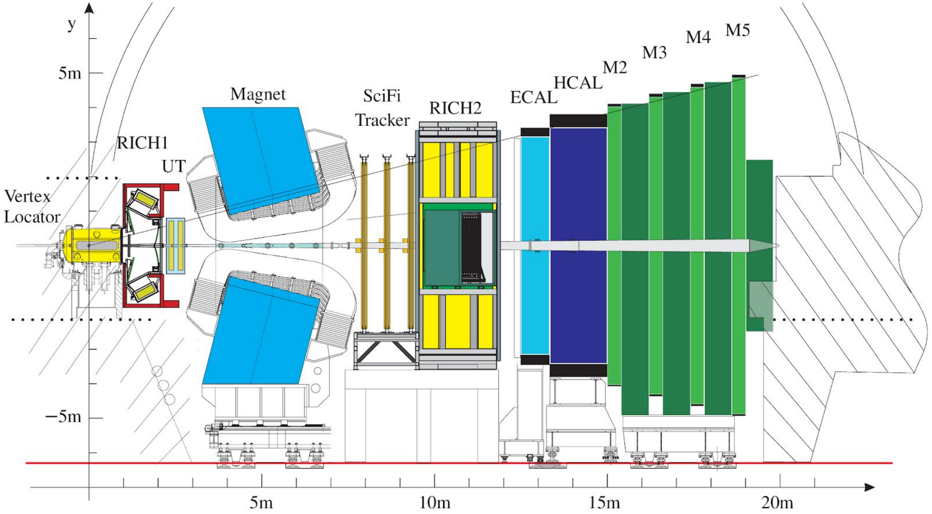
AI for Data Center Cooling and Resource Allocation in LHCb

