

基于碳意识的微服务部署以预算内实现最优用户体验

Kevin Kreutz*

kreutz@campus.tu-berlin.de

Technical University of Berlin

Berlin, Germany

Philipp Wiesner*

wiesner@tu-berlin.de

Technical University of Berlin

Berlin, Germany

Monica Vitali*

monica.vitali@polimi.it

Politecnico di Milano

Milan, Italy

摘要

数据中心的碳足迹最近已成为一个关键问题。到目前为止，大多数碳意识策略都集中在利用调度决策的灵活性来转移批处理工作负载的时间和地点。然而，这些方法不能应用于面向服务的云计算应用程序，因为它们必须在任何时间点都能访问，并且通常需要低延迟。

我们提出了一种碳意识的方法，用于在每小时碳预算下操作微服务。通过为每个微服务选择最合适的版本和水平扩展，我们的策略能够在遵守预算约束的同时最大化用户体验和收益。各种应用程序配置和碳预算下的实验表明，该方法能够适应不断变化的工作负载和碳强度。

Keywords

绿色应用、服务计算、能源效率

ACM Reference Format:

Kevin Kreutz, Philipp Wiesner, and Monica Vitali. 2025. 基于碳意识的微服务部署以预算内实现最优用户体验. In *Proceedings of 1st International Workshop on Low Carbon Computing (LOCO 2024)*. ACM, New York, NY, USA, 4 pages.

1 介绍

像欧盟这样的管理机构正在启动多项行动，旨在到2050年实现碳中和[6]。ICT部门在这项具有挑战性的任务

中扮演着重要角色，贡献了全球总排放量的3.9%，并且这一趋势在逐年增加[3]。为此，欧洲绿色数字联盟(EGDC)已让多家公司参与联合努力，以衡量并减少ICT行业的环境影响，并设定了到2040年实现碳中和的目标[4]。类似地，欧盟排放交易系统(EU ETS)[5]通过限定特定行业允许的总排放量来促进碳中和目标。

云服务提供商宣布了碳中和目标[1, 7, 11]。尽管基础设施层面的能源效率有所提高，但由于托管应用程序对计算能力需求的增长，云计算排放量仍在持续上升。因此，减少这些应用程序的环境影响非常重要。为此，通过增加自适应特性来增强应用程序[10, 15]可以在需要时帮助降低其排放。

在所提出的方案中，我们将每小时碳排放预算或基于时间的可再生能源证书[9, 14]应用于微服务架构。我们提出了一种能够根据当前预算、电力电网的能源结构以及工作负载持续重新配置微服务部署的算法。我们证明了我们的方法能够在满足预算约束的同时最大化用户体验和应用程序所有者的收益。

2 方法

所提出的方案在长期运行的云应用部署过程中包含了碳意识，利用灵活性[2, 8, 12, 13, 17, 18]根据当前电力电网的碳强度来调整应用程序的部署和扩展。

该方法的概述如Figure 1所示。其核心是一种优化算法，用于在避免超出碳预算的同时，选择应用程序的最佳配置并优化其他非功能性需求，例如用户体验质量(QoE)和应用程序所有者的收入。在本节中，

*All authors contributed equally to this research.

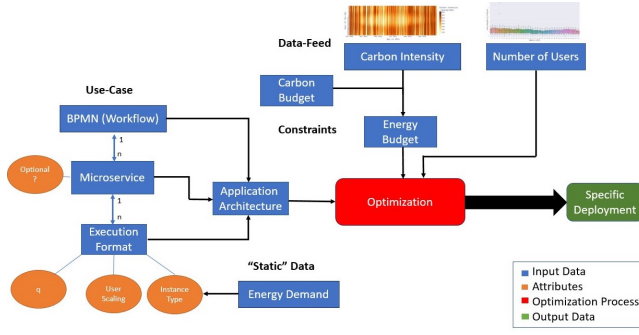


图 1: 方法概述

我们将提供更多关于该方法组件的详细信息，并对用于优化的目标函数进行建模。

应用架构. 本工作中考虑的应用类型是由多个微服务组成的长期运行的云原生应用。应用程序的工作流可以通过如 [15] 中讨论的 BPMN 进行建模和丰富。每个微服务可以提供多种版本，每种版本在能耗需求和向用户提供的体验质量方面具有不同的特性。例如，一个低功耗版本的微服务可能会移除一些计算密集型功能，从而影响用户的整体满意度。微服务也可以被标注为可选。可选的微服务不是达到应用程序总体目标所必需的。然而，它们可以提高用户的 QoE 并/或为应用程序所有者生成奖励（例如，广告）。

