranque} + n * T_{elemento}$$



$T_{arranque}$ = 10 ciclos

$T_{elemento}$ = 1 ciclo

T_{64} = 10 ciclos + 64 elementos * 1 ciclo = 74 ciclos

Figura 3.26: Tiempo de ejecución de una operación de multiplicación sobre 2 vectores de 64 elementos.

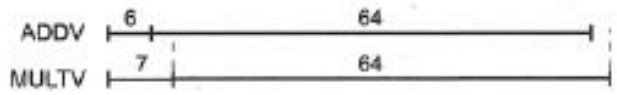
- Situaciones
  - Si no hay riesgos estructurales ni dependencias verdaderas
    - Planificación por convoy o paquetes
    - Convoy o paquete
      - Conjunto de instrucciones sin dependencias reales ni riesgos estructurales que pueden planificarse juntas
  - Si hay riesgos o interdependencias
    - Esperar a que los operandos estén disponibles.
- Otra medida de rendimiento Rn= número de operaciones de coma flotante (FLOP) por ciclo de reloj

R_n= (Operaciones en coma flotante * n elementos) / T_n → A mayor Rn mayor rendimiento

Sin dependencia

ADDV V1, V2, V3

MULTV V4, V5, V6



T_{64} = 7 + 64 = 71 ciclos

$T_{arranque}$ = 7 ciclos

$T_{elemento} = \frac{64 \text{ elementos} * 1 \text{ ciclo}}{64 \text{ elementos}} = 1 \text{ ciclo}$

$R_{64} = \frac{64 \text{ elementos} * 2 \text{ FLOP/elemento}}{71 \text{ ciclos}} = 1{,}8 \text{ FLOP / ciclo}$

(a)

ADDV V1, V2, V3

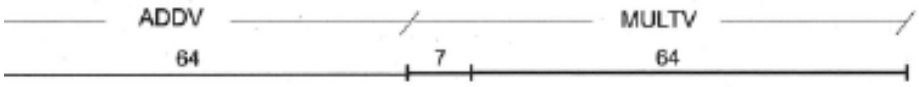
MULTV V4, V5, V1

Ejecución de un convoy de 2 instrucciones

Con dependencia

ADDV V1, V2, V3

MULTV V4, V5, V1



T_{64} = (6 + 64) + (7 + 64) = 141 ciclos

$T_{arranque}$ = 6 + 7 = 13 ciclos

$T_{elemento} = \frac{64 \text{ elementos} * 1 \text{ ciclo} + 64 \text{ elementos} * 1 \text{ ciclo}}{64 \text{ elementos}} = 2 \text{ ciclos}$

$R_{64} = \frac{64 \text{ elementos} * 2 \text{ FLOP/elemento}}{141 \text{ ciclos}} = 0{,}9 \text{ FLOP / ciclo}$

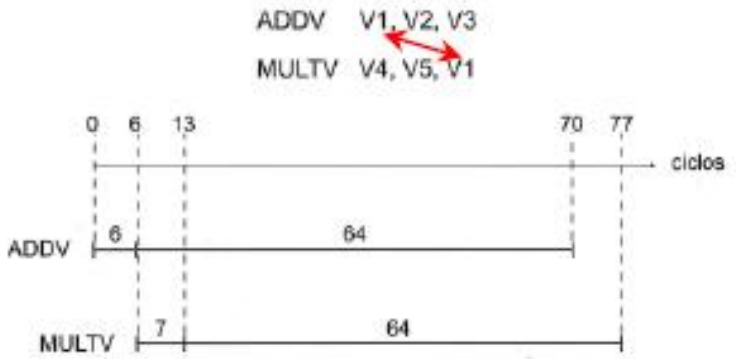
(b)

2 convoyes de una única instrucción (b)

Encadenamiento de resultados entre unidades funcionales (chaining)

- Permite a una unidad funcional empezar a operar tan pronto como los resultados de la unidad funcional de que depende estén disponible (pasado el tiempo de arranque)
  - Aunque dos operaciones sean dependientes, el encadenamiento permite que se realicen en paralelo sobre elementos diferentes de un vector (los que ya han sido procesados) y formen parte del mismo convoy

Con encadenamiento
Con dependendencia



$T_{64} = (6 + 7) + 64 = 77$ ciclos

$T_{arranque} = 6 + 7 = 13$ ciclos

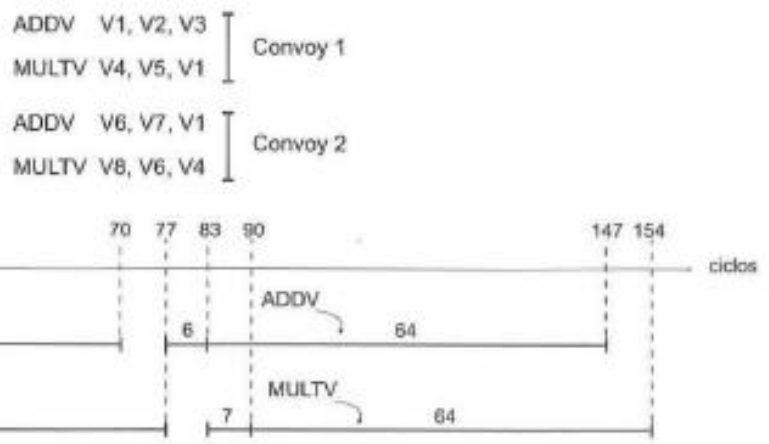
$$T_{elemento} = \frac{64 \text{ elementos} * 1 \text{ ciclo}}{64 \text{ elementos}} = 1 \text{ ciclo}$$

$$R_{64} = \frac{64 \text{ elementos} * 2 \text{ FLOP/elemento}}{77 \text{ ciclos}} = 1{,}66 \text{ FLOP / ciclo}$$

Figura 3.28: Ejecución de dos instrucciones con encadenamiento entre las unidades funcionales.

