Estructura de Computadores

Tema 6. Memoria Cache

Requisitos del sistema de memoria

- El programador requiere memoria que sea...
  - rápida
  - de gran capacidad
  - barata

Requisitos del sistema de memoria

- El programador requiere memoria que sea...
  - rápida
  - de gran capacidad
  - barata
- Ninguna tecnología cumple con todos los requisitos

Tecnología de memoria	Tiempo de acceso	$ por GB en 2008
SRAM	0.5-2.5 ns	$2000-$5000
DRAM	50-70 ns	$20-$75
Disco magnético	5,000,000-20,000,000 ns	$0.20-$2

Requisitos del sistema de memoria

- El programador requiere memoria que sea...
  - rápida
  - de gran capacidad
  - barata
- Ninguna tecnología cumple con todos los requisitos

Tecnología de memoria	Tiempo de acceso	$ por GB en 2008
SRAM	0.5-2.5 ns	$2000-$5000
DRAM	50-70 ns	$20-$75
Disco magnético	5,000,000-20,000,000 ns	$0.20-$2

- Podemos crear la ilusión de una gran cantidad de memoria muy rápida combinando múltiples tecnologías, gracias al principio de localidad

Principio de localidad

- Los programas acceden a una porción relativamente pequeña del espacio de direcciones en cada instante de tiempo
  - No todas las direcciones de memoria tienen la misma probabilidad de ser accedidas

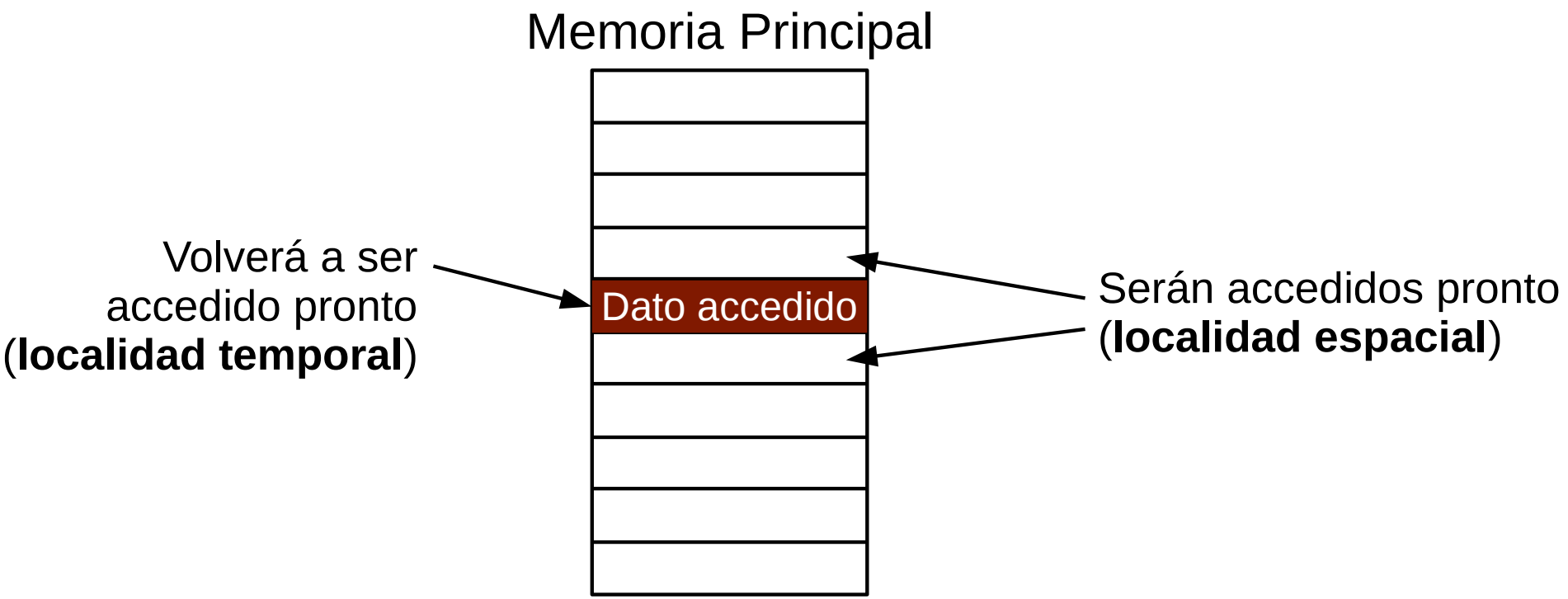
Principio de localidad

- Los programas acceden a una porción relativamente pequeña del espacio de direcciones en cada instante de tiempo
  - No todas las direcciones de memoria tienen la misma probabilidad de ser accedidas
- **Localidad temporal**
  - Cuando un programa accede a un dato, existe una elevada probabilidad de que ese mismo dato vuelva a ser accedido pronto

Principio de localidad

- Los programas acceden a una porción relativamente pequeña del espacio de direcciones en cada instante de tiempo
  - No todas las direcciones de memoria tienen la misma probabilidad de ser accedidas
- **Localidad temporal**
  - Cuando un programa accede a un dato, existe una elevada probabilidad de que ese mismo dato vuelva a ser accedido pronto
- **Localidad espacial**
  - Cuando un programa accede a un dato, existe una elevada probabilidad de que los datos cercanos sean accedidos pronto

Principio de localidad



Principio de localidad

- Supongamos que un programa accede a las siguientes direcciones de memoria:

Acceso	Dirección de memoria
1	0
2	1024
3	0
4	1024
5	0
6	1024
7	0
8	1024

- ¿Qué tipo de localidad tiene este programa?

Principio de localidad

- Supongamos que un programa accede a las siguientes direcciones de memoria:

Acceso	Dirección de memoria
1	0
2	4
3	8
4	12
5	16
6	20
7	24
8	28

- ¿Qué tipo de localidad tiene este programa?

