Exploiting Redundancy for Efficient Processing of Deep Neural Nets and Beyond

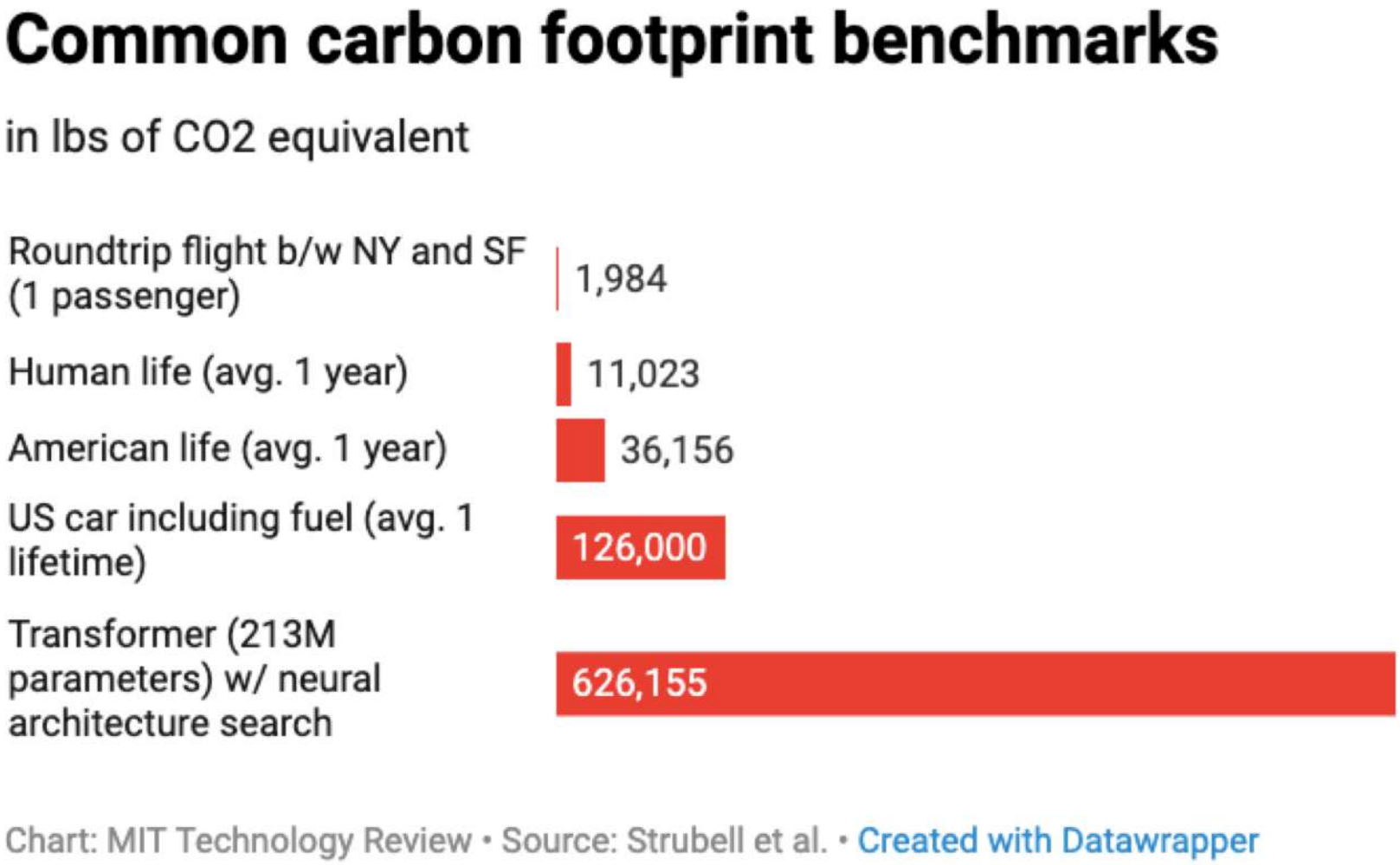
Vivienne Sze

Massachusetts Institute of Technology

In collaboration with Luca Carlone, Yu-Hsin Chen, Joel Emer, Sertac Karaman, Trevor Henderson, Amr Suleiman, Tien-Ju Yang, Zhengdong Zhang

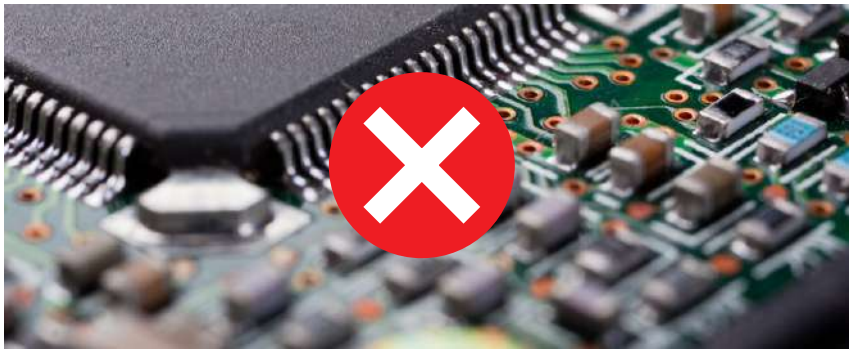
Contact Info
email: sze@mit.edu
website: www.rle.mit.edu/eems

Compute Demands for DNN



Existing Processors Consume Too Much Power

< 1 Watt

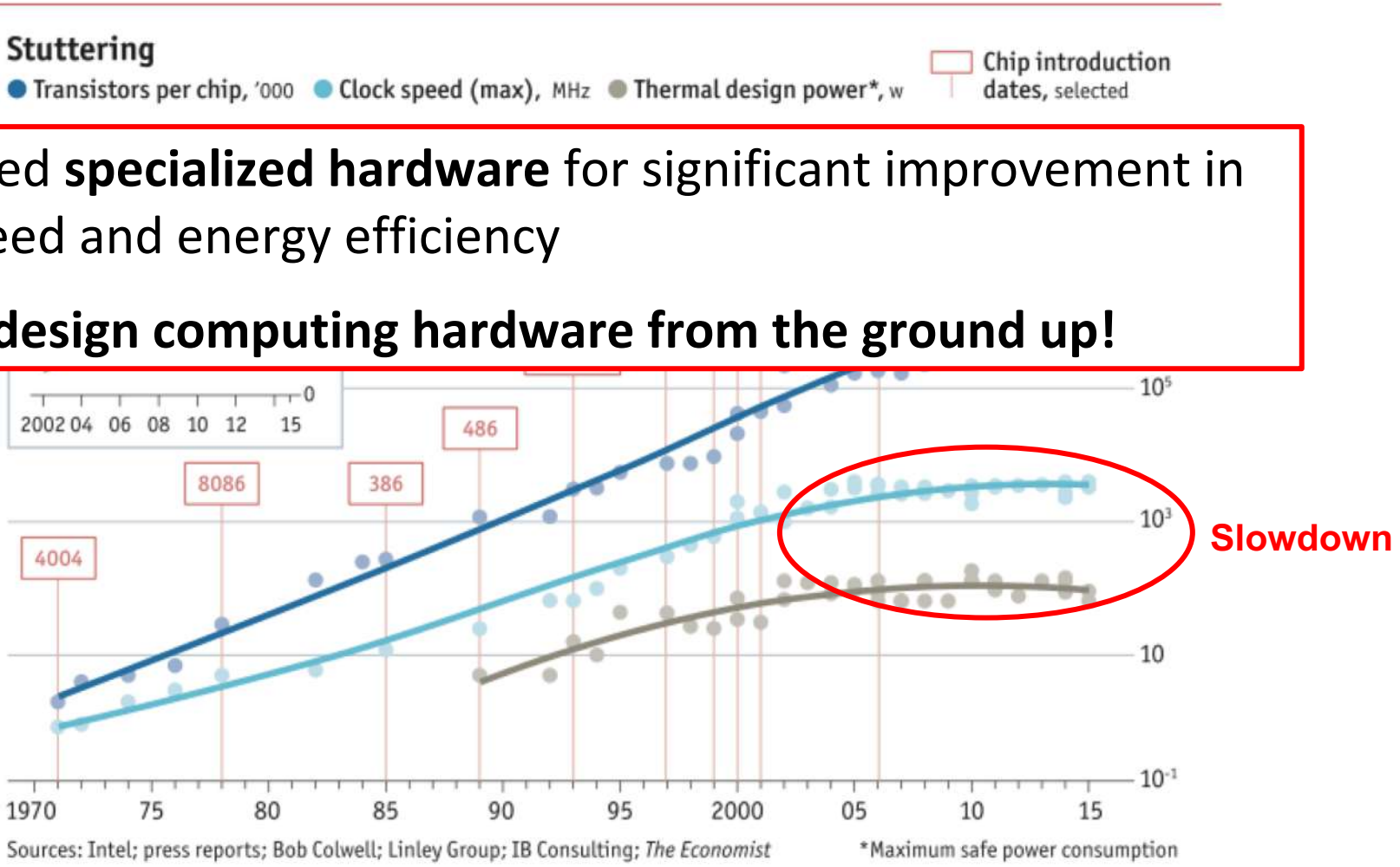


> 10 Watts

Transistors are NOT Getting More Efficient

Slow down of Moore's Law and Dennard Scaling

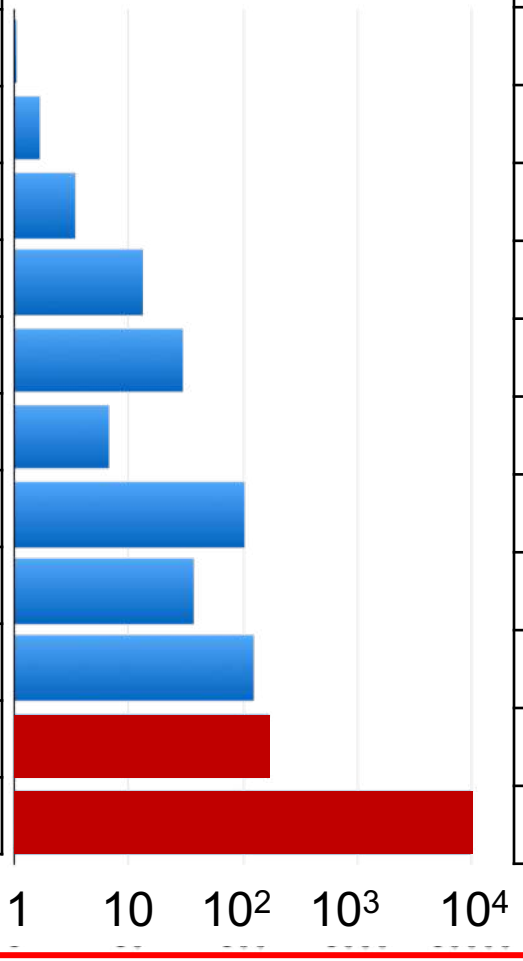
General purpose microprocessors not getting faster or more efficient

