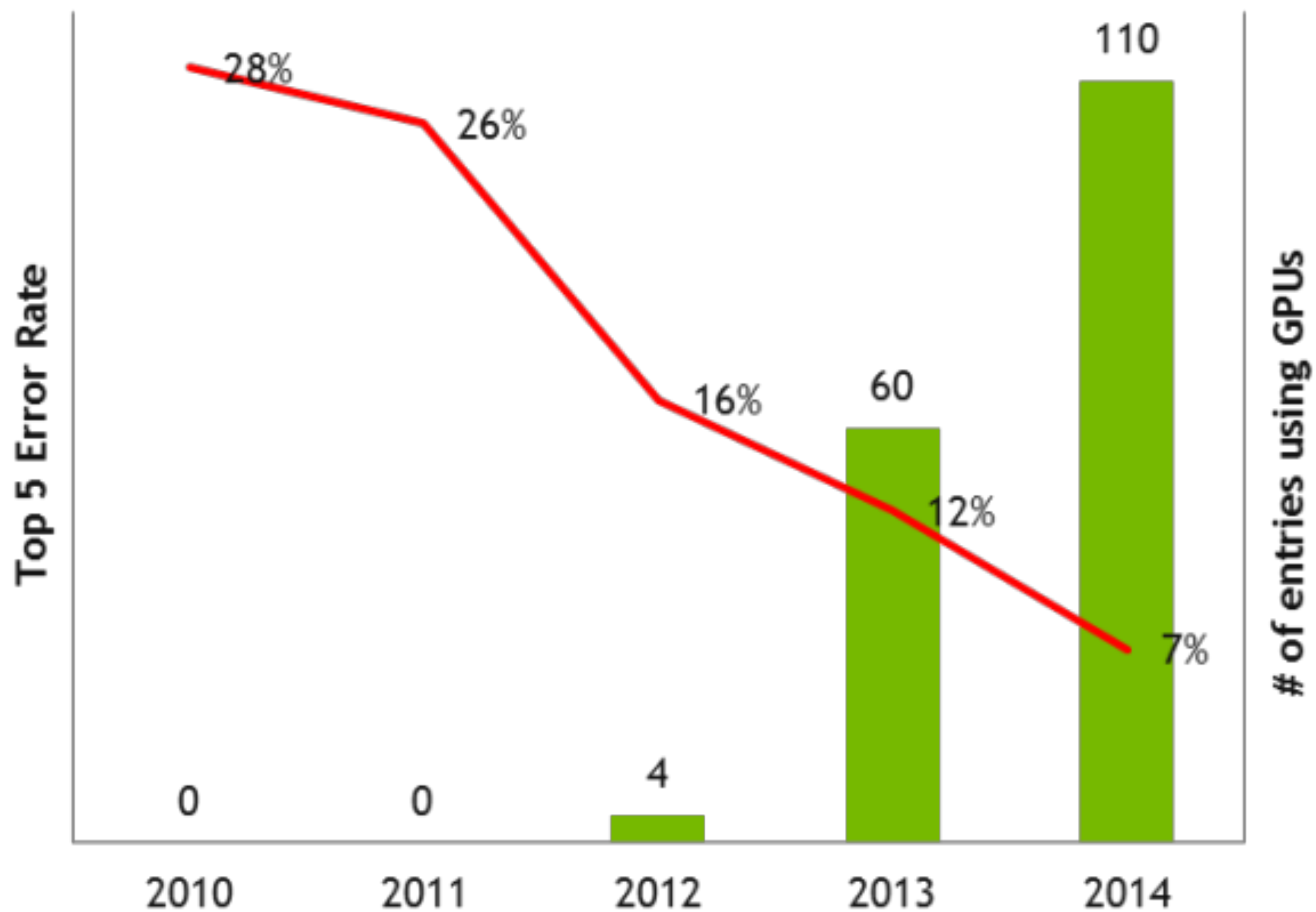
Accelerators (I)

Joel Emer

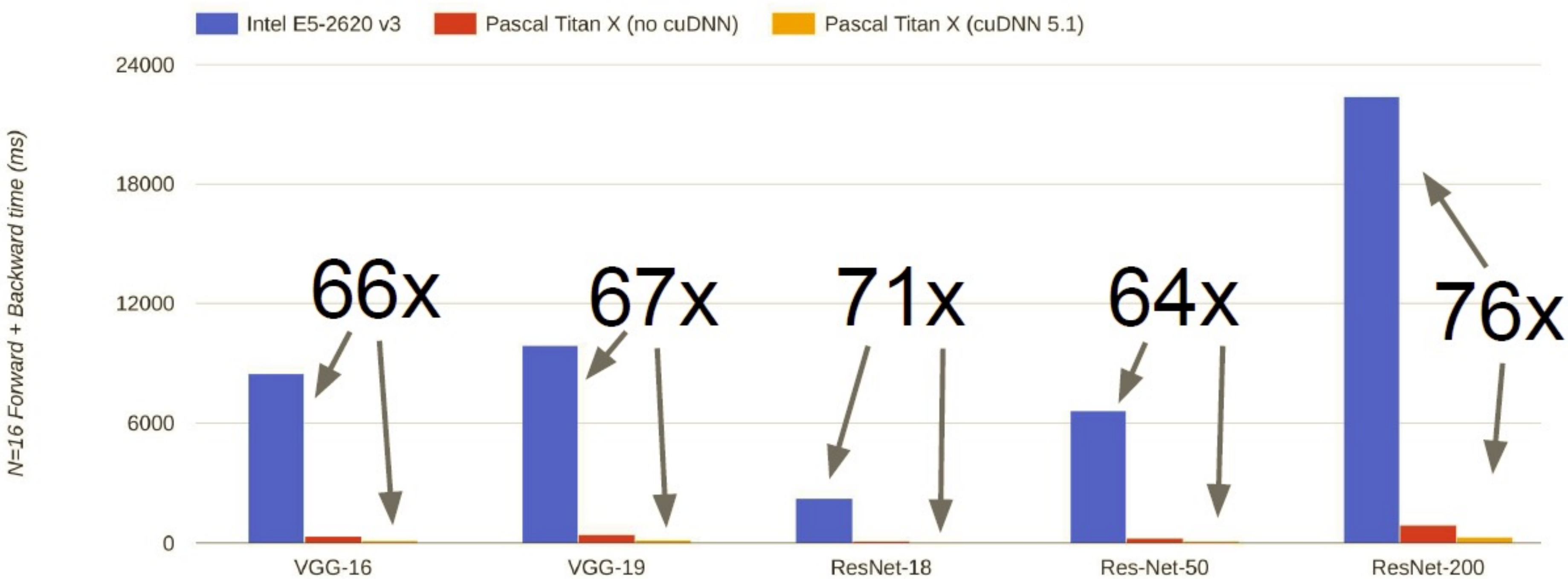
Massachusetts Institute of Technology
Electrical Engineering & Computer Science

"Compute has been the oxygen of deep learning"
– Ilya Sutskever (Open AI)

GPU Usage for ImageNet Challenge



CPU vs. GPU Performance



Data from https://github.com/jcjohnson/cnn-benchmarks

Ratio of (partially-optimized) CPU vs. CUDA library (cuDNN)

Source: Stanford CS231n

Opportunities

From EE Times – September 27, 2016

"Today the job of training machine learning models is limited by compute, if we had faster processors we'd run bigger models…in practice we train on a reasonable subset of data that can finish in a matter of months. We could use improvements of several orders of magnitude – 100x or greater."

– Greg Diamos, Senior Researcher, SVAIL, Baidu

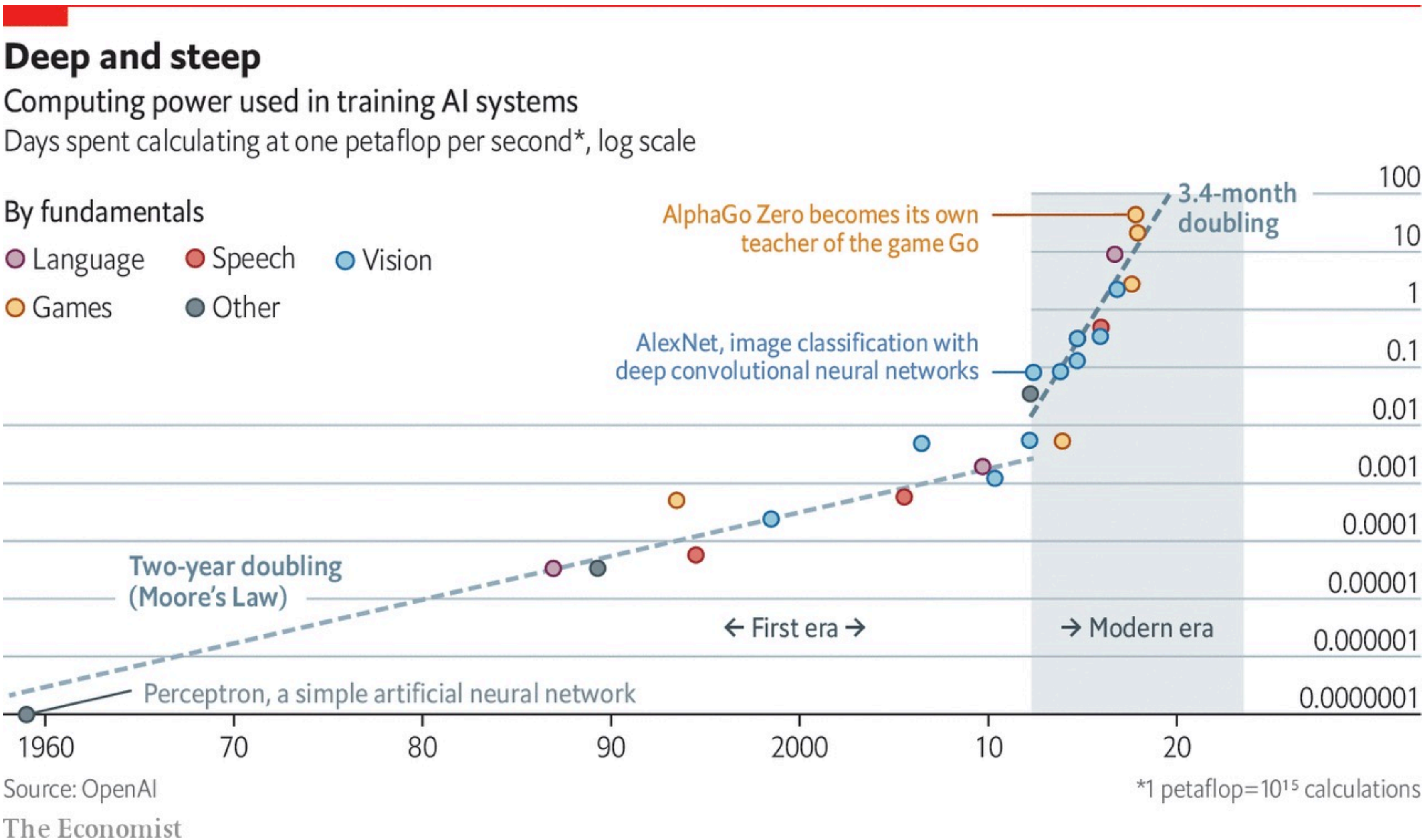
ACM's Celebration of 50 Years of the ACM Turing Award (June 2017)

"Compute has been the oxygen of deep learning"

– Ilya Sutskever, Research Director of Open AI

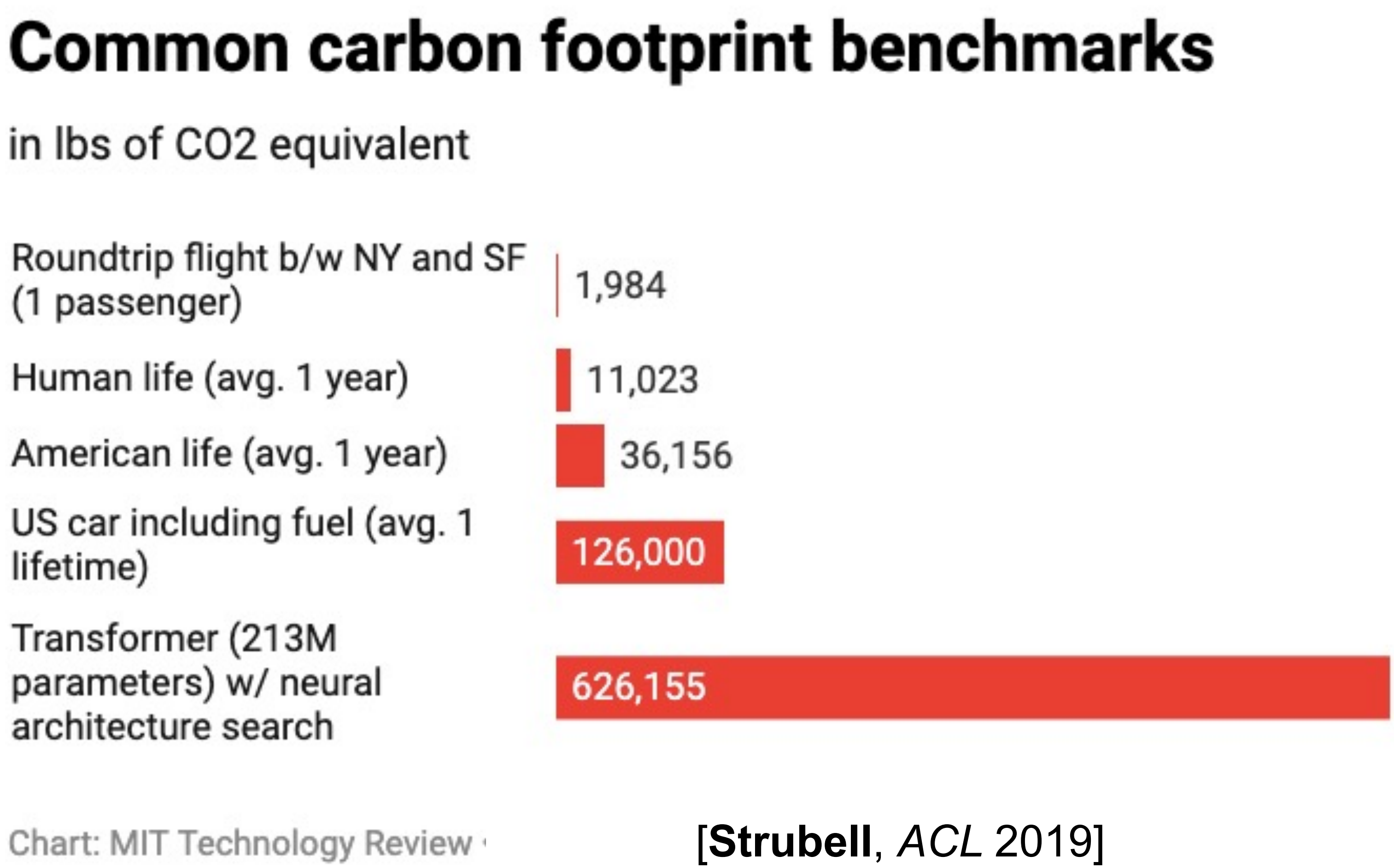
Compute Demands Growing Exponentially

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



Source: https://www.economist.com/technology-quarterly/2020/06/11/the-cost-of-training-machines-is-becoming-a-problem

Compute Demands for Deep Neural Networks



[**Strubell**, *ACL* 2019]

Compute Challenges for Self-Driving Cars

JACK STEWART TRANSPORTATION 02.06.18 08:00 AM

SELF-DRIVING CARS USE CRAZY AMOUNTS OF POWER, AND IT'S BECOMING A PROBLEM

Shelley, a self-driving Audi TT developed by Stanford University, uses the brains in the trunk to speed around a racetrack autonomously.

NIKKI KAHN/THE WASHINGTON POST/GETTY IMAGES

WIRED

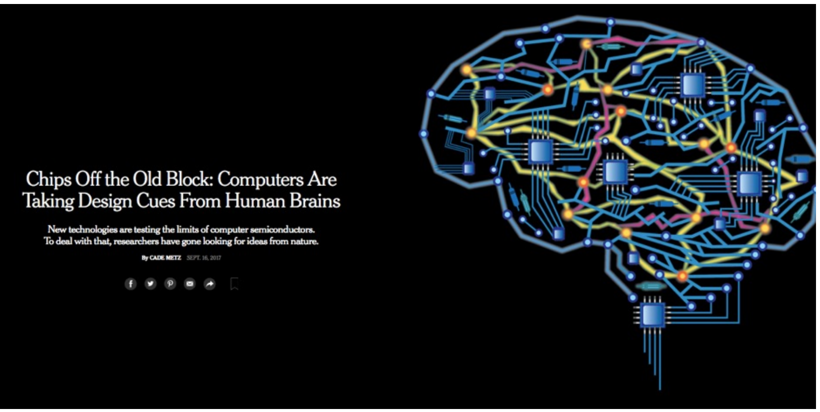
(Feb 2018)

Cameras and radar generate ~6 gigabytes of data every 30 seconds.

