System Energy Efficiency Lab

seelab.ucsd.edu

Hyperdimensional Computing & Applications

Prof. Tajana Simunic Rosing

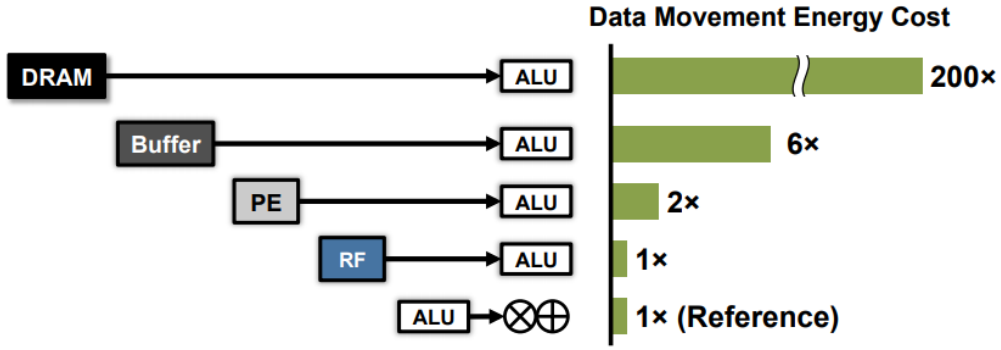
& amazing PhD students, postdocs and staff @ UCSD

Learning from the Big Data

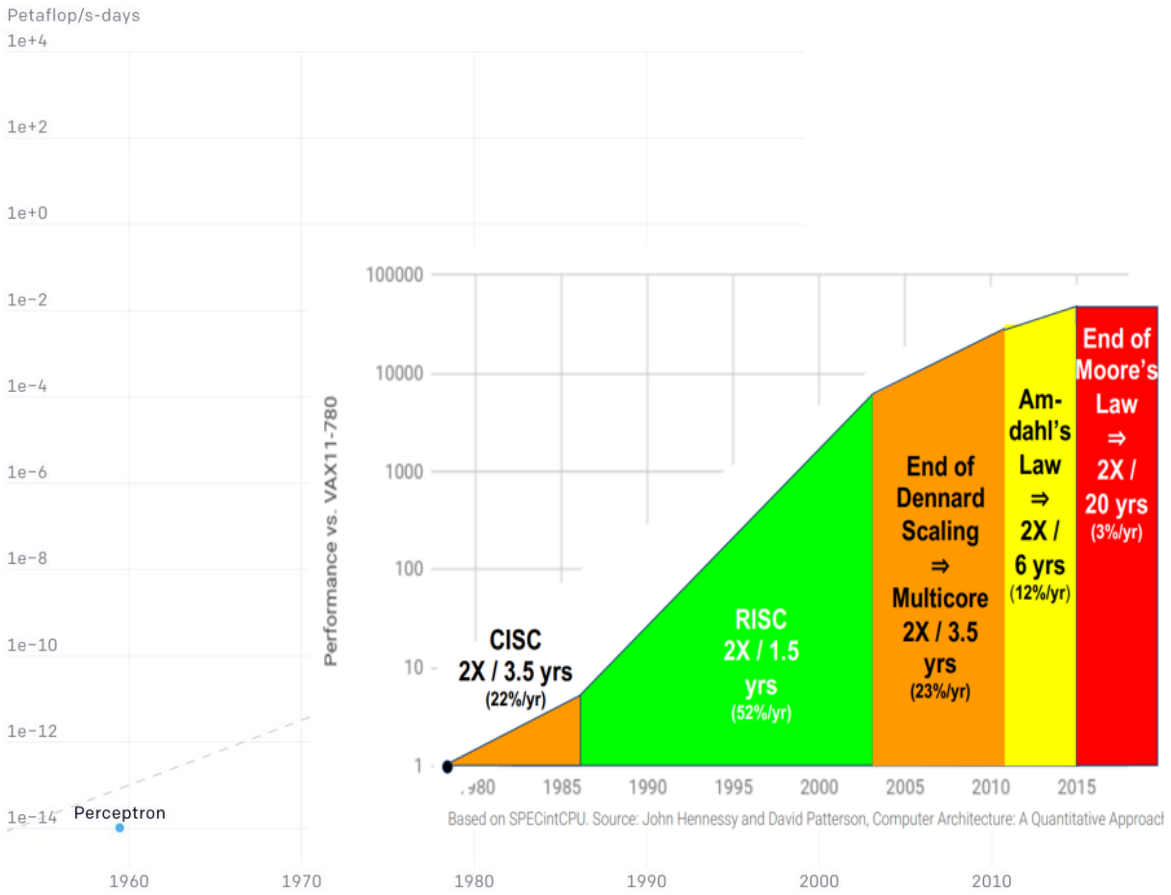
- Big data affects all areas of our lives
  - Compound annual growth rate of 14%, $116B by ‘27
- Two phases of growth in AI systems:
  - 1959-’12: 2 year doubling time
  - 2012 - 22 training at 10x/yr; inference 1.5x/yr
- Compute has not kept up =>
- Sizes of the models are exploding:

Models =>	GPT-1	GPT-2	GPT-3
Parameters	117 M	1.5 B	175 B

- Moving data to compute is expensive!



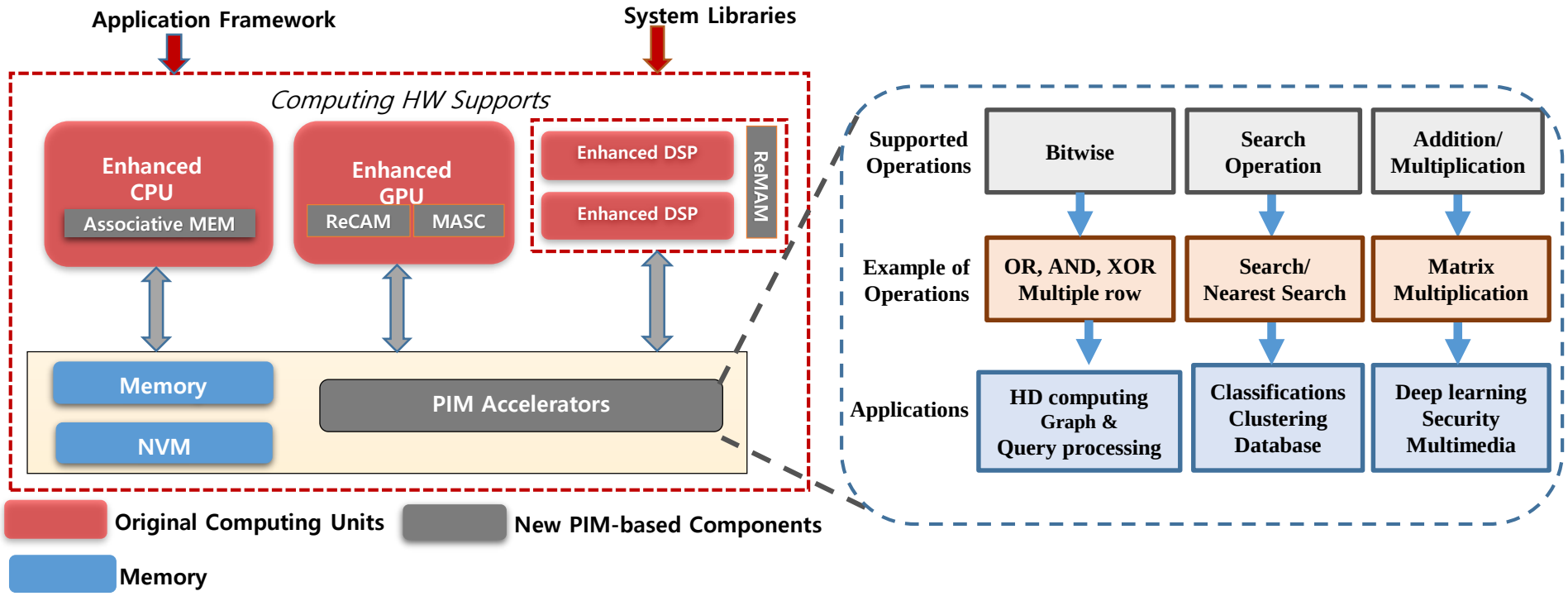
Computing in AI systems



Sources: https://openai.com/blog/ai-and-compute; Attention is all you need; David Patterson ‘22

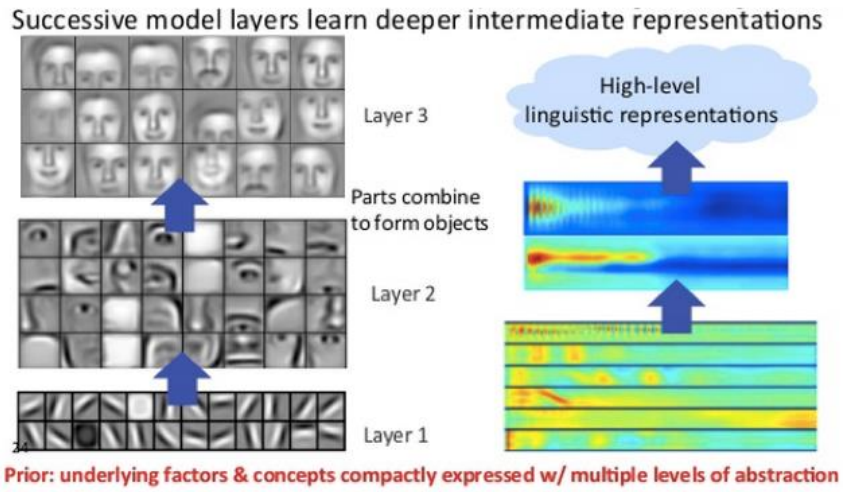
Accelerating Big Data Analysis

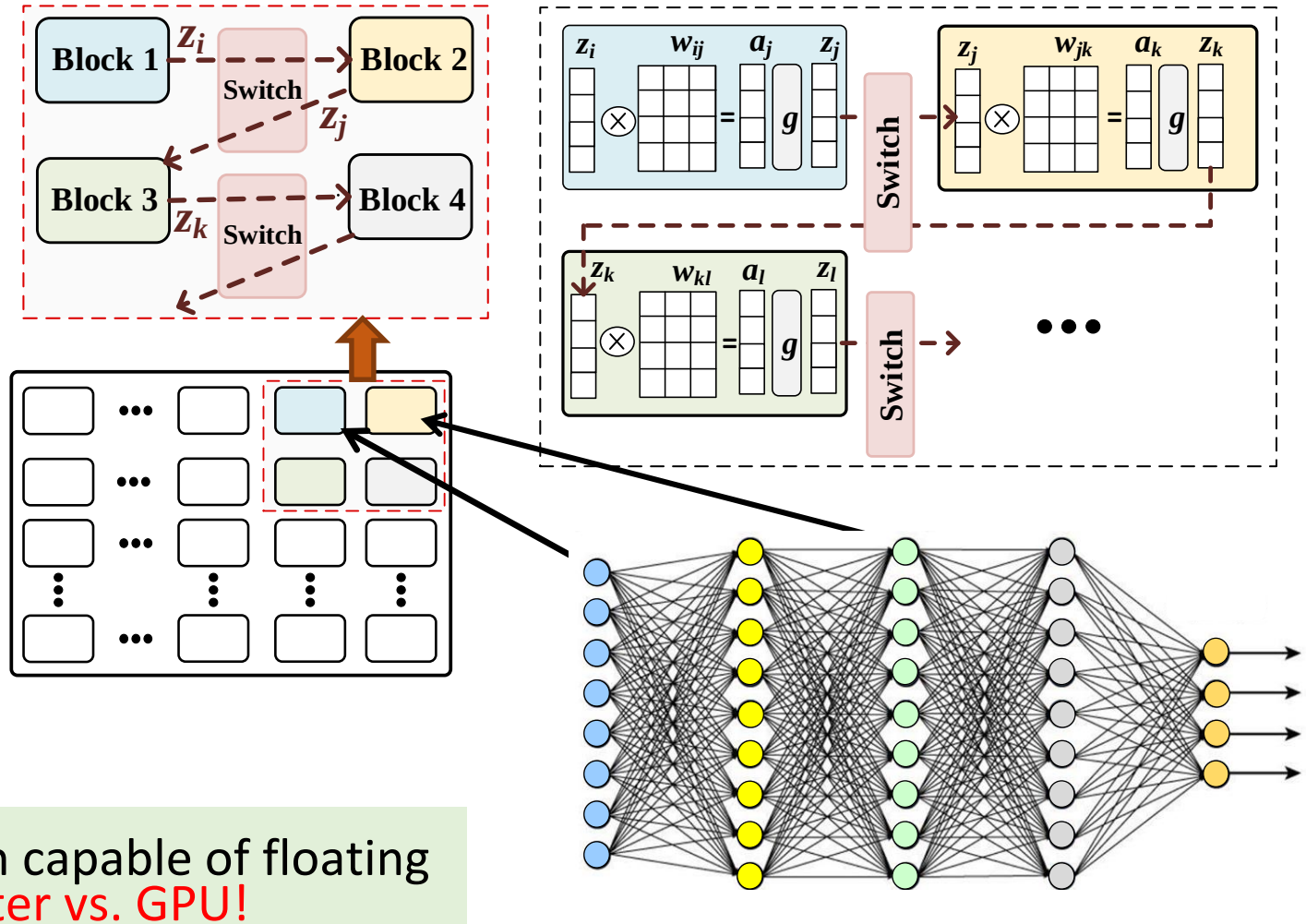
- Key insights:
  - As much as 90% cost of data processing is due to memory transfers
- Solutions:
  - Move compute to data: accelerate Processing In/near Memory (PIM) and storage



DNNs Processing In Memory (PIM)

- DNN performance depends on expensive training that requires a lot of data & hyperparameter tuning
- Google accelerates it with TPUs
  - Data movement costs are still there
- Instead accelerate in memory!



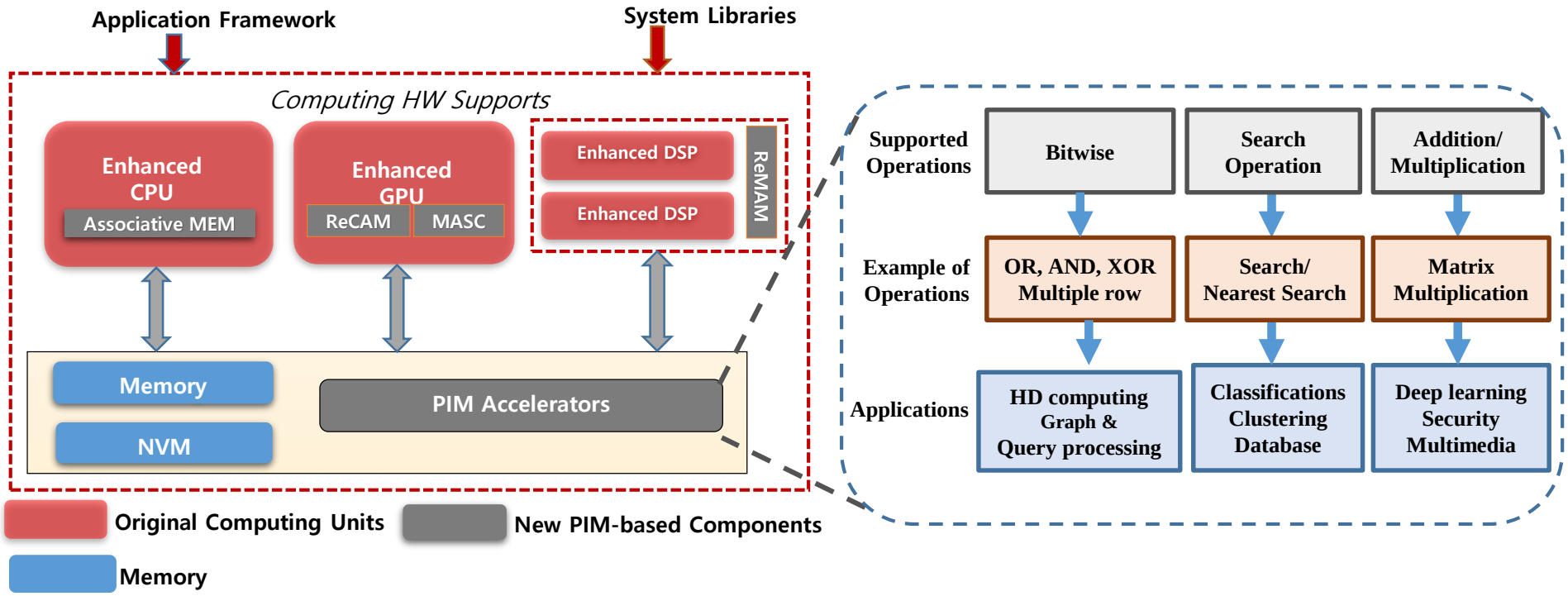


Our DNN PIM accelerator is the first design capable of floating point training and inference. It is 303x faster vs. GPU!

M. Imani, S. Gupta, Y. Kim, T. S. Rosing “FloatPIM: In-Memory Acceleration of Deep Neural Network Training with High Precision”, ISCA, 2019.

Accelerating Big Data Analysis

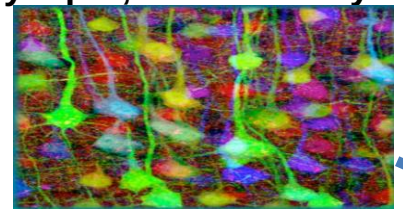
- Key insights:
  - As much as 90% cost of data processing is due to memory transfers
- Solutions:
  - Move compute to data: accelerate Processing In/near Memory (PIM) and storage
  - Brain-inspired algorithms better suited for PIM: Hyperdimensional computing



Human brain is built for learning!

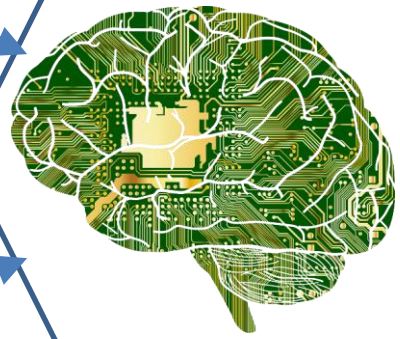
Fault tolerance

Noisy input, Neurons may die



Low power

~20 W power consumption

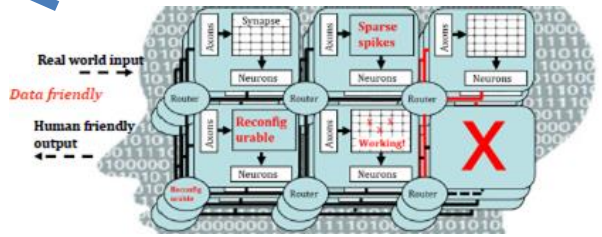


Learning ability

Supervised & unsupervised

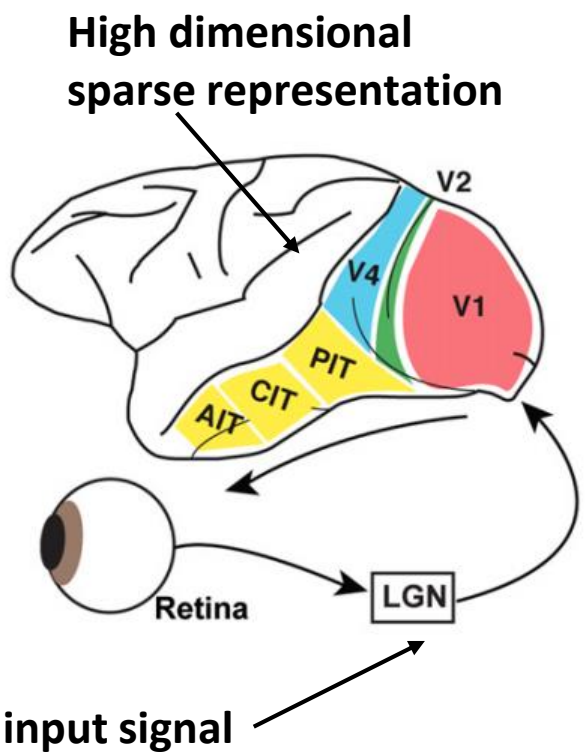
Highly parallel

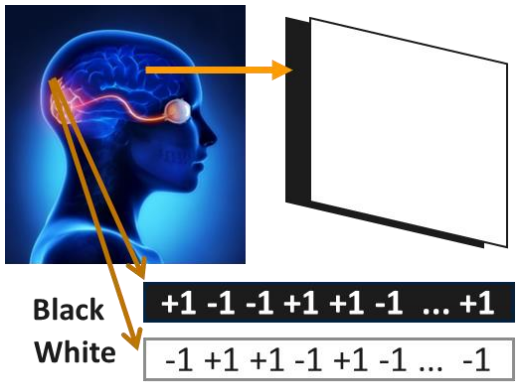
Extremely compact

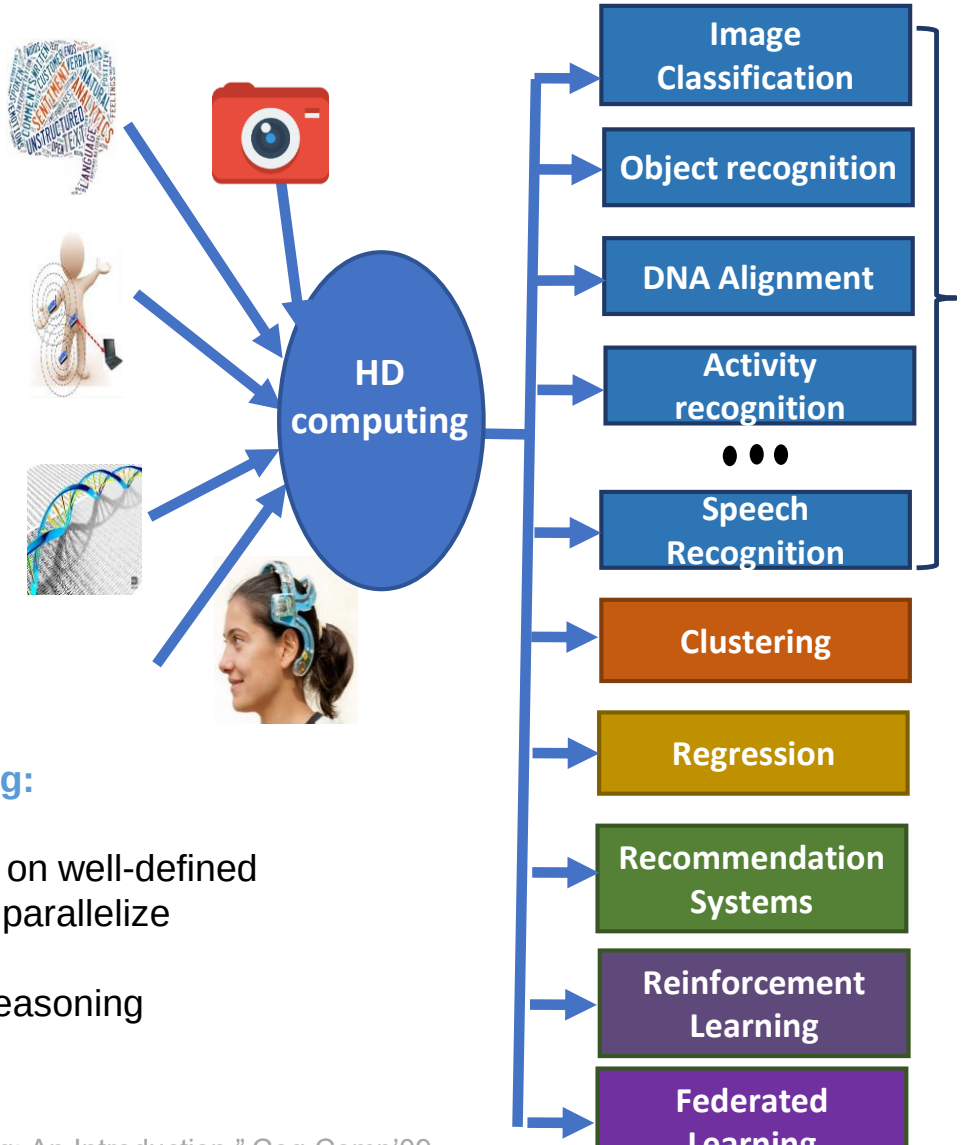


Brain-inspired Hyperdimensional Computing

Dense sensory input is mapped to
high-dimensional sparse representation
on which brain operates [Babadi, Sompolinsky 2014]





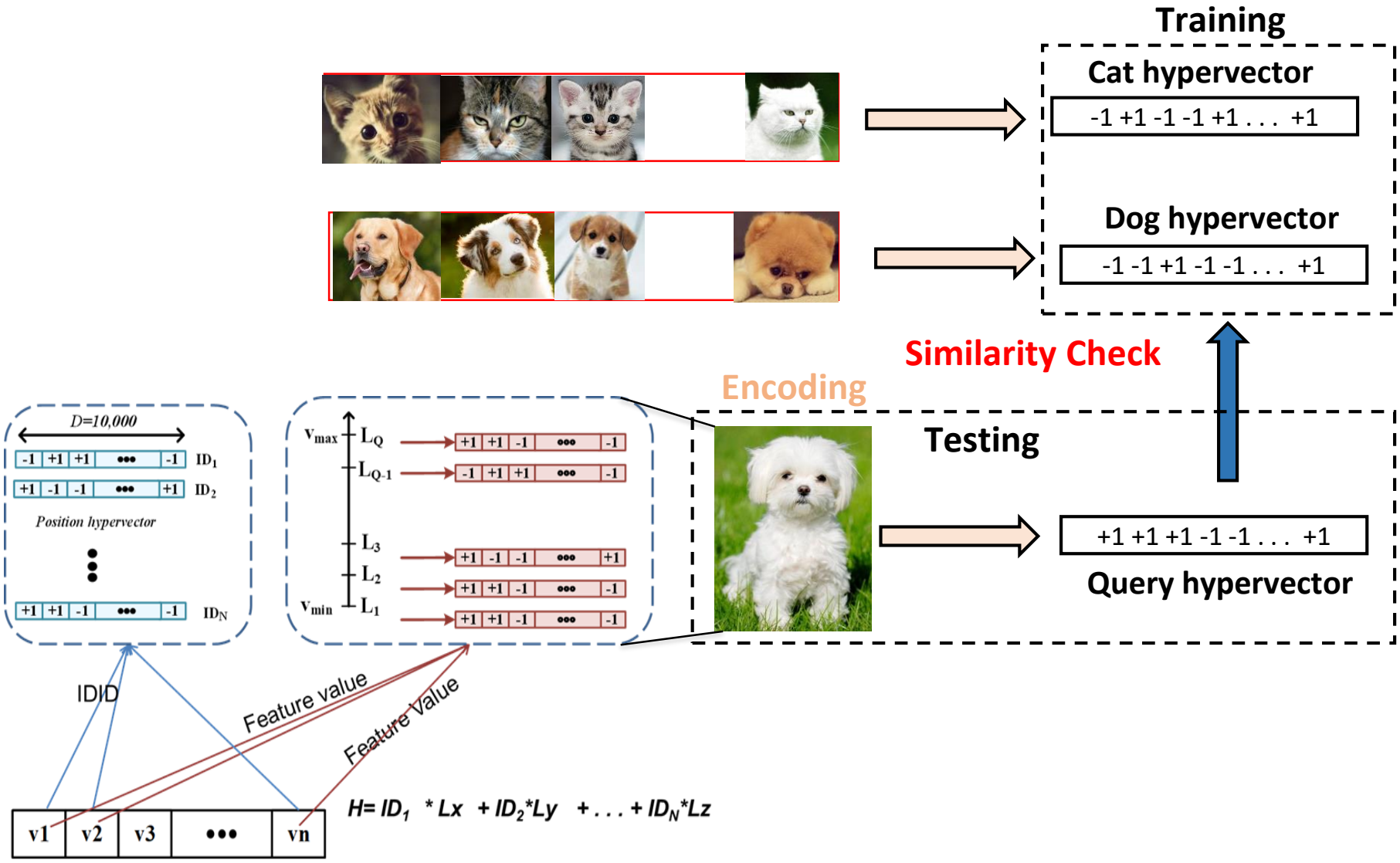


Hyperdimensional (HD) computing:

