# Green Flash

High performance computing for real-time science

European HPC summit week 2016
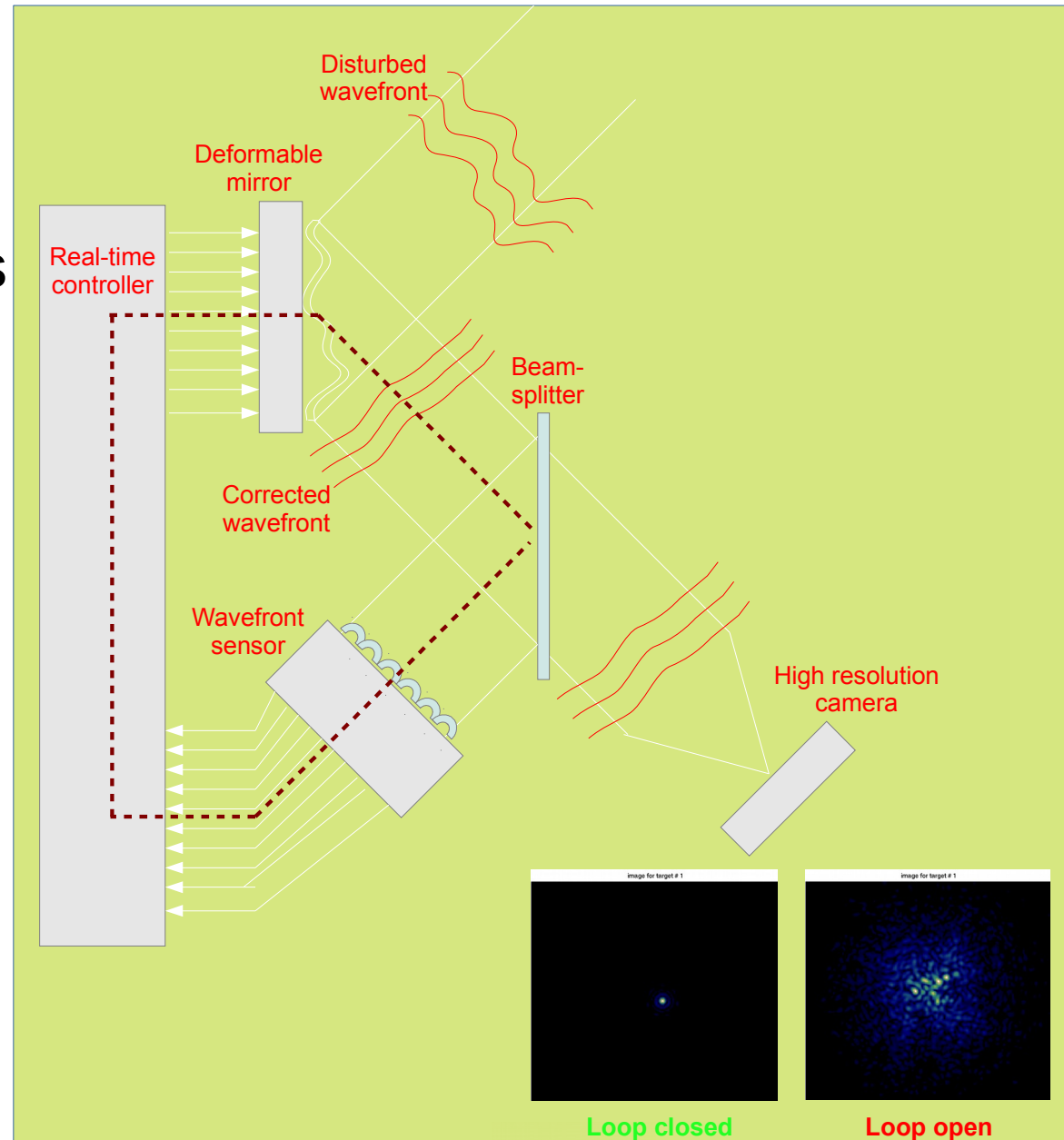
# European Extremely Large Telescope

- 39m diameter telescope : x5 in diameter => x25 in system complexity

  - 100m dome, 2800 tones structure rotating @ 360°, seismic safe (Chile)

  - 1.2 G€ project, first light foreseen in 2024

  - Construction led by ESO (European Southern Observatory), international organisation funded by 15 European countries

  - Telescope components + science instruments built by european research labs + industrial partners
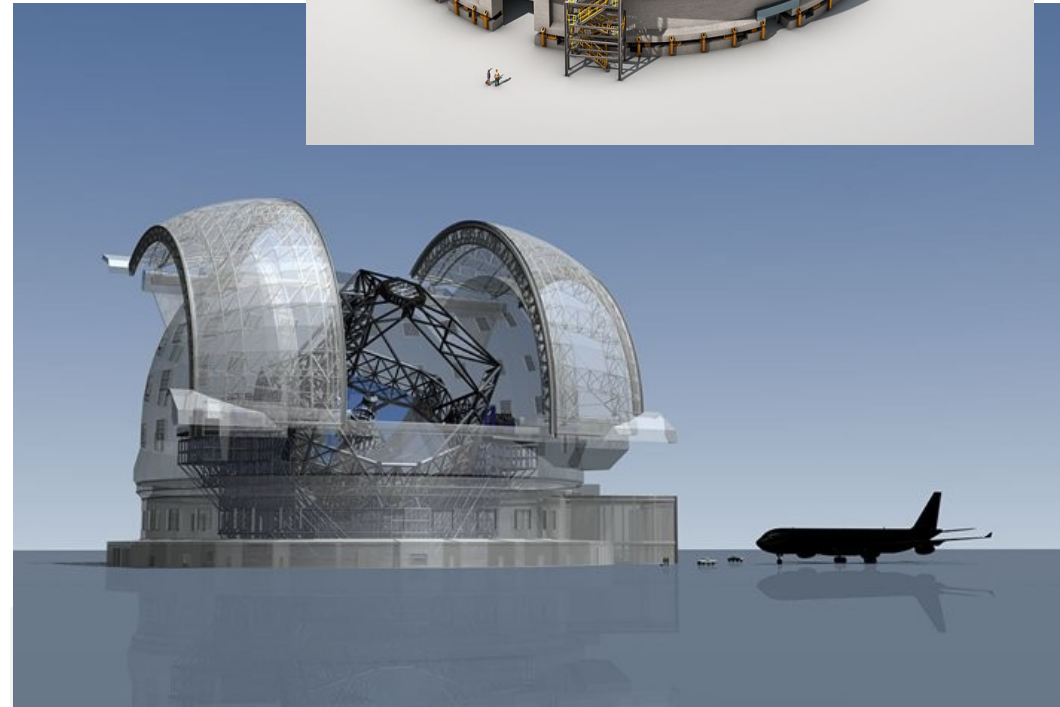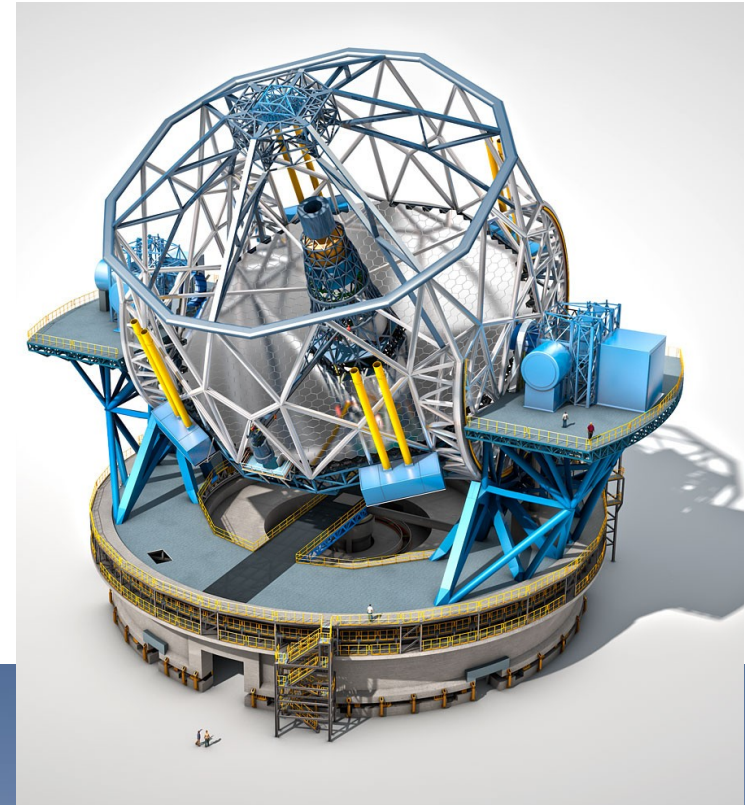
# Adaptive optics
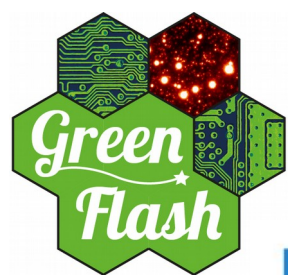
- Compensate in real-time the wavefront perturbations

