Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

AIRBUS
DEFENCE & SPACE

Embedded GPU benchmarking for High-Performance On-board Data Processing

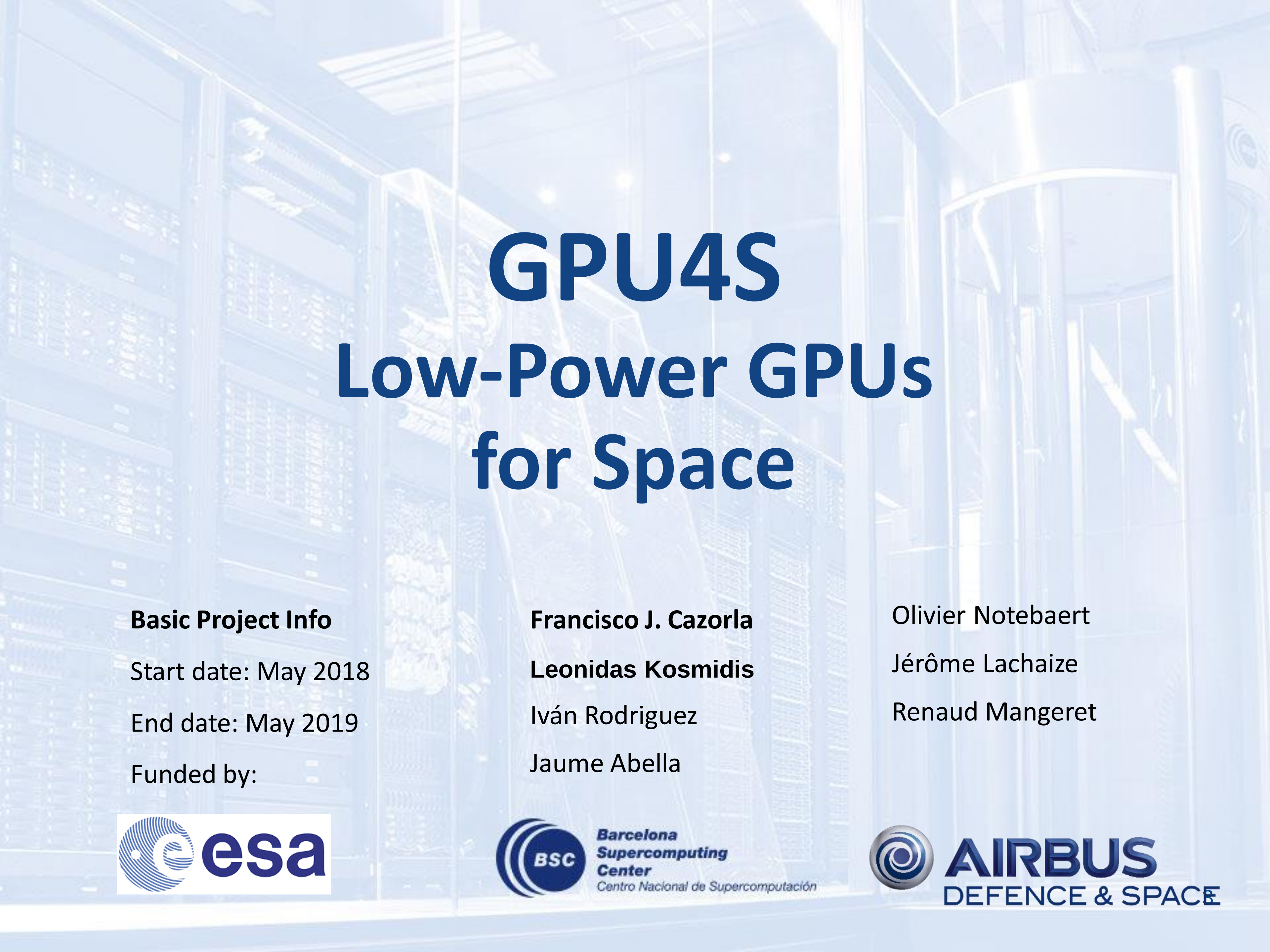
Leonidas Kosmidis, Iván Rodriguez, Jérôme Lachaize, Jaume Abella, Olivier Notebaert, Francisco Cazorla, David Steenari

27/02/2019

OBDP 2019

Outline

- Introduction to the GPU4S (GPU for Space) ESA Activity
- Development of the GPU4S Benchmarking Suite
- Methodology
- Preliminary Results
- Conclusions and Future Work



GPU4S
Low-Power GPUs for Space

Basic Project Info

Start date: May 2018

End date: May 2019

Funded by:

esa

Francisco J. Cazorla

Leonidas Kosmidis

Iván Rodriguez

Jaume Abella

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Olivier Notebaert

Jérôme Lachaize

Renaud Mangeret

AIRBUS
DEFENCE & SPACE

Project Overview

- Answer to Tender:
  - ESA ITT AO/1-9010/17/NL/AF
  - Low Power GPU Solutions For High Performance On-Board Data Processing
- Partners:
  - BSC (Coordinator)
    - CUDA Center of Excellence, extensive experience with embedded GPUs, critical systems and performance evaluation
  - Airbus Defense and Space (Toulouse)
    - Primary satellite supplier, experience with both hardware and software for space
- Project Duration
  - May 2018-May 2019

Previous Experience

- Extensive previous work on low-end embedded GPUs
- We can apply GPGPU on any embedded GPU
  - Use and benchmark devices **beyond** the few high-end CUDA/OpenCL GPUs in the embedded market
  - Majority of the embedded GPUs in the market are still low-end GPUs, supporting only embedded graphics, OpenGL ES 2 (basis of OpenGL SC 2)
  - We have developed GPGPU solutions on top of OpenGL ES 2 [1][2][3]
  - We are able both to achieve correct functionality [1] and optimise code for a given embedded GPU platform [2][4], while we can offer productivity and certifiability [3]

[1] Trompouki and Kosmidis, Towards General Purpose Computations on Low-End Mobile GPUs, DATE 2016

[2] Trompouki and Kosmidis, Optimisation opportunities and evaluation for GPGPU applications on low-end mobile GPUs, DATE 2017

[3] Trompouki and Kosmidis, High-Level Certification-Friendly Programming for GPU-powered Automotive Systems, DAC 2018

[4] Trompouki, Kosmidis, Navarro, An Open Benchmark Implementation for Multi-CPUMulti-GPU Pedestrian Detection, ICCAD 2017

Major Project Tasks and Expected Outcome

- Purpose:
  - Study the applicability of embedded GPUs in the space domain
  - Explore the possibility for ESA to build a hard rad GPU or to use a COTS one

- Perform a survey of the state of the art in
  - Existing embedded GPU, mainly European and major US (Nvidia, AMD)
  - Existing and future space algorithms amenable to GPGPU acceleration
- Select promising embedded GPUs
  - benchmark and compare them with existing on-board technologies
- Build a demo of a space application on the most appropriate candidate
- Define the roadmap for the adoption of embedded GPUs in space

