arXiv:2304.00404v1 [cs.DC] 1 Apr 2023

GreenScale：面向边缘计算的碳感知系统

Young Geun Kim[§*] Udit Gupta[δ] Andrew McCrabb[‡] Yonglak Son[§] Valeria Bertacco[‡] David Brooks[†δ] Carole-Jean Wu[δ]

[§]高丽大学 [†]哈佛大学 [‡]密歇根大学 [δ]Meta

摘要

为减轻计算需求不断增长带来的环境影响，未来的应用程序需要提升计算基础设施的碳效率。然而，最先进的方法并未考虑可再生能源的间歇性。在决定如何执行计算时，人们忽视了为计算供能的碳强度随时间 和位置而变化这一特点。这带来了新的挑战——决定何时 以及何地 在边缘的消费设备和云端服务器上运行应用程序。随机运行时方差以及日益增加的隐含排放的摊销，使这些调度决策更加复杂。本研究提出*GreenScale*，这是一个用于理解边缘云基础设施中绿色应用碳感知调度设计与优化空间的框架。基于对基础设施组件碳排放量的量化，我们证明，与优化性能和能效相比，以碳排放为优化目标会产生独特的调度方案。我们对三类具有代表性的应用程序（即人工智能、游戏和增强现实/虚拟现实）进行评估，结果表明，借助*GreenScale*，应用程序的碳排放量最多可减少29.1%；按边缘云应用用户规模（数百万计）估算，年均可减少232.7tCO_2, 碳排放量，相当于55辆汽车的平均年碳排放量。本研究还进一步提供了详细路线图，帮助边缘云应用开发者构建绿色应用。

1. 引言

随着信息与通信技术（ICT）的飞速发展，人工智能（AI）、扩展现实（XR）和加密货币等各种新型应用应运而生 [3,8,43,47,83]。尽管这些应用带来了诸多益处，信息与通信技术也在全球范围内造成了显著的能耗和环境负担。2022年，信息与通信技术产生的碳排放占全球碳排放的3% [5,51],，与航空业的碳排放相当。预计未来十年，这一比例甚至将升至全球排放量的8% [50]。因此，迫切需要设计可持续的绿色应用，将碳排放降至最低。

要设计绿色应用，开发者需要提高基础设施组件的碳效率。大规模计算基础设施已通过大规模计算基础设施设计 [9,44,81],得到了显著优化

*通讯请联系 Young Geun Kim *younggeun_kim@korea.ac.kr*

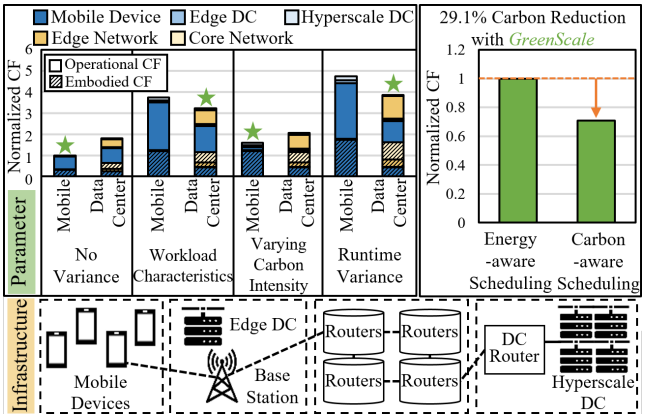


图1：边缘云基础设施的总体碳排放会随工作负载特征、碳强度变化和运行时方差而显著波动。通过审慎地为使用者选择执行目标，借助*GreenScale*，碳排放最多可显著减少29.1%。

运行效率优化，例如微处理器能效和电源使用效率（PUE），来自工业界 [41, 45, 80, 82] 和学术界 [74, 96, 134]。进一步提升运行能效正变得越来越具有挑战性 [13, 32, 53, 65, 69, 73, 102, 116, 117,118, 127, 132]。除了能效优化之外，太阳能和风能等可再生能源也越来越多地被采用，以降低计算的碳足迹（CF）[2, 54, 106, 114]。遗憾的是，可再生能源发电具有间歇性，无法在任何单一地点始终持续供应 [2, 106]。由于可再生能源供应间歇性，边缘客户端设备和大规模数据中心基础设施所处计算范围内的计算碳强度会随地点和使用时间而变化。

一个新的挑战随之而来——为了最大限度地减少计算产生的碳排放，需要决定在哪里运行应用以及何时运行。现代计算基础设施支持通过自动扩缩容灵活地执行计算。例如，许多个性化人工智能和娱乐用例都依托于由智能手机和云组成的协同执行环境 [35, 56, 94, 121]。此外，虚拟现实和增强现实系统可以由可穿戴电子设备、作为中转设备的智能手机以及云组成 [42, 46, 84, 92]。计算可以在许多不同节点上执行。然而，对于任何应用来说，决策过程都颇具挑战性，因为边缘云基础设施中每个执行目标的碳强度可能会有显著差异

这取决于计算执行目标在何处以及何时充电或供电，也取决于哪些能源发电来源。

要降低应用程序的碳排放、隐含排放和运行时方差，绿色应用的设计面临着额外挑战。近期对现代硬件碳足迹的研究表明，隐含碳排放（即制造处理器、内存和存储所产生的排放）正开始占据计算碳足迹的主要部分，[6,21,48,50,51],这归因于集成电路制造和数据中心建设的高昂开销。尽管应用程序无法直接改变隐含碳排放，但在应用运行期间，用户共同使用系统可以分摊隐含碳排放。因此，为最大限度地提高碳效率，应用程序在基础设施组件之间分配用户时，必须同时考虑运行和隐含排放。此外，移动执行本质上具有随机性 [37,38,72]。碳效率波动可能源于应用程序之间及应用程序内部的干扰 [123] 以及网络的稳定性 [22]。遗憾的是，最先进的方法，例如 [13,32,53,54,65,66,68,69,73,88,102,114,116,117,118,127,130,132],确定执行目标时，主要依赖性能和能耗等运行特征，而没有考虑上述特性（即依赖位置和时间的可再生能源可用性、隐含排放和运行时方差），因此仍有很大的改进空间（图 1）。这是由于缺乏一种能够量化和分析基础设施组件碳排放的工具，并将上述特性纳入考量；同时，系统栈之间的接口异构，以及基础设施组件由多个组织管理等问题，也带来了各种挑战。

为了为开发者构建绿色应用提供有价值的指导，本文提出了*GreenScale* ——一个碳排放设计空间探索与优化框架。*GreenScale* 通过考虑应用工作负载特征、电网中因地点和时间而异的可再生能源可用性、隐含排放的摊销 [25,120], 以及运行时方差 [50], , 量化边缘云调度领域的碳排放情况。基于量化后的碳排放，*GreenScale* 针对三类重要的边缘云应用——人工智能、游戏和增强现实/虚拟现实——在不同运行时环境中探索碳感知调度的设计空间。

这项工作的核心贡献包括：

- 我们构建了一个框架*GreenScale*，用于建模并量化碳感知绿色应用的设计空间。*GreenScale* 在考虑应用工作负载特征、因地点和时间而异的可再生能源可用性、隐含排放摊销以及运行时方差的影响后，实现了以最小环境足迹进行计算扩展（第 3 节）。

- 通过使用 *GreenScale*，我们探索了适用于典型边缘云应用类别的碳感知调度设计空间。基于特征分析结果，我们证明，碳感知调度 不同于传统的性能和/或能耗

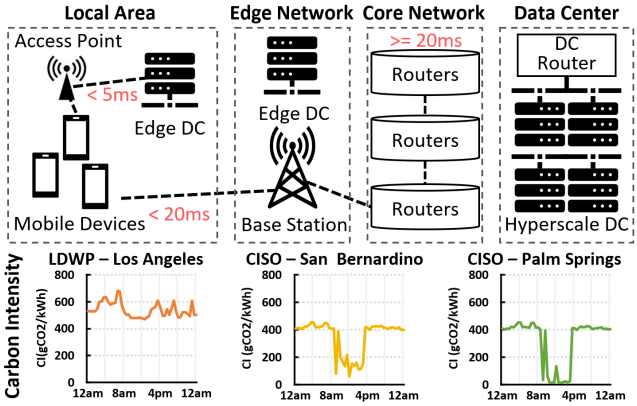


图 2：边缘云基础设施示例

感知调度，因此碳排放改善仍有很大空间（第 5 节）。

- 我们还提炼了绿色应-用程序开发的关键见解——为开发人-员提供指导，帮助他们在应用程序设计和系统基础设施参数（例如调度方法和资源配置，见第 5.4 节）方面做出碳知情决策，从而最大限度地减少应用程序的碳影响。

2. 边缘云基础设施

面向边缘客户端设备运行的应用程序，可以利用网络层级中的分布式异构资源，从网络边缘一直延伸到多跳远程云 [54,66,88,114,130]。图 2 展示了边缘云基础设施的一个示例。

边缘设备 是位于网络边缘的计算资源（即图 2 中的本地区域）。边缘设备的示例包括智能手机、智能手表、笔记本电脑、家用电器等。传统上，在许多移动服务中，边缘设备主要用作用户端传感器、用户界面，或兼具两者。近年来，随着高性能移动片上系统 (SoCs) 的发展 [52,58,118],，边缘设备现在可以在本地执行不同规模的计算 [102,116,117,118,122,127,132]。这样做可以缩短服务的响应时间，并消除数据传输开销 [13,32,53,65]。

边缘网络 将多用户接入网络中的边缘设备连接到另一个接入网络，或高速核心网络 [49,124,125]。对于移动边缘云环境，多用户接入网络可以是 Wi-Fi接入点网络（即图 2 中本地区域内的接入点），也可以是无线蜂窝网络（即图 2 中的边缘网络）。在典型的蜂窝网络中，每个小区内都会部署不同规模（例如宏基站、小基站和微微基站）的基站（BS），充当无线收发器 [124]。

边缘数据中心是指与边缘网络共同部署的小型数据中心（DC） [36,100,101]。如图 2 所示，边缘数据中心可以位于以下两类位置之一：1）城市区域（例如，由开发者或应用公司运营的内部小型服务器 [18],，或边缘规模云

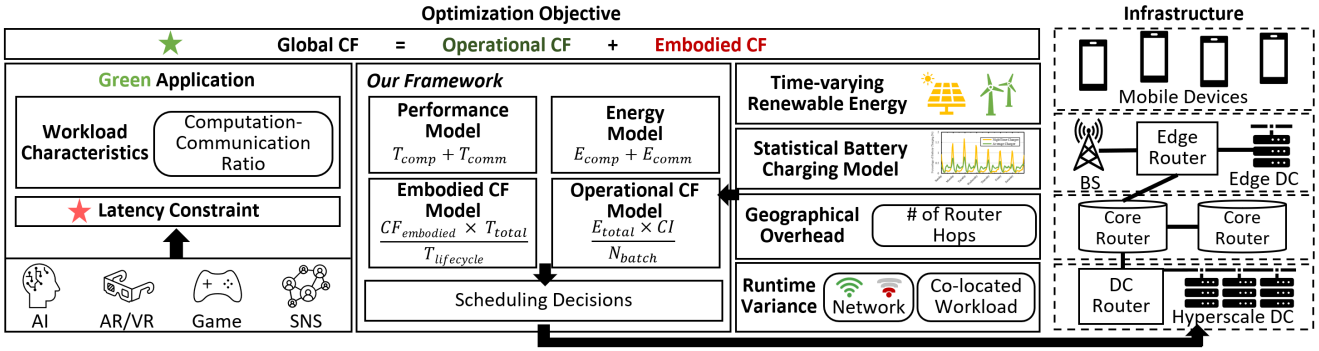


图 3：碳感知绿色应用的设计空间概览。

从云服务提供商处租用的服务器 [1,23,115]）；或 2）与基站相邻的农村地区 [31]。由于用户端设备只需经过边缘网络，边缘数据中心能够提供相对强大的计算资源，同时将数据传输开销控制在可接受范围内（如图 2 所示，根据边缘数据中心的位置不同，延迟为 5 毫秒至 20 毫秒）。然而，由于空间和可扩展性限制 [36,66],，边缘数据中心中的服务器通常比超大规模数据中心中的服务器具有更低的计算能力和效率 [18]。

