基于国产处理器架构的高能物理数据处理系统

程耀东[1,2,3]，程垚松[1]，毕玉江[1,3]，高宇[1,2]，李海波[1]，汪璐[1]，姚秋玲[1]
1. 中国科学院高能物理研究所，北京 100049；2. 中国科学院大学，北京 100049；
3. 四川天府新区宇宙线研究中心，四川 成都 610213

摘要
随着规模的不断扩大，高能物理实验产生了越来越多的科学数据，迫切需要先进的数据处理系统来支撑科学研究。目前，以ARM架构等为代表的国产处理器发展迅速，高能物理数据处理系统面临着新的机遇与挑战。首先总结了高能物理数据处理系统的需求及体系架构；然后描述了在国产处理器上开展的高能物理数据处理软件移植等相关工作，并提出了一种新的面向高能物理数据处理的可计算存储技术方案；最后给出了在国产处理器架构上的典型应用评测结果。

关键词
国产处理器；高能物理；数据处理；可计算存储

中图分类号：TP316 文献标识码：A doi: 10.11959/j.issn.2096-0271.2021046

Data processing system for HEP based on domestic processor architecture

CHENG Yaodong[1,2,3], CHENG Yaosong[1], BI Yujiang[1,3], GAO Yu[1,2], LI Haibo[1], WANG Lu[1], YAO Qiuling[1]
1. Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. TIANFU Cosmic Ray Research Center, Chengdu 610213, China

Abstract
More and more scientific data are produced by fast-developing high energy physics (HEP) experiments, which urgently require advanced data processing system to support scientific research. At present, HEP data processing system is facing new opportunities and challenges with the rapid development of domestic CPU such as ARM architecture. Firstly, a brief introduction to the requirements and architecture of HEP data processing system was given. Then the relevant work such as porting software to domestic CPU architecture was described. Additionally, a cutting-edge computational storage technology for HEP data processing was proposed. Finally, the evaluation results of typical HEP applications on domestic CPU architecture were given as well.

Key words
domestic CPU, high energy physics, data processing, computational storage

1 引言

随着装置复杂度的不断增加和规模的不断扩大，高能物理实验产生的实验数据越来越多，海量数据处理在计算规模、计算精度、即时性等方面的需求也越来越高，给传统计算体系架构带来前所未有的挑战，全球高能物理领域都在积极探索和研究最新的解决方案。ARM（advanced RISC machine）多核架构由于其自身的灵活性和自由性，逐渐成为业界研究的热点。近年来，以ARM为代表的国产架构服务器异军突起。ARM早期专注低功耗领域，在移动端处于领先位置，生态体系已经十分完善。随着多核异构计算时代和场景多样化计算时代的到来，国内服务器行业端生态逐步完善，以ARM为代表的国产架构服务器快速发展。为此，研究和开发基于国产多核架构的高能物理计算环境及软件有助于实现高能物理数据处理系统的自主可控及技术创新，从而促进高能物理计算架构演进，并加速科学发现。

当前，国内外高能物理实验的数据处理平台以x86 CPU架构为主。同时，图形处理器（graphics processing unit，GPU）、现场可编程逻辑门阵列（field programmable gate array，FPGA）、张量处理单元（tensor processing unit，TPU）等异构计算设备也开始受到重视，并被应用到高能物理数据处理系统中[1]。结合高能物理的数据处理需求以及IT的发展，本文基于国产处理器及服务器等硬件，建设了高能物理数据处理系统，包括系统及平台软件、基础应用软件框架以及应用软件等。此外，本文提出了面向高能物理数据处理的可计算存储技术架构，基于ARM和FPGA构建存储节点，修改数据分析框架软件ROOT[2]及数据存储软件EOS[3]等，把计算任务从计算节点卸载到存储节点，避免数据搬运，实现了绿色节能、运算高效的数据处理模式。

2 研究背景

2.1 高能物理实验

高能物理研究组成物质的基本粒子及其相互作用规律，高能物理实验是研究高能物理的重要手段。当前，高能物理实验的规模通常很大，需要成百上千位科学家参加，同时会产生海量的实验数据，一个大型实验往往产生PB级甚至EB级的数据。例如，目前大亚湾核反应堆中微子实验已经累积了2 PB的实验数据；北京正负电子对撞机重大改造工程（BEPCII）已经累积了10 PB的实验数据，并且数据量还在不断增加；江门中微子实验（JUNO）预计在2022年开始取数，每年将产生3 PB的原始实验数据；高海拔宇宙线观测站（LHAASO）边建设边运行，目前已经累积了近10 PB数据，预计2021年完全运行以后，每年将产生8 PB以上的原始数据；高能同步辐射光源（HEPS）一期建设的15条光束线实验站预计平均每天产生200 TB的原始实验数据，峰值可达每天500 TB；在欧洲的大型强子对撞机（LHC）升级改造后的HL-LHC阶段，仅ATLAS探测器的数据量就将是目前的10倍以上，在2030年左右将超过3 EB/年，计算量增长60倍以上。这实际上已经超出了目前信息技术的处理能力，迫切需要突破新的技术[4]。

因此，高能物理实验产生的海量实验数据需要借助先进的计算机技术进行处理和分析，同时实验的需求也助推了信息技术的不断发展，比如万维网、网格计算与云

计算以及大数据处理等。

2.2 离线数据处理流程

粒子在高能物理实验的探测器中的运动过程被捕获，产生了大量的电子学信号。然后，通过触发判选和在线选择的事例，由在线数据获取系统（data acquisition，DAQ）以二进制文件的形式记录下来。这种数据被称作原始数据，主要包含探测器电子学信号的时间和幅度信息。通过高速以太网，原始数据文件被传输到磁带库永久保存。对原始数据进行刻度和重建后，生成重建数据，供物理分析使用。

离线数据处理和物理分析的简化过程如图1所示。原始数据经过离线刻度，能够消除实验的各种外部条件（如温度、气压）和探测器本身条件（如探测器高压）对电子学信号与物理测量之间转换关系的影响。重建是离线数据处理的核心，数据重建算法使用刻度算法产生的刻度常数，将探测器记录的原始数据转化为粒子的动量、能量和运动方向等物理量，生成重建数据。物理研究还需要产生与真实数据数量相当的模拟数据，这部分数据也要进行重建。和原始数据一样，所有重建数据会被保存在磁带库中。物理分析人员利用物理分析工具（如运动学拟合、粒子衰变顶点寻找和粒子鉴别等软件）分析重建数据，得到物理研究结果。数据处理过程主要包括模拟计算、事例重建以及物理分析3种计算类型。

