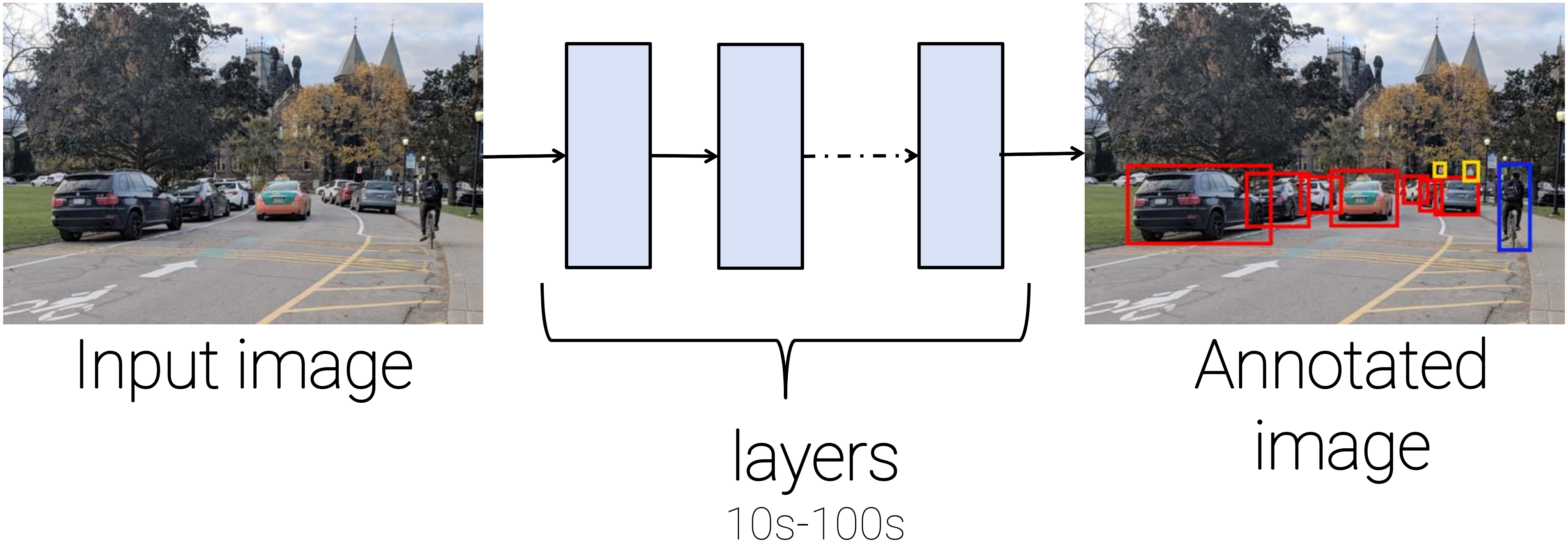
UNIVERSITY OF TORONTO

Bit-Tactical: Sparse/Dense Deep Learning Inference Accelerator(*)

Alberto Delmas, Patrick Judd, Mostafa Mahmoud,
Milos Nikolic, Zissis Poulos, Sayeh Sharify,
Kevin Siu, Dylan Stuart
Andreas Moshovos

(*) For licensing please contact the University of Toronto Innovation Office

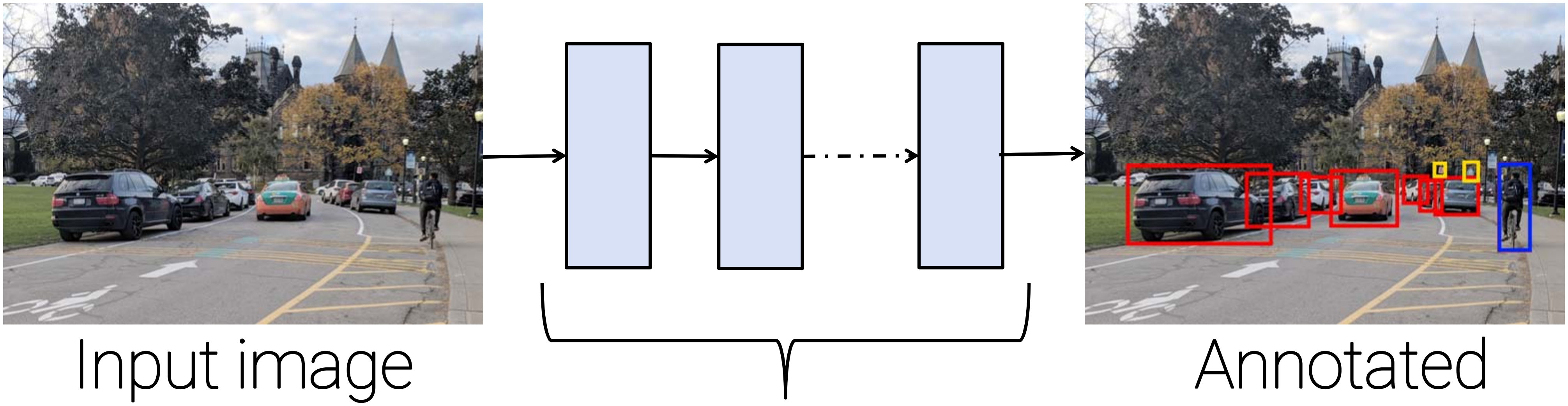


Input image

layers

10s-100s

Annotated image



Input image

layers

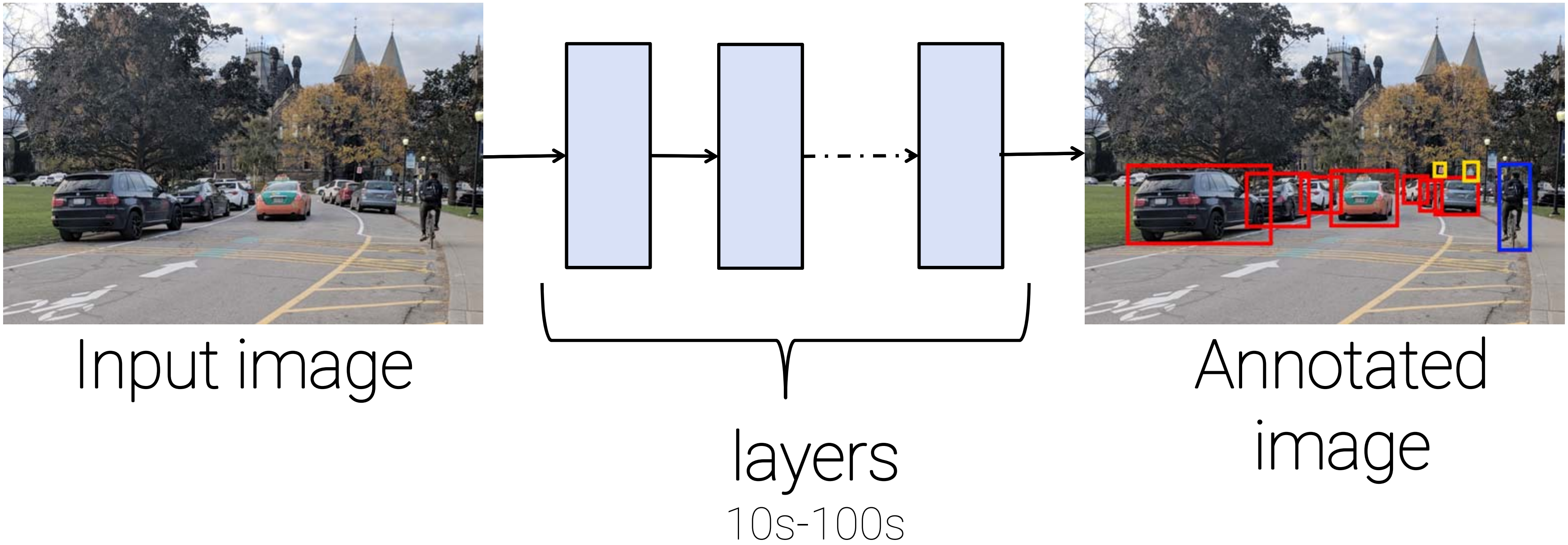
10s-100s

Annotated image

Faster

+

More Energy Efficient



Input image

layers

10s-100s

Annotated image

Fewer Operations

Fewer Data Transfers

Fewer Operations

Fewer Data Transfers

Fewer Operations
Fewer Data Transfers

$$\text{Out} \mathrel{+}= A_0 \times W_0$$

$$\text{Out} \mathrel{+}= A_1 \times W_1$$

$$\text{Out} \mathrel{+}= A_2 \times W_2$$

$$\text{Out} \mathrel{+}= A_3 \times W_3$$

$$\text{Out} \mathrel{+}= A_4 \times W_4$$

$$\vdots$$

Eliminate Ineffectual Operations

Out += A_0 x W_0

Out += A_1 x W_1

Out += A_2 x W_2

Out += A_3 x W_3

Out += A_4 x W_4

$\vdots$

Eliminate Ineffectual Operations

$\text{Out} \mathrel{+}= A_0 \times W_0$

$\text{Out} \mathrel{+}= 0 \times W_1$

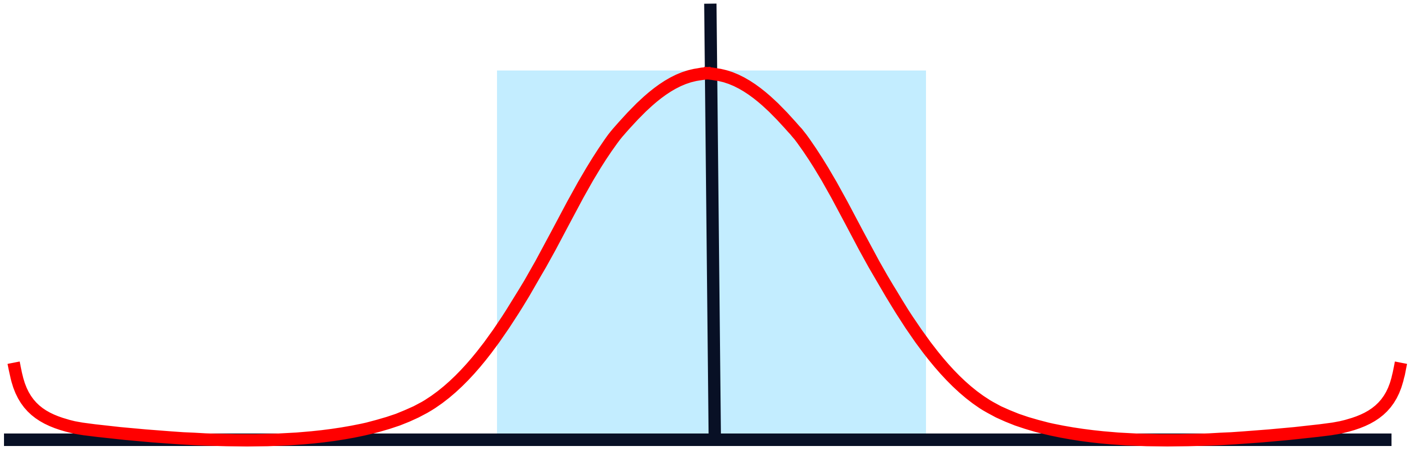
$\text{Out} \mathrel{+}= A_2 \times W_2$

