新一代交互式分析引擎hologres

金晓军

阿里云智能-计算平台事业部-交互式分析

极客时间 | 企业服务

这是一封正经“福利”贴

想做团队的领跑者 需要迈过这些“槛”

成长型企业，易忽视人才体系化培养 企业转型加快，团队能力又跟不上	VS	从基础到进阶，超100+一线实战 技术专家带你系统化学习成长
团队成员技能水平不一， 难以一“敌”百人需求	VS	解决从小白到资深技术人所遇到 80%的问题
寻求外部培训，奈何价更高且 集中式学习	VS	多样、灵活的学习方式，包括 音频、图文 和视频
学习效果难以统计，产生不良循环	VS	获取员工学习报告，查看学习 进度，形成闭环

课程顾问「橘子」

回复「QCon」
免费获取
学习解决方案

 # 极客时间企业账号 # 解决技术人成长路上的学习问题

自我介绍

大数据领域从业9年，曾担任阿里云数据平台架构师，从无到有设计研发Aliyun StreamCompute V1.0。

后担任网易数据科学中心大数据平台负责人，负责网易大数据平台建设、团队建设、人才培养，负责整体架构设计、自研系统研发与开源组件功能扩展与集成、大数据产品化输出。

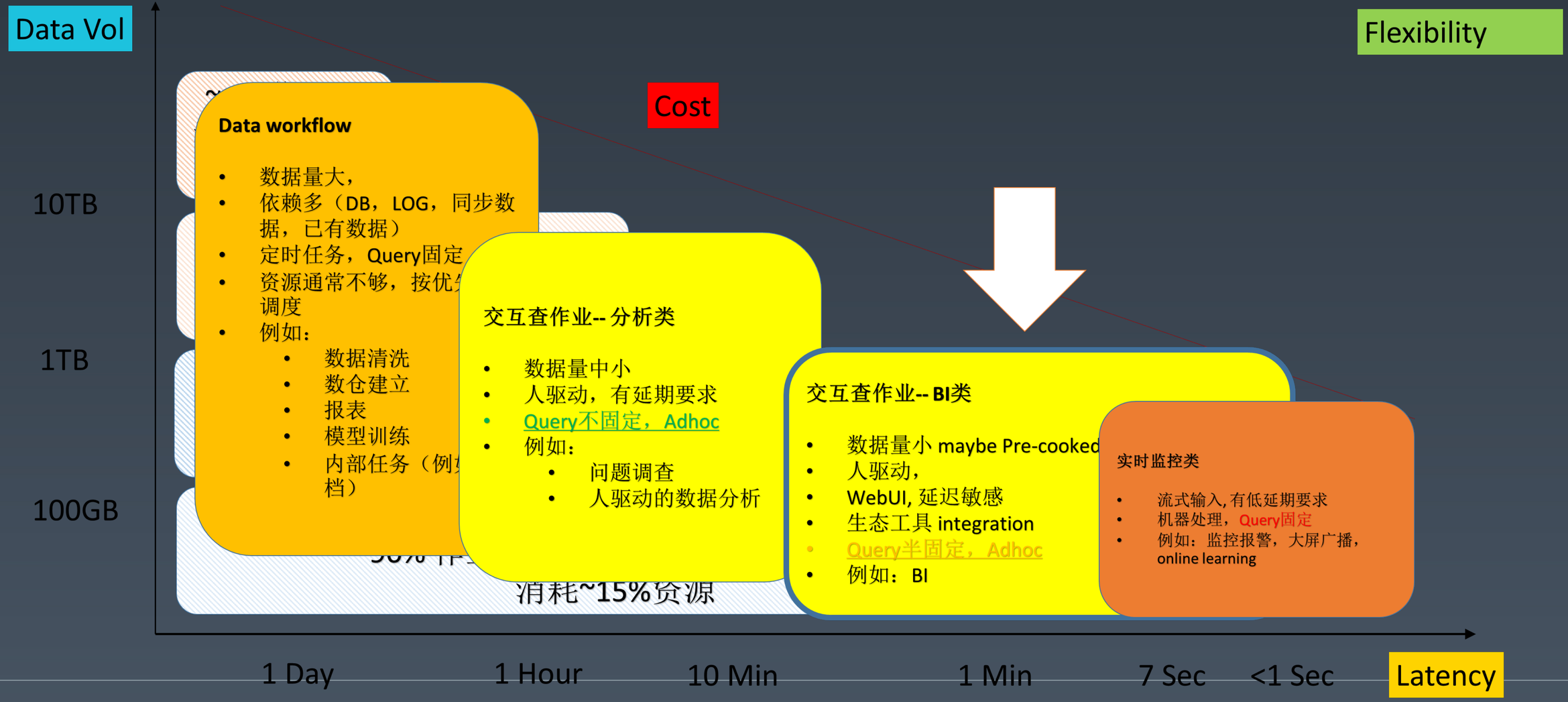
现就职于阿里巴巴计算平台事业部，从事交互式分析引擎hologres设计与研发工作。

