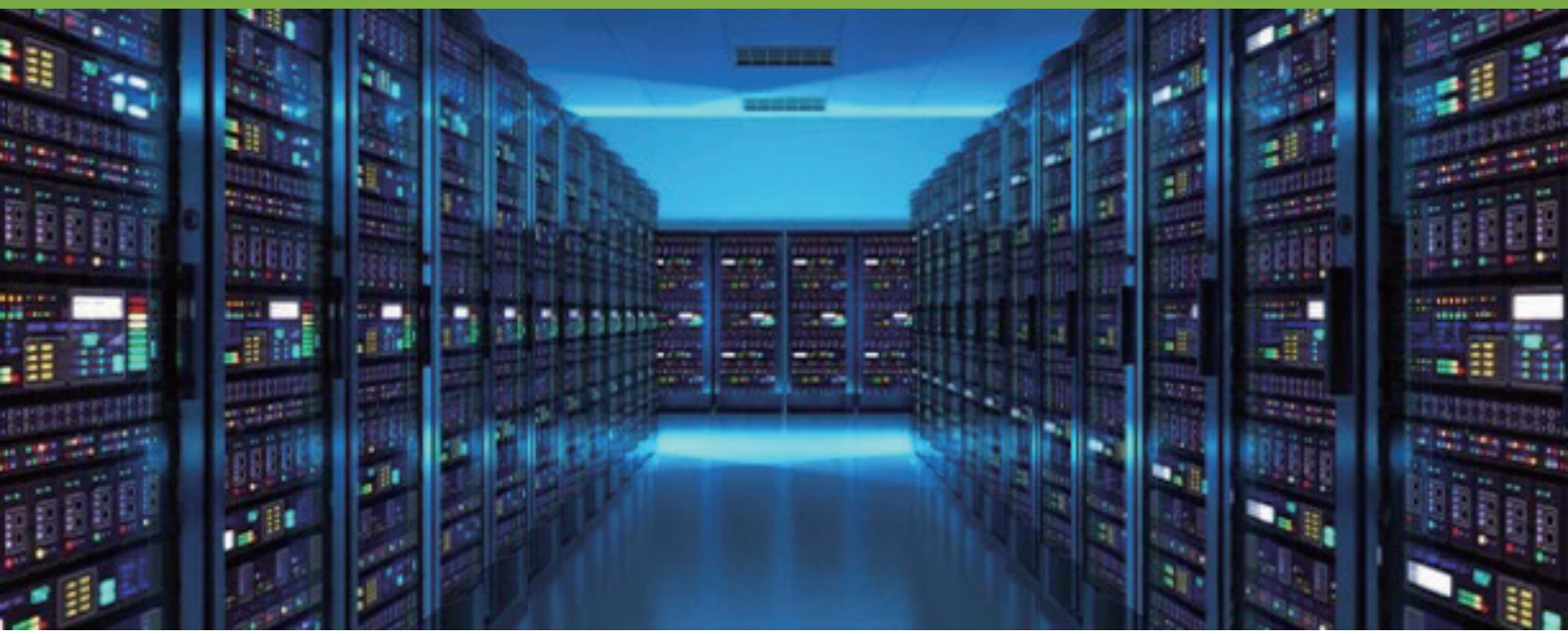
SUCCESS STORY | 极客天成

NVIDIA BlueField-2 DPU 助力极客天成构建云原生高速存储

极客天成 NVMatrix 云原生高性能存储解决方案全栈支撑大型金融行业客户的人工智能云平台



NVIDIA

SUMMARY

充分利用 NVIDIA BlueField-2 InfiniBand DPU 架构和 NVIDIA InfiniBand 网络的出色性能优势，融合极客天成软件定义块存储的技术优势，NVMatrix 云原生存储解决方案打破了传统存储的种种限制，实现了存储的云原生：

- 实现免装机流程：操作系统会在云盘预先安装好，秒级直接拷贝和克隆镜像副本；
- 分钟级物理机交付：操作系统云盘即用即挂载，网络即插即用，只有服务器启动时间；
- 故障恢复时间短：当服务器故障时，只需将云盘挂载到另一台服务器启动，即可实现分钟级别的快速恢复使用；
- 弹性易扩容：相比本地磁盘，高速云盘可随时快速进行磁盘扩容；
- 数据安全可靠：存储端采用分布式存储集群提供二或三副本，数据可靠性高。支持高级数据医生功能，通过快照技术数据可恢复到任意一秒。支持数据灾备功能，可以通过灾备将数据备份到异地机房。
- 云原生：基于 Kubernetes CSI 的云原生接口为计算节点提供动态弹性的高速存储服务，支持多租户， 高安全性和裸金属性能。

