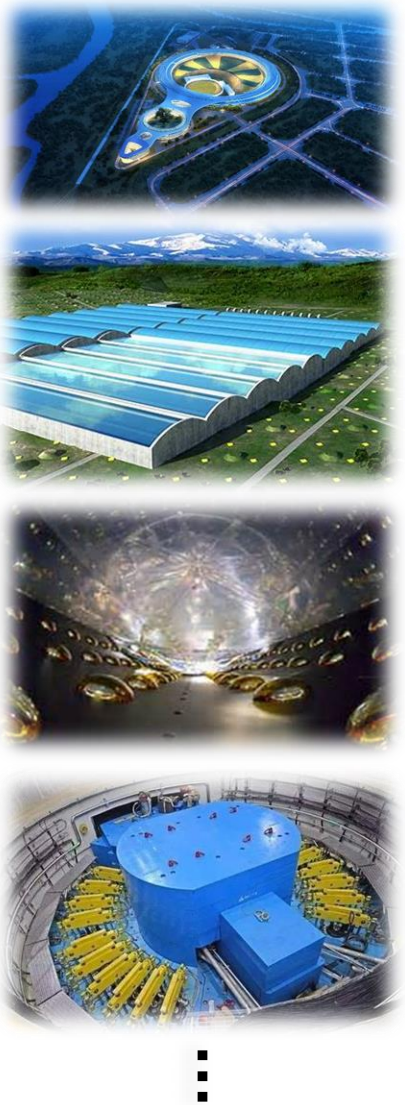
中国科学院高能物理研究所
Institute of High Energy Physics Chinese Academy of Sciences

