CAICT 中国信通院

集智 专题报告

DPU 发展分析报告

(2026 年)

中国信息通信研究院云计算与数字化研究所

2026年5月

版权声明

前　言

当前，全球数字经济加速发展，以大模型、智能体为代表的新一代人工智能技术加速落地，智算需求爆发式增长。作为智算设施的重要组成，数据处理器（Data Processing Unit，DPU）凭借软硬协同的专用加速能力，实现网络、存储、安全等任务的高效卸载，破解传统计算架构的性能瓶颈，助力算力资源的高效调度，降低系统总拥有成本，为模型训练、应用推理、边缘计算等场景提供支撑，被业界认为继 CPU、GPU 之后“第三颗主力芯片”。

DPU 产业正由高速发展向高质量发展全面演进，从政策方面看，全球主要经济体积极推动产业发展，美国、欧洲等国家在“创世纪计划”、《芯片法案 2.0》等中强调要推动涵盖 DPU 在内的半导体发展，我国通过《算力基础设施高质量发展行动计划》、算力强基揭榜行动等政策牵引 DPU 研发应用。从技术方面看，DPU 架构不断演进，通过硬件级卸载释放 CPU、GPU 潜能，依托高速接口技术实现数据高效传输，解决 AI 训练和推理中的网络瓶颈问题。从产业方面看，通用算力中心、智能算力集群、5G 边缘计算等多样化场景逐步应用 DPU，为产业发展提供有效的需求牵引。未来，DPU 架构将不断迭代创新，加速场景渗透，释放应用价值。

本报告立足产业发展，系统梳理 DPU 发展背景、行业动态与市场环境，研判未来趋势，并提出针对性建议，旨在全面呈现全球及我国 DPU 产业发展全貌，为政策制定、技术研发与生态构建提供参考，助力我国 DPU 产业发展，筑牢数字经济高质量发展的算力之基。

时间仓促，报告仍有诸多不足，恳请各界批评指正。后续我们将不断更新完善，如有意见建议请联系中国信通院研究团队：dceco@caict.ac.cn。

目 录

图 目 录

一、DPU 发展概述

当前，随着数字经济的持续发展，尤其是大语言模型、自动驾驶、具身智能等应用加速落地，智算需求爆发式增长。作为算力中心的“第三颗主力芯片”，DPU 重要性不断凸显，从早期面向网络、存储、安全等卸载功能，逐步升级为支撑算力基础设施建设的重要硬件底座。

（一）全球强化政策引领，夯实发展根基

美国、欧盟等相继出台产业战略，围绕半导体加大算力研发布局，抢占技术主导权。2025 年 11 月，**美国**联邦政府正式启动“创世纪计划”（Genesis Mission），12 月美国能源部联合 26 家科技企业和机构签署合作备忘录，计划将“半导体与微电子”列为优先突破领域之一，为包括 DPU 在内的相关芯片技术的研发、制造和应用提供顶层的战略支持和潜在的资源倾斜。同年，**欧盟**启动《芯片法案 2.0》修订进程，明确加大 AI 芯片、先进互连、光封装、算力卸载芯片资金倾斜，优化审批与人才体系，进一步强化数据中心专用微电子技术主权布局。2025 年 12 月，韩国公布 700 万亿韩元长期投资规划，目标建成全球最大芯片产业集群，抢占 AI 芯片竞争优势

我国高度重视 DPU 产业发展，已构建起国家与地方协同推进的政策体系，有力推动产业规模扩张与技术创新突破。国家层面，2023 年 10 月，工业和信息化部等六部门印发《算力基础设施高质量发展行动计划》提出开展 DPU、无损网络等技术升级与试点应用，实现算力中心网络高性能传输。2025 年 2 月，工业和信息化部启动算力强基揭榜行动，开展基于芯粒（Chiplet）和 RISC-V（Reduced Instruction

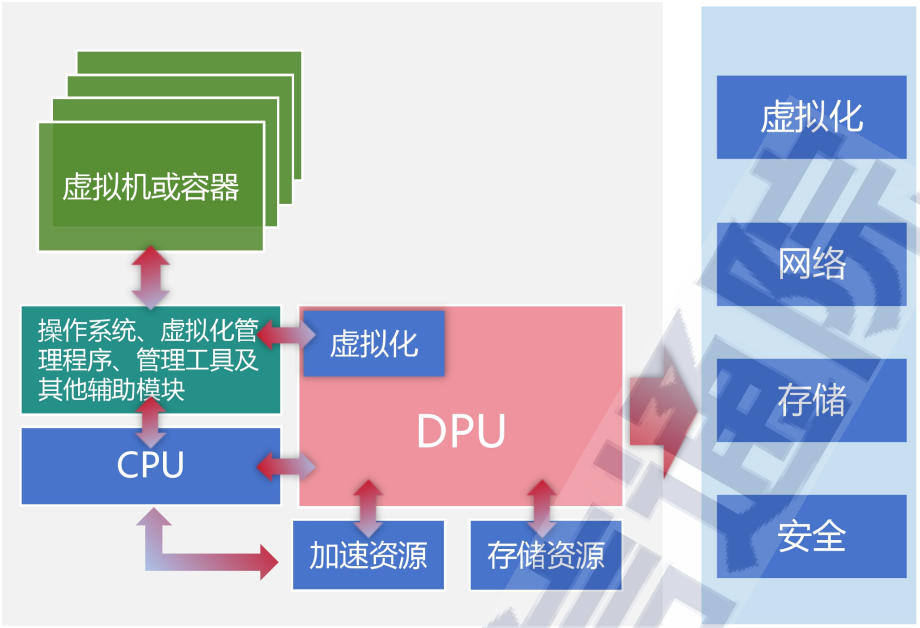
Set Computing-V，第五代精简指令集计算）技术的软硬件一体 DPU 芯片技术研究。2025 年 8 月，国务院发布《关于深入实施“人工智能+”行动的意见》，其中提到支持人工智能芯片攻坚创新与使能软件生态培育，加快超大规模智算集群技术突破和工程落地。2025 年 9 月，工业和信息化部征集“重点产品、工艺‘一条龙’应用计划”的推进机构，明确将 DPU 列为关键产品，提出要“围绕 DPU 应用需求，开展基于国产工艺的高性能可编程数据处理器的研制、操作系统（OS）子系统升级与硬件扩展”，并推动其在云计算、高性能计算、AI 等领域的应用。**地方层面**，上海、深圳等地纷纷出台行动计划。2025 年 10 月，上海市发布《上海市智能终端产业高质量发展行动方案（2026—2027 年）》，旨在加强端侧人工智能芯片的布局，推动产业规模突破。2026 年，深圳市印发《深圳市“人工智能+”先进制造业行动计划（2026—2027 年）》，构建人工智能与半导体产业“双向赋能”的生态，通过 AI 技术提升芯片设计、制造全流程效率，以破解高端自主不足的瓶颈。