Solapamiento

Sin solapamiento: hasta que no termina un convoy no puede empezar otro



$T_{64} = (6 + 7) + 64 + (6 + 7) + 64 = 154$ ciclos

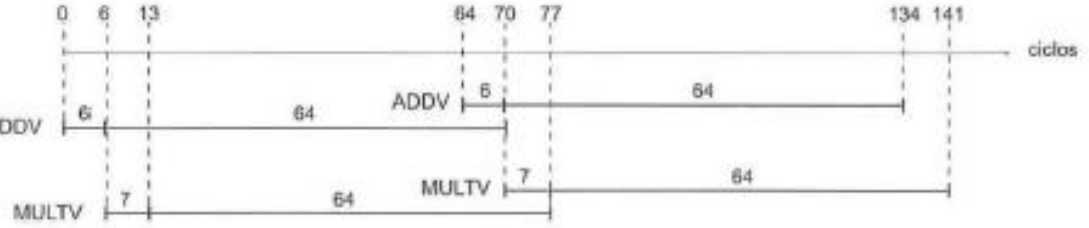
$T_{arranque} = 26$ ciclos

$$T_{elemento} = \frac{64 \text{ elementos} * 2 \text{ ciclos}}{64 \text{ elementos}} = 2 \text{ ciclos}$$

$$R_{64} = \frac{64 \text{ elementos} * 4 \text{ FLOP/elemento}}{154 \text{ ciclos}} = 1{,}66 \text{ FLOP / ciclo}$$

(a)

Ejecución solapada de varios convoy



$T_{64} = (6 + 7) + 64 + 64 = 141$ ciclos

$T_{arranque} = 13$ ciclos

$$T_{elemento} = \frac{64 \text{ elementos} * 2 \text{ ciclos}}{64 \text{ elementos}} = 2 \text{ ciclos}$$

$$R_{64} = \frac{64 \text{ elementos} * 4 \text{ FLOP/elemento}}{141 \text{ ciclos}} = 1{,}81 \text{ FLOP / ciclo}$$

(b)

3.16 La unidad funcional de carga/almacenamiento vectorial

- El elemento hardware más crítico en un procesador vectorial para poder alcanzar un rendimiento óptimo es la **unidad de carga/almacenamiento**.
  - Debe ser capaz de poder intercambiar datos con los registros vectoriales de forma sostenida y a una velocidad igual o superior a la que las unidades funcionales aritméticas consumen o producen nuevos elementos ($T_{elemento}$)
- Solución
  - Organizando físicamente la memoria en varios bancos o módulos y distribuyendo el espacio de direccionamiento de forma uniforme entre todos ellos.
- Ta= Tiempo de arranque
  - Tiempo que transcurre desde que se solicita el primer elemento del vector al sistema de memoria hasta que está disponible en el puerto de lectura del banco, para su posterior transferencia
- Operación de **carga** de un registro vectorial
  - Primero se solicita (Ta = tiempo de arranque) y después se almacenan los datos
  - El tiempo por elemento se considera como el número de ciclos que se consumen en transferir el dato ya disponible en el banco de memoria al registro vectorial (por lo general, inferior a un ciclo pero se iguala a uno para equipararlo al T elemento de las unidades funcionales).
- Operación de **escritura** en memoria (Almacenamiento):
  - 1 Transfiere los elementos del vector a los puertos de escritura de los bancos de memoria
  - 2 Tiempo que tarda en escribir el último elemento del vector en el banco de memoria
    - El tiempo por elemento se considera como el tiempo que se emplea en transferir un elemento desde un registro vectorial al puerto de escritura del banco de memoria.
    - El tiempo de arranque se puede ver como el tiempo que emplea el banco en escribir el último elemento del vector en una posición del banco de memoria.

- La lectura de los *n* elementos de que consta un vector supondría un coste de $n * T_a$ ciclos, lo que es totalmente **inadmisible.**
- Para poder **ocultar** toda esta **latencia** es fundamental **dimensionar** correctamente el **número** de **bancos** de **memoria** de forma que solo se vea el tiempo de acceso correspondiente al primer elemento del vector y que los tiempos de acceso de los demás elementos queden ocultos.
- Esto se consigue **distribuyendo los *n* elementos** de **un vector entre *m* bancos** de memoria para que se solapen.
  - Un dimensionamiento correcto · $m \geq Ta$ (*Ta* en ciclos de reloj)
- Existen dos formas de realizar el solapamiento de las latencias de acceso a los n datos de un vector:
  - De forma síncrona y
  - De forma asíncrona