Prototypes use around 2,500 Watts. Generates wasted heat and some prototypes need water-cooling!

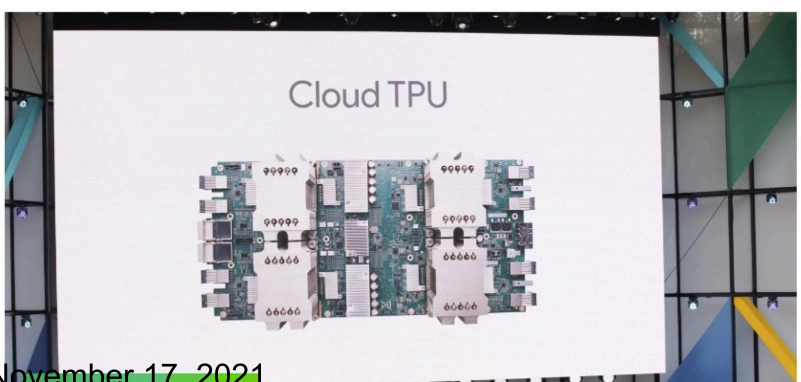
Software Companies are Building HW

The New York Times



Google's custom TPU machine learning accelerators are now available in beta

Frederic Lardinois @fredericl / Feb 12, 2018 Comment



Chips Off the Old Block: Computers Are Taking Design Cues From Human Brains *(September 16, 2017)*

After training a speech-recognition algorithm, for example, Microsoft offers it up as an online service, and it actually starts identifying commands that people speak into their smartphones. **G.P.U.s are not quite as efficient during this stage of the process. So, many companies are now building chips specifically to do what the other chips have learned.**

Google built its own specialty chip, a Tensor Processing Unit, or T.P.U. Nvidia is building a similar chip. **And Microsoft has reprogrammed specialized chips from Altera**, which was acquired by Intel, so that it too can run neural networks more easily.

HW Beyond Cloud Computing

WIRED

Musk Says Tesla Is Building Its Own Chip for Autopilot

TOM SIMONITE BUSINESS 12.08.17 01:09 PM

MUSK SAYS TESLA IS BUILDING ITS OWN CHIP FOR AUTOPILOT

Elon Musk disclosed plans for Tesla to design its own chip to power its self-driving function.

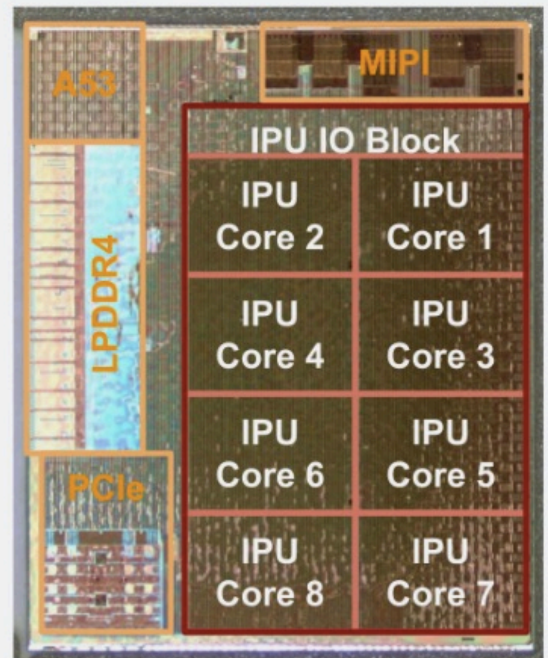
ars TECHNICA — BIZ & IT | TECH | SCIENCE | POLICY | CARS | GAMING & CULTURE

TWO SOCS IS BETTER THAN ONE —

Surprise! The Pixel 2 is hiding a custom Google SoC for image processing

Google's 8-core Image Processing Unit will be enabled with Android 8.1.

RON AMADEO - 10/17/2017, 9:00 AM



Enlarge / Google's Pixel Visual Core, an SoC designed for image processing and machine learning.

[Also Nvidia, Intel, Qualcomm…]

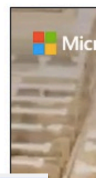
Growing Demand for HW Designers

Bloomberg Technology | Markets | Tech | Pursuits | Politics | Opinion | Businessweek

Facebook Is Forming a Team to Design Its Own Chips

By **Mark Gurman**, **Ian King**, and **Sarah Frier**
April 18, 2018, 3:49 PM EDT

- Social network could use semiconductors for consumer devices
- Move follows Apple's chip efforts, early work by Google



facebook Sign Up

Email or Ph

Work at Facebook | Teams | Locations | University Students | Benefits | Facebook Life | C

Infrastructure

ASIC & FPGA Design Engineer

(Menlo Park, CA)

Facebook's mission is to give people the power to build community and bring the world closer together. Through our family of apps and services, we're building a different kind of company that connects billions of people around the world, gives them ways to share what matters most to them, and helps bring people closer together. Whether we're creating new products or helping a small business expand its reach, people at Facebook are builders at heart. Our global teams are constantly iterating, solving problems, and working together to empower people around the world to build community and connect in meaningful ways. Together, we [illegible]ger communities — we're just getting started.

High-level Synthesis Design Engineer, Consumer Hardware
Google
Mountain View, CA, US
In this role, you will use your software engineering expertise to help solve complex problems, design and optimize algorithms (for example in the domains of machine learning, ... careers.google.com
44 connections work here
5 days ago

ASIC Design Verification Engineer, Consumer Hardware
Google
Mountain View, CA, US
Experience verifying digital logic at the Register Transfer Level (RTL) using SystemVerilog for FPGAs, ASICs, and/or SoCs. Experience with image processing, computer vision, and... careers.google.com
MIT 373 company alumni work here
5 days ago

Global ASIC/SoC Sourcing Manager, Consumer Hardware
Google
Mountain View, CA, US
7 years of experience of ASIC and/or SoC sourcing Management or supply chain management experience in commercial sourcing roles with particular experience in silicon and ... careers.google.com
44 connections work here
5 days ago

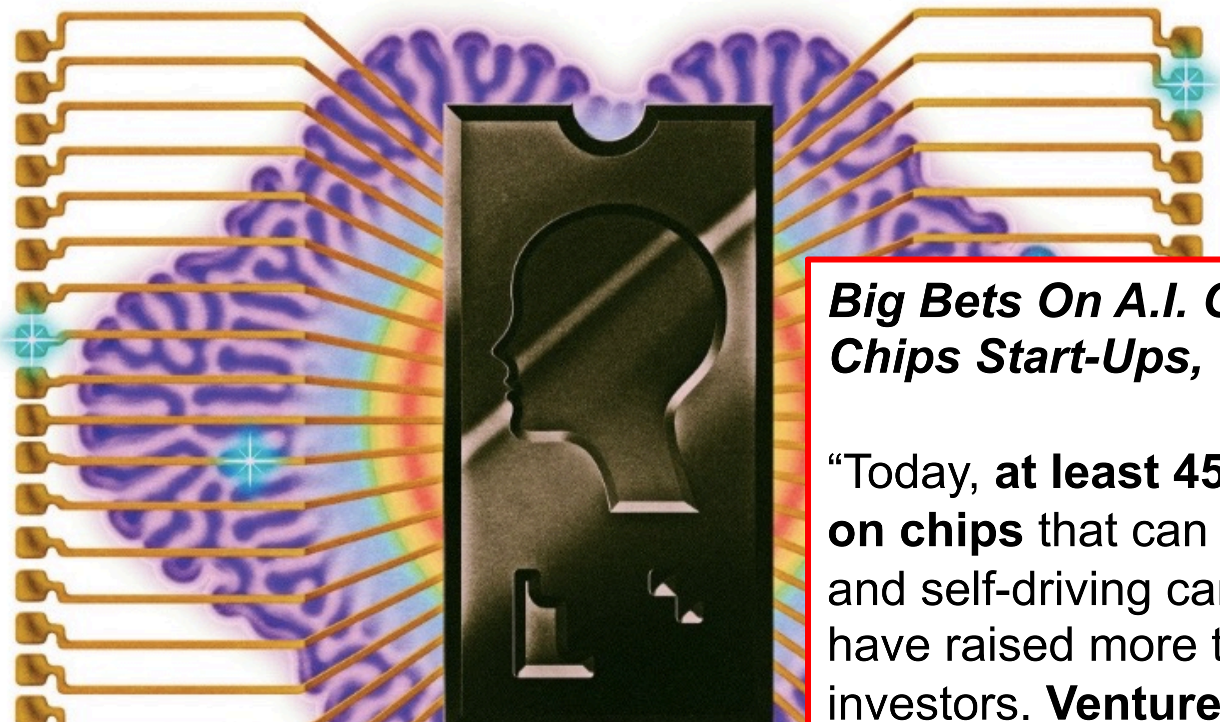
HW Development Manager, FPGA and ASIC IP design - CSI / Azure - Cloud Server Infrastructure
Microsoft
Bellevue, WA, US
Microsoft is seeking a highly motivated, FPGA and ASIC IP design engineering manager to build innovative FPGA-based computing systems within a large design team. The group will ... careers.microsoft.com
13 connections work here
1 month ago

Physical Design Engineer
Microsoft
Redmond, WA, US
1-2 years of experience in ASIC physical design flows and methodologies. Job responsibilities will entail taking RTL logic through a full ASIC design flow. Worked with toolsets ... careers.microsoft.com
13 connections work here
2 weeks ago

Startups Building Custom Hardware

By CADE METZ JAN. 14, 2018

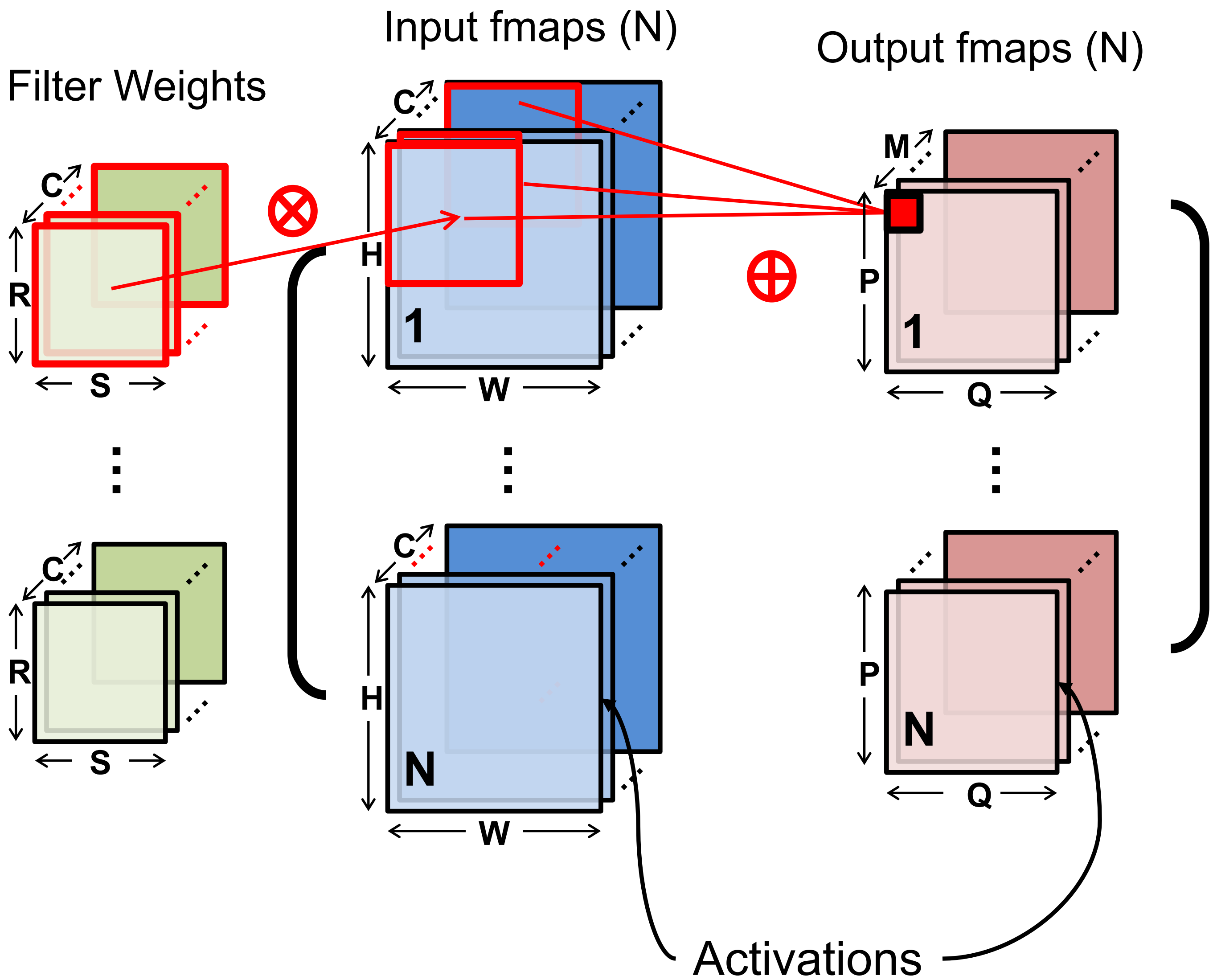
The New York Times



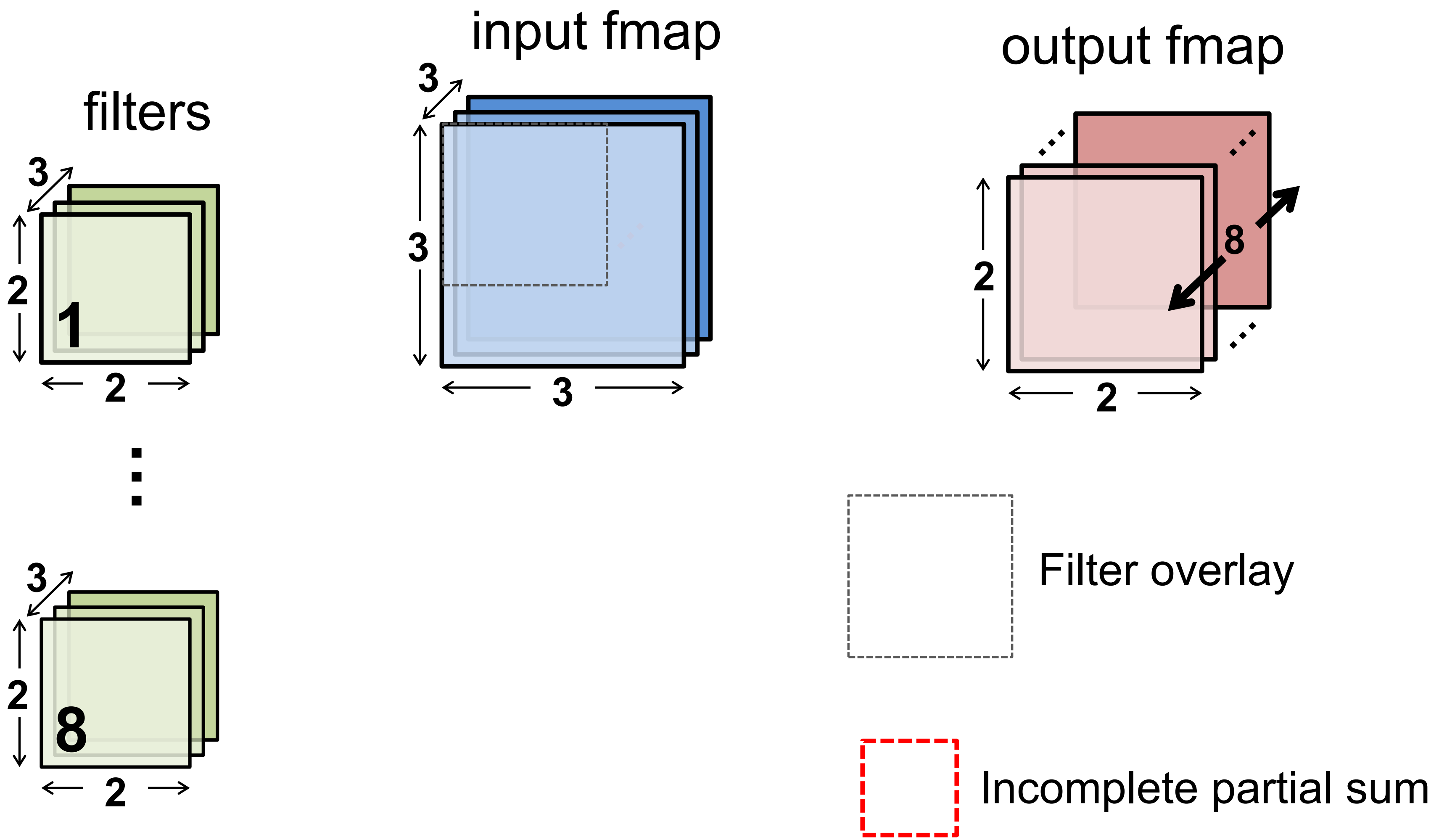
Big Bets On A.I. Open a New Frontier for Chips Start-Ups, Too. *(January 14, 2018)*

"Today, **at least 45 start-ups are working on chips** that can power tasks like speech and self-driving cars, and at least five of them have raised more than $100 million from investors. **Venture capitalists invested more than $1.5 billion in chip start-ups last year**, nearly doubling the investments made two years ago, according to the research firm CB Insights."