能量预算. 它表达了与碳相关的包括分配给应用程序的碳预算和能源电网的碳强度的信息。我们假设碳预算是按年分配的，但会根据预期的工作负载动态地每小时拆分。碳强度是当前电力网络中绿色能源和棕色能源的比例混合。此值可用于计算在时间槽内可以使用多少能量而不会超过碳预算。

用户数量. 该应用程序可以根据前几年收集的信息与一年中的估计请求数量（工作负载）相关联。

目标函数可以表示为公式 1

$$\max \left(\alpha \sum_{ms \in CA} \frac{QoE_{ms}}{\#ms} + \beta \sum_{ms \in CA} \frac{rev_{ms}}{revMax} \right), \quad (1)$$

其中 QoE_{ms} 是与微服务版本相关的 QoE, rev_{ms} 是与微服务相关的应用程序所有者的收入, α 和 β 是两个

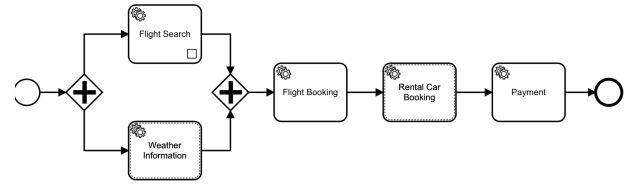


图 2: 航班预订应用程序在 BPMN 中

微服务	版本	可选	实例类型	教育	q	单位补码	用户体验质量	修订
航班搜索	Low Power	不	t3.micro	13.0	0.5	20000 or 5000	0.5 or 0.1	0
	Normal		t3.xlarge	39.9	0.7	20000 or 5000	0.75 or 0.3	0
	High Performance		g2.2xlarge	305.4	0.9	20000 or 5000	1	0
天气信息	Off	是	t3.micro	13.0	0.9	20000 or 5000	0	0
	Normal		t3.micro	13.0	1.0	20000 or 5000	1	0.2
航班预订	Low Power	不	t3.micro	13.0	0.5	20000 or 5000	0.5 or 0.1	0
	Normal		t3.xlarge	39.9	0.95	20000 or 5000	1	0
	Off		t3.micro	13.0	0.5	20000 or 5000	0.5 or 0.1	0
租车预订	Low Power	是	t3.micro	13.0	0.5	20000 or 5000	0.5 or 0.1	0
	Normal		t3.xlarge	39.9	0.9	20000 or 5000	0.75 or 0.3	2
	High Performance		g2.2xlarge	305.4	1.0	20000 or 5000	1	2
Payment	Normal	no	t3.xlarge	39.9	0.99	20000 or 5000	1	0

表 1: 航班预订微服务的注解

与这些方面对应用程序所有者相关性的权重。优化算法的完整实现对公众开放¹。

3 实验与结果

我们通过一个简单的应用验证了该方法，如 Figure 2 所示，使用了增强的 BPMN [16]。该应用由五个微服务组成。其中两个（天气信息和租车预订）是可选的。而航班搜索微服务则提供了多个版本。这五个微服务的具体注释总结在 Table 1 中，包括版本和可选标签。对于每个版本，都指定了部署所需的实例类型以及估计的能量需求 (ED)。每个版本都与一项收入 (rev, 仅适用于可选功能) 和一个 QoE 值相关联。参数 uc 表示单个实例处理的用户请求数量，而 q 表示继续进入下一个微服务的用户的百分比。在实验中，大多数参数保持不变，而改变了 uc 和 QoE 的值。

提出的算法（最优选择 (OS)）旨在为每个微服务选择最佳配置，确保在满足预算约束的同时优化目标函数 (Equation 1)。它与三个基线进行了比较：(i) 高性能 (HP) - 所有微服务都部署了高性能版本，并且可选的微服务默认被部署；(ii) 顺序碳意识 (SCA) - 应用程序根据可用预算使用三种固定的顺序配置进行调整；

¹<https://github.com/POLIMIGreenISE/carbonbudget.git>

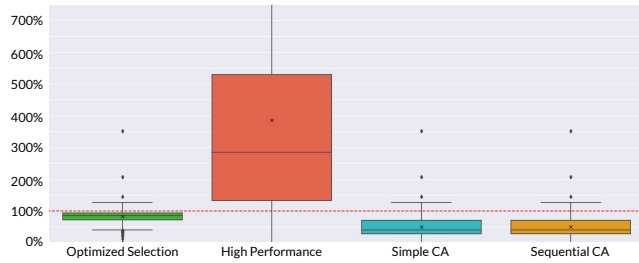


图 3: 碳预算利用

以及 (iii) 简单的碳意识 (CA) - 与单倍群 SCA 类似, 但三个配置不必是顺序的。

实验使用了不同的碳预算进行。例如, Figure 3 显示了在一年中将碳预算设定为应用在高性能模式下运行所需的排放量与低功耗模式所需排放量之间的结果。可以看到, HP 基线显著违反了碳预算, 而所有其他解决方案都能满足预算限制。所提出的方法有效地最大化了预算利用, 从而带来了更高的收入和 QoE (服务质量)。相比于碳感知基线, 操作系统算法平均提高了 5.44% 的 QoE, 并增加了 32.38% 的收入。

