INFN PADOVA
Istituto Nazionale di Fisica Nucleare
Sezione di Padova

INFN LNL
Istituto Nazionale di Fisica Nucleare
Laboratori Nazionali di Legnaro

FEROCE

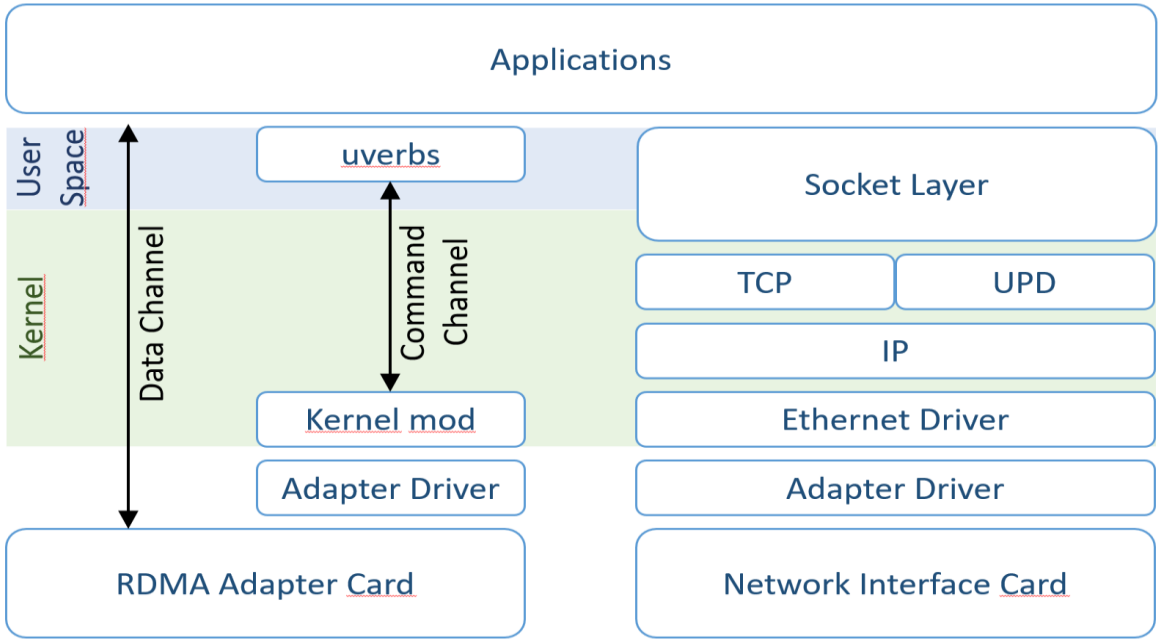
Front-End Rdma Over Converged Ethernet

Context

- Constant trend in producing larger and larger dataset in almost every field of experimental physics
  - Heightened in experiments without hardware trigger system
    - Less inefficiencies
    - Relaxed latency budget
  - LHCb is an example
    - Hardware pre-processing of the data stream followed by a software online selection similar to the one applied in offline analyses
  - CMS could follow
    - Calorimeter and muon system are candidates for streaming readout (bunch crossing rate)
    - Fast calibration or monitoring or even to allow the possibility of studying new physics channels
    - Modern data science techniques can be employed on accelerating hardware with a relaxed latency constraint
  - Beyond HEP
    - CTA advanced camera prototype
    - AGATA electronics upgrade to a FPGA based front-end

