

高能物理计算技术

第三届（2022年）高能物理计算暑期学校

石京燕

高能所计算中心

shijy@ihep.ac.cn

提纲

1

高能物理离线数据处理过程与特点

2

集群计算，网格计算

3

高通量计算与高性能计算

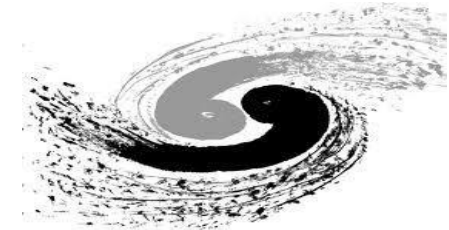
4

分布式计算

5

总结

高能物理实验过程



实验
装置

在线数据
实时获取

传到离线，
长期保存



在线
取数



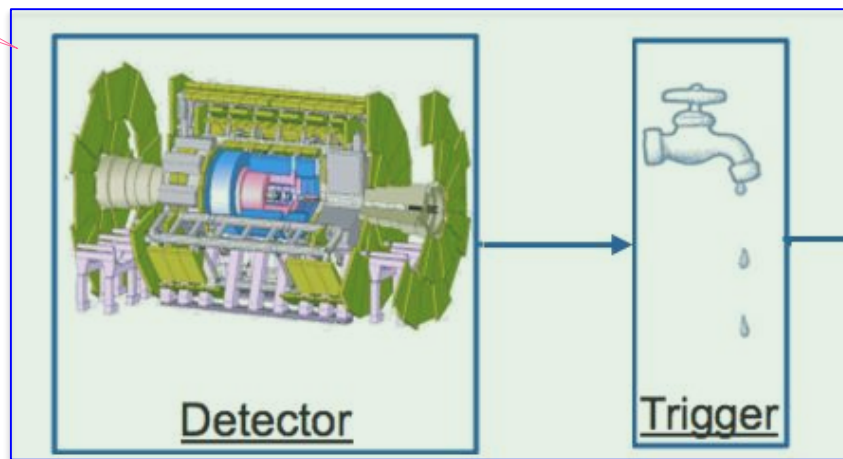
磁带库

- 高能物理实验装置运行
 - 在线获取探测器实验数据
 - 实验数据被长久保存
 - 实验数据被离线分析处理
- 物理成果

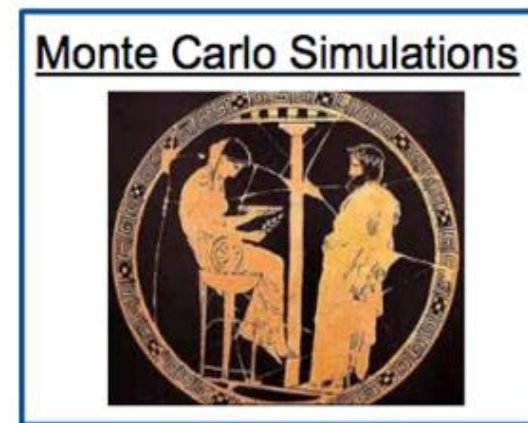
高能物理实验数据处理过程



在线
取数

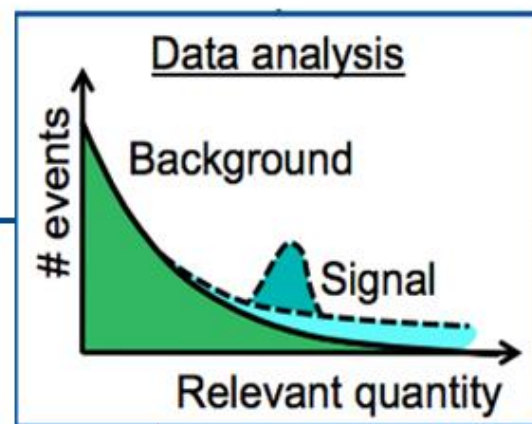
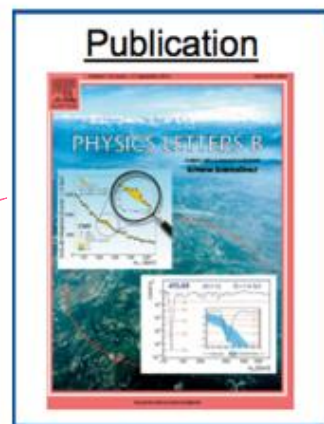


原始
数据

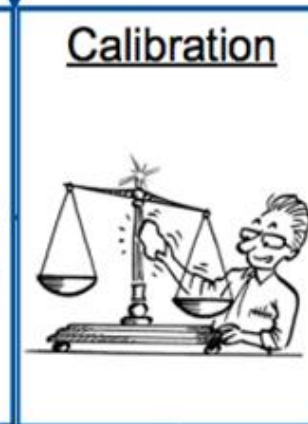
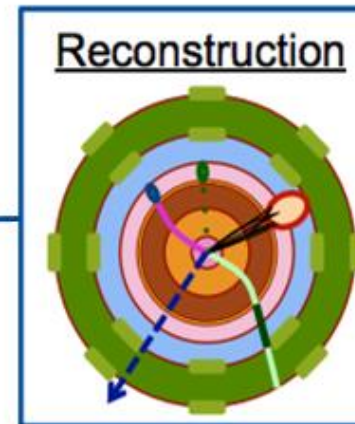


离线模拟
→ 模拟数据

物理
成果



不断反
复分析



刻度+重建
→ 重建数据

高能物理离线计算的特点及过程

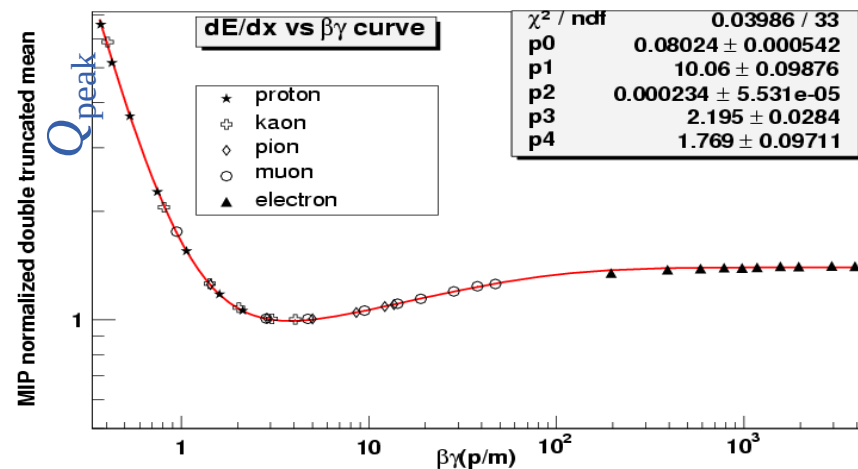
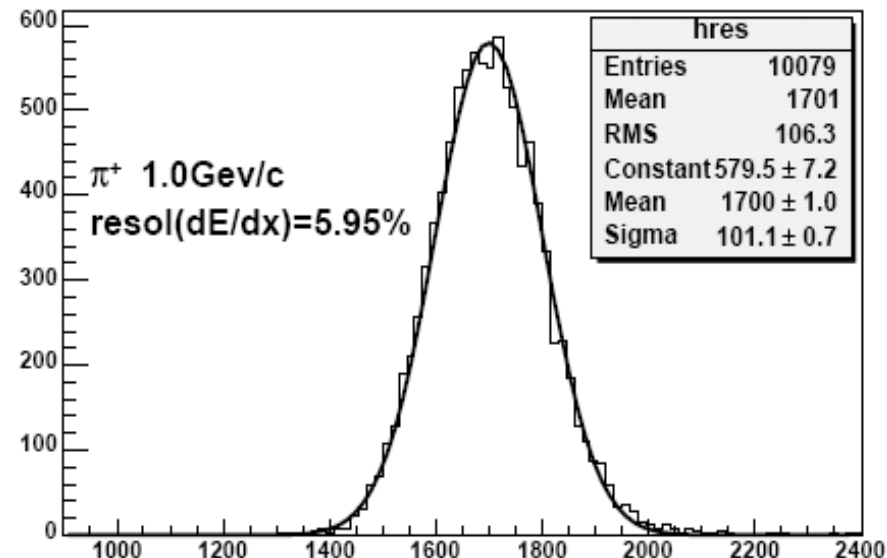


● 特点:

- 实时性不高，是实验的一项长期工作
- 计算过程被不断调整，优化
- 分析计算得到的数据及结果可重现

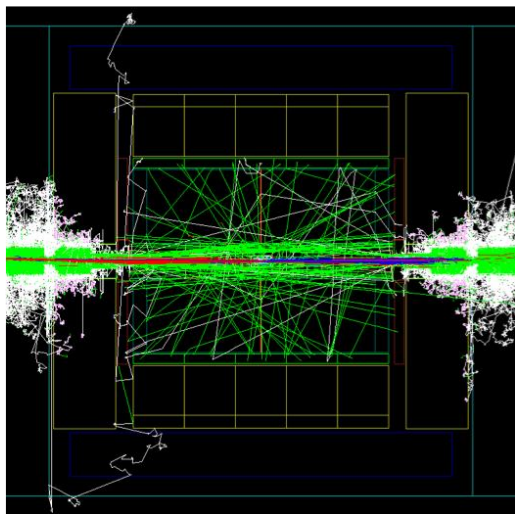
● 过程:

- 刻度 (Calibration) : 将原始数据转换为物理量
- 校准 (Alignment) : 定位探测器精度
- 事例模拟:
 - 模拟与真实数据最相近的人造事例
 - 需要背景噪声研究, 修正, 错误估算
- 事例重建:
 - 重建粒子轨迹和顶点
 - 标注粒子类型和衰变
 - 加入物理约束 (能量动量保留)
- 物理分析:
 - 从背景数据中找到物理信号、计算截面、分支比等
 - 需要大量培训, 具备高超物理数据分析技术



物理模拟

- 利用计算机来模拟物理过程-事例产生
 - 物理过程是随机过程
 - 事例的产生
 - 粒子在介质中的输运
- 信号：需要研究的物理过程
- 本底：除信号以外的其他物理过程以及探测器的噪声



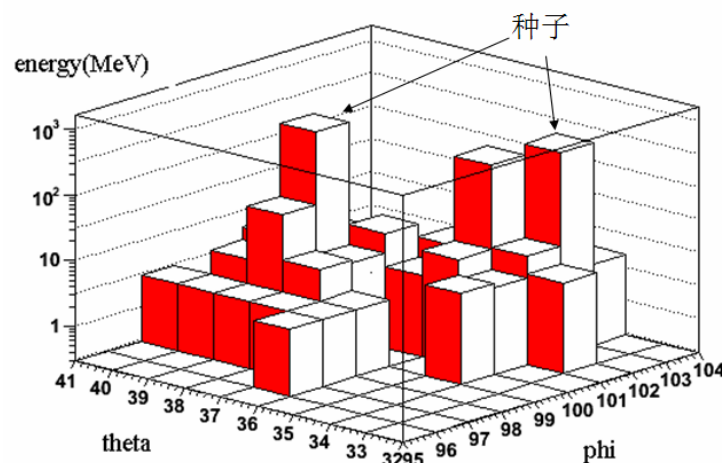
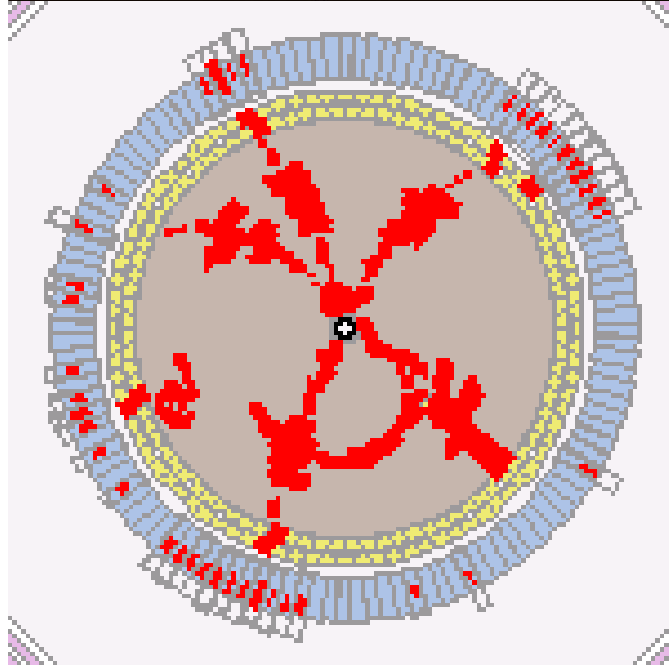
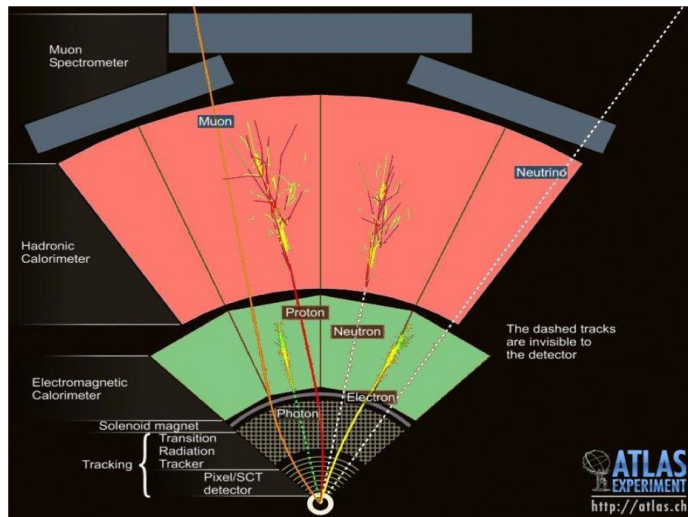
- 重要的模拟软件：GEANT4
- 基于'已知'物理理论模拟粒子在介质中的传输过程
 - 粒子衰变、光电效应、康普顿散射、粒子对产生、粒子输运过程中的能量损失、韧致辐射、强子相互作用