- Using a wavefront sensor to measure them

- Using a deformable mirror to reshape the wavefront

- Commands to the mirror must be computed in real-time (1ms rate)



Disturbed wavefront

Deformable mirror

Real-time controller

Beam-splitter

Corrected wavefront

Wavefront sensor

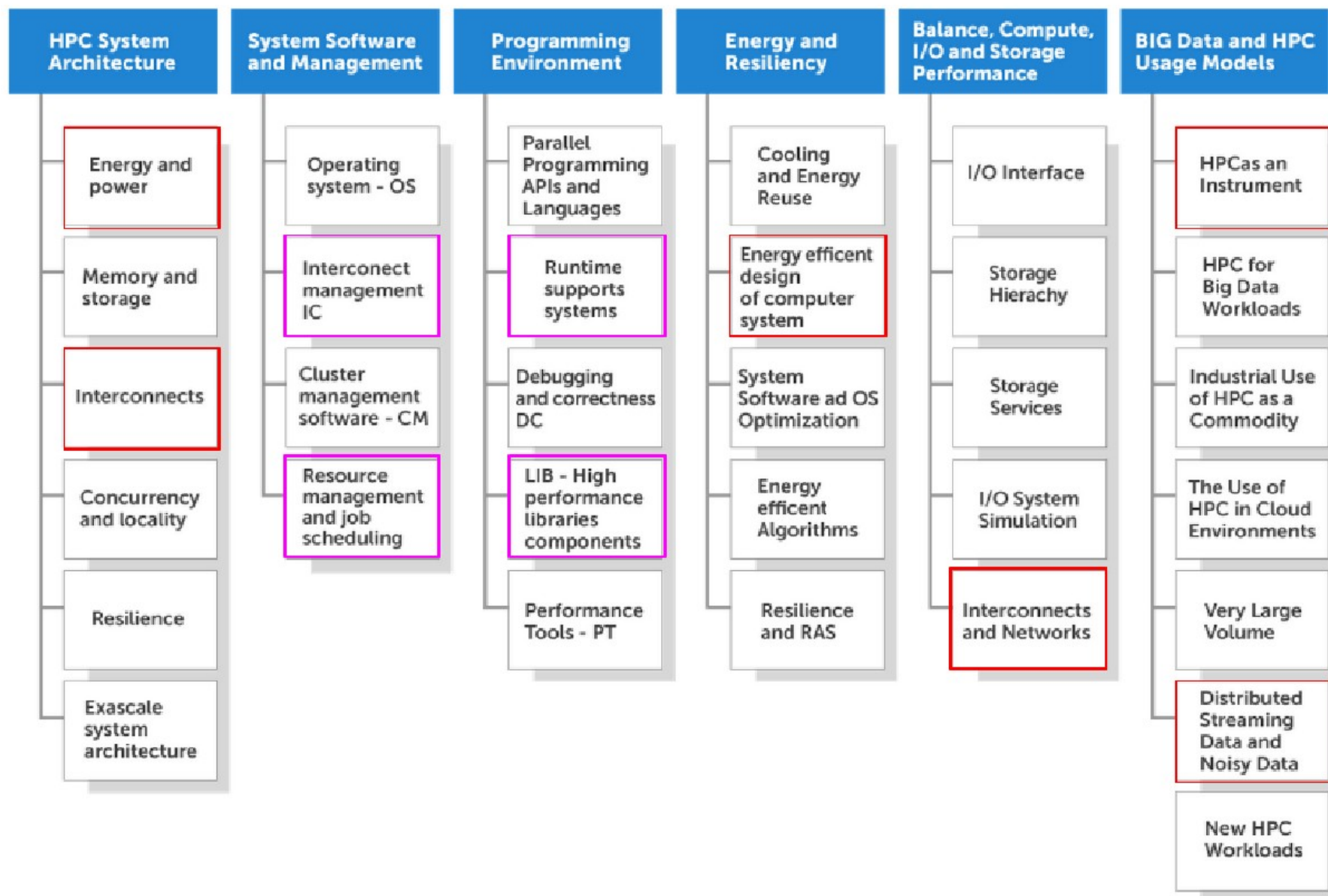High resolution camera

Loop closed

Loop open

# Introduction to Green Flash

- Program objectives: 3 research axes
  - 2 technological developments and 1 validation study
- Real-time HPC using accelerators and smart interconnects
  - Assess the determinism of accelerators performance
  - Develop a smart interconnect strategy to cope for strong data transfer bandwidth constraints
- Energy efficient platform based on FPGA for HPC
  - Prototype a main board, based on FPGA SoC and PCIe Gen3
  - Cluster such boards and assess performance in terms of energy efficiency and determinism
- AO RTC prototyping and performance assessment
  - Assemble a full functionality prototype for a scalable AO RTC targeting the MAORY system
  - Compare off-the-shelf solutions based on accelerators and new FPGA-based concept
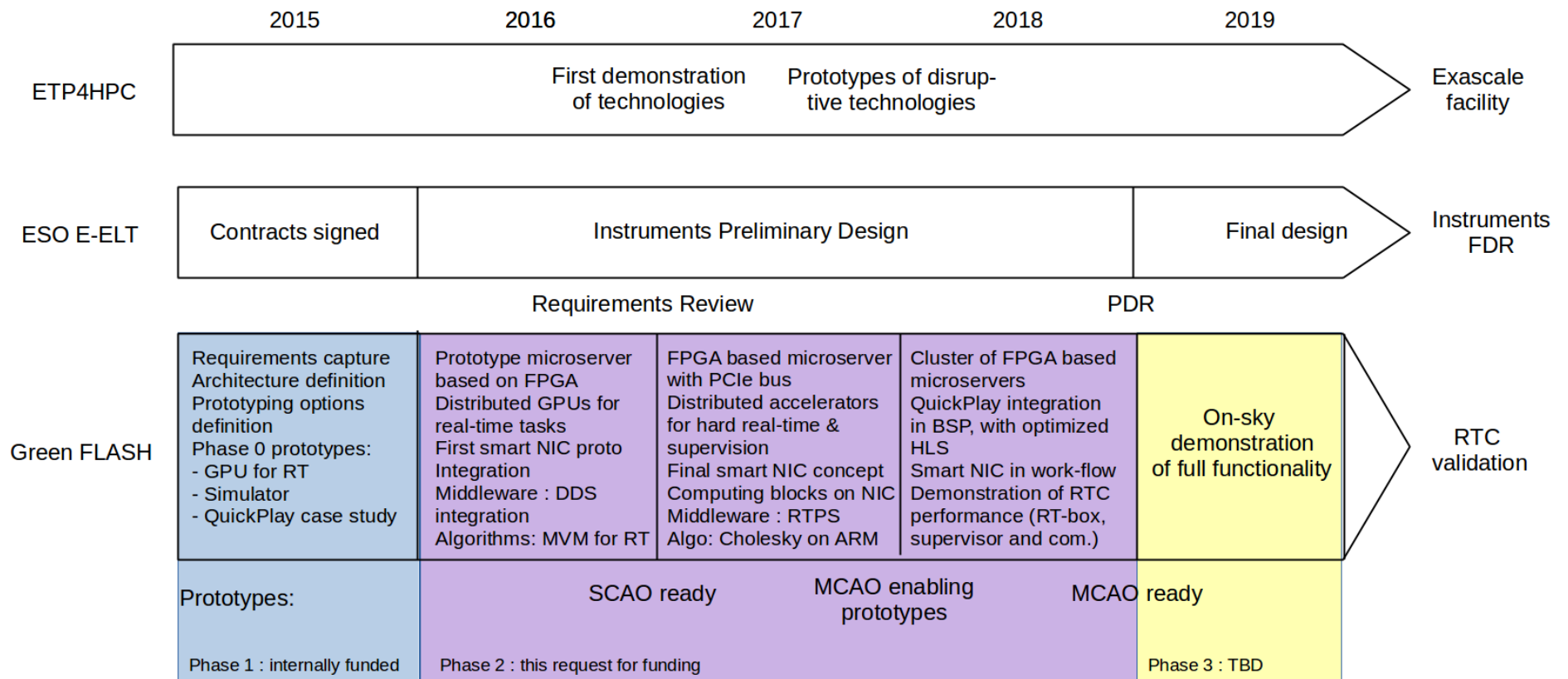
# Addressing HPC roadmap to exascale

