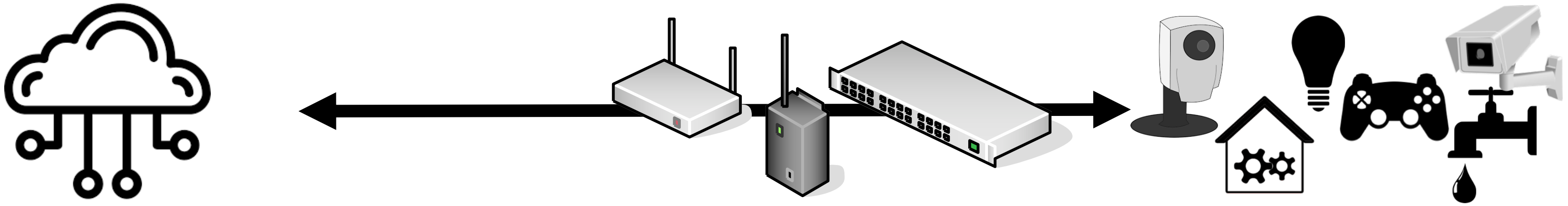
European Workshop on On-Board Data Processing
OBDP2019, 25-27/Feb/2019, ESA-ESTEC

# Using Heterogeneous Computing on GPU Accelerated Systems to Advance On-Board Data Processing

Nandinbaatar Tsog*, Mikael Sjödin*, Fredrik Bruhn*^

* Mälardalen University, Sweden
^ UNIBAP Publ. AB, Sweden

Dr. Harris Gasparakis - An AMD GPGPU, Computer Vision and Machine Learning technical expert and project manager, USA
Dr. Moris Behnam - Associate Professor, Mälardalen University, Sweden
Dr. Matthias Becker - Postdoc Researcher, KTH Royal Institute of Technology, Sweden



Cloud computing

Edge/Fog computing

IoT devices

Real-time properties:

- Edge/Fog computing

⇳

- Intelligent/Advanced On-Board Processing

Heterogeneous architectures & computing

# Outline

- Heterogeneous Processors in Space
  - Real-time Systems
  - Heterogeneous System Architecture
- Understanding of Heterogeneous Computing
  - Heterogeneous Segment
- In-Orbit Advanced Applications
  - MIOpen, AlexNet with Tensorflow, Hashcat
- Experiments & Results
- Conclusion
- Reference

