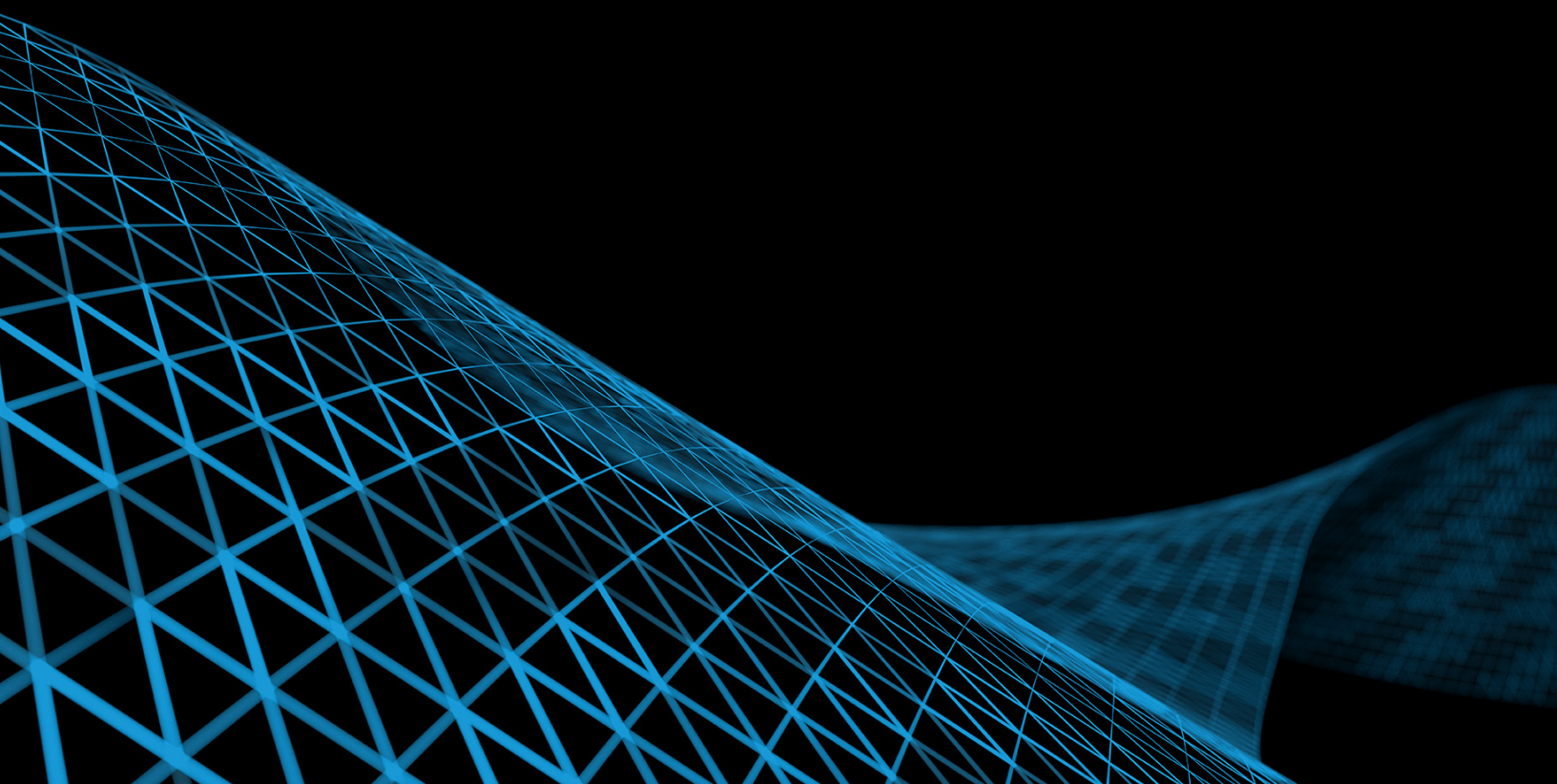
TechTarget | Custom Media
E-Book

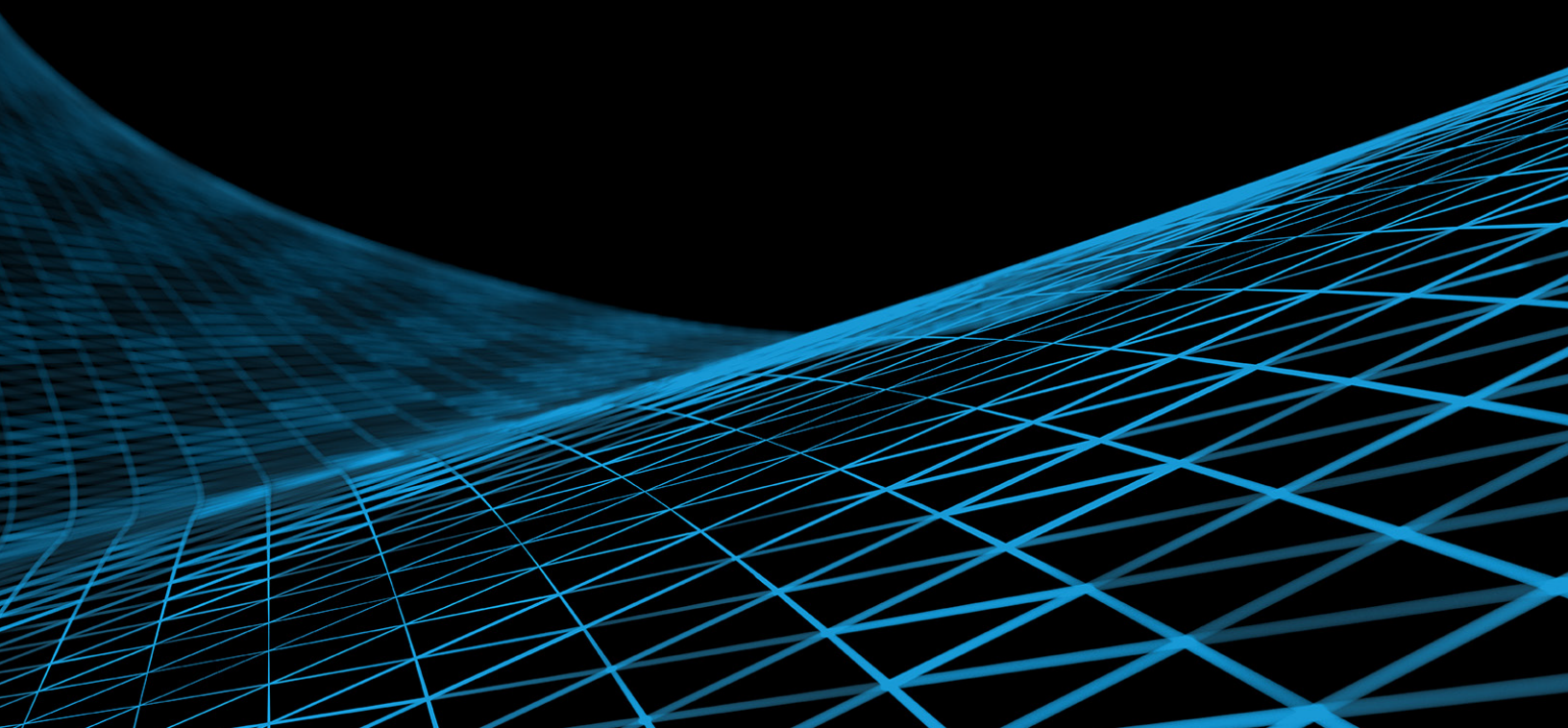
贯穿数据中心、边缘与云的 AI 变革

AI 的变革性影响极其深远。它正在从根本上改变云计算和边缘计算为企业创造价值的方式，并且范围覆盖各行各业、所有地区以及各种不同的应用场景。本电子书将从实用角度阐述 AI 在分布式云中日益重要的作用。

Akamai



目录



package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

前言

为什么本电子书很重要

无论您是 IT 决策者、企业高管、业务线利益相关者还是董事会成员，您很可能都在花费越来越多的时间去阅读、讨论和思考 AI 对企业的影响以及对个人生活和职业生涯的影响。

与先前大多数的热门技术相比，AI 的市场发展速度要快得多。AI 是近年来为数不多的几项关键技术之一，其早期的市场热议确实得到了印证，并且已经带来了切实可见的成果。（我们将分享一些数字来说明这一重要事实。）

不过，本电子书的作用远不止将 AI 的增长和潜力与现实环境联系起来，它还会通过一个重要的角度来分析 AI：靠近用户的 AI 和分布式云中的 AI。关于数据靠近用户和在云端的重要性，我们早已耳熟能详；现在，我们已进入去中心化 AI 推理和 AI 位于分布式云中的时代。