| HPC System Architecture | System Software and Management | Programming Environment | Energy and Resiliency | Balance, Compute, I/O and Storage Performance | BIG Data and HPC Usage Models |
|---|---|---|---|---|---|
| Energy and power | Operating system - OS | Parallel Programming APIs and Languages | Cooling and Energy Reuse | I/O Interface | HPCas an Instrument |
| Memory and storage | Interconect management IC | Runtime supports systems | Energy effient design of computer system | Storage Hierachy | HPC for Big Data Workloads |
| Interconnects | Cluster management software - CM | Debugging and correctness DC | System Software ad OS Optimization | Storage Services | Industrial Use of HPC as a Commodity |
| Concurrency and locality | Resource management and job scheduling | LIB - High performance libraries components | Energy effient Algorithms | I/O System Simulation | The Use of HPC in Cloud Environments |
| Resilience | | Performance Tools - PT | Resilience and RAS | Interconnects and Networks | Very Large Volume |
| Exascale system architecture | | | | | Distributed Streaming Data and Noisy Data |
| | | | | | New HPC Workloads |

# Green Flash project

- ## Partners
  - ### 2 academic partners
    - LESIA, Observatoire de Paris, P.I. Damien G.
    - CfAI, University of Durham
  - ### 2 industrial partners
    - Microgate : Italian SME designing FPGA solutions for various applications (including astronomical AO)
    - PLDA: french SME developing FPGA solutions (mostly IP cores, world leader in PCIe IP)

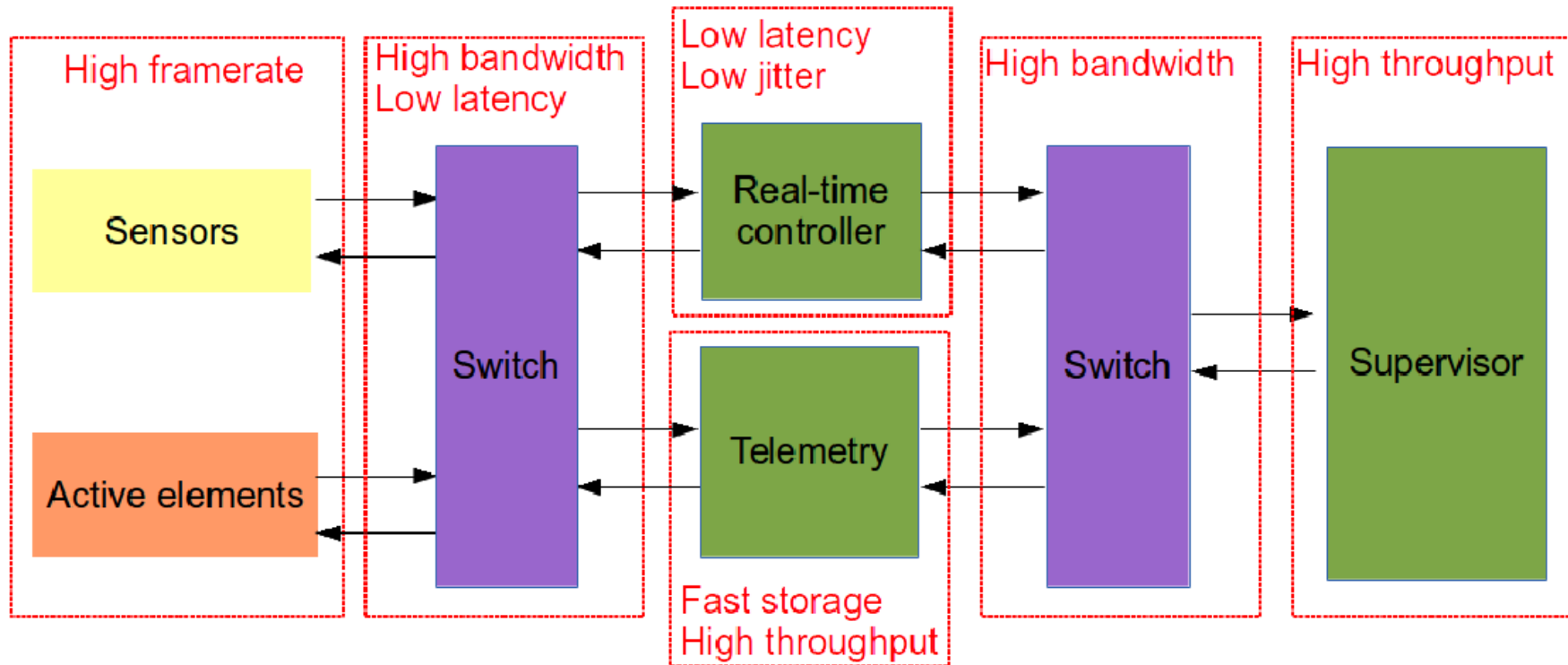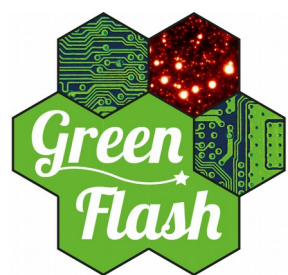- ## 3.8M€ grant, 36 months work plan

# Project Management
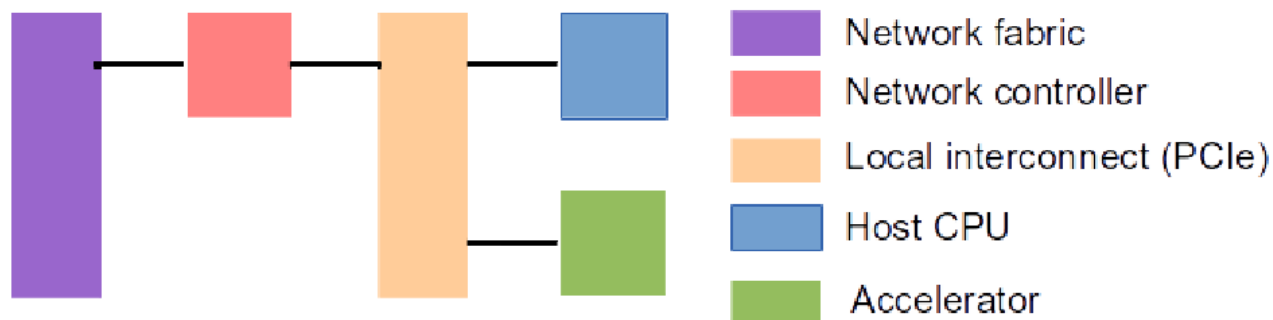
- Good convergence with H2020 ETP4HPC / E-ELT project timeline

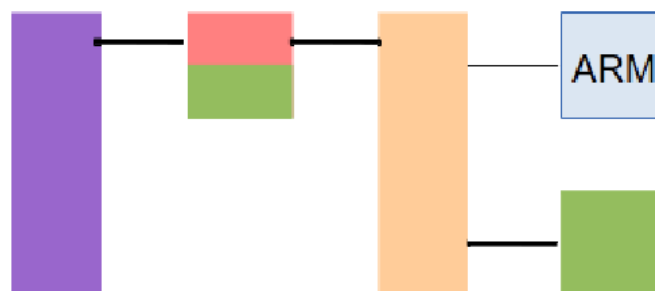| | 2015 | 2016 | 2017 | 2018 | 2019 | |
|---|---|---|---|---|---|---|
| ETP4HPC | | First demonstration of technologies | Prototypes of disruptive technologies | | | Exascale facility |
| ESO E-ELT | Contracts signed | Instruments Preliminary Design | | | Final design | Instruments FDR |
| | | Requirements Review | | PDR | | |
| Green FLASH | Requirements capture Architecture definition Prototyping options definition Phase 0 prototypes: - GPU for RT - Simulator - QuickPlay case study | Prototype microserver based on FPGA Distributed GPUs for real-time tasks First smart NIC proto Integration Middleware : DDS integration Algorithms: MVM for RT | FPGA based microserver with PCIe bus Distributed accelerators for hard real-time & supervision Final smart NIC concept Computing blocks on NIC Middleware : RTPS Algo: Cholesky on ARM | Cluster of FPGA based microservers QuickPlay integration in BSP, with optimized HLS Smart NIC in work-flow Demonstration of RTC performance (RT-box, supervisor and com.) | On-sky demonstration of full functionality | RTC validation |
| | Prototypes: | SCAO ready | MCAO enabling prototypes | MCAO ready | | |
| | Phase 1 : internally funded | Phase 2 : this request for funding | | | Phase 3 : TBD | |