（二）技术驱动架构升级，成为主力芯片

随着算力需求的多元化与高性能计算的快速迭代，传统 CPU＋GPU 的双核架构已难以适配高带宽、低时延、强安全的需求，推动计算架构向“通用+专用”协同转型。DPU 通过在硬件层面实现网络协议栈、低时延 RDMA 传输、存储 I/O、加密解密等功能的专用加速，并在 AI 集群中承担构建推理上下文内存存储平台、卸载非计算任务释放 GPU 算力、承担机架内数据神经中枢等功能，提供可编程

的算力中心基础设施软件栈，实现对算力资源的细粒度调度与动态优化。近年来，NVIDIA 多次在 GTC 等技术大会上强调 DPU 的重要性，并已在 DGX AI 超级计算机和 Spectrum-X 网络平台中规模化部署，DPU 在算力基础设施中的战略性地位持续提升。

从技术层面看，DPU 通过硬件级卸载网络、存储及安全任务，有效释放 CPU 与加速器的算力。DPU 可采用芯粒（Chiplet）设计，集成专用网络处理单元、可编程计算核心与高速 I/O 接口，通过片上互连网络实现内部功能模块的高效协同。在互联架构层面，DPU 重点支持单节点内 Scale-Up 与跨节点集群 Scale-Out 互联，依托远程直接内存访问（Remote Direct Memory Access，RDMA）等先进协议，实现 GPU 间的低时延、高吞吐数据交换，缓解万卡级 AI 集群的通信瓶颈。此外，DPU 依托软件定义技术，在云计算领域优化虚拟化效率。同时，通过集成网络处理、计算加速、本地存储及安全防控等功能，满足 5G、智算中心等高带宽、低时延、高吞吐应用场景需求，推动计算与网络深度融合。



来源：中国信息通信研究院

图 1 DPU 功能示意图

从应用价值看，DPU 已超越基础的“降本增效”。在 AI 推理与超大模型训练场景中，其核心作用演变为重构数据路径与内存层级，以系统级优化直接释放算力红利。DPU 通过构建池化的“推理上下文内存”，将海量 KV Cache 从 GPU-HBM 卸载至由其统一管理的大容量 NAND 资源池，解决了上下文长度快速扩张带来的 KV Cache 容量爆炸、访存压力剧增与算力效率急剧下降的问题，降低单次推理的数据搬运与预处理功耗，缩短训练周期并降低单词元（Token）推理成本。同时，DPU 作为智能的数据调度中枢，实现对存储 I/O、安全协议及虚拟化功能的彻底卸载与硬件加速，使 CPU 与 GPU 专注于核心计算本身，在提升算力中心整体性能与利用效率的同时，有效降低硬件冗余与总体能耗。

（三）核心场景率先突破，行业渗透深化

DPU 作为算力基础设施的核心组件，在互联网、金融、能源等关键行业实现规模化落地，已从单一功能加速设备，发展成为覆盖多场景、全链路的算力核心处理器之一。**在算力建设方面**，DPU 聚焦云计算、大数据和人工智能等场景，利用硬件卸载网络转发、存储虚拟化及安全隔离等基础设施负载，有效解决性能与安全难以兼顾的难题。目前，全球头部云厂商已完成规模化部署，国内阿里云、腾讯云、华为云等企业在新一代服务器中实现 DPU 标准化落地，推动 DPU 从可选加速部件向算力中心的基础标配器件转变。

在互联网领域，DPU 赋能高并发业务，提升服务体系效能。DPU 已广泛应用于搜索引擎、电商交易、在线视频、社交媒体等各类高并发在线服务体系，成为头部平台降本增效、优化用户体验的核心抓手。在搜索引擎场景中，通过硬件加速用户请求的转发、处理与响应，提升检索效率与结果反馈速度，同时降低后端服务器的算力开销，支撑海量用户的实时交互通信、内容分发与数据共享，优化终端用户体验。

在金融领域，DPU 筑牢安全防线，保障核心交易高效运行。针对高频交易、实时风控及跨境支付等核心场景，依托微秒级低时延、硬件级安全隔离及全链路可审计能力，精准匹配金融行业对效率与合规的要求。在证券高频交易中，DPU 将端到端时延压缩至亚微秒级，提升交易执行效率。目前，DPU 已广泛应用于银行、证券、保险机构的私有云及核心系统建设，成为金融关键信息基础设施的重要支撑。

在能源领域，DPU 助力数字化转型，强化电网智能调控能力。随着新型电力系统建设、新能源场站规模化发展与能源行业数字化转型的持续深化，DPU 在智能电网、新能源场站管控、油气勘探等场景的应用需求快速攀升。在智能电网场景中，通过部署于变电站边缘节点，实现通信协议转换、数据实时采集预处理及故障快速识别，将故障响应时延从秒级压缩至毫秒级，提升电网运行稳定性与应急处置能力。

此外，DPU 在教育、工业、医疗等行业同样具备广阔应用前景，为行业数字化转型提供关键支撑。

（四）多重瓶颈叠加制约，发展仍待突破

当前，我国 DPU 正处于从技术突破向规模化商用跨越的关键窗口期。在算力网络升级与人工智能发展的双重机遇下，仍面临技术、市场与供应链等关键问题，产业发展仍存在诸多瓶颈与挑战。