Convolution (CONV) Layer

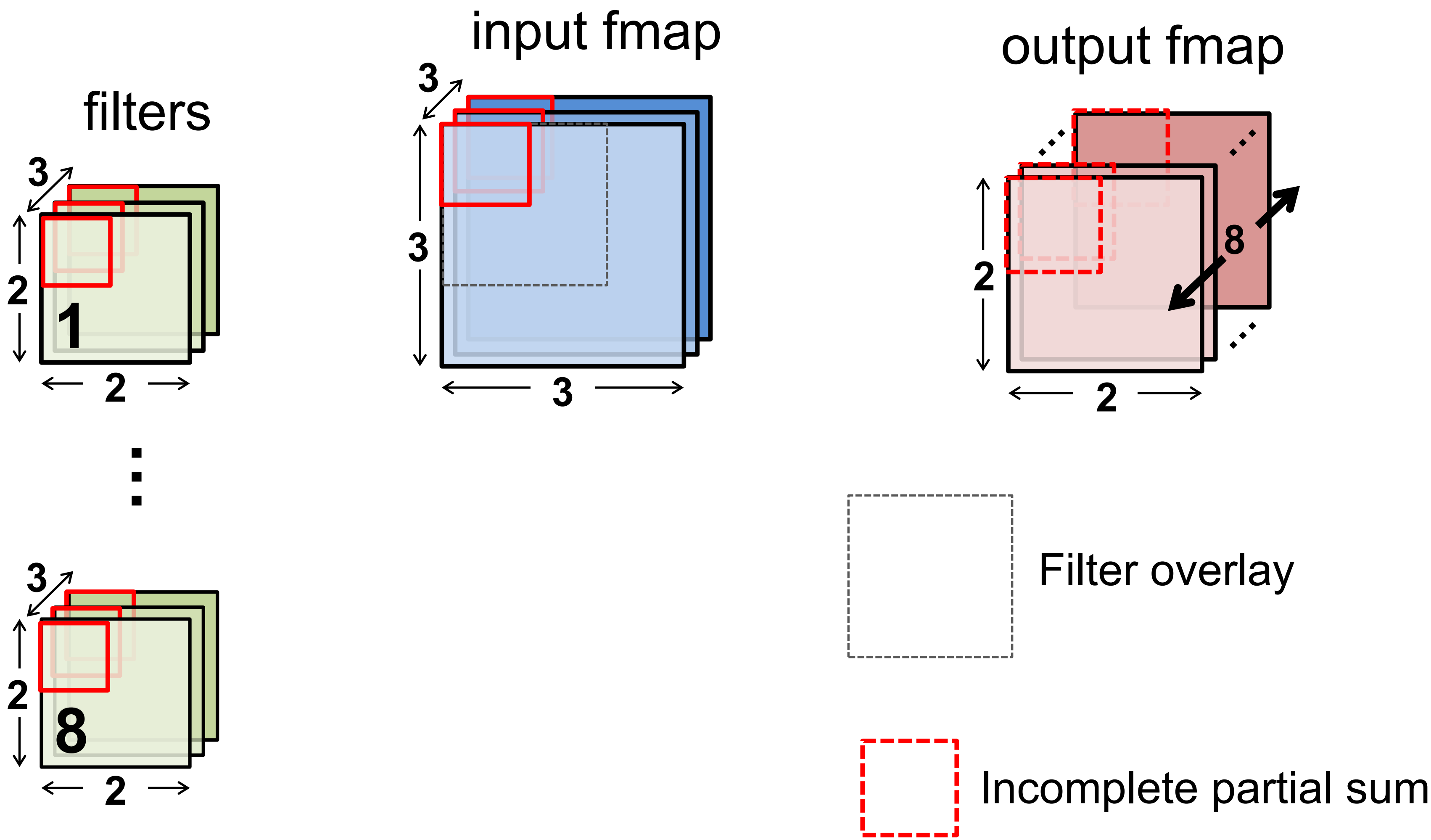


CONV Computation



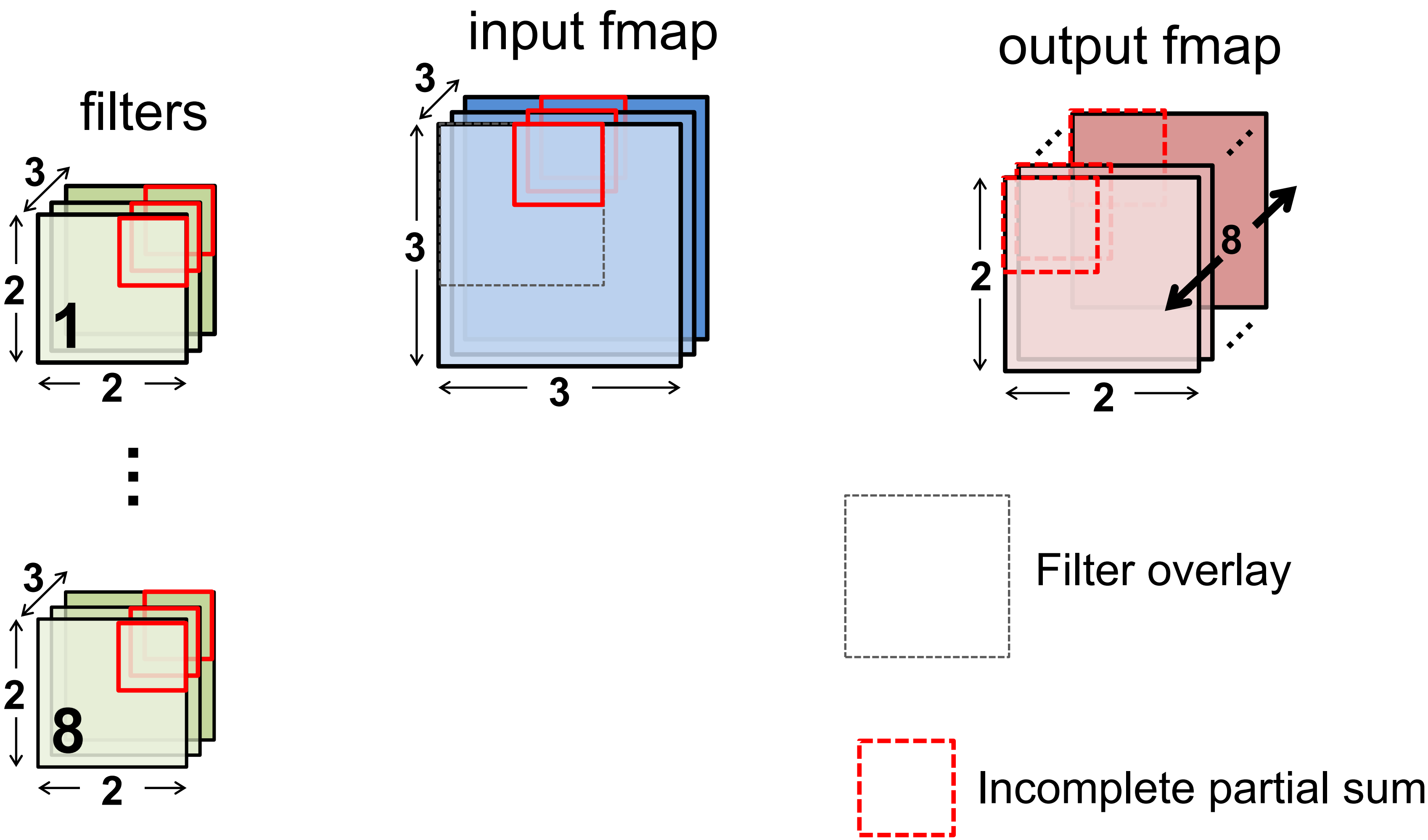
CONV Computation

Cycle through input fmap and weights (hold psum of output fmap)



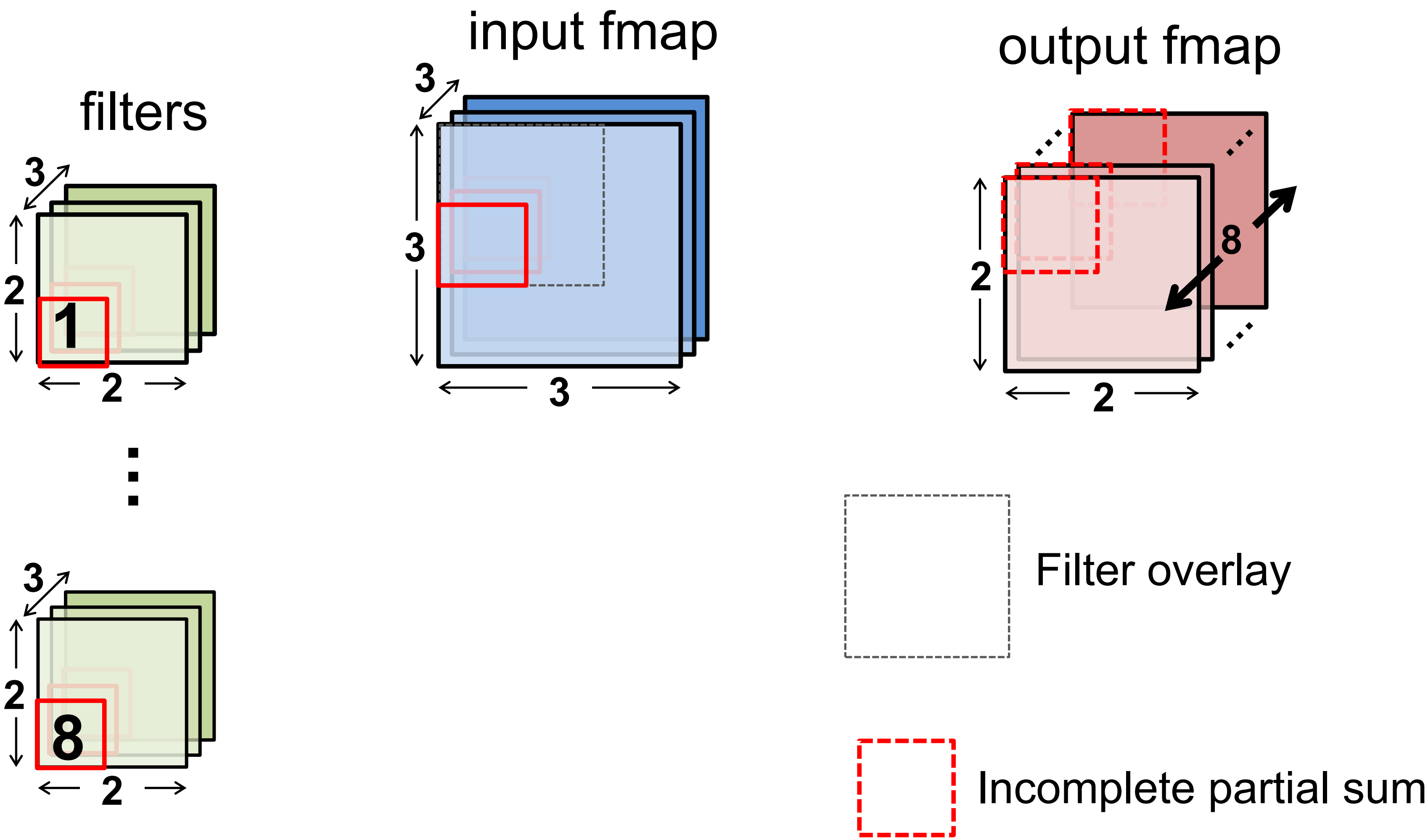
CONV Computation

Cycle through input fmap and weights (hold psum of output fmap)



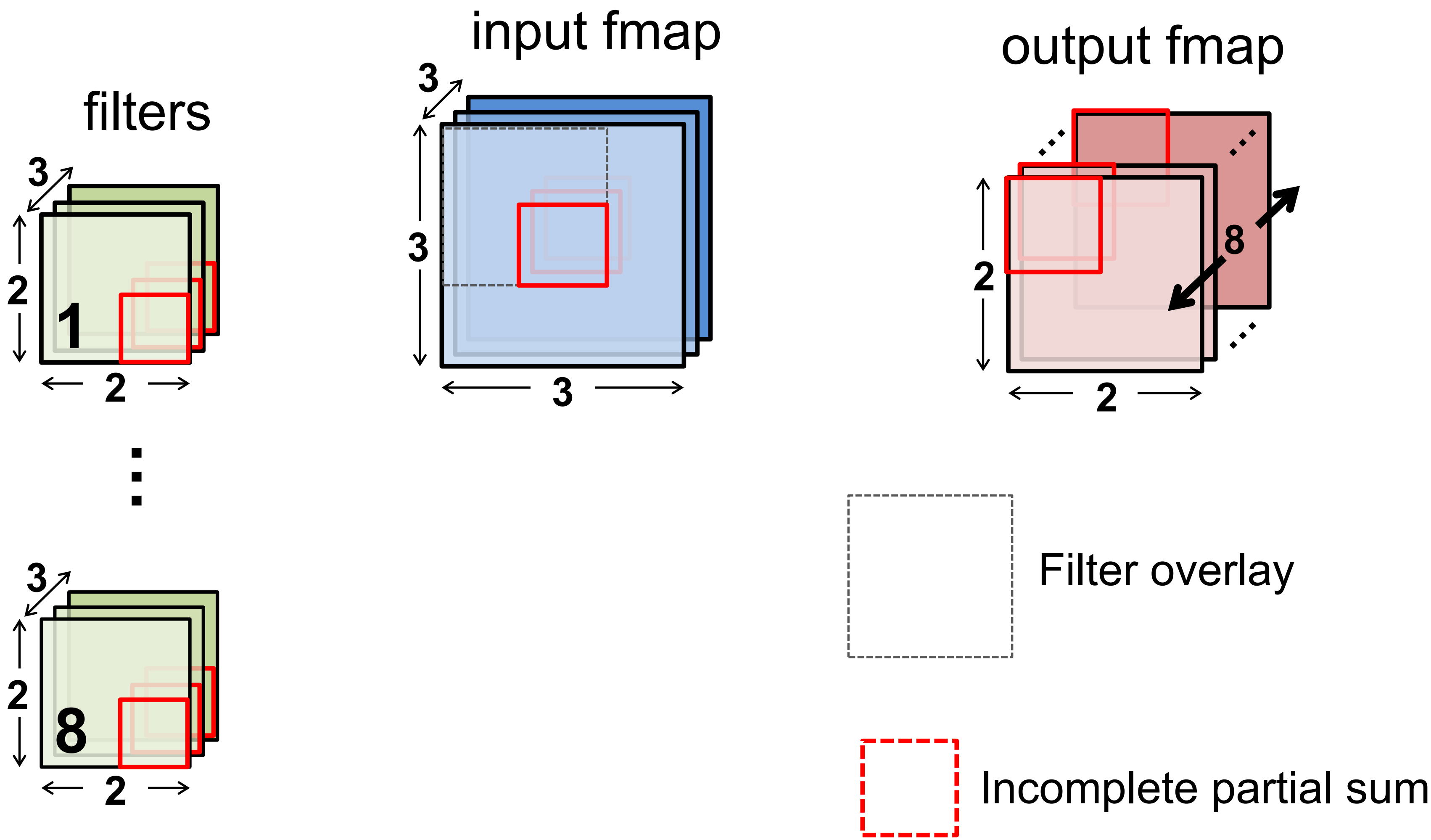
CONV Computation

Cycle through input fmap and weights (hold psum of output fmap)



