Simplified Cache Demo

For summation of 2D array

2D Array

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

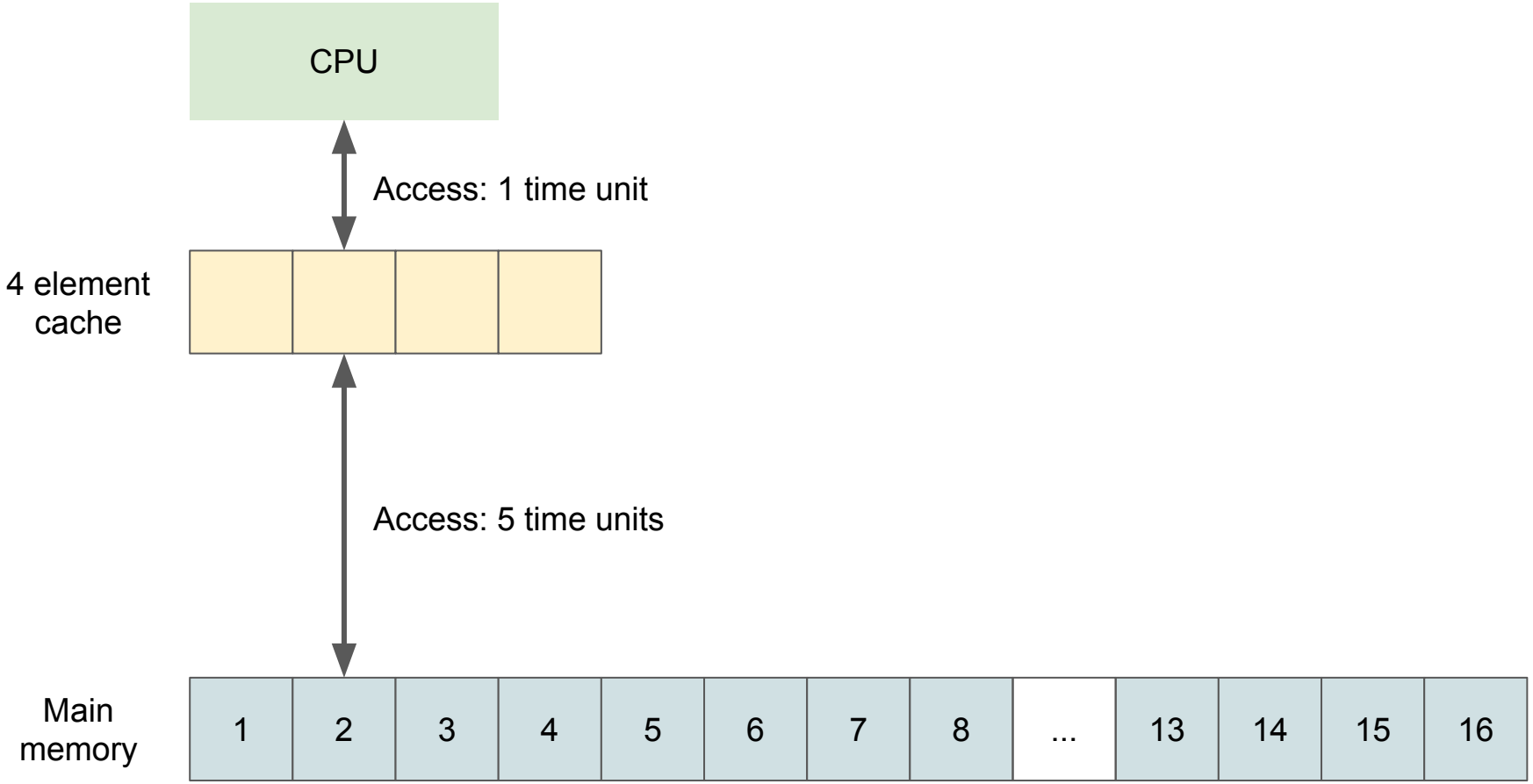
2D Array

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

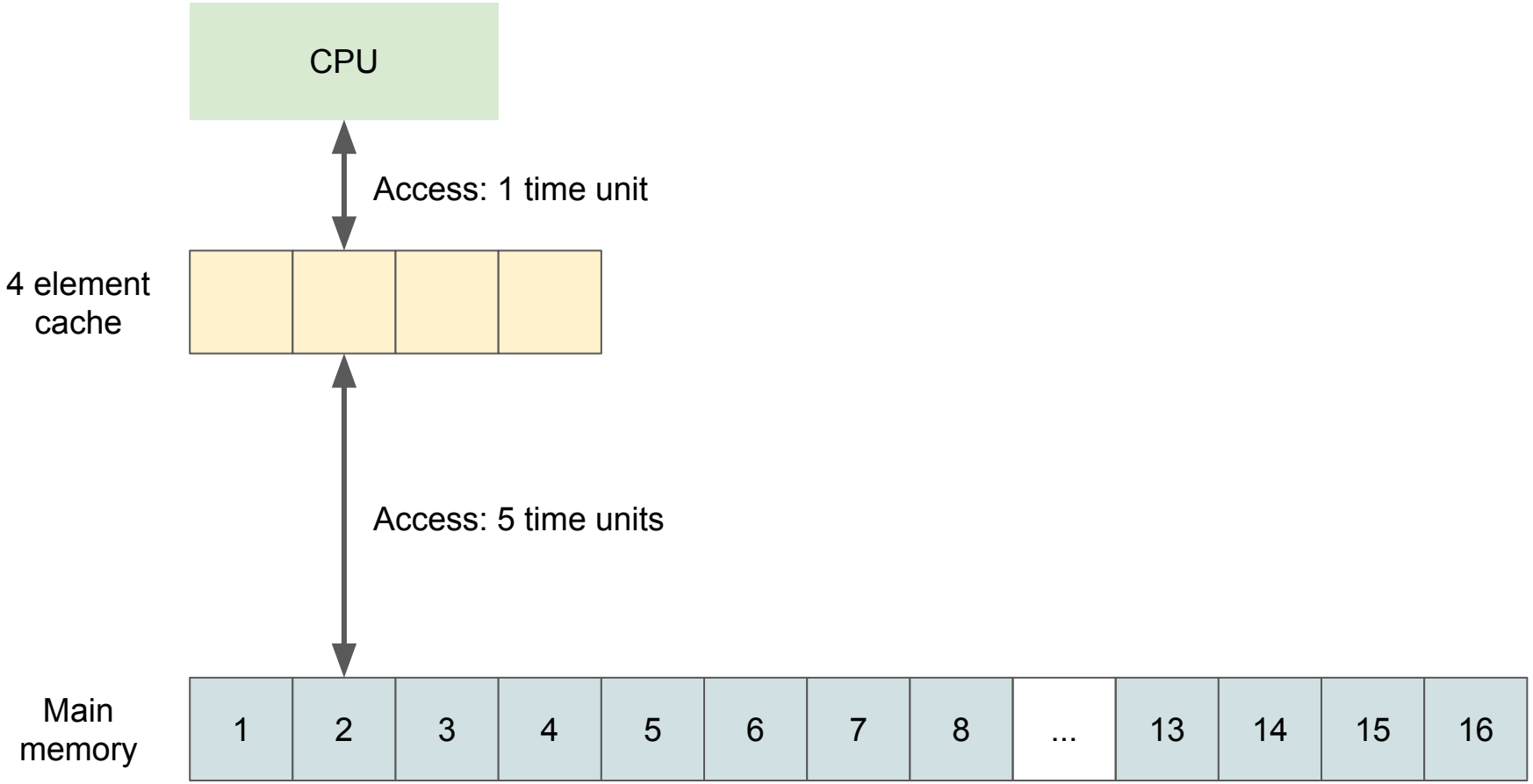
Representation in linear memory (row-major format)

