Green Flash

High performance computing for real-time science

Project overview & status

Project #671662 funded by European Commission under program H2020-EU.1.2.2 coordinated in H2020-FETHPC-2014

Green Flash @ AO4ELT5

1) Biasi et al. “*FPGA based microserver for high performance real-time computing in AO*” [**P3055**]
2) Reeves et al. “*The Green Flash Real-Time simulator*” [**P1052**]
3) Perret et al. “*A generic and scalable heterogeneous architecture for real-time computing and performance measurements in AO*” [**P3037**]
4) Doucet et al. “*Efficient supervision strategy for tomographic AO systems on ELTs*” [**P3054**]
5) Bernard et al. “*A GPU based RTC for the E-ELT AO: RTC prototype*” [**P3045**]
6) Jenkins et al. “*ELT scale real-time control on Intel Xeon Phi and manycore CPUs*” [**P3036**]
7) Ferreira et al. “*ROKET: erROr breaKdown Estimation Tool for adaptive optics systems*” [**P1057**]
8) Vidal et al. “*MICADO SCAO numerical simulations*” [**P1056**]
9) Petit et al. “RTC strategies for Harmoni SCAO and LTAO modes” [**P3035**]

l'Observatoire de Paris | LESIA
Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique

Durham University

MICROGATE

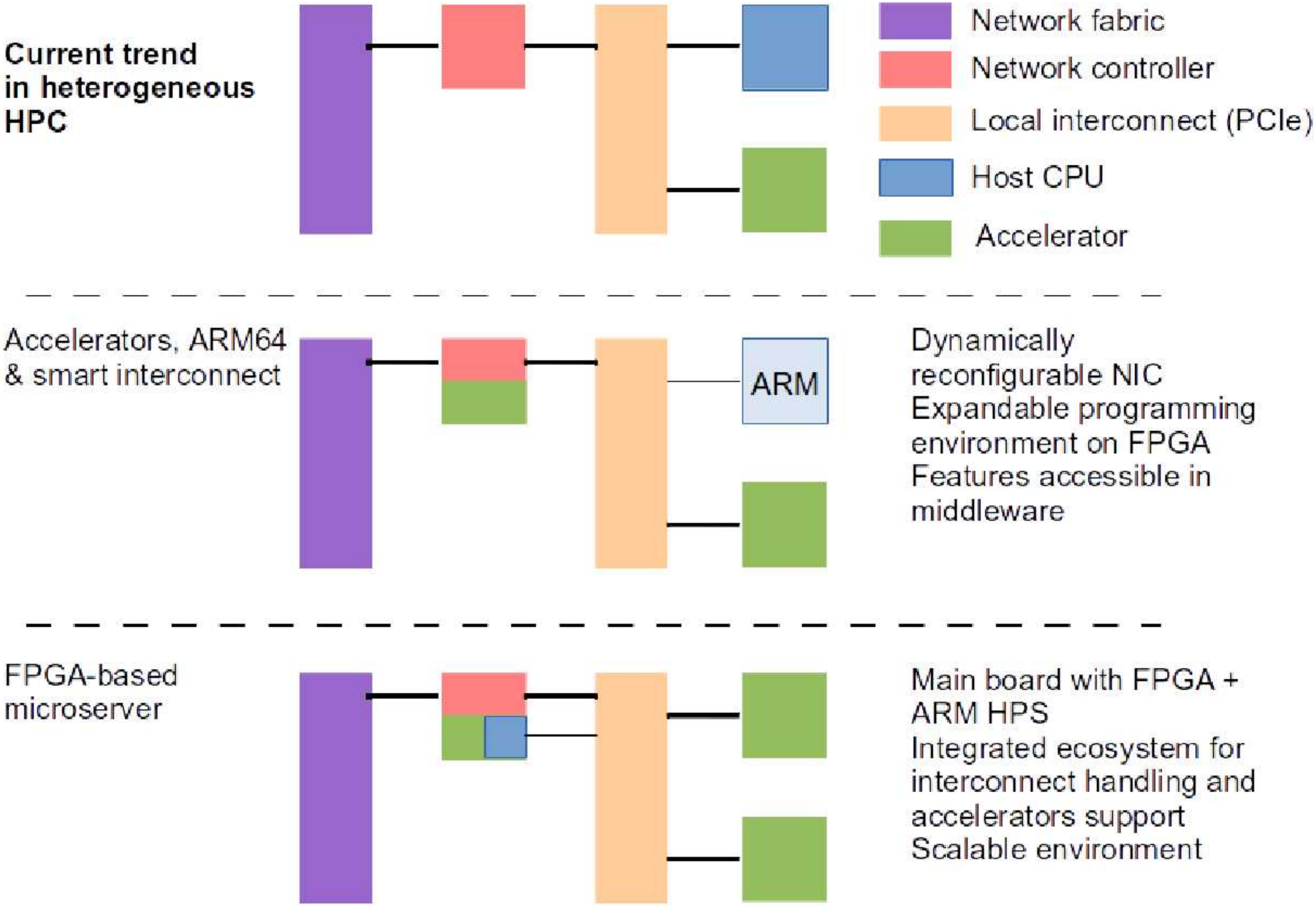
PLDA

What this is about … really

- Find the best trade-off for ELT sized AO systems RTC
  - Comprehensive assessment of existing technologies
  - Development of new custom solutions for comparison
  - Propose new development processes to reduce cost and increase maintainability
- Build one or several full featured RTC prototype at the largest scale possible
  - Technology down-selection from a number of criteria : performance, cost, compliance to standards, obsolescence, maintainability
  - State of the art systems to be assessed in the lab, with a simulator

Green Flash

Assessing new HPC concepts



Current trend in heterogeneous HPC

Accelerators, ARM64 & smart interconnect

Dynamically reconfigurable NIC
Expandable programming environment on FPGA
Features accessible in middleware

FPGA-based microserver

Main board with FPGA + ARM HPS
Integrated ecosystem for interconnect handling and accelerators support
Scalable environment

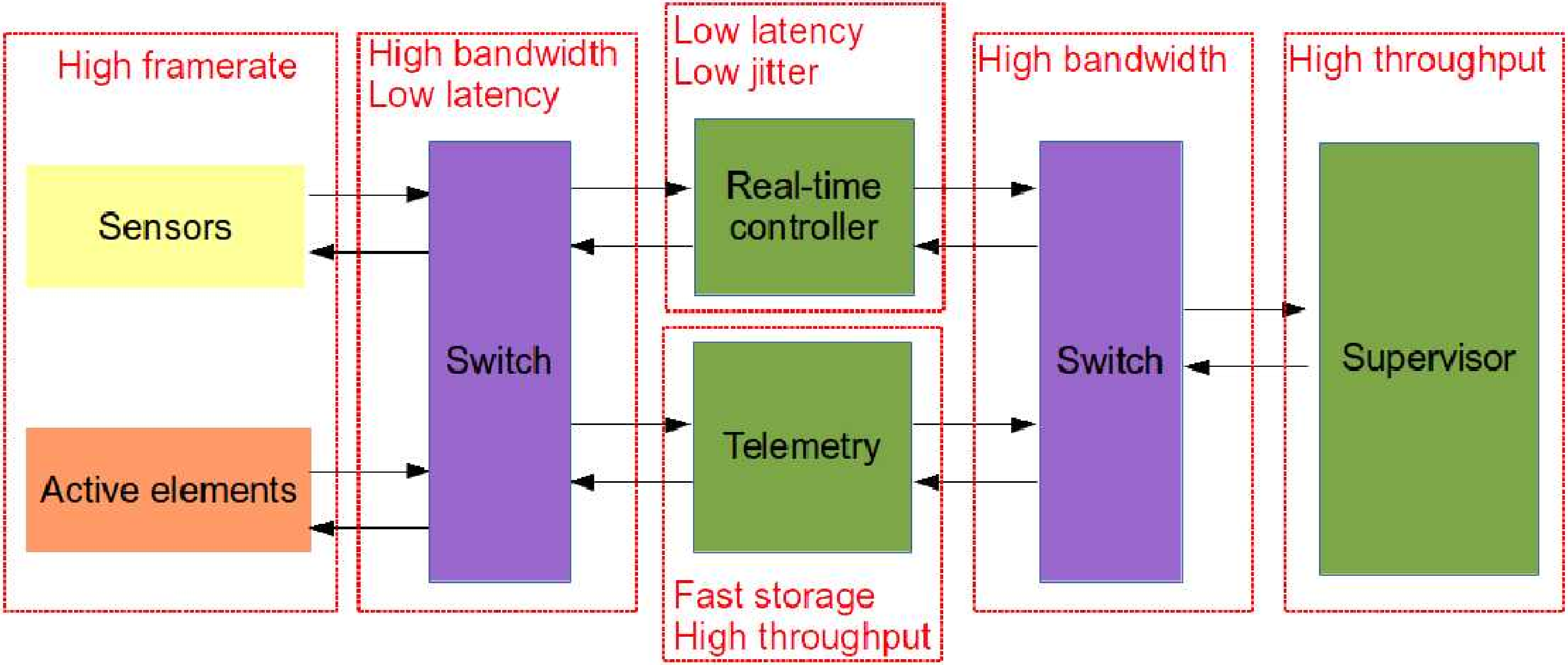
l'Observatoire de Paris — LESIA
Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique

Durham University

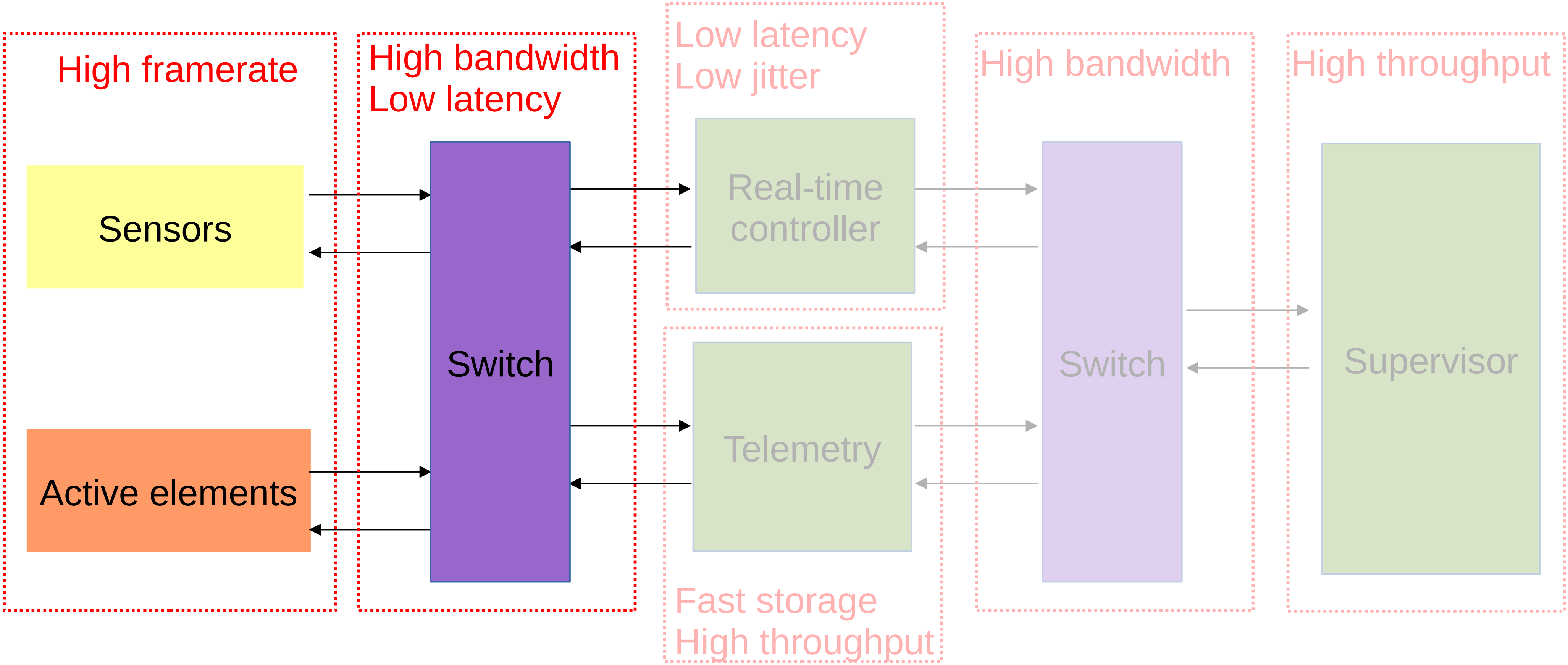
MICROGATE

PLDA

AO RTC concept

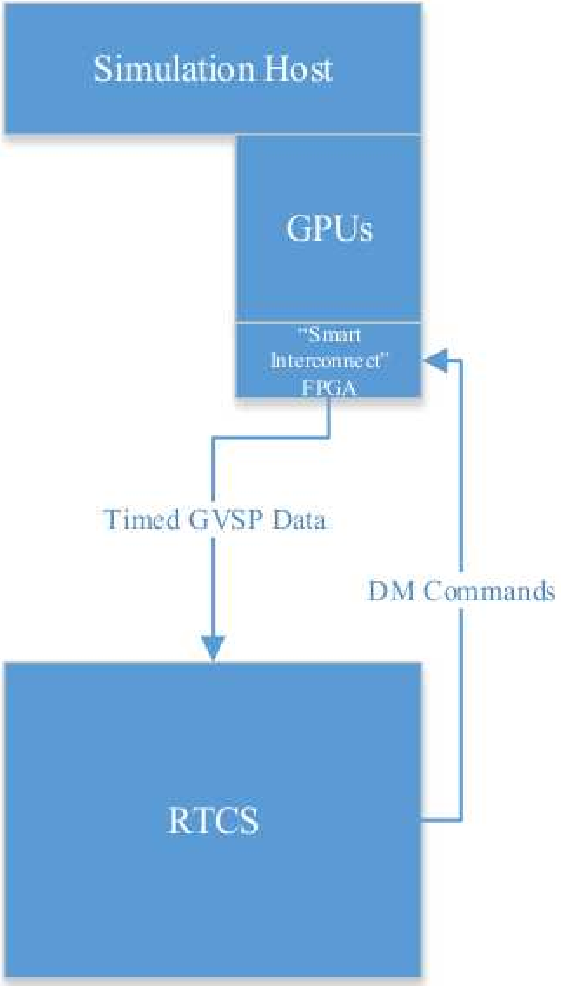


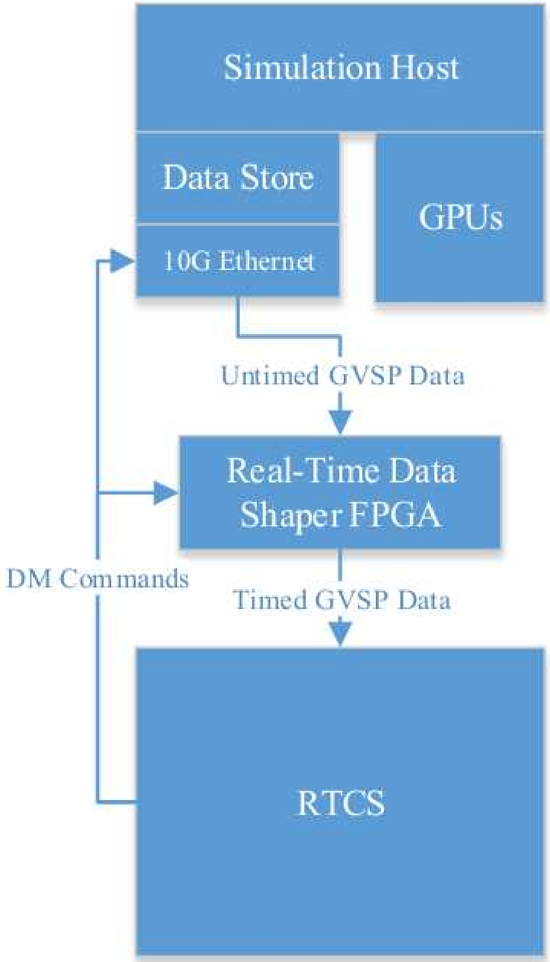
AO RTC concept : RT simulator



Real-time simulator concept

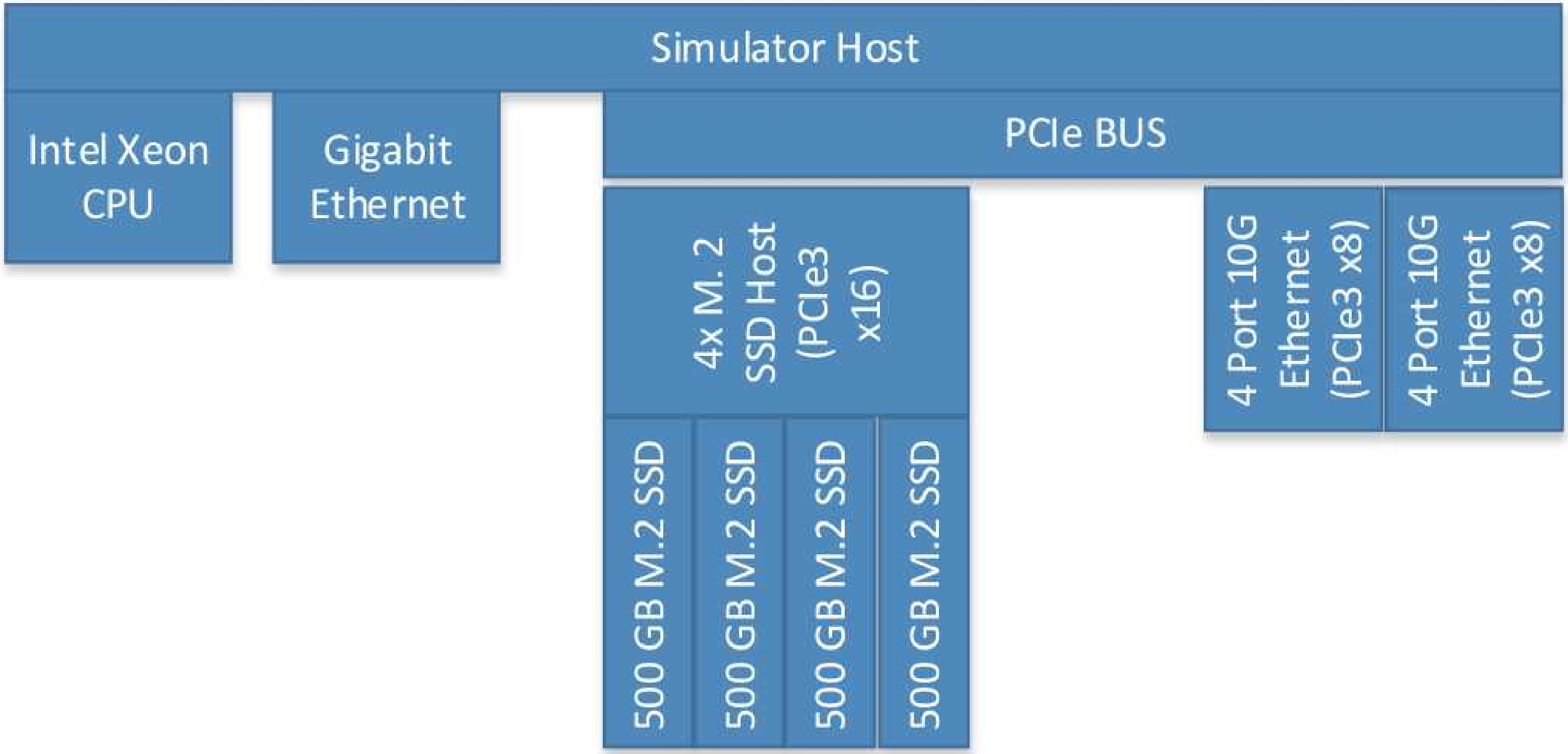
- 2 modes : simulation rate / real-time rate





Real-time simulator concept

- Data store concept
  - PCIe carrier with 4 SSDs (up to 12 GB/s)



Simulator Host

Intel Xeon CPU

Gigabit Ethernet

PCIe BUS

4x M. 2 SSD Host (PCIe3 x16)

