# Energy-Efficient Machine Learning Acceleration

**Jorge Albericio, Patrick Judd, Alberto Delmas, Mostafa Mahmoud, Milos Nikolic, Zissis Poulos, Sayeh Sharify, Kevin Siu, Dylan Stuart**

**Isak Edo, Omar Awad, Ali Hadi Zadeh**

*Tayler Hetherington, Tor Aamodt, Natalie Enright Jerger*

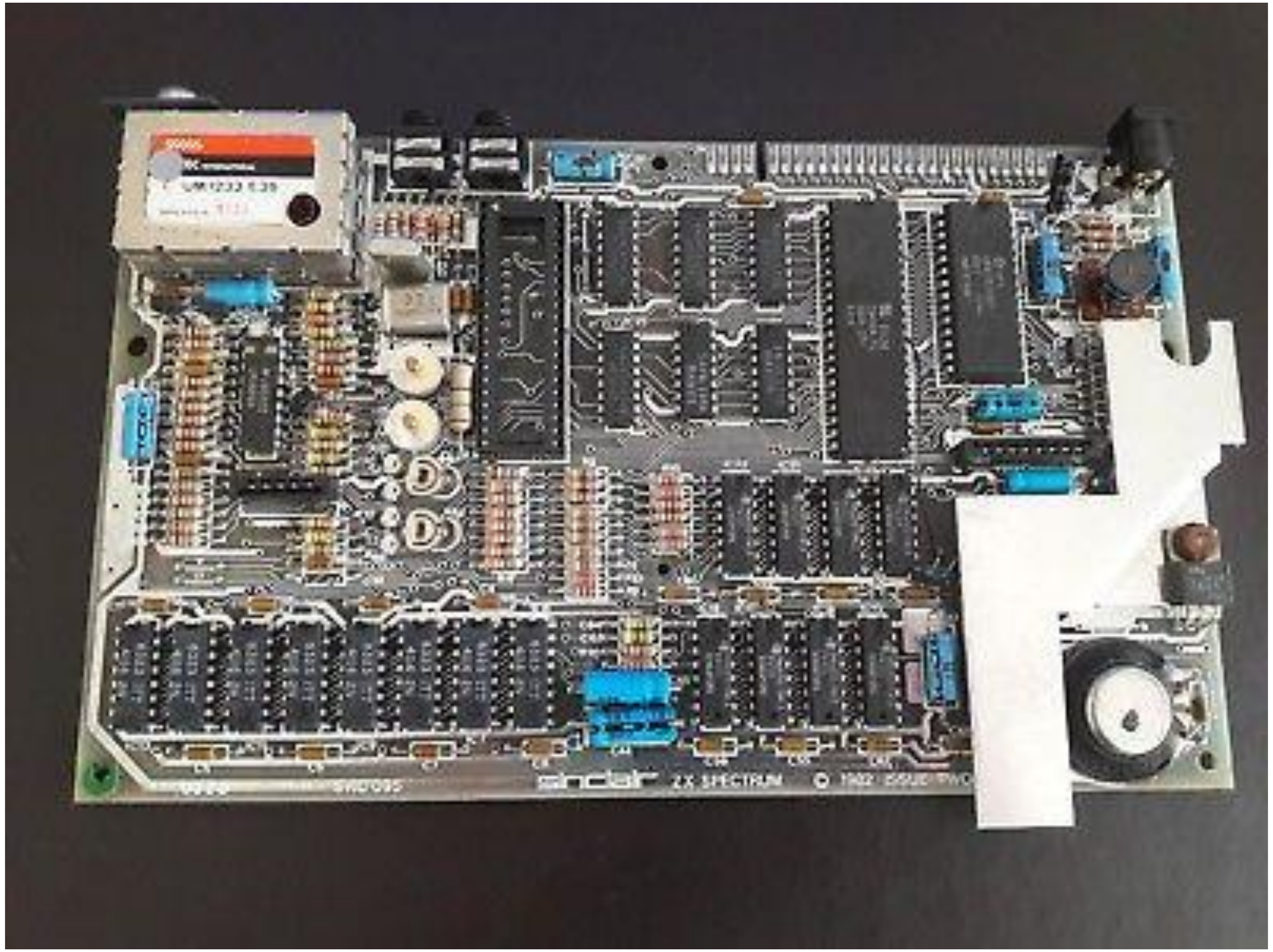
**Andreas Moshovos**

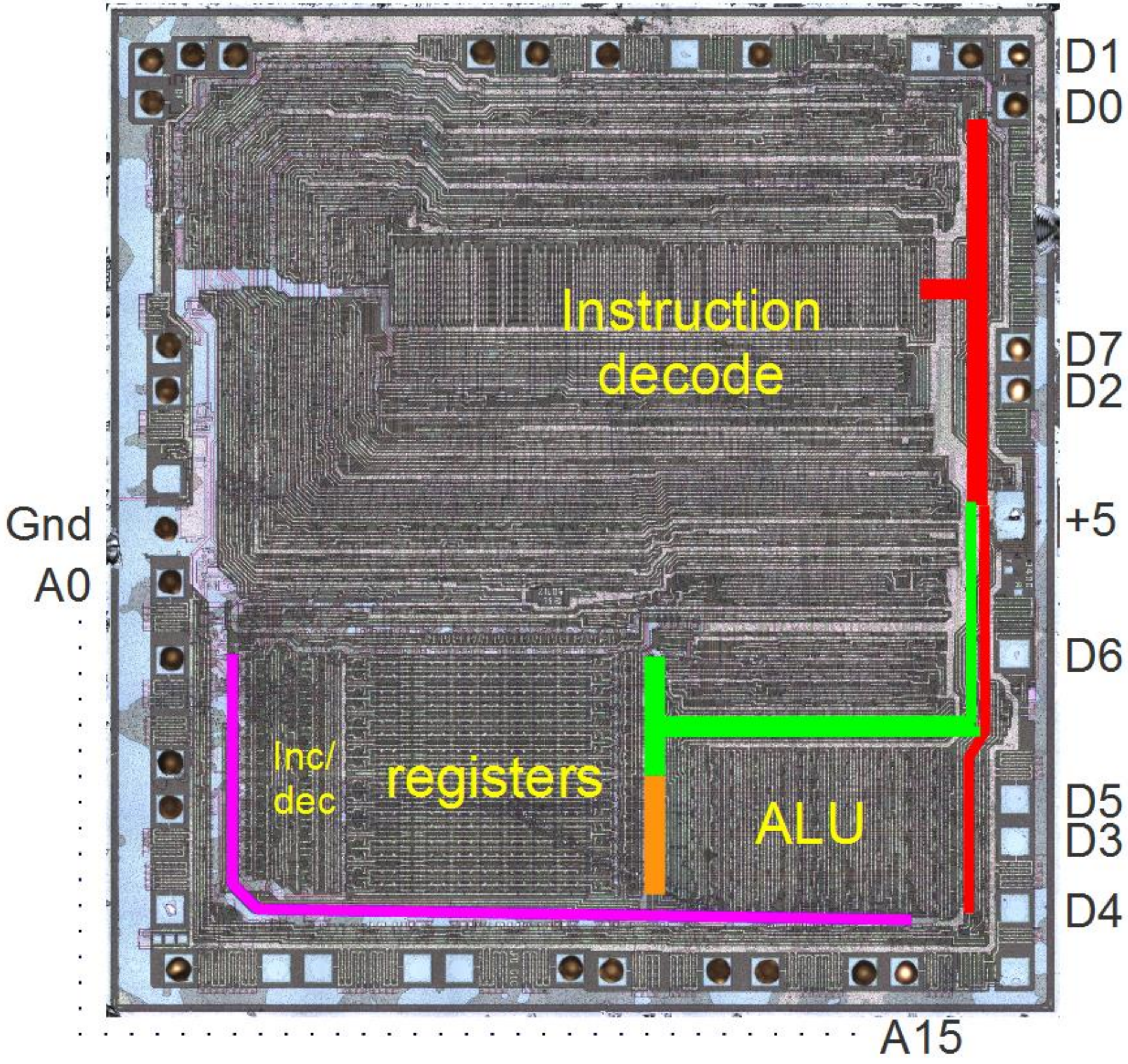
**moshovos@ece.utoronto.ca**


ZX Spectrum
BLUE EDIT 1 DEF FN
RED CAPS LOCK 2 FN
MAGENTA TRUE VIDEO 3 LINE
GREEN INV. VIDEO 4 OPEN #
CYAN 5 CLOSE #
YELLOW 6 MOVE
WHITE 7 ERASE
8 POINT
GRAPHICS 9 CAT
BLACK DELETE 0 FORMAT
SIN Q PLOT ASN
COS W DRAW ACS
TAN E REM ATN
INT R RUN VERIFY
RND T RAND MERGE
STR $ Y AND RETURN
CHR $ U OR IF
CODE I AT INPUT IN
PEEK O POKE OUT
TAB P PRINT
READ A STOP NEW
RESTORE S NOT SAVE
DATA D STEP DIM
SGN F TO FOR
ABS G THEN GOTO
SQR H GOSUB CIRCLE
VAL J LOAD VAL $
LEN K LIST SCREEN $
USR L LET ATTR
ENTER
CAPS SHIFT
LN Z COPY BEEP
EXP X CLEAR INK
L PRINT C CONT PAPER
L LIST V CLS FLASH
BIN B BORDER BRIGHT
IN KEY $ N NEXT OVER
PI M PAUSE INVERSE
SYMBOL SHIFT
BREAK SPACE

```
   10>LET a=10
   20 PRINT a
```

K





D1
D0
Instruction
decode
D7
D2
Gnd
+5
A0
D6
Inc/
dec
registers
ALU
D5
D3
D4
A15


12
10

## Computing Hardware

We build tools

Used by “everyone” for “everything”

Science, medicine, commerce, …

# www.eecg.toronto.edu/~moshovos/CUDA08

www.eecg.toronto.edu/~moshovos/CUDA08/doku.php?id=project_presentations_reports_source_code

ECE1742S: Programming Massively Parallel Multiprocessors Using CUDA

Trace: start → project_presentations_reports_source_code

## Convolutional Neural Networks for Object Classification


Alex Krizhevsky

Department of Computer Science, University of Toronto

akrizhevsky at gmail.com



**Abstract** I implemented a convolutional neural network with one layer of convolution. I tested it on the CIFAR-10 dataset, which consists of 6000 32×32 colour images in each of 10 classes. The convolutional net does well on the classification task and takes roughly 140x less time to train than a CPU implementation.


- Presentation: (pdf).
- Report: (pdf).
- Source code: (zip).

# **Our Goal:**

Hardware Acceleration of ML

**Why?**

Hope to… Further Enable ML:

