Green Flash

High performance computing for real-time science

Contribution from Observatoire de Paris on WP4

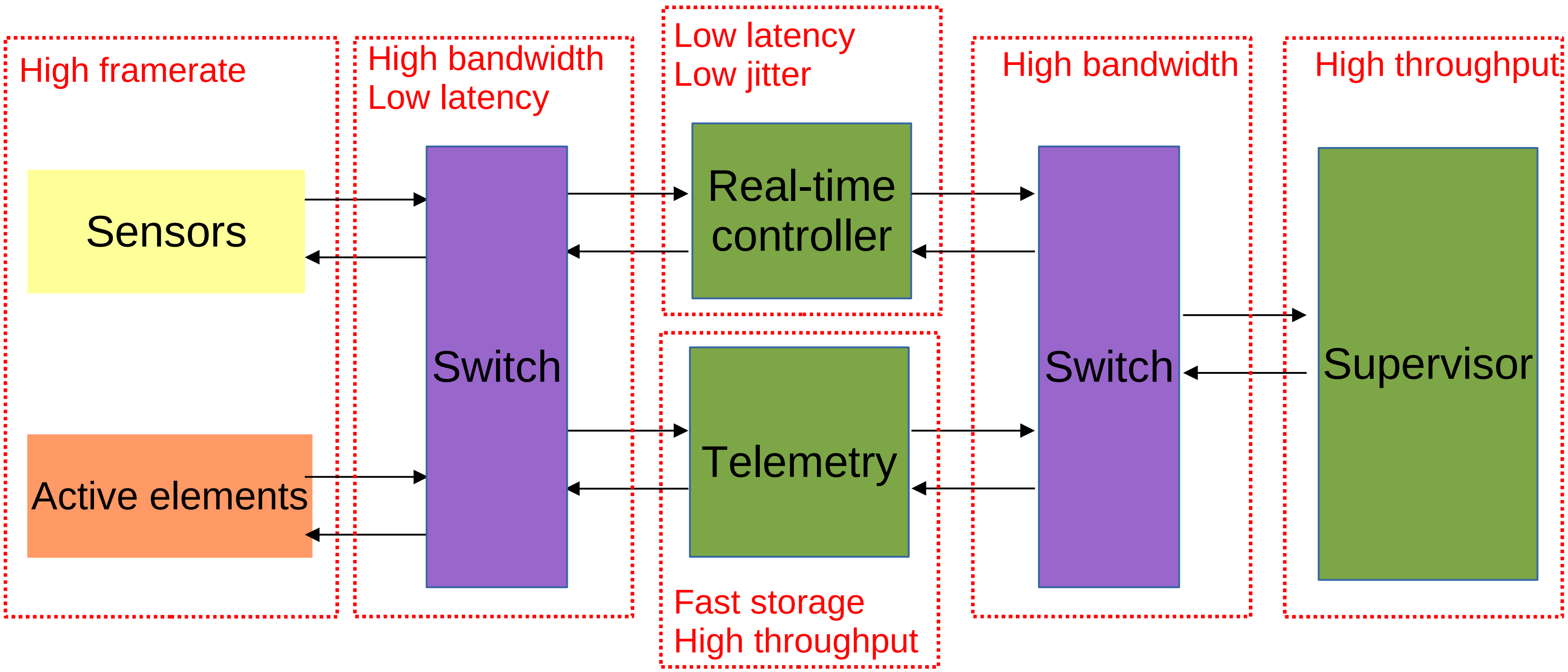
Final Design Review, April 6th 2018

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

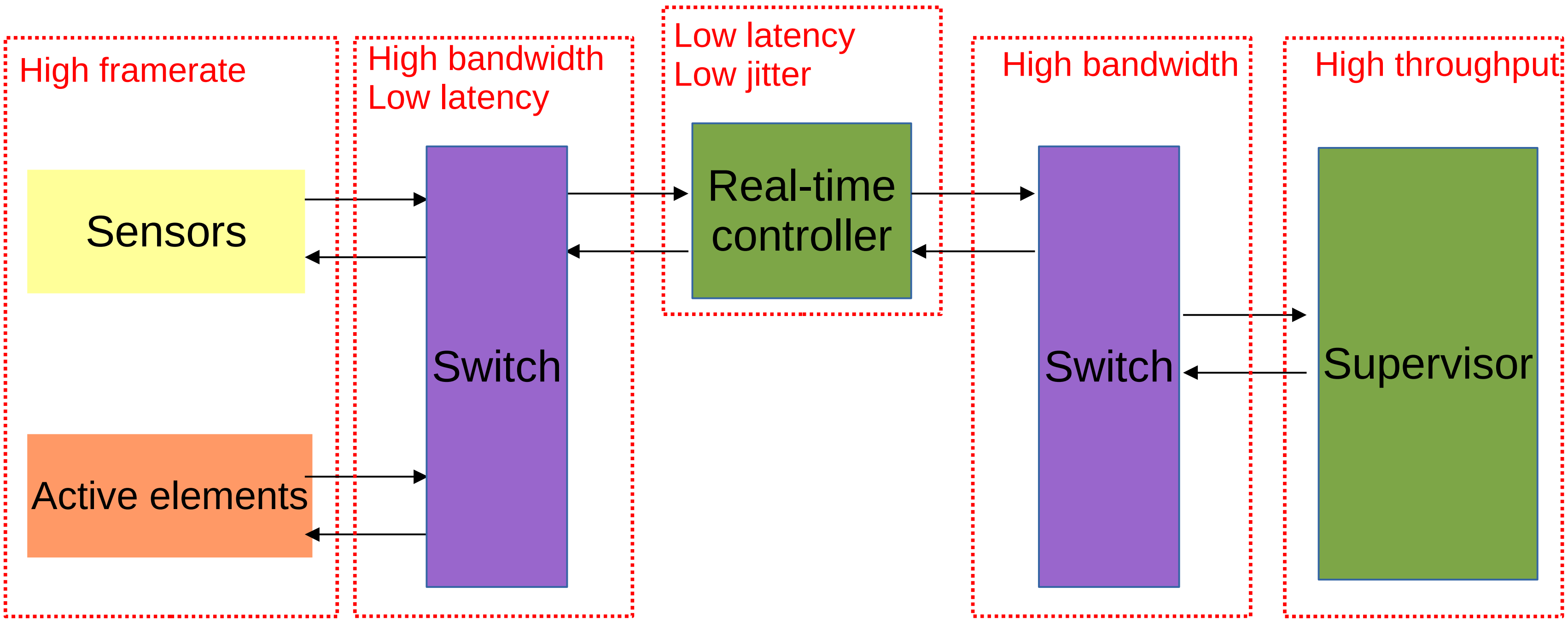
WP 4 : Accelerators for real-time HPC

- Assess various HW accelerator options on a real-time application
  - GPU : lead by OdP with contribution from UoD
  - Xeon Phi : lead by UoD
  - FPGA : lead by UoD with contribution from OdP
- Assess performance of same hardware on complex data pipeline
  - Supervisor module for AO : lead by OdP
  - Criterion optimization and large matrix inversion

WP 4


High framerate
Sensors
Active elements
High bandwidth
Low latency
Switch
Low latency
Low jitter
Real-time controller
Telemetry
Fast storage
High throughput
High bandwidth
Switch
High throughput
Supervisor

WP 4


High framerate
Sensors
Active elements
High bandwidth
Low latency
Switch
Low latency
Low jitter
Real-time controller
High bandwidth
Switch
High throughput
Supervisor

WP 4

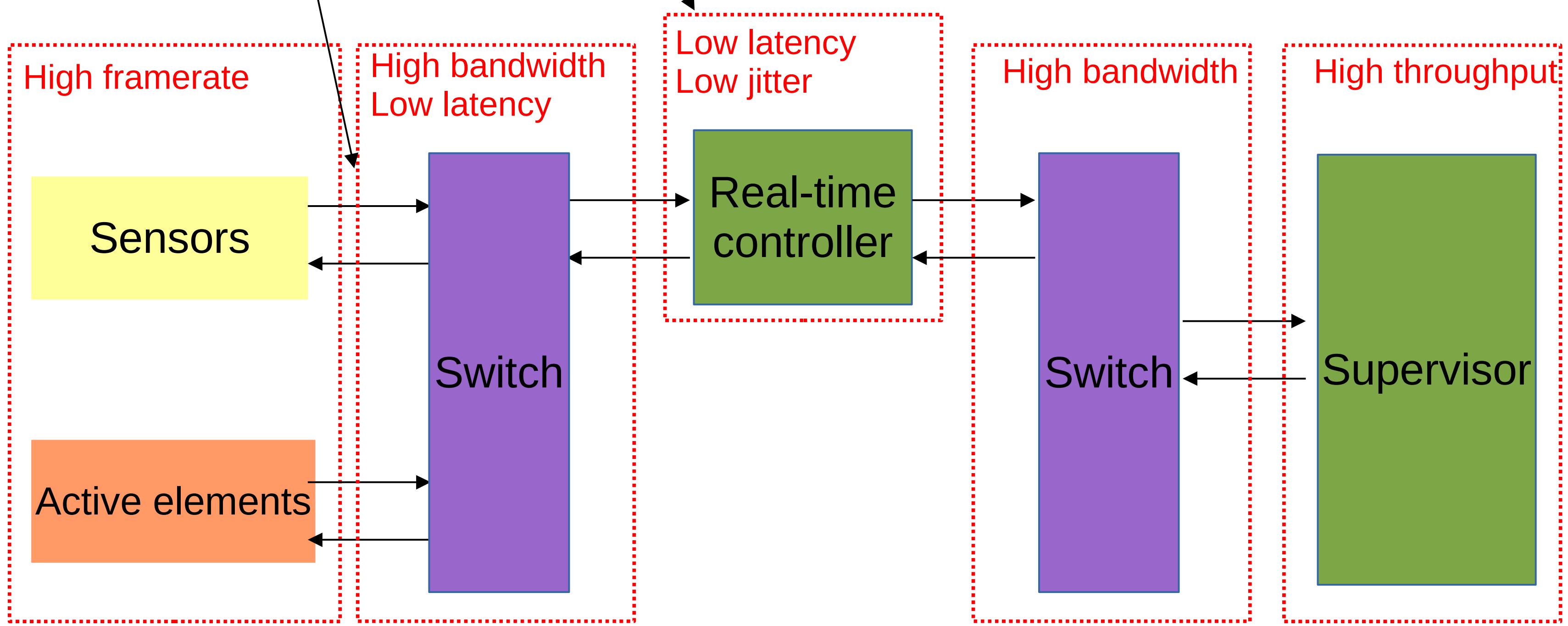
Real-time pipeline.
Includes sensors pixels streams processing and MVM for control of active elements

