

开发日语生成式 AI，实现数字 广告服务转型

CyberAgent, Inc. 采用搭载了 8 个 NVIDIA® H100 Tensor Core GPU 的戴尔 PowerEdge XE9680 服务器，加速开发生成式 AI 并提升广告效果。

业务需求

自 2016 年以来，CyberAgent, Inc. 一直积极投身 AI 的研发工作，并将 AI 运用到了其广告业务中。为了支持生成式 AI 的开发工作，公司需要配备稳定可靠且搭载先进 NVIDIA GPU 的本地服务器，让员工能够以经济实惠的方式快速使用这些资源。

业务成果



与上一代服务器相比，PowerEdge XE9680 服务器将大型语言模型 (LLM) 的性能提升了约 5.14 倍。



借助 NVIDIA Transformer 引擎优化，预计未来性能将提升 10 倍以上。



能够根据最新数据集，对机器学习模型进行快速微调。



与主流的 8U 外形规格相比，采用 6U 外形规格的服务器可节省数据中心的空间，并提升冷却效率。

解决方案概览

- 搭载 NVIDIA® H100 GPU 的戴尔 PowerEdge XE9680 服务器
- Dell ProSupport

CyberAgent, Inc. 是日本互联网广告行业的知名企业，经营着创新电视平台 ABEMA 等多项业务。该公司在 2016 年成立了一个名为 AI Lab 的 AI 研究组织，此后一直积极从事 AI 研发工作。2020 年，CyberAgent 推出了一款先进的预测式 AI，可以帮助公司更好地制作更具影响力的横幅广告标语和图像组合，从而提高广告的效果。

之后 CyberAgent 继续开发生成式 AI，创建了一个包含 130 亿个参数的日语大型语言模型 (LLM)。该 LLM 属于通用 AI 模型，适用于多种应用场景，同时还可进行微调，以便创作吸引不同广告平台用户的广告标语文案。CyberAgent 目前已将此日语 LLM 应用于旗下的 Kiwami Prediction AI、Kiwami Prediction TD 和 Kiwami Prediction LP 等 AI 服务，帮助制作创意广告并预测广告效果。未来，CyberAgent 计划开发一种能够同时处理日语 LLM 和图像的多模态 AI。

“以前使用公有云的时候，我们的内部研究人员或者无法获得 GPU 资源或者因为使用时间过长而被收取了高额费用，而现在他们可以获得更多 GPU 资源且无需担心费用。”

Daisuke Takahashi
CyberAgent, Inc. 集团 IT 部门 CIU 解决方案架构师

2023 年 5 月，CyberAgent 发布了一款名为 OpenCALM (Open CyberAgent Language Models) 的商用开源日语 LLM，其中包含多达 68 亿个参数。

虽然 ChatGPT 的聊天功能经过调优，但 OpenCALM 更像的是一款通用的日语语言模型，并且可以按用户需求进行调整。CyberAgent 之所以将 OpenCALM 作为一个开源项目，是因为比起在一个封闭的环境内开发日语 LLM，公司更希望从其他渠道获得反馈，并与其他公司合作推动日本 AI 技术的发展。

支持 CyberAgent 从事 AI 创新的基础架构

2016 年，CyberAgent 成立了 AI Lab，当时研究人员人手一台搭载有 GPU 的工作站。到了 2020 年，由于疫情迫使人们远程办公，研究人员很难再用到搭载 GPU 的工作站。为了确保研究人员获得所需的计算资源，该公司在得知新款 NVIDIA® A100 GPU 发布以后，开始考虑使用搭载 GPU 的服务器在数据中心或公有云中构建集中式机器学习 (ML) 平台。

CyberAgent, Inc. 集团 IT 部门 CIU 解决方案架构师 Daisuke Takahashi 表示：“如果只是想使用 GPU，我们可以选择使用公有云。但如果选择使用公有云，我们根本不会知道何时有一款 GPU 可供使用。另外，我们也无法确保在需要的时候有 GPU 可用。因此，我们决定在本地部署使用方便的 GPU 资源。为了灵活使用公有云和私有云作为基础架构，我们设计了尽可能接近公有云规范的用户界面。” CyberAgent 使用搭载 4 个 NVIDIA A100 GPU 的戴尔 PowerEdge XE8545 服务器构建了初始本地 ML 平台。

