工程与应用

DeepSeek背景下数据中心基础设施演进趋势与技术创新

刘洪，孙丽玫，朱丽，姜宇光，孙立峰，赵金铭，李雅婷
（中国移动通信集团设计院有限公司，北京 100080）

摘　要：DeepSeek技术创新颠覆了传统大模型的高投入模式，引发算力发展理念变革，从而对数据中心技术设施架构产生不可忽视的影响。基于需求牵引视角，系统分析了DeepSeek技术创新对算力需求的影响，并据此明确了数据中心基础设施演进趋势，提出集中节点与分布式节点的分层解决方案和未来发展建议。研究结论为数据中心基础设施规划、建设及运营方更好地适应人工智能技术发展及前瞻性建设数据中心基础设施，提供理论参考与实践指引。

关键词：智算推理；数据中心基础设施；工程产品化；产品工程化；运维智能化
中图分类号：TN915；TP393.4
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2026046

Evolution trend and technological innovation of data center infrastructure in the context of DeepSeek

Liu Hong, Sun Limei, Zhu Li, Jiang Yuguang, Sun Lifeng, Zhao Jinming, Li Yating
China Mobile Communications Group Design Institute Co., Ltd., Beijing 100080, China

Abstract: DeepSeek’s technological innovation has disrupted the high-investment model of traditional large models, triggering a transformation in the concept of computing power development and thus exerting an undeniable impact on the technical infrastructure architecture of data centers. From the perspective of demand traction, the impact of DeepSeek’s technological innovation on computing power demand was systematically analyzed. Based on this, the evolution trend of data center infrastructure was clarified, and a hierarchical solution of centralized nodes and distributed nodes as well as future development suggestions were proposed. The research conclusion provides theoretical references and practical guidance for the planning, construction and operation parties of data center infrastructure to better adapt to the development of artificial intelligence technology and build data center infrastructure in a forward-looking manner.

Key words: intelligent computing reasoning, data center infrastructure, engineering productization, product engineering, intelligent operation and maintenance

收稿日期：2025−04−30；**修回日期：**2025−09−23
通信作者：李雅婷，sjzwenyu@163.com
基金项目：北京市自然科学基金资助项目（No.L257012）
Foundation Item: The Beijing Natural Science Foundation (No.L257012)

0 引言

2025年1月，推理模型DeepSeek-R1以极低成本和受限算力实现了与OpenAI同级的性能，颠覆了传统人工智能（artificial intelligence，AI）高投入模式，引发全球科技领域的广泛关注。DeepSeek推动了开源生态建设，引爆了AI+应用的快速发展，激发了算力的泛在需求，从而对数据中心产业发展产生了深远影响，推动数据中心基础设施进入新的发展阶段。因此，深入分析DeepSeek案例，解析AI时代的创新规律和发展特征，探究算力需求变化趋势，进而提出数据中心基础设施未来演进方向与发展建议，不仅具有重要的理论价值，而且为推动我国数据中心基础设施创新发展提供了新的观察视角。

1 激增的AI算力需求与泛在化的算力部署

1.1 智能应用加速涌现，算力需求快速增长

DeepSeek通过开源模型架构、训练方法和核心技术要素，构建了包含200多项算法专利的开源知识库，吸引了全球开发者共同参与，加速技术迭代。这种开放模式催生了“开放底层架构+模块化开发平台”的创新生态，使企业能够基于自身需求灵活应用DeepSeek技术，加速产业数智化升级，催生蓬勃的产业应用。根据彭博社报道，截至2025年2月，DeepSeek应用位列全球140多个国家（地区）的应用商店榜首，显示出强大的市场影响力，超过200家头部企业完成了DeepSeek技术集成部署，覆盖能源、通信、汽车、金融等关键领域[1]。这种广泛的产业应用直接推动了对算力需求的增长。这一现象体现了“杰文斯悖论”[2]在AI领域的应用，即算力效率的提升降低了单位计算成本，反而刺激了总体算力需求的增长[3-4]，导致能源与基础设施资源总消耗上升。AI领域杰文斯悖论示意图如图1所示。

图1 AI领域杰文斯悖论示意图

近期，中国企业加速生成式人工智能布局和投入，智能算力发展水平增速高于预期。国际数据公司（IDC）调研结果显示，目前42%的中国企业已经开始进行大模型的初步测试和重点概念验证，17%的企业已经将AI技术引入生产阶段[5]，并应用于实际业务中，在未来18个月内，硬件升级将成为企业的首要投资目标。在旺盛的市场需求、丰富的应用场景的驱动下，中国智能算力规模呈现快速增长态势。

1.2 算力需求结构重塑，算力部署泛在化

DeepSeek的技术创新引发了行业对算力发展理念的转变，从追求算力规模转向注重算力性能效率和整体架构成本优化，推动了AI应用普惠、算力需求结构重塑和算力泛在化部署。算力需求结构变化主要体现在3方面：训推结构转变、边缘算力扩张和国产算力需求攀升。

首先，DeepSeek基于思维链技术，通过拆分问题提高AI回答问题的质量[6]，从而实现模型推理能力和可解释性的大幅提升，驱动算力需求从“预训练”走向“预训练+后训练+推理”，并呈现向“推理侧倾斜”的发展趋势。据IDC预测，未来5年国内训练、推理算力年复合增速分别为50%和190%，2028年推理算力规模将超过训练算力，AI服务器工作负载中推理占比将达到73%[7]。

其次，DeepSeek使用知识蒸馏技术，从复杂的大型“教师模型”中提取知识并转移到精简的小型“学生模型”中[8]，使得模型部署实现轻量化，对基础设施最小规模的要求大幅降低，激发了AI应用边缘部署需求。边缘智能算力能够有效满足实时性要求高的任务需求，通过本地化数据

处理降低网络时延、保障数据隐私。

最后，与其他主流模型相比，DeepSeek最大的特征在于构建了高性能的混合专家（mixture of experts，MOE）架构，显著减少了所需的参数规模，成功降低了大模型对高端芯片的依赖度，能够与国产芯片架构适配，从而推动了国产芯片需求的提升。在全球贸易摩擦加剧的背景下，芯片自主可控逻辑增强，国产芯片份额有望进一步提升。技术层面，目前主流国产芯片均已兼容适配DeepSeek技术，国产芯片的生态劣势得到一定弥补，受益于推理算力需求的爆发，国产芯片需求量将大幅提升。

