HUAWEI

迈向智能世界白皮书

计算

共建计算产业，共赢数智时代

构建万物互联的智能世界

序言

数字经济已经成为全球经济增长的主引擎，2021年我国数字经济规模达45.5万亿元占GDP比重达到39.8%，数字经济的快速发展，带来数字化、智能化的巨大发展机遇。

数字经济时代，数据是生产资料，算力是生产力。人均算力决定了数字经济的发展水平，算力基础设施成为新基建的核心，为数字经济发展提供新动能。

随着通用算力的普及，使能了各行各业的数字化，带动了数字经济发展。而数字经济增长又产生更多的数据，更多的数据，又需要更多的算力。我们预计，到2030年，全球通用计算算力相比2020年将增长10倍，AI算力将增长500倍。

计算从通用计算进入通用计算+AI计算的多样性计算时代。通用计算构建了数字经济发展的基础，AI计算将成为数字经济发展的加速器，从数字化到智能化，人工智能作为新的GPT（通用目的技术），将使能数字经济迈向新高度。

为驱动数字经济的高速发展和满足数字经济多样化的场景需求，三年前，华为发起了面向通用计算的鲲鹏产业、面向人工智能的昇腾产业，得到了全产业链伙伴的积极响应，发展迅速。目前已发展超过5000家合作伙伴，260万开发者，13000多个解决方案完成认证。

鲲鹏已经广泛应用于政府、金融、电信、电力、交通、制造、教育、医疗等行业核心应用系统，成为国计民生行业数字化转型的首选。

基于昇腾AI，在全国已构建了20多个人工智能计算中心，让算力成为一种公共基础设施，开创普惠AI新模式。并从计算中心走向算力网络，成为千行百业转型升级的智能根基。

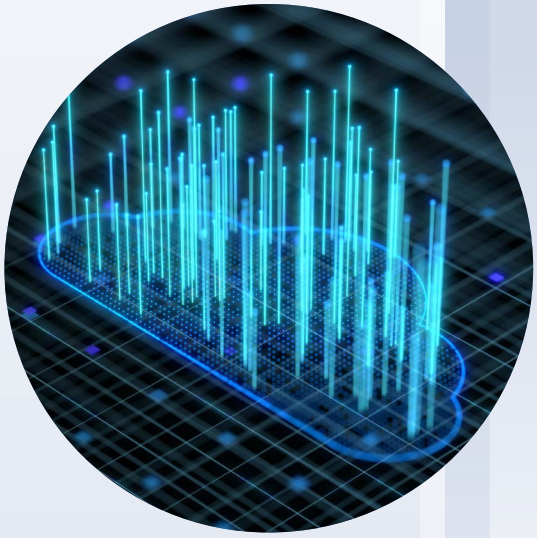
面向未来，华为将坚持围绕鲲鹏和昇腾，携手产业伙伴共建计算产业生态；坚持“硬件开放、软件开源、使能伙伴和发展人才”，和产业伙伴共同构筑坚实的算力底座。

共建计算产业，共赢数智时代。

目录

趋势一

ARM 成为多样性计算的重要选择



产业趋势

1.1.1 应用的多样化驱动算力多样性发展

1）随着自动驾驶、云游戏、VR/AR等应用的兴起，以及物联网、移动应用、短视频、个人娱乐、人工智能的爆炸式增长，应用越来越多样化，用户对应用体验的追求不断提高。数据中心侧，传统单一架构难以满足要求。

万物互联的智能时代，非结构化数据占比越来越大。相对应原来可以用数据库二维表结构来实现的结构化数据，海量、多种多样非结构化数据，如文本、图片、语音、视频等这类数据的加工、处理、传输，自然需要多样性的计算来匹配。举例来说，CPU处理大数据、Web等场景是非常匹配的，但是对于图形、图像的处理，就需要GPU来匹配；而日常生活中的图形/图像识别、智能搜索推荐等，就需要基于AI计算的NPU来处理了。所以说，业务应用场景的多样性、数据的多样化，使得计算进入多样性计算的新时代。

2）边缘计算兴起，未来超过70%的数据和应用将在边缘产生和处理。边缘和移动端设备受场景约束，处理能力和性能的提升受到限制，需要与云协同。随着5G的规模部署，网络传输时延、带宽、连接密度均得到数量级的提升，给端-边-云协同提供了基础保障。目前云、边、端的计算架构、开发模式存在较大差异，应用须多次开发和部署，开发复杂度、成本增加，同时，架构的差异导致协同时性能的损耗（大约在10%-20%），因此，反向驱动了中心侧计算架构与边、端侧的一致性。

1.1.2 多样性计算需求，加速算力格局转换，ARM算力从嵌入式场景快速延深至服务器场景

ARM算力是从最初的端侧起步，在智能手机、平板、智能电视等领域占据绝对领先的份额，但随着云、边、端协同的驱动、多样性计算的

发展，已经开始进入到算力更高的服务器领域，同时也表现出显著的优势：

1）在分布式数据库、大数据、Web前端等高并发应用场景，单芯片核数更多的ARM架构处理器相比传统处理器拥有更好的并发处理效率。

2）绝大多数移动终端采用ARM架构处理器，端云同构为开发人员在整个生态系统的编写与优化上提供便利，而且能够降低异构环境开发所造成的性能损失和潜在漏洞风险。随着云化进程的推进，大量基于ARM架构的终端业务与数据中心的云端业务维持同构，可以实现应用开发、部署和运行的无缝协同，大幅度降低开发者开发的难度。

3）ARM生态优势不断推动技术进步，近年来不断涌现出创新的服务器产品和解决方案，如亚马逊AWS对外提供的Amazon EC2服务就是基于ARM架构，华为基于ARM架构鲲鹏处理器打造了TaiShan系列服务器等。在高性能计算领域，以ARM、RISC-V 为代表的多样性计算平台也逐渐发挥重要作用，例如欧盟EPI（欧洲处理器计划）项目致力于打造本土基于ARM架构核心处理器和RISC-V架构加速器芯片的百亿亿级超级计算机；日本“富岳”超算系统采用自主开发的ARM架构处理器，成为全球首台基于ARM芯片的TOP500冠军超级计算机等。

4）ARM架构授权模式让伙伴既自主发展又共享生态平台，加速产业链多样化。ARM的商业模式不以出售芯片为主，而是架构授权。合作伙伴可以根据自身需求，灵活选择不同的授权模式：一是架构授权模式。基于ARM架构，可以自主扩充指令集并升级产品；二是CPU核授权模式（软核和硬核）。基于ARM CPU IP可实现设计生产，升级则需完成新CPU核授权的获取。

5)2000年x86占据市场第一的份额，总算力输出达到了70%。到了2020年，算力架构发生了逆转，世界上最大算力架构变成了ARM平台，基于ARM指令的处理器总算力输出占比超过80%。

1.1.3 中国市场，服务器侧ARM生态已逐步成熟，并全面应用于国计民生行业；

全球范围内，以ARM为核心架构的CPU已经开始显现出增长趋势。在中国，众多芯片厂商和云巨头也纷纷布局基于ARM架构的系列产品，鲲鹏、飞腾已耕耘多年，ARM服务器市场份额持续增加。

以鲲鹏为代表的ARM服务器，已经广泛应用于包括政府、金融、电信、电力、交通、制造、教育、医疗等行业核心场景；各行业生态已经建立，超过12000个行业应用完成适配认证，产业生态瓶颈已经消除。

行动建议

1.2.1 基于业务需求，识别适合ARM架构的业务场景，主动规划部署ARM架构服务器

数字化转型、人工智能和5G的在垂直行业的广泛应用带来了海量数据处理、高能效边缘计算等问题，尤其在电信、金融、政府、能源等重点行业，ARM架构能够更好的满足数字化应用对IT基础设施算力的严苛要求，在升级发展中发挥关键作用。

以电信行业为例，5G时代数据量爆发式增长、电信云面临从架构到底层硬件基础设施的全面升级，在容器化部署、分布式存储和边缘计算等关键场景都非常适合引入ARM架构，充分利用其多核高并发、大内存和高内存带宽等架构优势：

1）IT云方面， IT支撑系统业务逻辑更趋复杂，实时数据处理、高并发数据处理、大数据分析等技术需求不断扩大，容器化部署、分布式处理等场景加速向CRM、BOSS、MSS等核心系统渗透，需要底层IT基础设施在并行计算、内存容量和带宽等方面提供更好能力匹配；

2）网络云方面，5G核心网采用原生云化设计思路和微服务架构，将网元功能拆分为细颗粒度的网络服务，为差异化的业务场景提供敏捷的系统架构支持，核心网容器化、硬件资源池化成为发展方向，对底层计算架构的多样性、负载能力和计算效率提出新的要求；

3）在边缘节点方面，为应对大视频、物联网等各类高带宽和低时延的边缘计算类业务，电信云计算能力将向移动边缘节点下沉，边缘数据中心IT基础设施将面临计算、存储等网络能力的全面提升以实现大流量、高并发、低时延的本地数据处理能力。（图1）

围绕重点行业的计算诉求，主动推进ARM架构服务器的应用，依托ARM处理器多核高并发、高效可靠的硬件平台，以及在基础软件方面的领先优势和安全特性，在大数据、分布式存储、数据库和云平台等计算场景中构建安全可靠的算力底座；

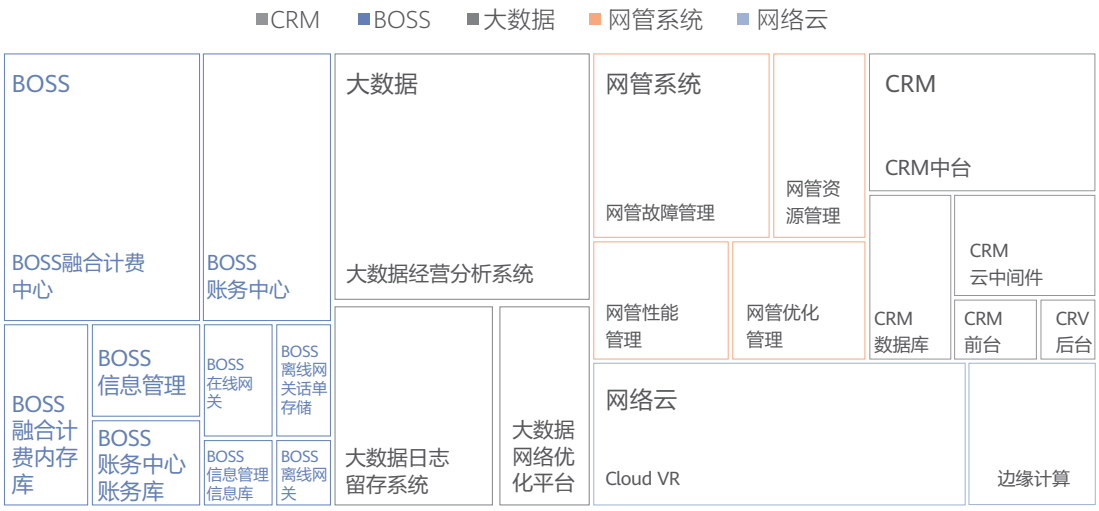


图1 电信行业主要计算场景

1.2.2 有节奏的开展现有应用适配、迁移，并基于ARM架构，持续开发原生应用

以电信行业为例，根据电信行业的业界专家评估绘制的《电信行业ARM架构迁移路径图》显示，ARM架构平均优势高，平均迁移难度较小，其中云核心网、大数据经营分析系统、大数据网络优化平台、CRM前台和中台、网关资源管理系统、网管性能管理系统、BOSS话单存储、Cloud VR等系统的ARM架构优势明显并且迁移难度偏低，均可优先考虑适配迁移。（图2）