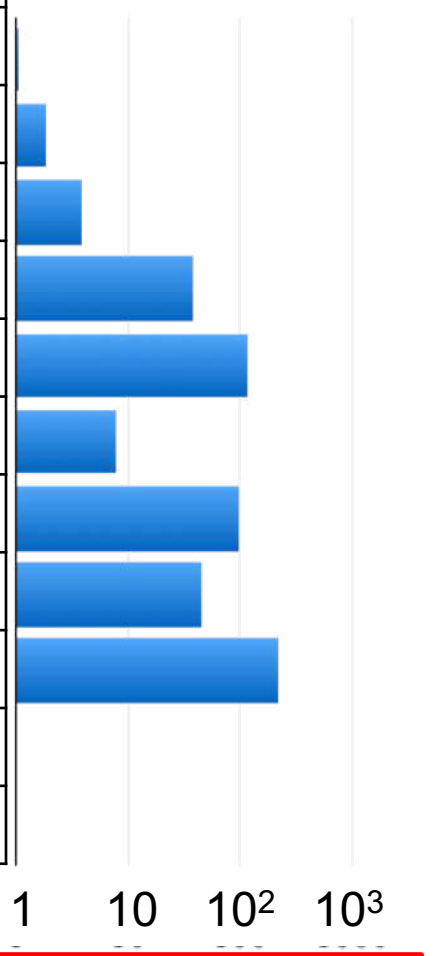


- Need **specialized hardware** for significant improvement in speed and energy efficiency
- **Redesign computing hardware from the ground up!**

Power Dominated by Data Movement

Operation:	Energy (pJ)	Area (μm^2)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A



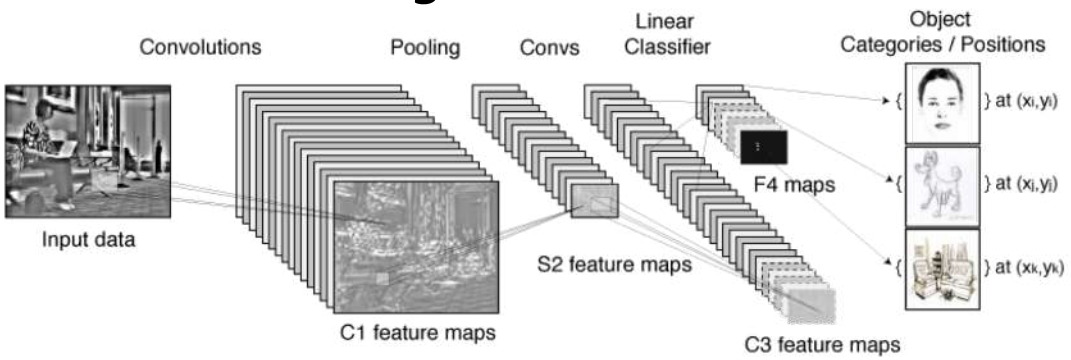


Memory access is **orders of magnitude** higher energy than compute

[Horowitz, ISSCC 2014]

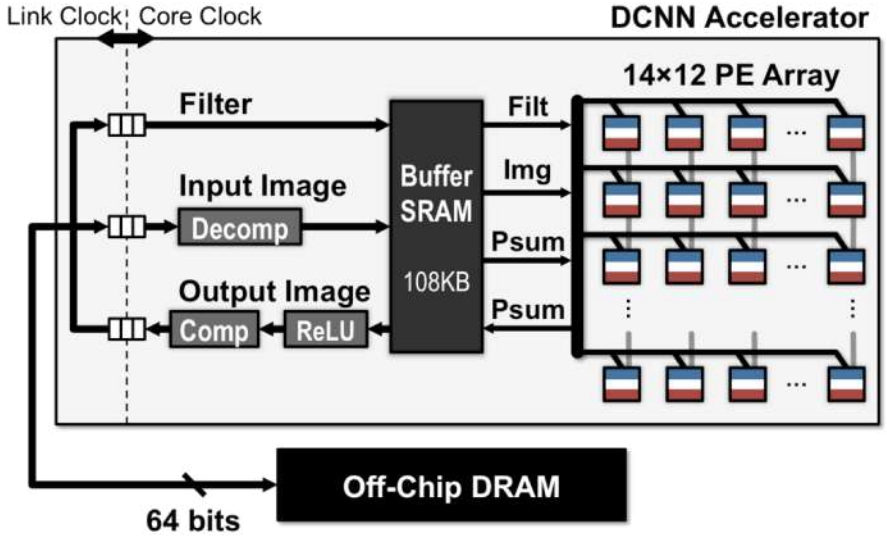
Efficient Computing with Cross-Layer Design

Algorithms

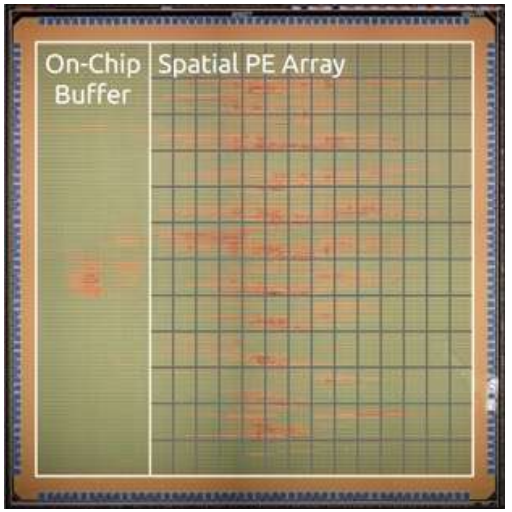


Systems

Architectures



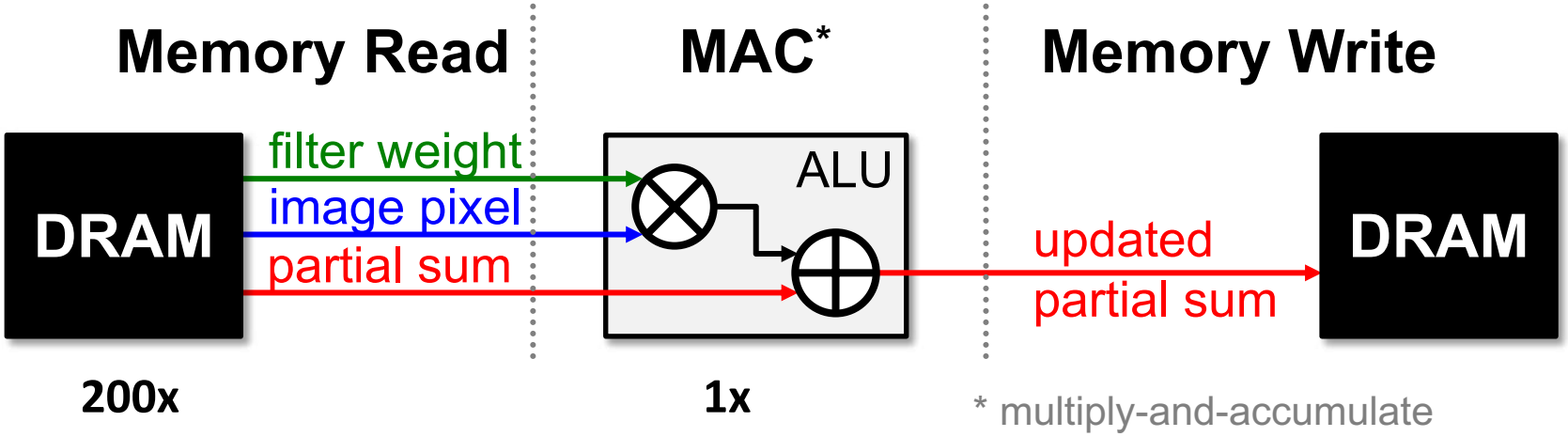
Circuits



Exploiting Reuse and Sparsity for Efficient DNN

Properties We Can Leverage

- Operations exhibit **high parallelism**
  → **high throughput** possible
- Memory Access is the Bottleneck



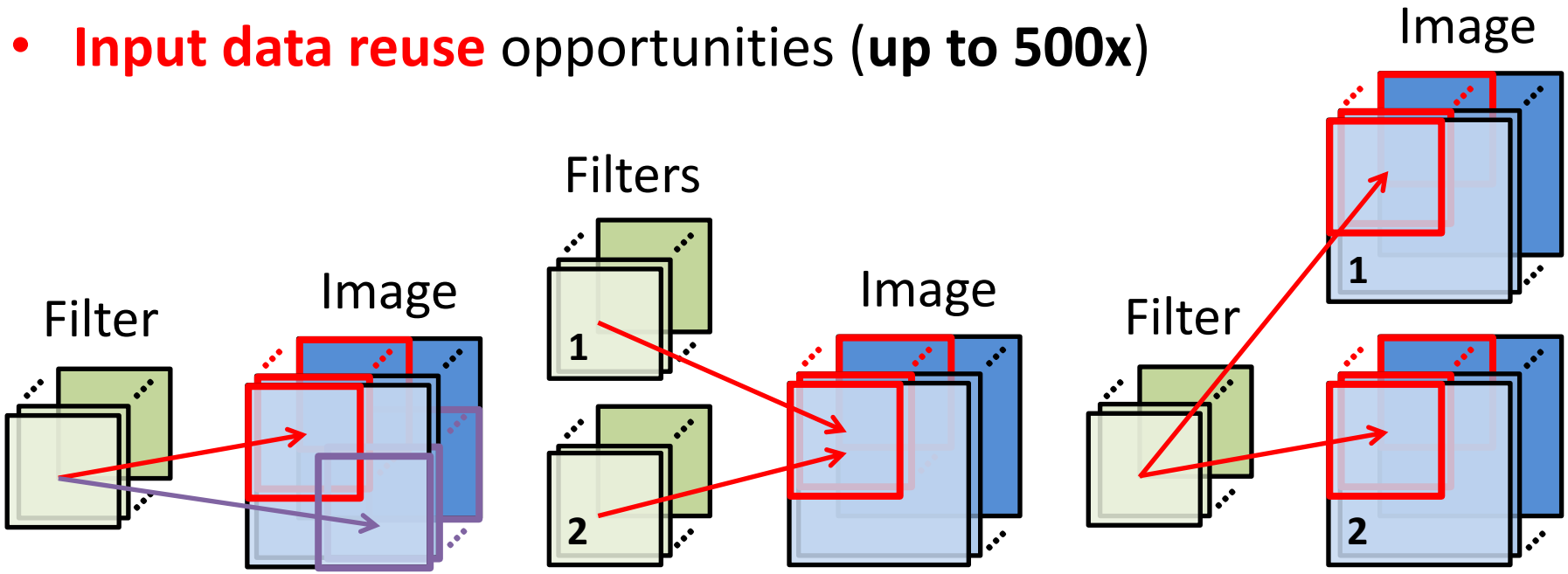
Worst Case: all memory R/W are **DRAM** accesses

