中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

A Distributed Memory Cache Pool-Centered Online Computing Framework for HEP

——*Upgrade of LHAASO Online Computing Framework*

Hangchang ZHANG

On behalf of LHAASO TDAQ Team

Nov. 2025

Guangzhou

中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

LHAASO

Outline

- Background
- Design of the framework
- Development Progress
- Summary

Large High Altitude Air Shower Observatory

Water Cherenkov Detector Array (WCDA)

With a total area of 78,000 square meters, it is designed for all-sky surveys of gamma-ray sources. It comprises three water ponds containing 3,120 detector units.

1 km^2 array (KM2A)

Consists of 5,195 electromagnetic detectors and 1,171 muon detectors, covering an area of one square kilometer

Wide Field-of-View Cherenkov Telescope Array (WFCTA)

Composed of 18 wide-field-of-view Cherenkov imaging telescopes

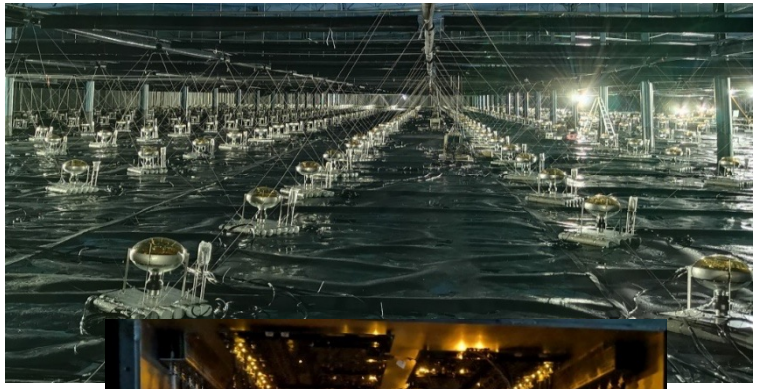
LACT *Under Construction*

32 imaging Cherenkov telescopes with a diameter of 6 meters each

LHAASO explores the origins of high-energy cosmic rays and conducts fundamental research on related high-energy radiation, celestial evolution, and dark matter distribution.

Detectors and Elecs

WCDA

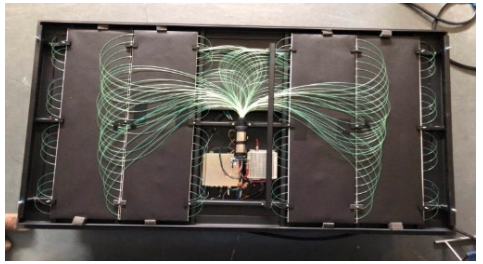


WCDA Elecs

WFCTA

SiPN Cameras

KM2A ED & MD



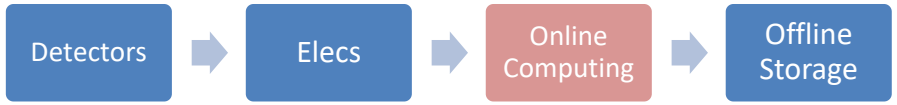
ED Detector Units

~3.0 GB/s Data

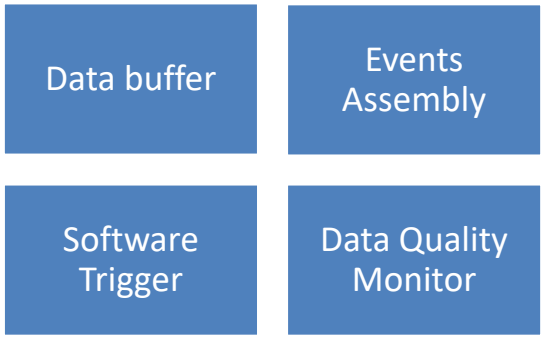
LHAASO Online Computing System

The online computing system of LHAASO handles data acquisition, software triggering, and real-time data processing

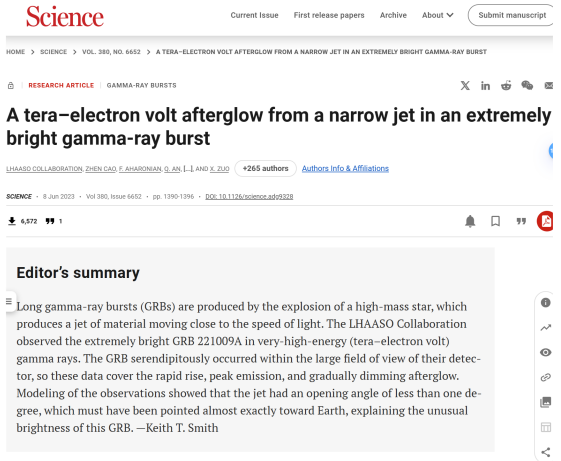
- Physical algorithms directly access raw data without hardware triggering.
- High real-time capability is vital for early warnings of special natural phenomena (such as GRBs and thunderstorms) and for monitoring experimental operational status.



The role of online computing system in High-Energy Physics Experiments



Primary functions of online computing system


Science

A tera–electron volt afterglow from a narrow jet in an extremely bright gamma-ray burst

Editor's summary

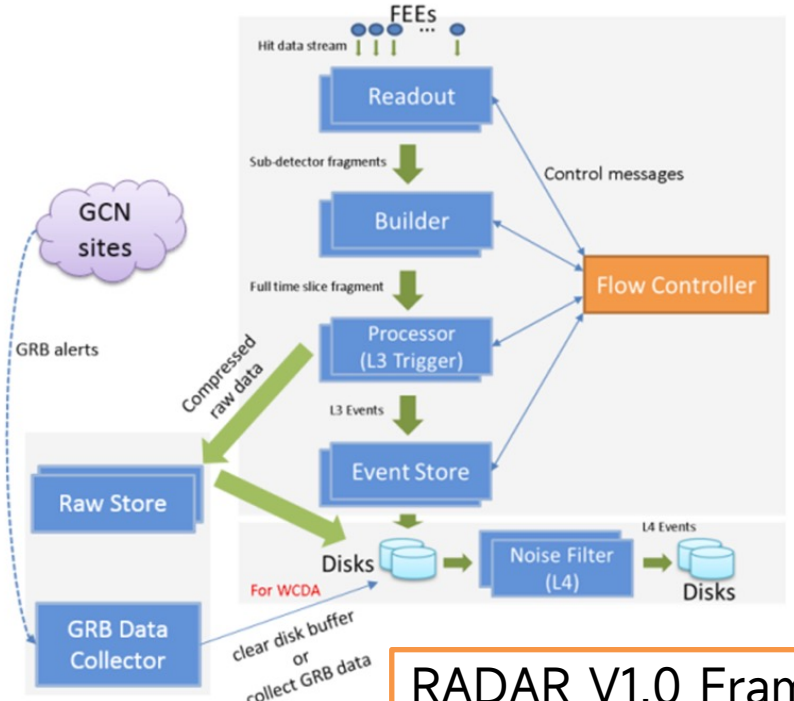
Long gamma-ray bursts (GRBs) are produced by the explosion of a high-mass star, which produces a jet of material moving close to the speed of light. The LHAASO Collaboration observed the extremely bright GRB 221009A in very-high-energy (tera–electron volt) gamma rays. The GRB serendipitously occurred within the large field of view of their detector, so these data cover the rapid rise, peak emission, and gradually dimming afterglow. Modeling of the observations showed that the jet had an opening angle of less than one degree, which must have been pointed almost exactly toward Earth, explaining the unusual brightness of this GRB. —Keith T. Smith

