算力产业发展方阵
Computing Power Advanced Matrix

CAICT 中国信通院

面向 AI 的智能网卡产业发展分析报告（2026 年）

算力产业发展方阵

中国信息通信研究院云计算与数字化研究所

2026 年 5 月

前　言

当前，全球数字经济与人工智能产业加速演进，算力已成为重塑全球经济结构和竞争格局的关键力量。作为连接计算、存储与网络的核心枢纽，智能网卡已从传统网络接口的辅助部件，演进为智算中心连接计算、存储与网络的关键基础设施。

我国智能网卡产业正迈向高质量发展的新阶段，从政策维度看，《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文件的出台，为智能网卡发展指明方向。从技术维度看，FPGA 可编程、ASIC 专用芯片、SoC 集成架构在差异化竞争中协同演进，PCIe、CPO、RDMA 等技术的突破，为智能网卡实现高吞吐、低时延、可编程奠定基础。从产业维度看，国际巨头凭借“硬件+软件”的生态壁垒，抬高市场准入门槛。国内产业界，加速从“单点突破”向“全链协同”转变。未来，我国智能网卡将不断迭代创新架构，加速场景渗透释放价值，持续创新发展。

本报告立足全球智算产业发展现状，系统梳理智能网卡的发展背景、技术演进和典型应用，分析智能网卡产业生态格局，提出前瞻性发展建议，旨在为产业界提供参考，助力智能网卡高质量发展。

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研究团队：dceco@caict.ac.cn。

目 录

图 目 录

表 目 录

一、智能网卡发展概况

（一）定义与核心特征

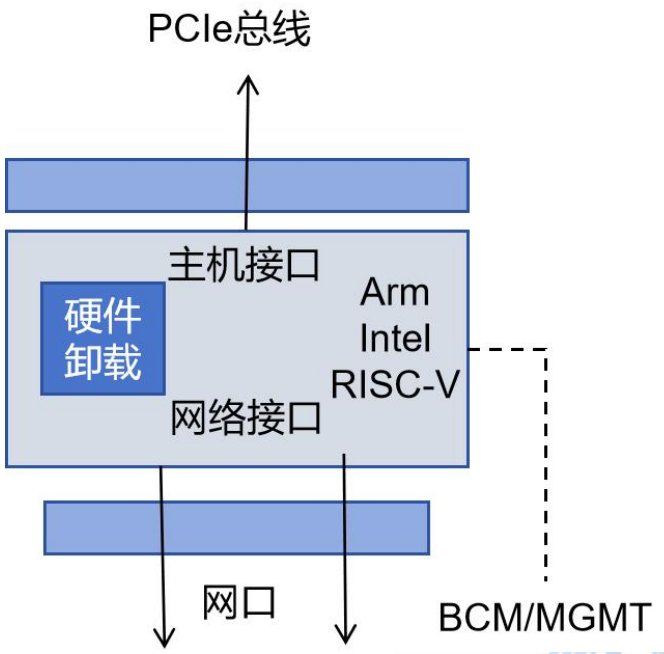
随着人工智能规模化落地及智算需求爆发，算力网络正从“计算驱动”向“算网融合”跃迁，对算力网络性能提出更高要求。算力网络需具备超低时延、超大带宽、确定性传输与全局灵活调度的能力，为算力高效利用与算网深度协同提供坚实支撑。在此背景下，智能网卡从算力基础设施的高速互联纽带，演进为支撑 AI 集群全栈加速、算网高效协同的关键核心载体。

智能网卡作为面向算力基础设施的新型网络适配器，兼具可编程能力与异构计算能力，高效卸载网络、存储及安全类负载。其主要采用 FPGA[1]可编程芯片、ASIC[2]专用芯片或 SoC[3]集成架构进行设计，灵活集成 Arm、x86 或 RISC-V 架构处理器，为网络任务处理提供异构算力支撑。同时搭载专用硬件卸载引擎，将数据加密、协议解析等密集型任务从主机 CPU 剥离并在网卡侧完成，有效释放主机算力、提升系统运行效率。智能网卡原生兼容 SDN、虚拟化等技术，适配高速网络与算网融合场景的多样化需求。此外，设备通过 BMC、MCTP、NC-SI、Redfish 或厂商自定义管理通道等接口实现配置下发与状态监控，保障高速数据传输与算力卸载的稳定性和高效性。

[1] 在可编程阵列逻辑（PAL）、通用阵列逻辑（GAL）、可擦除可编程逻辑器件（EPLD）等器件的基础上进一步发展的产物。它是一种可完成通用功能的可编程逻辑芯片，即可以对其进行编程实现某种逻辑处理功能。

[2] 即专用集成电路，是一种根据特定应用需求进行定制化设计和制造的集成电路芯片。与 CPU、GPU 等通用芯片相比，ASIC 针对特定算法或功能进行硬件级优化，因此在执行特定任务时具有更高的性能、更低的功耗和更小的体积。

[3] 专为特定应用场景定制的高度集成化系统级芯片。它将处理器核心、内存控制器、高速接口及专用硬件加速引擎等完整系统功能，高度集成于单一硅片之上，并兼顾了通用可编程的灵活性与专用硬件的高效能。。



来源：中国信通院，ODCC

图 1 智能网卡

智能网卡并非传统网卡的简单技术迭代升级。传统网卡仅承担数据链路层基础数据转发功能，智能网卡依托 ASIC 专用芯片、FPGA 可编程芯片或 SoC 集成架构，将原本由主机 CPU 承载的网络传输、信息安全、数据存储等业务负载，卸载至网卡硬件侧执行，显著降低主机 CPU 资源开销，有效释放主机端算力用于核心业务运行，实现算力卸载与网络内生智能的双重赋能。一方面，通过硬件级加速能力，将 CPU 从非计算密集型任务中剥离释放，间接提升集群整体有效算力利用率，优化算力资源配置。另一方面，依托 200G/400G/800G 的超高速互联接口，搭建节点间、芯片间的低时延“数据高速路”，着力破除 AI 大模型训练与推理中通信瓶颈，缓解算力集群因通信滞后带来的算力利用低效问题。

