The Best of Many Worlds: Scheduling Machine Learning Inference on CPU-GPU Integrated Architectures

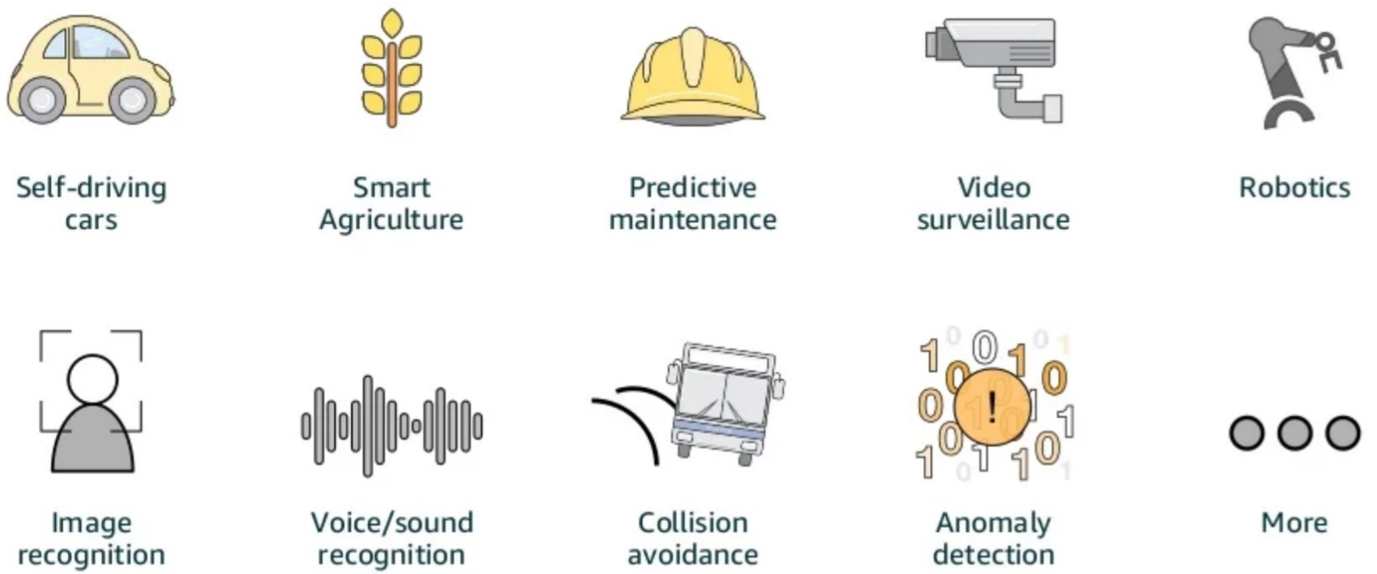
Giorgos Vasiliadis, Rafail Tsirbas, Sotiris Ioannidis