500 GB M.2 SSD

500 GB M.2 SSD

500 GB M.2 SSD

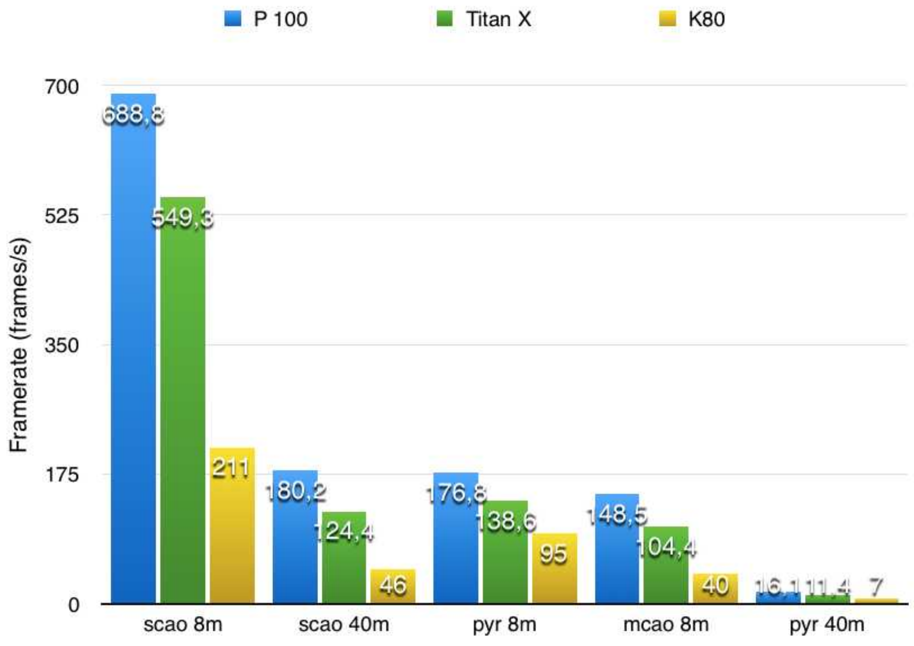
500 GB M.2 SSD

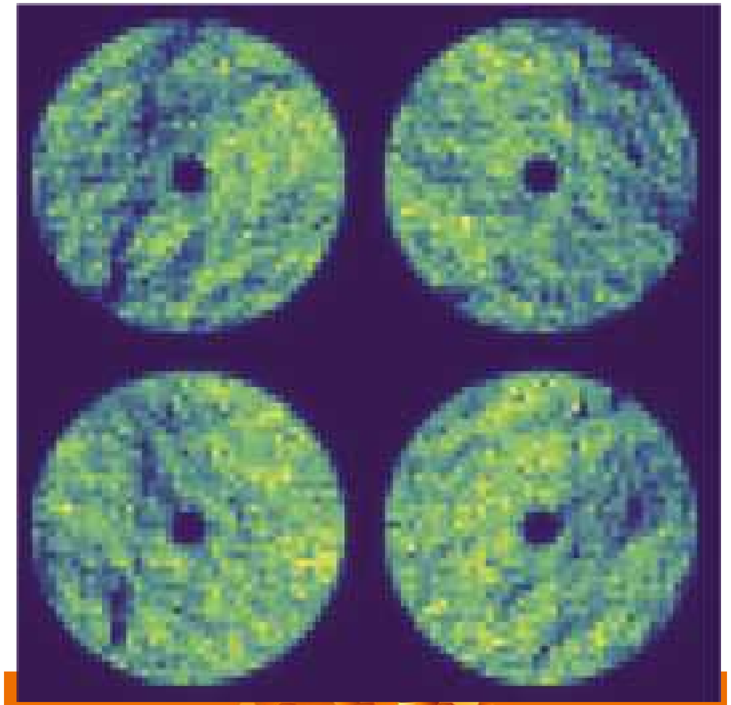
4 Port 10G Ethernet (PCIe3 x8)

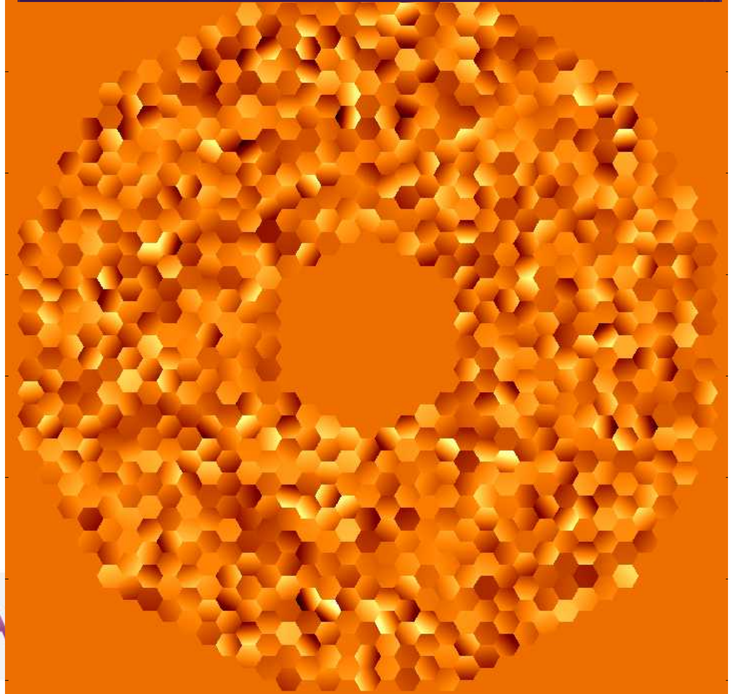
4 Port 10G Ethernet (PCIe3 x8)

Real-time simulator SW

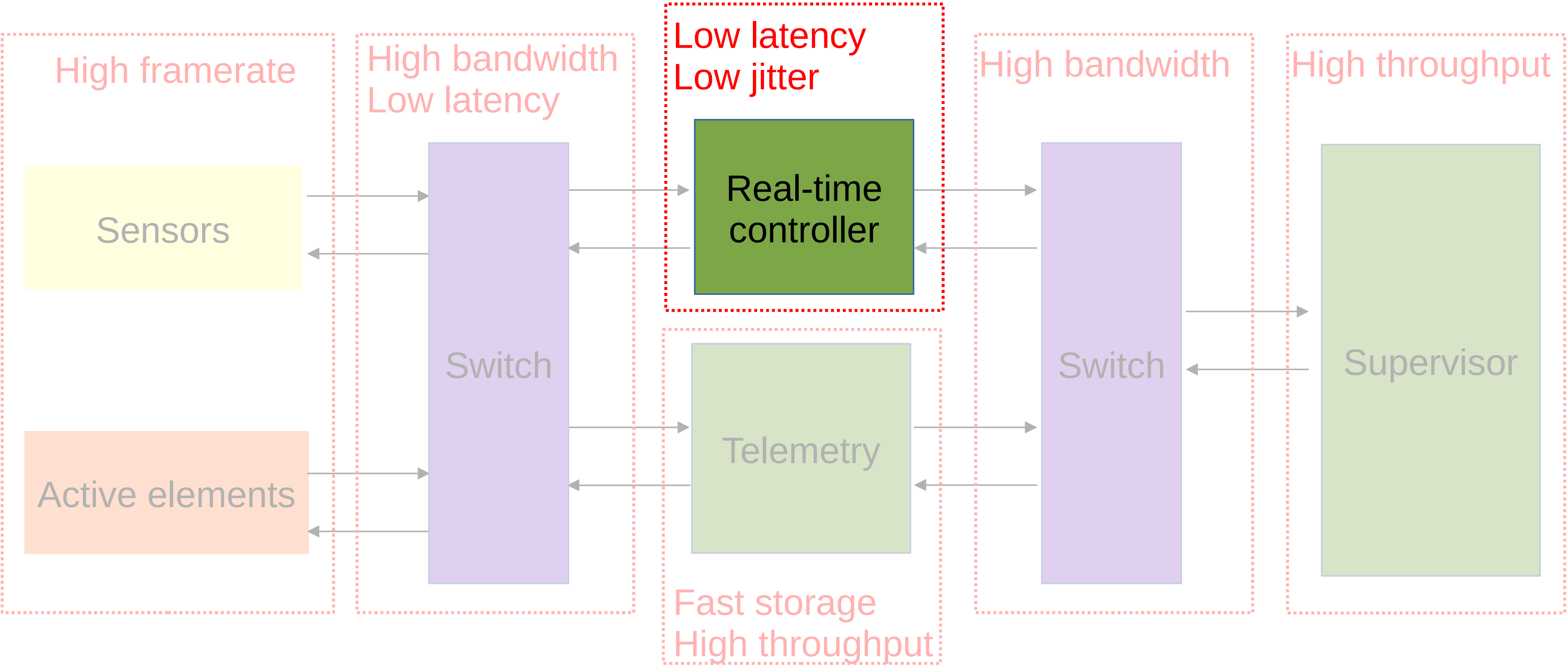
- COMPASS simulator
  - GPU based, scalable, versatile, very fast !



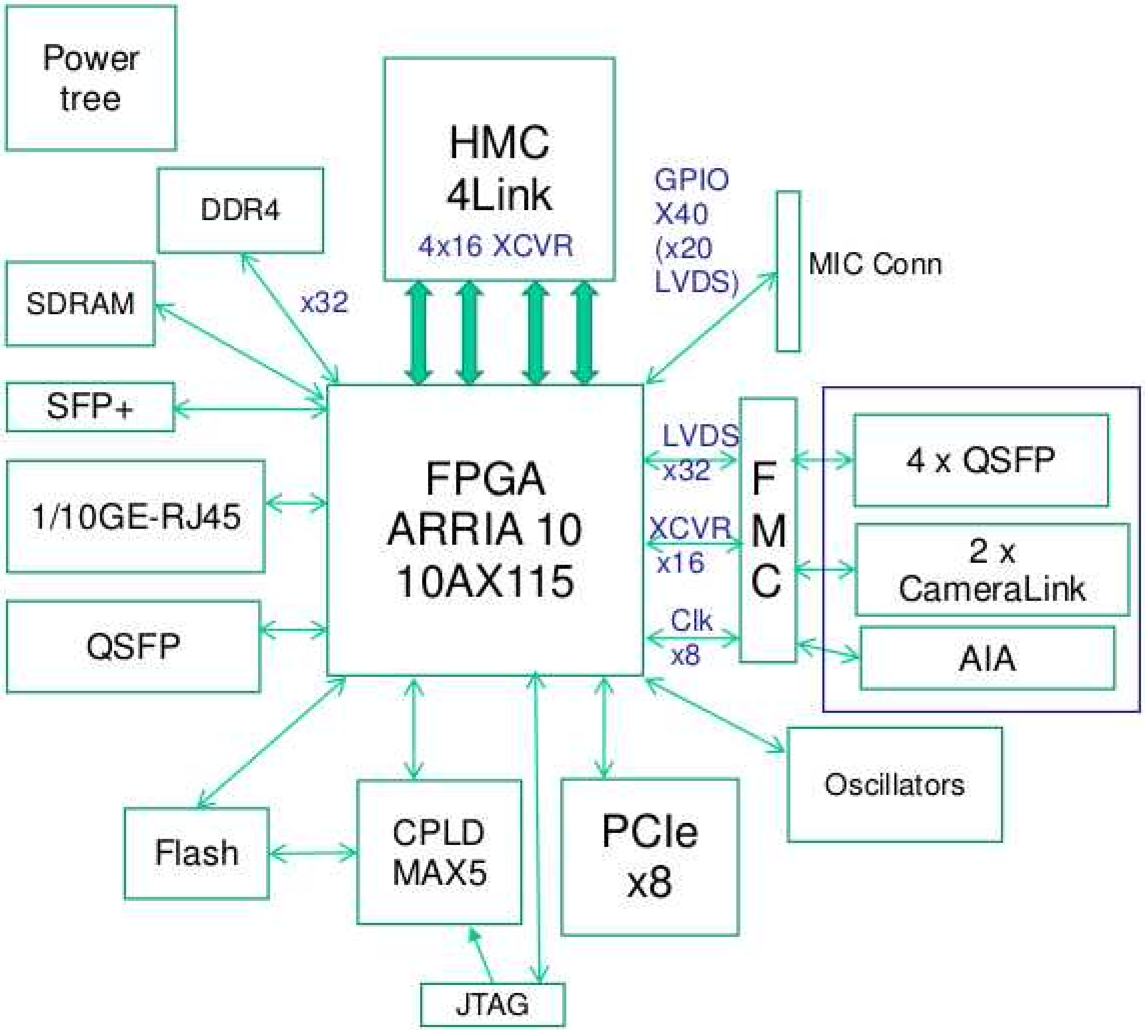




AO RTC concept : data pipeline



FPGA solutions : µXcomp



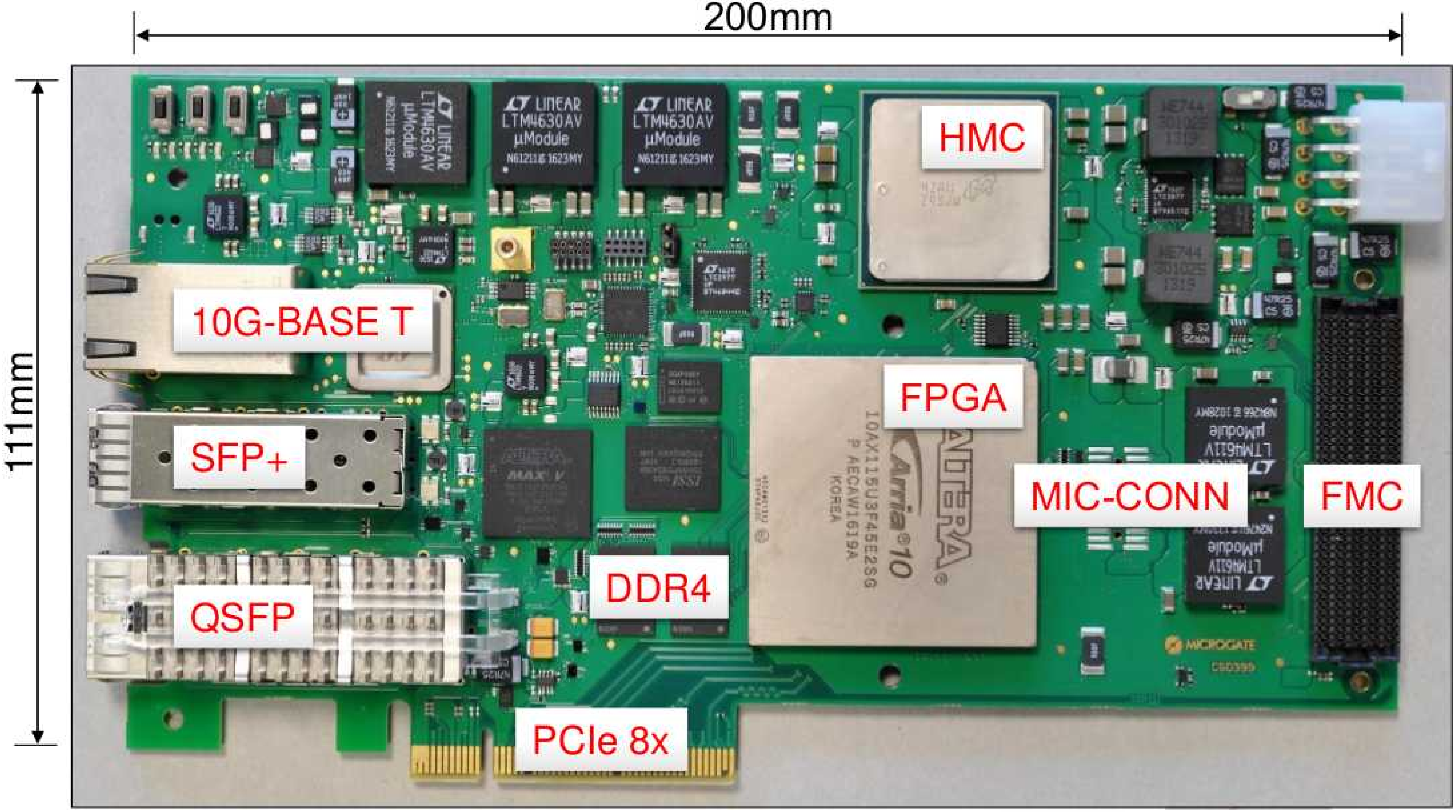
Based on ARRIA 10AX115:

- 1518 DSP blocks
- 6.6MB int. RAM
- 96 XCVR

Board features:

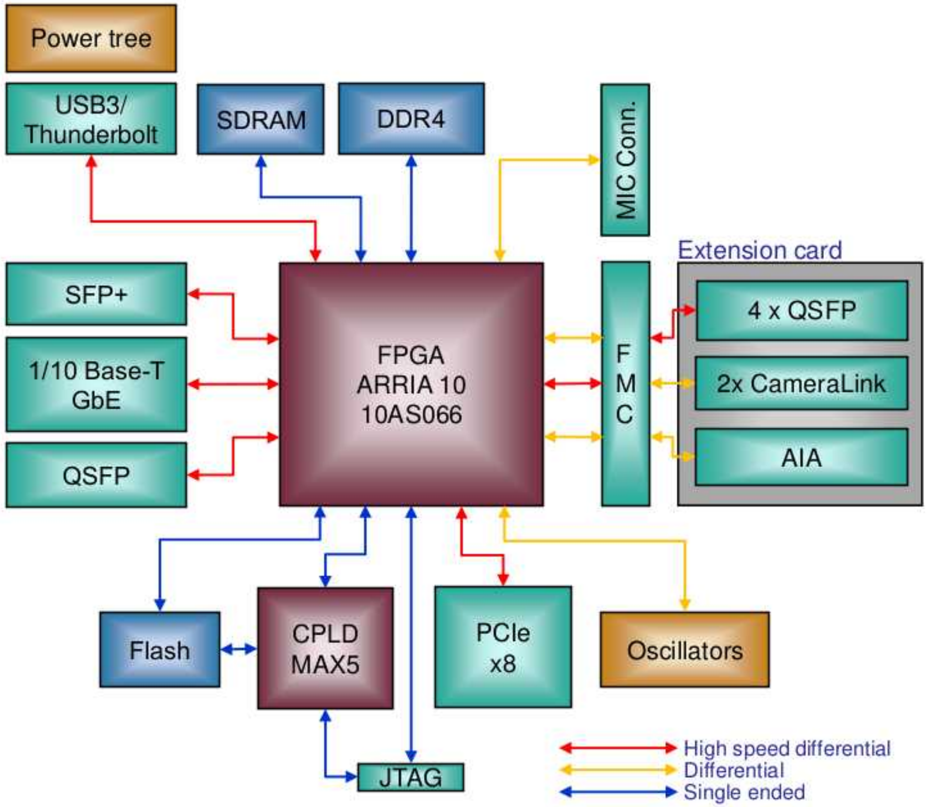
- Optimized for **heavy deterministic computation** in floating-point
- **Large Bandwidth between HMC and FPGA** - 4 links 16 lanes/link up to 15Gbps/lane = 120GB/s bidirectional
- Extremely **low jitter**
- More **power efficient** compared to GPUs
- Offers a lot of different interfaces on board or via the FMC connector and extension cards