先进制程获取受限、软件生态建设仍需完善。技术层面，先进制程获取受限成为制约 DPU 产业高质量发展的首要障碍。**软件生态层面，**国内厂商多聚焦硬件芯片设计，在数据面开发套件、虚拟化层、行业应用接口等上层软件能力建设方面处于培育期，仍有提升空间，整体呈现出“硬件先行、软件跟进”的阶段性特征。

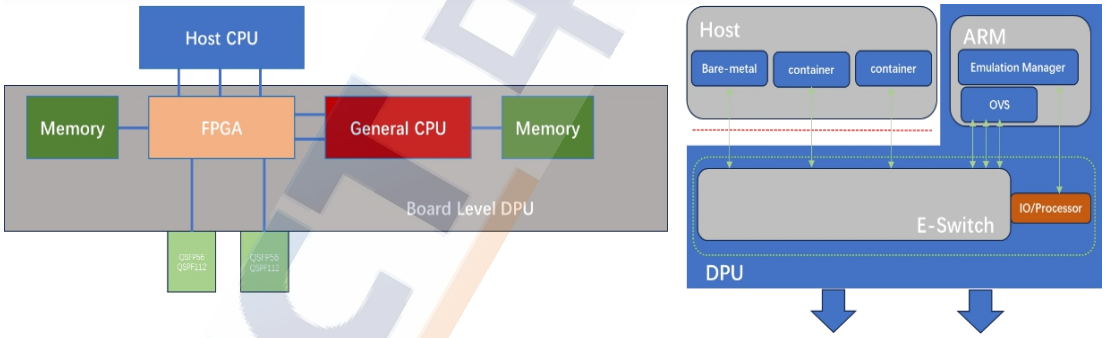
国际头部企业如 NVIDIA、Marvell、Intel 等已构建起涵盖芯片、软件栈、解决方案到客户落地的较为成熟的生态体系，形成了较高的生态粘性。国内 DPU 厂商在市场拓展过程中，需面对客户迁移成本、方案适配周期以及场景验证门槛等现实因素，商业化进程仍需完善。

此外，在新兴场景中，用户对 DPU 的功能定制化、场景适配性提出较高要求，需产业链上下游加强协同，通过联合技术验证与方案优化，提升产品成熟度，为实现规模化商用落地提供支撑。

二、DPU 关键技术发展分析

（一）计算架构持续迭代，体系效能全面升级

与 CPU 和 GPU 生态发展路径不同，DPU 发展的软硬件条件较为成熟。DPU 从发展初期便依托成熟的虚拟化软件基础和计算架构，实现了功能和性能的整体提升。当前，DPU 主要采用 CPU+FPGA、CPU+ASIC、SoC 等多核异构的主流计算架构。CPU+FPGA 架构凭借开发周期短、可编程性高、可快速进行方案验证的优势，满足用户定制化开发的需求。CPU+ASIC 架构具有性能高、功耗较低、大规模量产下成本较低的特点。



来源：中国信息通信研究院

图 2 通用 CPU+FPGA（左）和 CPU+ASIC（右）实现 DPU 示意图

SoC 凭借高集成度、性能优化以及软件协同设计等优势，已成为当前 DPU 发展的主流路线之一。在高度集成性方面，SoC 将处理器、存储器、接口等功能模块集成到单个芯片上，降低了系统的复杂性，提高了整体性能和能效比，满足智算中心等场景对高性能计算的需求。

在软硬件协同方面，通过优化软件算法和调度策略，SoC-based DPU 显著提升数据处理的速度和效率，实现高效的资源利用和更低的功耗。同时，基于 SoC 的 DPU 产品具备高性能、低功耗、低成本等优势，为智算中心等场景应用提供技术支撑。此外，随着 Chiplet 技术与先进封装技术的深度应用，正推动 DPU 向模块化设计、动态重构方向发展。当前，DPU 已从早期的网络协处理器，发展为融合控制面管理、数据面卸载与专用加速能力于一体的异构计算核心，其架构创新集中体现在软硬件协同上。

（二）端网融合深度推进，传输时延显著降低

在人工智能场景中，智算网络不仅需要支撑高速率传输，更需应对模型训练中复杂的网络场景，实现灵活的流量调度与深度的端网协同。**端网融合已成为智算产业从“智能互联”迈向“全域协同”的关键标志。**DPU 作为实现端网连接的“桥梁”，通过集成专用卸载引擎，实现网络、存储和安全等任务的硬件级卸载，降低主机性能开销。其支持 RDMA、NVMe over Fabrics 等高速协议，将主机与 GPU 或存储之间的数据传输延迟压缩至微秒级，同时借助流量整形技术优化带宽分配，有效避免网络拥塞。此外，DPU 通过硬件级虚拟化技术，打破传统处理器“单任务独占资源”的局限，实时感知终端算力负载与流量特征，为智算网络端网融合提供底层支撑。

为了满足智算网络端网融合的需求，网络接口速率已进入 400G 规模化部署，800G 商用落地的新阶段。当前主流 DPU 普遍集成单路或双路 400G 以太网接口，部分高端型号支持 800G 光模块。在物理

层，光模块普遍采用 PAM4 调制与链路自适应技术，根据传输距离动态调整信号强度与均衡参数，以优化误码率与能效。DPU 则通过标准管理接口参与链路状态监控、参数配置及重训练触发，实现端到端智能协同。此外，DPU 内置的可编程数据平面引擎支持细粒度流量调度、拥塞控制与虚拟交换功能，有效避免网络热点，构成“端网融合”的关键硬件基础。如 AMD Pensando Salina 400 DPU 支持双端口 400G 以太网，拥有完全可编程的 P4 数据路径以及 16 个 Arm Neoverse-N1 核心，满足企业、云计算和 AI 工作负载需求。云豹智能 Corsica 400G DPU 支持 PCIe Gen 5 接口，采用多级可编程架构，双通道 DDR5 内存，以及 16 个 N2 ARM 核心，为智算场景提供解决方案。