Current GPU4S Project Progress 1/2

- HW and SW Survey has been completed
- HW Survey:
  - Covered almost every embedded GPU vendor based in Europe including the two major US based
    - ARM, Imagination, Think Silicon, Broadcom, Nvidia, AMD
      - Several models and rugged products
    - Direct contact with embedded GPU companies
    - Explored the possibility of IP acquisition for rad hard SoC design or FPGA use
  - Covered open source GPUs for FPGA implementation
  - Covered High-Level Synthesis (OpenCL) for FPGAs
  - Covered GPU-like architectures, ie. many cores, DSP-like etc
    - HPDP, RC64, Kalray MPPA, RISC-V accelerators by Esperanto/Semi-Dynamics and the European Processor Initiative (EPI) for automotive
  - Focus on devices up to 10W TDP
  - Evaluated also their software stacks (tools, APIs, libraries, development/optimisation productivity), certifiability

Current GPU4S Project Progress 2/2

- SW Survey:
  - Covered several Space domains:
    - Image processing/vision, SW defined radio, neural networks, compression
  - Focus on both existing and mainly future space mission needs
  - Inputs from several ADS divisions
  - Identify applications that are potentially amenable to parallelisation
  - Identify algorithms that are good fit for the GPU programming model
    - Eg. coalesced memory accesses, no thread divergence etc.

Benchmarked HW

- Hardware selection
- 3 platforms selected for experimental evaluation:
  - ARM Mali-G72
    - Huawei HiKey 970, HiSilicon Kirin 970 SoC
  - Imagination Technologies PowerVR Series 6
    - Renesas RCAR H3, ASIL-B Certified
  - Nvidia Jetson Xavier, ASIL-D Certifiable
    - **Early access adopters' program, starting from mid-October**

Benchmarking Methodology

- No benchmark suite for GPUs
  - EEMBC ADASMark has been only recently released
    - OpenCL only
    - Not representative for space
- Benchmarks for space?
  - NGDSP mainly signal processing
  - Euclid NIR?
    - Yes, but are we using the same input? Is the input representative or random?
- Solution:
  - Develop an open source GPU benchmark suite for space
  - Extract algorithm building blocks used in several domains
    - Maximum domain coverage with reasonable effort
    - Representative inputs, reference outputs and CPU versions for validation
  - Chain building blocks to mimick complex application scenarios
  - Euclid NIR for comparison with existing ports (using the same input!)

Summary of Identified Building Blocks

- A matrix showing identified building blocks and the domains they represent

Domains Building Block	Compression	Vision Based Navigation	Image Processing	Neural Network Processing	Signal Processing
Fast Fourier Transform			GENEVIS		ADS-B, NGDSP
Finite Impulse Response Filter					ADS-B, NGDSP
Integer Wavelet Transform	CCSDS 122				
Pairwise Orthogonal Transform	CCSDS 122				
Predictor	CCSDS 123				
Matrix computation		GENEVIS (Solver)		Image classification	
Convolution Kernel		OpenCV	GO3S,GENEVIS	Image classification	
Correlation		OpenCV	GO3S,GENEVIS		ADS-B
Max detection			GO3S	Image classification	ADS-B
Synchronization mechanism		GENEVIS	EUCLID NIR, GO3S	TensorFlow	ADS-B, NGDSP
Memory Allocation		CERES Solver , OpenCV	EUCLID NIR, GO3S	TensorFLow	ADS-B, NGDSP

- Complex application: Image recognition pipeline, based on CIFAR-10
- The Euclid NIR (Near InfraRed) has been selected as a potential demonstrator application for the GPU4S. It has been ported on Leon3 multicore (simulated) and Kalray MPPA256 in an OpenMP implementation. It requires few type of mathematical operations.

Many configurations per benchmark

SW side:

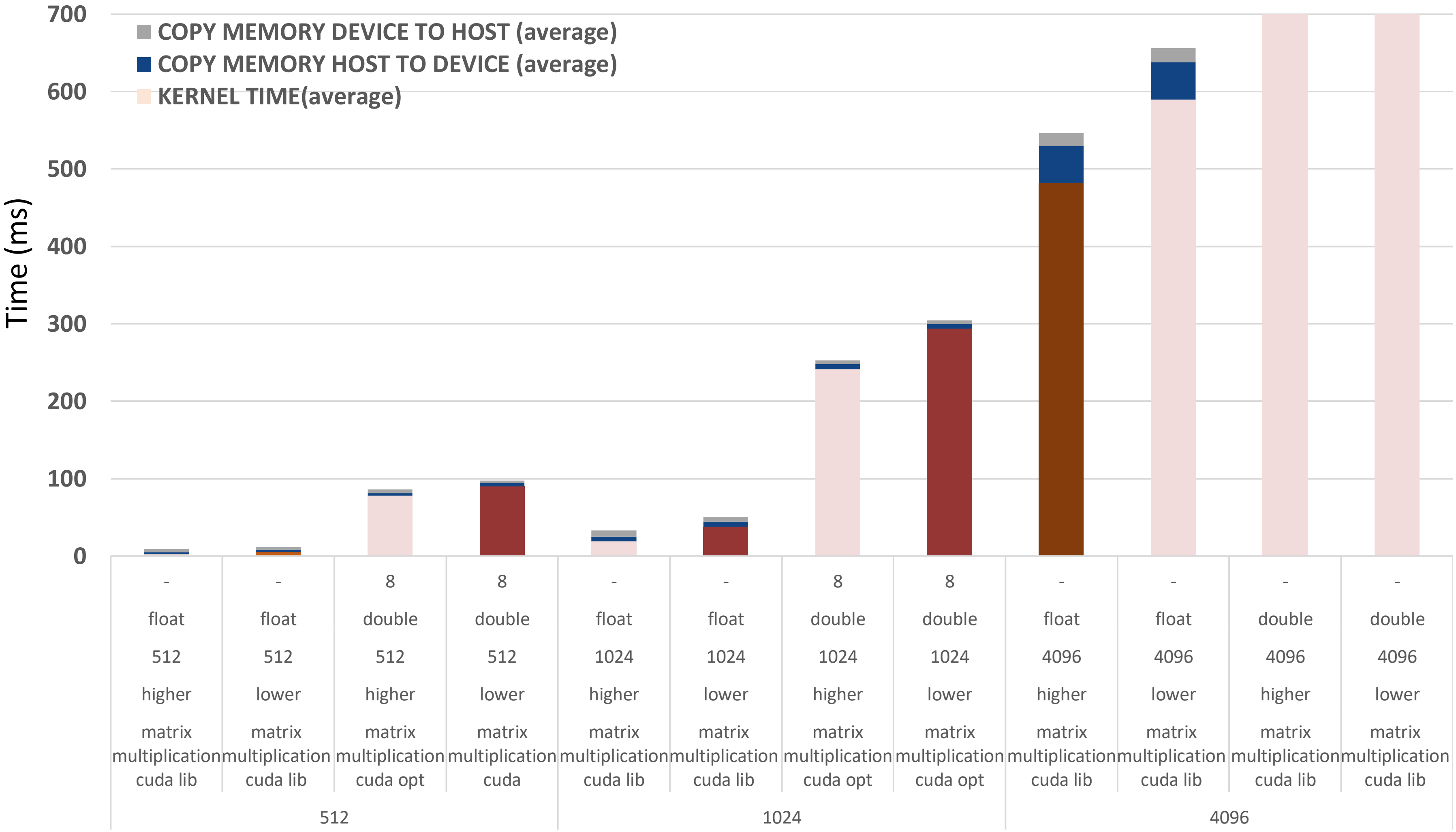
- Equivalent implementations in both CUDA and OpenCL, thanks to our carefully designed benchmark structure
- Variable, representative input sizes
- Several data types: 32-bit floating point, 16-bit floating point, double, integer
- Several benchmark versions:
  - Naïve parallelisation (straightforward)
  - Optimised handwritten implementation
    - E.g. memory blocking in the shared memory, thread coarsening etc.
    - Identify optimal blocking factor
  - Vendor provided library
- Several thread grid configurations
- Multiple executions to account for platform jitter (100 executions)