面向高能物理数据处理的可计算存储系统设计与实现

报告人：高 宇
导　师：程耀东 研究员

2022年8月11日

研究背景：典型高能物理计算平台



分布式计算系统

前端登录集群

本地计算集群

计算子系统

防火墙

高速核心网络

磁盘存储

磁带存储

数据备份

实验站

专线链路

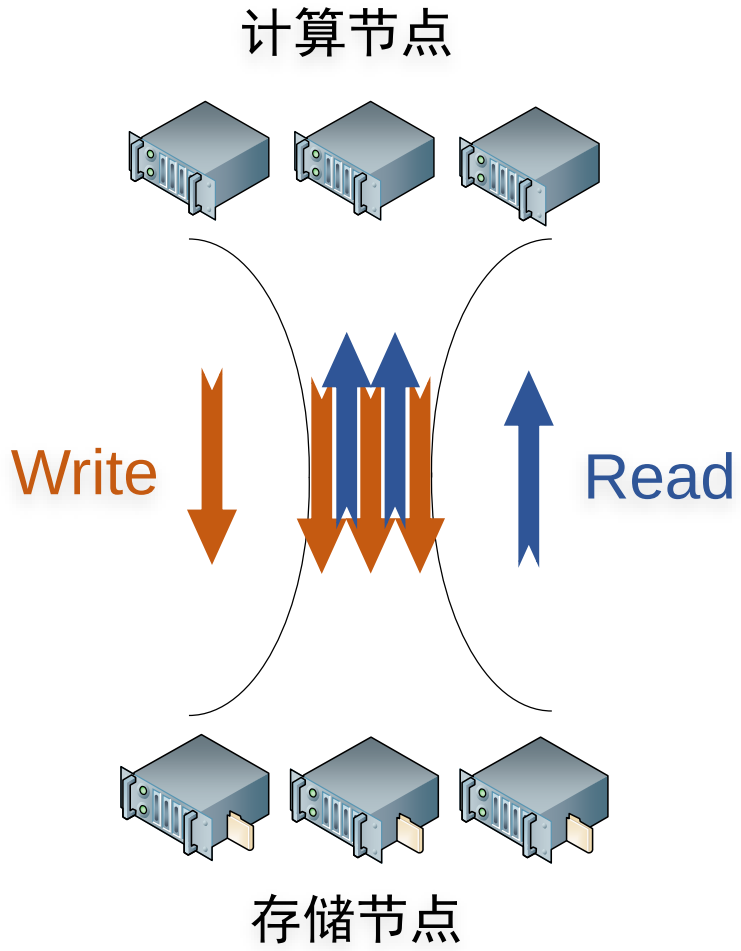
数据传输

网络子系统

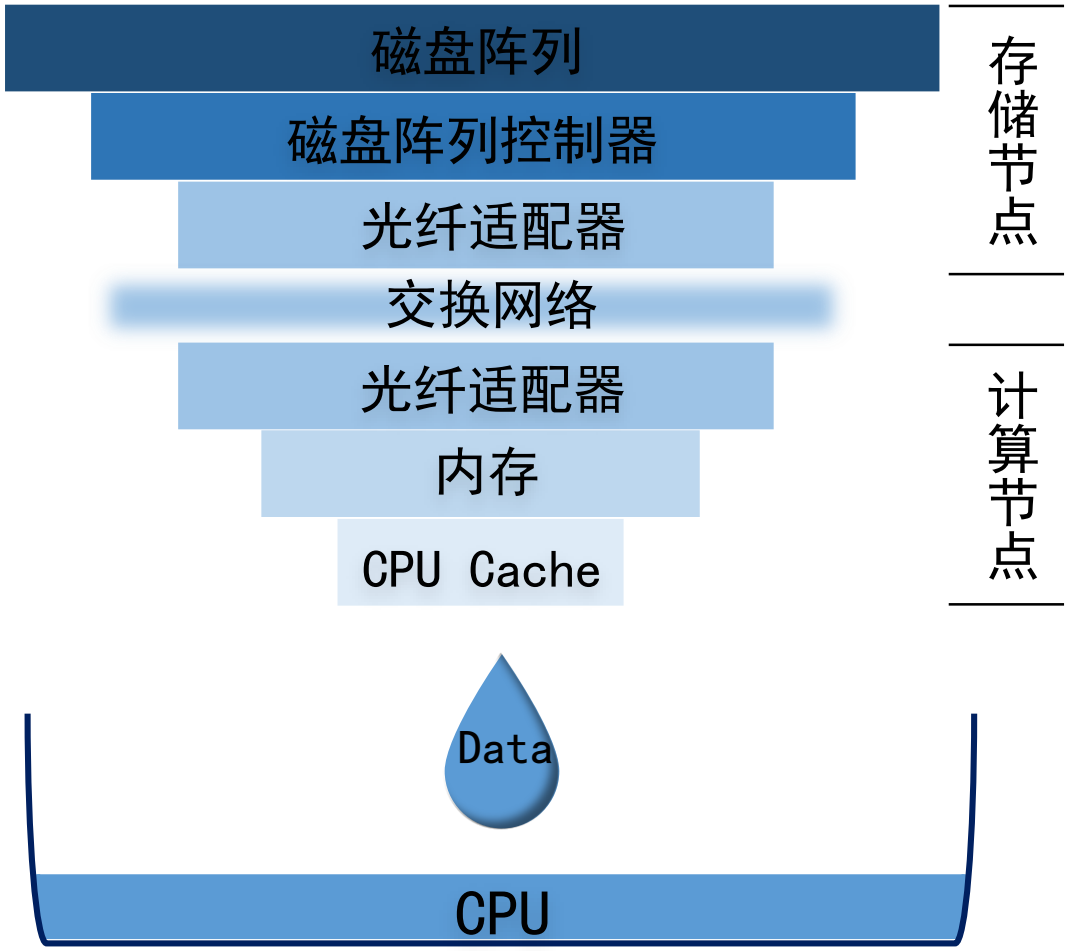
存储子系统

研究背景：存-算分离的弊端

1、网络瓶颈



2、存储瓶颈



研究背景：可计算存储

国内外现状