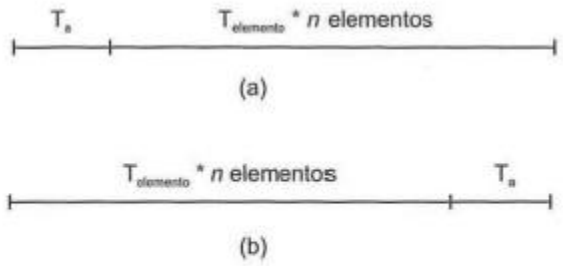
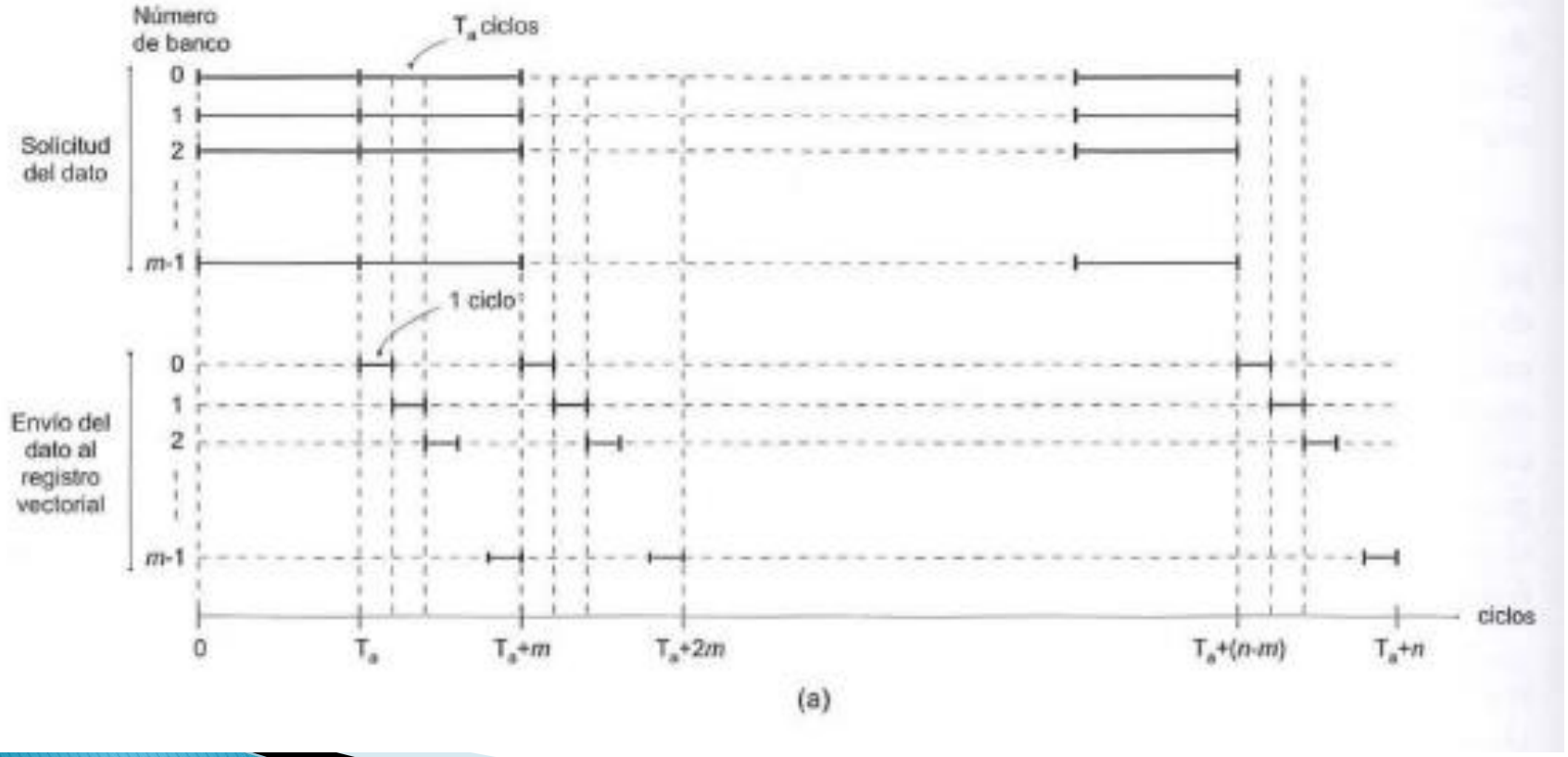


Figura 3.30: Diferencias entre instrucciones de carga (a) y almacenamiento (b) en la visibilidad de los tiempos de acceso y por elemento.

