



# Energy Estimation Framework for AI Systems on Silicon

**Murat Isik, Vedant Karia, Sadasivan Shankar**



NATIONAL  
ACCELERATOR  
LABORATORY

16<sup>th</sup> Aug 2023

# OUTLINE

- Motivation
- Goals
- Methodology for Energy Estimates along Three Vectors
  - Applications-NN Architectures-HW Architectures
- Results & Analysis
- Analysis Tools

# ACKNOWLEDGMENTS

- Dr. Dhireesha Kudithipudi, Anurag Daram

- Mike Davies, Dr. Tim Shea



- Ceyhun Kayan



- Steven Abreu



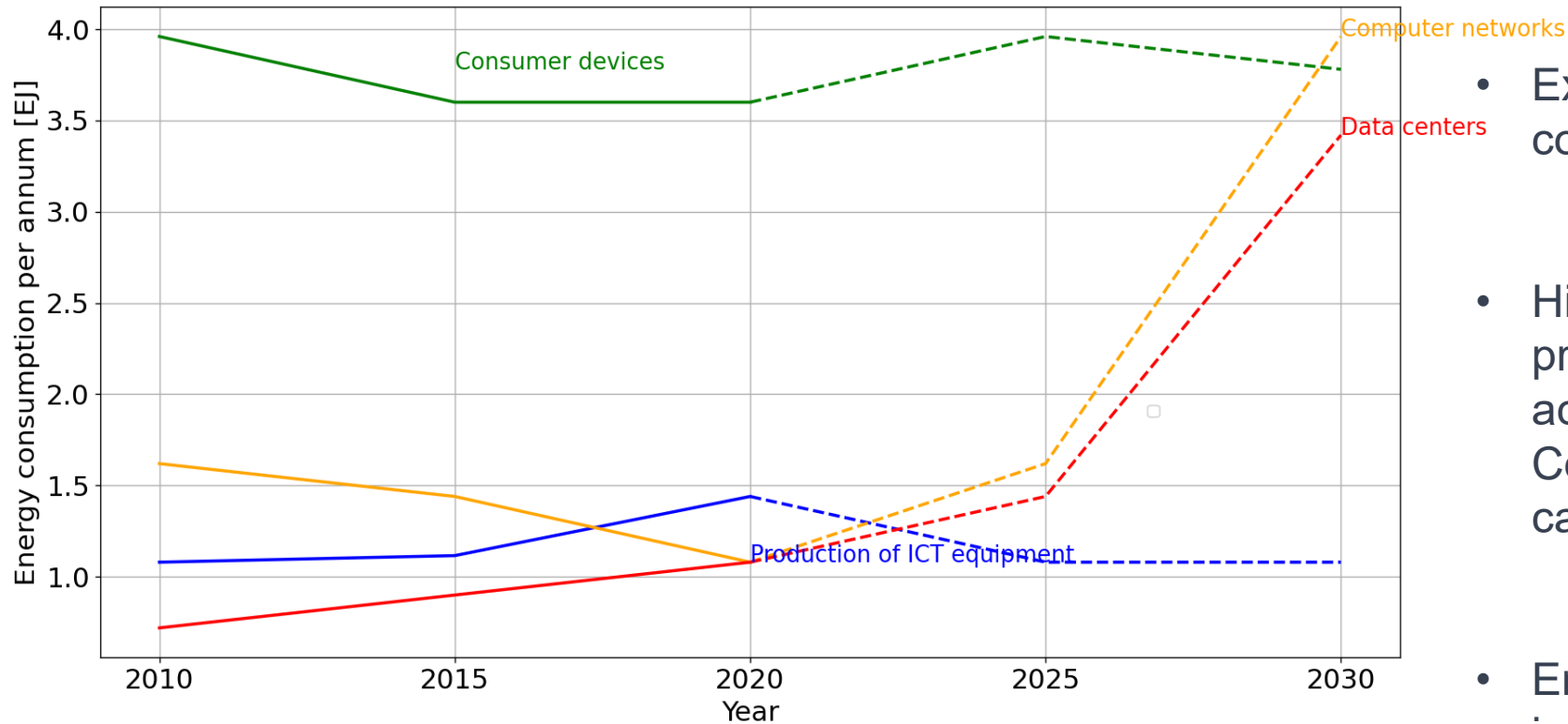
university of  
 groningen

- Sen Liu, Chris Tessone, Paul McIntyre (SLAC)

- Funding Support: AMMTO U.S. Department of Energy's Office of Science contract DE-AC02-76SF00515 with SLAC through an Annual Operating Plan agreement WBS 2.1.0.86



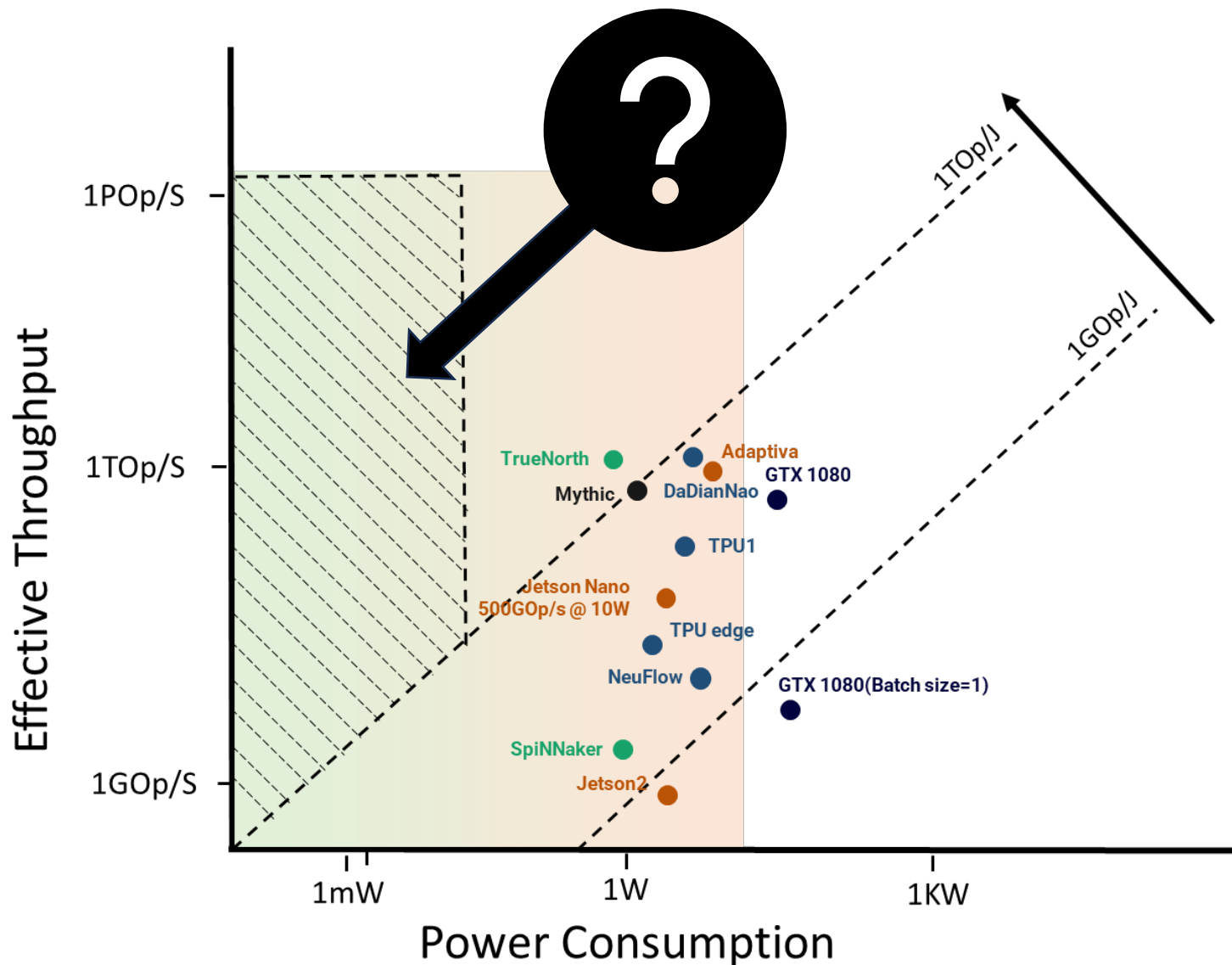
# MOTIVATION



- Expected in energy consumption in computer networks and data centers
- Highlights both the historical and projected energy consumption trends across various Information and Communication Technology (ICT) categories from 2010 to 2030
- Energy estimates show unsustainable increase over this decade unless some specific actions are undertaken

Estimated annual ICT energy consumption[1]


