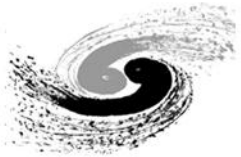


中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

国家高能物理科学数据中心
National HEP Data Center

高能所計算中心
IHEP Computing Center

面向HEPS的IO方法设计和优化

汇报人：符世园

<fusy@ihep.ac.cn>

中国科学院高能物理研究所

目录

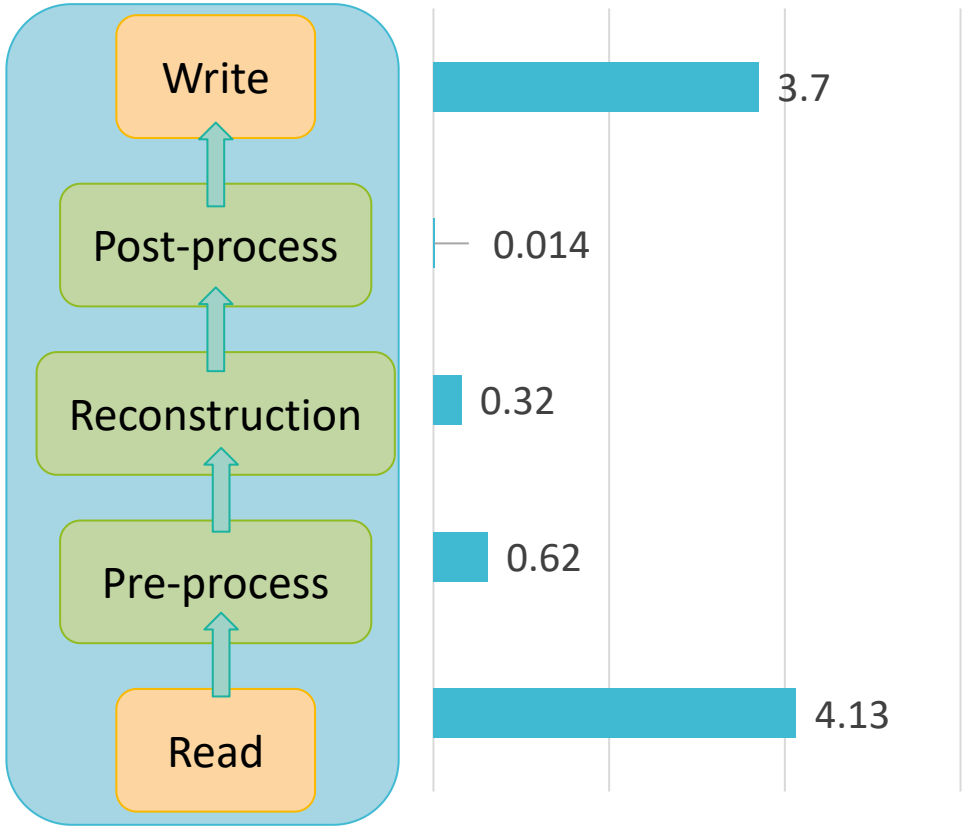
- 高能同步辐射光源HEPS
- 优化方法
- 架构设计
- 下一步计划

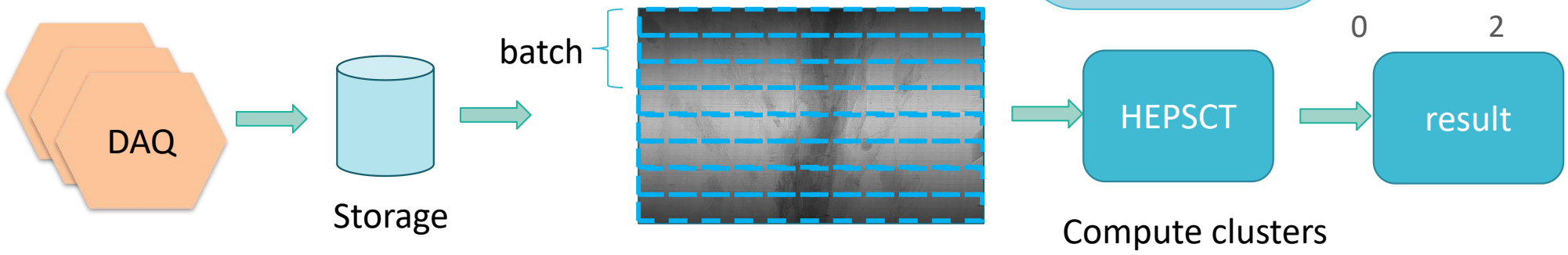
高能同步辐射光源（HEPS）

- HEPS预计产生海量原始实验数据
  - 包含14条线站
  - 多模态、跨尺度、高帧率
  - 面向多学科、多方法学

线站名称	每天数据量峰值（TB）	每天数据量均值（TB）
B1	600	200
B2	500	200
B3	8	3
B4	10	3
B5	10	1
B6	2	1
B7	1000	250
B8	80	10
B9	20	5
BA	35	10
BB	400	50
BC	1	0.2
BD	10	1
BE	25	11.2
BF	1000	60
总计		805

