High-Performance Neuromorphic Sensor Processing

Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE, AND RESILIENT COMPUTING (SHREC)

oneAPI HPC Developer Summit 2022

Dr. Alan George
Mickle Chair Professor of ECE
University of Pittsburgh

Luke Kljucaric
PhD Candidate
University of Pittsburgh

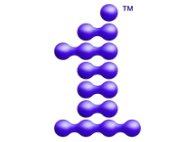
Dr. Ryad Benosman
Professor of ECE and Ophthalmology
University of Pittsburgh

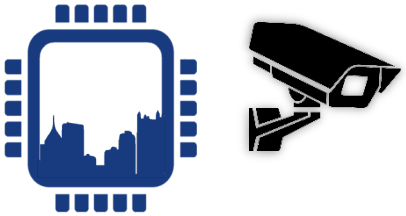
University of Pittsburgh
BYU Brigham Young University
Virginia Tech
UF University of Florida

Wednesday, December 7, 2022

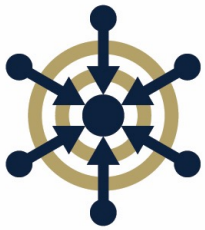
Overview

- What is SHREC?
- Background
  - HLS Technologies
  - Neuromorphic Technologies
- Motivations
- Approach
- Results
  - Performance
  - Resource Usage
- Conclusions





What is SHREC?

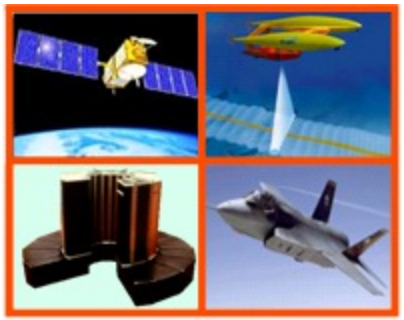


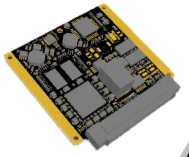
Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE, AND RESILIENT COMPUTING (SHREC)

- **NSF Center for Space, High-Performance, & Resilient Computing**
  - Founded in Sep. 2017, replacing highly successful ***NSF CHREC Center***
  - Leading ECE research groups @ four major universities
    - **University of Pittsburgh** (lead)
    - **Brigham Young University** (partner)
    - **University of Florida** (partner)
    - **Virginia Tech** (partner)
- **Under auspices of IUCRC Program at NSF**
  - **Industry-University Cooperative Research Centers**
    - Fostering university, agency, and industry R&D collaborations
  - **SHREC is both National Research Center and Consortium**
    - University groups serve as research base (faculty, students, staff)
    - Industry & government organizations are research partners, sponsors, collaborators, advisory board, & technology-transfer recipients

CHREC
NSF Center for High-Performance Reconfigurable Computing

Center Mission



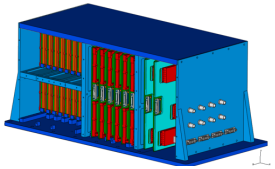


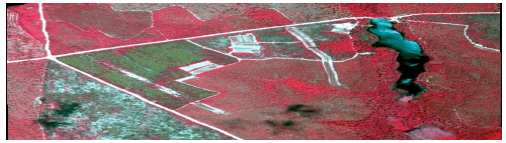
Space Computing

SHREC

High-Performance Computing

Resilient Computing





Theme: Mission-Critical Computing

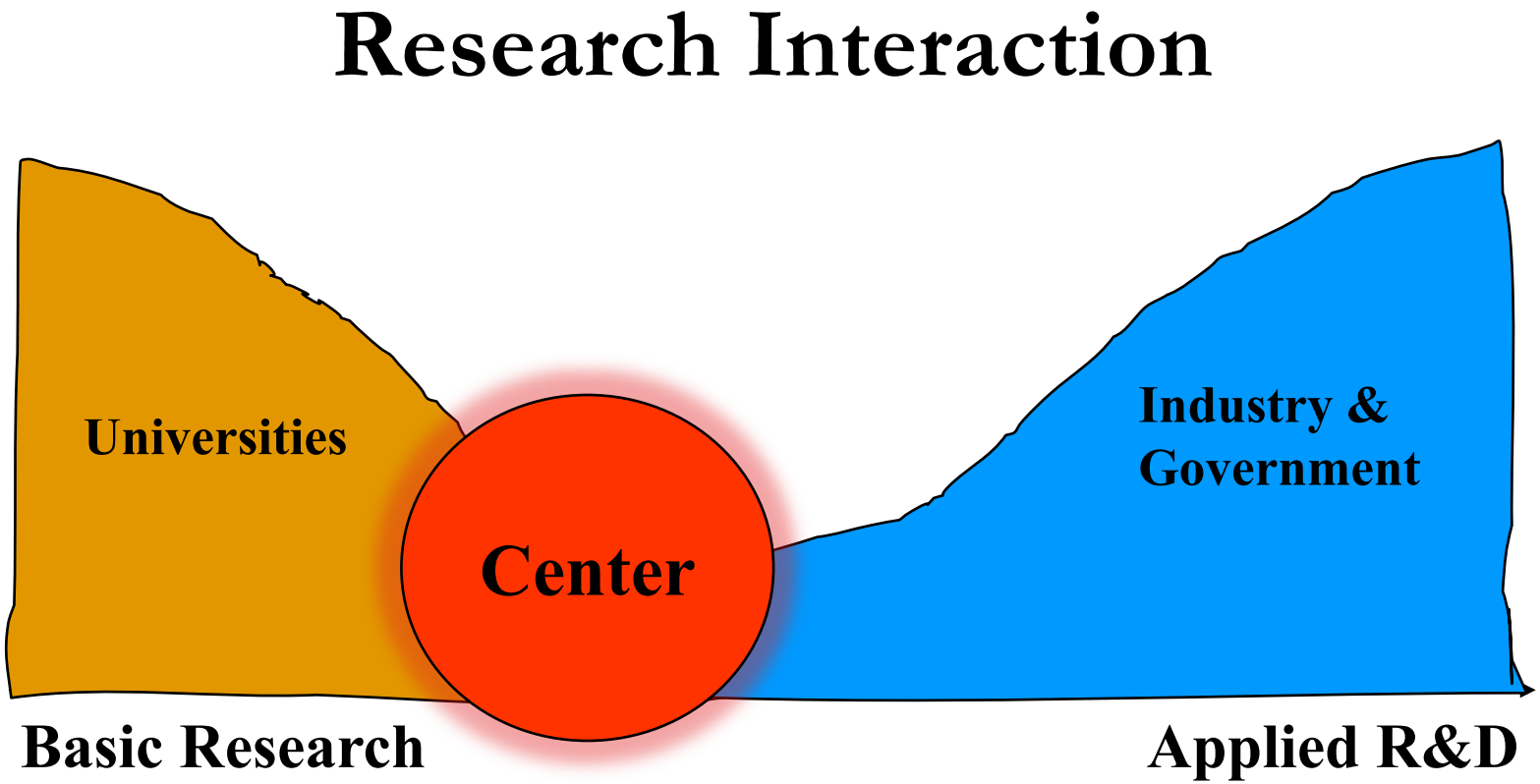
Basic and applied R&D to advance S&T on advanced computing.