Objectives

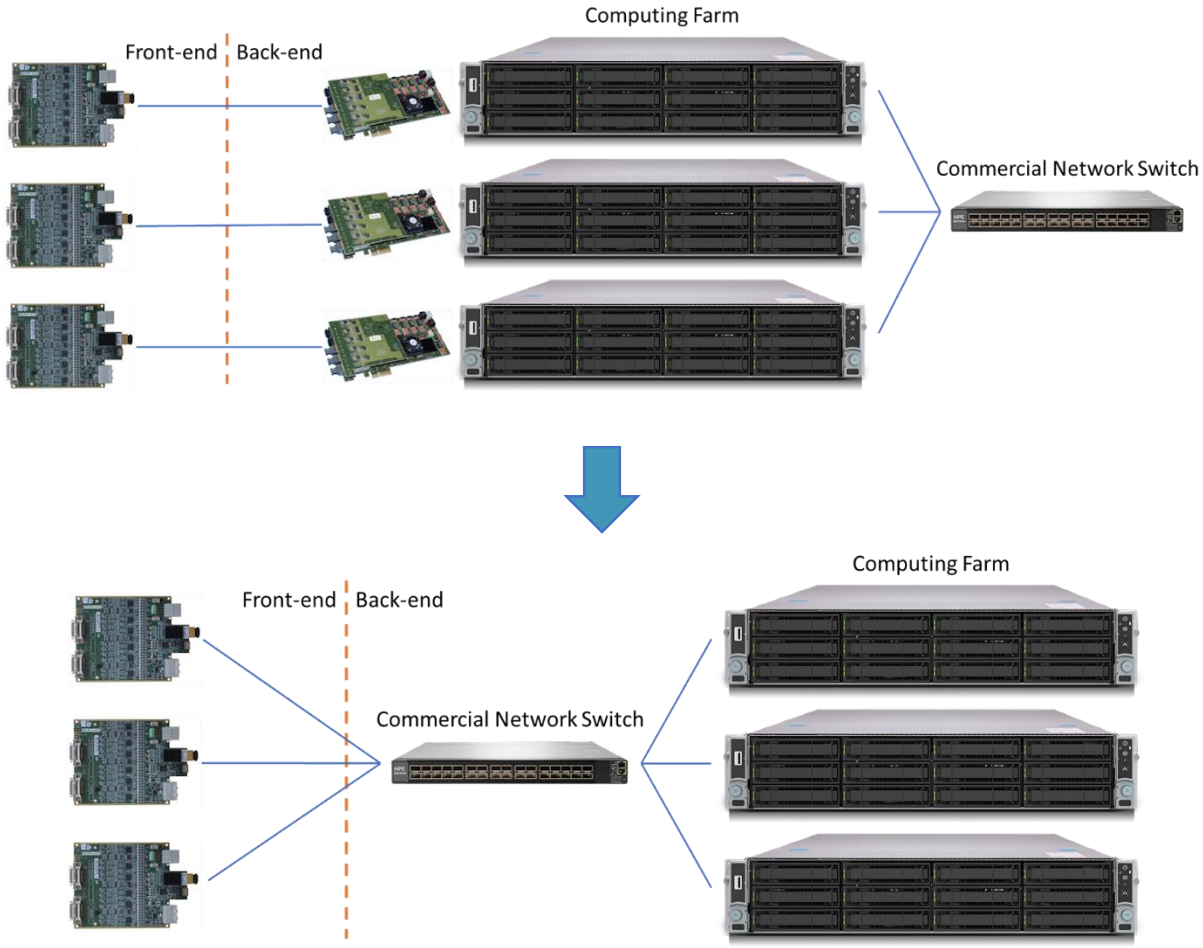
- Processing power is important as an efficient data movement
- In a DAQ system a large fraction of CPU is engaged in networking
  - Data manipulation (several copies)
  - Latency increase and throughput reduction
- Zero-copy is obtained by adding RDMA layer to the network stack



- FEROCE wants to move the adoption of the network protocol to the data producer
  - Front-end initiates the RDMA transfer
  - No point to point connection between front-end and back-end
  - Dynamical switching routing according to node availability

Methodology

- Several network stacks implementing RDMA
  - InfiniBand, RoCE, iWARP…
- RoCE (RDMA over Converged Ethernet)
  - Based on Ethernet networks
  - Industry-standard
  - Multi-vendor ecosystem
  - RoCE v2 packet switching (layer 2 and 3)
- FPGA are already used for implementing network stacks
  - Data center
  - ATLAS

