



算力经济时代·数字中国万里行 2023新型算力中心调研报告

张广彬 王海峰 张翼 | 著

益企研究院
E7 · RESEARCH

出品

智能计算
中国智能计算产业联盟

指导

更多免费行业报告

并购家

www.ipoipo.cn



特 别 鸣 谢

安谋科技
arm CHINA

目录 CONTENTS

序：算力经济发展趋势分析与展望	P001
1、从洞察算力到提出“算力经济”	P001
2、狭义算力经济与广义算力经济	P003
3、AGI 时代来临，模型服务（MaaS）商业模式呈现	P004
4、科学计算：传统科学与 AI 深度融合	P005
5、算网融合带来算力市场变局	P006
6、用算力服务标准确保算力服务健康发展	P006
 第一章 算力经济时代的基础设施新价值	P007
多类算力基础设施并行发展	P011
多元算力与高速互联	P016
高效绿色的数据存储与管理	P022
高安全数字基础设施是趋势	P024
绿色低碳持续推广	P027
能源与算力协同	P029
 第二章 多元算力：CPU + GPU	P031
GPU：大芯片与小芯片	P033
CPU：性能核与能效核	P034
摩尔谢幕，Chiplet 当道	P037
• 摩尔定律放缓	P037
• Chiplet 简史	P037
• 四等分：形似神不似	P040
Chiplet 与芯片布局	P041
• 网格架构：Arm 与 Intel	P043
Arm 新升：NVIDIA Grace 与 AmpereOne	P045
网格架构的两类 Chiplet	P049
• EMIB 及其带宽估算	P053

第三章 算存互连：Chiplet 与 CXL P055

SRAM 的面积律 P056

向上堆叠，翻越内存墙 P057

- 回首 eDRAM 时光 P061

HBM 崛起：从 GPU 到 CPU P063

- 中介层：CoWoS 与 EMIB P065

向下发展：基础层加持 P067

标准化：Chiplet 和 UCIe、CXL P072

- CXL：内存的解耦与扩展 P073

- UCIe 与异构算力 P080

- Chiplet 的中国力量 P084

- Chiplet 走出“初级阶段” P085

第四章 算力互连：由内及外，由小渐大 P087

为 GPU 而生的 CPU P088

NVLink 之 GPU 互连 P094

NVLink 组网超级集群 P096

InfiniBand 扩大规模 P102

第五章 绿色低碳和可持续发展 P105

液冷应用 高性能计算中心跨越功耗墙 P110

液冷实践 全栈数据center理念落地 P111

- 1、业务前置 模块化交付 P114

- 2、以全栈的视角 垂直整合 P115

- 3、产业生态融合演化 P117

智算中心 跑出液冷加速度 P119

节能减排新实践 重构排碳之源 P121

序：算力经济发展趋势分析与展望

全国政协委员
中国科学院计算技术研究所研究员
益企研究院首席专家顾问

张云泉

在今年两会期间，我递交了一份关于算力发展的提案：关于合理规划算力网建设，确保东数西算健康发展的提案（W01072），核心内容是算力网和东数西算。在《算力经济时代·数字中国万里行 2023 新型算力中心调研报告》出版之际，希望通过这篇文章来解释这份提案产生的背景，同时也对当前算力经济的发展做一些展望。

当人类社会从热力时代过渡到算力时代，计算也随之成为未来智能设备的关键驱动力，这点在数字经济时代尤为突出，算力经济名词也在 2018 年被提出。

1、从洞察算力到提出“算力经济”

算力经济最初定义的维度是比较简单的。

在从事超级计算 30 余年的过程中，我对计算技术的发展和應用有深刻的理解，早期的超算并不倾向于使用 GPU。2008 年英伟达提出 Fermi 架构，将显卡扩展为通用计算 GPU，希望用在超级计算机上，但在当时，GPU 在科学计算的应用都不是很成功，如超级计算机中的曙光星云、天河 1、天河 2 等在使用中的效果没有达到预想效果。

到 2010 年，我们团队整理中国高性能计算机 TOP100 排行榜的计算机结构后发现，CPU+GPU 正成为超级计算机的技术发展趋势。

这一趋势在 2015 年之后更为明显，AlphaGo 围棋大战之后，人工智能取得成功，发现 GPU 其实更适合深度学习，英伟达将 GPU 的应用重点从超级计算机转到人工智能上。

一直以来，超级计算主要是做科学计算和基础研究，需要具备长期投资的理念，很难直接和国民经济发生关系，地方政府在算经济账时，会考虑投资的回报率是多少、投资周期是多长？多少年能收回投资成本？能拉动多大的经济增长？因此说服政府投资超级计算平台很难。

2018 年，有了“算力”这个名词后，这一问题出现了转折点。起初算力这个词来源于区块链、挖矿领域，相对比较狭窄、有点偏负面。

但随着超级计算和人工智能、云计算的结合，甚至包括区块链和大数据的融合，“算力”似乎和国民经济的联系更密切了。过去面临的关于超算的经济回报问题，在人工智能时代（我们称之为“智能计算时代”），应该说清楚了。基于这个想法，在区块链的启发下，国内的专家们开始把超级计算的“计算”，泛化成“算力”。

2018 年，我参加地方政府的相关活动时提出“算力经济”这一理念，当时认为，随着超级计算技术的发展，大数据、云计算、人工智能、区块链彼此之间的融合创新，算力经济会成为经济发展的重要抓手，会成为地方政府新旧动能转换的重要手段。但在那时，“算力经济”其实还不太被社会接受。那时最热的是大数据、人工智能、区块链，但算力不热，没什么人谈“算力经济”。

这一观点在随后就得到印证，2018 年益企研究院（E 企研究院）开启数字中国万里行，实地考察了全国 8 个超大规模云数据中心，并出版了首个《中国超大规模云数据中心考察报告》聚焦数据中心架构创新和技术迭代，探索智能基础设施的上层应用，呈现新技术和新型算力基础设施的价值。

2019 年发布的中国高性能计算机 TOP100 排行榜中我们发现，这一年超算应用的领域也发生了极大的变化。过去超算主要集中于科学计算、政府行业、能源行业、电力行业以及气象领域。但随着许多互联网公司开始申报超级计算机，在 TOP100 中，有 30% 的系统都来自互联网行业，比如云计算、机器学习、人工智能、大数据分析以及短视频领域。这些领域对于计算需求的急剧上升，超级计算继续与互联网技术进行融合。

同时，算力基础设施中除了云数据中心和超算中心，还出现智能计算中心为代表的算力基础设施。其中较为典型的案例就是国家超算济南中心科技园与腾讯在上海松江打造的人工智能计算中心。

2020 年，益企研究院发起数字中国万里行第三年之际，我在接受益企研究院访谈中，正式提出：我们即将进入一个依靠算力的人工智能时代，这

也是未来发展的必然趋势之一,同时,随着用户对算力需求的不断增长,算力经济时代将登上历史舞台。

2020 年后,我在相关调研实践中不断总结,最后形成了“超算与人工智能融合创新的算力经济时代”的思考。

2、狭义算力经济与广义算力经济

中国高性能计算机 TOP100 排行榜已经发布了 20 多年,行业一直通过排行榜观察中国超级计算产业的发展趋势。到 2021 年,我们又发现一个新的现象:在 TOP100 的前 10 名有 7 台机器,它们不是专门服务某些行业,而且这些机器没有具体的应用目标,是公司买过来之后专门用于卖算力的,而且这些机器性能很强。面对这个新出现的状况, TOP100 的专家委员定义了一个新领域叫算力服务业。

当时间进入 2022 年,算力服务的性能指标相比上一年已经翻倍,增长速度很快。算力服务业在 2021~2022 年的异军突起,也意味着中国正式进入算力经济时代,其背后的原因是超级计算技术的发展,大数据、人工智能、区块链彼此之间的融合创新,而这些因素背后的核心要素就是算力。算力应用已经开始渗透到千行万业之中,这也是在 2018 年提出算力经济概念之后,我们观察到这个行业的极速变化。

算力经济最初定义的维度是比较简单的。首先计算要成为算力经济的核心,未来,以计算能力来衡量一个地方或地区的数字经济发展水平,使之成为一个很重要的指标。一个地区的算力产业是不是发达,也意味着数字经济是不是有机会,尤其在东数西算成为国家发展战略之后,算力经济也成为西部地区新一轮经济发展的强力抓手。就目前来说,针对算力还没有一个统一的定义,我们可以将其理解为硬件和软件的配合,共同执行某种计算需求的能力,这个定义现在看来不是很全面。

我认为狭义的算力经济定义是指与算力强关联的算力服务产业链,其中包括了 4 类参与者:一是算力生产者,二是算力调度者,三是算力服务商,四是算力消费者;他们共同闭环成为一种商业模式。

随着认识的深化,随后又有一个广义的“算力经济”,我们称之为算力+。这不是我一个人提出来的。凡是可以用到算力的国民经济的各个方向单元,都是算力经济的范围。只要以算力为核心生产要素,以算力为引擎,就都是广义的算力经济。这是数字经济很重要的一个组成部分,在数字经

济中的比重会越来越大。统计数据显示，在世界各国的算力排名中，中国排在世界第二，人均算力处于中等国家的水平，目前中国还是有很大的算力鸿沟。在我国，算力的需求毋庸置疑，人工智能、5G、区块链、元宇宙的发展都对算力提出了强烈的需求，其增长前景是没有问题的。

现在有各种各样新的概念，很多课题组也开展了很多研究。在针对算力研究的著作中，《算力：数字经济的新引擎》这本书正式把算力进行系统的研究，提出来无数据不经济，这个定义非常好，比算力经济最初的概念更近了一步，提出了引擎性自主创新驱动的先进计算产业以及算力赋能和服务衍生的新模式、新业态形成了算力经济，作者是经济学家，从经济学角度阐述了算力对于经济的巨大影响力。书中指出，算力经济是数字经济衍生的新经济形态，数据作为主要的生产要素通过算力、算法的技术创新，促进数据经济和实体经济的深度融合，实现效率、效能、质量提升和经济结构优化升级。

综上所述，围绕算力本身产生的算力服务产业中，我们看到里面有芯片、操作系统，我认为可以从狭义和广义两个角度来看算力经济，狭义的算力经济指算力服务业产业链；有更广义的算力经济叫数字产业化、产业数字化、城镇数字化这种提供各种基础设施、提供各种支撑保障的新模式、新业态，也就是是算力 + 产业。

3、AGI 时代来临，模型服务（MaaS）商业模式呈现

随着算力经济的发展，超级计算机技术和人工智能融合创新会产生一类新的基建，专门用于人工智能计算的中心，也成为当下非常热的资产中心。就在 ChatGPT 面世之前，我们还不知道大模型可以实现令科技界为之兴奋的应用水平，只是知道它可以写一点新闻、聊天、画画，这些简单的功能会在更多应用场景中带来价值。

从 GPT3 到 ChatGPT 的过程，是大模型技术发展的关键节点，也是中国人工智能之路和美国人工智能之路的分歧点。这两年大模型国内也有相当数量的公司参与其中，但我们追求的是参数量，从千亿级到万亿级很快的跃进，但是智能属性没有涌现。OpenAI 走了另外一条路，利用人工反馈的训练机制，通过标注、对齐高质量数据，最后把这条路走通了，用千亿参数的大模型把通用智能挖掘出来了，这个事情是值得国内科技界去反思的。

另外一条路是人工智能内容生成 AIGC，包括大家在微信朋友圈里看到各种画，也成为现在的热门赛道。在 AIGC 赛道国内已经有布局了，从上游、中游到下游都有一些中国公司在做。

这些都意味着人工智能进入通用人工智能（AGI: Artificial General Intelligence）时代，具备五个特性：涌现性（参数超过临界值，模型能力实现突变）、工程化、通用性、密集型、颠覆性。这里就不多展开阐述。

4、科学计算：传统科学与 AI 深度融合

当计算改变科学，人工智能生物算法反过来被融合到科技计算建模中，相当于把数据科学和计算科学（AI for Science）整合在一起，这时产生一个新的“智能科学”赛道。以前科学计算的四个范式分别是实验科学、理论科学、计算科学和数据科学，智能科学范式（AI 范式）被称之为第五范式。其代表是斩获 2020 年戈登贝尔奖的 Deep Potential 方法展示了 AI 和分子动力学模型的有效结合，在保证精度的同时，指数级地提升了物理模型的效率。

基于科学计算的深度学习怎么反哺科学计算、解决计算问题，AI 范式确实创造了新的科学计算的方向，尤其是制药这个行业特别有效，极大提高了科学计算的精度，降低了成本。比如近年来，AlphaFold 等人工智能（AI）工具的出现，在生命科学领域促成了多项突破性进展。蛋白质的功能预测与设计成为最先受益的领域之一，在《科学》（Science）杂志上，Baker 教授团队带来了蛋白质设计的又一项革命性突破：利用强化学习，“自上而下”（top-down）设计蛋白质复合物结构。在几年前，预测蛋白质三维结构都遥不可及，更不用说从头进行设计了。这套颠覆了传统方案的全新突破不仅可能为我们带来更有效的疫苗及药物，还有望引领蛋白质设计的全新时代。

AI for Science 的数据来自各个学科的数据积累；模型来自各领域科学家发现的科学原理和规律；算法源自机器学习算法和数值方法等方面的创新。需要多样算力融合的综合型智能计算平台，通过分布式异构并行体系结构，实现多样算力的融合、优势互补，为 AI 训练、AI 推理、数值模拟等不同应用提供不同算力，实现高精度到低精度算力的全覆盖、多种计算类型的全覆盖，以及 AI 训练 + 推理全覆盖。

5、算网融合带来算力市场变局

在算力布局方面，国内目前有很多算力中心，有超算中心、智算中心，还有超大规模云数据中心，我认为未来算力中心慢慢会融合到统一的形态上，只是功能不同。

随着算力中心的发展，我国的算网融合也取得了长足的进步。算力网络，是一种根据业务需求，在云、网、边之间按需分配和灵活调度计算资源、存储资源以及网络资源的新型信息基础设施。算力网络体系包括算力度量、算力感知、算力路由、算力编排、算力交易等内容。

目前，中国联通、中国移动、中国电信的算网融合战略很清晰，标准也很清楚，他们将通过实施算网融合战略转型为算力供应商。

6、用算力服务标准确保算力服务健康发展

对于未来的展望讲过很多，东数西算工程标志着算力经济时代正式的拉开帷幕。未来，算力将加速普及，类似于电力插座变成算力插座。我们使用算力不需要带一台电脑，随便一个卡或者一个东西，就可以通过一个标准的计量方式来使用算力。未来还可能会出现类似于发电厂的算力工厂，尤其在西部地区会出现，据说在煤矿、水电站的附近已经开始建设算力工厂，电力极其便宜，成本特别低。

工业时代有公路、电网，算力时代也有算力网络。随着算力服务的发展，未来在算网时代有三类不同角色：一是网络通信商，通过算网融合参与进来；另外超算的供应商、云计算供应商，通过超算互联网也会参与提供算力服务；还有国家电网通过建设发电厂，参与提供算力服务。三类角色从不同的技术途径抢占算力服务市场。基于此，市场也在呼唤算力服务标准，确保算力健康发展。新一年的数字中国万里行即将开启，希望有更多的力量参与到算力+产业的考察实践中，推动中国算力经济的发展和升级。





CHAPTER1

算力经济时代的 基础设施新价值

2023 新型算力中心调研报告

第一章

算力经济时代的基础设施新价值

2023 年始，ChatGPT 和 GPT-4 再次掀起了人工智能的热潮，并打开了海量的应用场景：生成应用和布局、搜索和数据分析、程序生成和分析、文本生成、内容创作……ChatGPT 基于其庞大的算力和算法分析，可覆盖教育、科研、新闻、游戏等行业。

从 2018 年第一代生成式预训练模型 GPT-1 诞生以来，GPT 系列模型几乎按照每年一代的速度进行迭代升级，2022 年以来，新的通用人工智能开始以更加高效的方式解决海量的开放式任务，它更加接近人的智能，而且能够产生有智慧的内容，也带来了新的研究范式——基于一个非常强大的多模态基础模型，通过强化学习和人的反馈，不断解锁模型的新能力。

ChatGPT 是 AI 大模型创新从量变到质变长期积累的结果，是通用人工智能（AGI, Artificial General Intelligence）发展的重要里程碑。以 GPT-4 为例，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，人们通过输入提示词和多模态内容，便可生成多模态数据。更重要的是，它可以用自然语言方式生成任务描述，以非常灵活的方式应对大量长尾问题和开放性任务，甚至是一些主观的描述。

“大模型 + 大算力 + 大数据”成为迈向通用人工智能的一条可行路径，比如大模型技术是自动驾驶行业近年的热议趋势。自动驾驶多模态大模型可以做到感知和决策一体化。在输出端，通过环境解码器可对 3D 环境进行重建，实现环境的可视化理解；行为解码可生成完整的路径规划；同时，动机解码器可以用自然语言描述推理的过程，进而使自动驾驶系统变得可以解释。

而大规模深度学习模型的参数和数据量达到了一定量级，超大规模 AI 大模型的训练一般必须在拥有成百上千加速卡的 AI 服务器集群上进行，需要相应算力的支撑。根据 OpenAI 的数据，GPT-3 XL 参数规模为 13.2 亿，训练所需算力为 27.5PFlop/s-day。由于 ChatGPT 是在 13 亿参数的 InstructGPT 基础上微调而来，参数量与 GPT-3 XL 接近，因此预计 ChatGPT 训练所需算力约为 27.5PFlop/s-day。



新的通用人工智能开始以更加高效的方式解决海量的开放式任务，它更加接近人的智能，而且能够产生有智慧的内容，也带来了新的研究范式——基于一个非常强大的多模态基础模型，通过强化学习和人的反馈，不断解锁模型的新能力。



国内 AI 大模型项目发布情况汇总

企业	AI 名称	发布情况	具体发布日期
百度	文心千帆	2023 年 3 月 16 日	2023 年 3 月 16 日
华为	盘古 NLP 模型	2023 年 4 月 10 日	未知
昆仑万维	天工 3.5	2023 年 4 月 17 日测试	未知
搜狗	百川智能	2023 年 4 月 10 日	预计 2023 年底
字节跳动	My AI	2023 年 4 月 11 日	2023 年 4 月 11 日
阿里巴巴	通义千问	2023 年 4 月 11 日	2023 年 4 月 11 日
360	360 智脑	2023 年 4 月 10 日	2023 年 4 月 10 日
商汤科技	日日新	2023 年 4 月 10 日	2023 年 4 月 10 日
腾讯	混元	2023 年 4 月	预计 2023 年内
科大讯飞	1+N 智能大模型	2022 年 12 月	2023 年 5 月 6 日
京东	言犀产业大模型	2023 年 2 月 10 日发布 125 计划	预计 2023 年内
清华大学	ChatGLM-6B	2023 年 3 月 28 日	2023 年 3 月 28 日
复旦大学	MOSS	2023 年 2 月 20 日	2023 年 2 月 20 日
达观数据	曹植	2023 年 3 月 18 日公布试用	未知
网易	玉言	2023 年 1 月 17 日测试	未知
澜舟科技	孟子	2023 年 3 月 14 日	2023 年 3 月 14 日
中科院自动化所	紫东太初	2021 年 9 月 27 日	2021 年 9 月 27 日
智源研究院	悟道 2.0	2021 年 6 月 1 日	2021 年 6 月 1 日
知乎	知海图 AI	2023 年 4 月 13 日发布	2023 年 4 月 13 日
心识宇宙	MindOS	2022 年 11 月内测	2023 年 1 月上线
MiniMax	Glow	2023 年 2 月 16 日	2023 年 2 月 16 日

(截止 4 月份国内 AI 大模型项目发布情况汇总, 信息来源网络 益企研究院整理)



同样，算力作为自动驾驶的基本要素，从视觉检测、传感器融合、轨迹预测到行车规划，上万个算法模型需要同时完成高并发的并行计算，需要更高性能的智算中心来完成训练、标注等工作。从 2022 年开始，人工智能算力成为主要增量，数字中国万里行考察期间，小鹏汽车和阿里云共同发布在乌兰察布合建当时国内最大的自动驾驶智算中心“扶摇”，专门用于自动驾驶模型训练，算力规模达 600PFLOPS，相当于每秒可以完成 60 亿亿次浮点运算。

从 2018 年开始，益企研究院（E 企研究院）开启数字中国万里行，几年来，数字中国万里行的足迹遍布“全国一体化大数据中心”体系下的 8 个枢纽节点，出发点切合了国家后来提出“新基建”，路线选择和洞察也与国家“东数西算”工程的规划高度契合，深入实地对风、光、储能的考察符合“双碳战略”。

结合算力经济时代的算力基础设施发展，我们认为以下几个方向值得讨论。



+ + + +

多类算力基础设施并行发展

迄今为止，数字中国万里行已经考察了位于全国一体化算力网络十大数据中心集群中的多个不同类型数据中心，包含：互联网 / 云计算数据中心、金融数据中心、运营商数据中心、第三方 IDC、超算中心、智算中心。2022 年，我国算力基础设施迎来了多样化发展的繁荣期，从数据中心承载的应用来看，需要多类算力基础设施并行发展，保障算力资源的多元供给。

1、云数据中心加速算力普惠

过去几年，云计算行业均处于蓬勃发展阶段，技术演进结合客户需求释放，推动市场规模加速增长，促使云服务商加大全球数据中心布局。从全球来看，在过去三年对数字化转型进行了持续的 IT 投资后，通货膨胀推动公共云成本不断上升，迫使企业客户优化公共云支出。宏观经济的不确定性导致信息技术预算采用更加保守的方案。越来越多的客户正在调整云策略，以提高效率和控制能力，在 2022 年，云基础设施服务的增长开始变缓。从 Canalys 的数据来看，2022 年全年，云基础设施服务总支出从 2021 的 1917 亿美元增长至 2471 亿美元，增幅达 29%。季度增长率放缓，2022 年第一季度为 34%，2022 年第四季度为 23%。Canalys 预计，在未来几个季度，云基础设施服务的增长速度将继续放缓。2023 年，全球云基础设施服务支出将增长 23%。

同样，Synergy Research Group 的数据显示，2022 年第四季度全球企业在云基础设施服务方面的支出超过 610 亿美元。从数据来看，比 2021 年第四季度增长了 100 多亿美元，前四季度的平均增长率为 31%。由于市场规模越来越大，Synergy 认为增长率的下降在一定程度上是意料之中的，但毫无疑问，当前的经济环境也产生了不利影响。

而对于中国市场而言，2022 年是保守的一年，传统云服务商市场增长了 10%，总额达到 303 亿美元。Canalys 数据显示，2022 年第四季度，云计算支出总额为 79 亿美元，同比增长 4%。与过去几年的强劲表现（前三年的年增长率超过 30%）相比，2022 年的增长率大幅下降。Canalys 预计，2023 年，中国云基础设施服务支出将增长 12%。

303 亿美元

10% ↑

2022 年，中国传统云服务商市场

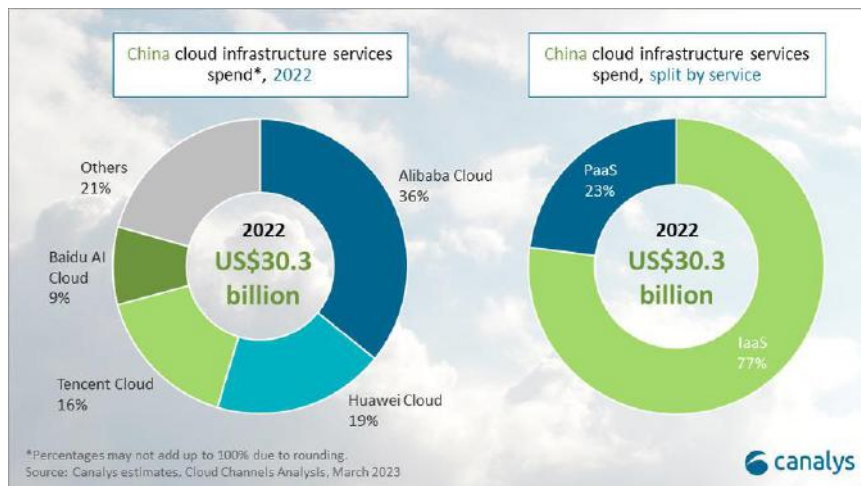
79 亿美元

4% ↑

季度增长率放缓

12% ↑

2023 年，中国云基础设施服务支出将增长 12%



<https://canalys.com/newsroom/china-cloud-market-Q4-2022>

疫情及其限制的影响不容忽视，但实际上，云计算行业增长动力逐步由互联网转向传统企业。政企客户对于云服务的安全、可控要求较高，再加上国资云、算力网络等新基建相关政策，电信运营商云接棒互联网巨头成为政企行业上云的 IaaS 服务主力军。

从中国移动、中国电信、中国联通 2022 年年报业绩来看，三家企业营收、净利润均实现增长，云计算成为拉动增长的主力，2022 年：

- 中国电信天翼云营收 579 亿元，同比增长 108%；
- 联通云营收 361 亿元，同比增长 121%；
- 移动云营收 503 亿元，同比增长 108%。

作为算力的聚集点，云数据中心的规模化效应使得算力得以普惠化，用户按需采购算力、存储、带宽即可开展业务。随着国内大模型市场的快速发展对我国的基础算力提出更高的要求，没有算力基础，算法等发展难以为继。此时，云计算厂商的算力基础设施优势凸显，大模型的爆发会导致训练的应用场景越来越多，对训练的需求大幅增长，如何保证算力不衰减，对算力的高带宽、存算一体等提出新要求，需要底层平台 + 分布式框架 + 加速算法的高效集成。2023 年，云计算厂商开始发布人工智能大模型，4 月份，阿里云通过官方微信公众号官宣了旗下的超大规模语言模型；华为云也介绍了华为盘古大模型的架构以及应用场景，还有在矿山、铁路、气象、医药分子等细分行业的应用。

579 亿元
108% ↑

361 亿元
121% ↑

503 亿元
108% ↑



53.9 亿美元

2021 年中国 AI 服务器市场规模

103.4 亿美元

预计 2025 年达到

17.7% ↑

2021~2025 年 CAGR 达

155.2 EFLOPS

2021 年中国智能算力规模

922.8 EFLOPS

预计 2025 年达到

56.15% ↑

2021~2025 年 CAGR 达

未来，云数据中心的的核心依然是：让算力更加普惠，促使 AI 大规模普及。全方位的算力服务能力依然是云服务商竞争力的基石，算力基础设施的使用效率，会直接影响到云服务商的创新能力和盈利能力。另外，大模型是一场“AI+ 云计算”的全方位竞争，超千亿参数的大模型研发，并不仅仅是算法问题，而是囊括了底层庞大算力、网络、大数据、机器学习等诸多领域的复杂系统性工程，需要有超大规模 AI 基础设施的支撑。因此，云服务商不断优化硬件基础设施提升算力效率，提供通用计算、智能计算能力，通过云统一管理多种算力，灵活调度算力资源，并形成完整的产业生态，推动新兴产业发展。

2、智算中心加快智能算力部署

智算中心是服务于人工智能的数据计算中心，采用领先的人工智能计算架构，提供人工智能应用所需算力服务、数据服务和算法服务的公共算力新型基础设施。2022 年，智算中心作为发展最快的一种算力供给形式，全球人工智能算力成为主要增量。据 IDC 统计，2021 年中国 AI 服务器市场规模为 53.9 亿美元，预计 2025 年达到 103.4 亿美元，2021~2025 年 CAGR 达 17.7%；2021 年中国智能算力规模为 155.2EFLOPS，预计 2025 年达 922.8EFLOPS，2021~2025 年 CAGR 达 56.15%。

在中国，智算中心发展尚处于初期阶段但发展迅速。从国家信息中心发布的《智能计算中心创新发展指南》来看，当前我国超过 30 个城



商汤上海临港人工智能计算中心（AIDC）

市正在建设或提出建设智算中心，整体布局以东部地区为主，并逐渐向中西部地区拓展。

智算中心建设目的促进产业 AI 化、AI 产业化，主要应用在城市治理、智能制造、自动驾驶等领域。2023 年火热的大模型计算的需求加速了算力的商业应用以及智算中心的发展。无论是智慧城市还是智能制造、无人驾驶、数字孪生等场景，除了要有数据支撑以外，还要和各领域、各场景的知识模型、机理模型甚至物理模型相叠加，形成基于人工智能的新应用和场景实现。以 AI 芯片为主的高效率、低成本、大规模的智能算力基础设施将成为训练 AI 大模型的前提。比如商汤科技发布多模态多任务通用大模型“书生（INTERN）2.5”，其图文跨模态开放任务处理能力可为自动驾驶、机器人等通用场景任务提供高效精准的感知和理解能力支持。

320 亿

参数规模的通用视觉模型

2000 亿

参数规模的大模型

1800 亿

参数的多模态大模型

多任务、多模态的能力需要强大的算力基础设施，以数字中国万里行参观的商汤上海临港人工智能计算中心（AIDC）一期为例，作为 SenseCore 商汤 AI 大装置的算力基座，AIDC 基于 2.7 万块 GPU 的并行计算系统实现了 5.0 exaFLOPS 的算力输出，可支持最多 20 个千亿参数量超大模型（以千卡并行）同时训练。目前商汤有 320 亿参数规模的通用视觉模型，在 NLP 领域也有接近 2000 亿参数的大模型，有能力去训练 1800 亿参数的多模态大模型。

大模型进一步促进智算中心的发展。智算中心有技术实现复杂、建设周期长、资源投入巨大、产业辐射面广的特点。一方面，智能算力需求呈现几何式增长，本地智算中心主要服务本地产业和科研机构，无

+ + + + +
+ + + + +



超算算力是基于超级计算机等计算集群所提供的高性能计算能力，通过各种互联技术将多个计算机系统连接在一起，利用所有被连接系统的综合计算能力来处理大型计算问题，所以又通常被称为高性能计算集群。目前已有 11 个国家级超算中心，多个省级超算中心和高校级超算中心。

法向全国提供算力服务。另一方面，为了提供相匹配的超大规模的算力支撑，通过算力的生产、聚合、调度和释放，支撑产业创新聚集，亟需构建云化的智能算力网络，通过情况和各地区的需求情况进行算力动态调配，确保已建成的人工智能计算中心保持高效运营。

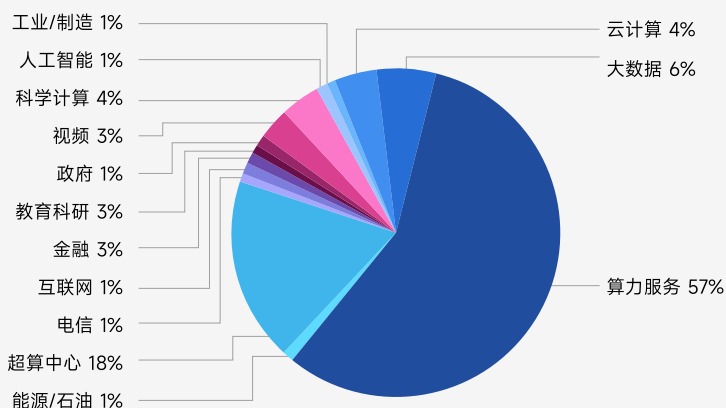
3、超算中心产业化

超算算力是基于超级计算机等计算集群所提供的高性能计算能力，通过各种互联技术将多个计算机系统连接在一起，利用所有被连接系统的综合计算能力来处理大型计算问题，所以又通常被称为高性能计算集群。目前已有 11 个国家级超算中心，多个省级超算中心和高校级超算中心。

一般来说，超算中心主要面向科研和科学计算进行计算密集型的任务处理，应用在基础学科研究、模拟仿真、气象环境、天文地理等领域。科学计算是大模型之外，AI 发展的另一重要方向，借助 HPC，科学计算对基础科学研究和行业发展起到重大的推动作用。随着业务场景越来越复杂，AI+HPC 的算力融合成为趋势。

2022 年，超算商业化进程不断提速，我国超算进入到以应用为需求导向的发展阶段。国内很多超算中心加强了商业化运行改革，算力服务异军突起，加速科研创新，以云服务方式提供通用超算资源，为拓展科学边界、推进技术创新提供了更强劲的动力。从 2022 年中国高性能计算机性能 TOP100 排行榜来看，应用于“算力服务”的系统性能份额占比达到 57%，超算中心、大数据、云计算、科学计算、视频应用分别以 18%、6%、4%、4%、3% 排在其后。

应用领域性能份额



+ + + +



在应用领域新增算力服务，充分反映了在大数据、人工智能算法和算力三驾马车协同配合时代中算力经济的发展，算力的多样化正成为高性能计算领域的发展趋势。

目前，国家也重视超算互联网工程，整合多个超算中心和云计算中心的软硬件资源，平衡算力的需求与供给，通过建设超算资源共享与交易平台，支持算力、数据、软件、应用等资源的共享与交易，同时向用户提供多样化的算力服务。

多元算力与高速互联

自动驾驶、云游戏、短视频、人工智能等应用场景呈现多样化，使得数据中心侧传统单一的结构难以满足要求。而随着非结构化数据占比增大，原来可以用数据库二维表结构实现的结构化数据，现在需要对海量、多种多样非结构化数据（如文本、图片、语音、视频）进行加工、处理，自然需要多样性计算来进行匹配。

多样性计算需求，加速算力格局变换。基于 x86 的通用计算继续构建数字经济发展的基础，依然保持计算的核心地位。一方面继续提供更强的核心和更多的核心数满足客户不同场景需求，如第四代 AMD EPYC 处理器基于业界领先的 5nm 的制程工艺，提供多达 96 个“Zen 4”架构核心、192 线程，以及最大 384MB 的 L3 缓存容量。另一方面，在 AI 应用的规模化部署和实践中发挥重要的作用。为了更加充分地利用 CPU 的资源，几年前英特尔就在 CPU 中内置针对 AI 进行加速的专用运算单元或指令集，英特尔第四代至强可扩展处理器新集成 5 种加速器，并搭配以更为简单易用、能够降低部署和优化难度的软件工具。

而在 Arm 阵营中，算力继续快速延伸至服务器市场，目前在国外，基于 Arm 指令集兼容架构的服务器芯片厂商主要有 NVIDIA、Ampere Computing、亚马逊和富士通。NVIDIA Grace CPU 基于最新的 Armv9 架构，为 AI、HPC、云计算和超大规模应用而设计。如 Ampere Computing（安晟培半导体）致力于为数据中心带来创新的云原生处理器，基于 Arm 架构的 Ampere Altra 产品系列包括 80 核的 Ampere Altra 和 128 核的 Ampere Altra Max，并最新推出基于 192 个自研核的 AmpereOne。目前国内腾讯云、阿里云、优刻得 UCloud、京东、字节跳动等多个超大规模客户的数据中心已在



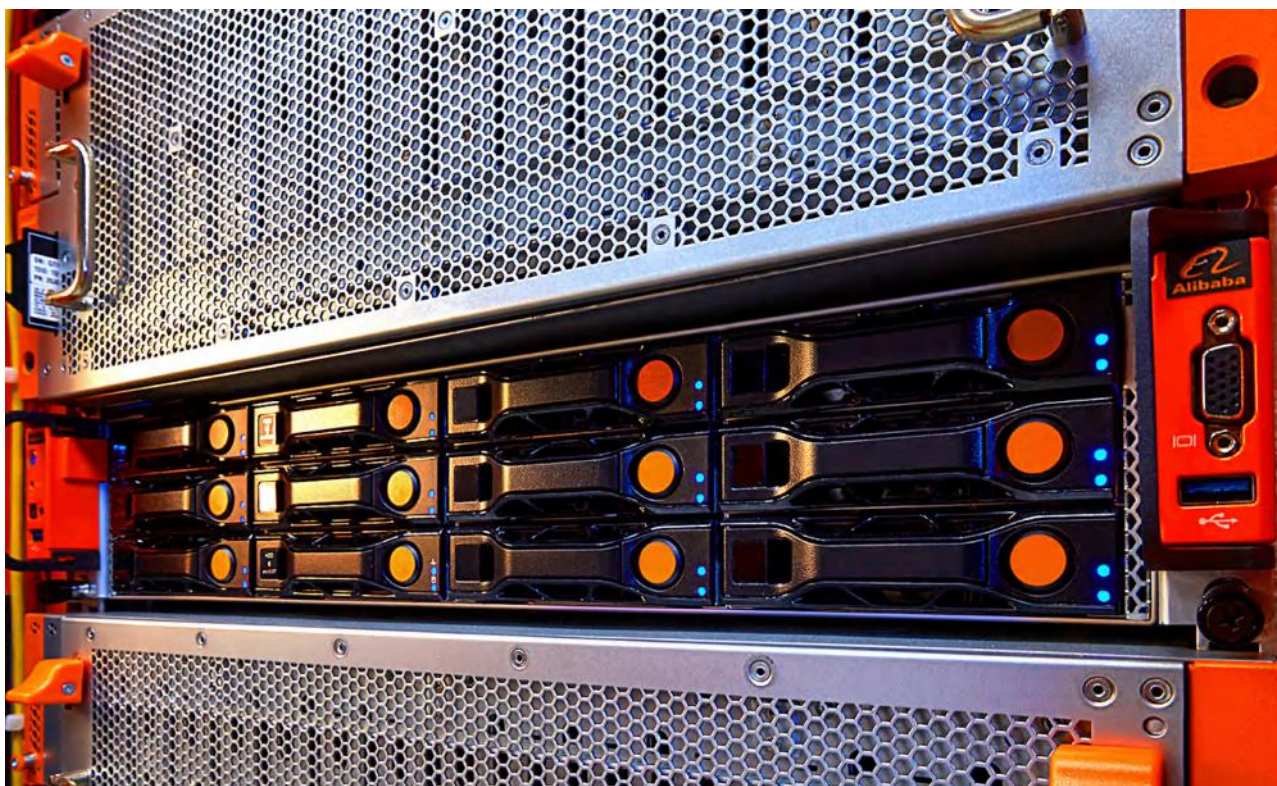
部署 Ampere Computing 产品。

亚马逊云科技 (AWS) 也发布采用了 Graviton 3 的 C7g 应用实例，成为业界首款采用 Arm Neoverse 架构并支援 DDR5 的云端应用实例。

在国内，鲲鹏、飞腾耕耘市场多年，Arm 服务器市场份额持续增加。同时，Arm 解决方案已经在云服务商、高性能计算领域发挥重要作用。

目前云数据中心领域正在进行 x86 + Arm 多元算力的布局。阿里巴巴浙江云计算仁和液冷数据中心已经大规模应用自研 CPU 芯片倚天 710 以及搭载倚天 710 的阿里云自研磐久服务器。在 2022 云栖大会期间，阿里云宣布搭载倚天 710 芯片的阿里云弹性计算实例正式上线，从现场官方公布的数据来看，在新型云计算架构体系下，倚天 + 飞天 + CIPU 的组合表现亮眼，在大数据和 AI 及高性能计算、视频编解码等场景下性能可提升 20% 以上。

腾讯云 CVM 标准型实例 SR1，基于主频达 2.8GHz 的 Ampere Altra 处理器，结合全新优化虚拟化平台，提供了平衡、稳定的计算、内存和网络资源。





