**基于加速器的粒子物理发展战略规划**

**高能量前沿-软件与计算**

**工**作组成员：

邓子艳，黄性涛，李刚，李卫东，林韬，孙胜森，尤郑昀，张晓梅，邹佳恒

(2020/7/29)

**主要内容**

[背景 2](#_Toc46908300)

[挑战与机遇 2](#_Toc46908301)

[软件平台 3](#_Toc46908302)

[模拟 7](#_Toc46908303)

[重建 9](#_Toc46908304)

[分析工具 10](#_Toc46908305)

[数据可视化技术 11](#_Toc46908306)

[高性能计算与网格 13](#_Toc46908307)

[数据和软件的持久化 16](#_Toc46908308)

[软件开发环境 17](#_Toc46908309)

[经费支持和配套措施 17](#_Toc46908310)

## 背景

希格斯玻色子发现后的高能粒子发展主要聚焦在通过高强度和高能量两个前沿方向来对新物理规律和新现象进行大范围扫描，其中希格斯玻色子本身就是一个通向新物理和新现象的物理的研究工具和窗口。在未来的十五年乃至更长时期，软件与计算领域的技术发展将为基于加速器的粒子物理发展提供技术支持和驱动力，实现在精度和灵敏度的方面超越以往的实验，为基于加速器的高能量粒子研究在深度和广度上获得突破奠定基础。

基于高能实验科学目标的需求，实验数据处理在数量和复杂性方面都有显著增长，同时速度要求更加严苛，只有对计算和软件的创新性发展给予大力的支持，进一步推进业界研究和发展的良性科研环境建设，激励与支持在计算和软件方面更强的自主创新能力培育与更加广泛的国际合作与交流，才能为把握当前国际高能物理实验发展的重大机遇，实现科学目标提供可能。过去三十年我国大科学装置科学研究的成功经验已经证明，来自科研机构和高等院校的参与，以及基金委和科学院的扶持是必不可少的。所需的支持不仅包括财政支持，同时也包括共享计算与数据中心等基础设施，对计算和软件研发人员的支持与认可，高效组织架构（如软件与计算实验室）的存在，构建与不同学科，乃至商用计算机科学行业之间的促进与共同发展关系。

在将来建立新的加速器与探测器设施的同时，需要相应的软件与计算研发投入，管理、处理和分析将要获取的大量数据。在未来高能量前沿实验中，挑战随着科学目标的复杂性而增长，在实验的路线图与软件目标和优先级规划上达成一致是至关重要的，并且这些科研工作是互为促进与补充的。基于这一思路，对软件与计算的研究发展进行了描述。

## 挑战与机遇

随着后摩尔时代的到来，单纯依靠增加商用硬件、格点计算等传统计算资源投入获得计算能力增加的情形已经迅速发生着改变。与此同时，随着未来高能量前沿实验规模的扩大和数据复杂程度的提高，数据的获取、处理与分析需要的计算量与当前实验相比有显著增长。计算资源与需求之间的矛盾不仅是计算和软件领域的问题，更是粒子物理实验科学产出面临的严重挑战。通过创新改进软件性能，大幅度提高现有和未来计算资源的利用效率，突破计算能力瓶颈是未来较长一个时期的主攻方向。

粒子物理实验中使用大量的遗留软件需要在内存使用和吞吐量方面进行重要改造。软件算法升级与发展采用先进计算技术和适应不同计算平台以获得最佳的数据处理效率，如并行计算，GPU硬件加速和超级计算机等。利用机器学习和人工智能等先进技术的优势为实现现有粒子物理实验软件性能的大幅提高提供了可能的选项。

粒子物理实验软件从业人员除计算机科学专业知识外，粒子物理研究和探测器知识背景也是必须具备的技能，较高的准入门槛使得人才储备不足。同时，由于软件工程的迅猛发展，早期从业人员更多关注运行中实验的维护，在新的软件算法体系结构的差距越来越大。这个问题同样不是我国独有的，国际高能物理实验面临的普遍难题。加强国际合作与交流，共同发开基于抽象架构的通用性软件，适应我国高能量前沿实验的发展，同时培养和锻炼人才，不断提高我国从业人员的自主创新能力。

## 软件平台

粒子物理离线软件是大型粒子物理实验的重要研究内容之一，为高效精确地处理和分析实验获取的海量数据以及开展蒙特卡罗模拟研究提供关键技术，保障实验的科学产出和成果发表。离线软件系统通常由软件平台、事例产生子软件、探测器模拟软件、刻度软件、重建软件以及物理分析软件等子系统组成。软件平台是离线软件系统的核心组成部分之一，是结合实验的不同需求，通过采用计算机技术、信息处理技术、高性能计算技术、大数据管理技术以及统计分析技术等构建起粒子物理实验离线数据处理和物理分析计算和软件环境，实现处理流程管理、数据管理、文件管理等核心功能，并提供标准的事例产出、探测器模拟、刻度、重建以及物理分析等应用开发接口和统一管理方式，在实验前期对探测器方案的优化、蒙特卡洛模拟研究以及实验运行后的数据处理过程提供关键技术和解决方案。