INTRODUCTION

在现代大型数据中心中，提升存储的性能和提高存储的利用率是一个永不过时的话题，各种存储解决方案迅速演进，从不同的角度来解决当前存储的各种不足之处，并满足每个客户日益增加的特殊需求。

那么如何构建低成本、高性能、低能耗、可扩展且安全的数据中心来承载大型机构庞大的数据存储，如何优化利用数据为他们的客户提供可靠、高效的服务，就成为亟待解决的问题。这不但对数据中心的性能提出了更高的要求，而且需要在云原生环境中为存储工作负载提供网络和安全加速的基础设施。作为英伟达初创加速计划 （NVIDIA Inception ）会员企业，极客天成 NVMatrix 云原生存储解决方案在数据中心存储方面，很好的解决了这一问题，它利用 NVIDIA BlueField-2 InfiniBand DPU 及 NVIDIA InfiniBand 网络构建了云原生高速存储，使存储中无法得到充分利用的容量变为有效容量，高效地弹性扩展数据，实现了数据中心存储的高性能和安全性。

CHALLENGE STATEMENT

国内某头部大型金融公司致力于打造行业领先的人工智能应用研发、部署及统一运行的云平台，提供统一的数据、算力和研发等服务。该金融公司的人工智能云平台是基于 Kubernetes 的云原生平台，以提供面向内部人工智能应用的基础设施云原生服务。提供安全、可靠、高性能且可扩展的持久性存储是人工智能云平台的关键指标之一。然而，要在Kubernetes环境中来达到这个目标是一项巨大的挑战。该金融公司原有的数据中心使用的是直连式 NVMe（Non-Volatile Memory Express 非易失性内存接口）存储，Kubernetes 编排支持使用 NVMe SSD 作为本地持久卷，但无法在编排层和容器层来提供数据保护，这就需要应用自身来提供高可用性，否则当容器从当前服务器迁移到另一台服务器上之后就无法访问它们以前写入的数据。

因此，要解决这个问题，需要一个集中式的、且带有冗余功能的存储解决方案，其性能与本地 NVMe SSD 基本相同，可与容器配合运行，并且满足数据冗余保护的要求。该金融公司考虑和尝试了多种替代解决方案，如各种传统的数据中心存储解决方案，发现有的方案性能不足，有的方案安全性不够，有的方案不可扩展，有的方案成本太高，有的方案实现起来太复杂等，故一直没能找到能和该金融公司成本结构及业务案例完全匹配的方案。

该金融公司经过多方调研和分析，为他们的存储解决方案明确了如下的需求：

1. 所需的持久性存储要能够提供与本地 NVMe SSD 闪存性能水平相当的　高可用性；
2. 必须能与 Kubernetes 集成，并通过容器存储接口（CSI–Container Storage Interface）规范与 Kubernetes 集成；
3. 必须支持多租户（Multi Tenancy），并支持多租户之间的安全隔离；
4. 支持不同操作系统（如采用 Linux 和 Windows 计算节点）在同一套存储平台下的高速数据存储和 OS 镜像加载；
5. 高扩展性，面向未来的海量数据可以轻松扩展，不必改变存储架构；
6. 对于业务的透明性，在不影响业务的前提下实现业务向新存储的迁移；
7. 高性价比，面向节能减碳。

SOFTWARE

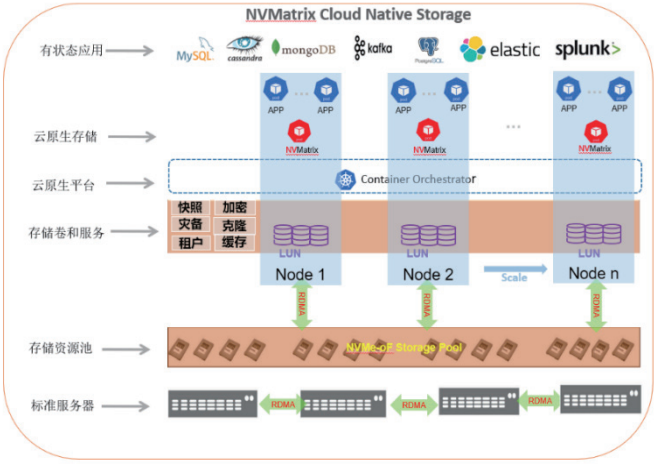
DOCA 软件开发套件

SOLUTION STATEMENT

经过极客天成（ScaleFlash）和该金融公司的多次技术沟通和 PoC（新概念验证），最终该公司选择了极客天成的基于 NVIDIA BlueField-2 InfiniBand DPU 的 NVMatrix 高性能存储解决方案，全面满足了上述的所有需求。