LHAASO detects GRB 221009A, the most energetic in human history

LHAASO Online Computing Framework

Current Status

Readout and Computing Nodes x172



RADAR V1.0 Framework

Achieved

- Massive TCP/IP data readout
- Clustered parallel stream processing
- Software triggering
- Automated operation

Challenges

- Difficulties in integrating offline algorithms
- Inability to achieve *"debugging while running"*
- Single points of failure and performance bottlenecks pose risks to operations.

After LHAASO started running, growing physical demands are difficult to meet.

Motivation for Upgrading LHAASO Online Computing System

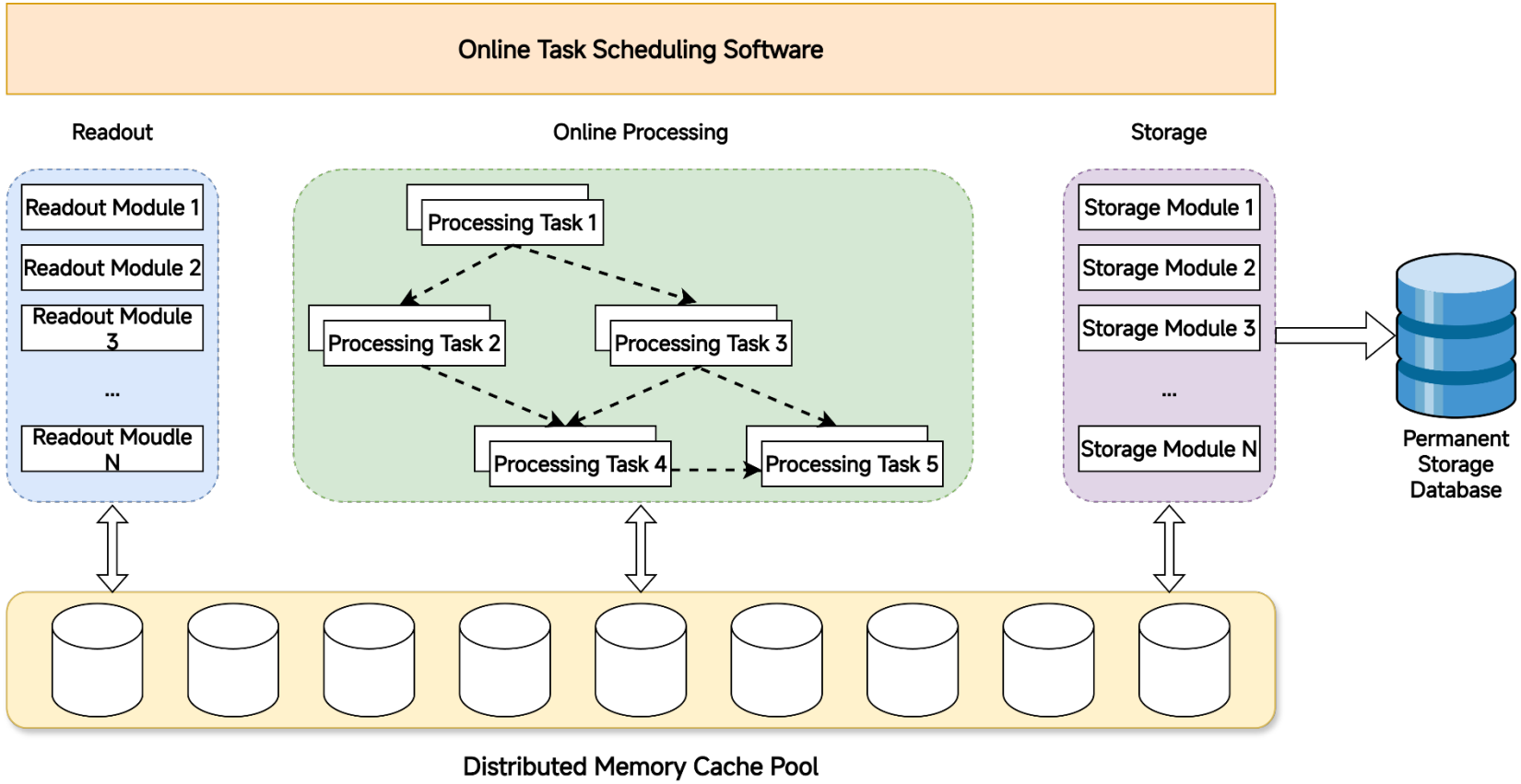
Flexible

- Offline users lead online algorithm development
- Provides file-based online data processing software development solutions for users

Reliable

- Ensures real-time data processing
- Guarantees data reliability
- Decoupled dataflow modules

