第3章

行业智能化参考架构

行业智能化转型是一个长期的、循序渐进的过程，如何选择转型道路、如何分层分级建设智能化ICT基础设施，将成为智能化转型的关键，需要有一个明确的指导思想来引领转型过程，在不同的阶段做出匹配的选择，避免走弯路、走错路，提升转型的效率。

第一，分层分级的行业智能化转型建设路径是行业智能化参考架构的基础。每个行业都有其独特的业务模式和流程，智能化转型需要针对行业特点制定合适的策略和方案。参考架构为行业提供了一个清晰、有序的转型蓝图，避免盲目投入和资源浪费。

第二，多技术融合是行业智能化的关键。行业智能化涉及算力、存储、网络、大数据、人工智能等多种技术，这些技术相互关联、相互支持。通过参考架构整合这些技术，形成协同效应。同时，参考架构还可以根据行业需求和技术发展趋势，不断优化多种技术组合，推动行业创新。

第三，AI创新与算法模型行业适配度是行业智能化的核心。智能化转型的关键在于利用AI技术提升传统行业业务效率和质量。随着AI技术的演进，通过参考架构模型算法层面整合最新的算法模型应用到行业场景中，提高智能化等级。此外，参考架构还需要考虑行业的特殊需求，定制符合行业特点的AI解决方案，提高AI模型的适配度和准确性。

第四，高质量行业数据集是行业智能化的基石。数据是AI技术的驱动力，高质量的数据集对于提升AI模型的性能至关重要。参考架构的数据层面关注数据的采集、清洗、标注等环节，确保数据的质量和准确性，还需要考虑数据的共享和流通，促进数据在行业内的有效利用。

第五，高效算力是行业智能化的保障。随着AI技术的不断发展，算力需求也在不断增加。参考架构算力层面需要关注算力的优化配置和高效利用，确保行业智能应用能够得到足够的算力支持，还需要关注算力的可持续发展，推动绿色计算和节能减排。

综上所述，基于在城市、金融、交通、制造等20多个行业智能化实践过程中的总结，华为提出具备分层开放、体系协同、敏捷高效、安全可信等特征的、全行业通用的行业智能化参考架构，联合行业伙伴共同构筑行业智能化的基础设施，使能百模千态的AI大模型，加速千行万业走向智能化，如图3-1所示。

行业智能化参考架构是系统化的架构，它包含智能感知、智能联接、智能底座、智能平台、AI大模型、千行万业共6层，这6层相互协同、相互促进。行业智能化参考架构是面向全行业的、能够服务不同智能化阶段的参考架构，通过分层分级建设，选取合适的技术和产品，提升行业的智能化水平，具备协同、开放、敏捷、可信的特征。