FORTH
INSTITUTE OF COMPUTER SCIENCE

HELLENIC MEDITERRANEAN UNIVERSITY

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

Use Cases



Self-driving cars

Smart Agriculture

Predictive maintenance

Video surveillance

Robotics

Image recognition

Voice/sound recognition

Collision avoidance

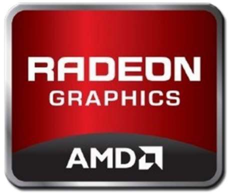
Anomaly detection

More

Commodity Processors

- Multi-core processors

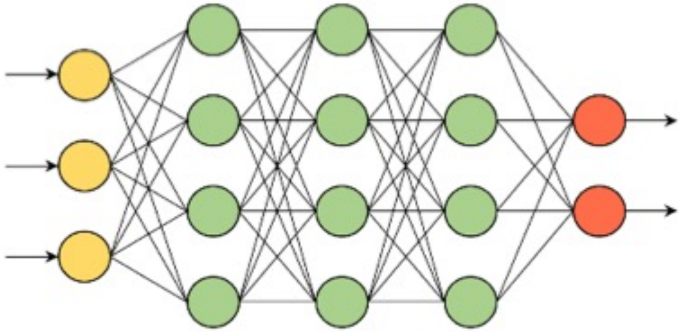
- Discrete accelerators



- System on Chip / Chip-integrated graphics units

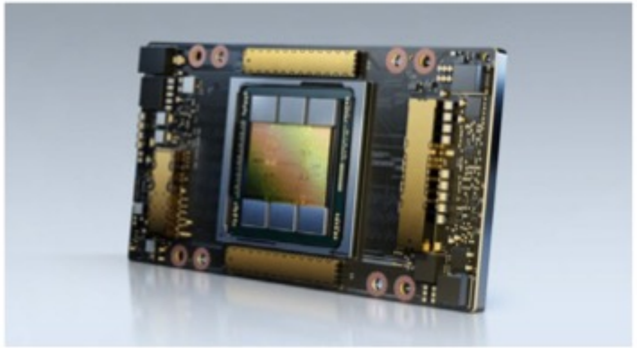
Motivation

- Programmers initial intuition when utilizing external accelerators



ML/DL Model

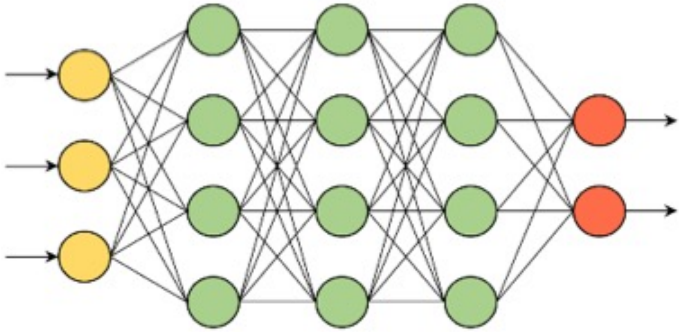
offload



Compute Device

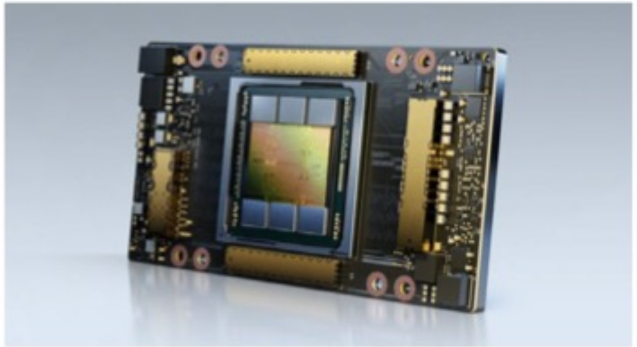
Motivation

- Programmers initial intuition when utilizing external accelerators



ML/DL Model

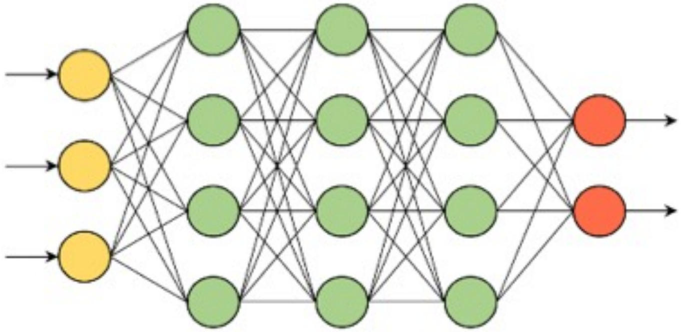
offload ??



Compute Device

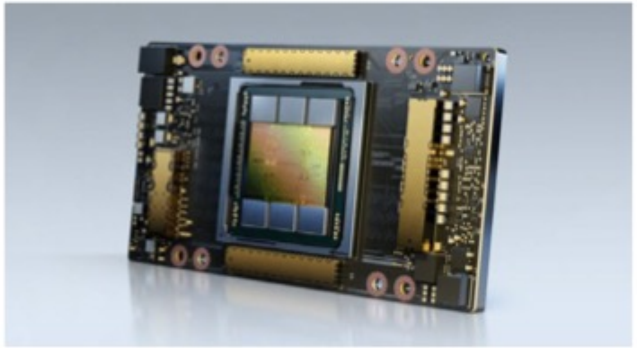
Motivation

- Programmers initial intuition when utilizing external accelerators



ML/DL Model

offload ??



Compute Device

- Performance fluctuations?
- Data variability?
- Data overloads?
- …

Performance Characterization

- **Workload:** Image classification on *three* different processors *

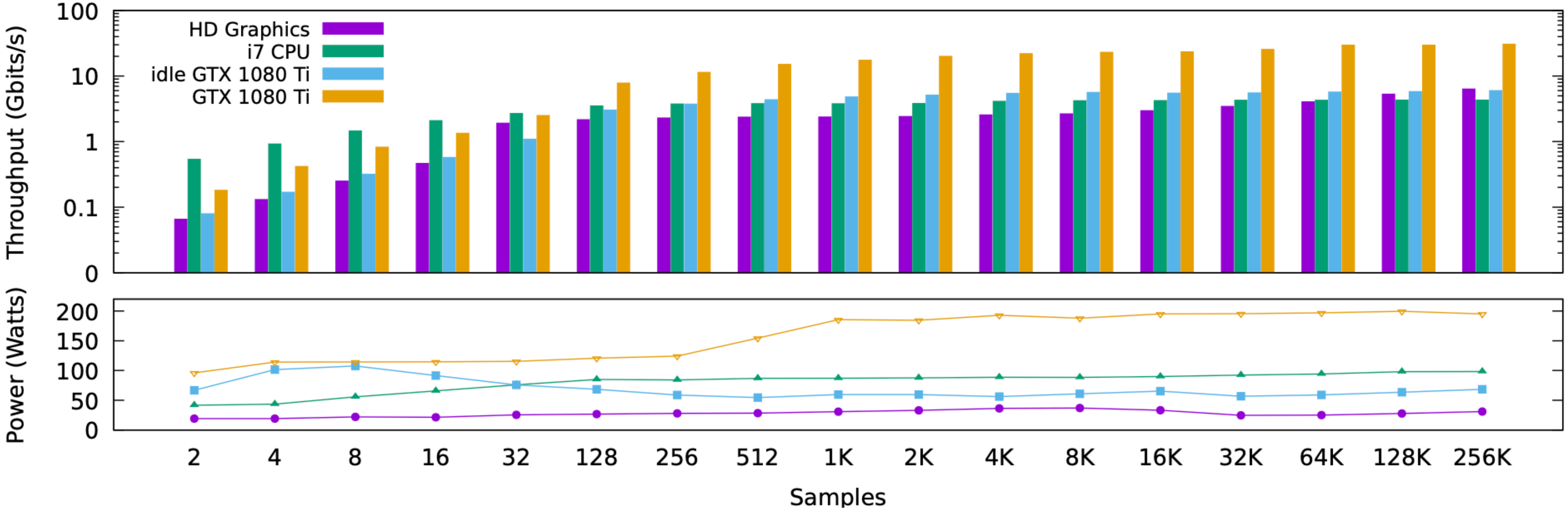
intel CORE i7 8th Gen | NVIDIA GEFORCE GTX 1080Ti | intel inside HD Graphics