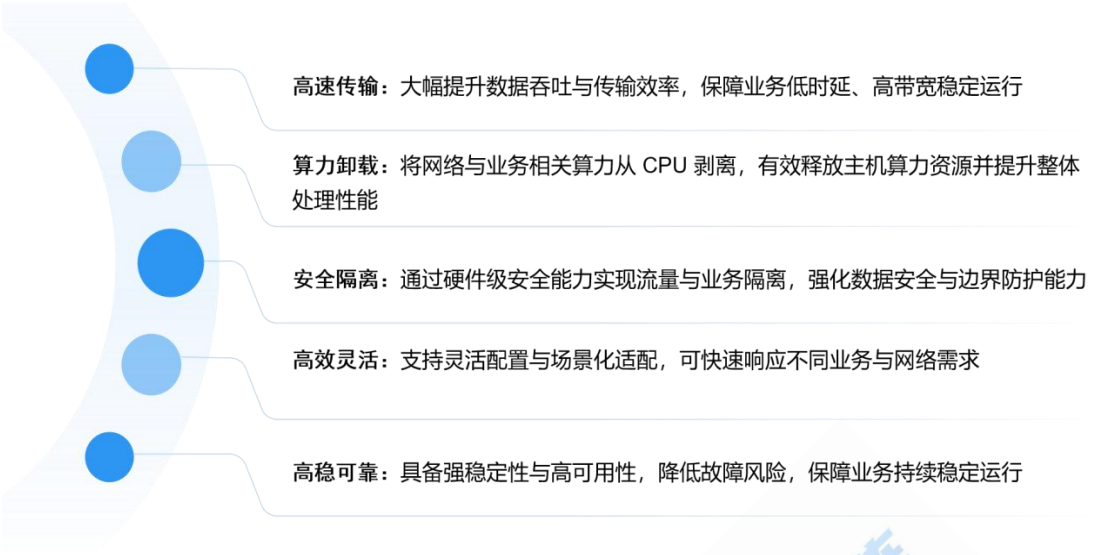
表 1 智能网卡与传统网卡对比

类别	传统网卡	智能网卡
数据处理能力	仅支持基础网络数据链路层转发	具备网络处理、协议加速、硬件加解密、存储加速等多元业务处理能力
CPU 负载	网络、虚拟化、安全等任务均由主机 CPU 承担，负载较高	通过硬件卸载实现网络与业务加速，显著降低主机 CPU 占用
可编程性	功能固化，无灵活可编程能力	支持 P4、C/C++等可编程框架，可按需定制网络功能
适用场景	通用服务器、常规网络接入等简单场景	智算中心、高性能计算、云计算、分布式存储、AI 大模型训练和推理等高性能场景

来源：IDC，中国信息通信研究院

智能网卡具有高速传输、算力卸载、安全隔离、高效灵活、高稳可靠五大核心特征，实现从传统网络适配器向智能算力加速引擎的跃迁。在高速传输方面，智能网卡依托 100G-800G 超高速接口，深度集成 RDMA（远程直接内存访问）、RoCEv2（基于融合以太网的 RDMA）等低延迟互联技术，支持微秒级低时延网络通信，为构建微秒级端到端数据传输通道提供硬件基础，提升跨节点、跨芯片之间的数据吞吐与稳定性，有效解决 AI 训练集群中大规模数据交互的时延与抖动问题，成为智算中心高速网络的核心硬件支撑。**在算力卸载方面，**智能网卡内置 ASIC 专用芯片、FPGA 可编程芯片或 SoC 集成架构的硬件加速引擎，可将虚拟化开销、网络协议解析、数据加解密、NVMe-oF 存储加速等算力密集型负载，从主机 CPU 卸载至网卡本地执行。此

模式下，主机 CPU 可聚焦于模型训练、业务逻辑处理等核心任务，显著提升算力中心整体算力密度与运行效率。**在安全隔离方面，**智能网卡在硬件层面实现安全策略的固化与内生执行，通过硬件级防火墙、精细化访问控制、端到端数据加密及可信执行环境等技术，为云环境多租户、智算集群等场景提供高强度安全隔离能力，从底层保障算力中心的网络安全与数据隐私。**在高效灵活方面，**智能网卡支持 P4、C/C++等多语言可编程框架，可根据业务场景需求灵活定制网络功能，实现快速迭代与动态适配，同时配合软件定义的资源调度，可在智算中心、边缘节点、云原生等多种环境下实现即插即用与弹性扩展。**在高稳可靠方面，**智能网卡采用高可用设计，具备硬件冗余、链路聚合、热插拔、故障自愈等能力，可在长周期、高负载运行场景下保持稳定可靠，确保大规模算力基础设施连续稳定运行。本质上，智能网卡正在从“网络接口设备”演进为“分布式算力基础设施控制节点”，成为 AI 时代算力中心架构的重要组成部分。