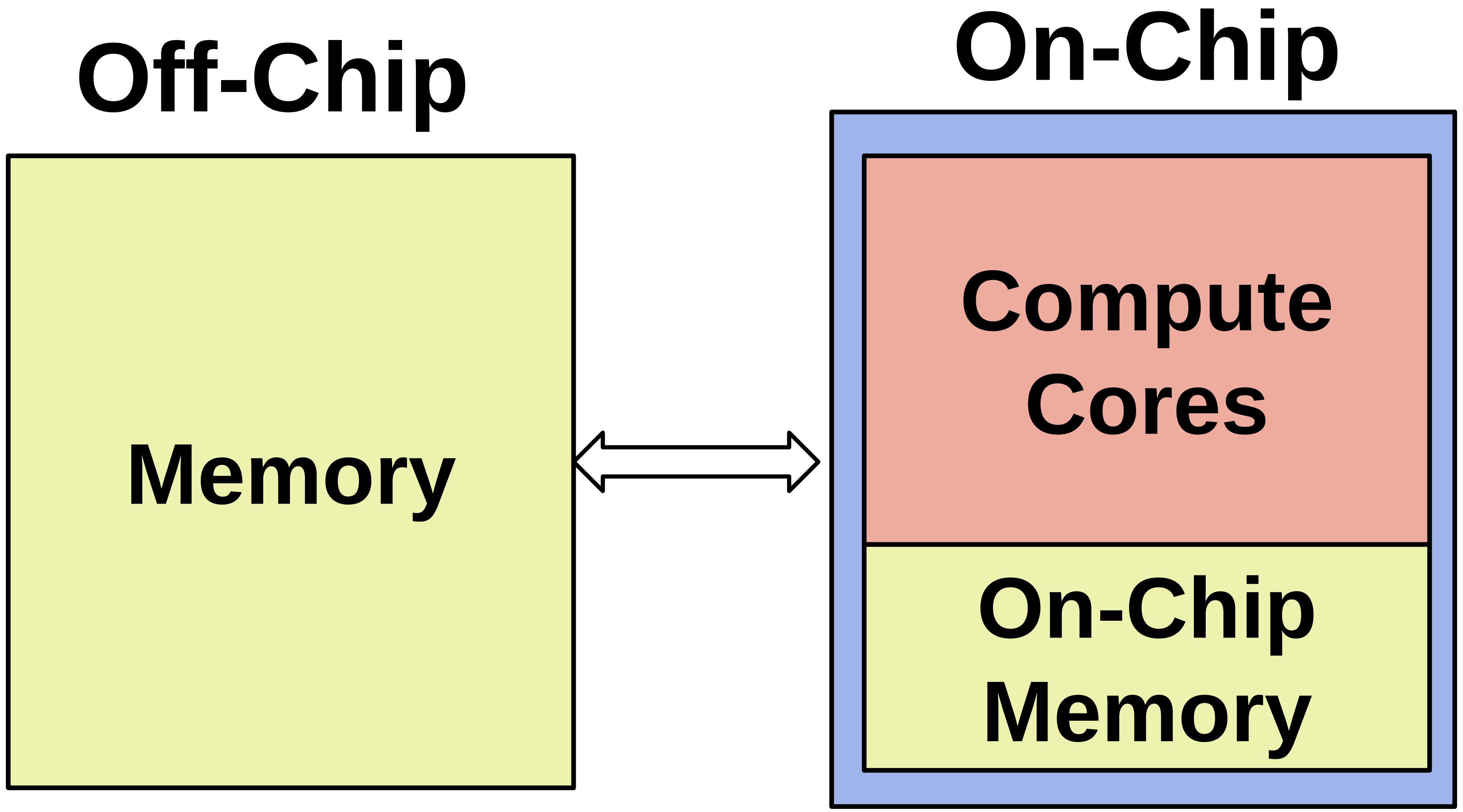
*Innovation*

*Applications*

# Computing Devices do…

Data "Transformation" → A + B

Data Movement → A = B




**On- vs. Off-Chip**

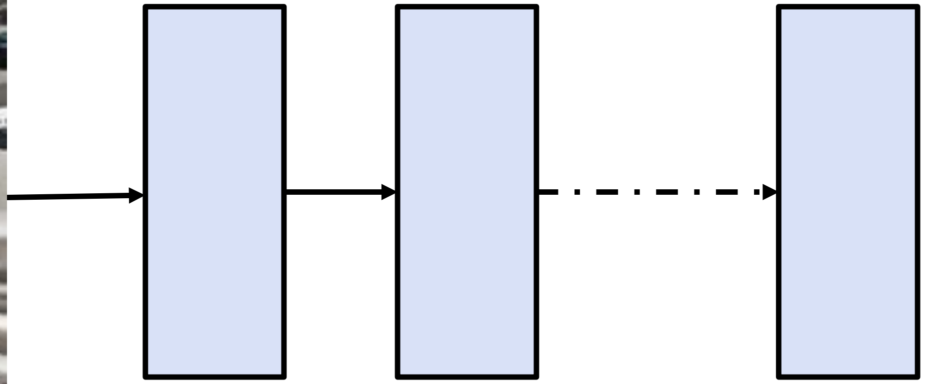
Energy: ~100x

Latency: ~50x

**Compute/Watt is the primary design constraint**

# Example: Convolutional Neural Networks

**Layers**



*LHC*
**maybe**

Tons of **Out += A x W**

For other types of networks too

$$Out_0 \mathrel{+}= A_0 \times W_0$$

$$Out_0 \mathrel{+}= A_1 \times W_1$$

$$Out_0 \mathrel{+}= A_2 \times W_2$$

$$Out_0 \mathrel{+}= A_3 \times W_3$$

$$Out_0 \mathrel{+}= A_4 \times W_4$$

$$\vdots$$

$Out_0 \mathrel{+}= A_0 \times W_0$

$Out_0 \mathrel{+}= A_1 \times W_1$

# Lots of Parallelism

$Out_0 \mathrel{+}= A_3 \times W_3$

$Out_0 \mathrel{+}= A_4 \times W_4$

$\vdots$

$Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times$

$Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times$

$Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times$

$Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times$

$Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times$

$Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times$

$Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times$

$Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times$

$Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times$

$Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times$

$Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times$

$Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times$

$Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times$

$Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times$

$Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times$

# Do as you are told?

$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0$ x

$Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1$ x

$Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2$ x

$Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3$ x

$Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4$ x

$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0$ x

$Out_0$ $= A_1$ x

$Out_0$ $= A_2$ x

$Out_0$ $= A_3$ x

$Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4$ x

$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0$ x

$Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1$ x

$Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2$ x

$Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3$ x

$Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4$ x

$Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0$ x

$Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1$ x

$Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2$ x

$Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3$ x

$Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4$ x

# Calculate the *same* output but … do less work

$Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0 \times W_0$ $Out_0 += A_0 \times W_0$ $Out_1 += A_0$ x

$Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1 \times W_1$ $Out_0 += A_1 \times W_1$ $Out_1 += A_1$ x

$Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2 \times W_2$ $Out_0 += A_2 \times W_2$ $Out_1 += A_2$ x

$Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3 \times W_3$ $Out_0 += A_3 \times W_3$ $Out_1 += A_3$ x

$Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4 \times W_4$ $Out_0 += A_4 \times W_4$ $Out_1 += A_4$ x

$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0$ x

$Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1$ x

$Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2$ x

$Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3$ x

$Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4$ x

$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0$ x

$Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1$ x

$Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2$ x

$Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3$ x

$Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4$ x

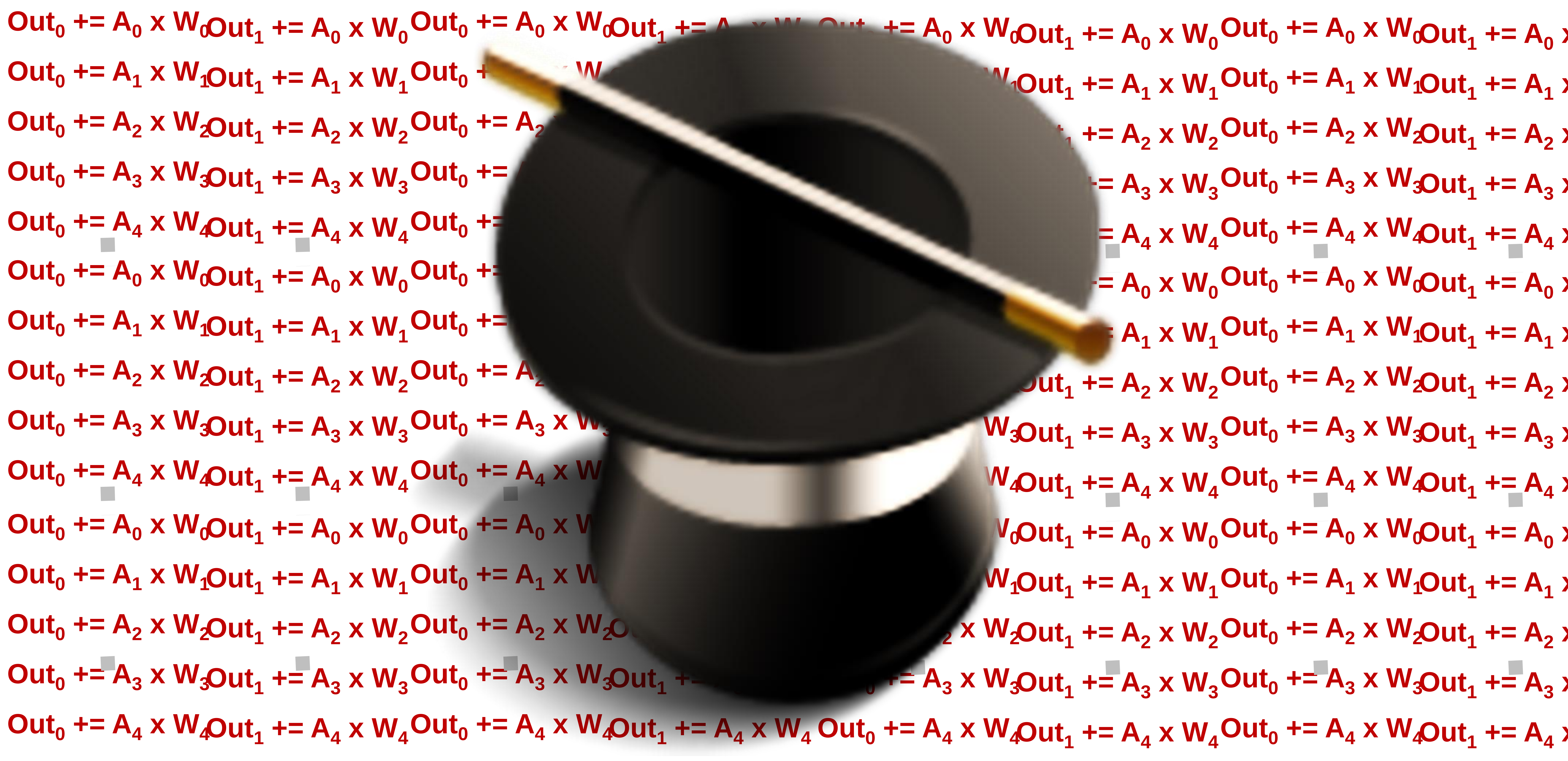
$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0$ x

$Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1$ x

$Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2 \times W_2$ $Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_2$ x

$Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3 \times W_3$ $Out_0 \mathrel{+}= A_3 \times W_3$ $Out_1 \mathrel{+}= A_3$ x

$Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times W_4$ $Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4$ x


Out0 += A0 x W0 Out1 += A0 x W0
Out0 += A1 x W1 Out1 += A1 x W1
Out0 += A2 x W2 Out1 += A2 x W2
Out0 += A3 x W3 Out1 += A3 x W3
Out0 += A4 x W4 Out1 += A4 x W4
Out0 += A0 x W0 Out1 += A0 x W0
Out0 += A1 x W1 Out1 += A1 x W1
Out0 += A2 x W2 Out1 += A2 x W2
Out0 += A3 x W3 Out1 += A3 x W3
Out0 += A4 x W4 Out1 += A4 x W4
Out0 += A0 x W0 Out1 += A0 x W0
Out0 += A1 x W1 Out1 += A1 x W1
Out0 += A2 x W2 Out1 += A2 x W2
Out0 += A3 x W3 Out1 += A3 x W3
Out0 += A4 x W4 Out1 += A4 x W4

$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times$

$Out_0 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times W_1$

$Out_1 \mathrel{+}= A_3 \times W_3$

$Out_1 \mathrel{+}= A_4 \times W_4$ $Out_1 \mathrel{+}= A_4 \times$

$Out_0 \mathrel{+}= A_4 \times W_4$

2

$Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times$

$Out_1 \mathrel{+}= A_3 \times W_3$

$Out_1 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_0 \mathrel{+}= A_1 \times W_1$ $Out_1 \mathrel{+}= A_1 \times$

$Out_0 \mathrel{+}= A_2 \times W_2$ $Out_1 \mathrel{+}= A_4 \times W_4$

$Out_0 \mathrel{+}= A_1 \times W_1$

$Out_0 \mathrel{+}= A_3 \times W_3$

$Out_0 \mathrel{+}= A_4 \times W_4$

$Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times$

$Out_0 \mathrel{+}= A_1 \times W_1$

$Out_1 \mathrel{+}= A_4 \times$

$Out_1 \mathrel{+}= A_4 \times W_4$

$Out_0 \mathrel{+}= A_4 \times W_4$

$Out_1 \mathrel{+}= A_4 \times W_4$

$\text{Out}_0 \mathrel{+}= A_0 \times W_0$

$\text{Out}_0 \mathrel{+}= A_4 \times W_4$

$\text{Out}_1 \mathrel{+}= A_0 \times W_0$

$\text{Out}_1 \mathrel{+}= A_3 \times W_3$

$\text{Out}_0 \mathrel{+}= A_0 \times W_0$

$\text{Out}_1 \mathrel{+}= A_0 \times W_0$

$\text{Out}_1 \mathrel{+}= A_1 \times W_1$

$\text{Out}_1 \mathrel{+}= A_4 \times W_4$

$\text{Out}_0 \mathrel{+}= A_0 \times W_0$

$\text{Out}_1 \mathrel{+}= A_0 \times W_0$

$\text{Out}_1 \mathrel{+}= A_1 \times W_1$

$\text{Out}_0 \mathrel{+}= A_4 \times W_4$

$\text{Out}_1 \mathrel{+}= A_0$ x

$\text{Out}_1 \mathrel{+}= A_4$ x

$\text{Out}_0 \mathrel{+}= A_2 \times W_2$

$\text{Out}_0 \mathrel{+}= A_3 \times W_3$

$\text{Out}_0 \mathrel{+}= A_4 \times W_4$

$\text{Out}_1 \mathrel{+}= A_0 \times W_0$

$\text{Out}_1 \mathrel{+}= A_1 \times W_1$

$\text{Out}_0 \mathrel{+}= A_0 \times W_0$

$\text{Out}_0 \mathrel{+}= A_1 \times W_1$

2

$\text{Out}_1 \mathrel{+}= A_3 \times W_3$

$\text{Out}_1 \mathrel{+}= A_4 \times W_4$

$\text{Out}_0 \mathrel{+}= A_0 \times W_0$

$\text{Out}_0 \mathrel{+}= A_1 \times W_1$

$\text{Out}_1 \mathrel{+}= A_0 \times W_0$

$\text{Out}_0 \mathrel{+}= A_1 \times W_1$

$\text{Out}_1 \mathrel{+}= A_0$ x

$\text{Out}_1 \mathrel{+}= A_1$ x

$\text{Out}_0 \mathrel{+}= A_4 \times W_4$

$\text{Out}_1 \mathrel{+}= A_0 \times W_0$

$\text{Out}_0 \mathrel{+}= A_0 \times W_0$

$\text{Out}_0 \mathrel{+}= A_1 \times W_1$

$\text{Out}_1 \mathrel{+}= A_0 \times W_0$

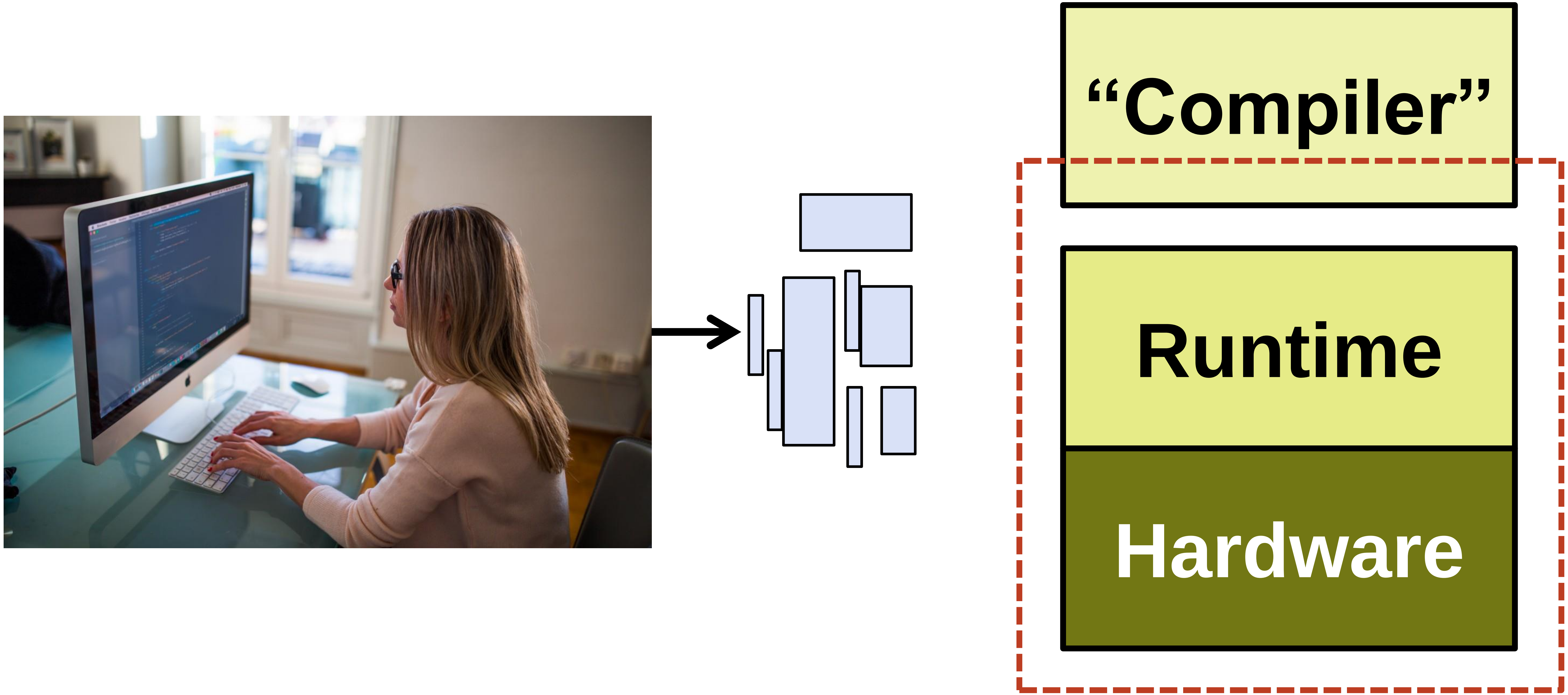
$\text{Out}_1 \mathrel{+}= A_4 \times W_4$