- **Performance metrics:**
  1. Throughput
  2. Latency
  3. Power consumption

* Experiments performed on the MNIST dataset. More workloads and datasets are analyzed in the paper.

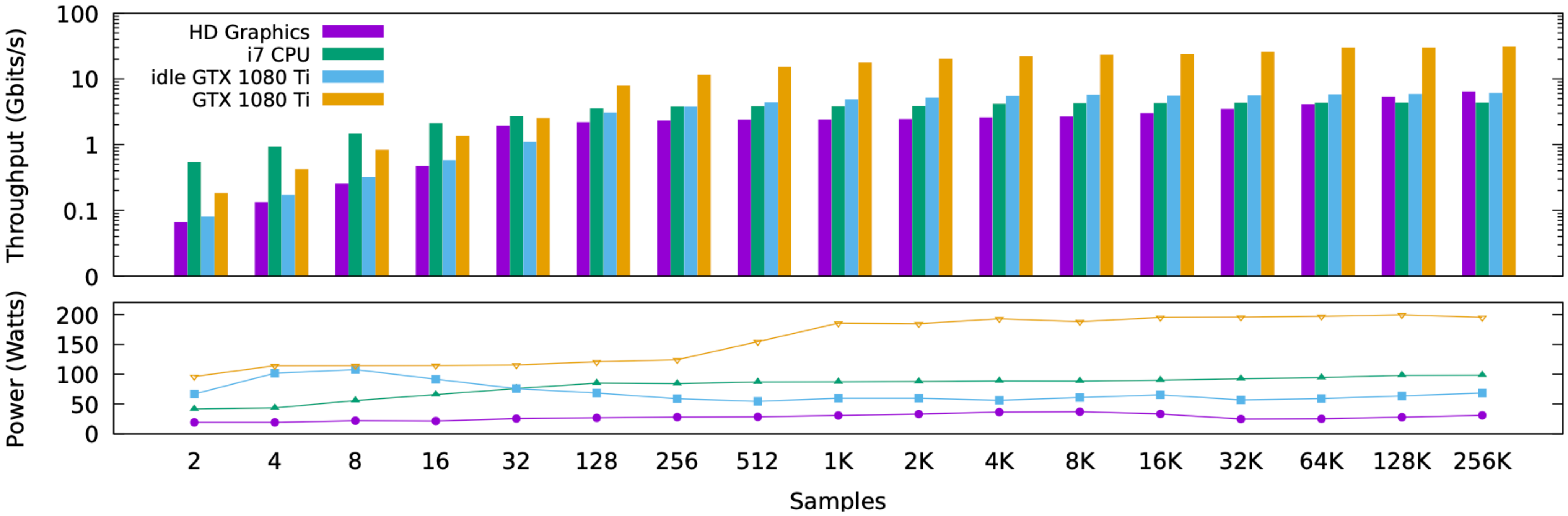
Performance Characterization

- **Workload:** Image classification on *three* different processors



