Emerging AI Applications: Moving Beyond ResNet50 on ImageNet

Amir Gholami, Michael Mahoney, Kurt Keutzer

and Pallas Group: Suresh Krishna, Ravi Krishna, Zhewei Yao, Zhen Dong,

Sehoon Kim, Tianren Gao, Bohan Zhai , Ani Nrusimha

Opening Keynote, Intel System Architecture Summit (ISAS), Feb. 2021

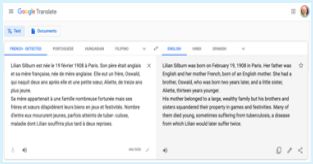
Our Group's Focus:
All Deep Learning All The Time

Image Classification

Object Detection

Image Segmentation

Computer Vision and Core ML

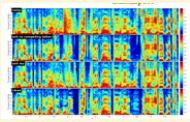


Translation

Question answering

Document Understanding

Natural Language Processing

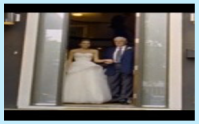


Audio Enhancement

Call-center Sentiment Analysis

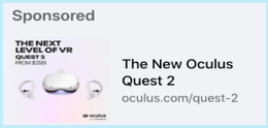
Speech Recognition

Audio Analysis



Video Recommendation

Music Recommendation



Ad Recommendation

Rec Systems

State-of-the-Art Solutions
Typically Rely on one DNN (or a few)

Bandwidth Bound Applications

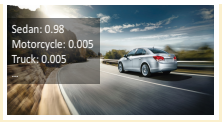
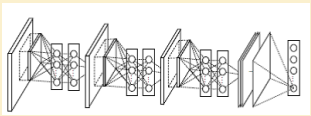


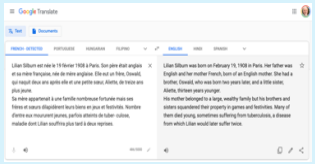
Image Classification

Object Detection

Image Segmentation

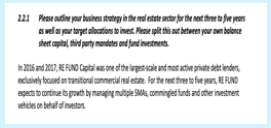


Convolutional NN

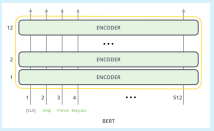


Translation

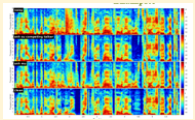
Question answering



Document Understanding



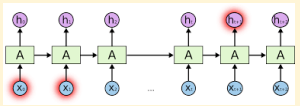
Transformer



Audio Enhancement

Call-center Sentiment Analysis

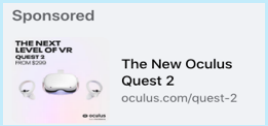
Speech Recognition



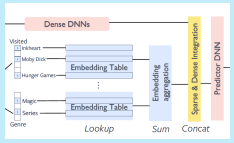
Recurrent NN (CTC/RNN-T)

Video Recommendation

Music Recommendation



Ad Recommendation



DLRM

Most Focus has been on CV

Bandwidth Bound Applications

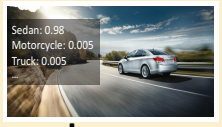
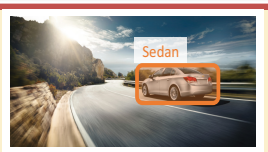
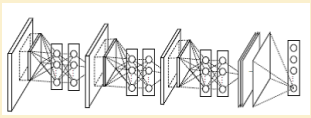


Image Classification

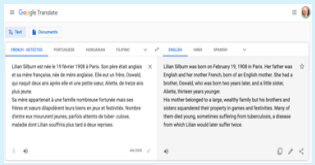


Object Detection

Image Segmentation



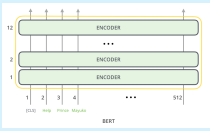
Convolutional NN



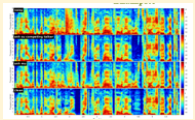
Translation

Question answering

Document Understanding



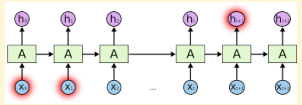
Transformer



Audio Enhancement

Call-center Sentiment Analysis

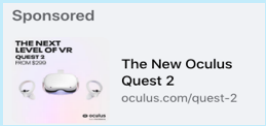
Speech Recognition



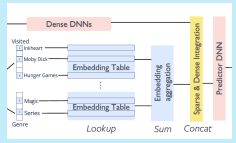
Recurrent NN (CTC/RNN-T)

Video Recommendation

Music Recommendation



Ad Recommendation



DLRM

CV workloads are not representative of other emerging AI applications

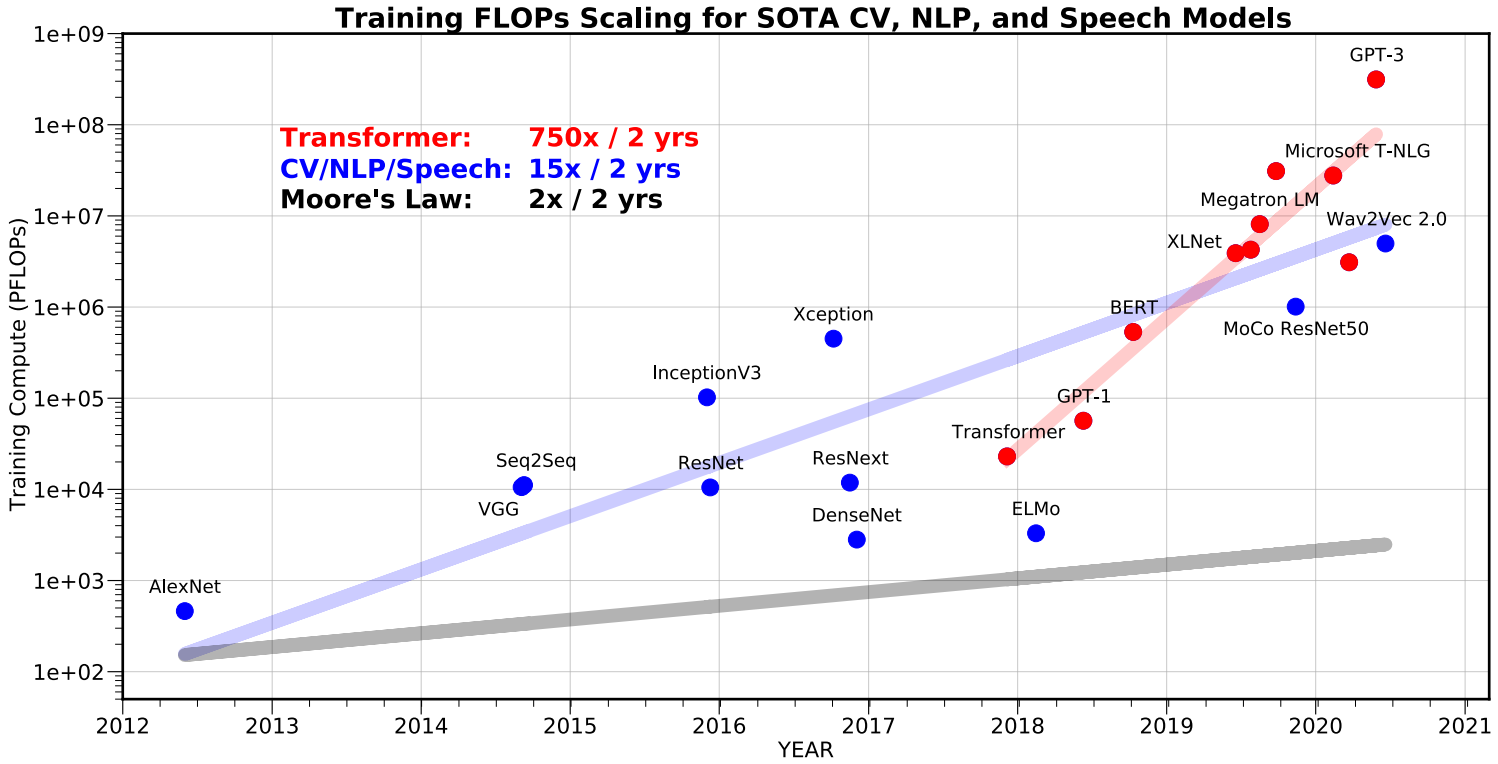
Executive Summary: New Opportunities for DSA

- **Emerging AI applications with low arithmetic intensity**
  - Recommendation Systems, that need DSA with
    - Large Memory Systems
    - Fast Interconnect
    - Efficient Prefetching and Cache Hierarchy
- **AI at the Edge:**
  - All AI domains (CV, NLP, RecSys, ASR, Robotics/RL, etc.)
    - Low-precision Inference
    - Unified software interface for better programmability
- **Emerging AI Optimization Algorithms:**
  - Moving beyond SGD based training to second-order methods
  - Need for DSA that supports fast Randomized algorithms
  - Important applications for Scientific ML/Computing

Executive Summary: New Opportunities for DSA

- **Emerging AI applications with low arithmetic intensity**
  - Recommendation Systems, that need DSA with
    - Large Memory Systems
    - Fast Interconnect
    - Prefetching/Caching Hierarchy
- AI at the Edge:
  - All AI domains (CV, NLP, RecSys, ASR, Robotics/RL, etc.)
    - Low-precision Inference
    - Unified software interface for programmability
    - HW/NN Co-design
- Emerging AI Optimization Algorithms:
  - Moving beyond SGD based training to second-order methods
  - Need for DSA that supports fast Randomized algorithms
  - Important applications for Scientific ML/Computing

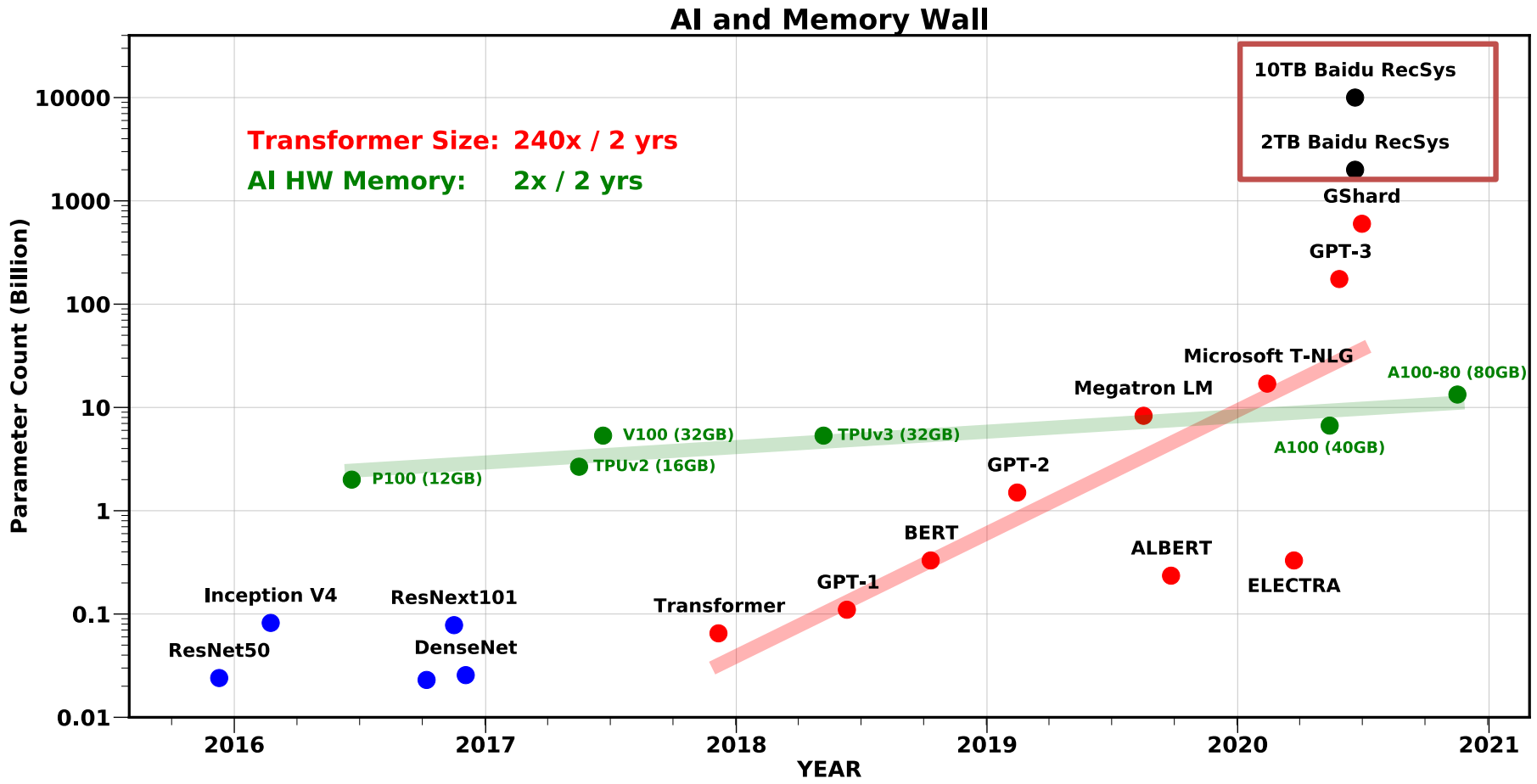
AI and Compute
(Not the Entire Picture)



- The wrong way to interpret this trend is to only focus on increasing peak FLOPs of AI accelerators => **Not optimal** for emerging AI applications