+ + + + +
+ + + + +

飞腾系列 CPU 也是基于 Arm 指令集兼容架构设计的处理器，共推出高性能服务器 CPU、高效能桌面 CPU 和高端嵌入式 CPU 等多个系列。

数字中国万里行在顺义考察了中国电子按照国家关键信息基础设施标准打造的中国电子信创云基地，支撑异构多节点云的管理，基于飞腾 Arm 架构和 x86 架构构建云平台资源池，其中国产化飞腾 Arm 体系满足国家安全规定，自主安全要求的信创基础设施资源池；x86 体系的资源，作为现有部分适配难度较大的业务运行的非信创过渡资源池，服务诸多央企和政府用户。

在高性能计算领域，从全球来看，全球超级计算机 TOP500 排行榜中，已有 5 台基于 Arm 指令集兼容架构处理器构建的超级计算机入围。同时，美国、日本、欧洲也都发布了多台基于 Arm 指令集兼容架构处理器的超级计算机建设计划，Arm 指令集兼容架构有望成为未来 HPC 的主流技术和发展趋势。

全球高性能计算机 TOP500 排行榜中基于 Arm 指令集兼容架构处理器的超级计算机

超算名称	SC2022 TOP500 排名	峰值性能	CPU 数量 / 核数	处理器型号	处理器架构	部署地	部署年份
Fugaku	2	537.21 PFlop/s	158976/48	A64FX 48C 2.2GHz	Armv8.2-A SVE 512 位	日本理研计算科学中心	2020
Wisteria	23	25.95 PFlop/s	7680/48	A64FX 48C 2.2GHz	Armv8.2-A SVE 512 位	日本东京大学信息技术中心	2021
TOKI-SORA	39	19.46 PFlop/s	5760/48	A64FX 48C 2.2GHz	Armv8.2-A SVE 512 位	日本宇宙航空工业振兴机构	2020
Flow	87	7.79 PFlop/s	2304/48	A64FX 48C 2.2GHz	Armv8.2-A SVE 512 位	日本名古屋大学信息技术中心	2020
Astra	467	2.30 PFlop/s	3990/36	Marvell ThunderX2 CN9975-2000 28C 2GHz	Armv8.1	美国桑迪亚国家实验室	2018

数据来源：根据 SC2022 TOP500 排名整理

基于 Arm 指令集兼容架构处理器的超级计算机进入全球超级计算机 TOP500 排行榜，已经很大程度上彰显出 Arm 指令集兼容架构在高性能计算领域的潜力。

在国内，以 Arm、RISC-V 为代表的多样性计算平台逐渐发挥重要作用。基于华为鲲鹏 920 CPU 的 TaiShan 服务器基于 Arm 指令集兼容架构的高性能处理器，面向高性能计算、大数据、分布式存储和 Arm 原生应用等场景，能够充分发挥 Arm 指令集兼容架构在多核、高能效等方面的优势。

+ + + +



+ + + + +
+ + + + +

比如上海交通大学“交我算”校级计算平台，在上海交通大学闵行校区的网络信息中心上线了国内高校首台基于鲲鹏处理器的集群系统。

“交我算”的鲲鹏集群共 100 个计算节点，节点采用双路 64 核华为鲲鹏 920 处理器，每个计算节点拥有 128 核处理器和 256GB 内存，总计 12800 核，系统理论双精度峰值性能达 133TFLOPS，覆盖了材料科学、生命科学和流体力学等多个高性能计算应用领域。

在智能计算场景领域，以 CPU + AI 芯片（GPU、FPGA、ASIC）提供的异构算力，并行计算能力优越、互联带宽高，可以支持 AI 计算效力实现最大化，成为智能计算的主流解决方案。人工智能算法需要从海量的图像、语音、视频等非结构化数据中挖掘信息。从大模型的训练、场景化的微调以及推理应用场景，都需要算力支撑。在大模型层面，以 GPU 等 AI 训练芯片为主，为 AI 计算提供更大的计算规模和更快的计算速度。



算力服务成为一种新的业态，将通用计算、智能计算、并行计算等多样性算力统一纳管和调度，屏蔽不同硬件架构差异，实现大规模异构计算资源的统一调度，实现算力的普惠化。

除了大模型，目前在 AI for Science 领域，人工智能正在给科学计算带来重大的范式革命。AI for Science 的数据来自各个学科的数据积累，模型来自各领域科学家发现的科学原理和规律；算法源自机器学习算法和数值方法等方面的创新；需要多样算力融合的综合型智能计算平台，通过分布式异构并行体系结构，实现多样算力的融合、优势互补，为 AI 训练、AI 推理、数值模拟等不同应用提供不同算力，实现高精度到低精度算力的全覆盖、多种计算类型的全覆盖，以及 AI 训练 + 推理全覆盖。

多元算力的多元开发生态体系相对独立，应用的跨架构开发和迁移困难，需通过开源、开放的方式建立可屏蔽底层硬件差异的统一异构开发平台。算力服务成为一种新的业态，将通用计算、智能计算、并行计算等多样性算力统一纳管和调度，屏蔽不同硬件架构差异，实现大规模异构计算资源的统一调度，实现算力的普惠化。同时，当算力和网络的发展呈现一体共生之势时，从算网协同到算网融合，业务需求的变化会通过芯片、计算和存储等 IT 设备传导到网络架构层面，即数据中心作为基础设施也会相应的产生自上而下的变化。为此，除了算力网络，数字中国万里行考察期间也重点关注 DPU/IPU 乃至芯片间的互连，展现数据中心基础设施如何应对这些变化与挑战，更好的服务于用户，并可持续的良性发展。



+ + + +



从“东数西存”到“东数西算”，促使更多行业和企业重视数据，带动数据存储、管理、使用的需求增长。用户对数据存储容量、数据传输速度、硬件设备性能等各方面有了新的认知。

+ + + + +
+ + + + +

高效绿色的数据存储与管理

2022 年的“东数西算”工程在实现数据中心一体化协同创新的战略价值被大家认同，东数西算是“全国一体化算力网络”下辖的一个子概念，而后者旨在推进技术融合、业务融合、数据融合，实现跨层级、跨地域、跨系统、跨部门、跨业务的协同管理和服务。从“东数西存”到“东数西算”，促使更多行业和企业重视数据，带动数据存储、管理、使用的需求增长。用户对数据存储容量、数据传输速度、硬件设备性能等各方面有了新的认知。

有了算力，业界也提出了“存力”这一概念。但实际上，定义存力这一概念较难，涉及的维度较多。我们可以认为：算力的底层支撑为计算芯片，同理，存力的底层支撑则为存储器介质（DRAM、NAND Flash SSD、硬盘等）。存力可以通过存储服务器或存储系统来承载，存力最基本的度量至少需要包括容量、性能两个维度。

尤其对于超大规模数据中心而言，需要突破 SSD/ 硬盘容量和瓶颈的同时提升服务器 / 存储系统的可扩展能力，需要构建高可靠、低成本的存储方案与服务，有效地激活数据价值。

在服务器中，大容量机械硬盘是海量数据的有效载体。机械硬盘的容量在持续增长。数字中国万里行发现，目前希捷的企业级银河系列 20TB 硬盘已经开始大量部署。预计 2023 年底，希捷将发布 30TB 容量的硬盘。随着硬盘容量的不断增长，系统散热、风扇设计、噪音振动等挑战接踵而至，对服务器架构的设计提出了更高的要求，硬盘厂商与服务器厂商需要更紧密协作，寻求硬盘和服务器的“更优兼容”，以保证整体解决方案的性能和稳定性。

而在有些场景中，机械硬盘无法满足现代工作负载对于数据访问更高速度的需求，同时机械硬盘还会占据数据中心的较大空间，会增加空间、电源、散热和备件更换方面的成本。为了追求更高的带宽、更短的延迟，SSD 的应用日趋广泛。SSD 擅长应用在高 IOPS、高吞吐量的场景中，常见的如数据库和云计算 / 虚拟化，以及热门的 AI、高性能计算，还有搜索等。

虽然 SSD 的性能增长速度、成本下降速度远远超过机械硬盘的发展速度，但到目前为止，SSD 的单位容量价格依然与硬盘有着数量级的差距。SSD 与硬盘各自的特点需要各自继续发展，而彼此之间的落差，也急需填补。



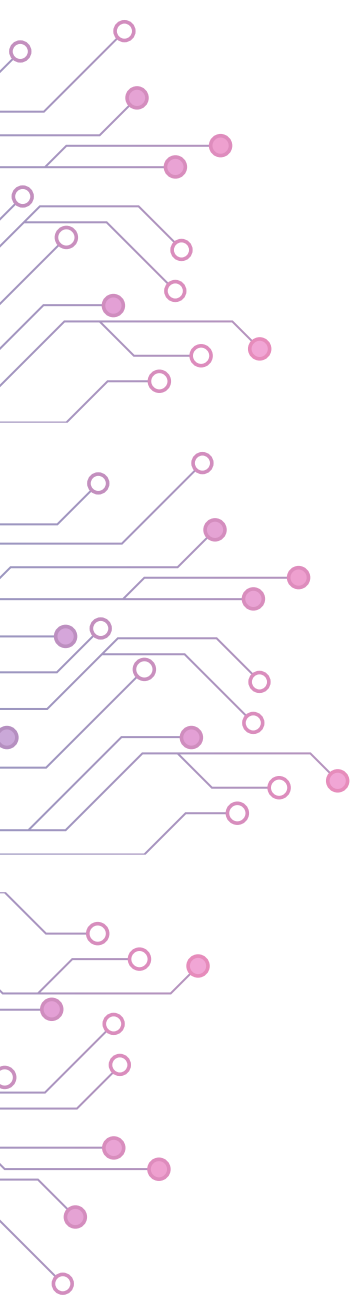
+ + + + +
+ + + + +

从硬盘角度，值得一提的是希捷的热辅助磁记录技术（英文缩写为“HAMR”），它是未来实现更高容量硬盘的关键技术。该技术通过不断增加硬盘的面密度和存储容量，打造新一代高性能硬盘。

HAMR 硬盘在读写速度、性能、可靠性、稳定性等方面均表现卓越，是非常重要的一个存储技术创新，增加了可用存储区域的数据存储量——即俗称的磁盘“磁密度”。磁密度的提升将有助于在下一个十年推动硬盘产品的发展和增长。

根据希捷最新公布的技术路线图，HAMR 技术将帮助希捷在未来四年中翻倍提升硬盘单碟容量，为市场提供更大容量的存储产品，同时降低数据存储成本。以今天 20TB 硬盘为例，目前每个碟片承载 2TB 的容量，需要 10 个碟片，而在 4 年后，5 个碟片就可以实现 20TB 的存储量，总体拥有成本将得到显著降低。

SSD 的发展也多面开花。其一，通过接口、控制器的迭代（如正在进入市场的 PCIe 5.0 接口）继续提升 SSD 的整体性能。其二，通过 NAND Flash 介质的 3D 堆叠进一步提升存储密度和单芯片的接口速度。譬如在 2022 年中，各大厂商普遍将堆叠层数推进到了 200 层以上，在 2023 年初的 ISSCC 会议上，SK 海力士还发表了关于 300 层产品的论文。其三，NAND 的多值化进一步提升了 SSD 的容量并降低单位容量成本，譬如 Intel 在 2020 年通过 QLC 技术将 SSD 单盘容量推至 30.72TB；在 2023 年，Solidigm 推出第四代 QLC NAND，堆叠 192 层，单芯片容量达到 1.3Tb，SSD 的单盘容量将会



进一步提升；面向未来的 PLC 技术，也已经有了样品，在刚刚结束的 CFMS2023 中，Solidigm 宣称 PLC SSD 进行 1000PE 擦写和高温老化后的数据保持能力依旧可以满足需求。不断追求高密度、低成本、低功耗，符合双碳政策引导之下的绿色数据中心需求。

从 SSD 厂商视角，“硬盘替代”是 QLC、PLC 甚至 HLC 等技术不断发展的驱动力。硬盘厂商则希望将单位容量成本的优势尽量延续。而对于用户而言，在成本制约下，“存力”的容量与性能两个维度是存在矛盾的——追求容量的场景，适合部署大量的硬盘，但需要付出空间和性能的代价；追求性能的场景，对 SSD 的使用需要精打细算、物尽其用。

当然，更现实的情况是结合 SSD 和硬盘的特点，进行混合部署。二者的结合，也从早期的“分工协作”（将不同特点的业务安排在不同的阵列 / 节点），逐步演化为“取长补短”（存储分层）。软件定义存储颠覆了传统的应用观念，存储性能的分层对最终用户趋于透明，主要基于硬件冗余实现的存储安全机制也被重构，从而释放出更多资源（容量、性能）供用户使用。可以说，如何充分释放“存力”的价值，应用水平、运维管理能力也是至关重要，可以将其视为度量“存力”的“隐形维度”。

数据中心高效的核心是算力和存力的协同调度，计算与存储高度融合，方能充分发挥生产力，真正形成核心竞争力。在从介质到数据中心的绵长产业链条当中，每一个环节都在思考如何为客户提供更大的价值。

高安全数字基础设施是趋势

《数字中国建设整体布局规划》明确，数字中国建设按照“2522”整体框架布局，强调强化数字中国关键能力，构筑自立自强的数字技术创新体系，筑牢可信可控的数字安全屏障。近年来，随着《网络安全法》、《数据安全法》、《个人信息保护法》出台，将我国数据安全保护及管理要求提升至新的高度。同时，“十四五”以来，国家出台多项政策要求加快培育数据要素市场，建立高效共享的普惠型数据要素市场。构建高安全可控的数字基础设施，是维护、夯实数字基础设施和数据资源体系的重要保障，是发展数字经济的重要技术支撑。

根据 IDC 2023 年全球数字政府十大预测，到 2024 年，由于经济和地缘政治事件，45% 的国家政府将认为“数字主权对于保护关键国

80%

2023 年，G5000 基础设施客户将采取积极的多源策略

70%

2025 年，G2000 客户将优先考虑主权云

65%

2026 年，买家将优先考虑基础设施即服务的消费模式

家基础设施以提高国家生存能力至关重要”。数字主权关系到国家的未来，数字主权上升到前所未有的高度。

注释：根据 IDC 的定义，数字主权涵盖多个层次，包括数据主权、技术主权、运营主权、业务可用主权、供应链主权和地域主权。通过多个层次的建设，达到数字主权的不同阶段，最终实现从自控（self-determination）、自足（Self-sufficiency）到自生（Survivability）。

在 IDC《2023 年全球数字基础设施未来十大预测》预测报告中也提到了业务导向、安全的相关的预测，包括：

- 到 2023 年，80% 的 G5000 基础设施客户将采取积极的多源策略来保护自己免受未来 IT 供应风险的影响。
- 到 2025 年，70% 的 G2000 客户将优先考虑主权云的可信基础设施，以确保特定敏感业务、数据的安全性和本地法规遵从性。
- 到 2026 年，65% 的技术买家将优先考虑基础设施即服务的消费模式，以帮助抑制 IT 支出增长，并填补 IT 人才缺口。

注释：G5000 指的是 global 5000，就是全球 5000 强的大公司。

随着国产处理器、国产操作系统、国产数据库的发展和成熟，在党政机构、能源、金融等关键行业领域，实现了高安全数字基础设施的“从无到有、从有到优”，高安全数字基础设施成为建设数字中国的重要力量。

数字中国万里行在考察调研多个政府数据中心和采访国内头部拥有自主技术的厂商后分析得出，高安全数字基础设施包含以下关键要素：

• 可信可控

具备高水平自立自强的数字创新体系，实现在云、计算、存储、网络、安全、数据、智能等关键核心技术攻关，拥有所有的技术资料、知识产权、源代码，云平台中不存在恶意后门并可以不断改进升级，不受制于其他技术壁垒。

• 原生安全

安全效果不能依靠单一技术或产品来解决，需要依靠“系统论”思想，进行体系性建设。通过搭建云平台原生安全、可信安全、云原生安全产品、合规安全等构建可信云原生安全架构。可信云原生安全架构具备四大核心原生安全能力：可信安全、云原生安全、数据原生安全、智能安全。

+ + + + + + +

• 统合算力

通过构建自主可控的算力调度服务平台，逐步开展异构云资源纳管，系统优化算力基础设施布局，对通用算力、超算、智算、边缘数据算力等算力资源进行统一调配，实现数据资源高效配置，数据要素加速流通，数据价值全面释放，数据安全有效保障。

• 数智融通

数据和人工智能是数实融合的关键，数智能力需要融入数字基础设施，构建云、网、智、算融合体系的数字经济基础底座。加大对大数据、人工智能、5G、区块链等数字技术的创新应用，利用 AI 技术激活数据价值，加快释放行业数字化生产力，实现质量、效率和动力变革。

以中国电子云为代表的中国信创云为例，依托中国电子自主计算产业体系，中国电子云走自主技术创新的道路，从云数融合、市场牵引到商业成功，秉承跟随到超越的产品体系理念，在数字基础设施建设运营、数据资源体系规划建设、数字技术的创新应用等方面全面布局，体现出以下优势：

- **全栈自研产品及自主技术。**依托中国电子自主计算体系及丰富的网信产业资源，中国电子云能够纵向打穿整个自主计算产业生态链，通过跨产线、跨企业的组合性产品解决方案，将各个单点优势再结合，形成电子云的整体优势，以云化形式对外输出中国电子整体自主核心技术和产品能力。
- **全栈分布式云原生架构。**中国电子云整体架构体系贯彻云原生、安全原生、数据原生的理念，打造具有分布式云原生、云数融合和原生安全三大关键技术优势的全栈分布式云，不断提升专属云运营质量。通过分布式云原生云操作系统 CCOS，以及软硬一体的“雨燕架构”共同支撑，提供统一技术服务底座。其中，云管理平台与云服务使用 Go 语言全面重构，实现内存开销减少 45%，CPU 开销减少 30%；基于容器微服务的系统高可用，实现云服务与云管平台的全 Operator 化；基于容器操作系统实现计算虚拟化产品，实现容器、虚拟机同平台管理和统一调度能力。
- **灵活部署与规模优势，功能全面和性能提升兼容并蓄。**中国电子云专属云 CECSTACK 可实现单集群、同架构从 3 台到 30 万台平滑线性进化，同时在多集群管理、多集群调度，以及在性能、损



东数西算是促进绿色节能，助力实现碳达峰、碳中和目标的重要手段。

“东数西算”工程聚焦创新节能，在集约化、规模化、绿色化方面着重发力，支持高效供配电技术、制冷技术、节能协同技术研发和应用，鼓励自发自用、微网直供、本地储能等手段提高可再生能源使用率。

耗和灵活性等方面具有优势。例如，通过对大数据计算集群基于云底座的容器化改造，合并大数据集群到云资源池，有效解决潮汐算力问题，提升算力利用率，降低存储空间。

- **落实“云数融合”。**中国电子云”现有产品体系包含三层，一是提供算力基础平台的产品，包括专属云 CECSTACK、超融合 CeaCube、云原生分布式存储 CeaStor、云原生安全 CeaSEC 等；二是提供数据管理平台的产品，包括飞瞰数据中台、飞思 AI 智能中台、云数据库平台 CeaSQL、大数据平台 CeaInsight 等；三是在业务层可提供各种商业模式和业务架构的分布式云全栈全域解决方案，包括运营云、专属云、分支云、边缘云等。同时产品性能具备国内国际竞争力，例如，中国电子云 Ceastor 18116E 全闪存储产品在 SPC-1 认证测试中集群（30 节点）性能 1000 万 IOPS、时延 500μs，在全球分布式存储厂商中位列世界第一。并且具备无限扩展的能力。

作为首个大型央企全栈信创云——数字 CEC，中国电子云采用全栈信创，成为中国信创云的“创新者+实践者”，通过构建安全、高效、协同的“数字 CEC 管理体系”，服务大型央企数字化，打造集团数字化底座，支撑中国电子集团及 687 家成员单位，服务 21 万中国电子员工。2022 年，中国电子云信创产品及技术已经演进为可支撑国家重大项目、支撑关键行业数字化，包括国家部委项目、省级信创云项目，能源、金融等关键行业。

绿色低碳持续推广

东数西算是促进绿色节能，助力实现碳达峰、碳中和目标的重要手段。

“东数西算”工程聚焦创新节能，在集约化、规模化、绿色化方面着重发力，支持高效供配电技术、制冷技术、节能协同技术研发和应用，鼓励自发自用、微网直供、本地储能等手段提高可再生能源使用率，改善数据中心电能利用率（PUE），引导其向清洁低碳、循环利用方向发展，推动数据中心与绿色低碳产业深度融合，建设绿色制造体系和服务体系，力争将绿色生产方式贯彻数据中心全行业全链条，助力我国在 2060 年前实现碳中和目标。

在“东数西算”政策引导下，部分计算业务将逐渐向西部迁移，而那些调用频次高、对网络时延要求极高的业务，又要求数据中心不能离



经济发达地区太远; 还有智能制造、科学探索、生物制药、自动驾驶、数字孪生等场景等基于人工智能的新应用和场景实现, 需要面向 AI 的算力基础设施, 仍需要本地数据中心承担。



强算力通常意味着高能耗。当数据中心的算力大幅度提升, CPU/GPU 功率和服务器的功耗也在增加。在双碳背景下, 数据中心迎来转型的关键期。

强算力通常意味着高能耗。当数据中心的算力大幅度提升, CPU/GPU 功率和服务器的功耗也在增加。在双碳背景下, 数据中心也迎来转型的关键期。

双碳不仅是环保概念, 更是决定技术路线。西部拥有丰富的可再生资源 (风能、太阳能等), 并可利用气候优势来帮助数据中心散热; 东部数据中心绿色化发展则更多需要从节能技术创新、优化节能模式入手, 来降低数据中心的能源消耗。作为更高效的冷却方式, 液冷日益受到广泛关注。

液冷是以液体作为热量传导媒介, 通过冷却液与服务器发热部件直接接触或者间接接触的方式换热, 将热量带走的一种服务器散热技术。目前数据中心液冷典型方式为冷板式液冷和浸没式液冷。

从液冷的优势来看, 可以有效提升服务器的使用效率和稳定性, 实现数据中心节能、降噪, 不受海拔和地域等环境影响, 液冷并有助于提高数据中心单位机柜的服务器密度, 大幅提升数据中心的运算效率, 更适合高密度功率且有节能要求的数据中心。

传统风冷冷却技术成熟, 冷板式冷却技术对数据中心架构和机柜结构所需改变较少, 未来一段时间内, 风液混合成为数据中心首选。



0 能耗

全程用于散热的能耗几乎为零

1.09

PUE 可低至 1.09

7000 万度

每年可节电 7000 万度

浸没式液冷技术需要对数据中心架构做较大调整，更适合新建设的数据中心。

大型互联网和云计算公司主导的超大规模数据中心，将对液冷服务器的普及产生决定性影响。以数字中国万里行团队实地考察的阿里巴巴浙江云计算仁和液冷数据中心为例，有一栋机房楼专用于部署单相浸没式液冷服务器，服务器被浸泡在特殊的绝缘冷却液里，运算产生的热量可被直接吸收，经过与外循环的交换带走，无需风扇、空调、冷机等，全程用于散热的能耗几乎为零——根据官方提供的数据，PUE 可低至 1.09，每年可节电 7000 万度，节约的电力可以供西湖周边所有路灯连续亮 8 年。

能源与算力协同

随着数据中心的计算和处理能力不断加强，对能源的需求也就越来越大。2022 年数字中国万里行考察中发现，云服务商通过技术驱动实现“数据中心节能”和“数据节能”，构建智能、绿色、高效能的基础设施以提升可持续性。

目前东部算力需求旺盛，但东部地区在气候、资源、环境等方面、不太利于低碳、绿色数据中心的建设。通过算力基础设施向西部迁移，可以充分发挥西部地区在气候、能源、环境等方面的优势，引导数据中心向西部资源丰富地区聚集，扩大可再生能源的供给。

当然东部区域也在尽其所能。以长三角区域为例，腾讯云仪征东升数据中心分布式光伏项目已经全容量并网发电。该项目充分利用 8 栋大平层机房楼的屋顶面积，共计安装光伏组件 2 万 8 千多块，总装机容量近 13 兆瓦，是江苏省目前最大的数据中心屋顶分布式光伏项目。每个屋顶还配有光伏组件自动清洗机器人，保持光伏组件清洁度，实现光伏系统的自动化高效运维。

在北京，中国电子信创云基地也最大化利用可再生能源，信创云基地在楼体立面布置了单晶光伏组件，为园区照明办公系统提供电能供应，由绿色能源保证了办公等辅助用电，为降低 PUE 做出了贡献。除此之外，水源热泵技术通过将信创云基地内服务器产生的热能进行回收再利用给办公等辅助区域供热使用，积极响应了国家“双碳”政策要求。

目前，对清洁能源的开发利用还有较大提升空间。由于光伏和风力等可再生能源的不稳定特点，我国西北部地区每年弃风弃光电量约 125





亿度，如果在这些地方依托电厂和电网布局就近建设大型以上数据中心，并利用储能系统和调度系统创新解决稳定负载的柔性供能问题，可以促进可再生能源开发利用，有效降低中西部地区弃风和弃光电量，进一步减少碳排放。

数据中心把能源转化为算力，瓦特转化为数字比特，成为数字化的基础设施。数字中国万里行考察中发现，基于云计算的发展，促进了能源行业的数字化和智能化的发展。加快能源数字化平台建设，可推动能源生产、传输、存储、销售、使用等整个数字化的升级过程，为各级政府“双碳”治理、产业绿色低碳发展提供强有力的支撑。比如中国电子云与华电电科院、华电南自华盾公司合作开发的国内首个行业级自主可控燃机智慧运维云平台正式上线。平台对标国际知名燃机诊断运维平台，全面采用了自主可控的基础软硬件产品和内生安全的中国电子云平台，以“云边部署，多级应用”的原则，采用“1+N”的云边协同架构，在电厂侧重点建设预警诊断、性能分析、运行优化等9大业务模块；在集团侧重点建设决策中心、监管中心等4大中心；在行业侧重点建设燃机诊断运维服务平台和生态，推动了燃气发电行业的数字化、智能化发展，助力传统电厂向智慧电厂升级，支持新产业、新模式、新业态的创新发展，为国家的“双碳”目标和能源安全做出央企应尽的责任和义务。

从2018年到2022年，数字中国万里行始终关注云计算、人工智能高速发展下的技术应用趋势和算力演进。进入算力经济时代，无论是人工智能大模型还是数字经济持续发展，对算力中心提出更高的挑战，建设高效集约、普适普惠的新型基础设施，不仅成为行业共识，行业从业者更是通过实践推动算力向绿色化和集约化方向加速演进。





CHAPTER2

多元算力 CPU + GPU

2023 新型算力中心调研报告

多元算力：CPU + GPU

超级计算（SuperComputing, SC），即人们常说的超算或者高性能计算（High Performance Computing, HPC），被誉为计算机界“皇冠上的明珠”，合称 ABC 的人工智能（Artificial Intelligence, AI）、大数据（Big data）和云计算（Cloud computing）都受益于超算领域的探索。

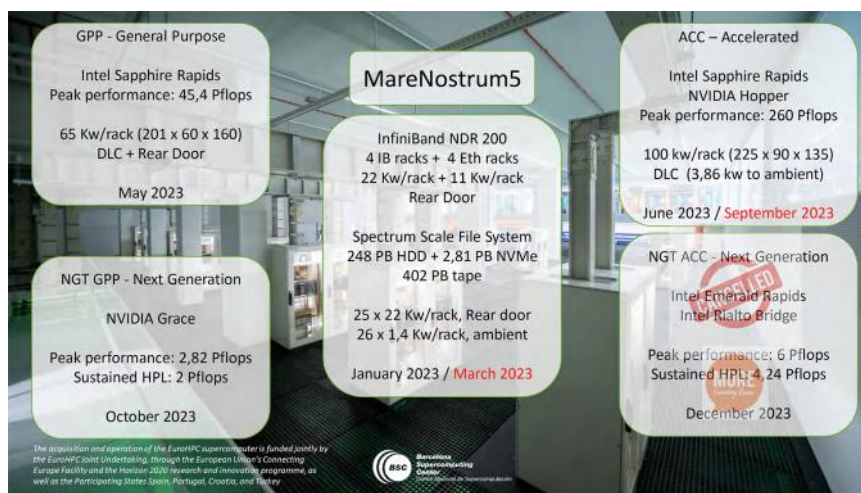
超算系统追求完成（特定）任务所需的算力和效率，为其构建的数据中心（超算中心）通常规模不是很大但具有很高的密度。从数据中心建设的角度，我们可以把云计算中心视为超算中心在通用算力方向上的大规模或超大规模版本，而智算中心与超算中心相比也有以（算力）精度换规模的成分。



超算系统追求完成（特定）任务所需的算力和效率，为其构建的数据中心（超算中心）通常规模不是很大但具有很高的密度。从数据中心建设的角度，我们可以把云计算中心视为超算中心在通用算力方向上的大规模或超大规模版本，而智算中心与超算中心相比也有以（算力）精度换规模的成分。

ChatGPT 的爆火让智算中心的热度再次走高，GPU 更是成为大厂们争抢的对象。GPU 不仅是智算中心的灵魂，在超算领域的应用也越来越普遍。在 2023 年 5 月下旬公布的最新一届 TOP500 榜单中：

- 使用加速器或协处理器的系统从上一届的 179 套增加到 185 套，其中 150 套使用了英伟达（NVIDIA）的 Volta（如 V100）或 Ampere（如 A100）GPU；
- 榜单前 10 名中有 7 套使用了 GPU，前 5 名中也只有第二名没有借力 GPU。



△ MareNostrum 5 的介绍有很多值得关注的信息，譬如 65 千瓦和 100 千瓦的单柜功率，以及冷板式液冷（DLC）和液冷后门

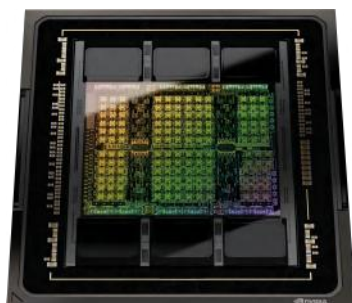


当然，CPU 依然不可或缺。仍以榜单前 10 名为例，AMD EPYC 家族处理器占了 4 套，英特尔至强家族处理器和 IBM 的 POWER9 各占 2 套，Arm 也有 1 套（富士通 A64FX）且高居第二。

通用算力与智能算力相辅相成，可以适应多变的算力需求。以欧洲高性能计算联合事业（EuroHPC JU）正在部署的 MareNostrum 5 为例：基于第四代英特尔至强可扩展处理器的通用算力计划于 2023 年 6 月开放服务，基于 NVIDIA Grace CPU 的“下一代”通用算力，以及第四代英特尔至强可扩展处理器与 NVIDIA Hopper GPU（如 H100）组成的加速算力，也将于 2023 年下半年投入使用。

GPU：大芯片与小芯片

英伟达在 GPU 市场上占据统治地位，不过 AMD 和英特尔也并未放弃。仍以最新的 TOP500 榜单前 10 名为例，4 套基于 AMD EPYC 家族处理器的系统中，搭配 AMD Instinct MI250X 与 NVIDIA A100 的各 2 套，前者的排名还靠前，分居第一、三位。



△ NVIDIA Hopper 架构的 H100 GPU 核心区（die）

但是英伟达 GPU 在 AI 应用上的优势就要显著得多，GTC2022 上发布的 NVIDIA H100 Tensor Core GPU 进一步巩固了其领先地位。H100 GPU 基于英伟达 Hopper 架构，采用台积电（TSMC）N4 制程，具有多达 800 亿晶体管，算、存、连全方位提升：

- 132 个 SM（Streaming Multiprocessor，流式多处理器）、第 4 代 Tensor Core，每时钟周期性能翻倍；
- 比前代更大的 50MB L2 缓存与升级到 HBM3 的显存，组成新的内存子系统；
- 第 4 代 NVLink，总带宽达 900GB/s，支持 NVLink 网络，PCIe 也升级到 5.0。

英特尔也终于在 2023 年 1 月，与第四代英特尔至强可扩展处理器和英特尔至强 CPU Max 系列一起，推出了代号 Ponte Vecchio 的英特尔数据中心 GPU Max 系列。英特尔数据中心 GPU Max 系列利用英特尔的 Foveros 和 EMIB 技术构建，在单个产品上整合 47 个小芯片，集成超过 1000 亿个晶体管，具有多达 408MB 的 L2 缓存和 128GB 的 HBM2e 显存，充分体现了 Chiplet 的理念。

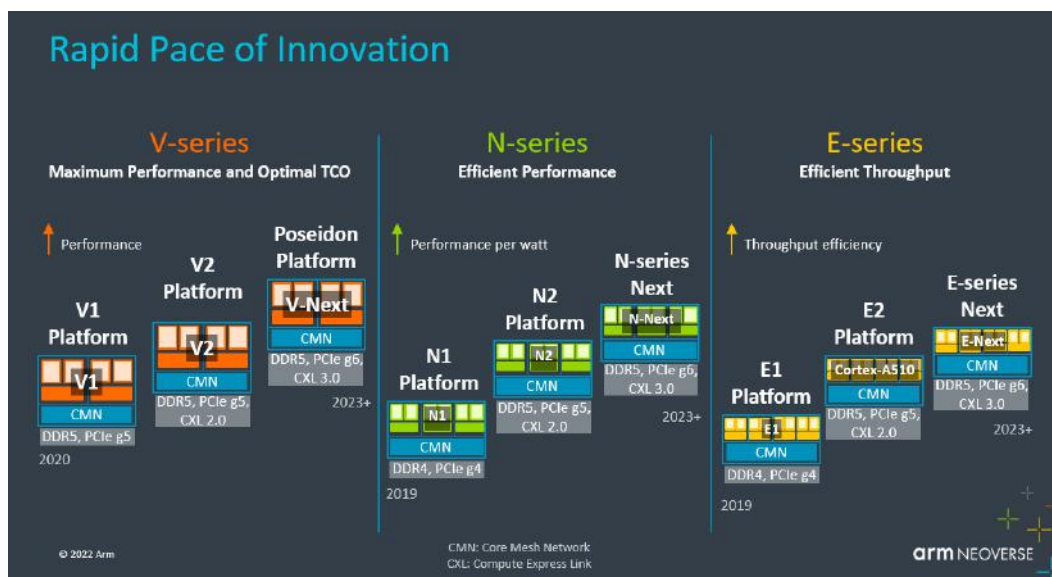


CPU：性能核与能效核

作为通用算力的代表，面对不同应用场景的需求，也渐呈多元化的趋势。先后在手机、PC（含笔记本电脑）等终端产品中得到验证的“大小核”架构，也开始在服务器 CPU 市场形成潮流。当然，服务器的特点是“集群”作战，并不（迫切）需要在同一款 CPU 内部实现大小核搭配，主流厂商正在用全是大核（突出单核性能，偏重纵向扩展）或小核（注重核数密度，偏重横向扩展）的 CPU 去满足不同的客户需求。

作为 big.LITTLE 技术的发明者，Arm 把异构核的理念带入服务器 CPU 市场，也已经有年头了。Arm 的 Neoverse 平台分为三大系列：

- Neoverse V 系列用于打造高性能 CPU，为追求高性能的计算和内存密集型应用程序的系统提供尽可能高的每核心性能。主要面向高性能计算（HPC）、人工智能 / 机器学习（AI/ML）加速等工作负载。
- Neoverse N 系列关注横向扩展性能，提供经过优化的平衡的 CPU 设计，以提供理想的每瓦性能。其主要面向横向扩展云、企业网络、智能网卡 / DPU 和定制 ASIC 加速器、5G 基础设施以及电源和空间受限的边缘场景。
- Neoverse E 系列期望以最小的功耗支持高数据吞吐量，面向网络数据平面处理器、低功耗网关的 5G 部署。



△ Arm Neoverse 三大系列核心架构

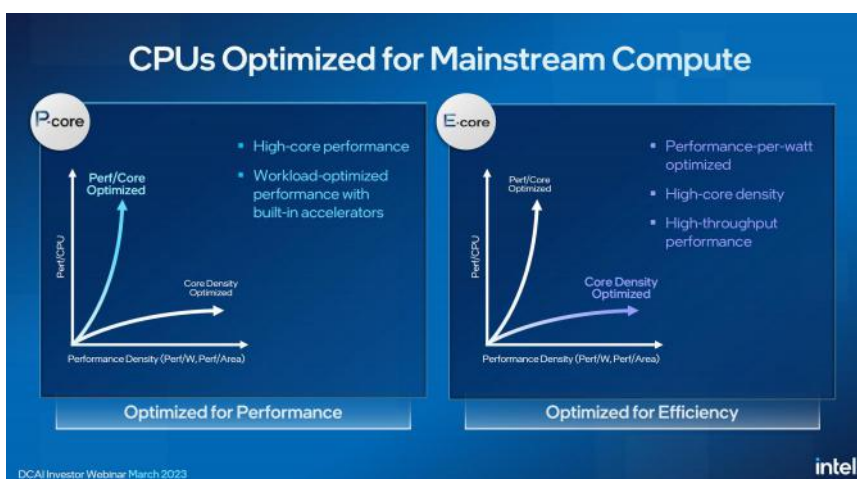


如果把应用场景限定在规模较大的云计算中心和智算中心、超算中心，相对侧重单核（纵向扩展，Scale-up）的 V 系列，与侧重多核（横向扩展，Scale-out）的 N 系列，完全可以视为大小核架构在数据中心市场的实践。

目前，比较有代表性的 V 系产品包括 64 核的 AWS Graviton3（推测 V1）和 72 核的 NVIDIA Grace CPU（V2），N 系产品除了 128 核的阿里云倚天 710（推测 N2），还在 DPU 中获得了较为广泛的应用。

2023 年 5 月中发布的 AmpereOne 采用 Ampere Computing 公司的自研（A1）核，从其最多 192 个核心来看，更接近 Neoverse N 系的风格。

英特尔在面向投资者的会议上也公布了类似的规划：



Emerald Rapids

预计 2023 年第四季度推出

Granite Rapids

预计 2024 年推出

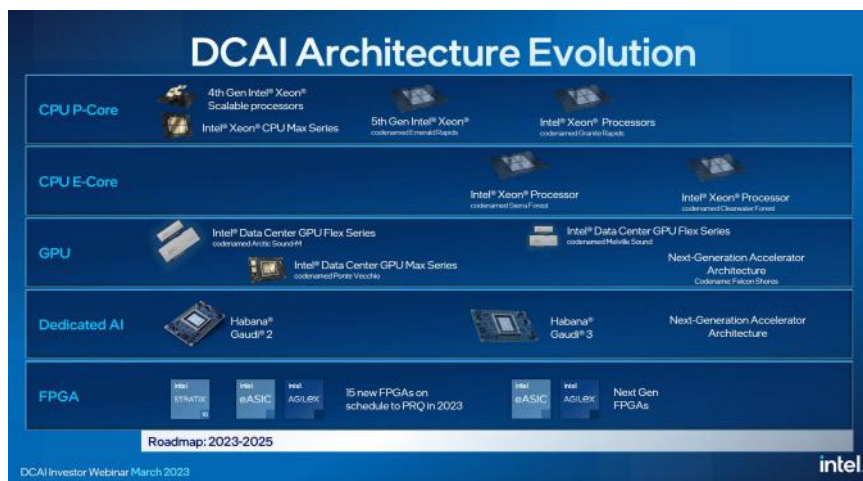
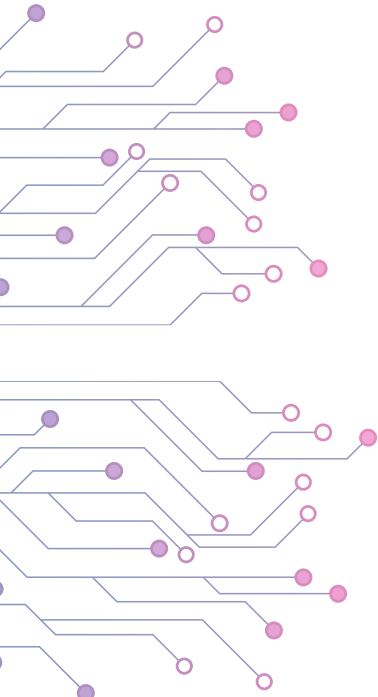
Sierra Forest

