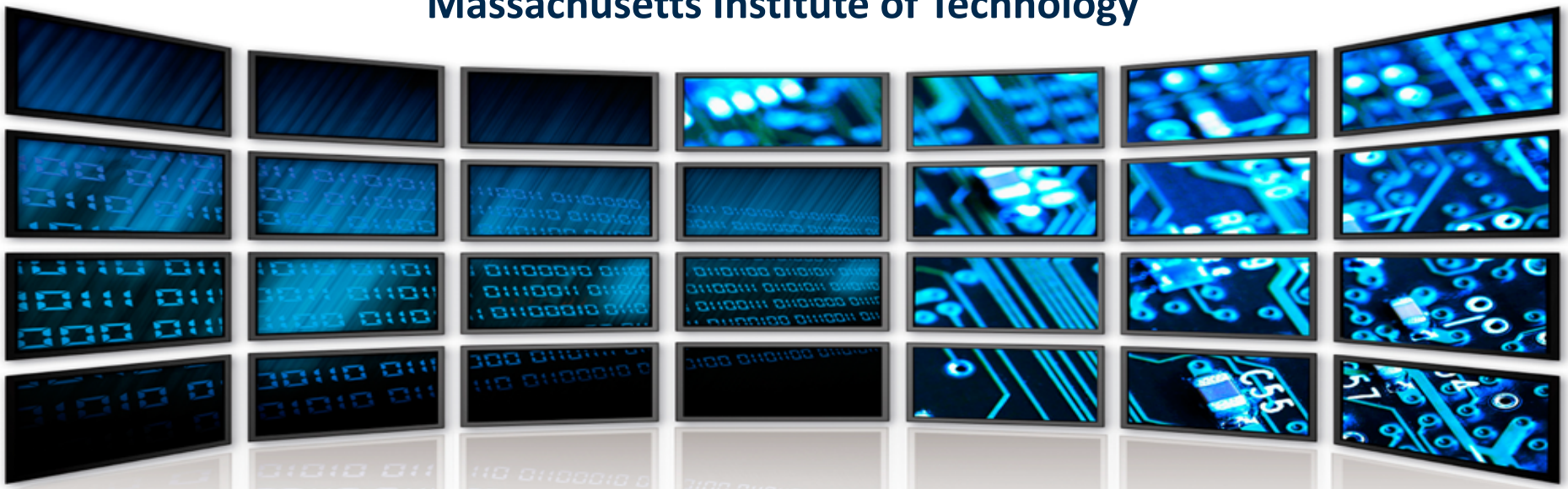
Domain-Specific Architectures for AI and Robotics: Opportunities and Challenges

Vivienne Sze

Massachusetts Institute of Technology

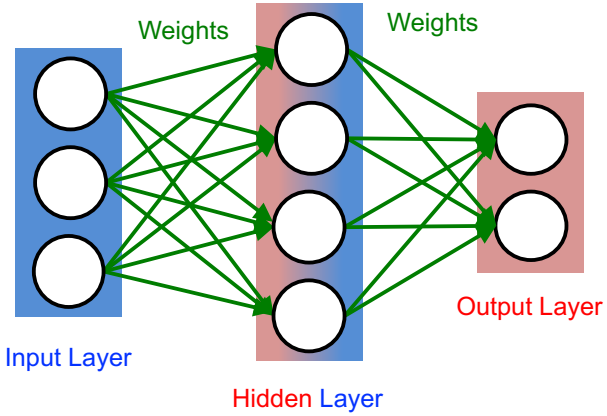


Contact Info
email: sze@mit.edu
website: www.rle.mit.edu/eems

Wide Range of Compute-Intensive Applications

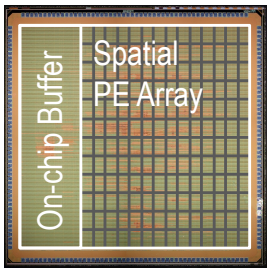
Video Compression

AI: Deep Neural Networks

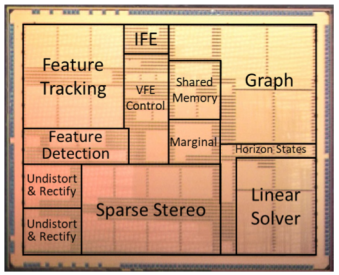


Robotics: Autonomous Navigation

Parallel CABAC [ISSCC 2011]



Eyeriss [ISSCC 2016]



Navion [VLSI 2018]

- *Rapidly growing volume of data to be processed*
- *Increasingly complex algorithms for higher quality of result*
- *Require high throughput/low latency and energy efficiency*

Need Domain Specific Architectures → 2 to 3 years to design!

Key Design Considerations

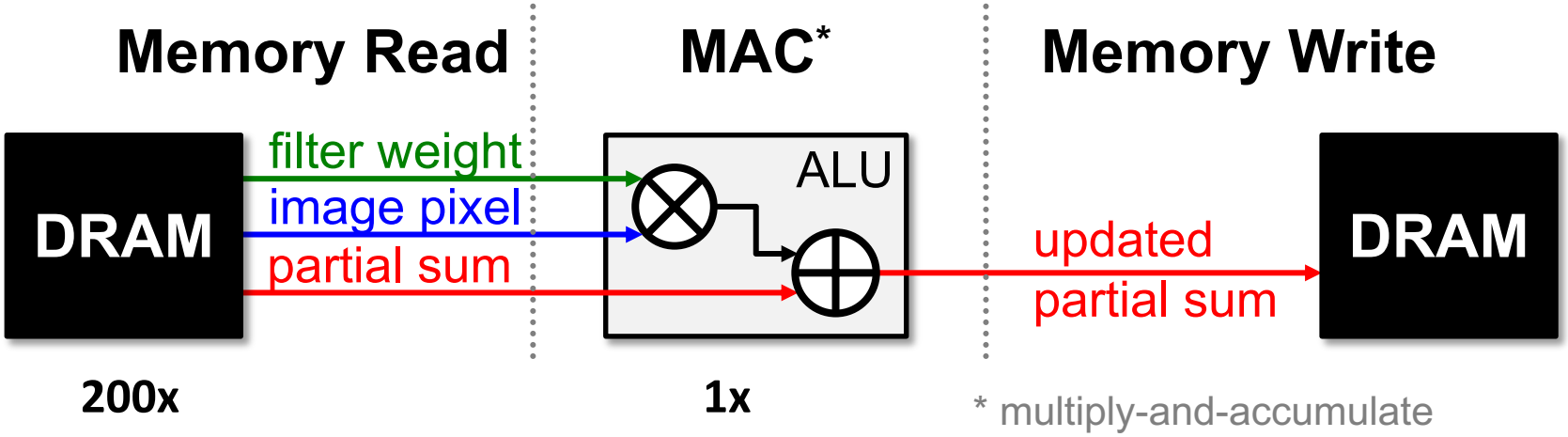
- **Exploit properties of workloads**
  - Specialized hardware translates parallelism, data access patterns and representation into increased throughput and energy efficiency
- **Design more efficient workloads**
  - Co-design of algorithms and hardware without affecting quality of result
- **Define range of workloads**
  - Balance flexibility and efficiency depending on application requirements

How can Agile and Open Hardware help accelerate the design process?

Deep Neural Networks (DNN)

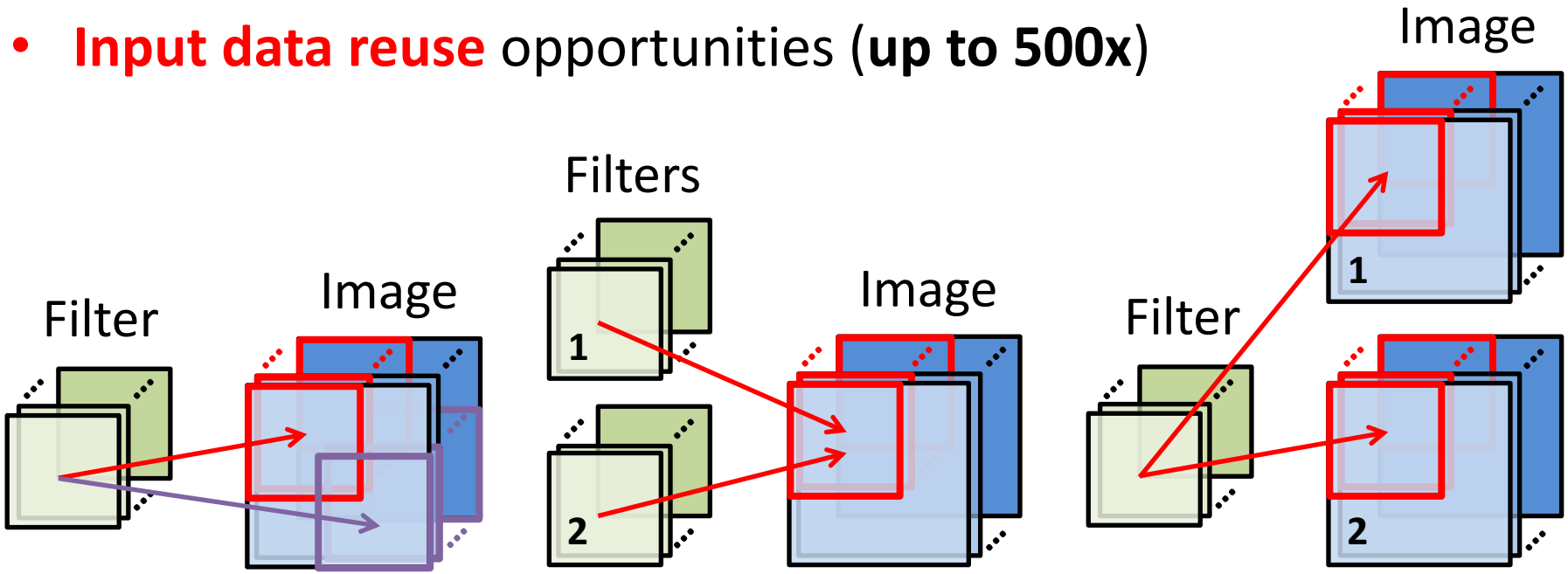
Properties We Can Leverage

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



Properties We Can Leverage

- Operations exhibit **high parallelism**
  → **high throughput** possible
- **Input data reuse** opportunities (**up to 500x**)

