



2025 | AI2AC

人工智能与先进计算融合创新学术会议

SYMPOSIUM ON INTEGRATED INNOVATIONS IN ARTIFICIAL INTELLIGENCE AND ADVANCED COMPUTING



人工智能与先进计算融合 创新关键技术与基础支撑体系研究 蓝皮书



蓝皮书电子版

复旦大学 主编

人工智能与先进计算融合 创新关键技术与基础支撑体系研究

主编单位

复旦大学

共同编写单位

清华大学
北京大学
浙江大学
天津大学
中国科学技术大学
国防科技大学
华中科技大学
信息工程大学
国家数字交换系统工程技术研究中心 (NDSC)
嵩山实验室
中国科学院半导体研究所
中国科学院计算技术研究所
中国科学院微电子研究所
中国电科智能科技研究院
中国电科三十二研究所
中国电科五十八研究所
中国电子信息产业集团有限公司
江苏长电科技股份有限公司
华进半导体封装先导技术研发中心
飞腾信息技术有限公司
长电集成电路 (绍兴) 有限公司
先进计算与关键软件 (信创) 海河实验室

版权声明

本蓝皮书版权属于编写发布单位，并受法律保护。转载、摘编或通过其他方式使用本蓝皮书文字或者观点应注明“来源：人工智能与先进计算融合创新学术会议暨中国工程院工程科技学术研讨会“生成式变结构计算与高效能先进计算新路径研究”，人工智能与先进计算融合创新关键技术与基础支撑体系研究蓝皮书，2025年8月”。违反上述声明者，版权方将追究其相关法律责任。

编委会主任

邬江兴

编委会（按姓氏笔划排序）

马凯学	叶 乐	石宣化	刘勤让	李 涛	汪志强
吴华强	张 帆	金 海	明雪飞	窦 强	梁新夫
韩银和	薛向阳	邹 宏			

编写人员（按姓氏笔划排序）

王钧彝	王 颖	王 刚	冯 枫	冯 丹	朱丹江
刘 强	刘丰满	许志伟	祁晓峰	余 杰	余新胜
吴庆波	张庆东	陈 艇	陈 曦	范 旺	尚玉婷
胡 杨	胡宇翔	胡小燕	贺 明	赵豪兵	查雨立
夏于桐	郭国平	郭 威	高彦钊	曹 伟	梁新夫
彭 聪	廖小飞	潘 纲			

人工智能与先进计算融合创新关键技术 与基础支撑体系研究

当前，人工智能与先进计算的融合创新正迎来关键历史机遇——国家深入推进“人工智能+”行动、加速人工智能规模化商业化应用，巨量场景对算力提出“高智能、低时延、零信任”的极致需求。达成这一需求，亟须突破传统计算困局。晶上生成式变结构计算为人工智能与先进计算 (AI2AC) 融合创新提供了全新路径。在关键技术层面，晶上生成式变结构计算通过动态重构硬件资源与任务的映射关系，以“结构内生多样性”化解“应用需求多样性”矛盾；软件定义互连打破协议与拓扑的刚性约束，构建“协议可编程、带宽可定义”的柔性连接网络；晶圆级封装 (SDSoW) 突破单一制程限制，在 28nm 工艺下实现算力密度较传统方案提升 3 个数量级，为变结构计算提供高密度物理基座。在支撑体系层面，超融合算力网络环境为大规模算力优化提供了关键支撑；AI 计算内生安全通过化解大模型对抗攻击、数据投毒等未知风险，实现安全与效能的协同演进；算能一体化通过新能源波动性与算力负载的精准匹配，大幅提升绿电利用率，为算力基础设施注入可持续动能。人工智能与先进计算的深度融合，将推动计算范式从“工具赋能”向“体系重构”跃迁，最终实现“智能驱动决策、安全融入基因、能效突破极限”的终极目标，为全球数字经济打造高可靠、高可信、高效能的智能底座。



目 录

第一章 生成式变结构计算	1
1.1 生成式变结构计算的架构设计	1
1.2 领域专用基础算核库生成	3
1.3 全栈式软件工具链设计	5
第二章 软件定义互连	8
2.1 软件定义互连 (SDI) 概述	8
2.2 SDI改变刚性计算结构瓶颈	11
2.3 生成式复杂网络的构建方法	12
2.4 SDI与变结构计算的协同演进	14
第三章 晶圆级封装与物理实现	16
3.1 先进封装进展	16
3.2 生成式结构计算的工程物理特征	19
第四章 超融合算力网络环境	22
4.1 算力网络发展面临的挑战	22
4.2 超融合算力网络环境概述	25
4.3 超融合算力网络环境的关键技术	28
4.4 发展愿景展望	31
第五章 AI计算应用系统内生安全	34
5.1 内生安全:先进计算范式的“免疫基因”	34
5.2 AI驱动:内生安全架构的“智能进化引擎”	36
5.3 安全筑基:AI可信应用的“保障体系”	39
第六章 智能调度高效柔性负荷微网能源系统	44
6.1 AI算力激增与新能源波动性挑战	44
6.2 EFLM解决算力与多能电源时空耦合问题	46
6.3 AI驱动构建EFLM“动态优化大脑”	49
6.4 能算结合实现AI可持续发展	50
参考文献	52

第一章

生成式变结构计算

近年来，人工智能推动领域专用计算体系结构迅速发展。受应用任务变化、制程工艺进步及设计思想演进等因素推动，计算系统设计不断发展，但现有体系架构如通用计算、专用集成电路、异构计算和可重构计算等均存在一定缺陷。为解决这些问题，通过动态重构硬件资源与计算任务的映射关系，以“结构内生多样性”应对“应用变化多样性”，推动计算系统从“固定架构适配任务”向“动态架构响应需求”转型，实现硬件资源与计算需求的实时适配，进而支撑大规模智能计算，实现智能、高效、绿色和经济的生成式变结构计算系统。

1.1 生成式变结构计算的架构设计

1.1.1 理论模型

计算架构的固定性无法满足不同场景下的多样化需求，导致在性能、效能和灵活性上无法同时达到最佳。如图1所示，优化一个或两个方面通常难以兼顾第三个方面，而设计一个全面优秀的计算架构极其复杂，本文称之为“EPF困境”。计算系统无论是采用虚拟化方案还是众核方案，都无法解决多样性的算力需求。要打破EPF困境，必须超越现有架构，寻找新的计算范式。这是推动计算技术持续发展的关键。

任何单一的计算体系架构在EPF三维都无法在应用场景多样化条件下实现优化能力。在问题的原有维度上无法求解时，生成式变结构计算通过增加时间维度进行求解，即在EPFT四维上实现先

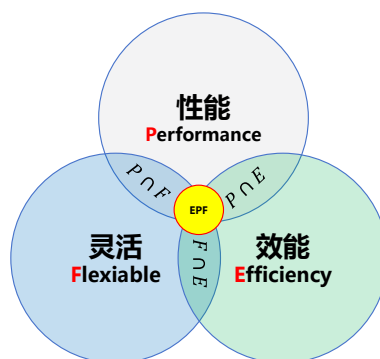


图1 EPF困境：单一刚性计算架构的“不可能三角”

进计算系统的优化。

对软硬件资源的元素化抽象处理是构建高效、灵活计算体系结构的关键。通过动态定义方式，如编排、调度等，在EPFT四维创建多样化不完全交集。这些交集在计算体系结构全生命周期中，通过动态调整和优化，形成EPF完备交集。计算体系结构的创建和演进过程，实质是元素化资源的可编程构建或释放回收过程，涉及资源动态分配管理及计算任务的高效响应处理。

在EPFT四维的“微分”时段，若能一体化展现多种任务的最优计算体系结构，那么在“积分”区间，我们就能等效地获得EPF完备交集。系统在极短时间间隔上迅速适应，优化计算资源分配，满足不同任务需求。生成式变结构计算需支持多种体系结构共存及多样化应用场景的演进，要求计算体系结构具有高度灵活性和可扩展性。这样，我们能构建出高效且可持续的计算生态系统，为未来信息技术应用打下坚实基础。

生成式变结构计算的关键是软硬件资源向确定性任务映射的共享表达方式。如图2所示，在EPFT时空上实现EPF完备交集时，不同的计算体系结构应具有不同的软硬件资源映射方式，将计算资源元素化表达

为如公式 (1) 所示。

$$R_{Arch_N}^t = \begin{bmatrix} A_{1N}^t & B_{1N}^t & C_{1N}^t \\ A_{2N}^t & B_{2N}^t & C_{2N}^t \\ \cdots & \cdots & \cdots \\ A_{sN}^t & B_{sN}^t & C_{sN}^t \end{bmatrix} \quad (1)$$

其中， t 为时间， N 为计算体系结构序号， s 为在计算体系结构时的任务序号， A ， B ， C 为不同类型的计算资源。在EPFT四维，不同时段计算体系结构所需计算资源的类型、数量、形态会根据系统服务质量评估结果发生动态变化，计算资源的“聚合”方式也应随之动态改变，引起资源的重构。记 t 时刻重构方法为 L^t ，服务质量的评估方法为 S^t ，则从时刻 t_i 到时刻 t_j ，重构方法动态变化函数如公式 (2) 和公式 (3) 所示。

$$L^{t_j} = f(L^{t_i}, S^{t_j}) \quad (2)$$

$$L^{t_i} = \begin{bmatrix} l_{1A}^{t_i} & l_{1B}^{t_i} & l_{1C}^{t_i} \\ l_{2A}^{t_i} & l_{2B}^{t_i} & l_{2C}^{t_i} \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & l_{NC}^{t_i} \end{bmatrix} \quad (3)$$

生成式变结构计算整合了资源和互连结构，通过分类应用需求并映射至特定领域，提炼形成软件定义节

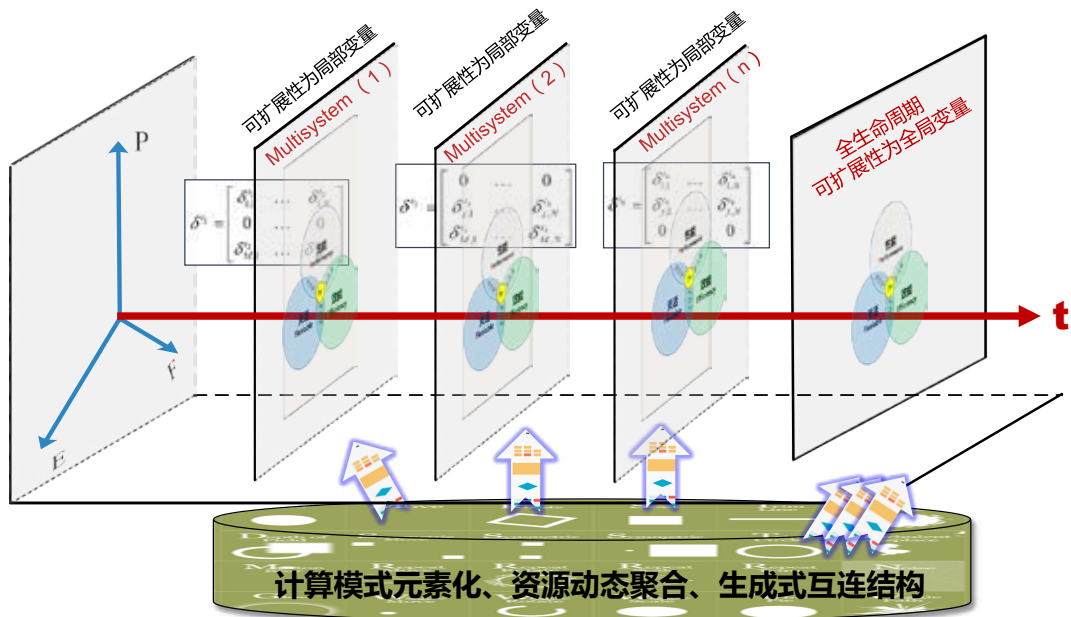


图2 生成式变结构计算的理论模型

点。这种互连结构为应用开发和部署提供框架支持，实现多节点协同作业，满足多样化计算和网络服务需求。资源动态聚合与分散，提高利用率，降低成本，增强系统性能和稳定性。

1.1.2 系统框架

构建生成式变结构计算系统遵循“计算模式元素化、资源动态聚合”的核心原则。该原则意味着将复杂的计算模式拆解为更微小、更易于管理的元素，此举不仅有助于简化复杂的计算模式，而且显著提升了系统的灵活性，使得这些元素能够更加便捷地进行组合与调度。进一步地，资源的动态聚合策略允许我们依据实际需求灵活地组合与分配资源，从而显著提升了资源的利用率，并且能够更好地适应工作负载的动态变化。通过采用生成式互连结构，能够满足各种计算任务的需求，实现资源的高效调度与优化，进而促进互连体制的共生，确保整个计算系统的高效、稳定运行。

生成式变结构计算设计根据应用需求动态生成和调整互连结构，具有动态性和可重构性，提供灵活性以适应多种计算任务。关键在于运用可重构和虚拟化技术，使系统能自我调整优化，确保高效完成计算任务。架构利用人工智能算法预测负载变化，提升响应速度和处理能力，为复杂计算环境提供稳定支持。如图3所示，生成式变结构计算的灵活性不仅体现在其对

资源池化聚合的支持上，还在于它能够通过智能算法和先进的管理技术，将计算资源、存储资源以及网络资源进行高效整合和池化管理。这种整合使得系统能够实现资源的动态分配和优化组合，从而达到系统性能的最优化。通过这种方式，资源池化聚合显著提高了资源的利用效率，同时为系统的可扩展性和可维护性提供了有力的保障。在面对不断变化的工作负载和业务需求时，计算系统能够快速响应，灵活调整资源分配策略，以适应新的挑战和需求。这种灵活性确保了系统能够持续提供高效的服务，同时也为未来技术的发展和升级预留了空间。

1.2 领域专用基础算核库生成

领域专用基础算核库是构建生成式变结构计算系统的基础。确定目标服务领域后，需分析并提取共性算子，通过融合这些算子，实现不同粒度的算核，建立领域算核库。

1.2.1 应用特征分析

应用特征分析涵盖算法特征分析、任务划分分析和算法共性特征分析。算法特征分析涉及领域算法逻辑，旨在获取算法的计算、存储和通信特征，以优化软硬件协同设计。核心在于算法复杂度分析，包括时



图3 生成式变结构计算系统框架

间复杂度和空间复杂度。时间复杂度分析评估算法运行时间，帮助决策算法设计和选择。空间复杂度分析量度算法运行时占用的存储空间，包括程序、输入数据等所占空间。通过复杂度分析，可从不同算法实现中选出性能较优方案。

任务划分分析依据任务的计算、控制、存储和I/O特征，将任务拆解并优化硬件设计，指导异构系统架构。它依赖于基于功能、性能和数据的实验分析数据，通过仿真实验评估算法在不同硬件上的性能，综合评估算法指标。异构硬件如CPU、GPU、FPGA的性能、功耗和面积是关键指标。任务划分分析支撑运行时系统任务调度，提升算法的可维护性和扩展性，优化调试，提升系统性能。基于功能的分析是常用方法，明确任务功能和目标，分解为子任务，优化功能模块，明确模块输入输出和实现方式，整合成算法流程，并进行测试优化。

分析领域应用算法的共性特征，以支持可重构系统架构的设计。该架构通过软硬件协同设计，实现模块化，增强系统的可扩展性和灵活性。核心在于统计学特征分析，评估算法的共性和特异性，为算核融合提供依据。主要步骤包括确定统计特征、数据收集、描述性与推论性统计分析、数据可视化、模型拟合及结果评估解释。

1.2.2 共性元素抽象

对领域应用的算法特征分析后，可以确定高效运算段的边界(对应高阶运算栈)，提取并构建具有共性元素的基础算子集合，进而设计算核。算核是构成高阶运算段的、完全由硬件实现的基本功能单元。算核集是充要一系列算核所构成的集合，可以支持完整的应用算法。

基础算子集集合的生成算法如下。

目标领域应用集为 $S = \{S_1, S_2, \dots, S_n\}$ 。

目标领域应用集中应用的 S_i 权重记为 α_i ， $\sum_{i=1}^n \alpha_i = 1$ 。

待求的基础算核集

$$CC = \{b_1 \cdot CC_1, b_2 \cdot CC_2, \dots, b_k \cdot CC_k\}。$$

系统可利用资源为 R ， CC_k 所占用资源为 r_i 。

总占用资源 $r = \sum_{i=1}^k b_i \times r_i$ 。

应用 S_i 重构所占用资源记为 r_{S_i} ，应用集 S 的重构总

代价为 $\sum_{i=1}^n r_{S_i}$ 。

设计约束条件： $(\sum_{i=1}^k b_i \times r_i + \sum_{i=1}^n r_{S_i}) \leq R$ 。

算核 $CC_{i,j}$ ， $1 \leq j \leq b$ ，被调用表示 $y_{i,j,l} = \begin{cases} 1, & \text{被调用} \\ 0, & \text{其他} \end{cases}。$

算核 $CC_{i,j}$ 的利用率为 $p_{i,j} = \sum_{l=1}^n y_{i,j,l} \times \alpha_l$ 。

一个合理的基础算核集，是在满足所选应用的重构需求的前提下，使各个算核在应用中被使用的概率基本一致，即算核利用率方差为，

$$\min \sum_{i,j} \left(p_{i,j} - \frac{\sum_{i,j} p_{i,j}}{\sum_{i=1}^k b_i} \right)^2$$

此时，对应的 $\{b_1 \cdot CC_1, b_2 \cdot CC_2, \dots, b_k \cdot CC_k\}$ 为所求解。

求解基础算核集的算法是一个迭代过程，从各应用所需算核集的简单并集出发，调整算核集合，使算核利用率的方差最小。

1.2.3 算核设计实现

在人工智能领域需要大量的并行计算资源，而基础算核是实现高效计算的关键。通过优化基础算核的算法和数据结构，可以提高人工智能系统的性能和效率。

基础算核设计是一个复杂而精细的过程，不同的算法具有不同的时间复杂度和空间复杂度，对于特定的计算任务，需要设计合适的算核以实现高效的计算。数据结构是基础算核设计中的一个重要方面。对于大规模矩阵运算，可以选择稀疏矩阵压缩存储技术来降低存储空间消耗。代码优化是基础算核设计的关键环节。通过优化代码，可以提高算核的执行速度和效率，减少内存消耗，提高程序的稳定性。代码优化包括算法优化、数据结构优化、代码布局优化等多个方面。例如，可以使用循环展开技术来减少循环次数，提高代码执行速度。并行计算在基础算核设计中是一个重要的研究方向。通过将计算任务分解为多个子任务，并分配给多个处理器同时执行，可以显著提高计算速度和效率。并行计算需要设计合理的并行算法和数据结构，以实现高效的并行计算。

基础算核的实现可以采用硬件设计和软件编程。通过设计集成电路来实现基础算核的功能，集成电路设计需要使用硬件描述语言如Verilog或VHDL等来定义算核的功能和行为，然后通过EDA工具进行仿真和验

证，生成可制造的集成电路版图。使用可编程逻辑器件来实现基础算核的功能，PLD器件可以通过配置不同的逻辑门来满足不同的计算需求，具有灵活性和可重构性。软件编程实现时，通过编写指令集来实现基础算核的功能。指令集是处理器体系结构中的基本单元，它规定了处理器可以执行的基本操作和指令。通过编写指令集，可以实现算核的基本功能。通过编译器将高级语言程序翻译成机器码来实现基础算核。编译器将源代码转换成目标代码，这个过程涉及到词法分析、语法分析、语义分析等多个步骤。使用编译器可以简化软件开发过程，提高开发效率。硬件设计和软件编程是相辅相成的，它们可以互相补充和优化。

1.3 全栈式软件工具链设计

生成式变结构计算依赖于一套软件工具链，以全面管理应用的全生命周期。工具链确保应用开发和部署各阶段的有效管理。在规划阶段，工具链协助需求分析和设计，确保应用与环境特性及业务需求匹配。在开发阶段，工具链优化代码，保证应用高效运行。在模拟仿真阶段，测试验证性能和可靠性，确保真实环境中的表现。在部署阶段，工具链支持应用快速稳定运行。

通过高效的编译、资源管理和动态优化，实现应用在变结构计算系统上的快速部署。工具链在软件定义计算方面，生成专用硬件逻辑；在编译部署优化方面，生成高效可执行代码；在运行时优化方面，提供应用主动感知、运行重构和协同执行能力。应用部署标准规范确保了开发和部署的质量与效率。这些规范包括应用特性、指标体系、测评方法和度量工具。应用特性定义了应用的核心属性，如可扩展性、灵活性和安全性。评估应用性能的指标体系，包括响应时间、吞吐量和资源利用率等关键指标。测评方法标准化了测试流程，确保应用性能和质量稳定。度量工具实时监控应用状态，收集性能数据，用于分析和优化，帮助开发者提升应用性能。

如图4所示，全栈式工具链框架由5部分组成：应用部署计算图生成引擎接受应用框架或领域专用描述

语言的应用场景处理流程描述作为输入，经过前端优化后，转换为基于基础算核库的应用部署计算图；任务划分引擎将算核计算图划分为静态和动态任务，生成静态任务计算图IR和动态任务IR；目标结构与代码自动生成引擎基于静态任务IR进行计算结构探索和代码生成；系统运行时引擎根据动态任务IR创建运行时资源管理和任务调度程序；模拟仿真平台贯穿编译流程，对目标代码和中间表示进行验证。

1.3.1 应用部署计算图生成引擎

应用场景用领域专用高级语言描述，涵盖数据类型如标量、向量、张量的存储与运算，以及数据流和控制流结构。该语言还包含性能要求，如响应时间、吞吐量和计算峰值，这些要求以值范围或概率分布表示，以适应性能需求的变化。高级语言描述的应用场景经解析后生成控制数据流图，分析其计算、存储和通信特征。统计应用算法中操作的类型、数量、依赖关系等，通过子图挖掘和算子融合提取基础算核，并与库单元匹配，形成部署计算图。性能要求标注在计算图节点上，指导任务划分和代码自动生成。

1.3.2 任务划分引擎

通过多层级划分方法简化应用部署计算图。首先，根据节点计算强度和连接关系，合并计算强度低且连接紧密的节点为粗粒度节点，以减少计算图规模和任务划分探索空间。其次，将粗粒度计算图细分为子计算图，作为资源分配的基础。然后，评估子计算图的资源需求，优化匹配高阶算核节点与系统资源，并分配相应资源（根据性能需求变化，资源分配可调整）。最后，根据子计算图的互连关系评估整体性能。采用演化算法设计任务划分，从初始方案开始迭代，直至优化目标达成，即最小化资源占用以满足性能需求。最终方案将提供静态任务计算图和动态控制信息，作为生成目标结构和代码的输入。