（三）高速存储加速成型，卸载能力关键跃升

DPU 的存储卸载能力是连接算力与存储服务高效协同的关键，为人工智能场景提供算力和存储资源的独立伸缩。随着 NVMe 等非易失存储介质访问延时降低，系统性能瓶颈逐渐从存储设备转向网络协议栈与主机 I/O 处理。DPU 通过硬件卸载机制，将任务从 CPU 迁移至专用数据平面，将远端 NVMe-oF 存储资源虚拟化为本地 PCIe NVMe 设备，对主机呈现为标准块设备。同时依托 NVMe-oF 协议硬卸载与 vDPA 技术，显著降低主机 CPU 开销，提升存储 I/O 效率与确定性。

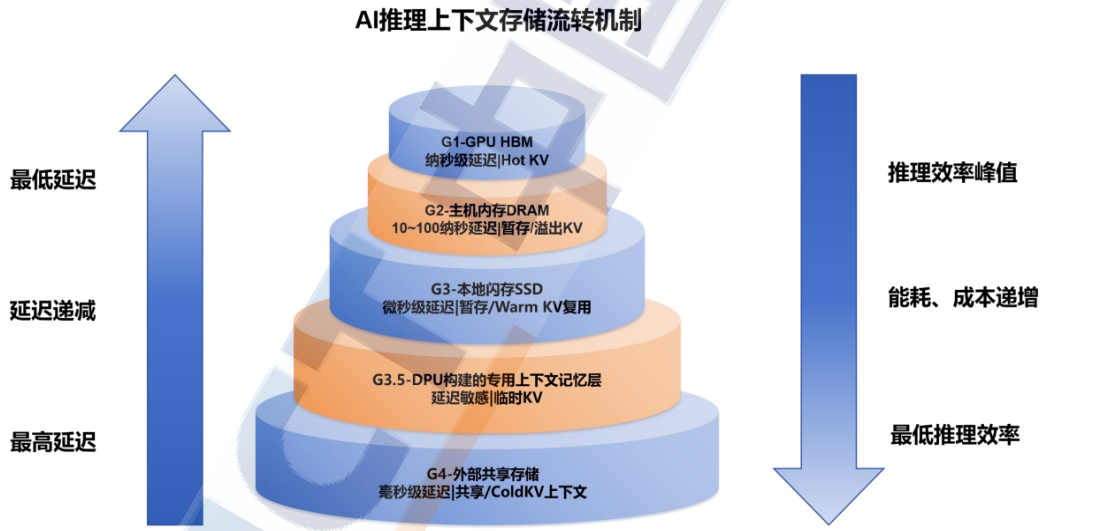
DPU 通过协议栈硬件化重构存算分离架构下的存储访问链路。当前主流 DPU 产品已实现 NVMe-oF 协议的全栈硬件卸载，将传统由主机 CPU 承担的 NVMe 队列管理、数据路径处理、加密校验等高

开销任务迁移至 DPU 内置的存储加速引擎，降低主机 CPU 占用率。根据卸载能力深度，业界实践可归纳为基础卸载、零拷贝卸载、全硬件卸载三个层级。其中，基础卸载基于 SPDK 框架支持多协议适配但数据路径仍依赖主机内核，时延较高。零拷贝卸载通过 RDMA QP 直接映射远端存储内存，消除 DMA 拷贝开销，提升吞吐与确定性；全硬件卸载数据平面由硬件处理，主机保留控制面配置，实现微秒级时延，但灵活性受限。例如天翼云研发的紫金 DPU 采用“零拷贝+硬件校验”混合卸载方案，在典型 AI 训练场景下实现存储 IOPS 的数量级提升。

DPU 存储效能提升依赖于硬件级数据处理能力的突破。在数据缩减方面，部分 DPU 产品支持硬件加速的透明压缩，在高带宽链路下显著降低有效数据传输量，减少 CPU 参与度与网络负载。在数据可靠性方面，DPU 集成端到端数据校验模块，通过 T10-PI（Protection Information）与 CRC（Cyclic Redundancy Check，循环冗余校验）机制，保障从主机内存到存储介质的数据一致性。同时，通过卸载垃圾回收元数据管理等任务，优化 SSD 写入模式，间接缓解写放大效应，延长存储设备使用寿命。对于金融、医疗等敏感场景，DPU 通过内存硬隔离技术构建存储安全沙箱，实现不同租户数据的物理隔离，满足安全监管合规要求。如 NVIDIA BlueField DPU 与 Vast Data Platform 的深度集成，通过存储协议处理，将存储服务从 CPU 解耦，释放主机算力用于 AI 训练等核心业务，满足 DPU 内置的安全引擎保障数据访问合规性需求。

（四）互联技术持续优化，集群算力高效协同

DPU 凭借在单节点内与跨节点间的双重技术架构优化，成为打通 AI 算力集群 Scale Up 与 Scale Out 瓶颈的关键核心器件。在 Scale Up 层面，业界正积极探索基于 CXL（Compute Express Link）协议的异构内存池化技术，以构建统一地址空间，降低跨芯片数据访问延迟。如 NVIDIA BlueField-4 DPU 赋能推理上下文记忆存储平台（ICMS），通过卸载 KV Cache 的 I/O 路径与元数据管理，构建高速、可共享的“记忆层”，有效加速大模型推理中的上下文复用，减少 GPU 间数据搬运开销。中科驭数等国内 DPU 厂商聚焦于低时延网络与可编程数据平面，在金融、AI 推理等场景中实现高性能数据处理。



来源：中国信息通信研究院

图 3 AI 推理上下文存储流转机制

在 Scale Out 层面，DPU 集成 400G/800G 高速网络接口，通过硬件加速实现 TCP/IP、RDMA 等协议的全栈卸载，将节点间数据传输的协议处理、流量调度等开销从 CPU 剥离，大幅提升网络吞吐、降低端到端时延，为 AI 集群提供高带宽、低延迟、高可靠的通信能力，