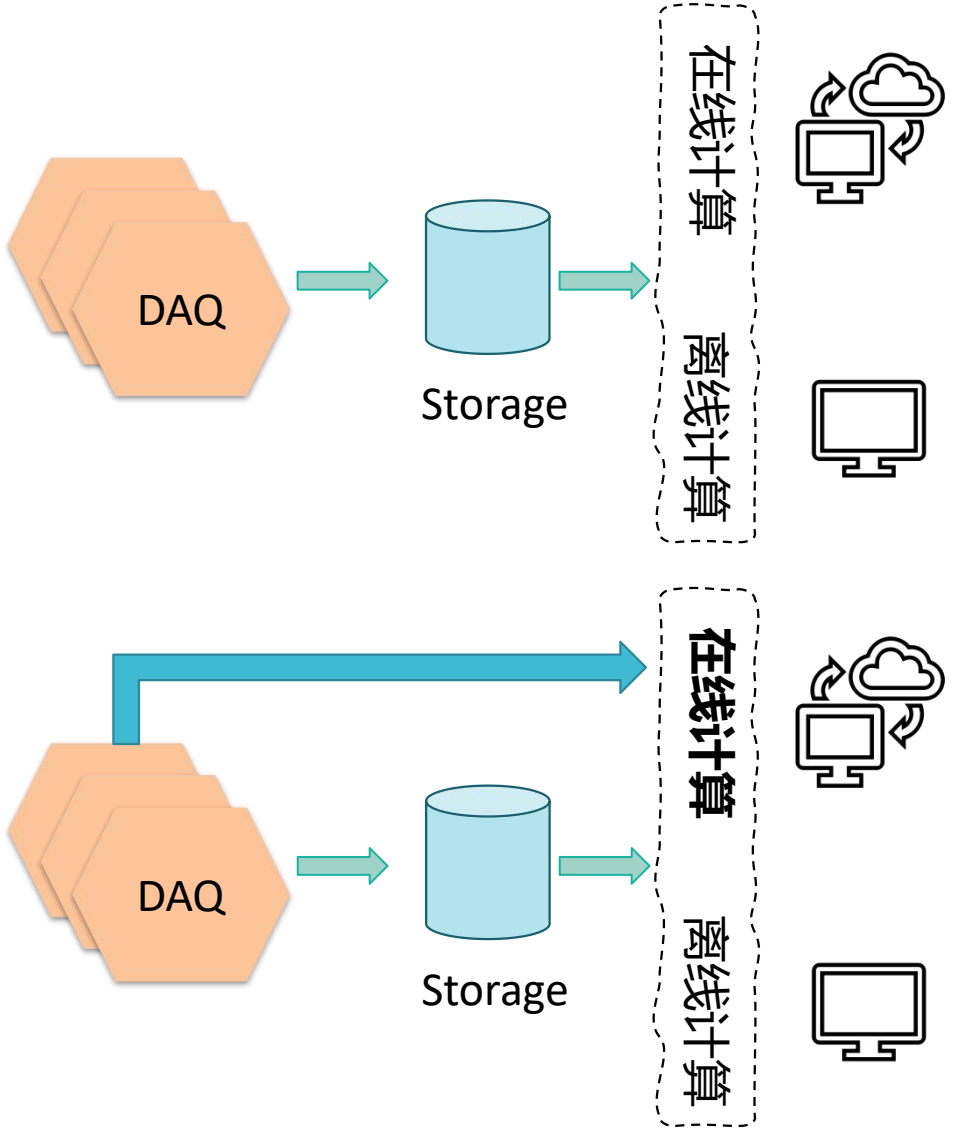
- 海量数据为科学计算带来极大压力
  - 目前的数据处理流程（以B7为例）
    - B7主要计算任务是CT重建（HEPSCT）
      - 重建最小单位（unit）：所有图像的同一行数据
    - 计算过程分为多个batch，相互独立
    - 读写占据绝大部分时间
  - IO瓶颈严重影响计算效率





•优化方法

- STEP 1. 批处理IO加速
  - 目前数据处理流程：数据落盘后再读取
  - 文件格式：HDF5 TIFF
  - 加速方法：异步、多线程、多进程等
- STEP 2. 对于在线计算引入流数据IO
  - 从DAQ直接获取数据发送到计算节点
  - 避免数据落盘再读取导致的时延和IO瓶颈
  - 在实验级别实现所见即所得

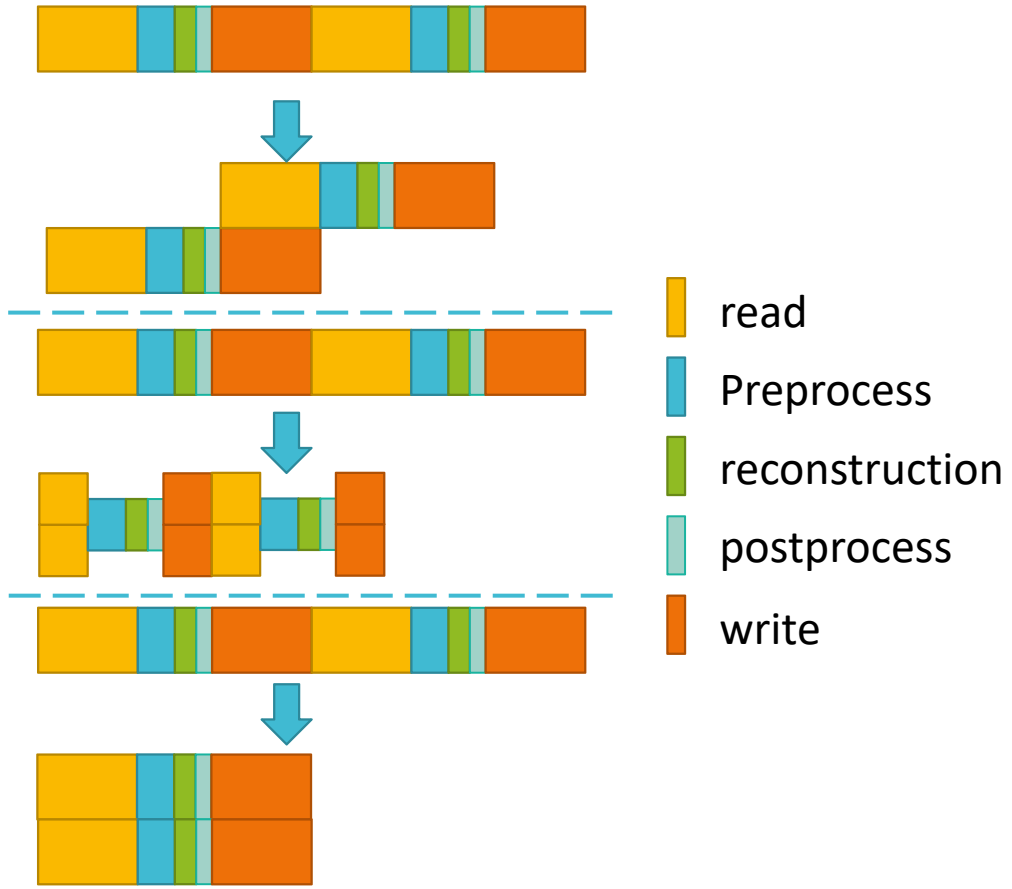


批处理IO加速

- 数据格式： HDF5、 TIFF
  - HDF5： 一种数据模型、文件格式和 I/O 库
    - 用于存储、交换、管理和归档复杂数据，包括科学、工程和遥感数据
  - TIFF： 图像格式
- IO加速方法：
  - 多进程：与并行计算结合，多个batch同时计算
  - 多线程：加速单个batch的数据读取
  - 异步IO：读写和处理过程异步执行，加速数据处理
  - HDF5分块存储
  - 具体实现需要结合具体计算任务

