基于容器技术的高性能计算公共服务平台建设

摘要：本文介绍了山东大学在新的高性能计算系统建设中，引入容器技术，实现对传统高性能应用模式的改进，使高性能计算系统成为初步具备公共服务特性的云计算系统所作的实践探索。文中介绍了“山东大学高性能计算云平台”的主要服务模式、总体解决方案、关键技术，以及基础硬件体系，并在最后介绍了“山东大学高性能计算云平台”的现状及未来的发展趋势。

文 / 万林[1] 包婵婵[2] 平凡[3]

近些年的科研实践表明，高性能计算在提高科研水平，加速科研成果产出方面具有无可替代的重要意义[1]。高校作为重要的科研力量，其在高性能计算系统的应用和建设方面投入了巨大的资源。其中以校级计算系统最具有代表性，包括北京大学的“未名一号”，上海交通大学的“π”[2]，以及中国科学技术大学的“超级计算中心”[3]等。山东大学从 2016 年开始组织筹备新的高性能计算系统，实践了容器技术在高性能计算环境中完成资源的管控，改变了高性能计算环境的公共服务能力与模式，并优化用户应用体验，取得了较满意的应用效果。

平台背景

山东大学科学计算的应用需求具有多样化碎片化的特点，其中既包括计算机、数学、信息、生命、物理、化学、医学、材料、机械等传统的科学计算相关专业，也包括经济、管理、文学、新闻传播、电气等有新兴的计算需求的专业。绝大多数计算和存储能力的需求不高，但个性化需求较多，工具软件繁杂，用户技术能力和经验参差不齐。此外山东大学具有异地办学的特点，校区众多、人员分散。新系统作为山东大学“大型设备公共技术平台”的一部分[4]，从项目建设伊始就以“统管共享”运行机制为基础，整合汇集全校的科研计算需求，以简便用户应用获得和使用为核心，而计算性能并不作为最关键的考核指标。基于上述现状，结合近些年 IT 技术的新进展和 IT 服务的新模式，形成了具有山东大学特色，以服务山东大学教学科研为目标的，具有公共服务属性的“山东大学高性能计算云平台”（文中简称“平台”）。

服务模式

平台与传统的高性能计算系统的最主要区别是用户服务模式的不同。首先，引入云计算的服务模式，将绝大多数的用户访问和工具软件安装部署等工作通过互联网转移到平台端，减少用户在具体设备和系统侧的工作量和技术难度；其次，通过容器技术对计算资源进行抽象，并封装工具软件和环境，实现工具软件安装部署工作的可复制化，减少重复性工作量，实现对资源的精细化管理调度。此外，实现计算资源抽象后，传统高性能计算系统中的资源独占、资源利用率低的问题就能够得到较好的解决，资源需求并不极端的多个用户可以同时共享计算资源；再次，引入互联网应用中较为普及的社交功能，平台用户可以通过专业领域和科研兴趣形成科研团队，交换计算应用实例和数据，交流平台使用经验，从而形成以科研和学习为目的的协作促进机制，促进多学科的交叉与融合式发展；最后，平台对于每个用户都是独立的学习工作空间，平台的易用程度远高于本地计算机，交互模式又近似于本地计算机，软件资源和数据资源也是即插即用式的，因此这样的环境对提高专业背景不同的用户在学习工作中的“获得感”和“满足感”，并促使其把更多的精力投入到具体的学习科研工作中，提高科研的产出率。因此这种具有公共服务属性的高性能计算系统服务模式就能在实现简便易用的同时，适配尽可能多的应用软件和应用场景。

目标场景

根据山东大学现有与科研计算相关的业务，平台确定服务如下的目标场景：

基于科研项目的中小规模的并行计算需求；

图 1 山东大学高性能计算云平台

个人的程序设计与测试运行；

教学实验课中的实践内容；

基于工作流的计算需求。

其中前两个场景可以直接替代传统高性能计算系统应用场景。后两个场景由于具有了流程化的特征，因此需要与具体的用户需求相结合并融合部分高校业务系统，同时进行较多的教学科研需求分析，目前尚处于需求调研分析阶段。