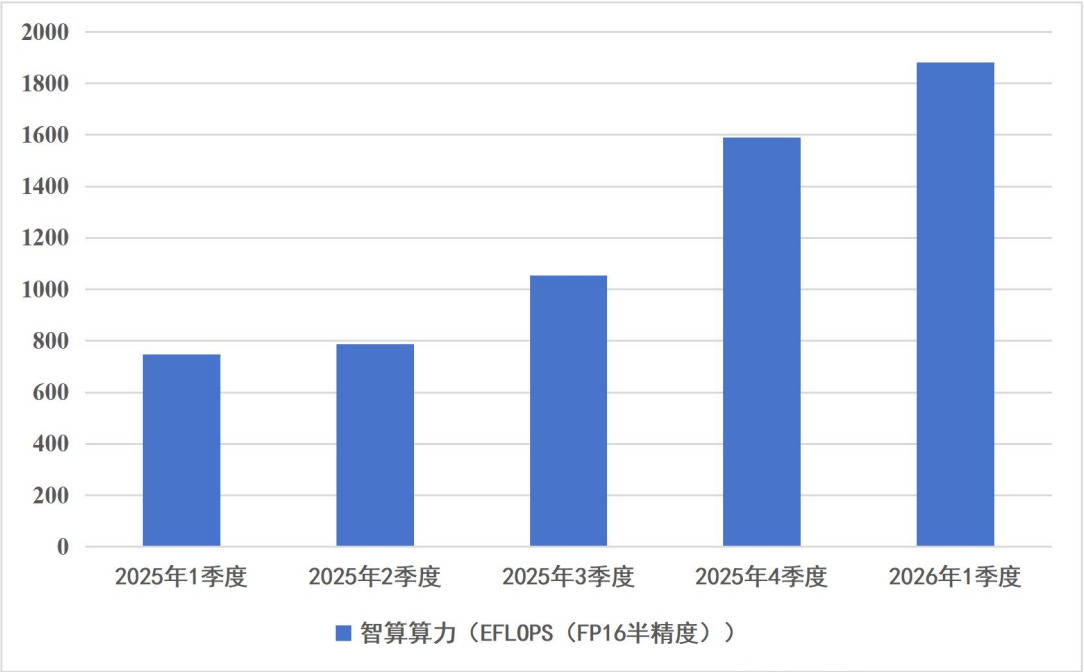
来源：中国信通院，ODCC

图 2 智能网卡核心特性

（二）发展背景

1、智算算力爆发催生硬件卸载刚性需求

随着数字经济纵深推进，我国智能算力规模持续突破，截至 3 月底，我国智能算力规模达 1882EFLOPS(FP16)。十万卡级集群、万亿参数量大模型、PB 级训练数据与 EB 级存储成为行业标配，算力供给能力实现量级跃升。此外，Open Claw、Open Manus 等自主智能体的全面爆发，7×24 小时自主执行、多工具联动、高频并发特性，带来词元（Token）消耗与算力需求的量级跃迁，推理算力需求较传统 AI 对话提升百倍以上，万卡/十万卡集群的网络转发、协议处理、安全隔离、存储加速等非核心负载被无限放大，成为制约算力高效释放的关键瓶颈。算力供需矛盾升级为“结构错配、效率瓶颈”的系统性难题。



来源：中国信息通信研究院

图 3 2025-2026 年 3 月我国智能算力规模

当前智算集群中，CPU 需同时承担核心计算与 RDMA 协议、虚拟交换、数据加解密、存储转发、虚拟化隔离等非核心负载，在超大规模并发场景下容易成为瓶颈。以十万卡集群为例，仅网络协议处理与存储转发一项，就占用大量的 CPU 算力。在此背景下，硬件卸载成为支撑超大规模智算集群的硬性刚需，以智能网卡为核心载体，内置 ASIC/FPGA 专用加速芯片，将 RDMA、vSwitch、NVMe-oF、安全加密、流量管控等非核心负载从 CPU/GPU 中完整剥离，实现核心算力专注 AI、专用硬件处理基础设施的分层解耦与专业化分工，可将 GPU 利用率大幅度提升，端到端时延降至微秒级，成为破解算力洪峰、保障智算集群高效稳定的核心技术路径。

2、智算网络升级上升为国家战略支点

美国、欧盟、日韩等主要经济体均将高性能算力与全域算网协同纳入国家科技与产业顶层规划。美国依托“星际之门”计划、《赢得竞赛：美国人工智能行动计划》等政策构建覆盖本土及全球节点的智算调度体系，支撑大模型训练、国防科研与产业创新全链条。2025 年 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，为包括智能网卡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。**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 1 月英国发布《人工智能机遇行动计划》，将建立多个“人工智能增长区”，以促进人工智能数据中心建设。2025 年 12 月，**韩国**公布 K 半导体战略，计划投资超过 700 万亿韩元，目标建成全球最大芯片产业集群，抢占 AI 芯片竞争优势。

面向全球人工智能产业竞争新格局，我国持续强化智算网络顶层设计与统筹推进。自 2023 年起，《算力基础设施高质量发展行动计划》与《关于深入实施“东数西算”工程加快构建全国一体化算力网的实施意见》等政策相继落地，确立了全国一体化算力网络的总体框架，智算网络正式迈入政策红利释放与市场需求爆发的黄金发展期。2025 年，工信部启动算力强基揭榜行动，围绕高性能数据处理器、基于 RoCE 的智算网络等关键方向部署揭榜任务；同年 8 月，国务院印

发《关于深入实施“人工智能+”行动的意见》，明确提出支持人工智能芯片技术攻关与使能软件生态建设，加快推进超大规模智算集群技术突破与工程化落地。当前，智算网络建设正加速完成从“单点算力中心”向“跨区域、跨层级、跨类型全域协同体系”转型，智能网卡作为智算集群的智算网络调度核心与数据交互枢纽，是打通算力节点、实现算力高效调度与弹性供给的核心载体，为算力资源高效协同调度提供坚实支撑。

3、AI 应用场景催生智算集群规模化发展

人工智能技术正从研发验证阶段全面走向规模化落地，以行业场景为牵引的应用突破成为产业发展主线，智能化转型向全域普及加速延伸。在生产制造、金融服务、智慧交通、医疗健康、政务服务、科研创新等关键领域，AI 深度融入业务流程与核心环节，大幅提升生产效率、决策精度与服务能力。以大模型、多智能体为代表的新一代 AI 技术持续下沉，对智算集群的算力效率、传输时延、安全隔离、协同能力提出更为严苛的要求，而智能网卡作为智算网络的核心硬件载体，成为适配场景需求、破解智算瓶颈的关键支撑。

从行业落地来看，AI 应用场景的规模化扩张，也推动智能网卡从“可选配件”升级为“必配硬件”，在万卡级智算集群、智算云平台、边缘智算节点等核心场景率先实现规模化部署，并逐步向全行业算力中心渗透。随着 AI 场景的持续破局与需求升级，进一步推动智能网卡在性能、功能、兼容性上持续迭代，推动其与 GPU、算力调度平台、

场景应用深度适配，加速智能网卡规模化部署进程，为 AI 全领域渗透筑牢智算网络硬件底座。

4、智能网卡成为超大规模智算集群的核心支撑

智能网卡的技术发展依托网络互联技术的迭代升级，传统算力中心以通用网卡为基础，仅完成基础 TCP/IP 转发，所有网络、存储等负载均由 CPU 处理，适配通用计算场景。随着云计算虚拟化普及，智能网卡应运而生，依托 FPGA/ASIC 实现部分协议卸载，初步释放 CPU 算力资源。进入智算时代，面对十万卡级集群、万亿参数大模型的极致算力需求，智能网卡作为迭代升级的核心硬件，集成专用加速引擎，搭载 200G/400G/800G 高速网络接口，采用 ASIC/FPGA/SoC 专用架构，实现全栈基础设施负载的硬件级卸载，最终形成 CPU 负责通用业务控制，GPU 专注 AI 核心计算，智能网卡处理网络与存储数据负载三层异构解耦架构，使其成为超大规模智算集群不可或缺的“数据调度中枢”，为超大规模智算集群高效稳定运行的筑牢硬件根基。

二、五大核心技术分析

智能网卡正通过高速无损通信、基础设施卸载/数据面加速、定制化架构、异构协同与生态适配等技术，系统性破解大模型训练与智算集群在传输时延、算力效率、调度协同和部署兼容等方面的瓶颈，全面支撑人工智能基础设施向高性能、高可靠、高融合方向演进。