Principio de localidad

- Supongamos que un programa accede a las siguientes direcciones de memoria:

Acceso	Dirección de memoria
1	0
2	4
3	8
4	12
5	0
6	4
7	8
8	12

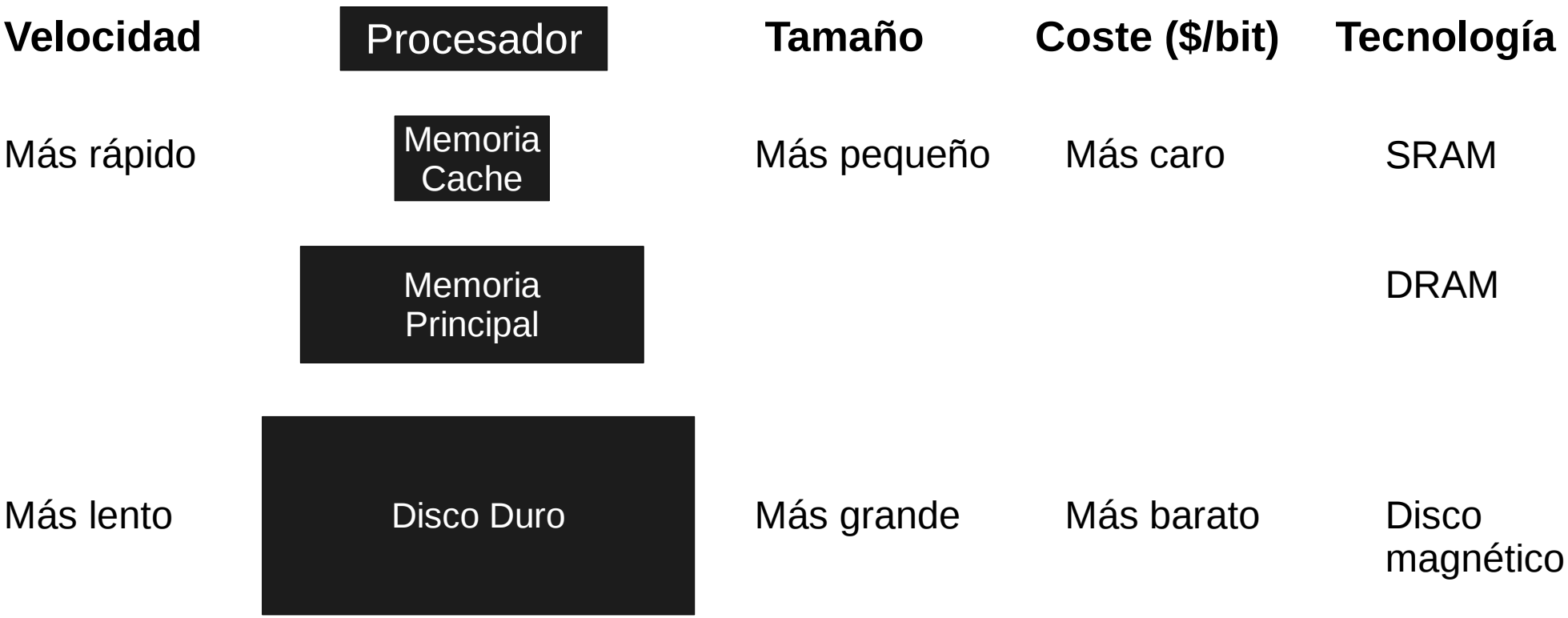
- ¿Qué tipo de localidad tiene este programa?

Principio de localidad

- La localidad aperece de forma natural en los programas
- La mayoría de progamas contienen bucles
  - Es muy probable que los mismos datos sean accedidos de forma repetida en cada iteración del bucle (localidad temporal)
- Muchos progamas acceden a vectores
  - El acceso a elementos consecutivos de un vector exhibe localidad espacial

Jerarquía de memoria

- Podemos aprovechar el principio de localidad para implementar el sistema de memoria como una jerarquía de niveles

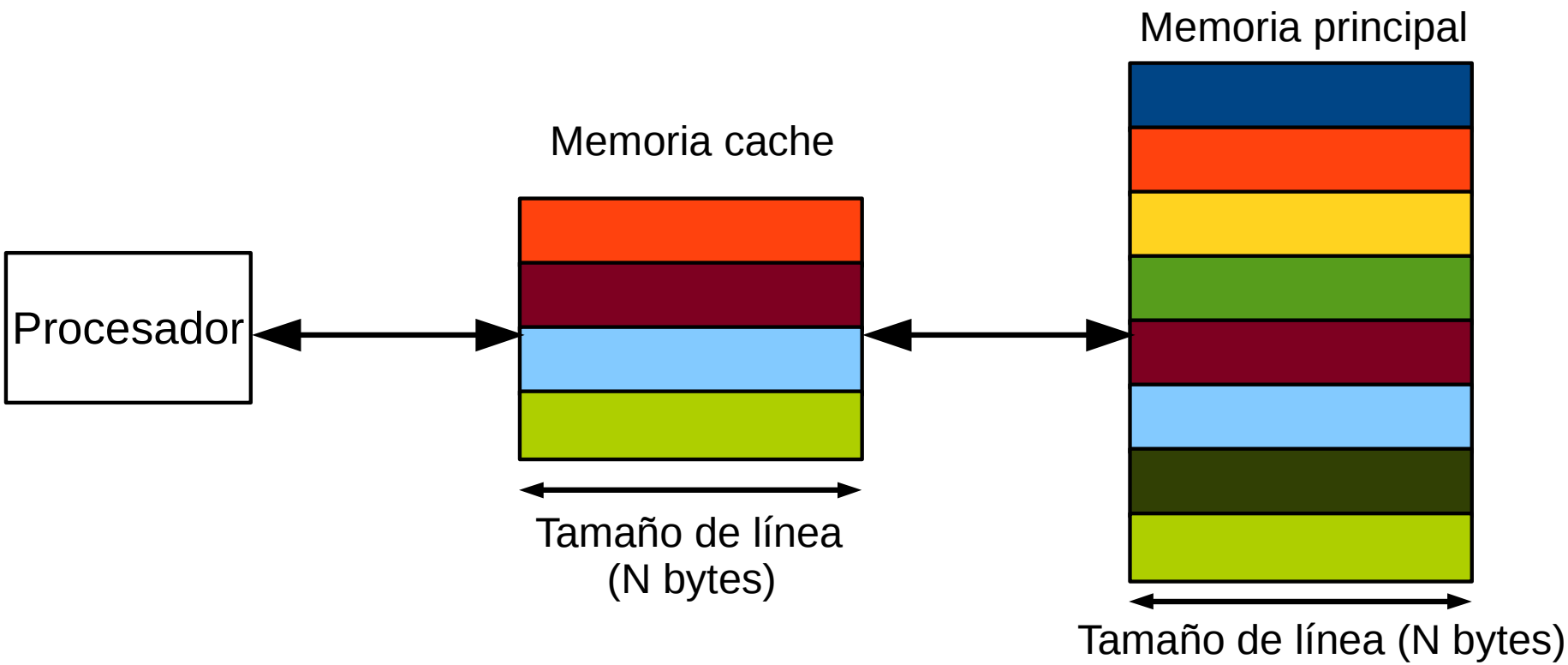


Jerarquía de memoria

- La memoria cache contiene los datos más recientemente accedidos
  - Tamaño pequeño para **reducir el coste**
  - Gran **velocidad de acceso**
  - La mayoría de accesos son servidos por la memoria cache (localidad)
- El sistema de memoria ofrece **gran capacidad** mediante el uso de DRAM (memoria principal) y el disco magnético
- Cuanto más lejos del procesador, mayor capacidad pero mayor tiempo de acceso

