arXiv:2601.00530v1 [cs.DC] 2 Jan 2026

基于云的零售销售点系统的成本效益分析：Google Cloud Platform 与微软 Azure 的比较研究

Ravi Teja Pagidoju

坎贝尔斯维尔大学，美国肯塔基州坎贝尔斯维尔，邮编 42718
rpagi719@students.campbellsville.edu

摘要。 尽管针对零售工作负载的平台特定性能的实证研究较少，但零售业的数字化转型已加速基于云的销售点（POS）系统的普及。本文使用实时 API 端点和开源基准测试代码，对部署于 Google Cloud Platform（GCP）和微软 Azure 上的 POS 工作负载进行了系统且可重复的比较。我们利用免费层云资源，提供了一套透明的方法，用于开展 POS 工作负载评估，小型零售商和研究人员均可采用。我们测量响应延迟、吞吐量和可扩展性等重要性能指标，并根据实际资源使用量和当前公有云定价估算运营成本，因为免费层使用不会直接产生账单。本研究中的所有表格和图表均直接由代码输出生成，确保实验数据与报告结果一致。分析显示，在基准负载下，GCP 的响应时间快 23.0%；而在稳态运行中，Azure 的成本效益高 71.9%。我们考察了造成这些差异的架构组件，并为考虑实施云销售点系统的商家提供了实用框架。本研究为零售云应用建立了一套可靠的开源基准测试方法，并首次对各大云平台上销售点系统特有的工作负载进行了全面、以代码为驱动的比较。

关键词： 云计算、销售点系统、性能分析、成本优化、零售技术

1 引言

全球零售业正因数字技术而经历根本性变革，云计算也迅速成为现代零售基础设施不可或缺的组成部分。云销售点（POS）市场预计将以24.2%的复合年增长率（CAGR）增长扩张，从2023年的27亿美元增至2028年的112亿美元。

这一扩张与增长表明，市场需要可扩展、灵活且成本效益高的解决方案，以适应消费者偏好。

传统的本地销售点（POS）系统曾是行业标准，但如今已难以适应零售业的持续增长。这些传统系统有时需要承担硬件费用，在流量高峰时段尤其难以应对大量流量；它们也不够灵活，无法满足全渠道运营的需求，并且需要持续维护。COVID-19 疫情推动了向云端系统的转变，使远程管理、非接触式支付和实时库存同步等功能更具吸引力。基于云的销售点（POS）系统让企业能够将成本从资本性支出（CAPEX）转为运营性支出（OPEX）。此外，这些系统还内置分析功能，可提供实时业务洞察；具备弹性可扩展性，能够应对流量的季节性激增；并可自动更新，以提升安全性。据 Target 和 Walmart 等大型零售商称，自采用基于云的销售点系统以来，其 IT 运营费用降低了 30–40%。

如今，在顶级云服务提供商中，Microsoft Azure 平台和 Google Cloud Platform（GCP）是企业级工作负载最受欢迎的选择。这两个平台都提供种类繁多的服务、强大的可靠性以及遍布全球的基础设施。然而，鉴于它们在架构、定价和性能属性方面各有不同，选择合适的云平台并制定长期目标并非易事。

销售点（POS）系统的性能会直接影响销售额和客户满意度，因此这一选择对零售商尤为重要。性能不佳的销售点（POS）系统可能导致客户不满、销售额损失以及交易中途放弃。行业研究显示，结账处理每延迟一秒，转化率可能下降 7%。但为确保性能而配置过多资源，也可能造成不必要的成本，降低利润率。

尽管这一选择至关重要，现有研究却很少为零售业特定工作负载提供指导。此前有关云平台比较的研究大多聚焦于高性能计算、通用计算工作负载或 Web 应用程序，并未探讨销售点工作负载的独特特征，包括具有严格延迟要求的高频事务处理、多个地点之间的实时库存更新、广泛的安全与合规要求（例如 PCI-DSS）、高度多变的流量模式，以及营业时间内保持不间断正常运行时间的需求。