Design of the new online computing framework



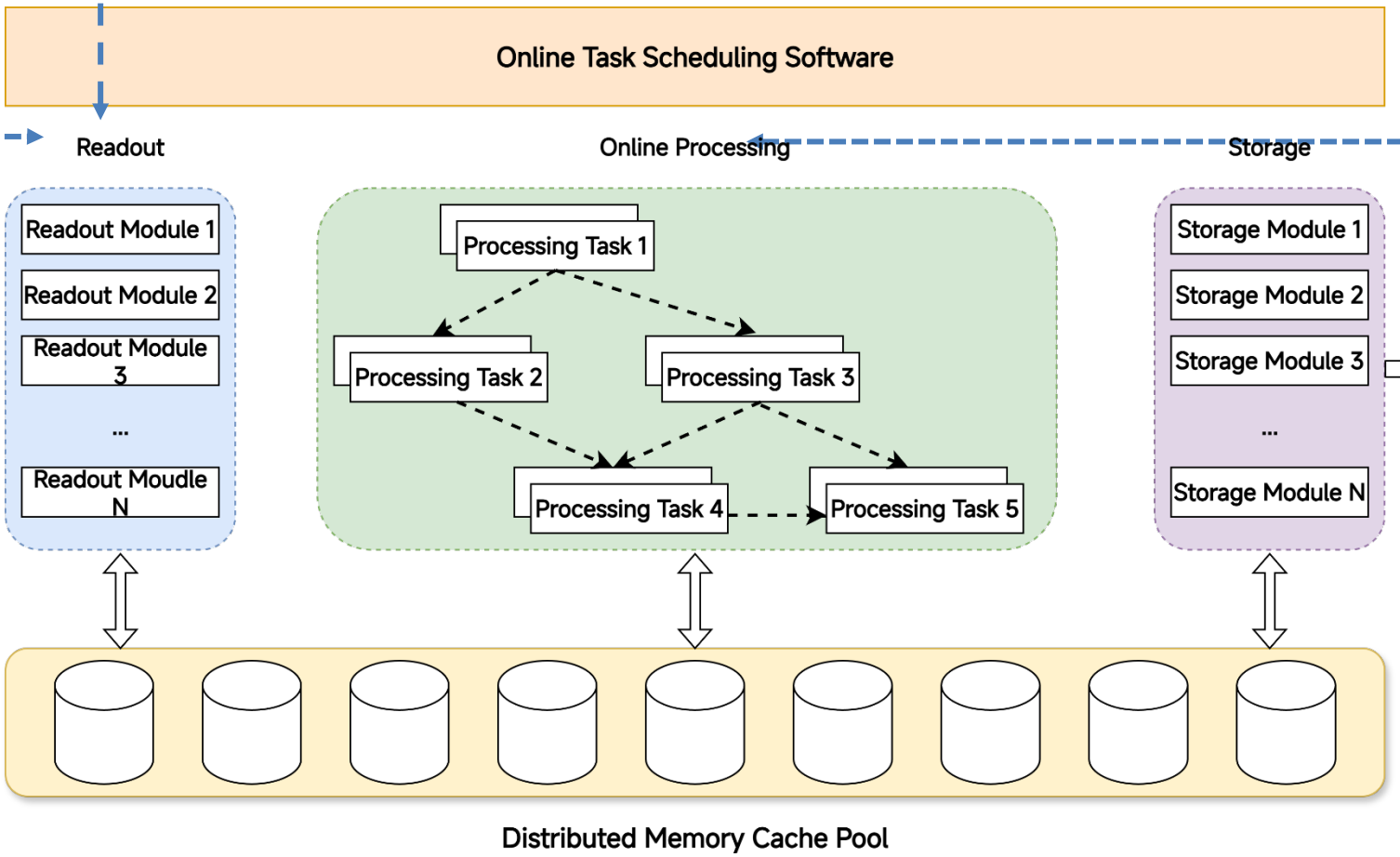
Module Decoupling

- Between experimental systems
- Between data flow modules
- Between algorithms

Data Fusion

- Perceptual data
- Diagnostic data
- Scientific data

Design of the new online computing framework



Decoupling of Readout & Processing

Offline algorithms can access raw data and run online.

Physically, computing and cache belong to the same computing cluster.

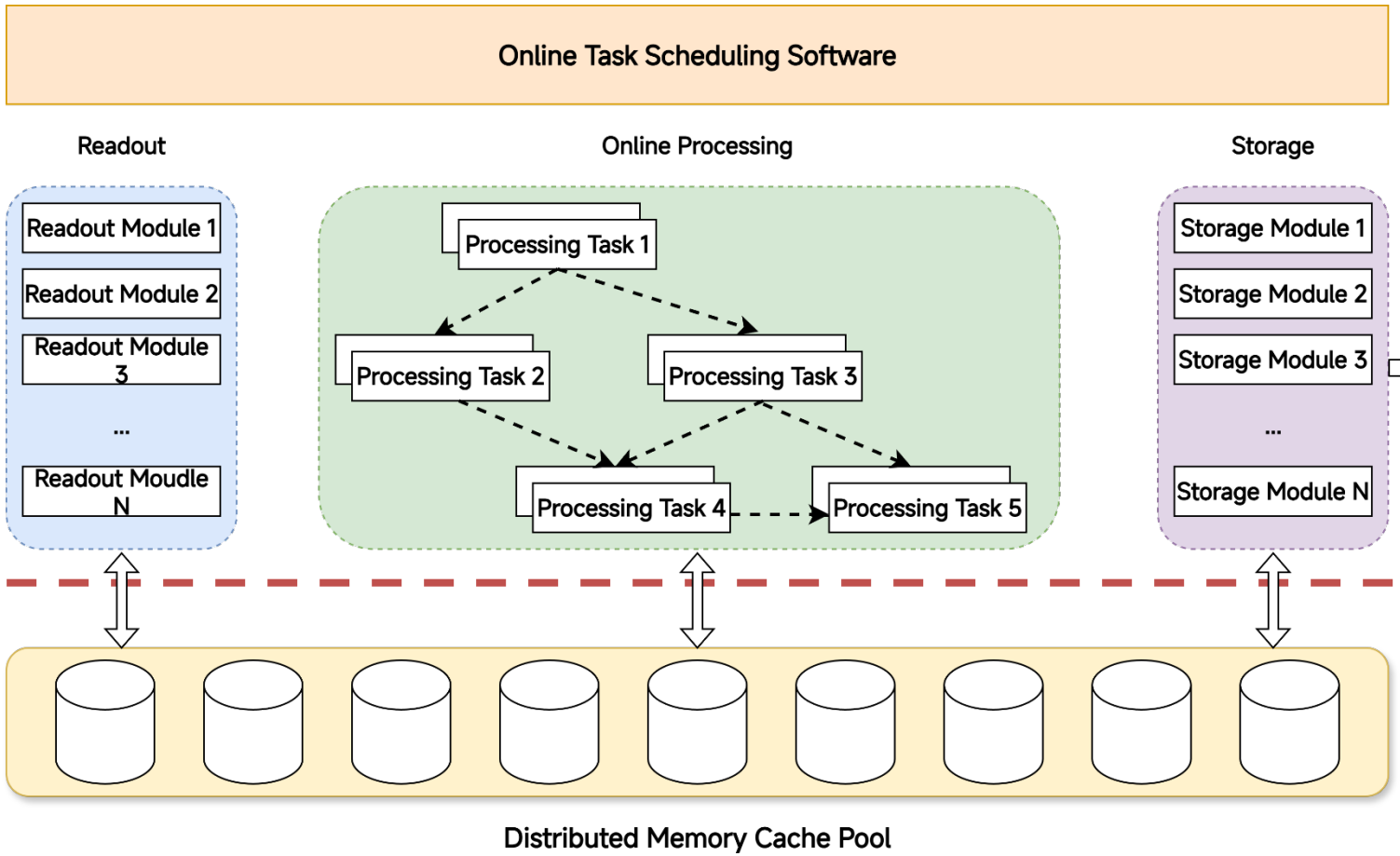
Distributed Memory Cache Pool exchange physical data and computational result