Jerarquía de memoria

- El espacio de direcciones de memoria se divide en bloques o líneas de memoria
  - Una línea de memoria es la mínima unidad de información que puede estar o no estar presente en la memoria cache



Memoria cache

- Hit (Acierto)
  - El dato accedido por el procesador está presente en la memoria cache

Memoria cache

- Hit (Acierto)
  - El dato accedido por el procesador está presente en la memoria cache
- Hit Ratio (Tasa de aciertos)
  - Porcentaje de accesos a memoria que encuentran los datos en la cache (porcentaje de aciertos)

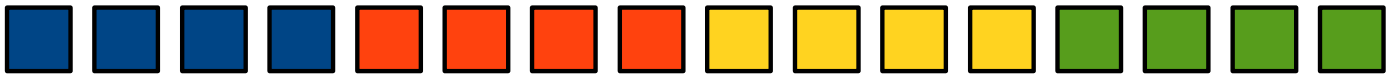
Memoria cache

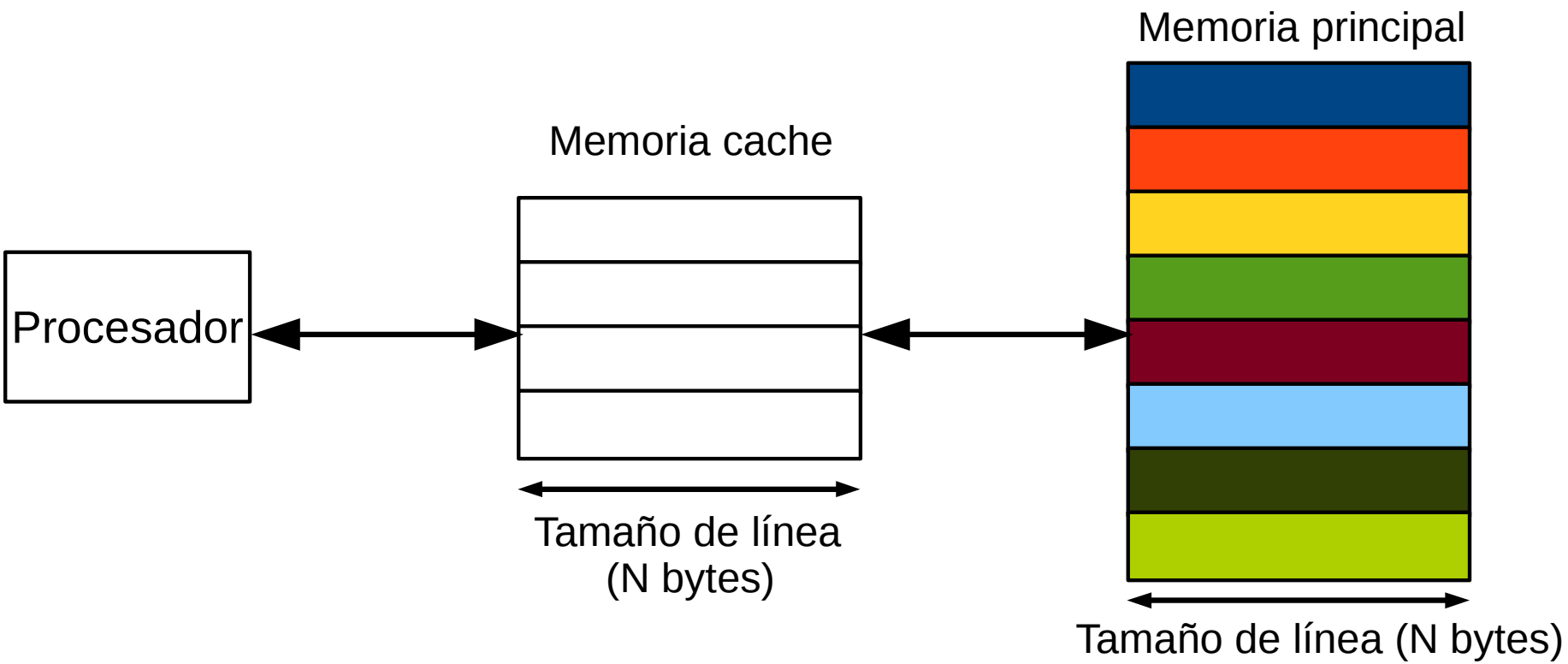
- Hit (Acierto)
  - El dato accedido por el procesador está presente en la memoria cache
- Hit Ratio (Tasa de aciertos)
  - Porcentaje de accesos a memoria que encuentran los datos en la cache (porcentaje de aciertos)
- Miss (Fallo)
  - El dato accedido por el procesador no está presente en la memoria cache
  - Hay que traer la línea de memoria principal