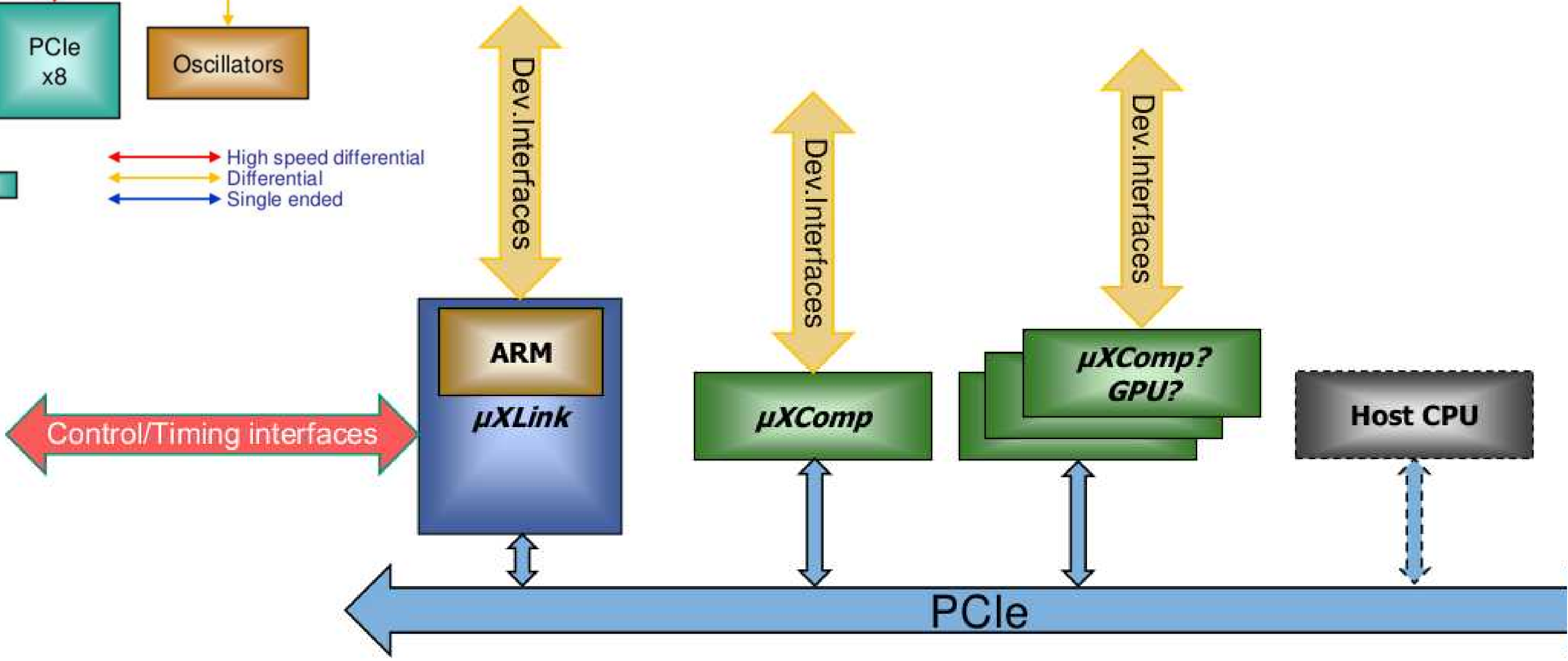
FPGA solutions: status



FPGA solutions: µserver concept

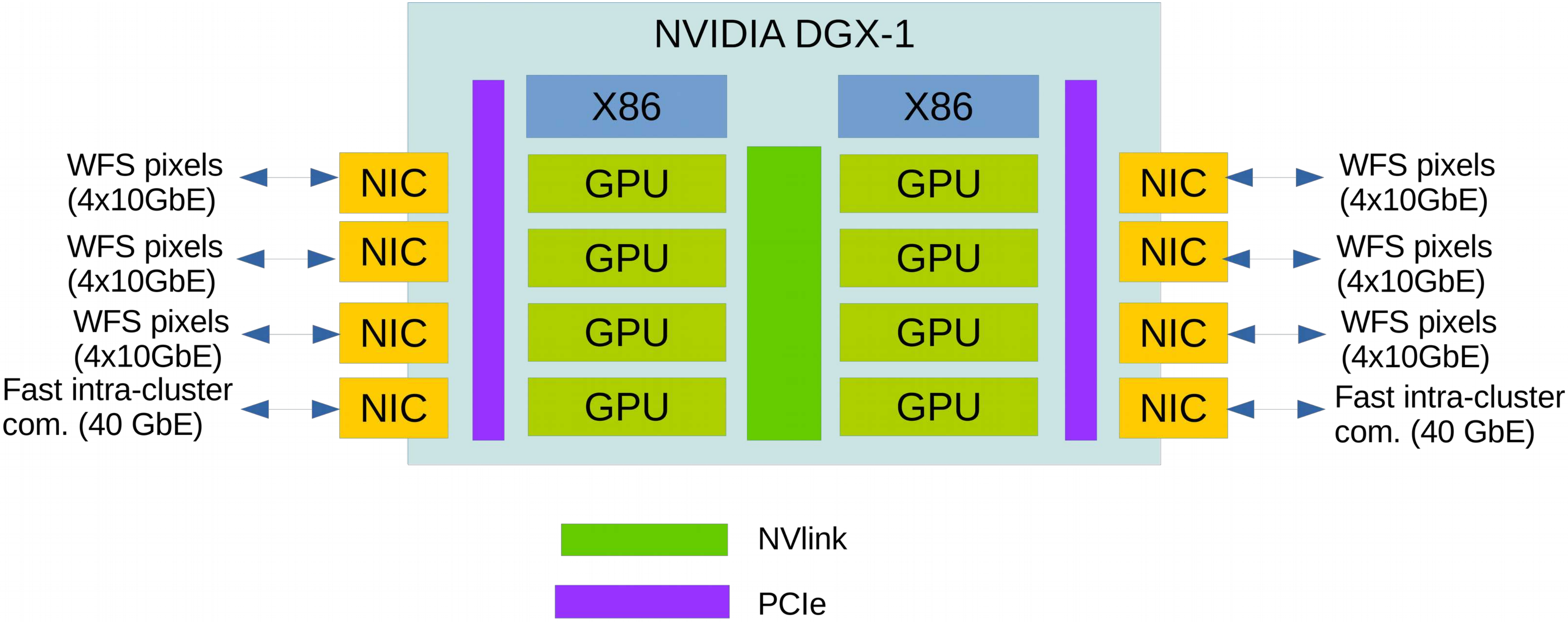
- Based on Arria 10 but no HMC
- To be used as a main unit in a cluster





RT data pipeline with GPUs

- Prototype using latest generation GPU server



NVIDIA DGX-1

X86

X86

WFS pixels (4x10GbE)

NIC

GPU

GPU

NIC

WFS pixels (4x10GbE)

WFS pixels (4x10GbE)

NIC

GPU

GPU

NIC

WFS pixels (4x10GbE)

WFS pixels (4x10GbE)

NIC

GPU

GPU

NIC

WFS pixels (4x10GbE)

Fast intra-cluster com. (40 GbE)

NIC

GPU

GPU

NIC

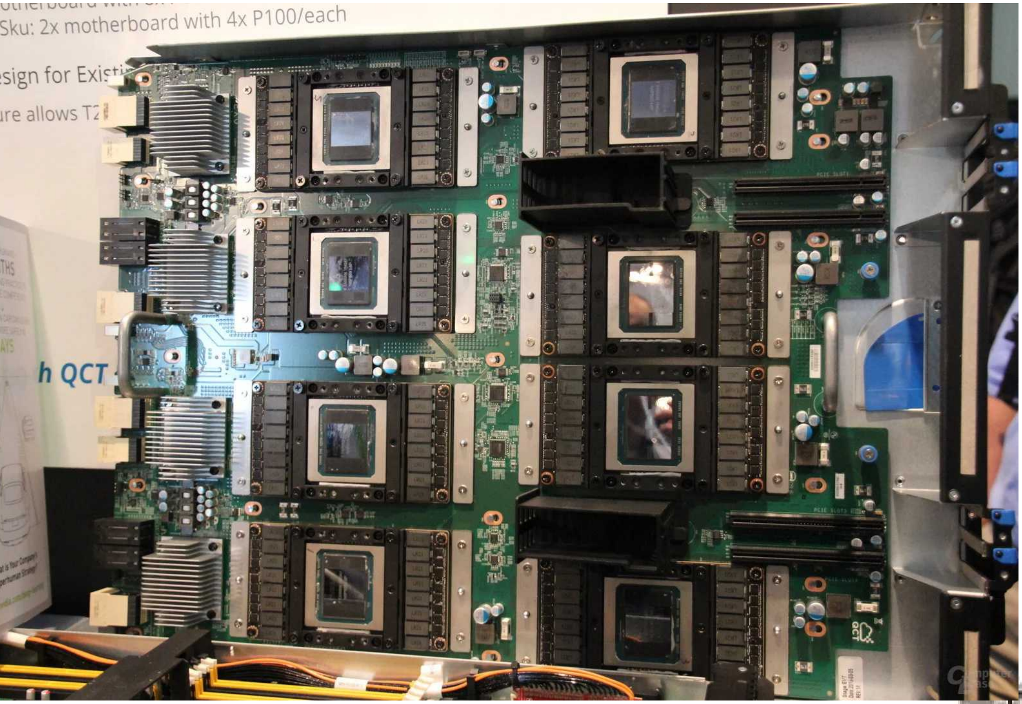
Fast intra-cluster com. (40 GbE)

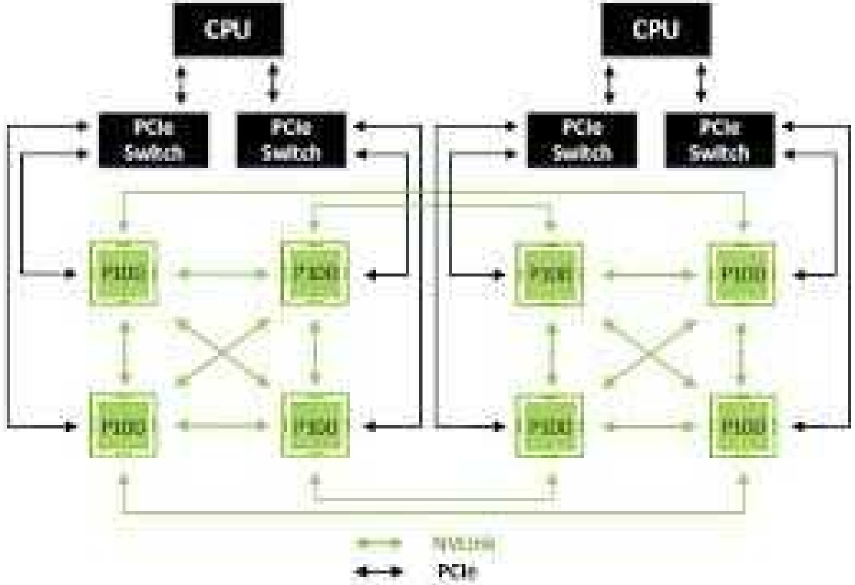
🟩 NVlink

🟪 PCIe

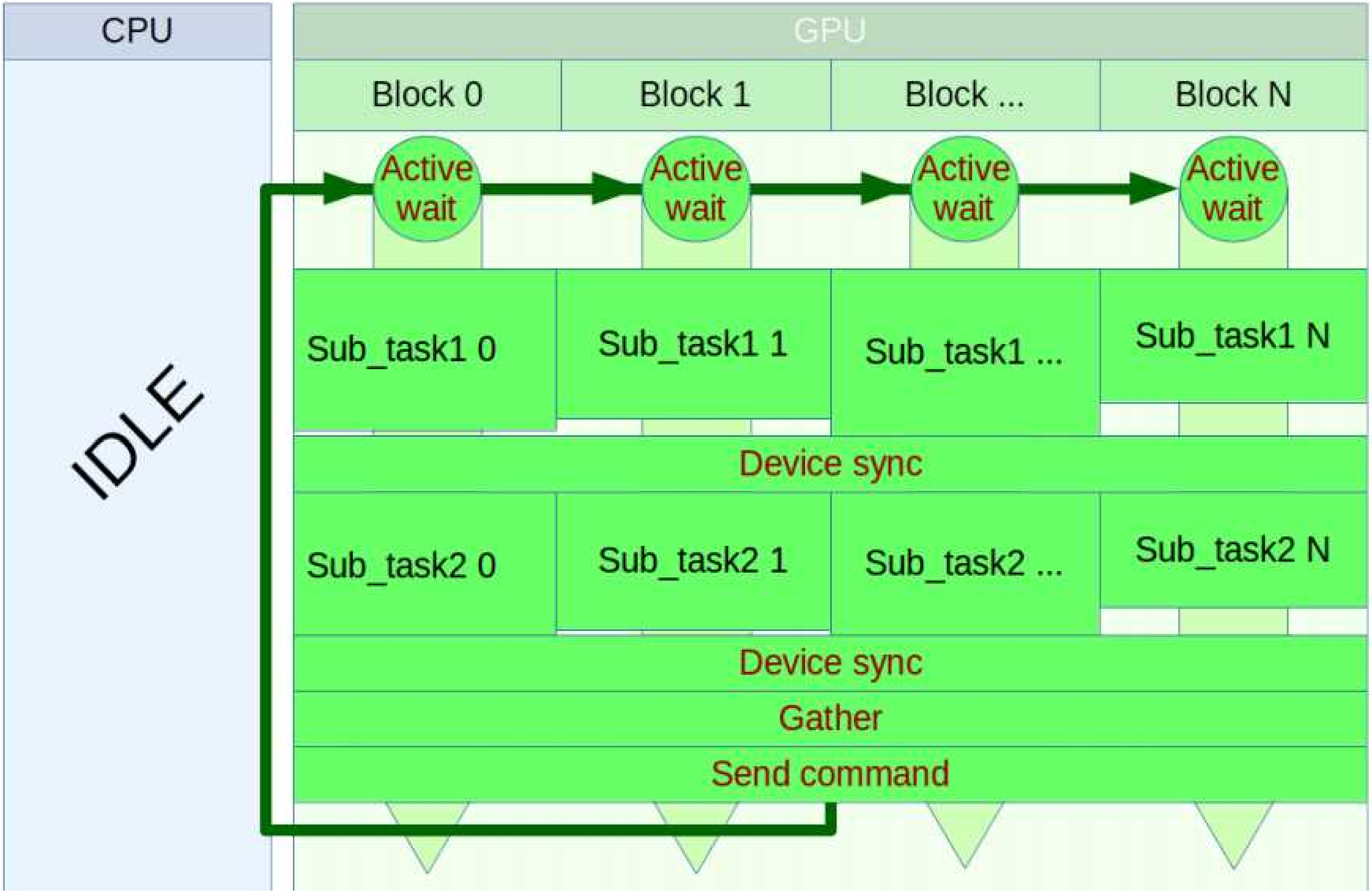
RT data pipeline with GPUs

- Prototype using latest generation GPU server

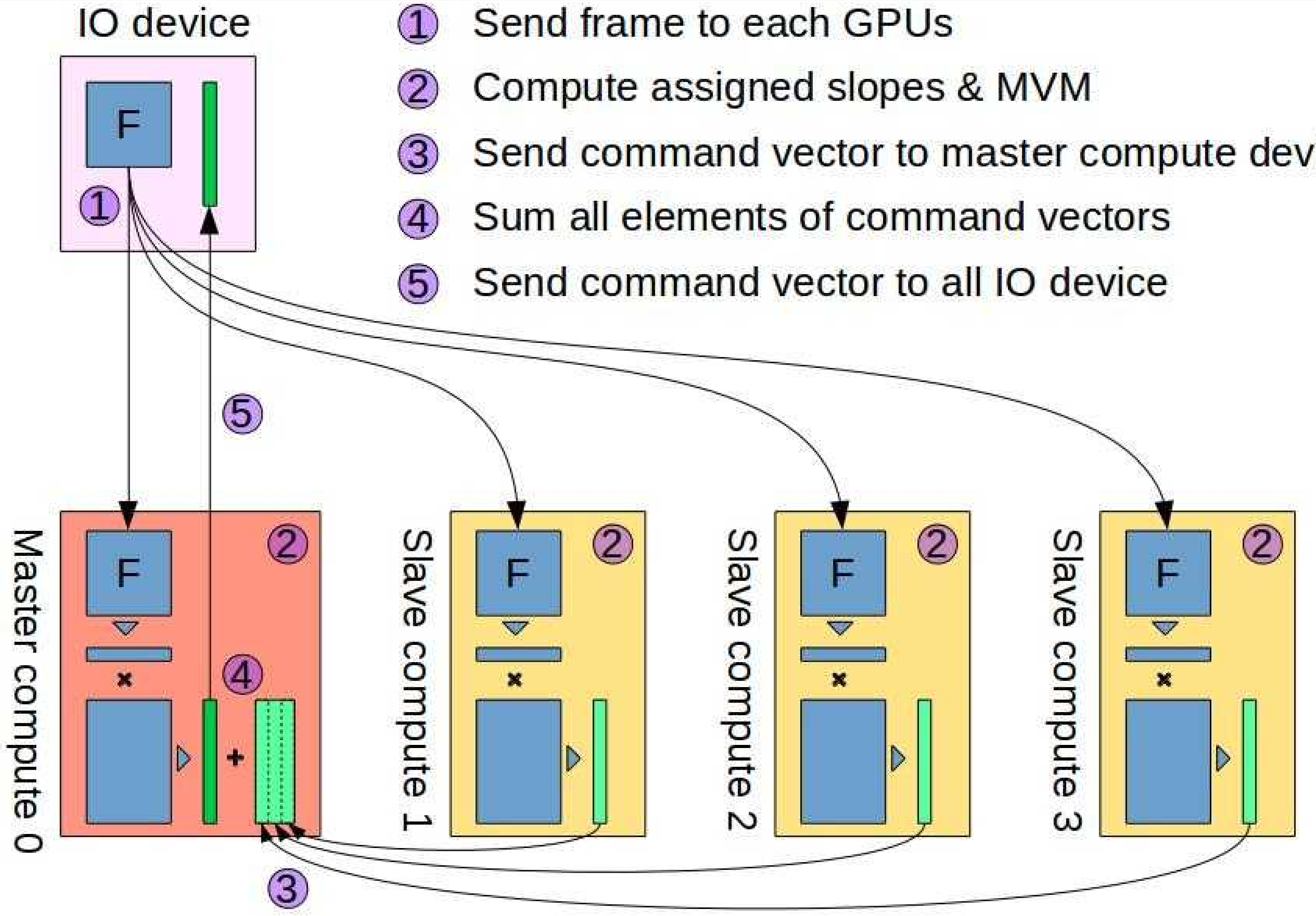




Persistent kernels

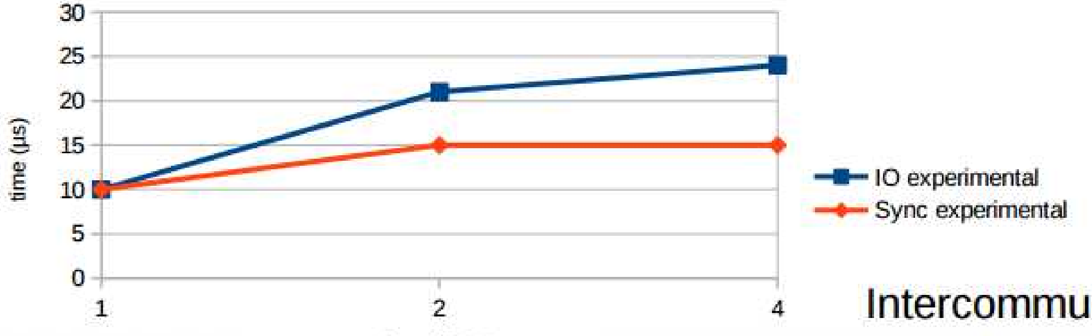


Multi-GPU prototype

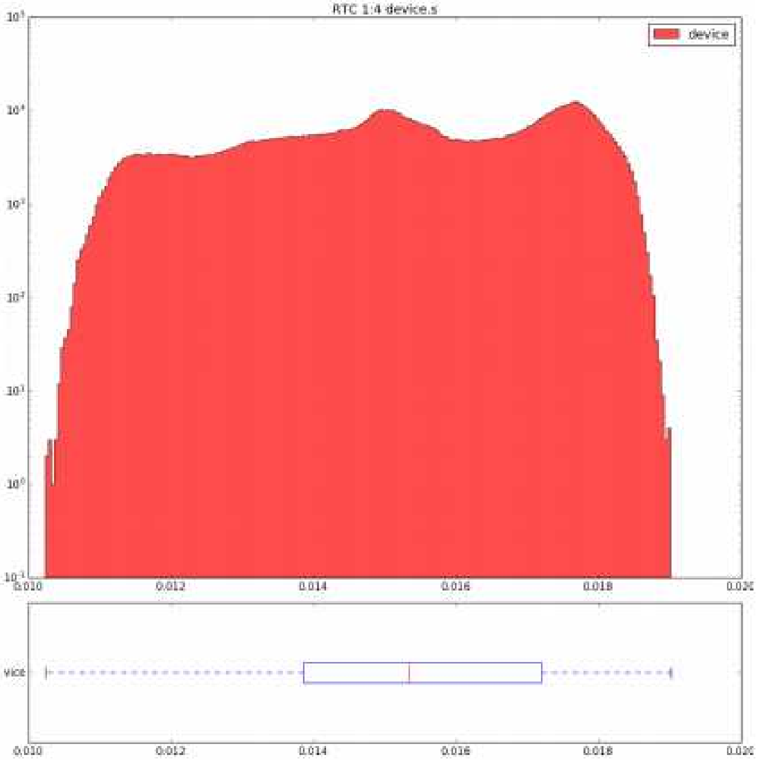


1. Send frame to each GPUs
2. Compute assigned slopes & MVM
3. Send command vector to master compute device
4. Sum all elements of command vectors
5. Send command vector to all IO device

