Deloitte.
Insights

随着生成式人工智能功耗日增，数据中心寻求更绿色可靠的能源解决方案

科技行业应优化基础设施、创新芯片设计，并与电力提供商合作，助力数据中心实现未来可持续发展。

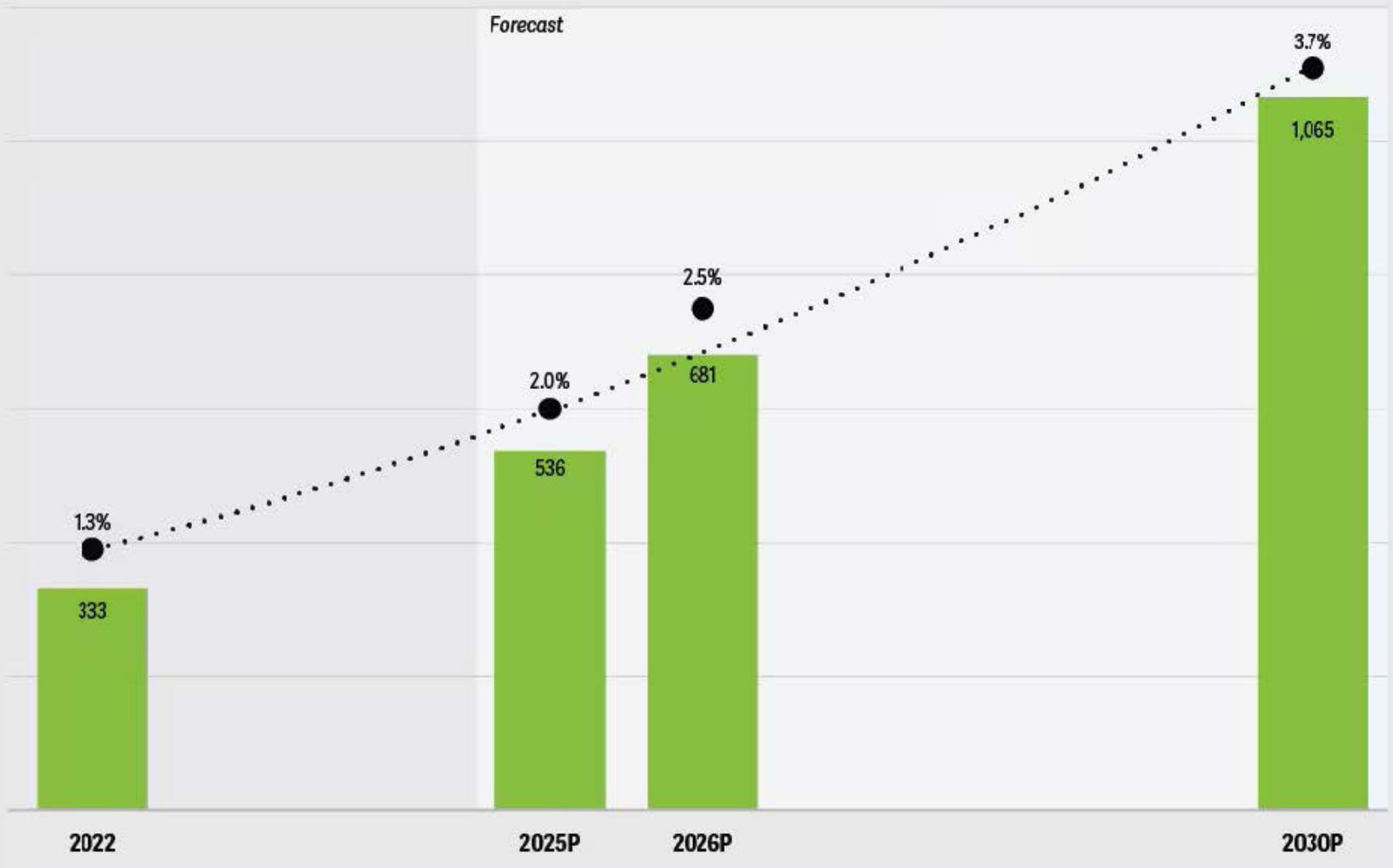
人工智能数据中心用电量将持续攀升，但事实上，数据中心用电量占全球电力需求的比重并不高。德勤预测，到2025年，数据中心用电量大约仅占全球用电量的2%，即536太瓦时（TWh）。然而，随着电力密集型生成式人工智能的训练和推理需求迅速增长，超过其他应用，预计到2030年全球数据中心的用电量将翻一番，约达1,065太瓦时（见图1）。[1]为保障数据中心供电并减少环境影响，许多公司正探索将数据中心的创新节能技术与更多无碳能源结合使用。

然而，发电和电网基础设施要满足人工智能数据中心激增的用电需求实属不易。由于电气化进程——运输、建筑和工业领域从化石燃料设备转向电力设备——以及其他因素，电力需求增长迅速。加之生成式人工智能的出现，导致电力需求增长超出预期。此外，数据中心往往对电力供应有特殊要求，需要具备高冗余度、高可靠性的全天候电力供应，并致力实现供电过程的碳中和。