事例刻度与重建

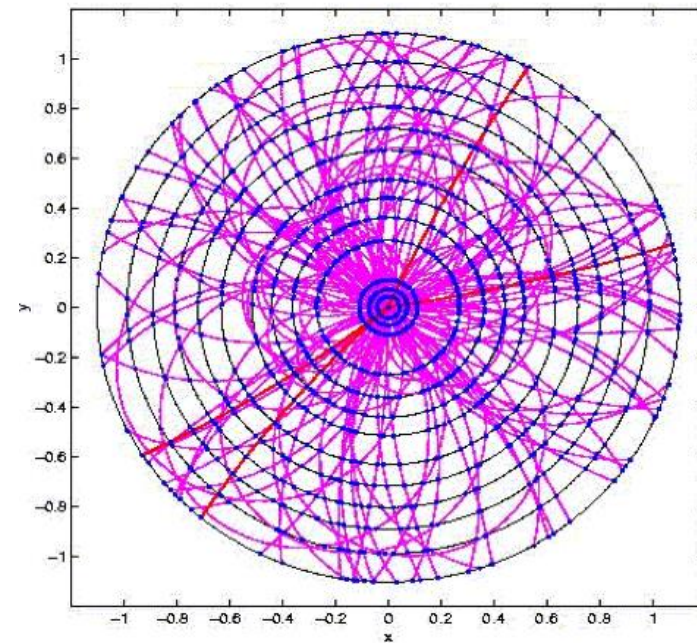


● 还原事例的基本物理信息

- 径迹重建
- 簇射簇重建
- 粒子鉴别
- 顶点重建
- ○ ○ ○ ○ ○



第三届 (2022年) 高能物理计算暑期学校



物理分析

- 物理动机
- MC模拟
- 程序编写
- 程序调试
- 初步分析
- 深入分析
- 输入输出检查
- 系统误差

初步想法

B. For $\psi' \rightarrow \pi^+\pi^-J/\psi, J/\psi \rightarrow l^+l^-$ process

- Track level cuts
 - $|V_z| < 10\text{cm}$
 - $|V_r| < 1\text{cm}$
 - $|\cos\theta| < 0.80$
 - $|\vec{p}| < 2.0\text{GeV}/c$
- nGood ≥ 4
- The charged tracks with $|\vec{p}| < 0.45\text{GeV}/c$ are assumed to be pion, select the $\pi^+\pi^-$ pair candidates by minimizing $|M_{\pi^+\pi^-}^{\text{rec}} - M_{J/\psi}|$
- $\cos\theta_{\pi^+\pi^-} < 0.95$
- $3.05\text{GeV}/c^2 \leq M_{\pi^+\pi^-}^{\text{rec}} \leq 3.15\text{GeV}/c^2$
- Take the two fastest positive and negative tracks as lepton candidates, identify the e/μ pair
 - $\mu^+\mu^-: [E/p]^+ < 0.26$ and $[E/p]^- < 0.26$
 - $e^+e^-: [E/p]^+ > 0.80$ or $[E/p]^- > 0.80$ or $\sqrt{([E/p]^+ - 1)^2 + ([E/p]^- - 1)^2} < 0.4$
- $\cos\theta_{l+l-} < -0.95$ in lab frame.
- $2.7\text{GeV}/c^2 < m_{l+l-} < 3.2\text{GeV}/c^2$ for $\pi^+\pi^-e^+e^-$ channel and $3.0\text{GeV}/c^2 < m_{l+l-} < 3.2\text{GeV}/c^2$ for $\pi^+\pi^-\mu^+\mu^-$ channel

程序实现

```
// With momentum method calculate the invariant mass of Jpsi
// actually we use the recoil mass
HepLorentzVector m_lv_recoil, m_lv_jpsi;
m_lv_recoil = m_lv_lab - m_lv_pionp - m_lv_pionm;
m_lv_jpsi = m_lv_lepp + m_lv_lepm;

m_mass_twopi = (m_lv_pionp + m_lv_pionm).m();
m_mass_recoil = m_lv_recoil.m();
m_mass_jpsi = m_lv_jpsi.m();

// Jpsi mass cut
if( m_mass_recoil < 3.05 || m_mass_recoil > 3.15 ) return sc;
if( m_mass_jpsi < 3.0 || m_mass_jpsi > 3.2 ) return sc;
m_cout_recoil++;

HepLorentzVector m_ttm(m_lv_jpsi + m_lv_pionp + m_lv_pionm);
if(m_ttm.m() > 4 || m_ttm.m() < 3) return sc;

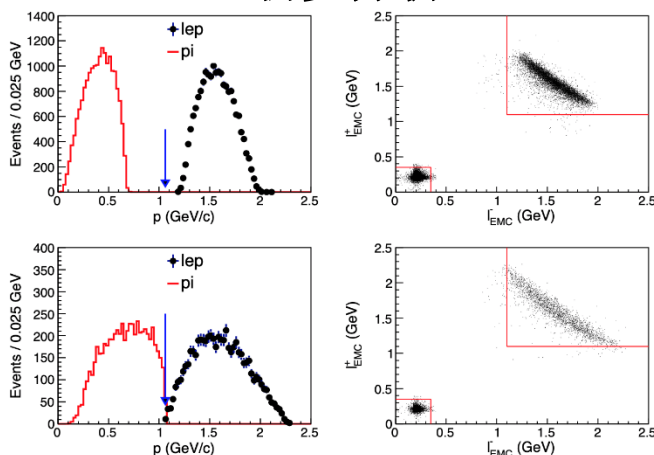
// dangle between pions, suppress gamma conversion
m_pi_pi_dang = m_lv_pionp.vect().cosTheta(m_lv_pionm.vect());

m_mom_pionp = m_lv_pionp.vect().mag();
m_mom_pionm = m_lv_pionm.vect().mag();
m_mom_lepp = m_lv_lepp.vect().mag();
m_mom_lepm = m_lv_lepm.vect().mag();
m_trans_ratio_lepp = m_lv_lepp.vect().perp().mag()/m_lv_lepp.vect().mag();
m_trans_ratio_lepm = m_lv_lepm.vect().perp().mag()/m_lv_lepm.vect().mag();
m_trans_ratio_pionp = m_lv_pionp.vect().perp().mag()/m_lv_pionp.vect().mag();
m_trans_ratio_pionm = m_lv_pionm.vect().perp().mag()/m_lv_pionm.vect().mag();

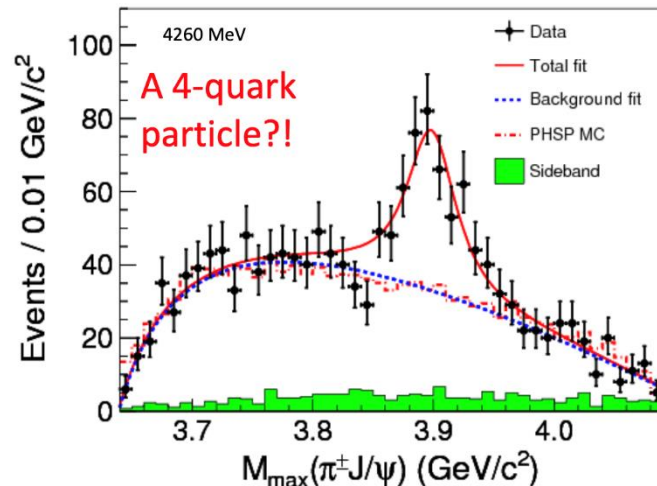
Hep3Vector m_boost_jpsi(m_lv_recoil.boostVector());
HepLorentzVector m_lv_cms_lepp(boostOf(m_lv_lepp, -m_boost_jpsi));
HepLorentzVector m_lv_cms_lepm(boostOf(m_lv_lepm, -m_boost_jpsi));
m_cms_lepp = m_lv_cms_lepp.vect().mag();
m_cms_lepm = m_lv_cms_lepm.vect().mag();
log << MSG::DEBUG << "Jpsi four momentum in cms " << m_lv_cms_lepp + m_lv_cms_lepm << endl;

m_inv_mass = m_ttm.m();
m_tot_e = m_ttm.e();
m_tot_px = m_ttm.px();
```