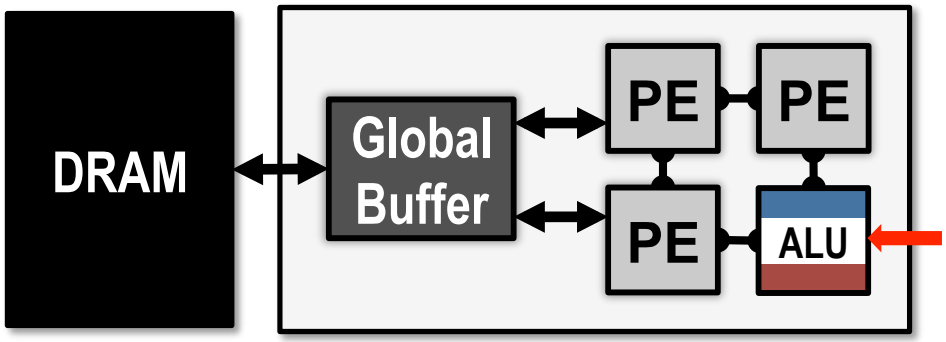


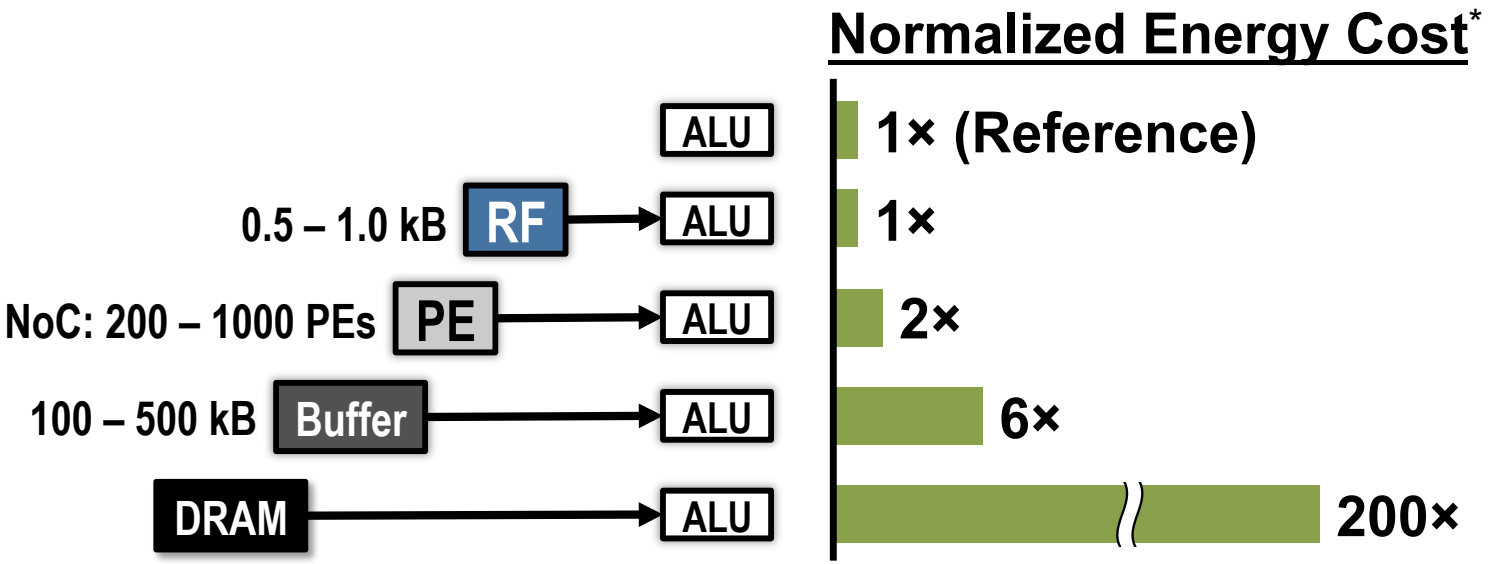
Convolutional Reuse (pixels, weights)

Image Reuse (pixels)

Filter Reuse (weights)

Data Movement is Expensive

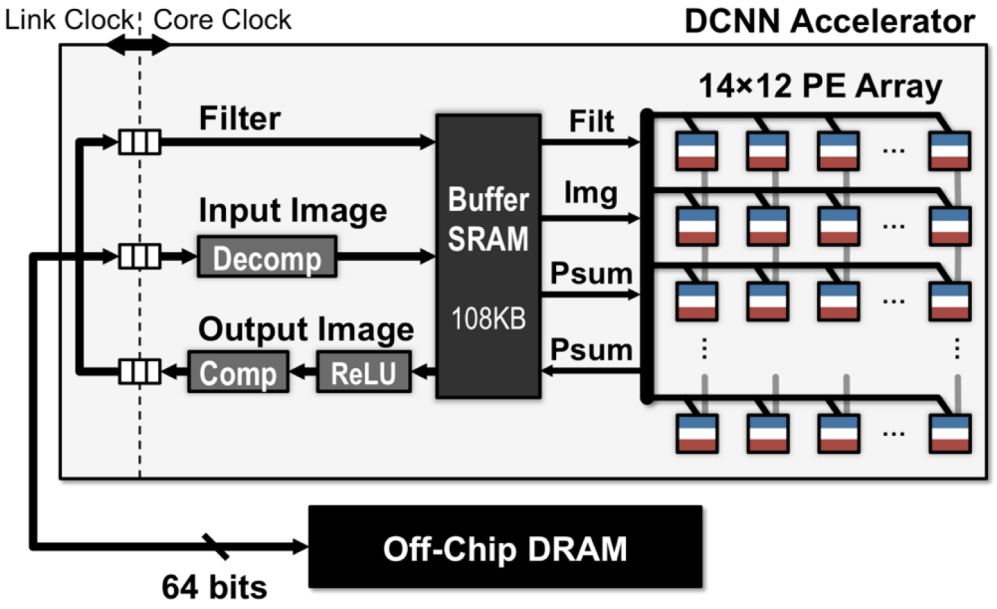


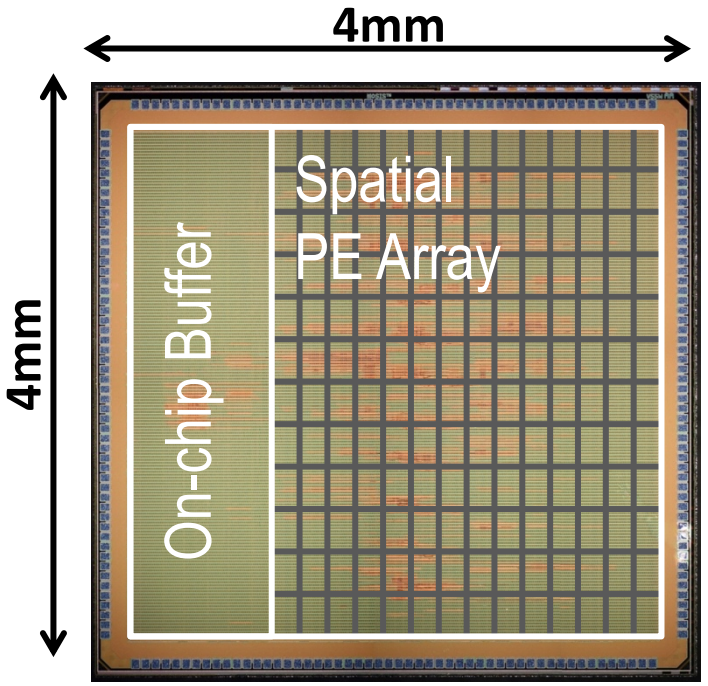


* measured from a commercial 65nm process

Design memory hierarchy and dataflow to **exploit data reuse at low cost memories**