由于多重变量的影响，要估算2030年及以后全球数据中心的用电量并非易事。据德勤评估，如果人工智能和数据中心的处理效率不断提升，到2030年全球数据中心的能耗水平将达约1,000太瓦时。然而，若预期的效率提升在未来几年内未能实现，则到2030年全球数据中心的能耗水平或将超过1,300太瓦时，这将直接影响到电力提供商，并阻碍气候中立目标的达成。[2]因此，未来十年里，推动人工智能创新和提高数据中心效率，将成为塑造可持续能源格局的关键。

图1

受能源密集型生成式人工智能模型的推动，预计2030年全球数据中心用电量将激增



注：P表示预测值。

资料来源：德勤基于公开资料来源以及与行业专家的讨论所做的分析。

分析方法：德勤基于美国能源信息署《2023年国际能源展望》中关于住宅、商业、工业和交通业终端用户总用电量的基础用电数据（参见表：按终端用户部门和燃料划分的能源消耗量），得出了2022年至2030年全球数据中心用电量（太瓦时）的估计值和预测值。德勤对数据中心用电量占全球总用电量比重的估计和预测，是基于对Semi Analysis、EPRI、高盛、彭博和Latitude Media等多方公开资料的分析，并通过与科技、能源和可持续发展行业专家的讨论予以进一步验证。

随着人工智能数据中心电力需求的日益增长，全球部分地区已面临发电和电网容量管理问题。[3]2023年至2026年间，全球数据中心关键组件（包括GPU 和CPU服务器、存储系统、冷却设备及网络交换机）所需电力预计将增长近一倍，到2026年将达96吉瓦（GW），其中仅人工智能运算就可能占用超40%的电力。[4]预计到2026年，全球人工智能数据中心的年用电量将达90太瓦时（约占届时全球数据中心预计用电量681太瓦时的七分之一），较2022年增长约十倍。[5]因此，生成式人工智能投资大幅推高了用电需求，例如，2024年第一季度全球人工智能数据中心的新增电力净需求约为2吉瓦，较2023年第四季度增长了25%，更是2023年第一季度的三倍多。[6]满足数据中心的用电需求颇具挑战，因为数据中心设施往往集中在特定区域（如美国），且其全天候电力需求将对现有电力基础设施造成负担。[7]

德勤预计，科技行业和电力行业将共同应对上述挑战，降低人工智能（尤其是生成式人工智能）对能源行业的影响。目前，许多大型科技公司和云服务提供商已着手进行无碳能源投资，并推动实现净零排放目标，[8]积极践行对可持续发展的坚定承诺。

超大规模云服务提供商拟大规模扩建生成式人工智能数据中心，以满足日益增长的客户需求

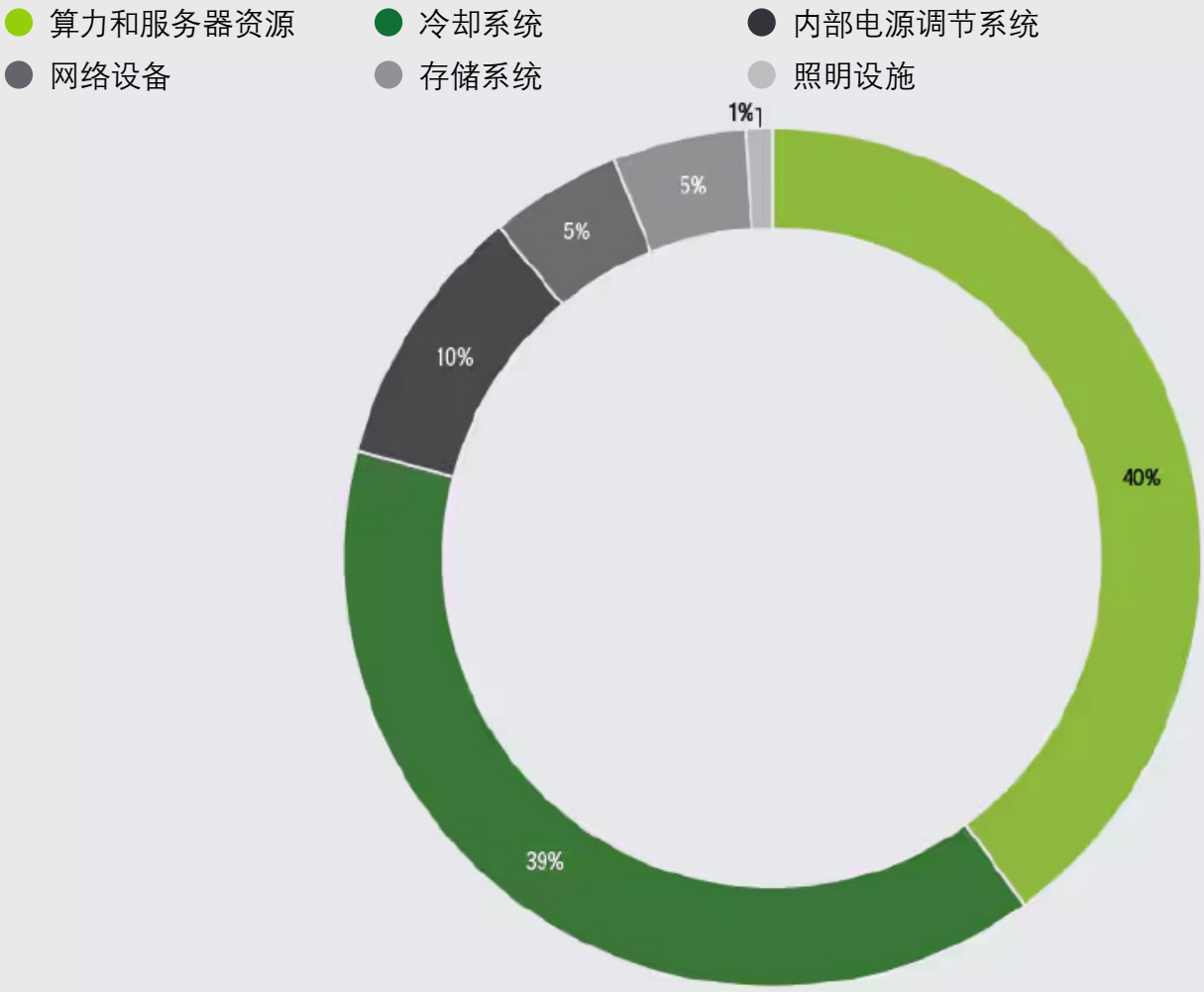
电力需求激增的主要原因在于超大规模云服务提供商计划拓展全球数据中心的容量。[9]随着人工智能（尤其是生成式人工智能）需求上升，企业和国家纷纷投入数据中心建设。各国政府也在打造主权人工智能能力，以保持其技术领先地位。[10]有数据显示，几家主要云服务提供商的数据中心建设投资已创历史新高，2024年资本支出约2,000亿美元，2025年或将超过2,200亿美元。[11]