# AO RTC concept

# Assessing new HPC concepts



**Current trend in heterogeneous HPC**

Legend:
- Network fabric
- Network controller
- Local interconnect (PCIe)
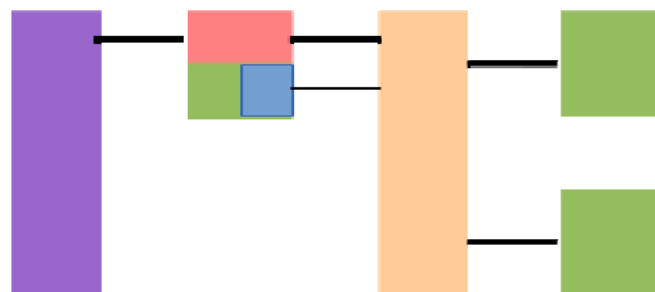- Host CPU
- Accelerator

**Accelerators, ARM64 & smart interconnect**

ARM

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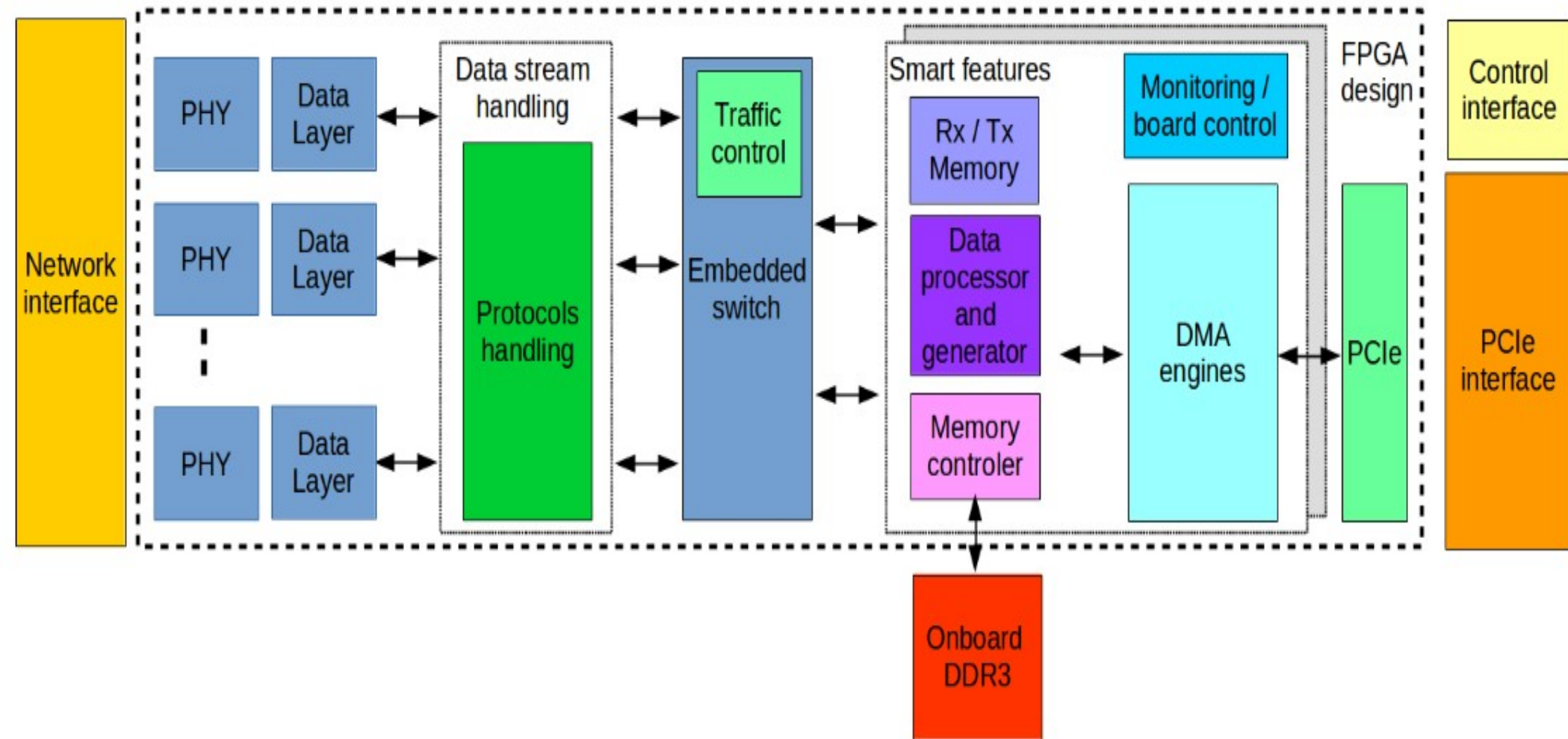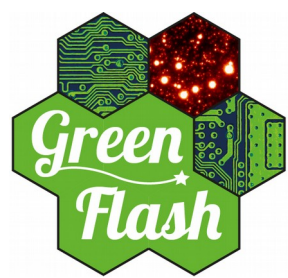Features accessible in middleware

**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

# Smart Interconnect Subsystem Architecture

# First prototype: hardware at work

- Reflex CES XpressKUS
  - Standard profile
  - XCKU040-2FFVA1156 (20 xceivers)
  - Up to PCIe 3.0 x8, with backward capability to PCIe 1.0 and 2.0
  - DDR3 SDRAM SODIMM support up to 8GB
  - HPC FMC interface
  - RoHS and REACH compliant

- Faster Technology FM-S14
  - Mezzanine Card (FMC) module
  - up to four SFP/SFP+ module
  - interfaces directly into Multi-Gigabit Transceivers (MGTs) of a Xilinx FPGA.

*(\*) Reflex CES is a PLDA Group company*

# Reducing Barriers to FPGA Adoption

## QuickPlay Platform

- C/C++ Entry & Debug
  (HDL Entry for Advanced Users)

- Hardware Abstraction
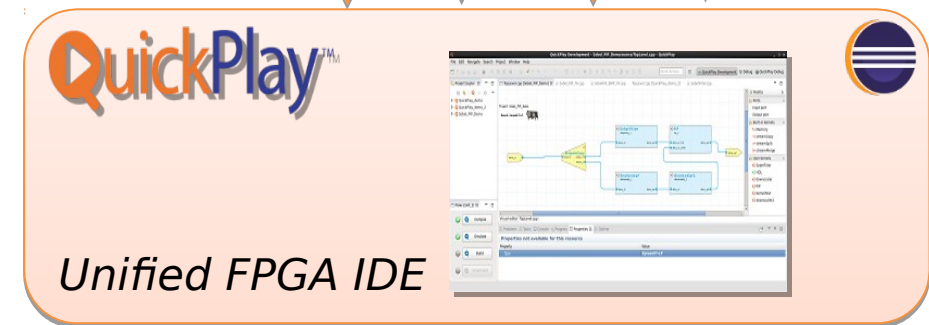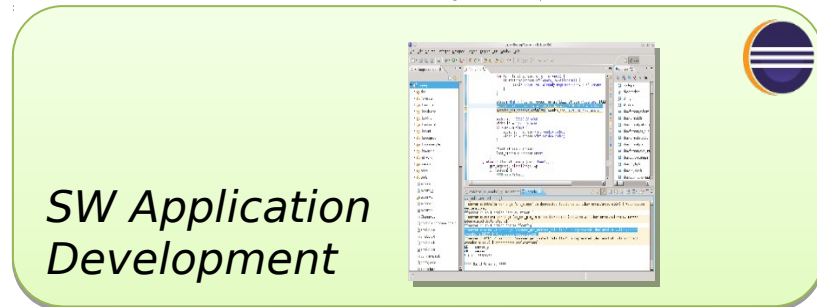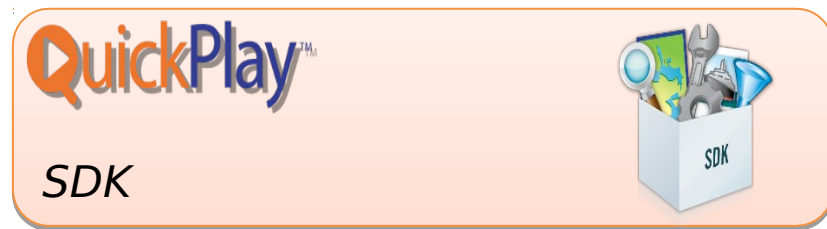
⊹ Bridge the HW/SW Gaps

## Ecosystem

- Boards, IP Cores, Libraries

- Value sharing

⊹ QuickPlay as a channel.