（一）高速率：无损通信，破解智算网络传输瓶颈

作为智算场景的核心传输支撑，高速无损通信是智能网卡赋能大模型高效运行的关键抓手，更是破解万卡级智算集群、万亿参数大模

型分布式训练通信困境的核心技术。智能网卡可有效缓解智算网络面临的时延高、丢包高、吞吐低等突出问题，为 AI 训练推理、跨域算力协同构建稳定、高效、低时延的传输场景。

在协议层面，相较于传统网卡的通用化协议设计，智能网卡采用“专用协议+硬件卸载”的创新架构，从根源上提升通信效率。一方面，原生适配 RoCEv2、InfiniBand 等高性能智算专用通信协议，精准匹配智算场景的高带宽、低时延需求，摒弃通用协议的冗余开销；另一方面，将 TCP/IP 协议栈、VXLAN 叠加网络封装/解封装、路由转发等核心任务，从主机 CPU 彻底卸载至网卡硬件层面处理，既能大幅降低主机 CPU 的资源占用，又显著缩短网络传输时延，实现“轻量传输、高效协同”。此外，智能网卡集成多重智能管控机制，为通信稳定性保驾护航。通过 PFC（优先级流控制）、ECN（显式拥塞通知）等技术，实时感知网络拥塞状态；结合动态流量调度、多路径选路功能，可根据网络负载动态调整数据传输路径与速率，从源头避免数据包丢失与重传，确保大模型训练过程中高吞吐、高可靠业务的连续稳定运行，杜绝因通信中断影响训练效率。

在硬件层面，智能网卡在硬件接口与互联适配上，全面对标智算场景的大带宽、高速率、低时延需求，构建全链路高速传输体系。在 PCIe 接口方面，普遍采用 PCIe Gen4/Gen5 及更高规格，为网卡与主机之间的数据交互提供充足带宽支撑。在网络端口配置上，支持 100G、200G、400G 甚至 800G 高速以太网互联，匹配智算集群东西向流量与上行互联的大带宽需求，实现多节点、多任务的并行高速通

信。在光模块适配层面，全面兼容 400G/800G 高速光模块，与智算网络高速互联体系深度协同，打通高速传输链路，从硬件底层破解大模型传输时延瓶颈，为智算场景的高效运转提供坚实的硬件保障。

（二）高性能：设施卸载，释放 CPU/GPU 有效算力

基础设施卸载是智能网卡与传统网卡形成本质区别的核心标志性技术。不同于传统网卡仅承担基础数据传输功能，智能网卡聚焦智算场景专属算力痛点，通过硬件级原生加速设计，围绕计算、存储构建全方位加速能力，专门卸载 GPU、CPU 的非核心负载，最大限度释放核心计算芯片的算力潜力，有效缓解智算场景算力供给与需求矛盾。

在计算层面，智能网卡可内置独立专用 AI 加速引擎，采用 ASIC/FPGA 等定制化硬件架构，摒弃通用架构的冗余设计，针对智算场景的核心计算负载进行架构级深度优化。可直接承接并完成部分轻量级推理与数据预处理等非核心 AI 任务，无需占用 GPU、CPU 的核心算力资源，让核心计算芯片能够专注于大模型训练、复杂推理等核心关键任务，最大化提升整体算力有效产出。

在存储层面，依托 AI 专用加速引擎的硬件能力，智能网卡支持 NVMe over Fabrics（NVMe-oF）硬件加速技术，打破传统存储访问的协议栈壁垒，降低主机协议栈处理开销，由智能网卡独立处理，无需占用主机核心算力。实现了存储与网络的高性能融合，有效打破存储与计算之间的传输瓶颈，进一步提升智算场景的数据读写效率，精准适配大模型训练推理中高频数据访问的核心需求。

（三）高适配：架构兼容，夯实算力卸载技术底座

定制化架构和可编程能力是智能网卡实现算力卸载、提升智算效率的核心技术支撑。通过面向智算场景的深度定制化设计与灵活适配能力，将 CPU/GPU 的非核心负载全面剥离并交由专用硬件独立处理，最终构建“通用算力减负、专用算力增效”的智算硬件体系，为智算集群高效运转筑牢架构根基。

在架构设计层面，智能网卡采用面向智算场景的专用定制化硬件架构，精准对标智算集群核心运行需求，定制开发多元专用硬件单元。全面覆盖网络协议处理、虚拟化交换、存储加速、安全加解密、流量管控、RDMA 加速等核心功能模块，实现各类非核心负载的线速级高效处理，从根源上杜绝非核心任务占用 CPU/GPU 核心算力，确保核心计算资源集中聚焦于大模型训练、复杂 AI 推理等关键场景，最大化发挥核心算力价值。同时，采用算力分层解耦设计，将网络传输、存储交互等非核心负载，与 AI 核心计算负载进行彻底分离，实现独立并行处理。这种设计不仅有效减少不同负载间的相互干扰，更从架构层面提升了智算集群的运行效率与稳定性，为大规模分布式训练提供坚实架构支撑。

在可编程能力层面，智能网卡基于 FPGA 或 ASIC 的可编程流水线，允许用户或开发者通过 P4 等高级语言定义数据包的处理逻辑，自主定义数据包的处理逻辑，无需改动硬件本身，即可实现网络功能的快速迭代与创新部署，大幅降低功能升级的时间成本与硬件成本，

适配智算场景技术快速迭代的需求。部分智能网卡支持在运行时加载不同的加速功能微码，以适应不断变化的工作负载需求。

（四）高协同：异构调度，打通集群智能调度壁垒

当前，GPU、NPU、CPU、TPU 等多元异构芯片共存成为常态，而异构算力之间的互联与协同调度成为制约智算集群效率的关键瓶颈。智能网卡作为异构算力协同的核心枢纽，通过异构互联与资源管理，有效打破调度壁垒，提升智算集群整体效能。

在异构互联层面，智能网卡集成多个通用 CPU 核心，用于运行控制平面软件、管理代理及虚拟交换机，保障系统稳定运行。同时，内置硬件级协议转换、地址映射与资源虚拟化模块，通过标准接口和驱动适配实现与多类计算芯片协同，实现跨芯片、跨节点、跨域的高效数据流转与通信。从根本上打破了异构算力之间因芯片架构差异导致的通信壁垒，为不同算力单元提供统一、高效的数据交换通道，确保海量数据在异构集群间的无损、低延迟传输。

