The Intersection of SSCS and AI

- A Tale of Two Journeys -

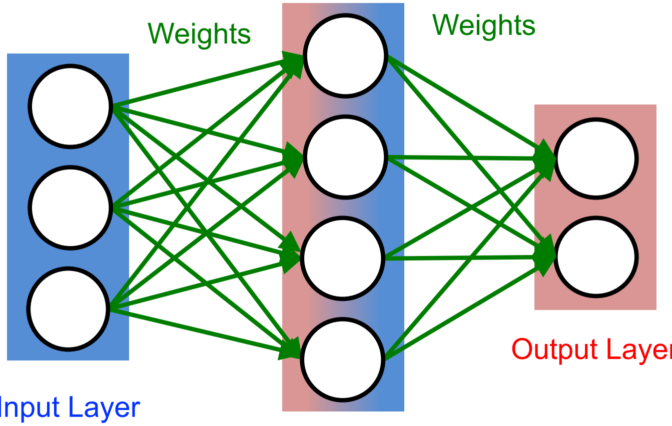
Vivienne Sze (@eems_mit)

Massachusetts Institute of Technology

In collaboration with Madhukar Budagavi, Luca Carlone, Anantha Chandrakasan, Yu-Hsin Chen, Joel Emer, Daniel Finchelstein, Sertac Karaman, Tushar Krishna, Thomas Heldt, Theia Henderson, Hsin-Yu Lai, Peter Li, Fangchang Ma, James Noraky, Gladynel Saavedra Peña, Mahmut Sinangil, Charlie Sodini, Amr Suleiman, Diana Wofk, Nellie Wu, Tien-Ju Yang, Zhengdong Zhang, Minhua Zhou

Slides available at
https://tinyurl.com/szeSSCStinyML

Wide Range of Compute-Intensive Applications

Video Compression	***AI: Deep Neural Networks***	***Robotics: Autonomous Navigation***
		

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

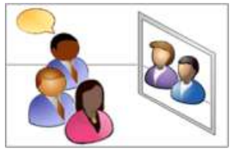
Co-design *across algorithms, architectures, circuits, and systems*

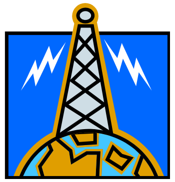
Compressing Pixels

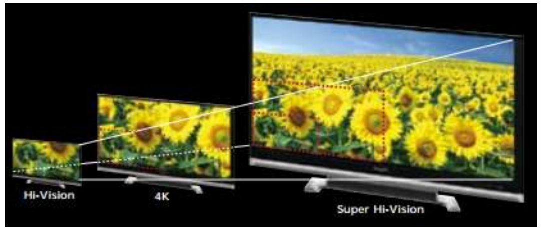
PhD at MIT (2006-2010)
Member of Technical Staff at Texas Instruments (2010-2013)
Goal: Make pixel compression ubiquitous on portable devices

Video is the Biggest Big Data

- Video accounts for over 70% of today's Internet traffic. Increase in applications, content, fidelity, etc.
  → **Need to compress well**
- Ultra-HD 4K televisions and 360º for virtual reality.
  → **Need to compress fast**
- Video is a "must have" on portable devices. Battery capacity is not keeping up with processing demands.
  → **Need to use less power to compress**



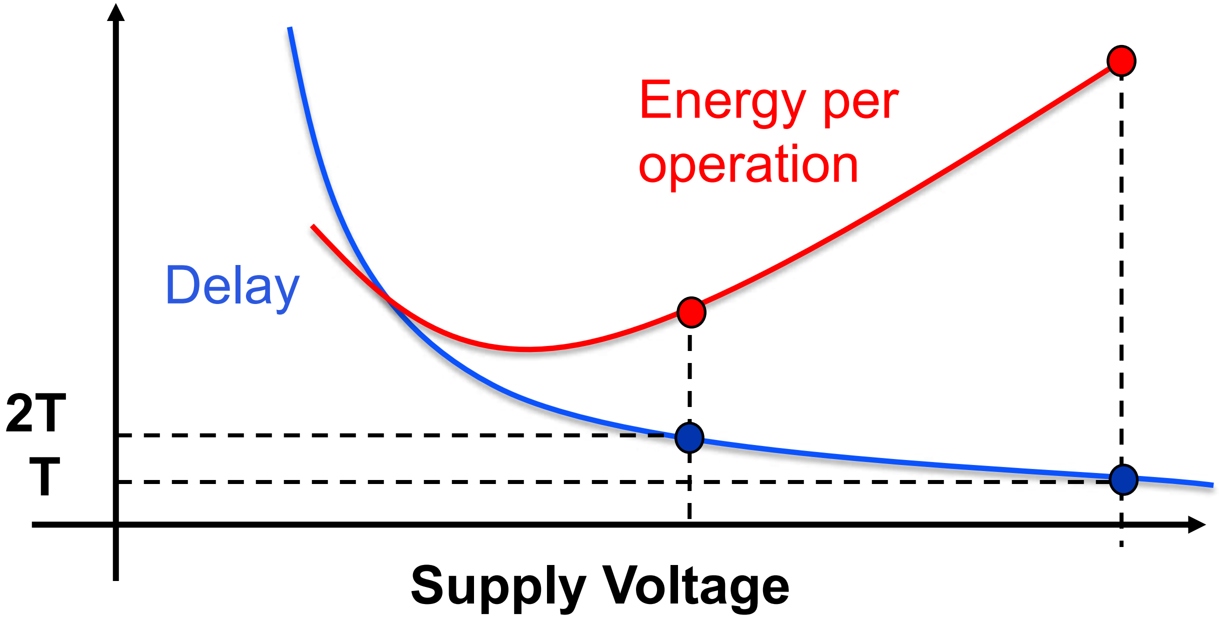


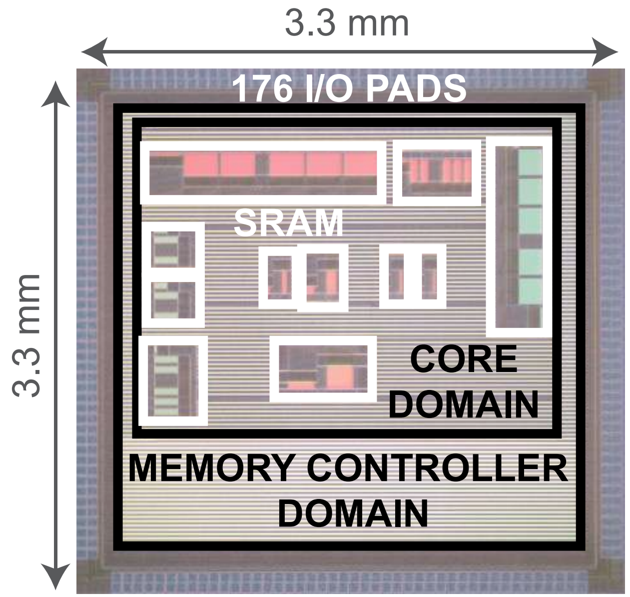


Sources: Cisco Visual Networking Index
Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update

Low Power Design for Video Compression

- H.264/AVC used to decode over 80% of video content online
- Voltage scaling and parallelism to reduce power consumption





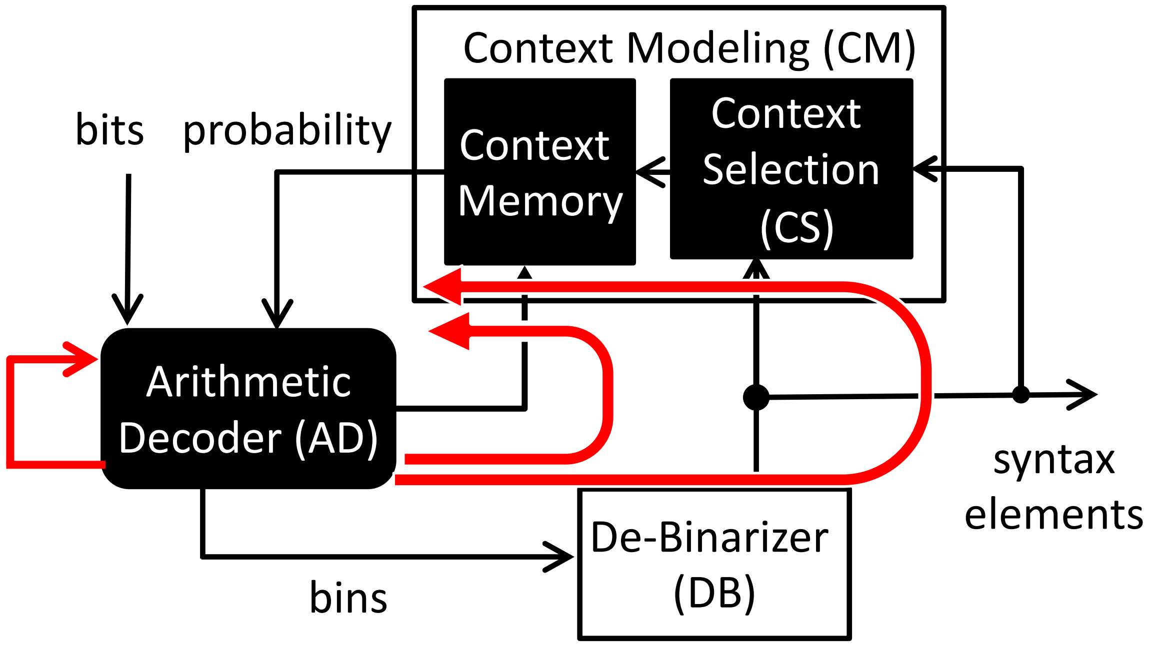
[**Sze**, *JSSC* 2009]

Achieves high definition (720p @ 30fps) decoding at under 2mW
Over 6x lower power than state-of-the-art

Parallelism Limited By Algorithm

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

Context-Adaptive Binary Arithmetic Coding (CABAC)