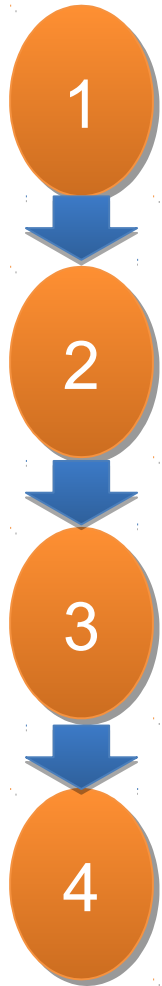## Business Model

- Online Market Place

- Disruptive Licensing Mechanism

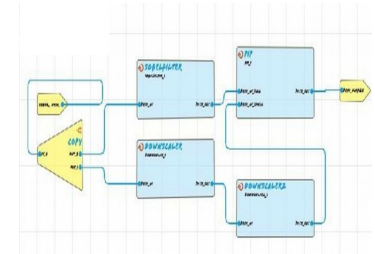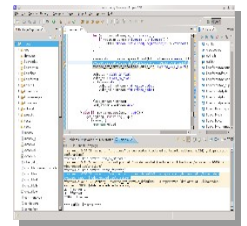⊹ Enable innovative business models by our partners

# The QuickPlay Development Platform

# FPGA Design with QuickPlay IDE

**1**
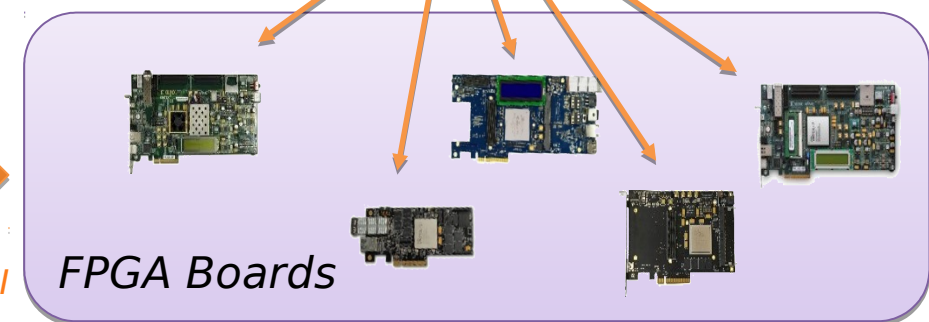
## MODEL
&ndash; C/C++ functional modeling

**2**

## VERIFY & VALIDATE
&ndash; Desktop execution of system functional model

**3**

## BUILD
&ndash; Hardware implementation: HLS, Logic Synthesis, P&R

**4**

## EXECUTE
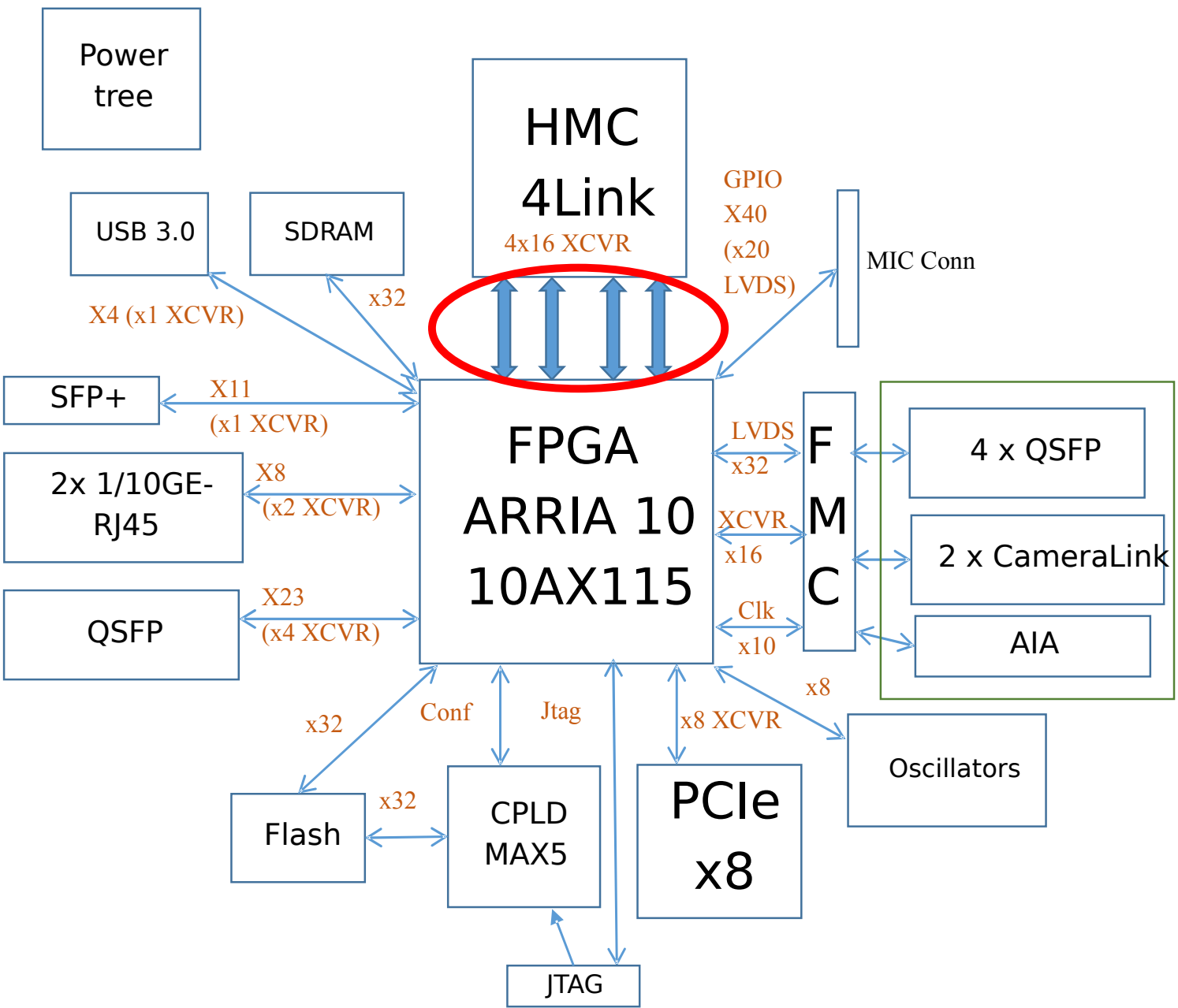&ndash; FPGA based system hardware execution

➡ "Design the way you think"