$\text{Out}_1 \mathrel{+}= A_4 \times W_4$

$\text{Out}_0 \mathrel{+}= A_0 \times W_0$

$\text{Out}_1 \mathrel{+}= A_0$ x

$\text{Out}_1 \mathrel{+}= A_4$ x



“Compiler”
Runtime
Hardware

# Value Properties to Exploit? Many ~0 values

- Out += $A_0$ x $W_0$

# 0 x W

# ~0 x W

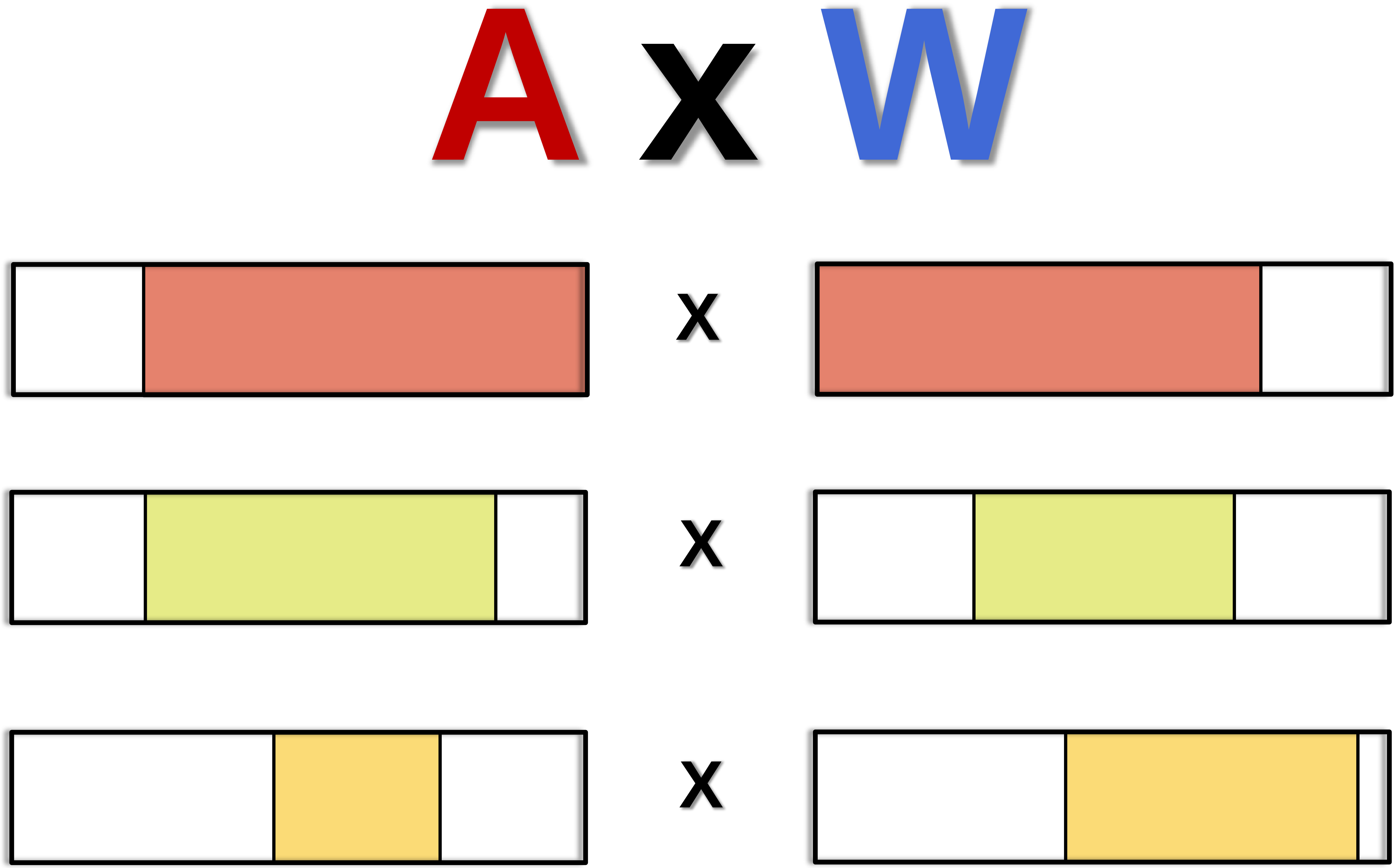
# Value Properties to Exploit? Many ~0 values

- Out += $A_0$ x $W_0$

0 x W

~50%

# Value Properties to Exploit? Varying Precision Needs




- **At Various Granularities: e.g., Layer or finer**

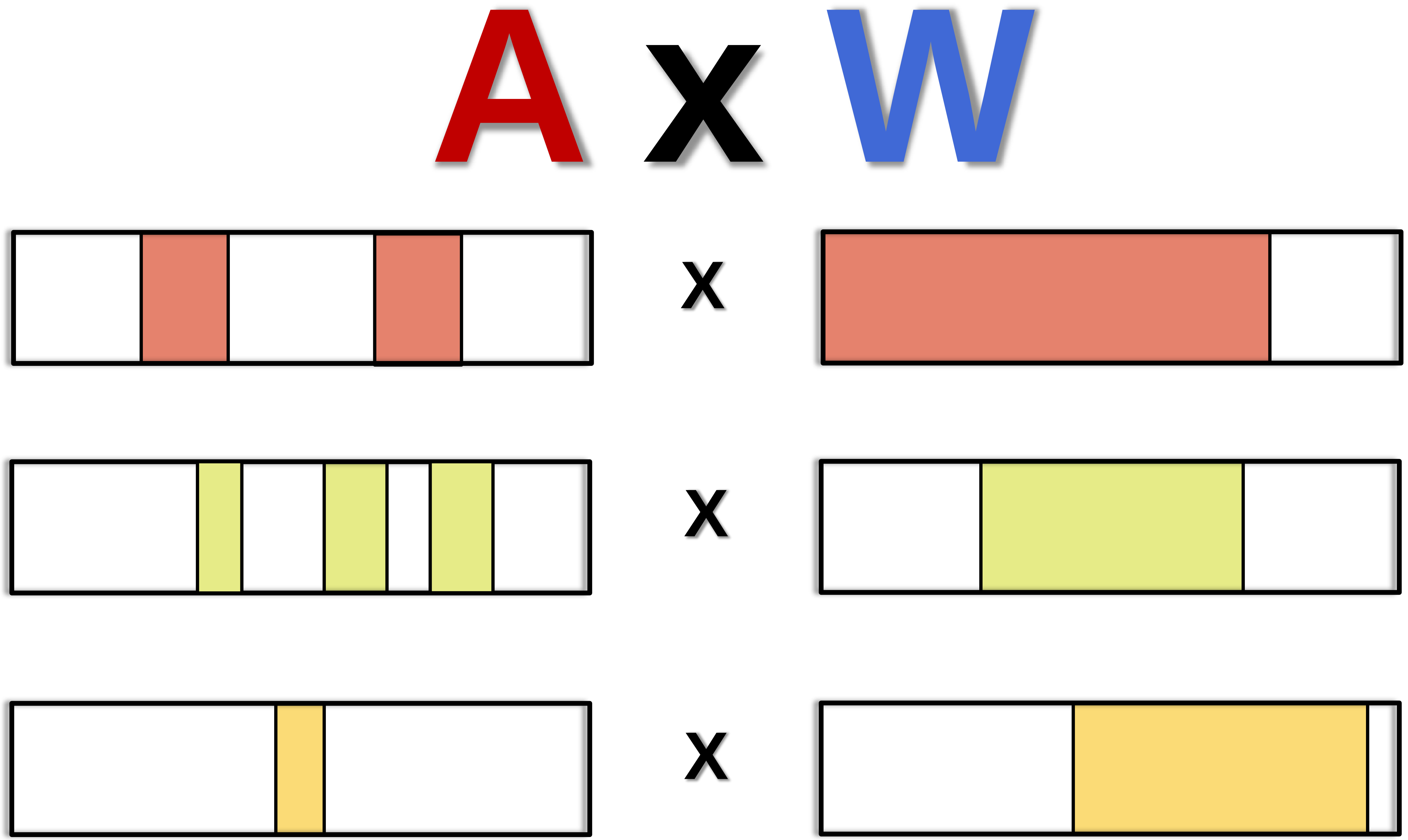
# Value Properties to Exploit? Varying Precision Needs

A x W

~60%-

- At Various Grar  arit  s: e.g.,  ay  r c  finer

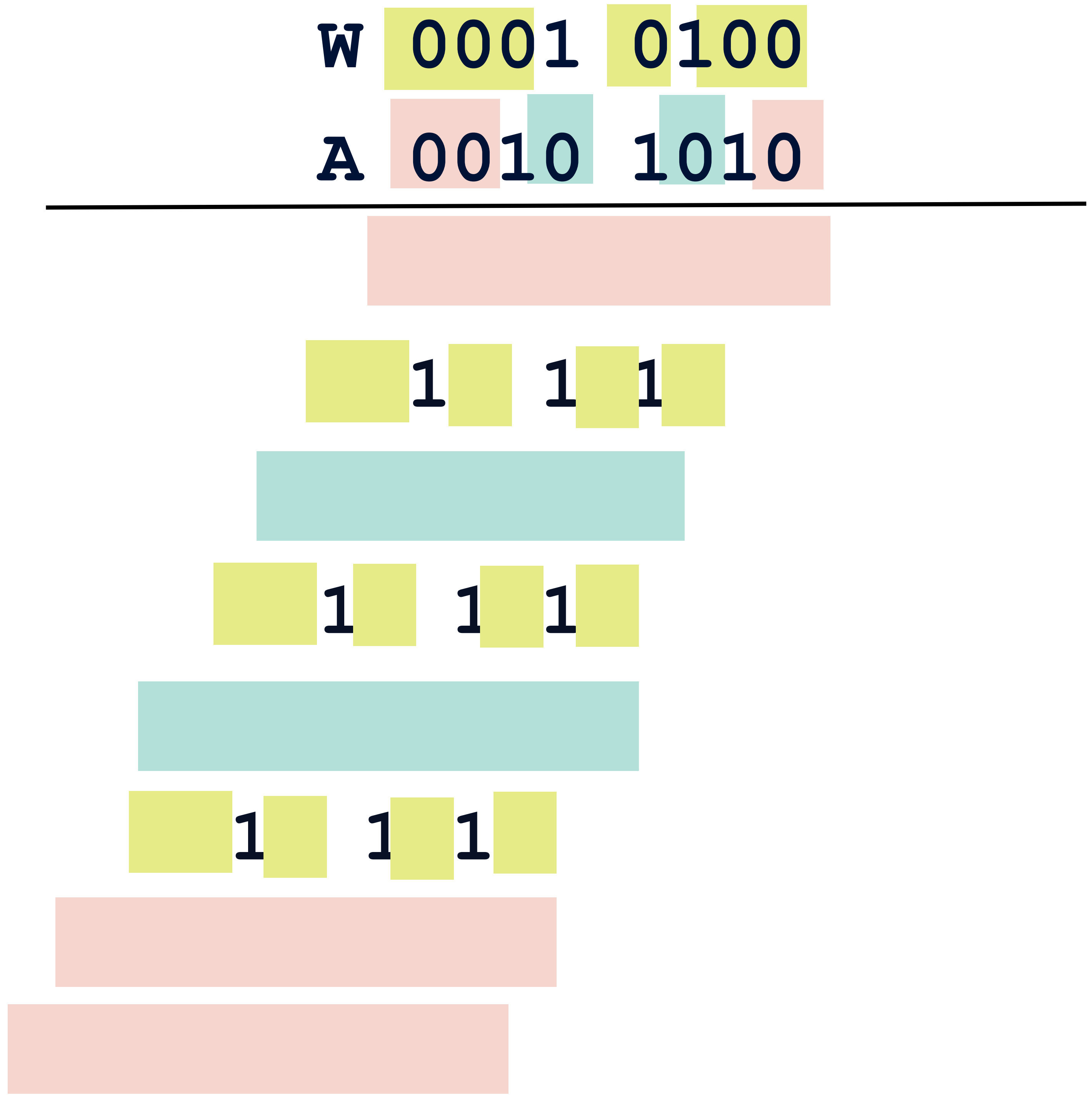
# Value Properties to Exploit? Bits that are "1"