全球网络存储工业协会SNIA成立了可计算存储工作组，制定技术标准

SNIA

国际高能物理领域成立数据管理组织DOMA

Computational Storage Technical Work Group

Data Organization, Management and Access (DOMA)

SkyHookDM：加州大学圣克鲁兹分校项目，基于CEPH实现通用数据管理功能的卸载，包括筛选、映射、聚合、索引等

SkyhookDM

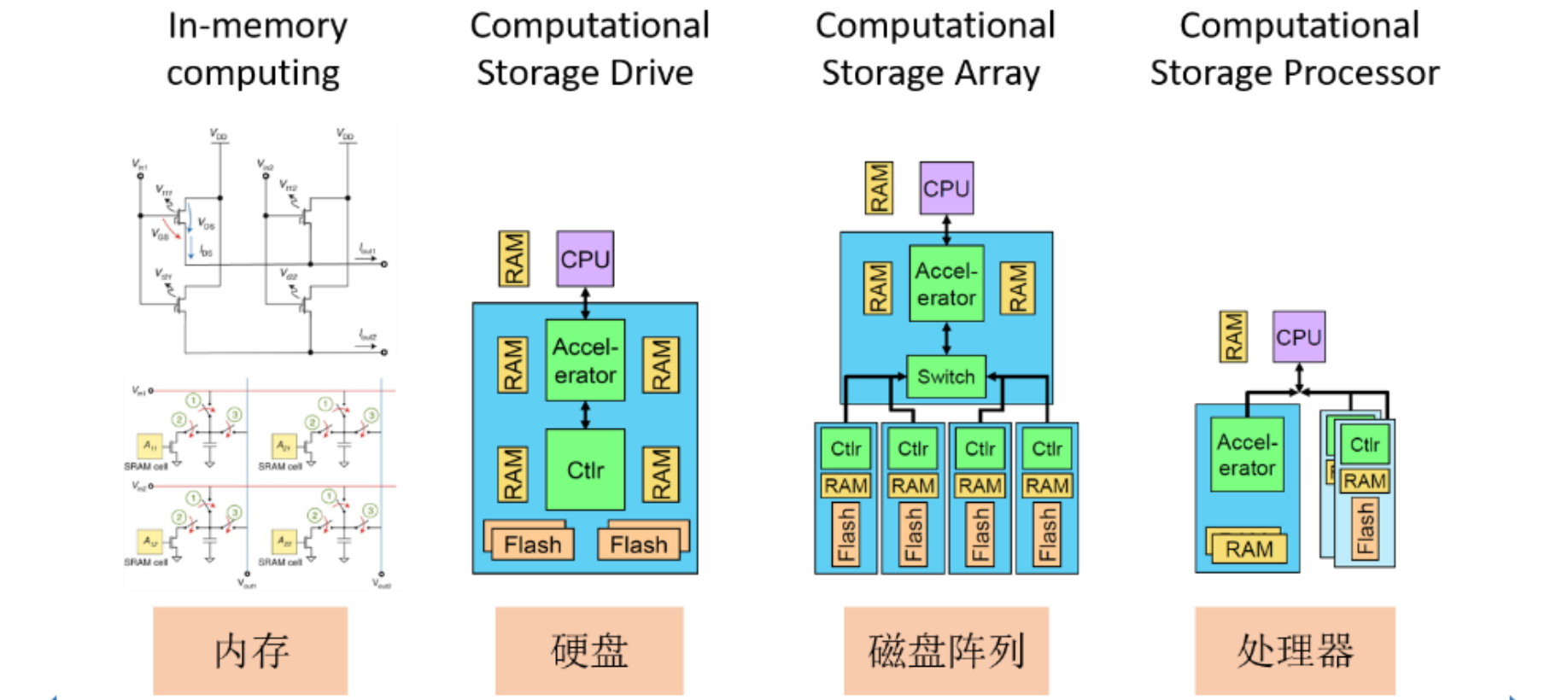
PolarDB：阿里云与利用scaleflux等具有计算存储驱动器的云本地数据库