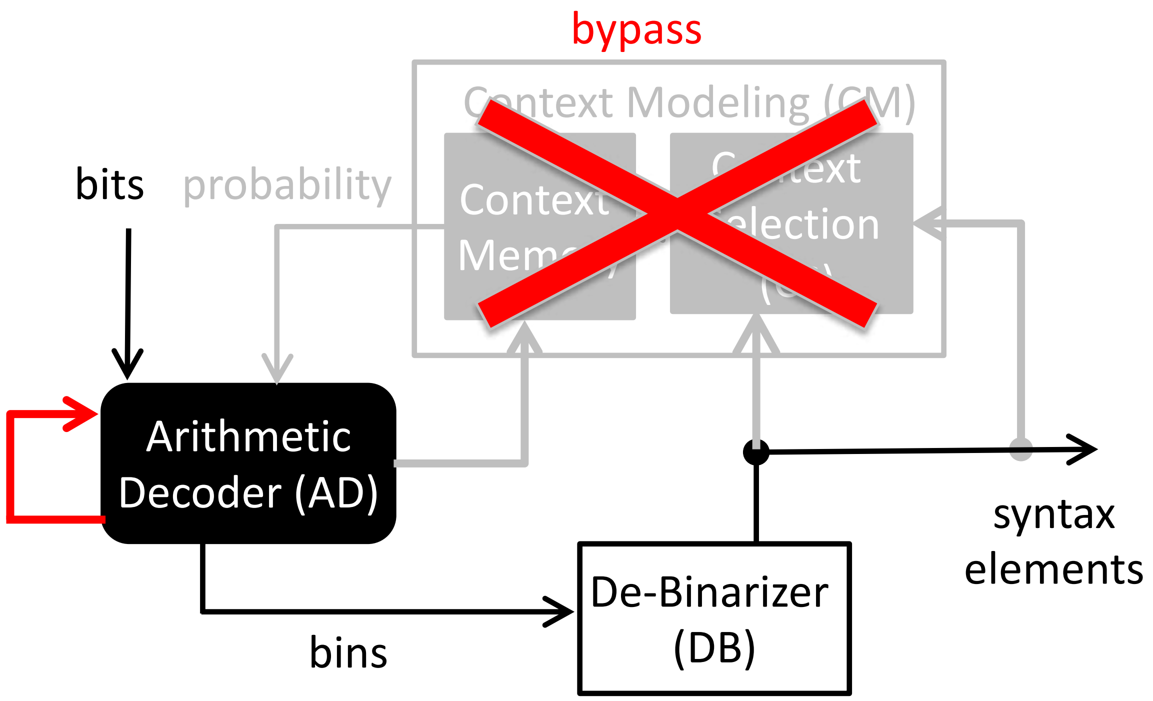


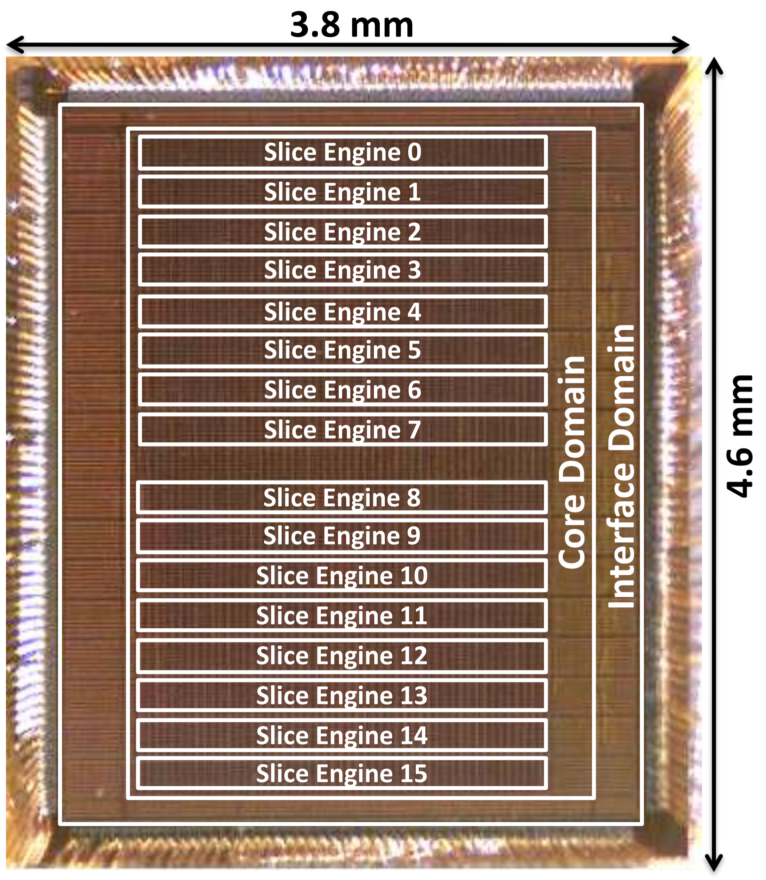
[Joint work with Anantha Chandrakasan]

Parallelism Limited By Algorithm

- Advanced algorithms more difficult to parallelize
- Co-design algorithms and hardware

Context-Adaptive Binary Arithmetic Coding (CABAC)





[**Sze**, *ISSCC* 2011]

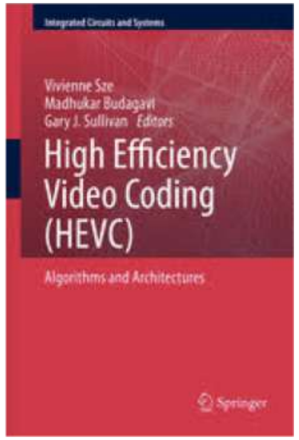
Parallel entropy coding algorithm gives **>10x higher throughput** than state-of-the-art with minimal impact on coding efficiency

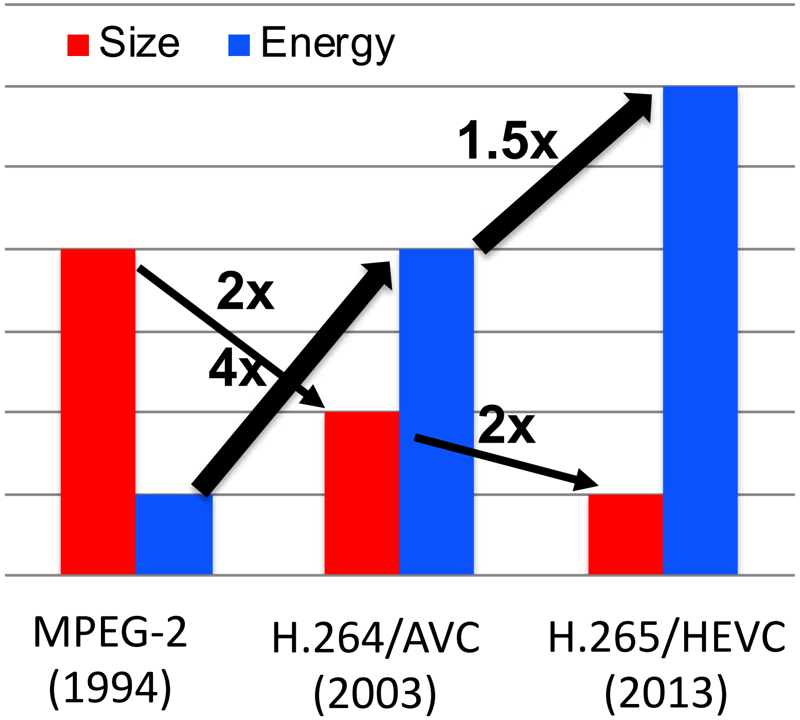
[Joint work with Anantha Chandrakasan]

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

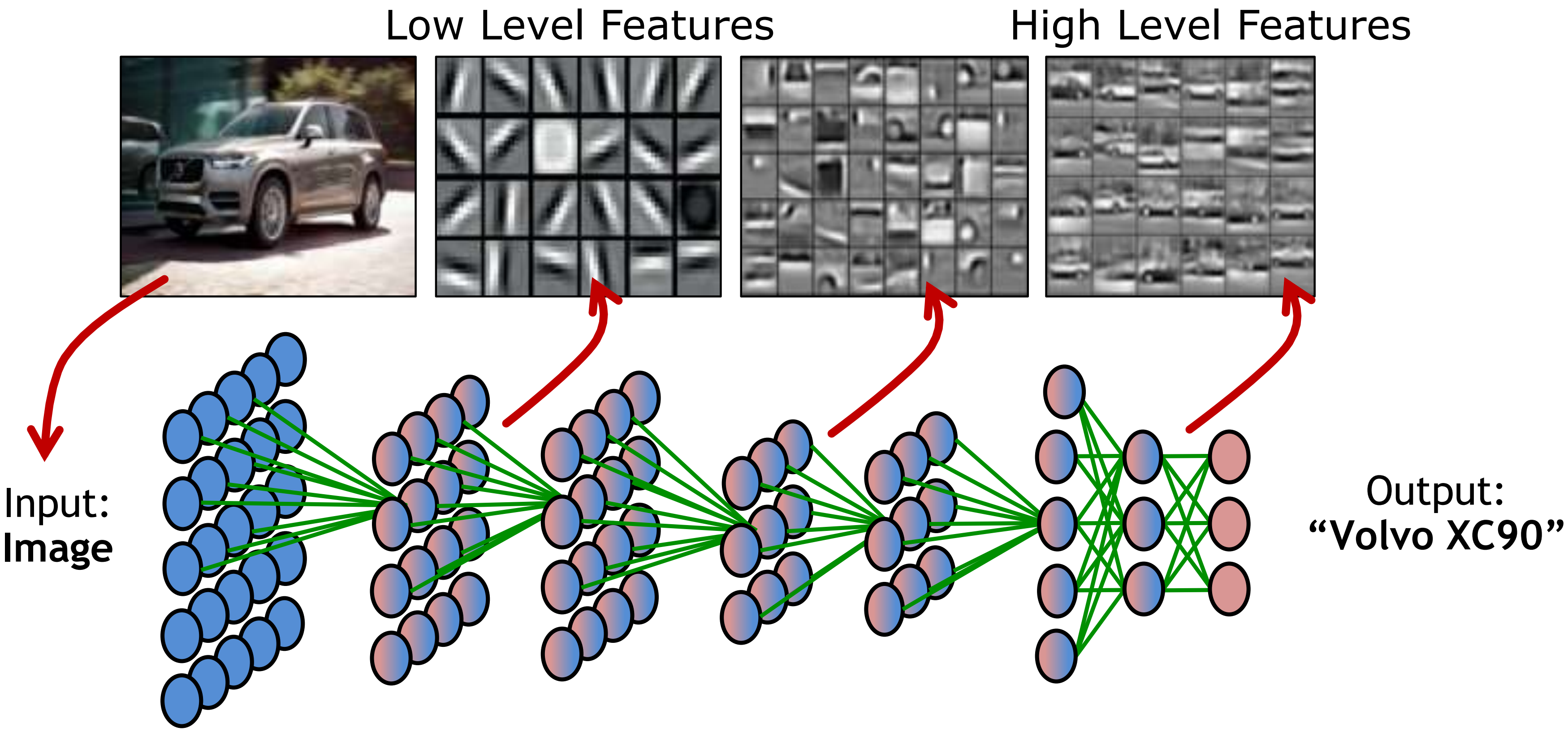
Co-design algorithm & hardware to address **coding efficiency, throughput and power challenges**

Understanding Pixels

Faculty at MIT (2013 - present)

Goal: Make understanding pixels as ubiquitous as compressing pixels

Deep Neural Networks



Modified Image Source: [**Lee**, *CACM* 2011]

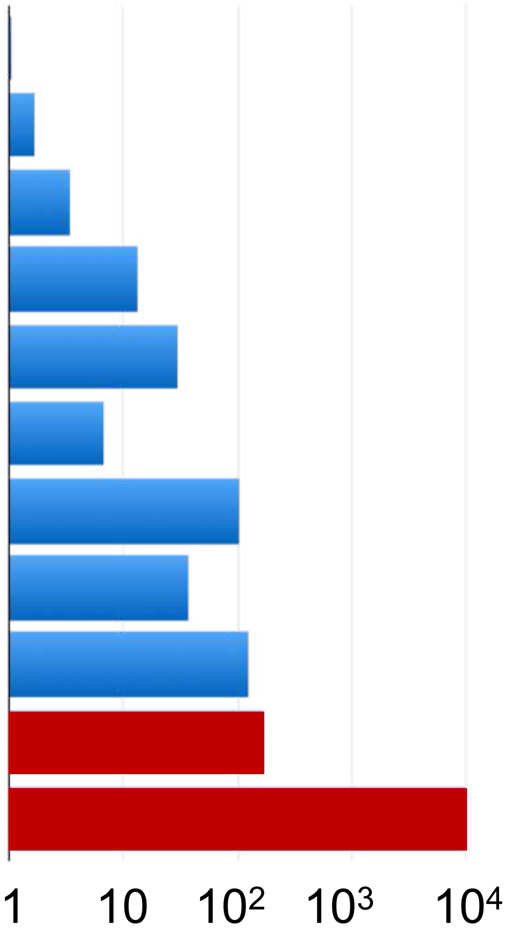
Deep Neural Networks (DNNs) delivers **state-of-the-art accuracy**,
but require **up to several hundred millions of operations and weights to compute!**