# Ineffectual Computations Explained

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

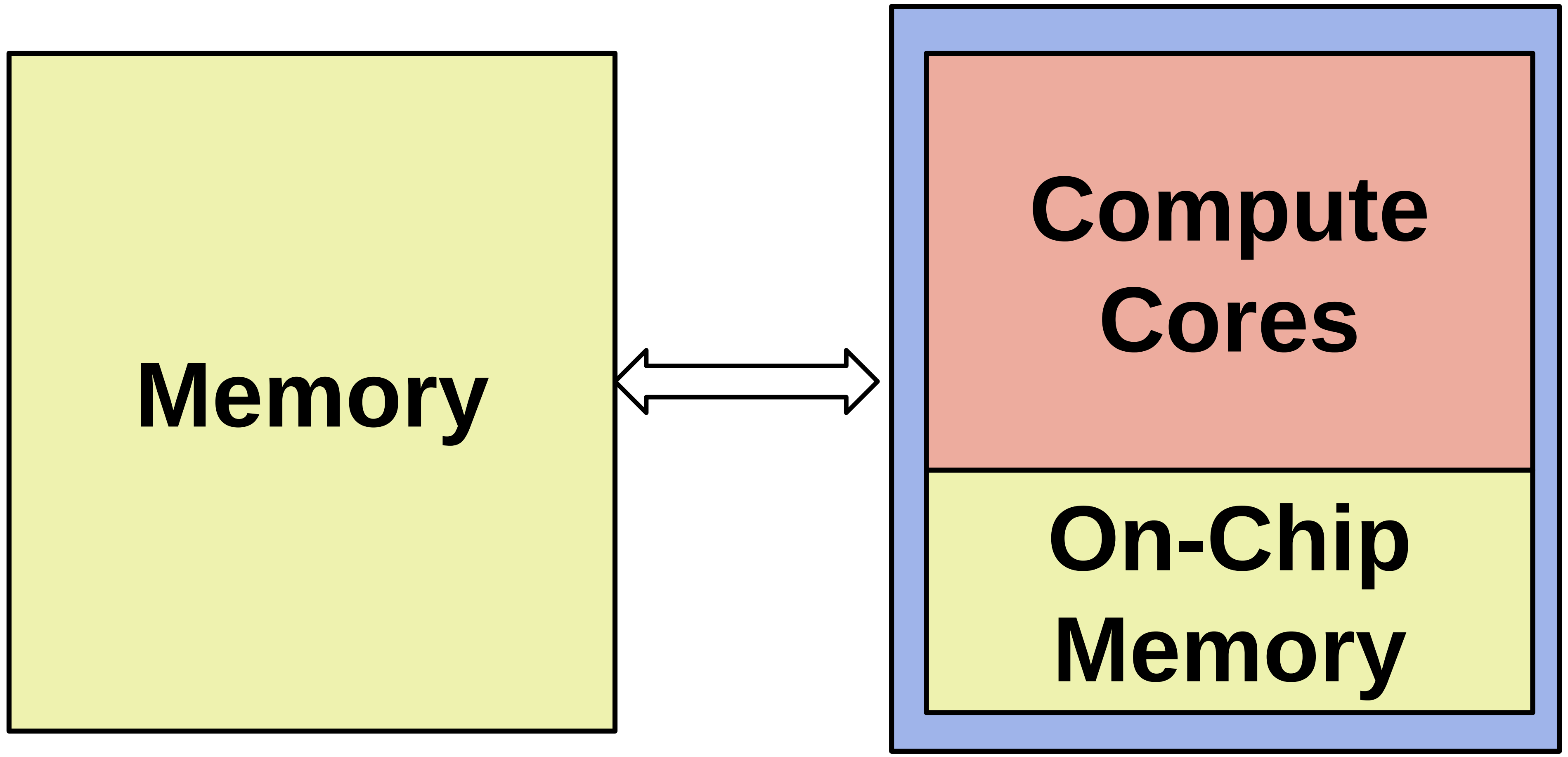
# Ineffectual Computations Explained



# Value Properties to Exploit? Bits that are "1"

**A x W**

**75%-99%**




**On- vs. Off-Chip**

Energy: ~100x

Latency: ~50x

$Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x

$Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x $W_1$ $Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x $W_1$ $Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x $W_1$ $Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x

$Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x $W_2$ $Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x $W_2$ $Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x $W_2$ $Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x

$Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x $W_3$ $Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x $W_3$ $Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x $W_3$ $Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x

$Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x $W_4$ $Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x $W_4$ $Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x $W_4$ $Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x

$Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x

$Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x $W_1$ $Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x $W_1$ $Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x $W_1$ $Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x

$Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x $W_2$ $Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x $W_2$ $Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x $W_2$ $Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x

$Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x $W_3$ $Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x $W_3$ $Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x $W_3$ $Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x

$Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x $W_4$ $Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x $W_4$ $Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x $W_4$ $Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x

$Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x

$Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x $W_1$ $Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x $W_1$ $Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x $W_1$ $Out_0$ += $A_1$ x $W_1$ $Out_1$ += $A_1$ x

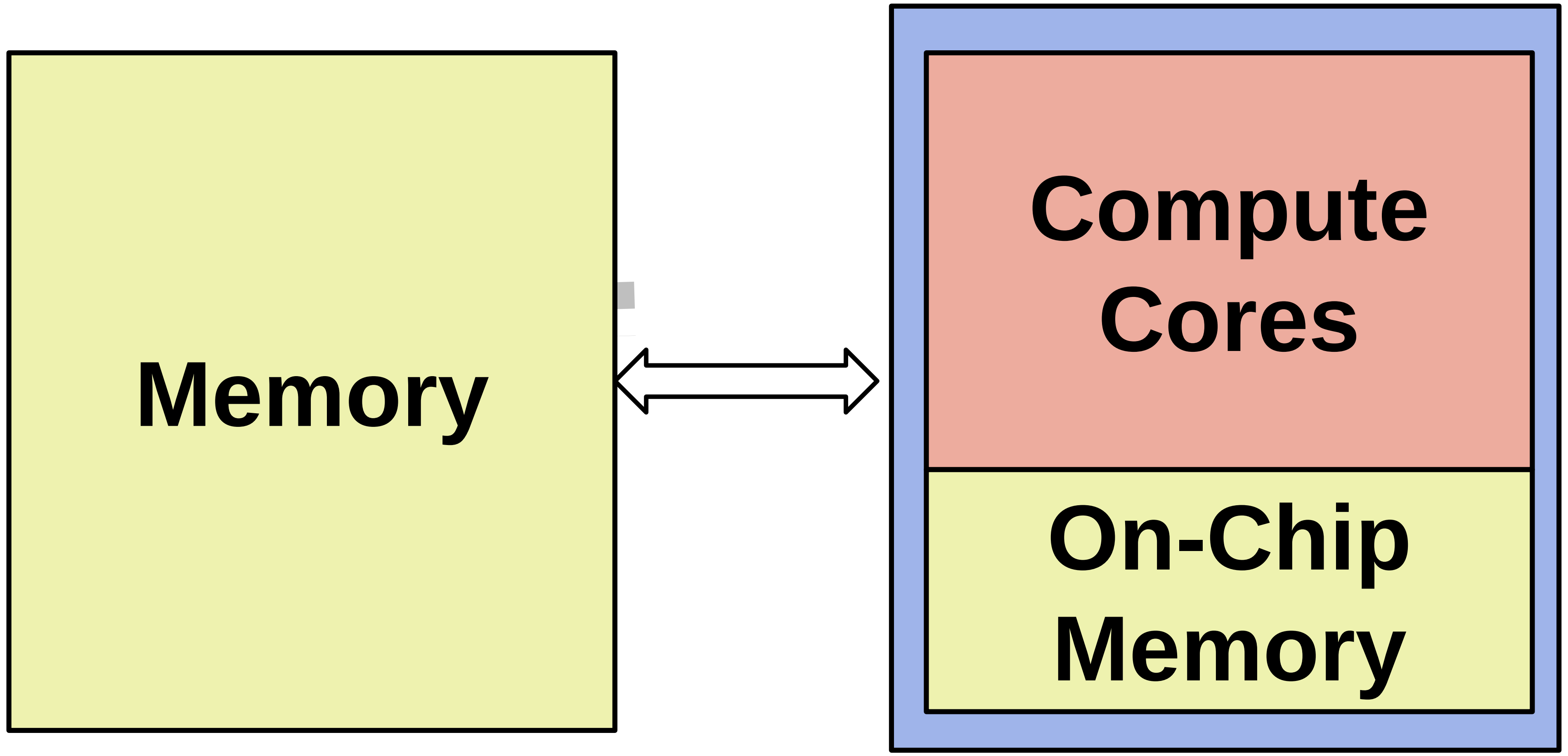
$Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x $W_2$ $Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x $W_2$ $Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x $W_2$ $Out_0$ += $A_2$ x $W_2$ $Out_1$ += $A_2$ x

$Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x $W_3$ $Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x $W_3$ $Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x $W_3$ $Out_0$ += $A_3$ x $W_3$ $Out_1$ += $A_3$ x

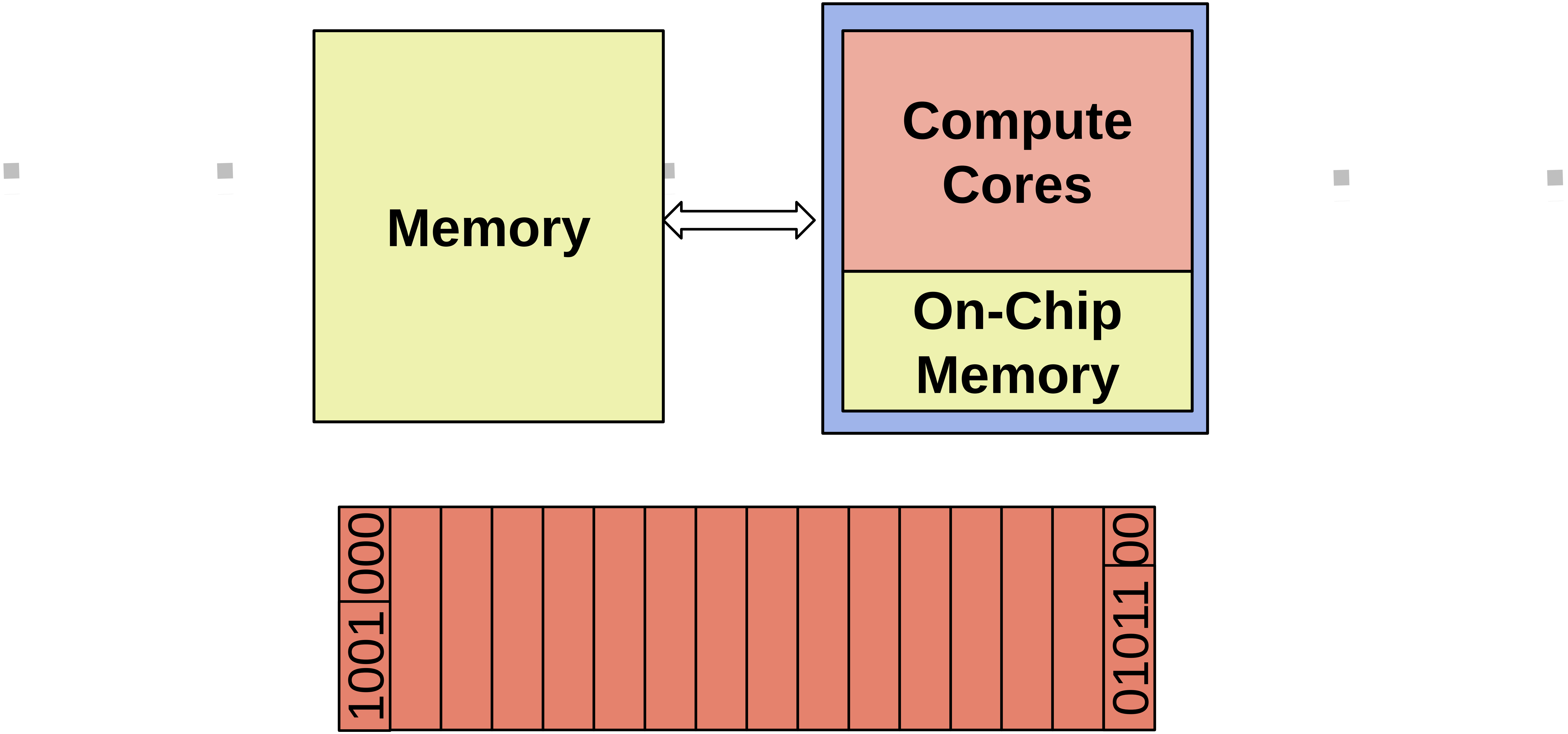
$Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x $W_4$ $Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x $W_4$ $Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x $W_4$ $Out_0$ += $A_4$ x $W_4$ $Out_1$ += $A_4$ x