目录

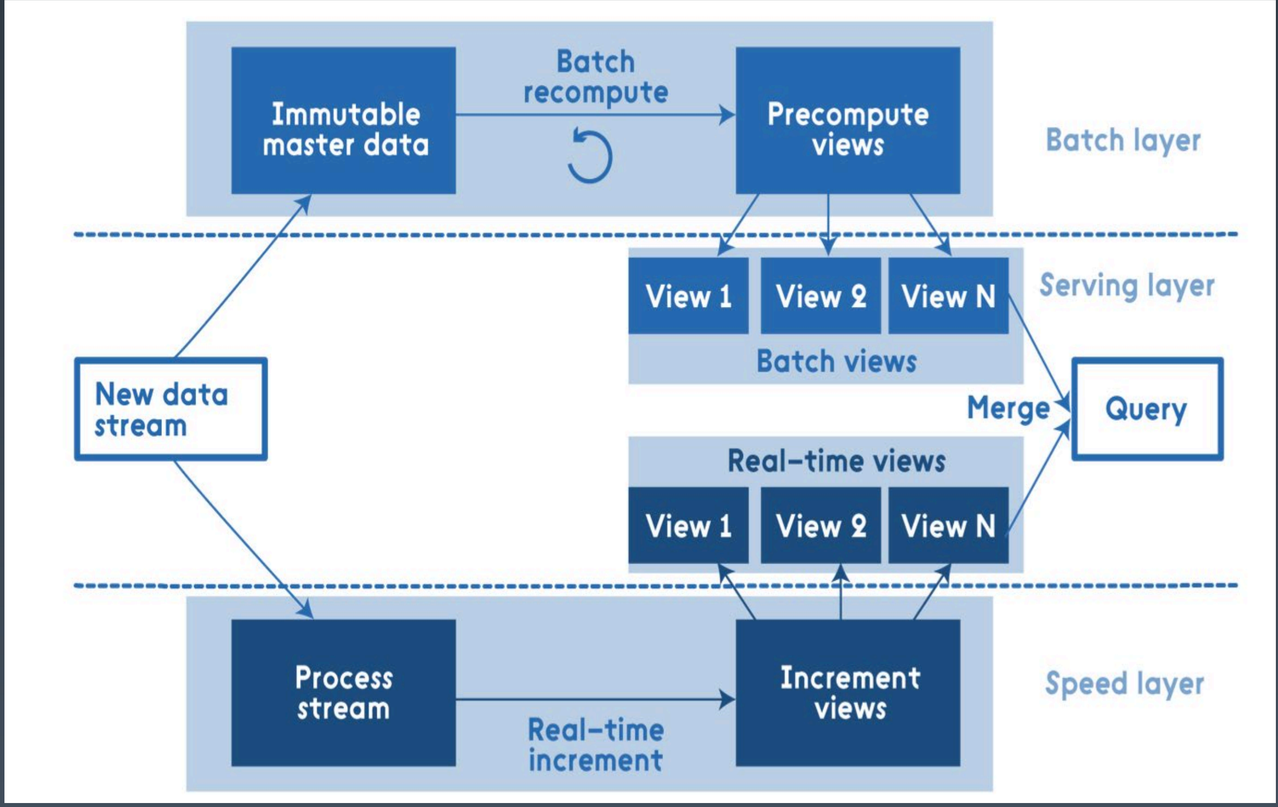
一. 背景介绍
二. 基础架构
三. 技术亮点
四. 典型场景介绍
五. 未来规划

