

AI 助力企业成就不凡： 开启自主发展的新纪元



目录

AI 助力企业把握行业转型机遇 1

AI 在各个行业中的应用 4

IT 决策者必须考虑的事项 5

入门：AI 技术剖析 5

关键选择 6

 性能..... 6

 数据安全..... 6

扩展您的解决方案 7

 兼顾成本与创新..... 7

 简单易用，灵活多用..... 7

 确保可解释性 7

现实场景 8

零售 8

医疗..... 9

我们的解决方案..... 10

 让 AI 人人可用：戴尔和 AMD 致力于普及 AI 10

 与 Hugging Face 合作..... 11

 AMD EPYC™ 处理器 11

 AMD Instinct™ MI300X 加速器..... 11

 AMD ROCm™ 6 开源软件平台 12

 戴尔 PowerEdge™ 服务器产品组合 12

总结..... 13

AI 助力企业把握行业转型机遇

AI 为企业带来了千载难逢的机会，可助力当今企业实现业务转型，为未来创新做好准备。根据 Accenture Vision Technology 2023 年报告收集的数据，98% 的全球高管认为，在未来三到五年内，AI 基础模型将在其组织战略中发挥重要作用。¹

AI 有助于更高效地完成任务、推动创新并改进决策流程，因此对零售、医疗和金融服务等领域的企业非常有用。尽管 AI 具有众多优势，但由于存在一些常见误区，企业在集成 AI 方面仍面临难以跨越的门槛。



您需要拥有一支 AI 开发人员团队，才能顺利采用这项技术：

虽然在开发高级 AI 解决方案和了解各项基本原则时，具备数据科学方面的专业知识仍很重要，但这不再是一项必备条件。易于用户使用的 AI 工具、Hugging Face 等平台 and 任务特定的模型层出不穷，这些工具显著降低了开发 AI 解决方案的复杂性。

您需要花费数千万美元投资硬件，才能取得成效：

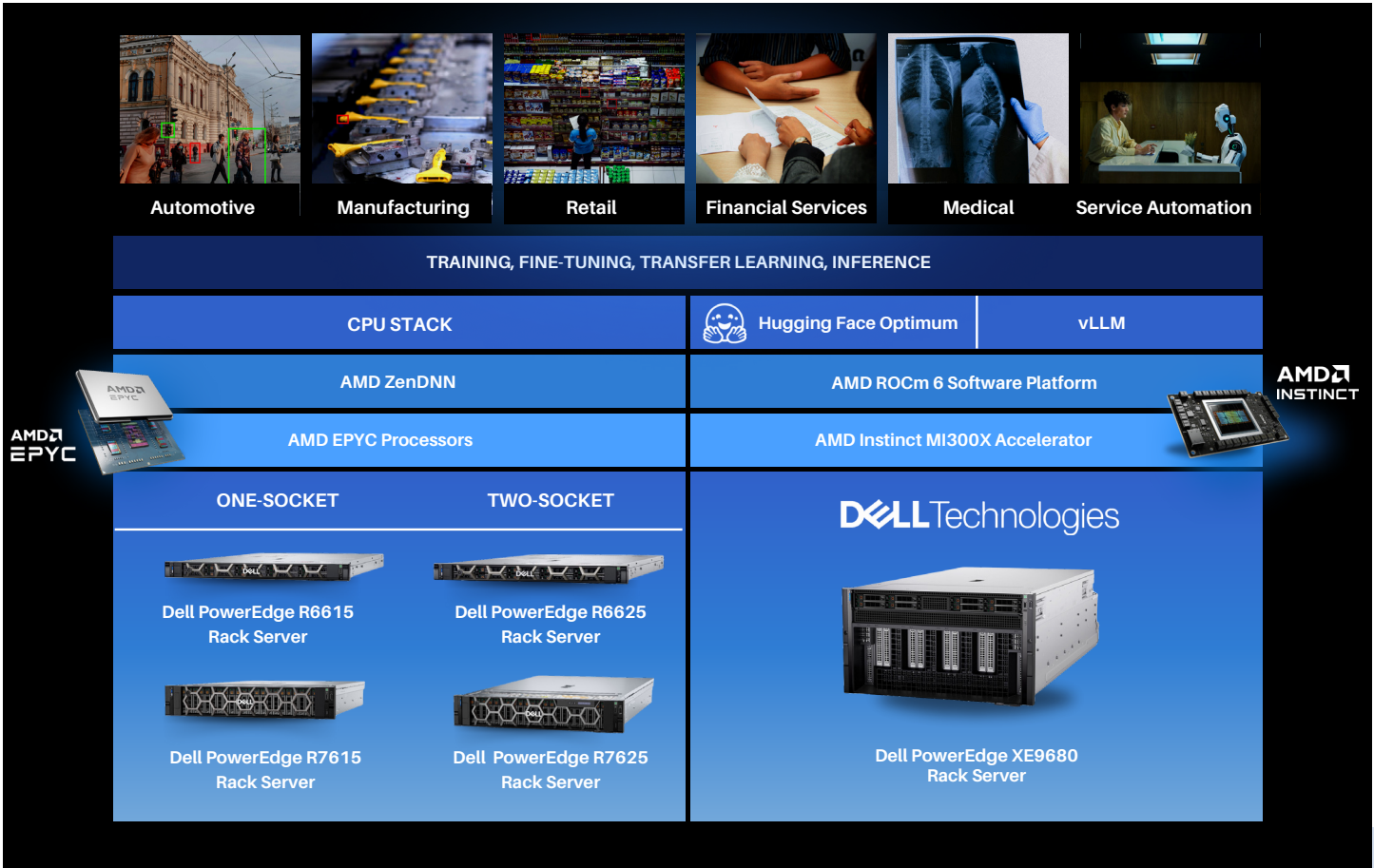
这种误区会极大地限制企业利用现今多样化的 AI 资源。虽然这些常见资源往往功能强大并具备充分支持，但它们不一定适合每个企业，也不一定是所有企业经济实惠的选择。

