医疗机构自主可控大数据科研平台建设方案

匡亚岚 李春漾 应志野

（四川大学华西医院生物医学大数据中心 成都 610000）

〔摘要〕 以多组学数据分析场景为例，从基础硬件设施、基础软件、应用系统 3 方面阐述医疗机构自主可控大数据科研平台国产化建设方案，为生物医学数据科研平台建设积累经验，为国内医疗科研机构提供参考。

〔关键词〕 自主可控；医疗大数据；医疗科研平台

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.03.013

Construction Scheme of Big Data Research Platform for Medical Institutions Based on Autonomous and Controllable Technology

KUANG Yalan, LI Chunyang, YING Zhiye, West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu 610000, China

〔**Abstract**〕 Taking the multi omics data analysis scenario as an example, the localization construction scheme of big data research platform for medical institutions based on autonomous and controllable technology is elaborated from three aspects of basic hardware facilities, basic software and application system, so as to accumulate experience for the construction of biomedical data research platform and provide references for domestic medical research institutions.

〔**Keywords**〕 autonomous and controllable; medical big data; medical research platform

1 引言

人工智能、云计算、移动互联等技术迅猛发展，为医疗信息化建设提供新思路的同时也带来压力。一方面，医疗数据呈爆炸式增长，数据管理、数据安全等问题逐渐凸显，需建立统一的大数据平台优化管理，加速实现数据资产价值转化，从而推动诊疗服务模式革新以及成果孵化创新[1]。另一方面，面对一些核心设备和元器件“卡脖子”现状，我国发展自主核心技术的需求越发迫切，复杂的国际关系也推动我国将信息安全和关键技术自主创新提升到国家战略高度[2－3]。建设基于自主可控技术的医疗卫生大数据科研平台符合产业、民生以及国家安全战略需要[4]。医疗大数据平台升级国产化应提前布局，确保关键设备自主可控，充分保障数据安全性。

2 自主可控技术发展现状

2.1 总体情况

在国家高度重视以及各类重大专项支持下，我国国产基础软硬件发展取得较大进展[5]。自主可控产业实现关键技术突破，覆盖芯片设计、整机生

〔修回日期〕 2021－07－19

〔作者简介〕 匡亚岚，硕士，初级工程师；通讯作者：李春漾，博士，助理研究员。

产、软件研发、系统集成、测试验证、运维服务等产业链环节，基本形成从底层芯片到基础应用软件的全栈生态，见图 1。

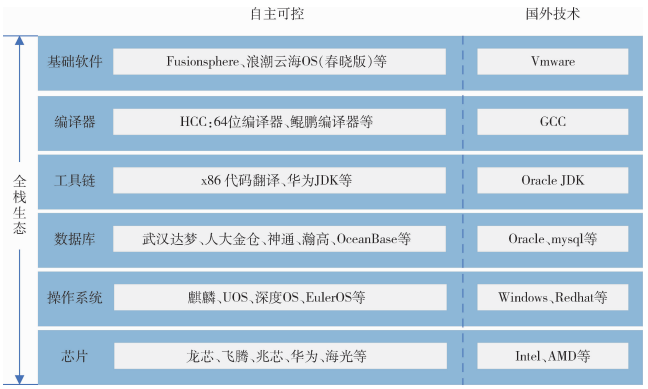

图 1 全栈生态自主可控替代

2.2 硬件方面

2.2.1 处理器 “十三五”期间，在我国政策引导和国际形势压力下芯片技术高速发展，形成以 MIPS、Alpha、ARM 和 X86 等为基础架构的处理器产业链[6-7]，性能已接近国际先进水平。图形处理器（Graphic Process Unit，GPU）方面，国内 GPU 芯片研制虽然可满足目前大多数图形应用需求，但在科学计算、人工智能及新型图形渲染技术方面仍然和国外领先水平存在较大差距[8]。

2.2.2 存储设备 为保障医疗数据安全，自主可控的存储系统需达到高可靠、高安全、高性能[9]。目前国内存储产业生态基本形成，已实现从存储芯片到存储介质、固态硬盘控制器、存储控制器，再到存储引擎全部自研的高端存储设备，达到国际先进水平[10-12]。