- Example: AlexNet has **724M** MACs
  → **2896M** DRAM accesses required

Properties We Can Leverage

- Operations exhibit **high parallelism**
  → **high throughput** possible

- **Input data reuse** opportunities (**up to 500x**)

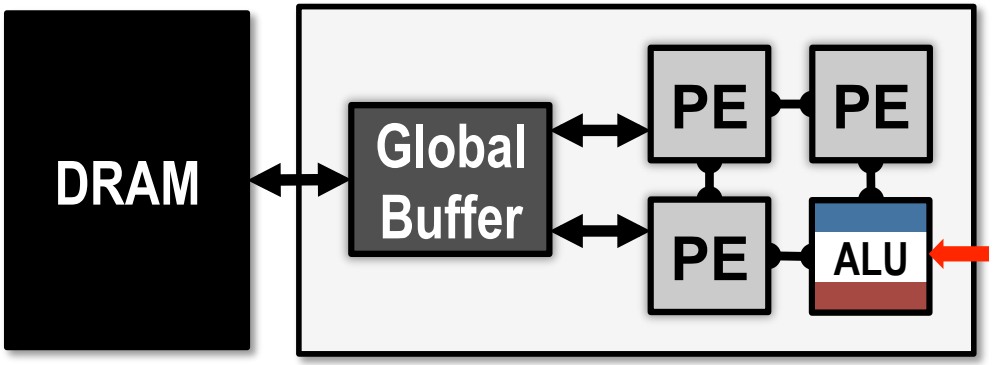


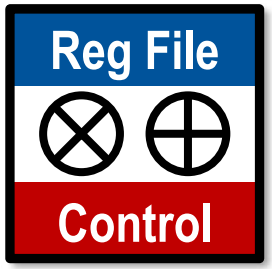
Convolutional Reuse (pixels, weights)

Image Reuse (pixels)

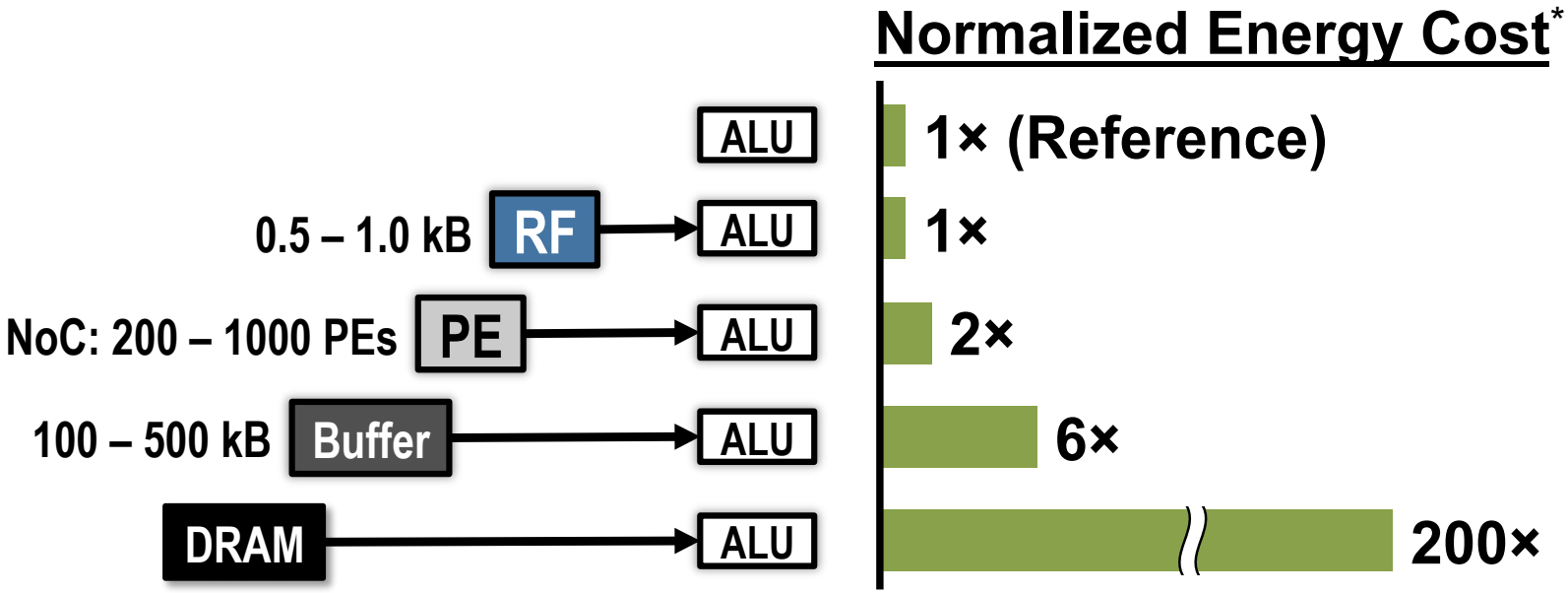
Filter Reuse (weights)

Exploit Data Reuse at Low-Cost Memories





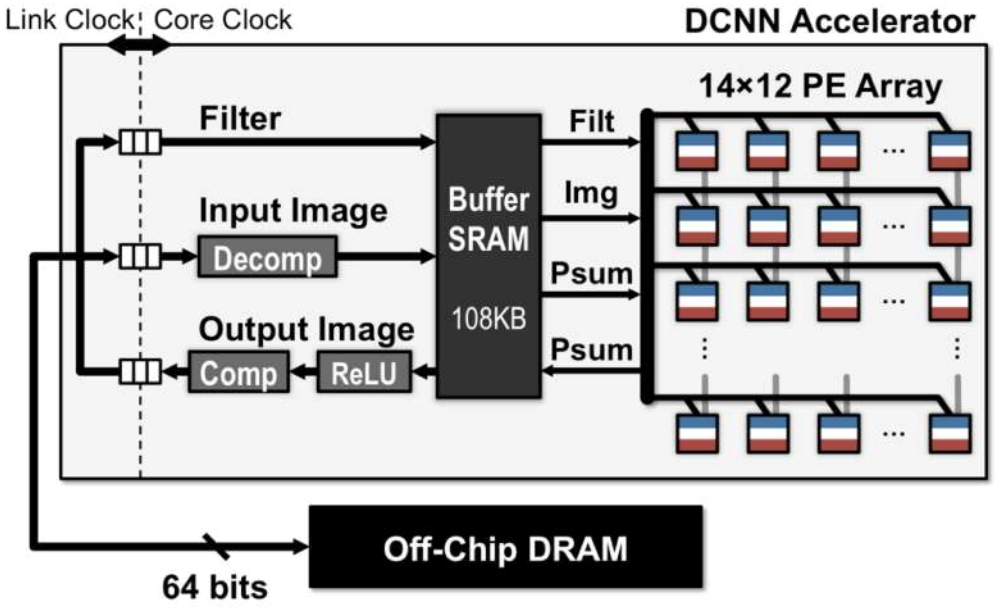
Specialized hardware with small (< 1kB) low cost memory near compute

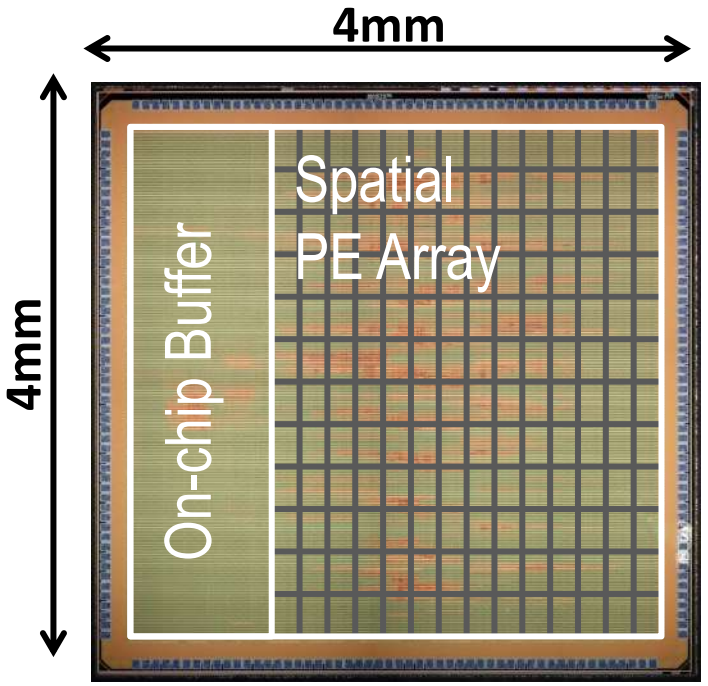


* measured from a commercial 65nm process

Farther and **larger** memories consume more power

Eyeriss: Deep Neural Network Accelerator





[Chen et al., ISSCC 2016, ISCA 2016]

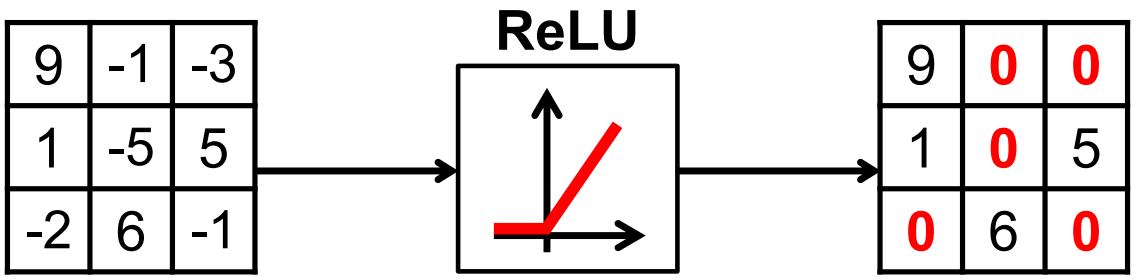
Exploits data reuse for **100x** reduction in memory accesses from global buffer and **1400x** reduction in memory accesses from off-chip DRAM

Overall **>10x energy reduction** compared to a mobile GPU (Nvidia TK1)

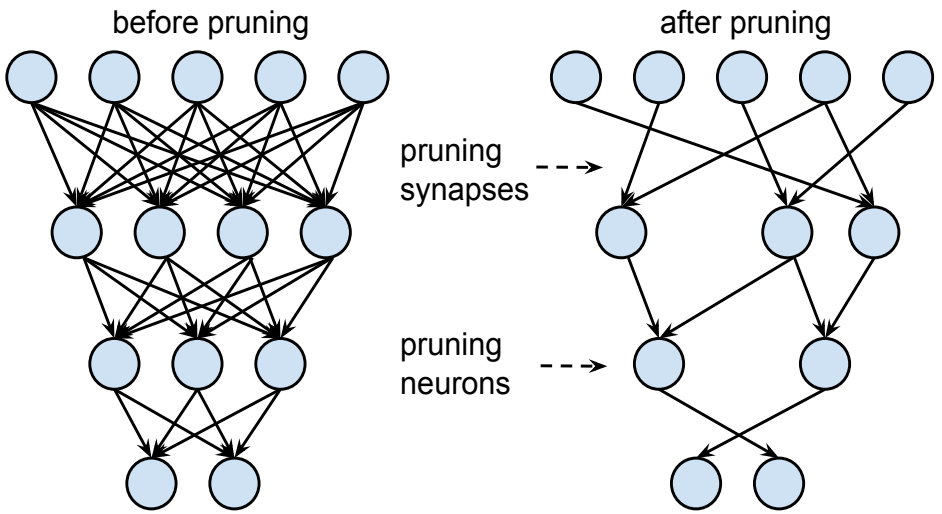
Results for AlexNet

Sparsity in DNN due to ReLU and Pruning

ReLU (activations)