**底层框架：**针对高能量对撞机实验的需求，目前广泛使用的是由LHCb实验组设计和开发的Gaudi框架，并被多个对撞机物理实验作为底层的框架软件所采用。最近几年，为了满足日益增长的实验数据或愈加复杂的物理分析研究，Gaudi框架本身也再不断的优化和扩展，比如增加了其对并行计算以及公共数据模型等方面的支持，并作为未来加速器物理实验公共框架软件的候选者加入了Key4hep软件站。但为了满足今后HL-LHC，CEPC， FCC等实验日益增长的数据量，需要进一步优化其自身的架构、实现对新兴计算技术、数据处理方法以及新的硬件架构的支持。

**数据管理：**数据管理是粒子物理实验离线数据处理和物理分析研究的关键因素，决定着离线软件系统的设计和开发以及整个离线软件系统的性能和功能。从数据的类型上分为事例数据和非事例数据。

事例数据指在事例产生、探测器模拟、刻度、重建以及物理分析过程中描述物理事例及其关联关系的信息，是不同数据处理过程和物理分析研究的对象。事例数据的管理包括数据格式、数据接口、数据存储以及数据在内存和文件中的管理等。目前ROOT是高能物理实验数据分析的通用软件，同时它还提供了数据的读写机制、动态类型信息、前后兼容以及数据压缩等高级功能，因此，ROOT格式成为当前粒子物理实验事例数据最为流行的格式。基于ROOT的事例数据管理方式也成为许多大型粒子物理实验的最自然的选择，通过采用一套事例数据模型，实现数据在处理过程和存储等环节中的共同，避免了数据类型的转化，在一定程度上缓解了数据的读写的瓶颈。

粒子物理实验沿着高能量、高亮度快速发展，实验数据量急剧增长、数据结构也越来越复杂，数据处理和物理分析研究的难度也越来越大。因此研究新的事例数据管理机制和技术成为必然，其内容主要包括：1）事例数据模型对并行计算的支持；2）通用事例数据模型的设计和开发。并行计算成为当前粒子物理实验离线数据数据和物理分析软件面临着首要技术挑战，而事例数据是实现并行计算的关键。传统的基于类的继承关系设计的分层数据模型成为制约并行计算的因素，因此采用扁平的数据格式成为当前事例模型设计和开发的主要方向，其结构简单、易分割和序列化等特点为在并行环境中的线程安全提供了可能。另外目前各大型粒子物理实验的规模越来越大、物理研究的要求越来越高，其离线数据处理和物理分析的技术也面临着巨大的挑战，各个实验之间的进一步协同开发、技术和方法共享成为当前粒子物理实验软件与技术方向共识，基于数据模型是联系各个数据处理环节和物理分析的主线，因此，设计和开发出国际上各个实验通用的数据模型是今后研究的重点，不仅能够减少重复研发，更为主要的是为重建和物理分析方法等方面的共享奠定了基础。

非事例数据包括探测的配置、刻度以及运行环境等过程中数据产生的数据，对事例数据的处理和物理分析研究发挥着重要作用。因现在粒子物理实验的规模越来越大、运行周期越来越长，合作单位越来越多，其非事例数据的管理也面临着新的挑战。传统上通过简单的采用数据库管理的方式其在非事例数据的访问、迁移、维护等方面的问题越来越多。因此，开展非事例数据管理新机制、新方法的研究也成为当前大型粒子物理实验的关注的重点。考虑到探测器的规模变大、颗粒度变小、运行周期边长等因素，需要对非事例数据的管理包括数据模型、管理流程、数据缓存以及用户接口等方面开展新的研究。在非事例数据模型设计方面，把条件数据分为刻度常数和元数据分开的方法，一般刻度常数数据量大采用文件的管理方式，元数据仅描述条件数据的产生或使用条件，数据量小，采用数据库直接管理，通过该中方法能够显著降低数据库管理的数据量，提高数据库运行、维护的性能。考虑合作组用户越来越多、且地域分布广泛，因此采用一些新软件技术或方法，实现数据的缓存机制，避免应用程序频繁的访问中心非事例数据管理服务器，也是当前需要解决的一个问题。另外还需提供给用户简单快捷的非事例数据上传、查看、统计等需求方面的网页接口或客户端，在提升非事例数据管理功能的同时，降低或维持用户对新技术的需求。

【1】 https://github.com/key4hep/EDM4hep/tree/master/edm4hep

【2】 D. Barberis, A. Formica, E.J. Gallas, G. Govi, G. Lehman Miotto, A. Pfeiffer, J. Phys. Conf. Ser. 664, 042015 (2015)

**高性能计算技术：**高能物理实验数据处理需要使用大量计算资源。当今计算机CPU主频提升遇到瓶颈，计算性能改善的主要途径已转变为多核心CPU以及GPU、FPGA等异构架构设备。相应于计算硬件的发展方向，并行和异构软件开发必然成为未来软件发展的重要途径。

并行计算方面，综合考虑高能物理数据特点，采用数据并行与任务并行相结合的并行模型。在大部分高能物理实验中，事例相互间独立无关，可以同时处理，具有天然的数据并行特征。而事例计算的不同阶段（算法）、以及I/O等其它不同任务，通过良好的规划设计也能够做到任务并行。具体实现上，构建多线程和MPI相结合的多层次并行软件系统，实现对大规模计算集群的高效利用。节点内以多线程技术为主，既可有效节约内存资源，又有利于不同任务间的高效通信和同步。节点间使用MPI技术实现Server-Client模式，由Server端利用高速网络对事例进行分发并回收结果，Client端负责具体的计算，从而将多个节点组织为一个有机整体，实现对集群资源的高效利用。再进一步，若能将多台节点内存进行全局管理和合理搭配，从而把足够多事例信息全部缓存到内存中，辅以Spark等大数据技术有望实现低延时、可交互的大规模数据分析。

