NetApp AIPod Mini - 利用NetApp和 Intel 进行企业 RAG 推理

NetApp artificial intelligence solutions

NetApp
August 20, 2025

This PDF was generated from https://docs.netapp.com/zh-cn/netapp-solutions-ai/infra/ai-minipod.html on August 20, 2025. Always check docs.netapp.com for the latest.

目录

NetApp AIPod Mini - 利用NetApp和 Intel 进行企业 RAG 推理 . . . 1
内容提要 . . . 1
英特尔存储合作伙伴验证 . . . 1
使用NetApp运行 RAG 系统的优势 . . . 1
目标受众 . . . 2
技术要求 . . . 2
硬件 . . . 2
软件 . . . 3
解决方案部署 . . . 4
软件堆栈 . . . 4
部署步骤 . . . 5
启用ONTAP S3 . . . 10
配置计划同步设置 . . . 10
更改卷访问模式 . . . 11
（可选）禁用 SSL 证书验证 . . . 11
尺寸指南 . . . 13
结束语 . . . 13
致谢 . . . 14
物料清单 . . . 14
在哪里可以找到更多信息 . . . 15

NetApp AIPod Mini - 利用NetApp和 Intel 进行企业 RAG 推理

本文介绍了NetApp AIPod for Enterprise RAG 的经过验证的参考设计，该设计采用了 Intel Xeon 6 处理器和NetApp数据管理解决方案的技术和组合功能。该解决方案演示了下游 ChatQnA 应用程序利用大型语言模型，为并发用户提供准确、上下文相关的响应。这些响应是通过隔离的 RAG 推理管道从组织的内部知识库中检索的。

Sathish Thyagarajan、Michael Oglesby、 NetApp

内容提要

越来越多的组织正在利用检索增强生成 (RAG) 应用程序和大型语言模型 (LLM) 来解释用户提示并生成响应，以提高生产力和商业价值。这些提示和响应可以包括从组织内部知识库、数据湖、代码存储库和文档存储库检索到的文本、代码、图像，甚至治疗蛋白质结构。本文介绍了NetApp AIPod Mini 解决方案的参考设计，包括NetApp AFF存储和搭载 Intel Xeon 6 处理器的服务器。它包括与英特尔高级矩阵扩展 (Intel AMX) 相结合的NetApp ONTAP数据管理软件，以及基于企业 AI 开放平台 (OPEA) 构建的英特尔企业检索增强生成 (RAG) 软件。适用于企业 RAG 的NetApp AIPod Mini 使组织能够将公共 LLM 增强为私有生成式 AI (GenAI) 推理解决方案。该解决方案展示了企业规模的高效且经济的 RAG 推理，旨在提高可靠性并让您更好地控制您的专有信息。

英特尔存储合作伙伴验证

搭载英特尔至强 6 处理器的服务器专为处理要求苛刻的 AI 推理工作负载而设计，并使用英特尔 AMX 实现最佳性能。为了实现最佳的存储性能和可扩展性，该解决方案已使用NetApp ONTAP成功验证，使企业能够满足 RAG 应用程序的需求。该验证是在配备 Intel Xeon 6 处理器的服务器上进行的。英特尔和NetApp建立了强大的合作伙伴关系，致力于提供优化、可扩展且符合客户业务需求的 AI 解决方案。

使用NetApp运行 RAG 系统的优势

RAG 应用程序涉及从公司各种类型的文档存储库中检索知识，例如 PDF、文本、CSV、Excel 或知识图。这些数据通常存储在诸如 S3 对象存储或 NFS 本地解决方案中作为数据源。 NetApp一直是边缘、数据中心和云生态系统中数据管理、数据移动性、数据治理和数据安全技术的领导者。 NetApp ONTAP数据管理提供企业级存储，以支持各种类型的 AI 工作负载，包括批量和实时推理，并提供以下一些优势：

- 速度和可扩展性。您可以高速处理大型数据集以进行版本控制，并能够独立扩展性能和容量。
- 数据访问。多协议支持允许客户端应用程序使用 S3、NFS 和 SMB 文件共享协议读取数据。 ONTAP S3 NAS 存储桶可以促进多模式 LLM 推理场景中的数据访问。
- 可靠性和保密性。 ONTAP提供数据保护、内置NetApp自主勒索软件保护 (ARP) 和动态存储配置，并提供基于软件和硬件的加密以增强机密性和安全性。 ONTAP 的所有 SSL 连接均符合 FIPS 140-2 标准。

目标受众

本文档适用于希望利用为提供企业 RAG 和 GenAI 解决方案而构建的基础设施的 AI 决策者、数据工程师、业务领导者和部门主管。对 AI 推理、LLM、Kubernetes 以及网络及其组件的先前了解将有助于实施阶段。

技术要求

硬件

英特尔人工智能技术

使用 Xeon 6 作为主机 CPU，加速系统可受益于高单线程性能；更高的内存带宽；更高的可靠性、可用性和可服务性 (RAS)；以及更多的 I/O 通道。英特尔 AMX 加速了 INT8 和 BF16 的推理，并支持 FP16 训练模型，INT8 每核每周期最多可进行 2,048 次浮点运算，BF16/FP16 每核每周期最多可进行 1,024 次浮点运算。要使用 Xeon 6 处理器部署 RAG 解决方案，通常建议至少使用 250GB 的 RAM 和 500GB 的磁盘空间。然而，这在很大程度上取决于 LLM 模型的大小。欲了解更多信息，请参阅英特尔 "Xeon 6处理器"产品简介。

图 1 - 搭载 Intel Xeon 6 处理器的计算服务

器

NetApp AFF 存储系统

入门级和中级NetApp AFF A 系列系统提供更强大的性能、密度和更高的效率。 NetApp AFF A20、 AFF A30 和AFF A50 系统提供真正的统一存储，支持块、文件和对象，基于单一操作系统，可以以最低的成本在混合云中无缝管理、保护和调动 RAG 应用程序的数据。