POLARDB

目前的工作多集中在数据库的操作，针对海量数据文件操作的研究较少

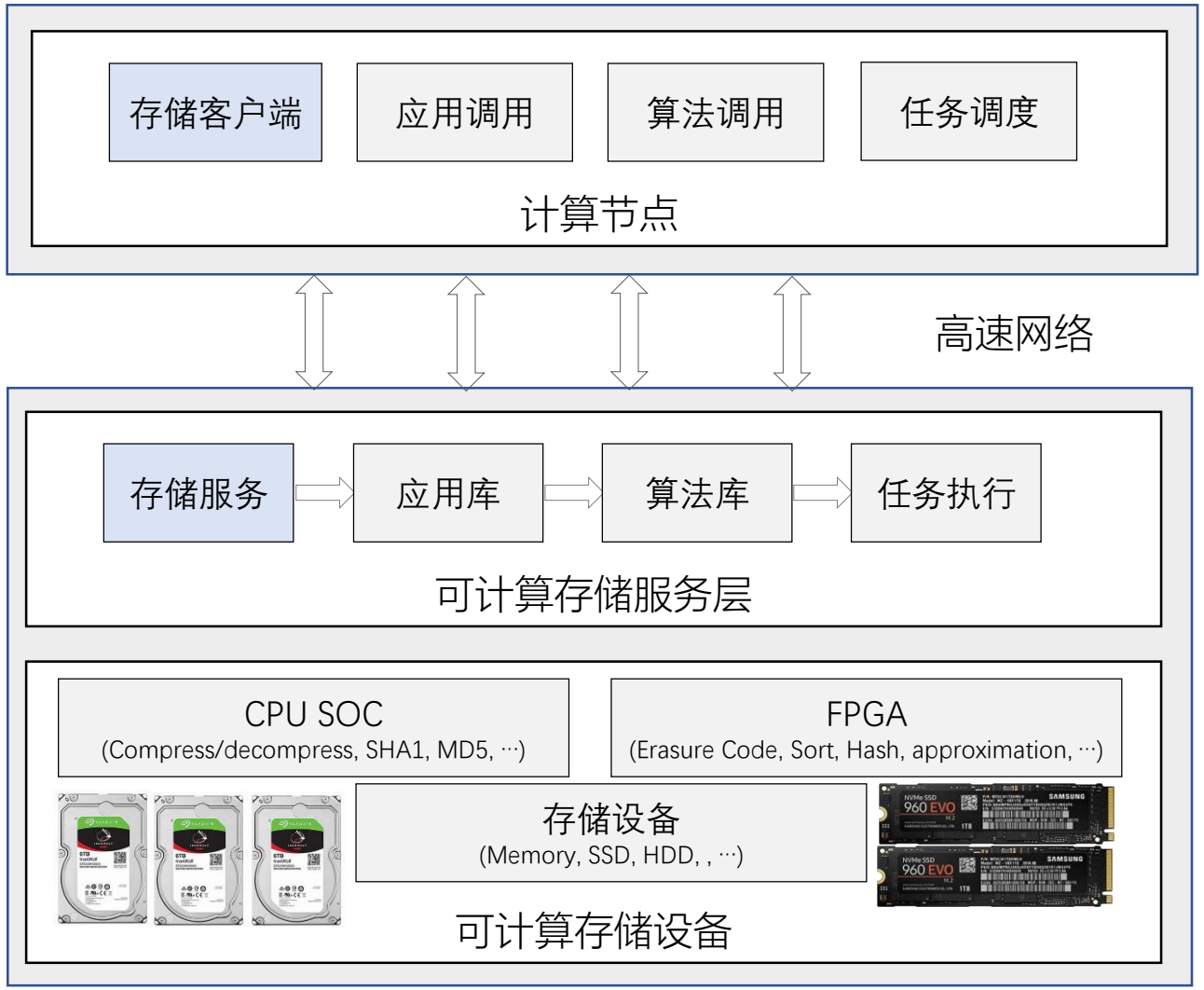
研究背景：可计算存储

可计算存储架构有多种不同层次的实现



总体框架

- 让存储具备“计算能力”
  - “就近计算”，降低网络传输压力
- 硬件研制
  - ARM CPU + Xilinx Kintex UltraScale+ FPGA
  - 形成面向高能物理的定制化硬件，绿色节能
- 软件研究
  - 存储软件，基于EOS实现
  - 中间件，基于ROOT实现
- 应用支撑
  - 通用数据压缩，算术编码，图片/视频压缩，机器学习
  - 排序，拟合，去弥散等常用ROOT数据处理卸载
- 目标
  - 效率高，软硬件协同的数据处理系统
  - 自主可控