总体解决方案

平台的系统体系基于传统的 x86 架构处理器，以容器技术对资源进行抽象和调度，同时容器实例对工具软件和环境进行封装。平台与山东大学校园网连接，并利用山东大学信息服务系统实现全校师生的统一认证登录，在校园网提供具有统一服务门户的计算服务。

用户应用方面，应用交互通过浏览器实现。由于部分工具软件和环境已被封装为容器应用模板，因此可以免去软件安装部署和环境配置的工作，简便用户使用。同时容器轻量化的特点，使其更容易实现定制化和迭代的版本管理，并且应用模板的管理也更灵活简便。模板既可以是所有用户通用的，通过类似应用市场分发的应用模板，也可以是用户自定义的私有应用模板。每个应用实例通过应用模板创建，创建实例时，部分应用可以定制化的设置资源使用量，包括实例个数、实例运行时长、内核使用量、内存使用量、异构计算资源的使用量等[5]。平台中每个用户自助创建的应用实例和数据都是私有独立的，并可以实现数据持久化，既用于存储作业完成后的结果数据，也用于作业运算期间结果的存储和数据缓存。用户与应用实例的交互可以通过命令行界面，也可以通过图形界面，其操作的方式与本地计算机类似，且用户本地的计算机无需安装任何软件。目前应用实例还只能适配基于 Linux 环境的操作规则和应用环境[6]。

关键技术

与传统的计算中心或数据中心不同，“山东大学高性能计算云平台”当中的容器实例不提供 Web 应用，而是模仿本地计算机提供用户交互。因此平台在软件开发期间对容器技术做了大量优化和扩充，以使容器实例适应这种类型的工作模式。这其中容器的网络与数据持久化是较为关键的技术。

平台中容器实例的网络既是人机交互的途径，也是实例与数据存储交互的途径，也是创建私有的计算集群时集群节点互联互通的途径。因此传统的以太网结构容易出现不同用户的应用实例网络能够联通，造成旁路数据泄露，并可能由于地址配置的混乱影响应用正常运行。平台借鉴虚拟网络设备的组网方式，为每个用户创建独立的私有网络（VPC），私有网络内包含两个独立的网络区域 Public Vxlan Network 和 Private Vxlan Network，并均与应用实例通过 veth 虚拟网络设备连接[7]，具体如图 2。

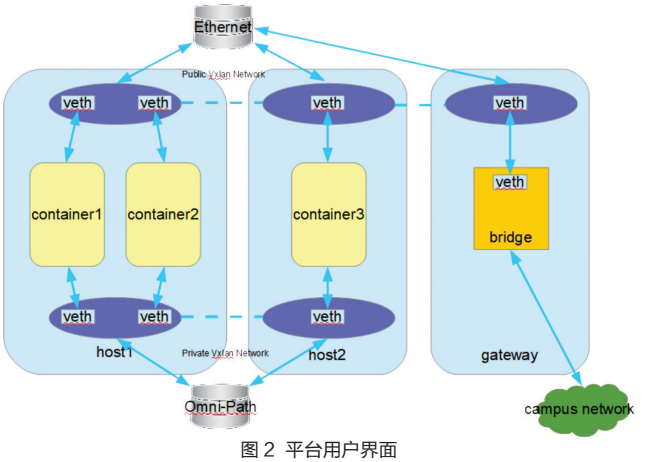

图 2 平台用户界面

其中 Public Vxlan Network 用于为应用实例提供人机交互和数据存储交互等需要容器与外部网络交互的服务（南北向网络），Private Vxlan Network 用于为应用实例间提供交互的能力（东西向网络），Public Vxlan Network 中设置一个网桥节点，以实现与外部网络的连接。在虚拟网络中，用 Vxlan 区分不同主机的网络接入，用 iptables 实现应用实例间的隔离。在物理网络中，Public Vxlan Network 通过每台服务器的 10Gbps 以太网实现连接，Private Vxlan Network 则通过服务器的 100Gbps Omni-Path 组成的网络承载的以太网实现连接，此种组网方式也符合传统高性能计算系统中光纤架构网络独立组网的配置应用习惯。