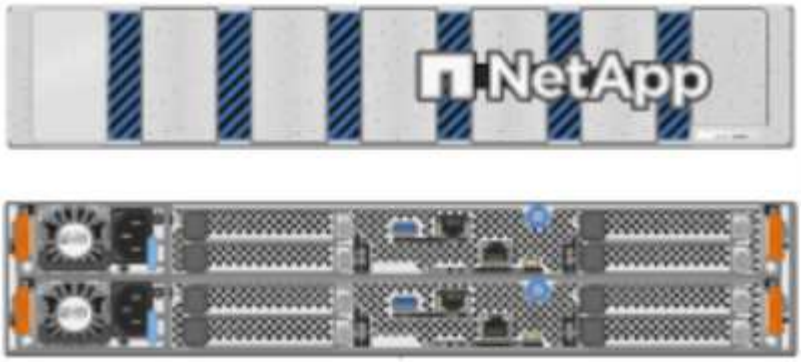


图 2 - NetApp AFF A 系列系统。

硬件	数量	评论
基于 Intel Xeon 6 的服务器	2	RAG 推理节点 - 配备双插槽 Intel Xeon 6900 系列或 Intel Xeon 6700 系列处理器以及 250GB 至 3TB RAM，配备 DDR5 (6400MHz) 或 MRDIMM (8800MHz)。 2U服务器。
带有英特尔处理器的控制平面服务器	1	Kubernetes 控制平面/1U 服务器。
100Gb 以太网交换机的选择	1	数据中心交换机。
NetApp AFF A20（或AFF A30；AFF A50）	1	最大存储容量：9.3PB。注意：网络：10/25/100 GbE 端口。

为了验证此参考设计，我们使用了 Supermicro 的 Intel Xeon 6 处理器服务器（222HA-TN-OTO-37）和 Arista 的 100GbE 交换机（7280R3A）。

软件

企业**AI**开放平台

企业 AI 开放平台 (OPEA) 是由英特尔与生态系统合作伙伴共同主导的一项开源计划。它提供了一个可组合构建块的模块化平台，旨在加速尖端生成 AI 系统的开发，重点关注 RAG。 OPEA 包括一个综合框架，该框架具有 LLM、数据存储、提示引擎、RAG 架构蓝图以及基于性能、特性、可信度和企业准备度评估生成式 AI 系统的四步评估方法。

OPEA 的核心包括两个关键部分：

- GenAIComps：由微服务组件组成的基于服务的工具包
- GenAIExamples：可立即部署的解决方案，例如 ChatQnA，可展示实际用例

有关详细信息，请参阅 "OPEA项目文档"

由 **OPEA** 提供支持的英特尔企业人工智能推理

适用于英特尔企业人工智能 RAG 的 OPEA 简化了将企业数据转化为可操作见解的过程。它由英特尔至强处理器提供支持，集成了来自行业合作伙伴的组件，为部署企业解决方案提供了简化的方法。它可以与成熟的编排框架无缝扩展，提供企业所需的灵活性和选择。

在 OPEA 的基础上，英特尔企业人工智能 RAG 通过增强可扩展性、安全性和用户体验的关键功能扩展了这一基础。这些功能包括与现代基于服务的架构无缝集成的服务网格功能、用于管道可靠性的生产就绪验证以及用于 RAG 即服务的功能丰富的 UI，从而可以轻松管理和监控工作流程。此外，英特尔和合作伙伴的支持提供了广泛的解决方案生态系统，结合集成的身份和访问管理 (IAM) 与 UI 和应用程序，实现安全且合规的操作。可编程护栏对管道行为提供细粒度的控制，实现定制的安全性和合规性设置。

NetApp ONTAP

NetApp ONTAP是 NetApp 关键数据存储解决方案的基础技术。 ONTAP包含各种数据管理和数据保护功能，例如针对网络攻击的自动勒索软件保护、内置数据传输功能和存储效率功能。这些优势适用于一系列架构，从本地到 NAS、SAN、对象和 LLM 部署的软件定义存储中的混合多云。您可以在ONTAP集群中使用ONTAP S3 对象存储服务器来部署 RAG 应用程序，从而利用通过授权用户和客户端应用程序提供的ONTAP的存储效率和安全性。有关详细信息，请参阅 "了解ONTAP S3 配置"

NetApp Trident

NetApp Trident软件是一款开源且完全受支持的存储编排器，适用于容器和 Kubernetes 发行版，包括 Red Hat OpenShift。 Trident可与整个NetApp存储产品组合配合使用，包括NetApp ONTAP ，并且还支持 NFS 和 iSCSI 连接。有关详细信息，请参阅 "Git 上的NetApp Trident"

软件	版本	评论
面向企业 RAG 的英特尔 AI 的 OPEA	1.1.2	基于OPEA微服务的企业RAG平台
容器存储接口（CSI驱动程序）	NetApp Trident 25.02	支持动态配置、 NetApp Snapshot 副本和卷。
Ubuntu	22.04.5	双节点集群上的操作系统
容器编排	Kubernetes 1.31.4	运行 RAG 框架的环境
ONTAP	ONTAP 9.16.1P4	AFF A20 上的存储操作系统。它具有 Vscan 和 ARP 功能。

解决方案部署

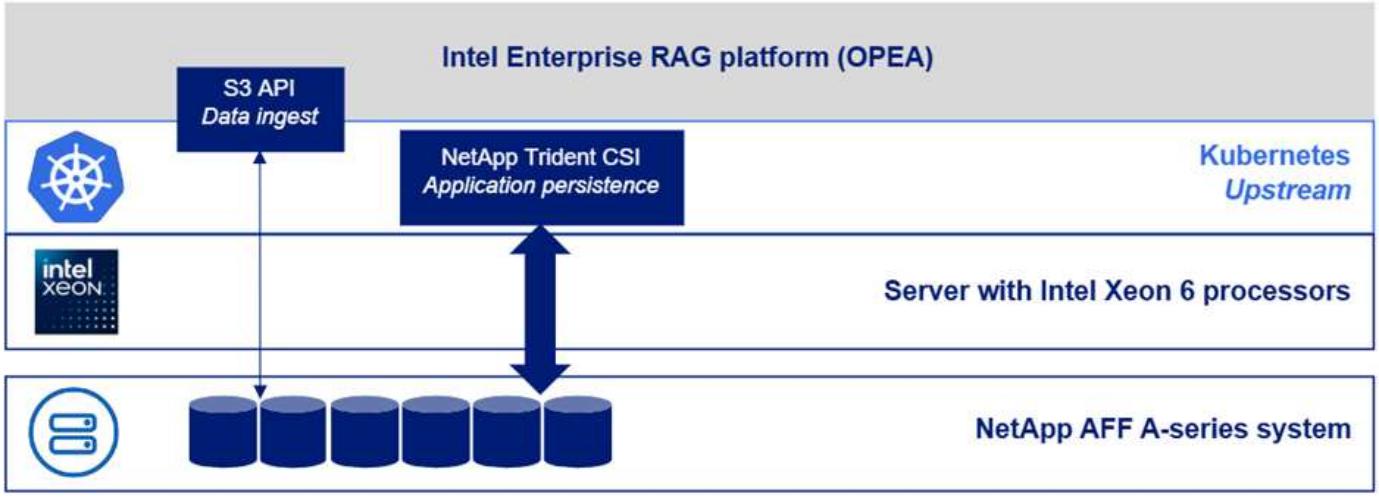
软件堆栈

该解决方案部署在由基于 Intel Xeon 的应用节点组成的 Kubernetes 集群上。至少需要三个节点才能实现 Kubernetes 控制平面的基本高可用性。我们使用以下集群布局验证了该解决方案。