Network Pruning (weights)

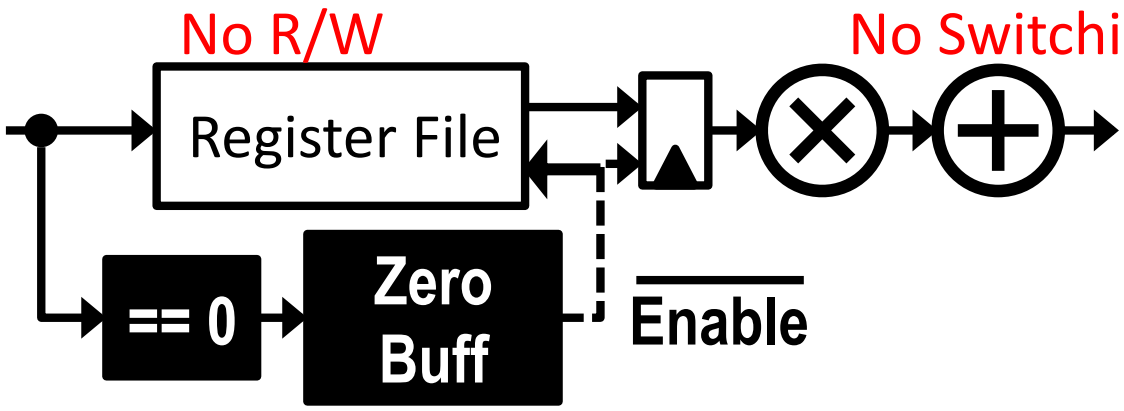


Compress sparse feature maps and weights

Exploit Sparsity for Speed and Energy

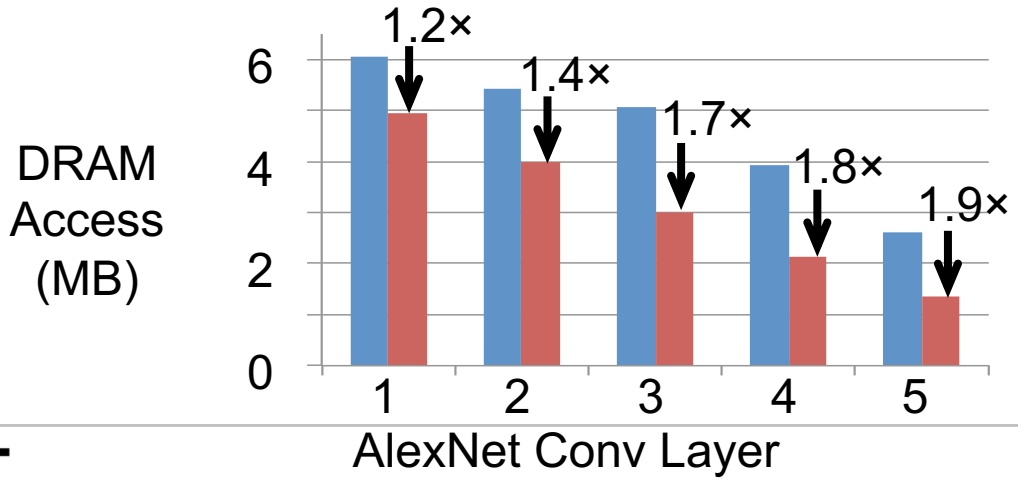
Method 1. Skip memory access and computation

(if in compressed format, can also save cycles)



45% power reduction

Method 2. Compress data to reduce storage and data movement



Run length encoding within 5-10% of entropy

Uncompressed Fmaps + Weights

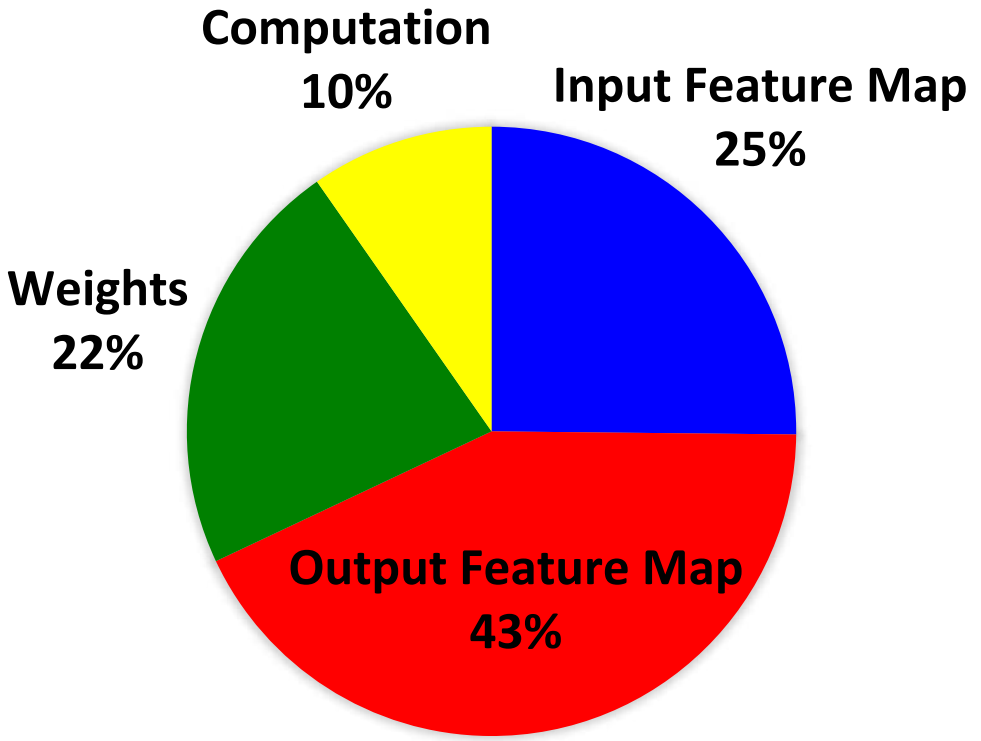
RLE Compressed Fmaps + Weights

[Chen et al., ISSCC 2016]

Maximizing Sparsity $\stackrel{?}{=}$ Minimize Energy

- Number of weights ***alone*** is not a good metric for energy
- **All data types** should be considered

Energy Consumption of GoogLeNet

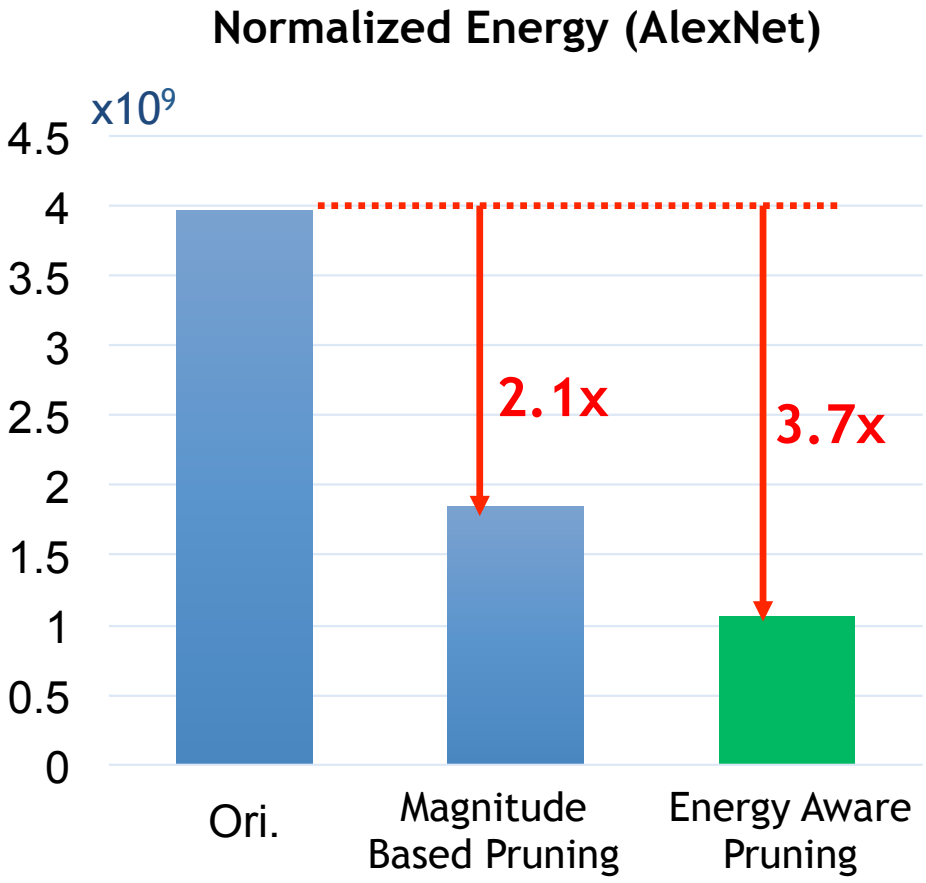


[Yang et al., CVPR 2017]

Energy-Aware Pruning

Directly target energy and incorporate it into the optimization of DNNs to provide greater energy savings

- Sort layers based on energy and prune layers that consume most energy first
- EAP reduces AlexNet energy by **3.7x** and outperforms the previous work that uses magnitude-based pruning by **1.7x**