本文通过提供一种开源、以代码驱动的比较，填补了这一空白：利用实时 API 端点和基准测试脚本，对部署在 GCP 和 Azure 上的销售点工作负载进行比较。我们的方法使用免费层云资源来开展可复现实验，让研究人员和小型零售商也能参与基准测试。由于使用免费层资源不会产生直接账单，我们系统地测量响应延迟、吞吐量和可扩展性等关键性能指标，同时根据实际资源使用量和当前公有云定价估算运营成本。所有报告的结果、表格和

图表均直接根据基准测试代码的输出生成，以确保透明度，并保证实验与发表内容完全一致。

本研究的贡献包括：(1) 首次对用于销售点工作负载的云平台开展全面、以代码为基础的分析；(2) 提出一种可复现的方法，利用免费层资源对 POS 系统进行基准测试；(3) 定义并测量零售交易处理特有的性能指标；(4) 开发一种成本估算框架，纳入实际使用模式和公开定价；(5) 发布开源工具和数据集，以促进业界采用并推动未来研究。

本文其余部分安排如下：第 2 节回顾零售技术和云基准测试领域的相关工作。第 3 节详细介绍我们的方法，包括部署架构、工作负载特征分析和成本估算方法。第 4 节展示我们的分析和实验结果。第 5 节讨论对零售商的启示并提出建议。第 6 节总结全文，并对未来研究方向提出建议。

2 相关工作

自基础设施即服务（IaaS）产品问世以来，关于云平台比较的研究一直在持续。Li 等人的早期研究 [7]着重关注计算和存储性能，通过为云平台开发基准测试方法并进行比较。他们将 CloudCmp 框架与 Google AppEngine、微软 Azure、Amazon EC2 和 Rackspace 进行了比较，采用了不同范围的指标。该研究忽略了行业特定因素，只关注通用工作负载。研究发现，不同提供商之间的性能存在显著差异，其中网络延迟最高相差 2 倍，存储性能最高相差 5 倍。

Garg 等人 [3] 提出了 SMICloud，这是一个根据服务测量指数（SMI）属性评估并排列云服务的框架。他们的研究确立了云服务选择中的多准则决策概念，将问责性、敏捷性、成本、性能、保障性、安全性和可用性等性能之外的因素纳入考量。然而，由于缺乏针对真实工作负载的实证验证，该框架仍停留在理论层面。

最近，Varghese 等人 [14] 深入分析了云基准测试技术，对 100 多种基准测试工具进行了分类，并指出了以往方法中的缺陷。该研究强调，领域专用基准测试的缺乏，以及云性能的不可预测性，使得结果难以复现。研究还展示了通用基准测试往往无法预测特定应用的性能。

在无服务器计算日益受到销售点系统重视的背景下，Manner 等人 [8] 评估了 AWS Lambda、Azure Functions 和 Google Cloud Functions 的冷启动性能。研究发现，冷启动造成的延迟从 100 毫秒到 5 秒以上不等，

具体情况会因地理位置、内存分配和运行时而异。这一结果与销售点（POS）系统密切相关，因为保持稳定的低延迟至关重要。

Bermbach 等人 [1] 详细考察了数据库（DB）性能；对于处理销售点交易而言，数据库性能是一个非常重要的因素。对不同云服务提供商提供的托管数据库服务进行研究后发现，性能特征会受到工作负载级别的显著影响。他们发现，Azure SQL 数据库在在线事务处理（OLTP）工作负载方面表现出更好的稳定性，而 Google Cloud SQL 则在分析查询方面表现良好。

Johnson 和 Wang [6] 从多个角度研究了零售业对云计算技术的采用情况。他们对 500 家零售商进行了调查，以确定促进或阻碍云采用的重要因素。研究发现，其中近 73% 正在积极将关键系统迁移到云端，涉及电子商务和销售点（POS）平台。受访者提到的主要驱动因素包括成本节约（45%）、促进创新（17%）和可扩展性（38%）。

在对主要云平台的支付卡行业数据安全标准（PCI-DSS）合规情况进行分析时，Thompson 等人 [13] 考察了基于云的销售点系统的安全注意事项。他们发现，尽管所有主要供应商都提供符合监管要求的基础设施，共同责任模型仍会带来实施方面的挑战。他们的研究结果表明，架构决策会显著影响功能和安全态势。