预计 2024 年上半年推出

- 定于 2023 年第四季度推出的第五代英特尔至强可扩展处理器（代号 Emerald Rapids），和预计 2024 年推出、代号 Granite Rapids 的更新一代产品，将延续目前的性能核（Performance-Core，P-Core）路线；
- 预计 2024 年上半年推出、代号 Sierra Forest 的 CPU，将是第一代能效核（Efficient-core，E-Core）至强处理器，具有 144 个核心。

第五代英特尔至强可扩展处理器与第四代共平台，易于迁移，而 Granite Rapids 和 Sierra Forest 都将采用 Intel 3 制程。

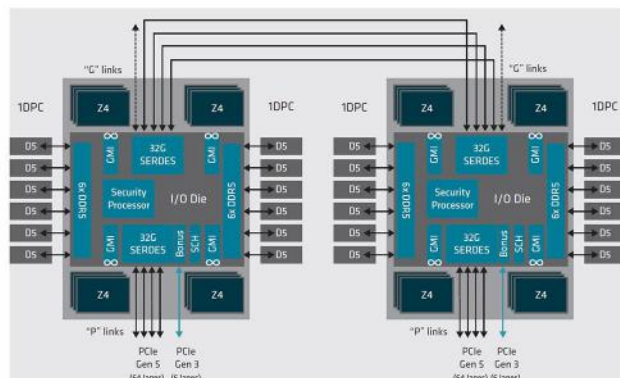
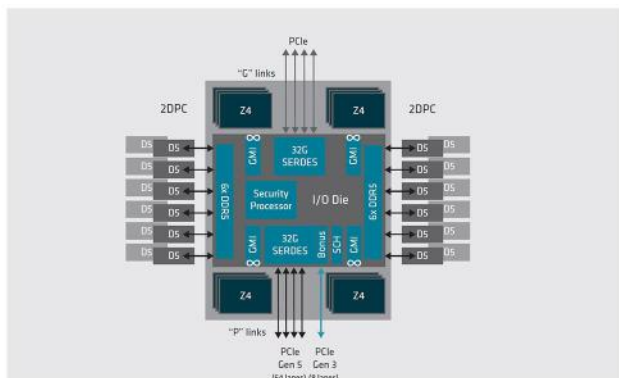
P-Core 与 E-Core 的组合已经在英特尔的客户端 CPU 上得到检验，两者之间一个很大的区别是有无超线程。E-Core 每核心只有一个线程且注重能效，适合追求更高（物理）核密度的云原生应用。



AMD 的策略大同小异。2022 年 11 月 AMD 发布代号 Genoa（热那亚）的第四代 EPYC 处理器，具有多达 96 个 5nm 的 Zen 4 核心；在 2023 年年中，AMD 将推出代号 Bergamo 的“云原生”处理器，据传有多达 128 个核心，通过缩小核心及缓存来提供更高的核心密度。

性能核与能效核这两条路线之间存在着（物理）核心数量的差异，但各自增加核心数则是共识。CPU 核心数量的持续增长要求更高的内存带宽，仅仅从 DDR4 升级到 DDR5 是不够的，AMD 第四代 EPYC 处理器（Genoa）已经把每 CPU 的 DDR 通道数量从 8 条扩充至 12 条，Ampere Computing 也有类似的规划。

100 多核的 CPU 已经超出了一些企业用户的实际需求，每 CPU 的 12 条内存通道，在双路配置下也给服务器主板设计提出了新的挑战。在多种因素作用下，单路服务器在数据中心市场的份额是否会迎来比较显著的增长？让我们拭目以待。



△ AMD 第四代 EPYC 处理器拥有 12 个 DDR5 内存通道，但单路（2DPC）和双路（1DPC）配置都不超过 24 个内存槽，比 8 内存通道 CPU 的双路配置（32 个内存槽）还要少。换言之，单 CPU 的内存通道数增加了，双路配置的内存槽数反而减少了



摩尔谢幕，Chiplet 当道

摩尔定律放缓

“摩尔定律已死”是近几年来半导体行业内不断被提起的话题，在提升晶体管密度的过程中，困难实在太多了，尤其是 EUV（Extreme UltraViolet，极紫外）光刻技术的量产曾遭遇多次延迟，大大拖慢了微缩工艺的发展。产业界、学术界在不断的碰壁、失败当中，难免发出这样的哀叹。

幸好半导体行业的增长动力不仅仅来自光刻技术的精进，封装技术的创新也提供了许多新的思路。譬如以 AMD EPYC 系列处理器为代表的“以小博大”，通过将较小的处理器核心进行组合，甚至将 I/O 单元分开制造，再封装为一体的方式，实现了不同工艺特性的解耦，并提升了良率，从而让核心数量的增长驶上了快车道。这种理念被称为：Chiplet（小芯片）。

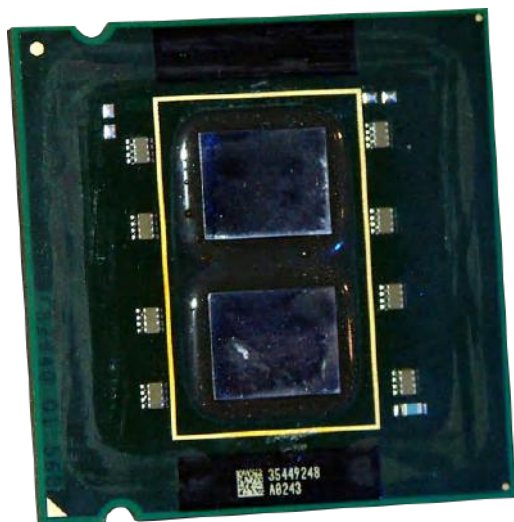
在数据中心 CPU 市场，AMD EPYC（霄龙）家族处理器的成功，使得 Chiplet 技术广为人知，也不可避免的产生了一些误解。然而，这种多个 die（芯粒、晶片）封装为一个整体的技术，并不是凭空出现的。

Chiplet 简史

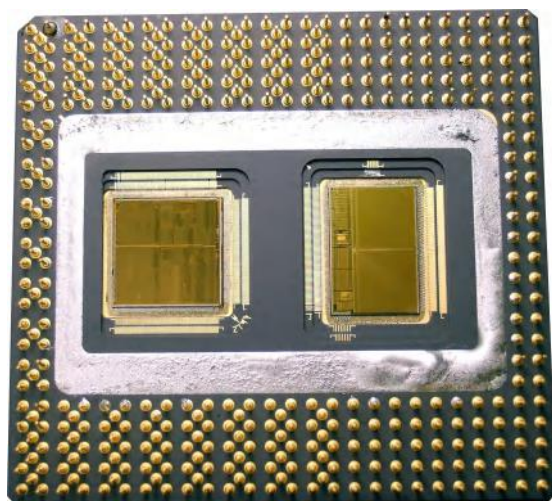
光刻技术之外的创新重新唤起了业界对半导体未来发展速度的期待，诸如 More Moore（深度摩尔）、More than Moore（超越摩尔）等等，当然，也包括材料等创新，所谓 Beyond CMOS（新器件）。

回到 Chiplet，“过来人”可能会认为：在一个封装基板上放置若干核心并不是什么新鲜事，譬如英特尔（Intel）在消费级的 Pentium D、Pentium Extreme Edition（EE）上就实现了“胶水双核”；再往前看，Pentium Pro 的处理器内核和 L2 Cache（缓存）也是两颗独立的裸晶封装在一起——这是 1995 年的事情。





△ 核心代号 Presler 的 Intel Pentium D 处理器



△ Intel Pentium Pro 处理器，封装左侧为核心，右侧为缓存

确实，从制造角度而言，Chiplet 算不上创新，MCM（Multi-Chip Module）、SiP（System in Package）已经存在多年了。先进封装是提升芯片规模的基础，而 Chiplet 则是一种设计理念。Chiplet 要做的是充分利用先进封装技术，实现芯片架构或系统架构的创新。创造 Chiplet 这个概念，其实是向以往单一追求晶体管微缩、追求晶体管规模的发展方式告别，更强调以合理的方式、合理的成本实现目标。

过去的 MCM 更像是一种权宜之计，当晶体管微缩能力进一步提升后，出于性能和成本的考虑，曾经分立的器件会再度被整合到一片裸晶之内，前面提到的 Pentium Pro、Pentium D 的形态，在之后十年并未复现。而现在的 Chiplet，则是一条明确的长期演进路线，多芯粒的组合将是常态。

Chiplet 之路不会反复的原因主要有：

1、高性能、高并发的需求使得数据中心、超算等领域对增加核心规模和数量的需求非常迫切，不论光刻工艺如何精进，顶级供应商都会倾向于将晶体管数量和裸晶面积堆砌到工程上难以负





高性能、高并发的需求使得数据中心、超算等领域对增加核心规模和数量的需求非常迫切，不论光刻工艺如何精进，顶级供应商都会倾向于将晶体管数量和裸晶面积堆砌到工程上难以负荷的程度。通过微缩减少裸晶面积、降低单位成本，并不是高性能产品主要的考虑方向。

荷的程度。通过微缩减少裸晶面积、降低单位成本，并不是高性能产品主要的考虑方向。

2、28nm 制造工艺之后，微缩已经无法降低单位晶体管的生产成本。另外，不同特点的器件在微缩中的收益也并不相同。譬如

- a) I/O 部分适用于较成熟的工艺；
- b) 运算器件可以明显受益于先进工艺；
- c) 存储器件介于上述二者之间，且主流存储器本质上是电容，即便使用先进工艺，也不能获得如逻辑器件那样显著的面积缩小效果。

而高性能处理器对存储带宽及容量、I/O 带宽及数量的要求也越来越高，SRAM、存储控制器、I/O 控制器及 PHY（物理层）所占用的晶体管数量、面积已经大到不可忽视的程度。

3、Chiplet 的芯粒可以应用到多款产品上，增加了产品开发的灵活性。譬如 AMD 的 CCD 和 IOD 可以按需组合，同代的消费级（Ryzen）和服务器 CPU（EPYC）可以使用相同的 CCD，但数量不同，并搭配不同规模的 IOD。随着业界对先进封装的应用越来越熟练，芯粒正在进一步细分，如 GPU、内存控制器、PHY 等单元都有独立出来的实例，一块芯片内封装十颗以上的芯粒将是常事。进一步的，IP 开发者可以不仅仅是向芯片设计者出售授权，而是可以将受欢迎的 IP 核“硬化”为芯粒，并将这些芯粒直接提供给封装环节。

4、芯粒的标准化集成也促进了标准化接口的产生。早期的 Chiplet 是芯片所有者的“家务事”，其使用自有接口、自有总线，捆绑特定晶圆厂、封装厂进行生产。但随着第三方 IP 的硬化和集成越来越多，芯粒之间 I/O 的标准化就成为必选项。

简而言之，芯粒的“通用化”和接口的“标准化”赋予 Chiplet 旺盛的生命力，Chiplet 不仅仅是顶级企业、顶级产品的专属，而会出现在广泛的产品当中。未来芯片的基板就如同过去的主板一般，将承载多种不同的芯粒。





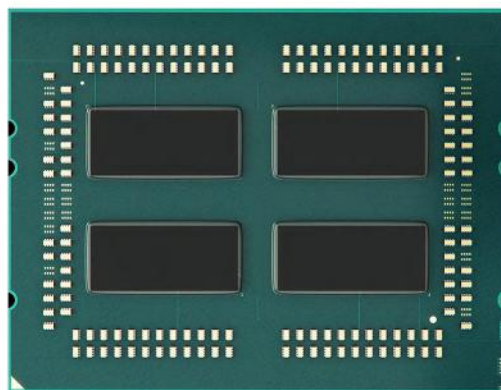
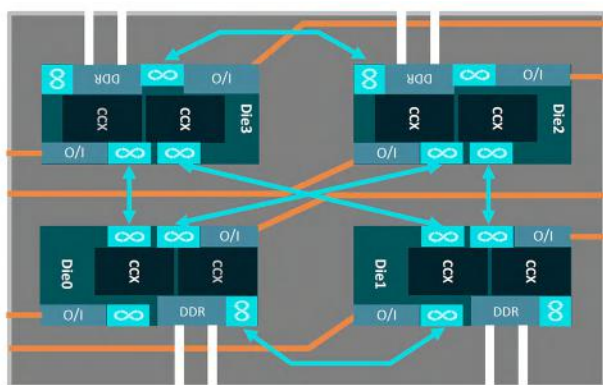
四等分：形似神不似

在《2021 中国云数据中心考察报告》的第二章“多元算力”篇，提到了代号 Naples（那不勒斯）的 AMD 第一代 EPYC 处理器，与代号 Sapphire Rapids（SPR）的第四代英特尔至强（Xeon）可扩展处理器，在四等分这个视角上的相似性。随着第四代英特尔至强处理器在 2023 年 1 月中旬正式发布，架构细节逐渐公开，下面简单比较一下异同。

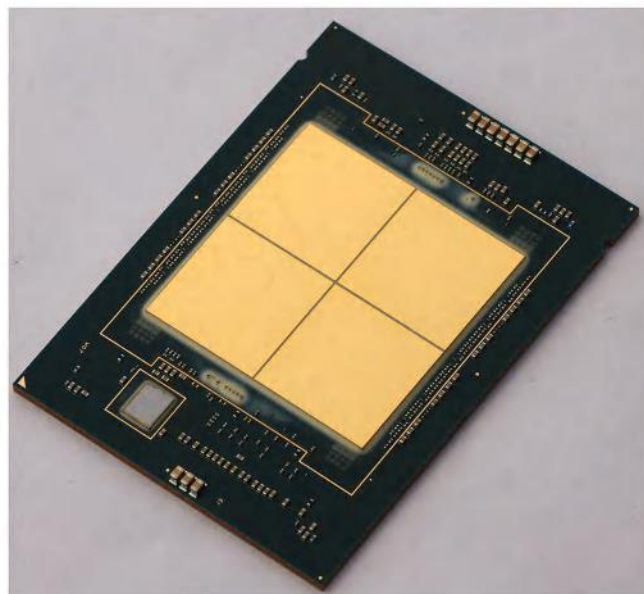
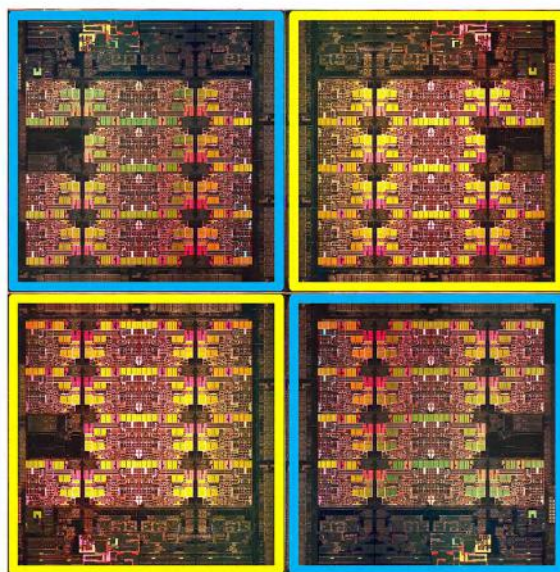
第一代 EPYC 处理器采用 14nm 制程，由 4 个 CCD（Core Complex Die，核心复合体）组成，CCD 的中间是 8 个 CPU 核心及其缓存（Cache），I/O 分布在外围，包括双通道 DDR 内存控制器、用于晶片间互联的 IFOP（Infinity Fabric On-Package）、PCIe 控制器或用于 CPU 之间互连的 IFIS（Infinity Fabric Inter-Socket）。

这 4 个 CCD 理论上是一样的，可以视为同一款（SKU）。在布局上，其中的半数要水平旋转 180°，以保证 4 个 CCD 上的 8 个 DDR 内存控制器“一致对外”，满足内存插槽物理布局的需要。代价是 PCIe 控制器或 IFIS 的走线不好布置，需要借助分层来避免交叉。

AMD 将上述架构命名为多芯片模块（Multi-Chip Module，MCM），同样由 4 个 die（晶片）对等拼接而成的第四代英特尔至强可扩展处理器就已经或主动或被动的归类为 Chiplet 了。这当然主要归因于历史的进程，但也有微小的“个体努力”造成的差异。



△ 第一代 EPYC 处理器用 1 种 die 满足了 4 die 组合的需求，代价是布线难度加大，各 die 也会空置一个 IFOP



△ 第四代英特尔至强可扩展处理器的 Chiplet 实现

第四代英特尔至强可扩展处理器采用 10nm 级别的 Intel 7 制程，分 MCC 和 XCC 两大构型，后者才是 Chiplet 版本：由 4 个 die 拼接而成，最多可达 56 ~ 60 核心。这 4 个 die 也排列为 2x2 的矩阵，但与第一代 EPYC 处理器的不同之处在于，XCC 构型的第四代至强可扩展处理器由 2 种互为镜像的晶片构成，在对角线上的 2 个是同一款（SKU）。

Chiplet 与芯片布局

在 CPU 的 Chiplet 实现上，AMD 和英特尔都和大家有“点”不一样。

从代号罗马（Rome）的第二代 EPYC 开始，AMD 将 DDR 内存控制器、Infinity Fabric 和 PCIe 控制器等 I/O 器件从 CCD 中“抽取”出来，集中到一个单独的 die 里，居中充当交换机的角色，即 IOD（I/O Die），这部分从制程提高到 7nm 中获益很小，仍然采用成熟的 14nm 制程；CCD 内部的 8 个核心加（L3）缓存所占面积由 56% 提高到 86%，可以从 7nm 制程中获得较大的收益。

IOD 和 CCD 分开制造，按需组合，“解耦”带来的优点有很多：

- **独立优化：**可以按照 I/O、运算、存储（SRAM）的不同要求分别选择成本适宜的制程，譬如代号 Genoa（热那亚）的第四代



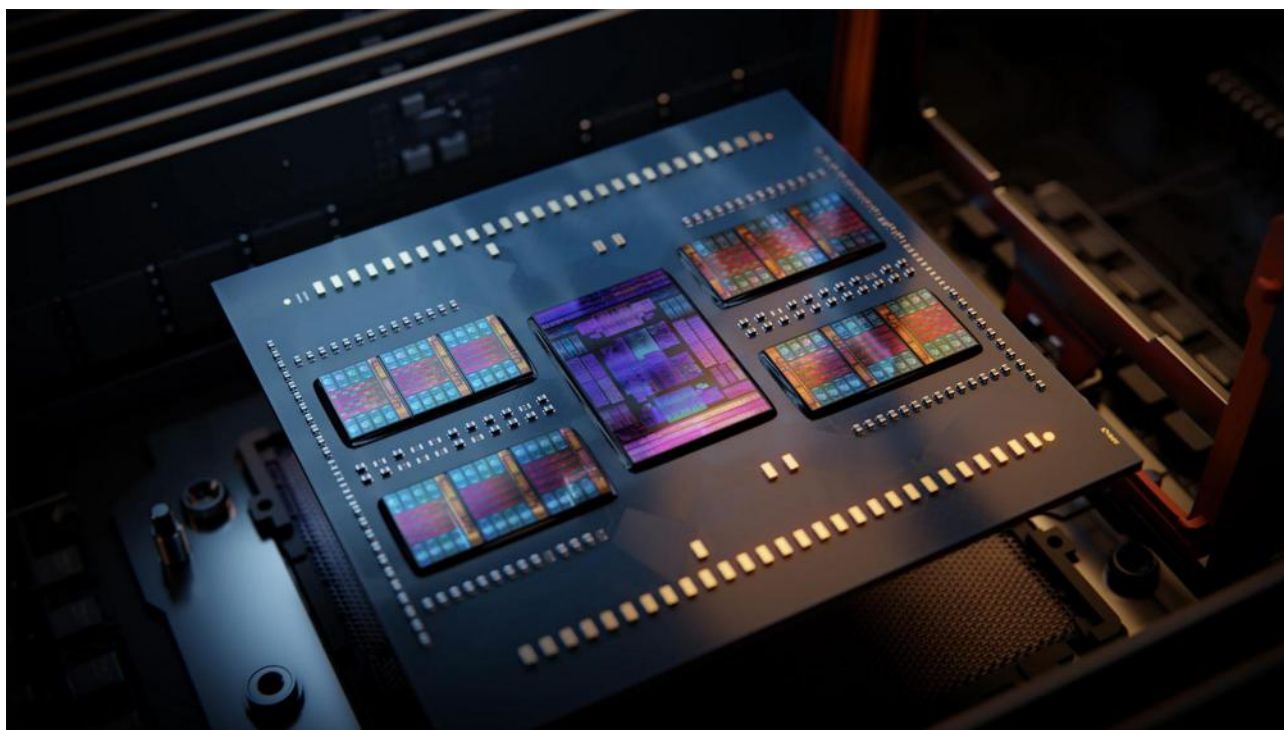


EPYC 处理器就分别“进化”为 5nm 制程的 CCD 搭配 6nm 制程的 IOD；

- **高度灵活：**1 个 IOD 可以搭配数量不等的 CCD，以提供不同的 CPU 核心数，譬如代号 Rome（罗马）的第二代 EPYC 处理器，最多支持 8 个 CCD，但也可以把数量减少到 6、4、2 个，总之能轻松自如的提供 8 ~ 64 个核心。

如果将这个 CCD 看作 8 核的 CPU，IOD 看作原来服务器中的北桥或 MCH（Memory Controller Hub），第二代 EPYC 就相当于一套微缩到封装里的八路服务器，用这种方法构建 64 核，难度比在单个 die 上提供 64 核要低多了，还有良率和灵活性上的优势。

扩大规模也更为容易：在 IOD 的支持下，通过增加 CCD 的数量，可以“简单粗暴”地获得更多的 CPU 核心，譬如第四代 EPYC 处理器就凭借 12 个 CCD 将核心数量扩展到 96 个。



△ AMD 第四代 EPYC 处理器，12 颗 CCD 环绕 1 颗 IOD



△ 代号 Genoa 的 AMD 第四代 EPYC 处理器

第二至四代 EPYC 以 IOD 为中心连接多个较小规模的 CCD，是比较典型的星形拓扑结构。这种架构的优势在于 IOD 及其成本，增加 PCIe、内存控制器的数量比较容易；劣势是任意核心与其他资源的距离太远，带宽和时延会受限。在 AMD 享有明显的制程优势（并体现在核数优势）的时候，EPYC 家族处理器即使单核性能略逊，多核性能依旧能相对优异。但随着英特尔的制造工艺改进，或者其他架构竞争者（如 Arm）的大核性能提升，AMD 的核数优势有被削弱的趋势，目前的多核性能优势恐难以保持。

在过去几年中，AMD 借助较小的 CCD 以较低成本实现了横向扩展（总核数提升），未来的可持续性尚待观察。目前其他几家的多核 CPU 在布局上普遍采取网格化的思路，实现核心、缓存、外部 I/O（包括内存、PCIe 等）的快速互联，减小任意核心与其他核心或 I/O 资源的访问距离，从而更有效地控制时延（latency）。

网格架构：Arm 与 Intel

作为 x86 阵营的带头大哥，英特尔从开启至强可扩展处理器系列至今，四代产品都基于网格（2D Mesh）架构。

致力于颠覆 x86 在服务器 CPU 市场霸主地位的 Arm 阵营，所采用的 Corelink 互连方案 CMN（Coherent Mesh Network，一致性网格网络），显然也是一种网格架构。

（2D）Mesh 是水平（X）和垂直（Y）方向的连线组成的二维交换矩阵，其中的一个个交叉点（Crosspoint，XP）用以连接 CPU/Cache、DDR/PCIe 控制器等设备。出于布线方便等考虑，内存控制器、PCIe 控制器、UPI/CCIX 等负责对外 I/O 的设备都布置在最外面一圈，其他交叉点留给 CPU 和缓存（Cache）等“核心资产”。

显然，网格的规模越大，交叉点就越多，可以布置的 CPU、缓存、I/O 资源也就随之增加。譬如：

- 至强可扩展处理器从第一代的 6×6 矩阵发展到第三代的 7×8 矩阵，核心数从 28 个扩展至 40 个，DDR 内存控制器和 PCIe 控制器的数量也有所增长；

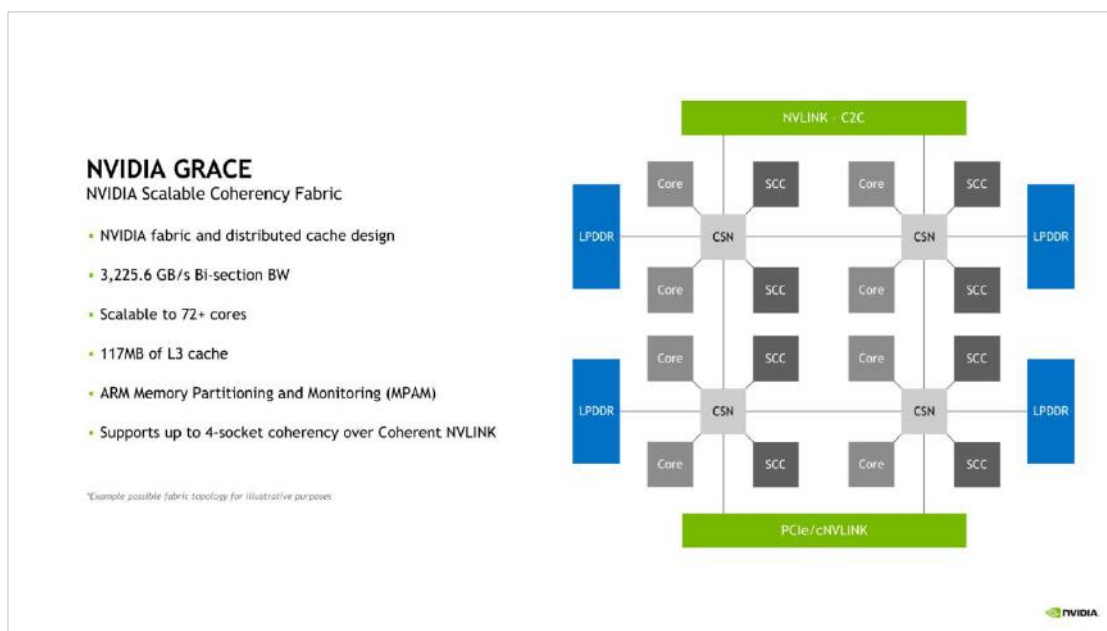


- Arm Neoverse 平台从 CMN-600 的 8×8 矩阵升级到 CMN-700 的 12×12 矩阵，支持的每 die 核心数从 64 个增长到 256 个，系统级缓存（System Level Cache, SLC）容量也从 128MB 提高到 512MB。

随着矩阵规模的扩大，居中的核心访问 I/O 资源的路径也会有所延长，但通过增加 I/O 资源数量并优化其分布及访问策略等手段，可以较好的抑制时延增长。

同样是网格架构，Arm 和英特尔在细节上还是有些值得注意的不同，主要体现在节点（交叉点）上。

CMN-700 每个交叉点上的设备从 CMN-600 的 2 个增加到 3 ~ 5 个，以英伟达（NVIDIA）基于 Arm Neoverse V2 的 Grace CPU 为例，每个节点通常会有 2 个 CPU 核心及对应的 2 片（slice）L3 缓存，在矩阵边上的节点还很可能会连接内存控制器、PCIe/NVLink 等 I/O 设备。



△ NVIDIA Grace 的 SCF 网络

注意：通过 Coherent NVLink，NVIDIA Grace 可支持多达四路 CPU 的一致性

英特尔至强可扩展处理器的每个（非 I/O）节点上只有 1 个 CPU 核心及其对应的 L3 Cache，考虑到每个 CPU 核心支持超线程（Hyper-Threading，HT），可以当作 2 个逻辑核心使用，在每个节点上论逻辑核心数量的话，和 Arm 倒是一样的。

Arm 新升：NVIDIA Grace 与 AmpereOne

Arm 在过去十年中一直期望能够在服务器市场获得一席之地。亚马逊、高通、华为等企业都推出了基于 Arm 指令集的服务器 CPU。随着亚马逊的 Graviton、Ampere Altra 等系列产品逐渐在市场站稳了脚跟，Arm 在服务器 CPU 市场渐入佳境。而且，随着异构计算的兴起，Arm 在高性能计算、AI/ML 算力基础设施中的影响力正在扩大——或许，随着 Neoverse V2 推出和英伟达加入战团，Arm 在服务器 CPU 领域有望从一个参与者变为领先者。



随着亚马逊的 Graviton、Ampere Altra 等系列产品逐渐在市场站稳了脚跟，Arm 在服务器 CPU 市场似乎渐入佳境。而且，随着异构计算的兴起，Arm 在高性能计算、AI/ML 算力基础设施中的影响力正在扩大——或许，随着 Neoverse V2 推出和英伟达加入战团，Arm 在服务器 CPU 领域有望从一个参与者变为领先者。

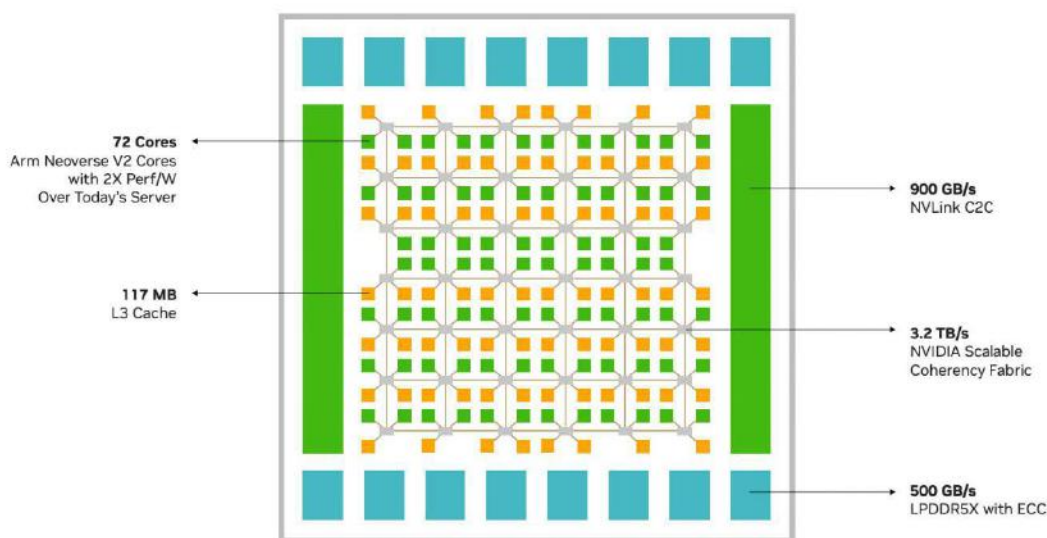
早在 2021 年，英伟达就对外介绍了基于 Arm Neoverse 架构的数据中心专属 CPU —— NVIDIA Grace，拥有 72 个核心。Grace CPU 超级芯片（Superchip）则由两个 Grace 芯片组成，它们之间通过 NVLink-C2C（Chip-2-Chip）连接在一起，可以在单插座内提供 144 个核心，以及 1TB LPDDR5X 内存。除了双 CPU 的组合，在 GTC2022 上，NVIDIA 还宣称 Grace 可以通过 NVLink-C2C 与 Hopper GPU 连接，组成 Grace Hopper 超级芯片。

NVIDIA Grace 是基于 Arm Neoverse V2 IP 的第一款重磅产品。目前还没看到 NVIDIA Grace 晶体管规模的相关资料，不妨先参照两位“前辈”的数据。据推测基于 Arm Neoverse V1 的 AWS Graviton 3 是 550 亿晶体管，对应 64 核、8 通道 DDR5 内存；据推测基于 Arm Neoverse N2 的阿里云倚天 710 是 600 亿晶体管，对应 128 核、8 通道 DDR5 内存、96 通道 PCIe 5.0。从 NVIDIA Grace Hopper 超级芯片的渲染图看，Grace 的芯片面积与 Hopper 近似，而已知后者为 800 亿晶体管，二者均基于台积电 N4 制程节点。据此推测 72 核的 Grace 芯片的晶体管规模大于 Graviton 3、倚天 710 是合理的，也与 Grace 基于 Neoverse V2（支持 Arm V9 指令集、SVE2）的情况相符。



+ + + + + +
+ + + + + +

Arm Neoverse V2 配套的互连方案是 CMN-700，在 NVIDIA Grace 这里称作 SCF（Scalable Coherency Fabric，可扩展一致性结构）。英伟达宣称 Grace 的网格支持超过 72 个 CPU 核心的扩展——实际上，在英伟达白皮书的配图中可以数出来 80 个 CPU 核心。每个核心 1MB L2 缓存，整个 CPU 有多达 117MB L3 缓存（合每个核心 1.625MB），明显高于其他同属“旗舰级”的 Arm 处理器。



△ NVIDIA Grace CPU 的网格布局

NVIDIA SCF 在芯片内的设备（如 CPU 核心、内存控制器、NVLink 等系统 I/O 控制器）之间提供 3.2 TB/s 的分段带宽。网格的节点称为 CSN（Cache Switch Nodes，缓存交换节点），每个 CSN 通常要连接 2 个核心及 2 个 SCC（SCF Cache partitions，SCF 缓存分区）。但从示意图来看，位于网格角落的 4 个 CSN 连接的是 2 个核心和 1 个 SCC，而位于中部两侧 4 个 CSN 连接的是 1 个核心和 2 个 SCC。整体而言，Grace 的网格内应该有 80 个核心和 76 个 SCC，其中 8 个核心应该是出于良率等因素而屏蔽。而网格外围“缺失”的 4 个核心和 8 个 SCC 对应的位置被用于连接 NVLink、NVLink-C2C、PCIe、LPDDR5X 内存控制器等。

NVIDIA Grace 支持 Arm 的许多管理特性，譬如服务器基础系统架构（SBSA）、服务器基础启动要求（SBBR）、内存分区与监控（MPAM）、性能监控单元（PMU）等等。通过 Arm 的内存分区

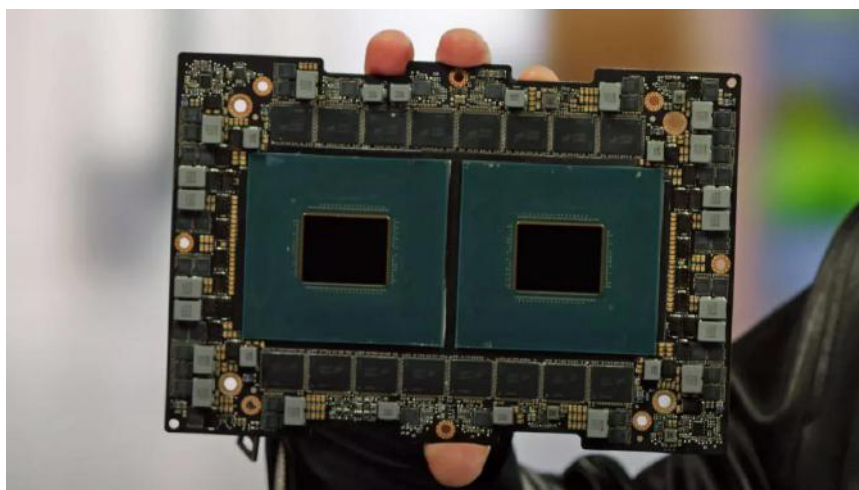


和监控 (Memory Partitioning and Monitoring, MPAM) 功能, 可以解决 CPU 访问缓存过程中因为共享资源的竞争导致的性能下降问题。高优先级的任务可以优先占用 L3 缓存, 或者根据虚拟机预先划分资源, 实现业务之间的性能隔离。



△ NVIDIA Grace CPU 超级芯片

NVIDIA Grace 作为已知的最新最强版本 Arm 架构核心 (Neoverse V2) 的实例, 再加上其必将获得自家 GPGPU 的深厚实力加持, 业界从一开始就给予了高度关注, 期待其在超算、机器学习等领域的表现。在 GTC2023 上, 人们终于看到了 Grace 的实物, 其实际市场表现仍需要一段时间的等待。



△ GTC2023 演讲中展示的 Grace 超级芯片实物

	NVIDIA Grace CPU Superchip architecture features
Core architecture	Neoverse V2 Cores: Armv9 with 4x128b SVE2
Core count	144
Cache	L1: 64 KB I-cache + 64 KB D-cache per core L2: 1 MB per core L3: 234 MB per superchip
Memory technology	LPDDR5X with ECC, co-packaged
Raw memory BW	Up to 1 TB/s
Memory size	Up to 960 GB
FP64 peak	7.1 TFLOPS
PCI express	8x PCIe Gen 5 x16 interfaces; option to bifurcate Total 1 TB/s PCIe bandwidth. Additional low-speed PCIe connectivity for management.
Power	500 W TDP with memory, 12 V supply

Table 1. NVIDIA Grace CPU Superchip specifications

作为 Arm Neoverse V1 的“后浪”，Neoverse V2 的升级可以说是全方位的，包括基于 Armv9-A 指令集、更高的性能和微架构能效，加上更多的核心数和更大的 L3 缓存，NVIDIA Grace CPU 在 Arm 版图中高于 Graviton3，是可以预期的。

英伟达毕竟是 Arm 服务器 CPU 领域的新手，在这方面资深的 Ampere Computing（安晟培半导体）经过多代产品积累之后，在 2023 年 5 月中正式发布拥有 192 个单线程自研核的 AmpereOne 系列处理器，这个核心数量也创下了（主流）服务器 CPU 的新纪录。

AmpereOne 采用台积电 5nm 制程，提供的 Ampere（A1）核数量覆盖 136 ~ 192 个的区间，每个核心配备 2MB L2 缓存，这一点与 Neoverse V2（的上限）相当，达到 Ampere Altra 和 Altra Max 系列的两倍。系统级缓存（SLC）为 64MB，分别是 Altra 和 Altra Max 系列的 2 ~ 4 倍，与 AWS Graviton3 持平。

其他如 8 通道 DDR5 内存和 128 个 PCIe 5.0 通道，都属于新一代





服务器 CPU 的正常水平。

由于每个核心相对不那么复杂，又采用了比较先进的制程，AmpereOne 系列的使用功耗在 200 ~ 350 瓦（W）之间，平均每核心不到 2 瓦。NVIDIA Grace CPU 的功耗也不算高，超级芯片加上内存的 TDP “才” 500 瓦，即单个（72 核的）Grace CPU 在 250 瓦以内。