核心网络 包括大容量核心路由器和高速光纤电缆，它们将消费者和企业连接到数据中心 [19,20,61,126]。来自边缘网络的数据包被传送至核心路由器，在路由器之间逐一转发，最终送达包含目标地址的另一个边缘网络。

数据中心 指通常由工业公司运营的大型数据中心 [4,43,85]。数据中心通过利用大量高效协处理器（例如图形处理器和张量处理单元（TPU）），在处理大型工作负载和共置（即批处理）方面受益于规模经济。然而，与边缘数据中心相比，用户端设备需要经过边缘网络以及核心网络中的多个路由器跳点，因此数据中心通常会产生更高的数据传输开销 [54,66,88,114,130]。

3. GREENSCALE

本节提出 *GreenScale*，这是一个碳设计空间探索与优化框架，旨在支持绿色应用开发。图 3 展示了碳感知边缘云调度的整体设计空间。该设计空间的优化目标是最大限度地减少应用程序的碳排放，同时满足延迟约束，并可将计算内核灵活地调度到边缘云计算范围内的各类处理器上。为实现这一优化目标，*GreenScale* 根据不同应用工作负载下的性能和能耗测量数据（并考虑运行时方差）、为基础设施组件供能的所有美国电网的每小时能源发电数据，以及隐含碳足迹建模工具，量化计算内核对应的基础设施组件的碳排放。基于量化后的碳排放，*GreenScale* 在不同环境中全面探索碳排放最优的调度决策。

为量化基础设施组件的碳排放，*GreenScale*使用四种模型：1）执行时间性能模型，2）能耗模型，3）运营碳足迹（CF）模型，以及 4）隐含碳足迹模型。其中，性能模型和能耗模型是根据不同应用工作负载下的测量结果，并考虑运行时方差得出的。运营碳足迹模型根据对每个基础设施组件能耗和碳强度的估算计算得出 [6,7,17,50,51,60,105,108,113]。隐含碳足迹模型采用了两种隐含碳足迹建模工具：ACT [50] 和生命周期评估（LCA） [7,60,105,108,113]。第 3.1 节将详细介绍这些碳足迹模型。

3.1 碳排放模型

正如许多先前的研究所探讨的那样，硬件生命周期中的碳排放可分为四个主要阶段：硬件制造、硬件运输、运行使用，以及报废处理与回收。在各阶段的排放中，*GreenScale* 考虑运行碳排放以及隐含碳排放（即硬件生命周期其余阶段产生的排放）。表 1 总结了基础设施组件的碳排放模型，表 2 则介绍了模型中使用的缩写和符号。

运行碳排放：每个执行目标的运行碳排放根据该执行目标所涉及组件的运行能耗及其各自的碳强度（表 I 中的碳强度 ）计算 [2,50,51,106]。

运行能耗包括三个组成部分 [70,71,129]：1）计算能耗；2）通信能耗；3）空闲能耗。计算能耗由两部分构成：在实际执行目标（移动设备、边缘数据中心或超大规模数据中心）上执行所需计算任务的执行时间（表 I 中的 T_{comp} ），以及执行期间消耗的功耗（表 I 中的 P_{comp} ）。通信能耗由两部分构成：将数据从用户端设备传输到执行目标所需的时间（表 I 中的T_{comm} ），以及各组件在数据传输期间消耗的功耗（表 I 中的P_{comm} ）。空闲能耗是未参与执行的执行目标所消耗的空闲能耗。空闲能耗开销根据应用程序运行时间内消耗的空闲功率（P_{idle}）计算，并除以每个组件的用户数

表 1：基础设施的碳排放模型

执行目标	组件	运行碳足迹	隐含碳足迹
移动设备	移动设备	$T_{comp_M} \times P_{comp_M} \times CI_M$	$ECF^M \times \frac{Tcomp_M}{\text{移动设备延迟阈值}}$
	边缘网络	-	-
	边缘数据中心	$\frac{Tcomp_M \times Pidle_E_\text{数据中心}}{Nuser_E} \times CI_E$	$\frac{ECFE_\text{数据中心}}{Nuser_E} \times \frac{Tcomp_M}{LTE_\text{数据中心}}$
	核心网络	-	-
	超大规模数据中心	$\frac{T_{\text{补偿}_M} \times P_{\text{空闲}_H}}{N_{\text{用户}_\text{数据中心}} \times CI_H}$	$\frac{ECF_H}{N_{\text{用户}_\text{数据中心}}} \times \frac{T_{\text{补偿}_M}}{LT_H}$
边缘数据中心	移动设备	$(T_{comm_E} \times P_{comm_M} + T_{comp_E_DC} \times Pidle_M) \times CI_M$	$ECF_M \times \frac{(Tcomm_E + Tcomp_E_DC)}{\text{移动设备延迟阈值}}$
	边缘网络	$\frac{T_{\text{通信}_E} \times P_{\text{计算}_\text{基站}} N_{\text{用户}}}{_\text{基站}} \times CI_E$	$\frac{\text{基站能耗系数}}{Nuser_\text{基站}} \times \frac{Tcomm_E}{LT_{BS}}$
	边缘数据中心	$\frac{T_{\text{压缩}_E_\text{数据中心}} \times P_{\text{压缩}_E_\text{数据中心}}}{N_{\text{用户}_E}} \times CI_E$	$\frac{ECF_\text{数据中心}}{Nuser_B} \times \frac{Tcomp_E_\text{数据中心}}{LTE_\text{数据中心}}$
	核心网络	-	-
	超大规模数据中心	$\frac{(Tcomm_E + Tcomp_E_\text{数据中心}) \times Pidle_H}{Nuser_\text{数据中心}} \times CI_H$	$\frac{ECF_H}{Nuser_\text{数据中心}} \times \frac{(Tcomm_E + Tcomp_E_\text{数据中心})}{LT_H}$
超大规模数据中心	移动设备	$(T_{comm_E} \times P_{comm_M} + (T_{comm_R} + Tcomp_H) \times Pidle_M) \times CI_M$	$ECF_M \times \frac{(\text{通信_能耗} + \text{通信_速率} + \text{计算_高})}{\text{移动设备延迟阈值}}$
	边缘网络	$\frac{T_{\text{通信_能耗}} \times P_{\text{计算_基站}}}{N_{\text{用户_基站}}} \times CI_E$	$\frac{\text{基站能耗系数}}{N\text{用户_基站}} \times \frac{T\text{通信_能耗}}{LT_{BS}}$
	边缘数据中心	$(T_{comm_E} + T_{comm_R} + T_{comp_H}) \times P_{idle_E_\text{数据中心}} N_{user_E} \times CI_E$	$\frac{ECFE_\text{数据中心}}{Nuser_E} \times \frac{(Tcomm_E + Tcomm_R + Tcomp_H)}{LTE_\text{数据中心}}$
	核心网络	$\frac{T_{\text{通信}_R} \times P_{\text{通信}_R} N_{\text{用户}_R}}{} \times CI_R$	$\frac{ECF_R}{N\text{用户_}R} \times \frac{T\text{通信_核心}}{LT_R}$
	超大规模数据中心	$\frac{T\text{计算_}H \times P\text{计算_}H}{N_B} \times CI_H$	$\frac{ECF_H}{N_B} \times \frac{Tcomp_H}{LT_H}$

（$Nuser$），假设应用开发者（或公司）与用户共同承担应用程序所使用计算组件产生的碳排放。

根据执行目标涉及的组件不同，执行目标的运行能耗也会有所不同。例如，当移动设备被视为执行目标时，移动设备的计算能耗以及数据中心的空闲能耗开销都会计入运行能耗。另一方面，当数据中心被视为执行目标时，客户端设备和网络组件（即基站或核心路由器）的通信能耗、数据中心的计算能耗，以及客户端设备的空闲能耗开销都会计入运行能耗。

每个组件的碳强度取决于为其供能的能源来源。各能源来源的碳强度汇总于表 3 [2,50,51,106] 中。通常，组件由距离最近的电网供能 [2,106,120]。由于各电网中各能源来源的每小时发电量在很大程度上取决于地理位置，因此每个组件的碳强度最终由其所在地和时间共同决定。为准确建模各组件的碳强度，我们收集了美国电网 [25,120] 的每小时发电数据，并将其用于 *GreenScale*。

隐含碳排放：隐含碳排放包括硬件生命周期中除运行使用之外产生的排放。组件的隐含碳足迹根据其总隐含碳足迹（ECF）和生命周期时间（LT）、应用运行时间（T_{comp} 或 T_{comm}），以及在应用运行时间内共同使用该组件的用户数量（N_{user} 或 N_B）计算，假设应用运行时间（即延迟）内消耗的隐含碳足迹可按组件生命周期内共同使用该组件的用户数量进行折算。

组件。这里，每个组件的 ECF 可使用生命周期分析工具（LCAs）[7,60,105,108,113] 或架构碳模型工具（ACT）[50]获取。这些模型根据与硬件生命周期相关的多种参数计算隐含碳足迹，例如片上系统面积、单位面积能耗、单位面积气体排放、良率、原材料等。

3.2 设计空间参数

工作负载特征：碳排放最优的执行目标取决于工作负载特征。这是因为执行目标的运行效率取决于工作负载的计算通信比。计算能耗和延迟由工作负载所需的计算量决定（例如，人工智能工作负载的浮点运算次数），而通信能耗和延迟则由传输数据的大小决定（例如，人工智能工作负载的输入图像或文本）。这里，能耗与基础设施组件的运行碳排放直接相关，延迟则与其隐含碳排放相关。

性能约束也取决于工作负载的特性。例如，对于人工智能工作负载，用户的服务质量（QoS）预期可以定义为特定的延迟（例如，视频帧率为每秒 30 帧时为 33.3 毫秒 [24,133],，或交互式应用程序为 50 毫秒 [26,78]），低于这一限值时，用户通常很难察觉差异。另一方面，对于游戏工作负载，则可以考虑更多类型的服务质量指标，例如延迟、每秒帧数、抖动等 [14,15,75],。类似的服务质量指标也适用于增强现实/虚拟现实工作负载 [131]。由于服务质量是移动端优化的一项关键指标，因此满足

表 2：缩写和符号

缩写和符号	描述
M	移动设备
E	Edge
DC	数据中心
H	超大规模数据中心
BS	基站
R	核心路由器
T_{comp}	计算时间
T_{comm}	通信时间
P_{comp}	计算功耗
P_{comm}	通信功耗
P_{idle}	空闲功耗
CI	碳强度
ECF	总隐含碳足迹
LT	生命周期时间
N_{user_DC}	每个数据中心的用户数
N_{user_edge}	边缘系统的用户数
N_{user_BS}	每个边缘基站的用户数
N_B	数据中心中的批处理用户数