Persistent kernels

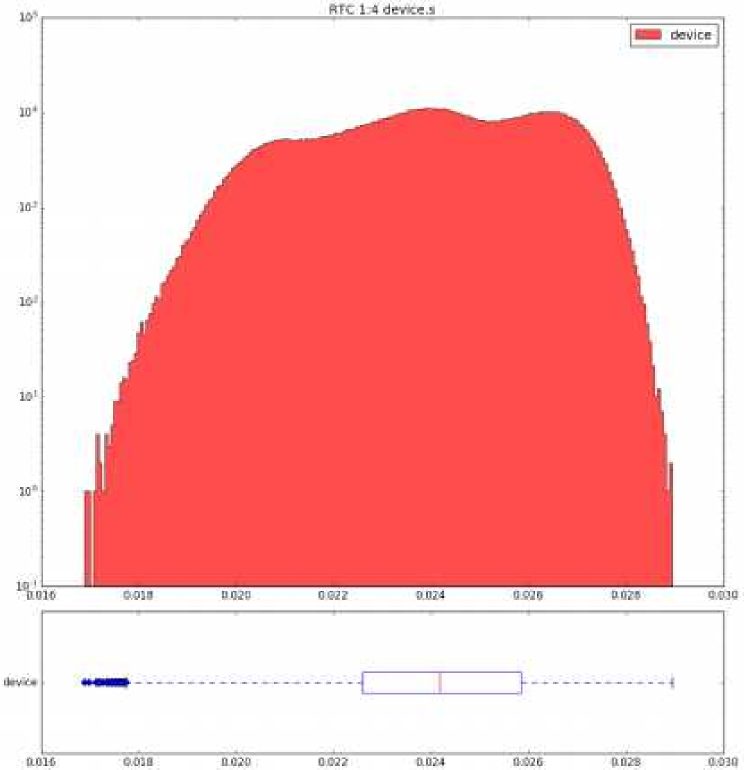


Synchronize jitter



Average : 15µs Jitter : 8.8µs

Intercommunication jitter

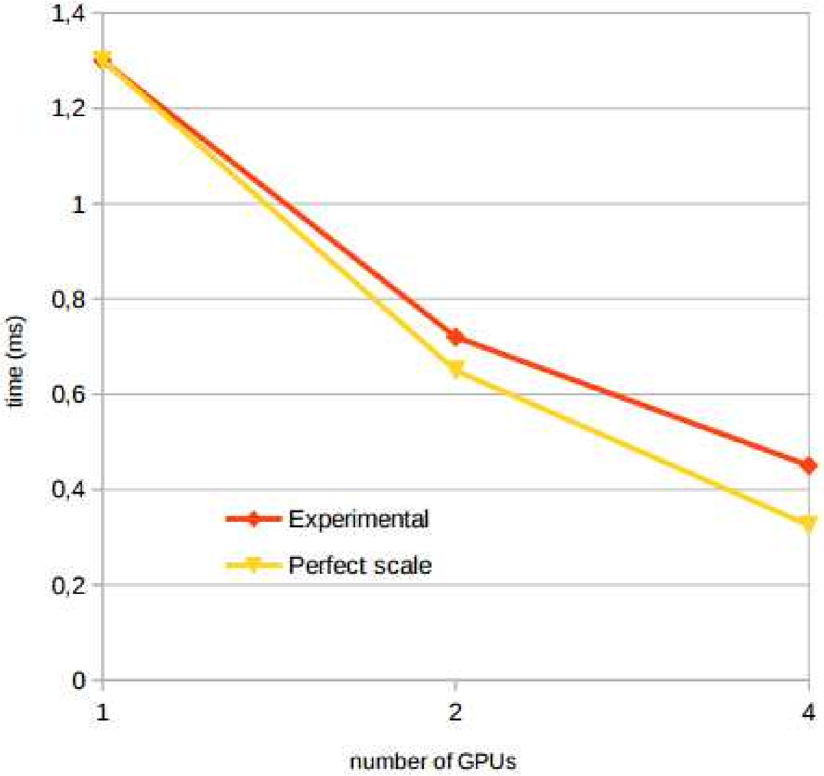


Average : 24µs Jitter : 12µs

Persistent kernels

Strong scalability

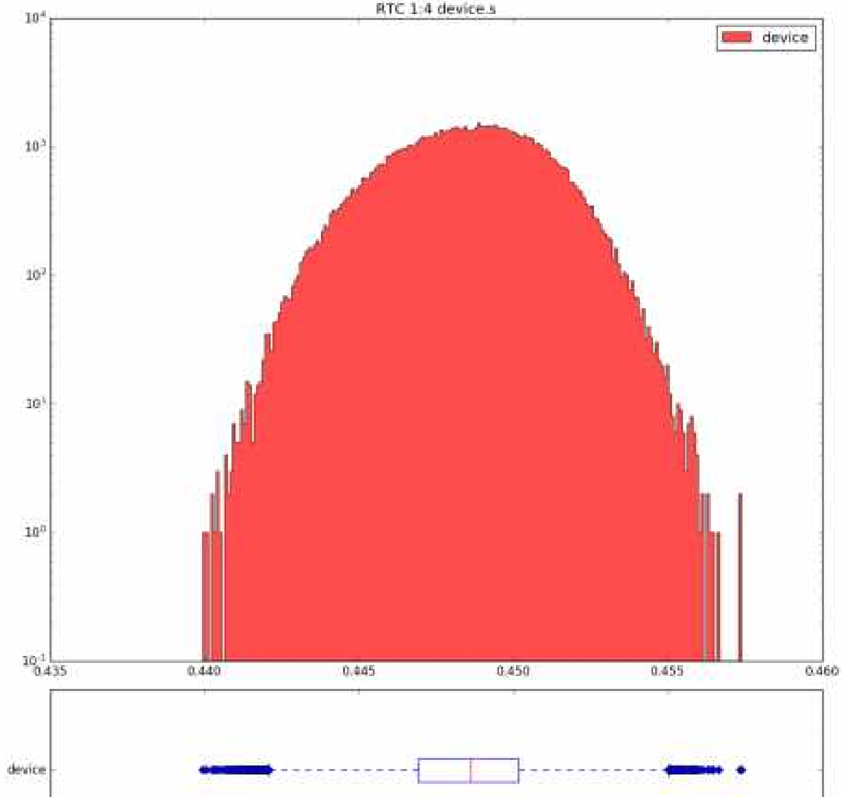
Constant case with **10,048 slopes x 15,000 commands**



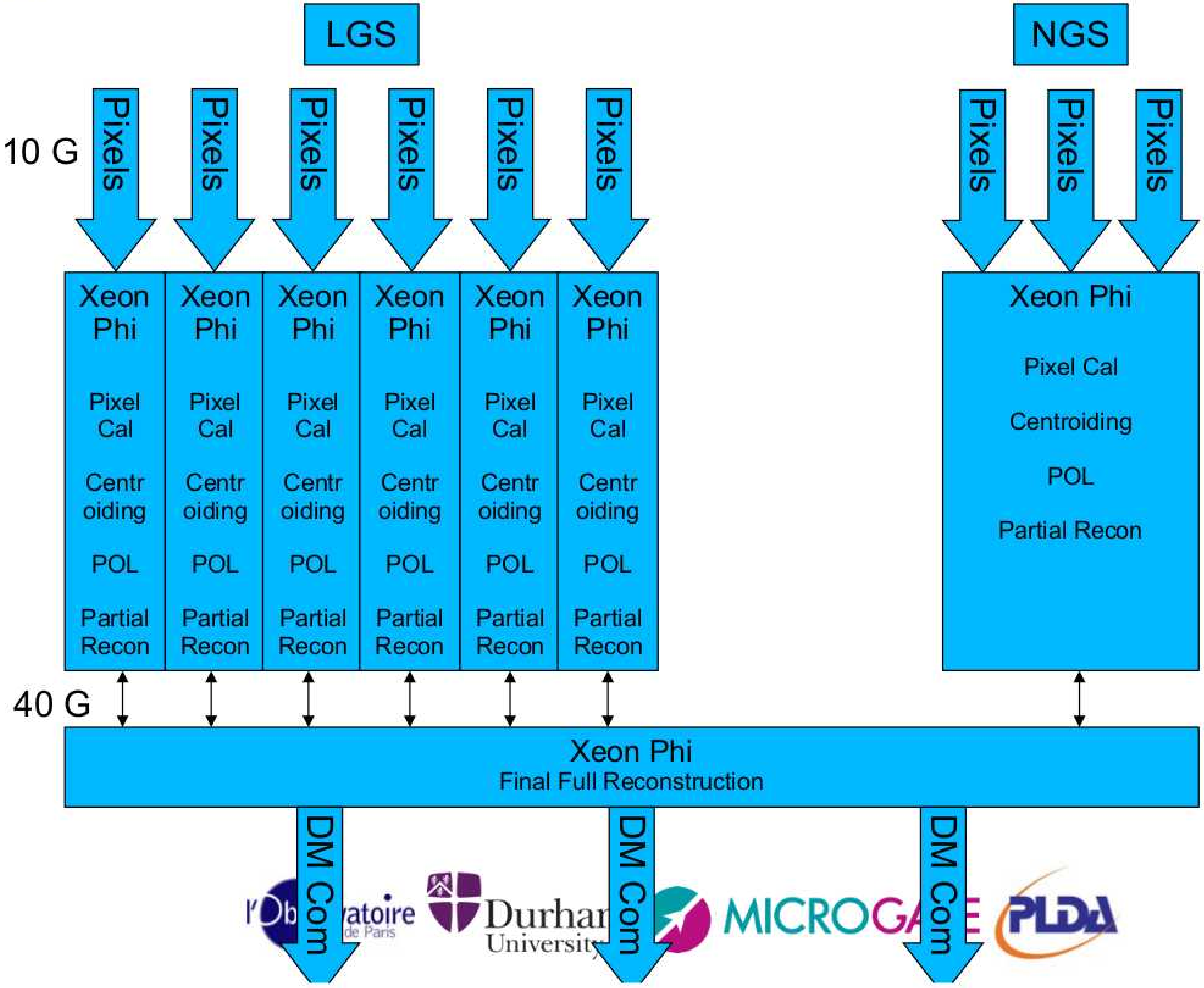
Histogram

Case with **10,048 slopes x 15,000 commands** on 4 devices

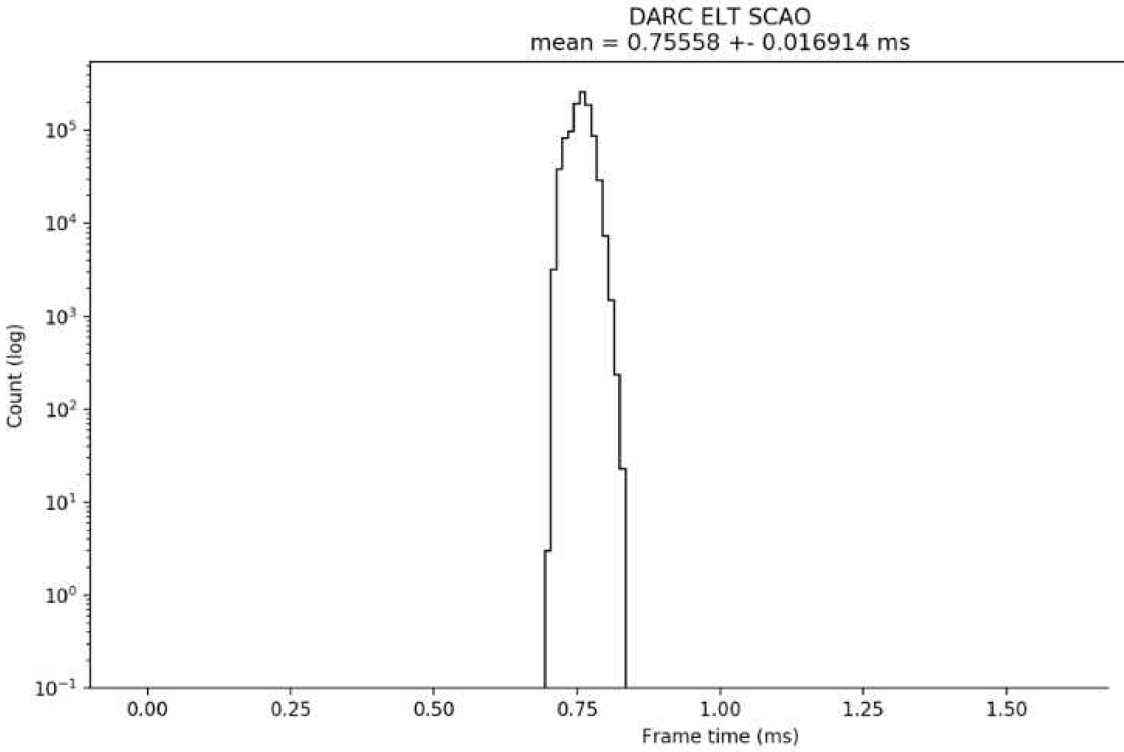
Average : 0.45ms   Jitter : 17µs



RT data pipeline with Xeon Phi

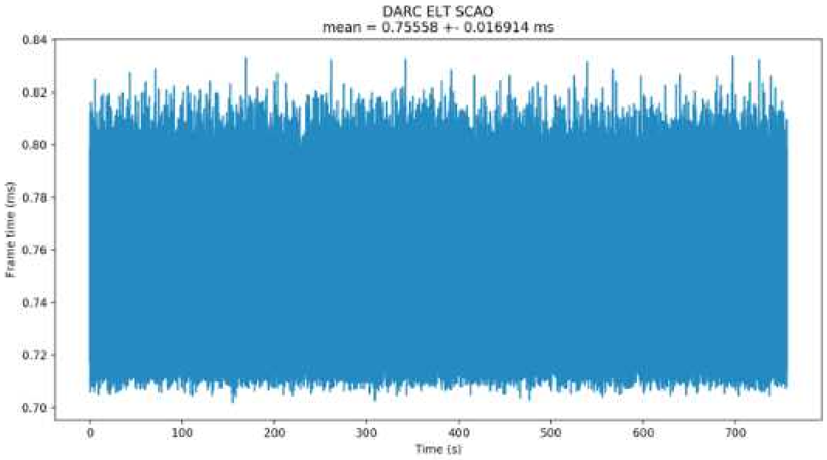


Xeon Phi solution

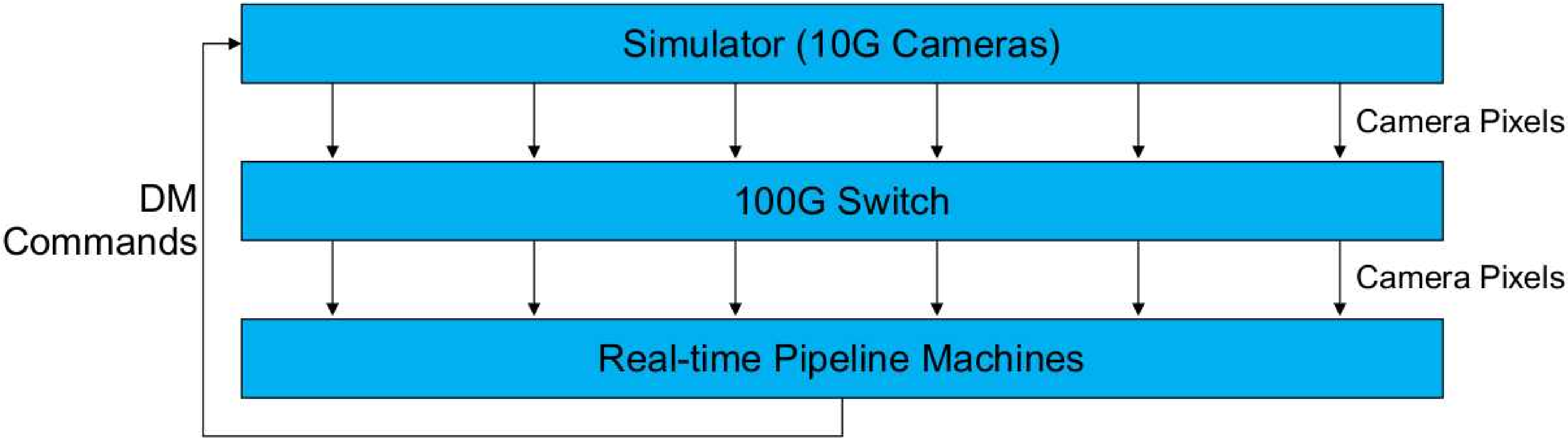


SCAO MVM on single Xeon Phi, Knights Landing

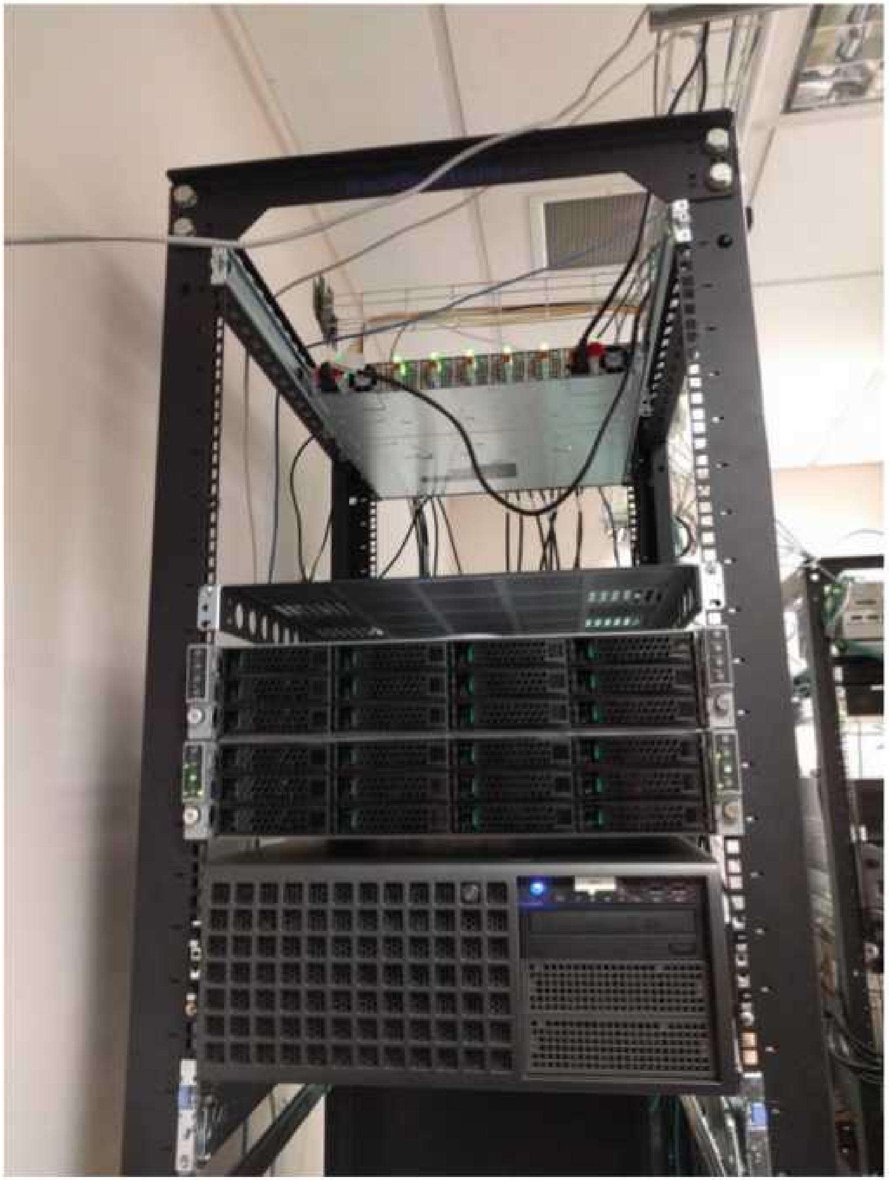
- 1 million iterations
- (~10 minutes)
- 750 +- 17μs



