目录 CONTENTS

P05 **第一章　超智融合发展背景**

P06 一、人工智能催生的巨大算力需求推动超算向适 AI 化升级
P09 二、算力赋能人工智能的关键在于算力设施向超算与智算融合演进

P19 **第二章　超智融合重塑计算格局**

P19 一、超智融合的内涵范畴
P22 二、超智融合的三个阶段
P26 三、超智融合的关键能力

P29 **第三章　超智融合技术路径创新**

P30 一、算力架构
P34 二、算力调度
P37 三、 算力服务和运营
P39 四、 数据安全和隐私保护

P41 **第四章　超智融合新型应用场景和实践案例**

P42 一、创新应用场景
P43 二、典型案例

P45 **第五章　超智融合进展和目标愿景**

CHAPTER1
超智融合发展背景

一、人工智能催生的巨大算力需求推动超算向适 AI 化升级

随着技术迭代发展，计算解决问题的范式也在不断变化。从早期的数学模型驱动，到数据驱动，再到 AI 赋能。超算与 AI 的融合正在重塑计算科学、IT 产业和人类社会发展格局。中国在超算领域拥有深厚的技术积累，而应用侧对算力结构转型的迫切需求进一步推动了超智融合技术的快速发展。

超智融合是一个循序渐进的发展过程，它融合了超算强大的数据处理能力与人工智能的算法优化能力，可有效解决算力瓶颈，推动计算技术创新发展。超智融合不仅涉及数据、算法、业务等层面的融合创新，也对底层的计算、存储、网络等基础设施提出了相应要求。

（一）超智融合的历史演进。 尽管人工神经网络早期的研究工作可以追溯至上世纪 40 年代，现代意义的超级计算机也在上世纪 70 年代问世，但在很长一段时间里，超级计算与人工智能似乎就象两条车道上并行奔跑的汽车，前者追求“算得快”，后者追求“算得巧”，相对独立发展。

进入本世纪 10 年代，随着英伟达公司开始推动 GPU 从传统图形渲染应用进入大规模并行计算领域，加上互联网大数据的爆发式增长，深度学习算法、CPU+GPU 异构并

行超级计算机、大规模数据集这三大要素开始汇集，进而开启了新一代人工智能日新月异的发展进程。超级计算机的强大算力与人工智能的深度神经网络算法，在大数据的催化下，开始走向融合。

在超智融合的早期探索阶段，一个具有里程碑意义的事件是在 2012 年多伦多大学研究团队使用 NVIDIA GTX 580 GPU 训练 AlexNet，并在 ILSVRC 挑战赛上大获成功，开创了一种新的计算模式。在这一时期，尽管已经出现了大规模的 CPU+GPU 异构集群，如 2010 年出现的世界首个达到 P 级算力规模的异构超算曙光星云（采用 2560 块 Nvidia C2050），但当时的 GPU 仍以双精度为主，低精度性能不足。

随着深度学习算法在图像识别、语音识别、自然语言处理等领域的大量应用，为提高各类模型的训练速度，并降低成本，以提供半精度、整型算力为主的专用 AI 芯片开始出现，如 2016 年谷歌发布的首个 AI 专用芯片 TPU，2018 年寒武纪发布的中国首款 AI 专用芯片 MLU 等。加之 2017 年谷歌提出 Transformer 架构，AI 从小模型进入大模型时代，对低精度算力需求出现指数级增长。专门面向 AI 训练和推理的智能计算系统开始出现，并与传统的高性能计算机或超级计算机独立发展。

然而，近几年来，随着人工智能的进一步发展，如多模态大模型、科学大模型、行业智能体的出现，尤其是 AI 在行业专业领域里的泛化和深入应用，如 AI for Science（科学智能），使得智能化任务更加复杂，应用场景日趋多样。无论是以双精度算力为主的传统超算，还是以半精度算力见长的智算系统都难以单独满足多元复杂的场景任务，对新型超智融合系统的需求日益迫切。2021 年 AlphaFold2 对 35 万种蛋白质结构的成功预测，开启了新的科研范式，同时也对计算系统提出新的要求。新的超智融合计算系统，需同时满足数值模拟和神经网络计算任务，支持高精度、低精度和混合精度计算模式，并从传统的以 CPU 为中心转向以 GPU 为中心进行系统重构，对计算、存储、网络等子系统进行协同优化设计，同时系统的管理、运维和调度也要更加智能化。