Eyeriss: Deep Neural Network Accelerator



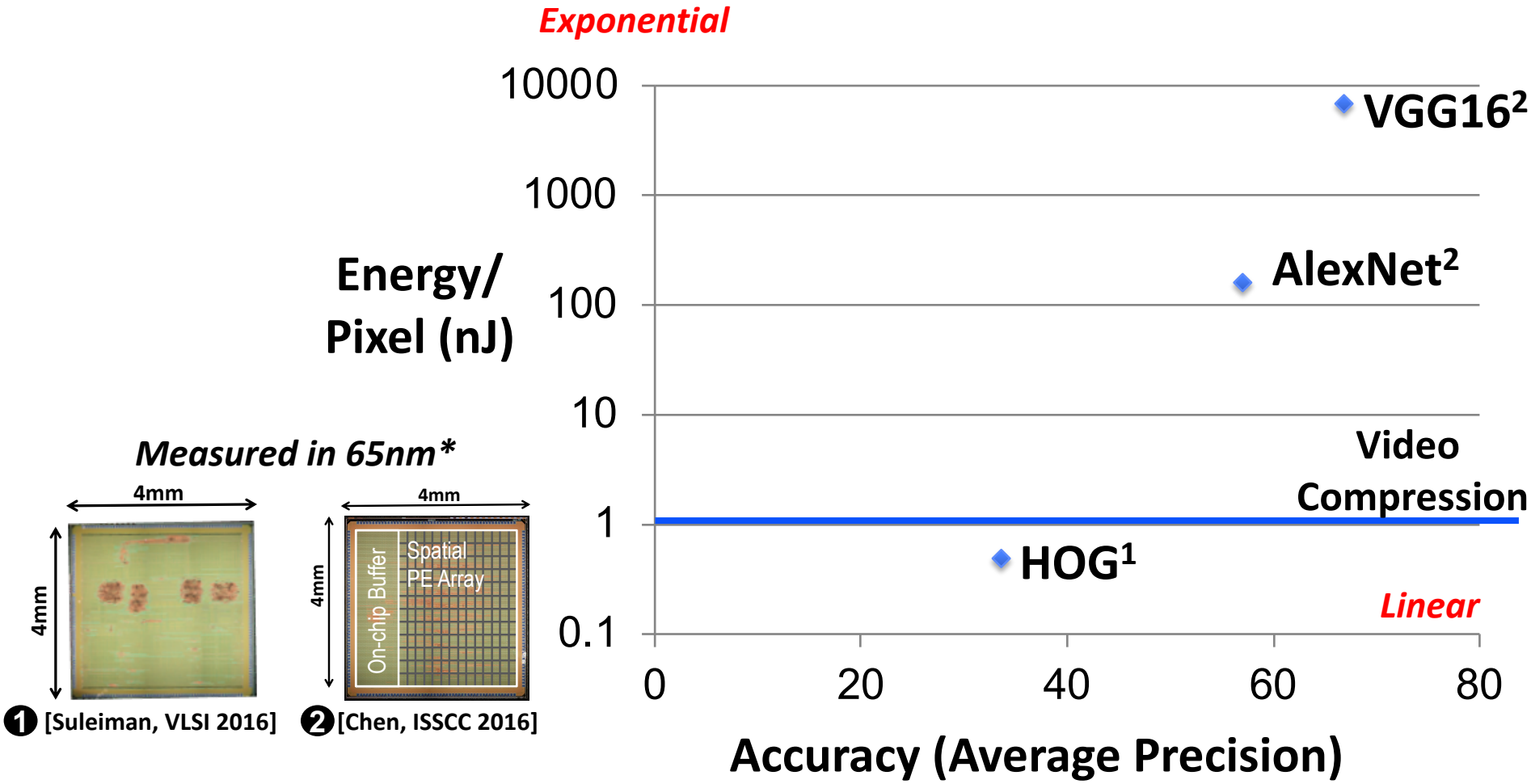


[Chen et al., ISSCC 2016, ISCA 2016]

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

Results for AlexNet

Features: Energy vs. Accuracy



** Only feature extraction. Does not include data, classification energy, augmentation and ensemble, etc.*

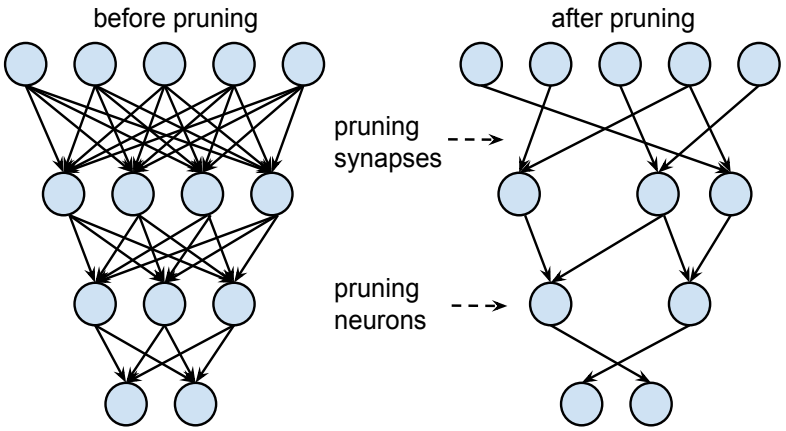
Measured in on VOC 2007 Dataset

1. DPM v5 [Girshick, 2012]
2. Fast R-CNN [Girshick, CVPR 2015]

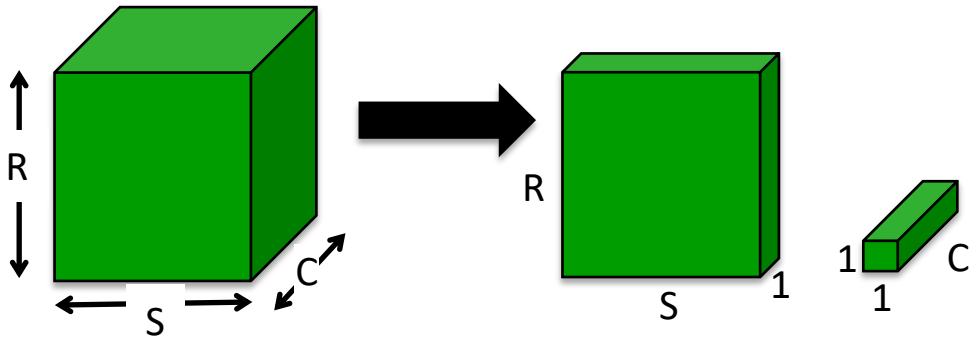
Design of Efficient DNN Algorithms

- Popular efficient DNN algorithm approaches

Network Pruning



Compact Network Architectures



Examples: SqueezeNet, MobileNet

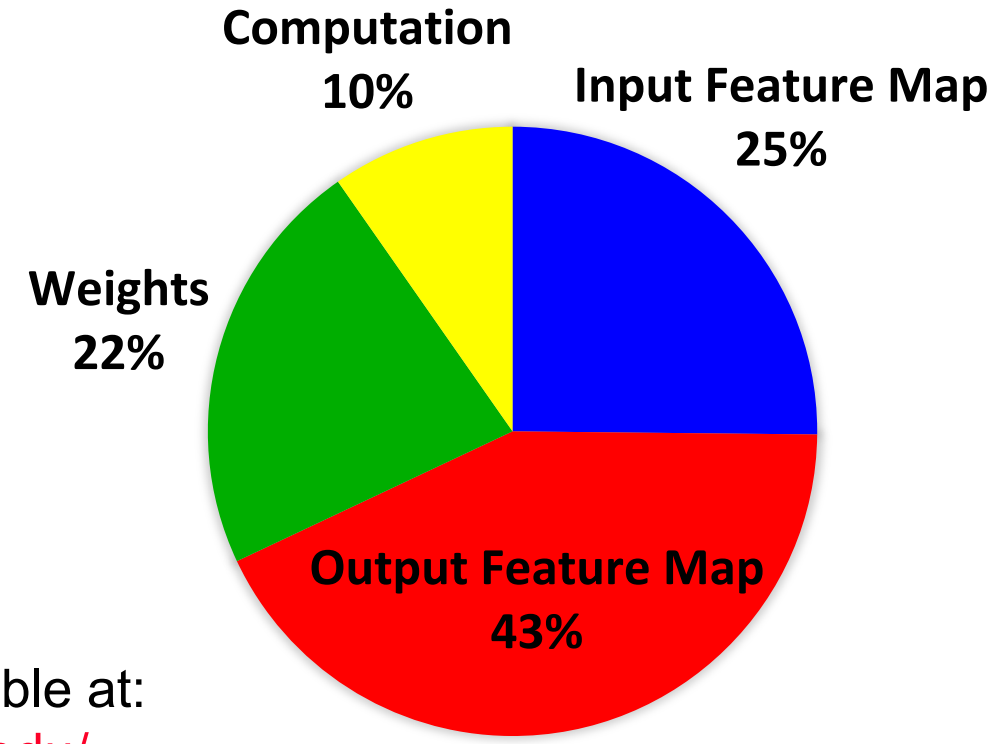
... also reduced precision

- Focus on reducing **number of MACs and weights**
- Does it translate to **energy savings** and reduced latency?

Key Observations

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

Energy Consumption of GoogLeNet



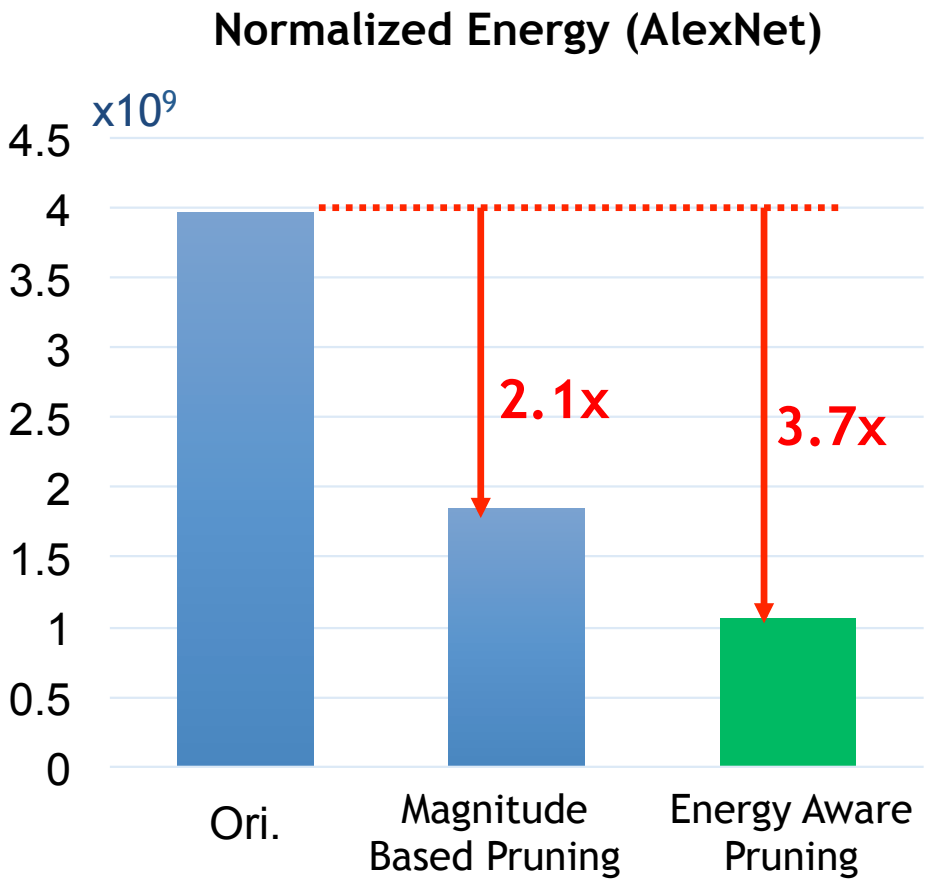
Energy estimation tool available at:
https://energyestimation.mit.edu/