1.3.3 目标结构与代码自动生成引擎

构建基于静态任务计算图的多目标编译自动化探索空间，结合应用功能、性能和效能等因素，通过

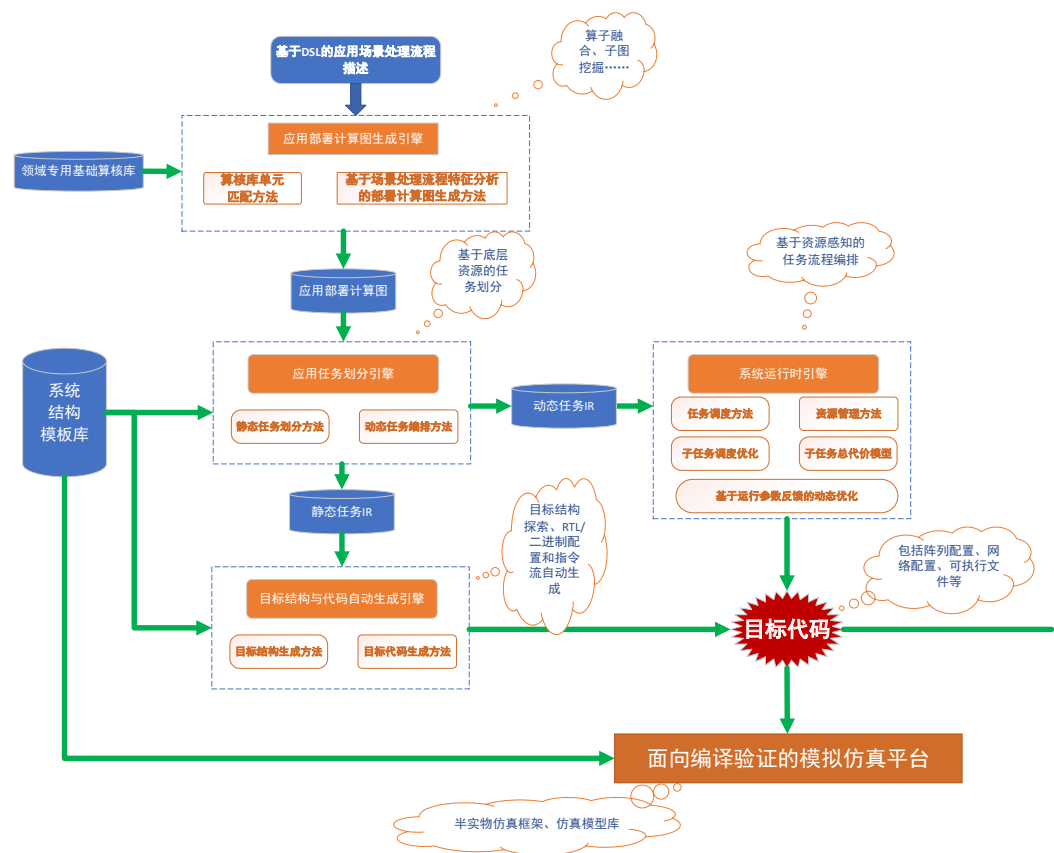


图4 全栈式工具链框架

多目标优化算法实现系统架构的帕累托最优解自动化求解，获取目标结构的参数集合。根据自动化探索结果，从模板库中选取单元组成目标结构模板，并在资源和时序约束下进行设计空间探索，以找到优化的目标结构并生成硬件RTL代码。设计空间探索通过优化结构设计和算核变换参数，旨在提升性能、功耗和面积。死代码消除依赖于目标结构的参数化设计，其设计空间大小受单元设计参数和基础算核变换优化参数定义和范围的影响。

目标结构优化涉及异构系统，包括标量、向量、空间计算单元，以及多层存储和互连单元。代码生成分为两部分：控制代码驱动标量单元生成控制信号和执行计算；配置代码设置单元执行、数据组织和传输模式。在数据流计算图中，复杂的高阶算核通过时间维展开，分阶段完成运算。每个阶段要求特定的执行

和数据传输模式。时域调度算法协调算核时序，生成控制代码，与标量计算合并，形成标量计算单元的指令流代码；不同阶段的配置形成配置代码。

1.3.4 系统运行时引擎

为高效管理任务和资源，需双向感知并决策，构建实时感知矩阵，分析管理任务资源，并进行预调度以满足静态需求，为动态调度提供基础方法。针对动态需求，采用最优化算法生成调度方案，结合动态任务信息和静态目标结构，实现资源与任务的协同优化管理，支持系统动态伸缩和资源重组。

1.3.5 面向编译验证的模拟仿真平台

该平台覆盖编译流程，仿真验证目标代码和中间表示。系统级模拟仿真平台由功能仿真模型和层次

化仿真库组成。功能仿真模型基于系统资源建模，包含解释引擎实现目标代码精确仿真。层次化仿真库基于领域特征和算法构建，提高仿真效率。模拟仿真模型采用多层中间表示对计算、存储和互连单元进行参数化建模。单元模型采用统一控制接口，描述数据和互连接口，提炼执行模式和参数空间。计算单元类型多样，参数包括规模、粒度等。存储单元类型包括寄存器文件、私有存储等，访问模式与算核特征相关。互连单元类型多样，参数包括连接类型、互连灵活度

等。层次化仿真库设计针对特定应用的快速功能仿真库单元，仿真阶段可替换目标代码段和资源功能模型，显著提升多节点仿真效率。

生成式变结构计算技术通过动态地重构硬件与任务之间的映射关系，成功地克服了传统计算架构在性能、效能以及灵活性方面所面临的诸多局限性。这种创新的方法为未来智能计算领域的发展提供了一种可持续的新范式，使得计算系统能够更加高效地适应不断变化的工作负载和应用需求。

第二章

软件定义互连

在当今数字化时代，人工智能与先进计算技术正以前所未有的速度发展，深刻改变着人类社会的方方面面。然而随着数据量的爆炸式增长、计算任务的日益复杂以及应用场景的多样化，传统的计算系统架构面临着诸多挑战。软件定义互连 (Software-Defined Interconnection, SDI) 作为一种新兴的技术理念，为解决这些问题提供新的思路和方法。本章将探讨软件定义互连如何支撑人工智能与先进计算，推动计算系统架构的变革。

2.1 软件定义互连 (SDI) 概述

相比于现有支持固定协议、固定带宽分配和固定交换结构的互连方式，软件定义互连在硬件层面实现了互连协议、端口类型、拓扑结构、带宽分配、互连模式等关键参数的软件定义，可以支持不同协议、不同带宽、不同互连模式的数据交互通信需求，因此，软件定义互连是一种面向网络互连领域的软件定义硬件技术，基于可重构和可编程技术，具有可扩展的硬件/软件架构，允许应用程序对硬件进行在线定义，以动态更改交换互连的功能、性能，实现对不同应用场景的最佳适配。

软件定义互连的实现结构如图5所示。通常包括软件定义协议控制器、软件定义转发器和软件定义交换结构三个部分组成，软件定义协议控制器可软件配置为支持不同频点、不同绑定模式、不同流控方式、不同协议类型的协议控制器；软件定义转发器可软件配置为不同协议的解析与封装、不同协议间的转换、不同协议的融合转发查表；软件定义交换结构可软件配置为不同规模、不同拓扑、不同交换模式、不同服务质量等级支持的交换结构。

2.1.1 互连端口可软件定义

互连端口是系统中不同设备与部件之间的硬件接口，为数据端设备提供传送数据的通路，形成适合数据传输需要的实体。如图6所示。

互连端口可软件定义包含了以下三方面的含义：

(1) 端口协议类型可软件定义：信息基础设施内部通常有多种通信协议，软件定义互连的每一个端口都可以根据应用需求，通过软件对其进行定义，可支持的软件定义内容包括：标准协议、行业专有协议和用户自定义协议。

(2) 端口工作模式可软件定义：基于端口类型，端口可软件定义为不同的类型；通道绑定模式可定义为1x、2x、4x、8x、16x等；通道和端口传输速率可软件定义为任意端口协议的典型频率；可对端口的速率及带宽利用度等进行定义。

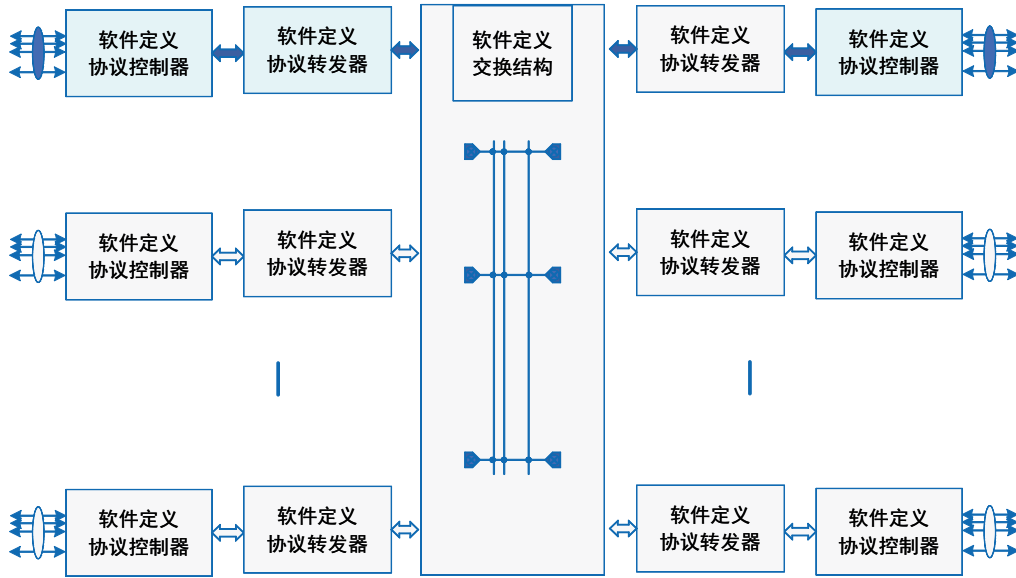


图5 软件定义互连实现结构

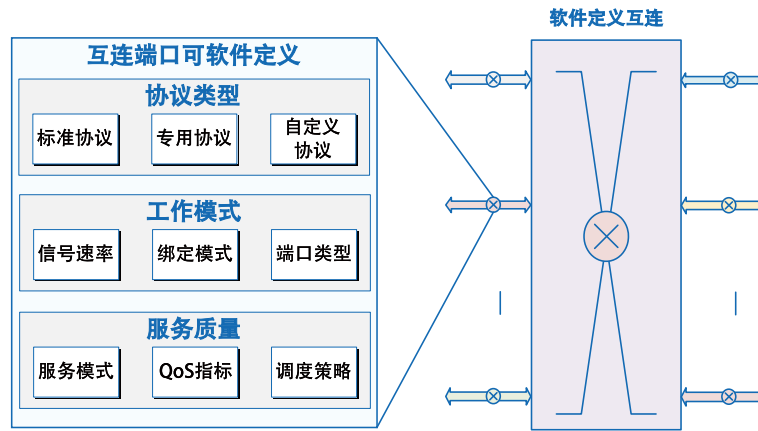


图6 互连端口可软件定义示意

(3) 服务质量策略可软件定义：QoS服务模型可软件定义；QoS的关键指标软件可定义，如带宽、时延、抖动、可靠性等；QoS调度策略可软件定义。

2.1.2 互连模式可软件定义

互连模式包含了交换方式、通信模式、拓扑结构三个方面，如图7所示。

互连模式可软件定义具体含义如下：

(1) 交换方式可软件定义：根据系统特点和组网需

求定义网络的交换方式为电路交换、报文交换、分组交换等；

(2) 通信模式可软件定义：可根据应用需求将网络节点之间通信方式定义为单播（一对一的通信模式）、组播（一对一组的通信模式）和广播（一对所有的通信模式）；

(3) 拓扑结构可软件定义：互连拓扑可依据通信业务模型软件定义为星型拓扑、树形拓扑、环形拓扑、基于Mesh的拓扑（分级互连拓扑、混合拓扑、不规则

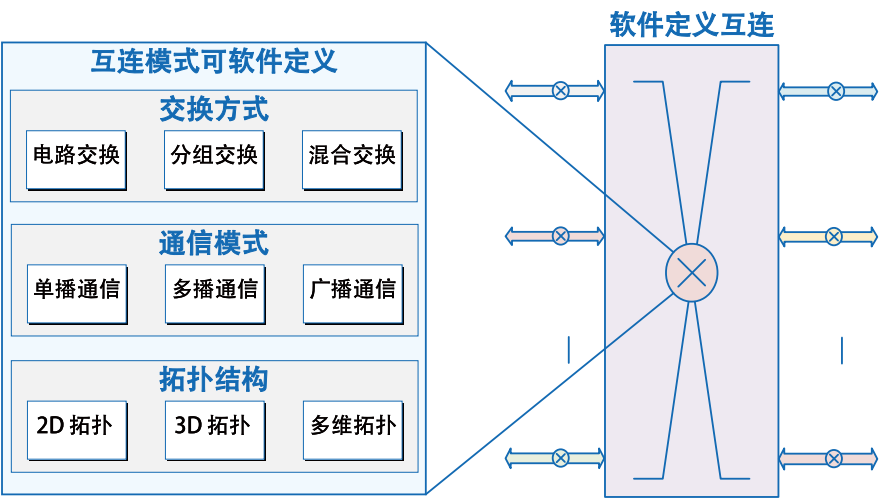


图7 互连模式可软件定义示意

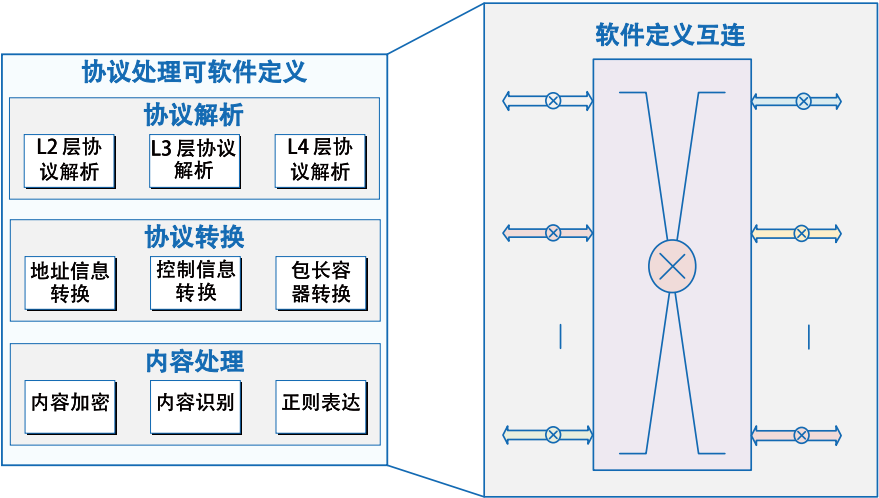


图8 协议处理可软件定义示意

拓扑等结构)。

2.1.3 协议处理可软件定义

软件定义互连的协议处理包含了协议包头解析、异构协议转换和协议内容处理三个方面的内容，如图8所示。

协议处理可软件定义的具体含义如下：

(1) 协议包头解析可软件定义：能够识别端口传输过来的不同通信协议类型，支持任意通信协议二层的

包头信息解析；可根据通信协议类型进一步对其他层的包头信息进行解析。

(2) 异构协议转换可软件定义：基于异构协议之间通信上层协议栈对等 (即协议数据载荷不发生变化) 的前提条件，支持系统内任意两种不同协议类型的端点设备通信。

(3) 内容处理可软件定义：可对转发载荷数据进行灵活的基于流表的处理操作，如端到端加密、安全监测、语法转换、基于正则表达式的协议识别、业务统

计等灵活的流式处理操作。

2.2 SDI改变刚性计算结构瓶颈

在人工智能、高性能计算与分布式系统深度融合的技术浪潮中，传统刚性计算架构因硬件互连模式固定、资源调度僵化，在动态负载适配、通信效率优化及系统扩展性方面的瓶颈日益凸显。面对动态负载的波动，其静态或预定义的服务与资源分配模式往往难以灵活响应：在流量激增时容易因资源不足而导致性能下降甚至服务中断；而在低峰期，又可能造成昂贵的硬件资源大量闲置，效率低下。同时，通信瓶颈问题突出，尤其是依赖于集中式通信枢纽（如单一数据库服务器或消息队列）的设计，极易成为系统性能的瓶颈点，随着请求量的增加，通信延迟会显著上升，并带来单点故障的巨大风险，严重制约系统的吞吐量和可靠性。此外，在扩展性方面，传统架构通常依赖于垂直扩展，即通过升级单台服务器的硬件（如增加CPU、内存）来提升性能。这种方式不仅成本高昂、存在物理上限，而且无法实现细粒度的资源按需分配，难以应对现代应用爆炸性增长的用户规模和业务需求。

因此，现代信息系统要实现自动化与智能化，

就必须搭建连接丰富、按需触发和动态变化的互连网络，以支持各个处理部件的灵活信息交互。互连系统作为一个复杂系统，要支持结构的演化，核心就是在于必须有一个支持跨层次、多样化和动态可变的高级互连网络。SDI技术通过解耦网络功能与硬件实体，构建起柔性可重构的计算资源交互体系，为突破刚性架构限制、提升复杂计算系统效能提供了系统性解决方案，实现从物理层、数据链路层、网络层到业务层全维软件定义的网络体系，从而将刚性网络变为柔性网络，构建可定义、可重组、可重构、可重建的全维定义互连架构，为新型数据中心、云计算等信息基础设施提供高效扁平化和灵活定义的互连，为自动驾驶、即时保密通信等情景网络动态服务提供按需而变的“柔性骨骼”，为物联网时代大规模传感节点的群体智能提供灵活定义的“按需连接”。

生成式互连结构在软件定义互连的基础上，将互连网络软件定义的层次从计算系统间软件定义互连结构向下延伸到领域专用软件定义芯片、FPGA芯片等不同颗粒度的可重构计算芯片内，实现上至系统，下至芯片的全维可定义的互连网络支撑环境，能够支持多种互连体系和网络突破的共存。如图9所示为典型软件定义互连系统应用示意图，在这个环境中，不同的互

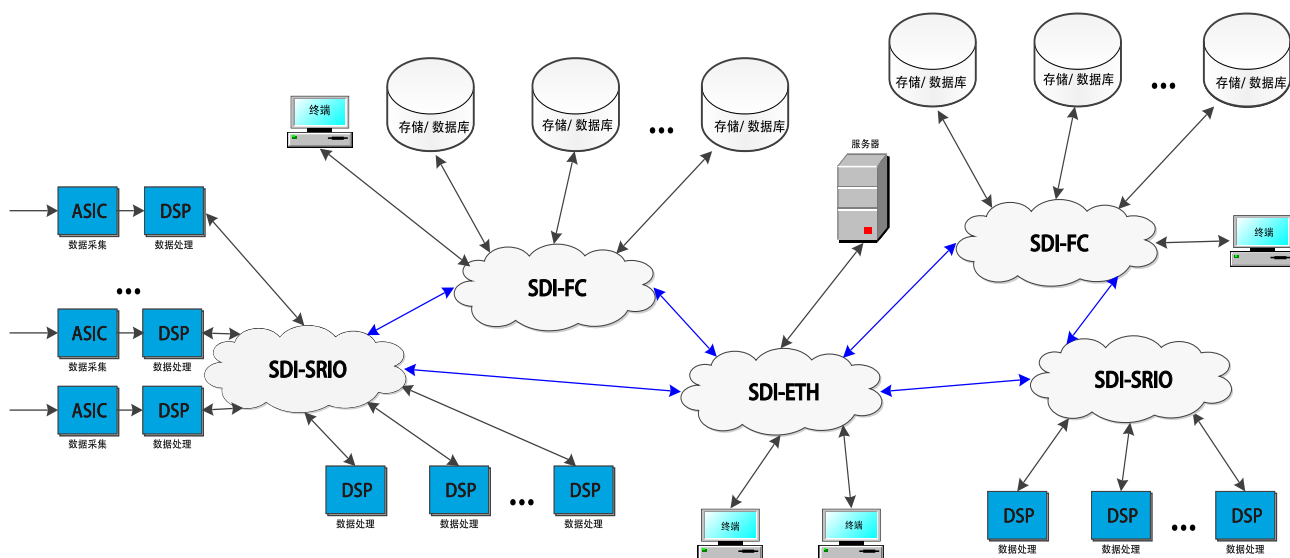


图9 典型软件定义互连系统应用示意

连体制可以像应用一样被开发、部署和升级，而无需对底层网络硬件进行大规模的改动就可以根据业务需求快速调整网络配置，实现资源的最优分配和利用。该环境不仅具备高度的可扩展性，允许根据业务需求灵活地添加或移除互连资源，以支持网络的扩展，而且还支持多种网络协议和标准。这种多协议支持确保了不同设备和系统之间的互操作性，从而实现了网络环境的多元化和兼容性。

相比于现有支持固定协议、固定带宽分配和固定交换结构的互连方式，生成式互连结构在硬件层面实现了互连协议、端口类型、拓扑结构、带宽分配、互连模式等关键参数的软件定义和按需生成，可以支持不同协议、不同带宽、不同互连模式的数据交互通信需求，实现网络资源的配置和优化，以满足不同计算节点的需求，达到互连协议可定义、互连协议可转换、互连拓扑可转换、互连带宽可调整、互连端口可定义、互连内容可处理以及交换模式可定义的效果，从而实现网络的自动化和智能化管理。

SDI技术通过对计算互连体系的软件定义重构，从根本上解决了传统架构在动态负载、通信效率、扩展性方面的刚性瓶颈，为大规模AI训练、异构算力协同、云边端融合等复杂场景提供了弹性高效的技术路径。随着计算范式向分布式、泛在化演进，SDI正成为支撑下一代智能计算架构的核心使能技术，推动算力网络从功能固定的硬件系统向灵活智能的软件定义体系加速转型。

2.3 生成式复杂网络的构建方法

在数字化转型与AI技术深度渗透的背景下，网络系统面临着任务需求动态变化、资源约束多样化等挑战，传统网络构建方式已难以满足复杂应用场景的需求。SDI技术的发展为网络架构创新提供了新路径，基于SDI驱动构建生成式复杂网络成为提升网络灵活性与适应性的关键。其构建过程包含需求驱动的网络生成算法、自适应演化机制、多目标优化设计三大关键技术，通过协同作用实现网络拓扑的动态生成与持续优化。与此同时，AI技术与SDI的深度融合，正推动计算资源分

配模式的革新，并为AI应用的实时处理能力提供底层架构支撑。

2.3.1 需求驱动的网络生成算法

需求驱动的网络生成算法是构建生成式复杂网络的基础，其核心在于通过AI赋能资源感知与拓扑生成机制，基于实时任务特征动态生成最优拓扑结构，以提升网络资源利用率。在实际应用中，不同的业务任务对网络带宽、延迟、可靠性等指标的需求差异显著。对于视频流传输场景，需要高带宽和低抖动的网络环境，而工业控制指令传输场景则更强调低延迟和高可靠性。传统网络构建往往采用静态拓扑设计，难以适应多样化的任务需求。

基于深度学习的流量预测模型可对异构计算节点（如GPU集群、NPU阵列）的通信模式进行建模，通过分析历史流量数据、任务负载特征及实时资源利用率，提前预判数据交互高峰时段与热点区域。例如，在大规模AI训练场景中，分布式训练框架的模型并行与数据并行策略会产生周期性的通信密集型需求，AI算法可根据训练阶段（如前向传播、反向传播）动态预测各节点间的数据交互量，从而触发SDI的拓扑重构机制，将高带宽链路优先分配给关键通信路径。

这种需求驱动的拓扑生成机制突破了传统“被动响应”模式，转向“主动预测”范式。以强化学习（RL）为例，将网络拓扑的生成过程建模为马尔可夫决策过程（MDP），以吞吐量最大化、延迟最小化为优化目标，通过与网络环境的持续交互学习最优策略。当检测到自动驾驶场景中传感器数据突发激增时，RL智能体可快速激活高速互连通道，将摄像头、雷达等设备的数据流优先路由至边缘计算节点，确保实时决策所需的低延迟通信。

2.3.2 生成式复杂网络的自适应演化机制

自适应演化机制是保障生成式复杂网络持续稳定运行的核心，它通过实时性能反馈动态调整网络连接关系，实现系统自愈与性能优化，如图10所示为网络拓扑随时间演化示意图。网络在运行过程中，受到硬件故障、流量波动、外部攻击等因素影响，其性能可

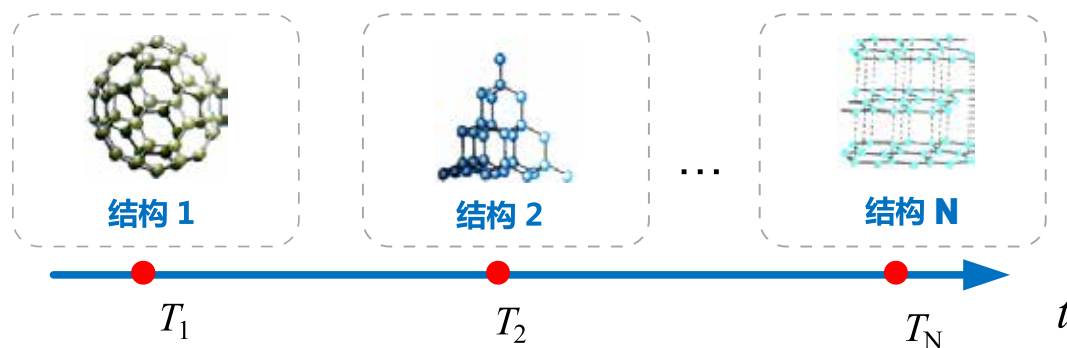


图10 自适应演化机制网络拓扑演化示意图

能会下降甚至出现故障。传统网络缺乏对自身状态的动态感知与调整能力，一旦出现问题，往往需要人工介入处理，响应速度慢且效率低。

网络拓扑自适应演化机制是指网络拓扑结构在时间上的动态变化过程，包括网络的生长、演化和衰退。这种机制通过分析网络节点和连接的变化，揭示网络的演变规律和趋势，有助于预测网络的未来发展。着眼于生成式复杂网络，网络拓扑的演化主要根据多应用场景需求、单应用场景持续优化需求及网络拓扑容错机制设计等进行，通过不同的需求反馈例如节点属性、节点增减、资源配置、数据传输时间等对网络拓扑进行动态持续优化，使得生成式复杂网络拓扑随时间的变化持续朝着满足当前应用场景需求发展。