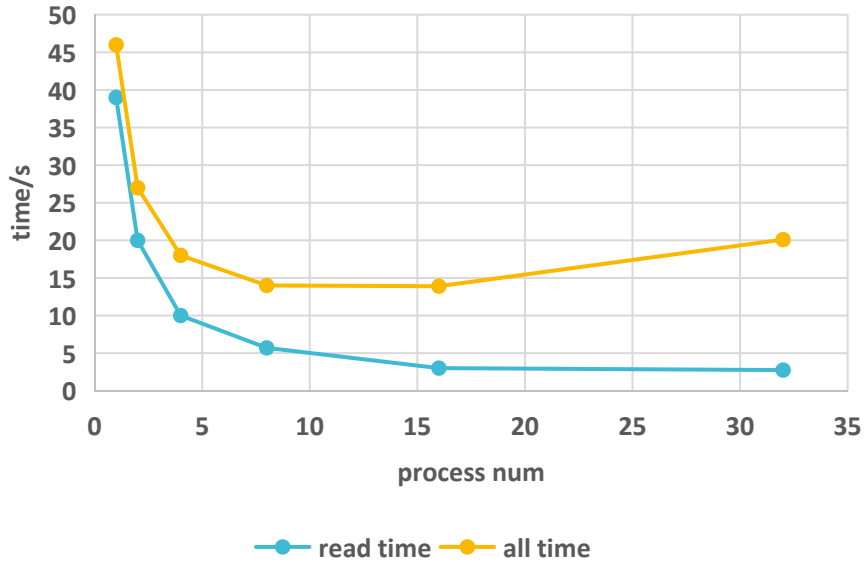
批处理IO加速

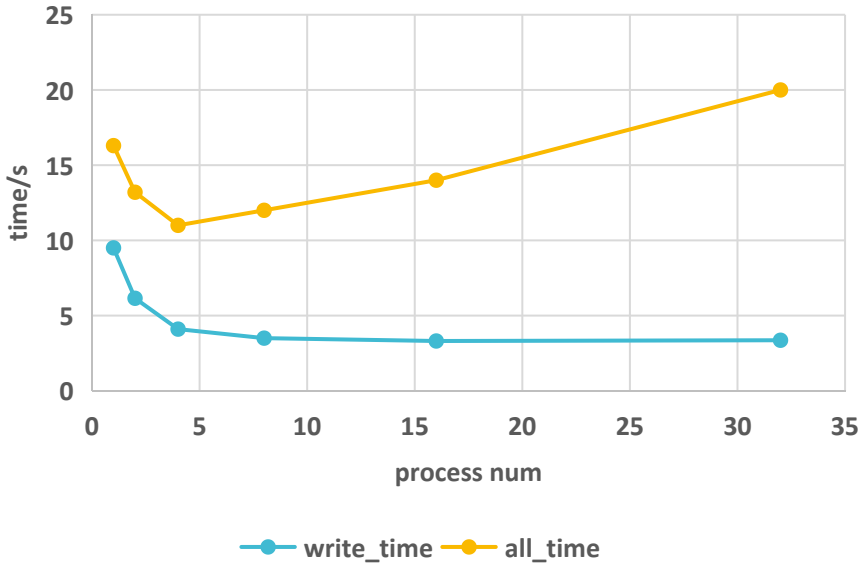
- 异步
  - HEPSCT中相邻batch读写操作可以异步执行
  - 加速效果：~100S-->~75s
- 多线程
  - 加速单个batch的数据读取（TIFF）
  - 第一个batch读取~10s，后续batch读取 < 1s
- Pipeline流水线技术
  - 与并行计算结合
  - 多个batch同时计算



批处理IO加速

- HDF5特性-并行
  - 基于MPI
  - 测试数据：5000(frames)×400(rows)×500(cols)，保存为10个HDF5文件
  - 每次读取的batch大小: 500×1×500 vs. 500×10×500



