

面向事例的高能物理大数据管理系统

Monday, 5 June 2017 15:00 (20 minutes)

摘要：新一代高能物理实验装置的建成和运行，产生了 PB 乃至 EB 量级的数据，这对数据采集、存储、传输与共享、分析与处理等数据管理技术提出了巨大挑战。事例是高能物理实验的基本数据单元，一次大型实验即可产生万亿级的事例。本文就高能物理事例的索引技术，事例跨域传输技术，事例缓存技术进行介绍。

传统高能物理数据处理以 ROOT 文件为基本存储和处理单位，每个 ROOT 文件可以包含数千至数亿个事例。这种基于文件的处理方式虽然降低了高能物理大数据管理系统的开发难度，但存在着很多问题。比如全数据扫描，筛选时间长。基于文件的缓存效率低，基于文件的传输 Overhead 高。在实际的高能物理数据分析过程中，大部分的数据都是物理学家们不感兴趣的数据，而且可以通过一些简单的条件即可过滤掉，如果条件设置得当，该系统能够帮助物理学家筛选掉甚至 99.9% 的不感兴趣的数据。这样不仅可以节省 I/O 资源，还能提高 CPU 利用率，减少数据分析耗时。提出一种面向事例的高能物理大数据管理方法，重点研究海量事例特征高效索引技术，在这种方法中，将物理学家感兴趣的事例特征量抽取出来建立专门的索引，存储在 NoSQL 数据库中。为便于物理分析处理，事例的原始数据仍然存放在 ROOT 文件中。最后，通过系统验证和分析表明，基于事例特征索引进行事例筛选是可行的，优化后的 HBase 系统可以满足事例索引的需求。

大型高能物理实验往往由国际合作单位共同贡献资源形成分布式计算系统，比如 WLCG[4]、BES Grid 等。传统的计算方式是事先将数据传输到目标站点，然后再将计算任务调度过去运行。随着网络带宽的提升，全网调度计算任务，数据远程访问成为未来的发展趋势。一般局域网的时延在 1ms 以下，而中国到欧洲的广域网时延能达到 200ms 左右，在这种情况下直接使用文件系统 I/O 访问基本无法工作，急需要研究高带宽的远程 I/O 访问技术。欧洲大型强子对撞机产生海量数据便是由 WLCG(World wide LHC Computing Grid) 负责存储和处理的。在 WLCG 的 Tier 结构中，数据并不是完全复制到所有的站点中，因此计算任务会被调度到存储数据的地方。如果某个站点需要分析感兴趣的数据，需要提前进行数据订阅，将数据预先传输到指定的站点。不同于 WLCG 预先传输文件，面向事例的数据传输系统仅传输物理分析程序所感兴趣的事例，所需数据量大幅降低，随着网络带宽不断提升，将可以支持计算任务实时传输数据。数据传输系统由数据传输服务器和数据传输客户端两部分组成，分别运行在不同的站点。数据传输服务器负责数据的存储和对请求的响应。在服务器端应用了多进程并发处理机制，实现高效的请求响应。运行在远程站点的高能物理数据处理软件在做物理分析时不用考虑数据是否在本地球点，它可以通过 ROOT 框架或者本地文件系统接口来访问所需要的事例数据。为提升数据访问性能，在数据传输客户端设置了基于事例和数据块的缓存系统。数据传输基于 HTTP 协议，支持分块、多流及断点续传等功能。并基于 OAuth 授权进行安全保障。系统测试结果表明，在网络带宽良好的环境里，带宽利用率可以达到 90% 左右。

设计实现了事例级高能物理实验数据的跨域访问缓存系统进行跨站点数据缓存。物理学家进行实验作业分析时，不需要将整个 DST 文件下载到本地。将事例请求发送至缓存服务器后，缓存服务器向远程站点发送请求，之后以事例为级别进行 HTTP 多流传输至本地缓存，并返回至客户端。对客户端来说，所有操作都是在缓存服务器上进行的，远程站点是透明化的。缓存服务器提供了按需访问、动态调度的新型高能物理数据跨域访问模式，系统访问及传输以事例为单位，大大的减少了资源浪费，提高了作业处理效率。同时缓存系统提供了统一数据管理、远程站点统一文件视图，为用户提供了本地化操作模式。缓存系统中设计了用户操作日志分析模块，以 syslog 模式抓取用户对于数据分析的记录，通过近期数据分析，实现数据预取来增强系统读性能。在整个缓存系统模块中应用了多进程并发处理机制，实现高效的请求响应和高性能的读写调度架构。系统中客户端与服务器端通信都采用了高能物理计算中通用的 XROOTD 架构，具有较强的普适性与通用性，更好的与高能物理实验分析作业相结合。

作为一种新型的高能物理事例管理系统架构，有效的解决了传统基于文件处理的资源浪费和效率低下问题，同时缓存服务器将远程站点的数据以本地化的模式提供给用户，提供了便捷高效的数据处理模式。整个系统为高能物理跨域计算提供了新型的架构，在高能物理计算环境中具有较好的应用发展前景。

Primary author: Ms 王, 聪 (IHEP)

Co-author: 徐琪 (高能所)

Presenter: Ms 王, 聪 (IHEP)

Session Classification: 高能物理计算机软件:BESIII&MOMENT

Track Classification: 高能物理计算机软件