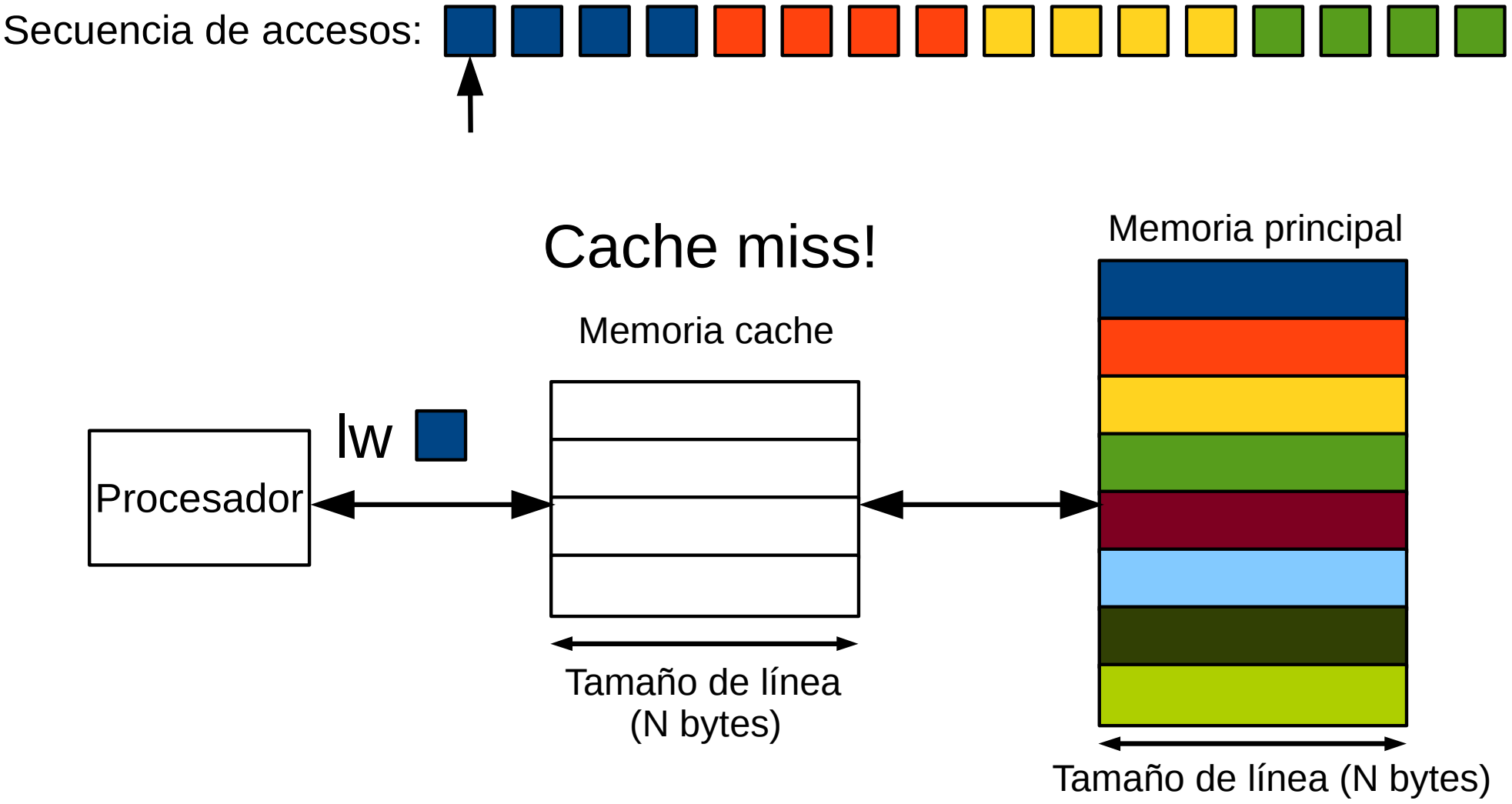
Memoria cache

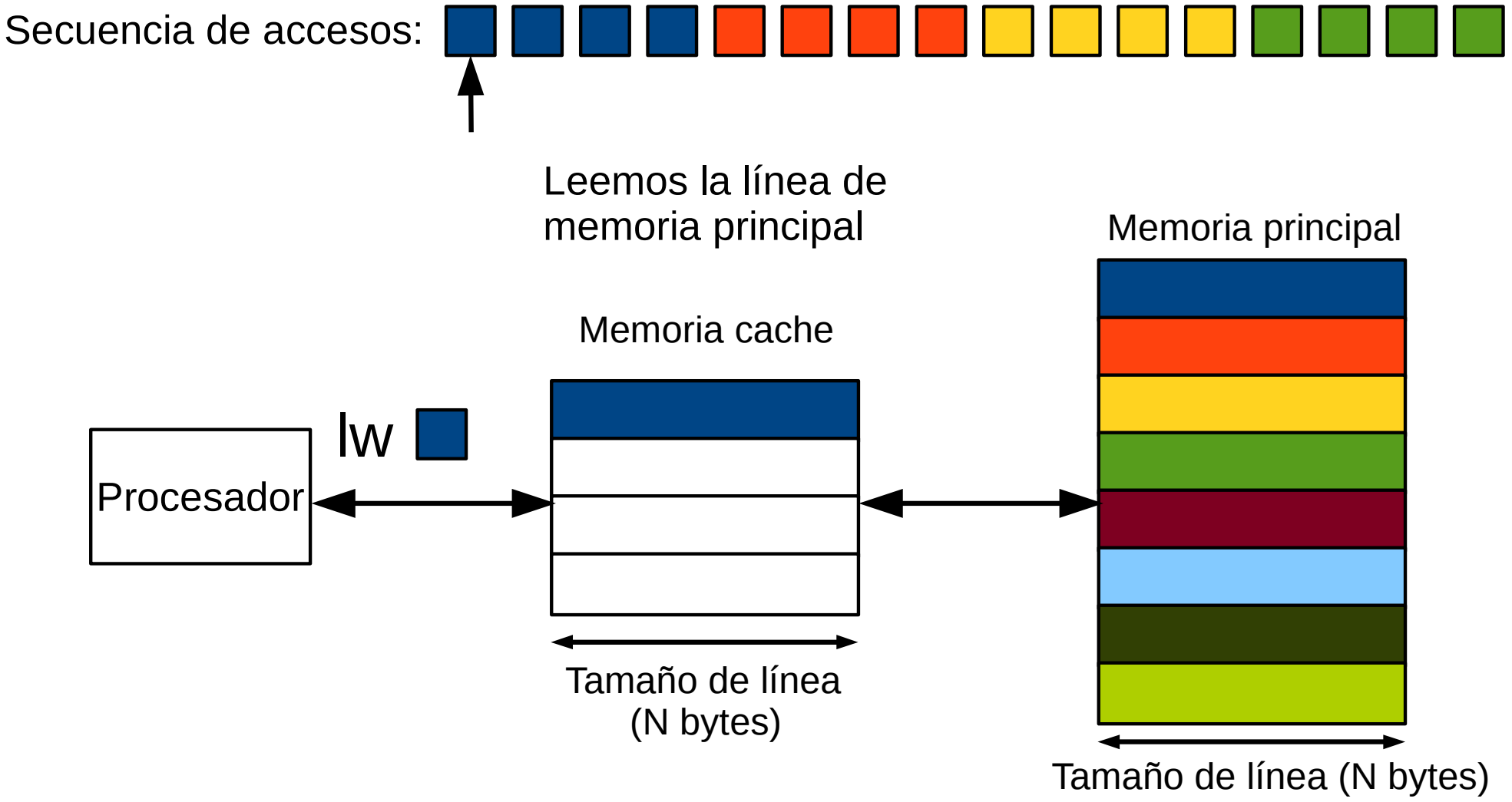
- Hit (Acierto)
  - El dato accedido por el procesador está presente en la memoria cache
- Hit Ratio (Tasa de aciertos)
  - Porcentaje de accesos a memoria que encuentran los datos en la cache (porcentaje de aciertos)
- Miss (Fallo)
  - El dato accedido por el procesador no está presente en la memoria cache
  - Hay que traer la línea de memoria principal
- Miss Ratio (Tasa de fallos)
  - Porcentaje de accessos a memoria que **NO** encuentran los datos en la cache (porcentaje de fallos)