➡ Pure functional modelling



C, HDL or IP

Untyped Streaming Channels

Virtual Streaming I/Os
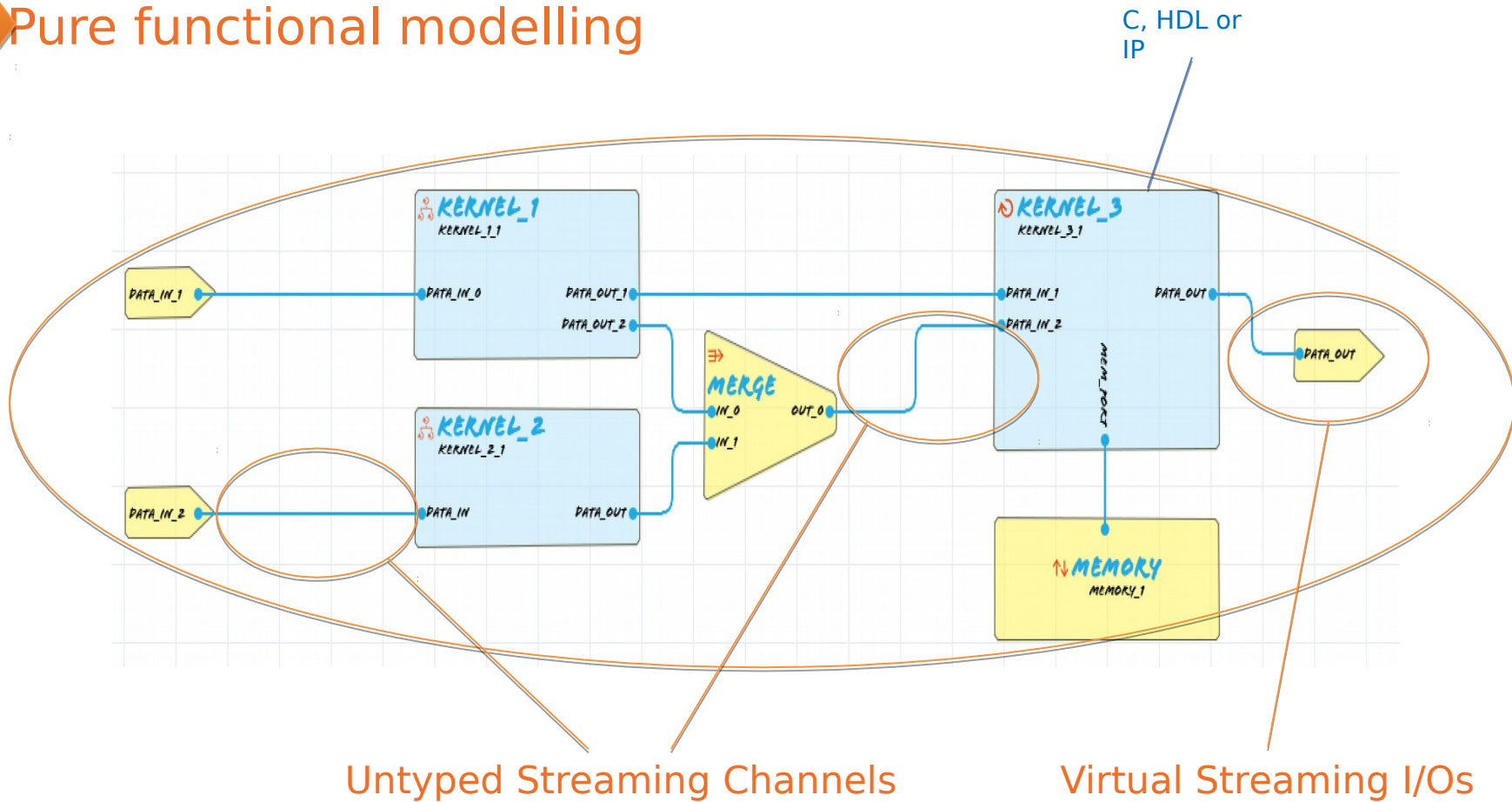
# Scheme 1 – Single FPGA system

4

**Power tree**

**USB 3.0**

**SDRAM**

## HMC 4Link
4x16 XCVR

GPIO X40 (x20 LVDS)

MIC Conn

X4 (x1 XCVR)

x32

**SFP+**

X11 (x1 XCVR)

**FPGA ARRIA 10 10AX115**

LVDS x32

**FMC**

**4 x QSFP**

**2x 1/10GE-RJ45**

X8 (x2 XCVR)

XCVR x16

**2 x CameraLink**

**QSFP**

X23 (x4 XCVR)

Clk x10

**AIA**

x8

x32

Conf

Jtag

x8 XCVR

**Oscillators**

**Flash**

x32

**CPLD MAX5**

**PCIe x8**

**JTAG**

ARRIA 10AX115
- 1518 DSP blocks
- 6.6MB int. RAM
- max. 96 XCVR

PROS:
- Max. Bandwidth between HMC and FPGA - 4 Links
- Easier power tree because only one FPGA and HMC
- Shorter length of board possible then 2 FPGA version

CONS:
- No ARM co-processor only NIOS II in logic
- No powerful PCIe root port because no ARM
- More expensive than scheme 2

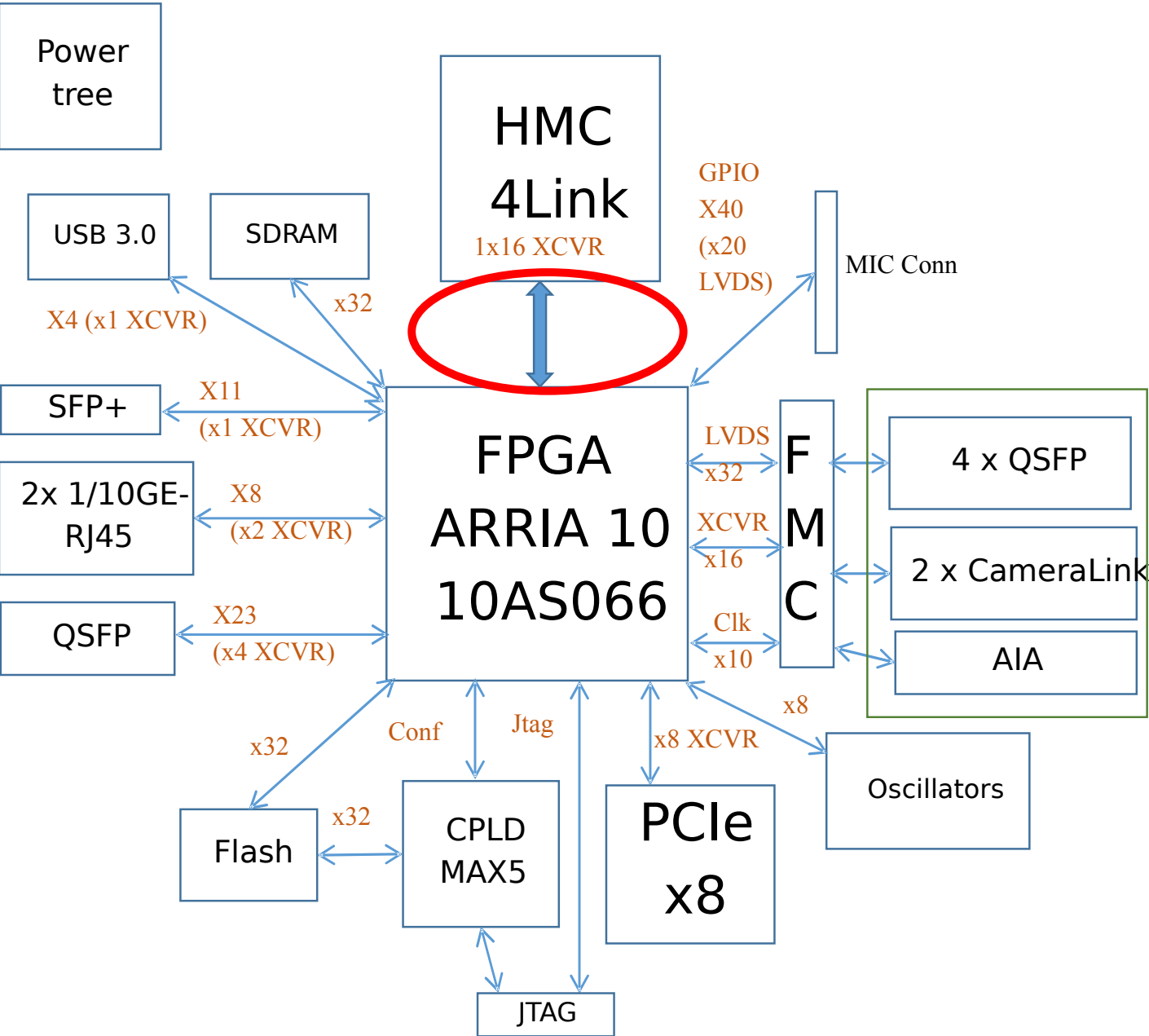# Scheme 2 – FPGA with ARM coprocessor system

