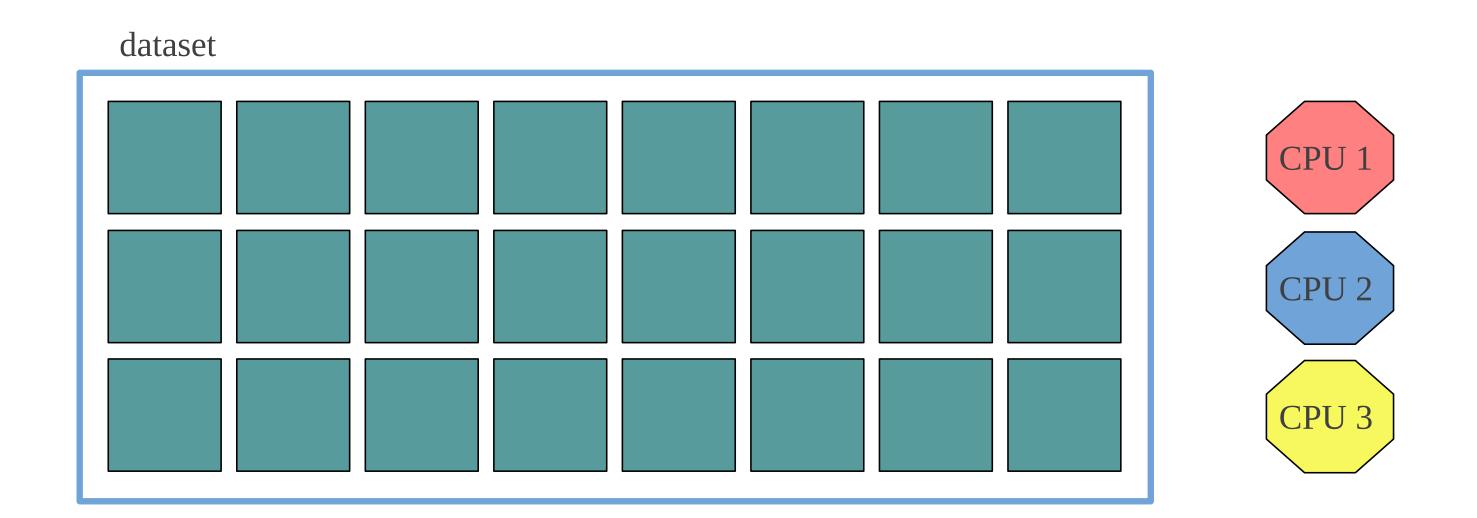
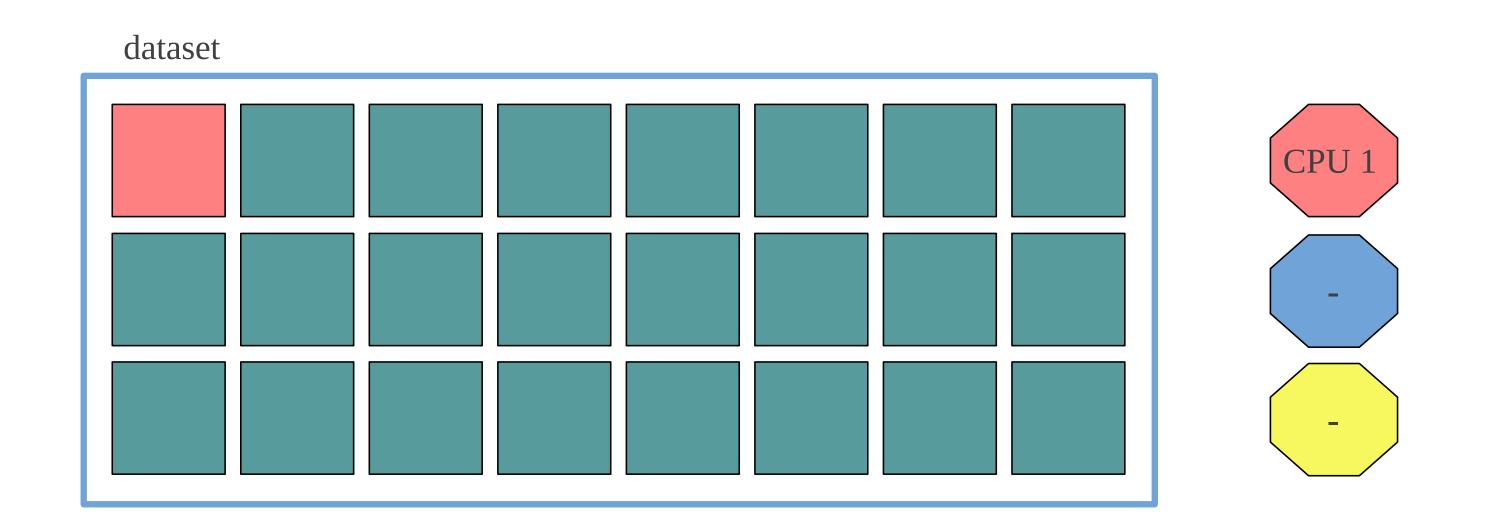