在资源管理层面，智能网卡支持网络队列、VF、带宽、存储路径和安全策略隔离等高级特性，将全域异构算力资源进行统一纳管、动态分配。能够根据不同 AI 任务的算力需求，灵活调度最优算力资源，实现算力按需分配、高效利用。这不仅避免算力资源的闲置浪费，更能确保各类任务获得精准的算力支撑，从全局层面最大化智算集群的资源利用率与整体运行效率。

（五）高融合：生态共建，降低行业落地部署流程

智能网卡的规模化落地，离不开与智算全产业链生态的深度协同。通过打通硬件、软件、平台全链路适配壁垒，实现与智算产业链各环节的深度融合，构建统一、开放、高效的智算生态，为智能网卡规模化应用奠定坚实基础。

在硬件适配层面，智能网卡立足全产业链兼容需求，全面适配当前智算场景主流硬件体系，有效破解硬件兼容难题。在 AI 芯片方面，全面兼容 NVIDIA、AMD、华为昇腾等主流 AI 芯片，无需额外进行硬件改造即可实现无缝对接；在架构适配方面，同步支持 x86、ARM、RISC-V 等多元架构，覆盖不同厂商、不同场景的硬件需求，大幅降低硬件兼容性调试成本，避免因硬件不兼容导致的部署延误，加快智算场景落地进度。

在软件与平台适配层面，围绕智算场景软件与平台协同需求，智能网卡深度融入全软件生态，实现与各类主流工具、平台的无缝对接。在 AI 框架适配方面，全面兼容 TensorFlow、PyTorch 等主流 AI 训练与推理框架，确保网卡加速能力与 AI 任务深度协同，提升模型训练推理效率；在云原生与调度平台适配方面，支持与 Kubernetes、OpenStack 等平台集成，支持与智算调度系统无缝对接，实现网卡管理、算力卸载、流量调度等功能的统一管控，打破生态碎片化壁垒，构建统一、开放的智算软件生态。

在安全防护层面，智能网卡集成专用硬件加密模块，依托硬件密码引擎和安全策略执行单元，可高效完成在线加密/解密、IPsec、

TLS/SSL 卸载等各类安全任务。这种硬件级加密设计，无需占用主机 CPU、GPU 的计算资源，在全面保障智算场景数据传输、存储安全的同时，不额外增加主机计算负担，真正实现智算场景高安全、高可靠的核心需求。

三、五大典型应用场景分析

（一）高互联：智算集群高速互联

当前，以大模型、多模态为代表的人工智能技术加速演进，模型参数规模已突破万亿级别，训练集群规模持续向万卡乃至十万卡量级扩展。智算中心作为新型信息基础设施的核心载体，其性能瓶颈已从单一计算单元的算力供给，转向集群内部海量加速器之间的高效协同与数据交换能力。传统基于 TCP/IP 协议栈的网络架构，在面对高并发、高吞吐的分布式训练通信需求时，普遍存在端到端时延高、丢包率高、带宽利用率低等问题，导致通信开销在整体训练周期中占比显著上升，制约了集群的线性扩展效率与训练任务稳定性，难以适配超大规模 AI 模型训练需求。智能网卡重点解决智算集群内海量加速器之间的数据传输效率瓶颈，聚焦分布式训练中梯度同步、参数广播等关键集合通信操作的高效落地，破解传统网络架构的通信梗阻问题，实现微秒级端到端时延、数百 Gb/s 稳定带宽，支撑万卡级集群实现线性扩展，保障长时间、高复杂度训练任务连续无故障运行。

智能网卡通过原生支持 RDMA、RoCEv2 等无损网络技术，构建端到端硬件加速的数据传输通路。其内置的拥塞控制、流量调度与错误恢复机制，在微秒级时延下实现数百 Gb/s 的稳定带宽，确保梯度

同步、参数广播等关键集合通信操作在超大规模集群中高效完成。该能力显著提升了集群的通信效率与扩展比，保障了长时间、高复杂度训练任务的连续性和可靠性。智能网卡在超大规模智算集群中的规模化部署，是提升我国人工智能基础设施全球竞争力的关键举措，也是实现算力资源高效聚合与调度的基础前提。此外，随着 CPO（光电共封装技术）的兴起，智能网卡正与光引擎深度融合，进一步降低功耗与信号损耗，成为未来智算中心互联的主流方案。

（二）深卸载：基础设施效能优化

随着云计算、大数据与人工智能的深度融合，算力中心承载的任务日益多元化。除核心业务计算外，需处理大量由虚拟化、网络、存储及安全等功能衍生的基础设施负载，如虚拟交换（OVS）、网络隧道封装（VXLAN/GRE）、存储协议解析（NVMe-oF）以及 TLS/IPsec 加解密等。任务传统上由 CPU 执行，导致通用计算资源被大量挤占，形成显著的“基础设施资源开销”，降低了服务器的有效算力密度，增加单位算力的能耗与运营成本，成为制约算力中心绿色低碳发展的瓶颈。

智能网卡作为基础设施任务优化的硬件加速单元，通过集成可编程数据处理引擎，将重复性高、规则性强的任务从主机 CPU 中高效卸载，使 CPU 能够专注于高价值业务逻辑计算，有效提升服务器资源利用率与能效水平。在规模化云环境中，单台服务器通过智能网卡卸载后，可释放 30%以上的 CPU 资源，有效支撑更高密度的虚拟机或容器部署，在不新增物理设备的前提下，提升算力中心的服务能力

与经济效益。智能网卡卸载能力不仅适用于大型云服务商和智算中心，也延伸至边缘计算节点等资源受限场景，通过 SR-IOV 和硬件级流量整形实现多业务隔离与确定性传输，是推动算力基础设施集约化、绿色化发展的关键技术路径。

（三）快训推：AI 训推全链加速

人工智能工作负载对底层网络提出高性能要求。在模型训练阶段，分布式训练需在每个迭代步中完成前向传播、反向传播及梯度同步等操作，其中梯度同步涉及全集群范围内的高并发数据交换，对网络带宽、时延与确定性提出严苛要求。网络抖动或丢包均可能导致训练收敛变慢。在模型推理阶段，尤其是面向智能客服、实时推荐、自动驾驶感知等在线服务的高并发场景，用户对响应延迟和吞吐量的要求敏感，毫秒级的网络延迟将影响服务质量与商业转化效率。因此，网络性能已成为衡量 AI 系统端到端效能的核心指标。智能网卡重点解决 AI 训练与推理过程中的网络性能瓶颈，优化训练侧梯度同步通信效率与推理侧网络 I/O 处理能力，实现 AI 训推全链路加速，保障训练任务高效推进与推理服务高质量交付。

