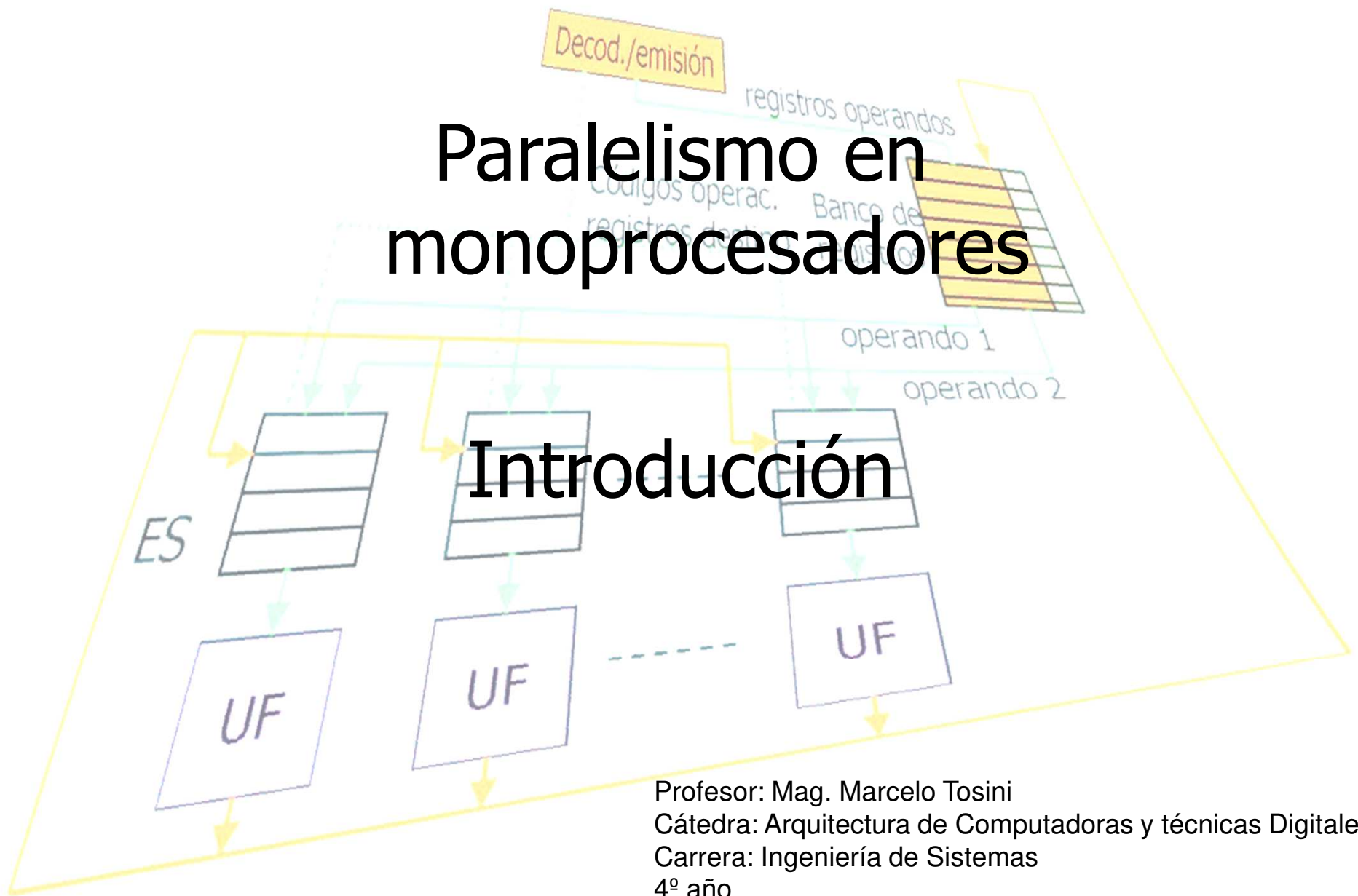


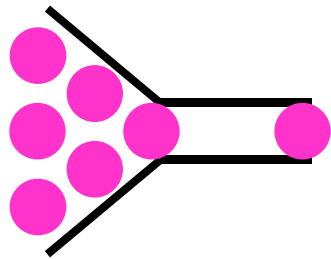
Paralelismo en monoprocesadores

Introducción

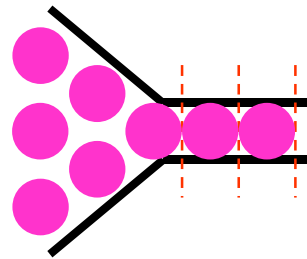


Profesor: Mag. Marcelo Tosini
Cátedra: Arquitectura de Computadoras y técnicas Digitales
Carrera: Ingeniería de Sistemas
4º año

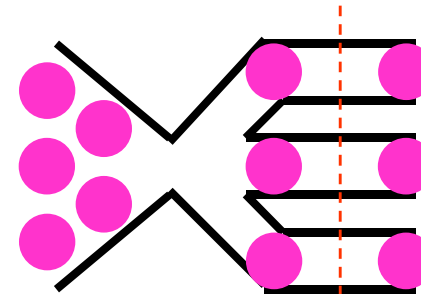
Técnicas para mejorar el rendimiento de un procesador



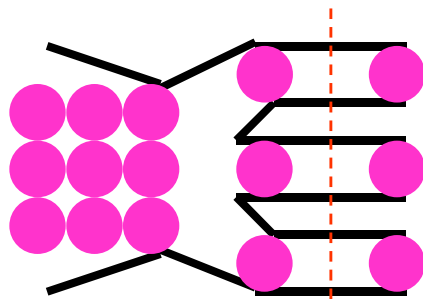
Sin mejoras



Segmentado
(Escalar)

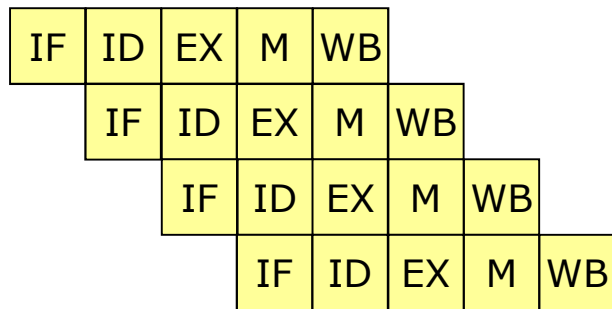


Superescalar



VLIW (EPIC)