协同。大模型时代，智能化产业的上下游产业多，产品能力复杂，需要各从业企业基于行业智能化参考架构来构建产品和能力，相互之间协同以形成合力，共同完成智能化体系的

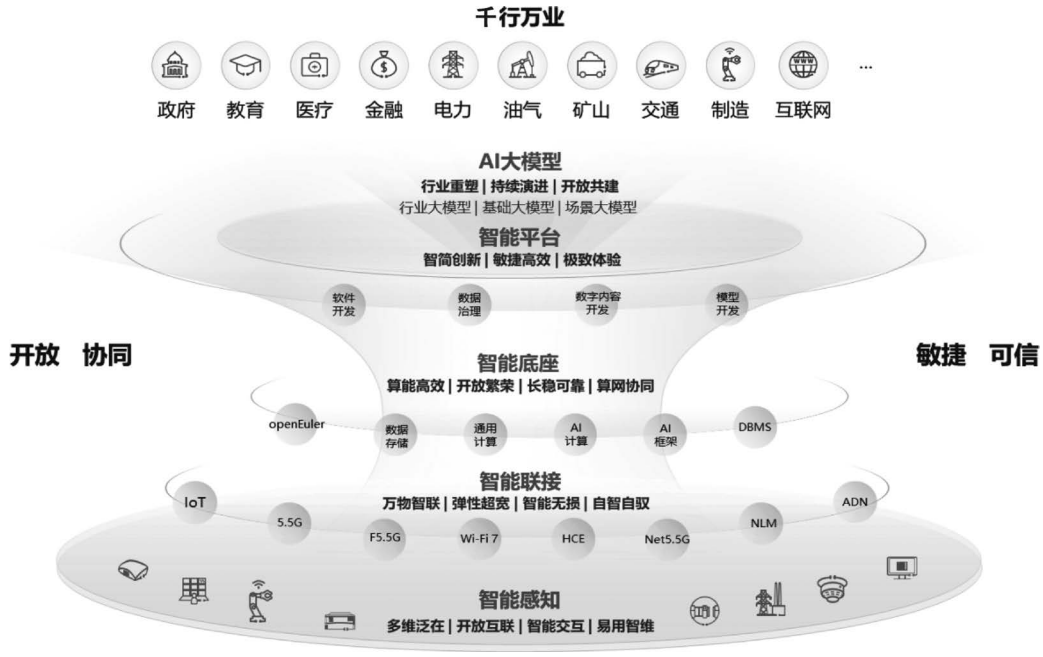

图 3-1　行业智能化参考架构

建设。各个行业的企业之间也需要协同，共同构筑有竞争力的基础大模型、行业大模型，服务于行业的智能化发展。在智能化过程中通过云、管、边、端的协同，业务信息实时同步，提升业务的处理效率，并通过应用、数据、AI 的协同，打通组织鸿沟，使能业务场景全面智能化。

开放。行业智能化发展是一个庞大的工程，需要众多的企业共同参与，以开放的架构助力行业智能化发展。通过算力开放，以丰富的框架能力支持各类大模型的开发，形成百模千态；通过感知开放，接入并打通品类丰富的感知设备，实现万物智联；通过模型开放，匹配千行万业的应用场景，实现行业智能。

敏捷。在智能化的过程中，可按照业务需要灵活匹配合适的 ICT 资源，并通过丰富、成熟的开发工具和框架构筑智能化业务，让业务人员直接参与智能化业务的开发，快速上线智慧应用。

可信。智能化系统必须是可信的，在系统安全性、韧性、隐私性、人身和环境安全性、可靠性、可用性等方面全面构筑可信赖的能力，并从文化、流程、技术 3 个层面确保在各场景中落地；智能化应用的运行过程必须是可信的，可追溯、防篡改，避免受到外部的恶意破坏。

3.1　智能感知

智能感知是物理世界与数字世界的纽带，它基于品类丰富、泛在部署的终端设备，对传统的感知能力进行智能化升级，构建一个无处不在的感知体系，具备多维泛在、开放互联、智能交互、易用智维等特点。

多维泛在。智能化时代，需要对事物进行全方位的感知，才能获取到完整、全面的信息，支撑后续的智能化业务处理。通过雷达、视频、温度传感、气压传感、光纤感知等多种类型的感知设备从不同的维度获取数据，进而汇总成为更全面的信息，支撑后续的智能分析和处置；同时，为了保证能够获取到准确且实时的信息，感知设备还需要贴近被感知的对象，并保持实时在线，充分获取感知数据，实时上传至处理节点，形成无处不在的感知。

开放互联。行业里各类感知终端种类繁多，协议七国八制导致数据难互通，难以支撑复杂的业务场景。因此，需要开放终端生态，通过鸿蒙或其他智联操作系统，将协议复杂、系统孤立的终端有机协同起来，实现对同一感知对象的联动感知能力，做到“一碰传、自动报”。开放应用生态，ICT 技术与场景化深度融合，实现精细化治理。

智能交互。随着各类智能终端的广泛应用，人与人之间、人与设备之间的协同也越来越广泛，视频会议、远程协作等交互场景在行业应用中得到了很大的推广。通过云边协同、AI 大模型等技术的应用，极大地提升设备认知与理解能力，实现软件、数据和 AI 算法在云边端自由流动，并通过包含智联操作系统的终端设备，基于对感知数据的处理结果，在物理世界中进行响应处理，实现智能的交互能力。

易用智维。行业的业务场景复杂，对感知的要求也有很大差异，感知设备有相当大的比例安装在不易于部署维护的地点，如荒野、山顶、铁路周界、建筑外围等，其中一部分设备在获取电力、网络资源时也存在一定的困难。因此，需要感知设备具备网算电一体集成、边缘网关融合接入等能力，实现感知设备智简部署、即插即用，智简运维平台和工具数字化、智能化，实现无人化、自动化的可视可管可维。