1	2	3	4	5	6	7	8	...	13	14	15	16

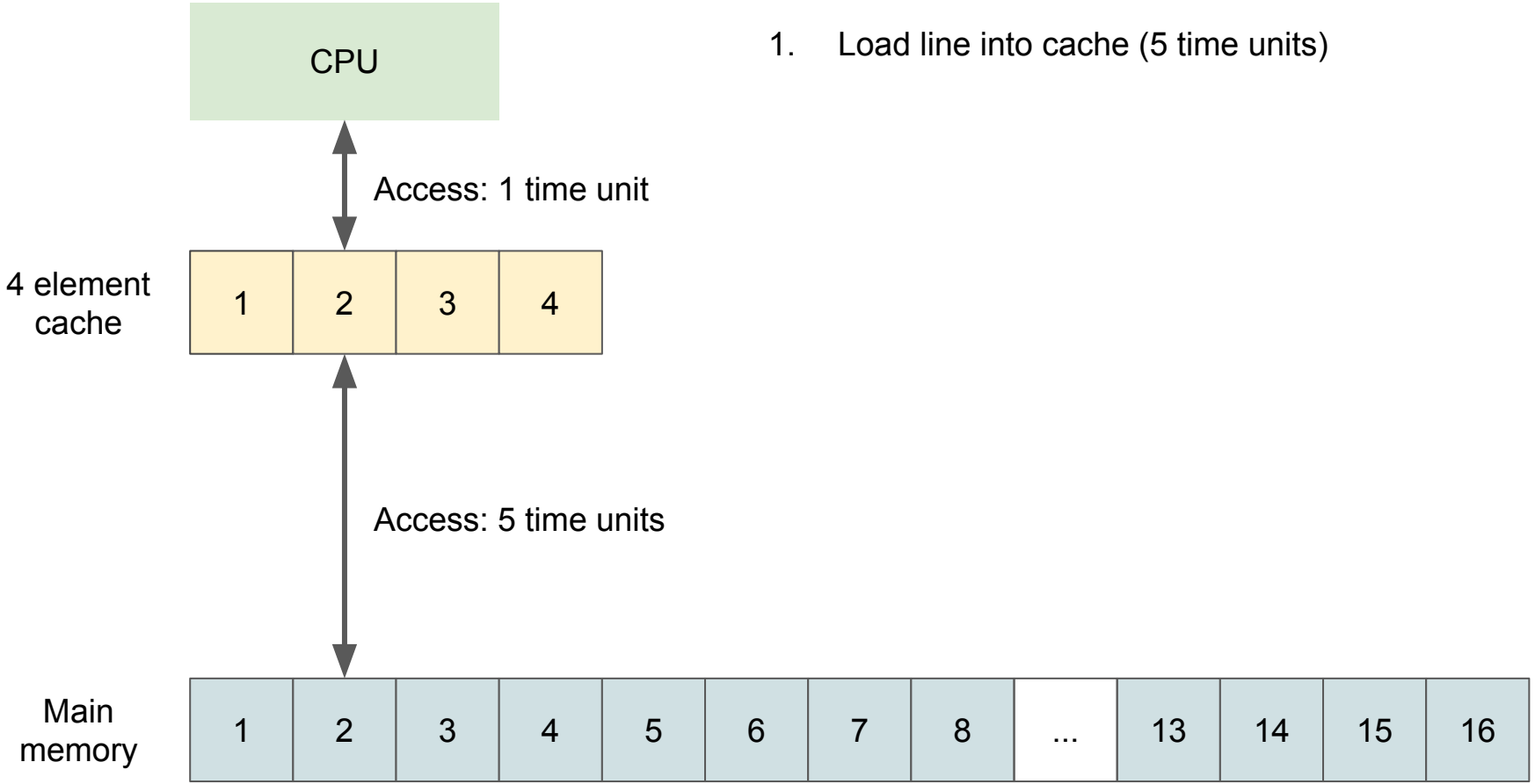
(low address) (high address)