智能网卡通过深度优化的数据面处理能力，为 AI 训练与推理提供端到端加速。在训练侧，其硬件引擎可对主流 AI 框架（如 PyTorch、TensorFlow）的通信模式进行针对性优化，结合自适应拥塞控制算法，确保在极端负载下维持网络无损特性，提升训练吞吐（samples/sec）与稳定性。**在推理侧，**智能网卡通过优化推理服务的网络 I/O 路径，减少数据拷贝和 CPU 协议处理开销，降低尾部延迟，同时支持动态

批处理与模型并行的高效调度，为提升智算中心服务效能、缩短模型迭代周期、保障线上服务质量提供核心支撑。

（四）严隔离：多租户安全隔离

当前，智算中心正加速向多租户共享服务平台演进，为金融、政务、医疗、能源等高敏感行业客户提供 AI 即服务（AIaaS）能力。多租户云平台共享场景要求不同租户的训练数据、模型资产及计算资源在共享基础设施上实现严格隔离，防止数据泄露、模型窃取与资源争抢。然而，传统的基于主机的防火墙、虚拟交换机 ACL 等软件定义安全策略存在两大短板：**一是**安全策略执行依赖主机 CPU，易受恶意租户攻击或绕过；**二是**缺乏硬件级可信根，难以满足《网络安全法》《数据安全法》及等级保护等合规要求，存在重大安全风险。智能网卡重点解决多租户共享场景下的安全隔离难题，构建独立于主机 CPU 的硬件执行环境，形成可信安全边界，实现租户间网络流量、存储访问与计算资源的硬件级强隔离，筑牢多租户安全防护底线。

智能网卡通过构建独立于主机 CPU 的硬件执行环境，形成可信安全边界。安全策略如微隔离、加密传输、访问控制、完整性度量等，均可在网卡侧的可信执行环境中直接编译、加载与执行，实现租户间网络流量、存储访问与计算资源硬件增强的强隔离。“硬件即安全”模式，有效降低跨租户数据泄露与侧信道攻击风险，并将安全策略执行开销控制在较低水平。同时，智能网卡与云管理平台深度协同，支持安全策略的集中下发、动态更新与审计追溯，为构建开放、共享、安

全、可信的 AI 服务平台奠定坚实基础，成为推动 AI 普惠与产业落地不可或缺的安全保障。

（五）智运维：智能网络全景运维

随着智算中心网络规模指数级扩张与业务复杂度的持续攀升，传统依赖人工巡检、被动响应的运维模式已难以适配高可用、高效率的运行需求，运维团队面临“规模激增与人力有限”的结构性矛盾，网络故障定位耗时久、隐患预警滞后等问题，成为制约算力资源稳定释放的关键瓶颈。智能网卡凭借硬件级感知与可编程能力，作为推动网络运维从“人工排查”向“智能自治”演进的技术底座，解决智算中心网络运维效率低、故障定位慢、隐患预警滞后等问题，构建智能网络运维体系。

智能网卡通过三大核心能力赋能智能运维体系落地。一是实时遥测能力，依托 INT（带内网络遥测）技术，在数据包装发路径中嵌入遥测信息，实现细粒度、实时网络状态感知，精准捕捉微突发流量、链路时延波动等关键指标，构建全链路可视化的网络运行视图。**二是**故障自愈能力，借助硬件级错误检测与上报机制，可快速定位网卡掉线、链路异常、慢节点等常见故障，并结合预设策略自动触发隔离、切换等修复动作，降低人工干预频次。**三是**数字孪生支撑能力，通过 P4 等可编程接口动态采集网络行为数据，为构建网络虚拟映射提供实时数据源，支持对网络拥塞、故障扩散等场景的仿真推演与趋势预测，实现运维从“事后补救”向“事前预防”的跨越。智能网卡的应用提升了智算中心的运维效能与网络稳定性，将平均故障定位时间从小时

级缩短至分钟级，降低智算中心运维成本。同时，其支撑的智能运维体系可适配超大规模集群、多租户云平台等多种复杂场景，为算力资源的持续稳定供给提供坚实保障。

四、智能网卡产业生态建设分析

（一）全球格局：巨头突出，壁垒较高

1、市场扩容提速，算力成为核心引擎

全球智能网卡市场正处于高速增长的关键阶段，算力需求已成为核心驱动力。2025 年全球智能网卡市场规模约 56 亿美元，预计 2026 到 2030 年复合年增长率约为 20%-35%，2030 年市场规模约达 139-251 亿美元，行业增长势头强劲，市场发展潜力较大。依托 AI 大模型训练与推理对高带宽、低时延、高吞吐的刚性需求牵引，智算中心已成为智能网卡的主要消费场景，其需求占比预计将超过 60%。智能网卡通过硬件卸载、数据加速等技术，有效缓解 CPU 算力瓶颈，已从单纯的网络连接部件，成为智算中心提升整体算效的关键组件。未来，随着 AI 应用场景的不断拓展和算力需求的持续爆发，智能网卡市场将迎来更广阔的增长空间。

智能网卡全球市场主要按照芯片实现方案，细分为 FPGA 可编程方案、ASIC 专用芯片方案以及集成化 SoC 架构方案。2025 年，基于 FPGA 的智能网卡占整体市场的 16%，基于 ASIC 的智能网卡约占整体市场的 74%，基于 SoC 集成构架的智能网卡约占整体市场的 10%，尤其在 AI 推理流程中应用广泛，有效满足基础机器学习、深度学习及大语言模型推理的算力需求。

表 2 智能网卡技术架构对比

对比维度	FPGA 可编程型智能网卡	ASIC 专用芯片型智能网卡	SoC 集成架构型智能网卡
核心优势	高可编程性：支持逻辑重构，适应自定义数据包处理管道。	高能效比：针对特定任务优化，性能与功耗表现优异。	通用计算能力：集成 ARM 或 RISC-V 内核，支持复杂工作负载。
部署规模	集群常部署 500 到 3000 个单元。	超大规模云集群部署量超 10000 台。	边缘计算节点集群部署 50 到 500 个单元。
关键性能	灵活适应实时数据流管理。	吞吐量达 200-400Gbps，CPU 负载大幅降低。	每秒处理超 200 万个数据包。
典型应用	网络安全：加速深度数据包检查与异常检测。	算力中心/虚拟化：大幅降低 CPU 负载，提升吞吐量。	AI 推理/边缘计算：处理元数据过滤、AI 推理任务。