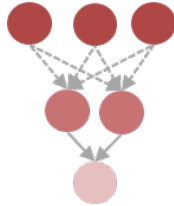
# MOTIVATION



- Specifications provided for GPUs tend to show very high performance
- Batch size is the number of samples processed before the model parameters are updated
- Reducing the batch size of training the ResNet neural network, throughput of **GTX 1080 GPU** drops 30 times from **900 GOps/W** to **30 GOps/W**

Quantification of energy efficiency is needed for different Architectures for different Applications

# GOALS

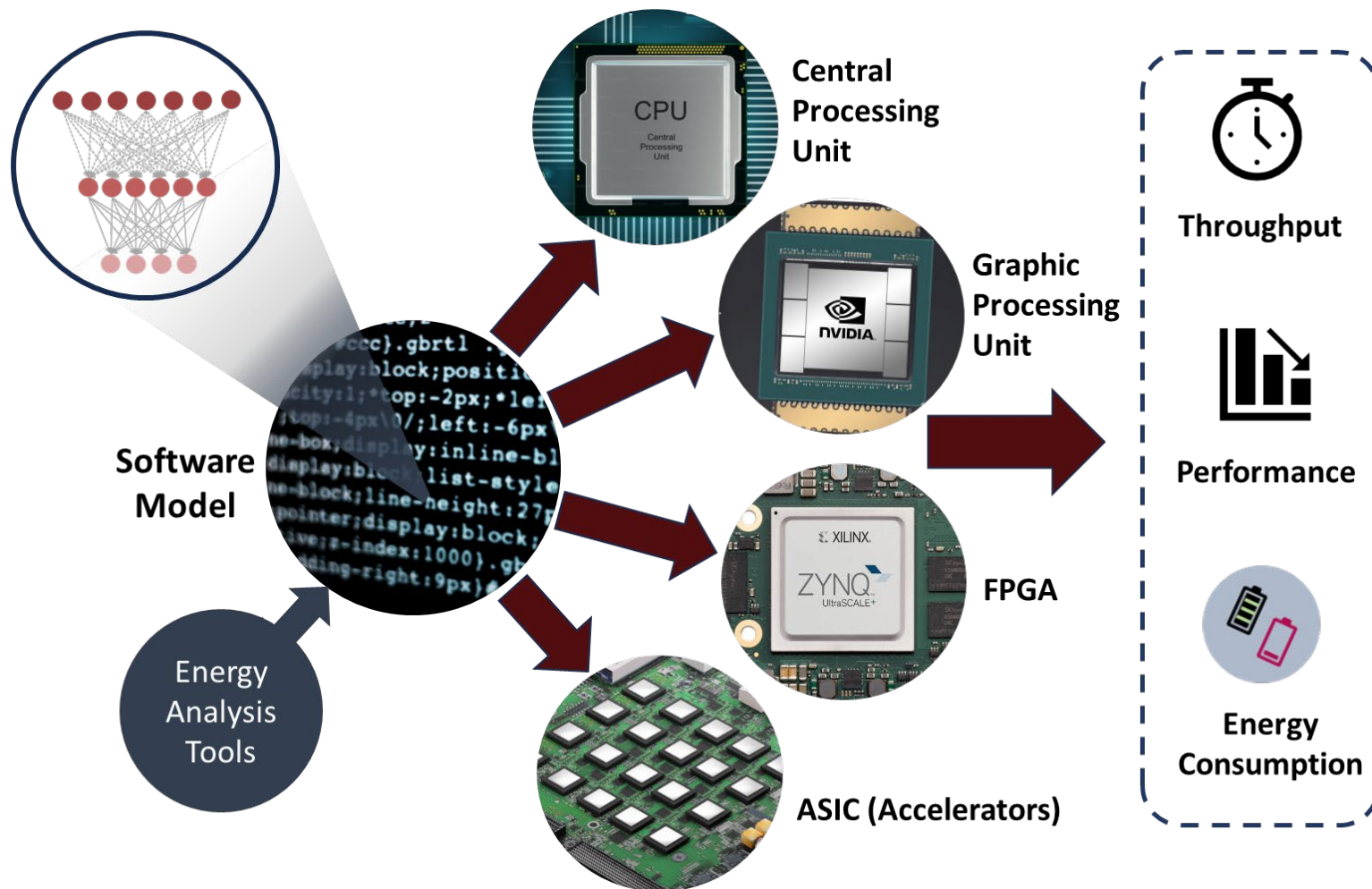
- Estimate the **energy consumption and throughput** performance of a Neural Network on various hardware architectures
  - CPU, GPU, FPGA, ASIC, Heterogenous Architectures 
- Assess the performance of **several applications** on various architectures as well as their energy consumption
  - Natural language processing
  - Computer Vision
  - Time Series
- Analyze the trends in energy consumption of wide range of commonly used **neural network architectures**
  - Spiking Neural Networks 
  - Convolutional Neural Networks
  - Transformers



# Methodology

Applications, Neural Network Architectures, Hardware Architectures

# METHODOLOGY



- **Methodology 1:**

- Tool : NVIDIA SMI & Intel power gadget
- Estimates the peak power of process which further used to calculate energy

- **Methodology 2:**

- Tool : pyJoules
- Estimates energy & execution time of a process
- Running Average Power Limit technology from Intel to estimate energy.



# Applications

*Computer  
vision*

*Natural  
Language  
Processing*

*Time  
Series data*

- **Computer Vision**

- Computer Vision was selected due to its computational intensity in tasks like image recognition, its ubiquitous applications from healthcare to autonomous driving

- **Natural Language Processing**

- Natural Language Processing was chosen due to its central role in human-machine communication, its computational demand in handling complex language structures, its applications ranging from sentiment analysis to language translation, and its contemporary relevance with models like Chat-GPT

- **Time-Series**

- Time-Series analysis was selected for its complexity in processing sequential data requiring specialized algorithms and continuous monitoring, its industrial applications such as in manufacturing (e.g., SSRL radiography for additive manufacturing, Data collection)

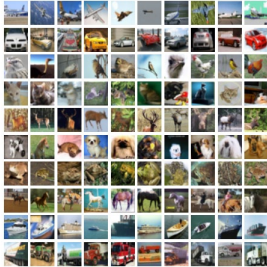
**Applications**

# Applications

## Computer Vision



airplane  
automobile  
bird  
cat  
deer  
dog  
frog  
horse  
ship  
truck



MNIST dataset[1]

CIFAR10 dataset[2]

- MNIST dataset:
  - 60,000 Images of Handwritten digits
  - Image size: 28x28 pixels
- CIFAR10 dataset
  - 60,000 Images of various objects
  - 32 x 32 pixels of RGB images

## Natural Language Processing

Roughly, how much oxygen makes up the Earth crust?  
Ground Truth Answers: almost half | almost half | half | almost half | half  
Prediction: half

What is the atomic number of the element oxygen?  
Ground Truth Answers: 8 | 8 | 8 | 8 | 8  
Prediction: 8

Of what group in the periodic table is oxygen a member?  
Ground Truth Answers: chalcogen | chalcogen | chalcogen | the chalcogen group  
Prediction: chalcogen

- Question & Answer dataset [4]
  - 690,000 words worth of cleaned text from Wikipedia that was used to generate the questions

## Time Series data



Google Speech Command sample[3]

- Google speech command dataset:
  - 65,000 utterances of one-second long of 30 short words