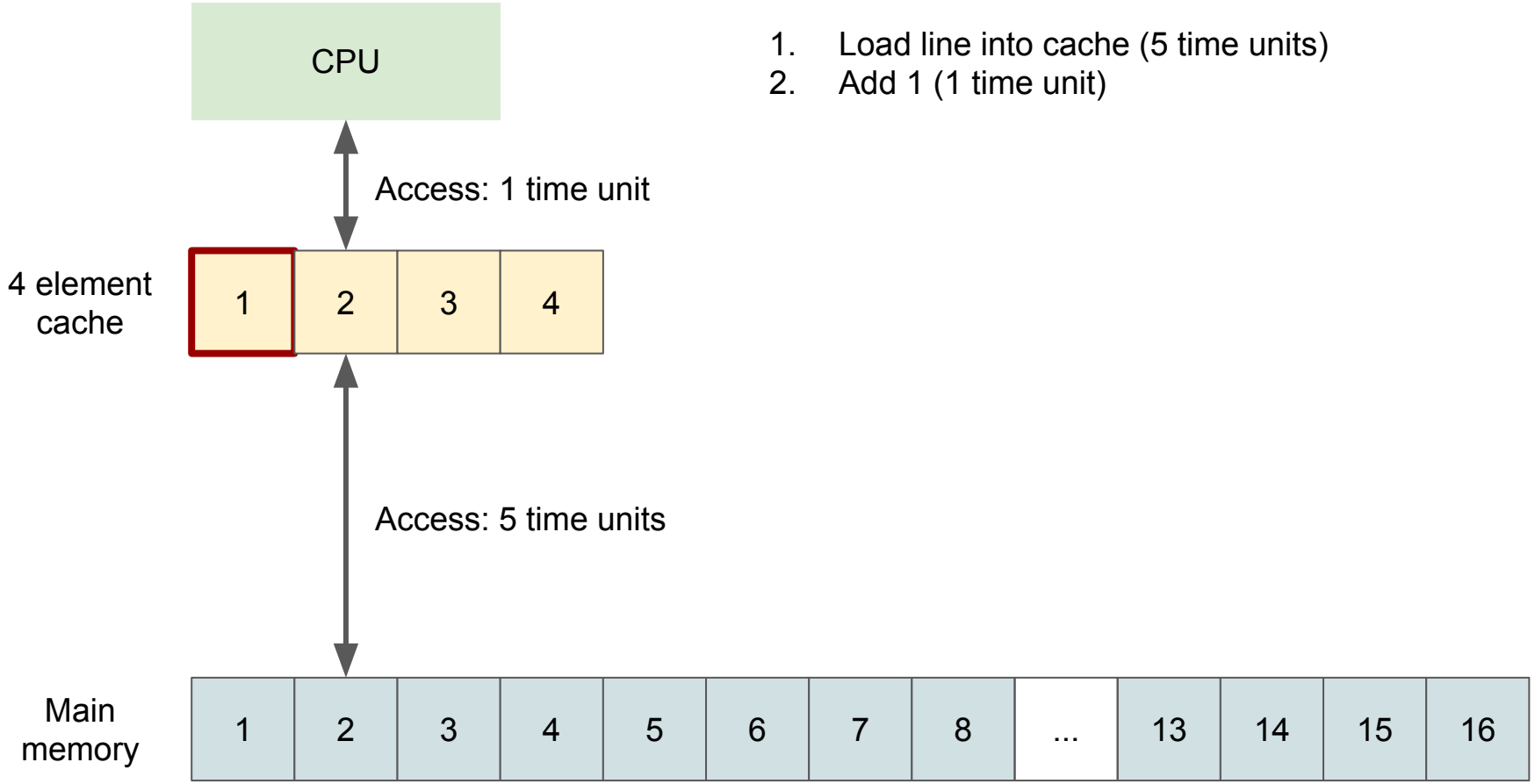

CPU
Access: 1 time unit
4 element cache
Access: 5 time units
Main memory
1
2
3
4
5
6
7
8
...
13
14
15
16

Sum over row

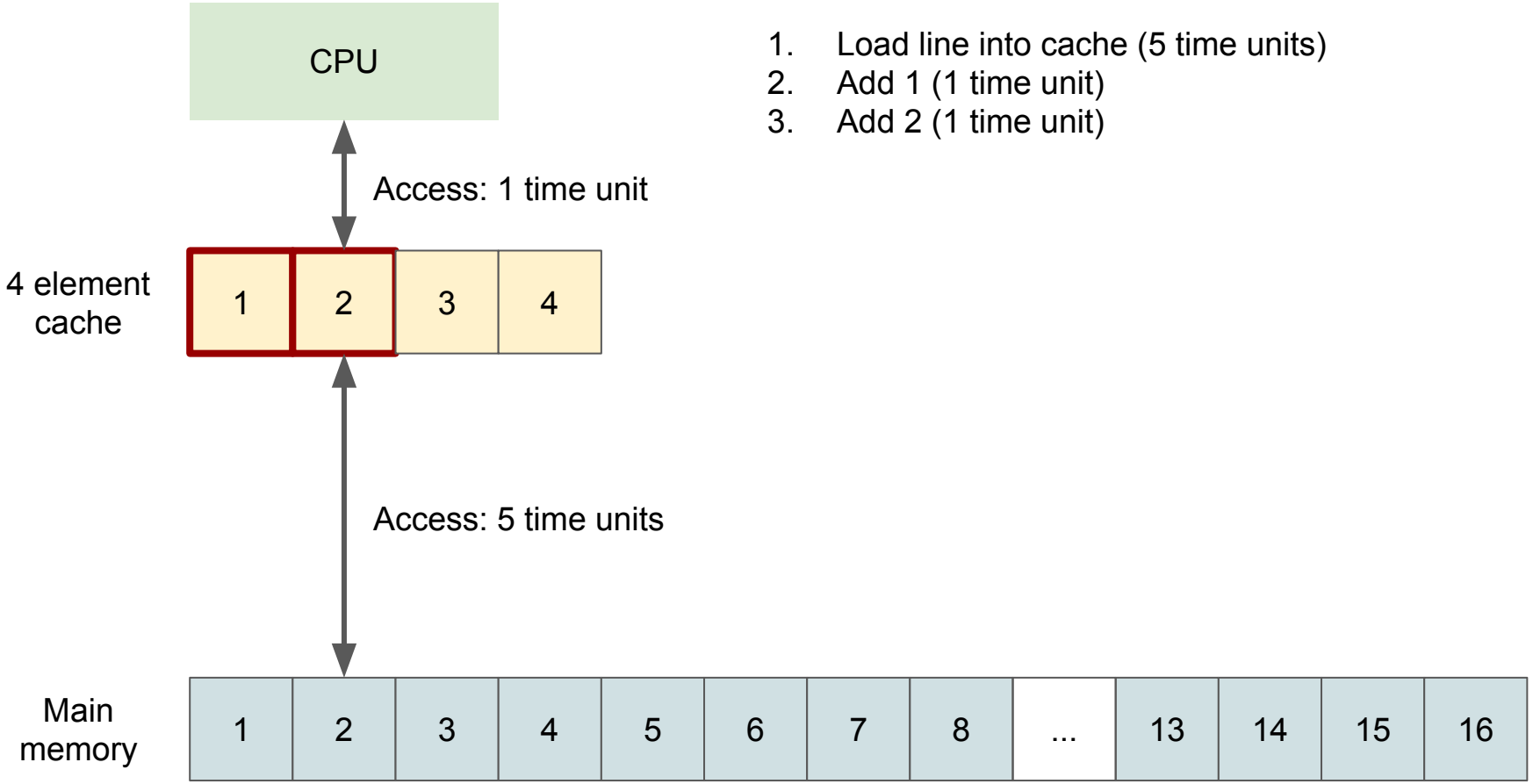

CPU
Access: 1 time unit
4 element cache
Access: 5 time units
Main memory
1
2
3
4
5
6
7
8
...
13
14
15
16