Pixel Streams processing, 2 options:
* tens of GFLOPS in simple arithmetics or
* hundreds of GFLOPS in batched Fourier Transform

MVM : up to 3 TFLOP/s (1.5 TMAC/s)

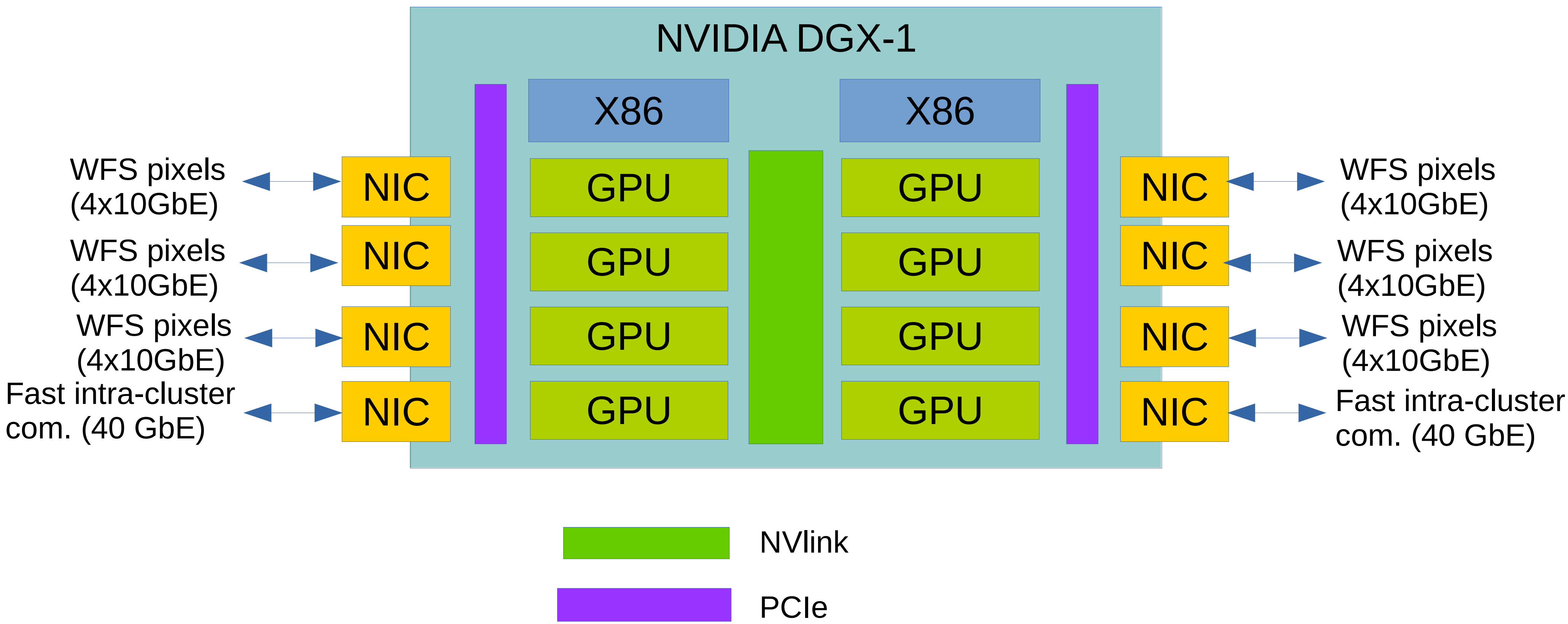
Up to 100 Gb/s of streaming data

Performance must be deterministic, max jitter : 100µs



RT data pipeline with GPUs

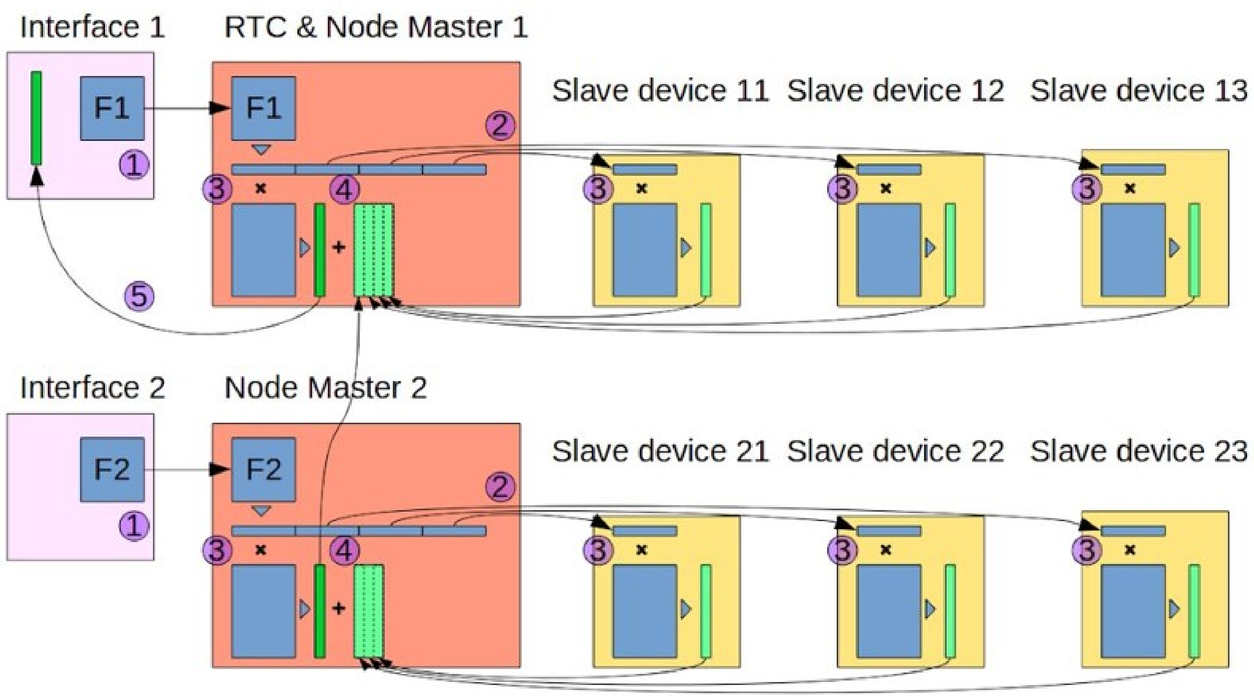
- Prototype using latest generation GPU cluster