Comparación de procesadores



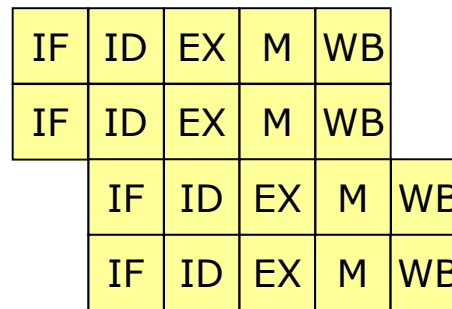
Segmentado

Características

- Emisión de instrucciones cada ciclo

Limitaciones

- Frecuencia de emisión
- Paradas de UF (Riesgos)
- Profundidad de UF



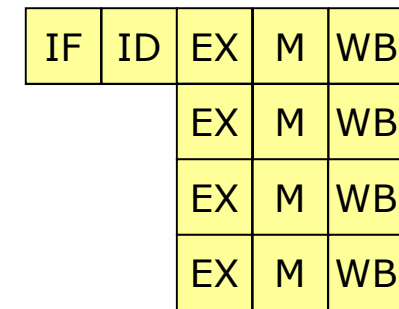
Superescalar

Características

- Emisión múltiple de instrucciones

Limitaciones

- Resolución de riesgos
- Búsqueda de ILP



VLIW

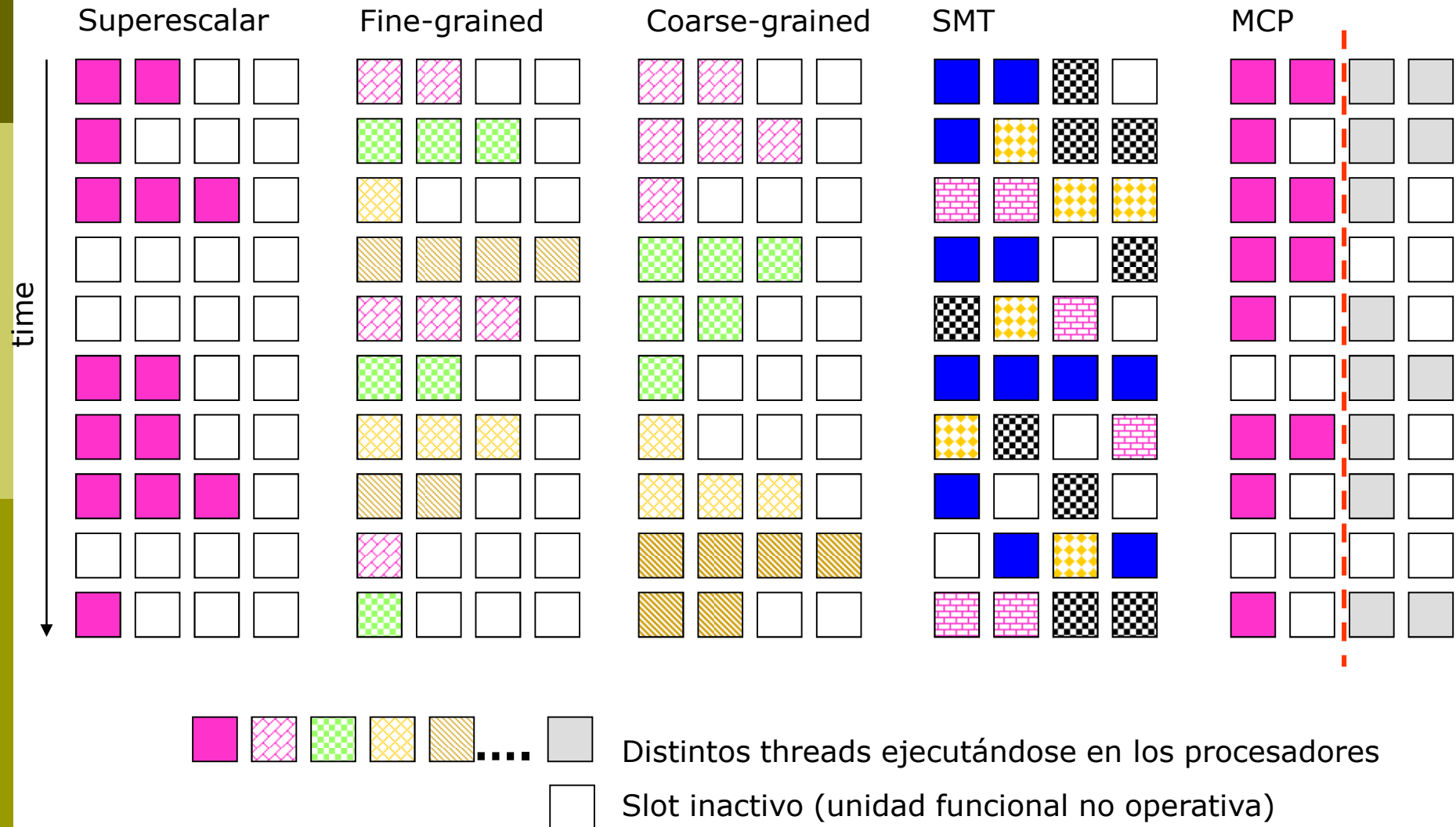
Características

- El compilador determina el grado de paralelismo
- Una instrucción implica múltiples operaciones

Limitaciones

- Empaquetamiento
- Compatibilidad de código

Ejecución en las distintas arquitecturas



Dependencias de un programa

Una dependencia entre dos sentencias de un programa es algún tipo de conflicto que evita que las mismas puedan ejecutarse concurrentemente.

Clasificación:

- Estructurales (de recursos)

consecuencia de las limitaciones de hardware disponible en un sistema de computación. ocurre cuando dos sentencias intentan usar el mismo recurso simultáneamente

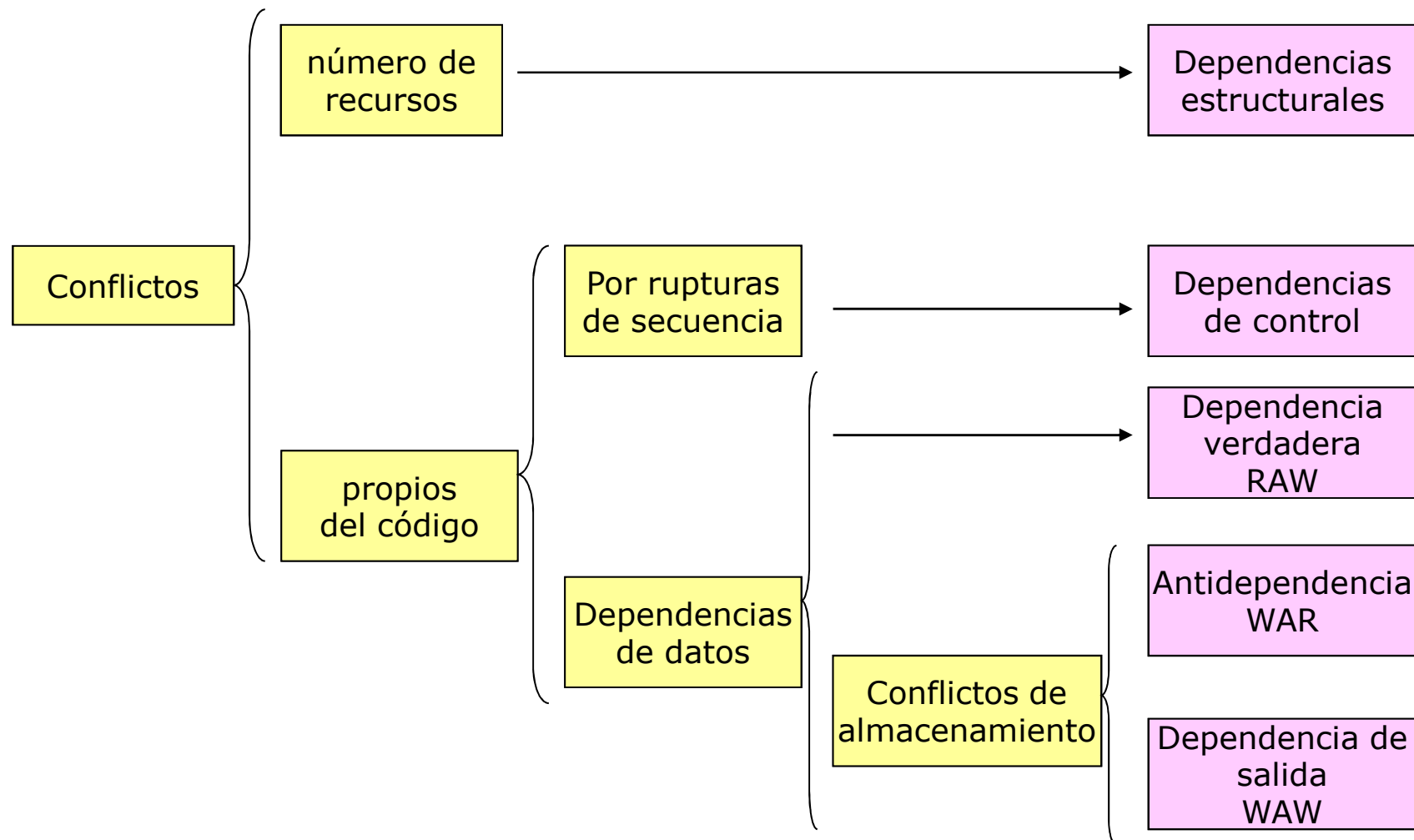
- de control

Una dependencia de la sentencia S_i a S_j existe cuando la sentencia S_j debiera ser ejecutada sólo si S_i produce un cierto resultado.

- de datos

se da entre dos sentencias cuando ambas referencian la misma posición de memoria o acceden a un mismo registro

Limitaciones en la ejecución de instrucciones

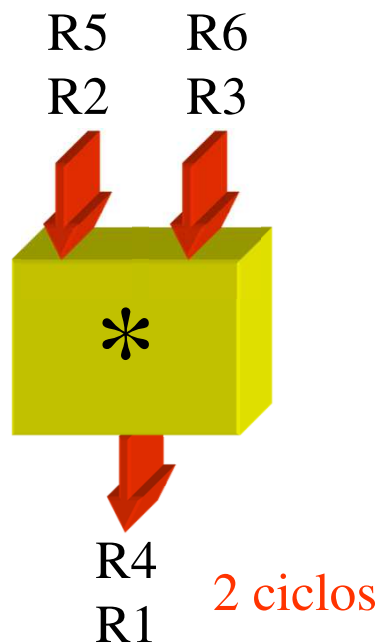


Dependencias estructurales

Con recursos limitados (uno)

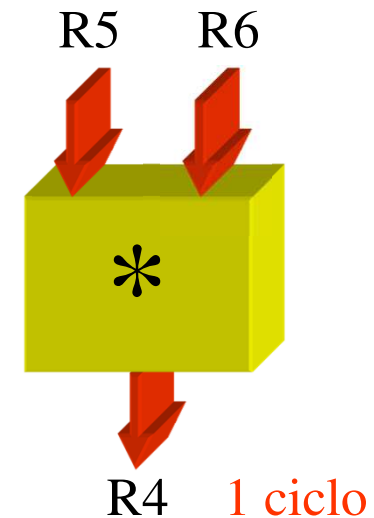
S1 : mul R1, R2, R3

S2 : mul R4, R5, R6



Con recursos ilimitados (dos)