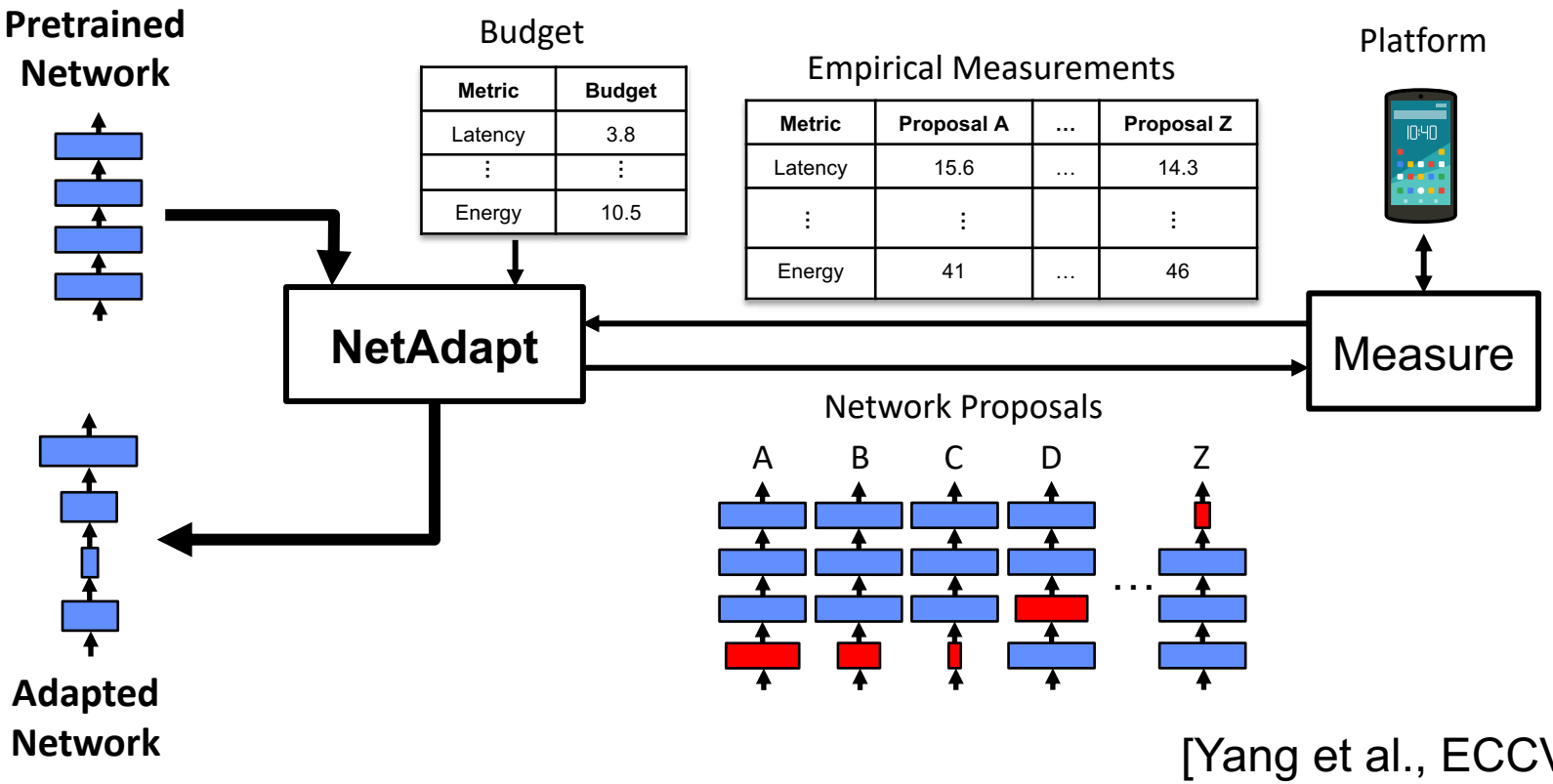


Pruned models available at
http://eyeriss.mit.edu/energy.html

[Yang et al., CVPR 2017]

NetAdapt: Platform-Aware DNN Adaptation

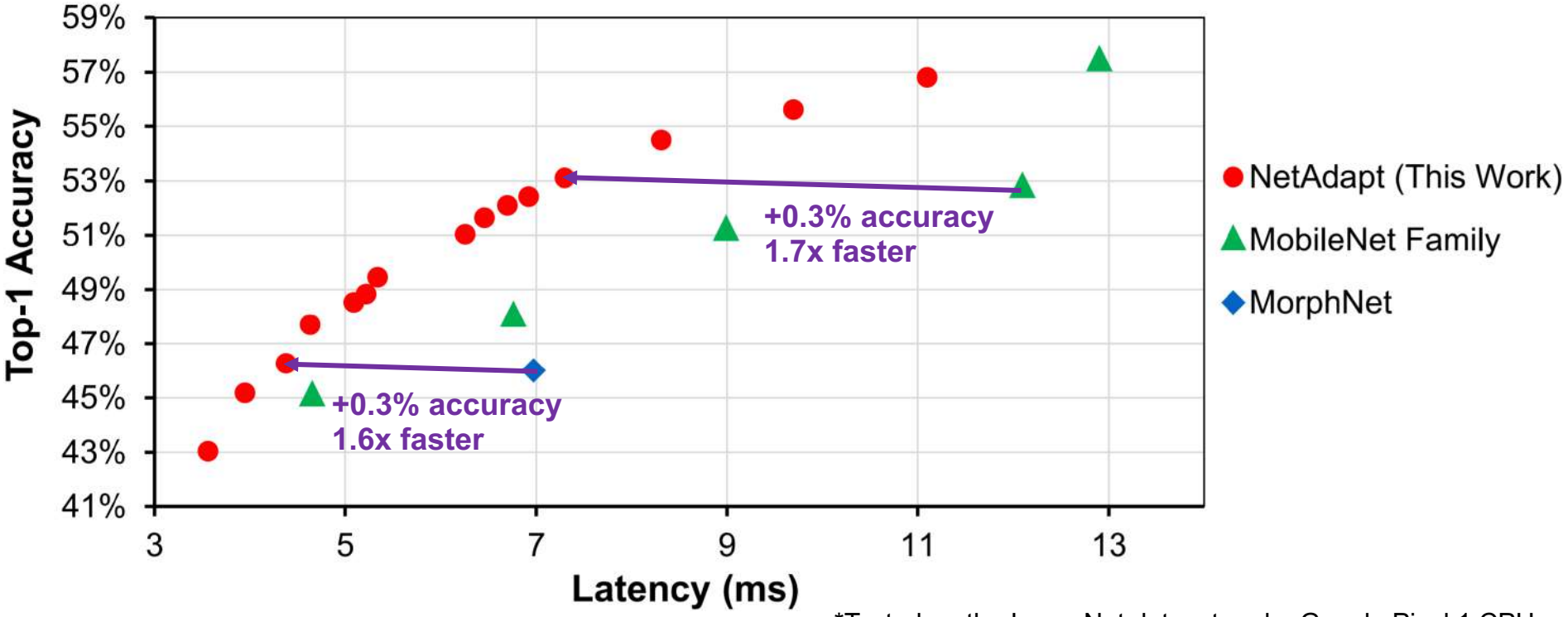
- **Automatically adapt DNN** to a mobile platform to reach a target latency or energy budget
- Use **empirical measurements** to guide optimization (avoid modeling of tool chain or platform architecture)



[Yang et al., ECCV 2018]

Improved Latency vs. Accuracy Tradeoff

- NetAdapt boosts **the real inference speed** of MobileNet by up to 1.7x with higher accuracy

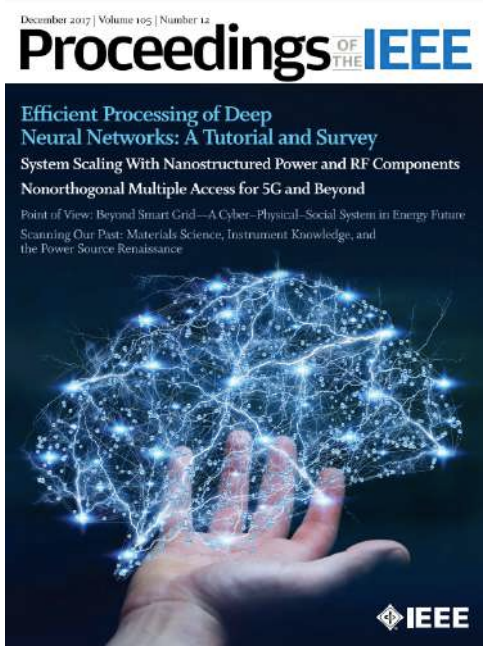


*Tested on the ImageNet dataset and a Google Pixel 1 CPU

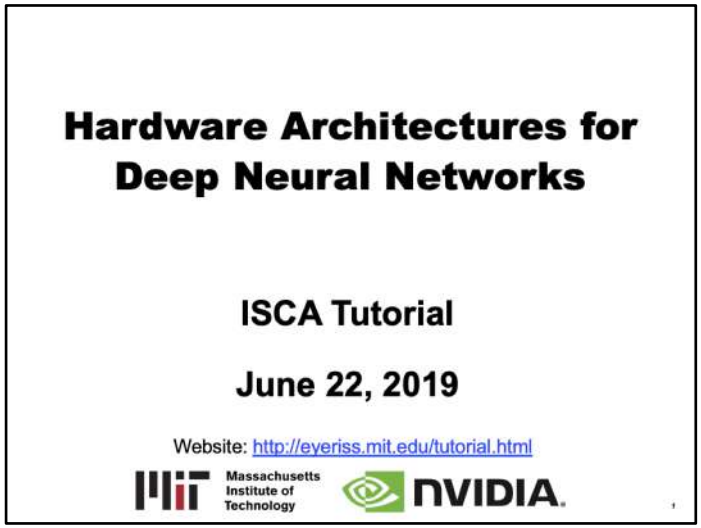
Reference:
MobileNet: Howard et al, "Mobilenets: Efficient convolutional neural networks for mobile vision applications", arXiv 2017
MorphNet: Gordon et al., "Morphnet: Fast & simple resource-constrained structure learning of deep networks", CVPR 2018

Tutorial Material on Efficient DNNs

A significant amount of algorithm and hardware research on energy-efficient processing of DNNs



V. Sze, Y.-H. Chen, T-J. Yang, J. Emer, "***Efficient Processing of Deep Neural Networks: A Tutorial and Survey***," Proceedings of the IEEE, Dec. 2017

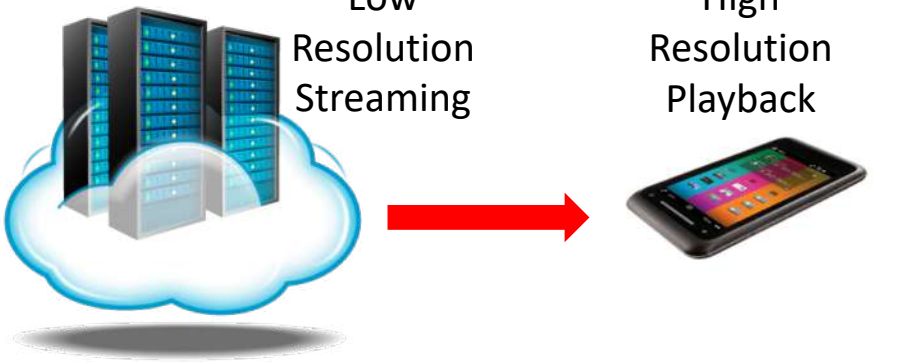


http://eyeriss.mit.edu/tutorial.html

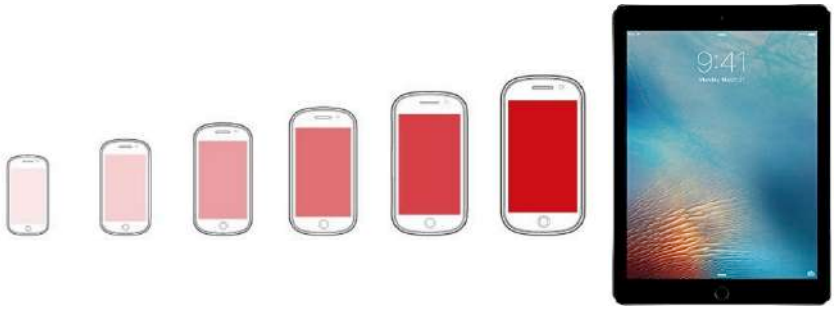
Looking Beyond the DNN Accelerator for Acceleration