# Real-Time Systems

- Timing constraints

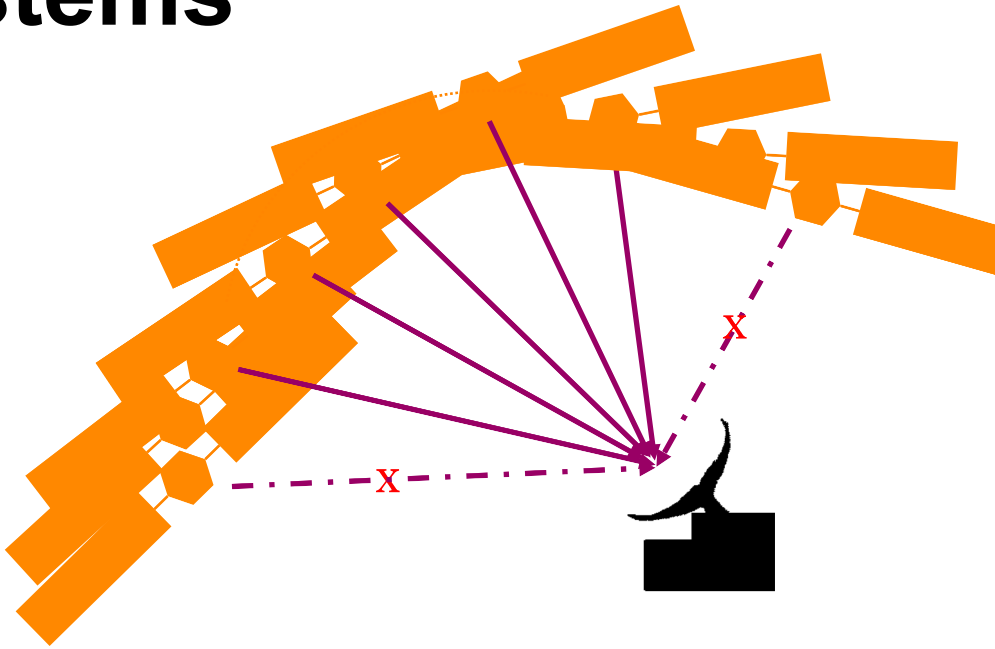
Deadline

- Worst-Case Scenarios

- Image processing
  - Video frame rate
    - 60fps

      17ms
    - 20fps

      50ms



# Heterogeneous Processors

- CPU + FPGA

How to access to the memory

Data consistency!

Communication latency!

- Several techniques / methods
  - Pipeline
  - Pinned Memory
  - Asynchronous Transfers
  - Persistent kernel/thread

Heterogeneous System Architecture (HSA)

- GPU
  - Embedded in SoC
  - Integrated GPU or Accelerated Processing Unit (APU)

Radiation?

- GIMME3
  - Invented at Mälardalen University and Unibap
  - Heterogeneous System Architecture (HSA) compliant GPU with FPGA

HyTI - Hyperspectral Thermal Imager (NASA)

GIMME 4/e22xx families by Unibap

*Ref 1. Tsog et al.*

# Heterogeneous System Architecture (HSA)

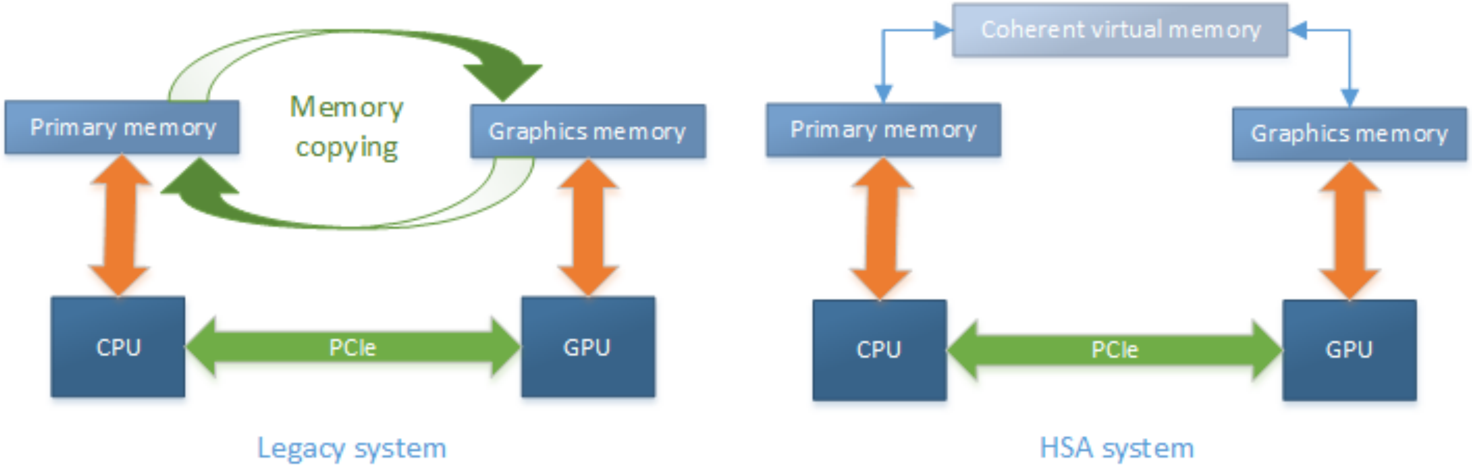
- HSA Foundation – Founders are AMD, ARM, Imagination, MediaTek, Qualcomm and Samsung.

- Challenges / Features of HSA
  - Memory handling
  - Queuing
  - Instruction Set Architecture

HSA includes/simplifies the techniques
Pinned memory, pipeline etc.