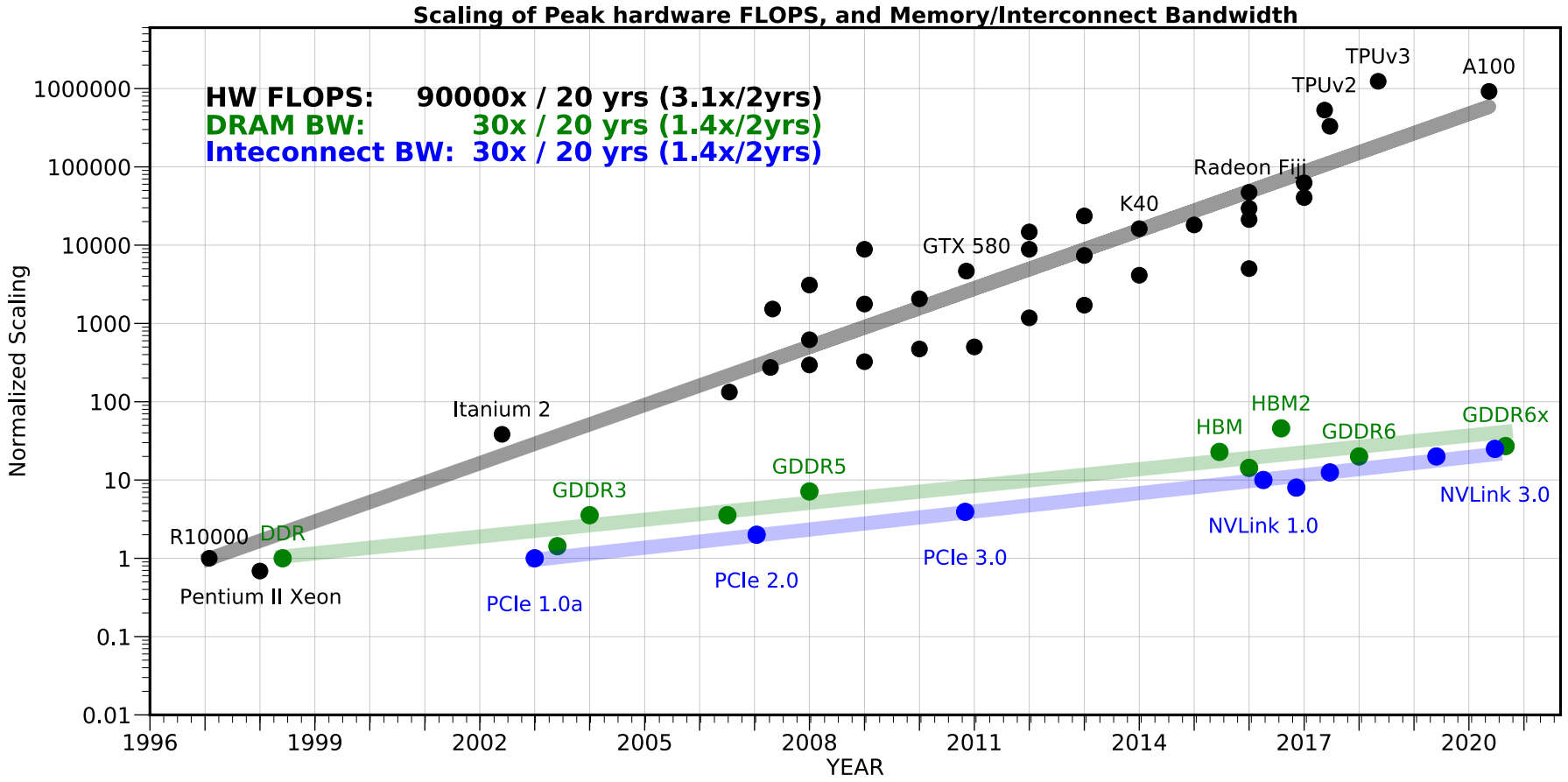
Amir Gholami, Zhewei Yao, Sehoon Kim, Michael W. Mahoney, Kurt Keutzer, AI and Memory Wall, Riselab Medium Blogpost, 2021.

AI and Memory Wall



Amir Gholami, Zhewei Yao, Sehoon Kim, Michael W. Mahoney, Kurt Keutzer, AI and Memory Wall, Riselab Medium Blogpost, 2021.

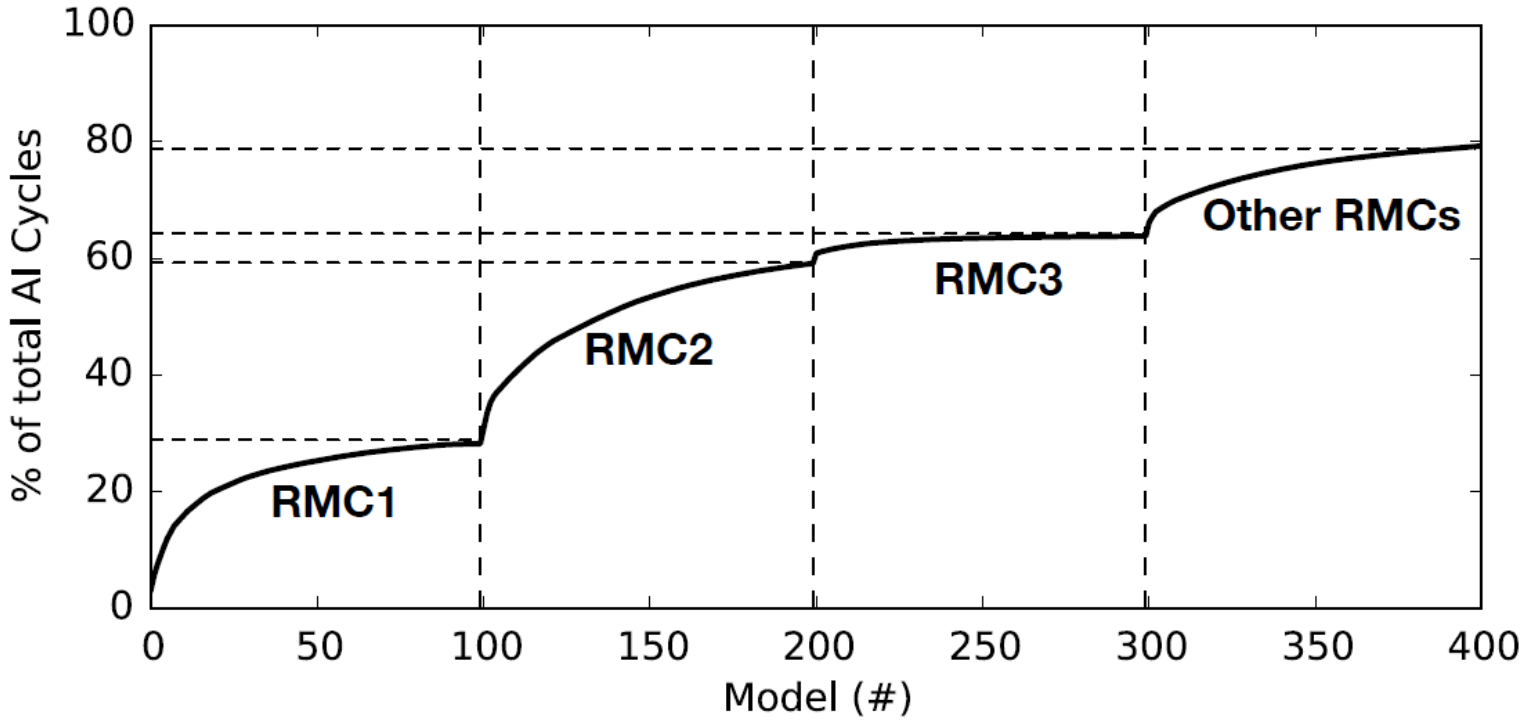
AI and Memory Wall



Amir Gholami, Zhewei Yao, Sehoon Kim, Michael W. Mahoney, Kurt Keutzer, AI and Memory Wall, Riselab Medium Blogpost, 2021.

Majority of AI Workloads in Datacenter is RecSys (not NLP or CV)

- Recommendation systems, account for more than 80% of AI Cycles in FB.



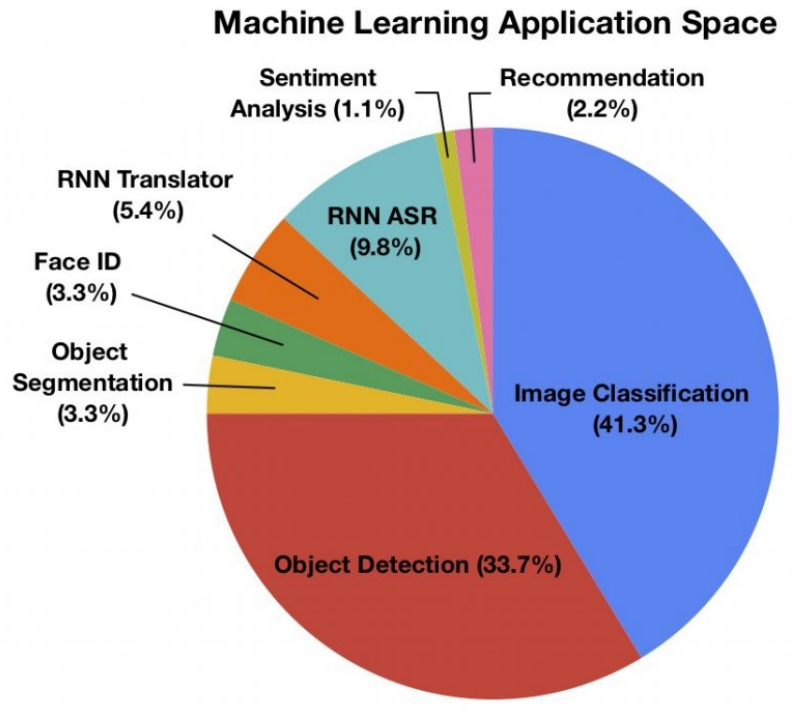
Gupta U, et al. The Architectural Implications of Facebook's DNN-based Personalized Recommendation, HPCA'20.

Majority of AI Workloads in Datacenter is RecSys (not NLP or CV)

- Most academic papers have been focusing on optimizing:
  - CV: **81.6% papers**
  - NLP: **6.5% papers**

- **Only 2%** of papers in top conferences are considering **recommendation systems**

But this is rapidly changing. RecSys is now part of MLCommons/MLPerf benchmark.

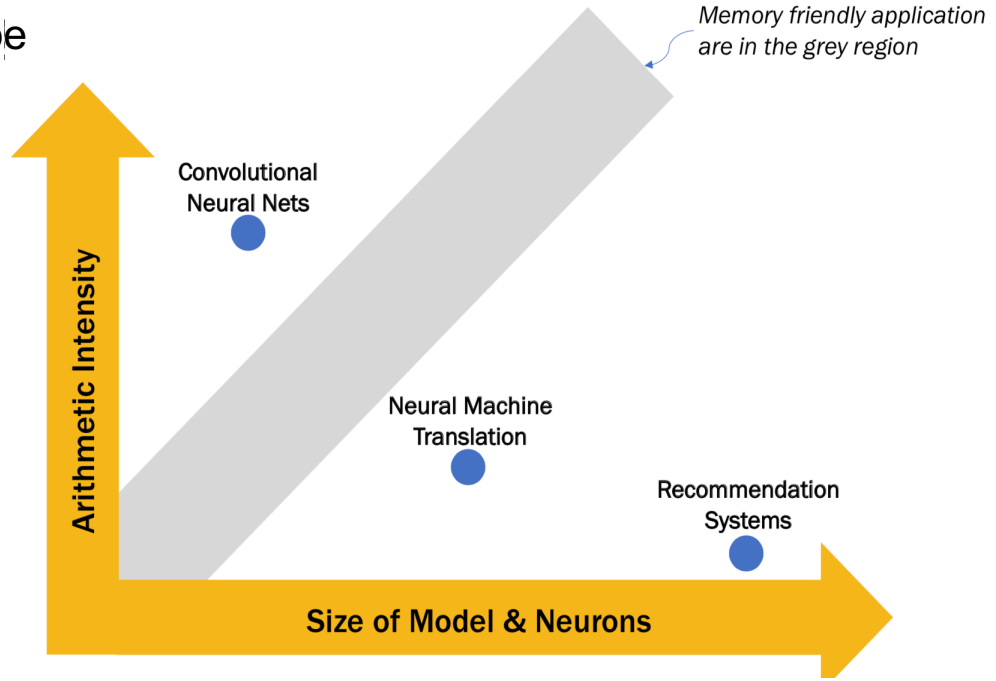


The breakdown of application spaces of ML related papers published in HPCA, ASPLOS, ISCA, and MICRO over the last five years. Source: C. Wu et al.