目前，针对网络拓扑演化机制的数学模型主要有BA模型、WS模型、ER模型等，但是这些演化机制最终指向的网络模型有特定限制，例如BA模型最终会生成无标度网络，WS模型则是将规则网络朝着小世界网络的方向优化，这些对于生成式网络来说有着一定的局限性，由于生成式网络的构建是面向多场景应用的，所以朝着固定网络拓扑优化不能满足灵活多变的应用需求，需要建立一个基于应用场景进行参数反馈的演化机制。

自适应演化机制通过AI与SDI的深度融合形成了“感知-决策-执行-反馈”的闭环优化体系。在感知层，部署于网络节点和链路中的传感器及监测设备，实时采集网络的吞吐量、延迟、丢包率等性能指标数据；决策层利用图神经网络（GNN）构建全网资源图

谱，将计算节点、存储单元、通信链路抽象为图结构中的节点与边，通过GNN的消息传递机制计算最优资源分配方案；执行层则通过SDI的可编程交换结构，动态调整带宽分配比例、协议转换规则及拓扑连接方式；反馈层将实际通信性能（如延迟、丢包率）与AI预测结果进行比对，通过在线学习持续优化决策模型。

2.3.3 生成式复杂网络的多目标优化设计

多目标优化设计是平衡网络性能、成本、能效等多维度约束，生成满足复杂场景需求拓扑方案的关键。在网络构建与运行过程中，单一追求高性能可能导致成本过高和能耗增加，而过度关注成本控制又可能影响网络性能。因此，需要综合考虑多个目标，找到最佳平衡点。

多目标优化设计通过建立包含性能指标（如带宽、延迟、吞吐量）、成本指标（如设备采购、部署、运维成本）、能效指标（如单位数据传输能耗、设备待机功耗）等在内的多目标函数模型。利用多目标粒子群算法、NSGA-II算法等智能优化算法，在满足约束条件的解空间中搜索帕累托最优解集。

在复杂场景下，满足需求的网络架构往往存在多种可能。例如在智能交通系统中，既要保障车联网实时数据传输的低延迟（<100ms）和高带宽（≥100Mbps），以支撑自动驾驶决策，又需控制路侧单元部署成本与边缘服务器集群的能耗。多目标优化过程中，先将性能指标设定为刚性约束边界，同

时把成本与能效指标纳入动态优化范围。通过算法迭代,可获得一系列非支配解:有的方案通过增加冗余链路提升网络可靠性,但成本和能耗较高;有的方案采用轻量化拓扑与节能设备,虽降低运营成本,但可能牺牲部分极端场景下的性能。

这些帕累托最优解构成的解集,为决策者提供了权衡取舍的空间。决策者可根据实际预算、政策导向或业务优先级,从解集中筛选出最符合当前需求的网络架构,使生成式复杂网络在不同场景下,都能实现性能、成本与能效的动态平衡,达成综合效益最大化。

基于SDI驱动的生成式复杂网络构建方法,通过需求驱动的网络生成算法、自适应演化机制和多目标优化设计三大关键技术的协同应用,实现了网络拓扑的动态生成、持续优化与综合效益提升。在云计算、物联网、智能交通等复杂应用场景中,该方法展现出强大的适应性和竞争力,为未来网络架构的发展提供了重要的技术支撑与实践方向。

2.4 SDI与变结构计算的协同演进

2.4.1 软件定义互连技术发展目标

传统网络架构以通信为中心,难以满足当前分布式计算的需求,尤其在处理异构计算单元(CPU、GPU、NPU、XPU和ASIC等)、多样化存储介质(SRAM、DRAM、SSD和SCM等)及复杂网络连接资源方面显得复杂且易出错。同时,AI任务需求的爆炸性增长,对算力、数据流通效率和系统灵活性提出了前所未有的高要求。未来的网络体系必须向多元化、专业化、智能化演进,亟须颠覆性创新技术实现网络资源的深度可定义与灵活编排。

软件定义互连技术是为了满足未来多样化需求提出的一种新型网络技术。它代表了交换与互连技术及产业的未来演进形态,是体系架构的核心技术之一,能从架构技术上保证信息系统的高效能与灵活性,为大规模AI训练与推理提供底层支撑。软件定义互连的目标是构建一种面向未来的智慧信息基础设施,也称之为I3(Intelligent Information Infrastructure):

软件定义互连技术的发展方向有两个主要关注点。一方面将重心下移,更关注网络结构中的基础物理硬件。这意味着在软件定义互连中,对于网络的底层基础设施和硬件的发展和优化非常重要。另一方面更关注网络体系架构方面的核心技术。通过聚焦国家信息基础设施技术升级和国防武器装备体系结构升级换代,聚集国内优势力量开展协同攻关。

(1) 是着力突破软件定义互连技术基础理论问题、基础支撑技术和共性关键技术问题。这包括研究和解决软件定义互连的核心算法、协议、路由机制等方面的理论问题,以及开发支持软件定义互连的基础支撑技术,如高速网络交换芯片、高性能服务器、光纤通信等。通过解决这些问题,可以形成一套体系化的软件定义互连成果,涵盖标准规范、芯片、设备、网络平台、软件、工具等方面,提升我国在新型体系结构方向的技术引领地位,建成国内领先、国际一流的软件定义互连科研与学科方向;

(2) 瞄准国家信息基础设施的技术升级,重点发力数据中心、边缘计算等AI核心应用场景。通过构建以SDI为核心的新型数据中心体系,实现高性能、高灵活性的AI算力平台,满足多样化的模型训练与推理需求,赋能我国AI产业跨越式发展。

(3) 聚焦国防领域支撑智能作战能力,通过SDI实现电子信息装备的软件定义生成与互联互通,快速构建和迭代AI赋能的智能作战系统,提升信息化作战的适应性与先发优势。

2.4.2 软件定义互连技术发展路线

尽管软件定义互连在技术体系上具有颠覆性,但是在实现上还必须要考虑兼容现有网络。因此按照“继承发展、演进升级”的发展思路,遵循“先实现网络升级、再实现终端渗透、最后搭建全新生态”的原则,着力从芯片突破为抓手,打造涵盖芯片、设备、平台、网络、解决方案、标准的全新软件定义互连生态,服务国家信息基础设施的升级和国防电子信息装备的升级换代。

软件定义互连技术在实现上实行三步走的策略,如图11所示:第一步实现软件可定义的标准协



图11 软件定义互连发展路线

议，首先支持当前主流的高速接口协议，如RapidIO协议、FC协议、以太网协议、PCIe协议、IB协议等，通过这些协议能够以软件方式进行定义和配置，实现灵活性和可编程性。其中，世界上首款支持RapidIO、FC和以太网协议的软件定义互连芯片SDI3210已经成功问世，为软件定义互连技术的发展奠定了基础；第二步是制订软件可定义的专用协议，这些协议包括各种军标协议和行业应用协议。例如AI集群内部高速互连协议，满足AI基础设施的定制化高性能需求。通过将这些专用协议纳入软件定义互连技术的范畴，能够满足不同领域和行业的特定需求，提供定制化的解决方案；第三步是实现用户可自由定义的互连协议，通过提供灵活的编程接口和工具，用户可以根据自己的要求定义和配置互连协议，以实现更高的个性化和定制化。

通过以上三步走的策略，软件定义互连技术可以逐步实现与现有网络的兼容，并在此基础上推进技术的发展。这种渐进的方式可以确保软件定义互连技术的平稳过渡和广泛应用，同时为国家信息基础设施的升级和国防电子信息装备的升级换代提供支持。这也为未来构建更灵活、可编程和可定制的互连网络奠定了基础，满足不断变化的应用需求。

2.4.3 软件定义互连技术核心应用

软件定义互连技术可以实现异构协议的互连互

通，使其在人工智能领域展现出强大的赋能优势，成为关键的加速器。在高性能AI计算集群方面，SDI为千卡/万卡级GPU/NPU集群提供超高带宽、超低延迟的数据交换，优化模型并行、数据并行等策略下的通信效率，最大化集群算力利用率，显著缩短大模型训练周期。针对大规模数据处理与流式分析需求，SDI能在异构存储与计算单元间建立高效数据通道，加速数据供给与预处理，保障实时流式AI分析所需的高吞吐、低时延数据交互。在异构AI算力融合场景下，SDI无缝连接CPU、GPU、NPU及各类AI加速卡，实现任务动态调度和资源池化，提升整体系统效率与灵活性，支持混合精度计算等复杂需求。更进一步，SDI本身的灵活可编程性使其成为部署网络AI（如智能流量调度、异常检测）的理想平台，形成“AI优化网络，网络支撑AI”的良性循环；在边缘AI与智能终端协同领域，SDI高效连接边缘节点、云端与物联网终端，为分布式AI推理、联邦学习等提供可靠灵活的数据互连基础，支撑实时智能决策。

综上所述，SDI技术通过其协议级可定义、硬件级高性能和架构级灵活性，正成为人工智能基础设施不可或缺的关键使能技术。它直接赋能AI计算集群的高效协同、海量数据的实时流动、异构算力的无缝融合，并支撑网络自身的智能化演进，是突破当前AI算力与数据流通瓶颈、释放人工智能巨大潜力的核心基石。

晶圆级封装与物理实现

对2.5D、3D封装技术及晶圆级封装技术 (SoW) 进行介绍，特别是国内SDSoW及国外SoW的技术现状和生成式结构计算的需求结合，导出生成式结构计算的工程物理载体新需求。然后从“规模法则”出发，描述生成式结构计算的工程物理特征，对生成式结构计算对工程物理载体的需求进行总结。针对生成式结构计算的工程物理特征，结合当前晶圆级封装的物理实现技术，给出采用SDSoW的技术内涵、实现方式、结构优势，得出晶上生成式结构计算的先进计算新路径SDSoW，对生成式结构计算系统与芯片先进制程技术弱绑定的可行性进行证明。

3.1 先进封装进展

随着半导体技术的持续演进，先进封装已成为驱动产业发展的关键领域之一。当前主流的先进封装技术，如2.5D封装、3D封装和晶圆级封装 (WLP)，主要面向同构同质芯片的集成或局限于较小规模的集成，为应对大规模异构异质芯片集成的挑战，国内研究团队提出了软件定义晶上系统 (Software Defined System on Wafer, SDSoW) 技术，旨在实现晶圆级规模的异构异质芯片高效集成。

3.1.1 2.5D封装

2.5D封装可以实现多个芯片的高密度互连 (图12)，通过中介层 (interposer) 将多颗芯片集成为一个封装。在2.5D封装中，芯片并排放置在中介层顶部，通过芯片的微凸块 (uBump) 和中

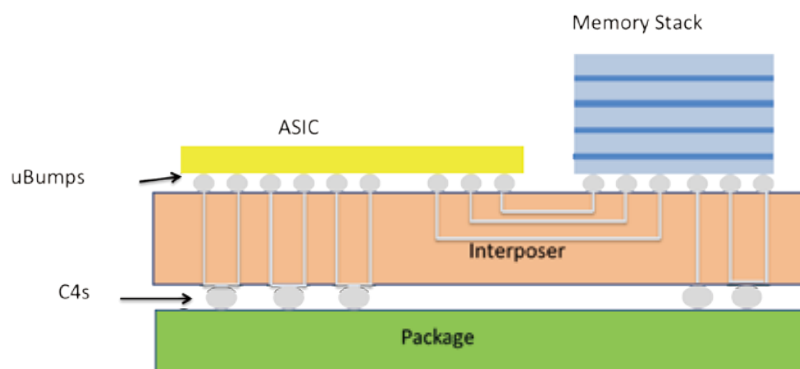


图12 2.5D封装示意图

介层中的布线实现互连。中介层通过硅通孔 (TSV) 实现上下层的互连, 再通过锡球 (C4) 焊接至传统2D的封装基板上。

其中中介层 (interposer) 是一种由硅和有机材料制成的硅基板, 是先进封装中多芯片模块传递电信号的管道, 可以实现芯片间的互连, 也可以实现与封装基板的互连, 充当多颗裸片和电路板之间的桥梁。硅中介层是一种经过验证的技术, 具有较高的细间距布线能力和可靠的TSV能力, 可以实现高密度I/O需求, 在2.5D封装中扮演着关键角色。而硅通孔 (TSV) 是2.5D封装解决方案的关键实现技术, 是在晶圆中填充以铜, 提供贯通硅晶圆裸片的垂直互连, 用最短路径将硅片一侧和另一侧进行电气连通。

当下火热的TSMC的CoWoS (Chip on wafer on substrate) 封装技术就是2.5D先进封装的代表。CoWoS技术主要有三种技术路线: CoWoS-S、CoWoS-R、CoWoS-L。

(1) CoWoS-S

CoWoS-S封装示意图如图13所示。

CoWoS-S采用硅中介层作为互连基板, 利用硅通孔 (TSV) 和再分布层 (RDL) 实现芯片间垂直与水平连接。中介层充当通信层, 可提供超高密度互连 (微凸块间距30~60微米), 能够显著缩短信号传输路径, 降低延迟和功耗, 同时硅中介层支持嵌入深沟电容器 (eDTC) 以优化电源完整性和信号稳定性。

CoWoS-S适合AI服务器芯片、高效能运算 (HPC) 产品, 例如NVIDIA Hopper H100、H200和AMD MI300。但由于硅材料较脆弱, 当中介层尺寸变大时良率难以提升。

(2) CoWoS-R

CoWoS-R封装示意图如图14所示。

CoWoS-R的中介层由多层铜布线 (最多6层) 构成, 最小线宽/间距达 $2\mu\text{m}/4\mu\text{m}$, 支持高密度互连。RDL采用柔性聚合物材料, 相比硅中介层具备更好的机械缓冲性, 可缓解芯片与基板间的热膨胀系数 (CTE) 失配问题, 显著降低C4凸点应力, 提升封装可靠性, 但其互联密度低于CoWoS-S, 适合对性能要求不高的应用场景。

(3) CoWoS-L

CoWoS-L的封装示意图如图15所示。

CoWoS-L在芯片关键区域 (如GPU与HBM之间) 嵌入小型硅桥 (见图15中的局部硅互连 (LSI)), 通过 $40\mu\text{m}$ 间距的微凸块实现超高密度互连 (密度达CoWoS-R的5倍), 满足AI芯片的TB/s级带宽需求。在非关键区域采用低成本聚合物布线层 (见图15中的有机RDL层), CoWoS-L封装可大幅减少硅中介层面积, 缓解热膨胀失配问题, 提升封装可靠性。

CoWoS依赖硅中介层实现高密度互连, 但在AI时代, CoWoS暴露出明显局限。最突出的问题是光罩尺寸限制, 这直接制约了单个模块的最大封装面积。在此背景之下, 台积电对CoWoS进行重大架构升级, 推出了面板级的CoPoS。CoPoS保留了CoWoS的2.5D架构, 但

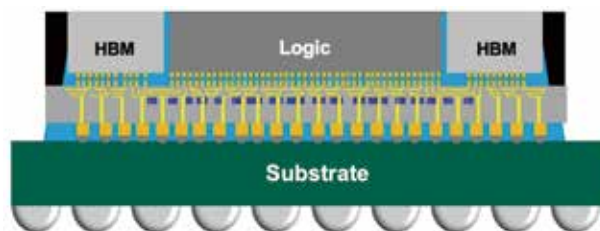


图13 CoWoS-S示意图

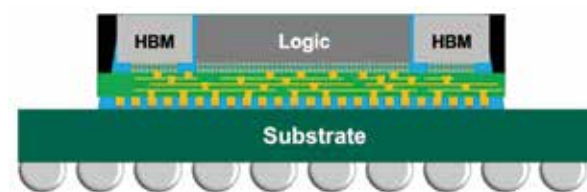


图14 CoWoS-R封装示意

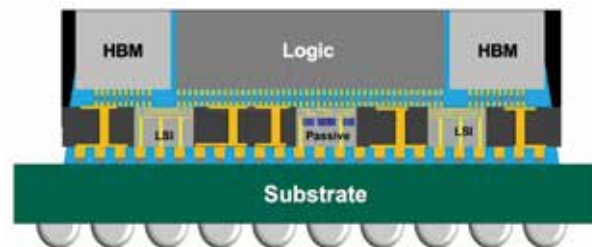


图15 CoWoS-L的封装示意

其用面板级硅/玻璃中介层代替了晶圆级硅中介层，芯片通过微凸点 (Micro-bump) 连接中介层。CoPoS突破了封装尺寸限制，能够支持更大芯片/HBM集成，尤其适合集成多颗高性能、高功耗芯片 (如GPU与HBM)。CoPoS的中介层材料从传统硅转向玻璃，玻璃具备更高的成本效益和热稳定性，其热膨胀系数 (CTE) 匹配度更优，翘曲率更低，可提升大面积整合的良率与制程控制，但需要高精度TGV (玻璃通孔) 技术。

3.1.2 3D封装

3D封装通过TSV (硅通孔) 技术，在芯片上刻蚀垂直通孔，并填充金属，以此来完成多个晶粒的上下堆叠封装，属于纵向封装。相较于2D封装芯片平铺在基板上进行互连的方式，3D封装能够合理利用封装内的垂直方向空间，通过TSV等垂直互联技术，将芯片在垂直方向上进行互连，实现更短互联距离和更小封装尺寸。3D封装示意图如图16所示。

3.1.3 晶圆级封装 (SOW)

晶圆级封装 (SOW) 是指将多颗同质同构的芯片贴装在晶圆上，通过晶圆上的高密度布线实现多颗同质同构芯片间的互连。如图17所示为特斯拉的Dojo系统的核心，其采用TSMC的Info-SOW封装技术将25颗芯片集成在一个晶圆上。

3.1.4 软件定义晶上系统

当前主流的先进封装技术主要面向同构同质芯片的集成或局限于较小规模的集成，为应对大规模异构异质芯片集成

的挑战，国内研究团队在2019年提出了软件定义晶上系统 (SDSoW) 技术，旨在实现晶圆级规模的异构异质芯片高效集成。

SDSoW的结构如图18所示，通过异构异质的晶圆级集成可以将2.5D/3D小规模封装升级至晶圆级大规模集成，将单一工艺拓展至多种工艺，将硅基材料拓展至多种异质材料，打破了先进工艺制程的限制，在28nm工艺下实现算力密度较传统方案提升3个数量级的巨大优势。

SDSoW具有晶圆级大规模高密度封装/布线资



图16 3D封装示意



图17 特斯拉Dojo系统的核心晶圆

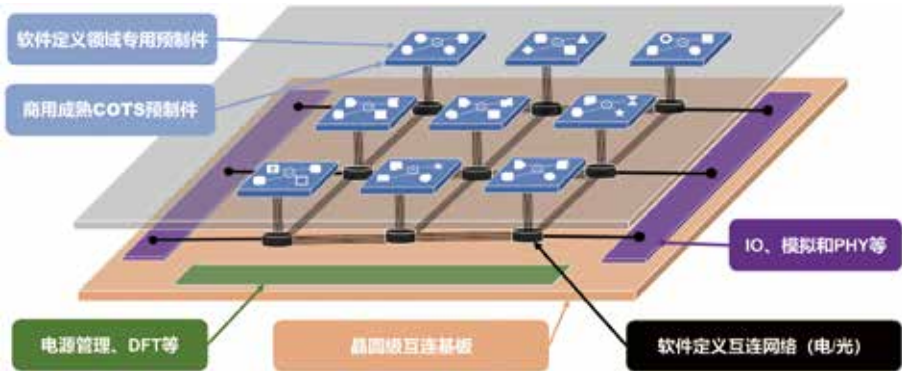


图18 SDSoW示意图

源，通过软件定义互连可支持“异构芯粒动态重组”机制，能够很好的支持生成式结构计算所需的规模法则与多样性需求。

3.2 生成式结构计算的工程物理特征

生成式结构计算的核心特征为面向领域应用特征优化的基本元素由生成式互连结构相连，实现针对不同应用任务计算需求的计算结构生成，从而实现先进计算的发展愿景。

而传统封裸芯成芯片、载芯片与板卡、聚板卡成机柜、堆机柜成系统的计算系统构建工程技术路线伴随着物理空间的扩张、散热结构、供电网络、机械支撑等非计算资源的引入导致单位体积内有效计算资源的比例（即资源密度）逐级下降，同时信号在长距离物理介质中传输的固有延迟、信号完整性问题以及层级接口的处理开销，使得跨层级、跨模块的通信时延呈非线性增长。因此，即使通过生成式互连结构将基本元素聚合起来形成了与应用任务适配的计算结构，但由于其天然的缺陷势必也难以满足生成式结构计算的要求。

资源规模的大小决定了生成式互连结构的可变空间规模。如果资源规模小，根据幂律（非线性）分布，则生成式互连结构的可变空间必然减小，当资源规模小到一定程度（如计算基本元素的个数小于10），则生成式互连结构的可变空间会急剧萎缩，此时基本元素间的互连关系或互连选择必然会趋于固定，基于元素“键合效应”的多样性结构将无法获得，此时生成式互连结构将趋于刚性结构进而失去其意义，计算系统也将失去必要多样性，无法在面对多样化的计算需求下支撑多样化计算结构的生成。

而资源密度大小决定了生成互连结构的时延，若资源密度小，生成式互连结构即使能适配不同的计算需求生成相应的计算结构，但由于资源稀疏带来较长的信号互连长度，这必然会导致通信时延的增加，进而会带来“存储墙”和“通信墙”问题，生成式结构计算低通信时延的优势也将无法得到发挥。

综上所述，从“规模法则”出发，生成式结构计

算需满足两大核心需求，如图19所示：一是高密度封装资源，二是生成式互连结构。高密度的封装资源可提供高阶的互连密度、高异构兼容性，降低通信延时的基础上为变结构计算提供大规模的基本元素；而生成式互连结构具有高密度的互连资源，可提供高动态重构能力，可极大的拓展计算结构的可变空间。

因此，作为生成式结构计算的理想物理载体，其必须具备高密度与大规模的资源集成的特征以满足多样化的计算需求。

3.3 SDSoW: 生成式结构计算的工程载体

在深刻剖析集成电路发展困局与综合考量我国技术优势的基础上，结合生成式结构计算理想物理载体的需求，邬江兴院士提出了软件定义晶上系统（Software Defined System on Wafer, SDSoW）技术。

生成式变结构计算的核心在于按算法需求动态重构计算架构，要求算力的硬件载体能够实现以软件驱动物理结构变化，大幅度提升特定计算结构对特定算法的执行效能。SDSoW就是要推动计算架构从“刚性流水线”向“软件可塑形”跃迁，打通生成式变结构计算从理论到技术物理实现的闭环，使得基于二流器件或部件构建一流系统成为可能，其结构如图20所示，实现形态如图21所示。

软件定义晶上系统包含两个层面技术内涵：

(1) 软件定义体系结构，即软件定义系统、系统之系统。软件定义晶上系统以面向领域应用的高性能、高效能与高灵活为目标，遵从“结构适配应用”的软硬件协同技术路线，以生成式互连和软件定义节点为核心特征，通过分析领域应用特性，设计领域专用的软硬件协同系统架构与实现结构。该结构既有领域内的灵活性，又有专用级高性能、高效能，并天然匹配