- Concept studied at LESIA

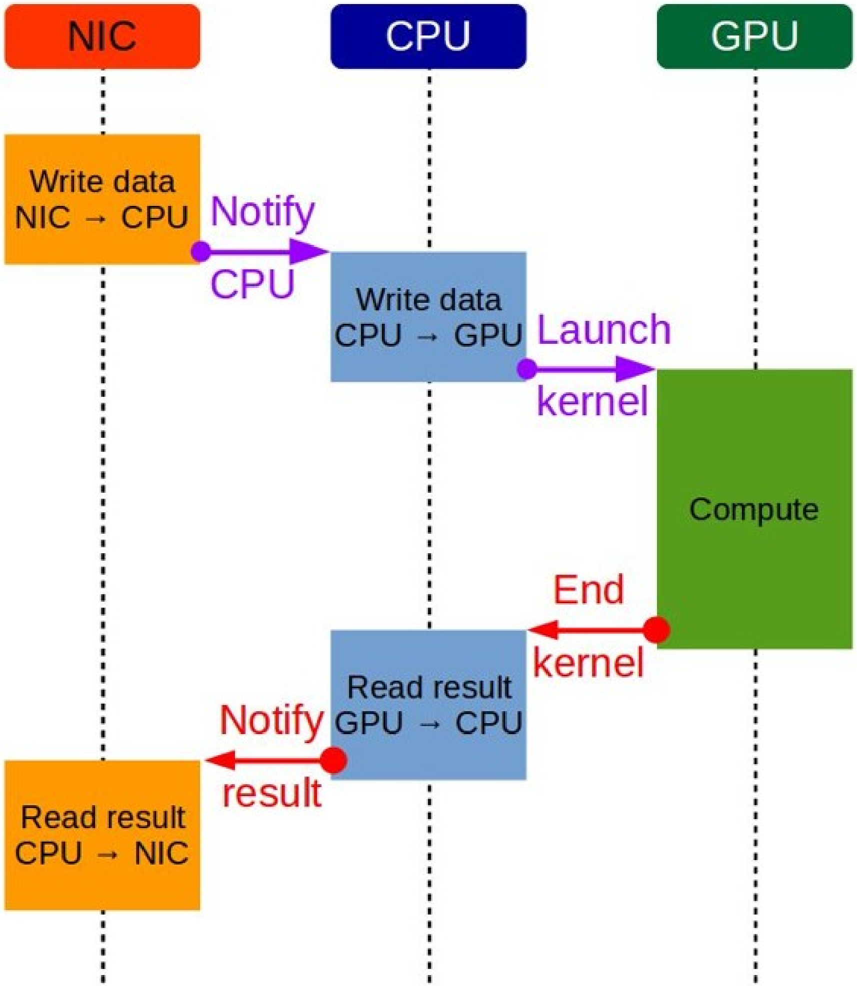
System architecture

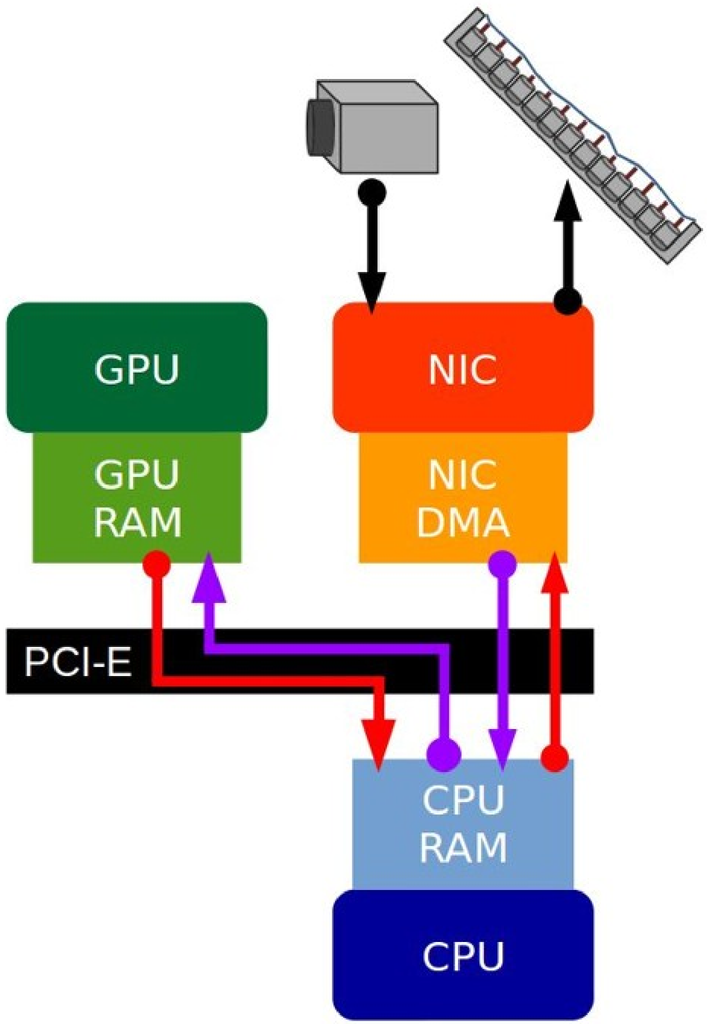
- Master-Slave approach



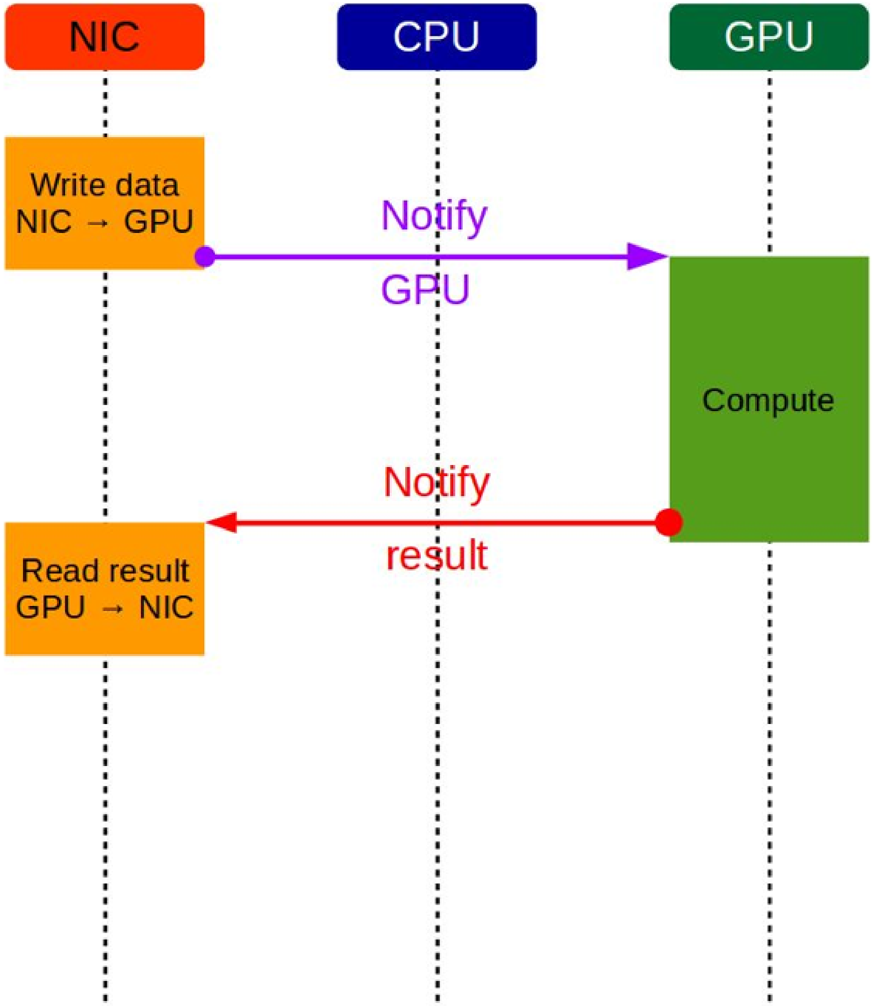
1. Send frame to master devices
2. Send slopes to each slave device
3. Compute assigned slopes & MVM
4. Gather and sum all command vectors
5. Send back command vector to interface

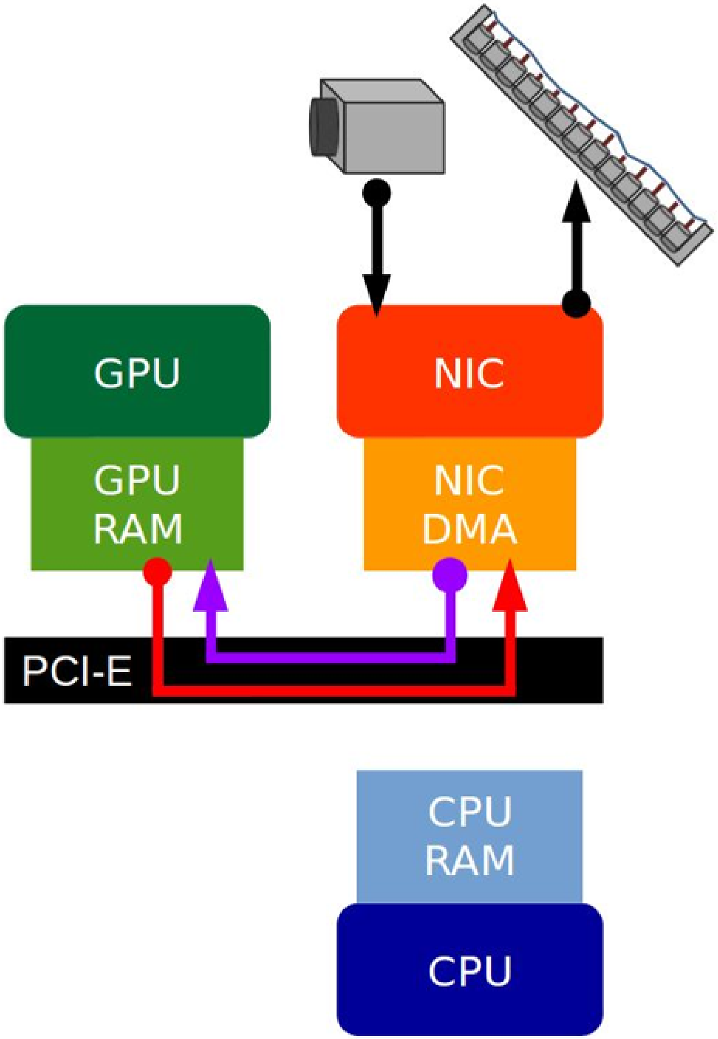
Standard GPU programming implementation





Persistent kernels and DMA



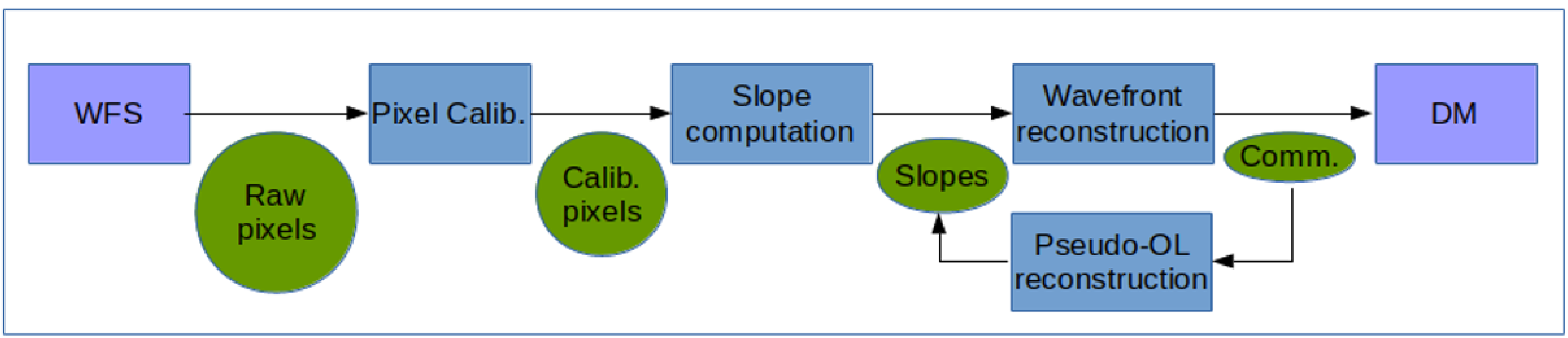


l'Observatoire de Paris — LESIA
Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique
Durham University
MICROGATE
PLDA

AO pipeline

Profile is dominated by MVM

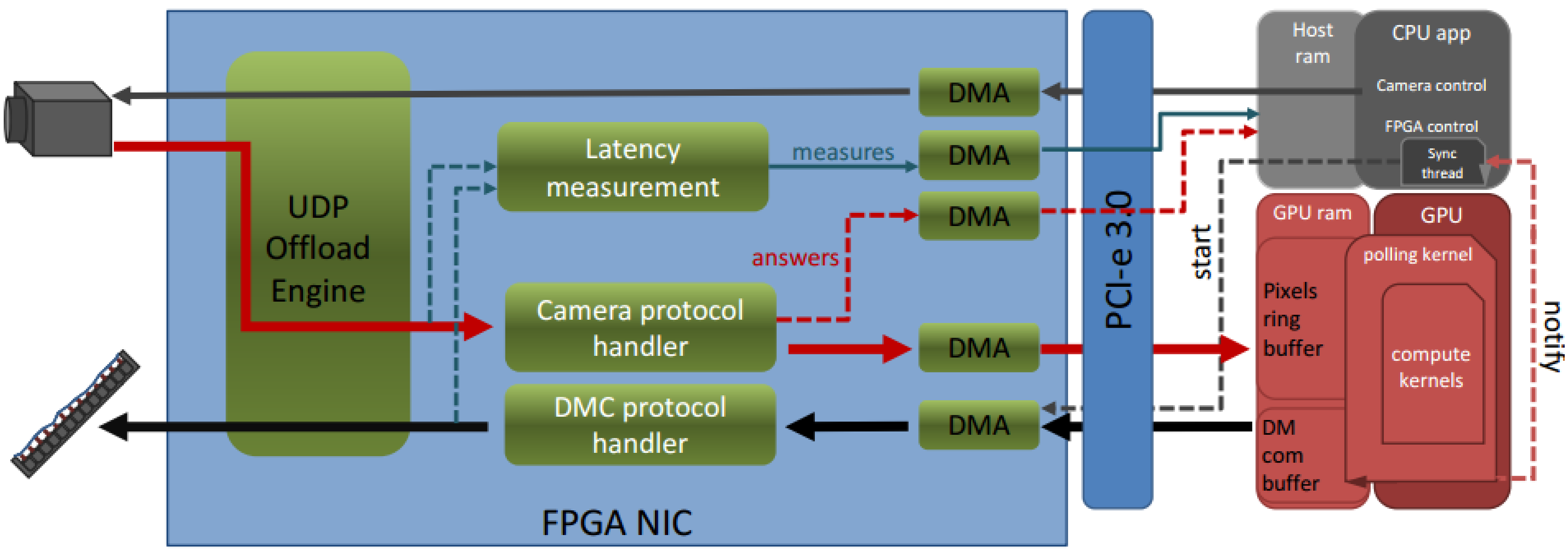
- Master GPU receives the data from WFS (simplest datapath), compute the slopes and distribute over slave GPUs
- In the case of several nodes, data from WFS is shared between the two node masters but a single RTC master will collect the data