Performance Characterization

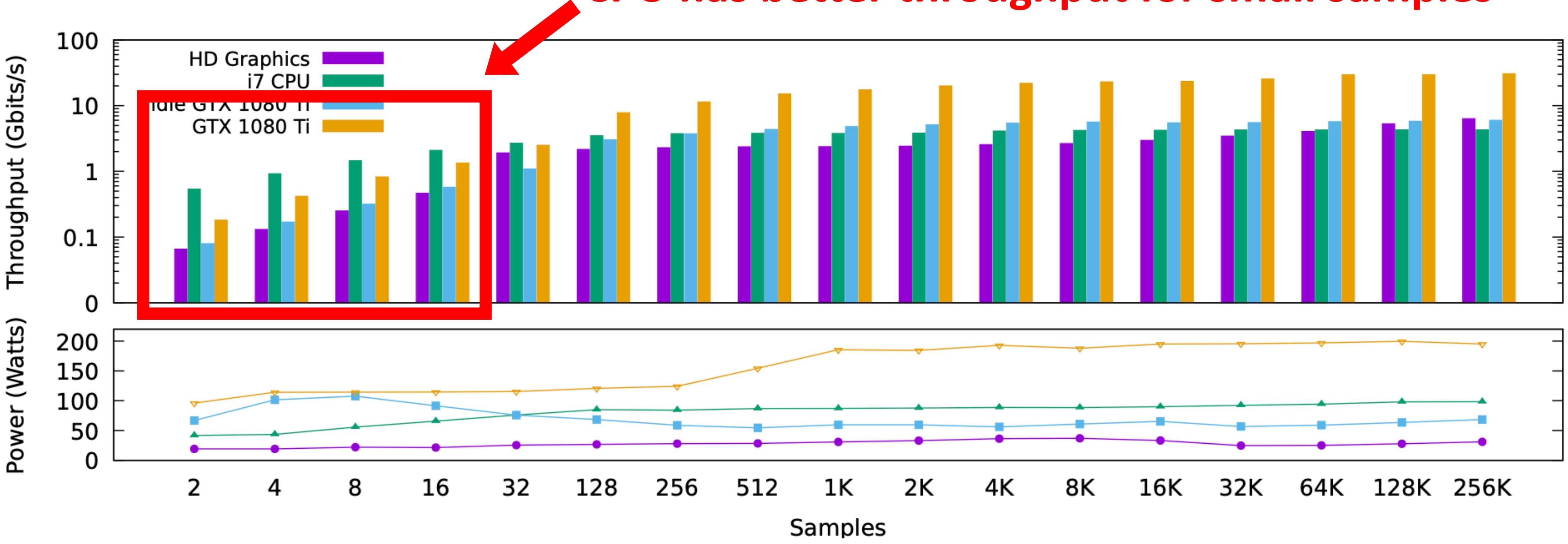
- **Workload:** Image classification on *three* different processors



Performance Characterization

- **Workload:** Image classification on *three* different processors

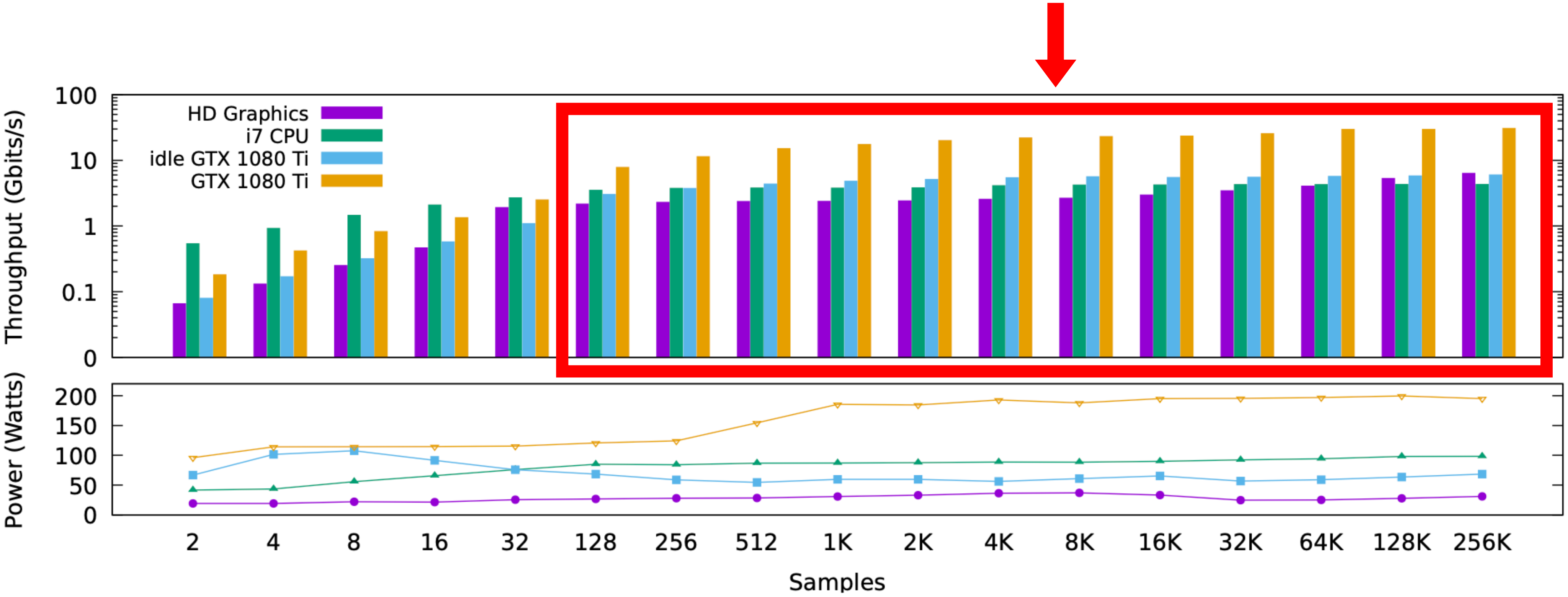
CPU has better throughput for small samples