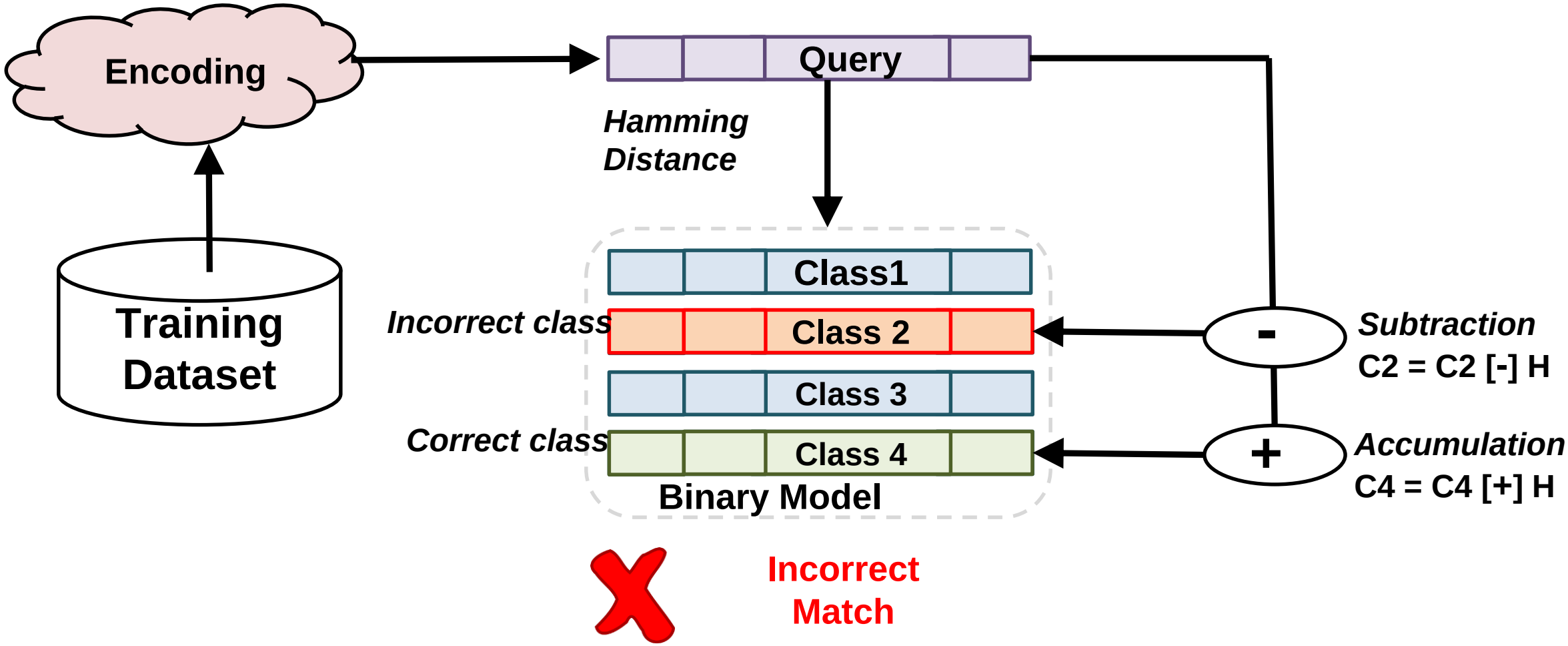
- Encodes data into hypervectors
- Leverages full algebra and works on well-defined set of operations that are easy to parallelize
- Fast single-pass training
- Supports real-time learning and reasoning
- Energy-efficient & robust to noise

Sources: DiCarlo et al "How does the brain solve visual object recognition?." *Neuron* 12; Kanerva, "HD Computing: An Introduction," Cog.Comp'09

HD Computing Classification: Encoding, Training & Inference

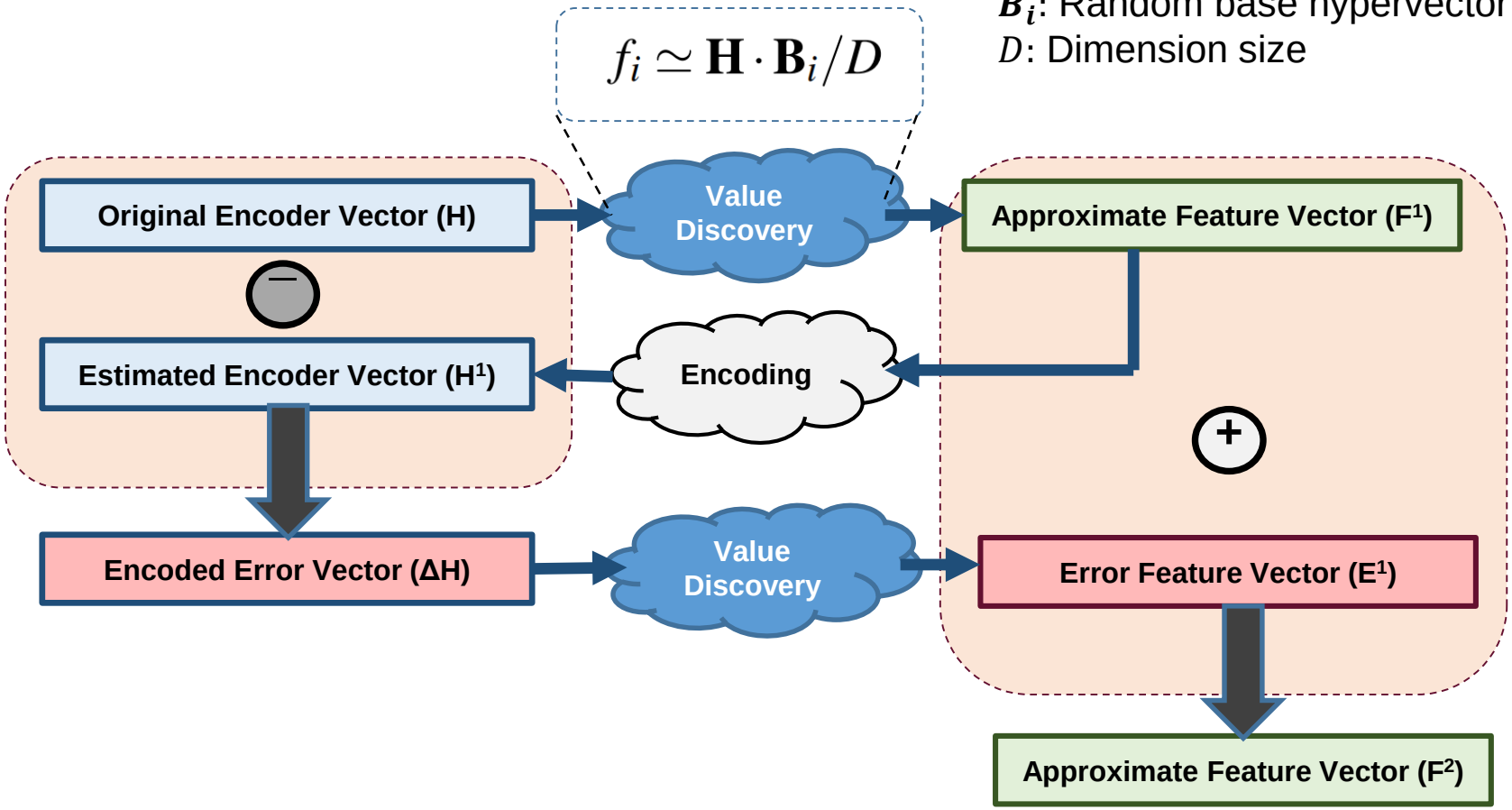


Online Retraining/Model Adjustment



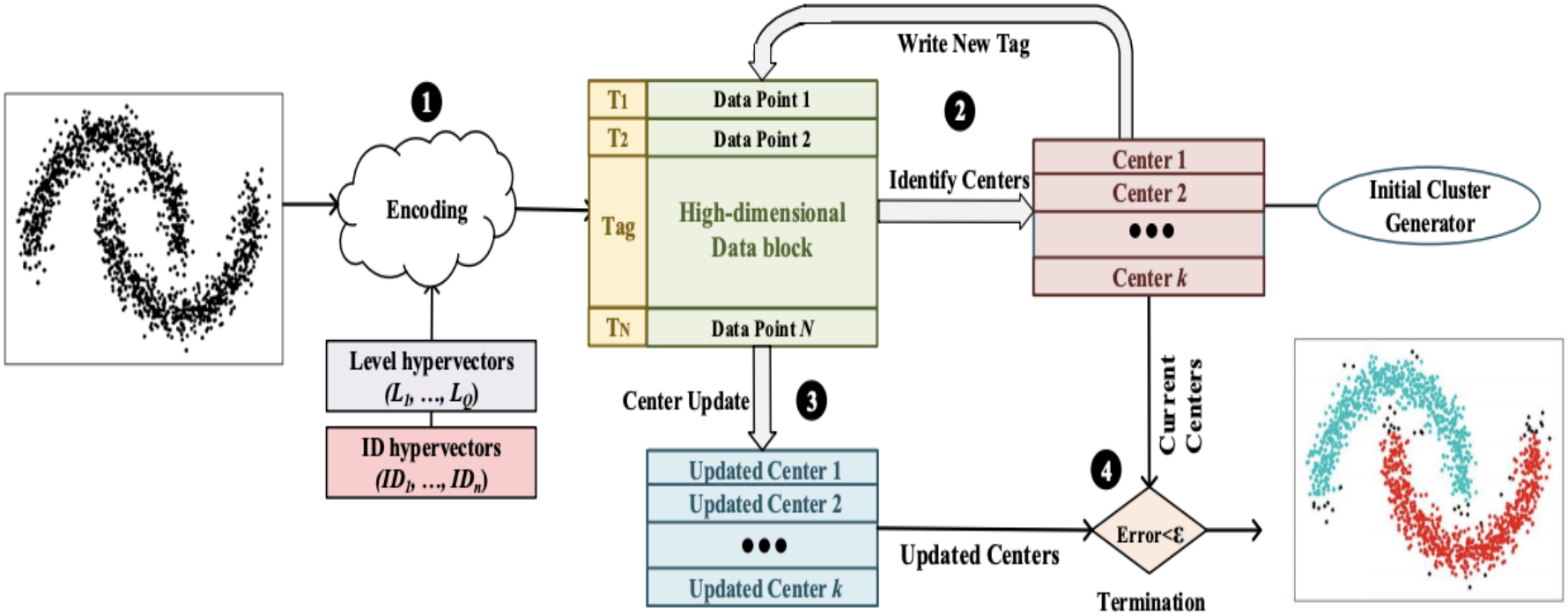
HD Decoding & Interpreting Inference Results

f_i: original feature value
$\boldsymbol{B_i}$: Random base hypervector
D: Dimension size



HDC Clustering

- Encode data in HD space, cluster using similarity metric
- Generalizes popular K-Means algorithms to HD encoded data



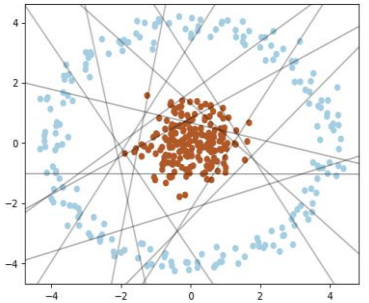
M. Imani, S. Pampana, S. Gupta, M. Zhou, Y. Kim, T. Rosing, "DUAL: Acceleration of Clustering Algorithms using Digital-based PIM," MICRO 2020.

Theoretical Results on Hypervector Dimensionality

Random Projection Encoding:

Encoding dimension depends on desired level of sparsity

To ensure a k-sparse representation which separates classes with probability 1-α

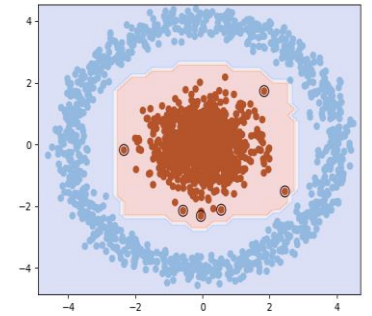


$$d \geq k \frac{\log \alpha}{\log 1-\theta} \text{ where } \theta \approx O\left(1 - \frac{16}{k^2}\right)^n \text{ when } k \text{ is} \geq 10$$

ID-Level Encoding:

Preserves the geometry of data up to additive distortion

To ensure separability is preserved $d > \frac{4n^2}{\delta^2} \log \frac{2(nm)^2}{\alpha}$



n: dimensionality of the original data

m: quantization levels

δ: distance between classes or centroids in low dimensionality

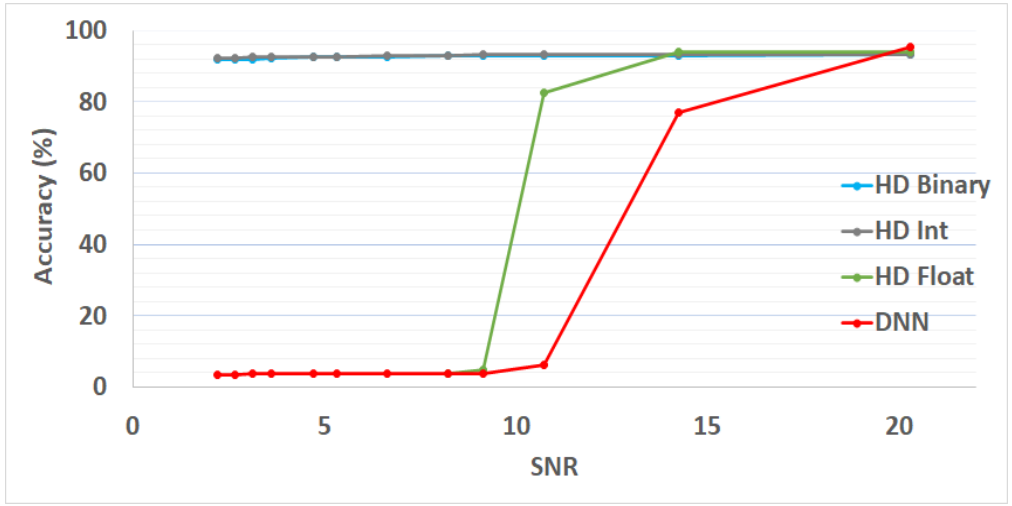
Effects of Noise:

For N symbols drawn from alphabet of size M with noise bound ω: $d_{hv} = O(\omega N \log M)$

A.Thomas, S. Dasgupta, T. Rosing, "A Theoretical Perspective on Hyperdimensional Computing," JAIR'21.

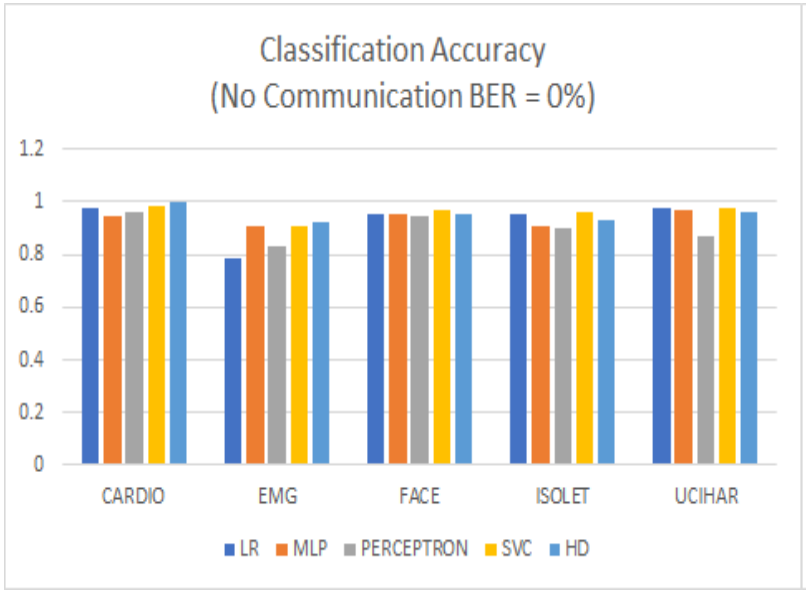
HD computing Robustness to Noise

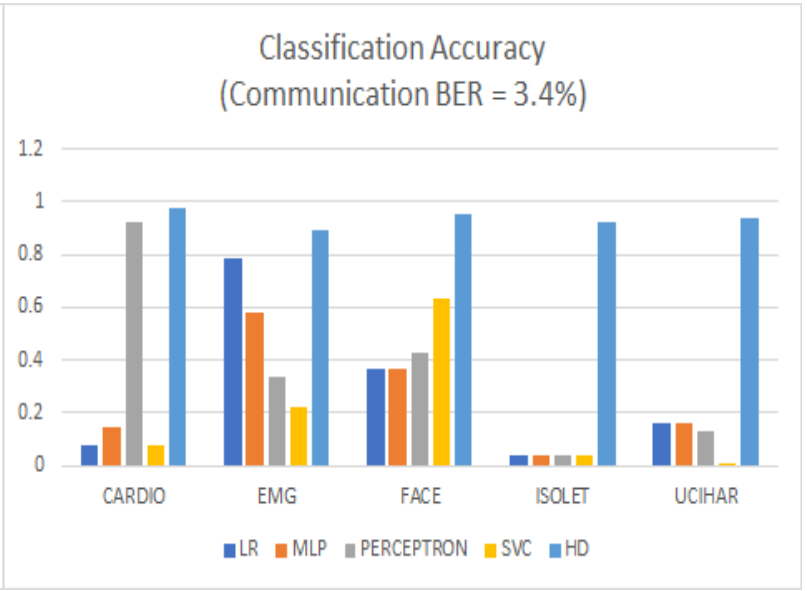
HD vs. DNN accuracy & robustness



HDC accuracy drop vs. D at BER=3.4%

Accuracy Loss %	10,000	8,000	6,000	4,000	2,000
Classification	0.58%	0.82%	1.44%	1.89%	2.39%
Clustering	0.66%	2.48%	2.52%	2.79%	3.13%



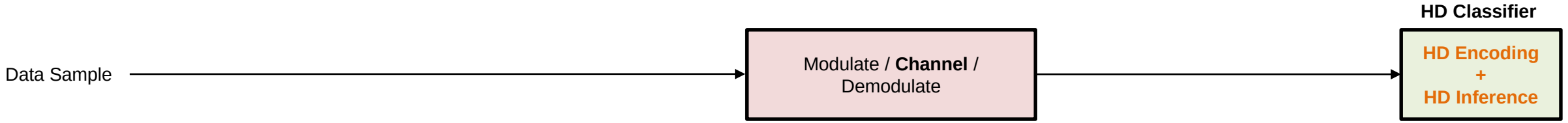


HDC has comparable accuracy to other light-weight classifiers with no noise, but is much more robust at increased BER

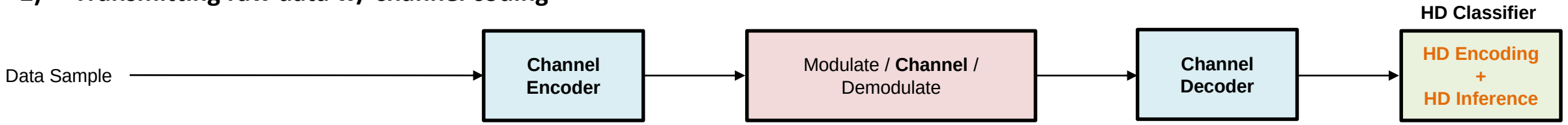
J. Morris, K. Ergun, B. Khaleghi, M. Imani, B. Aksanli, T. Rosing, "HyDREA: HD Computing For Robust and Efficient Machine Learning System," ACM TECS'22

HDC Communication Robustness Analysis

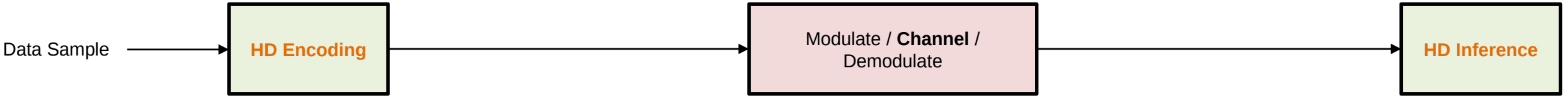
1) **Transmitting raw data**



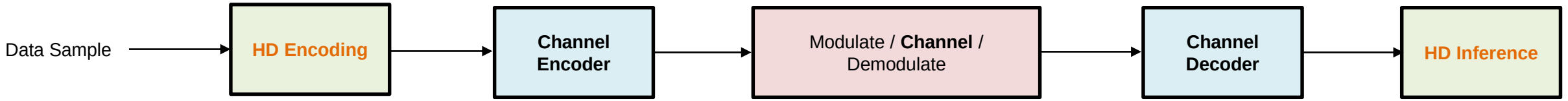
2) **Transmitting raw data w/ channel coding**



3) **HD encoding at the transmitter**



4) **HD encoding at the transmitter w/ channel coding**



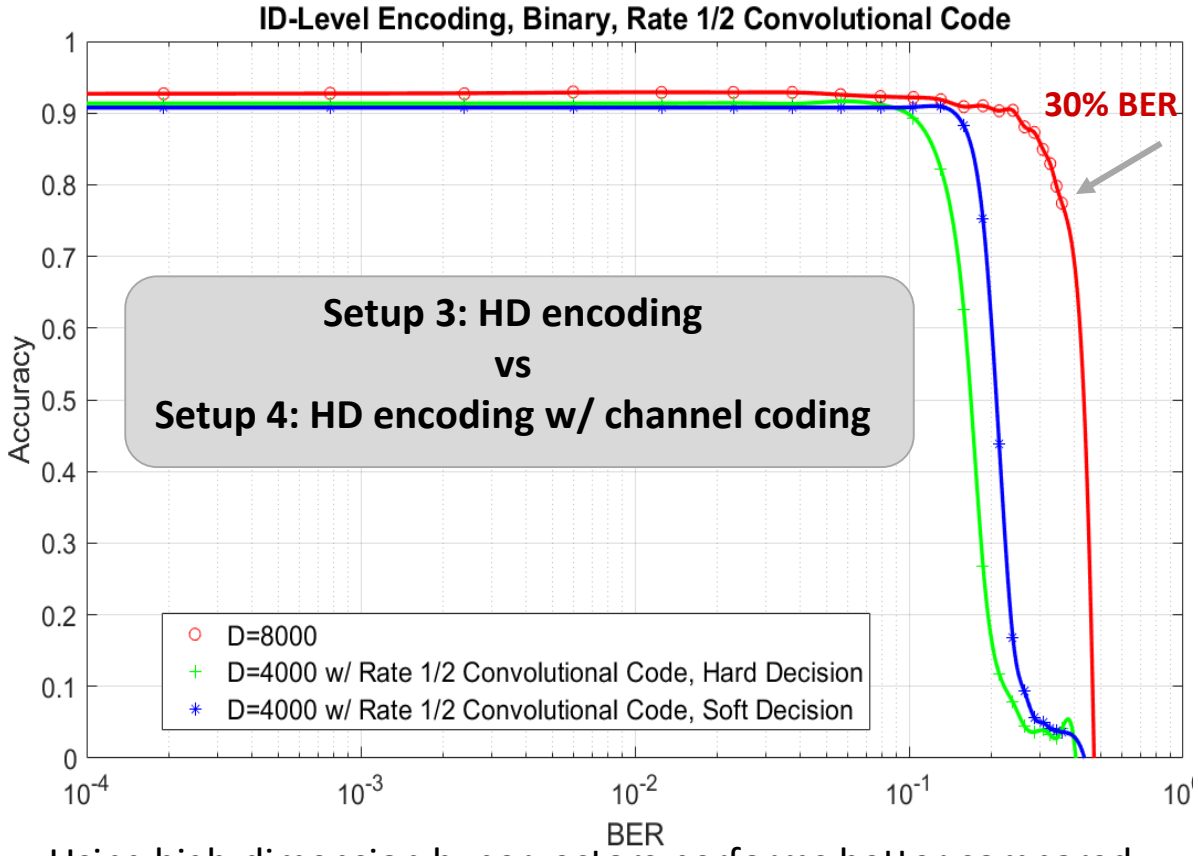
HDC Communication Robustness Analysis

ISOLET Speech Recognition with 617 features

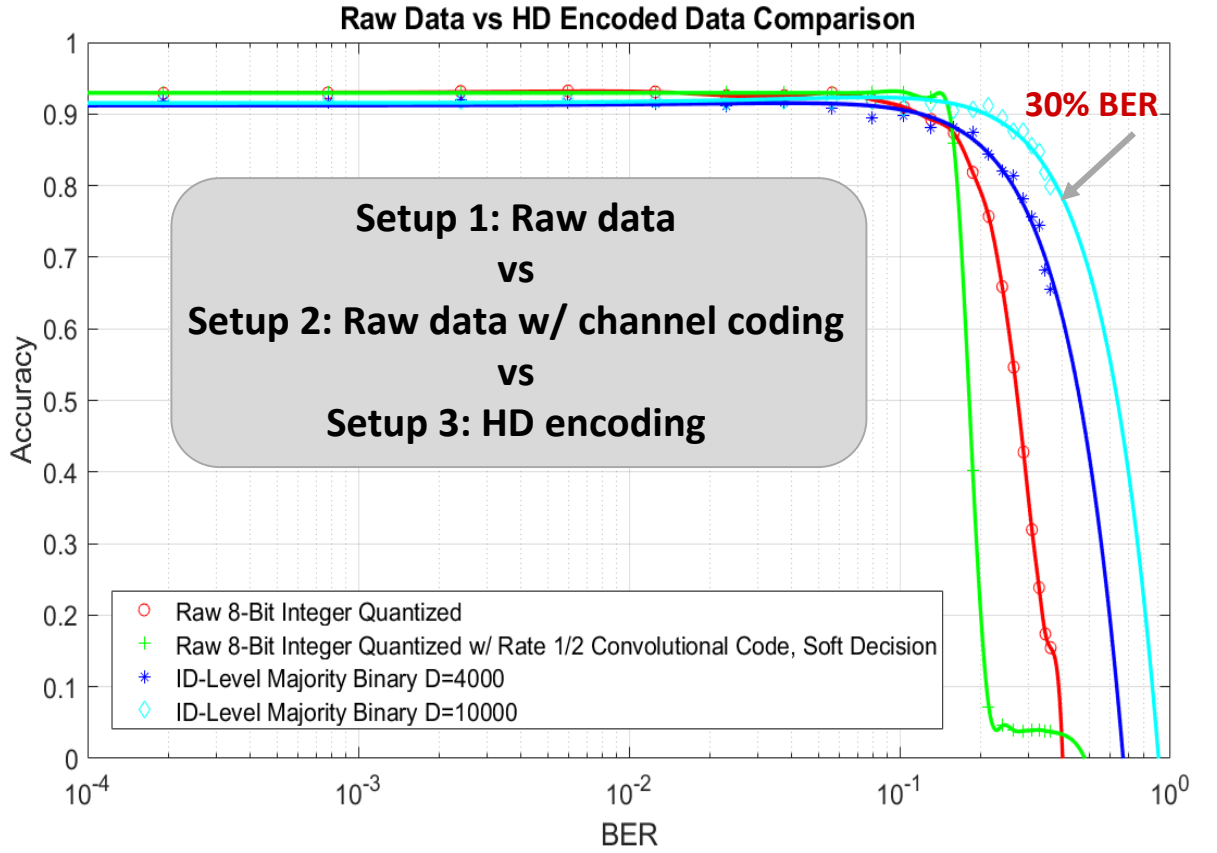
Communication Params: QAM modulation & AWGN channel

Bits Communicated:

Raw 8-bit Integer:	4,936bits
Raw 8-bit Integer w/ ½ Conv Code:	9,872bits
HDC ID-Level Encoding:	4,000 & 10,000bits