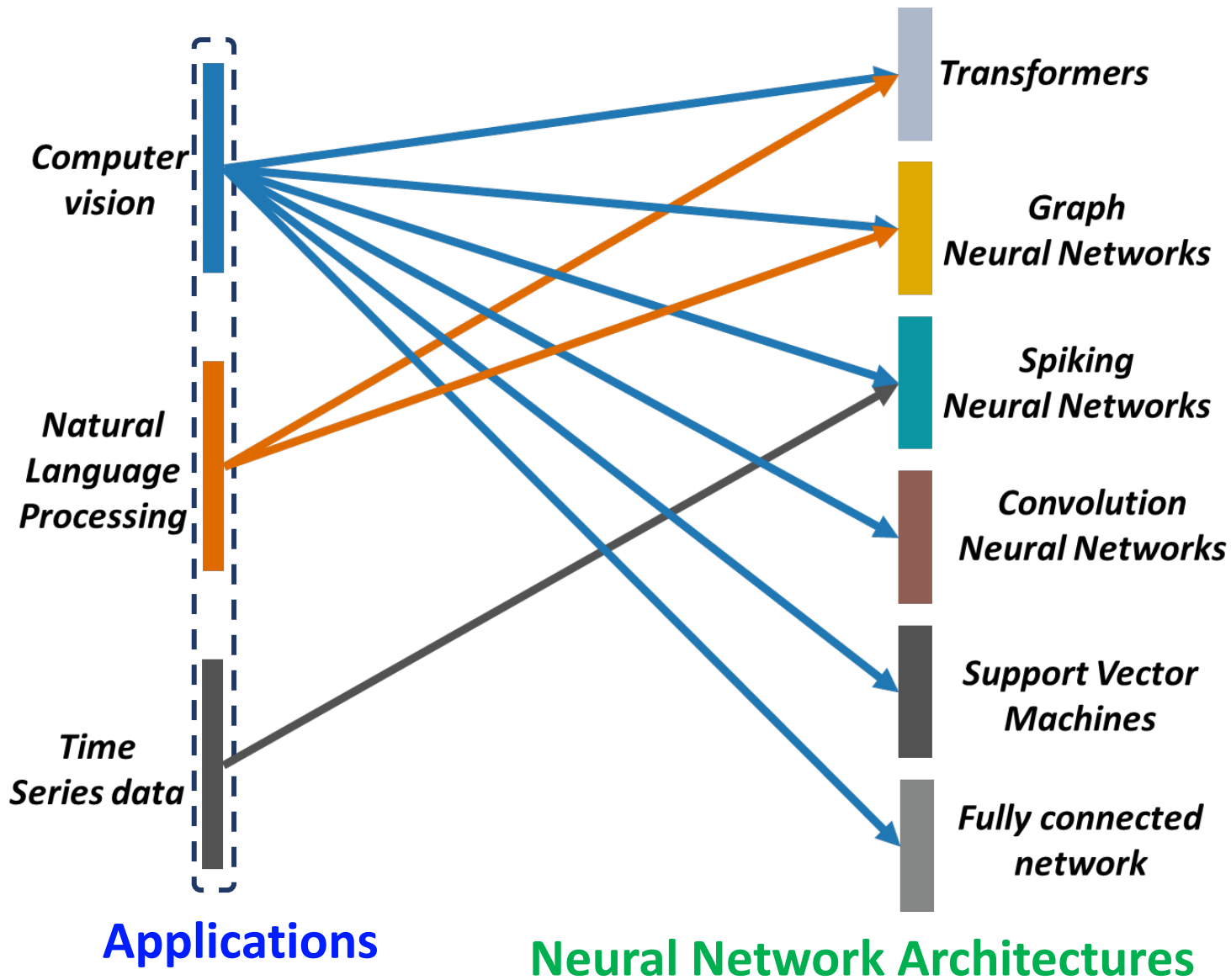
[1] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[2] Krizhevsky, A., Nair, V., & Hinton, G. (2010). Cifar-10 (Canadian institute for advanced research). URL <http://www.cs.toronto.edu/kriz/cifar.html>

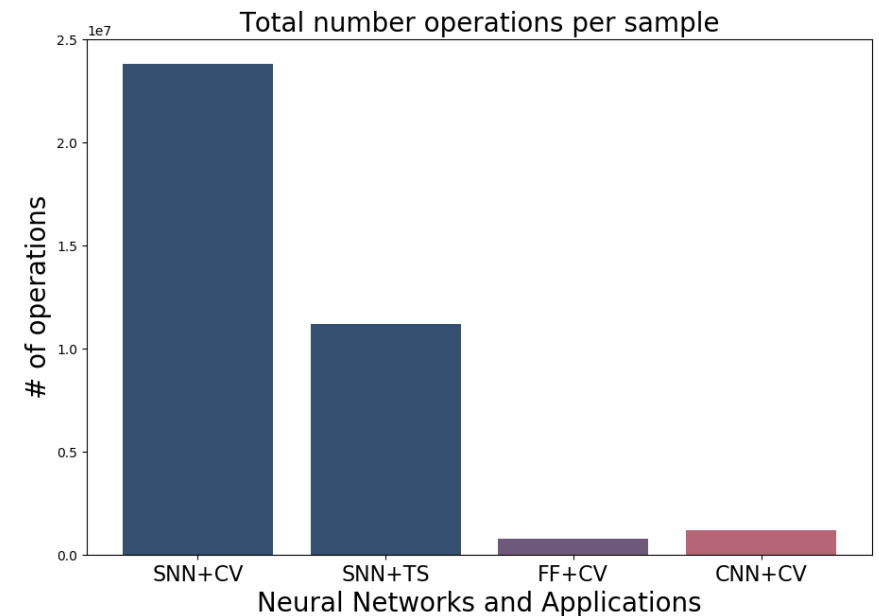
[3] Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209* (2018).

[4] Noah A Smith, Michael Heilman, and Rebecca Hwa. Question generation as a competitive undergraduate course project.

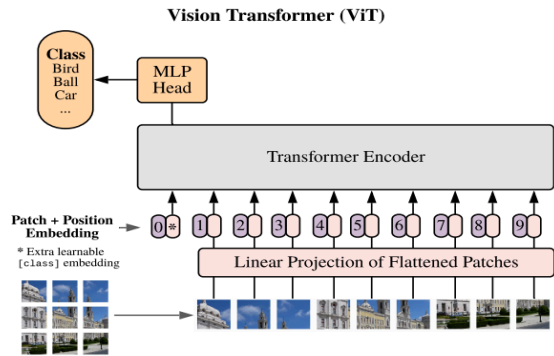
# Neural Network Architectures



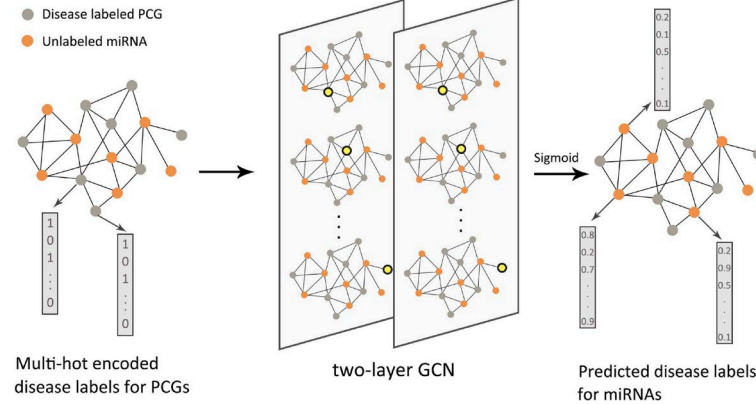
- **Five types of Neural Networks** chosen to analyze energy consumption
- Networks are organized in order of number of computations from bottom to top