智能感知层的关键技术和部件包括鸿蒙感知、多维感知、通感一体等。鸿蒙感知是以鸿蒙智联操作系统为核心的智能终端系统，具备接入简单，一插入网、一跳入云，安全性强等特点；多维感知是通过雷视拟合、光视联动等技术融合创新，提高全场景感知精准性；通感一体是通过有线和无线组合，实现无处不在、无时不在的感知。

3.2 智能联接

行业智能化的场景复杂多样，智能联接用于智能终端和数据中心的联接、数据中心之间的联接、数据中心内部的联接等，解决数据上传、数据分发、模型训练等问题。各种场景对联接都有不同的要求。例如某个工业园区场景中智能终端和数据中心的联接，AOI 机器视觉质检要求实时推理交互，软件包下载要求高峰值带宽，视频会议要求稳定带宽，需要借助网络切片保障不同流量的互不干扰。在数据中心中，AI 训练集群网络丢包率会极大影响算力效率，1/10000 的丢包率会导致算力降低 10%，而 1/1000 的丢包率会导致算力降低 30%。因此，行业智能化需要万物智联、弹性超宽、智能无损、自智自驭的智能联接。

万物智联。在行业智能化时代，种类繁多的感知终端（如雷达、行业感知、光纤感知、温度感知、气压感知等）都需要通过网络自动上传感知数据，以支撑各种类型的业务系统。数据上传需要实时、准确，不能有丢失。智能联接综合采用 5G-A、F5G Advanced、Wi-Fi 7、超

融合以太(HCE)、IPv6+等多种网络技术,推进全场景、全触点、无缝覆盖的泛在联接,支撑数据采集汇聚,推进智能应用普及,为智能化参考架构的持续进化构筑万物智联的基础。

弹性超宽。随着行业智能化不断发展、感知能力不断丰富与增强,生成的业务信息量也在极速增长,支撑大模型的训练数据更加丰富完善,训练出的模型更加精准。训练出的模型也要迅速下发,推动业务处理更加智能。面向PB(Petabyte,拍字节)级样本训练数据上传、TB(Terabyte,太字节)级大模型文件分发的突发性、周期性、超宽带联接需求,需要建设大带宽、低时延、智能调度的网络;基于时延地图和带宽地图动态选择最优路径,实现极速推理和实时交互,为行业智能化参考架构打通"数据上得来、智能下得去"的持续进化循环。

智能无损。面向超大规模AI集群互联需求,以400GE/800GE超融合以太、网络级负载均衡等技术实现大规模、高吞吐、零丢包、高可靠的智能无损计算互联;智算数据中心网络升级,以网强算,通过算、网、存深度协同优化,支撑万亿级参数的模型训练,让智能化参考架构越来越智能。面向海量智能感知终端连云入算、AI助理以云助端等场景,基于网络大模型(NLM)实现智能感知应用类型、智能优化联接体验、智能保障网络质量,为极速推理、协同工作、音视频会议等各种应用提供智能无损的高品质联接,让智能持续进化,服务更多的生产、生活场景。

自智自驭。基于网络大模型识别应用与终端类型,准确生成配置与仿真验证、准确预测故障与安全风险,并实现网络零中断(智能预测网络拥塞准确率为99.9%)、安全零事故(智能预测未知威胁)、体验零卡顿(智能识别应用类型并保障体验),加速网络自动驾驶向L4级迈进,实现网络的自智自驭,提升智能化参考架构的整体运转效率。

智能联接主要涉及接入网络、广域网络、数据中心网络。

接入网络。承担着感知设备的接入及汇聚到数据中心网络或广域网络的职责。接入网络通过5G-A、F5G Advanced、Wi-Fi 7、超融合以太(HCE)、IPv6+等技术,实现稳定、可靠、低时延的感知设备接入;同时,接入网络还承载着多种业务类型,如实时业务处理、训练数据采集与上传、推理模型下发至边缘计算节点等,需要接入网络能够根据业务类型分别设置网络资源,为不同的业务数据设置不同的资源优先级。