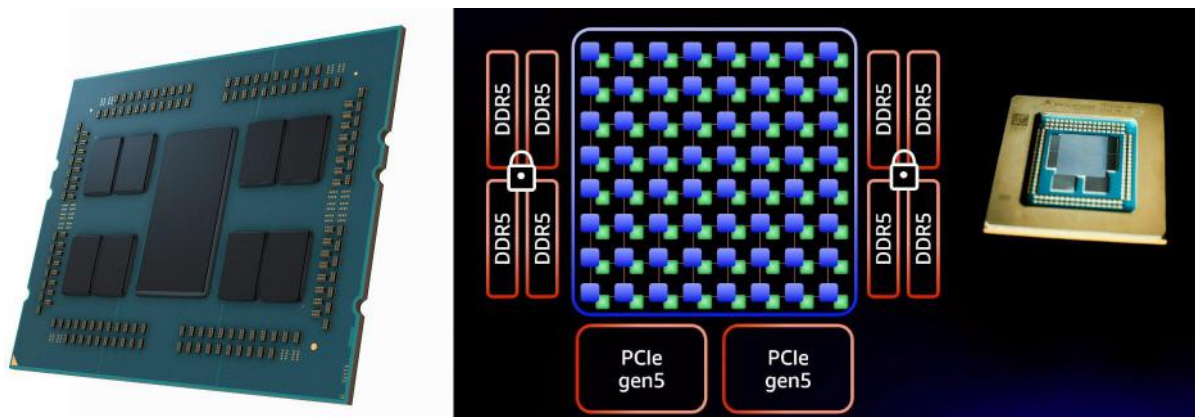
	Ampere® Altra® Family					AmpereOne™ Family				
Cores	32	64	80	96	128	136	144	160	176	192
L2 Private Cache	1MB/Core					2MB/Core				
Memory Config	8 Channel DDR4					8 Channel DDR5				
IO Config	128 Lanes PCIe Gen4					128 Lanes PCIe Gen5				
Usage Power	40-180W					200-350W				
Cloud Native Features	• Single Threaded Cores • Consistent Frequency • Large Private L2 Cache • 2x128b Vector Units, FP16, Int16, Int8 • Interrupt & IO Virtualization					• Bfloat16 • Performance Consistency • Mesh Congestion Management • Memory & SLC QoS Enforcement • Nested Virtualization • Scalable Management • Fine Grained Power Management • Advanced Droop Detection • Process Aging Monitors • Security • Secure Virtualization • Single-Key Memory Encryption • Memory Tagging				

尽管从核心微架构到外部 I/O 都获得了全面的升级，但 AmpereOne 并没有取代 Altra 和 Altra Max 系列的任务，Altra Max 系列继续提供 128 核与 96 核，Altra 系列覆盖 80 核及以下的的需求。我们认为，这种布局也与网络架构的特性有关，我们接下来讨论这个话题。

网络架构的两类 Chiplet

AmpereOne 毕竟有多达 192 个核心和 384MB L2 缓存，采用渐趋流行的 Chiplet 技术并不出人意料。目前比较普遍的推测是做法与 AWS Graviton3 类似，即 CPU 及缓存单独一个 die，DDR 控制器的 die 在其两侧，PCIe 控制器的 die 在其下方。

把 CPU 核心及缓存，与负责外部 I/O 的控制器，分离在不同的 die 上，是服务器 CPU 实现 Chiplet 的主流做法。



△ IOD 居中的 AMD 第二代 EPYC 处理器，与核心 die 居中的 AWS Graviton3 处理器

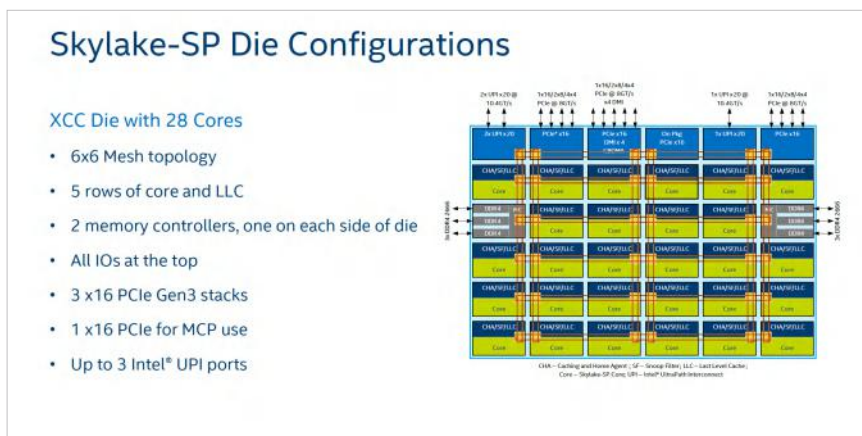
前面已经提到，AMD EPYC 家族处理器采取星形拓扑，把 I/O 部分集中放在 1 个 IOD 上，CPU 核心及缓存（CCD）环绕四周的设计。网格架构的特性决定了 CPU 核心及缓存必须在中间，I/O 部分分散在外围，所以拆分开时就是一个相反的布局。

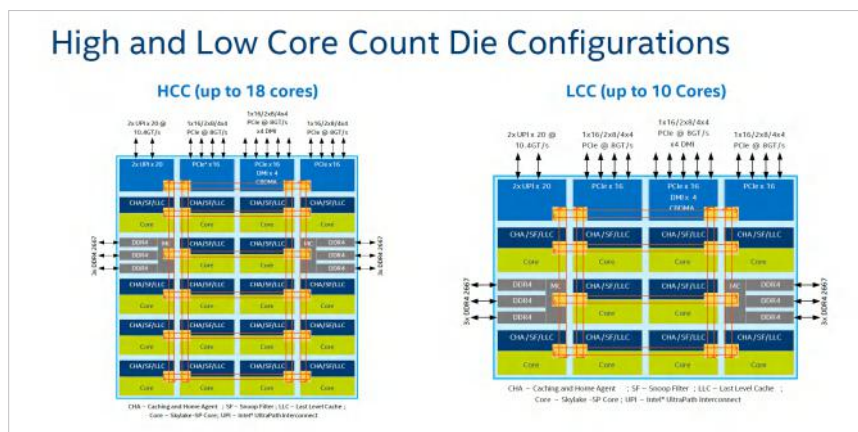
共同点是中间的 die 大，四周的 die 小。

与 EPYC 家族的架构比，网格架构的整体性比较强，天生的单体式（Monolithic）结构，不太利于拆分。

网格架构必须考虑交叉点（节点）的利用率问题，如果有太多的交叉点空置，会造成很大的资源浪费，不如缩小网格的规模。

以初代英特尔至强可扩展处理器为例，为了满足从 4 ~ 28 个的核





数（Core Count, CC）变化范围，提供了 3 种不同构型的 die（die chop），分别是：

- 6×6 的 XCC (eXtreme CC, 最多核 or 极多核), 可支持到 28 个核心;
- 6×4 的 HCC (High CC, 高核数), 可支持到 18 个核心;
- 4×4 的 LCC (Low CC, 低核数), 可支持到 10 个核心。

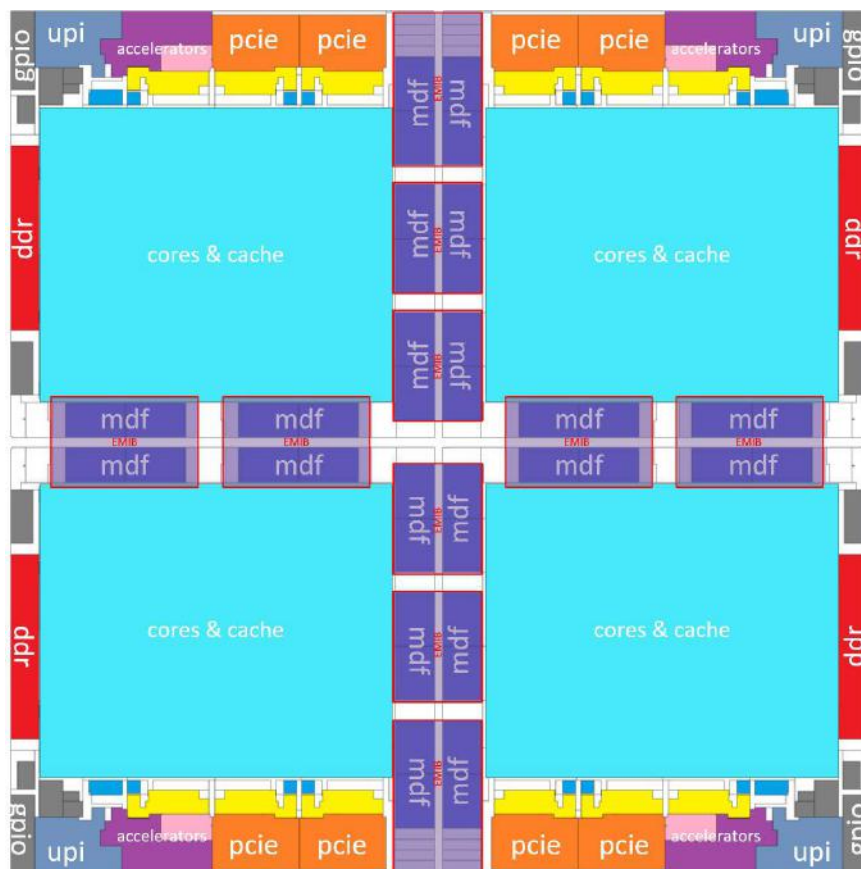
从这个角度来看, AmpereOne 不支持 128 核及以下也是合理的, 除非增加 die 的构型, 而这又离不开公司规模和出货量的支持——量的问题还得量来解决。

第四代英特尔至强可扩展处理器就提供了 2 种构型的 die, 其中, MCC (Medium CC, 中等核数) 主要满足 32 核及以下的需求, 这个核数要求比代号 Ice Lake 的第三代英特尔至强可扩展处理器的 40 核还要低, 所以网格的规模也比后者的 7×8 还少 1 列, 为 7×7, 在布局上最多可以安置 34 个核心及其缓存。

36 ~ 60 个核心的需求就必须由 XCC 来满足了, 它就是前面提到过的 Chiplet 版本, 把网格架构从中间切成了 4 等分, 可谓独树一帜。

XCC 版的第四代英特尔至强可扩展处理器由 2 种互为镜像的 die 拼成 2×2 的（大）矩阵, 所以这个整体高度对称——上下、左右都对称, 前三代产品和同代的 MCC 版都没有如此对称。

+ + + + + +
+ + + + + +



英特尔认为（XCC 版）的第四代英特尔至强可扩展处理器 4 个 die 拼接的效果是一个准单体式（quasi-monolithic）的 die。单体式不难理解，常见的网格架构就是如此，第四代英特尔至强可扩展处理器外圈的左右有 DDR 内存控制器，上下是 PCIe 控制器和集成的加速器（DSA/QAT/DLB/IAA），UPI 位于四角，也是典型的网格架构布局。

单体式前面的“准”是怎么达成的呢？就要看网格结构内部的“缝合”技术了。

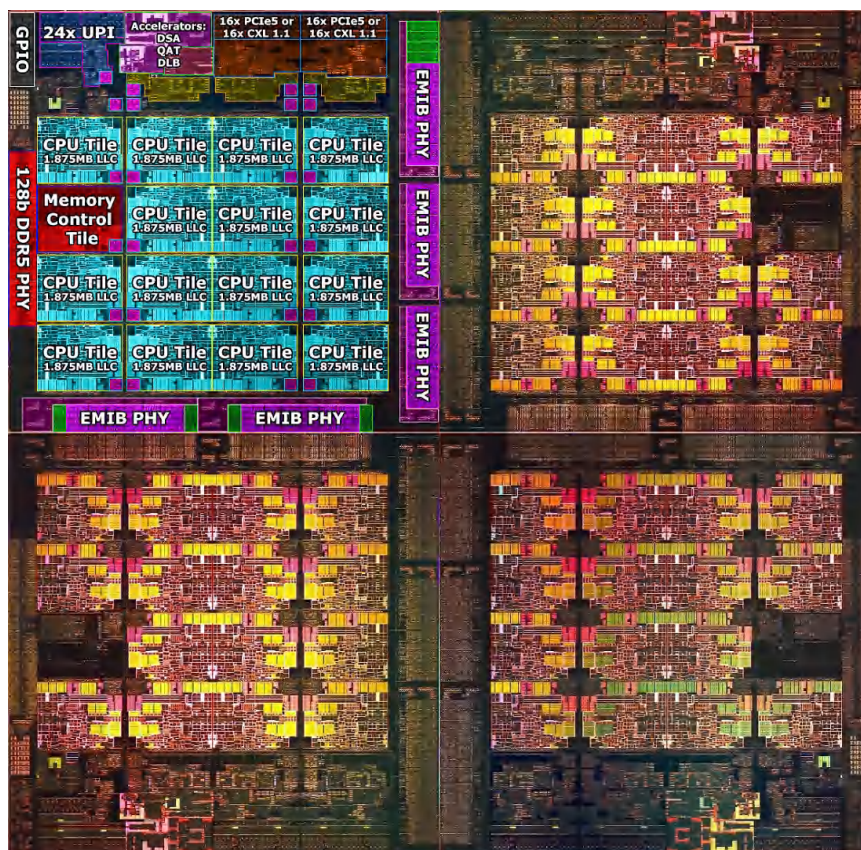
+ + + + + +
 + + + + + +

EMIB 及其带宽估算

如果没有采用 Chiplet 技术, XCC 版本的第四代英特尔至强可扩展处理器应该是一个 10×8 的网格架构, 最多 60 个核心, 留下 20 个(节点) 给 I/O。

如果直接把这个单体式的 die 四等分, 那每一部分就应该是一个 5×4 的小网格。

但事实是这 4 个 die 要连为一体, 就要为它们增加一行一列的连接点, 其中多出来的一行有 4 个, 一列有 6 个。4 个 die 对接到一起, 就用 20 个交叉点形成了 10 个 EMIB 的“桥”。



△ 第四代英特尔至强可扩展处理器的 EMIB 连接

EMIB（Embedded Multi-die Interconnect Bridge，嵌入式多芯片互连桥接）是英特尔用于实现 2.5D 封装的技术。第四代英特尔至强可扩展处理器内部封装了 4 个 XCC 的 die，每个 die 拥有 14 条 EMIB 互联，其中 4 条用于对外连接 HBM2e 内存，10 条（6 横 4 纵）用于相邻 XCC Tile 之间的互联。目前英特尔尚未公布 die 层面 EMIB 互联的具体带宽，我们仅能从工艺角度获知 EMIB 互联总线每对触点可以提供 5.4Gb/s 以上的带宽（Pin Speed），凸块间距为 55 μ m（微米），die 之间的距离为 100 μ m。

我们可以通过间接的方式进行估算。每 die 的 4 条 EMIB 对应 16GB 8-Hi HBM2e。HBM2e 每个引脚的数据传输率为 3.2Gb/s，每堆栈（Stack）为 1024bit 位宽，总带宽为 400GB/s 量级。因此，每条连接 HBM2e 的 EMIB PHY 至少可以提供约 100GB/s 的带宽。将每堆栈 HBM2e 的 1024bit 位宽均摊到 4 条 EMIB，则为每条至少 256bit。将 EMIB 每 pin 5.4Gb/s 的带宽代入，则每条 EMIB 的理论带宽起码可以达到 173GB/s。

将上述估算套回 XCC 的 die 间互联，则可知第四代英特尔至强可扩展处理器每个 XCC 的互联带宽约为 1 ~ 2TB/s 量级（1TB ~ 1.7TB/s），相邻两个 XCC 之间的互联为 600GB/s ~ 1TB/s（左右向 6 组 PHY）或 400GB/s ~ 691GB/s（上下向 4 组 PHY）。





CHAPTER3

算存互连

Chiplet 与 CXL

2023 新型算力中心调研报告



距离 CPU 核心最近的存储器，非基于 SRAM 的各级 Cache（缓存）莫属。不过，既然都分级了，那还是有远近之分。在现代 CPU 中，L1 和 L2 Cache 已经属于核心的一部分，需要为占地面积发愁的，主要是 L3 Cache。

算存互连：Chiplet 与 CXL

“东数西存”是“东数西算”的基础、前奏，还是子集？这牵涉到数据、存储与计算之间的关系。

数据在人口密集的东部产生，在地广人稀的西部存储，主要的难点是如何较低成本的完成数据传输。

计算需要频繁的访问数据，在跨地域的情况下，网络的带宽和时延就成为难以逾越的障碍。

与数据的传输和计算相比，存储不算耗能，但很占地。核心区域永远是稀缺资源，就像核心城市的核心地段不会用来建设超大规模数据中心，CPU 的核心区能留给存储器的硅片面积也是相当有限。

“东数西算”并非一日之功，超大规模数据中心与核心城市也是渐行渐远，而且不是越远越好。同理，围绕 CPU 早已构筑了一套分层的存储体系，虽然从 Cache 到内存都是易失性的存储器（Memory），但往往越是那些处于中间状态的数据，对访问时延的要求越高，也就需要离核心更近——如果真是需要长期保存的数据，距离远一些反倒无妨，访问频率很低的还可以“西存”嘛。

距离 CPU 核心最近的存储器，非基于 SRAM 的各级 Cache（缓存）莫属。不过，既然都分级了，那还是有远近之分。在现代 CPU 中，L1 和 L2 Cache 已经属于核心的一部分，需要为占地面积发愁的，主要是 L3 Cache。

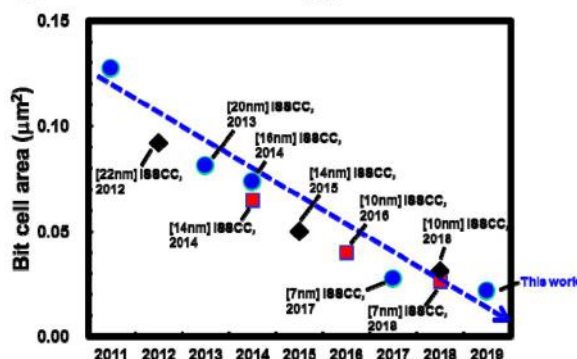
SRAM 的面积律

在 IEDM 2019 上，台积电展示了其引入 EUV 的 5nm 制程成果。当时业界便留意到一个问题：芯片的逻辑密度提高了 1.84 倍，而 SRAM 密度仅提高了 1.35 倍。在 ISSCC2020 中，关于 5nm SRAM 的论文还展示了 2011 ~ 2019 年 SRAM 面积的演进过程。在下图中可以很明显看出：

2017 年之前，SRAM 的面积缩减基本上与制程改进同步；

SRAM Bit Cell Area Scaling Trend

- Report $0.021\mu\text{m}^2$ for high density SRAM bit cell with leading edge 5nm technology



© 2020 IEEE
International Solid-State Circuits Conference

15.1: A 5m 135Mb SRAM in EUV and High-Mobility-Channel FinFET Technology with Metal Coupling and Charge-Sharing Write-Assist Circuitry Schemes for High-Density and Low-VMIN Applications

10 of 40

+ + + + +
+ + + + +

之后，SRAM 面积的缩减近乎停滞，即使应用了 EUV 技术，改善也不明显。

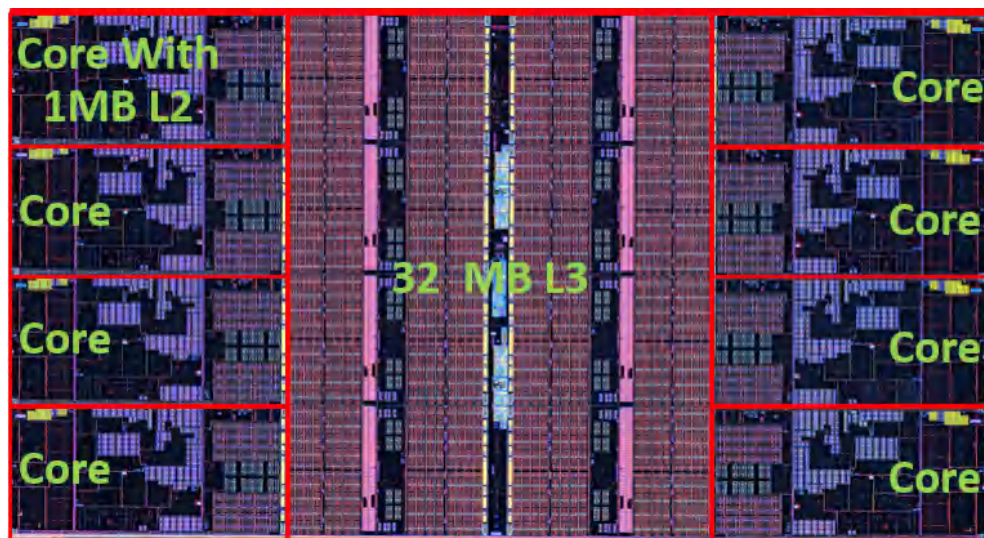
现在是 2023 年，制造工艺正在向 3nm 迈进。台积电公布其 N3 制程的 SRAM 单元面积为 0.0199 平方微米，相比 N5 制程的面积为 0.021 平方微米，只缩小了 5%。更要命的是，由于良率和成本问题，预计 N3 并不是台积电的主力工艺，客户们更关注第二代 3nm 工艺 N3E。而 N3E 的 SRAM 单元面积为 0.021 平方微米，和 N5 工艺完全相同。至于成本方面，据传 N3 一片晶圆是 2 万美元，N5 报价是 1.6 万美元，意味着 N3 的 SRAM 比 N5 贵 25%。

作为参考，Intel 7 制程（10nm）的 SRAM 面积为 0.0312 平方微米，Intel 4 制程（7nm）的 SRAM 面积为 0.024 平方毫米，和台积电的 N5、N3E 工艺差不多。

半导体制造商们的报价是商业机密，但 SRAM 越来越贵，密度也难再提高，终究是事实。于是，将 SRAM 单独制造再次变为合理，且可以配合先进封装实现高带宽、低时延。

向上堆叠，翻越内存墙

积极引入新制程生产 CCD 的 AMD 对 SRAM 成本的感受显然比较深刻，在基于台积电 5nm 制程的 Zen 4 架构 CCD 中，L2、L3 Cache 占用的面积已经达到整体的约一半比例。



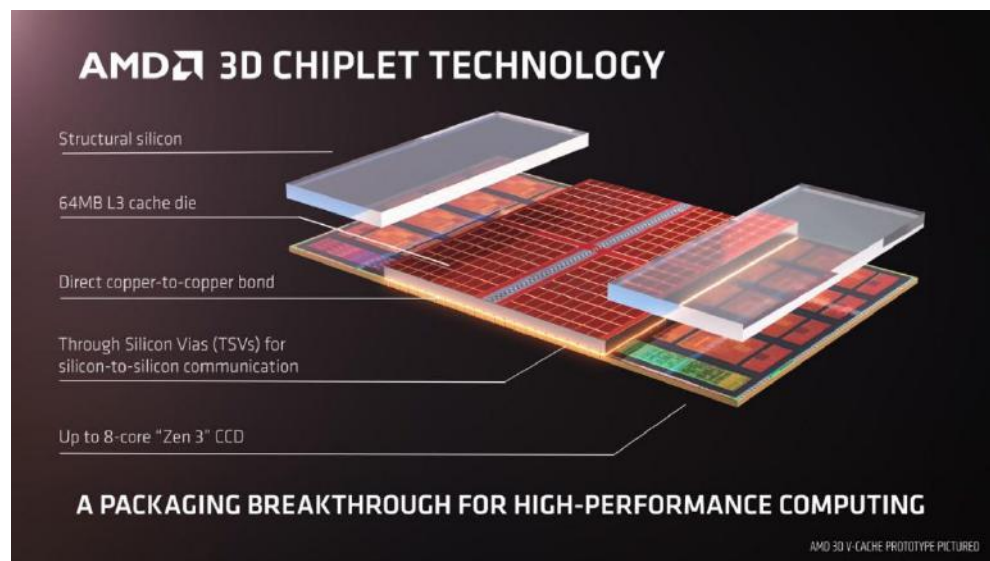
△ Zen4 CCD 的布局，请感受一下 L3 Cache 的面积

AMD 当前架构面临内存性能落后的问题，其原因包括核心数量较多导致的平均每核心的内存带宽偏小、核心与内存的“距离”较远导致延迟偏大、跨 CCD 的带宽过小等。这就促使 AMD 需要用较大规模的 L3 Cache 来弥补访问内存的劣势。而从 Zen 2 到 Zen 4 架构，AMD 每个 CCD 的 L3 Cache 都为 32MB，并没有“与时俱进”。为了解决 SRAM 规模拖后腿的问题，AMD 决定将 SRAM 扩容的机会独立于 CPU 之外。

AMD 在代号 Milan-X 的 EPYC 7003X 系列处理器上应用了第一代 3D V-Cache 技术。这些处理器采用 Zen 3 架构核心，每片 Cache（L3 Cache Die，简称 L3D）为 64MB 容量，面积约 41mm²，采用 7nm 工艺制造——回顾 ISSCC2020 的论文，7nm 恰恰是 SRAM 的微缩之路遇挫的拐点。

缓存芯片通过混合键合、TSV（Through Silicon Vias，硅通孔）工艺与 CCD（背面）垂直连接，该单元包含 4 个组成部分：最下层的 CCD、上层中间部分 L3D，以及上层两侧的支撑结构——采用硅材质，将整组结构在垂直方向找平，并将下方 CCX（Core Complex，核心复合体）部分的热量传导到顶盖。

AMD 在 Zen 3 架构核心设计之初就备了这一手，预留了必要的逻辑电路以及 TSV 电路，相关部分大约使 CCD 增加了 4% 的面积。L3D 堆叠的位置正好位于 CCD 的 L2/L3 Cache 区域上方，这一方



△ 3D V-Cache 结构示意图

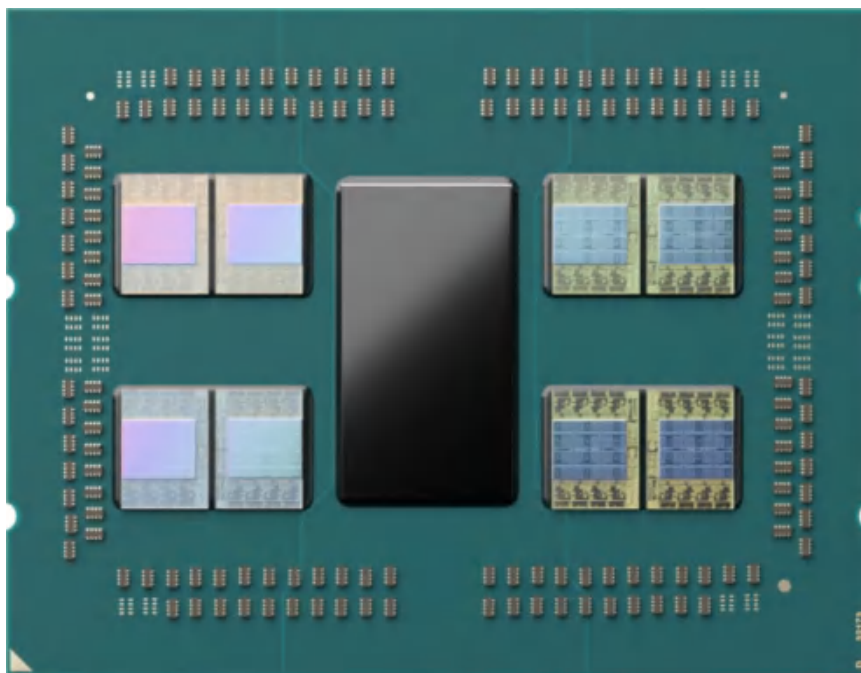
面匹配了双向环形总线的 CCD 内的 Cache 居中、CPU 核心分居两侧的布局，另一方面是考虑到（L3）Cache 的功率密度相对低于 CPU 核心，有利于控制整个 Cache 区域的发热量。

Zen 3 的 L3 Cache 为 8 个切片（Slice），每片 4MB；L3D 也设计为 8 个切片，每片 8MB。两组 Cache 的每个切片之间是 1024 个 TSV 连接，总共 8192 个连接。AMD 宣称这外加的 L3 Cache 只增加 4 个周期的时延。

随着 Zen 4 架构处理器进入市场，第二代 3D V-Cache 也粉墨登场，其带宽从上一代的 2TB/s 提升到 2.5TB/s，容量依旧为 64MB，制程依旧为 7nm，但面积缩减为 36mm²。缩减的面积主要是来自 TSV 部分，AMD 宣称基于上一代积累的经验和改进，在 TSV 最小间距没有缩小的情况下，相关区域的面积缩小了 50%。代号 Genoa-X 的 EPYC 系列产品预计在 2023 年中发布。

SRAM 容量增加可以大幅提高 Cache 命中率，减少内存延迟对性能的拖累。AMD 3D V-Cache 以比较合理的成本，实现了 Cache 容量的巨大提升（在 CCD 内 L3 Cache 基础上增加 2 倍），对性能的改进也确实是相当明显。代价方面，3D V-Cache 限制了处理器整体功耗和核心频率的提升，在丰富了产品矩阵的同时，用户需要根据自己的实际应用特点进行抉择。

那么，堆叠 SRAM 会是 Chiplet 大潮中的主流吗？



△ 应用 3D V-Cache 的 AMD EPYC 7003X 处理器



对于数据中心，核数是硬指标。表面上，目前 3D V-Cache 很适合与规模较小的 CCD 匹配，毕竟一片 L3D 只有几十平方毫米的大小。但其他高性能处理器的内核尺寸比 CCD 大得多，在垂直方向堆叠 SRAM 似乎不太匹配。

说到这里，其实是为了提出一个外部 SRAM 必须考虑的问题：更好的外形兼容性。堆叠于处理器顶部是兼容性最差的形态，堆叠于侧面的性能会有所限制，堆叠于底部则需要 3D 封装的进一步普及。对于第三种情况，使用硅基础层的门槛还是比较高的，可以看作是 Chiplet 的一个重大阶段。以目前 AMD 通过 IC 载板布线水平封装 CCD 和 IOD 的模式，将 SRAM 置于 CCD 底部是不可行的。至于未来 Zen 5、Zen 6 的组织架构何时出现重大变更还暂时未知。

对于数据中心，核数是硬指标。表面上，目前 3D V-Cache 很适合与规模较小的 CCD 匹配，毕竟一片 L3D 只有几十平方毫米的大小。但其他高性能处理器的内核尺寸比 CCD 大得多，在垂直方向堆叠 SRAM 似乎不太匹配。但实际上，这个是处理器内部总线的特征决定的问题：垂直堆叠 SRAM，不论其角色是 L2 还是 L3 Cache，都更适合 Cache 集中布置的环形总线架构。

对于面积更大的处理器，怎么突破 SRAM 的成本约束呢？

不但要找 SRAM 的（廉价）替代品，还要解决“放在哪儿”的问题。



回首 eDRAM 时光

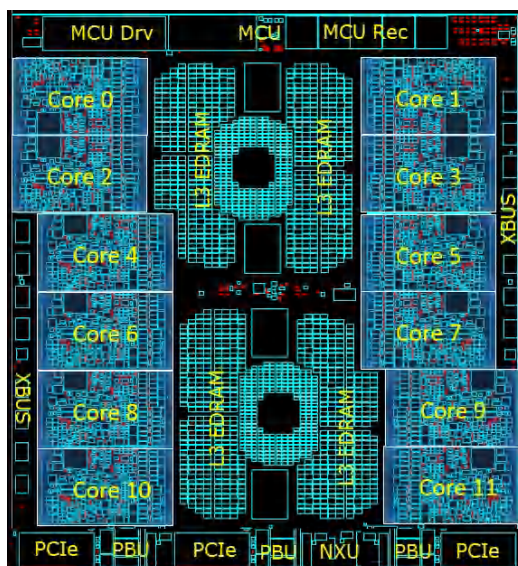
缓存容量的问题，本质上是弥补内存的性能落差。SRAM 快但是贵，DRAM 便宜但是慢。如果 SRAM 已经很难更快（频率、容量被限制），且越来越贵，那么，为什么不把增加的成本用在 DRAM 上呢？能不能找到更贵但更快的 DRAM？答案是肯定的。因此，最务实的思路就是提升内存性能，以及拉近内存与核心的距离。

提升 DRAM 性能的一种比较著名的尝试是 eDRAM（embedded DRAM，嵌入式 DRAM）。由于每单元 SRAM 需要由 4 或 6 个晶体管构成，其面积必然偏大，密度不如 DRAM，成本也比 eDRAM 更高一些。

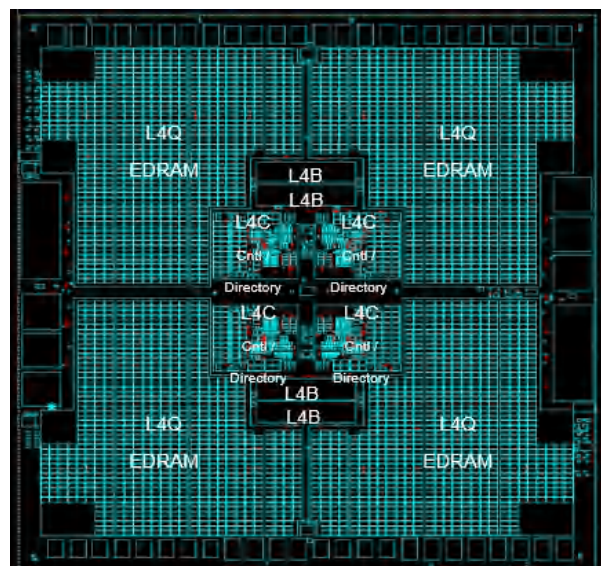
IBM 是 eDRAM 的积极应用者，宣称 eDRAM 的每 Mb 面积约为 SRAM 的三分之一，并从 2004 年的 PowerPC 440 就开始内嵌 eDRAM 作为 L3 Cache 使用。之后的 Power7 到 Power9，eDRAM 都被用作 L3 Cache 使用，于是“只有” 12/24 核的 Power9 处理器，L3 Cache 容量已经高达 120MB。

这种爱好蔓延到了 IBM Z15 这样的主机处理器。2019 年发布的 Z System 大型机使用的中央处理器（Central Processor, CP Chip）有 12 个核心，面积 696mm^2 ，其 L2、L3 Cache 均由 eDRAM 构成，其中 L2 Cache 为 $(4+4) \times 12 = 96\text{MB}$ ，L3 Cache 为 256MB。然后，Z15 还可以通过系统控制器（System Controller, SC Chip）提供 960MB L4 Cache，SC 的面积也是 696mm^2 。上一代的 Z14 也是类似的架构，L3 和 L4 Cache 分别为 128MB 和 672MB。两代芯片均采用 14nm SOI 制程。

格芯和 IBM 宣称基于 14nm 制程的 eDRAM 每单元面积为 0.0174 平方微米，比 5nm 的 SRAM 还要小。当然，任何技术优势在竞争压力面前都会被压榨到极限，eDRAM 的单位成本虽低，也架不住堆量。因此，IBM 用 eDRAM 作为 Cache 的实际代价其实也是很大的：大家可以从图片中看到 L3、L4 eDRAM 在 Z15 的 CP 和 SC 中占用的面积。



△ Z15 中央处理器



△ Z15 系统控制器

x86 服务器 CPU 对 eDRAM 则没有什么兴趣。

在处理器内部，其面积占用依旧不可忽视，且其本质是 DRAM，目前仍未看到 DRAM 能够推进到 10nm 以下制程。IBM 的 Power10 基于三星的 7nm 制程，便不再提及 eDRAM 的问题。

在处理器外部，eDRAM 并非业界广泛认可的标准化产品，市场规模小，成本偏高，性能和容量也相对有限。

后起之秀 HBM（High Bandwidth Memory，高带宽内存）则很好的解决了上述问题：

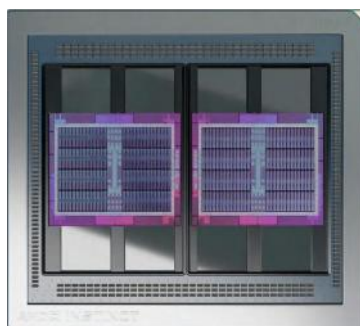
- 首先，不去 CPU 所在的 die 里抢地盘；
- 其次，纵向堆叠封装，可通过提升存储密度实现扩容；
- 最后，在前两条的基础上，较好的实现了标准化。

HBM 的好处都是通过与 CPU 核心解耦实现的，代价是生态位更靠近内存而不是 Cache，以时延换容量，很科学。



HBM 崛起：从 GPU 到 CPU

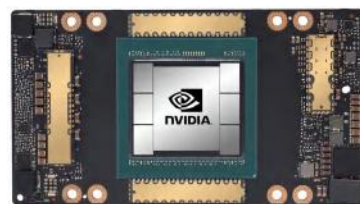
HBM 是 2014 年 AMD、SK 海力士共同发布的，使用 TSV 技术将数个 DRAM Die（晶片）堆叠起来，大幅提高了容量和数据传输速率。随后三星、美光、NVIDIA、Synopsys 等企业积极参与这个技术路线，标准化组织 JEDEC 也将从 HBM2 列入标准（JESD235A），并陆续迭代了 HBM2e（JESD235B），以及 HBM3（JESD235C）。得益于堆叠封装，以及巨大的位宽（单封装 1024bit），HBM 提供了远超其他常见内存形态（DDR DRAM、LPDDR、GDDR 等）的带宽和容量。



△ AMD Instinct MI250X



△ NVIDIA Tesla P100



△ NVIDIA GPU A100

	HBM	HBM2	HBM2E	HBM3
传速率（单 pin）	1Gbps	3.2Gbps	3.65Gbps	6.4Gbps
封装内堆叠数量	4	2/4/8	4/8/12	4/8/12/16
最大封装容量	4GB	16GB	24GB	64GB
带宽（1024bit）	128GBps	410GBps	460GBps	819GBps

典型的实现方式是通过 2.5D 封装将 HBM 与处理器核心连接，这在 CPU、GPU 等产品中均有应用。早期也有观点把 HBM 视作 L4 Cache，从 TB/s 级的带宽角度看，也算合理。而从容量角度，HBM 就比 SRAM 或 eDRAM 大太多了。由此，HBM 既可以胜任（一部分）Cache 的工作，也可以当做高性能内存使用。

AMD 是 HBM 的早期使用者，发展至今，AMD Instinct MI250X 计算卡在单一封装内集成了 2 颗计算核心和 8 颗 HBM2e，容量共 128GB，带宽达到 3276.8GB/s。