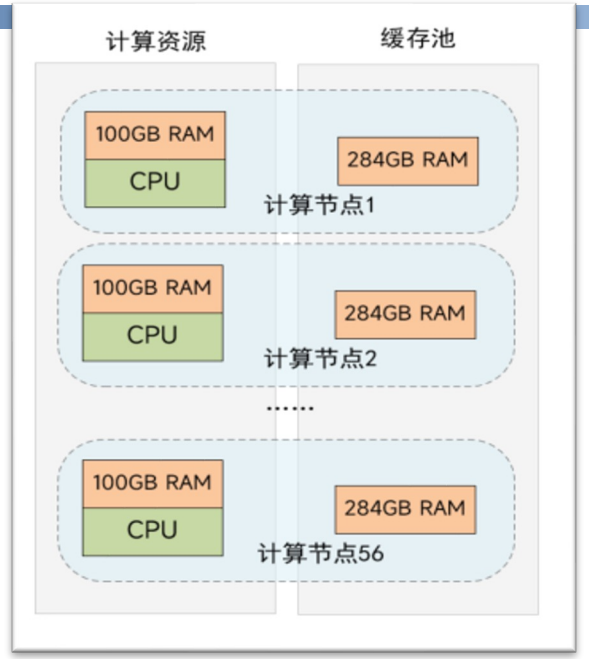
Readout modules offer strong extensibility, enabling dynamic addition or removal.

During the construction and commissioning of LACT, normal operation of LHAASO must not be affected.

Design of the new framework

Distributed Memory Cache Pool





Cache pools and online tasks share computing resources.

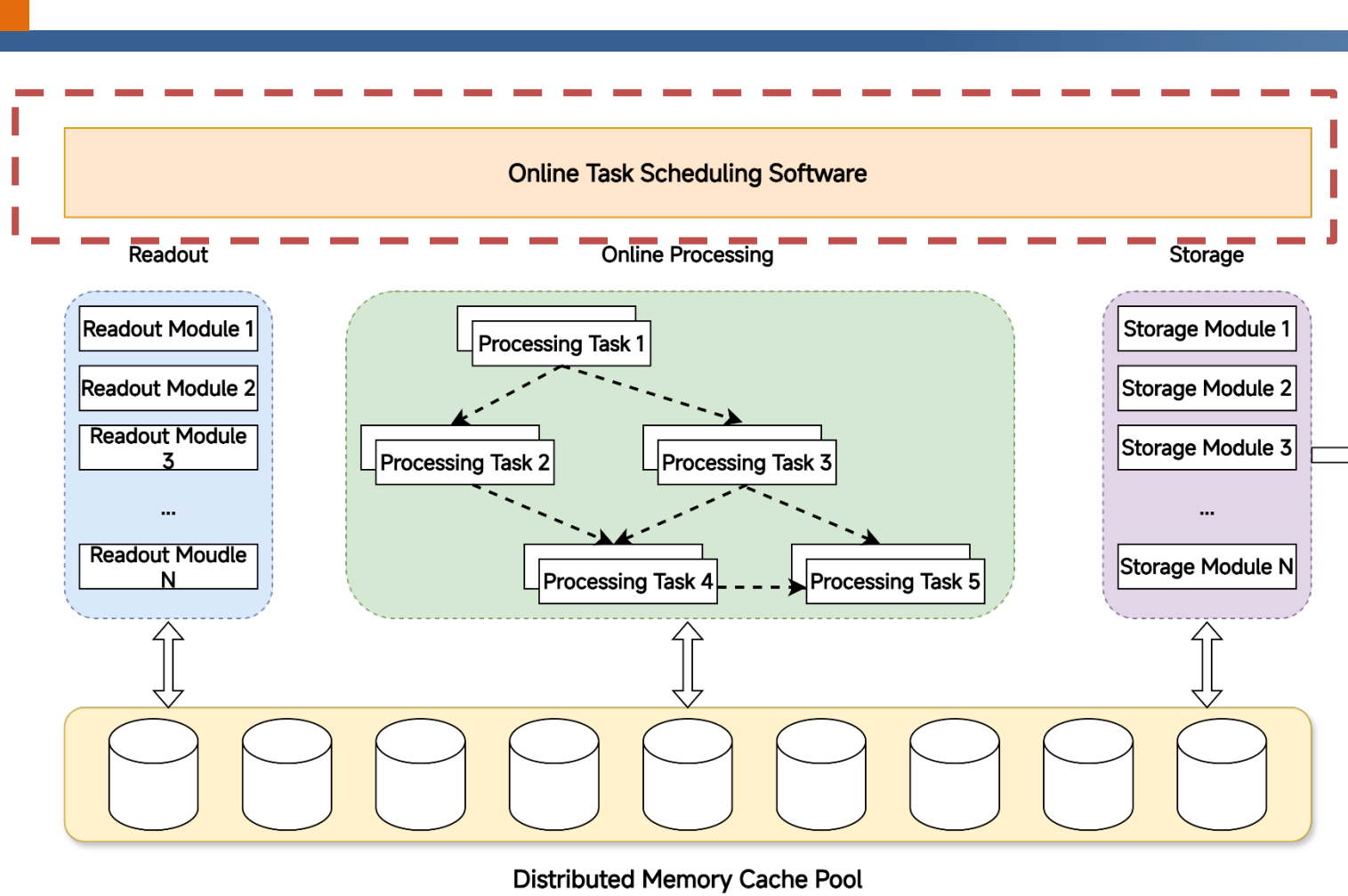
The throughput of DMCP determines the throughput of this system
The reliability of DMCP determines the reliability of this system