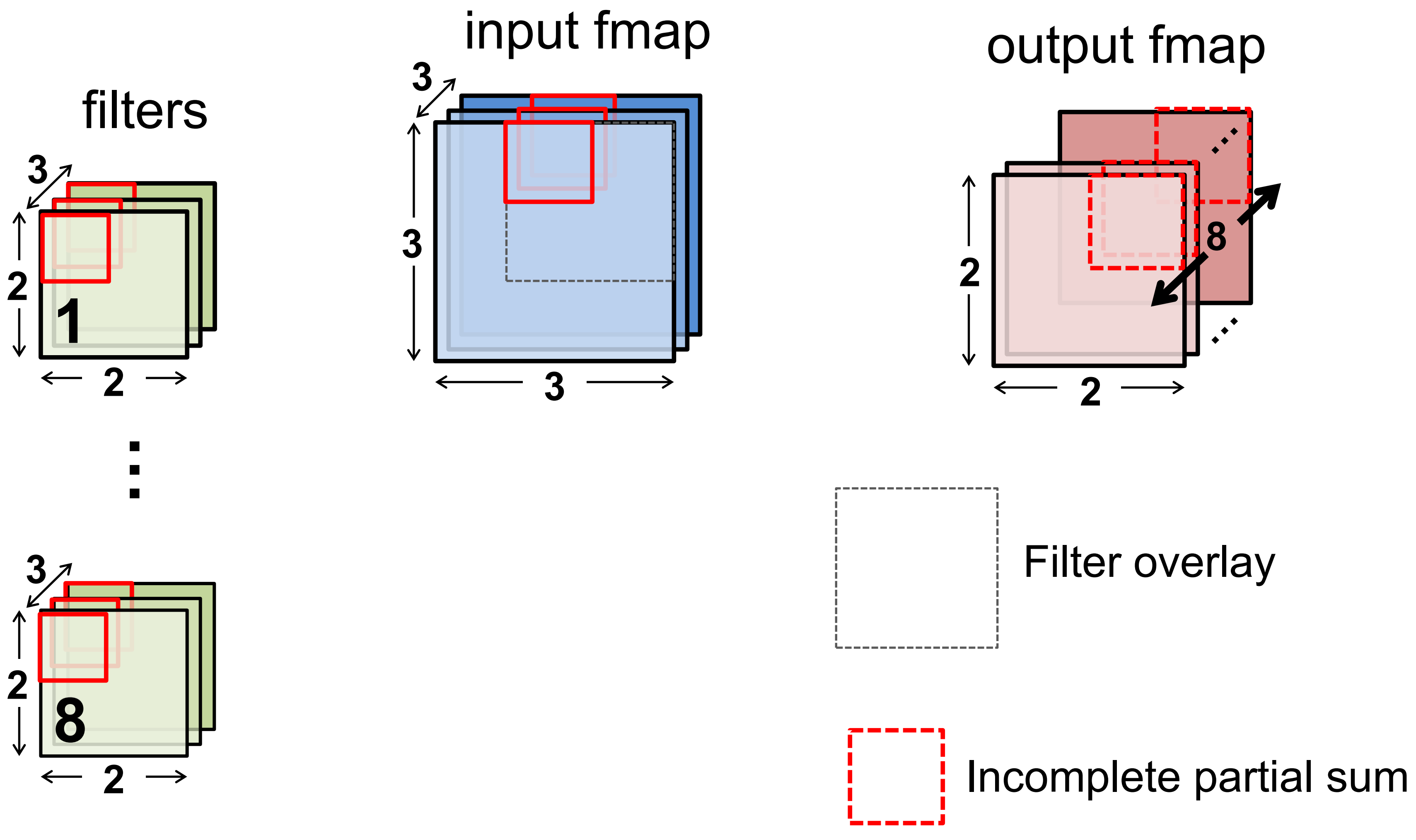
CONV Computation

Cycle through input fmap and weights (hold psum of output fmap)



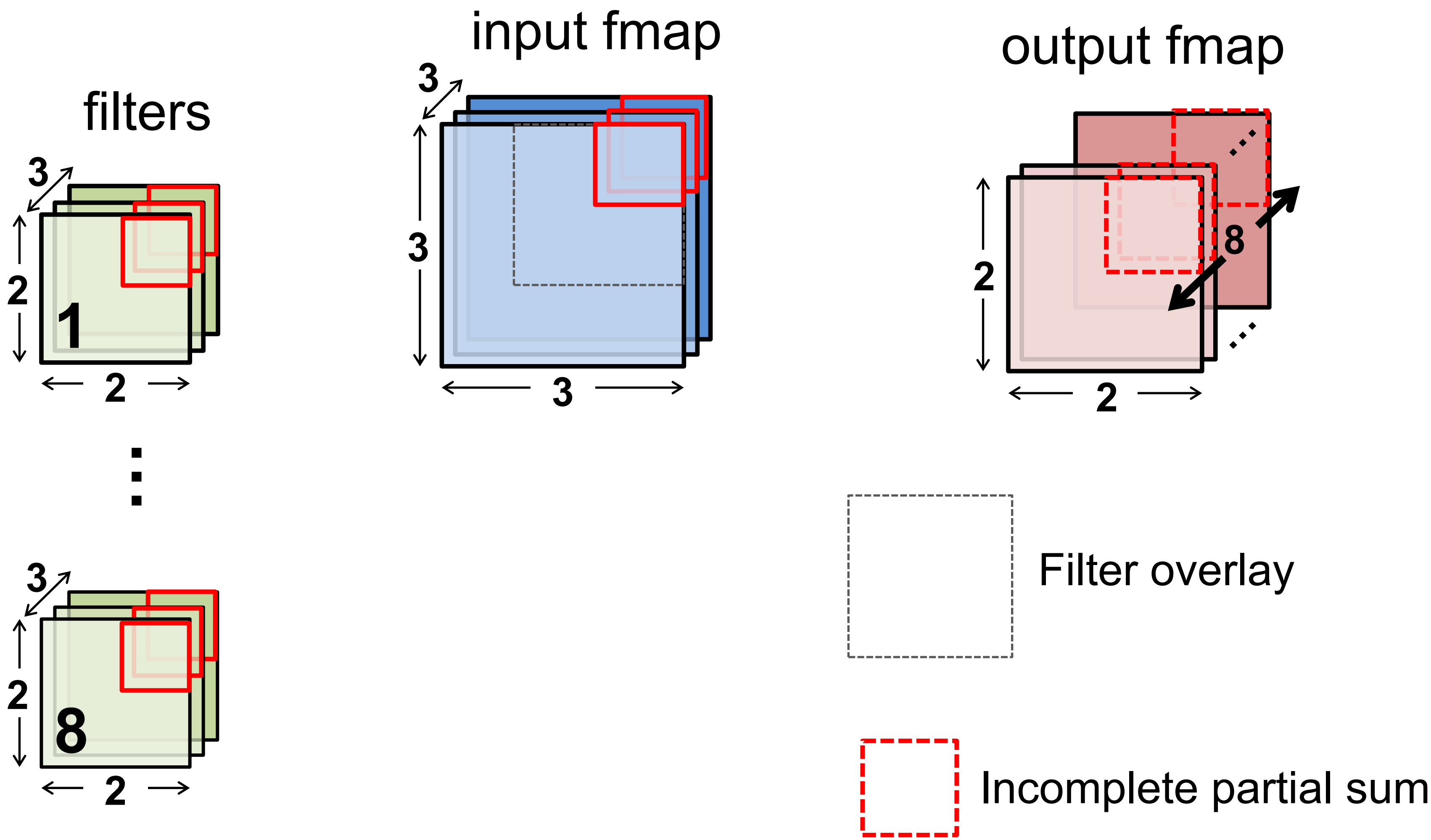
CONV Computation

Start processing next output feature activations



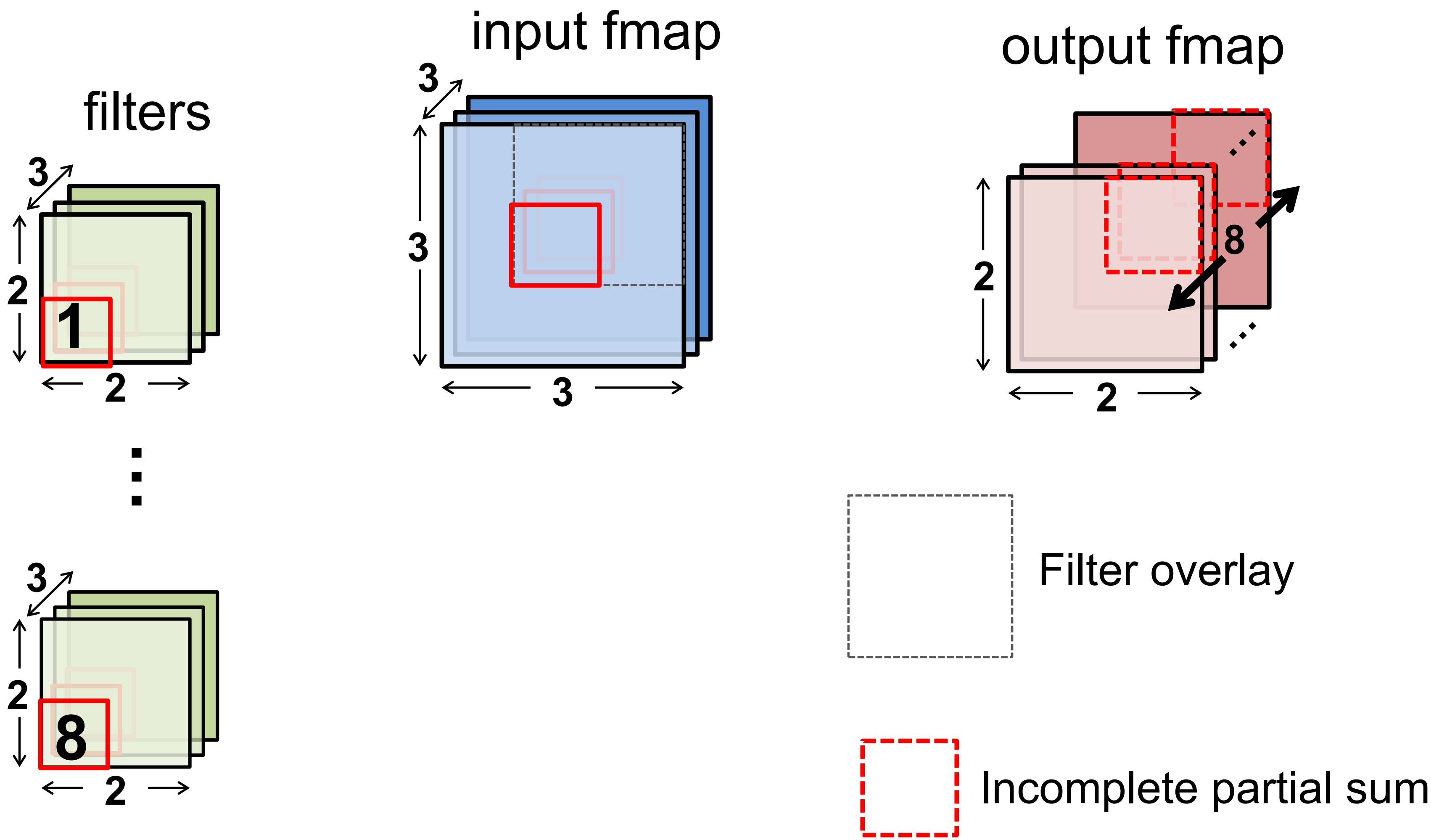
CONV Computation

Cycle through input fmap and weights (hold psum of output fmap)



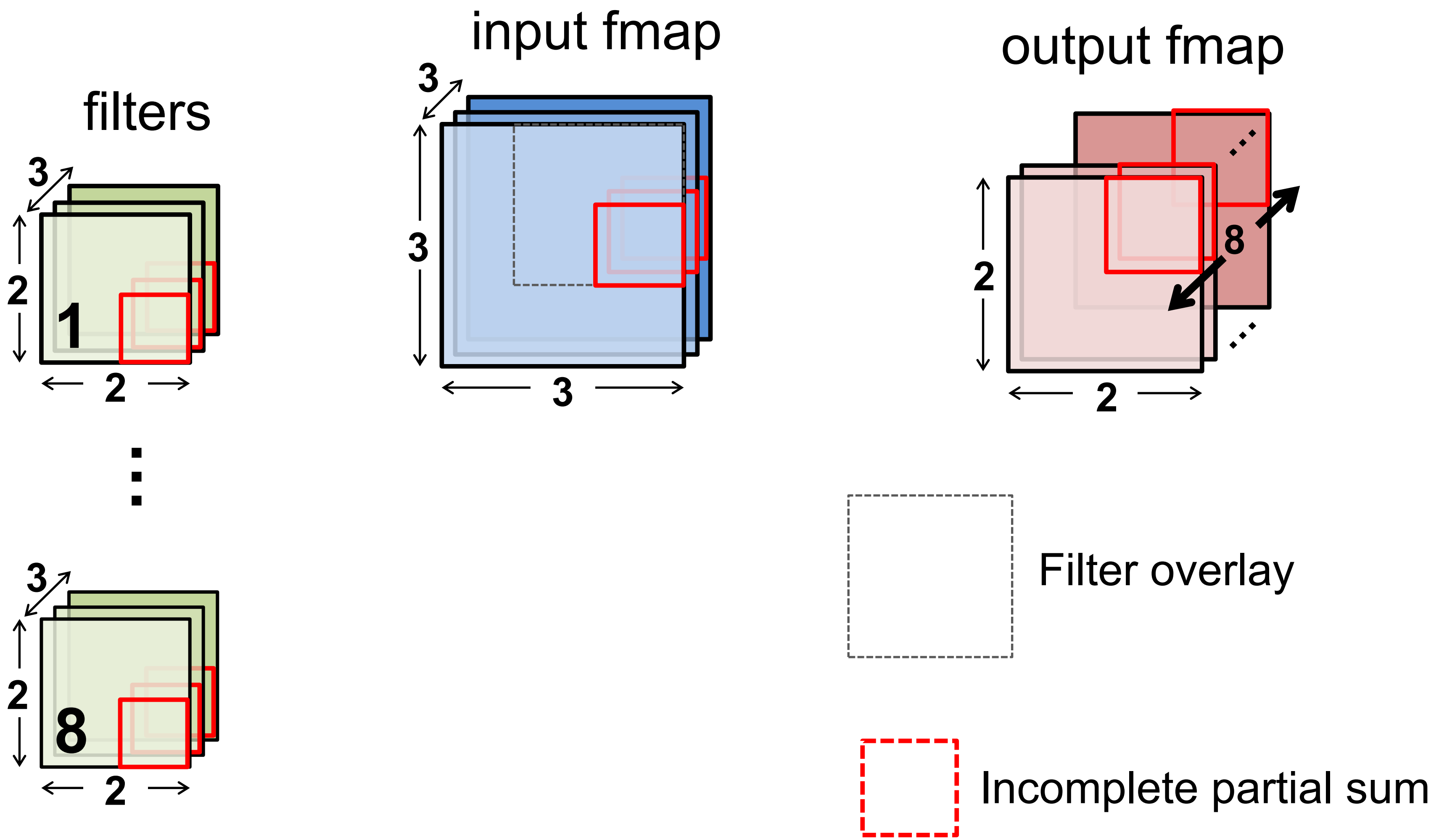
CONV Computation

Cycle through input fmap and weights (hold psum of output fmap)



CONV Computation

Cycle through input fmap and weights (hold psum of output fmap)



CONV Layer Implementation

Naïve 7-layer for-loop implementation:

for each output fmap value

convolve a window and apply activation

```
for n in [0..N):
    for m in [0..M):
        for q in [0..):
            for p in [0..P):
                O[n][m][p][q] = B[m];
                for r in [0..R):
                    for s in [0..S):
                        for c in [0..C):
                            O[n][m][p][q] += I[n][c][Up+r][Uq+s] × F[m][c][r][s];
                        }
                    }
                }

                O[n][m][p][q] = Activation(O[n][m][p][q]);
            }
        }
    }
}
```

CNN Decoder Ring

- N – Number of input fmaps/output fmaps (batch size)
- C – Number of channels in input fmaps (activations) & filters (weights)
- H – Height of input fmap (activations)
- W – Width of input fmap (activations)
- R – Height of filter (weights)
- S – Width of filter (weights)
- M – Number of channels in output fmaps (activations)
- P – Height of output fmap (activations)
- Q – Width of output fmap (activations)
- U – Stride of convolution

CONV Variants

- Depthwise layer - $M == C\ and\ \forall_{c\ != m} F_{c,m,r,s} = 0$
- Pointwise layer - $R == S == 1$
- Matrix multiply - $R == S == 1\ and\ flatten\ H/W$
- Compress (pointwise) - $M < C\ and\ R == S == 1$
- Expand (pointwise) - $M > C\ and\ R == S == 1$

Compress…Expand sequences are called a "bottleneck"

Architecture Metrics

- Speed – The rate at which the hardware finishes tasks. Limited by the number of computation units and their utilization.
- Energy – The total energy, e.g., in Joules, consumed to perform a task. Often constrained by battery capacity or desire to reduce carbon footprint.
- Power – The rate at which energy is consumed. Often limited by delivery or packaging constraints
- Accuracy – The precision of the results produced. Can be dictated by bit width of compute units.
- Flexibility – The range of problems that can be solved, which is constrained by the limitations of the architecture.