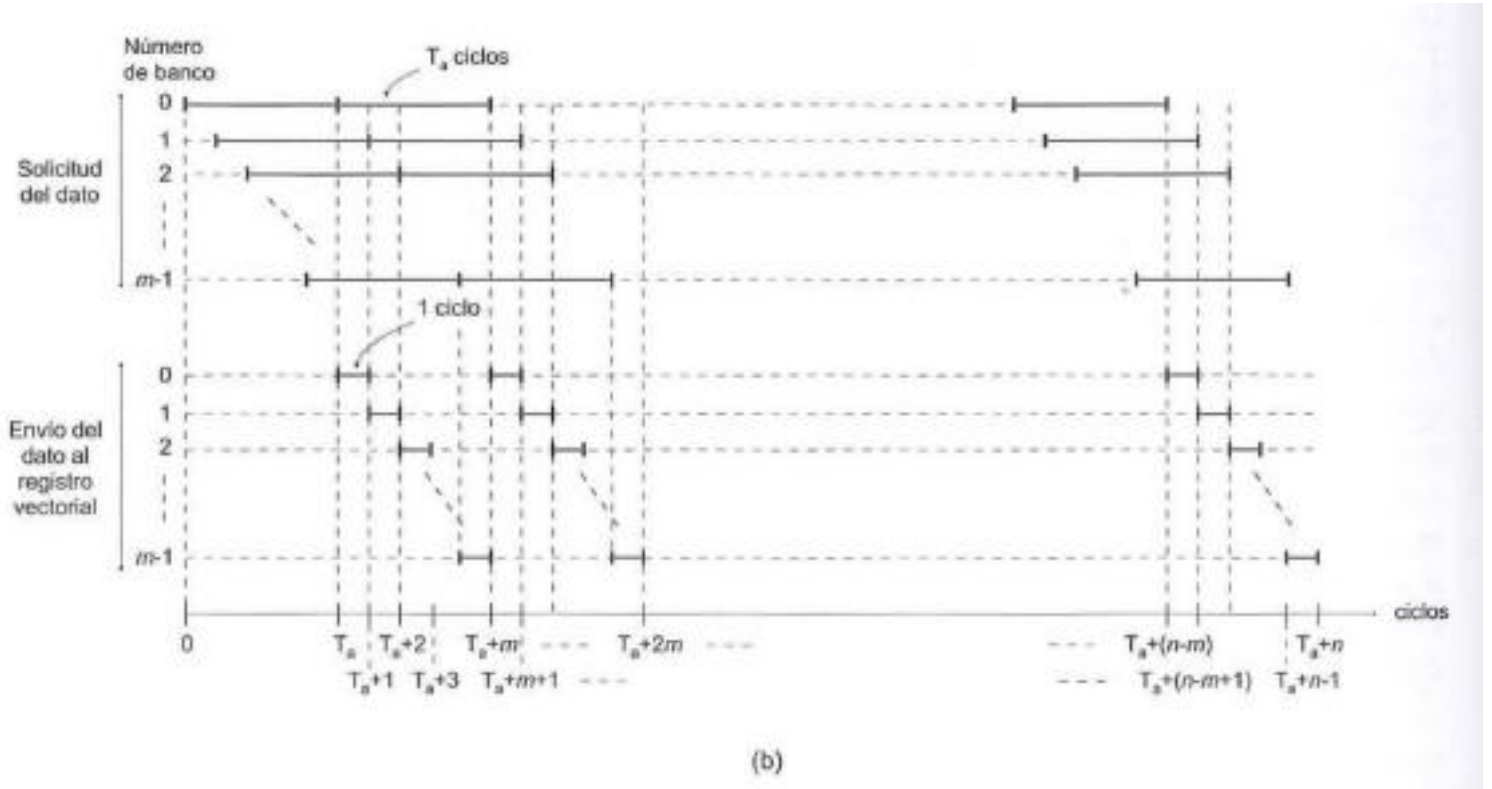
Acceso síncrono

- Solicita simultáneamente un dato a los *m* bancos cada T_a ciclos
- Cada Ta ciclos se realizan dos acciones
  - 1 Efectuar una nueva petición simultáneas a los todos los bancos para extraer los m elementos siguientes del vector
  - 2 Se comienza a transferir ciclo a ciclo los m elementos obtenidos en la fase anterior



Acceso asíncrono

- Solicitar los elementos de que costa el vector a cada uno de los m bancos de forma periódica con periodos de Ta con un desfase entre bancos consecutivos de $T_{elemento}$ ciclo
- De esta forma; comenzando en el ciclo 0 y cada *Ta* ciclos el banco 0 solicita un dato, en el ciclo 1 y cada *Ta* ciclos el banco 1 solicita un nuevo dato, y así sucesivamente hasta el banco m-1.
- En el momento en que un banco tiene el dato disponible realiza dos acciones:
  - 1. Efectúa la nueva solicitud de dato.
  - 2. Transfiere el dato ya disponible al registro vectorial



Almacenamiento

- Para que todo funcione correctamente las palabras que componen un vector deben estar distribuidas correctamente entre los bancos de memoria
- La distribución de los elementos de un vector en los bancos de memoria, se realiza de forma consecutiva y cíclica a partir de una posición de memoria inicial que es múltiplo del ancho de palabras en bytes
- Cuando se va a leer un vector de memoria, se analiza la posición de memoria del primer elemento para así conocer el número del banco en que se encuentra, tras lo que se inicia la lectura de todo el vector
- Dada una dirección de memoria, el banco de memoria en que se encuentra está determinado por los bits de orden inferior de la dirección de memoria

136: 0x10**001**000 (banco 1)
144: 0x10**010**000 (banco 2)
152: 0x10**011**000 (banco 3)
160: 0x10**100**000 (banco 4)
168: 0x10**101**000 (banco 5)
176: 0x10**110**000 (banco 6)
184: 0x10**111**000 (banco 7)
192: 0x11**000**000 (banco 0)