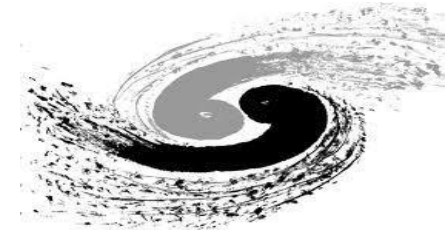
初步分析



深入分析



提纲



1

高能物理离线数据处理过程与特点

2

集群计算，网格计算

3

高通量计算与高性能计算

4

分布式计算

5

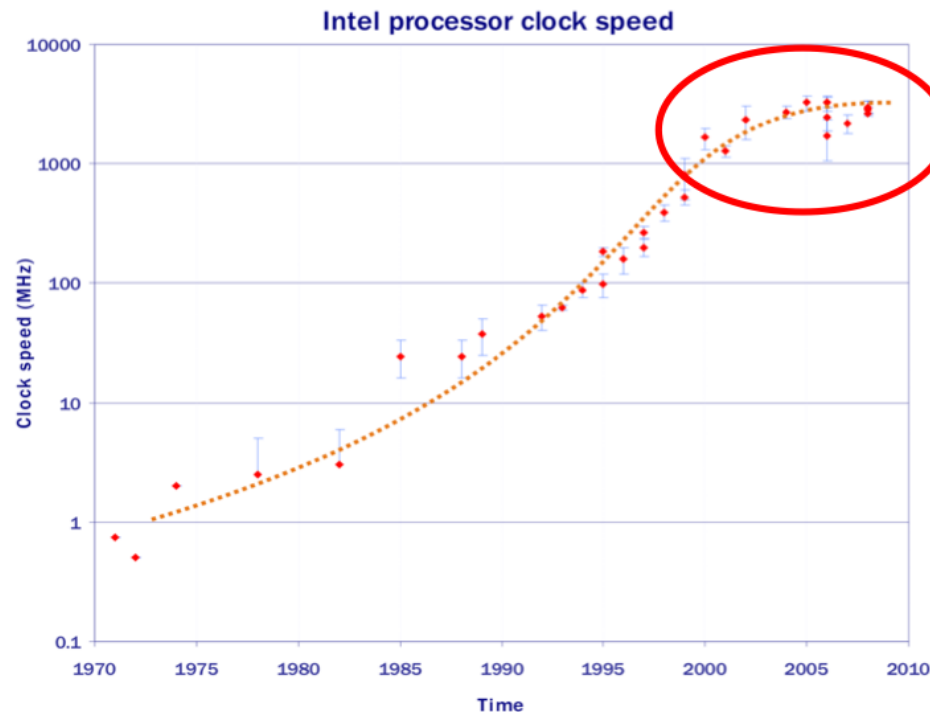
总结

计算技术的演变

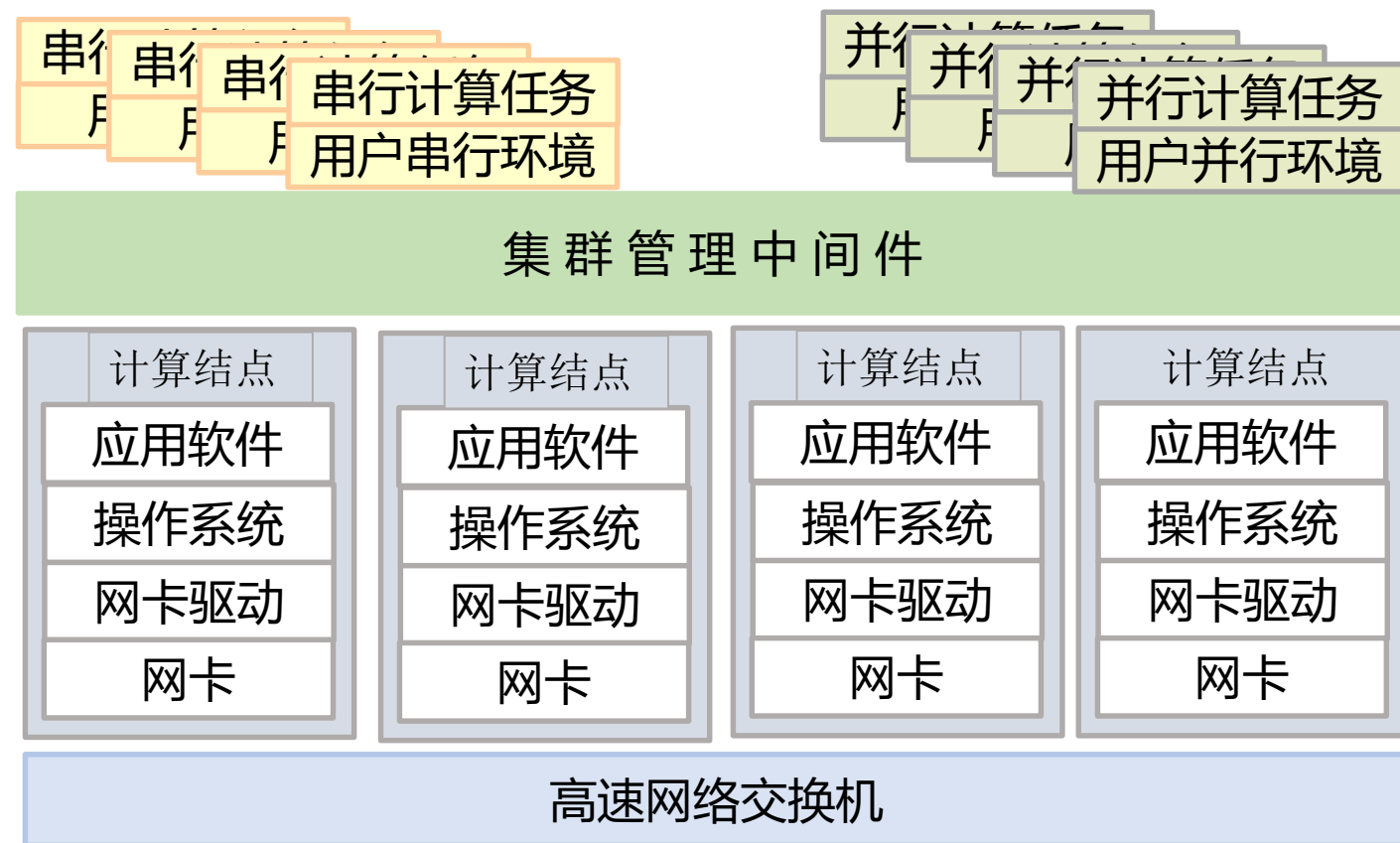


- 过去40年：
 - 芯片性能增长一直符合摩尔定律：每个芯片的晶体管数**每两年**增加一倍
 - 但是计算能力并没有增长这么快：缓存, 内存, 网络...
 - 新问题：更高的能耗和制冷
- 科学计算的技术演变
 - 20年前：本地计算集群（今天仍是主流!）
 - 15年前：广域网上的分布式计算
 - 不断发展出的新技术：虚拟化计算，志愿计算，GPU加速，AI数据分析，量子计算...

The free ride is over

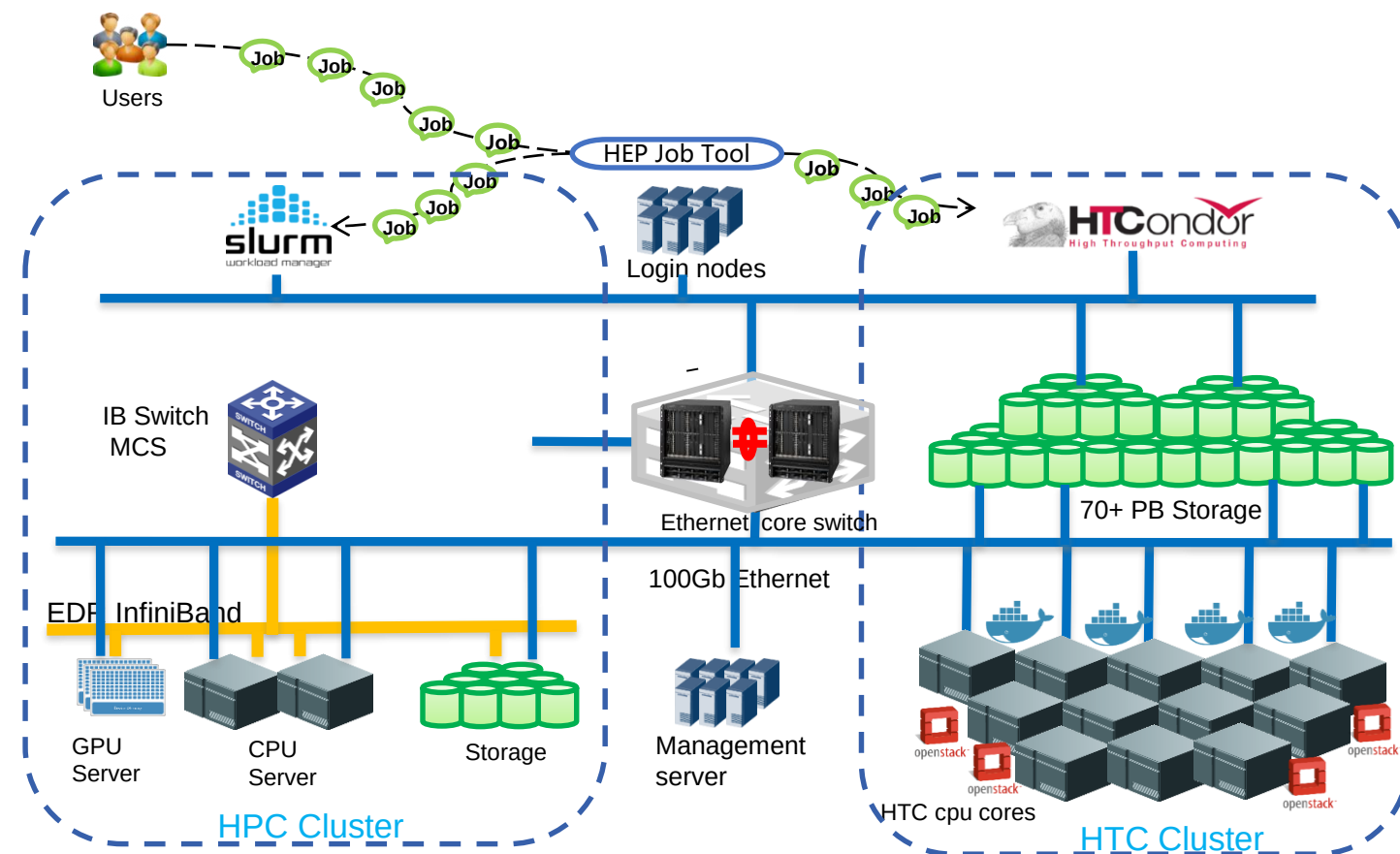


集群计算（本地计算）



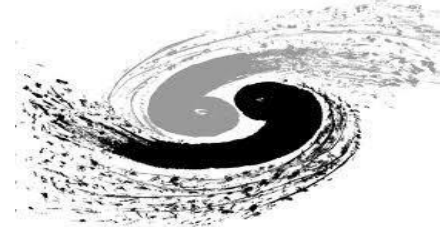
- 集群：将一组松散的同构或异构计算机的软硬件通过局域网连接起来紧密协同工作
 - 多用户：提供众多用户同时使用，用户之间无关联
 - 物理位置一致：计算机一般位于同一计算站点
 - 局域网连接：网络延时较低
 - 用户采用非交互的作业形式运行计算任务
 - 集群管理中间件：为用户的计算作业分配合适的计算资源（作业调度与资源分配），负责照管每个作业的完整生命周期
 - 一般具有统一用户名空间和存储空间

高能所计算集群



- 登录结点
 - 用户交互用服务器—coding, debugging
- 调度服务器
 - HTCondor, Slurm
- 计算结点
 - CPU, GPU
- 存储空间
 - 用户数据共享存储空间

网格计算



● 什么是网格？

- 网格技术将**全球地理上分散**的计算资源有机的整合起来，协同工作，为大型科学实验研究提供计算支持，能够完成单个集群无法完成的大规模计算任务。

● 网格技术的主要体现

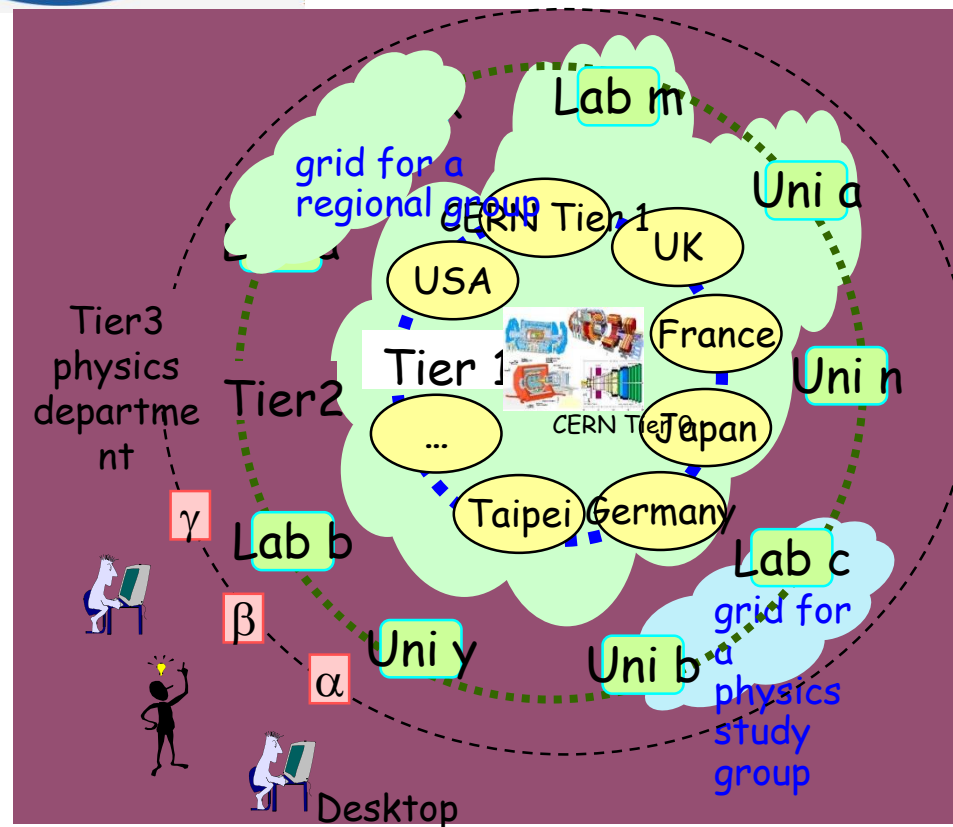
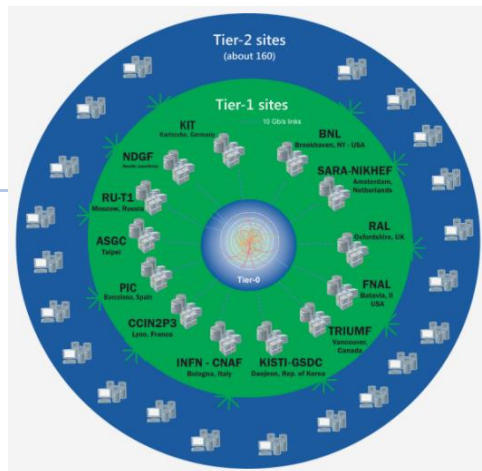
- 网格是一台超级计算机，拥有大量的计算资源与存储资源，用户可以透明使用
- 通过网格，最大程度的实现任务调度与资源共享，在全球范围内合理的分配资源
- 在共享资源的同时最大程度的保证网格资源的安全性。

