arXiv:2611.32731v1 [cs.SE] 24 Nov 2026

作为服务级目标的能耗：事务级能耗归因

Julian Legler[0000−0001−8715−0156]
指导教师：Stefan Tai

柏林工业大学，德国柏林 信息系统工程
julian.legler@tu-berlin.de

摘要。 服务工程师会根据功能需求以及可靠性、性能和可用性等质量属性评估每项变更。能耗很少被纳入此类评估。当变更导致能耗上升时，团队无法判断这种增加是否值得调查、是否值得接受，抑或两者皆非。对追踪、重试或共享组件的变更，可能会对能耗产生远超被修改组件本身的影响。本论文探讨如何使能耗目标成为可执行的工程工具，而不只是报告指标。站点可靠性工程（SRE）曾通过服务级指标、目标和响应策略，对可靠性提出了类似的解决方案。我们以 SRE 作为指导性研究愿景，研究这一类比在能耗领域何处成立、何处失效。例如，一个能耗目标可以规定在给定工作负载条件下每次成功事务的能耗上限。本论文将三个研究方向整合为一种经过实验评估的方法论。有效工作语义定义什么是成功事务。能耗归因将共享能耗和辅助能耗分配给各个事务。服务级工程决策则依据由此得到的能耗目标采取行动。现有研究涵盖归因方法、测量框架和可持续性目标，但大多停留在组件层面，且未与决策建立联系。我们此前关于服务级能耗建模的研究表明，辅助服务的能耗经常被忽视，而且难以测量。计划开展的实验将在受控改变可观测性配置和注入故障的条件下，对比汇总指标、请求归一化指标和基于有效工作的指标。事务级归因是否是做出这些决策所必需的，或者采用带有结果计数的服务级归因是否已经足够，仍是一个开放问题；这些实验旨在对此作出回答。

关键词： 能耗归因 · 有效工作 · 服务级别目标 · 站点可靠性工程· 微服务

1 引言

服务工程师每天都会对代码、配置和架构做出决策。回归测试会根据选定的功能需求评估这些变更，或
系统属性，但测试只能检测其所观测到的属性。设想一个团队提高追踪采样率以改善可观测性。这一变更可能会增加应用及其

可观测性后端中的计算资源、网络流量和存储开销。在额外工作使延迟和可用性仍处于各自目标范围内的部署中，仅凭这些指标无法表明存在能耗回归。有关监控开销的实证研究表明，其对能耗和性能的影响往往被忽视 [6, 10]。

检测到能耗增加还不够。哪些信息能够帮助团队判断，增强可观测性是否值得付出相应代价，确定谁能够采取应对措施，并评估应对效果？降低可观测性可能会减少能耗，但也会带来盲点。因此，能耗指标必须结合服务级别目标、工作负载，以及与整个系统其他质量属性之间的权衡来解读。

可靠性工程过去也面临类似问题。站点可靠性工程（SRE）通过定义服务级别目标，并将其与运营决策和职责关联起来，解决了这些问题 [3]。因此，我们以 SRE 作为指导性研究愿景。本论文探讨服务级别指标（SLI）、服务级别目标（SLO）和响应策略如何将能耗测量与决策联系起来。我们重点比较采用共享能耗归因的有效工作核算，与每请求能耗等更简单的请求归一化指标。现有文献已提供生成能耗信号的方法，以及利用这些信号进行服务级别决策的方法，但两者尚未建立联系。本论文的贡献在于提出一种方法，将有效工作语义、能耗归因和服务级别工程决策联系起来，并通过实验进行评估。预期结果将说明，能耗目标何时能成为可执行的工程工具，而非仅用于报告的指标。具体而言，研究将揭示何时需要事务级能耗归因来支持决策，何时简单指标就已足够。