How does this impact AI Hardware?
Rec Systems have extremely low arithmetic intensity

This trend is changing fast because:

- RecSys models are no longer using old NCF type models
- New models are using DNN based approaches similar to NLP but with much large model sizes
  - Importantly they have orders of magnitude **smaller Arithmetic Intensity**



$$\text{Arithmetic Intensity} = \frac{\#\text{FLOPs}}{\#\text{Memory OPS}}$$

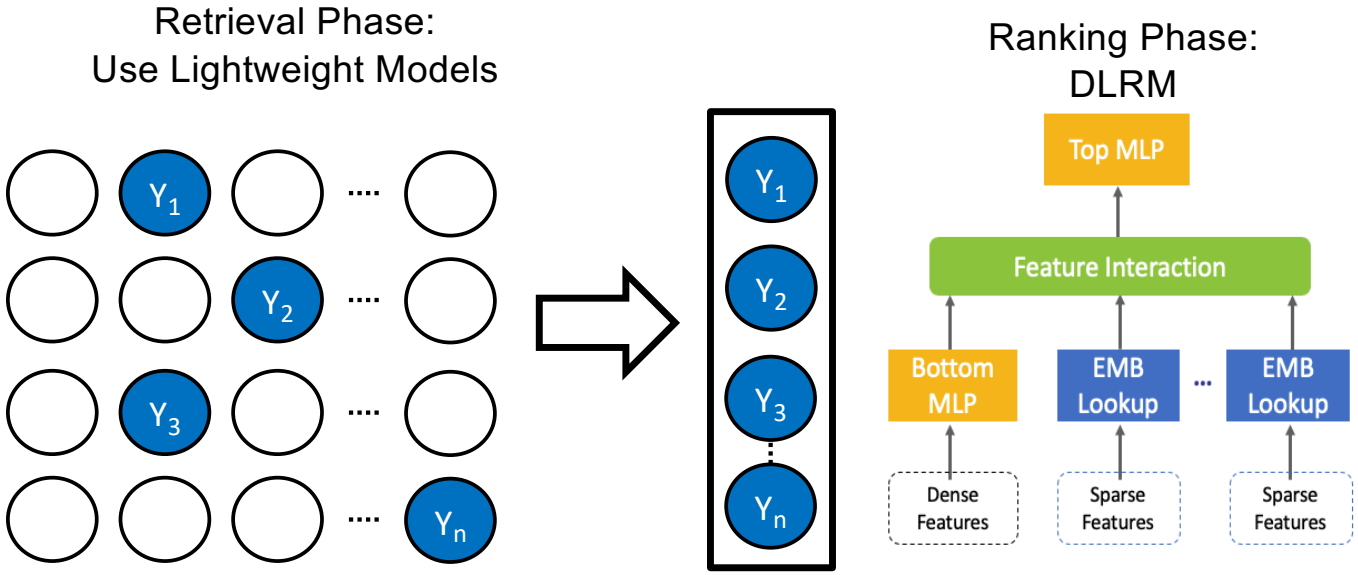
Image: Courtesy of M. Smelyanskiy

Structure of a Recommendation System (DLRM)

- Find the most appropriate ad (Y), for a user (u) with a past history (X_i)

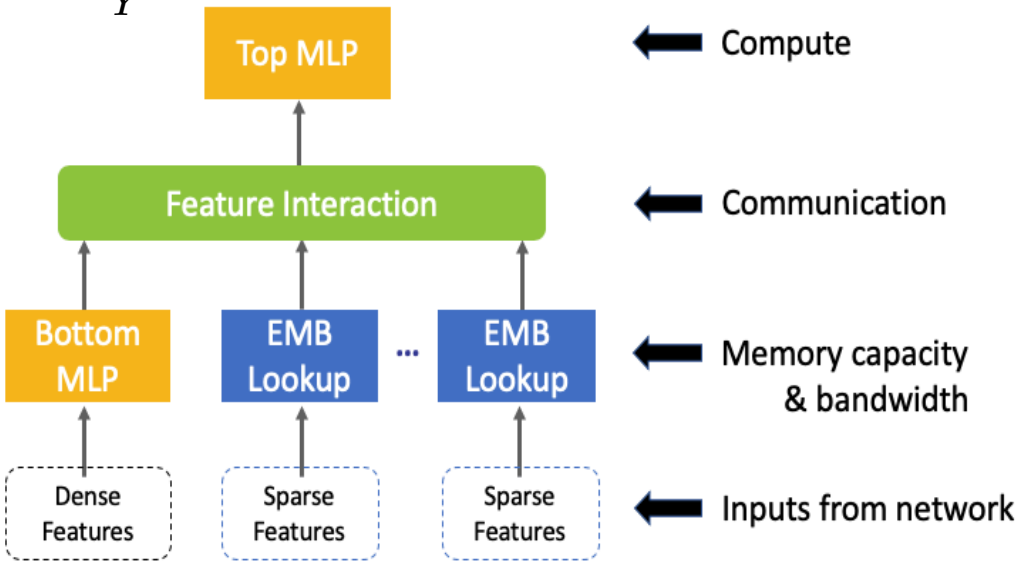
$$\underset{Y}{\operatorname{argmax}} \; P(Y|u, X_i)$$

- This is done in two phases: Retrieval and Ranking



Ranking Phase (DLRM)

$$\underset{Y}{\operatorname{argmax}} \; P(Y|u, X_i)$$



Structure of a modern Recommendation System

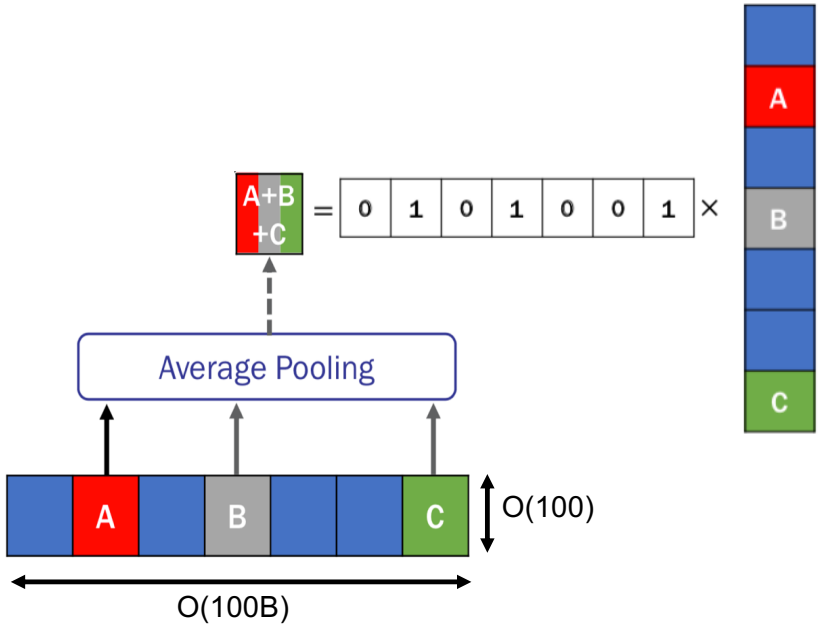
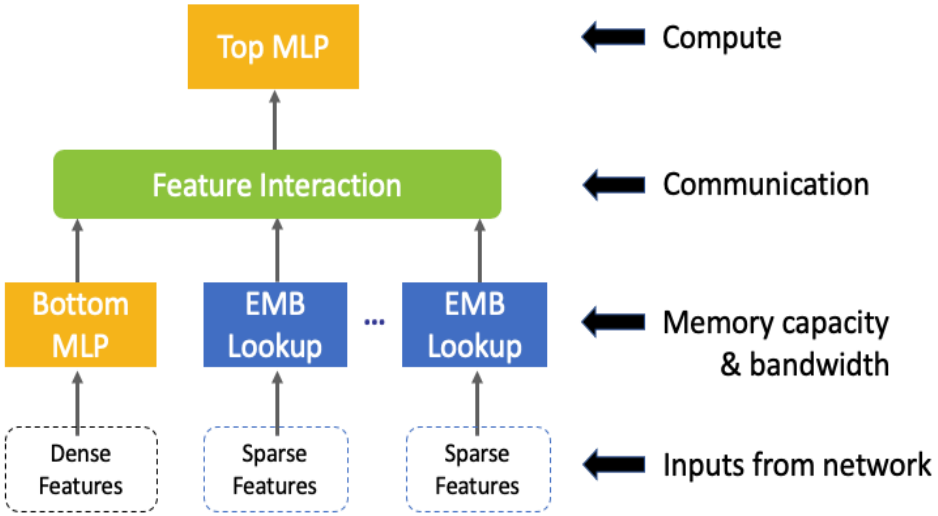


Illustration of Embedding lookup

Naumov M, Kim J, Mudigere D, Sridharan S, Wang X, Zhao W, Yilmaz S, Kim C, Yuen H, Ozdal M, Nair K. Deep Learning Training in Facebook Data Centers: Design of Scale-up and Scale-out Systems. arXiv:2003.09518.

RecSys is orders of magnitude Bigger in size than CV (and NLP)

- Recommendation Systems have very low arithmetic intensity



	#Non-zeros	#Sparse	#Dense	Size (GB)	MPI
A	100	8×10^9	7×10^5	300	100
B	100	2×10^{10}	2×10^4	600	80
C	500	6×10^{10}	2×10^6	2,000	75
D	500	1×10^{11}	4×10^6	6,000	150
E	500	2×10^{11}	7×10^6	10,000	128

Existing model sizes are as large as 10TB!

Mikhail Smelyanskiy, AI at Facebook Datacenter Scale, Invited Lecture in UC Berkeley EE 290, 2020.
Baidu: Distributed Hierarchical GPU Parameter Server for Massive Scale Deep Learning Ads Systems, W. Zhao et al.

Comparison with Other Tasks

Category	Model Type	Model Size (#Params)	Arithmetic Intensity	Maximum Live Activations
Computer Vision	ResNeXt101-32x4-48	43-829M	300 (Avg) 100 (Min)	2-29M
Language	GRU/LSTM/Transformer	10M-1B	2-60	> 100K
Recommendation	Fully Connected	1-10M	20-200	> 10 K
	Embeddings	**> 10 Billion**	**1-2**	> 10 K

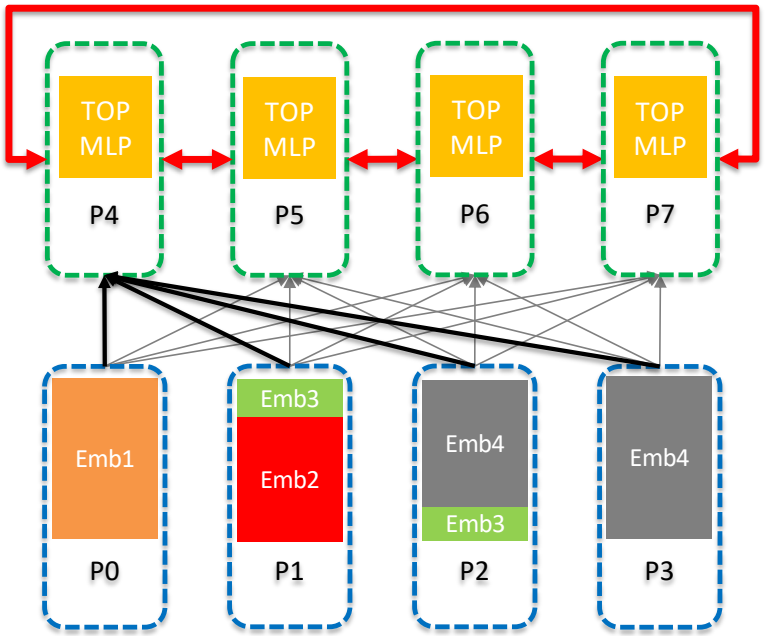
Mikhail Smelyanskiy, AI at Facebook Datacenter Scale, Invited Lecture in UC Berkeley EE 290, 2020.

Training Recommendation Systems

For model parallel training we will have need **allreduce** and **alltoall** communication

- Optimized **HW** needs to have:
  - **Fast interconnect to reduce alltoall communication overhead**
  - **Intelligent Caching and Prefetching**
  - **Large Capacity High Bandwidth Memory**
- We also need optimized ML **algorithms** to enable:
  - **Ultra low precision quantization and fast Structured/Unstructured pruning to speed up inference**
  - **Robust optimization algorithms that are require less tuning and have lower overhead**

All Reduce to aggregate gradients for MLP layers

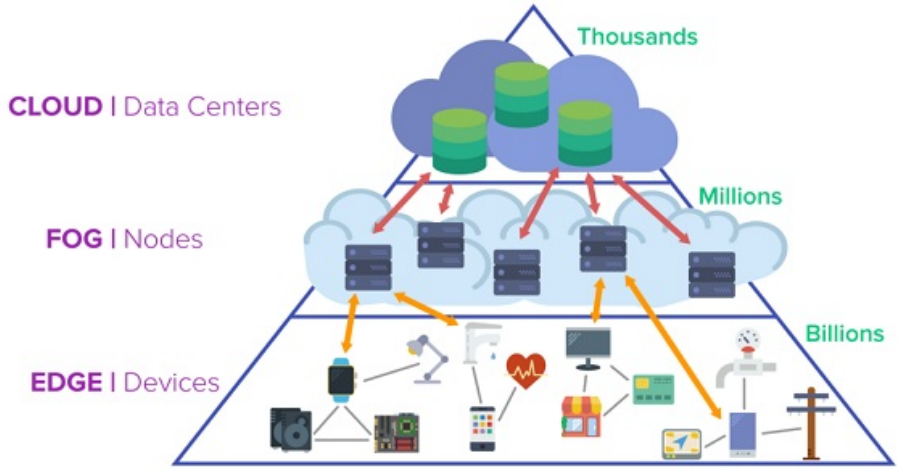


- Gholami, Amir, et al. "Integrated model, batch, and domain parallelism in training neural networks." SPAA'18.
- Yao, Zhewei, et al. "HAWQV3: Dyadic Neural Network Quantization." arXiv preprint arXiv:2011.10680.
- Z. Dong*, Z. Yao*, A. Gholami*, M. Mahoney, K. Keutzer, HAWQ: Hessian Aware Quantization of Neural Networks With Mixed Precision, ICCV'19.
- Z. Dong, Z. Yao, Y. Cai, D. Arfeen, A. Gholami, M. Mahoney, K. Keutzer HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks, NeurIPS'20.
- Shen, Sheng, et al. "Q-bert: Hessian based ultra low precision quantization of BERT." AAAI'20.
- Kim, Sehoon, et al. "I-BERT: Integer-only BERT Quantization." arXiv preprint arXiv:2101.01321 (2021).
- Naumov, Maxim, et al. "Deep learning training in facebook data centers: Design of scale-up and scale-out systems." arXiv preprint arXiv:2003.09518 (2020).
- Krishna S, Krishna R. Accelerating Recommender Systems via Hardware" scale-in". arXiv:2009.05230, 2020.