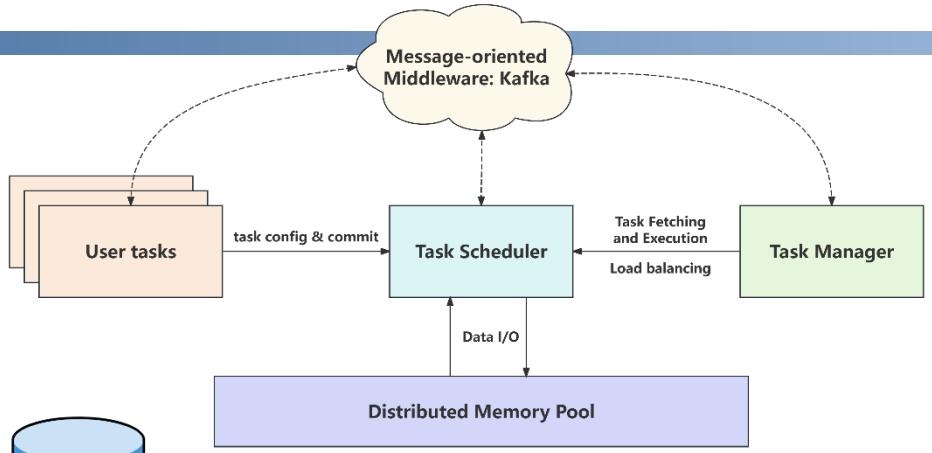
Reasons for using memory as a caching medium:

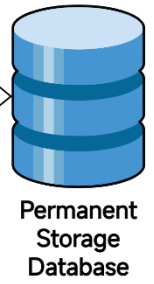
- Caching capacity requirements < Throughput requirements
- Integrating computation with caching to optimize resource utilization

Design of the new framework

Online Task Scheduling Software





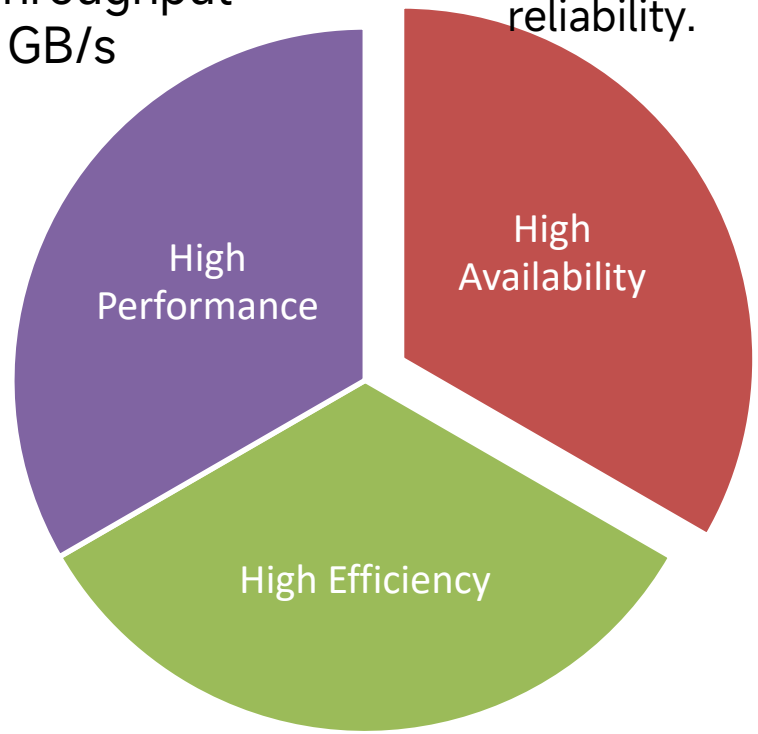


- Deployment of computational tasks
- Loading balancing for online clusters
- Task anomaly detection and handling
- Task status information logging
- User-friendly algorithm deployment methods for scientists

Development Progress: In-house Memory Cache Pool Software

Total Throughput
>> 3.5 GB/s

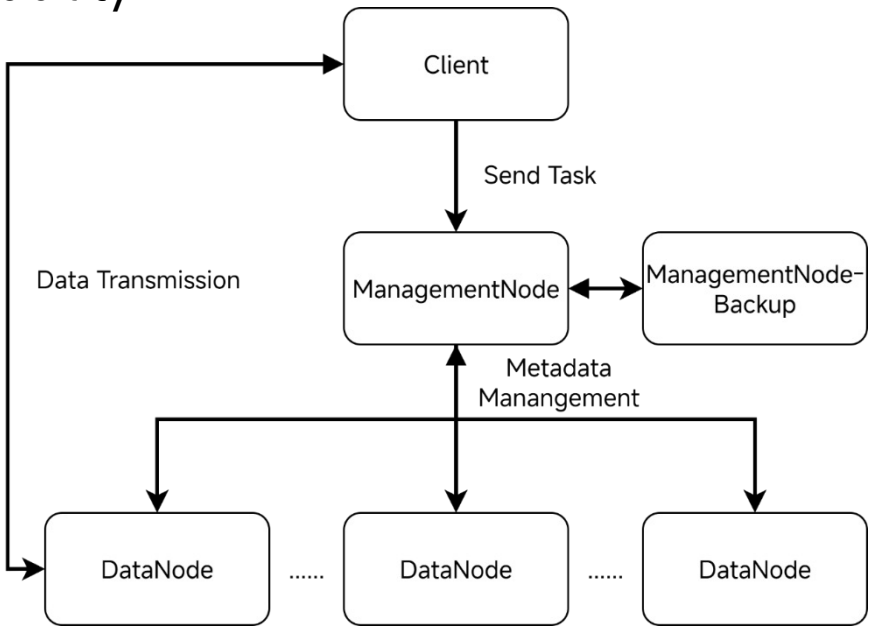
A single point of failure does not affect the service, ensuring data reliability.



Utilizes minimal CPU resources

After researching and testing various open-source tools, we found that none fully met our requirements, leading us to develop our own solution.

- The software caches files to facilitate offline users in migrating algorithms to an online environment.
- Multiple redundancies are implemented to ensure reliability.