5

ARRIA 10AS066 SoC

- 1.5GHz ARM dual-core Cortex-A9 co-processor
- 1855 DSP blocks
- 5.2MB int. RAM
- max. 48 XCVR

PROS:

- ARM Co-Processor for stand-alone real-time box
- Powerful PCIe root port because of ARM and OS
- Shorter length of board possible then 2 FPGA version
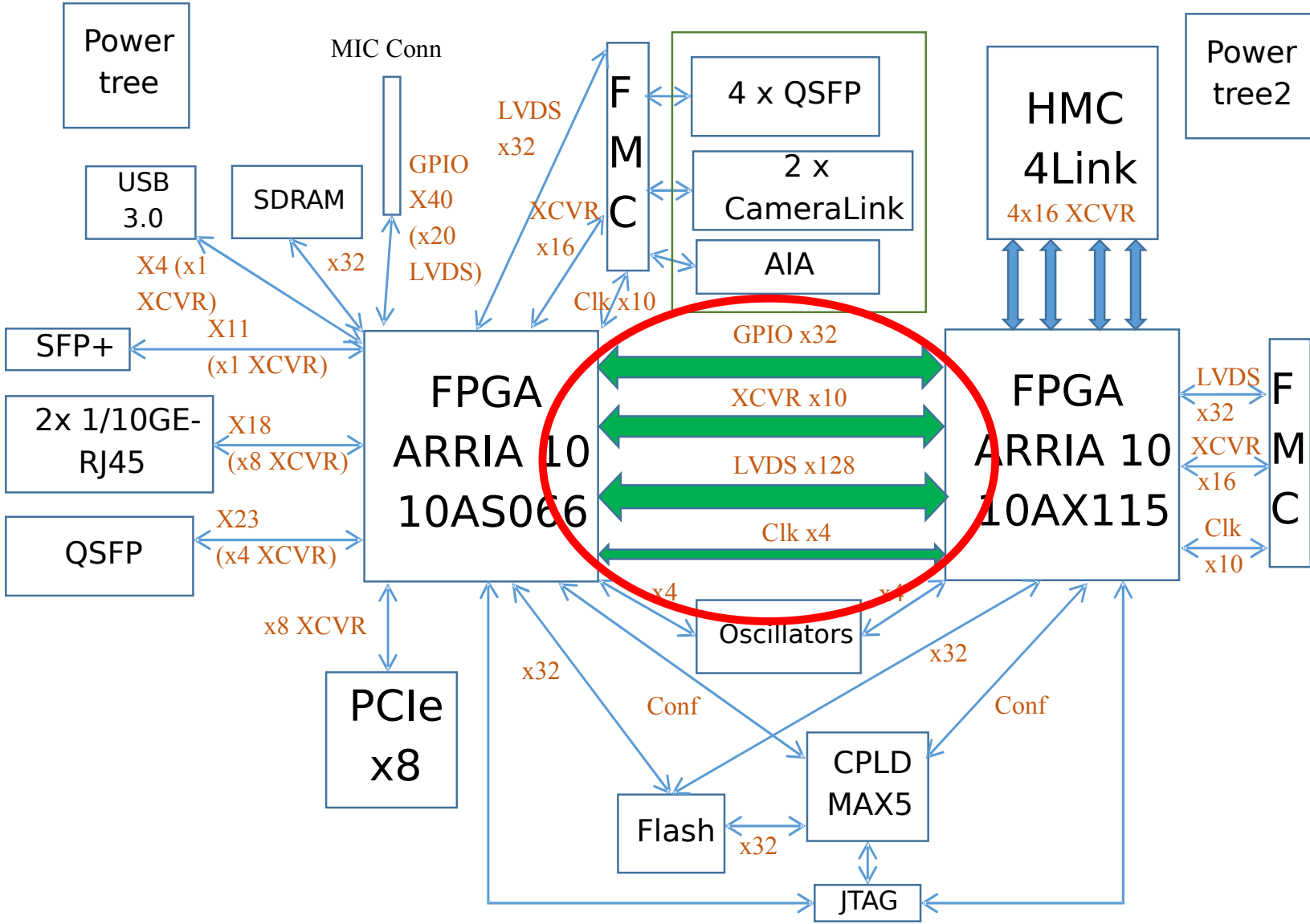- Less expensive than Scheme 1

CONS:

- Less Transceiver
- 1/4 Bandwidth between HMC and FPGA - 1 Link
- Reduced computation because ext. mem bottleneck, less int. mem

Power tree

USB 3.0    SDRAM

HMC 4Link

1x16 XCVR

GPIO X40 (x20 LVDS)

MIC Conn

X4 (x1 XCVR)    x32

SFP+    X11 (x1 XCVR)

2x 1/10GE-RJ45    X8 (x2 XCVR)

QSFP    X23 (x4 XCVR)

FPGA ARRIA 10 10AS066

LVDS x32    F M C

XCVR x16

Clk x10

4 x QSFP

2 x CameraLink

AIA

x8

Conf    Jtag    x8 XCVR

x32    x32

Flash    x32    CPLD MAX5    PCIe x8

Oscillators

JTAG

# Scheme 3 – Two FPGA System

6



Power tree

MIC Conn

USB 3.0

SDRAM

GPIO X40 (x20 LVDS)

LVDS x32

XCVR x16

Clk x10

X4 (x1 XCVR)

X11 (x1 XCVR)

SFP+

2x 1/10GE-RJ45

X18 (x8 XCVR)

X23 (x4 XCVR)