2 问题与研究问题

如果同期请求率上升，即使进行了节能重构，也可能无法降低总能耗。每请求能耗降低，反过来可能掩盖完成同一用户任务所需的额外尝试次数。总能耗、每服务能耗、每服务器能耗、每次尝试能耗、每请求能耗、每小时能耗以及每个已接受结果的能耗等指标，分别回答不同的问题。如果不了解预期结果和总体工作负载，这些指标中的任何一个都可能产生误导。

我们将 事务 作为一个工作术语，用于指代由用户或其他系统发起的应用级活动或工作单元。它可能跨越多个服务。我们以组织为核算边界，并将其他组织的服务所完成的工作报告为未归属，因为我们至多只能看到它们导出的跨度，却没有关于这些跨度的能耗数据。在组织内部如何划定核算边界，尤其是对于空闲能耗和支撑系统而言，是我们模型面临的关键未决问题之一。

“成功事务”是我们对有效工作结果的工作术语。它指以满足应用功能和质量需求的方式完成预期活动，例如在时限内返回正确数据。什么样的事务算作成功事务仍是一个开放问题，答案可能因使用场景而异。重试是为实现成功事务而进行的尝试，而不是新的事务。

“中止或取消事务”是指已经启动但未成功完成的事务，原因可能是用户取消了事务，也可能是系统发生故障。因此，包含多次重试的失败尝试会增加能耗，但不计为成功事务。

每次成功事务能耗 是目前计划用于衡量面向服务的系统能效的指标。能源服务等级目标随后会对该指标设定上限，例如在规定的工作负载条件下，限定每次成功交易的最大能耗。第 4 节将讨论此类目标的候选形式。能否从追踪记录中推导出成功事务，或者产品负责人是否必须针对每项活动声明判定标准，仍是一个悬而未决的问题。我们预计，这些标准必须针对每个工作流分别声明，且不存在通用规则，但目前尚未对此进行测试。另一个假设是，只有跨多个服务才能判断有效工作，因为单个服务很少知道用户的意图是否已经实现。这就是我们称之为成功事务，而不是成功请求的原因。另一个有待解决的问题是，如何处理空闲能耗和支撑系统（例如可观测性后端）的能耗，以及应将这部分能耗分配给事务，还是单独报告。

因此，该模型需要一个能耗分配策略，明确如何在事务或服务之间分摊共享能耗，以及如何处理空闲能耗和支撑系统。

我们的总体问题是：如何使能耗目标成为可执行的工程工具，而不仅仅是报告指标。其技术核心在于：在面向服务的系统中，成功事务、共享能耗分配和估计不确定性在何种条件下会改变能耗回归判断与干预选择。我们将其分解为三个研究问题。

RQ1： 对于包含重试、失败和支撑性工作的服务活动，如何规定成功事务和核算边界?
RQ2： 归因值和回归判断对分配策略、工作负载组合及估计不确定性的敏感程度如何?
RQ3： 事务感知能耗指标和服务级别目标如何指导工程决策，以及在何种条件下更简单的指标已足够?

3 相关工作

已有研究已在能耗指标的语境下定义了有效工作。Guldner 等人提出的绿色软件度量模型（GSMM）将有效工作视为特定于应用程序的 [8]。只有当两个系统执行可比较的工作并产生等效结果时，才能进行比较。软件碳强度（SCI）规范已被标准化为 ISO/IEC 21031，该规范定义了每功能单位的运营碳排放和隐含碳排放，并将能耗作为其中一项输入 [7]。系统的所有组件都应使用相同的功能单位。SCI 特意不限定功能单位，以适应不同的使用场景。GSMM 和 SCI 都没有提供具体方法来判断哪些结果算作有效，也没有说明如何在面向服务的系统中分摊共享成本。

其他研究则将能耗归因于单个事务。我们自己的 GOXN 工作对每项服务的 CPU、内存、网络和存储能耗进行建模，其中包括辅助服务 [10]。Brunnert 和 Gutzy 扩展了 OpenTelemetry Java 代理（OTJAE），根据事务的资源活动估算其能耗，并计算每个事务的软件碳强度（SCI） [5]。Batista 等人的 Spanergy 将一项服务的能耗平均分配给同时运行的各个跨度 [2]。这种方法保留了所有活动跨度的能耗总量，但在这些跨度之外消耗的能耗却没有归因给任何对象。