图19 生成式结构计算的工程物理特征

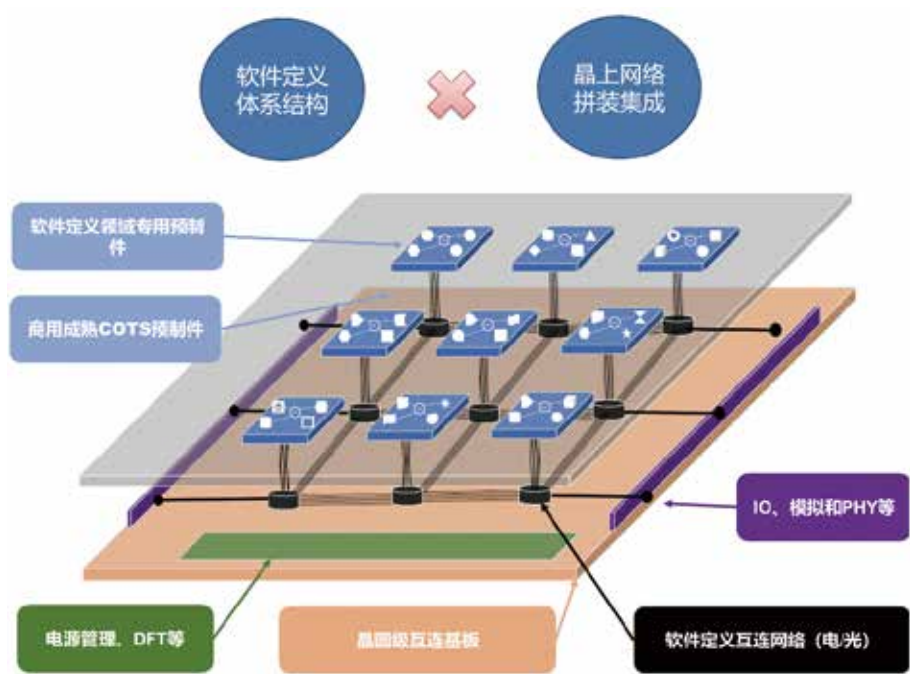


图20 SDSOW结构图

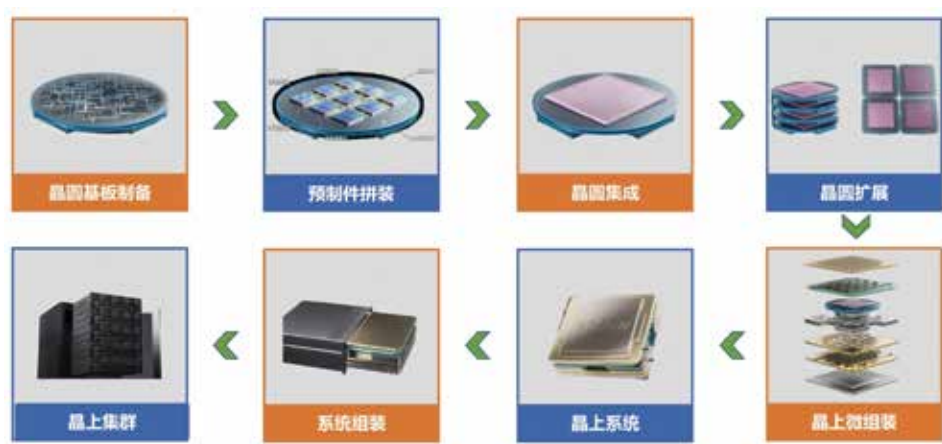


图21 SDSOW实现形态

开放复杂性系统乃至智能系统的灵活高效结构内涵，拓展信息平台面向未知威胁的作战能力生成模式。

(2) 晶上拼装集成。软件定义晶上系统按照系统科学与工程思维，以系统级的整体最优为目标，以目前人类最具规模经济性与成熟度的晶圆互连网络基板为集成载体，实现不同材料（氧化硅、砷化镓、氮化镓等）、不同工艺（130nm、65nm、7nm等）、

不同功能（传感、通信、计算、存储等）、不同架构（CPU、GPU、NPU、DSP等）的芯粒（Dielets）按照2D/2.5D/3D进行物理形态的拼装集成。该系统极大提升资源集成规模和密度，克服芯片封装与逐级堆砌集成对复杂系统带来的性能损耗，在系统层面实现摩尔定律延续，极大缩小芯片间通信半径，避免带宽下降、时延增加与能效下降，形成以应用场景垂直

整合、随阅历数据自我演化的新一代智能集成电路设计、工艺和应用技术。既可增维体系架构创新带来的性能和效能增益，又能扩维晶圆互连带来的性能和密度增益，可将摩尔定律的微缩增益“无损”拓展至系统维度，使得系统整体性能不再单纯依赖于部件级的指标先进性。

同时，SDSoW还具备五大能力：

(1) 系统破局能力。SDSoW摆脱“核心器件决定论”思维，改变“芯粒、芯片、模组、机匣、机架、系统”的逐级堆砌、逐级插损式工程技术路线，通过晶圆级异构集成实现功能解构—晶上重组，达成功能等价、系统最优的目标，将制程工艺短板通过系统工程方法转化为非主要矛盾。

(2) 整体增效能力。借助晶圆级高密度互连、超短距离、异质异构封装，获得高带宽、低延迟、低功耗的系统增益，SDSoW系统的带宽可提升一个数量级，时延可缩小一个数量级，功耗可降低一个数量级，系统效能可提升三个数量级。

(3) 通专融合能力。基于晶圆级系统硬件可编程/重定义架构，可通过软件实时配置或由AI大模型生成SDSoW功能和性能。在同一物理载体上，能根据不同应用需求或使用场景，实现“一平台、多样性”生成式变结构计算，既可满足专用场景特殊算力需求，又

能兼顾领域内相对灵活的通用算力要求。

(4) 开源协同能力。通过建立SDSoW开源社区，发布基础互连协议、动态控制器及生成式变结构计算工具链等，SDSoW可构建“中国引领、全球参与”的生态环境，以开放破垄断，形成相对Chiplet (小芯片) 路线的比较优势。

(5) 内生安全能力。SDSoW可从源头上应对开放生态带来的新域新质安全挑战，通过引入内生安全架构实现开放可控，即便供应链欠安全，仍能保证开放条件下该系统具有“开箱即用、默认安全”的网络弹性。

变结构计算工程物理实现是大规模高密度资源集成，而SDSoW是高效结构计算系统在SMVT四维空间的新结构。SDSoW的算力收益在工艺差两代的情况下，仍能与现有传统工艺技术方法相比具有相当甚至更优秀的性能，而在同等工艺条件下能提升4到6个数量级。

晶上生成式变结构计算可为我国打破算力芯片“制程工艺茧房”开辟新路径，走出一条用“三流材料、二流工艺”达成“一流能力”的系统工程级创新路线。通过应用与设计、算法与算力的垂直整合，解耦当前我国算力基础设施产品对芯片先进制程工艺的强依赖关系，最大程度获取体系结构与工艺进步的综合增益。晶上生成式变结构计算也能为全球智算普惠提供中国方案。

第四章

超融合算力网络环境

聚焦网络支撑环境与应用网络体制解耦的新发展范式，构建AI赋能的高度融合、高效协同、安全可信的超融合算力网络环境，用超融合基础设施解决“算网存”割裂困境，可快速、经济地生成满足异构异质多样性算网服务需求的应用网络，彻底变革网络传统升级换代模式，形成“算力-数据-网络”全要素协同的智能计算环境。

4.1 算力网络发展面临的挑战

4.1.1 网络发展范式的历史沿革

网络发展范式是网络技术创新发展的理论基础和实践规范，将范式理论实践于信息网络领域拥有巨大的研究价值。目前来看，网络技术体系发展历史可以大致分为以下两种范式。

(1) 第一范式：以网络业务为中心

在通信网络发展初期，网络建设以业务需求为驱动，采用以硬件为主的布线逻辑，专用基础设施架构仅提供特定的服务与应用，如图22所示，电话通信网主要面向语音服务，而广播电视网则专注于视频/音频混合信号传递。在这一时期，网络技术体制呈现出功能单一、服务高效、彼此隔离的特点。受限于当时的通信传输技术的发展水平，网络终端以电话机、电报机、电视机等只能完成收发功能的简单设备为主，从而可以降低终端推广应用的门槛，上层应用与底层架构紧密耦合，导致通信设备与协议功能紧密耦合，网络交换芯片及其设备制造业逐渐形



电话网



广播电视网



公共电报网
(自动电报网)



卫星通信网

图22 以网络业务为中心的发展范式

成较高的行业壁垒。相应地，由于网络只需要提供特定的服务与应用，一种网络使用一种通信技术，能够针对业务特征设计最适配的网络设施、编码方式和交换协议，因此可以最大程度的提高服务质量和网络效能。各通信网络的隔离性体现在承载流量特征单一，封闭的路由控制协议对外不兼容，以及传输媒介与接入设备的不同。

“业务即网络，网络即业务”的模式带来服务质量高、网络效能高等优势。然而，专用网络意味着基础建设代价高、运行维护成本高、升级换代难度高，无论是业务的扩展还是网络的扩展性都受到了严重的制约，随着新兴业务需求的不断涌现，为每一类业务新建网络设施并设计相应协议的方式已经不能再适应网络的发展规律。

(2) 第二范式:以业务数据透明转发为中心

随着民用计算机的迅速普及，网络终端由简单设备转换为计算机、服务器、手机、平板电脑等能够进行数据处理的复杂设备，大量智能化终端迫切的联网通信需求催生了遵循“以业务数据透明转发为中心”发展范式的互联网的诞生，如图23所示。互联网采用TCP/IP协议进行高效的数据传输，业务处理集中在各类智能化终端，网络提供不区分业务的端到端连接和“尽力而为”的服务模式，统一以服务质量 (Quality of Service, QoS) 来衡量。通信业务开始抽象为异构设

备互联和数据分组交互，促使上层业务与底层网络分离，打破了各种高层业务应用之间的隔离与壁垒，推动信息网络基础设施互联互通和资源共享，有利于促进消费升级、产业转型和民生改善。在互联网发展的初期，人们对网络的需求集中在网际互连和数据传输服务，网络与业务分离的发展范式提高了网络资源利用率，降低了网络的建设与部署成本，在一定程度上释放了网络业务的创新活力。

然而，IP协议作为大规模、异构网络的互联互通的关键，仅仅提供了单一的寻址方式，并不适用于所有数据传输服务。进入21世纪，5G、云计算、物联网、机器学习、大数据等通信与计算技术发展带动网络技术革新，互联网繁荣呈现全新态势。以IP为中心的互联网瘦腰模型已无法满足多元化业务需求，陈旧的设计目标、排他性的技术标准、低效能的发展模式等，使之很难满足“网络空间时代”日新月异的应用需求。

4.1.2 传统网络发展范式面临的困境

通过对网络技术体系发展范式演变路线的梳理，可以看出，当前的网络体系结构面临诸多困境，主要集中在以下四个方面：

困境一：对单一网络体系的演进式修补使得系统工程复杂度失控。传统消费型互联网以变长分组交

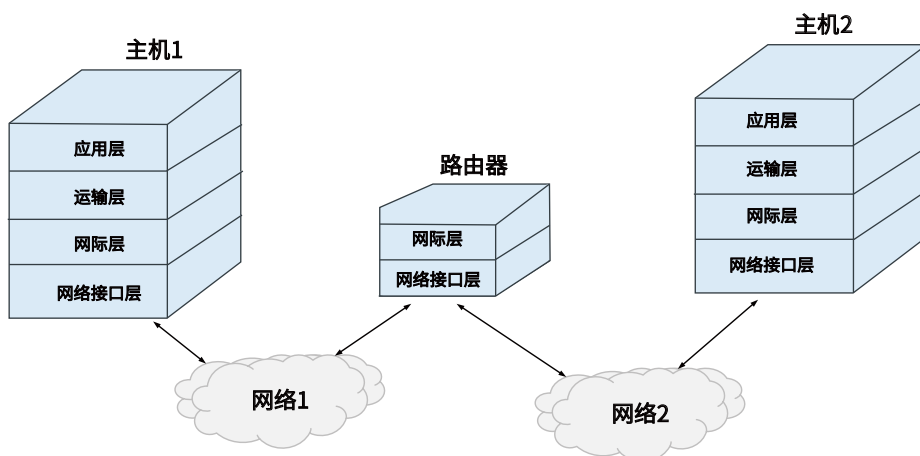


图23 以业务数据透明转发为中心的发展范式

换、尽力而为传送、最短路径寻址和TCP/IP协议等为基础，靠堆砌带宽和处理资源实现多媒体通信功能，在经历了五十余年的辉煌后，其“基因缺陷”已暴露无遗。为支持多样化、个性化应用和网络安全需求，不得不对现有的单一承载网络体系（即TCP/IP）进行持续的增量修补，网络复杂度急剧膨胀。截至2020年年底，RFC数量已超8000多个，仅核心路由器的软件代码量就达到上亿行，早已背离了“Simple is Best”的网络设计初衷。

困境二：单一IP体系在众多方面制约着网络的传输性能。万物互联时代，大带宽、高容量不再是评价网络质量的唯一指标，精细化的应用场景对网络服务能力提出更全面的要求。面对上述需求，现有基于IP的网络架构给出的方案只能是虚拟专网、网络切片、边缘计算等虚拟化方案作为过渡，无法从根本上解决问题。事实证明，网络与业务功能确实可以分离，但性能与功能并不总是可以分离的，业务功能往往可用虚拟化形式来实现，但性能常常与网络技术体系、寻址/编址方式、协议体系、路由控制、交换模式、传送方式等底层基线技术强相关。

困境三：堆叠式的复杂网络技术体系凸显广义安全危机。堆叠式发展的复杂网络，使得功能安全与网络安全问题相互交织，原本清晰的功能边界变得模糊，传统可靠性假设前提条件已不复存在。以往将安全措施集中部署在网络边缘或特定区域的做法已不足以应对当前的安全威胁，攻击者可以利用系统间的接口、协议漏洞或配置错误，轻易绕过传统防线，直达核心系统。无论是基于云—网—边—端设施中软硬件代码中存在的未知漏洞后门、病毒木马等未知安全威胁，还是随着网络软硬件规模的迅猛扩展其随机性失效引发的可靠性问题日趋严重，迄今为止还没有一体化的理论与技术解决办法。

困境四：网络与端应用的严格分界难以适应计算、存储、转发一体化趋势。在算力网络时代，仅以信息或数据的端到端传送为目标的互联网，已远远不能满足包括云计算、边缘计算和分布式计算在内的一系列新兴服务的需要。未来发展时代中，互联网不再仅仅是一个单纯的数据传输管道，而是需要进化为

一个能够高效整合、调度和利用各个区域分布的计算资源的平台，将计算、存储等算力资源在云边端之间进行有效的连接和调度，满足算力“随时、随地、随心”的服务要求。

总之，当前主流的网络发展范式正在遭遇难以克服的挑战，包括如何支持人—机—物大规模自由乃至智能组网、全时空随机接入、服务性能可确保、安全性能可量化设计与验证度量、大数据或分布式计算应用等。因此，当前发展范式的方法论必须得到彻底的改变，源于消费互联网时代的IP网络既无法靠一味地增加各种传输资源扩展传输带宽、无节制地堆砌软硬件代码量和粗放式地使用能源方式来满足产业互联网时代新的发展要求，也不能再指望CPU性能大幅提升和刚性网络技术架构以及深度虚拟化技术，以尽力而为的方式来满足数字经济时代各种垂直行业对高性能、高效能、高安全、高可用、个性化应用等服务的需求。“一网独大，意欲解决所有应用是不可能的”。

4.1.3 算力网络发展的创新需求

算力资源的跨域联网，提供普遍服务，已成为当今时代的必然趋势。不同的应用场景对算力网络提出了差异化、个性化的需求。超算网络追求微秒级的延迟与无损传输，以支持科研的精准与高效；通算网络则注重高并发与梯度同步，确保数据处理的流畅与准确；智算网络通过算力动态切分与业务链编排，实现资源的灵活调度与优化；而新兴计算网络，如量子计算、类脑计算、光子计算等，更是以低成本广覆盖、通用协议边缘轻量化等特性，引领着未来计算技术的发展方向。从算力资源基础性、易获得性、技术经济性而言，必须通过网络化部署提高算力应用效率，然而传统信息基础网络的功能与性能已难以满足算力优先等新兴应用场景的需求。

因此，算力网络必须从“专业应用系统”向“存算通智一体化”基础设施转型。这不仅是化解多样性专业应用场景与统一基础设施矛盾的关键，更是提升算力资源基础性、易获得性、技术经济性的必由之路。为此，本文提出通过超融合基础设施构建更加智能、灵活、高效的算力网络，化解多样性专业应用

场景与统一基础设施的矛盾，更好地应对未来数字化时代挑战。

4.2 超融合算力网络环境概述

4.2.1 基本思想

迄今为止，无论是理论模型还是实用方案上均未发现哪种单一网络体制，既能有效承载应用场景明确、有经济性和功能/性能保障的预期业务，又能在整个生命周期内适配应用场景可不断丰富的未来业务。超融合算力网络环境以“网络环境与技术体制分离”的多模态网络发展新范式为基本思想，从思维视角、方法论和实践规范三个角度具体展开如下。

(1) 思维视角

信息通信基础设施向云网融合、地球空间智联网络方向转型已成为业界广泛共识。基于分布式异构计算架构和虚拟化等相关技术的云计算平台，正在使传统计算机与信息通信网络技术朝着“一切皆为服务”的融合模式发展，除了回归网络即服务外还引入了基础设施即服务、平台即服务、软件即服务等在内的新应用体验，试图将广义计算能力(计算、存储、交换、传输等)通过互联网作为商品资源进行流通，必然会产生许多与当前大众或普遍服务类型和提供方式不同的新兴应用需求。任何单一体制网络或转发模式都不可能具备应对发展过程中不确定因素影响的适应与进化能力，新范式以“统一网络支撑环境”为全生命周期内所需的多样化网络体制或转发模式及相关业务，提供计算、存储、交换、管理等可扩展的网络基线技术元素及动态聚合软硬件资源，给网络技术体系及其业务和性能发展提供更为充分的创新空间。

(2) 方法论

数字经济、海陆空天多维万物智联时代必然会对网络技术体制提出更加多样化的需求，新的方法论需要为“如何在一个技术物理环境中支持多元化网络技术体制及服务的共生共存、演进与变革兼容并蓄”问题提供“方法之方法”，并不涉及任何应用网络体制的系统架构、寻址方式、路由控制、传送制式、协议体系、业务形态乃至演进路线等实践规范层面的内

容。面向全新的思维视角，新范式将网络支撑环境与应用网络体制或转发模式及业务相分离，以网络基线技术元素化和资源动态化聚合为基础，以软件定义互连和软硬件协同编排及内生安全架构赋能的网络弹性工程等为基本手段，构建一个多维多元共生的融合环境，支撑其上的各种网络技术体制及相关业务功能和性能，满足多样化的服务需求。

(3) 实践规范

范式决定基础理论，基础理论决定具体理论，理论决定模型，模型决定模式。与传统的网络发展范式研究坐标不同，新范式将研究坐标锁定在“如何为各种类型应用网络提供一体化、可共享、安全的网络基础设施资源环境”上，即如何建构一个能够支撑各类应用网络及相关业务“快速生长或部署”之上的“开放、透明、统一、弹性、智能、安全隔离”的网络基础设施，以及可健康持续发展的技术与产业生态环境。与现有网络发展范式相比，新范式支持各种应用网络以类似“应用插件”的形式进行快速部署(或撤收)，极大地降低新兴网络技术及应用快速进入市场的门槛，以网络“平滑升级换代”的新模式取代“one by one”的陈旧换代模式，打造存量资产利用、网络技术创新、产业发展与市场应用“四位一体”和谐发展格局，营造网络模态与业务应用以及网络基础设施一体化增量发展的全新生态环境，为网络体制及关联服务充分释放创新活力。

4.2.2 总体架构

针对超融合算力网络环境对灵活性、柔性化、弹性化的要求，采用元素化方法对网络功能、资源进行细粒度拆分，自顶向下逐层进行需求抽象和逐层映射，形成了基于“业务域、技术体制域、功能域、资源域”的网络功能元素化建模机制。这一机制确保了网络资源的“元素化”，使得在统一的网络物理资源基础上，能够通过资源组合重构的方式实现或生成多种多样的应用网络，如图24所示。

业务域：根据网络所服务的业务进行分类，例如电子商务、视频流媒体、云计算等。按照业务类型进行归类，便于业务聚类、管理和优化。



图24 网络功能元素化建模分层结构

模态域：将业务需要的网络寻址方式、调度策略、安全机制、服务质量、运维管理等定义为一种网络模态，一般是为满足某一部分垂直行业特殊应用需求而由领域专家设计的。

功能域：将各个网络模态需要的网络功能进行基线能力池化，支持基线能力动态释放、重组。例如：分组调度、流量控制、带宽分配、队列管理、缓存管理等。

资源域：将基线能力所需要的资源进行元素化拆分，包括CPU、FPGA、NPU、多核处理卡、SRAM、存储、表项等类型资源。

4.2.3 运行逻辑

超融合算力网络通过生成式网络架构支持各种网络体制以类似“应用插件APP”的形式进行快速部署或撤收。这种架构适配已有或未来的网络体制，满

足多样化的垂直行业应用需求，同时充分保障多种或多个网络体制之间有足够的隔离度，确保网络环境的稳定可靠。该架构运行逻辑包括以下两个关键环节。

第一个环节是多维异构标识网络体制按需生成，负责完成网络体制所需的语义表征，实现面向业务的网络体制需求与公共云网资源可识别标识之间的转译映射。这项技术不仅实现了不同网络体制之间的标识转译映射，还建立并支撑了网络体制承载与保障优化所需的参数信息规范化模型表征，如图25所示。

在多维异构标识网络体制按需生成技术中，多维度网络体制的异构标识融合转译方法是一个重要组成部分。这种方法针对网络体制生成过程中，多维度网络体制业务标识体系在公共云网资源环境下的语义识别问题进行了研究。它实现了多维度网络体制在业务与网络之间的标识融合转译，以及网络体制之间的转

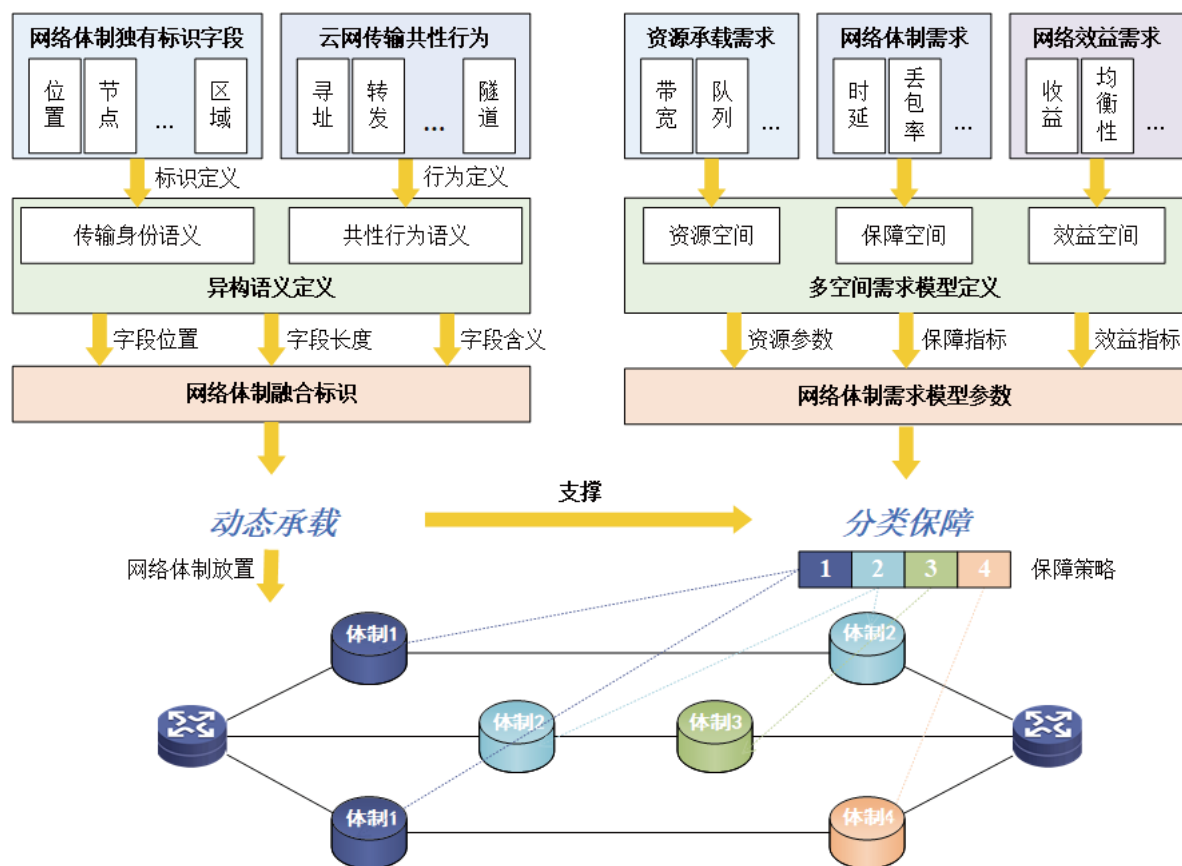


图25 多维异构标识语义表征

译。具体来说，就是从网络体制的标识语义出发，对多维度的业务网络体制标识进行映射，并根据网络体制的传输需求，将其所涉及到的转发、路由、封装、隧道等传输行为转译为云网资源层中具备共识的、可识别的、可执行的多网络体制标识。