$\text{Out} \mathrel{+}= A_3 \times 0$

$\text{Out} \mathrel{+}= A_4 \times W_4$

⋮

Takeaway #1

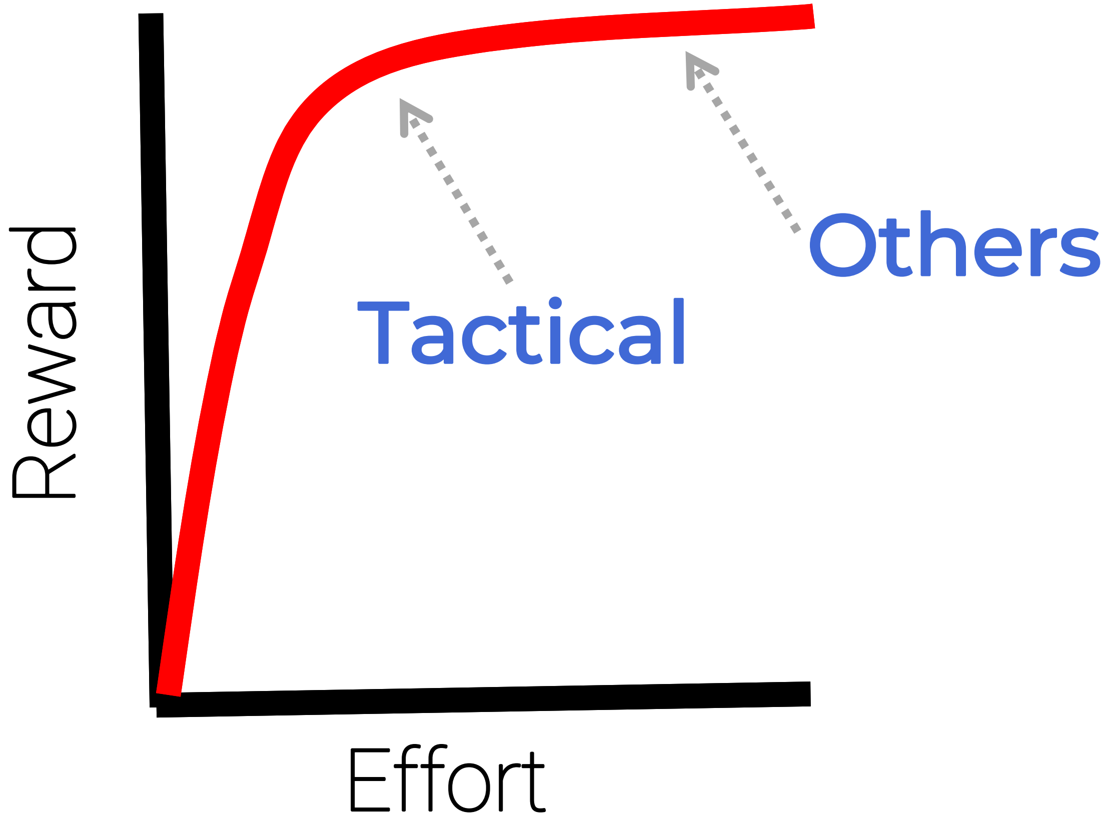


0.5x – 0.2x Off-Chip Transfers On-Chip Storage

DPRed: Making Typical Activation and Weight Values Matter In Deep Learning Computing, Delmas et al., **https://arxiv.org/abs/1804.06732**

Takeaway #2

Out += A x 0



Takeaway #2

Out += A x 0

Aim to get **most** not **all**

Simple Design
Software Scheduler

Takeaway #3

Out += 0 x W

Don’t!

Only ~50% A==0

Takeaway #3

Out += 0100001 x W

Go for Bit Sparsity

> 90%

Takeaway #4

vs. SCNN

Speed	Energy
5.3x	0.85x

Do as you are told?

$$\text{Out} \mathrel{+}= A_0 \times W_0$$

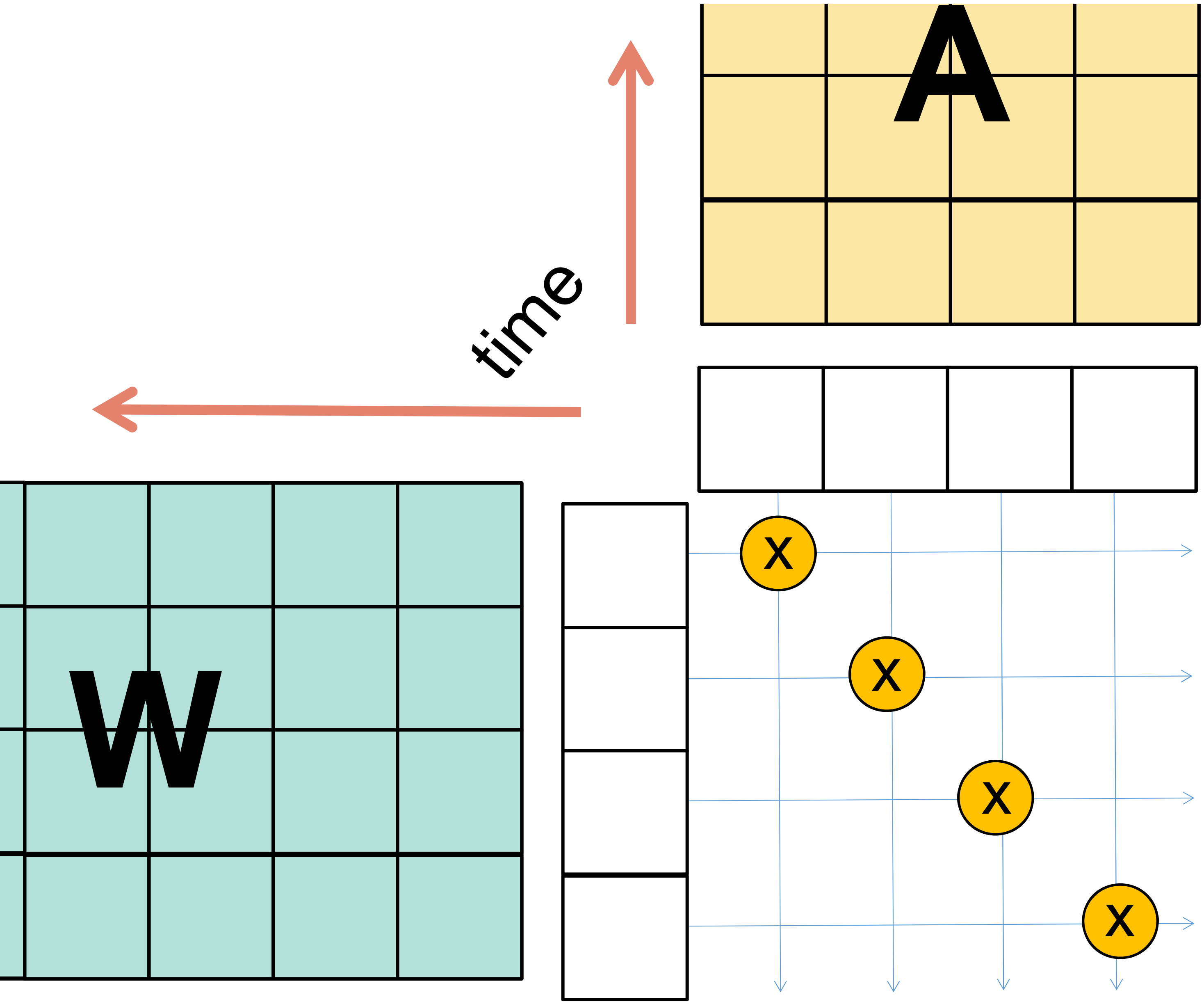
$$\text{Out} \mathrel{+}= A_1 \times W_1$$

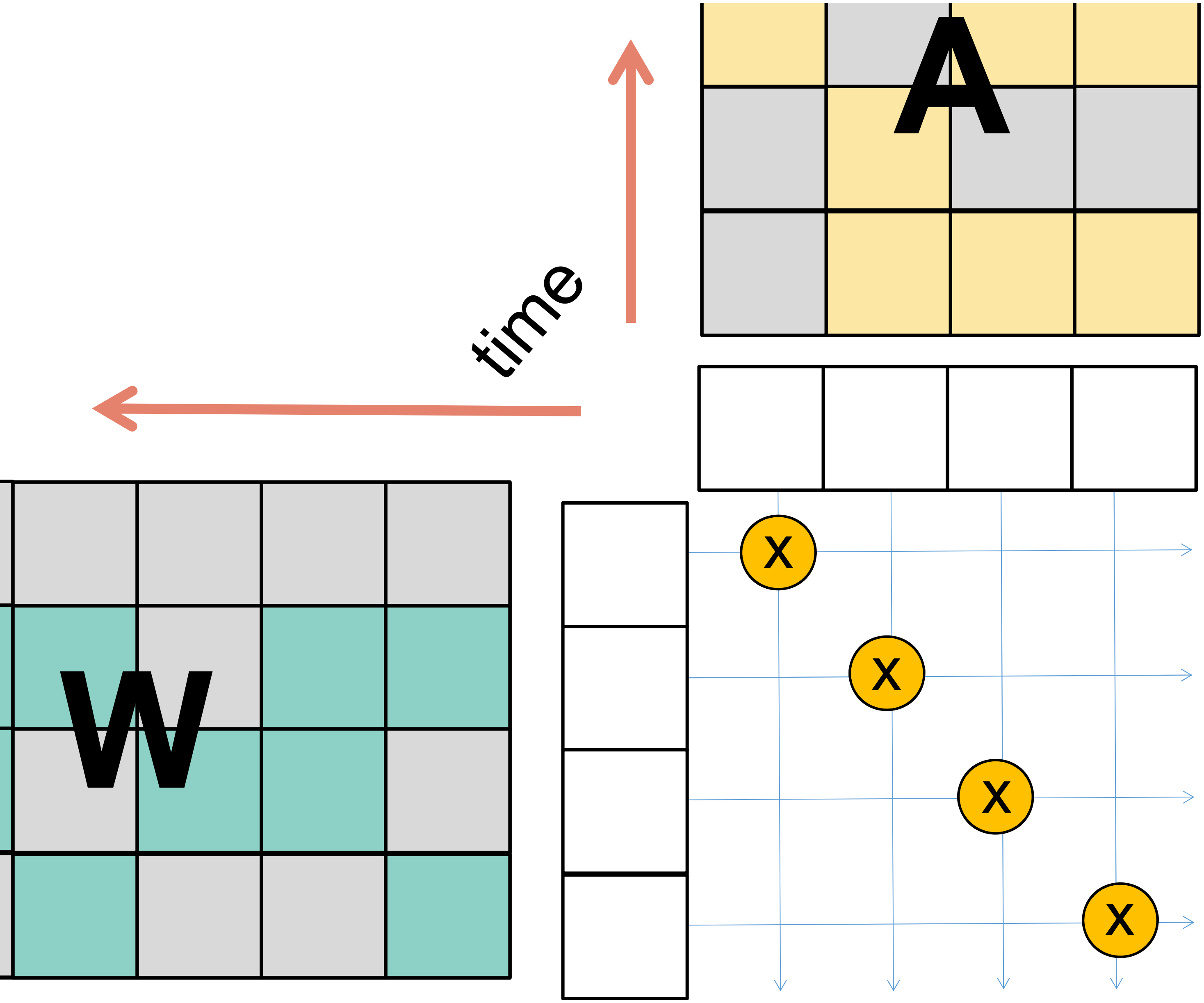
$$\text{Out} \mathrel{+}= A_2 \times W_2$$