HW side:

- Different performance modes (TDP): low and high

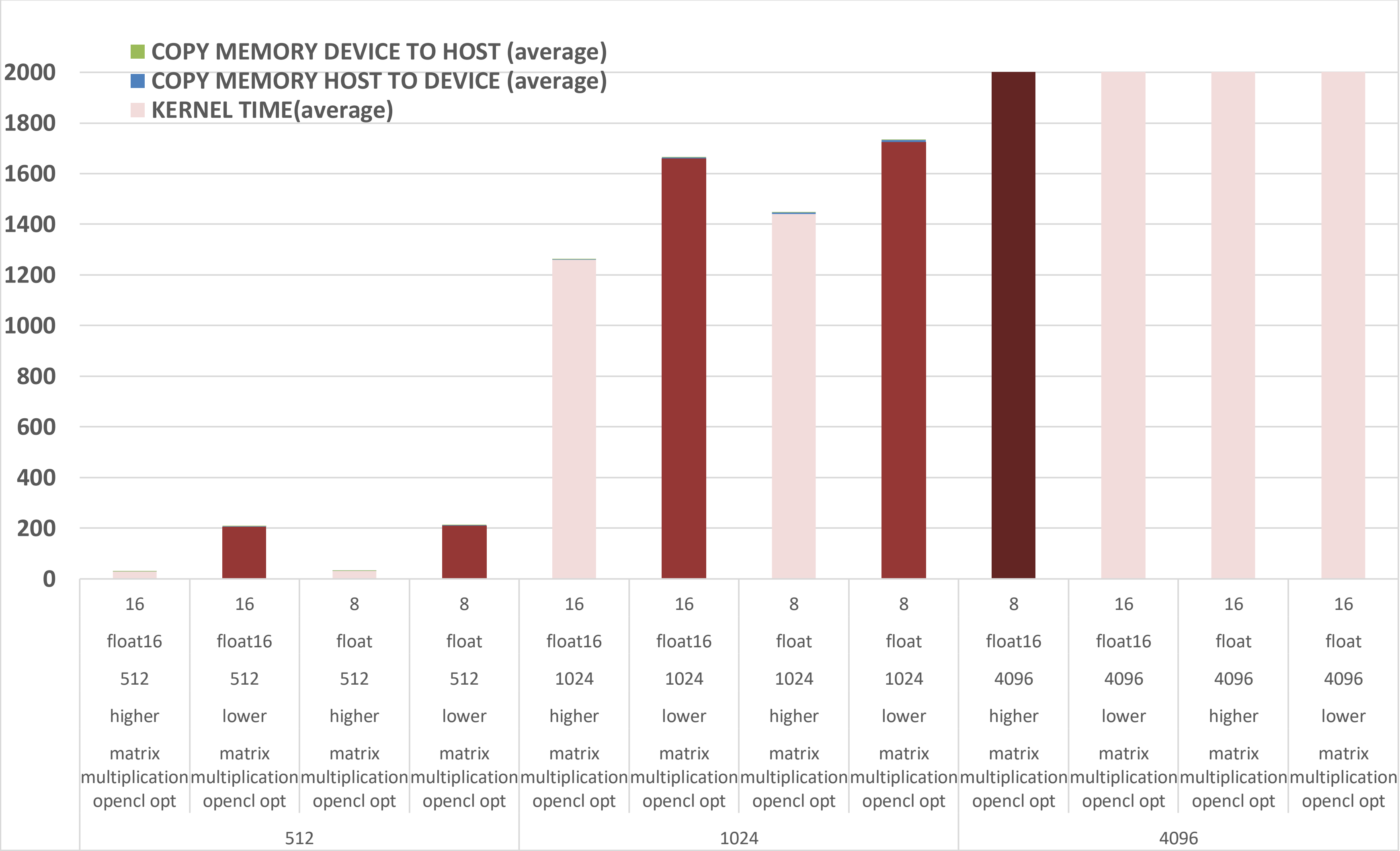
Nvidia Xavier: Low 10W, "High" 15W
Mali-G72 (HiKey 970): Low unknown, High 10W