依赖工作负载的性能约束。

碳强度的变化：影响碳排放最优执行目标的另一个重要因素，是组件所使用能源的碳强度。如表3所示，可再生能源（即风能和太阳能）的碳强度远低于煤炭等其他能源。由于碳强度存在差异，组件的运营排放在很大程度上取决于能源的可用性，进而影响碳排放最优执行目标。例如，对于运营排放高于移动系统的服务器级计算系统，使用可再生能源可通过降低运营排放的碳强度，大幅减少全球碳足迹。

然而，电网中各类能源的发电量在很大程度上取决于其所在位置和时间 [25, 120]。例如，对于加利福尼亚州的电网（即图4中的CISO），太阳能仅在白天可用，因此电网的碳强度只有在白天才较低。另一方面，对于纽约州的电网（即图4中的NYISO），风能的供应具有间歇性，因此电网的碳强度会在一天中波动。

由于能源来源的可用性会随时间和位置而变化，组件的实际碳强度也可能有所不同。对于边缘的客户端设备而言，碳强度取决于用户何时以及如何为其设备的电池充电 [106]。图4展示了两种不同用户的电池充电模型，以及加利福尼亚州电网（即 CISO）和纽约州电网（NYISO）的每小时碳强度：1）夜间充电器，通常在夜间为电池充电；2）平均充电用户，在一天中按需为电池充电。如图4所示，夜间充电用户的移动设备可能具有更高的碳强度，因为他们大多在碳强度较高时为电池充电。碳强度还可能取决于用户所在的位置。对于数据中心而言，碳强度可以由其供电电网的平均碳强度，或企业通过能够追踪可再生能源信用的复杂会计框架所购买的可再生能源数量来

表3：能源的运行碳强度

Type	gCO_2eq/千瓦时
Wind	11
太阳能	41
水	24
Oil	650
天然气	490
Coal	820
核能	12
其他（生物燃料等）	230

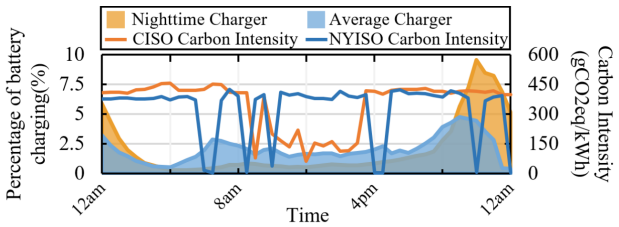


图4：根据用户为电池充电的时间不同，移动设备的碳强度也会有所变化；夜间充电用户主要在碳强度较高时为电池充电（黄色区域），而平均充电用户则几乎均匀地为电池充电（蓝色区域）。

确定。由于碳排放最优执行目标可能取决于组件的实际碳强度，因此绿色应用必须考虑随时间变化且依赖位置的碳强度。

地理权衡：考虑到可再生能源可用性随时间变化，也可以考虑由另一电网供电的其他执行目标。例如，为了降低碳强度，我们可以将任务卸载到距离更远的执行目标，即使这会导致更长的网络延迟（例如，选择农村地区可再生能源充足的边缘数据中心，而非图2中的移动设备或距离较近的边缘数据中心）。由于网络延迟与碳强度之间的权衡程度可能取决于传输数据量，因此必须仔细考虑这种地理权衡。

运行时方差：边缘云执行环境本质上具有随机性 [37, 38, 72]。在真实的执行环境中，不仅服务器规模的数据中心，移动设备上也可能运行多个共置工作负载——近期的移动设备支持多任务功能，例如同时共享多个应用程序的屏幕。共置工作负载带来的资源干扰会影响各组件的计算效率，从而改变碳排放最优执行目标。边缘网络中的无线信号强度变化以及核心网络中的拥塞所导致的网络波动，也会显著影响通信效率。例如，随着边缘设备用户移动，无线蜂窝网络的信号强度可能发生显著变化——用户在日常生活中会经历明显的信号强度波动（43% 的数据是在信号较弱时传输的 [22]）。此外，核心网络中的排队延迟也会随流量变化而动态变化。

表 4：神经网络推理工作负载

类别	NN	浮点运算次数	参数	输入/输出 (KB)
视觉	MobileNet	0.31G	3.5M	150.5
	SqueezeNet	0.82G	1.2M	150.5
	ResNet 50	4.09G	25.6M	150.5
	MobileNet-SSD	0.8G	6.8M	270
	Inception	5.71G	23.8M	268.2
Text	BERT	25.3G	17.5M	1

表 5：游戏工作负载

类别	Name	输入/输出 (MB)	FPS Req.	延迟请求（毫秒）
第一人称 Game	Fortnite	3.2	60	100
第三人称角色扮演 Game	原神 原神	3.0	60	500
Omnipresent Strategy	云顶之弈	1.9	60	1,000

表 6：增强现实/虚拟现实 工作负载

类别	Name	任务	输入/输出 (KB)	性能 Req. (ms)
3D 世界	Sponza	1) 输入 2) 感知 3) 视觉/音频	540.47	97.83
3D 材质	材质			
3D 卡通	平台跳跃游戏			
AR	增强现实演示			

4. 方法论

4.1 工作负载

为了了解 GreenScale 所面向的绿色应用程序的碳设计空间，*GreenScale*，我们在边缘云基础设施上运行了三种最先进的工作负载：人工智能（AI）、游戏和增强现实/虚拟现实（AR/VR）。我们重点研究人工智能工作负载，因为它们广泛应用于许多近期的智能服务和应用程序 [53, 79, 98, 119]。此外，我们还研究游戏应用，因为它们占据了 2022 [109]应用市场的 63.5%。我们也研究增强现实/虚拟现实应用，因为它们近期在消费者和研究社区中都获得了广泛关注 [131] 这得益于高效计算技术、高速通信和专用硬件平台的发展。

对于人工智能工作负载，我们运行最先进神经网络（NNs）的推理 [53,79,98,119]，这些网络广泛应用于移动智能服务 [35, 42, 46, 56, 84, 92, 94, 121]。表 4 总结了这些神经网络。如表 4 所示，神经网络的计算特性各不相同，具体取决于浮点运算（FLOPs）次数和参数规模。此外，不同类别的神经网络具有不同的输入/输出规模。不同类型的神经网络也有不同的性能要求。对于视觉神经网络，我们将 30 FPS 作为性能要求——只要每秒帧数超过30 [24, 133]，用户通常不会察觉到服务质量的差异。对于文本工作负载，我们将 100 毫秒作为性能要求 [98]。

对于游戏工作负载，我们运行表 5 总结的三种不同类型的游戏，这些游戏广泛应用于移动设备用户群体 [14, 15, 75]。如表 5 所示，不同类型的游戏有不同的性能要求。当执行目标为移动设备时，我们直接在移动设备上运行 Android 应用程序，同时测量设备的延迟、每秒帧数和功耗。另一方面，当执行目标为数据中心（DC）时，我们在移动设备上运行云游戏 Android 应用 NVIDIA Geforce Now [91], , 同时测量设备的延迟、每秒帧数和功耗。为了测量云服务器的功耗，我们还在数据中心基础设施上运行游戏的桌面版本。

对于 AR/VR 工作负载，我们使用 IL-LIXR 中的四种工作负载，如表 6 所示：VR - 3D World、VR - 3D Material、VR - 3D 卡通和 AR。AR/VR 工作负载包含四个子任务，其中包括：1）感知，读取传感器输入并了解当前周围环境；2）视觉，将虚拟信息与物理世界相结合并渲染最终帧；以及 3）音频，计算并播放音频。对于所有工作负载，输入均为摄像头数据和运动跟踪数据，输出则为要流式传输到客户端设备的计算帧。

4.2 边缘云基础设施

本研究所采用的边缘云基础设施包括：移动设备、边缘数据中心、数据中心、边缘网络和核心网络。对于人工智能和游戏工作负载，我们使用现成的安卓智能手机 Pixel 3 [48], 作为移动设备，该设备配备了骁龙 845系统级芯片 [95]。另一方面，对于 AR/VR 工作负载，我们使用配备 NVIDIA Volta图形处理器的 NVIDIA Jetson 设备，因为 AR/VR 工作负载无法在安卓设备上运行。设备规格汇总于表7。我们使用外部功率测量设备 Monsoon 功率监测仪 [90]测量设备的功耗。

对于边缘数据中心和超大规模数据中心，我们分别使用 AWS 的 p3.2xlarge 和 p4d.24xlarge 实例，其规格与 MLCommons [87,100,101]中使用的实例相近。详细规格见表8。

对于边缘网络的基站，我们使用热设计功耗为 1000W [49] 的宏基站。我们根据宏基站可服务的用户数量计算单用户功耗 [49]。对于核心路由器，我们采用市售路由器的典型功耗和带宽 [19, 20]。

4.3 GreenScale 参数

随时间变化且依赖位置的碳强度：正如我们在第 3.1 节中所述，组件的碳强度取决于它们由何时、何地以及如何 供电——

表7：移动设备规格

设备	CPU	GPU	加速器	RAM (GB)
Pixel 3	Cortex A75 2.4GHz	Adreno 630 0.7GHz	Hexagon 685	4
Jetson AGX Xavier	英伟达 Carmel 2.4GHz	英伟达 Volta 1.4GHz	NVDLA	64

表8：服务器规格

Type	CPU	加速器	RAM
p3.2xlarge	英特尔至强 E5-2686 v4	英伟达 Tesla V100	64GB
p4d.24xlarge	英特尔至强 P-8275CL	英伟达 A100 x8	1152GB

电。为探究组件实际的碳强度，我们结合美国电网的可再生能源发电小时报告，考虑合理的情景 [25, 120]。

对于移动设备，我们基于先前的用户研究 [34, 93, 103]，考虑用户充电行为的三种情景：1）夜间充电器仅在夜间为电池充电；2）平均充电用户在一天中均匀地为电池充电；3）智能充电器仅在附近电网有可再生能源时为电池充电。利用 [34, 93, 103]，中的统计模型，我们计算移动设备在不同地点的平均碳强度。

对于边缘数据中心和基站，我们根据其地点 [49, 124] 考虑两种情景：1）城市区域，可再生能源发电量相对较少（数据传输延迟较短）；2）农村地区，可再生能源来源充足（数据传输延迟相对较长）。对于核心路由器，由于核心路由器通常分布在全州各地以传输数据包 [19, 20, 64]，我们采用该州电网的平均碳强度。

对于数据中心，我们根据其供电方式考虑两种情形 [2, 106]：1）电网组合，即数据中心由最近电网产生的混合能源供电；2）无碳[1]，即数据中心完全由可再生能源供电。

运行时方差：我们探讨用户使用应用程序时普遍会遇到的两种运行时方差 [37, 38, 72]：共置工作负载造成的干扰和网络稳定性。对于共置工作负载造成的干扰，我们在运行其他工作负载的同时，测量每个组件的计算延迟。为模拟不稳定的边缘网络，我们在无线信号强度较差的情况下测量数据传输延迟 [70], [72]。同样，我们根据 [10, 12, 61, 62]中建模的变化（例如拥塞）的影响，对不稳定的核心网络进行建模。