[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**

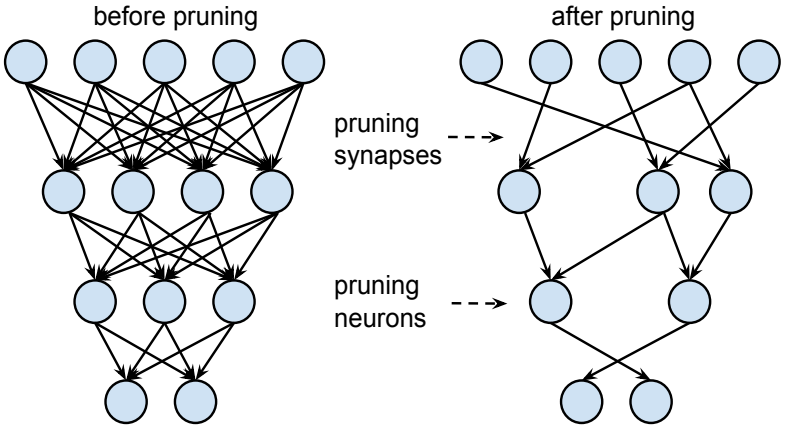


Pruned models available at
http://eyeriss.mit.edu/energy.html

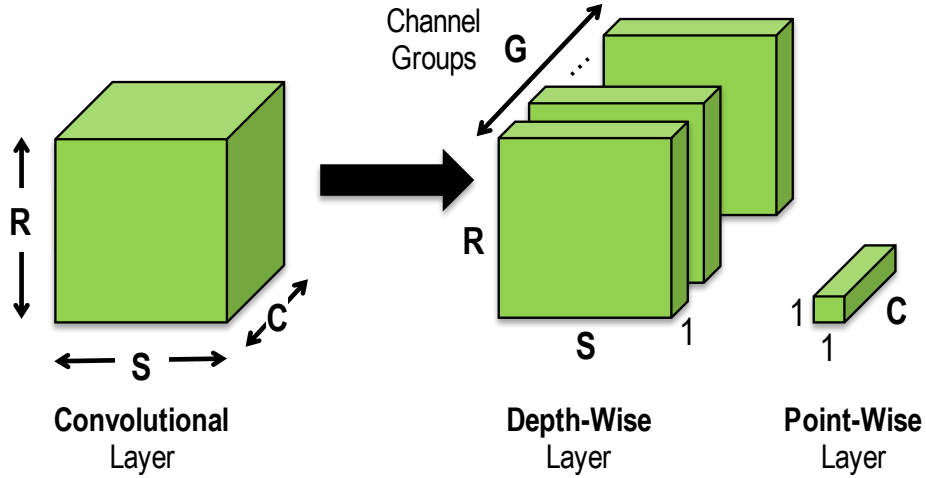
[Yang et al., CVPR 2017]

Many Efficient DNN Design Approaches

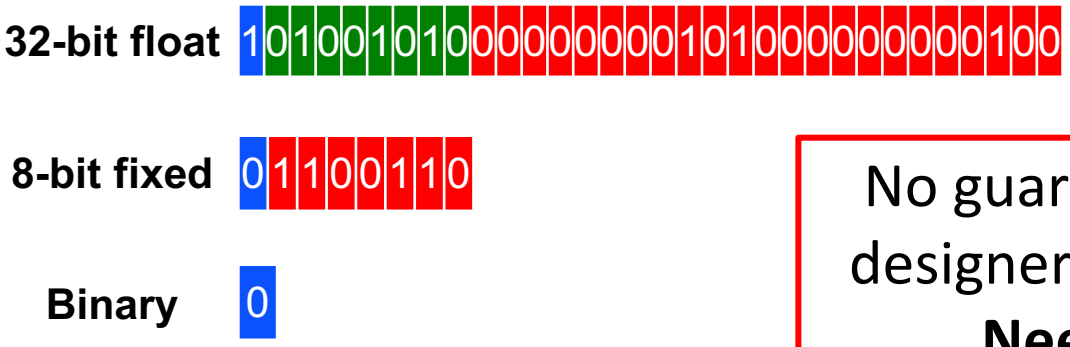
Network Pruning



Compact Network Architectures



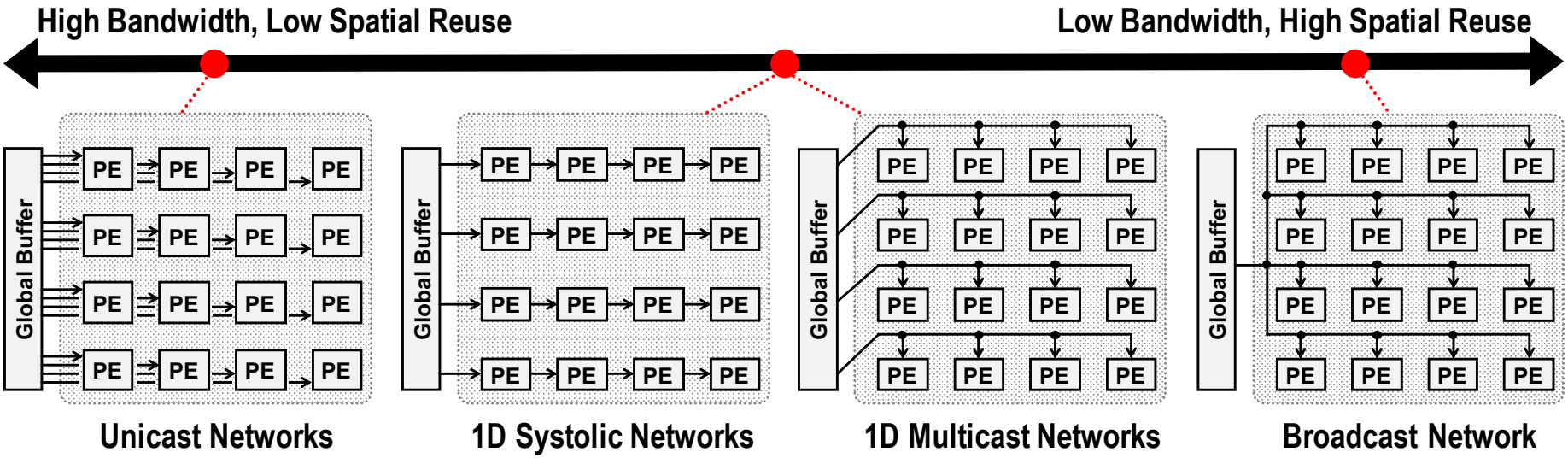
Reduce Precision



No guarantee that DNN algorithm designer will use a given approach.
Need flexible hardware!

Need Flexible NoC for Varying Reuse

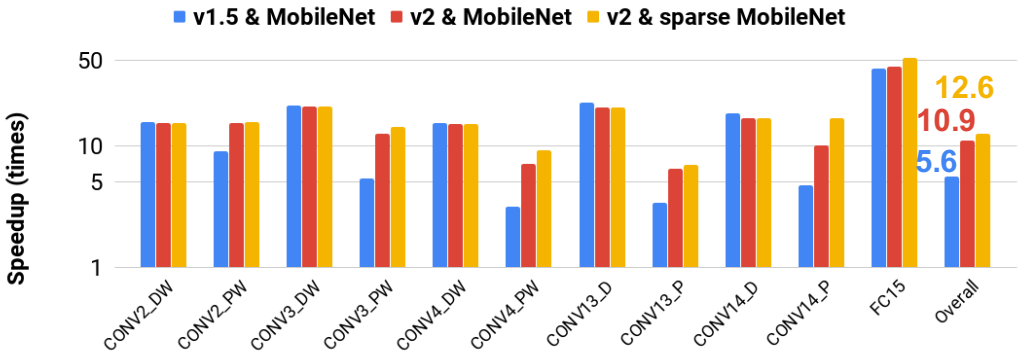
- When reuse available, need **multicast** to exploit spatial data reuse for energy efficiency and high array utilization
- When reuse not available, need **unicast** for high BW for weights for FC and weights & activations for high PE utilization
- An **all-to-all** satisfies above but too expensive and not scalable



Eyeriss v2: Balancing Flexibility and Efficiency

Efficiently supports

- Wide range of filter shapes
  - Large **and** Compact
- Different Layers
  - **CONV, FC, depth wise, etc.**
- Wide range of sparsity
  - Dense **and** Sparse
- **Scalable architecture**



Speed up over Eyeriss v1 scales with number of PEs

# of PEs	256	1024	16384
AlexNet	17.9x	71.5x	1086.7x
GoogLeNet	10.4x	37.8x	448.8x
MobileNet	15.7x	57.9x	873.0x

Over an order of magnitude faster and more energy efficient than Eyeriss v1

[Chen et al., JETCAS 2019]