● 网格的现状

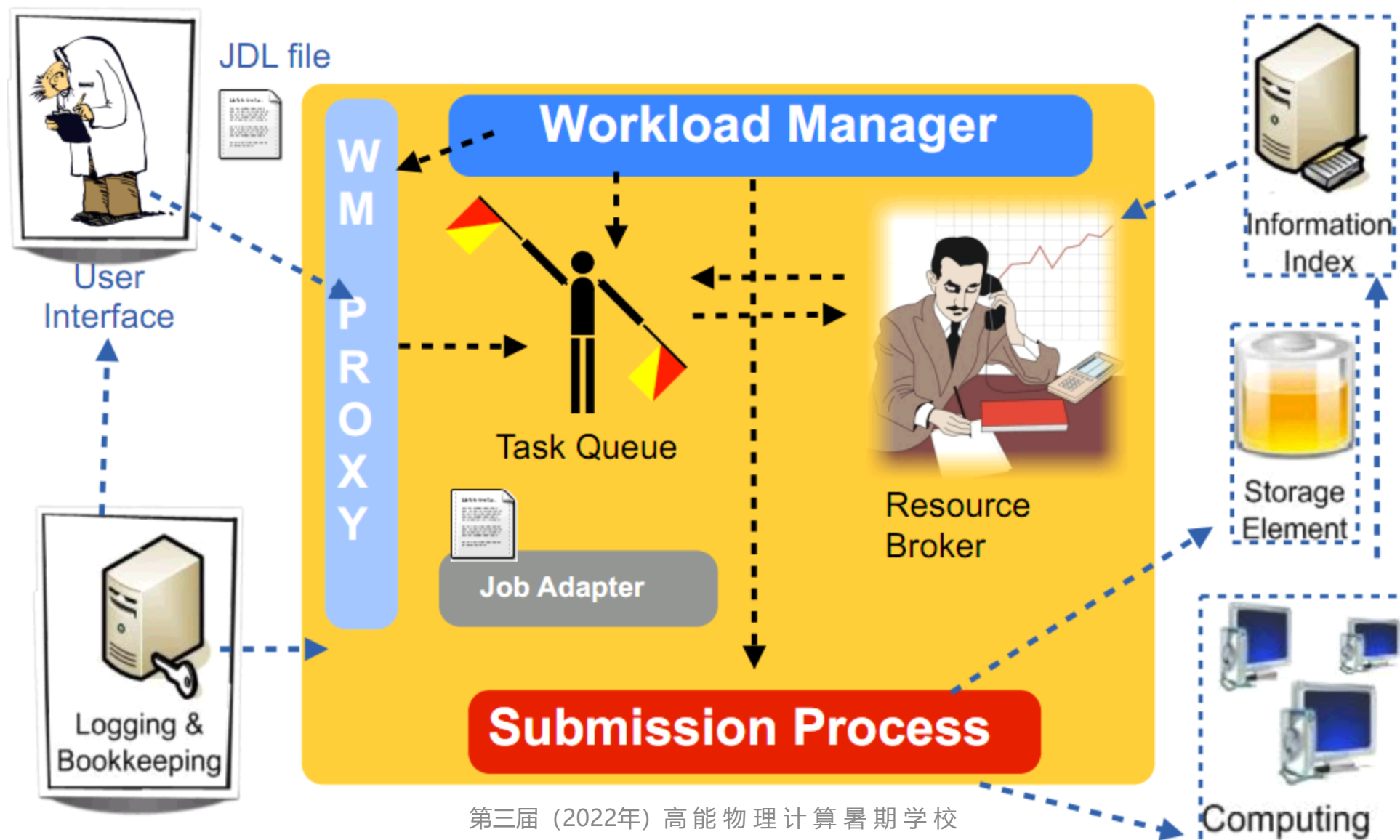
- 网格技术是成熟技术，得到广泛应用。世界各国或者国际间合作启动了多个网格项目。网格已成为高能物理、生物医学等领域科学家日常使用的计算基础设施，发挥了极其重要的作用

LHC 网格现状

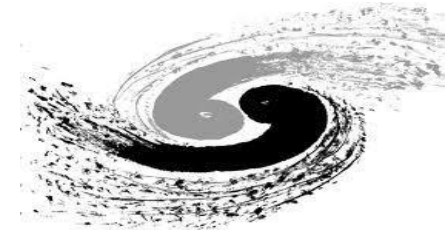
- 世界最大的网格项目
- 42个国家，170个计算中心
- 1M CPU 核每天运行2M计算任务
- 1EB的数据存储
- 分层
 - Tier 0: Cern 数据中心
 - Custodial 存储
 - First-pass 重建
 - Tier 1: 13个大型计算中心
 - Custodial 存储
 - Reprocessing
 - Tier 2: 遍布全球的160个计算中心
 - 事例模拟
 - 最终用户（物理学家）分析
 - Tier 3: 小型计算中心
 - 小型计算设施



网格重要组件



提纲



1

高能物理离线数据处理过程与特点

2

集群计算，网格计算

3

高通量计算与高性能计算

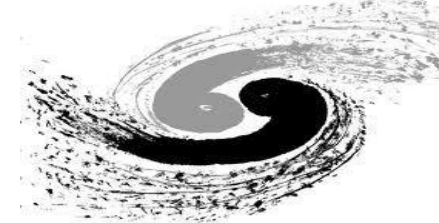
4

分布式计算

5

总结

高通量计算 vs. 高性能计算



- 高通量计算: **High Throughput Computing**

- 同时在多个处理器上运行大量相互独立的软件实例 (作业)
- 长时间 (一年多年)
- 大规模稳定的计算 -- 快速吞吐

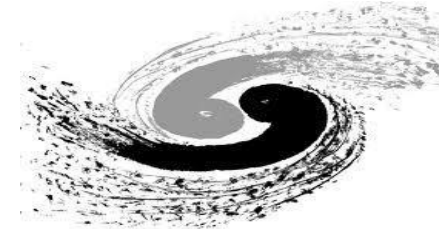


- 高性能计算: **High Performance Computing**

- 同时在多个处理器上运行大并行软件
- 作业数量少, 但是作业对CPU能力要求高
- 短时间需求

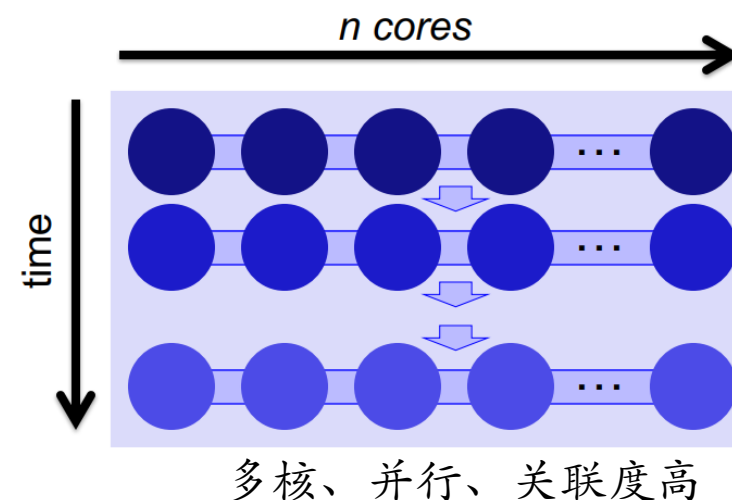
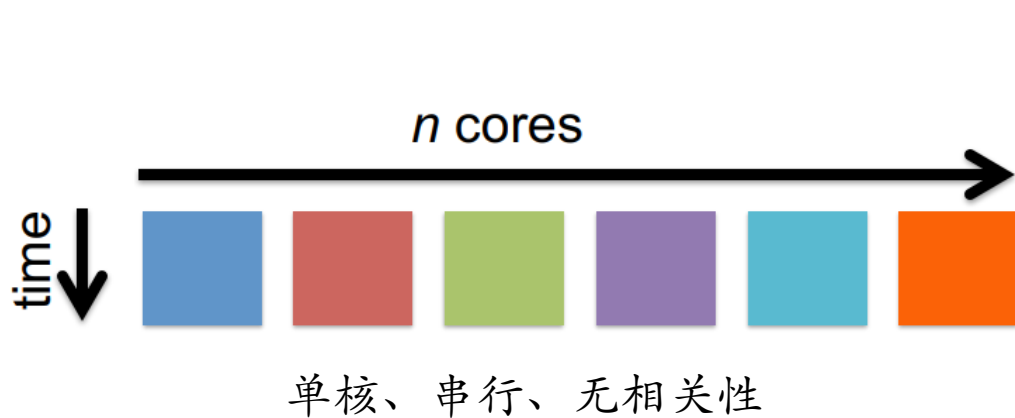


高通量计算 vs. 高性能计算



- HTC: 在一个较长时间提供大量计算资源。HTC不关注每秒完成的操作,而是每月、每年完成的操作。其目标是在一个很长的周期内完成的尽可能多的作业量,而非可以多快时间内完成时间单个作业。

	作业槽数/ 作业量	单作业进程 并行度	作业槽异构/ 同构	文件系统	节点网络速度
HTC	多	小	异构	共享文件系统 /数据传输	不要求
HPC	少	多	同构	共享文件系统	快