背景介绍

背景:典型场景分析（业务需求，数据/计算量与资源消耗的分布）



背景：典型开源架构



Lambda架构的问题：

1. 使用多种引擎和系统去组合，开发和维护成本高，学习生成高
2. 数据在不同的View中存储多份，空间浪费，数据一致性的问题如何解决
3. 从使用上来说，Batch,Streaming及Query均使用不同的language，使用起来并不容易

背景：技术和业务背景

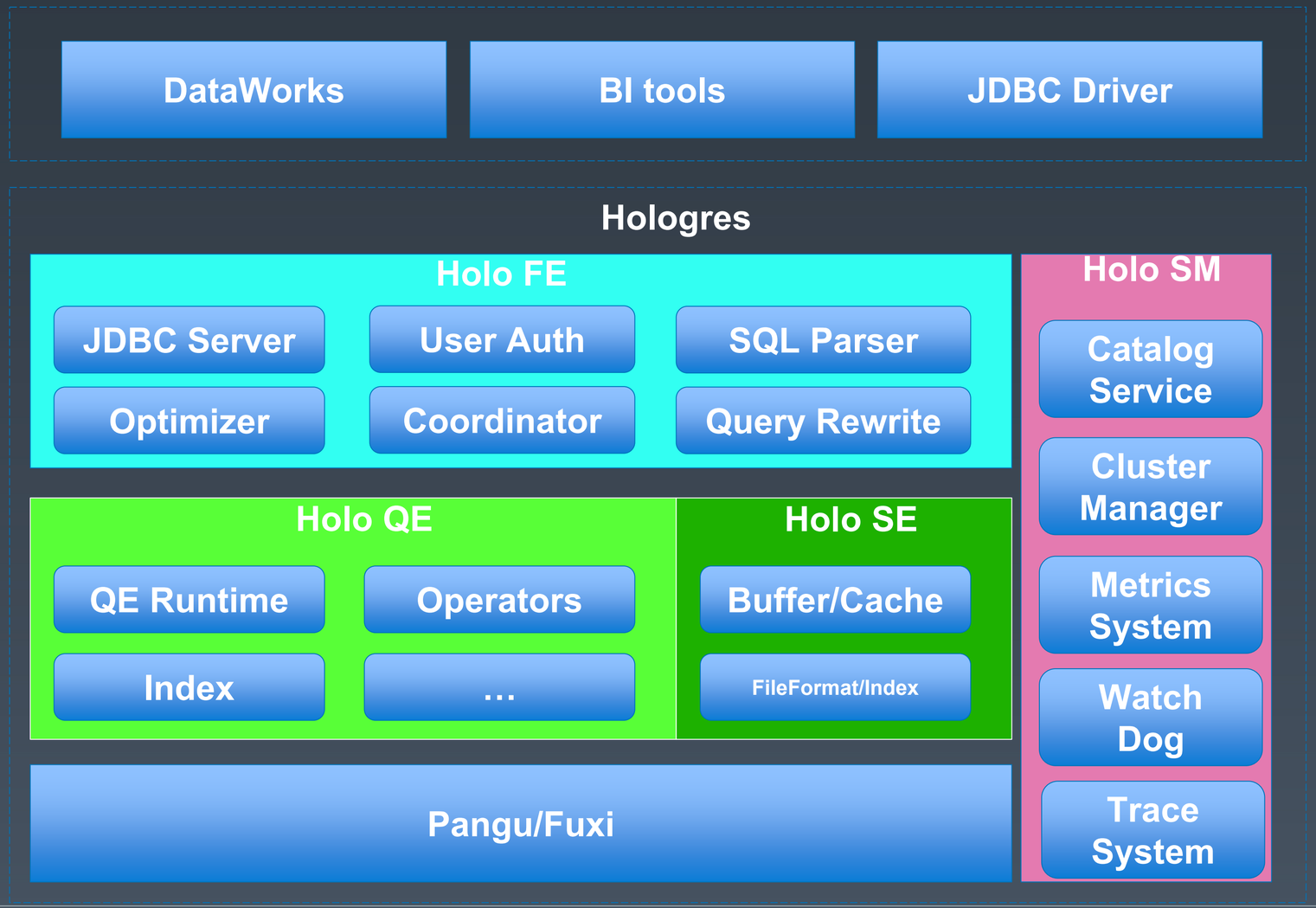
- 技术源于流计算+数据库技术，落地于搜索与广告业务
- 开源的解决方案无法满足阿里巴巴复杂的业务场景
- 实时数据中台建设的需求：一个入口，一份数据，一种查询语言
- 成本，易用性，实时数据中台架构的统一性
- Alibaba Blink(Flink内部版，已开源)创始人量仔老师牵头打造新一代交互式分析引擎