2.3 离线数据处理环境

高能物理探测器产生的原始数据经过复杂的处理后，被转化为可用于物理分析的数据。实验数据的处理和分析都是在离线计算环境中进行的，包括数据存储、数据传输、数据分析等。典型的高能物理计算环境如图2所示，其核心是一个高速、高可靠的网络，其余子系统连接到这个核心网络上，包括前端登录集群、海量存储系统、计算节点集群、备份与分级存储系统、管理系统等。不同的子系统具有不同的功能和配置，功能上相互独立，整体上协同工作[5]。

海量存储系统包括磁盘存储和磁带存储等，分别采用EOS、Lustre、CASTOR等存储软件进行管理。其中，EOS是高能物理领域常用的分布式存储系统之一，通过XRootD协议透明支持ROOT等数据分析框架。计算节点集群由大量的工作节点组成，通过作业管理系统（如HTCondor等）

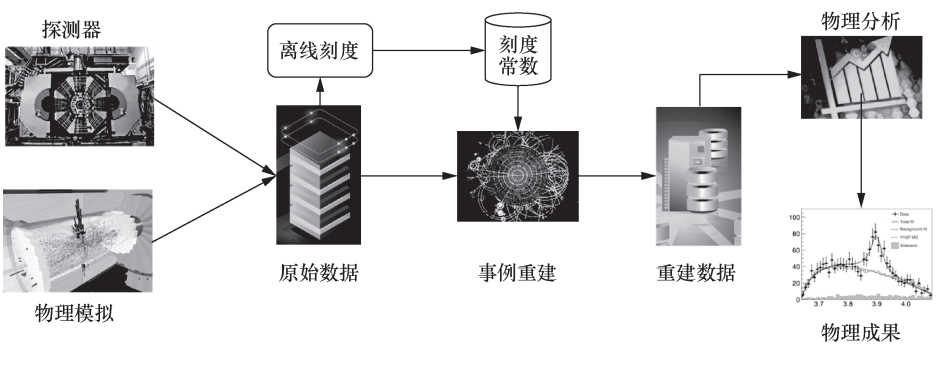

图1 高能物理数据处理的基本流程

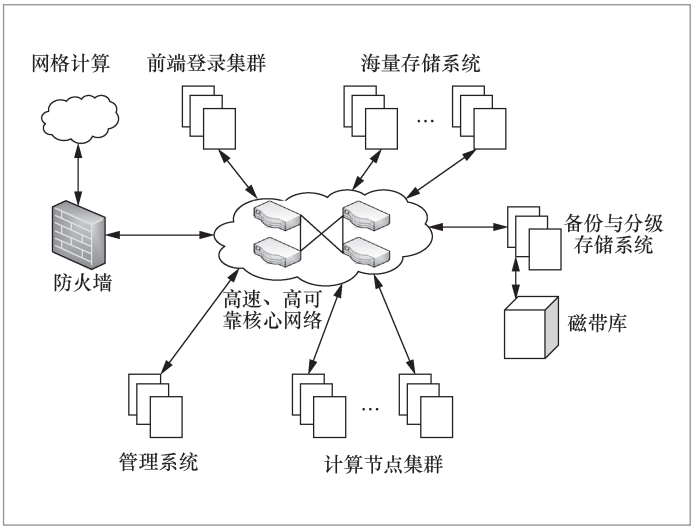

图 2 典型的高能物理计算环境及其组件

形成统一的资源池。工作节点是最终运行应用的机器，由CPU、内存、硬盘、网卡等硬件和操作系统、应用程序等软件组成。网格计算将分布在全球的计算资源统一管理起来，是“集群之集群”，形成更大的资源池。管理计算环境的这些系统软件通常被称为“系统软件”，目前大部分运行在x86架构上。在这种计算与存储分离的架构中，计算节点需要从存储节点读取数据进行计算，计算结束后再将输出结果写回存储节点，从而导致数据的大量搬运，这成为目前数据处理的主要瓶颈之一。

2.4 相关工作

由于ARM等架构的快速发展，高能物理应用软件开始从x86架构迁移到其他架构。从1978年启动首台巨型机“银河-I”研制到2010年“天河一号”首次摘下全球超级计算机500强榜单第一名再到今天，我国的超级计算机经历了40多年的“超常速”发展。随着我国超级计算机的发展，相应的软件、算法及优化方法也在不断跟进。2016年，运行于“神威·太湖之光”之上的应用“千万核可扩展大气动力学全隐式模拟”获得戈登贝尔奖，实现了我国在此大奖上零的突破。何晓斌等人[6]对面向大数据异构系统的神威并行存储系统展开研究，经过优化的系统使得某些应用获得10倍以上的性能提升。胡正丁等人[7]研究面向异构众核超级计算机的大规模稀疏计算性能优化问题，为高能物理格点量子色动力学（lattice quantum chromo dynamics，LQCD）等应用提供借鉴。高能物理研究所与合作单位针对“神威·太湖之光”超级计算机的主从核架构，自主开发了LQCD中的D-slash热点程序的申威版本，并针对申威架构进行了优化，然后将其集成到国际通用的LQCD开源程序Chroma中，形成了能够在神威机器上大规模运行的完整软件系统[8]。该软件在神威机器上实现了32 768个申威处理器、约852万核的大规模消息传递接口（message passing interface，MPI）并行计算，取得了良好的应用效果。

在国外，大型强子对撞机的数据处理正在经历重大变化，欧洲核子研究中心（CERN）启动了相关项目来移植优化软件，以应对LHC Run 3的数据处理。CERN计划将LHCb堆栈从x86_64体系结构移植到AArch64（ARM架构）和ppc64le（PowerPC架构）两种体系结构[9]，旨在评估高级触发器（high level trigger，HLT）的计算基础架构的性能和成本。在所有软件包中，最大的挑战是向量化的日益广泛使用。目前，许多向量化库专用于x86架构，并且不支持其他架构。尽管存在这些挑战，CERN仍已成功将LHCb高级触发器代码移植到AArch64和ppc64le。根据在LHCb上进行的测量，与x86架构相比，使用ARM架构的物理结果在可接受的精度上数值正确，较小的差异可能是由舍入误差和体系结构中位数不同（ARMv7和x86 64分