Some Preliminary Results:
Matrix Mult on Nvidia Xavier

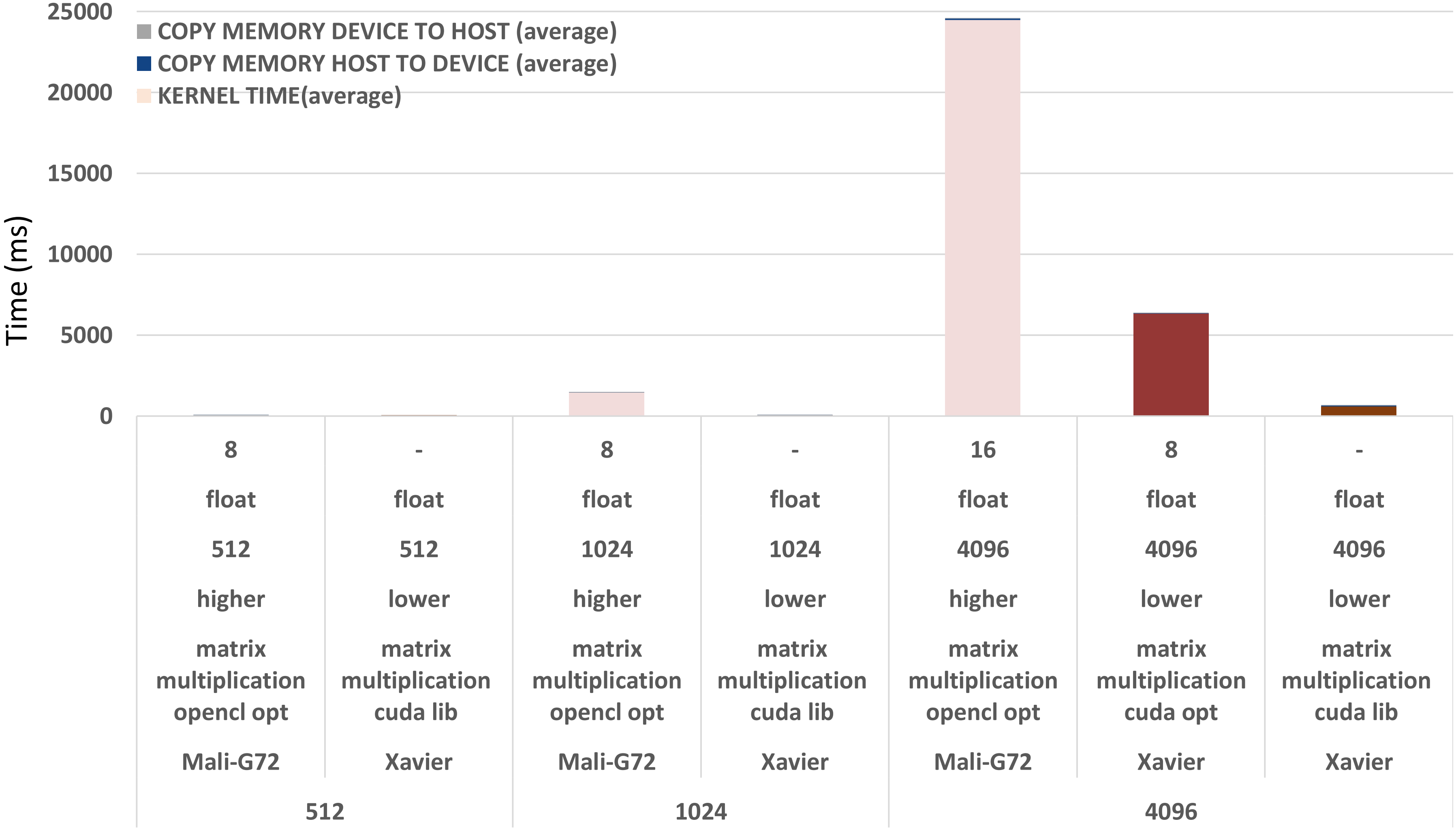


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

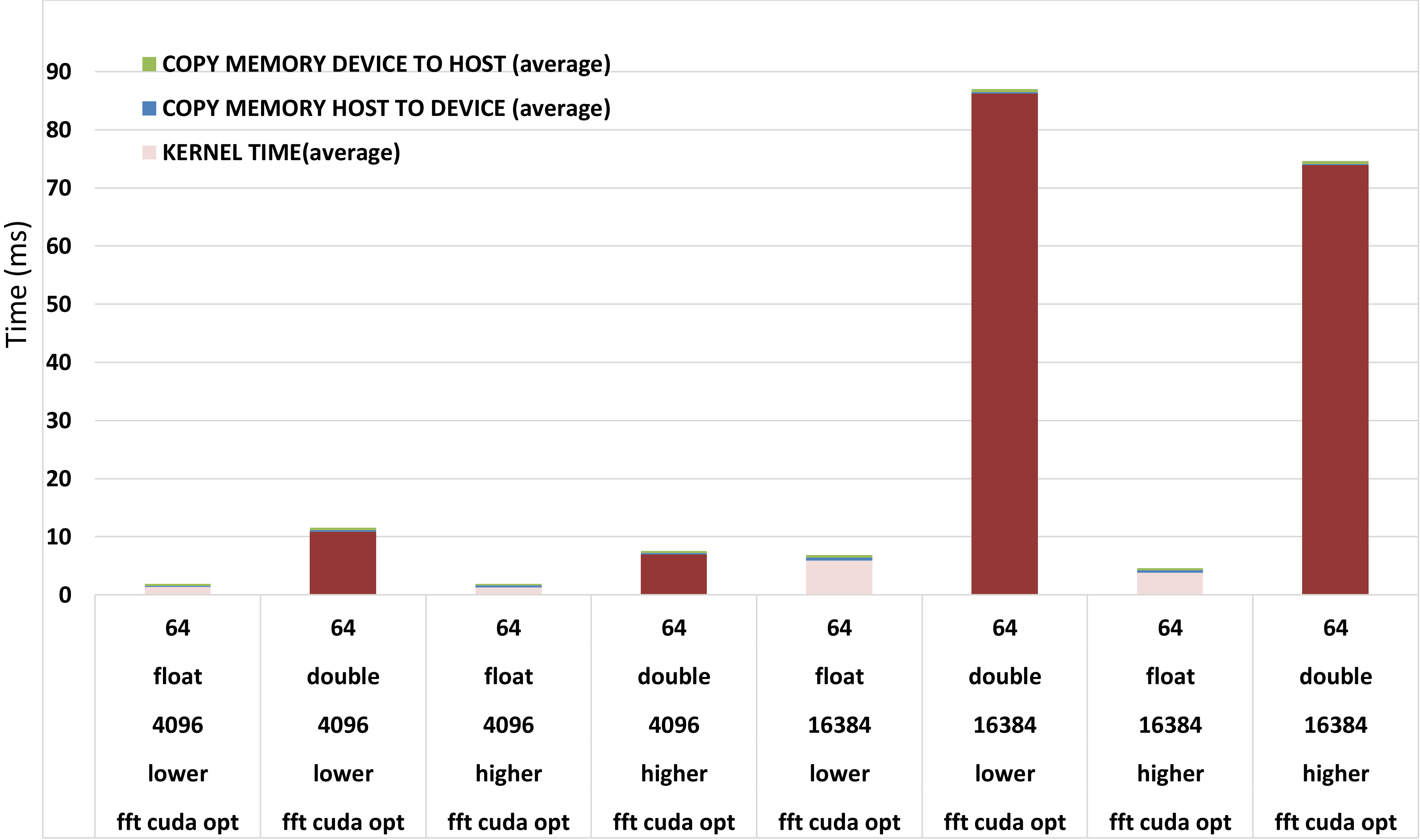
Some Preliminary Results:
Matrix Mult on HiKey 970 (Mali-G72)