此外，多空间的需求网络体制生成表征模型也是多维异构标识网络体制按需生成技术中的重要一环。这个模型针对多维度的网络体制在资源与服务需求表征上存在的差异化问题进行了研究。它建立了网络体制从语义到资源与保障需求表征的能力，通过对网络体制的具体需求进行解构表述，将网络体制的承载需求与优化需求拆分为具体的资源描述与服务质量描述。这与标识融合转译共同构建了多维度网络体制面向公共云网资源的标识体系、承载体系与编排体系的模型表征能力，完成了网络体制从需求到云网可识别

表征模型的实例化生成。

第二个环节是网络体制动态承载，根据网络体制的需求，决策对应于网络体制语义模型的承载策略，并对网络体制的配置信息进行下放置。在网络体制的放置过程中，这项技术实现了纵向层面的分类网络体制业务需求承载与横向层面的网络体制均衡承载，如图26所示。其中，异构网络体制数据流特性与资源需求分析技术是一个关键部分。它面向网络体制多维度与异构需求的分析问题，研究了多样化的数据流特性与资源需求分析技术，引入了多队列理论、网络微积分来模拟网络中承载的不同类型数据流，分析了网络中不同网络体制和服务质量类型流量的资源占用情况，并模型化表述了各业务之间的需求冲突模型，为网络体制承载放置决策提供了需求信息输入。

基于这些分析，线性与非线性融合网络体制动态

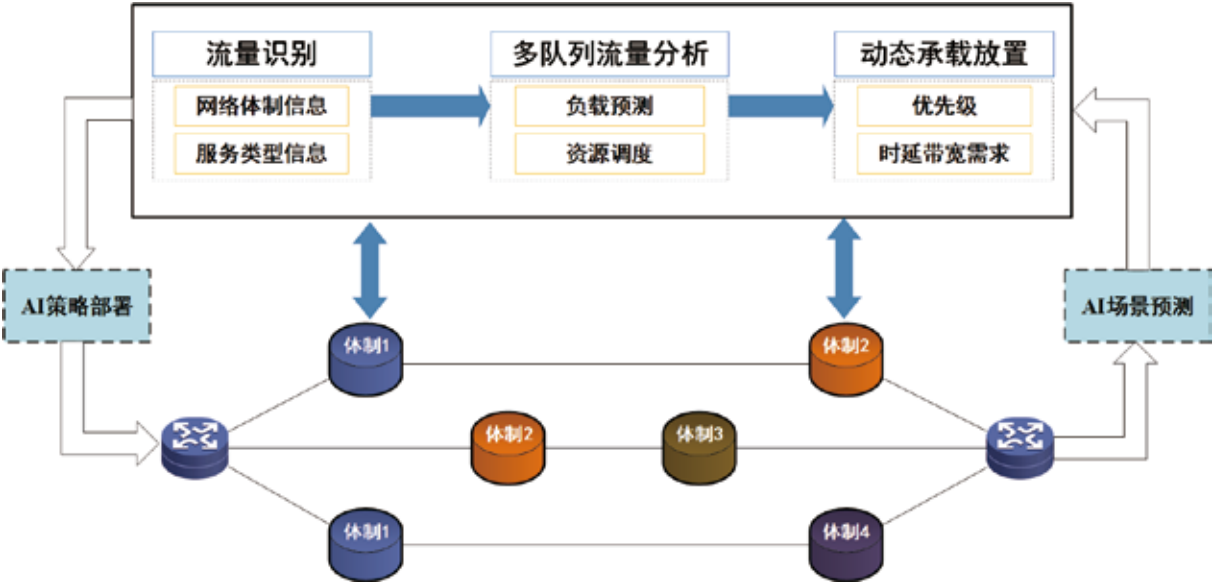


图26 网络体制动态承载机制

承载放置算法应运而生。这项算法面向异构网络体制业务的优先级、时延需求和带宽等要求，根据不同业务场景和流量特点，动态调整和优化网络资源的分配方式，实现了网络资源的集中管理和动态配置，以适应不同业务的实时需求变化，保障了自动化的网络体制放置，支撑了网络体制的无感承载。

4.3 超融合算力网络环境的关键技术

4.3.1 网络全维可定义技术

网络全维可定义技术通过打破传统网络刚性架构，建立网络各层功能全维度可定义的开放、灵活、通用架构，通过将“软件定义”思想从服务层下沉到控制层、数据层，能够对开放架构下基础网络的软/硬件、协议、接口、芯片等进行全维可定义，真正在各层面实现功能可定义，资源高效自动适配，支持各层功能在全维可定义功能平台上的动态加载和演进发展，支持异构网络的柔性互连，通过网络结构自组织和业务自适应来动态灵活适配业务发展需求，从而为网络的柔性化组织和功能多模态呈现提供平台支撑。如图27所示，网络全维可定义技术具体包括软件定义



图27 网络全维可定义的技术内涵

硬件 (Software-Defined Hardware, SDH)、软件定义互连 (Software-Defined Interconnection, SDI)、软件定义计算 (Software-Defined Computation, SDC)、软件定义功能编排 (Software-Defined Orchestration, SDO)、软件定义转发软件定义功能 (Software-Defined Forwarding, SDF)、软件定义协议 (Software-Defined Protocol, SDP)、拟态计算 (Mimic Computing, MSC) /领域专用架

构 (Domain-Specific Architectures, DSAs) 等。

全维可定义技术要求网络必须具备三大能力。首先,为了保护现有业务和网络设施能够适应网络的演进,开放架构下的全维可定义构件及其运行平台必须具有良好的稳定性,也就是说构件能够在保持新的网络协议或应用增量部署的同时也能够保证原有应用的正常运行;其次,为了保证网络能够在多维评价指标上呈现良好的变化以满足业务的变化以及新型应用的部署,构件必须具有可变化的内在结构,也就是说包处理的方式以及网络协议的运行方式可以动态改变;最后,在构件结构可变的基础上,构件要能够以某种“柔性”的方式对其结构进行调整。进一步地,柔性是构件针对应用要求对其内在结构、资源做出隐性调整,以实现网络服务效果对应用需求的动态、紧密跟随。

全维可定义技术可改变当前网络固化的运行模式,由其作为“网络功能构件市场”机制的提供者,允许多样化应用动态装配、构建与之匹配的定制化服务构件链,动态部署个性化应用,实现全网业务的灵动适配。此时,网络的功能和工作机理将不再受制于具体的协议,其服务能力不再依赖节点的初始植入设计,其服务能力空间将超越已知协议和机制的限制,

从而实现由已知协议定义的网络到需求驱动的网络形态和服务能力转变,充分满足新型网络不断演进的业务需求。

4.3.2 网络技术体制编译与加载技术

以全维可定义的网络环境为支撑,各种技术体制可以通过标准化软硬件接口从网络支撑环境中获得所需的计算/存储/交换资源,除了能够实施期望的软件编程操作外,还可以对网络技术体制所分配的硬件资源进行构造/算法/性能等的软件定义。各种技术体制在超融合网络环境中自持发展,可充分发挥各种网络体制的内在特性和优势,并不断发展、完善。

如图28所示,通过标准化的软硬件接口,超融合网络环境可以通过智能、自动或手动等加载/卸载方式,配置各种网络协议,构成各种可能的网络寻址方式、路由方式、转发方式、安全保障方式等,从而能支持各种可能的网络应用,包括工业、军事、能源、交通、金融、政务、电力、农业、移动通信、卫星通信等,提供高效能、高通量、时敏性、可靠性、移动性、安全性、可扩展性等多样化、个性化的领域专用需求。

超融合网络环境支持各种网络技术体制以类似

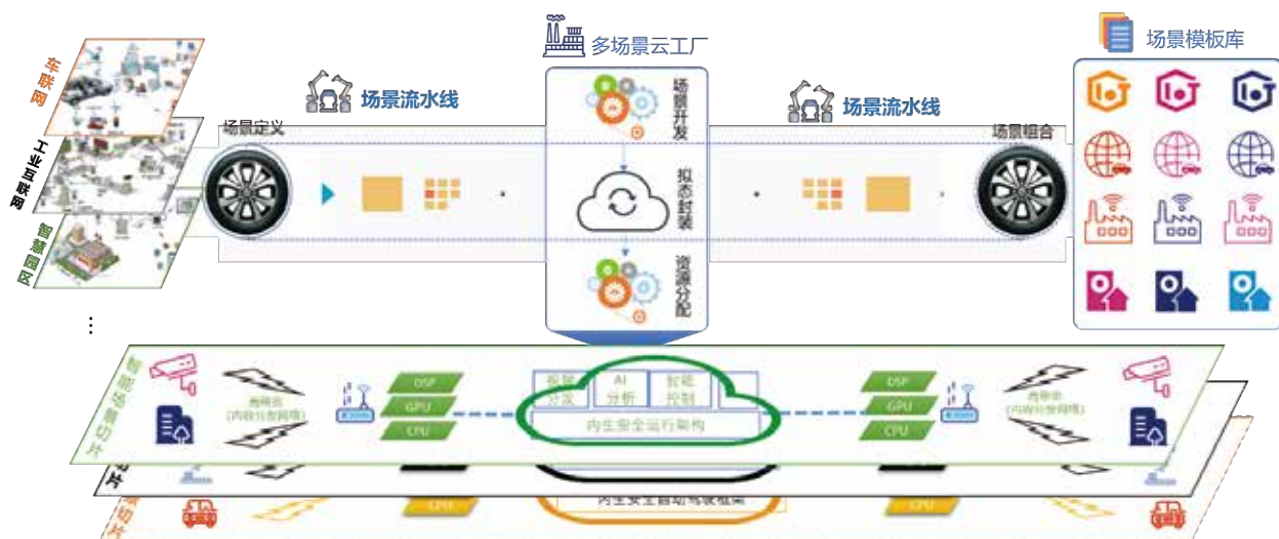


图28 多种网络体制动态加卸载载入

“应用插件APP”的形式进行快速部署或撤收，各种网络体制的技术演进、业务发展、应用创新及运维管理均属于网络技术体系的创建者或拥有者，各自均可以独立部署或演进发展。超融合网络环境不但能适配已有或未来的网络技术体制以及多样化的垂直行业应用需求，而且能充分保障多种或多个网络体制间有足够的隔离度。

4.3.3 多模态路由与寻址技术

多模态寻址与路由技术在网络体系结构中内嵌新型的寻址与路由功能要素，支持IP、内容标识、身份标识、地理空间标识等多模态标识的共存与协同，采用计算和数据分离方法将基于上述标识的多种寻址与路由机制自然配合、联合起效、融为一体，从而在根本上突破传统网络IP单一承载的制约瓶颈，满足多元化和专业化的高效服务需求。具体来说，多模态寻址与路由技术面向多模态标识空间灵活组网与高效扩展需求，支持跨模态的资源协同与资源分解，实现面向多样化标识空间的协议定义与适配，实现对网络状态、用户需求、服务类型、安全需求具有不同粒度要求的多模态寻址路由的按需切换和互联互通。

多模态寻址与路由的基本逻辑示意如图29所示，支持终端以多模态标识混合接入，基于全维可定义网络功能基础平台实施基于多模态标识的混合寻址与路由。在路由建立阶段，用户终端可直接指定标识类型，或由网络根据资源状态、业务类型、安全需求等反馈标识类型选择结果；在数据传输阶段，网络亦可智能进行标识空间切换。

多模态寻址与路由互通技术通过多种标准寻址路由协议以及自定义协议的互联互通，形成“领域专用、功能互补、多模共存、确定性能”的寻址与路由机制，通过多样化标识空间的协同技术创新和优势互补，解决现有网络存在的诸多弊端，综合有效提升网络服务能力、安全性、移动性、资源利用率等。

多模态寻址与路由系统可支持身份标识空间与现有网络融合，提升多样化终端接入的灵活性和服务一致性，从本质上解决终端移动性问题，有效提高网络管控能力；支持内容标识空间与现有网络融合，在网络节点中融入计算、存储和传输的能力，实现以“内容”为中心的网络数据自适应流动和汇聚机制，使用户获得高效的内容获取体验；支持地理空间标识空间与现有网络融合，建立地理空间标识与网络编码

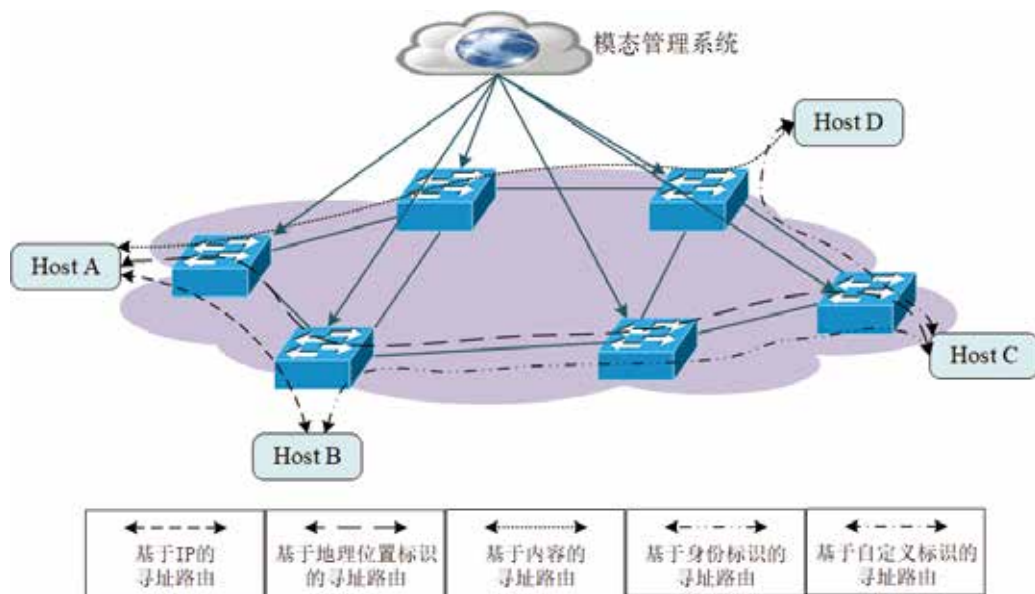


图29 多模态寻址与路由逻辑示意图

的转换关系，实现不同地理空间位置的互联互通，为跨平台的多源空间数据高效交互提供支持；支持用户自定义标识空间与现有网络融合，在网络基础寻址与路由功能层面上体现出多元融合发展，并由此构建出基于基因信息表达的体系化网络功能生成和按需演化机制，为新型网络技术创新和功能的快速部署提供基础环境支撑。

4.4 发展愿景展望

4.4.1 技术发展愿景

依托超融合算力网络环境，面向人-网-物泛在互联的行业垂直应用场景，将数据驱动的智能计算融入网络之中，预期将在以下三个方向形成技术优势。

(1) 多样化协议支撑

超融合算力网络环境具备灵活、动态的多样化协议处理和计算能力，部署轻量级智联计算网元前端设备，解决协议不兼容，功能逻辑不一致等问题，支持各类型应用业务和终端的“即插即用”和“共同部署”。数据平面提供通用或专用的应用程序接口，同时将大规模定制服务基于行业场景进行封装，将智能终端、边缘设备及云计算服务器等实体进行灵活接入，通过物联网及智联计算网络的深度融合，实现对行业应用服务场景的异构海量数据的实时处理。超融合算力网络环境通过可定义的协议支撑能力，可以适应多种接入标准与技术的要求，支撑多样化的设备连接和服务需求。超融合算力网络环境具备网络感知能力，可以主动地感知各类接入终端，感知网络状态和用户需求等，再依据网络当前的系统资源智能地配置网络资源，按需提供服务。

(2) 网算存一体化控制

超融合算力网络环境支持数据处理过程中的存储计算与转发的一体操作，支持分级量化计算，优化处理海量内容的可靠传输以及高质量服务保证与网络基础设施所能提供的存储、计算和传输能力之间的关系。超融合算力网络环境数据层采用存储计算与转发一体协同技术，支持加速处理的软硬件协同处理，控制和调度物理资源以优化应用性能和系统性能。同

时，超融合算力网络环境能够协同调度各种网算存资源，通过应用驱动部署分级计算模型，计算任务调度数据归集传送，使得数据在网络的高效存储计算与转发。超融合算力网络环境将通信纳入计算的资源体系，突破计算的网络连接瓶颈，实现数据的按算归集、分级智能处理。网元设备同时具有“计算+连接”能力，实现计算调度与网络连接的精准、一致匹配。

(3) 服务功能智能编排

超融合算力网络环境以多样应用为核心驱动力增强网络感知能力，包括对用户行为、意图及业务的感知、分析及按需部署，实现网络运行、功能编排等自适应的承载，从覆盖范围、性能指标、业务场景等角度，动态自主调度网络资源，实现对用户的精准及个性化按需服务。将用户网络业务需求进行抽象化建模，依据业务基本参数与效益期望等指标对用户发起的业务进行详细规划，并依赖业务与服务之间的拟合关系实现业务需求到智慧化服务策略的映射，采用服务动态编排与业务自适应承载等机制，形成智慧化网络服务策略，并根据当前网络状态对已有决策进行迭代优化更新。按需服务的超融合算力网络环境可以无缝切换服务方式及内容，降低应用响应时延，简化应用的部署流程，实现高适应性和灵活敏捷的弹性网络。

4.4.2 应用部署愿景

随着物联网、车联网、VR/AR的蓬勃发展，互联网不断涌现新的业务形态，也不断产生出越来越多样化和个性化的网络需求。同时，伴随互联网与工业、能源、交通、医疗等传统领域的深度融合，网络的实时性、服务能力、服务等级等方面也面临一系列新的挑战。如图30所示，在超融合网络环境的基础上，通过网算存智一体编排和调度，面向多样化算网场景构建服务化、定制化的应用网络，并可拓展适配各类垂直行业需求，打造多样化算网应用生态。下面以智慧园区网络、垂直行业网络、数据中心网络三大典型算网融合应用场景为例，明确超融合算力网络环境的应用模式。

(1) 智慧园区网络场景

智慧园区作为一种信息化服务网络平台，网络是

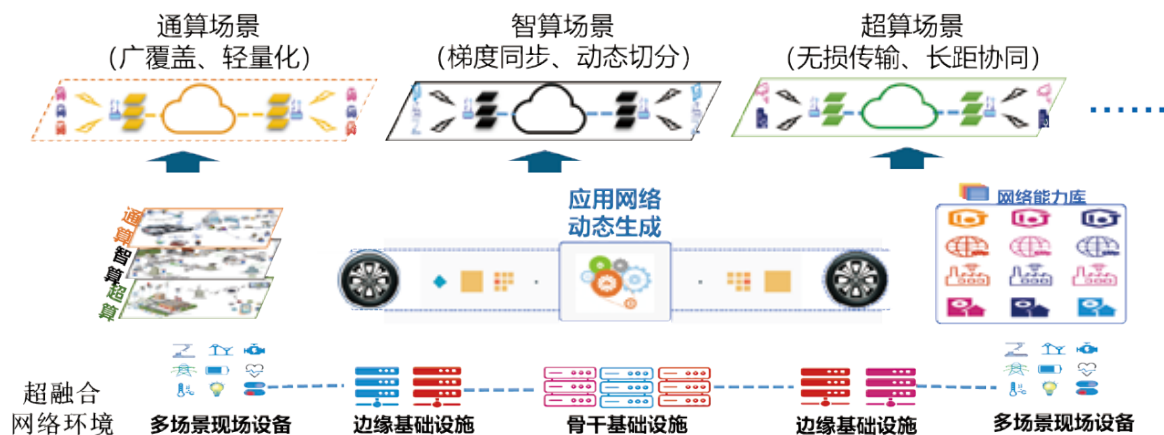


图30 超融合算力网络环境部署形态

建设智慧园区的基础。园区内相关资源和人员通过互联网、移动通信等信息技术实现有效地整合与衔接，促进园区内人、物间的协调发展。

在超融合算力网络环境“智慧园区”应用模式中，网络环境支持多样化边缘网络协议按需定制，实现园区网络柔性互联，有效提高园区资源配置效率，降低园区运营成本。通过部署轻量级网元前端设备，解决协议不兼容、功能逻辑不一致等问题，支持园区内各类检测业务和终端的“即插即用”和“共网部署”。在靠近业务现场的“边缘”实施网内计算、存储等处理操作，减少园区内业务处理时延，节省网络带宽，实现实时高效和低能耗及简化部署。超融合算力网络环境支持园区网络各类型终端的接入可控，基于零信任和边界安全架构实施网络安全隐私保护，满足园区对网络信息服务质量和能力的要求。

(2) 垂直行业网络场景

新型行业应用场景如沉浸式城市空间服务、精准医疗、智能工业物联网、智能电网等全新业务将广泛深入应用，进一步实现社会治理精准化、公共服务高效化、人民生活多样化。其中，智能电网在电网信息化加速发展的浪潮中已成为国家抢占未来低碳经济制高点的重要战略措施。

在超融合算力网络环境“智能电网”应用模式中，利用其计算与网络按需定义、逐级计算、分级部

署、多样化业务的整形和调度等特性对网络、存储及计算资源按需灵活部署，提升面向智能电网场景的计算效率及用户响应速度等性能。利用存储计算与转发一体协同调度，使边缘设备具备数据采集、处理、现场控制三大能力，将大大削减通过网络骨干部分上传的流量，避免数据上传下达所产生的时延弊端，提升网络处理能力。通过多样化异构终端灵活接入技术，提供通用或专用的应用程序接口，同时将大规模定制服务基于行业场景进行封装，将智能电网终端、边缘设备及云计算服务器等实体进行灵活接入，通过物联网及智联计算网络的深度融合，实现对智能电网场景中的异构海量数据的实时处理、同时提供分布式感知及控制能力，实现对网络的智能化配置、态势预测及安全控制。

(3) 数据中心网络场景

随着云计算数据中心的快速发展，云计算应用对数据中心网络延迟和带宽提出了越来越严格的要求，减少延迟，缓解拥塞就成为关键问题。然而传统数据中心网络节点只能完成数据交换功能，所有的计算任务都要在服务器上执行，对于All-reduce、参数聚合等复杂的集合运算而言，网络中需要传输大量的数据，这将增加延时和消耗更多的CPU周期，并且显著降低应用的可扩展性。

将超融合算力网络环境引入到数据中心网络中，

赋予交换节点计算和存储的能力，将数据规约卸载到交换节点，服务器将数据发送给交换节点，交换节点在网络中执行数据规约操作，再将规约结果返回给服务器，无论系统大小如何，网络延时都可以保持在很低且可预测的状态，不仅可以降低数据规约计算的整体延时，而且可以大幅降低网络中传输数据的规模，

有效缓解网络拥塞。此外，超融合算力网络环境可以充分利用网元节点“存、算、网”一体的优势，将分布式机器学习中的参数服务器的功能卸载到智联计算网元平台，利用交换节点高吞吐率来加速参数更新，对于增强学习等需要进行大量参数迭代的应用可以起到显著的加速效果。

AI计算应用系统内生安全

聚焦AI计算与内生安全的双向赋能机理，阐述安全、可信的AI应用体系设计及可量化设计、可验证度量的标准规范。内生安全构架为AI应用系统提供“安全基座”，化解软硬件系统漏洞后门等共性风险，克服大模型个性化的黑盒效应导致的AI算法个性风险，保障自动驾驶、智能医疗等关键场景高可靠、高可信。同时，可靠的AI应用系统作为内生安全架构构建过程的“便捷工具”，为构建异构性、动态性提供快速高效实现途径。

5.1 内生安全:先进计算范式的“免疫基因”

本小节系统阐述内生安全作为先进计算范式核心属性的本质特征，强调其通过架构设计构建具备自免疫、自修复、自验证能力的原生安全体系，如图31所示；强调区别于传统外挂式安全防护，内生安全从系统根源屏蔽漏洞后门等共性风险，为复杂计算环境下的功能正确性、数据完整性及服务持续性提供根本保障，成为衡量先进计算范式安全性的核心标尺。