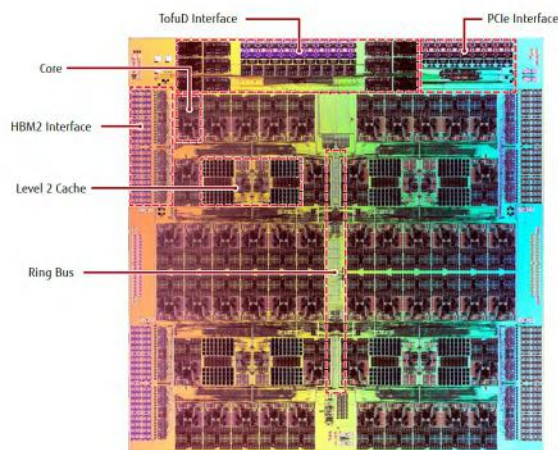
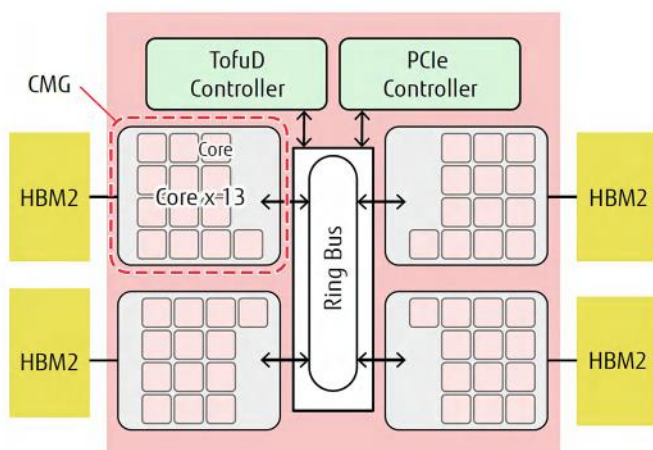
NVIDIA 应用 HBM 的主要是专业卡，其 2016 年的 TESLA P100 的 HBM 版搭配了 16GB HBM2，随后的 V100 搭配了 32GB HBM2。目前当红的 A100 和 H100 也都有 HBM 版，前者最大提供 80GB HBM2e、带宽约 2TB/s；后者升级到 HBM3，带宽约 3.9TB/s。

华为的昇腾 910 处理器也集成了 4 颗 HBM。对于计算卡、智能网

+ + + + + +
+ + + + + +

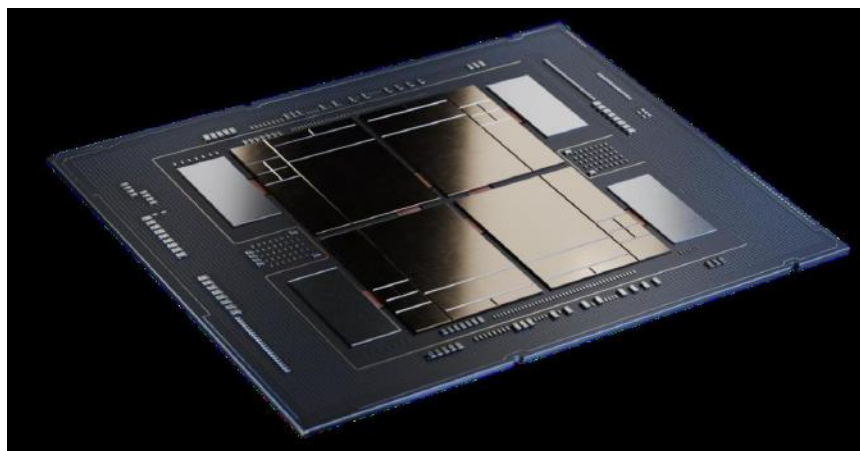
卡、高速 FPGA 等产品，HBM 作为一种 GDDR 的替代品，应用已经非常成熟了。

CPU 也已开始集成 HBM，其中最突出的案例是曾经问鼎超算 TOP500 的富岳（Fugaku），使用富士通研发的 A64FX 处理器。A64FX 基于 Armv8.2-A，采用 7nm 制程，每封装内集成了 4 颗 HBM2，容量 32GB，带宽 1TB/s。



△ 富士通 A64FX CPU

英特尔在 2023 年 1 月中与第四代至强可扩展处理器一同推出的至强 Max 系列，在前者的基础上集成了 64GB 的 HBM2e。这些 HBM2e 可以作为内存独立使用（HBM Only 模式），也可以搭配 DDR5 内存共同使用（HBM Flat Mode 和 HBM Caching Mode 两种工作模式）。

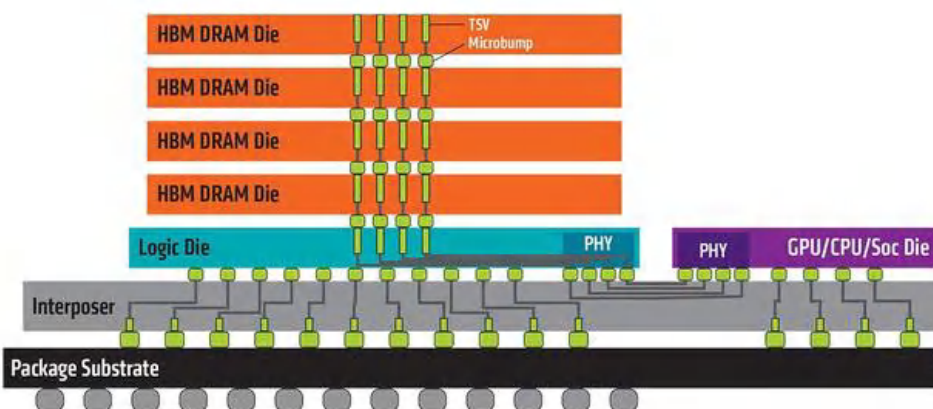


△ Intel Xeon Max 系列，注意外围的 4 颗 HBM 芯片



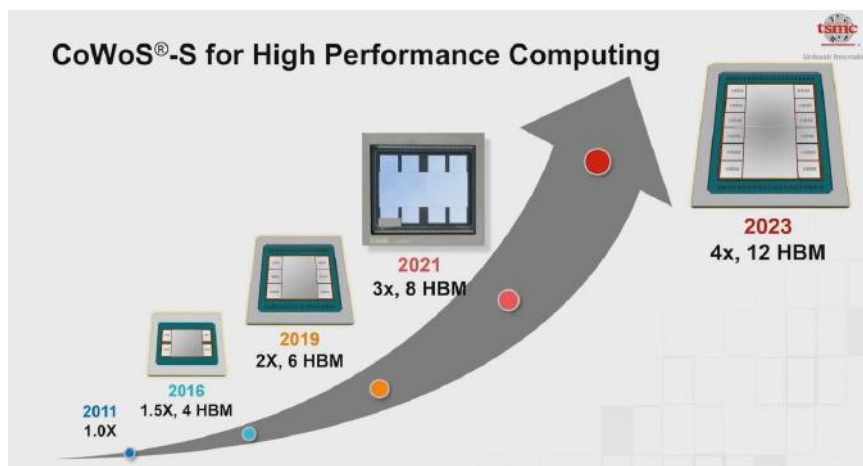
中介层：CoWoS 与 EMIB

值得一提的是，目前 HBM 与处理器“组装”在一起都需要借助硅中介层。传统的 ABS 材质基板等难以胜任超高密度的触点数量和高频率。但硅中介层有两种技术思路，代表是台积电的 CoWoS（chip-on-wafer-on-substrate）和英特尔的 EMIB（Embedded Multi-die Interconnect Bridge）。



△ HBM 的基本结构。左侧彩色的 5 层结构为 HBM 封装。灰色为中介层

台积电 CoWoS-S 通过硅中介层承载处理器和 HBM。其硅中介层也被称为硅基础层，因为中介层会完全承载其他芯片。换句话说，处理器和若干 HBM 的投影面积决定了硅基础层的大小，而基础层的面积会限制 HBM 的使用数量（常见的就是 4 颗）。硅中介层使用 65nm 之类的成熟工艺制造，其成本并不高昂，但尺寸受限于光刻掩膜尺寸。这就成为了早期 HBM 应用的瓶颈——需要 HBM 的往往是高性能的大芯片，而大芯片的规模本身就已经逼近了掩膜尺寸极限，给 HBM 留下的面积非常有限。到了 2016 年，台积电终于突破了这个限制，实现 1.5 倍于掩膜尺寸的中介层，从此单芯片内部可封装 4 颗 HBM，这就是当前市场上的主流形态了。



△ 台积电 CoWoS-S 发展路线

“

英特尔认为只需要通过硅中介层连接内存和处理器的 PHY 部分，其他信号依然可以直通基板。整体而言，EMIB 充分利用了硅中介层和有机载板的技术特点和电气特性，但也存在组装成本高的缺点。

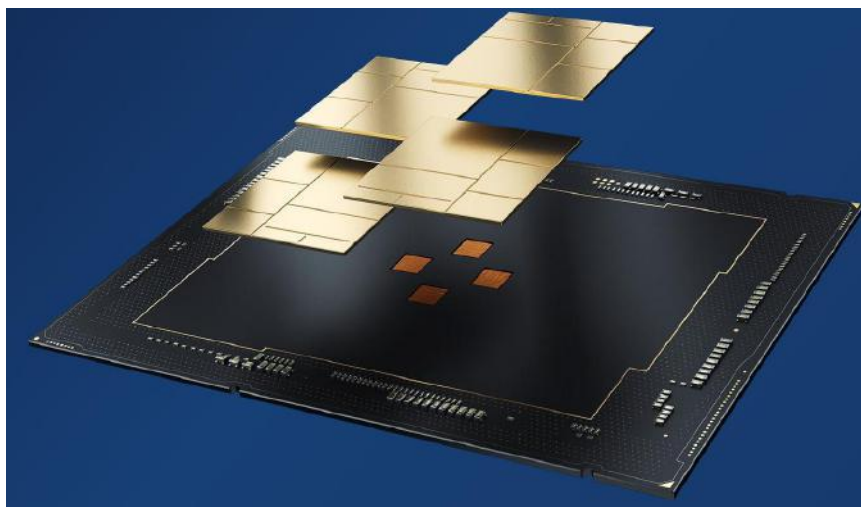
2019 年，台积电宣称实现 2 倍掩膜尺寸，可以支持 6 颗 HBM 了。很快，2020 年发布的 NEC SX-Aurora TSUBASA 矢量处理器，集成 6 颗共 48GB HBM2；同年的英伟达 A100 则是 6 颗共 40GB HBM2e（有一颗 HBM 未启用）。

至于可以封装 12 颗 HBM 的巨型芯片，预计面积将达到 3200 平方毫米。硅中介层的面积如此发展，下一个瓶颈就是硅晶圆的切割效率了。

另一种思路是英特尔的 EMIB，使用的硅中介层要小得多。以第四代英特尔至强可扩展处理器的渲染图为例，棕色的小方块就是 EMIB 的“桥”，用以将 4 个 XCC 的 die 拼为一个整体；而在至强 Max 系列中，每个 die 还需要通过 EMIB 去连接对应的 HBM 芯片。结合 HBM 的架构示意图可以看出，英特尔认为只需要通过硅中介层连接内存和处理器的 PHY 部分，其他信号依然可以直通基板。整体而言，EMIB 充分利用了硅中介层和有机载板的技术特点和电气特性，但也存在组装成本高的缺点（需要在有机载板中镶嵌，增加了工艺复杂度，限制了载板的选择）。

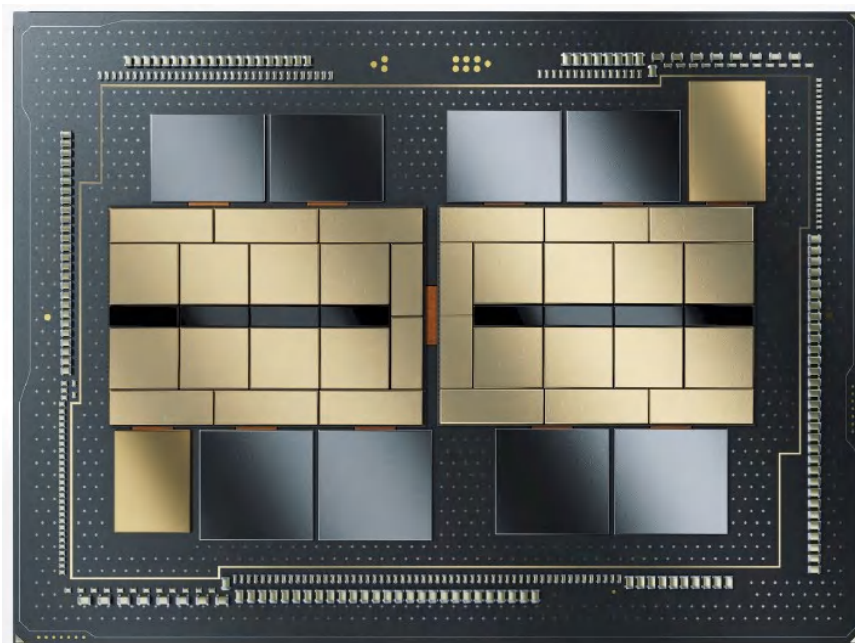
当然，对于更复杂的“组装”，英特尔也有对应的方案，如代号 Ponte Vecchio 的英特尔数据中心 GPU Max 系列整合了基于 5 种制造工艺生产的 47 个小芯片，其中的基础层（Base Die）的面积为 650mm²。该产品综合了 Foveros 3D 封装和 EMIB 2.5D 封装的特点，纵向横向齐发展。

+ + + + +
+ + + + +



向下发展：基础层加持

英特尔数据中心 Max GPU 系列引入了 Base Tile 的概念，姑且称之为基础芯片。相对于中介层的概念，我们也可以把基础芯片看做是基础层。基础层表面上看与硅中介层功能类似，都是承载计算核心、高速 I/O（如 HBM），但实际上功能要多得多。硅中介层的本质是利用成熟的半导体光刻、沉积等工艺（65nm 等级），在硅上形成超高密度的电气连接。而基础层更进一步：既然都要加工多层图案，为什么不把逻辑电路之类的也做进去呢？

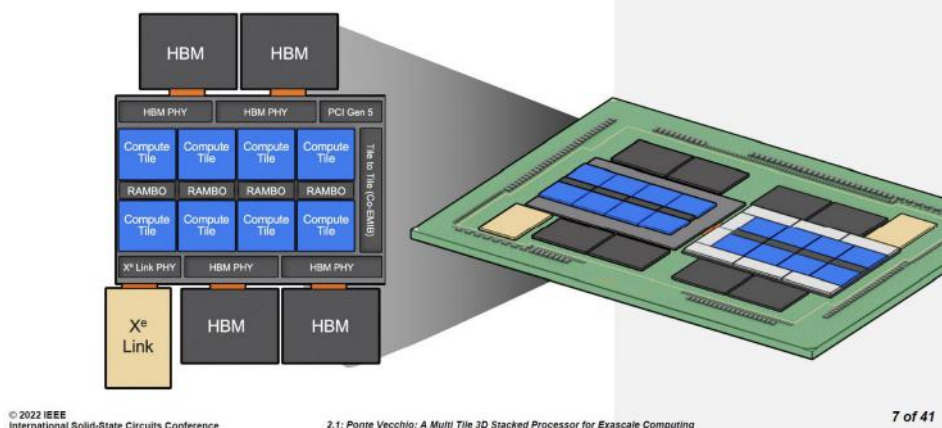


△ 英特尔数据中心 Max GPU



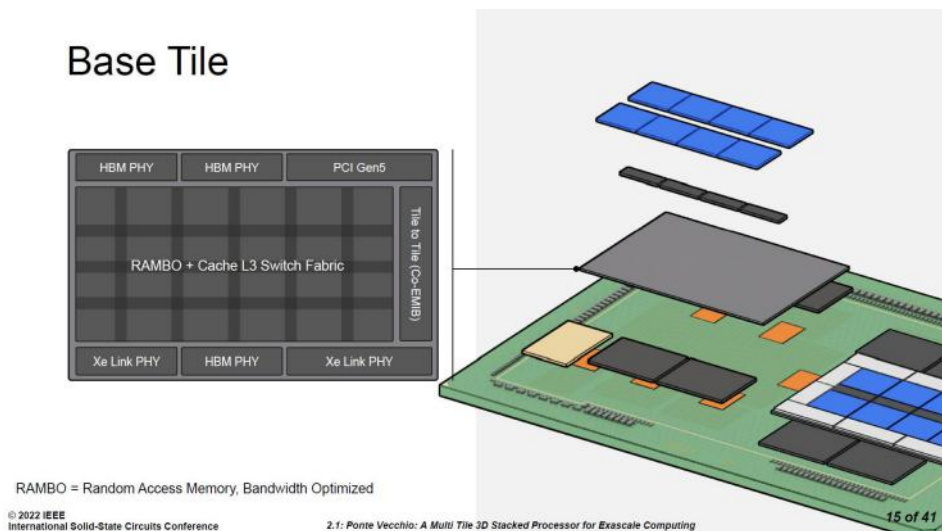
Intel 在 ISSCC2022 中展示了英特尔数据中心 Max GPU 的 Chiplet（小芯片）架构，其中，基础芯片面积为 640mm^2 ，采用了 Intel 7 制程——这是目前 Intel 用于主流处理器的先进制程。为何在“基础”芯片上就需要使用先进制程呢？因为 Intel 将高速 I/O 的 SerDes 都集成在基础芯片中了，其作用有点儿类似 AMD 的 IOD。这些高速 IO 包括 HBM PHY、Xe Link PHY、PCIe 5.0，以及这一节的重点：Cache。这些电路都比较适合 5nm 以上的工艺制造，将它们与计算核心解耦后重新打包在一个制程之内是相当合理的选择。

Multi-Tile Architecture



△ 英特尔数据中心 Max GPU 的 Chiplet 架构

Base Tile



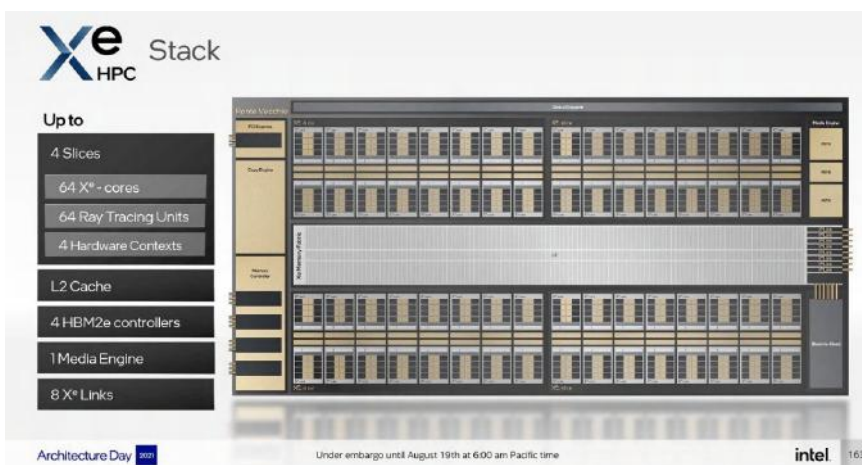
△ 英特尔数据中心 Max GPU 的基础芯片。注意，此图中的两组 Xe Link PHY 应是笔误。芯片下方应为两个 HBM PHY 和一个 Xe Link PHY



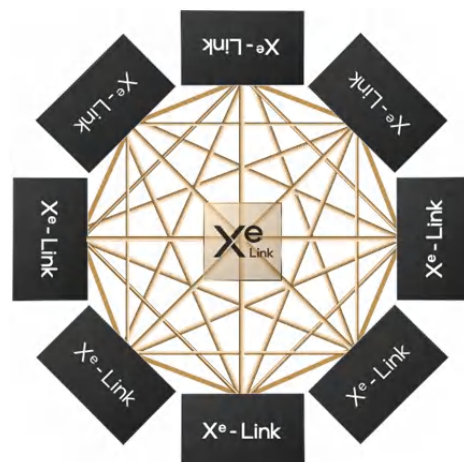
英特尔数据中心 Max GPU 系列通过 Foveros 封装技术在基础芯片上方叠加 8 颗计算芯片 (Compute Tile)、4 颗 RAMBO 芯片 (RAMBO Tile)。计算芯片采用台积电 N5 工艺制造，每颗芯片自有 4MB L1 Cache。RAMBO 是 “Random Access Memory, Bandwidth Optimized” 的缩写，即为带宽优化的随机访问存储器。独立的 RAMBO 芯片基于 Intel 7 制程，每颗有 4 个 3.75MB 的 Bank，共 15MB。每组 4 颗 RAMBO 共提供了 60MB 的 L3 Cache。此外，在基础芯片中也有 RAMBO，容量 144MB，外加 L3 Cache 的交换网络 (Switch Fabric)。

因此，在英特尔数据中心 Max GPU 中，基础芯片通过 Cache 交换网络，将基础层内的 144MB Cache，与 8 颗计算芯片、4 颗 RAMBO 芯片的 60MB Cache 组织在一起，总共 204MB L2/L3 Cache，整个封装是两组，就是 408MB L2/L3 Cache。

英特尔数据中心 Max GPU 的每组处理单元都通过 Xe Link Tile 与另外 7 组进行连接。Xe Link 芯片采用台积电 N7 工艺制造。



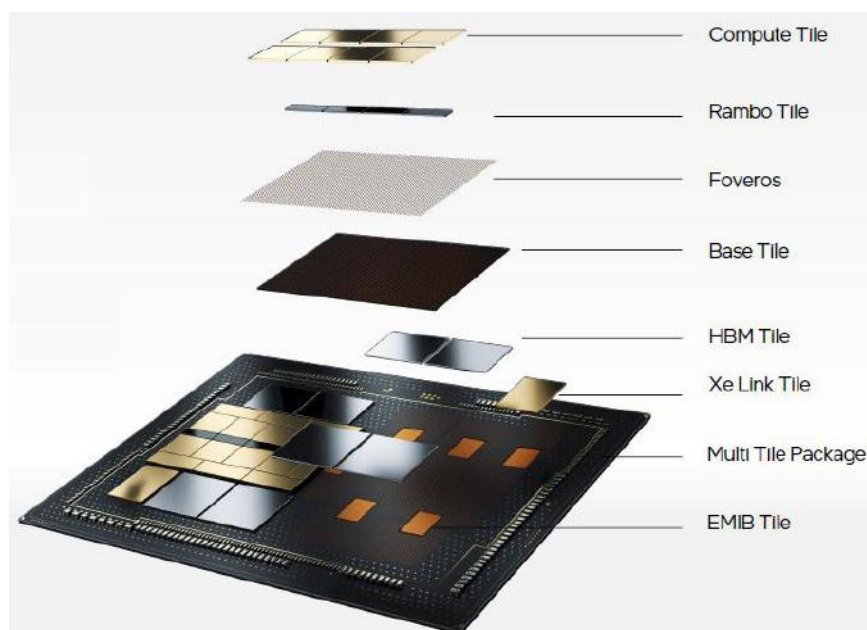
△ Xe HPC 的逻辑架构



◁ Xe Link 的网状连接



前面已经提到，I/O 芯片独立是大势所趋，共享 Cache 与 I/O 拉近也是趋势。英特尔数据中心 Max GPU 将 Cache 与各种高速 I/O 的 PHY 集成在同一芯片内，正是前述趋势的集大成者。至于 HBM、Xe Link 芯片，以及同一封装内相邻的基础芯片，则通过 EMIB（爆炸图中的橙色部分）连接在一起。



△ 英特尔数据中心 Max GPU 爆炸图

根据英特尔在 HotChips 上公布的数据，英特尔数据中心 Max GPU 的 L2 Cache 总带宽可以达到 13TB/s。考虑到封装了两组基础芯片和计算芯片，我们给带宽打个对折，基础芯片和 4 颗 RAMBO 芯片的带宽是 6.5TB/s，依旧远远超过了目前至强和 EPYC 的 L2、L3 Cache 的带宽。其实之前 AMD 已经通过指甲盖大小的 3D V-Cache 证明了 3D 封装的性能，那就更不用说英特尔数据中心 Max GPU 的 RAMBO 及基础芯片的面积了。

Ponte Vecchio 2-Stack	Register File	L1 Cache	L2 Cache	HBM
Maximum Size	64 MB	64 MB	408 MB	128 GB
				
Peak Read Bandwidth	419 TB/s	105 TB/s	13 TB/s	3.2 TB/s
				

△ 英特尔数据中心 Max GPU 的存储带宽



回顾一下 3D V-Cache 的弱点——“散热”不良，我们还发现将 Cache 集成到基础芯片当中还有一个优点：将高功耗的计算核心安排在整个封装的上层，更有利于散热。

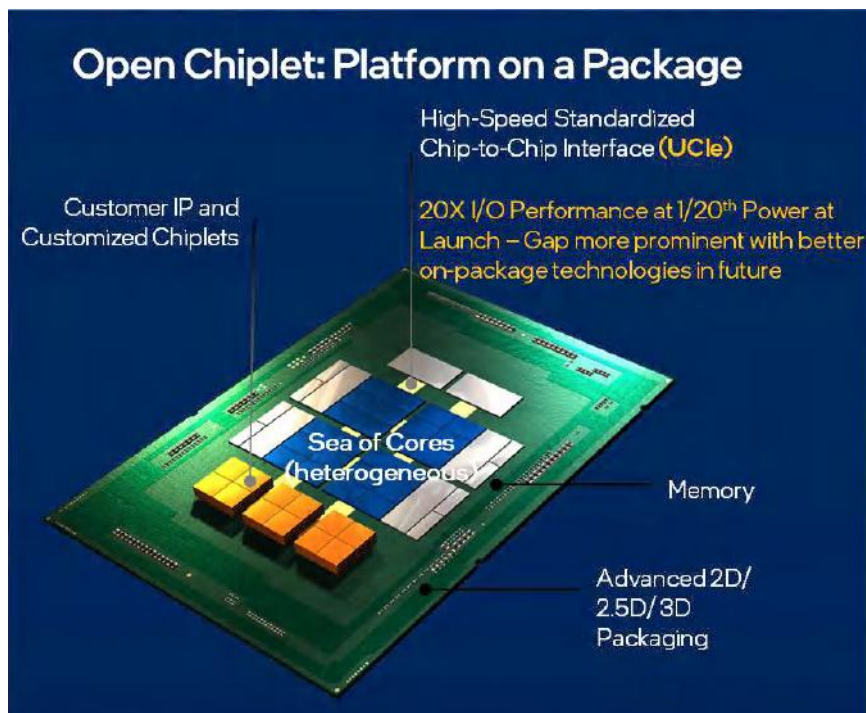
再往远一些看，在网格化的处理器架构中，L3 Cache 并非简单的若干个块（切片），而是分成数十甚至上百单元，分别挂在网格节点上的。基础芯片在垂直方向可以完全覆盖（或容纳）处理器芯片，其中的 SRAM 可以分成等量的单元与处理器的网格节点相连。换句话说，对于网格化的处理器，将 L3 Cache 移出到基础芯片是有合理性的。目前已经成熟的 3D 封装技术的凸点间距在 30 ~ 50 微米的量级，足够胜任每平方毫米内数百至数千个连接的需要，可以满足当前网格节点带宽的需求。更高密度的连接当然也是可行的，10 微米甚至亚微米的技术正在推进当中，但优先的场景是 HBM、3D NAND 这种高度定制化的内部堆栈的混合键合，未必适合 Chiplet 对灵活性的要求。



标准化：Chiplet 与 UCle

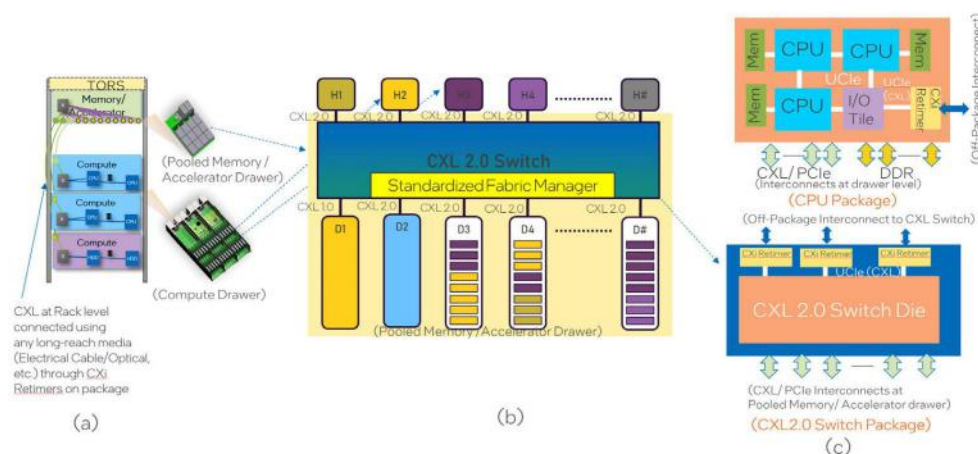
Chiplet 的优势已经获得了充分的验证，接下来的问题就是通用化、标准化。通过标准化，来自不同供应商的芯片可以更容易地实现封装内的互联，在这个前提下，部分 IP 可以固化为芯片，而不再需要分别集成到不同客户的芯片中，也不需要适配太多版本的生产工艺。

在此愿景之下，2022 年 3 月，通用处理器市场的核心玩家 Intel、AMD、Arm 等联合发布了新的互联标准 UCle（Universal Chiplet Interconnect Express，通用小芯片互连通道），希望解决 Chiplet 的行业标准问题。



由于标准的主导者与 PCIe 和 CXL（Compute Express Link）已有千丝万缕的关系，因此，UCle 非常强调与 PCIe/CXL 的协同，在协议层本地端提供 PCIe 和 CXL 协议映射。

与 CXL 的协同，说明 UCle 的目标不仅仅是解决芯片制造中的互联互通问题，而是希望芯片与设备、设备与设备之间的交互是无缝的。在 UCle 1.0 标准中，即展现了两种层面的应用：Chiplet（In package）和 Rack space（Off package）。



△ UCle 规划的机架连接交给了 CXL

CXL：内存的解耦与扩展

PCIe 经过十年的发展，已经是最为广泛的板卡互连协议。这种兼容性基础正在向节点外扩展，也就是 UCle 所称的 Rack（机柜）空间。随着新一代 Arm 和 x86 架构服务器处理器平台（第四代英特尔至强可扩展处理器和 AMD 第四代 EPYC 处理器）进入市场，CXL 协议有望获得广泛的支持。

当前 CXL 1.1 的物理层基于成熟的 PCIe 5.0。以第四代英特尔至强可扩展处理器公开宣称的支持 CXL Type 1、Type 2 Device 看，首先从 CXL 获益的将是 GPGPU、智能网卡、计算卡等设备。而非常有趣的是，AMD 第四代 EPYC 处理器则完全相反，声称支持 CXL Type 3 Device，也就是 CXL 内存模块，而不支持 Type 1、Type 2 Device。

Driving Platform Innovation

Broad Ecosystem Readiness To Deploy Today

DDR5

- 8ch DDR5 (per CPU): up to 4800 MT/s
- 9x4 RDIMM support
- 3DS RDIMM support
- New RAS features
- Enhanced ECC
- Error Check and Scrub
- Both 2DPC and 1DPC today

Up to **1.5x**
higher memory bandwidth
vs. DDR4

PCIe 5.0

- 80 lanes
- Improved DDIO and QoS capabilities
- X2 bifurcation @ Gen 4

Up to **2x**
Increased I/O Bandwidth
vs. PCIe 4.0

CXL 1.1

- Next Gen I/O
- Up to 4 CXL devices supported per CPU
- Type 1 device: CXL.io and CXL.cache (e.g. SmartNIC)
- Type 2 device: CXL.io, CXL.cache and CXL.mem (e.g. GPU, ASIC, FPGA)

Type 1 and Type 2

UPI 2.0

- Up to 4 UPI links @ 16 GT/s
- New 8S-4 UPI Performance Optimized Topology

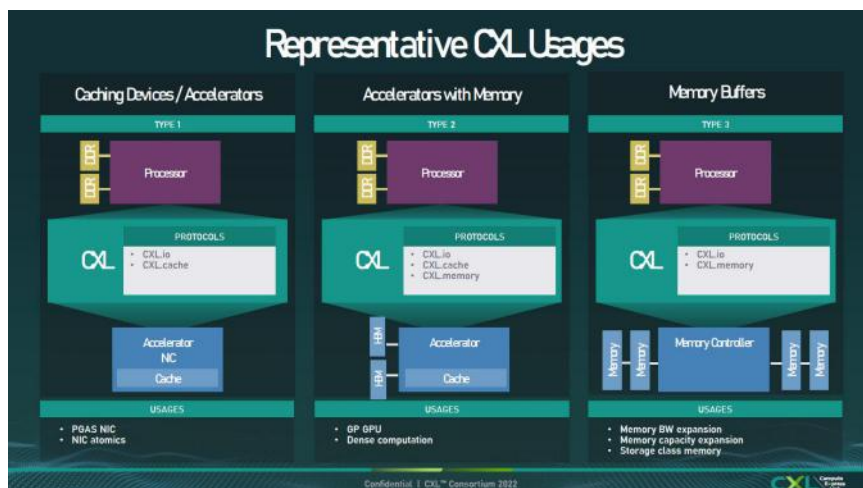
Up to **1.9x**
Increased inter-socket
bandwidth vs. prior gen

intel
xeon Accelerate with Xeon

See backup for workloads and configurations. Results may vary.

△ 第四代英特尔至强可扩展处理器正式支持 CXL 1.1 中的 Type 1、Type 2

+ + + + +
+ + + + +



△ CXL 定义的三种类型设备

相对于 PCIe，CXL 最重要的价值是减少了各子系统内存的访问延迟（理论上 PCIe 协议的延迟为 100ns 量级，CXL 为 10ns 量级），譬如 GPU 访问系统内存，这对于设备间的大容量数据交换至关重要。这种改进主要来源于两方面：

首先，PCIe 在设计之初没有考虑缓存一致性问题，通过 PCIe DMA 跨设备读写数据时，在操作延迟期间，内存数据可能已经发生变化，因此需要额外加入验证过程，这增加了指令复杂度和延迟。而 CXL 通过 CXL.cache 和 CXL.memory 协议解决了缓存一致性问题，简化了操作，也减少了延迟。

其次，PCIe 的初衷是大流量，针对大数据块（512B、1KB、2KB、4KB）进行优化，希望减少指令开销。CXL 则针对 64B 传输进行优化，对于固定大小的数据块而言，操作延迟较低。换言之，PCIe 发展至今，其协议特点更适合用于 NVMe SSD 为代表的块存储设备，而对于看重字节级寻址能力的计算型设备，CXL 更为适合。

除了充分释放异构计算的算力，CXL 还让内存池化的愿景看到了标准化的希望。CXL Type 3 Device 的用途就是 Memory Buffer（内存缓冲），利用 CXL.io 和 CXL.memory 的协议实现扩展远端内存。在扩展后，系统内存的带宽和容量即为本地内存和 CXL 内存模块的叠加。

在新一代 CPU 较普遍支持的 CXL 1.0/1.1 中，CXL 内存模块先实现了主机级的内存扩展，试图突破传统 CPU 内存控制器的发展瓶颈，CPU

+ + + + + +
+ + + + + +

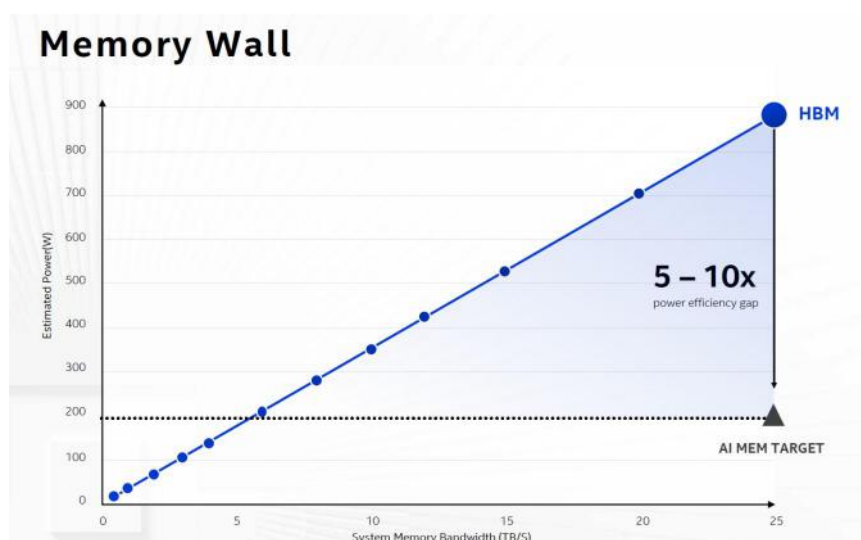


过去十年间，CPU 的核心数量从 8 ~ 12 个的水平，增长到了 60 乃至 96 核，Arm 已有 192 核的产品。而每插槽 CPU 的内存通道数仅从 4 通道增加到 8 或 12 通道。每通道的内存在此期间也经过了三次大的迭代，带宽大概增加 1.5 ~ 2 倍，存储密度大约为 4 倍。

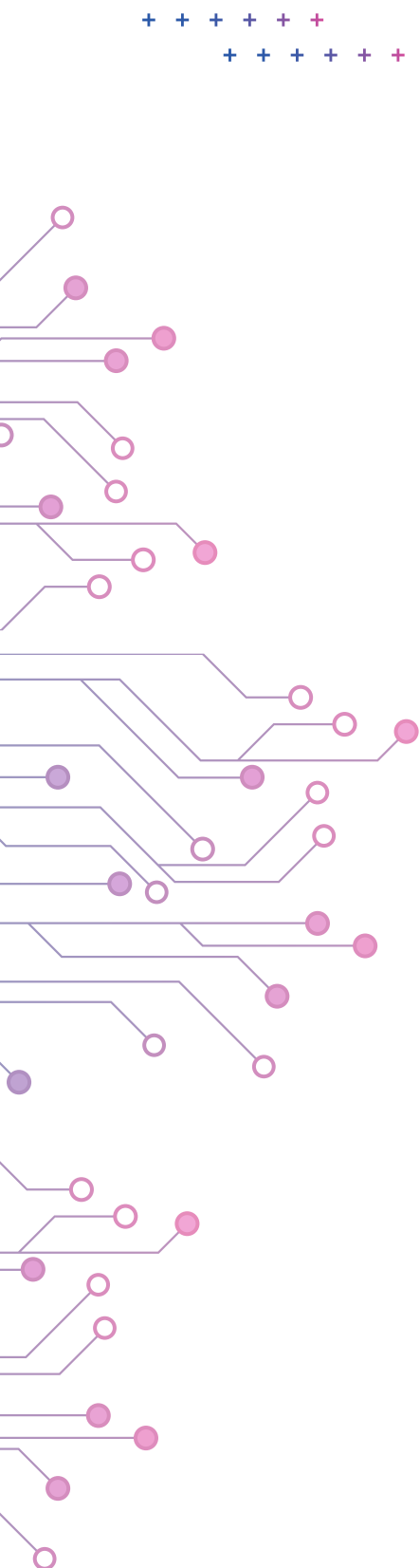
+ + + + + +
+ + + + + +

核心数量增长的速度远远快于内存通道的增加速度是原因之一。

过去十年间，CPU 的核心数量从 8 ~ 12 个的水平，增长到了 60 乃至 96 核，Arm 已有 192 核的产品，而每插槽 CPU 的内存通道数仅从 4 通道增加到 8 或 12 通道。每通道的内存在此期间也经过了三次大的迭代，带宽大概增加 1.5 ~ 2 倍，存储密度大约为 4 倍。从发展趋势来看，每个 CPU 核心所能分配到的内存通道数量在明显下降，每核心可以分配的内存容量和内存带宽其实也有所下降。这是内存墙的一种表现形式，导致 CPU 核心因为不能充分得到数据来处于满负荷的运行状态，会导致整体计算效率下降。