2.2.3 网络 医学科研平台对网络传输效率和质量提出更高要求。在高性能计算和人工智能算法训练等场景中，多采用存储和计算分离策略，网络传输主要使用光纤（Fibre Channel，FC）交换机和 IB（Infiniband）交换机。由于专利壁垒，这两项技术被国外厂商垄断，基于传统以太网协议的网络设备还无法实现国产化替代[13]。RoCE 协议的发展为自主创新以太交换机替代 FC 和 IB 提供了机会[14]。

2.3 软件方面

2.3.1 操作系统 是软件体系的基础。微软相继对 Windows XP 和 Windows 7 操作系统停止服务支持引发公众对信息安全的担忧，也给国产操作系统发展带来契机。国内厂商持续推进国产操作系统生态构建，已实现多款具有内生安全体系的操作系统并与国产芯片完成适配，全面兼容主流软硬件产品[15-16]。

2.3.2 数据库 我国数据库产品相对成熟，在性能、稳定性、安全性等方面有突出表现，接近国际先进水平。国产数据库在国内市场份额已提升至 8% ~10%[17]，在政府、金融、电信等多个领域都有广泛应用。目前我国已成功研发出具有自主知识产权的关系型数据库以及分布式数据库软件[18-19]。

3 医疗大数据科研平台自主可控替代方案

3.1 平台总体架构

2018 年国家卫健委对全国公立医院开展大数据相关工作提出要求，三级医院需在 5 ~10 年内建设医疗大数据科研平台[20]。目前各医院医疗大数据科研平台服务器、存储、数据库等产品大部分采用国外厂商核心技术，特别是用于前沿医学科研的高性能计算平台，其软硬件设备国产化程度不高，缺少核心软硬件平台国产化探索。现有针对医疗大数据科研平台的研究大多集中在数据收集、数据处理、模型搭建等软件框架，基础设施层研究较少。本文以生物信息科研应用场景为例，提出多组学大数据科研平台自主可控建设方案。平台总体架构，见图 2。底层为包含各类硬件与基础软件的基础设施层，为上层应用提供大算力、高性能存储以及数据安全方面保障。搭建“大而全”的底层数据中心，在此基础上通过多组学数据整合、治理与集成构建科研数据湖，形成多组学数据资产，进而结合机器学习、人工智能等方法从多组学数据中挖掘有价值信息，为疾病发生发展的分子机制发现、药物研发以及个性化诊疗等提供辅助支持。本文聚焦基础设施层、核心软硬件自主可控，从源头上保障医学科研

环境和医疗数据安全性，应对被“停服”或“禁售”等供应链风险，同时兼顾平台性能和兼容性，为科研平台多元化应用场景提供重要技术支持。

图 2 医疗大数据科研平台总体架构

3.2 基础硬件设施

3.2.1 *计算资源* 基础硬件设施自主可控主要包括计算资源、存储资源、网络资源构建，为上层应用提供算力支持。由于多组学科研分析中各阶段对计算性能的需求大多以计算密集型为主，为提升计算效率、提高资源利用率，方案采用 CPU + GPU 异构方式。目前国内已研发多款拥有自主知识产权的系列芯片，在基于 X86 或 ARM 两种指令集的国产芯片中已有能满足临床科研大数据需求的产品。GPU 国产芯片也有可商用的产品。在满足性能要求的前提下，各医院根据科研实际需求选择 ARM 或 X86 架构的国产化芯片，利用不同的异构平台来实现性能最大化，可充分保障科研分析对计算的需求。

3.2.2 *存储资源* 医疗数据涉及病例数据完整性和数据隐私保护，对存储安全要求较高。方案选择芯片、控制器、操作系统均为自主可控的存储设备，操作系统定期进行安全加固，且支持数据加密传输，有效保护医疗敏感数据安全性[21]。在部署时采用分布式系统结构，利用多台存储服务器分担存储负荷并配置数据备份节点，不但提高系统可靠性、存取效率和安全性，还易于扩展，将通用硬件引入的不稳定性降到最低，可高效应对本方案中异构复杂场景对存储性能和安全性的要求。

3.2.3 *网络资源* 多组学分析计算过程中计算节点间、存储节点与计算节点间需要交换大量数据，对网络带宽要求高，网络无阻塞性、低丢包率很重要。目前市面上常用的 IB 协议从硬件级别保证可靠传输、高吞吐量，技术先进但成本高昂。RoCE 协议在性能上与 IB 相当，稳定性较好且成本低[22]，国内市场已推出基于 RoCE 协议的网络和存储设备，具备国产化替代条件。在部署方式上可采用管理网和业务网分离模式，管理网主要负责传输管理节点与存储节点、计算节点之间的管理任务，业务网主要为节点间数据传输提供网络支撑，大吞吐量、稳定的数据传输是保障高算力的关键。