广域网络。具备多分支机构的大型企业存在大量的数据跨分支机构互传的场景,如训练数据上传、算法模型下发、业务应用下发、业务数据传输等,相应地需要在分支机构之间提供稳定、大带宽的广域网络。企业可根据自身的实际情况,选择租用运营商网络或自建广域网络的方式,获取稳定、可靠、高带宽的多分支机构间的网络联接能力。

数据中心网络。随着AI大模型的兴起,大模型训练成为数据中心的一个重要职责,其超大规模的数据分析对数据中心的网络也带来了新的挑战,传统的基于计算机总线的数据中心网络技术已无法满足大模型训练的要求。因此,数据中心网络需要新的网络架构,能够打通各协议间的壁垒,"内存访问"直达存储和设备,并统一芯片侧高速接口,打破"带宽墙",使能端口复用。数据中心的业务类型也是多样的,例如在大模型训练时就存在参数面、业务面、存储面等网络平面,需要能够按照业务类型建立网络平面,并相互隔离。

3.3 智能底座

智能底座提供大规模 AI 算力、海量存储及并行计算框架，支撑大模型训练，提升训练效率，提供高性能的存算网协同。根据场景需求不同，提供系列化的算力能力。适应不同场景，提供系列化、分层、友好的开放能力。另外，智能底座还包含品类多样的边缘计算设备，支撑边缘推理和数据分析等业务场景。具备算能高效、开放繁荣、长稳可靠、算网协同等特点，以更好地支撑行业智能化。

算能高效。随着大模型训练的参数规模不断增长、训练数据集不断增大，大模型训练过程中需要的硬件资源越来越多，时间也越来越长。需要通过硬件调度、软件编译优化等方式，实现最优的能力封装，为大模型的训练加速，提升算能的利用率。同时，针对基础大模型、行业大模型、场景大模型的训练算力需求，以及中心推理、边缘推理的算力需求，提供系列化的训练及推理算力基础设施配置，根据业务场景按需选择，确保资源价值得到最大化的利用。在数据存储方面，闪存技术具备高速读写能力和低延迟特性，并伴随着其堆叠层数与颗粒类型方面的突破，带来成本的持续走低，使其成为处理 AI 大模型的理想选择。通过全局的数据可视、跨域跨系统的数据按需调度，实现业务无感、业务性能无损的数据最优排布，满足来自多个源头的价值数据快速归集和流动，以提升海量复杂数据的管理效率，直接减少 AI 训练端到端周期。

开放繁荣。不同场景、不同类型的大模型，根据大模型的参数规模、数据量规模，需要的算力有着很大的差异；在推理场景，中心推理和边缘推理对算力的要求也不一样。行业用户可以根据实际业务场景选择不同的模组、板卡、整机、集群，获取匹配的算力，并可在品类丰富的开源操作系统、数据库、框架、开发工具等软件中进行选择，屏蔽不同硬件体系产品的差异，帮助用户在繁荣的生态中选择合适的产品和能力，共同形成行业智能化的底座。

长稳可靠。大模型业务场景下，一次模型训练往往要耗费数天甚至数月的时间，如果中间出现异常，将会有大量的工作成果被浪费，耗费宝贵的时间和计算资源。为减少异常导致的训练中断、资源浪费，要保证训练集群长期稳定，提升集群的稳定性；同时，在出现极端情况时，可以使用过程数据恢复训练，降低因外部因素带来的影响。

算网协同。随着大模型的参数数量、训练数据规模的不断增长，模型训练所消耗的时间也不断增加，逐渐变得不可接受。传统的计算机总线＋网络的数据传输方式已成为瓶颈，难以继续提升效率。因此，需要算网协同的传输架构，提升数据的传输效率和模型训练速度。同时，网络需要参与计算，减少计算节点交互次数，提升 AI 训练性能。

同样，在大模型训练过程中，数据在存储、内存、CPU(Central Processing Unit，中央处理器)间移动，占用大量的计算和网络资源。为减少资源占用，需要存算协同架构，通过近存计算、以存强算的能力，让数据在存储侧完成部分处理，将算力卸载下沉进存储实现随路计算，减少对 AI 计算能力的占用。智能底座的主要技术特征有：