在迁移过程中，针对行业应用跨架构迁移周期长、工作量大的问题，通过ARM架构配套的应用迁移工具，将代码修改、汇编语言翻译、兼容文件替换、编译调试、调优诊断等迁移关键步骤在工具辅助下自动完成，降低开发人员技术门槛、提升应用迁移效率，引导行业加快应用迁移进展。

迁移完成之后，在后续版本迭代及新功能开发过程中，通过ARM架构配套的开发工具，帮助开发人员便捷获取和使用ARM架构优势特性，开发出高性能软件，同时自动完成典型场景下的应用包构建和执行，提升开发效率和体验，引导开发人员持续基于ARM架构原生开发行业应用，深入构建行业软件生态。

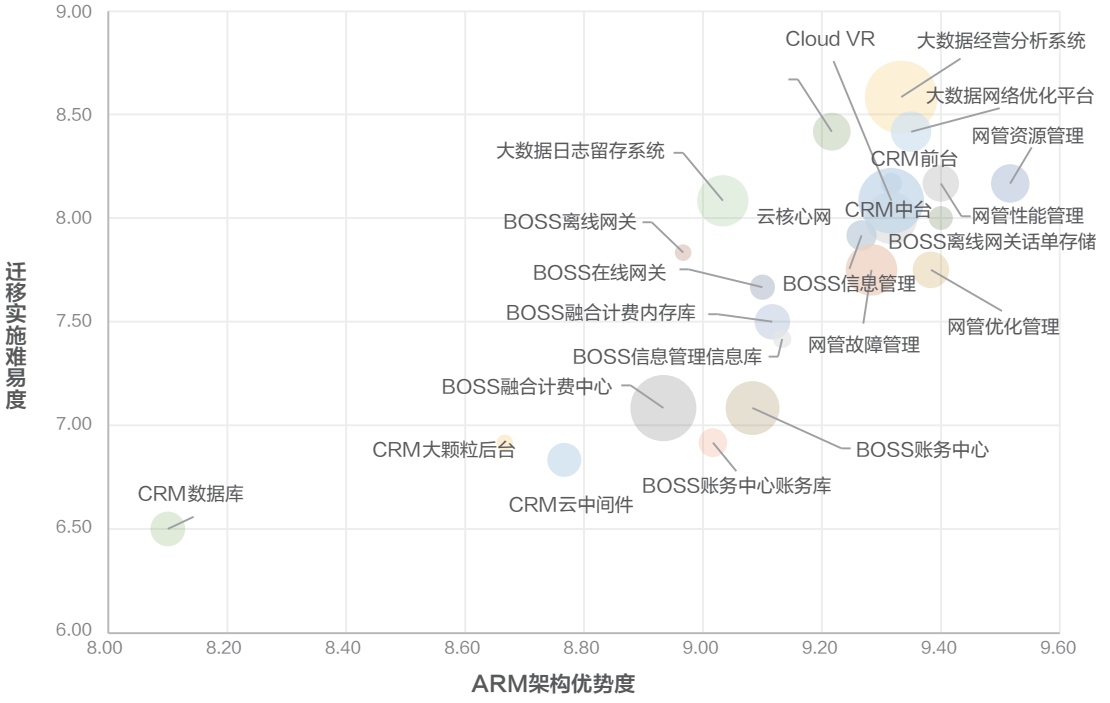


图2 电信行业ARM架构迁移路径图

1.2.3 通过全栈软硬件优化，充分释放多样算力，发挥极致性能

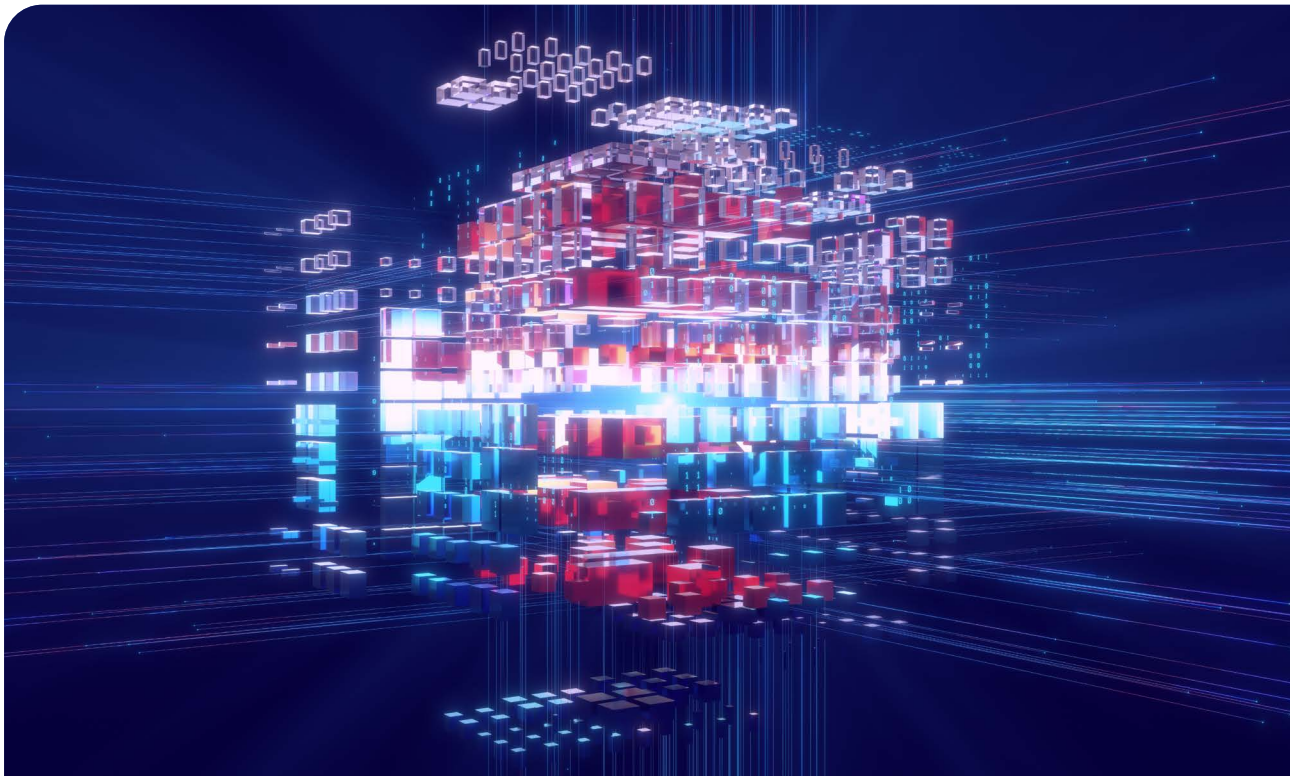
为了适应行业应用快速创新及多样性计算的需求，进一步提升软件运行性能，面向ARM架构的全栈优化能力必不可少，通过使用包括一系列的硬件加速库、软件加速包、开源加速组件、典型场景的性能优化解决方案等，围绕硬件、基础软件，到场景化应用开展全栈优化，充分发挥应用极致性能。

1）硬件加速：提供CPU、内存、磁盘、网络子系统等硬件基础性能优化参考，包括系统硬件配置优化方法及硬件加速库，消除性能瓶颈，提升硬件资源利用率；

2）软件加速：围绕系统指令、媒体转码、数学算法、存储网络等方向，提供一系列软件加速包，优化大数据加解密、分布式存储压缩、视频转码等常用软件性能；

3）基础软件优化：开源软件作为最重要的软件开发模式之一，是软件生态的核心，让开源软件与ARM平台进行充分的适配和优化尤为重要，持续在开源社区贡献关键性能优化成果，提供典型场景下的开源加速组件，让主流开源软件能够在ARM架构上发挥最佳性能；

4）典型场景优化：面向大数据、分布式存储和数据库等行业应用的典型计算场景，提供加速数据处理、优化存储访问和提升算力部署密度的场景优化方案，有针对性的提升行业应用性能；



解决方案

华为提供基于ARM架构的鲲鹏全栈基础软件平台解决方案

鲲鹏计算产业是基于鲲鹏处理器（基于ARM架构）的基础软硬件设施、行业应用及服务，涵盖从底层硬件、基础软件到上层行业应用的全产业链条。纵观鲲鹏计算产业生态全景，硬件方面，围绕鲲鹏处理器，涵盖包括智能网卡芯片、底板管理控制器（BMC）芯片、固态硬盘（SSD）、磁盘阵列卡（RAID卡）、主板等部件以及个人计算机、服务器、存储等整机产品。基础软件方面，涵盖操作系统、虚拟化软件、数据库、中间件、存储软件、大数据平台、数据保护和云服务等基础软件及平台软件。行业应用方面，鲲鹏计算产业生态覆盖政府、金融、电信、能源、大企业等各大行业应用，提供全面、完整、一体化的信息化解决方案。

鲲鹏计算产业从2019年正式起航，在全球鲲鹏计算产业伙伴的共同努力下，已经构筑了完整的基础软硬件生态和人才发展体系，并在各大国计民生行业实现了规模商用落地，为行业数字化变革和应用创新提供了强大稳定的算力支持。

作为鲲鹏计算产业的发起者和重要成员，华为秉持“硬件开放、软件开源、使能伙伴，发展人才”的策略，通过战略性、长周期的研发投入，吸纳全球计算产业的优秀人才和先进技术，和产业伙伴一起，持续推进全栈计算技术的创新发展，构筑面向多样性计算的全球开源体系与产业标准，推动鲲鹏生态全面发展。（图3）

截止目前，已发展10家鲲鹏整机伙伴，打造共计150余款整机产品。20+家用户/伙伴基于openEuler社区发行版，打造了自用或商用的操作系统版本；13家用户/伙伴基于openGauss社区发行版，打造了自用或商用的数据库版本；总计发展了超4000家鲲鹏合作伙伴，认证了12000+个鲲鹏解决方案，实现国计民生行业的核心应用场景全覆盖。

图3 鲲鹏全栈开放，使能全产业伙伴创新

1.3.1 鲲鹏主板开放，伙伴优先，使能商业成功

2019年华为面向伙伴开放基于鲲鹏处理器的主板、网卡、硬盘等标准部件，帮助整机合作伙伴快速推出自有品牌的服务器产品。

2020年华为发布了主板开放2.0，通过基础板+扩展板的开放模式，基础板沉淀共性，减少伙伴重复开发；扩展板实现创新，使能伙伴差异化竞争力；同时结合BIOS/BMC软件开放，支持伙伴自行开发差异化部件，打造创新整机产品。当前，鲲鹏主板走向更加开放，华为仅聚焦“CPU+内存”最小计算单元，通过全量组件化方式，实现从使能伙伴创新走向伙伴主导创新；

此外，在鲲鹏主板开放的同时，华为从研发、制造、采购&供应、服务、商业模式、解决方案、市场、人力资源、财务、文化十大方面，全方面对伙伴进行赋能，帮助伙伴快速成长，使能合作伙伴打造更有竞争力的鲲鹏计算产品。

市场上，华为践行伙伴优先，将自有品牌TaiShan服务器逐步退出市场，和伙伴不竞争，把市场空间让出来，支持伙伴商业成功，2022年1到10月，伙伴出货占比已达95%以上。（图4）

