arXiv:2303.05296v1 [cs.DC] 9 Mar 2023

高能物理实验专用分析设施

G´aborB´ır´o[1,2], Gergely G´abor Barnaf¨oldi[1], P´eter L´evai[1][1]Wigner物理研究中心，匈牙利布达佩斯Konkoly–Thege Mikl´os Str. 29–33号，邮编H-1121。[2]E¨otv¨os Lor´and大学物理研究所，P´azm´any P´eter S´et´any 1/A号，匈牙利布达佩斯，邮编H-1117。

电子邮箱：biro.gabor@wigner.hu；barnafoldi.gergely@wigner.hu；levai.peter@wigner.hu 2023年3月10日

摘要。高能物理（HEP）产生的数据量不断增长。为了分析这些数据，在持续提升计算能力的同时，还需要扩大存储容量。这些发展对该领域的未来至关重要；然而，也可以通过优化现有IT资源来实现。

高能物理软件工作流程中，探测器模拟和数据分析是计算能力的主要消耗项之一。为优化实现这些目标所需的资源，欧洲核子研究中心的 ALICE实验提出了在第三次运行期间设立专用分析设施（分析设施）的构想。这些分析设施是专门的计算中心，配备中央处理器和高速互联的磁盘存储模块，可针对一部分专用数据快速完成分析任务。这进而有助于优化分析流程和代码，以便在全球大型强子对撞机计算网格的大规模数据样本上开展分析。

本文介绍了 Wigner 分析设施（Wigner 分析设施）的结构，并总结了其在2020—2022年期间的进展。

1. 引言

大型强子对撞机（LHC）上的大型探测器在2019-2022 [1]期间的第二次长期停机（LS2）中进行了重大升级。探测器灵敏度、读出硬件，以及相关的在线和离线软件均得到更换和现代化改造。研发活动的目标是使各实验能够在第三次运行数据采集期（2022—2025年）及以后开展新的物理研究。

为实现这些目标，我们针对大型强子对撞机第一次运行和第二次运行的大规模数据样本，研究了高效的数据处理。我们还测试并优化了蒙特卡罗模拟的性能，使其能够在全球大型强子对撞机计算网格（WLCG）上进行大规模并行事件生成。最终目标是在维持合理维护和运行成本的同时，尽可能提升计算性能，尽管现有硬件组件已较为老旧。

Wigner数据中心（WDC）的分析设施与原有设施一同建立，继承自前欧洲核子研究中心第0级站点（布达佩斯，

匈牙利） [2]。基于现有硬件，根据实验需求对拓扑结构和模块进行了进一步优化，以满足快速数据处理活动的需要。离线分析设施（AF） [3]的最初构想首先在 Wigner 分析设施中大规模应用，最新的多核分析软件框架 Hyperloop [4] 也在此进行了测试。

2. 分析设施的结构

Wigner 分析设施是 Wigner 科学计算实验室（WSCLAB）的一部分，实际位于 WDC。Wigner 分析设施的大部分硬件由布达佩斯 Tier 0 计算中心的旧有硬件构成，主要是 AMD Opteron 6276 处理器 [5]。分析设施的主要用途是每天高效处理大量数据，数据量达 $\mathcal{O}$(PB)，同时每年将资源扩充 $\mathcal{O}$(15%)。因此，存储和计算部分都采用模块化设计，并确保两者之间具有高带宽通信，这一点至关重要。

经过多轮硬件测试和带宽优化，最终选定双机架式“单元”作为 Wigner 分析设施的可扩展单元。这种独立运行单元由计算、前端、存储和服务组件构成，如图 1 所示。每个 8 个计算机箱 均包含 4 台双处理器机器，每个单元共计 1024 个线程。这些单元处理通过专用接口 VO Box [6], 提交的分析作业；VO Box 是从全球 WLCG 系统进入分析设施的入口。随后，作业会被传递至 HTCondor [7] 和 HTCondor-CE [8] 服务器，由它们分发给相连的工作节点。