3.3 基础软件

3.3.1 *操作系统层面* 鉴于 Linux 系统在服务器领域的稳定性和安全性优势，选择基于开源 Linux 开发的国产操作系统，基于现有 Linux 生态支持部分生物信息分析软件安装和使用，减少适配工作量。

3.3.2 *数据库层面* 在医疗科研平台中，Oracle 等国外产品长期占有大量市场份额。目前国内已拥有一系列完全自主可控的数据库产品，与大部分国产芯片、操作系统完成适配，稳定性和高可用性接近国外主流产品，支持在医疗领域的推广应用。数据迁移方面需要根据国产数据库特性，从数据结构、类型和使用场景 3 方面制定迁移实施方案。通过语法、语义对比找到静态数据差异，根据数据使用场景将业务流程梳理清晰，明确业务与数据库之间的调用方式，重点需要对一些复杂的长事务操作进行程序适配，例如特殊分析函数、存储过程等，避免出现数据丢失等问题。

3.3.3 *中间件层面* 主要是 JDK 适配，由于国产化平台不支持 SUN JDK，应用系统业务不能直接迁移到国产化平台，在原 X86 平台下预编译的 JSP 文件需重新修改编译。

3.4 应用系统

构建自主可控的应用系统首要任务是平滑迁移，即从国外分析工具生态迁移到国产化生态后力

求对用户行为习惯影响最小。在生物信息科研场景应用分析中，从数据采集到数据归档，需要通过大规模计算分析从海量数据信息中辨识有用基因及其序列，最终获取遗传信息。这一过程常用的分析软件繁多，在进行国产化平台迁移时，每个阶段所用到的软件都需要根据底层国产化软硬件做兼容性测试，必要时还需开发同类型可替代的分析工具，见表 1。目前大部分多组学分析软件都基于 X86 架构开发，在 ARM 等其他架构的自主可控平台上还没有完整的适用于生物信息科研分析的应用软件生态，尚待进一步完善。

表 1 生物信息科研典型应用软件

应用分类	描述	典型分析软件
基因拼接	一般运行于 32、64 位操作系统下，至少需要 5G 物理内存，人类基因组等较大基因组一般需要 150G 内存。其特点如下：内存需求超高；I/O 需求高；不同物种基因组对硬件平台需求有较大差异	GATK、CANU、Falcon、HGAP4、ALLPATHS－LG 等
基因对比	序列比对软件运行特征是：检索、查询、比较、输出。其特点如下：对系统的 I/O 的要求高；程序消耗内存大	BWA、bowtie、bowtie2、Blast、soap、blasr 等
全基因组关联分析	全基因组关联研究（Genome Wide Association Study，GWAS）是利用各类分离群体研究定位与目标性状相关联的基因区域的一种方法。其特点如下：内存需求较高；I/O 需求较高；计算量大	Plink、MACH、IMPUTE、BEAGLE、fastPHASE 等
RNA 等其他组学数据分析	基因测序根据样本类型及实验目的可进行基因组学、转录组学、表观组学、蛋白组学等组学研究，甚至通过多组学联合研究，挖掘一个完整的生物学功能过程。这些组学分析比较复杂，主流分析软件也较多，其特点如下：内存密集型；I/O 密集型；计算密集型	RSEM、rMATS、FusionMap、STAR－fusion、deFuse、Cufflinks、Cuffdiff、DESeq2、edgeR、topGO 等

4 问题与建议

4.1 性能

目前能支撑医疗大数据高性能计算的国产化芯片产品相对单一，性能与国外产品尚有差距。国内厂商已具备一定芯片设计能力，但芯片生产关键技术还有待突破。国产 CPU 架构大多采用国外技术，受专利壁垒限制，一旦架构更新将面临重新授权的问题，能否自主研发具有更高性能的 CPU 内核成为关键。在提升单颗 CPU 主频成为瓶颈时，国产化替代需要从架构创新寻找突破，针对特定领域的特定需求，设计不同的异构计算平台，实现专用性能扩展，或能成为性能提升的新方向。