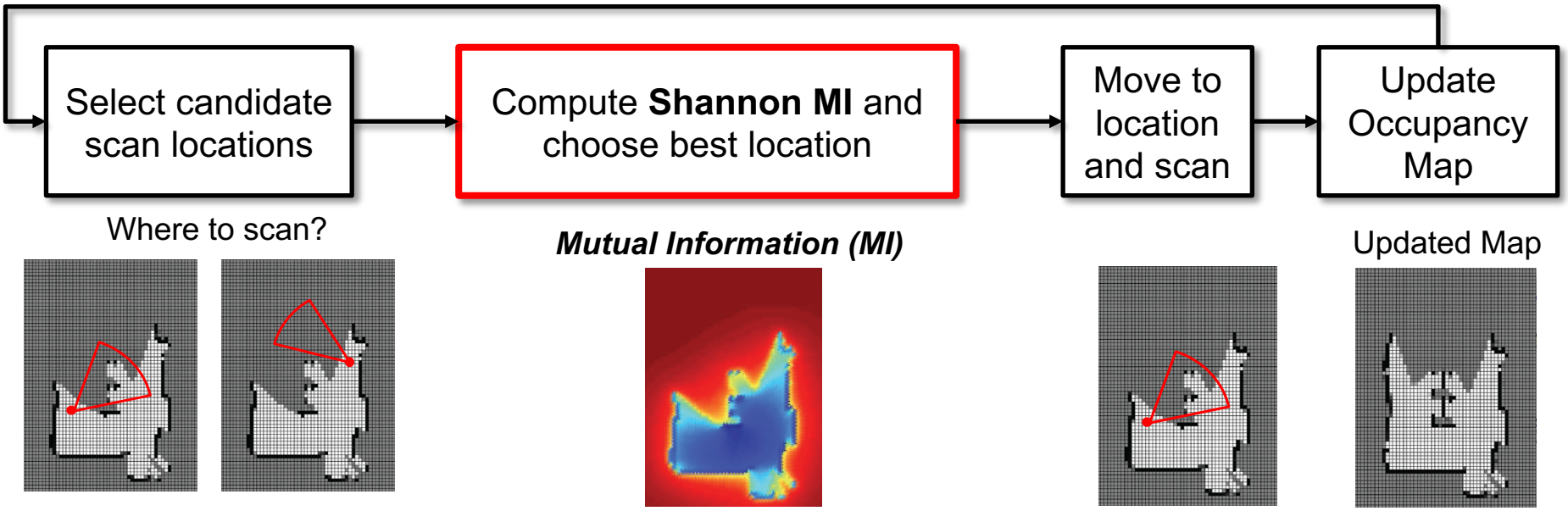
DNN Design Considerations

- **Exploit properties of workloads**
  - Efficient memory hierarchy and dataflow for data reuse; exploit natural sparsity in activation
- **Design more efficient workloads**
  - Design efficient DNN models with increased sparsity, reduced precision, and compact network architectures
  - Drive design of algorithms with direct metrics (i.e., energy, latency) rather than indirect metrics (i.e., # of ops, weights)
- **Define range of workloads**
  - Flexibility to support a wide range of DNNs, including different efficient DNN approaches