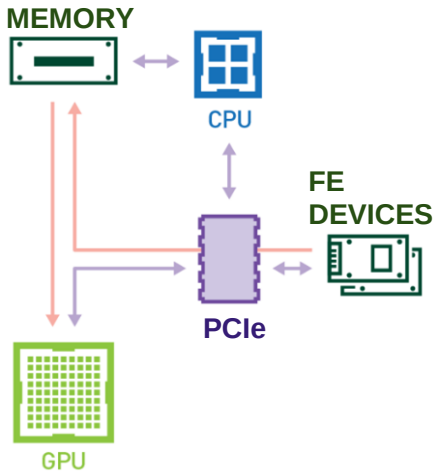


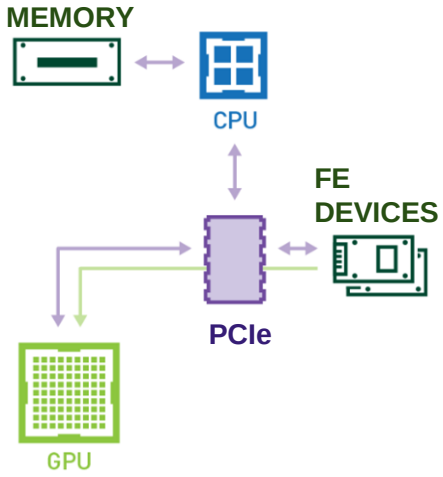
FW activities

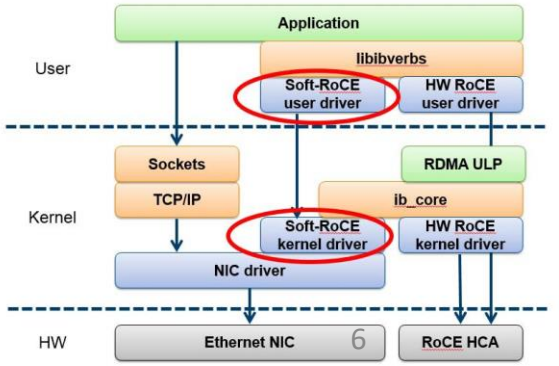
- Development of the firmware block needed by a RoCE endpoint
  - Small appliance of RoCE stack
    - Small scale FPGAs
    - No external memory or soft-core processor
  - Setup a global-oriented simulation environment
    - Language heterogeneity
    - Dynamic simulation
  - Test of the RoCE stack on evaluation platform
  - Capability of scale down the datapath
    - 100GbE/10GbE/1GbE
    - Many small distributed endpoints
    - Commercial switch port aggregation
  - Porting the front-end RoCE core to flash-based FPGA (radiation environment)
    - No technology dependent IPs or proprietary compilation tools

Networking activities

- Real-time processing of streaming data from FE devices with zero-copy access to memory
- Unburdening CPU by serving data directly to GPU memory
  - RDMA to GPU (GPUDirect)
  - Enabling CUDA-based applications on front-end data
- Expected activities
  - Emulating the RoCE transport in SW (via Soft-RoCE drivers) to enable testing of RDMA over Ethernet-enabled servers before the RoCE networking HW procurement
  - Deployment of networking hardware devices and qualification of the infrastructure in terms of throughput, latency and congestion avoidance
  - Test the commercial-available SW to move and control the data flow from RoCE to GPU







Project organization

- Two working package
  - WP1 Application layer and emulation
  - WP2 Front-end firmware core
- Four milestones
  - 31 Dec 2023 M1.1 Testing a RoCE network based on COTS products (WP1)
  - 31 Dec 2023 M1.2 Setting up of a global-oriented simulation environment for a small appliance of the RoCE stack (WP2)
  - 31 Dec 2024 M2 Complete test of the scalable RoCE firmware stack for front-end (WP1+WP2)
  - 31 Dec 2025 M3 Complete test of the developed firmware on a radiation tolerant front-end FPGA (WP1+WP2)

		2023			2024			2025		
		Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
WP1	**Application layer and emulation**	Soft RoCE	Setting up RoCE network	Test of RoCE network	Acquisition system		Test FE RoCE on network	GPUdirect	Test of GPUdirect	Test on flash FPGA
WP2	**Front-end firmware core**	Development light RoCE		Simulation light RoCE	Implementing different data paths		Test FE RoCE on network	Porting to flash technology		Test on flash FPGA

Research group

- Padova
  - Marco Bellato (0.3)
  - Antonio Bergnoli (0.2)
  - Daniele Mengoni (0.15)
  - Matteo Migliorini (0.3)
  - Fabio Montecassiano (0.2)
  - Jacopo Pazzini (0.15)
  - Andrea Triossi (0.3)
  - Marco Zanetti (0.15)
- LNL
  - Damiano Bortolato (0.15)

Total FTE: 1.9

Richiesta finanziaria

Anno	Missioni	Inventariabile	Consumo	Licenze e Specifiche tecniche	Totale
2023	**1**	**12**	**4**	**5**	**22**
2024	1	9			10
2025	1	10			11
Totali	3	31		5	43

La parte di apparati è dettagliata come segue:

2023	2x adapter NVIDIA ConnectX-6	4kE
	1x 100GbE switch 8 porte	12kE
	Specifiche RoCE	5kE
2024	1x scheda di valutazione FPGA Xilinx	8kE
	1x scheda di valutazione FPGA Microchip	1kE
2025	1x GPU NVIDIA	10kE