Performance Characterization

- **Workload:** Image classification on *three* different processors

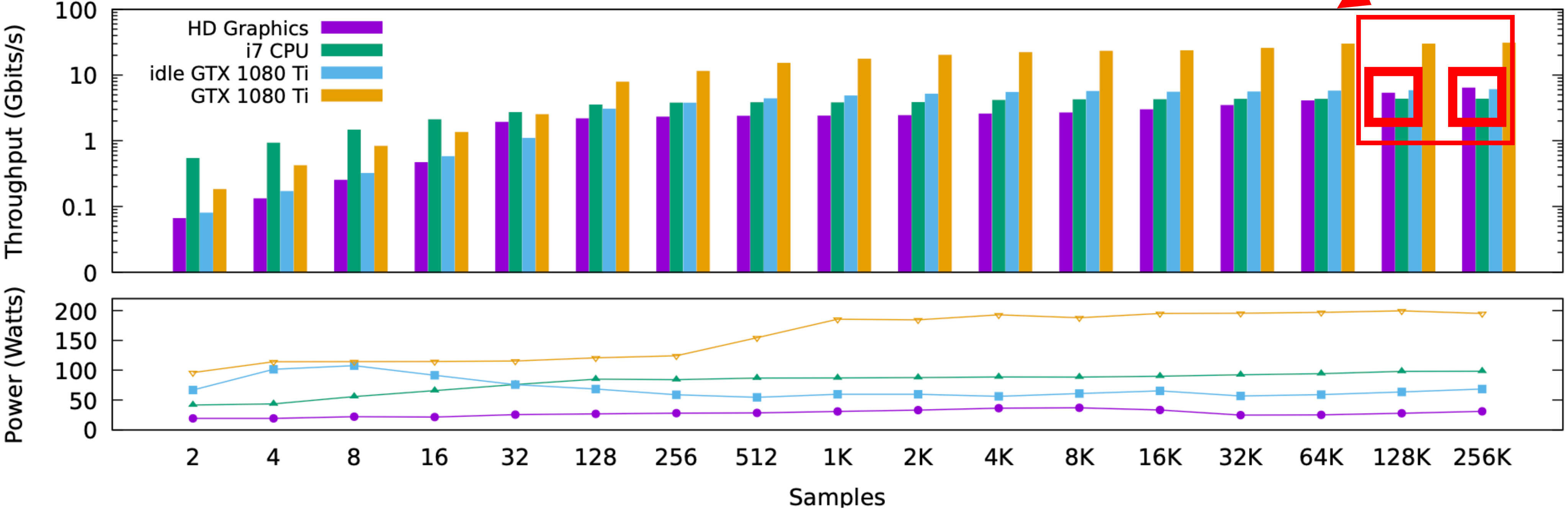
GPU is better for big samples



Performance Characterization

- **Workload:** Image classification on *three* different processors

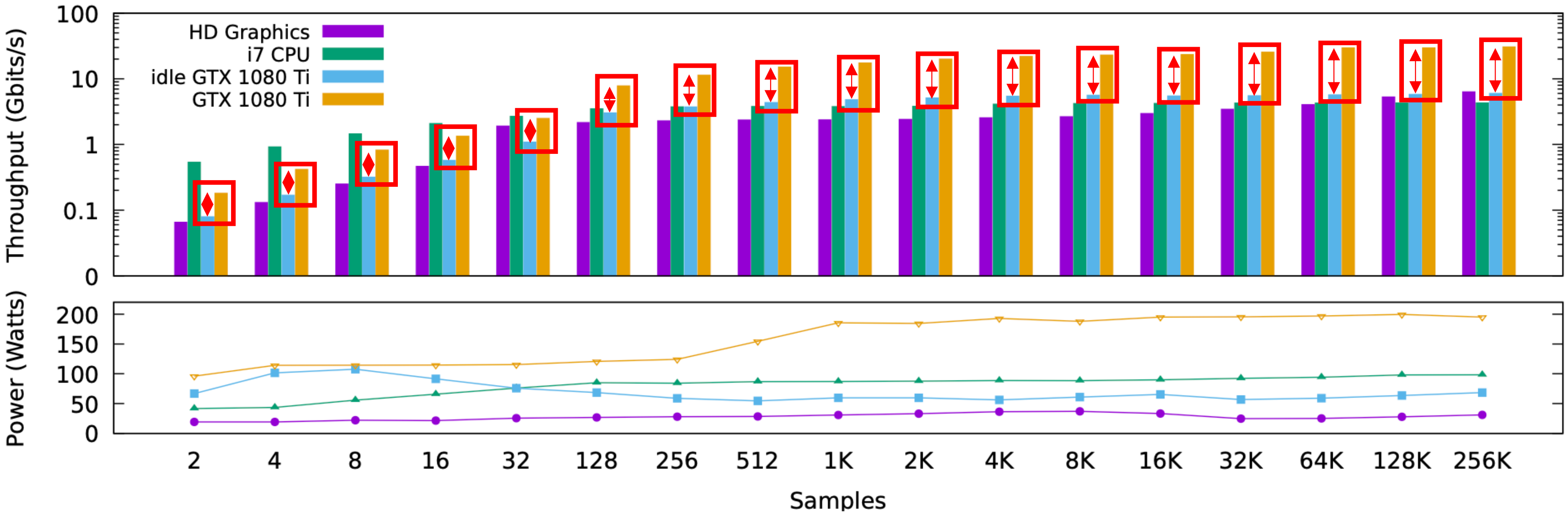
iGPU becomes better than CPU for very big samples



