FET Innovation Launchpad (FETOPEN-03-2018-2019-2020)

The EDRA project: Bringing Efficient Reconfigurable Architectures to the Amazon Cloud

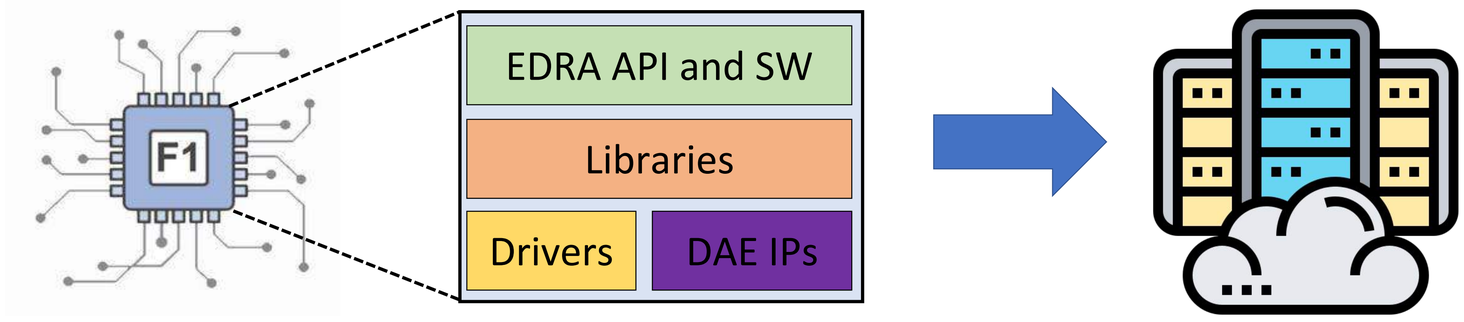
Dimitris Theodoropoulos, Nikolaos Alachiotis, Andreas Brokalakis, Dionisios Pnevmatikatos
dtheodoropoulos@isc.tuc.gr, nalachiotis@mhl.tuc.gr, abrokalakis@mhl.tuc.gr, pnevmati@ece.tuc.gr

This project is funded from the European Union's Horizon 2020 research and innovation programme "FET Innovation Launchpad" under grant agreement No #851631. **Project budget:** 100,000€ **Duration:** 1/5/2019 – 31/10/2020 https://edra-project.eu/

EDRA in a nutshell

1. Automate software workflow for generating Decoupled Access / Execute (DAE) hardware accelerators
2. Integrate DAE-compatible SW / HW interface with Amazon's FPGA cloud infrastructure
3. Publish Amazon Machine Images hosting DAE accelerators in the domain of Bioinformatics

EDRA machine image



FPGAs in datacenters

- **X**aaS ➔ **X** is Infrastructure, Platform, Software, Function…
- Cloud service revenue is forecasted from **182.4** b$ in 2018 to reach **331 b$ by 2022**
- FPGA-as-a-Service

Data Center Accelerator Market size was 3.5 billion US$ and it is expected to reach 35 billion US$ by the end of 2025, with a CAGR of 38.7% during 2019-2025