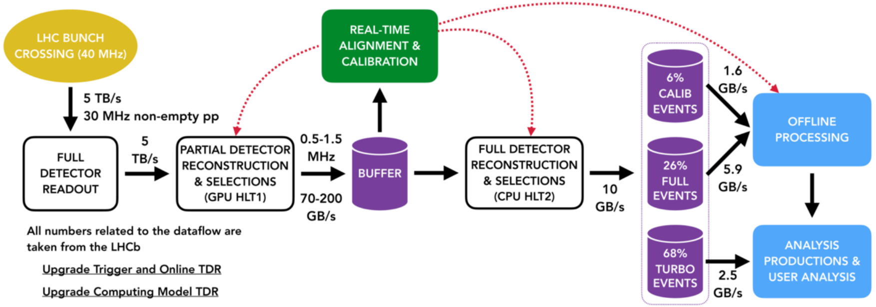
Pierfrancesco Cifra

Digital Twins for Nuclear and Particle physics
University of Messina

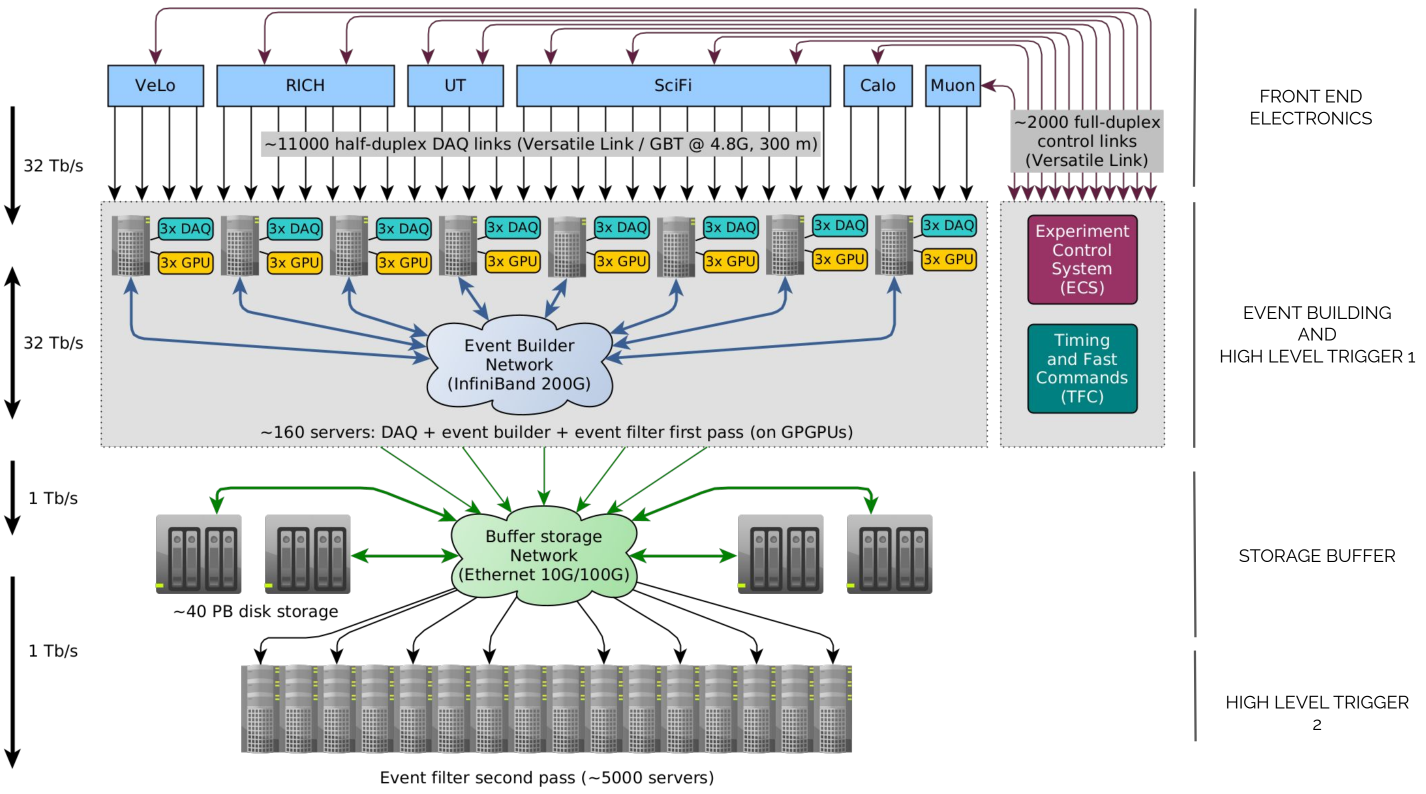
LHCb Experiment

- One of the 4 large HEP experiment at CERN
- Produces large amount of data per second
  - 32 Tb/s data redout from front-end electronics
  - Impossible to store permanently all the raw
  - Data is filtered and reduced to about 10 GB/s written to tape
- Requires large computing infrastructure to support:
  - Data Acquisition
  - Analysis
  - Simulation





LHCb DAQ System



LHCb Computing Infrastructure

- **Modular Data Centre** consisting of single row modules (shaped like a container)
    - 24 standard computing racks (48U) per module
- LHCb site composed by **6 containers**
    - 2 for Event Building + HLT1
    - 4 for HLT2, Alignement, Simulation
- In numbers:
    - ~5300 servers
    - ~270.000 CPUs
    - ~5000 HDDs (50 PB of raw storage) in JBODs
    - ~600 GPGPUs
    - ~600 Readout Boards (PCiE 40)
    - ~200 Network Switches
    - And more...