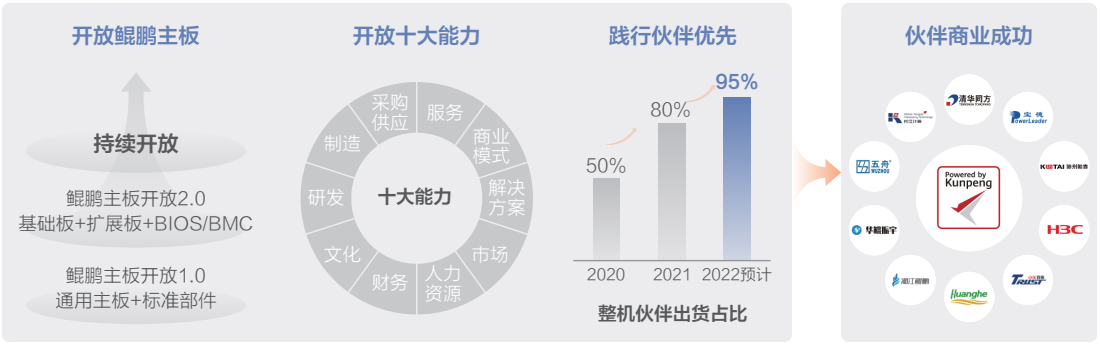


图4 硬件开放，伙伴优先，使能商业成功

1.3.2 基础软件开源，持续创新，实现最佳支持鲲鹏

基础软件方面，华为坚定开源，把自身多年来构建的操作系统能力和数据库能力开源出来，让合作伙伴能够在此基础上做增量开发，由此来提升中国的基础软件产业水平，和伙伴共建生态。并创建了openEuler开源社区（中文：欧拉开源社区）和openGauss开源社区，以社区运作的方式，同产业伙伴和广大开发者共同构建基础软件生态。

2021年底华为联合伙伴，把欧拉贡献给了中国开放原子开源基金会，从企业主导到产业共建，让欧拉快速跨越生态拐点。截止2022年11月，在中国服务器操作系统领域，欧拉新增市场份额已经达到22%，到2022年底预计将达到25%，有望加速成为中国新增市场份额第一。

当前无论是openEuler，抑或是openGauss，均在鲲鹏服务器上做了大量的性能优化工作，最终实现最佳支持鲲鹏，为鲲鹏生态的体系构建，奠定了基础。

以openGauss为例，通过NUMA-Aware优化，Inplace-Update融合引擎，多存储引擎架构，软硬协同优化等技术为用户带来多样化业务场景下极致、稳定的数据业务处理能力，在鲲鹏2路服务器上实现性能达150W TpmC，鲲鹏4路服务器上达230W TpmC，单节点处理能力业界领先，同时保持内核在高负载情况下性能抖动小于5%，业界稳定性最优。在2022年4月openGauss3.0版本，发布分布式解决方案，在性能方面持续精进，16节点性能达到1000万tpmC，领先目前竞品性能2倍。

1.3.3 使能极简开发，极致性能，繁荣应用生态

鲲鹏开发套件（鲲鹏DevKit，英文名：Kunpeng DevKit）使能应用极简开发

鲲鹏生态发展的关键挑战是应用软件迁移，为了帮助开发者加速应用迁移和算力升级，华为提供鲲鹏开发套件DevKit，包括代码迁移、开发调试、编译、测试、调优&诊断等一系列工具套件。

鲲鹏DevKit主要面向不同计算平台间的应用迁移以及鲲鹏平台原生开发，当前实现1-2人天/应用无忧迁移。2022年鲲鹏DevKit2.0聚焦原生开发能力增强，面向全研发作业流程，提供鲲鹏开发框架&场景化SDK、鲲鹏编译工具、鲲鹏调试器、云测服务、以及面向全场景性能分析和调优，让开发者更便捷高效的基于鲲鹏原生开发，效率提升50%+。（图5）

图5 鲲鹏DevKit：从“应用迁移”走向“原生开发”，开发效率提升50%+

鲲鹏应用使能套件（鲲鹏BoostKit，英文名：Kunpeng BoostKit）使能应用极致性能

鲲鹏BoostKit，从硬件、基础软件，到场景化应用开展全栈优化，主要面向伙伴和客户的开发者，提供高性能开源组件、基础加速软件包、应用加速软件包，使能应用极致性能。其中，高性能开源组件由伙伴从开源社区、鲲鹏社区获取，直接编译/部署，目前90%主流开源软件已支持鲲鹏，实现开源软件在鲲鹏上开箱即用。 基础加速软件包，面向伙伴开源、开放丰富的基础性能优化方法、加速库、加速算法，释放鲲鹏算力。应用加速软件包，联合伙伴开展解决方案创新，提供业界领先的加速组件、算法，实现应用性能倍增。

鲲鹏BoostKit 1.0面向鲲鹏聚焦的八大主力场景，把鲲鹏算力性能发挥到极致。在很多传统计算负载中，CPU的实际利用率并不高，大量有效计算能力浪费在等待数据的加载上。2021年全新推出的BoostKit 2.0，提供五大类“数据亲和”加速组件，包括数据就近计算，数据加速传输，数据并行化处理，数据安全等，对数据全处理流程进行负载优化，从而大幅提升应用性能。（图6）

图6 鲲鹏BoostKit：“数据亲和”五大加速组件，使能应用性能倍增

图7 极简开发，极致性能，繁荣应用生态

通过使能极简开发、极致性能，鲲鹏在国计民生行业的技术生态满足度从19年的仅9%，逐年稳步提升，22年底预计达70%以上，生态兼容性的瓶颈已基本消除，初步构建起繁荣的鲲鹏应用生态。（图7）

1.3.4 全栈协同，加速行业规模应用

鲲鹏与伙伴、开发者一路前行，全面进入国计民生行业核心应用场景。

在政府，鲲鹏与北明、超图、太极和神州软件等伙伴服务于各省市政务云；

在金融，鲲鹏携手长亮、麒麟软件、科蓝、华锐等伙伴服务于大行和金融机构的核心交易系统。

在电信运营商，鲲鹏和亚信、浩瀚深度、东方国信等伙伴服务于三大运营商的网络云、IT云与公有云。

在电力，鲲鹏携手南瑞集团、许继、麒麟信安和岳能科技等伙伴服务于国网、南网电力调度系统。

邮储银行基于鲲鹏全栈打造新一代分布式金融新核心

邮储银行作为拥有百年历史的金融机构，在中国有6亿用户，4万个营销网点，是国家普惠金融的主力，为国民经济发展做了突出贡献。现有的核心系统采用经典的大型机+商业软件搭建而成，支撑了邮储银行初期信息化，电子金融。但随着金融服务在线化，小额交易频次越来越高等这些服务场景的变化，对传统的核心系统带来了巨烈的冲击，尤其在交易热点时段，现有系统弹性不足，造成交易缓慢。商业软件架构与技术封闭，迭代慢，在应对金融创新乏力。无法继续支撑邮储银行向前发展。

因此，邮储银行从19年初开始启动下一代金融核心的预研，为了保持持续创力的能力和可

能，邮储银行决定基于通用计算平台加开源软件技术构建分布基础IT能力。整个不仅保持灵活的资源扩缩容能力，还具有丰富的开源软件生态，使未来的技术获取等方面成本更低。同时邮储采用企业级业务建模，对邮储上千种业务进行抽像建模，使用业务逻辑关系更清晰。

同时基于鲲鹏服务器和openGauss的原型验证，结果超越客户预期。并引入了微服务，容器等业界先进成熟的技术。经过一年多的建设，并于21年4月18日技术平台上线，开始接入生产系统进行镜像验证，于6月上线分布式运维系统，利用AI技术解决海量节点带来的运维复杂度。系统于22年3月份全量投产，支持邮储日均20亿笔的交易和未来10年的业务发展。

邮储银行是国内**首个建成新一代个人业务新核心的国有大行**，证明了鲲鹏和openGauss在金融这种的对可靠性和性能要求极高的场景，不但可以胜任， 而且可以很好，鲲鹏的多核、高并发，结合openGauss高性能、高可用及智能运维等内核能力，助力邮储个人新核心业务处理能力5倍提升，支取和查询等核心业务场景的性能提升25%以上， 这些数据都可以提升客户的使用体验与感知，提升满意度，加强邮储银行服务竞争力。

邮储银行通过分布式金融新核心建设，在金融服务技术上已走到同行前列，相信凭借邮储银行人的勇于开拓的创新精神，未来会持续领先，为同业树立新的标杆和为用户带来更好的服务。

趋势二

数字化走向深入，操作系统走向多样性算力和全场景的协同

数字化走向深入，操作系统走向多样性算力和全场景的协同

操作系统作为计算产业中最基础的软件，承担着抽象底层硬件，向上层应用提供统一接口的核心功能，是计算产业的关键环节。面向多样性计算和海量应用场景，操作系统应支持多样算力和多种应用的协同，成为数字产业的可靠软件底座。

产业趋势

在IT产业的全栈系统中，处理器是硬件的基础，操作系统是所有软件的根基。（图8）

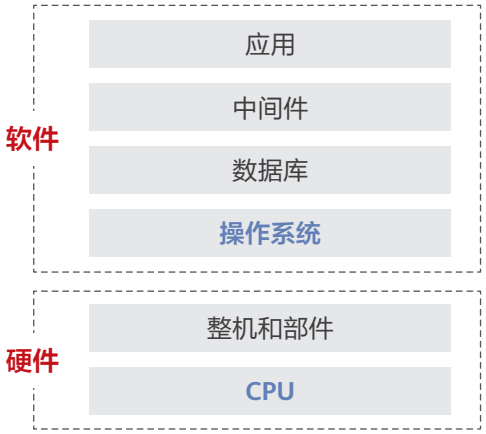


图8 操作系统是软件的根

操作系统作为连接底层基础硬件（处理器，整机/部件）和上层应用的最基础软件，被称为IT产业的魂：硬件提供算力的供给，应用软件是算力价值的实现，而操作系统则完成算力释放。一方面，操作系统面向硬件系统，提供更好更高效的硬件资源管理能力；另一方面，操作系统面向应用和用户，沉淀应用领域共性，提供更为便利易用的人机交互。

2.1.1 多样性计算时代，呼唤面向数字基础设施的操作系统

计算产业从通用计算已经进入到通用计算+AI计算的多样性计算时代。多种算力协同发展，对操作系统提出了新的要求。

首先，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，要完成不同计算架构、不同硬件的兼容适配，提供良好的兼容性，为应软件用的部署提供尽可能的便利。

其次，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，提升性能。比如，基于ARM架构的处理器，其典型特征是核数更多，这使得ARM处理器在

高并发应用场景，更具竞争力。因此，操作系统需要针对多核的处理器进行优化，确保多核任务并发时的任务调度更加合理，避免任务冲突，提高系统整体性能。

此外，除了针对不同架构的CPU优化，CPU和GPU、NPU等其他特定用途之间处理器之间的协同，也是影响系统效率的关键因素。操作系统层面，在处理CPU任务和GPU、NPU任务时，协调好这些任务的调度，成为必要的能力。

2.1.2 数字化走向深入，操作系统面向云管边端全场景应用协同发展

随着云计算的快速发展，云计算和云服务已经成为各企业进行数字化转型的优先选择。无论是高科技行业还是传统行业，无论是大企业还是小企业，都可以通过云服务随时随地获取数字化转型所必需的计算、存储等硬件资源，大数据、AI、IoT 等技术资源，以及凝结了领先企业大量投入的经验知识资源，极大提升了企业运行效率。

云上的应用与其他场景的应用协同场景越来越丰富，比如AI应用在云和边缘的协同。通过云端充足强大的算力进行AI训练，而且云端能很好的支持多种不同的服务和AI框架，此外云端可以简化训练的开发，无需软件下载、无需配置、无需安装。边缘端则利用靠近数据产生和采集的优势，在边缘端可以迅速把采集的数据拿去做推理，快速得到推理结果的同时，避免了向云端传输大量数据带来的高成本。

因此，操作系统通过在软件底层实现应用在云、管、边、端、数据的高效、可靠、安全交换，是可以大幅提升系统整体效率和安全性的。

2.1.3 开源成为主流软件开发模式，操作系统开源共建成为产业共识

开源已经成为主流软件开发模式。从全球范围来看，过去一年，开源整体呈现高速发展的趋势。据最新官方报告，2021 年全球最大开源代码平台GitHub活跃用户数和活跃代码仓库数量均有明显增长，其中新增活跃用户数超过 1600 万、新增活跃代码仓库数量超过 6100 万。中国开源贡献者占比明显提升，从2015年的7%的占比，快速提升至2021年的11%。开源模式越来越成为全球软件技术和产业创新的主导模式。