隐含碳足迹模型：尽管隐含碳排放近来日益受到关注 [50, 51], ，隐含碳足迹（CF）模型的部署仍处于起步阶段。为探究隐含碳足迹模型对绿色应用调度决策的影响，我们使用了两种隐含碳足迹建模工具：ACT 和生命周期评估工具。

对于计算组件，我们首先使用以下方法计算计算组件的隐含碳排放：

ACT [50]。我们还根据制造企业提供的生命周期评估报告 [7, 60, 105, 108, 113] 获取其隐含碳排放。例如，对于移动设备，我们采用 Google Pixel 可持续发展报告中给出的隐含碳足迹 3 [48]。另一方面，对于服务器级系统，我们采用戴尔 R740 [21] 可持续发展报告中给出的隐含碳足迹。请注意，根据 [51], ，这两种建模工具估算的隐含碳排放相差 28%。

对于基站和路由器，我们不使用 ACT，因为 ACT 未对收发器等网络组件建模。相反，我们根据网络基础设施公司提供的生命周期评估报告来获取其隐含碳排放 [19, 20, 27, 28, 29, 30, 31]。

5. 评估结果与分析

本节中，我们使用 *GreenScale* 探索代表性边缘云应用的碳感知调度设计空间。该设计空间包括 1）调度决策（即跨边缘云基础设施分配用户请求），以及 2）资源分配（即调整租用的服务器实例数量）。对于每个应用类别，设计空间的规模为 ~200K，由工作负载特征、变化的碳强度、运行时方差和可用的调度决策组合而成。

5.1 碳设计空间结果概述

图 5 展示了边缘云基础设施中各种应用用例的碳优化设计空间结果，包括 (a) 性能、(b) 能耗和 (c) 碳足迹。x 轴表示每种工作负载可用的执行目标；(a)、(b) 和 (c) 的 y 轴则分别表示这些执行目标的归一化延迟、能耗和碳足迹。此处，移动设备、边缘数据中心和超大规模数据中心的碳强度模型分别为夜间充电器、城市区域和电网混合。ACT 用作隐含碳足迹建模工具，且不存在运行时方差。

碳最优执行目标并不总是与性能最优或能耗最优执行目标相同。性能、能耗和碳排放最优的执行目标（图 *5* 中的绿色星号）会随用例而变化。

对于人工智能工作负载，延迟、能耗和碳排放最优的执行目标取决于神经网络的计算与通信比。例如，尽管 MobileNet、SqueezeNet 和 ResNet 的传输数据量相同，但碳最优执行目标各不相同（即，

[1]注意，云服务提供商声称实现净零排放（每年通过可再生能源信用抵消数据中心的能耗），并不意味着其实际能耗是无碳的 [2] —— 当可再生能源供应在某些时段不足时，数据中心仍会消耗高碳能源。

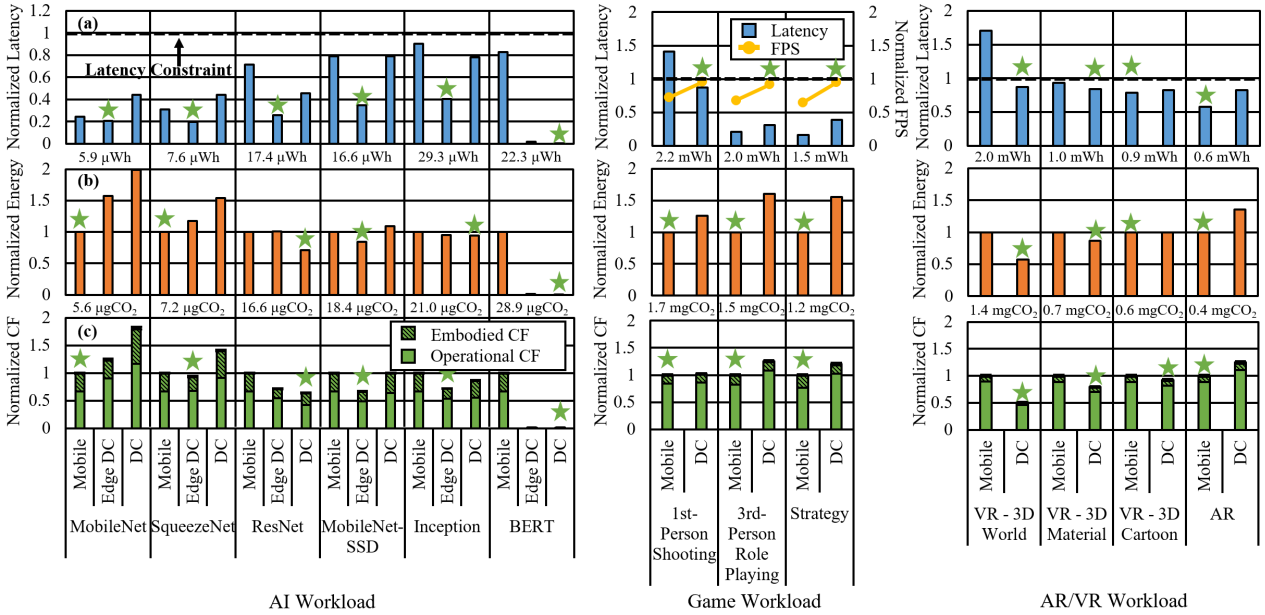


图 5：边缘云基础设施中的碳优化设计空间：性能、能耗和碳排放最优的执行目标因用例而异。请注意，延迟按各工作负载对应的时延约束归一化，而能耗和碳足迹则按各工作负载中移动设备的能耗和碳足迹归一化——(b) 和 (c) 上方的数值分别表示移动设备处理各工作负载时的能耗和碳足迹。

分别为移动设备、边缘数据中心和数据中心），这是由于它们的浮点运算次数不同。另一方面，MobileNet-SSD 和 Inception 的碳排放最优执行目标均为边缘数据中心，因为与 MobileNet 相比，它们的数据传输开销更大，浮点运算次数也更多。对于 BERT，由于数据传输开销最小，延迟、能耗和碳排放最优的执行目标均为数据中心。

对于游戏工作负载，数据中心（即云游戏服务）通过提供更高且满足时延约束的每秒帧数，相较于移动设备（即 Android 应用）提供更好的体验质量（QoE）。然而，由于数据传输带来的运行能耗开销较高，移动设备的碳排放始终低于数据中心——数据中心需要持续将捕获的画面传输到移动设备。

对于 AR/VR 工作负载，延迟、能耗和碳排放最优的执行目标也取决于工作负载的计算通信比。事实上，所有 AR/VR 工作负载的数据传输量相同，详见表 6。然而，与其他工作负载相比，VR - 3D World 需要更多计算。因此，它在移动设备上无法满足时延约束，最终由数据中心执行时碳足迹更低。另一方面，其余工作负载由移动设备执行时碳足迹更低。

这项关于 *GreenScale* 的工作展示了智能碳感知计算调度的潜力。通过将计算范围扩展到本地客户端设备之外，可以将应用的碳排放成本降低 *29.1%*。

组件碳强度的变化和运行时方差使碳优化设计空间变得更加复杂。由于设计空间复杂，碳优化有别于传统的性能和能耗优化，为碳效率提升留下了显著空间——碳感知

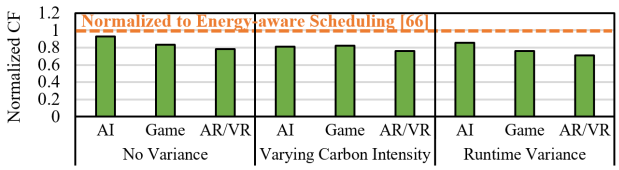


图 6：碳优化不同于性能优化和能耗优化，因此碳排放改善仍有很大空间。横轴表示设计参数，纵轴表示碳感知调度相对于能耗感知调度的归一化碳足迹 [72]。

调度器（即明确考虑组件碳排放特征、在各组件间分配用户请求的调度器）比最先进的调度器 [72]最多可提升 29.1%（图 6 中最右侧的柱形）。

5.2 不同碳强度的影响

移动设备的碳强度： 图 7 展示了在不同充电场景下，移动设备运行 ResNet 的碳足迹。横轴表示三种具有不同电池充电场景的执行目标（即每列中的夜间充电器、平均充电用户和智能充电器），纵轴表示每个执行目标的累计碳足迹。请注意，边缘数据中心和超大规模数据中心的碳强度模型分别为城市区域和电网组合。ACT 用作隐含碳足迹建模工具，且不存在运行时方差。

客户端设备的智能碳感知电池充电

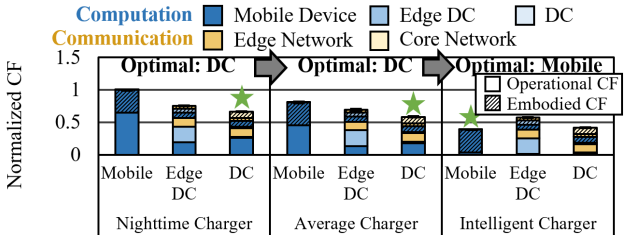


图7：Mobile 在不同充电场景下运行 ResNet 的碳足迹。请注意，执行目标的碳足迹已归一化为 Mobile 使用夜间充电器时的碳足迹。Mobile 的碳强度取决于移动设备用户何时以及如何为电池充电。如果用户根据最近电网的可再生能源可用性智能地为电池充电（即使用智能充电器），便可大幅降低碳足迹（61.2%）。

碳排放最多可减少 *61.2%*。考虑到客户端设备的规模，从智能手机到可穿戴设备，数量达到数十亿 *[110]*, 其碳影响十分显著。

使用夜间充电器时，Mobile 的碳强度通常较高，因为电池是在夜间充电的，而此时可再生能源较少。当用户从夜间充电器改用智能充电器时，Mobile 的碳强度会显著降低，使整体碳足迹最多减少 61.2%（即 Mobile 柱状条所示）——请注意，其他工作负载也呈现类似趋势。在这种情况下，碳排放最优执行目标会从数据中心转移到 Mobile。

数据中心的碳强度：图8展示了在边缘数据中心和基站的不同碳强度场景下，(a) ResNet 和 (b) MobileNet-SSD 的延迟和碳足迹。副 y 轴显示归一化延迟。请注意，移动设备和超大规模数据中心的碳强度模型分别为夜间充电器和电网组合。隐含碳足迹建模工具采用 ACT，且运行时方差为零。当边缘数据中心和基站位于城市区域时，由于可用可再生能源较少，其碳强度通常较高 [120]。相反，边缘网络延迟相对较短。另一方面，当边缘数据中心和基站位于农村地区时，由于可用可再生能源充足，其碳强度较低 [120]——边缘网络延迟则会相应增加。

考虑到这种地理权衡，碳排放最优的执行目标取决于边缘数据中心和基站的位置，以及工作负载特征。对于计算量相对较大、数据传输开销适中的 ResNet，农村地区带来的碳足迹收益超过了网络延迟增加造成的损失（即边缘数据中心对应的柱状条）。另一方面，对于传输数据量较大的 MobileNet-SSD，由于网络延迟增加，在农村地区无法满足延迟约束。这一结果表明，选择基础设施组件（例如城市区域的边缘数据中心与农村地区的边缘数据中心）时，需要仔细考虑地理权衡以及工作负载特征。