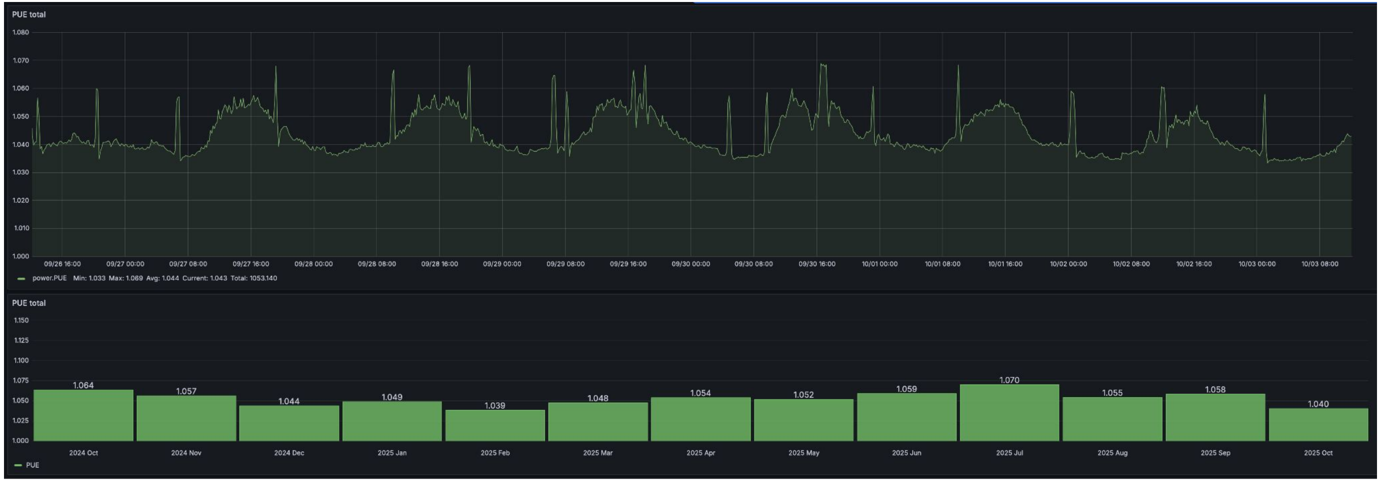
LHCb Computing Infrastructure

- Each module is design to operate at a **maximum power consumption** of **500 kW**
  - Total Power consumption of the site 2.3 MW
  - 20 kW / rack on average (up to 40 kW /rack)
  - Produces a lot of heat that needs to be moved out of the Data Centre
- **Free cooling**: 3+1 Air Handler Units (AHU) placed on top of each container
  - AHU: it use external air and adiabatic water cooling to lower DC intake temperature
  - Geneva temperature allow this solution to be effective
  - Each module is fully redundant: normal operation can be maintained with any 3 of the 4 AHUs
  - It adds some overhead in the total power consumption
- Fan speed and water-pump frequency are regulated by a **standard PID controller**
  - Proportional Integral Derivative controller largely used in the industry, easy to implement and tune
  - Requires retuning when operating conditions change and lacks predictive/optimal control

LHCb Computing Infrastructure

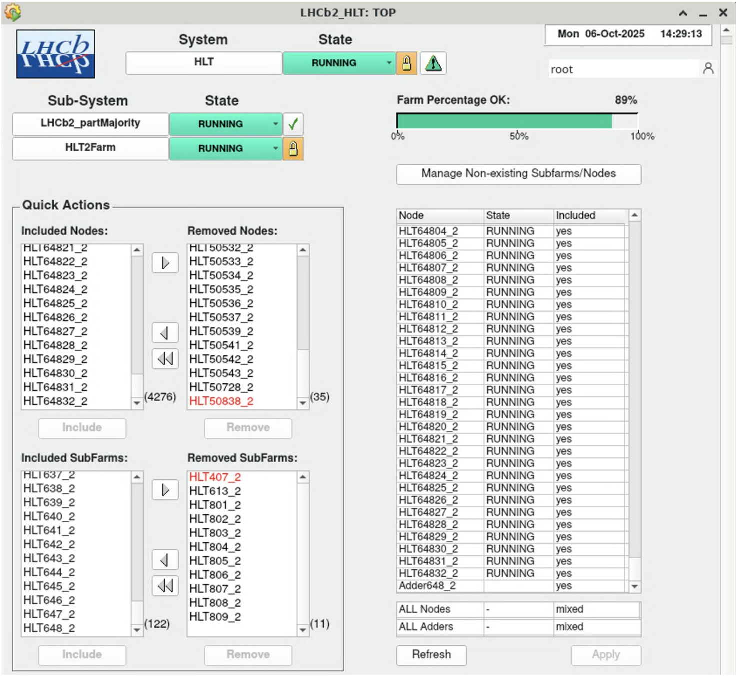
- Key metric for energy efficiency: **Power Usage Effectiveness**
  - PUE = Total Facility Energy / IT Equipment Energy

- LHCb Data Centre is **overall already very efficient**
  - Average yearly PUE below 1.1 (less than 10% of total energy used for cooling)
  - Reports place the industry-wide average around 1.5