此外，德勤调研报告《企业生成式人工智能应用现状》指出，截至目前，多数企业的人工智能应用尚处于试点和实验阶段。[12]但在探索生成式人工智能价值的过程中，受访企业已看到切实成果，因此打算在试点和概念验证之后迅速扩大其应用规模。随着生成式人工智能技术的成熟和使用量的增加，预计到2025年和2026年，云服务提供商的资本支出将继续保持高位。

数据中心的电力消耗主要集中在两大领域：算力和服务器资源（如服务器系统，约占总电力消耗的40%）以及冷却系统（约占38%~40%）。即便在人工智能数据中心，这两者亦是能耗最高的部分，持续推高整体电力消耗。此外，内部电源调节系统占8%~10%，网络通信设备和存储系统各占约5%，照明设施则通常仅占1%~2%（见图2）。[13]考虑到生成式人工智能的高电力需求，超大规模云服务提供商、数据中心运营商等数据中心提供商在设计数据中心时，应考虑采用替代能源、创新冷却技术和更节能的解决方案。目前，多项相关工作已在进行中。

图2

算力和冷却系统是人工智能数据中心能耗主因



资料来源：德勤基于ScienceDirect（2023年）和IEEE Access（2021年）等公开调研报告所做的分析。

生成式人工智能带来电力需求增长

自2023年以来，数据中心能耗因人工智能需求的激增而持续攀升。[14]部署先进的人工智能系统需要大量芯片和处理能力，而训练复杂的生成式人工智能模型更需要数千个GPU。

为此，支持人工智能和高性能计算的超大规模云服务提供商和大型数据中心运营商，必须构建高密度基础设施以保证算力。过去，数据中心主要依靠CPU，每块芯片的运行功率约在150瓦至200瓦之间。[15]2022年，用于人工智能的GPU运行功率为400瓦，而2023年用于生成式人工智能的最先进GPU运行功率已达700瓦；预计2024年，新一代芯片的运行功率将高达1,200瓦。[16]相较几年前的传统数据中心设计，新一代芯片（约8个）组成的刀片机架（每个机架10个刀片），每平方米占地面积的耗电量和发热量都更大。[17]截至2024年初，数据中心的机架功率普遍超过20千瓦。预计到2027年，单个服务器机架的平均功率密度将从2023年的36千瓦增至50千瓦。[18]

自生成式人工智能问世以来，以每秒浮点运算次数（FLOPS）衡量的人工智能总算力亦呈指数级增长。自2023年第一季度以来，全球人工智能总算力每季度增长50%~60%，并预计到2025年第一季度仍将保持这一增速。[19]但数据中心不仅以FLOPS来衡量算力，还以兆瓦时（MWh）和太瓦时（TWh）作为衡量标准。

包含数十亿参数的生成式人工智能大型语言模型及其巨额功耗

生成式人工智能的大型语言模型（LLM）日益精密，其参数（即实现人工智能学习和预测功能的变量）数量也在逐步增加。2021年至2022年间，问世的初始模型拥有1,000亿至2,000亿个参数，而到2024年中期，先进的大型语言模型已扩展至近两万亿个参数，能够解读和解码复杂的图像。[20]此外，全球正竞相发布十万亿参数级大型语言模型。由于人工智能必须经过训练和部署，更多的参数也将增加数据处理和算力需求。这将进一步加大对生成式人工智能处理器和加速器的需求，增加耗电量。

此外，大型语言模型的训练过程极为耗能。研究表明，对于参数量超过1,750亿的大型语言模型，单次训练的耗电量在324兆瓦时至1,287兆瓦时之间……而且模型通常需要多次训练。[21]