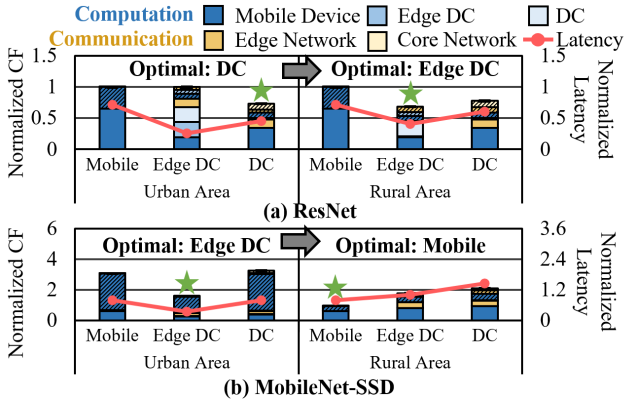


图8：(a) ResNet 和 (b) MobileNet-SSD 在边缘数据中心和基站不同碳强度场景下的延迟和碳足迹。请注意，延迟和碳足迹均以城市区域中移动设备的对应值为基准进行归一化。边缘数据中心的碳强度取决于其位置（即，可再生能源较少的城市区域和可再生能源充足的农村地区），但网络延迟也取决于位置。

应用程序的碳成本取决于整个计算体系中各类基础设施所用电力的碳排放强度，这些基础设施包括边缘、数据中心、基站，以及连接边缘与数据中心的网络基础设施。

图 9 展示了在数据中心的不同碳强度场景下，(a) MobileNet-SSD 和 (b) 增强现实的碳足迹。请注意，移动设备和边缘数据中心的碳强度模型分别为夜间充电器和城市区域。隐含碳足迹建模工具采用 ACT，且不存在运行时方差。如图 9 所示，碳强度对碳最优决策的影响取决于工作负载。对于 MobileNet-SSD，即使数据中心的能耗完全由可再生能源供给（即无碳），总体碳足迹也不会显著变化。这是因为，由于其他组件的网络开销和空闲开销较高，数据中心的运行碳足迹仅占碳足迹的一小部分。另一方面，对于增强现实，无碳仍然有益（即数据中心对应的柱形），因为大量计算可以受益于数据中心。

碳强度不确定性：在现实环境中，由于环境因素（例如云湍流、太阳辐射等），可再生能源发电即使在一小时内也可能波动 [57, 63, 67, 86, 112]。这类波动可能导致碳强度不确定性。

为了解调度决策中不确定性的影响，我们根据可再生能源发电的统计分布，将不确定性注入 GreenScale——根据 [16, 33], , 太阳能和风能发电的波动分别可用贝塔分布和威布尔分布建模。即使碳强度存在 16.8% 的波动，*GreenScale* 得出的调度决策仍保持一致（见图中的误差棒）。这一结果表明，*Green-Scale*得出的总体结论通常适用于现实环境。

5.3 运行时方差的影响

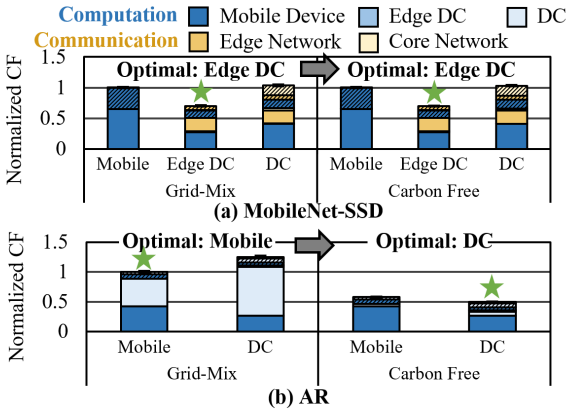


图9：不同数据中心碳强度场景下，(a) MobileNet-SSD 和 (b) 增强现实的碳足迹。数据中心的碳强度场景（即每列中的电网组合和无碳）也会影响碳排放最优执行目标。数据中心碳强度场景对碳排放最优执行目标的影响，则因工作负载特征而异。

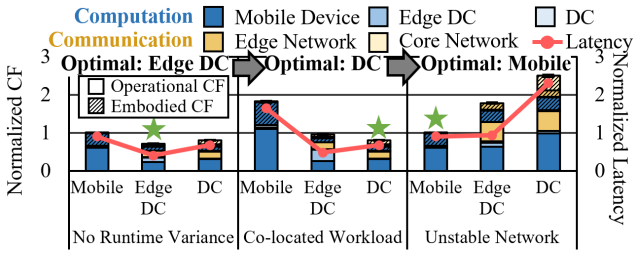


图10：在 (a) 没有运行时方差、(b) 计算机系统上存在共置工作负载以及 (c) 网络不稳定时，Inception 的延迟和碳足迹。各种来源的运行时方差会改变碳排放最优执行目标。

图 10 展示了 Inception 的碳足迹（CF）和延迟：(a) 不存在运行时方差时；(b) 存在共置工作负载时；(c) 网络不稳定时。请注意，移动设备、边缘数据中心和超大规模数据中心的碳强度模型分别为夜间充电器、城市区域和 Grid- Mix，并且使用了 ACT。

运行时方差，尤其是在边缘云计算中很常见，会对应用程序的碳成本产生显著影响。

当不存在运行时方差时，碳排放最优执行目标是边缘数据中心。然而，当存在共置工作负载时，碳排放最优执行目标会转向数据中心。这是因为共置工作负载带来的负面影响取决于各组件的计算能力和内存能力，而数据中心在所有计算组件中拥有最强的计算能力和内存能力。另一方面，当边缘网络或核心网络不稳定时，由于数据传输开销显著增加，碳排放最优执行目标会转向移动设备

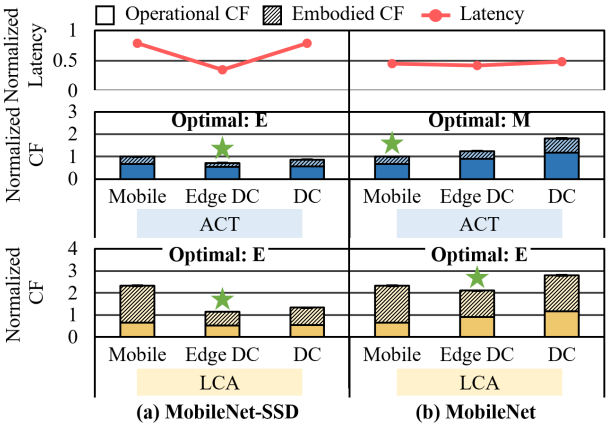


图11：使用两种不同的隐含碳足迹建模工具时，(a) MobileNet-SSD 和 (b) MobileNet 的延迟和碳足迹。根据估算的隐含碳足迹量，碳最优决策可能有所不同。

由于数据传输开销显著增加。

5.4 应用程序的设计考量

隐含碳足迹模型：鉴于组件的隐含碳足迹不断增加，本研究探讨了两种隐含碳足迹建模工具，即 ACT 和 LCA；它们分别对组件的隐含碳足迹进行较低和较高的估算。图 11 展示了执行目标在 (a) MobileNet-SSD 和 (b) MobileNet 上的延迟和碳足迹（采用两种不同的建模工具）。请注意，移动设备、边缘数据中心和超大规模数据中心的碳强度模型分别为 NighttimeCharger、城市区域和电网组合，且不存在运行时差异。

对于 MobileNet-SSD，两种不同的建模工具均将边缘数据中心判定为碳排放最优执行目标。这是因为边缘数据中心的延迟最短（延迟会线性影响隐含碳足迹），同时运行能耗最低。另一方面，对于 MobileNet，ACT 将移动设备判定为碳排放最优执行目标，而 LCA 则将边缘数据中心判定为碳排放最优执行目标。这种差异源于各执行目标在延迟和运行能耗方面的不同模式——对于适用于移动设备的 MobileNet，移动设备的运行能效得到了进一步优化。考虑到 LCA 对隐含碳足迹的估算高于 ACT，这一结果表明，未来设计绿色应用程序时，应进一步考虑不断增加的隐含碳足迹。

租用的服务器数量：传统上，应用设计人员会根据用户数量及其请求，尝试优化从云服务提供商处租用的服务器数量。然而，在碳优化设计空间中，效率与碳足迹之间的权衡同样至关重要。为提供这一参数的指导，我们将探讨其影响。

GreenScale 使应用开发者能够在配置和租赁云基础设施时做出碳排放知情决策。碳排放知情的云资源规划可以

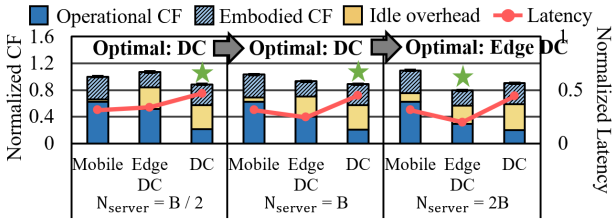


图12：从数据中心租用不同数量的服务器时，SqueezeNet 的延迟和碳足迹。延迟已按延迟约束归一化，碳足迹已按移动设备在$N_{servers}$为$B/2$时的情况归一化。请注意，$N_{servers}$和B分别表示租用的服务器数量和最优批量大小（即 1024）。随着租用服务器数量增加（从左到右），延迟和运行效率均有所提升。延迟的改善也降低了空闲开销和隐含碳足迹开销。

碳成本最多降低 *24.9%*。

图12展示了数据中心中服务器数量不同时，SqueezeNet 的延迟和碳足迹。请注意，移动设备、边缘数据中心和超大规模数据中心的碳强度模型分别为夜间充电器、城市区域和电网组合。此处采用 ACT，且不存在运行时方差。如图12所示，随着服务器数量增加，延迟和运行效率均有所改善。尽管使用的服务器数量更多，但由于延迟得到改善，空闲开销和隐含碳足迹开销也有所改善。由于小规模边缘数据中心的改善幅度更大，碳排放最优执行目标从数据中心转向边缘数据中心。

工作负载相关参数：正如我们在第 5.1 节中所述，碳排放最优执行目标取决于工作负载的计算通信比。开发者可以利用工作负载相关参数改变其计算通信比，从而改善碳足迹。

应用程序级优化手段，例如帧分辨率设置（游戏）和任务划分（*AR/VR*流水线），分别可实现最高 *31.1%* 和 *14.8%* 的碳减排。

对于游戏应用，可以牺牲画面质量，以减少执行目标的计算量和数据传输量。图13(a)展示了不同分辨率选项下第三人称角色扮演游戏应用的碳足迹。将分辨率选项从全高清调至高清后，随着计算量和数据传输开销的减少，整体碳足迹也降低了（31.1%）。不过，降低分辨率可能会对用户的 QoE 产生不利影响。

对于 AR/VR 应用程序，可以根据各流水线阶段所需的计算量以及中间结果的大小，在移动设备和数据中心之间划分不同的流水线阶段。图13(b)展示了不对任务进行划分，以及在移动设备和数据中心之间划分流水线阶段时 AR 的碳足迹。任务划分后，整体碳足迹降低了 14.8%。原因有二：1）各阶段中间结果的大小小于输入的数据量，从而减少了数据传输开销；2）资

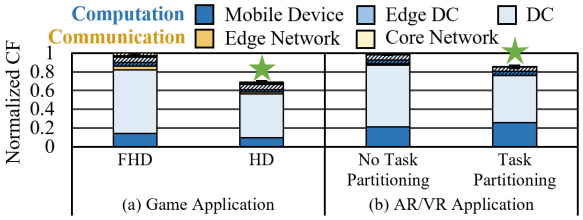


图13：通过利用依赖于工作负载的参数，可以改变计算与通信的比例，从而提高碳效率。请注意，移动设备、边缘数据中心和超大规模数据中心的碳强度模型分别为夜间充电器、城市区域和电网组合。ACT 用作隐含碳足迹建模工具。