由于容器的工作特性，必须重新调整容器内的数据存储模式，才能够实现数据存储的功能。平台中每个容器应用模板利用 OverlayFS 重新组织了 RootFS8，实现了类似镜像的模板功能。需要实现数据持久化的存放在独立的外部存储设备上。在应用安装配置时，指定由 Lustre 节点组成的用户存储空间用于存放中间结果，由 NFS 组成的用户存储空间用于存放结果数据。

此外平台还用 Cgroup 实现对不同用户应用实例的资源分配和约束，并在未来结合 Kubernetes 实现资源调度和应用融合。

硬件体系

平台的硬件设备以一体化规划建设和运维为原则，着力打造整体化的硬件体系。其中全部设备设施集中在一套 FusionModule 微模块数据中心内（图 3），包括高性能计算、通用计算、大数据分析、人工智能、胖节点、运维管理等各型服务器约 160 余台，存储设备 2 套，全部服务器和存储节点具备 10Gbps 以太网互联和 100Gbps Omni-Path 光纤架构互联（图 4）。此外作为主要设备供应商的 Huawei 还提供了其 eSight 统一运维管理系统，配合带外的 IPMI 设备运维网络，从而实现整个数据中心全部设备设施运维的集中化，绝

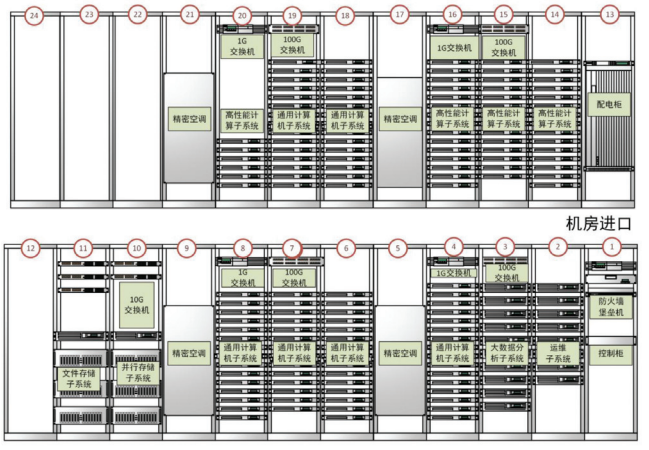

图 3 平台数据中心设备分布

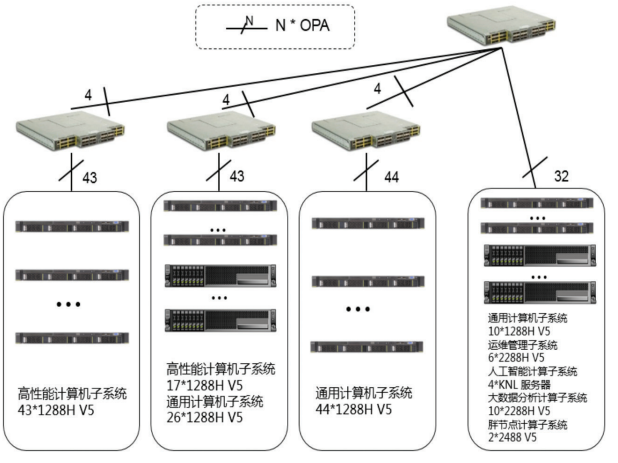

图 4 平台 Omni-Path 网络示意

大多数运维工作无需在数据中心现场完成。最后平台的规划建设期间也重视网络安全的保障能力，配置了专用的多功能防火墙和运维审计设备（堡垒机），网络出口采用了地址转换和端口映射，配合最小化服务的白名单机制，缩小平台的网络暴露范围。

应用现状

平台在 2017 年底开始试运行部分基础性功能，并利用山东大学信息服务系统实现全校师生的统一认证登录，面向全校师生进行试用。经过近一年的测试，平台积累并发布了 20 个应用模板，其中既包括 Gromacs、MOCAT2 等开源软件应用，也包含 MATLAB 这样的大型商用软件。专业领域囊括了数学、生命、材料、化学、计算机、信息、控制、机械、流体、电气、经济等诸多专业，累计创建的应用实例数千余个，累计登陆用户超过 200 人。测试期间整个平台 IT 设备峰值功耗高达 48 千瓦，服务器的 CPU 平均负载接近 40%。