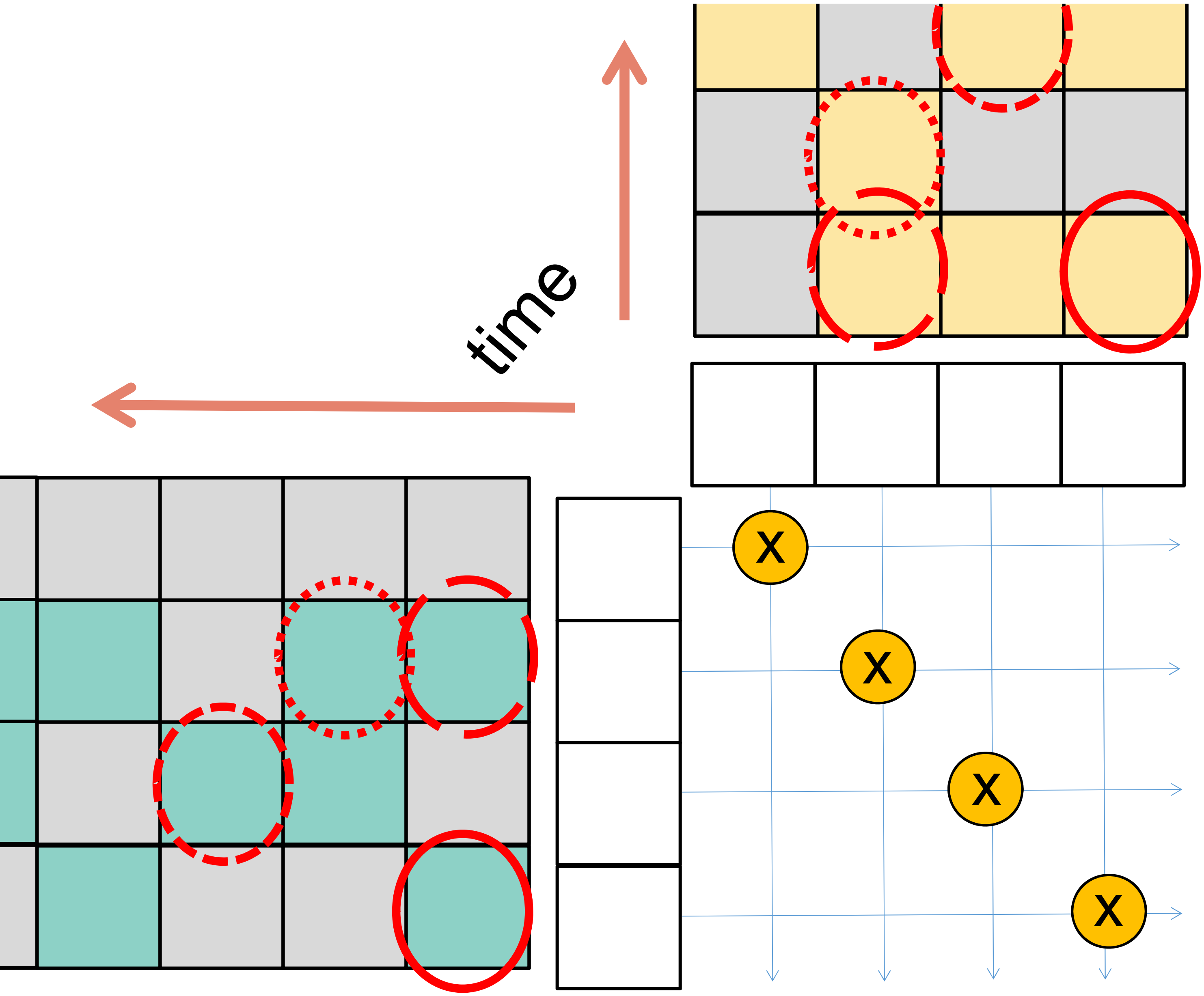
$$\text{Out} \mathrel{+}= A_3 \times W_3$$

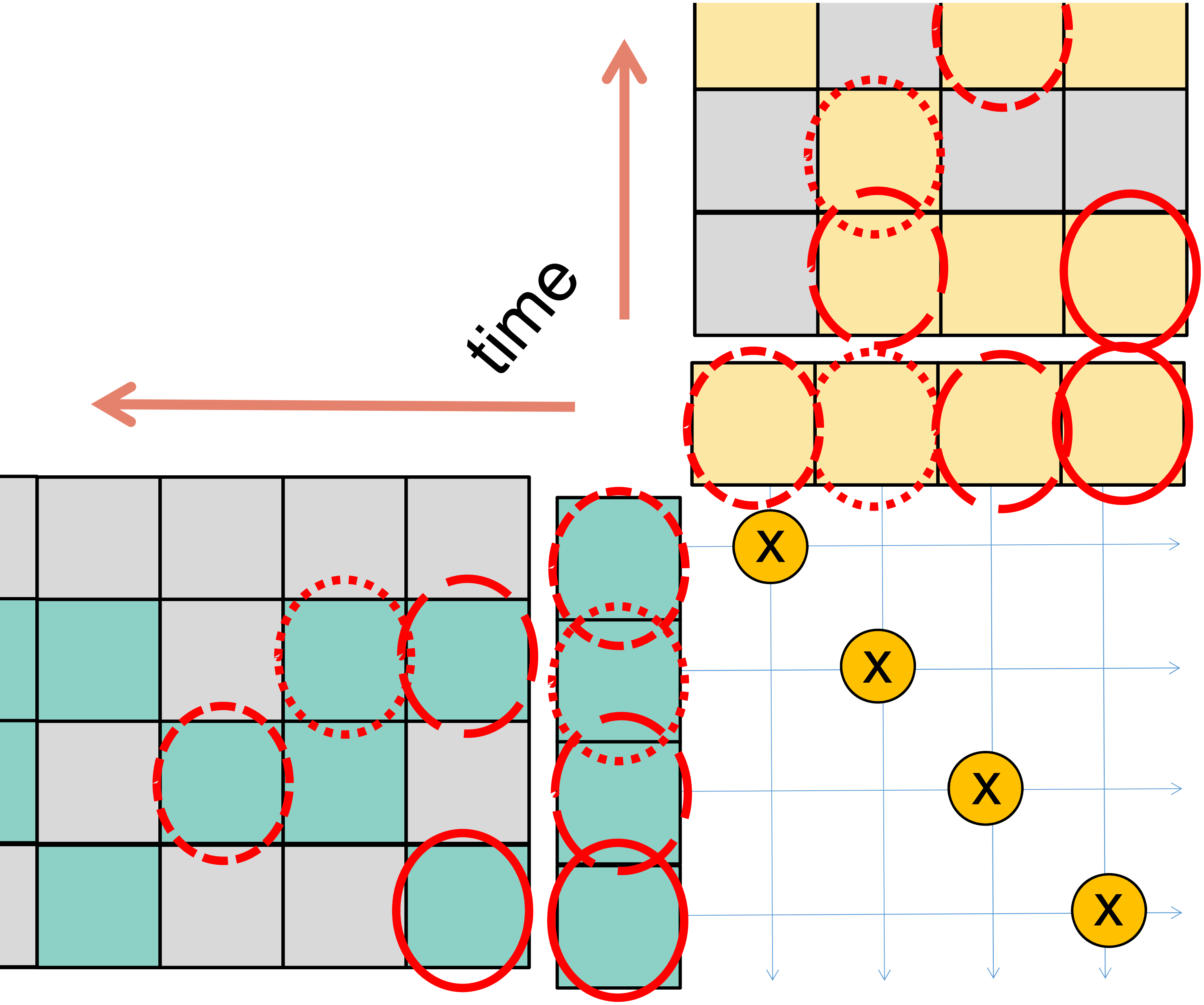
$$\text{Out} \mathrel{+}= A_4 \times W_4$$