Hologres架构介绍

Hologres介绍

- 新一代海量数据交互式分析引擎
- 一套引擎支持Point Query(hbase场景)，Ad-hoc Query(Druid场景)，OLAP Query(Impala场景)
- 快
- 存储计算分离
- 支持实时数据与批量数据导入
- 支持External Storage，与阿里云大数据产品无缝对接

Hologres架构


DataWorks
BI tools
JDBC Driver
Hologres
Holo FE
JDBC Server
User Auth
SQL Parser
Optimizer
Coordinator
Query Rewrite
Holo QE
QE Runtime
Operators
Index
…
Holo SE
Buffer/Cache
FileFormat/Index
Pangu/Fuxi
Holo SM
Catalog Service
Cluster Manager
Metrics System
Watch Dog
Trace System

Storage Engine(SE)

- 存储计算分离的架构
- 内置存储引擎(行存，列存)
  - 行存：整行数据连续存放，更新高效，对point query和批量scan更友好(Hbase场景)
  - 列存：相关列的数据连续存放，按列做聚合更高效，压缩更高效，适合分析型场景
- External Table

Query Engine(QE)

- 自研QE(性能卓越)
  - 异步执行引擎
  - 向量化计算
  - 支持Filter/Agg计算的pushdown
- PostgreSQL QE(兼容生态)
  - 兼容PostgreSQL生态
  - 与生态合作开发

Frontend(FE)

- PostgreSQL协议及SQL语法的兼容
- 更加智能的优化器，提供Query Federation的能力
- 调度，流控，反压

Hologres技术亮点

hologres技术亮点 - 统一引擎架构

- Why？ 大数据业务Hbase中数据存一份，Druid里存一份，XXX里存一份
  - 浪费!
  - 数据一致性怎么保证?
  - 学习成本高，成天学习新系统的使用
- 功能
  - 内置支持两种存储格式，创建表的时候选其一或者都选，数据一致有保证
  - QE提供两个版本，自研和开源
  - 能够替换现有业务的Hbase，Druid和impala，且性能更好
    - 阿里巴巴业务已得到验证
    - 团队十多名Flink commiter，两名Hbase PMC，多名Hbase/Druid/Kylin等开源系统commiter