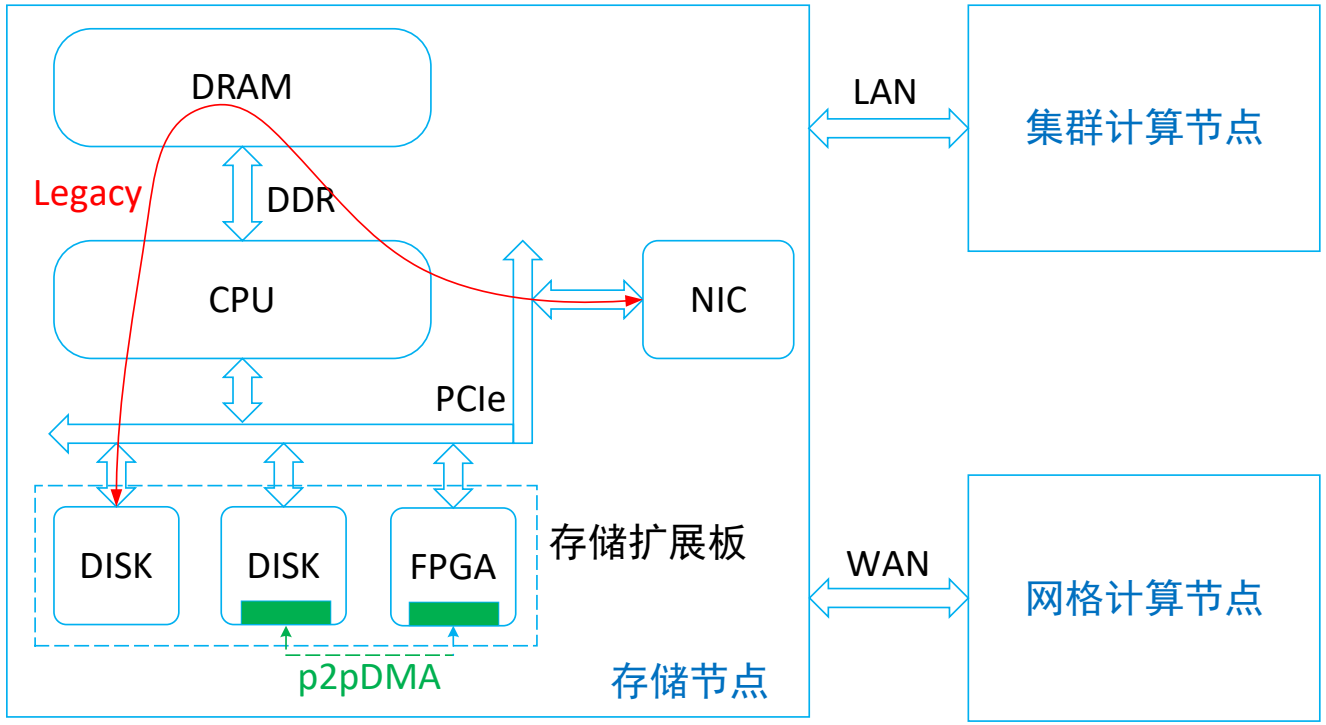


硬件框架

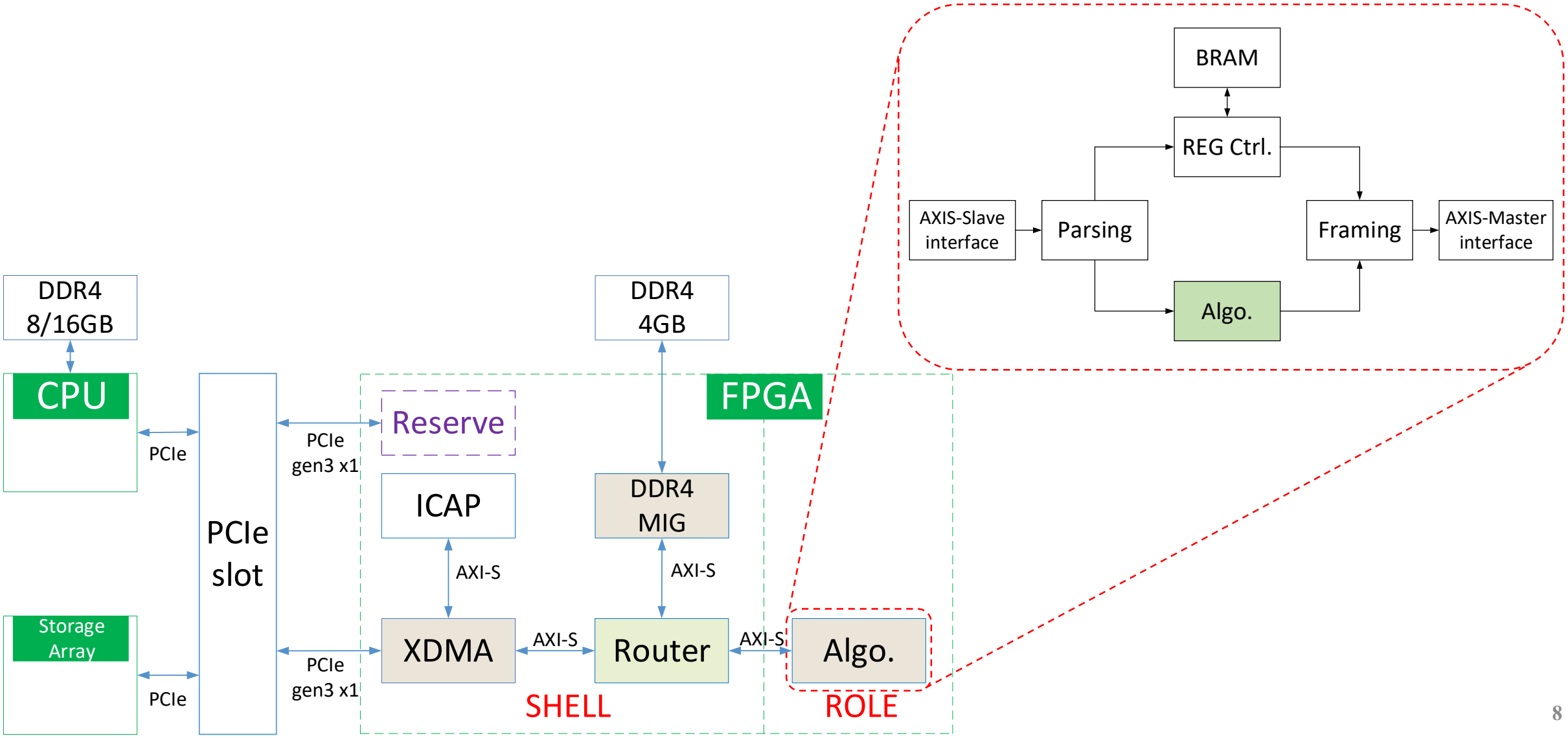
CPU通过PCIe总线访问硬盘阵列和FPGA

可使用CPU进行计算，或将计算任务下发至FPGA实现异构计算

硬盘与FPGA之间可直接交互数据，减少数据拷贝开销



FPGA框架

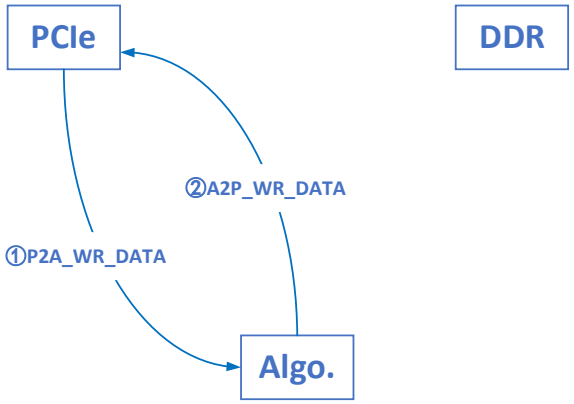

BRAM
REG Ctrl.
AXIS-Slave interface
Parsing
Framing
AXIS-Master interface
Algo.
DDR4 8/16GB
CPU
PCIe
PCIe slot
PCIe gen3 x1
Reserve
ICAP
AXI-S
DDR4 4GB
FPGA
DDR4 MIG
AXI-S
Storage Array
PCIe
PCIe gen3 x1
XDMA
AXI-S
Router
AXI-S
Algo.
SHELL
ROLE

FPGA数据帧

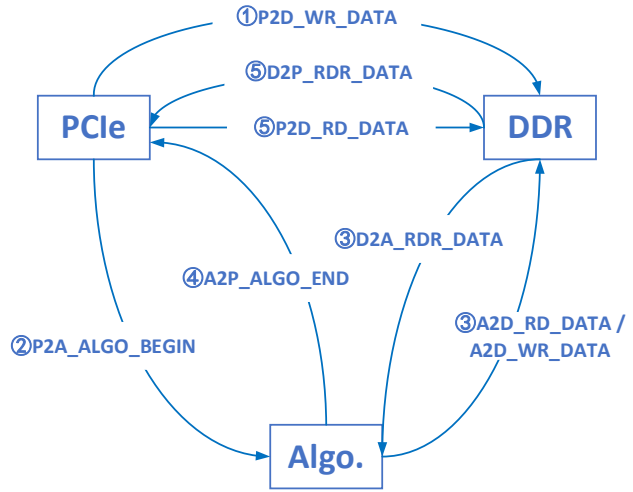
SRC	DST	OPT	TSK	ADDR	LEN
RSV					
< 256K data					