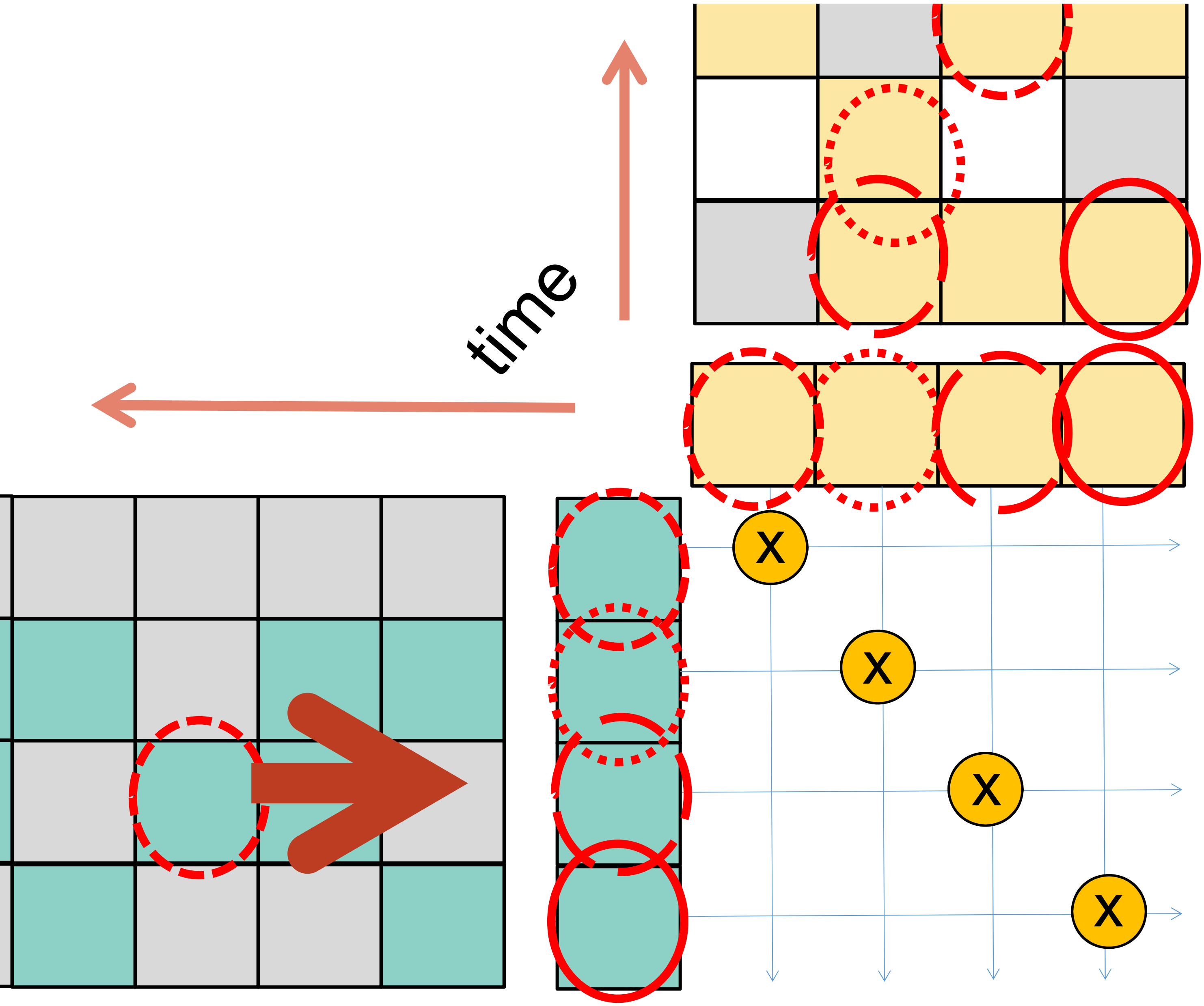
⋮


A
time
W
X
X
X
X

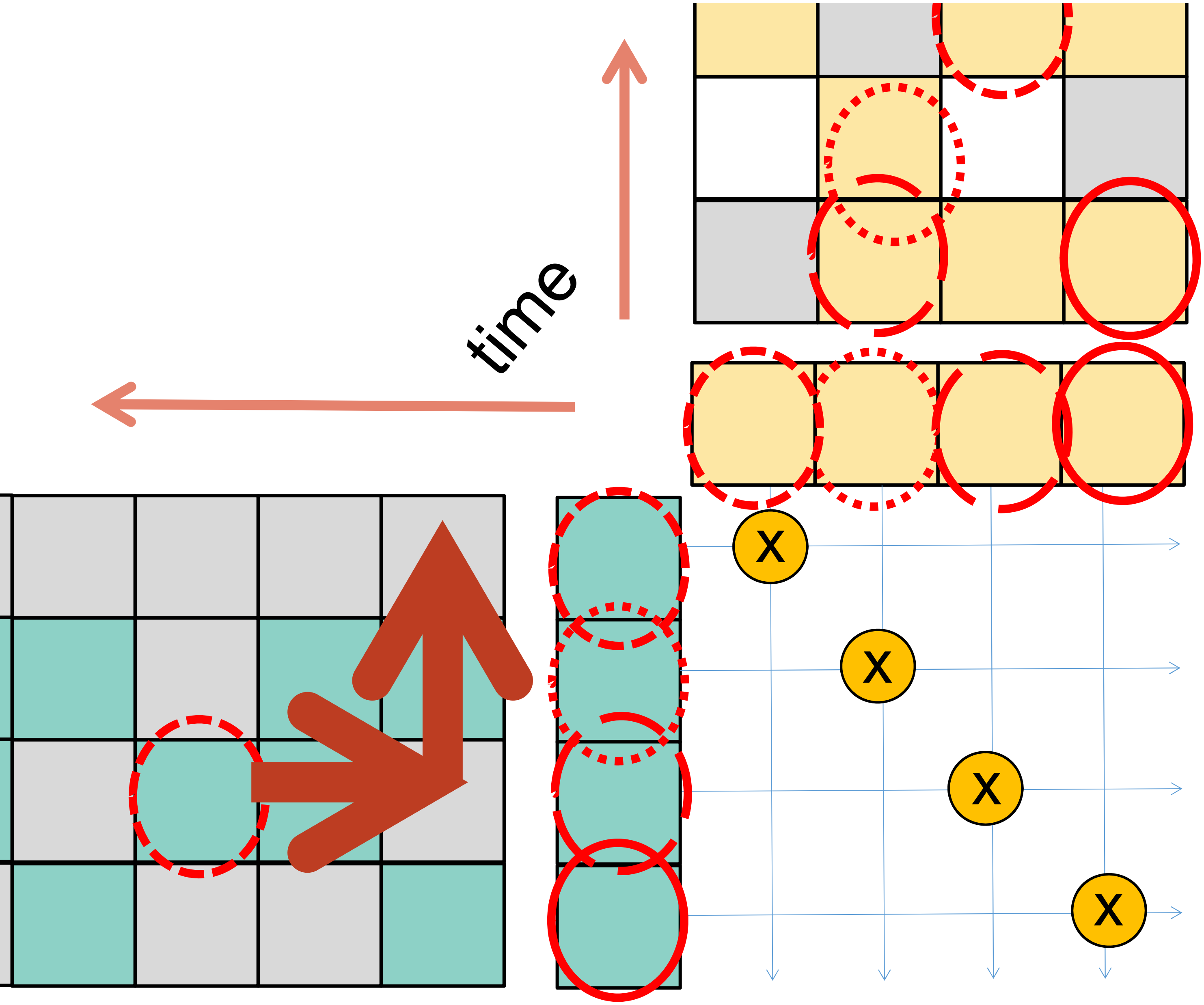

A
time
W
x
x
x
x


time
X
X
X
X

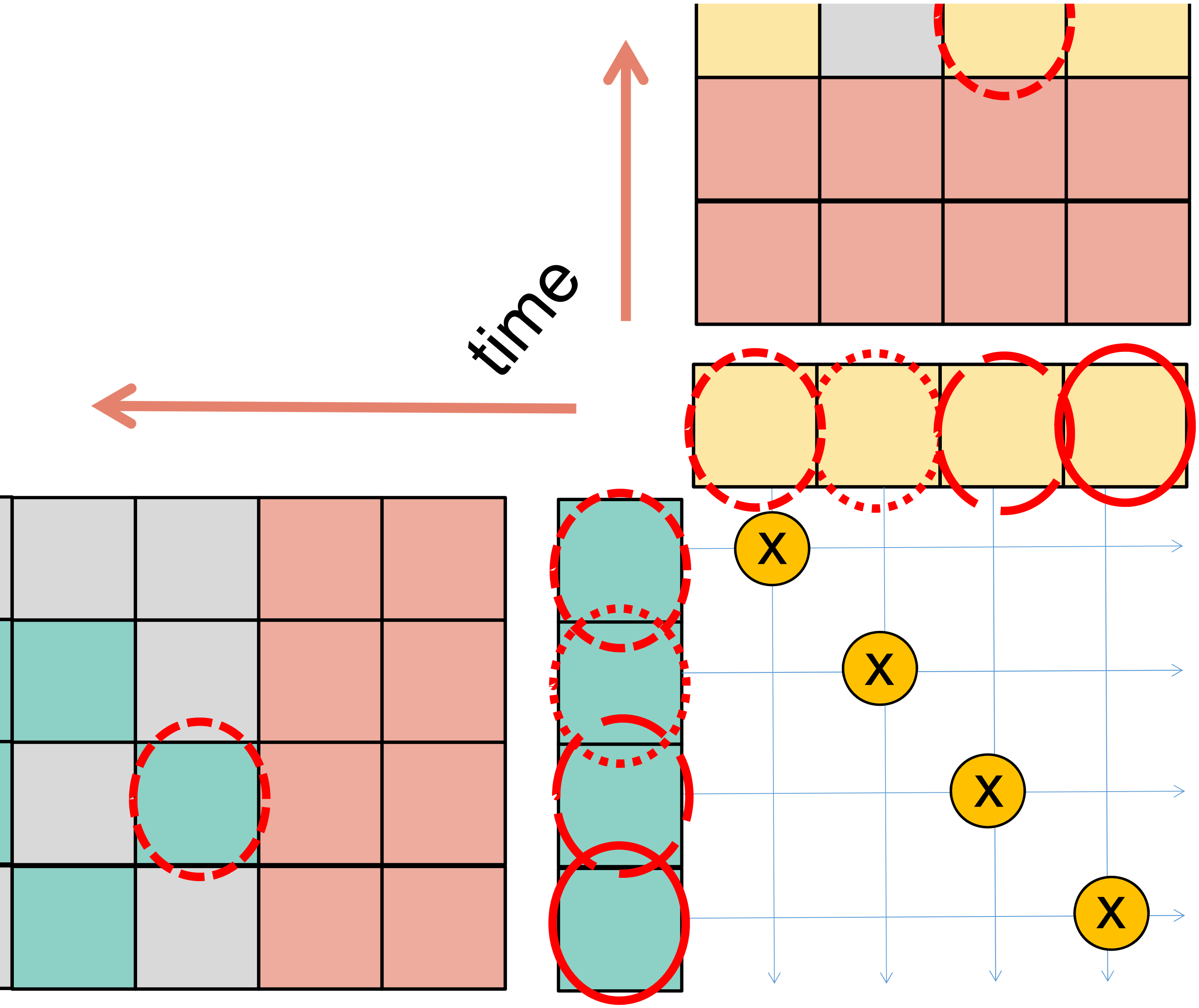

time
X
X
X
X



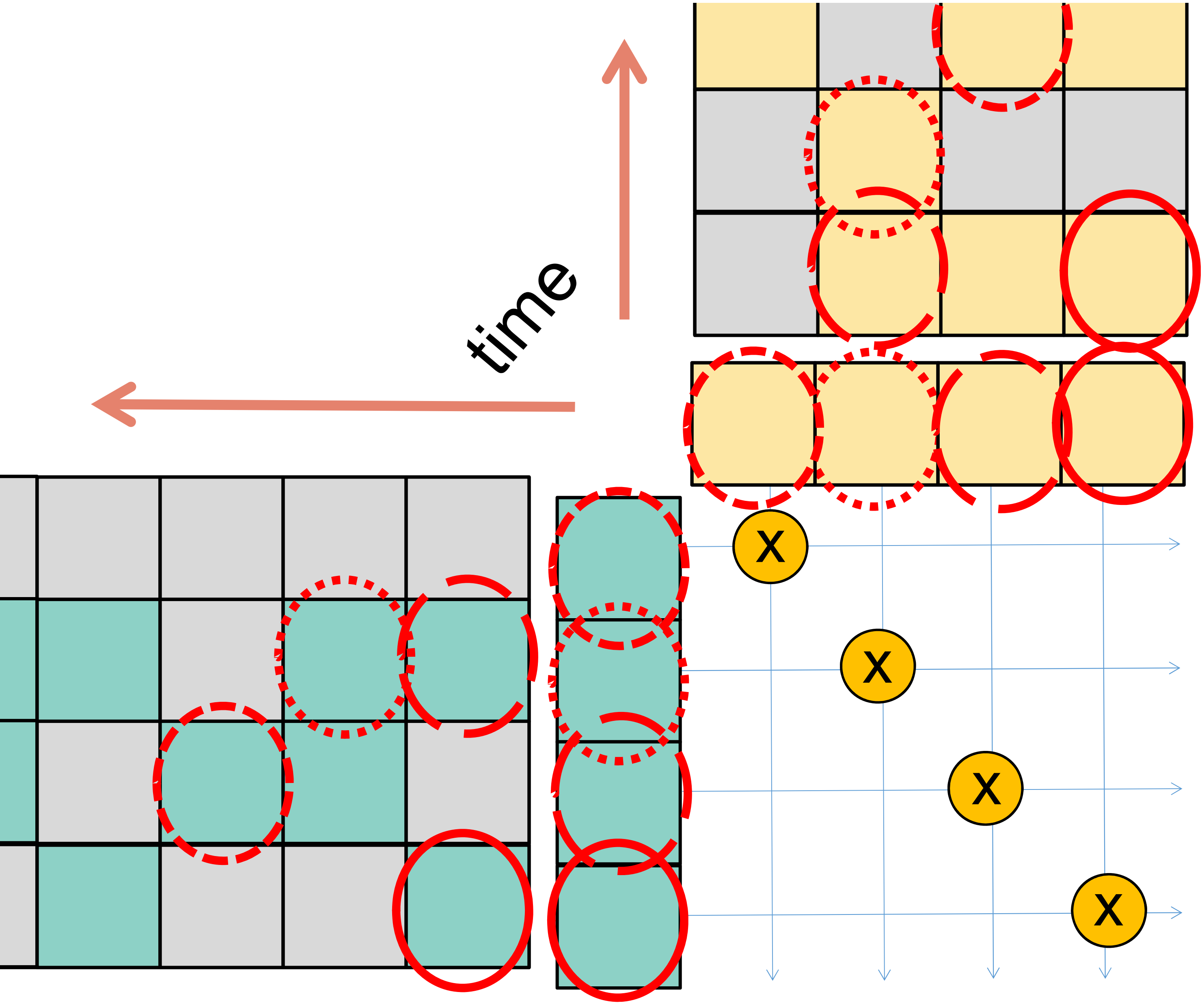
2 steps in time



Another 2 steps in space

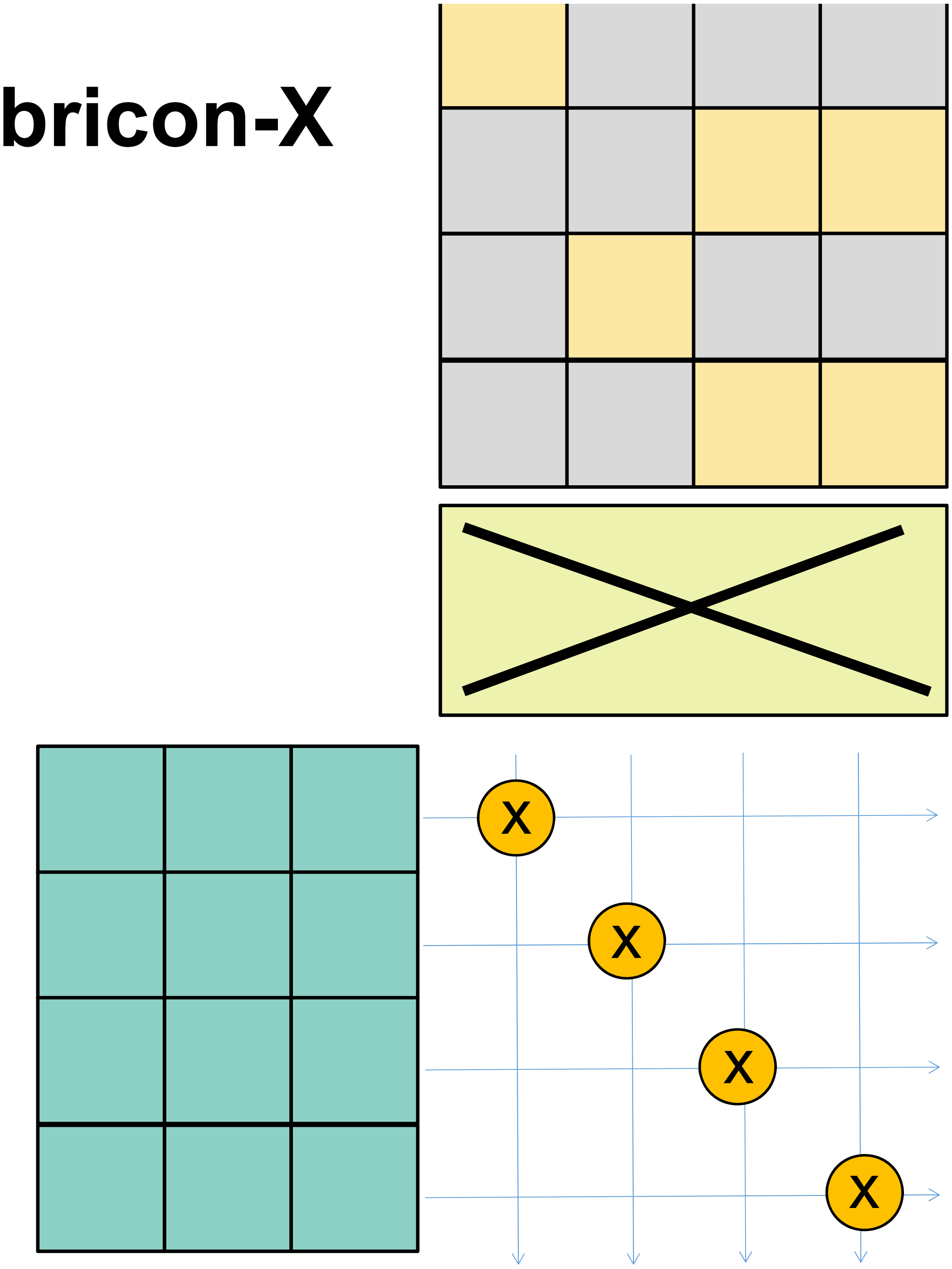


“unrestricted” motion



A: Dynamic Sparsity Pattern

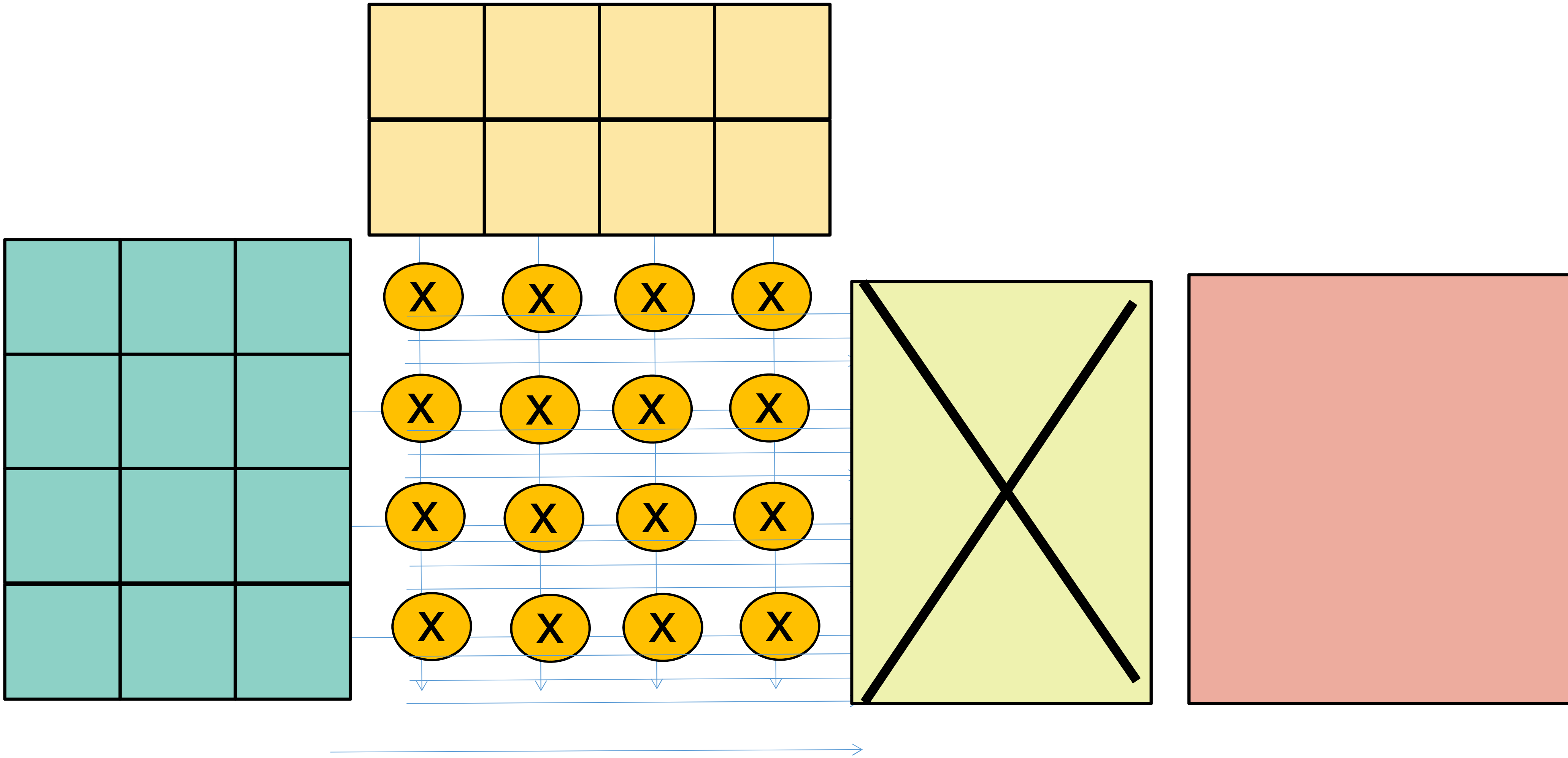
Cambricon-X



time
“any”

space
“any”

Nvidia's SCNN



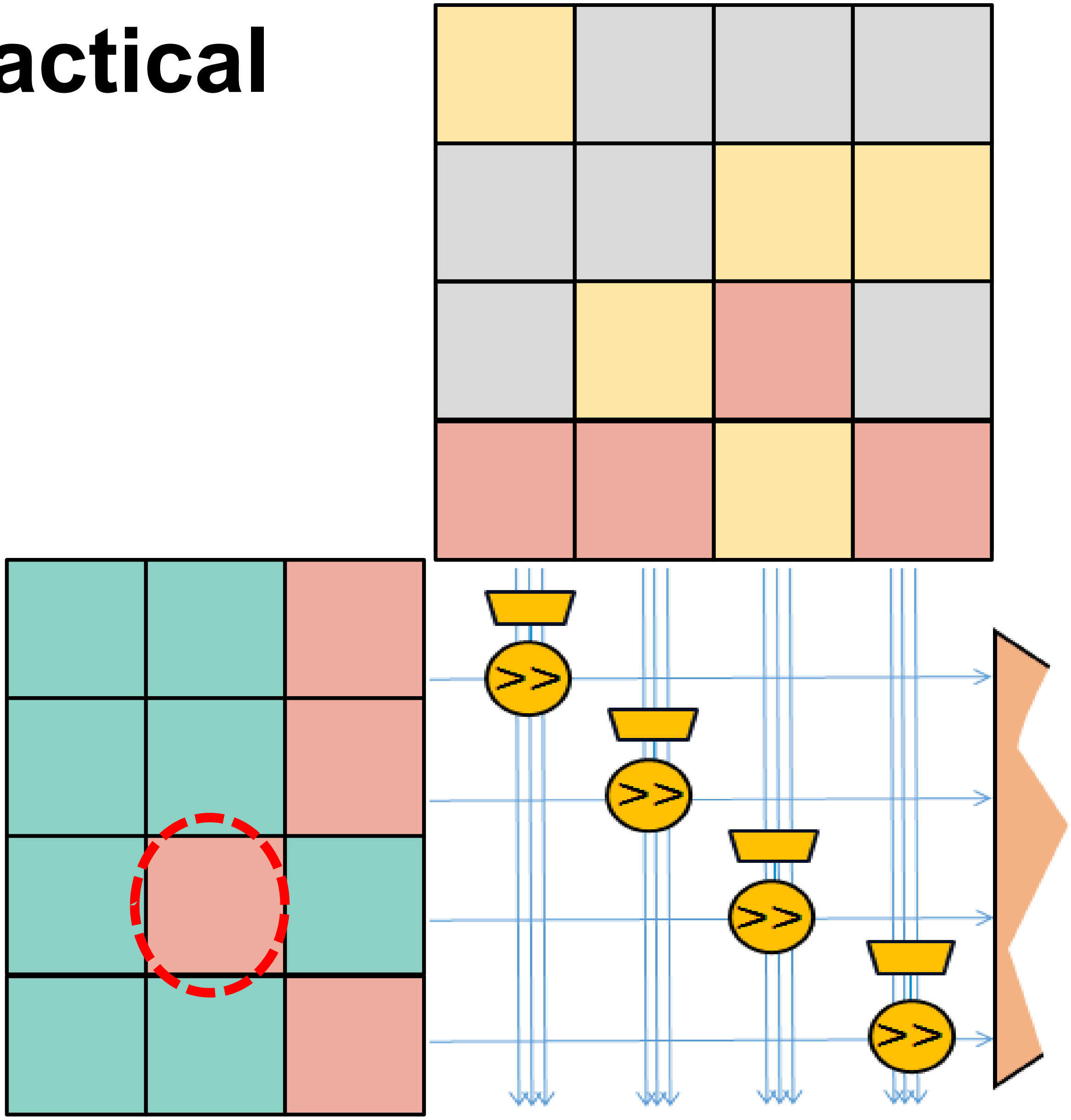
time

"any"

space

"any"

Bit-Tactical



Weights

time

1-2 steps

space

2-5 places

Do as you are told?

Out += A_0 x W_0

Out += A_2 x W_2

Out += 0 x W_3

Out += A_4 x W_4

⋮

Do as you are told?

Out += A_0 x W_0

Out += A_2 x W_2

~~Out += 0 x W_3~~

Out += A_4 x W_4

⋮

Do as you are told?

Out += A_0 x W_0

Out += A_2 x W_2

Out += A_3 x W_3

Out += A_4 x W_4

⋮

Do as you are told?

Out += 000100100 x W_0

Out += 000110100 x W_2

Out += 000010000 x W_3

Out += 010000010 x W_4

⋮

Do as you are told?

Out += 000100100 x W_0

Out += 000110100 x W_2

Out += 000010000 x W_3

Out += 010000010 x W_4

⋮

Do as you are told?