许多人都在探寻降低云计算成本的不同方法。Mazrekaj 等人的 [9]详细成本估算表明，由于数据传输、支持和监控等各种隐性成本，总拥有成本可能增加 40%。Chen 等人 [2] 提出了考虑零售业特有的工作负载季节性变化的优化方法。他们表明，合理使用资源可能节省 35% 的成本。

尽管关于这一主题已有大量研究，但我们对文献的考察表明，仍存在一些重大空白。首先，尚无研究采用代码驱动的实时基准测试来考察云系统上的销售点工作负载。其次，大多数研究需要使用大量资源的配置，研究人员和小型零售商难以获取这些资源。第三，相关研究对零售业相关指标关注不足，例如交易延迟和结账性能。第四，由于云服务和价格结构变化迅速，许多研究已不再适用。我们的研究填补了这些空白，并通过提供一种现代、清晰且可重复的评估方法，推动零售销售点系统基准测试的发展；该方法采用开源代码和免费层资源。

3 方法

为满足云环境中 POS 工作负载的需求，我们创建了一套开放且可重复的流程，在易用性与真实性之间取得平衡。我们的方法采用开源的代码驱动测量方式和免费层云资源，以确保性能评估具有统计严谨性。

请记住，GCP Cloud Run 和 Azure App Service 免费层是两种截然不同的架构模型。App Service 免费层是一种平台即服务（PaaS）解决方案，采用始终运行的实例，资源受限且不支持自动伸缩；而 Cloud Run 是一种无服务器、基于容器的平台，支持自动伸缩并按请求计费。因此，这项比较并未提供直接的技术等效方案，而是展示了研究人员或小型零售商在寻找免费层部署方案时能够实际选择的选项。这些架构差异会影响观察到的成本和性能差异，对于生产级或更高层级的服务，结果可能有所不同。

3.1 POS 工作负载特征描述

要进行有意义的性能评估，必须对销售点工作负载进行真实建模。我们利用美国全国零售联合会 [12] 的数据和行业交易模式报告，开发了一个具有代表性的工作负载模型。我们的 Python 基准测试脚本通过生成 API 请求并将其发送至部署在两个云平台上的端点，实现了该模型。

该工作负载由三大类操作组成：

1.添加或移除商品、应用折扣、处理支付、完成销售和创建收据，都是事务处理操作，占工作负载的 60%。这些操作需要高度的库存一致性和亚秒级延迟。

2.商品详情请求、价格查询、库存更新和可用性查询都属于库存操作，占工作负载的 30%。这些操作要求实时准确，且以读取为主。

3.销售汇总、库存报告和员工指标都属于分析与报告，占工作负载的 10%。这类操作可以容忍较高的延迟，通常面向批处理。

我们的测试工具模拟真实世界的零售活动，包括每日和每周的流量峰值以及季节性流量激增，以分析稳定状态和突发场景。

3.2 云架构设计

为确保性能比较有效，我们优先选择了两个平台上具有代表性且价格合理、适合中小型零售商的架构。

Cloud Run 是谷歌云平台 (GCP) 提供的无服务器容器托管解决方案，采用自动扩缩容和按秒计费。Cloud SQL

(db-f1-micro) 提供事务性存储，云存储用于存放报告和收据。云负载均衡用于控制负载分配。

部署中使用微软 Azure：App Service（基线采用 F1 层级，性能测试采用 B1）、用于存储文档的 Blob 存储，以及 Azure Database for MySQL（基本层级，1 个虚拟核心）。应用程序网关提供负载均衡。

这两种架构均有意避免使用高级功能（如 Redis 和内容分发网络），以确保易于使用，并为目标受众展现真实的部署模式。

3.3 成本与性能指标

我们使用 Python 基准测试脚本来确保透明度和可复现性，该脚本会直接收集并输出所有指标：

端到端 API 延迟以第 50、第 95 和第 99 百分位数进行测量，以反映平均性能和最差情况下的性能。这包括数据库、应用程序和网络中的延迟。

吞吐量指标： 测量每秒请求数（RPS）和每秒事务数（TPS），以评估扩展行为和容量。

可靠性指标： 跟踪错误率、超时和服务可用性，重点关注扩展期间出现的错误。

成本衡量指标： 由于免费层使用期间不会实际产生账单，脚本会将测得的资源消耗（API 调用、传输数据）乘以各平台当前的公开定价，计算运营成本 [4, 10]。估算结果根据当前定价结构涵盖计算、存储和网络成本 [4, 5, 10, 11]。