Deep Learning Platforms

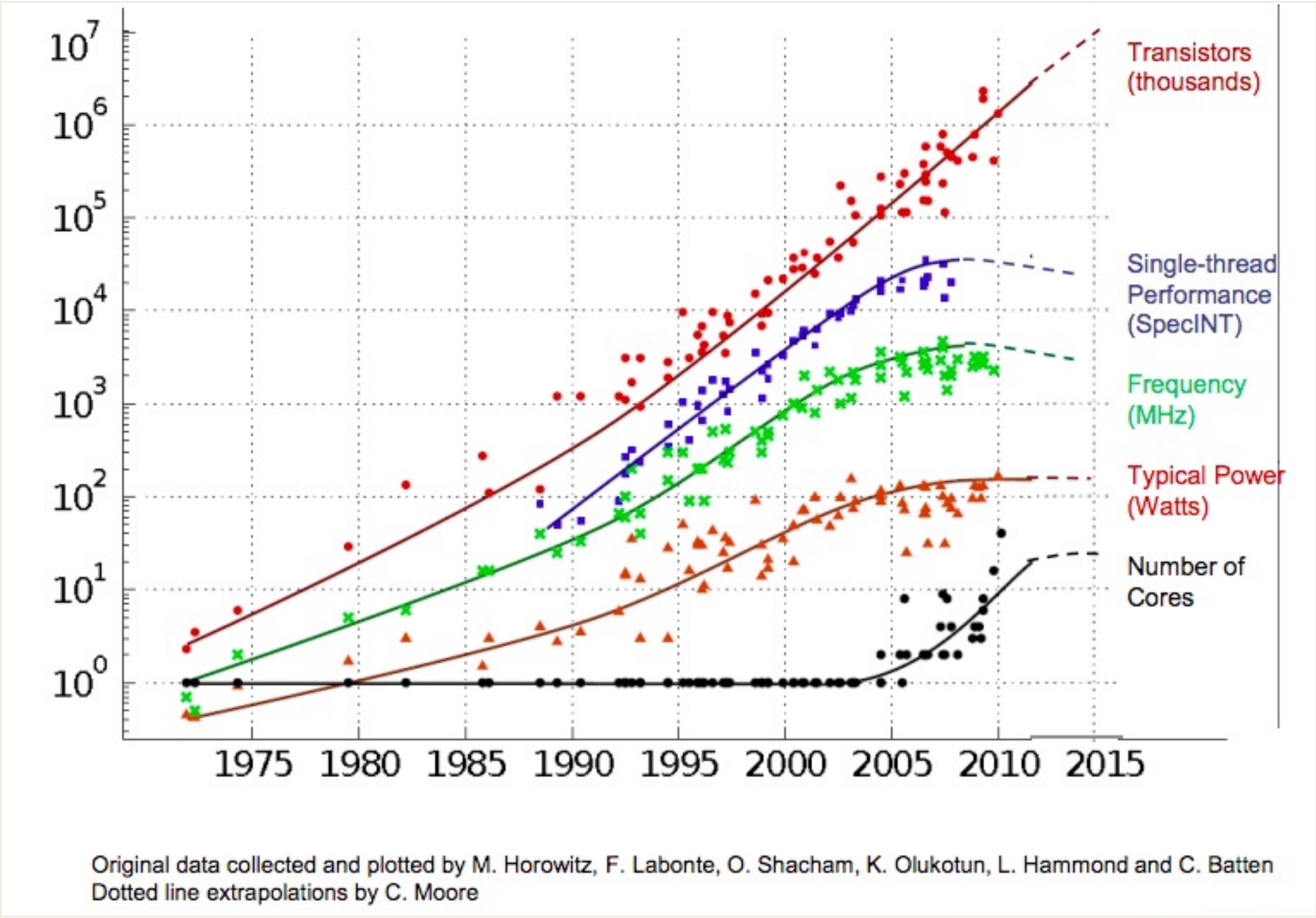
- CPU
  - Intel, ARM, AMD…
- GPU
  - NVIDIA, AMD…
- Fine Grained Reconfigurable (FPGA)
  - Microsoft BrainWave
- Coarse Grained Programmable/Reconfigurable
  - Wave Computing, Plasticine, Graphcore…
- Application Specific
  - Neuflow, *DianNao, Eyeriss, TPU, Cnvlutin, SCNN, …

What is Moore's Law

- CPU performance will double every two years*
- Chip performance will double every two years*
- The speed of transistors will double every two years*
- Transistors will shrink to half size every two years*
- Gate width will shrink in half every two years*
- Transistors per die will double every two years*
- The economic sweet spot for the number of devices on a chip will double every two years*

* Or 18 months…

Technology Trends

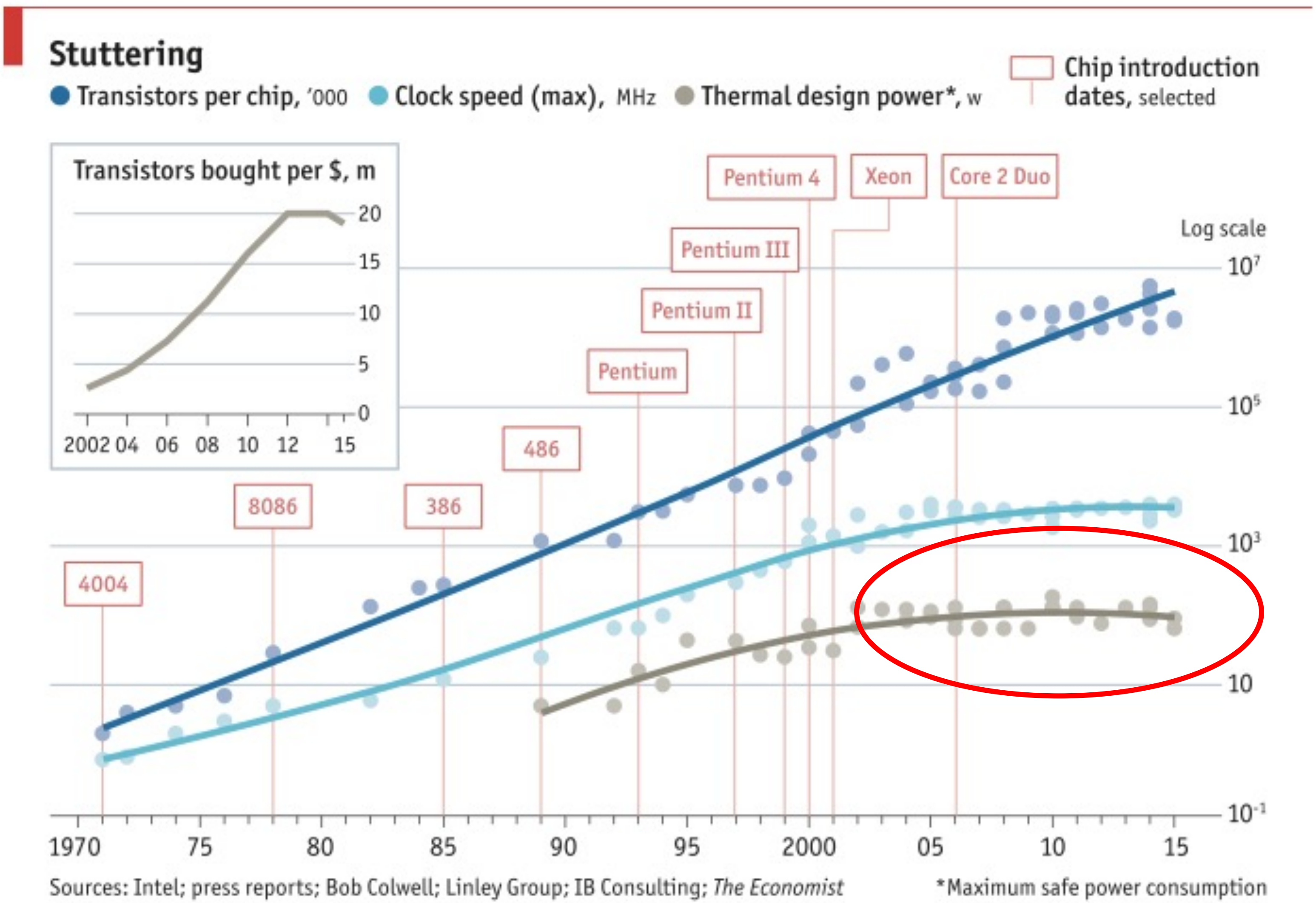


Source: C Moore, Data Processing in ExaScale-ClassComputer Systems, Salishan, April 2011

During the Moore + Dennard's Law Era

- Instruction-level parallelism (ILP) was largely mined out by early 2000s
- Voltage (Dennard) scaling ended in 2005
- Hit the power limit wall in 2005
- Performance is coming from parallelism using more transistors since ~2007
- But….

Technology Trends



The High Cost of Data Movement

Fetching operands more expensive than computing on them

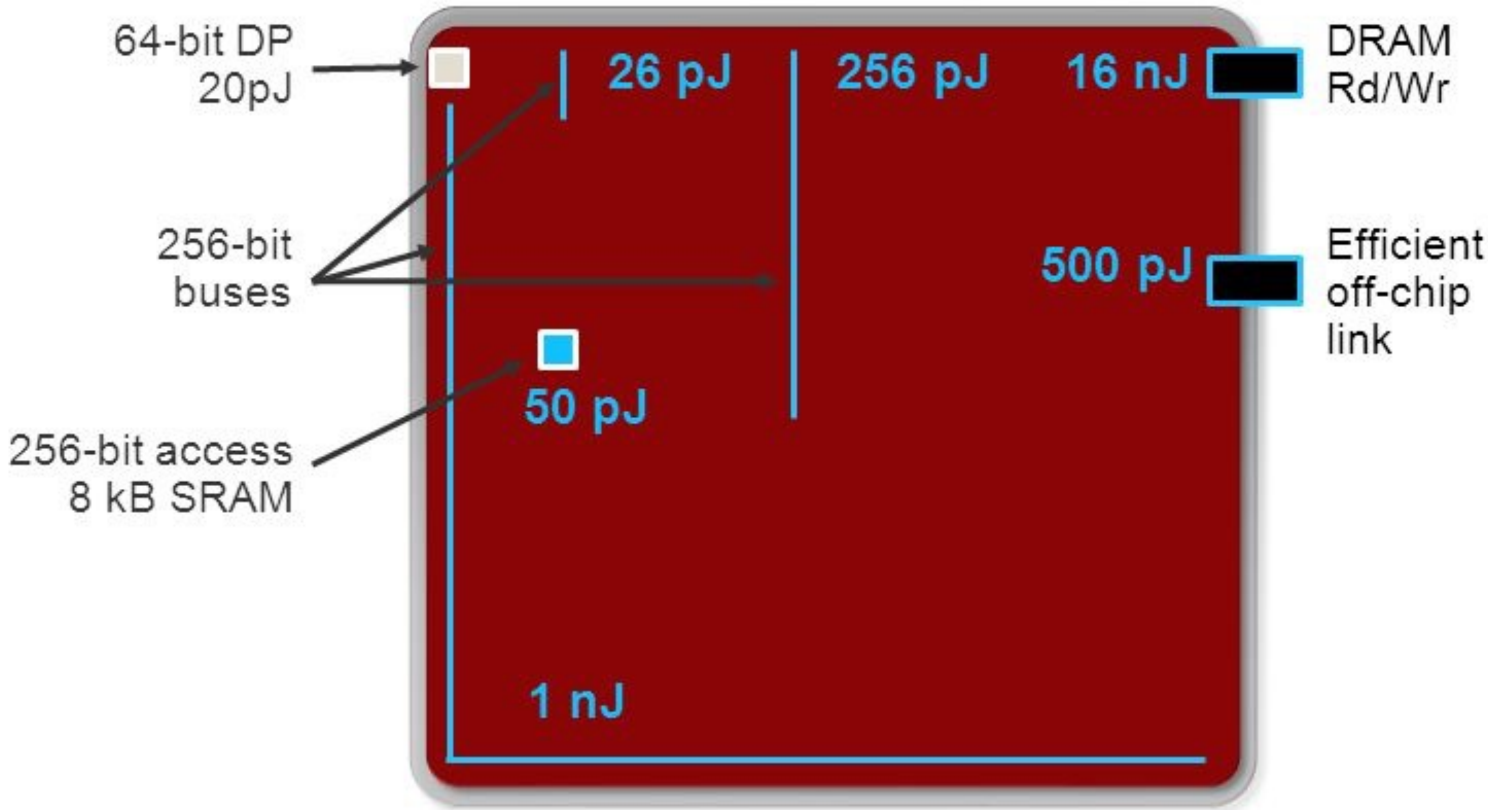
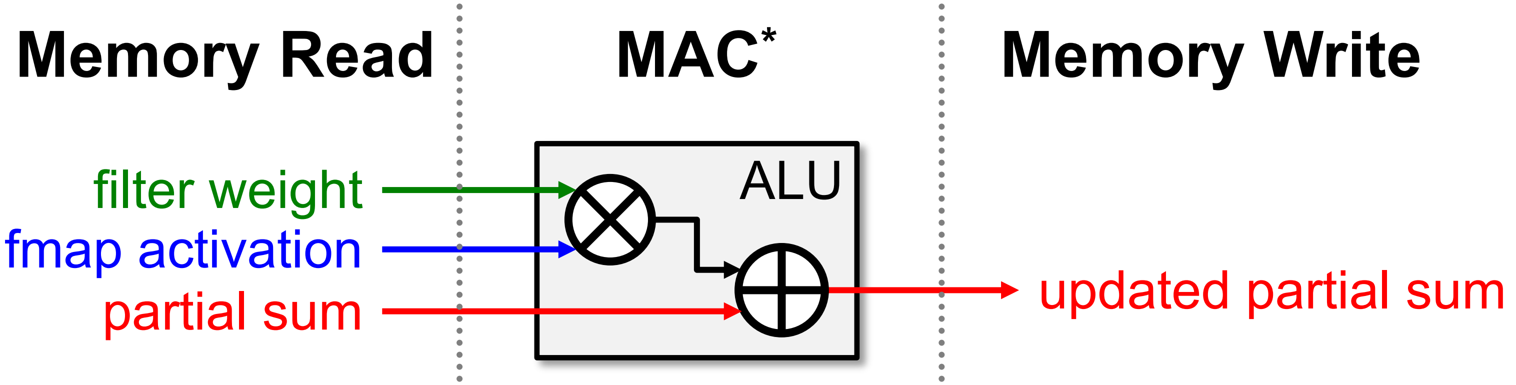


Image source: Bill Daly

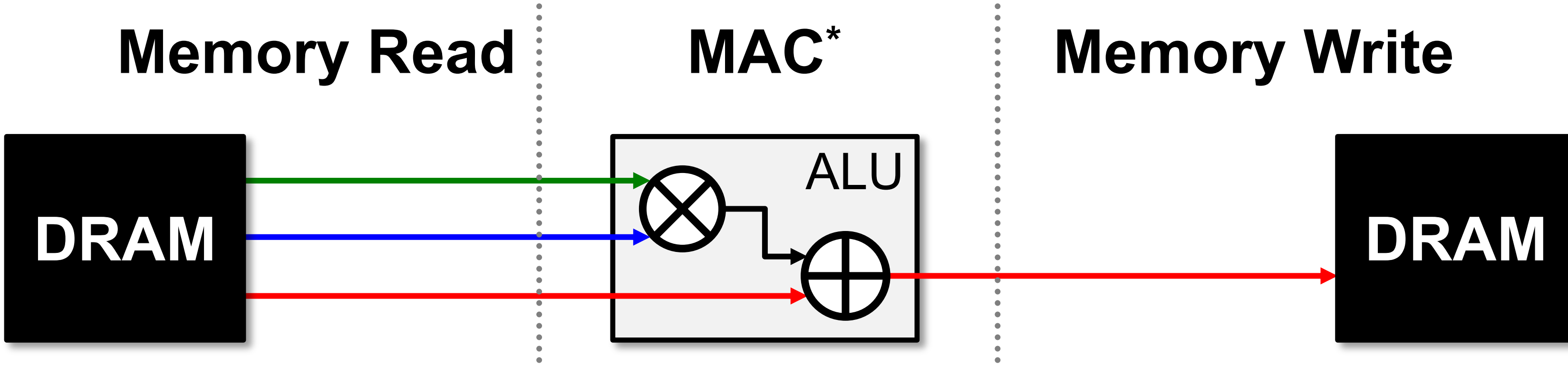
Now the key is how we use our transistors most effectively.