已成为 AI 服务器 Scale-out 分布式架构后端网络的核心标配组件。腾讯云星脉网络结合 DPU 通过硬件级 RDMA 卸载与自研流量调度算法，在万卡级 AI 集群中实现微秒级节点通信。百度智能云通过百舸异构计算平台与 DPU 协同，优化大规模参数服务器架构下的梯度同步效率，提升分布式训练吞吐。

（五）软硬协同创新突破，技术标准加速落地

以大模型训练、分布式推理为核心的新一代智算中心体系中，DPU 已成为软硬协同设计的核心枢纽与关键载体。软硬协同不再局限于辅助优化，通过“硬件可编程+软件可定义”的深度融合，重构计算、网络、存储与管理的全栈技术体系，为 AI 基础设施提供高效率、高可靠、高兼容的底层支撑。在软硬协同方面，DPU 推动关键系统功能从主机 CPU 向专用处理器迁移，实现控制面与数据面的深度解耦。在网络加速场景中，传统算力中心依赖 CPU 完成网络协议处理与流量转发，在高带宽、高并发场景下易形成性能瓶颈。依托 DPU 硬件转发引擎与配套软件栈优化，可将 TCP/IP、RoCE、SR-IOV、OVS(Open vSwitch，开源虚拟交换机)等网络功能全面卸载，将端到端通信延迟降至微秒级。上层软件需定义网络拓扑、通信规则与 QoS 策略，由 DPU 硬件完成高速转发、流量隔离、拥塞控制等操作，使跨节点通信效率提升，为万卡级 AI 集群提供稳定无损的高速互联通道。

在平台建设方面，推动云管理软件安全位置从主机 CPU 向 DPU 处理器迁移，实现管理平面与加速引擎的解耦。通过 DPU OS 与 Host OS 的协同机制，支持传统云平台软件在 DPU 上的高效运行，降低传

统软件向 DPU 迁移的适配成本。

在标准化建设方面，针对当前 DPU 产品接口碎片化、生态粘性不足等问题，产业界正加速推进统一规范。在软件层面，围绕管理、网络、存储、安全等核心功能，构建统一的交互接口与功能规范体系。在硬件层面，以统一整机结构、边带信号及管理运维为导向，解决异厂家设备适配难题。目前，中国信通院联合开放数据中心委员会（ODCC）牵头开展 DPU 技术研究、测试验证和产业生态工作，围绕 DPU 参考架构、CXL 内存扩展接口、网络卸载能力等重点方向，先后发布《基于 DPU 的新一代存算分离存储架构：重构算力中心存储范式》《基于 DPU 的高性能存储网络技术报告》等一系列成果，为构建开放、协同、可互操作的 DPU 产业生态提供坚实支撑。

三、DPU 产业生态发展分析

随着大模型训练推理、边缘计算等场景规模化落地，全球 DPU 市场进入爆发增长期。沙利文 2026 年报告数据显示，全球 DPU 市场规模由 2021 年的 649.93 亿元增长至 2025 年的 1964.91 亿元，年复合增长率为 31.86%；预计到 2030 年将以 17.07%的年复合增长率增长至人民币 4362.39 亿元。从产业发展基础来看，DPU 规模化应用依托硬件组件生态的持续成熟。IO Die 作为 I/O 操作处理核心单元，正与 DPU SoC 深度集成，持续提升数据吞吐效率。Switch 模块通过开放标准推广，有效增强产品互操作性。

（一）国内：政策、技术与需求协同驱动的快速发展

1.政策体系完善，释放市场增长红利

国家层面持续出台重磅政策，推动 DPU 产业规模化发展。《算力基础设施高质量发展行动计划》《关于深入实施“人工智能+”行动的意见》等政策的实施，引导算力基础设施的规模化建设，推动 DPU 技术创新。**地方政府**同步发力，北京、上海、深圳、成都等重点城市纷纷出台配套政策，通过税收减免、设立专项基金等方式，为 DPU 研发企业提供人才、资金等全方位支持。沙利文 2026 年报告数据显示，2025 年中国 DPU 市场规模达到约人民币 500 亿元，预计到 2030 年将增长至超人民币 1,200 亿元，是算力基础设施领域中增长迅速、成长空间广阔的细分市场之一。

2.技术创新突破，多元主体协同创新

国内 DPU 产业已形成多元主体协同创新的格局，技术突破成效显著。大型互联网企业依托云计算场景优势，研发定制化 DPU 产品。华为自研 SP900 系列 DPU 搭载 24 核 Hi1822 处理器，具备网络、存储加速及虚拟化卸载功能，提升大数据查询性能 40%。百度推出太行 DPU，基于自研芯片架构，具备多平台兼容、全栈功能卸载、混合场景适配、流量智能调度四大核心能力。腾讯自研水杉和银杉两代 DPU，聚焦高性能网络卸载与安全隔离，在视频直播、金融交易等高并发场景中显著降低主机 CPU 开销，提升服务稳定性。**新兴 DPU 研发企业**立足技术差异化路径，在芯片设计与功能集成环节持续创新。云豹智能旗下云霄 S10DPU 智能网卡具备 400G 全线速转发能力，在吞吐、延迟、卸载效率等关键性能指标上，已进入国际主流竞争力梯队，面向下一代高速网络场景，新一代产品 800Gbps/1.6Tbps 高速率的 DPU