4 结论与未来工作

我们提出了一种碳感知方法, 根据可再生能源的可用性、碳预算和当前工作负载来调整云原生应用程序的操作, 同时最大化 QoE 和收入。我们的方法在动态工作负载和波动的能源混合下进行了测试, 展示了其满足碳预算并最大化利用率的能力, 从而提高 QoE 和收入。在未来的工作中, 我们计划探索额外的配置, 并实验目标函数中收入和 QoE 的不同权重。

Acknowledgments

本工作得到了 FREEDA 项目的资助, 该项目由框架 PRIN (意大利教育部、大学与研究部) 和欧盟下一代 (CUP: I53D23003550006) 提供资金支持, 并且还获得了意大利高性能计算、大数据和量子计算研究中心 (ICSC) 的 “FutureHPC & BigData” 分支的资金支持, 该中心由 MUR Missione 4 - 欧盟下一代 (NGEU) 资助。

References

- [1] Amazon. 2023. Amazon Sustainability Report. <https://sustainability.aboutamazon.com/2023-amazon-sustainability-report.pdf>
- [2] Tayebah Bahreini, Asser Tantawi, and Alaa Youssef. 2022. An Approximation Algorithm for Minimizing the Cloud Carbon Footprint through Workload Scheduling. In *2022 IEEE 15th International Conference on Cloud Computing (CLOUD)*. 522–531. <https://doi.org/10.1109/CLOUD55607.2022.00075>
- [3] European Commission. 2024. Rolling Plan for ICT standardisation. <https://joinup.ec.europa.eu/collection/rolling-plan-ict-standardisation/ict-environmental-impact-rp2024>
- [4] Ministerial Declaration. 2021. A Green and Digital Transformation of the EU.
- [5] Ministerial Declaration. 2023. European Trading System Directive. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02003L0087-20230605>
- [6] Constanze Fetting. 2020. The European green deal. *ESDN Report, December 2, 9 (2020)*.
- [7] Google. 2024. Environmental Report. <https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf>
- [8] Conor Kelly, Eleni Mangina, and Antonio Ruzelli. 2011. Putting a CO2 figure on a piece of computation. In *11th International Conference on Electrical Power Quality and Utilisation*. 1–7. <https://doi.org/10.1109/EPQU.2011.6128960>
- [9] Michael Kelly and David Gonzalez. 2023. *The Role of Hourly EAC Markets in Facilitating the Clean Energy Transition*. Technical Report. Google.
- [10] Justin J Meza, Thote Gowda, Ahmed Eid, Tomiwa Ijaware, Dmitry Chernyshev, Yi Yu, Md Nazim Uddin, Rohan Das, Chad Nachiappan, Sari Tran, et al. 2023. Defcon: Preventing Overload with Graceful Feature Degradation. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*. 607–622.
- [11] Microsoft. 2024. How can we advance sustainability? 2024 Environmental Sustainability Report. <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lMjE>
- [12] Tobias Piontek, Kawsar Haghshenas, and Marco Aiello. 2023. Carbon emission-aware job scheduling for Kubernetes deployments. *The Journal of Supercomputing* (06 2023), 1–21. <https://doi.org/10.1007/s11227-023-05506-7>
- [13] Thanathorn Sukprasert, Abel Souza, Noman Bashir, David E Irwin, and Prashant J Shenoy. 2023. Quantifying the benefits of carbon-aware temporal and spatial workload shifting in the cloud. *CoRR* (2023).
- [14] Maud Texier. 2022. Timely progress towards around-the-clock carbon-free energy. Google. <https://cloud.google.com/blog/topics/sustainability/t-eacs-help-drive-around-the-clock-carbon-free-energy>
- [15] Monica Vitali. 2022. Towards greener applications: enabling sustainable-aware cloud native applications design. In *International Conference on Advanced Information Systems Engineering*. Springer, 93–108.
- [16] Monica Vitali, Paul Schmiedmayer, and Valentin Bootz. 2023. Enriching Cloud-native Applications with Sustainability Features. In *2023 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 21–31.
- [17] Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. 2021. Let’s Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud. In *22nd International*

Middleware Conference. ACM. <https://doi.org/10.1145/3464298.3493399>

- [18] Philipp Wiesner, Dennis Grinwald, Philipp Weiß, Patrick Wilhelm, Ramin Khalili, and Odej Kao. 2024. Quality Time: Carbon-Aware Quality Adaptation for Energy-Intensive Services. *arXiv preprint* (2024).