极客天成的 NVMatrix 存储解决方案提供了可与 Kubernetes 集成的完全云原生持久性存储，并通过 NVIDIA BlueField-2 InfiniBand DPU 将计算和存储解耦，以集群式存储的方式实现了前所未有的可扩展性和高可用性。凭借优化的 NVMe-oF（NVMe Over Fabric）前端和智能的后端分布式存储管理，充分发挥了 RDMA 技术和 NVIDIA InfiniBand 网络的在存储中技术优势，达到了极高的 IOPS 性能，从应用的角度来看，可与直连式NVMe 固态硬盘（SSD）媲美。

极客天成的云原生存储 NVMatrix 如下图所示：



极客天成的基于 NVIDIA BlueField-2 InfiniBand DPU 打造的 NVMatrix 云原生存储解决方案在架构上主要包含了极客天成的支持容器卷的高性能软件定义存储模块，NVIDIA InfiniBand 端到端网络（含 NVIDIA NVMe SNAP 软件）及高性能的 NVMe SSD 盘，具体配置如下：

计算节点：

- CPU：2 颗 Intel Xeon 6346 16 核 3.1GHz
- 内存：256GDDR4
- 操作系统：CentOS 7.8 或者 Windows Server 2016
- DPU卡：NVIDIA BlueFied-2 InfiniBand DPU
- 软件：极客天成智能网卡嵌入式系统软件 v3.0

存储节点：

- CPU：2 颗 Intel Xeon 6346 16 核 3.1GHz
- 内存：512GDDR4
- 操作系统：CentOS 7.8
- 网卡：NVIDIA ConnectX-5 InfiniBand 双端口网卡
- 存储：4 块 3.84TB NVMe SSD
- 软件：极客天成分布式存储软件 v3.0

交换网络：

- NVIDIA MSB7800 InfiniBand 交换机

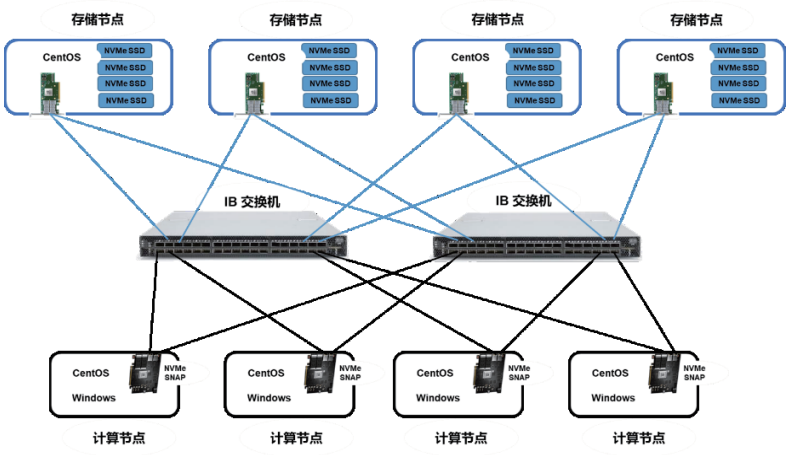
HARDWARE

NVIDIA BlueField-2 DPU

NVIDIA ConnectX-5 InfiniBand 双端口网卡

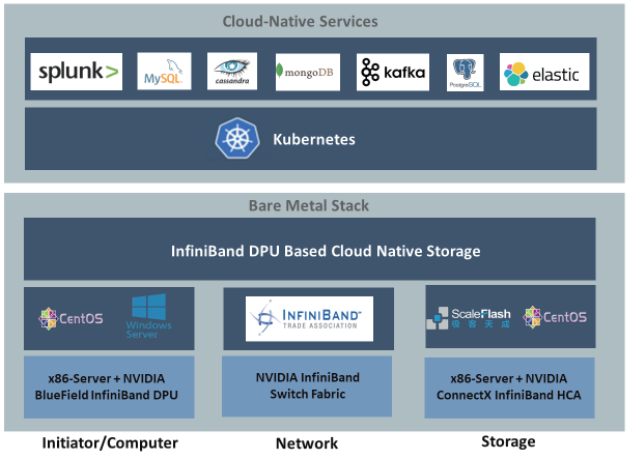
NVIDIA MSB7800 InfiniBand 交换机

极客天成基于 NVIDIA InfiniBand 网络的 NVMatrix 云原生存储系统拓扑如下图所示：



通过采用双端口的 NVIDIA BlueField-2 InfiniBand DPU 卡和 NVIDIA ConnectX-5 InfiniBand 网卡，可以自动利用的 NVIDIA InfiniBand 网络的 Multi-Rail 技术实现双端口的流量均衡，通过双交换机实现了全局的冗余。在计算节点上，DPU 将扮演 InfiniBand 网卡和 NVMe SSD 盘的双重角色，对于主机操作系统来讲，实现对于 NVMeoF 的完全透明，即在主机 OS 下只需安装网卡的驱动，即可以实现通过NVMeoF对存储的远程访问。在存储节点上，也可以利用 InfiniBand 网卡的 NVMeoF Target 卸载功能，进一步提升存储的性能和降低对于 CPU 的消耗。