别为32位和64位）引起的。虽然现代ARM处理器的功能仍然不如传统x86处理器强大，但是就能效比而言，ARM的表现更好。CERN近期的一个报告显示，整个LHC计算网格（LHC computing grid，LCG）堆栈都可以基于AArch64构建，并计划将ARM版本的LCG堆栈安装到共享文件系统CVMFS（cern virtual machine file system）[10]上，提供给全球的用户使用。

3 基于国产处理器的数据处理系统

3.1 系统组成

基于国产处理器及服务器等硬件，通过移植和开发相应的系统软件和应用软件，笔者团队构建了高能物理计算环境及数据处理系统，其组成如图3所示。

在图3中，底层是国产处理器及服务器硬件，如鲲鹏处理器、飞腾处理器、申威处理器以及泰山服务器、“神威·太湖之光”等超级计算机等。在此之上安装系统软件与计算平台软件，包括CentOS等操作系统，HTCondor、Slurm等作业调度软件，Lustre、EOS、CVMFS等数据存储软件等。基于国产硬件和系统软件等运行环境，开发和移植基础应用软件与框架，包括ROOT、GEANT等粒子物理模拟与分析软件，Chroma等理论物理计算软件，Gromacs、NAMD等分子动力学模拟软件。在最上层，支持高能物理实验及应用软件，包括LHAASO、BES、JUNO等粒子物理和天体物理实验，LQCD等理论物理应用，HEPS、CSNS、纳米生物等多学科应用。

3.2 关键软件的移植

国产处理器及服务器通常有相匹配的操作系统及编译器，到目前为止，大部分常用的高能物理基础软件和应用软件

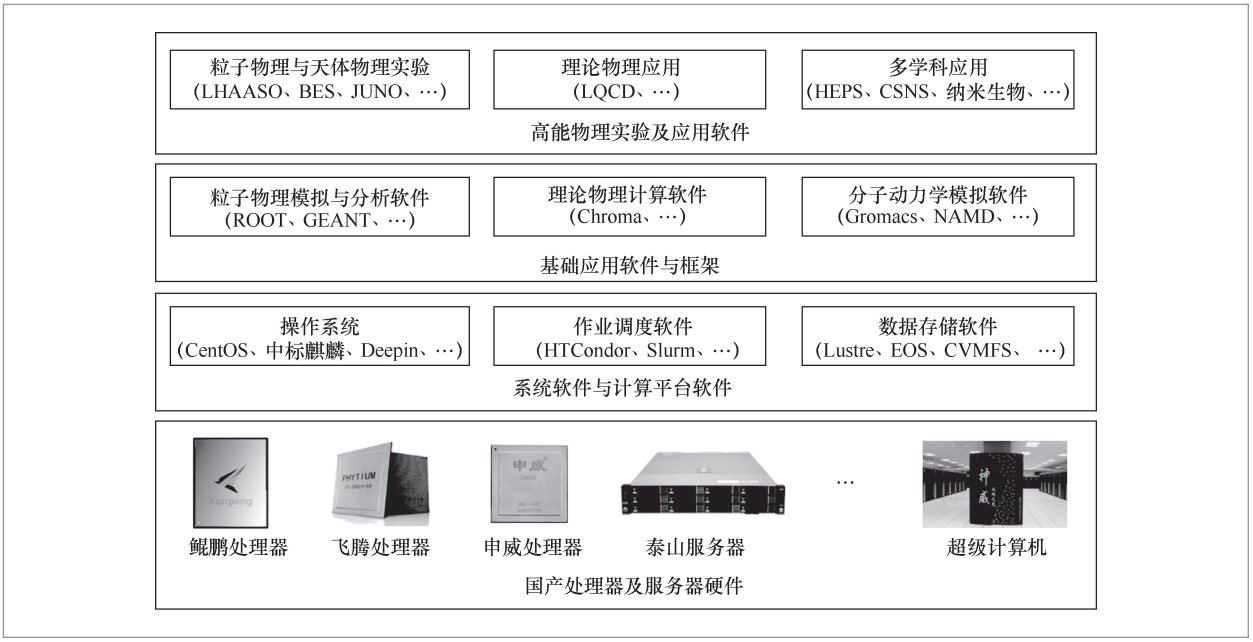

图 3 基于国产处理器的高能物理数据处理系统组成

已经被移植到ARM V8架构（AArch64）上运行，包括EOS、ROOT、GEANT、Chroma、Gromacs、NAMD、LHAASO数据重建等。其中，ROOT是由CERN开发的一个模块化科学软件工具包，提供了大数据处理、统计分析、可视化和存储所需的多项功能，是高能物理数据处理的基础。GEANT4是由CERN开发的一个蒙特卡洛应用软件包[11]，用于模拟粒子在物质中输运的物理过程，其主要应用领域为高能物理、核物理、加速器物理、核医学和太空科学等。这些软件大多基于C/C++编写，属于编译型语言。x86 CPU属于复杂指令集计算机（complex instruction set computer，CISC），ARM属于精简指令集计算机（reduced instruction set computer，RISC），另外，x86和ARM使用的向量寄存器也不同，向量指令集存在差异，因此原来在x86上开发的程序必须重新编译才能在ARM架构上运行。C/C++源码需要由编译器、汇编器翻译成机器指令，再通过链接器链接库函数生成机器语言程序。在软件移植的过程中，主要步骤包括：获取源码；准备GCC等编译环境；修改配置文件（configure或者CMakeList.txt），生成编译规则文件（Makefile）；重新编译或者替换x86平台的动态链接库；执行编译过程，生成可执行程序。在代码的移植过程中，重点需要修改如下内容。