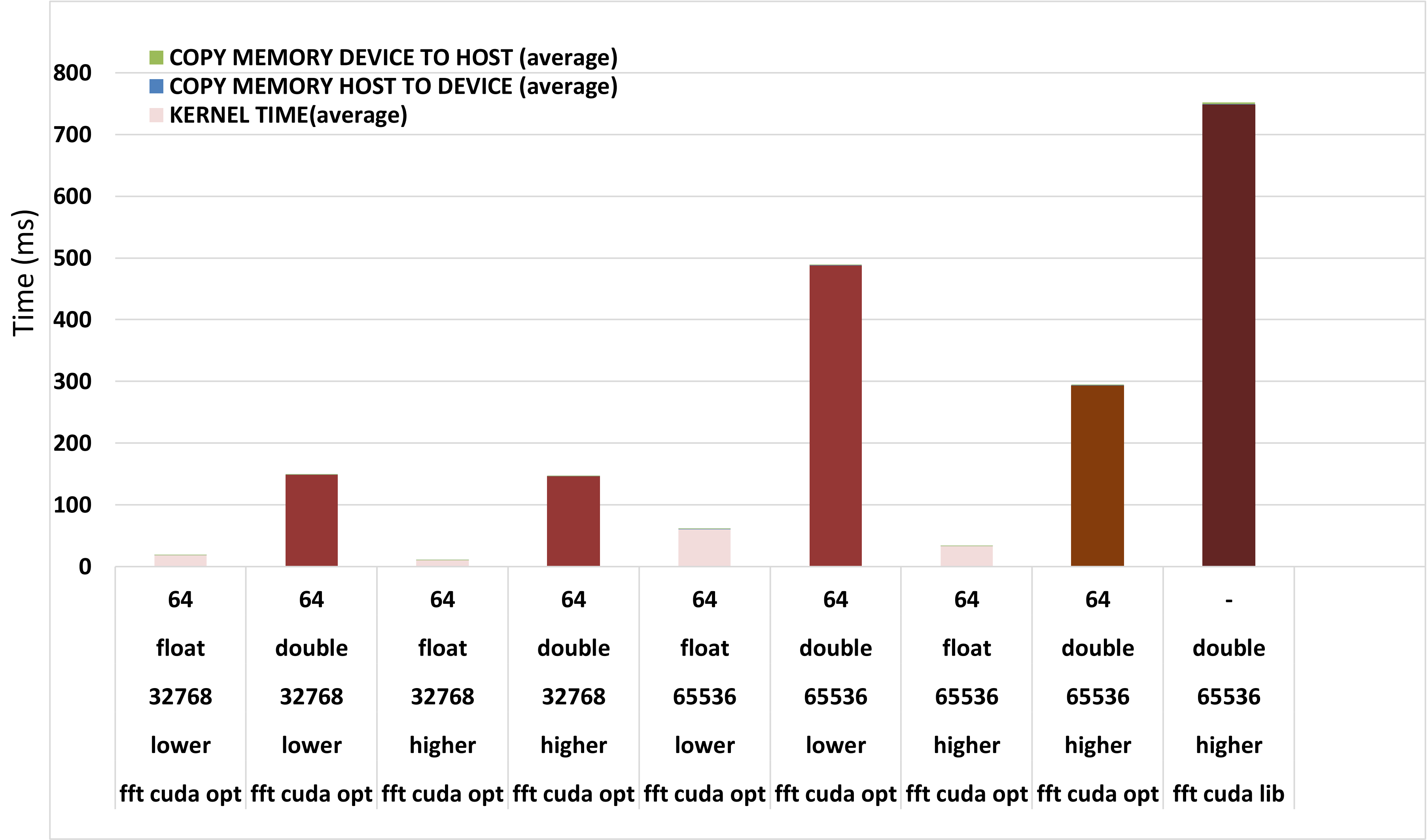
Some Preliminary Results: Matrix Mult Nvidia Xavier vs HiKey 970 (Mali-G72)



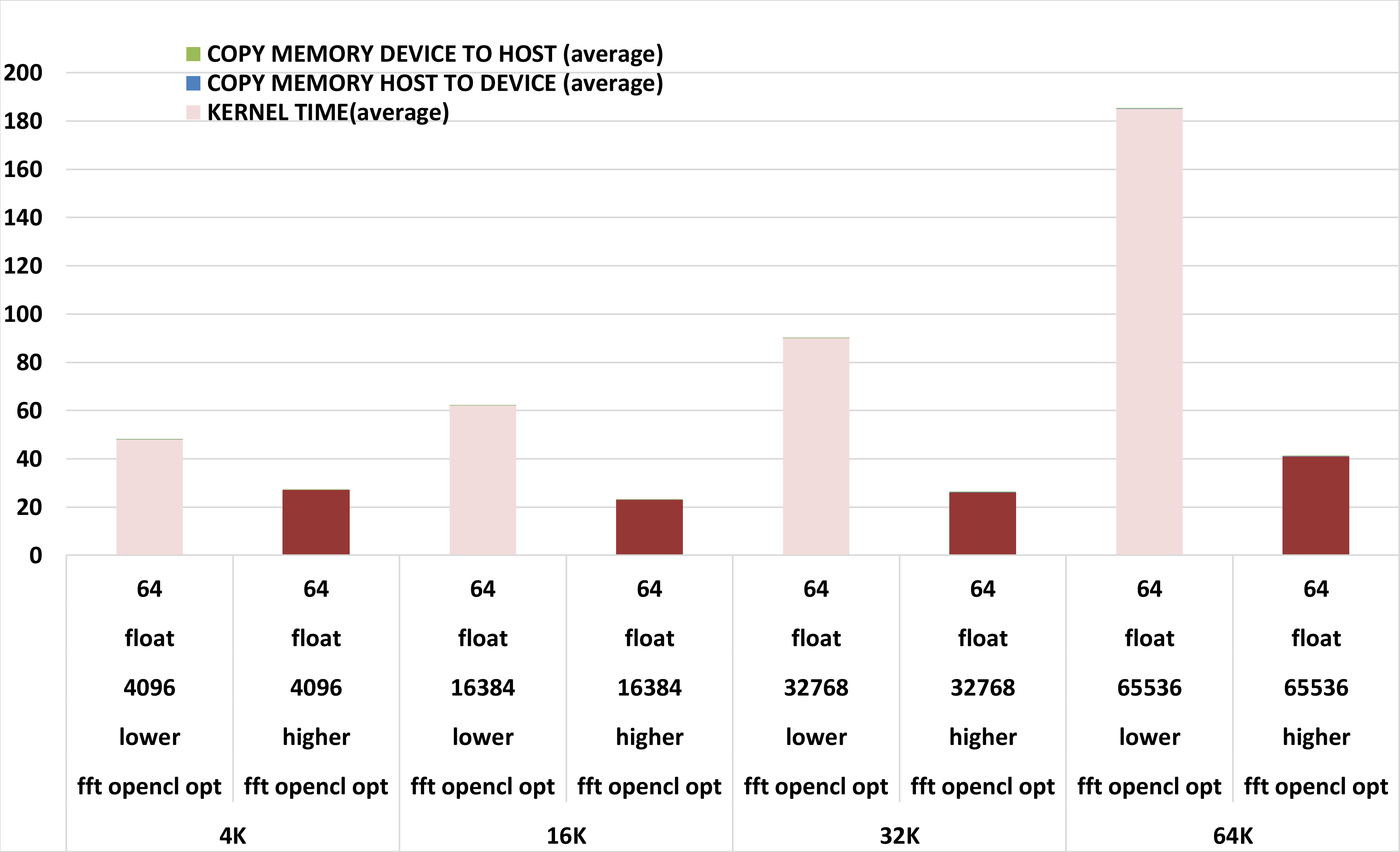
Some Preliminary Results: FFT on Nvidia Xavier 1/2