同时，开放开源是软件技术创新，特别是发展操作系统这类基础软件的重要途径，充分利用开源，参与开源，支持开源，发展操作系统，联合做大做强是当前最为可行之路。构建根植于中国的开源社区，培养良好的土壤和与环境，可以为产业打造可持续发展的创新之地。

行动建议

2.2.1 规划部署支持数字基础设施多样算力的操作系统，使能全场景应用协同创新

通过规划部署支持不同应用场景、支持多样性算力的统一数字基础设施操作系统，打通不同硬件架构和多种场景应用，实现更优的性能，业务更好的协同。

在企业的各类数字应用场景中，通常部署了各种不同类型的计算设备，典型的包括服务器、边缘设备，嵌入式等等。不同设备安装各类不同的操作系统，给整体系统运行运维带来挑战；设备间的互联互通复杂度也因此显著提升；不同应用之间的可靠、安全的交互，协同相对繁琐。

统一的数字基础设施操作系统，可以实现从操作系统底层完成设备间的连接、数据交互，从而大幅提升运行运维效率。

2.2.2 分析应用迁移策略，制定应用迁移计划，完成应用高效迁移

部署新的操作系统，应选择具备可持续演进性、基础兼容性和支持应用快速迁移的能力的技术路线。可持续演进，是指除了可靠、稳定、安全等基础能力外，所选择的技术路线有具备独立维护、长期演进的机制和能力；基础兼容性，是指在操作系统南向各类处理器、整机、板卡的兼容性支持，以及北向的各类应用的适配性；应用迁移能力，是指需要提供包括兼容性识别、应用迁移与调优，系统测试等全流程的自动化工具和技术支持文档。

操作系统迁移是一个系统工程，包括从技术路线选型、系统分析、方案设计、移植适配、迁移实施和测试上线等全流程，因此需要组建合理的团队、详细的计划，有节奏分阶段实施，在确保业务持续稳定运行的情况下，有序开展。

2.2.3 加入开源操作系统社区，积极拥抱开源、回馈开源

通过主动加入开源社区，与社区核心组织和成员的运作与沟通，保持与业界各类领先技术的同步，可获取最新的技术趋势、业务方向以及关键支撑。企业、高校、操作系统厂商等组织单位加入操作系统开源社区，加强交流，合作共建共赢，共同发展。

更为重要的是，操作系统开源社区提供各类开发工具，硬件资源，技术指导以及各类在线服务，企业鼓励有能力的开发人员加入社区，获取最新的开发手册、技术补丁以及开发平台，可以大幅度提升应用开发效率。成熟的开源社区，除了提供日常代码审核、提交的工具之外，还包括大量满足开发全流程所需要的各类资源，同时社区是技术人员积累能力、了解技术趋势、解决技术难题，互相交流的平台。

解决方案

华为是全球领先的ICT（信息与通信）基础设施和智能终端提供商，在ICT领域提供包括服务器、存储、云服务、边缘计算、基站、路由器、工业控制等各类产品和解决方案。在多年的全系列产品研发过程中，不断累积软件根技术，全面布局操作系统等基础软件，满足自身业务发展需要。

2019年12月，华为创立欧拉开源项目，通过开源的方式，把积累的操作系统能力开放出来，携手产业伙伴共同发展操作系统产业，得到了产业积极响应。

目前，欧拉开源操作系统发展迅速，生态快速构建，已累计实现超过245万套装机，国内新增市场份额超过22%，跨越生态拐点，成为企业数字化转型、应用创新、构筑安全可靠操作系统的首选技术路线。

2.3.1 欧拉，面向数字基础设施的开源操作系统

欧拉（英文：openEuler）是面向数字基础设施的开源操作系统，支持服务器、云计算、边缘计算、嵌入式等应用场景，支持多样性计算，致力于提供安全、稳定、易用的操作系统。通过为应用提供确定性保障能力，支持OT领域应用及OT与ICT的融合。（图9）

欧拉持续丰富南向多样性设备支持，北向使能IT、CT和OT全场景应用。

当前欧拉已经实现主流计算架构100%覆盖，支持包括ARM、x86、RISC-V等全部主流CPU指令集，同时支持NPU、GPU和DPU等多种异构算力，适配超过100款整机、300款板卡，成为最佳支持多样性算力的开源操作系统。

在北向应用生态上，与伙伴协作，适配了一万多款应用，主流应用场景100%支持，满足各行业不同应用需求。

图9 欧拉，面向数字基础设施的开源操作系统

统一操作系统，支持多设备

通过一套操作系统架构，支持多样性设备。

欧拉采用全量组件原子化，支持内核灵活组合，全栈服务化按需构建，可以根据设备不同的资源能力和业务需求灵活裁剪，按需构建不同的操作系统版本，满足不同设备对于操作系统的要求。

同时，欧拉支持构建服务自助化，支持“菜单式”配置内核和系统服务，可以针对软件包、文件级、函数级的不同层级分级灵活组合，自动化、简化版本构建。

进一步，欧拉还提供多设备协同套件，来实现不同设备间的能力互助和资源共享。

应用一次开发，覆盖全场景

欧拉通过一套标准API，ICT+OT全场景提供统一API，这样就实现了操作系统与应用之间交互语言的统一；

同时，通过欧拉SDK，把各种应用所需数据能力、音视频能力、安全等能力，进行统一封装，使能极简开发；

欧拉Devkit开发套件，还可以方便的集成到各种主流应用开发平台。

欧拉与鸿蒙能力共享，生态互通

欧拉是数字基础设施开源操作系统，鸿蒙是面向万物互联的智能终端操作系统，欧拉和鸿蒙，进一步打通，就可以更好地服务数字全场景。

欧拉和鸿蒙已经实现了内核技术和分布式能力共享。通过共享分布式套件，实现了欧拉和鸿蒙的互通，两大开源操作系统打通，欧拉覆盖云管边，鸿蒙覆盖端，欧拉+鸿蒙共同服务全场景数字应用。（图10）

未来进一步在安全OS、设备驱动框架、以及新编程语言等方面实现共享，通过能力共享，实现生态互通。

图10 欧拉与鸿蒙能力共享 生态互通

2.3.2 欧拉开源共建，已构建成熟的产业生态

欧拉开源项目

2019年12月，华为创建欧拉开源项目，成立欧拉开源社区（https://openeuler.org/），开源代码上线。

2021年11月，在操作系统产业峰会2021上，在“政产学研用”各方代表的共同见证下，华为携手社区全体伙伴，将欧拉开源操作系统全量代码、品牌商标、社区基础设施等相关资产贡献给中国开放原子开源基金会，具体包括：华为自己开发的数百万的自研代码，它的版权和知识产权许可，超过8000多个经过华为和社区验证的软件包，openEuler以及相关项目的中英文的商标品牌共30多个，域名4个，以及构建服务与测试体系代码托管，社区运营平台等社区基础设施。这实现了欧拉开源操作系统从企业主导到产业主导的重要转变，有利于促进欧拉开源项目从开放治理走向自治繁荣。（图11）

图11 欧拉开源操作系统捐赠签约

欧拉欧拉生态繁荣，欧拉社区已成为国内最具活力开源社区

截止目前，国内外10+家主流操作系统厂商（麒麟、统信、麒麟信安、SUSE等）均已发布了欧拉路线的操作系统商业发行版；社区当前已有超过400家企业加入，汇聚了从处理器、整机、到基础软件、应用软件、行业客户等全产业链核心伙伴；社区已经有超过1万名开源贡献者，创立近100个SIG组（特别兴趣小组），社区维护的核心软件包达到8000多个。

欧拉创新领先的技术，良好的硬件兼容性，丰富的应用软件生态，和覆盖全场景的部署能力，为欧拉的规模部署提供了充分的条件。

截至目前，欧拉技术路线的操作系统，已经在数字政府、电信、金融、电力等多个行业实现大规模部署，应用在核心系统中，为各行业提供稳定、可靠的数字根基，累计部署量超过245万套，国内新增市场份额占比达到22%，在数字政府、金融行业增速第一。

部署欧拉路线操作系统的用户包括三大运营商（中国移动、中国电信、中国联通）、两大电网（国家电网、南方电网）以及多个大型国有和商业银行（建设银行、工商银行、中信银行、中国银联等）等。典型应用案例包括：中移在线大数据平台、中国电信云平台、国家电网核心调度系统、中国建设银行信用卡核心系统等等。

2.3.3 欧拉，以发展根技术引领操作系统创新

欧拉以内核级创新，打造最佳多样性算力支持、全场景数字基础设施操作系统，成为企业数字化转型、应用创新的首选可靠操作系统技术路线。

欧拉引领操作系统内核创新。作为社区主要成员，华为自 2012 年以来向 Linux Kernel 社区贡献，在 Linux Kernel 5.10版本中，华为内核代码贡献排名第一。

欧拉最佳支持多样性算力。支持鲲鹏、x86、飞腾、龙芯、申威、RISC-V等多种处理器架构，并且性能相比主流操作系统更佳。

欧拉打破不同场景操作系统生态壁垒，成为首个全场景数字基础设施操作系统。欧拉统一支持服务器、云计算、边缘计算、嵌入式等等应用场景。（图12）

截止目前，欧拉已经发布2个LTS（长生命周期支持）版本和4个创新版本。

华为不做欧拉商业发行版，通过社区使能伙伴商业发行版、企业自用版、社区发行版等多种形式，促进操作系统产业健康、高速发展。华为持续在欧拉开源项目贡献，包括技术创新、社区运营、生态建设等。华为联接、计算和云等各领域继续全面使用欧拉技术路线，以社区版本为基线，构筑华为自用操作系统版本。

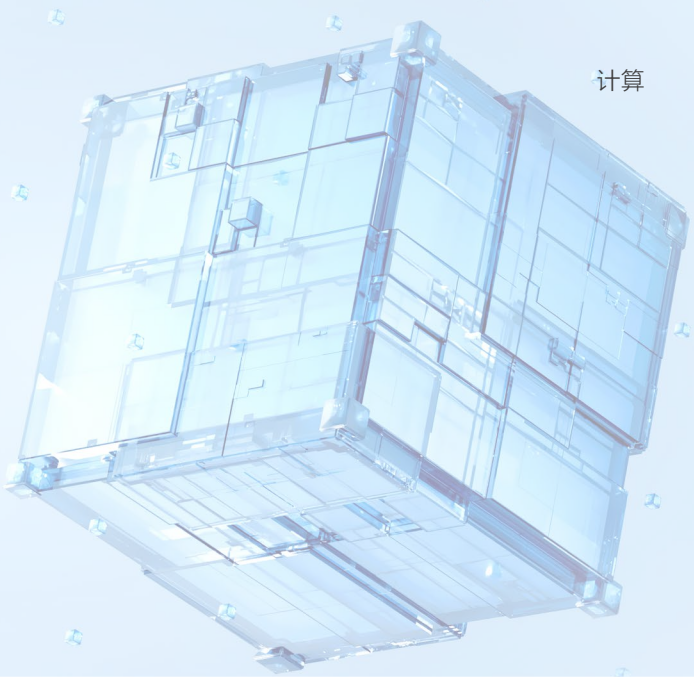
欧拉技术路线的操作系统，主要包括以下集中形式：

- **社区发行版：**由欧拉社区成员和社区开发者共同构建发布的开源操作系统版本，以免费的形式通过社区提供。社区每2年发布一个长周期（LTS：Long Term Support）版本，比如：openEuler 20.03 LTS版，openEuler 22.03 LTS版。
- **商业发行版：**操作系统产业伙伴（即OSV），结合各自的优势，基于欧拉的社区版，开发自己的商业发行版操作系统，面向最终用户提供和销售有竞争力的产品。比如麒麟软件有限公司的银河麒麟高级服务器操作系统V10、统信服务器操作系统V20（1020e），麒麟信安操作系统V3（欧拉版），SUSE数硕Linux等。
- **企业自用版：**具备自研能力的企业，基于欧拉的社区发行版，开发自用的操作系统版本（非独立销售或不销售）。比如华为公司通信设备搭载的自研操作系统、中国移动BC-Linux for Euler、中国电信CTyunOS，中国联通CULinux、百度 Linux 智能云操作系统等。