$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W$

$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W$

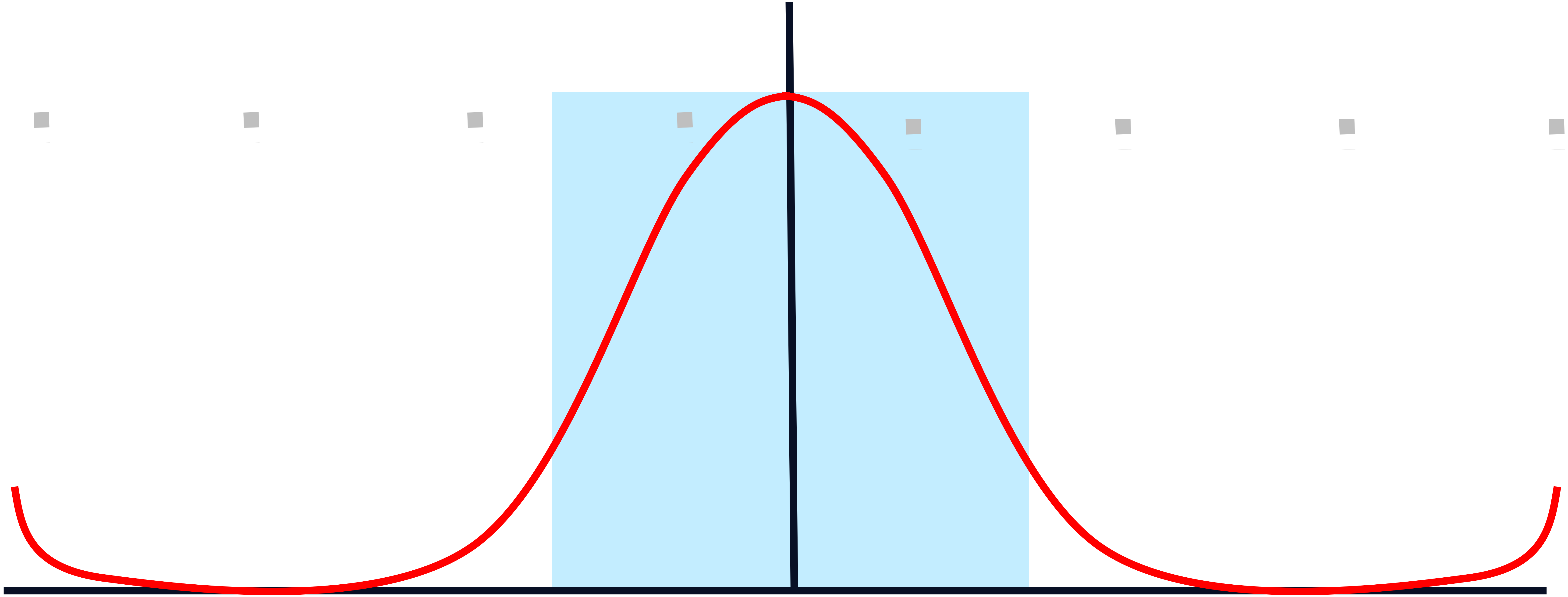


$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W$



# Making Typical Values Matter

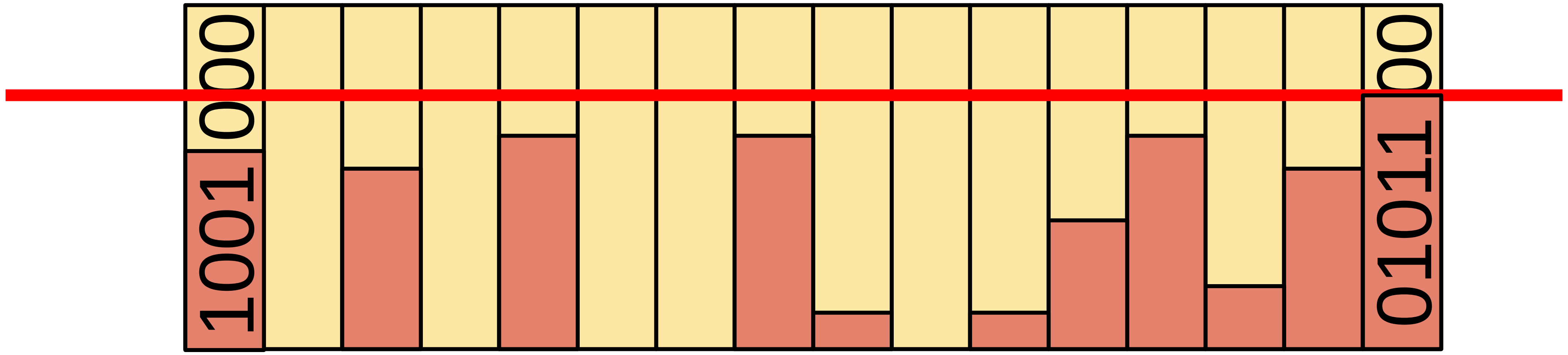
$Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x $W_0$ $Out_0$ += $A_0$ x $W_0$ $Out_1$ += $A_0$ x W



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

# Making Typical Values Matter

$Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W_0$ $Out_0 \mathrel{+}= A_0 \times W_0$ $Out_1 \mathrel{+}= A_0 \times W$




# Reduces traffic to 50% -- 25% of the original

# Effective Precisions w/ Per Group Adaptation: 16b case

Precisions are much lower

| | |
|---|---|
| GoogleNet | 6.19-5.94-5.74-6.77-6.91-6.77-6.86-6.77<br>-6.92-6.31-5.96-6.31-6.00-6.31-6.55-5.33<br>-5.33-5.33-5.33-5.33-5.48-6.74-6.33-6.74<br>-6.51-6.74-7.07-6.35-6.17-6.35-5.88-6.35<br>-6.56-5.07-4.69-5.07-4.82-5.07-5.31-5.53<br>-4.89-5.53-5.70-5.53-5.86-7.88-7.62-7.88<br>-8.07-7.88-8.31-4.97-3.85-4.97-3.61-4.97-5.36 |