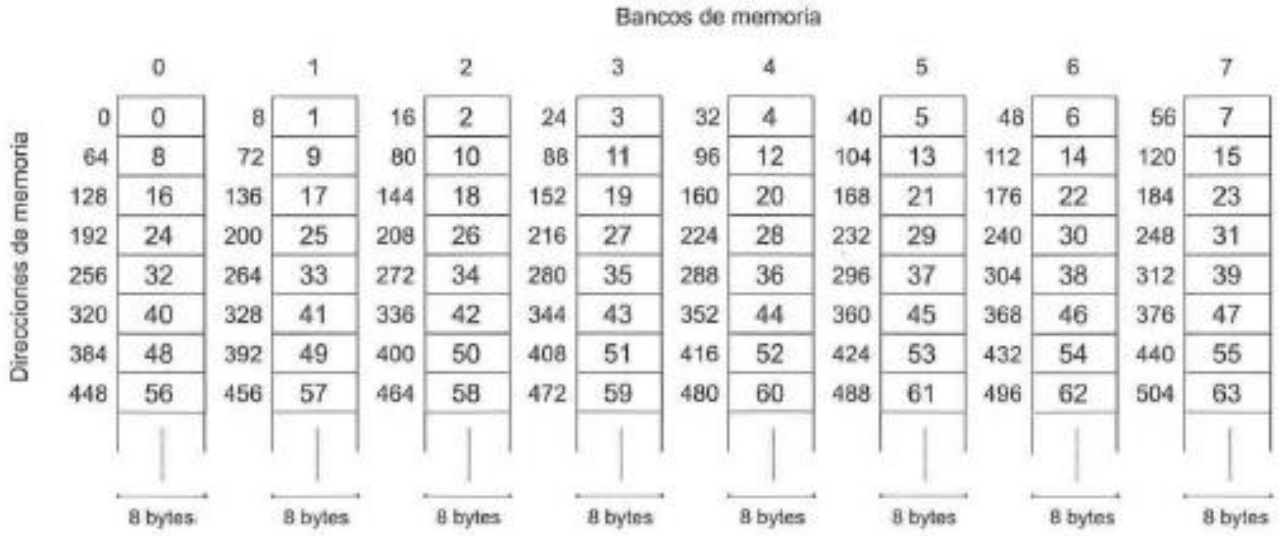


Figura 3.32: Almacenamiento en 8 bancos de un vector de 64 elementos de doble precisión a partir de la dirección de memoria 0. Los números contenidos en las celdas representan el orden del elemento en el vector, no su valor.

Comienzo no en banco cero

- En los ejemplos anteriores el primer elemento de un vector se ubicaba en el banco O, el segundo elemento en el banco 1 y así sucesivamente, lo habitual es que el primer elemento del vector se ubique en cualquier otro banco
  - Comenzando en la dirección 80 con Ta de 6 ciclos
  - 80 =0x01010000, el banco 2

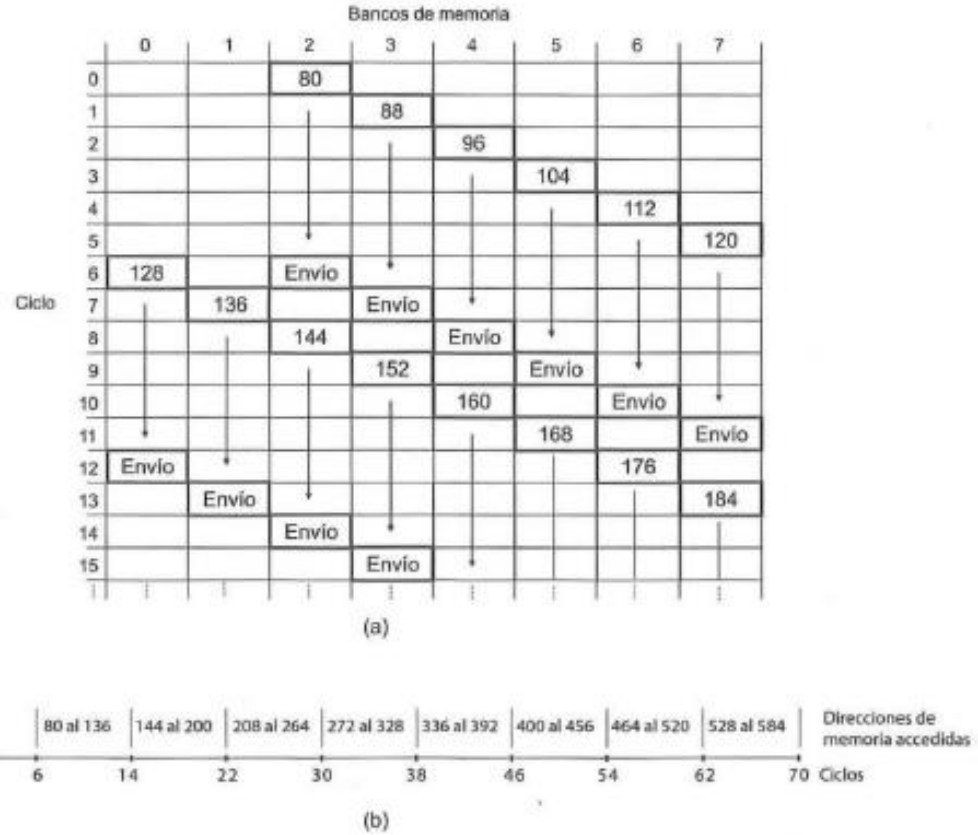


Figura 3.33: Esquema de la transferencia de un vector de 64 elementos ubicado a partir de la posición de memoria 80 en un sistema de memoria organizado en 8 bancos con acceso asíncrono y con T_a de 6 ciclos (a). Relación temporal entre los accesos a los bancos y las palabras enviadas (b).