综上，在全球范围内AI基础设施建设持续加速，美国等各国加大投资，而国内市场在DeepSeek技术的催化下正加速技术迭代和基础设施升级，AI推理应用的爆发，将推动算力资源泛在化部署。

2　智能推理资源对数据中心基础设施的需求

2.1　智能推理技术发展趋势

随着DeepSeek的推出，其优异的性能引发了全球范围内的应用热潮。用户量和算力需求迅速增长导致算力不足，DeepSeek官方网站服务频繁卡顿。为了快速提供算力并提升推理服务体验，产业界正在推动异构算力、训推一体、云计算和云智算融合等技术的发展，以应对算力需求爆发的挑战。

在处理大规模模型和应对多样化的应用场景时，单一硬件平台已难以兼顾高能效、低时延、低成本等多重需求，算力瓶颈问题日益凸显。异构算力融合架构通过整合中央处理器（central processing unit，CPU）、图形处理单元（graphics processing unit，GPU）、神经网络处理器（neural processing unit，NPU）、现场可编程门阵列（field programmable gate array，FPGA）、专用集成电路（application specific integrated circuit，ASIC）等不同类型的计算单元，能够针对不同的推理任务进行精细化的资源分配和加速，从而显著提升计算效率并降低能耗[9]。例如，在需要进行大规模内容或图像生成时，通常会选择GPU进行推理计算；对于相对简单的推理任务，如语音识别等，CPU有时也能成为高效的推理引擎；FPGA和ASIC则在特定场景下展现出卓越的推理能力。异构算力融合化架构创新，突破了算力“性能–能效”瓶颈。

与传统模型训练和推理分离部署不同，训推一体化将模型训练与推理过程深度融合，从而提升整体效率和性能[10]。训推一体化的资源部署，通过共享模型参数和优化训练数据集，显著降低了推理阶段的计算负载；通过简化从模型开发到推理应用的流程，合并训练与推理硬件资源池，有效减少了数据迁移开销，规避了训练节点和推理节点间数据传输带来的潜在风险。例如，基于某国产生态的推理全流程工具链，可以实现训练与推理状态的无缝切换，提升模型部署效率。

DeepSeek-V3发布后，各类用户争相在本地部署DeepSeek，云服务商也纷纷加大投入，推动云计算向云智算升级。为满足智算、通算、存储等资源互通，以及智算资源的多样性、泛在化和可扩展性等要求，智算资源可与传统云计算深度融合[11]。智算与通算统一规划、共池建设，提供一体化算网资源、全栈式开发环境、一站式模型服务、多样化场景应用的新型云服务，使云计算从以CPU为主的“云计算”向以GPU为主的“云智算”演进。

2.2　推理资源部署对数据中心基础设施的需求分析

基于网络时延需求，AI推理应用场景主要分为在线推理和离线推理两种类型。在工业物联网、金融风控等实时或近实时场景下，为了即时处理输入数据并迅速返回推理结果，这类对低时

延和高响应速度有严格要求的应用场景被定义为在线推理。与此相对，批量生成、数据清洗等周期性任务或对时延不敏感的场景，则可以在非实时环境下对批量数据进行处理和分析，这类场景被称为离线推理。因此，为了高效支撑各类推理应用的动态需求，数据中心基础设施应做好能力规划，合理布局，弹性部署，满足日益提升的高密功率需求。

2.2.1 集中节点和分布式节点协同部署

AI算力资源的集中节点和分布式节点协同部署，将提高计算效率，满足不同场景的算力需求。集中节点和分布式节点协同部署示意图如图2所示。

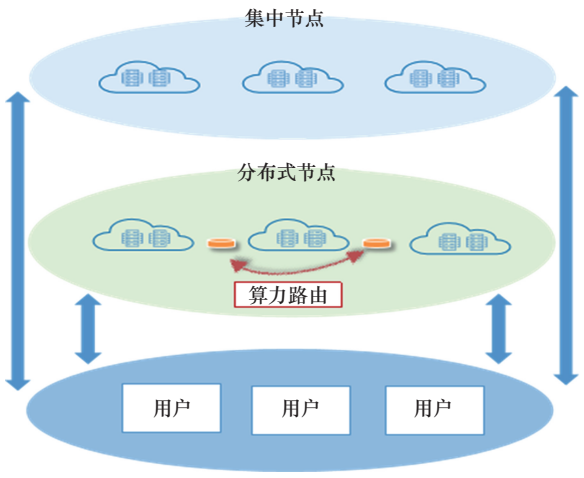

图2 集中节点和分布式节点协同部署示意图

集中节点有明确的功能定位，即能够满足大型模型推理需求或作为高并发推理应用、共享的推理资源池，服务于多个应用或用户，提高资源利用率和管理效率。在服务器芯片典型配置方面，集中节点采用高规格硬件，如配备多颗高性能GPU、张量处理单元（tensor processing unit，TPU）或专用AI加速器，用于处理复杂的、计算密集型的推理任务。关于部署位置，对于离线推理场景，可优先选择电力供应充足、成本较低的地区，如水电、风电等能源丰富的地区；对于在线推理场景，或共享的推理资源池、数据源集中在某个区域时，可考虑将集中节点部署在更靠近用户或数据源的位置，以降低网络时延、减少数据传输需求。在规模预估上，在低成本地区，集中节点可布局机房楼或数据中心园区，单栋机房楼具备支持万卡以上能力，园区支持向10万卡演进的机房设施；在业务热点和AI发展基础较好的城市，可布局单栋万卡级，园区支持万卡到数万卡的大型算力中心，从而满足行业大模型对海量数据和复杂计算的需求。