您需要不懈努力才能获取加速器：

虽然加速器在处理繁重的 AI 工作负载方面表现出色，但企业的 AI 应用程序可能不需要这么强大的计算能力。为了获得性能卓越的加速器而长时间观望，显然也不切实际。在许多情况下，AI 优化型 CPU 其实就可以提供实时生成 AI 辅助分析和决策所需的性能和效率，这是一种更加经济实惠且适应性更强的解决方案。

¹ Accenture, 2023 年 3 月 30 日, 《Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds》, <https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds>

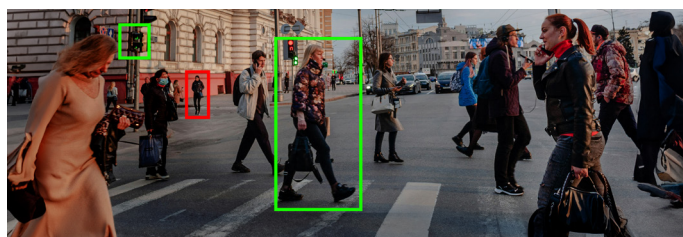
戴尔与 AMD 携手合作，提供统一的硬件和软件生态系统，来支持开发人员打造端到端 AI 解决方案，以便轻松高效地进行迁移学习、微调 and 推理。在 **Hugging Face** 的支持下，我们现在拥有越来越多的模型产品组合，这些模型在搭载 AMD EPYC™ 处理器或 AMD Instinct™ MI300X 加速器的戴尔 PowerEdge 服务器上运行，以便开发人员可以进行微调、应用迁移学习并部署推理。通过对 AMD ROCm™ 和 AMD ZenDNN™ 进行投资并采用 PyTorch、Tensorflow 和 ONNX Runtime 框架，我们可以从根本上帮助应用 AI 开发人员体验 AI 的普及。下面的堆栈图详细展示了构成戴尔和 AMD 统一 AI 生态系统的各个部分。



AI 在各个行业中的应用

随着资源的多样化以及企业对开源创新的重视，AI 正在渗透到许多不同的行业，包括客户服务、金融和银行业、医疗和零售业等。而在这些行业中，AI 均可帮助组织发掘其专有数据的潜力，并重塑 AI 工作流，为实现此目标，它提供了以下几项关键功能：数据分析、自动化、个性化设置和预测性分析。此外，AMD ROCm 和 ZenDNN 库还可加快这些 AI 工作流，从而近乎实时地取得成果。

下面深入了解一下 AI 对各个行业的具体影响。



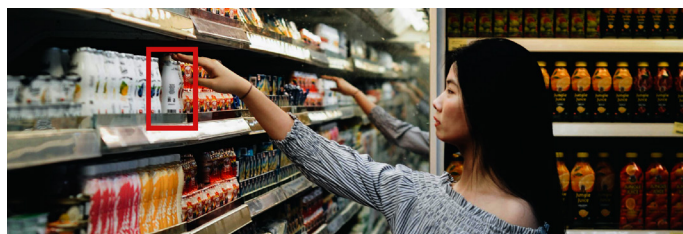
汽车

AI 用于在自动驾驶汽车中进行物体检测、车道跟踪和决策。AI 还可以预测汽车组件何时可能发生故障，从而进行主动维护并缩短停机时间。



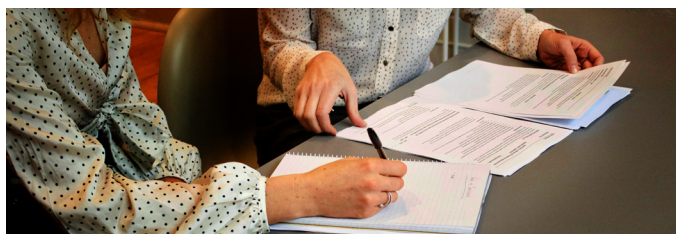
制造业和工业

AI 可在制造业和工业中用于进行预测性维护、质量控制、流程优化和供应链管理，从而提高效率并缩短停机时间。



零售

AI 可以分析客户行为，以提供个性化的产品建议，从而改进客户互动并增加销售额。它还可以预测需求并尽可能减少积压或缺货情况，从而优化库存水平。



金融服务

在金融和银行业中，AI 可用于进行欺诈检测、风险评估和投资分析以及提供客户服务，从而提高安全性并帮助做出更明智的决策。



医疗

在医疗行业中，AI 可用于各种应用，包括医学影像分析、疾病诊断、个性化的治疗计划和药物研发，从而改善患者治疗效果并降低成本。



服务自动化行业

采用 AI 技术的聊天机器人可以处理客户问询并提供支持，从而减少人为干预的需求。AI 还可以自动执行重复性任务，例如数据输入或文档处理，从而提高效率并减少错误。

IT 决策者必须考虑的事项

入门：AI 技术剖析

在探索这些应用场景之前，我们先来深入了解一下 AI 生命周期。AI（人工智能）生命周期是指开发、部署和维护 AI 系统所涉及的各个阶段。虽然具体的方法和术语可能有所不同，但典型的 AI 生命周期始终包括模型训练和推理。