平均而言，生成式人工智能提示请求的电力消耗是普通网络搜索的10到100倍。[22]德勤预测，如果全球每天有5%的网络搜索采用生成式人工智能提示，则需要约20,000台服务器（每台服务器配备8个专用GPU）来满足这些请求，每台服务器平均耗电6.5千瓦，这意味着日均用电量可达3.12吉瓦时，年用电量达1.14太瓦时[23]——相当于约108,450个美国家庭一年的用电总量。[24]

数据中心用电需求对电力行业转型是一把双刃剑

电力行业已着手制定相应计划，以满足不断增长的用电需求。业内人士普遍预测，到2050年，部分国家的用电量将增加两倍之多。[25]但近期，由于数据中心用电需求激增，部分地区的用电量增速明显加快。多个国家曾做出预测，电气化进程推进、数据中心用电量增长以及整体经济增长将导致电力需求持续上升。但近期数据中心用电需求的激增或许仅是冰山一角，供电压力日渐凸显。[26]

随着电力公司对电网基础设施进行建设和升级，以及推进去碳化和数字化，电力行业步入长达数十年的转型期。在许多地区，电力公司还在加固设备，以应对日益严峻的气候事件，并保护网络免受与日增加的网络安全威胁。[27]部分国家的电网难以满足电力需求，尤其是对低碳或零碳电力的需求。2026年，美国数据中心的用电量预计将占全国总用电量的6% (约260太瓦时) 。[28]由于人工智能的发展，英国数据中心的电力需求或在短短十年内增长六倍。[29]到2026年，中国数据中心 (包括人工智能数据中心) 预计将占全国电力需求的6%。[30]此外，数据中心的电力需求，对于中国能源转型来说是一个新的变量，只有与清洁能源发电相结合，才有利于推动能源转型和双碳目标的实现。[31]

面对数据中心用电需求的不断增长，部分国家正在制定相关法规予以应对。例如，爱尔兰现有数据中心的用电量占全国总用电量的五分之一，随着人工智能数据中心不断涌现，该比例或将进一步提高；家庭用电甚至出现下降。[32]爱尔兰一度叫停接入电网的新数据中心建设计划，但后来改变了这一决策。[33]与爱尔兰一样，荷兰阿姆斯特丹亦暂停了新数据中心的建设，以支持城市的可持续发展。[34]新加坡则针对数据中心推出了全新的可持续发展标准，要求运营商逐步提高设施运行温度至26摄氏度或以上，以减少冷却需求并降低功耗，但这将缩短芯片的使用寿命。[35]

数据中心需求的紧迫性、地域集中度以及对全天候无碳能源的需求，使得科技公司和电力提供商面临更加严峻的挑战。电气化、制造业等领域也将产生新的用电需求。弗吉尼亚州北部是全球最大的数据中心市场，[36]本地公用事业公司Dominion Energy预计，未来15年弗吉尼亚州北部的电力需求将增长约85%，其中数据中心的用电需求将翻两番。[37]许多科技公司难以在短期内获得全天候无碳电力。电力提供商正想方设法，以满足需求并维持电力供应的可靠性和可负担性。除开发新的可再生能源和电池储能技术外，多家电力提供商还计划建设含碳的天然气发电厂，[38]这或将加大公用事业、州乃至国家实现脱碳目标的难度。[39]

尽管人工智能将消耗大量清洁能源，但亦有助于加速清洁能源转型：部分公用事业公司已开始利用人工智能改进天气预报、电力负荷预测、电网管理、可再生资产性能、风暴后恢复和野火风险评估等，从而降低电网的运行成本、提高运行效率和可靠性。[40]

数据中心冷却系统耗水量巨大

新一代CPU和GPU较上一代具有更高的热密度。与此同时，为迎合高性能计算和人工智能应用的强劲需求，部分服务器提供商在每个服务器机架上安装更多高能耗芯片。这样的高密度机架需水量更大，尤其是冷却生成式人工智能芯片。到2027年，人工智能数据中心的淡水需求量最高可达1.7万亿加仑。[41]如果一个超大规模数据中心计划采用空气冷却和饮用水蒸发冷却技术来控制过热，其年用水量将超过5,000万加仑 (约为制造14,700部智能手机的用水量[42]) ，且这些水量无法回流到含水层、水库或供水处。[43]

在普通数据中心中，仅空气冷却技术的耗电量就高达40%。因此，数据中心正在寻找传统空气冷却方式的替代方案，首选即是液体冷却技术，原因是液冷技术具有更高的热传导性能，有助于冷却高密度服务器机架，且相较于空气冷却方式，可减少高达90%的用电量。[44]液冷技术还可对服务器机架进行直接冷却，因此可支持50至100千瓦或更高功率的密集机架。[45]此外，液冷技术还有助于减少对传统冷却器的依赖。

尽管液冷技术在降低数据中心整体能耗方面颇具潜力，[46]但其应用仍处于早期阶段，尚未在全球范围内的人工智能数据中心广泛采用。[47]此外，水作为一种有限资源，其成本和供给状况预计将决定液冷技术的未来应用。

科技行业正朝着更可持续的解决方案和无碳能源的方向发展

科技巨头持续通过购电协议 (PPA) 或与可再生能源提供商签订长期合同，积极寻求可再生能源，以加速利用无碳能源为人工智能数据中心供电。[48]该等交易为可再生能源项目提供了融资支持。同时，科技公司还与电力提供商和创新企业展开合作，以帮助测试和推广有前景的能源技术，如先进的地热能、风能、太阳能和水电技术，甚至探索建立海底数据中心。