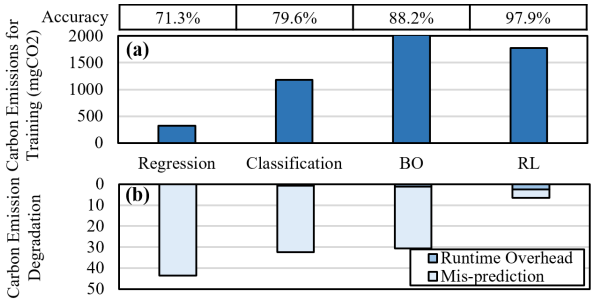


图14：不同调度方法的开销和准确性各不相同（a）。由于运行时开销和误预测会导致碳排放增加（b），因此设计绿色应用程序时需要谨慎选择调度方法。

源（即移动设备和数据中心）的利用率提高，使空闲碳足迹降低了 55.3%。

调度方法：由于碳最优执行目标会随各种参数而变化，因此选择能够适应这些参数的调度方法至关重要。在边缘云执行环境中，有多种可用的调度方法。由于这些技术在碳开销和准确性方面各不相同，因此需要进行谨慎选择。

为提供选择指南，我们定制并实现了以下调度方法，并使用*GreenScale*评估其碳开销和准确性[2]：

- 回归：根据基于回归的模型预测的碳足迹确定执行目标 [104]，
- 分类：根据基于分类的模型确定执行目标 [111, 128]，
- 贝叶斯优化（BO）：根据通过 BO 获得的预测模型确定执行目标 [107]，，以及
- 强化学习（RL）：通过考虑环境特征，使用自行学习的策略确定执行目标 [72]。

请注意，为了公平比较，我们对所有调度方法的参数（例如强化学习的超参数）都进行了微调。

[2]准确率和碳足迹结果通过 *GreenScale* 获得，而开销则通过真实系统实现进行测量。

图 14 展示了（a）各项技术的训练碳开销和预测准确率，以及（b）与最优调度器相比，各项技术造成的碳排放劣化。总的来说，只要基础设施组件没有变化，训练开销就是设计绿色应用程序时的一次性成本。然而，预测准确率会显著影响应用程序的碳效率，因为误预测会导致碳足迹劣化。

在图 14(b)中，运行时开销越大，碳排放劣化通常越小，这通常是误预测所致。在这些方法中，强化学习通过自我学习策略来适应碳强度和运行时方差等特征，以牺牲一定开销为代价实现了最高的准确性；而其他方法未能准确建模这些特征之间的非线性关系——不过，强化学习的运行时开销仅占每帧执行时间的 2.4%。*GreenScale* 的碳知情决策可通过多种具有不同准确性与开销之间的权衡。

y-

6. 相关工作

边缘云应用的优化： 从移动设备用户的角度来看，优化 QoE（即用户端设备能效与响应时间性能的乘积）至关重要。因此，以用户为中心的优化技术致力于找出能够最大限度降低用户端设备运行能耗、同时满足性能约束的最佳执行目标 [13, 32, 53, 65, 69, 73, 117, 118, 132]。

从数据中心运营商的角度来看，在不影响服务级别协议 (SLA) 约束的前提下，优化资源利用率、运行效率和执行成本至关重要。因此，云中心化优化技术尝试将计算分配到边缘规模和超大规模服务器上，从而提高基础设施的运行效率，同时满足服务级别协议中的性能约束 [54, 66, 88, 114, 130],。其中，一些技术尝试考虑数据中心 [54, 114] 中可再生能源可用性的时变特性。例如，[76, 77] 尝试根据不同地点数据中心的碳强度，在数据中心之间切换计算任务；而 [99] 则尝试根据数据中心内的可再生能源情况安排批处理作业。此外，也有针对数据中心的碳感知能源容量规划研究 [11]，考虑了场内和场外可再生能源发电及其成本。

尽管已有许多技术，但它们往往无法做出全局碳最优决策，因为这些技术没有考虑碳优化的独特特性，例如边缘侧碳强度的变化，以及隐含碳排放的摊销。这主要是由于缺乏能够量化并分析基础设施组件碳排放的框架，而该框架需考虑上述特性。

碳建模工具： 鉴于信息与通信技术的碳足迹日益增加，学术界和工业界提出了多种工具，用于量化硬件生命周期各阶段的碳排放。基于火用的工具采用能量平衡方法，量化服务器在制造和使用过程中的环境影响——期间的环境影响 [55, 97]。尽管能量平衡方法可以简化可持续系统的设计空间，但基于火用的工具并未考虑可再生能源在制造和使用过程中的影响。生命周期评估工具量化产品在生命周期各阶段的碳足迹 [60, 105, 108], , 但并未转向计算系统。生命周期评估工具使用粗粒度信息（例如电子设备的经济成本、系统的物料清单等）。尽管工业界发布的产品环境报告基于生命周期评估工具 [6, 7, 113],，但这些报告尚未用于系统或硬件组件之间的比较分析，以指导设计空间探索。作为对上述工具的补充，ACT 被提出用于考虑计算系统硬件制造和运营使用所产生的直接碳足迹 [50]。

基于上述工具，本研究提出了*GreenScale*，这是一个考虑碳优化独特特征的碳设计空间探索与优化框架。通过使用*GreenScale*，我们展示了碳优化有别于性能和能耗优化，这是由碳强度的变化、运行时方差以及隐含碳排放所致。基于设计空间探索，本研究为设计绿色应用提供了指导。

7. 结论

鉴于信息与通信技术基础设施的碳排放日益增长，设计绿色应用 以提升基础设施组件的碳效率至关重要。本文提出了*GreenScale* ，一个设计空间探索与优化框架。通过对边缘云基础设施中最先进的应用进行深入的碳特征分析，*GreenScale* 表明，碳优化的调度决策取决于多种特征，例如工作负载特征、基于时间 和位置的碳强度、随机运行时方差，以及隐含碳排放的摊销。这些特征使碳优化有别于性能和能耗优化，也为提升碳效率留下了很大空间（与最先进的边缘云调度器相比，最高可提升 29.1%）。我们认为，*GreenScale* 可作为切实可行的解决方案，为开发者提供详细指导，帮助他们设计和实现绿色应用 ，从而实现可持续运行。

参考文献

[1] 365数据中心。 [在线]。可访问: https://365datacenters.com/cloud-solutions[2] B. Acun、B. C. Lee、F. Kazhamiaka、K. Maeng、M. Chakkaravarthy、U. Gupta、D. Brooks 和 C. J. Wu，“设计碳排放感知数据中心的整体方法”，载于国际编程语言与操作系统架构支持会议（*ASPLOS*）论文集，2023年。[3] Amazon，“Alexa”。 [在线]。可访问: https://developer.amazon.com/en-US/alexa[4] Amazon，“AWS”。 [在线]。可访问: https://aws.amazon.com[5] A. S. G. Andrae 和 T. Edler，“通信技术的全球电力使用量：截至2030年的趋势”，*Challenges 6*，第1卷，第117–157页，2015年。[6] Apple，“Apple 环境责任报告”。 [在线]。可访问: https://www.apple.com/environment/pdf/Apple¯环境_责任_报告_2019.pdf

[7] Apple，“产品环境报告 iPad”。 [在线]。可访问：https://www.apple.com/environment/pdf/products/ipad/iPad_PER_sept2019.pdf

[8] Apple，《Siri 。》 [在线]. 可通过以下网址获取：https://www .apple .com/siri

[9] AWS，《云计算、服务器利用率与环境 .”[在线]。可访问：https://aws.amazon.com/ko/blogs/aws/cloud-computing-server-utilization-the-environment/

[10] U . Baldoni, F. Antonelli 和 M. Ambriola，“核心网络延迟研究”，载于*2020*年 第*30*届国际电信网络与应用会议 *(ITNAC)*。IEEE，2020年，第 1–6 页。

[11] R . Bianchini，“在数据中心利用可再生能源：现状与未来”，载于国际高性能并行与分布式计算研讨会论文集 *(HPDC)*，2012年，第 135–136 页。

[12] J . Botha，《OSP土建工程指南》，2015年 . [在线] . 获取地址：https://www.thefoa.org/tech/ref/1pstandards/OSP% 20Civil%20Works%20Guide-FOA.pdf

[13] E. Cai、D.-C. Juan、D. Stamoulis 和 D. Maculescu，《Neuralpower：预测并部署节能型卷积神经网络》，载于亚洲机器学习会议（*ACML*）论文集，2017年。

[14] W . Cai, R. Shea, C.-Y. Huang, K.-T. Chen, J. Liu, V. C. M. Leung 和 C. H. Hsu，《云游戏：电脑游戏的未来综述》，*IEEE Access*，第 4 卷，第 7605–7620 页，2016 年。

[15] M . Casado 和 P. Barford，《流量分析》《计算机通信》，第 188 卷，116，2022年 99页。–。

[16] A . N. Celik，《基于威布尔模型和瑞利模型对土耳其南部地区风功率密度的统计分析》《可再生能源》，第 29 卷，第 4 期，第 593–604 页，2004年。

[17] J. Chang、J. Meza、P. Ranganathan、A. Shah、R. Shih 和 C. Bash，“全面绿色：评估并设计服务器的生命周期环境影响”，《ACM SIGPLAN 通知》，第 47 卷，第 4 期，第 25–36 页，2012年。[18] B. Charyyev、E. Arslan 和 M. H. Gunes，“云数据中心与边缘服务器的延迟比较”，载于 IEEE 全球通信大会（GLOBECOM）论文集，2020年。[19] 思科。 [在线]。可访问：https://www.cisco.com/c/en/us/products/collateral/routers/8000-series-routers/datasheet-c78-742571.html[20] 思科。 [在线]。可访问：https://www.cisco.com/c/en/us/products/collateral/routers/catalyst-8300-series-edge-platforms/datasheet-c78-744088.html[21] 戴尔，《Dell R740 的生命周期评估》。 [在线]。可访问：https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/Full_生命周期评估_戴尔_Dell R740.pdf[22] N. Ding、D. Wagner、X. Chen、A. Pathak、Y. C. Hu 和 A. Rice，“无线信号强度对智能手机电池耗电的影响：特征分析与建模”，载于计算机系统测量与建模国际会议（SIGMETRICS）论文集，2013年，第 29–40 页。[23] EdgeConneX。 [在线]。可访问：https://edgeconnex.com[24] B. Egilmez、M. Schuchhardt、G. Memik、R. Ayoub、N. Soundararajan 和 M. Kishinevsky，“安卓智能手机中面向用户的帧率管理”，ACM Transactions on Embedded Computing Systems，第 16 卷，第 5s 期，第 1–17 页，2017年。[25] electricityMaps。 [在线]。可访问：https://www.elertricitymaps.org [26] Y. Endo、Z. Wang、J. B. Chen 和 M. I. Seltzer，“使用延迟评估交互式系统性能”，载于 USENIX 操作系统设计与实现研讨会（OSDI）论文集，1996年。[27] 爱立信。 [在线]。可访问：https://docbox.etsi.org/workshop/ 2015/201506_EEWORKSHOP/SESSION03_生命周期评估/生命周期评估 of LTE based station primary data Bergmark Ericsson.pdf[28] 爱立信。 [在线]。可访问：https://www.diva- portal.org/smash/record.jsf?pid=diva2%3A1540115&dswid=4215[29] 爱立信。 [在线]。可访问：https://www.ngmn.org/wp-content/uploads/210719-NGMN_绿色-未来网络_生态设计-v1.0.pdf