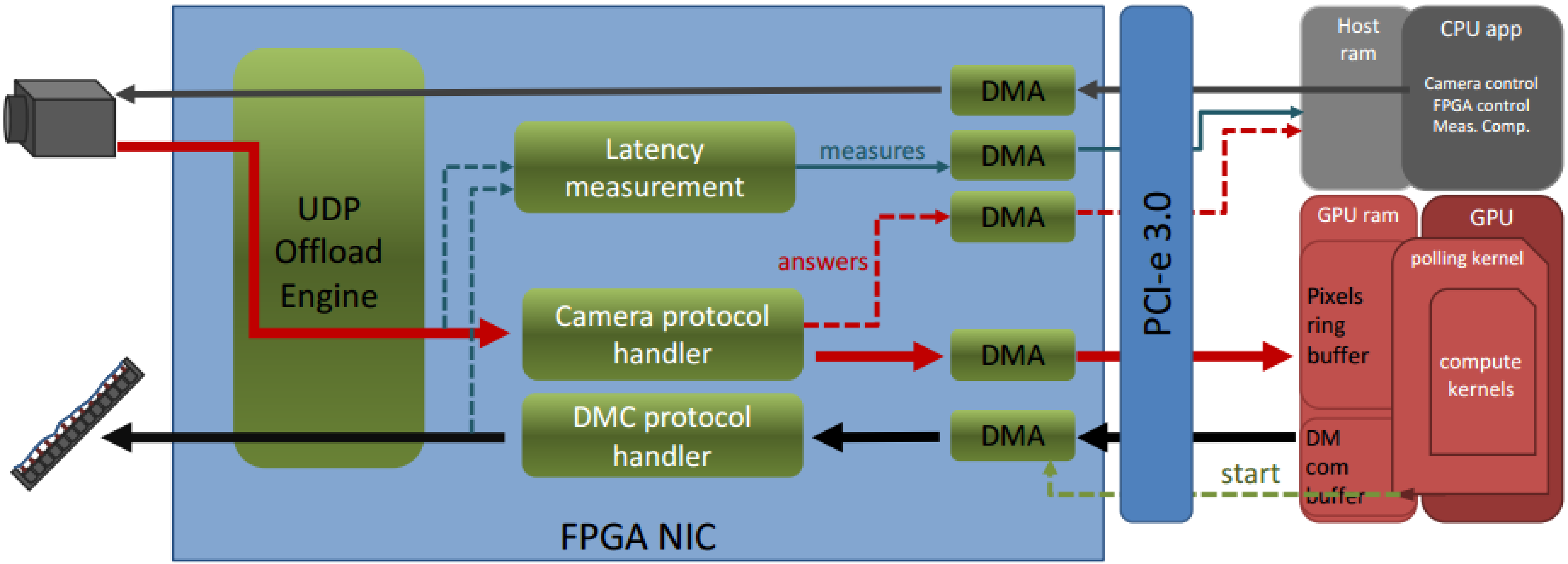
Optimizing GPU-FPGA sync

FPGA writes/reads directly to/from GPU memory

Using only writes would be better though



Optimizing GPU-FPGA sync



Little to no improvements, but CPU free for other kind of computations

Initial results

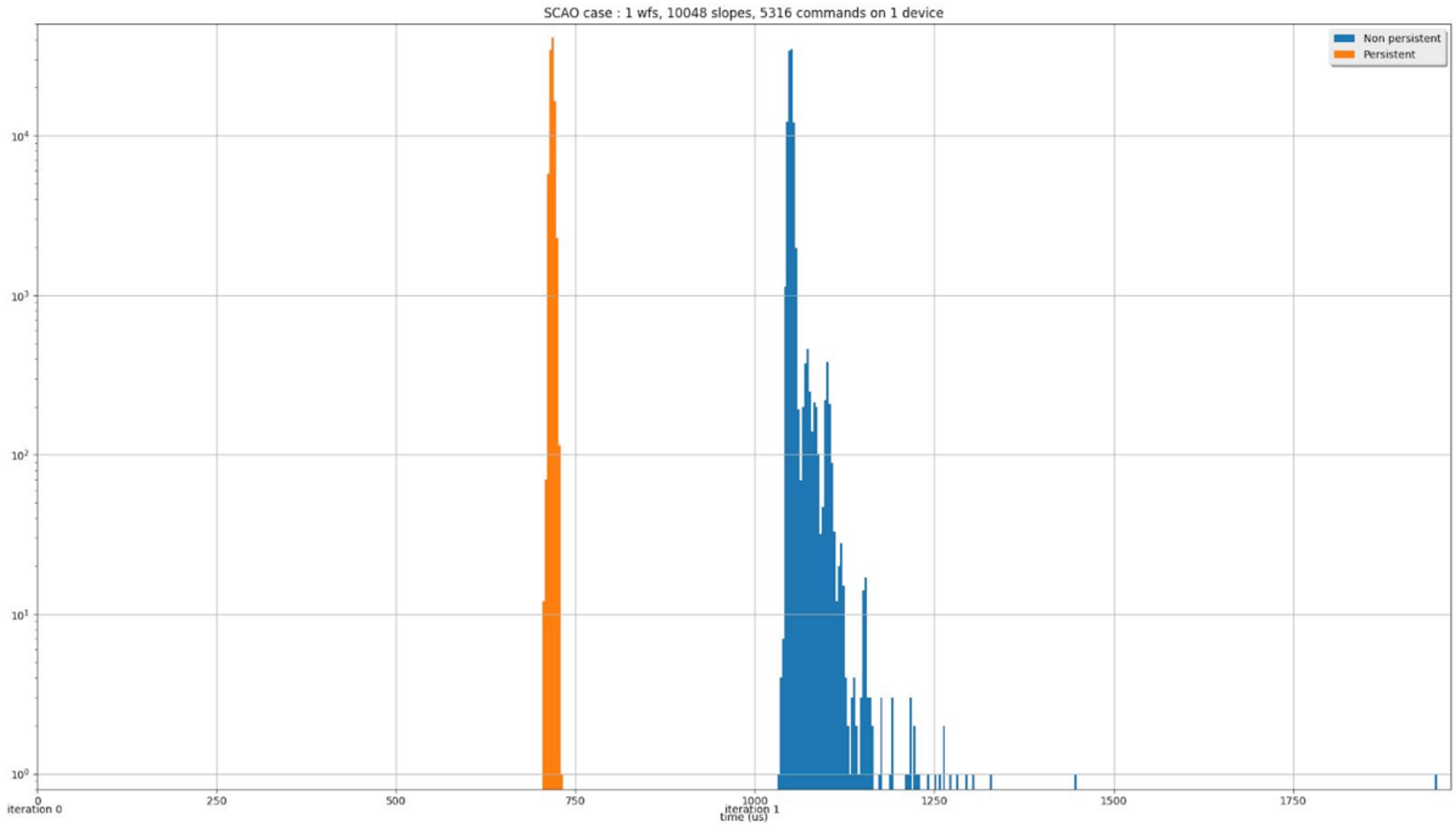
Tested various system configurations

Tests were performed on a DGX-1 platform (only 1 GPU for SCAO)

Name	N slopes	N actuators	Goal frame rate
SCAO	10048	5316	1000 FPS
LTAO	60288	5316	500 FPS
MCAO	60288	15316	500 FPS