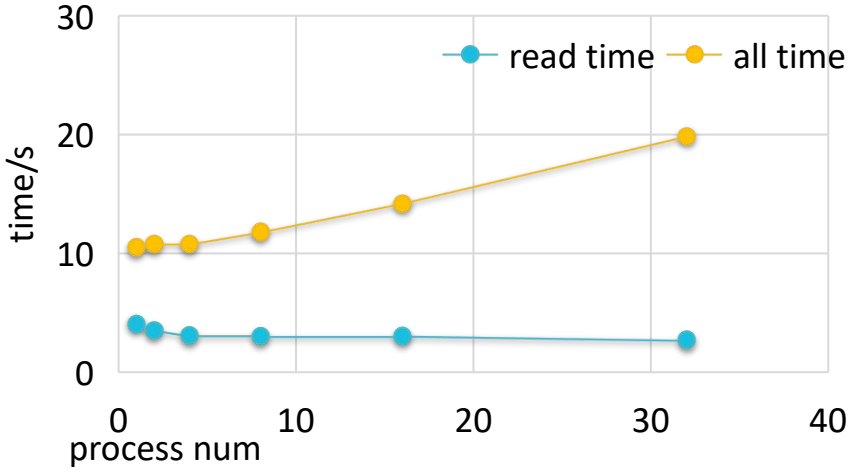


  - 结论：多进程可以有效加速HDF5读取速度

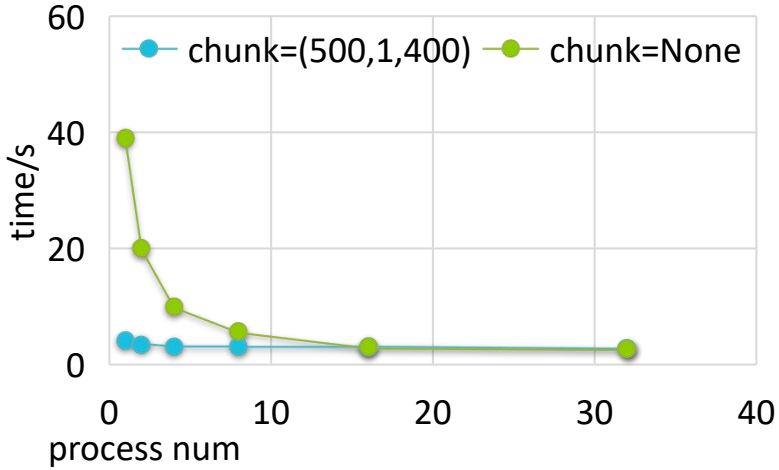
批处理IO加速

- HDF5特性-分块chunk
  - Data: 5000(frames)×400(rows)×500(cols)
  - chunk和读取batch大小: 500×1×500

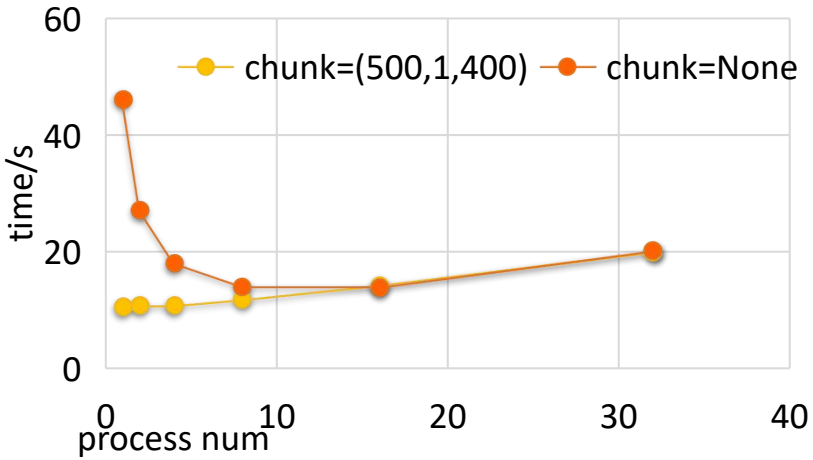
分chunk后读时间



分chunk前后read time对比



分chunk前后all time对比



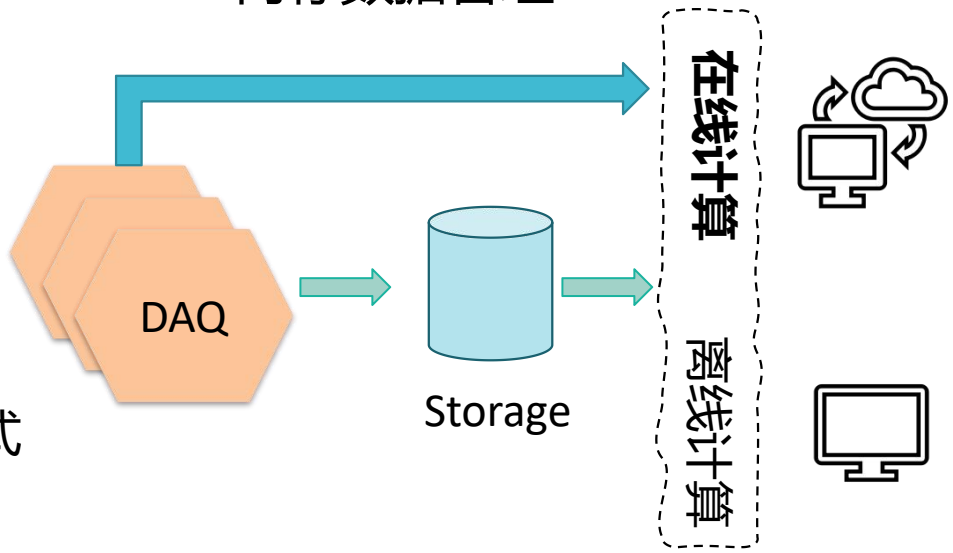
- 结论：块大小和读取batch一致时可以以更少的进程数大幅加速数据读速度
  - 局限性：分块属性需要在文件写入时指定，影响文件写入效率；块大小的精确设置
  - 解决方法：多线程、多进程；块大小设置的模糊值

流数据IO-需求

- 流数据IO连接DAQ与计算平台
  - 面向DAQ：
    - 稳定的流数据接收器
    - 大内存存储从DAQ接收的数据
  - 面向计算平台
    - 流数据IO接口与批处理IO接口的统一性
      - 不同计算任务、不同数据来源、不同数据格式
    - 统一内存数据模型
      - 满足跨语言跨平台计算任务需求