Many common challenges, technologies, & benefits, in terms of performance, power, adaptivity, productivity, cost, size, etc.

From architectures to applications to design concepts and tools.

From spacecraft to supercomputers!

NSF Model for IUCRC Centers



CY22 Center Members

Each member funds 1 or more memberships

1. AFRL Space Vehicles Directorate
2. Army Research Laboratory
3. Astrobotic
4. BAE Systems
5. Ball Aerospace
6. Blue Origin
7. Dell EMC
8. Intel *
9. Fermilab
10. Genesis Engineering Solutions
11. GSI Technology
12. Honeywell
13. JHUAPL *
14. JPL
15. L3Harris
16. Laboratory for Physical Sciences
17. Lockheed Martin
18. Microchip *
19. MIT Lincoln Laboratory
20. MITRE
21. NASA Ames Research Center
22. NASA Goddard Space Flight Center
23. NASA IV&V Facility
24. NASA Johnson Space Center
25. National Security Agency
26. Naval Research Laboratory
27. Northrop Grumman *
28. Raytheon
29. Renaissance Associates
30. Satlantis
31. Space Micro

oneAPI For FPGAs

- Single-Source Programming Model
  - **Host** – CPU based, (Data-Parallel) **SYCL**
    - Configures buffers (data transferred to/from FPGA)
      - **Abstracts DMA/memory subsystems**
    - Kernel (accelerator) **code expressed inline**
    - **Unified shared memory**
    - Large set of library APIs to **efficiently leverage accelerator** devices
    - Can perform work concurrently (inherent **heterogeneous design**)
  - **Accelerator – SYCL**
    - DPC++ – **C++ and SYCL**
      - **Single-source** programming model
      - High-level abstraction for large **productivity** gains over RTL

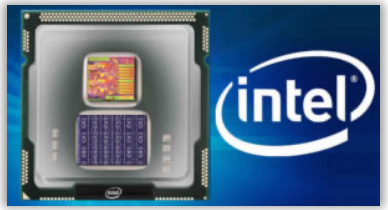
Neuromorphic Technology

- Asynchronous event-based vision sensors
  - **Mimic human-eye** functionality
    - Records individual **changes** in relative pixel **luminance**
  - Data sent as **stream of events**
    - Does not capture frames like conventional camera sensors
  - Events captured at **1 µs resolution**
    - State of every pixel is known at each event
    - Small number of events in real-world apps
    - **1M FPS effective frame rate**
    - **Efficient data rate**

PROPHESEE
METAVISION FOR MACHINES

Neuromorphic Technology

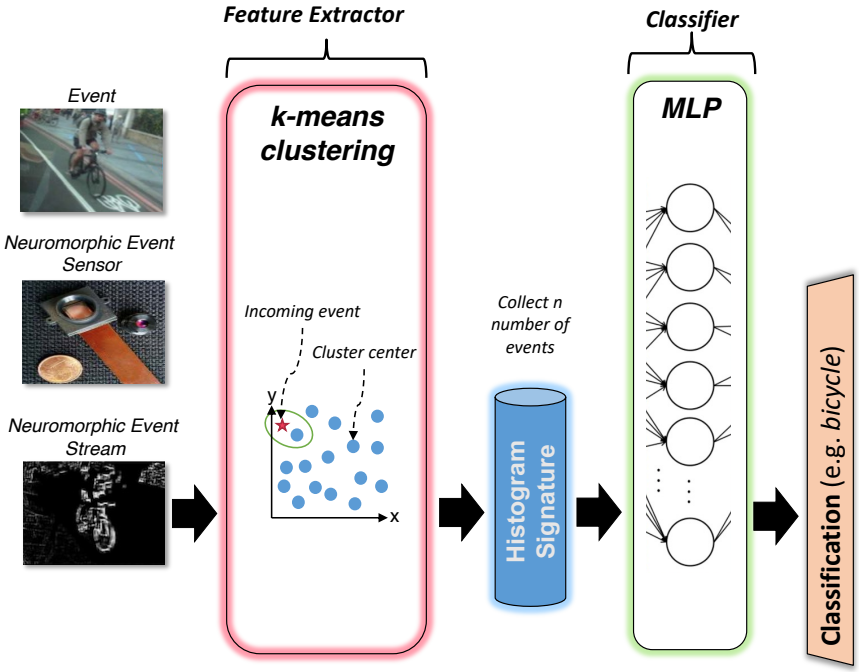
- Neuromorphic Architectures
  - **Mimic human-brain** functionality
    - Operates on **spikes** with many **interconnected neurons**
    - **Temporal- and rate-encoding** provides more information
  - Well suited for apps using **neuromorphic sensors**
    - Event-based data carries "spikes" with time
  - Intel Labs **Loihi** Architecture
- Algorithms
  - How do we **classify objects**?
  - Reconstruct frames over time?
    - Use CNNs?