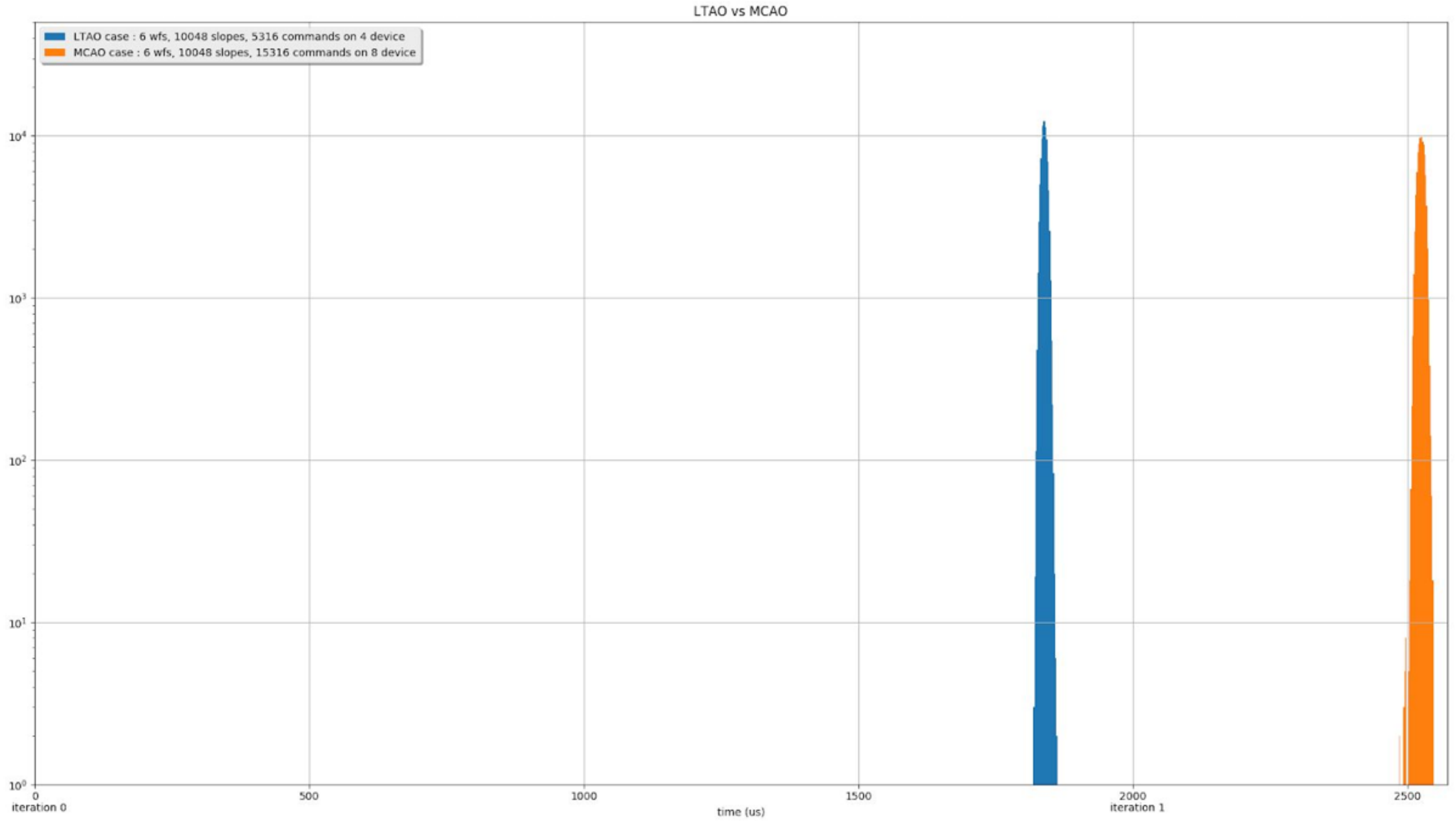
Initial results: SCAO

Persistent kernels versus standard kernels



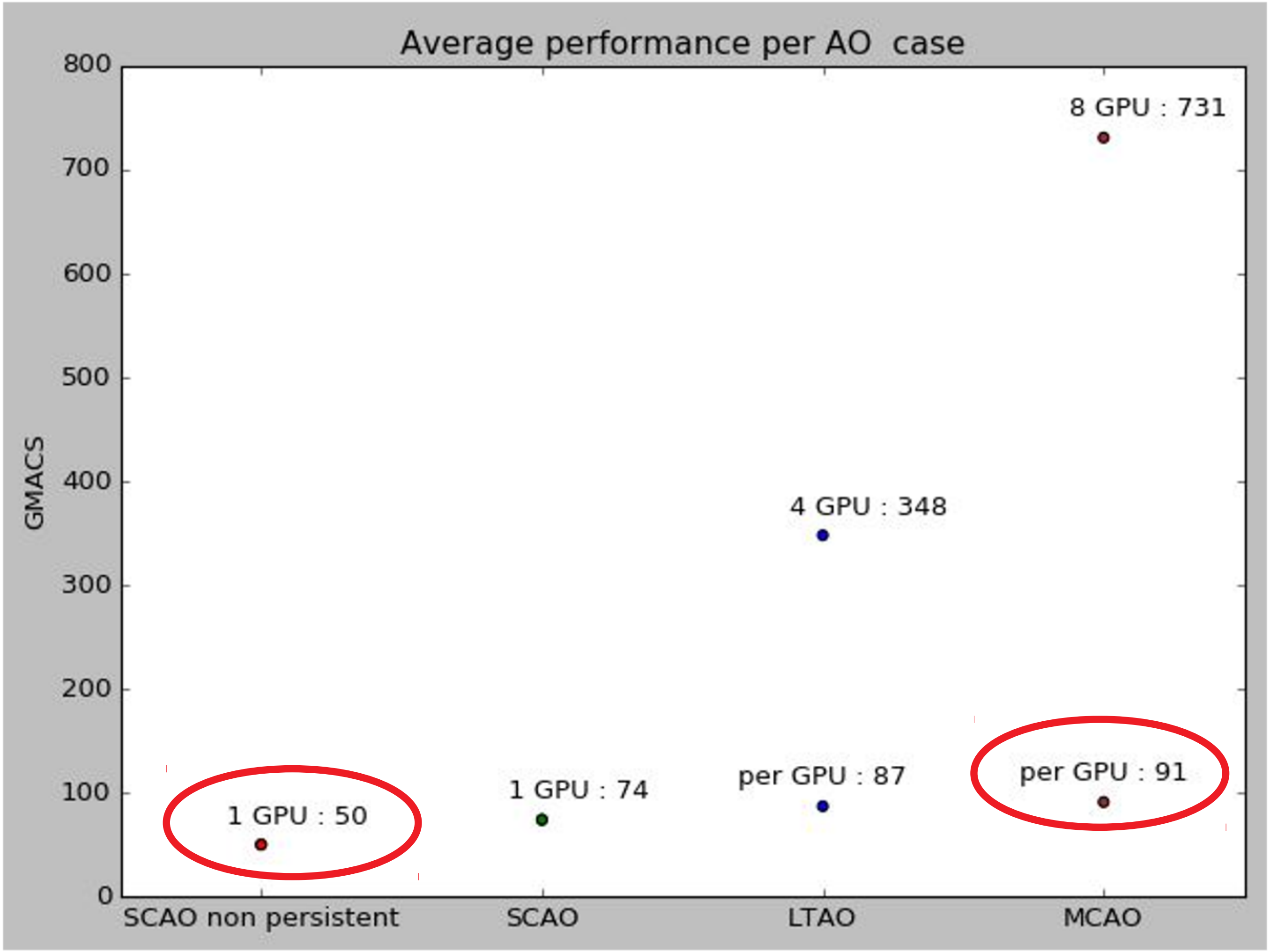
Initial results: LTAO/MCAO

LTAO running on 4 GPUs, MCAO on 8 GPUs



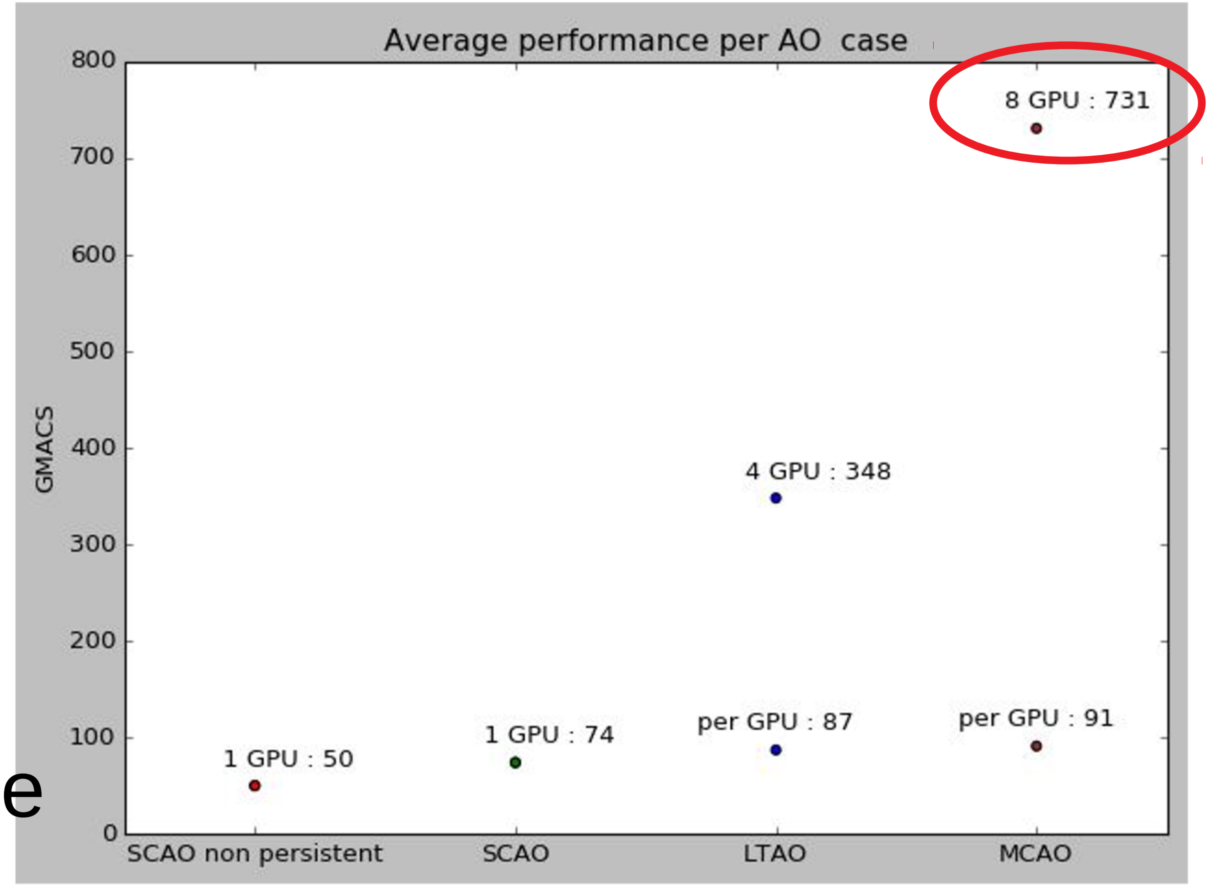
Initial results: throughput

SCAO case is not large enough to feed a GPU !