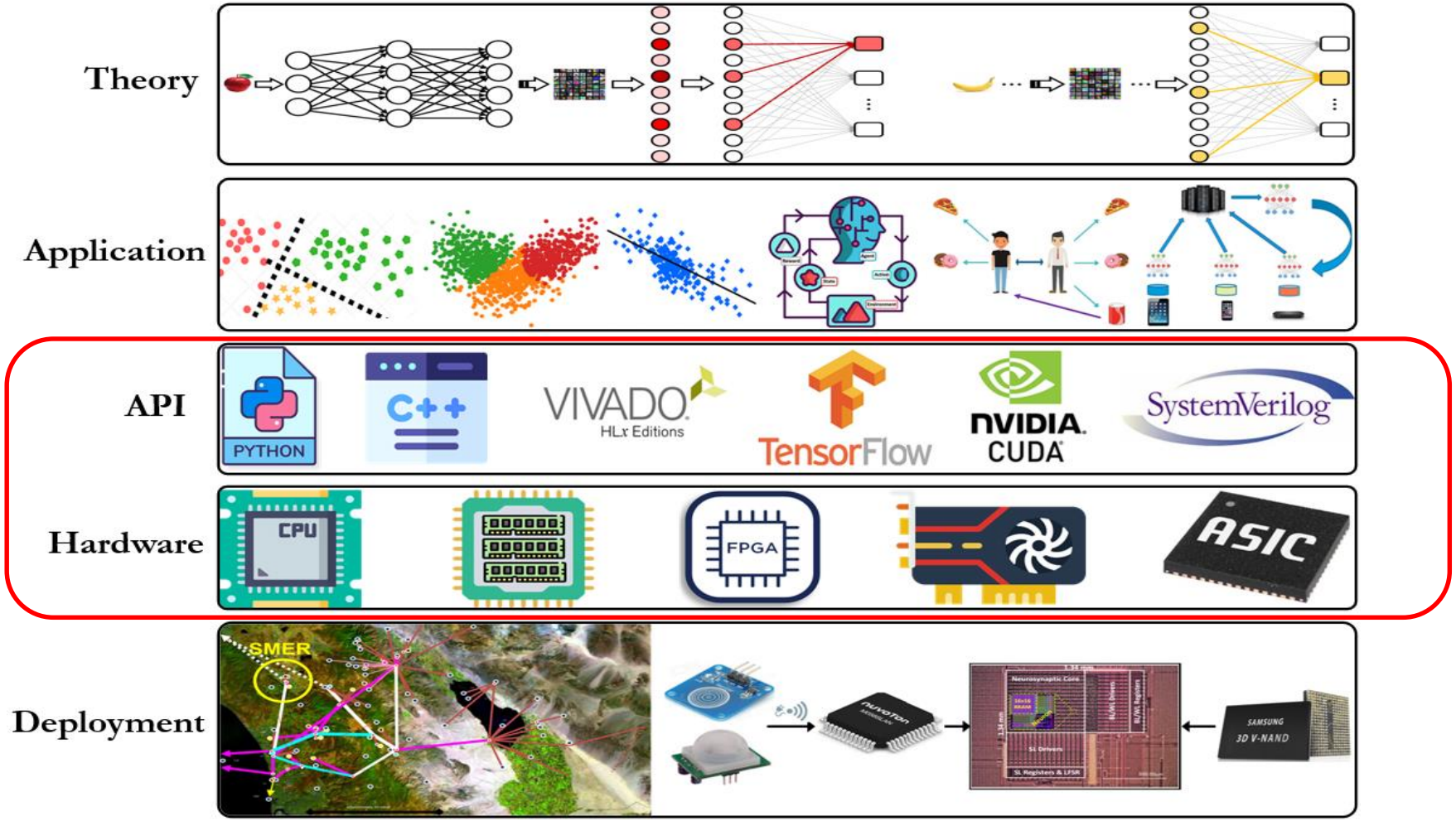


Using high-dimension hypervectors performs better compared to channel coding lower-dimension hypervectors



Transmitting HD encoded hypervectors is more robust to noise compared to channel coded raw data for the same number of communicated bits

Building HD Computing Systems



HD Computing Acceleration in Hardware: Experimental Setup

- HDC has been implemented on various hardware platforms
  - CPU: Intel i7-8700K with 16GB RAM
  - GPU: Nvidia GTX 1080 Ti with 11GB VRAM
  - FPGA: Kintex-7 (KC705) & Xilinx U280 FPGA
  - ASIC: HD encoding with PIM in ReRAM memory blocks
    - Simulations on 45 nm technology node in Cadence Virtuoso
    - VTEAM ReRAM model: Ron/Roff = 10k/10MΩ, SET/RESET time = 1.1ns

Release of HDC SW for CPU, GPU & HD2FPGA tool for FPGAs

https://github.com/UCSD-SEELab/HD-Classification

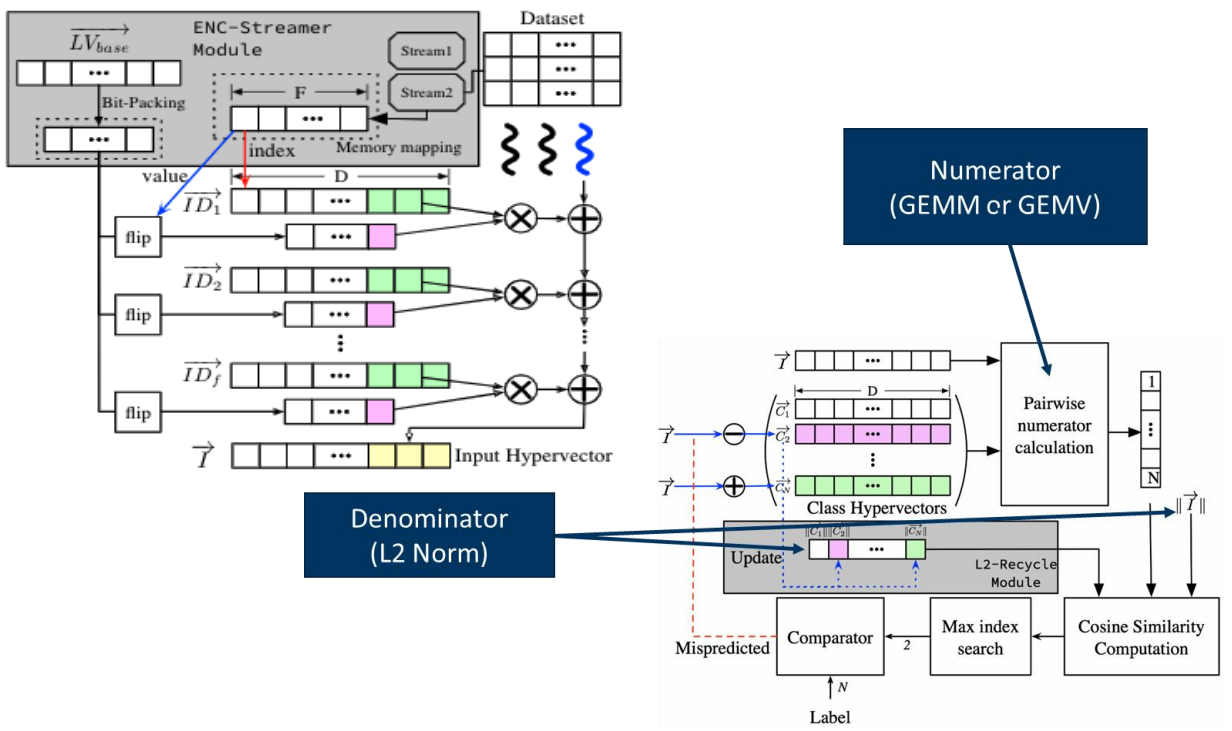
https://github.com/UCSD-SEELab/HD-Clustering

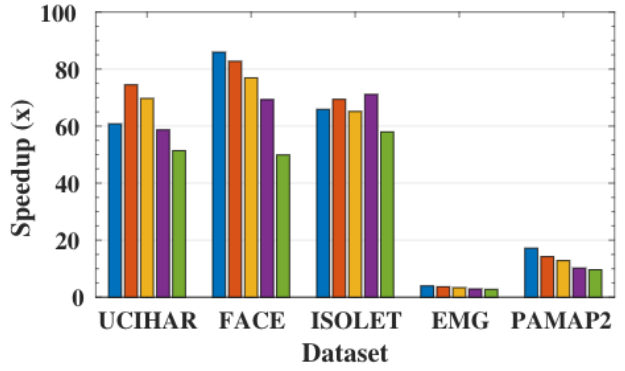
Classification Dataset	Description
ISOLET	Speech Recognition
UCIHAR	Activity Recognition
EMG5	Hand Gesture Recognition
Cardio3	Medical Diagnosis
Face	Face Detection

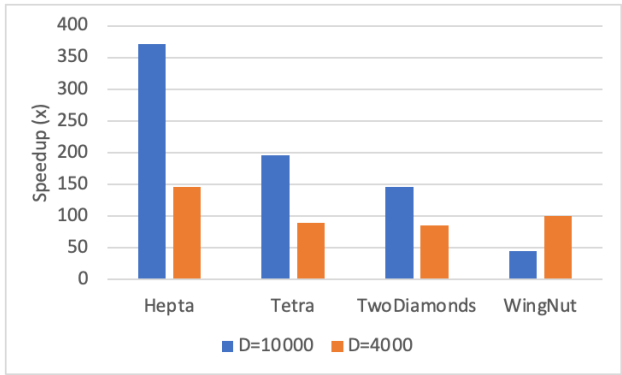
Clustering Dataset	Description
FCPS Hepta	Fundamental Clustering Ploblem Suite
FCPS Theta	
FCPS TwoDiamands	
FCPS WingNut	
Iris	Flower Clustering

OpenHD: Accelerating HDC Classification & Clustering on GPUs

- HDC classification & clustering only need two key components running on GPUs
  - Encoding
  - Similarity search & update
- HD classification is 86x faster and has 172x better energy efficiency vs. CPU
  - Encoding is 65x faster & training is 46x faster on average
- HD clustering is up to 371x faster with 133x better energy efficiency vs. CPU



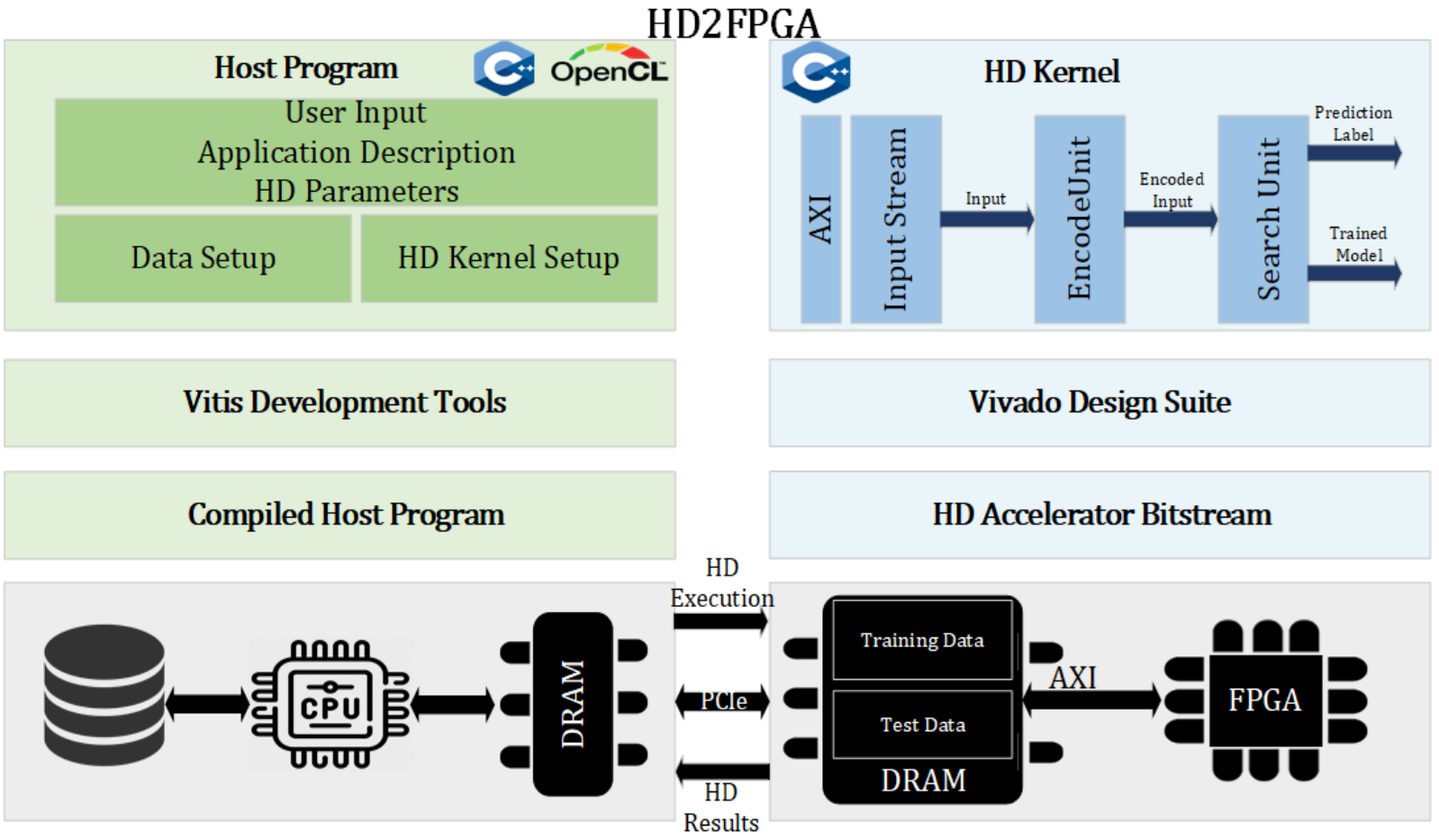




J. Kang, B. Khaleghi, Y. Kim, T. Rosing, "OpenHD: A GPU-Powered Framework for Hyperdimensional Computing," IEEE TC'22.

HD2FPGA

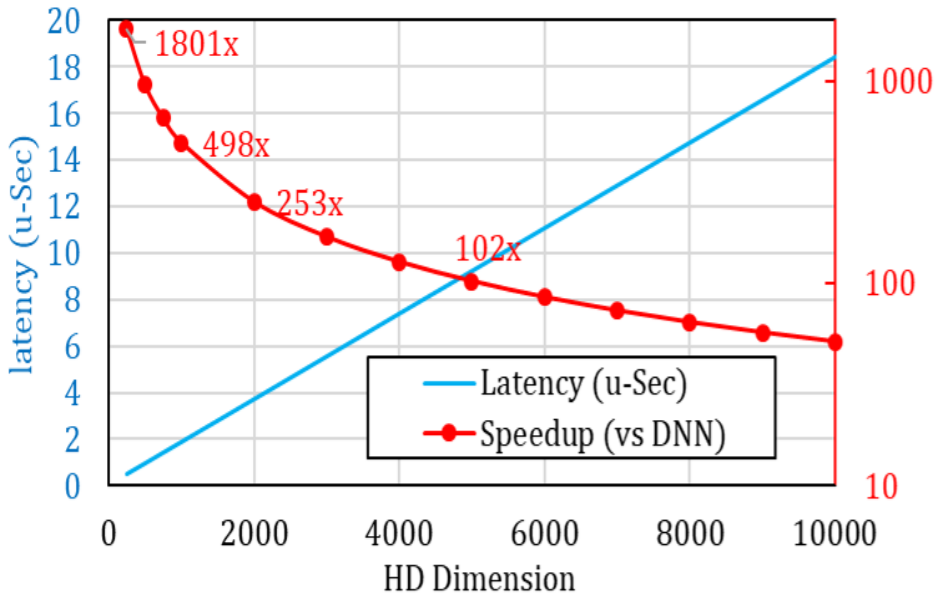
Automates mapping of HD classification & clustering to FPGAs

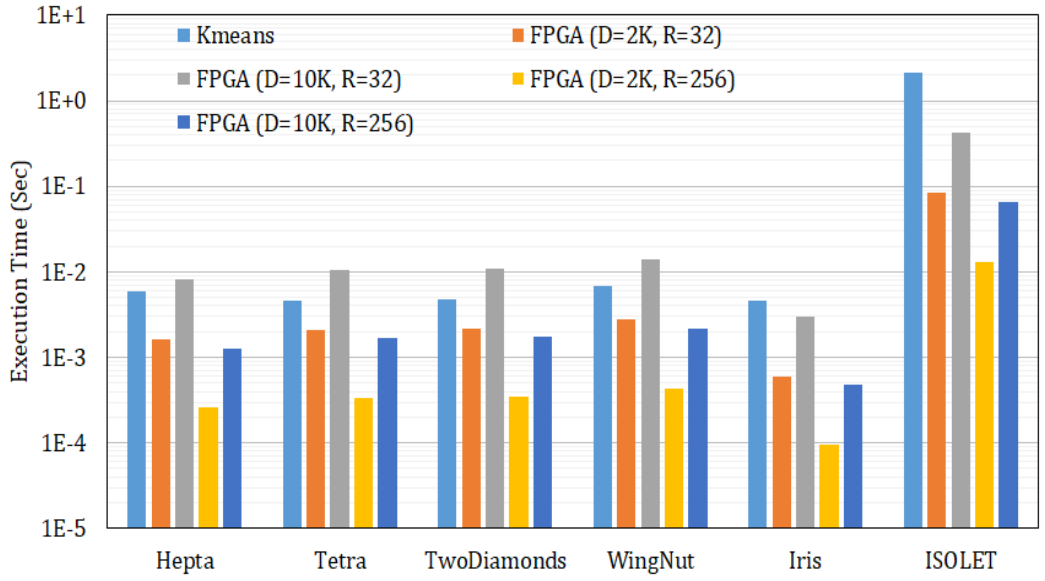


M Imani, Z Zou, S Bosch, S Anantha Rao, S Salamat, V Kumar, Y Kim, T Rosing, "Revisiting HD Learning for FPGA and Low-Power Architectures," HPCA'21
S Salamat, M Imani, T Rosing, "Accelerating hyperdimensional computing on FPGAs by exploiting computational reuse," IEEE TC, 2020.

HD2FPGA Classification & Clustering

- Classification: HD2FPGA with D=2k & R=256 vs. FPGA-based DNN accelerator[1] is 320x faster & consumes 127x less energy at comparable accuracy for ISOLET dataset
- Clustering: HD2FPGA is 46x faster vs. FPGA-based Kmeans [1] with D= 2k, Row=256



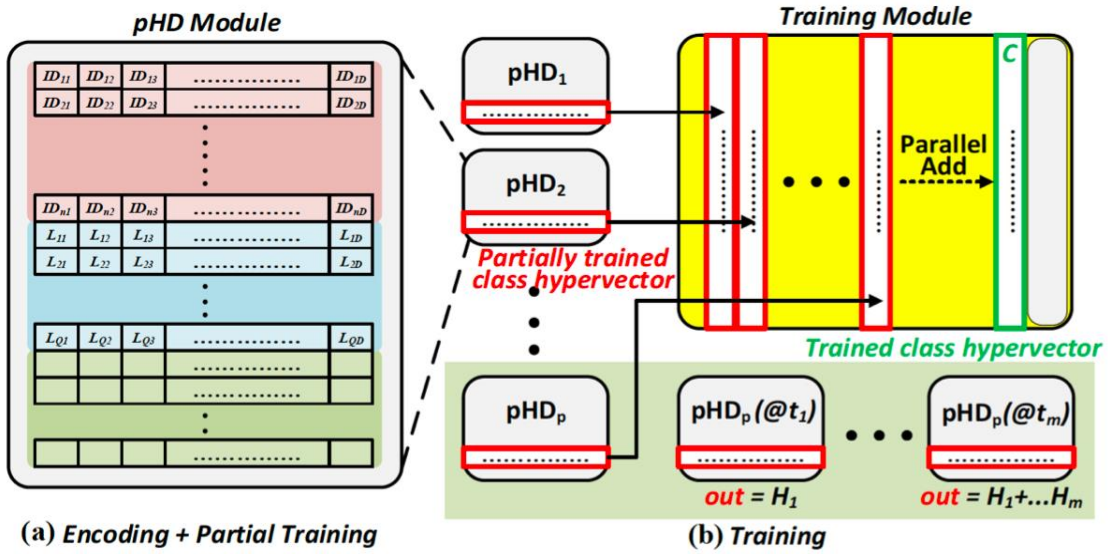


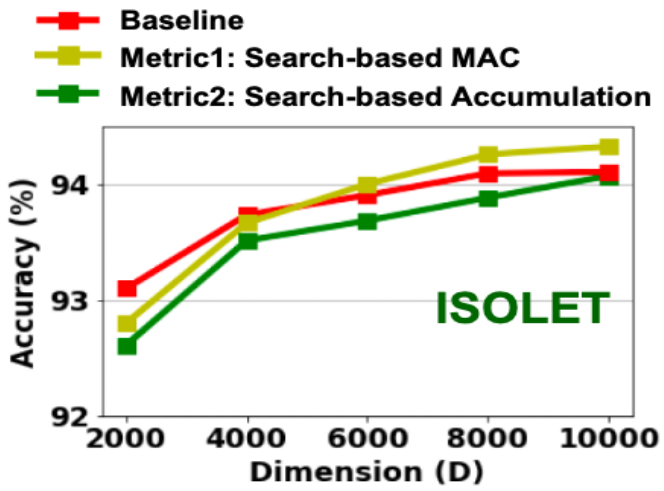
[1][Samragh'17]

Tri-HD: Train, Re-train, and Infer with HD in ReRAM

Accelerates HD encoding, training, retraining, and inference using processing in memory

- Dimension-wide and class-wise parallelism
- Partial training leveraged to fuse encoding and training phases together
- New approximate similarity metrics implemented with efficient search operations



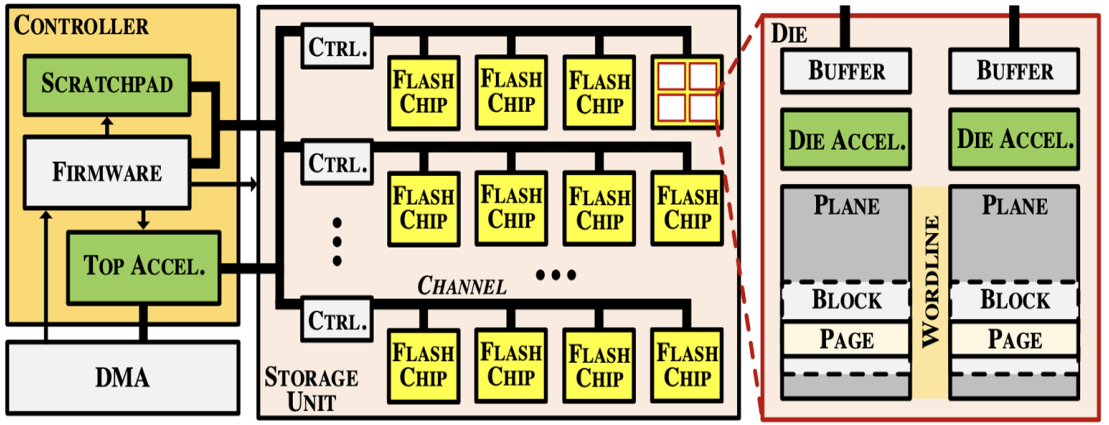


Effect of Approximate Similarity Metrics

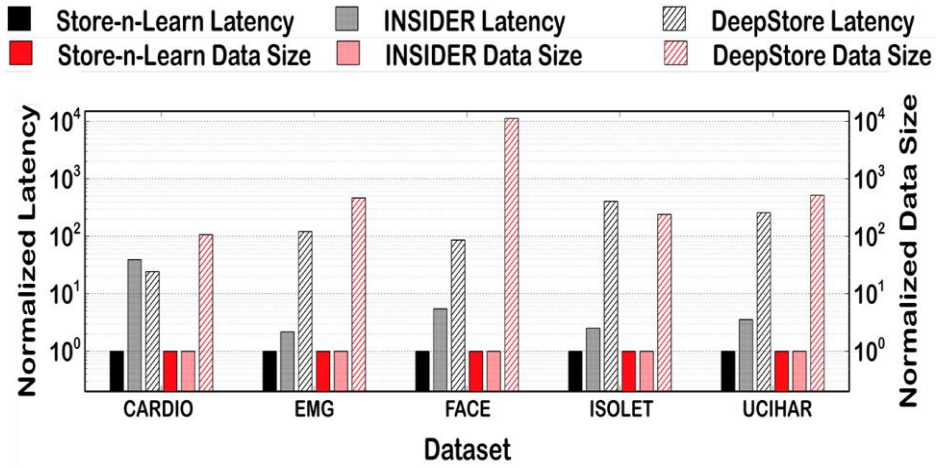
Tri-HD [TCAD'22] is **2,137x** faster & **3.6x** more energy-efficient vs. FloatPIM [ISCA'19] with comparable accuracy

Store-n-Learn: Accelerating HD Computing Classification and Clustering in Storage

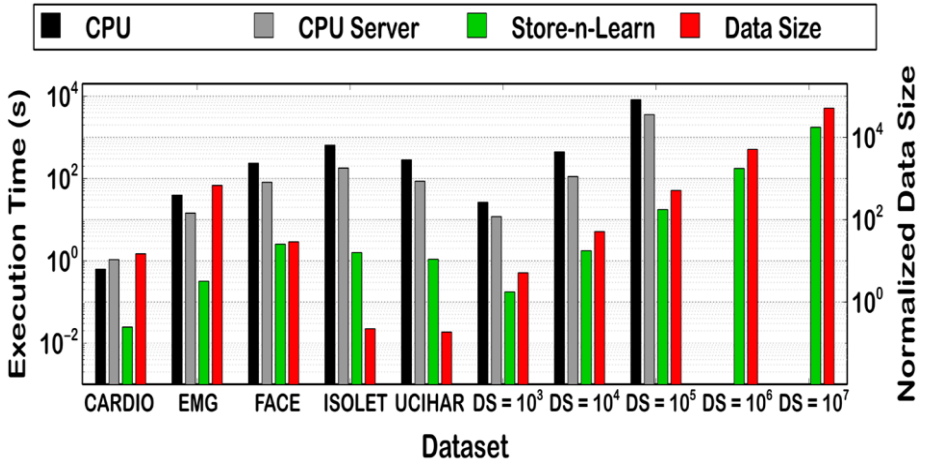
- Store-n-Learn leverages SSD hierarchy
  - HD Encoding accelerator at Flash plane
  - Top-level FPGA accelerator runs INSIDER [ATC'19] to do HD training, retraining, inference, and clustering



Store-n-Learn is 11x faster than INSIDER [ATC'19] and 179x faster than DeepStore [MICRO'21]



Store-n-Learn inference is 222x faster vs. CPU & transfers ~5000x less data



S. Gupta, B. Khaleghi, S. Salamat, J. Morris, R. Ramkumar, J. Yu, A. Tiwari, M. Imani, B. Aksanli, T. Rosing, "Store-n-Learn: Classification and Clustering with HD Computing across Flash Hierarchy," ACM TECS'22

HD Computing HW: Efficiency Improvements

- Our team developed HD libraries for CPU & GPU, HD2FPGA tool for mapping HD onto FPGAs; HD ASIC & PIM
  - Measurements on NVIDIA GPU & Xilinx u200 FPGA are compared to HDC running on Intel Xeon CPU; PIM is simulated using VTEAM model:

k_{on}	$-216.2m/sec$	$V_{T,ON}$	$-1.5V$	x_{off}	$3nm$
k_{off}	$0.091m/sec$	$V_{T,OFF}$	$0.3V$	R_{ON}	$10k\Omega$
$\alpha_{on}, \alpha_{off}$	4	x_{on}	0	R_{OFF}	$10M\Omega$

HD classification inference in PIM is **167,000x faster** vs. perceptron NN

Classification	Encoding		Training		Inference	
	Speedup	Energy	Speedup	Energy	Speedup	Energy
GPU	194x	34x	33x	14x	117x	21x
FPGA	1,930x	68,000x	580x	1,070x	280x	922x
PIM	**1,007,990x**	**594,523x**	**88,065x***	**164,718x***	**98,778x***	**82,200,185x***

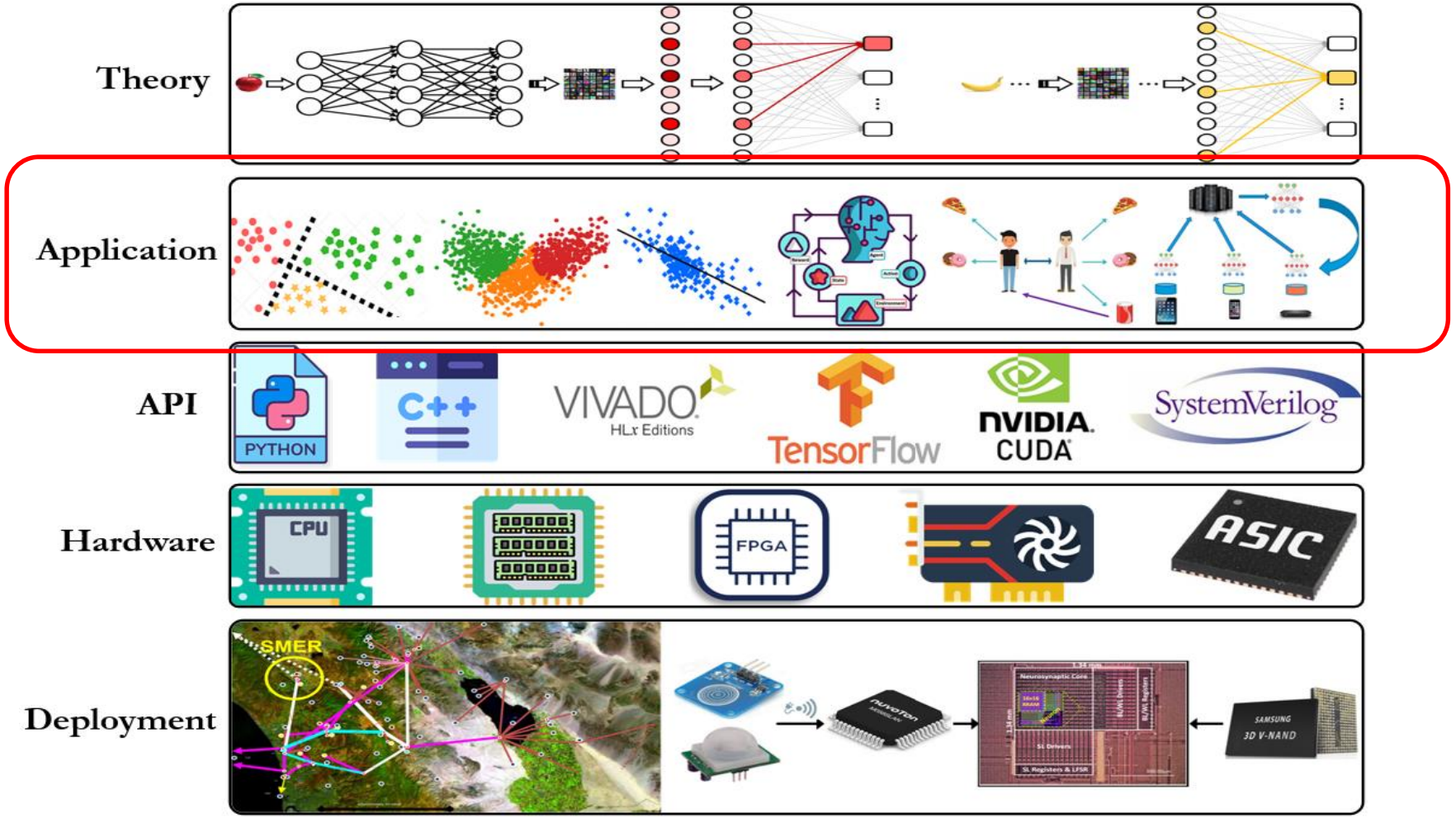
HD clustering in PIM is **855,000x faster vs. HD on CPU**