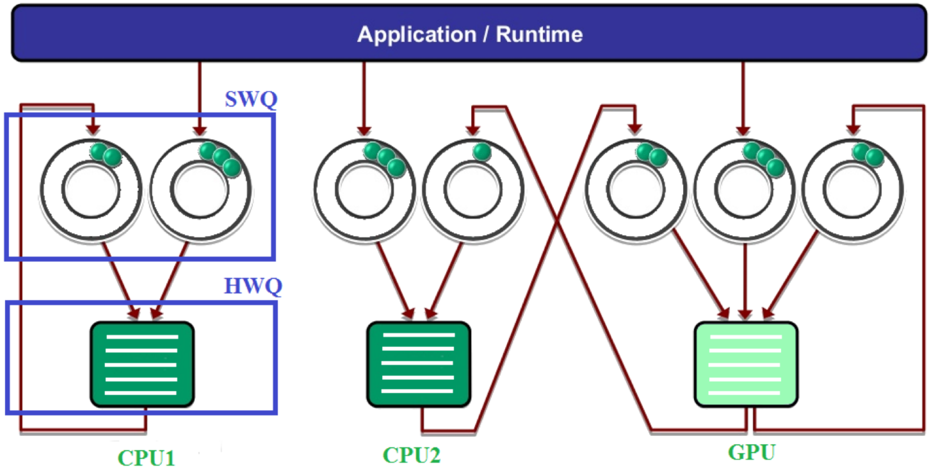
# No memory copy in HSA

- No memory copying between memories of Compute Units



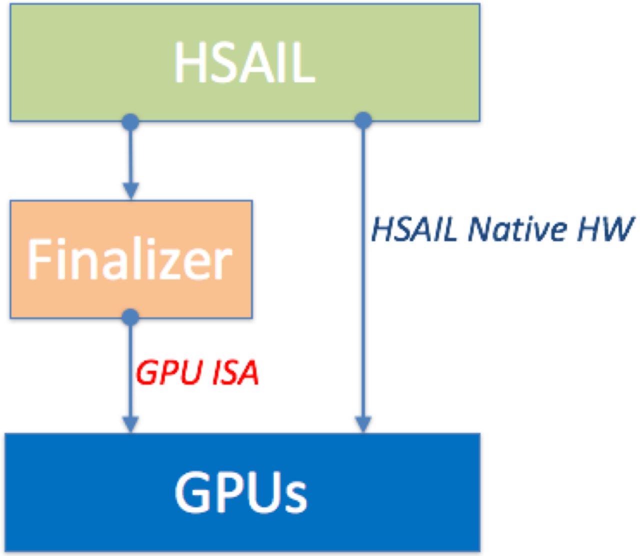
# Queuing in HSA

- Hardware queue structure in a HSA system



# Instruction Set Architecture in HSA

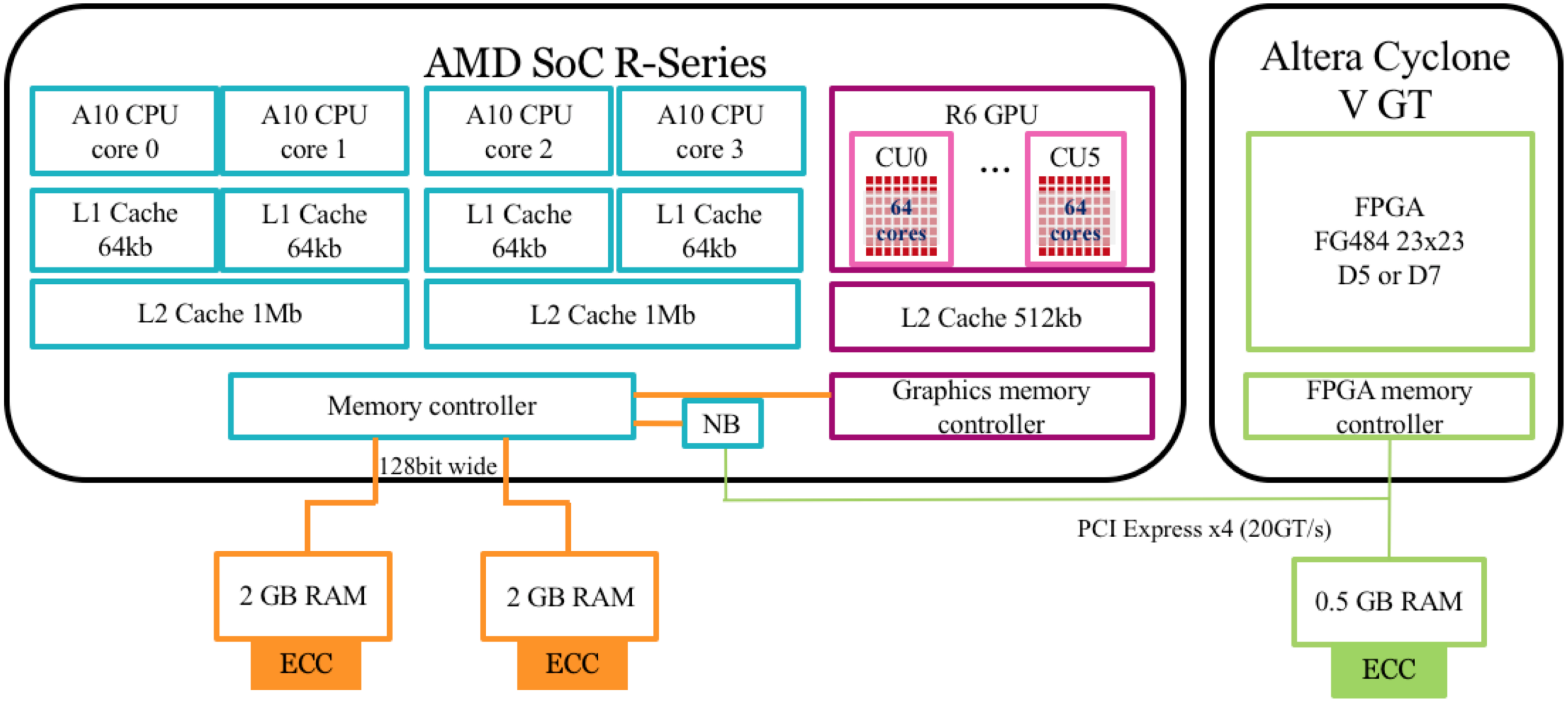
- Instruction Set Architecture
  - HSA Intermediate Language (HSAIL)
    - A low-level intermediate representation
    - Vendor- and ISA-independent
    - Generated by high-level compiler