Performance Characterization

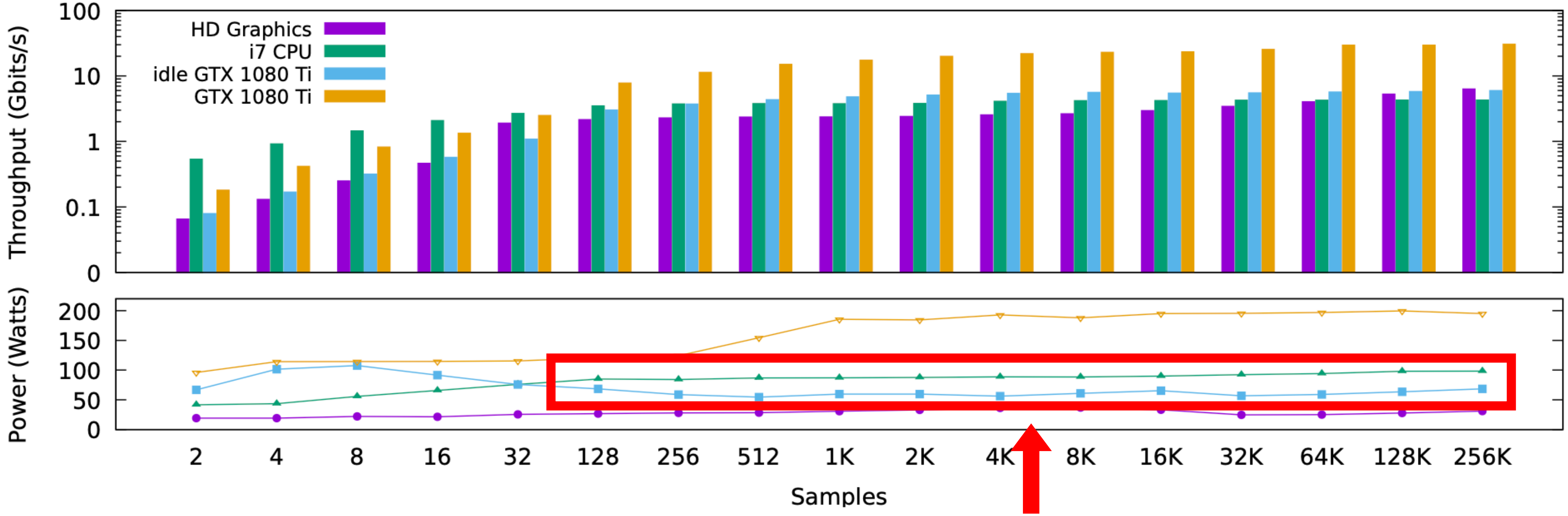
- **Workload:** Image classification on *three* different processors

GPU performance varies up to 7x times due to “power-saving” state



Performance Characterization

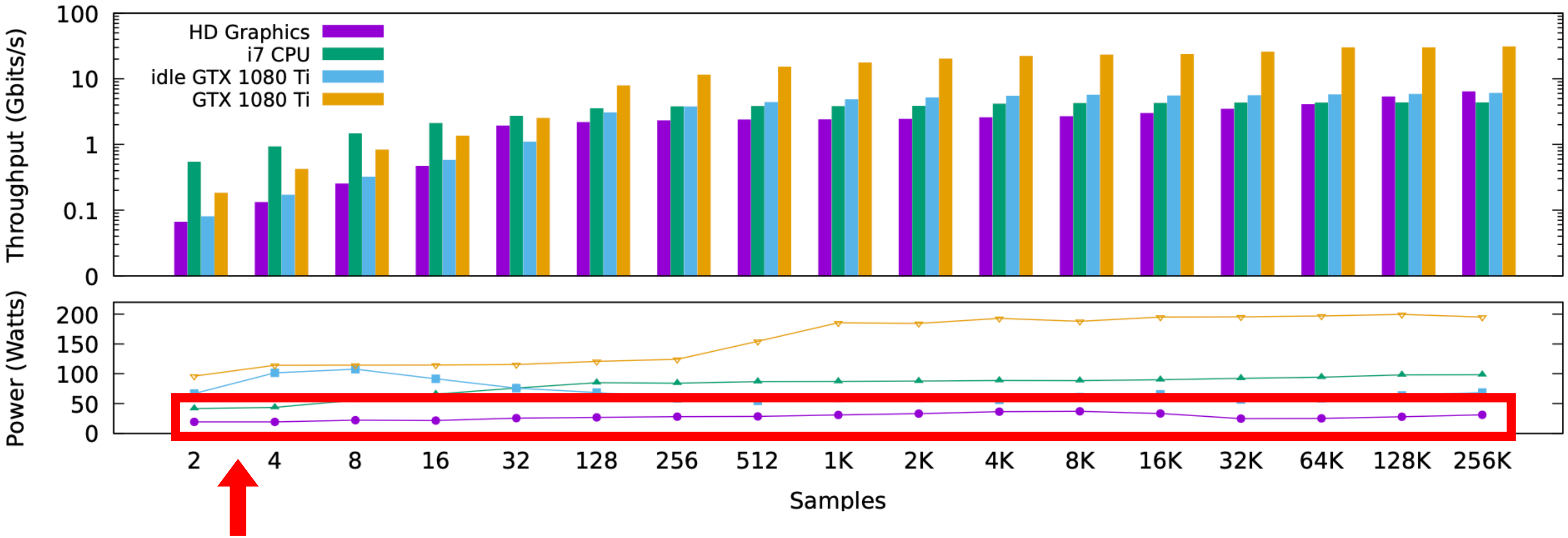
- **Workload:** Image classification on *three* different processors