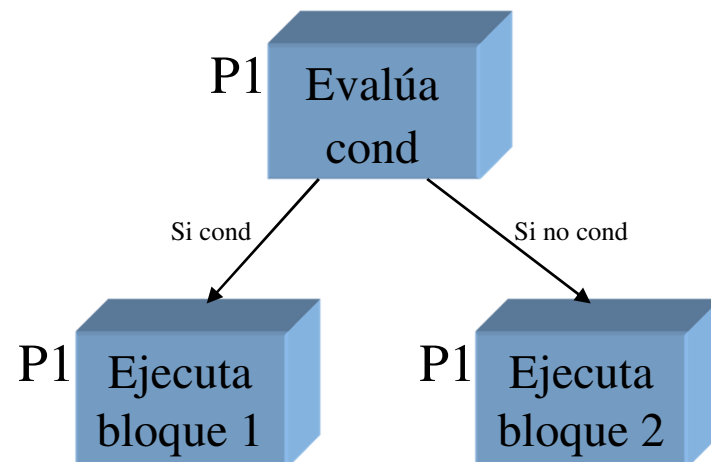
S1 : mul R1, R2, R3 S2 : mul R4, R5, R6



Dependencias de control

Monoprocesador

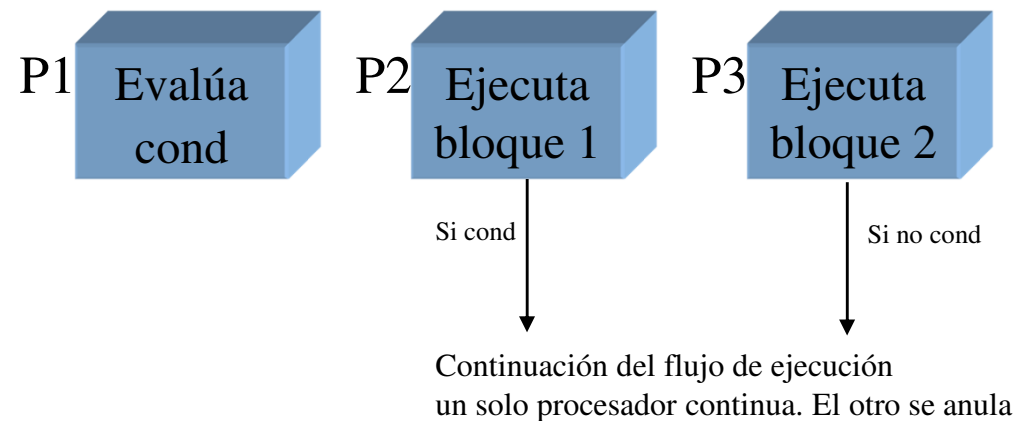
IF (cond)
bloque 1
else
bloque2



2 ciclos

Multiprocesador

IF (cond)
bloque 1
else
bloque2

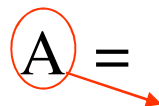


1 ciclo

Dependencias de datos

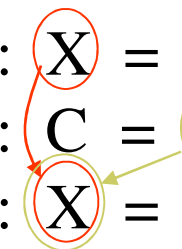
Flujo (read after Write hazard): Se da entre S1 y S2 cuando S2 necesita el como operando el resultado de S1

S1 : $\textcircled{A} = B + C$
S2 : $D = \textcircled{A} - E$



Salida (Write after Write hazard): Se da entre S1 y S3 cuando ambas escriben en la misma posición de memoria

S1 : $\textcircled{X} = Y + Z$
S2 : $C = \textcircled{X} * 22$
S3 : $\textcircled{X} = A - B$



Antidependencia (Write after read hazard): Se da entre S2 y S3 ya que S2 necesita leer un valor antes que S3 lo escriba

Dependencias de datos

NOTAS

- Las dependencias de flujo son las únicas dependencias verdaderas en las que el resultado producido por la primer sentencia es usado como valor por la segunda sentencia.
- Las antidependencias y la dependencia de salida ocurren cuando el programador o el compilador reusan espacio de memoria para reducir el espacio de variables del programa.

En estos casos, renombrar variables es una buena política para eliminar estas dependencias.

Eliminación de dep. de datos

Sin renombre

aux = B + C
X = aux * 2
write (X)
aux = J + 8
X = aux + J
write (X)

P1	P2
aux = B + C X = aux * 2 write (X)	aux = J + 8 X = aux + J write (X)

5 ciclos

Con renombre

aux = B + C
X = aux * 2
write (X)
aux2 = J + 8
X2 = aux2 + J
write (X2)

P1	P2
aux = B + C X = aux * 2 write (X)	aux2 = J + 8 X2 = aux2 + J write (X2)

3 ciclos

Máxima aceleración de LOOPS

2 motivos principales:

- Los loops se repiten varias veces por lo que su optimización representa un gran aumento de performance
- Permite comparar diferentes técnicas de paralelización sin importar la arquitectura subyacente

Premisas de análisis

El análisis se concentrará en las dependencias de datos

Se puede suponer que las dependencias de recursos se eliminan con hardware adicional

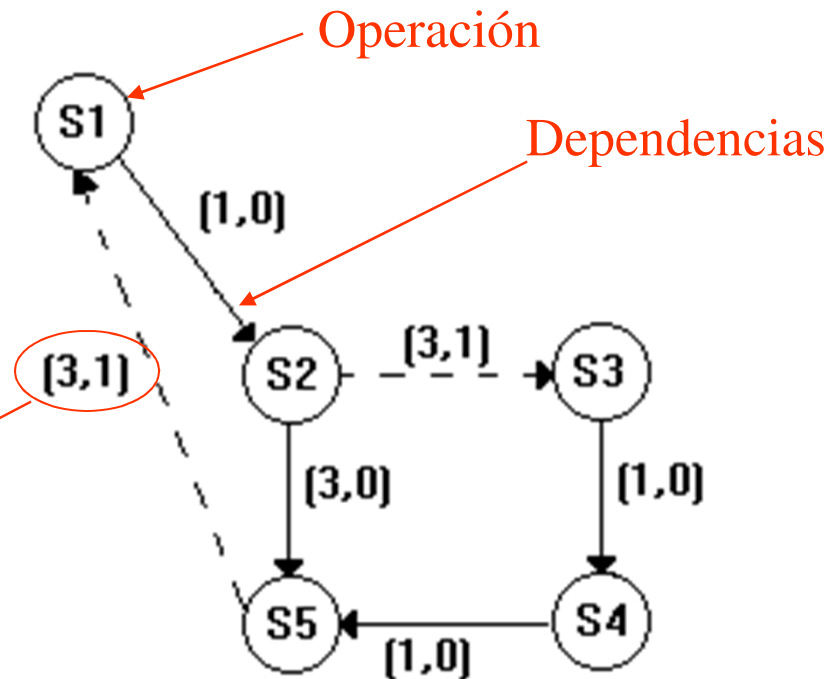
Por simplicidad analizaremos ciclos normalizados con índices de 1 hasta N con incrementos de 1 paso, aunque el análisis puede extenderse a ciclos más complejos posteriormente.

Máxima aceleración de LOOPS

Grafos orientados representan las dependencias de datos

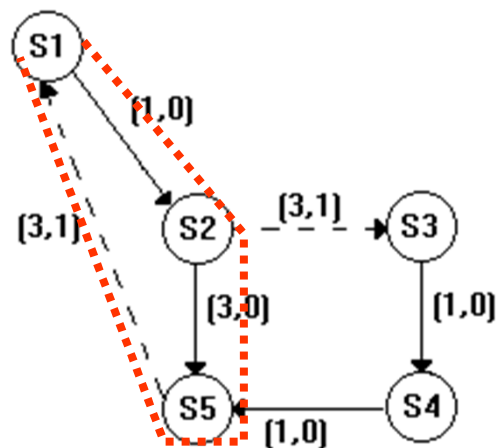
```
For (i=1; i<=N ; i++) {  
  S1 : A(i) = E(i-1) + 6  
  S2 : B(i) = A(i) * Z  
  S3 : C(i) = B(i-1) + X  
  S4 : D(i) = C(i) + Y  
  S5 : E(i) = B(i) * D(i)  
}
```

(T_k, C_k)
entre S_i y S_j
 T_k : tiempo de ejecución de S_i
 C_k : cantidad de iteraciones posteriores a S_i
en que se ejecuta S_j

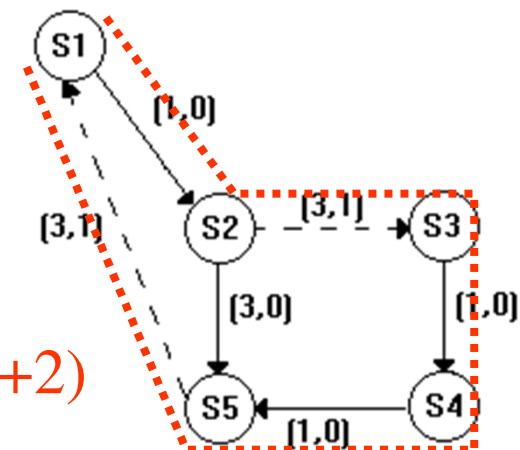


Análisis del ejemplo...