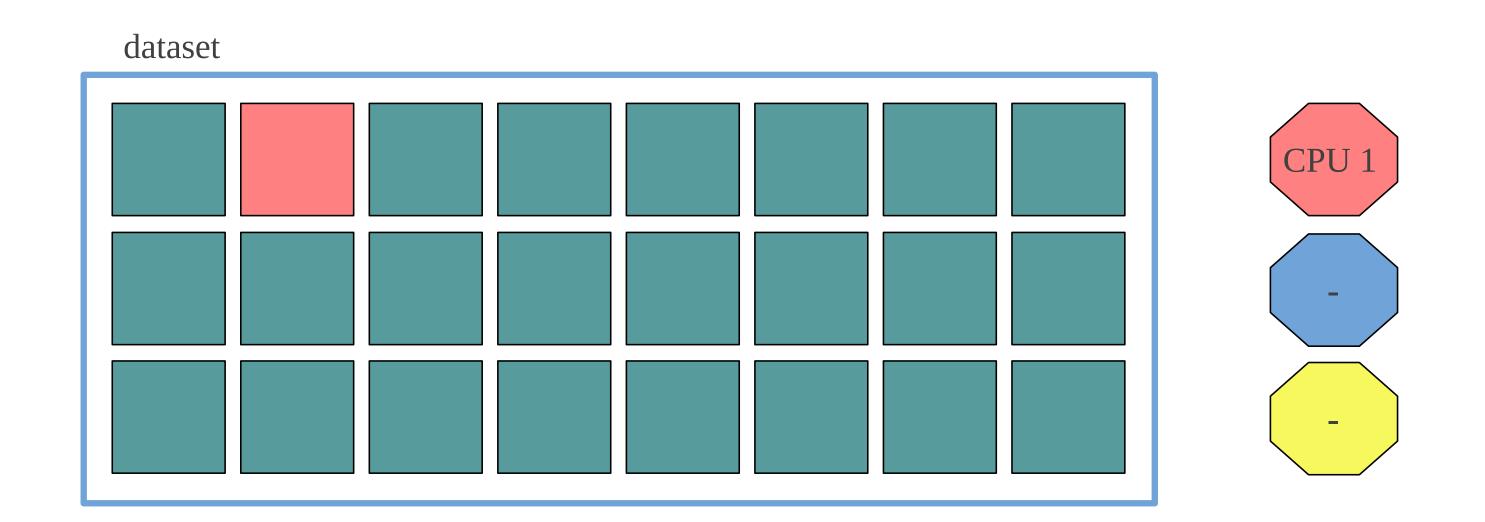
# Data processing



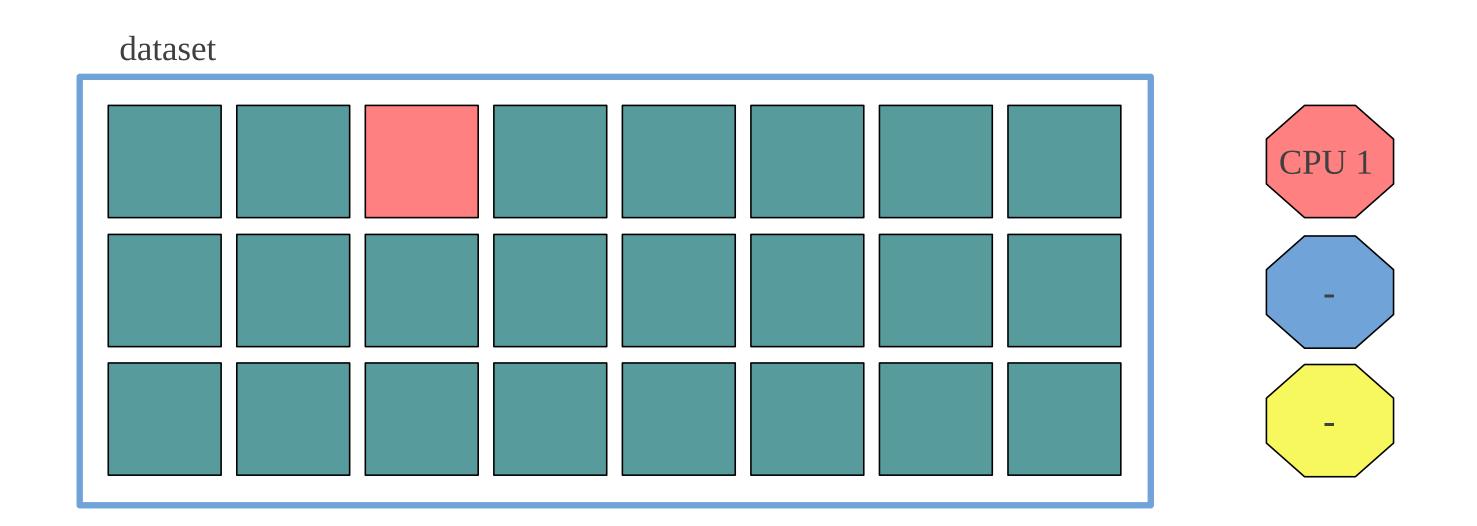
# Sequential iteration



# Sequential iteration



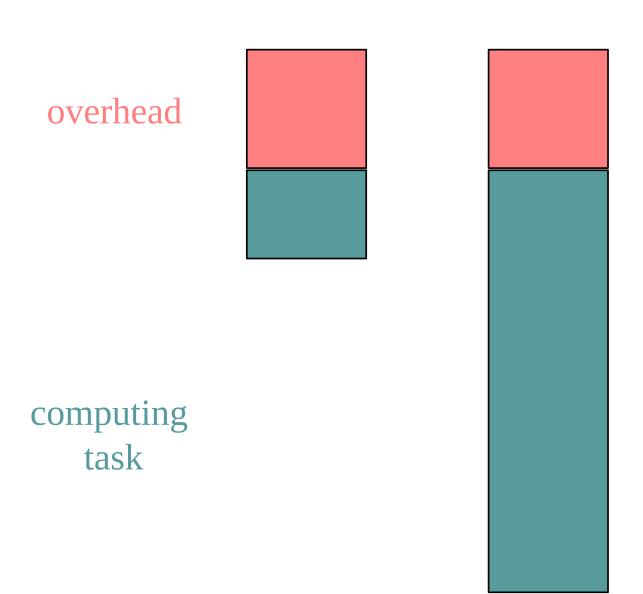
# Sequential iteration



dataset partition 1 datacenter 1 Node partition 2 Node partition 3 datacenter 2 Node

#### Important points

1. Size the partitions appropriately



#### Important points

- 1. Size the partitions appropriately
- 2. Benchmark and configure

#### Important points

- 1. Size the partitions appropriately
- 2. Benchmark and configure
- 3. Parallel process leads to the same result as the sequential one



### foldLeft

### All partitions folded

### foldLeft



#### monoFoldLeft use n + 1 default value

FoldLeft per partition

### monoFoldLeft requires combine to be associative

```
(a1 combine a2 combine a3)
(b1 combine b2 combine b3)
```

#### monoFoldLeft requires combine to be associative

```
(a1 combine a2 combine a3) combine (b1 combine b2 combine b3)
```

#### monoFoldLeft requires combine to be associative

foldLeft per partition

```
(a combine b combine c) combine
(d combine e combine f)
```

==

```
(a combine b combine c combine d) combine
(e combine f)
```