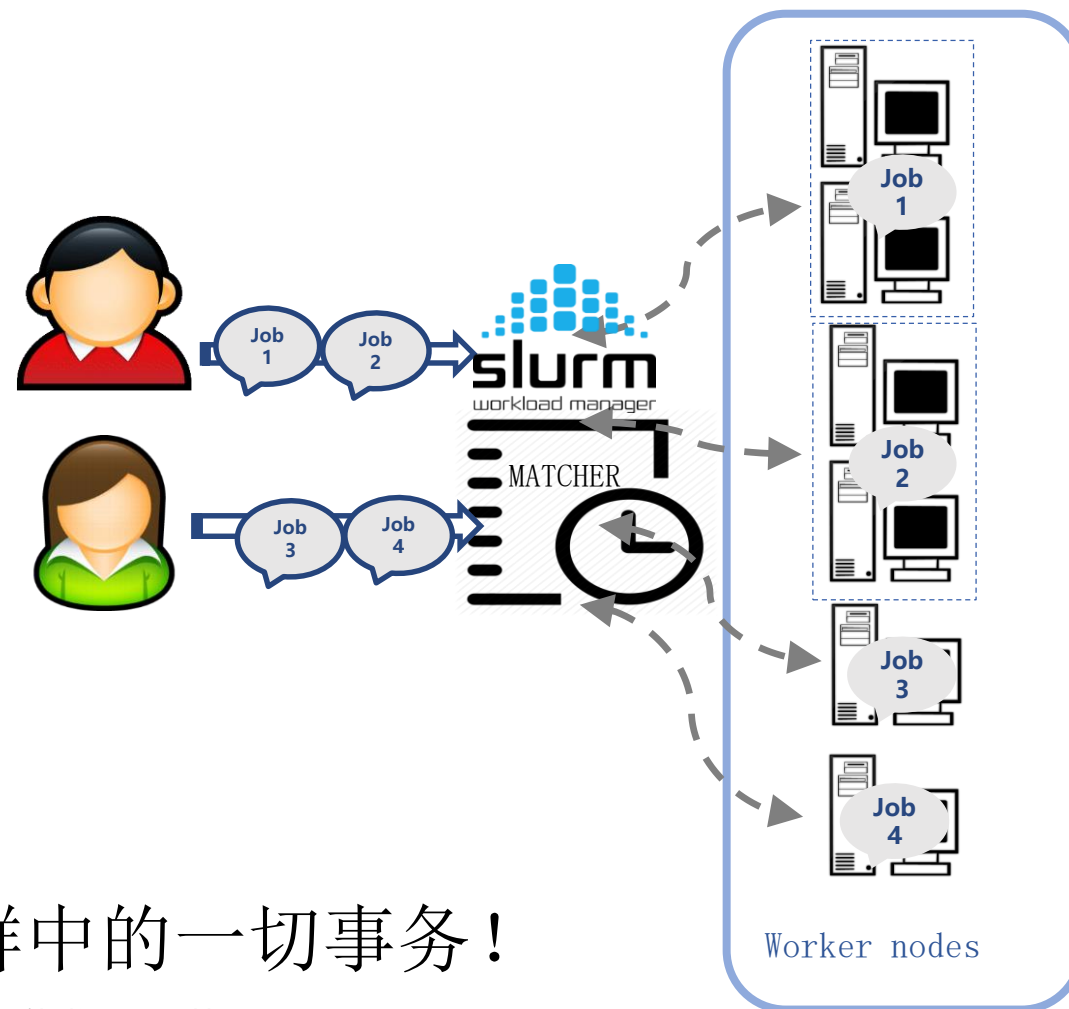


HPC作业调度与资源管理 -- SLURM



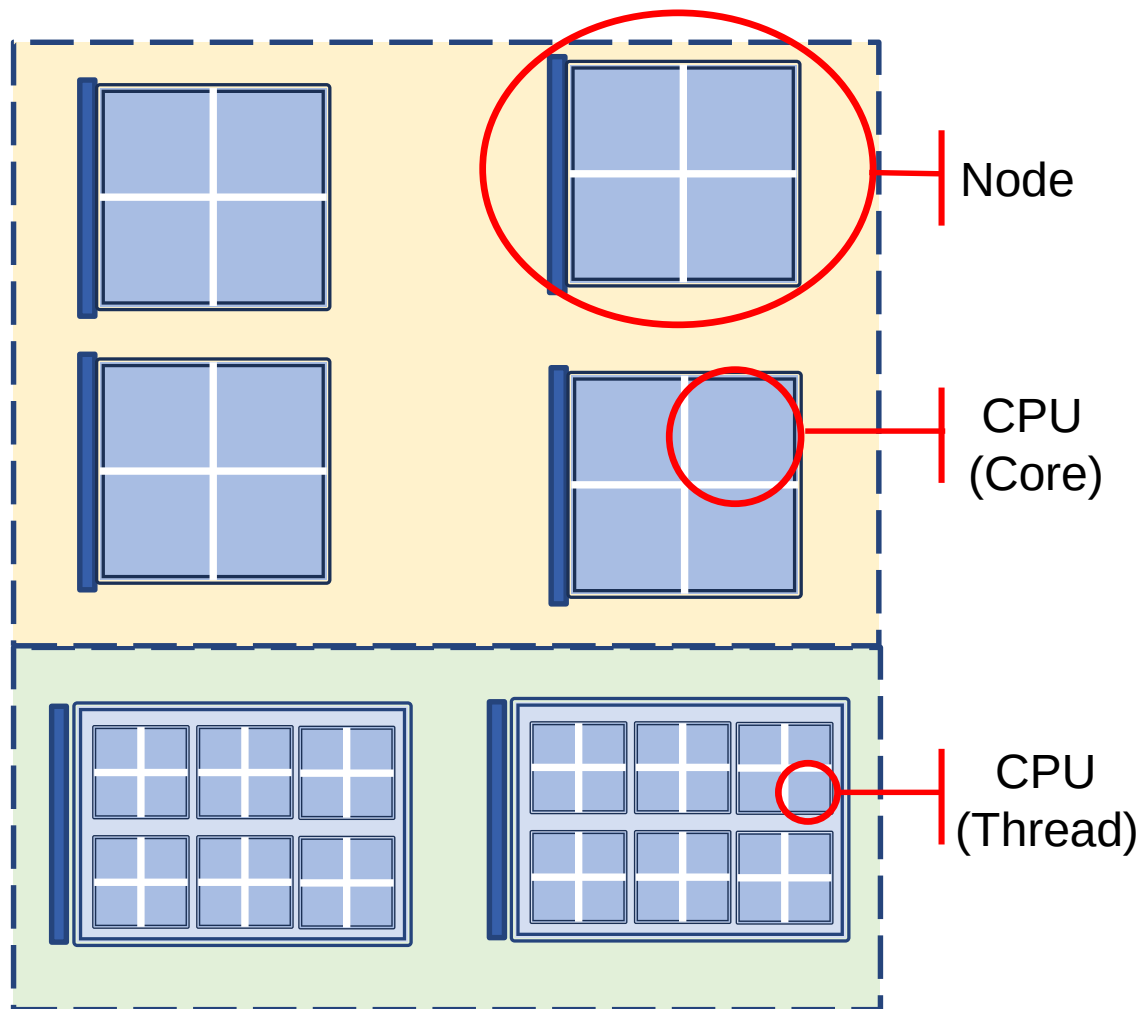
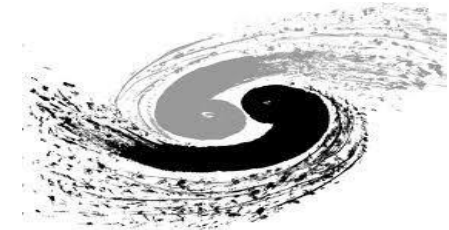
Slurm -- Simple Linux Utility for Resource Management

- 开源，容错，可扩展软件
- 提供作业的开始，执行，
监视整个生命周期操作
- 连接计算资源和队列中的作业
- 对大并行作业支持良好
- 在超算中心得到广泛应用
 - 被应用于大量非高能物理领域
 - 高能物理领域使用越来越多



调度器负责并控制集群中的一切事务！

SLURM 资源管理



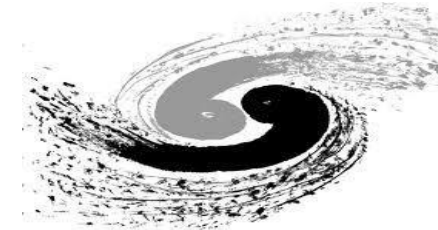
计算节点

- CPUs
 - Cores
 - threads
- 内存
- 状态
 - Idle
 - Mix
 - Alloc
 - Completing
 - Draining
 - Down