来源：IDC，中国信息通信研究院

2、巨头全链掌控，垂直生态锁定用户

国际巨头构筑全产业链壁垒，市场集中格局显著。英伟达、英特尔、博通三大企业凭借先发优势，占据全球市场较高的份额。**在核心技术层面**，三个企业牢牢掌握了高端 ASIC 与 FPGA 芯片的设计主导权，并通过软硬件深度集成的垂直生态服务用户。以英伟达为例，其通过“硬件筑基+软件锁场”战略，将 CUDA 生态打造为行业事实标准。截至 2026 年，CUDA 架构已发展 20 周年，全球累计接入设备数亿，形成了“装机量－开发者－生态”的闭环，使得开发者对英伟达硬件产生深度依赖。**在产业层面，**三大巨头通过供应链强化其主导地位。英伟达在 2025 年底至 2026 年初，斥资 80 亿美元投资新思科技、

Lumentum 等上游关键环节企业，将 EDA 工具、高速光通信等核心资源纳入其版图，实现了从芯片设计到硬件供应的全链条把控。

3、协议规范设限，标准准入门槛抬升

技术标准与生态叠加，行业准入门槛持续抬升。超以太网联盟（UEC）、IEEE 802.3 等关键行业标准主要由国际头部企业引领，在接口速率、协议规范等核心领域发展重要作用。如 IEEE 802.3 标准持续更新迭代，衍生出涵盖 2.5G/5G 传输速率的 IEEE 802.3bz 等细分标准。UEC 联盟汇聚微软、Meta 等众多科技力量，聚焦于 AI 时代的超大规模网络技术探索。同时，部分企业推行私有协议，在一定程度上阻碍了异构硬件间的兼容互通，增加了不同厂商产品的适配成本，间接提高新进入者的技术与适配门槛，促使各国重视技术研发与生态体系建设，降低对外部技术和标准体系的依赖，推动智能网卡产业高质量发展。

（二）国内态势：政策护航，蓝海待拓

1、央地联动施策，筑牢产业发展根基

国家及地方层面正密集出台重磅政策，全方位推动智能网卡产业的规模化发展与国产化替代进程。《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》等国家级政策，不仅引导算力基础设施规模化建设，更为智能网卡技术创新指明方向。此外《网络安全法》《数据安全法》等法规强化硬件级安全功能集成要求，为市场规范化、自主化发展奠定基础。**地方层面，**北京、上海、深圳、

成都等重点城市同步响应，通过税收减免、设立专项产业基金等手段，为研发企业提供人才、资金等全方位支持，形成上下联动的政策合力。

2、技术加速突破，标准生态协同推进

在技术突破与标准协同双轮驱动下，国内智能网卡产业加速迈向高质量发展。在技术层面，在 PCIe 6.0 接口、CPO 等前沿技术领域，部分产品性能已接近国际同类水平，配套固件、驱动软件与国产服务器、操作系统的适配兼容性得到大幅改善，有效解决了长期存在的“适配难、不稳定”问题，为智能网卡生态发展提供了坚实的技术支撑。在标准化建设方面，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，中国信通院依托开放数据中心委员会（ODCC）牵头开展智能网卡技术报告编制以及测试体系构建，围绕智能网卡参考架构、CXL 内存扩展接口、网络卸载能力等重点方向，先后发布《运营商智能网卡需求场景报告》等一系列成果，为构建开放、协同、可互操作的智能网卡产业生态提供坚实支撑。

3、潜力加速释放，企业竞争力分层显现[4]

中国智能网卡市场整体呈现较为集中的态势，产品结构呈现明显的分层化、场景化特征。2025 年全球智能网卡市场规模约 56 亿美元，我国国内市场规模约占全球总量的 30%。从技术路线结构来看，国内智能网卡市场主要按照芯片实现方案细分为 FPGA 可编程方案、ASIC

[4] 备注：本章涉及到企业排名，均为按照拼音字母先后顺序排列，排名不分先后。

专用芯片方案以及集成化 SoC 架构方案。2025 年，基于 FPGA 的智能网卡占整体市场的 16%，基于 ASIC 的智能网卡约占整体市场的 74%，基于 SoC 集成构架的智能网卡约占整体市场的 10%。从传输速率维度来看，100G 以下（不含）约占国内智能网卡整体市场的 22%，100-400G（不含）约占国内智能网卡整体市场的 31%，400G（含）以上约占国内智能网卡整体市场的 47%。综合技术架构与速率结构，以 ASIC 架构为例，智能网卡速率 100G 以下（不含）国内智能网卡市场规模约 19 亿元，100-400G（不含）国内智能网卡市场规模约 27 亿元，400G（含）以上国内智能网卡市场规模约 42 亿元。

为规范智能网卡产品性能评价，建立统一、科学、可量化的智能网卡性能评价体系，本报告围绕企业估值、投融资、性能成熟度、功能成熟度、创新能力、市场规模、市场渗透、产品兼容性、行业生态参与度等指标开展综合评估，全景式剖析我国智能网卡企业产业格局，为智能网卡选型、测试与认证提供依据。从整体产业格局来看，企业立足自身资源禀赋和发展定位，在投融资水平、市场规模、技术创新以及产品生态等方向差异化布局，构筑起多层次、立体化的产业生态体系。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，共同推动我国智能网卡产业迈向高质量、可持续发展新阶段。

投融资方面，国内智能网卡企业凭借雄厚的资本规模与活跃的投融资表现，彰显出强劲的资本潜力与抗风险能力。综合分析，企业按发展阶段可清晰划分为领航者、远见者、筑基者三大类型。其中，领

航者企业兼具高资本规模与高投融资水平，拥有成熟的商业模式与广泛的市场认可度，主要企业为楠菲微电子、星云智联、翼华科技、云豹智能、云脉芯联、中科驭数等；远见者企业依托核心技术壁垒与广阔市场前景获得资本青睐，成长潜力突出，主要企业为沐创集成电路、网讯科技等；筑基者企业尚处于发展初期，重点聚焦技术打磨与市场验证，稳步积累发展动能，主要企业为灵可达、益思芯科技等；

在市场规模与渗透层面，企业按市场表现可划分为领航者、远见者、筑基者三大类型，各类企业依托差异化市场策略，在算力中心、云计算等核心场景构建竞争优势，推动智能网卡技术规模化落地。具体而言，领航者企业凭借领先的市场规模与广泛的市场渗透，在行业发展中占据主导地位，引领技术迭代与产业升级，如楠菲微电子、星云智联、翼华科技、云豹智能、云脉芯联、中科驭数等；远见者企业依托规模化效应，在特定区域或细分场景形成核心竞争壁垒，如沐创集成电路、网讯科技、益思芯科技等企业；筑基者企业则以差异化路线切入市场，逐步拓展市场份额，完善产业布局，如灵可达等企业。