1. Load line into cache (5 time units)

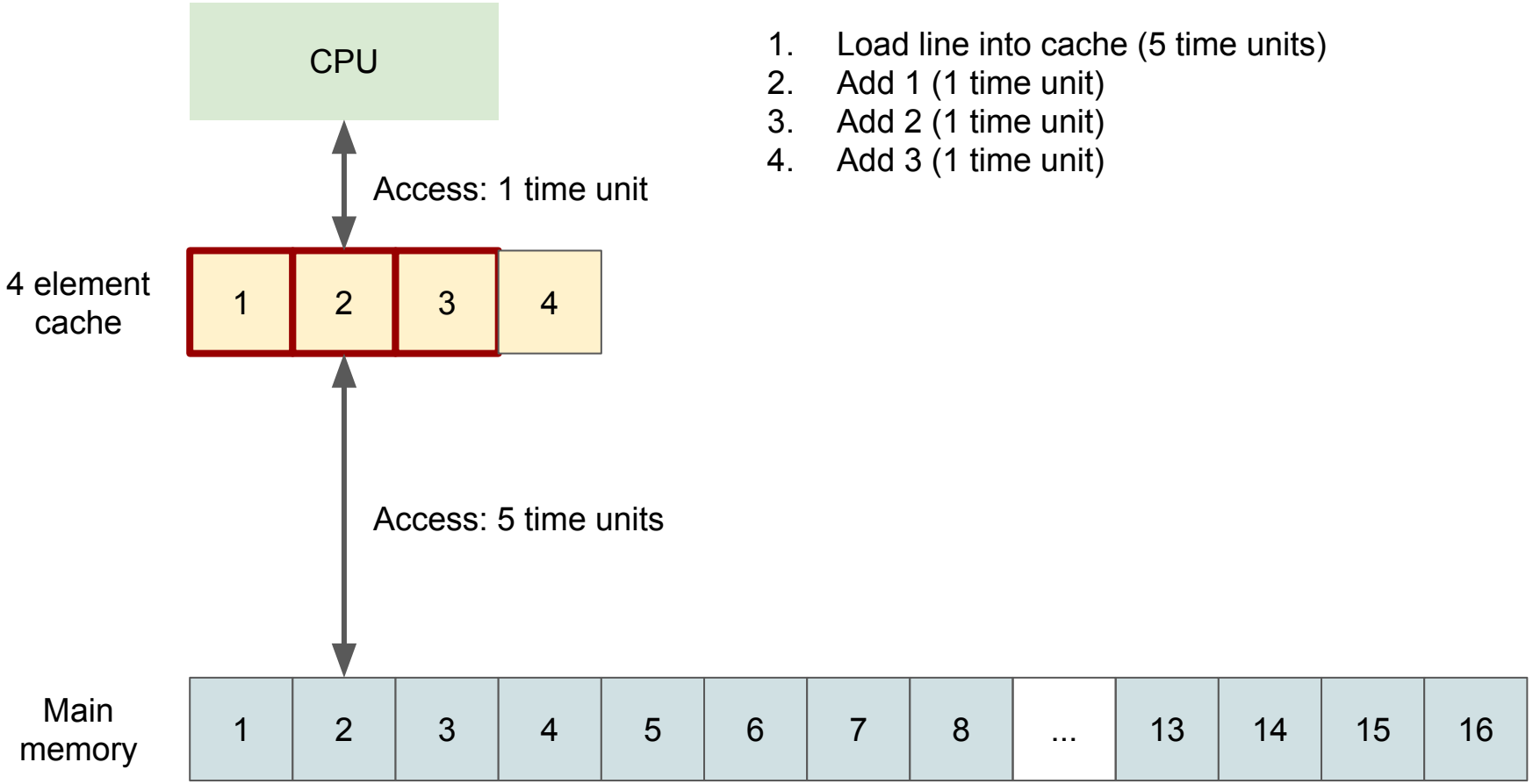




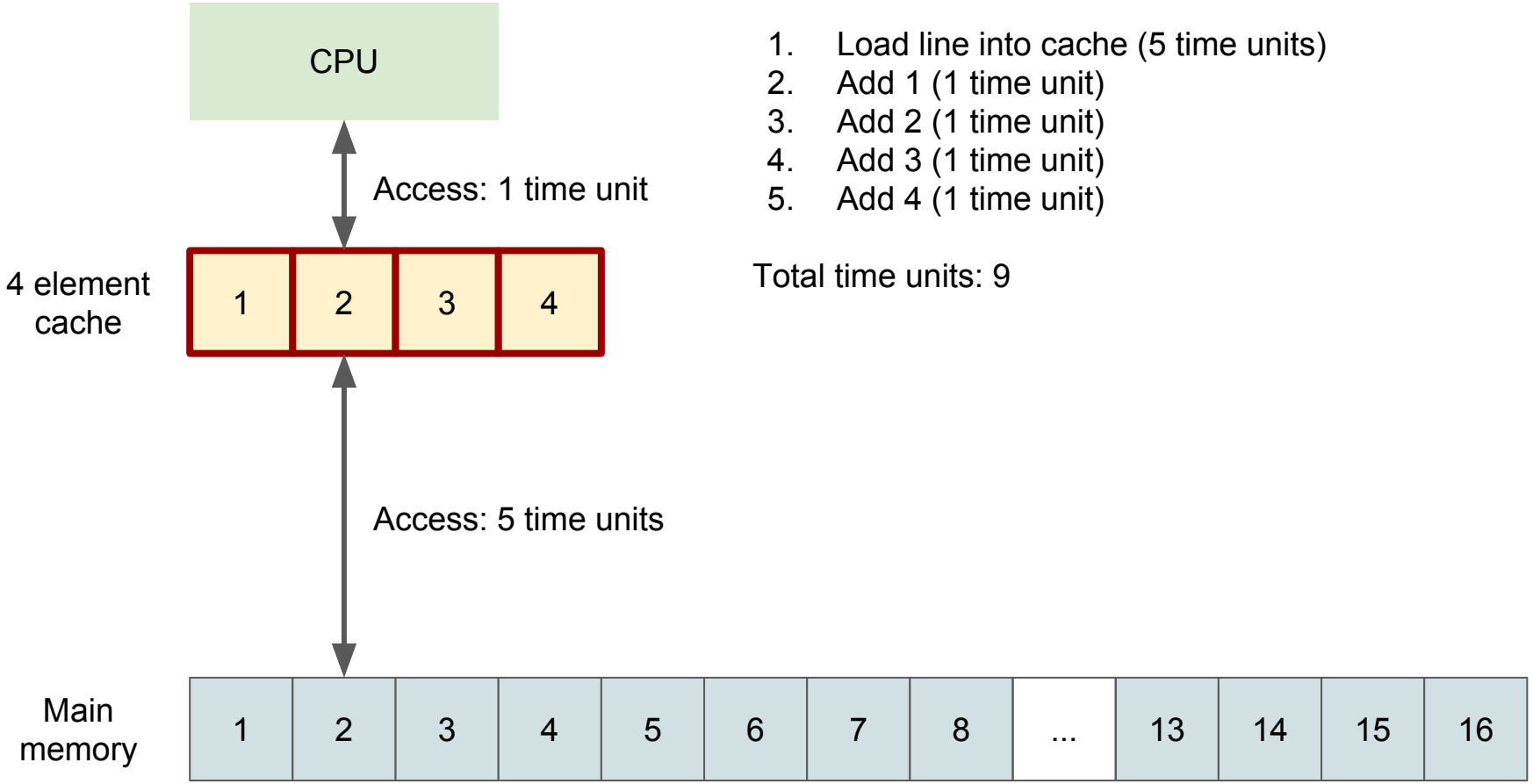
1. Load line into cache (5 time units)
2. Add 1 (1 time unit)