**Also: Computation Speed:**

# Ineffectual Computations Explained

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

# Ineffectual Computations Explained

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

# Ineffectual Computations Explained

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

```
0010 1010
```

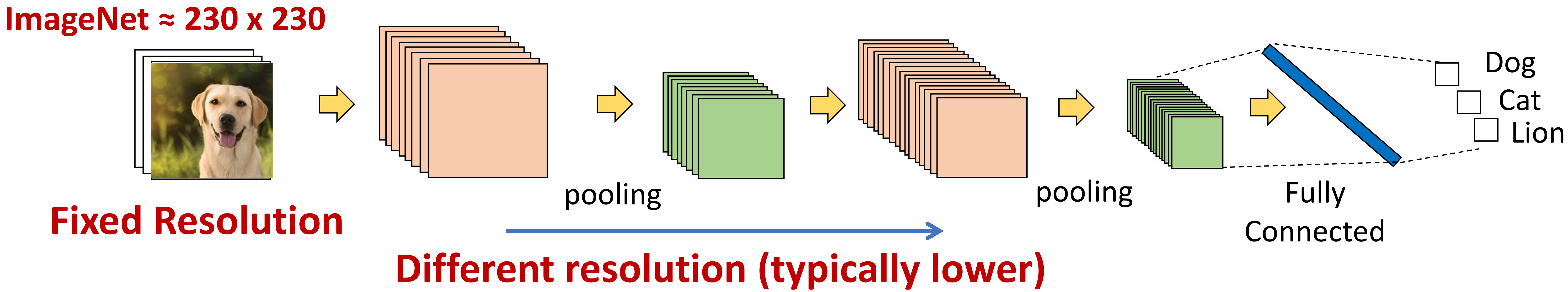
```
0010 1010
```

```
0010 1010
```

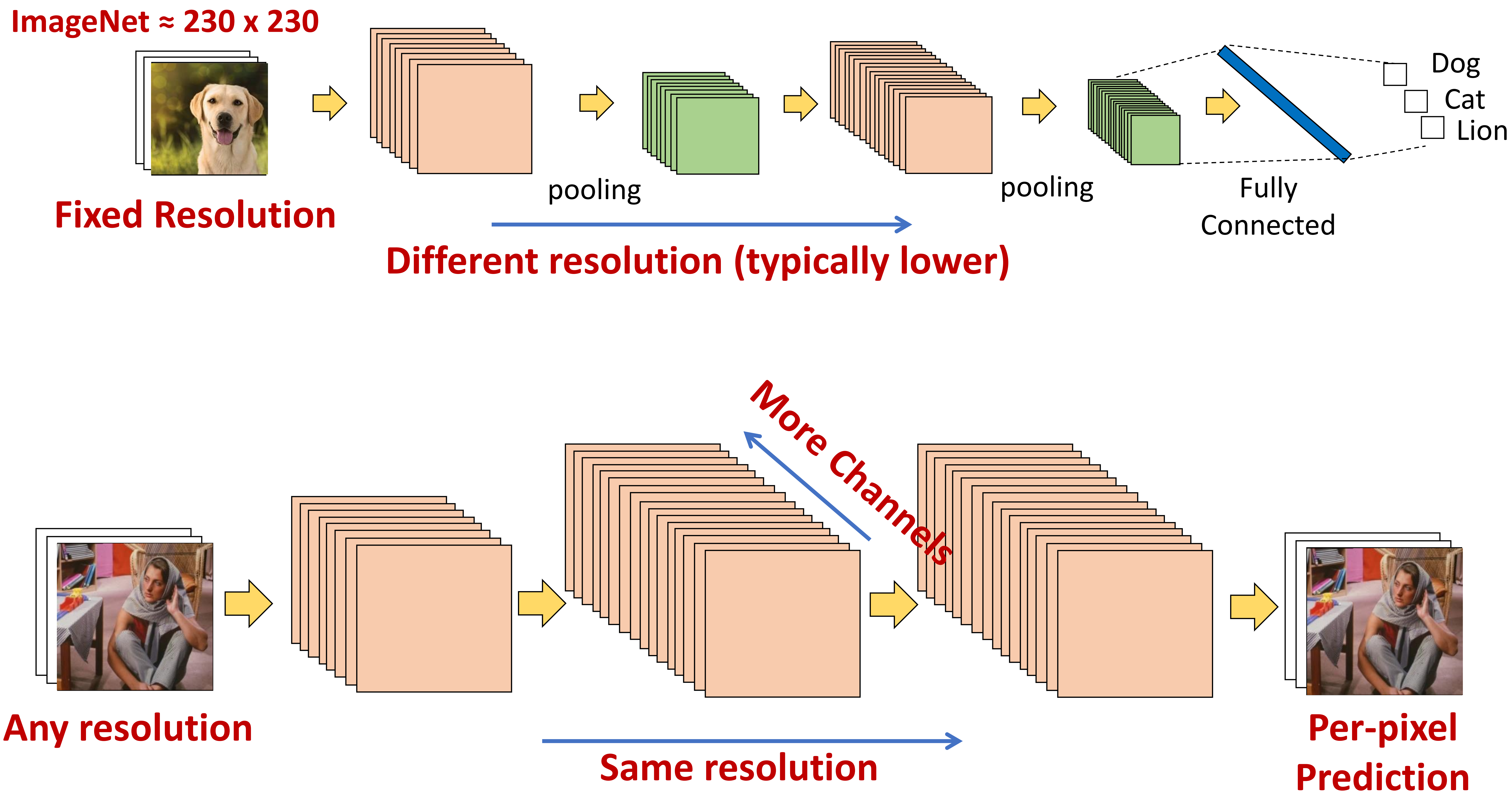
# Ineffectual Computations Explained

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

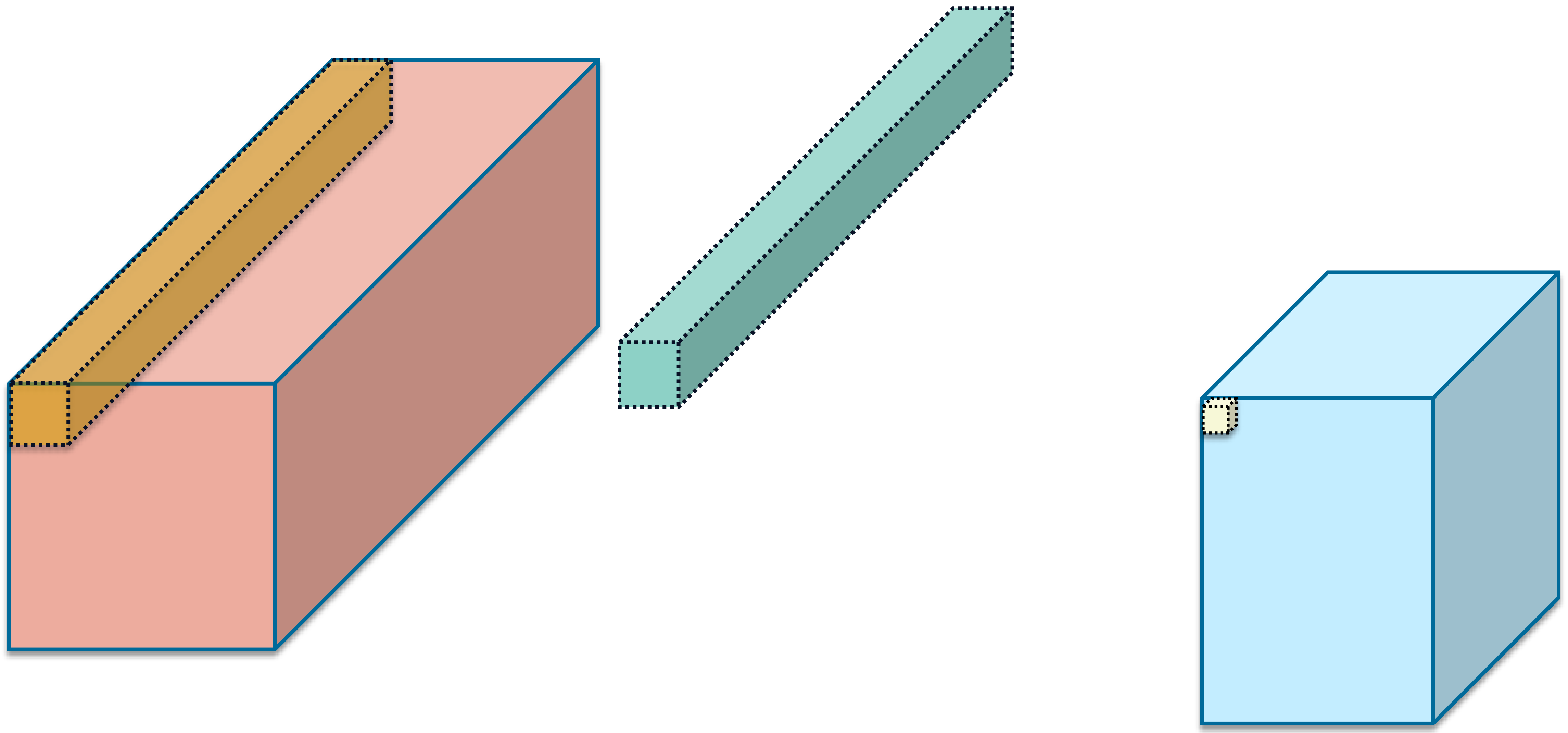
**75%-95%**



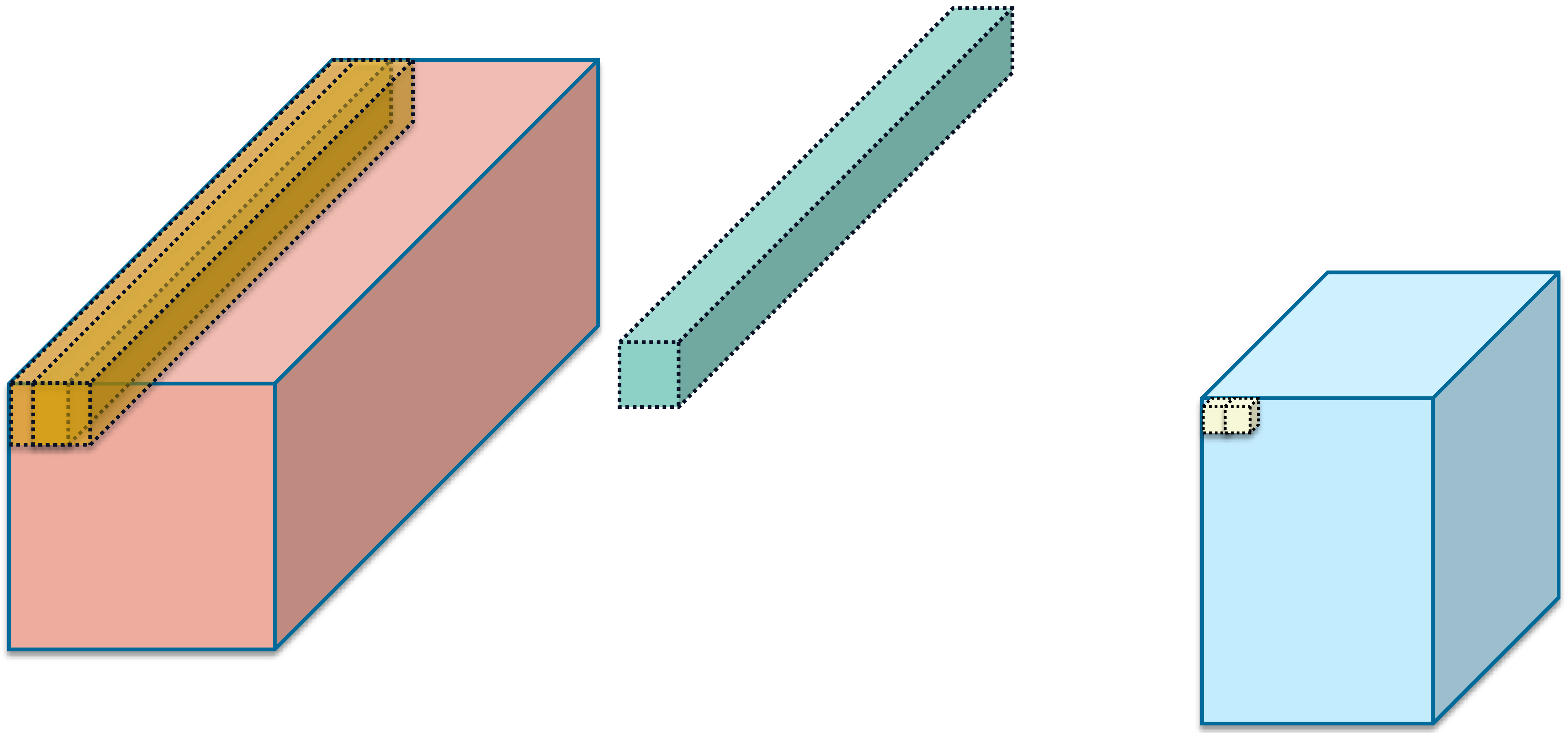
ImageNet ≈ 230 x 230
Fixed Resolution
pooling
pooling
Fully
Connected
Dog
Cat
Lion
Different resolution (typically lower)


ImageNet ≈ 230 x 230
Fixed Resolution
pooling
pooling
Fully Connected
Dog
Cat
Lion
Different resolution (typically lower)
More Channels
Any resolution
Same resolution
Per-pixel Prediction