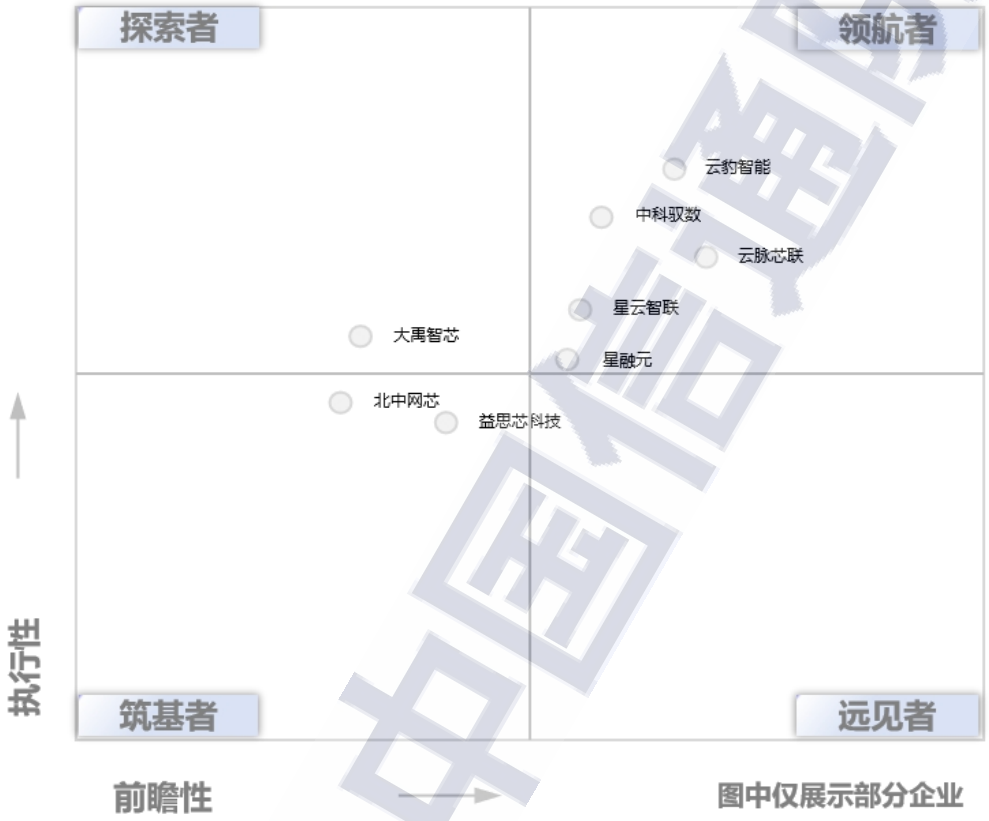
产品即将推出。云脉芯联推出 metaVisor-200 高性能 AI DPU 产品，支持云盘启动、VPC 网络卸载加速、RoCEv2 Overlay 和 RDMA 统一纳管运维监控等核心能力。中科驭数 DPU 芯片 K2-Pro，采用纯自研 KPU 架构，融合了网络卸载、存储卸载、安全卸载、计算卸载等功能。

3.产业梯队分化，生态协同持续演进

中国 DPU 市场整体呈现较为集中的态势，国际头部 DPU 厂商占据主要市场份额。沙利文 2026 年报告数据显示，NVIDIA 凭借长期的芯片架构积累、成熟的数据面处理能力以及完善的软硬件生态体系，占据市场首位，云豹智能排名第二，并在 DPU 厂商中位列第一，是目前国内可以实现全功能 DPU 芯片大规模量产的 DPU 厂商。

为科学量化目前 DPU 企业发展状况及相关产品情况，本报告构建综合评价体系，细化核心产品与服务、市场响应能力、客户体验、运营管理、营销战略、销售策略、业务模式以及科技创新能力等指标，全景式剖析我国 DPU 企业产业格局。从整体产业格局来看，企业立足自身资源禀赋和发展定位，在技术创新、场景落地和基础支撑等方向差异化布局，构筑起多层次、立体化的产业生态体系。技术创新方面，部分企业聚焦芯片研发和算网融合等方向，加快核心技术攻关，构建标准化平台，打造标杆应用场景，引领行业技术迭代与模式创新。场景赋能方面，企业凭借对市场需求的快速响应和扎实的落地能力，在工业、交通等细分领域推动 DPU 技术与实体经济深度融合，加速技术向商业价值的转化。基础支撑方面，企业聚焦于基础技术研发和设施建设，在算力芯片、网络设备等底层支撑领域持续深耕，为应用

创新提供坚实保障。未来，随着技术成熟度提升、标准体系完善和市场需求释放，各类主体有望在协同中进一步强化优势，共同推动我国 DPU 产业迈向高质量、可持续发展新阶段。



来源：中国信息通信研究院

图 4 DPU 产业魔力象限

（二）国外：需求、硬件与生态协同支撑的规模应用

1.实施战略并购，构建产业闭环

国外 DPU 市场起步早于中国，由 NVIDIA、Marvell、Intel、AWS 等国际巨头主导，技术路径成熟，应用场景聚焦高端算力与新兴领域，目前正处于规模化扩张阶段。**海外巨头通过战略收购整合资源，提前布局**DPU **领域并构建竞争壁垒。**AWS 于 2015 年收购 Annapurna Labs，推出了 DPU 早期形态的 Nitro 芯片；2020 年，NVIDIA 收购 Mellanox

Technologies，同年推出 BlueField-2 DPU；Intel 在 2015 年收购 Altera 后，于 2024 年 5 月发布了 IPU E2100 DPU；微软于 2023 年收购 Fungible，依托其存储优化与低延迟 DPU 技术补齐基础设施短板，并在 2024 年 11 月推出了自研的 Azure Boost DPU。海外巨头通过并购整合与技术迭代，构建了完整的产业闭环，持续抢占 DPU 产业生态的主导权。

2.主导高端市场，技术迭代加速

全球 DPU 市场呈现高度集中态势，NVIDIA、博通、Intel 三大厂商占据高端市场主导地位。AMD、Marvell、AWS、Microsoft 等厂商持续推出 DPU 及同类架构产品，技术迭代速度加快。如 NVIDIA 的 BlueField 系列芯片已到达第四代，支持 800 Gb/s 带宽，计算能力较前代提升 6 倍，使用 ARM 内核，为存储、网络和安全提供硬件卸载。2024 年 10 月 AMD 推出了 Pensando Salina 400 DPU 产品，可高效处理海量 AI 数据负载、传输及高算力业务需求。