计算能力。计算能力简称算力，实现的核心是 GPU(Graphics Processing Unit，图形处

理器)/NPU(Network Process Unit,网络处理器)、CPU、FPGA(Field Programmable Gate Array,现场可编程门阵列)、ASIC(Application-Specific Integrated Circuit,专用集成电路)等各类计算芯片,以及对应的计算架构。AI算力以GPU/NPU服务器为主。算力由计算机、服务器、高性能计算集群和各类智能终端等承载。算力需要支持系列化部署,训练需要支撑不同规格(万卡、千卡、百卡等)的训练集群、边缘训练服务器;推理需要支持云上推理、边缘推理、高性价比板卡、模组和套件。并行计算架构需要北向支持业界主流AI框架,南向支持系列化芯片的硬件差异,通过软硬协同,充分释放硬件的澎湃算力。

数据存储。复杂多样的业务场景,带来了复杂多样的数据类型。数据存储需对不同类型的数据,通过全闪存存储、全对称分布式架构等技术手段,为不同的业务场景提供海量、稳定高性能和极低时延的数据存储服务;为特定业务场景提供专属数据访问能力,如直通GPU/NPU缩短训练数据加载时间至毫秒级;具备数据的备份恢复机制,以及防勒索机制等安全能力,确保数据的安全、可用。

操作系统。操作系统对上层应用,要屏蔽不同硬件的差异,提供统一的接口,完成不同硬件的兼容适配,提供良好的兼容性,为应用软件的部署提供尽可能的便利;针对不同的硬件的特征,操作系统需要针对性的优化,确保能充分发挥硬件的能力;在多CPU、CPU和GPU、NPU协同的情景下,操作系统如何协调调度,也是一个关键的能力。

数据库。海量、格式多样的数据,追求极致的业务性能,对数据库也带来了新的挑战。为了适应业务的变化,数据库需要高性能,海量数据管理,并提供大规模并发访问能力;高可扩展性、高可靠性、高可用性、高安全性、极速备份与恢复能力,都是对数据库的基本要求。

云基础服务。智能底座上运行的各种应用、服务,在不同的时间段对应的业务量是有差异的,为了合理利用智能底座的硬件资源,智能底座通过虚拟化、容器化、弹性伸缩、SDN(软件定义网络)等技术,对外提供云基础服务能力,提升资源的利用效率。

3.4　智能平台

在海量的数据从感知层生成、经过联接层的运输,汇聚到智能平台,通过数据治理与开发、模型开发与训练,积累行业经验,最终服务智能应用的构建。

智能平台理解数据、驱动AI,支撑基于AI大模型的智慧应用的快速开发和部署,使能行业智能化,具备智简创新、敏捷高效、极致体验等特点。

智简创新。围绕软件、数据治理、模型、数字内容等生产线能力,提供一系列的开发使能工具,并通过数据、AI、应用的协同,让智慧应用的构建更高效、更便捷,让行业应用的创新更简单、更智能。

敏捷高效。智能化的开发生产线能力,为业务人员提供了多样化的业务开发方式选项;强大的DevOps(Development和Operations的组合词,是一组过程、方法与系统的统称)能力让业务迭代开发过程更敏捷,一键发布能力让业务上线速度更快,效率更高。

极致体验。具备简单易用的低代码、零代码业务配置能力,开发门槛低,业务人员可以

直接参与到模型开发、数据治理、应用开发中；为不同的用户提供个性化的操作界面，提升使用者的体验。

智能平台层的主要技术特征包括数据治理生产线、AI 开发生产线、软件开发生产线以及数字内容生产线。智能平台支持 AI 模型在不同框架以及不同技术领域的开发和大规模训练。

数据治理生产线。核心是从数据的集成、开发、治理到数据应用消费的全生命周期智能管理。一站式实现从数据入湖、数据准备、数据质量到数据应用等全流程的数据治理，同时融合智能化治理能力，帮助数据开发者大幅提升效率。