- 修改C/C++代码工程的编译脚本和编译选项：以x86下的-m64代码为例，其主要功能是将应用程序编译为64位，对应到华为ARM上则采用-mabi=lp64的编译选项。此外，x86平台上默认的char类型是一种有符号的类型，对应到华为ARM上则是无符号类型。在移植过程中需要显示定义，并将char类型定义为有符号类型。
- 移植编译宏：编译宏的作用是让编译器知道编译哪些分支代码能够在不同架构下达到最优性能。x86代码上有些编译器自带自定义宏，比如与smd属性相关的宏在x86上是SSE开头的宏，对应到华为ARM平台上需要自定义它的编译宏和相对应的分支。
- 移植builtin函数：builtin函数是编译器自带的函数，其在实际迁移项目中相当常见，主要是CRC32校验值的计算，大部分需移植的builtin函数集中在SSE intrinsic函数内。
- 移植内联汇编函数：第一种是指令替换，x86上对应的是BSWAP指令，ARM上对应的是rev指令，其他操作和寄存器是基于内联汇编的语法规则进行替换的。第二种是builtin函数的替换，以x86的指令popcount为例，其是对二进制数里面的1进行计数，对应到ARM平台上替换的是popcountll。
- 移植向量指令函数：SIMD是一种单指令处理多数据流的并行处理技术，能够在批量数据操作时进行向量化运算加速，具有较高的执行效率。Intel的SIMD技术实现包括MMX、SSE、AVX等。ARM的SIMD技术主要通过开源的NEON库等来实现。

根据以上规则和方法，笔者团队成功地将ROOT 6.20、GEANT4 10.6、EOS 4.7.7移植到华为鲲鹏920处理器及CentOS 7.6上，并部署到高能物理计算平台上，提供给用户使用。举例来说，EOS是CERN采用C++开发的一套分布式存储系统，依赖于XRootD、sparsehash、ncurses、Protobuf3、ISA-L、Folly C++ library、isa-l_crypto、RocksDB、c-ares等数十个开源软件包，需要提前找到这些软件所需的版本，并进行编译，最终生成应用软件包。而且，在EOS中还使用了汇编语言来处理数据CRC校验等操作。为了保持代

码的一致性，笔者团队重新定义了相关汇编指令，具体如下。

```
#ifdef __aarch64__
#define __builtin_ia32_crc32di __builtin_aarch64_crc32cx
#define __builtin_ia32_crc32si __builtin_aarch64_crc32cw
#define __builtin_ia32_crc32hi __builtin_aarch64_crc32ch
#define __builtin_ia32_crc32qi __builtin_aarch64_crc32cb
#endif // GCC_AARCH64_H
```

3.3 典型应用评测

目前，基于ARM的高能物理计算环境中已经装配了100台华为泰山200K服务器，每台服务器配置两个48核鲲鹏920 5251K处理器或者64核鲲鹏920-6426处理器，安装CentOS 7.6操作系统以及HTCondor、Slurm作业调度软件和CVMFS、EOS等数据存储软件。ROOT、GEANT4等基础软件库以及Chroma、LHAASO等应用软件全部移植成功，并部署在CVMFS上，所有计算节点均可共享。基于ARM的计算环境与x86、GPU等其他硬件统一管理，形成异构的计算资源池，共同支持各类高能物理实验和应用。在该计算环境中，开展了相关的应用测试，主要包括以下几方面。

(1) HS06 (HEP-SPEC06) 基准测试

这是高能物理领域用来评测CPU性能的标准工具。测试时，ARM CPU采用华为鲲鹏920-6426@2.6 GHz处理器，x86 CPU采用6核Intel E5-2620@2.0 GHz (命名为x86-1) 和20核Intel Gold 6230@2.1 GHz (命名为x86-2)。测试结果如**图4**所示，可以看出，ARM单核CPU性能比x86略低，但是整机性能是Intel E5-2620@2.0 GHz的7.27倍以及Intel Gold 6230@2.1 GHz的2.02倍。

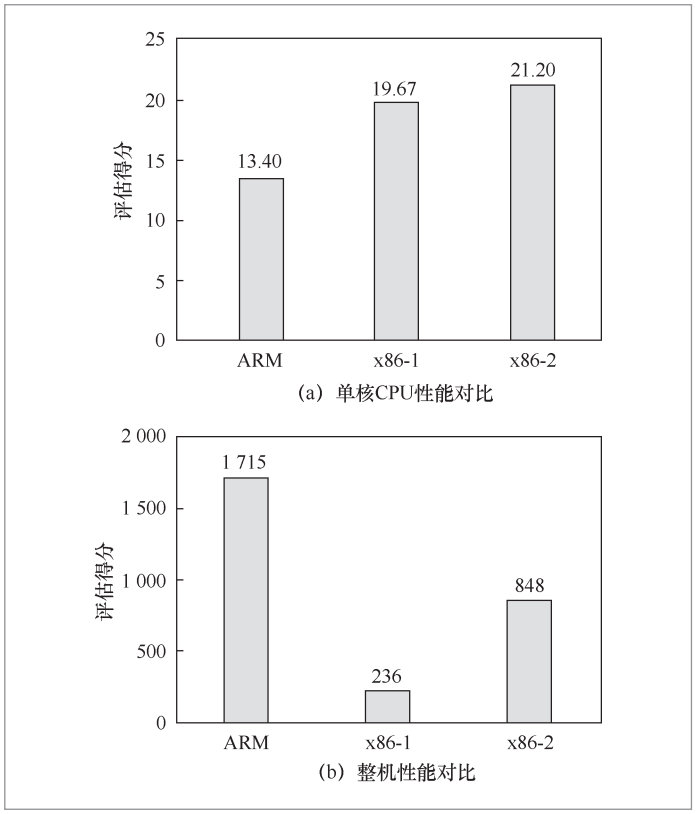

图 4 HS06 CPU 基准测试结果

(2) LHAASO事例重建测试

采用实际的LHAASO WCDA探测器事例数据进行重建。高海拔宇宙线观测站安装了3种类型的探测器，包括水切伦科夫探测器阵列 (WCDA)、地面簇射粒子阵列 (KM2A) 以及广角切伦科夫望远镜阵列 (WFCTA)。WCDA事例重建程序基于ROOT数据分析框架编写，从探测器获取的原始数据经过大量计算后构建出具有物理意义的事例。重建过程是一个典型的数据密集型计算，需要输入和输出大量数据。本测试对一个WCDA原始数据进行重建，包含418 816个事例，文件大小为1.1 GB。测试结果如**图5**所示，在minnhit=800和minnhit=1 500两种重建条件下，ARM单