GPU consumes less energy than CPU

Performance Characterization

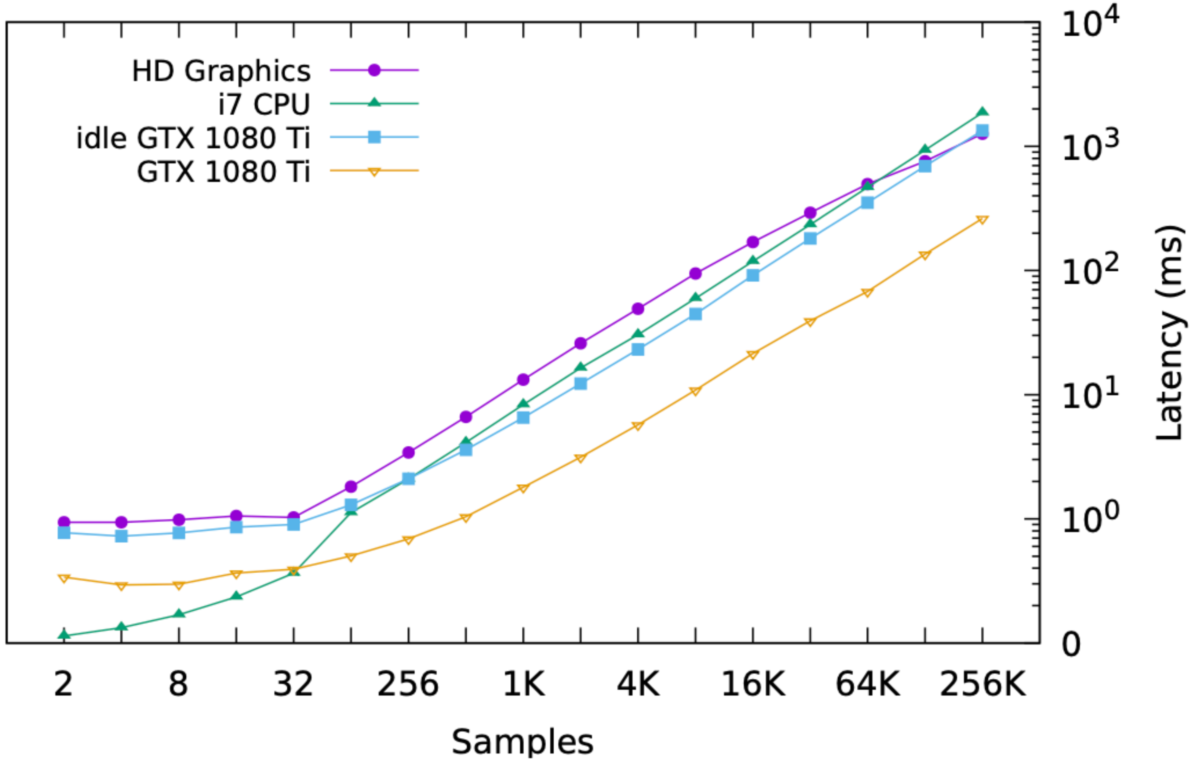
- **Workload:** Image classification on *three* different processors