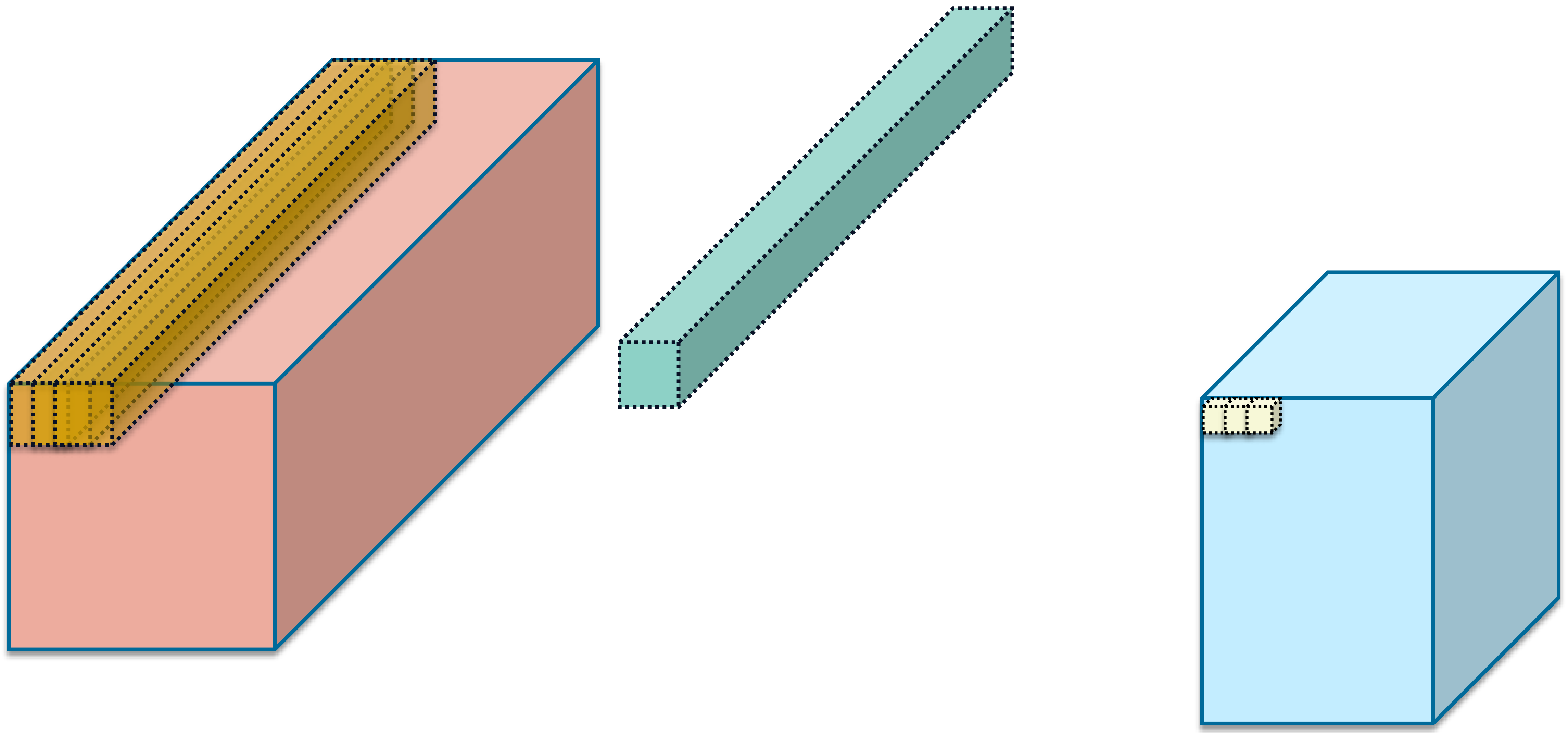
# Deep Learning: Convolutional Networks



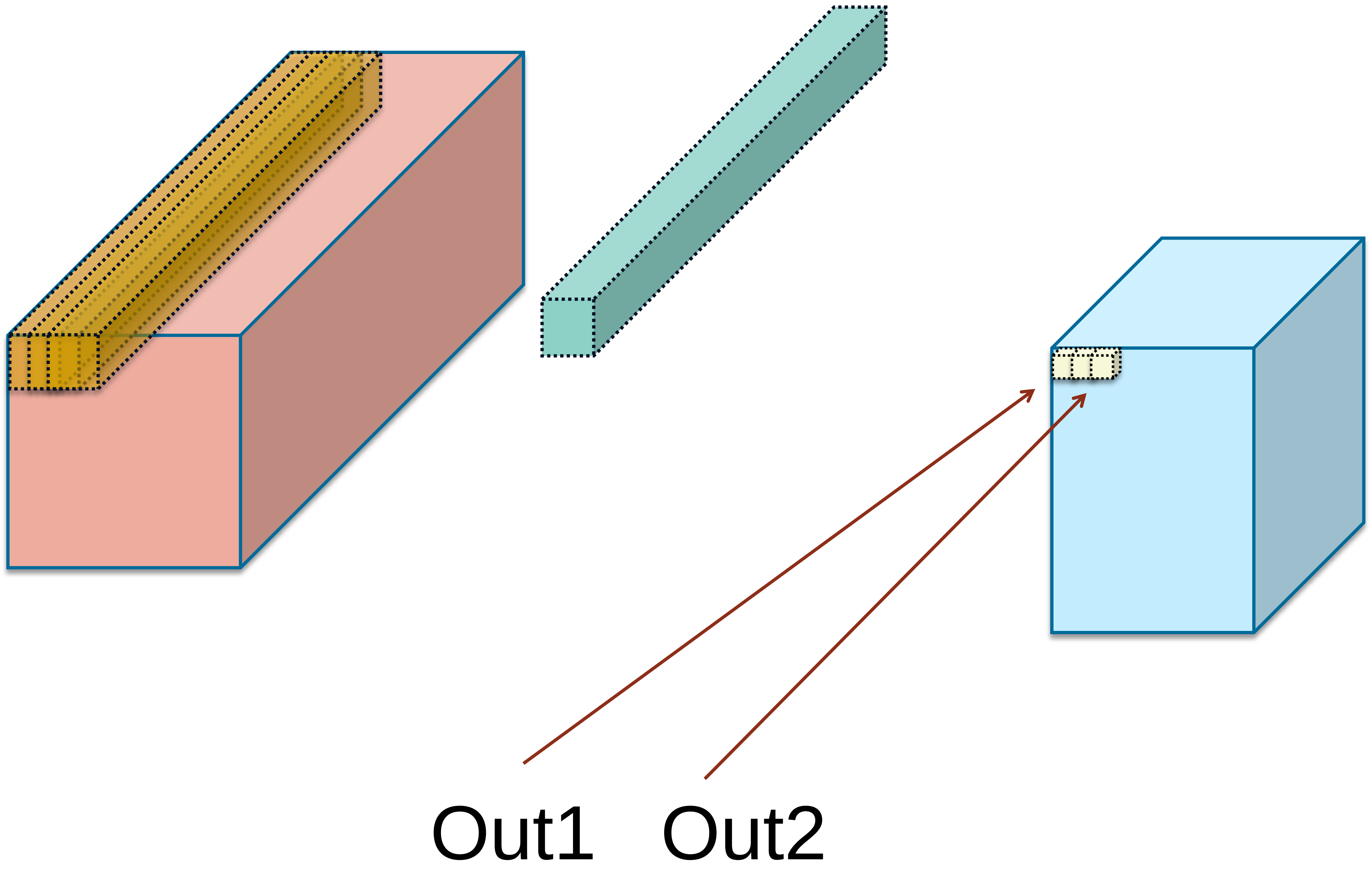
# Deep Learning: Convolutional Networks



# Deep Learning: Convolutional Networks



# Deep Learning: Convolutional Networks

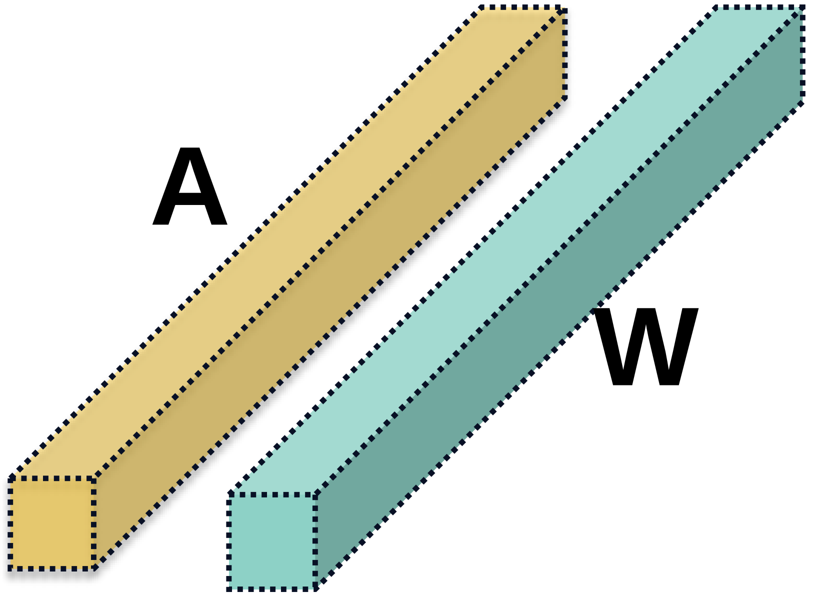


$\text{Out1} \mathrel{+}= A_0 \times W_0$

$\text{Out1} \mathrel{+}= A_1 \times W_1$

$\text{Out1} \mathrel{+}= A_2 \times W_2$

$\text{Out1} \mathrel{+}= A_3 \times W_3$




⋮

$\text{Out2} \mathrel{+}= A'_0 \times W_0$

$\text{Out2} \mathrel{+}= A'_1 \times W_1$

$\text{Out2} \mathrel{+}= A'_2 \times W_2$

$\text{Out2} \mathrel{+}= A'_3 \times W_3$

⋮

$$\text{Out1} \mathrel{+}= A_0 \times \boxed{W_0}$$

$$\text{Out2} \mathrel{+}= A'_0 \times \boxed{W_0}$$

$$\mathbf{Out1} \mathrel{+}= A_0 \times W_0 = 0\underline{11}0 \times W_0$$

Cost = 2

$$\mathbf{Out2} \mathrel{+}= A'_0 \times W_0 = 0111 \times W_0$$

**Cost = 3**

$$\text{Out1} \mathrel{+}= A_0 \times \boxed{W_0} = 0110 \times W_0$$

Cost = 2

$$\text{Out2} \mathrel{+}= A'_0 \times \boxed{W_0} = 0111 \times W_0$$

$$\text{Out2} \mathrel{+}= \text{Out1} + (A'_0 - A_0) \times W_0$$

$$\text{Out1} \mathrel{+}= A_0 \times \boxed{W_0} = 0110 \times W_0$$

Cost = 2

$$\text{Out2} \mathrel{+}= A'_0 \times \boxed{W_0} = 0111 \times W_0$$

$$\text{Out2} \mathrel{+}= \text{Out1} + (A'_0 - A_0) \times W_0$$

$$\text{Out2} \mathrel{+}= \text{Out1} + (0111 - 0110) \times W_0$$

$$\text{Out1} \mathrel{+}= A_0 \times \boxed{W_0} = 0110 \times W_0$$

Cost = 2

$$\text{Out2} \mathrel{+}= A'_0 \times \boxed{W_0} = 0111 \times W_0$$

$$\text{Out2} \mathrel{+}= \text{Out1} + (A'_0 - A_0) \times W_0$$

$$\text{Out2} \mathrel{+}= \text{Out1} + (0111 - 0110) \times W_0$$

$$\text{Out2} \mathrel{+}= \text{Out1} + 0001 \times W_0$$

Cost = 1

# Pruning and Sparsity

**Do as you are told?**

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

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