intel labs

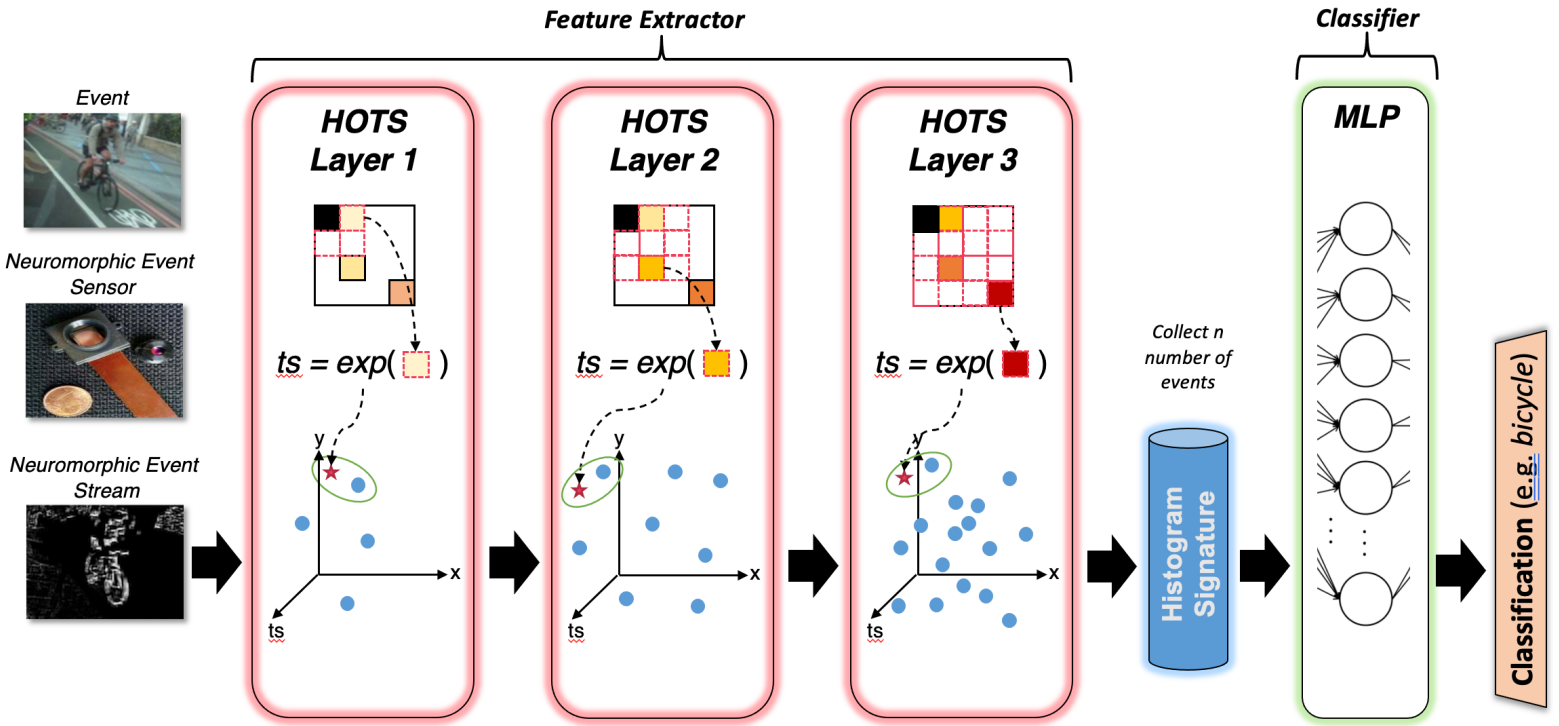
Event Clustering

- *k*-means Clustering
  - Events often manifest as clusters
    - **Common clusters** can be recognized as **features**
    - Potential issues with similar features in different locations



HOTS

- A Hierarchy Of event-based Time Surfaces
  - **Relates spatial and temporal** information
    - **Extracts object features** for classification



T1) Overview

Goals

- Create **low-latency**, **scalable** FPGA designs for **neuromorphic** algorithms
- Explore benefits of multiple algorithms for **neuromorphic feature extraction**

Challenges

- Neuromorphic sensors require **sub-microsecond** processing latency
- Tradeoffs between **compact designs** for scalability and **parallel performance**