iGPU consumes less energy in every case

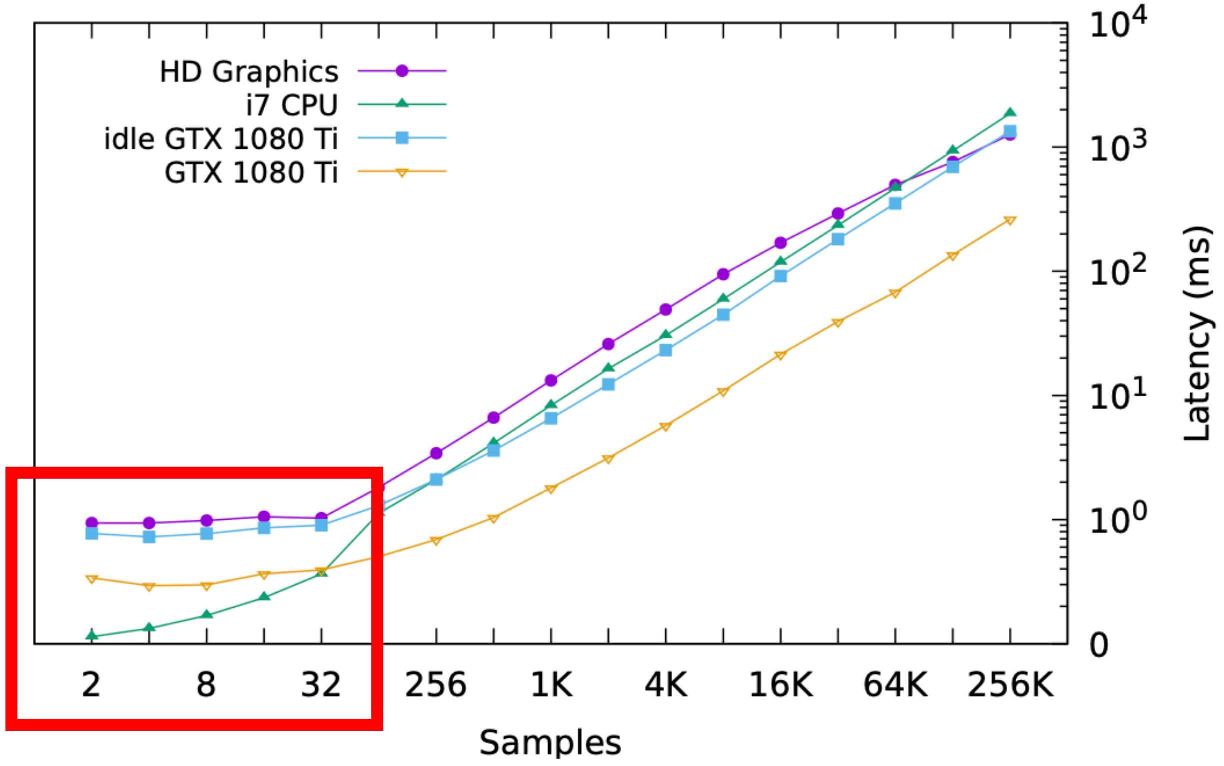
Performance Characterization

- **Workload:** Image classification on *three* different processors



Performance Characterization

- **Workload:** Image classification on *three* different processors



CPU provides the best latency

No single configuration is good for all

- Workload – Performance variability
  - Size of samples (Batch size)
  - Computational characteristics (i.e., structure) of ML model
- Hardware characteristics
  - GPU: High throughput comes with high latency
  - CPU: Low latency and good throughput
  - iGPU: Energy efficient and good throughput
- Harware state
  - Power saver states overthrow things:
    - e.g., GPU becomes more energy efficient than CPU

Search Space is Huge…

- Which device?
- How many samples?
- How many work groups / threads?
- How to partition datasets / workload?
- What memory to use?
- Power saver idle state?
- …

Choosing the right configuration

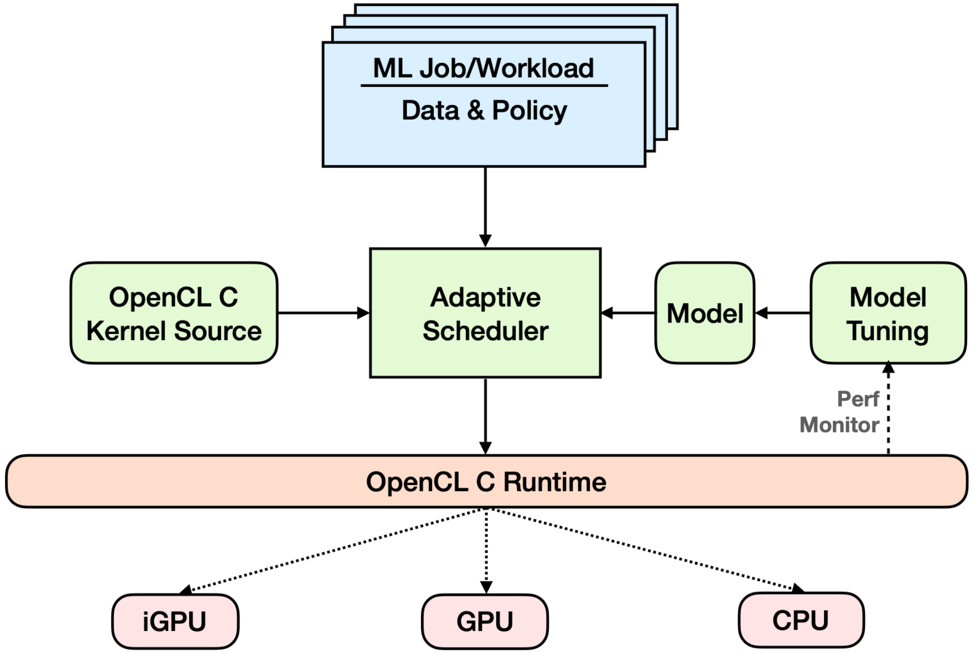
Hard to find the best choice manually

Need adaptive mechanisms to automatically select the most efficient processing device available

Adaptive Scheduling

- The scheduler is based on machine learning to make decisions
- Our aim is to train a model that would be able to learn and predict the appropriate device on which a classification model will run

- Online Tuning
  - Measure performance continuously
  - Update/tune model



Evaluation and Conclusions

- Our proposed scheduler is able to predict the appropriate device with an **accuracy of 92.5%**, while consuming up to **10% less energy**

- Adaptive schedulers is a promising solution to tackle performance variability

- Our proposed scheduler is able to utilize ***efficiently*** the computational capacity of its resources ***on demand***:
    - respond to relative performance changes
    - improve the energy efficiency