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

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

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

$$\vdots$$

## Do as you are told?

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

$$Out \mathrel{+}= A_1 \times 0$$

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

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

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

⋮

**Do as you are told?**

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

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

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

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

$$\vdots$$

**Do as you are told?**

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

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

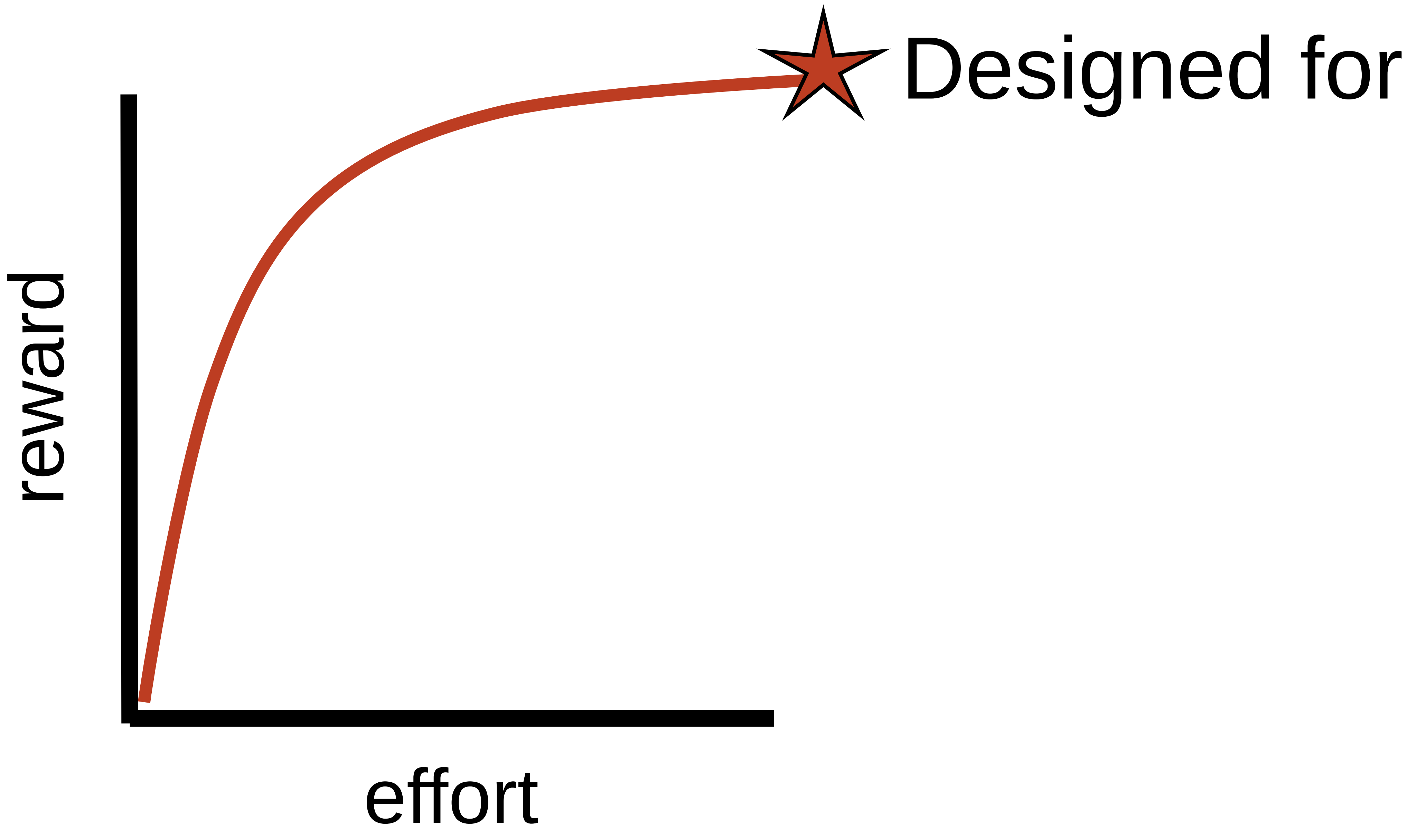
$$Out \mathrel{+}= A_3 \times W_3$$

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

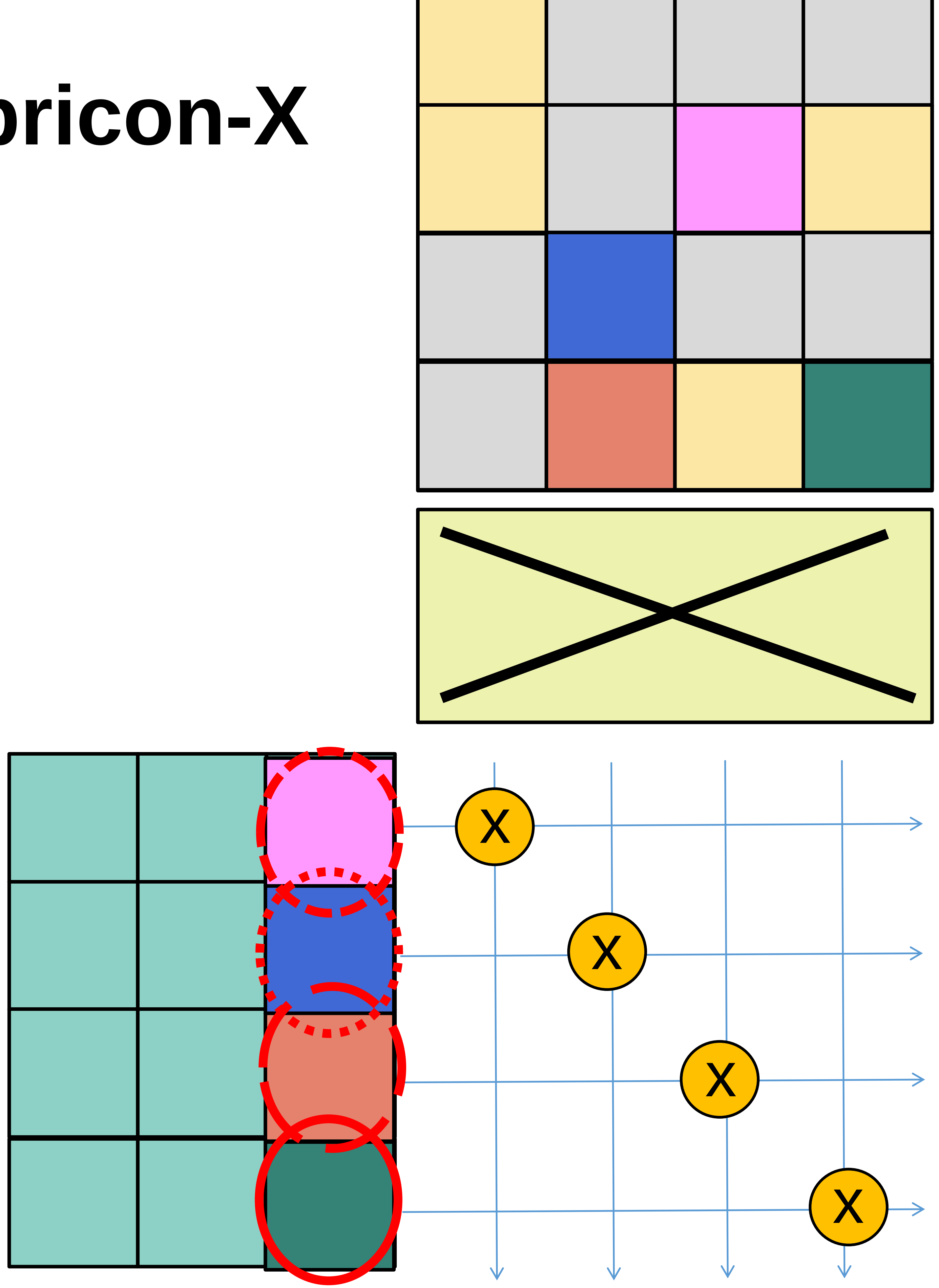
$$\vdots$$

# Removing Zero Weights

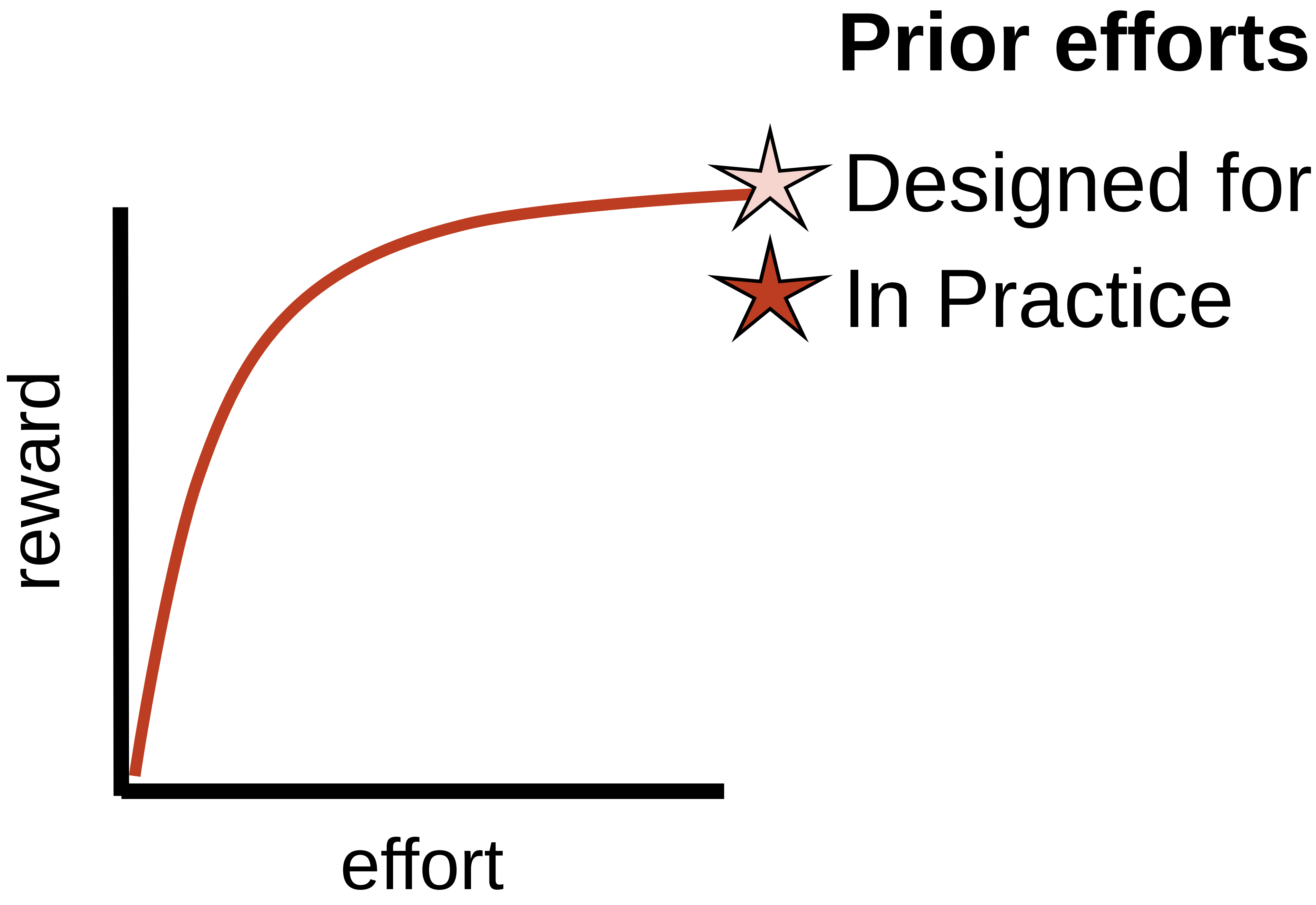
**Prior efforts**



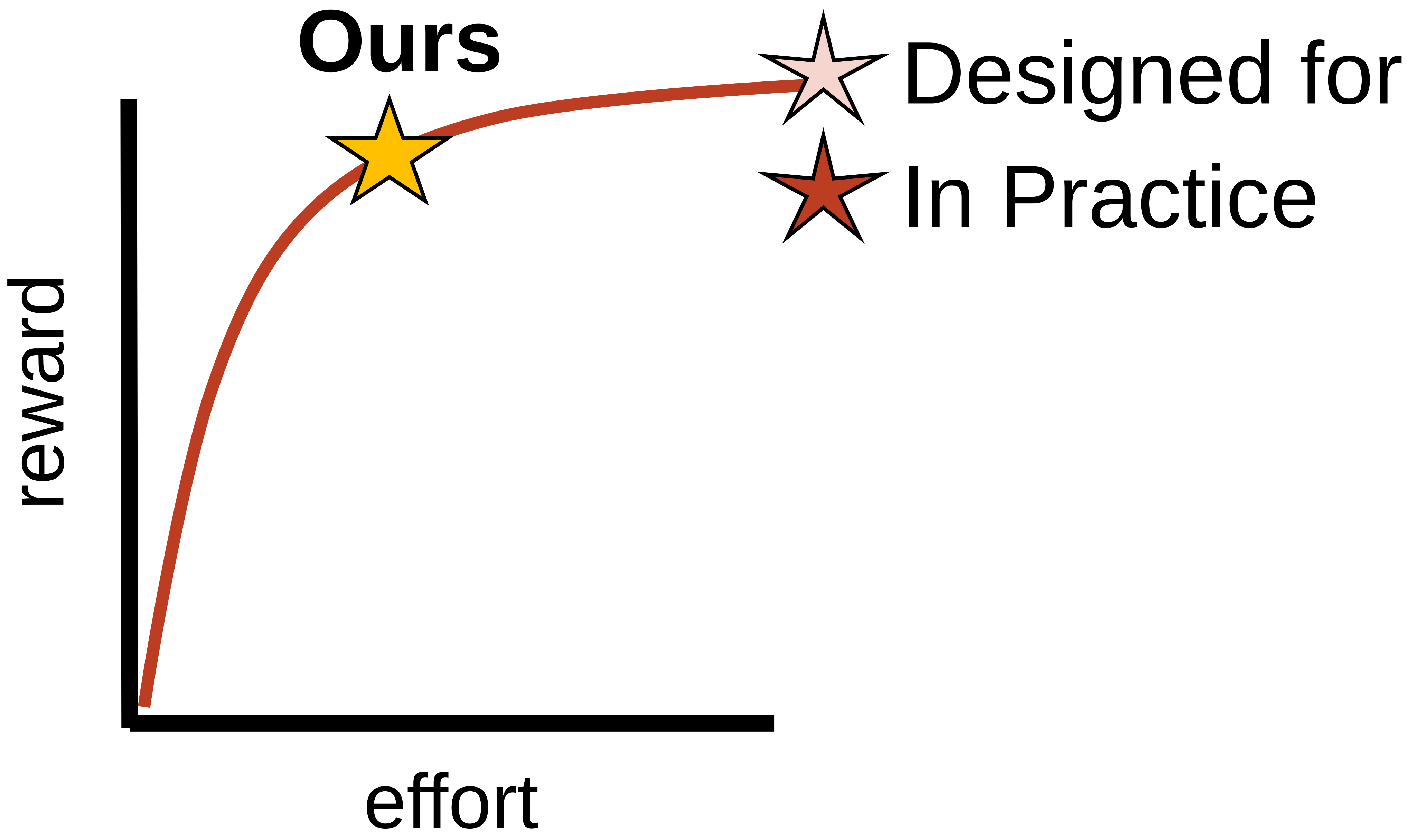
# Cambricon-X



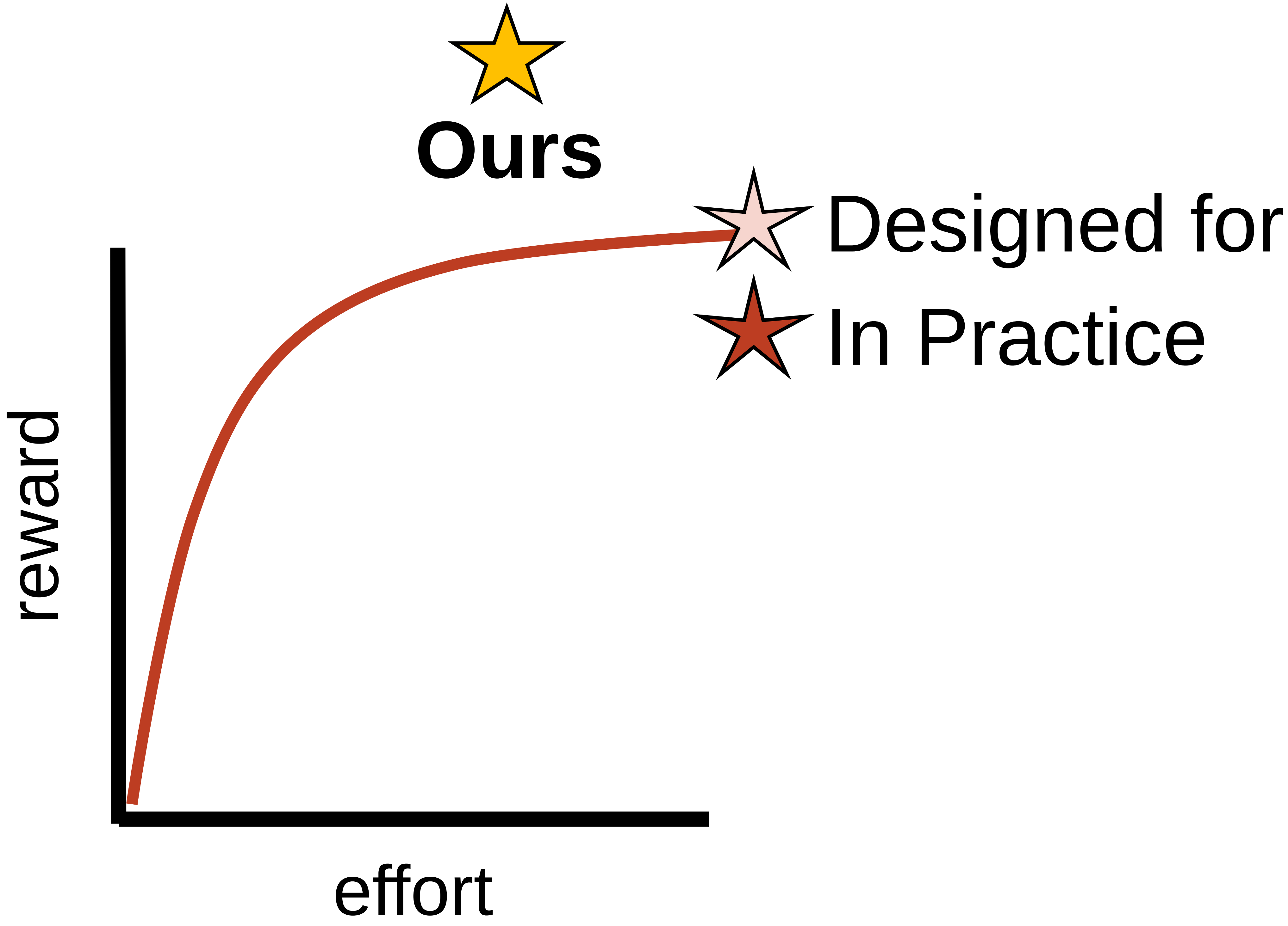
# Removing Zero Weights



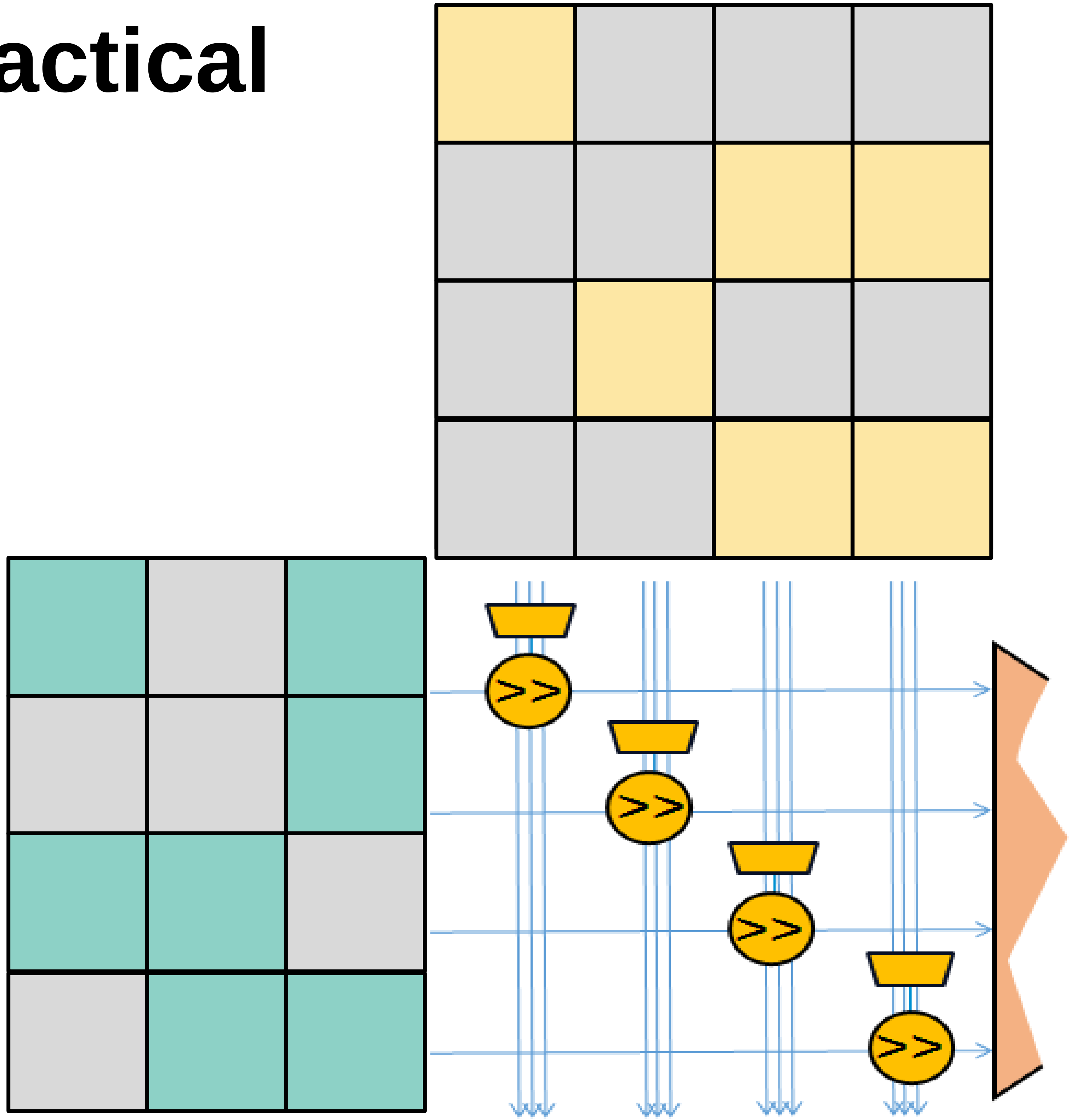
# Removing Zero Weights



# Removing Zero Weights



# Bit-Tactical



# Computing Hardware

- We build tools
- Used by "everyone" for "everything"
- Science, medicine, commerce, …

**Provide Immediate Benefits**

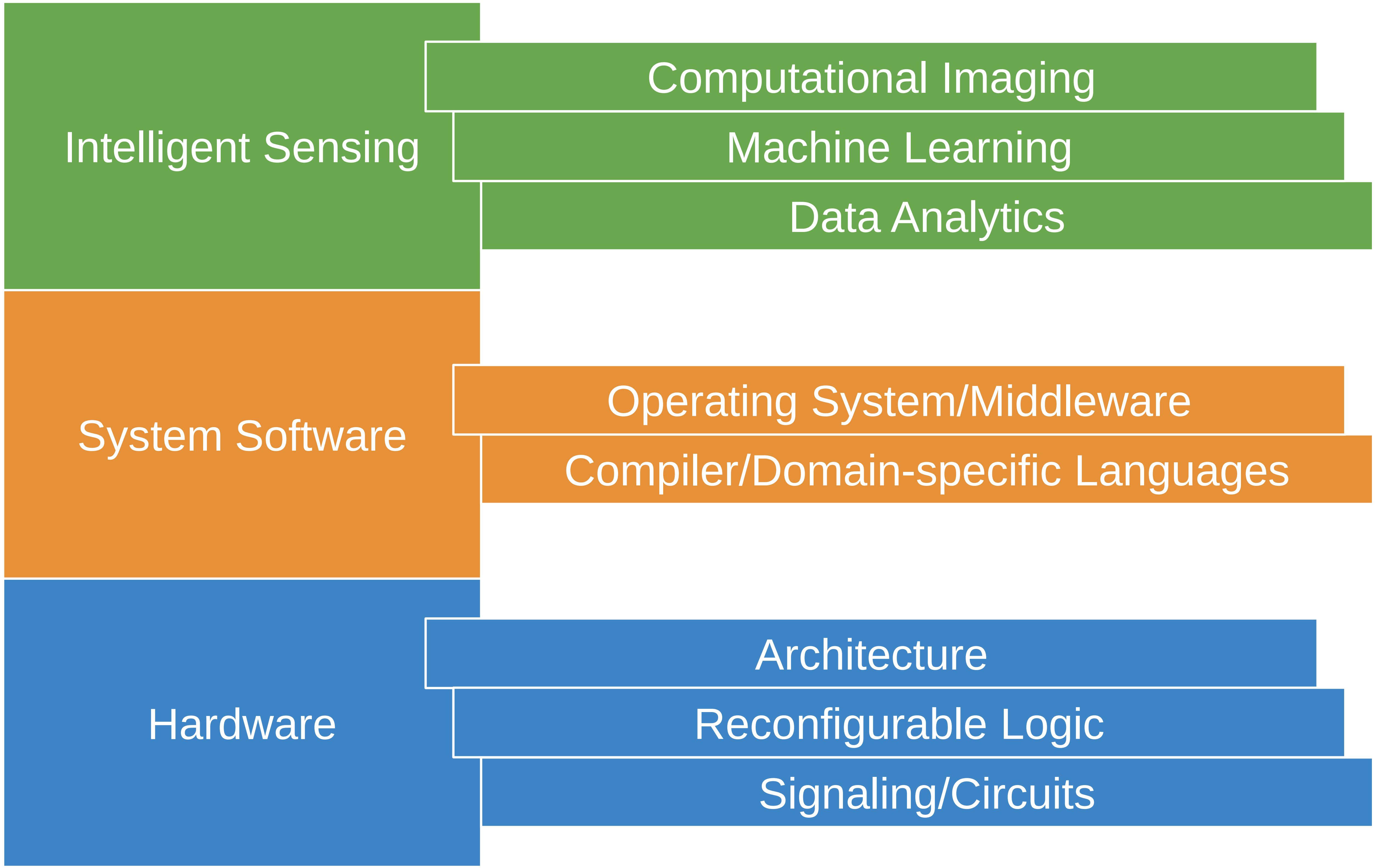
**Reward but do not Require Model Optimization**

**Path to Innovation**


UNIVERSITY OF TORONTO
AMD
ALTERA
intel
QUALCOMM
POLY MTL
Université de Montréal
McGill
UNIVERSITY OF ALBERTA
UBC
SFU
University of Victoria
CMC MICROSYSTEMS
HUAWEI
IIROC OCRCVM
Google
Mitacs

COHESA

# NSERC COHESA



Intelligent Sensing

System Software

Hardware

Sanja Fidler

Chris Pal

Raquel Urtasun

Yoshua Bengio