Optimize

- Optimize FPGA designs with **oneAPI**

- **Reduce resource usage** and leverage emerging FPGA technology

Investigate

- Investigate tradeoffs of accelerating **SYCL** with **FPGAs** and **CPUs**

- Evaluate novel event-based datasets for **robust algorithm testing**

Scale

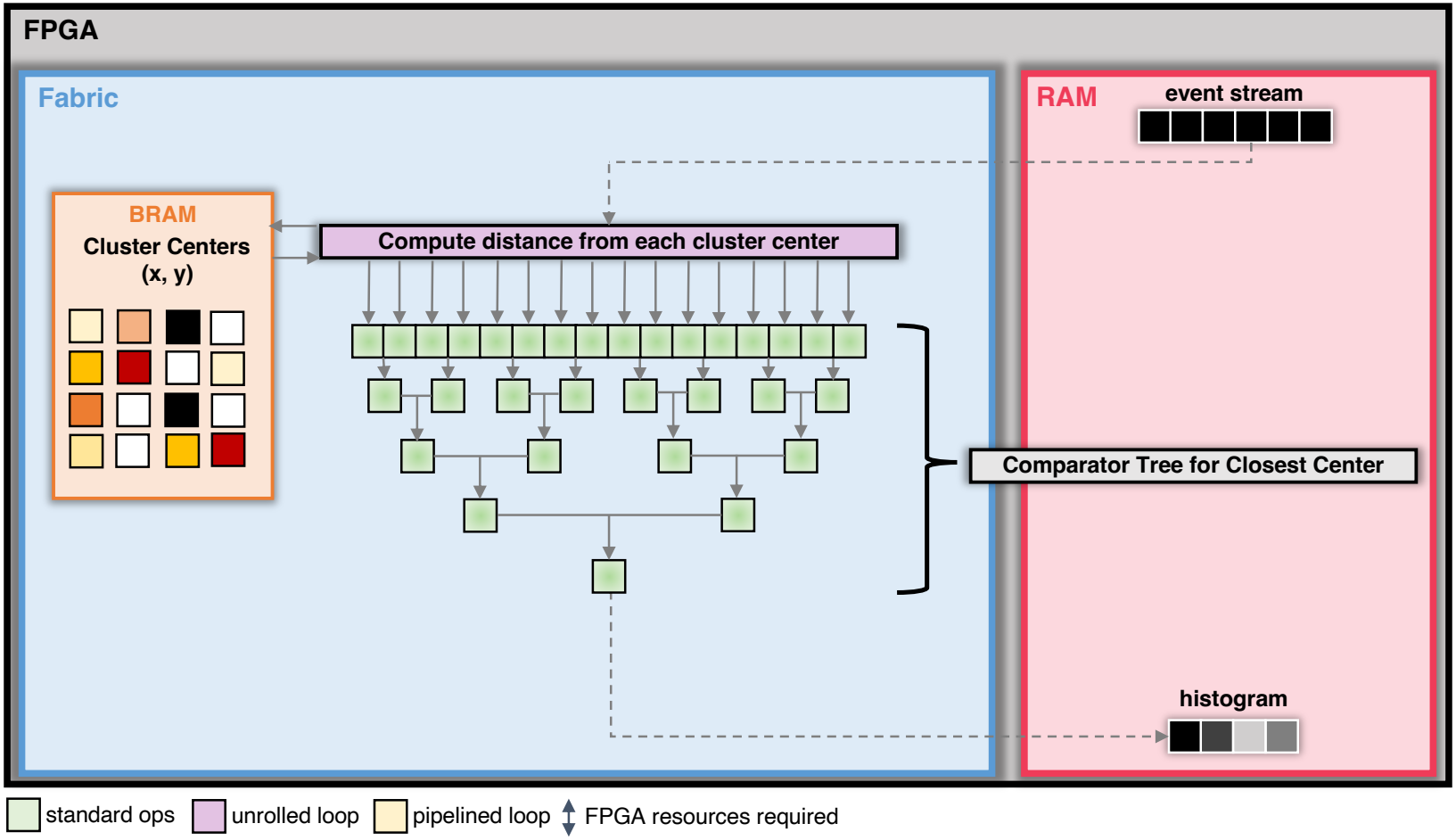
- Scale FPGA designs for **efficient resource usage**