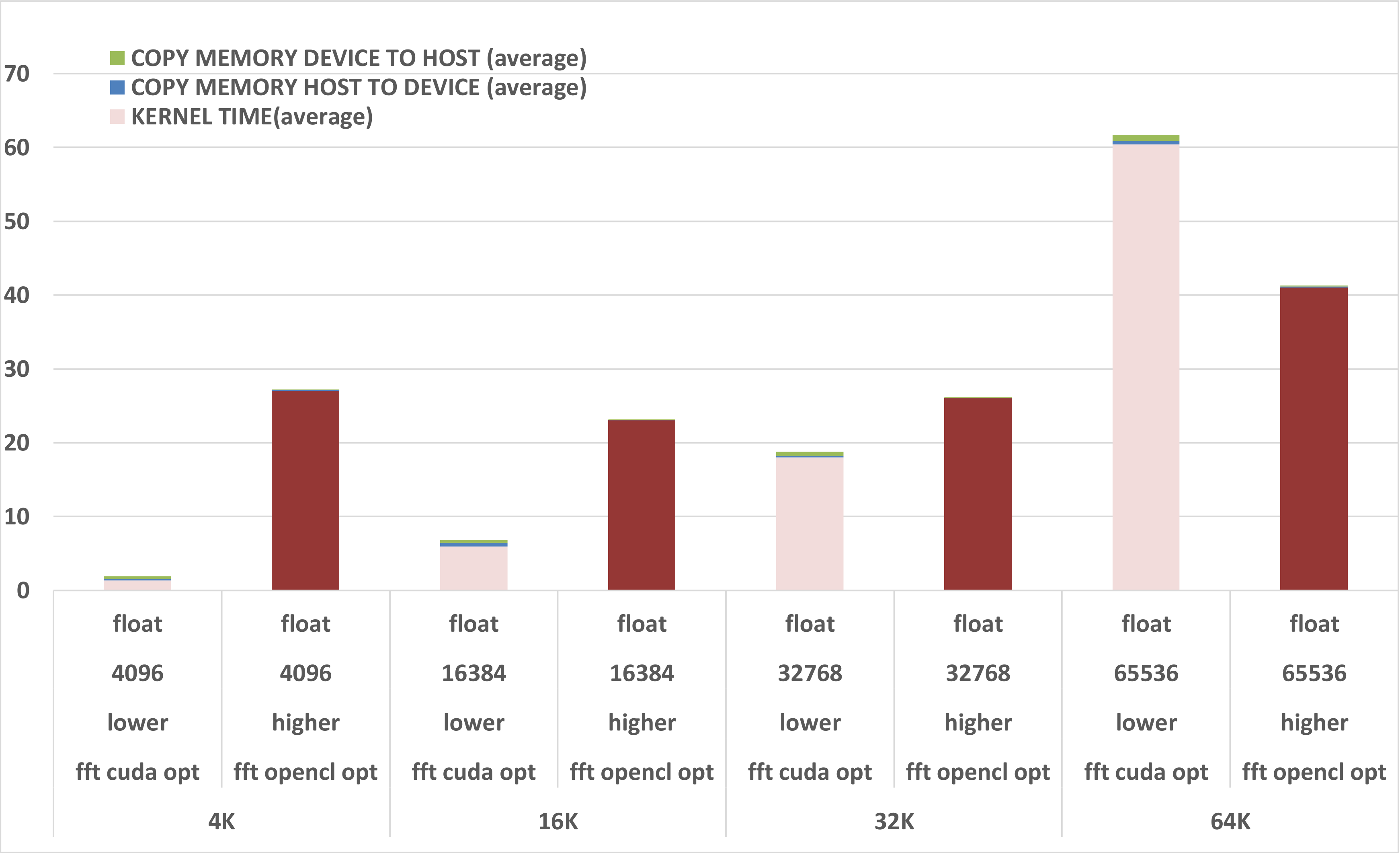
Some Preliminary Results: FFT on Nvidia Xavier 2/2



Some Preliminary Results: FFT on HiKey 970 (Mali-G72)



Some Preliminary Results: FFT Nvidia Xavier vs HiKey 970 (Mali-G72)



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Conclusion and Future Work

- Open source GPU benchmark suite for space developed
  - To be released after the end of the project
- Nvidia's Xavier provides high performance at the 15W
  - But vendor provided libraries are not perfect and not for every purpose
- ARM's Mali-G72 is competitive for the same 10W power budget
- Imagination's ASIL-B GPU will also be evaluated when the Renesas R-CAR H3 is delivered
- Euclid NIR will be ported soon
  - To be released as open source, too
- Normalisation of results to 65nm to be performed
- Include comparison on paper with other space technologies based on published results
  - Do you want to include your space hardware in the comparison?
  - Please contact us with results at leonidas.kosmidis@bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

AIRBUS DEFENCE & SPACE

Embedded GPU benchmarking for High-Performance On-board Data Processing

Leonidas Kosmidis, Iván Rodriguez, Jérôme Lachaize, Jaume Abella, Olivier Notebaert, Francisco Cazorla, David Steenari

27/02/2019

OBDP 2019

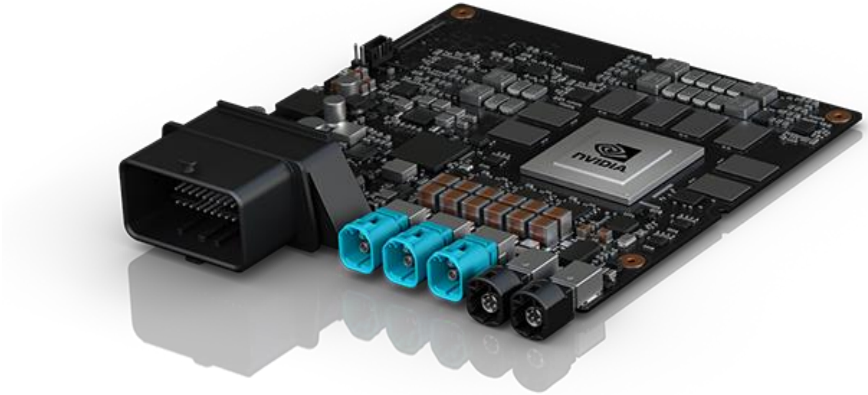
Backup Slides

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

NVIDIA

Xavier

- AutoChauffeur replacement in a single SoC
  - New GPU architecture, Tesla
  - New CPUs, 8 Carmel ARMv8 from Nvidia
  - Deep Learning Accelerator and Tensor Processing Core
- 1.3 TFLOPS
- TSMC 12nm FinFET
- LPDDR4 Memory
- TDP: 30W
- Designed to comply with ISO-26262
  - Targets ASIL-C
  - But not certified yet, neither information about it