这就引出了我们最初的标题：为什么本电子书很重要。我们列出了决策者、知名人物及维护者需要阅读本电子书的四条理由：

1. **边缘计算和云计算是 AI 创新和价值创造的新领域。**通过将 AI 能力延伸至数据中心之外进行创建和部署，企业不仅能发掘新的洞察、业务模式与解决方案，获享它们所带来的优势，更能把握与过去十年 IT 市场主流发展方向相一致的机遇。
2. **AI 的监管和治理是一项艰巨的挑战。**若您认为在 AI 主要局限于本地沙盒部署时，“负责任的 AI”的使用和实践已属不易，不妨考虑一下靠近用户的 AI 和分布式云中的 AI 所带来的合规性、数据保护和安全问题。企业需要在这方面严格把关，否则可能会遇到很多问题。
3. **AI 在改变劳动力模式和实践方面的影响将远超预期。**毫无疑问，很多机械式的 IT 职能会由人类员工大幅转向 AI，但将 AI 转移到数据中心之外也会为 IT 和企业员工带来令人兴奋，甚至令人鼓舞的机会。
4. **您现在必须争分夺秒。**AI 计划正迅速占领各企业战略优先事项清单的榜首位置。如果您的企业尚未充分利用本地部署的 AI 来测试新的应用场景，也没有详细了解 AI 的功能，那么毋庸置疑您已经处于落后位置。但是，向靠近用户的 AI 和分布式云中的 AI 的转变可能代表着一个新的机遇。