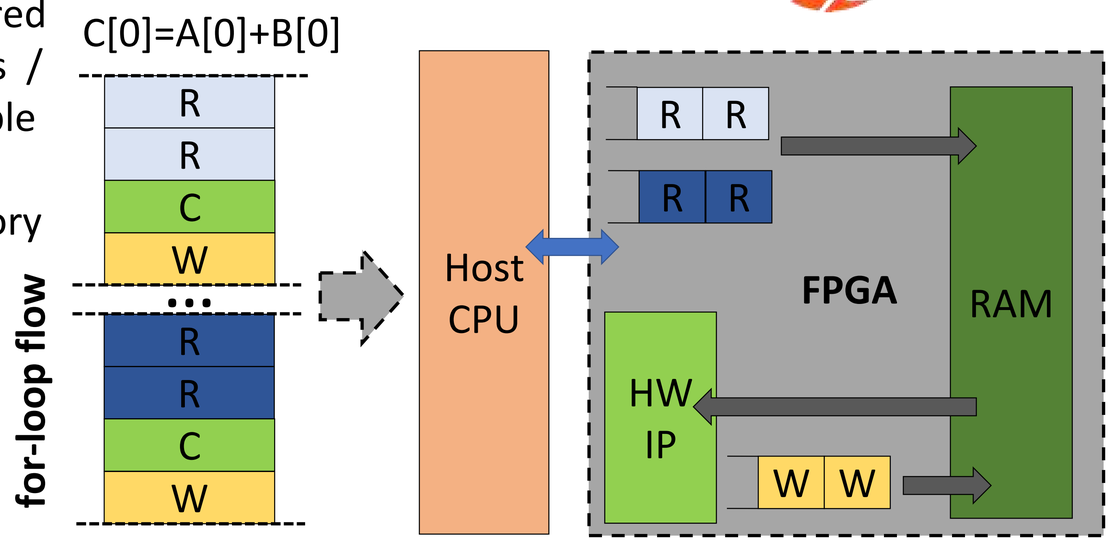
- Current providers

Background

https://www.extrahpc.eu

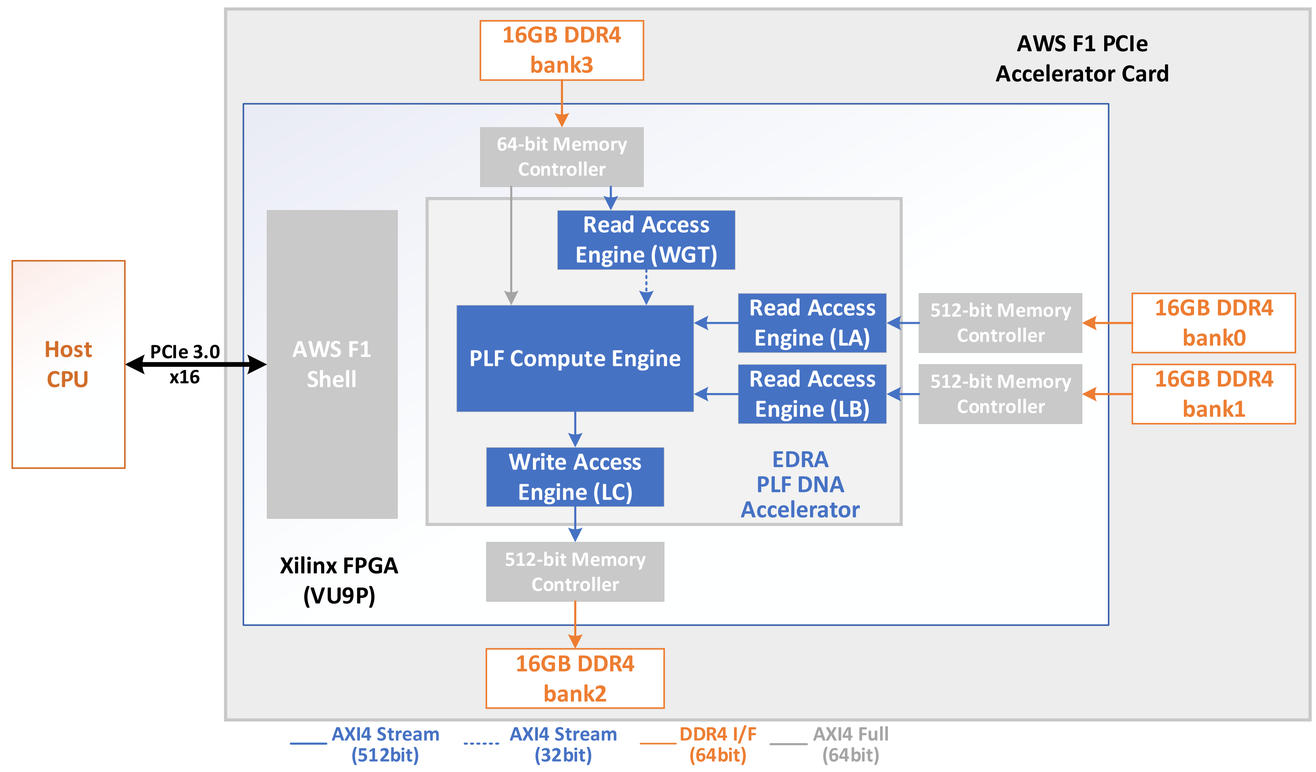
The EXTRA FET HPC H2020 project delivered the Decoupled Access / Execute Reconfigurable framework