El grafo de dependencia de la figura anterior posee dos ciclos



$S1(i) - S2(i) - S5(i) - S1(i+1)$



$S3(i) - S4(i) - S5(i) - S1(i+1) - S2(i+1) - S3(i+2)$

Ciclo 1...

S1(i) - S2(i) - S5(i) - S1(i+1)

Tiempo	Sentencia
1	S1(1)
2	S2(1)
3	
4	
5	S5(1)
6	
7	
8	S1(2)
9	S2(2)
10	
11	
12	S5(2)
13	
14	
15	S1(3)
...	

Las sentencias se ejecutan N veces

Cada repetición del ciclo requiere

$$T_c = \sum T_k = 7 \text{ unidades de tiempo}$$

El tiempo total para ejecutar las sentencias será

$$T_{t1} = N * T_c = N * 7 \text{ unidades de tiempo.}$$

Ciclo 2...

$S3(i) - S4(i) - S5(i) - S1(i+1) - S2(i+1) - S3(i+2)$

Tiempo	Sentencia
1	S3(1)
2	S4(1)
3	S5(1)
4	
5	
6	S1(2)
7	S2(2)
8	
9	
10	S3(3)
11	S4(3)
12	S5(3)
13	
14	
15	S1(4)
16	S2(4)
17	
18	
19	S3(5)
...	

La ejecución de las sentencias requiere

$$T_c = \sum T_k = 9 \text{ unidades de tiempo}$$

Puesto que el patrón se repite cada dos iteraciones,
la versión desenrollada del ciclo se ejecuta $N/2$ veces

El tiempo total de ejecución para este ciclo es

$$T_{t2} = T_c * (N/2) = 9N/2 \text{ unidades de tiempo.}$$

En general

El máximo paralelismo en un loop con un ciclo en su grafo de dependencias está limitado por el factor de dependencia crítico, T_{crit} , entonces el menor valor de II debido a dependencias de datos es

$$T_{crit}^{datos} (DDG) = MAX_{\forall ciclo \in DDG} \frac{l(ciclo)}{d(ciclo)} = \frac{l(crit)}{d(crit)}$$

Siendo $l(p)$ la suma de todas las latencias de una trayectoria y
 $d(p)$ la suma de todas las distancias de una trayectoria
DDG: Data Dependence Graph

Entonces:

$$T_{crit}^{ciclo1} = \frac{l(ciclo)}{d(ciclo)} = \frac{7}{1}$$

$$T_{crit}^{ciclo2} = \frac{l(ciclo)}{d(ciclo)} = \frac{9}{2}$$

$$Tt1 = T_{crit}^{ciclo1} * N = 7/1 * N$$

$$Tt2 = T_{crit}^{ciclo2} * N = 9/2 * N$$

Ejecución paralela del ciclo 2

En general, para un ciclo determinado,
Se puede inferir el grado de paralelismo
En funcion de la distancia de si tiempo
crítico

$$T_{crit}ciclo2 = \frac{l(ciclo)}{d(ciclo)} = \frac{9}{2}$$

Procesador 1

S1(1)

S2(1)

S3(1)

S4(1)

S5(1)

S1(3)

S2(3)

S3(3)

S4(3)

S5(3)

procesador 2

S1(2)

S2(2)

S3(2)

S4(2)

S5(2)

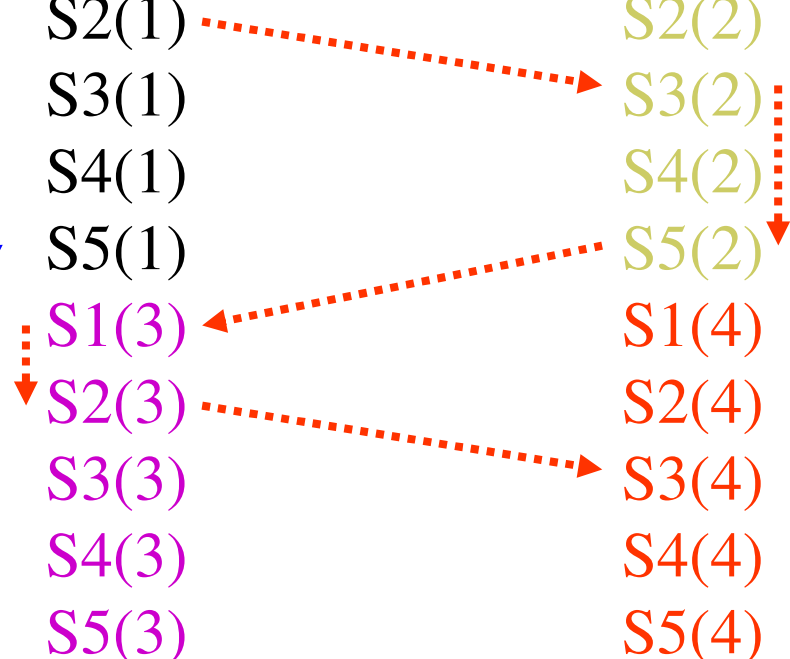
S1(4)

S2(4)

S3(4)

S4(4)

S5(4)



Tiempo de dependencia crítico

trayectoria crítica total (total critical path): cadena de dependencia más larga producida al desenrollar los ciclos de un loop. Su tiempo de ejecución se denota como TT_{crit} .

$$TT_{\text{crit}} = \max(T_{ti}) = \max(7N, 9/2N) = 7N \text{ unidades de tiempo}$$

Aceleración máxima: cociente entre el tiempo de la versión secuencial original y el tiempo de ejecución de la trayectoria crítica

$$T_L = 9$$

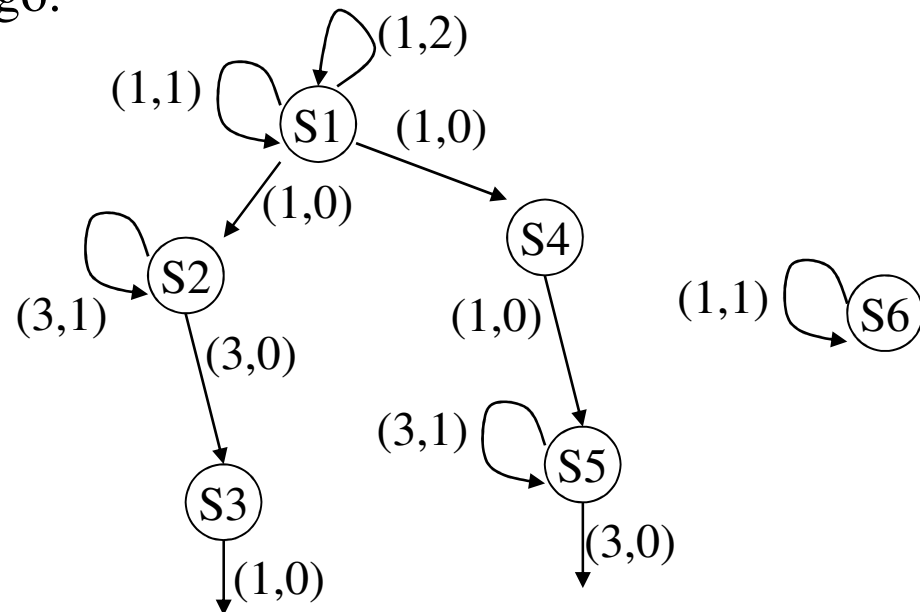
$$S_{\text{max}} = (NT_L)/T_{\text{crit}} = 9N/7N = 9/7 \approx 1,28$$

Segundo ejemplo

NOTA: Vale aclarar que lo que realmente se busca son las distintas trazas de código desde alguna instrucción de las primeras iteraciones hasta alguna instrucción de las últimas iteraciones

Sea el siguiente fragmento de código:

```
For (i=2; i<=6 ; i++) {  
    S1 : A(i) = A(i-1) + A(i-2)  
    S2 : B(i) = A(i) * B(i-1)  
    S3 : C(i) = B(i) + X  
    S4 : D(i) = A(i) + Y  
    S5 : E(i) = E(i-1) * D(i)  
    S6 : F(i) = F(i-1) + 5  
}
```



Segundo ejemplo

Las trazas son:

C1: $N * Ts1 + Ts2 + Ts3$	$5*1+3+1 = 9$
C2: $(N/2) * Ts1 + Ts2 + Ts3$	$2,5*1+3+1 = 6,5$
C3: $Ts1 + N * Ts2 + Ts3$	$1+5*3+1 = 17$
C4: $N * Ts1 + Ts4 + Ts5$	$5*1+1+3 = 9$
C5: $(N/2) * Ts1 + Ts4 + Ts5$	$2,5*1+1+3 = 6,5$
C6: $Ts1 + Ts4 + N * Ts5$	$1+1+5*3 = 17$
C7: $N * Ts6$	$5*1 = 5$