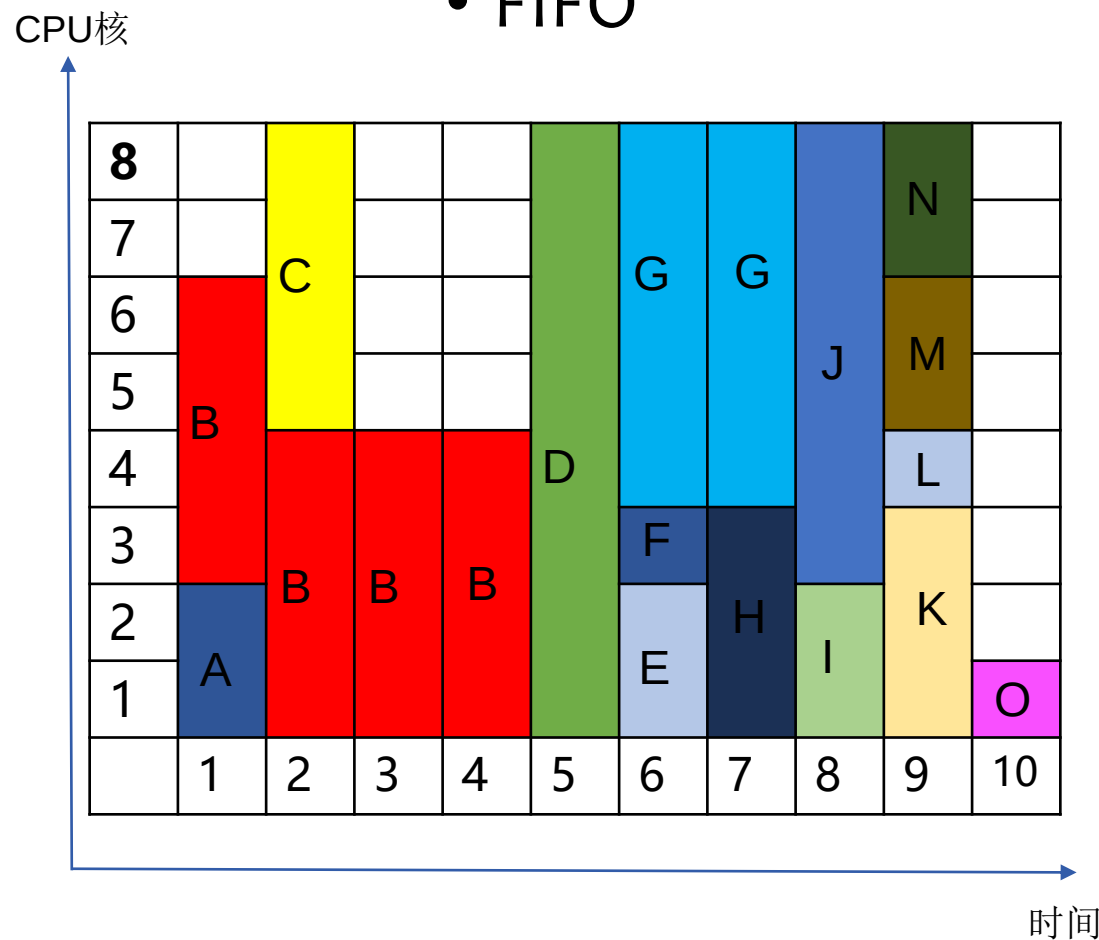
Partition

- 设定的节点集
 - 同构
- 状态
 - Idle
 - Mix
 - Alloc
 - Completing
 - Draining
 - Down

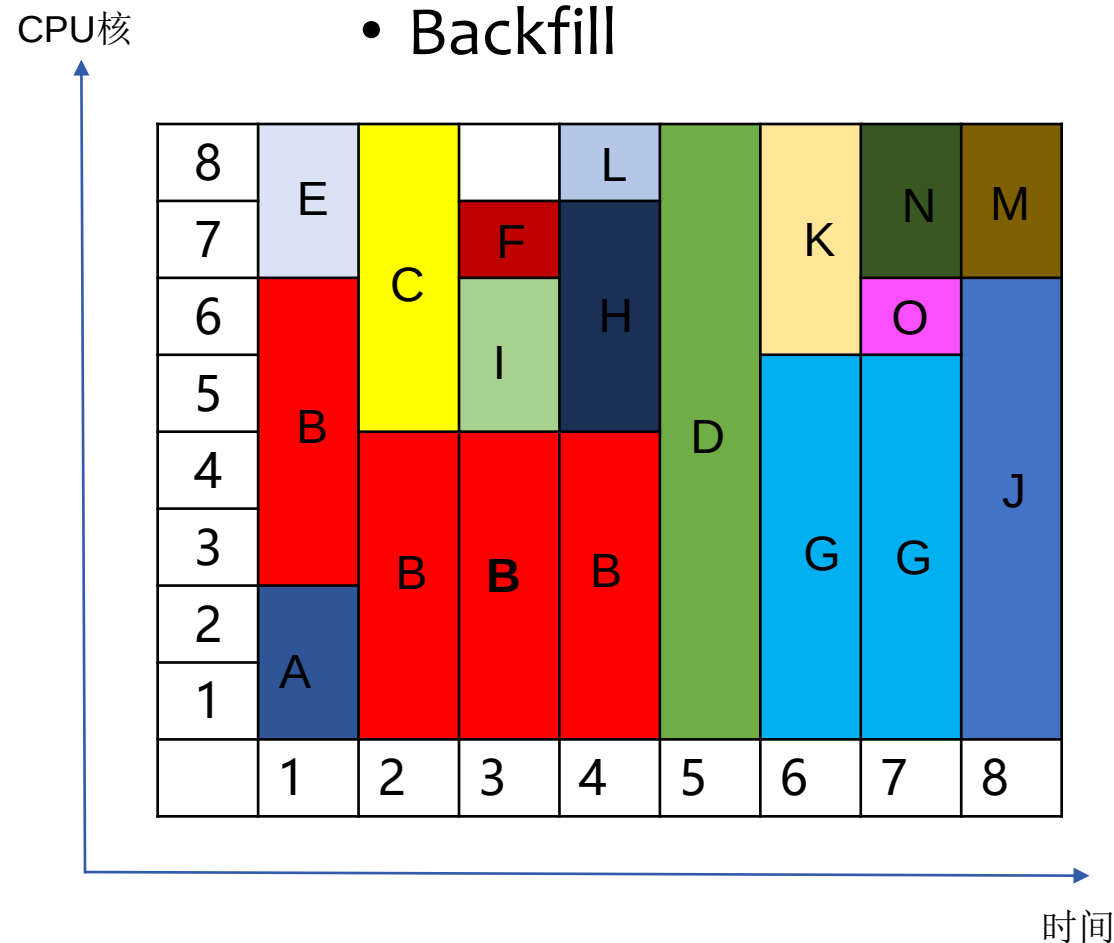
Slurm具有丰富的调度策略



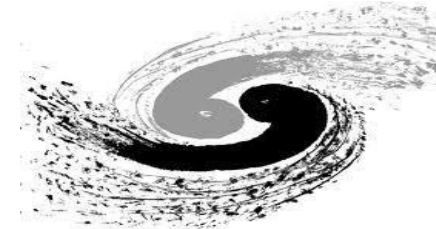
• FIFO



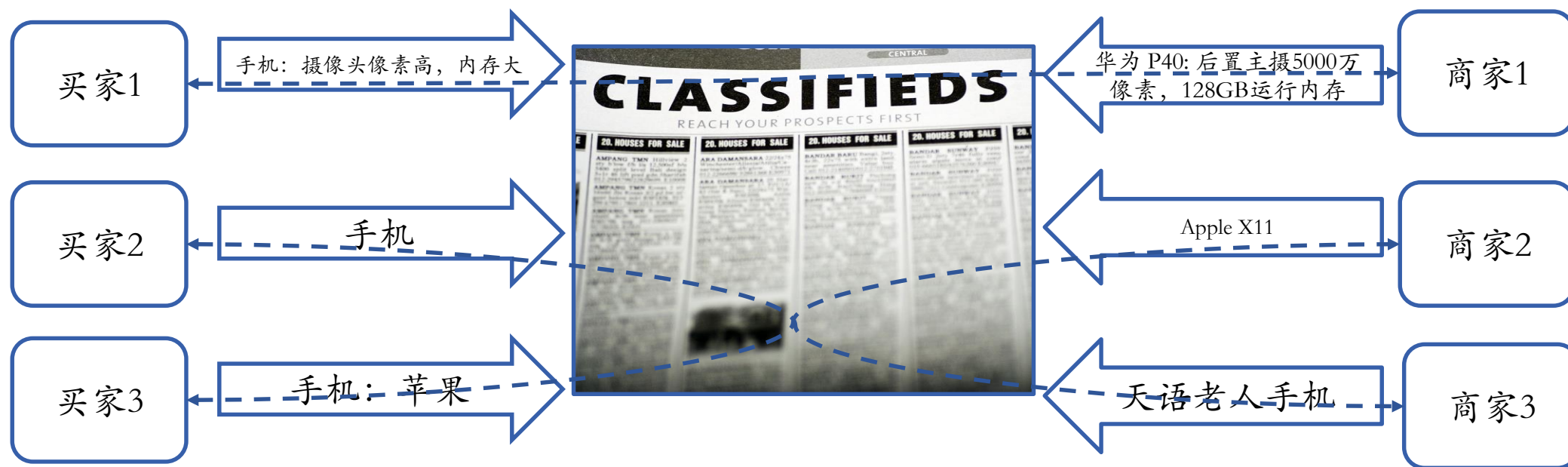
• Backfill



HTC的作业调度与资源管理



- HTCondor -- 一个示例
 - 买家/商家：购买/销售手机



HTCondor的作业调度



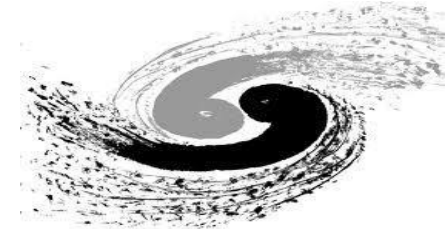
Job Ad

```
Universe = "Vanilla"  
AcctGroup = "physics"  
RequestOS = "CentOS"  
RequestMemory = 8000  
Requirements =  
    (OPSYS == RequestOS) &&  
    (Cpus >= 1) &&  
    (Memory > RequestMemory)
```

Job Slot Ad

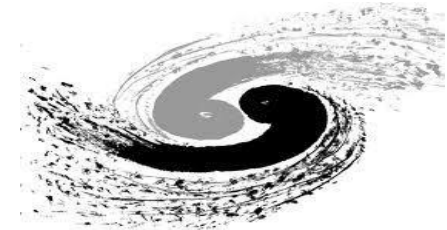
```
OpSysName = "CentOS"  
Memory = 10531  
Requirements =  
    (Universe == "vanilla") &&  
    (AcctGroup == "physics" ||  
     AcctGroup=="lhaaso" ||  
     AcctGroup=="juno")  
Rank = physics
```

HTCondor 调度算法的理念 - 公平共享



- 用户使用计算资源的公平性
 - 用户作业优先级 -- 动态调整
 - 基于最近时间过内用户使用资源的总和进行优先级排序
 - 跑得越多的用户，其作业优先级越低
 - 优先级按指定周期内作业累计运行时间总和半衰递减
- 不同用户组使用资源的公平性
 - 用户组的资源共享
 - 用户组可用资源受限于各组资源份额
 - Surplus模式：
 - 当有空闲资源时，可突破本组份额，占用更多资源
 - 份额未用足的用户组作业优先级最高
 - 共享池子越大 → 利用率越高，份额越容易保证

HTCondor 调度算法的理念 - 公平共享

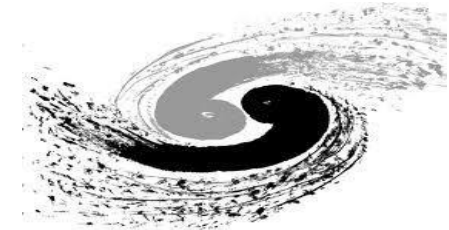


$$\frac{\sum Job\ Runtimes}{Wall\ Time} \Rightarrow \frac{\sum Completed\ Job\ Runtime}{Wall\ Time} \Rightarrow \left(\frac{\sum Completed\ Job\ Runtime}{Wall\ Time} \right)^*$$

* Subject to some notion of fairness

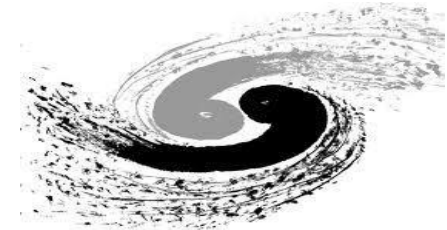
- 不同实验资源份额
- 公共账号→高优先级
- 作业运行时因素

HTCondor的运行模式特点 - 信任



- HTCondor对计算资源的管理：计算节点自己汇报运行策略，作业结束状态，Condor不干涉（杀死节点），只收集节点拥有的资源情况
- HTCondor对作业的管理：作业自己向HTCondor汇报自身需求和具备条件
- HTCondor对数据的管理：
 - 对存储系统无要求
 - 支持 共享文件系统/sandbox传输/第三方传输协议
- HTCondor安全易扩展：轻量级、无状态的中央管理机制，支持SSL, Kerberos, GSI, HostIP, password, security session
- HTCondor 对作业的管理：多种作业类型，详细作业生命周期日志