极客天成完整的 NVMatrix 云原生存储解决方案软件栈如下图所示：



从上图可以看出，基于软件定义块存储的极客天成 NVMatrix 高性能软件栈，以及 NVIDIA BlueField-2 InfiniBand DPU 和端到端 NVIDIA InfiniBand 高速网络，是实现这个高性能存储架构的核心。

随着网络、存储等各种 IO 服务的带宽不断增加，各种相关的 IO 处理对 CPU 的消耗呈现快速增长的局面。这样，底层基础设施负载所占的 CPU 资源越来越多，留给用户应用的 CPU 资源越来越少。通过在计算节点上使用 NVIDIA BlueField-2 InfiniBand DPU 取代了传统的网卡，实现了把基础设施层的任务从 Host CPU 转移到了 DPU 中，把完整的 CPU 资源都交给了业务，达到了业务与管理、通信和安全的分离。NVIDIA BuleField-2 InfiniBand DPU 以其 100Gb/s 的高速传输速度可以实现主机端（Initiator）

和存储目标端（Target）的高速通信；InfiniBand RDMA 技术实现了在存储通信过程中对于 DPU 上 ARM CPU 的零消耗。

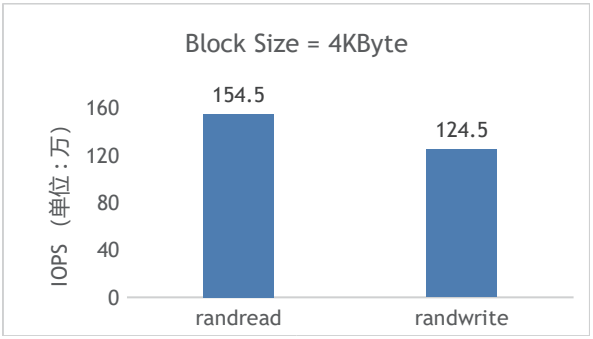
NVIDIA BlueField-2 InfiniBand DPU 还可被用于屏蔽操作系统差异性，构建统一计算资源池，实现计算和存储的分离，并支持容器的 Linux 环境和 Windows 环境的动态切换。 通过 DPU 的 NVMe SNAP 技术，可以在主机内通过 DPU 模拟成 NVMe 设备，数据层借助 DPU 的 ASIC 芯片高速转发能力，配合极客天 成的软件定义块存 储软件栈，通过高速 InfiniBand 无损网络来实现 NVMe-oF 从主机端到存储端的数据传输，直接连接到目标端的高通量分布式存储集群，并通过极客天成的高性能存储软件栈实现了数据的高可用性，达到了用户期望的计算节点物理机可以无缝地接入灵活可扩展的高速云盘的目标。

同时从主机端到存储端采用 NVIDIA InfiniBand 网络的端到端连接也是性能保障的关键，InfiniBand 网络是天然的无损网络，且在使用过程中无需对网络做任何配置，就可以保障整个存储网的端到端无损连接，真正实现了网络的即插即用。这种天然的无损网络保障也使 InfiniBand 网络达到了系统总线一级的传输性能，这对于在实现存储和计算解耦后的性能保障至关重要。

RESULT STATEMENT

从用户的实测性能来看，极客天成的基于 NVIDIA BlueField-2 InfiniBand DPU 的存储解决方案达到了裸金属云盘超过一百万 IOPS 的超高性能，真正实现了灵活性与性能的兼顾。

极客天成 NVMatrix 云原生存储解决方案的性能表现：



（基于上述计算节点、存储节点和网络配置）

云原生计算已经成为了趋势，随着云原生计算的发展，云原生存储将逐渐取代传统的存储架构，成为支撑未来数据中心即计算单元的重要支柱之一，极客天成 NVMatrix 云原生高性能存储解决方案为云上超级算力提供高性能、高性价比、高可用、易扩展及跨平台的存储架构。

了解更多

To learn more about NVIDIA BlueField DPU product and NVIDIA Inception , visit: **https://www.nvidia.cn**
For more information on partner, visit: **www.scaleflash.com**