UNIVERSITÀ DEGLI STUDI DI PADOVA

J. Pazzini
PADOVA UNIVERSITY, INFN

PROGETTI WP4 - UNIVERSITÀ DI PADOVA

Kick-off meeting del WP4/Spoke2/CN1

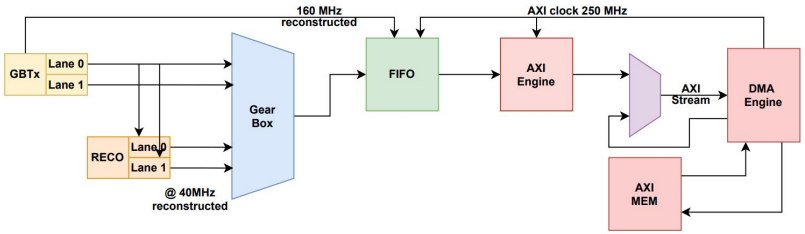
30 Ottobre 2022

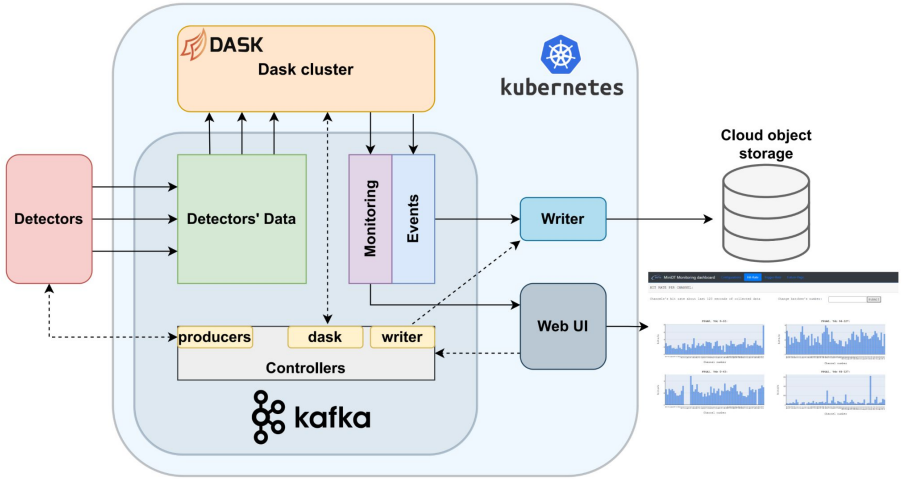
DISTRIBUTED STREAMING DATA PROCESSING

IDEA:
Abilitare il processing dello stream dei dati da detector minimizzando selezioni HW, processando on-the-fly con strumenti di calcolo distribuito i dati disponibili a livello di Back End dei rivelatori.

IMPLEMENTAZIONE:

- Implementazione di DMA in FPGA commerciali server-mount su interfaccia PCIe
- Processing distribuito con tool ormai comuni per fisici (Apache Spark / Dask), per permettere di definire workflow di analisi ad alto livello
- Data ingestion efficiente tramite tool di data brokerage distribuito come Apache Kafka
- Containerizzazione e Orchestrazione di tutte le risorse per facilitare il deployment in vari contesti (bare-metal, public or private cloud resources)
- Utilizzo di librerie disponibili (e.g. Rapids) per accelerazione su GPU per processing in cluster eterogenei





DISTRIBUTED STREAMING DATA PROCESSING

IDEA:
Abilitare il processing dello stream dei dati da detector minimizzando selezioni HW, processando on-the-fly con strumenti di calcolo distribuito i dati disponibili a livello di Back End dei rivelatori.

IMPLEMENTAZIONE:
- Implementazione di DMA in FPGA commerciali server-mount su interfaccia PCIe
- Processing distribuito con tool ormai comuni per fisici (Apache Spark / Dask), per permettere di definire workflow di analisi ad alto livello
- Data ingestion efficiente tramite tool di data brokerage distribuito come Apache Kafka
- Containerizzazione e Orchestrazione di tutte le risorse per facilitare il deployment in vari contesti (bare-metal, public or private cloud resources)
- Utilizzo di librerie disponibili (e.g. Rapids) per accelerazione su GPU per processing in cluster eterogenei

STATO:
Lavoro gia' avviato.
E' stato pubblicato un primo paper che descrive l'idea e primi test di benchmark con un tesbed dedicato (muon telescope, mock-up delle camere a deriva di CMS).
Primi test di integrazione a LHC in CMS negli scorsi mesi, con l'acquisizione e processing trigger-less di una porzione delle camere a deriva.
Altri usecase in via di definizione.
Lavoro da fare verso l'esplorazione e l'integrazione di librerie per l'utilizzo di GPU nel workflow.