- Analyze **performance vs. accuracy** scaling across both algorithms

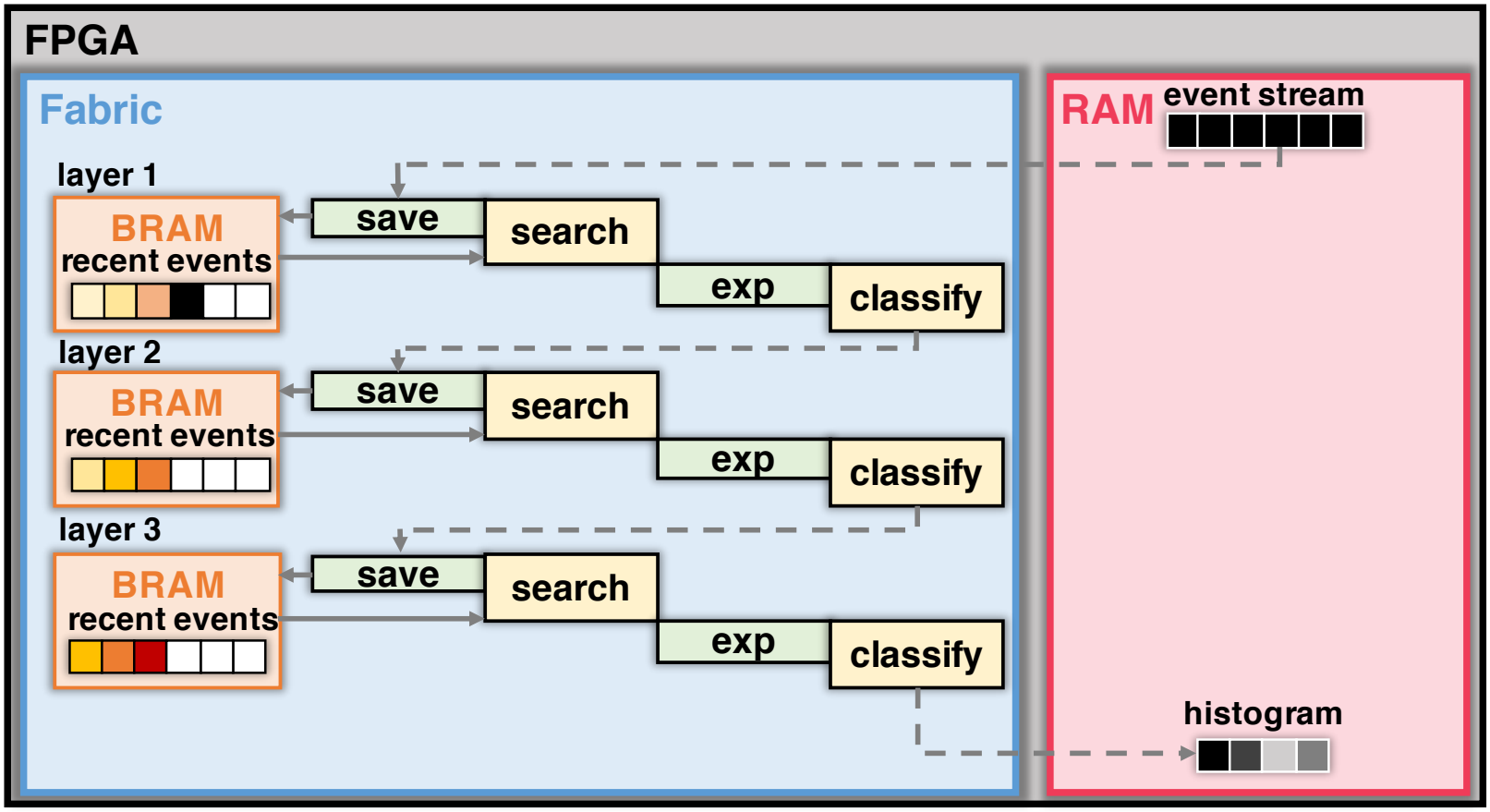
Approach

- Datasets
  - **N-MNIST** – Neuromorphic 1-to-1 MNIST
  - **N-Traffic** – Custom traffic-based dataset
    - Created with **dynamic vision sensor** for real-world data
- Classifiers
  - **Histogram** – How close features match class signature
  - **Multi-layer perceptron** – higher accuracy
- FPGA specifics
  - Events-per-stream (EPS)
    - Larger yields increased accuracy, smaller yields lower latency