技术创新是智能网卡产业发展的核心驱动力，企业可划分为三大梯队。其中，领航者企业在产品性能与功能成熟度上均处于行业领先水平，创新能力尤为突出，如星云智联、翼华科技、益思芯科技、云豹智能、云脉芯联、中科驭数等企业；远见者企业在性能或功能领域实现关键突破，正快速向高端化迈进，如沐创集成电路、网讯科技等企业；筑基者企业聚焦特定技术方向，稳步推进产品迭代与创新升级，如灵可达、楠菲微电子等企业。当前，部分企业重点聚焦芯片研发和

算网融合等核心方向，加快核心技术攻关，构建标准化技术平台，打造标杆应用场景，引领行业技术迭代与模式创新。如翼华科技、云豹智能、云脉芯联、中科驭数等企业，在高性能芯片、可编程数据平面、软硬协同架构等领域取得实质性进展，成为产业生态构建的重要力量。

产品兼容性与行业生态参与度是衡量智能网卡企业核心竞争力的重要维度。领航者企业兼具高产品兼容性与高生态参与度，在产业生态建设中发挥引领作用，推动形成开放协同的标准体系，如沐创集成电路、星云智联、翼华科技、云豹智能、云脉芯联、中科驭数等企业；远见者企业深度参与行业生态建设，通过标准制定、产业合作强化生态体系，如网讯科技等企业；筑基者企业聚焦特定平台或场景，实现深度适配与优化，如灵可达、楠菲微电子、益思芯科技等企业。

五、智能网卡发展趋势与建议

当前，智能网卡作为算力基础设施关键支撑，正处于技术迭代与规模应用的关键阶段，需加快突破 PCIe 7.0、CPO、可编程架构等核心技术，提升性能与能效；强化芯片、整机、云平台协同，构建全链条产业生态；推进开放标准体系建设，打破私有壁垒；深化智算中心、边缘计算、电信等场景应用，推动从“可用”向“必用”转变，推动产业高质量发展，夯实国家新型算力底座。

（一）技术迭代提速，突破核心性能与能效瓶颈

当前，智能网卡产业正处于技术代际跃升的关键窗口期，亟须强化前沿技术攻关，夯实产业升级基础。

一是紧扣核心方向，提升处理效能。聚焦接口速率、NPU 算力与

可编程能力，加快推进 PCIe 6.0/7.0 标准的研发适配与产业化应用，为数据高速搬运提供底层支撑。大力发展支持运行时动态编译的可编程架构，增强对新协议、新业务的快速响应能力，有效降低硬件迭代成本。**二是深化光电融合技术迭代，夯实高速互联底层支撑。**充分发挥光纤链路低损耗、大带宽、低比特能耗的综合优势，按照模块化封装、共封装集成、片上光互联梯次演进路径，推动互联形态由板间互联向芯片级、裸片级集成互联纵深拓展。抢抓 CPO 共封装光学产业化机遇，加快技术规模化推广应用，持续压降链路传输能耗，破解算力网络长距互联、绿色低碳发展瓶颈。**三是深化产业协同，构建绿色生态。**鼓励龙头企业联合上下游加快产品的工程化验证与规模部署，在巩固我国高端光模块全球优势的同时，通过技术创新与标准引领，构建绿色高效的下一代智能网卡产业生态。

（二）产业生态优化，构建全链条协同发展格局

以智能网卡为算力网络关键枢纽，统筹芯片设计、硬件研发、系统适配、场景应用及服务支撑等全产业链资源，破除上下游协同壁垒，构建产业生态体系。**一是强化产业链深度协同。**支持芯片企业、硬件厂商、服务器制造商、云平台服务商及行业用户开展深度合作，打通“核心器件、整机集成、系统软件、行业解决方案”的一体化发展链条。重点推动芯片、网卡、交换机与智算集群的协同设计，优化软硬件适配流程，提升端网协同效率，切实降低重复研发与适配成本。**二是培育多层次产业主体。**充分发挥龙头企业的技术与市场牵引作用，带动中小配套企业在专用芯片、固件开发、驱动软件、安全加速等细分领

域走“专精特新”发展之路，形成大中小企业融通创新的良好格局。**三是搭建高效对接平台。**建立产业供需合作机制，促进技术成果高效转化与资源要素优化配置，推动产业从单点技术突破向整体生态跃升转变，全面提升智能网卡产业自主可控水平与国际竞争力，为算力基础设施高质量发展筑牢产业根基。推动智能网卡产业上下游企业供需对接，鼓励网卡芯片企业、设备整机企业、行业应用企业以及高校科研机构组建产学研用联盟，通过项目共建、技术转移、人才交流、标准共建等多元化活动，促进智能网卡技术创新与成果转化，强化产业链上下游的紧密联系与协同发展，提升产业整体竞争力。

（三）标准体系完善，推动开放兼容与异构解耦

针对当前智能网卡领域协议标准不统一、烟囱式部署、软硬件强耦合等问题，要加快构建统一、开放、可互操作的标准体系。**一是构建全方位技术标准体系。**加快研究覆盖智能网卡硬件加速、网络虚拟化、存储卸载、安全加密及无损传输等关键领域的标准规范，注重与国内在研标准的协调衔接，同时充分考量与国际标准的互补兼容，避免技术孤岛。联合国内智能网卡、交换机、智算集群等关键环节龙头企业，推动形成覆盖全面的智能网卡端网协同统一标准。**二是共建开放技术生态。**深化产学研协作，探索驱动开发、协议适配、算力调度等关键环节的开源路径。鼓励骨干企业牵头参与 IETF、IEEE、OIF、ODCC 等工作，同步推进第三方测试认证体系建设，建立覆盖功能、性能、安全、能效的评测机制，切实降低用户部署成本与技术风险。

（四）应用场景深耕，实现全域规模化渗透落地

以场景创新驱动产业规模化发展为核心目标，坚持需求牵引，推动产品从“可用”向“好用”“必用”跨越。聚焦智算中心、边缘计算、工业互联网、电信网络等关键领域，系统挖掘并推广智能网卡赋能新质生产力的典型应用案例。重点打造一批具有示范效应的标杆项目，展示其在提升算力传输效率、降低能耗方面的显著成效。强化以点带面，实现深层次落地。建立绿色算力与网络效能评价机制，将智能网卡的能效优势转化为通用标准，推动智能网卡技术在更广范围、更高层次实现规模化渗透。