表 3 - Kubernetes 集群布局

节点	角色	数量
配备 Intel Xeon 6 处理器和 1TB RAM 的服务器	应用节点、控制平面节点	2
通用服务器	控制平面节点	1

下图描述了该解决方案的“软件堆栈视图”
。



部署步骤

部署**ONTAP**存储设备

部署和配置您的NetApp ONTAP存储设备。请参阅 "ONTAP硬件系统文档"了解详情。

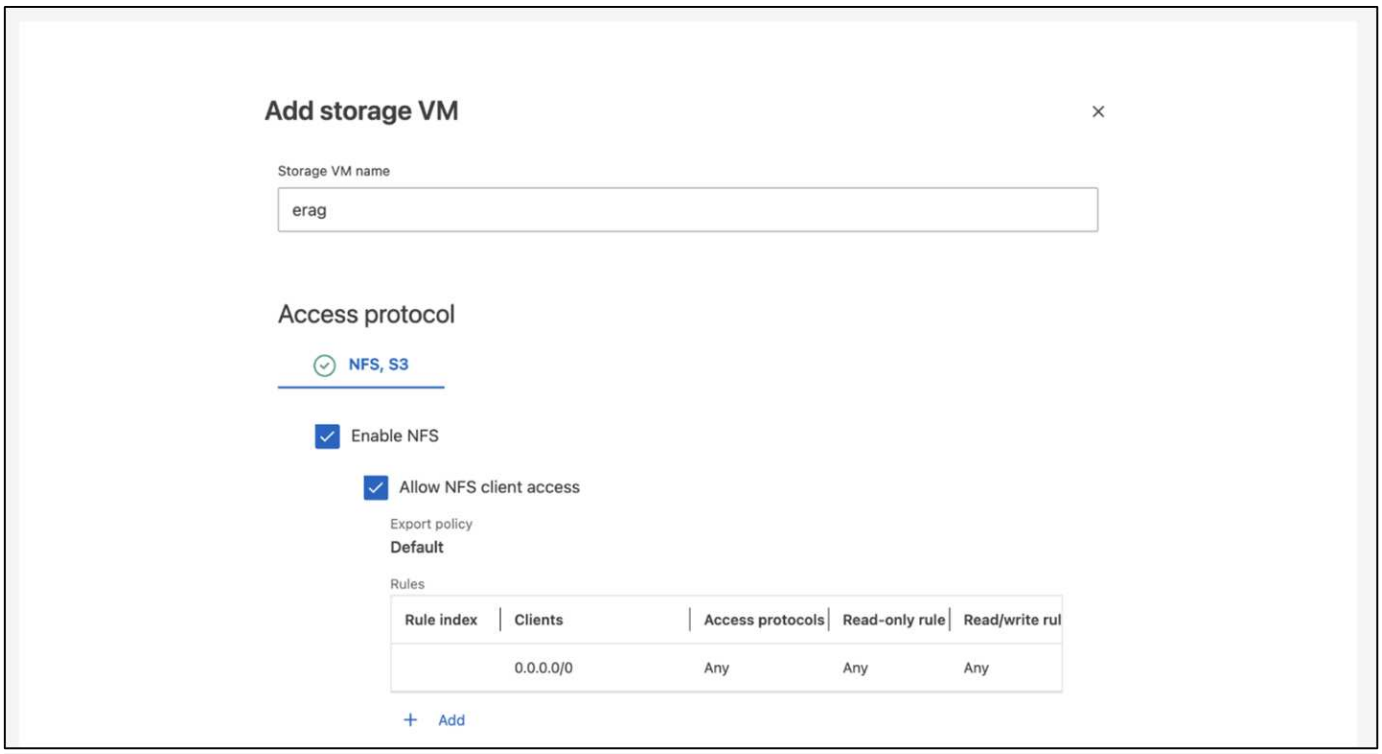
配置**ONTAP SVM** 以进行 **NFS** 和 **S3** 访问

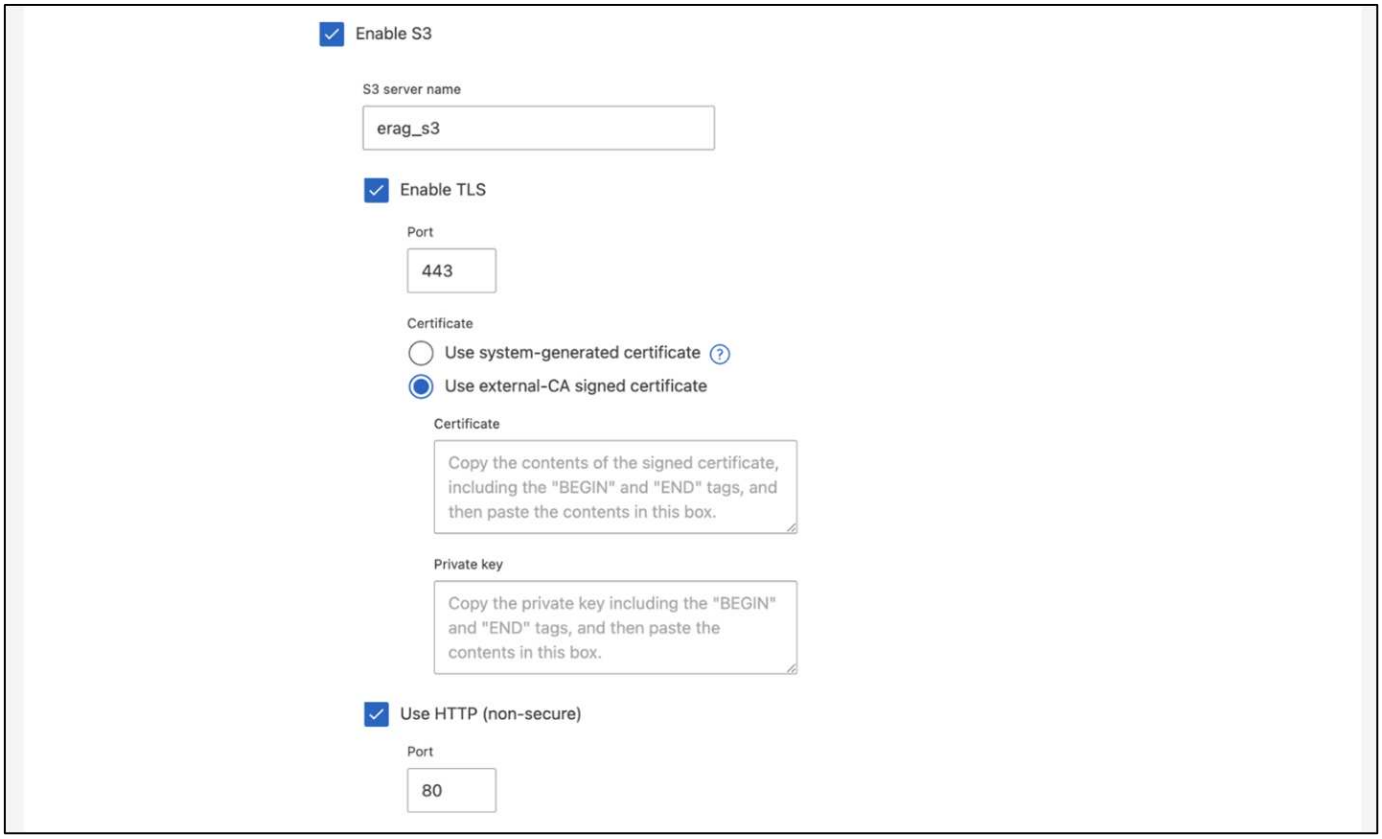
在 Kubernetes 节点可访问的网络上配置ONTAP存储虚拟机 (SVM) 以进行 NFS 和 S3 访问。