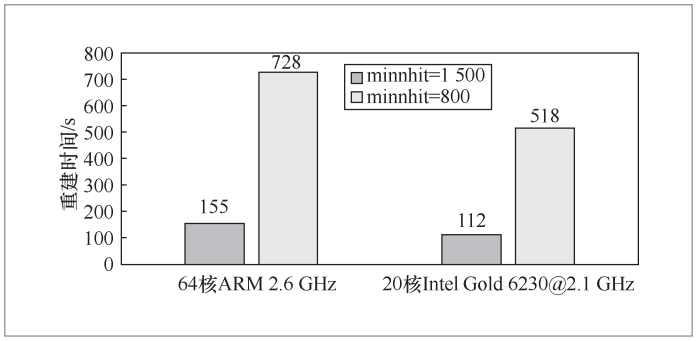

图 5 LHAASO WCDA 事例重建时间评测

核(64核ARM 2.6 GHz)运行时间比20核Intel Gold 6230@2.1 GHz的长40%左右。考虑到LHAASO重建过程具有很好的数据并行性,每个文件对应一个重建作业,如果根据核数来计算整机性能,那么ARM整机(128核)是x86机器(40核)的2.72倍,与HS06评测结果吻合。

(3)LQCD应用测试

LQCD是从第一原理出发求解QCD的非微扰方法,除标准模型基本参数之外,没有任何额外模型参数,因而其计算结果被认为是对强相互作用现象的可靠描述,格点计算对QCD理论研究意义重大。LQCD计算是典型的计算密集型应用,毕玉江等人[12]在“天河三号”原型上进行了ARM架构的移植与详细测试分析,在开启向量化后,Grid呈现了非常好的可扩展性。此外,本文还对比了x86与ARM的性能,使用两台服务器,一台采用2个Intel(R) Xeon(R) Gold 6248R CPU@3.00 GHz,另外一台采用2个华为鲲鹏920 7260@2.60 GHz CPU,两台机器都配置了256 GB DDR4 2933 MT/s内存。测试结果如**图6**所示。在整机满核情况下(Intel x86 48核,华为ARM 128核),x86实测性能为39 GFlops,ARM测试性能为43 GFlops,ARM约高出10%。但是,考虑到x86单核CPU理论性能更强,则ARM的CPU利用率更高,为x86 CPU的2倍。

从内存带宽基准测试工具STREAM的测试结果来看,华为鲲鹏ARM CPU的内存带宽比Intel(R) Xeon(R) Gold 6248R CPU高出20%以上(见**表1**),内存带宽的提升更有利于LQCD各个计算任务之间的通信,从而提升CPU利用率。

从以上基准测试、数据密集型应用的测试以及计算密集型应用的测试可以看出,ARM CPU能够满足高能物理数据处理的需求。虽然在单核性能上ARM CPU弱于x86 CPU,但是ARM CPU拥有更多的核心数与较高的内存访问带宽,从而提高了整机的计算性能。

4 可计算存储技术

如第2节所述,高能物理数据处理系统采用计算与存储分离的模式以及经典的冯·诺伊曼计算机体系架构,在进行数据处理时需要通过网络从存储节点读取计算节点,运算后再把结果写回存储节点,难以适应大数据驱动的科学计算需求。频繁的数据搬运导致的算力瓶颈以及功耗瓶颈已经成为高能物理数据处理系统对更先进算法进行探索的限制因素。基于这个问题,本文提出了面向高能物理数据处理的可计算存储技术解决方案。可计算存储也被称为存算一体化,它将数据存储单元和计算单元融为一体,能显著减少数据搬运,极大地提高计算并行度和能效。可计算存储的潜力引起了众多公司和标准机构的关注。全球网络存储工业协会(SNIA)成立了一个工作组[13],目标是建立可计算存储设备之间的互操作性标准。可计算存储可在软硬件协同的基础上解决大规模数据处理问题,如数据库查询优化[14]、key-value数据压缩[15]等。本文的可计算存储服务主要基于国产ARM

CPU片上系统(system on chip, SOC)的硬件加速引擎能力实现。

4.1 技术架构

可计算存储的本质还是存储，因此通过传统的存储系统接口仍然可以访问底层的存储设备。其计算能力通过扩展存储系统接口或者增加参数实现。比如，通过Open系统调用可以打开一个数据文件，然后进行正常的读写操作。同时，可以扩展特殊的参数，如在文件名后加上"&cssapp=decode"，即可完成相应的应用调用。该方式只需要通过存储系统调用即可实现完全本地化的数据处理，实现调度与存储的完全统一，从而解决高能物理数据分析处理的I/O瓶颈问题，其架构如**图7**所示。其中，最底层是可计算存储设备，由内存、固态硬盘、机械硬盘等存储介质和CPU SOC等计算资源构成。然后，在可计算存储设备上启动可计算存储服务，由存储服务管理存储介质，同时根据存储客户端的请求调用安装在可计算存储设备上的应用库和算法库，最终执行任务。计算节点通过高速网络连接到可计算存储系统，基于存储客户端实现应用调用、算法调用及简单的任务调度功能。

由于ARM架构的开放性，ARM CPU厂商在设计CPU时，一般还集成一些SOC，以提供相关的功能。以华为鲲鹏920为例，除了CPU外，它还集成了RoCE网卡、SAS控制器和南桥，以及加速引擎，包含了KAE加解密、KAEzip等，分别用于加速SSL/TLS应用和数据压缩，可以显著降低处理器消耗，提高处理器效率。

4.2 设计与实现

可计算存储设备的实现有多种，包括在磁盘控制器、磁盘阵列控制器、存储服务器、存储服务等进程中叠加计算功能，分别称为可计算存储驱动器(computational storage drive, CSD)、可计算存储阵列(computational storage array, CSA)、

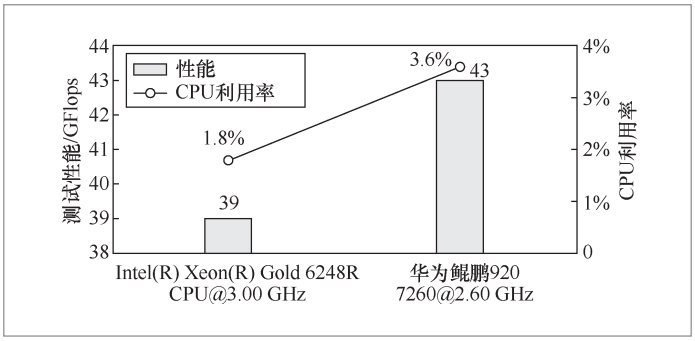

图 6 LQCD 应用测试对比

表 1 STREAM 内存带宽测试对比

进程数	内存带宽/(MB/s)			
	Copy	Scale	Add	Triad
48 (Intel)	201 232.0	201 855.5	221 104.4	220 547.3
128 (华为)	264 284.0	264 468.4	273 077.8	274 061.3