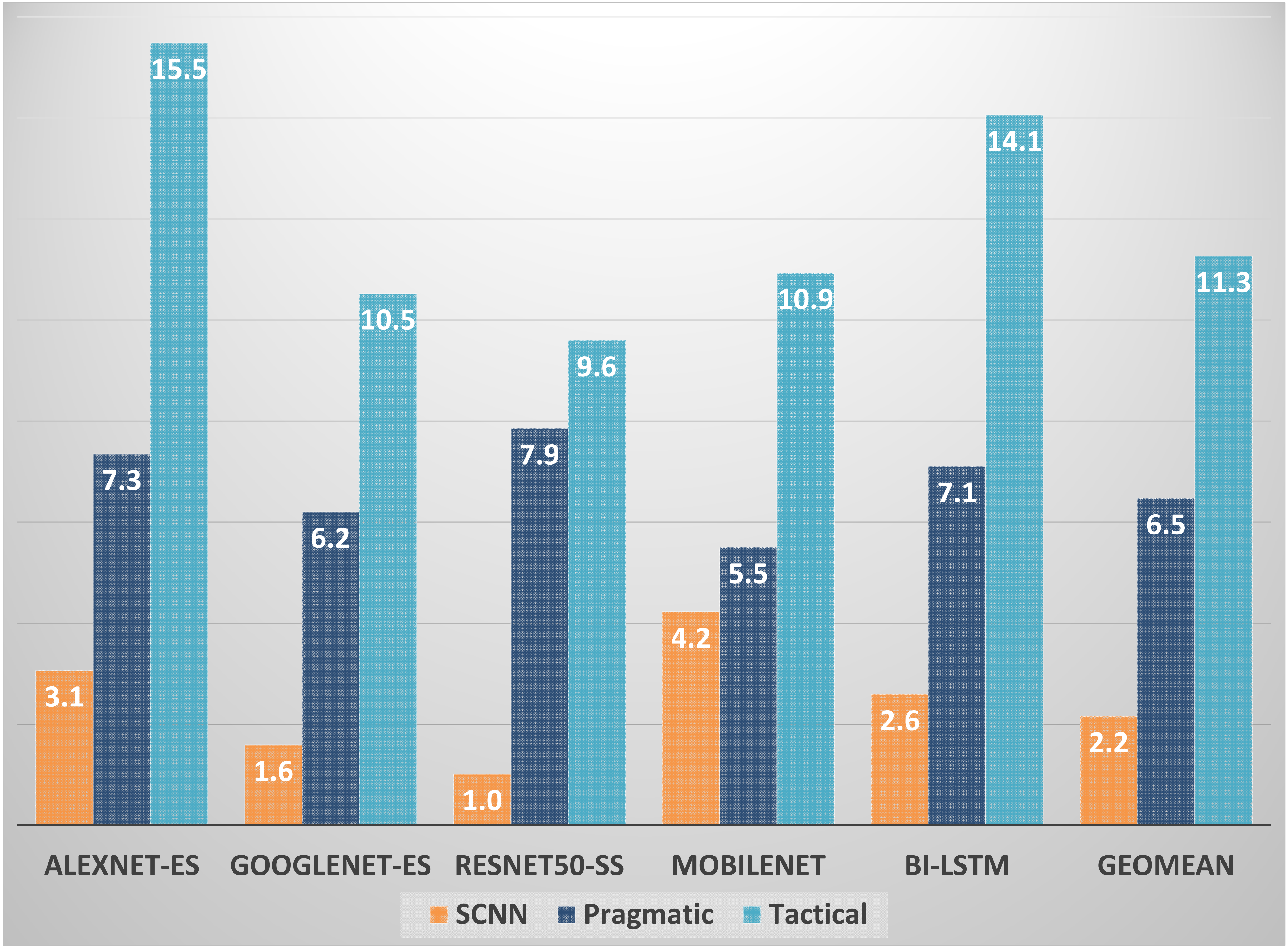
Out += 1 1 x W_0

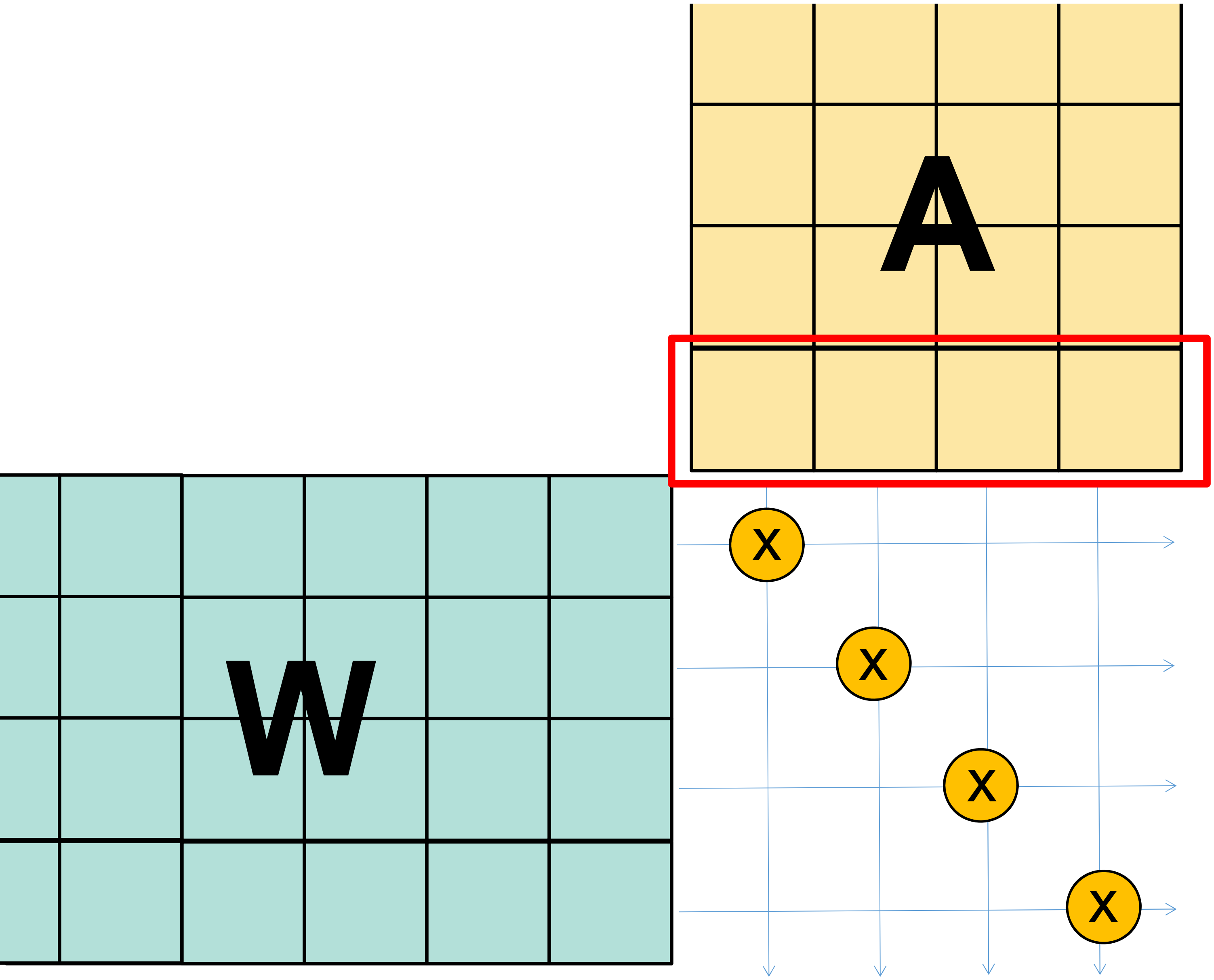
Out += 11 1 x W_2

Out += 1 x W_3

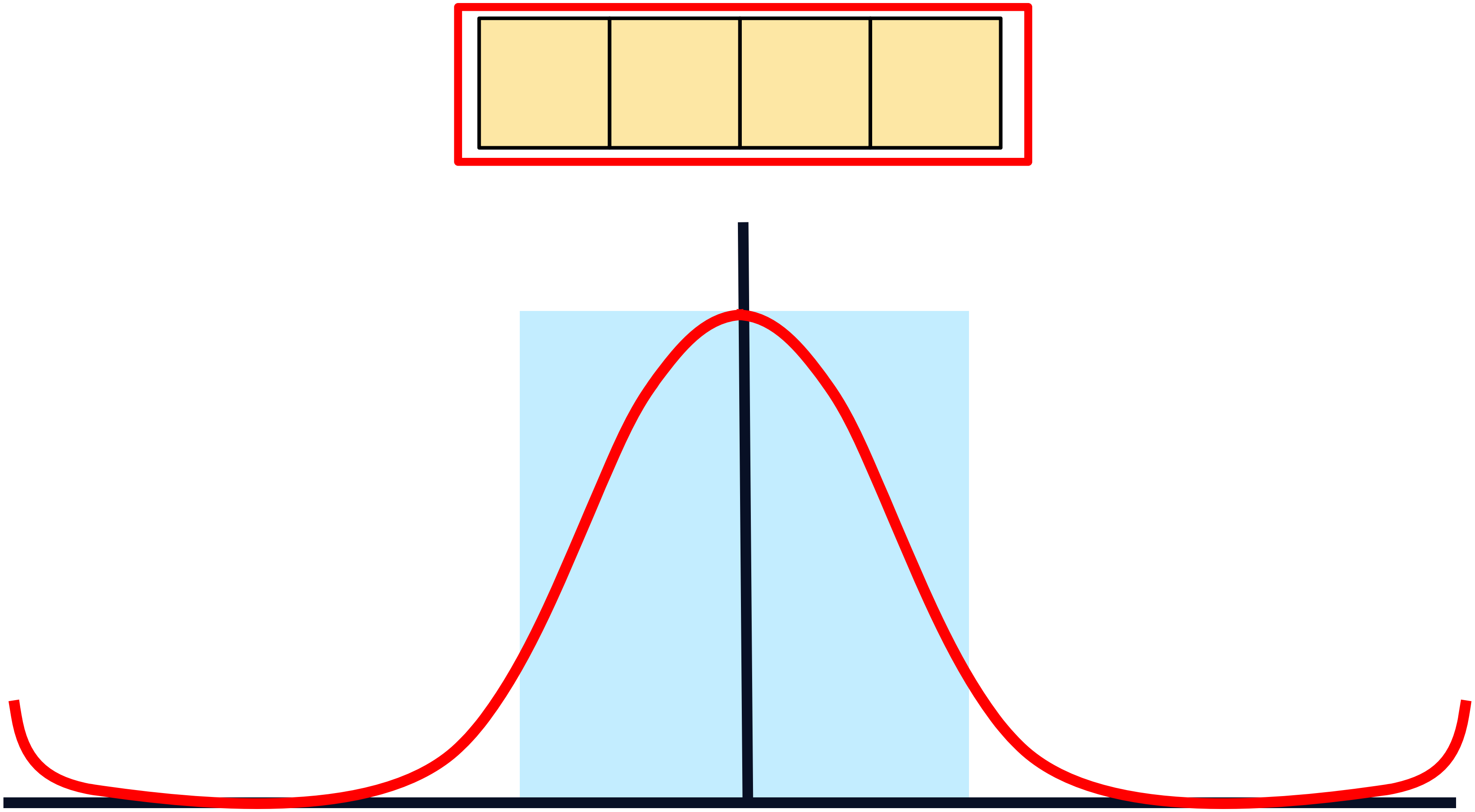
Out += 1 1 x W_4

⋮


15.5
7.3
3.1
10.5
6.2
1.6
9.6
7.9
1.0
10.9
5.5
4.2
14.1
7.1
2.6
11.3
6.5
2.2
ALEXNET-ES
GOOGLENET-ES
RESNET50-SS
MOBILENET
BI-LSTM
GEOMEAN
SCNN
Pragmatic
Tactical