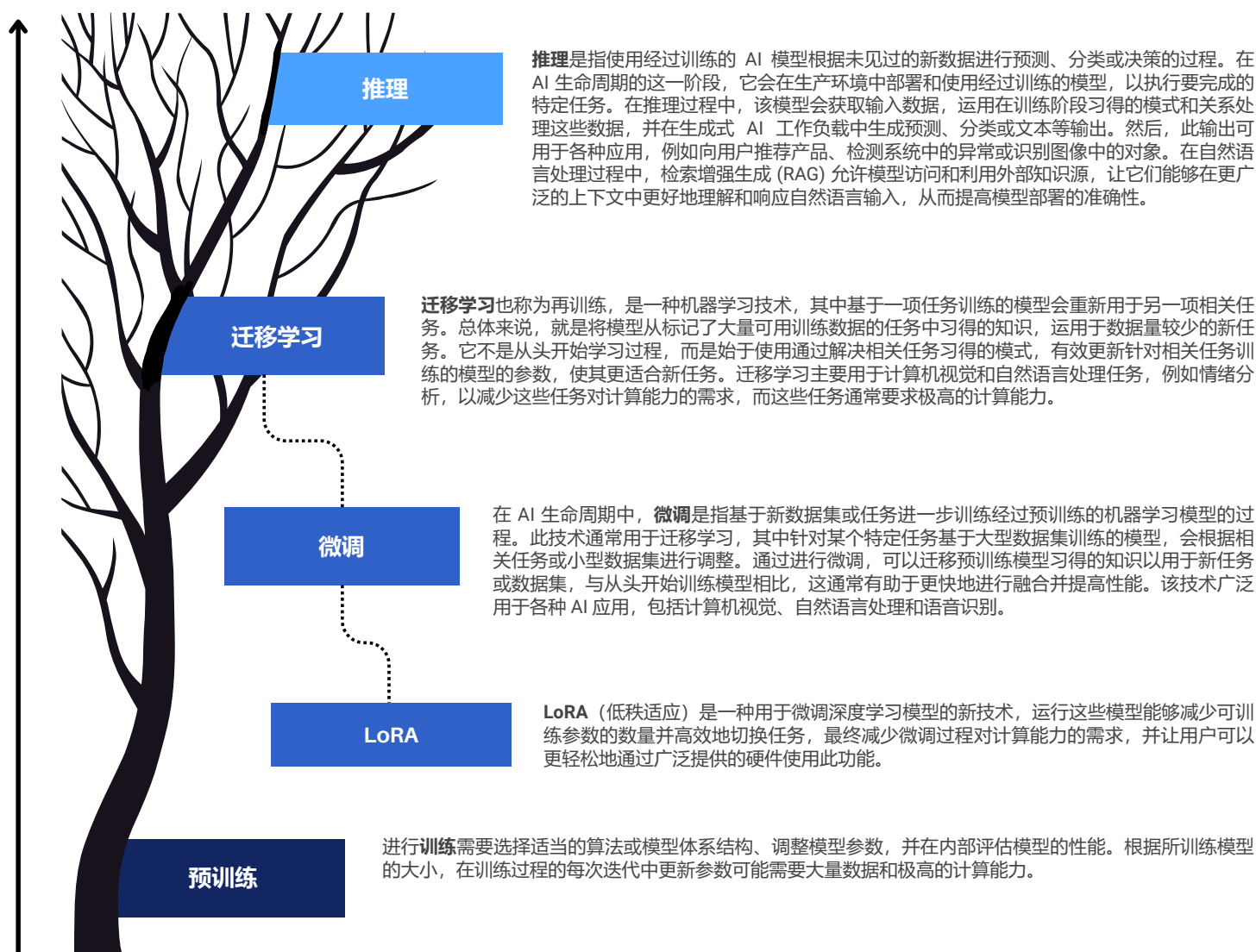


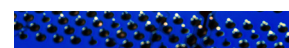
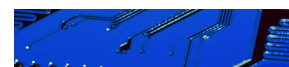
图 1：AI 生命周期



关键选择

| 性能

在很多实际应用中，实时或近乎实时地做出决策是成功与否的关键所在。例如，必须及时识别金融交易或保险索赔中的欺诈活动，以防发生财务损失并保护企业资产。在制造业场景中，必须动态监控装配线或工厂环境中存在的缺陷，以保证质量。实际上，处理推理工作负载的处理器必须经过优化，以便快速高效地处理传入数据流。戴尔 PowerEdge 服务器与 AMD EPYC 处理器搭配使用可实现多种功能，非常适合用于处理边缘推理工作负载，以及涉及高性能计算、云计算和大数据分析的任务。



| 数据安全

数据安全是 AI 系统成功与否的关键所在，对于利用生成式 AI 的系统来说尤为如此。这是力求将 AI 融入其业务运营的优秀技术型企业重点关注的一个问题。AI 系统通常依赖海量数据，其中可能包括个人详细信息、财务数据或专有信息等敏感和机密信息。保护这些数据至关重要，有助于防止未经授权的访问或数据失窃，以及确保 AI 模型和预测的精确性、可靠性和一致性。