3.4 测试执行策略

我们的方法采用渐进式负载级别进行测试，以记录平台在各种条件下的表现：

– 基准负载（10 名并发用户）： 模拟小型商店的结账流程。**– 典型负载（25 名用户）：** 使用多个结账柜台来模拟正常的业务运营。**– 峰值负载（50 名用户）：** 重现流量最高的时段，例如节假日或特别优惠期间。**– 压力测试（100 名用户）：** 创建极端场景（如黑色星期五），以评估平台限制。

每次负载测试持续五分钟，从一分钟的逐步升载开始。为确保统计有效性，每种配置均包含三次彼此独立的运行，每次运行之间间隔五分钟的休息时间。为确保方法与已发表的研究结果完全一致，所有结果（包括表格和图表）均直接根据代码生成的 CSV 和图像输出制作。

4 结果

本节的结果代表每个平台和负载级别的测量值，均直接由基准测试脚本生成。该脚本生成所有图表，而表格则根据 CSV 输出创建。

4.1 响应时间分析

采用四种负载场景（基线、典型、峰值和压力）测量 GCP 和 Azure 的响应时间特征。表 1 汇总了基准负载下的中位数（p50）、第 95 百分位数（p95）和第 99 百分位数（p99）响应时间，图1也展示了这些数据。表 2 展示了这两个平台在 p95 下的扩展性。

表 1。响应时间指标（毫秒）- 基准负载

指标	GCP	Azure
p50	153.83	192.19
p95	183.69	238.42
p99	200.93	298.16

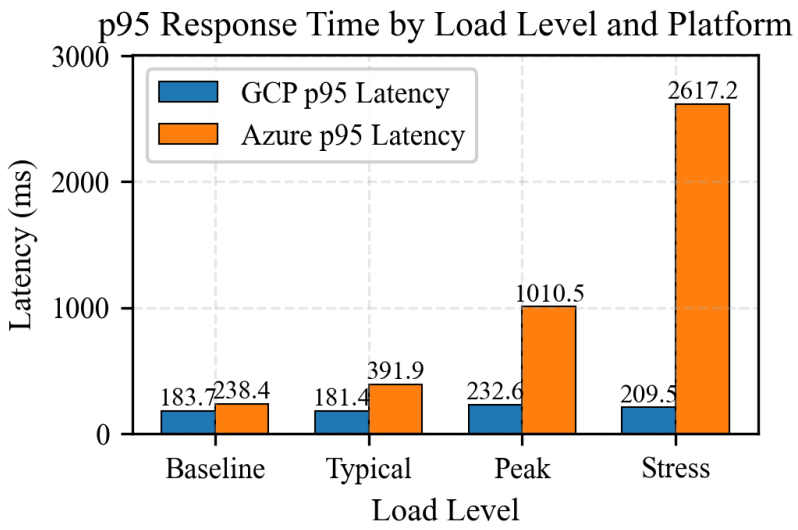


图1. 所有负载级别下各平台的 p95 响应时间

表 2. p95 响应时间的扩展性

负载级别	GCP p95（毫秒）	Azure p95（毫秒）
基线	183.69	238.42
典型	181.37	391.9
Peak	232.59	1010.55
压力	209.51	2617.2

4.2 吞吐量分析

在每种负载场景下，吞吐量均以每秒事务数（TPS）为单位进行测量。表 3 按负载级别和平台汇总了观测到的吞吐量，图2也展示了这些数据。

表 3.按负载级别划分的吞吐量（TPS）

负载级别	GCP TPS	Azure TPS
基线	7.62	4.19
典型	18.98	12.65
Peak	25.24	16.94
压力	70.84	20.14