Executive Summary: New Opportunities for DSA

- Emerging AI applications with low arithmetic intensity
  - Recommendation Systems, that need DSA with
    - Large Memory Systems
    - Fast Interconnect
    - Prefetching/Caching Hierarchy
- **AI at the Edge:**
  - All AI domains (CV, NLP, RecSys, ASR, Robotics/RL, etc.)
    - Low-precision Inference
    - Unified software interface for programmability
    - HW/NN Co-design
- Emerging AI Optimization Algorithms:
  - Moving beyond SGD based training to second-order methods
  - Need for HW that supports fast Randomized algorithms
  - Important applications for Scientific ML/Computing

Moving from the Cloud to the Edge!

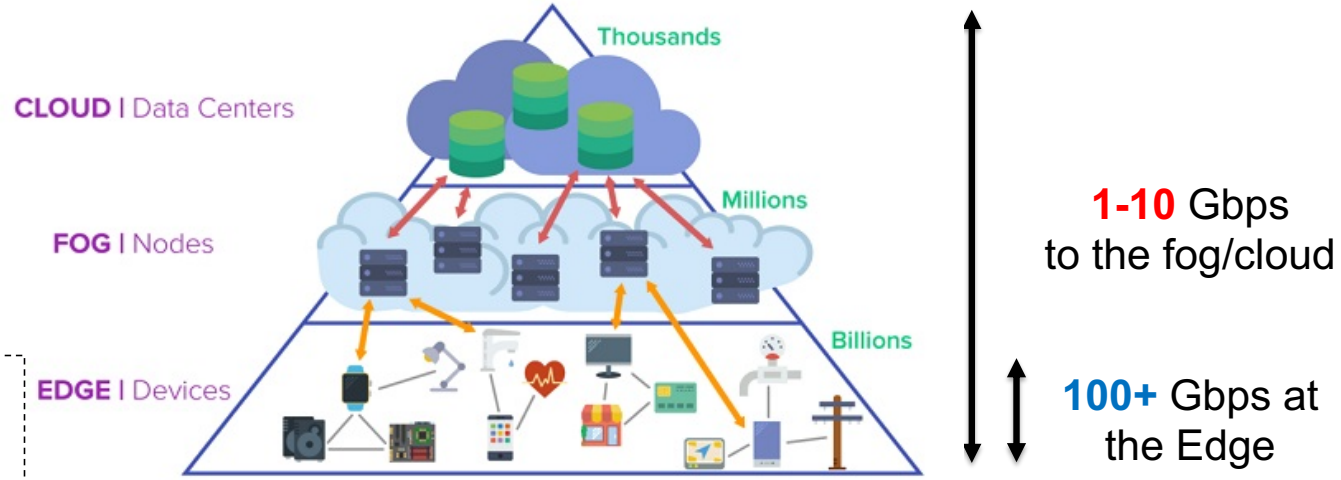


- We observed a big migration to cloud in the past decade.
- We are observing a significant migration back to the device driven by **privacy concerns**, and **need for real-time AI**

Image Credit: Qualcomm, Lance Eliot

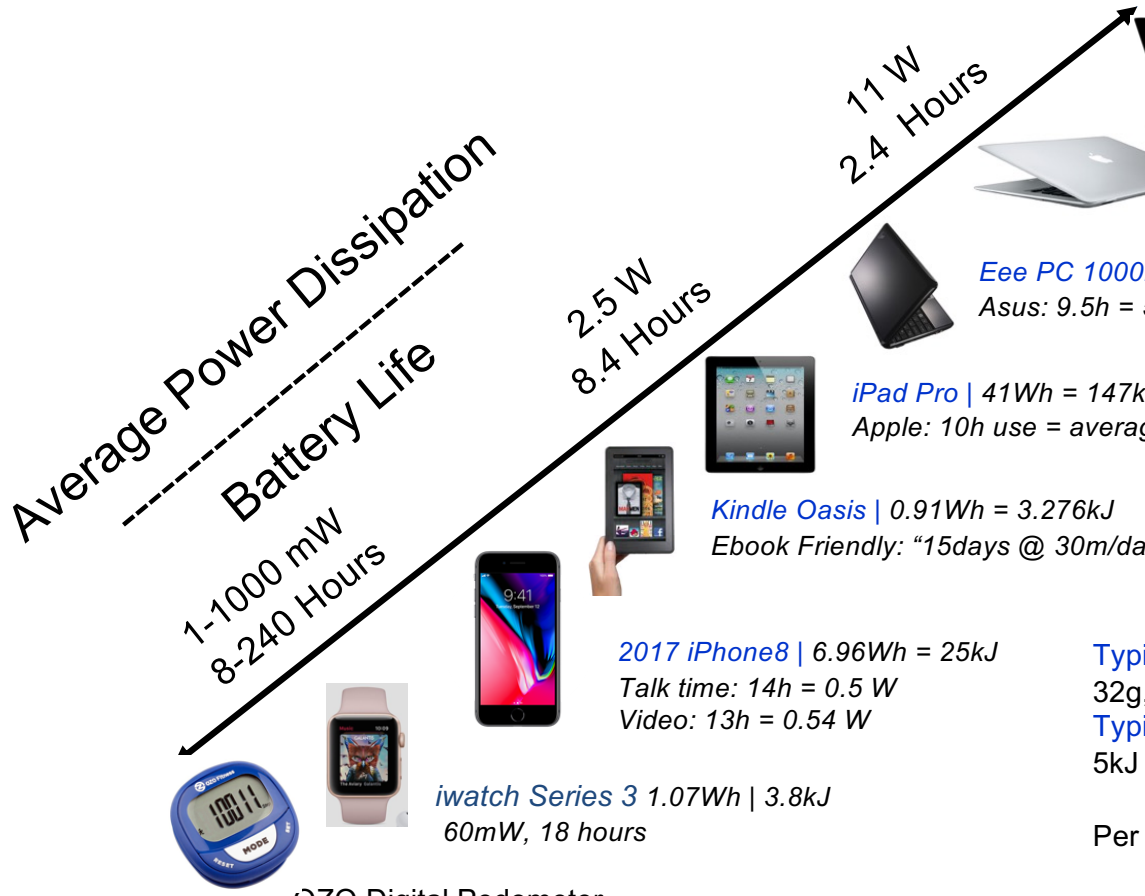
AI at the Edge

- Inference at the server is **unreliable** and challenging for **latency/energy constrained applications**
  - Example: Autonomous cars produce O(10) Gbps data and need latency <100ms under <100 Watts
- **User privacy** can be comprised when data is sent to the cloud
  - Example: More than 25% of smart speaker users do not want their data to be sent to the cloud [1]



[1] Malkin N, Deatrick J, Tong A, Wijesekera P, Egelman S, Wagner D. Privacy attitudes of smart speaker users. Proceedings on Privacy Enhancing Technologies. 2019.

We Want to Operate Across a Broad Range of Hosts at the Edge with Limited Resources



15 inch Macbook Pro |
76Wh = 273.6kJ
Apple: 10h = 7.6 W

13 inch Macbook Air | 54Wh = 194.4kJ
Apple: 12h = 4.5 W

Eee PC 1000HE | 49Wh = 176kJ
Asus: 9.5h = 5.2 W

iPad Pro | 41Wh = 147kJ
Apple: 10h use = average 4.1W

Kindle Oasis | 0.91Wh = 3.276kJ
Ebook Friendly: "15days @ 30m/day" = 7.5h @ 0.12 W average

2017 iPhone8 | 6.96Wh = 25kJ
Talk time: 14h = 0.5 W
Video: 13h = 0.54 W

iwatch Series 3 1.07Wh | 3.8kJ
60mW, 18 hours

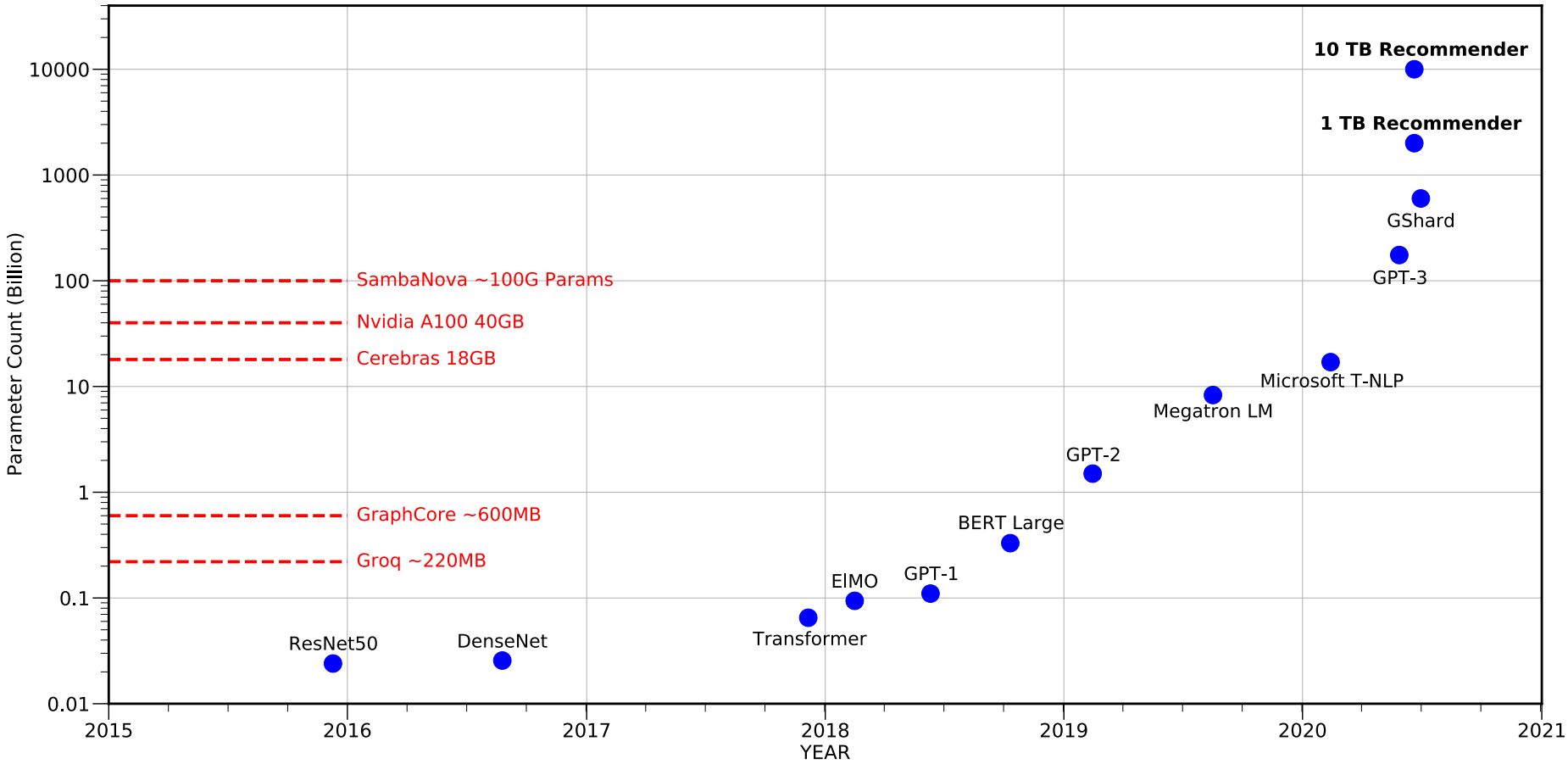
OZO Digital Pedometer
80µW, 0.72Wh | 1 year

Typical handset
32g, 13cc, 5.5Wh = 19.8 kJ
Typical usage
5kJ active + 12kJ standby = 1 battery charge
Per Ljung – Nokia, 2012

1 Wh = 3.6 kJ

Slide: Courtesy of Prof. Keutzer

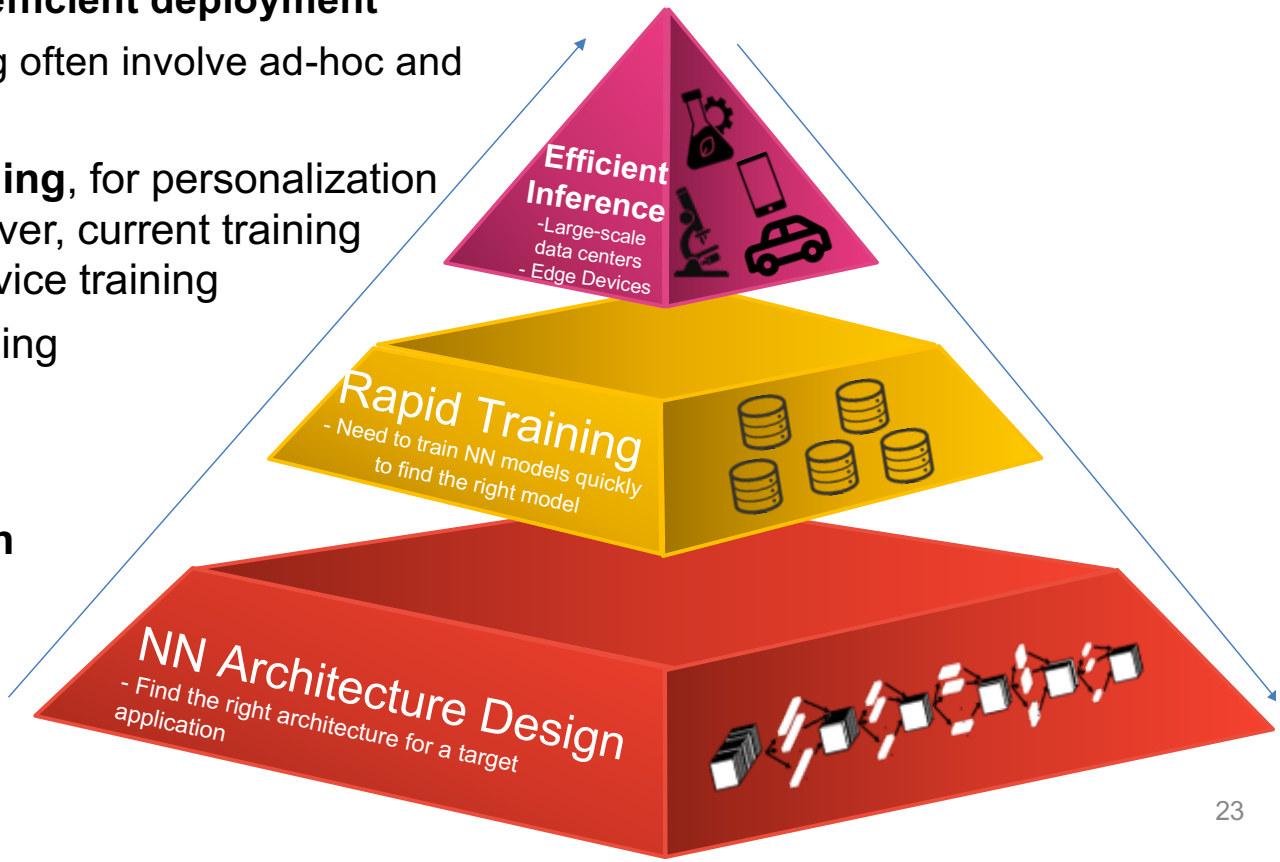
We cannot naively use general DNN models