5.1.1 内生安全计算架构

(1) 概念内涵:安全与系统的基因融合

内生安全 (Endogenous Security) 是一种在系统设计阶段即深度集成安全机制的理念，其核心是通过架构设计赋予系统“内生免疫力”，而非依赖外部附加防护。

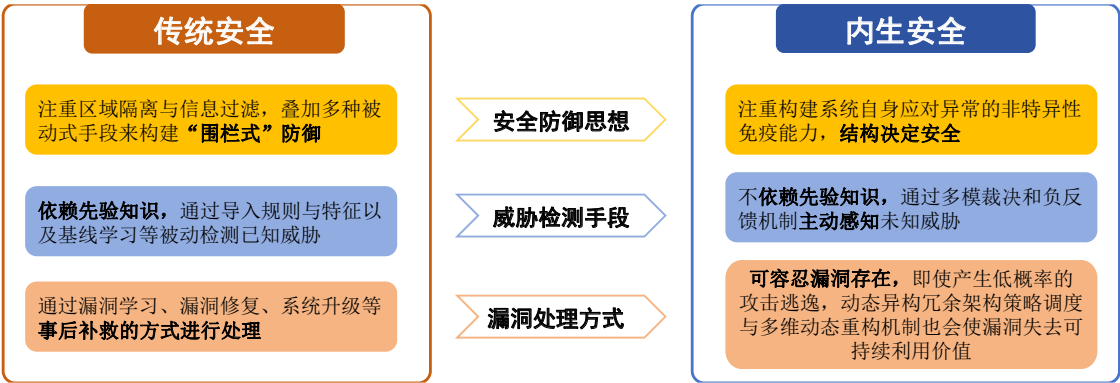


图31 传统安全与内生安全

定义：利用系统自身构造产生内源性安全效应，实现对未知威胁的主动防御。

核心思想：

构造决定安全：安全能力源于系统架构的固有属性(如异构性、冗余性)，而非事后修补；容忍未知风险：允许存在“已知/未知漏洞”，通过动态机制使漏洞难以被持续利用；全生命周期防护：覆盖数据存储、计算、传输全流程。

(2) 核心技术:动态异构冗余 (DHR) 架构

原理：部署多个功能等效但实现异构的执行体(如不同算法模型、硬件平台)，通过输入代理分发任务，由多模裁决器比对输出结果。正常输入输出一致；对抗样本引发“差模输出”，触发告警或丢弃结果。其核心流程如下：

输入代理分发：统一接收外部输入(如用户请求、数据样本)，通过负载均衡或随机策略将任务分发给多个异构执行体(如不同算法模型、硬件架构的计算节点)。

多模裁决器校验：正常场景：若各执行体输出结果一致(“共模输出”)，视为可信，输出最终结果。攻击场景：如构造的恶意输入导致执行体输出不一致(“差模输出”)，触发异常告警，拒绝可疑结果；关键逻辑：利用异构组件漏洞集合的非重叠性，使单一攻击难以突破所有执行体，从而暴露攻击行为。

内生安全效应：通过三重机制产生“结构免疫”效应。异构性防御：不同执行体存在非重叠漏洞，攻击者难以同时攻破所有组件(协同攻击难度指数级增长)；动态性迷雾：执行体组合的随机切换使攻击链无法持久维持；冗余性容错：多数裁决机制屏蔽局部故障，保障系统持续服务。

障，保障系统持续服务。

(3) 与传统安全架构的区别

传统架构依赖“边界防护+补丁修复”的外挂式防御，内生安全则将安全能力嵌入系统基因(表1)，通过结构设计(如动态异构冗余、拟态防御)构建主动免疫机制。在防御逻辑上，传统架构被动应对已知威胁，内生安全通过多维异构、动态变化、冗余裁决主动识别“差模异常”，可抵御未知攻击，从“亡羊补牢”转向“未雨绸缪”。

5.1.2 内生安全为先进计算提供安全基因

(1) 内生安全架构:从外挂防御到原生免疫的范式革新

传统安全架构依赖防火墙、入侵检测系统(IDS)、漏洞补丁等外部防护手段，形成“边界隔离-威胁检测-应急响应”的线性防御链条。这种模式存在两大根本性缺陷：其一，依赖已知威胁特征库，难以应对0day漏洞与未知攻击；其二，安全组件与系统内核分离，攻击者可通过供应链漏洞绕过防护体系。例如，2020年SolarWinds供应链攻击事件中，恶意代码通过合法软件更新通道渗透，致使传统安全防护完全失效。

内生安全则突破传统思维，将安全能力深度嵌入系统设计的基因中，构建“结构免疫”体系。以DHR架构为代表，其核心逻辑在于通过多维异构执行体(算法、硬件、软件)的动态协同与冗余裁决，将安全防护从“外部加装”转变为“内部生长”。当攻击发生时，异构执行体对恶意输入的响应差异(差模输出)会被系统自动识别，触发告警与自修复机制，使

表1 内生安全架构与传统安全架构的对比

维度	传统安全架构	内生安全架构
设计理念	外挂式防护(防火墙、入侵监测)	架构内生融合(安全即基础属性)
威胁应对	依赖特征库, 仅防御已知威胁	不依赖先验知识, 容忍未知漏洞
防护机制	静态规则、被动响应	动态异构冗余、主动裁决
漏洞处理	事后修补(打补丁)	动态重构使漏洞失效
信任模型	基于网络边界信任	基于异构运行空间交叉验证

系统具备类似生物免疫系统的“自我防御”能力。这种架构从根本上消除了“单一漏洞被重复利用”的风险，将安全防护从“被动堵漏”提升至“主动免疫”层面。

(2) DHR 架构:构建自免疫、自修复、自验证的原生安全体系

自免疫：异构冗余阻断攻击链。DHR架构通过四维异构技术(数据集、模型结构、训练方法、推理方法)构建多样化的执行体集合，不同执行体的漏洞空间呈现非重叠特性。攻击者若想突破系统，需同时攻克多个异构组件，协同攻击难度呈指数级增长。例如，在AI计算场景中，并行部署基于Transformer与CNN架构的异构模型，针对某一模型的对抗样本攻击难以对其他模型生效，攻击成功率从传统单模型的80%骤降至不足5%。这种“动态迷雾”效应使得攻击者难以形成持续有效的攻击路径，实现对未知威胁的主动免疫。

自修复：冗余裁决保障服务持续性。内生安全架构通过多数表决机制实现系统的自修复能力。当部分执行体遭受攻击或出现故障时，多模裁决器可通过对比异构执行体的输出结果，自动屏蔽异常组件，基于健康组件的输出维持服务连续性。在工业控制系统中，采用DHR架构的控制器可容忍单个节点的恶意篡改，通过三模冗余表决确保控制指令的正确性，避免因单点故障导致的生产线瘫痪。这种机制不仅提升了系统的安全性，更显著增强了其可靠性与容错能力。

自验证：差模检测实现动态安全审计。内生安全系统通过实时监测执行体输出的一致性，实现对计算结果的自验证。正常状态下，各执行体输出相同(共模输出)；当遭受攻击或出现异常时，差模输出会触发告警与追溯机制。例如，在区块链共识节点中，异构执行体对交易数据的验证结果差异可立即定位恶意节点，同时系统自动回溯攻击路径，生成安全审计日志。这种“以攻验防”的机制，使系统能够在运行过程中持续验证自身安全性，形成闭环的安全防护体系。

(3) 内生安全:先进计算安全的核心标尺

在量子计算、联邦学习、云计算等复杂场景下，

传统安全防护已无法满足先进计算对功能正确性、数据完整性及服务持续性的严苛要求。内生安全架构凭借其“从根源消除共性风险”的特性，成为衡量先进计算范式安全性的核心标准。

在量子计算领域，内生安全可抵御量子算法对传统加密体系的破解威胁，通过动态切换加密算法与硬件执行体，构建抗量子攻击的密码系统；在联邦学习场景中，DHR架构可防止数据投毒攻击，确保多方数据协同计算的安全性与隐私性；在云计算环境下，内生安全通过异构虚拟化层的动态调度，有效防御虚拟机逃逸、侧信道攻击等新型威胁。这些应用案例表明，内生安全已成为先进计算技术实现规模化商用的“安全底座”。

(4) 未来展望:内生安全驱动计算安全新生态

随着人工智能、物联网与先进计算技术的深度融合，安全威胁将呈现智能化、隐蔽化、跨领域协同的特征。内生安全架构作为一种“以变制变”的主动防御范式，将在以下方向持续演进：一是通过AI驱动的智能调度算法，进一步提升异构执行体的动态性与适应性；二是推动内生安全技术的标准化与模块化，降低在不同计算场景中的部署门槛；三是构建跨系统、跨领域的内生安全协同机制，形成多层次、立体化的安全防护网络。

综上所述，内生安全以DHR架构为核心，将安全能力从“外部防护”内化为“系统基因”，为先进计算构建起具备自免疫、自修复、自验证能力的原生安全体系。这种变革性的安全架构，不仅解决了传统防护体系的固有缺陷，更为未来计算技术的创新发展提供了坚实的安全保障，成为数字时代安全防护的必然选择。

5.2 AI驱动:内生安全架构的“智能进化引擎”

在数字化转型的浪潮中，人工智能技术正成为推动网络安全范式变革的核心动力。内生安全架构作为新一代主动防御体系，其核心优势在于通过异构冗余和动态重构机制实现对未知威胁的主动防御。然而，传统的内生安全实现方式在面对日益复杂的安全威胁

和系统规模时，存在着开发效率低下、动态响应能力不足等挑战。AI技术的深度融合为内生安全架构注入了“智能基因”，使其从静态的防御体系演进为具有自主学习、自适应进化能力的智能安全系统。

如图32所示，AI技术对内生安全架构的赋能体现在两个关键维度：一是通过智能化手段提升异构化实现的效率和质量，二是通过机器学习算法增强动态调度机制的智能决策能力。这种深度融合不仅解决了内生安全架构在工程实现中的技术难题，更重要的是赋予了系统持续进化的能力，使其能够在对抗中不断优化防御策略，真正实现了安全能力的内生性增长。

5.2.1 AI赋能内生安全异构化实现

异构化是内生安全架构的基石，通过构建功能等价但实现差异的多个执行体，形成对共模故障和恶意攻击的天然免疫力。然而，传统的异构化编程开发面临着开发周期长、成本高、质量难以保证等挑战。AI技术的引入为异构化实现开辟了全新路径。

(1) 代码智能生成与优化

AI技术首先在代码生成和优化层面展现出强大能力。基于大语言模型的代码生成技术能够理解高层次的功能需求描述，自动生成高质量的实现代码。通

过训练海量的优质代码库，AI系统掌握了各种编程范式、设计模式和最佳实践，能够根据具体的安全需求生成符合规范的代码实现。更重要的是，AI驱动的代码优化引擎能够自动识别代码中的性能瓶颈和潜在安全隐患。例如，通过程序分析技术识别可能导致缓冲区溢出的代码片段，自动插入边界检查；识别可能存在的竞态条件，自动添加同步机制；发现可能的信息泄露路径，自动实施数据脱敏处理。这种智能化的代码生成和优化不仅提高了开发效率，更从源头上提升了代码的安全性和可靠性。

(2) 多维度异构化实现

AI技术极大地简化了异构化实现的复杂度。在保证功能等价的前提下，AI系统能够自动生成基于不同编程语言、不同算法逻辑、不同实现架构的多个版本。通过深度学习技术，AI系统能够理解程序的语义本质，将其抽象为与具体实现无关的功能规约，然后基于这个规约生成多种不同的实现方案。在编程语言层面，同一功能可以分别用C/C++、Java、Python、Rust等不同语言实现，每种语言的内存管理机制、类型系统、异常处理方式都存在本质差异，这种差异性为攻击者制造了天然屏障。在算法逻辑层面，AI系统能够为同一计算任务选择不同的算法策

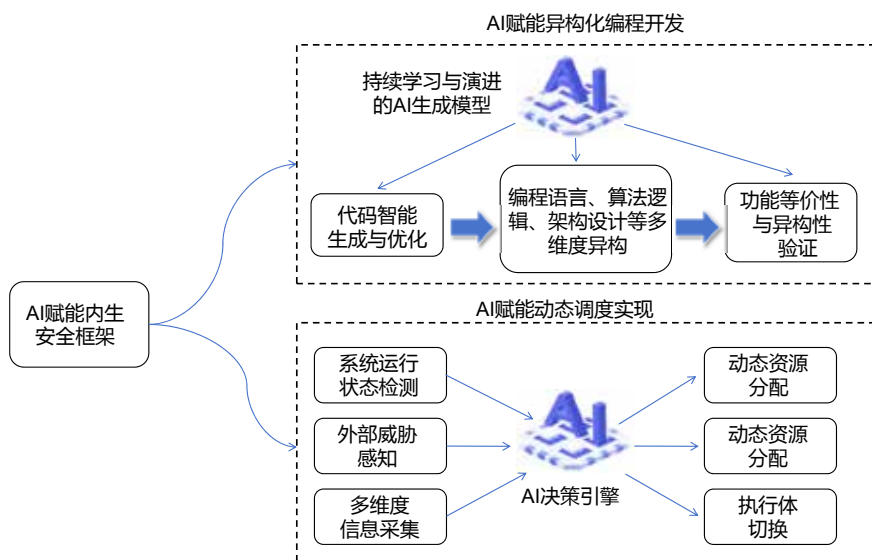


图32 AI技术赋能内生安全架构

略，如对于排序功能，可以选择快速排序、归并排序、堆排序等不同算法；对于查找功能，可以选择二分查找、哈希查找、B树查找等不同数据结构。在实现架构层面，可以采用面向对象、函数式编程、事件驱动等不同的编程范式，使用不同的设计模式如单例模式、工厂模式、观察者模式等，形成架构级别的异构性。

(3) 功能等价性与异构性验证评估

确保异构执行体的功能等价性是内生安全的关键要求。AI技术提供了强大的验证评估能力。基于形式化方法的AI验证系统能够自动证明不同实现版本的功能等价性。通过符号执行和模型检测技术，AI系统能够穷举所有可能的执行路径，验证在相同输入下各个异构版本是否产生相同的输出。同时，基于深度学习的语义相似度分析能够从更高层次评估代码的功能一致性。在异构性评估方面，AI系统通过多维度的差异性度量来量化异构程度。这包括静态特征差异（如代码结构、控制流图、数据流图的差异）、动态特征差异（如执行路径、内存访问模式、系统调用序列的差异）、以及攻击面差异（如不同版本对特定攻击手段的敏感程度差异）。通过构建综合的异构性评分模型，AI系统能够确保生成的异构体既保持功能等价，又具有足够的差异性来抵御共模攻击。

(4) 持续学习与智能进化

AI模型的持续训练是实现异构化智能进化的核心机制。通过构建专门的安全知识库和威胁情报系统，AI不断学习最新的攻击技术、漏洞模式和防御策略。基于对抗学习的训练框架，AI系统在虚拟环境中模拟攻击者行为，尝试突破现有的异构化防护，从而发现系统的薄弱环节。这种“自我攻击”的过程产生的数据被用于强化模型的防御能力。更重要的是，AI系统能够分析历史攻击数据，识别出哪些代码模式、实现方式更容易成为攻击目标。例如，通过分析大量的漏洞利用案例，AI系统发现某些特定的内存操作模式、特定的API调用序列更容易被攻击者利用，从而在生成异构体时对这些高风险部分进行重点异构化处理。通过联邦学习机制，部署在不同环境中的AI系统可以共享学习成果，形成集体智慧。这种持续学习能

力使得AI系统能够预测未来可能的攻击趋势，主动调整异构化策略，始终保持对新型威胁的防御优势。

5.2.2 AI赋能内生安全动态化机制

动态化是内生安全架构的灵魂，通过实时调整系统配置和资源分配，实现对威胁的主动响应和自适应防御。AI技术为动态化机制注入了智能决策能力，使系统能够基于复杂的环境感知和威胁分析做出最优的调度决策。

(1) 实时态势感知

AI赋能的实时感知能力是动态化机制的基础。通过部署分布式的智能监测节点，系统构建了全方位的感知网络。在系统运行状态监测方面，AI系统实时采集CPU利用率、内存占用、网络带宽、响应延迟等性能指标，通过时序分析模型识别系统运行的异常模式。在外部威胁感知方面，基于深度学习的异常检测算法能够从网络流量中识别出细微的攻击特征，包括零日攻击、APT攻击等高级威胁。系统通过多维度信息融合技术，综合分析用户行为模式、系统日志、安全事件等异构数据源，构建完整的安全态势图谱。时序预测模型能够捕捉攻击行为的演化特征，预测攻击者的下一步行动。通过构建多层次的威胁评估模型，AI系统能够准确评估当前面临的安全风险等级，为后续的动态调度决策提供精准的输入。

(2) 动态资源分配

基于实时感知的结果，AI系统通过智能化算法动态优化资源配置。系统将资源分配问题建模为约束优化问题，采用强化学习算法在计算资源、存储资源、网络资源等多维约束条件下寻找最优解。当检测到高风险威胁时，AI调度器会自动向关键防护节点倾斜资源，增加安全监测和防护模块的计算能力。对于低风险区域，系统会适当降低资源投入，实现资源的高效利用。AI系统还能预测未来的资源需求趋势，提前进行资源预留和调配。例如，在检测到DDoS攻击初期征兆时，系统会预先扩展网络带宽和计算资源，为即将到来的流量冲击做好准备。这种基于AI的动态资源分配机制确保了系统在面对各种威胁时都能保持最优的防御态势。

(3) 任务调度策略优化

AI技术使任务调度从静态规则驱动转变为智能自适应优化。系统通过深度强化学习构建任务调度模型，将调度决策过程建模为马尔可夫决策过程。AI调度器综合考虑任务优先级、执行时间、资源需求、安全风险等多个维度，动态生成最优的任务执行序列。基于历史数据的学习，系统能够识别不同类型任务的执行模式，预测任务完成时间，优化任务间的依赖关系。当面临安全威胁时，AI系统会自动调整任务优先级，确保关键安全任务得到优先执行。同时，系统采用多目标优化算法，在保证安全性的前提下，最大化系统吞吐量和响应速度。通过持续的效果评估和反馈学习，每次调度决策的结果都会被记录分析，成功的策略被强化，失败的案例用于改进模型，使调度策略在实践中不断进化优化。

(4) 执行体切换

执行体切换是内生安全动态防御的关键机制，AI技术赋予了这一机制智能化的决策能力。系统维护了一个异构执行体池，每个执行体都有其独特的实现特征和安全属性。AI决策引擎根据当前的威胁态势和系统状态，智能选择最适合的执行体组合。当检测到针对特定执行体的攻击时，系统会立即切换到其他异构版本，使攻击失效。切换决策不仅考虑安全因素，还综合评估性能影响、资源消耗等多个维度。AI系统通过分析历史攻击数据，学习不同执行体的防御效果，构建执行体安全性评分模型。基于预测的主动切换策略能够在攻击实际发生前完成执行体轮换，始终保持系统的不可预测性。多智能体协同机制使得分布式部署的内生安全节点能够协调执行体切换时机，避免全局同步切换带来的性能冲击，实现平滑的动态防御。

AI技术与内生安全架构的深度融合，标志着网络安全防御体系进入了智能化时代。通过AI赋能的异构化实现和动态化机制，内生安全架构不仅提升了防御能力和运行效率，更重要的是获得了持续进化的能力。这种“智能进化引擎”使得内生安全系统能够在与攻击者的持续对抗中不断成长，始终保持防御优势，为关键信息基础设施提供可靠的安全保障。

5.3 安全筑基:AI可信应用的“保障体系”

本小节重点介绍内生安全对AI可信应用的支撑作用，分析AI应用技术面临的风险，提出AI可信应用系统的构建方法，统性解决AI算法模型与运行底座存在的安全隐患，列举AI可行应用系统的实践场景，形成可复制的行业解决方案。

5.3.1 AI应用技术面临的风险

当前人工智能技术快速发展，对经济、社会治理和民众生活产生深远影响，为世界发展创造了巨大机遇。但与此同时，AI技术也带来了日趋严峻复杂的风险挑战。按照风险产生的源头，AI应用系统的风险可分为固有风险和应用风险两类。应用风险主要关注AI技术部署后可能引发的安全问题，与外部应用环境和政策法规密切相关，本小节将重点分析AI应用系统的固有风险。

AI应用系统的固有风险，是指在系统开发和运行过程中，源自程序代码、算法模型和环境交互等方面的内在风险。这类风险植根于系统内部的设计和实现环节，涵盖程序代码的逻辑缺陷、AI算法理论的固有局限、模型缺乏可解释性以及环境因素的干扰等多个层面。正因如此，这些安全问题属于系统的内生结构性问题，具有不可分割性，无法从根本上彻底消除，只能在持续的演进中寻求转化和化解。人工智能应用系统的个性化内生安全风险，特指由AI算法模型自身特性所引发的独特安全隐患。这类风险源于AI模型在数据处理、训练学习和推理决策各环节中固有的多重不确定性。典型的风险表现包括：对抗样本的客观存在使得攻击者能够通过精心构造的输入样本误导模型判断，从而暴露系统的鲁棒性缺陷；模型对训练数据的深度依赖性导致数据偏差、噪声等质量问题直接影响系统可靠性；此外还有模型泛化能力不足、决策边界模糊等由AI技术内在机制决定的风险特征。

然而，目前AI应用系统应对内生安全风险防护思路和技术方案，仍然受制于“尽力而为、力求完美”的传统观念，主要沿用“事后补救、经验积累”的经典防务理念。现有的对抗训练、漏洞挖掘、补丁

修复等层层累加的防护手段，在增强安全功能的过程中往往会引入新的内生安全隐患，因此无法从根本上打破“反复踩坑”的恶性循环。此外，传统防护策略需要以明确的威胁类型和行为模式为基础，面对具有广泛不确定性的安全威胁时显得力不从心。

5.3.2 AI可信应用系统构建方法

可信AI应用系统构造的重点不在于穷尽根除理论上不可避免的内生安全问题，而在于如何保障系统在复杂安全背景下仍可以稳定运行。因此，可信AI应用系统需要实现内生问题与系统功能的“和解”，从体制机制上管控或规避内生安全风险的威胁和破坏，保证系统在“有毒带菌”环境下主体功能的稳定，为可能出现的各种未知安全风险留有足够“安全余地”，这就要求可信AI应用系统应具有内生安全特性的构造。

内生安全特性的构造需具有以下性质或属性：(1) 具有开放性，允许架构内存在“已知的未知”或“未知的未知”内生安全共性问题；(2) 能以一体化形态获得体系化的安全增益；(3) 能将多样性、动态性和冗余性等防御要素或功能，以不可分割的一体

化方式赋予数字产品或信息物理系统内生安全的网络弹性；(4) 能从机理上有效瓦解任何形式的试错攻击或盲攻击，能为AI应用系统带来稳定鲁棒性与品质鲁棒性；(5) 可以自然地接纳已有或未来可能拥有的、以附加或内置或内嵌方式差异化 (或策略性) 部署的专业化的安全防护技术；(6) 对架构内存在的软硬件漏洞后门等广义功能安全问题的具体细节不敏感。

为此，具备内生安全特性的DHR架构可以有效赋能可信AI应用系统构建。从AI系统的数据、模型、训练和部署全生命周期出发，开展数据集异构、模型结构异构、训练方法异构以及应用推理异构等诸多研究，从而筑牢人工智能应用系统的安全“底座”，提供受信任的服务环境，具体思路方法如图33所示。

DHR架构的核心机理可以归纳为：(1) 将形式多样、种类繁多、不断发展的功能安全和网络安全 (乃至信息安全) 三重交织问题影响映射成动态异构冗余空间以差模或共模方式表达的安全事件；(2) 将工程实现复杂度高的全局静态异构冗余转变为具有技术经济性优势的局部动态异构冗余；(3) 将无法区别差模与有感共模的大数表决机制替换为基于策略性、多参量的迭代裁