3.依托核心技术，构建生态壁垒

国外厂商在 DPU 核心技术与软件生态领域形成显著优势。技术层面，在制程工艺、Chiplet 集成、AI 加速、高速互连等领域占据代际优势。台积电占据了全球 5nm 及以下先进制程 87%的市场份额，3nm 工艺订单近乎全部份额，NVIDIA BlueField-4、AMD Pensando Salina、Marvell Octeon 10 等头部 DPU 均采用其先进制程。**生态层面，**头部厂商积极参与 OpenCAPI、CXL 等开放标准制定，推动成熟软件生态形成。NVIDIA 通过 DOCA 软件栈降低开发门槛，构建全栈应用

生态。AWS 打造 Nitro 系统闭环，实现与云服务的深度绑定。Linux 基金会宣布新的开放可编程基础设施（OPI）项目，为基于 DPU 和 IPU 技术的下一代架构和框架培育一个社区驱动的、基于标准的开放生态系统，但不同生态间仍存在兼容壁垒，未完全打破割裂的格局。

四、DPU 典型应用场景

DPU 已成为算力中心重要的组成部分，其应用场景正从算力中心核心延伸至边缘计算、智能驾驶等新兴领域。DPU 通过集成高性能多核处理器、高速网络接口及可编程加速引擎，实现网络虚拟化、存储加速和安全功能的硬件级处理，为多样化应用场景提供了坚实的底层支撑。

（一）通用算力中心：破解虚拟化税，重塑原生架构

在云计算领域，DPU 破解传统架构下的“虚拟化税”问题，推动云算力中心架构向云原生转型。DPU 通过将虚拟交换机、存储控制器、安全策略执行等基础设施服务，从服务器主机 CPU 卸载至 DPU，大幅提升主机资源利用率，降低虚拟化损耗。同时，强化安全隔离能力，保障业务运行稳定，为产业高质量发展筑牢底层支撑。

目前，DPU 已在国内外主流云服务商中实现规模化部署，实践成效显著。国内方面，天翼云自研 DPU 芯片“紫金”采用 SOC+FPGA 设计，通过芯片级的硬件加速，提升网络 PPS 性能以及存储 IOPS 性能，网络时延降低至原来的四分之一。百度智能云基于太行 DPU2.0，打造新一代太行弹性计算产品，具备多平台兼容、多场景适配、多协议支持、多业务承载四大核心能力，可高效支撑超大规模智算集群各

类应用需求。国外方面，AWS Nitro、微软 Azure Boost 系统通过 DPU 全面卸载虚拟化功能，提升云服务器性能及传输效率。微软发布了首款数据处理器 Azure Boost DPU，集成 Azure 硬件安全模块用于加密功能，功耗约为传统 CPU 的三分之一，同时提供高达四倍的性能，保障云服务高效安全。

（二）智能算力集群：重构内存与通信，定义新智算底座

在人工智能与高性能计算（HPC）领域，DPU 是破解“数据搬运墙”、“I/O 墙”等性能瓶颈，释放 GPU 算力潜力的核心组件，支撑超大规模 AI 模型训练。在大规模 AI 集群训练过程中，GPU 之间需频繁交换梯度等中间数据，传统网络架构易出现数据传输阻塞、延迟偏高问题，严重制约训练效率，而 DPU 通过无损 RDMA 网络，保障数据在数千张 GPU 卡间高效、无阻塞流动，提升训练效率，为超大规模 AI 模型研发提供有力支撑。

在超大规模 AI 训练集群中，互联瓶颈的本质正从单纯的“网络带宽”问题，演进为复杂的“数据供给与协同效率”挑战。以 NVIDIA BlueField-4 DPU 及其 Vera Rubin 架构为代表，DPU 不再仅优化数据流动，而是主动重构内存层次，通过驱动池化 NAND 存储资源，在 GPU HBM（热数据）与远端冷存储之间创建一个由 DPU 统一调度的“温数据层”中间层，为检查点、激活值等海量临时数据提供近计算侧的高速存取空间，大幅减少跨机房 I/O 流量，缩短训练周期，支持灵活的模型/数据并行策略，使万卡集群能动态共享中间状态，提升资源调度效率。全新架构下，DPU 已演化为智算集群的“全局数据中枢”。

通过深度协同网络、存储与计算资源，解决单点瓶颈，重塑了高性能智算底座的系统范式。

（三）5G 边缘计算：赋能低时延业务，推动算网融合

DPU 凭借低功耗、小体积、高性能的优势，在资源受限的边缘节点提供数据处理能力。DPU 支撑 5G 用户面功能下沉、移动边缘计算节点部署，高效完成数据包高速转发、流量整形、网络切片策略执行等任务，满足工业互联网、VR/AR、车联网等领域对超低时延、本地化处理的苛刻要求，为边缘计算场景规模化落地提供坚实底层支撑。

在 5G 网络边缘部署方面，国内外企业积极开展实践，成效显著。国内方面，中兴通讯联合上海电信、浦东公交，采用自研 DPU 构建 5G-A 智能网络，支撑智慧城市公交云巡检应用，通过“5G+AI 云边端”协同架构，优化城市道路维护与交通管理效率，为智慧市政建设提供有力支撑。中国联通边缘算力平台中引入 DPU 作为 CPU 卸载引擎，将网络虚拟化、存储 I/O、安全加解密及硬件资源池化等基础设施层任务，通过硬件级卸载从主机 CPU 剥离，提升算力利用率、降低端到端时延。**国外方面**，红帽研发由 DPU 驱动的可组合计算基础设施，可精准适配边缘数据中心对高性能、低时延与高安全等级的严苛部署要求。

（四）新兴智能场景：支撑技术迭代，拓展应用边界

DPU 正逐步成为连接物理世界与数字空间、支撑智能技术规模化应用的关键基础设施，重点赋能智能驾驶、元宇宙、数字孪生、量子计算等场景，对拓展数字应用边界具有重要意义。随着智能技术的

快速发展，新兴场景对数据处理效率、时延控制、资源协同能力的要求持续提升，DPU 凭借专用硬件架构的独特优势，可有效破解各类应用瓶颈，为新兴智能场景规模化落地提供坚实支撑。