EdgeAI/TinyML: An Emerging Paradigm for Efficient AI Deployment

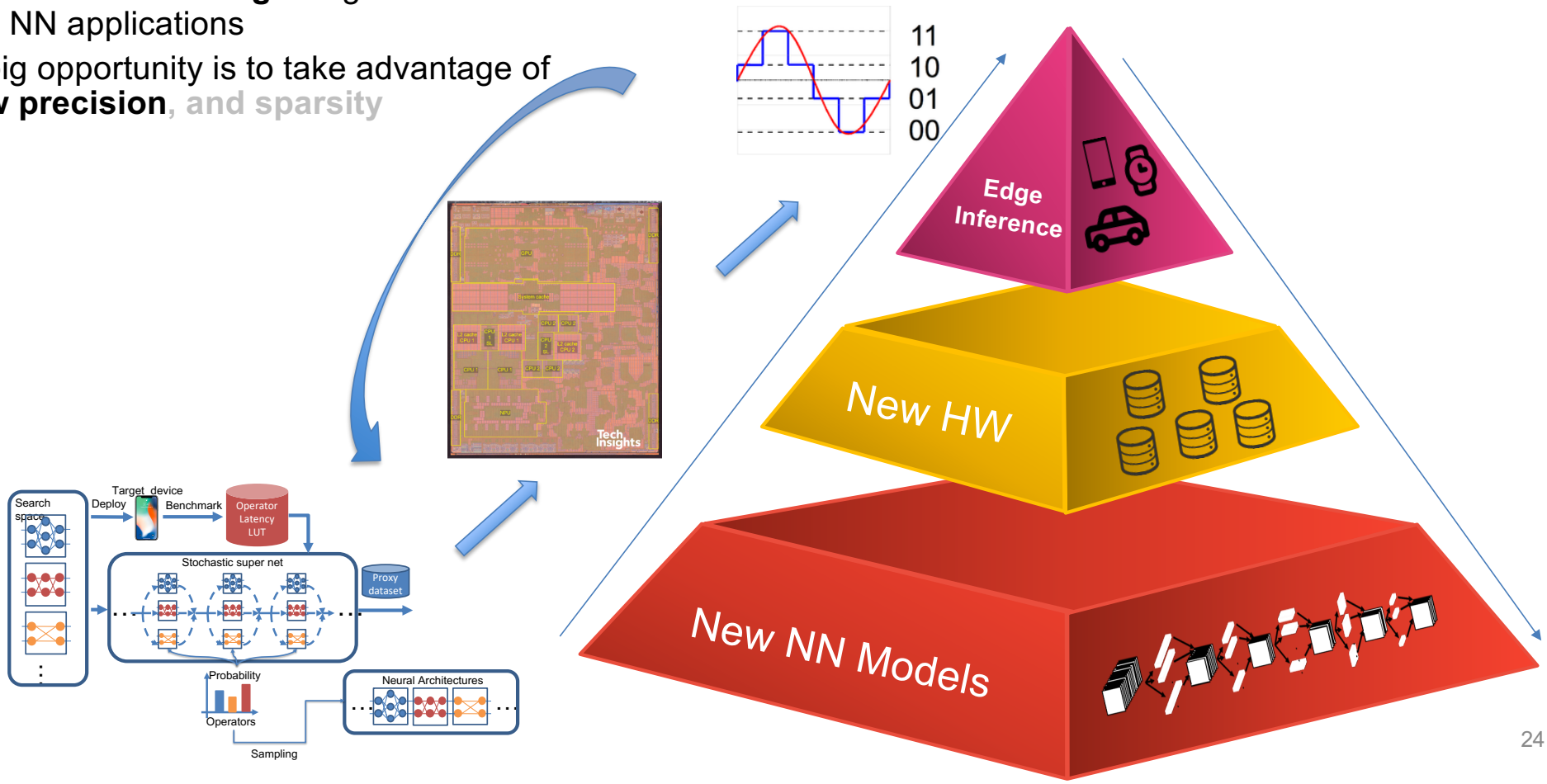
Challenges on the Machine Learning side:

- We need **systematic methods for efficient deployment**
  - Existing quantization and pruning often involve ad-hoc and do not work on new tasks.
- We need **efficient on-device training**, for personalization or online learning scenarios. However, current training methods are not suitable for on-device training
  - Brute force hyperparameter tuning
  - Prohibitive memory footprint
- We need to **rethink and co-design** the NN architecture for efficient edge deployment



Summary: Three Elements of Efficiency at the Edge

- We need to **co-design** Edge HW with the NN applications
- A big opportunity is to take advantage of **low precision**, and sparsity



Restricted Energy and Compute at the Edge: Use Low Precision

- Memory accesses are the principal cost in both latency and energy
- Lower precision weights in DNN mean each memory access brings more data values
- More data values few accesses overall

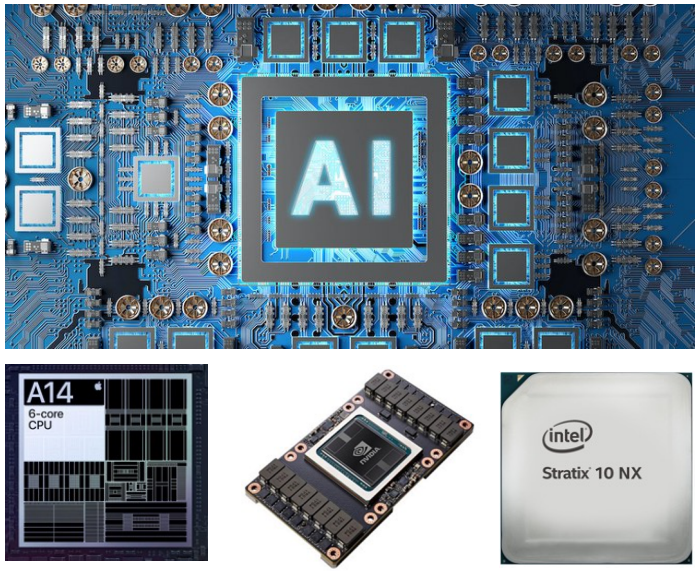
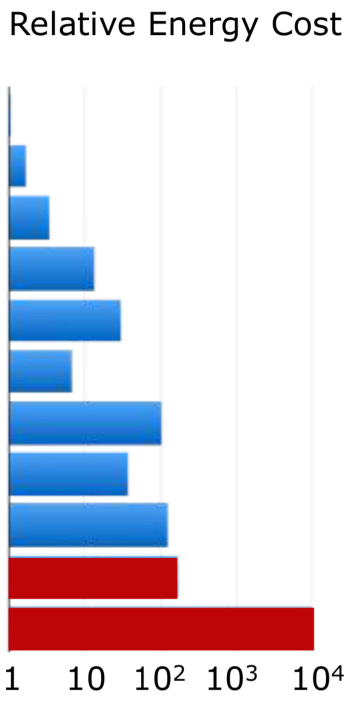


Image Credit: Sdxcentral, nvidia
Table credit: Mark Horowitz

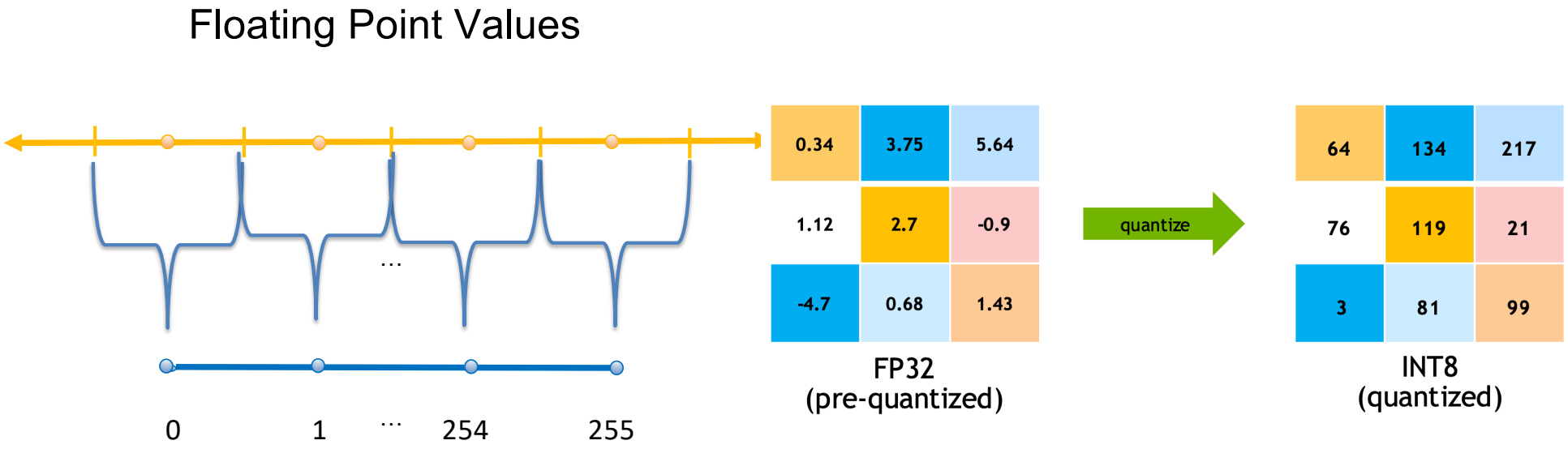
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Multiply	0.2
32b Multiply	3.1
16b FP Multiply	1.1
32b FP Multiply	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

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



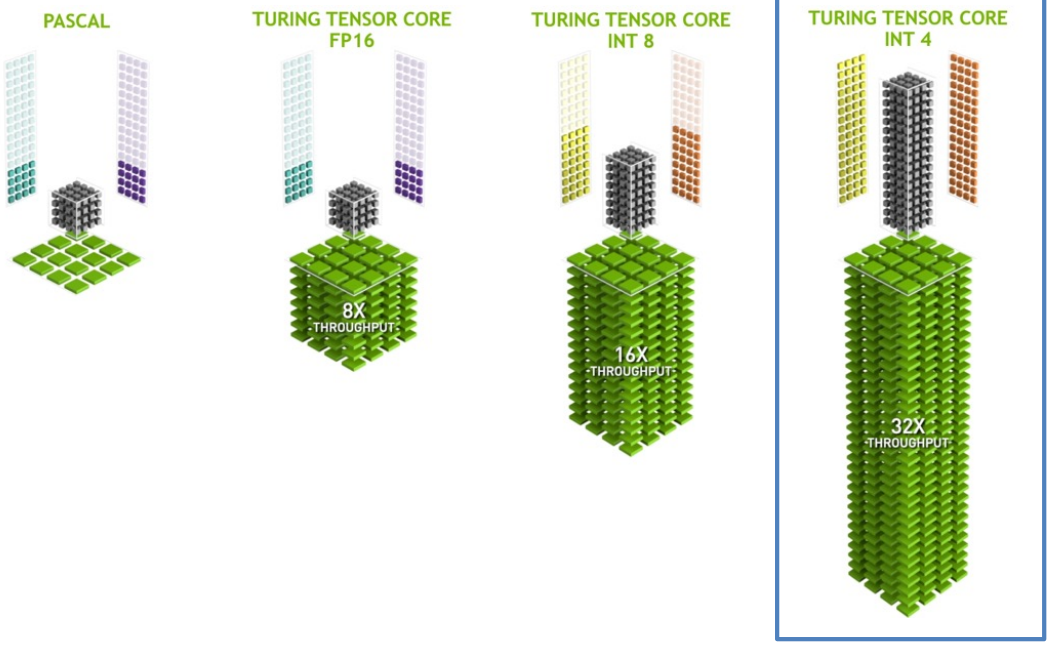
Quantization: Workhorse for Efficient Inference

- Uniform quantization is a linear mapping from floating point values to quantized integer values



Lower precision Multiply-Acc Reduces Energy

- Lower precision weights mean less energy per Multiply-Accumulate
- Also enables putting more MAC units per unit of silicon