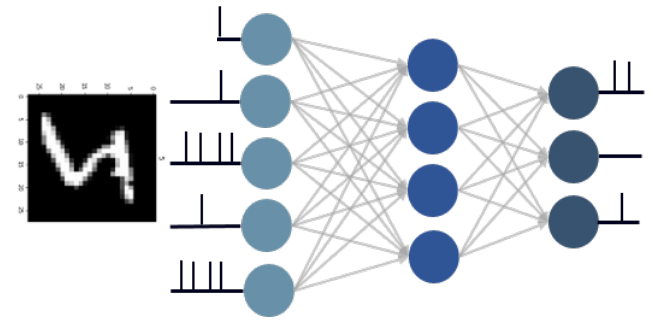
# Neural Network Architectures



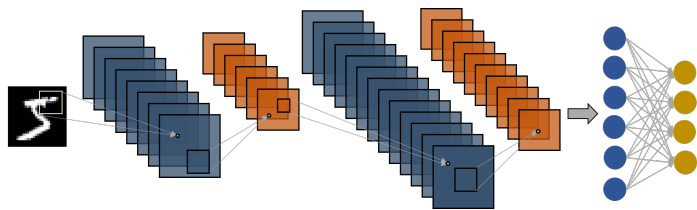
Transformers



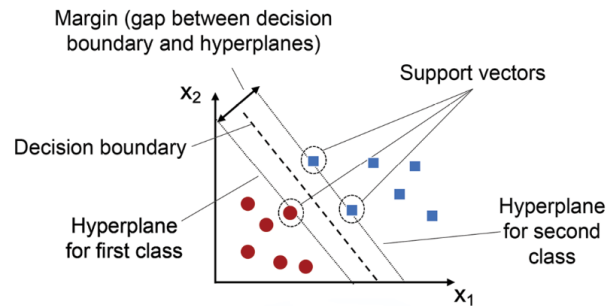
Graph Neural Networks



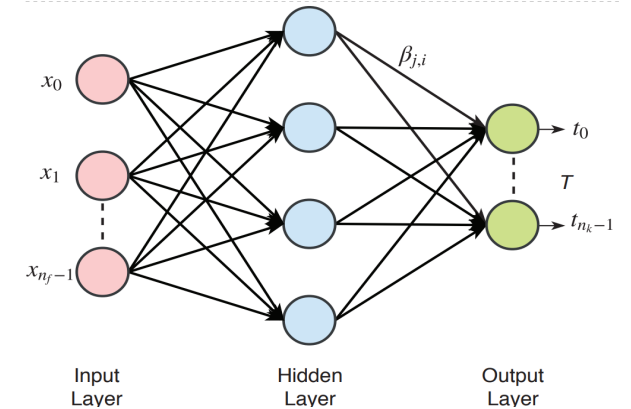
Spiking Neural Networks



Convolutional Neural Networks



Support Vector Machine

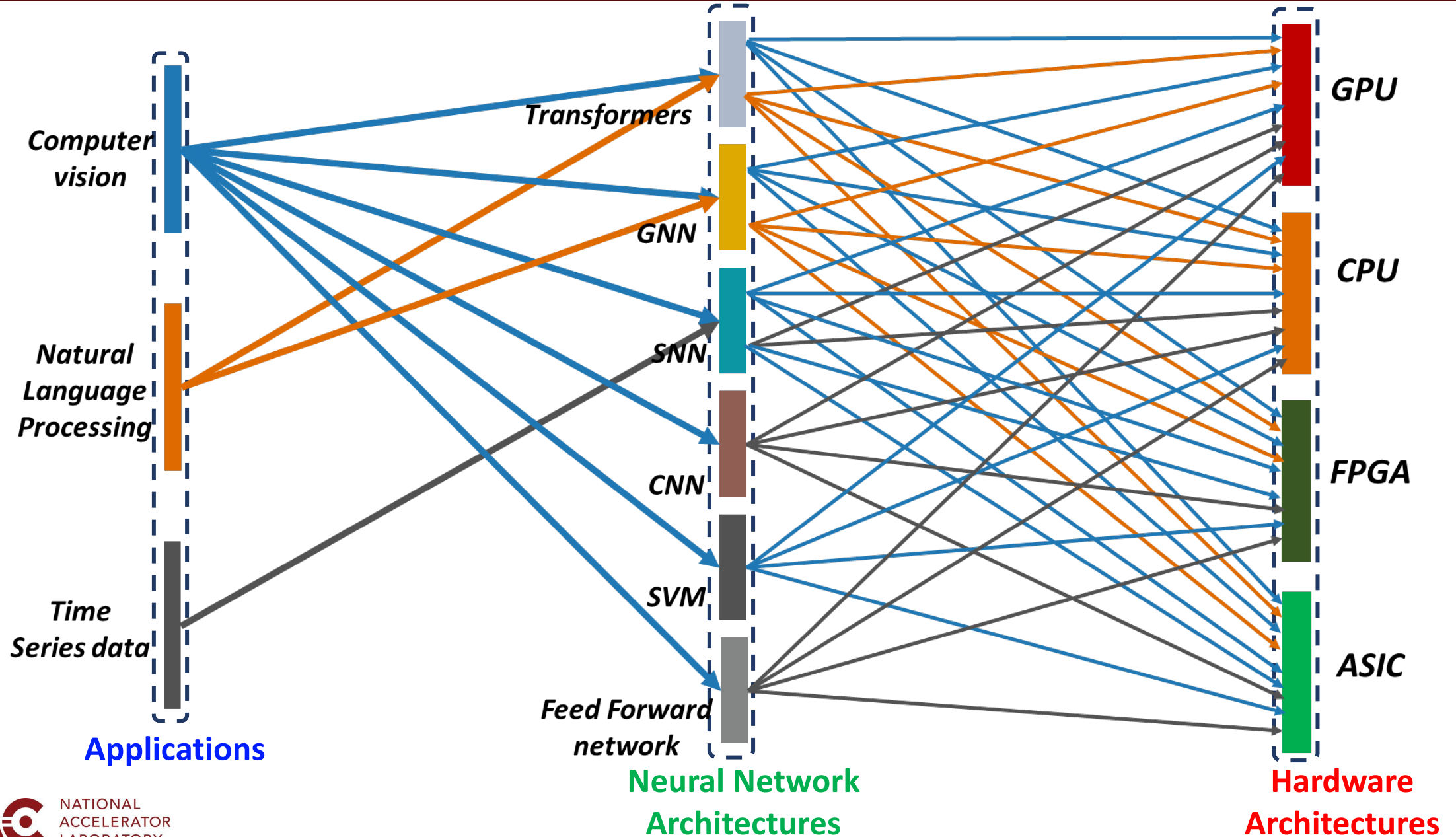


Feed Forward Neural Networks

# Neural Network Architectures

Architecture	Advantages	Drawbacks
<b>Transformers</b>	Highly parallelizable, handles long-range dependencies, powerful for translation & summarization	Computationally intensive, memory-hungry
<b>Graph Neural Networks</b>	Handles non-Euclidean data, captures relationships between nodes in a graph	Computationally expensive, requires careful preprocessing
<b>Spiking Neural Networks</b>	Emulates biological neural behavior of spiking for activity, potential energy efficiency, understanding of biological brain functions	Complex to train, asynchronous nature can lead to difficulties
<b>Convolutional Neural Networks</b>	Exceptional at image/video processing, automatically learns spatial hierarchies of features	Primarily for grid-like data, not suitable for non-image tasks without adaptation
<b>Support Vector Machine</b>	Effective in high-dimensional spaces, memory efficient, versatile, suitable for linear/nonlinear data	Kernel selection difficult, sensitive to noise, can be inefficient, computationally intensive for large datasets
<b>Feed Forward Neural Networks</b>	Simplicity, broad applicability, ease of training	No temporal behavior, limited in complex sequential tasks

# Overall Summary: *Applications*-*NN Architectures*-*Hardware Architectures*



# Hardware Architectures

## Advantages

- Optimized for **complex serial processing**
- High user programmability

## Disadvantages

- **Low** Throughput
- Low **core count** for parallel computing

**CPU**

## Advantages

- **High** throughput
- **Larger number of cores** for parallel operations

## Disadvantages

- High **Programming** complexity
- **High** operational **energy cost**

**GPU**

## Advantages

- **High** Reconfigurability
- Low **cost of design**
- **Low latency**
- **Power** efficient

## Disadvantages

- **Limited** software support
- High **development time**
- Low operating **frequency** unlike CPU, GPU

**FPGA**

## Advantages

- High **throughput**
- Low power consumption
- Application specific accelerators

## Disadvantages

- High development cost
- Low user programmability and user reconfigurability

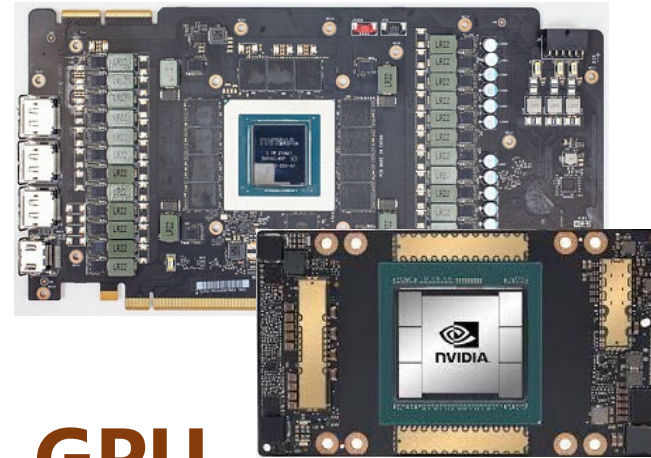
**ASIC**

# Hardware Architectures

- AMD threadripper 5995wx
  - TSMC 7nm technology node
  - Freq: 2.7GHz
  - # of cores: 64
- Intel i9 12900HA
  - 8nm technology node
  - Freq: 3.7GHz
  - # of cores: 14



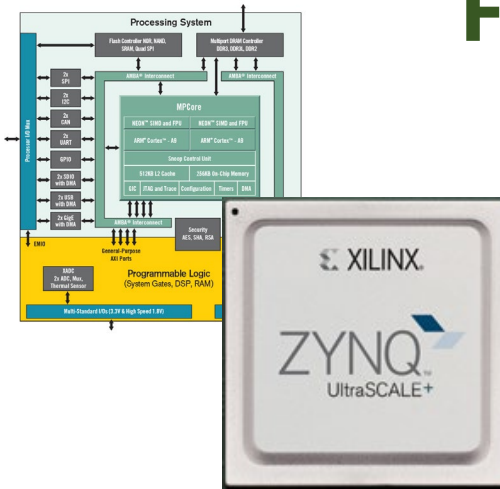
CPU



GPU

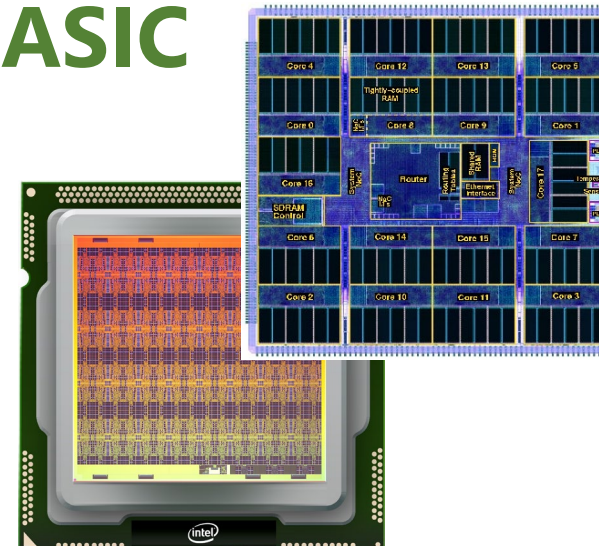
- NVIDIA RTX 3090
  - Samsung 8nm Technology node
  - Freq: 1.7GHz
  - # of cuda cores: 10,496
- NVIDIA RTX 3060
  - Samsung 8nm Technology node
  - Freq: 900 MHz
  - # of cuda cores: 3584