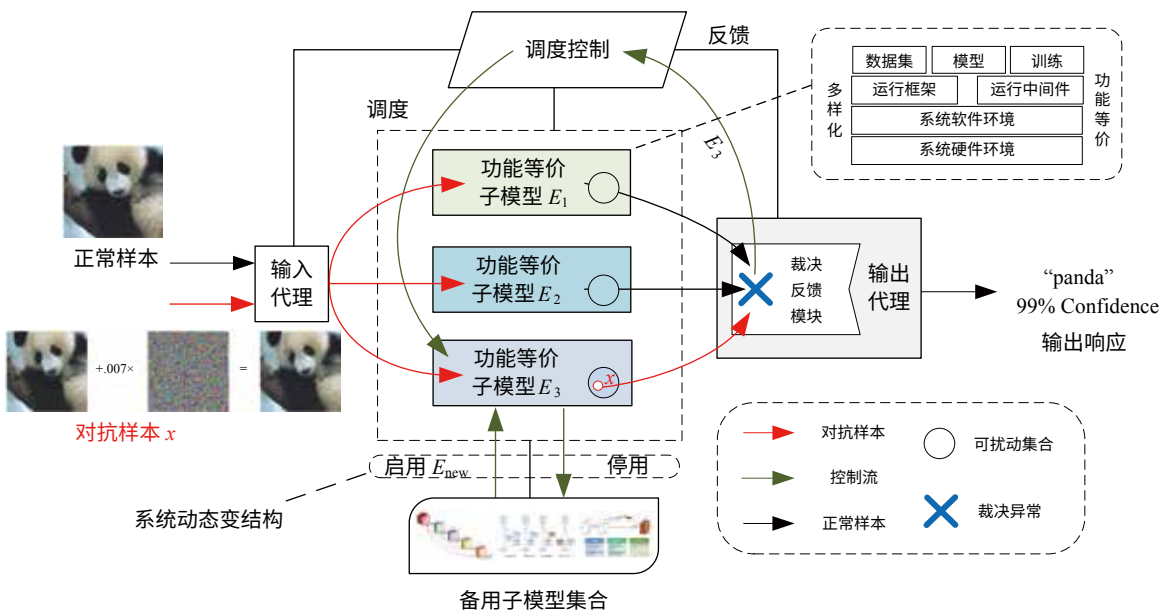


图33 基于DHR架构的AI可信应用系统框架

决；(4) 将“穷尽问题”或“打补丁”的处理机制转变为屏蔽纠错输出状态、问题场景时空迁移和安全矛盾性质变换等“兵来将挡水来土掩”的问题规避策略。

内生安全赋能人工智能应用系统构建，其基本内核是多模型的异构性。因此如何构建具有差异性的AI模型，是保障人工智能应用系统能够在“有毒带菌”环境下正常运行的关键，具体包括从数据集异构、模型结构异构、训练方法异构和推理方法异构等多个角度：

数据集异构。数据是AI模型构建的基础，且数据将直接影响AI模型算法的性能。数据集异构是指在保证模型任务一致的情况下使得每个模型学习到的数据特征呈现差异性，从而使得模型在面对对抗样本攻击时能够呈现鲁棒性。

模型结构异构。AI模型内部多由多层神经网络构成，这些层的神经网络内部结构多种多样并直接参与模型的前向推理与反向传播，是人工智能模型训练与推理的核心要素。AI模型结构异构是指在保证模型任务相同的情况下，使模型对输入信息的传播方式发生改变。

训练方法异构。AI模型训练过程本质上是一个基于最优化思想的拟合过程，训练方法的设计决定了模型最终的收敛方向和模型性能特征。AI模型优化目标是通过不同的损失函数来进行定义的，训练方法异构的核心在于通过定义不同的损失角度将使模型在训练过程中收敛于不同的特征角度。

应用推理异构。当AI模型训练完成并进行部署时，应用推理异构方法是一种有效的安全策略，它通过结合不同类型的人工智能推理模型，来提高系统的整体性能和鲁棒性。通过应用推理异构，可以在保持高精度和泛化能力的同时，增强系统对异常情况的识别和处理能力，减少单一模型可能存在的偏差和漏洞，从而在面对复杂多变的实际应用场景时，能够更加稳健地应对各种安全威胁和挑战，确保人工智能系统的内生安全性和可靠性。

5.3.3 AI可信应用系统实践场景

内生安全理论通过将安全防护深度融入人工智能应用系统核心架构，在人脸识别、智能驾驶、低空经济等领域构建起自适应防御体系，推动智能应用实现

安全性与可靠性的双重突破。

(1) 内生安全赋能人脸识别系统

在人脸识别应用场景中如图34所示，内生安全理论从根本上重构了系统的防护逻辑。传统人脸识别技术往往采用外挂式安全模块，容易受到深度伪造对抗、面具攻击等威胁的冲击。内生安全框架的人脸识别系统将模型异构深度融入到特征提取和比对过程中，形成人脸特征差异化描述，显著提升了系统对抗恶意伪造的能力。

在实际应用中，紫金山实验室首次研制了具备内生安全属性的人脸识别攻防验证平台。平台包含人脸检测与分析、人脸五官定位、人脸对比、人脸搜索等基础功能，支持人脸混淆、人脸伪装等多种攻击方式，支持开展多维度的模型鲁棒性评估，持续集成DHR拟态防御方法，通过构造攻防对抗场景，比较传统方式与拟态DHR构造的防御能力，验证拟态DHR构造的防御效果。实验证明，内生安全赋能的人脸识别系统具备更为稳健的人脸识别效果，能够有效抵御多种对抗攻击威胁。

(2) 内生安全赋能智能驾驶系统

智能驾驶内生安全的核心挑战源于功能安全与网络安全的深度交织风险如图35所示。传统架构将车辆电子电气系统与网联功能割裂防护，导致功能层控制逻辑缺陷与网络层漏洞后门形成复合攻击面。具体表现为：自动驾驶决策算法易受传感器欺骗干扰，车联网通信协议暴露于中间人攻击风险，而传统安全机制在资源受限的车规级系统中难以平衡防御强度与实时性要求。更关键的是，安全防护与车辆控制系统的紧耦合缺失，使得网络攻击可直接穿透至车辆执行机构，引发系统性失效风险，形成“攻防机制割裂、安全能力碎片化”的典型内生安全隐患。

内生安全理论通过动态异构冗余架构破解上述困局，开创功能安全与网络安全一体化增强的创新路径。基于广义鲁棒控制机制，将网络攻击建模为广义扰动，构建非对称执行体组合与精准裁决体系：在感知层部署多模态传感器交叉验证，决策层采用近似共识算法实现异构AI模型的同步裁决，通信层通过轻量化拟态防御构造抗干扰信道。这种架构突破性地实现

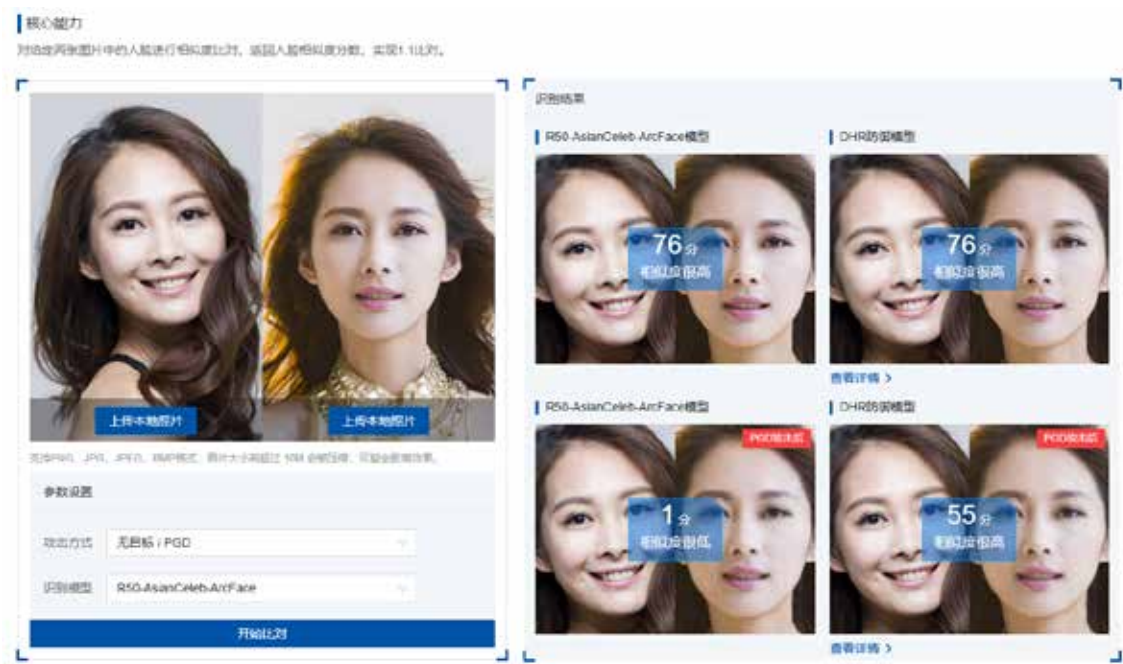


图34 内生安全人脸识别系统



图35 L3级、L4级内生安全试验车

安全防护与车辆控制的原生融合，形成”漏洞不相关、攻击难协同”的防御优势。其创新价值体现在三方面：首次建立功能-网络安全联合分析模型，攻克安全需求分解的技术瓶颈；通过动态冗余机制在有限资源下实现安全效能指数级提升；创造性地将车规级系统的可靠性转化为网络安全韧性。这不仅破解了传统防护”打补丁”式滞后性困局，更为智能驾驶系统构建了具备自主进化能力的免疫体系，为我国汽车产业

突破技术封锁、重构安全范式提供战略支点。

(3) 内生安全赋能低空经济

低空经济的内生安全问题源于低空智联网作为复杂信息物理系统的固有脆弱性如图36所示。无人机端侧存在开源软件漏洞、无线通信易受攻击等缺陷；网侧因通信制式多样、协议不统一导致数据链路开放易受干扰；云侧面临异常管控难、AI算法不可解释等挑战。同时，外部复杂环境中的电磁干扰、网电攻击、

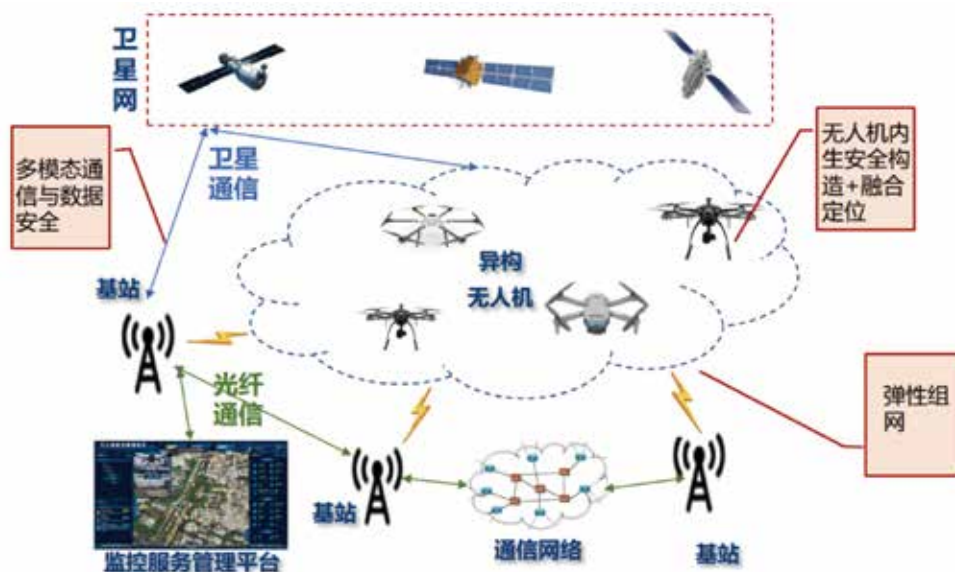


图36 低空智联网内生安全关键技术

样本攻击等威胁，与系统自身漏洞叠加，可能引发无人机失控、敏感数据泄露等严重后果，威胁人身财产安全乃至国家安全。

内生安全理论通过动态异构冗余 (DHR) 架构提供创新解决方案。其核心路径包括：硬件层面部署异构处理器与冗余模块，结合故障诊断实时容错；软件层面运行多版本飞控程序，遇攻击时动态切换等价功能模块；通信层面融合卫星、5G/6G及自组网链路，智能

择优保障连接；网络层面构建弹性拓扑，实现攻击自愈与资源动态调配。创新点在于将“拟态防御”思想融入系统基因，通过异构性、冗余性和动态性，无需依赖先验知识即可防御未知攻击。优势在于打破传统被动防护局限，形成“预防-防御-恢复-适应”的主动免疫闭环，轻量化实现端到端纵深防御，并兼容传统安全技术获得非线性防御增益，为低空经济提供根本性安全范式转变。

智能调度高效柔性负荷微网能源系统

随着人工智能应用爆发式增长，算力需求迅猛增加，已成为推动产业转型发展新的关键生产力。算力与电力的强耦合关系，决定了电力可靠且充足的供应成为制约算力设施建设、推动产业数智化转型的关键因素。目前算力支撑主要依赖传统能源，多数对能耗、低碳化、绿色化发展未能足够重视，且由于新能源发电具有波动性、随机性、间歇性、时空错配等特性，无法如传统电源般稳定提取利用，导致AI算力与低碳绿色电能协同受制，无法产生绿色高效的新质算力。

本章节针对新能源与高效AI算力协同，提出基于高效柔性负荷微网系统 (Efficient Flexible Load Microgrid system, EFLM) 的算电一体、协同发展解决路径；以能量时间片复用，多源多荷柔性连接方式，调和并优化新能源供给，解决新能源高波动、高随机特性下源荷适配难题，为智能计算与可再生能源结合以及新能源高效利用提供新路径，通过EFLM优化算电协同，达成与人工智能、先进计算双向赋能、可持续发展的共同目标。

6.1 AI算力激增与新能源波动性挑战

计算是信息通过能源转化及应用的过程，能源是算力的主要生产资料，能源供给制约算力的发展。自从2012年卷积神经网络 (CNN) 等模型在图像识别、语音合成等领域实现商业化应用，开启了人工智能的十年黄金发展期，至2022年起，OpenAI引领的全球通用人工智能 (AGI) 的发展，触发了AI算力的迅猛扩张，2025年，DeepSeek在全球AI领域掀起巨浪，而作为支撑数字经济发展的基石，算力基础设施投资也迎来了新一轮浪潮。近年来，我国AI算力需求呈指数级增长；2024年，我国智能算力规模达725.3百亿亿次/秒 (EFLOPS)，同比增长74.1%，增幅是同期通用算力增幅 (20.6%) 的3倍以上 (IDC，中国人工智能算力发展评估报告)。

AI算力的崛起催生旺盛的电力需求，目前AI算力依赖的主流GPU服务器的能耗为CPU服务器的4倍；GPT-3大模型单次训练的耗电量为128.7万千瓦时，相当于430个家庭一整年的生活用电量，而GPT-4更高达42.4吉瓦时，为前一代的数十倍。因此美国人工智能专家不断表达对算力背后电力需求的担忧；OpenAI的创始人奥特曼于2024年达沃斯论坛即警告下一代生成式AI系统消耗的电能将远远超过预期，能源系统将难以应对；黄仁勋2024年于斯坦福大学讲亦宣告 “我们不能只想着算力，如果只考虑计算机，我们需要烧掉14个地球的能源。超级AI将成为电力需求的无底洞”。各国国情差异下，针对未来算力与能源应对措施亦不相同。国外企业大多各自投资及合作配套能源项目，如OpenAI于2025年公布的Stargate超级智算中心项目预估15吉瓦能耗，并且计划以小型核电站与光储一体新能源部署供应能源。

我国由于能源种类及地理分布上存在先天匹配差距，资源“富煤贫油少气”，并且水电及可再生能源大量集中西部省份，而东部地区因经济发展迅速，对能源的需求量大，因此开启了自1996年以来的“西电东送”工程；而由于智能算力的大幅增长，西部的丰沛能源供应造就了2021年国家发展改革委明确的“东数西算”发展，打造围绕能源供给充裕地区的国家算力节点；但由于AI算力增速及规模效应巨大，传统能源未来将无法应对，且考虑国家AI算力发展要在3060碳达峰目标的约束下追赶美国，算力与电力问题需要重估；国家发展改革委提出《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》，要求2025年底前国家枢纽节点新建数据中心绿电占比超过80%，并且将数据中心与钢铁、电解铝、水泥等八大传统高耗能行业一体纳入节能减碳的重点行业。

算力绿色化不仅为我国，亦为世界趋势；谷歌承诺到2030年，其全球所有智算中心和园区将实现24/7无碳能源运营，并通过购买无碳能源、优化能源数据和使用，开发下一代能源技术以及微电网等主要举措来实现。然而其2023数据中心的碳排放量仍然同比增加了13%。以上非预期增量并非新能源发电总量无法应对算力需求，原因主要在于算力高度集中下新能源供给时空失配，以及传统新能源利用方式与新型电力系统发展无法应对AI算力快速迭代的倍增需求的结果。

我国除了算力绿色低碳挑战外，美国主导的

芯片“硅幕”造成芯片工艺与能耗的代差，需要更高的电力供给总量及效率；华为2025年5月发布的CloudMatrix 384超节点比肩英伟达NV72，但能耗为其2.6倍；我国于此发展制约条件下需要比其他国家更优化的芯片工艺维度，包含芯片集群效率及能源利用。

综合以上算力绿色化发展趋势以及AI算力迅猛发展导致传统能源无法应对快速的能源需求增量，结合我国能源战略及特殊时空条件，以可再生能源应对未来AI算力电能需求为必然发展路径，正所谓“AI的尽头是光伏板”，如果新能源能够有效利用于算力，AI所带来的海量计算，将与可再生能源一起，推动新一轮信息与能源革命。

然而新能源作为电源，并不如传统电源般具备高质量且稳定的特性(图37)；新能源发电因其固有机理具备高波动与高随机性特征，并且由于发电方式形成的低系统惯量，导致其难以提供稳定出力，其日出力曲线对比传统火电如下图。由于缺乏维持供电质量的鲁棒控制能力，无法长时期独立供电，常常会因为全时段供电质量问题而无法被有效利用。而电力体系为将新能源纳入利用，现行技术路径通过组织微电网，将域内新能源通过电力电子变换设备调制成为可接入条件(电压，频率及相位)，并且通过公共连接点(Point of Common Coupling, PCC)并联接入大电网，同时通过储能平滑新能源波动及时空差异，保证供电质量。

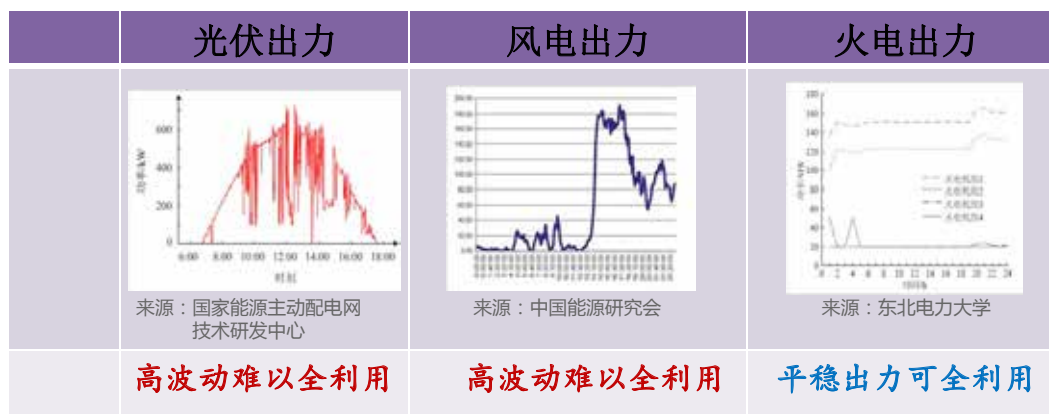


图37 传统能源出力与新能源对比

以微电网为核心的新能源接入模式存在技术路径上的挑战；首先，微电网如何面对动态范围过大的变化源荷，实现多时间尺度、多粒度源荷自适应匹配、降低电压功率调控损耗、提高新能源的直接利用率及“应发尽发”能力；其次，储电容量如何兼顾经济性与供电质量的平衡，并根据域内电源特性、负荷需求和外部环境条件，设计合理的能量管理策略，实现能源经济高效利用；再者，如何在大规模新能源接入与分布式连接、复杂潮流控制和技术经济性条件下，确保自身运行效率与安全性，以及与电网供电之间的平衡关系，并一体化地实现物理安全和网络安全，提高微电网的经济效益和市场竞争能力。

以上新能源利用及微电网技术路径的挑战，其根源为新能源的发电机理相较传统电源，为本质上的变化。其多时间尺度的变化性与随机性，与百年来电力输配体系中电源稳定可调的前提条件相悖。于前提条件发生变化，需要考虑路径依赖是否已然成为发展的制约条件；诸如基于共同接入点的交直流母线形态、基于脉宽调制的电力电子调频调压、基于开关技术的升降压环节、基于复杂协同的潮流控制、基于互联网模式的分布式控制、用能经济性与电网价格政策强相关等等，以上问题无法依靠单一的技术路径，抑或调整生产关系如新能源配置、使用方式、利用时间，抑或加大投入等非技术措施能够解决之。

因此本章节着重提出构建“能网一体计算环境”的必要性，针对新能源的发电机理及多时间尺度的变化特性，提出合适的技术路径，动态且智能的匹配与

算力算网之间的源荷关系，提高全时段的新能源利用率；通过绿色、便宜、安全的新能源，实现可持续算电协同发展。

6.2 EFLM解决算力与多能电源时空耦合问题

新能源利用的工程技术问题，依照广义特勒根定理，在一个互联系统，其内部存在能量守恒关系，亦即在任一时间片段，系统耦合能量与各节点动态能量之合为零 (如图38)。

而新能源具有高波动，高随机特性；大量非稳新能源接入网侧造成系统耦合破缺，即电网链路无法平衡节点的能量波动；加大耦合投入，例如增加储能投入或无功补偿装置，扩充末端线路容量等，皆需一定建设与投运周期，且损害经济性；于此状况下必须舍弃一部分节点耦合以维持能量平衡，因此造成弃风、弃光等电能无法实时利用状况，导致利用率下降，影响技术经济性；其推演如图39所示。

上述耦合破缺关系的缓解方式，除了加大网侧耦合投入外，另一种解决路径为节点本地实时自平衡；电能具备光速传输特性，域内源荷如无法直接平衡，则需寻求网侧耦合转移，如无法于节点外部耦合，则需将其转为其他能量，寻求自然消耗，造成发电利用率下降；因此在网侧耦合条件不具备状况下，源荷于节点内做到实时平衡为新能源直接利用率提升的充要条件。

目前新能源依循的微网系统基于共同接入点的功

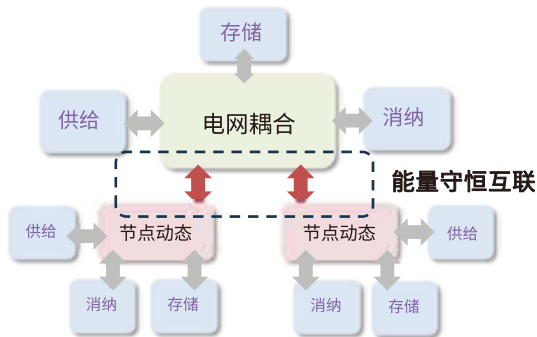


图38 能量守恒互联系统



图39 传统能源出力与新能源对比

率平衡为空分非稳关系，其满足供需平衡约束的充要条件为多路电源功率通过电力电子变换及母线分配，恒等于各负载的功率需求，以矩阵表达即为 $[L]=[C] \times [P]$ ，其中 L 为负载功率矩阵， P 为新能源(含网侧功率)出力矩阵， C 为源侧接入PCC的出力配平(含损耗)矩阵；而于传统微电网条件下， C 的取得为计算-迭代过程，源荷波动条件下，大规模网络无法实时求解，造成电源侧和负荷侧时间与空间尺度失配问题。以下垂控制为例，此过程为依据给定变换器各自的下垂曲线，进行对电压与电流的协同控制，但 $[C]$ 与在各自出力条件下系统等效阻抗有直接关系，在高波动电源下，与量测及其他相关电源均互为变量，于追求较小的电压偏差和较高的分流精度时，存在着固有矛盾，导致无法实时求解，造成电源侧和负荷侧时间与空间尺度失配问题。