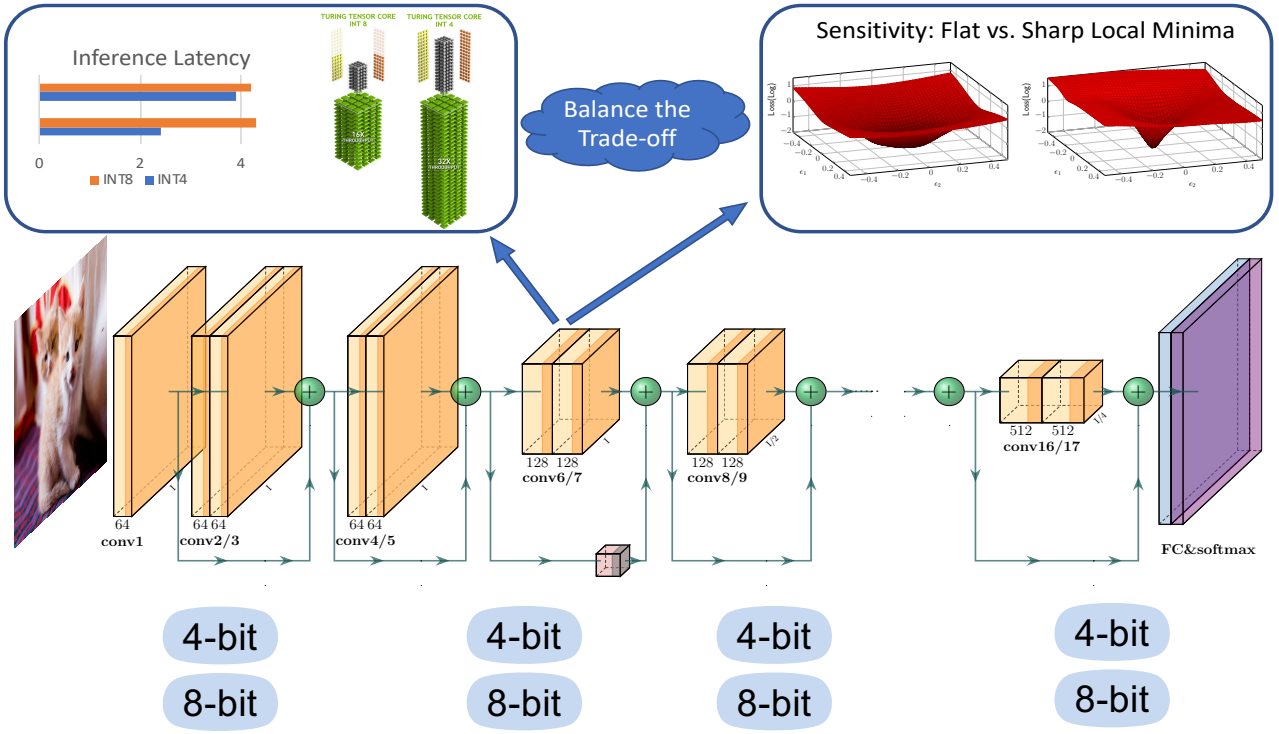


Bits in integer MAC	TOPS /Watt ~45nm
16	0.5 TOPS/Watt
8	1 (2x)
4	5 (10x)
2	10 (20x)

Data from Marian Verhelst, KU Leuven

Big opportunity to enable lower bit precision inference!

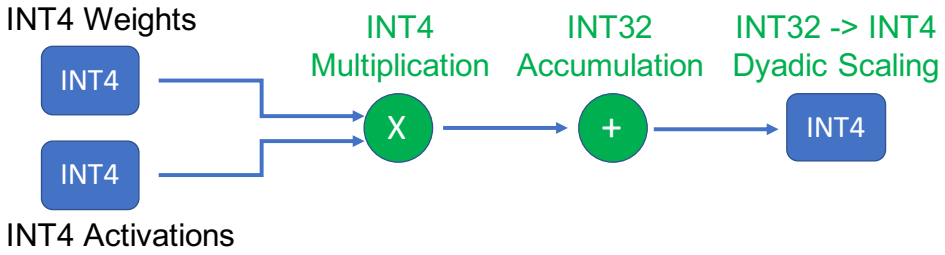
Mixed Precision INT4/8 Quantization Works!



Yao, Zhewei, et al. "HAWQV3: Dyadic Neural Network Quantization." arXiv preprint arXiv:2011.10680.
Z. Dong*, Z. Yao*, A. Gholami*, M. Mahoney, K. Keutzer, HAWQ: Hessian Aware Quantization of Neural Networks With Mixed Precision, **ICCV'19**.
Z. Dong, Z. Yao, Y. Cai, D. Arfeen, A. Gholami, M. Mahoney, K. Keutzer HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks, **NeurIPS'20**.
Shen, Sheng, et al. "Q-bert: Hessian based ultra low precision quantization of BERT." **AAAI'20**.
Kim, Sehoon, et al. "I-BERT: Integer-only BERT Quantization." arXiv preprint arXiv:2101.01321 (2021).

Integer-only Quantization

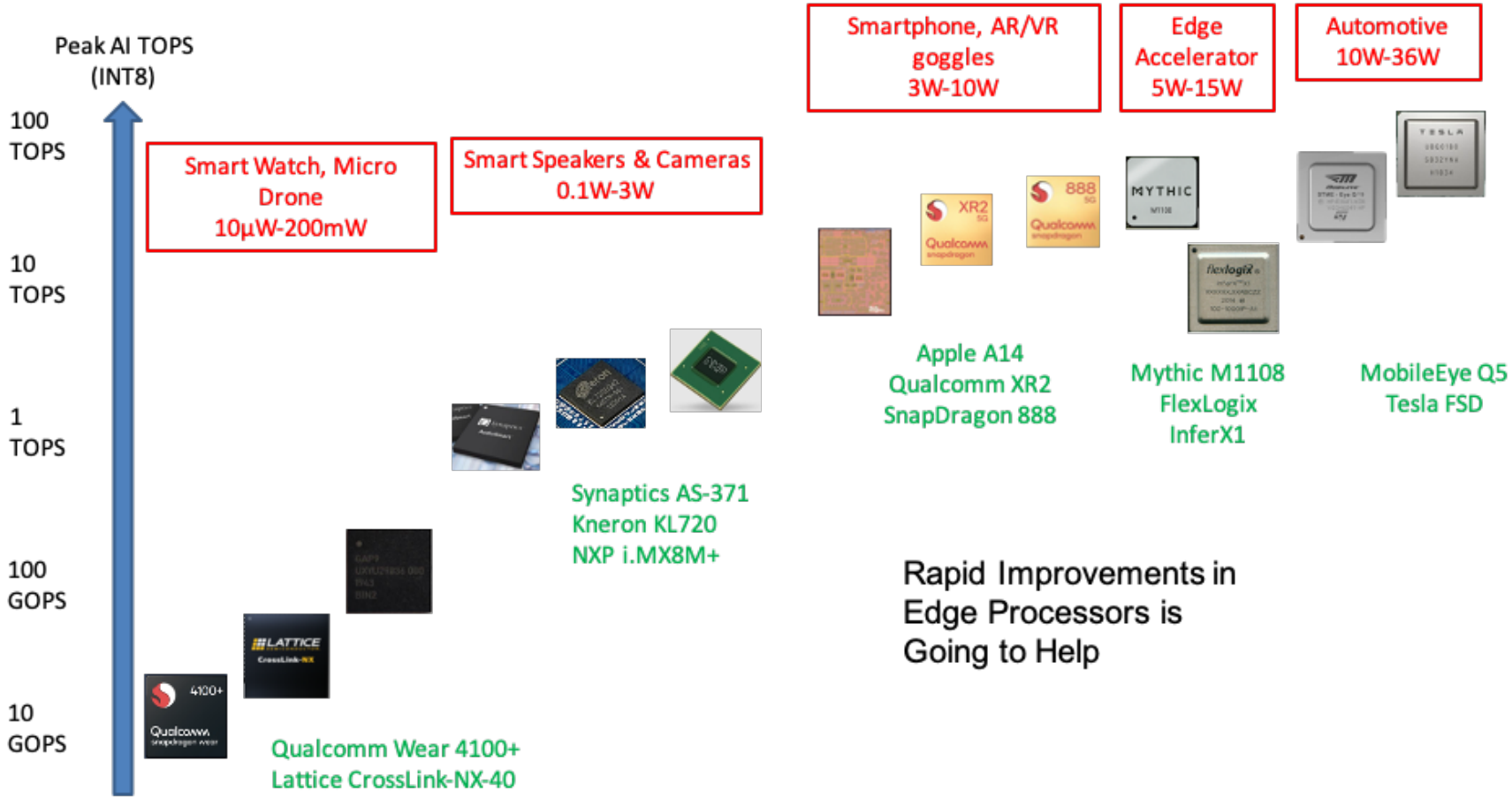
- It is possible to perform integer-only quantization algorithm with only INT multiplication, addition, and bit shifting



- **No accuracy degradation for INT8 (5% higher than prior art)**
- **Direct hardware implementation and verification**
  - Up to **1.5x speed up compared to INT8 quantization**

Yao, Zhewei, et al. "HAWQV3: Dyadic Neural Network Quantization." arXiv preprint arXiv:2011.10680.
Kim, Sehoon, et al. "I-BERT: Integer-only BERT Quantization." arXiv preprint arXiv:2101.01321 (2021).

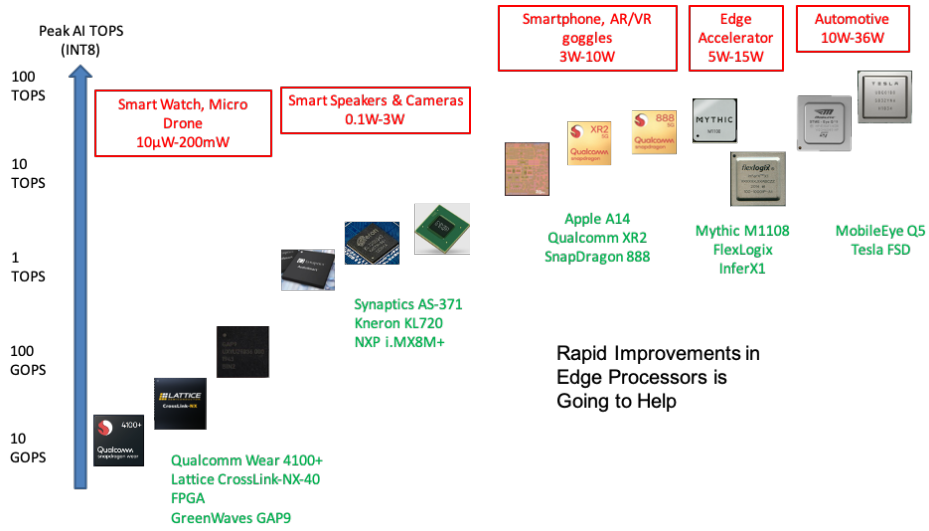
Lower Precision Can Improve Edge HW



Slide: Courtesy of Prof. Keutzer

EdgeAI: Challenging to Deploy

- An important challenge at the edge, is the wide variety of the HW, and the lack of good software that could help accelerate deployment
  - Existing solutions such as TVM are still not fully developed
  - Great opportunity for Intel to provide a standard family of edge HW along with integrated software

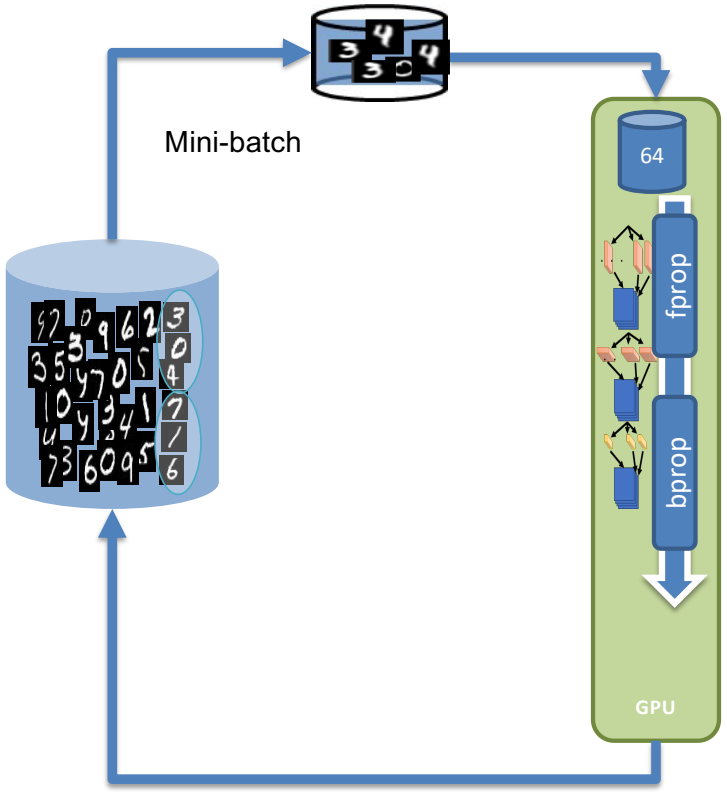


Executive Summary: New Opportunities for DSA

- Emerging AI applications with low arithmetic intensity
  - Recommendation Systems, that need DSA with
    - Large Memory Systems
    - Fast Interconnect
    - Prefetching/Caching Hierarchy
- AI at the Edge:
  - All AI domains (CV, NLP, RecSys, ASR, Robotics/RL, etc.)
    - Low-precision Inference
    - Unified software interface for programmability
    - HW/NN Co-design
- **Emerging AI Optimization Algorithms:**
  - Moving beyond SGD based training to second-order methods
  - Need for HW that supports fast Randomized algorithms
  - Important applications for Scientific ML/Computing

Background:
Stochastic Gradient Descent (SGD)

In every iteration of SGD we load a **random mini-batch of training** data, and compute the gradient.