Data Movement is the Challenge



* multiply-and-accumulate

Data Movement is the Challenge

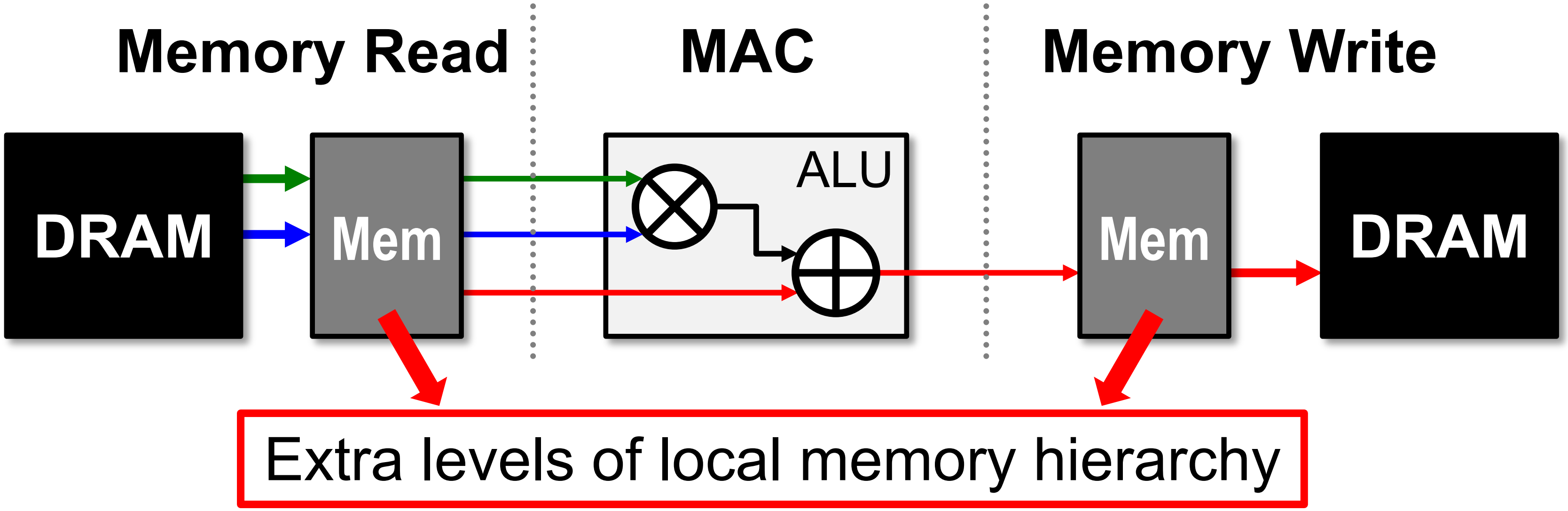


* multiply-and-accumulate

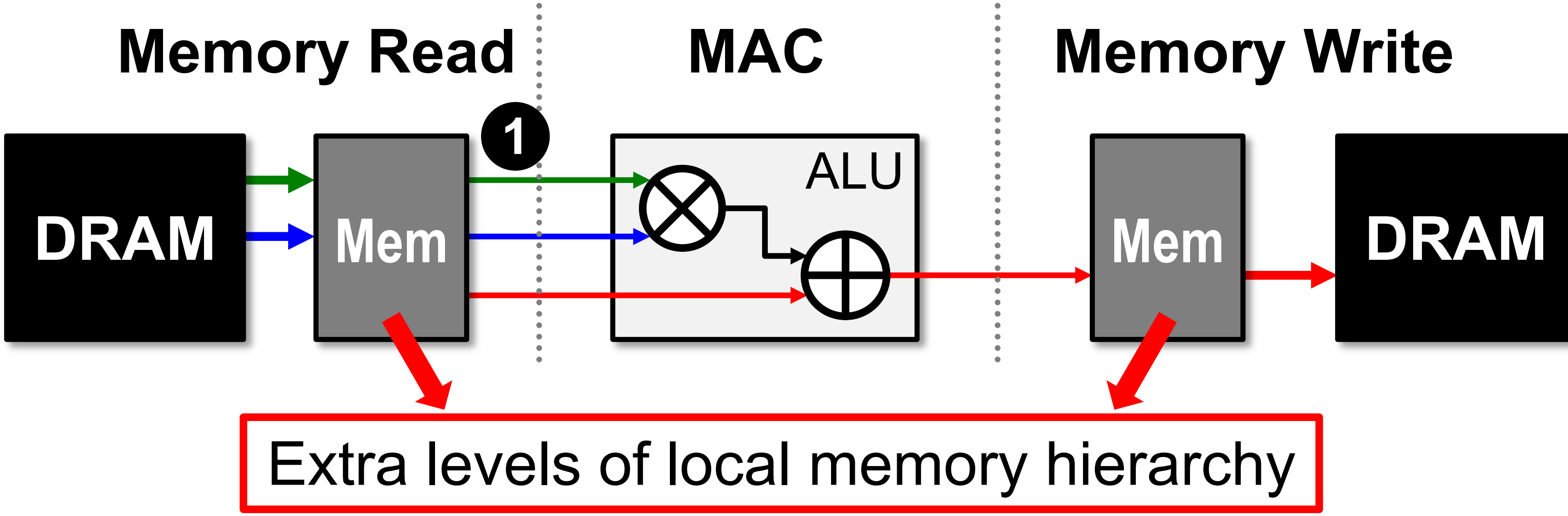
<u>Worst Case</u>: all memory R/W are **DRAM** accesses

- Example: AlexNet [NeurIPS 2012] has **724M** MACs
  → **2896M** DRAM accesses required

Data Movement is the Challenge



Data Movement is the Challenge

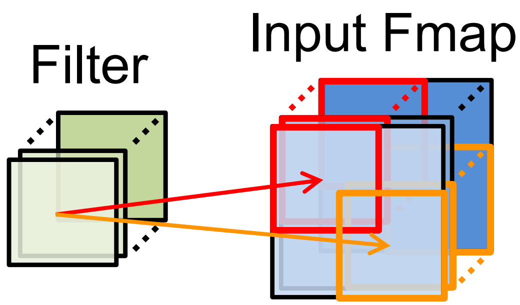


Opportunities: ❶ **data reuse**

Types of Data Reuse in DNN

Convolutional Reuse

CONV layers only
(sliding window)

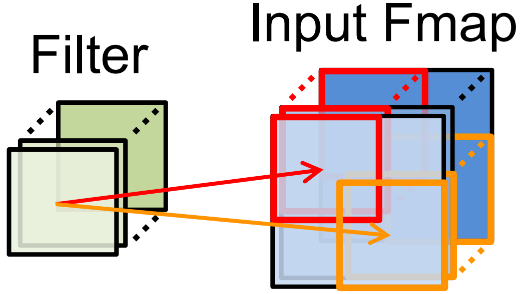


Reuse: Activations
Filter weights

Types of Data Reuse in DNN

Convolutional Reuse

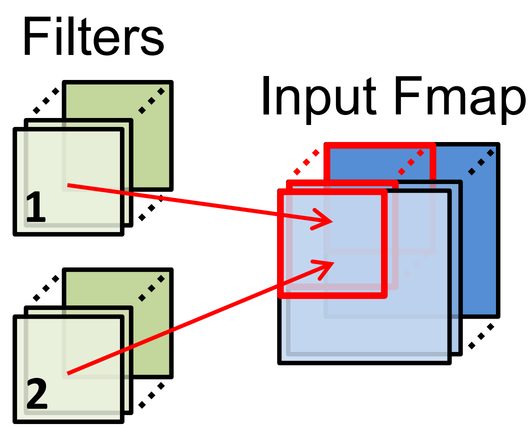
CONV layers only
(sliding window)



Reuse: Activations
Filter weights

Fmap Reuse

CONV and FC layers

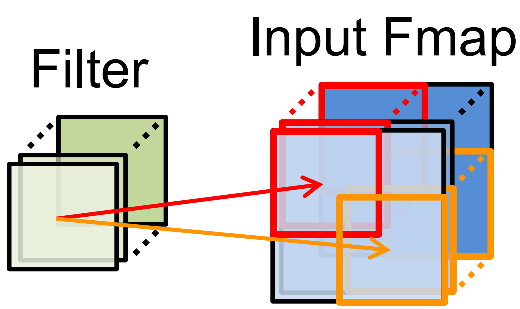


Reuse: Activations

Types of Data Reuse in DNN

Convolutional Reuse

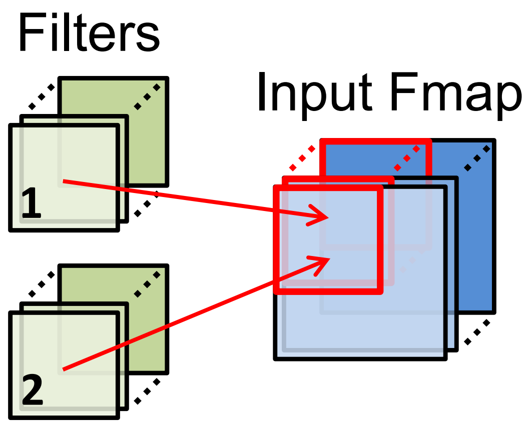
CONV layers only
(sliding window)



Reuse: Activations
Filter weights

Fmap Reuse

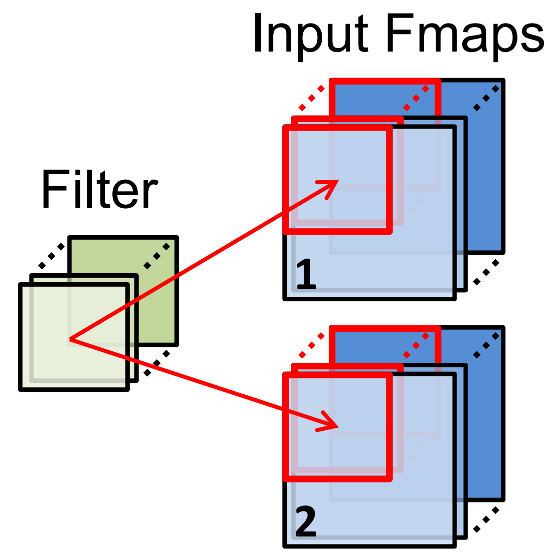
CONV and FC layers



Reuse: Activations

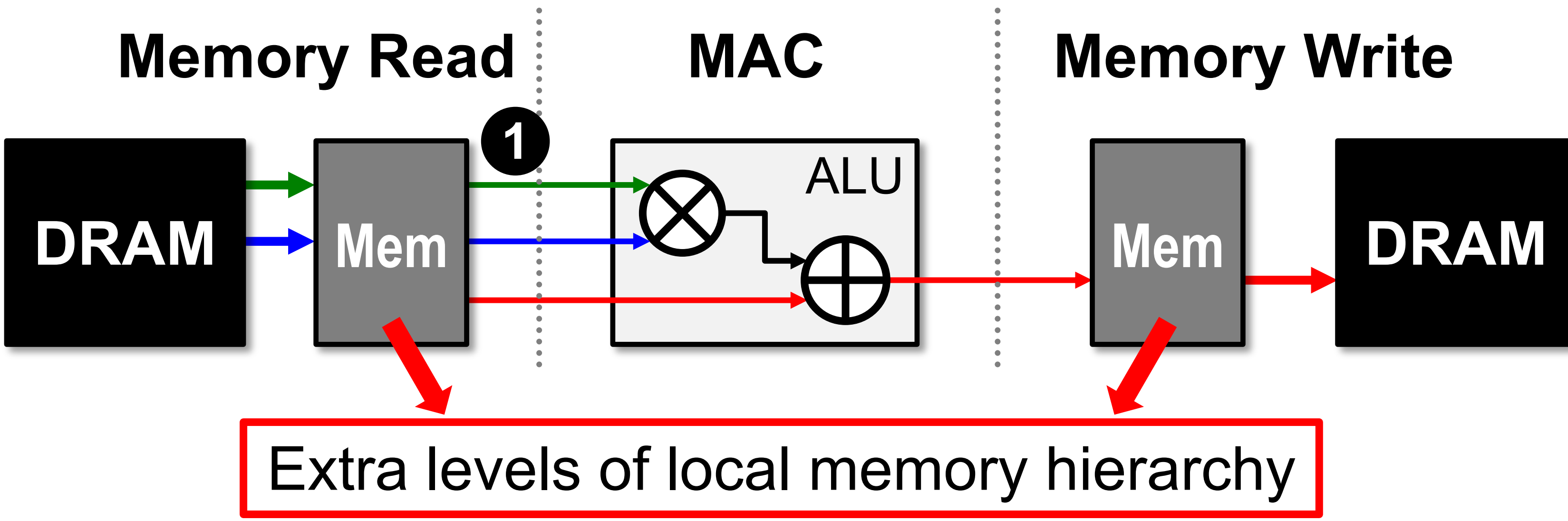
Filter Reuse

CONV and FC layers
(batch size > 1)



Reuse: Filter weights

Data Movement is the Challenge

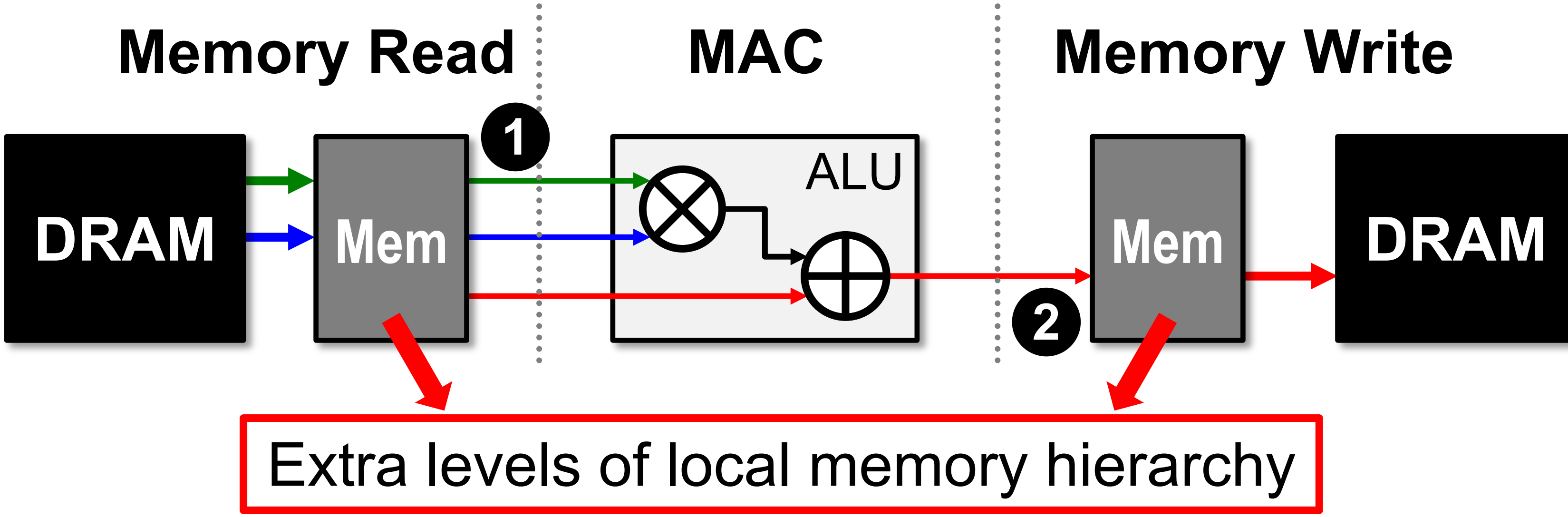


Opportunities: ❶ **data reuse**

❶ Can reduce DRAM reads of filter/fmap by up to **500×**[**]