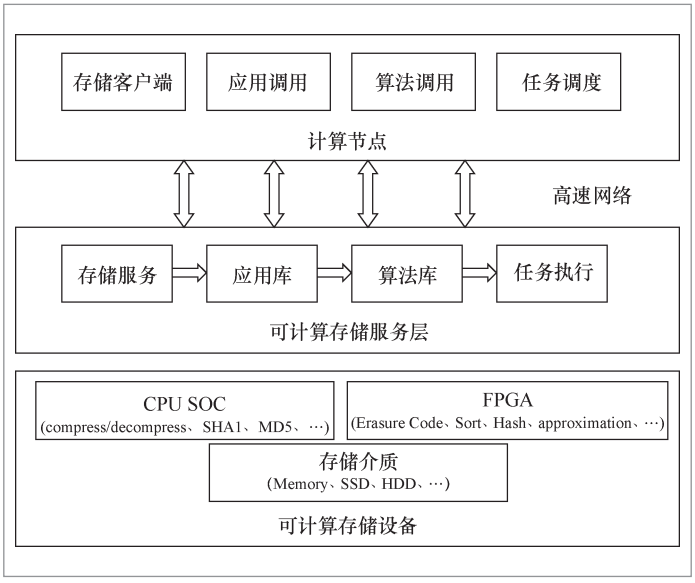

图 7 面向高能物理数据处理的可计算存储技术架构

可计算存储处理器（computational storage processor，CSP）、可计算存储服务（computational storage service，CSS）等。根据高能物理计算的特点，本文基于EOS分布式存储系统设计和研发了可计算存储系统。EOS分布式存储系统由CERN开发，在高能物理领域应用广泛，主要由元数据服务器（MGM）和I/O存储服务器（FST）构成，可以支持EB级的数据存储，客户端通过标准的XRootD协议[16]访问数据。对于用户或者应用来说，只要支持XRootD协议，就可以在本地甚至跨地域访问EOS存储。因此，该系统不需要修改客户端访问模式，原有的应用程序也不需要任何改变，如果用户希望利用存储设备上的计算能力，只需要在文件名后加上特定的参数就可以，其基本架构如**图8**所示。

在**图8**中，EOS的FST组件与其上安装的硬盘，再加上ARM CPU，共同构成可计算存储设备。具体的实现方式是编写两个动态加载的插件（eoscssmgm.so和eoscssfst.so）加在EOS配置文件中，不需要修改EOS的任何代码，从而保证系统的可扩展性和可移植性。eoscssmgm.so用来解析元数据的参数，它根据需要访问的文件路径，将文件访问请求分配到文件实体所在的FST上。eoscssfst.so用来解析数据I/O参数，从中分离出需要调用的应用或者算法，进而调用FST上CPU或者FPGA的计算能力，任务执行完后将计算结果返回给客户端。这样，所有数据I/O只在本地硬盘进行，完全避免了从存储节点（FST）到计算节点之间的网络流量。

高能物理数据处理程序通常基于ROOT软件框架实现，因此通过修改ROOT软件，可将其中某些计算能力卸载到存储节点，使可计算存储技术更加通用，如压缩/解压缩、排序、数据查找、数据索引、数据拟合等。目前，笔者团队已经把ROOT移植到ARM CPU上运行，并通过修改ROOT代码，实现了调用华为鲲鹏硬件压缩的能力，从而大幅提升ROOT数据压缩写入和读取性能。其他的数据处理功能（如数据索引、数据查找）也可以通过修改ROOT，将