- ZYNQ 7000 series FPGA
  - AMD manufactured
  - ARM A9 processor
  - 13,300 LUTs
- AMD ZCU104 board
  - Zynq Ultrascale+ FPGA
  - ARM cortex A53 processor
  - 504,000 LUTs



FPGA

ASIC



- Intel Loihi[1]
  - 14nm Technology node
  - Neurons: 131,072
  - Synapses: 130,000,000
- SpiNNaker[2]
  - 18 ARM A9 processors
  - Freq: 200MHz

[1] Davies, Mike, et al. "Loihi: A neuromorphic manycore processor with on-chip learning." *Ieee Micro* 38.1 (2018): 82-99.  
 [2] Furber, Steve B., et al. "The spinnaker project." *Proceedings of the IEEE* 102.5 (2014): 652-665.  
 Picture credits AMD, Nvidia & Intel.





# Preliminary Analysis

Applications, Neural Network Architectures, Hardware Architectures

# Methods for Estimating Energy

## • Method 1:

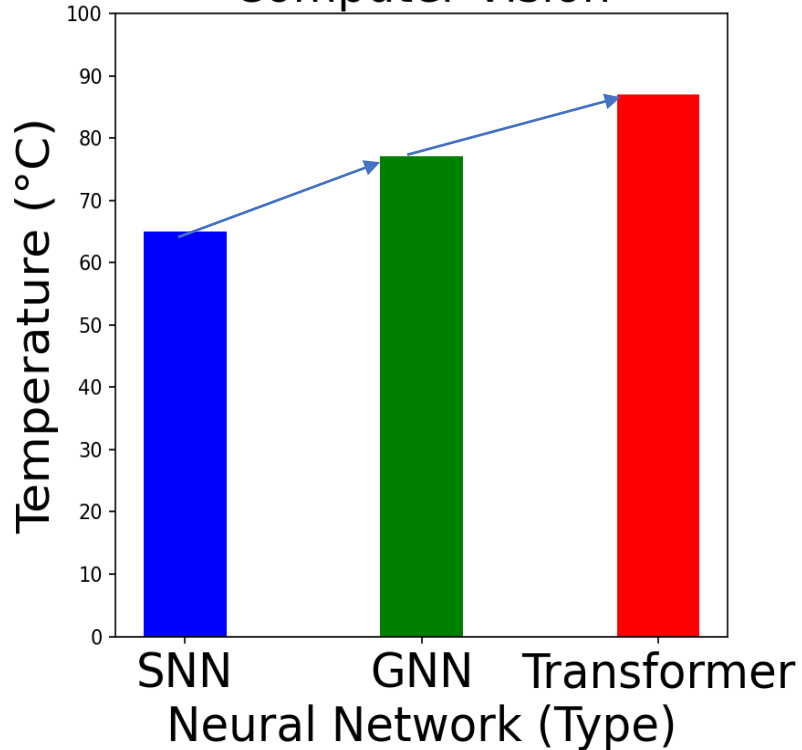
- It uses a hardware manually monitoring tool **NVIDIA SMI** to analyze the power and temperature of the chip
- It uses **peek power** to estimate the overall energy of the system
- It was evaluated on **Transformers, GPUs & SNN**
- Other processes on system do not affect this methodology
- **CPU: Intel I9 12900H**
- **GPU: Nvidia RTX 3060**

## • Method 2:

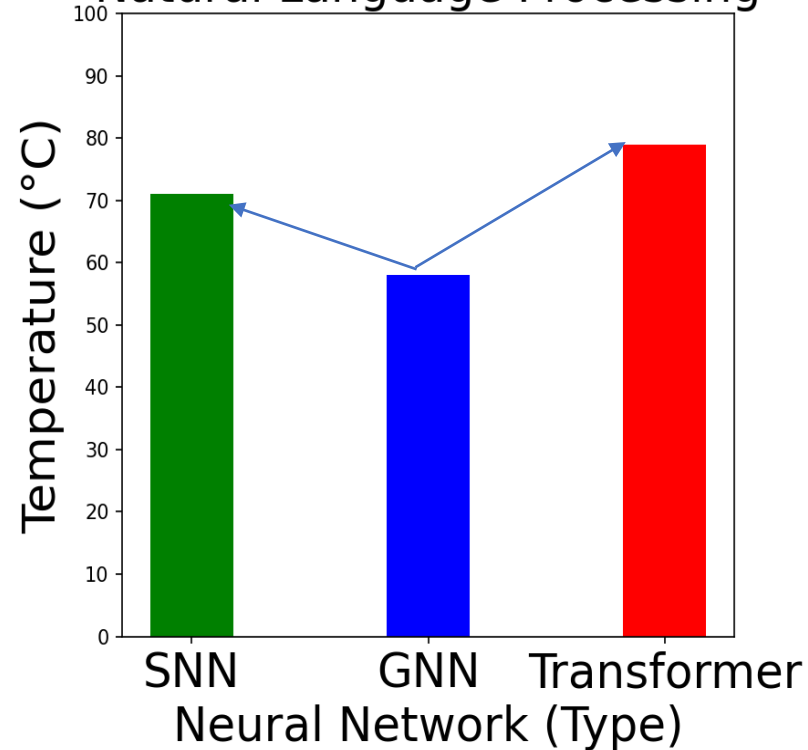
- It is a software tool which uses Intel's **Running Average Power Limit** application to estimate energy.
- It computes **average power** instead of peek power to estimate energy
- It was used for **CNN, FFNN & SNN** and on CV and Timeseries dataset
- It being a software tool, it incurs the **overhead caused by other applications.**
- **CPU: AMD threadripper 5995wx**
- **GPU: Nvidia RTX 3090**