（二）人工智能将助力超级计算机突破性能发展瓶颈。当前摩尔定律逼近物理极限，单一计算架构难堪重负，导致全球超算发展遇到瓶颈。全球超级计算机 500 强（Top500）数据显示，超算增长从过去每 10 至 11 年增长 1000 倍降至增长 100 倍以下。特别

传统超级计算机提供的是超强双精度浮点运算能力，主要用于解决数值模拟和第一性原理计算等科学与工程计算问题，开展预测性科学研究。与之不同的是， 智算系统提供的是半精度浮点数或整数运算，主要面向人工神经网络模型的训练和推理。

是由于系统规模受到能效指标约束、量子计算机等颠覆性技术距离实用尚有距离、新原理的计算和存储器件缺少突破、自主高端处理器研制受制于人、超算应用软件对外依赖度较高等原因，我国追求超算机器性能世界领先存在一定难度，将高性能计算的发展目标从机器性能世界领先转向应用成效世界领先成为赢得主动的关键。以应用成效为目标要求应用软件充分发挥并行硬件的优势，着力突破软硬结合、应用优化。人工智能催生的新软件能极大丰富传统超算的软件资源，提高其解决复杂挑战性问题的能力。抓住 AI 发展的契机，能够带动超算领域硬件、算法、软件、应用和系统的协同创新，提升超算的应用成效。

（三）人工智能驱动超级计算机增强 AI 特征运算能力。 随着人工智能技术快速发展，千亿参数人工智能大模型的训练催生了巨大的算力需求，例如 OpenAI 训练一次 1750 亿参数的 GPT-3 模型所需算力约为 3640PFlops-day（即每秒运算 1000 万亿次，运行 3640 天），而 GPT-4 的训练算力需求更是高达 GPT-3 的 68 倍。AI 改变计算解决问题的范式，“规模效应”猛增的大模型正成为名副其实的“算力黑洞”。为此 OpenAI 与微软公司计划构建十万乃至百万级 GPU 的算力集群，以满足 GPT-6 的训练需求。

相比之下，当前我国大模型训练面临着巨大的算力缺口。除了被限制的英伟达 GPU 产品外，我国目前有两类算力集群可以支持大模型训练。一类是基于国产 AI 芯片的集群系统，但由于国产 AI 芯片的生态系统尚不完

善，迈向应用的路途道阻且长；另一类是国家超级计算设施，其超大规模的异构计算集群是巨量参数规模的大模型绝佳的训练场。

通常来说，传统超级计算机提供的是超强双精度浮点运算能力，主要用于解决数值模拟和第一性原理计算等科学与工程计算问题，开展预测性科学研究。与之不同的是，智算系统提供的是半精度浮点数或整数运算，主要面向人工神经网络模型的训练和推理。随着人工智能算法不断渗入各行业应用领域，应用场景变得更加复杂，纯粹的半精、整型算力环境已难以满足应用落地的需求。例如在蛋白质结构预测、新材料设计、天气预报、大规模分子模拟等 AI for Science 场景中，只有超智融合才能实现最优解，甚至是只有超智融合才能使问题变得可解。

二、算力赋能人工智能的关键在于算力设施向超算与智算融合演进

超智融合涉及数据、算法、业务及算力设施等层面，其中基础设施层面的融合非常关键，是支撑上层应用融合得以顺利进行和持续发展的重要条件。推进超智融合不单是缓解大模型“算力荒”的有效之策，更是顺应智能时代发展的应有之义。

超算、智算融合发展具有三方面特点：

一是强化多元算力资源协同调度。 由于当前应用侧对算力结构转型存在迫切需求，基础算力、智算算力、超算算力等应用的多元化发展催生“超智融合”，即采用混合型算力资源或融合型算力体系，对异构算力资源进行池化管理与统一调度，同时满足多种不同算力的应用需求。多元算力融合可构建更加适应人工智能时代需求的新型算力生态系统，未来发展应从供需两侧做好算力资源和业务应用的统筹衔接，避免有效应用需求不足、缺乏网络服务质量保证、没有成熟调度体系的普遍性算力互联，更不能脱离实际应用需求进行异地计算和远地算力设施布局。

二是强化软硬件协同。 “超智融合”技术路径上需要底层技术与体系结构进行软硬件协同创新，例如解决覆盖全精度算力供给，友好通用、零移植开销的软件栈，存算传紧耦合分布式异构体系结构，全局高速共享存储、浪涌 IO 优化设计，卡间、节点间高速互连协同，多元融合的算力调度系统，大规模智能系统管理，支持大规模并行计算的 AI 框架、优化通信效率的中间件等。同时，由于不同的人工智能应用场景需要不

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/675020034043012023