DNNs are >100x more complex than video compression

Power Dominated by Data Movement

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

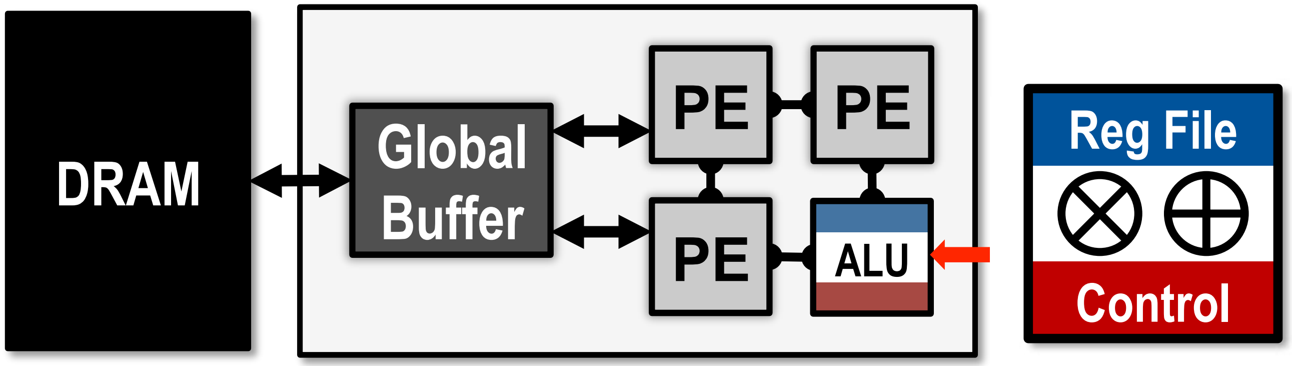
Relative Energy Cost



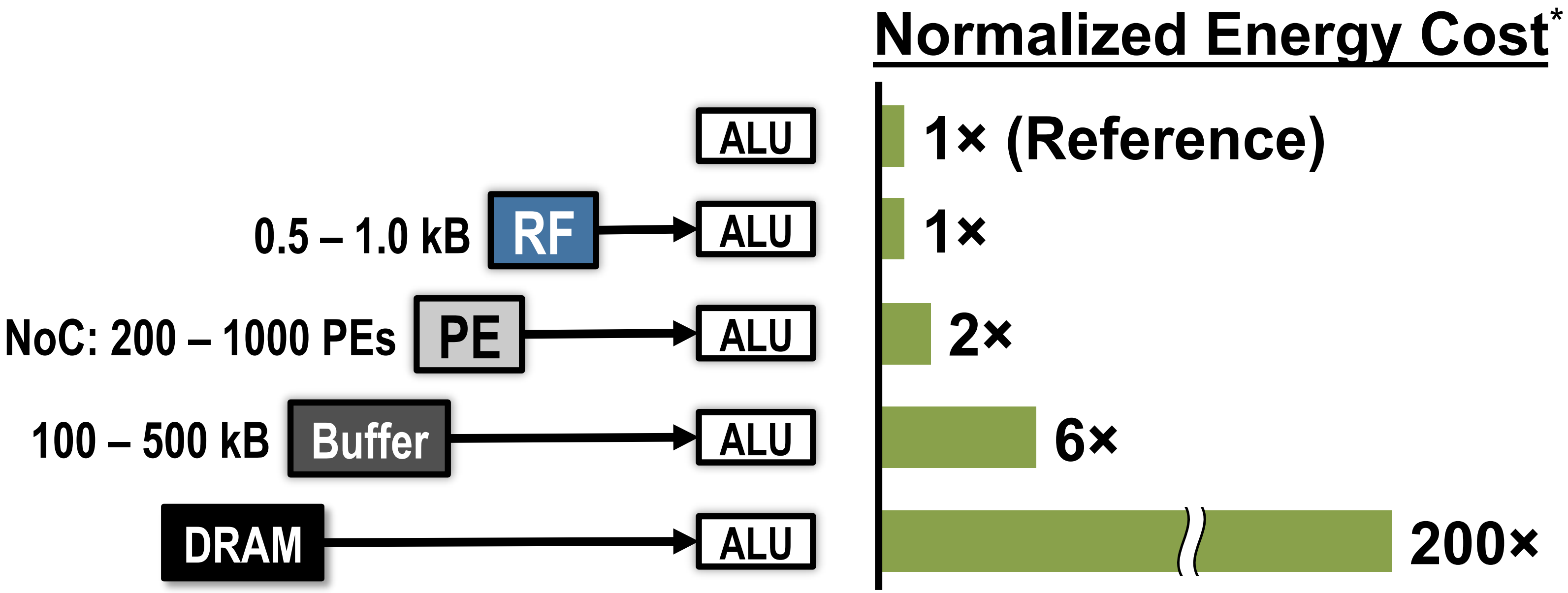
Memory access is **orders of magnitude** higher energy than compute

[**Horowitz**, *ISSCC* 2014]

Exploit Data Reuse at Low-Cost Memories



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

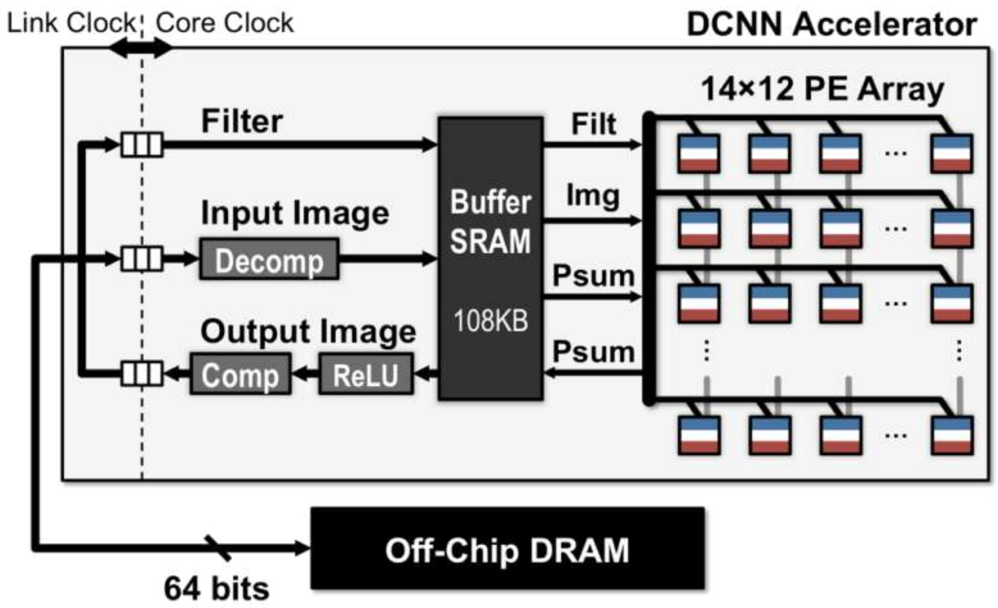


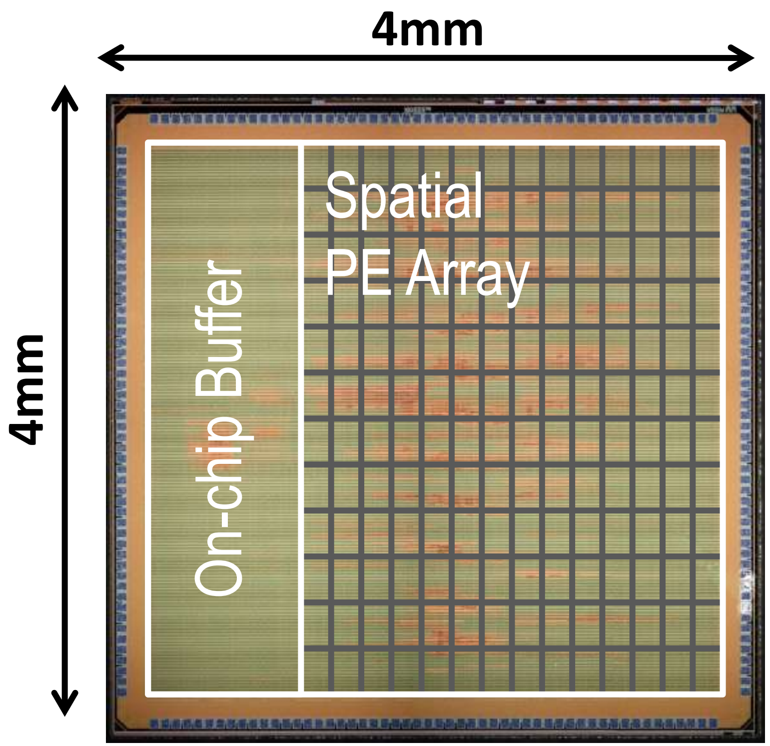
Farther and **larger** memories consume more power

* measured from a commercial 65nm process

Flexible and Efficient DNN Processor

Eyeriss





Eyeriss Project Website: http://eyeriss.mit.edu

[**Chen**, *ISSCC* 2016],[**Chen**, *ISCA* 2016] **Micro Top Picks**

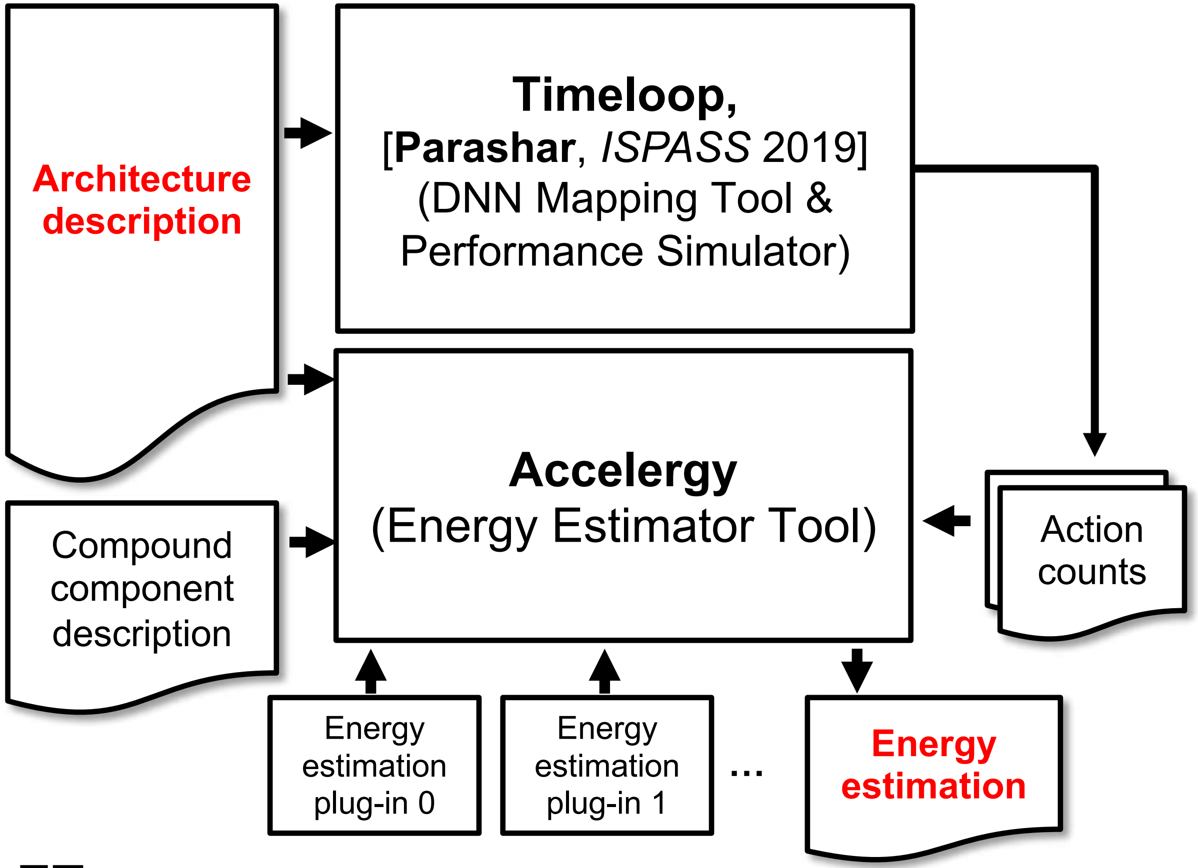
Exploits data reuse for **100x** reduction in memory accesses from global buffer and **1400x** reduction in memory accesses from off-chip DRAM

Overall **>10x energy reduction** compared to a mobile GPU

DNN Processor Evaluation Tools

- Provide a systematic way to
  - Evaluate and compare wide range of DNN processor designs
  - Rapidly explore design space

Use tool set to bridge architectures, circuits, and ***devices (e.g., in-memory processing)***