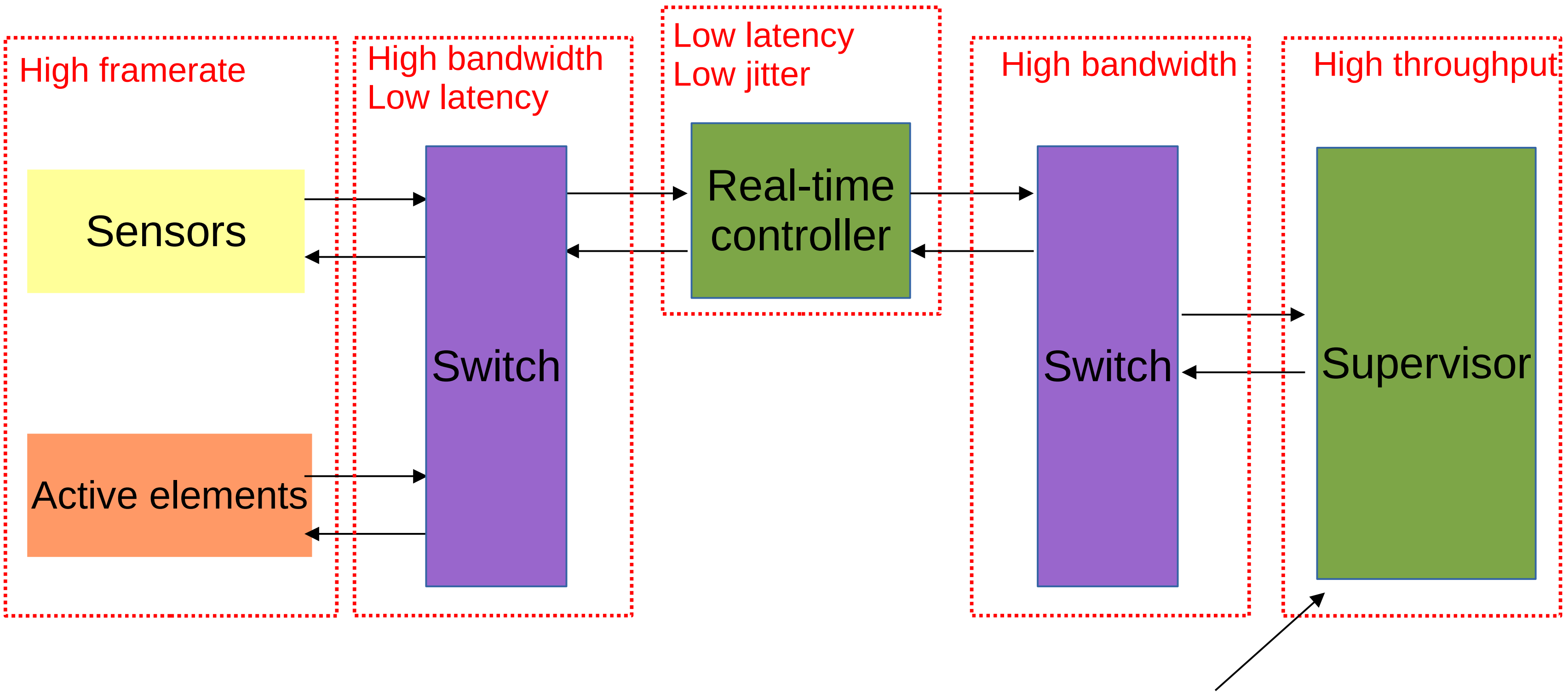
Initial results: throughput

- Reaching “only” 731 GMAC/s today (8xP100).
- x2 speedup to be expected with newer GPU generations in 1-2 years timeframe (faster HBM)



l'Observatoire de Paris — LESIA
Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique
Durham University
MICROGATE
PLDA

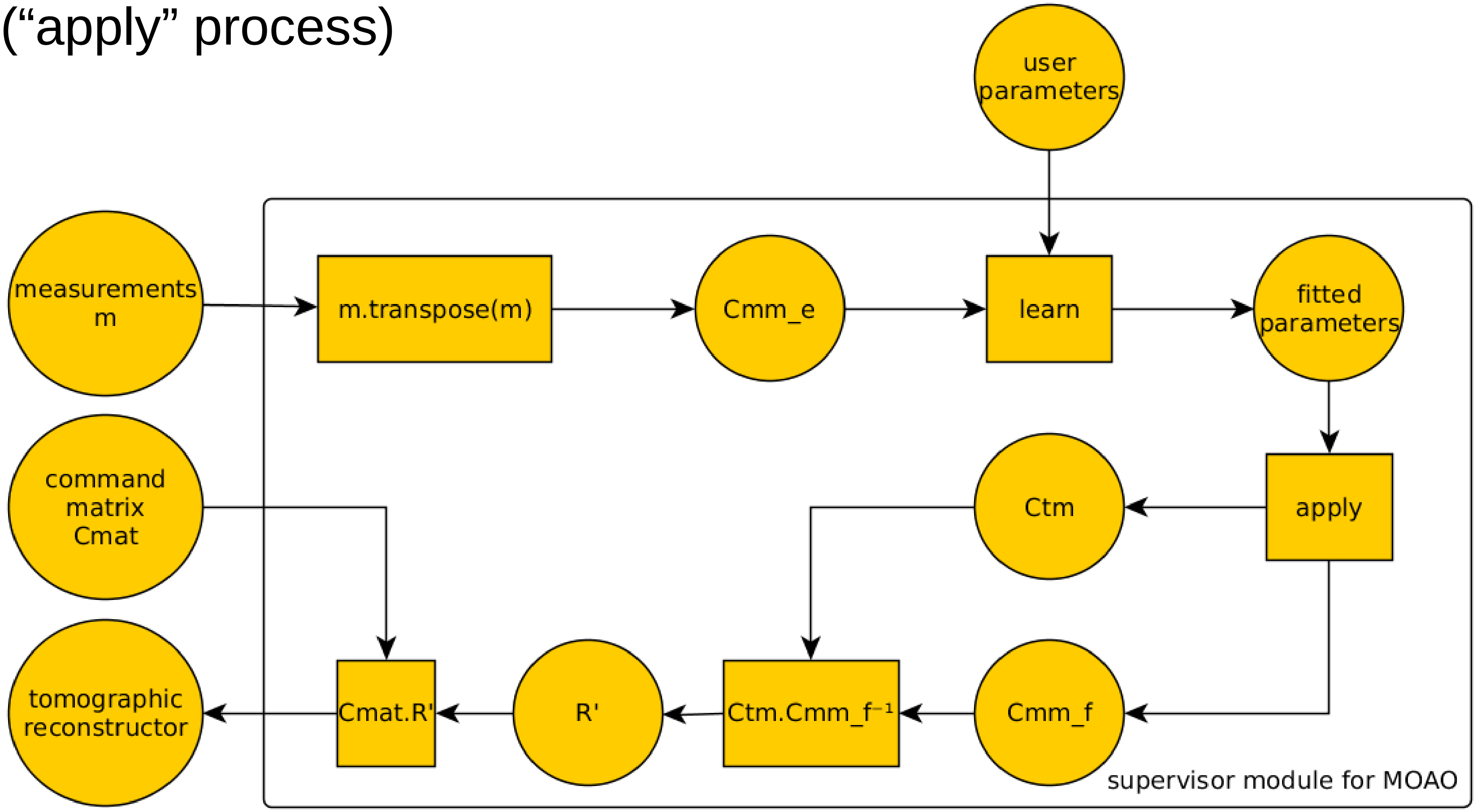
WP 4



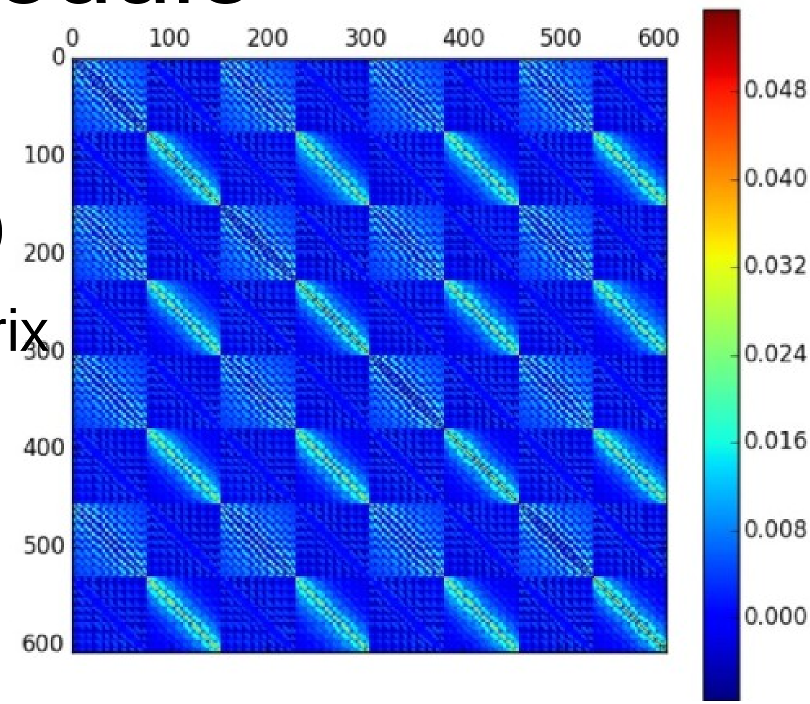
Supervisory module. Use the output data stream from RT pipeline to re-optimize the control matrix
2 stages : function optimization (gradient descent) and Choleski inversion : up to 100 TFLOP/s

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

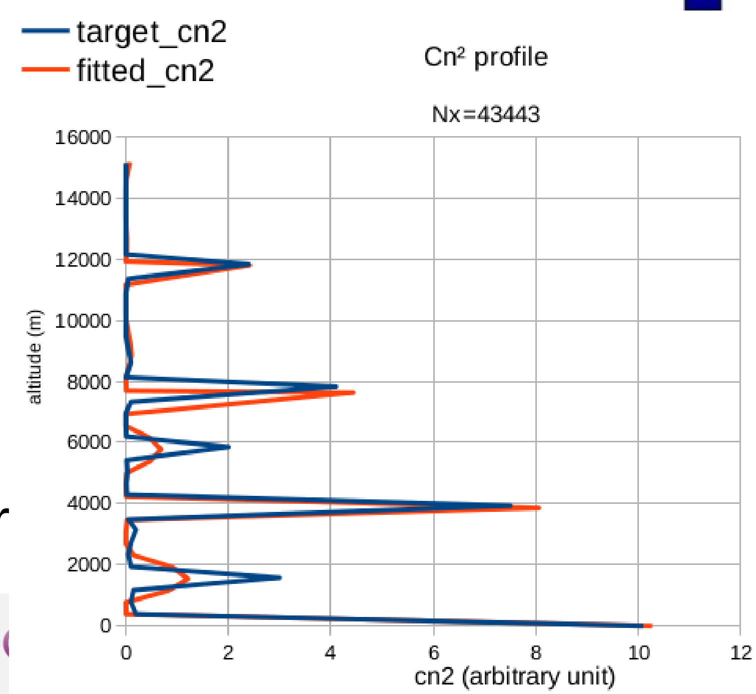


Parameters identification ("Learn" process)