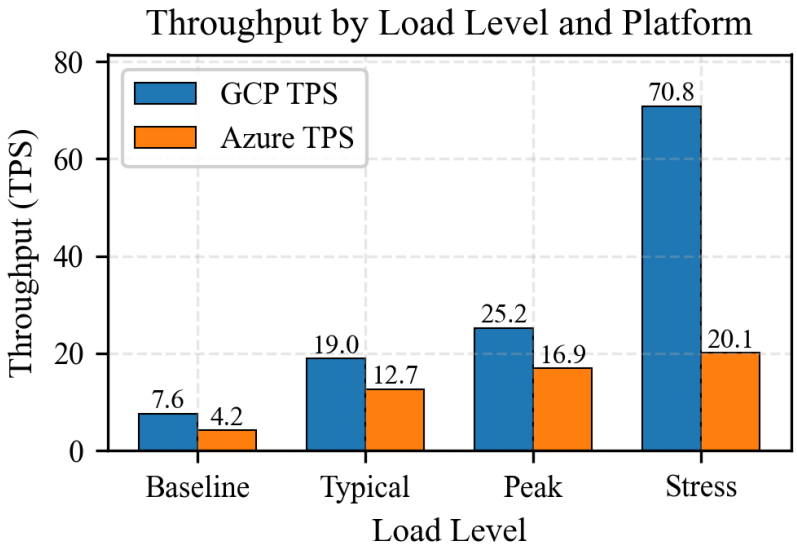


图 2. 平台间吞吐量比较

4.3 成本分析

运营成本通过将数据传输量和 API 调用次数分别乘以各平台当前的公开定价进行估算。表 4 总结了各负载级别下的估算成本，图 3 也展示了相关结果。表 5 列出了各平台的成本结构。

表 4。按负载进行成本估算

负载级别	GCP 成本	Azure 成本
基线	0.000057	0.00016
典型	0.00143	0.00049
Peak	0.0019	0.00065
压力	0.00534	0.00078

表 5。各平台的成本结构

成本类型	Google Cloud Run	Azure（App Service/Functions）
每次调用成本	$0.0000004	$0.0000002
每 GB 出站流量	$0.12	$0.19