AI 开发生产线。它是 AI 开发的一站式平台。提供从算力资源调度、AI 业务编排、AI 资产管理以及 AI 应用部署，提供数据处理、算法开发、模型训练、模型管理、模型部署等 AI 应用开发全流程技术能力。同时，AI 应用开发框架，屏蔽掉底层软硬件差异，实现 AI 应用一次开发、全场景部署，缩短跨平台开发适配周期，并提升推理性能。

软件开发生产线。提供一站式开发运维能力，面向应用全生命周期，打通需求、开发、测试、部署等全流程。提供全代码、低代码和零代码等各种开发模式。面向各类业务场景提供一体化开发体验。

数字内容生产线。提供 2D、3D 数字内容开发，应用开发和实时互动框架。根据用户需求，生成服务，如数字人等。使用者无需专业设备，即可使用的内容生产工具。

3.5 AI 大模型

AI 大模型分为 3 层，即基础大模型、行业大模型、场景模型。基础大模型（L0）提供通用基础能力，主要在海量数据上抽取知识学习通用表达，一般由业界的 L0 大模型供应商提供；行业大模型（L1）是基于 L0 基础大模型，结合行业知识构建，利用特定行业数据，面向具体行业的预训练大模型，无监督自主学习了该行业的海量知识，一般由行业头部企业构建；场景大模型（L2）指面向更加细分场景的推理模型，是实际场景部署模型，是通过 L1 模型生产出来的满足部署的各种模型。

AI 大模型在发展过程中呈现出了行业重塑、持续演进、开放共建等特点。

行业重塑。AI 大模型叠加行业场景，赋予行业场景更智能的处理能力，提升业务效率，降低企业成本，促进行业创新，为行业的发展注入新的生命力，重塑行业的智能化进程。

持续演进。行业场景使用大模型提升业务效率的同时，也会产生大量的业务数据，这些数据再对大模型进行训练，让大模型的能力越来越强大，推理越来越准确，成为行业智能化的有力支撑。

开放共建。行业客户与大模型供应商共同打造多样化多层级的大模型，构筑满足各类场景中需要的大模型，为不同行业场景提供多样化的选择，服务行业智能化发展。

大模型聚焦行业，从 L0、L1 到 L2，遵从由“通”到“专”的分层级模式，可实现从 L0 通用模型到 L1 行业模型再到 L2 专用模型的快速开发流程。

在建设大模型体系时，要依照企业的规模、能力、组织结构和需求因地制宜，层层落实，要充分考虑云网边端协同、网算存的协同，让 AI 上行下达。大模型可以分层分级建设，从 L0 到 L1，再到 L2，不断有行业数据加入来提升模型的训练效果，同时也需要模型压缩来节约推理资源。模型压缩是实现大模型小型化的关键技术，大模型通过压缩技术可以达到 10～20 倍参数量级压缩，使千亿模型单卡推理成为可能，节省推理成本；同时，模型压缩降低计算复杂度，提升推理性能。

在实际应用中，需要结合业务场景变化，迭代演进 AI 大模型能力，边学边用，越用越好。对于 NLP 大模型，可以结合自监督训练方式，进行二次训练，不断补充行业知识；在具体任务场景下，可以使用有监督训练方法进行微调，快速获得需要达到的效果；进一步地，可以基于自有训练后的模型，进行强化学习，获得更出色的模型。对于 CV 大模型，企业/行业用户可以结合自有行业数据，进行二次训练，迭代获得适配与自身行业的 L1 预训练大模型；同时，在具体细分场景中，可以提供小样本，基于行业预训练的 L1 模型进行微调，快速获得适配自身业务的迭代模型，小样本量，迭代也更快速。

大模型的三级模型之间可以交互优化。L0 模型可以为 L1 模型提供初始化加速收敛，L1 可以通过模型抽取蒸馏产生更强的 L2 模型，L2 也能够在实际问题中通过积累案例数据或者行业经验反哺 L1。

3.6　智赋万业

千行万业的智慧应用是行业智能化参考架构的价值呈现，每个个体所能感受到的个性化、主动化服务体验都来自应用。智慧应用的发展关键是探索可落地场景，对准其痛点，通过 ICT 和行业/场景 AI 大模型的结合，快速创造价值。所有这些场景汇聚起来，便能涓滴成河，逐步完成全场景智慧的宏伟蓝图。