  - Finalizer
    - To translate HSAIL code into appropriate machine code (ISA)
    - Used for the HW component which does not support HSAIL natively



# Advantages of HSA

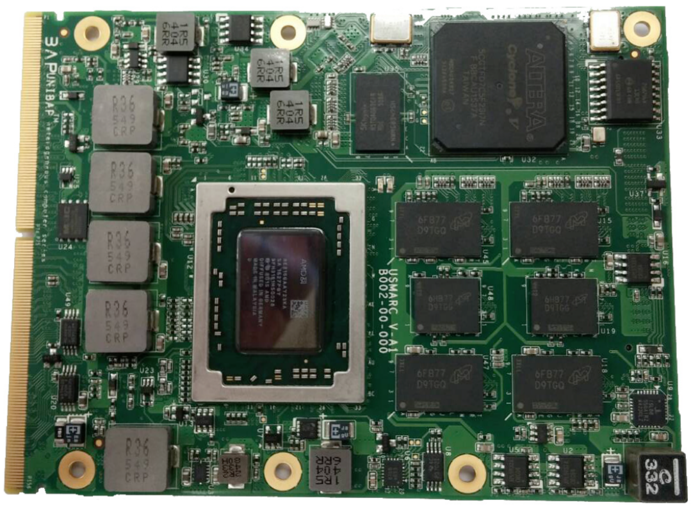
- Compilers
  - HCC based on LLVM/Clang
  - GNU 7 or later
- Drivers
  - Open-source and proprietary source drivers
  - ROCm, amdgpu-pro/radeon, Mesa, Catalyst
- Libraries
  - Machine Intelligent = MIOpen
  - OpenVX, OpenCV
  - Caffe, Tensorflow
  - Vulkan