CyberAgent 为什么选择搭载了 NVIDIA H100 GPU 的 PowerEdge XE9680 服务器

CyberAgent 长期关注 GPU 创新，尤其关注新款 NVIDIA H100 GPU。“这款 GPU 之所以吸引我们，不光是因为它的性能更强劲，还因为它采用了可以加速特定计算算法的 Transformer 引擎等机制。” Takahashi 解释道，“根据 NVIDIA 的说法，与上一代 NVIDIA A100 GPU 相比，Transformer 引擎可以将 LLM 的 AI 训练加速高达 9 倍，将 AI 推理加速高达 30 倍。”

CyberAgent 选择了搭载 8 个 NVIDIA H100 GPU 的 PowerEdge XE9680 服务器型号。Takahashi 解释道：“当我们听说戴尔即将发布搭载 NVIDIA H100 GPU 的戴尔 PowerEdge XE9680 服务器时，就决定尽快采用这款服务器。当时，我们与 Dell Technologies 就即将推出的 PowerEdge XE9680 服务器和 GPU 的配置进行了密切沟通。我们希望用尽可能少的设备来增加正常运行时间，因此我们很高兴 Dell Technologies 能够以合理的价格为我们提供高水准的维护服务，包括四小时上门服务。”



**目前包含 130 亿个参数的 LLM 加速了 5.14 倍，
未来将加速超过 10 倍。**

Takahashi 继续说道：“我们选择 PowerEdge XE9680 服务器的另一个原因在于，我们之前安装的 PowerEdge XE8545 服务器性能稳定且易于维护。此外，我们也很看重戴尔 iDRAC 管理工具方便易用的特性，可以轻松地使用这款工具来安全地管理本地和远程服务器。”

Takahashi 先生还对交货速度表示了赞赏，他们在 2023 年 3 月下单之后短短一个多月，也就是 5 月中旬就收到了货。“在供应链因疫情而纷纷陷入混乱的情况下，Dell Technologies 相对稳定的供应链也让我更加放心。很高兴能看到他们在如此短的时间内完成交付。”

Dell Technologies 在交付后的构建过程中完成了许多创新。Takahashi 回忆道：“由于我们的 LLM 拥有大量参数，因此需要使用多个 GPU。我们在每台服务器中安装了 8 个 400 Gbps 的网卡 (NIC)，并使用 RDMA (远程直接内存访问) 技术来创建服务器之间的高速互连。GPU 服务器会产生大量热量，因此在设计时必须考虑冷却效率。PowerEdge XE9680 服务器采用 6U 外形规格，可以做到稳定散热，这一点也值得称赞。此外，我们还将数据中心搬到了新的位置，并在该位置配备了后门热交换器。通过在机架后部安装水冷后门热交换器，而不是为数据中心所在的整个房间制冷，实现了高效冷却。”

借助 Transformer 引擎优化，让广告标语更准确

安装 PowerEdge XE9680 服务器为 CyberAgent 带来了多项优势。“我们希望能够在性能显著提升的良好态势下，更快更频繁地更新日语 LLM。” Takahashi 先生说，“日语 LLM 的发展速度也会变得更快。此外，与搭载 4 个 NVIDIA A100 GPU 的 PowerEdge XE8545 服务器相比，搭载 8 个 NVIDIA H100 GPU 的 PowerEdge XE9680 服务器的性能提升了约 5.14 倍。我们

还预计，对 NVIDIA Transformer 引擎进行优化之后，未来服务器的性能将提升 10 倍以上。现在，我们还能根据最新数据集对 ML 模型进行快速微调，因此能够更轻松地响应改进服务的请求，提高广告标语的准确度并提供更有效的内容。”

采用 PowerEdge XE9680 服务器的 ML 基础架构获得了用户的高度评价。“我们从内部研究人员那里了解到，以前使用公有云的时候，他们或者无法获得 GPU 资源或者因为使用时间过长而被收取了高额费用，而现在他们可以获得更多 GPU 资源且无需担心费用。” Takahashi 先生说道，“另一项优势是我们现在能够提供高规格的基础架构，包括互连技术，因此用户可以更轻松地取得业务成果。”