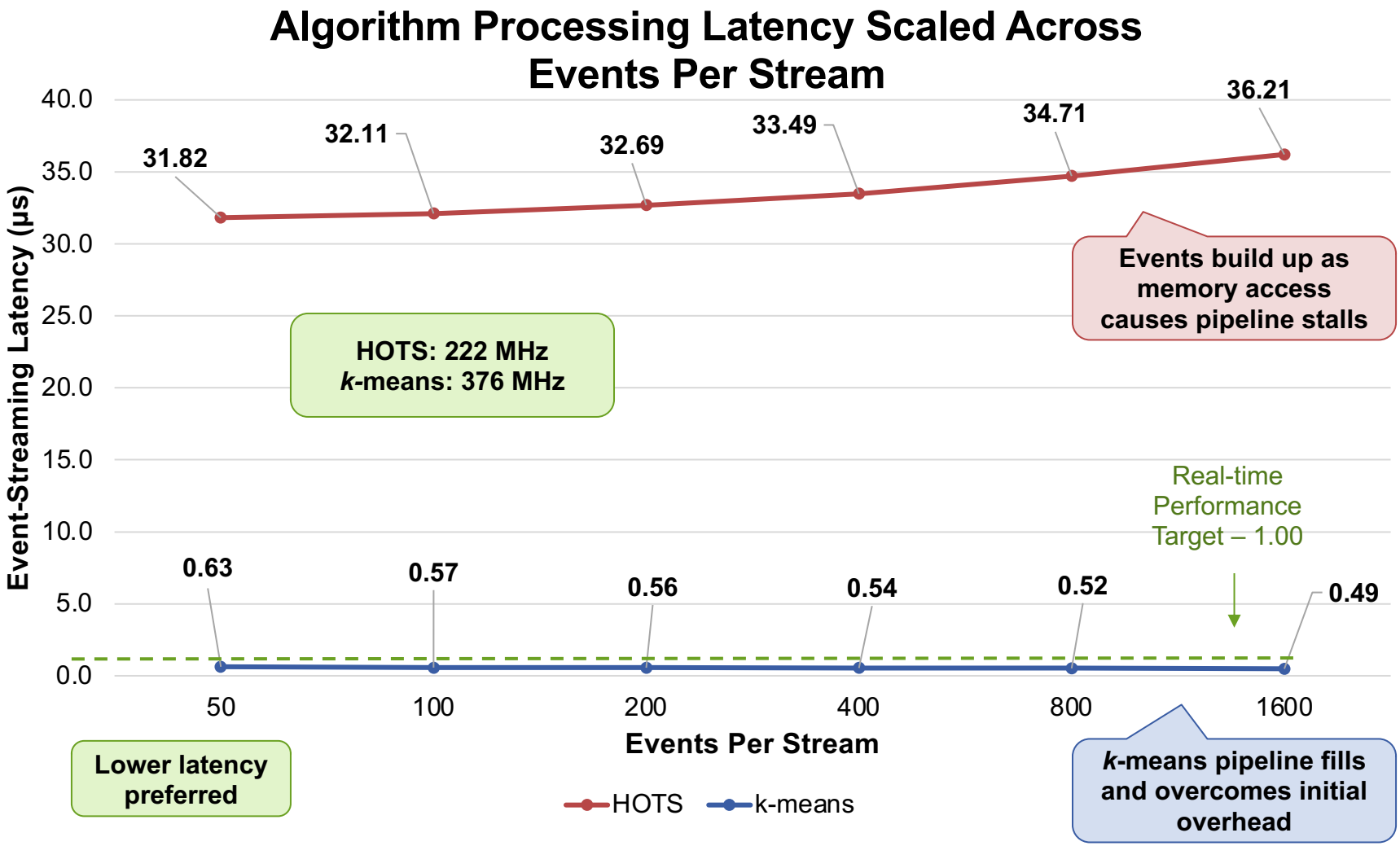
k-means FPGA Design



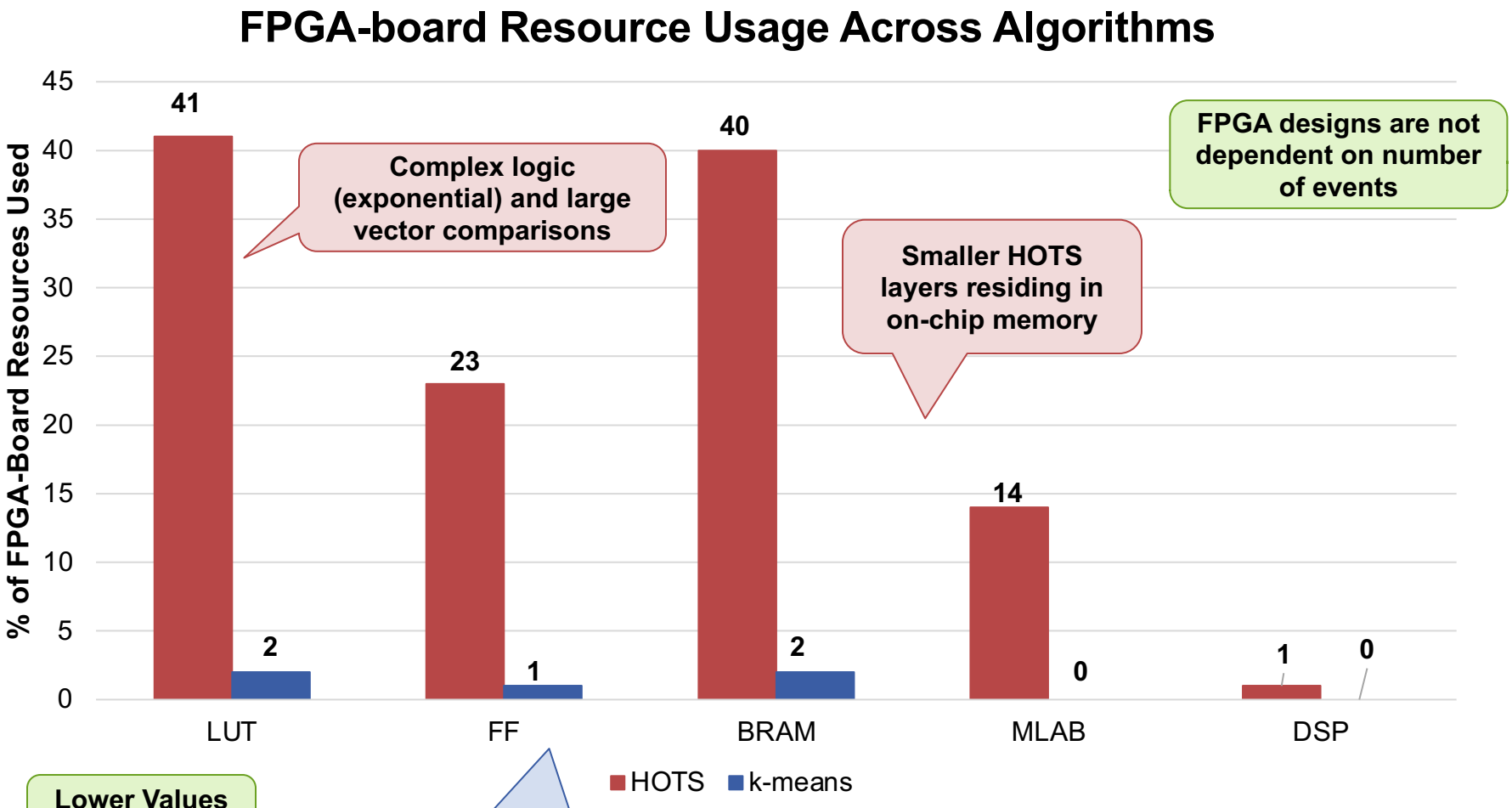
HOTS FPGA Design



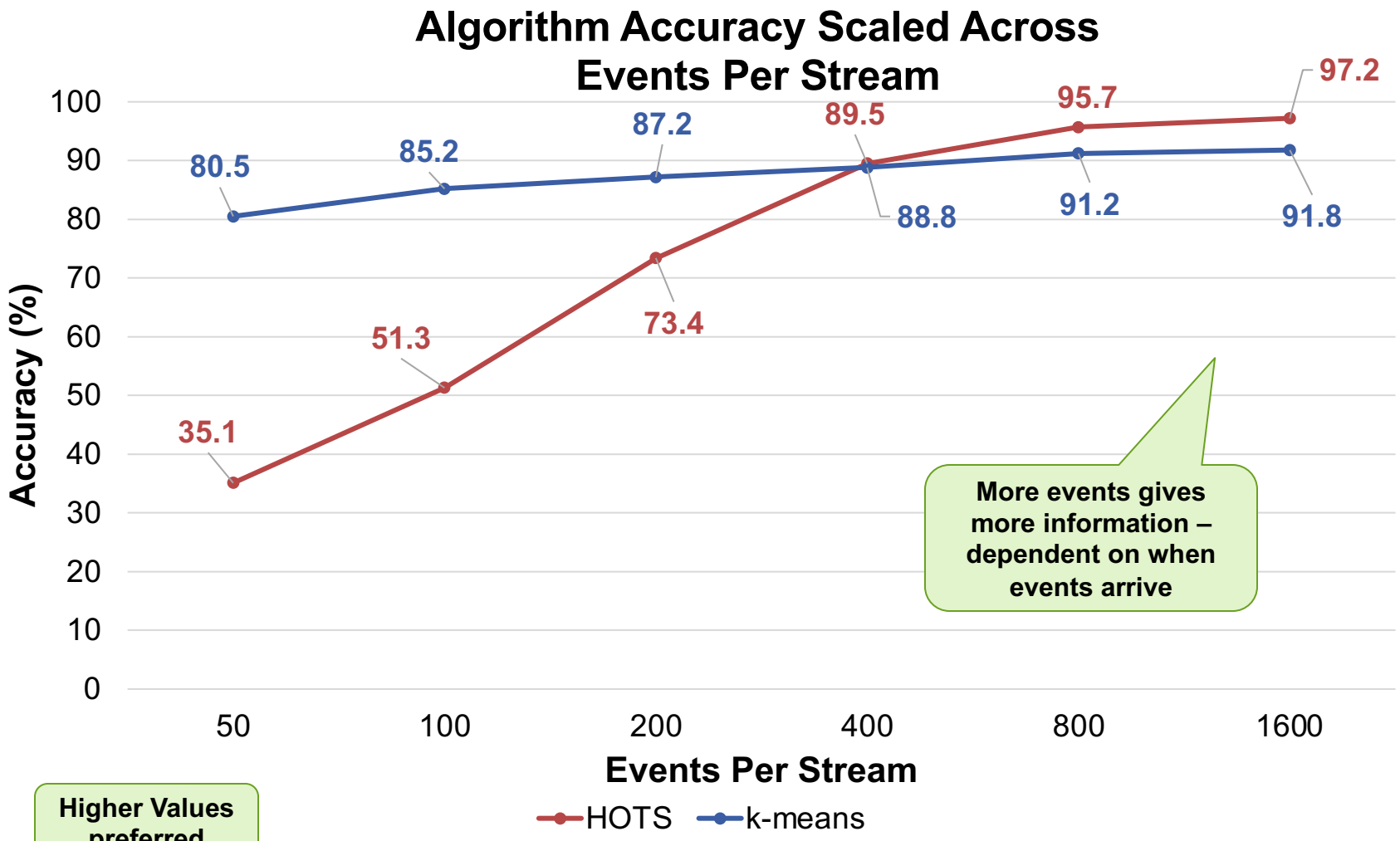
Performance Scaling (1/2)



Performance Scaling (2/2)



Accuracy Scaling



Discussion

- **Performance**
  - Realtime processing with *k*-means
    - Less complexity, smaller pipeline
    - HOTS exponential calculation ~ critical path
  - Large feature clustering with HOTS
    - 2D x and y points with k-means
    - 3-stage clustering in HOTS with 25-, 81-, and 289-point features
    - HOTS clusters are too large for fabric memory (costly stalls)
      - **Could be alleviated with HBM2 devices**
  - Better scalability with k-means
    - Large feature points require more resources, deeper pipeline
    - <5% resource with k-means can process multiple sensor inputs

Discussion

- **Accuracy**
  - K-means accuracy is better at lower EPS
    - Depth of HOTS requires more information
  - As number of events increase, HOTS accuracy improves
    - Does not support improvement to justify performance loss
  - Features in different spatial locations
    - Should present a problem in k-means