要使用ONTAP系统管理器创建 SVM，请导航到“存储”>“存储虚拟机”，然后单击“+ 添加”按钮。为您的 SVM 启用 S3 访问时，请选择使用外部 CA（证书颁发机构）签名的证书，而不是系统生成的证书。您可以使用自签名证书或由公众信任的 CA 签名的证书。有关更多详细信息，请参阅 "ONTAP文档。"

以下屏幕截图展示了使用ONTAP系统管理器创建 SVM 的过程。根据您的环境根据需要修改详细信息。

图 4 — 使用ONTAP系统管理器创建 SVM
。





配置 **S3** 权限

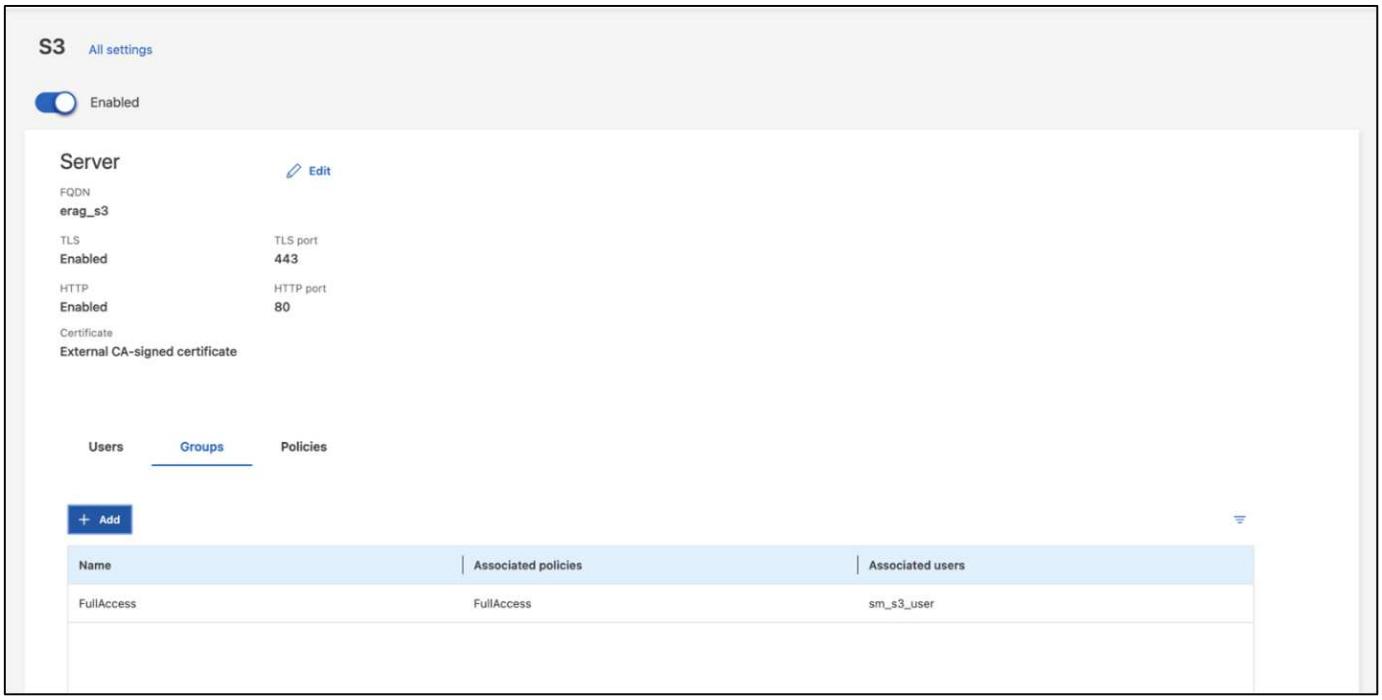
为您在上一步中创建的 SVM 配置 S3 用户/组设置。确保您拥有对该 SVM 的所有 S3 API 操作具有完全访问权限的用户。有关详细信息，请参阅ONTAP S3 文档。

注意：英特尔 AI for Enterprise RAG 应用程序的数据提取服务需要此用户。如果您使用ONTAP系统管理器创建了 SVM，系统管理器将自动创建一个名为 `sm_s3_user`以及一个名为 `FullAccess`当您创建 SVM 时，

但尚未分配任何权限 `sm_s3_user。

要编辑此用户的权限，请导航至“存储”>“存储虚拟机”，单击您在上一步中创建的 SVM 的名称，单击“设置”，然后单击“S3”旁边的铅笔图标。给予 `sm_s3_user`拥有所有 S3 API 操作的完全访问权限，创建一个关联 `sm_s3_user`与 `FullAccess`策略如下面的屏幕截图所示。