# Temperature Estimates (in GPU)

## Computer Vision

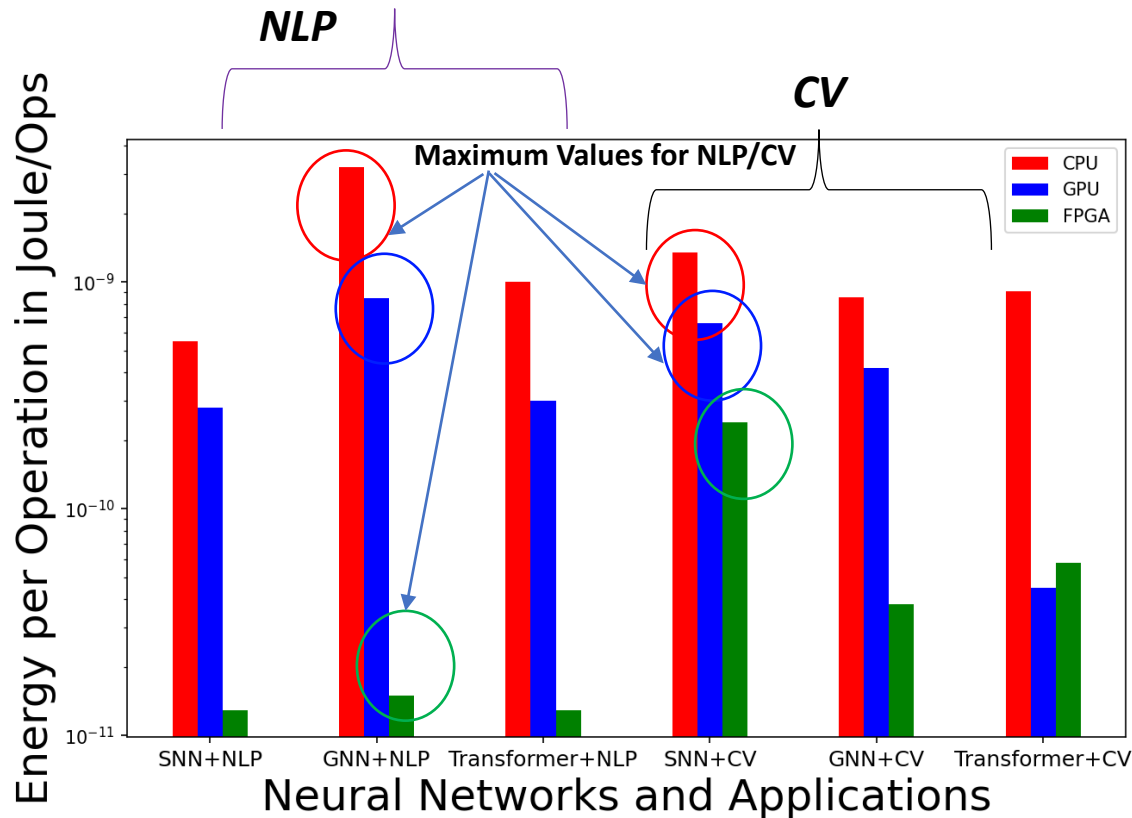


## Natural Language Processing



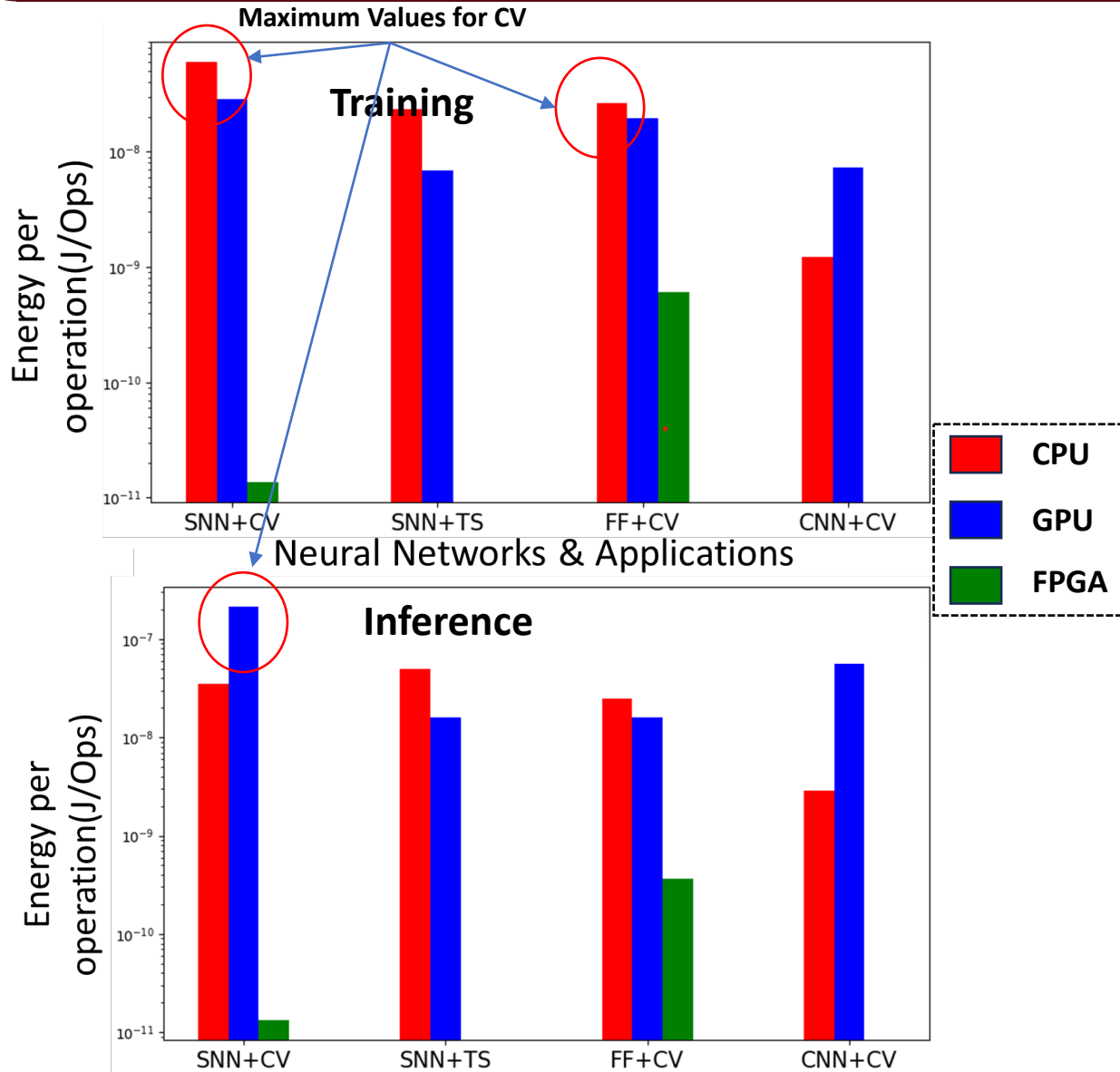
- Method 1 used to plot graphs present temperatures for three neural networks in **NLP** and **CV**, with color-coding from **blue** (lowest) to **red** (highest).
- Depending on the **Applications**, different NN architectures lead to difference in chip temperatures of  $\sim 20$  C (*between lowest and highest*)
  - Transformer appears to be most energy intensive
  - SNN appears to be more energy efficient

# Energy Estimates (Method 1)



- **FPGA** is ~76 times more efficient than CPU and 23 times more efficient than GPU (e.g. *Transformer for NLP*)
- **CPU** generally consumes more energy per operation than **GPU and FPGA** due to large computation time and memory access
- For **Natural Language Processing**:
  - GNN is more energy intensive for CPU-GPU-FPGA
  - FPGA is most energy efficient
- For **Computer Vision**:
  - Computer vision application consume **approx 1.5-2 times** more energy due to a larger number of operations

# Energy Estimates (Method 2)



- **CPU** generally consumes more energy per operation than **GPU and FPGA** due to large computation time and memory access
- **Training** more energy intensive than **Inference** for a given precision
- **Time series data tends** to consume more energy compared to **Computer Vision** application due to larger number of operations
  - Computer vision application consume **approx 10-15 times** less energy
- **Spiking neural networks** are less efficient on CPU and GPU
  - SNNs not optimized on GPU and CPU
  - SNNs requires parameter updates several times in one sample which demands more **communication between CPU and GPU**, reducing the **energy efficiency**

# Summary (1)

- **Two different models used to estimate energies across Applications-NN Architectures-HW Architectures**
  - Due to the different tools with different implementations
  - Will try to reconcile the differences
- **Transformer Models appear for hardware implementation the most energy intensive**
  - **1.15 times** higher than SNN and GNN
- **SNNs consumes more energy and produces least throughput on CPU and GPU**
  - SNNs needs weight update multiple times in one sample which increase the data movement between CPU and GPU which reduces the throughput of system.
  - SNNs requires **approx. 3-4 times more computation for image classification** application due to converting the static images to spike trains.

[1]Nurvitadhi, Eriko, et al. "Can FPGAs beat GPUs in accelerating next-generation deep neural networks?." *Proceedings of the 2017 ACM/SIGDA international symposium on field-programmable gate arrays*. 2017.

[2]Zhao, Bo, et al. "Feedforward categorization on AER motion events using cortex-like features in a spiking neural network." *IEEE transactions on neural networks and learning systems* 26.9 (2014): 1963-1978.

