用户案例

寻因生物

寻因生物：依托生命科学云上大内存解决方案，单细胞测序任务效率提升5倍

寻因生物 SEEKGENE

“这能够使我们的单细胞数据分析业务完全消除IO瓶颈，并在实际的分析任务中将持久内存的大容量能力充分利用起来，让任务的并发能力提升了5倍以上，且该方案能让多细胞数，大样本任务能顺利地运行成功，对提高我们生信用户的业务吞吐能力和工作效率有非常大的助力。”

——寻因生物生信部门 张广鑫

2009年，单细胞测序技术首次问世。四年后，单细胞测序技术被Nature Methods评为年度技术。2015年，单细胞测序技术再度登上Science转化医学封面。目前，单细胞测序的全球潜在科研市场体量已经达到130亿美元。

单细胞测序到底有什么好？传统的研究方法在多细胞水平进行，因此，最终得到的信号值，其实是多个细胞的平均，丢失了异质性信息。从2018年起，单细胞基因测序技术就开始飞速发展。作为一项高效的医疗辅助手段，基因测序在预防出生缺陷、检测遗传性疾病、肿瘤用药等领域提供了有效帮助。

寻因生物是一家专注于单细胞技术的生物科技企业，致力于通过自主研发的高通量单细胞产品、实验及分析全链条服务，将单细胞技术普适化，助力临床诊断和药物研发。

为了进一步提升云上单细胞测序数据分析的效率及用户体验，寻因生物采用了阿里云最新的单细胞测序分析任务的云上解决方案，将业务部署在了基于英特尔® 傲腾™ 持久内存的阿里云i4p持久内存型实例中，并在实例中使用了MemVerge公司开发的 Memory Machine大内存软件，大幅缩短了单细胞测序分析任务的效率，极大提升了用户体验。

挑战：超大细胞样本量及分析复杂性对“大内存”的需求

在生物信息行业，仅一个单细胞测序文件的大小可达100GB以上，而随着一个单细胞项目包含的样本量越来越多，细胞数据级别往往达数百GB甚至TB。随着业务的发展，寻因生物面临生信行业普遍的超大数据量和分析复杂性带来的任务并发度低、数据加载速率慢两大难题。

海量单细胞样本，任务并发度低：在单细胞测序数据分析过程中，每个细胞的表达量数据高达数十万条读取 (reads)，产生的数据更是要大得多，这种海量级的数据分析对云主机的内存容量提出了更高的要求。而通用的云主机的内存容量与CPU配比有限，单细胞的分析任务常会出现因内存不足而导致运行失败；而选用传统的大内存云主机，不仅要付出更加高昂的成本，而且会造成CPU算力的浪费。所以，内存容量的限制使寻因生物不得不采用生信分析的惯用操作——将样本参数调低、或者仅运行一个比较大型的单细胞分析任务。但在测序任务多的情况下，多个单细胞分析项目只能排队执行。

分析复杂性高，数据加载速率慢：单细胞数据的分析复杂，需要反复做数据读取和参数调整。在测序数据分析过程中，每次临时数据在磁盘上的导出和加载（IO）过程长达1000秒，随着数据集的持续增长，这种处理速度阻碍了预期的研究发现时间。当样品量上来，各个样品之间又要做各种关联或者是更复杂的计算，所以对算力的消耗量就会非常大。现在，逐渐又出了很多多组学的检测，在普通单细胞的维度上又加了很多维度，对算力的需求会来到一个更高的水平。

考虑到单细胞的检测和分析将会科研和药物研发领域越来越普及，所需要分析的数据和维度都在增加的情况，寻因生物不得不寻求更优化的计算架构。寻因生物希望通过利用云上最新的技术，减少在基础设施层面的精力，将更多的技术创新资源聚焦在单细胞测序数据分析业务上。

解决方案： 基于英特尔® 傲腾™ 持久内存的阿里云持久内存型实例 i4p 与 MemVerge 大内存软件助力寻因生物测序分析任务提速

2019年，寻因生物与阿里云开展了合作，选择了阿里云ecs.g5、g6、g7三代产品。为了能够更好地解决“内存墙”和“功耗墙”问题、提升测序数据分析的效率，寻因生物将业务迁移至基于英特尔® 傲腾™ 持久内存的阿里云i4p持久内存型实例中，同时使用了阿里云合作伙伴MemVerge公司开发的 Memory Machine大内存软件，具体如图1所示。成功地运行了多细胞数、大样本的测序数据分析任务。

阿里云i4p持久内存实例是基于英特尔® 傲腾™ 持久内存推出的第二代持久内存实例，傲腾持久内存让高性价比的大容量内存与对数据持久性的支持巧妙地结合在一起，将更多数据保存在更靠近CPU的地方，加速了大内存计算， 可以说重新定义了传统的两级存储架构。

除基本vCPU和内存外，阿里云i4p实例还配置了持久内存资源，极大地扩展了主机的内存容量，让内存中可以存放更多数据用于测序数据分析，同时并发运行更多的测序任务，相对于传统普通大内存实例，i4p持久内存实例可以帮助用户打破“内存墙”藩篱，获得更高性能的同时，有效降低整体IT基础设施拥有成本（TCO）。

MemVerge开发的Memory Machine大内存虚拟化软件，可运行在i4p持久内存实例中，将其中的持久内存和普通内存进行融合，可以透明地使用大内存资源，无需对应用进行改造，即可充分发挥持久内存的全部性能；其软件的高级功能“ZeroIO内存快照”，可以完全避免临时数据的磁盘IO过程，实现客户应用性能的飞跃。同时通过阿里云计算巢还实现了Memory Machine大内存虚拟化软件与云平台的标准化集成，实现快速的软件交付部署和标准化的运维管理，大幅提升了业务效率。