Sharma 和 Fuerst 将整个系统的能耗拆分为无服务器函数调用的能耗，并计入共享硬件和软件成本 [12]。Schneider 和 Mattia 根据预留分配空闲能耗、根据使用情况分配动态能耗，然后沿服务链将共享服务的成本转嫁给消费者 [11]。Anand 等人展示了组件之间的权衡，并提出了每个请求的能耗预算 [1]。这支持端到端核算，但没有提供通用的结果定义或分配规则。这些方法各自采取了不同的核算决策。因此，一个事务的足迹取决于测量边界以及共享消耗的分配策略。我们尚未决定是基于其中一种方法构建，还是在 GOXN 之上自行实现归因。

《SRE 工作手册》将服务级目标与错误预算策略联系起来，其中明确了响应措施和负责人员 [3]。这一过程就是我们的研究愿景。

Botero 等人的 GreenSLAs 将服务的能耗表达为服务级目标，并将合同与灵活性请求、奖励和惩罚联系起来 [4]。Herrera 等人的 CASCA 将遥测数据和声明的服务级目标范围与运行时重新配置联系起来 [9]。在他们对媒体流服务的评估中，会根据质量和可持续性信号调整编码资源。Vitali 等人将年度碳预算分解为每小时的能耗预算 [14]。在这些预算范围内，他们选择组件版本、扩缩策略和可选功能，以平衡用户体验与收入。他们的仿真使用的是建模能耗而非测量值，也没有展示此类策略会如何影响真实服务团队的决策。

4 研究方法与方法论

SRE 使用错误预算来平衡变更速度与系统可靠性 [13]。我们希望迁移的是一种包含明确目标、测量窗口、响应策略和清晰职责的流程。这种方法是否适用于能耗，以及类比在哪些方面失效，是本研究的一个假设。与可靠性服务级目标的比较并非在所有方面都成立。可靠性服务级指标通常是良好事件与有效事件的比率。它的取值范围介于 0% 和 100% 之间，用于描述用户能够感知的症状。能耗测量本身不具备这两个特性。能耗没有自然上限，而且用户无法直接感知。能耗预算究竟应当重置、累积，还是应当表示为一个阈值，目前也尚不明确。谁应当设定能耗目标，仍是一个开放性问题。

我们目前认为能源服务等级目标有三种候选形式。第一种是强度阈值，例如在一个时间窗口内评估的每次成功事务的最大能耗。这种形式保留了 SRE 的时间窗口和响应策略。第二种是在既定工作负载条件下对能效的约束。第三种是违规

预算，即每个周期内允许超过阈值的时间窗口数量。这种形式在结构上最接近错误预算，因为它将能耗重新转化为优质和劣质时间窗口的计数。什么样的时间窗口算作优质或劣质，仍有待决定。哪种形式合适，以及能源预算这一概念本身是否有用，属于 RQ3 的研究内容。

明确成功事务和能耗目标的定义，也应明确各项响应分别由哪个角色负责。

RQ1. 我们计划为微服务系统中的选定工作流定义成功事务的标准，并在受控实验中记录尝试、失败和完成状态。目前，我们的 GOXN 工具在 Kubernetes 上运行 OpenTelemetry astronomy-shop 演示，并使用 Kepler 测量能耗。我们计划增加更多实验系统（SUE）和工作负载，以检验结果的普适性。我们使用分布式追踪将跨度关联成事务，并通过服务级测量补充共享能耗和辅助能耗。是否继续在 GOXN 中开展已启动的追踪采样率实验，尚未确定。

RQ2. 我们计划比较不同事务归因方法下的指标，例如总能耗、每次尝试能耗和每次成功事务能耗。实验一次只改变一个因素，其中一个因素是在注入故障时的重试行为。每次试验中，工作负载、请求组合和验收标准保持不变。后续实验可以采用其他实验系统、其他工作负载以及更复杂的事务。