Z. Zhang, V. Sze, "**FAST: A Framework to Accelerate Super-Resolution Processing on Compressed Videos**," *CVPRW* 2017

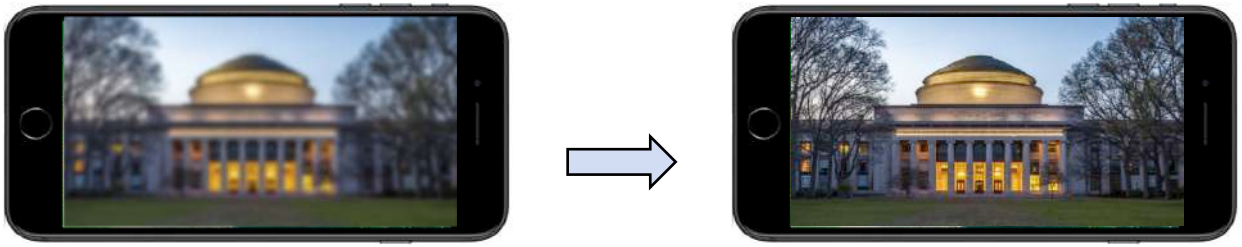
Super-Resolution on Mobile Devices



Transmit low resolution for lower bandwidth

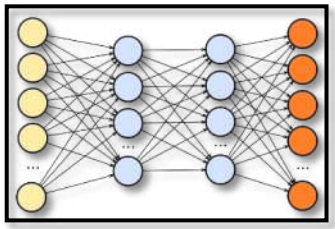


Screens are getting larger



Use **super-resolution** to improve the viewing experience of lower-resolution content (*reduce communication bandwidth*)

FAST: A Framework to Accelerate SuperRes



SR algorithm

Compressed video

FAST

SR
15x faster

Real-time

A framework that accelerates **any SR** algorithm by up to **15x** when running on compressed videos

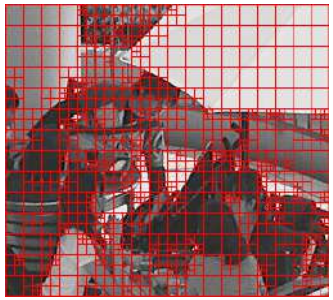
[Zhang et al., CVPRW 2017]

Free Information in Compressed Videos

Compressed video

Decode

Pixels



Block-structure

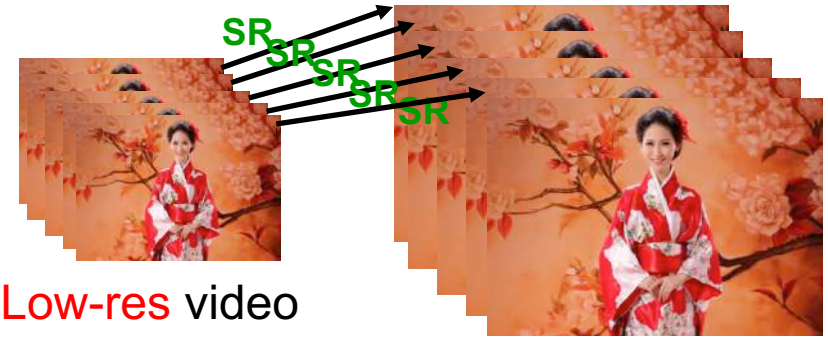
Motion-compensation

Video as a stack of pixels

Representation in compressed video

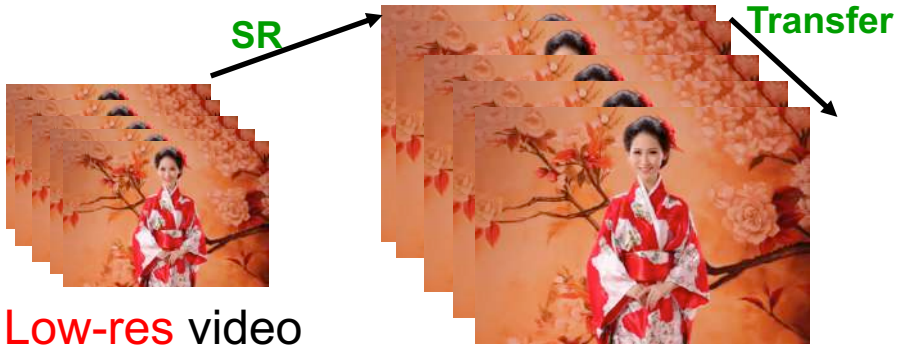
This representation can help **accelerate** super-resolution

Transfer is Lightweight



Low-res video

High-res video



Low-res video

High-res video

Transfer allows SR to run on only **a subset of frames**

Fractional Interpolation + **Bicubic Interpolation** =

Skip Flag

The complexity of the transfer is comparable to bicubic interpolation.
Transfer **N** frames, accelerate by **N**

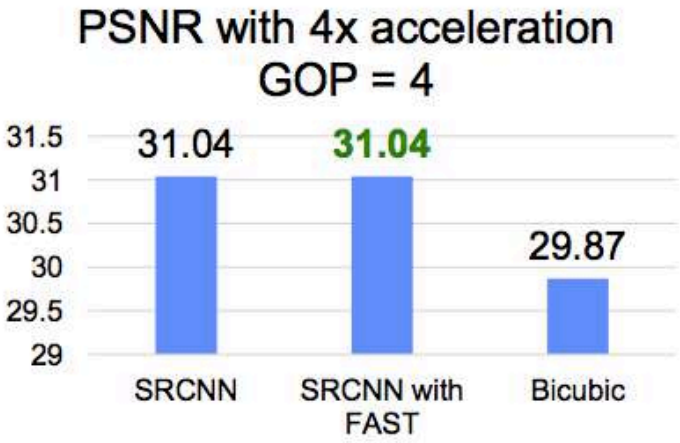
Evaluation: Accelerating SRCNN

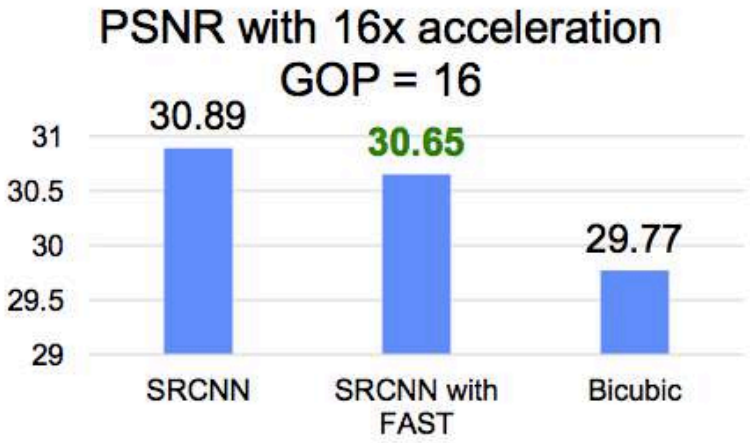
PartyScene

RaceHorse

BasketballPass

Examples of videos in the test set (20 videos for HEVC development)





4 × acceleration with NO PSNR LOSS. 16 × acceleration with 0.2 dB loss of PSNR

Visual Evaluation

SRCNN | **FAST + SRCNN** | **Bicubic**

Look ***beyond*** the DNN accelerator for opportunities to accelerate DNN processing (e.g., structure of data and temporal correlation)

Code released at www.rle.mit.edu/eems/fast

[Zhang et al., CVPRW 2017]

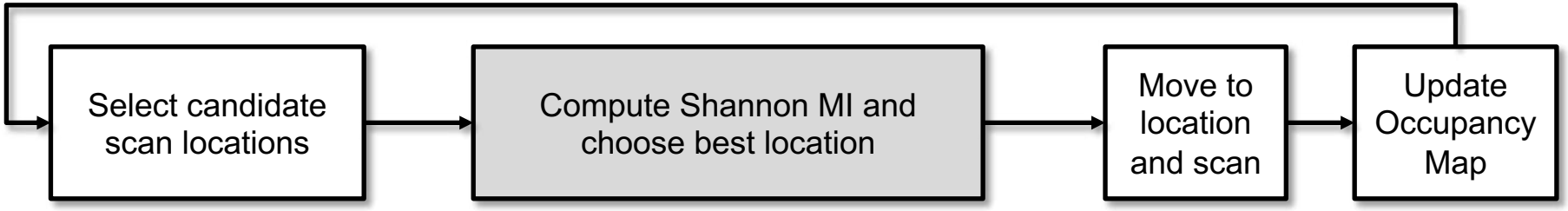
Beyond Deep Neural Networks

Z. Zhang et al., "**FSMI: Fast computation of Shannon Mutual Information for information-theoretic mapping**," *ICRA 2019*

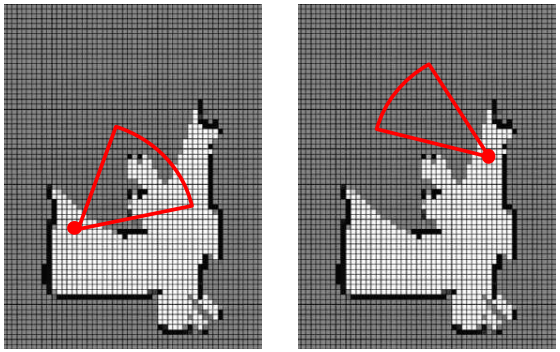
Extended version arXiv 2019 http://arxiv.org/abs/1905.02238

Where to Go Next: Planning and Mapping

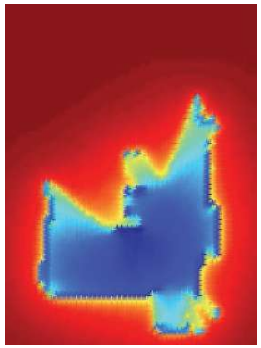
Robot Exploration: *Decide where to go by computing Shannon Mutual Information*



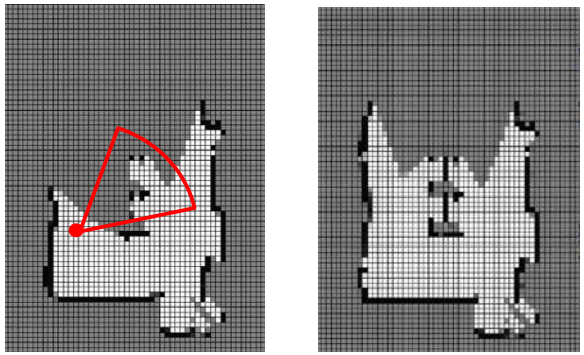
Where to scan?