hologres技术亮点 - 存储计算分离

- Why ?
  - 用户只关心自己有多少计算资源，根本不关心自己的机器是什么
  - 已经申请的计算资源可否利用，如ODPS/Blink
  - 新的NVME SSD盘可以达到150000IOPS，磁盘IO不再是性能瓶颈，问题转变为如何把CPU高效利用起来
  - 存储计算分离是未来大势所趋，存储和计算非对齐采购，成本更低，部署运维更方便
- 功能
  - 存储使用Pangu 2.0，由存储团队维护，QE和SE可运行在K8S及飞天集群中
  - 全异步的存储和计算引擎，吃尽所有CPU计算能力
  - 灵活扩容，缺存储扩存储，缺计算扩计算

hologres技术亮点 - 更加聪明的Optimizer

- Why ?
  - 用户写好Query如何去调优?
  - 一套引擎中支持多套QE，查询计划如何去生成?
  - 多种文件格式，不同版本的operator多种实现方案，如何去选择?
  - 如何更高效的去生成上述查询计划?
- 功能
  - 支持多引擎的查询优化器，能够很容易与各种QE结合
  - 基于代价的优化器模型，支持各种index，predicted pushdown

hologres技术亮点 - 新技术

- Why ?
  - 近几年硬件性能提升的很快，N年前的技术方案不一定能够很好的利用现在的硬件性能发挥到极致
  - 技术追求，没有最好，只有更好
- 功能
  - 全异步框架(Thread-per-core架构)，把CPU利用到极致
  - vectorization(细节很多坑)，集团内大规模使用向量化计算技术加速计算(1个量级)
  - 各种Index的实现
  - 精细化的Cache

技术亮点举例------ 为什么要用全异步架构?

- 传统存
- 最新硬
- Open大

I/O Is Faster Than the CPU – Let's Partition Resources and Eliminate (Most) OS Abstractions