图 5 - S3 权限。

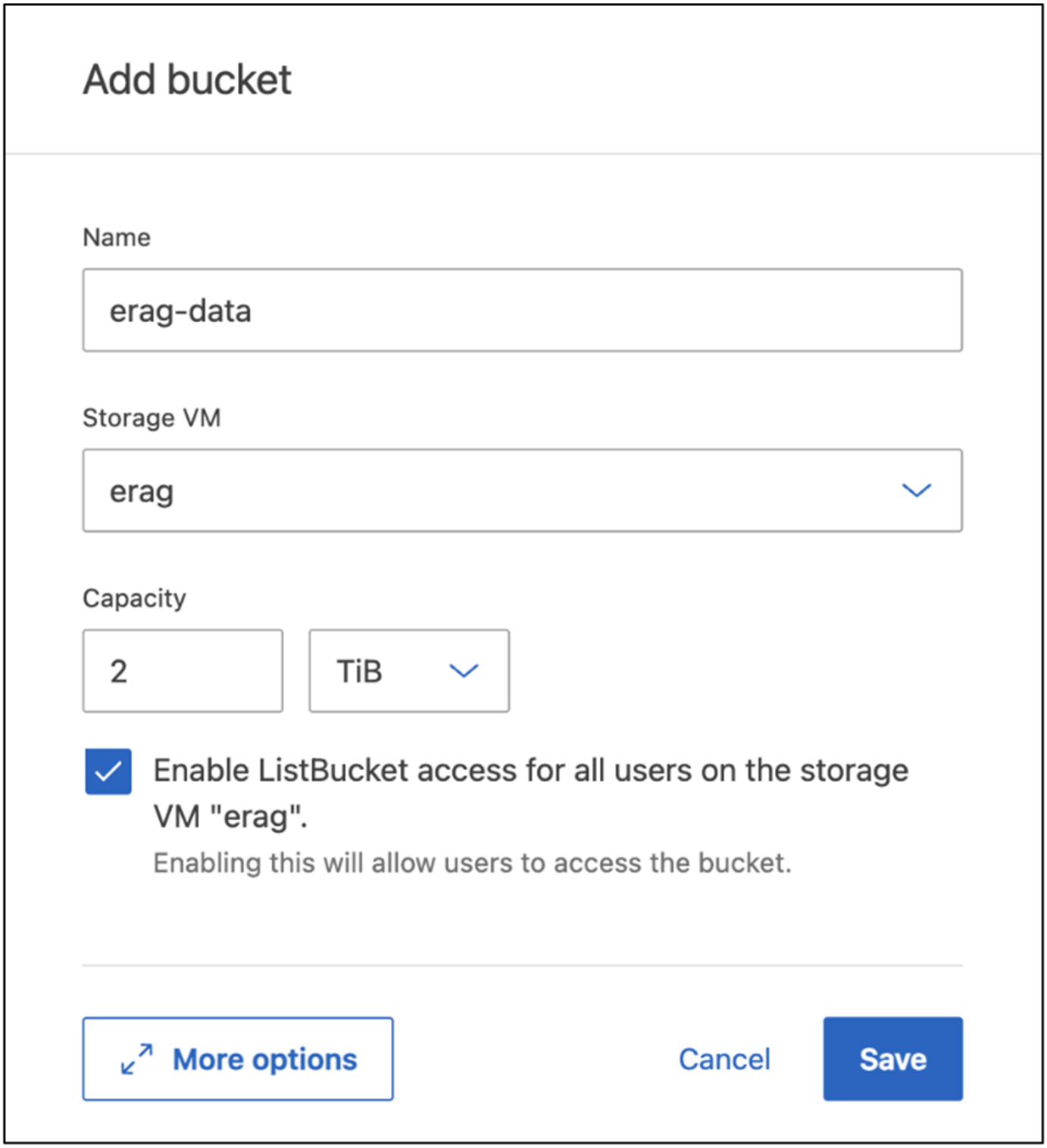


创建 **S3** 存储桶

在您之前创建的 SVM 内创建一个 S3 存储桶。要使用ONTAP系统管理器创建 SVM，请导航到“存储”>“存储桶”，然后单击“+ 添加”按钮。有关更多详细信息，请参阅ONTAP S3 文档。

以下屏幕截图展示了使用ONTAP系统管理器创建 S3 存储桶的过程。

图 6 - 创建 S3 存储
桶。



配置 **S3** 存储桶权限

为您在上一步中创建的 S3 存储桶配置权限。确保您在上一步中配置的用户具有以下权限： `GetObject`, `PutObject`, `DeleteObject`, `ListBucket`, `GetBucketAcl`, `GetObjectAcl`, `ListBucketMultipartUploads`, `ListMultipartUploadParts`, `GetObjectTagging`, `PutObjectTagging`, `DeleteObjectTagging`, `GetBucketLocation`, `GetBucketVersioning`, `PutBucketVersioning`, `ListBucketVersions`, `GetBucketPolicy`, `PutBucketPolicy`, `DeleteBucketPolicy`, `PutLifecycleConfiguration`, `GetLifecycleConfiguration`, `GetBucketCORS`, `PutBucketCORS`.

要使用ONTAP系统管理器编辑 S3 存储桶权限，请导航到“存储”>“存储桶”，单击存储桶的名称，单击“权限”，然后单击“编辑”。请参阅 "ONTAP S3 文档"了解更多详细信息。

以下屏幕截图展示了ONTAP系统管理器中必要的存储桶权限。

图 7 - S3 存储桶权限。

创建 **bucket** 跨域资源共享规则

使用ONTAP CLI，为您在上一步中创建的存储桶创建存储桶跨域资源共享 (CORS) 规则：

```
ontap::> bucket cors-rule create -vserver erag -bucket erag-data -allowed
-origins *erag.com -allowed-methods GET,HEAD,PUT,DELETE,POST -allowed
-headers *
```

此规则允许英特尔 AI for Enterprise RAG Web 应用程序的 OPEA 从 Web 浏览器内与存储桶进行交互。

部署服务器

部署您的服务器并在每台服务器上安装 Ubuntu 22.04 LTS。安装 Ubuntu 后，在每台服务器上安装 NFS 实用程序。要安装 NFS 实用程序，请运行以下命令：

```
apt-get update && apt-get install nfs-common
```