Mutual Information

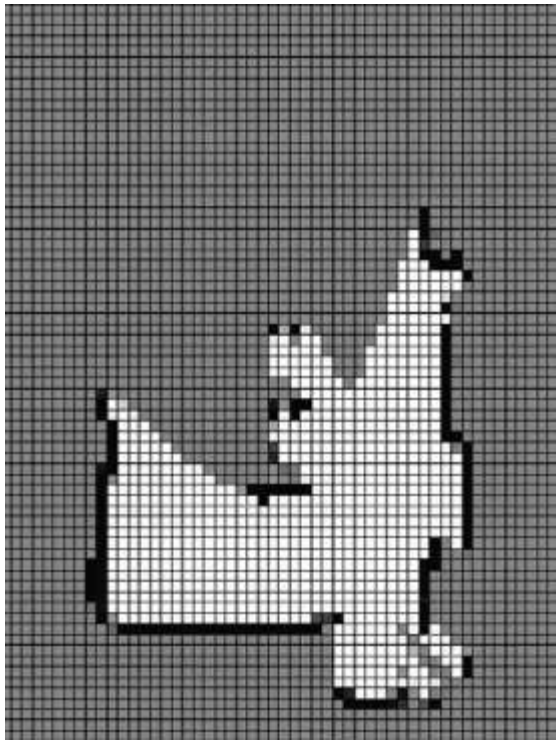


Updated Map

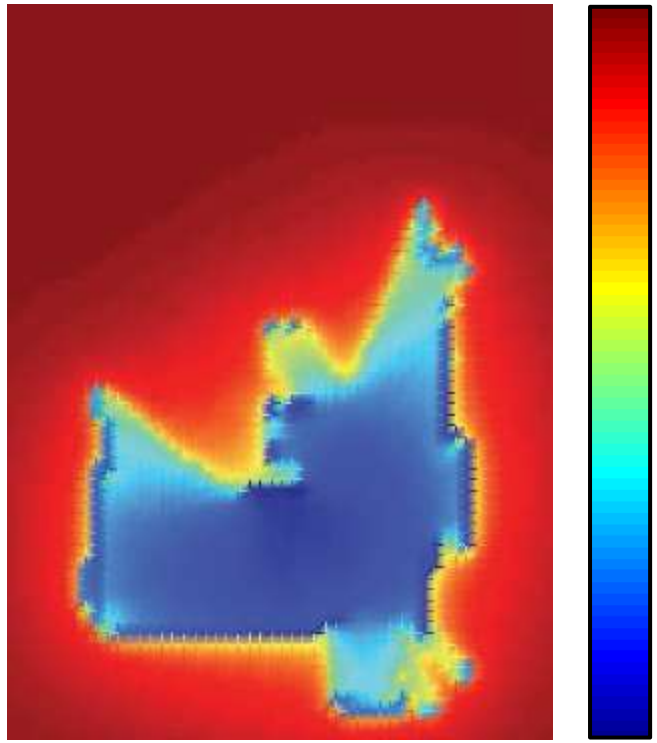


[Joint work with Sertac Karaman]

Information Theoretic Mapping



Occupancy grid map, M



Mutual information map, $I(M;Z)$

$$H(M|Z) = H(M) - I(M;Z)$$

Perspective updated map entropy

Current map entropy

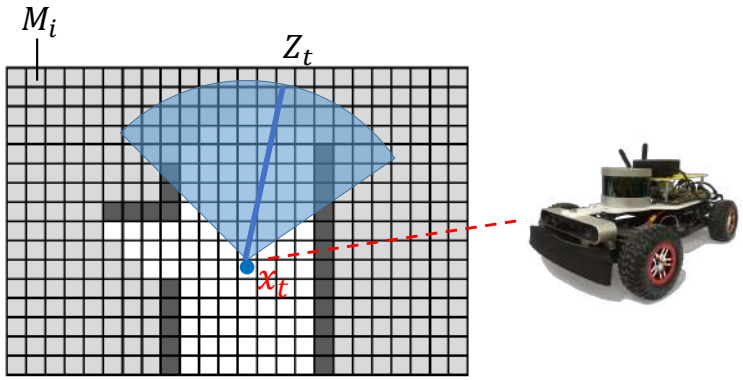
Mutual information

FSMI: Fast Shannon Mutual Information

Shannon Mutual Information
(between beam Z and map M)
[Julian et al., IJRR 2014]

$$I(M;Z) = \sum_{i=1}^{n} \int_{z \geq 0} P(z) f(\delta_i(z), r_i) dz$$

No closed form solution. Requires expensive **numerical integration at resolution** λ_z. $\boldsymbol{O(n^2 \lambda_z)}$



FSMI: Fast Shannon Mutual Information

$$I(M;Z) = \sum_{j=1}^{n} \sum_{k=1}^{n} P(e_j) C_k G_{k,j}$$

Evaluate MI for all cells in entire beam altogether **removes numerical integration**. $\boldsymbol{O(n^2)}$

Approximate FSMI

$$I(M;Z) = \sum_{j=1}^{n} \sum_{k=j-\Delta}^{j+\Delta} P(e_j) C_k G_{k,j}$$

Approximate noise model of depth sensor with **truncated Gaussian***. $\boldsymbol{O(n)}$

*Charrow et al., ICRA 2015

Experimental Results (4x Real Time)

Occupancy map with planned path using RRT* (compute MI on all possible paths)

MI surface

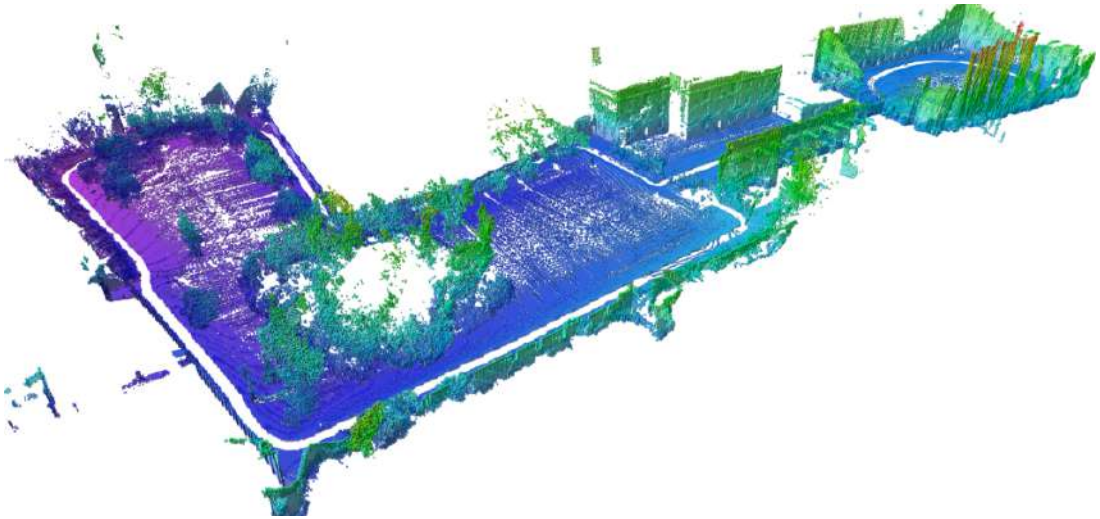
Exploration with a mini race car using motion capture for localization

Approximate FSMI is **over 1000x faster** than original MI
and **1.7 – 2.8x faster** than previous state-of-the-art methods (e.g., CSQMI)

[Zhang et al., ICRA 2019]

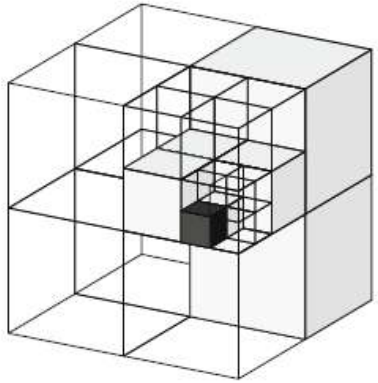
Computing MI in 3D Environments

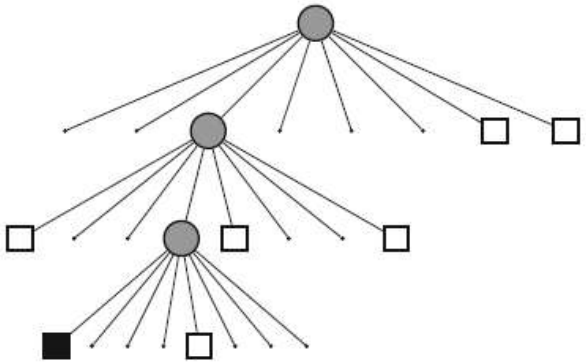
Computing MI on a **3D map** requires significant amounts of storage and compute



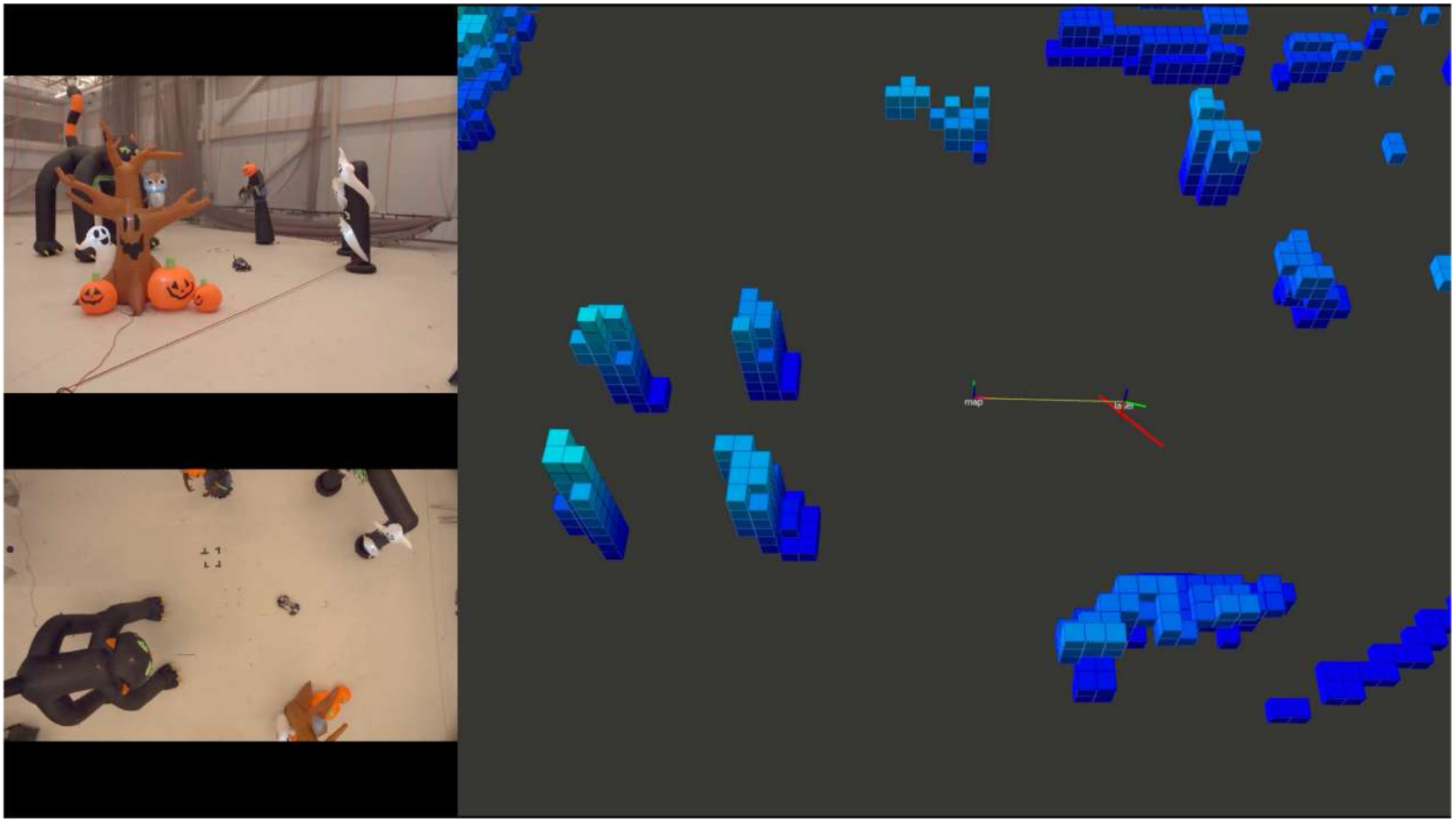
Compress map with OctoMap

[Hornung, et al., Autonomous Robots, 2013]