机密计算是一种有助于在安全飞地中处理数据的技术，可保护数据免遭未经授权方（包括云提供商和其他用户）未经授权的访问或操作。² 在数据处理过程中，它使用加密和其他安全措施来隔离数据。AMD Infinity Guard 提供一系列集成到 AMD EPYC 处理器中的先进安全功能，并采用安全加密虚拟化 (SEV) 来支持机密计算。SEV 使用仅处理器知道的密钥对虚拟机 (VM) 进行加密。这些服务旨在使用 AMD SEV 安全嵌套分页 (SEV-SNP) 提供基于硬件的可信执行环境，从而增强来宾保护以帮助抵御外部威胁。

联合学习是维护数据安全的另一种方法。该方法用于在分散的设备或服务器中训练中央模型。³ 每台设备均在本地训练模型，而不是将所有数据传输到一个中央位置，并且仅共享模型更新。这种方法可保护隐私并实现协作式学习，而无需共享原始数据。借助 Dell Technologies 的联合 AI 平台，收集数据时可在网络边缘针对数据集运行计算流程、AI 和 ML 算法，并通过网络与其他边缘设备、数据中心或云仅共享数学模型、元数据和查询结果。以这种方式进行数据交换可近乎实时地从大型分布式数据集中提取切实可行的见解，而不会泄露数据和任何知识产权，从而取得更好的成果。

² Advanced Micro Devices, Inc., 2023 年 8 月 30 日, 《AMD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings》, <https://www.AMD.com/en/newsroom/press-releases/2023-8-30-AMD-shares-the-technical-details-of-technology-pow.html>
Advanced Micro Devices, Inc., 2021 年, 《Data Center Solutions, Confidential Computing》解决方案简介, <https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf>

³ Analytics Vidhya, 2023 年 12 月, 《Federated Learning: A Beginner's Guide》, <https://www.analyticsvidhya.com/blog/2021/05/federated-learning-a-beginners-guide/#:~:text=Federated%20learning%20works%20by%20training,learning%20without%20sharing%20raw%20data>
Dell Technologies, 2021 年, 《A federated learning platform for real-time artificial intelligence》解决方案简介, <https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf>

扩展您的解决方案

I 兼顾成本与创新

适当兼顾成本和创新，确保 AI 解决方案不仅在经济上可行，还可发挥影响力，从而切实为企业和用户创造价值。要做到两者兼顾，一个关键因素在于找到一款符合您应用场景的硬件，并且该硬件可轻松集成到现有基础架构中。在现代 AI 硬件市场中，各个行业对加速器的需求不断增长。此外，企业还面临产能限制、物流挑战和半导体短缺问题。这些因素都造成了加速器短缺。

但是，CPU 已成为大多数数据中心的标准组件，因此与添加全新的加速器硬件相比，集成起来更简单、经济。AI 优化型 CPU 可以利用现有的软件和工具，这就减少了对广泛重新配置工具和再训练的需求。CPU 还有助于更灵活高效地执行 AI 领域之外的各种任务，以便更广泛地使用数据中心内的资源。使用运行 AMD EPYC 处理器的戴尔 PowerEdge 服务器更新数据中心，有助于满足您的现有工作负载需求，同时在 AI 的支持下随时准备开拓创新并提高效率。

I 简单易用，灵活多用

AI 系统应简单、灵活，从长远来看，这对于构建有效、适应性强且可扩展的 AI 解决方案至关重要。通过获取一系列软件框架和优化作为硬件的补充，可以提高性能，而无需花费额外的时间和精力来进行跨平台集成。这些特质对于处理混合 AI 工作负载尤为重要，此类工作负载整合了不同类型的 AI 任务，例如训练、推理和数据处理任务。

AMD 和 Dell Technologies 通过结合使用硬件和软件解决方案来处理混合 AI 工作负载。AMD EPYC 高核心数处理器提供高性能计算能力，以及同步多线程 (SMT) 等功能，支持高效地并行处理 AI 工作负载。这些处理器针对 AI 任务进行了优化，可提供强劲的性能来处理训练和推理工作负载。戴尔 PowerEdge 服务器搭载 AMD EPYC 处理器，提供可扩展的灵活平台来部署 AI 工作负载。此外，Dell OpenManage 软件套件还提供多种管理工具，用于优化混合 AI 工作负载的资源分配和性能监控。

AMD 还推出了 Unified Inference Frontend (UIF)，该产品可利用当前每个软件堆栈的性能增强版本，并依托 AMD EPYC 处理器的 AMD ZenDNN 库、用于 AMD Instinct 加速器的开源 AMD ROCm 堆栈，以及用于 AMD 自适应 SoC 的软件堆栈。此外，AMD ROCm 也经过专门设计，可使用各种 AMD CPU 和加速器（包括专业产品和消费级产品）。

I 确保可解释性

可解释的 AI 在确保人工智能应用程序的透明度、可信度和有效性方面发挥着关键作用。可解释的 AI 揭示了底层因素和推理过程，阐释了 AI 模型如何做出决策。这种透明度对于获取利益相关者的信任至关重要，尤其是在医疗、金融和刑事司法等敏感领域，在这些领域所做的决策会直接影响个人的生活。