异构计算方面，重点针对软件中高耗时的计算密集型算法模块，进行专门的优化开发。当代GPU具有高达数千的计算核心，在特定的场景下，对应用可能实现上千倍的性能提升。FPGA作为一种可编程的半定制集成电路，在一些对性能有严苛要求的领域也有成熟和广泛的应用。目前GPU和FPGA的使用和优化需要针对具体的计算问题，进行专门的代码级的优化开发。未来如何减少代码在不同设备间的差异，实现不同类型计算设备的融合，是一项富有挑战性的工作。此外，ARM在服务器CPU领域的快速发展也不容忽视，其相较X86 CPU具有能耗比等优势。兼容X86和ARM的软件生态也应当是我们未来的努力方向。

在高能物理数据处理过程中，特别对于物理分析任务，数据I/O较易成为系统瓶颈，为整个计算环境带来压力。该问题成因与数据特点和计算模式密切相关。实验数据由极巨量的事例组成，事例之间相互独立。物理学家面对某一特定的分析任务，其关心的事例往往只占很小比例，但这些事例以均匀的概率密度散布于所有事例之间。传统做法中，每个物理分析任务都需要读取所有事例数据，根据事例信息筛选感兴趣的事例，然后再进行细致的计算和分析。对于未通过筛选的绝大部分事例，其读取过程耗用大量磁盘带宽，在造成资源浪费的同时也显著降低了数据分析速度。针对该问题成因，对事例建立索引并提供有效的事例判选预处理技术，通过提高对目标事例读取的命中率、减少不必要的事例数据读取，就能改善磁盘I/O压力并提升数据处理的效率。

**机器学习：**机器学习是一种从海量数据中发现统计特征和结构、提取经验的人工智能算法。实践表明，机器学习算法是解决分类、数据回归、概率密度估计、信号压缩、特征提取等科学计算任务的有效手段。近年来，以深度学习为代表的机器学习算法取得了突破性进展，在高能物理、天文等领域出现了大量的应用，未来将在大科学装置的数据处理过程中继续深化应用。一方面，基于机器学习算法改进的数据处理过程，可以应用在大科学装置数据的信号背景分类、研究对象分类、数据重建中的模式识别、在线计算的信号背景过滤、快速数据模拟、探测器运行状态监控和异常检测等方面，这将直接提高科学数据的产出。另一方面，通过基于机器学习算法优化计算资源的管理，包括虚拟机调度、数据放置和迁移策略、自动化参数配置、异常检测、网络路由选择和优化、网络入侵分析等,提高计算效率，可以间接提高科学数据的处理效率，促进科学成果产出。

随着功能更强大的硬件和性能更高的机器学习算法的出现，机器学习工具集将越来越多地被用于开发潜在的应用软件，有可能的应用场景应用包括：

1）大科学装置监控及运行状态的异常检测：大科学装置是非常敏感、昂贵而且复杂的机器，往往包括上亿条数据通道，数千个连续不断取数的性能监控探针。根据现有大科学装置运行的经验，通过机器学习算法对加速器和探测器系统的故障进行预测是可行的。这些算法可以是基于递归神经网络、长短记忆网络等模型的机器学习算法应用。

2）大科学装置数据质量检查：当前大科学装置的数据质量一般通过7x24人工值班来进行监控。人工监控存在的问题首先是值班人员的经验差异很大，其次是现有程序对新问题不敏感。当实验装置的运行状态发生改变时，很有可能会出现新问题。这些问题可以通过机器学习中的异常检测算法来解决。通过与工业界的技术合作，能够实现与数据质量检测相关的经验和成果的转移，从而帮助科学界构建大规模、全局的自动化质量监控基础设施。

3）大规模数据模拟：在大科学装置数据过程中，从大科学装置的研发、到数据分析策略设计、再到真实数据与理论模型的比较等都离不开数据模拟，未来10年至少需要100倍以上当前量级的模拟数据。而实际上，并没有对应的计算资源做匹配。因此，改进现有的模拟算法，提高计算的并行度变得尤其重要。机器学习算法有望用一个以深度学习模型为代表的黑盒子来代替传统数据模拟和计算密集型的蒙特卡洛计算，从而加快模拟的速度。在技术上可以采用多目标的数据回归方法和计算机视觉中最先进的对抗生成网络方法等。

4）粒子识别和基于图片的事例识别：将基于深度学习的，用于解决计算机视觉领域的物体识别问题的算法和模型，用于粒子识别和事例识别，解决计算实时性等问题。

5）数据压缩：未来15年及更长时间，大科学装置数据将增长100倍以上，如何从海量数据中提取最有效的信息，这是高能物理与媒体、电子商务，以及其它工业界的应用共同面临的挑战。典型应用包括数据分析前的压缩，比如在数个小时内将1PB数据压缩1000倍到1TB，同时加上可重用的划分和索引信息等。