# Architecture of GIMME4 Platform

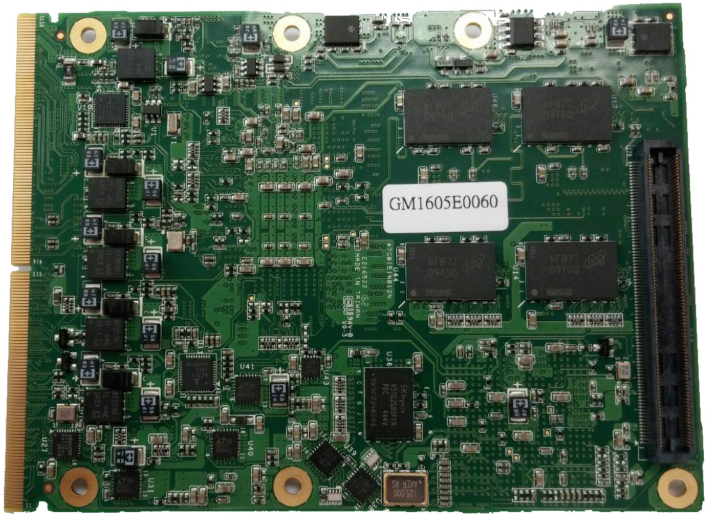


*Ref 2. Tsog et al.*

# Platform



The top side Unibap e2250 prototype module based on the GIMME-4 architecture featuring an AMD R-series SoC, dual DDR4 memory banks with ECC, and Intel Altera Cyclone V FPGA.



Photograph of the bottom side Unibap e2250 prototype module showing on the right the expansion connector with 180 IO for additional features.

### GIMME4 platform with A10-8700p APU

- 85g
- 82 mm x 110mm
- 12-35Watt (TDP – 15Watt)

### A10-8700p APU

- 28nm

- CPU: (4 cores, 1.8GHz)
- GPU: (6/8 CUs, 800MHz
  384/512 shaders
  533/819 GFLOPS)