我们相信您会发现本电子书有价值的另一个原因是：在本电子书的结尾，我们提供了具备实际借鉴价值的建议，可帮助您的企业充分利用向分布式云中 AI 的转变。

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh

第 1 章

AI 变革以及它如何改变了市场格局

鉴于 AI 实际上已存在数十年之久，将当今时代定义为“AI 革命”是否准确？抑或是，AI 仅仅是在其既有的形式与功能上发展演进——只不过是速度大大加快了而已？这些问题绝非是随口一问。我们的行业倾向于谈论那些似乎在一夜之间突然出现的技术变革，但除了那些已经在实验室中进行过实验的技术专家之外，这些技术变革几乎无人知晓。

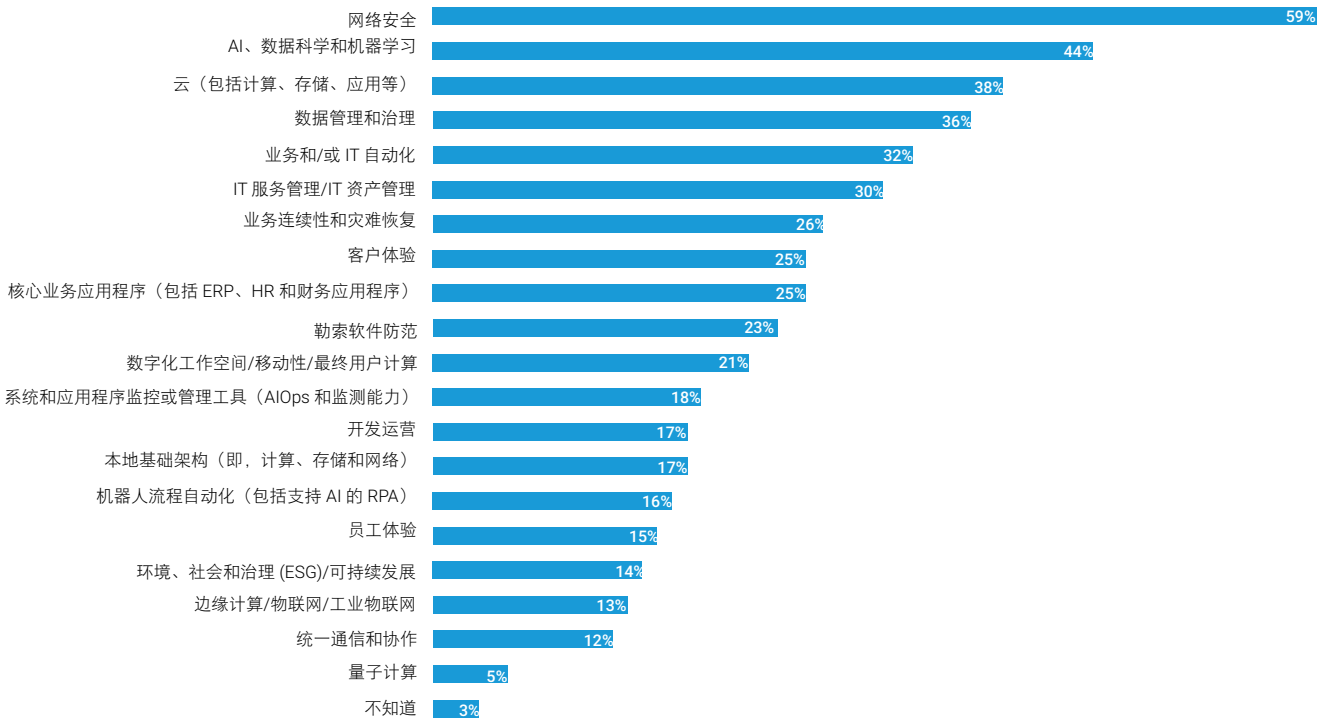
但出于多种原因，我们可以合理地称之为一场 AI 变革。首先，不可否认的是，AI 市场的增长以及企业负责人现在对这项技术的重视程度令人惊叹。其增长速度和当前的市场影响力远远超过了个人电脑、局域网、云计算和物联网等长期以来被视为 IT 行业发展里程碑的技术运动。

以下数据点说明了 AI 变革的影响、重要性和潜力：

- 到 2024 年底，全球在 AI 相关硬件、软件和服务上的支出超过了 6000 亿美元；到 2032 年，此支出将飙升至超过 2.7 万亿美元，年复合增长率超过 20%。[1]
- 97% 的企业负责人表示，其企业在 AI 投资方面看到了积极的回报。[2]
- 在短短一年内，将 AI 视为其两大技术优先事项之一的 CEO 所占比例从 4% 猛增至 24%。[3]
- 如下图所示，在过去两年里，几乎一半的美国企业都表示 AI 对企业的未来变得更加重要；在企业举措方面，只有网络安全的重要性高于 AI。[4]