HD Clustering	Encoding		Clustering		Encoding + Clustering	
	Speedup	Energy	Speedup	Energy	Speedup	Energy
GPU	95x	15x	945x	226x	127x	21x
FPGA	141,800x	950,000x	55x	203x	196x	731x
PIM	2,177,000x	69,000x	332,000x	763x	855,000x	2,700x

HDC speedup vs. SoA @same accuracy

State of the Art	SVM	NN	MLP
Training	3	1.9	7.3
Testing	14	1.7	4.8

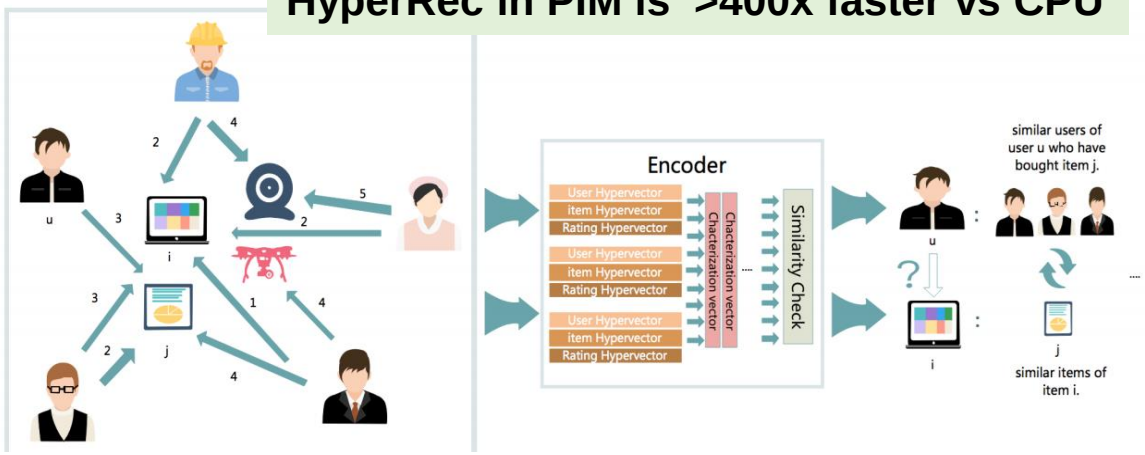
Building HD Computing Systems

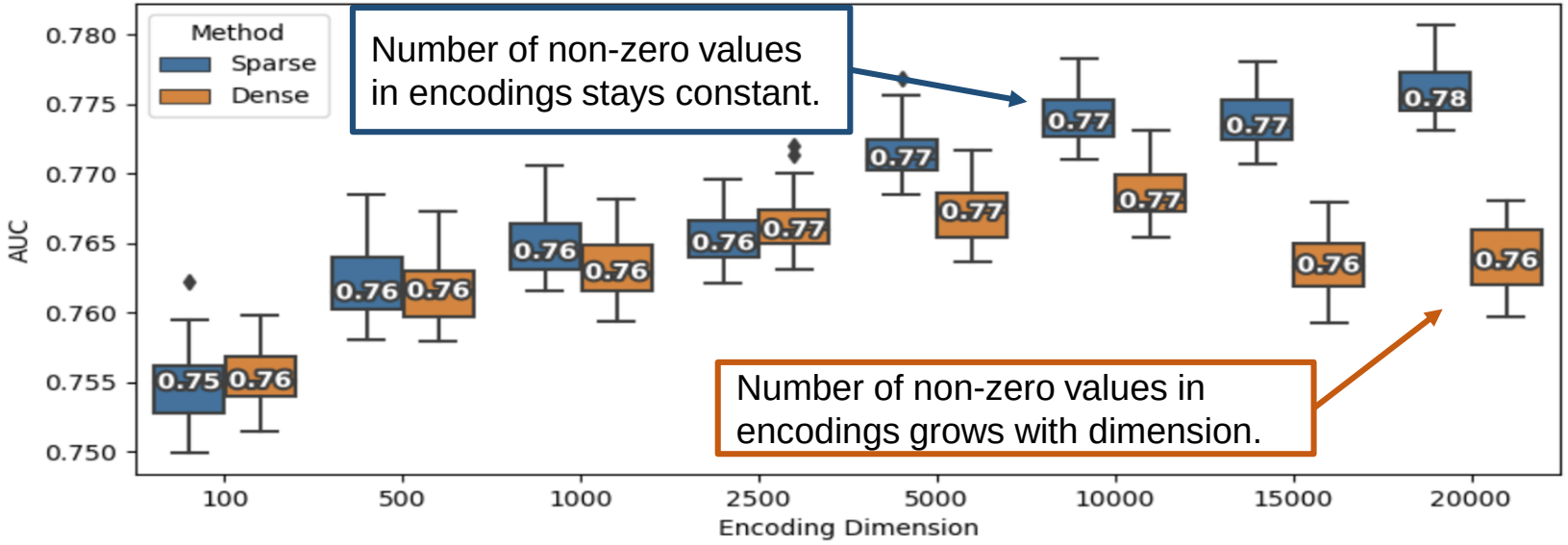


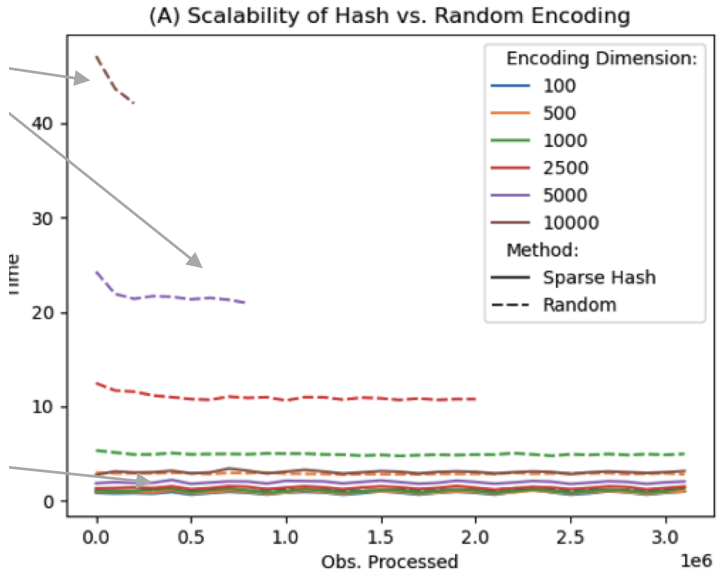
HyperRec: HD Computing Recommendation Systems at Scale

- HD Recommendation systems identify similar users & items by using their HD characterization vectors
  - Problem: generate HD encoding for tens of millions of symbols
  - Solution: Instead of storing codewords, we construct them “on-the-fly” by evaluating a handful of hash-functions
- Successfully tested on 1TB of data
  - Dataset has categorical features defined over a very large alphabet with hundreds of millions of symbols

HyperRec in PIM is >400x faster vs CPU



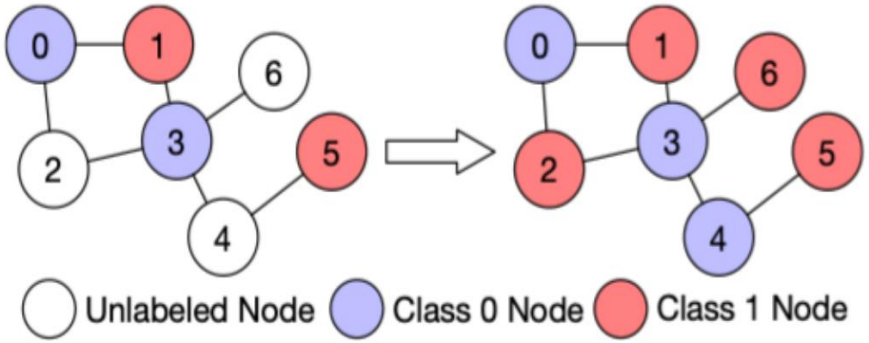




Y. Guo, S. Gupta, M. Imani, Y. Kim, J. Morris, T. Rosing, "HyperRec: Efficient Recommender Systems with Hyperdimensional Computing," ASPDAC 2021 & new paper on ArXiv

RelHD: Graph-based Learning with HDC

- Goal: Predict unlabeled node class with graph topology, labeled nodes, and node features
  - *Encoding*: most HD encoding works well
  - *Relation HV embedding*
    - Node hypervector **N**: Node feature encoding $\mathbf{I} = \mathbf{N} \odot \phi_0 + \mathbf{H}^1 \odot \phi_1 + \mathbf{H}^2 \odot \phi_2$
    - 1-hop $\mathbf{H^1}$ & 2-hop $\mathbf{H^2}$: Bundle node hypervectors of 1 & 2-hop neighbors
  - *HDC-based training:*
    - Aggregate relation hypervectors *I* corresponding to each class $\mathbf{C}_i = \sum_{j \in L} \mathbf{I}_j$
  - *Inference* predicts the label of an unlabeled node



RelHD-PIM is up to **82,000x faster** vs CPU; 10x faster & 990x more energy efficient vs PIM-GCN [ICCAD'21]

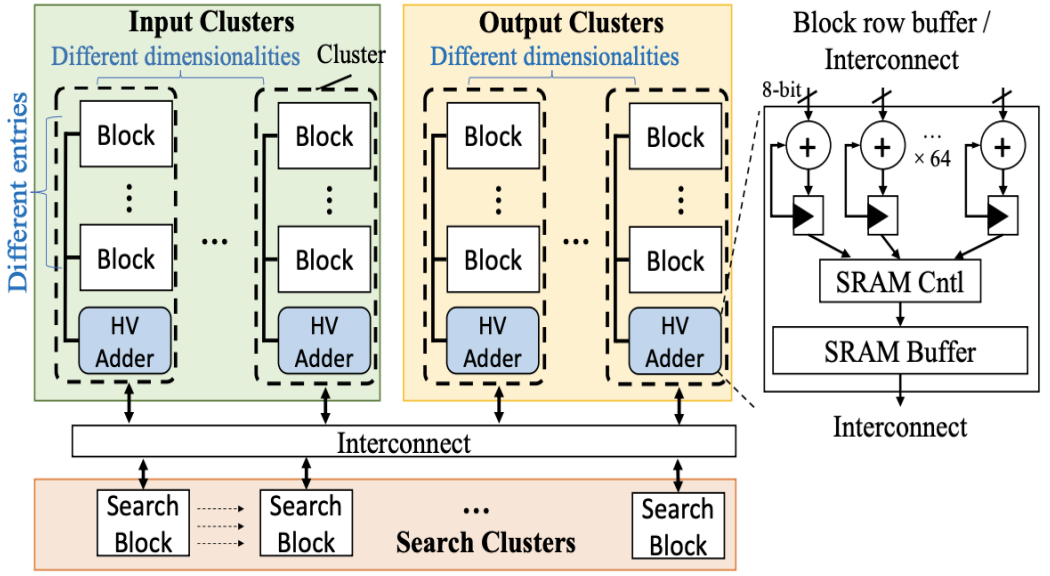
show comparable accuracy between RelHD & baselines:

- Graph convolutional network (GCN) [ICLR'17]
- Graph attention network (GAT) [ICLR'18]
- Node2Vec [SIGKDD'16]

Speedup of RelHD vs. Baseline Algorithms on GPU

RelHD vs ->	Node2Vec	GAT	GCN
Training	2,475x	82x	13x
Inference	35x	6.7x	5x

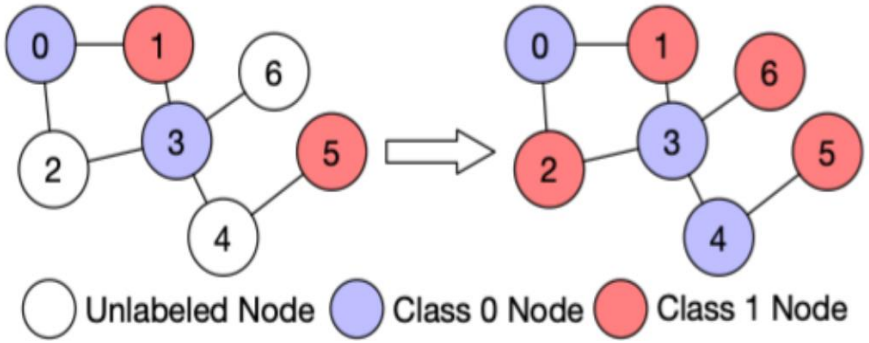
RelHD accelerated in memory using FeFET



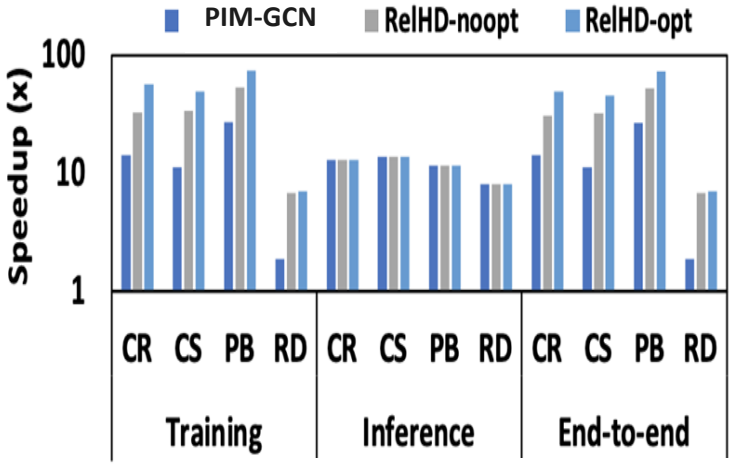
J. Kang, M. Zhou, A. Bhansali, W. Xu, A. Thomas, T. Rosing, "RelHD: A Graph-based Learning on FeFET with Hyperdimensional Computing," ICCD'22

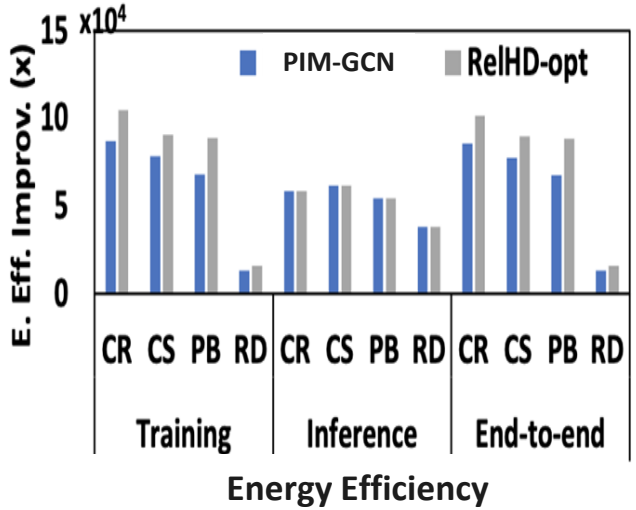
RelHD: Graph-based Learning with HDC

- Goal: Predict unlabeled node class with graph topology, labeled nodes, and node features
  - *Encoding*: most HD encoding works well
  - *Relation HV embedding*
    - Node hypervector **N**: Node feature encoding $\mathbf{I} = \mathbf{N} \odot \phi_0 + \mathbf{H}^1 \odot \phi_1 + \mathbf{H}^2 \odot \phi_2$
    - 1-hop $\mathbf{H}^1$ & 2-hop $\mathbf{H}^2$: Bundle node hypervectors of 1 & 2-hop neighbors
  - *HDC-based training:*
    - Aggregate relation hypervectors *I* corresponding to each class $\mathbf{C}_i = \sum_{j \in L} \mathbf{I}_j$
  - *Inference* predicts the label of an unlabeled node

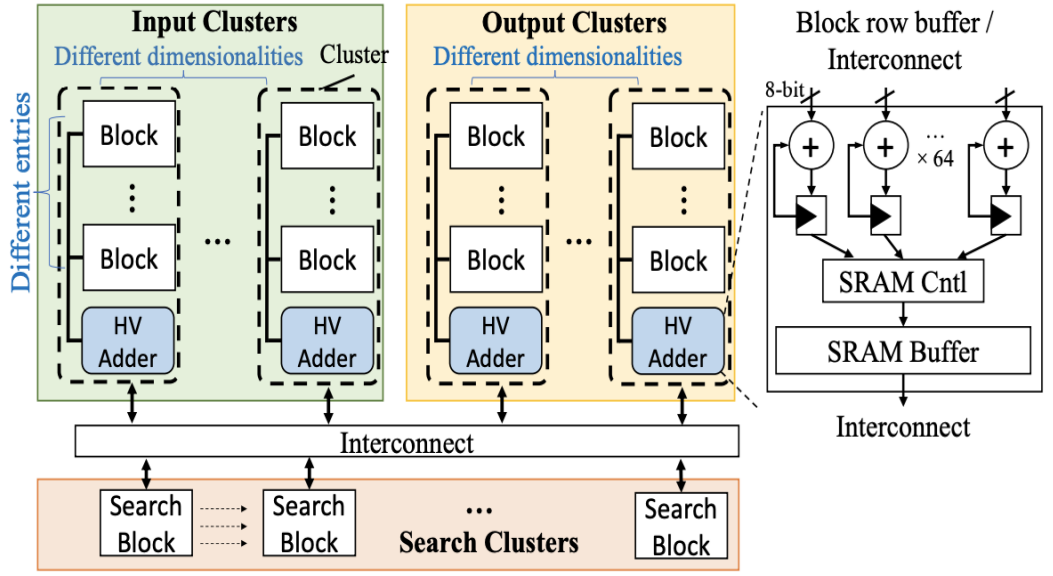


RelHD-PIM is up to **82,000x faster** vs CPU; 10x faster & 990x more energy efficient vs PIM-GCN [ICCAD'21]





RelHD accelerated in memory using FeFET



J. Kang, M. Zhou, A. Bhansali, W. Xu, A. Thomas, T. Rosing, "RelHD: A Graph-based Learning on FeFET with Hyperdimensional Computing," ICCD'22

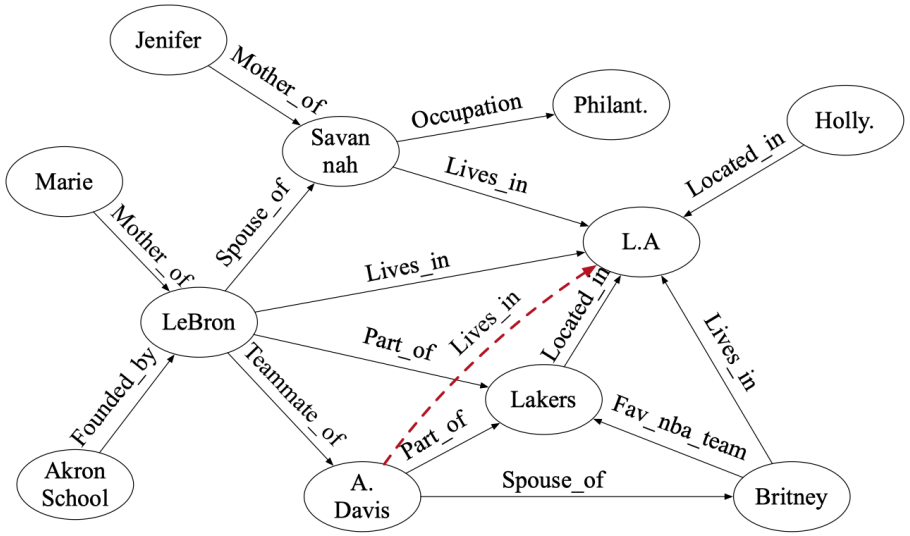
Symbolic Reasoning: Information Retrieval from Knowledge Graphs with HDC

- "Relation linking": Predict unseen nodes
  - GNN-like algorithms can help to extract underlying relationships
  - Inherit characteristics from source to target

- Approach: Inspired by the message-passing methodology of GNN
  - (1) Bind HV of edge and HV of a source node
  - (2) Bundle the result with the HV of the target node
  - (3) Iterate until pairwise similarities between all nodes converge

- Empirically, within three iterations on small graphs
  - Embedding relationship between nodes
  - Define HV(Lives_in) as HV(Part_of) * HV(Located_in)

$$\arg\max_i \delta(HV(node_i), HV(A.Davis) * HV(Lives\_in)) = L.A.$$



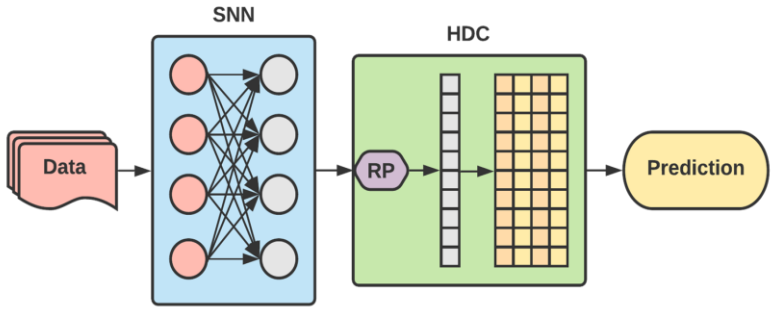
An example of knowledge graph completion:
Query relation: Lives_in, head entity: A. Davis,
Reasoning result: L.A

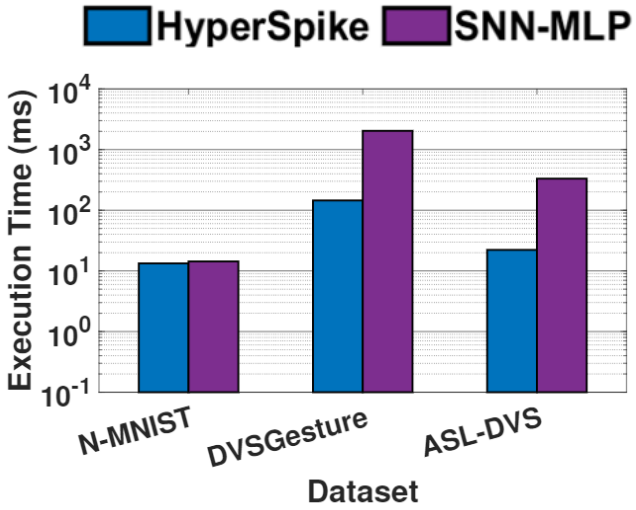
An example of knowledge graph question answering:
Question: Where do the spouses of the teammates of Lakers usually live?
Reasoning result: L.A

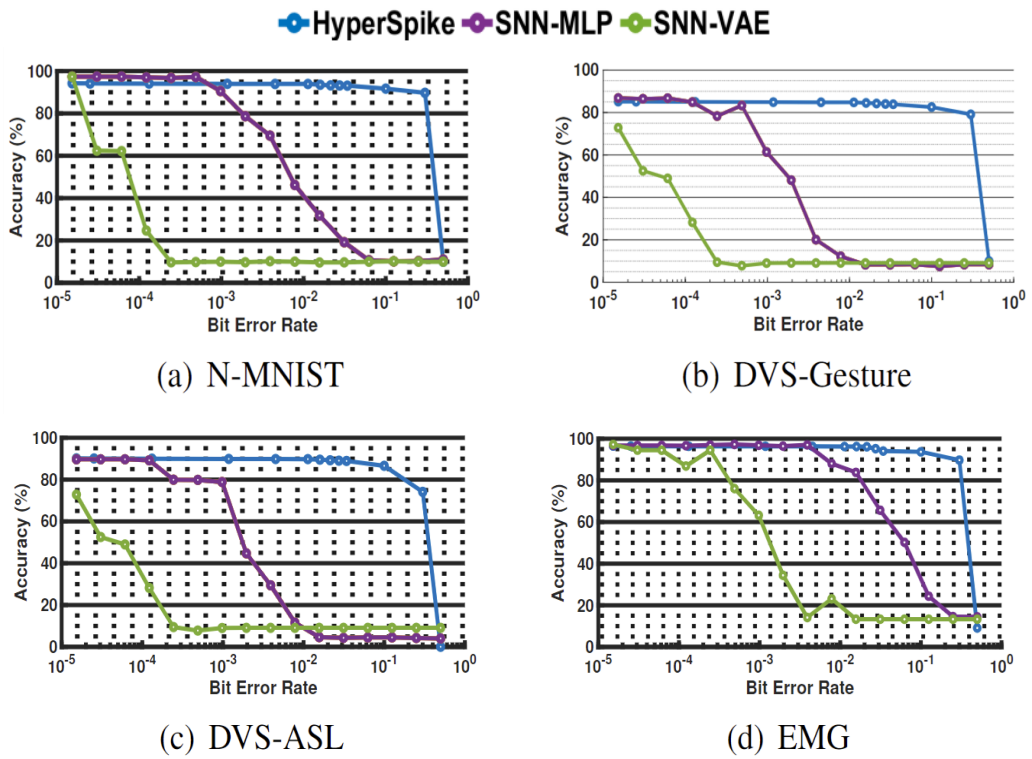
Figure source: Zhang et al., Neural, Symbolic and Neural-Symbolic Reasoning on Knowledge Graphs, AI Open Journal 2021

HyperSpike: HD Computing with Spiking Neural Networks

- Single layer **untrained SNN** extracts features
- HDC does reasoning
  - Random projection, binary quantization, Hamming distance
- Implemented using Intel Loihi & TinyHD [DATE'21]
  - **15x faster** and 4.6x more energy efficient
  - **58x more robust** at 3.4% BER







J Morris, HW Lui, K Stewart, B Khaleghi, A Thomas, T Marback, B Aksanli, E Neftci, T Rosing, "HyperSpike: HyperDimensional Computing for More Efficient and Robust Spiking Neural Networks, DATE'22.

HDnn: Image Classification

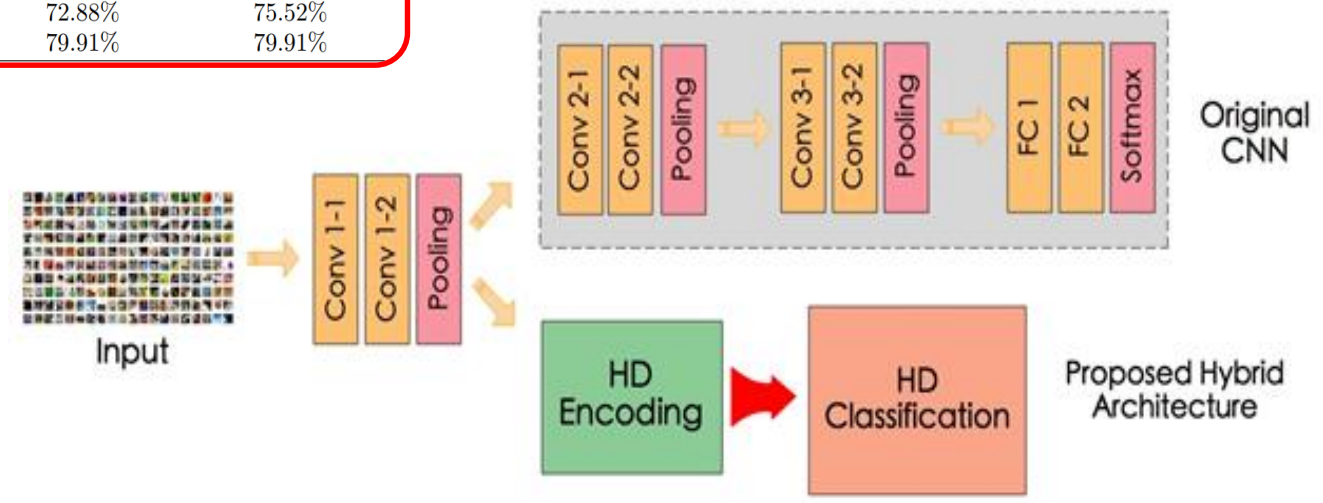
- Combine HD with a feature extractor derived from the CNN
  - Prune and cut many of CNN layers; add HD as the last layer

Accuracy comparison of HD, CNN, and HDnn.