图12 欧拉已发布2个LTS版本和4个创新版本

趋势三

数字经济发展引发算力需求爆炸式增长，人工智能算力增长是主要增量



宏观趋势

3.1.1 数字经济飞速发展将催生强劲算力需求，人工智能算力是主要增量

当前，数字经济正在成为全球经济的主要增长点，算力作为数字经济时代新的生产力，是支撑数字经济发展的坚实基础。算力已成为全球战略竞争新焦点，是国民经济发展的重要引擎，全球各国的算力水平与经济发展水平也呈现出显著的正相关。

与此同时，人工智能技术的不断发展带来了远远超越摩尔定律的算力需求。从2011年深度学习技术兴起到今天，对人工智能算力的需求一直是指数级增长的，每隔3-4个月算力需求翻一倍。2020年，自然语言处理模型GPT-3参数量达到1750亿，算力需求是3640PD（PD代表以千万亿次每秒的算力计算一天所用的浮点计算量）；2021年，鹏程.盘古——业界首个全开源 2000亿参数中文预训练语言模型，使用E级AI算力的鹏城云脑II算了50天，算力需求达到了 25000PD；到2023年，这种大模型的算力需求能到百万PD，这就对现有计算处理能力提出了严峻的考验。

在蓬勃的需求带动下，全球算力发展水平正在持续扩大，而在这其中，人工智能算力成为主要增量。华为预测，到2030年，人类将进入YB数据时代，全球通用算力将增长10倍达到3.3 ZFLOPS(FP32)，人工智能算力将增长500倍，超过105 ZFLOPS(FP16)。

3.1.2 人工智能正日益快速渗透行业应用的核心场景

人工智能技术的落地为行业带来更多价值，不仅提高了企业的运作效率、生产效率，还推动了企业创新的能力。调研发现，采用人工智能三年以上的企业，已经在多方面获得显著的收益，实现收入增加和生产效率提升。在2021的行业调研中，TOP3的行业人工智能渗透度均超过了50%，最高的渗透度甚至超过了80%。

人工智能将在城市、交通、制造、能源、医药、教育、农业等行业持续渗透，为衣食住行带来更智能的体验：

在城市领域，城市的智慧化治理成为实现城市可持续发展的必然选择。未来，每一个物理实体都将有一个数字孪生，如城市楼宇、水资源、基础设施等将组成城市数字孪生，实现更加智能的城市管理。城市智慧治理将带来100倍的社会数据聚集，人工智能技术将实现高效城市治理。

在交通领域，预计2030年，全球道路上的电动汽车、面包车、重型卡车和公共汽车数量将达到1.45亿辆。每辆汽车行驶中产生的数据需要在汽车与城市之间频繁进行数据交换，借助智慧交通基础设施的AI分析能力，城市通勤时间将得到大幅提升（日均通勤缩短15-30分钟），交通事故和汽车对城市碳排放量也随之大幅降低。智能带来的交通安全、效率、体验的提升，必将释放出新的生产力，推动社会经济的发展。

在制造领域，AI可以帮助制造企业实现智慧化运营管理、海量数据分析挖掘以及低时延诊断预警。中国制造2025提出，制造业重点领域全面实现智能化，试点示范项目运营成本将降低50%，产品生产周期缩短50%，不良品率降低50%。

人工智能将融入千行百业的核心场景，实现多个人工智能场景的落地，带来源源不断的创新与无所不及的智能。

行动建议

3.2.1 加速AI基础设施建设，让AI算力成为像水和电一样的公共资源

作为新基建的重要组成部分，人工智能已经成为数字经济发展的重要驱动力，人工智能产业在发展过程中仍面临诸多挑战：

1）人工智能产业布局不均衡：通过对多个城市的实际调研发现，人工智能产业存在基础薄弱和研发实力弱等情况，无法匹配产业发展规划和战略布局。

2）企业使用AI算力成本高：大规模预训练模型的参数量越来越大，需要的算力也从TFLOPS级别增加到PFLOPS级别，多数企业表示当前算力不足。企业使用人工智能算力成本昂贵，仅算力成本就占据企业开发成本的15%～25%。

3）AI人才缺口大：大量人工智能企业表示缺乏AI技术人才，在全国各地都缺乏核心AI技术人才的背景下，加强培育本土AI人才非常必要。

鉴于上述挑战，建议大力发展以人工智能计算中心为代表的新型基础设施，让AI成为水和电一样的基础公共资源，为数字经济发展提供新动能。

人工智能计算中心建设具有技术实现复杂、建设周期长、资源投入巨大、产业辐射面广的特点，需要进一步强化战略统筹和政策保障，进行系统的组织机制和体制创新，加强关键核心技术攻关和标准化建设，以加快推动人工智能计算中心的高质量发展和网络化建设。系统总结已建成的人工智能计算中心的建设经验，持续加强人工智能计算中心的统筹建设，在确保

已建成的人工智能计算中心保持高效运营的同时，顺应人工智能发展趋势和产业落地的需求，坚持以应用为导向，坚持自主创新技术路线，加强人工智能计算中心建设。

1）持续推进计算中心高效运营和可持续发展。要强化洞察人工智能产业发展现状、调研算力需求的能力，实施算力普惠政策，为行业用户及应用开发企业、科研机构、高校提供普惠算力服务等功能。联合产业组织面向人工智能应用场景的项目机会清单，面向人工智能企业、高校院所、科研机构进行公开发布，鼓励开展人工智能先导性应用开发和场景试验，牵引科技创新成果进行商用转化，打造一批有影响力、有实际效果的应用示范项目，形成围绕大模型的产业集群，进一步带动产业的智能化升级。

2）坚持自主创新技术路线与推动开放开源并重。在当前日益复杂的国际竞争环境下，在推动人工智能计算中心建设的过程中，要继续坚持自主创新，进一步强化政策支持，广泛吸纳产学研用各方参与，共同提升相关产业链供应链现代化水平。同时，要以积极开放的态度拥抱开源开放，在全球范围内推动形成共建共享的人工智能算力与创新生态。

3.2.2 加速人工智能进入行业关键场景，使能行业智能化升级

促进人工智能与各行业融合创新，在城市、交通、制造、能源、医疗、金融等重点行业和领域开展人工智能应用试点示范，推动人工智能规模化应用，全面提升产业发展智能化水平。

智慧城市。构建适用于政府服务与决策的人工智能融合赋能平台，实现AI在智慧城市建设中“大脑”般的智慧，将人工智能技术与城市应用场景深度融合，实现城市在各类场景下的高效治理。研制面向开放环境的决策引擎，在复杂社会问题研判、政策评估、风险预警、应急处置等重大战略决策方面推广应用。

智慧交通。针对高速公路自由流收费、收费稽核、视频云联网、车路协同等典型交通AI应用场景，打造智慧交通解决方案，用人工智能技术对车辆、轨迹等进行智能分析，让出行管理更高效，让通行更通畅。

智慧制造。打造数字工厂AI使能解决方案，为制造行业量身定制的质量检测、厂区安全等应用领域的一站式、高精度、支持快速换线、开箱即用的AI解决方案，打通AI落地制造行业的“最后一公里”，加速AI应用在工厂规模化部署，把AI带入每一条产线，为工厂生产和运营提质增效。

智慧巡检。用人工智能的分析取代传统的人工巡检，让巡检更安全，效率和准确率更高。结合智能电网、智能油气和智能矿山的发展需求，以AI技术为基础，打造智慧巡检解决方案，为输电线路、变电站、配电房、油气田、加油站和煤矿等场景提供区域智能感知。

智慧医疗。打造传染病AI监测预警平台、紧密型县域医共体AI解决方案、智慧医院AI 解决方案等，助力医疗行业智能化升级，将AI科技进步服务于人类健康。

智慧金融。面向金融行业提供更加高效、安全、个性化的综合性金融解决方案，贯穿于金融服务垂直全流程，为银行智慧网点、金融OCR、智能双录等AI应用场景提供智慧化解决方案。

解决方案

3.3.1 以昇腾AI基础软硬件平台，构筑智能根基

昇腾AI产业是以昇腾AI基础软硬件平台为基础，坚持“硬件开放、软件开源、使能伙伴、发展人才”，联合技术和商业伙伴，打造“共建、共享、共赢”的人工智能产业，致力于让AI“用得起、用得好、用得放心”，以人工智能赋能社会发展与产业升级，为人类社会发展带来价值。

昇腾AI产业(简称昇腾AI/昇腾)是以昇腾AI基础软硬件平台为基础构建的人工智能计算产业。

昇腾AI基础软硬件平台包含Atlas系列硬件及伙伴硬件、异构计算架构CANN、全场景AI框架昇思MindSpore、昇腾应用使能MindX、全流程开发工具链MindStudio和一站式AI开发平台ModelArts等。

1）Atlas系列硬件及伙伴硬件

基于昇腾AI处理器，通过模组、板卡、小站、服务器、集群等丰富的产品形态，打造面向“云、边、端”的全场景昇腾AI基础设施解决方案。

2）异构计算架构CANN

异构计算架构CANN（Compute Architecture for Neural Networks），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。

3）全场景AI框架昇思MindSpore

全场景AI框架昇思MindSpore，致力于成为全球主流AI框架，具备一次开发云边端全场景部署、原生支持大模型训练、支持AI+科学计算等关键特性，加速科研创新和产业应用。

4）MindX昇腾应用使能

昇腾应用使能MindX包含深度学习使能MindX DL、智能边缘使能MindX Edge、模型优选库ModelZoo，和行业应用开发套件MindX SDK，旨在沉淀行业知识，使能行业应用极简开发，加速人工智能应用落地。

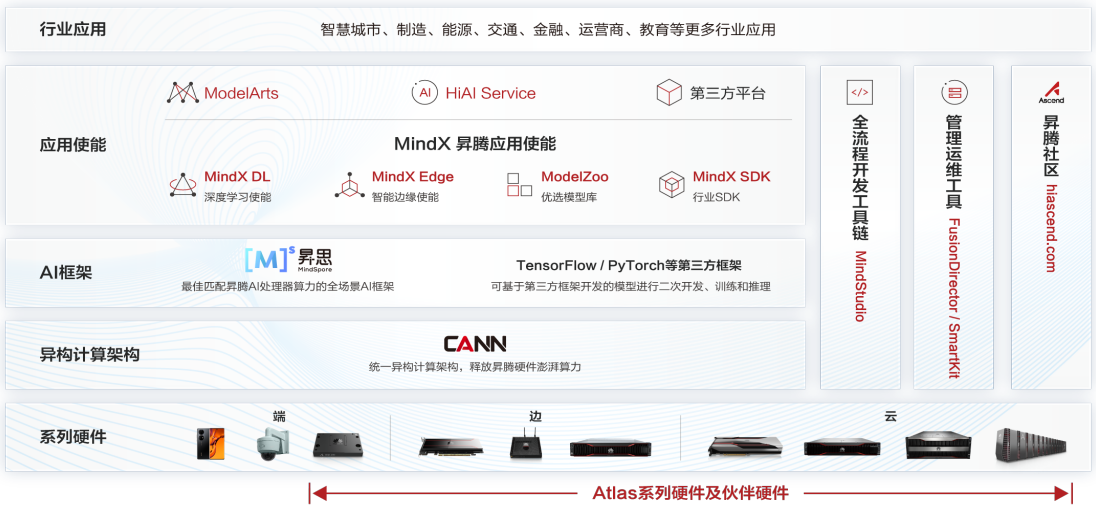


图13 昇腾AI产业

3.3.2 以“一中心四平台”建设人工智能计算中心，打造人工智能算力基础设施

人工智能计算中心（AICC，AI Computing Center）是专注于AI计算的新型城市基础设施，它以昇腾AI基础软硬件平台为基础，是涵盖了从基建基础设施、硬件基础设施到软件基础设施的完整系统。