人机回圈 AI 系统利用人类智慧来提升 AI 性能并减少算法偏差。通过纳入人工监督，这些系统可以更有效地处理模棱两可的复杂情况，确保 AI 解决方案符合道德和社会规范。此外，人工参与有助于基于真实反馈持续优化和调整 AI 模型，从而促进迭代改进并实现长期可靠性。这些方法对于打造负责任的包容性 AI 系统至关重要，从而造福社会。

现实场景

Scalers AI 与戴尔和 AMD 携手合作，展示了搭载 AMD 处理器的戴尔 PowerEdge 服务器所具有的功能。了解如何在零售和医疗场景中利用这些技术进行训练、迁移学习和推理。

零售

Scalers AI 打造了零售库存管理参考解决方案，该系统可通过实施对象检测 AI 模型来监控和管理零售货架的库存水平。此参考解决方案利用 SSD_MobileNet_V2 模型来确定和识别商店货架上的产品，最终实现自动库存盘点并精确监控库存水平。该模型使用 SKU110K 图像数据集（包括 Roboflow 提供的 23,000 张图像）进行迁移学习。利用计算机视觉和机器学习算法，系统可以检测到商品何时库存不足或缺货，从而向店员发出警报，以便及时补充库存或补货。

此解决方案使用搭载 AMD EPYC 9354P 32 核处理器的戴尔 PowerEdge R7615 服务器。

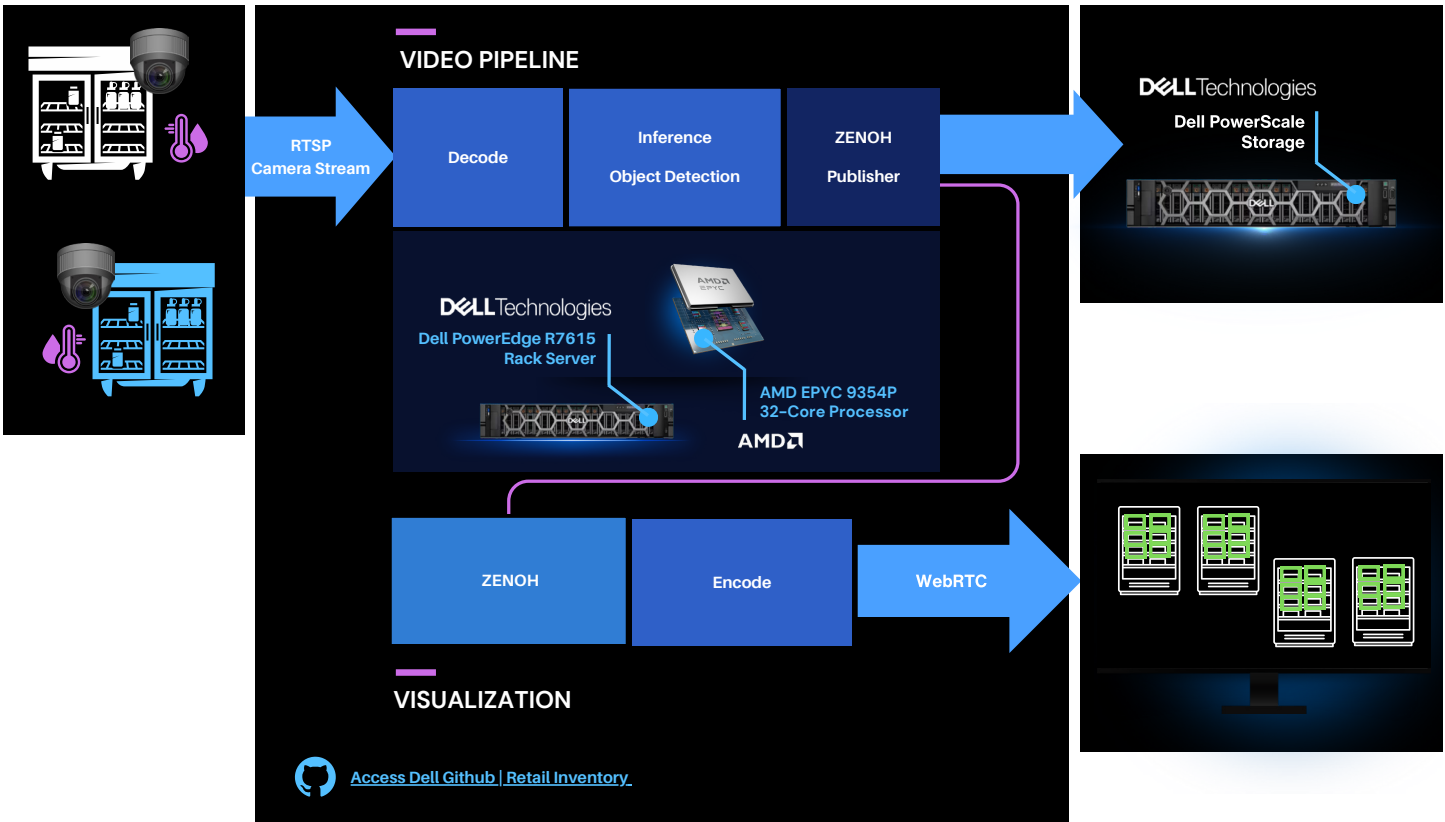


图 2：零售库存管理参考解决方案的体系结构图

医疗

AI 支持的医疗成像技术具有重要作用，它可提高诊断的准确性和效率，让医疗专业人员能够精确了解肉眼可能难以察觉的疾病，从而提升医疗水平。通过自动分析医疗影像，AI 可缩短诊断所需的时间，支持更快地做出治疗决策，并最终改善患者的治疗效果。