Pekka Enberg
University of Helsinki

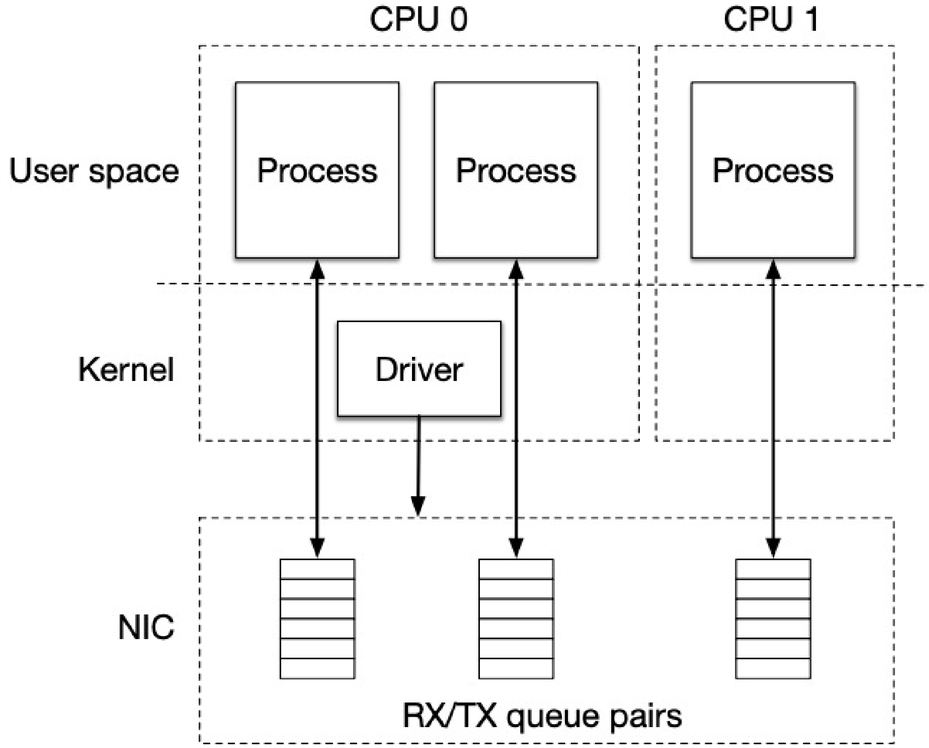
Ashwin Rao
University of Helsinki

Sasu Tarkoma
University of Helsinki

Abstract

I/O is getting faster in servers that have fast programmable NICs and non-volatile main memory operating close to the speed of DRAM, but single-threaded CPU speeds have stagnated. Applications cannot take advantage of modern hardware capabilities when using interfaces built around abstractions that assume I/O to be slow. We therefore propose a structure for an OS called *parakernel*, which eliminates most OS abstractions and provides interfaces for applications to leverage the full potential of the underlying hardware. The parakernel facilitates application-level parallelism by securely partitioning the resources and multiplexing only those resources that are not partitioned.

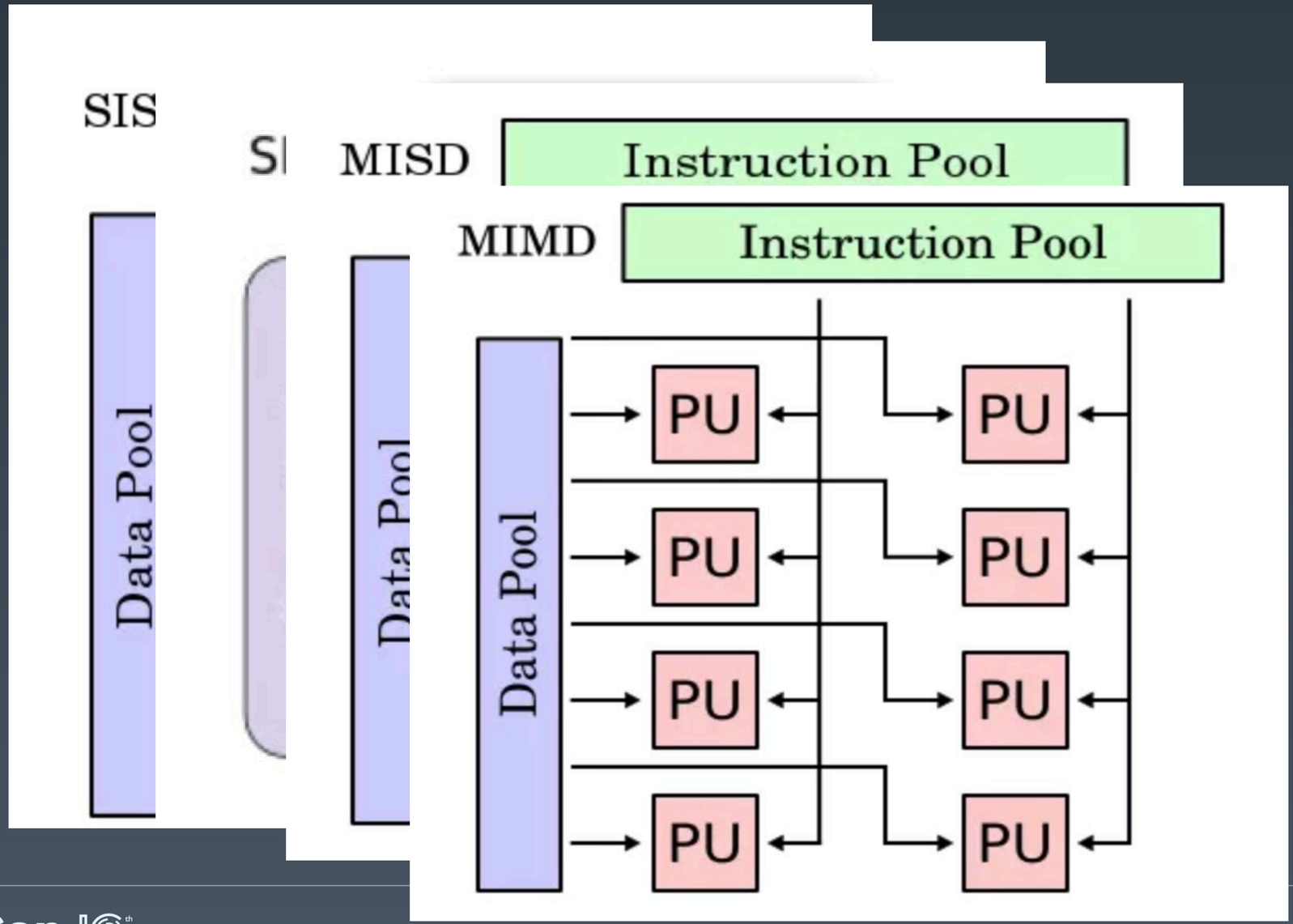
ACM Reference Format:
Pekka Enberg, Ashwin Rao, and Sasu Tarkoma. 2019. I/O Is Faster Than the CPU – Let's Partition Resources and Eliminate (Most)