Takahashi 先生还对已经使用了一段时间的 Dell Technologies 的 iDRAC 管理工具表示了赞赏，认为它的的确确简化了管理工作。“我们的工作人员并不总是驻扎在数据中心，而 iDRAC 可以让我们在无需访问操作系统的前提下，远程完成检查 GPU 温度和状态以及更新固件等工作，非常实用。”



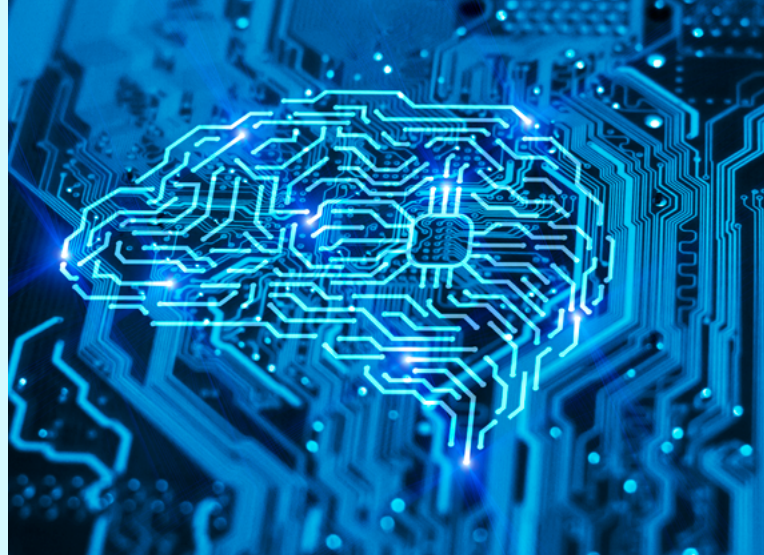
PowerEdge XE9680 服务器采用 6U 外形规格，可以做到稳定散热，这一点也值得称赞。”

Daisuke Takahashi
CyberAgent, Inc. 集团 IT 部门 CIU 解决方案架构师

“我们希望能够更快地更新日语 LLM。搭载 8 个 NVIDIA H100 GPU 的 PowerEdge XE9680 服务器的性能提升了约 5.14 倍。”

Daisuke Takahashi

CyberAgent, Inc. 集团 IT 部门 CIU 解决方案架构师



专注于 LLM、GPU 和基础架构

CyberAgent 计划在未来利用通过 OpenCALM 收集的反馈和数据来改进员工使用的 LLM。CyberAgent 还尝试通过 OpenCALM 与广告行业以外的公司和组织开展合作。例如，CyberAgent 已着手与零售和金融从业者讨论建立基于相关行业数据的行业特定 LLM。

同时，Takahashi 先生表示，他将持续关注新款 GPU 和相关新技术，了解它们的商业化应用情况。“我们也期待看到其他供应商努力创建类似于 NVIDIA 所实现的软件生态系统。我还对 NVIDIA NVLink-C2C 的实施以及连接 CPU 和 GPU 的 CXL (Compute eXpress Link) 等新标准感兴趣，因为 PCIe 总线可能成为 GPU 性能的瓶颈。我期待 Dell Technologies 能够一如既往地快速采用新技术并设计高性能产品。”

CyberAgent 的 AI 研发团队将通过采用经济实惠的新款 GPU，继续提供用户需要的 ML 基础架构，并在 AI 道路上不断取得进步。此外，随着日语 LLM 的进一步发展，CyberAgent 将不仅仅在自身的广告业务方面受人瞩目，还将成为日语 AI 市场中的一颗璀璨明星。

此内容由 Dell Technologies 翻译自日语版本。

“我们希望用尽可能少的设备来增加正常运行时间，因此我们很高兴 Dell Technologies 能够以合理的价格为我们提供高水准的维护服务，包括四小时上门服务。”

Daisuke Takahashi

CyberAgent, Inc. 集团 IT 部门 CIU 解决方案架构师

详细了解 Dell Technologies 生成式 AI 解决方案。

在社交平台上关注我们。



DELLTechnologies

版权所有 © 2023 Dell Inc. 或其子公司。保留所有权利。Dell Technologies、Dell 和其他商标均为 Dell Inc. 或其子公司的商标。其他商标可能是其各自所有者的商标。本案例分析仅供参考。戴尔认为本案例分析中的信息截至 2023 年 9 月发布之时是准确的。如有更改，恕不另行通知。戴尔对本案例分析不作任何明示或暗示的担保。