分布式节点的功能定位主要是满足大规模并行推理需求，它可以将推理任务分散到多个节点就近处理，提高业务感知能力。例如，部署在更接近数据源或用户的边缘设备上，满足实时性需求或安全需求。在服务器芯片典型配置方面，设备配置较为多样化，通常包括多颗高性能GPU、TPU或专用AI加速硬件、CPU等硬件，还可集成到服务器或边缘设备中的专门用于推理的FPGA或ASIC等。分布式节点的部署位置也具有多样化特点，如对于区域型的算力中心，可承载一定区域内的推理需求，如覆盖多个或一个地市范围；对于靠近用户现场型的算力中心，可满足现场实时业务、大数据量或数据不出场的需求。在规模预估方面，在AI发展较快、基础较好的重点城市，可能布局千卡到万卡的中大型算力中心，用于部署实时性、高价值业务或跨区域集中业务；同时，地市级可结合智慧城市、智慧工业等本地化多样化计算需求，部署百卡甚至千卡规模的算力中心、10～20架模块级算力资源，或几个机架的更小规模算力资源。

2.2.2 高密制冷和敏捷部署的需求

在算力基础设施建设中，高效的制冷系统至关重要。目前，推理服务器的功率存在差异，散热方案也呈现多样化。主流的推理服务器功率通常在3 kW以内，以高密风冷散热方式为主；训推一体型或高端推理服务器，功率显著提升，需要考虑风液融合的技术架构。推理服务器和机架

功率分析见表1。

同时，在实际场景中应用落地，推理任务的算力需求具有波动性，难以预测和管理，智算推理中心需要快速敏捷部署，并具备可扩展性。

面对智能推理资源部署需求，现有数据中心基础设施存在明显短板，难以适配算力扩张与能效提升需求。制冷方面，智算高功率服务器和高密度机柜散热需求激增，传统风冷效率不足且成本高；供电方面，算力密度攀升使电力机房空间需求剧增，需要系统统筹；运维层面，智能运维水平有限，AI能源管理等技术未规模化应用，管控精细化不足；建设周期方面，传统模式周期长、管控节点多，难以满足快速部署需求。

3　智算推理资源部署的数据中心基础设施解决方案

3.1　基础设施模型构建

数据中心基础设施技术架构，包括L0层（土建设施层）和L1层（机电设施层）。智能化系统部署了安防、电力监控、空调监控等系统或平台，覆盖L0层到L1层的运行监控。数据中心基础设施层的架构如图3所示。

为适配智算推理场景高效部署，数据中心应结合算力目标要素，构建空间布局（space）、供电（power）、制冷（cooling）、算力（floating-point operations per second，FLOPS）架构模型SPCF，实现基础设施层与算力设施层的联动，从而制订最优基础设施解决方案。

智算推理数据中心SPCF架构模型表达式如下。

$$S_{空间}=f(F_{算力},P_{功率},C_{功率}) \tag{1}$$

$$\mathrm{IT}_{功率}=F_{算力}/\mathrm{CE} \tag{2}$$

$$C_{功率}=\beta\times\frac{\mathrm{IT}_{功率}}{\mathrm{GCOP}} \tag{3}$$

$$P_{功率}=\alpha\times\frac{\mathrm{IT}_{功率}+C_{功率}+P_{建筑其他}}{\eta} \tag{4}$$

其中，$S_{空间}$表示数据中心建筑面积，单位为m^2；$F_{算力}$表示数据中心算力规模，单位为FLOPS；$\mathrm{IT}_{功率}$表示数据中心IT设备功率，单位为kW；

表1　推理服务器和机架功率分析

应用场景	单台服务器功率	单机柜功率/kW	制冷方式
模型参数量较小，实时性要求低	3 kW以内	10～20	普通风冷
大模型推理，训推一体	5 kW、10 kW等	20～40	高密风冷/液冷
大模型推理，训推一体	5 kW、10 kW、14 kW等	40～120	高密液冷

图3　数据中心基础设施层的架构

CE表示数据中心服务器的算效水平，单位为GFLOPS/W；$C_{功率}$表示数据中心制冷系统各设备的功率之和，单位为kW；$P_{功率}$表示数据中心电力系统各设备的功率之和，单位为kW；$P_{建筑其他}$表示数据中心其他电力负荷，单位为kW；β为制冷负荷冗余系数，在制冷系统中，为了确保系统在各种工况下都能稳定可靠运行，该系数是实际配置的制冷设备容量与计算得出的理论制冷负荷之间的比值；GCOP为数据中心空调系统额定能效系数；α为配电冗余系数，在配电系统中，为了保证供电的可靠性和稳定性，该系数是实际配置的配电容量与实际所需配电容量的比值；η为配电系统效率。

式（1）反映了数据中心建筑面积$S_{空间}$与算力规模$F_{算力}$、电力系统功率$P_{功率}$和制冷系统功率$C_{功率}$之间的内在联系。建筑面积与算力规模正相关，并与电力系统和制冷系统的功率水平紧密关联。在实际应用中，数据中心的建筑面积需求受到诸多因素的综合影响，包括数据中心算力规模目标、服务器设备的单位算力功耗水平、建筑平面布局、供电系统的冗余等级，以及冷却系统的能效比（coefficient of performance，COP）等。在有限的建筑空间内，电力供应、制冷能力和物理空间必须高度协同，才能确保高功率算力服务器的顺利部署。例如，目前一些数据中心由于制冷系统容量不足，即使拥有充足的物理空间和电力供应，仍然无法大规模部署高功率算力服务器，从而限制了数据中心算力的快速提升。

式（2）计算得出智算中心的IT设备功率；式（3）计算得出数据中心的电力容量；式（4）计算得出数据中心的制冷容量。

SPCF模型已在大量实践中得到应用。以搭建300 PFLOPS的推理数据中心为例，基于服务器算效600 GFLOPS/W得出$IT_{功率}$为500 kW。配电冗余系数α取1.85（供电系统按2N架构考虑），配电系统效率η取95%，制冷负荷冗余系数β取1.2，数据中心空调系统额定能效系数GCOP取3。根据式（3）得出$C_{功率}$约200 kW（采用高温水冷冻水空调系统）；根据式（4）得出$P_{功率}$约1 400 kW；根据IT设备选型及布放方案，以及制冷、供电等系统的设计要求和设备选型，确定IT设备、供电系统、制冷系统和辅助空间各自的建筑面积，计算得出，$S_{空间}$约500 m^2。