Xeon Phi prototype



Xeon Phi testing facility



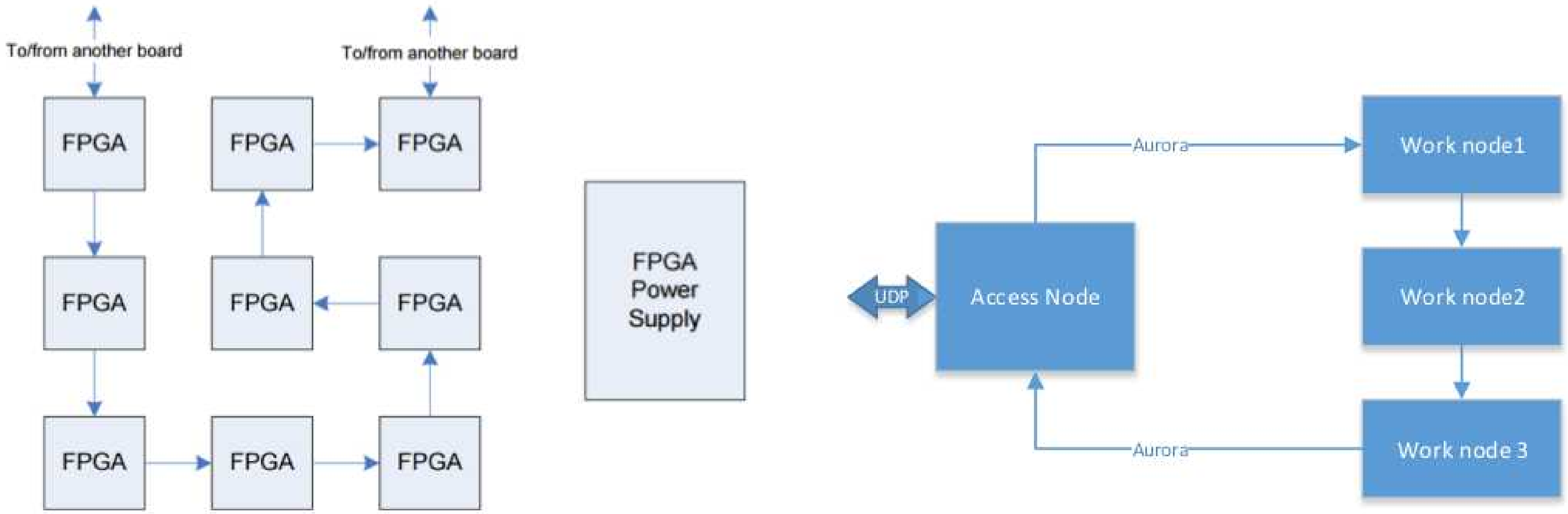
100G Switch

Xeon Phi Cluster

Simulator

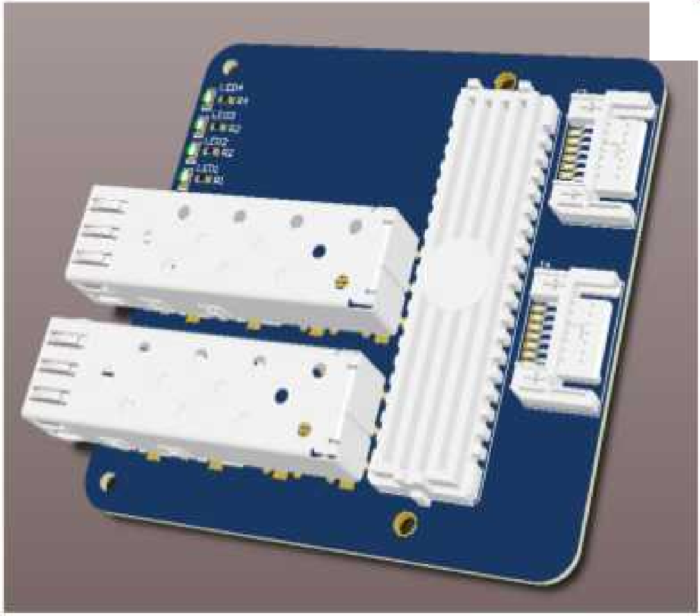
RT data pipeline with COTS FPGA

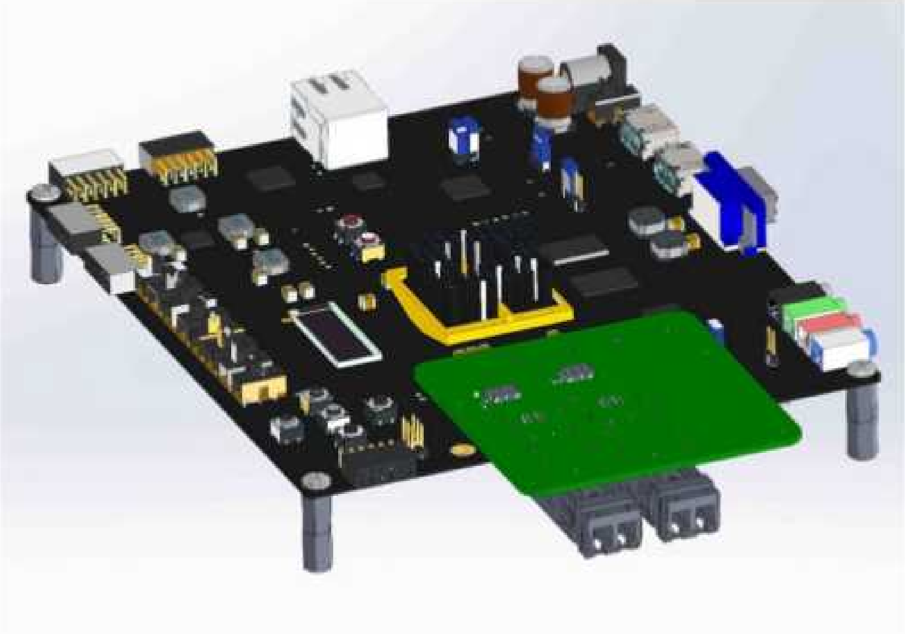
COTS FPGA cluster



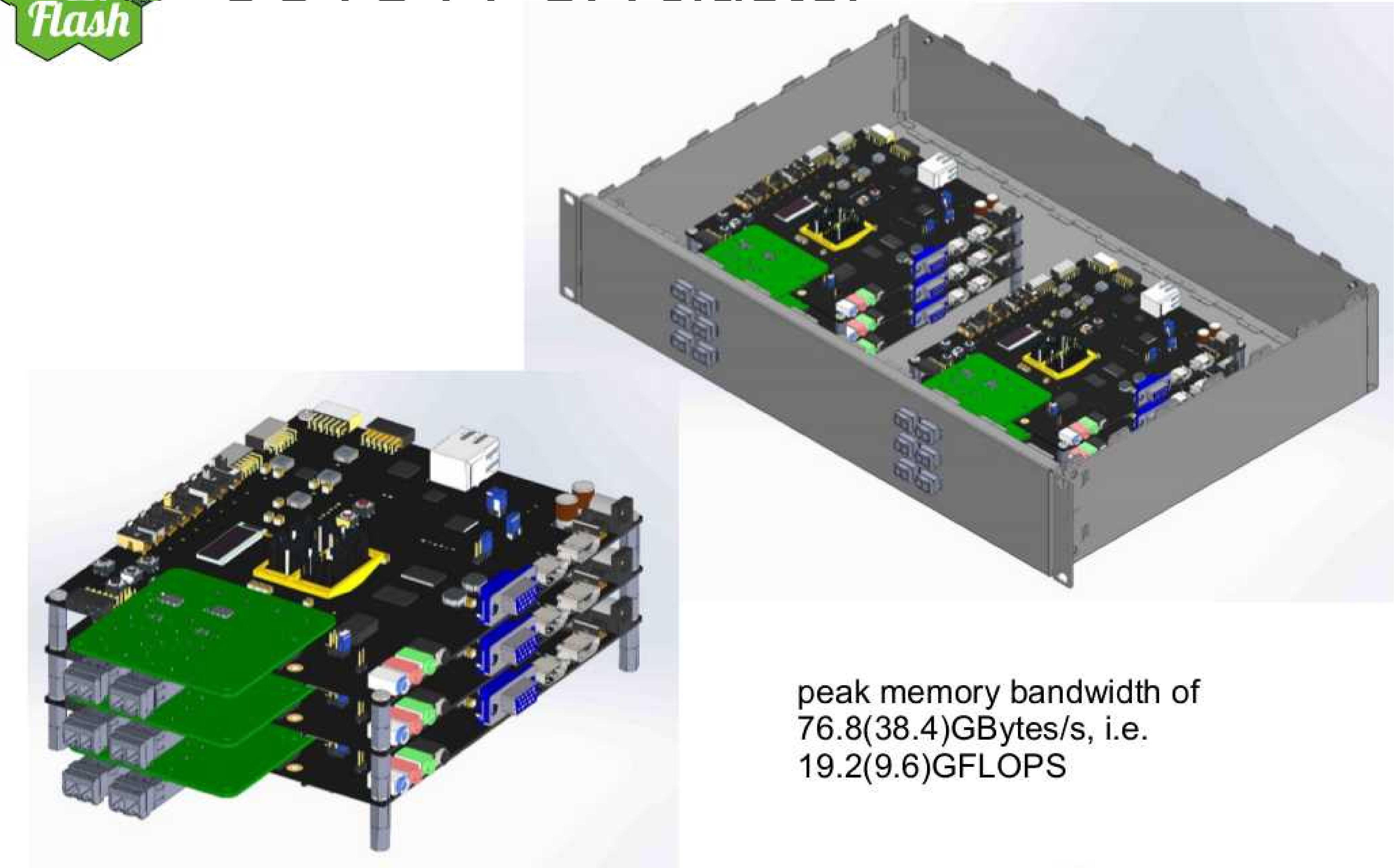
FMC to 10GbE

- 2 SFP+ 10GbE interface
- 2 SATA like internal connection interface
- Based on the FMC HPC connection



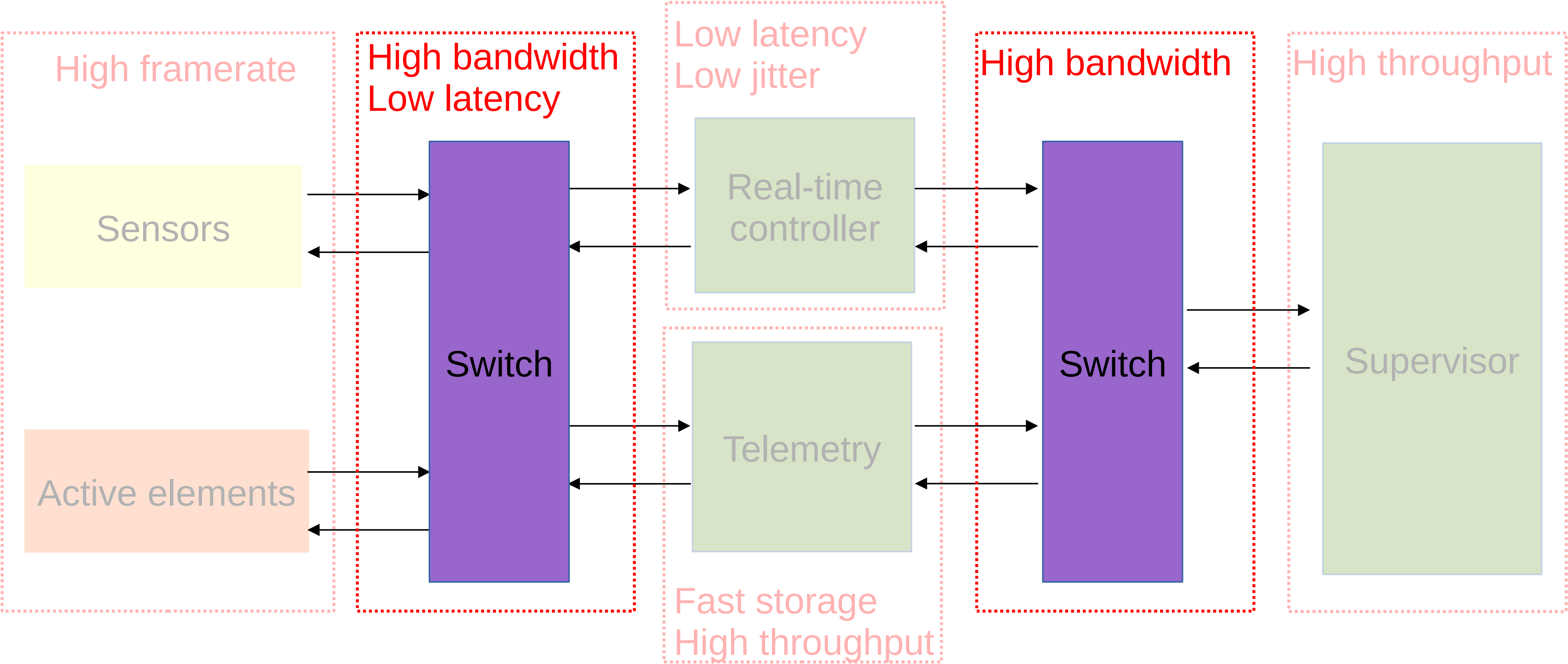


COTS FPGA cluster

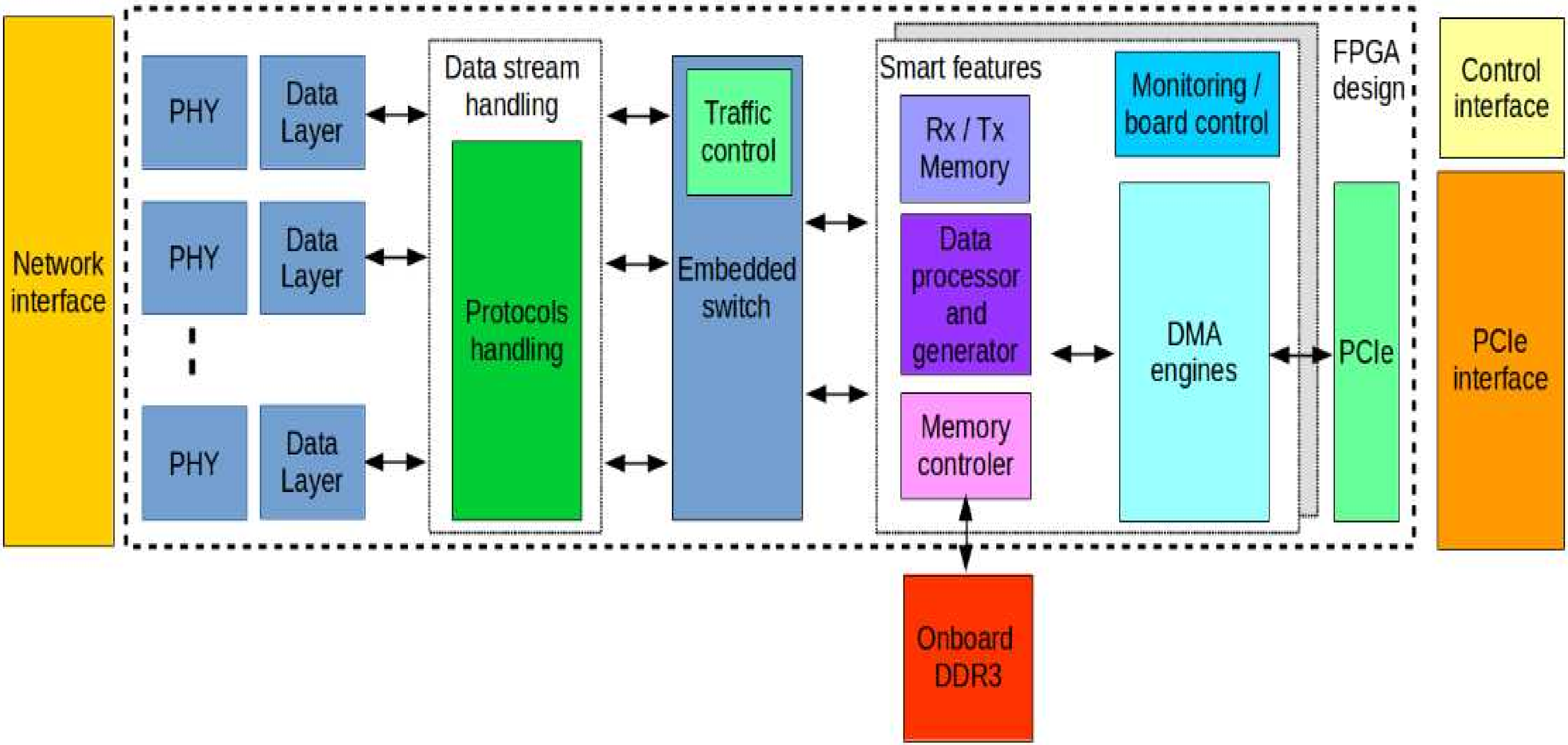


peak memory bandwidth of
76.8(38.4)GBytes/s, i.e.
19.2(9.6)GFLOPS

AO RTC concept : smart interconnect

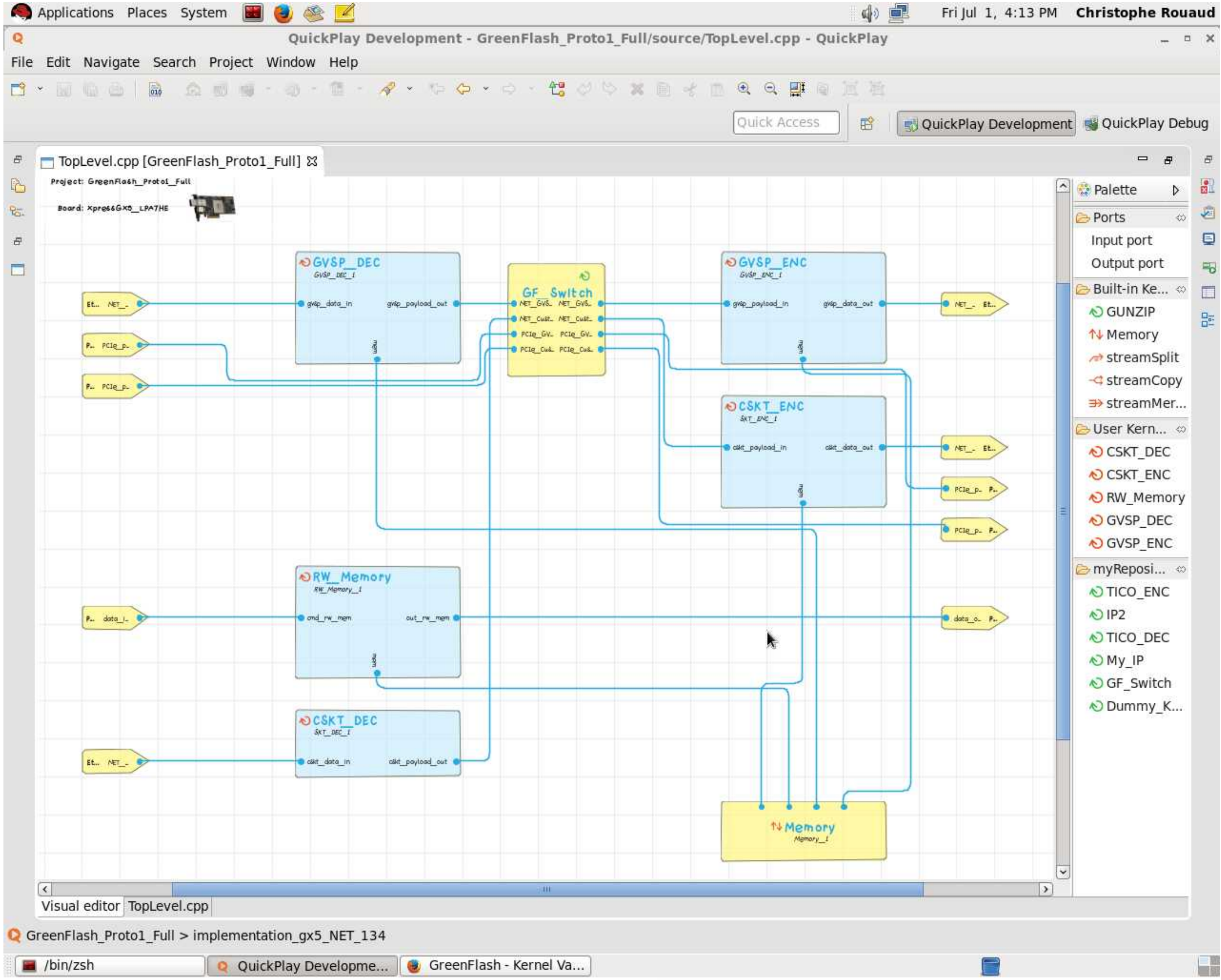


Smart interconnect concept



Smart interconnect concept

- Eased devel. process using the QuickPlay tool from PLDA

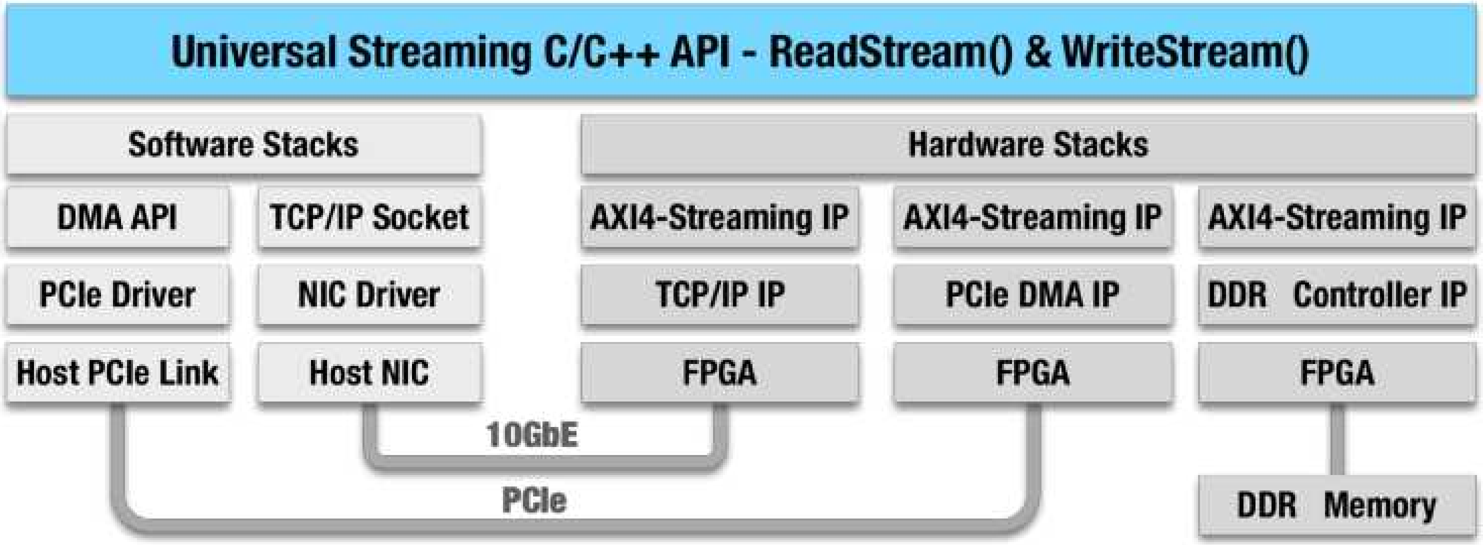


QuickPlay™

QuickStore™

QuickPlay

QuickPlay™ Hardware Accelerator Abstraction Layer



Universal Streaming C/C++ API - ReadStream() & WriteStream()