Tutorial ***this*** Friday, May 29 @ 10 AM ET
http://accelergy.mit.edu/isca20_tutorial.html

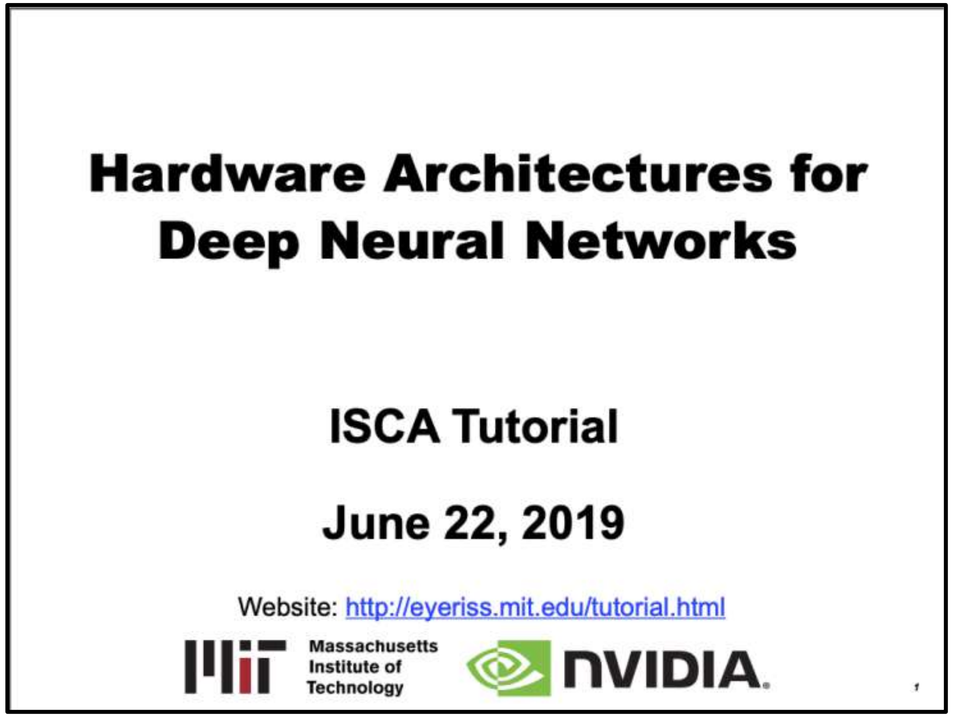
Open-source code available at:
http://accelergy.mit.edu

[**Wu**, *ICCAD* 2019], [**Wu**, *ISPASS* 2020]

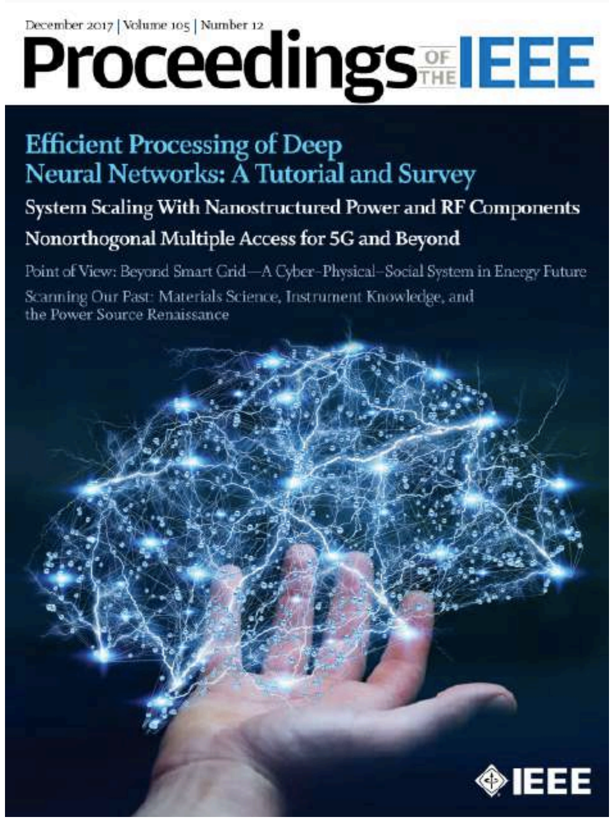
[Joint work with Joel Emer]

Energy-Efficient Processing of DNNs

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



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



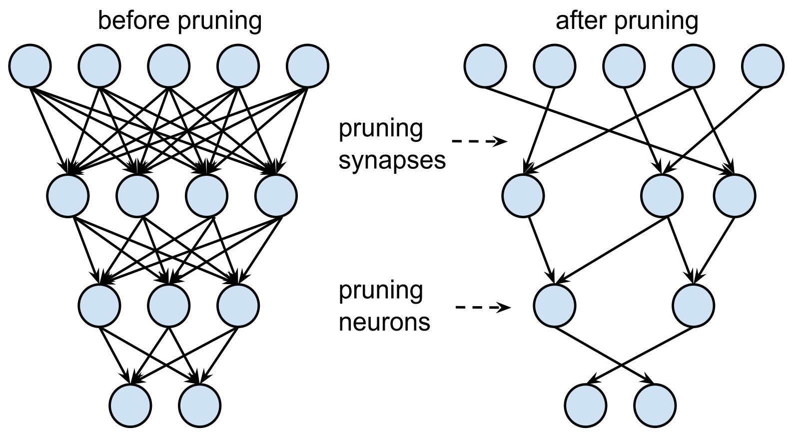
V. Sze, Y.-H. Chen, T-J. Yang, J. Emer, "***Efficient Processing of Deep Neural Networks: A Tutorial and Survey***," Proceedings of the IEEE, Dec. 2017

We identified various limitations to existing approaches

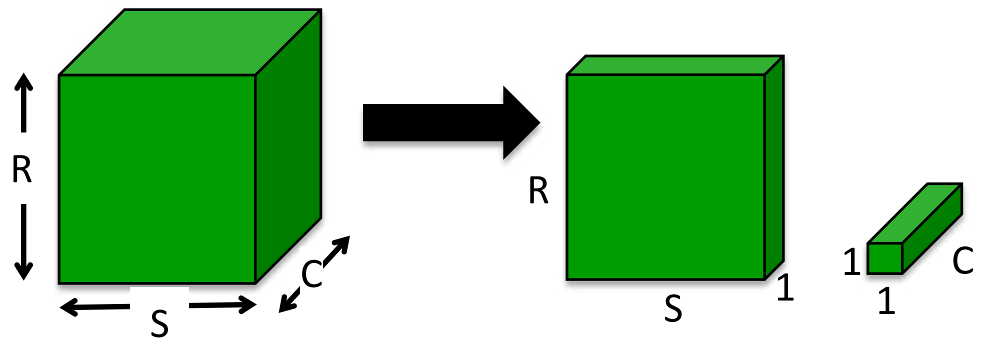
Design of Efficient DNN Algorithms

Popular efficient DNN algorithm approaches

Network Pruning



Efficient Network Architectures



Examples: SqueezeNet, MobileNet

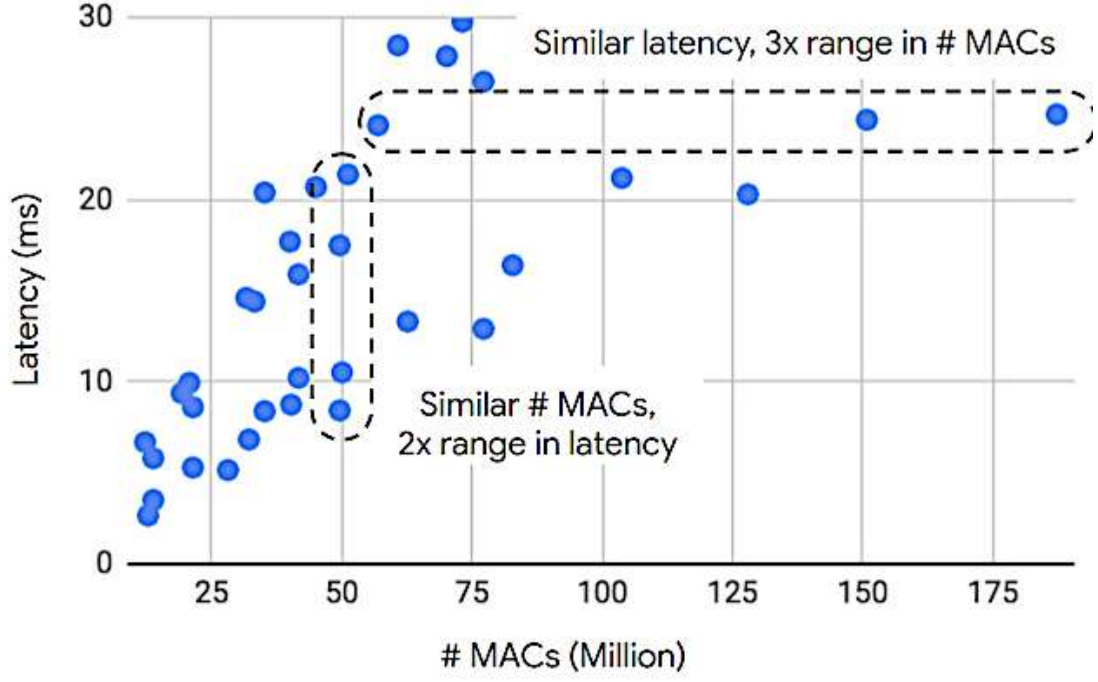
... also reduced precision

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

[**Chen*, Yang***, *SysML* 2018]

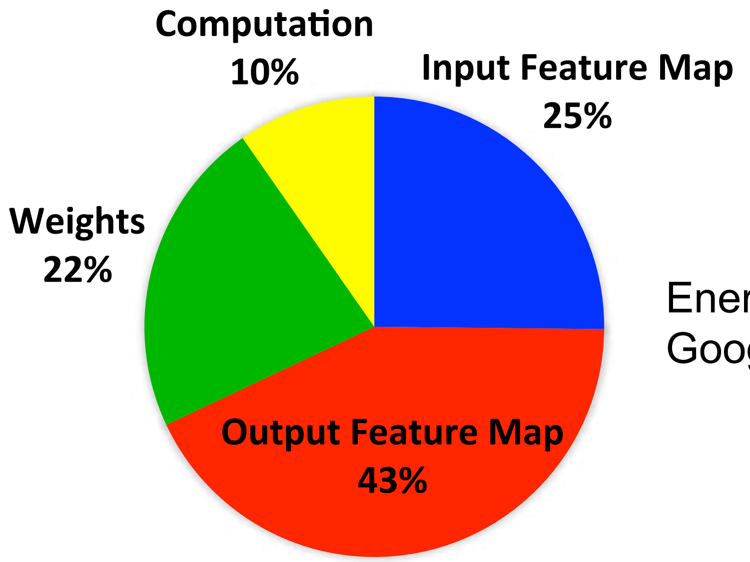
Number of MACs and Weights are Not Good Proxies

of operations (MACs) does not approximate latency well



Source: Google (https://ai.googleblog.com/2018/04/introducing-cvpr-2018-on-device-visual.html)