作为一体化城市人工智能新型基础设施，AICC承载着“一中心四平台”的产业模式创新，解决算力普惠、科研创新、应用孵化与落地、人才培养等AI发展关键问题，旨在让AI算力像水和电一样成为城市公共基础资源，为数字经济发展提供新动能，让智能无所不及。

公共算力服务平台：通过产业政策牵引，将人工智能计算中心的算力资源有序、高效、普惠地开放给当地的企业、科研机构和高校，解决当地AI技术发展和产业智能升级的算力和服务需求。

应用创新孵化平台：各地AI企业、高校、科研机构，针对各地特色的AI应用场景项目机会，依托人工智能计算中心，进行科技创新成果商用转化、形成有本地特色的的重大产品创新和示范应用。

产业聚合发展平台：依托计算中心，配套相关产业政策、吸引和招募AI产业链上的各类公司（算法公司、数据处理公司、行业集成公司等）入驻形成完整产业闭环，促进和推动AI产业集约集聚发展。

科研创新和人才培养平台：基于人工智能计算中心充沛的算力资源，促进高校院所联合行业龙头企业，围绕产业技术创新需求，开展人工智能技术研发、科技成果转化等重点工作，落地科技创新成果的、培养关键人才。

当前，在国家统筹规划下，已有20多个城市在规划和建设人工智能计算中心，华为也积极参与其中。深圳、武汉、中原、西安、成都、南京、杭州、沈阳、青岛、重庆已相继上线或试运营，已经累计为1200+企业、120+高校、70+科研单位提供了算力服务。

深圳“鹏城云脑II”于2020年10月正式上线，实现上线即饱和运营，其三项打榜获得世界第一：2021年7月，在IO500排行榜中，蝉联全系统输入输出和10节点规模系统两项世界冠军。其中，全系统输入输出性能超越第二名近20倍，至今仍保持榜单第一。2021年11月，在AI Perf500排行榜中，保持世界第一。依托鹏城云脑II E级的澎湃算力，鹏城实验室与华为联合研发了全开源开放的两千亿参数中文NLP大模型鹏程.盘古，以及赋能生物医药探索的大模型鹏程.神农。

图14 鹏城云脑II机房内部

武汉人工智能计算中心基于昇腾AI基础软硬件建设，于2021年5月31日正式竣工并投入运营，上线即算力资源满负荷使用。于2022年2月7日完成扩容，总算力达200P，并再次饱和运营。率先实践“一中心四平台”，开创“武汉模式”。5个月从进场施工到正式投运，让业界见证了“武汉速度”，打造了全国人工智能示范标杆。目前，基于武汉人工智能计算中心，孵化了全球全球首个三模态大模型——紫东.太初， 全球首个遥感影像智能解译专用框架——武汉.LuojiaNet，业界最大遥感影像样本数据集——武汉.LuojiaSet，并成立多模态人工智能产业联盟和智能遥感开源生态联盟，为武汉孵化数百亿级智能遥感和多模态产业（大于300亿），截止到2022年9月底，已服务企业120+，孵化AI创新解决方案130+。

图15 武汉人工智能计算中心

3.3.3 产学研携手，共筑人工智能产业生态

华为开放昇腾AI基础软硬件平台，包括Atlas系列硬件、异构计算架构CANN、全场景AI框架昇思MindSpore、昇腾应用使能MindX以及一站式开发平台ModelArts等，帮助伙伴和开发者高效使用AI能力，创新场景化AI应用，加速千行百业智能升级。

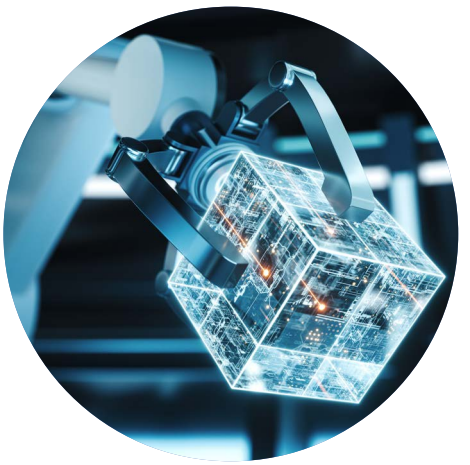
完成昇腾AI生态的初步构建，目前发展了100万+开发者，在100多所高校开设昇腾AI相关的人工智能课程，发展700+行业合作伙伴，共同孵化了超过1600+解决方案，为中国人工智能产业繁荣提供一个强健、稳固的基石。

全场景AI框架昇思MindSpore是业界首个全自动并行的框架，且具备全场景协同和全流程极简的特点。华为于2020年3月28日开源昇思MindSpore框架，开源后获得国内外开发者的积极响应，访问量数千万，超过320万用户下载安装使用，在码云千万开源项目中综合排名第一，服务企业数量超过5500家，高校授课数量超过140所，超过40所科研机构选择昇思进行科研创新，社区贡献者达8000+，ModelZoo支持模型350+，获得业界首个AI框架类产品级CC安全认证和AI可信开源社区认证，成为国内最具创新活力的AI开源社区。

昇腾众智计划是华为围绕昇腾基础软件平台推出的一项生态合作计划，旨在汇聚高校、科研院所、企业等组织和机构的开发团队，通过项目合作方式，基于昇腾基础软硬件平台开发算子、网络模型及行业参考设计，不断丰富昇腾计算产业生态，为加速千行百业智能化升级贡献智慧与力量。目前，通过昇腾众智计划，已经完成4000多个AI模型、算子等。2022年，将继续投入2亿人民币激励基金，推出超过4000个众智任务。

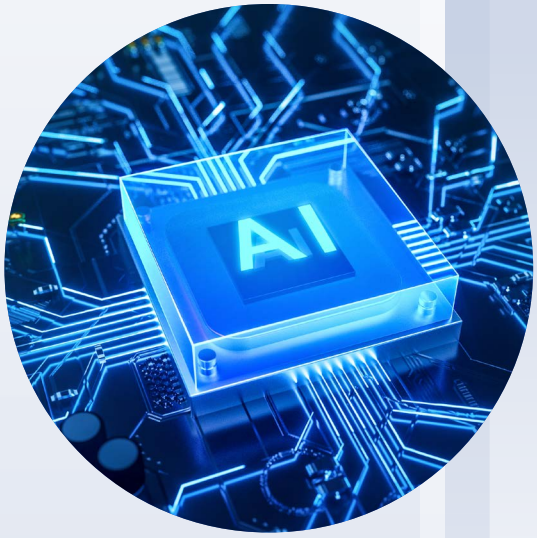
此外，在人才培养方面，教育部-华为“智能基座”产教融合协同育人基地项目由教育部、华为于2020年底联合发起，首批布局72所高校。

华为联合72所高校持续深化“智能基座”项目，在理工科专业深入实践，产教融合，把鲲鹏、昇腾、欧拉、高斯、昇思等根技术融入高校教学。目前，已赋能3000多个老师，开设1500多门课程，覆盖了30多万学生，成立了72个智能基座社团，出版约20本教材教辅书籍和12门精品慕课，并推出“智能基座”优秀教学资源奖励计划，激励更多教师百花齐放，自主开发教材和慕课。华为联合教育部已建设17个教育部智能基座课程虚拟教研室。



趋势四

大模型成为 AI 规模应用重要途径，科学计算正在进入科学智能新阶段



宏观趋势

4.1.1 “大算力+大数据”正在催生大模型的快速发展，孵化系列行业新应用

当前人工智能领域，大规模预训练模型得到长足发展和广泛关注，以大数据和大算力优势取代了一些小的算法模型，“大模型+大数据+大算力”成为迈向通用人工智能的一条可行路径。以GPT-3为代表的超大规模预训练模型，展示了一条通向通用人工智能的可能方向。

在此背景下，我国超大规模预训练模型的发展如火如荼。2021年以来，国内相继发布了一系列大模型，华为与鹏城实验室联合发布了“鹏程.盘古”系列超大规模预训练稠密模型，中科院自动化所发布了全球首个三模态大模型“紫东.太初”，以及北京智源人工智能研究院发布了“悟道2.0”稀疏模型等。

人工智能大模型可以实现在众多场景通用、泛化和规模化复制，减少对数据标注的依赖。随着超大规模预训练模型系统的开放，预训练基线智能水平大幅提升，行业人工智能应用不必从零开始开发，只需结合某个行业的领域数据进行调整，即可生成某个领域的相关模型，且得到良好的精度和性能。华为云发布的盘古预训练大模型已经在多个行业、100多个场景成功验证，包括能源、零售、金融、工业、医疗、环境、物流等等。其中，在能源领域，盘古预训练大模型帮助行业客户实现设备能耗的智能控制，可以节约电力成本50%；在金融行业中的异常财务检测，让模型精度提升20%以上；在尘肺检测中，病例识别准确率提升22%等等。行业应用和算法高效流通可以让人工智能应用和场景快速复制。

4.1.2 科学计算正在从传统HPC进入科学智能新阶段

科学计算是继大模型之后，AI发展的另一重要方向。此前，借助HPC(High Performance Computing，高性能计算)技术，科学计算对基础科学研究和国计民生行业发展起到重大推动作用。但是，随着求解问题不断复杂化、高维化，科学计算仍然面临着维数灾难、计算尺度受限、理论突破与工程方法创新缓慢三大挑战。

因此，越来越多的科学家正在将AI技术引入到科学计算，科学计算正在从传统HPC进入到科学智能的新阶段。科学智能同时覆盖HPC与AI两大技术领域，包含AI赋能机理计算、数据驱动AI计算、机理计算与AI计算相融合三大计算场景。

第一个场景是AI赋能机理计算，它是将AI计算嵌入到机理计算中，实现AI对机理计算的加速。

第二个场景是数据驱动的AI计算，它则不依赖于数学机理，通过大量的数据输入，获得AI模型，通过AI计算获得结果。

第三个场景则是机理计算与AI计算相结合，它提升了科学计算的准确率和计算效率。

目前，科学计算已经进入科学智能新阶段，其创新技术已经在气象、新材料研发、生物信息等领域中得到应用。

建议

4.2.1 汇聚大模型发展要素，使能大模型从规划到落地

当前人工智能技术趋势正朝着通用大模型方向发展，大模型具备更强泛化能力、可覆盖多业务场景，发展大模型也成为产学研各界共识。为了更好的推动大模型的发展，倡议汇聚大模型的发展要素，构建从规划、开发到产业化的大模型全流程使能体系，与产业界共筑中国大模型生态。

1、以大模型地图，统筹大模型有序发展

首先，建议统筹规划大模型发展布局，汇聚大模型发展要素，在算力方面加强发展人工智能计算中心和算力网络，塑造我国人工智能大模型人才培养体系，同时以自主创新的人工智能根技术发展我国大模型；其次，强化场景创新，提升大模型的活跃度和影响力；最后，强化政府支持，鼓励产学研各界携手在产业条件具备的行业和区域加速大模型的产业落地。

2、打造大模型开发使能平台，让大模型易开发、易适配、易部署

针对基础模型开发，建议打造大模型开发套件，通过算法开发、并行计算、存储优化等能力，实现大模型的高效开发；此外，建议开发大模型微调组件来适配行业应用，实现一键式微调和调优功能；在模型推理部署方面，还需要提供大模型部署套件，以实现分布式推理服务化、模型轻量化和动态加密部署功能。