图 1.安全引领 AI、云和数据管理成为关键技术举措

在过去两年里，以下哪些广泛的技术举措对贵企业的未来变得更加重要？（受访者百分比，N=1, 351，多项回答除外）



1 来源：《财富商业洞察》2024 年 12 月的报告“Artificial Intelligence market size ··· 2024-2032”。
2 来源：Lizzie McWilliams 于 2024 年 12 月 10 日在 EY.com 上发表的文章“Artificial intelligence investments set to remain strong in 2025, but senior leaders recognize emerging risks”。
3 来源：2024 年 5 月 Gartner 的 LinkedIn 动态，2024 年 CEO 调查显示 AI 是最重要的战略优先事项。
4 来源：Enterprise Strategy Group 于 2024 年 12 月发布的报告“Complete Survey Results: *2025 Technology Spending Intentions Survey*”。

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

事实上，AI 采用和资本投资的激增甚至超过了云计算，而后者是 IT 领域中最重要、增长最快的部分之一。AI 增长速度超过云计算的原因很简单：虽然云计算代表了一种开发和交付 IT 服务的强大且高度可利用的方式，但 AI 正在从根本上改变我们的工作和生活方式以及生活中几乎各个方面的互动方式。

此外，AI 的快速采用也暴露出传统云计算架构的局限性。例如，集中式云模型难以满足 AI 的实时、数据密集型特性对性能的高需求。从本质上来说，在性能、成本、人员技能、安全性和数据治理等问题上，AI 已显著改变了人们对云架构的期望。

是什么推动了 AI 市场的飞速增长和市场影响？一个主要因素是企业在过去一年中对 AI 模型的使用更加精明和成熟。由于 AI 模型的实用性和易用性迅速提高，并且很多企业已获得可观的投资回报，因此一些重要的应用场景应运而生，它们既加强了对如何使用 AI 的初步思考，也为如何从 AI 模型中受益开辟了全新的思路。

例如，不妨考虑以下情况：

- **业务预测**始终是成功企业不可或缺的一部分，而 AI 让这项功能更加精确、准确、及时并且更具可操作性。使用预测性 AI 模型等工具来预测未来趋势比十年前企业采用大数据分析时更加先进，因为现在的 AI 模型利用企业自身的数据进行了增强。与传统模型相比，这使得数据科学家和业务主管能够更快、更准确地筛选海量数据。
- **文档总结**过去是白领员工的工作内容，但生成式 AI (GenAI) 彻底改变了这一现状。现在，可以在很短的时间内对长篇幅的文档（包括那些充满神秘内容和技术性很强的内容的文档）进行自动审核、分析和总结。这不仅可以提供更深入的洞察，还能够使 IT 和业务专业人员以更具创新性和战略性的方式发挥作用，从而提升生产效率并改善员工体验。
- **客户流失建模**现在是客户关系管理中的关键一环，可帮助企业作出更明智、更符合实际情境的决策，从而预测客户流失的可能方式、时间和原因或者客户增加购买的可能性。
- **AI 驱动的聊天机器人**通过赋能大量日常请求的智能化、实时自助服务，正在推动客户服务、IT 服务管理和人力资源职能的转型。这些聊天机器人甚至可用于技术性很强的技术工程和计算机科学应用场景，例如代码生成和新功能建模。

这些应用场景帮助孕育了很多其他 AI 应用、工作流程和应用场景，例如网络安全威胁检测、数据分析/商业智能、竞争分析、快速原型设计、合规性管理等。

当然，任何新技术都会面临开发、部署和管理方面的挑战，即便是 AI 这样势头强劲、前景广阔的技术也不例外。这些挑战可能是技术挑战，例如缺乏数据质量、构建正确的基础架构或确保驱动大语言模型 (LLM) 所需的高性能，也可能是与业务相关或风险相关的挑战，例如数据治理、确保负责任地使用 AI 或克服巨大且不断增长的 AI 技能差距。

但初步迹象表明，各企业都了解这些挑战，并且正在利用适当的资源以及在高管层和董事会的支持下应对这些挑战。在下一章中，我们将介绍使 AI 超越科学项目的一项关键发展：将 AI 的重点从模型训练转向推理。

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

第 2 章

AI 的范式转变：从训练到推理

AI 模型训练曾经是市场发展的一项关键因素，也必然是各企业最初强调 AI 计划的重点。各企业很早便认识到，缺乏数据质量（或者说，缺乏正确类型和数量的数据）会影响 AI 模型的开发。Informa TechTarget 的 Enterprise Strategy Group 进行的研究发现，37% 的企业将数据质量问题视为实施 GenAI 时的最大挑战，这使其成为仅次于 AI 技能差距的第二大挑战。[5]

由于企业认识到利用自己的数据是解决数据质量问题和创建更准确、情境感知能力更强且响应速度更快的 LLM 的最佳方法，因此，现在构建正确的 AI 模型对企业来说已是水到渠成。此方法通常比购买使用第三方数据训练的现成商用 AI 模型更好，因为后者可能会造成数据偏见和不准确，或缺乏定制能力。作为一项标准实践，大多数企业会将其 LLM 训练集中在集中式云环境中，这对于 LLM 初始开发工作来说是一种有效方法。

AI 模型训练已成熟，并且质量和完整性已趋于稳定。目前，各企业已开始为其内部构建的训练模型分配更少的资源，尤其是当 AI 调整和优化减少了对计算密集型、以 GPU 为中心的基础架构的需求时。