LHCb Computing Infrastructure

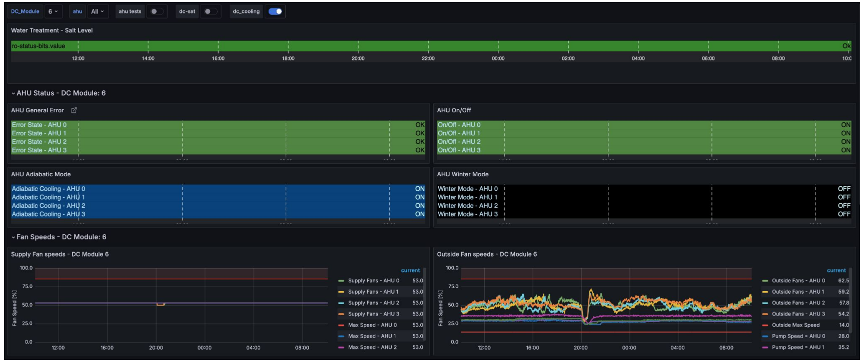
- Computing resources are currently **allocated by the experts** according to the needs
  - Using WinCC OA or HTCondor as interface to run jobs
  - HLT2, Alignment and Calibration jobs driven by the LHC efficiency and Disk Buffer status
  - Opportunistic Monte Carlo simulations
  - Use of CPUs and GPUs for other purposes

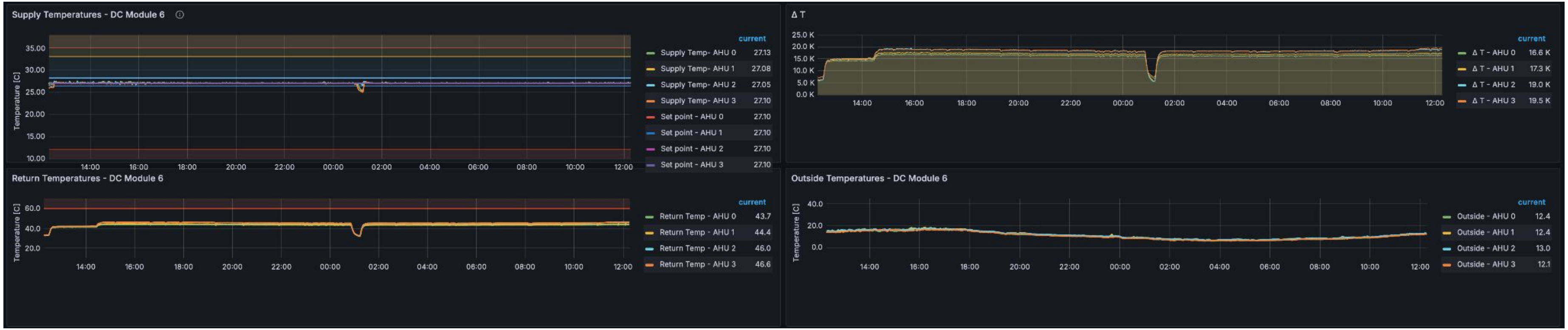


- Manual approach is **not optimal** as it relies on expert coordination and manual decisions
  - Lack of dynamic adaptation to workload changes or real-time priorities
  - Resources are often idling when operations are stopped (waste of energy)

Monitoring Data

- Custom **monitoring system**:
  - Sensors for temperature, humidity, AHU status, power consumption, water usage
  - 5 years of history, 30 seconds granularity, ~300 GB of raw data
  - Overall suitable and easy adaptable for training AI and ML algorithms



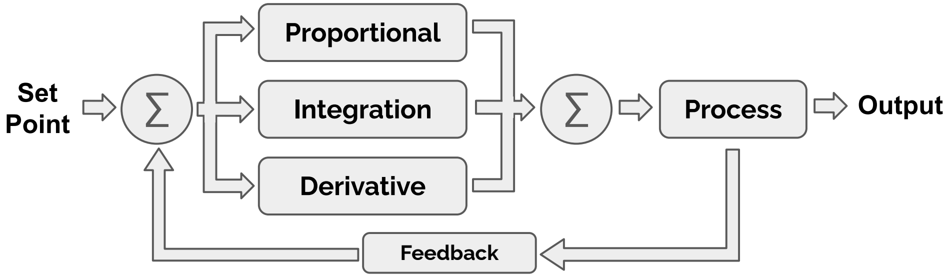


AI for Cooling

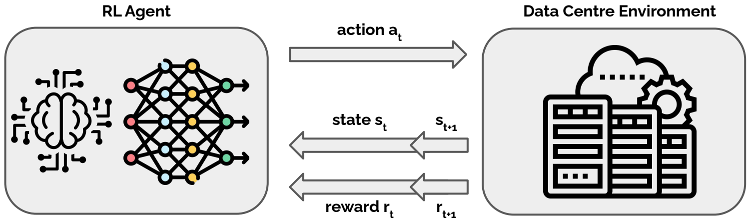
- **Improve the efficiency** of the whole facility
  - Lowering the PUE
  - Lowering water consumption
  - Improve temperature stability