3、成立大模型产业联盟，推动大模型应用落地

技术维度端到端打通后，大模型下一个最为关键

的问题是产业化落地。为了打通科研创新和产业应用的断点、促进大模型产业化落地，建议围绕大模型打通产学研用，建立大模型产业联盟，促进产业伙伴直接基于大模型孵化行业应用，实现产业聚集，让大模型真正赋能产业。

同时，产业联盟模式可以加速大模型从科研创新到行业落地的进程，在这样的大模型产业化落地过程中，各行业领域可以以更为丰富的数据和参数、更泛化的应用场景，来反哺大模型基础能力，让大模型更智能、场景适用性更好，从而迭代升级，为行业应用提供更大的支持，从而形成大模型创新-应用-迭代创新的产业正循环，开启了“炼大模型”的新范式。

4.2.2 打造科学智能基础平台、携手产学研构筑科学智能生态，加速产业闭环

过去单一、烟囱状的软硬件平台已无法满足科学智能需求。因此，华为建议打造原生科学智能基础软硬件平台，以实现极致性能、极简开发。华为认为，该基础平台在硬件方面应当拥有面向多样性算力的液冷整机柜，在软件方面包含业界领先的融合编程语言、编译器和操作系统，在开发使能方面则需要全场景统一的工具链，应用使能方面需要AI与HPC融合的框架和调度器。从底层硬件到上层应用协同创新，为科学研究提供“AI范式”。

对于科学智能的产业生态建设，华为倡议成立科学智能创新联合体，汇聚政策、科研和产业优质资源，携手产学研伙伴，以科学智能新范式，拓展科学边界，助力技术创新，加速科研创新到产业落地进程，加强交叉学科建设和人才培养，构筑中国科学智能领先格局。

解决方案

4.3.1 基于大模型全流程使能体系，使能大模型规划、开发、产业化

华为的人工智能大模型全流程使能体系，包含从大模型规划、大模型开发到大模型产业化的全流程，可端到端加速大模型产业落地，是以大模型产业化推动AI产业化的新范式。

规划大模型沙盘，与产业界共筑中国大模型创新高地

从2020年开始，国内外顶尖公司的AI技术发展，越来越像一场比拼资金与人才的军备竞赛，推动AI竞争从2018年前后兴起的“大炼（小）模型”，进入到今天的“炼大模型”时代。大模型的优势不言而喻，但动则上百亿的大参数，也带来了训练成本太昂贵，模型修正不容易等难题，导致本来定位于“不再重复造轮子”的大模型，面临重新陷入粗放式发展的境地。华为看到这一问题，积极联合产业界规划大模型沙盘，牵引产业界建设真正需要的大模型，共筑中国大模型创新高地。（图16）

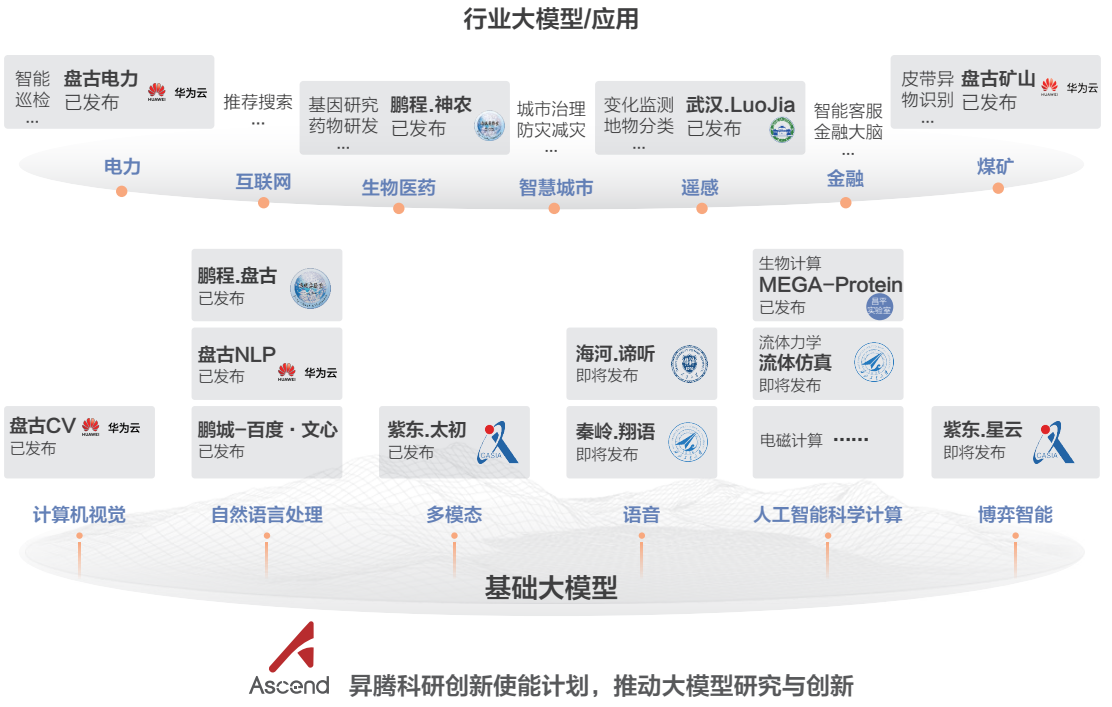


图16 昇腾大模型沙盘

从任务和应用类别两个维度出发，过去的一年，华为携手产业界伙伴基于昇腾AI先后推出了各个领域有影响力的大模型，形成了基础大模型+行业大模型的整体布局。基础大模型面向多行业领域通用需求，行业大模型面向特定行业多应用场景，类似“新基建”中的信息基础设施+融合基础设施，形成既有横向，也有纵深的立体支撑。

值得一提的是，考虑到“炼大模型”对大算力的强需求，华为与产业界在规划大模型沙盘的同时，全国20多个城市也都规划和建设了人工智能计算中心，并已开始将部分算力中心连点成片构建中国算力网——智算网络，以便基于它们的超强算力孵化AI大模型，大幅缩短大模型的训练时间。鹏程、武汉、秦岭、金陵系列大模型的快速推出，正得益于这一布局的强力支持。反过来，这些带有一定地域特色的大模型，又能够结合本地AI算力更好地服务产业。

打造大模型开发使能平台，让大模型易开发、易适配、易部署

依托长期的根技术积累，华为建立起了完整的大模型开发使能平台，加速从基础模型开发到推理部署的全流程，让大模型易开发、易适配、易部署。

首先，在基础模型开发方面，华为推出大模型开发套件，通过算法开发、并行计算、存储优化、断点续训重磅特性支撑大模型的高效开发。这其中，作为人工智能之“魂”，昇思MindSpore自诞生起就有着鲜明的产业导向，可以在云、边、端等不同环境下进行开发部署，是并行维度业界最多、模型切分支持结构最全、单机容纳模型参数业界最强的的AI框架，这使其原生支持AI大模型训练，具备实现开发并行代码量降低80%、系统调整时间下降60%、仅用512卡就能完成十万亿模型参数训

练的超强能力。

其次，在行业应用适配方面，华为推出基于MindX的大模型微调组件，其预置典型行业任务微调模板，通过小样本学习等手段，实现一键式微调和低参数调优，可以快速适配各种行业应用。目前紫东太初大模型就基于微调套件，提供了开放服务平台，已有40多个企业在平台上孵化了近60个产品解决方案，可以快捷的完成场景适配。

最后，在推理部署方面，推出基于MindStudio的大模型部署套件，其提供量化、剪枝、蒸馏等模型小型化能力，实现10倍级模型压缩率，同时分布式推理服务化能力还大幅提高吞吐率，此外模型动态加密技术，可在保证模型性能的同时对部署的模型进行加密，保护开发者的模型资产。

从科研创新到行业落地，开创人工智能产业聚集新模式

技术维度端到端打通后，大模型下一个最为关键的问题是产业化落地。去年底，基于全球首个智能遥感框架及数据集武汉.LuoJia和全球首个三模态大模型紫东.太初，产业各界成立了智能遥感开源生态联盟和多模态人工智能产业联盟，如今60余家伙伴已陆续孵化出多个行业解决方案。

千博信息与中科院自动化所、华为三方联手，基于昇腾AI基础软硬件平台以及紫东.太初三模态大模型，打造出手语多模态模型并发布手语教考一体机，大幅改善了特殊人群的学习环境。此外，长安汽车、新华社技术局、浙江移动、爱奇艺等多模态人工智能产业联盟成员也分别打造了自己的多模态+智能座舱、多模态+新媒体内容检索平台、多模态+南宋御街数字人、多模态+视频摘要智能平台等场景化大模型及行业应用。智能遥感开源生态联盟下，基于武汉.LuoJia的自然资源大脑、全场景类脑遥感矩阵、耕地保护自然监测平台、智能遥感解译平台等创新成果也不断涌现。

大模型是AI产业加快发展的必然，也是科研创新走向产业应用的关键。华为联合产业界基于昇腾AI开启的“炼大模型”新范式，首次从大模型规划、开发到产业化构建了大模型全流程使能体系，拉通了技术生态与商业生态之间的桥梁，将加速我国大模型产业化发展，进而推动AI产业化和产业AI化，加速智能世界到来。

4.3.2 打造原生支持科学智能的基础软硬件平台，原生构建科学智能新生态

华为基于鲲鹏和昇腾AI，融合HPC和AI两大技术优势，通过创新的计算架构，打造原生科学智能基础软硬件平台，以全栈的创新实现科学智能基础设施的极致性能、极简开发。（图17）

在硬件方面，华为推出科学智能全场景液冷“天成”多样性算力平台，其支持多样性算力灵活弹性部署，可实现液冷级能效，整系统TCO降低20%，性能提升20~30%；在基础软件方面，华为发布毕昇C++编程语言并全面升级毕昇编译器，实现系统开发效率提升一倍，系统性能提升30~50%；在开发使能方面，华为升级全场景统一工具MindStudio，实现软件融合编程、编译和调优，可使科学智能全场景开发效率提升50%；在应用使能方面，昇思MindSpore 2.0升级为AI融合框架，原生支持科学智能以及多瑙融合调度器，其内嵌科学智能套件，让科学智能应用的开发、部署和调度更高效，应用性能提升10~20倍，系统资源利用率提升15%。

目前，科学智能基础软硬件平台已在新材料研发、大飞机设计、蛋白质结构预测等领域中应用。

科学智能要实现产业化落地，还需要突破科研理论，创新工程方法，并构建产业生态，聚焦产业价值场景，打通科研创新、应用示范到产业推广的通道。

在华为全联接大会2022中，华为倡议成立科学智能创新联合体，呼吁产学研各方共同携手，为大力发展科学智能生态奠定基础。

图17 科学智能基础软硬件平台

趋势五

绿色高效成为算力基础设施建设的关键诉求

产业趋势

5.1.1 在双碳目标下，算力基础设施的建设更加注重能耗

未来算力将爆炸式增长，而数据中心是算力的主要载体，是新型基础设施节能降耗的关键环节，也是促进全社会降碳增效的有力抓手。传统数据中心能耗高、算力利用率低，在“3060双碳”目标牵引下，国家对数据中心能耗提出更严格的要求，各省也出台了能耗指标及PUE要求，算力爆发式增长和降低碳排放的矛盾愈发突出，数据中心绿色化转型升级势在必行，算力基础设施的建设更加注重绿色高效。

5.1.2 从单领域创新走向系统级创新，实现绿色高效

传统数据中心能耗控制往往是单领域创新优化，比如材料优化、供配电优化、空调制冷优化等，但提升效果有限，因此需要通过系统工程的创新，包括提升集成度、多领域全栈协同优化，比如通过AI技术对设备功率进行动态控制、IT设备与供配电及制冷设备全栈协同联动等，解决大规模数据中心建设能耗的难题，降低能耗，提高能效比和系统性能，实现绿色高效。