Enterprise Strategy Group 的 AI、分析和商业智能实践总监 Mike Leone 针对从训练到推理的这一转变提出了独立的观点：

他表示："随着预训练模型和实时 AI 应用日趋成熟和普及，推理在高效、经济地实现 AI 的承诺价值方面发挥着核心作用。这种重心转变使各行各业都能进行可扩展的部署，而其主要驱动因素是硬件和软件堆栈中取得的持续进步。供应商们高度关注 AI 基础架构的合理调整和优化、提升模型效率和管理持续运营成本。当然，在推理向大众普及 AI 的同时，它也带来了可扩展性、隐私问题和监管审查等挑战。"

从 AI 模型训练向推理驱动型 AI 应用及具有重大影响的应用场景的转变正在顺利进行中。IBM 公司 AI 平台副总裁 Armand Ruiz 简洁而有力地指出："在我看来，推理将主导 AI 工作负载。在模型得到适当训练后，理论上它已不再需要进一步训练。但是，推理会持续进行。"[6]

当市场发生转型时，资本的投资方式和投资目标会不可避免地发生不成比例的波动。例如，一项估计指出，AI 硅芯片上的支出中大约有 15% 用于 AI 训练，有 45% 用于基于数据中心的 AI 推理，而剩余的 45% 用在基于边缘的推理上。[7]

5 来源：Enterprise Strategy Group 于 2024 年 9 月发布的报告"Complete Survey Results: *The State of the Generative AI Market: Widespread Transformation Continues*"。

6 来源：2024 年 12 月 Armand Ruiz 的 LinkedIn 动态。

7 来源：Jonathan Goldberg 于 2023 年 5 月 30 日在 TechSpot.com 上发表的文章"The AI chip market landscape: Choose your battles carefully"。

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


另一项研究指出了向推理的类似转变，而 Google 的一项研究以另一种方式强调了这种转变：Google 指出，其以机器学习为中心的能耗中有 60% 用于推理，而只有 40% 用于训练。[8]

这并不让人感到意外，这种从 AI 模型训练向 AI 推理的转变对 AI 基础架构和云架构具有重大影响。对于寻找适合的云环境和架构来开发、部署及管理 AI 应用的企业来说，此转变尤为重要。

传统的集中式云模型仍然非常适合处理很多业务工作负载，但 AI 对云架构提出了很高的要求，以满足大规模的 AI 性能需求。这些挑战很可能表现为令人沮丧的延迟问题，即查询和响应之间存在显著的时间差。

计算基础架构也是企业从 AI 模型训练转向 AI 推理时需要考虑的一个关键问题。以 AI 为中心的 GPU 凭借其强大的 LLM 训练能力而备受赞誉，因为它们能够提供很高的计算性能和内存带宽。相比之下，推理需要不同的基础架构思维，即注重低延迟和高效的资源利用。

在下一章中，我们将仔细研究企业如何通过将推理资源部署在更靠近用户的位置以及分布式云架构中来帮助满足 AI 推理的性能需求。

8 来源：Scientific American 网站 2023 年 10 月 13 日的文章“The AI boom could use a shocking amount of electricity”。

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

第 3 章

去中心化 AI 推理和分布式云的兴起

正如我们在上一章中阐述的，传统的集中式云架构给那些希望优化 AI 应用以创建突破性和变革性应用的企业带来了会引发问题的限制和挑战。随着 AI 需求从模型训练转向推理，人们应当考虑延迟、网络带宽限制、确保大规模性能和管理复杂性等所有问题并进行相应的规划。

现代和下一代 AI 推理应用需要能够在数据驻留的位置开发、部署和利用推理。这越来越趋向于靠近用户或在敏捷、可扩展且灵活的分布式云架构中。

在我们进行深入探讨之前，明确一些定义会很有帮助。

- *去中心化 AI 推理*是一种 AI 工作流程设计范式，这些流程跨越集中式数据中心（云）和云外部更靠近人类及物理事物（边缘）的设备。这与更常见的做法形成鲜明对比，即 AI 应用完全在云端开发和运行，人们开始将其称为*云 AI*。此外，它与早期的 AI 开发方法也有所不同，在这些早期的方法中人们会在台式机上编写 AI 算法，然后将其部署在台式机或专用硬件上，以完成读取支票号码之类的任务。

云服务的分布可帮助企业满足应用程序性能和响应时间、边缘计算、监管要求或其他需求的目标...

- 作为优秀的云、安全和内容交付平台提供商，Akamai 将*分布式云*定义为“一种云计算形式，企业可利用多个地理位置的公有云基础架构，而运营、治理和更新由单个公有云服务提供商集中管理。分布式云服务有助于企业满足应用程序性能和响应时间的目标、边缘计算需求、监管法规的要求或其他需求，这些需求可以通过从更靠近最终用户和设备的位置提供云服务来满足。之所以要使用分布式云计算，是源自于物联网 (IoT) 网络、人工智能和其他必须实时处理大量数据的应用场景的兴起。”

去中心化 AI 推理和分布式云对于 AI 推理的开发及利用都至关重要，因为它们代表了现代、高效、安全且可靠的方法，可以克服先前提到的挑战：高性能、大规模性能、计算的高效使用、存储和网络资源，以及克服与通常包含大量非结构化数据的大型数据集相关的延迟问题。