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

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

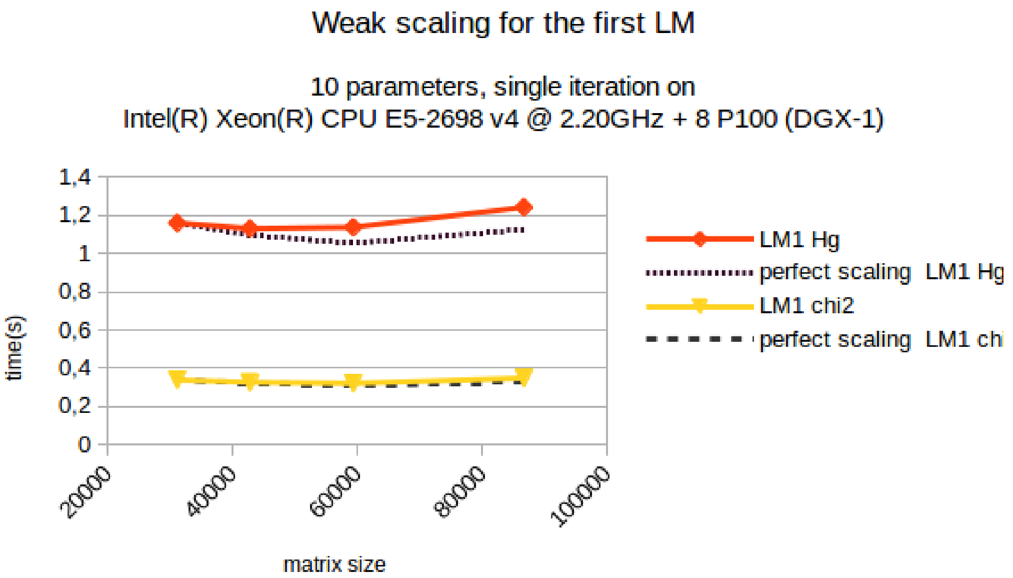
- Levenberg-Marquardt algorithm for function optimization
- Exemple of turbulence profile reconstruction
- Dual stage process (5 layers + 40 layer

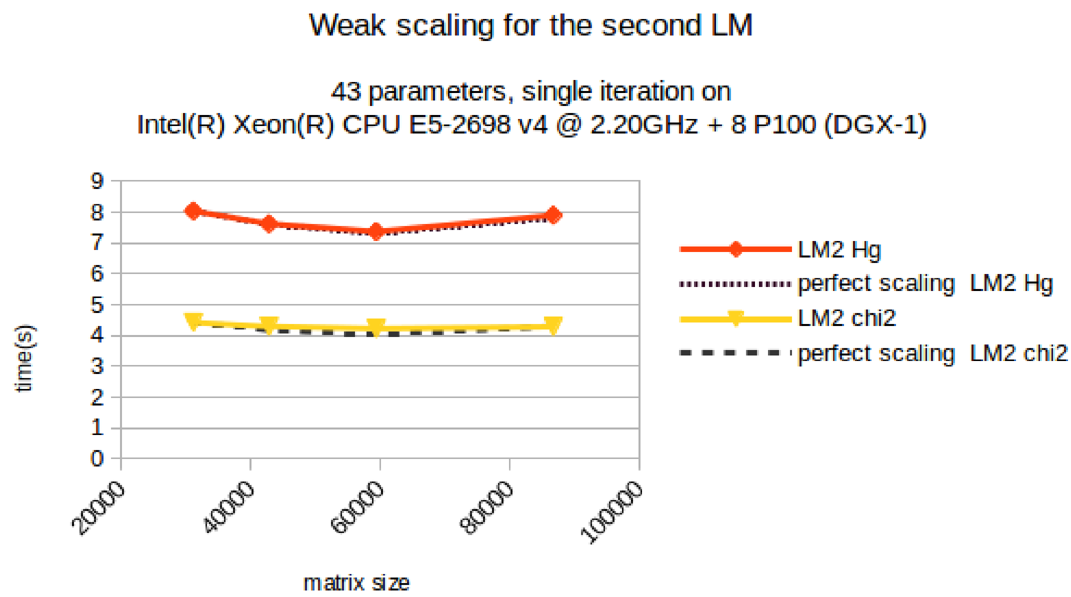


l'Observatoire de Paris — LESIA
Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique
Durham University
MICRO

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





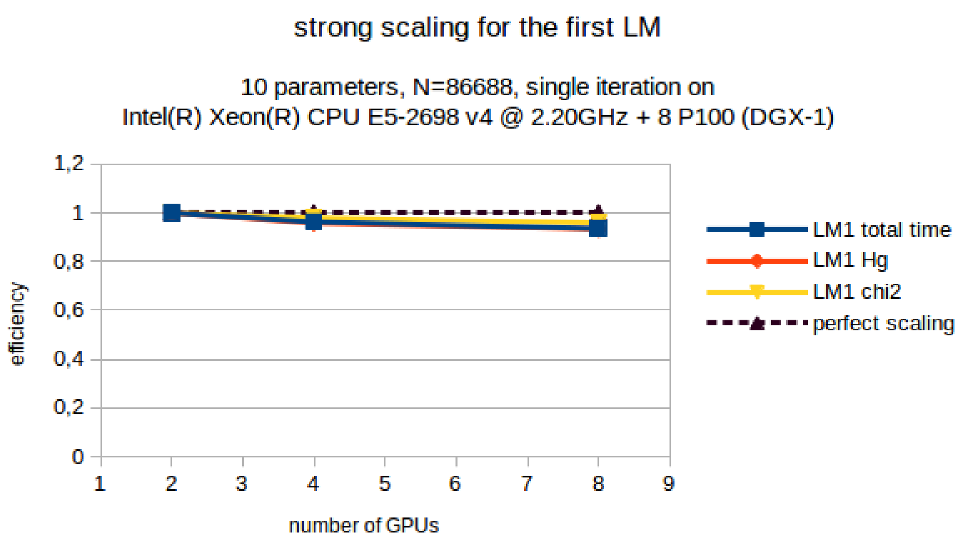
Loop supervision module

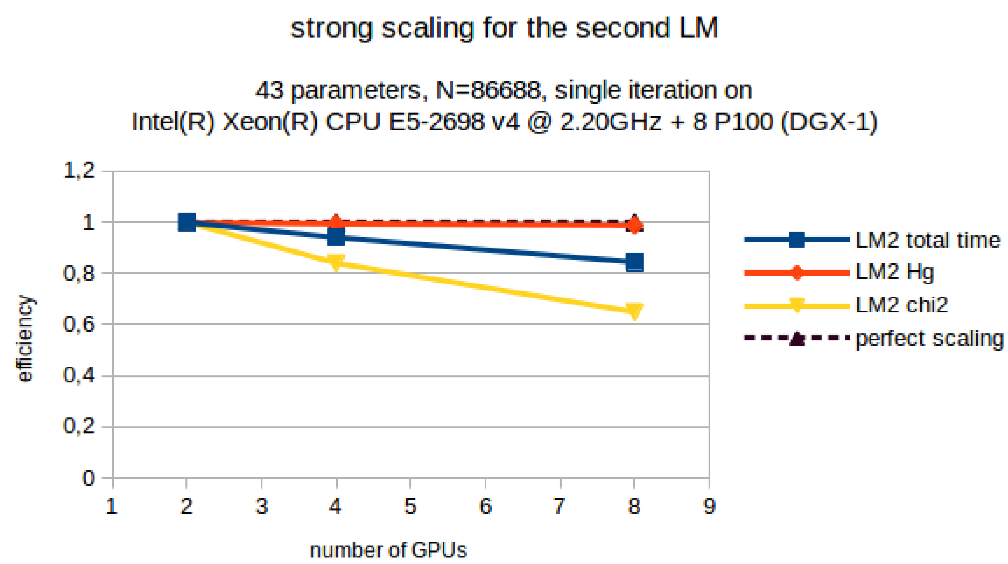
Performance for parameters identification (“Learn” process)

Multi-GPU process, including matrix generation and LM fit

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

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





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

Durham University

MICROGATE

PLDA

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

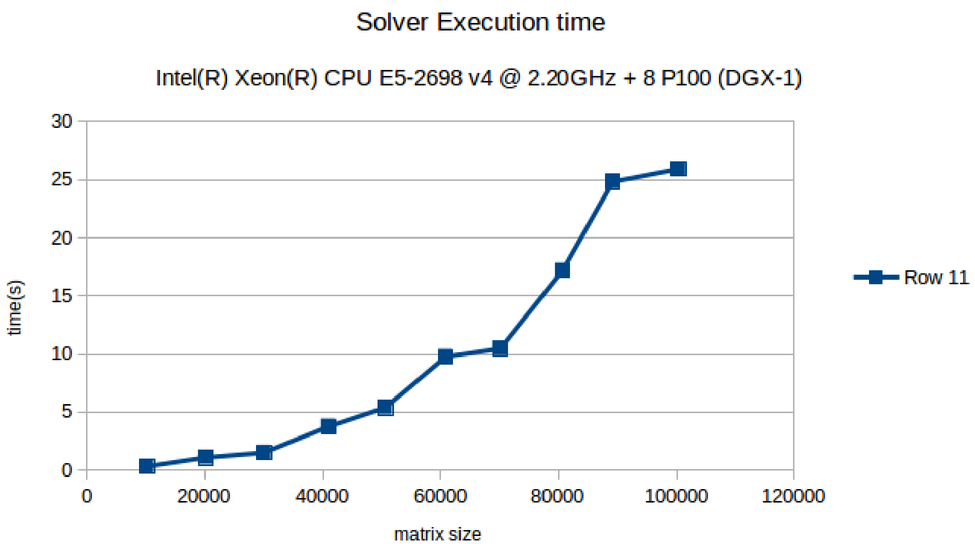
$$R' = Ctm \cdot Cmm_f^{-1}$$

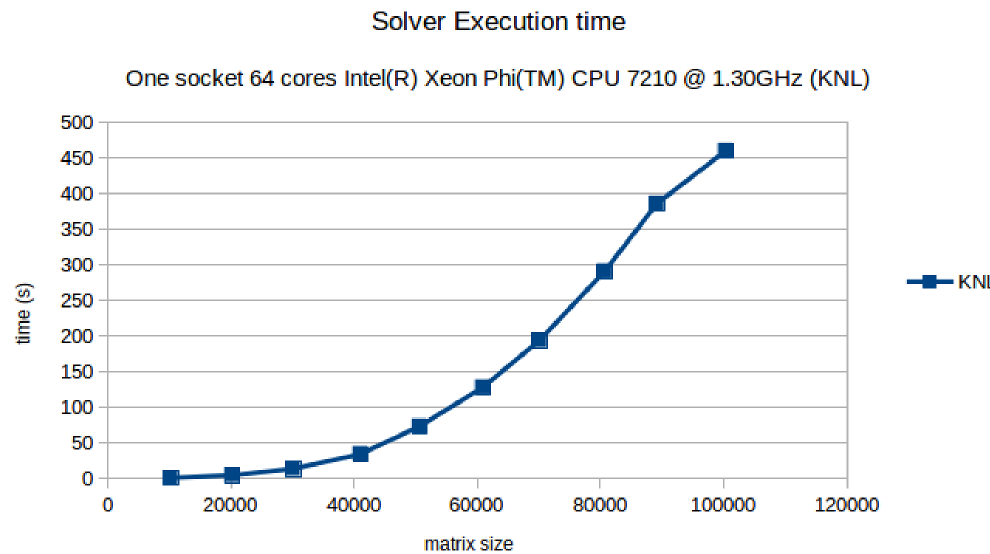
- Can use various methods. "Brute" force : direct solver
- Standard Lapack routine : "posv" : mostly compute-bound, high level of scalability
- Highly portable code : explore various architectures by using standard vendor provided maths libraries