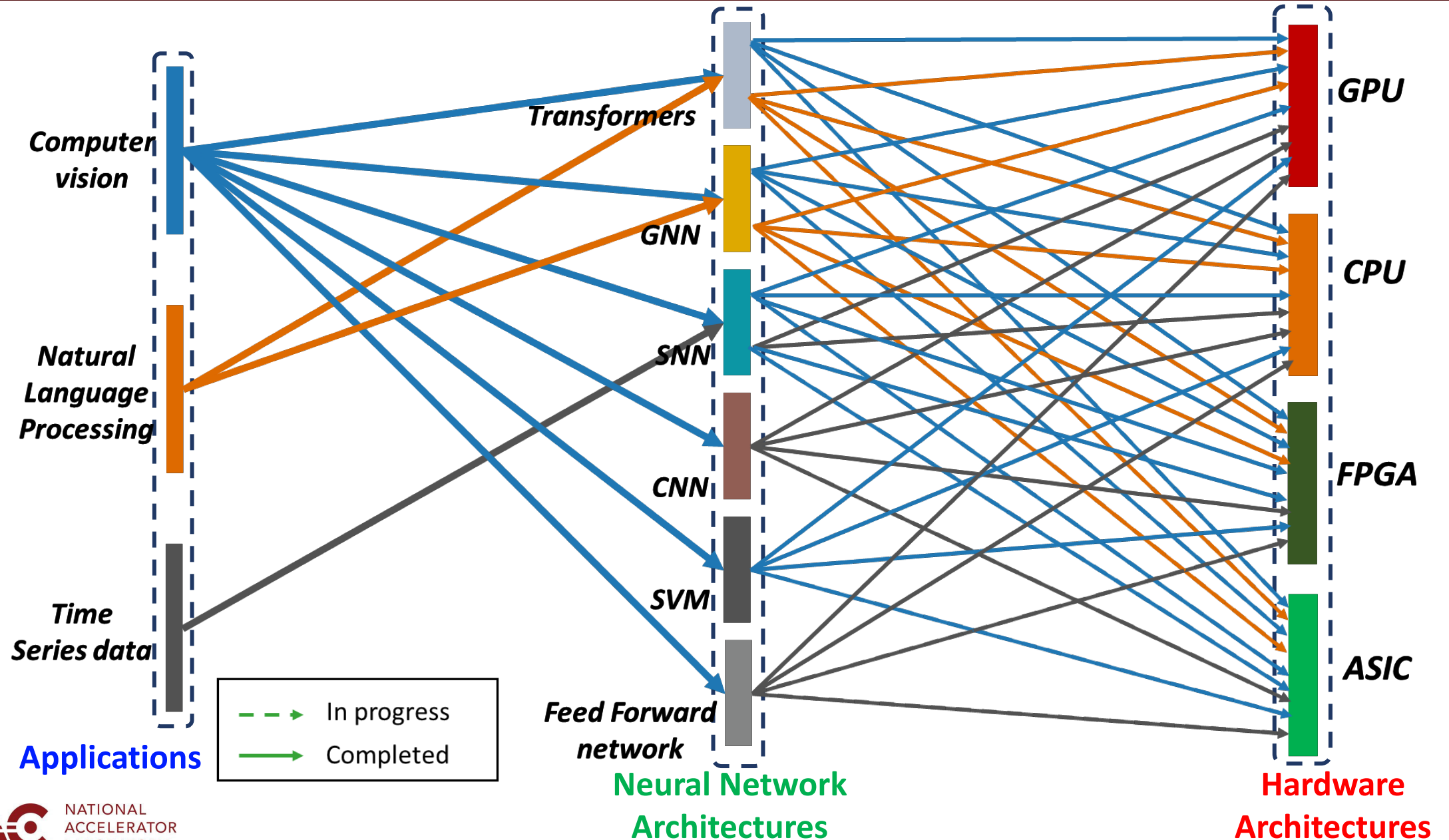
# Summary (2)

- **FPGA is more energy efficient than CPU and GPU on SNN for computer vision tasks**
  - FPGAs uses custom architecture to compute spiking neural networks
  - **Skip zero computation**[1], **Address-Event representations(AER)**[2] and **16-bit fixed point format** which boosts performance on FPGA
- **Generally, GPU has high throughput and energy per operation on CNNs compared to CPU due to its low latency, but is dependent on specific example**
  - Power consumption of GPU compared to CPU is **approx. ~1.5 times** but the **latency is 3 times lower** which consumes comparatively lower energy per operation (backup slides)
- **Continue Testing on other systems including Hybrid/Heterogeneous Systems**

[1]Nurvitadhi, Eriko, et al. "Can FPGAs beat GPUs in accelerating next-generation deep neural networks?." *Proceedings of the 2017 ACM/SIGDA international symposium on field-programmable gate arrays*. 2017.

[2]Zhao, Bo, et al. "Feedforward categorization on AER motion events using cortex-like features in a spiking neural network." *IEEE transactions on neural networks and learning systems* 26.9 (2014): 1963-1978.

# Ongoing Work







# **Analysis Tools & Future Work**

Energy Estimation, Workload Analysis & Existing tools

# Analysis Tools: *Review of Existing Energy Estimation Tools*

- **Marcher[1]:**
  - Measure the power of different components of a computer, such as **CPUs, DRAMs, disks, Coprocessor**
  - Overlooks power analysis of custom **accelerators** and applications specific design
- **HP CACTI:**
  - Uses Bottom-up approach for power estimation of memory components
  - Analyzes cache, memory access time, area, latency and dynamic power estimation tool
  - Lacks estimation of several other components and workload analysis of a system architecture
- **Accelergy[2]:**
  - Incorporates energy calculator based on the workload and the primitive component power table
  - It performs bottom-up approach and requires a simulator of the accelerator for workload analysis and component energy details
- **Watch[3] & GPUWattch:**
  - Framework to analyze microprocessor power dissipation at architecture level.
  - Limited to CPUs and GPUs for component power analysis.

**How to estimate energy consumption of any Neural Network on multiple hardware platforms?**

[1] Zong, Ziliang, Rong Ge, and Qijun Gu. "Marcher: A heterogeneous system supporting energy-aware high performance computing and big data analytics." *Big data research* 8 (2017): 27-38.

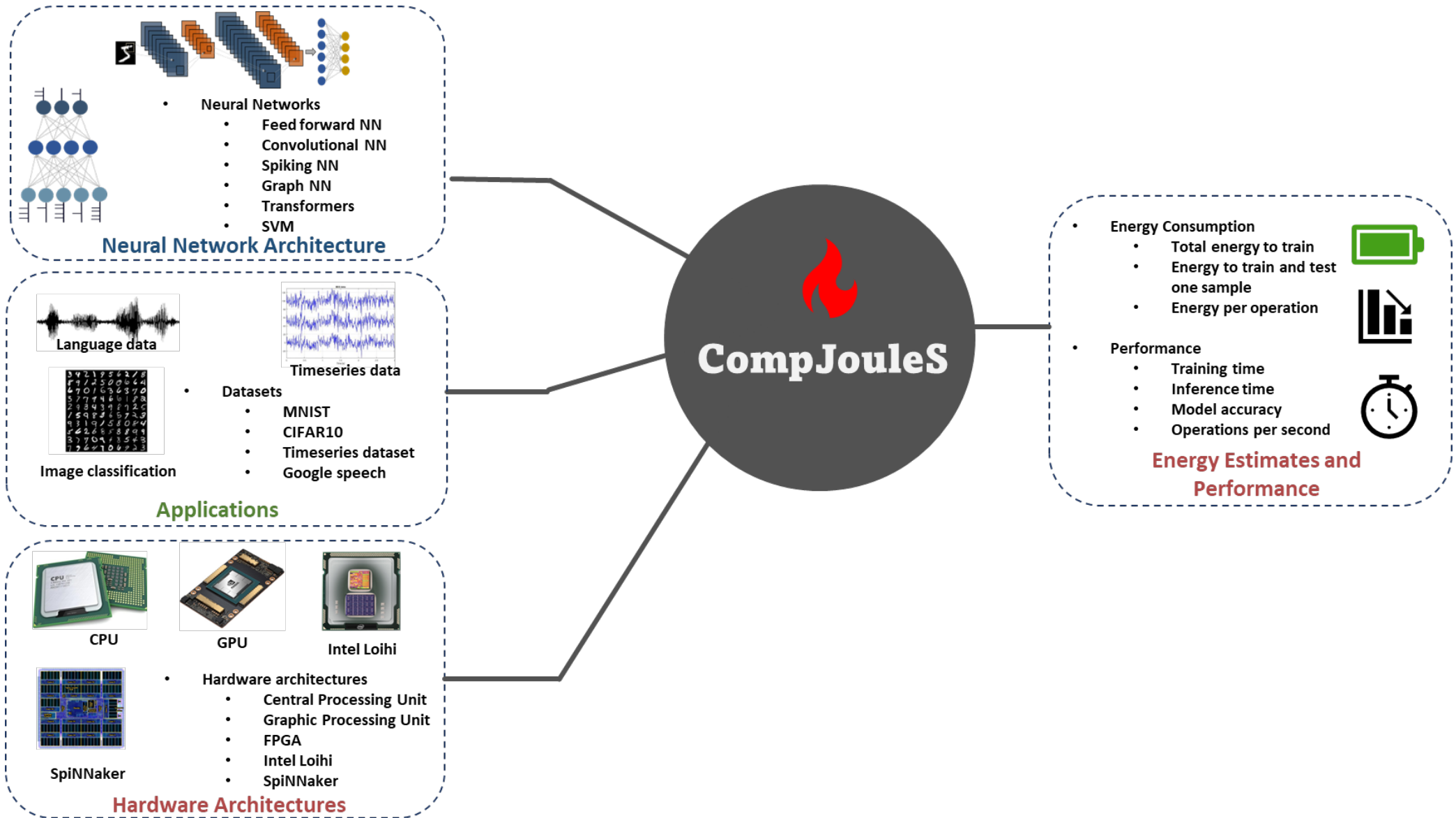
[2] Wu, Yannan Nellie, Joel S. Emer, and Vivienne Sze. "Accelergy: An architecture-level energy estimation methodology for accelerator designs." *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2019.