这在现实生活中如何运作？某家公司希望将个性化打造成其客户体验要求的标志，并需要确保他们从各种设备和数据源中捕获数据。他们决定使用依托 LLM 的 AI 聊天机器人，将本地化数据分发到靠近用户的位置。这降低了延迟并缩短了响应时间，从而显著提升了性能并实现了更快、更具个性化的客户支持。

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

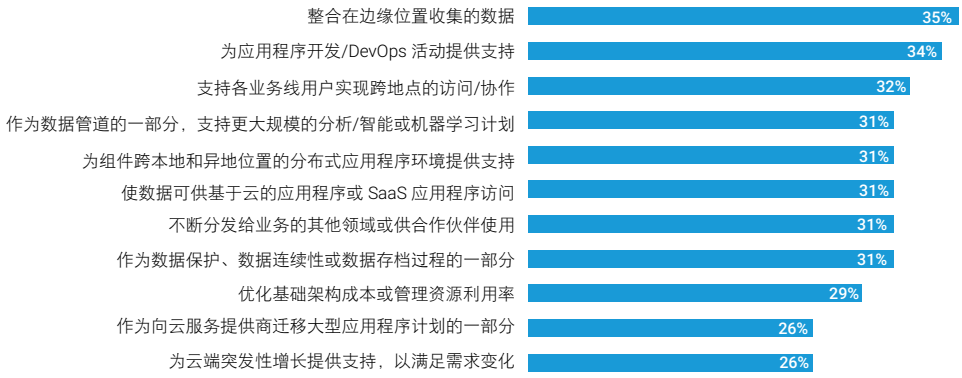

为利用 AI 推理应用所带来的增长机遇，越来越多的企业采用这一方法。为了解决受计算基础架构限制和延迟问题影响的性能瓶颈，AI 工作负载正在通过去中心化 AI 推理和分布式云架构向更广泛的最终用户所利用的数据源靠近。

例如，此模型广泛用在设施分布广泛且地理位置分散的行业中，包括零售、电信和媒体。此方法正逐渐成为 AI 推理的标准架构模型，因为它能够克服与复杂的 AI 推理要求相关的延迟和大规模性能挑战。

促使 AI 推理转向分布式云模型的一个主要驱动因素是，企业越来越多地需要使用边缘在数据中心与云服务提供商 (CSP) 之间移动数据。Enterprise Strategy Group 发现，定期在数据中心与 CSP 的基础架构之间移动数据的企业中有 35% 的企业会这样做，以便整合在边缘位置收集的数据，这成为此类数据移动的首要动机。[9]

图 2.边缘是在数据中心与公有云服务之间移动数据的首要驱动因素

您表示自己的企业会定期在数据中心与公有云服务之间移动数据。贵企业出于什么目的在数据中心与公有云服务之间移动数据？（受访者百分比，N=170，mult）



9 来源：Enterprise Strategy Group 于 2023 年 11 月发布的报告“Research Report: *Distributed Cloud Series: The State of Infrastructure Modernization Across the Distributed Cloud*”。

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

第 4 章

为什么未来以去中心化 AI 推理为主导

若想帮助企业从推理中获取最大收益，关键在于利用来自边缘位置的数据构建和部署 AI 推理。随着智能手机、专用手持设备、数字签名和物联网设备等各种边缘设备产生并存储的数据越来越多，企业必须采用边缘思维才能实现对 AI 推理不可或缺的实时数据处理。

转向去中心化 AI 推理以满足 AI 推理要求，这种转变迅速而又影响深远。研究表明，在靠近用户的推理系统上的支出将增长十倍，从 2024 年的 270 亿美元增长至 2032 年的近 2700 亿美元。[10]

在帮助企业开发和部署 AI 推理应用方面，去中心化 AI 推理有很多重要的考虑因素，例如设备原生的 AI 处理、针对边缘设备优化的集成安全框架、低能耗和低延迟。这些特性会实现更高的大规模性能、更准确且更深刻的数据洞察以及更低的带宽投资。这进而使分散在各地的企业成员能够利用 AI 推理，在更靠近数据生成、存储、处理和访问位置的地方作出更明智的实时决策。

以 Bloomfield 为例，这家具有远见卓识的企业利用去中心化 AI 推理来改善基本运营并简化数据管理。对于经常在网络访问受限情况下运营的农场主来说，他们需要以高效、经济且安全的方式管理来自其分布式智能摄像头的海量数据。通过使用去中心化 AI 推理架构模型，Bloomfield 利用 AI 推理来帮助确定如何以最佳方式组织、存储数据并呈现给农场主。这简化了作物健康管理、提高了产量，并且在远离本地数据中心的地方实现了更快、更准确的决策。

去中心化 AI 推理能够在各种分布式位置为 AI 推理提供支持，这使其自然而然地适用于以下应用场景：

- **床边医疗和远程医疗应用**——在这些应用场景下，临床医师可以使用智能手机、物联网设备和其他轻便但功能强大的便携设备随时随地作出决策。去中心化 AI 推理也非常适合于使用医疗保健行业的智能设备（例如输液泵、起搏器、心脏监护仪和药物分析仪）进行实时决策。
- **智慧城市**——在此应用场景下，边缘系统可以执行从控制交通流量、监控公共事业系统行为到简化选民登记、提升公共安全等各项任务。
- **无人驾驶汽车**——它们使用智能摄像头和车载传感器数据来评估路线选择、监控并改进燃油使用情况、记录 GPS 变化，以及与执法机构和公共安全机构进行通信。
- **零售**——支持从 IP 视频监控、库存管理到防盗、客户店内流量模式等各种 AI 推理应用。