提纲



1

高能物理离线数据处理过程与特点

2

集群计算，网格计算，分布式计算

3

高通量计算与高性能计算

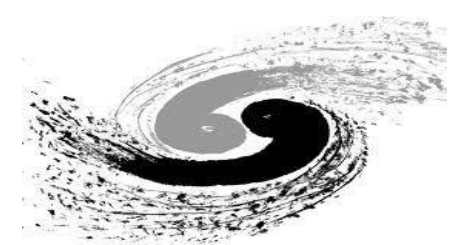
4

分布式计算

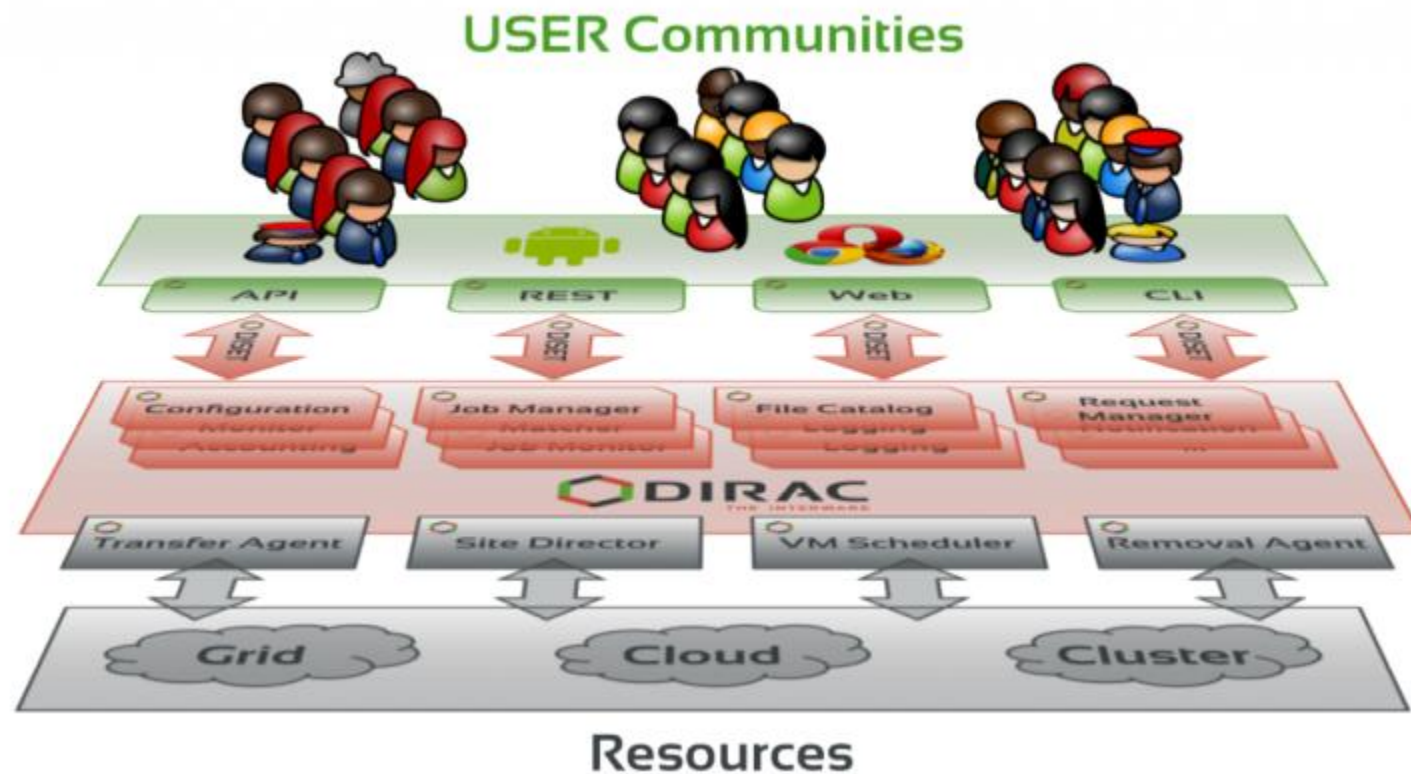
5

总结

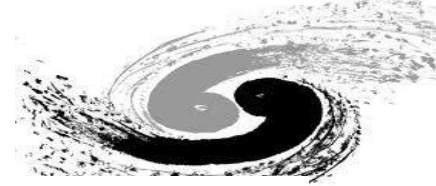
分布式计算 -- Dirac



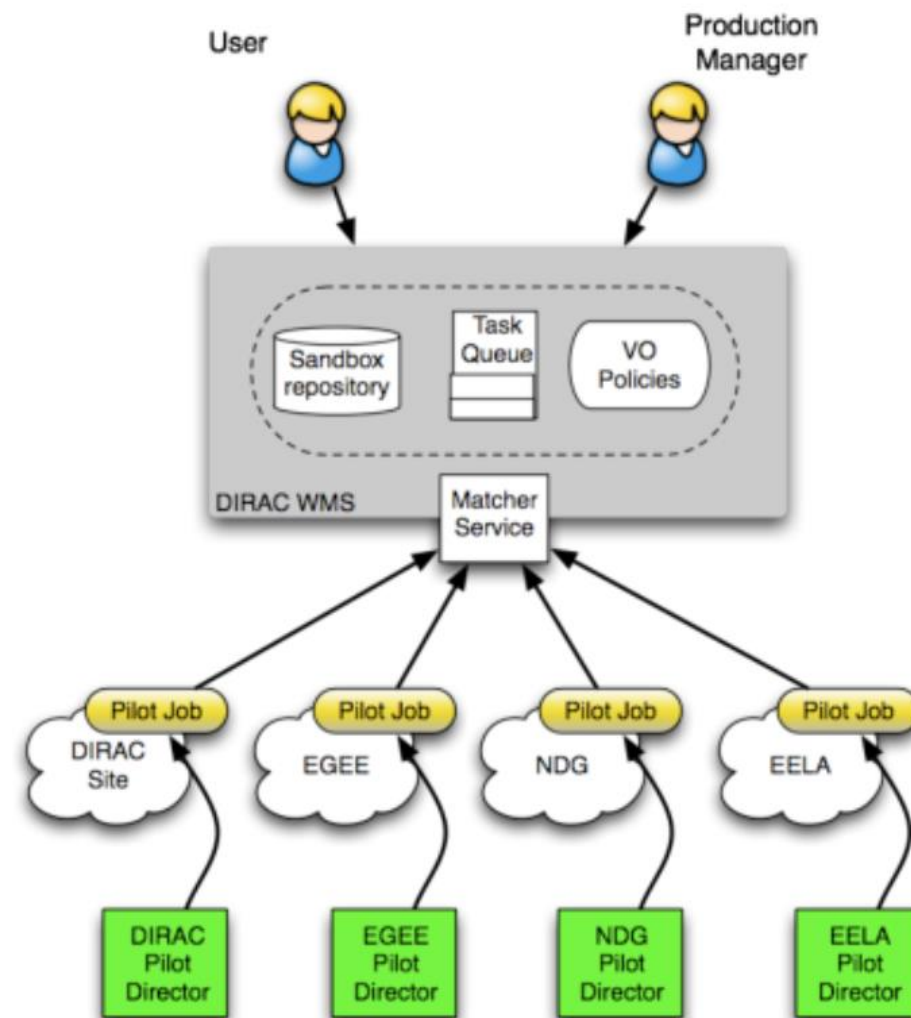
- 开源分布式软件框架
- 在用户和资源之间的中间层
- 与实验无关
 - 可扩展，灵活性高
- 可整合多种资源
 - 集群，网格，商业云
- 功能完备
 - 基于面向服务架构
 - 资源调度、数据管理、文件传输、
监视记账。。。。



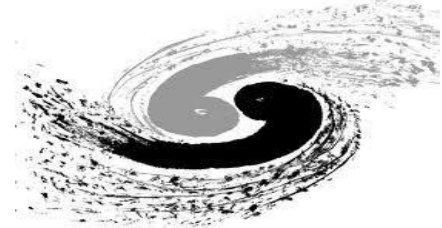
Dirac 作业调度 -- 拉作业



- “拉作业”的pilot机制
- Pilot用于为作业准备所有运行环境及相关配置
- 用户以证书及VO将作业提交到 Central Task Queue
- Directors向Grid WMS提交指定用户角色的pilot作业
- Pilot作业从TaskQueue里拉取用户作业
- 作业被pilot监管下执行。

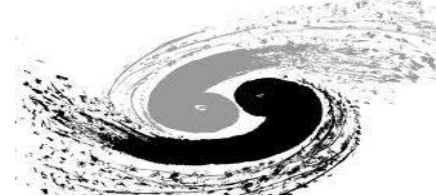


Dirac 的应用



- Dirac由LHCb实验开发，用于WLCG资源整合，成功应用于LHCb, BelleII等多个高能物理实验，用户群体包括LHCb, ILC, Belle II 等实验，还拓展应用于天文，生物领域，包括T2K, CTA, Pierre Auger Observatory, Eiscat 3D, BioMed 等
- 2014年，高能所自主建成基于DIRAC的分布式计算系统，为 BESIII、JUNO、CEPC 等中国高能实验服务，现已整合包括美国、俄罗斯、意大利、英国、土耳其、法国、台湾、北大、北航、国科大、中科大、上海交大、武汉大学、等十多个分布式站点，资源类型覆盖集群、云、网格以及志愿者计算等

基于HTCondor集群扩展的分布式计算



- 高能物理计算平台：跨地域、跨学科的异构计算平台

- 南北两个区域中心

- 高能物理研究所计算中心

- 高通量计算/高性能计算/网格站点
 - Lustre / EOS 文件系统/磁带库

- 东莞大科学智能计算中心

- 高性能计算/云计算
 - 华为OceanStor9000

- 众多远程站点

- 北京、广东、四川等多个大型高能物理实验装置配套IT资源
 - 实验取数使用，初步分析
 - 山东大学、中国科技大学等众多合作组成员计算资源
 - 边缘站点：计算及存储资源规模等不稳定
 - 可弹性接入国家超算中心、商业云等领域外资源

- 以高能所计算中心集群为中心向外扩展，集成其它站点资源

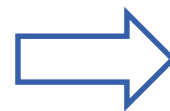
- 轻量级，对于远程站点最小影响和变动，无额外运维负担

- 保持用户原有的数据分析习惯 → 去网格化

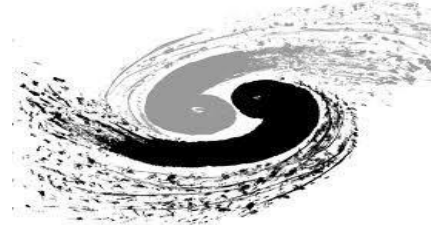


- 需要解决

- 对实验用户作业细致分类
 - 用户服务的认证透明
 - 集群规模动态弹性扩展
 - 对高能所站点存储的透明访问



站点分类与作业的分类



- 将作业分发至最适合的站点

- 每个实验的各类作业有自身特点

- 输入/输出数据大小;
- IO性能需求高低
- 特殊服务支持
- 实验作业量多寡

- 各个站的资源条件不尽相同

- 网络带宽条件
- 计算节点CPU强弱情况
- 长期/临时存储空间

- 站点分类

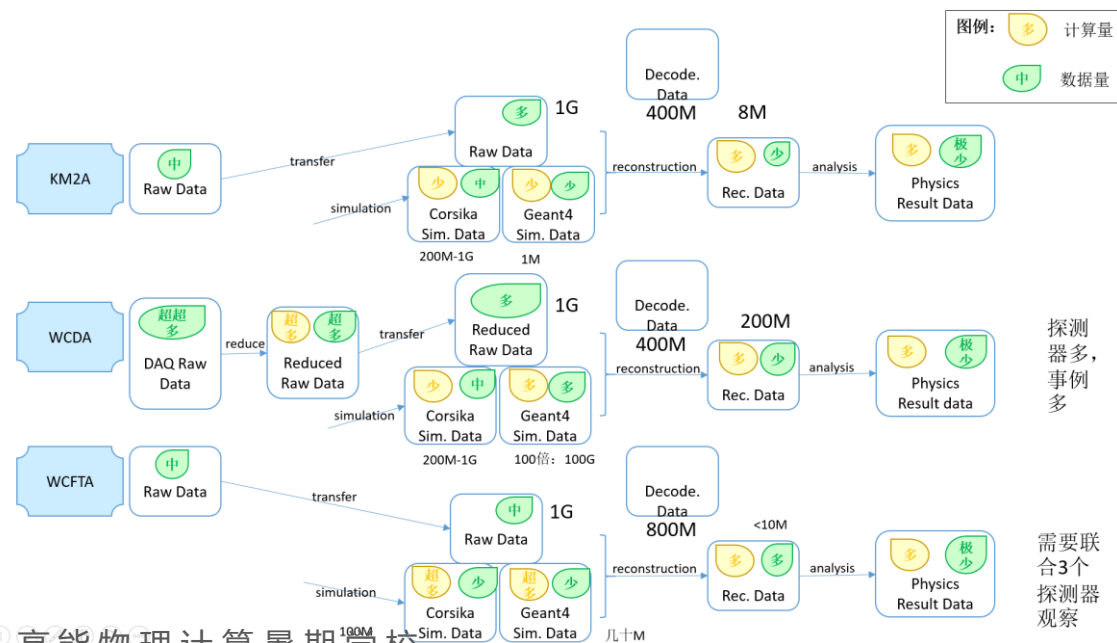
- 区域中心 - 北京/东莞
 - 长期稳定的计算、存储资源
 - 稳定可靠的网络连接
- 边缘站点 - 合作组成员
 - 无长期可用存储, 可用计算资源变化频繁
 - 有限的可用网络带宽
- 临时站点 - 商业资源
 - 峰值需求的短期使用

Scheduler

Tag and Match

- 作业分类

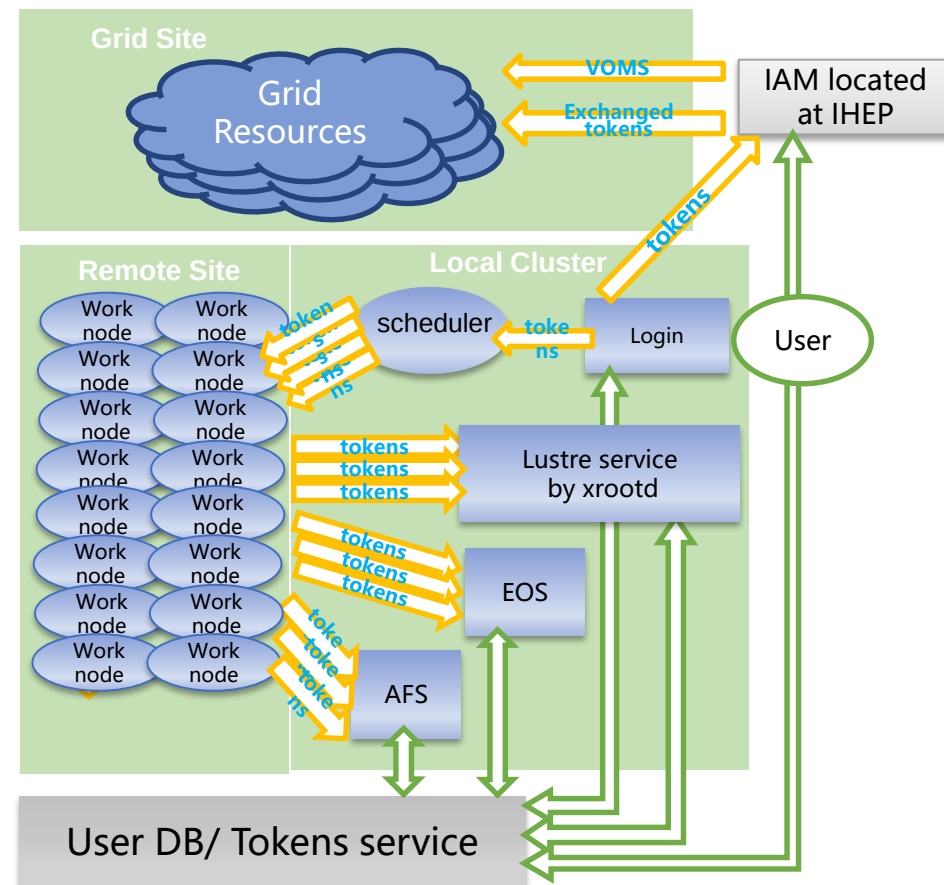
- 不同实验计算作业不同
 - 模拟重建计算占比不同
 - IO需求不同
 - 网络带宽要求不同
 - 临时存储空间需求不同