由于时空尺度失配，新能源的本质波动问题无法通过以上模型得到完整源荷匹配解决，需要对 $[C]$ 的取得过程做简化，将出力变化做为可实时适配的系统内禀属性。本章节提出的EFLM (Efficient Flexible Load Microgrid), 为新能源的源荷实时匹配提供可行路径，其原理为将加入基于时间片的开关矩阵 $[S]$ ，并借此将 $[C]$ 简化为单元对角矩阵 $[I]$ ，亦即出力不需要经过潮流控制，各源间为正交关系；同时将单一源与转换矩阵 $[S]$ 之间的分配系数构成 $[C']$ ，得到 $[L]=[C'] \times [P] \times [S]$ ，(其中 C' 为简化出力平衡矩阵， S 为基于时间片的开关矩阵)实现时分可重构动态源荷连接，将多源多荷平衡问题，解耦成多个一源多荷问题；由于通过 S 矩阵转换为单向潮流，消除了各源间平衡需求，使求解成为线性问题(单纯的功率叠加关系，没有系统潮流关系)，同时降低源荷间因空分复用造成的交互影响及损耗；如图40为原有微电网物理并联，通过 S 矩阵维度转换及解耦，形成可重构动态连接，实时适配源荷动态。

以上机理推演出EFLM体系化实践架构，核心为以电力电子开关器件(如IGBT, IGCT, 碳化硅等)建构可重构连接的开关矩阵TMFC (Time division Multiplexing Flexible Connection)，执行上述 $[S]$ 矩阵的时分连接能力；总体设备与控制集成于敏捷电能分配器 (Clever Power Distributor, CPD)，耦合控

制与信息体系，进行源荷基于能量时间片的调度(如图41)。

由传统微电网并联调节形态，演变为EFLM时分能量调度机制，应用于AI算力作为用能方，对比于传统工业恒定不可调节负载，更能体现EFLM源荷实时智能适配优势。AI算力与其基础设施，作为新能源的消纳及负荷方，其特性为高质量且集中的电力需求，且各使用阶段(训练，推理，调优等)及内容(图形，视频，文本等)与能耗亦为强关联；例如于通用大模型训练阶段，由于算力集群于本地执行大量并行运算，造成单点的大量电力消耗；EFLM的本质为基于能量时间片的实时优化能源调度体系，通过电能时间片优化调度，结合各AI算力的任务及阶段需求，提供依照功率需求的高质量的电能供应，而非依循传统微电网额定功率齐头式供能，使得能源调度利用更优化且经济；并通过透明电能传送，及多尺度、多粒度源荷平衡消纳机制，依照负荷实时条件尽量匹配新能源，使新能源发电“应用尽用”，于保证AI算力能源供应前提下，得



图40 由物理并联转化为可重构时分动态连接

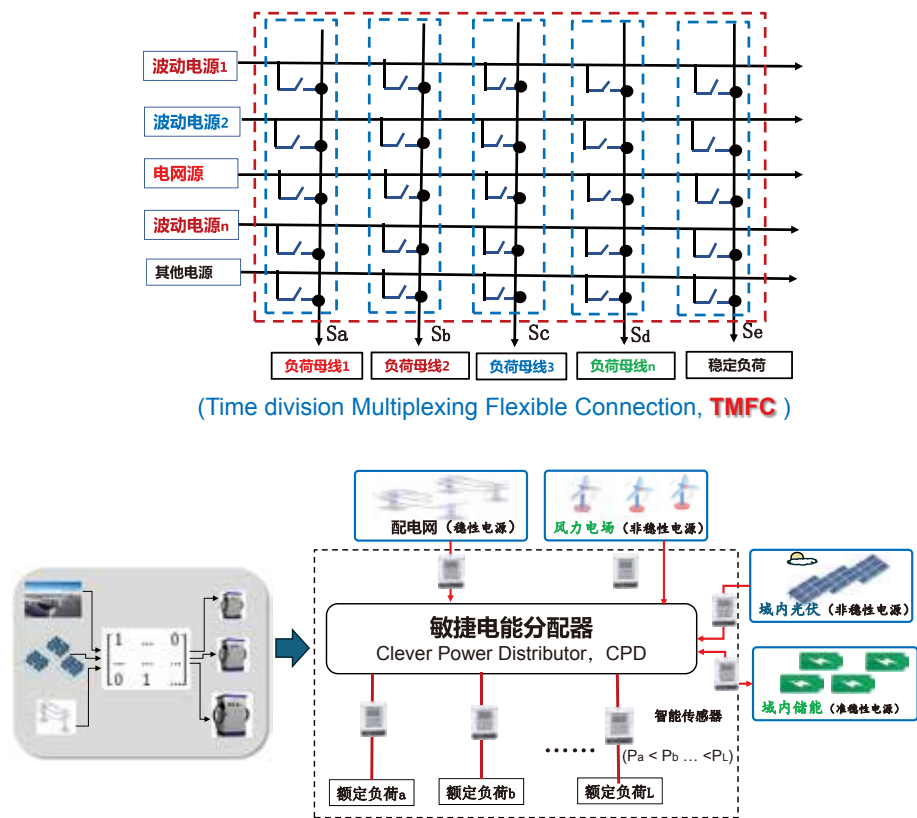


图41 由TMFC开关矩阵构成的敏捷能量分配器

到安全，便宜且绿色的电能。

其次于技术经济性考量及算电时空尺度方面，新能源具有时空特殊性，算力亦有其区位及时空要求，能算于时空重叠非常态现象，但算力无法失去能源支撑，时空不匹配亟须解决手段：依照AI算力匹配能源基础设施具有技术经济性风险，电力基础设施的生命周期为数十年，新能源光伏年限约25年，储能则为7年，而基于GPU的AI算力中心依照目前的迭代速度生命周期仅有3年，相差数倍至十倍的年限无法依照新能源基础建设投入构建具备经济意义的回收周期；且大模型训练周期最长数月，并不如工业生产(如冶炼行业或化工)般持续数十年以上，如何平衡短周期用能，并向长周期过渡需要同时具备技术经济性与合理时空配置的能源供给手段。EFLM基于柔性连接，可适配功率及连接形态不确定的对象，如能源车，电氢转换设备，储热及热交换等各种转换及抵近应用；以上转换

对接运输及能源配送机制，不仅可于新能源发电侧与智算中心侧提供灵活能源接口，同时利用交通线路及能源载体，对中短期集中高耗能的AI算力基础设施形成多条虚拟能源专线，提供绿电抵近服务，降低电力基础设施投入，并符合技术经济性要求，解决AI算力用能时空匹配问题。

此外，垂直行业大模型崛起，带有行业属性的大模型的要求更专业且需要避免通用大模型带来的AI幻觉；为达到与行业及领域专业知识深度结合的目的，需要结合通用模型，针对专业领域进行二次预训练，及后续微调。以上面向领域的AI训练为短时(可能数周)高强度的分布式AI算力需求，于推理阶段则可以低强度算力替代。而领域专用大模型用户多半为具有核心知识产权的企业用户，不合适将其核心知识于通用模型平台上分享；因此垂直行业大模型的应用方式，为短时且高强度的AI算力，结合通用模型与算力的抵

近服务，并配合本地私有核心数据进行二次训练，并将专用模型数据留存本地以供后续推理及本地应用；而基于EFLM柔性连接提供的能源专线，则可扩展为能算一体的AI算电抵近服务，延伸绿电的可及性，并同时消除AI算力及能源时空失配问题，一体化解决未来因应各行业需求，庞大数量的垂直领域大模型面对短时训练算电资源不足的困境。

6.3 AI驱动构建EFLM“动态优化大脑”

EFLM系统本质上是针对新能源本地利用的优化与精进技术；传统微电网针对新能源的调控为基于运行脚本的粗粒度控制，根据运行优化原则，并参考本

地天气预测及运营综合数据得到日前小时级规划，并且依照当日变化作小时至十五分钟调整，各电源则依靠电力电子变换器作母线并联前作出力调节；由于其调控机理与源荷变化的秒级至毫秒级时间尺度无法匹配，无法保证新能源是否“应发尽发”，负荷是否将新能源“应用尽用”。

EFLM对域内新能源作精细化管理，扩大出力范围并提高新能源利用率；与传统微电网比较，EFLM精进之处在于将能量分配及控制依照功率时间片形成可精细管理对象，并且准确做到“能量-信息-控制”三流合一，以此对应的云网一体三级调度控制体系为秒级至毫秒级精准控制提供手段(如图42所示)。理论上控制时间精度愈细，能量时间切片愈窄，可利用能量边

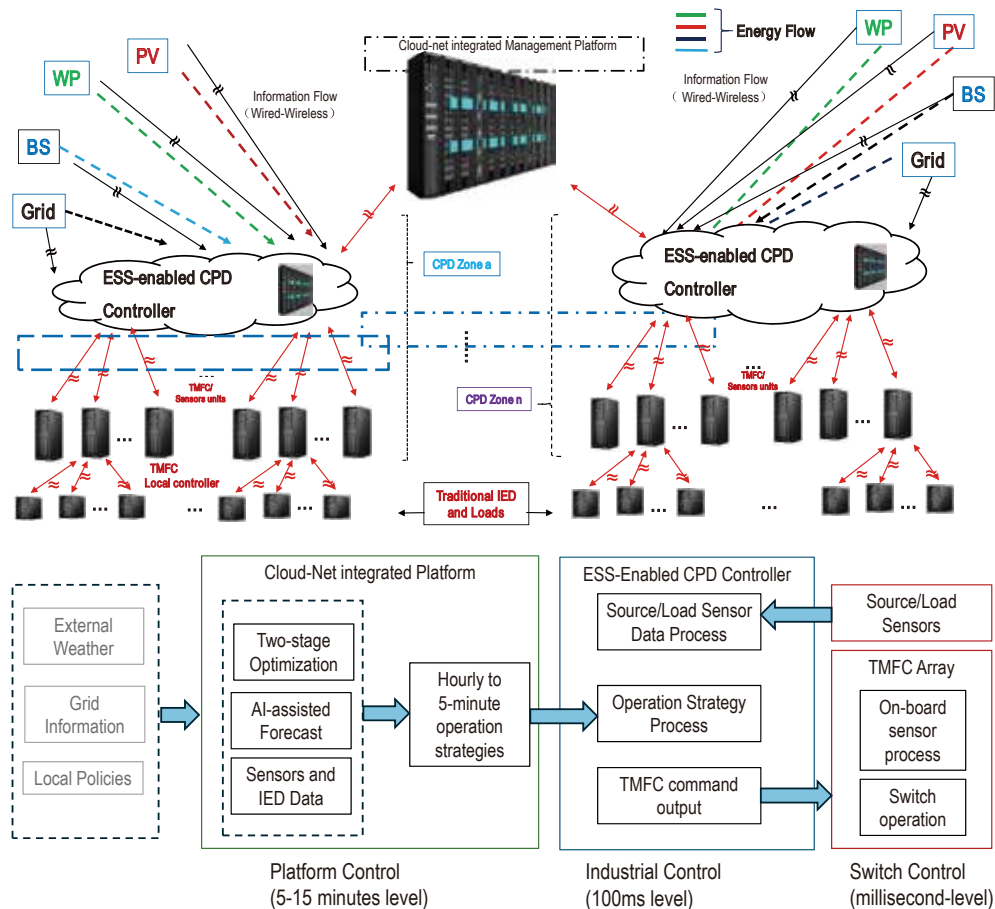


图42 EFLM云网一体三级调度控制体系

界愈宽，新能源利用率愈趋近理论最大化边界。

上述秒级至毫秒级调度需要强大的调控模式及信息体系；首先，EFLM的可控域可延伸至上百至上千组本地控制器，及至多上万组高功率IGBT开关组成的协同阵列，其源荷适配各自为多目标（绿能优先，电价，成本，损耗，电能质量等）离散优化的结果，并且需针对各自的前端传感器及变化率作相应处理及功率预测，在考虑协同及整体系统稳定性边界条件下，机理为上万个短时模型预测控制（MPC，Model Predictive Control），并且于各自目标中于总体系统侧优化及协同。传统微电网优化控制已证明为NP-Hard问题，无法于有限时间求得最佳解，故需要应用AI中启发式及神经网络算法优化，而EFLM的实时控制体系时间尺度及物理单元皆数量级提升，需要更大规模AI算力；同时系统需要具备云边AI算力协同机制，将区域中千级边端AI协同，类似运控中心与千辆自驾车的关系，由云到端形成AI算力为EFLM赋能的“动态优化大脑”。

此外，新能源出力与气象强相关，而气象大模型自2018年ECMWF的初步尝试，至最近国内发布的伏羲、风清、盘古等更在实际应用中展现出广阔前景。EFLM基于动态源荷调节的体系与传统新能源发电调控机理相异，对预测的需求亦不相同。传统新能源发电对短时天气变化需求不大，因其调控时空尺度不及应对，反之EFLM需求本地，短时且细空间粒度（分钟级-百米立方）的气象大数据，至于长周期预测则可用于大范围运行优化，此与传统微电网需求一致，但其与EFLM调控体系非为耦合关系，而属于长时策略的一部分。AI算力可赋能本地的短时-细空间粒度预测，并依照本地历史数据，结合大模型训练，提供更精准及个性化本地预测，借以修正本地短时传感及气象数据的外推误差，使控制更贴近新能源实际发电变化。

6.4 能算结合实现AI可持续发展

AI算力与能源体系的相互发展关系，不仅于其迅速增长的能源需求，或我国能源地理分布限制下的“东数西算”大尺度区域调节，而在于考虑于全天候无碳能源（CFE）的终极目标下，如何能够创造新能源

与AI算力的源荷互动场景，让新能源发电可靠可用，让智算中心用能可测可控，推进算力与新能源深度结合，才能建构可持续的能算关系，并进一步利用AI算力，协助发展新一代的新能源调控技术，加速电力电子变革，服务未来新能源抵近服务及交易模式创新，使带来新能源挑战的AI算力，转化为新能源利用方式变革的推进力量，创造新能源与AI算力源荷互动，能算结合新型范式。此一范式变革需要以EFLM作为催化剂，于AI算力用能侧，透过EFLM能量时间切片提供解决新能源随机波动的最佳应对方式，作为能源与AI算力的源荷互动基础；于新能源调控技术变革方面，EFLM利用柔性时分复接构网技术，将能量流与信息控制结合，于“能源-信息-控制”三位一体的精确调度，则需要借重AI算力进行秒级海量计算，创造以AI算力引领驱动EFLM精进，进入精细化新能源调控新境界。

以上结合关系可总结为能算一体化发展循环（如图43），要推进能算一体的发展目标，首先需针对AI算力扩展需求的三个主要维度，即单点算力增加，算力基础设施整体扩展扩容，及算力应用边界扩展，分析发展过程面临的挑战；单点算力扩展主要直面能源供给支撑问题，如前文所述，“硅幕”导致人为工艺及能效代差，需要更大功率支撑；AI算力基础设施扩展，除区域能源体系支撑，需要面对“双碳”及绿电比例

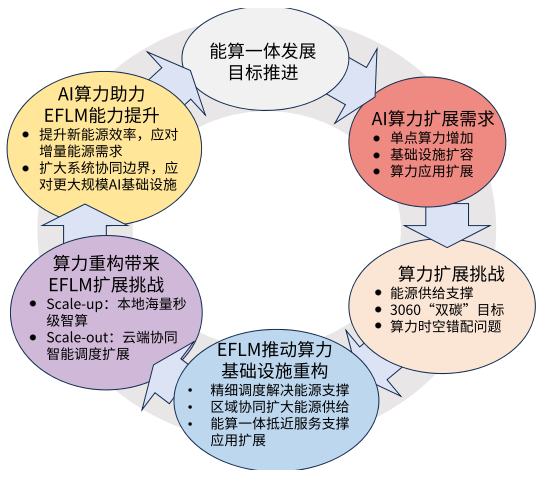


图43 能算一体化发展循环

要求；算力应用边界扩展，则需要深入分布式AI算力及垂直领域大模型广泛应用的数量级增长压力，与带来的算力-电力时空错配问题。

EFLM催化下的能算结合推动算力基础设施重构，利用精细化调控，解决能源供给支撑问题，面对3060“双碳”及绿色算力目标，提供新能源尽发尽用的路径，并通过区域协同调度扩大能源供给边界，同时提出能算一体抵近服务支撑应用增长，及消弭能算时空错配问题。

而由以上算力重构带动了EFLM的能力提升挑战，首先于纵向扩展 (Scale-up) 方面，单点AI智算功耗提升需要更大功率及更细时间尺度的能源调度，进一步

推升EFLM边端算力；首先于横向扩展 (Scale-out) 方面，AI算力基础设施规模扩张带来能源调控边界的扩张，推升EFLM云端协同边界的扩展，导致运控体系协同算力增加，进一步推升对AI算力的需求。

由以上AI算力助力EFLM边界扩展，则可转化应对下一阶段的能算一体发展，循环而至“以算调能、以能促算”的发展目标，利用EFLM能源精细化调控解决区域AI算力绿电不好用，用不好的痛点，并依托AI算力解决EFLM测不精，控不细难题；由EFLM主导的新型能源体系支撑绿色算力中心的规模化部署，为AI算力提供“内生可持续”的能源保障，推动数字技术与能源革命的双向突破。

参考文献

- L. a. P. J. a. Y. W. a. R. X. a. S. X. Cai, "Self-Evolving and Transformative Protocol Architecture for 6G," *IEEE Wireless Communications*, pp. 178–186, 2023.
- J. a. H. W. a. X. F. a. X. P. a. H. K. a. I. B. a. F. P. a. Z. A. Liang, "Code as Policies: Language Model Programs for Embodied Control," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500, 2023.
- L. a. Y. P. a. W. G. a. T. W. a. C. Z. Qian, "Self-Evolving Wireless Communications: A Novel Intelligence Trend for 6G and Beyond," *2024 International Conference on Future Communications and Networks (FCN)*, pp. 1–6, 2024.
- Y. X. a. Y. L. a. B. Z. a. P. C. a. H. Z. a. E. H. a. J. Z. a. W. H. a. X. J. a. X. Z. a. H. J. a. D. M. a. Q. Z. a. H. T. a. P. L. a. W. K. a. G. Pan, "Bio-plausible reconfigurable spiking neuron for neuromorphic computing," *Science Advances*, p. 6733, 2025.
- F. a. W. Y. a. W. Z. a. L. L. a. D. Y. a. K. B. a. L. L. a. W. S. a. X. Y. a. Y. S. Tu, "ReDCIM: Reconfigurable Digital Computing- In -Memory Processor With Unified FP/INT Pipeline for Cloud AI Acceleration," *IEEE Journal of Solid-State Circuits*, pp. 243–255, 2023.
- M. a. A. G. S. a. S. A. a. K. N. a. G. S. Singh, "Deep-Learning-Based Blockchain Framework for Secure Software-Defined Industrial Networks," *IEEE Transactions on Industrial Informatics*, pp. 606–616, 2021.
- J. G. N. B.-h. Ali, "An Intelligent Blockchain-based Secure Link Failure Recovery Framework for Software-defined Internet-of-Things," *Journal of Grid Computing*, p. 57, 2023.
- T. T. a. K. N. Hieu, "Swift Recovery: An Innovative Routing Approach for Multi-Failure Protection in Software-Defined Networks," *IEEE Access*, pp. 158685–158702, 2024.
- D. a. R. F. M. V. a. V. P. E. a. R. C. E. a. A. S. a. U. S. Kreutz, "Software-Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE*, pp. 14–76, 2015.
- J. a. C. L. a. Y. S. a. S. X. S. Pan, "Network for AI and AI for Network: Challenges and Opportunities for Learning-Oriented Networks," *IEEE Network*, pp. 270–277, 2021.
- X. a. W. X. a. Y. B. a. H. Q. a. H. M. Chen, "Deep Learning-Based Traffic Prediction for Energy Efficiency Optimization in Software-Defined Networking," *IEEE Systems Journal*, pp. 5583–5594, 2021.
- S.-R. a. K. T.-H. a. T. H.-Y. a. L. C.-S. a. W. C.-T. a. H. J.-S. a. L. T.-S. a. K. T. a. Y. D. Chun, "InFO_SoW

- (System-on-Wafer) for High Performance Computing,” *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, pp. 1–6, 2020.
- Y. a. L. X. a. W. H. a. H. Z. a. Y. X. a. Z. J. a. Y. Q. a. X. Z. a. G. S. a. F. J. a. S. H. a. T. X. a. D. X. a. W. S. a. Y. S. Hu, “Wafer-Scale Computing: Advancements, Challenges, and Future Perspectives [Feature],” *IEEE Circuits and Systems Magazine*, pp. 52–81, 2024.
- G. Lauterbach, “The Path to Successful Wafer-Scale Integration: The Cerebras Story,” *IEEE Micro*, pp. 52-57, 2021.
- X. W. Y. W. Y. X. Y. Ma, “Survey on chiplets: interface, interconnect and integration methodology,” *CCF Transactions on High Performance Computing*, pp. 43–52, 2022.
- Q. L. S. L. S. Z. L. C. L. Z. L. G. W. Z. Z. G. H. Q. D. Y. Z. Z. P. Z. Jiangxing WU, “From SoC to SDSoW: a new paradigm for microelectronics development,” *SCIENTIA SINICA Informationis*, p. 1350, 2024.
- X. a. C. C. a. W. Y. a. Z. S. a. L. Y. a. L. M. a. H. T. Tang, “Computing power network: The architecture of convergence of computing and networking towards 6G requirement,” *China Communications*, pp. 175–185, 2021.
- H. a. H. S. a. X. M. a. G. D. a. W. X. a. L. V. C. a. W. W. Zhang, “How Far Have Edge Clouds Gone? A Spatial-Temporal Analysis of Edge Network Latency In the Wild,” *2023 IEEE/ACM 31st International Symposium on Quality of Service (IWQoS)*, pp. 1–10, 2023.
- S. a. B. L. a. J. L. a. H. H. a. X. Z. a. J. P. a. W. W. Yukun, “Computing power network: A survey,” *China Communications*, pp. 109–145, 2024.
- S. M. D. O. a. D. K. Michał Król, “Compute First Networking: Distributed Computing meets ICN,” *In Proceedings of the 6th ACM Conference on Information-Centric Networking*, pp. 67–77, 2019.
- J. Wu, *Cyberspace Mimic Defense*, Springer, 2020.
- J. a. A. J. a. B. M. a. D. C. a. C. J. a. C. F. J. a. E. C. a. T. M. a. N. G. a. F. J. L. Perez-Cerrolaza, “Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey,” *Association for Computing Machinery*, p. 40, 2024.
- Z. a. K. X. a. S. L. a. Z. L. a. X. C. Zhou, “Endogenous Security Defense against Deductive Attack: When Artificial Intelligence Meets Active Defense for Online Service,” *IEEE Communications Magazine*, pp. 58–64, 2020.
- K. a. T. S. Dhanushkodi, “AI Enabled Threat Detection: Leveraging Artificial Intelligence for Advanced Security and Cyber Threat Mitigation,” *IEEE Access*, pp. 173127–173136, 2024.
- K. a. K. S. a. P. G. a. V. K. a. A. K. a. S. P. Chokkanathan, “AI-Driven Zero Trust Architecture: Enhancing Cyber-Security Resilience,” *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 1–6, 2024.
- 贺力真, 汪振涛. 能源与人工智能的协同进化:从算力革命到可持续发展路径[J]. 信息通信技术与政策, 2025, 51(6): 15-20.
- S. a. O. K. a. D. M. Zhu, “Green AI for IIoT: Energy Efficient Intelligent Edge Computing for Industrial Internet of Things,” *IEEE Transactions on Green Communications and Networking*, pp. 79–88, 2022.

- S. a. S. Y. a. L. B. a. Q. B. a. S. K. a. L. Y. a. T. X. Zheng, "Incentive-Based Integrated Demand Response for Multiple Energy Carriers Considering Behavioral Coupling Effect of Consumers," *IEEE Transactions on Smart Grid*, pp. 3231–3245, 2020.
- C. a. K. Y. Zhang, "Low-Carbon Economy Optimization of Integrated Energy System Considering Electric Vehicles Charging Mode and Multi-Energy Coupling," *IEEE Transactions on Power Systems*, pp. 3649–3660, 2024.
- G. C. J. H. F. Z. Yinbiao Shu, "Building a New Electric Power System Based on New Energy Sources," *Strategic Study of CAE*, pp. 61–69, 2021.