# of weights ***alone*** is not a good metric for energy (**All data types** should be considered)



Energy breakdown of GoogLeNet

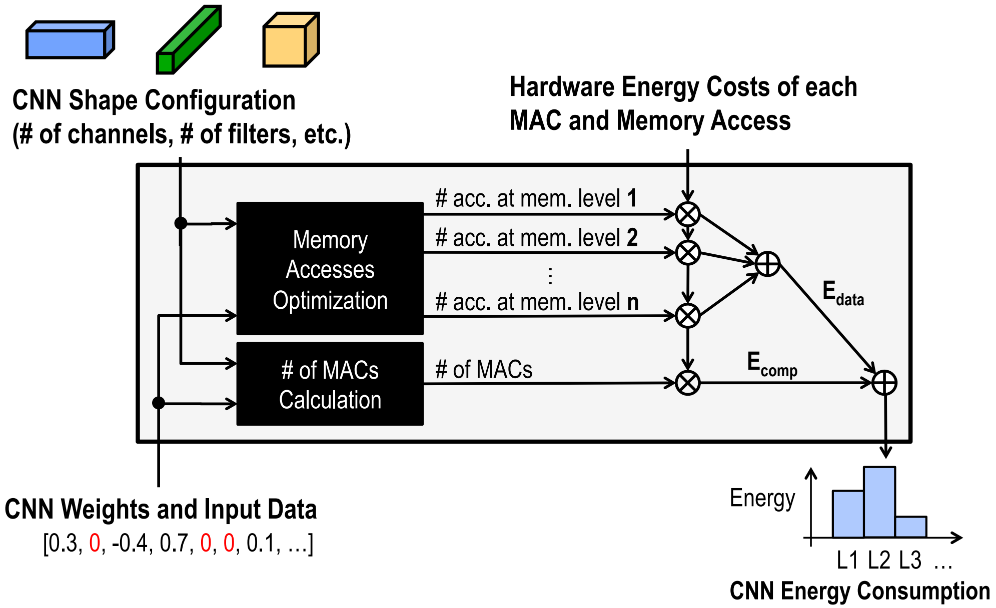
https://energyestimation.mit.edu/

[**Yang**, *CVPR* 2017]

Designing Energy-Efficient DNN Models

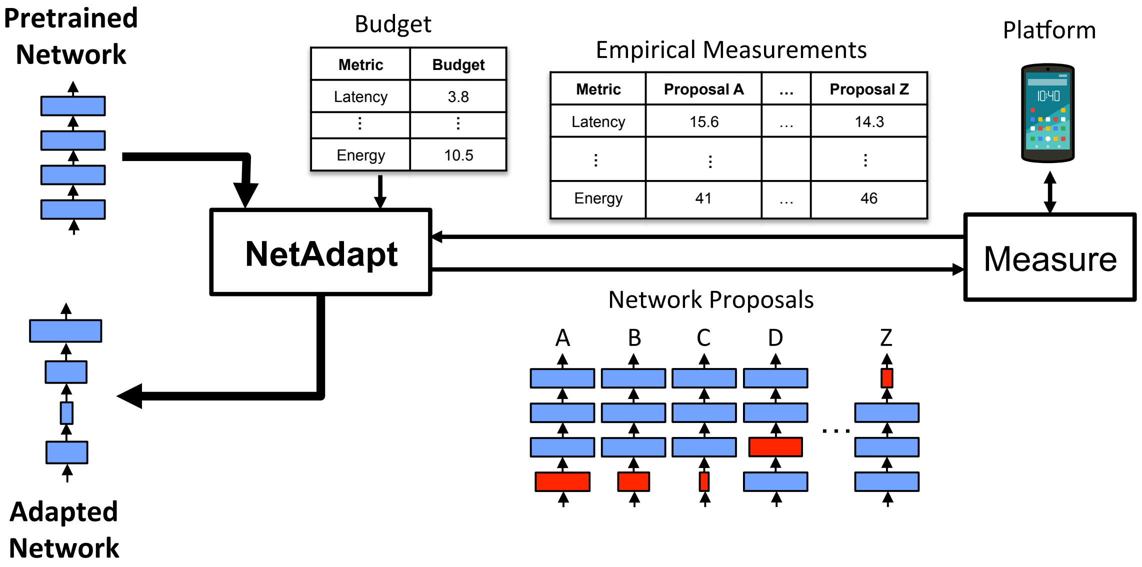
Directly integrate hardware metrics into algorithm design

Energy-Aware Pruning



[**Yang**, *CVPR* 2017]
Pruned models available at
http://eyeriss.mit.edu/energy.html

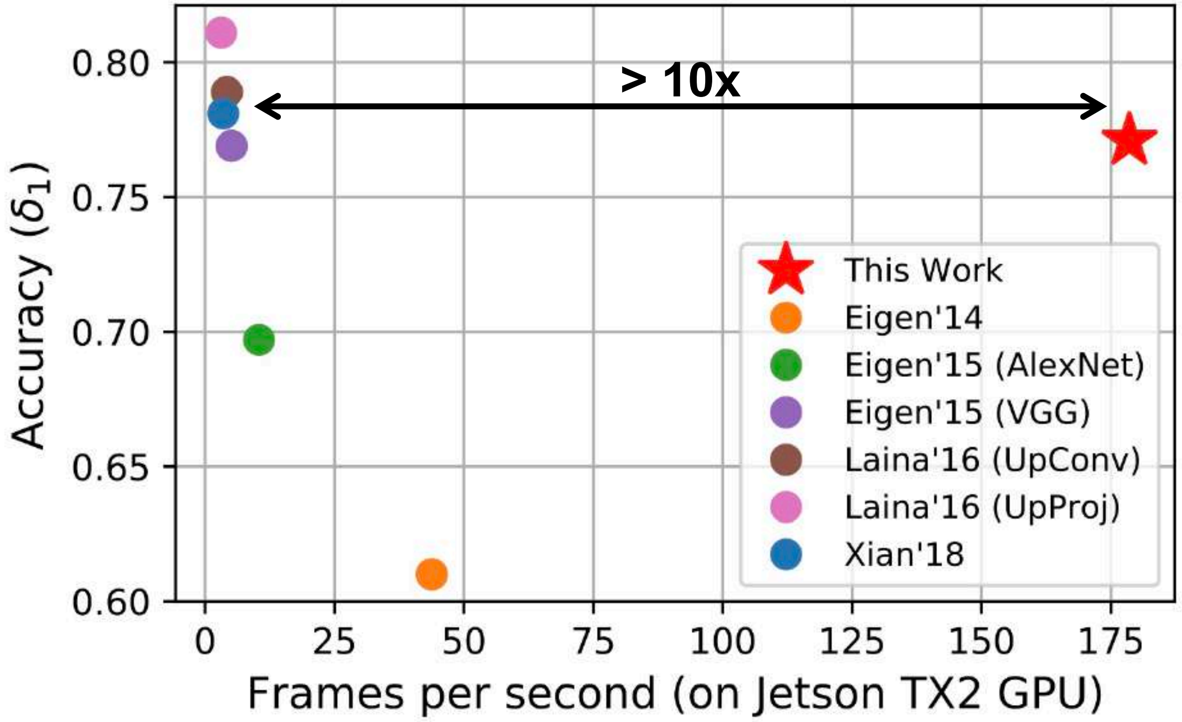
NetAdapt: Platform-Aware DNN



[**Yang**, *ECCV* 2018]
Code available at http://netadapt.mit.edu
In collaboration with Google's Mobile Vision Team

FastDepth: Fast Monocular Depth Estimation

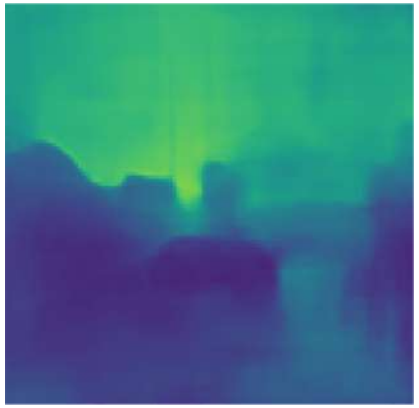
Depth estimation from a single RGB image desirable, due to the relatively low cost and size of monocular cameras.



Configuration: Batch size of one (32-bit float)

RGB

Prediction



~40fps on an iPhone

Models available at http://fastdepth.mit.edu

[**Wofk*, Ma***, *ICRA* 2019]

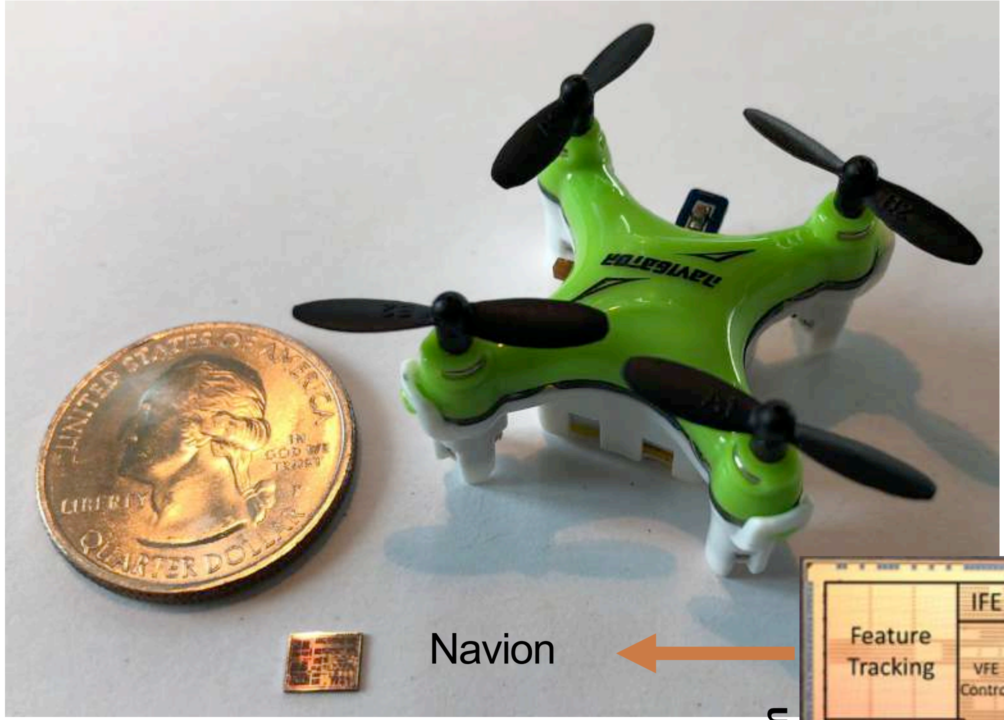
Understanding Accuracy → Application

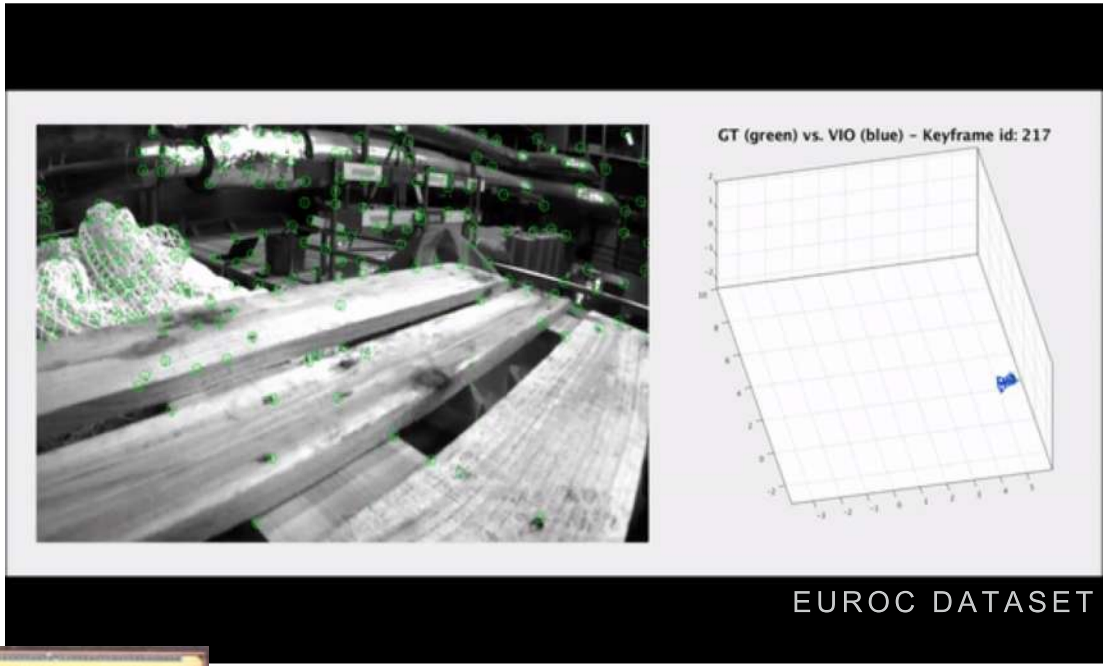
Faculty at MIT (2013 - present)