Software Stacks

DMA API

TCP/IP Socket

PCIe Driver

NIC Driver

Host PCIe Link

Host NIC

Hardware Stacks

AXI4-Streaming IP

TCP/IP IP

FPGA

AXI4-Streaming IP

PCIe DMA IP

FPGA

AXI4-Streaming IP

DDR Controller IP

FPGA

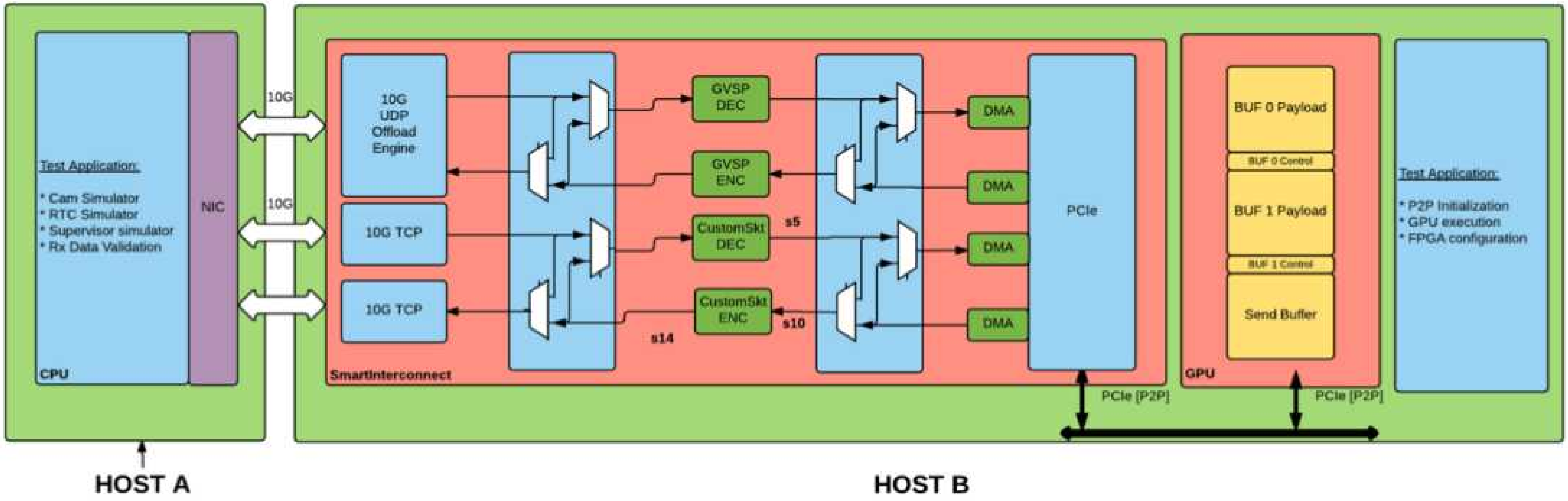
DDR Memory

10GbE

PCIe

Smart interconnect prototype

- Link with high level API / application



Smart interconnect prototype

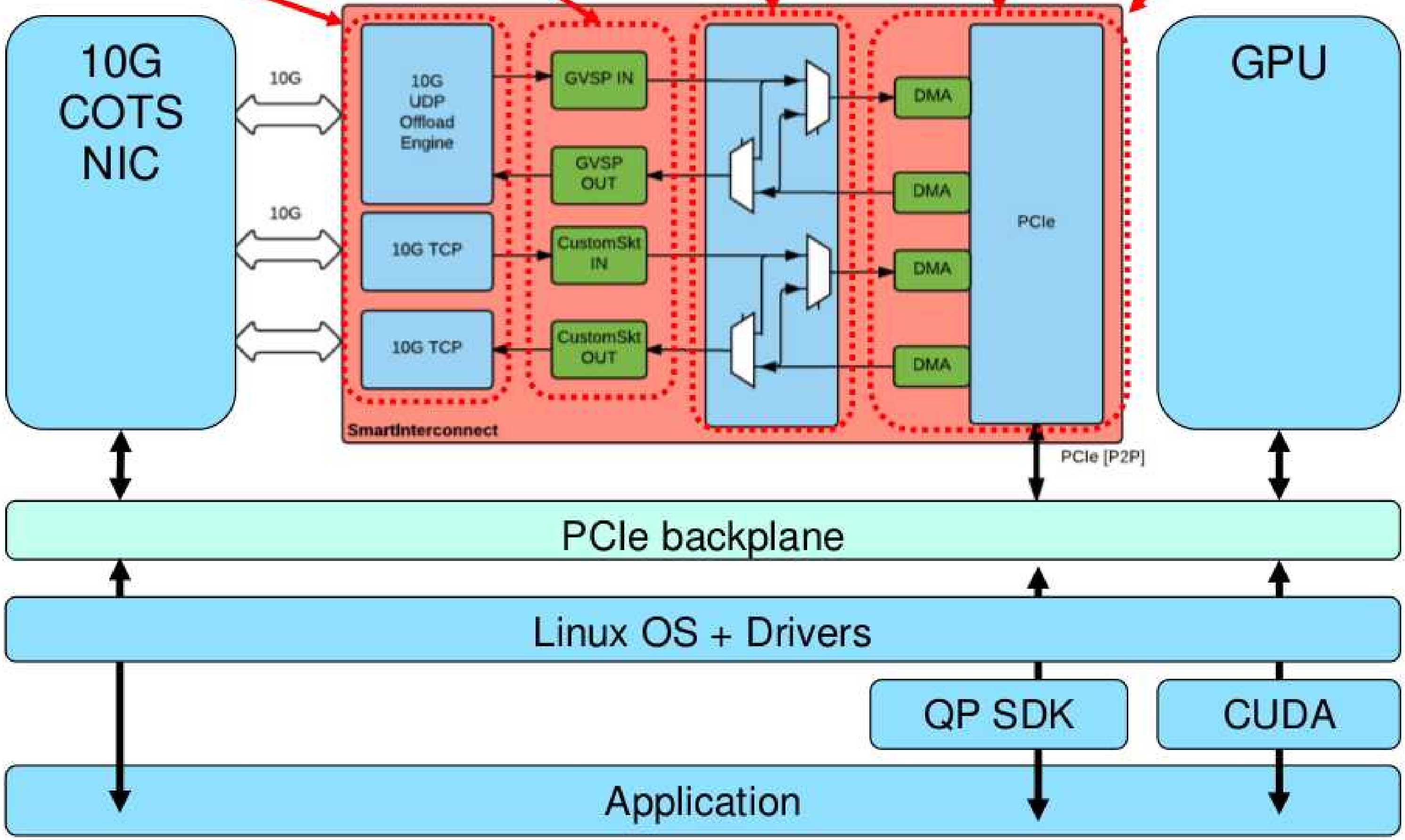
3x10G Ethernet ports with TCP and UDP offloading IP Cores

GeV partial support (GVSP) and Custom Socket support (CSKT)

Elementary Switch (HDL)

Configurable PCIe with P2P capability

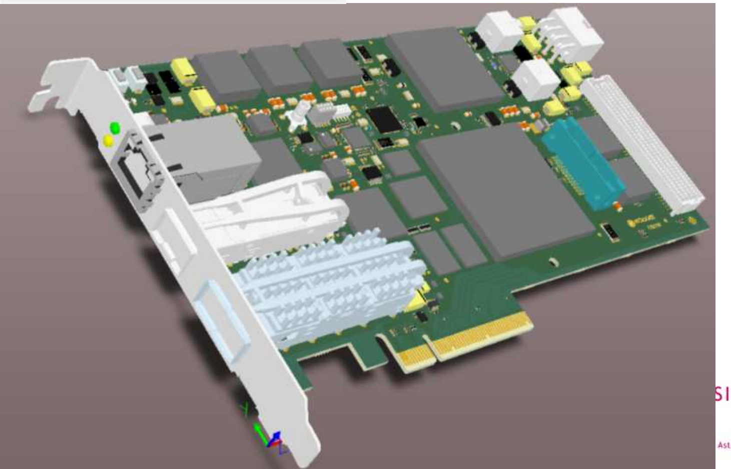
XpressGX5 XpressKUS boards supported



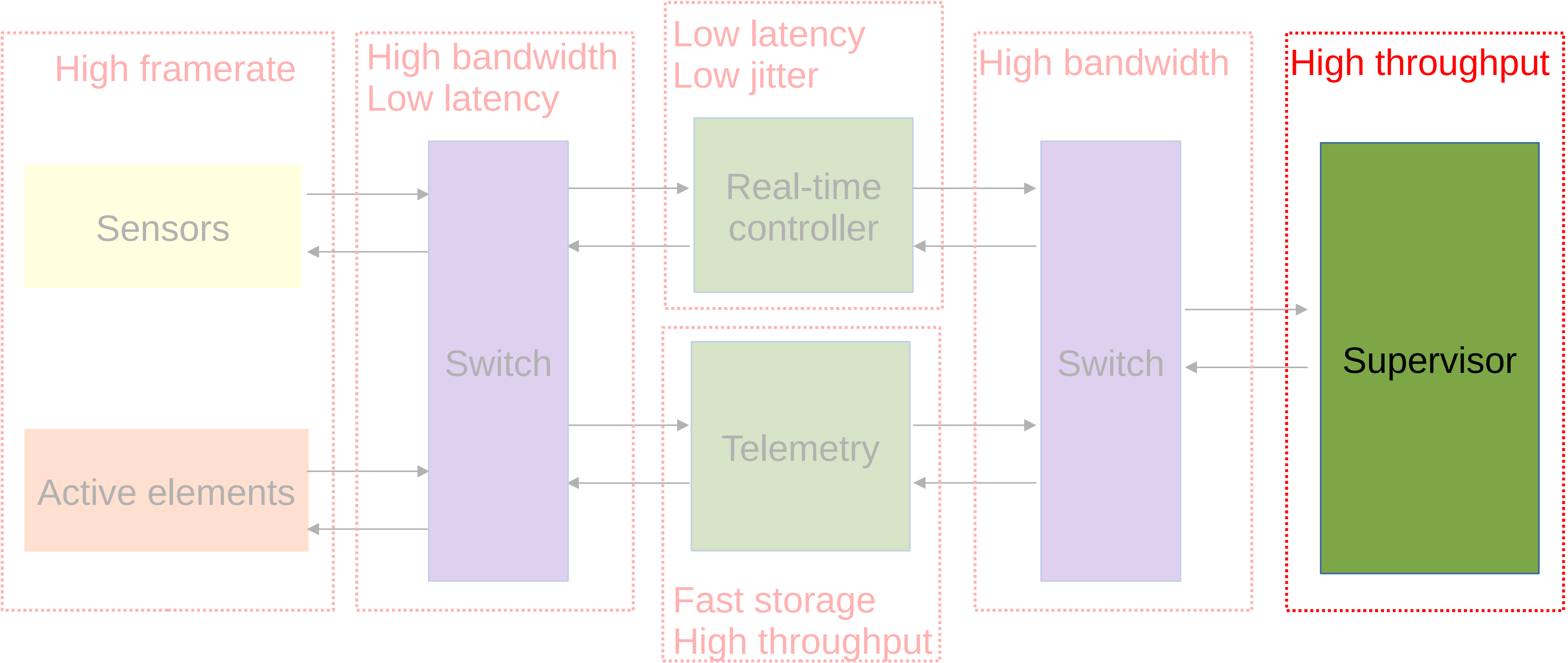
Smart interconnect prototyping

- Single generic design / multiple target boards

FPGA family	Board name
Stratix V	Reflex XpressGX5
Kintex-7	Reflex XpressK7 160/325 ***(v2.0)***
	Xilinx KC705 ***(v2.1)***
Kintex UltraScale	Reflex XpressKUS ***(v2.0)***
	Xilinx KCU105 ***(v2.1)***
Arria10	Microgate µXComp ***(2017.5)***
	Reflex XpressGXA10 ***(2017.5)***
	Bittware A10PL4 ***(2017.5)***

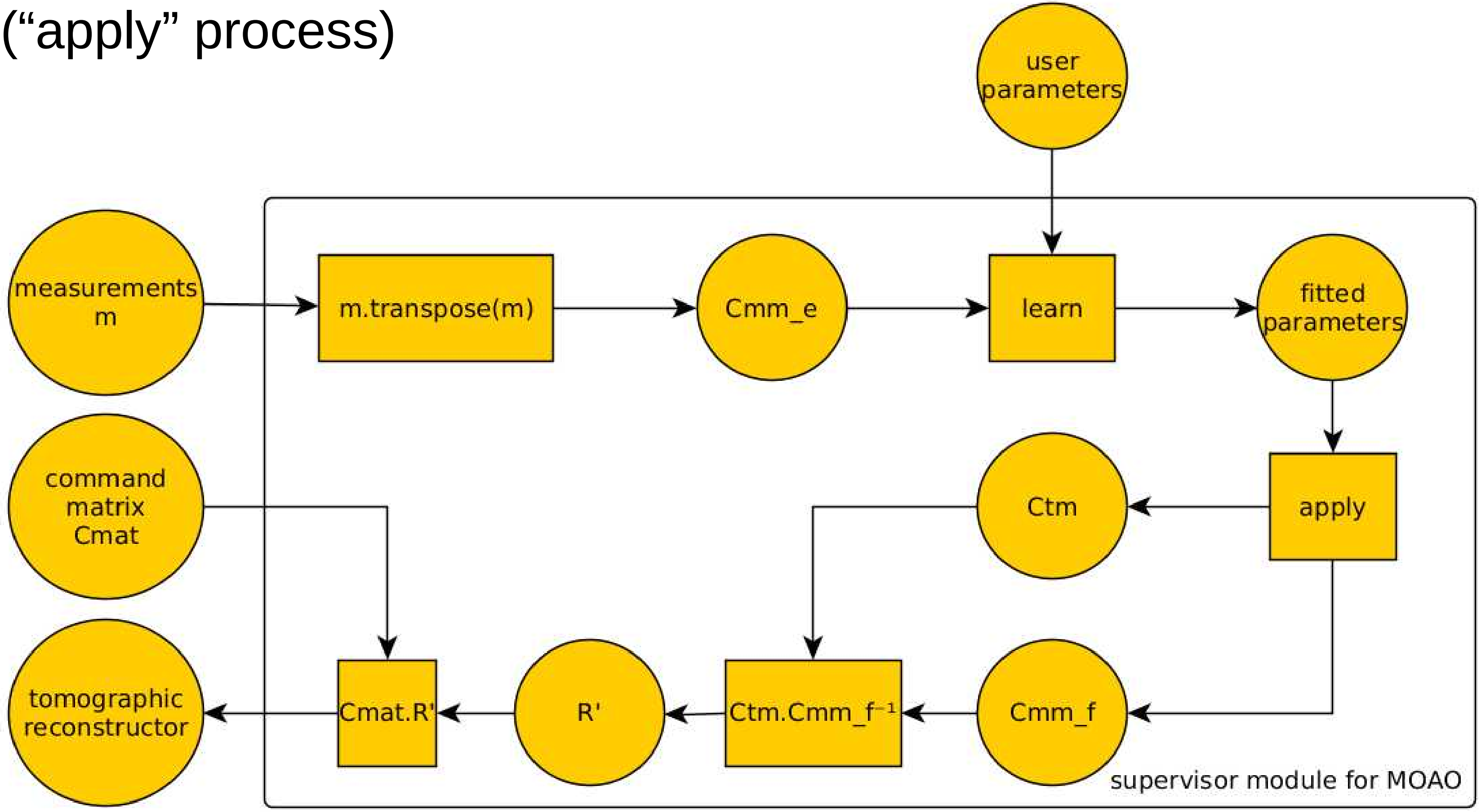


AO RTC concept : supervisor

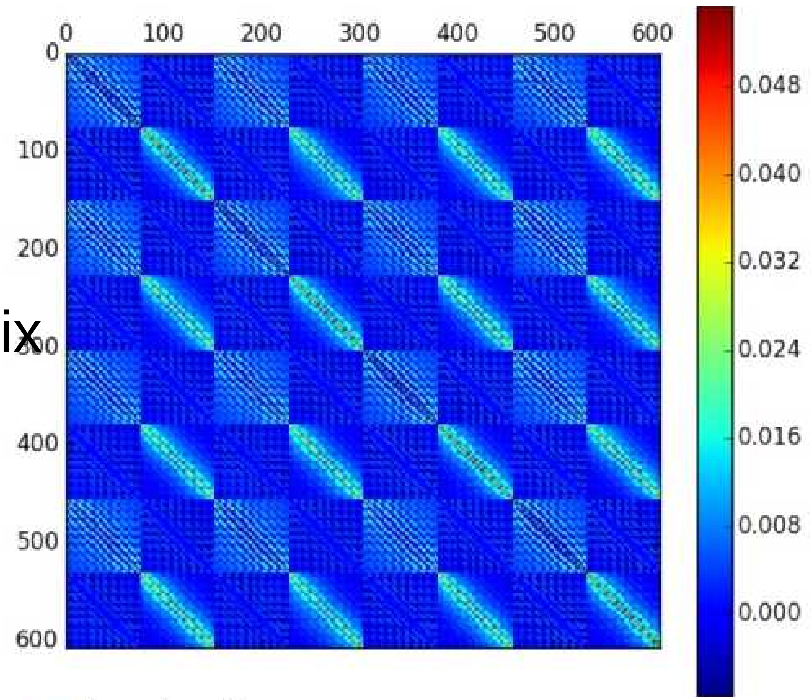


Loop supervision module

Mix of cost function optimization for parameters identification (“Learn” process) and linear algebra for reconstructor matrix computation (“apply” process)