6）计算环境的运行效率优化：例如基于作业I/O模式分类的计算资源调度、数据放置和系统计算资源长期规划等。

7）数据工程：机器学习等算法的训练导致大规模的数据密集型计算，需要研究工业界的大数据处理平台技术及流式计算技术与机器学习算法的结合，以及与现有计算框架的结合等。

## 模拟

探测器模拟在探测器设计阶段和物理分析过程中扮演越来越重要的角色。在探测器预制研究和建造阶段，完整的探测器模拟对于估算本底水平和优化探测器性能都有重要的作用，也可用于评估不同探测器几何设计对计算资源的消耗。探测器建成并开始取数后，离线软件负责产生模拟数据并处理实验数据从而进行物理分析工作。探测器模拟是整个离线软件系统的重要组成部分，它的可靠与否直接影响到最终物理结果的正确性和准确性。一个与真实实验符合较好的模拟系统是影响物理结果和降低系统误差的重要因素。探测器模拟是高能物理实验中消耗CPU时间最多的任务，海量模拟样本的产生总是受限于可用的CPU资源。因此，模拟软件的准确性和运行速度是关键性能指标。

探测器模拟包括产生子接口、探测器的精确描述、粒子在探测器中的传输和相互作用、探测器响应机制的研究、探测器运行的真实化模拟、调试与测试。目前国际上的高能物理实验大多采用通用的底层模拟工具Geant4来开发探测器模拟软件。作为离线软件的一部分，模拟软件框架要结合Gaudi和Geant4的流程来设计相应的算法，保证事例产生子、探测器输运、探测器响应三个部分的独立性。

1. 产生子接口：为满足物理分析对多种产生子的需求，产生子接口要能适应不同格式的事例产生子，例如HepEvt，StdHep， HepMC格式。
2. 几何管理：探测器设计阶段，多种方案并存，每种方案也会对几何进行多次优化设计以达到最优的性能，几何管理对探测器设计至关重要。要实现Geant4与通用几何管理工具DD4Hep的有机整合，有利于探测器几何的灵活配置；并保证几何在离线软件系统内多个环节的一致性，包括模拟、重建、事例显示。
3. 探测器响应和数字化：要结合具体的探测器类型深入研究各种相互作用过程以及探测器响应机制；未来的高能物理实验将采用一系列新型探测器，例如硅径迹探测器、时间投影室、高颗粒度量能器、契伦科夫探测器等，每种探测器的响应机制不同，这是模拟软件中最重要也是最难的部分，直接关系到模拟的准确性。
4. 真实化模拟： 为尽可能保证模拟样本的准确性，要开展束流本底相关的模拟研究，并开发本底混合算法将信号与本底进行混合。
5. **MC Truth**：结合物理分析的需求，调整MC Truth信息记录的策略，结合离线软件框架(例如Gaudi)以及Geant4的软件框架设计灵活的接口，例如采用Gaudi框架中的Tool工具，在模拟的不同阶段(例如Event, Track, Step)记录MC真实信息。
6. 并行化模拟：相比于现有的高能物理实验，未来高能物理实验探测器结构更复杂，颗粒度更高，探测器模拟将消耗更多的计算资源，包括CPU时间和内存。综合考虑异构资源上的模拟软件开发，例如采用GPU对模拟进行加速，Geant4的相互作用例如标准电磁作用过程等需要基于CUDA重新开发，并重新设计几何和数据结构，以适用GPU上的数据处理；现代CPU体系结构的发展趋势是核数增多，同等内存条件下每个核平均可利用的内存有限，开展多线程模拟研究，不同线程之间共享几何和相互作用模型，降低总体内存使用量。
7. 快速模拟：快速模拟是另一种加速模拟速度的方式，针对不同类型的探测器，使用参数化、机器学习等方式替代全模拟，在精度满足物理分析要求的情况下，提高时间运行速度。模拟框架应提供全模拟和快速模拟的灵活配置，并能够在一个事例的模拟中混合使用全模拟和快速模拟。基于参数化的模拟方法通过对特征量的概率密度函数进行抽样产生最终的探测器响应，概率密度函数除了使用公式描述，还可以使用预先产生的直方图替代。需要通过对称性、机器学习等技术压缩这些直方图数据，从而保证快速模拟中可以加载。研究机器学习等技术在模拟中的应用，特别是基于生成式对抗神经网络(GAN)的机器学习方法，可作为快速模拟的研究方法之一。

参考文献：

HSF CWP: <https://arxiv.org/abs/1712.06982>

The Fast Simulation Chain for ATLAS: <https://iopscience.iop.org/article/10.1088/1742-6596/898/4/042016/pdf>

The Parameterized Simulation of Electromagnetic Showers in Homogeneous and Sampling Calorimeters: <https://arxiv.org/pdf/hep-ex/0001020.pdf>

## 重建

现代高能物理实验的事例重建是一项非常复杂、具有挑战性的任务，把感兴趣的物理事例通过探测器中记录的末态粒子击中信息再现出来提供给物理学家进行下一步分析。精确测量末态粒子的四动量要求优秀的探测器硬件与性能优良的重建软件缺一不可。伴随着现有和未来高能物理实验的发展，数据的规模和复杂度不断提高；与此同时，摩尔定律的逐渐失效造成计算资源日益无法满足数据处理与分析的需求。如何更加高效地利用数据资源获得科学产出在未来将成为较为突出的矛盾，毫无疑问，数据的处理是其中最为关键的环节之一，新的软件研发和升级为充分利用计算资源，采用新的并行计算，异构计算等更高效的数据处理策略提供了可能的选项，因此被业界寄予厚望。