在某些地区，受当地电网的制约以及新能源与电池储能设施接入用时过长的影响，该等设施的并网进程出现延误。[49]在美国，由于用电需求旺盛且输电基础设施不足，常常导致电力供应延误，延误时间最长可达五年。因此，科技公司正积极寻求现场或离网的能源解决方案[50]，并投资于长时储能 (LDES) 和小型模块化核反应堆 (SMR) 等新技术，以应对该等挑战。同时，科技公司与公用事业公司计划开展协作，推动新型清洁能源技术的规模化应用，从而惠及更多客户，并加速电网脱碳进程。[51]其中许多研发项目、试点项目和其他清洁能源投资需要数年时间才能显现效益和商业化潜力。[52]例如，小型模块化核反应堆目前仍处于早期开发阶段，短期内可能非理想的零碳解决方案。[53]

科技行业在美国企业的可再生能源采购中始终占据主导地位。在2024年2月28日前的12个月内，美国企业达成近200项可再生能源采购交易，合同容量约19吉瓦，其中科技行业占68%以上。[54]同样，印度的超大规模云服务提供商和数据中心运营商亦日益依赖于风能和太阳能为其数据中心供电。[55]若无这些采购合同，许多可再生能源项目恐难落地。[56]

因此，科技行业在为清洁能源技术提供资金支持、推动其规模化发展方面将继续发挥重要作用。科技行业不仅直接与创新企业和可再生能源生产商合作，还与公用事业公司合作。[57]重要的是，创新企业和电力行业通常无法比肩科技行业的资金实力，因此科技公司如何注资以助推清洁能源转型，显得尤为关键。

中国通过能源转型与优化资源配置助力数据中心可持续发展

目前，中国数据中心耗电量约占全社会用电量的3%。其中，AI驱动的数据中心将成为重要的能源消耗来源，随着生成式AI技术的不断发展和大规模部署，未来数据中心的电力需求将呈现几何级增长。预计到2025年中国数据中心用电量将突破4,000亿千瓦时，占全社会用电量4.1%。因此推动清洁能源的应用以及优化数据中心的能效成为当务之急。中国的新可再生能源计划强调了在数据中心建设中逐步增加可再生能源的使用比例，支持在有冷水资源的国家枢纽节点建设数据中心，并逐步对旧基站和分散的小型数据中心进行绿色技术升级，并通过“东数西算”工程优化资源配置。未来，中国将更加积极应对平衡AI驱动的电力需求激增与能源可持续发展的挑战，进一步促进算力与电力协同，推动数据中心可持续发展。

小结

广大科技行业、超大规模云服务提供商、数据中心运营商、公用事业公司和监管机构应如何行动，以推动生成式人工智能的可持续发展？以下是超大规模云服务提供商和科技行业需考量的几点，它们与德勤全球在2021年关于云迁移预测中提出的观点不谋而合。[58]尽管市场需求驱动因素或有转变，且变化节奏加快，但基本需求大致相同，所以可持续发展的基本要求和重点依旧不变：控制数据中心不断增长的能源需求，并寻求更可持续的方式为人工智能（尤其是生成式人工智能）供电。

1. 提升生成式人工智能芯片能效：目前，新一代人工智能芯片可在90天内完成人工智能训练，耗电8.6吉瓦时，不到上一代芯片在同等情形下所需能耗的十分之一。[59]芯片公司应与半导体生态深化合作，以聚焦并提高每瓦特性能，让未来芯片能在更低能耗情况下训练出远超现有规模的人工智能系统。

2. 优化生成式人工智能应用，实现数据处理边缘化：评估数据中心与边缘设备在训练和推理方面的能耗差异，据此调整数据中心的设备配置。边缘计算不仅适用于时间敏感型应用，还能有效处理敏感数据和满足高隐私需求。边缘设备还有助于节省网络和服务器带宽，将生成式人工智能的工作负载导向本地和近地或主机托管设备，只将必要的人工智能工作负载传输到数据中心。[60]

3. 改变生成式人工智能算法，调整人工智能工作负载：我们是否应一味追求建立更大的基础模型（例如，万亿参数级模型），还是转而使用更具可持续性的小模型？目前，初创企业正在开发端侧的多模态人工智能模型，该等模型无需依赖高能耗的云端计算。[61]客户应根据实际业务需求，精准调整人工智能工作负载，并选取适当的人工智能模型（包括现成模型，仅在需要时才进行训练），以最大限度地减少能耗。此外，CPU可根据人工智能推理的具体需求（例如，实时推理和低延迟推理）充分发挥优势、提升效率。[62]

4. 建立战略合作伙伴关系，满足地方和集群级人工智能数据中心的需求：部分中小型客户（如大学）可能难以获得足够的生成式人工智能数据中心资源，因此应与专业数据中心运营商和云服务提供商开展合作，后者专注于为小型HPC GPU集群主机托管提供HPC解决方案。[63]因此，数据中心应主动监测自身使用情况和资源可用性，发掘潜在商机和需求洼地，以满足短期主机托管服务需求。