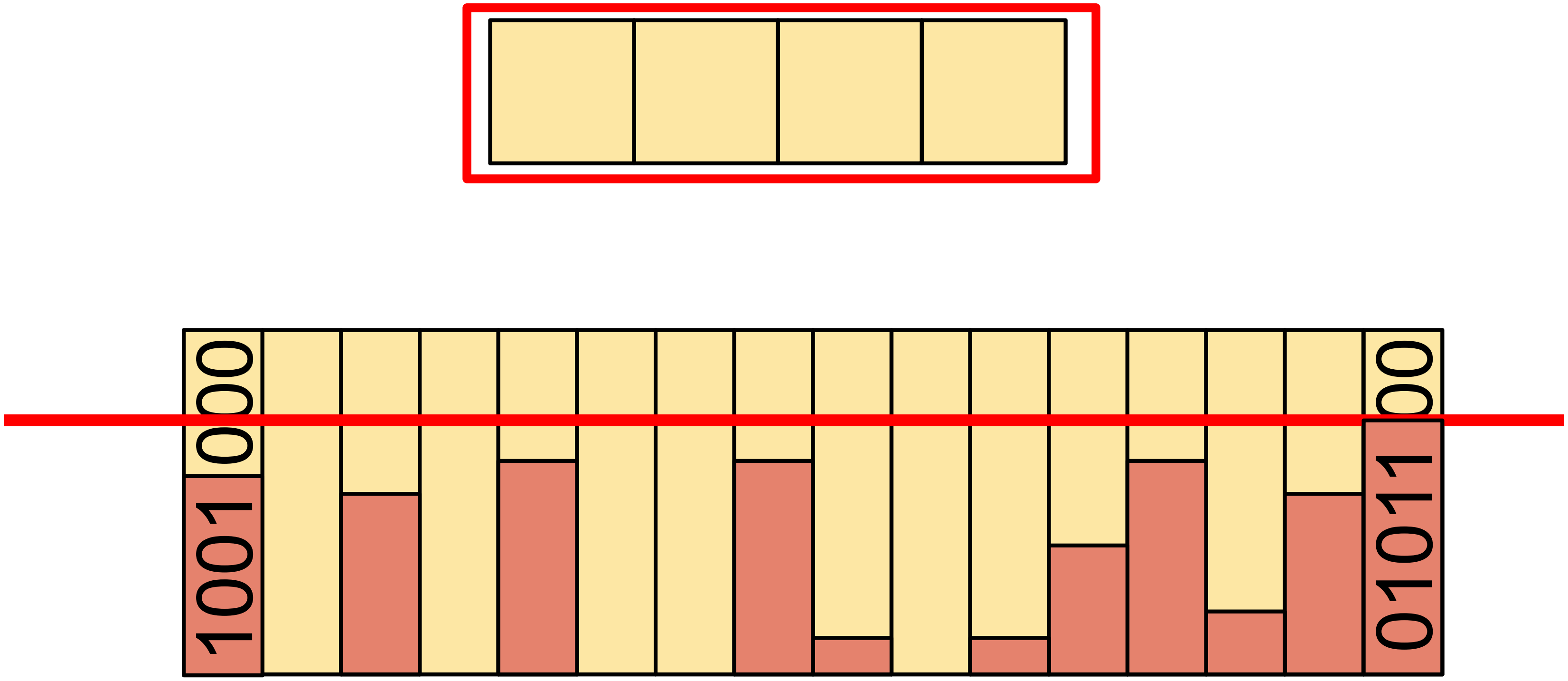

A
W
x
x
x
x

Making Typical Values Matter



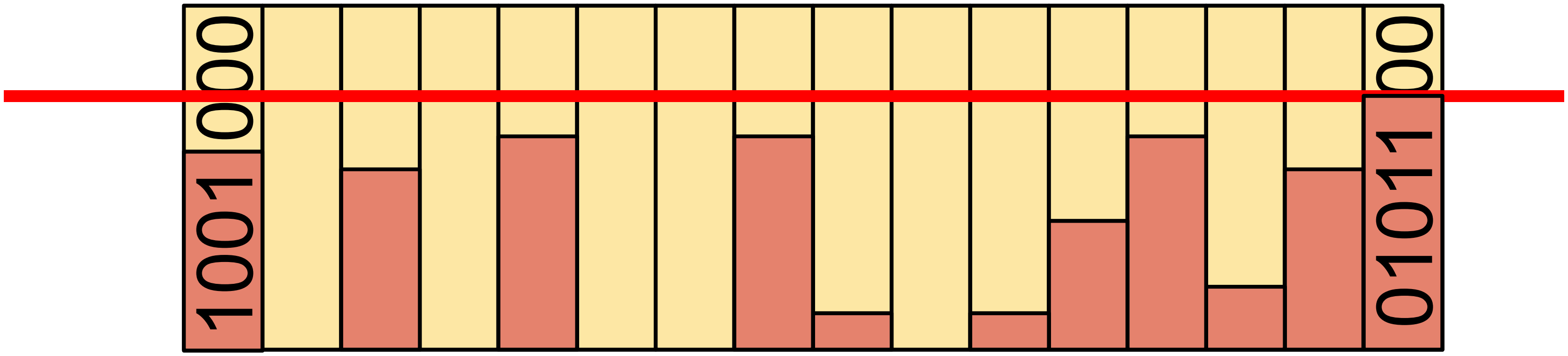
DPRed: Making Typical Activation and Weight Values Matter In Deep Learning Computing, Delmas et al.,
https://arxiv.org/abs/1804.06732

Fine-Grain Precision Adaptation

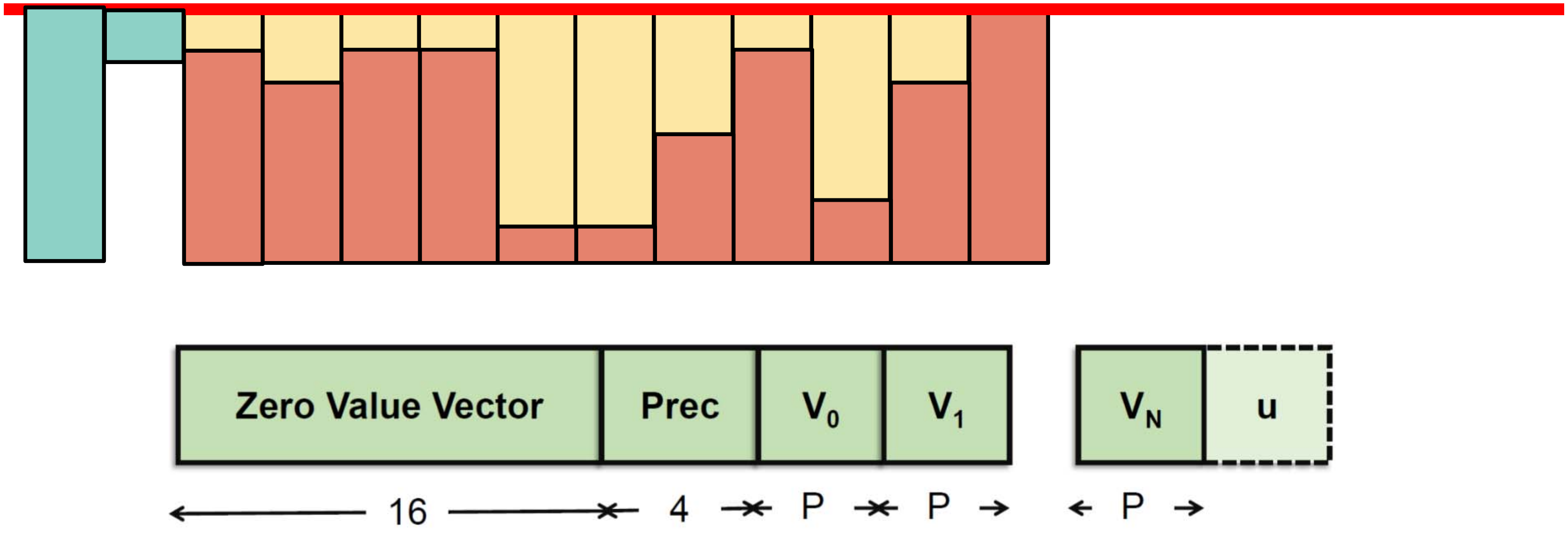


Per Group Precision Detection/Encoding

Detect Precisions: On-the-fly for Activations + Statically for Weights



In Memory



Takeaways

Don’t go after all W = 0

- Go for enough
- Software Scheduler + Restricted Motions are enough

Don’t go after A = 0

- Go for bit sparsity 50% vs 90%+

Don’t let the loud values dominate the data type

- Encode in groups

Bit-Tactical: Exploiting Ineffectual Computations in Convolutional Neural Networks: Which, Why, and How, Alberto Delmas, Patrick Judd, Dylan Malone Stuart, Zissis Poulos, Mostafa Mahmoud, Sayeh Sharify, Milos Nikolic, Andreas Moshovos, arXiv:1803.03688

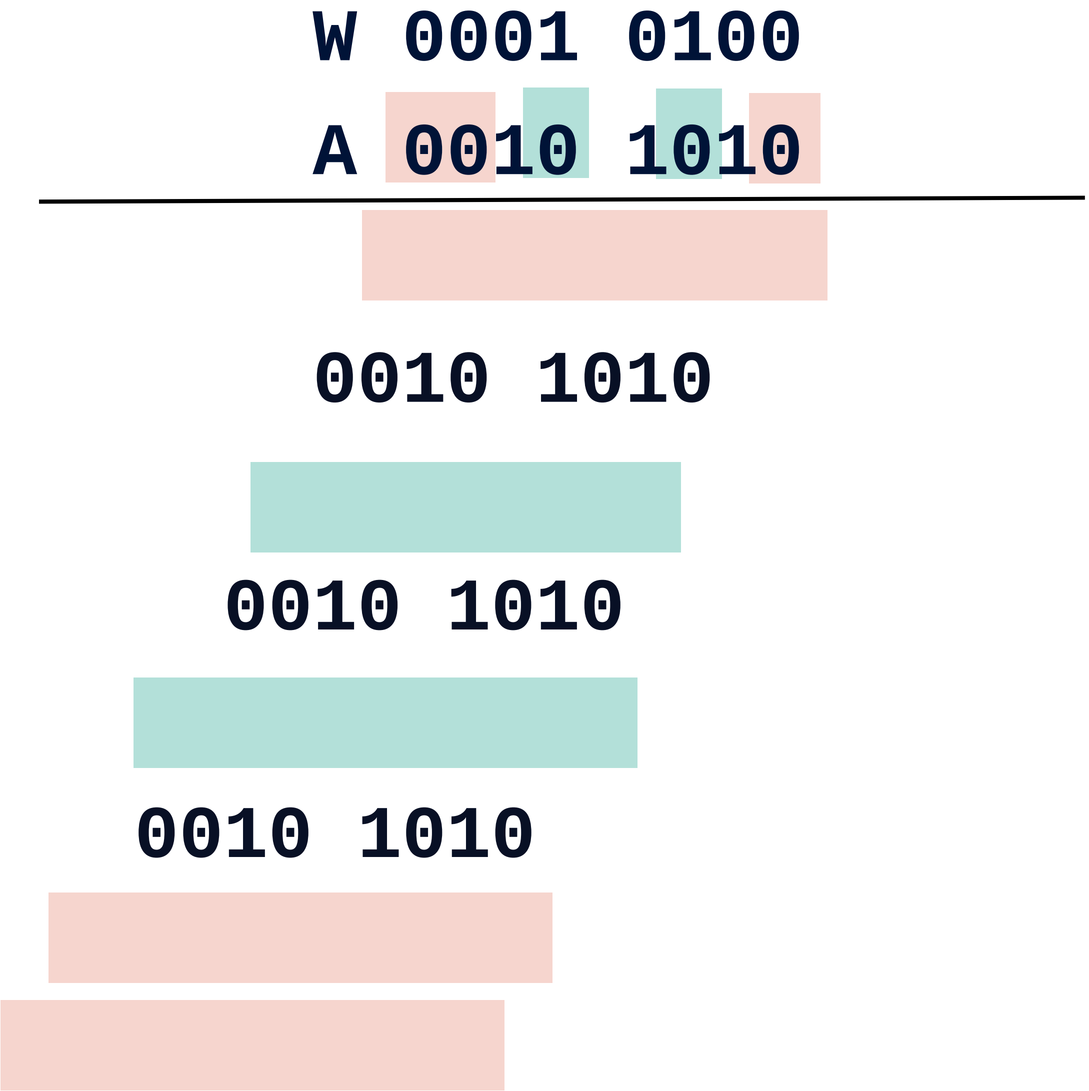
Laconic: Goal

```
      W 0001 0100
      A 0010 1010
-----------------------
       0000 0000
      0010 1010
     0000 0000
    0010 1010
   0000 0000
  0010 1010
 0000 0000
0000 0000
```