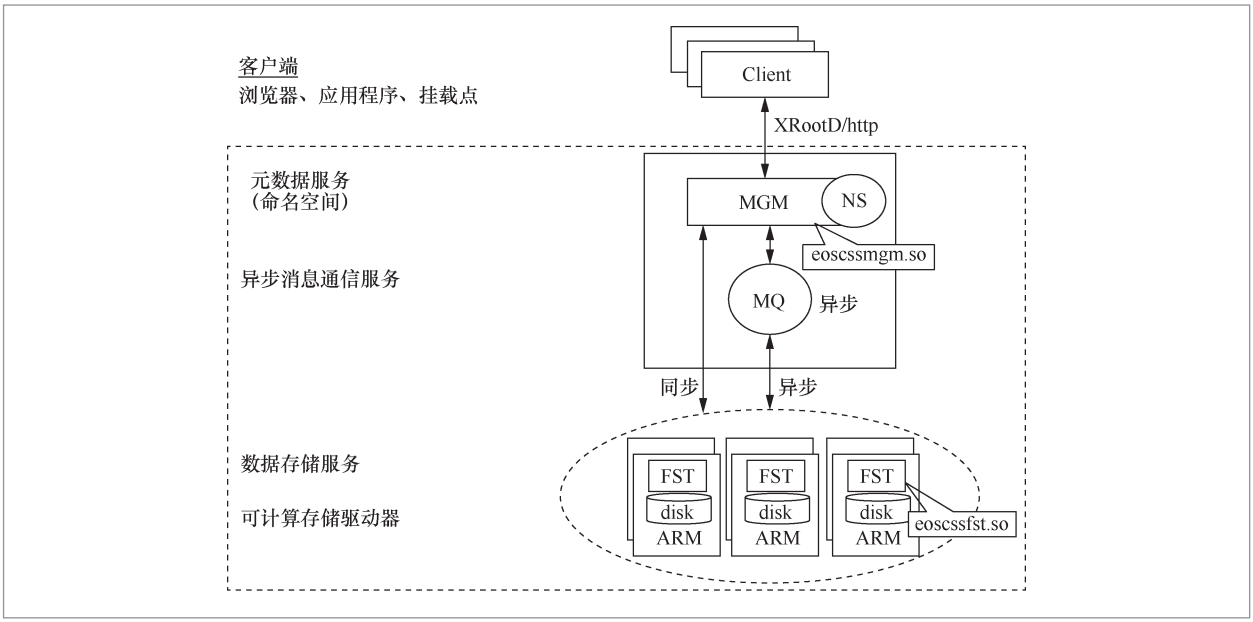

图 8 高能物理可计算存储实现框架

其从计算节点卸载到存储节点(FST)上实现，从而大大节省数据传输时间，提高数据分析效率。

4.3 典型应用评测

本文主要介绍两类可计算存储的应用评测，一类是LHAASO KM2A探测器事例解码，另一类是基于ARM SOC的压缩。LHAASO KM2A事例解码是将探测器的原始数据转换成具有物理意义的格式的过程。在传统的分析模式中，解码程序在计算节点上运行，通过网络读入二进制格式的探测器数据，任务结束后输出ROOT格式的原始数据。一个解码过程所需读带宽约为10 MB/s，写带宽约为4 MB/s。本文实现了可计算存储的模式，在计算节点打开文件时加上“&css=decode”选项，存储节点就会在FST存储服务器上调用解码程序，直接从本地硬盘读取文件，然后将输出的ROOT写到本地硬盘，并注册到分布式存储文件系统元数据中，从而实现“零传输”的数据分析模式。测试结果如**图9**所示，原计算模式下的执行速度随着并发进程数的增加逐步变慢，特别是达到10个进程后，数据访问速度达到1 Gbit/s网络带宽的限制，执行过程从10个进程165.177 s上升到40个进程489.321 s。可计算存储模式下的执行速度随并发进程数的增加变化不大，当并发进程数达到40个左右时，可计算存储模式的执行速度开始变慢，从30个进程的79.623 s增加40个进程的99.632 s，这是因为数据访问速度达到了本地硬盘性能的限制。测试结果显示，可计算存储模式可以支持更多的并发解码进程。

在ARM CPU中，一般会集成SOC系统级芯片，比如华为鲲鹏920 CPU内置了压缩引擎，本文通过修改ROOT软件实现对华为KAE zlib压缩加速库的调用。华为

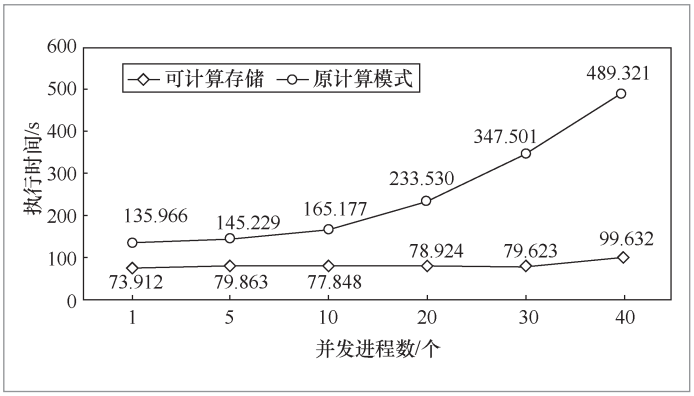

图 9 LHAASO KM2A Decode 应用评测

鲲鹏920 CPU KAE zlib压缩引擎单处理器理论上最大压缩带宽为7 GB/s，最大解压带宽为8 GB/s。在该应用评测中选用LHAASO KM2A事例解码程序。解码前，原始文件尺寸为1 GB。解码在输出时，可以设置数据压缩算法，由ROOT自动调用压缩算法生成ROOT格式的文件。因为华为鲲鹏ARM CPU仅支持zlib压缩算法，所以这里仅列出ROOT调用zlib压缩算法的运行时间对比，如**图10**所示。图例中压缩算法后的3位数字，第一位代表算法序号，第二位无意义，第三位代表压缩等级，同一种压缩算法中，压缩等级越高，速度越慢，压缩率越高。从**图10**可以看出，当进程数在20个以内时，采用华为KAE zlib加速库，可以将解码程序的执行时间减少30%以上。当进程数增加到40个以后，性能与CPU执行zlib相当。

5 结束语

本文基于ARM国产处理器及服务器等硬件，构建了完整的高能物理数据处理系统，包括EOS等数据存储软件、HTCondor等作业调度软件、ROOT/GEANT等基础软件库、LHAASO/LQCD

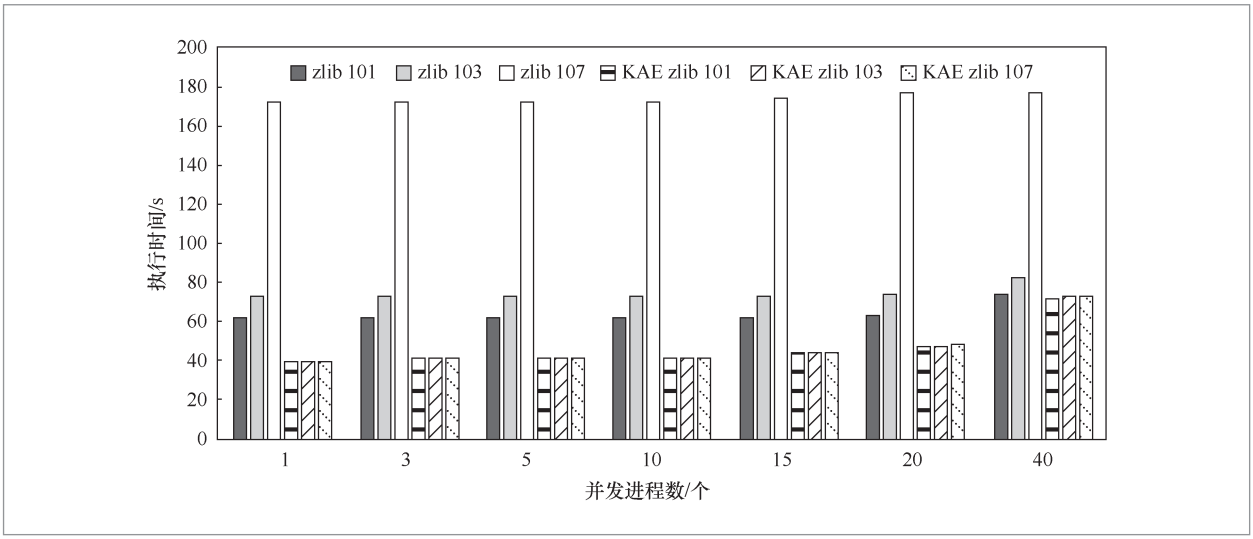

图 10 华为泰山服务器上各种压缩算法评测

等应用软件，并实现了大规模的运行。同时，本文还从超大数据处理挑战出发，提出了可计算存储技术方案，有效地解决了计算过程中因数据搬运带来的I/O瓶颈问题。LHAASO事例重建与解码等典型应用评测结果说明，基于国产处理架构的高能物理数据处理系统运行正确，多核架构的整机性能突出。可计算存储方式能够有效地利用存储节点的硬件能力，实现典型计算任务的卸载，避免数据多次搬运，提高计算效率。