1. Load line into cache (5 time units)
2. Add 1 (1 time unit)
3. Add 2 (1 time unit)



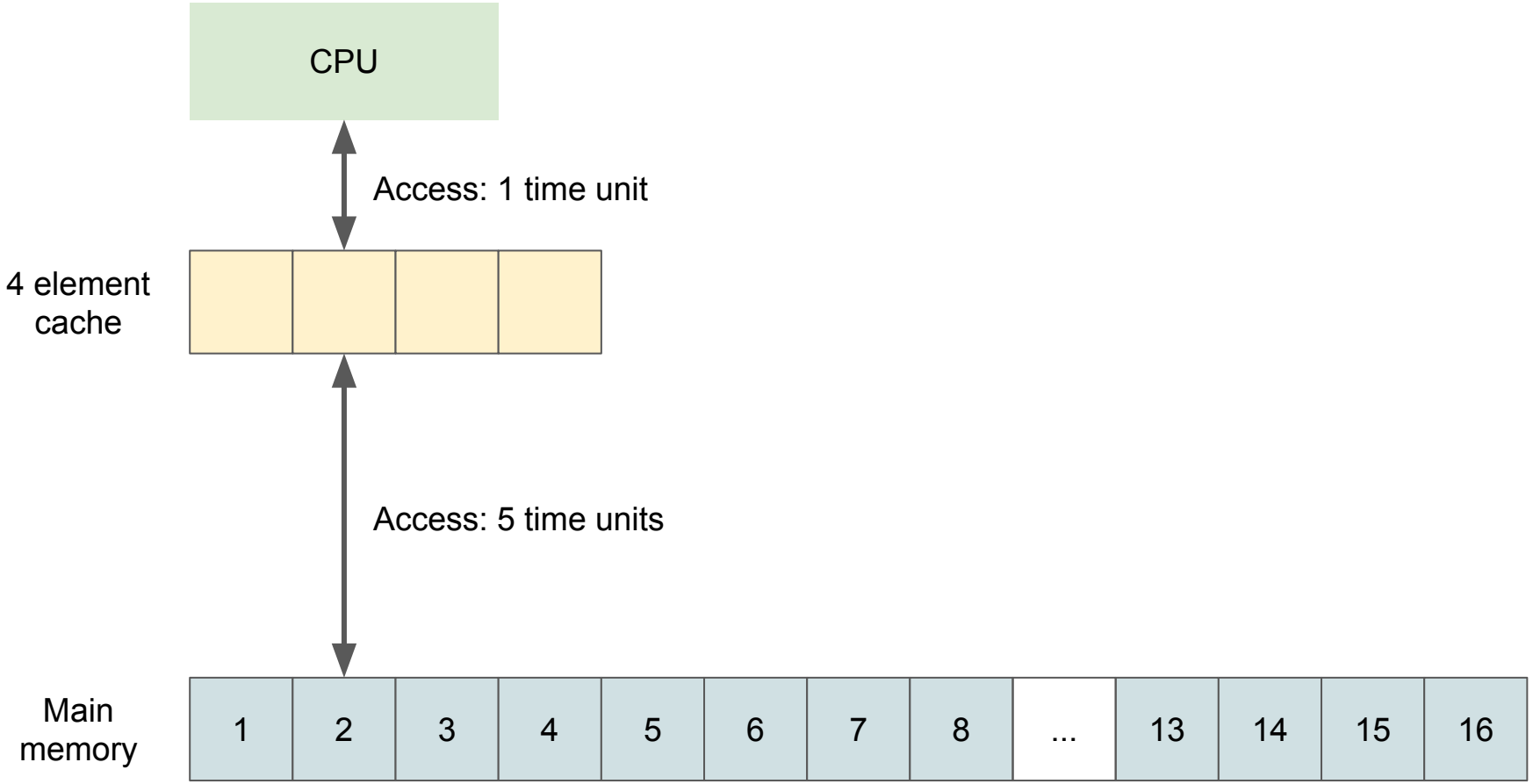
1. Load line into cache (5 time units)
2. Add 1 (1 time unit)
3. Add 2 (1 time unit)
4. Add 3 (1 time unit)