- Can AI techniques **outperform** standard PID regulations?
  - In safety and efficiency
  - AI-driven control can model complex thermal dynamics and interactions

Schematic of a standard PID control system



RL control approach



AI for Cooling

- **Reinforcement Learning Agent**:
  - Given a state, it takes an action to regulate the data center cooling system
  - Using RL where actions are guided by reward (lower PUE, critical parameters must staying within safety limits)

- **No Direct Interaction** with the Environment
  - The agent cannot interact directly with the real data center for training
  - Safety constraint: temperatures and operating conditions must remain within safe limits

- **Offline** Reinforcement Learning Approach
  - Learn policies directly from the historical operational data
  - No exploration in order to ensures safety and stability

AI for Cooling

- Decision-making problem formulated by a **Markov Decision Process**
  - $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma)$
  - State space: system conditions (temperatures, power, etc.), each expressed as a temporal vector
  - Action space: Controllable variables for the AHUs (fan speed and water pump frequency)
  - Transition function: Model system dynamics
  - Reward function: Reward function to balance energy saving and temperature regulation
  - Discount factor: Future rewards

- Learn **optimized policy**:
  - Based on offline Dataset
  - That maximize the discounted cumulative return

$$R(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, r(s_t, a_t)\right]$$

AI for Cooling

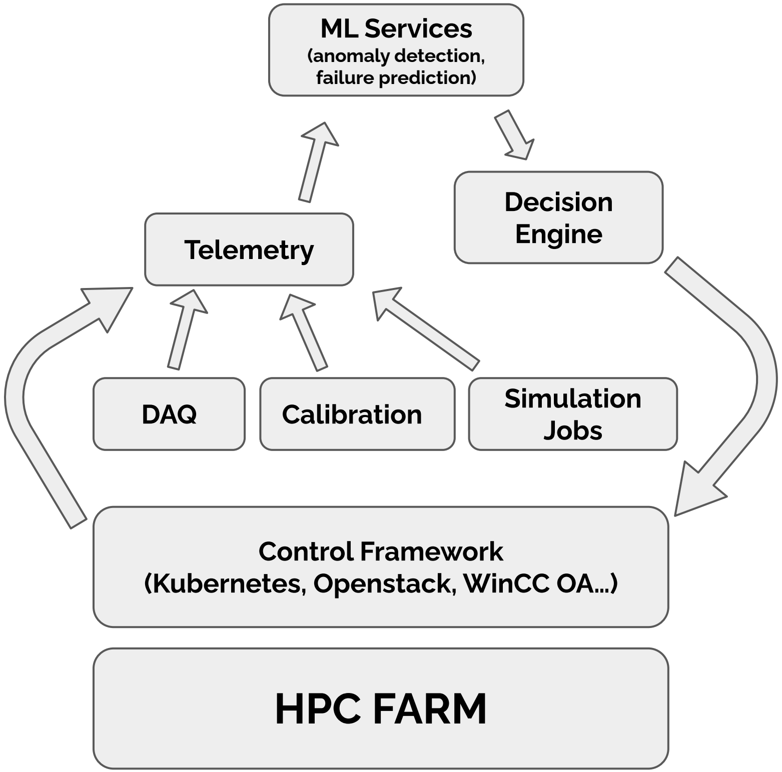
- **Data Centre Digital Twin** for Simulation and Evaluation:
  - Virtual replica of the Data Centre
  - Enable simulation, testing, and further optimization of cooling

- Developing accurate Digital Twins is **challenging**:
  - Complex physics: airflow, heat transfer, fluid dynamics (CFD)
  - Dynamic workloads: variable compute demand and scheduling
  - External influences

- There are already some initiatives and frameworks for developing **Data Centre Digital Twins**
  - Nvidia omniverse
  - Intertwin (EU project)

AI for Smart Allocation

- AI-driven system that **dynamically manages computing resources**
  - Workload forecasting
  - Real-time utilization
  - Priority handling (DAQ > Simulation)
  - Anomaly detection and failure prediction