NVIDIA

Xavier Jetson (~$1300)

- Multiple operating modes: 10W, 15W, 30W
- 10X energy efficiency over the TX2
- 20X performance over the TX2
- Not automotive grade

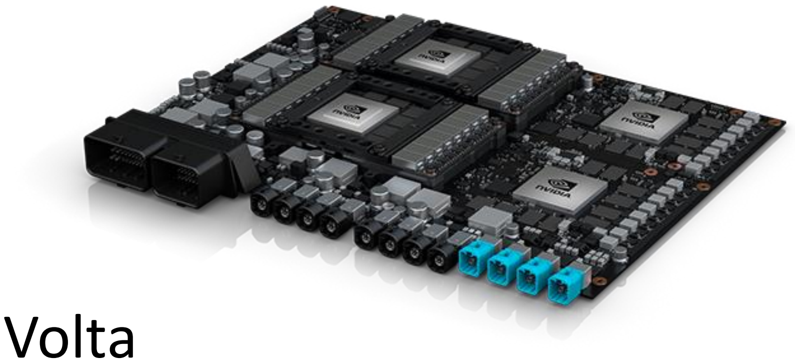
Dual Xavier ($20K)

- 60 TOPS

NVIDIA

Drive Pegasus ($50K)

- Performance requirements of level 5 autonomy (self driving)
- Designed for ASIL-D certification
- Two Xavier SoCs with 2 discrete post Volta GPUs
- TDP: 500W
- 320 TOPS



NVIDIA

Xavier Summary

- No indications of use in space
- Automotive platforms available from NVIDIA
- But no indication regarding certification

NVIDIA

Pros

- Most used programming language
- High-performance

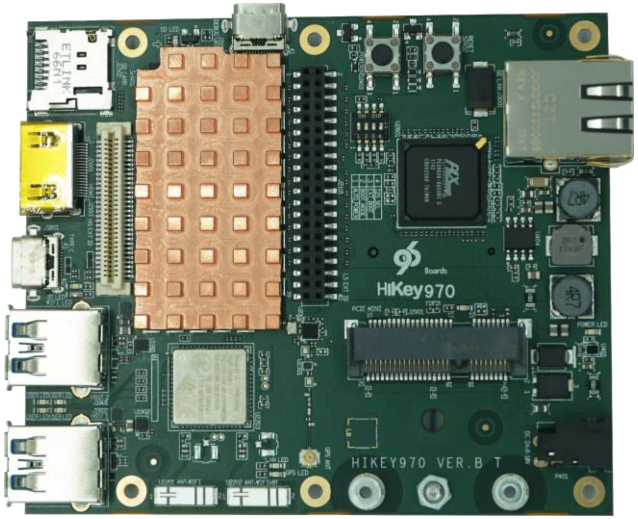
Cons

- Platforms change rapidly with older ones not receiving support and running out of stock
- Closed source ISA and tools, No OpenCL on Tegra
- Proprietary language
- No certified driver yet
- Some platforms exceed the 10W limit

ARM Updates

Suggested COTS board to be evaluated

- HiKey 970 ($299.00) 9/2017
  - ARM big.LITTLE processor TSMC 10nm
  - 4 ARM Cortex-A73, 4 Cortex-A53 cores
  - 6GB of LPDDR4X SDRAM memory
  - Mali G72 MP12 GPU (746MHz)
  - 64GB of UFS 2.1 flash storage



Imagination Updates

Critical Markets

- Already GPUs in automotive products
- Interested to target space requirements

- Renesas R-CAR H3 ($859)
- ASIL-B Certified
- Arm A57, Arm A53
- Series6XT GX6650, 6 Unified Shading Clusters (USCs) and 192 cores
- 4GB LPDDR4
- Introduced 12/2015
- Mass production started 3/2018

Imagination Updates

No COTS board available for newer GPU families

- Series 8 in automotive products but silicon only available to OEMs
- Series 9 in the fabs, products in ~1 year in the market
- Series 6 can be still become available for licensing and to be maintained if it is licensed
- Basic design and features are the same e.g. Virtualisation

ARM/Imagination Comparison

	Imagination			ARM		
	GX6650	7XTP	GX6250	HiKey 970	HiKey 960	Mali-400MP2
Price	$859	350 eur	£399.99	$299.00	$239.00	
Power				<10W	10W	
Performance	384GFLOPS	204GFLOPS		346GFLOPS	272GFLOPS	up to 10/20GFLOPS

NVIDIA/AMD Comparison

Nvidia											AMD			
K1	TX1	Drive PX	CX	TX2	TX2i	AutoC	AutoCl	Xavier	Jetson Xa	Pegasus	E8860	Unibap	Ryzen V1605B	AMD GX-210HA
$339	~500 eur				£699.00				$1300	$50K				
5-20W	10-15W	20W		<7.5W		10W	250W	30W		500W	37W	4-20W	12-25W	9W
384GF	1TF/16	1TF/16		1TF16		1.3TF	8TF	1.3 TF	60TOPS	320TOPS	769GF	77GF	1TFL/16	85GF

Minimum BSP

Software on GPU

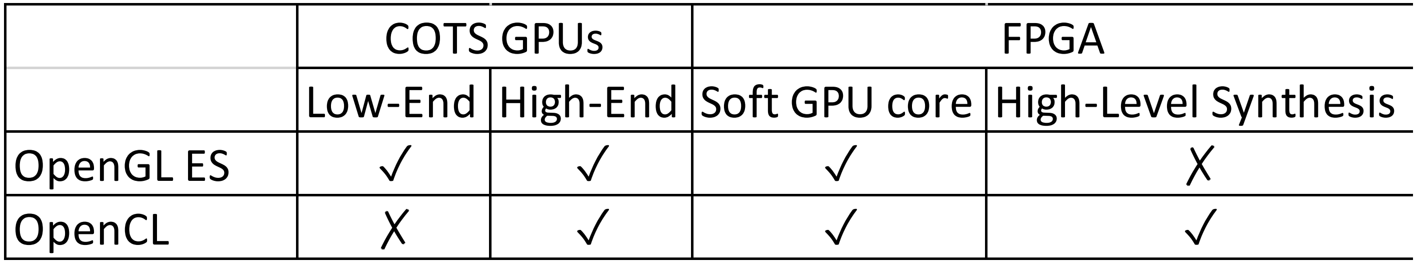
- No strong dependence on API
- All supported APIs are new for this domain
  - OpenGL ES 2, OpenCL, CUDA
- Operating Systems:
  - RTOS desirable but not hard requirement
  - Linux is universally supported and it is acceptable
  - Increasing adoption in space
- No dependence on External Libraries
  - Libraries such as OpenCV are only used for prototyping
  - At deployment replaced by custom optimised code for the target