不同的应用对数据传输和处理方式有着不同的要求。

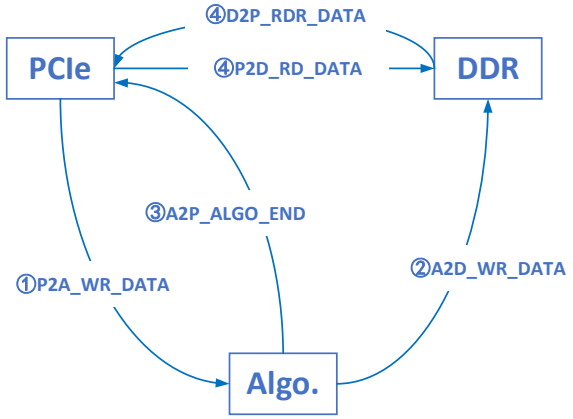
算法模块和DDR控制器可主动发起数据传输，通过数据帧传输到目的模块。



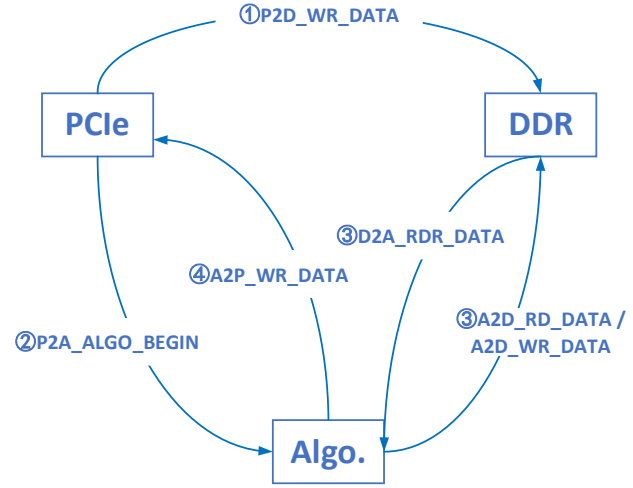
(a) 纯流式处理



(b) 纯批处理



(c) 半流式处理半批处理



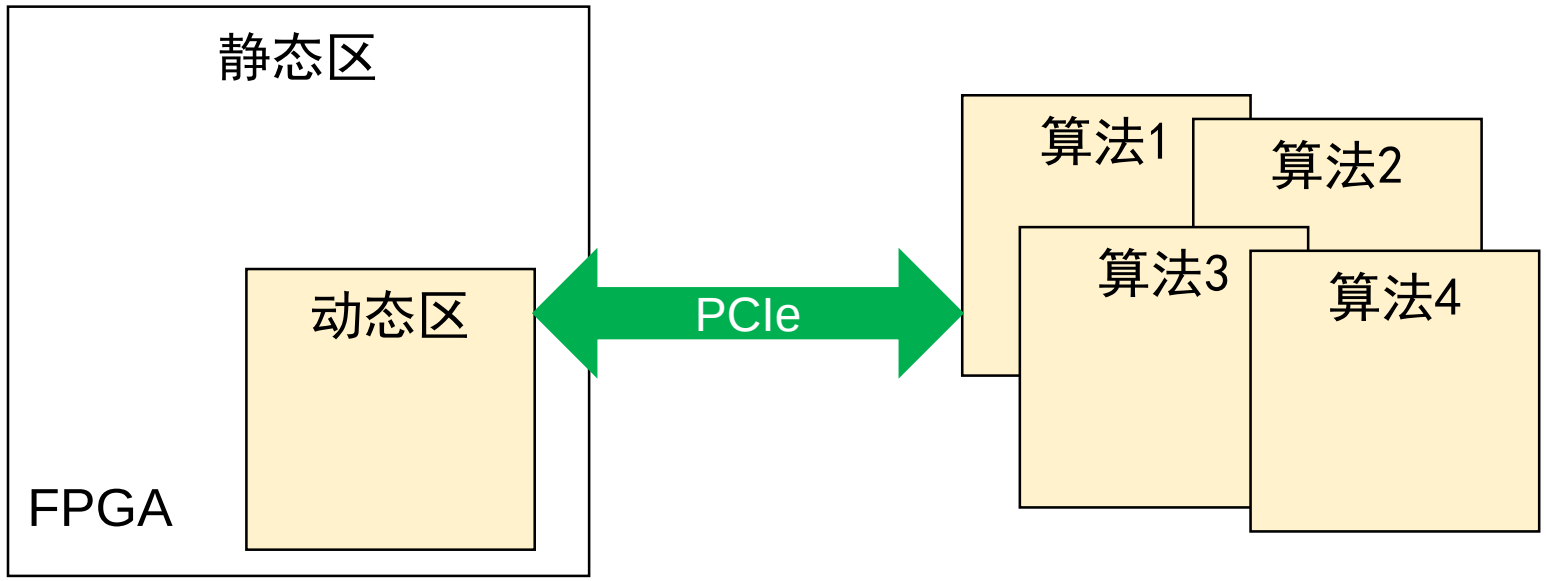
(d) 半批处理半流式处理

FPGA应用的动态加载

在FPGA中，将逻辑区域划分为动态区和静态区：

- 静态区包含PCIe接口逻辑，DDR4控制逻辑等
- 动态区用于实现具体的算法

下发新的计算加速任务时，CPU首先通过PCIe总线对FPGA的动态区进行更新，由此实现算法应用的动态切换