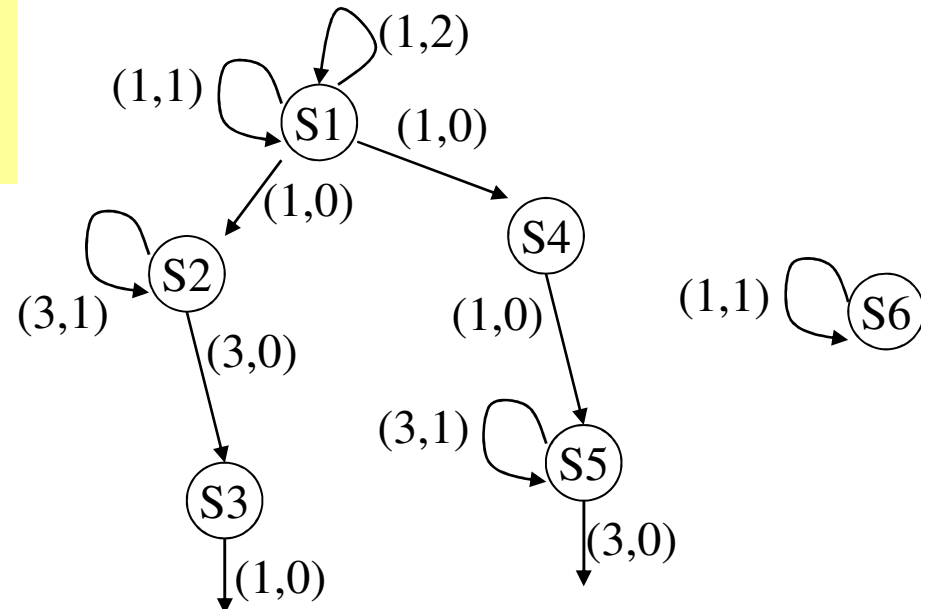
Por ejemplo:

Para C1, se calcula:

$S1(2) - S1(3) - S1(4) - S1(5) - S1(6) - S2(6) - S3(6)$

Para C3:

$S1(2) - S2(2) - S2(3) - S2(4) - S2(5) - S2(6) - S3(6)$



Tercer ejemplo

Dado el código a continuación, arme los grafos de dependencia de datos que detecte, muestre todas las trazas relevantes y calcule la aceleración que la traza crítica produce por sobre la ejecución secuencial del bucle.

En cuántas unidades de cálculo podría ejecutarse el loop luego de optimizarlo?
(NOTA: se sugiere armar el gráfico de iteración espacial)

```
for (int i = 3; i < 1003; i++){  
S1:    A[i-1] = E[i-2] + D[i-3];  
S2:    B      = B * B + A[i-1];  
S3:    C[i]   = A[i] * B;  
S4:    D[i]   = C[i+2] / 2;  
S5:    E[i]   = C[i-1] * 5;  
}
```

Suponga los siguientes parámetros

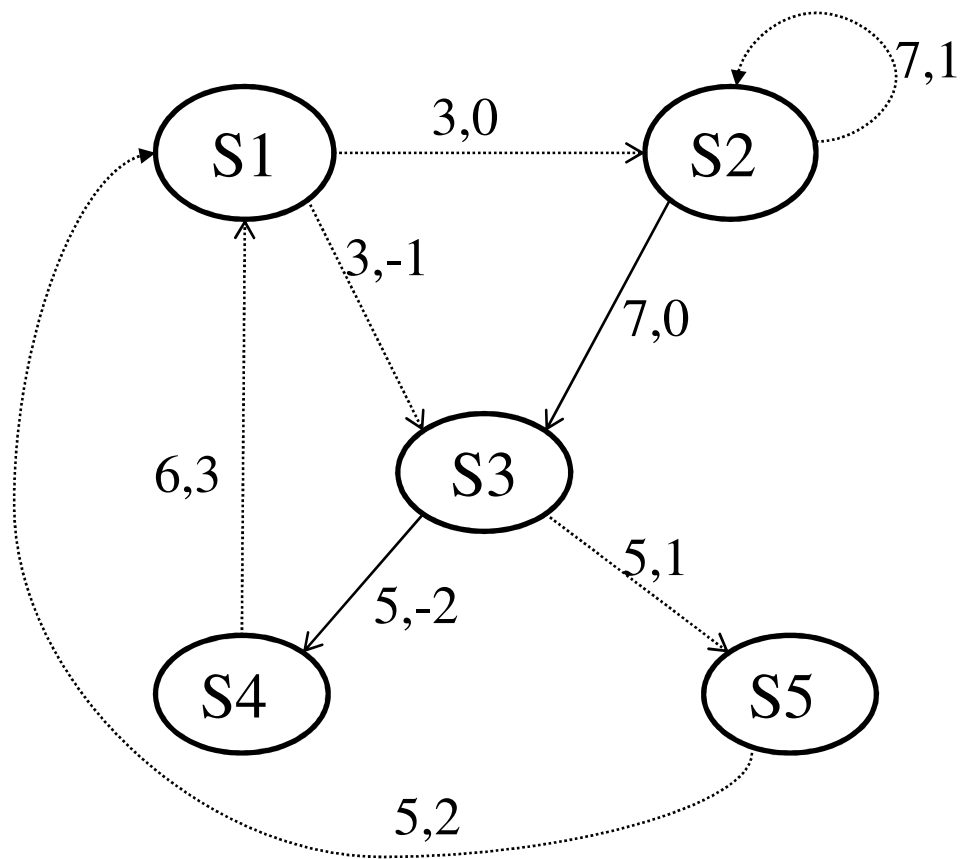
Suma: 2 ciclos

Multiplicación: 4 ciclos

División: 5 ciclos

Asignación: 1 ciclo

Solución



Grafo temporal

Gráfico de iteración espacial

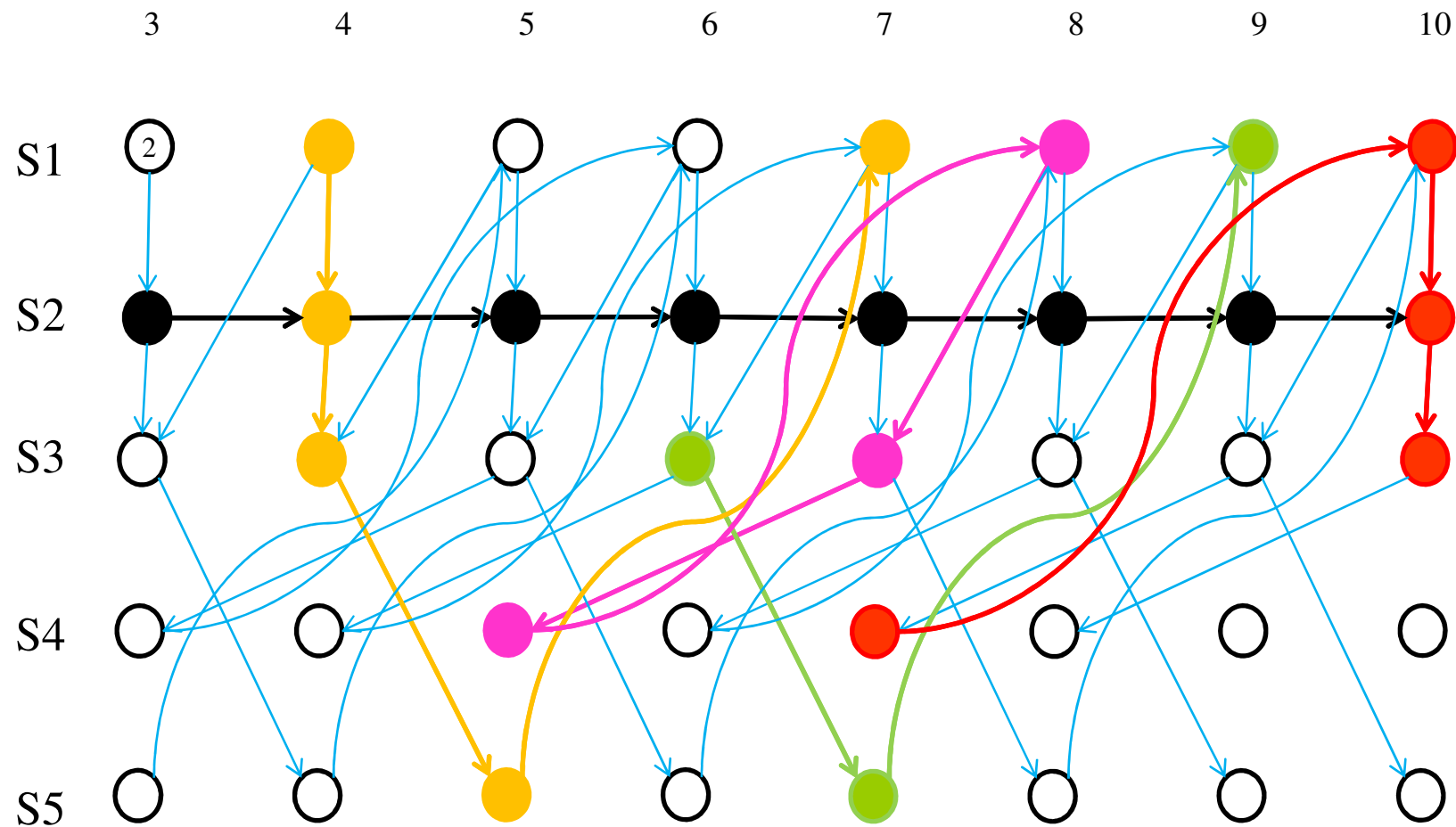


Gráfico de iteración espacial

$S1(i) - S3(i-1) - S5(i) - S1(i+2) = (3+5+5) = 13$ (hay N de estas)