安装 **Kubernetes**

使用 Kubespray 在您的服务器上安装 Kubernetes。请参阅 "Kubespray 文档"了解详情。

安装**Trident CSI** 驱动程序

在您的 Kubernetes 集群中安装NetApp Trident CSI 驱动程序。请参阅 "Trident安装文档"了解详情。

创建**Trident**后端

为您之前创建的 SVM 创建Trident后端。创建后端时，使用 `ontap-nas`司机。请参阅 "Trident后端文档"了解详情。

创建存储类

创建与您在上一步中创建的Trident后端相对应的 Kubernetes 存储类。有关详细信息，请参阅Trident存储类文档。

面向企业 **RAG** 的英特尔 **AI** 的 **OPEA**

在您的 Kubernetes 集群中安装适用于 Intel AI for Enterprise RAG 的 OPEA。请参阅 "英特尔 AI 企业版 RAG 部署"文档以了解详细信息。请务必记下本文后面描述的所需的配置文件修改。您必须在执行安装手册之前进行这些修改，以便 Intel AI for Enterprise RAG 应用程序能够与您的ONTAP存储系统正确配合使用。

启用ONTAP S3

为 Intel AI for Enterprise RAG 安装 OPEA 时，编辑主配置文件以允许使用ONTAP S3 作为源数据存储库。

要启用ONTAP S3，请在 `edp`部分。

注意：默认情况下，Intel AI for Enterprise RAG 应用程序会从 SVM 中所有现有存储桶中提取数据。如果您的 SVM 中有多个存储桶，则可以修改 `bucketNameRegexFilter`字段，以便仅从某些存储桶中提取数据。

```
edp:
  enabled: true
  namespace: edp
  dpGuard:
    enabled: false
  storageType: s3compatible
  s3compatible:
    region: "us-east-1"
    accessKeyId: "<your_access_key>"
    secretAccessKey: "<your_secret_key>"
    internalUrl: "https://<your_ONTAP_S3_interface>"
    externalUrl: "https://<your_ONTAP_S3_interface>"
    bucketNameRegexFilter: ".*"
```

配置计划同步设置

安装英特尔企业人工智能 RAG 应用程序的 OPEA 时，启用 `scheduledSync`以便应用程序自动从您的 S3 存储桶中提取新的或更新的文件。

什么时候 `scheduledSync`启用后，应用程序会自动检查源 S3 存储桶中是否有新文件或更新的文件。在此同步过程中发现的任何新文件或更新文件都会自动提取并添加到 RAG 知识库中。应用程序根据预设的时间间隔检查您的源存储桶。默认时间间隔为 60 秒，这意味着应用程序每 60 秒检查一次更改。您可能希望更改此间隔以满足您的特定需求。

启用 `scheduledSync`并设置同步间隔，在 `deployment/components/edp/values.yaml`:

```
celery:
  config:
    scheduledSync:
      enabled: true
      syncPeriodSeconds: "60"
```

更改卷访问模式

在 `deployment/components/gmc/microservices-connector/helm/values.yaml`，对于每个卷 `pvc`列表，更改 `accessMode`到 `ReadWriteMany`。

```
pvc:
  modelLlm:
    name: model-volume-llm
    accessMode: ReadWriteMany
    storage: 100Gi
  modelEmbedding:
    name: model-volume-embedding
    accessMode: ReadWriteMany
    storage: 20Gi
  modelReranker:
    name: model-volume-reranker
    accessMode: ReadWriteMany
    storage: 10Gi
  vectorStore:
    name: vector-store-data
    accessMode: ReadWriteMany
    storage: 20Gi
```

（可选）禁用 SSL 证书验证

如果您在为 SVM 启用 S3 访问时使用了自签名证书，则必须禁用 SSL 证书验证。如果您使用由公众信任的 CA 签名的证书，则可以跳过此步骤。

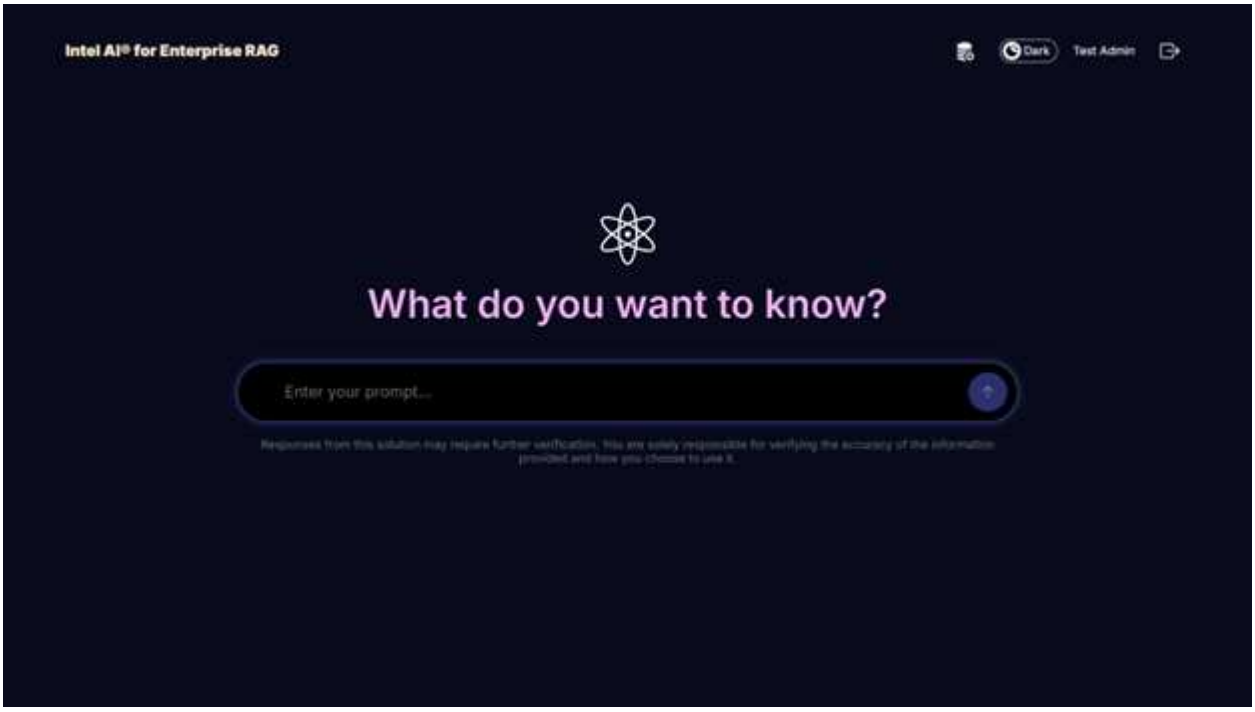
要禁用 SSL 证书验证，请在 `deployment/components/edp/values.yaml`：

```
edpExternalUrl: "https://s3.erag.com"
edpExternalSecure: "true"
edpExternalCertVerify: "false"
edpInternalUrl: "edp-minio:9000"
edpInternalSecure: "true"
edpInternalCertVerify: "false"
```