** AlexNet CONV layers

Data Movement is the Challenge

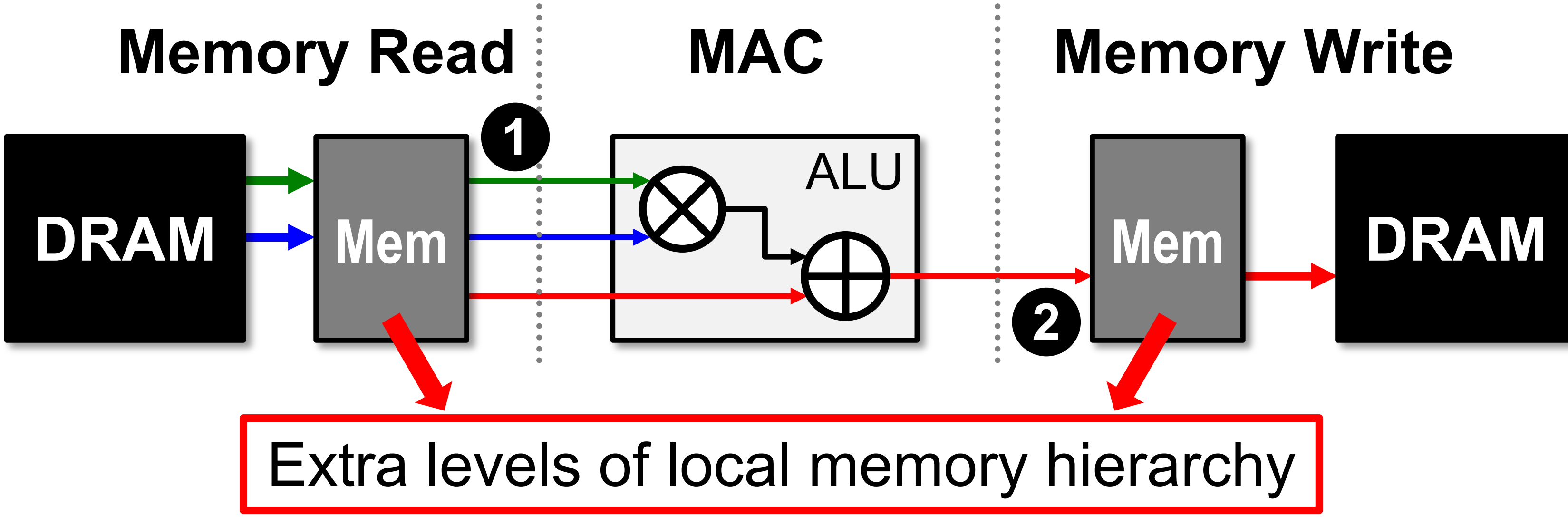


Opportunities: ❶ **data reuse** ❷ **local accumulation**

❶ Can reduce DRAM reads of filter/fmap by up to **500×**

❷ Partial sum accumulation does **NOT** have to access DRAM

Data Movement is the Challenge

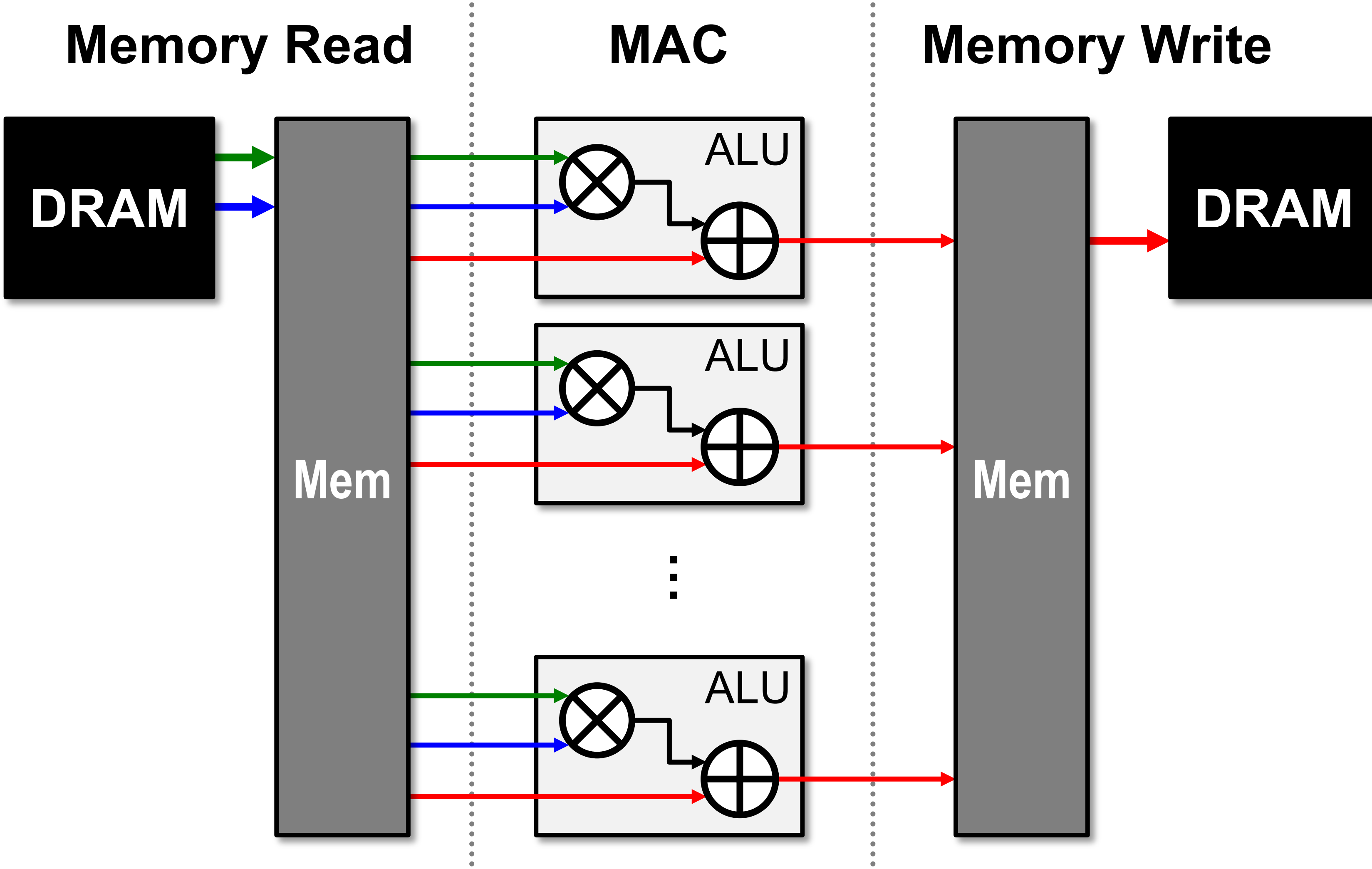


Opportunities: ❶ **data reuse** ❷ **local accumulation**

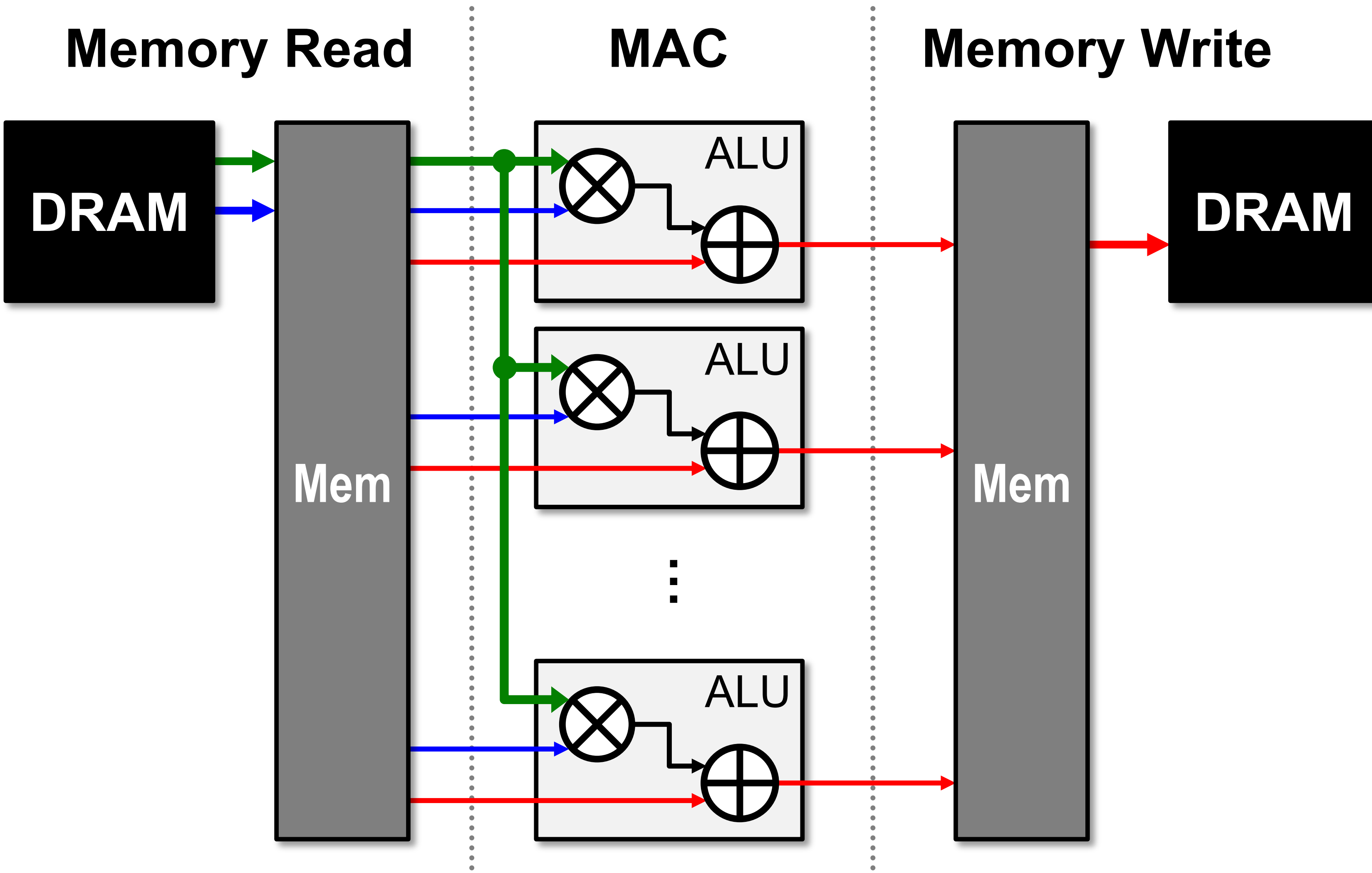
❶ Can reduce DRAM reads of filter/fmap by up to **500×**
❷ Partial sum accumulation does **NOT** have to access DRAM

- Example: DRAM access in AlexNet can be reduced from **2896M** to **61M** (best case)

Leverage Parallelism for Higher Performance



Leverage Parallelism for *Spatial* Data Reuse



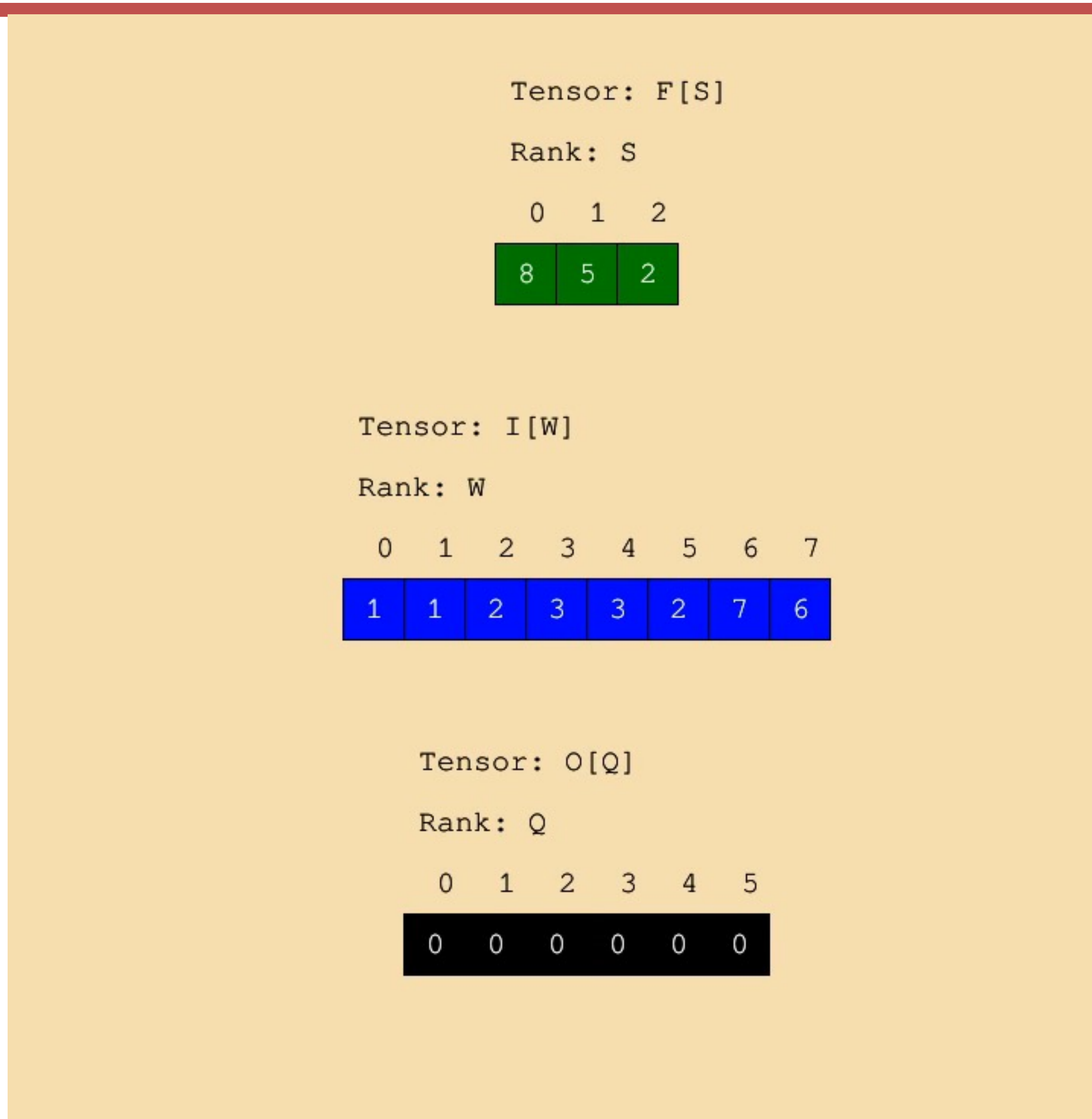
1-D Convolution


```
int i[W];      # Input activations
int f[S];      # Filter weights
int o[Q];      # Output activations

for q in [0, Q):
    for s in [0, S):
        w = q+s
        o[q] += i[w]*f[s];
```

† Assuming: ‘valid’ style convolution

Output Stationary - Movie



Output Stationary – Spacetime View

Tensor: F[S, T]

Rank: T

Rank: S	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Tensor: I[W, T]

Rank: T

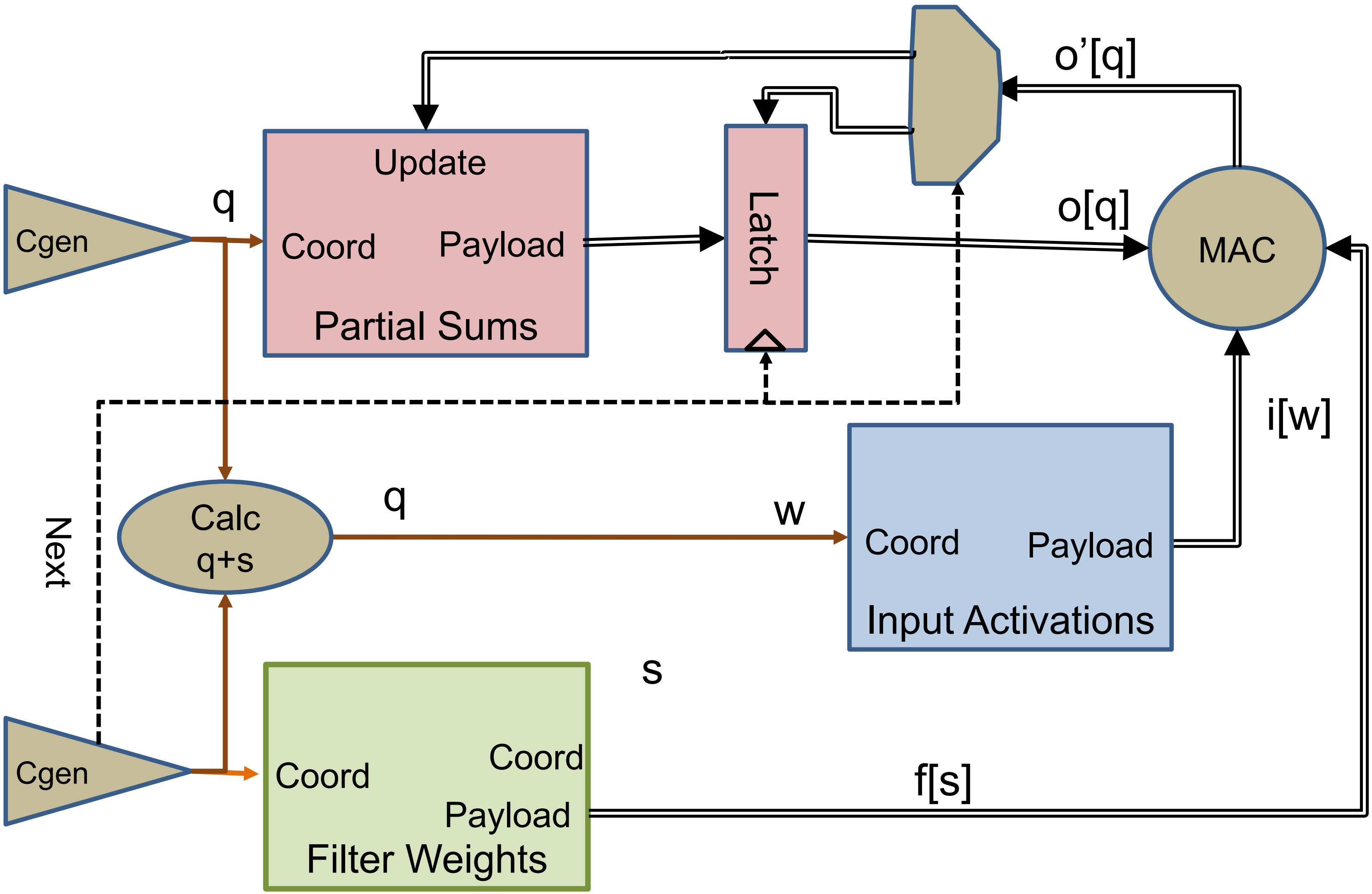
Rank: W	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
5	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
7	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6