随着企业想方设法来降低实时查询响应的延迟，去中心化 AI 推理将成为更高效、更明智且更准确的业务决策来源的关键要求。AI 应用无须再向遥远的数据中心发送或请求数据，而是将在全球分布式云上处理和共享数据。

一家大型广告技术公司的工程主管指出：“我们花费了很多时间来优化我们的 AI 模型。现在，我们认识到，不能事后才考虑推理。”

企业已别无选择，必须将 AI 推理作为其业务战略的重要组成部分。他们必须增强紧迫感，并将去中心化 AI 推理用于推理任务，因为他们的很多竞争对手已经在靠近用户的位置使用 AI 推理并取得了重大进展。

10 来源：《财富商业洞察》2024 年 12 月的报告“Edge AI Market Size ··· 2024-2032”。

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

第 5 章

对 AI、边缘和分布式云的战略性展望：发展方向和重要意义

评估、选择成熟的技术合作伙伴并与之进行合作是企业的 AI 推理计划取得成功的根本要素，但它难以实现。

它之所以很重要，是因为以下不可否认的事实：在 AI 基础架构、云架构、推理应用场景、应用优化和完美部署方面存在重大 AI 技能差距。

但是，它也因为缺少适合的候选方案而变得非常棘手。如上所述，成功的 AI 推理工作必须将技术知识、领域专业知识、对安全、治理和风险管理的理解以及对实施和持续管理中细节的狂热关注完美地融合在一起。

为了在整个物理和虚拟企业中开发并部署 AI 推理（从核心、边缘到云端再到核心），企业需要与有相关经验的提供商合作。拥有一个了解成功 AI 项目的技术和业务要素的合作伙伴必不可少。

Akamai 是先进的云计算、安全和内容交付解决方案的出色供应商，在构建、实施和管理 AI 计划（从模型训练到推理）方面拥有良好的记录。它在复杂、计算密集型工作负载方面拥有丰富的实践经验，让它能够与不同应用场景、行业和地区的各种客户合作。

为了帮助企业顺利地满足自身的 AI 推理需求，Akamai 在一致、可靠且安全的基础上成功地部署了一个分布式云架构，使企业能够在更靠近最终用户的位置处理 AI 工作负载。这对于降低延迟并确保大规模的高性能来说非常重要，因为这两点对计算密集型 AI 推理工作负载来说必不可少。

Akamai 分布式云架构的主要优势之一是，让开发人员能够在基于开放云标准的云原生框架中完全专注于和平台无关的开发。

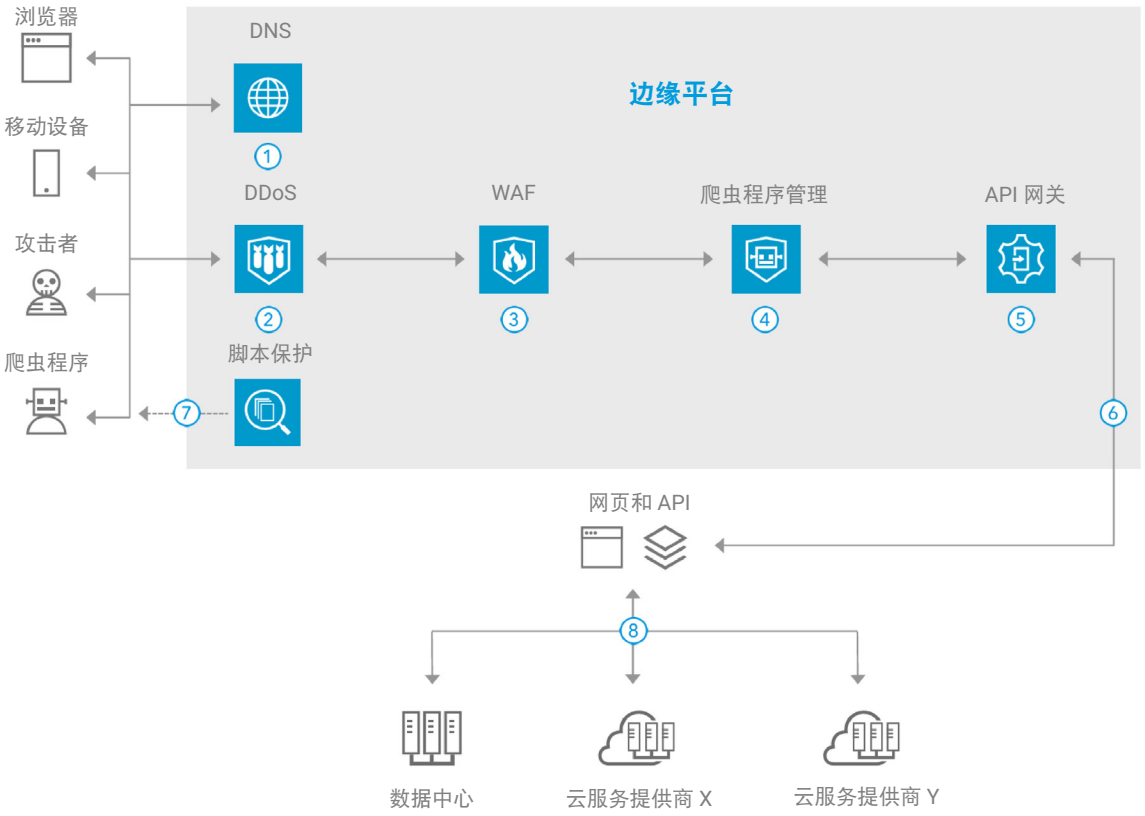
为了实现大规模的高性能，Akamai 采用了多种不同的缓存技术，并与开源软件集成，以确保提供足够的吞吐量并降低延迟。例如，很多使用 Akamai 分布式云架构的客户都表示，他们在面向客户的聊天机器人这样非常重要且热门的推理应用场景中将响应时间缩短了 50%。借助此类近乎实时的响应，各个行业（包括零售、医疗保健、金融服务和高科技）的企业都能够以更专业的方式更快地提供针对其独特需求量身定制的客户服务。