基于旧的计算架构或过时的软件工程思想的重建软件，对计算资源的消耗是十分惊人的；相反，性能良好的软件通过算法改进与计算资源的特点与优势相结合，改进计算和存储策略等，在有限资源投入的前提下，能够大幅度提高数据处理速度，为物理目标的实现提供技术驱动。为了实现上述利用先进硬件和技术的目标，大量的高能物理软件和算法需要重新设计和代码实现。必须清醒地认识到，在带有预见性的人力投入预期下，这项研究任务无法由我国科学家独立完成。欧洲核子中心同样没有足够的人力完成HL-LHC和FCC-ee所需的软件开发任务。在这样的形势下，2019年建立了Key4HEP项目，联合规划中的高能量前沿实验ILC，CLIC，FCC，CEPC，LHC，SCTF和HSF等实验和项目，致力于联合开发适用于未来高能量前沿实验通用的软件站。Key4HEP软件使用抽象的架构，针对各个具体的实验和探测器设计，仅需要最低程度的维护和调试得到高水平的性能。

重建软件主要包括带电径迹的重建和量能器的重建，主要包括一种通用的径迹软件（ACTS）和粒子流算法（如PandoraPFA，Arbor等等）。软件通过抽象的架构把探测器概念与软件流程、算法、工具、服务和应用分离，实现软件良好的通用性和可移植性。在应用层面，需要根据探测器和物理过程的特点对流程和参数进行调节，获得最佳的性能。

径迹探测器的重建与量能器的簇射寻找算法中都需要利用模式识别技术，近些年兴起的机器学习和人工智能技术在处理这一类复杂问题中已经表现出优异的性能。同时高能物理实验将获取高统计量的真实数据样本，这也为机器学习等算法的提供了训练样本。利用机器学习和人工智能等先进技术，进一步挖掘探测器固有潜力，联合使用各个子探测器的信息，进一步优化探测器性能，对促进高能量实验的科学产出具有重要意义。

## 分析工具

数据分析工具是终端用户得到结合了探测器和软件性能的最终数据样本以及最终物理结果的直接依赖，鉴于加速器亮度、探测器的粒度和数据维度的增大，数据分析工具随着大数据和机器学习领域的协同发展成为必然发展趋势。下面描述我们可能需要重点发展和培育的数据分析工具发展方向。

1. 可视化的数据处理:今年来计算机视觉的发展迅猛，但在于高能物理数据分析领域，因为物理和探测器的复杂性，数据本身有几个特征：纬度高、复杂度高、本底高。如果先基于一些简单的预处理，比如主成分分析（PCA）降维之类的，再比如本底预减除等方案，再根据进行可视化，这样有可能大大提高对于数据主要特征变量选择效率，从而提高数据分析的进程。
2. 方便使用的机器学习框架接口用于事例分类、统计估计和假设检验等。机器学习行业已经发展出各种比较成熟的机器学习(深度学习)框架和平台，主要用于满足该领域计算机视觉、自然预研处理和各种自动控制的需求。而对于高能物理，需求基本类似，仅就数据分析环节，侧重方面稍有不同。目前数据分析中对于机器学习的基本需求是事例分类，如果拓展到重建环节，除了分类外，模式识别、回归等的需求更为广泛。从发展分析软件工具角度来说，针对高能物理需求较为密集的几种场景把市场上的各种机器学习框架进行针对性的包装，提供相应的数据输入、训练模型的的自动部署等功能会大大提高数据分析的效率和效果，亦即更好的发现灵敏度和测量精度。
3. 声明性分析:常规数据分析方式是在一定数据分析框架、模拟产生和数据管理规范下，通过编写程序得到初选的 Ntuple 和直方图等“预选”数据在 root 等分析工具下，独立的调用各种分析、计算和绘图工具,进一步优化事例选择，并在最终进行统计估计和假设检验得到物理结果。声明性分析主要特征是基于对于数据和常用算法的良好封装，对终端用户隐藏非常重要的、但和物理无关的信息，比如某批次的数据的文件位置和数量等，或者某个子算法的实现细节。这样用户可以用简单、明确的描述进行数据分析，比如:
   1. Analyze RUN XX data to obtain the invariant mass distribution of Z  e+e- process (分析 XX 批数据得到Z 到两轻子末态的不变质量)
   2. Filter all RUN XX data and save all jets information when at least 2 jets and when the Pt>20 GeV to ntuple (对 XX 批次数据进行初选并保存横动量大于 20GeV、至少两喷注事例的全部喷注信息到 ntuple)

因为主要容易出错的信息被封装的对用户友好，而经过大量测试的成熟算法具有非常简单的接口可以灵活被调用，从而大大减少终端用户的出错几率和工作效率，从细节和重复性操作中解放时间和精力，从而可以研究更多、更深刻的物理课题。

## 数据可视化技术

在高能物理实验中，数据可视化技术在多方面有重要应用，主要包括：探测器设计与优化、探测器模拟、事例产生、数据质量监控、事例重建、事例显示、物理分析，以及整个高能物理领域的科普、教育工作与对外宣传等。