3.17 Medidas del rendimiento de un bucle vectorizado

- Pseudo-código muestra las operaciones que son necesarias para vectorizar el bucle DAXPY de *n* elementos con la técnica *strip mining:*
  - Strip mining = troceado del vector
    - VLR (vector Length Register)
    - MVL (maximun Vector Length)

Bucle DAXPY de *n* elementos con la técnica *strip mining*

```
1:  primero:=1;                          % Primer elemento de la sección
2:  secciones:=n/MVL;                    % Secciones de MVL elementos
3:  VLR:=n mod MVL;                      % Longitud de la primera sección
4:  for (i=0;i<=secciones;i++)           % Bucle exterior
5:     último:=primero+VLR-1;            % Último elemento de la sección
6:     for (j=primero; j<=último;j++)    % Bucle interior
7:        Y[i]:=a*X[i]+Y[i];             % Operaciones vectoriales
8:     end for;
9:     primero:=primero+VLR;             % Primer elemento de la nueva sección
10:    VLR:=MVL;                         % Inicio longitud de la nueva sección
11: end for;
```

Nº elementos del vector

Posición en el vector del 1º elemento

Nº total de secciones de MVL elementos

Componentes de los costes de ejecución de instrucciones vectoriales y escalares

Componentes del coste	T_{base}: Es el tiempo que consumen las instrucciones escalares de preparación antes de abordar el bucle exterior.	Líneas 1, 2 y 3
	T_{bucle}: Son los costes derivados de ejecutar en cada iteración del bucle exterior las instrucciones escalares necesarias para realizar el seccionamiento.	Líneas 4, 5, 9, 10 y 11
	$T_{arranque}$: Es la suma de los tiempos de arranque visibles de las unidades funcionales que se utilizan en cada convoy de instrucciones.	
	$T_{elemento}$: Es igual al número de convoyes en que se organizan las instrucciones vectoriales que se derivan del bucle interior	Líneas 6, 7 y 8.

$$T_n = T_{base} + \left\lceil \frac{n}{\text{MVL}} \right\rceil * \left(T_{bucle} + T_{arranque}\right) + n * T_{elemento}$$

Valor entero inmediatamente superior

$$R_\infty = \lim_{n\to\infty}\left(\frac{\text{Operaciones vectoriales} * n}{T_n}\right)$$

```
1: primero:=1;                          % Primer elemento de la sección
2: secciones:=n/MVL;                    % Secciones de MVL elementos
3: VLR:=n mod MVL;                      % Longitud de la primera sección
4: for (i=0;i<=secciones;i++)           % Bucle exterior
5:    último:=primero+VLR-1;            % Último elemento de la sección
6:    for (j=primero; j<=último;j++)    % Bucle interior
7:       Y[i]:=a*X[i]+Y[i];             % Operaciones vectoriales
8:    end for;
9:    primero:=primero+VLR;             % Primer elemento de la nueva sección
10:   VLR:=MVL;                         % Inicio longitud de la nueva sección
11:end for;
```

Ejemplo

Análisis del rendimiento de un procesador vectorial al ejecutar el código obtenido de vectorizar el conocido bucle DAXPY para vectores de 11 elementos.

Ejemplo

Bucle DAXPY para vectores de *n* elementos.

- Una unidad de suma (6 ciclos de latencia).
- Una unidad de multiplicación (7 ciclos de latencia).
- Una unidad de carga/almacenamiento (12 ciclos de latencia).
- MVL es 64.
- La frecuencia de reloj es 500 MHz.
- VLR = 64

El fragmento de código vectorial que se genera para realizar las operaciones

Y(i): =a*X(i)+Y(i)

```
LV      V1,R1       % Carga de una sección de X
MULTSV  V2,V1,F0    % Operación vectorial a*X
LV      V3,R2       % Carga de una sección de Y
ADDV    V4,V3,V2    % Operación vectorial a*X+Y
SV      R2,V4       % Almacenamiento sección de Y
```

$T_{base} = 10$ ciclos y $T_{bucle} = 15$ ciclos

Caso 1: sin encadenamiento de resultados entre unidades

```
Convoy 1:   LV       V1,R1
Convoy 2:   MULTSV   V2,V1,F0
            LV       V3,R2
Convoy 3:   ADDV     V4,V3,V2
Convoy 4:   SV       R2,V4
```

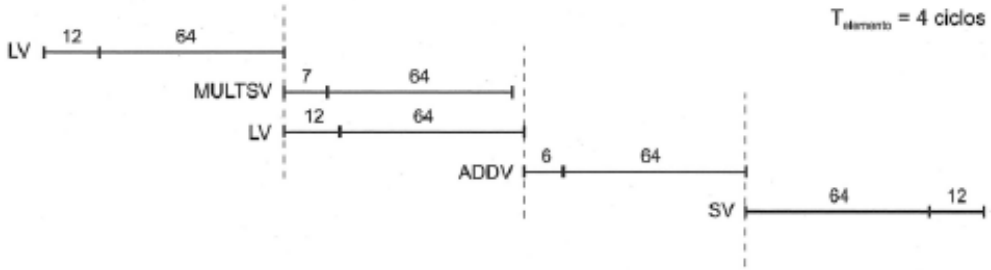


Figura 3.34: Secuencia de ejecución de los cuatro convoyes con VLR = 64.