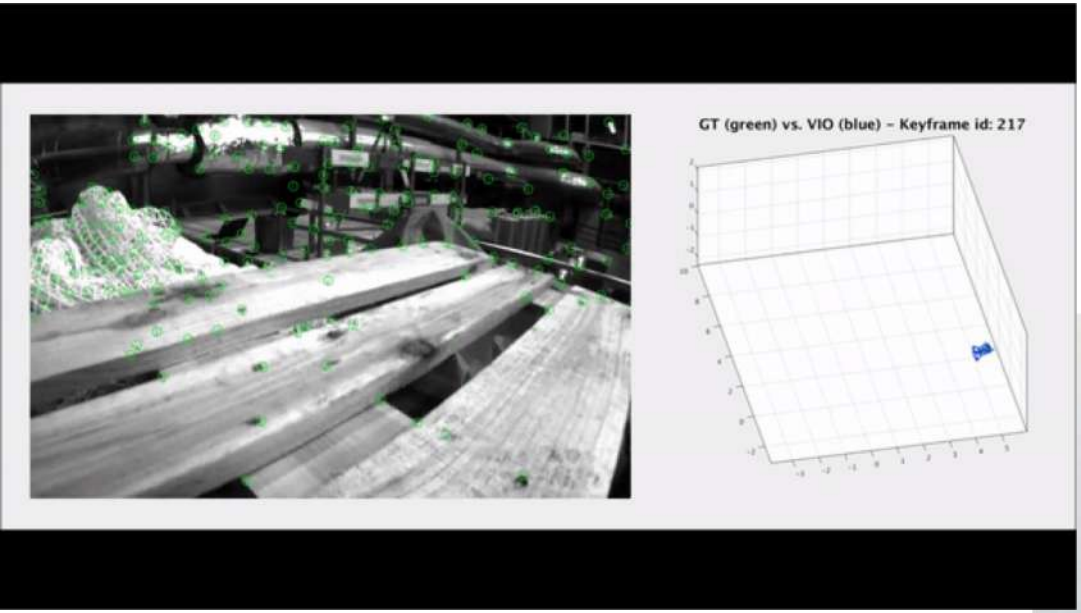
Experiments of 3D FSMI (4x Real Time)



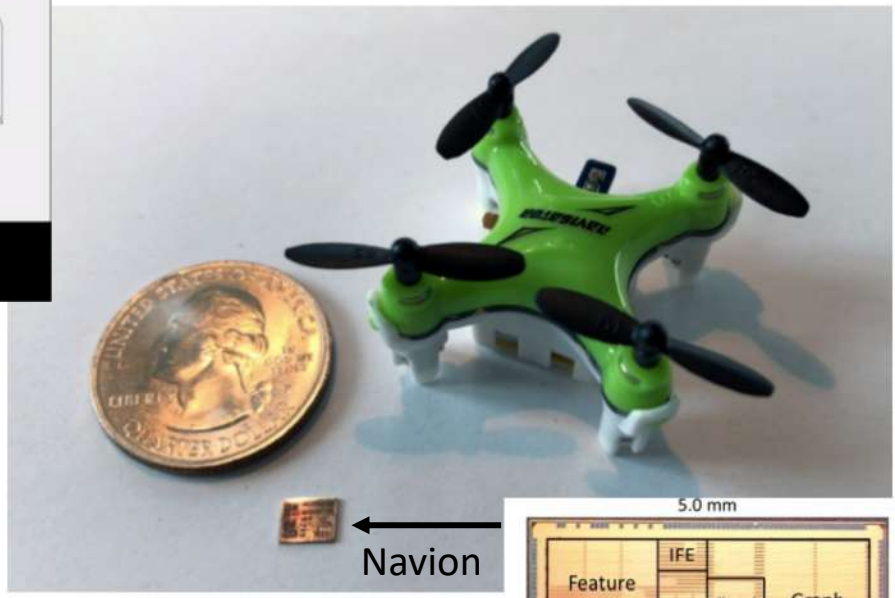
We achieve an average compression ratio of around $18\times$, with an acceleration ratio of $8\times$

Z. Zhang et al., FSMI: Fast computation of Shannon Mutual Information for information-theoretic mapping, *arXiv 2019* http://arxiv.org/abs/1905.02238

Localization on Chip



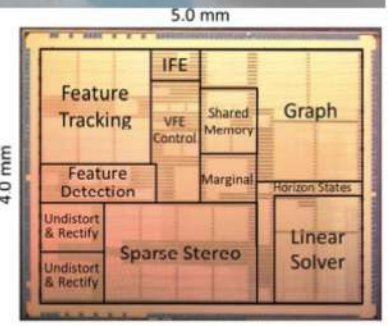
Localization is a key step in autonomous navigation (also AR and VR)



Entire system fully integrated on chip. Use compression/sparsity to reduce storage to under 1MB

Consumes **684× and 1582×** less energy than mobile and desktop CPUs, respectively

Technology	65nm CMOS	Supply	1 V
Chip area (mm²)	4.0 x 5.0	Resolution	752x480
Core area (mm²)	3.54 x 4.54	Camera rate	28 - 171 fps
Logic gates	2,043 kgates	Keyframe rate	16 - 90 fps
SRAM	854KB	Average Power	24 mW
VFE Frequency	62.5 MHz	GOPS	10.5 – 59.1
BE Frequency	83.3 MHz	GFLOPS	1 – 5.7



http://navion.mit.edu

[Zhang et al., RSS 2017], [Suleiman et al., VLSI 2018]

[Joint work with Sertac Karaman]

Summary of Key Insights

- **Data movement dominates energy consumption**
  - Exploit data reuse to reduce data movement cost
- **Design considerations for co-design of algorithm and hardware**
  - Incorporate ***direct metrics*** into algorithm design for improved efficiency
- **Accelerate deep learning by looking beyond the accelerator**
  - Exploit data representation for FAST Super-Resolution
- **Processing in compressed domain can accelerate many applications**

Acknowledgements

Joel Emer

Sertac Karaman

Thomas Heldt

Research conducted in the **MIT Energy-Efficient Multimedia Systems Group** would not be possible without the support of the following organizations:

AFOSR AIR FORCE OFFICE OF SCIENTIFIC RESEARCH · NSF · ANALOG DEVICES · BROADCOM · Google · intel · IBM · DARPA · SRC · 3M · NVIDIA · QUALCOMM · SAMSUNG · Texas Instruments · tsmc

References

- **Efficient Processing for Deep Neural Networks**
  - **Project website:** *http://eyeriss.mit.edu*
  - *Y.-H. Chen, T. Krishna, J. Emer, V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," IEEE Journal of Solid State Circuits (JSSC), ISSCC Special Issue, Vol. 52, No. 1, pp. 127-138, January 2017.*
  - *Y.-H. Chen, J. Emer, V. Sze, "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," International Symposium on Computer Architecture (ISCA), pp. 367-379, June 2016.*
  - *V. Sze, Y.-H. Chen, T.-J. Yang, J. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, December 2017.*
  - *Hardware Architecture for Deep Neural Networks: http://eyeriss.mit.edu/tutorial.html*
- **Co-Design of Algorithms and Hardware for Deep Neural Networks**
  - *T.-J. Yang, Y.-H. Chen, V. Sze, "Designing Energy-Efficient Convolutional Neural Networks using Energy-Aware Pruning," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.*
  - *Energy estimation tool:* http://eyeriss.mit.edu/energy.html
  - *T.-J. Yang, A. Howard, B. Chen, X. Zhang, A. Go, V. Sze, H. Adam, "NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications," European Conference on Computer Vision (ECCV), 2018.*

References

- **Fast Shannon Mutual Information for Robot Exploration**
  - *Z. Zhang, T. Henderson, V. Sze, S. Karaman, "FSMI: Fast computation of Shannon Mutual Information for information-theoretic mapping," IEEE International Conference on Robotics and Automation (ICRA), May 2019.*
  - *P. Li*, Z. Zhang*, S. Karaman, V. Sze, "High-throughput Computation of Shannon Mutual Information on Chip," Robotics: Science and Systems (RSS), June 2019*
  - *Z. Zhang, T. Henderson, S. Karaman, V. Sze, "FSMI: Fast computation of Shannon Mutual Information for information-theoretic mapping," extended preprint on arXiv, May 2019*
    http://arxiv.org/abs/1905.02238
- **Energy-Efficient Visual Inertial Localization**
  - **Project website:** *http://navion.mit.edu*
  - *A. Suleiman, Z. Zhang, L. Carlone, S. Karaman, V. Sze, "Navion: A Fully Integrated Energy-Efficient Visual-Inertial Odometry Accelerator for Autonomous Navigation of Nano Drones," IEEE Symposium on VLSI Circuits (VLSI-Circuits), June 2018.*
  - *Z. Zhang*, A. Suleiman*, L. Carlone, V. Sze, S. Karaman, "Visual-Inertial Odometry on Chip: An Algorithm-and-Hardware Co-design Approach," Robotics: Science and Systems (RSS), July 2017.*
  - *A. Suleiman, Z. Zhang, L. Carlone, S. Karaman, V. Sze, "Navion: A 2mW Fully Integrated Real-Time Visual-Inertial Odometry Accelerator for Autonomous Navigation of Nano Drones," IEEE Journal of Solid State Circuits (JSSC), VLSI Symposia Special Issue, Vol. 54, No. 4, pp. 1106-1119, April 2019.*