如何更好地展示高能物理实验中的数据，比如探测器几何数据的显示、事例数据的显示、物理分析数据的显示，如何开发更通用、集成度更高、跨平台性能更好的数据显示工具，是未来十年内下一代高能物理实验界在软件和计算方面需要考虑的问题。基于此，国际高能物理界对未来数据显示技术的发展和相关软件的开发近年来有一些相关的讨论和共识[1][2]，以下将就几个方向分别讨论。

1. 探测器几何描述与可视化:探测器几何描述根据其详细程度可分为三级:
   1. 模拟级别几何，包括灵敏探测器信息与支撑结构；
   2. 重建级别几何，只包括灵敏探测器如量能器晶体、漂移室的信号丝等;
   3. 提供显示功能的简化几何。

目前提供几何描述的基本框架包括，ROOT，Geant4以及计划为未来高能对撞机提供几何描述的DD4HEP等。未来的探测器几何描述应致力于提供同类型实验可通用的几何描述工具，有利于不同探测器设计方案的共享，同时提供与工业界通用的几何描述格式相应的接口，比如将Geant4的几何描述转化为VRML或FBX格式的几何格式，从而能够被工业界的几何显示软件所支持和使用。

1. 事例可视化:事例显示需要解决两方面的问题：事例信息的获取和显示平台的选择。目前实验中事例显示工具一般分两种模式：一种是内嵌于实验软件框架的事例显示，另一种模式则是独立于软件框架的事例显示。内嵌模式的优点是能够直接获取事例和几何最详细的信息，缺点是可使用的显示软件不多，比如ROOT，并受限于整个软件框架的安装，需要与服务器通讯，单独将事例显示功能移植出来比较困难，需要在高能物理实验方面比较有经验 的专业用户才能开发和使用。独立模拟则是将几何数据与事例数据导出为中间格式，从而可基于其他更多元化的工业软件开发事例显示，比较方便为领域外非本专业人士开发并用于教育、宣传等，缺点是有一层额外的中间数据转化，在事例数据比较 庞大时，不容易筛选出感兴趣的事例进行分析。未来事例可视化的发展方向，应尽量支持与工业界软件的互通，特别应注意工业软件的长期性与稳定性，避免其开发和维护终止所带来的额外工作。同时提高与虚拟现实、增强现实等工具的融合，从而使其更方便地服务于高能物理专业分析，以及教育、科普、宣传等方向。
2. 数据处理分析可视化:高能物理实验中数据处理和物理分析的显示，目前一般是由基于C++的ROOT完成的。虽然ROOT是专门设计为高能物理和核物理数据处理使用，到目前为止也已经很好地提供了实验数据分析所需的各种功能，但是在领域外的应用不多。Python是今年来逐渐流行的编程语言，在机器学习等领域有广泛应用。同时，Python也提供了很多具有更强显示功能的数据可视化软件。将Python与ROOT相结合的PyROOT，能够在数据分析和显示两方面具有一定的优势。未来数据可视化的发展方向之一，是将数据的处理分析与数据显示功能逐渐剥离开，实现数据内容与显示风格的弱耦合，从而可以使用更广泛的应用以增强显示功能与效果。另外，由于基于机器视觉的深度学习技术，比如卷积神经网络的发展，实现事例数据处理和分析的可视化，从而利用人工智能辅助实验中的事例重建、物理分析等工作，也是数据可视化的一个重要发展方向。
3. 显示平台与设备:由于历史发展的原因，高能物理实验一般基于Linux平台完成。科研人员习惯于在台式机或笔记本电脑环境下工作。随着移动设备、网络技术的发展，对跨平台下的数据处理和显示，也提出了更高的要求。

未来高能物理数据可视化平台的发展方向，一方面需要支持不同的操作系统，比如Linux、Windows、OSX、iOS、Android、网络浏览器模式、VR平台等不同的平台；另一方面，也要考虑对不同设备，如手机、平板、可穿戴设备等的支持。目前，在这方面已经有一些尝试，比如通用的Qt3D，OpenGL显示引擎，Unity 游戏开发引擎，Unreal Engine等，能够通过API接口与部署，部分实现跨平台的功能，同时通过与穿戴式设备如Oculus，HTC Vive，Lenovo等的接口，实现跨设备的虚拟现实和增强现实显示功能。

在国际交流合作方面，中国应更积极地参与计算和软件相关的国际合作，参加相关的国际会议如HEP Visualization Workshop，不仅在高能物理的实验层面、还要在软件计算技术方面，与其他学科如地球物理、天体物理等领域的研究者加强共享与交流，以实现最新最好的数据可视化功能，服务于我们目前所参与的以及规划中的各个高能物理实验。

参考文献：

1. Bellis, Matthew, and et al. Mon . "HEP Software Foundation Community White Paper Working Group – Visualization". United States, arXiv:1811.10309 [physics.comp-ph].

2. Antonio Augusto Alves Jr et al. A Roadmap for HEP Software and Computing R&D in the 2020s. Tech. rep. HSF-CWP-2017-01. HEP Software Foundation, 2017. arXiv:1712.06982 [physics.comp-ph].