在智能驾驶研发领域，DPU 通过高效数据处理能力重构技术发展路径，有效解决海量传感器数据处理效率低、模型迭代周期长等问题。NVIDIA 在其下一代智能驾驶的 NVIDIA DRIVE 平台中，集成了 DPU 芯片。通过 DPU 实现高效的数据交换、网络加速和存储卸载，从而降低端到端时延，满足 L3 及以上级别自动驾驶对实时性要求。在元宇宙、数字孪生领域，西门子能源利用 NVIDIA Omniverse 创建数字孪生，来预测热回收蒸汽发生器部件的腐蚀情况，将原本需要数周的计算缩短到几小时。在量子计算领域，DPU 作为经典计算集群的协处理器，高效执行量子测量数据的筛选、错误缓解与预处理任务，降低主计算节点的负载和系统整体延迟。

五、DPU 发展趋势与建议

DPU 作为算力基础设施的核心组件，其发展需在技术、应用、标准等方面协同推进。通过技术创新、场景落地和标准引领，我国有望在 DPU 领域实现从“跟跑”到“并跑”再到“领跑”的跨越，为数字经济高质量发展提供坚实支撑。

（一）架构革新与智能融合并进，突破效能瓶颈制约

DPU **从硬件架构、软件体系及软硬协同三方面优化发展，强化性能与适配性，支撑场景高效运行。硬件层面**，多核异构已成为 DPU 硬件设计的主流方向，通过可编程网络处理引擎、专用安全加速模块、

高速存储控制器等模块化组件的集成，实现网络、存储、安全任务的专业化处理。Chiplet 技术与先进封装技术的深度应用，推动 DPU 向模块化设计、动态重构方向发展，支持 200Gbps 以上线速处理能力，部分高端产品已支持 800G 带宽。同时，硬件设计聚焦场景适配需求，在高端智算场景强化多接口集成与数据吞吐能力，在边缘场景推进低功耗芯片设计，形成覆盖不同算力需求的硬件产品矩阵。**软件层面**，DPU 软件架构已逐步形成硬件抽象、加速框架、虚拟化、编排管理的分层体系，通过标准化 API 实现与上层系统的灵活对接，降低应用适配成本。可编程能力持续增强，微码开发框架与调试工具链日趋成熟，支持 TCP/IP、OVS 等协议的硬件加速模板封装，满足不同场景的定制化需求。**软件与硬件协同层面**，通过硬件功能模块化与软件定义能力结合，实现 DPU 的“按需配置”。在异构计算场景中，DPU 通过高速网络优化 GPU 集群间通信效率，显著降低跨节点数据传输延迟。

智能与协同能力升级，重塑算力调度格局。DPU 通过硬件级可编程流水线与实时遥测能力，为上层智能运维系统提供高精度网络状态数据，支撑流量预测、异常检测与自动化故障恢复等智能运维功能。例如，基于历史流量模式，上层系统可动态调整带宽分配，当检测到链路异常时，自动切换至冗余路径，提升系统可靠性。在异构计算场景中，DPU 通过卸载 RoCE 等高速网络协议栈，优化 GPU 集群间的跨节点数据传输效率，显著降低通信时延。通过构建 CPU、GPU、

DPU 协同的异构算力架构，实现控制面、计算面与数据面的深度解耦，为 AI 大模型训练与推理提供支撑。

（二）场景渗透与价值释放加速，赋能行业转型升级

DPU 的应用场景正从传统的算力中心领域，逐步向多个新兴领域拓展。在算力中心场景中，DPU 作为主机的数据出入口，具备标准网卡能力，有效解决云主机实例与虚拟化软件共享计算资源时出现的资源争抢限制、计算特性损失以及裸金属管理等难题。在网络协议处理、存储性能优化和安全防护体系等方面发挥重要作用，实现虚拟交换、流量均衡、QoS 保障，提升存储集群的 IOPS。**在细分场景中**，DPU 成为破解“算力浪费”难题的关键。通过卸载虚拟化、网络协议栈等底层任务，释放 CPU 算力，解决传统“算力中心税”问题，提升单服务器的有效算力输出。如在 AI 大模型推理中，DPU 通过高效数据搬运与安全加密，将部分 KV Cache 卸载至低成本存储介质，显著降低推理成本。

DPU 的应用场景正加速向智能驾驶、工业互联网等新兴领域拓展。在智能驾驶领域，DPU 作为车载安全网关，支撑 V2X 通信与远程 OTA 安全更新。在工业互联网场景，DPU 通过时间敏感网络与确定性调度，满足亚毫秒级实时通信需求，支撑智能制造设备协同运作。

（三）体系构建与生态协同提速，规范产业发展秩序

构建完备的标准体系，助力 DPU 规范化、标准化发展。DPU 标准处于发展的初级阶段，缺乏行业规范或开源标准。需要围绕 DPU 硬件、软件，打造面向不同类型客户的标准化产品，助力 DPU 行业

的发展。**硬件方面**，从服务器结构、功耗、运维策略等维度推进整机兼容性设计，适配多厂商生态。其中架构方面，加速 DPU 与主机间的管理接口、控制面通信协议标准化，并制定 CXL 在 DPU 场景下的应用指南，解决跨平台兼容性问题。通过 ODCC、CCSA 等组织推动 Scale Up/Out 互联标准，降低 AI 集群部署成本。**软件方面**，聚焦 RDMA 接口标准化、NVMe-oF 协议卸载、虚拟化卸载、安全功能卸载等关键技术，满足 AI 训练场景低时延网络、大数据高性能存储读写等场景需求。

生态协同方面，政策与市场双轮驱动，推动跨厂商统一编程模型与硬件抽象标准制定。开源社区成为标准推广的重要载体，通过联合头部企业、高校及科研机构培育 DPU 开源生态，将标准化接口与编程框架融入开源项目，促进标准在云原生、AI 原生等场景的规模化应用。在国际合作中，应积极对接全球技术体系，在吸收国际先进实践的同时，立足本土算力架构与行业应用特点，发展兼具兼容性与自主性的 DPU 标准路径，为国产技术融入全球生态奠定基础。

中国信息通信研究院 云计算与数字化研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62300095

传真：010-62304958

网址：www.caict.ac.cn