为什么增加内存通道如此缓慢？因为增加内存通道不仅仅需要增加芯片面积，还需要扩展对外接口，在电气连接方式没有根本性改变的情况下，触点数量的大量增加会导致 CPU 封装面积剧增。10 年前的英特尔至强（Intel Xeon）处理器的 LGA2011 封装尺寸为 52.5mm×45.0mm（毫米），当前 Xeon 所用 LGA 4677 封装尺寸为 77.5mm×56.5mm，触点数量增加了 1.33 倍，封装面积增加了 1.85 倍。而 AMD 第四代 EPYC 启用的新封装 SP5 更大，有 6096 个触点，封装面积达到 75.4mm×72mm，跟一张扑克牌差不多大了，毕竟它的内存通道数量达到了 12 个。为了与 AMD 和 Arm 继续“核战”，英特尔代号 Granite Rapids 和 Sierra Forest 的下一代 Xeon 将启用 LGA 7529 插槽，尺寸 105mm×70.5mm。作为参考，iPhone 4 的正面尺寸是 115.2mm×58.6mm，iPhone 8 则为 138.4mm×67.3mm。



△ LGA 4677 已接近信用卡大小

同时，主板上内存相关的走线数量和距离也需要相应增加，保证信号质量的难度加大。CPU 插槽面积增加、内存槽数量增加，还受到主板面积的限制。按照英特尔和 AMD 的通用处理器的这个发展趋势，双路服务器的主板布局将会愈加困难，其市场份额可能会逐步下降。

通过 CXL 扩展内存，可以将 CPU 与内存从沿革多年的紧耦合关系变为松耦合，利用 PCIe/CXL 通道的物理带宽增加内存总带宽，而不仅仅限于内存控制器自身的通道总数（即使前者的带宽相对较低，但也是增量），利用机箱的立体空间容纳更大容量的内存，而不再受主板面积的约束。



△ CXL 内存

+ + + + + +
+ + + + + +

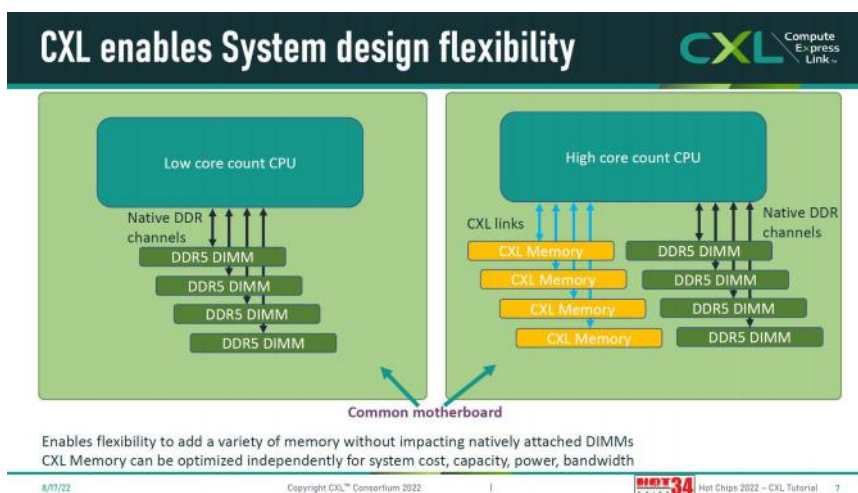
“

有约 50% 的服务器的实际内存利用率不到一半。这是由于内存的分配是与 CPU 核心绑定的，当客户按照预设的实例配置租用资源时，每个核心便搭配了固定容量的内存，譬如 2GB。当主机的 CPU 核心数量被分配完毕后，未被搭配的内存便被闲置了。

考虑到人工智能，尤其是机器学习领域的发展，模型容量在过去 5 年间大致增加了 50 倍，内存容量的扩展方式确实值得突破一下。不过这也不是一蹴而就的，毕竟第四代英特尔至强可扩展处理器每插槽 CPU 只支持 4 个 CXL 设备，给计算卡之类的一分就没了。所以也就不用纠结它暂时没有宣布支持 CXL Type 3 Device (Memory Buffer)。

在第四代可扩展至强处理器平台上，如果支持 CXL 1.1 的加速卡 / 计算卡 / 智能网卡能够提供比 PCIe 5.0 更好的性能，稍微拉近跟 SMX 接口 (NVLink) 的性能落差，那就非常开心了。而 AMD 则反过来，处理器大核确实多，而且不论单路还是双路处理器，内存槽上限都是 24 条，如果不优先另辟蹊径扩展内存容量，每个核心能够分配到的内存资源其实反而会落了下风，补短板看起来更迫切。但是，AMD 同样也会面临内存扩展与计算卡抢 PCIe 通道数量的问题。

总之，不论这两家通用处理器具体各怀啥心思，CXL 的第一轮普及工作就是不尽如人意，顾此失彼。甚至现在还不到纠结内存扩展的时候，即使 CXL 内存模组已然是各种技术论坛中样品最接近现实的 CXL 设备。在这个阶段，解决 CXL 设备的有无问题，借机逐步导入 EDSFF，初步形成生态环境，就算是成功。至于内存的大事情，且得看下一代平台以及更新版本的 CXL。



△ CXL 的本地内存扩展



当 CPU 本地内存不足时，再到内存池调用。这虽然增加了一些访问延迟，但会降低内存的总成本。如果减少 10% 的内存搭配数量，对于大型数据中心而言也是数以亿计的资金节约。微软预计通过 CXL 和内存池化，可以为云数据中心减少 4 ~ 5% 的成本。

到了 CXL 2.0，通过 CXL Switch，内存扩展将可以跨 CPU 实现。这个阶段将构建机柜级的资源池化。这其中的好处多多，此处主要集中在云服务的需求角度去看。

微软曾调研了 Azure 公有云数据中心的内存使用情况，其结论是：有约 50% 的服务器的实际内存利用率不到一半。这是由于内存的分配是与 CPU 核心绑定的，当客户按照预设的实例配置租用资源时，每个核心便搭配了固定容量的内存，譬如 2GB。当主机的 CPU 核心数量被分配完毕后，未被搭配的内存便被闲置了。考虑到预先配置的内存容量相对核心数量必然是超配的，譬如 56 核的至强，搭配 128GB 内存，每个实例配 2GB 内存的话，那注定有 $128 - 2 \times 56 = 16\text{GB}$ 内存将会被闲置。如果服务器核心未被充分利用，被闲置的内存将会更多。而运行中的实例，其实际内存占用率通常也不高。由此，无从分配的、未被分配的、分配但未充分使用的，这三种性质的浪费叠加之后，主机的实际内存浪费相当惊人。

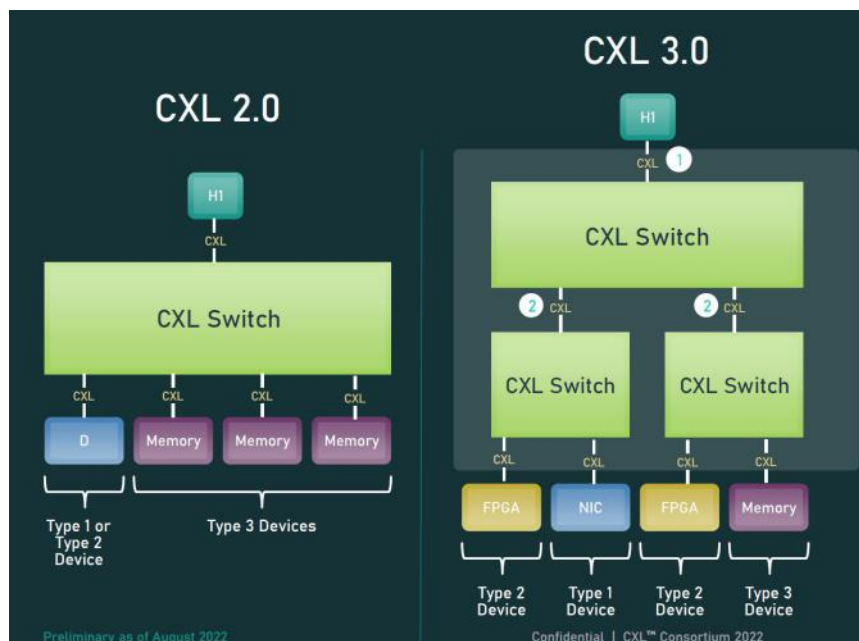
由此，微软提出通过内存池来解决这个问题。各主机搭配容量较少的内存，其余内存放入内存资源池。当 CPU 本地内存不足时，再到内存池调用。这虽然增加了一些访问延迟，但会降低内存的总成本。如果减少 10% 的内存搭配数量，对于大型数据中心而言也是数以亿计的资金节约。微软预计通过 CXL 和内存池化，可以为云数据中心减少 4 ~ 5% 的成本。

除了节约总内存投入，内存池化还可以带来内存持久化、内存故障热迁移等等新的功能特性以供业界进一步挖掘，此处暂不展开。

CXL 的完整愿景，需要到 CXL 3.0 规范才能实现。

首先是带宽，CXL 3.0 基于 PCIe 6.0，更换了 PCIe 沿革多年的 NRZ 调制方案，变为 PAM-4 脉冲幅度调制编码，在电气特性变化不大的情况下，链路带宽翻倍，从 32GT/s 提升到了 64GT/s。

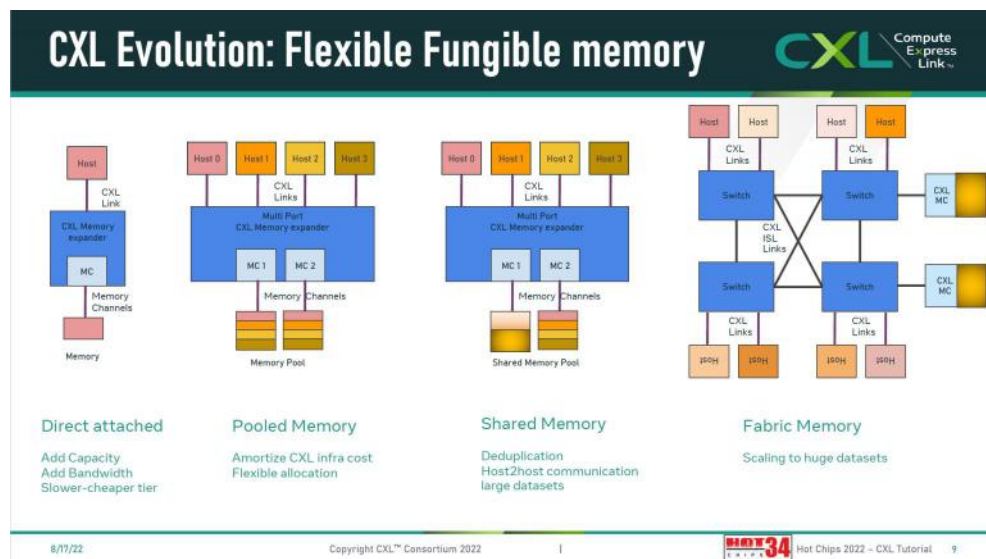
其次，CXL 3.0 增加了对二层交换机的支持，也就是叶脊（Leaf-Spine）网络架构，资源池化也不再局限于内存，而是可以实现 CPU 资源池、加速器资源池、网卡资源池等。



△ CXL 3.0 将改变资源的组织方式

CXL 2.0 实现的是机柜内的池化，CXL 3.0 除了可以在一个机柜内实现计算资源和存储资源的解耦和池化，还可以在多个机柜之间建立更大的资源池。跨主机、跨机柜调度规模巨大的计算资源，已经是超算的范畴了。然后，CXL 3.0 网络可以支持 4096 个 CXL 节点！单纯从数量上看，这远远超过了 NVLink 网络 256 个节点的规模（见下一章）。这将是 CXL 对私有但标榜高性能的 NVLink 最有力的挑战。当然，CXL 3.0 依旧暂时还未落地，而 NVIDIA 新一代的系统已经正式发布了。二者在机柜互联方面的带宽远超 400G InfiniBand 或者以太网，实际运行效率都是非常值得期待的。

另外，考虑到 CPU 和加速器都可以从内存池访问数据，那么，CPU 确实不需要再去（替其他设备）管理那么多本地内存。毕竟，计算卡通过 CXL 访问 CPU 内存控制器下的内存，和访问内存资源池，瓶颈都在 CXL，性能上没有本质差异。因此，CPU 可以搭配容量更小，但速度更高的内存，例如 HBM 等。如此一来，CPU 就可以作为一种更高效的计算资源存在，而不再负担统筹的工作。到这一层次的时候，这几年时不时被谈起的诸如 CPU 为中心、DPU 为中心之类的话题也就没有太大意义了。



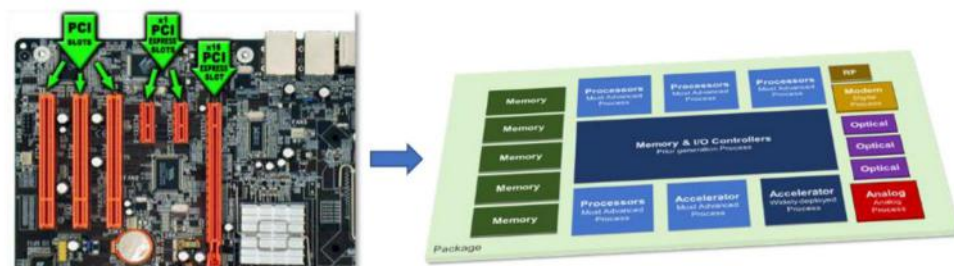
△ CXL 规划了多种内存组织方式

UCIe 与异构算力

UCIe 的 In package 本质就是将整个芯片封装视作主板，在基板上组装大量的芯粒，包括各种处理器、收发器，以及硬化的 IP。整体而言，UCIe 是一个基于并行连接的高性能系统接口，主要是面向 PCIe/CXL 设备（芯片）的“组装”，如 CPU、GPU、DSA、FPGA、ASIC 等的互联。

随着人工智能时代的到来，异构计算已经是显学，原则上，只要功率密度允许，这些异构计算单元的高密度集成可以交给 UCIe 完成。除了集成度的考虑，标准化的 Chiplet 也带来了功能和成本的灵活性，对于不需要的单元，在制造时不参与封装即可——而对于传统的处理器而言，对部分用户无用的单元常常成为无用的“暗硅”，意味着成本的浪费。一个典型的例子就是 DSA，如英特尔第四代可扩展至强处理器中的若干加速器，用户可以付费开启，但是，如果用户不付费呢？这些 DSA 其实已经制造出来了。

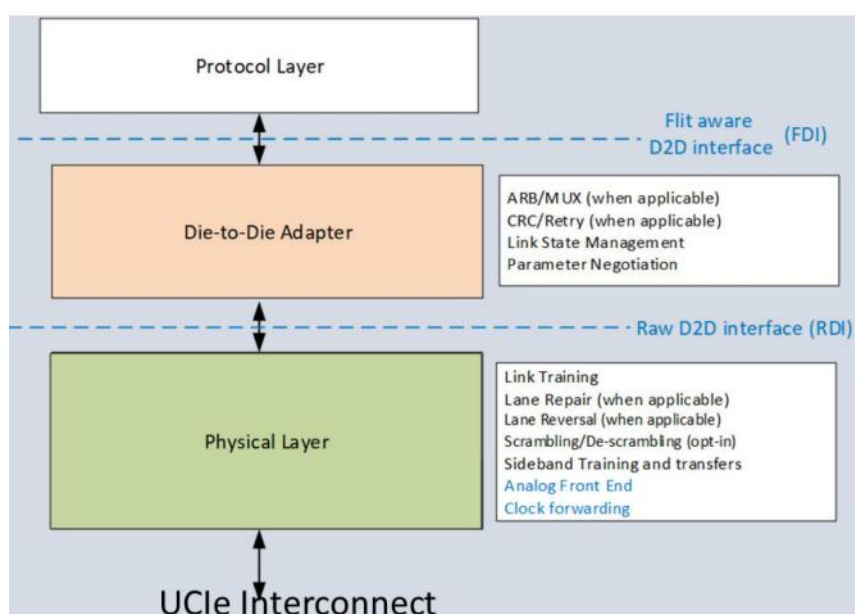
+ + + + + +
+ + + + + +

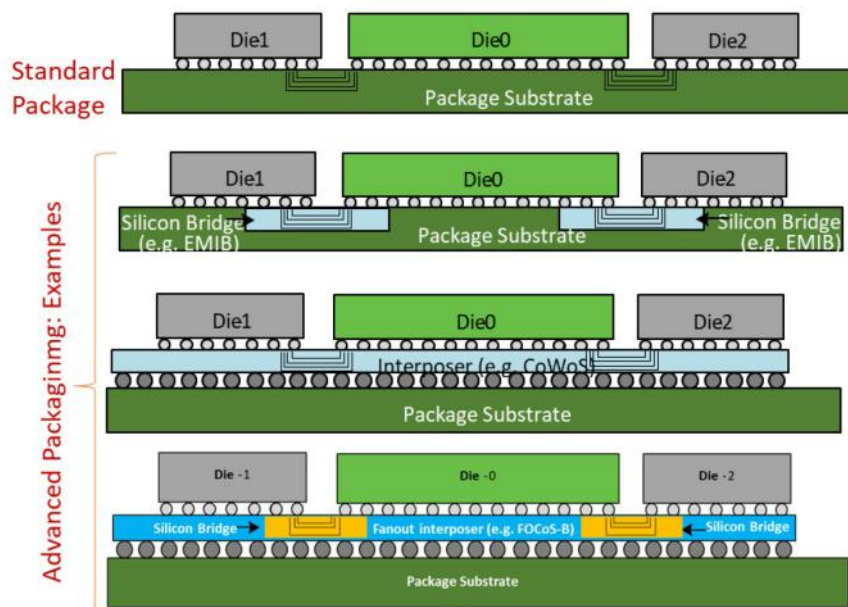


(a. Board level to Package level integration)

△ UCle 的 In package 本质就是将整个芯片封装视作主板

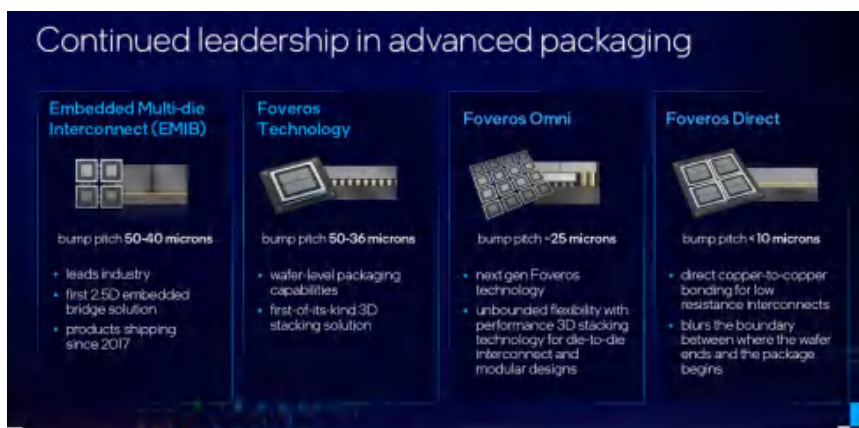
UCle 包括协议层 (Protocol Layer)、适配层 (Adapter Layer) 和物理层 (Physical Layer)。协议层支持 PCIe 6.0、CXL 2.0 和 CXL 3.0, 也支持用户自定义。根据不同的封装等级, UCle 也有不同的 Package module。通过用 UCle 的适配层和 PHY 来替换 PCIe/CXL 的 PHY 和数据包, 就可以实现更低功耗和性能更优的 Die-to-Die 互连接口。





△ UCle 对两种封装的划分

UCle 考虑了两种不同等级的封装：标准封装（Standard Package）和先进封装（Advanced Package），凸块间距、传输距离和能耗将有数量级的差异。譬如对于先进封装，凸块间距（Bump Pitch）为 $25 \sim 55\mu\text{m}$ ，对应的是采用硅中介层为代表的 2.5D 封装技术的特点。以英特尔的 EMIB 为例，当前的凸块间距即为 $50\mu\text{m}$ 左右，未来将向 $25\mu\text{m}$ ，甚至 $10\mu\text{m}$ 演进。台积电的 InFO、CoWoS 也会有类似的规格和演进。而标准封装（2D）的规格对应的是目前应用最为广泛的有机载板。



△ 英特尔先进封装的凸块间距演进



不同封装的信号密度也是有本质差异的，如标准封装模块对应的是 16 对数据线（TX、RX），而高级封装模块包含 64 对数据线，每 32 个数据管脚还提供 2 个额外的管脚用于 Lane 修复。如果需要更大的带宽，可以扩展更多的模块，且模块的频率是可以独立的。

Table 1: UCle 1.0 Characteristics and Key Metrics

Characteristics / KPIs	Standard Package	Advanced Package
Characteristics		
Data Rate (GT/s)	4, 8, 12, 16, 24, 32	
Width (each cluster)	16	64
Bump Pitch (um)	100 – 130	25 - 55
Channel Reach (mm)	<= 25	<=2
Target for Key Metrics		
B/W Shoreline (GB/s/mm)	28 – 224	165 – 1317
B/W Density (GB/s/mm ²)	22-125	188-1350
Power Efficiency target (pJ/b)	0.5	0.25
Low-power entry/exit	0.5ns <=16G, 0.5-1ns >=24G	
Latency (Tx + Rx)	< 2ns	
Reliability (FIT)	0 < FIT (Failure In Time) << 1	

△ UCle 规划了两种等级封装的性能目标

当然，UCle 没有必要急于跟进封装技术的极限，更高密度的键合通常还是为私有（协议）接口准备的，典型的如存储器（SRAM、HMB、3D NAND）的内部。UCle 能够满足通用总线的连接需求即可，如 PCIe、UPI、NVLink 等。

值得一提的是，UCle 对高速 PCIe 的深度捆绑，注定了它“嫌贫爱富”的格局。实际上，SoC（System on Chip）是一个相当宽泛的概念，UCle 面向的可以看做是宏系统集成（Macro-System on Chip）。而在传统观念中适合低成本、高密度的 SoC 可能需要集成大量的收发器、传感器、块存储设备等等。再譬如，一些面向边缘场景的推理应用、视频流处理的 IP 设计企业相当活跃，这些 IP

可能需要更灵活的商品化落地方式。既然相对低速设备的集成不在 UCle 的考虑范围内，低速、低成本接口的标准化尚有空间。

Chiplet 的中国力量

在国际大厂合纵连横推出 UCle 为代表的 Chiplet 连接标准之际，中国也并未缺席这一技术潮流，而是基于国内产业界资源，积极制定本土的相关标准。



在国际大厂合纵连横推出 UCle 为代表的 Chiplet 连接标准之际，中国也并未缺席这一技术潮流，而是基于国内产业界资源，积极制定本土的相关标准。

《小芯片接口总线技术要求》

早在 2020 年 8 月，中科院计算所牵头成立了中国计算机互连技术联盟（CCITA），重点围绕 Chiplet 小芯片和微电子芯片光 I/O 成立了两个标准工作组，并于 2021 年 6 月在工信部中国电子工业标准化技术协会立项了《小芯片接口总线技术》和《微电子芯片光互连接口技术》两项团体标准。其中小芯片项目集结了国内产业链上下游六十多家单位共同参与研究。

2022 年 3 月，由中科院计算所、工信部电子四院以及多家国内芯片厂商合作，《小芯片接口总线技术要求》完成草案并公示。

2022 年 12 月 16 日，在第二届中国互连技术与产业大会上，《小芯片接口总线技术要求》团体标准正式面向世界发布。

2023 年 2 月，由中国电子工业标准化技术协会审订，首个由中国企业和专家主导制订的 Chiplet 技术标准《小芯片接口总线技术要求》（T/CESA 1248-2023）正式实施。

《小芯片接口总线技术要求》兼顾了 PCIe 等现有协议的支持，包括并行总线接口技术、差分串行总线接口技术和单端串行总线接口技术三种，采用 DC 耦合方式以简化 PHY IP 和封装基本实现复杂度，速率 5 ~ 32GT/s，目标误码率为 1E-15。CCITA 已经在考虑和 UCle 在物理层上兼容，以降低 IP 厂商支持多种 Chiplet 标准的成本。



ICS 31.200
CCS L 56

团 体 标 准

T/CESA 1248—2023

小芯片接口总线技术要求

Technical requirements for chiplet interface bus

2023-01-13 发布

2023-02-13 实施

中国电子工业标准化技术协会 发布

Chiplet 走出“初级阶段”

为了满足板内甚至封装内高速互联的需要，半导体大厂（设计、代工）都有相关的互联总线协议和接口标准。譬如板内的有 Intel 的 QPI/UPI、AMD 的 Infinity Fabric、NVIDIA 的 NVLink，这些通常是私有协议；面向高级封装的有 Intel 的 AIB、IEEE 的 MDIO、TSMC 的 LIPINCON 和 OCP 的 BoW 等，这些大多是开放协议。一些 IP 企业，如 Rambus、Kandou、Cadence 等，也提出了一些方案，而且主要是基于串行连接方式——选择串行方案，通常意味着相对较低的成本、较



远的传输距离，有利于吸引生态圈内更多（更弱势）的参与者。国内学界和部分企业也在试图建立自己的标准，争夺话语权，绝大多数处于草案甚至立项阶段。

不论是大厂，还是产业界的老面孔，亦或是学界，积极探索 Chiplet 技术带来了百花齐放百家争鸣的局面，也会带来资源浪费。湮没在历史长河中的标准，不计其数。

目前是 Chiplet 发展的早期阶段，主要是解决技术瓶颈和成本约束的问题。这个阶段内，Chiplet 考虑的主要是芯片的切分问题，譬如由大拆小、功能与制程的匹配等。应用这种思路的主要是服务器处理器为代表的“大芯片”，不论它们是来自老牌大厂，还是互联网新贵。

如果第一阶段可以称为“实现”，那么，Chiplet 第二阶段的目标则是“复用”。进入这个阶段的企业还不太多。其中的成功典型是 AMD，其核心 IP（CCD、IOD）都实现复用，可以满足不同产品线甚至跨代产品线的需要，有效摊薄设计投入，也降低了生产成本。另一个能称得上复用的例子是 Apple 的 M1 Max/Ultra、M2 Max/Ultra 这类产品。AWS Graviton3 的内存、PCIe 控制器可能在未来的产品中也会被复用，尚待观察。

第三阶段就是本章开头提到的愿景了，IP 硬化、芯粒商品化、货架化，不同厂商（而不是代工方）的芯片可以通用。这不仅需要包括 UCle、BoW 在内的多种标准完成竞合，出现若干主导性的标准，还需要整个产业界探索出新的设计、验证流程，明确生产中的责任归属，甚至在安全性方面也会有巨大的挑战。

国内产业界则将 Chiplet 视为“弯道超车”的机会。如果从第一阶段角度看，在国外大厂面临生产技术瓶颈的时候，国内部分互联网大厂、独角兽企业确实有机会通过 Chiplet 以相对合理的成本推出有竞争力的明星产品。但是，国内企业需要有能力和决心、有市场进行长期投资，让旗下产品持续迭代，产品矩阵羽翼丰满，才有可能进入第二阶段。至于第三阶段，要的不仅仅是脚踏实地发展的耐心，还要有大格局。



CHAPTER4

算力互连 由内及外，由小渐大

2023 新型算力中心调研报告

算力互连：由内及外，由小渐大

随着“东数西算”工程的推进，诸如“东数西渲”、“东数西训”等细分场景也逐渐被提起。

视频渲染和人工智能（Artificial Intelligence, AI）/ 机器学习（Machine Learning, ML）的训练任务，本质上都属于离线计算或批处理性质，完全可以在“东数西存”的基础上，即原始素材或历史数据传输到位于西部地区的数据中心之后，就地独立完成计算过程，中间极少与东部地区的数据中心交互，因此可以不受跨地域的时延影响。

换言之，“东数西渲”、“东数西训”的业务逻辑能够成立，是因为计算与存储仍是就近耦合的，不需要面对跨地域的“存算分离”挑战。

在服务器内部，CPU 与 GPU 存在着类似而又不同的关系。以目前火热的大模型为例，对计算性能和内存容量都有很高的要求，而 CPU 与 GPU 在这方面偏偏存在“错配”的现象：GPU 的（AI）算力明显高于 CPU，但是直属的内存（显存）容量基本不超过 100GB，与 CPU 动辄 TB 级的内存容量相比，相差一个数量级。

好在，CPU 与 GPU 之间的距离可以缩短，带宽可以提升。消除互连瓶颈之后，可以大量减少不必要的数据移动，提高 GPU 的利用率。

为 GPU 而生的 CPU

NVIDIA Grace CPU 的核心基于 Arm Neoverse V2，互连架构 SCF（Scalable Coherency Fabric，可扩展一致性结构）也可以看作是 Arm CMN-700 网络的定制版。但是在对外 I/O 的部分，NVIDIA Grace CPU 与其他 Arm 和 x86 服务器都有很大的不同，体现出英伟达做这款 CPU 的主要意图——为需要高速访问大内存的 GPU 服务。

内存方面，Grace CPU 有 16 个 LPDDR5X 内存控制器，这些内存控制器对应着 CPU 外面封装在一起的 8 个 LPDDR5X 芯片，裸容量 512GB，扣除 ECC 开销后，可用容量为 480GB。这样看来，有 1 个内存控制器及其对应的 LPDDR5X 内存 die 被用于 ECC。

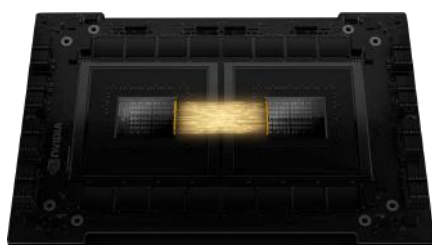
在英伟达的官方资料里，与 512GB 内存容量同时出现的内存带宽

参数是 546GB/s，而与 480GB（w/ ECC）一同出现的是（约）500GB/s，实际的内存带宽应该是 512GB/s 左右。

PCIe 控制器是一定要有的，Arm CPU 的惯例是有一部分 PCIe 通道会与 CCIX 复用，但这样的 CCIX 互连带宽太弱了，还不如英特尔专用于 CPU 间互连的 QPI/UPI，英伟达肯定是看不上。



△ NVIDIA Grace 的 I/O

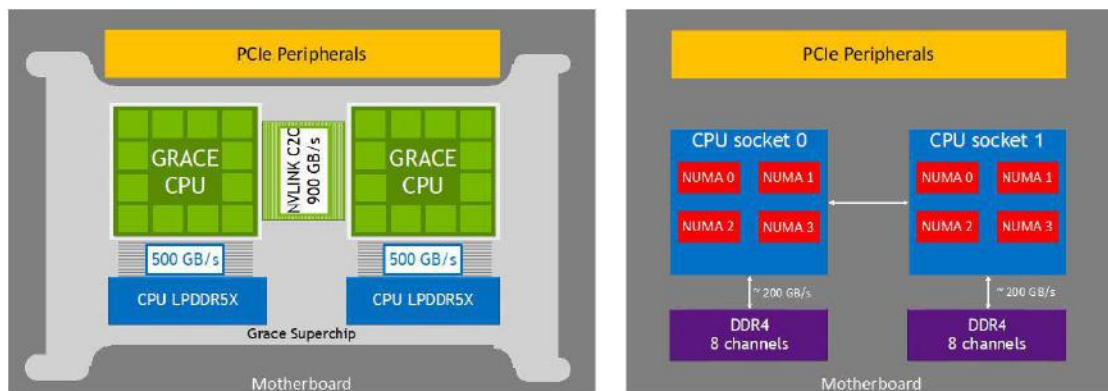


△ NVLink-C2C

Grace CPU 提供 68 个 PCIe 5.0 通道，其中有 2 个 x16 也可以用作 12 通道一致性 NVLink（coherent NVLINK，cNVLINK）。真正用于芯片（CPU/GPU）之间互连的，是与 cNVLINK/PCIe 隔“核”相望的 NVLink-C2C 接口，带宽高达 900GB/s。

NVLink-C2C，其中的 C2C 就是 chip to chip 之意。根据 NVIDIA 在 ISSCC2023 中的论文，NVLink-C2C 由 10 组连接（每组 9 对信号和 1 对时钟），共 200 个 I/O 构成，NRZ 调制，工作频率 20GHz，总带宽为 900GB/s。每个封装内的传输距离为 30mm，PCB 上的传输距离为 60mm。对于 NVIDIA Grace CPU 超级芯片，用 NVLink-C2C 连接两个 CPU，构成一个 144 核的模块；对于 NVIDIA Grace Hopper Superchip（超级芯片），那就是把 Grace CPU 和 Hopper GPU 互联。



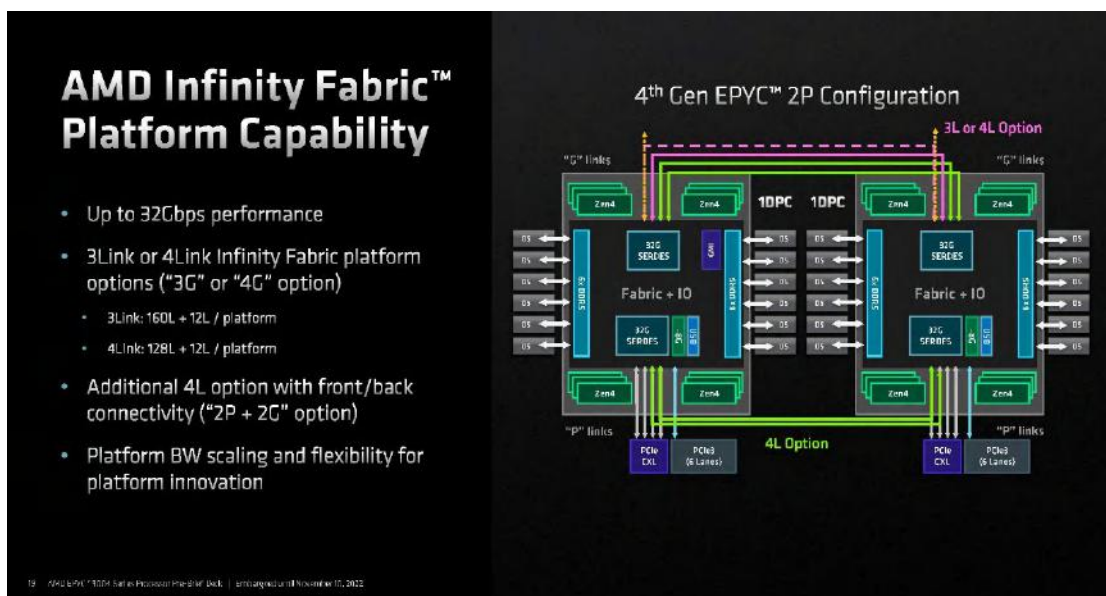


△ NVIDIA Grace 的处理器、内存互联带宽非常可观

NVLink-C2C 的带宽为 900GB/s，这是一个相当惊人的数据。作为参考：

Intel 代号 Sapphire Rapids 的第四代至强可扩展处理器包含 3 或 4 组 x24 UPI 2.0 (@16GT/s)，多路处理器间互联的总带宽接近 200GB/s；

AMD 第四代 EPYC 用于处理器内 CCD 与 IOD 互联的 GMI3 接口带宽为 36GB/s，CPU 间互联的 Infinity Fabric 相当于 16 通道 PCIe 5.0，带宽为 32GB/s。双路 EPYC 9004 之间可以选择使用 3 或 4 组 Infinity Fabric 互联，4 组的总带宽为 128GB/s。

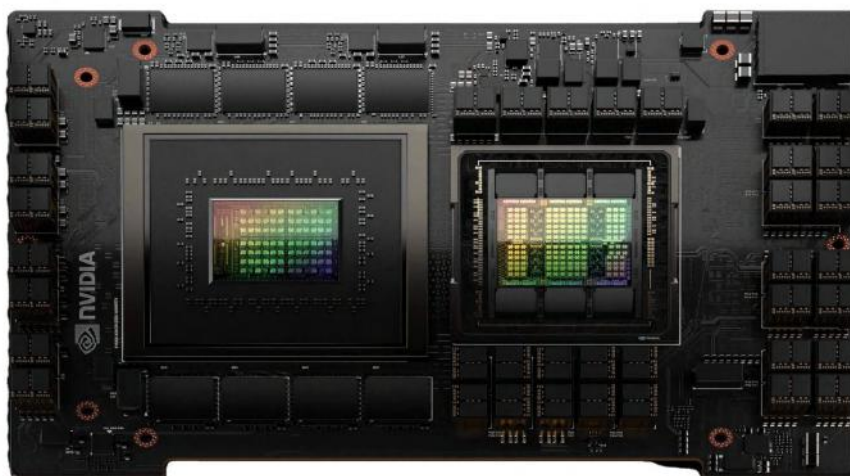


△ AMD Infinity Fabric

+ + + + + +
+ + + + + +

通过巨大的带宽，两颗 Grace CPU 被紧密联系在一起，其“紧密”程度远超传统的多路处理器系统，已足以匹敌现有的基于有机载板的多数 Chiplet 封装方案（2D 封装）。要超越这个带宽，需要硅中介层（2.5D 封装）的出马，例如 Apple M1 Ultra 的 Ultra Fusion 架构是利用硅中介层来连接两颗 M1 Max 芯粒。苹果宣称 Ultra Fusion 可同时传输超过 10,000 个信号，从而实现高达 2.5TB/s 低延迟处理器互联带宽。Intel 的 EMIB 也是 2.5D 封装的一种，其芯粒间的互联带宽也应当是 TB 级。

NVLink-C2C 另一个重要应用案例是 GH200 Grace Hopper 超级芯片，将一颗 Grace CPU 与一颗 Hopper GPU 互联。格蕾丝·霍波（Grace Hopper）是世界上第一位著名女程序员，“bug”术语的发明者。因此，NVIDIA 将这一代 CPU 和 GPU 分别命名为 Grace 和 Hopper，其实是有深意的，充分说明在前期规划中，二者便是强绑定的关系。

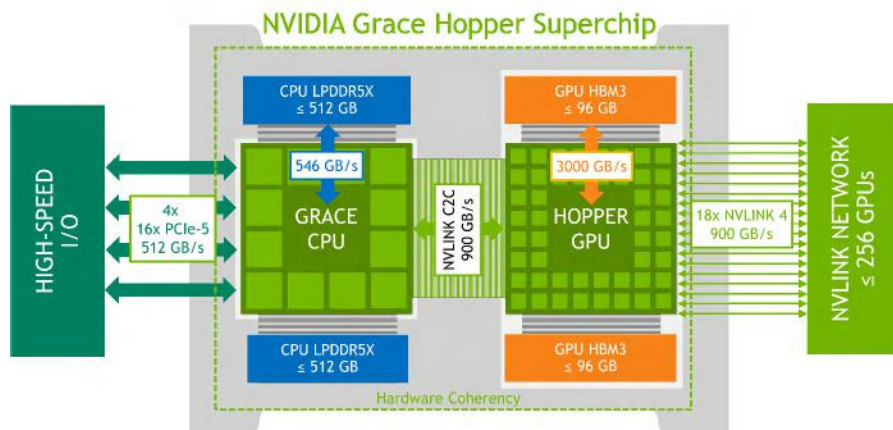


△ NVIDIA Grace Hopper 超级芯片

Feature	Description
Grace CPU cores (number)	Up to 72 cores
CPU LPDDR5X bandwidth (GB/s)	Up to 546 GB/s
GPU HBM3 bandwidth (GB/s)	3 TB/s
NVLink C2C bandwidth (GB/s)	900 GB/s total, 450 GB/s per direction
CPU LPDDR5X capacity (GB)	Up to 512 GB
GPU HBM3 capacity (GB)	Up to 96 GB
PCIe Gen 5 Lanes	64x