数据流稳定接收和暂存

1. 数据流接收
2. 内存数据管理

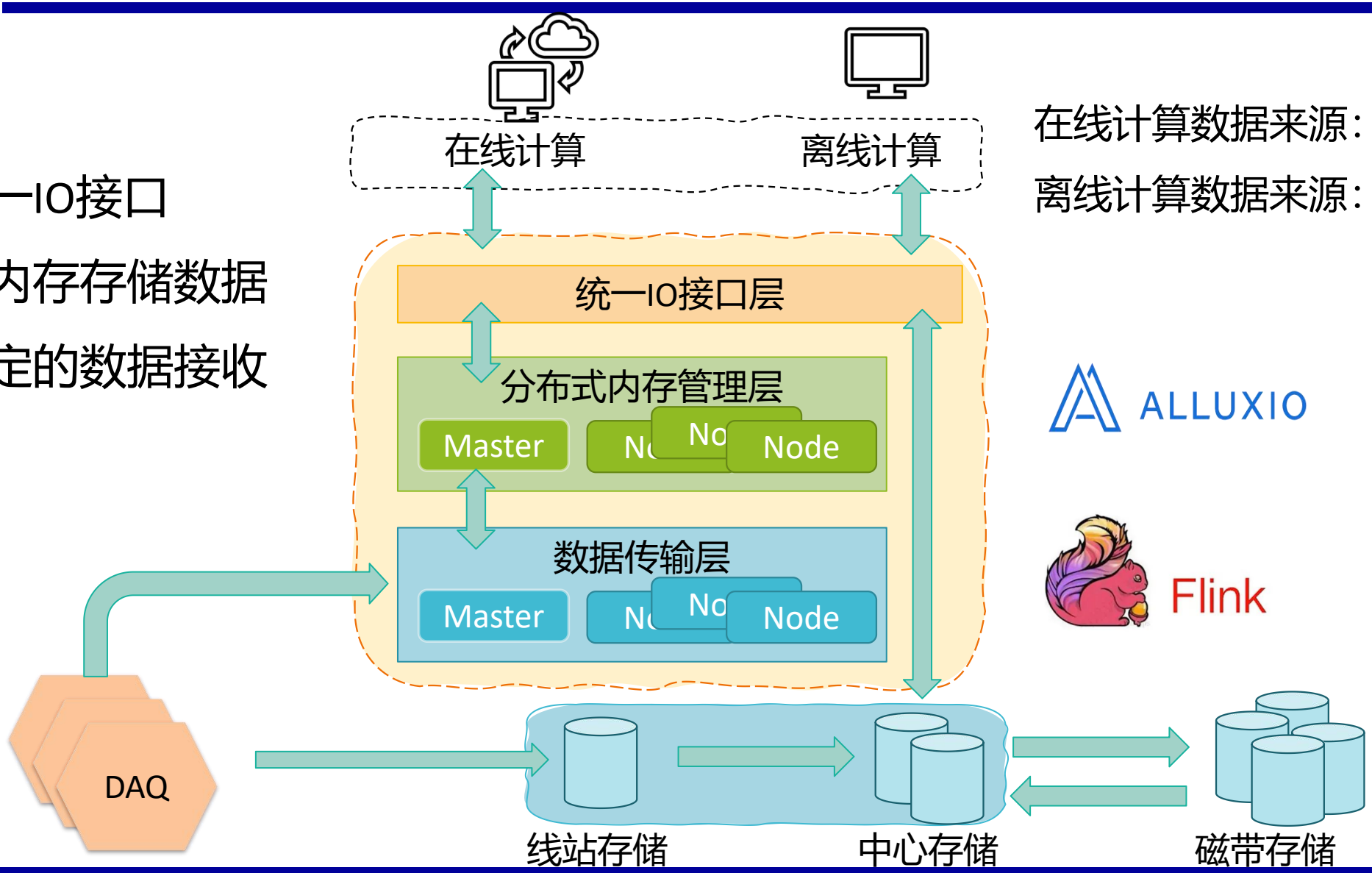


数据访问统一化

1. 数据接口一致
2. 数据模型跨平台跨语言访问
3. 兼容不同数据格式

整体架构设计

- **目标**：
  - 统一IO接口
  - 大内存存储数据
  - 稳定的数据接收



在线计算数据来源：DAQ和线站存储

离线计算数据来源：中心存储

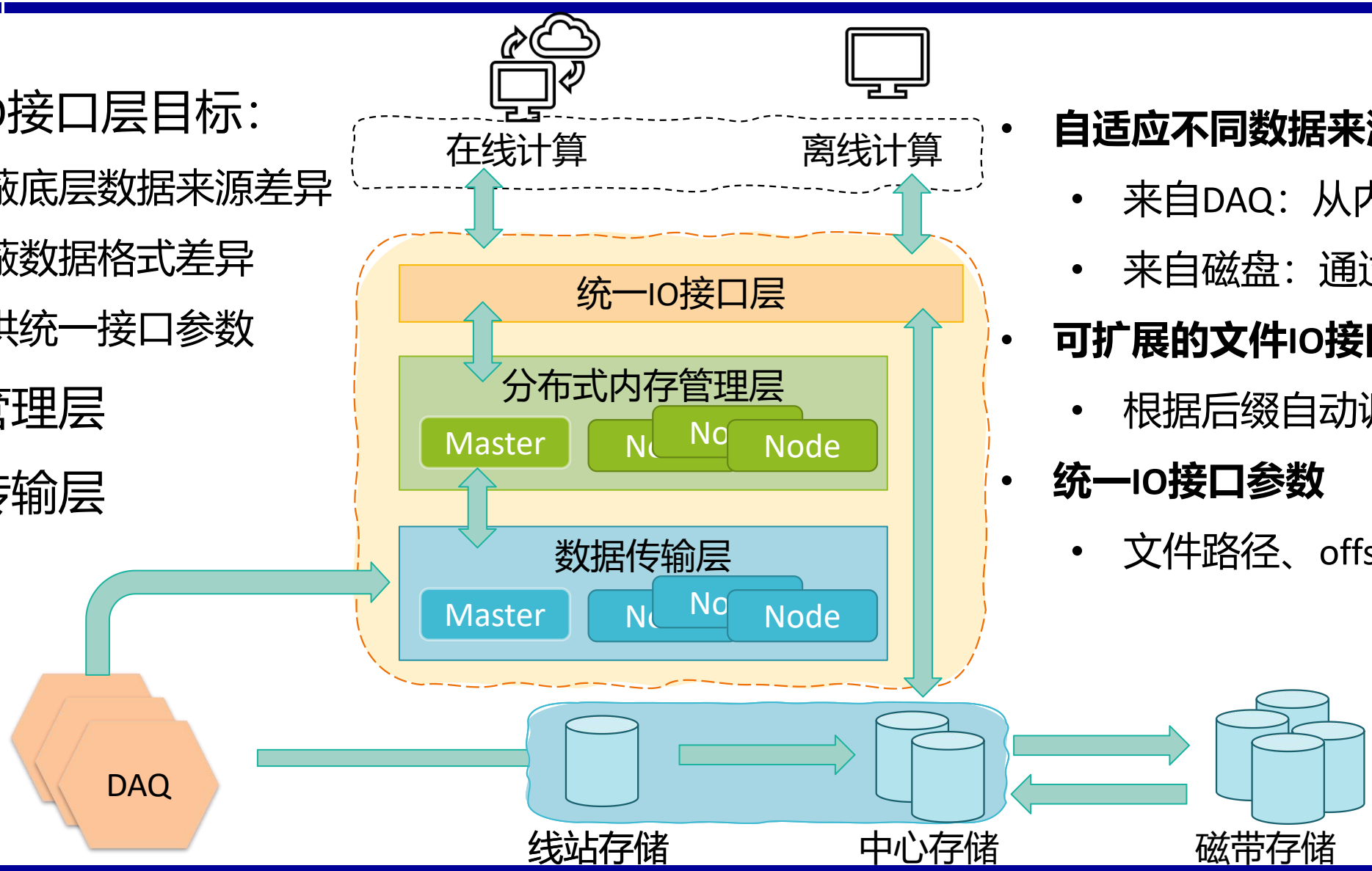
ALLUXIO

APACHE ARROW