Laconic: Goal

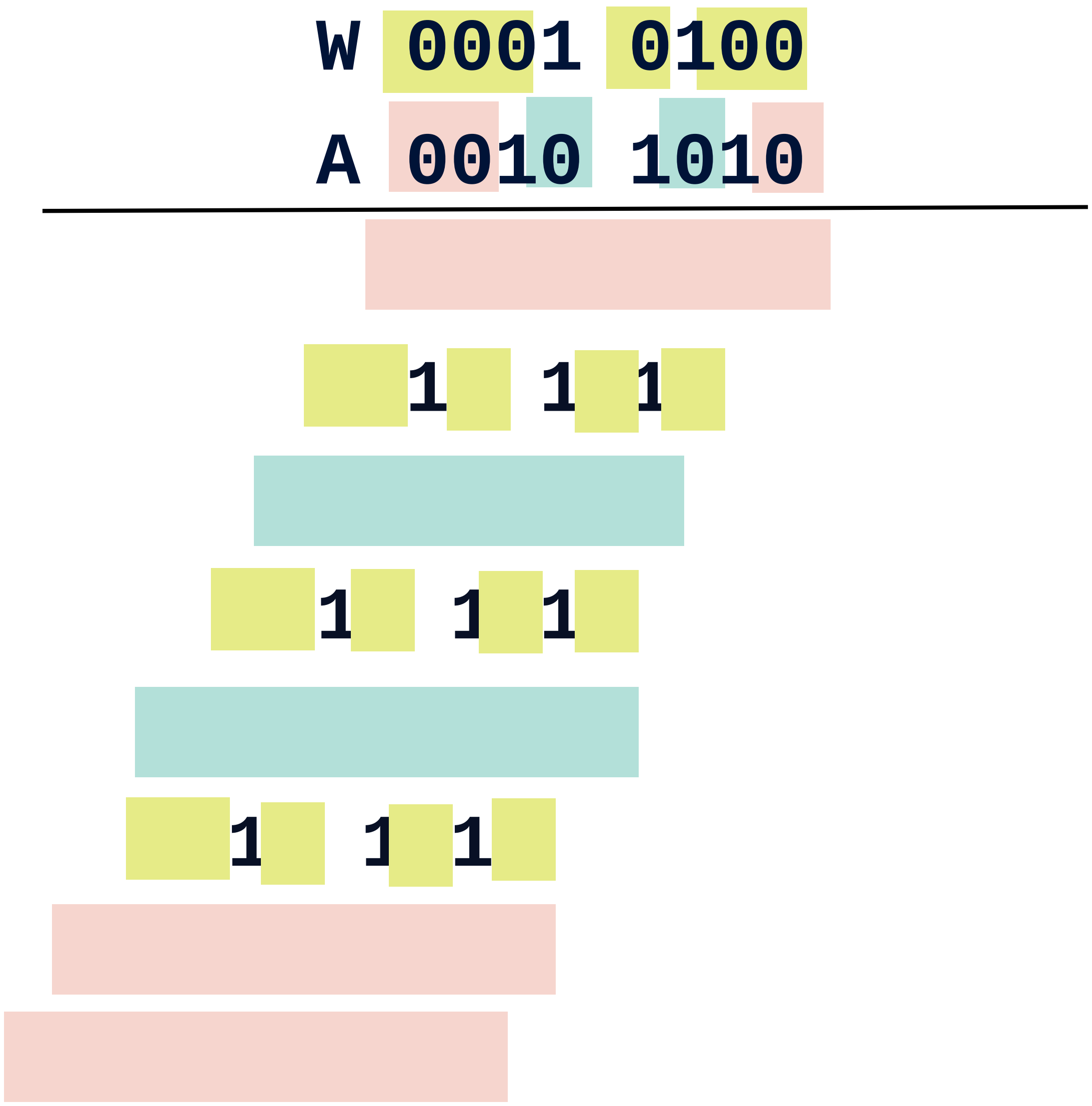
```
    W 0001 0100
    A 0010 1010

      0010 1010
     0000 0000
    0010 1010
   0000 0000
  0010 1010
```

Laconic: Goal



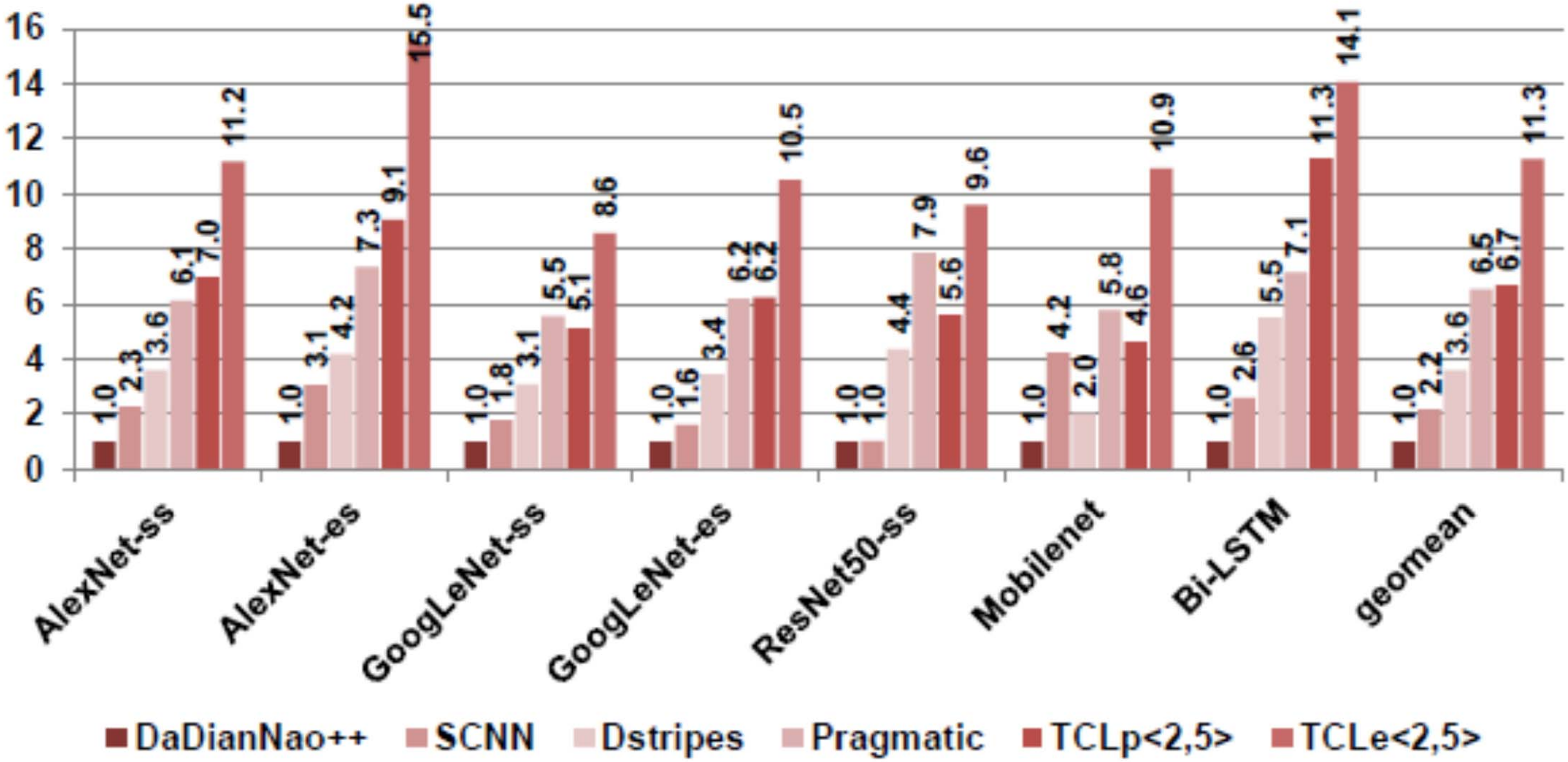
```
      W 0001 0100
      A 0010 1010
 -------------------------
      0010 1010
    0010 1010
  0010 1010
```

Laconic: Goal



Laconic

- **Best for edge devices**
- **Laconic Deep Learning Computing,** *Sayeh Sharify, Mostafa Mahmoud, Alberto Delmas Lascorz, Milos Nikolic, Andreas Moshovos, Arxiv,* **arXiv:1805.04513**


16
14
12
10
8
6
4
2
0
AlexNet-ss: 1.0, 2.3, 3.6, 6.1, 7.0, 11.2
AlexNet-es: 1.0, 3.1, 4.2, 7.3, 9.1, 15.5
GoogLeNet-ss: 1.0, 1.8, 3.1, 5.5, 5.1, 8.6
GoogLeNet-es: 1.0, 1.6, 3.4, 6.2, 6.2, 10.5
ResNet50-ss: 1.0, 1.0, 4.4, 7.9, 5.6, 9.6
Mobilenet: 1.0, 4.2, 2.0, 5.8, 4.6, 10.9
Bi-LSTM: 1.0, 2.6, 5.5, 7.1, 11.3, 14.1
geomean: 1.0, 2.2, 3.6, 6.5, 6.7, 11.3
DaDianNao++
SCNN
Dstripes
Pragmatic
TCLp<2,5>
TCLe<2,5>

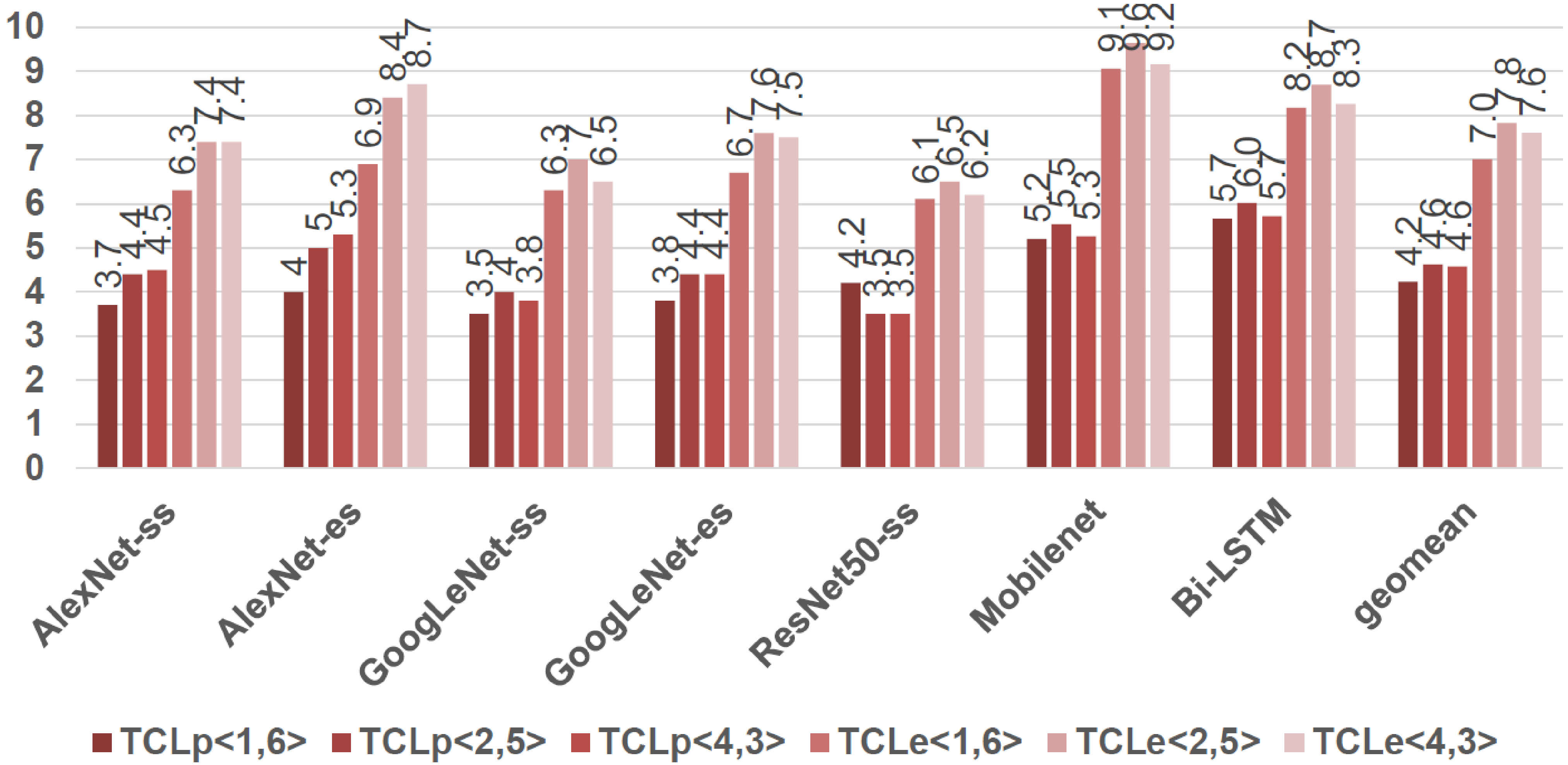


Figure 12: Speedup with 8b quantization