未来发展

近一年的测试证明，山东大学高性能计算云平台为教学和科研提供公共计算服务的实践是成功的，使广大师生能够在相对低的技术难度和成本付出的情况下使用到高性能的计算资源。未来平台规划在如下几个方面进行优化和扩展：

网络资源：在现有网络条件下，借助国家超算的网络传输优化相关课题的成果，对平台网络的接入访问和数据传输性能进行优化提升，增强服务能力和数据吞吐能力，方便大数据量计算任务的承接能力。同时增加 IPv6 的访问接入，使其成为具有山东大学特色的 IPv6 网络服务，并吸引更多的用户在教学科研中应用 IPv6 网络，为 IPv6 网络的发展贡献绵薄之力。

计算和应用资源：整合校内的资金和需求，扩充平台的计算、网络、存储资源，尤其在计算方面引入基于 Ascend 和 Dhyana 等自主可控计算设备。此外扩充应用规模，加入更多的面向更多专业领域的开源或商用软件，规划用户自定义应用模板的相关功能，并丰富软件的不同版本。

计算模式：开放容器集群计算应用，探索不同应用软件在容器集群环境下的优化和应用模式。

服务场景：扩展服务场景。包括平台在教学活动中的应用，提供实训和动手实践，进行新的应用组合，实践在原有环境下不能实现的计算功能。此外还包括利用容器轻量化功能化的特点，将有工作流性质的科研活动在平台中构建并运行，实现流程的可复制，从而促进科研活动的深层次化，使平台成为不同学科领域交叉融合的载体，促进科研活动中人与计算资源，以及人与人之间的交流碰撞。

山东大学高性能计算云平台的建设是一次用互联网思维对传统应用领域的优化，使得人们传统上认为阳春白雪的高性能计算，能够以一种更亲民的姿态呈现。虽然与社交相关的功能在测试期间还没体现出价值，但它却让每个应用有了自己的个性。在目前互联网席卷并重塑各行各业的大背景下，教育和科研，这两个在早期就推动了计算机网络发展成熟的元素，也需要融入新的思维，探索新的思路，以谋求自身在互联网大潮下的自我变革和自我发展。希望我们进行的这些实践对这一过程能够提供一点益处。CEN（责编：王左利）

（作者单位：1 为山东大学软件学院，2 为山东大学资产与实验室管理部，3 为山东大学齐鲁医院信息网络中心）

参考文献

[1] 杨慧，李慎涛，薛冰．冷冻电镜技术：从原子尺度看生命——2017 年诺贝尔化学奖简介 [J]. 首都医科大学学报，2017,(5):770–776. DOI:10.3969/j.issn.1006–7795.2017.05.027.
[2] 林新华，顾一众．上海交通大学高性能计算建设的理念与实践 [J]. 华东师范大学学报（自然科学版），2015,(z1):298–303. DOI:10.3969/j.issn.1000–5641.2015.z1.048.
[3] 沈瑜，李会民，刘晓辉．第一性原理计算软件包 ABACUS 中格点积分的优化 [J]. 科研信息化技术与应用，2015,(5):12–21. DOI:10.11871/j.issn.1674–9480.2015.05.002.
[4] 王文君，刘淑云，白志学，等．高校公共技术服务平台资源共享体系研究 [J]. 实验室研究与探索，2015,(4):259–262,270. DOI:10.3969/j.issn.1006–7167.2015.04.068.
[5] 葛虎．基于 Linux 容器实现 NFV 平台的研究 [J]. 电子技术，2016,(8):10–14. DOI:10.3969/j.issn.1000–0755.206.08.004.
[6] 易升海，彭江强，卿勇军，等．浅析 Docker 容器技术的发展前景 [J]. 电信工程技术与标准化，2018,(6):88–91.
[7] Linux– 虚拟网络设备 –veth pair，https://www.aliyun.com/jiaocheng/123463.html
[8] 深入理解 overlayfs（一）：初识，https://www.aliyun.com/jiaocheng/1375313.html