Model↓ Dataset→	MNIST	CIFAR-10	CIFAR-100	Flowers
HD (RP) [9]	94%	26.9%	9%	19.6%
HD (non-linear) [10]	97%	45.5%	27.7%	31.5%
StocHD [11]	98%	N/A	N/A	N/A

Dataset	Baseline VGG16	binary HDnn
22 super categories of ImageNet	72.88%	75.52%
22 vehicle subcategories of ImageNet	79.91%	79.91%

HDnn runs ImageNet & CIFAR as accurately as CNNs at lower cost



[9] M. Imani, J. Morris, *et al.*, "Bric: Locality-based encoding for energy-efficient brain-inspired hyperdimensional computing," in *56th Annual Design Automation Conference*, pp. 1–6, 2019.

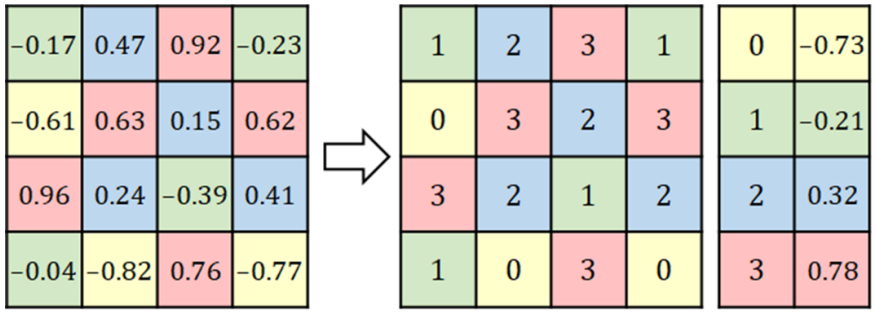
[10] Z. Zou, Y. Kim, M. H. Najafi, and M. Imani, "Manihd: Efficient hyper-dimensional learning using manifold trainable encoder," in *2021 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 850–855, 2021.

[11] P. Poduval, Z. Zou, H. Najafi, H. Homayoun, and M. Imani, "Stochd: Stochastic hyperdimensional system for efficient and robust learning from raw data," in *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pp. 1195–1200, 2021.

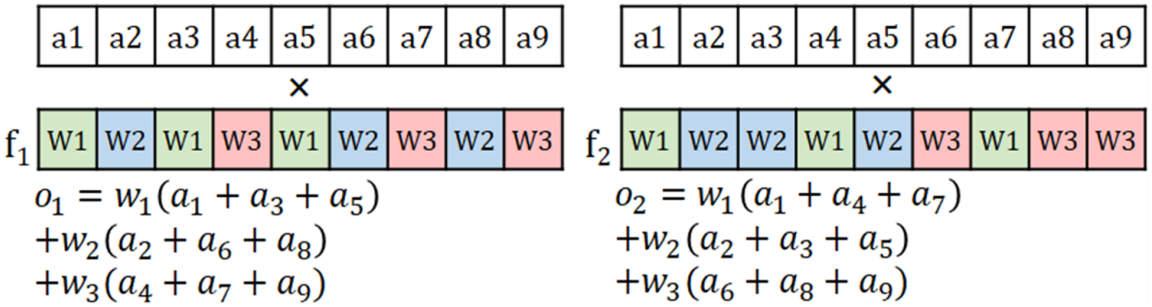
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Compact Feature Extractor: PatterNet [DAC'22]

Weight clustering concept



Conventional weight clustering

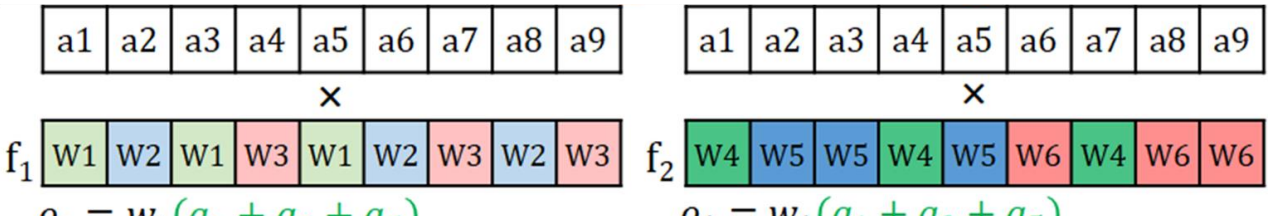


$o_1 = w_1(a_1 + a_3 + a_5)$
$+w_2(a_2 + a_6 + a_8)$
$+w_3(a_4 + a_7 + a_9)$

$o_2 = w_1(a_1 + a_4 + a_7)$
$+w_2(a_2 + a_3 + a_5)$
$+w_3(a_6 + a_8 + a_9)$

Total F×N ADD, F×W MUL, F×N index memory

Patterned weight clustering



$o_1 = w_1(a_1 + a_3 + a_5)$
$+w_2(a_2 + a_6 + a_8)$
$+w_3(a_4 + a_7 + a_9)$

$o_4 = w_4(a_1 + a_3 + a_5)$
$+w_5(a_2 + a_6 + a_8)$
$+w_6(a_4 + a_7 + a_9)$

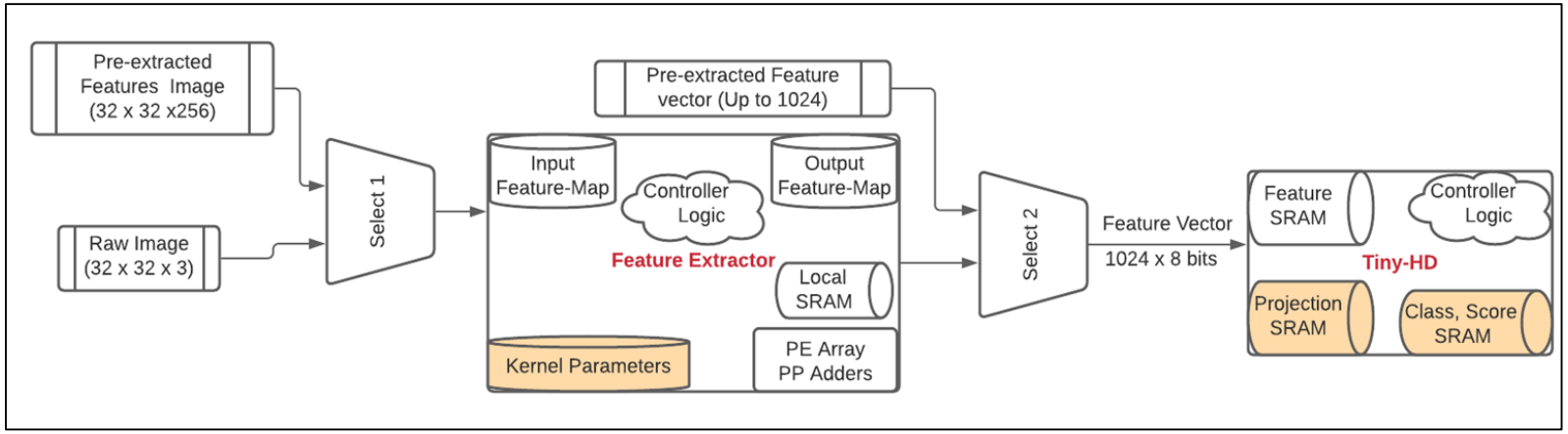
Total N ADD, F×W MUL, N index memory

	Model	Accuracy			Operation (M)			Parameters (MB)		
		Base	Hrank	PatterNet	Base	Hrank	PatterNet	Base	Hrank	PatterNet
Mini-ImageNet	VGG16	56.95%	53.16%	55.90%	1272	549 (56.8%)	355 (72.0%)	22.79	14.1 (38.2%)	11.7 (48.4%)
	Resnet18	62.28%	60.97%	62.20%	2221	1364 (38.6%)	854 (61.5%)	10.74	5.52 (48.5%)	3.14 (70.7%)
	Resnet50	64.20%	62.65%	63.88%	5192	2727 (47.5%)	1629 (68.6%)	22.75	12.3 (45.8%)	8.31 (63.4%)
	MobV2	55.19%	53.17%	54.53%	218.8	136.1 (37.8%)	140.9 (35.6%)	2.34	1.53 (34.6%)	1.23 (47.4%)

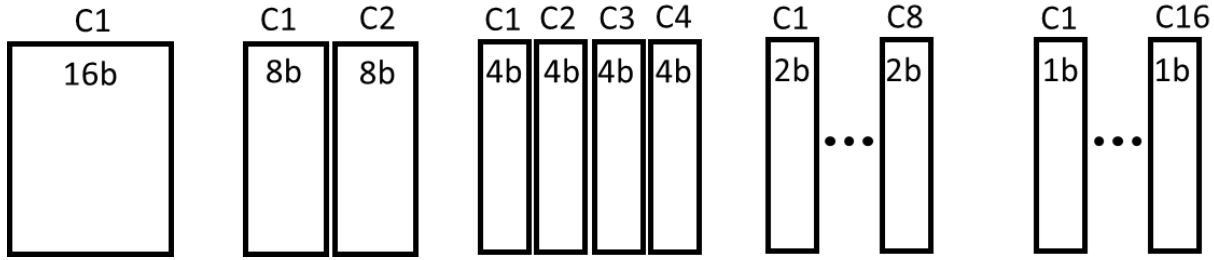
Dataset	MAC Reduction	Parameter Reduction	Accuracy
Mini-ImageNet	36% – 72%	47% – 71%	–0.53% (vs –2.2%)
CIFAR10	64% – 72%	64% – 78%	–0.1% (vs –1.0%)
CIFAR100	56% – 73%	64% – 77%	–0.5%

Behnam Khaleghi, Uday Mallappa, Duygu Nur Yaldiz, Haichao Yang, Monil Shah, Jaeyoung Kang and Tajana Rosing, "PatterNet: Explore and Exploit Filter Patterns for Efficient Deep Neural Networks," DAC'22

GENERIC: Flexible & Efficient HDC in CMOS [DAC'22]



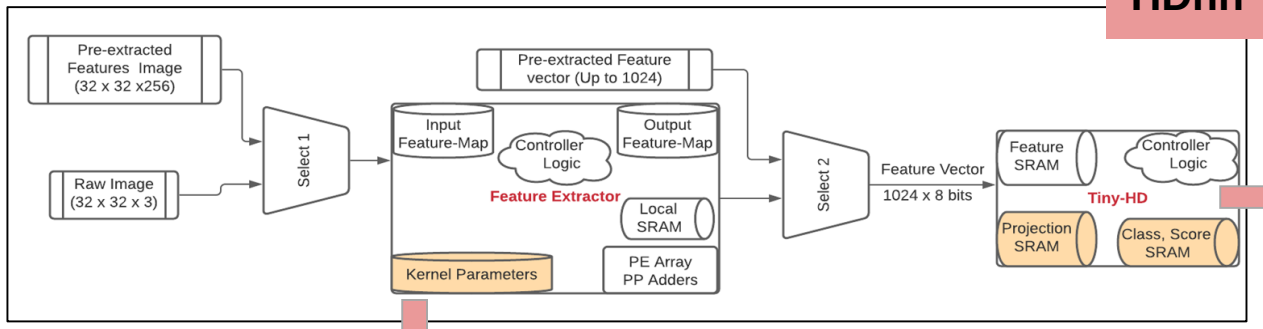
- For the given total memory capacity, we can configure arbitrary vector length (D), class count (C), and class precision (W)
  - D×C×W = 256KB
- Training needs W=16b
  - 128 classes of 1K, 32 classes of 4K

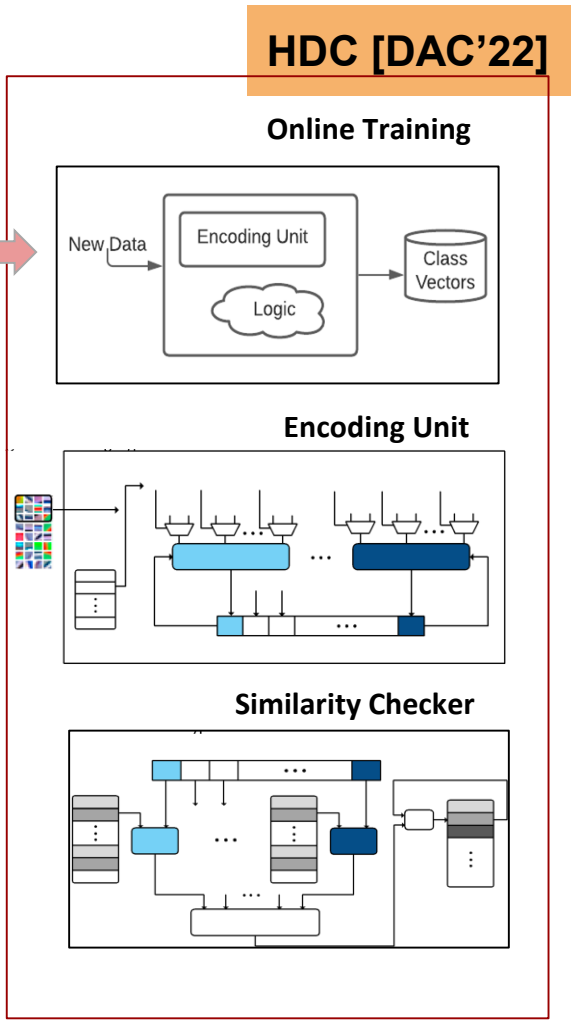


- GENERIC design 1.06mm^2, peak 9mW at 250MHz, processes 18,000 MNIST/second
  - 0.6mm^2 with 128KB memory, much less non-peak and inference power (1–2mW)

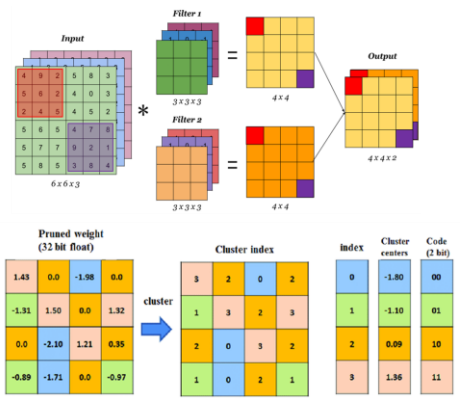
Behnam Khaleghi, Jaeyoung Kang, Hanyang Xu, Justin Morris and Tajana Rosing, "GENERIC: Highly Efficient Learning Engine on Edge using Hyperdimensional Computing," DAC'22.

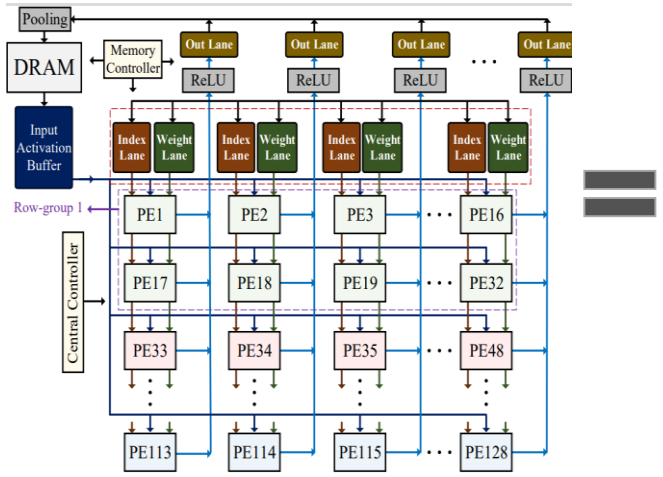
HDnn ASIC overview



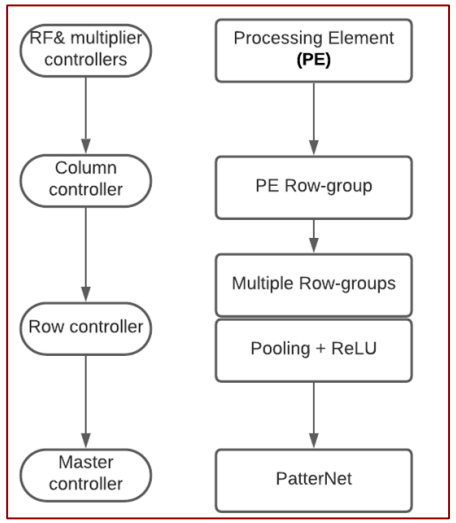


PatterNet [DAC'22]





Hierarchy of modules



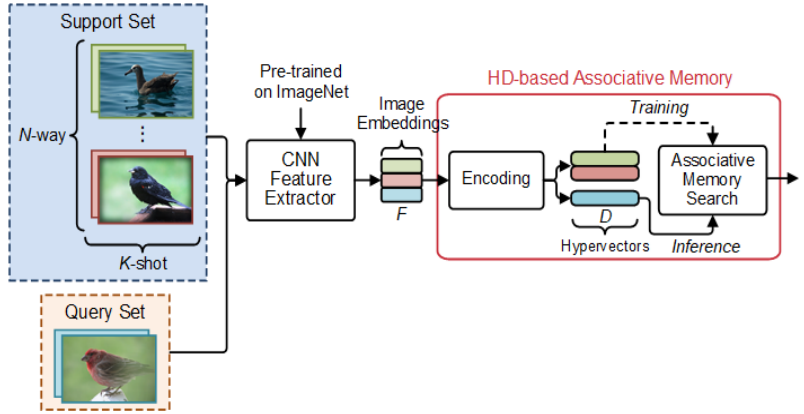
HDnn runs ImageNet at 3.2 TOPS/W with accuracy of up to 79%

Behnam Khaleghi, Uday Mallappa, Duygu Nur Yaldiz, Haichao Yang, Monil Shah, Jaeyoung Kang and Tajana Rosing, "PatterNet: Explore and Exploit Filter Patterns for Efficient Deep Neural Networks," DAC'22

Behnam Khaleghi, Jaeyoung Kang, Hanyang Xu, Justin Morris and Tajana Rosing, "GENERIC: Highly Efficient Learning Engine on Edge using Hyperdimensional Computing," DAC'22.

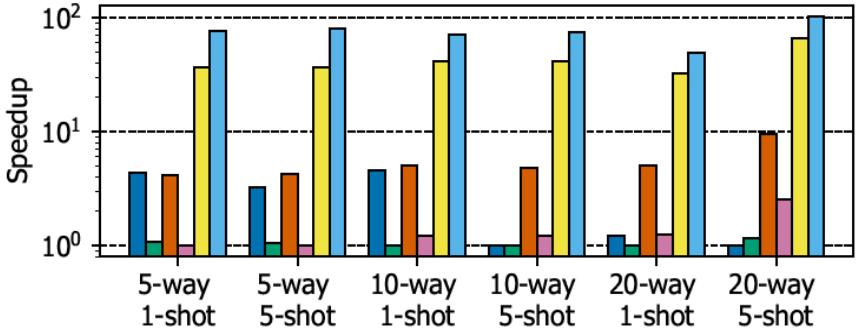
FSL-HD: Few Shot Learning with HDC [DATE'23]

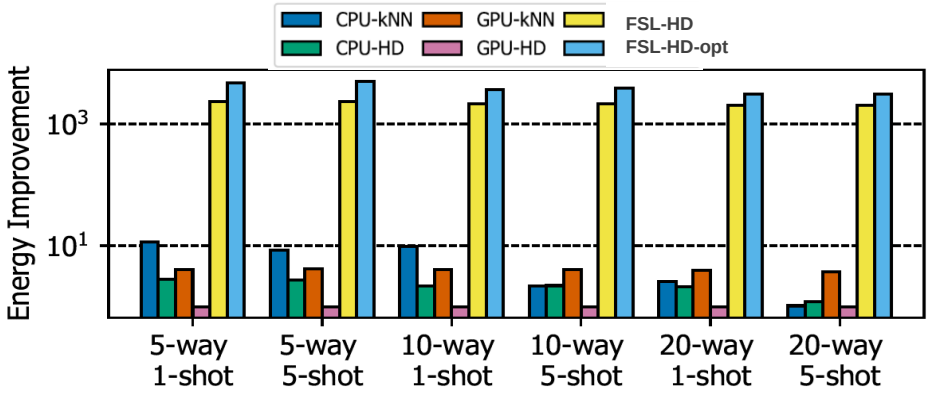
- **Few-shot learning (FSL):** classifying new data with only a few training samples
  - **N**-way, **K**-shot: **N** new classes & **K** training samples for each class
  - **CNN feature extractor** with fixed pre-trained weights
  - **HDC-based classifier** as the associative memory



- **Experimental setup:**
  - **Baselines:** MLP[ICCV'21], kNN-L1[ITED'21], kNN-Cosine[ICLR'19]
  - **Datasets:** CIFAR-100, Describable Textures DTD, Caltech-UCSD Birds 200, Traffic Sign, Omniglot
  - **Technology:** 40nm 1T1R ReRAM array [ISSCC'22] and current sense amplifier [JSSC'16] scaled to 40nm; other circuits are synthesized with ARM 40nm library

FSL-HD & FSL-HD-opt (with progressive search)
20x faster & 3,000x more energy efficient
vs. state of the art at comparable accuracy





W. Xu, J. Kang, T. Rosing, "FSL-HD: Accelerating Few-Shot Learning on ReRAM using Hyperdimensional Computing," DATE'23.

FSL Accuracy vs. Bit Error Rate

- Experimental Setup:
  - FSL setting: 10-way, 5-shot
  - Front-end CNN model: ResNet-18 with 512 features
  - Back-end HD classifier: D=2,048; binary
- Datasets: CIFAR-100 and Caltech-UCSD Birds 200
  - CIFAR-100 has 100 classes
  - Caltech-UCSD Birds has 200 classes for birds
- Bit-error rate (BER) is introduced to the HD hypervectors

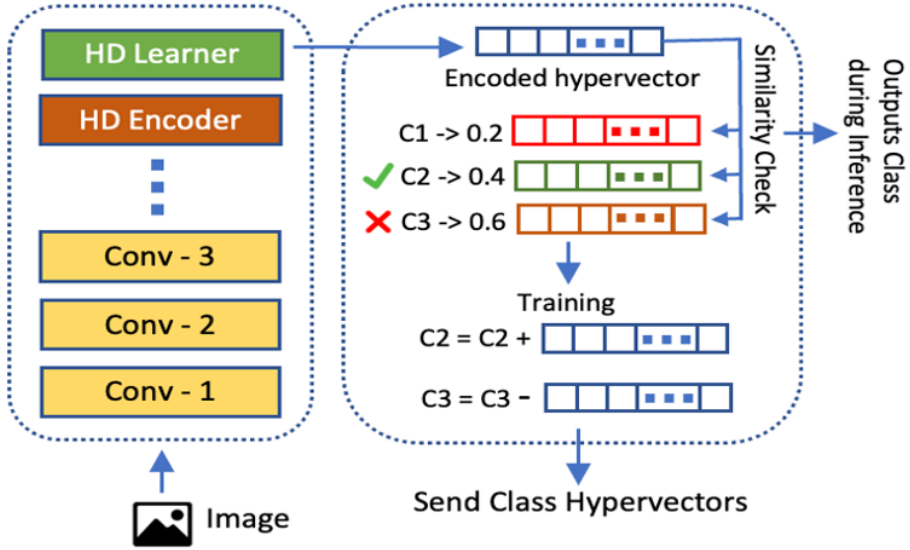
BER	0	2%	6.5%	11.0%	15%
CIFAR-100	61.4%	61.2%	61.0%	59.2%	54.5%
CU-Birds 200	85.1%	84.8%	84.3%	82.6%	77.9%

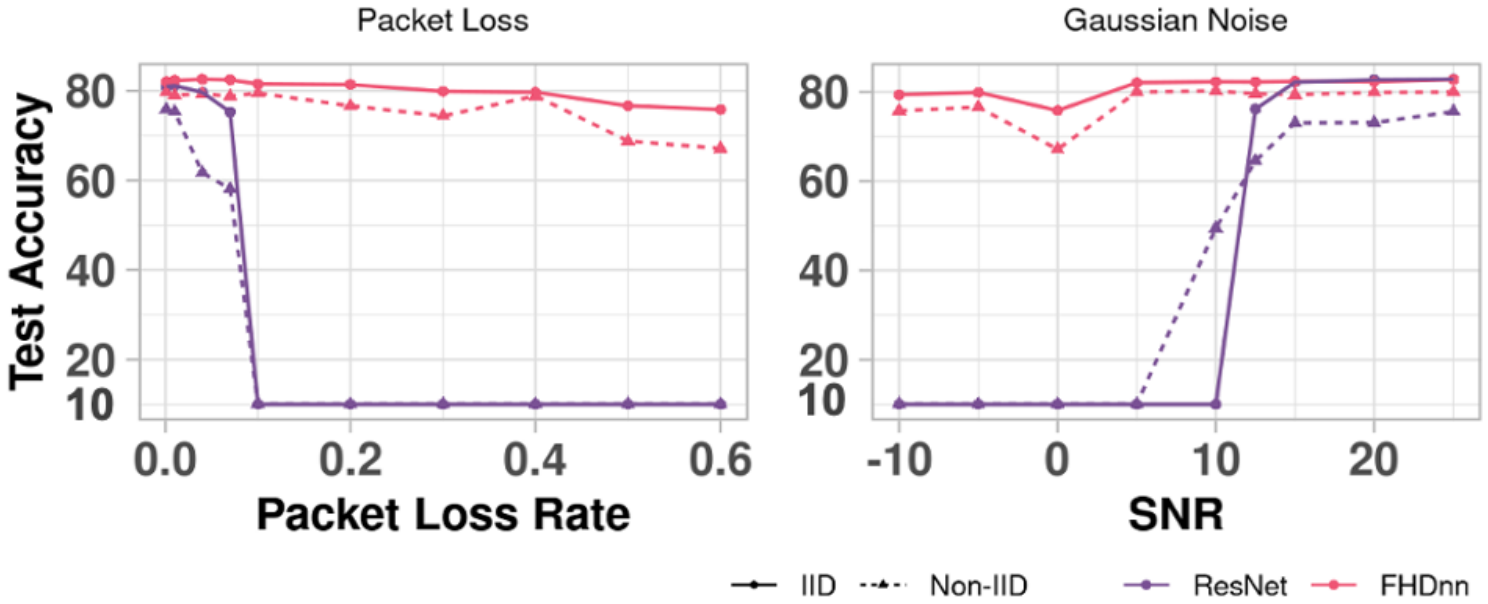
Amazing resilience of HD -> accuracy stable even in very high BER regimes!

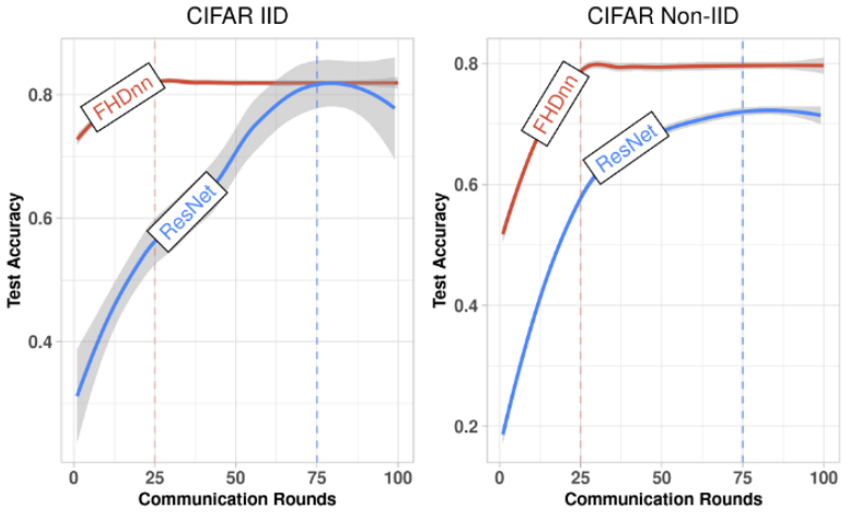
W. Xu, J. Kang, T. Rosing, “FSL-HD: Accelerating Few-Shot Learning on ReRAM using Hyperdimensional Computing,” DATE’23.