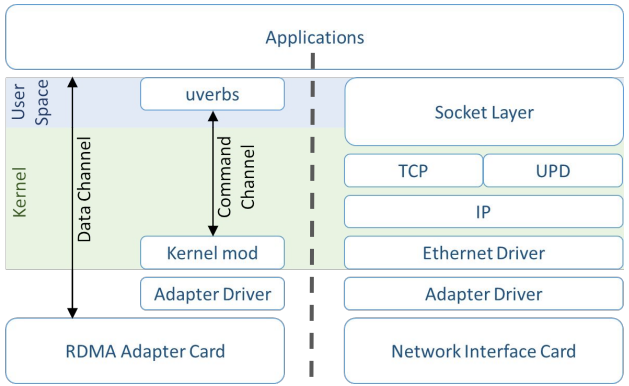
REMOTE DMA FROM DETECTORS' FRONTEND

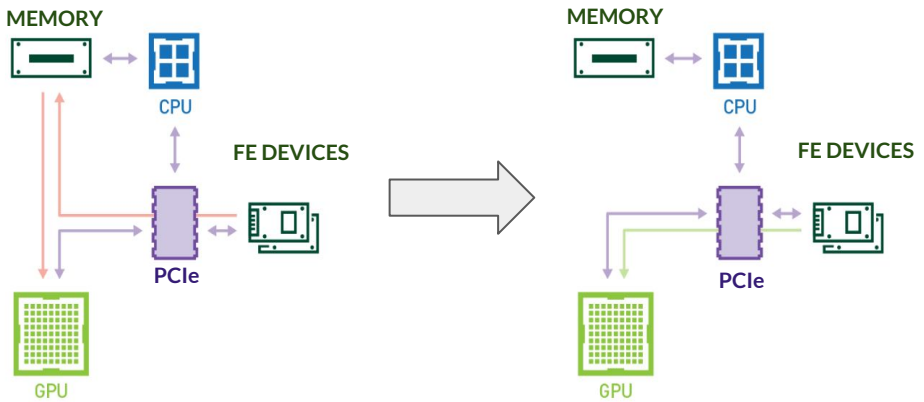
IDEA:

Portare i dati direttamente dal Front End dei detector verso nodi di calcolo attraverso Remote DMA Over Converged Ethernet (ROCE), rimuovendo la necessita' di connessioni dirette tra Front End e Back End dedicati, e minimizzando l'uso di CPUs per operazioni di data-copy.

IMPLEMENTAZIONE:

- Implementazione in FW di una versione "light" dello stack ROCEv2 con target le principali FPGA commerciali e rad-hard (Xilinx, Microchip).
- Distribuzione e redirezione dei pacchetti da Front End verso i nodi di calcolo tramite switch Connect-X
- Sviluppo del SW per la gestione e aggregazione dello stream di pacchetti ("event-merging")
- Test di driver commerciali disponibili per il data offload verso la memoria di GPU (e.g. GPUDirect) per accelerazione, e implementazione di librerie dedicate dove necessario





REMOTE DMA FROM DETECTORS' FRONTEND

IDEA:
Portare i dati direttamente dal Front End dei detector verso nodi di calcolo attraverso Remote DMA Over Converged Ethernet (ROCE), rimuovendo la necessita' di connessioni dirette tra Front End e Back End dedicati, e minimizzando l'uso di CPUs per operazioni di data-copy.

IMPLEMENTAZIONE:
- Implementazione in FW di una versione "light" dello stack ROCEv2 con target le principali FPGA commerciali e rad-hard (Xilinx, Microchip).
- Distribuzione e redirezione dei pacchetti da Front End verso i nodi di calcolo tramite switch Connect-X
- Sviluppo del SW per la gestione e aggregazione dello stream di pacchetti ("event-merging")
- Test di driver commerciali disponibili per il data offload verso la memoria di GPU (e.g. GPUDirect) per accelerazione, e implementazione di librerie dedicate dove necessario

STATO:
Attivita' in partenza a PD.
Attualmente gia' svolti dei primi test di implementazione di ROCE basati su uno stack HLS reso disponibile da ETH.
Nel prossimo futuro e' previsto lo sviluppo di un simulatore per permettere test delle implementazioni FW di "light-ROCE".
In attesa di tempi di procurement delle risorse di networking (NIC, switches, …) si intende procedere allo sviluppo SW attraverso l'emulazione dello stack RDMA in SW (Soft-RoCE).

PERSONPOWER

	Mesi/persona
PA	2
RTDb	3
RTDb	3
RTDa	1
*PhD**	*12*
Totale	9 (*+12*)

* E' previsto il reclutamento di 1 PhD student con fondi PNRR da destinare al 100% sulle attivita' WP4.

Personpower distribuito su entrambe le attivita', ad oggi principalmente attivo sul progetto di Distributed Streaming Data Processing.

E' pero' prevista la transizione degli FTE verso il progetto di Remote DMA con il ramp-up dell'attivita'.