基于Tokens的用户认证实现不同服务的认证



- 计算集群采用Token 认证，并集成于高能所SSO
- 用户作业通过自己的Token使用分布式计算平台提供的服务
- （本地/异地） 站点提供的服务均支持Tokens认证
- 用户登录高能所集群即生成自己的Token
- 用户的Token被HTCondor从高能所集群透明传送至远程站点，供作业运行期间使用计算平台的各种服务

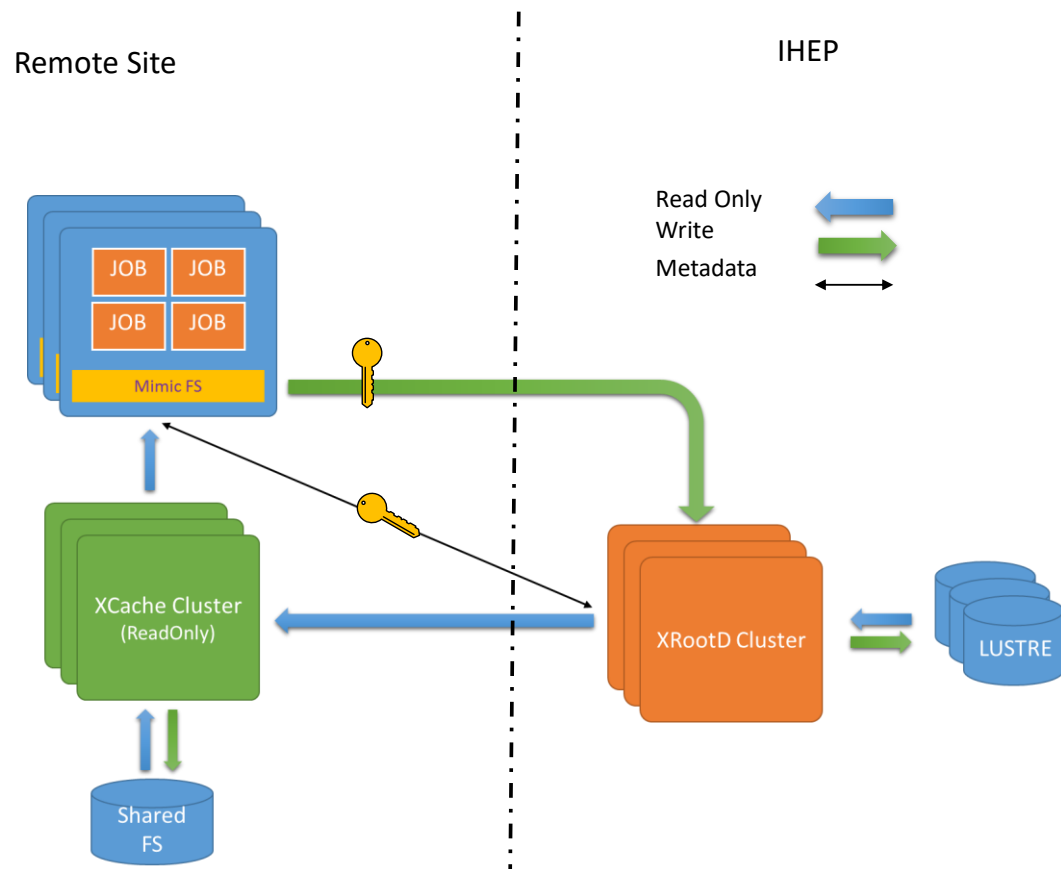


基于Token认证架构

透明数据访问模式



- 开发面向Lustre Mimic文件系统并部署在远程站点
 - 从linux kernel 接收访问请求
 - 文件路径
 - 访问操作及参数
 - 上下文处理
 - 将访问请求转为XrootD URL并发送给XrootD服务器
 - 接收XrootD的响应，并传给linux内核
 - 将URL转换成读/写元数据的操作
- XrootD协议
 - 在高能物理计算领域被广泛应用的数据访问协议
 - 支持Tokens认证
 - 具备性能线性扩展能力
- XrootD Server@IHEP
 - I/O 代理
 - 用户镜像
 - 从Token中读取用户名
 - 使用用户Token访问lustre文件系统



基于XrootD的posix数据透明访问

基于HTCondor计算集群的扩展

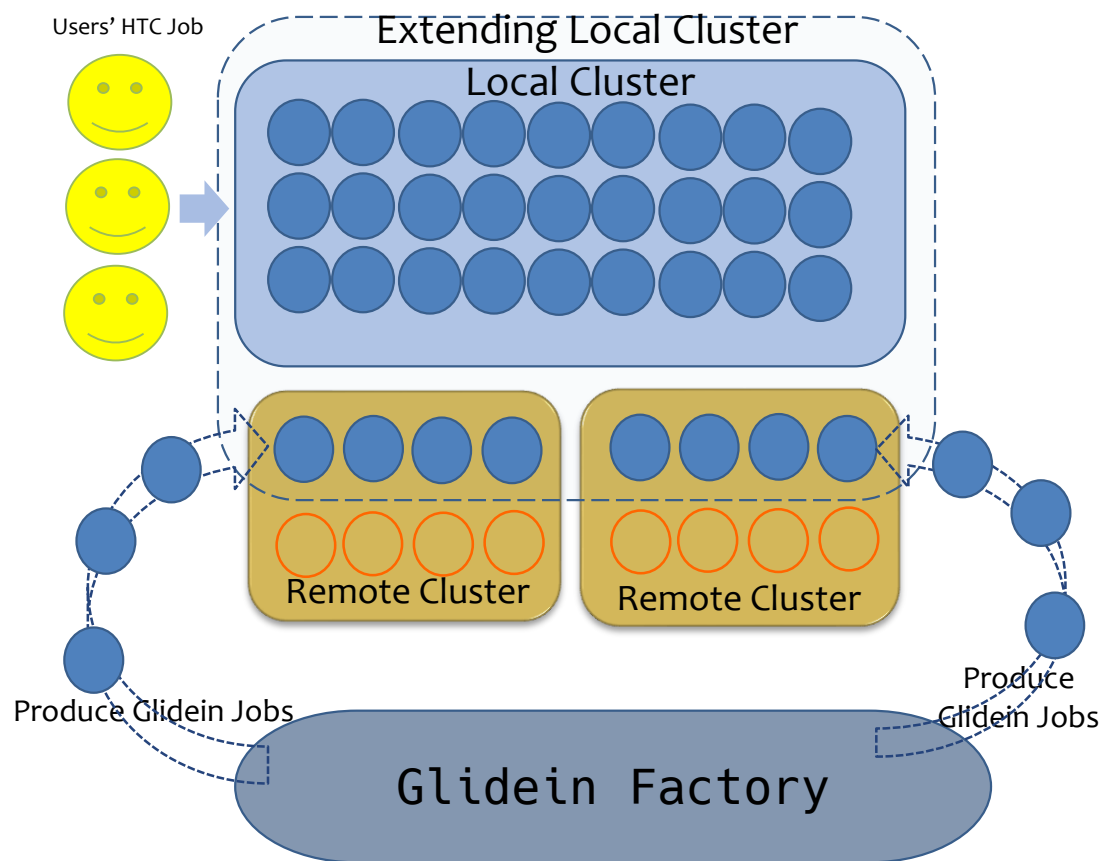


● HTCondor集群的透明扩展步骤

- ① 用户通过HEPJOB工具向高能所集群提交作业
- ② HEPJOB工具分析作业情况，为作业添加分类标签
- ③ 根据作业分类及排队情况，Glidein Factory向远程站点发送glidein作业
- ④ Glidein作业被远程站点当做其本地作业执行
- ⑤ Glidein作业启动高能所集群客户端，生成作业槽加入高能所集群
- ⑥ 用户作业被分配至在远程站点上的“高能所集群作业槽”运行

● 扩展集群的调度管理

- 根据扩展集群调度策略，将作业分发至合适的远程站点的合适作业槽
- 为不同实验各类作业定制不同的wrapper脚本，使作业认为自己是在高能所集群上运行

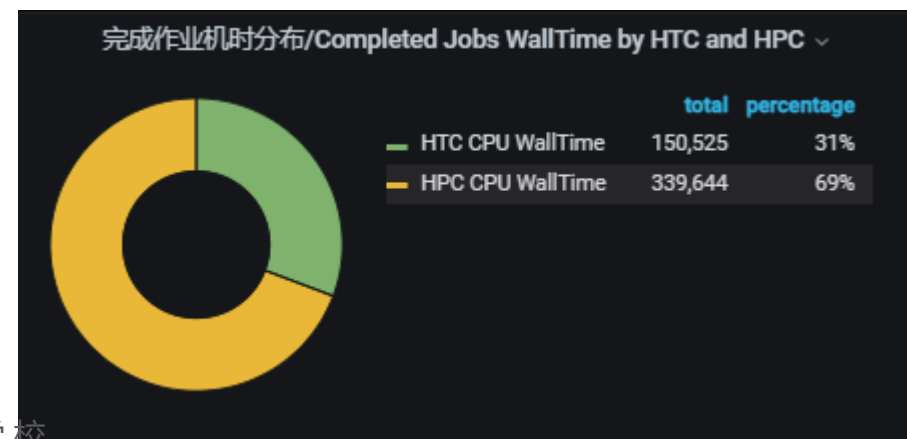


“HTCondor集群扩展”架构

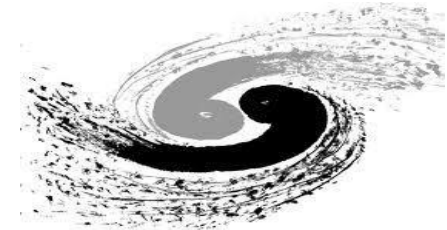
基于HTCondor计算集群的扩展现状



- 集群扩展模式
 - BES实验与LHAASO实验的本地集群作业被分发至东莞区域中心运行
 - 用户视角：使用高能所本地集群
 - 本地和远程集群均使用同一作业脚本和参数
 - HTCondor的wrapper屏蔽了环境和数据访问的差异性
 - Mimic文件系统提供了posix的透明访问
 - 提供有限资源的边缘站点接收本地集群中特定类型作业运行



提纲



1

高能物理离线数据处理过程与特点

2

集群计算，网格计算

3

高通量计算与高性能计算

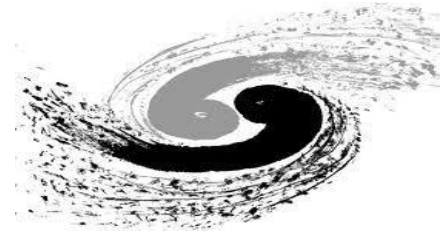
4

分布式计算

5

总结

总结



- 高能物理离线处理需要有力的计算服务支持
- HTC, HPC计算是高能物理计算的基础
- 分布式计算整合更多高能物理资源协同工作
- 计算平台的建构与运维已成为高能物理实验重要组成部分
- 在需求的牵引下，计算平台在不断发展与完善