### Bus bandwidth

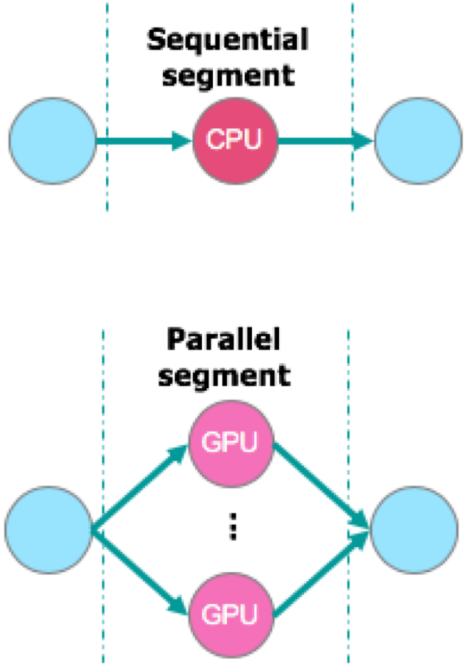
- Between CPU and GPU

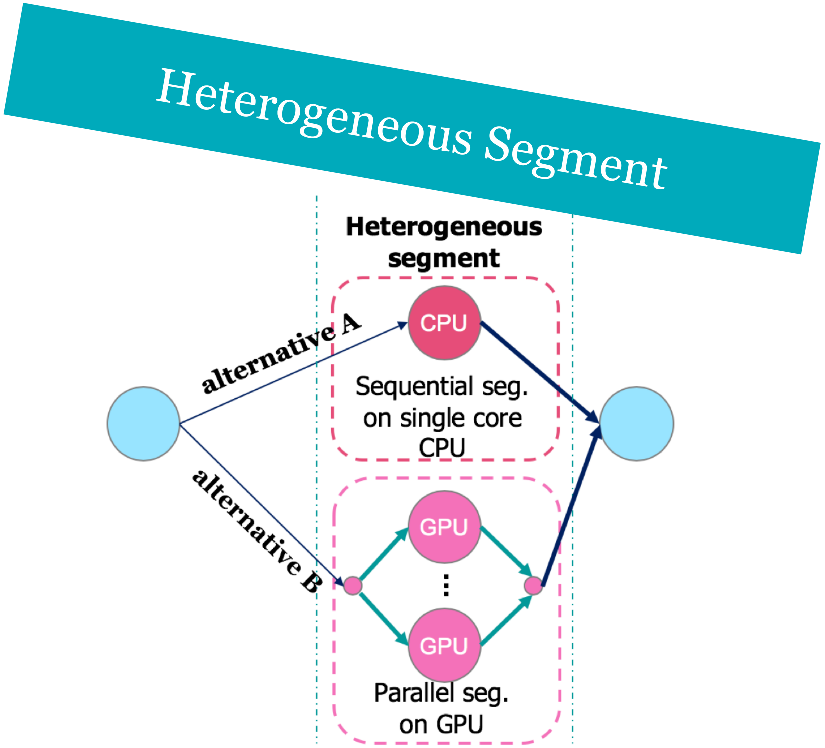
  At least 100GBps communication between CPU and GPU caches (128bit wide)
- Between APU and FPGA PCI Express x4 (20GT/s)

# Heterogeneous Computing

- 28nm -> 7nm (AMD) -> 5nm (Apple)
- Parallelism
- Use of
  - multiple numbers of processing units
  - different processing units

# Heterogeneous segment

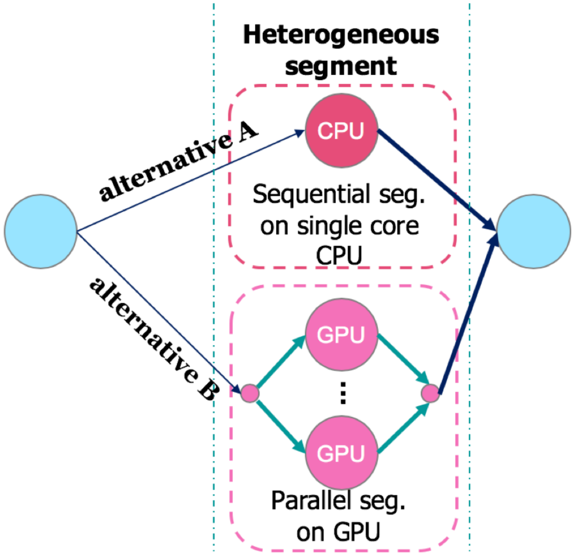




*Ref 3. Tsog et al.*

# Heterogeneous segment

- OpenMP
  - A host device & target devices
  - Implicitly on host (target device is not able)
- OpenCL
  - A host processor & accelerators
  - Explicitly using
    - *clCreateContextFromType + if condition*
- CUDA
  - A host (CPU) & devices (NVIDIA's GPU)
  - Explicitly using 3 qualifiers/space-specifiers
    - __global__, __device__, __host__
- C++AMP
  - A host & accelerators
  - Implicitly