3.2 基础设施建设关键技术

3.2.1 高效制冷

根据服务器的功率密度和散热需求不同，冷却技术可分为风冷散热技术和液冷散热技术两大类[11]。

服务器风冷散热技术适用于低密度推理服务器或训推一体服务器，单机柜功率在30 kW以下。近年来，风冷散热技术逐渐向高制冷能力、高自然冷源利用率、高集成度及就近冷却方向发展[12]，以满足集中节点高密度、低能耗和快速部署的需求。当前的主流技术包括高温水冷冻水技术、风侧间接蒸发冷却技术及磁悬浮相变冷却技术等[13]，在全国范围部署后，其电能利用效率（power usage effectiveness，PUE）可降至1.25以下。分布式节点的部署场景较多，对于千卡级及以下规模的节点，容量规模较小，可采用氟泵变频空调技术，末端可采用列间空调、背板空调或机架内置空调等形式，机架功率可高低搭配，这种配置集成度高，能够实现快速部署，且维护便利，同时室外机采用静音型风机，可有效降低噪音影响。从实践经验看，在全国范围部署，其PUE可降至1.3及以下，几乎无耗水。

服务器液冷散热技术适用于高密度推理服务器或训推一体服务器，单机柜功率可达40～120 kW甚至更高。当前的主流液冷技术可分为冷板式与浸没式两种[14-15]。冷板式液冷技术是通过冷板中的冷却液与高密度芯片间接接触带走热量，同时服务器风冷和液冷并存，需要采用风液融合制冷

方案，PUE 可降至 1.15 以下。浸没式液冷技术是通过冷却液与服务器直接接触带走热量，该技术成熟度较低，成本较高，应用较少，需要做好服务器定制。该技术的 PUE 可降至 1.1 及以下。

高效制冷技术特点及应用场景建议见表 2。

3.2.2　高效供配电

供配电系统应以智算需求为导向，围绕“可靠性、高效性、集约性”三大核心维度规划系统配置，实现能源的安全供给与设备的高效运行。

可靠性：合理选用供配电系统架构，配置关键不间断电源及备用电源设备，保障能源的安全可靠供给，避免因供电中断、波动或人为因素影响算力设施的稳定运行。基础设施资源池化[16]处理，不间断电源系统采用并机池化供给方案，提升系统整体的抗冲击能力，应对算力业务的突发电力需求与单台设备故障。

高效性：保障变配电设备在全生命周期高效运行，重点关注设备选型和运行管理两个方面，优先选用高效节能型变配电设备及高效运行模式。例如，采用 1 级能效的变压器或采用高转换效率的不间断电源等；在市电质量稳定的前提下，直接利用市电为智算设施供电，降低供电系统损耗；启用模块休眠功能，维持供配电系统在最优效率区间运行。

集约性：整合变配电设备，推动变配电系统集约化，提升电力设备空间利用率。随着算力密度的不断提升，电力机房空间需求激增。供配电系统应兼顾空间的高效利用，通过供电架构的极简化和设备的高度集成化实现集约高效。

3.2.3　智能运维

数据中心的智能化运维需要构建“全域感知–智能决策–动态调优”的一体化架构，通过部署多维度传感器网络实时采集设备与环境数据，实现基础设施运行状态的全局监控与健康度评估；基于知识图谱的故障预测模型，分析设备退化曲线与历史告警数据，预判潜在风险并自动生成预防性维护工单，大幅降低人工干预频率[17]；构建多物理场耦合动态调优引擎，实时解析 IT 负载热力图与制冷系统状态，通过混合整数规划框架对冷冻水温度、冷却塔风机转速等多参数全局寻优；依托数字孪生底座构建跨域优化知识库，持续学习历史调优数据与异常事件，通过强化学习算法迭代控制策略，形成“策略生成–效果验证–模型进化”的自适应闭环。对于分布式节点的小规模数据中心，可通过轻量化技术栈实现运维与调优能力的深度耦合，实时整合设备状态、环境参数与业务负载数据，构建动态运行画像，完成风险预警和能效调优。

表 2　高效制冷技术特点及应用场景建议

技术	高温水冷冻水技术	风侧间接蒸发冷却技术	磁悬浮相变冷却技术	氟泵变频空调	冷板式液冷技术	浸没式液冷技术
节能性	较高	较高	高	中	高	高
节水性	耗水量大	耗水量较少	耗水量较大	几乎无耗水	耗水量大（冷却塔冷源）	耗水量大（冷却塔冷源）
装机率	较高	多层建筑：低 单层或双层建筑：中	多层建筑：中 单层或双层建筑：较高	中	高	高
成熟度	高	较高	中	高	中	低
应用场景建议	建议在水源充足的万卡级及以上规模的智能推理中心应用	建议在土地资源充足、水资源有限且需要快速部署的万卡级及以上规模的智能推理中心应用	建议在千卡、万卡级规模的智能推理中心应用	建议在千卡及以下分布式节点部署	建议在高密度推理和训推一体数据中心应用，优先推荐采用冷板式液冷	
适配单机柜功率	30 kW 以内	12 kW 以内	30 kW 以内	30 kW 以内	40～120 kW 甚至更高	

注：“装机率”主要指单位建筑面积机架功率产出率，单位为 kW/m^2。装机率越高代表建筑空间利用率越高。

3.2.4 快速建设模式

随着智算部署对敏捷交付的高要求，“产品工程化”模式和“工程产品化”模式应运而生。数据中心的建设模式，正逐步从传统建设模式向“产品工程化”模式、“工程产品化”模式演进。

传统建设模式（通过工程建设、设备采购安装等分阶段完成的建设模式）通常建设周期长，质量较难控制。“产品工程化”模式，是将智算中心的核心功能（如供配电、制冷、机柜、监控等子系统）设计并制造成标准化、模块化的集成产品，再通过“搭积木”的方式完成整体工程建设，将原本分散采购、现场施工的系统，转化为预设计、预制造、预测试的“产品”进行安装或组装，缩短建设周期，提高工程质量。“工程产品化”模式是指将整个智算中心工程作为一个完整的产品，以交付为中心，用户不需要参与设计、施工、调试等环节，最终获得一个可直接使用的“算力产品”。