5. 与多方利益相关者和行业合作，对整体环境产生积极影响：超大规模云服务提供商及其客户、第三方数据中心运营商、主机托管服务提供商、电力提供商、地方监管机构和市政当局以及房地产公司等生态系统参与者，应围绕商业、环境和社会效益展开持续对话。[64]合作内容应涵盖多个方面，包括：确定潜在的战略主机托管服务需求（即数据中心公司向一家或多家公司出租计算和服务器资源）、评估冷却需求（如液冷系统的适当温度）、制定热能和废水管理解决方案以及回收利用策略。例如，在欧洲，已有数据中心运营商利用余热为附近泳池供暖。[65]电力提供商应与科技行业开展更密切合作，以保障数据中心的能源供给，并确定科技公司如何资助和推广新能源技术，这对推动清洁能源上网尤为重要。

长远来看，超大规模云服务提供商和电力提供商在提升数据中心（包括专为生成式人工智能而建的数据中心）的无碳能源利用比重、满足其电力需求方面所做的全面努力预计将取得成效。

By

Karthik Ramachandran
India

Duncan Stewart
Canada

Roger Chung
China

Kate Hardin
United States

Gillian Crossan
United States

尾注

1. Deloitte analysis based on publicly available information sources and conversations with industry specialists. We used base electricity consumption data from the US Energy Information Administration's (EIA) International Energy Outlook 2023 data on total electricity usage across residential, commercial, industrial, and transportation end uses (a reference to US Energy Information Administration, "*Table: Delivered energy consumption by end-use sector and fuel*," accessed Nov. 4, 2024) to arrive at estimates and prediction values for global data centers' electricity consumption (TWh) between 2022 and 2030. Our estimates and projections for data centers' percent electricity consumption of global total are based on our research of multiple publicly available sources including SemiAnalysis, EPRI, Goldman Sachs, Bloomberg, and Latitude Media, and further validated based on our conversations with subject matter specialists in the areas of technology, energy, and sustainability. Total energy consumption by end-use sector and fuel (as noted from the aforementioned table from EIA's International Energy Outlook 2023 data), globally, is estimated and forecast at 26,787 TWh in 2025, 27,256 TWh in 2026, and 29,160 TWh in 2030— increasing from 25,585 TWh back in 2022.

2. As noted in endnote 1 above, we arrived at 2022 to 2030 data, estimates, and predictions based on a combination of in-depth secondary research of multiple publicly available sources, and validated further from our discussions with subject matter specialists. Also, see Prof. Dr. Bernhard Lorentz, Dr. Johannes Trüby, and Geoff Tuff, "*Powering artificial intelligence,*" Deloitte Global, November 2024."

3. One-fifth of Ireland's electricity is consumed by data centers, and this is expected to grow, even as households are lowering their electricity use. To read further, see: Chris Baraniuk, "*Electricity grids creak as AI demands soar*," BBC, May 21, 2024.

4. Dylan Patel, Daniel Nishball, and Jeremie Eliahou Ontiveros, "*AI data center energy dilemma: Race for AI data center space*," SemiAnalysis, March 13, 2024.

5. Ibid.

6. Data center BMO report, Communications Infrastructure, "1Q24 data center leasing: Records are made to be broken," April 28, 2024; Moreover, due to strong demand from cloud providers and AI workloads, the data center primary market supply in the United States alone was up 26% year over year to 5.2 GW in 2023, and more are under construction. See further: CBRE, "*North America data center trends H2 2023*," March 6, 2024.

7. Lisa Martine Jenkins and Phoebe Skok, "*Mapping the data center power demand problem, in three charts*," Latitude Media, May 31, 2024.

8. Based on our analysis of multiple publicly available information and reports from what companies self-report, and further validated from third-party sources.

9. For context, hyperscalers are large cloud service providers and data centers that offer huge amounts of computing and storage resources typically at enterprise scale. See: Synergy Research Group, "*Hyperscale operators and colocation continue to drive huge changes in data center capacity trends*," Aug. 7, 2024.

10. Yifan Yu, "*AI's looming climate cost: Energy demand surges amid data center race*," Nikkei Asia, June 12, 2024.

11. Data center BMO report, Communications Infrastructure, "1Q24 data center leasing: Records are made to be broken," April 28, 2024. Further, Deloitte analysis based on information from select tech companies' publicly available sources such as earnings releases and Dell'Oro Group's market research data on data center IT capital expenditure shows that if we consider the capital expenditure spending of other data center providers, including third-party operators and outsourced cloud service providers, data centers' aggregate capital expenditure spending could be at least US$250 billion in 2025. See: Baron Fung, "*Market research on data center IT capex*," Dell'oro Group, accessed Nov. 4, 2024.

12. Nitin Mittal, Costi Perricos, Brenna Sniderman, Kate Schmidt, and David Jarvis, "*Now decides next: Getting real about generative AI*," Deloitte's State of Generative AI in the Enterprise quarter two report, Deloitte, April 2024.

13. Deloitte analysis based on publicly available research reports including: Wania Khan, Davide De Chiara, Ah-Lian Kor, and Marta Chinnici, "*Advanced data analytics modeling for evidence-based data center energy management*," *Physica A* 624, 2023; Kazi Main Uddin Ahmed, Math H. J. Bollen, and Manuel Alvarez, "*A review of data centers energy consumption and reliability modeling*," in *IEEE Access* 9, 2021: pp. 152536–152563.