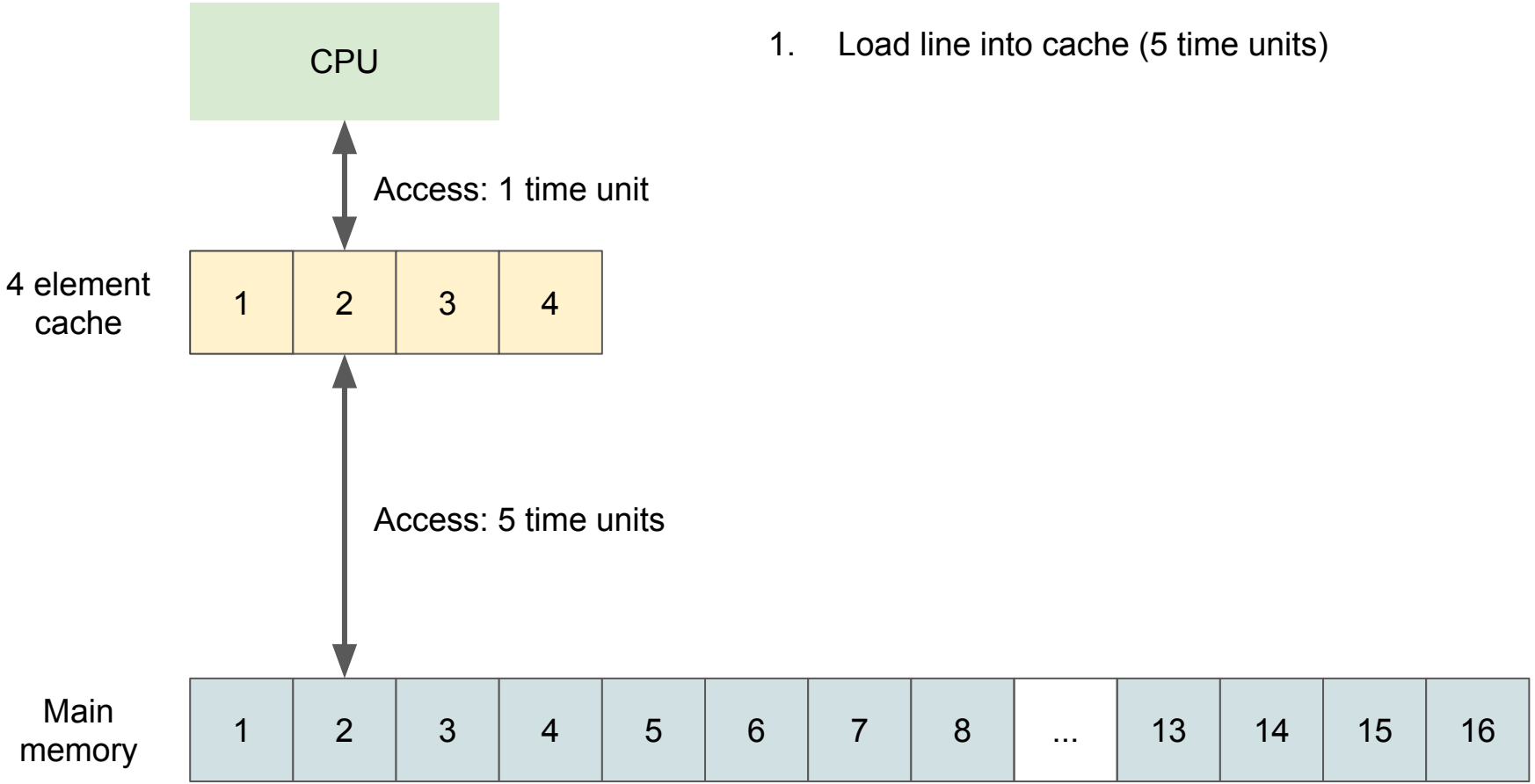
1. Load line into cache (5 time units)
2. Add 1 (1 time unit)
3. Add 2 (1 time unit)
4. Add 3 (1 time unit)
5. Add 4 (1 time unit)