FHDnn: Federated Learning with HD Computing

- HDC federated learning has a few key steps:
  1. Get features and encode into hypervectors
     - Leverage contrastive learning framework via SimCLR
  2. Local training via simple perceptron-like method
  3. Update global model via federated bundling
- FHDnn gets comparable accuracy with ResNet, but is faster and more robust to noise



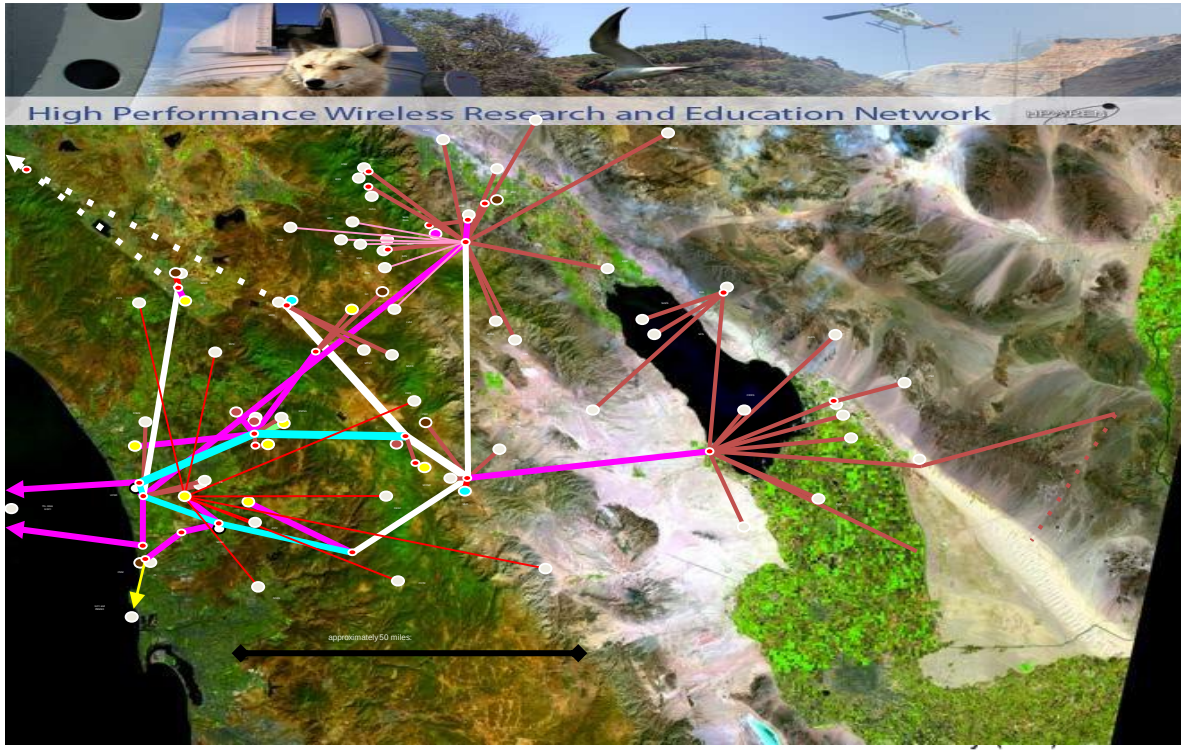


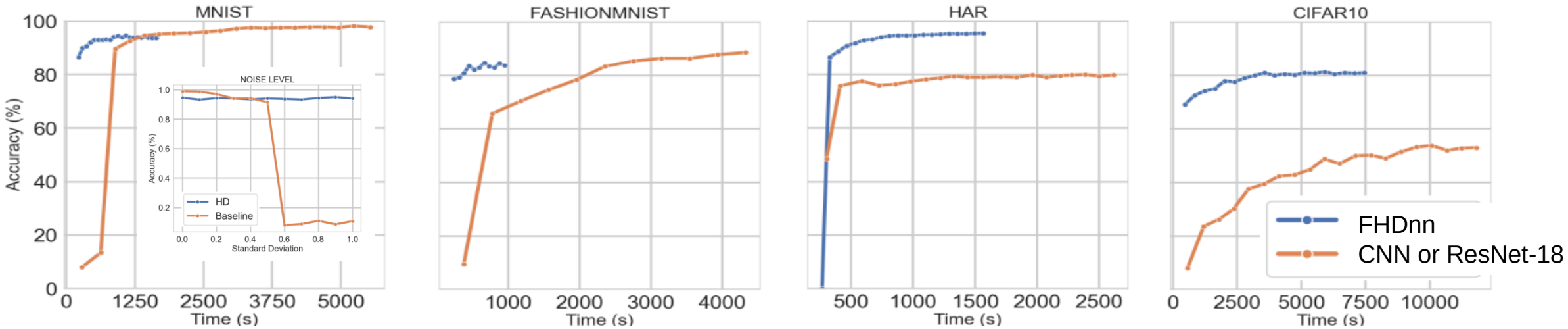


R. Chandrakasaran, X. Yu, W. Xu, T. Rosing, "FHDnn: Communication Efficient And Robust Federated Learning for AIoT Networks," DAC'22

FHDnn in San Diego

- Deployment covered 150 sq miles of San Diego area & used **Raspberry PI**s in local homes
  - All RPIs communicate with **UCSD** as the cloud
  - Latencies vary due to networking & computation
- HD improves FL efficiency in real networks
  - **5x** better communication efficiency
  - **3.2x** faster convergence vs state-of-the-art CNNs
  - FedHD is very robust in unreliable networks



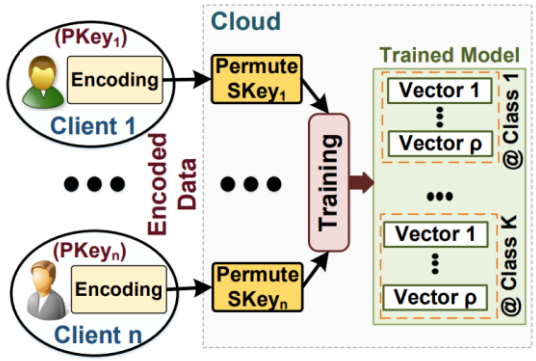


Code is available at https://github.com/QuanlingZhao/FedHD

Q. Zhao, R. Chandrakasaran, X. Yu, W. Xu, T. Rosing, "FedHD - Federated Learning with Hyperdimensional Computing," demo at Mobicom'22

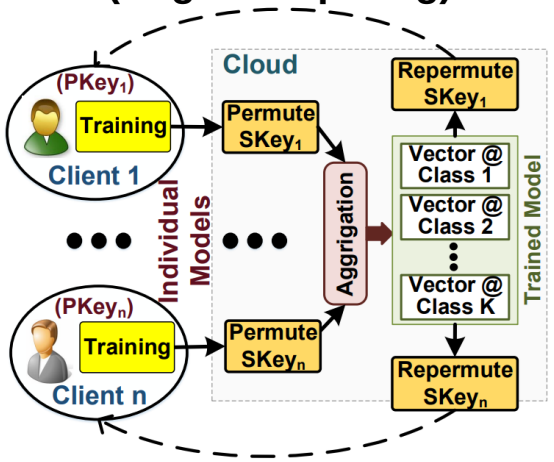
Secure Distributed HD Learning

Centralized Learning (Cloud Computing)



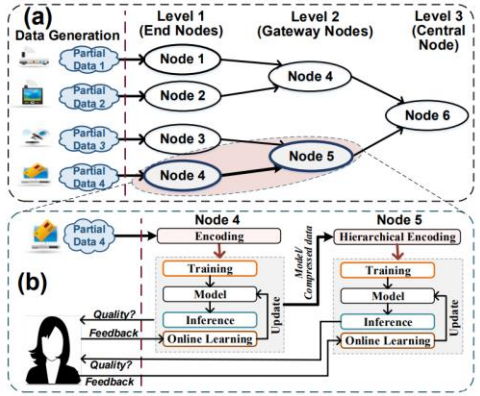
✓Send hypervectors during training
✓High speed learning with large data☺

Federated Learning (Edge Computing)



✓Send pretrained hypervectors for a single data type
✓Drastically lower bandwidth ☺

Hierarchical Learning (Hybrid Computing)



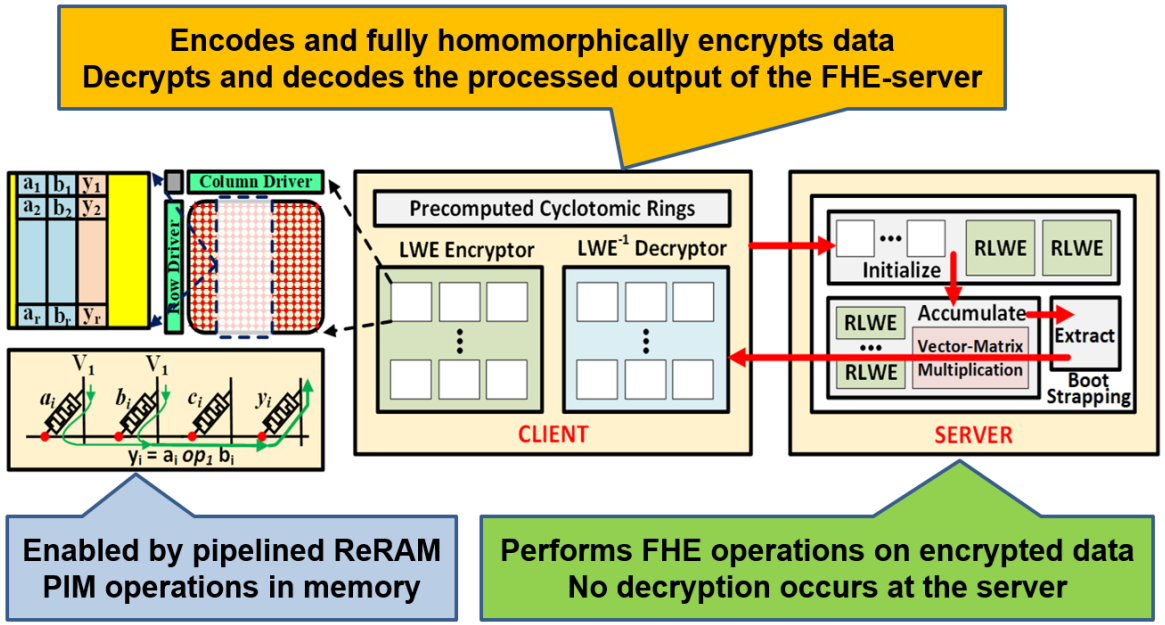
✓Distributed learning using different data types
✓Combine hypervectors to aggregate information

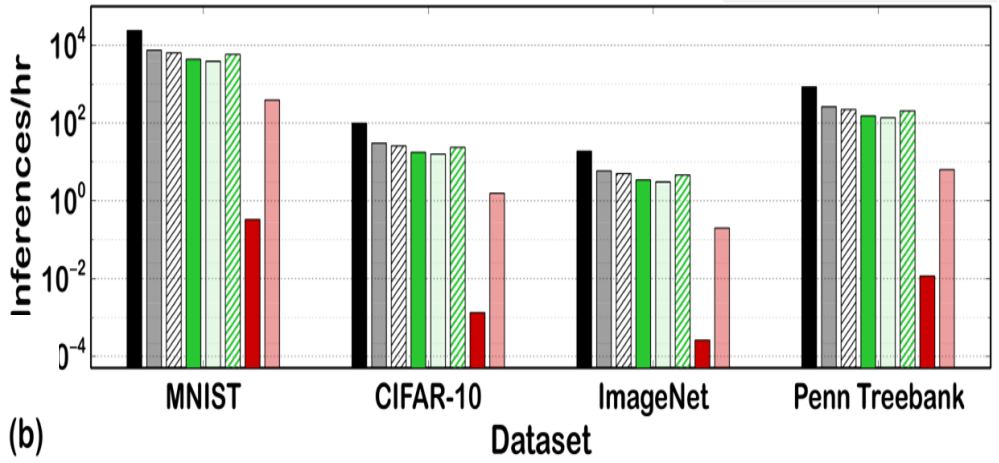
SecureHD **encoding is 146× faster** and **decoding is 6.8× faster** vs homomorphic encryption [Microsoft SEAL]

M. Imani, Y Kim, S Riazi, J Merssely, P Liu, F Koushanfar, T. Rosing, "A Framework for Collaborative Learning in Secure High-Dimensional Space," IEEE Cloud Computing (CLOUD), 2019

FHE computing with HDC in PIM - MemFHE

- Fully homomorphic encryption removes the need for decryption => all processing in encrypted domain
  - Problem: Explosion of data and operations; e.g. int turns into 20kB, int multiply takes >10M ops
- MemFHE design implements 3rd generation fully homomorphic encryption in 1TB ReRAM PIM
- We compare MemFHE with TDNN-FHE (TDNN-Lvl) [NeurIPS'19] with 163-bit (152-bit) classical security that runs on Intel Xeon E7-4850 CPU, 1TB DRAM



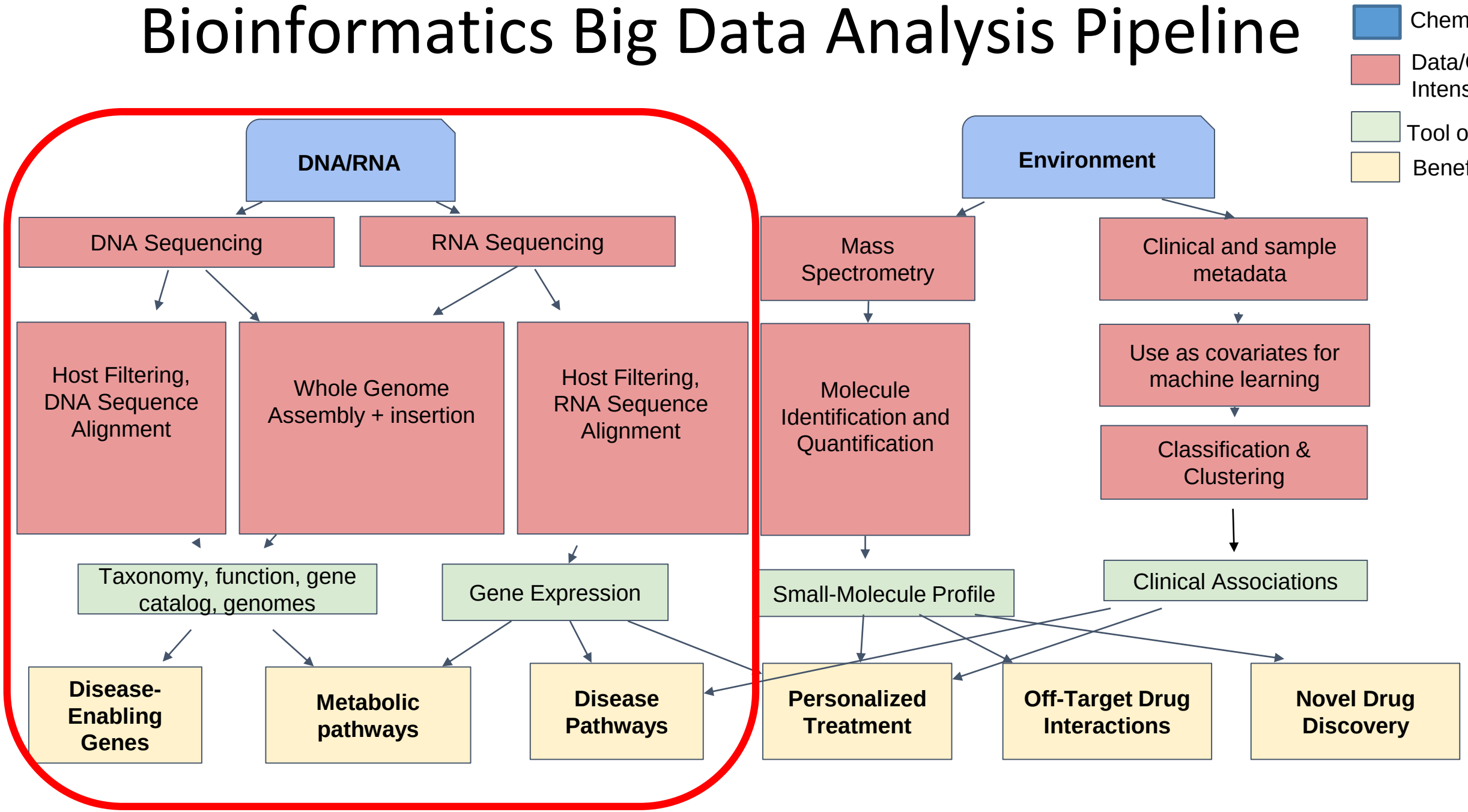


MemFHE DNN vs. TDNN-FHE [NeurIPS'19]
- **Normal mode: 36,007x higher throughput**
- **Quantum-safe mode: 15,000x higher throughput**

MemFHE + HDC is 1,000,000x faster than TDNN-FHE [NeurIPS'19]

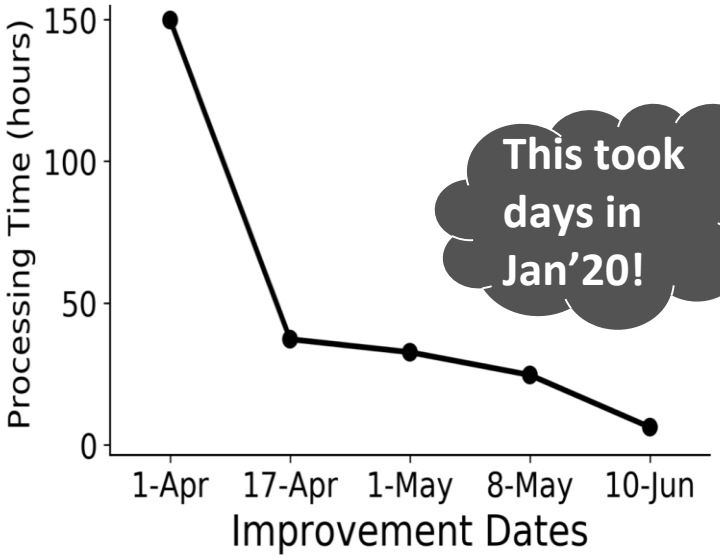
S. Gupta, T. Rosing et al "MemFHE: FHE acceleration in memory," ACM TECS'22.

Bioinformatics Big Data Analysis Pipeline



End-to-End COVID-19 Workflow Optimization

Jointly with Micron, Rob Knight & Niema Moshiri @ UCSD

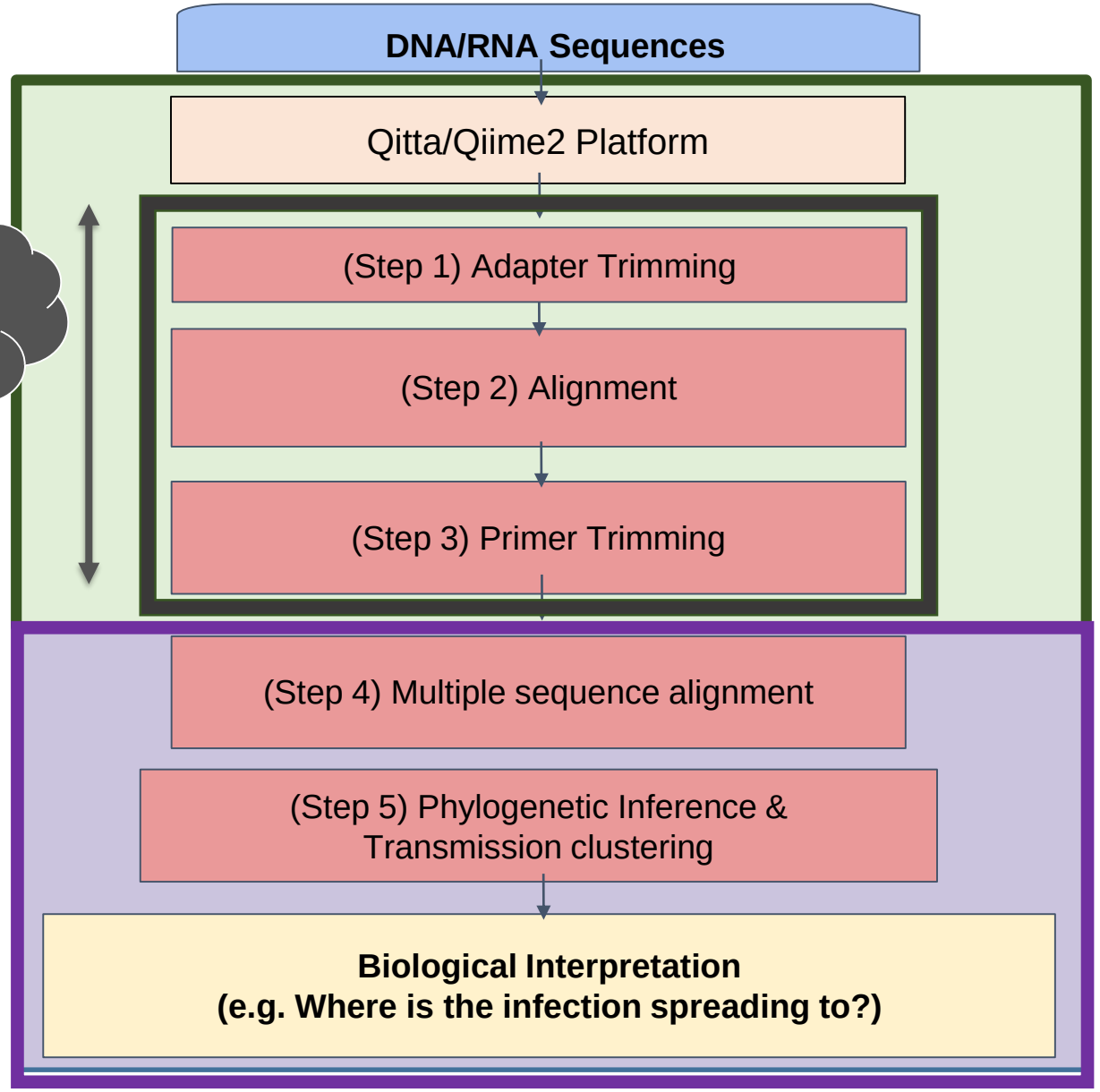


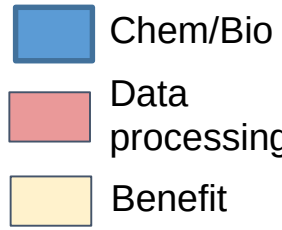
This took days in Jan'20!

CPU Pipeline optimization:

- **START: 1 week to process**
- **17-April** Massive parallelization
- **1-May** Better piping strategy
- **8-May** Threading optimization
- **10-Jun** Tool/Method optimization

Down to hours on CPU => needs further optimization

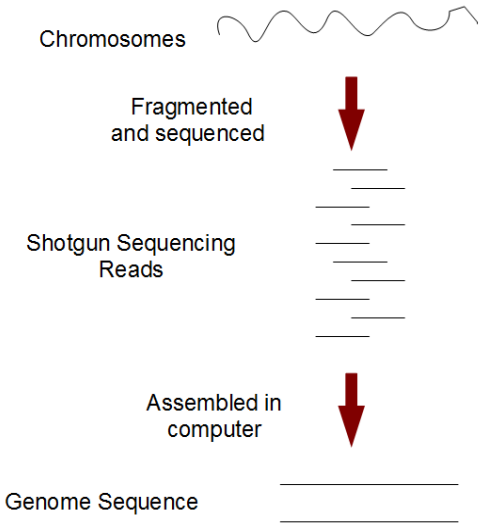




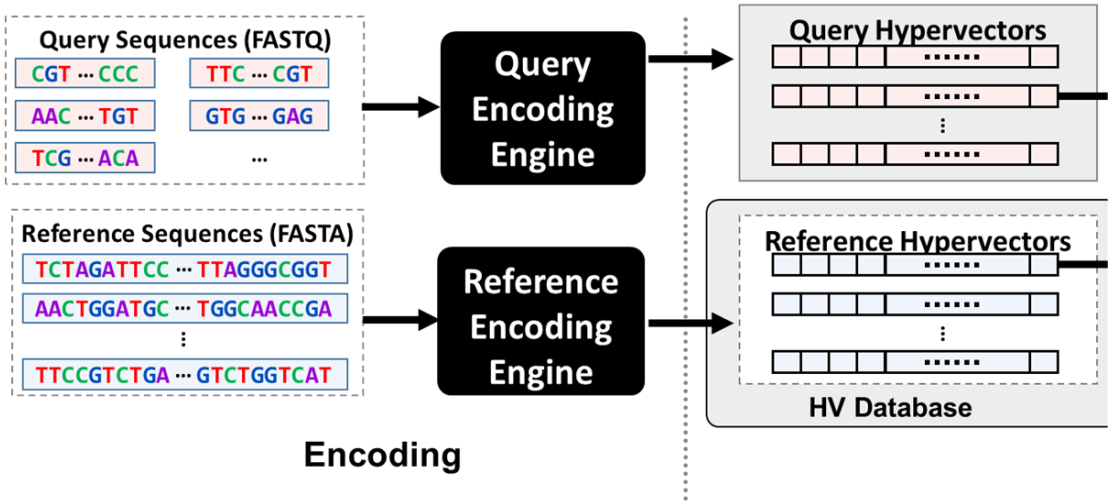
Preprocessing to get aligned sequences; common to all our genomics applications

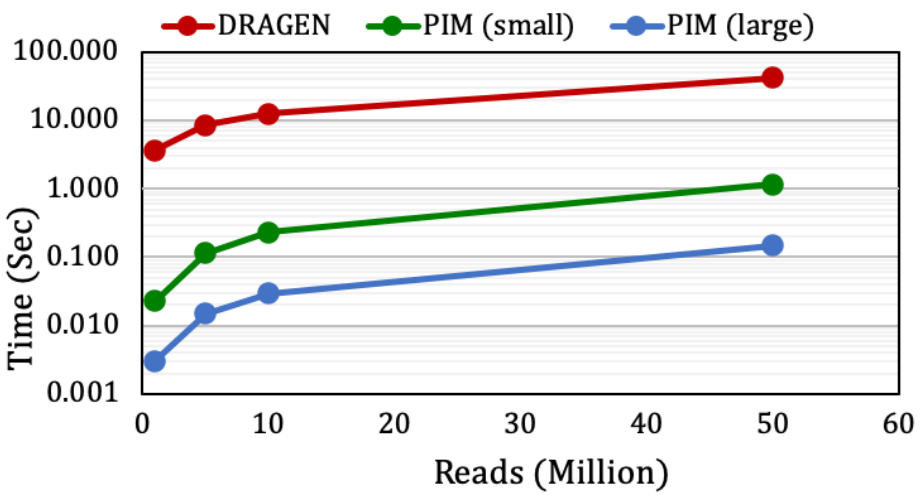
Used by San Diego County for all of COVID-19 analysis!

GenieHD & RAPID Alignment



- Genome Identity Extractor using HyperDimensional Computing
  - Encode DNA into hypervectors
  - Combine ~1,000 segments of the reference DNA into a hypervector
  - Find the existence of DNA patterns using similarity computation
- RAPID – short sequence alignment accelerated in memory
  - Large PIM can fit human genome and is **1,900x faster** vs. Minimap on CPU, 253x faster vs. DRAGEN FPGA





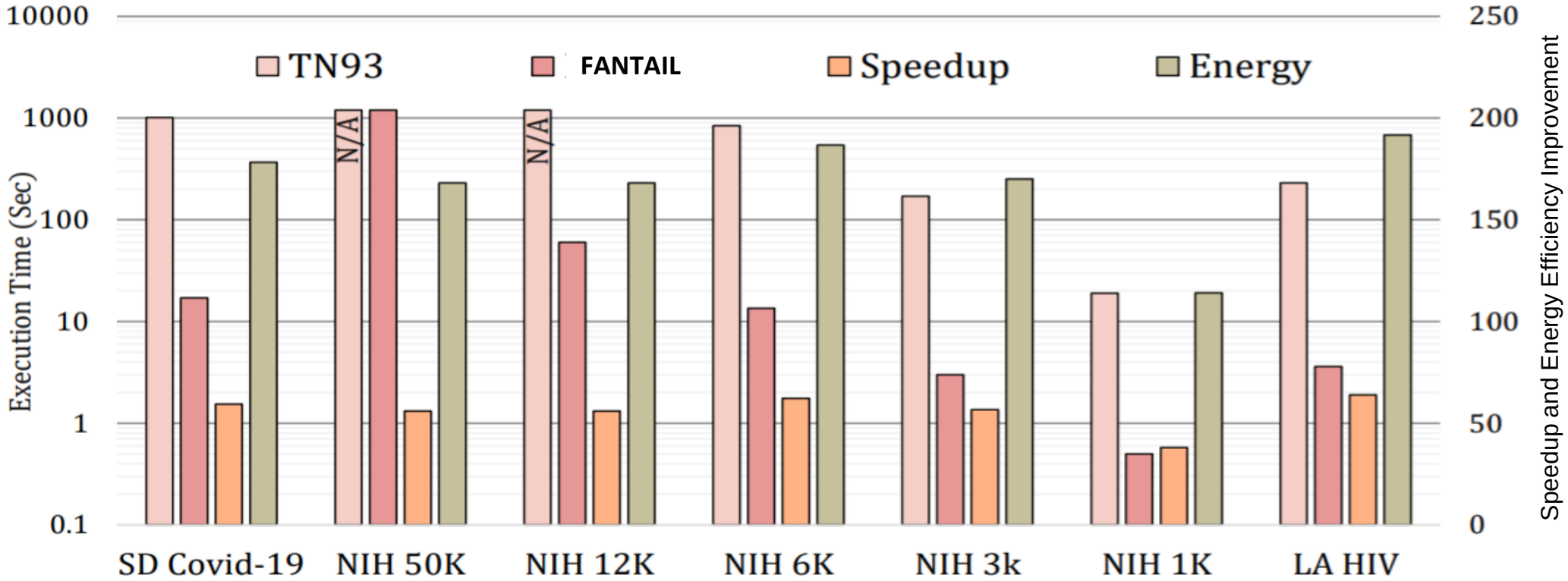
Y. Kim, M. Imani, N. Moshiri, T. Rosing, “GenieHD: Efficient DNA Pattern Matching Accelerator Using HD Computing”, Best Paper at DATE’20

Gupta, S, T. Rosing, et al. "RAPID: A ReRAM processing in-memory architecture for DNA sequence alignment." ISLPED’19.