[30] 爱立信。 [在线]。可访问：https://www.diva-portal.org/smash/get/diva2:1540115/FULLTEXT01.pdf

[31] 爱立信，《边缘计算——分阶段介绍 .》 [在线]. 可访问：https://ericsson.com/en/edge-computing

[32] A . E. Fahim 和 M. Pedram，《移动云计算环境中深度神经网络的能耗与高性能计算卸载》，载于大湖区超大规模集成电路研讨会（*GVLSI*）论文集，2018年，第111–116页。

[33] F . Y. Bouroumi，《利用太阳能测量数据进行统计分析》，可再生能源，第26卷，第1期，第47–67页，2002年。

[34] D . Ferreira, A. K. Dey 和 V. Kostakos，《理解人类与智能手机相关的关注问题：一项关于电池续航时间的研究》普适计算，第 19–33 页，2011 年。

[35] Fitbit，《Fitbit Flex 2》。 [在线]。网址：https://www.fitbit.com/in/flex2

[36] T . L. 基础设施，《边缘状态 [在线]. 其途径，https://stateoftheedge.com

[37] B. Gaudette、C.-J. Wu 和 S. Vrudhula，“通过概率性服务质量保证平衡性能与能耗，改善智能手机用户体验”，2016。

[38] B . Gaudette, C. Wu 和 S. Vrudhula，《在存在不确定性来源下优化移动工作负载的用户满意度》，*IEEE*移动计算汇刊，第18卷，第12期，第2941–2953页，2019。

[39] Godot网址《材质测试器》，2020年 . [在线] . https://github.com/godotengine/godot-demo-projects/tree/master/3d/material_testers

[40] Godot，“3D平台游戏”，2020年。 [在线]。可访问：https://github.com/godotengine/godot-demo-projects/tree/master/3d/platformer[41] Google，“云端TPU”。 [在线]。可访问：https://cloud.google.com/tpu[42] Google，“Google Cardboard”。 [在线]。可访问：https://arvr.google.com/cardboard/[43] Google，“Google Cloud Vision”。 [在线]。可访问：https://cloud.google.com/vision[44] Google，“Google数据中心”。 [在线]。可访问：https://www.google.com/about/datacenters/ [45] Google，“Google数据中心能效”。 [在线]。可访问：https://google.com/about/datacenters/efficiency[46] Google，“Google Daydream”。 [在线]。可访问：https://arvr.google.com/daydream/[47] Google，“Google翻译”。 [在线]。可访问：https://translate.google.com[48] Google，“Google Pixel 3产品环境报告”，2019。[49] A. Gupta、I. K. Jain 和 D. Bharadia，“ 多个小型基站比单个大功率基站更环保：无线蜂窝网络的密集化”，载于第一届可持续计算机系统设计与实现研讨会（*HotCarbon*），2022。[50] U. Gupta、M. Elgamal、G. Hills、G. Y. Wei、H. H. S. Lee、D. Brooks 和 C. J. Wu，“ACT：借助体系结构碳建模工具设计可持续计算机系统”，载于国际计算机体系结构研讨会论文集（*ISCA*），2022。[51] U. Gupta、Y. G. Kim、S. Lee、J. Tse、H. H. S. Lee、G. Y. Wei、D. Brooks 和 C. J. Wu，“追踪碳排放：计算领域难以捉摸的环境足迹”，载于*IEEE*高性能计算机体系结构国际研讨会论文集（*HPCA*），2021，第854–867页。[52] M. Halpern、Y. Zhu 和 V. J. Reddi，“移动CPU的崛起：量化代际移动CPU设计趋势对性能、能耗和用户满意度的影响”，载于*IEEE*高性能计算机体系结构国际研讨会论文集（*HPCA*），2016，第64–76页。[53] M. Han、J. Hyun、S. Park、J. Park 和 W. Baek，“Mosaic：面向异构性、通信和约束的模型切片与执行，实现准确高效的推理”，载于并行体系结构与编译技术国际会议论文集（*PACT*）

编译技术（*PACT*），2019，第165–177页。

[54] R. Han, S. Wen, C. Liu, Y. Zhan, Z. Wang, and L. Chen, "Edgetuner：面向动态边缘云工作负载和资源的快速调度算法调优"，载于*IEEE*国际计算机通信会议（*INFOCOM*）论文集，2022年。

[55] C. Patel, R. Bash, and P. Ranganathan, "企业服务器的全生命周期火用消耗"，载于《国际电子与环境研讨会论文集》，2008年。

[56] C. Healthcare, 《智能医疗平台》。[在线]。获取地址：https://www.changehealthcare.com/about/innovation/intelligent-healthcare-platform

[57] Y. Tani, H. Hara 和 T《环境因素对光伏组件输出变化的影响——I. 太阳光谱辐射对光伏组件输出的影响》，《太阳能》，第 55 卷，第 6 期，第 463–468 页，1995年。

[58] L. N. Huynh、Y. Lee 和 R.K. Balan，《Deepmon：面向连续视觉应用的移动 GPU 深度学习框架》，载于国际移动系统、应用与服务会议（*MobiSys*）论文集，2017年，第 82–95 页。

[59] M. Huzaifa、R. Desai、S. Grayson、X. Jiang、Y. Jing、J. Lee、F. Lu、Y. Pang、J. Ravichandran、F. Sinclair、B. Tian、H. Yuan、J. Zhang 和 S. V. Adve，《Illixr：推动端到端扩展现实研究》，载于*IEEE*国际工作负载表征研讨会（*IISWC*）论文集，2021。

[60] C. M. U. G. D. 研究所，"云计算生命周期研究所（eiolca）"，[在线]。网址：http://www.eiolca.net

[61] M. Casetti, C. Chiasserini, 等，"无线基础设施的重要：测量与经验教训"，*ACM SIGCOMM*计算机通信评论，第51卷，第4期，第2–13页，2021。

[62] 国际电信联盟，"以尽量减少现有基础设施为目标的光纤电缆安装标准"，2018。

[63] W. Marwali 和 R. Kumar，"移动云团对采用分布式光伏发电的电力公司的影响"，*IEEE*能源转换汇刊，第4期，第570–576页，1987年。

[64] V. Chapman，《Cisco IP及下一代网络的QoS部署：权威指南》，*Morgan Kaufmann*，2009。

[65] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars 和 L. Tang，《Neurosurgeon：在移动边缘之间的协同智能》，载于国际计算机体系结构支持编程语言与操作系统会议论文集（*ASPLOS*），2017年，页码 615–629。

[66] K. Kaur、S. Gang、G. Kaddoum、S. H. Ahmed 和 M. Atiquzzaman，《Keids：面向边缘云生态系统中工业物联网的基于 Kubernetes 的能量与干扰驱动调度器》，*IEEE Internet of Thins Journal*，第 7 卷，第 5 期，2020年。

[67] E. C. Kern、E. M. Gulachenski 和 G. A. Kern，《云对分布式光伏发电影响：马萨诸塞州 Gardner 光伏实验中的缓慢瞬态》，*IEEE*能源转换汇刊，第 4 卷，第 2 期，页码 184–190，1989年。

[68] Y. Kim, J. Kim, D. Chae, D. Kim 和 J. Kim，《μLayer：利用协同单层加速和处理器友好型量化实现低延迟设备端推理》，载于欧洲计算机系统会议（*EuroSys*）论文集，2019。

[69] Y. Kim, J. Kim, D. Chae, D. Kim 和 J. Kim，《μLayer：利用协同单层加速和处理器友好型量化实现低延迟设备端推理》，载于欧洲计算机系统会议（*EuroSys*）论文集，2019。

[70] Y. Kim, J. Lee, J. Kim, H. Jung 和 S. Ha，《异构移动多核处理器上多媒体应用的能效》，*IEEE Transactions on Computers*，第 66 卷，第 11 期，第 1878–1889 页，2017年。

[71] Y. Kim, J. Lee, J. Kim, H. Jung 和 S. Ha，《异构移动多核处理器上多媒体应用的能效》，*IEEE Transactions on Computers*，第 66 卷，第 11 期，第 1878–1889 页，2017年。

[72] Y. Kim, J. Lee, J. Kim, H. Jung 和 S. Ha，《异构移动多核处理器上多媒体应用的能效》，*IEEE Transactions on Computers*，第 66 卷，第 11 期，第 1878–1889 页，2017年。

[73] N. D. Lane, S. Bhattacharya, A. Mathur, P. Georgiev, C. Forlivesi 和 F. Kawsar，《Deepx：一款用于移动设备低功耗深度学习推理的软件加速器》，载于传感器网络信息处理国际会议（*IPSN*）论文集，2016年，第98–107页。[74] Y. Li、D. Brooks、Z. Hu 和 K. Skadron，"SMT和CMP架构中的性能、能耗和热考量"，载于高性能计算机体系结构国际研讨会（*HPCA*）论文集，2005年。[75] T. Liu、S. He、S. Huang、D. Tsang、L. Tang、J. Mars 和 W. Wang，"云端交互式三维应用的基准测试框架"，载于国际微体系结构研讨会（*MICRO*）论文集，2020年。[76] Z. Liu、M. Lin、A. Wierman、S. H. Low 和 L. L. H. Andrew，"采用可再生能源的地理负载均衡"，*ACM SIGMETRICS*性能评估评论，第39卷，第3期，第62–66页，2011年。[77] Z. Liu、M. Lin、A. Wierman、S. H. Low 和 L. L. H. Andrew，"绿色地理负载均衡"，*ACM SIGMETRICS*性能评估评论，第39卷，第1期，第193–204页，2011年。

[78] D. Lo, T. Shiping, C. Huang, C. Delimitrou, 交互应用的指导性进化，能耗和负载均衡"，载于国际微体系结构研讨会（*MICRO*）论文集，2015年，第508–520页。

[79] P. Mattson, V. J. Reddi, C. Cheng, C. Coleman, G. Diamos, D. Kanter, P. Micikevicius, D. Patterson, G. Schmuelling, H. Tang, G. Wei 和 C.-J. Wu，"MLPerf：用于衡量机器学习性能的行业标准基准测试套件"，*IEEE Micro*，2020年。

[80] Meta，"利用专用硬件加速 Facebook 的基础设施。"[在线]。链接：https://engineering.fb.com/2019/03/14/data-center-engineering/accelerating-infrastructure/

[81] Meta，可访问：https://datacenters.fb.com/ "Meta 数据中心》[在线]。

[82] Meta，《软件、服务器、系统、传感器与科学：Facebook 打造超高效数据中心的秘诀。》[在线]。可访问：https: // tech.fb.com/engineering/2020/01/hyperefficient-data-centers/