14. Tom Dotan and Asa Fitch, "*Why the AI industry's thirst for new data centers can't be satisfied*," *The Wall Street Journal*, April 24, 2024.

15. Noam Brousard, "*Examining the impact of chip power reduction on data center economics*," Semiconductor Engineering, March 12, 2024.

16. Based on our analysis of multiple publicly available sources including: Michael Studer, "*The energy challenge of powering AI chips*," Robeco, Nov. 6, 2023; Agam Shah, "*Generative AI to account for 1.5% of world's power consumption by 2029*," HPCwire, July 8, 2024.

17. From our study and analysis of select gen AI data center chip solutions offered by major AI chip vendors, further corroborated with publicly available third-party sources including: Beth Kindig, "*AI power consumption: Rapidly becoming mission-critical*," *Forbes*, June 20, 2024.

18. Jones Lang LaSalle, "*Data centers 2024 global outlook*," Jan. 31, 2024; Doug Eadline, "*The gen AI data center squeeze is here*," HPCwire, Feb. 1, 2024; Per IDC, besides graphics processing unit, servers, data centers also need to grapple with a corresponding growth in storage capacity, which is likely to double between 2023 and 2027 to reach 21 zettabytes in 2027. See: John Rydning, "*Worldwide Global StorageSphere forecast, 2023 to 2027: Despite decreased petabyte demand near term, the installed base of storage capacity continues to grow long term*," IDC Corporate, May 2023.

19. Patel, Nishball, and Eliahou Ontiveros, "*AI data center energy dilemma*."

20. Sean Michael Kerner, "*What are large language models?*" TechTarget, May 2024; Yu, "*AI's looming climate cost*."

21. Alex de Vries, " *The growing energy footprint of artificial intelligence*," *Joule* 7, no. 10 (2023): pp. 2191–2194.

22. Eren Çam, Zoe Hungerford, Niklas Schoch, Francys Pinto Miranda, and Carlos David Yáñez de León, "*Electricity 2024: Analysis and forecast to 2026 report*," International Energy Agency, accessed Nov. 4, 2024.

23. Deloitte analysis based on publicly available reports and sources including: de Vries "*The growing energy footprint of artificial intelligence*," pp. 2191–2194.

24. Deloitte analysis based on data related to energy use and electricity consumption in homes in the United States. See: US Energy Information Administration, "*Use of energy explained*," accessed Dec. 18, 2023.

25. Darren Sweeney, "*Utility execs prepare for 'tripling' of electricity demand by 2050*," S&P Global, April 19, 2023.

26. Robert Walton, "*US electricity load growth forecast jumps 81% led by data centers*," Utility Dive, Dec. 13, 2023.

27. Aaron Larson, "*How utilities are planning for extreme weather events and mitigating risks*," *POWER*, March 13, 2024.

28. Çam, Hungerford, Schoch, Miranda, and de León, "*Electricity 2024*."

29. Baraniuk, "*Electricity grids creak as AI demands soar*."

30. Yu, "*AI's looming climate cost*."

31. Data on China's energy use and CO2 emissions sourced from International Energy Agency, accessed September 25, 2024. See: International Energy Agency, "*China's energy use*," accessed Nov. 4, 2024; International Energy Agency, "*China's CO2 emissions*," accessed Nov. 4, 2024.

32. Baraniuk, "*Electricity grids creak as AI demands soar*."

33. Paul O'Donoghue, "*Build it and they will hum: What next for Ireland and data centers?*" *The Journal*, Sept. 2, 2024.

34. Hosting Journalist, "*City of Amsterdam puts halt to new data center construction*," Dec. 21, 2023.

35. With every 1C increase, operators could save 2% to 5% on the energy they use for cooling equipment. To read further, see: Inno Flores, "*Singapore unveils green data center road map amid AI boom that strains energy resources*," Tech Times, May 30, 2024.

36. Julie R. Peasley, "*Ranked: Top 50 data center markets by power consumption*," Visual Capitalist, Jan. 10, 2024.

37. Whitney Pipkin, "*Energy demands for Northern Virginia data centers almost too big to compute*," Bay Journal, June 18, 2024.

38. Zach Bright, "*Southeast utilities have a 'very big ask': More gas*," E&E News, Jan. 22, 2024.

39. Ibid.

40. Robert Walton, "*AI is enhancing electric grids, but surging energy use and security risks are key concerns*," Utility Dive, Oct. 23, 2023.

41. Karen Hao, "*AI is taking water from the desert*," *The Atlantic*, March 1, 2024.

42. Deloitte analysis based on publicly available information sources including: Jennifer Billock, "*Photos: How much water it takes to create 30 common items*," North Shore News, Jan. 19, 2023.

43. Hao, "*AI is taking water from the desert*"; One case in point is China—where its data centers' annual water consumption is expected to increase from around 1.3 billion cubic meter as of 2023 to over 3 billion cubic meter by 2030. To read further, see: Yu, "*AI's looming climate cost*."

44. Eadline, "*The gen AI data center squeeze is here*."

45. Diana Goovaerts, "*Data center operators want to run chips at higher temps. Here's why.*" Fierce Network, June 11, 2024.