“产品工程化”模式更适用于集中节点和分布式节点中千卡、万卡及以上规模的智算推理基础设施部署。对于分布式节点中百卡及以下、规模较小的智算资源部署，则可采用“工程产品化”模式，将整个工程以产品的形式交付，提高标准化程度。快速建设模式应用场景示意图如图4所示。

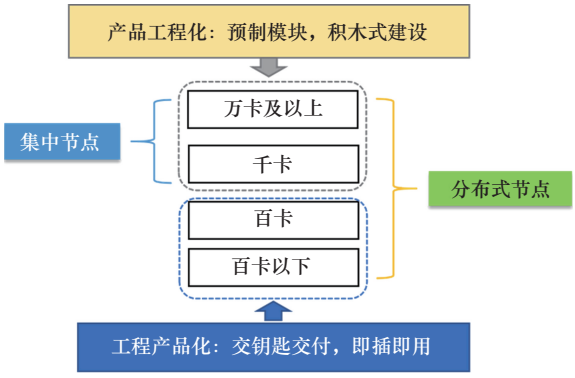

图4 快速建设模式应用场景示意图

3.3 解决方案及典型案例

数据中心基础设施解决方案以“智能敏捷、绿色低碳、安全可控”为原则，通过高效制冷、供电、智能运维及快速交付体系，构建弹性扩展、自主进化的绿色数字信息基础设施，满足智能推理资源即插即用、按需扩容与可持续运营的需求。

3.3.1 集中节点

集中节点数据中心基础设施架构如图5所示。集中节点的智算推理基础设施主要通过“产品工程化”的模式实现部署，分为改造和新建两种场景。改造场景应基于“最小化改造干预、最大化能效提升”的原则，结合智算需求，优化空间布局和提升供电、制冷能力。新建场景应遵循“前瞻性规划、全栈化协同”的核心理念，支撑未来

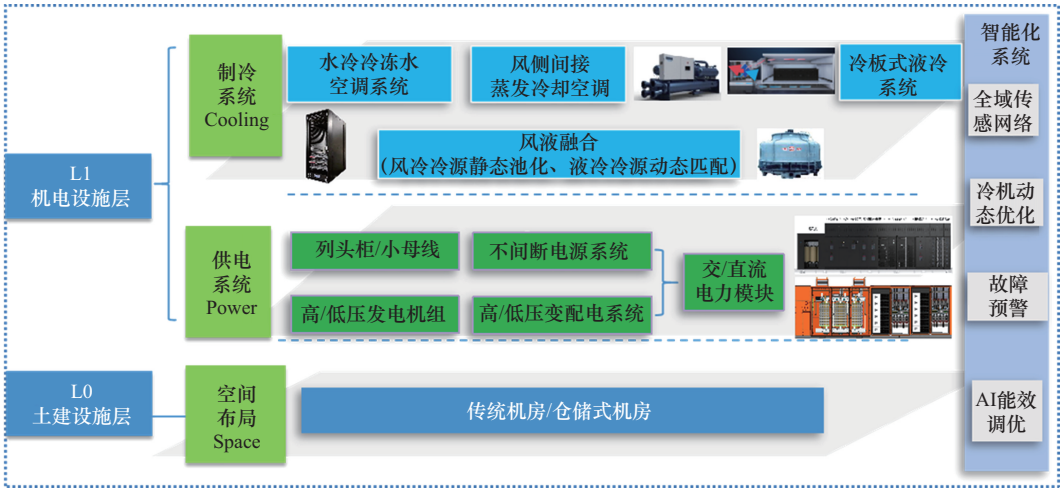

图5 集中节点数据中心基础设施架构

设备演进和新技术的应用，构建面向通用人工智能（artificial general intelligence，AGI）时代的智算基础设施。

（1）存量数据中心改造场景建设方案

存量数据中心改造应统筹考虑功能定位和装机需求，结合原机房规划制订具体改造方案，优先对现有资源进行容量提升，挖掘既有空间、供电和制冷能力。

空间布局重构：考虑高密机柜的散热要求，做好冷热通道规划，在保障现网业务正常运行的同时，减少改造量。

供电系统重构：采用模块化电力设备、母线和智能布线系统，支撑运行功率动态调节；结合原系统配置特点，优选市电+不间断电源的架构，采用并机运行方案实现不间断电源容量资源池化，提升设备的抗冲击能力。对于空间不足的场景，可选用一体化电力设备来满足智算设备的高密部署需求。

制冷系统重构：对于风冷推理服务器，由于功率密度增加，原有的风冷空调末端及冷源需要扩容，水管管径若不满足供冷需求时，也需要改造扩大管径，原房间级空调末端需要改造为列间级或机柜级就近制冷末端，同时封闭热通道；对于液冷服务器，需要在屋面或室外地面增加液冷泵站和冷塔设备，增设管路，带走液冷负荷。

智能化重构：通过部署轻量化感知网络与智能分析平台，实现全域设备状态监控与异常预警。基于历史数据与运行特征构建动态优化模型，对制冷系统参数（如冷冻水温度、风机转速）及供电策略进行自适应调整，以匹配实时负载波动。

（2）新建数据中心场景建设方案

新建数据中心建设应结合推理业务需求，在建筑结构、动力配套、制冷散热、智能化系统等方面协同化布局、产品化部署，并预留高密机柜和普通机架互弹条件，实现灵活匹配。

空间布局规划：充分考虑空间弹性转换，在荷载、空间高度方面做好预留，满足近远期推理业务需求和设备演进需求，同时满足设备安装和运维需求。

供电系统构建：根据智算资源的配电需求，采用不间断电源2N或市电+不间断电源的配电架构，采用并机运行方案提升不间断电源系统资源池容量，保障设备稳定运行。对于需要连续运行保障的液冷设施，采用市电+不间断电源的架构配置供配电系统。

制冷系统构建：采用“风冷+液冷”融合模式，风冷散热可以采用水冷冷冻水空调系统或风侧间接蒸发冷却空调，液冷散热采用冷板式液冷系统，通过风冷冷源静态池化、液冷冷源动态匹配，实现风液融合，适配不同风液服务器需求。