Scalers AI 利用配备 AMD EPYC 9554 64 核处理器的戴尔 PowerEdge R7625 服务器提供的功能，创建了用于检测肺炎的 AI 支持的医疗成像解决方案。此解决方案使用先进的算法和机器学习技术来分析医疗影像（例如 X 光片或 CT 扫描），有助于提高患者肺炎诊断的速度和准确性。最后，它引入了额外的计算机辅助检视层，有望帮助医疗专业人员更高效地处理大量成像数据。

此参考解决方案利用 ResNet50 模型来分析从 NIH 临床中心数据集中获取的 X 光胸片影像。其主要目的是检测患者是否患有肺炎，本质上就是执行二元分类。该模型使用 NIH 临床中心数据集中的 X 光片 DICOM 数据集进行训练，并使用 ResNet50 体系结构进行迁移学习。

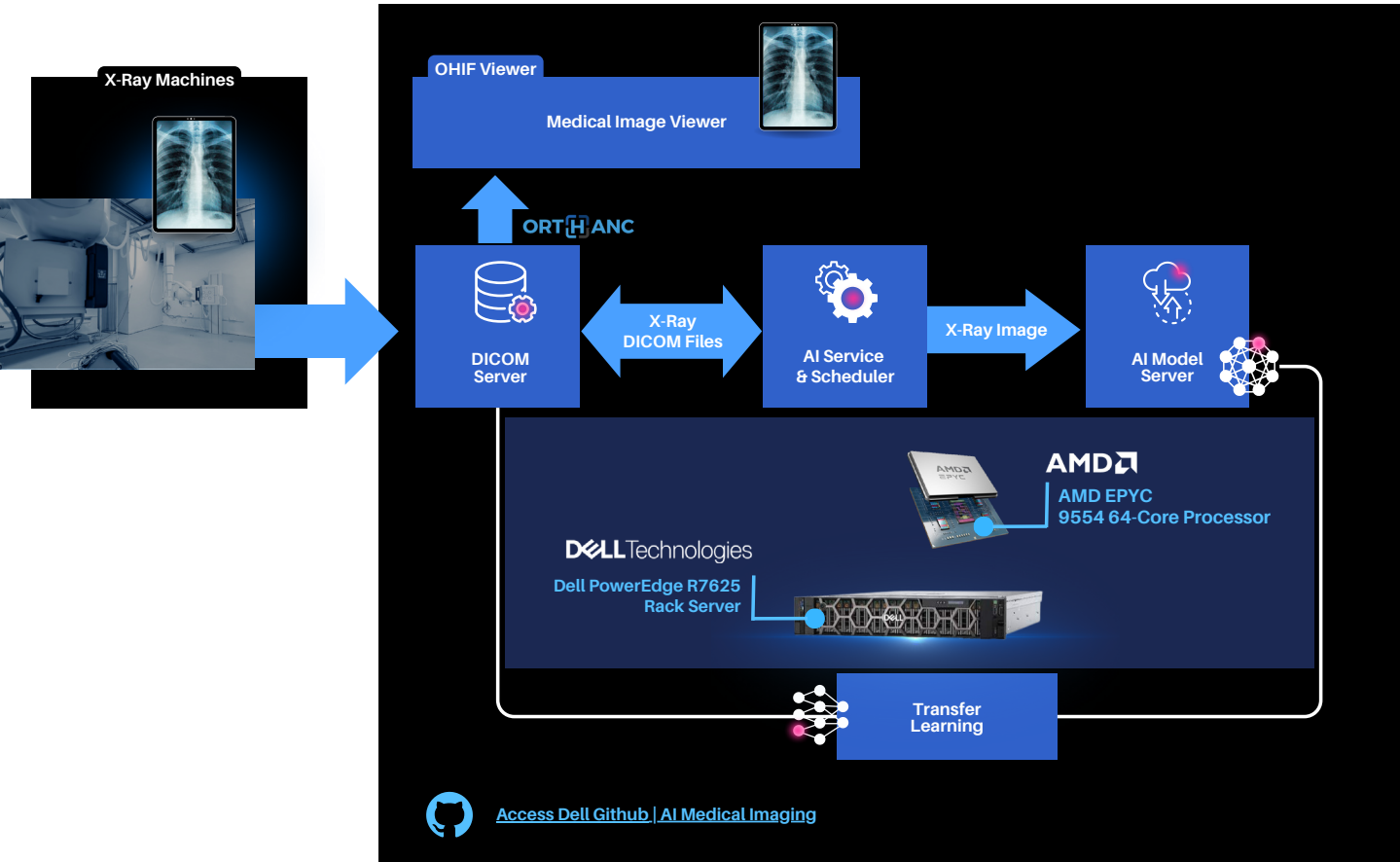
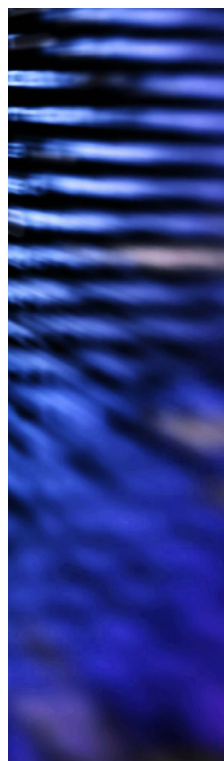
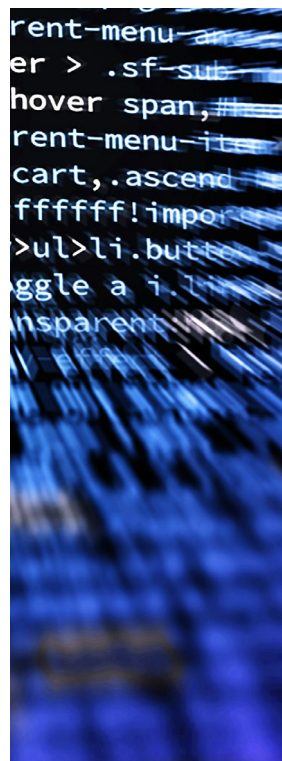