Goal: Understand what is an acceptable accuracy tradeoff, which is application dependent

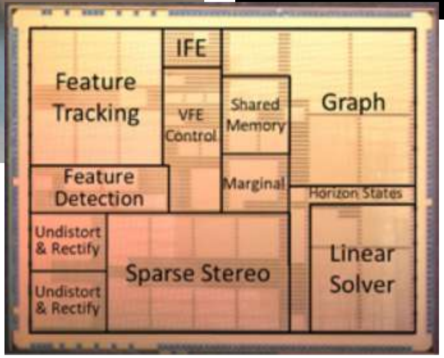
Robot Localization

Determine location/orientation of robot from images and IMU (also used for AR/VR)





4mm



5mm

Navion Project Website
http://navion.mit.edu

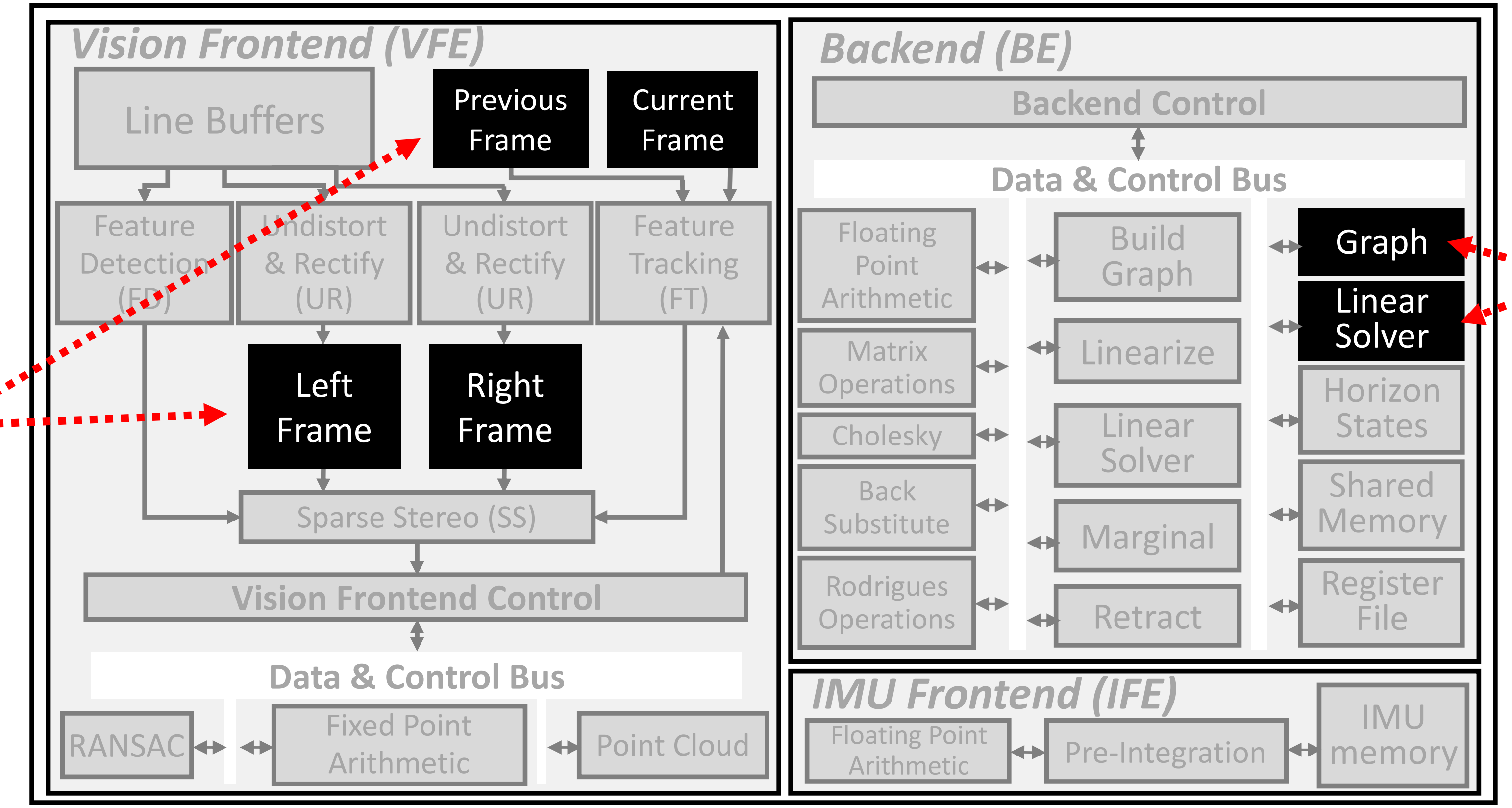
Consumes **684× and 1582×** less energy than mobile and desktop CPUs, respectively

[**Zhang**, *RSS* 2017], [**Suleiman**, *VLSI-C* 2018]

Key Methods to Reduce Data Size

***Navion:** Fully integrated system – no off-chip processing or storage*

Apply Low Cost Frame Compression



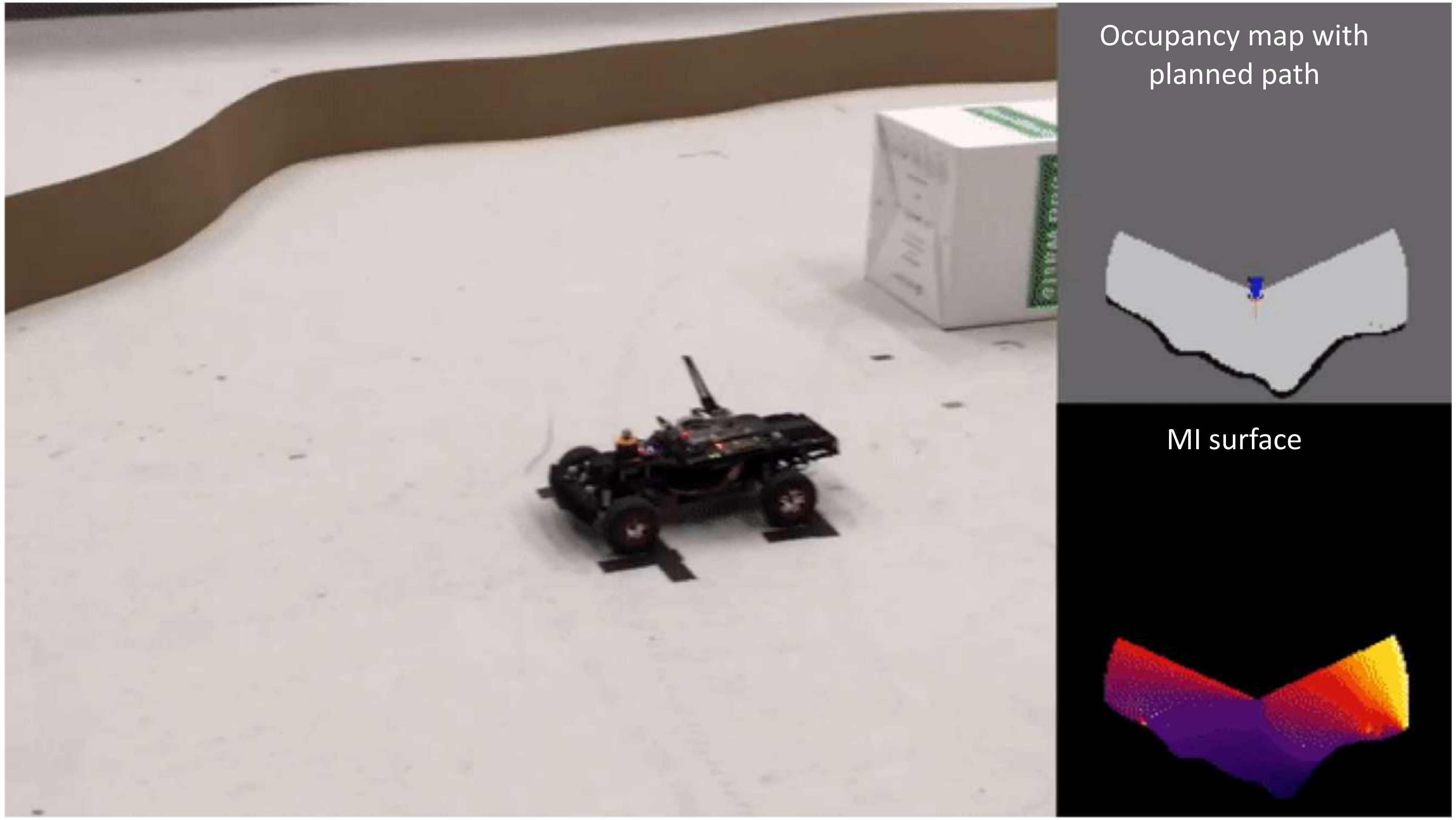
Exploit Sparsity in Graph and Linear Solver

Use **compression** and **exploit sparsity** to reduce memory down to 854kB

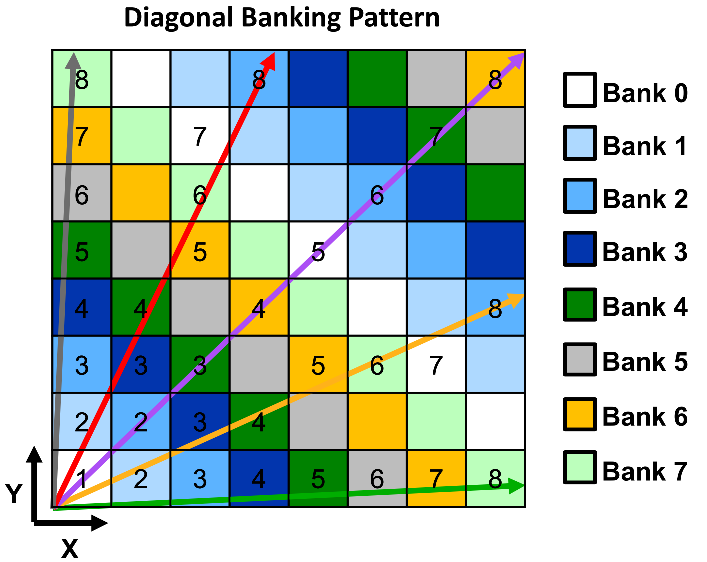
[**Suleiman**, *VLSI-C* 2018] **Best Student Paper Award**

Robot Exploration

Decide where to go by computing Shannon Mutual Information (MI)



[**Zhang**, *ICRA* 2019], [**Henderson**, *ICRA* 2020]



Compute the mutual information for an **entire map** of 20m x 20m at 0.1m resolution **in under a second** → a 100x speed up versus CPU for 1/10th of the power.

[**Li**, *RSS* 2019]

Monitoring Neurodegenerative Disorders

Dementia affects 50 million people worldwide today (75 million in 10 years) [World Alzheimer's Report]

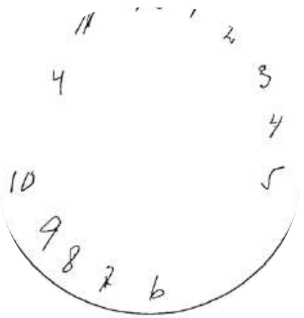
Mini-Mental State Examination (MMSE)

Q1. What is the year? Season? Date?

Q2. Where are you now? State? Floor?

Q3. Could you count backward from 100 by sevens? (93, 86, …)

Clock-drawing test



Agrell et al. *Age and Ageing*, 1998.