Tensor: O[Q, T]

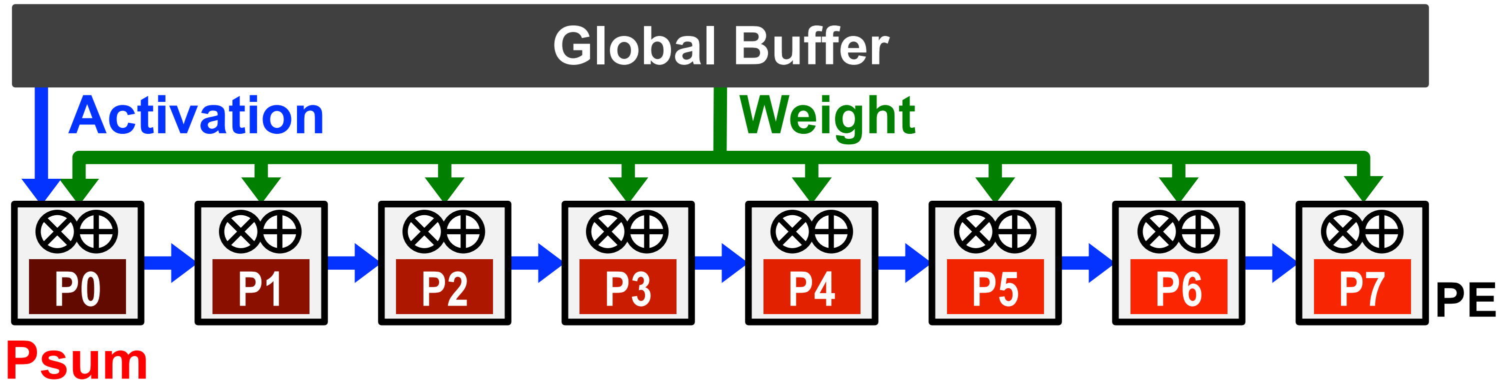
Rank: T

Rank: Q	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	8	13	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17
1	0	0	0	8	18	24	24	24	24	24	24	24	24	24	24	24	24	24
2	0	0	0	0	0	0	16	31	37	37	37	37	37	37	37	37	37	37
3	0	0	0	0	0	0	0	0	0	24	39	43	43	43	43	43	43	43
4	0	0	0	0	0	0	0	0	0	0	0	0	24	34	48	48	48	48
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	51	63

1-D Output Stationary



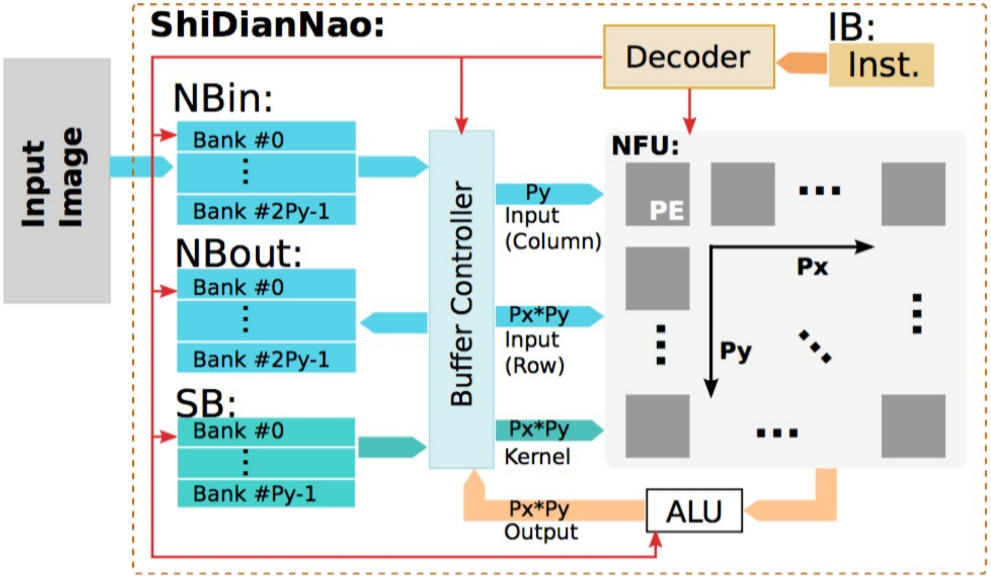
Output Stationary (OS)



- **Minimize partial sum** R/W energy consumption
  - maximize local accumulation
- **Broadcast/Multicast filter weights** and **reuse activations spatially** across the PE array

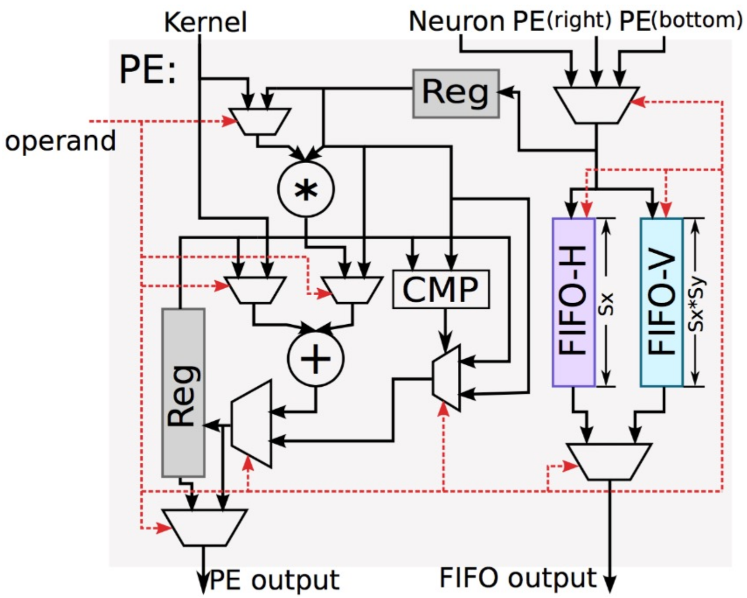
OS Example: ShiDianNao

Top-Level Architecture



PE Architecture

weights activations

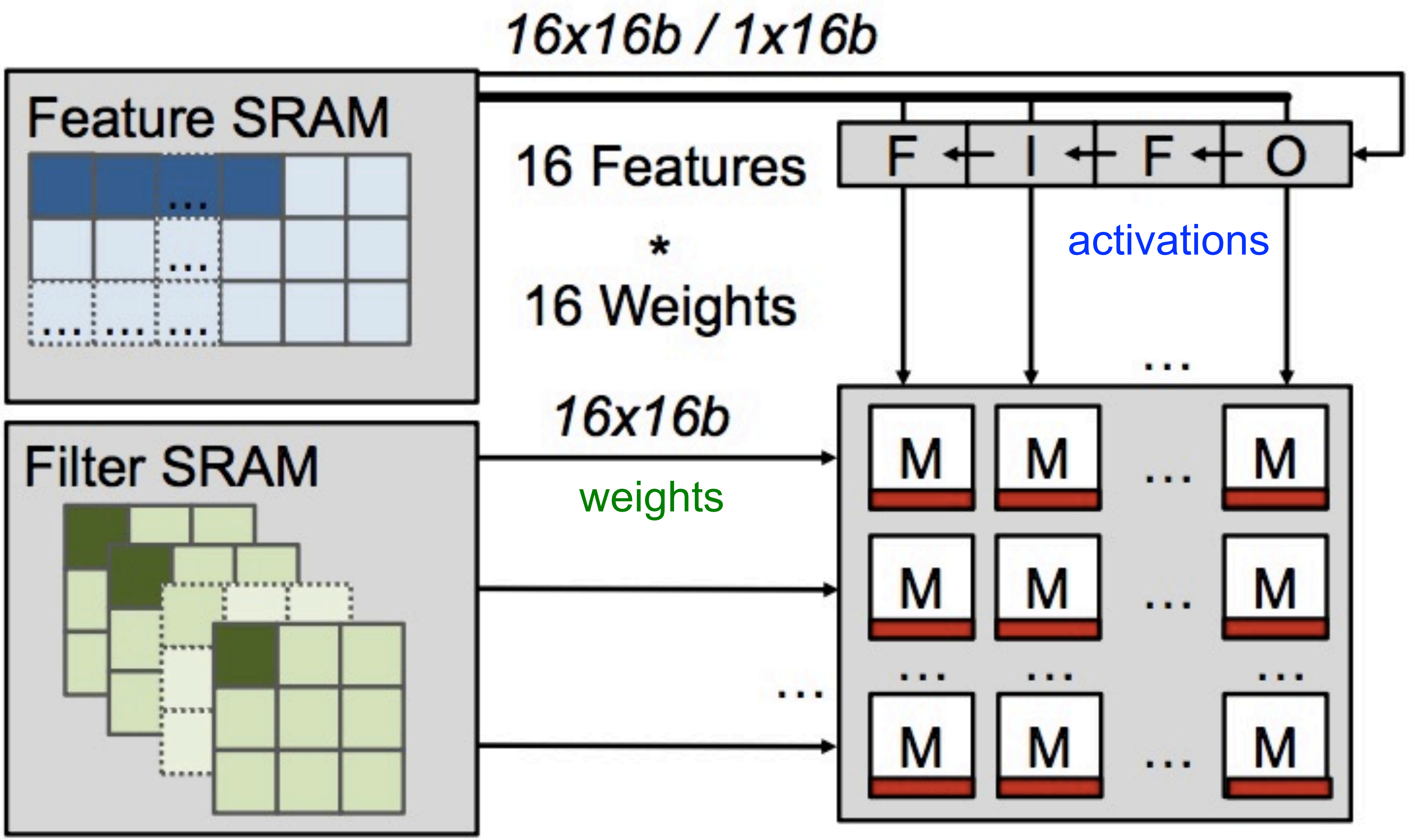


psums

- Inputs streamed through array
- Weights broadcast
- Partial sums accumulated in PE and streamed out

[Du et al., ISCA 2015]

OS Example: KU Leuven



[Moons et al., VLSI 2016, ISSCC 2017]

Many Dataflows

- **Output Stationary (OS)**

  [**Peemen**, *ICCD* 2013] [**ShiDianNao**, *ISCA* 2015]
  [**Gupta**, *ICML* 2015] [**Moons**, *VLSI* 2016] [**Thinker**, *VLSI* 2017]

- **Weight Stationary (WS)**

  [**Chakradhar**, *ISCA* 2010] [**nn-X (NeuFlow)**, *CVPRW* 2014]
  [**Park**, *ISSCC* 2015] [**ISAAC**, *ISCA* 2016] [**PRIME**, *ISCA* 2016]
  [**TPU**, *ISCA* 2017]

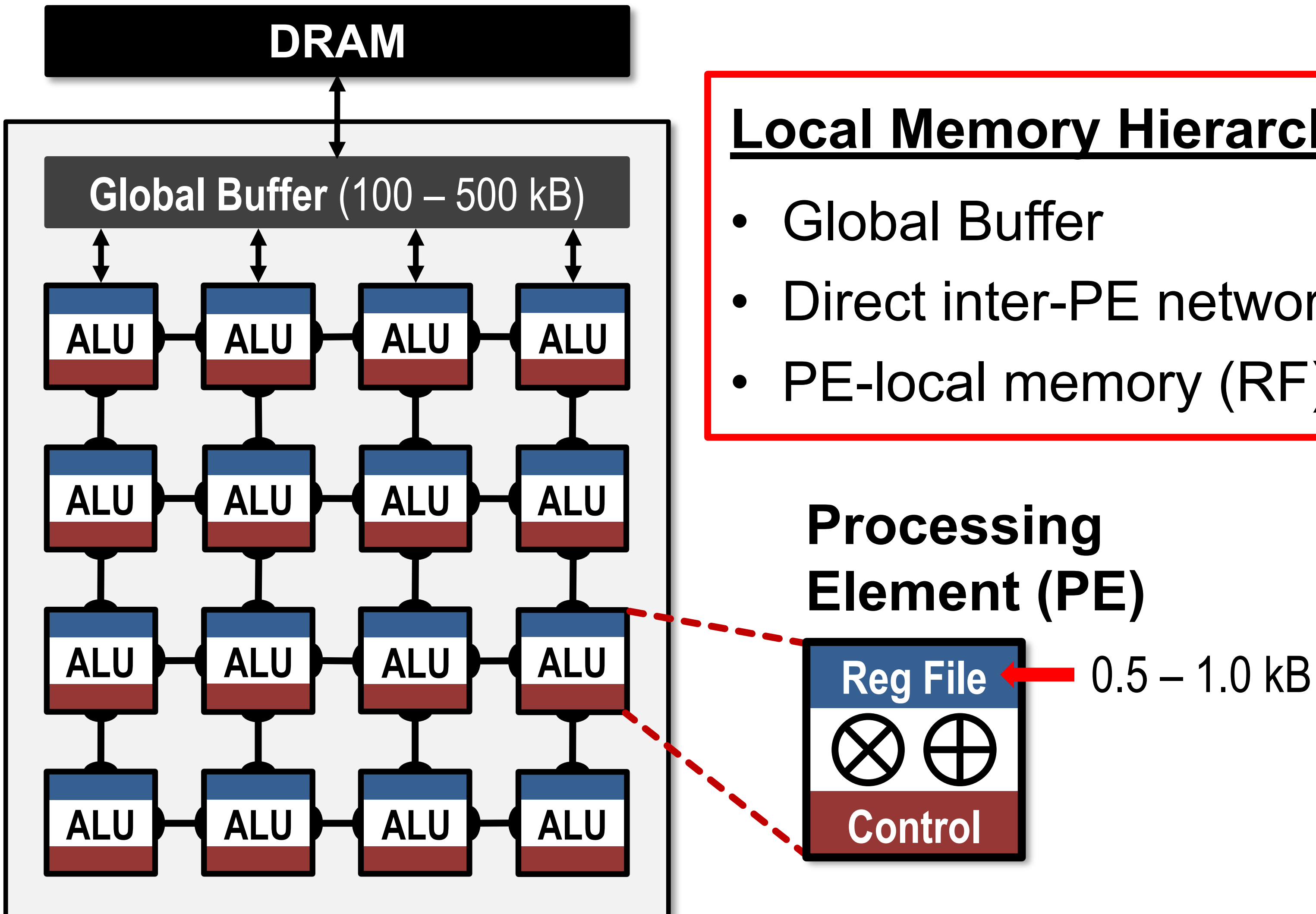
- **Input Stationary (IS)**

  [**Parashar (SCNN)**, ISCA 2017]

- **Row Stationary (IS)**

  [**Eyeriss**, ISCA 2016] [**Tetris** ASPLOS 2017] [**Eyeriss2**, JETCAT 2019]

Spatial Architecture for DNN



Local Memory Hierarchy

- Global Buffer
- Direct inter-PE network
- PE-local memory (RF)

And Other Design Options

Per storage level cross product of:

- **Dataflow**
- **Split/Shared storage**
- **Tiling in time**
- **Tiling in space**
- **Bypassing**

Plus

- **Scale up**
- **Precision/Quantization**
- **….**

And flexibility!

Thank you!

Next Lecture:
Accelerators (II)