46. Scott Wilson, "*Is immersion cooling the answer to sustainable data centers?*" Ramboll, Dec. 13, 2023.

47. David Eisenband, "*100+ kW per rack in data centers: The evolution and revolution of power density*," Ramboll, March 13, 2024; Direct-to-chip cooling (also known as cold plate liquid cooling or direct liquid cooling) cools down servers by distributing heat directly to server components, while, immersion cooling involves submerging servers and components in a liquid dielectric coolant that also helps prevent electric discharge.

48. Based on Deloitte's analysis of developments and announcements from select major cloud hyperscalers and tech companies— and information gathered from publicly available sources (time period: 2023 and first three quarters of 2024).

49. Joseph Rand, Nick Manderlink, Will Gorman, Ryan Wiser, Joachim Seel, Julie Mulvaney Kemp, Seongeun Jeong, and Fritz Kahrl, "*Queued up: 2024 edition*," Lawrence Berkeley National Laboratory, April 2024.

50. Based on Deloitte's analysis of developments and announcements from select major cloud hyperscalers and tech companies— and information gathered from publicly available information sources between the first quarter of 2023 and the third quarter of 2024.

51. Julian Spector, "*Duke Energy wants to help Big Tech buy the 24/7 clean energy it needs*," Canary Media, June 11, 2024.

52. For example, it's not easy to submerge and drop a 1,300-ton data center unit underwater, especially since it demands special equipment to withstand pressure and corrosion caused by seawater. Moreover, there are concerns related to its impact on marine life.

53. David Schlissel and Dennis Wamsted, "*Small modular reactors: Still too expensive, too slow, and too risky*," Institute for Energy Economics and Financial Analysis, May 2024.

54. Deloitte's analysis of data and information gathered from multiple reports from S&P Global Market Intelligence, published during March and August 2024.

55. Manish Kumar, "*India's data center boom opens up a fresh segment for green developers*," Saur Energy International, July 1, 2024.

56. Naureen S. Malik and Bloomberg, "*With AI forcing data centers to consume more energy, software that hunts for clean electricity across the globe gains currency*," *Fortune*, Feb. 25, 2024.

57. Based on Deloitte's analysis of developments and announcements from select major cloud hyperscalers, tech companies, and power and utility players—on publicly available information sources during 2023 and the first three quarters of 2024.

58. Duncan Stewart, Nobuo Okubo, Patrick Jehu, and Michael Liu, "*The cloud migration forecast: Cloudy with a chance of clouds*," *Deloitte Insights*, Dec. 7, 2020.

59. Wylie Wong, "*Nvidia launched next-generation Blackwell GPUs amid AI 'arms race'*," Data Center Knowledge, March 19, 2024; For instance, Nvidia notes that it can train a very large AI model using 2,000 Grace Blackwell chips in 90 days, consuming 4 MW power. In comparison, it would take as much as 8,000 of the previous generation chips to do the same work within the same time, consuming 15 MW power.

60. To read further, see section "Generative AI comes to the enterprise edge: 'On prem AI' is alive and well" in our 2025 TMT Predictions chapter on "*Updates*"; Additionally, see:
Sabuzima Nayak, Ripon Patgiri, Lilapati Waikhom, and Arif Ahmed, "*A review on edge analytics: Issues, challenges, opportunities, promises, future directions, and applications*," *Digital Communications and Networks* 10, no. 3 (2024): pp. 783–804.

61. Yu, "*AI's looming climate cost*."

62. Luke Cavanagh, "*GPUs vs. CPUs in the context of AI and web hosting platforms*," Liquid Web, Aug. 20, 2024.

63. Eadline, "*The gen AI data center squeeze is here*."

64. Goovaerts, "*Data center operators want to run chips at higher temps. Here's why*."

65. Baraniuk, "*Electricity grids creak as AI demands soar*."

致谢

The authors would like to thank **Dilip Krishna, Marlene Motyka, Jim Thomson, Adrienne Himmelberger, Thomas Schlaak, Freedom-Kai Phillips, Johannes Truby, Clement Cabot, Negina Rood, Ankit Dhameja, Suzanna Sanborn,** and **Akash Cha erji** for their contributions to this article.

Cover image by: **Jaime Austin;** Getty Images, Adobe Stock

关于德勤

Deloitte（“德勤”）泛指一家或多家德勤有限公司，以及其全球成员所网络和它们的关联机构（统称为“德勤组织”）。德勤有限公司（又称“德勤全球”）及其每一家成员所和它们的关联机构均为具有独立法律地位的法律实体，相互之间不因第三方而承担任何责任或约束对方。德勤有限公司及其每一家成员所和它们的关联机构仅对自身行为承担责任，而对相互的行为不承担任何法律责任。德勤有限公司并不向客户提供服务。请参阅www.deloitte.com/cn/about了解更多信息。

本通讯中所含内容乃一般性信息，任何德勤有限公司、其全球成员所网络或它们的关联机构并不因此构成提供任何专业建议或服务。在作出任何可能影响您的财务或业务的决策或采取任何相关行动前，您应咨询合资格的专业顾问。

我们并未对本通讯所含信息的准确性或完整性作出任何（明示或暗示）陈述、保证或承诺。任何德勤有限公司、其成员所、关联机构、员工或代理方均不对任何方因使用本通讯而直接或间接导致的任何损失或损害承担责任。

CQ-001CN-25