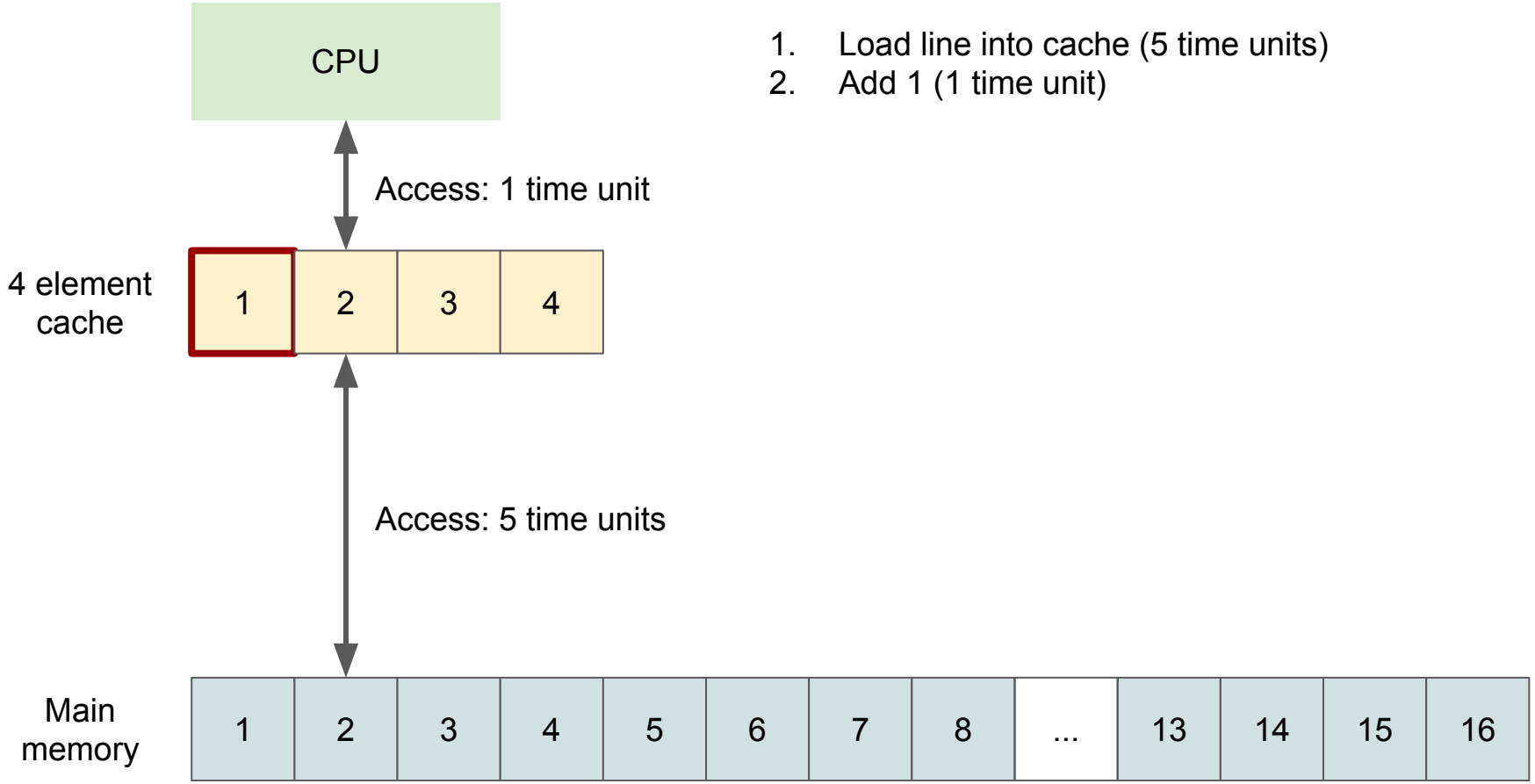
Total time units: 9

Sum over column

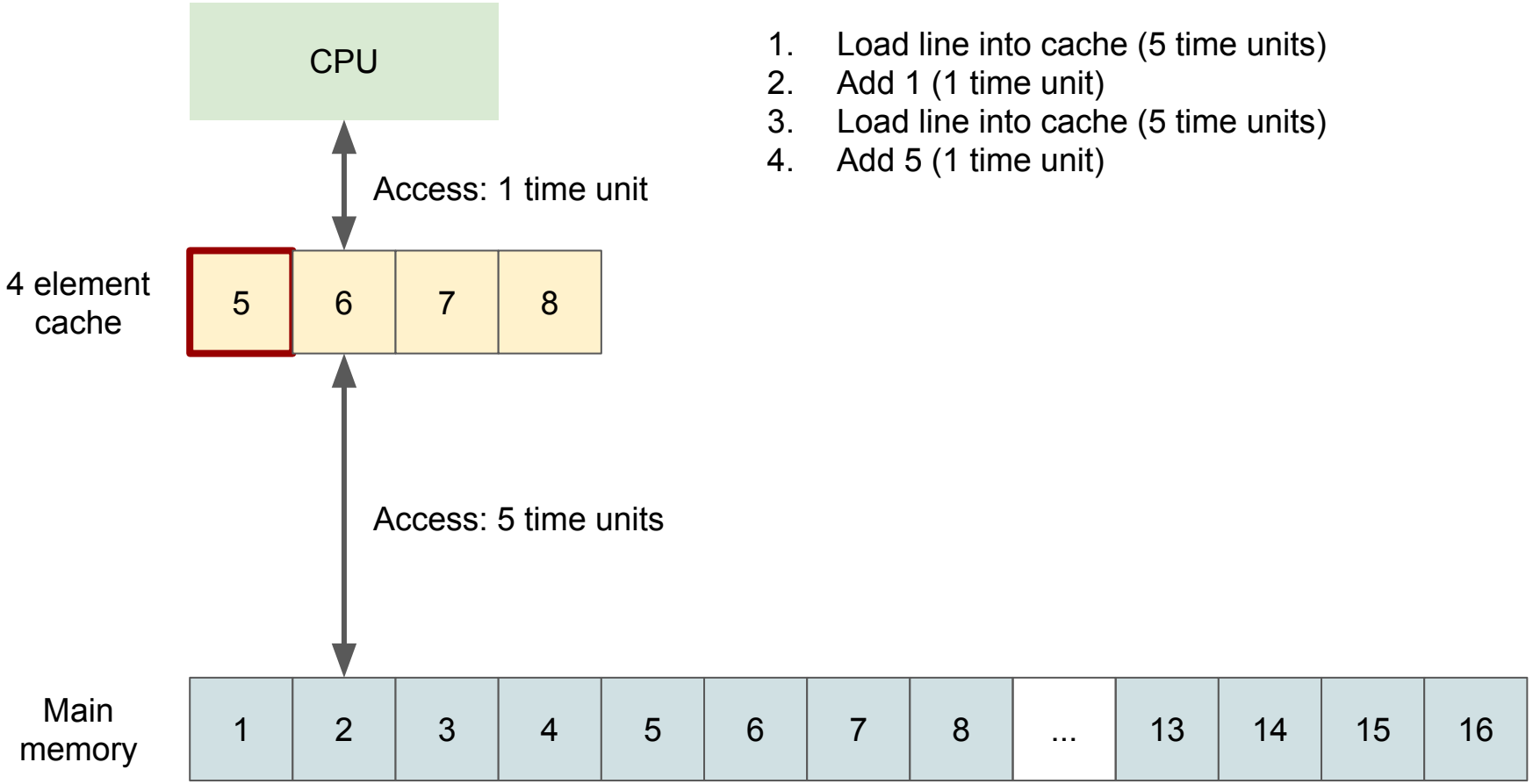

CPU
Access: 1 time unit
4 element cache
Access: 5 time units
Main memory
1
2
3
4
5
6
7
8
...
13
14
15
16