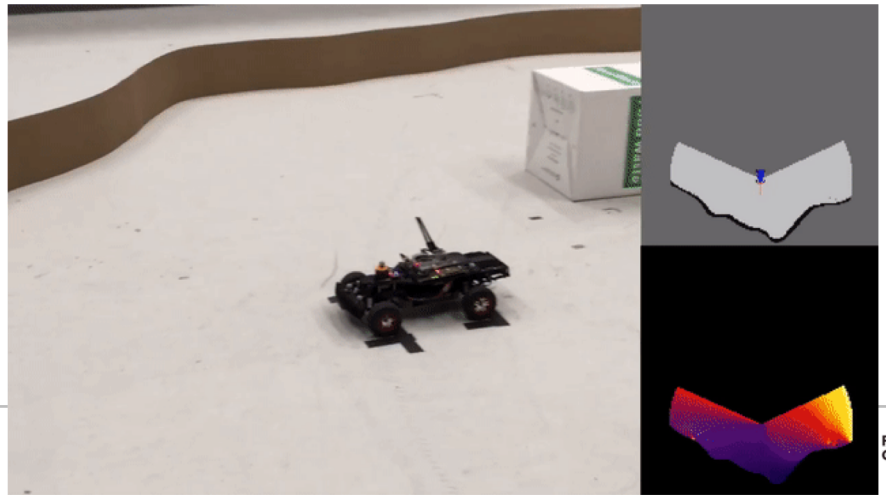
Autonomous Navigation

Robot Exploration

*Decide where to go by computing **Shannon Mutual Information***



Exploration with a mini race car using motion capture for localization

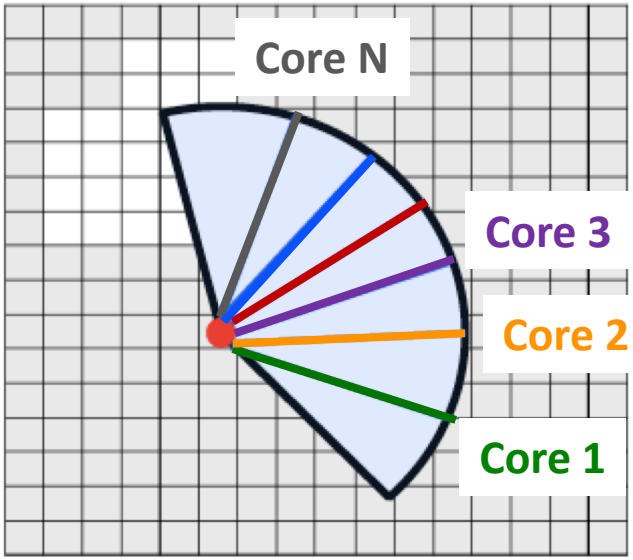


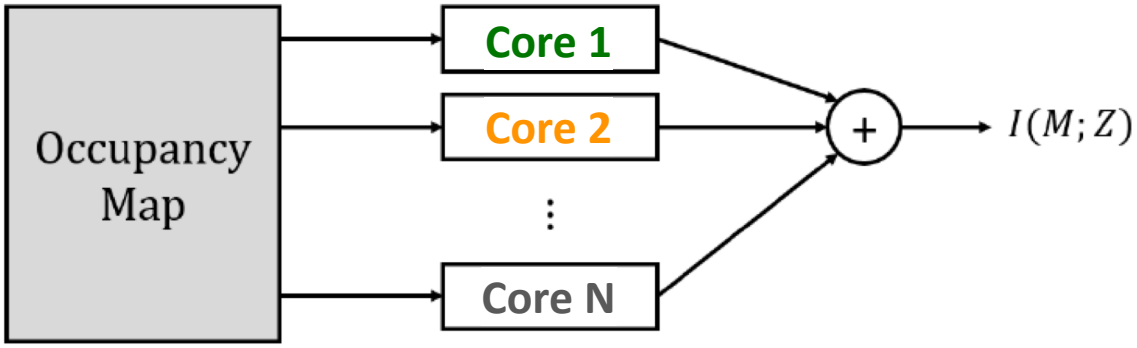
Occupancy map with planned path

MI surface

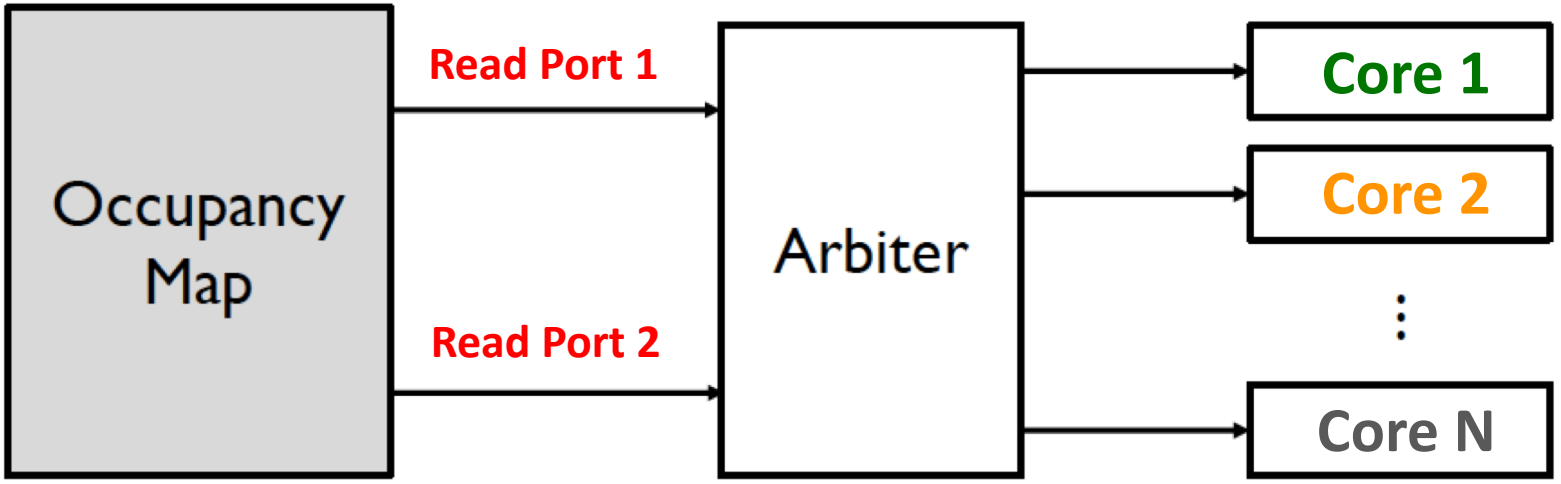
Challenge is Data Delivery to All Cores

Process multiple beams in parallel





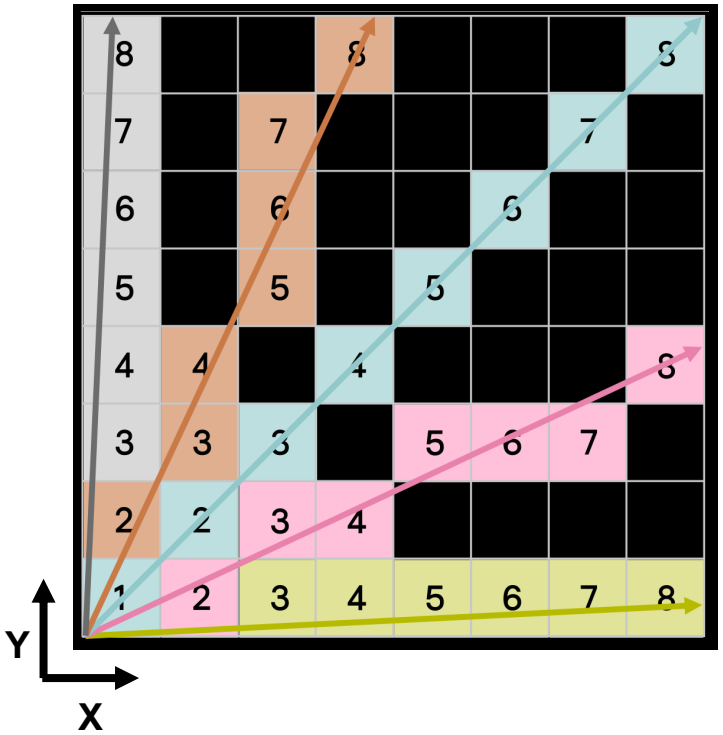
Data delivery from memory is limited



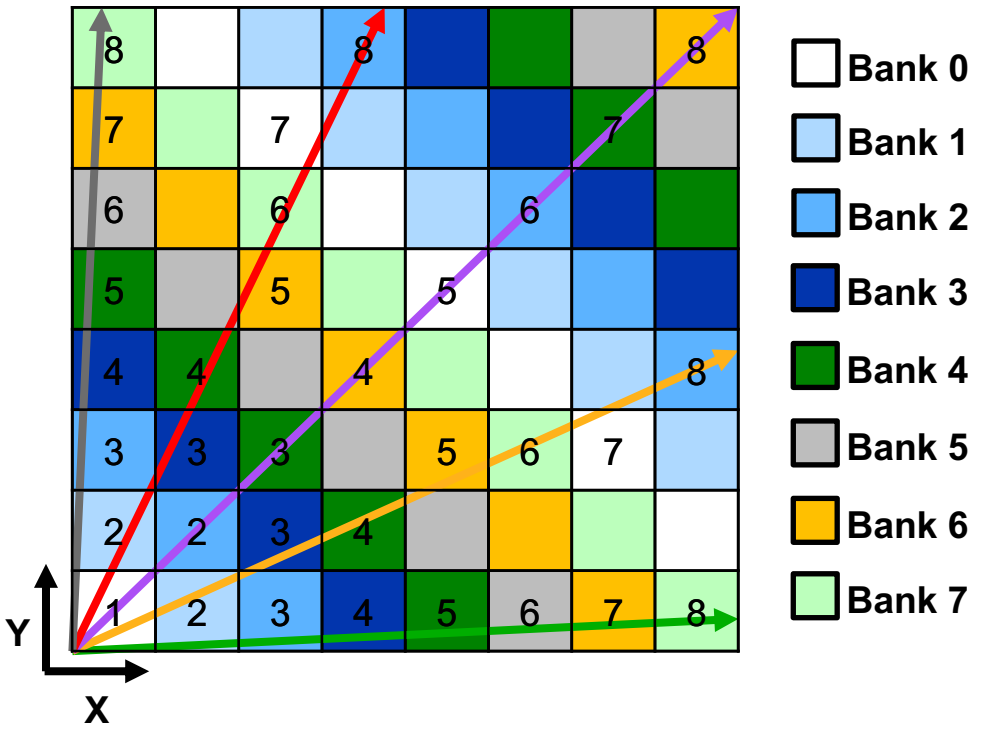
Specialized Memory Architecture

Break up map into **separate memory banks** and novel storage pattern to minimize read conflicts when processing different beams in parallel.

Memory Access Pattern

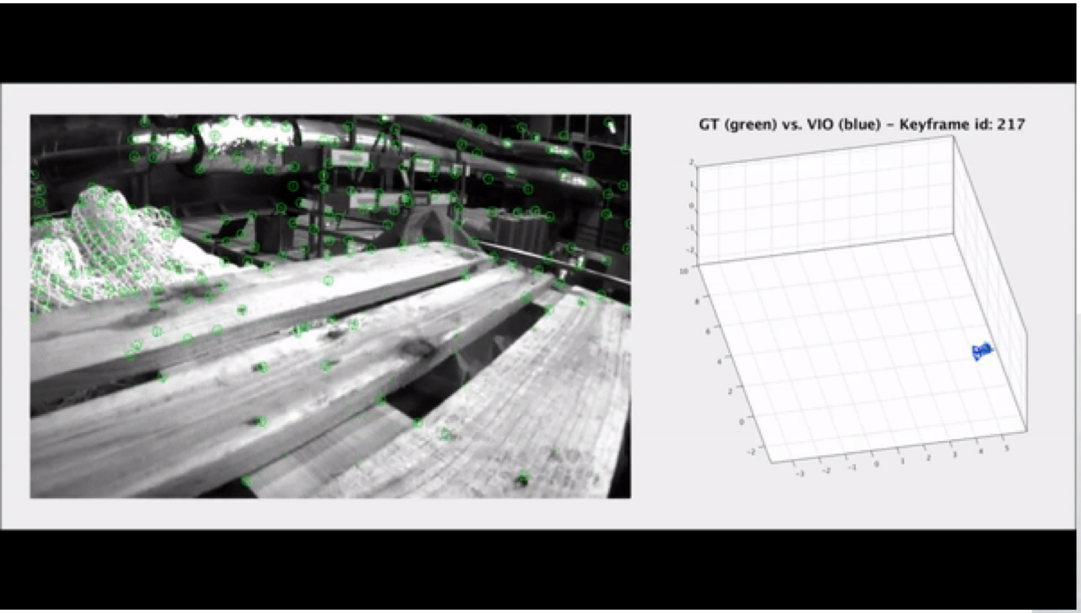


Diagonal Banking Pattern

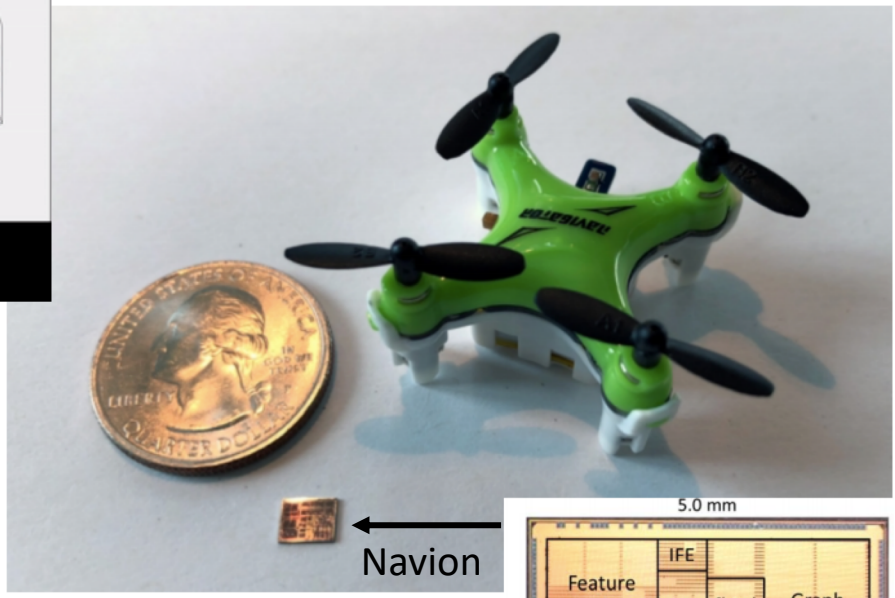


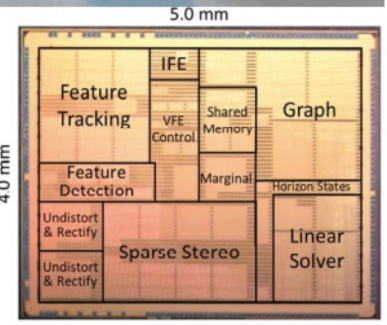
Achieves throughput **within 94% of theoretical limit** (unlimited bandwidth).
Compute ***entire 20mx20m map in under a second!***

Robot Localization in Under 25mW



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





Entire system fully integrated on chip. Use compression/sparsity to reduce total storage to **854kB!**

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

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

http://navion.mit.edu

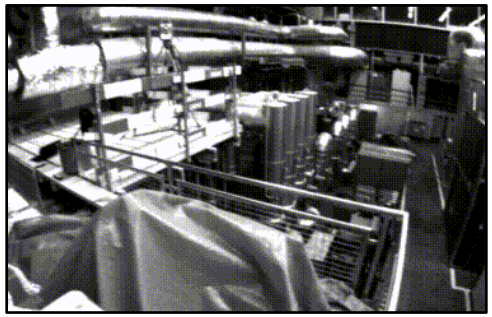
[Joint work with Sertac Karaman]

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

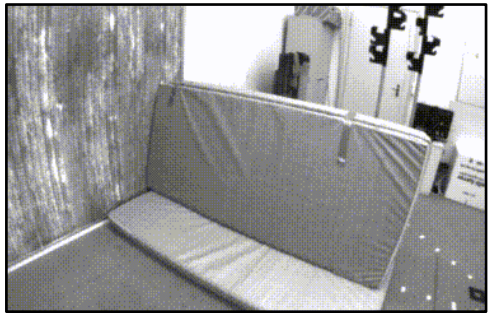
Configurable for Different Environments

EuRoC dataset is a very challenging, and widely used UAV dataset
11 sequences with three categories: easy, medium & difficult