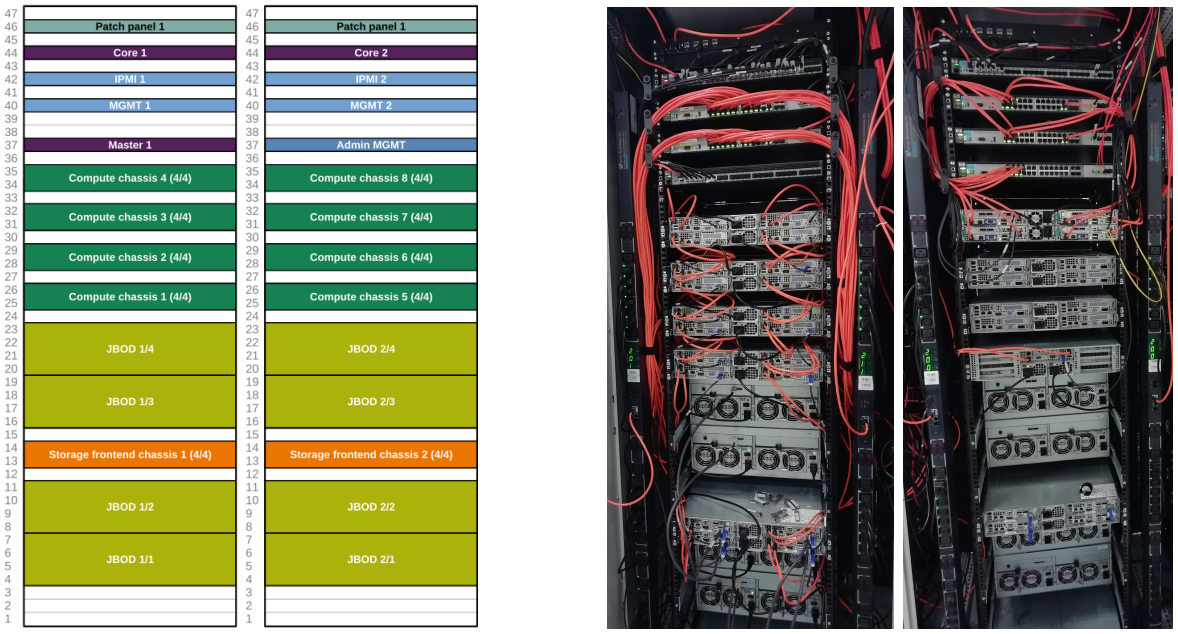

图 1. Wigner 分析设施中单个单元的结构。

一个单元的存储单元由一个配备 24 块磁盘的 JBOD机箱构成，并通过 XRootD [9] 和 EOS [10, 11] 服务及守护进程，由前端机箱中的机器进行控制。此类文件存储服务器（FST）节点由管理服务器负责管理。上述每台提供管理服务的服务器机器都配有可信网格服务器证书，以确保与 WLCG基础设施无缝连接。

机器的操作系统级编排通过 Canonical 开发的数据中心自动化系统 Metal-As-A-Service (MAAS) 实现 [12]。将计算、存储和网络节点部署在同一单元中，可确保分析工作流所需的快速数据传输。节点之间的高速内部通信由 HP ProCurve 6600-24XG (J9265A) 交换机保障 [13]。利用 SFP+ 10GbE 端口，单元内部以及不同单元之间均可实现 10 Gbps 的高带宽。

第一个工作单元中有两个机箱专门用于特殊用途。为满足未来发展需求，一个计算机箱专用于配备图形加速卡（图形处理器）的机器，另一个计算机箱中的机器则承担管理职能。

3. 计算能力和网络连接

分析设施的概念在欧洲核子研究中心 ALICE 实验框架内进行了测试，测试地点为 WSCLAB，共设有 4 个单元（总计 8 个机架），规模可与中型 Tier 2站点相当 [2]。该站点位于匈牙利布达佩斯的 KFKI园区，隶属于 Wigner数据中心；该数据中心通过支持 100 Gbps 链路的新设备连接到 GEANT网络。该站点的总存储容量和计算能力汇总于表 1。

表 1. Wigner 分析设施总资源概览

总存储容量	总计算资源
32 个 FST 节点	单核和多核作业队列
每个 FST 节点的原始容量为 24×3 TB	128 个工作节点
总原始容量： $\sim$2.2 PB	每个节点配备 32 个虚拟 CPU 和 64 GB 内存
使用 RAID-1 后的可用容量： $\sim$1.1 PB	总计 4096 个逻辑核心

3.1. 计算资源基准测试

在进行硬件安装的同时，也开展了一系列性能和优化测试。首批试运行作业于2020年末执行，当时分析设施仍处于软硬件配置阶段；而作业成功率较高的生产期则于2021年2月开始。针对采用真实分析工作负载的8机架配置进行的性能测试于2021年9月开展，并于2022年2月再次进行（见图2）。因此，输入/输出速率的提升，也就是性能的提升，