Technology development

# In-Orbit Advanced Applications

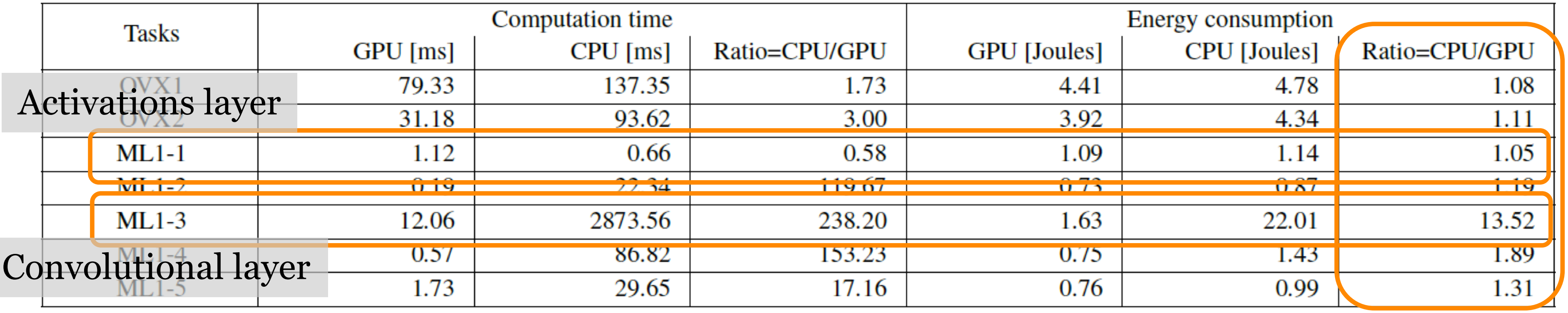
- MIOpen – Convolutional Neural Network acceleration
  - An alternative to Nvidia’s CuDNN;
  - Supported different layers:
    - Activations, Batch Normalization, CNN, RNN, Local Response Normalization, Pooling, Softmax
- AlexNet with Tensorflow
  - The key role to bring Deep Learning era
- Computer Vision applications
  - Combination of Optical Flow and Harris feature detection alg.
- HashCat

# Experiments

- Exp A
  - An investigation of the computational performance and power consumption in CPU and GPU
  - Activations (ML1-1), Batch Normalization (ML1-2), CNN (ML1-3), LR Normalization (ML1-4), Pooling (ML1-5) and Combination of Optical Flow and Harris Feature Detection Algorithm (OVX1,2)
- Exp B
  - "Balanced Use" of CPU and GPU using Heterogeneous Segment idea
- Exp C
  - Heterogeneous Computing of AlexNet and Harris Edge Detector Application

# Evaluation

- Exp A



| Tasks | Computation time | | | Energy consumption | | |
|---|---|---|---|---|---|---|
| | GPU [ms] | CPU [ms] | Ratio=CPU/GPU | GPU [Joules] | CPU [Joules] | Ratio=CPU/GPU |
| OVX1 | 79.33 | 137.35 | 1.73 | 4.41 | 4.78 | 1.08 |
| OVX2 | 31.18 | 93.62 | 3.00 | 3.92 | 4.34 | 1.11 |
| ML1-1 | 1.12 | 0.66 | 0.58 | 1.09 | 1.14 | 1.05 |
| ML1-2 | 0.19 | 22.34 | 119.67 | 0.73 | 0.87 | 1.19 |
| ML1-3 | 12.06 | 2873.56 | 238.20 | 1.63 | 22.01 | 13.52 |
| ML1-4 | 0.57 | 86.82 | 153.23 | 0.75 | 1.43 | 1.89 |
| ML1-5 | 1.73 | 29.65 | 17.16 | 0.76 | 0.99 | 1.31 |

Activations layer
Convolutional layer

Speed up ratio up to 238 times (Conv. layer)