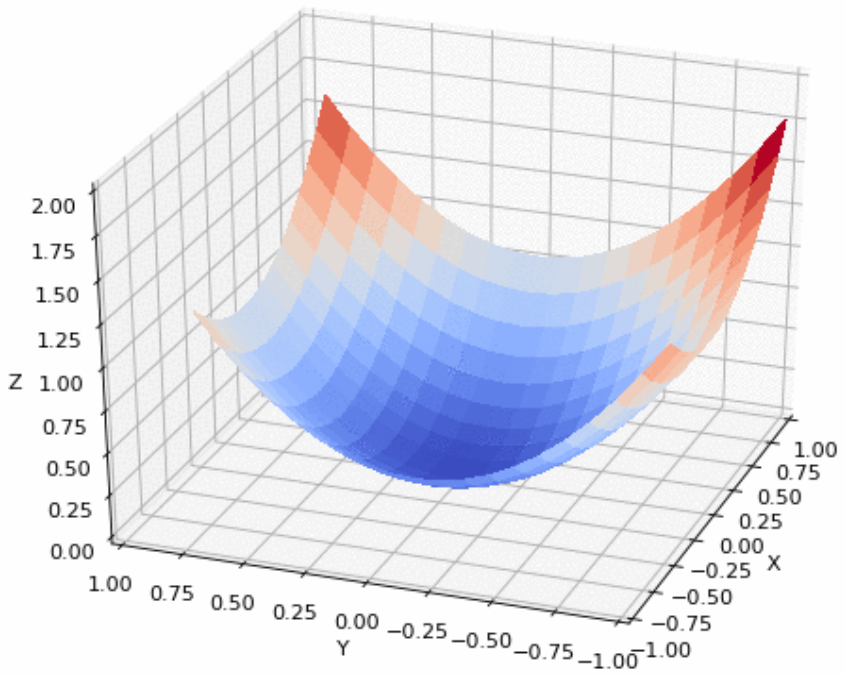
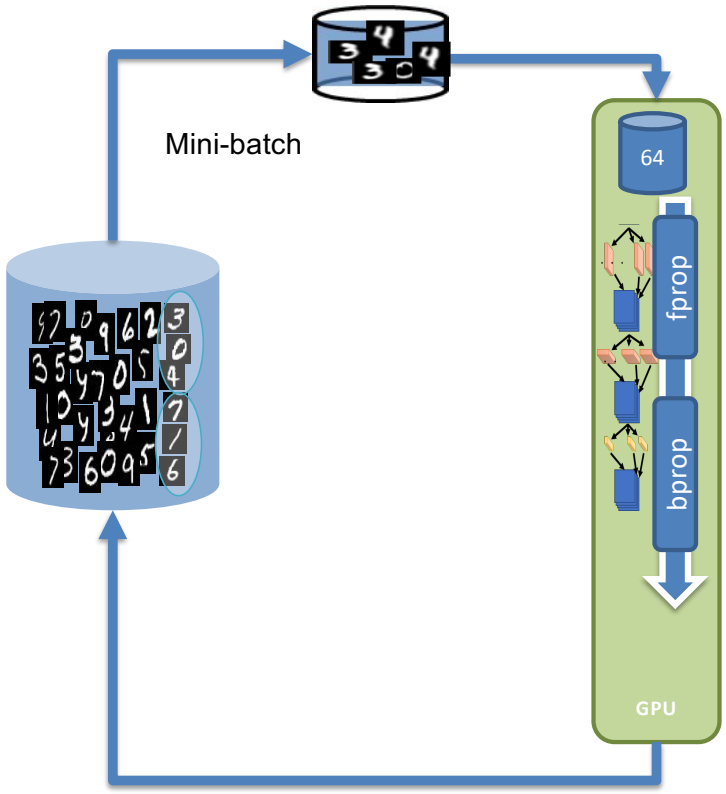


Image Source

Background: Stochastic Gradient Descent (SGD)

In every iteration of SGD we load a **random mini-batch of training** data, and compute the gradient.



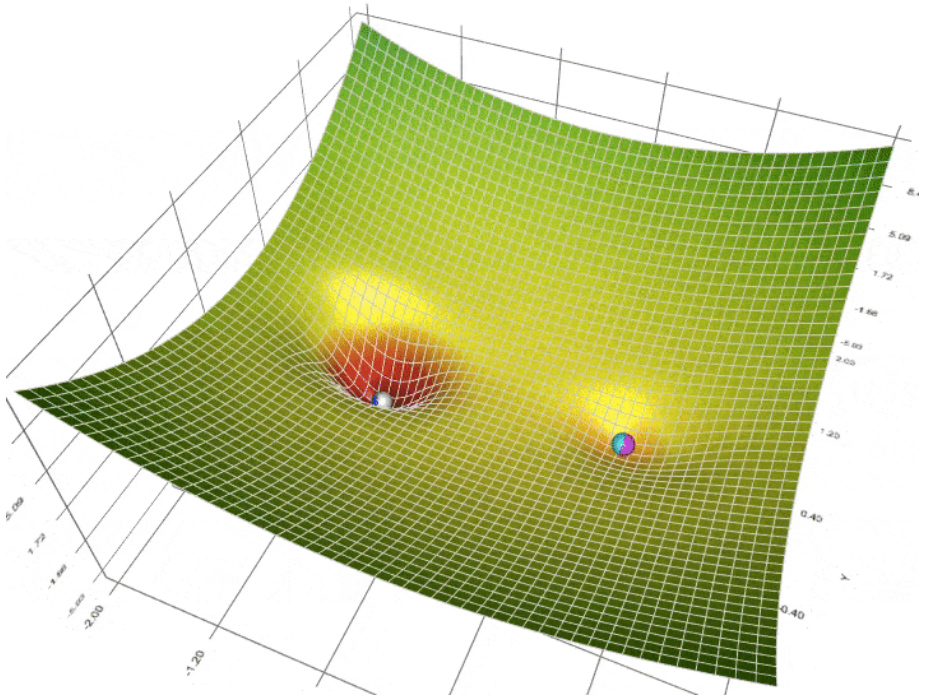


Image Source

Rapid Training of NNs

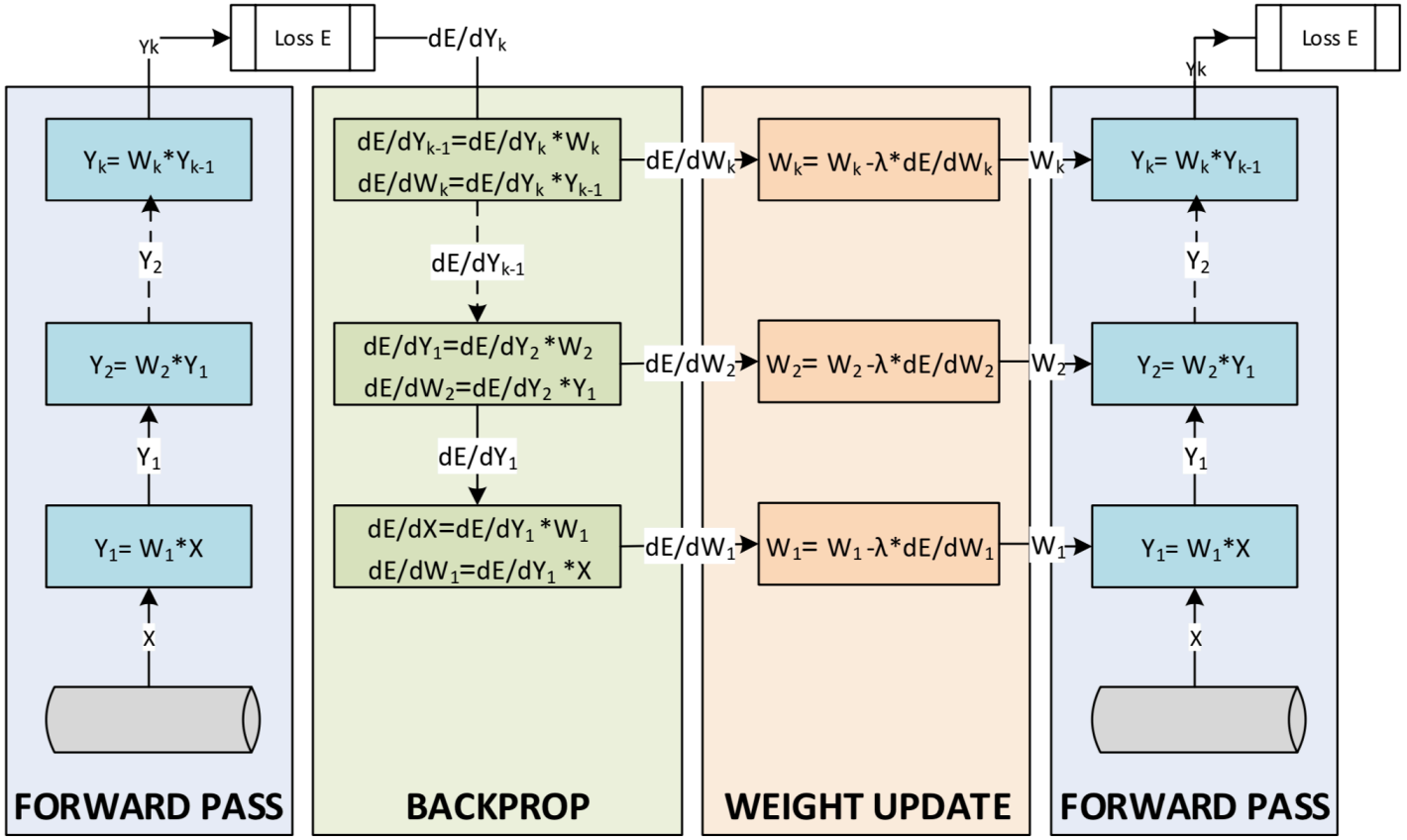
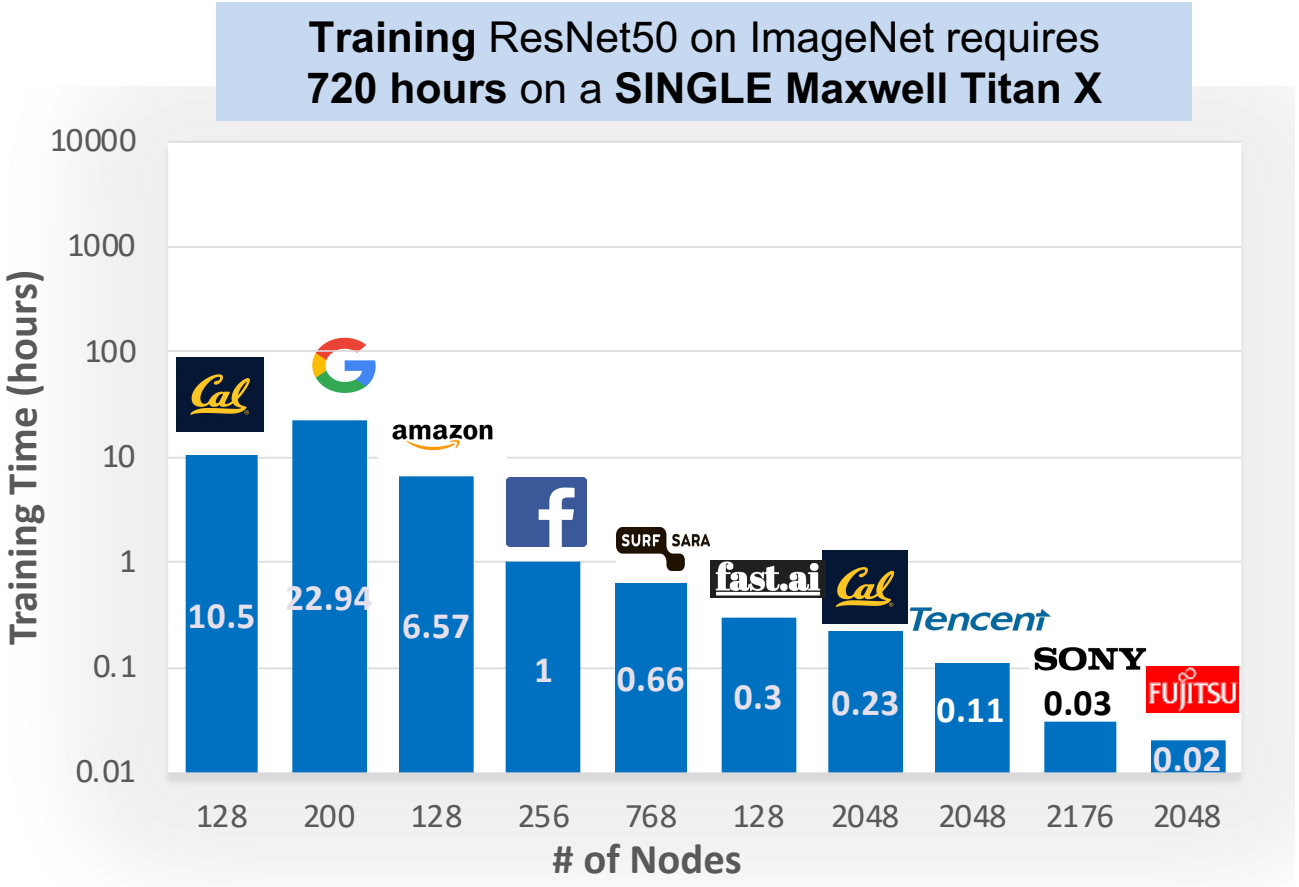


Illustration from Nvidia (B. Ginsburg)

ImageNet Scaling!

- Main benchmark for hardware performance in ML

Great Progress on Scaling ResNet50 on ImageNet



- **<1 minute**
- **>50,000x speedup**

Despite the great progress the solutions **do not work** on problems other than ResNet50 on ImageNet

ResNet50 on ImageNet is too Simple!