1. Load line into cache (5 time units)

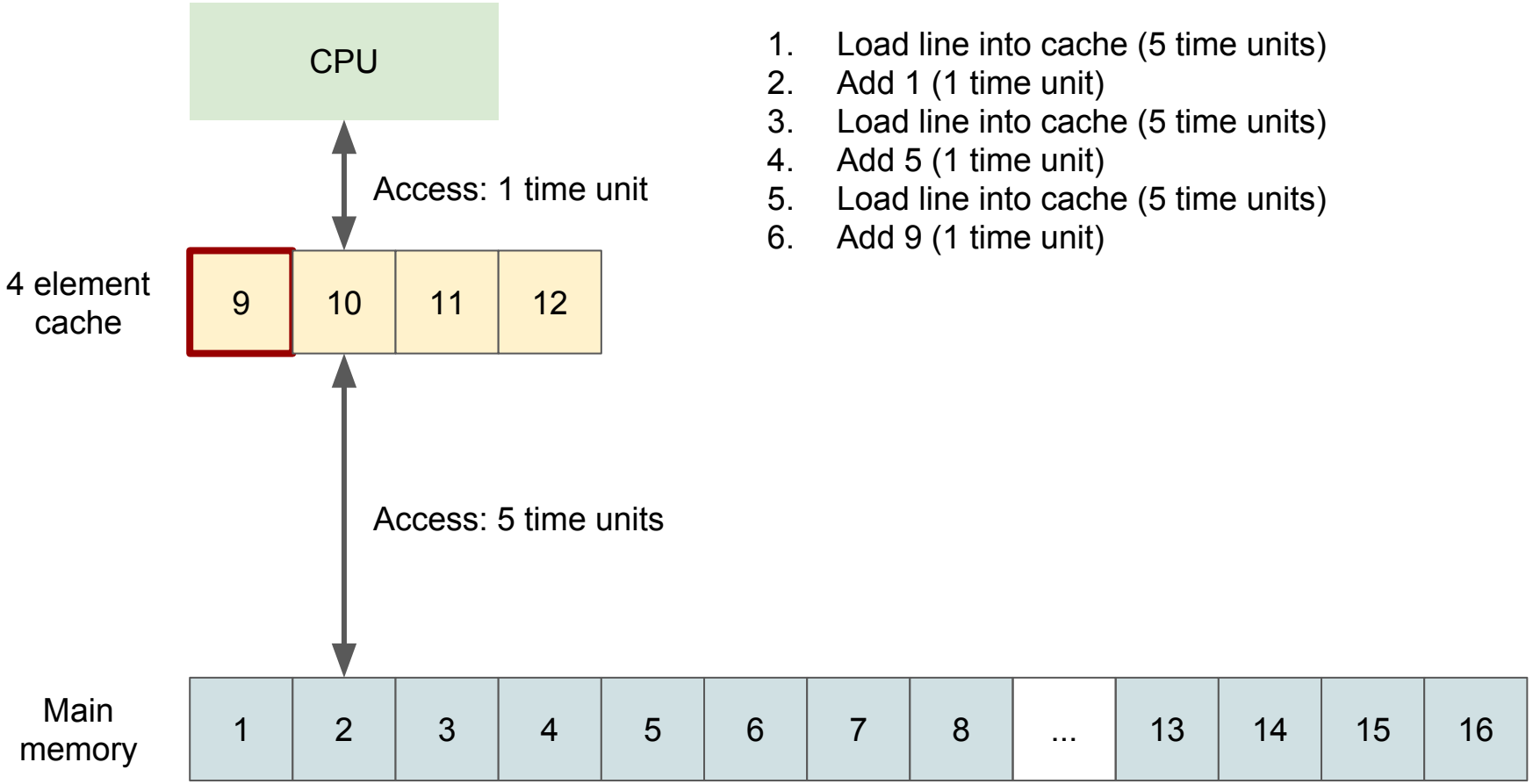




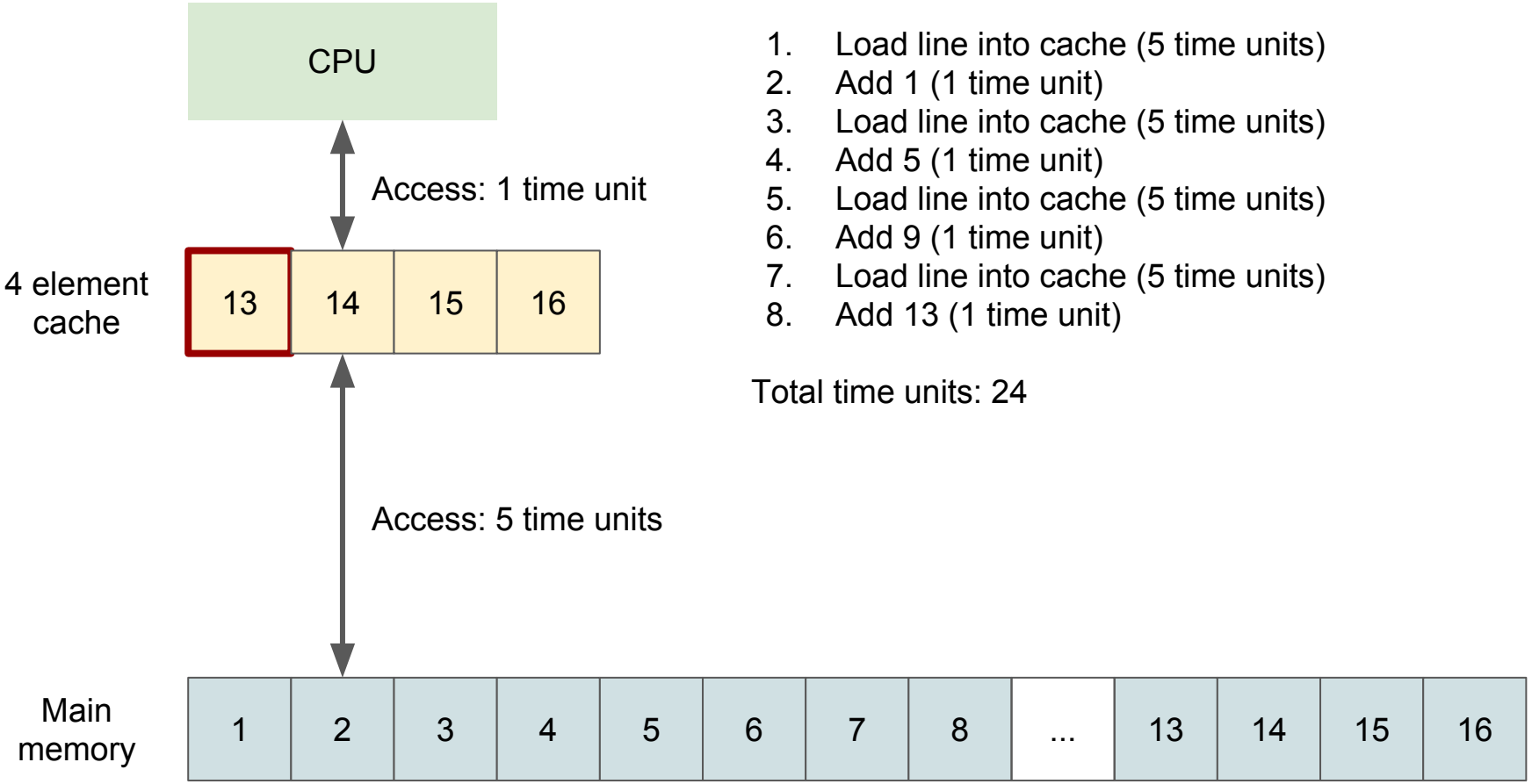
1. Load line into cache (5 time units)
2. Add 1 (1 time unit)



1. Load line into cache (5 time units)
2. Add 1 (1 time unit)
3. Load line into cache (5 time units)
4. Add 5 (1 time unit)



1. Load line into cache (5 time units)
2. Add 1 (1 time unit)
3. Load line into cache (5 time units)
4. Add 5 (1 time unit)
5. Load line into cache (5 time units)
6. Add 9 (1 time unit)



1. Load line into cache (5 time units)
2. Add 1 (1 time unit)
3. Load line into cache (5 time units)
4. Add 5 (1 time unit)
5. Load line into cache (5 time units)
6. Add 9 (1 time unit)
7. Load line into cache (5 time units)
8. Add 13 (1 time unit)

Total time units: 24