- Decoupling memory access from computation
- Multiple read / write streaming requests
- Hardware-accelerated data processing



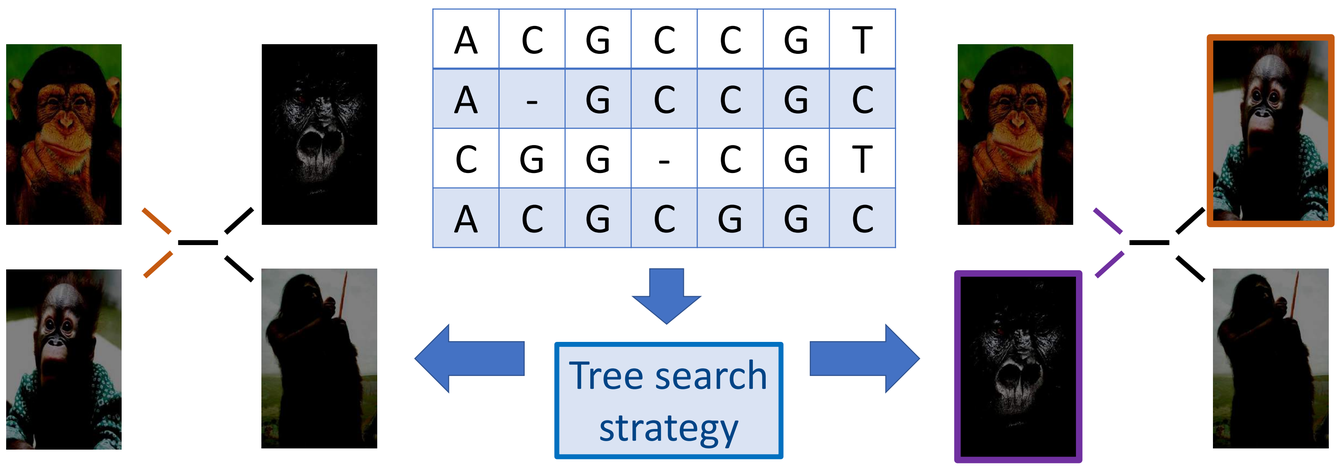
The EDRA architecture

- Hosted on Amazon's F1 machine instances
- AWS shell interface (SH)
  - PCIe communication with host CPU
  - Configuration / management to control accelerator from host s/w
- Custom Logic (CL) – The EDRA Accelerator
  - Wide Read / write data streaming ports to memory controllers
  - Parallel access to all DDR4 memory banks to maximize effective throughput
  - Double buffering to minimize data transfer latency between host and accelerator



First results

DAE accelerator for Phylogenetic tree reconstruction



- Resources Used (overall, including shell resources):
  - **LUTs: 27% / REGs: 23% / DSPs: 21% / BRAMs: 17%**
- Clock Frequency: **222MHz**
- *Speedup achieved vs software implementation*: **5,7x**
- Comparison parameters:
  - Software baseline refers to official sequential RAxML code compiled with supplied optimization flags and executed on Xeon E5v4 @2,7GHz