△ NVIDIA Grace Hopper 超级芯片主要规格

+ + + + + +
+ + + + + +



△ NVIDIA Grace Hopper 超级芯片的互联架构

“

简而言之，CPU 拥有的内存容量是 GPU 不能比的，带宽也还可以，但 GPU 到 CPU 之间的互连（PCIe）才是瓶颈所在。要改变这一点，亲自下场做 CPU 是最直接的。

考虑到 CPU + GPU 的异构组合，二者之间交换数据的效率（带宽、延迟）就是一个非常值得重视的问题，尤其是超大机器学习模型的时代——GPU 本地显存过于昂贵，容量实在捉襟见肘。

NVIDIA 为 Hopper GPU 配备了大容量的高速显存，为该系列的满配 6 组显存控制器全开，容量 96GB，显存位宽 6144bit，带宽达到 3TB/s。作为对比，独立的 GPU 卡 H100，根据不同版本，其显存配置有 80GB HBM2e（H100 PCIe）、80GB HBM3（H100 SXM），以及 GTC2023 上刚发布的双卡组合 H100 NVL 的 188GB HBM3。其中前二者均只启用了 5 组显存控制器。

Grace CPU 则搭载了 480GB 的 LPDDR5X 内存，带宽略超 500GB/s。这个内存配置有省电及空间紧凑的优势，但付出了可扩展性（容量）的代价。表面上看，Grace 的内存带宽与使用 DDR5 内存的竞品处于同一水平。譬如 AMD EPYC 9004 系列，12 通道 DDR5 4800 内存可以提供 461GB/s 的带宽，双路系统则可以实现超过 900GB/s 的内存带宽。但是，相比于内存带宽上的这点差异，GPU 与 CPU 之间的互连才是决定性的——典型的 x86 CPU，到 GPU 只能通过 PCIe，这个带宽比 NVLink-C2C 至少低一个数量级！

与 PCIe 相比，NVLink 还有缓存一致性的优势，CPU 与 GPU 之间、GPU 与 GPU 之间是可以互相寻址内存的。通过 NVLink-C2C，Hopper GPU 可以顺畅地访问 CPU 内存，这不仅是 H100 PCIe 无法企及的，就连 H100 SXM 都会羡慕——以 NVIDIA HGX 4GPU 为基准，Grace



Hopper 超级芯片中每 GPU 可分配的带宽为 3.5 倍。另外，高带宽的直接寻址还可以转化为容量优势：Grace Hopper 超级芯片中的 GPU 可以寻址 576GB (480GB+96GB) 本地内存。更高的内存带宽、更低的延迟有利于超大模型的机器学习等应用。

简而言之，CPU 拥有的内存容量是 GPU 不能比的，带宽也还可以，但 GPU 到 CPU 之间的互连（PCIe）才是瓶颈所在。要改变这一点，亲自下场做 CPU 是最直接的。

NVLink-C2C 的带宽足以匹配（CPU 的）内存，访问内存的友好度也超过 PCIe，都是 GH200 Grace Hopper 超级芯片相对 x86+GPU 方案的核心优势。NVLink-C2C 的另一个亮点是能效比，英伟达宣称 NVLink-C2C 每传输 1 比特数据仅消耗 1.3 皮焦耳能量，大约是 PCIe 5.0 的五分之一，再考虑速率，那就有 25 倍的能效差异了。这种比较当然不够公平，毕竟 PCIe 是板间的通讯，传输距离有本质的区别。但这个数据也有助于理解 NVLink-C2C 相对 NVLink 的能效差异，后者大概参考 PCIe 的量级来看即可。在能效方面，传输距离和封装方式与 NVLink-C2C 类似的接口总线是 AMD 用于 EPYC 的 Infinity Fabric，大概是 1.5pJ/b。至于 2.5D、3D Chiplet 使用的接口，如 UCIe、EMIB 等的能耗还要再低一个数量级，大致的情况可以参考下面的表格。

互联接口	能耗
Infinity Fabric	~1.5 pJ/b
NVLink-C2C	1.3 pJ/b
UCIe 高级封装	0.25 pJ/b
UCIe 标准封装	0.5 pJ/b
TSMC CoWoS	0.56 pJ/b
Foveros	0.2 pJ/b
EMIB	0.3 pJ/b

NVLink 最初是为满足 GPU 之间高速交换数据而生的，在 NVSwitch 的帮助下，可以把服务器内部的多个 GPU 连为一体，获得容量成倍增加的显存池。



NVLink 之 GPU 互连

NVLink 的目标是突破 PCIe 接口的带宽瓶颈，提高 GPU 之间交换数据的效率。2016 年发布的 P100 搭载了第一代 NVLink，提供 160 GB/s 的带宽，相当于当时 PCIe 3.0 x16 带宽的 5 倍。V100 搭载的 NVLink2 将带宽提升到了 300 GB/s，接近 PCIe 4.0 x16 的 5 倍。A100 搭载了 NVLink3，带宽为 600 GB/s。

H100 搭载的则是 NVLink4。相对 NVLink3，NVLink4 不仅增加了链接数量，内涵也有比较重大的变化。NVLink3 中，每个链接通道使用 4 个 50Gb/s 差分对，每通道单向 25GB/s，双向 50GB/s。A100 使用 12 个 NVLink3 链接，总共构成了 600GB/s 的带宽。NVLink4 则改为每链接通道使用 2 个 100Gb/s 差分对，每通道双向带宽依旧为 50GB/s，但线路数量减少了。在 H100 上可以提供 18 个 NVLink4 链接，总共 900GB/s 带宽。

NVIDIA 的 GPU 大多提供了 NVLink 接口，其中 PCIe 版本可以通过 NVLink Bridge 互联，但规模有限。

更大规模的互联还是得通过主板 / 基板上的 NVLink 进行组织，与之对应的 GPU 有 NVIDIA 私有的规格 SXM。SXM 规格 NVIDIA GPU 主要应用于数据中心场景，其基本形态为长方形，正面看不到金手指，属于一种 mezzanine 卡，采用类似 CPU 插座的水平安装方式“扣”在主板上，通常是 4-GPU 或 8-GPU 一组。其中 4-GPU 的系统可以不通过 NVSwitch 即可彼此直连，而 8-GPU 系统需要使用 NVSwitch。

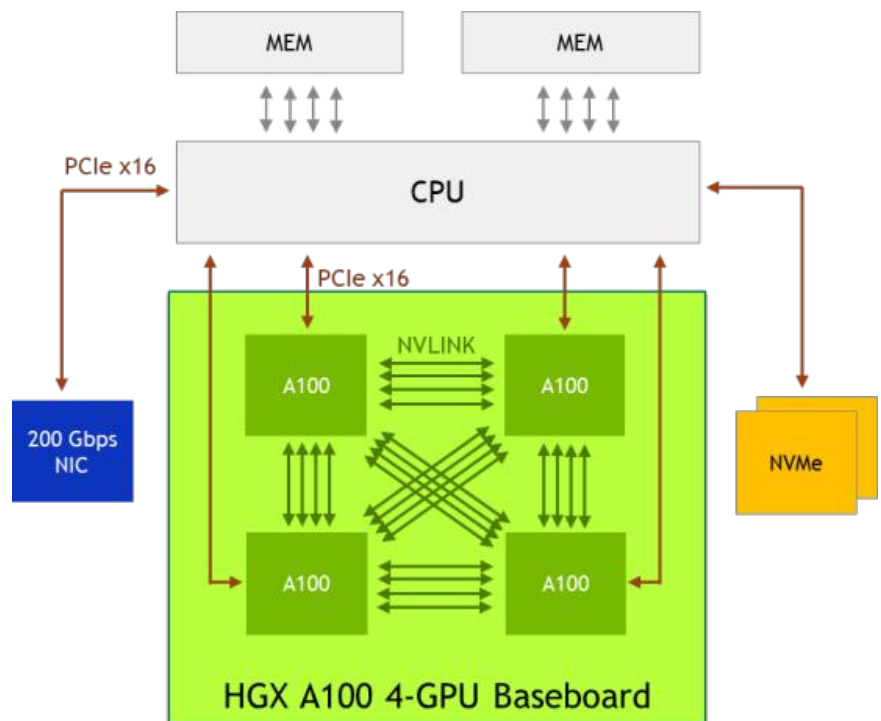


△ NVIDIA V100 SXM2 版本正反面，提供 NVLink2 连接

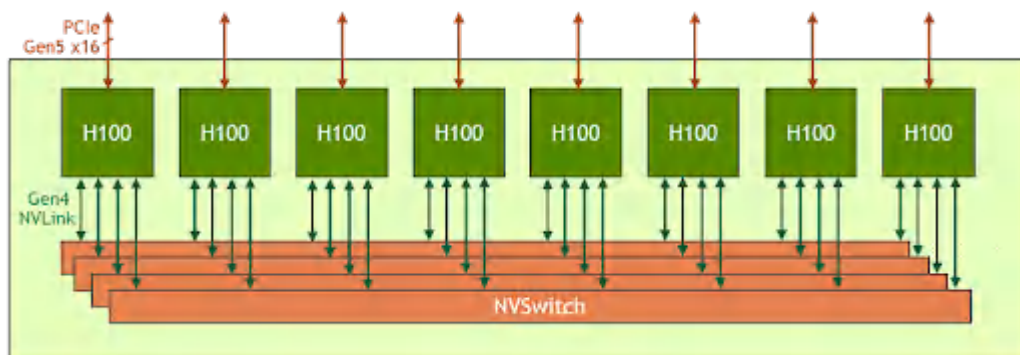
+ + + + +
+ + + + +



△ NVIDIA HGX A100 8-GPU 系统。此图完整展现了主要结构、安装形式和散热。其中右侧的两块 A100 SXM 没有安装散热器。右上角未覆盖散热器的细长矩形芯片即为 NVSwitch



△ NVIDIA HGX A100 4-GPU 系统的组织结构。每个 A100 的 12 条 NVLink 被均分为 3 组，分别与其他 3 个 A100 直联



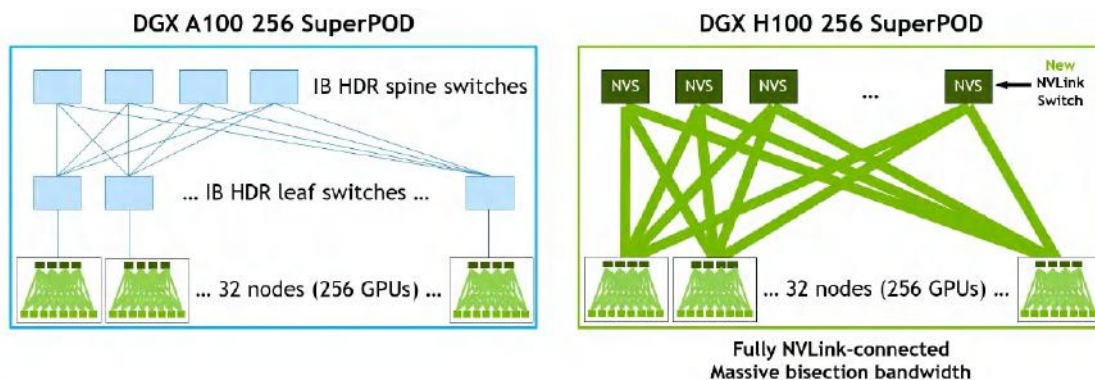
△ NVIDIA HGX H100 8-GPU 系统的组织结构。每个 H100 的 18 条 NVLink 被分为 4 组，分别与 4 个 NVSwitch 互联。

经过多代发展之后，NVLink 日趋成熟，已经开始应用于 GPU 服务器之间的互连，进一步扩大 GPU（及其显存的）集群规模。

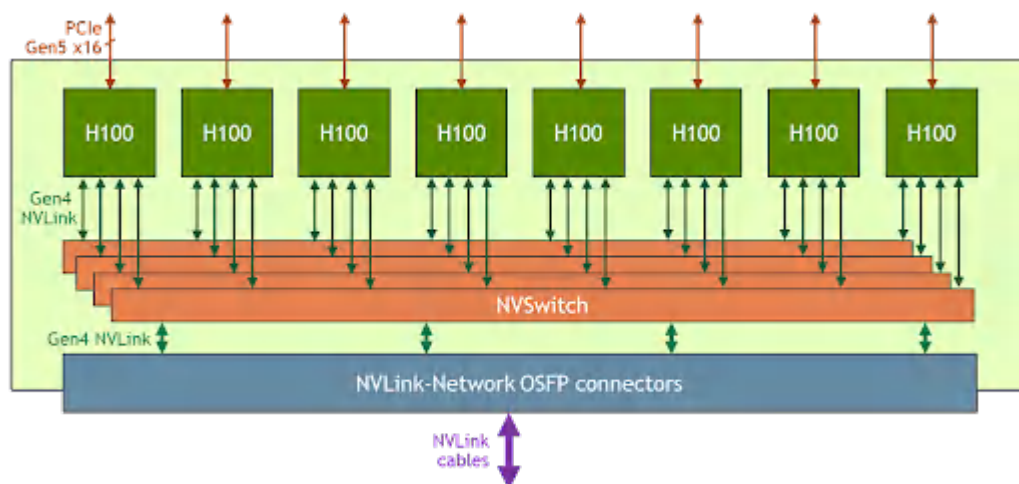
NVLink 组网超级集群

在 2023 年 5 月底召开的 COMPUTEX 上，英伟达公布了 256 个 Grace Hopper 超级芯片组成的集群，GPU 内存总量达 144TB。以 GPT 为代表的大语言模型（Large Language Model, LLM）对显存的容量需求极其迫切，巨量显存将迎合大模型的发展趋势。那么，这个前所未见的容量是如何达成的？

NVLink4 Networks 是一个重大创新，让 NVLink 可以扩展到节点之外。通过 DGX A100 和 DGX H100 各自构建 256-GPU SuperPOD 的架构图，可以直观感受到 NVLink4 Networks 的特点。在 DGX A100 SuperPOD 中，每个 DGX 节点的 8-GPU 是通过 NVLink3 互联的，而 32 个节点则需要通过 HDR InfiniBand 200G 网卡和 Quantum QM8790 交换机互联。在 DGX H100 SuperPOD 中，节点内部是 NVLink4 互联 8-GPU，节点之间通过 NVLink4 Network 互联，各节点接入称为 NVLink Switch 的设备。

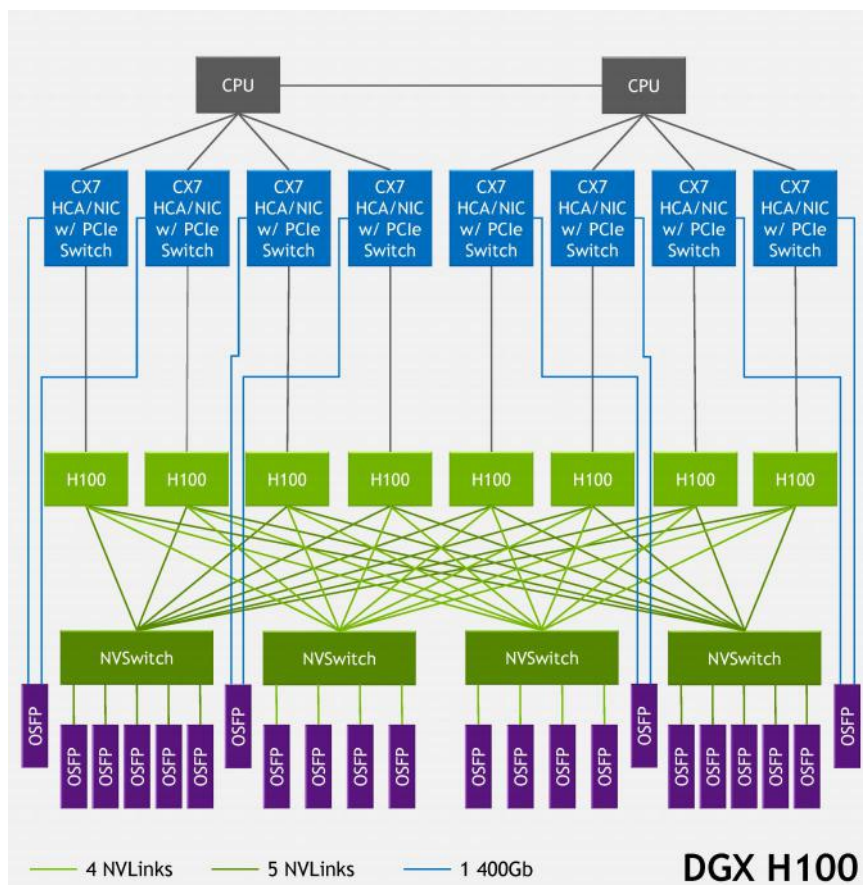


△ DGX A100 和 DGX H100 256 SuperPOD 架构



△ HGX H100 8-GPU 的 NVLink-Network 连接

在 NVIDIA 提供的架构信息中，NVLink Network 支持了 OSFP (Octal Small Form Factor Pluggable) 光口。这也符合 NVIDIA 宣称的线缆长度从 5 米增加至 20 米的说法。DGX H100 SuperPOD 使用的 NVLink Switch 规格为：端口数量 128 个，32 个 OSFP 笼 (cage)，总带宽 6.4TB/s。



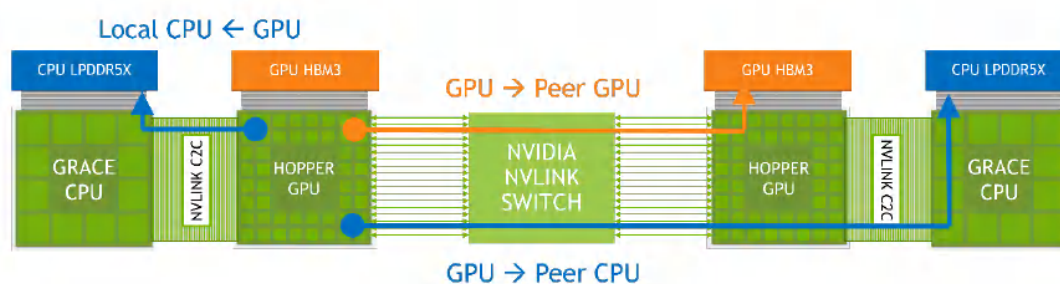
△ DGX H100 SuperPOD 节点内部的网络架构

每个 8-GPU 节点内部有 4 个 NVSwitch，对于 DGX H100 SuperPOD，每个 NVSwitch 都通过 4 或 5 条 NVLink 对外连接。每条 NVLink 是 50GB/s 带宽，对应一个 OSFP 则相当于 400Gb/s，是非常成熟的。每个节点总共需要连接 18 个 OSFP 接口，32 个节点共需要 576 个连接，对应 18 台 NVLink Switch。

DGX H100 也可以（仅）通过 InfiniBand 互联，参考 DGX H100 BasePOD 的配置，其中的 DGX H100 系统配置了 8 个 H100、双路 56 核第四代英特尔至强可扩展处理器、2TB DDR5 内存，搭配了 4 块 ConnectX-7 网卡——其中 3 块双端口卡为管理和存储服务，还有一块 4 OSFP 口的用于计算网络。

回到 Grace Hopper 超级芯片，NVIDIA 提供了一个简化的示意图，其中的 Hopper GPU 上的 18 条 NVLink4 与 NVLink Switch 相连。NVLink Switch 连接了“两组” Grace Hopper 超级芯片。任何 GPU 都可以通过 NVLink-C2C 和 NVLink Switch 访问网络内其他 CPU、GPU 的内存。

NVLink4 Networks 的规模是 256 个 GPU——注意，是 GPU，而不是超级芯片，因为 NVLink4 连接是通过 H100 GPU 提供的。对于 Grace Hopper 超级芯片，这个集群的内存上限就是：（480GB 内存 + 96GB 显存）× 256 节点 = 147456GB，即 144TB 的规模。假如 NVIDIA 推出了 GTC2022 中提到的 Grace + 2Hopper，那么，按照 NVLink Switch 的接入能力，那就是 128 个 Grace 和 256 个 Hopper，整个集群的内存容量将下降至约 80TB 量级。



△ Grace Hopper 超级芯片之间的互联

在 COMPUTEX 2023 期间，NVIDIA 宣布 Grace Hopper 超级芯片已经量产，并发布了基于此的 DGX GH200 超级计算机。NVIDIA DGX GH200 使用了 256 组 Grace Hopper 超级芯片，以及 NVLink 互联，整个集群提供高达 144TB 的可共享的“显存”，以满足超大模型的需求。

先列几个数字来感受一下，NVIDIA 以一己之力打造的 E 级超算系统。

算力：1 exa Flops (FP8)

光纤总长度：150 英里

风扇数量：2112 个 (60mm)

风量：7 万立方英尺 / 分钟 (CFM)

重量：4 万磅

显存：144 TB

NVLink 带宽：230 TB/s

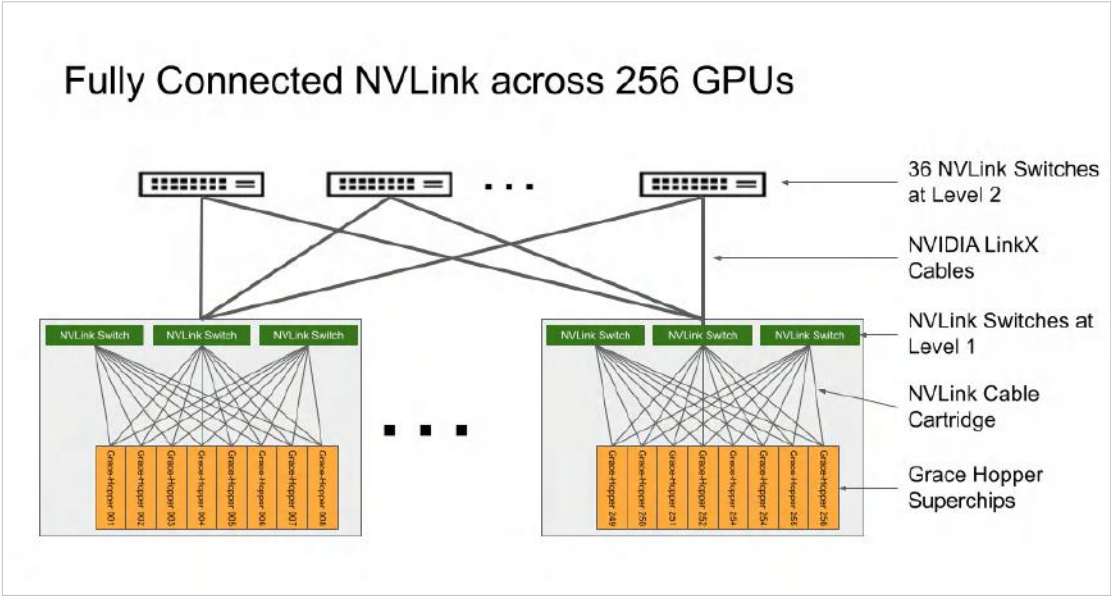
+ + + + + +
+ + + + + +

+ + + + + +
+ + + + + +

从 150 英里的光纤长度，我们就可以感受其网络复杂度。这个集群的整体网络资源如下：

Networking	256x OSFP single-port NVIDIA ConnectX7 VPI with 400Gb/s InfiniBand
	256x dual-port NVIDIA BlueField3 VPI with 200Gb/s InfiniBand and Ethernet
	24x NVIDIA Quantum-2 QM9700 InfiniBand Switches
	20x NVIDIA Spectrum SN2201 Ethernet Switches
	22x NVIDIA Spectrum SN3700 Ethernet Switches
NVIDIA NVLink Switch System	96x L1 NVIDIA NVLink Switches
	36x L2 NVIDIA NVLink Switches

由于 Grace Hopper 芯片上只有 CPU 和 GPU 各一，GPU 数量远少于 DGX H100，同样达到 256 个 GPU 所需的节点数大为增加，导致 NVLink Network 的架构复杂很多：



△ NVIDIA DGX GH200 集群内的 NVLink 网络架构



GH200 的每个节点有 3 组 NVLink 对外连接，每个 NVLink Switch 连接 8 个节点。256 个节点总共分为 32 组，每组 8 个节点搭配 3 台 L1 NVLink Switch，共需要使用 96 台交换机。这 32 组网络还要通过 36 台 L2 NVLink Switch 组织在一起。

相比 DGX H100 SuperPOD，GH200 的节点数量大幅增加，NVLink Network 的复杂度明显提高了。二者的对比如下：

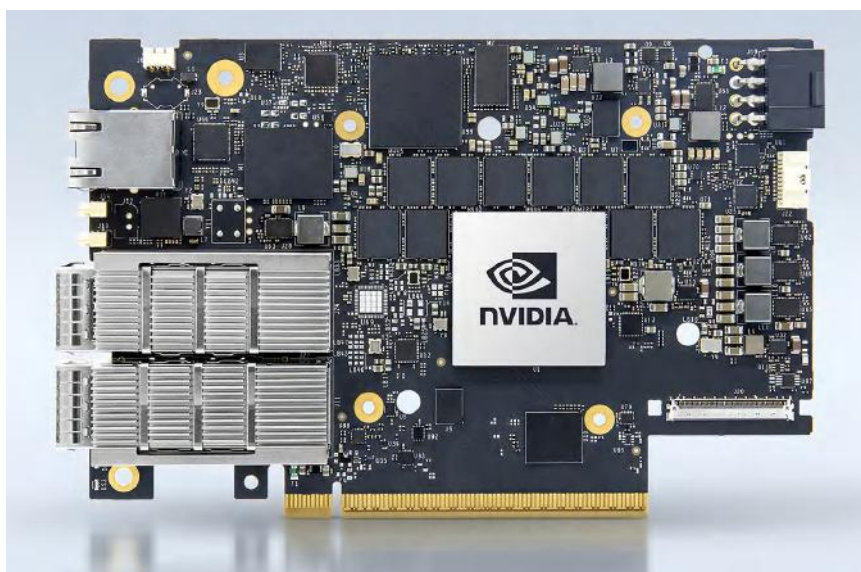
	DGX H100 SuperPOD	DGX GH200	差异
节点数量	32	256	8x
交换机数量	18	96+36	7.3x
节点 NVLink 出口（理论）	576	4608	8x





InfiniBand 扩大规模

如果需要更大规模（超过 256 个 GPU）的集群，那就得 InfiniBand 交换机上场了。对于 Grace Hopper 超级芯片的大规模集群，NVIDIA 的建议是采用 Quantum-2 交换机组网，提供 NDR 400 Gb/s 端口；每个节点配置 BlueField-3 DPU（已经集成了 ConnectX-7），每 DPU 都提供 2 个 400Gb/s 端口，总带宽就是 100GB/s。理论上，使用以太网连接也能有类似的带宽水平，但既然 NVIDIA 收购了 Mellanox，偏爱 InfiniBand 完全可以理解。



△ NVIDIA BlueField-3 DPU

基于 InfiniBand NDR400 组织的 Grace Hopper 超级芯片集群有两种架构。一种是完全采用 InfiniBand 连接，另一种是混合配置 NVLink Switch 和 InfiniBand 连接。二者的共同点是：各节点均通过双端口（共 800Gbps）连接 InfiniBand 交换机，DPU 占用 x32 的 PCIe 5.0，由 Grace CPU 提供 PCIe 连接。二者的区别是：后者每个节点还通过 GPU 接入 NVLink Switch 连接，构成若干 NVLink 子集群。

很显然，混合配置 InfiniBand 和 NVLink Switch 的方案性能更好，毕竟部分 GPU 之间拥有更大的带宽，以及对内存的原子操作。譬如 NVIDIA 计划打造超级计算机 Helios，将由 4 个 DGX GH200 系统组成——通过 Quantum-2 InfiniBand 400 Gb/s 网络组织起来。



混合配置 InfiniBand 和 NVLink Switch 的方案性能更好，毕竟部分 GPU 之间拥有更大的带宽，以及对内存的原子操作。譬如 NVIDIA 计划打造超级计算机 Helios，将由 4 个 DGX GH200 系统组成——通过 Quantum-2 InfiniBand 400 Gb/s 网络组织起来。

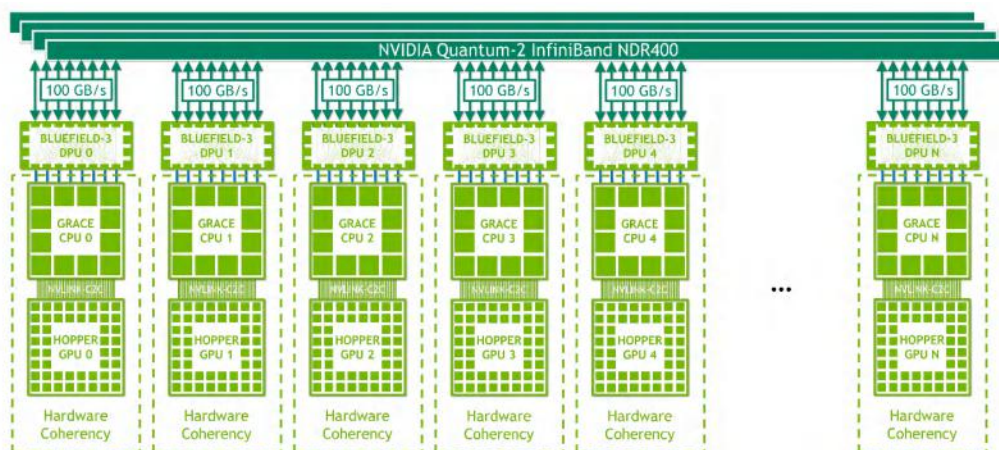


Figure 3. NVIDIA HGX Grace Hopper Superchip system with InfiniBand networking for scale-out ML and HPC workloads

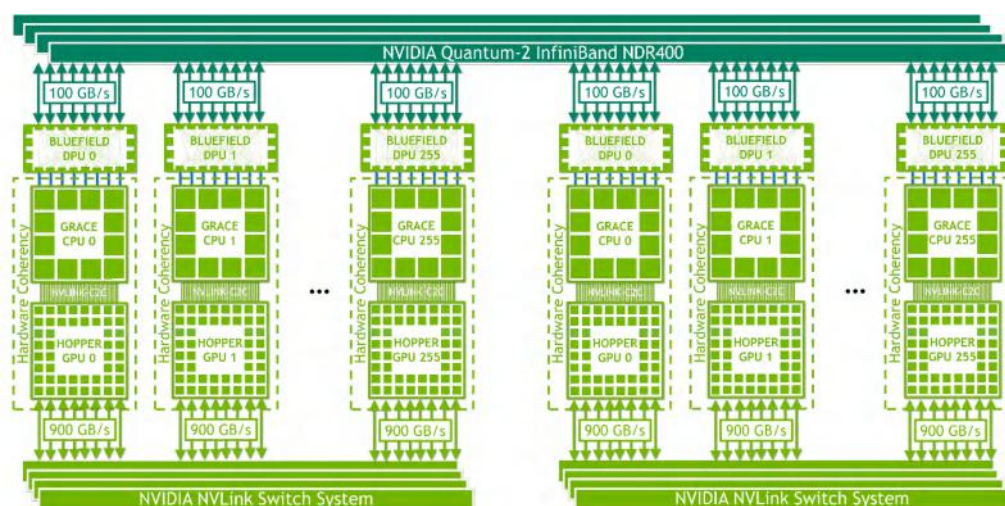
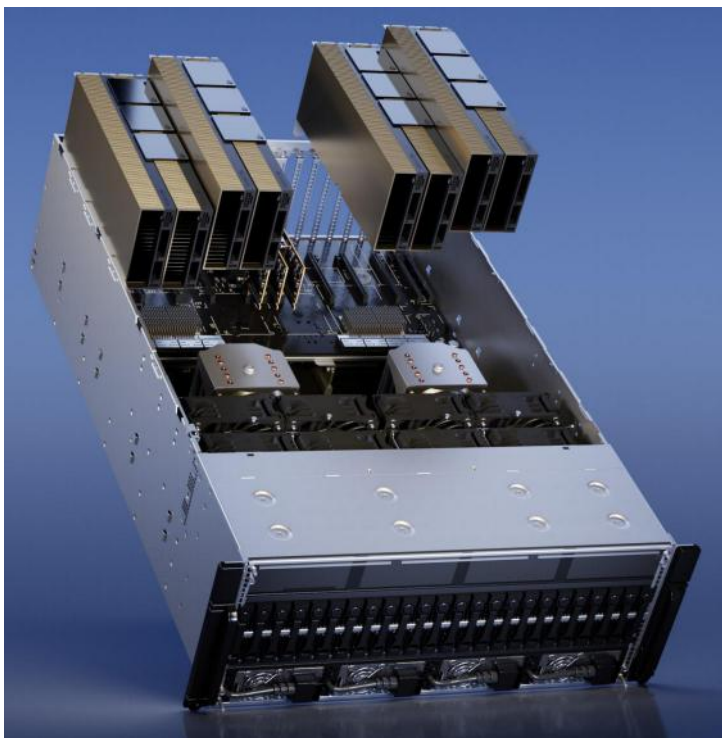


Figure 4. NVIDIA HGX Grace Hopper Superchip System with NVLink Switch System for strong-scaling giant ML and HPC

+ + + + + +
+ + + + + +

从 H100 NVL 的角度再看 NVLink

在 GTC2023 上，英伟达发布了面向大语言模型部署的 NVIDIA H100 NVL，与 H100 家族的另外两个版本单卡（SXM、PCIe）相比，它有两大特别之处：



△ NVIDIA H100 NVL

首先，H100 NVL 相当于两张 H100 PCIe 通过 3 块 NVLink bridge 连接在一起；

其次，每张卡有接近足额的 94GB 显存，连 H100 SXM5 都没有这样的待遇。

按照英伟达官方文档的介绍，H100 PCIe 的双插槽 NVLink 桥接沿用自上一代的 A100 PCIe，因此 H100 NVL 的 NVLink 互连带宽为 600GB/s，仍有通过 PCIe 5.0 互连（128GB/s）的 4 倍以上。

H100 NVL 由两张 H100 PCIe 卡拼合的颗粒度和产品形态，适合推理应用，经高速 NVLink 连为一体的显存容量高达 188GB，以满足大语言模型的（推理）需求。

如果把 H100 NVL 的 NVLink 互连视为缩水版的 NVLink-C2C，应该有助于对 NVLink 通过算力单元互连加速内存访问的理解。事实上，与 H100 NVL 一同发布的还有 3 款推理卡，其中就有面向推荐模型的 NVIDIA Grace Hopper。



CHAPTER5

绿色低碳和可持续发展

2023 新型算力中心调研报告

520万+

2021 年底，数据中心机架总规模

55%+

平均上架率

268
百亿亿次每秒
(EFLOPS)

2022 年智能算力规模

52.3%

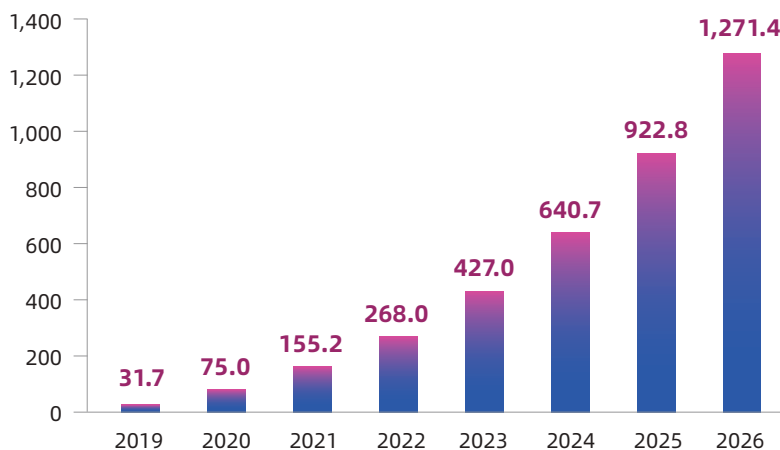
未来 5 年我国智能算力
规模的年复合增长率

目前，我国算力基础设施迎来了多样化发展的繁荣期，结合不同应用场景需求的异构化布局将加快推进。在超级算力方面，2023 年 4 月 17 日，国家超算互联网联合体成立，北京、贵州、上海、惠州、天津等地算力基础设施计划持续落地，算力建设持续提速。传统的高性能计算，更偏向于天气预报、大型工程设计和基础科学研究等应用场景，而未来，超算互联网是由各大超算中心提供算力，以软件和服务等形式提供给科研机构、公司企业。

在通用算力方面，工信部数据显示，截止 2021 年底，我国在用数据中心机架总规模超过 520 万标准机架，平均上架率超过 55%。在智能算力方面，根据《智能计算中心创新发展指南》，2022 年我国智能算力规模快速增长，达到 268 百亿亿次每秒（EFLOPS），超过通用算力规模，预计未来 5 年中国智能算力规模的年复合增长率将达 52.3%。

中国智能算力发展趋势

百亿亿次浮点运算 / 秒 (EFLOPS)

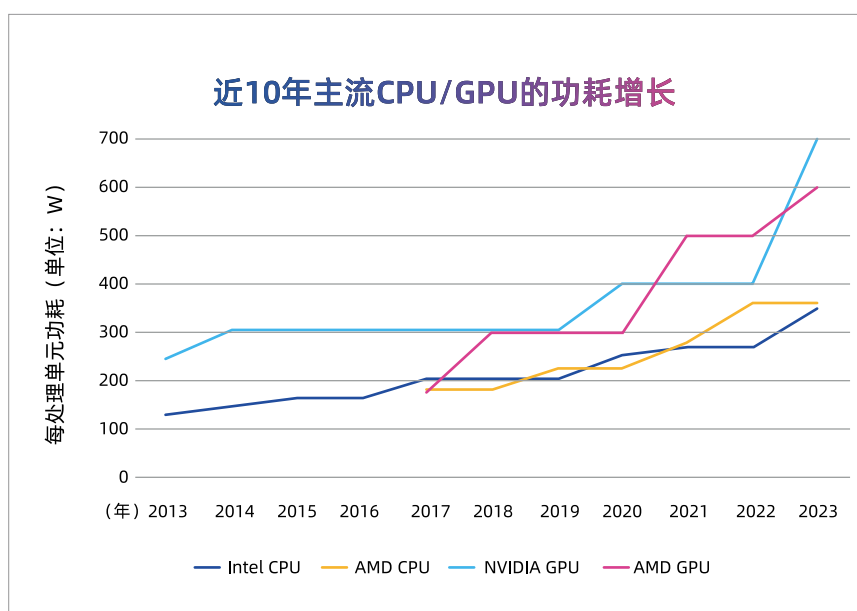


数据来源：《智能计算中心创新发展指南》

对于算力中心而言，算力规模持续增长，随之而来的是散热压力和节能挑战。目前，作为服务器关键部件的 CPU/GPU，随着性能提升功耗增加非常显著。CPU 方面，第四代英特尔至强可扩展处理器的核心数最多可达 60 个，比代号 Ice Lake(-SP) 的第三代至强可扩展处理器高出 50%。相应的，公开款的 TDP 指标上限，也从 270 瓦 (W) 一跃而至 350 瓦。AMD EPYC 9004 系列处理器，最大功率可达 400W。