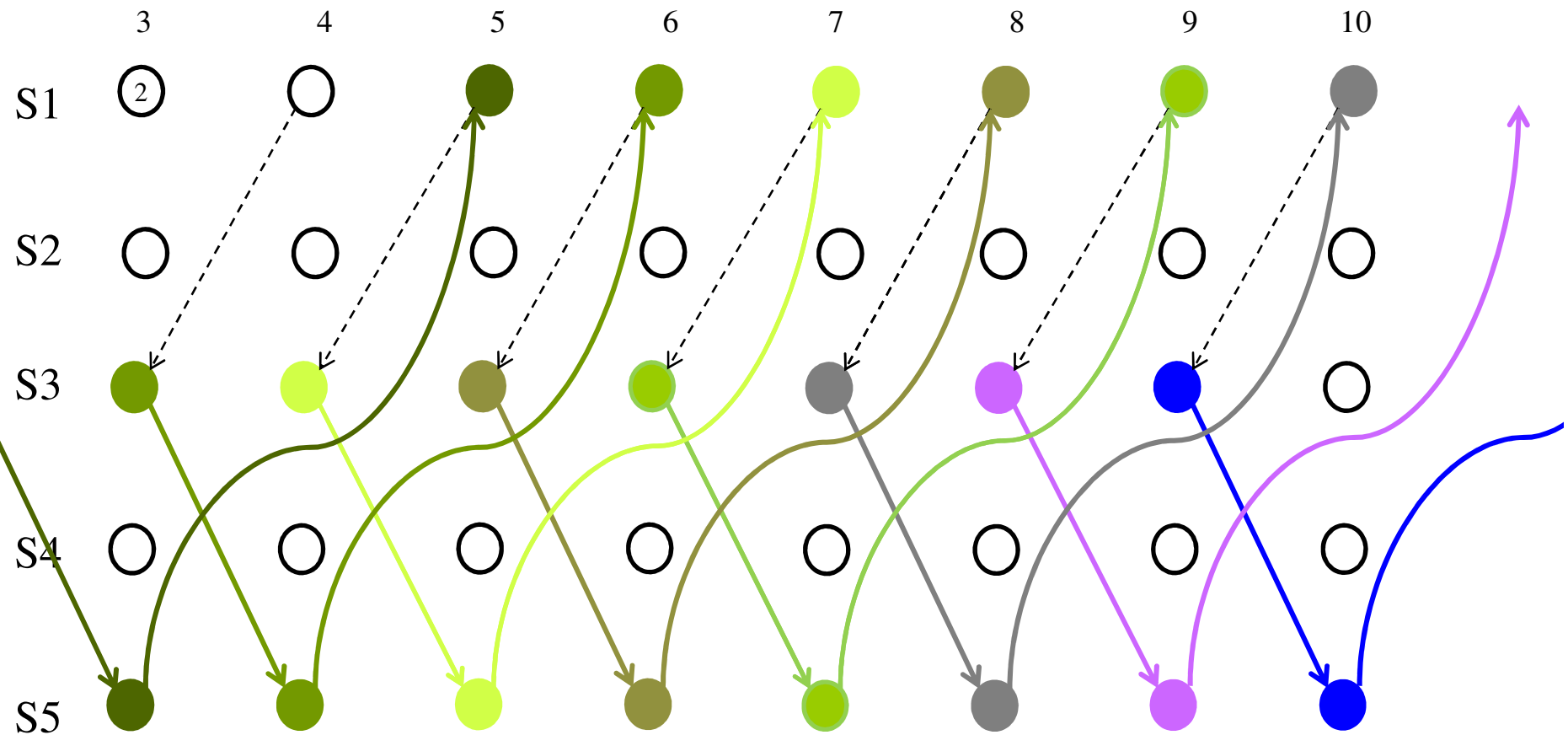


Gráfico de iteración espacial

$S1(i) - S3(i-1) - S4(i-3) - S1(i) = (3+0+6) = 9$ (hay N de estas)

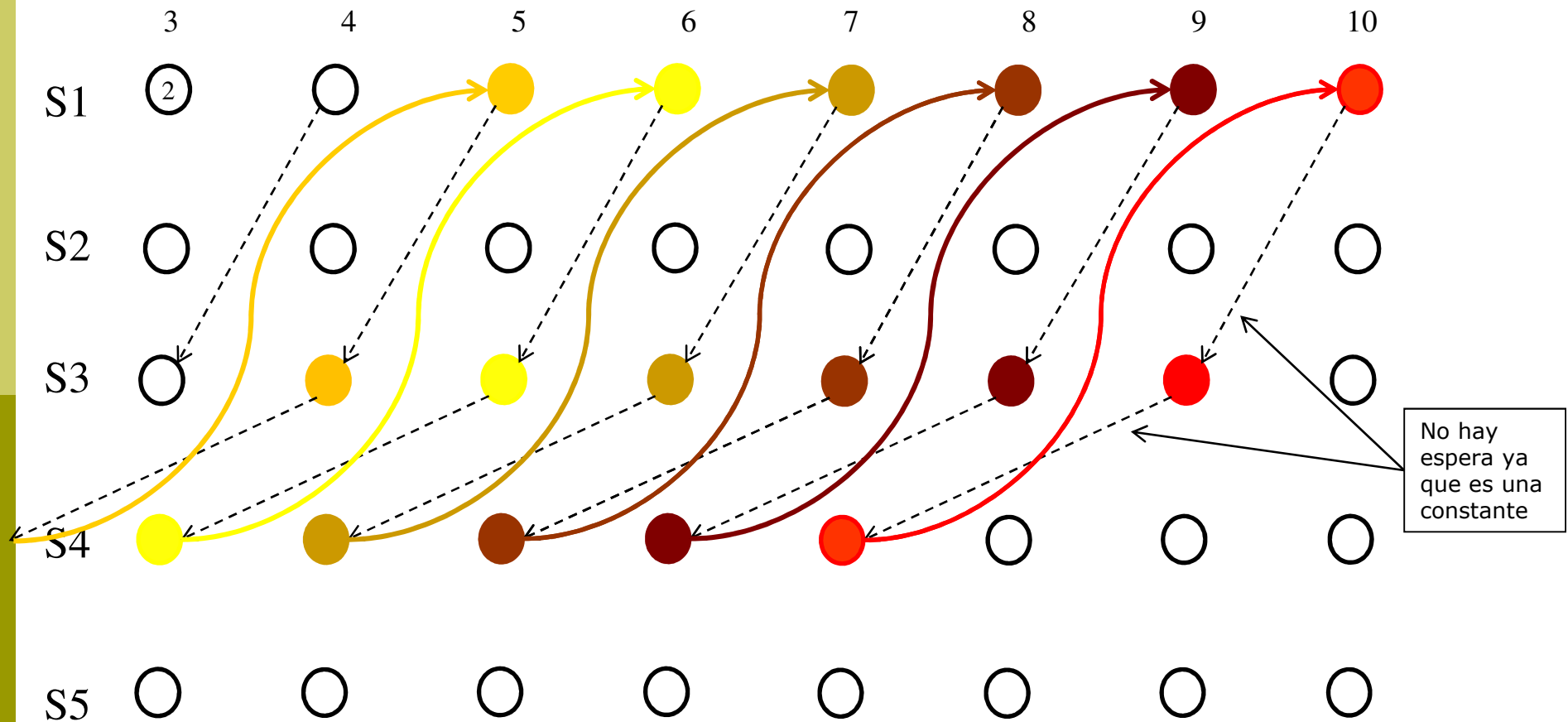


Gráfico de iteración espacial

$S1(i) - S2(i) - S3(i) - S4(i-2) - S1(i+1) = (3+7+5+6) = 21$ (hay N de estas)