Flink

架构设计-统一IO接口

- 统一IO接口层目标：
  - 屏蔽底层数据来源差异
  - 屏蔽数据格式差异
  - 提供统一接口参数
- 内存管理层
- 数据传输层

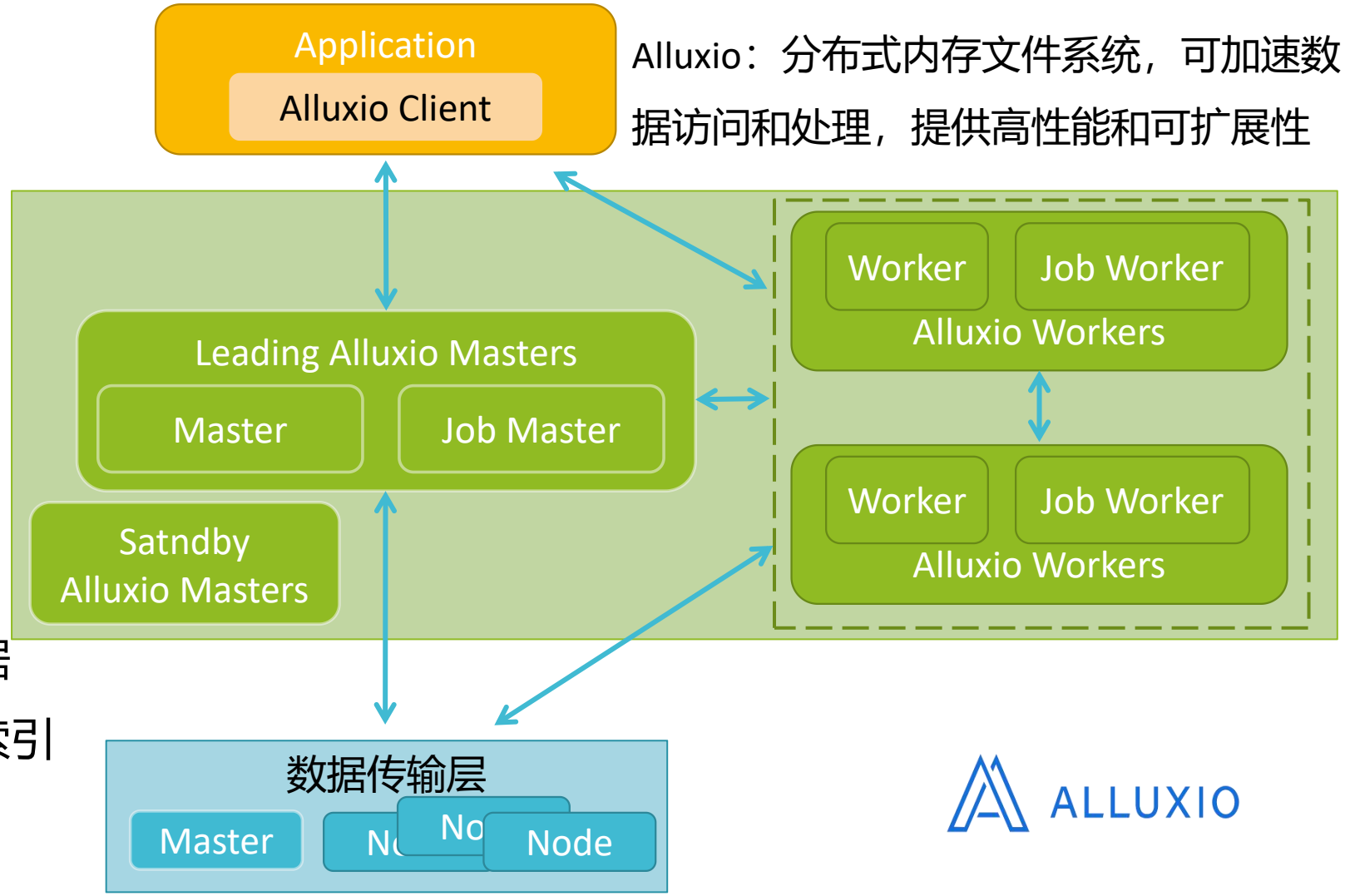


- **自适应不同数据来源**
  - 来自DAQ：从内存中获取数据
  - 来自磁盘：通过挂载的方式获取
- **可扩展的文件IO接口**
  - 根据后缀自动调用对应IO接口
- **统一IO接口参数**
  - 文件路径、offset、batch_size等