FPGA数据传输性能

PCIe Gen3.0 x1（8b/10b编码）理论最大带宽 819.2MB/s

非数据帧读写BRAM，单独读写的情况下：
（AXI-Stream总线64bit位宽，125MHz主频）
- 单独读最大带宽约 590MB/s
- 单独写最大带宽约 670MB/s

使用数据帧读写DDR4（1600MHz），混合读写的情况下：
（AXI-Stream总线128bit位宽，100MHz主频）
- 读最大带宽约 330MB/s
- 写最大带宽约 360MB/s

三种服务方式

算法层次：

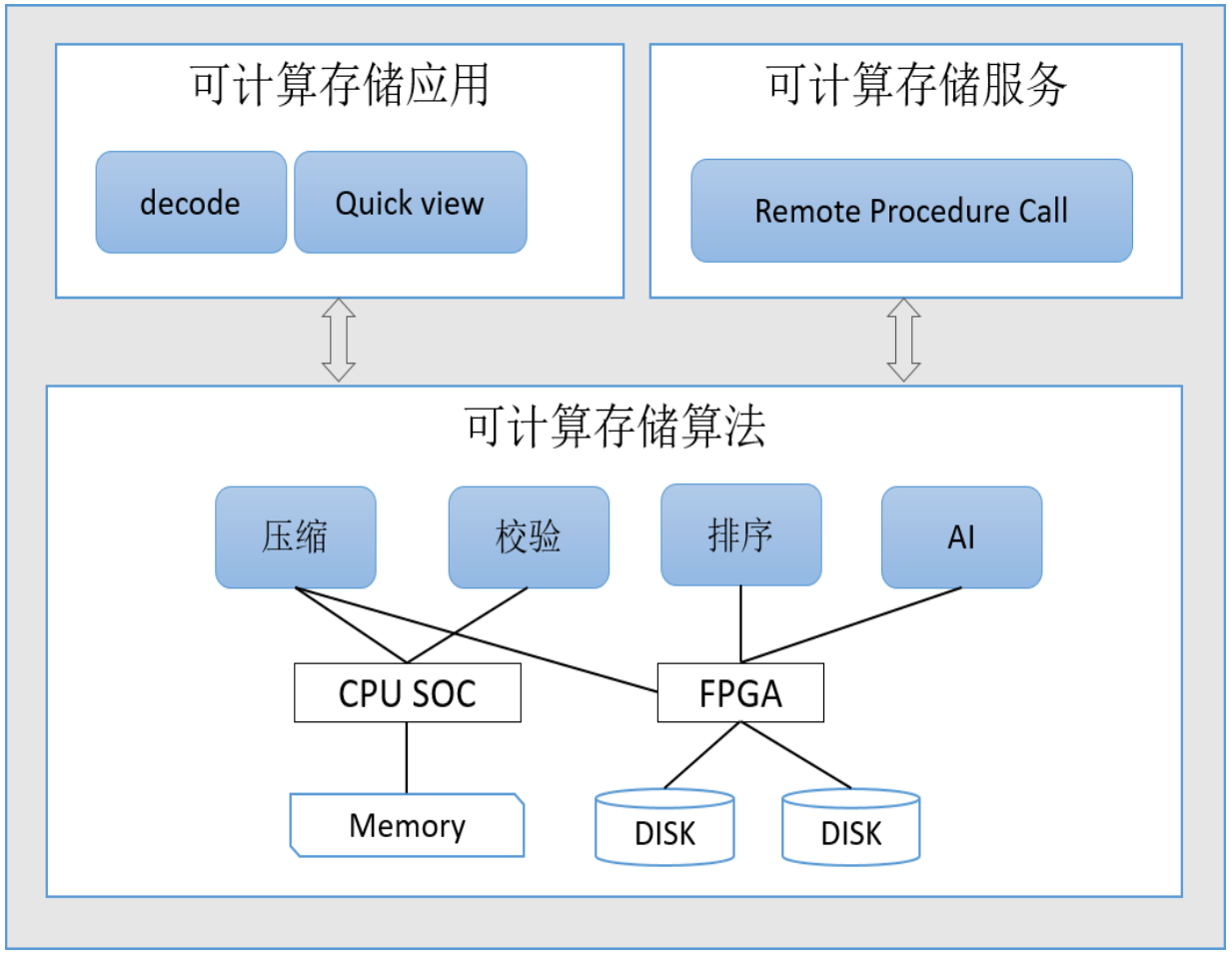
对于常用的基础算法，如压缩、校验、排序及各种AI常用算法等，以算子的形式提供服务