- Neuropsychological assessments are **time consuming** and **require a trained specialist**
- Repeat **medical assessments** are **sparse**, mostly **qualitative**, and suffer from **high retest variability**

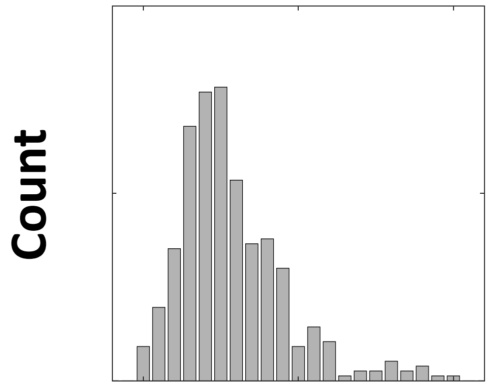
[Joint work with Thomas Heldt and Charlie Sodini]

Use Eye Movements for Quantitative Evaluation

Eye movements can be used to **quantitatively evaluate severity, progression or regression** of neurodegenerative diseases

Eye movements

Smartphone



We are investigating how to perform eye movement tests on a smart phone in order to **enable low-cost, in-home measurements**

[**Saavedra Peña**, *EMBC* 2018] [**Lai**, *ICIP* 2018]

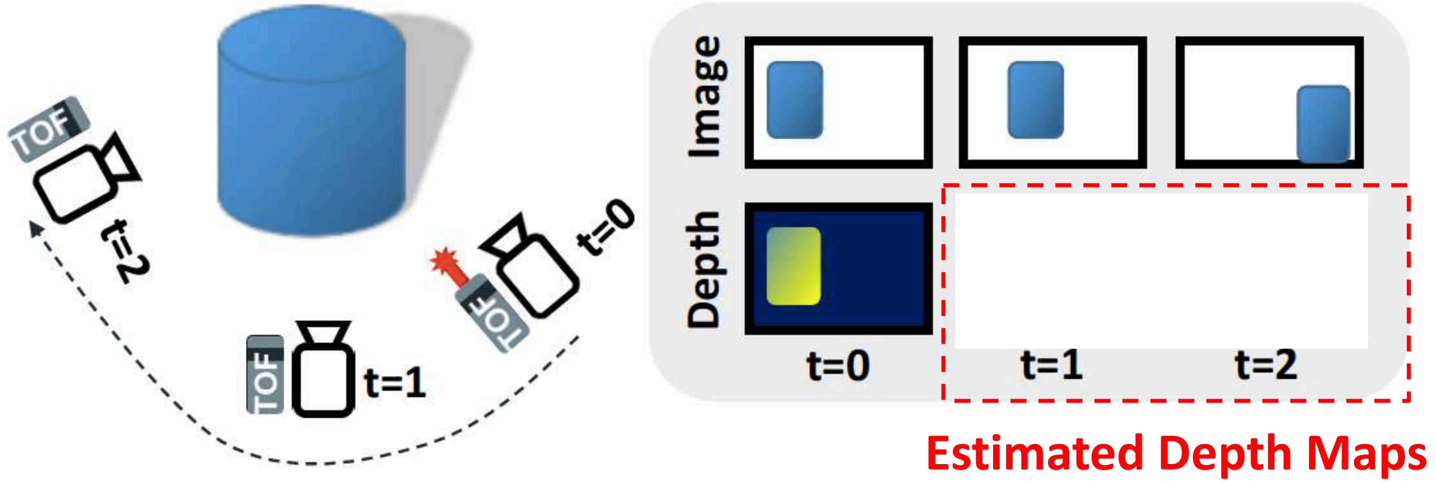
Consider the Entire System

Faculty at MIT (2013 - present)

Goal: Optimized energy efficiency of the ***entire system***

Low Power 3D Time of Flight Imaging

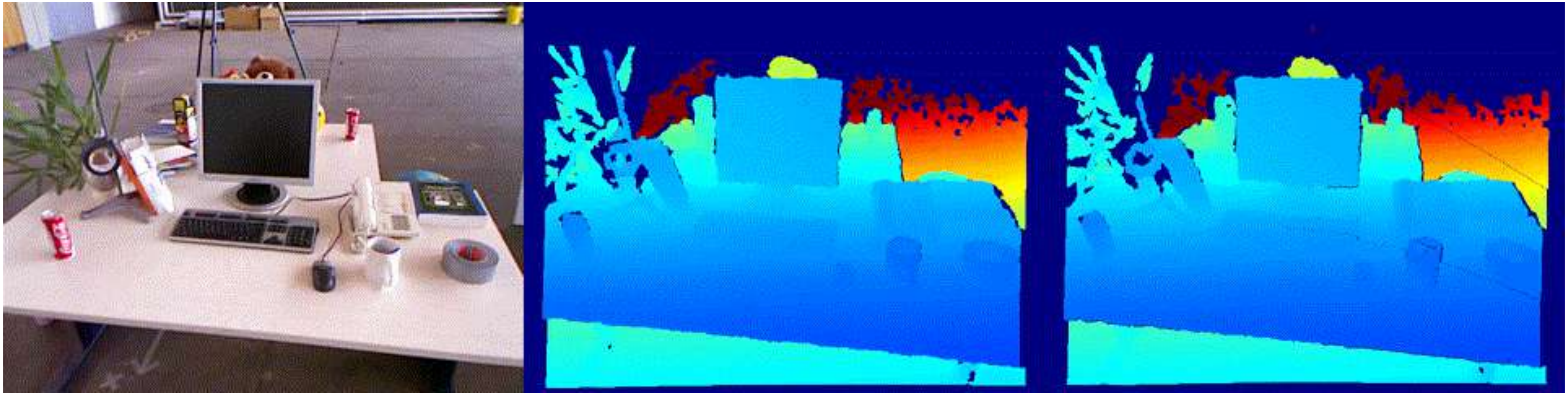
- Pulsed Time of Flight: Measure distance using round trip time of laser light for each image pixel
  - **Illumination + Imager Power: 2.5 – 20 W for range from 1 - 8 m**
- Use computer vision techniques and passive images to estimate changes in depth without turning on laser
  - **CMOS Imaging Sensor Power: < 350 mW**



Real-time Performance on Embedded Processor
VGA @ 30 fps on Cortex-A7
(< 0.5W active power)

[**Noraky**, *ICIP* 2017], [**Noraky**, *TCSVT 2020*]

Results of Low Power Depth ToF Imaging



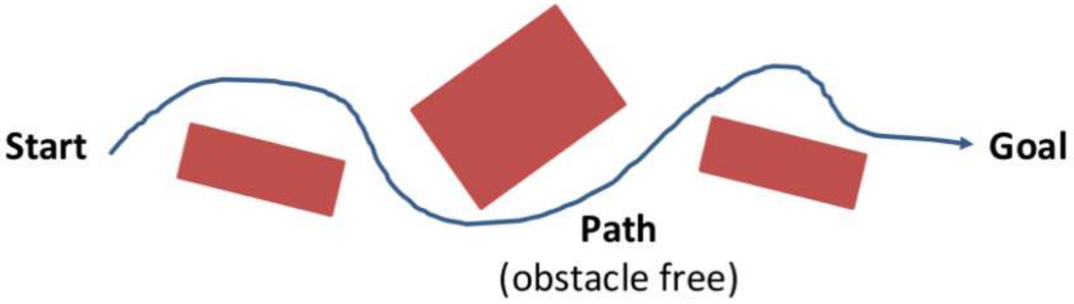
RGB Image | Depth Map **Ground Truth** | Depth Map **Estimated**

Mean Relative Error: 0.7%
Duty Cycle (on-time of laser): 11%

[**Noraky**, *ICIP* 2017], [**Noraky**, *TCSVT 2020*]

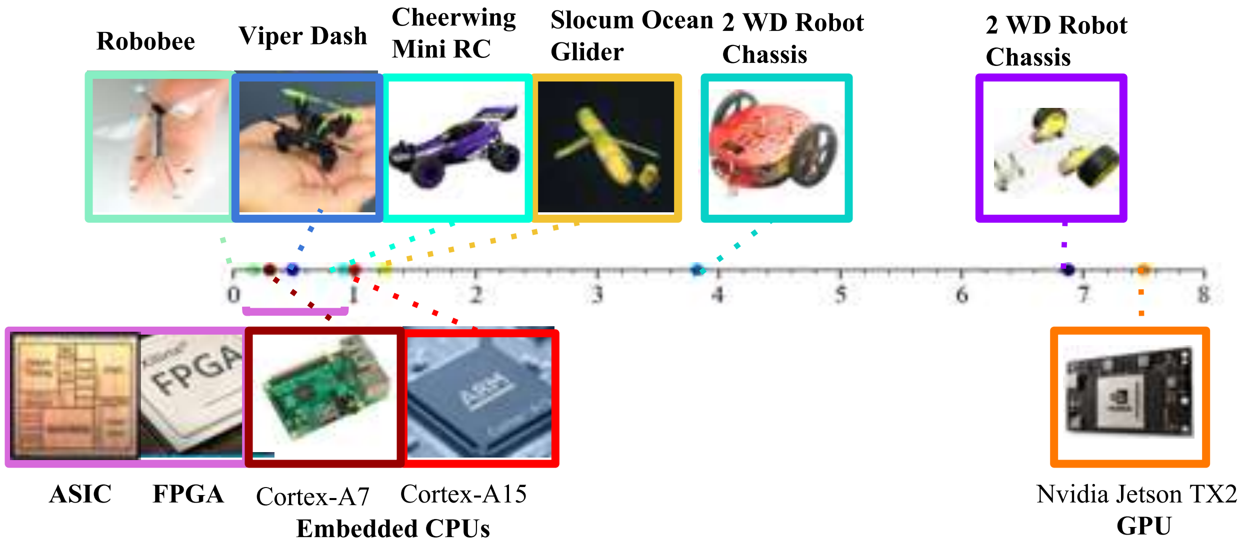
Balancing Actuation and Computing Energy

Motion Planning
Find a feasible (obstacle-free) path
[typically optimize for shortest path]



Energy to move 1 more meter (P_a/v [W/(m/s)])

Low-power Robotics
Actuation and computing energy are similar order of magnitude

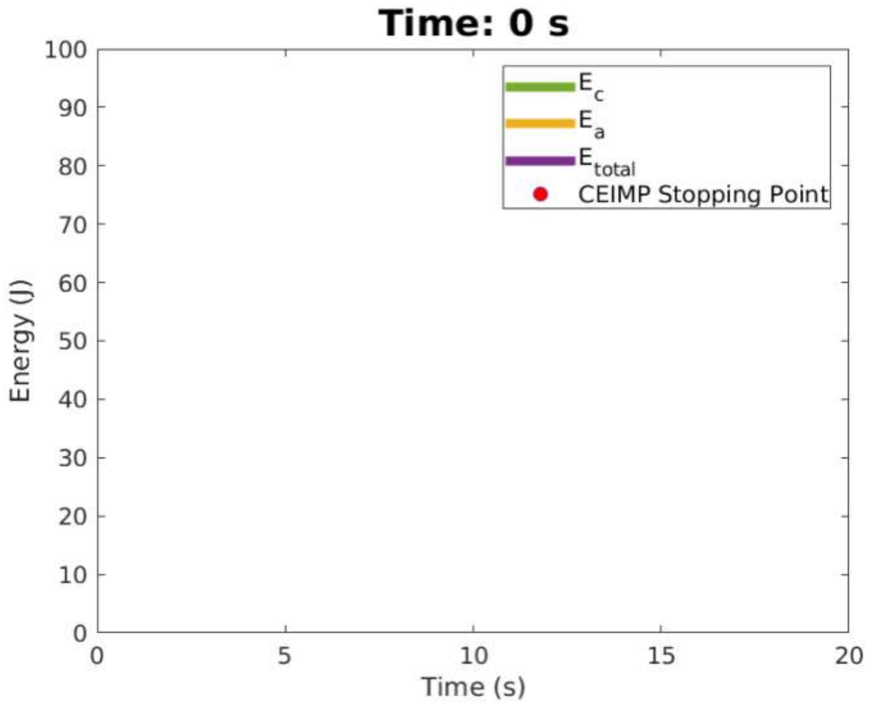


Energy to compute 1 more second (P_c [W])