随后，我们在此设置中实现事务级归因，并将其与服务级归因进行比较。我们检验归因值之和是否等于聚合测量结果，以及干预是否会带来可测量的变化。目标是区分可归因与不可归因的能耗部分，并证明归因值与系统级测量结果保持一致。

RQ3. 我们仍然每次只改变一个因素，并保持其他条件不变。目标是观察各因素如何影响在既定质量约束下作出的决策，以及事务级归因是否会导致与服务级归因不同的决策。一种做法是设定能耗预算，并观察归因方法如何改变为遵守该预算而采取的决策。

在测量噪声下检测受控回归，可能是这项评估面临的主要挑战。希望这能进一步证明，在现实条件下，归因方法在指导节能决策方面具有实际效用。

5 初步及正在进行的工作

我们已发表的 GOXN 研究提出了一种服务级能耗模型，涵盖 CPU、内存、网络、存储和辅助服务，并提供了实验工具 [10]。在评估的配置中，忽略网络和存储会导致辅助服务的能耗被低估，最高达 63%。可观测性变更的大部分能耗发生在服务调用图之外，消耗于存储和追踪后端。

GOXN 实验还表明，该环境中存在显著的测量噪声和变异性，这使得将能耗归因于单个服务和操作变得更加困难。通过谨慎的实验设计和重复测量，仍然得到了可用的结果。

READ 是我们正在开展的、将 GPU 能耗归因于大语言模型推理中各跨度的工作，该工作确认了两项可延续至本论文的观察结果。归因策略在如何处理空闲能耗方面存在差异，而分配的公平性与其物理准确性是两个相互独立的问题。共享推理本身不在本论文的范围内。

6 研究计划与讨论

本文是这一研究议程的首个书面表述，以 SRE 作为指导性研究愿景。这个想法源于发表于 ICSOC 2025 [10] 的 GOXN 论文中的观察与挑战，以及个人对 SRE 的兴趣。在接下来的六个月中，我们希望确定 astronomy-shop 工作流的结果与能耗核算语义，扩展 GOXN 以实现事务级归因，并建立可复现基线（RQ1）。分配策略与指标比较（RQ2）可以在基线稳定后立即开始，并将与第一阶段重叠。目标与响应策略评估（RQ3）取决于 RQ2 的结果，也是目前最不确定的部分。除 astronomy-shop 演示之外，我们还计划采用其他 SUE，以检验可推广性。通过一项正在开展的研究项目（SoftwareCampus）进行的实践者研究属于可选扩展，并非论文所必需。