[3] Brooks, David, Vivek Tiwari, and Margaret Martonosi. "Wattch: A framework for architectural-level power analysis and optimizations." *ACM SIGARCH Computer Architecture News* 28.2 (2000): 83-94.

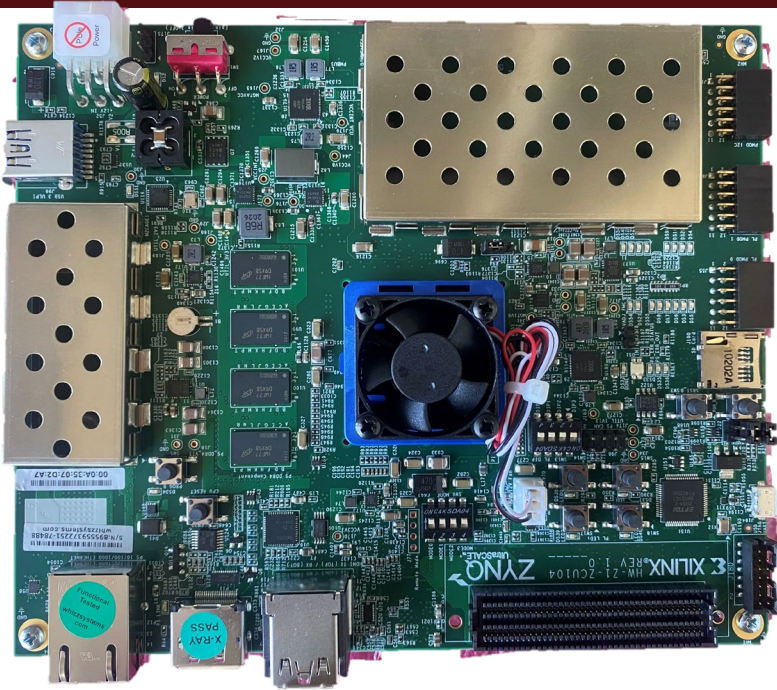
# Analysis Tools: *Need for Additional Capabilities*

- Current tools exist, but need work to be compared along the three vectors
- Flexible Tools for Energy Estimates across the different layers of computing:
  - *Applications-Neural Network Architectures-Hardware Architectures*
- Top-down and Bottom-up Analysis can identify bounds in energy estimates
- Once calibrated,
  - can enable identification of specific attributes that could be used for energy efficiency
  - can estimate energy of heterogeneous systems with different architectures, mixed precision computations etc....

# Analysis Tool: **CompJouleS**

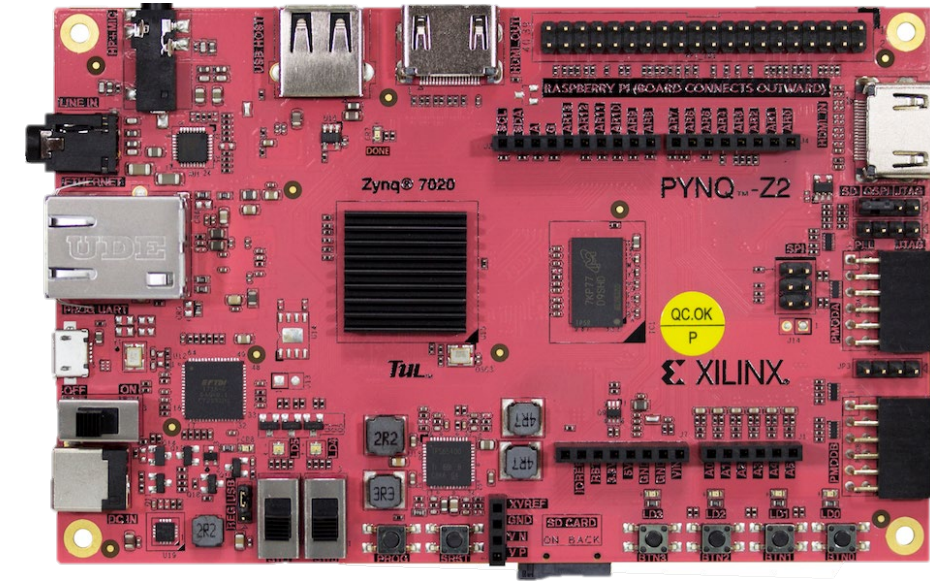
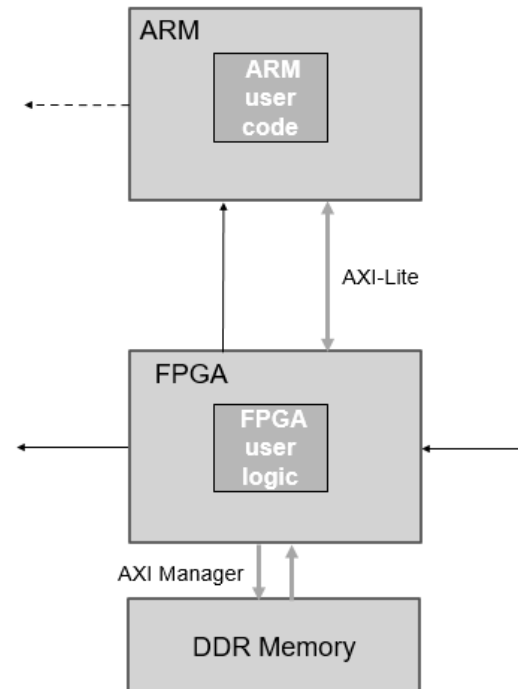


# Analysis Tool: Application to Hybrid architectures (Future work)



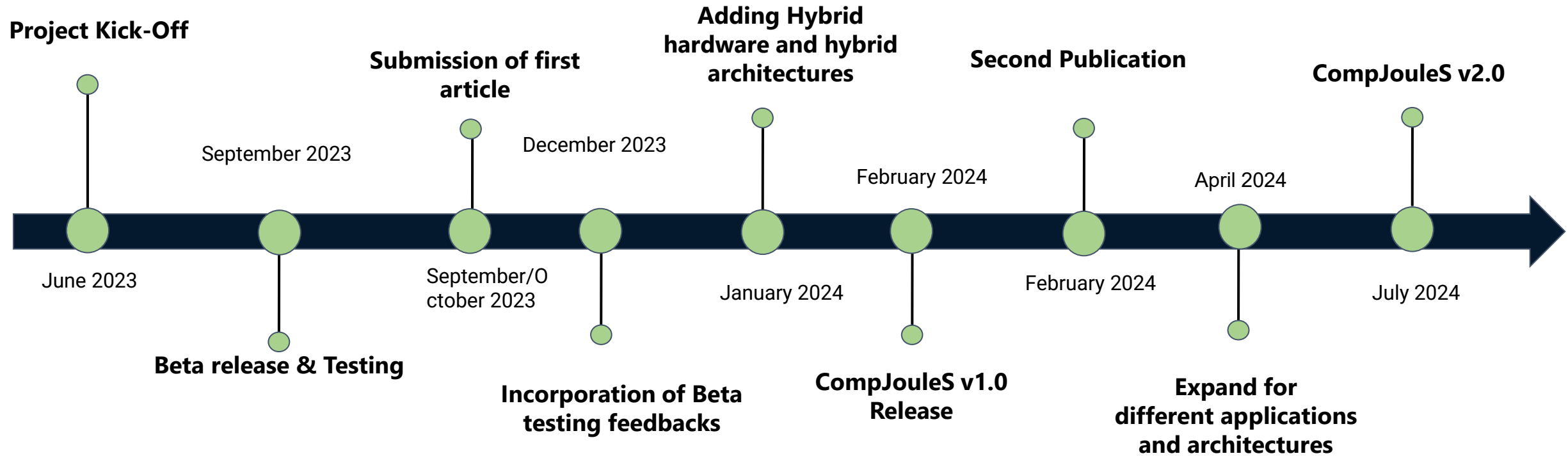
- AMD ZCU104 board
  - XCZU7EV-2FFVC1156 MPSoC
  - ARM cortex A53 processor
  - 16nm FinFET+ programmable logic
  - 504,000 LUTs

## System Architecture



- PYNQ Z1 board
  - ZYNQ XC7Z020-1CLG400C
  - 650MHz dual-core Cortex-A9 processor.
  - 28nm technology node
  - 13,300 LUTs

# TIMELINE: **CompJouleS**



# Thank You

[sshankar@slac.stanford.edu](mailto:sshankar@slac.stanford.edu)