图 3：AI 医疗成像解决方案的体系结构图

我们的解决方案

让 AI 人人可用：戴尔和 AMD 致力于普及 AI

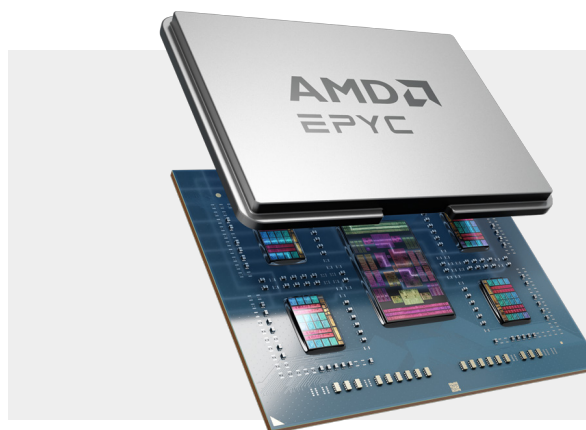
此次合作为普及 AI 奠定了基础，这对于促进创新和提高 AI 生态系统的包容性至关重要。为了取得这一成效，戴尔和 AMD 为个人和组织提供了一组强大的服务器，这些服务器配备先进的 AMD CPU 和加速器技术，可帮助他们利用 AI 并解决在各自领域中面临的独特挑战。搭载 AMD Instinct MI300X 加速器的戴尔 PowerEdge 服务器能够处理大型 AI 工作负载，例如训练和微调大型语言模型 (LLM)，而配备 AMD EPYC 处理器的戴尔 PowerEdge 服务器则擅长处理边缘推理工作负载。除了底层硬件平台之外，AMD 还提供了 ZenDNN 软件库和 AMD ROCm 软件库，前者用于针对 AMD CPU 优化深度学习推理，后者用于针对 AMD Instinct 加速器改进训练、微调和推理功能。所有这些选项都在 AMD 的统一推理模型 (UIF) 中无缝地绑定在一起，用户可以通过该模型构建端到端 AI 解决方案，并灵活地选择软件框架、软件优化和硬件平台。



与 HUGGING FACE 合作

Hugging Face 是一个致力于数据科学和机器学习的开源平台，渴望采用 AI 的企业可以直接从该平台入手，利用根据其特定需求量身定制的已有模型或 AI 工作流。AMD 已与 Hugging Face 达成了合作，他们制定了一个共同目标，力求通过向已与 AMD 平台无缝集成的软件库和框架添加 AMD 特定的软件优化功能，来提供卓越的 Transformer 性能。Hugging Face 正在与 AMD 的工程团队积极开展合作，力求优化关键模型以实现卓越性能，将 AMD ROCm 整合到 Transformer 库中，并改进专为 AMD 平台设计的 Optimum-AMD 库。这样一来，Hugging Face 用户只需进行极少的代码更改即可利用这些工具。

Dell Technologies 最近也与 Hugging Face 联手，以简化企业流程，让他们可以更轻松地使用 Hugging Face 社区开发、微调和应用自己的生成式 AI (GenAI) 开源模型，而这一切都是通过业界卓越的戴尔基础架构产品和服务实现的。我们正在 Hugging Face 平台上开发一个新的戴尔门户，其中将包括专用的自定义容器和脚本，可帮助用户使用戴尔的服务器和数据存储系统，安全轻松地部署 Hugging Face 上提供的开源模型。企业现在可充分利用 Hugging Face 资源，直接在搭载 AMD 处理器的戴尔 PowerEdge 服务器上部署模型，并使用自己的专有数据构建端到端 AI 解决方案。