- The loss landscape of ResNet50 is too simple, and not representative of new workloads

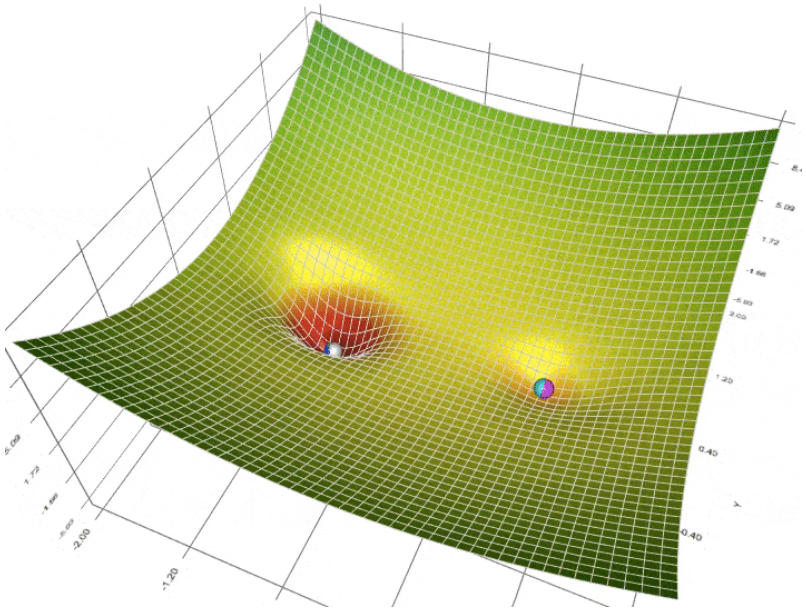
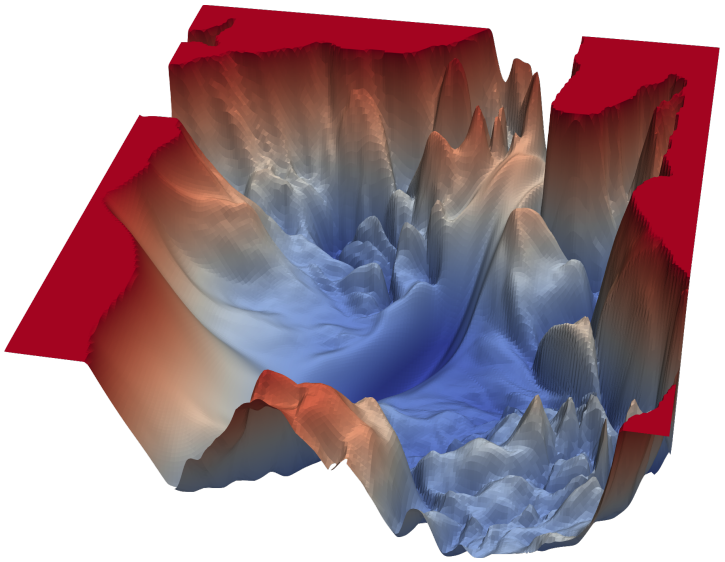
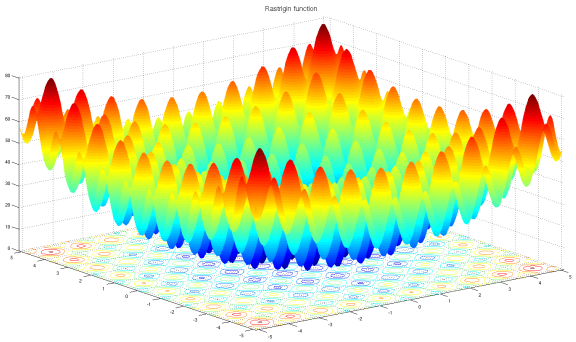


Image Source

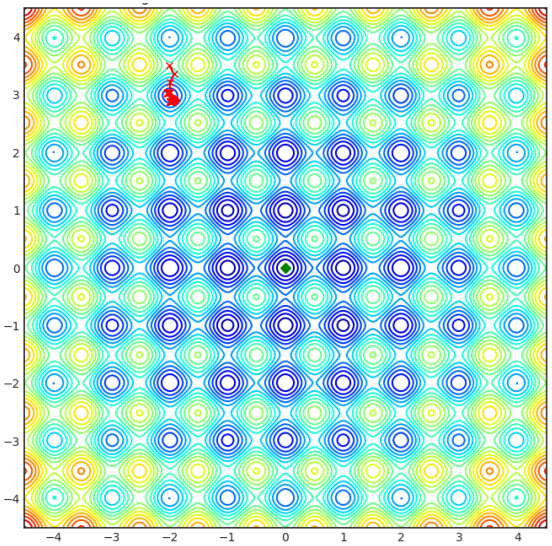


Li H, Xu Z, Taylor G, Studer C, Goldstein T. Visualizing the loss landscape of neural nets. NeurIPS'18, 2018

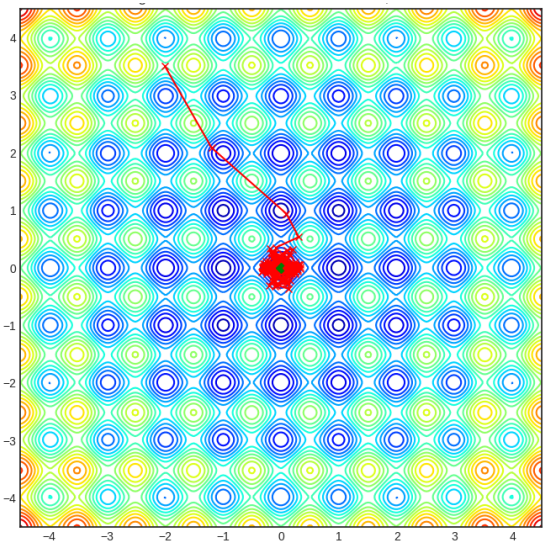
Emerging Solutions:
Going Beyond Simple SGD



Loss Function



First-Order Methods
(SGD)



Second-Order Methods
(AdaHessian)

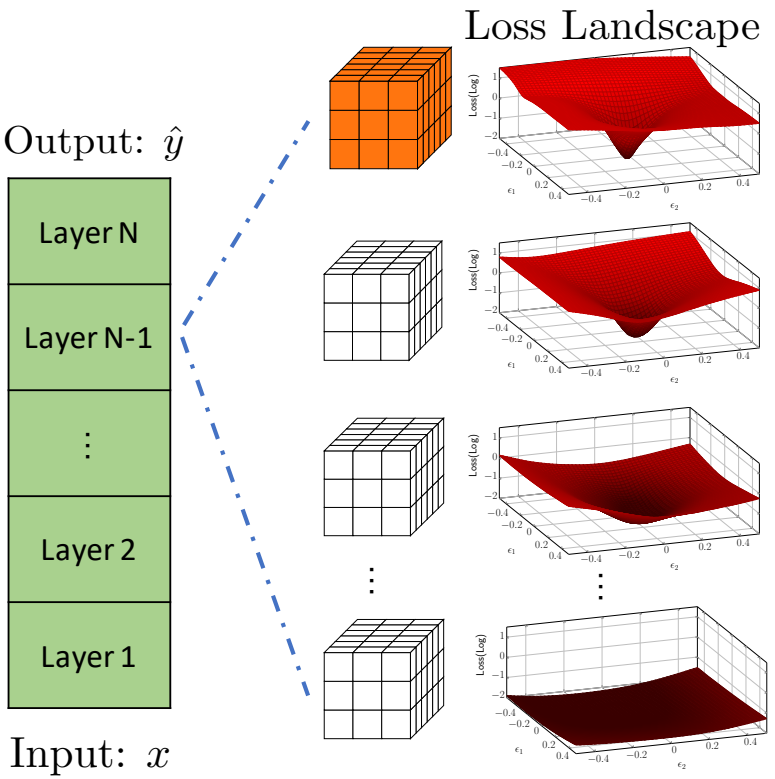
Benchmark Link: https://github.com/jettify/pytorch-optimizer

Emerging Solutions: Going Beyond Simple SGD

- New second-order algorithms are emerging for training NNs

How would this impact HW design?

- Need fast Randomized Linear Algebra
  - Randomized Matrix-Matrix Operations
  - RBLAS

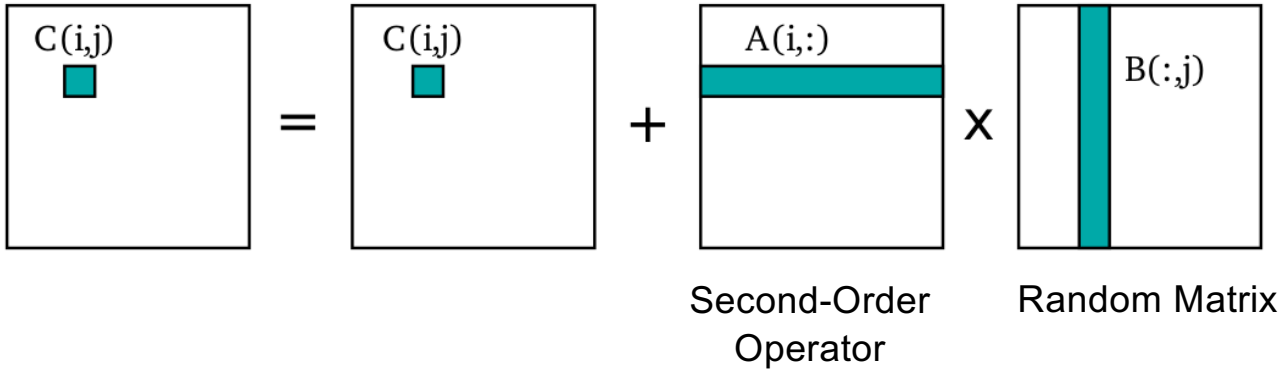


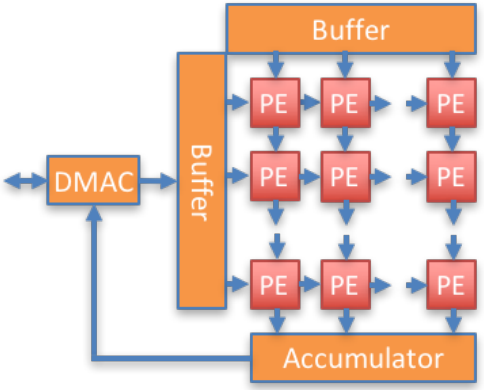
Z Yao, A Gholami, S Shen, M Mustafa, K Keutzer, MW Mahoney, ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning, **AAAI'21**.
Z. Yao*, A. Gholami*, Q. Lei, K. Keutzer, M. Mahoney, Hessian-based Analysis of Large Batch Training and Robustness to Adversaries, **NeurIPS'18**, 2018.
Z. Yao*, A. Gholami*, K. Keutzer, M. Mahoney, PyHessian: Neural Networks Through the Lens of the Hessian **Spotlight at ICML'20 workshop** on Beyond First-Order Optimization Methods in Machine Learning, 2020.
Code: https://github.com/amirgholami/PyHessian
Code: https://github.com/amirgholami/AdaHessian

DSA with Fast Randomized Matrix Operations

- The core of these algorithms is multiplying a matrix with a random matrix
  - Accelerating Randomized operations can have a huge impact

```
% inner product approach
for i = 1:I
    for j = 1:J
        for k = 1:K
            C(i,j) = C(i,j) + A(i,k)*B(k,j);
```



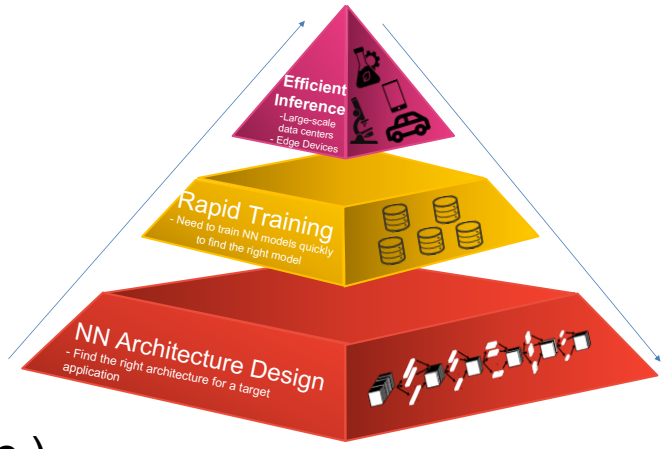


Fast RNG can significantly improve performance

A. Gholami, K. Kwon, B. Wu, Z. Tai, X. Yue, P. Jin, S. Zhao, K. Keutzer, SqueezeNext: Hardware-Aware Neural Network Design, ECV Workshop at CVPR'18, 2018.
K. Kwon, A. Amid, **A. Gholami**, B. Wu, K. Keutzer, Co-Design of Deep Neural Nets and Neural Net Accelerators for Embedded Vision Applications, Design Automation Conference (DAC'18), 2018.
Matrix Image Source: https://patterns.eecs.berkeley.edu/?page_id=158
NSF Ballistic Project (UCB, UoT, ORNL, UOM, CU)

Summary: New Opportunities for DSA

- **Emerging AI applications with low arithmetic intensity**
  - Recommendation Systems, that need DSA with
    - Large Memory Systems
    - Fast Interconnect
    - Efficient Prefetching and Cache Hierarchy
- **AI at the Edge:**
  - All AI domains (CV, NLP, RecSys, ASR, Robotics/RL, etc.)
    - Low-precision Inference
    - Unified software interface for better programmability
- **Emerging AI Optimization Algorithms:**
  - Moving beyond SGD based training to second-order methods
  - Need for DSA that supports fast Randomized algorithms
  - Important applications for Scientific ML/Computing



Thanks for Listening

For any feedback/questions please contact

amirgh@berkeley.edu

Berkeley
UNIVERSITY OF CALIFORNIA