AI for Smart Allocation

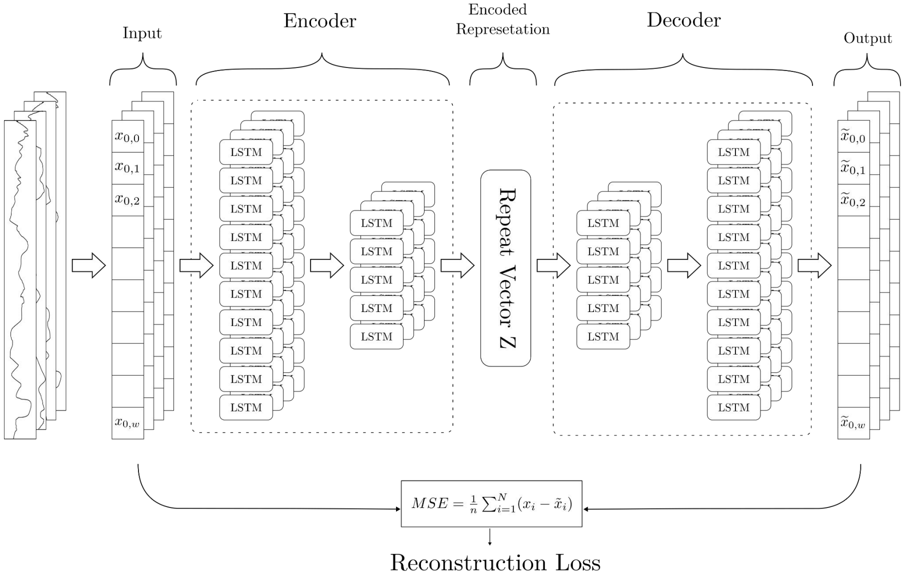
- Goal: **maximize efficiency** and **reduce total power consumption** without affecting performance
  - Predict idle periods and shut down unused resources to save energy
  - Quickly react to changing conditions (DAQ is subjected to rapid changes according to LHC coordination)
  - Automatically exclude components that failed or that are about to fail

- Using telemetry data from **different sources**:
  - LHC efficiency and schedule
  - Disk buffer utilization
  - Network links usage
  - And more…

Anomaly Detection

- **Time Series Anomaly Detection** techniques
  - Time Series Metrics from computing nodes and AHUs units
  - Unsupervised problem

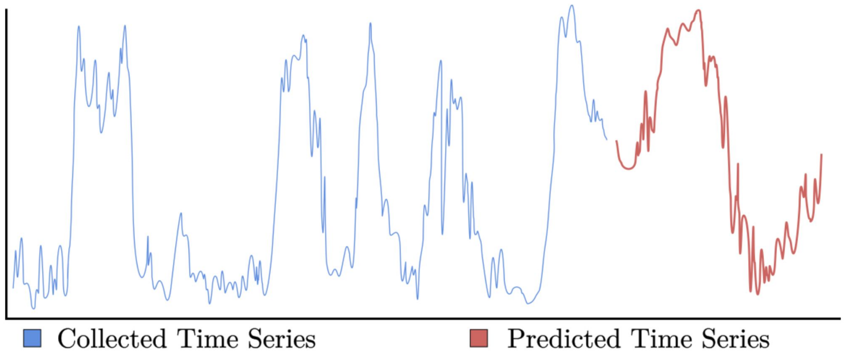
- **Autoencoder based** Neural Network
  - Using Multivariate Time Windows as input
  - Learn the standard behaviour of a computing node
  - Higher the reconstruction loss, higher the probability of having an anomaly
  - ~96% accuracy on a benchmark dataset



Failure Prediction

- **Failure prediction**:
  - Time Series forecasting
  - Similar approach to the anomaly detection architecture
  - The Neural Network learn to reconstruct the next Window

- **Implement Predictive maintenance** strategies:
  - We run old hardware
  - At this scale we have device failing on a daily basis
  - Reduce downtime
  - Improve reliability of the whole infrastructure



Conclusions

- AI can support all operational aspects essential to running large-scale experiments
- Intelligent control can significantly reduce power usage and improve the environmental sustainability of data centers.
- Dynamic, data-driven allocation enhances operational efficiency and minimizes idle resources
- Anomaly Detection and Predictive models can improve system reliability and reduce downtime

Thank you for your attention

Digital Twins for Nuclear and Particle physics
University of Messina

ODISSEE