Dado que hay cuatro convoyes, $T_{elemento}$ es 4 ciclos

$T_{arranque}$ total es igual a la suma de los tiempos de arranque visibles de los cuatro convoyes.

$$T_{arranque} = 2 * T_{arranqueLV} + T_{arranqueADDV} + T_{arranqueSV}$$

$$T_{arranque} = (2 * 12 + 6 + 12)\ \text{ciclos} = 42\ \text{ciclos}$$

$$T_n = T_{base} + \left\lceil \frac{n}{\text{MVL}} \right\rceil * \left(T_{bucle} + T_{arranque}\right) + n * T_{elemento}$$

$$T_n = 10 + \left\lceil \frac{n}{64} \right\rceil * (15 + 42) + 4 * n$$

$$T_{1000} = 10 + \left\lceil \frac{1000}{64} \right\rceil * (15 + 42) + 4 * 1000$$

Para n=1000 $\Rightarrow$

$$T_{1000} = 10 + 16 * (15 + 42) + 4 * 1000$$

$$T_{1000} = 4922\ \text{ciclos}$$

FLOPs/ciclo

$$R_\infty = \lim_{n\to\infty}\left(\frac{\text{Operaciones vectoriales} * n}{T_n}\right)$$

Para simplificar los cálculos, la expresión $\lceil n/64 \rceil$ se puede reemplazar por una cota superior dada por $(n/64 + 1)$.
Nº oper. Vectoriales = 2 (1 ADDV y 1 MULTSV)

$$R_\infty = \lim_{n\to\infty}\left(\frac{2 * n}{10 + \left(\frac{n}{64} + 1\right) * (15 + 42) + 4 * n}\right)$$

$$R_\infty = \lim_{n\to\infty}\left(\frac{2 * n}{67 + 4,89 * n}\right)$$

$$R_\infty = 0,409\ \text{FLOP/ciclo}$$

$$R_\infty = 0,409\ \text{FLOP/ciclo} * (500 * 10^6)\ \text{Hz}$$

$$R_\infty = 204,5\ \text{MFLOPS}$$

Caso 2: Con encadenamiento de resultados entre unidades

```
Convoy 1:   LV      V1,R1
            MULTSV  V2,V1,F0
Convoy 2:   LV      V3,R2
            ADDV    V4,V3,V2
Convoy 3:   SV      R2,V4
```

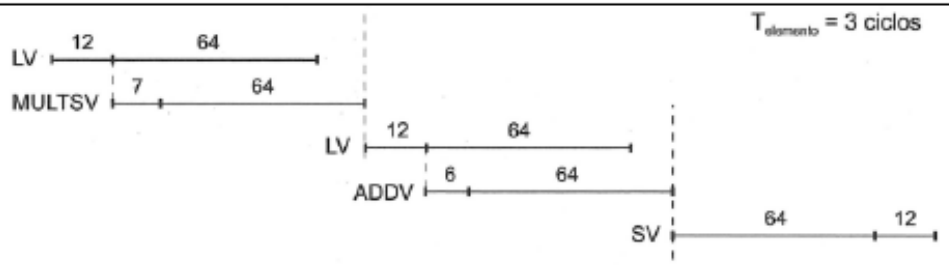


Figura 3.35: Secuencia de ejecución de los tres convoyes con VLR = 64

El $T_{elemento}$ ha pasado a ser de 3 ciclos dado que ahora se tienen 3 convoyes. El $T_{arranque}$ total se obtiene de sumar los tiempos de arranque visibles de las unidades funcionales. Si se analiza la Figura 3.35 se tiene

$$T_{arranque} = 2 * T_{arranqueLV} + T_{arranqueMULTSV} + T_{arranqueADDV} + T_{arranqueSV}$$

$$T_{arranque} = (2 * 12 + 7 + 6 + 12)\,\text{ciclos} = 49\,\text{ciclos}$$

Con estos valores la expresión del tiempo total de ejecución queda

$$T_n = 10 + \left\lceil \frac{n}{64} \right\rceil * (15 + 49) + 3 * n$$

que para el caso particular de $n = 1000$

$$T_n = T_{base} + \left\lceil \frac{n}{\text{MVL}} \right\rceil * \left(T_{bucle} + T_{arranque}\right) + n * T_{elemento}$$

$$T_{1000} = 10 + \left\lceil \frac{1000}{64} \right\rceil * (15 + 49) + 3 * 1000$$

$$T_{1000} = 10 + 16 * (15 + 49) + 3 * 1000$$

$$T_{1000} = 4034\,\text{ciclos}$$

Con respecto al caso 1, el permitir encadenamiento de resultados entre unidades funcionales ha reducido el tiempo de ejecución un 18 %, pasando de 4922 a 4034 ciclos. En lo que respecta al rendimiento expresado en FLOP por ciclo

$$R_\infty = \lim_{n\to\infty}\left(\frac{2*n}{T_n}\right)$$

$$R_\infty = \lim_{n\to\infty}\left(\frac{2*n}{10+\left\lceil\frac{n}{64}\right\rceil*(15+49)+3*n}\right)$$

$$R_\infty = \lim_{n\to\infty}\left(\frac{2*n}{10+\left(\frac{n}{64}+1\right)*(15+49)+3*n}\right)$$

$$R_\infty = \lim_{n\to\infty}\left(\frac{2*n}{74+4*n}\right)$$

$$R_\infty = 0,5\,\text{FLOP/ciclo}$$

Para expresar R_∞ en FLOPS hay que multiplicar el valor anterior por la frecuencia del procesador. Se tiene así

$$R_\infty = 0,5\,\text{FLOP/ciclo} * (500 * 10^6)\,\text{Hz}$$

$$R_\infty = 250\,\text{MFLOPS}$$

Claramente se aprecia la mejora en el rendimiento del procesador gracias al encadenamiento de los resultados entre las unidades funcionales.