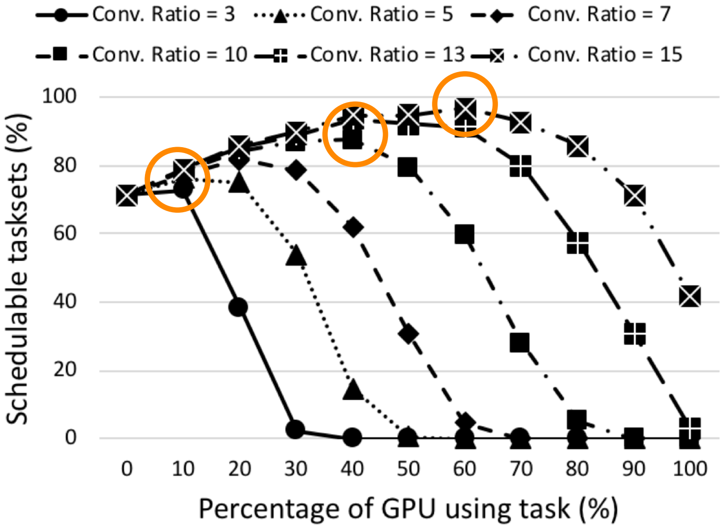
GPU consumes less energy than CPU

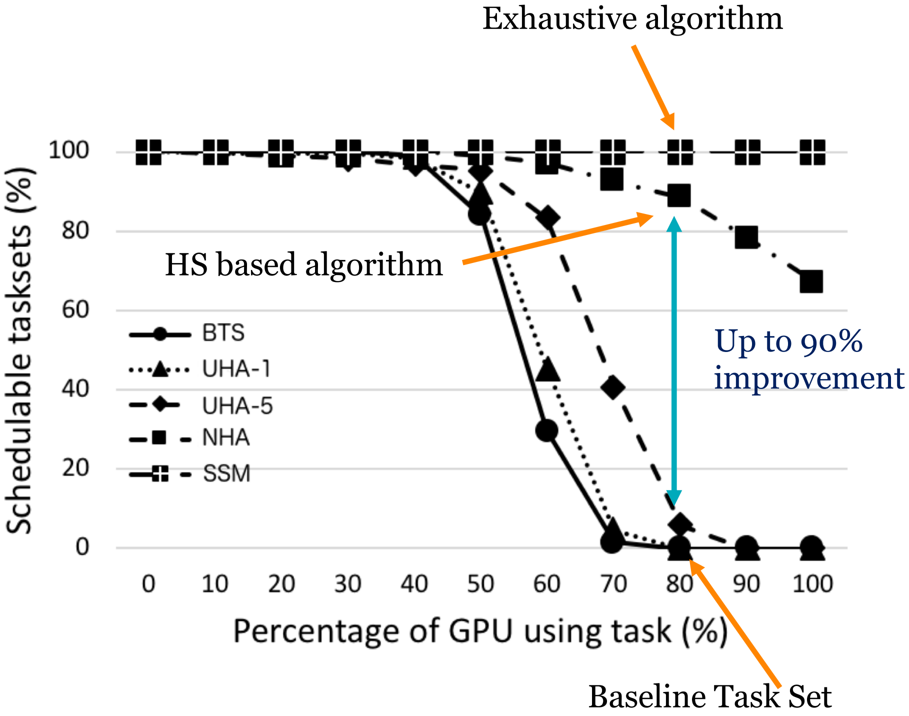
GPU consumes 13.52 times less energy than CPU (Conv. layer)

*Ref 1. Tsog et al.*

# Evaluation

- Exp B





Exhaustive algorithm

HS based algorithm

Up to 90% improvement

Baseline Task Set

*Ref 3. Tsog et al.*

# Evaluation

- Exp C

| Execution time [s] | AlexNet with TensorFlow | | Harris Edge Detector | |
|---|---|---|---|---|
| | Mean | WCRT | Mean | WCRT |
| Stand Alone | 7.875 | 8.036 | 1.649 | 1.87 |
| Together | 7.906 | 8.104 | 1.821 | 1.897 |

CPU-GPU communication

| Execution time [s] | AlexNet with TensorFlow | |
|---|---|---|
| | Mean | WCRT |
| Stand Alone | 12.355 | 12.366 |
| Together | **12.348** | 12.374 |

No data transfer loss

# Conclusion

- On-board processing of GPU embedded satellite
  - Consumes up to 13.52 times less energy and computes up to 238 times faster
- Using Heterogeneous Segment improves schedulability of tasksets up to 90%
- Heterogeneous computing performances well on GIMME4 platform

# References

- 1. N. Tsog, M. Behnam, M. Sjödin, and F. Bruhn. Intelligent data processing using in-orbit advanced algorithms on heterogeneous system architecture. In IEEE Aerospace Conference, pages 1–8, March 2018.

- 2. N. Tsog, M. Sjödin, and F. Bruhn. Advancing on-board big data processing using heterogeneous system architecture. In ESA/CNES 4S Symposium 4S, April 2018.

- 3. N. Tsog, M. Becker, F. Bruhn, M. Behnam, and M.Sjödin. Static Allocation of Parallel Tasks to Improve Schedulability in GPU Accelerated Real-Time Systems. In 31st Conference on Real-Time Systems (ECRTS'19). (Submitted)

# Thank you!

MÄLARDALEN UNIVERSITY
SWEDEN