访问适用于企业 **RAG UI** 的英特尔 **AI** 的 **OPEA**

访问英特尔企业人工智能 RAG UI 的 OPEA。请参阅 "英特尔企业人工智能 RAG 部署文档"了解详情。

图 8 - 适用于企业 RAG UI 的英特尔 AI 的 OPEA



。

为 **RAG** 提取数据

您现在可以提取文件以包含在基于 RAG 的查询扩充中。有多种提取文件的选项。根据您的需要选择适当的选项。

注意：提取文件后，英特尔 AI for Enterprise RAG 应用程序的 OPEA 会自动检查文件的更新并相应地提取更新。

*选项 1：直接上传到您的 S3 存储桶 要一次提取多个文件，我们建议使用您选择的 S3 客户端将文件上传到您的 S3 存储桶（您之前创建的存储桶）。流行的 S3 客户端包括 AWS CLI、Amazon SDK for Python（Boto3）、s3cmd、S3 浏览器、Cyberduck 和 Commander One。如果文件属于受支持的类型，则您上传到 S3 存储桶的任何文件都将由英特尔 AI for Enterprise RAG 应用程序的 OPEA 自动提取。

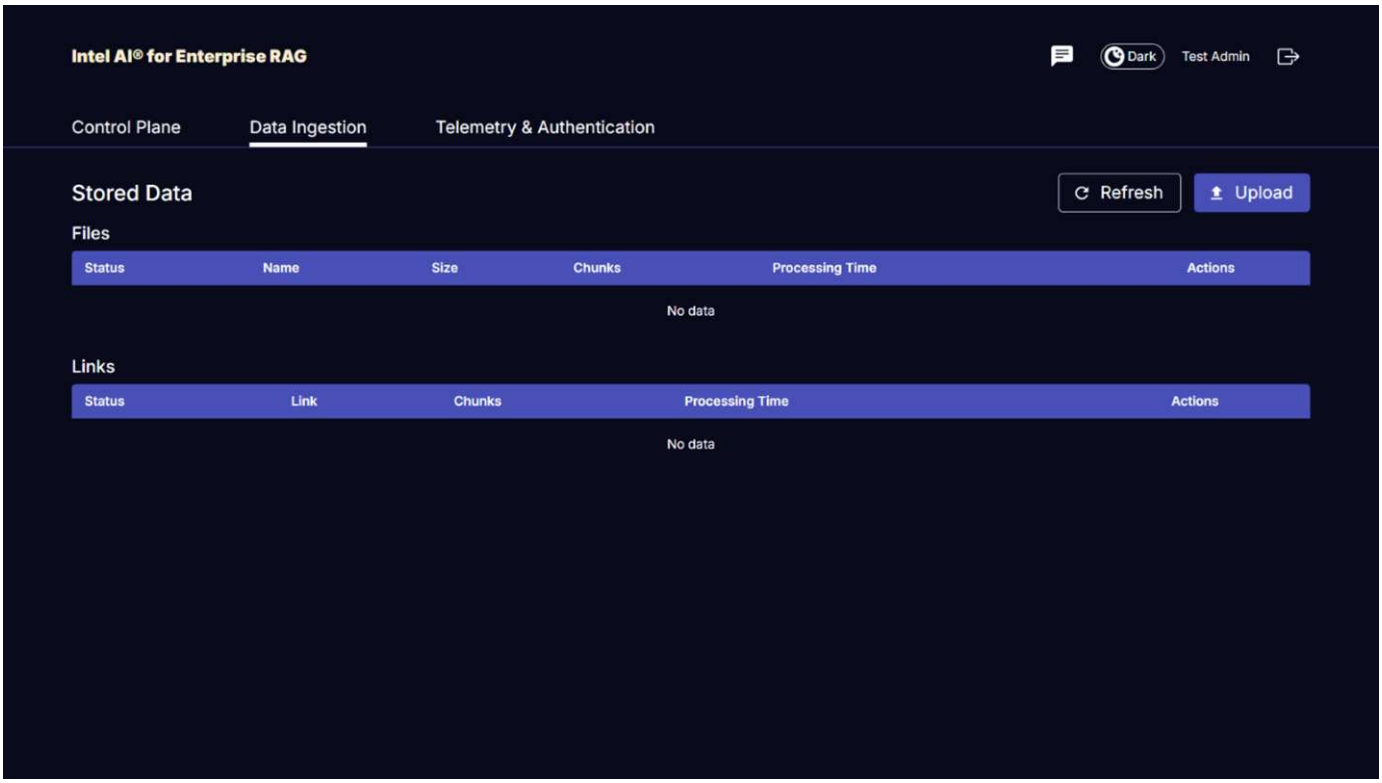
注意：在撰写本文时，支持以下文件类型：PDF、HTML、TXT、DOC、DOCX、PPT、PPTX、MD、XML、JSON、JSONL、YAML、XLS、XLSX、CSV、TIFF、JPG、JPEG、PNG 和 SVG。

您可以使用 OPEA for Intel AI for Enterprise RAG UI 来确认您的文件是否已正确提取。有关详细信息，请参阅英特尔 AI for Enterprise RAG UI 文档。请注意，应用程序可能需要一些时间来提取大量文件。

*选项 2：使用 UI 上传 如果您只需要提取少量文件，则可以使用 OPEA for Intel AI for Enterprise RAG UI 来提取它们。有关详细信息，请参阅英特尔 AI for Enterprise RAG UI 文档。

图 9-数据提取 UI

。



执行聊天查询

您现在可以使用附带的聊天 UI 与英特尔企业人工智能 RAG 应用程序的 OPEA“聊天”。在响应您的查询时，应用程序会使用您提取的文件执行 RAG。这意味着应用程序会自动在您摄取的文件中搜索相关信息，并在响应您的查询时合并这些信息。

尺寸指南

作为验证工作的一部分，我们与英特尔合作进行了性能测试。此次测试得出了下表中列出的尺寸指导。

特征	值	注释
模型尺寸	200亿个参数	Llama-8B、Llama-13B、Mistral 7B、Qwen 14B、DeepSeek Distill 8B
输入尺寸	约2000个代币	约4页
输出尺寸	约2000个代币	约4页
并发用户	32	“并发用户”是指同时提交查询的提示请求。