智能化系统构建：通过全域传感网络与数字孪生底座实时映射物理设施状态，结合冷机动态优化与供配电冗余可视化与预测性告警机制，实现制冷系统参数的自适应调优并保障供电可靠性。集成AI能效平台，基于IT负载预测与实时电价信号动态调整运行模式，自动平衡能效与经济性目标，同时通过故障预测–隔离机制保障系统稳定性。

（3）典型案例

某智算中心项目采用国产化服务器超万卡，算力规模$F_{算力}$约1 EFLOPS，以满足大模型训练和集中推理算力需求。该项目采用SPCF模型估算，结合算效水平和各系统配置架构和技术方案得出IT$_{功率}$约4 700 kW，制冷系统负荷冗余系数β取1.2，GCOP取3，可得出制冷各设备功率之和$C_{功率}$约1 880 kW，配电冗余系数α取1.85（供电系统为2N架构），配电系统效率η取95%，电力系统各设备的功率之和$P_{功率}$约13 200 kW。该智算中心由现有数据中心改造实现，改造面积$S_{空间}$约1 700 m^2。

技术方案方面，该项目的供电系统为新增

集装箱式通信用10 kV输入的交流不间断电源系统及智能配电母线；机柜风冷散热系统采用高温水冷冻水系统，冷源配置为冷机+板换+冷塔形式，空调末端采用列间空调；机柜液冷散热系统采用解耦型冷板液冷技术，液冷冷源采用预制集成泵站模式，部署在原有制冷站的屋面。

建设模式方面，该项目采用设计–采购–施工（engineering-procurement-construction，EPC）总承包模式，建设周期共计5个月，具备全部上线条件，实现了质量可控、成本可控和建设周期可控的目标。

节能措施方面，该项目利用液冷技术，全年实现自然冷却，不使用无压缩机机械制冷；风冷散热高温水冷冻水系统，可大幅延长自然冷却时间，降低冷却能耗。此外，项目还采用了AI能效调优、余热回收、自建分布式光伏系统等技术。项目现已投产运行，年均PUE为1.17，达到了较好的能效水平。

3.3.2 分布式节点

分布式节点的智算推理资源需求多样，规模差异较大，基础设施的建设可通过“产品工程化”与“工程产品化”等多种方式实现。分布式节点数据中心基础设施架构如图6所示。

（1）产品工程化解决方案

对于千卡、万卡级规模的分布式节点机房建设场景，可参考集中节点解决方案；对于百卡级及以下规模的分布式节点，可直接利用或改造现有机房资源，如采用“机架级、机列级微数据中心”模块化改造方案。制冷方案为将房间级空调末端改造为风冷氟泵列间空调或利用现有风冷散热系统列间空调，充分利用自然冷源，降低制冷能耗，同时封闭热通道，提高制冷效率。供配电系统宜以每列机柜功率为基准，选用容量匹配的高效电源设备，按市电+不间断电源架构配置供配电系统。

（2）工程产品化解决方案

面向自动驾驶、工业机器人等典型推理场景的中小型智算中心，可结合客户需求，提供多维度的集成产品服务，覆盖基础设施、算力设施、开发平台、客户特殊需求应用等全栈能力。例如，某企业的一体化智算中心产品“MINI智算中心”是业界首款一体化集成基础设施、智算资源、智算开发平台、安全解决方案的智算一体机产品，单柜功率10 kW+，可实现“上电上网即可用”，方便教育、医疗、科研、政府等行业政企客户快速构建企业级独享智算能力。

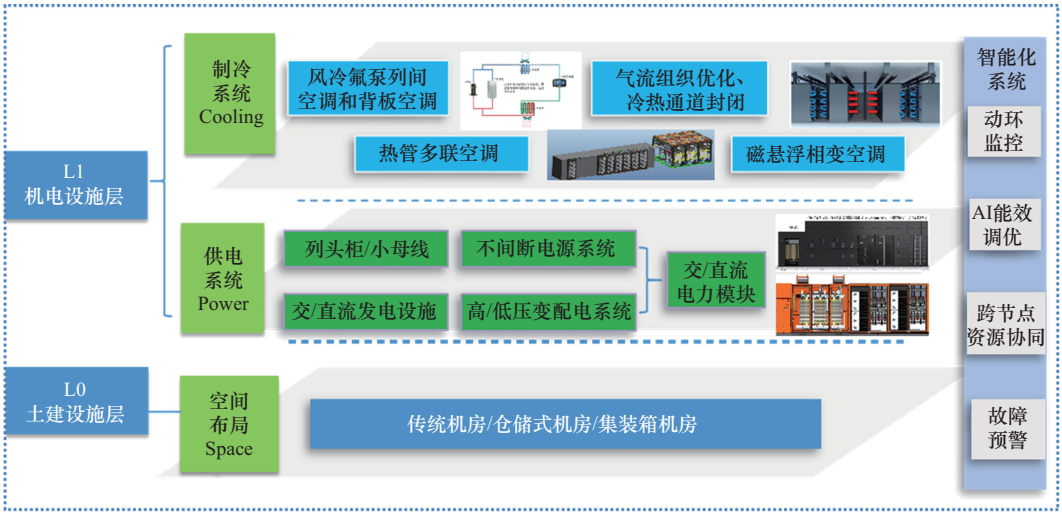

图6 分布式节点数据中心基础设施架构

（3）典型案例

某分布式节点项目通过对现有机房改造，使风冷智算机房实现千卡集群能力。算力规模目标约600 PFLOPS。结合算效水平和各系统配置架构和技术方案，得出$IT_{功率}$约2 800 kW，制冷系统负荷冗余系数β取1.2，GCOP取2.5，可得出制冷各设备功率之和$C_{功率}$约1 344 kW，配电冗余系数α取1.85（供电系统为2N架构），配电系统效率η取95%，电力系统各设备的功率之和$P_{功率}$约8 000 kW，改造面积$S_{空间}$约1 000 m^2。