QSFP

x8 XCVR

PCIe x8

x32

Conf

FPGA ARRIA 10 10AS066

x32

FMC

4 x QSFP

2 x CameraLink

AIA

GPIO x32

XCVR x10

LVDS x128

Clk x4

Oscillators

x4

Flash

x32

CPLD MAX5

JTAG

HMC 4Link

4x16 XCVR

Power tree2

FPGA ARRIA 10 10AX115

LVDS x32

XCVR x16

Clk x10
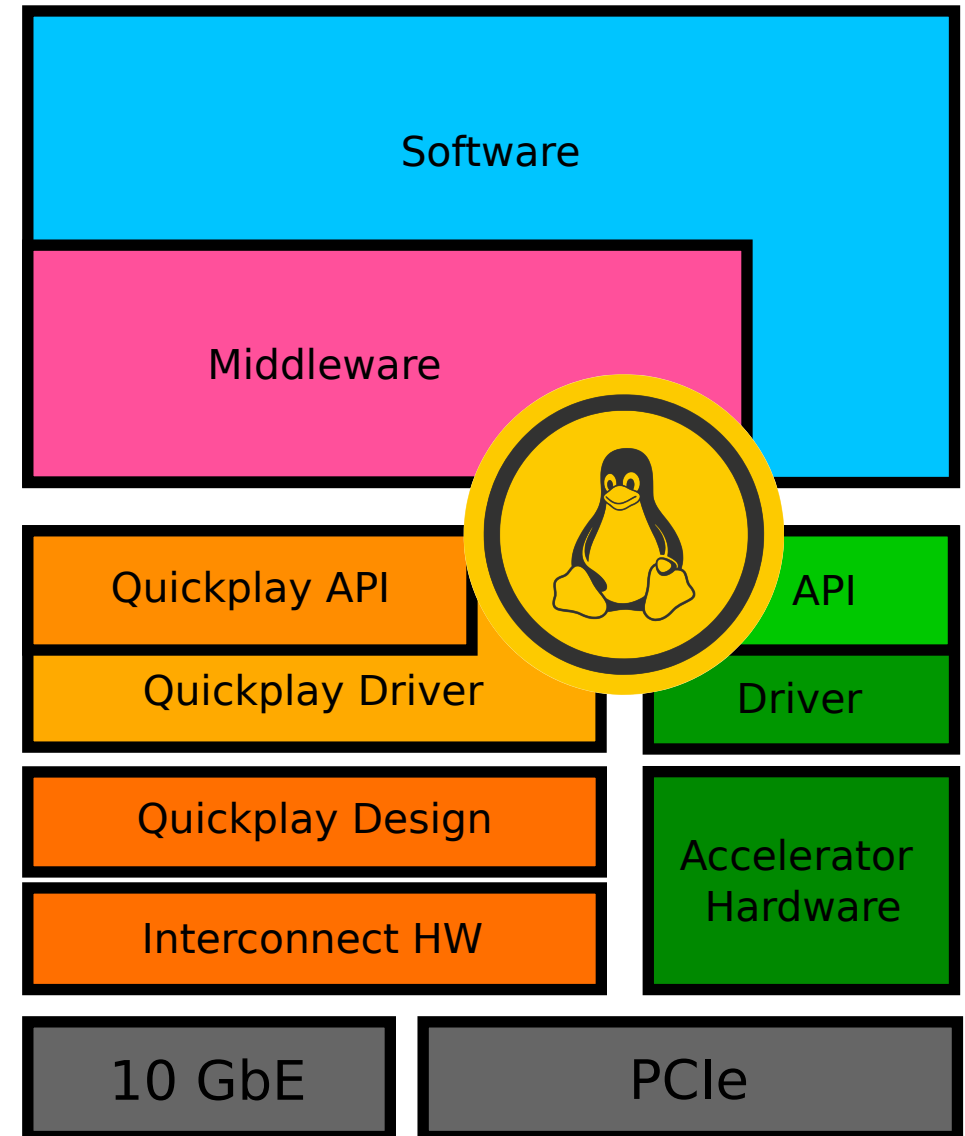
FMC

x32

Conf

# FPGA design of microserver board

- Integration of Quick Play

- Coordination of efforts between 2 european SMEs

- Bring together smart interconnect concept with energy efficient compute platform
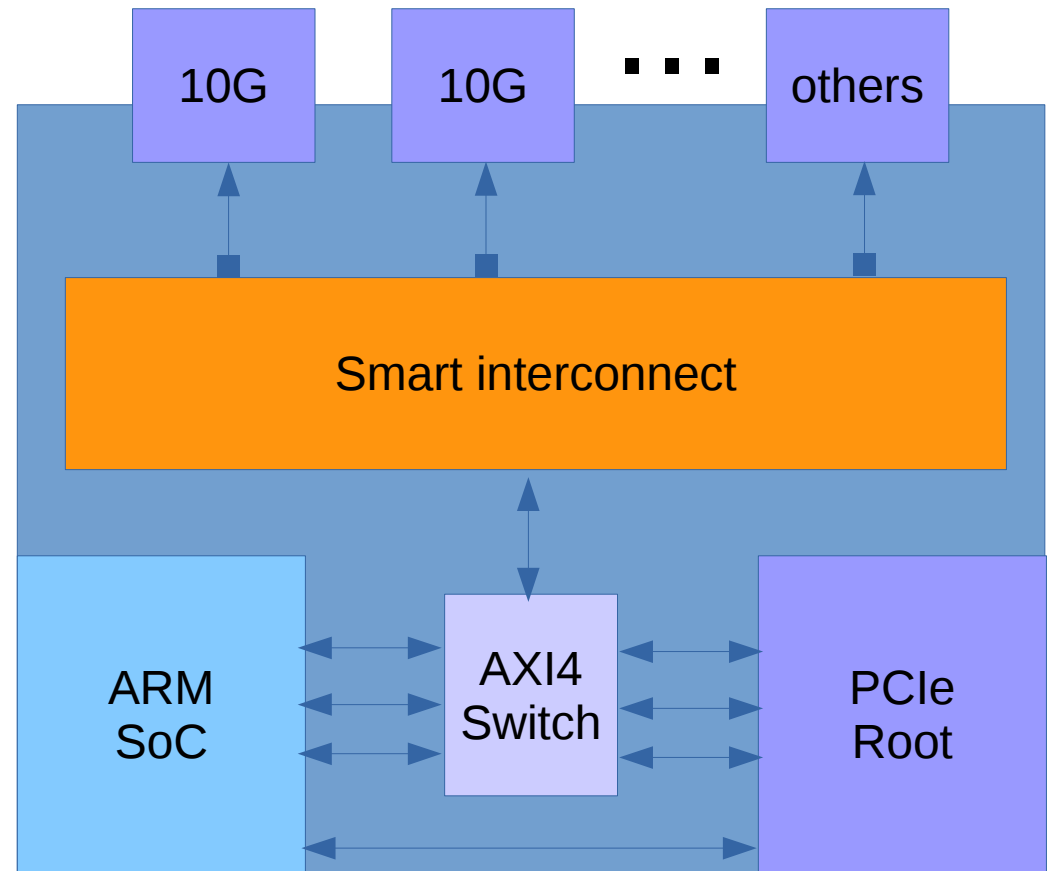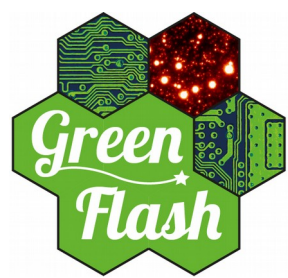
# SW / MW stack

- Under development at academic partners

- Run a standard HPC ecosystem on the prototype board

- Performance assessed w/r AO application (mainly linear algebra)



Software

Middleware

Quickplay API

Quickplay Driver

Quickplay Design

Interconnect HW

API

Driver

Accelerator Hardware

10 GbE

PCIe

# Working on the application side

- 2016 Gordon Bell award submission : 3.5 PFLOPs (60% of peak perf) on custom application

**Extreme Scale Simulations of Multi-object Adaptive Optics for the European Extremely Large Telescope**

Hatem Ltaief[1], Damien Gratadour[2,3], Ali Charara[1], Saber Feki[4], Eric Gendron[2], Bilel Hadri[4], Ahmad Abdelfattah[5], and David Keyes[1]

[1] Extreme Computing Research Center, KAUST, Saudi Arabia

[2] LESIA, Observatoire de Paris, France

[3] Université Paris Diderot, France

[4] KAUST Supercomputing Lab, KAUST, Saudi Arabia

[5] Innovative Computing Laboratory, Knoxville TN, USA

**ABSTRACT**
The European Extremely Large Telescope (E-ELT) is one of today's most challenging projects in ground based astronomy. Addressing one of the key science cases for the E-ELT,

**1. PERFORMANCE ATTRIBUTES**

| Category of achievement | Scalability Time-to-solution Peak performance |
|---|---|
| Type of method | Direct solve |
| Results | Whole application |
| Precision | Double precision |
| System scale | two third of full system |
| Metrics | Time Flops/s |

Table 1: List of Performance Attributes.
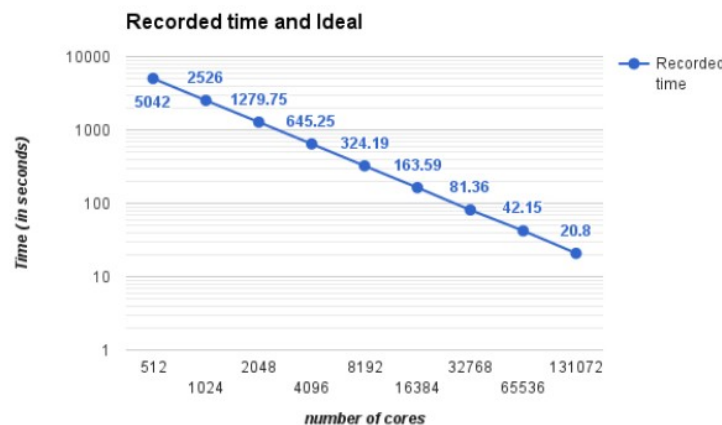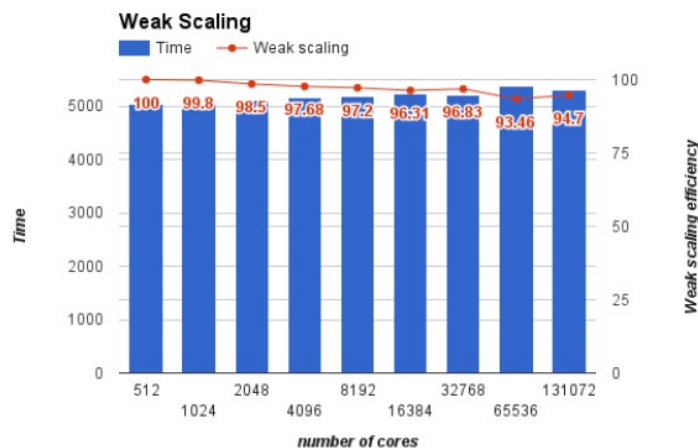


**Weak Scaling**

(a) Nearly perfect weak scalability of the code from 512 up to



**Recorded time and Ideal**

(b) Time to solution per snapshot: weak scalability from 512

# That's it for today ! Thank you