4.2 兼容性

自主可控技术在电子政务、金融、交通等领域得到很好的应用，但在医疗大数据领域还欠缺行业解决方案。完整、商业化的生态体系构建是自主可控技术发展的关键，尤其是高成熟度的解决方案。自主可控技术在医疗大数据领域的推广应用阶段，芯片与下游产业融合发展，增强与医疗机构和高校的合作，鼓励创新，从用户实际需求出发，积极推进医疗分析应用软件的优化适配工作，在满足科研平台多样化应用需求的基础上形成软硬件结合的平台解决方案，健全医疗行业科研应用软件生态，填补医疗大数据领域国产化分析软件空白。在此过程中应用和研发同步走，做好国产化和非国产化两类环境的双向适配及融合解决方案，鼓励重点医疗机构首批应用试点及规模推广，在市场应用过程中不断发现问题、解决问题，做强生态。

5 结语

医疗行业安全关系民生，基于自主可控技术的医疗卫生信息化建设、大数据健康产业及智慧医疗

推进将惠及亿万公众。“十四五”规划已明确要求面向人民生命健康深入实施科技自立自强，未来几年将是自主可控技术大规模推广和应用的关键时期。相信在不久的将来，随着关键技术突破、产业生态完善，在政府的鼓励和应用牵引下，我国医疗健康领域信息化发展将进入以自主创新为主的新局面。

参考文献

1 郭强，王丛，衡反修. 医疗大数据平台建设机遇、挑战及其发展［J］. 医学信息学杂志，2021，42（1）：2－8.

2 倪光南. 核心科技乃国之重器网信产业发展离不开自主可控［J］. 信息安全与通信保密，2018（11）：16－23.

3 倪光南. 坚持信创科技自立自强建设网络强国和数字中国［J］. 信息安全研究，2021，7（1）：2－3.

4 吴绍波，卢思羽. 新形势下中国信息产业创新生态的国产化替代战略研究［J］. 决策咨询，2020（1）：1－6，12.

5 石菲. 2020 年信创产业生态稳步向前［J］. 中国信息化，2020（11）：33－36.

6 曾宪荣. 中国国产化处理器进展综述［J］. 集成电路应用，2018，35（1）：13－18.

7 胡伟武. 龙芯 CPU 15 年研发历程［J］. 中国经济周刊，2018（17）：18－21.

8 李科奕. 大时代背景下抓住国产处理器的发展机遇［N］. 经济参考报，2019－06－06（7）.

9 姜斌. 高性能计算系统在大数据分析中的应用探究［J］. 电子元器件与信息技术，2021，5（2）：201－202.

10 张晔嘉. 国内信息系统自主可控生态环境分析［J］. 电子质量，2019（7）：58－61.

11 华为技术有限公司. 鲲鹏计算产业发展白皮书［EB/OL］.［2021－02－20］. http：//www. huawei. com.

12 龙芯中科. 腾凌科技存储［EB/OL］.［2020－08－30］. http：//www. loongson. cn.

13 钱德沛，王锐. E 级计算的几个问题［J］. 中国科学：信息科学，2020，50（9）：1303－1326.

14 金浩，杨洪章. RDMA 网络传输技术研究综述［J］. 科技风，2020（18）：131.

15 李艳. 希望之光！国产操作系统银河麒麟 V10 发布［EB/OL］.［2020－08－20］. https：//baijiahao. baidu. com/s? id=1675499913821217529&wfr=spider&for=pc.

16 统信软件技术有限公司. 统一操作系统 UOS［EB/OL］.［2021－02－20］. https：//www. uniontech. com.

17 智慧 IT. 国产数据库发展研究报告［EB/OL］.［2020－07－08］. https：//max. book118. com.

18 郑善双. 人大金仓：从传统数据库到新兴大数据［J］. 软件和集成电路，2017（6）：82－85.

19 周亚洁. 数据库国产化替代面临的问题及对策研究［J］. 信息安全研究，2018，4（1）：24－30.

20 国家卫生健康委员会. 全国医院信息化建设标准与规范（试行）［EB/OL］.［2018－04－13］. http：//www. nhc. gov. cn/.

21 华为技术有限公司. 华为 OceanStor 智能混合闪存存储系统技术白皮书［EB/OL］.［2021－06－30］. http：//www. huawei. com.

22 华为技术有限公司. RoCE、IB 和 TCP 等网络的基本知识及差异对比［EB/OL］.［2021－06－30］. http：//www. huawei. com.