该项目供电方案采用240 V直流电源系统，在电力室建设总容量2 000 kVA、输入10 kV的高效直流电源系统；制冷系统通过利旧现有风冷系统冷源，扩容改造列间空调末端容量及管径，以适配负荷增加后的制冷需求，改造后实现约230个5.6～25 kW风冷机架能力。

针对分布式节点弹性装机的需求，该项目进行了多种颗粒度的拓展预留。项目针对空调末端、配电列头柜等设备，按每列的总容量进行弹性装机配置，确保满足弹性装机需求，对机柜电源分配单元（power distribution unit，PDU）等进行最高功率配置，以降低后续扩容的布线难度并减少对在线设备运行的干扰。此外，项目还规划预留了液冷机房，可满足40 kW冷板式液冷高密机柜的部署需求，在空间、管路等方面做好预留。

4　未来发展建议

面对以DeepSeek为代表的AI技术加速涌现和算力需求结构深刻变革，数据中心正迈入能力跃升的关键阶段。未来数据中心基础设施创新需要坚持“资源协同化、弹性高效化、能效绿色化、运维智能化”的发展方向，以支撑AI技术的持续演进和应用拓展。

4.1　强化基础设施全栈协同水平

在AI应用深化的牵引下，IDC商业模式正在重构，从传统的资源式服务向模型即服务（model as a service，MaaS）、AI软件即服务（software as a service，SaaS）拓展升级，对数据中心基础设施的协同性提出了更高要求，不仅要深化物理基础设施层、算力设施层、平台层以及应用层的纵向协同，更要推动跨层次、跨模块的横向协同。加强基础设施与算力架构的深度耦合，要从服务器与液冷系统的接口兼容性、冷却液材料匹配性等问题入手，实现基础设施与算力设施的无缝衔接，确保高密度算力下的稳定可靠运行；提升面向MaaS、AI SaaS的基础设施预置与优化水平，对于MaaS、AI SaaS场景，基础设施将需要与上层应用进行更深度的融合，如针对特定行业和应用场景，预置匹配的算力资源，并集成优化的预训练模型和行业应用，打造“开箱即用”的整体解决方案，提升服务的快速交付能力和用户体验。

4.2　提升基础设施高效弹性能力

数据中心高效化是可持续发展的基石。在算力需求持续增长和高密度部署的驱动下，数据中心必须持续深化高效化技术的创新与应用，以提升资源和能源利用效率，并应对DeepSeek技术发展带来的算力需求波动。例如，通过软硬件协同优化供配电系统，提升其能效、功率密度和运行可靠性；采用冷、电盲插式的解耦液冷机柜，可实现液冷服务器的快速部署和极简维护。此外，模块化、预制化和标准化的弹性基础设施模块，是应对多代算力演进和多元化算力混合部署的理想选择，能够显著提升基础设施的快速部署能力。

4.3　推进基础设施极致能效绿色化

面对日益严峻的能源挑战和环境压力，引入绿色能源、构建绿色可持续的推理资源是实现AI长期健康发展的关键保障。大力提升可再生能源（如太阳能、风能、氢能等）的使用比例，探索市场化绿色电力采购、自建分布式可再生能源发电设施（如光伏发电、风力发电），以及储能技术的应用，构建多元化绿色能源供应体系。对于

大型集中式推理中心，可协同布局在大型可再生能源发电厂区域附近，配置储能设施，保障可再生能源的完全消纳。对数据中心全生命周期的碳足迹进行管理，从建设阶段的绿色建材选用到运行阶段的能源优化，再到报废阶段的设备回收和处理，都纳入绿色化考量，实现数据中心全生命周期的绿色可持续发展。

4.4 推动基础设施运维自主智能化

构建AI驱动的智能运维平台是应对数据中心复杂性和提升运维效率的关键所在。数据中心基础设施的智能化应以“自动感知–闭环控制–智能调优–预测维护”为基础，逐步推进。具体而言，实现全状态感知与实时监测，完善基础设施关键链路传感器与联动组件部署，确保故障可即时发现、告警可及时下达、应急可自动切换；集成平台汇聚多维运行数据，进行数据治理、模型融合和可视化分析；构建面向设施全生命周期的数字映射模型，对负荷、温控、气流等状态进行在线演算与仿真，为优化方案提供“镜像现场”验证环境；引入机器学习算法，建立能耗、容量、故障趋势的自适应策略库，实现PUE持续下降、容量弹性提升和故障趋势前移[18]。最终形成从检测到处置的全流程自动闭环，显著释放运维人力、降低风险。为确保落地效果，应同步建立跨部门数据标准与持续验证机制，使流程、组织与技术同频迭代，不断巩固安全、效率与绿色三重价值。

5 结束语

综上所述，本文系统提出了数据中心基础设施“资源协同化、弹性高效化、能效绿色化、运维智能化”的技术发展方向与整体解决方案，相关技术已在工程实践中得到初步验证，展现出良好的应用前景。未来，将进一步深化这些技术在大规模工程实践中的应用与推广，特别是在智能驾驶、智慧城市、智慧医疗等重点行业开展最佳实践，形成可复制、可推广的行业解决方案，助力各领域实现智能化升级。

展望未来，随着AI技术的不断演进和多元算力需求的持续增长，数据中心基础设施的创新之路将更加广阔而深远。因此，需要持续探索新技术、新能源、新配置、新模式，不断优化数据中心基础设施的资源配置，以更好地满足多代算力的混合部署需求，为智能时代的数字经济蓬勃发展奠定坚实基础。

参考文献：

[1] 魏钰明, 贾开, 曾润喜, 等. DeepSeek突破效应下的人工智能创新发展与治理变革[J]. 电子政务, 2025(3): 2-39.
Wei Y M, Jia K, Zeng R X, et al. Innovative development and governance transformation of artificial intelligence under the breakthrough effect of DeepSeek[J]. E-Government, 2025(3): 2-39.

[2] Jevons W S, Flux A W. The coal question: an inquiry concerning the progress of the nation, and the probable exhaustion of our coal-mines[M]. 3rd ed. New York: Augustus M. Kelley, 1965.