架构设计-内存管理

- 内存管理层目标：
  - ✓ 暂存实验数据-->分布式内存
  - ✓ 建立临时索引-->元数据管理
  - 统一内存数据模型

- **基于Alluxio的分布式内存管理**
  - 通过分布式内存节点存储流数据
  - 注册临时目录为数据建立唯一索引



Alluxio：分布式内存文件系统，可加速数据访问和处理，提供高性能和可扩展性

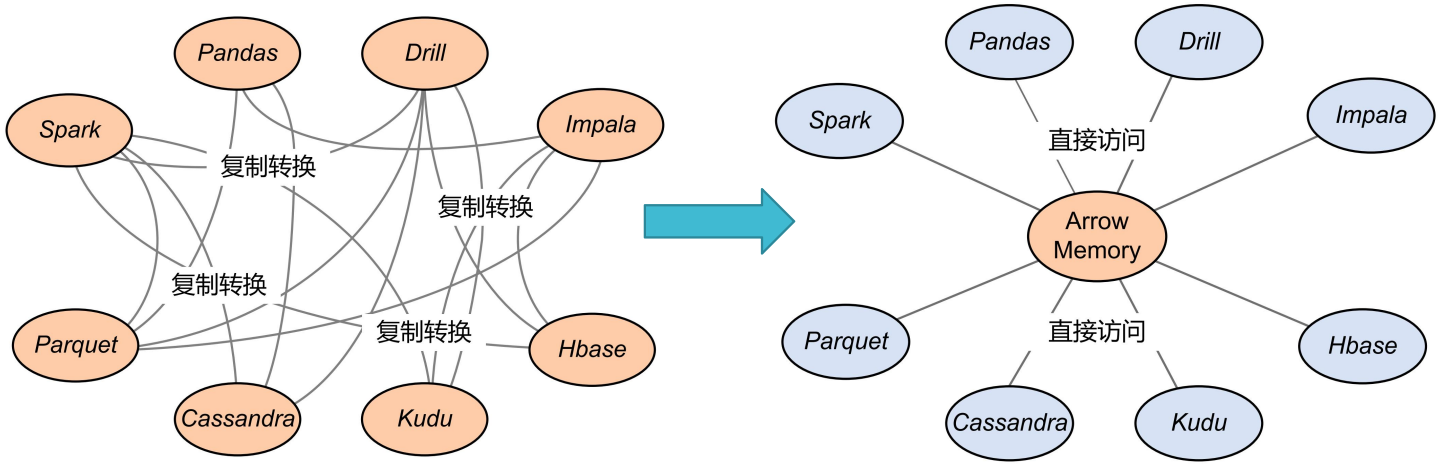
ALLUXIO

架构设计-内存管理

- 内存管理层目标：
  - 暂存实验数据
  - 建立临时索引
  - ✓ 统一内存数据模型-->跨语言跨平台

APACHE ARROW

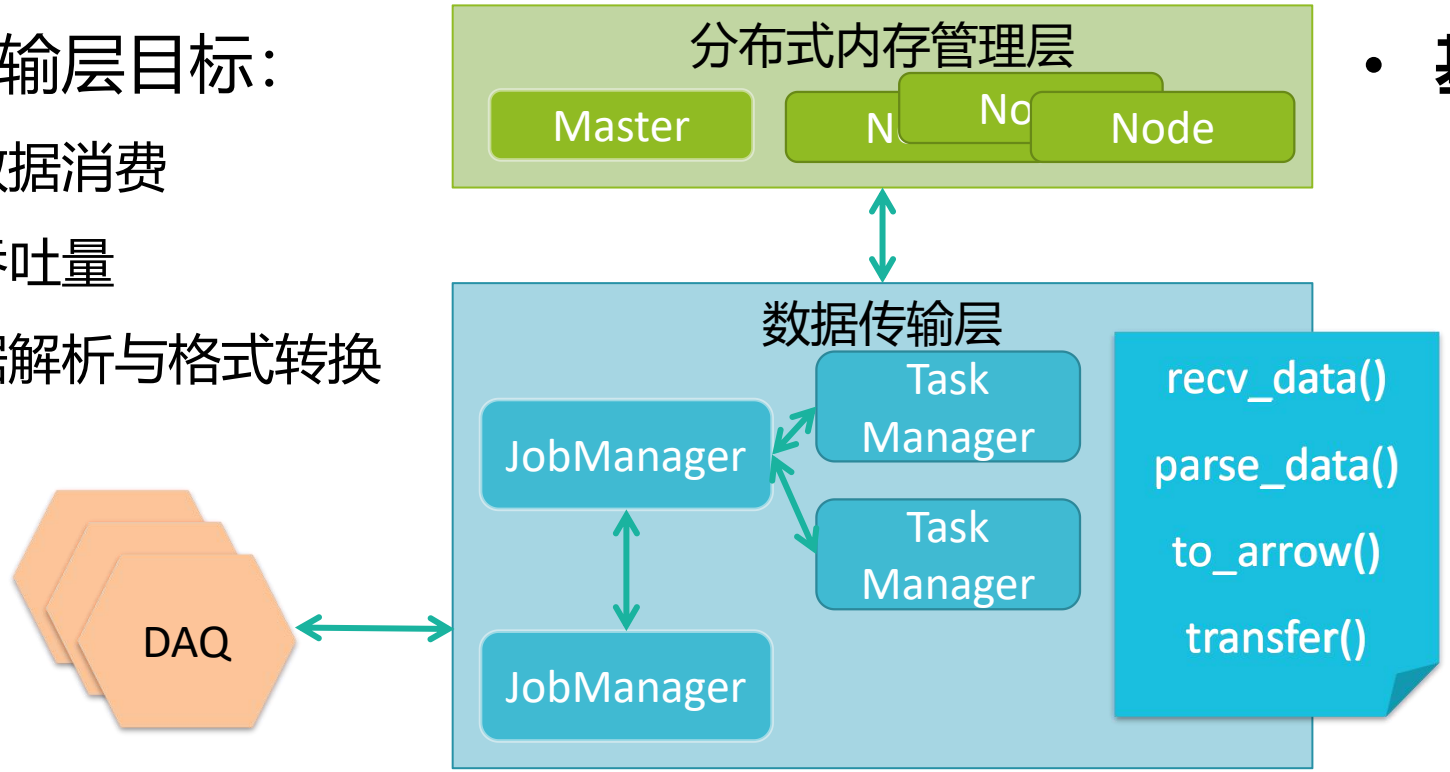
- 高性能：使用zerocopy技术减少数据序列化反序列化开销，
- 跨平台和跨语言：可在不同的计算引擎和编程语言之间共享数据



- **基于Arrow的共享内存数据模型**
  - 在Python中，可使用pyarrow库将NumPy数组转换为Apache Arrow格式
    - arrow_array = pyarrow.array(numpy_array)
  - 在转换过程中，Arrow并不会复制整个数据数组，而是创建一个元数据对象来引用NumPy数组的数据
  - 下一步需要根据具体情况进行性能测试和优化