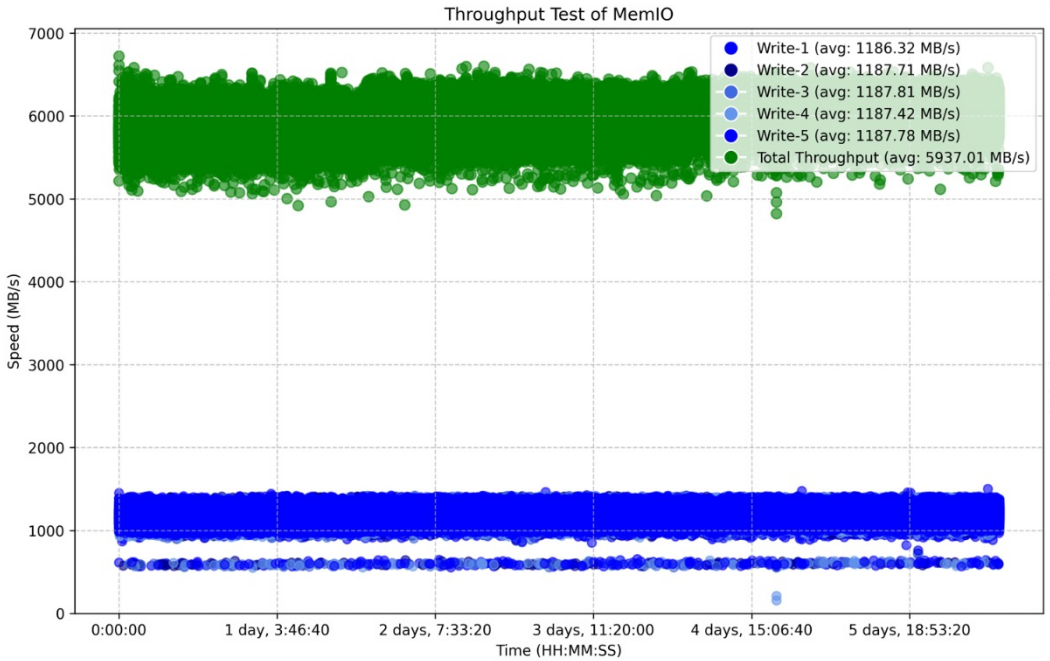
Development Progress: Applied in LHAASO

Cluster Validation at IHEP

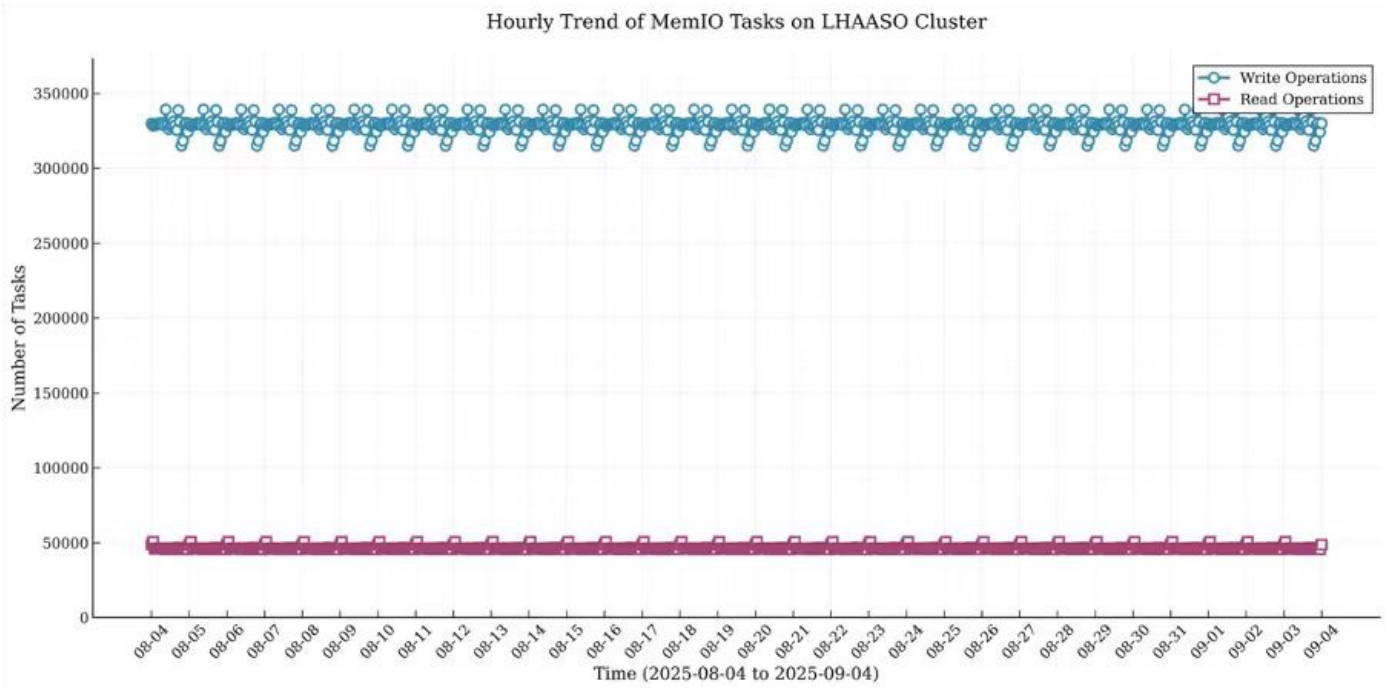
- 1 management node, 5 data nodes
- 100 GbE network
- Memory Pool capacity: 1000 GB

Deployment in LHAASO

- 1 management node, 6 data nodes
- Memory pool capacity: 1200 GB
- KM2A & WFCTA access