Examples of Easy Sequences

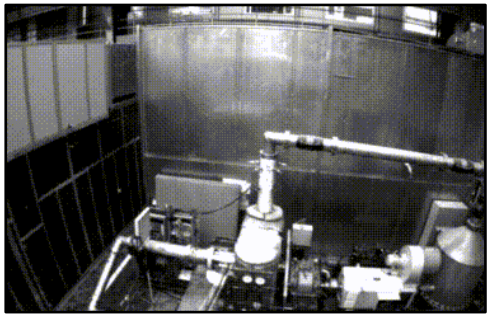


MH_1



V1_1

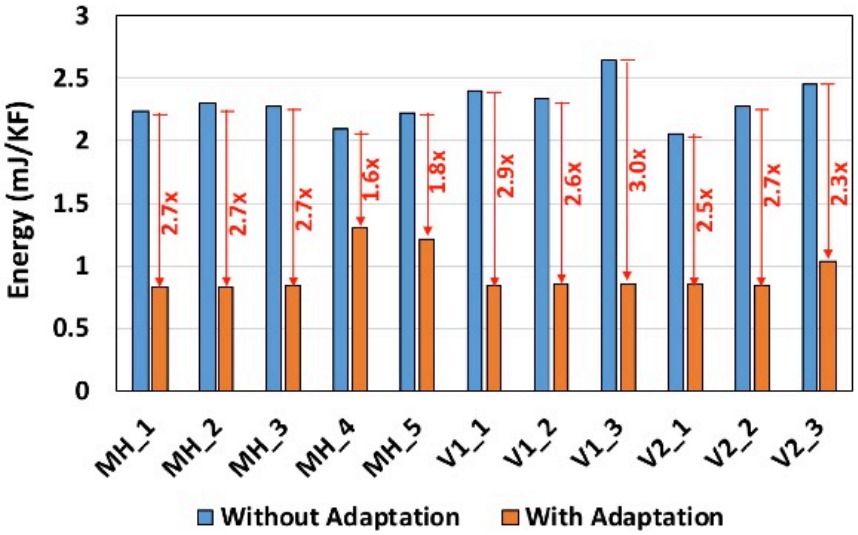
Examples of Difficult Sequences



Dark scenes (MH_4)

Motion blur (V2_3)

***Navion has over 250 configurable parameters** to adapt to different sensors and environments*



***Adapting to the environment** results in a 2 - 3x energy reduction*

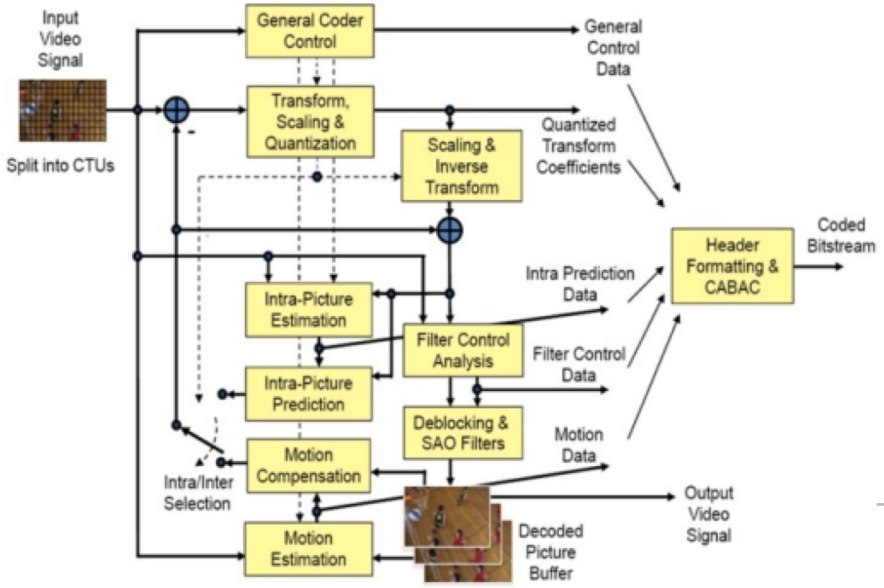
Autonomous Navigation Design Considerations

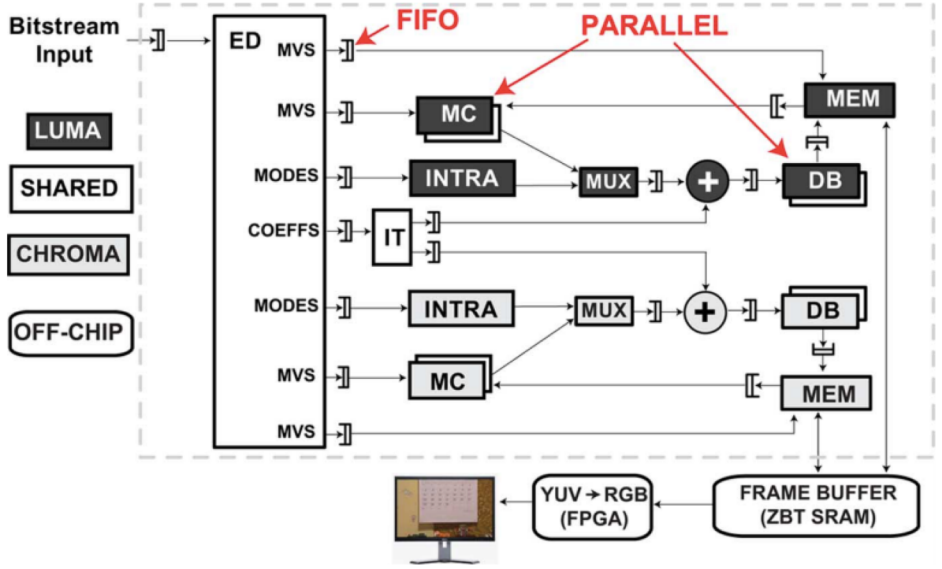
- **Exploit properties of workloads**
  - Optimized memory banking and mapping to meet memory bandwidth requirements for high throughput parallel processing
- **Design more efficient workloads**
  - Compact data representation to reduce data movement, storage and accelerate computation
- **Define range of workloads**
  - Adapt to changing environment for improved efficiency

Video Compression

Video Compression

- Video codec composed of multiple heterogenous modules
  - entropy coding, transform, motion comp, intra coding, deblocking, etc.
- Specialized hardware for each module
  - Hardcode values for parameters defined by video coding standard (e.g., weights of interpolation filter and coefficients of transform)
  - Dedicated optimized memories and dataflow for each module
- Parallel and pipeline across and within modules

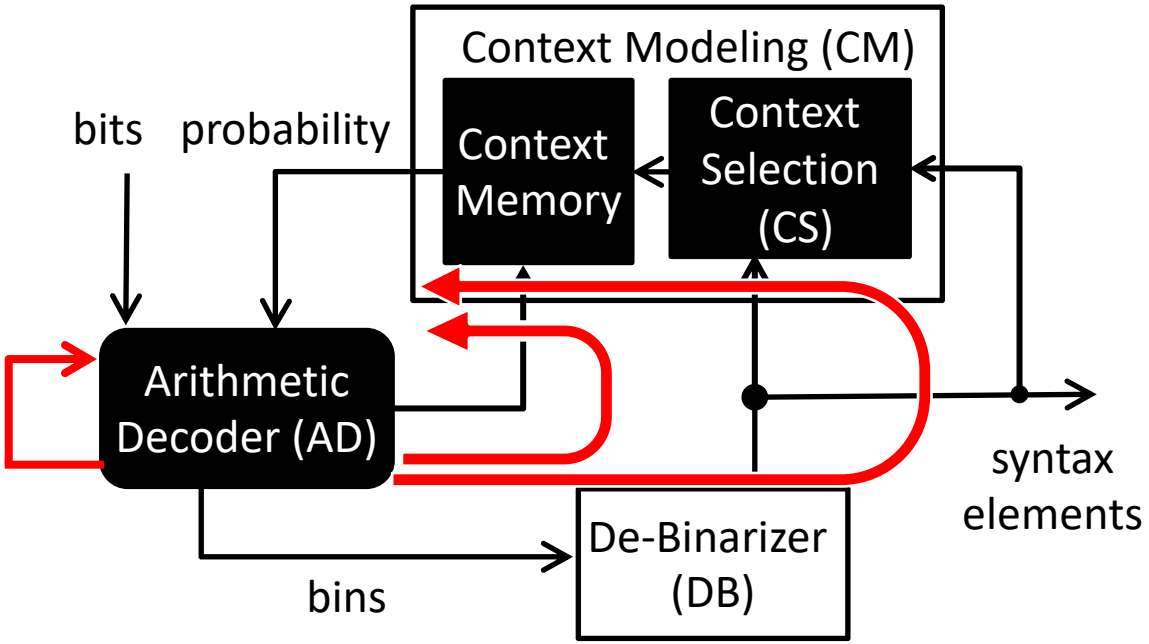




Parallelism Limited By Algorithm

- Advanced algorithms more difficult to parallelize
  - Limits throughput due to Amdahl's law

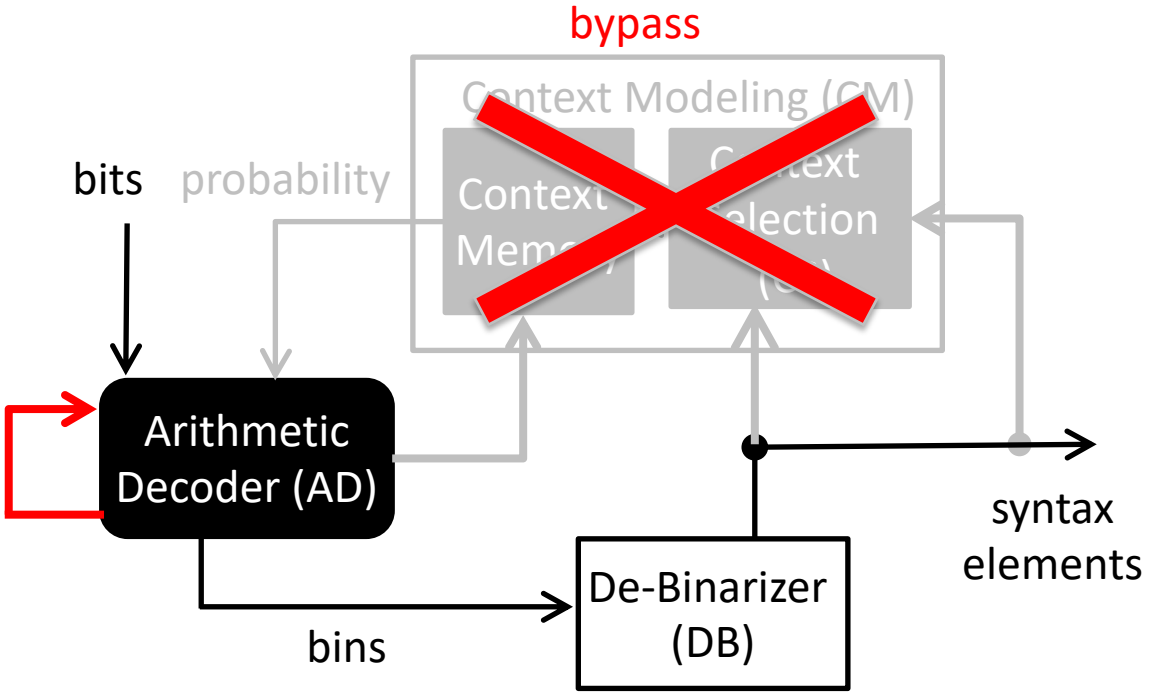
Context-Adaptive Binary Arithmetic Coding (CABAC)