Xu W, Gupta S, Moshiri N, Rosing T. “RAPIDx: High-performance ReRAM Processing in-Memory Accelerator for Sequence Alignment” tbd TCAD’23

FANTAIL: What mutation is it?

- FANTAIL is a highly parallelized FPGA-based Accelerator for computing pairwise distaNce for viral TrAnsmIssion cLustering based on Tamura-Nei 93 (TN93) model but is generalizable to other nucleotide substitution models
- **Fantail on FPGA takes only 4 seconds** and **consumes 168× less energy vs. TN93** running on an 8-core Intel Core-i7 CPU, while providing the same quality of results



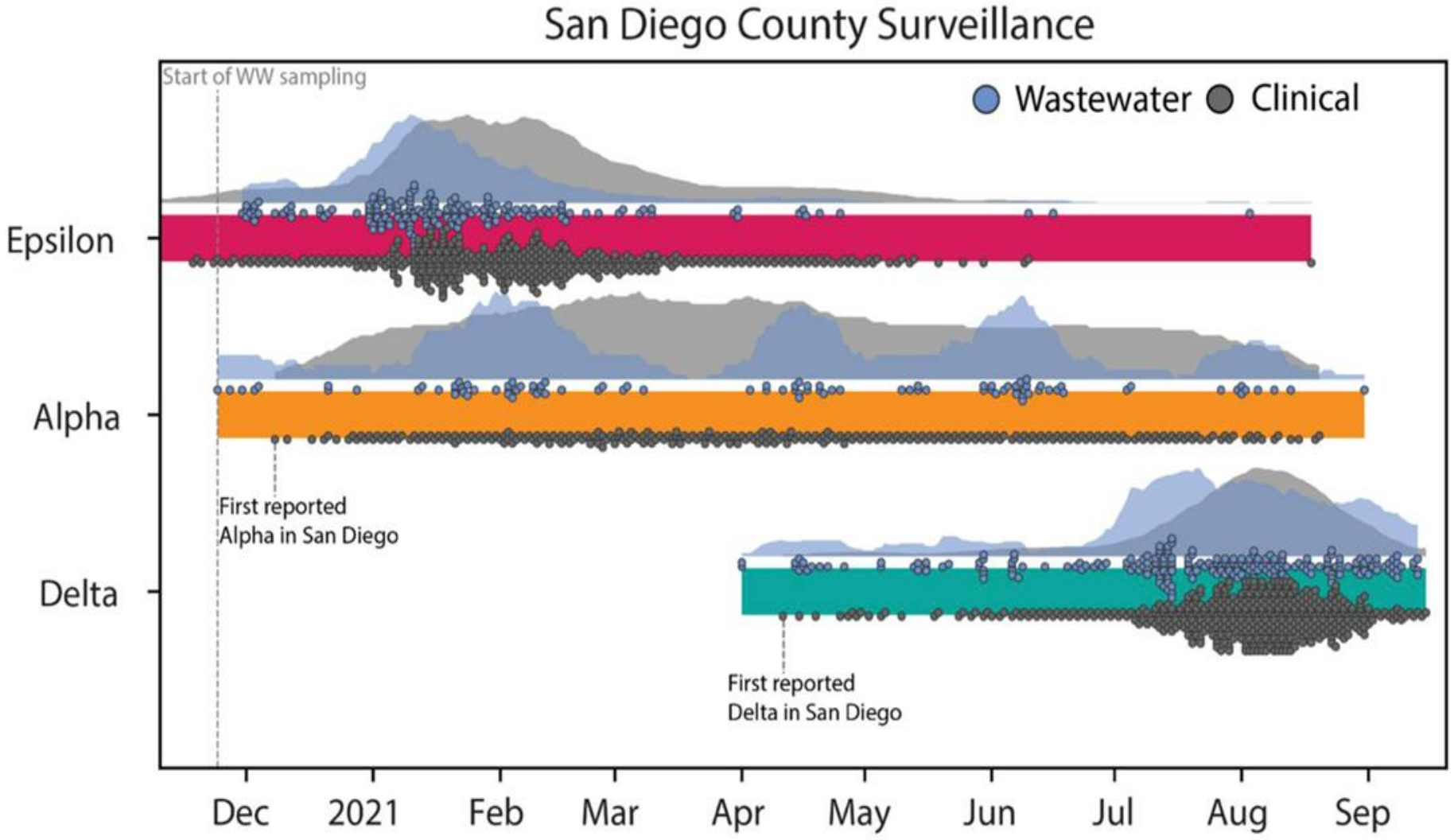
Salamat, Moshiri, Rosing “FANTAIL: Acceleration of TN93 Pairwise Distances” [BioCAS’21]

End-to-End COVID-19 Workflow Optimization

Jointly with Micron, Rob Knight & Niema Moshiri @ UCSD

Our optimized tools are used by UCSD Medical Center for all COVID-19 analysis in San Diego County

Chem/Bio
Data Processing
Benefit



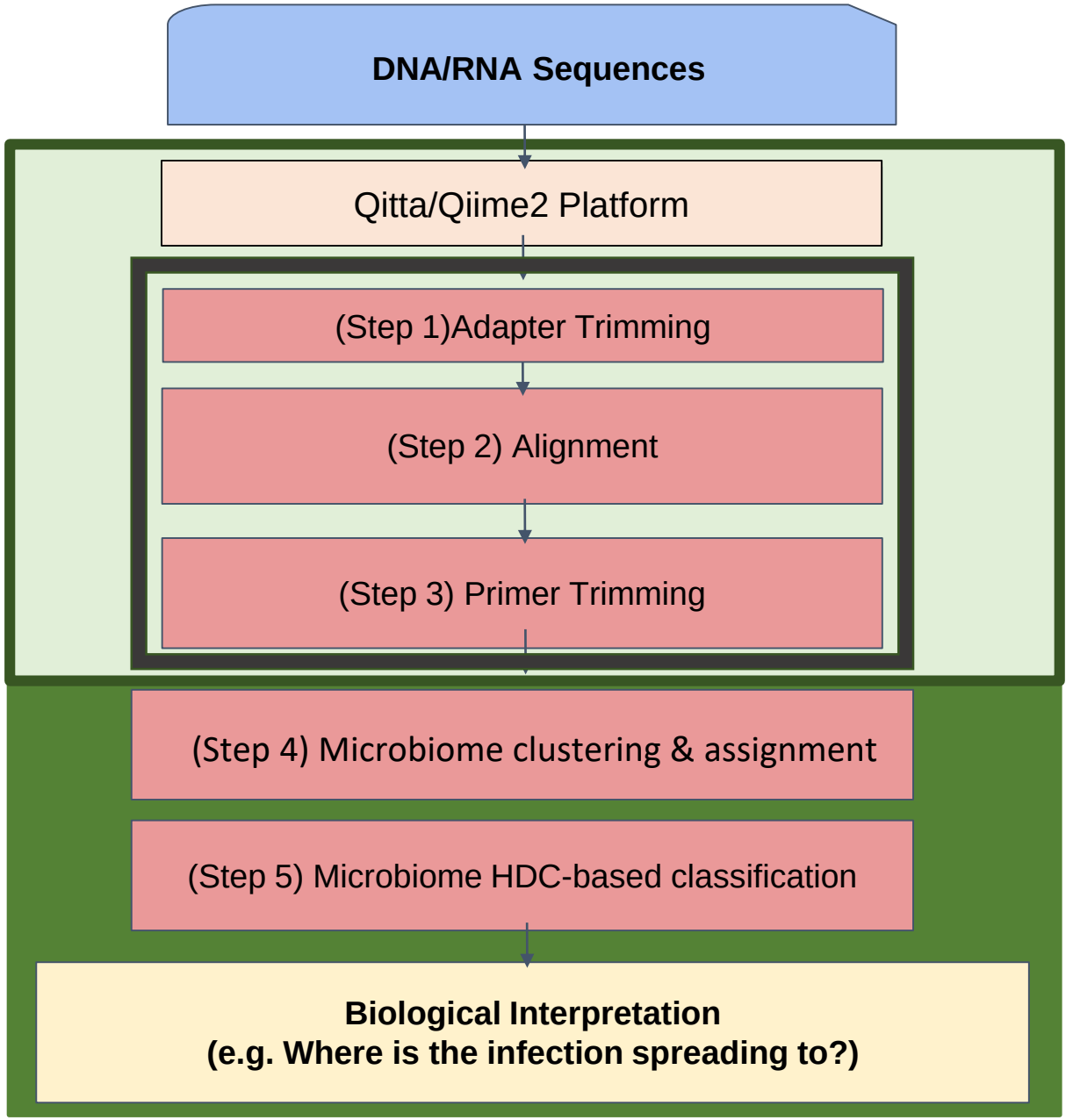
Wastewater samples show earlier appearance of VOCs

Data from 31,149 nasal swab sequences from SD county and 1200 wastewater sequences

Microbiome Analysis Acceleration

Get aligned sequences in seconds via RAPID PIM design

Microbiome classification



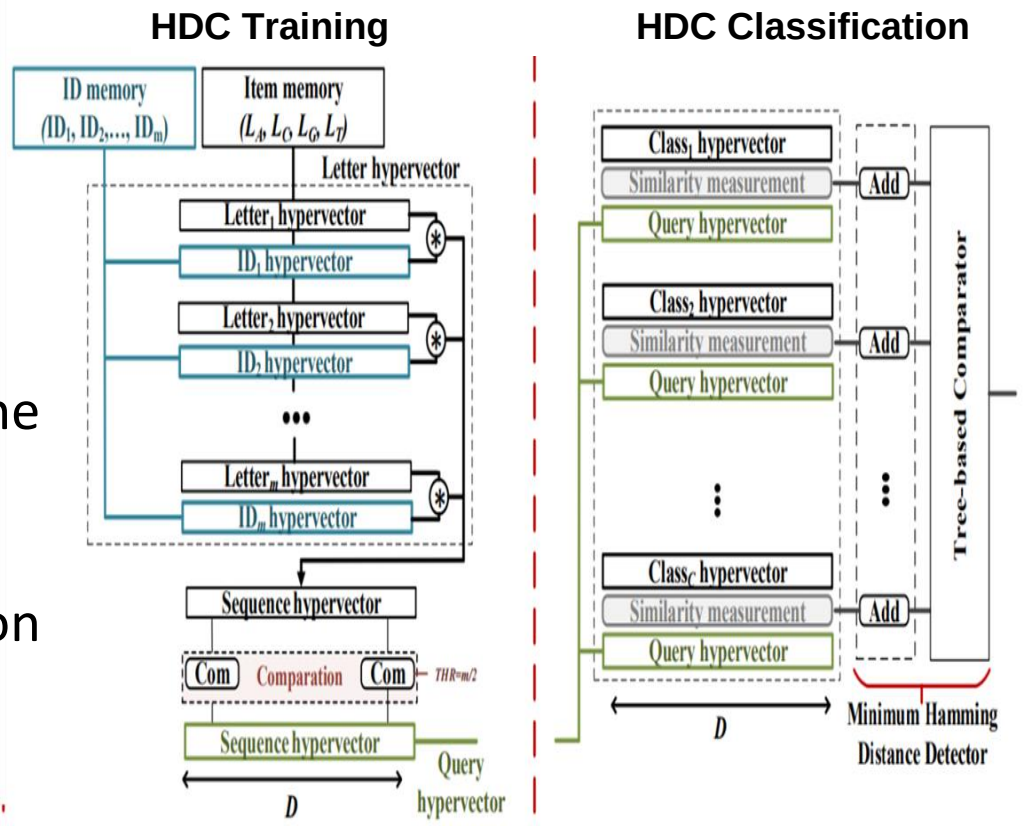
Chem/Bio

Data Processing

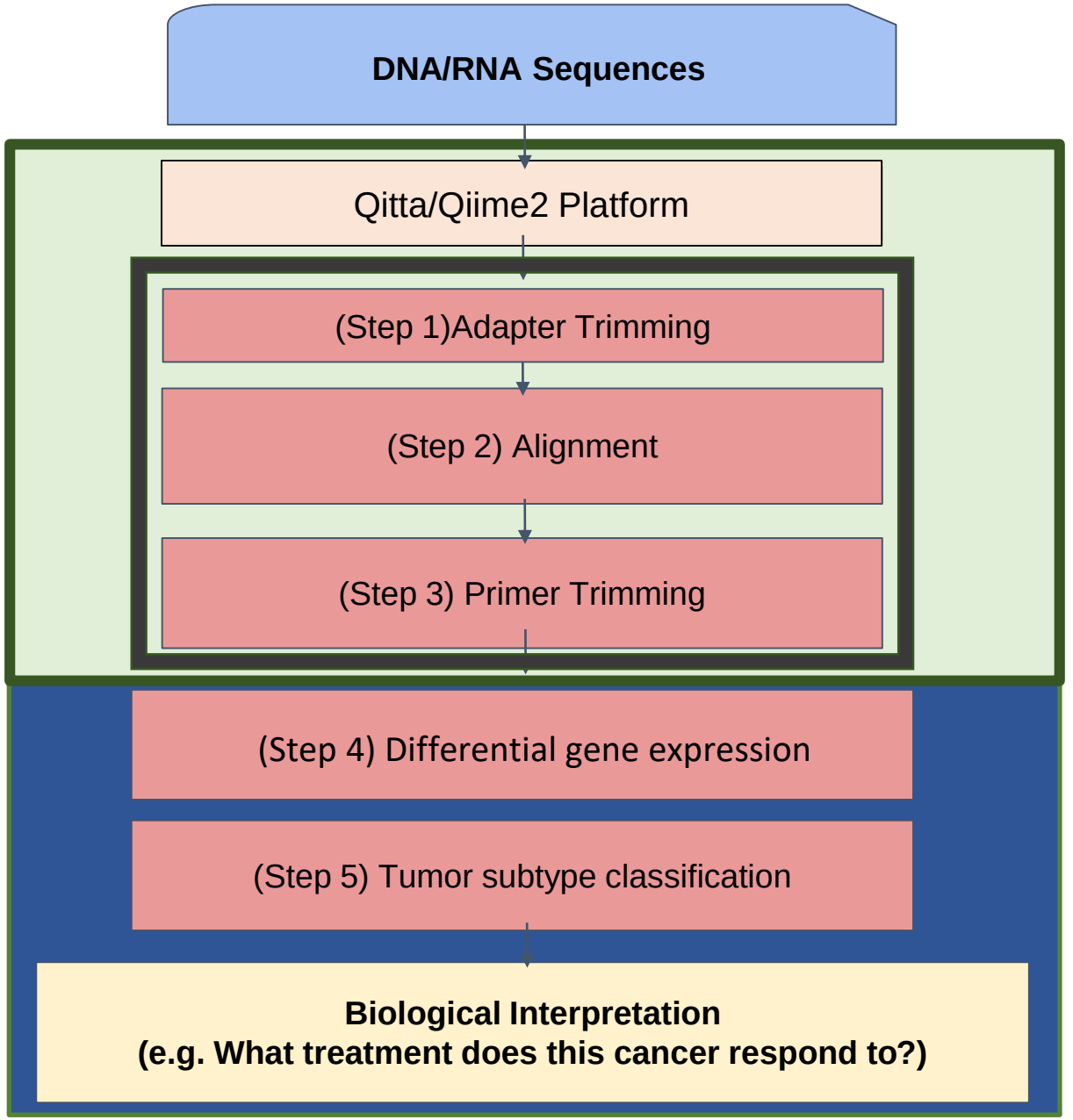
Benefit

Hyperdimensional Microbiome DNA Classification

- Microbiome is key to understanding & treating many diseases, from infections and antibiotic selection to types and success of cancer treatments
- Hyperdimensional (HD) classification of microbiome
  - Training:
    - Hypervectors represent each alphabet of DNA ; classes are created by combining hypervectors
  - Online classification:
    - HD encoded query is compared to all classes & the best is selected
- Results:
  - Comparable accuracy to the state of the art running on CPU - Bayes, SVM or k-NN, MPL and Perceptron
  - HDC PIM is **multiple orders of magnitude faster** than state of the art on CPU



Cancer Variant Acceleration



Get aligned sequences in seconds via RAPID PIM design

Preprocessing to get aligned sequences; common to all our genomics applications

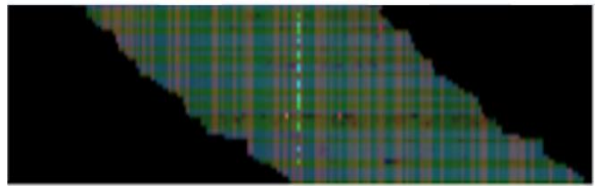
Cancer variant detection

DeepVariant & HDnn Acceleration

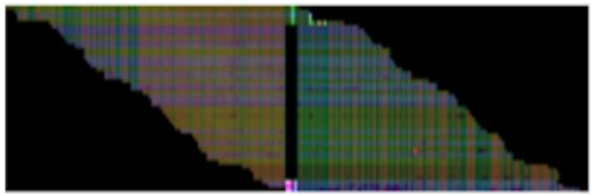
- Variant calling finds changes in genomes
  - e.g. mutations in cancer
- DeepVariant[Nature'18] uses CNNs
  - Converts aligned sequence reads into images & then detects variants
  - HDnn accelerates image classification in memory

Examples of DeepVariant Images

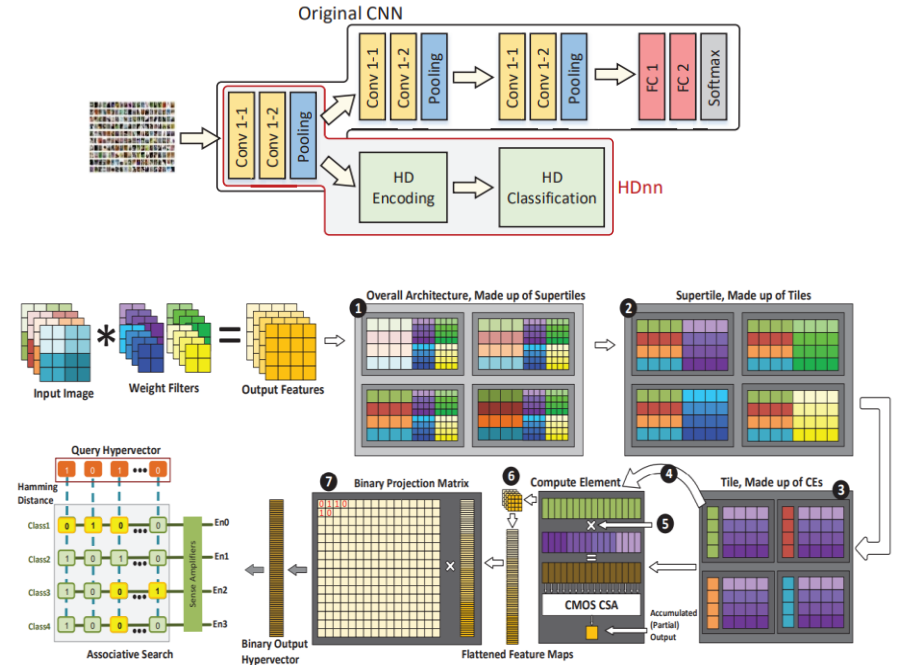
No variants



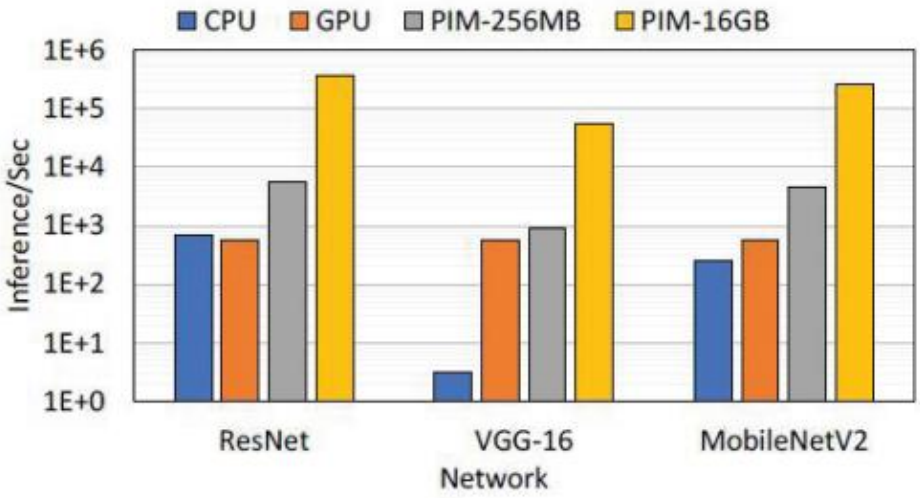
Variants in two chromosomes



HDnn Accelerated in ReRAM

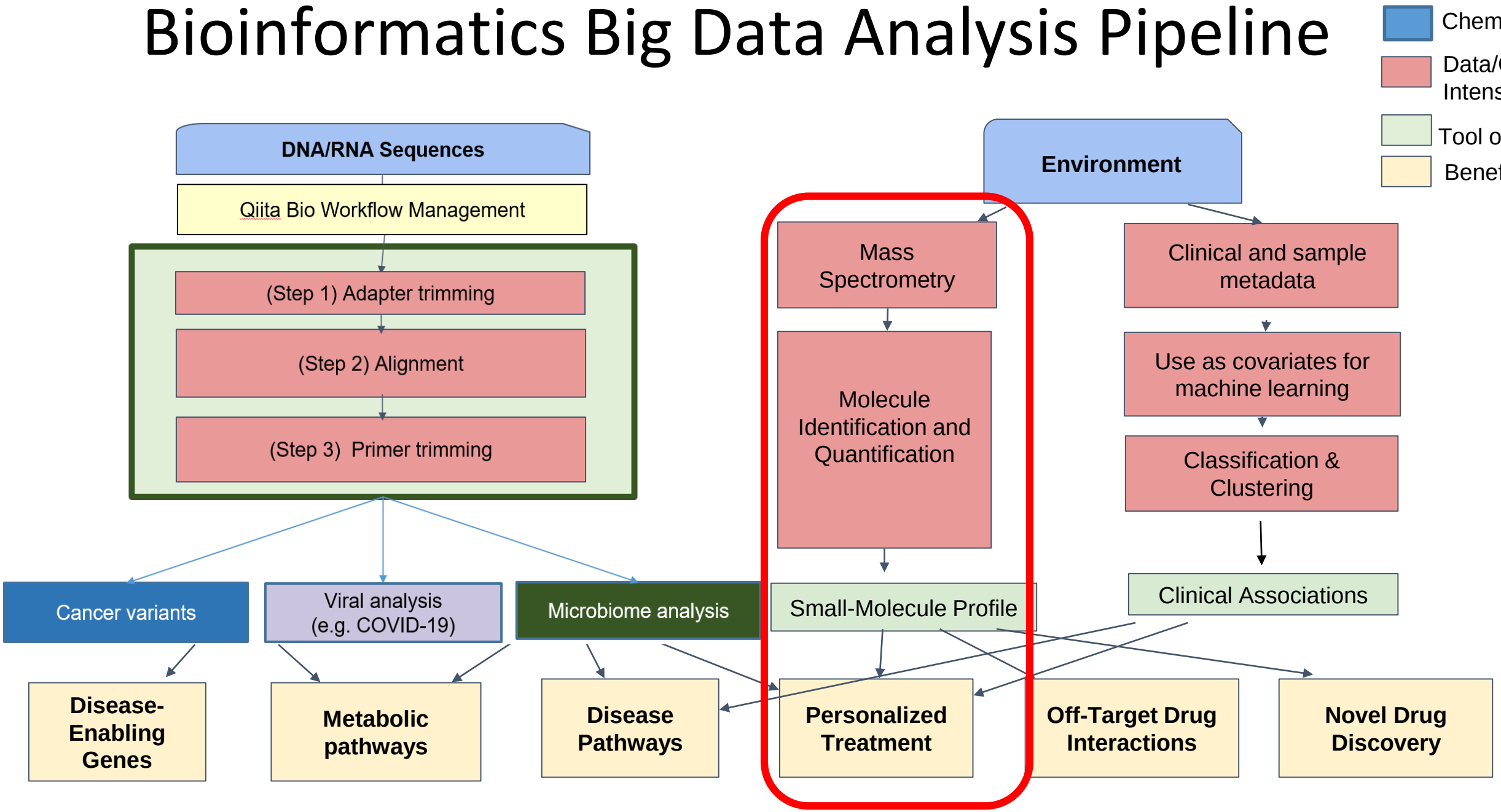


At least **233x** higher throughput vs GPU



A. Dutta, Gupta, S., Rosing, T et al. "HDnn-PIM: Efficient in Memory Design of Hyperdimensional Computing with Feature Extraction," GLVLSI'22
Poplin, Ryan, et al. "A universal SNP and small-indel variant caller using deep neural networks." Nature biotechnology'18 – source of variant images

Bioinformatics Big Data Analysis Pipeline

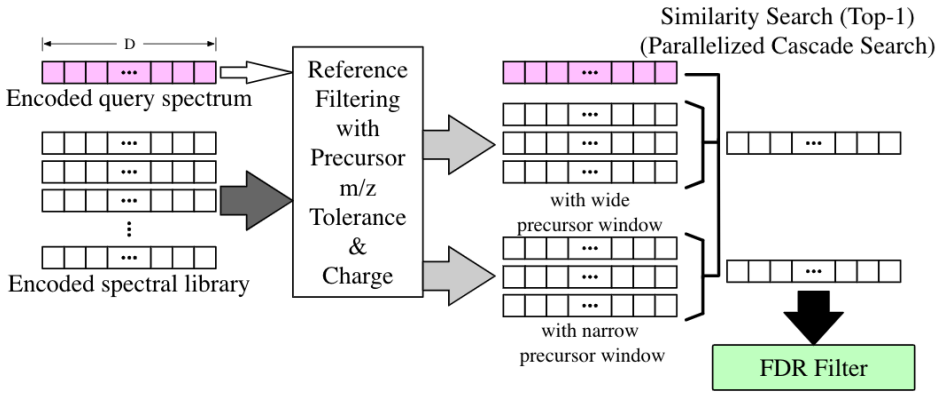


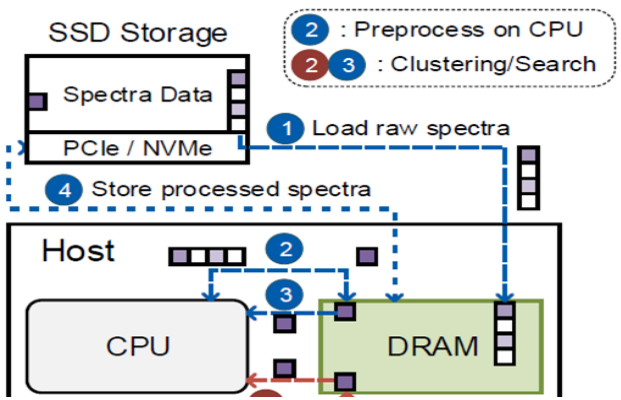
Mass Spectrometry with HD Computing

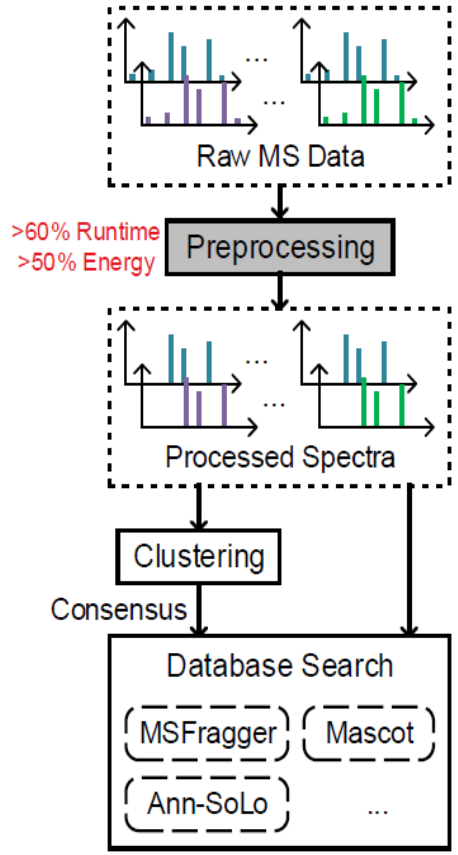
- Mass spectrometry data is growing rapidly
  - UCSD's MassIVE database has >400 TB of protein and peptide spectra
- We use HD computing to accelerate analysis of Mass Spectrometry data, including clustering and classification
  - ID-level hypervector encoding is used for raw spectra
  - Preprocessing was accelerated in storage at the channel level, and is 187x faster than msCRUSH clustering algorithm [JProteome'18]
  - HD-MS open search is 150x faster than AnnSoLo [JProteome'19] on GPU

MassIVE Repository Statistics

Public Datasets:	11,538	Proteins:	20,370
Number of Files:	6,093,453	Peptides:	5,849,004
Total Size:	343.08 TB	Peptide Variants:	15,855,998
Spectra:	3,837,777,086	PSMs:	581,619,799
Dataset Subscriptions:	6,782	Modifications:	1,027

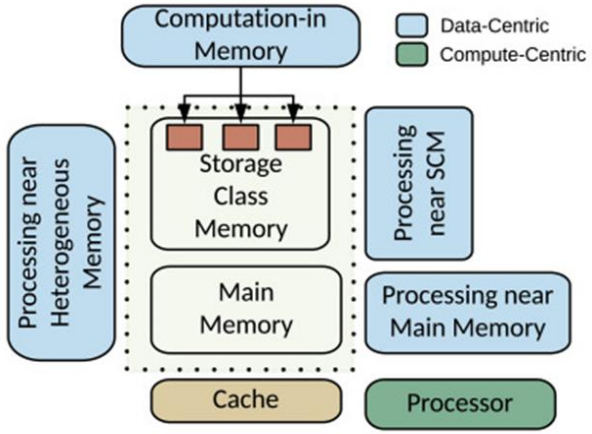




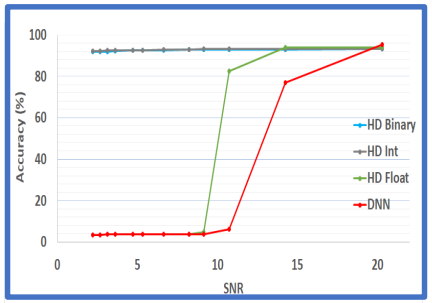


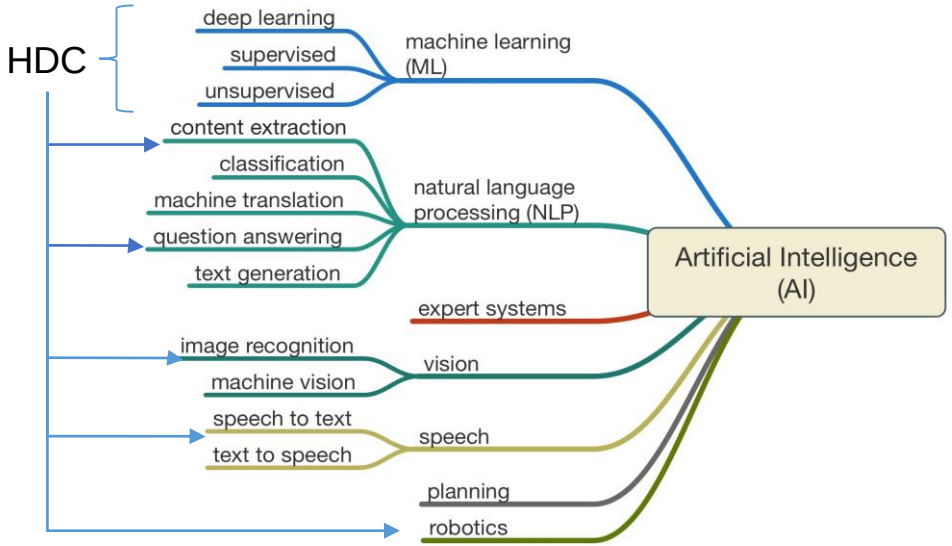
J. Kang, W. Xu, W. Bittremieux, T. Rosing, "Massively Parallel Open Modification Spectral Library Searching with HD Computing," PACT'22.