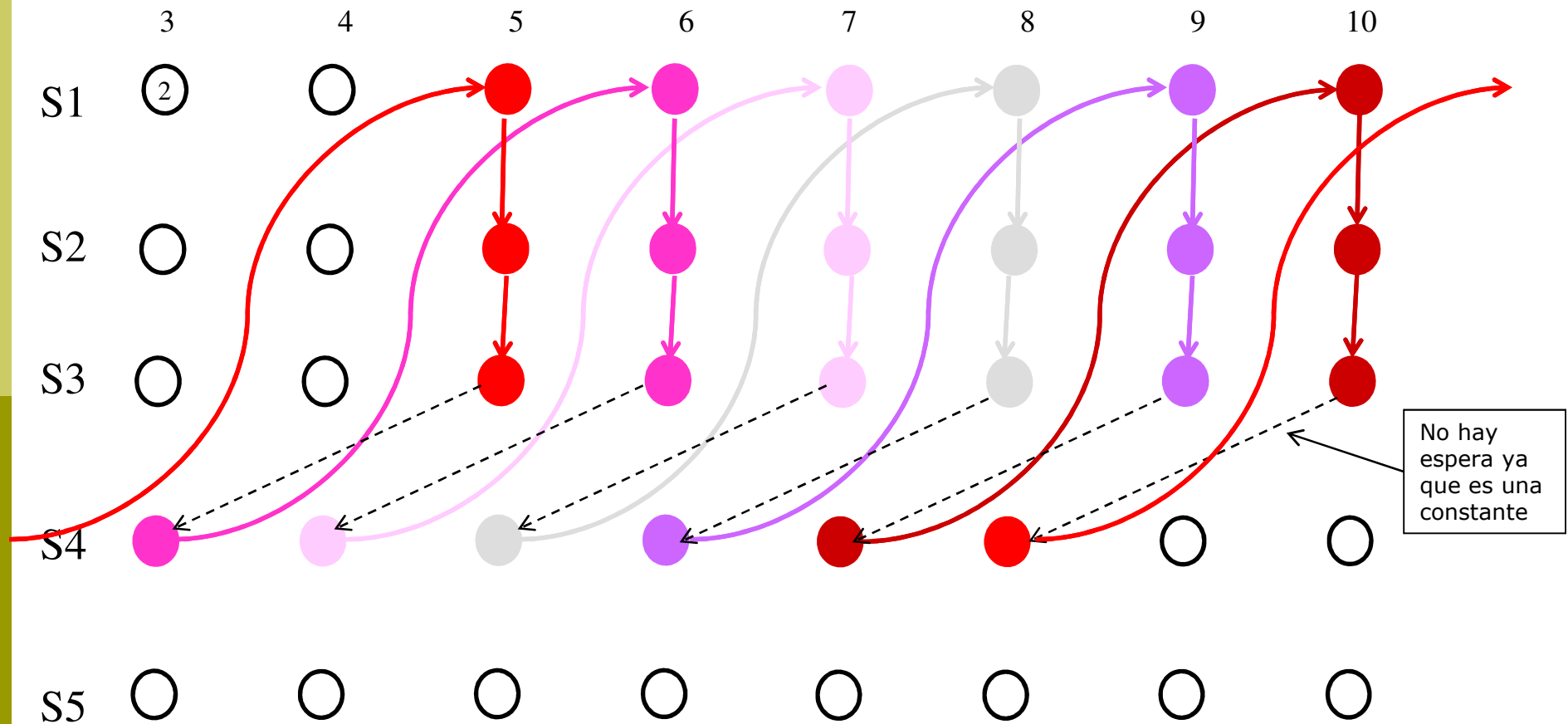


Gráfico de iteración espacial

$$S1(i) - S2(i) - S3(i) - S5(i+1) - S1(i+3) = (3+7+5+5) * (1000/3) = 6666.66$$

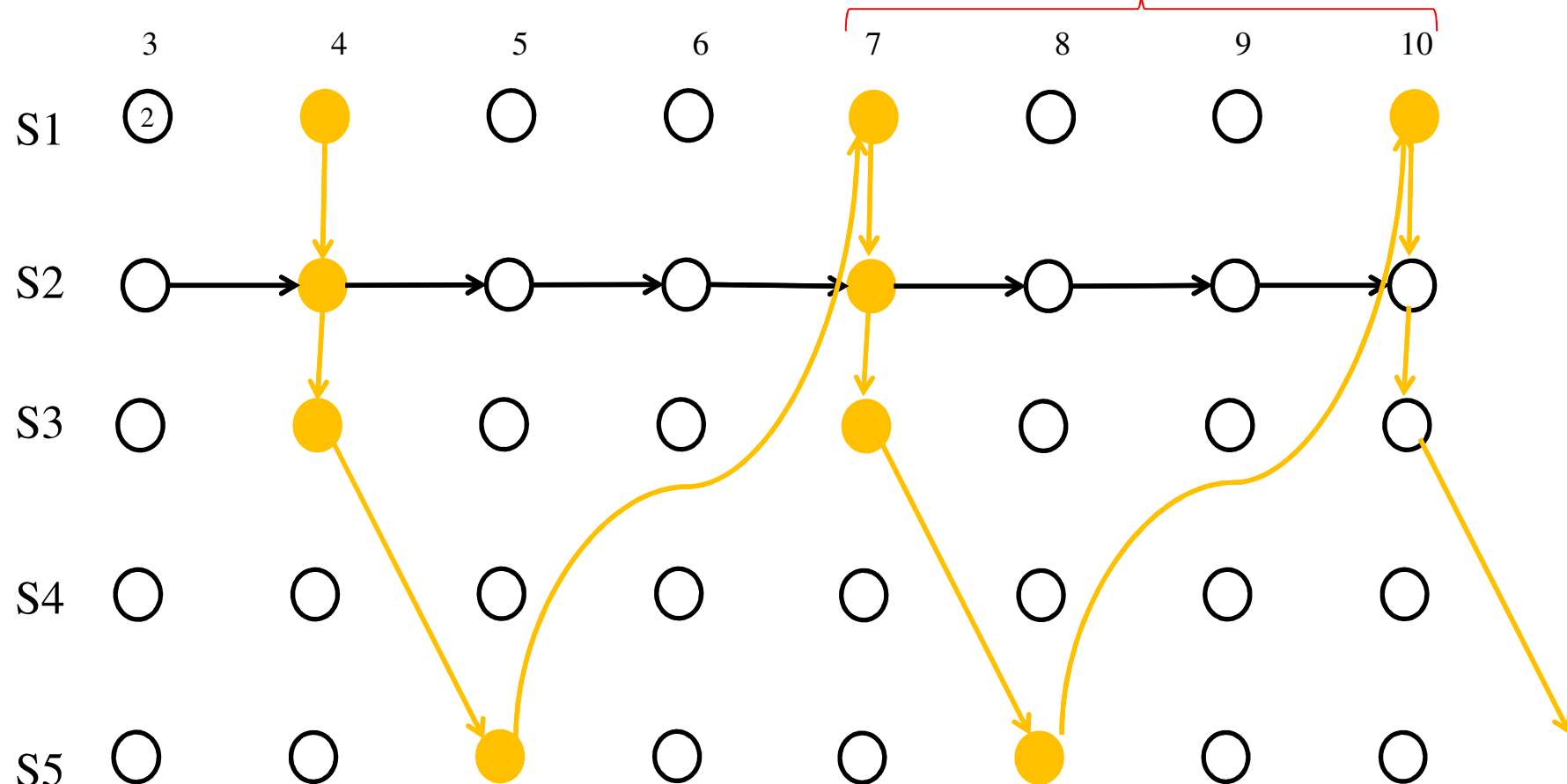
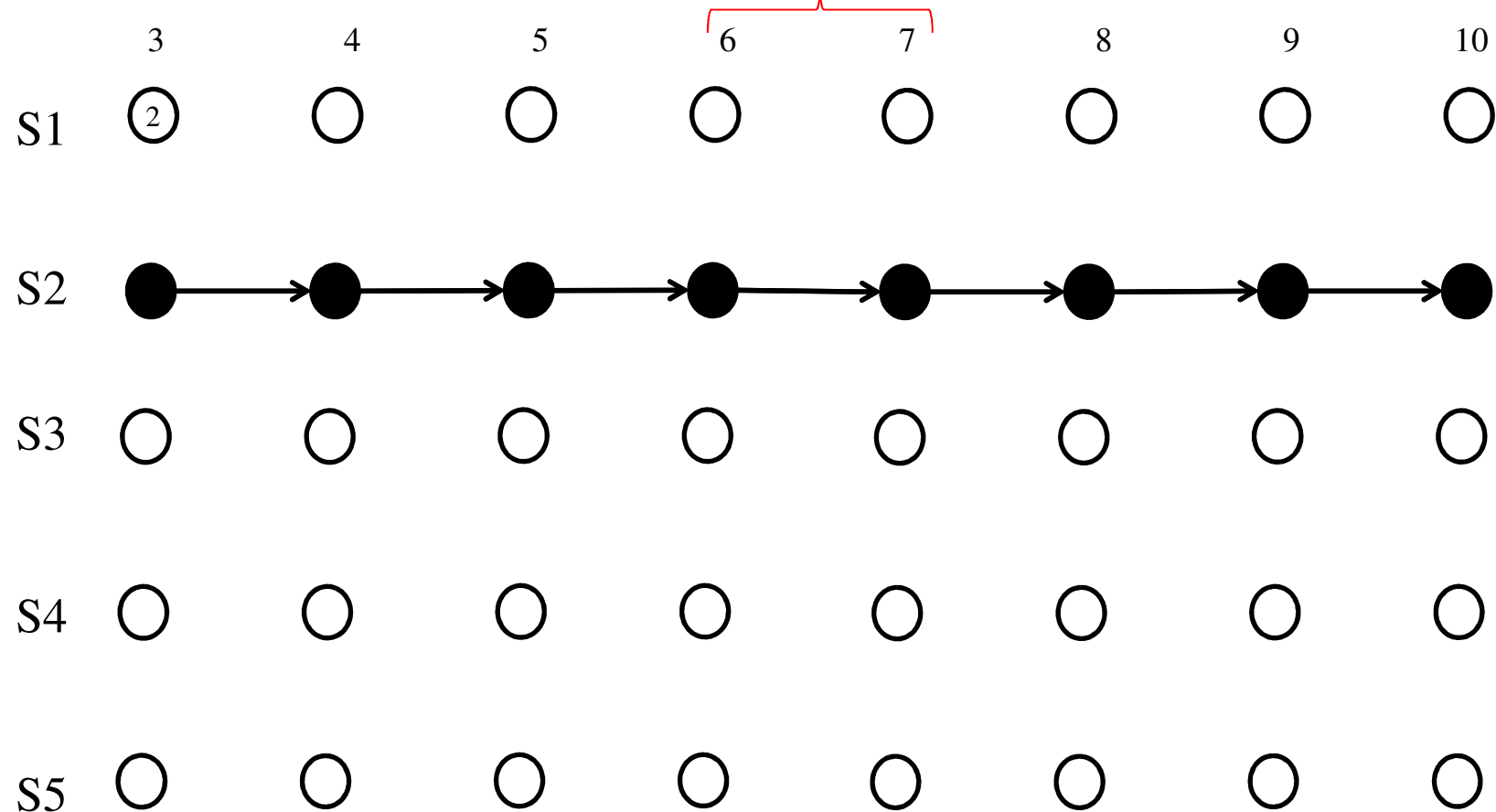