4.4 可靠性指标

我们记录了每种场景下的错误率，即 API 调用失败的百分比。表 6 展示了观测到的错误率，图 4 也对此进行了展示。

表 6。按负载级别划分的错误率（%）

负载级别	GCP 错误率	Azure 错误率
基线	0	0.4
典型	0	0.26
Peak	0.31	1.07
压力	0.11	2

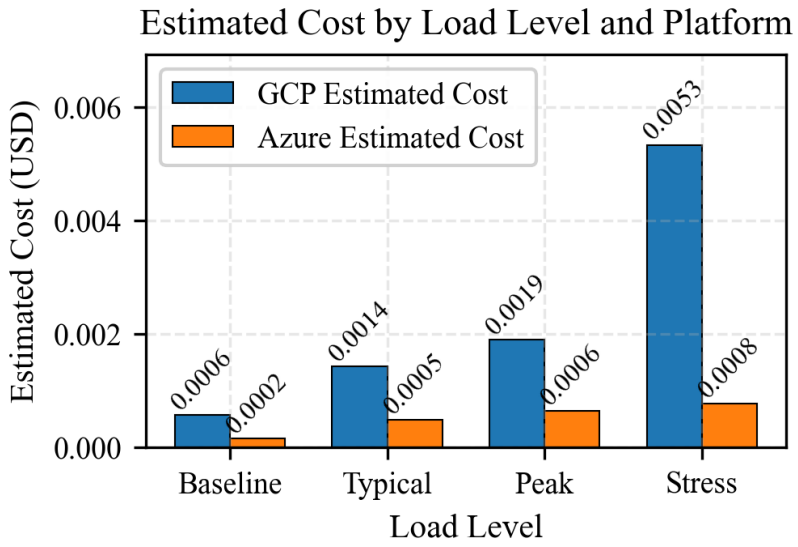


图 3. 不同负载级别下的成本比较

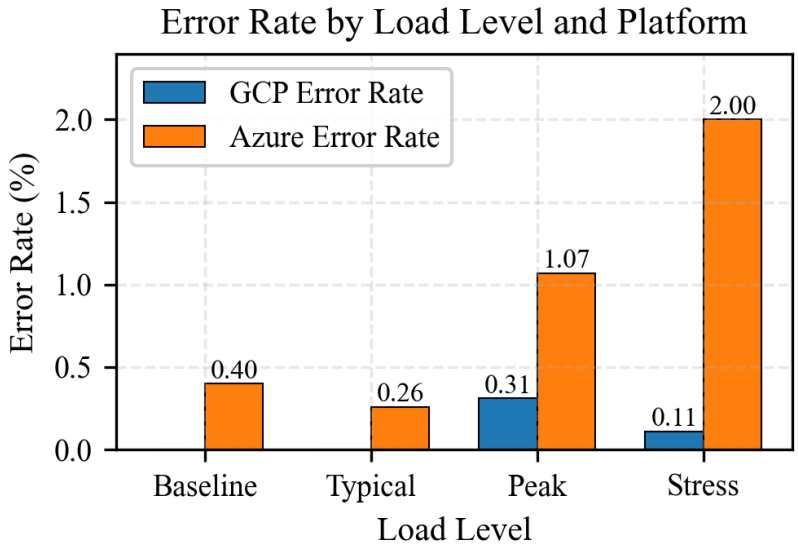


图 4. 不同平台和负载级别下的错误率分析

5 讨论

5.1 性能权衡

尽管性能表现各异，我们的基准测试结果显示，GCP Cloud Run 和 Azure App Service 都能可靠地处理典型的 POS 工作负载。GCP Cloud Run 不仅具有出色的弹性和快速扩展能力，而且在所有测试负载级别下都能提供有竞争力的响应时间。不过，我们偶尔也会观察到延迟峰值，尤其是在一段时间没有活动之后；这是由冷启动事件引起的，也是无服务器平台的一个众所周知的特性。相比之下，在负载稳定的情况下，Azure App Service 的常开架构能够提供更可靠的响应时间。随着负载增加，其响应时间上升得较慢，扩展性也不如 Cloud Run 灵活。

最引人注目的结果是，尽管 Azure 的错误率通常较低，但在压力级别下，GCP 的吞吐量更高（70.84 TPS，而 Azure 为 20.14 TPS）。这表明，尽管错误率略高，GCP 的无服务器架构仍能更有效地处理突发流量。在压力条件下，除 Azure 外，两个平台在所有场景中的错误率都保持在可接受的 2% 以下。

解读我们的研究结果时，务必牢记我们使用的是免费层服务。尽管这些结果为小规模或试点部署提供了有益的见解，但在较高服务层级下直接比较平台即服务与平台即服务、无服务器与无服务器产品，可能会得出不同结果。未来的研究应纳入此类比较，从而更全面地了解企业级部署。

5.2 成本因素

根据测试期间实际发生的 API 调用和传输数据进行的成本估算表明，对于所测试的工作负载，两个平台的运营成本都很低，尤其是在免费层限额内。Azure 较低的按请求计价结构是其在基准负载下成本低 71.9% 的主要原因。对于流量不规律或难以预测的应用程序，GCP 的按请求计费模式可能尤其划算，因为用户只需为处理请求所花费的实际时间付费。对于运行规律且流量较大的工作负载，Azure App Service 的固定费率定价或许更具可预测性。

5.3 平台选择指南

我们的研究结果支持以下实用建议：

如果您符合以下情况，请选择 GCP Cloud Run：(1) 有容器化部署和 Google Cloud 生态系统方面的经验；(2) 重视精细化扩缩容和按请求计费；或 (3) 工作负载变化很大，且流量容易出现不可预测的激增。

如果您需要始终在线且不会发生冷启动，工作负载稳定且可预测，或者使用微软的企业工具和服务，请选择 Azure App Service。

充分利用各平台优势，将不同平台用于特定组件的多云策略，可能有利于需求各异的企业。

5.4 限制

本研究存在若干局限。首先，工作负载是人为构造的，尽管具有代表性，仍可能无法准确反映真实销售点流量的细微特征。其次，由于所有测试都在同一地理区域进行，其他地区的结果可能会因网络延迟或区域容量而有所不同。第三，我们的部署采用了技术含量较低、较为简单的架构，未使用缓存或 CDN 集成等可进一步优化成本或性能的高级功能。最后，由于云服务发展迅速，这些结果仅反映了特定时间点的情况。