    - Accuracy remains high through 60K samples tested
  - Problem with current state-of-the-art datasets?
    - Not enough spatial information?
    - IBM Gesture dataset
    - Custom frequency-based dataset

Conclusions

- *Lower-latency performance with k-means*
  - **~74× faster** than HOTS with slightly less accuracy
- *More compact designs with k-means*
  - **23.8× less resources** used versus HOTS
- *Large-vector clustering degrades performance*
  - More-complex logic, more resources, longer latency, lower clock frequency – due to more memory accesses
- ***k*-means enables real-time neuromorphic event processing**
  - Stratix designs use **<5% of all available FPGA resources**, making designs **highly scalable**

Future Work

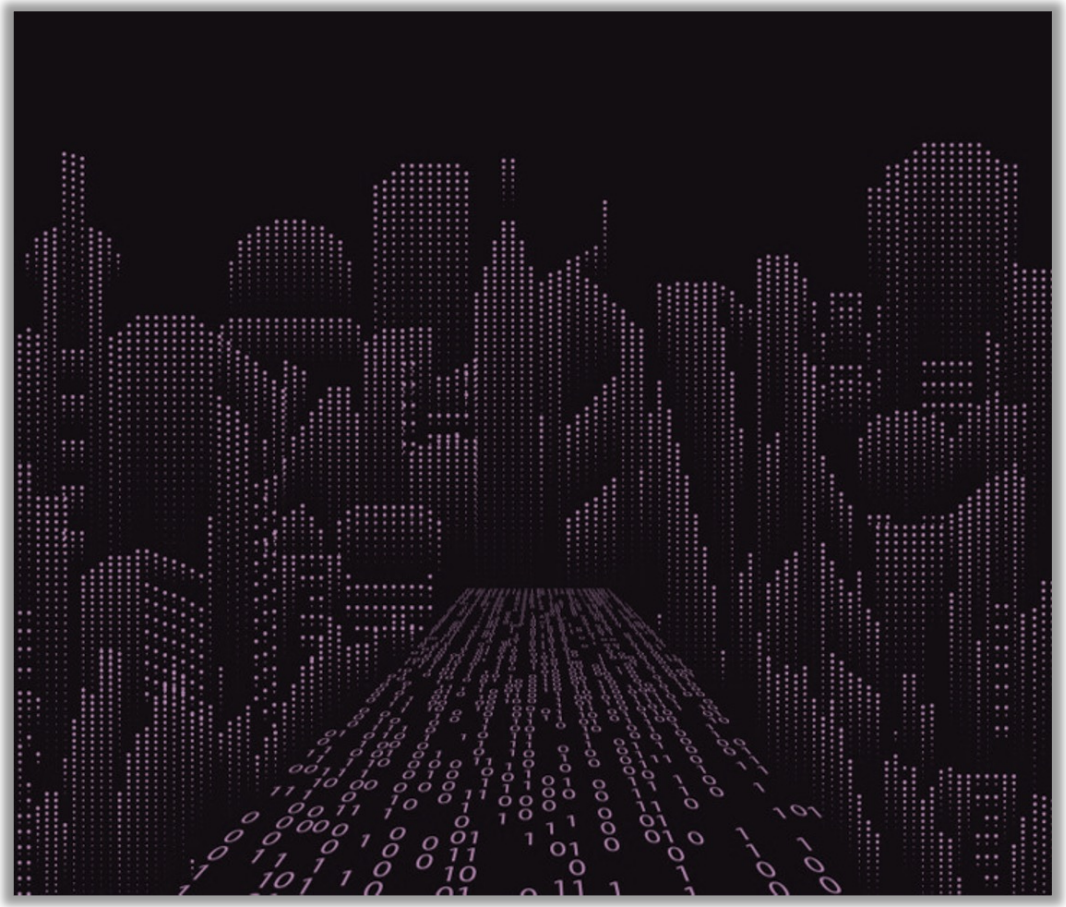
- **New dataset exploration**
  - IBM Gesture dataset
  - Custom dataset – Pendulum frequency classifier
  - Validate spatial and temporal relationships
- **HOTS optimizations**
  - Discretization of exponential calculation
    - SYCL implementation is efficient
  - Boost Trees over clustering?
- **New FPGA Technology**
  - Explore the use of on-chip HBM2 memory
    - Less memory penalties for clustering