Gráfico de iteración espacial

- $S2(i) - S2(i+1) = (7) * (1000/1) = \mathbf{7000}$



Trazas

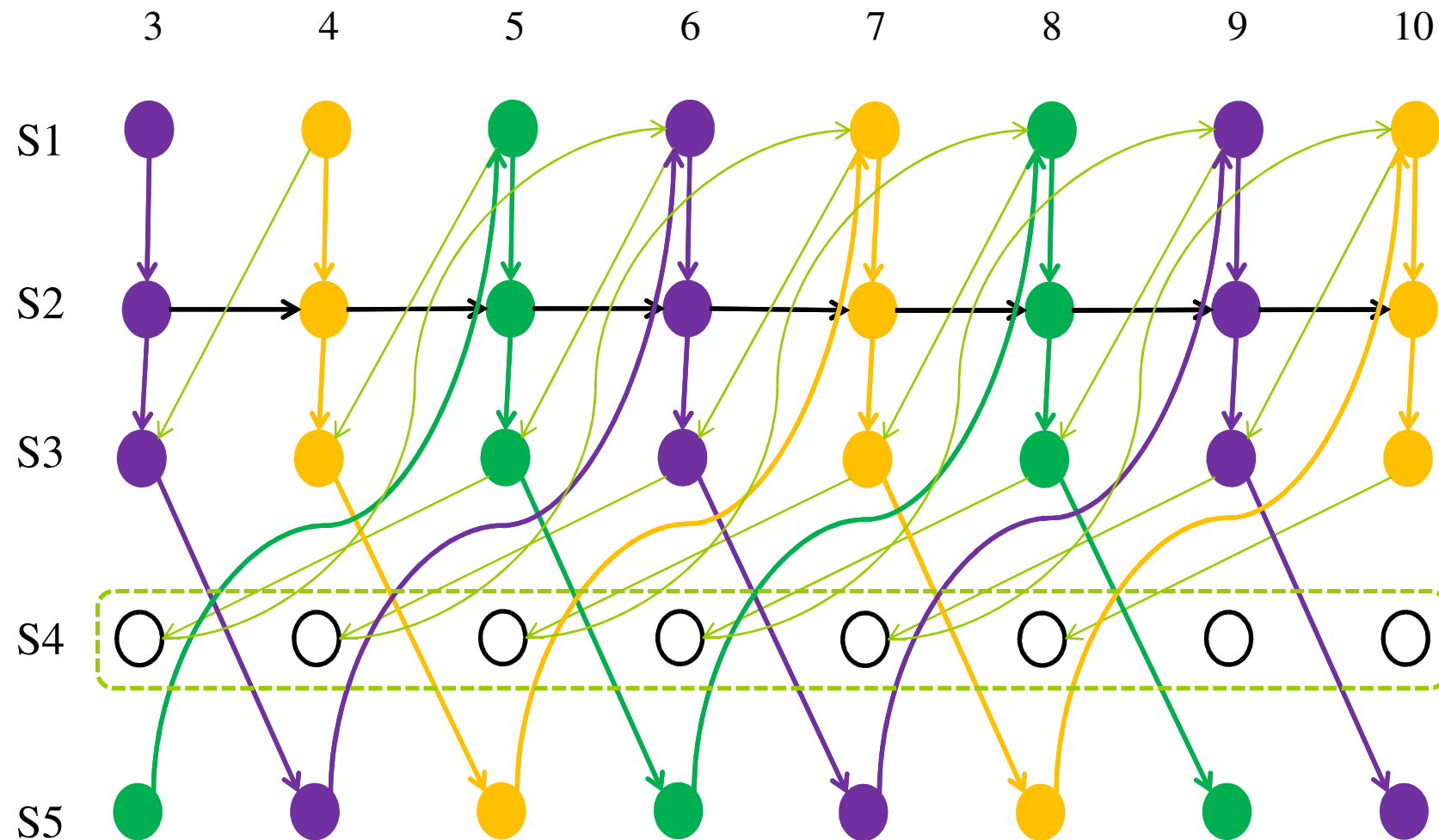
- $S1(i) - S3(i-1) - S5(i) - S1(i+2) = (3+5+5) = 13$ (hay N de estas)
- $S1(i) - S3(i-1) - S4(i-3) - S1(i) = (3+0+6) = 9$ (hay N de estas)
- $S1(i) - S2(i) - S3(i) - S4(i-2) - S1(i+1) = (3+7+5+6) = 21$ (hay N de estas)
- $S1(i) - S2(i) - S3(i) - S5(i+1) - S1(i+3) = (3+7+5+5)*(1000/3) = 6666.66$
- $S2(i) - S2(i+1) = (7)*(1000/1) = \mathbf{7000}$

$$\text{Aceleración} = TL/TT_{\text{crit}} = (26*1000)/ 7000 = \mathbf{3.714}$$

NOTA: Las 3 primeras trazas (ayudarse con filmina 24) no recorren todas las iteraciones. Si se presta atención son muchas trazas iguales de 3 ó 4 instrucciones. Pe, la traza naranja tiene las instrucciones S1(I), S2(I), S3(I), S5(I+1) y cierra en S1(I+3). De esa traza hay N iguales, por ejemplo: S1(7)-S2(7)-S3(7)-S5(8)-S1(10); S1(20)-S2(20)-S3(20)-S5(21)-S1(23), S1(51)-S2(51)-S3(51)-S5(52)-S1(54), etc. Como las 3 puestas de ejemplo hay N trazas iguales pero cortitas.

En cuantos procesadores se puede hacer??

Ejecución en procesadores



Se puede ejecutar en 4 procesadores

Problema 1...

Para el siguiente fragmento de código,

1. arme los grafos de dependencia de datos que detecte,
2. muestre todas las trazas existentes y
3. calcule la aceleración que la traza crítica produce por sobre la ejecución secuencial del bucle.
4. En cuántos procesadores podría ejecutarse el loop luego de optimizarlo?

```
for (int i = 3; i < 1003; i++){  
S1:   A[i]    = E[i-2] + D[i-3];  
S2:   B       = B * B + A[i-1];  
S3:   C[i-1] = A[i] * B;  
S4:   D[i]    = C[i+2] / 2;  
S5:   E[i]    = C[i-1] * 5;
```

Suponga los siguientes parámetros
Suma: 2 ciclos
Multiplicación: 4 ciclos
División: 5 ciclos
Asignación: 1 ciclo