尽管存在这些局限，我们开源、由代码驱动的方法为研究人员和零售商提供了透明且可复现的框架，使他们能够使用自己的工作负载和配置来评估云销售点系统的性能和成本。

6 结论

本研究使用免费层资源和开源基准测试脚本，对 Google Cloud Platform Cloud Run 与微软 Azure App Service 在销售点（POS）工作负载下的表现进行了透明、可重复的比较。研究人员和零售商可以采用这种方法，根据吞吐量、错误率、实时性能和估算运营成本，对云平台进行实证评估。所有结果均来自代码输出。

我们的结果表明，在所测试的规模下，GCP 和 Azure 均能保持可扩展的吞吐量和较低的错误率，并可靠地处理销售点工作负载。对于流量高度波动或难以预测的零售商，GCP Cloud Run 尤为适用，因为它在压力条件下表现出更高的吞吐量（70.84 TPS 对比 20.14 TPS），且在基准负载下响应时间快 23.0%。另一方面，对于流量模式可预测且运营成本敏感的零售商，Azure App Service 的运营成本低 71.9%，并且在稳定负载下响应时间更加稳定。

本研究创建的开源、代码驱动框架可供任何零售商复制这些评估、调整工作负载模型以适应自身情况，并就云采用作出明智决策。通过使用自动化测量和免费层资源，各种规模的企业都可以采用这一策略。

未来的研究将评估缓存、自动扩缩容和 CDN 集成等高级功能；还将研究面向大型零售商的多区域部署；并将扩展这一基准测试框架，以纳入其他云服务提供商，例如亚马逊云科技。未来的研究将重点关注随着服务变化而实现持续平台评估的自动化，与行业合作伙伴合作纳入真实的生产工作负载，并拓展方法，以涵盖安全、合规以及混合云或多云策略。

随着零售业持续推进数字化转型，易于使用且可重复的基准测试工具对于最大化技术投资至关重要。本研究通过构建持续开展数据驱动云平台评估的框架，帮助零售商在快速变化的环境中平衡成本、性能和运营韧性。

参考文献

1. Bermbach, D., Wittern, E., Tai, S.：《云服务基准测试：从客户端视角衡量云服务质量》。Springer（2017） 2. Chen, L., Zhang, Y., Liu, W.：季节性零售工作负载的成本优化框架。《IEEE 服务计算汇刊》**14**(5), 1423–1436（2021） 3. Garg, S.K., Versteeg, S., Buyya, R.：云计算服务排名框架。《未来一代计算机系统》**29**(4), 1012–1023（2013） 4. 谷歌云：Cloud Run 定价。https://cloud.google.com/run/pricing。访问日期：2025年6月29日 5. 谷歌云：VPC 网络定价。https://cloud.google.com/vpc/ network-pricing#internet egress。访问日期：2025年6月29日 6. Johnson, M., Wang, L.：零售业的云采用：综合调查。《国际零售与分销管理期刊》**51**(3), 321–339（2023） 7. Li, A., Yang, X., Kandula, S., Zhang, M.：CloudCmp：比较公有云服务提供商。载于：第10届 ACM SIGCOMM 互联网测量会议论文集，第1–14页。澳大利亚墨尔本（2010） 8. Manner, J., Endreß, M., Heckel, T., Wirtz, G.：影响函数即服务冷启动的因素。载于：IEEE/ACM 国际效用计算与云计算会议伴随会议论文集，第181–188页。苏黎世（2018） 9. Mazrekaj, A., Shabani, I., Sejdiu, B.：云计算中的定价方案：概述。《国际高级计算机科学与应用期刊》**7**(2), 80–86（2016） 10. 微软 Azure：App Service 定价。https://azure.microsoft.com/en-us/pricing/ details/app-service/。访问日期：2025年6月29日 11. 微软 Azure：带宽定价。https://azure.microsoft.com/en-us/pricing/ details/bandwidth/。访问日期：2025年6月29日 12. 美国全国零售联合会：《零售技术现状报告》。技术报告，NRF，华盛顿特区（2023） 13. Thompson, K., Miller, D., Anderson, J.：基于云的 POS 系统安全考量。IEEE《安全与隐私》**20**(4), 67–75（2022） 14. Varghese, B., Akgun, O., Miguel, I., Thai, L., Barker, A.：面向科学应用性能最大化的云基准测试。《IEEE 云计算汇刊》**7**(1), 170–182（2019）