架构设计-流数据消费

- 数据传输层目标：
  - 流数据消费
  - 高吞吐量
  - 数据解析与格式转换

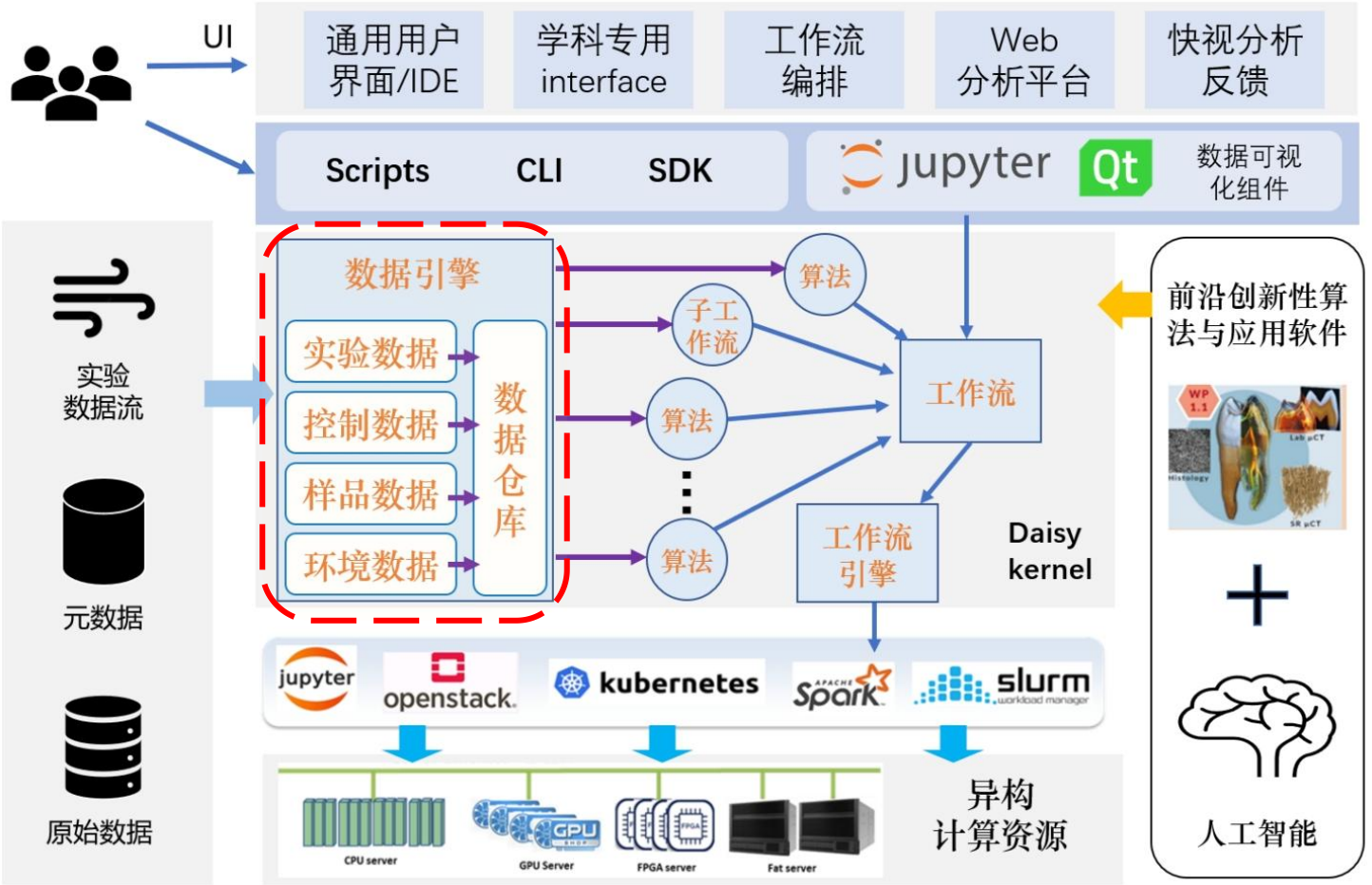


- Flink特点：
  - 通过优化的流处理引擎和内存管理，实现了高吞吐和低延迟的数据处理。
  - 可以与各种数据源（如Kafka、Hadoop、Amazon S3等）和数据接收器进行集成

- **基于Flink的流数据消费与转换**
  - 高可用性架构
    - 部署多个JobManager
    - 稳定高效消费数据流
  - 流数据格式自动解析
    - 不同线站流数据内容不同
  - 格式转换与存储
    - 解析完毕的数据转换为Arrow格式并存储到Alluxio

Flink

Daisy-数据引擎

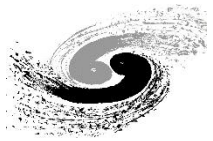


- 作为Daisy中的数据引擎
  - 接收来自DAQ和存储系统的数据
    - 数据流和文件
  - 向计算引擎提供所需数据
    - 使用接口即可，屏蔽底层差异
  - 提供流数据处理和批处理能力
  - 优化IO速度
  - 屏蔽地层差异

DAISY(Data analysis integrated software system) 是为先进光源在线数据处理而设计开发的软件框架。 目标设计实现一个通用的、具有良好扩展性的基础软件架构，集成多种方法学算法和工具，屏蔽计算架构的复杂性和计算资源的多样性，为上层应用软件和用户提供统一和简单的调用接口，并在此基础上开发数据可视化和分析桌面等通用组件，以期形成一个丰富和繁荣的软件生态环境。

Daisy详见：

《HEPS数据处理软件框架Daisy的进展及规划》胡誉



下一步计划

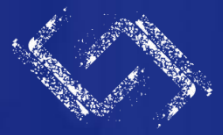
- 批处理优化
  - HDF5多线程IO
  - 与并行计算的结合
- IO框架性能测试
  - Flink消费流数据的能力是否满足HEPS产生的数据量
  - Arrow内存测试及优化
  - 稳定性测试

谢谢!

敬请批评指正!

国家高能物理科学数据中心

National HEP Science Data Center



高能所計算中心

IHEP Computing Center