我们目前看到的主要风险包括：难以获取跨服务边界的测量数据；归因模型复杂且难以验证；以及由于成功事务的定义各不相同，这些模型难以广泛适用于不同工作流和平台。

参考文献

1. Anand, V., Xie, Z., Stolet, M., De Viti, R., Davidson, T., Karimipour, R., Alzayat, S., Mace, J.: 与众不同：能源并不像其他指标。ACM SIGENERGY能源信息学评论 **3**(3), 71–77 (2023). https://doi.org/10.1145/3630614.3630627, https://stolet.github.io/papers/odd-one-out.pdf 2. Batista, C.P., Conan, D., Chabridon, S.: Spanergy：面向微服务的能量感知分布式追踪。载于：2026 IEEE第26届集群、云与互联网计算国际研讨会（CCGrid）论文集。第276–286页。IEEE（2026年5月）。https://doi.org/10 00036 3 1109 68966 2026 . /CCGrid . . . Beyer, B., Murphy, N.R., Rensin, D.K., Kawahara, K., Thorne, S.（编）：《网站可靠性工程手册：实施SRE的实用方法》。O'Reilly Media（2018）， https: //sre.google/workbook/implementing-slos/ 4. Botero, J.F., Klingert, S., Hesselbach-Serra, X., Falcone, A., Giuliani, G.: GreenSLAs：通过能量感知合同为数据中心提供能源消耗灵活性。载于：第二届智能电网与绿色信息技术系统国际会议论文集。第119–122页。SciTePress（2013 0004379401190122 5 10 5220 ）。https://doi.org/ . / . Brunnert, A., Gutzy, F.: 扩展OpenTelemetry Java自动检测代理，以发布绿色软件指标。Softwaretechnik-Trends **44**(4), 51–53 (2024), https://dl. gi.de/handle/20 45524 6 500 12116. . / . Dinga, M., Malavolta, I., Giamattei, L., Guerriero, A., Pietrantuono, R.: 基于Docker的系统中监控工具的能源与性能开销实证评估。载于：Monti, F., Rinderle-Ma, S., Ruiz Cortés, A., Zheng, Z., Mecella, M.（编）《面向服务的计算》。第181–196页。《计算机科学讲义》，Springer Nature Switzerland，Cham（2023 10 1007 978 3 031 48421 6_13 ）。7 https://doi.org/ . / - - - - . 绿色软件基金会：软件碳强度（SCI）规范。https://sci.greensoftware.foundation/ (2024)，版本1.1.0；访问日期：2026-10-10 8. Guldner, A., Bender, R., Calero, C., Fernando, G.S., Funke, M., Gröger, J., Hilty, L.M., Hörnschemeyer, J., Hoffmann, G.D., Junger, D., Kennes, T., Kreten, S., Lago, P., Mai, F., Malavolta, I., Murach, J., Obergöker, K., Schmidt, B., Tarara, A., De Veaugh-Geiss, J.P., Weber, S., Westing, M., Wohlgemuth, V., Naumann, S.: 绿色软件度量模型（GSMM）：用于评估软件产品和组件资源效率与能效的参考测量模型的开发与评估。Future Generation Computer Systems **155**, 402 033 9 418 2024 10 1016 2024 01 – ( ). https://doi.org/ . /j.future. . . . Herrera, J.L., Wang, D., Dustdar, S.: 用于可持续且智能地履行服务级目标和进行服务管理的基于微服务的平台（2026）。https://doi.org/10.48550/arXiv.2602.12875, https://arxiv.org/abs/2602.12875, 预印本 10. Legler, J., Werner, S., Borges, M.C., Tai, S.: 云原生微服务的服务级能源建模与实验。载于：《面向服务的计算：第23届国际会议，ICSOC 2025，中国深圳，2025年12月1–4日，论文集，第一部分》。《计算机科学讲义》，第16320卷，第230–247页。Springer Nature Singapore，新加坡（2026 10 1007 978 981 95 5012 8_17 ）。https://doi.org/ . / - - - - 11. Schneider, I., Mattia, T.: 云中的碳核算：一种在数据中心用户之间分配排放的方法（2024）。https://doi.org/10.48550/arXiv.2406.09645, https://arxiv.org/abs/2406 2024 12 09645, 1, 14 . 预印本；版本日期为2024年6月。Sharma, P., Fuerst, A.: 无服务器函数的可核算碳足迹与能源分析。载于：2024年ACM云计算研讨会论文集。第522–541页。ACM（2024）。https://doi.org/10.1145/3698038.3698531, https://doi.org/10. 1145/3698038.3698531

13. Sloss, B.T.：引言。载于：Beyer, B.、Jones, C.、Petoff, J.、Murphy, N.R.（编）：《网站可靠性工程：Google 如何运行生产系统》，第 1 章。O’Reilly Media（2016），https://sre.google/sre-book/introduction/ 14. Vitali, M.、Wiesner, P.、Kreutz, K.、Gandola, R.：自适应绿色云应用：通过近似计算平衡排放、收入和用户体验。《Future Generation Computer Systems》**176**，108143（2026年）。https://doi.org/10.1016/j.future.2025.108143，https://doi.org/10.1016/j.future.2025.108143