[3] 陈永伟. 从规模定律到规模经济: DeepSeek的创新、机遇与挑战[J]. 山东大学学报(哲学社会科学版), 2025(5): 140-151.
Chen Y W. From law of scale to scale economy: innovations, opportunities, and challenges of DeepSeek[J]. Journal of Shandong University (Philosophy and Social Sciences), 2025(5): 140-151.

[4] 黄晓野, 代栓平, 李克. 技术革命周期与我国算力竞争战略选择: 基于DeepSeek复杂经济系统的思考[J]. 工业技术经济, 2025, 44(4): 25-31.
Huang X Y, Dai S P, Li K. The technological revolution cycle and China’s strategic choices for computing power competition: thoughts based on DeepSeek complex economic system[J]. Journal of Industrial Technology and Economy, 2025, 44(4): 25-31.

[5] 国际数据公司(IDC), 浪潮信息. 2025年中国人工智能计算力发展评估报告[R]. 2025.
International Data Corporation, Inspur Information. Evaluation report on the development of artificial intelligence computing power in China in 2025[R]. 2025.

[6] Wei J, Wang X Z, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM Press, 2022: 24824-24837.

[7] 梁秉豪, 张传刚. 训推一体平台架构设计与关键技术研究[J]. 计算机科学与应用, 2023,13(9): 1748-1755.
Liang B H, Zhang C G. Architecture design and key technology research of training and push integrated platform[J]. Computer Science and Applications, 2023,13(9): 1748-1755.

[8] Gou J P, Yu B S, Maybank S J, et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[9] 王彦伟, 李仁刚, 徐冉, 等. 基于可重构架构的数据中心异构加速软硬件系统级平台[J]. 计算机研究与发展, 2025, 62(4): 963-977.
Wang Y W, Li R G, Xu R, et al. Data center heterogeneous acceleration software-hardware system-level platform based on reconfigurable architecture[J]. Journal of Computer Research and Development, 2025, 62(4): 963-977.

[10] 陈星延, 张雪松, 谢志龙, 等. 面向“云-边-端”算力系统的计算和传输联合优化方法[J]. 计算机研究与发展, 2023, 60(4): 719-734.
Chen X Y, Zhang X S, Xie Z L, et al. A computing and transmission integrated optimization method for cloud-edge-end computing first system[J]. Journal of Computer Research and Development, 2023, 60(4): 719-734.

[11] 陈心拓, 周黎旸, 张程宾, 等. 绿色高能效数据中心散热冷却技术研究现状及发展趋势[J]. 中国工程科学, 2022, 24(4): 94-104.
Chen X T, Zhou L Y, Zhang C B, et al. Research status and future development of cooling technologies for green and energy-efficient data centers[J]. Strategic Study of CAE, 2022, 24(4): 94-104.

[12] Gao P, Liu H, Luo H L, et al. Discussion on the technical path of data center information and communication thermal management[J]. Energy Reports, 2024, 11: 2704-2714.

[13] 姜宇光. 数据中心冷水空调系统与间接蒸发冷却空调系统建模对比分析[J]. 暖通空调, 2021, 51(1): 76-83.
Jiang Y G. Modeling and comparative analysis of chilled water air conditioning systems and indirect evaporative cooling air conditioning systems in data centers[J]. Heating Ventilating & Air Conditioning, 2021, 51(1): 76-83.

[14] 朱宸, 魏东黎, 余海生, 等. 冷板式液冷技术应用于智算中心高密机房的方案分析[J]. 电信工程技术与标准化, 2024, 37(增1): 174-179.
Zhu C, Wei D L, Yu H S, et al. Analysis of the application scheme of cold plate liquid cooling technology in high-density computer rooms of intelligent computing centers[J]. Telecom Engineering Technics and Standardization, 2024, 37(S1): 174-179.

[15] 谢文韬, 余承学, 谢昕言, 等. 数据中心冷却节能技术及余热回收技术研究进展[J]. 暖通空调, 2025, 55(2): 1-9, 25.
Xie W T, Yu C X, Xie X Y, et al. Research progress of cooling energy-saving technology and waste heat recovery technology of data centers[J]. Heating Ventilating & Air Conditioning, 2025, 55(2): 1-9, 25.

[16] 郭亮. 数据中心热点技术剖析[M]. 北京: 人民邮电出版社, 2019.
Guo L. Analysis of hot technologies in data centers[M]. Beijing: Posts & Telecom Press, 2019.

[17] 中国信息通信研究院, 开放数据中心委员会. 数据中心智能化运维发展研究报告(2023 年)[R]. 2023.
China Academy of Information and Communications Technology, Open Data Center Committee. Research report on the development of intelligent operation and maintenance of data centers (2023)[R]. 2023.

[18] 胡燕妮, 丁赟, 蔡幸波. “双碳”目标下, 智慧金融数据中心的能效优化与智能运维策略[J]. 智能建筑电气技术, 2025, 19(1): 5-9.
Hu Y N, Ding Y, Cai X B. Energy efficiency optimization and intelligent operation and maintenance strategies for intelligent financial data centers under the “dual carbon” goal[J]. Electrical Technology of Intelligent Buildings, 2025, 19(1): 5-9.

[作者简介]

刘洪（1973– ），男，中国移动通信集团设计院有限公司算力设施工程研究中心主任、正高级工程师、国务院特殊津贴专家，主要研究方向为数据中心基础设施新型技术。

孙丽玫（1971– ），女，中国移动通信集团设计院有限公司正高级工程师，主要研究方向为数据中心通信工艺、数据中心绿色低碳节能技术等。

朱丽（1980– ），女，中国移动通信集团设计院有限公司正高级工程师，主要研究方向为数据中心咨询、设计。

姜宇光（1988– ），男，中国移动通信集团设计院有限公司高级工程师，主要研究方向为算力中心制冷空调、绿色低碳。

赵金铭（1986– ），男，中国移动通信集团设计院有限公司高级工程师，主要研究方向为智算中心组网架构、机房工艺设计。

孙立峰（1980– ），男，中国移动通信集团设计院有限公司高级工程师，主要研究方向为算力中心供配电系统、智能化系统。

李雅婷（1996– ），女，中国移动通信集团设计院有限公司工程师，主要研究方向为算力产业、算力经济。