Loop supervision module

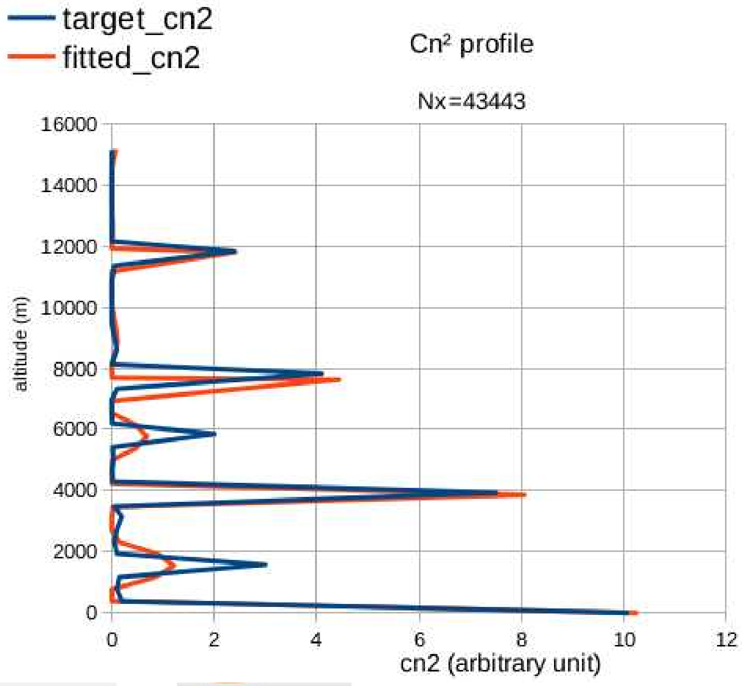


Parameters identification (“Learn” process)

- Fitting measurements covariance matrix on a model including system and turbulence parameters
- Using a score function

$$F(x)=\sum_{k=1}^{N^2}\left[Cmm_k-f_k(x)\right]^2$$

- Levenberg-Marquardt algorithm for function optimization
- Exemple of turbulence profile reconstruction
- Dual stage process (5 layers + 40 layer



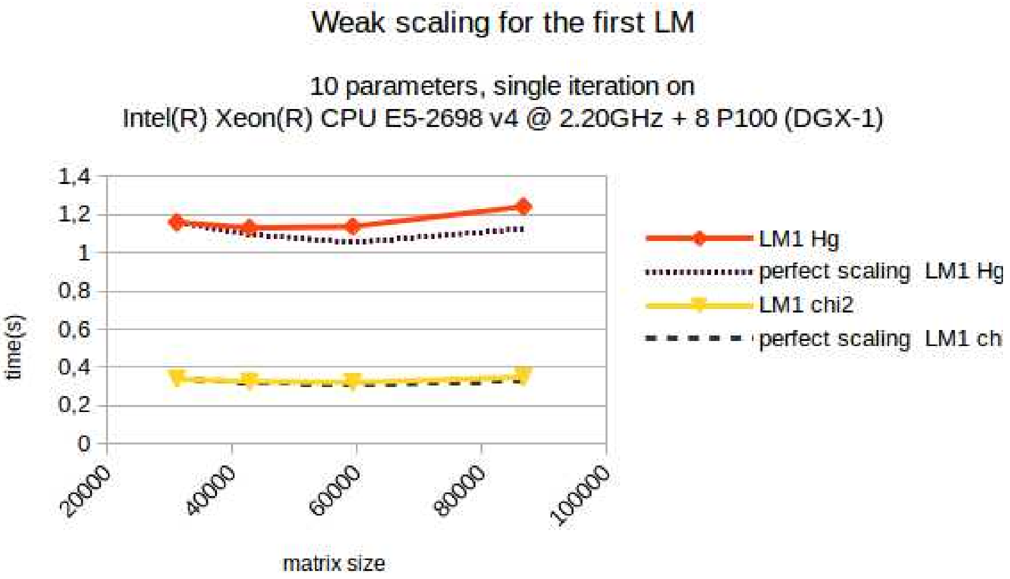
Loop supervision module

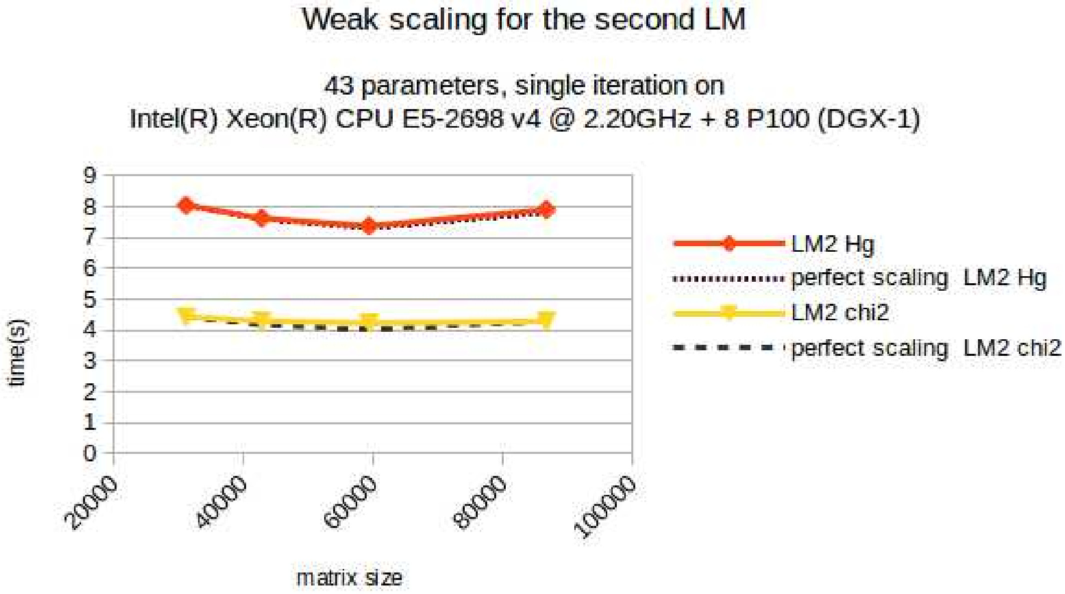
Performance for parameters identification ("Learn" process)

Multi-GPU process, including matrix generation and LM fit

Time to solution for a matrix size of 86k :240s (4 minutes)

- first pass (5 layers) : 25s
- Second pass (40 layers) : 213s





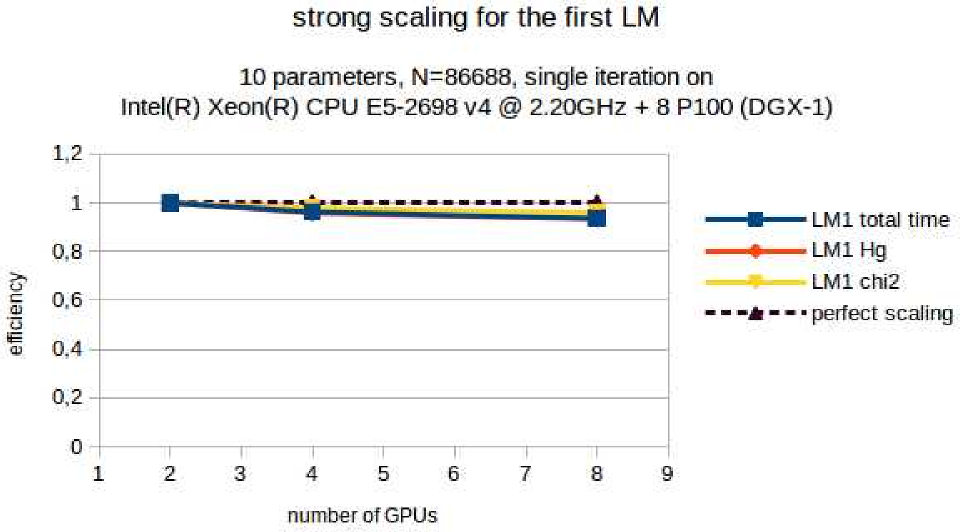
Loop supervision module

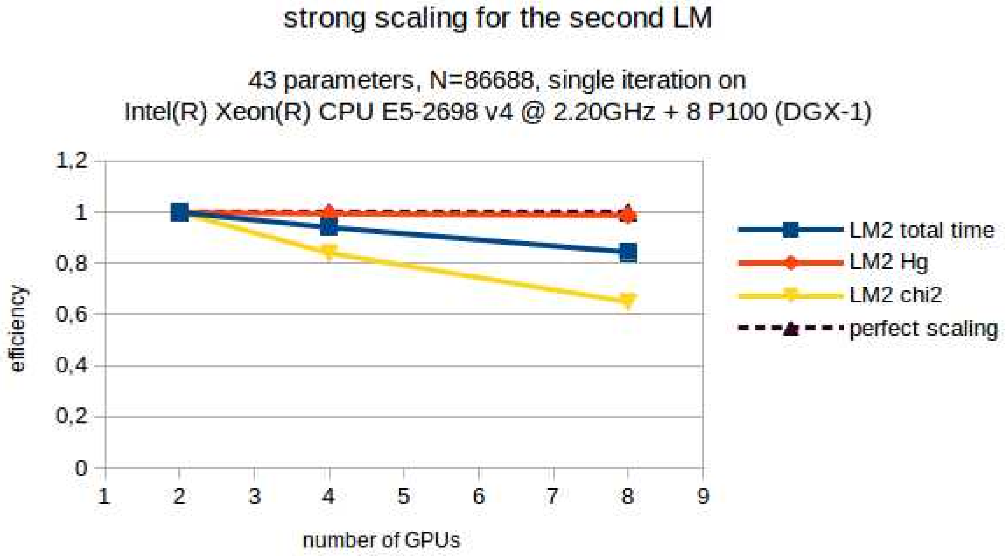
Performance for parameters identification (“Learn” process)

Multi-GPU process, including matrix generation and LM fit

Time to solution for a matrix size of 86k :

- first pass (5 layers) : 25sec
- Second pass (40 layers) : 213sec





Loop supervision module

Reconstructor matrix computation (“apply” process)

- Compute the tomographic reconstructor matrix using covarince matrix between “truth” sensor and other WFS and invert of measurements covariance matrix

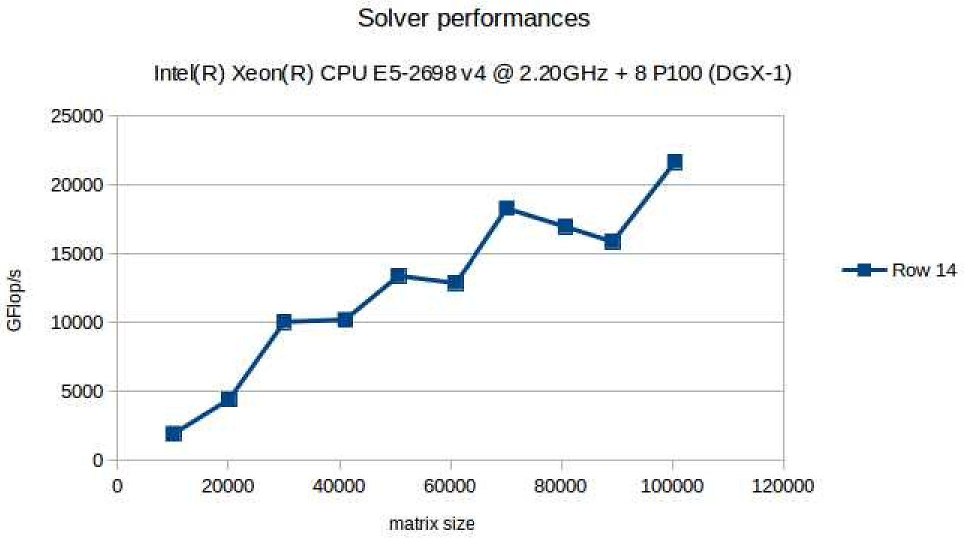
$$R' = Ctm \cdot Cmm_f^{-1}$$

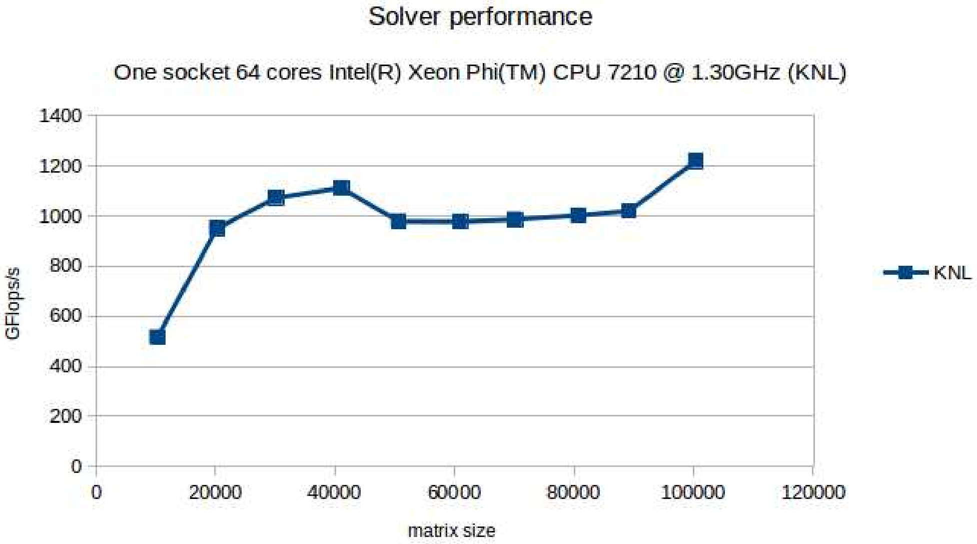
- Can use various methods. “Brute” force : direct solver
- Standard Lapack routine : “posv” : mostly compute-bound, high level of scalability
- Highly portable code : explore various architectures by using standard vendor provided maths libraries

Loop supervision module

Performance for reconstructor matrix computation ("apply" process)

Comparing last generation of GPU (NVIDIA P100) and last generation of Intel Xeon Phi (KNL)



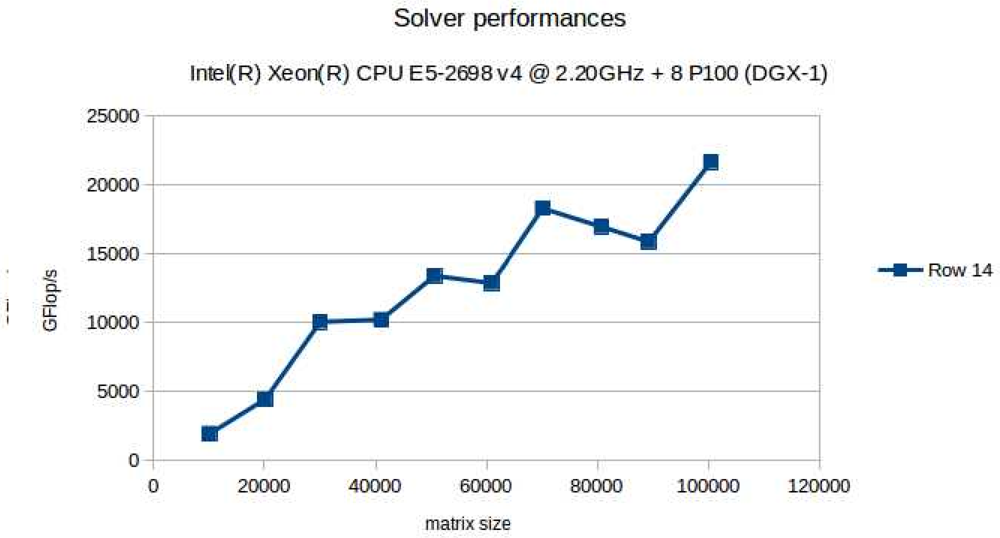


8 GPUs together reach more than 21 TFLOP/s while a single KNL can only reach about 1.2 TFLOP/s in peak performance

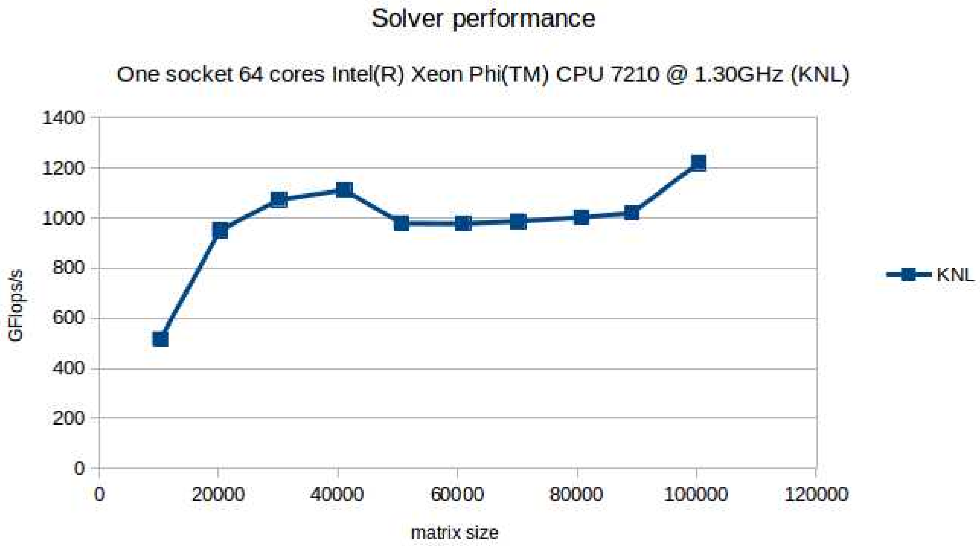
Loop supervision module

Performance for reconstructor matrix computation (“apply” process)

Comparing last generation of GPU (NVIDIA P100) and last generation of Intel Xeon Phi (KNL)


Solver performances

Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz + 8 P100 (DGX-1)


Solver performance

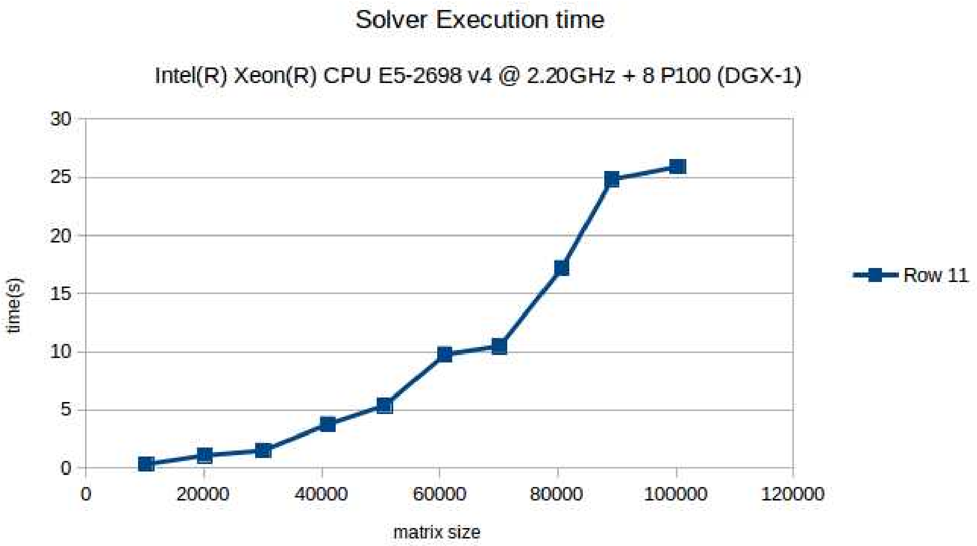
One socket 64 cores Intel(R) Xeon Phi(TM) CPU 7210 @ 1.30GHz (KNL)