Acknowledgment

- **Intel FPGA DevCloud**
  - Jeffery Nigh and Larry Landis
    - Resource access, debugging, guidance
- **NSF SHREC**
  - Michael Ing and Devon Callanan
    - Collaboration
- **This research was supported by SHREC industry and agency members and by the IUCRC Program of the National Science Foundation under Grant No. CNS-1738783.**

Questions?

- Thanks for your attention!



References

- [1] B. Jenkins, “Intel FPGA Design Workshop,” Intel Programmable Solu-tions Group. University of Pittsburgh Center for Research Computing.[Workshop on FPGA design using Intel Tools]. April 2019.
- [2] L. Kljucaric, “Deep-Learning Inferencing with High-Performance Hard-ware Accelerators,” 2019 IEEE High Performance extreme ComputingConference (HPEC), Waltham, MA, 2019.
- [3] X. Lagorce, G. Orchard, Et al., “HOTS : A Hierarchy Of event-basedTime-Surfaces for pattern recognition”, IEEE Trans. Pattern AnalysisMachine Intelligence, July 2016. doi :10.1109/TPAMI.2016.25747
- [4] G. Martin, G. Smith, “High-level synthesis: Past, present, and future,”IEEE Design & Test of Computers, vol. 26, issue 4, pp. 18–25, August2009.
- [5] F. Vahid, “Digital Design with RTL Design, Verilog and VHDL” (2nded.). John Wiley and Sons. p. 247. ISBN 978-0-470-53108-2. 2010.
- [6] Intel Corporation, “oneAPI Specification,” Intel oneAPI Specification.2020. https://spec.oneapi.com/versions/latest/index.html
- [7] J. Deng, W. Dong, Et al., ”ImageNet: A Large-Scale Hierarchical ImageDatabase,” CVPR09, 2009. http://image-net.org/about-publication
- [8] J. Deng, W. Dong, Et al., ”ImageNet: A Large-Scale Hierarchical ImageDatabase,” CVPR09, 2009. http://image-net.org/about-stats
- [9] S.-C. Liu, T. Delbruck, “Neuromorphic sensory systems,” Current Opin-ion in Neurobiology, vol. 20, issue 3, pp 288–295. June 2010.
- [10] Y. Lecun, L. Bottou, Et al., “Gradient-based learning appliedto document recognition.” Proc. IEEE 86, 2278–2324. 1998. doi:10.1109/5.726791
- [11] G. Orchard, G. Cohen, Et al., “Converting Static Image Datasets to Spik-ing Neuromorphic Datasets Using Saccades”, Frontiers in Neuroscience,vol.9, issue 437, Oct. 2015
- [12] J. Lu, J. Dong, Et al., “An Event-based Categorization ModelUsing Spatio-temporal Features in a Spiking Neural Network,”2020 12th International Conference on Advanced ComputationalIntelligence(ICACI),Dali,China,2020,pp.385-390,doi:10.1109/ICACI49185.2020.9177628
- [13] A. Sironi, M. Brambilla, Et al. “HATS: Histograms of Averaged TimeSurfaces for Robust Event-based Object Classification”. To appear inIEEE Conference on Computer Vision and Pattern Recognition (CVPR),2018. arXiv:1803.07913
- [14] R. Ghosh, A. Mishra, Et. al, “Real-time object recognition and ori-entation estimation using an event-based camera and CNN,” 2014IEEE Biomedical Circuits and Systems Conference (BioCAS) Pro-ceedings, Lausanne, Switzerland, 2014, pp. 544-547, doi: 10.1109/Bio-CAS.2014.6981783
- [15] M. Hofst¨atter, M. Litzenberger, Et al., “Hardware-accelerated address-event processing for high-speed visual object recognition,” 2011 18thIEEE International Conference on Electronics, Circuits, and Systems,Beirut, Lebanon, 2011, pp. 89-92, doi: 10.1109/ICECS.2011.6122221.
- [16] E. Luebbers, S. Liu, Et al., “Simplify Software Integration for FPGA Accelerators with OPAE” Intel FPGA White Paper. 2021.https://01.org/sites/default/files/downloads/opae/open-programmable-acceleration-engine-paper.pdf
- [17] Intel Corporation , “Intel FPGA Programmable Acceleration Card D5005 Data Sheet” Intel Product Data Sheet. 2019. https://www.intel.com/content/www/us/en/programmable/products/boardsandkits/dev-kits/altera/intel-fpga-pac-d5005/documentation.html