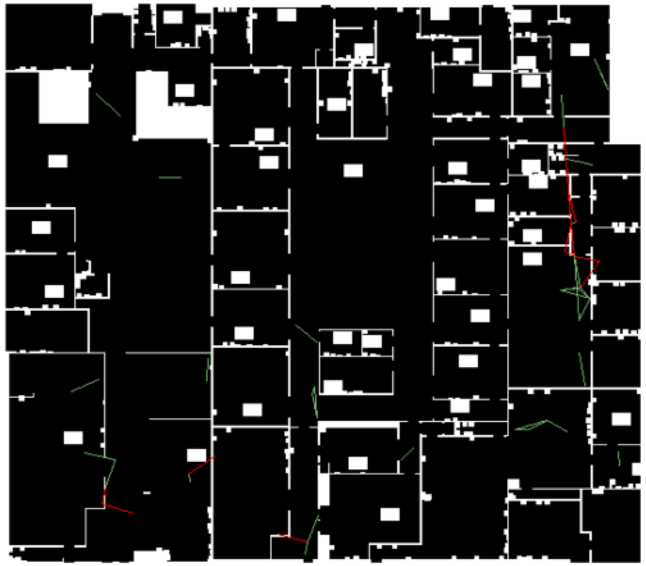
[**Sudhakar**, *ICRA* 2020]

Balancing Actuation and Computing Energy

Baseline
(compute 20,000 samples)



CEIMP



Compute Energy Included Motion Planning (CEIMP)
A framework to balance the energy spent on **computing** a path and the energy spent on **moving** along that path **(Don't think too hard!)**

[**Sudhakar**, *ICRA* 2020]

Key Takeaways

- **Look beyond traditional boundaries**
  – Opportunities lie at the intersection of different areas of research: build bridges
  – Co-design approach applied across different applications
- **How to identify research opportunities**
  – Is this an important problem?
  – What are the main challenges or bottlenecks?
  – What is the skill set needed to address the challenges or bottlenecks?
  – Do I have or can I learn that skill set?
    - Always be learning
    - Collaborate

Acknowledgements

Anantha Chandrakasan

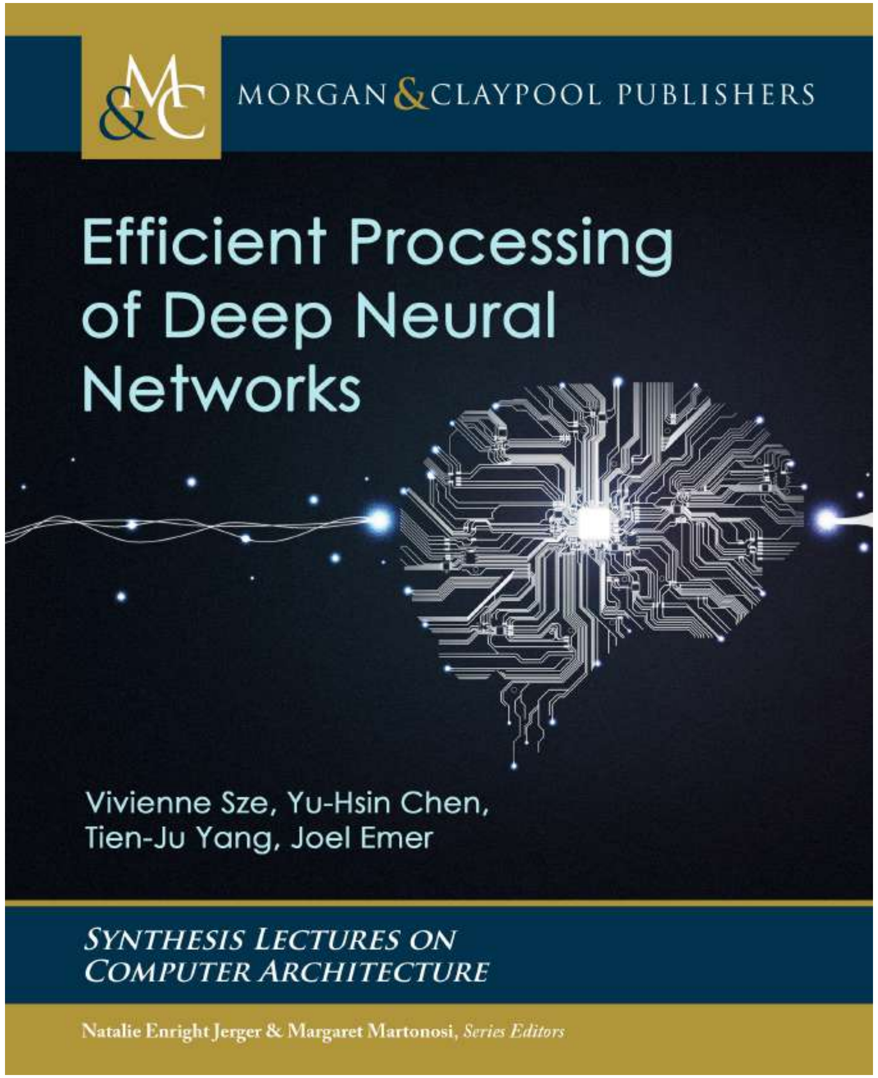
Joel Emer

Thomas Heldt

Sertac Karaman

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

Book on Efficient Processing of DNNs



Part I Understanding Deep Neural Networks
Introduction
Overview of Deep Neural Networks

Part II Design of Hardware for Processing DNNs
Key Metrics and Design Objectives
Kernel Computation
Designing DNN Accelerators
Operation Mapping on Specialized Hardware

Part III Co-Design of DNN Hardware and Algorithms
Reducing Precision
Exploiting Sparsity
Designing Efficient DNN Models
Advanced Technologies

https://tinyurl.com/EfficientDNNBook

Additional Resources

MIT Professional Education Course on

"Designing Efficient Deep Learning Systems"

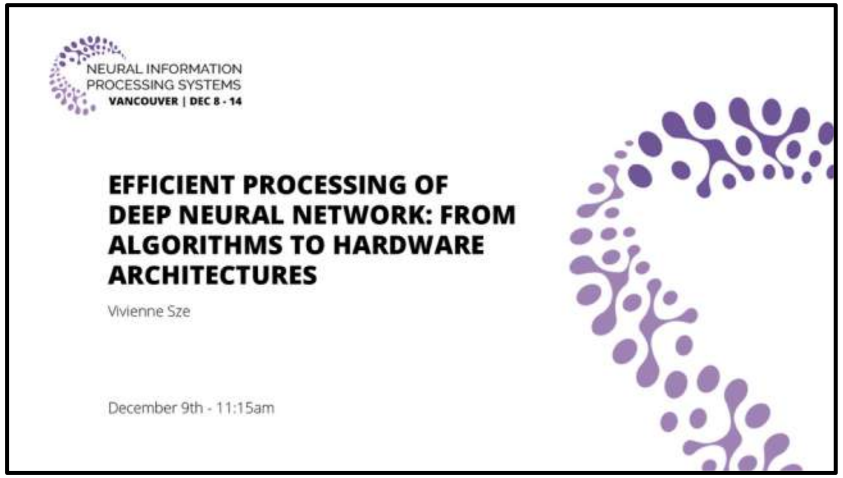
http://shortprograms.mit.edu/dls

Next Offering: July 20-21, 2020 (Live Virtual)

Additional Resources

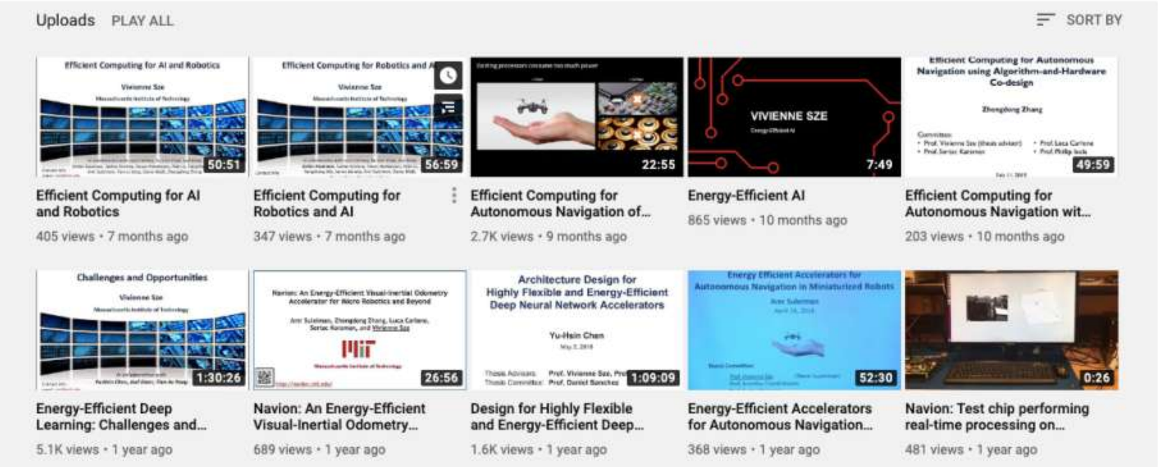
Talks and Tutorial Available Online

https://www.rle.mit.edu/eems/publications/tutorials/



YouTube Channel

EEMS Group – PI: Vivienne Sze



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.

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.
  - Y. N. Wu, J. S. Emer, V. Sze, "Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs," International Conference on Computer Aided Design (ICCAD), November 2019. http://accelergy.mit.edu/
  - Y. N. Wu, V. Sze, J. S. Emer, "An Architecture-Level Energy and Area Estimator for Processing-In-Memory Accelerator Designs," to appear in IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), April 2020.
  - 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. http://netadapt.mit.edu
  - D. Wofk*, F. Ma*, T.-J. Yang, S. Karaman, V. Sze, "FastDepth: Fast Monocular Depth Estimation on Embedded Systems," IEEE International Conference on Robotics and Automation (ICRA), May 2019. http://fastdepth.mit.edu/
- **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

- **Fast Shannon Mutual Information for Robot Exploration**
  - **Project website:** http://lean.mit.edu

  - 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," to appear in International Journal of Robotics Research (IJRR). http://arxiv.org/abs/1905.02238
  - T. Henderson, V. Sze, S. Karaman, "An Efficient and Continuous Approach to Information-Theoretic Exploration," IEEE International Conference on Robotics and Automation (ICRA), May 2020.

- **Monitoring Neurodegenerative Disorders Using a Phone**

  - H.-Y. Lai, G. Saavedra Peña, C. Sodini, T. Heldt, V. Sze, "Enabling Saccade Latency Measurements with Consumer-Grade Cameras," IEEE International Conference on Image Processing (ICIP), October 2018.
  - G. Saavedra Peña, H.-Y. Lai, V. Sze, T. Heldt, "Determination of saccade latency distributions using video recordings from consumer-grade devices," IEEE International Engineering in Medicine and Biology Conference (EMBC), 2018.
  - H.-Y. Lai, G. Saavedra Peña, C. Sodini, V. Sze, T. Heldt, "Measuring Saccade Latency Using Smartphone Cameras," IEEE Journal of Biomedical and Health Informatics (JBHI), March 2020.

References

- **Low Power Time of Flight Imaging**
  - J. Noraky, V. Sze, "Low Power Depth Estimation of Rigid Objects for Time-of-Flight Imaging," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2019.
  - J. Noraky, V. Sze, "Depth Map Estimation of Dynamic Scenes Using Prior Depth Information," arXiv, February 2020. https://arxiv.org/abs/2002.00297
  - J. Noraky, V. Sze, "Depth Estimation of Non-Rigid Objects For Time-Of-Flight Imaging," IEEE International Conference on Image Processing (ICIP), October 2018.
  - J. Noraky, V. Sze, "Low Power Depth Estimation for Time-of-Flight Imaging," IEEE International Conference on Image Processing (ICIP), September 2017.
- **Balancing Actuation and Computation**
  - **Project website:** http://lean.mit.edu
  - S. Sudhakar, S. Karaman, V. Sze, "Balancing Actuation and Computing Energy in Motion Planning," IEEE International Conference on Robotics and Automation (ICRA), May 2020