[83] Microsoft，可用：https://azure.microsoft.com/en-us/free/ai/ "Azure 人工智能》[在线]。

[84] Microsoft，可用：《HoloLens 2 》[在线]。https://www.microsoft.com/en-us/hololens

[85] Microsoft，《微软 Azure：云计算服务 》[在线]。可访问：https://azure.microsoft.com

[86] P. Peierls，"准粒子谱的波动性"，《物理评论快报》，第 110 卷，第 13 期。

[87] MLCommons。[在线]。可访问：https://mlcommons.org/

[88] R. Mohan, J. Gluhak, M. Imran 和 L. Xu，"物联网边缘云计算中的异构资源调度，"载于*IEEE*国际云计算会议（*CLOUD*）论文集，2020年。

[89] Monado，"Godot 中使用 OpenXR 插件的 Sponza 场景"，2019。[在线]。可访问：https://github.freedesktop.org/monado/demos/godot-sponza-openxr

[90] Monsoon，"高压电源监测器 。"[在线]。https://www.msoon.com/high-voltage-power-monitor

[91] 英伟达，"GeForce Now 。"[在线]。https://www.nvidia.com/en-us/geforce-now/memberships/?ncid=pa-srch-goog-180712&gclid=CjwKCAiAjs2bBhACEiwALTBWZU3scU9_rGhn4E9ib2_6eTs2-FUY1A0ZwD6O40EcQ0E04lbSbHfrihoCyXMQAvD_BwE#cid=ct0005_pa-srch-goog_en-us

[92] Oculus，"把世界变成你的游戏云。"[在线]。可用：https://www.oculus.com/?locale=en_US

[93] E. Kellner，《智能手机能量水平预测的实证方法》，2011 年，第 345–354 页。

[94] OMRON，"时间指南针。"[在线]。网址：https://omronhealthcare.com/products/heartguide-wearable-blood-pressure-monitor-bp8000m/

[95] 高通，《骁龙845移动平台 》[在线]。可用：

https://www.qualcomm.com/products/snapdragon-845-移动设备-平台

[96] K. Brooks, 《线程迁移：面向多核系统的细粒度功耗管理》，*ACM SIGARCH* 计算机体系结构新闻，第37卷，2009年。

[97] P. Ranganathan, 《绿色云：下一个前沿》，工程前沿桥梁冬季刊，第40卷，19、2010年4月12日，第–页。

[98] V. J. Reddi, C. Cheng, D. Kanter, P. Mattson, G. Schmuelling, C.-J. Wu, B. Anderson, M. Breughe, M. Charlebois, W. Chou, R. Chukka, C. Coleman, S. Davis, G. Diamos, J. Duke, D. Fick, J. S. Gardner, I. Hubara, S. Idgunji, T. B. Jablin, J. Jiao, T. S. John, P. Kanwar, D. Lee, J. Liao, A. Lokhmotov, F. Massa, P. Meng, P. Mcikevicius, C. Osborne, G. Pekhimenko, A. T. R. Rajan, D. Sequeria, A. Sirasao, F. Sun, H. Tang, M. Thomson, F. Wei, E. Wu, L. Xu, K. Yamada, B. Yu, G. Yuan, A. Zhong, P. Zhang 和 Y. Zhou, 《MLPerf 推理基准测试》，*arXiv:1911.02549*，2019年。

[99] C. Reddy, D. Wang, R. Urgaonkar, 《面向数据中心的碳感知能源容量规划》，载于计算机与电信系统建模、分析与仿真国际研讨会论文集（*MASCOTS*），2012年。

[100] D. Richins, D. Doshi, M. Blackmore, A. T. Nair, N. Pathapati, A. Patel, B. Daguman, D. Dobrijalowski, R. Illikkal, K. Long, D. Zimmerman, 和 V. J. Reddi, 《只见树木，不见森林：边缘数据中心中的端到端人工智能应用性能》，载于*IEEE* 高性能计算机体系结构国际研讨会论文集，2020年。

[101] D. Richins, D. Doshi, M. Blackmore, A. T. Nair, N. Pathapati, A. Patel, B. Daguman, D. Dobrijalowski, R. Illikkal, K. Long, D. Zimmerman, 和 V. J. Reddi, 《AI税：人工智能数据中心应用程序的隐藏成本》，*ACM*计算机系统汇刊，第37卷，2021年。

[102] A. Samanta 和 J. Tang, 《Dyme：边缘计算中支持物联网的动态微服务调度》，*IEEE*物联网期刊，第 7 卷，第 7 期，2020年。

[103] S. Sapra, S. Sandhu, 和 F. Pecht, 《便携式电子设备中的电池：用户视角》，*IEEE* 工业电子杂志，第 11 卷，第 35–44 页，2017年。

[104] G. A. F. Seber 和 A. J. Lee, 《线性回归分析》，第2版。约翰·威立父子出版公司，2012年。[105] SimaPro, 《助您推动变革的生命周期评估软件。》 [在线]。网址：https://simapro.com[106] A. Souza、N. Bashir、J. Murillo、W. Hanafy、Q. Liang、D. Irwin 和 P. Shenoy, 《Ecovisor：面向低碳应用的虚拟能源系统》，载于国际编程语言与操作系统架构支持会议（*ASPLOS*）论文集，2023年。[107] A. Souza、L. Nardi、L. Oliveira、K. Olukotun、M. Lindauer 和 F. Hutter, 《先验引导的贝叶斯优化》，载于元学习研讨会（*NeurIPS*）论文集，2020年。[108] Sphera, 《生命周期评估软件。》 [在线]。网址：https://gabi.sphera.com/america/index/[109] Statista, 《应用程序——全球。》 [在线]。网址：https://www.statista.com/outlook/dmo/app/worldwide[110] Statista, 《2020年至2025年全球移动用户数量预测。》 [在线]。网址：https://statista.com/statistics/218984/number-of-global-mobile-users-since-2010[111] J. A. K. Suykens 和 J. Vandewalle, 《最小二乘支持向量机分类器》，《神经处理快报》，第9卷，300，1999年，293页。– 。[112] Y. T. Tan 和 D. S. Kirschen, 《大规模光伏发电对电力系统的影响》，载于《IEEE电力工程学会年会》，2007年，第1–8页。[113] 台积电, 《台积电企业社会责任报告。》 [在线]。网址：https://esg.tsmc.com/download/csr/2018_tsmc_csr_report_published_May_2019/english/pdf/e_all.pdf[114] S. Tuli、G. Casale 和 N. R. Jennings, 《MCDS：人工智能增强的移动边缘云计算系统工作流调度》，*IEEE*并行与分布式系统汇刊，第33卷，第11期，2022年。

[115] VaporIO. [在线]. 网址：https://vapor.io

[116] L. Beuvry, 《协作移动边缘计算中的任务分配与调度》，载于*IEEE*国际网络协议会议论文集（*ICNP*），2018年。

[117] S. Pathania 和 T. Mitra, “嵌入式 ARM big.LITTLE 多核处理器上的高吞吐量卷积神经网络推理”，*IEEE* 集成电路与系统计算机辅助设计汇刊，2020年。

[118] S. Pathania, “异构片上系统上的神经网络推理”，*IEEE*《设计与测试》，2020年。

[119] Y. E. Wang, C.-J. Wu, X. Wang, K. Hazelwood 和 D. Brooks, 《利用深度学习框架中的并行机会》，*arXiv:1908.04705*，2019。

[120] Watttime, 《按电网划分的电网排放强度.》 [在线]. 获取地址：https://watttime.org/explorer/#3/41.23/-97.64

[121] WITHINGS, 《Move ECG 可检测心房颤动的模拟手表。》 [在线]。获取地址：https://www.withings.com/us/en/move-ecg?utm_source=CJ&utm_medium=Affiliate&utm_campaign=affiliation- Skimlinks&utm_content=7099101-Home+Page+US- 13184200&CJEVENT=3669c89e7f8511ea83c700830a1c0e13

[122] C.-J. Wu, D. Brooks, K. Chen, D. Chen, S. Choudhury, M. Dukhan, K. Hazelwood, E. Isaac, Y. Jia, B. Jia, T. Leyvand, H. Lu, Y. Lu, L. Qiao, B. Reagen, J. Spisak, F. Sun, A. Tulloch, P. Vajda,X. Wang, Y. Wang, B. Wasti, Y. Wu, R. Xian, S. Yoo, 和 P. Zhang, 《Facebook 边缘机器学习推理：理解》，载于*IEEE* 高性能计算机体系结构国际研讨会（*HPCA*）论文集，2019，第 331–344 页。

[123] C.-J. Wu, 和 M. Martonosi, 《应用内缓存干扰的特征分析与动态缓解》，载于*IEEE*系统与软件性能分析国际研讨会论文集，ISPASS ’11 系列，2011 年。

[124] D. Xu, Z. Liu, X. Zhang, C. Gao, S. Chen, J. Li, 和 T. Xu, 《理解商用5G：关于其覆盖范围、性能和能耗的首项测量研究”，载于*ACM*计算机通信特别兴趣小组数据通信年度会议论文集：计算机通信的应用、技术、体系结构与协议（*SIGCOMM*），2020年，第479–494页。

[125] M. Yan、C. A. Chan、A. F. Gygax、J. Yan、L. Campbell、A. Nirmalathas 和 C. Beckie, 《移动网络服务和应用程序总能耗建模》，*Energies*，第12卷第1期，2019年。[126] M. Yan、C. A. Chan、W. Li、L. Lei、Q. Shuai、A. F. Gygax 和 C. L. I, 《评估5G无线边缘缓存的能耗》，载于国际通信会议研讨会（*ICC*研讨会），2019年。[127] H. Yu、Q. Wang 和 S. Gun, 《移动边缘计算中的节能任务卸载与资源调度》，载于*IEEE*网络、架构与存储国际会议论文集（*NAS*），2018年。[128] B. Zhang 和 S. N. Srihari, 《使用基于聚类的树进行快速k近邻分类》，*IEEE*模式分析与机器智能汇刊，第26卷第4期，第525–528页，2004年。[129] L. Zhang、B. Tiwana、Z. Qian、Z. Wang、R. P. Dick、Z. M. Mao 和 L. Yang, 《面向智能手机的精确在线功耗估计与基于电池行为的自动功耗模型生成》，载于硬件/软件协同设计与系统综合国际会议论文集，2010年，第105–114页。[130] Y. Zhang、X. Chen、Y. Chen、Z. Li 和 J. Huang, 《边缘计算系统中时延敏感任务的成本高效调度》，载于*IEEE International Conference onService Computing (SCC)*论文集，2018年。[131] S. Zhao、H. Zhang、C. S. Mishra、S. Byuyan、Z. Ying、M. T. Kandemir、A. Sivasubramaniam 和 C. Das, 《Holoar：面向增强现实的三维全息处理即时优化》，载于国际微体系结构研讨会论文集（*MICRO*），2021年。[132] G. Zhong、A. Dubey、C. Tan 和 T. Mitra, 《Synergy：面向嵌入式异构系统高吞吐量卷积神经网络的硬件/软件框架

SoC》，*ACM Transactions on Embedded Computing Systems*，第18卷第2期，第1–23页，2019年。[133] Y. Zhu、M. Halpern 和 V. J. Reddi，《移动 Web 应用程序中节能服务质量（EQoS）的基于事件的调度》，载于*IEEE* 高性能计算机体系结构国际研讨会（*HPCA*）论文集，2015年，第137–149页。[134] V. Zyuban、D. Brooks、V. Srinivasan、M. Gschwind、P. Bose、P. N. Strenski 和 P. G. Emma，《流水线微处理器的功耗与性能综合分析》，*IEEE Transactions on Computers*，第53卷，2004年。