6 GB/s stable operation for over 6 days
200% LHAASO



Stable operation > 30 days

Development Progress: Computing driven by data

DEMO

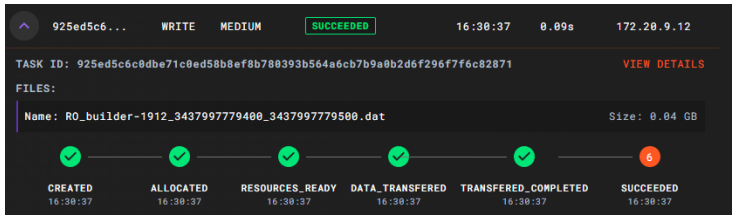
KM2A Spectrum Analysis

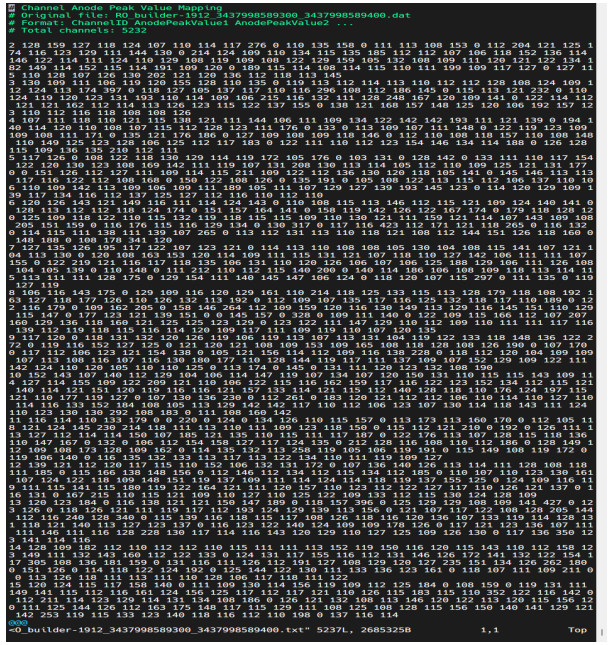
Read raw data from the cache pool into local memory

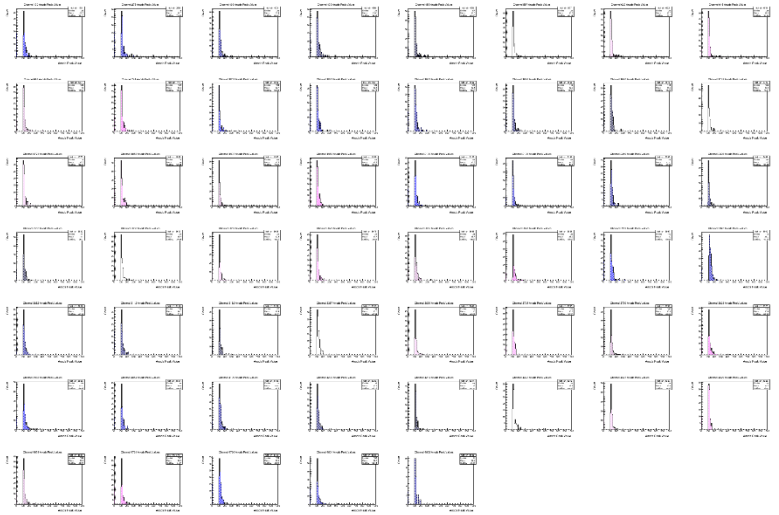
Parse the data to obtain {Channel: Peak Values}

Perform data analysis

User-friendly read/write interface
memio::read(file_name)
memio::write(file_name,local_filename)







Verified new framework
- Data-driven
- Decoupled task execution

Outlook: AI, LHAASO, CEPC

Goal: Enough minutes of cache time, NIC-level single-point throughput

Module Decoupling

- Between experimental systems
- Between data flow modules
- Between algorithms

Data Fusion

- Perceptual data
- Diagnostic data
- Scientific data

Once we've increased the throughput limit

HYLIA: Hyper Link Accelerator

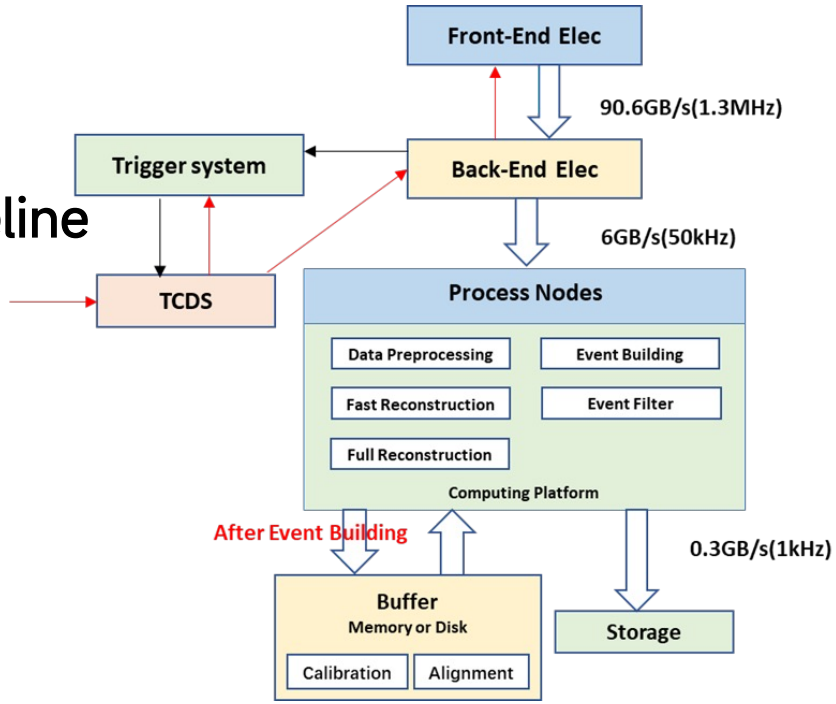
- C++ wrappers for high-performance network transport application layers
- Currently implemented wrappers for RoCE v2, TCP
- Goal: Unified interface, simple and easy to use
- Independent from distributed memory caching software

	Hylia / TCP	Hylia / RoCE v2
Speed	2.2 GB/s	5.2 GB/s
CPU Usage	100%	80%
Latency	512us	167 us

LHAASO: From Automation to Intelligence

- Provide I/O acceleration for heterogeneous computing and AI algorithms to enhance real-time performance
- Explore taking on experiment control and fault recovery with large language models.