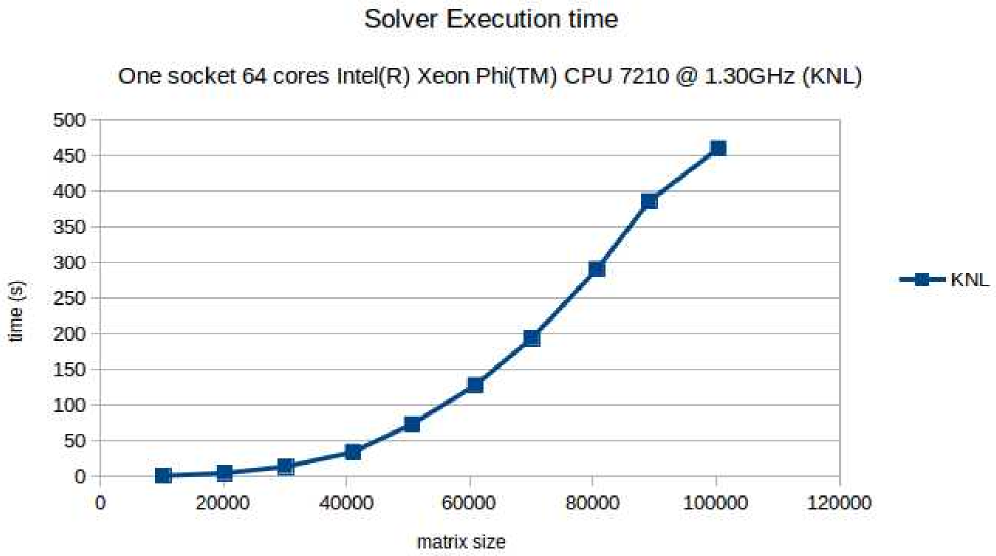
GPUs can deliver better peak perf. (saturation not reached, expect >2.5 or more) and the NVlink interconnect seems to perform very well

Loop supervision module

Performance for reconstructor matrix computation (“apply” process)

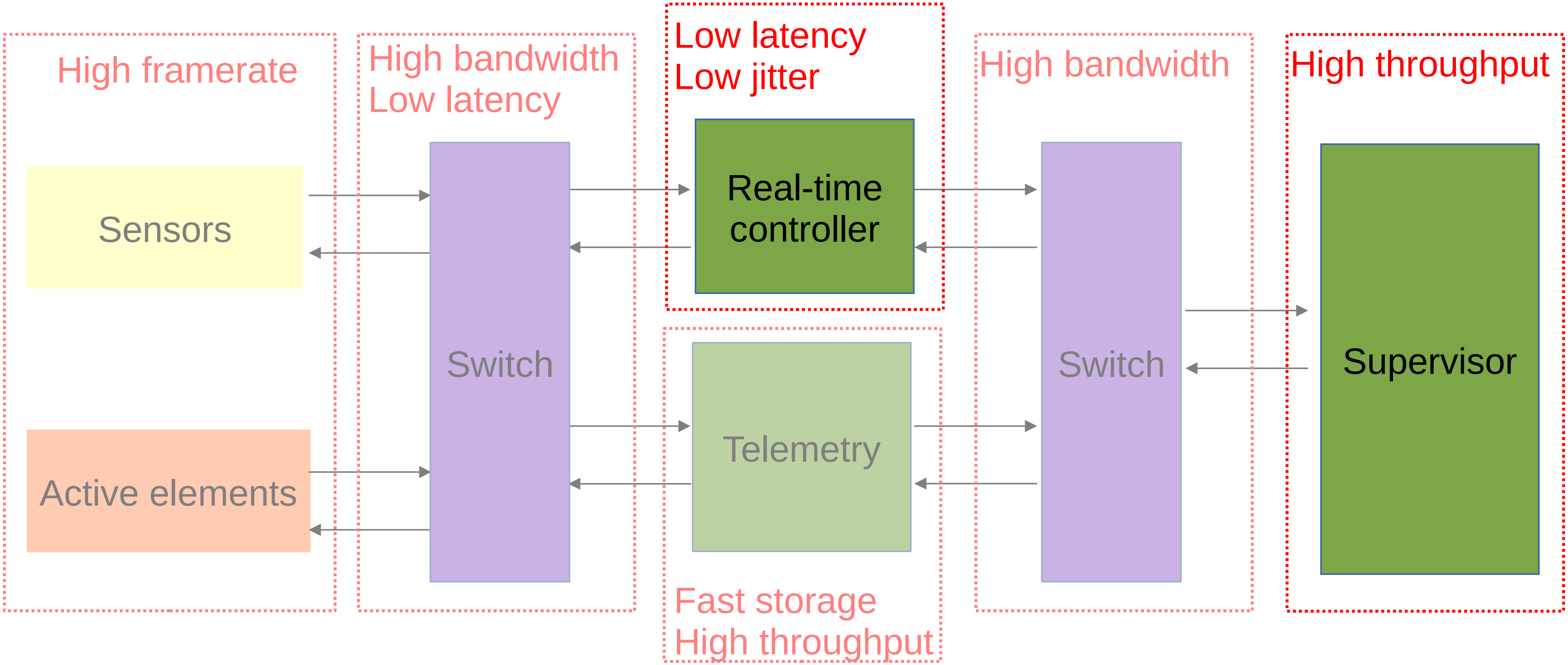
- Comparing last generation of GPU (NVIDIA P100) and last generation of Intel Xeon Phi (KNL)





- Record time-to-solution on DGX-1 : MAORY / HARMONI full scale (100k x 100k matrix) : 25sec to compute tomographic reconstructor

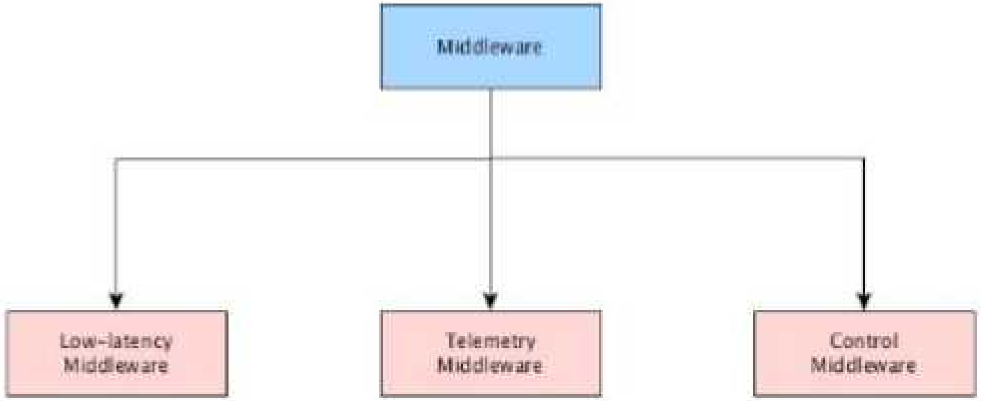
AO RTC concept : SW & MW

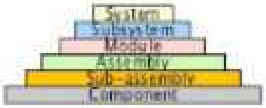


Middleware

3 Middleware domains:

- Control
- Telemetry
- Low-latency pipeline

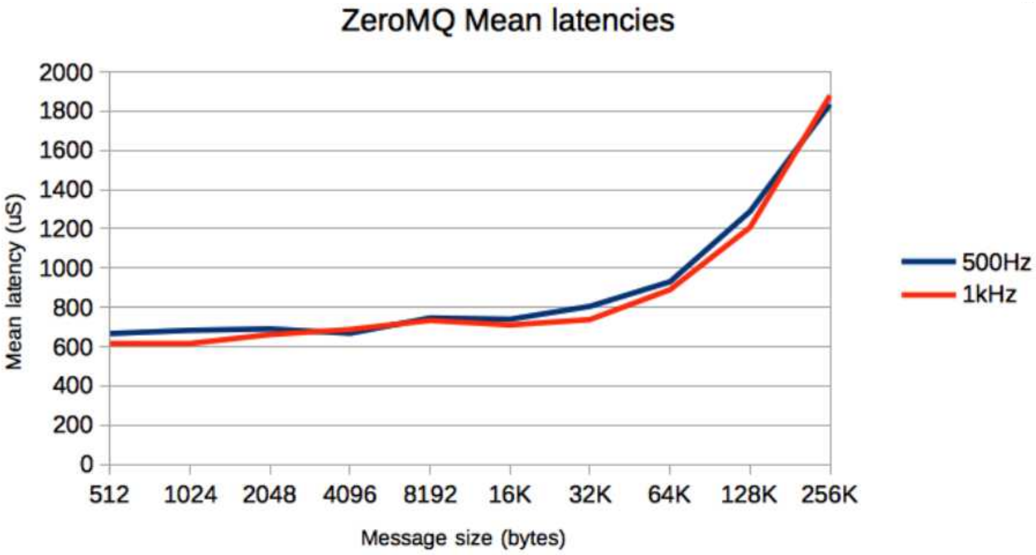


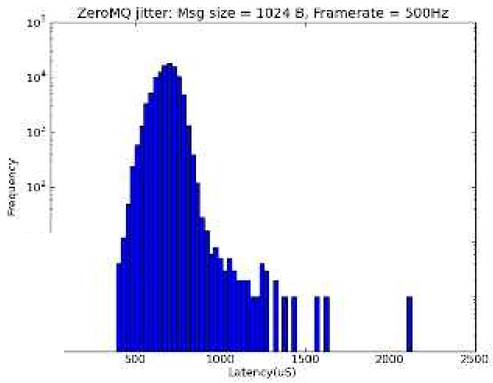


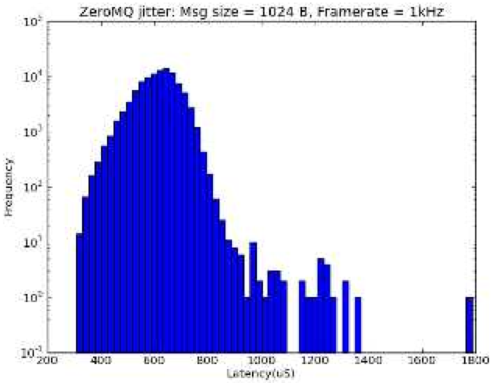
Middleware: ZeroMQ

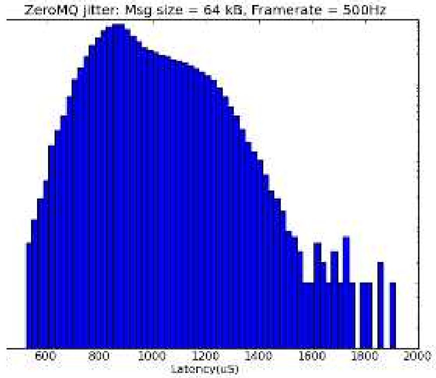
Unsuitable for RT data pipeline

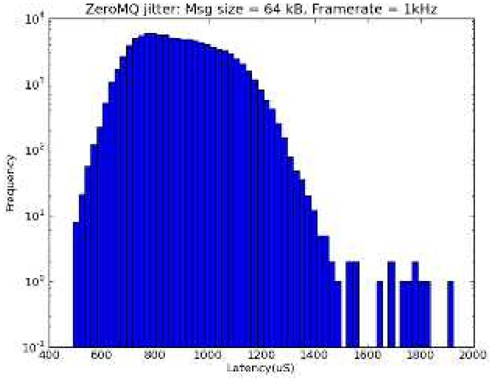
- Excessive latency x3 budget
- Probably due to internal buffering







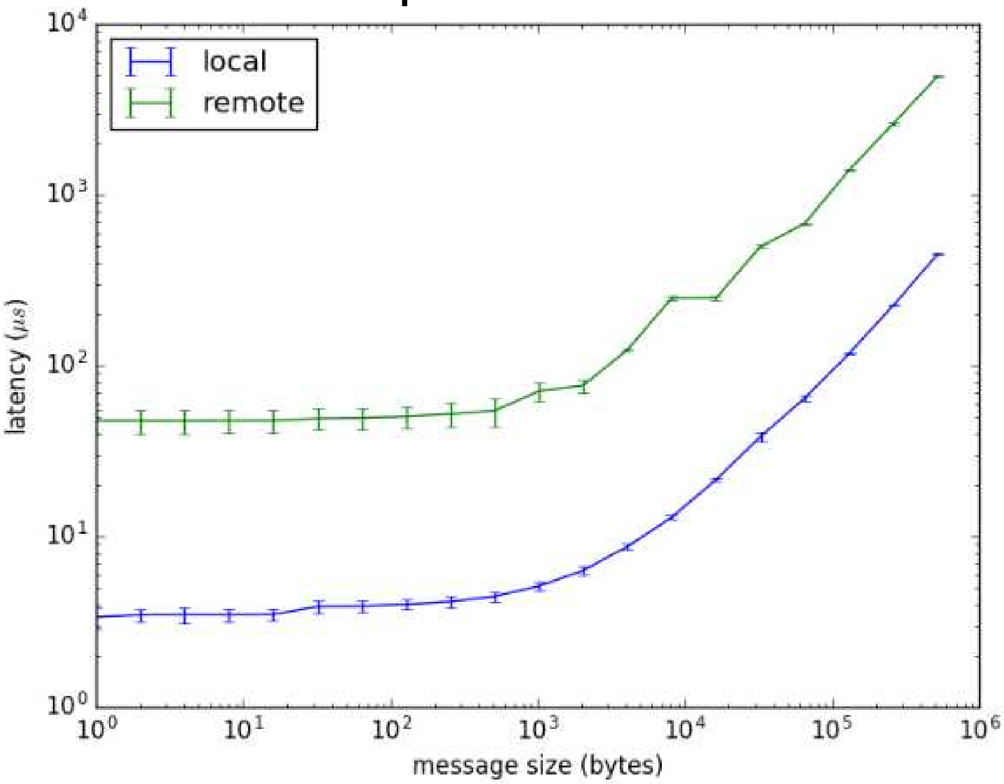


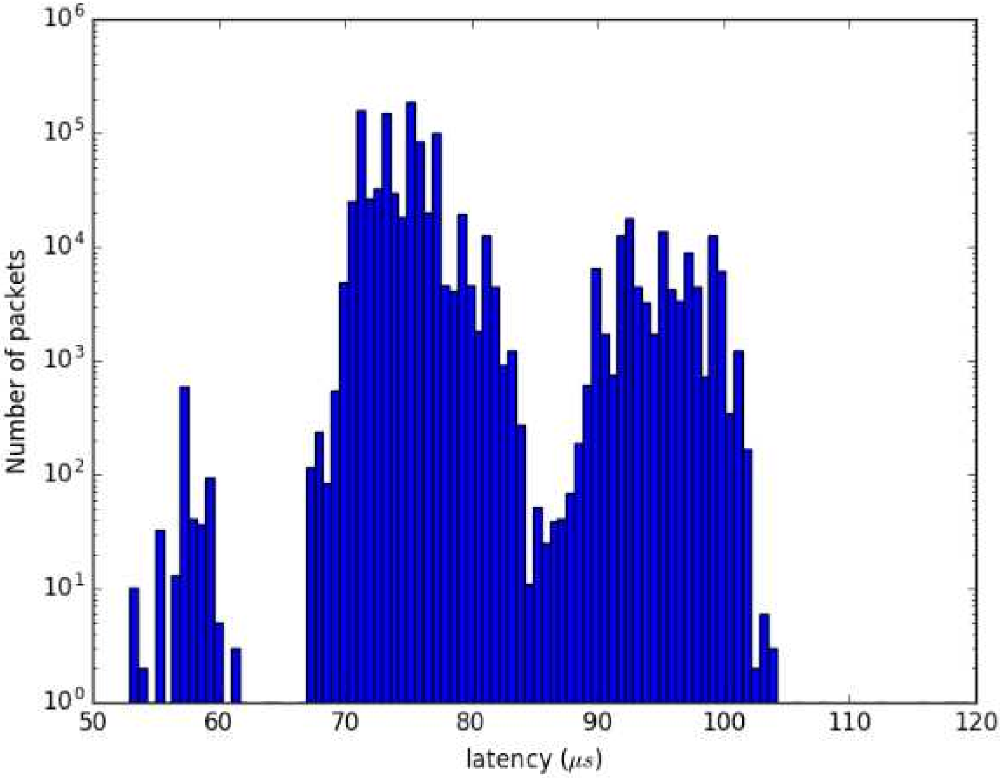


Middleware: MPI

Latency & jitter acceptable

- ~5% of budget for small message sizes
- But limited NW hops allowed, some constraints on implementation





Middleware: down-selection

ID	Criterion	Description	Weighting
DS-MW-1	Reliability	The middleware should be able to guarantee delivery of uncorrupted data, or at the least, detect and signal non-delivery or data corruption.	3
DS-MW-2	Latency	2mS (goal: 1mS) between first pixel received and last actuator demand delivered. This is the total latency budget for the pipeline, the majority of which must be available to be expended on processing; a nominal 10% of the budget has been allowed in this assessment for communications.	3
DS-MW-3	Jitter	100uS peak-to-peak; as in the case of latency, this is the budget for the pipeline. Contributions to jitter from different sources (processing, communication,...) sum quadratically; a nominal 30% of total jitter has been allowed.	3
DS-MW-4	Throughput	Within the pipeline: the most demanding case int terms of aggregate throughput is METIS LTAO mode, with a frame rate of 1kHz and 6 LGS/3 NGS WFS. The input bandwidth for pixel data is ~ 200Gb/s (25 Gb/s). However, this is not carried by a single connection, and pixel input data is not carried by the middleware. A more realistic requirement on bandwidth per link *within* the pipeline is the transport of pixel data for a single WFS, from a calibration module to a centroider module; for a single LGS WFS, the required bandwidth is 2.19 Gb/s (274 MB/s). If calibration and centroiding are perfomed within the same hardware module and data is not required to be transported on the network at pixel rates, the requirement is reduced to transporting frames of centroids from a single WFS, and for a LGS WFS at 1kHz this is evaluates to 350 Mb/s (44 MB/s).	3

Middleware: down-selection

Criterion	Weighting	Technology	Remarks	Score	Weighted score
DS-MW-1 Reliability	3	ZeroMQ	No guaranteed delivery	0	0
		MPI	Reliable QoS available	3	9
DS-MW-2 Latency	3	ZeroMQ	Unable to meet requirement	0	0
		MPI	Required performance achieved in testing	3	9
DS-MW-3 Jitter	3	ZeroMQ	Unable to meet requirement	0	0
		MPI	Required performance achieved in testing	3	9
DS-MW-4 Throughput	3	ZeroMQ	Required performance achieved in testing	3	9
		MPI	Required performance achieved in testing	3	9
DS-G-1 Cost	1	ZeroMQ	Available free/open source	3	3
		MPI	Available free/open source	3	3
DS-G-2 Ease-of-use	1	ZeroMQ	Commensurate with facilities provided	2	2
		MPI	Commensurate with facilities provided	2	2
DS-G-3 Long-term support	2	ZeroMQ	Single supplier; commercial support available	2	4
		MPI	Several implementations available, and very widely used.	2	4
DS-G-4 Standards compliance	2	ZeroMQ	No standard	0	0
		MPI	De-facto HPC standard	1	2
DS-G-5 Familiarity	2	ZeroMQ	Expertise in consortium	1	2
		MPI	Expertise in responsible partner	2	4
DS-G-8 Source of supply	2	ZeroMQ	Single supplier	1	2
		MPI	Multiple implementations	3	6
Overall Score		ZeroMQ			22
		MPI			57

Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique

University

Summary

Project on track

- PDR occurred in Jan. 2016 and MTR in Feb. 2017 with feedback from community
- Prototyping activities are entering final phase with downselection and final prototype(s) architecture to be defined by end 2017 during FDR

Collaborations initiated

- Good feedback from the community on different aspects (HPC + instrumentation)
- Evaluate the convergence and minimize additional effort
- More than happy to collaborate more !

Excellent feedback for European Commission

- Mid-term progress review in Brussels last week

Already enhancing the readiness level of commercial solutions

- Contribution to QuickPlay development environment
- Design of innovative FPGA boards (see Roberto's poster)

Green Flash @ AO4ELT5

1) Biasi et al. “*FPGA based microserver for high performance real-time computing in AO*” [**P3055**]
2) Reeves et al. “*The Green Flash Real-Time simulator*” [**P1052**]
3) Perret et al. “*A generic and scalable heterogeneous architecture for real-time computing and performance measurements in AO*” [**P3037**]
4) Doucet et al. “*Efficient supervision strategy for tomographic AO systems on ELTs*” [**P3054**]
5) Bernard et al. “*A GPU based RTC for the E-ELT AO: RTC prototype*” [**P3045**]
6) Jenkins et al. “*ELT scale real-time control on Intel Xeon Phi and manycore CPUs*” [**P3036**]
7) Ferreira et al. “*ROKET: erROr breaKdown Estimation Tool for adaptive optics systems*” [**P1057**]
8) Vidal et al. “*MICADO SCAO numerical simulations*” [**P1056**]
9) Petit et al. “RTC strategies for Harmoni SCAO and LTAO modes” [**P3035**]

l'Observatoire de Paris | LESIA
Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique

Durham University

MICROGATE

PLDA