Caso 3: Con encadenamiento y dos unidades de carga/almacenamiento

```
                LV      V1,R1
                MULTSV  V2,V1,F0
Convoy 1:       LV      V3,R2
                ADDV    V4,V3,V2
Convoy 2:       SV      R2,V4
```

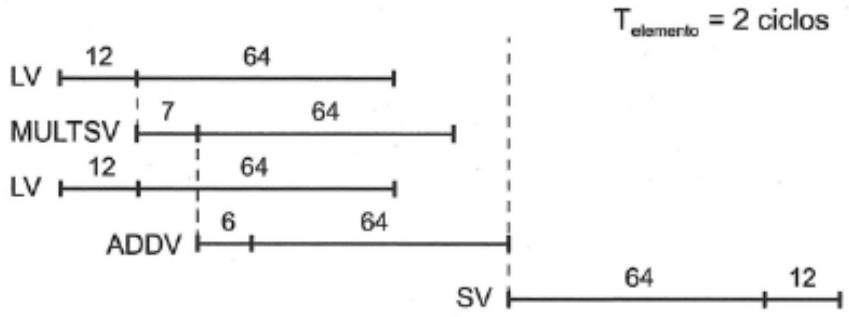


Figura 3.36: Secuencia de ejecución de los dos convoyes con VLR = 64.

$$T_{arranque} = T_{arranqueLV} + T_{arranqueMULTSV} + T_{arranqueADDV} + T_{arranqueSV}$$

$$T_{arranque} = (12 + 7 + 6 + 12)\,\text{ciclos} = 37\,\text{ciclos}$$

Se tiene ahora

$$T_n = 10 + \left\lceil \frac{n}{64} \right\rceil * (15 + 37) + 2 * n$$

y para $n = 1000$

$$T_{1000} = 10 + \left\lceil \frac{1000}{64} \right\rceil * (15 + 37) + 2 * 1000$$

$$T_{1000} = 10 + 16 * (15 + 37) + 2 * 1000$$

$$T_{1000} = 2842\,\text{ciclos}$$

Con respecto al caso 1 la mejora es del 73 % ya que el total de ciclos consumidos para procesar el bucle DAXPY con 1000 elementos se ha reducido de 4922 a 2842. El rendimiento en $^{\text{FLOP}}/_{\text{ciclo}}$ es

$$R_\infty = \lim_{n\to\infty}\left(\frac{2*n}{T_n}\right)$$

$$R_\infty = \lim_{n\to\infty}\left(\frac{2*n}{10+\left\lceil\frac{n}{64}\right\rceil*(15+37)+2*n}\right)$$

$$R_\infty = \lim_{n\to\infty}\left(\frac{2*n}{10+\left(\frac{n}{64}+1\right)*(15+37)+2*n}\right)$$

$$R_\infty = \lim_{n\to\infty}\left(\frac{2*n}{62+2,8125*n}\right)$$

$$R_\infty = 0,711\ ^{\text{FLOP}}/_{\text{ciclo}}$$

Si se expresa en FLOPS, se obtiene

$$R_\infty = 0,711\ ^{\text{FLOP}}/_{\text{ciclo}} * (500 * 10^6)\ \text{Hz}$$

$$R_\infty = 355,55\ \text{MFLOPS}$$

Caso 4: Con encadenamiento, dos unidades de carga/almacenamiento y solapamiento entre convoyes dentro de la misma iteración

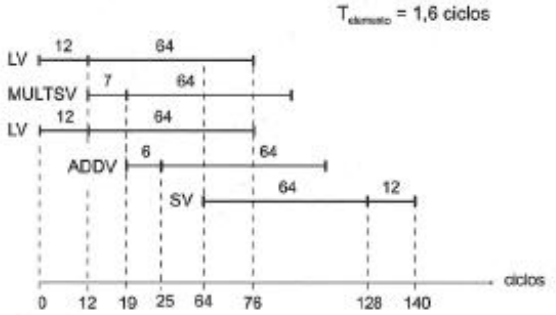


Figura 3.37: Secuencia de ejecución de los dos convoyes con VLR = 64.

$$T_{elemento} = \left(T_n - T_{arranque}\right)/n$$

Dado que para un valor de VL de 64, el tiempo de ejecución de las instrucciones vectoriales en este caso es de 140 ciclos (Figura 3.37) y los tiempos de arranque son

$$T_{arranque} = T_{arranqueLV} + T_{arranqueMULTSV} + T_{arranqueADDV} + T_{arranqueSV}$$

$$T_{arranque} = (12 + 7 + 6 + 12)\,\text{ciclos} = 37\,\text{ciclos}$$

Se tiene así

$$T_{elemento} = \left(T_{64} - T_{arranque}\right)/64$$

$$T_{elemento} = (140 - 37)/64$$

$$T_{elemento} = 1,6\,\text{ciclos}$$

Como T_{bucle} es cero dado su solapamiento con el código vectorial, el tiempo de ejecución del bucle vectorizado para n elementos es

$$T_n = 10 + \left\lceil \frac{n}{64} \right\rceil * 37 + 1,6 * n$$

y particularizando para un vector de 1000 elementos

$$T_{1000} = 10 + \left\lceil \frac{1000}{64} \right\rceil * 37 + 1,6 * 1000$$

$$T_{1000} = 10 + 16 * 37 + 1600$$

$$T_{1000} = 2202\,\text{ciclos}$$

La mejora que se obtiene con respecto al caso 1 es del 123 % al reducirse el total de ciclos consumidos de 4922 a 2202. El rendimiento en FLOP/ciclo es

$$R_\infty = \lim_{n \to \infty} \left(\frac{2 * n}{T_n} \right)$$

$$R_\infty = \lim_{n \to \infty} \left(\frac{2 * n}{10 + \left\lceil \frac{n}{64} \right\rceil * 37 + 1,6 * n} \right)$$