## 高性能计算与网格

目前，LHC实验大约有EB级的数据量，磁盘和磁带的存储大约各占一半。未来HL-LHC 阶段由于事例率和事例大小的显著增加，产生的数据将是目前数据量的10倍以上。针对数十EB数据量给高能物理计算带来的前所未有的挑战，我们将在LHC软件和计算白皮书的框架下，对数据中心架构、数据存储、云计算、高性能计算、分布式计算以及网络等的关键新技术开展研究。

**数据中心架构：**随着网络带宽的快速增长，未来网络带宽将趋于无限，从战略和运营的角度来看，大科学装置数据中心将迁移到运营成本更低的地方，比如欧洲核子研究中心CERN在匈牙利布达佩斯建立了第二个数据中心。对于建立一个新的数据中心，必须制定明确的目标。其中包括低于1.1的PUE（一种能源效率指标）、每机架15-20千瓦的相对较高的功率密度、模块化设计使得分级功率随时间增长（比如从最初的4 MW到后期的12-16 MW）。未来的增长率和设备更换周期的需求也必须根据预期的技术演变进行重新评估。此外，增加使用专用硬件来运行计算密集型任务的趋势要求未来数据中心能够适应不同种类的设备

**数据存储技术：**目前大科学装置的存储领域已经进入EB级时代，随着HL-LHC等装置的运行，未来10年左右将达到ZB级别。同时，由于云计算、并行计算、异构计算等计算模型的发展，将在数据规模、访问模式、访问速度等方面面临巨大的挑战，需要研究和设计新的存储模型，主要包括：1）新兴的可计算存储（Computational Storage）将部分计算能力下沉到存储设备，将有利于解决大数据带来的IO瓶颈问题；2）网络带宽的快速增长，使得CPU和数据可以不在一个站点，利用数据缓存技术实现远程文件直接访问，构建全球数据统一命名空间，比传统的数据传输系统更加灵活和高效。3）随着数据容量和文件数量的增加，面向对象的存储方式将成为未来构建大规模存储系统主要的发展趋势之一。这要求物理软件适应新的数据访问方式；4）基于事例的数据流比传统的基于数据集或文件的数据访问粒度更细，更有优势，发展很快；5）存储技术的另外一个趋势是异构，采用人工智能驱动的策略存储支持内存、SSD、磁盘、磁带、本地站点、远程站点、共有云等位置的透明迁移和备份，实现高性价比和高性能的统一存储。

当前的大科学装置海量数据存储系统由少量固态硬盘、大量机械硬盘和海量磁带组成，固态硬盘主要作为快速数据分析存储，机械硬盘是主力存储支撑大规模计算，磁带作为近线存储设备，需要时可直接访问。由于实验数据分析对于数据访问性能需求量剧增以及业界存储介质的发展，固态硬盘将大量代替现有机械硬盘作为主力存储，现有的机械硬盘将出现单盘100TB以上的容量，逐渐变成近线存储，而磁带逐渐变成单纯的备份归档设备。因此，在大规模数据存储设计过程中，将研究适应未来存储介质的容量增长模型，充分考虑存储系统的容量、性能及成本，并在新的分布式计算环境中整合已有数据分析工具链，保证系统的平滑升级。

**云计算：**在最近几年，国内外很多大科学装置数据与计算平台都引入了云计算和虚拟化技术，特别是中国科学院信息化专项支持了多个科技领域云，越来越多的资源都加入到各大装置的私有云平台中，采用Openstack等云资源管理软件进行有效的组织和管理，比如CERNCloud管理的CPU核达到25万颗以上，并且每年还在新增成千上万个CPU核。未来十五年，大科学装置的基础设施规模继续扩大，将是现在的50到100倍，还将出现各种新型异构的计算平台。为保证系统规模的高可扩展性，云计算技术和资源使用策略会出现新的发展。一方面，通过相同的用户界面和管理系统提供虚拟机、容器和物理节点来简化用户体验的研究，目标是为用户提供更多潜在的服务，例如文件系统即服务、秘钥即服务、物理节点管理、针对所有计算资源的单一库存等。总之，简化用户体验，为用户选择硬件资源方面提供更大的灵活性是云计算目前乃至今后的发展趋势。

**高性能计算技术：**向量寄存器的宽度增加、CPU核数的增加、GPU、协处理器以及众核等新硬件技术的发展，使得大型软件的优化更为复杂。将程序代码模块化，更易于开展有针对性的代码优化，以更好发挥硬件的优势。降低存储时延，应用并发处理模块，采用模准C++库，软件性能分析工具，软件质量自动化检查等方面都可提高软件运行效率。

分布式计算技术可将大型计算任务分布在不同的站点上执行，但目前大都是在同构的计算环境中运行。随着计算需求的不断增加，大科学装置的数据处理需要寻求更多的计算资源，这要求计算任务可以无缝应用于不同的硬件架构。所以高性能计算将向异构平台和架构多样化发展。除了传统的X86架构的CPU外，GPU、TPU、FPGA、众核以及将新出现的硬件技术会被更广泛的应用，特别是针对一些特定的计算场景，这类技术的应用可极大提升计算性能。将代码模块化以最大限度利用GPU、MIC等硬件的优势是应用软件发展的趋势。这需要更为系统的评价计算与数据处理的工作流，研究新技术如何应用于计算任务的最适合的环节以及如何将各个环节连接。在CPU级别开发的可被共用的算法，采用容器等轻量级的虚拟化技术实现代码优化，针对一些算法设计专用的GPU等硬件都可极大提高软件性能。