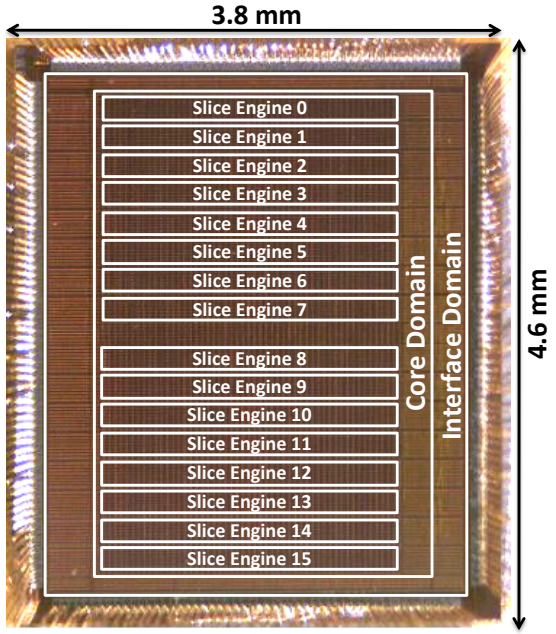


Parallelism Limited By Algorithm

- Advanced algorithms more difficult to parallelize
- Re-design algorithms to be more hardware-friendly

Context-Adaptive Binary Arithmetic Coding (CABAC)





[Sze et al., ISSCC 2011]

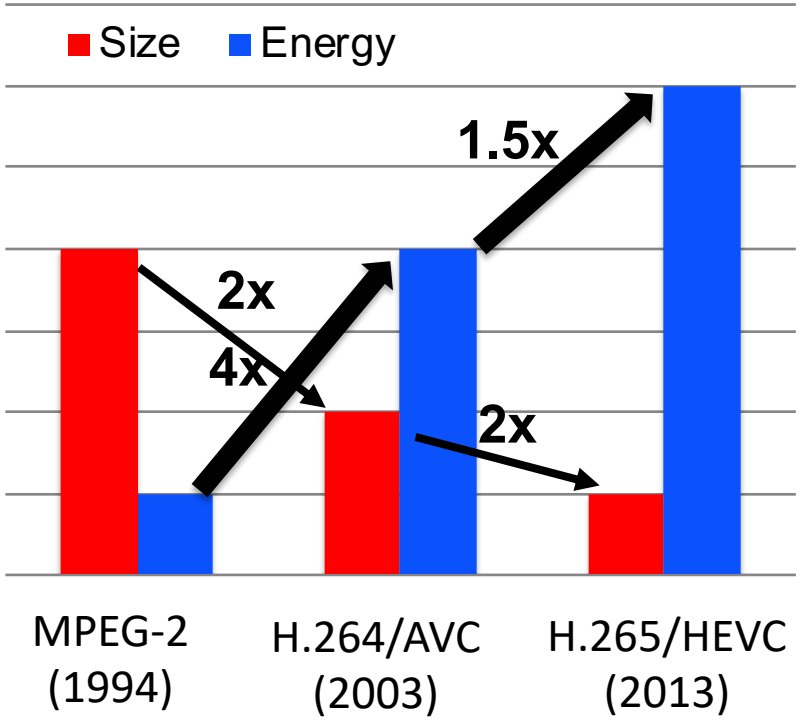
Parallel entropy coding algorithm gives **>10x higher throughput** than state-of-the-art with minimal impact on coding loss

[Joint work with Anantha Chandrakasan]

High Efficiency Video Coding (HEVC)

- H.265/HEVC is the successor to H.264/AVC
- **Achieves 2x higher compression than H.264/AVC**
- High throughput (Ultra-HD 8K @ 120fps) & low power

Primetime Emmy



	Coding Efficiency	Efficient Implementation
Larger and Flexible Coding Block Size	X	
More Sophisticated Intra Prediction	X	
Larger Interpolation for Motion Comp.	X	
Larger Transform Size	X	
Parallel Deblocking Filter		X
Sample Adaptive Offset	X	
High-Throughput CABAC	X	X
High Level Parallel Tools		X

Co-design of algorithm and hardware to address **coding efficiency, throughput and power challenges**

Flexibility Needed for Video Compression

- ***Support multiple standards***

Legacy (2003, 2013)

H.264 MPEG-4/AVC

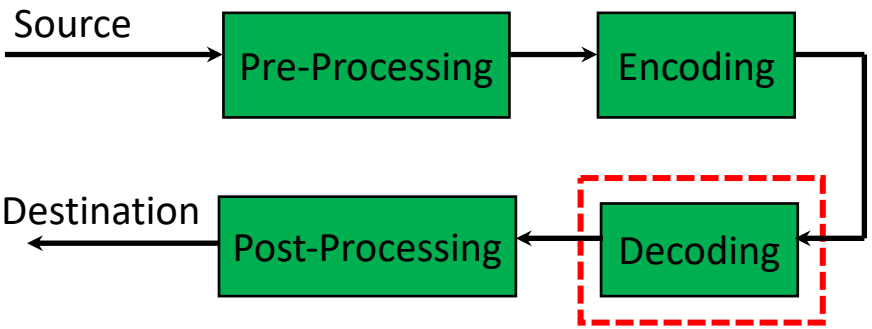
HEVC H.265 - HIGH EFFICIENCY VIDEO CODING

Emerging (2019, 2020)

AV1 VVC

Shared resources (e.g., cache for motion compensation), but modules tend to be hardcoded due to tight power and speed requirements

- ***Encoder must be flexible***



While decoder is standardize, encoder allows for product differentiation (e.g., better motion estimation)

Video Compression Design Considerations

- **Exploit properties of workloads**
  - Specialized hardware for heterogenous set of modules with hardcoded parameters; exploit parallelism and pipelining
- **Design more efficient workloads**
  - Parallel entropy coding algorithm to remove compute bottleneck
  - Co-design of algorithms and hardware in HEVC standard
- **Define range of workloads**
  - Flexibility to support multiple video standards and algorithm changes in encoder

Summary

- Domain-specific hardware can address the rising compute demands for many existing and emerging applications
- **Opportunities**
  - Exploit properties of workloads (e.g., parallelism, access patterns, representation)
  - Design efficient workloads using co-design of algorithms and hardware without affecting quality of result
- **Challenges**
  - Define range of workloads to support based on flexibility versus efficiency tradeoff
  - Workloads will evolve over time and across use cases/environments
- Agile design can be used for rapid exploration of workloads and tradeoff
- Open hardware can allows for rapid system development with **shared building blocks**
  - May need to configure for given application requirements
  - What is the granularity of the blocks?

Acknowledgements

Joel Emer

Anantha Chandrakasan

Sertac Karaman

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.-J Yang, J. Emer, V. Sze, "Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices," IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), Vol. 9, No. 2, pp. 292-308, June 2019.*
  - *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.*
  - *Y.-H. Chen*, T.-J. Yang*, J. Emer, V. Sze, "Understanding the Limitations of Existing Energy-Efficient Design Approaches for Deep Neural Networks," SysML Conference, February 2018.*
  - *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.*
  - *A. Suleiman*, Y.-H. Chen*, J. Emer, V. Sze, "Towards Closing the Energy Gap Between HOG and CNN Features for Embedded Vision," IEEE International Symposium of Circuits and Systems (ISCAS), Invited Paper, May 2017.*
  - *Hardware Architecture for Deep Neural Networks: http://eyeriss.mit.edu/tutorial.html*

References

- **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.*

References

- **Video Compression**
  - V. Sze, A. P. Chandrakasan, "A Highly Parallel and Scalable CABAC Decoder for Next-Generation Video Coding," IEEE Journal of Solid-State Circuits (JSSC), ISSCC Special Issue, Vol. 47, No. 1, pp. 8-22, January 2012.
  - V. Sze, M. Budagavi, "High Throughput CABAC Entropy Coding in HEVC," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), Vol. 22, No. 12, pp. 1778-1791, December 2012.
  - V. Sze, A. P. Chandrakasan, "Joint Algorithm-Architecture Optimization of CABAC to Increase Speed and Reduce Area Cost," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1577–1580, May 2011.
  - V. Sze, A. P. Chandrakasan, "A High Throughput CABAC Algorithm Using Syntax Element Partitioning," *IEEE International Conference on Image Processing (ICIP)*, pp. 773-776, November 2009.
  - V. Sze, M. Budagavi, A. P. Chandrakasan, M. Zhou, "Parallel CABAC for Low Power Video Coding," *IEEE International Conference on Image Processing (ICIP)*, pp. 2096-2099, October 2008.
  - V. Sze, D. F. Finchelstein, M. E. Sinangil, A. P. Chandrakasan, "A 0.7-V 1.8-mW H.264/AVC 720p Video Decoder," IEEE Journal of Solid State Circuits (JSSC), A-SSCC Special Issue , Vol. 44, No. 11, pp. 2943-2956, November 2009.
  - V. Sze, M. Budagavi, G. J. Sullivan (Editors), High Efficiency Video Coding (HEVC): Algorithms and Architectures, Springer, 2014.