技术亮点举例------ 全异步架构实现有何难点

- 传染性，系统整体执行流程需要全部异常化编程
- 更加极致的利用cpu？cpu调度,thread-per-core
- 代码中不能有blocker，tracing，debugging

技术亮点举例------ 向量化计算



技术亮点举例------ 向量化计算

- Flynn分类法: SISD, SIMD,MISD,MIMD
- SIMD：intel MMX->SSE->AVX
- 并非新技术，但对于大数据处理却非常有用
- 如何更多的实现向量化版本的function?
- 重剑无锋，大巧不工。细节

技术亮点举例------ 优化器

- operater实现可能同时存在行存版本，列存版本，向量化版本
- query如何高效的执行，如何去选择不同的operater实现
- 如何去支持各种 index
- 如何去支持多种QE
- 重剑无锋，大巧不工

典型业务场景介绍

服务场景 - 海量数据复杂查询

- 用户需求
  - 大数据复杂准实时分析 T+1 （亿级别）
  - 对latency敏感 但可以接受资源消耗的成本
  - 查询复杂， 需要支持完善的SQL语义(join/distinct/topk/window) 以及方便的接入协议（jdbc）
- 优势
  - 完备的SQL支持
  - 支持实时和批量导入，性能远超同类产品
  - 便捷性+性能
  - 与MaxCompute共享资源(计算/存储)， 错峰调度

服务场景 – 海量数据点查询(Hbase)

- 用户需求
  - 海量数据 PB级别存储+Billion级记录
  - 高频写入，高频查询，计算简单
  - 典型客户， 搜索广告，集团安全部， 支付宝风控 （平台型用户）
- 现有方案 （Hbase)
  - 导入任务难以维护 + 浪费存储 + 导入性能极低 （5-8个小时）
  - 无SQL接口
  - 海量存储成本极高
- 优势
  - 统一存储， 无需导入操作
  - 提供SQL接口， 方便开发集成

服务场景 - 小表直读 (RDS)

- 用户需求
  - 需要完备SQL，并支持JDBC以及开源BI工具， 方便开发
  - 报表展现，Latency敏感，数据量可以控制到非常小，如百万级别
- 当前解决方案
  - MaxCompute做好处理， 产出报表需要的结果表
  - 将结果表通过datax/dts导入到rds， 后续通过rds查询
- 问题
  - 维护多套服务的同步任务太复杂， 且数据时效性一致性不好保证， 浪费存储
- 阿里云 odps.pop 日均查询 2w+
- 阿里体育 直接通过PHP接入 对对业务侵入小 云账号打通 数据安全

未来规划

产品规划 - 阿里云交互

一. Hologres是alibaba Blink创始人对
队集结了众多在分布式存储计算深
本的交互式分析产品。
二. Hologres从诞生到现在，已经在集
2019年6月将正式登陆阿里云，为
三. 在此次Qcon会上首次对外公开Ho
术细节
四. https://www.aliyun.com/produ

品，团
能低成

。在

多的技

THANKS!

QCon 10th