目前，本系统已经支持典型的高能物理应用，证明了在国产处理器架构上开展高能物理数据处理的可行性和可推广性。下一步，将移植和优化更多的应用，基于可计算存储技术架构实现更多的计算任务卸载，并进一步将经验推广到其他相关领域。

致谢

本论文的工作得到国家高能物理科学数据中心在数据处理环境及科学数据等方面的支持，在此表示感谢！

参考文献:

[1] BOCCALI T. Computing models in high energy physics[J]. Reviews in Physics, 2019, 4: 100034.

[2] ANTCHEVA I, BALLINTIJN M, BELLENOT B, et al. ROOT: a C++ framework for petabyte data storage, statistical analysis and visualization[J]. Computer Physics Communications, 2011, 182(6): 1384–1385.

[3] PETERS A J, SINDRILARU E A, ADDE G. EOS as the present and future solution for data storage at CERN[J]. Journal of Physics: Conference Series, 2015, 664(4): 042042.

[4] KLIMENTOV A, BENJAMIN D, GIROLAMO A D, et al. Enabling data intensive science on supercomputers for high energy physics R&D projects in HL-LHC era[J]. EPJ Web of Conferences, 2020, 226(1): 01007.

[5] 程耀东, 石京燕, 陈刚. 高能物理计算环境概述[J]. 科研信息化技术与应用, 2014, 5(3): 3–10.
CHENG Y D, SHI J Y, CHEN G. A survey of high energy physics computing system[J]. e-Science Technology &

Application, 2014, 5(3): 3–10
[6] 何晓斌, 蒋金虎. 面向大数据异构系统的神威并行存储系统[J]. 大数据, 2020(4): 30–39.
HE X B, JIANG J H. Sunway parallel storage system for big data heterogeneous system[J]. Big Data Research, 2020(4): 30–39.
[7] 胡正丁, 薛巍. 面向异构众核超级计算机的大规模稀疏计算性能优化研究[J]. 大数据, 2020(4): 40–55.
HU Z D, XUE W. Research on performance optimization for large-scale sparse computation over many-core heterogenous supercomputer[J]. Big Data Research, 2020(4): 40–55.
[8] 张淼, 周宇, 陈建海, 等. LQCD Dslash在神威·太湖之光上的研究分析与MPI实现[J]. 计算机科学与探索, 2019, 13(10): 1664–1676.
ZHANG M, ZHOU Y, CHEN J H, et al. Analysis and MPI implementation of LQCD Dslash on Sunway TaihuLight[J]. Journal of Frontiers of Computer Science & Technology, 2019, 13(10): 1664–1676.
[9] PROMBERGER L, CLEMENCIC M, COUTURIER B, et al. Porting the LHCb stack from x86 (Intel) to AArch64 (ARM) and ppc64le (PowerPC)[C]//Proceedigs of the EPJ Web of Conferences. [S.l.]: EDP Sciences, 2019: 05016.
[10] MARIK M. Porting the LCG software stack to the ARM architecture[Z]. 2019.
[11] AGOSTINELLI S, ALLISON J, AMAKO K, et al. GEANT4: a simulation toolkit[J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2003, 506(3): 250–303.
[12] 毕玉江, 周超, 吴郁非, 等. 格点量子色动力学 Grid 数值模拟软件的并行计算特征分析[J]. 计算机系统应用, 2020, 29(7): 199–204.
BI Y J, ZHOU C, WU Y F, et al. Parallel computing feature analysis of grid numerical simulation software for lattice quantum chromodynamics[J]. Computer Systems & Applications, 2020, 29(7): 199–204.
[13] SNIA. Computational storage, computational storage architecture and programming model[R]. 2020.
[14] CAO W, LIU Y, CHENG Z S, et al. POLARDB meets computational storage: efficiently support analytical workloads in cloud-native relational database[C]// Proceedings of the 18th USENIX Conference on File and Storage Technologies. [S.l.:s.n.], 2020: 29–41.
[15] ZHANG T, WANG J Y, CHENG X T, et al. FPGA-accelerated compactions for lsm-based key-value store[C]//Proceedings of the 18th USENIX Conference on File and Storage Technologies. [S.l.:s.n.], 2020: 225–237.
[16] DORIGO A, ELMER P, FURANO F, et al. XRootD-a highly scalable architecture for data access[J]. WSEAS Transactions on Computers, 2005, 4(4): 348–353.

作者简介

程耀东(1977–), 男, 博士, 中国科学院高能物理研究所研究员、博士生导师, 主要研究方向为高性能计算、分布式存储、可计算存储等。

程垚松(1995–), 男, 中国科学院高能物理研究所助理工程师, 主要研究方向为高性能计算、分布式存储等。

毕玉江（1990– ），男，博士，中国科学院高能物理研究所助理研究员，主要研究方向为高性能计算、分布式存储、LQCD、量子计算等。

高宇（1994– ），男，中国科学院高能物理研究所硕士生，主要研究方向为分布式存储、可计算存储等。

李海波（1984– ），男，中国科学院高能物理研究所副研究员，主要研究方向为海量数据存储、大数据处理等。

汪璐（1983– ），女，博士，中国科学院高能物理研究所副研究员，主要研究方向为分布式文件系统、云存储和机器学习等技术在高能物理计算中的应用。

姚秋玲（1978– ），女，中国科学院高能物理研究所高级工程师，主要研究方向为海量数据存储、数据备份等。

收稿日期： 2021-05-30

通信作者： 程耀东，chyd@ihep.ac.cn

基金项目： 国家重点研发计划资助项目（No.2017YFB0203200）；国家自然科学基金资助项目（No.12075268，No.11575223）；中国科学院高能物理研究所科技创新项目（No.E15451U2）

Foundation Items: The National Key Research and Development Program of China(No.2017YFB0203200), The National Natural Science Foundation of China(No.12075268, No.11575223), The Science and Technology Innovation Project of IHEP(No.E15451U2)