_注意：上面提供的尺寸指导基于使用 96 核 Intel Xeon 6 处理器收集的性能验证和测试结果。对于具有类似 I/O 令牌和模型大小要求的客户，我们建议使用具有 96 或 128 个内核的 Xeon 6 处理器的服务器。

结束语

企业 RAG 系统和 LLM 是协同工作的技术，可帮助组织提供准确且情境感知的响应。这些回应涉及基于大量私人和内部企业数据的信息检索。通过使用 RAG、API、向量嵌入和高性能存储系统来查询包含公司数据的文档存

储库，可以更快、更安全地处理数据。 NetApp AIPod Mini 将 NetApp 的智能数据基础架构与ONTAP数据管理功能以及 Intel Xeon 6 处理器、Intel AI for Enterprise RAG 和 OPEA 软件堆栈相结合，帮助部署高性能 RAG 应用程序并让组织走上 AI 领导之路。

致谢

本文档由NetApp解决方案工程团队成员 Sathish Thyagarajan 和 Michael Ogelsby 编写。作者还要感谢英特尔企业 AI 产品团队（Ajay Mungara、Mikolaj Zyczynski、Igor Konopko、Ramakrishna Karamsetty、Michal Prostko、Shreejan Mistry 和 Ned Fiori）以及NetApp的其他团队成员（Lawrence Bunka、Bobby Oommen 和 Jeff Liborio）在验证此解决方案期间给予的持续支持和帮助。

物料清单

以下是用于该解决方案功能验证的BOM，可供参考。可以使用符合以下配置的任何服务器或网络组件（甚至是最好具有 100GbE 带宽的现有网络）。

对于应用服务器：

零件编号	产品描述	数量
222HA-TN-OTO-37	超级服务器 SYS-222HA-TN /2U	2
P4X-GNR6980P-SRPL2-UCC	英特尔至强 6980P 2P 128C 2G 504M 500W SGX512	4
RAM	MEM-DR564MC-ER64(x16)64GB DDR5-6400 2RX4 (16Gb) ECC RDIMM	32
	HDS-M2N4-960G0-E1-TXD-NON-080(x2) SSD M.2 NVMe PCIe4 960GB 1DWPD TLC D，80 毫米	2
	WS-1K63A-1R(x2)1U 692W/1600W 冗余单输出电源。散热量为 2361 BTU/Hr，最高温度为 59 C（约）	4

对于控制服务器：

零件编号	产品描述	数量
511R-M-OTO-17	优化了 1U X13SCH-SYS、CSE-813MF2TS-R0RCNBP、PWS-602A-1R	1
	RPL-E 6369P IP 8C/16T 3.3G 24MB 95W 1700 BO	1
RAM	MEM-DR516MB-EU48(x2)16GB DDR5-4800 1Rx8 (16Gb) ECC UDIMM	1
	HDS-M2N4-960G0-E1-TXD-NON-080(x2) SSD M.2 NVMe PCIe4 960GB 1DWPD TLC D，80 毫米	2

对于网络交换机：

零件编号	产品描述	数量
DCS-7280CR3A	Arista 7280R3A 28x100 GbE	1

NetApp AFF存储：

零件编号	产品描述	数量
AFF-A20A-100-C	AFF A20 HA 系统，-C	1
X800-42U-R6-C	跳线 Crd，驾驶室内，C13-C14，-C	2
X97602A-C	电源，1600W，钛金，-C	2
X66211B-2-N-C	电缆，100GbE，QSFP28-QSFP28，铜，2米，-C	4
X66240A-05-N-C	电缆，25GbE，SFP28-SFP28，铜，0.5米，-C	2
X5532A-N-C	导轨，4 柱，薄，圆形/方孔，小，可调节，24-32，-C	1
X4024A-2-A-C	驱动器包 2X1.92TB，NVMe4，SED，-C	6
X60130A-C	IO 模块，2PT，100GbE，-C	2
X60132A-C	IO 模块，4PT，10/25GbE，-C	2
SW-ONTAPB-FLASH-A20-C	SW、 ONTAP基础包、每 TB、闪存、A20、-C	23

在哪里可以找到更多信息

要了解有关本文档中描述的信息的更多信息，请查看以下文档和/或网站：

"NetApp产品文档"

"OPEA项目"

"OPEA Enterprise RAG 部署手册"

版权信息

版权所有 © 2025 NetApp, Inc.。保留所有权利。中国印刷。未经版权所有者事先书面许可，本文档中受版权保护的任何部分不得以任何形式或通过任何手段（图片、电子或机械方式，包括影印、录音、录像或存储在电子检索系统中）进行复制。

从受版权保护的 NetApp 资料派生的软件受以下许可和免责声明的约束：

本软件由 NetApp 按“原样”提供，不含任何明示或暗示担保，包括但不限于适销性以及针对特定用途的适用性的隐含担保，特此声明不承担任何责任。在任何情况下，对于因使用本软件而以任何方式造成的任何直接性、间接性、偶然性、特殊性、惩罚性或后果性损失（包括但不限于购买替代商品或服务；使用、数据或利润方面的损失；或者业务中断），无论原因如何以及基于何种责任理论，无论出于合同、严格责任或侵权行为（包括疏忽或其他行为），NetApp 均不承担责任，即使已被告知存在上述损失的可能性。

NetApp 保留在不另行通知的情况下随时对本文档所述的任何产品进行更改的权利。除非 NetApp 以书面形式明确同意，否则 NetApp 不承担因使用本文档所述产品而产生的任何责任或义务。使用或购买本产品不表示获得 NetApp 的任何专利权、商标权或任何其他知识产权许可。

本手册中描述的产品可能受一项或多项美国专利、外国专利或正在申请的专利的保护。

有限权利说明：政府使用、复制或公开本文档受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中“技术数据权利 — 非商用”条款第 (b)(3) 条规定的限制条件的约束。

本文档中所含数据与商业产品和/或商业服务（定义见 FAR 2.101）相关，属于 NetApp, Inc. 的专有信息。根据本协议提供的所有 NetApp 技术数据和计算机软件具有商业性质，并完全由私人出资开发。 美国政府对这些数据的使用权具有非排他性、全球性、受限且不可撤销的许可，该许可既不可转让，也不可再许可，但仅限在与交付数据所依据的美国政府合同有关且受合同支持的情况下使用。除本文档规定的情形外，未经 NetApp, Inc. 事先书面批准，不得使用、披露、复制、修改、操作或显示这些数据。美国政府对国防部的授权仅限于 DFARS 的第 252.227-7015(b)（2014 年 2 月）条款中明确的权利。

商标信息

NetApp、NetApp 标识和 http://www.netapp.com/TM 上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。