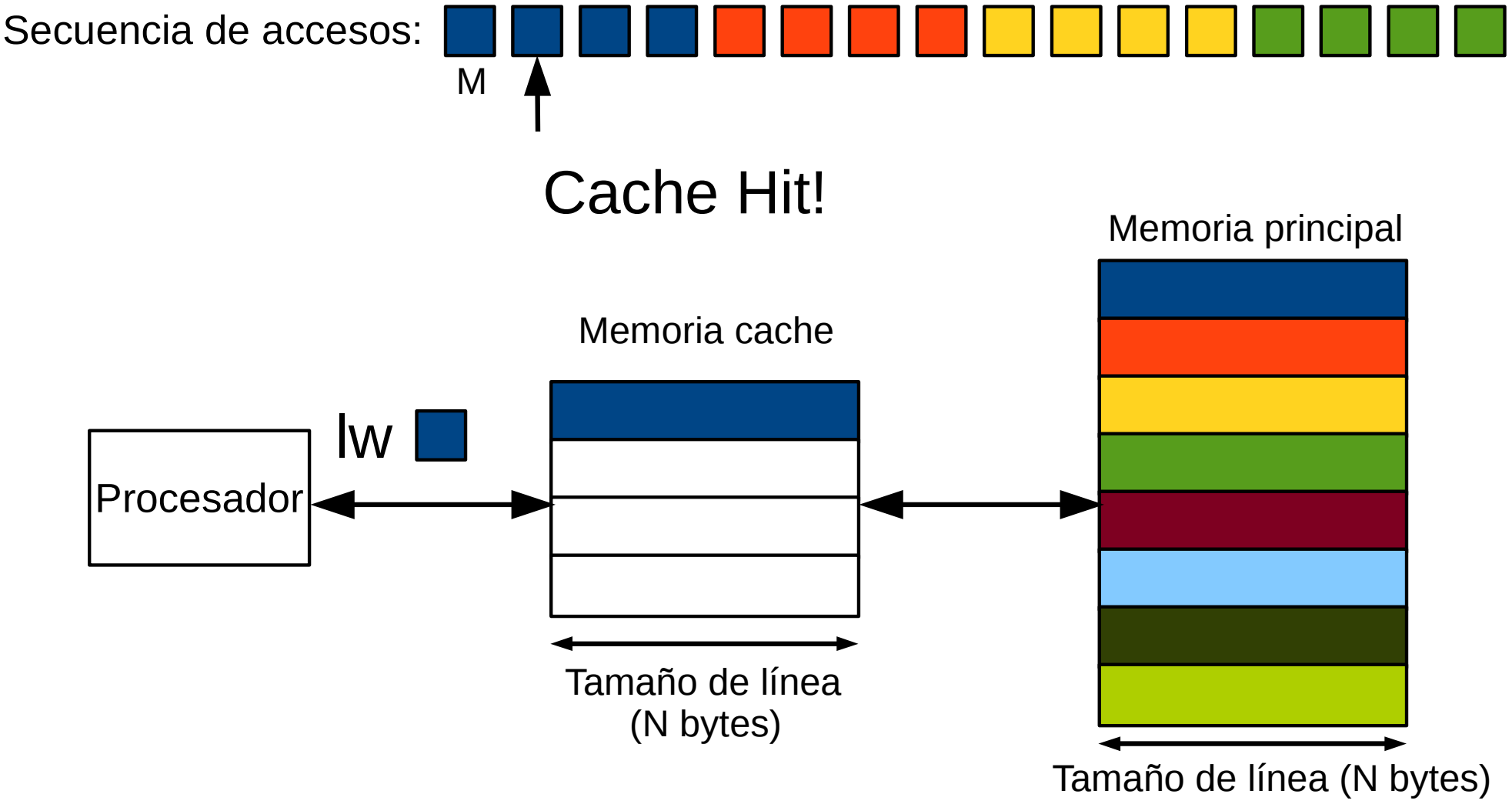
Secuencia de accesos:

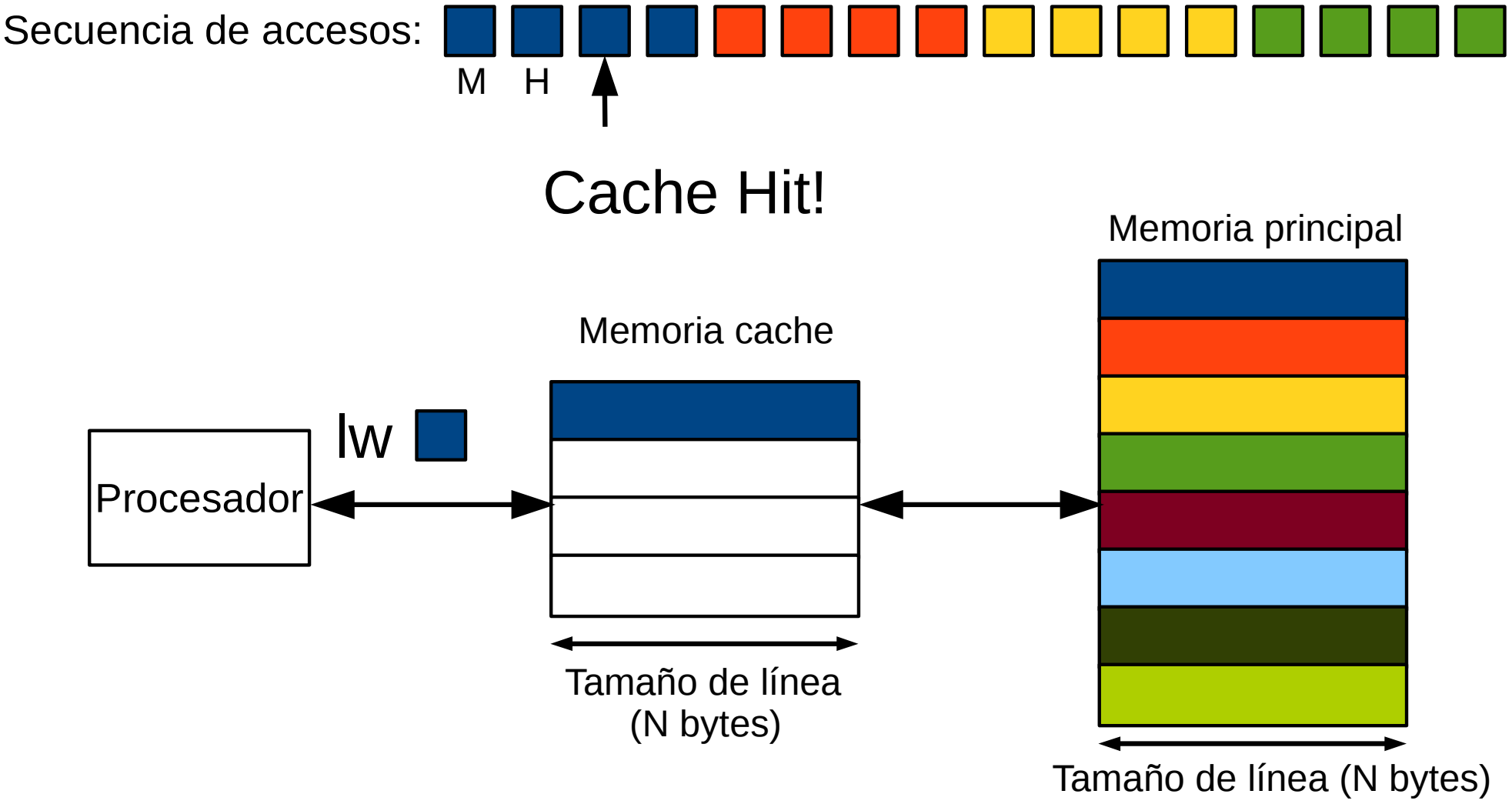


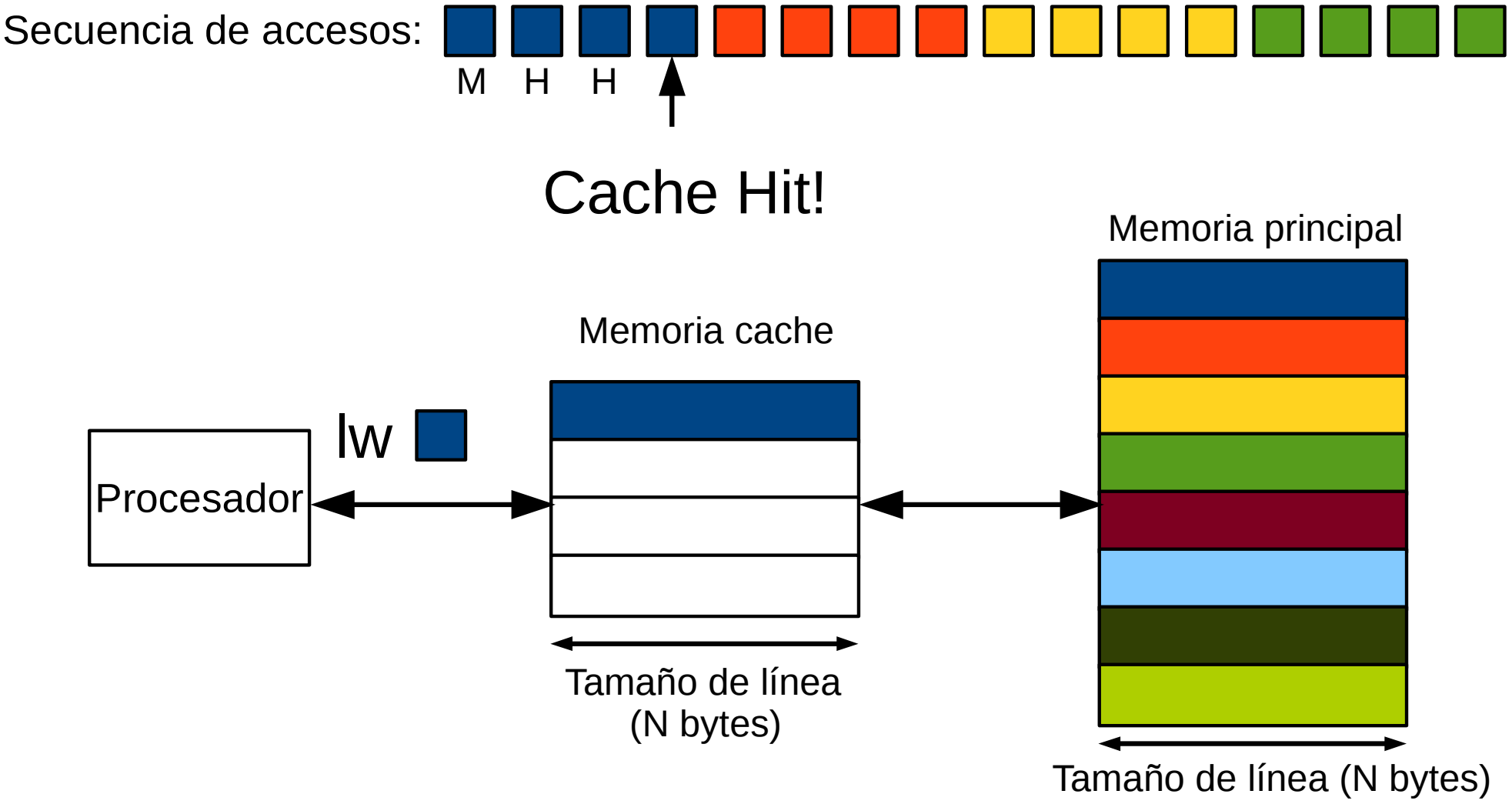


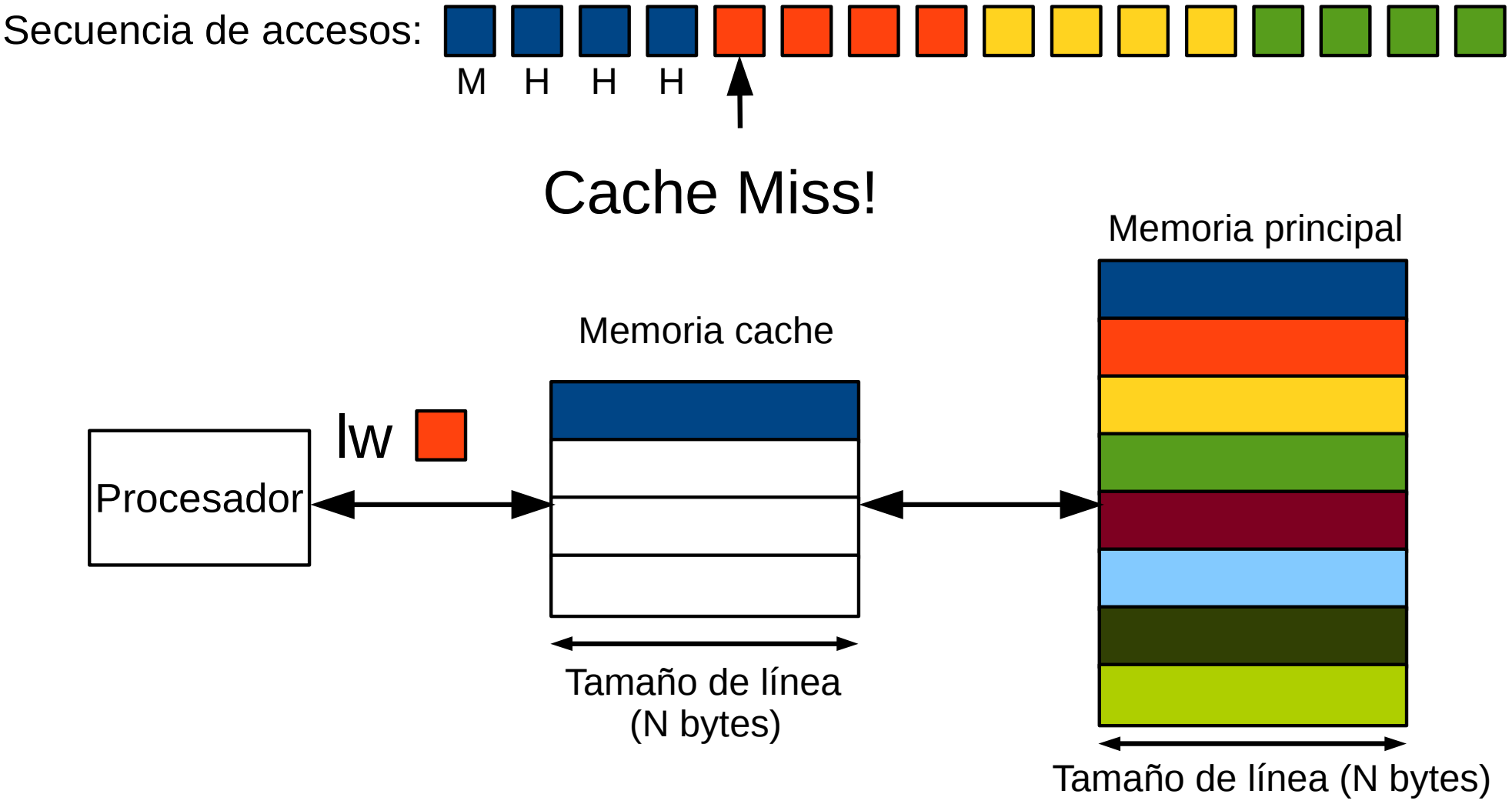

Secuencia de accesos:
Cache miss!
Memoria principal
Memoria cache
lw
Procesador
Tamaño de línea
(N bytes)
Tamaño de línea (N bytes)