**分布式计算：**随着研究目标不断的深化和对精确度的提高，高能物理实验计算的数据和规模将每十年数十倍的增长（按照高能物理实验LHC近10年的增长趋势预测）。与之对应，存放和处理这些数据的存储系统和计算系统也会出现同等规模的增长。除了规模化外，在未来的五年到十五年，分布式计算还将面临的挑战和变化：分布式计算运维的成本不断增加；网络带宽不断提高将使我们重新考虑更有效的计算模型；资源的异构性更为明显，不同处理器体系结构的混合型将是趋势（CPU、GPU、FPGA、量子计算以及后续可能出现更新的体系结构）；未来处理器的处理速度增长将超过存储设备的容量的增长速度，需要充分考虑到存储资源的成本等。

**网格：**高能物理研究所建立了ATLAS/CMS/LHCb 二级站点，作为WLCG网格的一部分，成功完成了数据存储、模拟数据产生、数据分析等任务，站点的稳定性位列所有二级站点前列。为了迎接Run3和HL-LHC的数据量的挑战，WLCG技术正面临新一轮的重大变革，本项目将紧紧围绕LHC实验的数据处理需求，参与WLCG 第三方传输协议、基于token的新型认证模式，数据联盟等新技术的研究和应用。 另外，本项目将研究统一的多实验支持的分布式计算平台，通过虚拟化技术消除底层硬件和系统的差异，构建统一资源池；基于分布式计算技术实现对来自全球的异构异地资源有效地整合和实验数据共享，并根据实验需求进行合理调度，使得资源和数据可以为实验所高效利用和共享。最终形成高可管理、高可扩展的统一的大型计算平台，成为中国LHC实验研究的重要计算设施。同时随着各种新型异构计算平台的出现以及高性能计算资源的需求愈发明显，还将研究如何通过统一的平台进行异构计算平台和高性能计算资源的融合，实现用户需求和资源的精确匹配。

近10年来，虽然计算机芯片中的单位面积的晶体管数目还在遵循摩尔定理增加，但能耗限制了计算机的时钟速度等速增长，这限制了单个CPU核的运算速度的增长。因此，通过并行计算来提升软件运行速度已经成为计算和软件的未来发展趋势。当前，高性能计算技术不断有新的突破，为社会以及科学的各个领域带来了深刻影响。一方面超级计算机/计算集群规模也越来越大，应用场景越来越多；另一方面， 新型的GPU等异构设备相较CPU具备了更强的计算能力，且使用范围日趋广泛；。异构平台架构的优化整合、大规模细粒度的资源管理、并行计算软件应用以及大数据分析处理等方面的研究已经成为高能物理计算领域的研究热点问题，对这些新技术的研究可以挖掘高性能计算技术潜力，将能缓解LHC等实验所面临的巨大计算压力，有效促进高能物理持续发展。

**网络与安全：**随着LHC实验数据量的快速增长，以及全球大科研模式的形成，现有的网络体系结构在数据交换效率、网络质量保障、网络信息安全以及网络资源调度等方面存在的问题也逐渐凸显。网络与安全将深入研究网络的高效性、安全性、可扩展性及灵活性，提升网络服务质量和数据交换效率，为保障LHC实验数据共享和处理工作的顺利开展有重要意义。

## 数据和软件的持久化

数据与软件长期保存指在实验开始后直至实验结束后的数十年内，通过各种技术手段，保证实验数据能够被访问，并且实验数据的处理环境（软件、文档、元数据）能够被还原，对数据的处理过程以及物理成果生成过程能够被重复。其意义在于：随着理论研究的进步和分析手段的提高，可能从旧实验数据得到新科学发现。对新旧数据的联合分析和交叉验证可提高物理结果的精度和可信度。对旧数据重新分析和结果对比，可验证新数据处理方法的正确性和性能。通过重复已有科学成果的产生过程，可培训新加入实验成员快速掌握实验处理方法。

## 软件开发环境

高能物理实验软件的管理和质量控制非常重要，在实验合作组中定义明确的软件开发模型，规范化标准的软件开发流程。特别是软件的生命周期比实验运行的生命周期更长，涉及实验研制阶段、实验运行阶段和实验停止后持续时间更久的物理分析阶段，因此面临极大的挑战，需要开发可长期维护的软件。现代的开发更加国际化，开发人员分散在世界各地贡献代码，支持分布式的开发和协作将愈发重要。利用持续集成和持续部署等先进的软件开发方式对开发人员提交的代码进行自动化构建和测试，从而尽早发现错误。随着实验使用软件包数量的增加，软件包的管理也更加重要，特别是软件安装和部署时的兼容性、可重复性。软件的测试和检查也非常关键，需要涵盖各个层次，例如单元测试、集成测试、大规模数据产生测试。

借鉴工业界软件开发和部署的方式，使用对实验软件开发有用的工具和经验。特别是考虑以下几个方面。

1. 软件管理和质量控制。开发模块化和可重用的组件、软件包。可长期维护性。考虑软件、硬件平台的影响。考虑采用容器等技术支持不同版本的操作系统。
2. 分布式开发模型。考虑使用合适的工具进行版本控制，例如采用Git进行分布式协作。同时，使用相应的平台实现问题汇报和跟踪，例如GitHub和GitLab等平台。考虑软件仓库的备份和安全性。
3. 软件部署。考虑使用合适的包管理工具，例如Spack。考虑自动化的编译和部署，尤其是在CVMFS文件系统的软件部署。
4. 软件测试和检查。注重测试用例的开发。考虑软件性能的测量和优化。

## 经费支持和配套措施