CEPC Baseline



Summary

Motivation

- Developing a new generation of **flexible and reliable** online computing software framework

The distributed memory cache-based online computing framework will achieve:

- Decoupling of readout modules from online processing algorithm modules
- Flexible access methods for online processing techniques provided to physics users
- Ready for AI and heterogeneous computing

Next Plan:

- By 2026, LHAASO will fully transition to the new architecture
- Realize TB/s total throughput through advanced networking protocols

backup

HYLIA: Hyper Link Accelerator

- C++ wrappers for high-performance network transport application layers
- Currently implemented wrappers for RoCE v2, TCP Epoll, and ZeroMQ
- Goal: Unified interface, simple and easy to use
- Independent from distributed memory caching software

```
HyliaComm tcpClient;
tcpClient.init("tcp", "0.0.0.0", 12345, "client")
tcpClient.send(data, size);
```

```
HyliaComm rdmaServer;
rdmaServer.init("rdma", "0.0.0.0", 12345, "server")
rdmaServer.listen();
rdmaServer.poll();
```

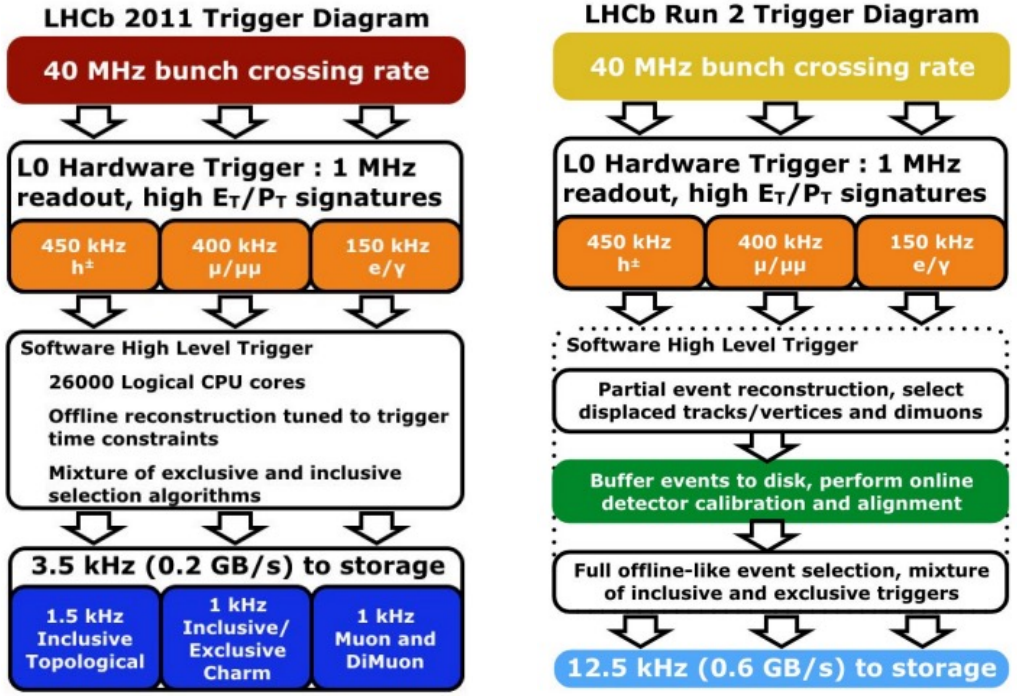
	1GB File		128KB Data Packet			
	Hylia / TCP	Hylia / RDMA	Hylia / TCP	Hylia / RDMA	QPERF / TCP	QPERF / RDMA
Speed	2.2 GB/s	5.2 GB/s	2.2 GB/s	5.3 GB/s	2.5 GB/s	9 GB/s
CPU Usage	100%	80%	100%	60%	100%	-
Latency	512us	167 us	512us	167 us	320us	160 us

Test Env.: 100GbE

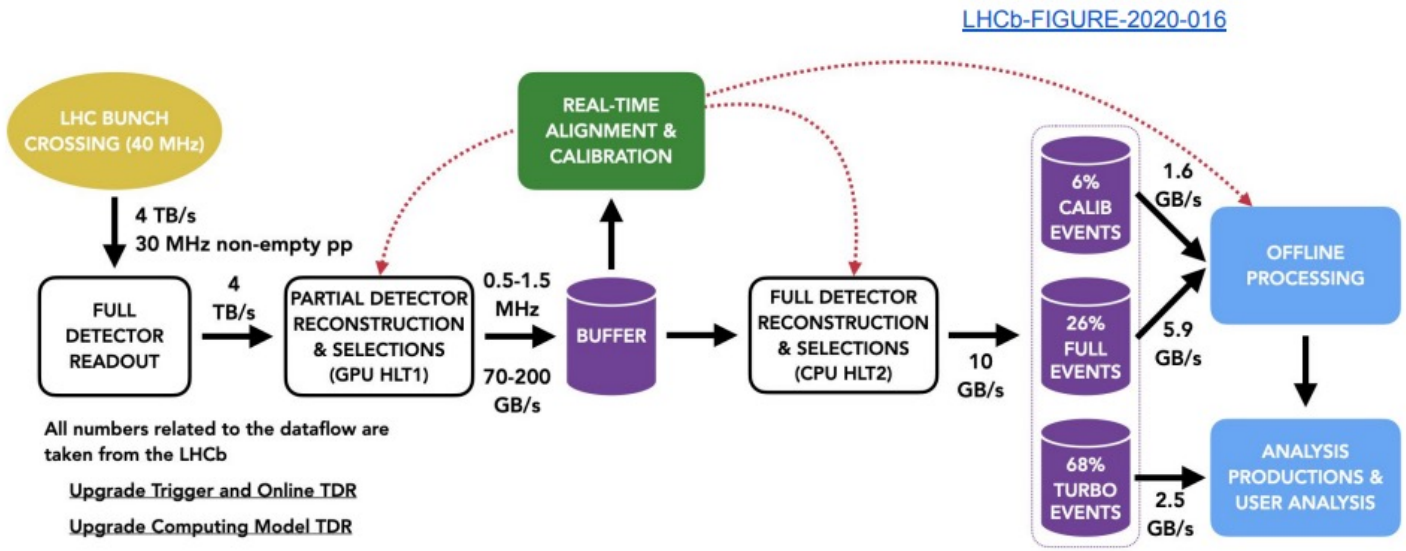
High-performance network protocols are poised to address TB/s throughput demands.

在线计算发展趋势

Run 1 → 2011 → Run 2 → 2015 → Run 3 → Now → Future



"Triggering Tb/s of data: LHCb perspective", CHEP2024



基于缓存的数据处理处理，增加了在线计算系统的灵活性

buffer between the two software levels. The flexibility in trigger processing that is provided by this buffer system allows the execution of high-quality alignment and calibration between HLT1 and HLT2.

"Tesla: an application for real-time data analysis in High Energy Physics."Computer Physics Communications 208 (2016)

- Run 2 开始在HTL1和HLT2间增加了磁盘缓存，使两级数据处理可以独立运行