分析设施的结果表明，优化分析框架软件对于充分利用底层硬件的能力至关重要。可以清楚地看到，优化后的架构使单核作业的输入/输出和 +20%分析吞吐量保持不变；而对于八核作业，输入/输出为 +10%，分析吞吐量几乎相同。

我们的测试表明，当前配置的理论吞吐量（包括数据输入/输出开销）峰值为 1.1 拍字节/天。平均而言，Wigner 分析设施的分析性能为 ~ 4 兆字节/秒；单核和八核作业的输入/输出速率分别为 18 MB/s 和 43 Mb/s。



图2。 单核和八核作业的估算输入/输出速率与分析吞吐量。

4. 结论

本文介绍的 Wigner 分析设施是专用吞吐量型计算设施的早期实例之一。未来，这类设施将在高能物理实验中日益普及。该设施针对欧洲核子研究中心 ALICE 实验提供的分析框架和数据，对所测试的特定任务进行了优化。

考虑到基础设施成本（包括电力和高吞吐量计算硬件），Wigner 分析设施可为未来的计算挑战提供可维护且可扩展的解决方案，例如 LIGO/VIRGO 的引力波分析，以及 EUPRAXIA 项目中的核数据库处理。这些经验和该设施也向其他大型合作项目开放。

5. 致谢

本研究得到了匈牙利国家研究、发展与创新办公室（NKFIH）的资助，合同编号为 OTKA K135515、2019-2.1.6-NEMZ KI-2019-00011、2022-4.1.2-NEMZ KI-2022-00009 和 2022-4.1.2-

´NEMZ KI-2022-00008。作者谨向 Ad´am Pint´er、J´ozsef Kadlecsik 以及 Wigner数据中心的技术人员表达谢意，感谢他们搭建分析设施的硬件。我们感谢 WLCG 管理层的支持。

参考资料

[1] “大型强子对撞机”。载于：JINST 3（2008）。Lyndon Evans 和 Philip Bryant 编，S08001。 DOI： 10.1088/1748-0221/3/08/S08001。[2] “Wigner ALICE 分析设施”。载于：（2021）。 URL： https://cds.cern.ch/record/2791181。[3] Schwarz、Kilian 等。“GSI 的 ALICE 分析设施原型”。载于：EPJ Web Conf. 214（2019），第 08027 页。 DOI： 10.1051/epjconf/201921408027。 URL： https://doi.org/10.1051/epjconf/201921408027。[4] Raquel Quishpe 等。“Hyperloop——第三次运行的 ALICE 分析列车系统”。载于：第九届大型强子对撞机物理学会议。2021 年 9 月。arXiv：2109.09594[physics.ins-det]。[5] AMD Opteron 6276。访问日期：2023 年 3 月 2 日。 URL： https://www.amd.com/en/products/cpu/6276。[6] WLCG VOBOX。访问日期：2023 年 3 月 2 日。 URL： https://twiki.cern.ch/twiki/bin/view/LCG/WLCGvoboxDeployment。[7] HTCondor。访问日期：2023 年 3 月 2 日。 URL： https://research.cs.wisc.edu/htcondor/。[8] HTCondor-CE。访问日期：2023 年 3 月 2 日。 URL： https://htcondor.github.io/htcondor-ce/。[9] Fabrizio Furano 和 Andrew Hanushevsky。“Scalla/xrootd 广域网全球化工具：现状如何”。载于：《物理学杂志：会议系列》219.7（2010），第 072005 页。 DOI： 10.1088/1742-6596/219/7/072005。 URL： https://doi.org/10.1088/1742-6596/219/7/072005。[10] Andreas J. Peters 和 Lukasz Janyst。“欧洲核子研究中心的艾字节级存储”。载于：J. Phys. Conf. Ser. 331（2011）。Simon C. Lin 编，第 052015 页。 DOI： 10.1088/1742-6596/331/5/052015。[11] Geoffray Adde 等。“EOS 文件系统的最新演进”。载于：《物理学杂志：会议系列》608（2015），第 012009 页。 DOI： 10.1088/1742-6596/608/1/012009。 URL： https://doi.org/10.1088/1742-6596/608/1/012009。[12] 裸金属即服务。访问日期：2023 年 3 月 2 日。 URL： https://maas.io/。[13] HP ProCurve 6600-24XG (J9265A)。访问日期：2023 年 3 月 2 日。 URL： https://support.hpe.com/hpesc/public/docDisplay?docId=emr_na-c01832969。