Minimum BSP

Software around GPU

- Strong dependence on development tools
  - Ease of programming
- Debugging
- Profiling/Inspection
  - Necessary for validation
- Performance tuning resources
  - Performance counters
  - Optimisation hints
    - Tools
    - Documentation
- Tool support of operating systems (desirable)
  - Possibility of updates of the toolchain
  - Possibility to use them on newer operating systems

Taxonomy



	COTS GPUs		FPGA	
	Low-End	High-End	Soft GPU core	High-Level Synthesis
OpenGL ES	✓	✓	✓	✗
OpenCL	✗	✓	✓	✓

Taxonomy

	COTS GPUs		FPGA		Many Cores/
	Low-End	High-End	Soft GPU core	High-Level Synthesis	GPU-like
OpenGL ES	✓	✓	✓	✗	✗
OpenCL	✗	✓	✓	✓	✓ (some)
OpenMP	✗	✗	✗	✗	✓ (some)
Custom Programming Models	✗	✗	✗	✗	✓

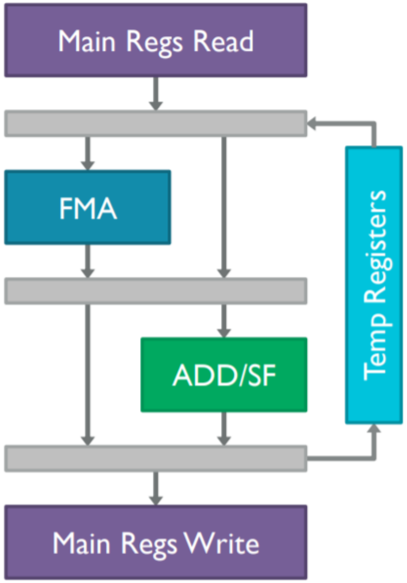
Taxonomy

	COTS GPUs		FPGA		Many Cores/
	Low-End	High-End	Soft GPU core	High-Level Synthesis	GPU-like
OpenGL ES	✓	✓	✓	✗	✗
OpenCL	✗	✓	✓	✓	✓ (some)
OpenMP	✗	✗	✗	✗	✓ (some)
Custom Programming Models	✗	✗	✗	✗	✓

ARM

Mali-Gxx (Bifrost)

- Latest GPU version
- OpenCL 2:
  - Shared Virtual Memory
  - Easier programmability in heterogeneous architectures
- More expensive IP
- Scalar design (“NVIDIA” like), with 4 threads
  - Higher utilisation than previous ARM GPU generations
  - But thread divergence problems
  - Although less probability to experience it with only 4 threads
- Dual-Issue
- Configurable number of cores and cache



ARM

arm

Overall Evaluation

Pros:

- Most licensed embedded GPU
- 50% of the entire market
- Can provide optimised cell-libraries in addition to IP
  - 3 versions: Low-power and low leakage, low-cost and high density and high performance and high speed
  - Not 65 nm
- Previous experience with space related projects
  - DAHLIA
  - Only IP has been licensed but ported to ST process libraries
    - 28nm FDSOI
  - Easier to integrate

Imagination

Both PowerVR Furian and Rogue:

- Unified Shader Architecture
- Scalar Architectures
- Tile-based Deferred Rendering
- Support for both Graphics (OpenGL ES 2) and Compute (OpenCL)
- Hardware Virtualisation
  - Enables partitioning for up to 8 Oses
  - Security
  - Critical systems eg. Automotive
- Microkernel
  - Dedicated microcontroller
  - Enables debugging
  - Managing Interrupts from CPU
  - Customisation for different markets:
    - Eg. advanced DVFS or power-gating based on workload information

Imagination

PowerVR Rogue

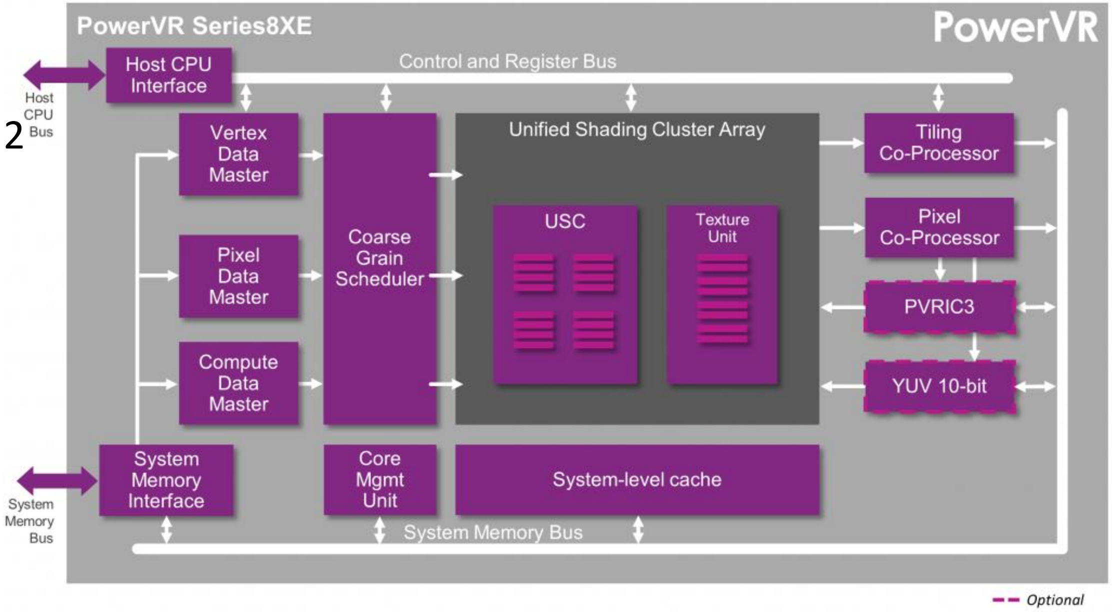
- Series 8 and Series 9

Low-end: Series 8XE and Series 9XM

- Optimised for limited area
- Customisable numbers of pixel processing per cycle

Mid range: Series 9

- Better performance/mm^2
- New MMU
- 36-bit addressing



Imagination

PowerVR Furian (Series8XT 2017):

- Higher performance
- Two-level MMU
  - Shadow page table support
- Specifically designed for Automotive (ADAS)
  - Can run a different tasks in each of the 2 Scalable Processing Units
  - Mixed Criticality
- Optimised for sub-14nm
  - Shorter paths
  - Less congestion
- Support for Shared Virtual Memory (OpenCL 2.0)
  - Easier programmability

Imagination

Overall Evaluation

Pros:

- Long experience in the market
- Proven Integration with several processors
  - ARM in Apple products, SH for DreamCast (Hitachi) and Automotive (Renesas), x86 in Atom
- Hardware Virtualisation is attractive for critical systems
- Automotive/Certification Oriented
  - R-Car H3 from Renesas (Series 5 and 6) reached ASIL-B
  - Furian designed to reach ASIL-D
  - Interested to specifically target space standards