建议

5.2.1 建设模式从传统的部件堆叠逐步走向集群全栈一体化

传统的数据中心都是分层建设、部件堆砌，导致建设周期长、能耗高、算力利用率低；集群计算中心为代表的新建数据中心，采用全栈一体化设计，从L0到L3整系统创新和协同优化，集中化建设、集约化使用，达到多样算力融合、模块化快速部署、液冷绿色高效，实现DC as a Computer。（图18）

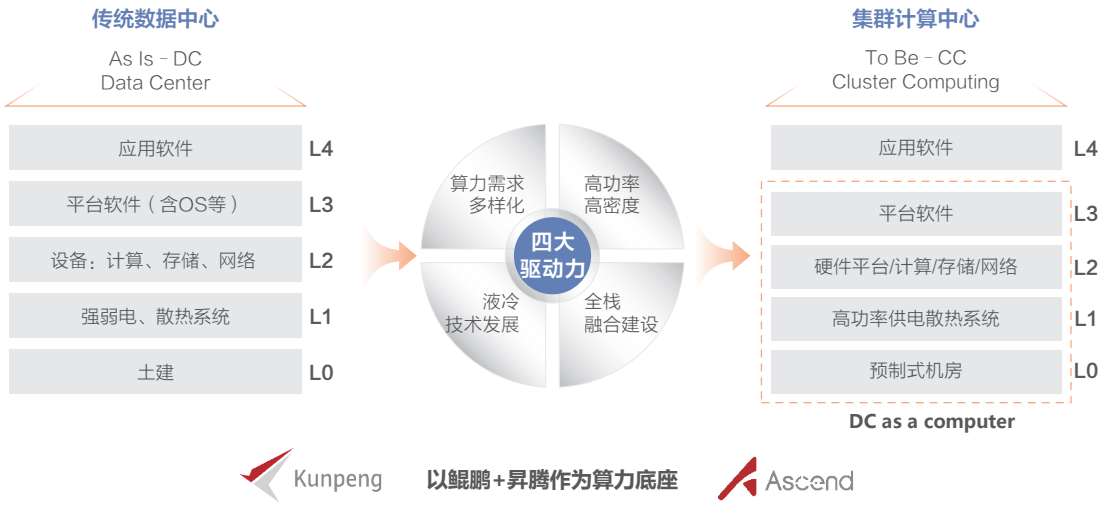


图18 数据中心建设从部件堆叠走向全栈一体化

5.2.2 散热方式逐步从传统风冷走向风液混合或全液冷

数字经济时代，对高性能、高密度的计算需求逐渐增多。芯片和单机柜功率密度不断增大，传统散热方式难以为继，房间级空调方案，受限于物理空间和空气比热容低，难以支持每柜12KW以上机柜；行级空调方案，单机柜超过12KW时，需冗余配置空调以增加换热量，影响机房出柜率和TCO；超过15KW，风冷换热效率不足，难以支撑高功率元器件散热负荷，无法满足散热需求，液冷技术逐渐普及。液冷提供了高能效、高可靠、低碳环保的散热技术，逐渐成为算力基础设施的主流散热方式。

5.2.3 算力评估逐步从面向硬件的裸算力，走向面向业务的有效算力

传统算力度量采用裸算力或部件级算力评估，如规格算力（芯片标称的算力规格）指标，单机或单服务器的性能评测，只关注IT计算设备的单台设备理论性能，无法完全体现集群系统或者算力中心整体性能。算力中心的真实性能需要综合考虑芯片、存储、网络以及平台软件各层协调所呈现的综合业务性能，也就是“有效算力”。有效算力通过评测真实业务性能表现，来衡量算力基础设施对业务的支撑效果，也就是业务实际可获得的算力水平。

通过有效算力的模式来牵引算力基础设施的建设，提升算力的利用率，推动算力建设绿色高效诉求的落地，更好地支撑当地产业的发展。

解决方案

5.3.1 集群计算全栈协同优化，实现DC as a computer

集群计算解决方案，通过系统级工程创新，采用软硬件协同设计，包括应用软件与平台软件的协同优化，基础硬件平台及供电散热系统与平台软件的协同优化，实现从应用到平台到基础硬件平台、供电散热系统的纵向业务联动，数据中心全栈优化DC基础架构；采用数据中心整体设计，包括计算、存储、互联等各子系统协同优化，结合基础架构及通信网络优化使能平台及中间件持续提升，CPU/NPU/xPU多样性算力平台及融合调度，实现横向资源整合，突破硬件基础算力瓶颈。（图19）

通过上述措施，软硬协同、纵向业务联动；整体优化、横向资源整合，提升数据中心的有效算力，提高能效比，实现DC as a Computer。

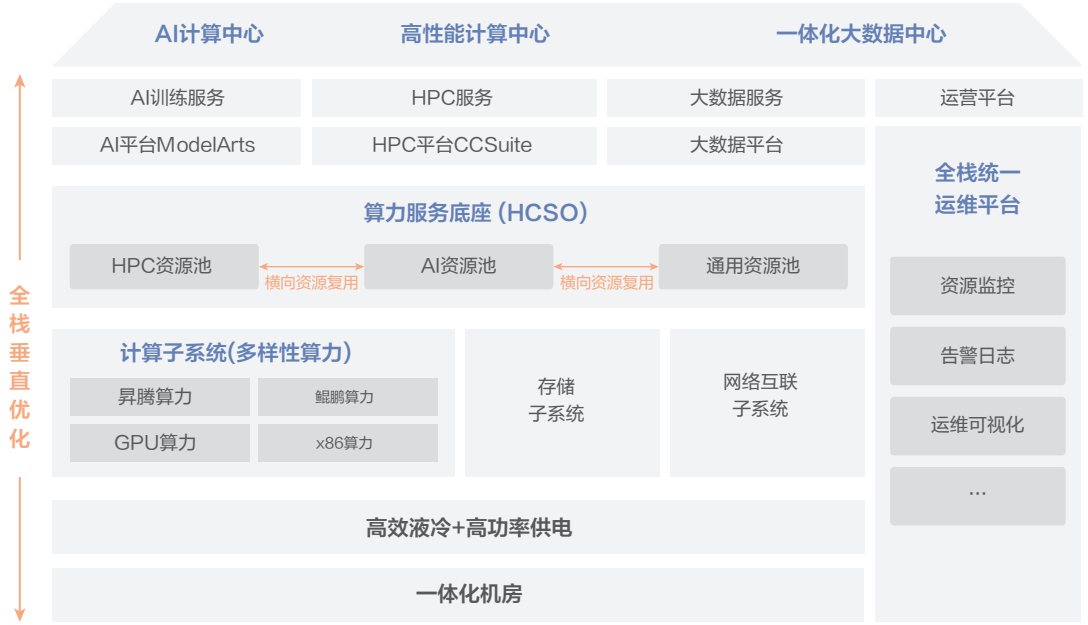


图19 集群计算解决方案整体架构

趋势六

算力网络将成为重要的算力供给方式

产业趋势

6.1.1 算力建设从分散化走向集约化

在“东数西算”“网络强国”等战略的牵引下，在“3060双碳”目标牵引下，原来传统的分散化算力建设的弊端也越来越突出，建设周期长、能耗高、利用率低，不符合绿色高效的算力发展趋势。以人工智能计算中心、超算中心、一体化大数据中心等为代表的算力基础设施，成为国家新基建的重要组成，算力建设走向集约化，建设周期短、能耗低、算力利用率高。各地集中进行算力中心的建设，让算力像水和电一样，成为城市新型基础设施和公共资源。就像过去每个核心城市标配有机场、有高铁站，未来数字经济发展、智能化发展，核心城市都将标配一个公共算力中心，来以算力赋能科研创新和产业发展。

6.1.2 从算力中心，走向算力网络

各地算力中心/算力基础设施陆续建成后，结合网络基础设施，就可以连成一张算力网络。像过去有电力网、通信网一样，在数字世界也一定会有一张算力网。以人工智能算力为例，2021年，中国科学技术信息研究所、新一代人工智能产业技术创新战略联盟（AITISA）、鹏城实验室共同发布《人工智能计算中心发展白皮书2.0》，指出了人工智能中心发展的新阶段——从人工智能计算中心走向人工智能算力网络。2021年底，在科技部的指导下，鹏城实验室牵头成立了人工智能算力网络推进联盟，推进各地上线的人工智能计算中心连接成网上线运行。2022年6月，“中国算力网—智算网络”一期正式上线，这是中国算力网络建设迈出的关键一步。各地的算力建设，开始从单独的算力中心，走向全国范围内的算力网络。

行动建议

6.2.1 加速算力基础设施的建设

集约化建设绿色高效的算力基础设施，既是响应国家产业政策的需要，也是区域社会经济发展的需要。算力基础设施建设，需要结合当地的产业布局、科研实力及数字经济发展情况，以应用为导向，以信息技术与制造等传统技术深度融合为主线，推动人工智能计算、超级计算等先进技术的产业化与集成应用，发展高端智能产品，夯实核心基础，提升智能制造水平。促进算力服务相关各基础设施的建设，完善公共支撑体系，促进产业发展，推动制造强国和网络强国建设，助力实体经济转型升级。

结合各地实际情况，联合高校、科研院所、企业等行业技术力量，适度超前、加速建设算力基础设施，可以服务于千行百业，满足高校、科研院所、企业不断增长的算力需求，以充沛算力，促进本地各行各业发展的诉求；同时，承担国家和区域里涉及国际民生的关键行业科研诉求，带来良好的经济效益和社会效益。

6.2.2 积极加入中国算力网，实现算力汇聚共享

2022年6月，在科技部指导下，由鹏城实验室牵头的“中国算力网-智算网络”正式上线，伴随各地算力基础设施的不断建设。截止2022年11月，鹏城云脑、北京、成都、中原、合肥、武汉、西安、济南、青岛、沈阳、广州、重庆、昆明、福州、长沙、河北（廊坊）等20多个节点已接入中国算力网。多个人工智能计算中心间的AI算力调度与协同训练已完成初步验证，全国AI算力一张网初具雏形。

未来，各地的人工智能计算中心、超算中心、一体化大数据中心、算力枢纽、以及社会泛在云算力中心都可以接入中国算力网，共同构建一个支撑中国数字经济发展的强大算力底座，汇聚多种社会算力，实现绿色高效布局、泛在算力协同和全网交易流通，以东数西存、东数西算、东数西训为牵引，将逐步形成绿色集约的算力布局；汇聚多种社会算力，形成更加泛在的算力协同，并通过全网的算力交易流通，弹性满足全网范围内的算力需求。让算力成为与水电一样，可“一点接入、即取即用”的社会级服务。

解决方案

6.3.1 算力网络架构创新，打造全网一台计算机

算力网络需要以终为始，站在最终用户使用者的角度，打造全网一台计算机的架构，实现全程全网的社会级算力服务。算力网络的参考架构包括算网大脑及运营层、算网基础设施及使能层。（图20）

单域自治：使能层通过算力使能、网络使能和数据使能实现算力、网络和数据的单域管理与调度，确保单域独立交付与演进；

跨域编排：实现跨域跨厂家的业务编排与调度，负责多云管理；

北向接口：制定统一接口标准，各单域使能以服务化形式（云服务或Restful API）对外，供上层调用；

以云调算：云纳管算，通过云服务来调度各种算力，重用云在大规模、跨域和异构算力的统一调度能力。非云化资源池由云管纳管，不参与全局调度；

通过单域自治、跨域编排、北向接口、以云调算，实现“全网一台计算机”，为用户提供无所不在的算力服务。

图20 算力网络架构创新，打造全网一台计算机

华为技术有限公司
深圳市龙岗区坂田华为基地
电话: (0755) 28780808
邮编: 518129
www.huawei.com