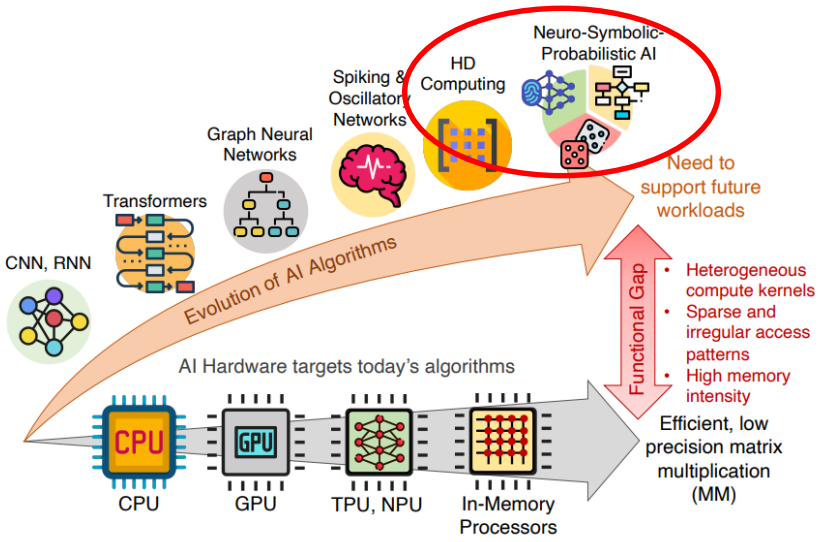
Where to next? Real-Time, Lightweight, Robust & Secure Data Analytics at Scale



- **Vision: Create novel intelligent memory and storage architectures that**
  - Answer when, where and how to store and process which data
  - Seamlessly integrate diversity of memory, storage, compute & software
  - Optimize for best performance, power, area and cost tradeoffs
- **HD Computing (HDC) is a promising solution for future systems**
  - Learns adaptively due to fast training & inference => secure lifelong & federated learning at scale
  - Handles big data => e.g. recommendation systems, mass spectrometry, image processing, ML
  - Inherently robust => excellent for new memory and storage devices & error-prone communication
  - Efficiently combines neuro-symbolic reasoning with probabilistic learning while being explainable







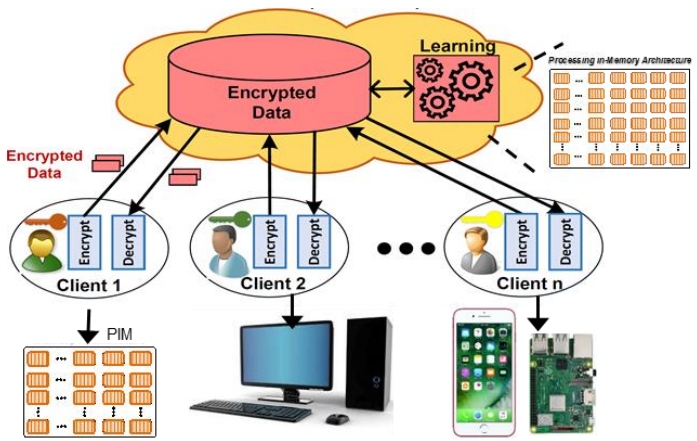


Figure sources: JUMP 2.0 Centers PRISM & CoCoSys

HD Computing publications

JOURNAL PAPERS

1. J. Kang, B. Khaleghi, Y. Kim, T. Rosing, “OpenHD: A GPU-Powered Framework for Hyperdimensional Computing,” Special Issue on Software, Hardware, and Applications for Neuromorphic Computing in IEEE Transactions on Computing, 2022.
2. J. Morris, K. Ergun, B. Khaleghi, M. Imani, B. Aksanli, T. Rosing, “HyDREA:Utilizing Hyperdimensional Computing For A More Robust and Efficient Machine Learning System,” Special issue of ACM TECS’22.
3. S. Gupta, B. Khaleghi, S. Salamt, J. Morris, R. Ramkumar, J. Yu, A. Tiwari, M. Imani, B. Aksanli, T. Rosing, “Store-n-Learn: Classification and Clustering with Hyperdimensional Computing across Flash Hierarchy,” Special issue of ACM TECS, 2022.
4. Justin Morris, Yilun Hao, Saransh Gupta, Behnam Khaleghi, Baris Aksanli and Tajana Rosing, “Stochastic-HD: Leveraging Stochastic Computing on the Hyper-Dimensional Computing Pipeline,” Special Issue of Frontiers in Neuroscience, section on Neuromorphic Engineering, 2022.
5. A. Thomas, S. Dasgupta, T. Rosing,”A Theoretical Perspective on Hyperdimensional Computing,” JAIR, 2021.
6. S. Gupta et al, “COSMO: Computing with Stochastic Numbers in Memory,” JETC, 2021.
7. J. Morris, Y. Hao, M. Imani, B. Aksanli, T. Rosing, “Locality-based Encoder and Model Quantization for Efficient Hyper-Dimensional Computing,” IEEE TCAD 2020.
8. X. Yu, K. Ergun, L. Cherkasova, T. Rosing, “Optimizing Sensor Deployment and Maintenance
9. Costs for Large-Scale Environmental Monitoring,” IEEE TCAD, 2020.
10. Sahand Salamat, Mohsen Imani, and Tajana Rosing, "Accelerating hyperdimensional computing on FPGAs by exploiting computational reuse," IEEE Transactions on Computing, 2020.
11. M. Imani, S. Bosch, S. Datta, S. Ramakrishna, S. Salamat, J. Rabaey, T. Rosing, "QuantHD: A Quantization Framework for Hyperdimensional Computing'', IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2019.
12. M. Imani, S. Gupta, S. Sarama, T. Rosing, "NVQuery: Efficient Query Processing in Non-Volatile Memory," IEEE TCAD, 2019.
13. M. Imani, X. Yin, J, Messerly, S. Gupta, M. Nemier, X. S. Hu, T. Rosing, "SearcHD: A Memory-Centric Hyperdimensional Computing with Stochastic Training", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2019.
14. M. Imani, J. Morris, H. Shu, S. Li, T. Rosing, “Efficient Associative Search in Brain-Inspired Hyperdimensional Computing”, IEEE Design & Test (D&T), 2019.
15. M. Imani, R. Garcia, S. Gupta, T. Rosing, “Hardware-Software Co-design to Accelerate Neural Network Applications”, ACM Journal on Emerging Technologies in Computing (JETC), 2019.
16. Yeseong Kim, Mohsen Imani, and Tajana S. Rosing, "Image Recognition Accelerator Design Using In-Memory Processing," IEEE MICRO, 2018.
17. M. Imani, A. Rahimi, Hwang, T. S. Rosing, J.M. Rabaey, “Low-Power Sparse Hyperdimensional Encoder for Language Recognition,” IEEE Design & Test, 2017.

HD Computing publications to date

CONFERENCE PAPERS

1. J. Kang, M. Zhou, A. Bhansali, W. Xu, A. Thomas and T. Rosing, “RelHD: A Lightweight Graph-based Learning with Hyperdimensional Computing”, ICCD 2022.
2. U. Mallappa, P. Gangwar, B. Khaleghi, and T. Rosing, “TermiNETor: Early Convolution Termination for Efficient Deep Neural Networks”, ICCD 2022
3. A. Thomas, S. Dasgupta, T. Rosing, “A Theoretical Perspective on Hyperdimensional Computing,” **Invited paper to a special session** of the International Joint Conference on Artificial Intelligence, IJCAI 2022
4. Weihong Xu, Jaeyoung Kang and Tajana Rosing, “A Near-Storage Framework for Boosted Data Preprocessing of Mass Spectrum Clustering,” DAC, 2022
5. Behnam Khaleghi, Uday Mallappa, Duygu Nur Yaldiz, Haichao Yang, Monil Shah, Jaeyoung Kang and Tajana Rosing, “PatterNet: Explore and Exploit Filter Patterns for Efficient Deep Neural Networks,” DAC’22
6. Rishikanth Chandrasekaran, Kazim Ergun, Jihyun (Lucy) Lee, Dhanush Nanjunda, Jaeyoung Kang and Tajana Rosing, “FHDnn: Communication Efficient and Robust Federated Learning for AIoT Networks,” DAC’22.
7. Behnam Khaleghi, Jaeyoung Kang, Hanyang Xu, Justin Morris and Tajana Rosing, “GENERIC: Highly Efficient Learning Engine on Edge using Hyperdimensional Computing,” DAC’22.
8. Yang Ni, Yeseong Kim, Tajana Rosing and Mohsen Imani,” Algorithm-Hardware Co-Design for Efficient Brain-Inspired Hyperdimensional Learning on Edge,” **Best Paper Award** at DATE’22.
9. Justin Morris, Hin Wai Lui, Kenneth Stewart, Behnam Khaleghi, Anthony Thomas, Thiago Marback, Baris Aksanli, Emre Neftci, and Tajana Rosing, “HyperSpike: HyperDimensional Computing for More Efficient and Robust Spiking Neural Networks, DATE’22.
10. J. Kang, B. Khaleghi, Y. Kim, T. Rosing, “XCelHD: Efficient GPU-Powered Hyperdimensional Computing with Parallelized Training,” ASPDAC’22
11. Q. Zhao, K. Lee, J. Liu, M. Huzaifa, X. Yu, T. Rosing, “FedHD - Federated Learning with Hyperdimensional Computing Demo”, Mobicom, 2022
12. Arpan Dutta, Saransh Gupta, Behnam Khaleghi, Rishikanth Chandrasekaran, Weihong Xu, Tajana Rosing,”HDnn PIM: Efficient in Memory Design of Hyperdimensional Computing with Feature Extraction,” GLVLSI’22.

HD Computing publications to date

CONFERENCE PAPERS

1. Yeseong Kim, M. Imani, S. Gupta, M. Zhou, T. Rosing, “CHOIR: Massively Parallel Big Data Classification on a Programmable Processing In-Memory Architecture,” ICCAD 21.
2. Justin Morris, Si Thu Kaung Set, Gadi Rosen, Mohsen Imani, Baris Aksanli and Tajana Rosing,” AdaptBit-HD: Adaptive Model Bitwidth for Hyperdimensional Computing,” ICCD’21
3. Yilun Hao, Saransh Gupta, Justin Morris, Behnam Khaleghi, Baris Aksanli and Tajana Rosing:”Stochastic-HD: Leveraging Stochastic Computing on Hyper-Dimensional Computing,” ICCD’21.
4. Mohsen Imani*, Zhuowen Zou, Samuel Bosch, Sanjay Anantha Rao, Sahand Salamat, Venkatesh Kumar, Yeseong Kim*, and Tajana Rosing, "Revisiting HyperDimensional Learning for FPGA and Low-Power Architectures," HPCA’21.
5. Namiko Matsumoto, Anthony Thomas, Tara Javidi and Tajana Rosing, “Hyperdimensional Computing and Spectral Learning,” CogArch’21.
6. Behnam K., Hanyang Xu, Justin Morris, Tajana S. Rosing, “tiny-HD: Ultra-Efficient Hyperdimensional Computing Engine for IoT Applications,” DATE’21.
7. Justin Morris, Kazim Ergun, Behnam Khaleghi, Mohsen Imani, Baris Aksanli, Tajana Rosing, “HyDREA: Towards More Robust and Efficient Machine Learning Systems with Hyperdimensional Computing,” DATE’21.
8. M. Zhou, S. Gupta, M. Imani, Y. Kim, T. Rosing,"DP-Sim:A Full-stack Simulation Infrastructure for Digital Processing In-Memory Architecture," ASPDAC 2021.
9. Y. Guo, S. Gupta, M. Imani, Y. Kim, J. Morris, T. Rosing, "HyperRec: Efficient Recommender Systems with Hyperdimensional Computing," ASPDAC 2021.
10. M. Imani, S. Pampana, S. Gupta, M. Zhou, Y. Kim, T. Rosing, “DUAL: Acceleration of Clustering Algorithms using Digital-based Processing In-Memory,” MICRO 2020.
11. J. Morris, Y. Hao, S. Gupta, R. Ramkumar, J. Yu, M. Imani, B. Aksanli, T. Rosing, “Multi-label HD Classification in 3D Flash,” VLSI-SOC 2020, **Invited paper**.

HD Computing publications to date

CONFERENCE PAPERS

1. B. Khaleghi, S. Salamat, A. Thomas, F. Asgarinejad, Y. Kim, T. Rosing, "SHEARer: Highly-Efficient Hyperdimensional Computing by Software-Hardware Enabled Multifold AppRoximation", ISLPED, 2020.
2. Saransh Gupta, Justin Morris, Mohsen Imani, Ranganathan Ramkumar, Jeffrey Yu, Aniket Tiwari, Baris Aksanli, Tajana Rosing , "THRIFTY: Training with Hyperdimensional Computing across Flash Hierarchy," ICCAD 2020.
3. M. Imani, M. Samragh, Y. Kim, S. Gupta, F. Koushanfar, T. Rosing “Deep Learning Acceleration with Neuron-to-Memory Transformation”, HPCA, 2020.
4. B. Khaleghi, M. Imani, T. Rosing “Prive-HD: Privacy-Preserved Hyperdimensional Computing”, DAC, 2020.
5. S. Gupta, M. Imani, J. Sim, A. Huang, F. Wu, H. Najafi, T. Rosing, “SCRIMP: A General Stochastic Computing Architecture using ReRAM in-Memory Processing”, DATE, 2020.
6. Y. Kim, M. Imani, N. Moshiri*, T. Rosing, “GenieHD: Efficient DNA Pattern Matching Accelerator Using Hyperdimensional Computing”, **best paper nomination**, DATE, 2020.
7. J.Morris, M. Imani, S. Bosch, A. Thomas, H. Shu, T. Rosing, “CompHD: Efficient Hyperdimensional Computing Using Model Compression”. IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2019.
8. M. Imani, J. Morris, S. Bosch, H. Shu, G. De Micheli, T. S. Rosing, “AdaptHD: Adaptive Efficient Training for Brain-Inspired Hyperdimensional Computing,” BioCAS, 2019.
9. Mohsen Imani, Yeseong Kim, Sadegh Riyazi, John Merssely, Patrick Liu, Farinaz Koushanfar, and Tajana S. Rosing, "A Framework for Collaborative Learning in Secure High-Dimensional Space", IEEE Cloud Computing (CLOUD), 2019
10. Mohsen Imani, Yeseong Kim, Thomas Worley, Sarangh Gupta, and Tajana S. Rosing, "HDCluster: An Accurate Clustering Using Brain-Inspired High-Dimensional Computing", IEEE/ACM Design Automation and Test in Europe Conference (DATE), 2019
11. Mohsen Imani, John Merssely, Fan Wu, Wang Pi, and Tajana S. Rosing, "A Binary Learning Framework for Hyperdimensional Computing", IEEE/ACM Design Automation and Test in Europe Conference (DATE), 2019

HD Computing publications to date

CONFERENCE PAPERS

1. M. Imani, S. Gupta, T. Rosing “Digital-based Processing In-Memory: A Highly-Parallel Accelerator for Data Intensive Applications”, ACM International Symposium on Memory Systems (MEMSYS), 2019.
2. Mohsen Imani, Tarek Nassar, Tajana Rosing, "Moving Toward Real-Time Diagnostics using Brain-Inspired Hyperdimensional Computing", AACR conference on Artificial Intelligence, Big Data, and Prediction in Cancer
3. Mohsen Imani, Tarek Nassar, Tajana Rosing, "Brain-Inspired Hyperdimensional Computing for Real-Time Health Analysis", IEEE International Conference on Biomedical and Health Informatics (BHI), 2019
4. Mohsen Imani, Saransh Gupta, Yeseong Kim, Minxuan Zhou, and Tajana S. Rosing, "DigitalPIM: Digital-based Processing In-Memory for Big Data Acceleration", ACM Great lakes symposium on VLSI (GLSVLSI), 2019
5. Saransh Gupta, Mohsen Imani, and Tajana Rosing, "Exploring Processing In-Memory for Different Technologies", ACM Great lakes symposium on VLSI (GLSVLSI), 2019
6. Mohsen Imani, Justin Morris, John Merssely, Helen Shu, Yaobang Deng, and Tajana S. Rosing, "BRIC: Locality-based Encoding for Energy-Efficient Brain-Inspired Hyperdimensional Computing", IEEE/ACM Design Automation Conference (DAC), 2019. **Best paper nomination**
7. Mohsen Imani, Sahand Salamat, Saransh Gupta, Jiani Huang, and Tajana S. Rosing, "FACH: FPGA-based Acceleration of Hyperdimensional Computing by Reducing Computational Complexity", IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), 2019
8. Mohsen Imani, Sahand Salamat, Behnam Khaleghi, Mohammad Samragh, Farinaz Koushanfar, and Tajana S. Rosing, "SparseHD: Algorithm-Hardware Co-Optimization for Efficient High-Dimensional Computing", International Symposium on Field-Programmable Custom Computing Machines (FCCM), 2019
9. Sahand Salamat, Mohsen Imani, Behnam Khaleghi, and Tajana S. Rosing, "F5-HD: Fast Flexible FPGA-based Framework for Refreshing Hyperdimensional Computing", ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), 2019

HD Computing publications to date

CONFERENCE PAPERS

1. Joonseop Sim, Saransh Gupta, Mohsen Imani, Yeseong Kim, and Tajana S. Rosing, "UPIM : Unipolar Switching Logic for High Density Processing-in-Memory Applications", ACM Great lakes symposium on VLSI (GLSVLSI), 2019
2. Joonseop Sim, Minsu Kim, Yeseong Kim, Saransh Gupta, Behnam Khaleghi and Tajana Rosing, "MAPIM: Mat Parallelism for High Performance Processing in Non-volatile Memory Architecture", IEEE International Symposium on Quality Electronic Design (ISQED), 2019
3. Mohsen Imani, Sahand Salamat, Jiani Huang, Saransh Gupta, Tajana Rosing, "FACH: FPGA-based Acceleration of Hyperdimensional Computing by Reducing Computational Complexity", IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), 2019
4. Joonseop Sim, Mohsen Imani, Woojin Choi, Yeseong Kim, Tajana Rosing, "LUPIS: Latch-Up Based Ultra Efficient Processing-in-Memory System", **Best paper nomination**, ISQED'18.
5. M. Imani, Y. Kim, T. Rosing, "Visual Object Recognition Accelerator Based on Approximate In-Memory Processing", Non-Volatile Memory Workshop (NVMW), 2018.
6. M. Imani, A. Rahimi, D. Kong, T. Rosing, Jan Rabaey, "Non-Volatile Associative Memory to Accelerate Brain-inspired Hyperdimensional Computing" Non-Volatile Memory Workshop (NVMW), 2018 .
7. M. Imani, C. Huang , D. Kong, T. Rosing, "Hierarchical Hyperdimensional Computing for Energy Efficient Classification", IEEE/ACM Design Automation Conference (DAC), 2018.
8. S. Gupta, M. Imani, T. Rosing, "Processing In-Memory Architecture for Multiple Memory Technology", IEEE/ACM Design Automation Conference (DAC), 2018.
9. M. Imani, S. Gupta, T. Rosing, "GenPIM: Generalized Processing In-Memory to Accelerate Data Intensive Applications ", IEEE/ACM Design Automation and Test in Europe Conference (DATE), 2018.
10. S. Gupta, M. Imani, T. Rosing, "FELIX: Fast and Energy-Efficient Logic in Memory" IEEE/ACM International Conference On Computer Aided Design (ICCAD), 2018.
11. M. Imani, Y. Kim, T. Rosing, "Brain-Inspired Hyperdimensional Computing: An Efficient Classifier for Embedded Devices" IEEE/ACM International Conference On Computer Aided Design (ICCAD), 2018.
12. Y. Kim, M. Imani, T. Rosing, "Efficient Human Activity Recognition Using Hyperdimensional Computing", IEEE Conference on Internet of Things (IoT), 2018.
13. M. Imani, T. Nassar, T. Rosing, "Moving Toward Real-Time Diagnostics using Brain-Inspired Hyperdimensional Computing", AACR conference on Artificial Intelligence, Big Data, and Prediction in Cancer, 2018.
14. M. Imani, A. Rahimi, D. Kong, T. Rosing, J. M. Rabaey "Exploring Hyperdimensional Associative Memory", HPCA'17.
15. M. Imani, S. Gupta, T. Rosing "Ultra-Efficient Processing In-Memory for Data Intensive Applications", DAC'17.
16. Mohsen Imani, Yeseong Kim, and Tajana S. Rosing, "Brain-Inspired Hyperdimensional Computing: An Efficient Classifier for Embedded Devices," ICCAD'17.
17. M. Imani, A. Rahimi, D. Kong, T. Rosing, J. M. Rabaey "Hardware Acceleration of Brain-inspired Hyperdimensional Computing", ICCAD VMC 2017.

Bio Related Publications

- B. Khaleghi, T. Zhang, C. Martino, G. Armstrong, Ameen Akel, Ken Curewitz, Justin Eno, Sean Eilert, Rob Knight, N. Moshiri, Tajana Rosing, “SALIENT: Ultra-Fast FPGA-based Short Read Alignment,” FPT’22
- J. Kang, W. Xu, W. Bittremieux, T. Rosing, “Massively Parallel Open Modification Spectral Library Searching with HD Computing,” PACT’22.
- Khaleghi B, Zhang T, Shao N, Akel A, Curewitz K, Eno J, Eilert S, Moshiri N, Rosing T (2022). "FAST: FPGA-based Acceleration of Genomic Sequence Trimming." BioCAS 2022.
- Weihong Xu, Jaeyoung Kang and Tajana Rosing, “A Near-Storage Framework for Boosted Data Preprocessing of Mass Spectrum Clustering,” DAC, 2022
- M. Zhou, Y. Guo, W. Xu, B. Li, K. Eliceiri, T. Rosing, “MAT: Processing In-Memory Acceleration for Long-Sequence Attention,” DAC’21.
- Armstrong G, Martino C, Morris J, Khaleghi B, Kang J, Dereus J, Zhu Q, Roush D, McDonald D, Gonzalez A, Shaffer J, Carpenter C, Estaki M, Wandro S, Eilert S, Akel A, Eno J, Curewitz K, Swafford A, Moshiri N, Rosing T, Knight R (2022). "Swapping Metagenomics Preprocessing Pipeline Components Offers Speed and Sensitivity Increases." mSystems’21
- Salamat S, Moshiri N, Rosing T (2021). "FPGA Acceleration of Pairwise Distance Calculation for Viral Transmission Clustering." BioCAS 2021.
- Zhou M, Wu L, Li M, Moshiri N, Skadron K, Rosing T (2021). "Ultra Efficient Acceleration for De Novo Genome Assembly via Near-Memory Computing." PACT’21
- Salamat S, Kang J, Kim Y, Imani M, Moshiri N, Rosing T "FPGA Acceleration of Protein Back-Translation and Alignment." DATE’21
- Salamat S, Moshiri N, Rosing T "FPGA Acceleration of Pairwise Distance Calculation for Viral Transmission Clustering." 28th International Dynamics & Evolution of Human Viruses
- Kang J, Young C, Morris J, Akel A, Eilert S, Eno J, Curewitz K, Moshiri N, Rosing T (2021). "A GPU-Powered Phylogenetic Analysis for Large-scale Genomic Sequences." 28th International Dynamics & Evolution of Human Viruses
- Khaleghi B, Akel A, Curewitz K, Eno J, Eilert S, Moshiri N, Rosing T "FPGA-based acceleration of primer trimming." Intl. Dynamics & Evolution of Human Viruses’21.
- Sfiligoi, Igor, Daniel McDonald, and Rob Knight. "Porting and optimizing UniFrac for GPUs: Reducing microbiome analysis runtimes by orders of magnitude." Practice and Experience in Advanced Research Computing. 2020
- Y. Kim, M. Imani, N. Moshiri*, T. Rosing, “GenieHD: Efficient DNA Pattern Matching Accelerator Using HD Computing”, Best paper, DATE’20.
- Gupta S, Imani M, Khaleghi B, Moshiri N, Rosing T (2020). "RAPIDx: A ReRAM Processing in-Memory Architecture for DNA Short Read Alignment." ASHG’20
- M. Imani, J. Morris, S. Bosch, H. Shu, G. De Micheli, T. S. Rosing, “AdaptHD: Adaptive Efficient Training for Brain-Inspired Hyperdimensional Computing,” BioCAS’19.
- Imani, M., Gupta, S., Kim, Y., & Rosing, T. “Floatpim: In-memory acceleration of deep neural network training with high precision,” ISCA’19
- M. Imani, T. Nassar, T. Rosing, "Moving Toward Real-Time Diagnostics using Brain-Inspired Hyperdimensional Computing", AACR in Cancer, 2019.
- Gupta, Saransh, et al. T. Rosing "RAPID: A ReRAM processing in-memory architecture for DNA sequence alignment." ISLPED’19.