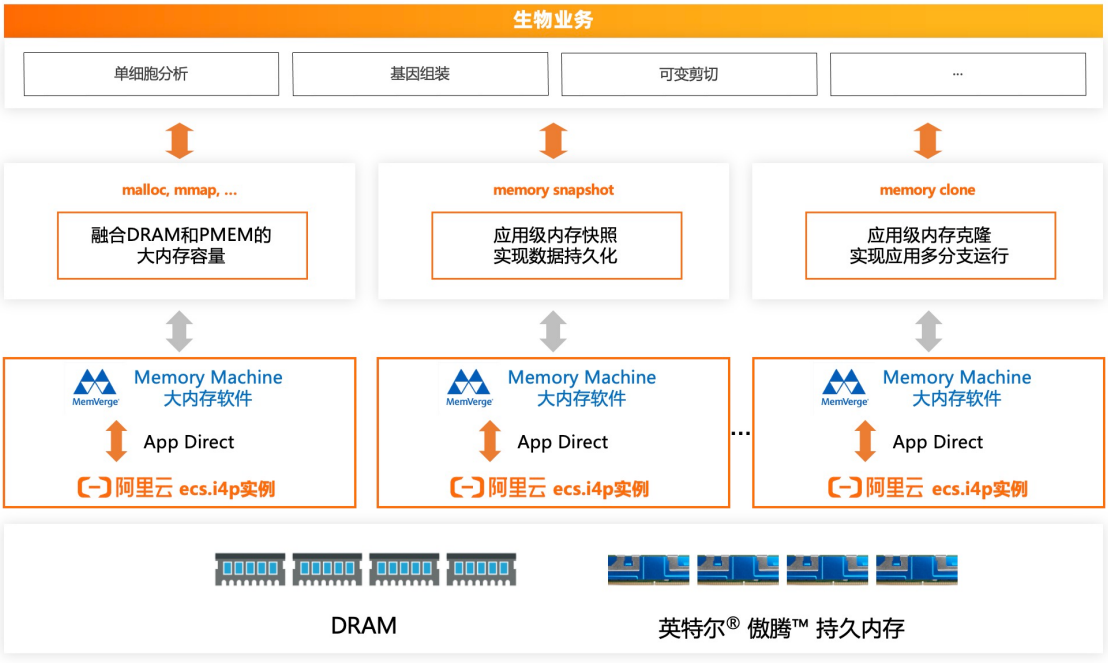

图 1： 生命科学大内存解决方案架构图

业务效果： 寻因生物测序分析任务效率提升 5 倍

正是有过阿里云多代云服务器产品的使用，寻因生物生物部门张广鑫对“上云”的评价直接了当：算得快、成本低。

最终经过测算，寻因生物的单细胞基因测序，数据加载和导出性能从1000秒缩至2.5秒；单任务的样本规模是原来的2倍。在运行时间和单任务的运行时间几乎差不多的情况下，测序任务的并发运行数由原来的1个提升到了5个，任务处理效率提升了5倍之多。

展望

寻因生物透露，目前正在测试阿里云E-HPC相关产品，希望通过E-HPC平台来简化编写流程、监控任务投递，以及任务运算的过程；同时也考虑进一步加强云上资源的精细化管理，更好的用好云平台的弹性和高并发能力，构建更加高效的分析平台，将单细胞技术普适化，助力临床诊断和药物研发。

阿里云也将与英特尔、Memverge等合作伙伴加强合作，以技术创新推动云服务产品和方案的差异化创新，进而推动行业数字化升级。

关于寻因生物

寻因生物是一家专注于单细胞技术的生物科技企业， 由资深转化医学和技术研发经验的李宗文博士和焦少灼联合创立，致力于通过自主研发的高通量单细胞产品、 实验及分析全链条服务， 将单细胞技术普适化， 助力临床诊断和药物研发， 推动精准医疗进入 2.0 时代。 寻因拥有完全自主知识产权的 SeekOne® 高通量单细胞建库平台、 SeekGene® Online 自动化在线数据分析平台， 以及国内唯一的液滴法 + 微孔法自研双平台测序能力。

www.seekgene.com

关于英特尔

英特尔（NASDAQ：INTC）作为行业引领者，创造改变世界的技术，推动全球进步并让生活丰富多彩。在摩尔定律的启迪下，我们不断致力于推进半导体设计与制造，帮助我们的客户应对最重大的挑战。通过将智能融入云、网络、边缘和各种计算设备，我们释放数据潜能，助力商业和社会变得更美好。

www.intel.cn

关于 MemVerge

MemVerge是大内存计算和大内存云技术的先行者，其研发的Memory Machine 软件是业界第一款虚拟化内存硬件的软件，用于对容量、性能、可用性等进行精细化的资源调配。在透明内存服务的基础上，Memory Machine还提供了ZeroIO内存快照，可以在几秒钟内封装数TB的应用程序状态，并以内存速度实现数据管理。

www.memverge.net

阿里云

阿里云

阿里云创立于 2009 年，是全球领先的云计算及人工智能科技公司，为 200 多个国家和地区的企业、开发者和政府机构提供服务。阿里云致力于以在线公共服务的方式，提供安全、可靠的计算和数据处理能力，让计算和人工智能成为普惠科技。2017 年 1 月阿里云成为奥运会全球指定云服务商。