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)



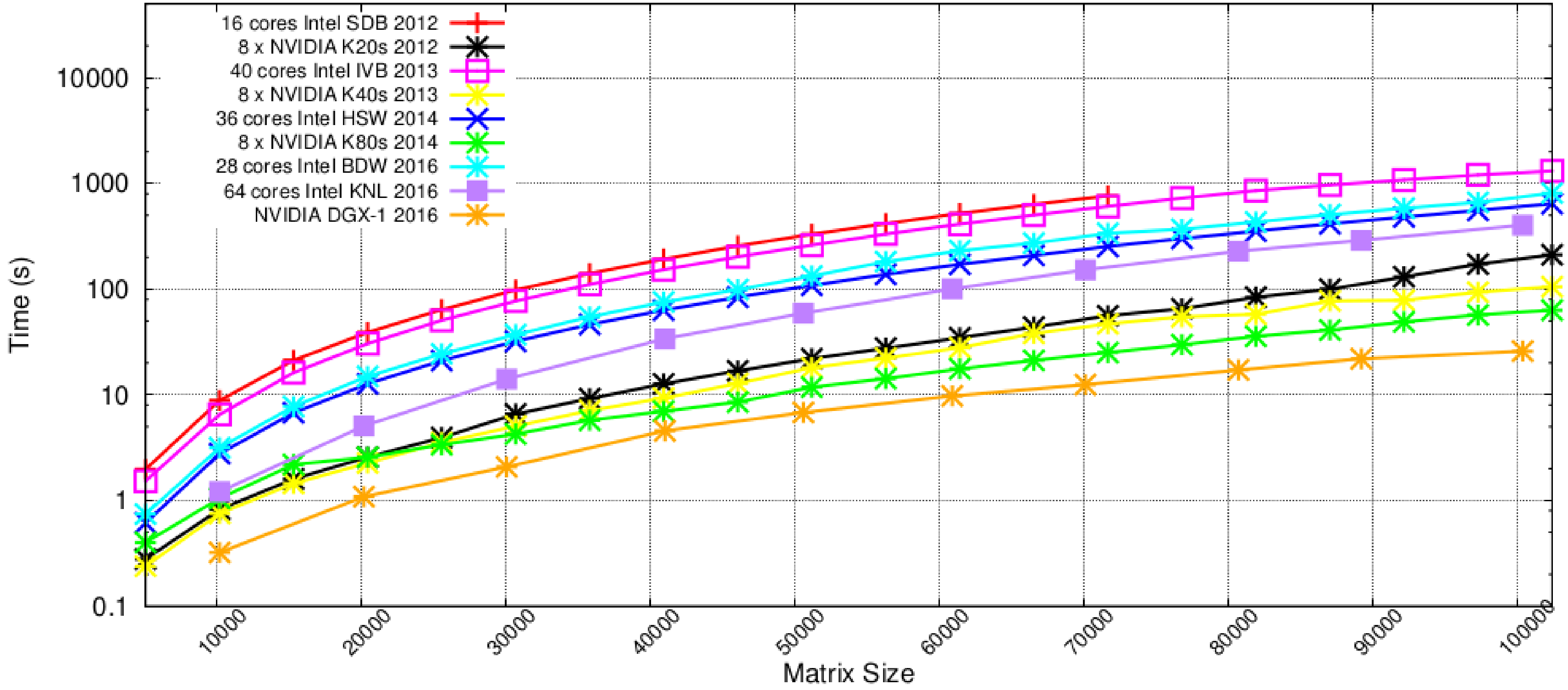


- Record time-to-solution on DGX-1 : MAORY / HARMONI full scale (100k x 100k matrix) : 25sec to compute tomographic reconstructor

Loop supervision module

Performance evolution over time on different platforms

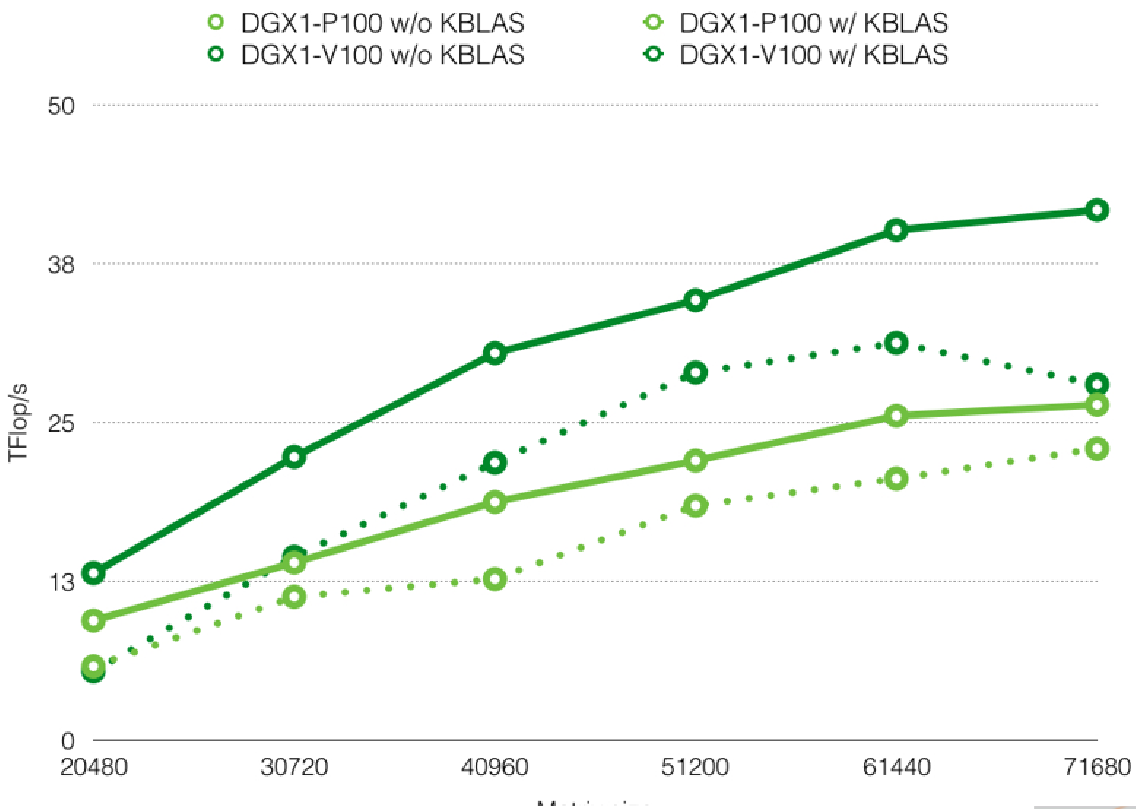
- Comparing generations of GPU and CPUs (+Xeon Phi)



Loop supervision module

State of the art performance on NVIDIA DGX-1 with V100

- Versus P100 using BLAS library from KAUST: x1.6

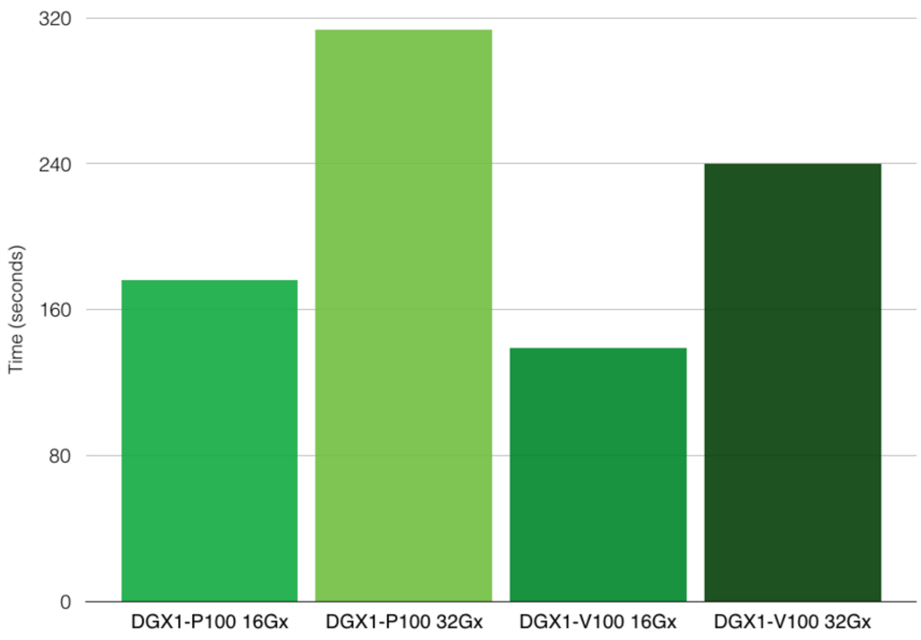


l'Observatoire de Paris — LESIA
Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique
Durham University
MICROGATE
PLDA

Loop supervision module

Time to solution to compute x16 and x32 tomo reconstructors in parallel

- 10s/reconstructor with P100 and 7.5s with V100 !
- Brute force computation of optimal M4 control matrix (averaging over the FoV) is feasible within few minutes
- Here again, we demonstrate that typical system scales are not large enough to feed the newest generations of GPUs with workload efficiently



WP4: deliverables

Task 4.1 (OdP):

- D4.1: GPU cluster for RT-box design and test report (OdP – M6 – submitted)
- D4.2: GPU cluster for RT-box prototype (OdP – M24– submitted)

Task 4.2 (OdP):

- D4.3: GPU cluster for supervisor design and test report (OdP – M6 – submitted)

Task 4.3 (UoD):

- D4.4: Intel Xeon Phi cluster for RT-box prototype design and test report (UoD – submitted)
- D4.5: Intel Xeon Phi cluster for RT-box prototype (UoD – M24 – delayed to M30)

Task 4.4 (UoD):

- D4.6 FPGA cluster for RT-box prototype design and test report (UoD – M24 – submitted)
- D4.7: FPGA cluster for RT-box prototype (UoD – M36)