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Akamai 还提供了一个 AI 参考架构（请参见下图 3），使企业能够快速上手并实现较高的经济和运营价值。此参考架构非常全面，通过一组熟悉的浏览器和用户定义的设备，企业可以部署从核心数据中心到边缘、云端再到核心数据中心的企业 AI 解决方案。这可以让企业中的所有各方受益，包括开发人员、IT 专家、基础架构管理人员、云架构师和业务线利益相关者。

图 3.

随着技术的飞速进步以及 AI 工作负载的复杂性和价值不断增加，企业应迅速、果断地采取行动，超越竞争对手。通过采用去中心化 AI 推理和分布式云架构来充分发挥 AI 推理的潜力至关重要，在由 Akamai 这样公认的出色企业进行规划、开发、部署和管理时尤其如此。



```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


结语

企业接下来应该做什么

人们认为，AI 已经彻底改变了卓越 IT 和业务转型的现状。但是，优化企业 AI 投资价值的关键在于快速从 AI 模型训练转向 AI 推理。

为了充分利用 AI 推理所提供的一切，企业必须在更靠近工作完成的地方以及关键人员可以利用系统实现卓越成果的地方，开发和部署其推理应用及工作负载。这意味着，从集中式云计算模型转变为在边缘由 AI 深度塑造的分布式云模型。

如本电子书中其他地方所述，企业需要考虑的问题有很多，需要了解并克服的挑战也有很多。对尚未将 AI 计划确定为战略优先事项的企业来说，这或许看起来令人望而生畏，但他们可以并应该执行一些重要步骤来取得成功并保持领先地位。

Enterprise Strategy Group 的 Leone 强调道：“那些在 AI 领域奋起直追的企业不应该因为害怕错过追赶技术的机会而受到影响。重点应该放在建立正确的基础上，可通过采取确定 AI 可以解决的具体业务问题、评估现有的数据基础架构和内部技能以及确定可以弥补差距的关键合作伙伴等步骤来实现。在确定一到两个影响深远并且能够使用现有的优质数据来展示明确价值的试点项目和应用场景时，他们应该采取战略性措施。”

任何 AI 推理计划要想取得成功，知道要衡量的对象以及如何衡量它都至关重要。

“在进行所有这些工作的同时，他们应该培养内部人才并建立治理框架，以确保负责任的使用。关键在于避免仓促实施，而要采取战略性步骤，建立基础性的并且最好是能够长期存在的能力。”

为此，您和您的同事应该了解并尽快做到以下四点：

1. **重视应用场景的选择。**这是至关重要的第一步，因为您需要对自己正确部署 AI 推理的能力树立信心。此外，还必须通过早期取得的一些成功在您的业务利益相关者与财务审批者之间树立信心。
2. **基础架构选择很重要。**很显然，AI 推理是计算密集型工作，因此最简单的决定是采用性能卓越的 GPU 作为您系统的核心。但是，对于 GPU 和其他 AI 基础架构来说，有很多绝佳选择。花时间把事情做好（或者，更有可能的是，依靠有相关经验的合作伙伴的经验和专业知识）。
3. **获得管理层的支持。**重点是要关注每一位高管层和董事会成员都关心的两个核心问题：我们的机会是什么？如何平衡潜在的风险？确保您的项目和应用场景与主要业务目标一致，例如改善客户体验、更快地确定并利用市场机遇以及创建更准确的 beta 测试流程。与此同时，从 IT 部门、业务线部门以及董事会成员等群体中寻找并招募高层支持者。
4. **定义成功指标并取得成功。**任何 AI 推理计划要想取得成功，知道要衡量的对象以及如何衡量它都至关重要。选择一些现实、相关且足够稳健的指标（或者与您的合作伙伴一起开发定制指标）来定义“重大胜利”，尤其是在您通过初步的 AI 应用场景取得一些容易实现的成果之后。

本内容由 Akamai 委托 TechTarget Inc. 制作。