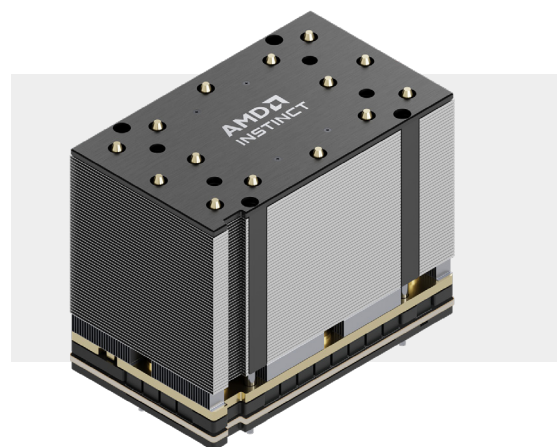


AMD EPYC 处理器

AMD 通过其 AMD EPYC 处理器，为基于云的现代数据中心带来了所需的技术进步。这些处理器是一种从头开始设计的片上系统 (SoC)，可有效满足当前和未来数据中心的需求。AMD EPYC 9000 系列处理器可为数据中心提供多达 128 个核心、256 个线程、12 个内存通道（每个插槽支持高达 6 TB 的内存）和 128 个 PCIe Gen5 通道。它同时采用业界开创性的硬件嵌入式 x86 服务器安全解决方案。通过将必要的计算、内存、I/O 和安全资源集成到 SoC 中，AMD EPYC 处理器可提供卓越的性能并有助于降低总拥有成本 (TCO)。

AMD INSTINCT MI300X 加速器

AMD Instinct MI300X 加速器基于先进的 AMD CDNA 3 体系结构构建而成，可为密集型 AI 和 HPC 应用程序提供业界卓越的效率和性能。它配备 304 个高性能计算单元，并具有 AI 特定的功能，例如支持新数据类型以及照片和视频解码。单个加速器具有 192 GB 的超大 HBM3 内存。

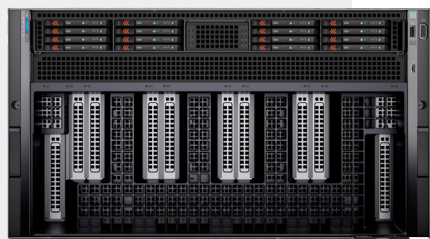


AMD ROCm 6 开源软件平台

AMD ROCm 6 开源软件平台经过优化，可通过 AMD Instinct MI300X 加速器，更大幅度地提升高性能计算 (HPC) 和 AI 工作负载的性能。它还扩展了对 AMD Instinct MI300X 加速器的支持，确保该加速器可与行业软件框架兼容。AMD ROCm 平台封装了多种驱动程序、开发工具和 API，可促进从内核级别应用程序到终端用户应用程序的加速器编程，并可根据您的特定要求量身定制。AMD ROCm 特别适用于高性能计算 (HPC)、人工智能 (AI) 和科学计算领域的应用程序。此外，AMD ROCm 平台还支持多加速器计算，包括用于服务器节点通信的远程直接内存访问 (RDMA)。

AMD
ROCm

戴尔 POWEREDGE 服务器产品组合



通过投资 AMD 产品，戴尔在市场上提供了一个有助于普及 AI 的重要选择，他们推出了四个搭载 EPYC 的服务器平台，以及搭载 AMD Instinct MI300X 加速器的旗舰戴尔 PowerEdge XE9680 机架式服务器，正是这一点的有力例证。新一代戴尔 PowerEdge 服务器配备 AMD EPYC 处理器，可支持数据库和分析、虚拟化、软件定义的存储、虚拟桌面基础架构 (VDI)、容器化、高性能计算 (HPC)、AI 和机器学习 (ML) 等变革性工作负载，从而提高业务敏捷性并有助于缩短产品上市时间。他们的单插槽 (单 CPU) 机架式服务器不仅价格经济实惠，同时兼顾性能和存储容量，可随着业务增长无缝扩展，而双插槽 (双 CPU) 机架式服务器则具备多种功能，可适应要求更严苛的工作负载。

戴尔 PowerEdge XE9680 机架式服务器是专为 AI 任务打造的强大数据处理机型。它支持八个加速器，非常适合机器学习 (ML)/深度学习 (DL) 训练和推理工作负载，特别适合用于训练大型语言模型 (LLM)。搭载 AMD Instinct MI300X 加速器的戴尔 PowerEdge XE9680 机架式服务器配备八个 MI300X 加速器，每个加速器具有 192 GB 5.3 TB/s 高带宽内存 (HBM3)，每台服务器的总 HBM3 容量为 1.5 TB，FP16 性能超过 21 petaFLOPS，可进一步将生成式 AI 的应用扩展到企业。这样，他们就可以训练更大的模型、尽可能减少数据中心占用空间、降低 TCO 并获得竞争优势。

总结

AI 加快了创新步伐，与任何其他技术转型相比，可以更快地推动数据中心工作负载革新。为了支持这些技术进步，戴尔和 AMD 正在竭力打造一个更具包容性、创新性并以合乎道德的方式开发的 AI 生态系统，鼓励各行各业的开发人员携手开发开源资源，推动当今企业在生成式 AI 领域进行创新。无论您的 AI 解决方案是通过 AMD EPYC 处理器，还是通过配备 AMD Instinct 加速器的服务器达到您的性能要求，我们都能让您灵活地在各种硬件平台上运行 AI 工作负载，尽享戴尔和 AMD 提供的众多优势。

参考资料

AMD 图像：AMD.com，AMD 合作伙伴资源库，<https://www.amd.com/en/partner/resources/resource-library.html>

戴尔图像：[Dell.com](https://www.dell.com)