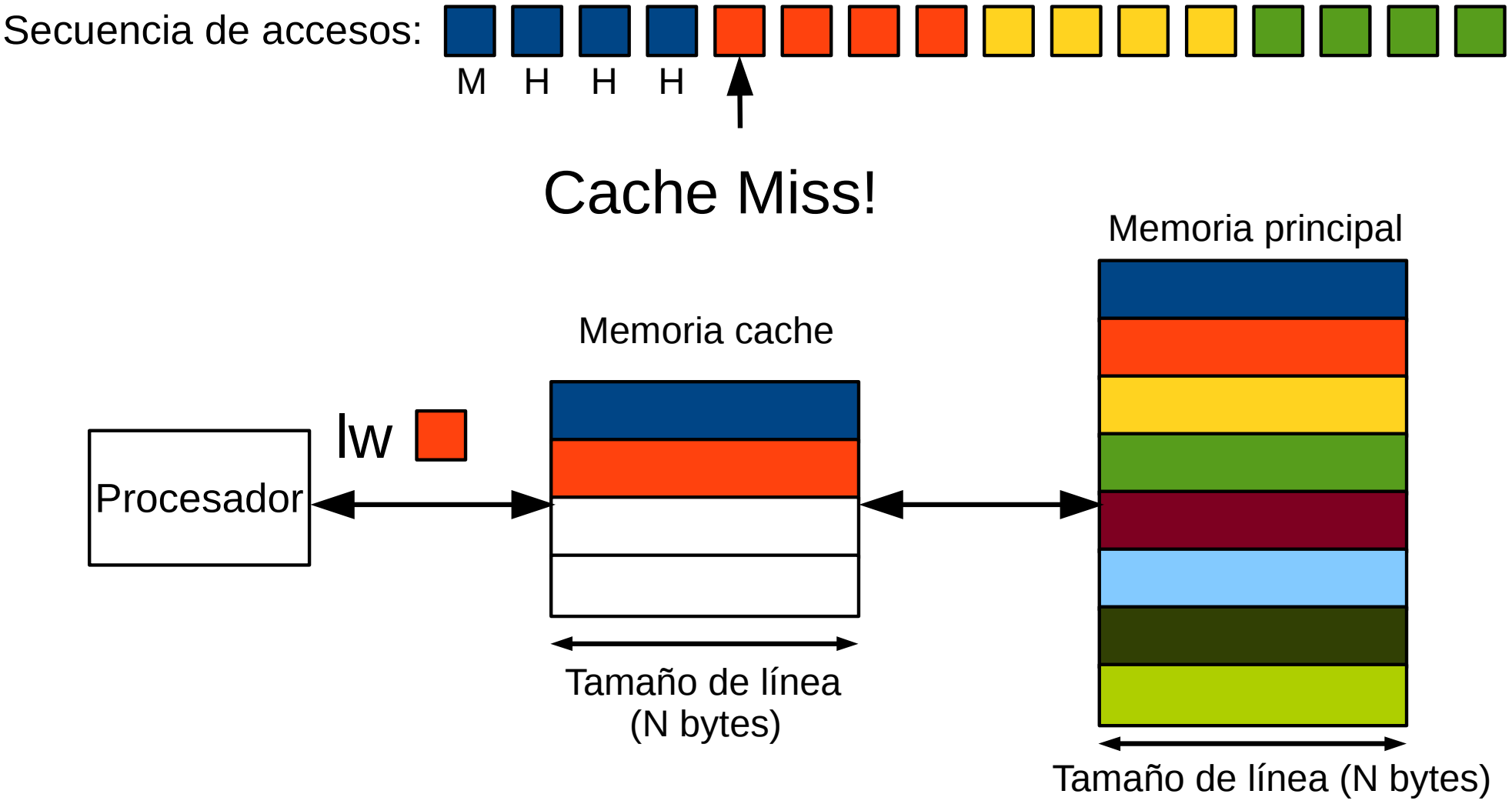

Secuencia de accesos:
Leemos la línea de memoria principal
Memoria principal
Memoria cache
lw
Procesador
Tamaño de línea (N bytes)
Tamaño de línea (N bytes)

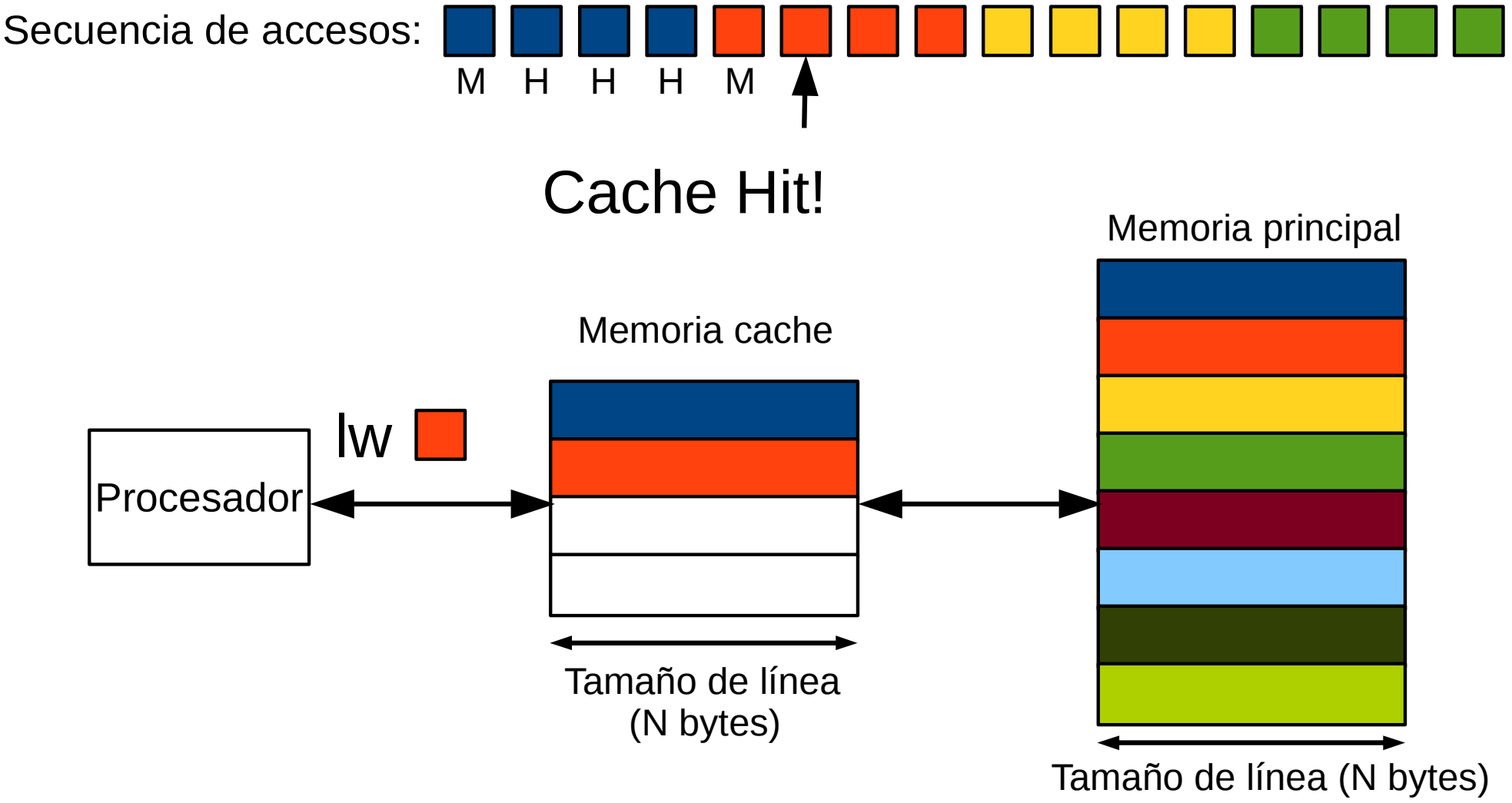

Secuencia de accesos:
M
Cache Hit!
Memoria principal
Memoria cache
lw
Procesador
Tamaño de línea
(N bytes)
Tamaño de línea (N bytes)


Secuencia de accesos:
M
H
Cache Hit!
Memoria principal
Memoria cache
lw
Procesador
Tamaño de línea
(N bytes)
Tamaño de línea (N bytes)


Secuencia de accesos:
M H H
Cache Hit!
Memoria principal
Memoria cache
lw
Procesador
Tamaño de línea
(N bytes)
Tamaño de línea (N bytes)


Secuencia de accesos:
M H H H
Cache Miss!
Memoria principal
Memoria cache
lw
Procesador
Tamaño de línea
(N bytes)
Tamaño de línea (N bytes)


Secuencia de accesos:
M H H H
Cache Miss!
Memoria principal
Memoria cache
lw
Procesador
Tamaño de línea
(N bytes)
Tamaño de línea (N bytes)


Secuencia de accesos:
M H H H M
Cache Hit!
Memoria principal
Memoria cache
lw
Procesador
Tamaño de línea
(N bytes)
Tamaño de línea (N bytes)

Secuencia de accesos:

Tasa de aciertos = 75%
Tasa de fallos = 25%