应用层次：

由若干子算法组成的数据处理任务，如数据解码任务，在可计算存储服务器本地完成

服务层次：

某些必须在计算集群进行复杂计算的任务，例如实验模拟、数据重建等，其中需要大量I/O操作的子任务，卸载到可计算存储服务器，即通过网络来调用可计算存储服务



调用方式示例

存储客户端：

兼容标准访问协议，通过增加参数调用，例如：

cat runid.txt&css_alg=sort；

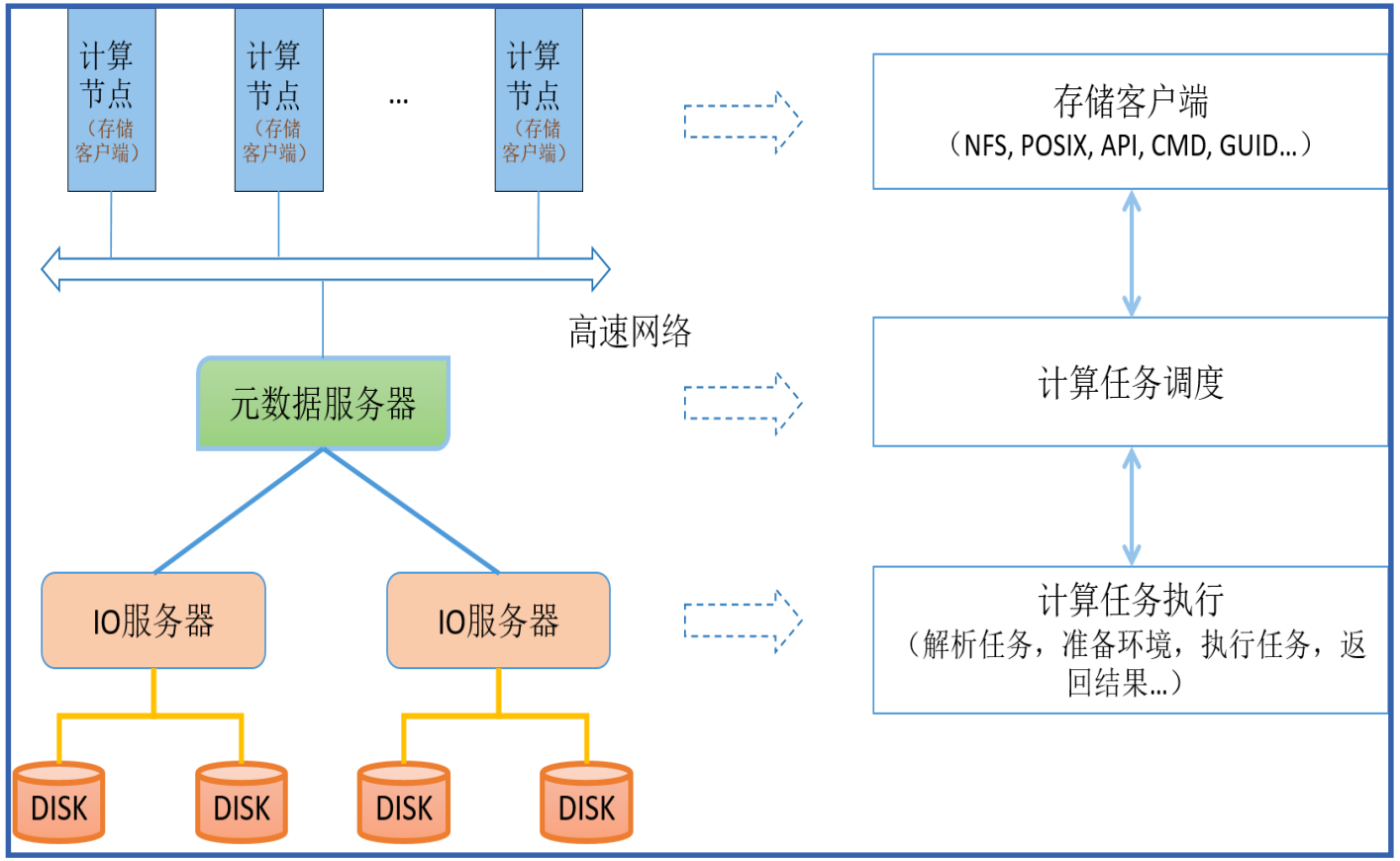
open(“run011.dat&css_app=decode”, O_RDONLY)

计算任务调度：

通过元数据查找物理文件位置

IO服务：

任务解析，执行等



总结

- “存算分离”的冯·诺依曼架构存在网络传输压力和“存储墙”问题
- 可计算存储作为一种“存算一体”技术，将部分计算任务卸载到数据存储部分，减少数据移动，提高计算效率
- 基于高能物理实验数据处理的现实问题，设计并实现了基于ARM CPU和FPGA的可计算存储服务器
- FPGA部分的性能有待提升，支持的算法、应用需要进一步扩展

敬请批评指正！

报告人：高 宇
email ：gaoyu94@ihep.ac.cn
导 师：程耀东 研究员

2022年8月11日