GPU 方面，2022 年英伟达于 GTC 大会上发布针对数据中心的新一代 Hopper 架构的 GPU 芯片单颗功耗达到 700 瓦，挑战传统风冷系统散热的能力边界。相比于传统服务器，AI 服务器的功耗更高，随着 AI 大模型与训练需求的持续增长，AI 服务器的市场规模将会继续扩大。根据 IDC 数据，2022 年全球 AI 服务器市场规模达 202 亿美元，同比增长 29.8%，占服务器市场规模的比例为 16.4%。



数据整理：益企研究院



从能源效率（能效）来看，芯片功耗提升，数据中心功率密度增高，产生更多热量，需要部署更多的空调控制机房温度，空调本身的用电也会上升，使数据中心能源效率变低，PUE 居高不下。

核心器件功耗的持续攀升给数据中心带来散热问题和能源效率挑战。传统的风冷主要依靠的就是散热面积和风量，在服务器内部的有限空间内，散热面积难以扩展，需要更大的风量，意味着提高风扇转速，不仅让风扇的功耗上升，同时风扇产生的震动和噪音也会严重影响机械硬盘（HDD）的性能。

从能源效率（能效）来看，芯片功耗提升，数据中心功率密度增高，产生更多热量，需要部署更多的空调控制机房温度，空调本身的用电也会上升，使数据中心能源效率变低，PUE 居高不下。

提高服务器的能效有助于节能。益企研究院出品的《2018 年中国超大规模云数据中心考察报告》指出，在数据中心层面，更重要的是将 IT 和基础设施作为一个整体考虑，提升数据中心整体的能效，达到进一步降低数据中心 PUE 的目的。

在国家政策的指引下，传统数据中心加快向高算力、高能效、低功耗，更



提高服务器的能效有助于节能。益企研究院出品的《2018 年中国超大规模云数据中心考察报告》指出，在数据中心层面，更重要的是将 IT 和基础设施作为一个整体考虑，提升数据中心整体的能效，达到进一步降低数据中心 PUE 的目的。



绿色特征的新型数据中心演进。

- 2019 年：工信部、国管局和国家能源局发布《关于加强绿色数据中心建设的指导意见》中提到 2022 年，PUE 达到 1.4 以下，改造使电能使用效率值不高于 1.8；
- 2021 年：工信部发布《新型数据中心三年行动计划（2021-2023）》中提到 2021 年底，新建数据中心 PUE 降低到 1.35 以下，到 2023 年底降低到 1.3 以下，严寒和寒冷地区力争降低到 1.25 以下；
- 2021 年：工信部、国管局和国家能源局发布《贯彻落实碳达峰碳中和目标要求推动数据中心和 5G 等新型基础设施绿色高质量发展实施方案》中提到 2025 年，全国新建数据中心 PUE 降到 1.3 以下，国家枢纽节点进一步降到 1.25 以下；
- 2022 年：工信部、发改委、财政部等六部门联合发布《工业能效提升行动计划》中提到 2025 年，新建大型、超大型数据中心 PUE 优于 1.3。

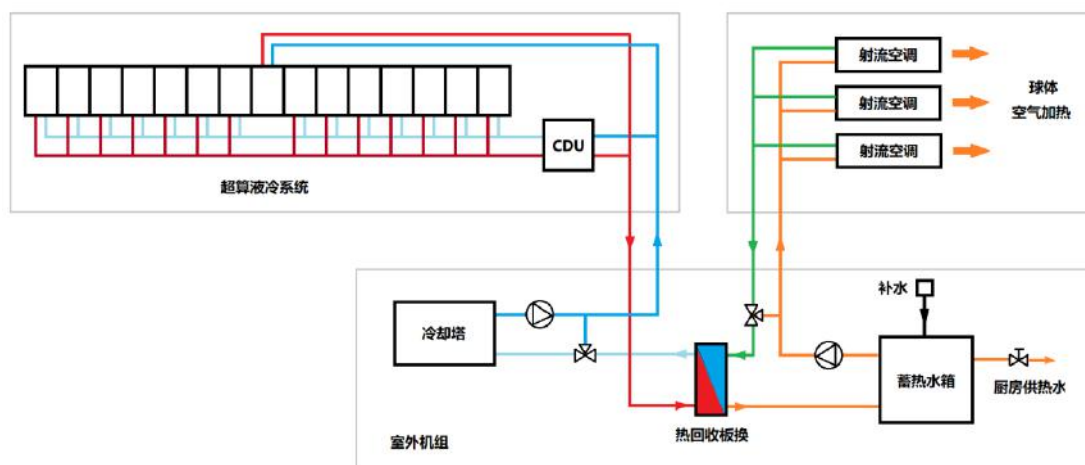
通过上述国家政策的出台可以看到对数据中心 PUE 指标控制更加严格，很多地区要求 PUE 做到 1.3 及 1.25，甚至 1.2 以下。而在应对新一轮低碳技术带来整体数据中心的技术变革中，液冷技术成为降低 PUE 的有效方式之一。

液冷的优势在于，单位体积的液体带走热量的能力通常比空气强得多，可以用较缓慢的流速冷却更高发热量的部件，而且工作温度也可以相对高一些。这就意味着液冷即使在气温较高的地区也可以更多的利用自然冷源，减少对电能的使用，具有更好的节能效果。

液冷数据中心可以提供更高温度的余热，充分利用这些余热可以实现供暖、提供卫生热水等等，可以有效减少供热设备能耗，大大降低了建筑和园区碳排放。比如上海交通大学计算中心的“思源一号”，除了能提供非常强大算力之外，还是国内唯一采用了热回收技术的超算中心，采用温水冷却技术，回收超算产生的热量。冷却水经 CDU（冷液分配单元）后流入热回收板式换热器，与球体大厅、地下室、实验室的相关空调系统回水热交换后进入蓄热水箱，一部分供给厨房生活热水系统，一部分供给球体大厅、地下室、实验室的相关空调系统。

通过余热回收替代原有消耗的电力包括燃气能源等，每年能够实现多达 950 吨、约 10% 比例的额外碳排放补偿。

+ + + + + +
+ + + + + +



△ 思源一号热回收原理图

1.1^{PUE}
9216 节点

联想 NeXtScale System

2,897,000
万亿次 (Gflops)

峰值运算速度

90.95%

整机效能

液冷应用 高性能计算中心跨越功耗墙

人类对宇宙探索的好奇心与对问题规模和精度的追求，决定了高性能计算能力的需求持续增长，而随着运算速度的不断改善，高性能计算中心成为液冷技术的早期用户，毕竟对于超级计算机这样的庞然大物来说，能耗是非常棘手的问题，超级计算机的耗电量又很大。在跨越功耗墙的进程中，早些年，美国国家安全局、美国空军、CGG、ORANGE、VIENNA 科学计算集群、日本东京工业大学就使用了 Green Revolution Cooling (GRC) 的浸没式液冷技术，美国 AFRL、ERDL、法国 TULAL、欧洲 AWE 等使用了 SGI 的液冷服务器。而在水冷技术层面，我们曾在 2016 年全球超算大会 (2016 ISC) 期间参观位于德国莱布尼茨实验室的 SuperMUC，号称首个采用温水水冷技术的 HPC 集群，联想 NeXtScale System 在该实验室部署了 9216 节点，峰值运算速度 2,897,000 万亿次 (Gflops)，整机效能高达 90.95%，PUE 低至 1.1。

在中国，神威·太湖之光全方位的绿色节能也是一大突破，采用液冷技术，功耗远低于早期的其他超算中心。2018 年 1 月 3 日北京大学高性能计算校级公共平台正式揭牌启用，“未名一号”、“未名教学一号”和“未名生科一号”等多套集群陆续投入运行，主要是面向全校提供数学、深度学习、大气海洋环境、新能源新材料、天文地球物理、生物医药健康等领域提供高性能科学与工程计算服务，作为国内第一个温水水冷的大规模超算集群，计算峰值达 3.65PFLOPS，存储容量 14PB，节能效果显著，LINPACK 效率达到 92.6%，PUE 值达到 1.1。

而在浸没式液冷技术的应用上，据公开资料显示，华中科技大学成为了中国首个成功实现商业化应用的全浸没液冷高性能计算平台和数据中心。



△ 北京大学高性能计算中心



液冷实践 全栈数据中心理念落地

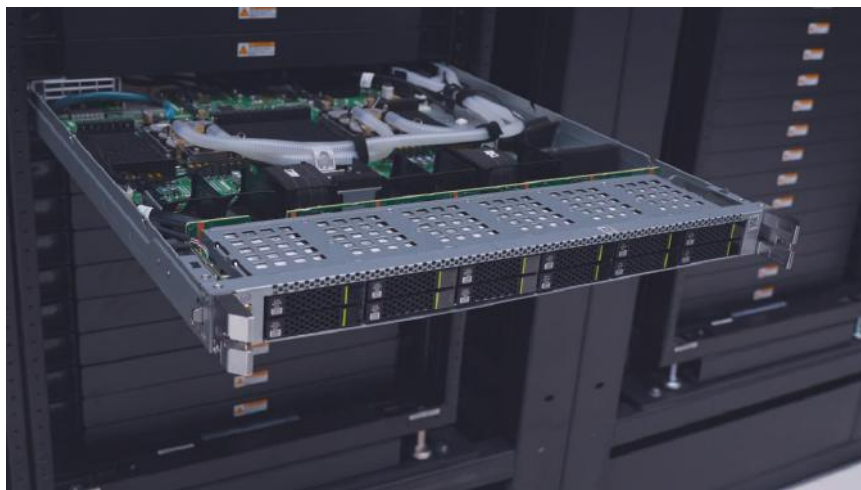
随着中国互联网、云计算的发展，超大规模数据中心应用的体量增加，数据中心的建设理念发生变化，最典型的是数据中心的建设与 IT 设备结合更紧密。大型互联网公司对数据中心行业的改变也是真正从机房建设，到 IT 设备的设计，再到上层的应用程序，将产品技术与应用贯穿了数据中心的全流程，将数据中心基础设施与 IT 基础架构作为整体优化。为了更快的满足业务需求，提高数据中心能效，大型互联网公司将液冷技术规模应用在数据中心，继而促进了价值链重构和产业生态演化。

为此，益企研究院提出并完善“全栈数据中心”理念。全栈数据中心是纵贯 IT 基础架构与数据中心基础设施，把芯片、计算、存储、网络等技术和数据中心风火水电作为一个整体看待；上层业务需求的变化会通过芯片、计算和存储等 IT 设备传导到网络架构层面，即数据中心作为基础设施也会相应的产生自上而下的变化。这也意味着服务器等 IT 设备的设计和液冷等先进技术的应用，以业务的视角实现应用与技术联动，以数据中心整体的视角将制冷、供电以及监控运维实现垂直整合。

从 2018 年始，数字中国万里行团队见证了液冷技术在云数据中心的应用，并在《2018 年中国超大规模云数据中心考察报告》中加以介绍。

常见的数据中心液冷方式主要包括喷淋式、冷板式和浸没式三种。

冷板式液冷相对成熟，虽然各家形态不同，但技术上差异不大。冷板式液冷是指采用液体作为传热工质在冷板内部流道流动，通过热传递对热源实现冷却的非接触液体冷却技术。通过对 CPU 和内存覆





盖冷板，液体直接带走这两个高发热部件的热量。液体在冷板内流动把 CPU 和内存的热量带走，自身温度达到 45℃，之后经过与数据中心冷却水交换后降低到 35℃ 返回，继续冷却。液体主要有不导电、不结垢的去离子水或不导电、不腐蚀的特殊液体两种。用户可根据自身需求进行选择，业界普遍认为前者更经济，而后者更安全。

冷板式液冷服务器对于目前的数据中心的架构影响不大，不需要对机柜的形态进行大幅度的改变，具有低噪音，高能效以及低总体拥有成本 (TCO) 的特点，可带来传统风冷数据中心所不具备的优势，使得耗能可以大幅度下降，同时又给 CPU 和内存提供了更好的工作环境和温度。

浸没式液冷总体方向比冷板式更进一步，给元器件提供更可靠和稳定的工作温度，并具有更高的能效。冷板式的服务器是的风冷和液冷混合，浸没式则是可以完全去除空调的全液冷的数据中心。

浸没式液冷把所有的 IT 设备所有器件浸泡在液体里。主要分为相变式液冷和单相浸没液冷。

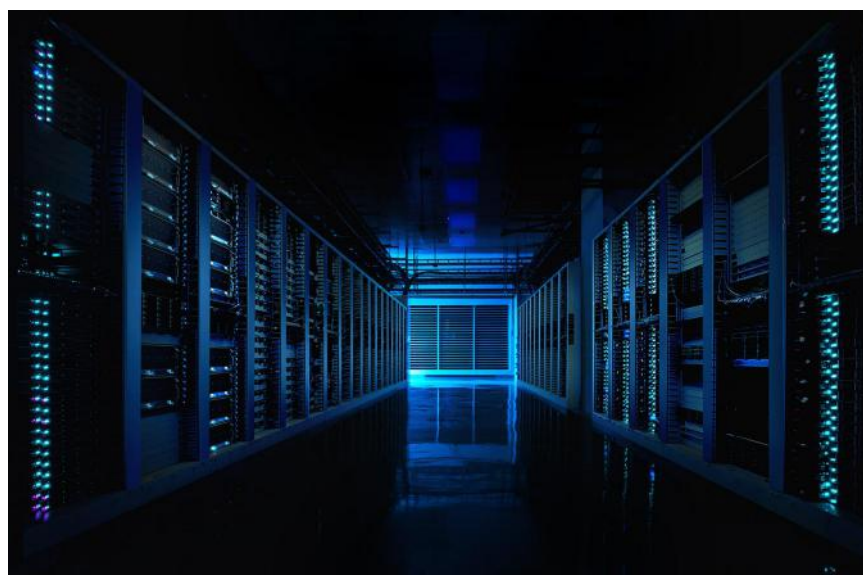


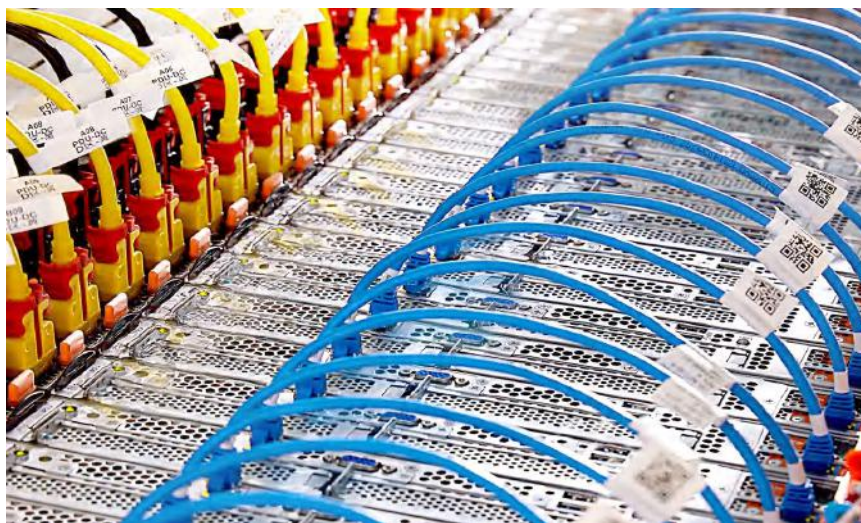
相变式液冷，采用沸点低、易挥发的液体作为冷媒，利用 CPU 等器件工作发热使冷媒沸腾带走热量，制冷剂蒸汽在换热器处冷凝，完成制冷循环，可以把冷却系统的能耗降到最低。如曙光研发的相变液冷方案，就可让数据中心实现全地域全年自然冷却。而从产品形态上来看，相变液冷产品还可分为缸式相变液冷方案，以及刀片式相变液冷技术方案。刀片式相变液冷方案，因为其具有更高的计算密度，更易维护性以及可按需灵活增减计算节点等优势。同时该方案对系统自动化供电、减压等有诸多技术要求，目前国内中科曙光实现了刀片式相变液冷方案的大规模部署。



△ 曙光研发的浸没式相变液冷系统

单相式浸没通过液体升温带走热量，不需要发生相变，在整个过程中就可以把换热设施和机柜实现分离，从而对换热系统进行一定的冗余设置就可实现在线维护。两种不同的设计方式也直接影响了维护方式。目前超算中心应用相变式的浸没液冷较多，单相式浸没液冷还更容易实现在线维护，适合通用型的云计算数据中心。





△ 阿里仁和数据中心单相式浸没液冷系统



过去几年，整机柜服务器的设计已经跳出机柜本身，以数据中心乃至整个基础设施的视角，与数据中心的风火水电基础设施紧密协同，同时也能够与上层的应用和业务结合。

液冷技术的推广应用，是全栈数据中心的最佳落地实践。举例来说，液冷就很适合通过整机柜（服务器）的形式交付。传统上在数据中心，机柜是基础设施团队（风火水电、场地）与 IT 业务部门的分界线。基础设施团队通常不会关注机柜里产品技术的演进（比如服务器产品）；IT 业务团队也很少了解基础设施的细节。互联网和云计算公司较多把机柜和服务器等 IT 设备做一个整体考虑。比如说阿里、腾讯、字节，服务器保有量都是百万台量级，在这样的规模下把服务器和机柜作为整体设计进行优化，哪怕效率提升 1% 都可以节省一大笔支出。而液冷技术天然适合整机柜交付模式，毕竟液冷更适合集中部署，需要突破服务器与整机柜界边界。

1、业务前置 模块化交付

过去几年，整机柜服务器的设计已经跳出机柜本身，以数据中心乃至整个基础设施的视角，与数据中心的风火水电基础设施紧密协同，同时也能够与上层的应用和业务结合。以京东云自研液冷整机柜服务器为例，基于业务的视角给应用端提供各种各样的可能性。京东业务涉及零售、金融、物流等多领域的服务，所以在整机柜设计时聚焦承载高 CPU 算力的通用算力平台，可以承载热存储和温存储的应用。对于冷存储、异构等应用，只是预留一些设计，以备未来有需要的时候可以开发。整机柜交付可提高交付效率、降低包材用量以及运输所损耗的燃料，可大幅降低碳排放。



△ 京东云自研液冷整机柜服务器

京东云自研液冷整机柜服务器尽量把业务功能涉及的模块放在前面，比如存储模块、IO 模块等业务功能前置，前出线使得维护更容易。而散热和供电基础设施后置，并预留支持能力，满足 CPU 的散热需求，风冷可以支持到 500 瓦，液冷可以支持到 800 瓦，甚至更高，如果需要更高功耗，可通过改变冷板设备等来实现。考虑数据中心生命周期很长，尤其是液冷技术的支持，预留三代平台的支持，确保整机柜能够在各种各样的部署环境下使用，既可以在自建新机房使用，也可在液冷机房部署，支持各种各样的设备类型和平台。

2、以全栈的视角 垂直整合

数据中心基础设施层面的能耗主要来自于制冷和供电模块的损耗。以典型冷冻水数据中心举例，从内到外包含有冷却塔、冷却水泵、冷水机组、冷冻水泵、空调等，都是用电设备；同样数据中心供电架构从市电到一级转化再到 UPS 到机柜，经历几次转化后也会有供电损耗。

整机柜服务器可以整合供电，不用 PDU 或者很少用 PDU，只起转接不起配电的作用，把电给到电源箱，电源箱到铜排（busbar）上配电，原来在服务器里的电源（PSU，供电单元）集中到电源箱里，成为机柜的一个组成部分。比如一个机柜 30 台服务器，每台服务器两个电源就是 60 个，但是如果把电源集成到机柜上，就用不到 10 个



电源，而且从 1+1 的冗余变成 N+1 的冗余——原来 30 个处于准浪费的状态，现在大大减少浪费，只提供必要的冗余就可以了；电源的数量少了，每个电源的功率比较大，负载也会比较高，电源在负载比较高的时候，转换效率也比较好。

以数字中国万里行团队考察某云数据中心为例，机房里部署了 20 千瓦的液冷整机柜服务器 FusionPoD，园区内还有相对独立的小型液冷机房 FusionCell，由类似集装箱体的供配电、机柜和制冷模块各一组成。

在产品形态上，超聚变液冷整机柜服务器 FusionPoD 类似于数据中心一个 PoD，作为一个天然物理分区，集成了供电、制冷、网络，同时兼容各种各样的服务器，比如为云场景打造的 FusionPoD 600 系列有分布式备电，数据中心使用这个系列可以去掉 UPS，提升供电效率。

FusionPoD 的特点是集成度高，集成了液冷并兼容 1U 的节点设计。从算力密度来看，在 1U 里面最大可以支持 4 个 CPU，风冷服务器通常只部署一半的柜位空间，整机柜可以布满，相对传统的机架服务器算力密度可以提高 8 倍。FusionPoD 机柜是一个平台，天生支持多元算力，机柜里的服务器可以集成计算型、计算存储型包括异构型服务器。FusionPoD 的另外一个特点是全部采用盲插，服务器背后从供液到供电、网络连接，在机柜后方部署有三条总线称之为全盲插，机柜内不用连线，整个部署效率能大幅提升。





盲插的技术难题在于有可能在插拔的时候出现漏液，为了提高可靠性，FusionPoD 在盲插 Manifold 上做了一个防喷射结构，当用户把节点插进来的时候，盲插 Manifold 上的防喷射结构把它封住。同时机柜底下有漏液告警。

同样，FusionPoD 选择冷板式液冷技术路线可兼容现有的基础设施部署，也可应用于新建液冷数据中心。采用混合液冷设计，对服务器里关键发热器件比如 CPU、内存、硬盘、电源等等做了可选的液冷适配并匹配了液冷后门（液冷门），液冷门也是来自于冷塔的供水，把机柜里所有的热量通过液体带走，去掉机房空调和冷机做到全液冷。FusionPoD 保留风扇给一些不太容易做冷板式液冷的小器件，液冷门也是选配，便于客户灵活搭配，利用现有的空调。在泄漏告警、隔离和处理上 FusionPoD 做了相应的设计，比如把节点做成天然能够支持故障隔离的设计，无论通过它的围挡结构的设计还是导流设计，最后对接盲插 Manifold 的设计，当一个节点出现泄漏只会顺着导流槽流往机柜积液盘，不会影响下一个节点，当然前文说的漏液告警监控也属标配。

在智能监控环节，FusionPoD 板内的水晶绳的监控通过服务机 BMC 上传到公司的 Fusion Director，机柜的漏液告警通过机柜顶上 RMU 监控模块也上报给 Fusion Director，由于供水温度很低液冷门出现冷凝水时，冷凝水的漏液告警到 Fusion Director 平台。Fusion Director 能对所有的信息全部汇聚监控进行统一处理。

3、产业生态融合演化

浸没式液冷也成为一套复杂的系统工程，需要在可靠性、经济性和能效之间取得平衡，要解决散热问题的同时解决冷却液和系统中所有部件兼容性、IT 设备高速信号问题。而在系统设计层面，要兼顾服务器和机柜的设计、冷却和监控系统的可靠性，从这个意义来说，液冷不仅是制冷方式的改变，也可能变革数据中心生态。

2018 年 8 月数字中国万里行团队考察了位于张北的阿里云数据中心，这里已经开始部署浸没式液冷服务器集群；2020 年阿里仁和数据中心投入运营，成为更大规模浸没液冷技术的典型实践案例，2022 年，数字中国万里行团队在杭州考察了阿里仁和数据中心。在杭州仁和数据中心部署了阿里云在云网技术、软硬一体探索后新一代智能计算产品：“灵骏”智能算力系统。灵骏智算产品是软硬件一体化设计的算力集群服务，具备公共云、专有云等多种产品

形态，灵骏的底层硬件核心组件由磐久服务器和自研高性能 RDMA 高速网络两部分组成，不仅拥有异构计算弹性能力，还以低通信延时、高并行计算效率为特征提供系统化的高密度计算服务。在浸没式液冷的场景下，整个系统所有的器件都是需要根据适配这种场景做一些调整的，IT 设备需要上插拔上接线和上维护，服务器不是放在立式的机柜里面，传统立式机柜改造成卧式，（整个机柜加上下面的高度不超过 1.2 米），换热设施也需要就近布置，IT 设备需要适配，例如光模块的密封，实际上主板的设计和排布并没有大调整，只是在信号排布和密封方式以及某些连接器做出了一些微小的调整。



阿里浸没式液冷数据中心主要功耗集中在泵与室外散热系统，搭载阿里自研液冷监控系统，能够全自动与负载率相匹配，始终保持系统高效运行。据官方介绍，磐久高性能计算一体机的单位面积算力可达 $8\text{PFLOPS}/\text{m}^2$ (FP16 AI 算力)，单位功耗算力可达 $0.4\text{PFLOPS}/\text{kW}$ 。浸没式液冷从原理上去除了室内部分的空调风机和服务器风机双侧流体驱动系统，彻底排除了空气流动的需求，这样 IT 故障率大幅下降减少维护量、系统热交换次数下降、全自动调泵风机部件运行情况、自主故障预测与调优预测运行，持续保持恒温恒湿环境，有效屏蔽了外界绝大部分不利因素。

新一轮低碳技术带来整体数据中心的技术变革，随着液冷技术在云



计算数据中心的应用，算力服务成本也将进一步降低，惠及更多终端用户。云计算数据中心基于规模和应用需求的优势，对数据中心建设也有足够的掌控力，将会整体数据中心的技术变革、价值链重构和产业生态演化。IT 架构和数据中心基础设施冷却也必将深度融合，构建全栈数据中心成为新趋势，产业链的垂直整合也会成为可能。风液冷也必将在很长一段时间之内共存。

智算中心 跑出液冷加速度

眼下，AI 模型运算量增长速度不断加快，推动硬件算力增长，在 AI 算力持续演进进程中，模型越来越复杂，训练算力需求的增长速度远超摩尔定律，导致处理器功耗持续增加，传统数据中心散热设计极限备受挑战，不断攀升的巨额算力成本也给社会 AI 创新造成巨大负担。

以 2023 年中国数据中心液冷技术峰会上 OPPO 所展示的浸没液冷智算中心实践为例，OPPO 全新基于全高速互联的浸没液冷训练集群，单柜密度提升了逾 400%，机房噪音降低了近 40%。在计算性能方面，TFLOPS 算力相较风冷环境提升约 8%，跑 Bert 模型时间缩短 8%。受液冷的基础设施架构调整的影响，为更好地提升液冷数据中心的利用效率及使用体验，OPPO 自研便捷运维车，并进行了热回收的探索研究。



△ 西部（重庆）科学城先进数据中心

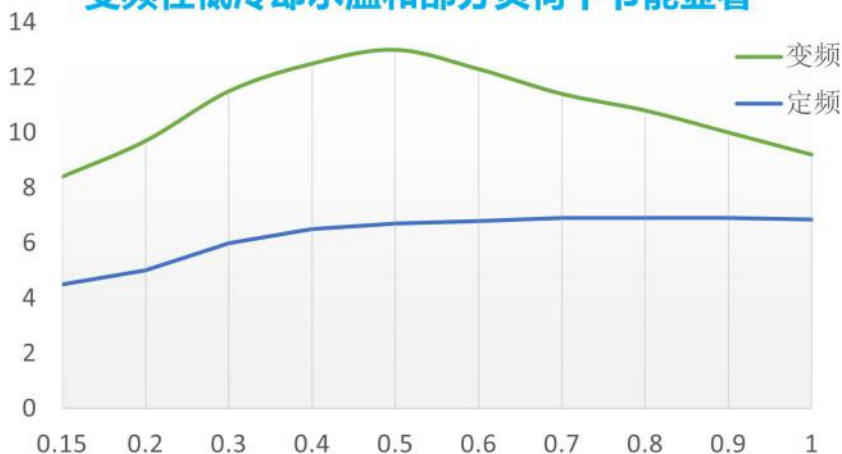
+++++

+++++

同样以“东数西算”成渝枢纽节点内的曙光承建的西部（重庆）科学城先进数据中心为例，该数据中心采用了浸没相变液冷技术、余热回收、绿色建筑、清洁能源（光伏）等多种相关技术，做到了从能源的使用、机架的合理选用、到散热的合理规划、机房设计、布局和使用等多方面的合理布局，全面提高机房散热效率，降低机房的整体能耗，最终达到节能减排的目标。据西部（重庆）科学城先进数据中心运营中心官方数据称，采用了节能技术之后，项目年均 PUE 可达到 1.144，相比传统风冷模式年节省用电约为 14624.8 MWh，年节省标准煤 4870 吨，年减少二氧化碳排放 13149 吨。

数字中国万里行考察的商汤上海临港人工智能计算中心（AIDC），在能源、技术和管理等层面，为 AIDC 采取了多种能源优化措施，年均 PUE 优化至 1.28，其综合搭配采用各项技术，包括液冷、AHU、微模块、高效变频离心机、高温冷冻水、高效供电架构及设备。相比于传统建设方式预计年节约耗电量约 5000 多万度，年减少碳排放约 5600 吨。同时用数字化、智能化调度完成重复而关键的任务，降低运维成本，减少冗余备份（降低运维人员数量超过 20%，降低运维人员工作负荷超过 20%，有效解决 70% 误操作）。商汤上海临港人工智能计算中心节能创新方案中，顶层选用 AHU 间接蒸发冷，相比传统冷冻水系统 + 板换，减少了热交换次数，提升了热交换及制冷效率，年节约用电量约 187 万度。冷板式液冷服务器的液体的热传递效率是空气的 20~30 倍，系统无需冷机压缩机、末端空调风机，节能效果显著，年节约用电量约 176 万度。高效变频离心机相比定频离心机，室外湿球温度低时，COP 大大提升，降低能耗，年节约用电量约 1172 万度。

变频在低冷却水温和部分负荷下节能显著





未来，智算中心为大规模 AI 模型创新与训练提供充裕算力，同时减少闲置浪费，通过算力共享模式，大幅降低社会 AI 算力成本，支持更广泛的 AI 创新研究和应用。智算中心绿色化、集约化发展趋势下，液冷正逐渐成为一个优选项。

未来，智算中心为大规模 AI 模型创新与训练提供充裕算力，同时减少闲置浪费，通过算力共享模式，大幅降低社会 AI 算力成本，支持更广泛的 AI 创新研究和应用。在智算中心绿色化、集约化发展趋势下，液冷正逐渐成为一个优选项。

节能减排新实践 重构排碳之源

数字技术与电子电子技术，成为驱动能源产业变革的重要引擎。

作为支撑数字经济发展的坚实底座，数据中心的计算和处理能力不断加强，“超大规模”数据中心的门槛已经从十万台（服务器）量级向百万台过渡，对部署速度的要求也随之提高，对能源的需求也就越来越大。液冷、蓄冷、高压直流、余热利用、蓄能电站等技术应用，以及太阳能，风能等可再生能源利用，进一步降低数据中心能耗及碳排放。云服务商通过技术驱动实现数据中心节能，构建智能、绿色、高效能的基础设施提升竞争力。



△ 腾讯云仪征东升云计算数据中心 8 栋大平层仓储式机房楼屋顶共计安装光伏组件 28000 多块



2022 年 9 月，数字中国万里行团队参观的腾讯云仪征东升云计算数据中心，占地约 350 亩，8 栋大平层仓储式机房楼，从土建到机电整个建设周期仅用了一年时间。仪征地处长江三角洲西北部，是长三角重点滨江工业城市，腾讯云仪征东升云计算数据中心是目前腾讯在华东地区最大的自建数据中心，计划部署超过 30 万台服务器，包括腾讯云自研星海服务器。在这些强大计算力的基础上，腾讯云超大规模、快速部署、弹性配置的能力支撑各项新型服务，辐射江苏省及长三角地区的产业数字化升级。



△ 腾讯云仪征东升数据中心分布式光伏

在绿色节能方面，借助间接蒸发冷却、气流优化、AI 调优等腾讯多年积累的技术优势，仪征数据中心的整体 PUE 低于 1.25，符合“东数西算”的要求。除了超大规模、快速部署、高效可靠、弹性配置之外，绿色节能也成为刚需。可再生能源可以从电力输入的“源头”上减排，符合双碳战略和东数西算的核心要求。



△ 屋顶光伏板自动清洗机器人

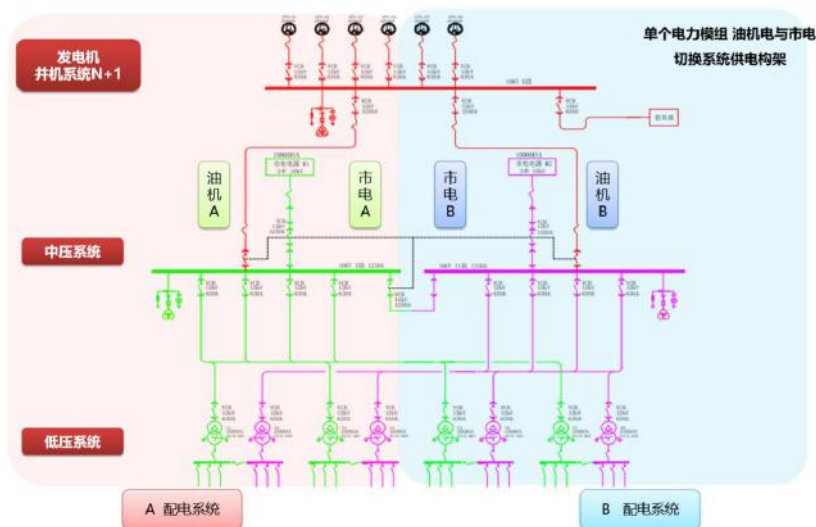
2022 年 2 月，腾讯云仪征东升数据中心分布式光伏项目正式全容量并网发电。该项目充分利用 8 栋大平层机房楼的屋顶面积，共计安装光伏组件 28000 多块，总装机容量达到 12.92 兆瓦，是江苏省目前最大的数据中心屋顶分布式光伏项目。每个屋顶还配有光伏组件自动清洗机器人，保持光伏组件清洁度，实现光伏系统的自动化高效运维。

项目采用“自发自用”的并网方式，近 3 万块单晶硅高效光伏组件产生的直流电经逆变器、变压器等流程处理后接入数据中心的中压电力方仓，将这些可再生电力就地消纳。项目平均年发电量超 1210 万度，每年可节约标煤约 3800 吨，对应减少约 1 万吨二氧化碳排放量，是推动数据中心与绿色低碳产业融合的又一实践。

在中国电子信创云顺义基地，在基础设施层面应用高效变压器、高效 UPS 等技术，在提升数据中心供电质量的基础上，还能够降低电能损耗。在制冷系统方面，磁悬浮冷机、间接蒸发冷却、高效换热器等技术，在保证数据中心制冷的同时，进一步提高了节能效率。

中国电子信创云基地也最大化利用可再生能源，信创云基地在楼体南侧立面布置了单晶光伏组件，为园区照明办公系统提供电能供应，不仅保证了办公等辅助用电，还为降低 PUE 做出了贡献。

商汤上海临港人工智能计算中心节能供电系统架构采用 220kV 直变 10kV 高压供电系统架构以及分散式低压配电系统架构，降低线损约 50%，年节约用电量约 200 万度。采用 SCB13 二级能效变压器，相比 SCB12 减少损耗约 10%，年节约用电量约 156 万度。高效 UPS 单路效率从 95% 提升至 99%，双路平均从 95% 提升至 97%，年节约用电量约 655 万度。



商汤上海临港人工智能计算中心采用 LED 节能灯比传统节能灯节约 75%，降低照明功耗，年节约用电量约 107 万度。冷冻水蓄冷在保障系统不间断运行的同时，能有效利用峰谷电价，削峰填谷。除制冷主机外，冷冻水泵、冷却水泵、冷却塔风机、末端空调风机均采用变频技术，降低其运行能耗和对供电系统的冲击。

算力基础设施功率密度不断提升，算力设施整体能耗偏高，绿色低碳应用需要持续推广，推动数据中心的可持续发展成为必选项。可持续发展是一个长期的价值创造过程，需要将可持续发展的理念纳入选址设计、优化供配电和制冷架构，贯彻“全栈数据中心”理念，加快液冷技术、新型节能新技术的应用和实践，加速新型算力基础设施的绿色升级，进而推动绿色能源革命进程。

能碳融合数字化技术是实现“碳中和”的引擎。在能源革命进程中，能源数字化是当前能源产业变革的一大特征，从自然资源依赖型向技术驱动型转变，采用科技手段开发可再生能源，构筑更经济、更稳定、更安全的发电网络，从源头上降碳更为关键。

在我国电源结构中，煤依然占据主导地位。从发电量看，根据 GE

+++++

+++++



△ 江天数据“环京大数据产业天津基地”1号数据中心在柴发楼南立面安装光伏板

1.5 亿千瓦

国家“十四五”电力发展规划中
气电装机将达

11171 万千瓦

2022 年 7 月
我国气电装机容量

8.1%

从 2017 年到 2022 年 7 月
我国气电装机容量年复合增速

4.55%

2022 年 7 月
气电在我国总发装机容量中占比

Gas Power 发布的《加速天然气发电增长，迈向零碳未来》，2020 年我国气电发量为 2470 亿千瓦时，仅占当年总发电量的 3.3%，远低于其他发达国家。按照国际经验，气电在能源转型中发挥着重要的作用，以日本、美国、英国为例，根据 Gas Power 数据，2019 年日本天然气发电量占总发电量比重达 37%；根据 BP 披露，美国和英国的天然气发电量分别占各自总发电量的 38.63% 与 40.1%。相比而言，我国天然气发电未来增长空间较大。而从 2017 年到 2022 年 7 月，我国气电装机容量从 7570 万千瓦增长至 11171 万千瓦，年复合增速为 8.1%；截至 2022 年 7 月，气电在我国总发装机容量中占比仍较低，仅为 4.55%（欧美、日本等发达国家占比 30% 以上）。国家“十四五”电力发展规划中，将调峰电源作为“十四五”气电发展的主要方向，气电装机将达到 1.5 亿千瓦。随着气源开发和天然气管线建设逐步加快，气电在未来仍有很大的增长潜力，国内燃机服务市场规模也将快速增长。

中国华电作为我国拥有最大燃气发电装机资源的中央企业，通过数字化和智能化转型升级保障燃气电厂本质安全、提升运营效率和创新发展的，比如华电电科院与中国电子云、华电南自华盾公司合作开发的国内首个行业级自主可控燃机智慧运维云平台，通过在电厂、集团、行业三个维度的协同，打造一流的国家级燃机智慧运维平台，为国家的“双碳”目标和能源安全承担央企应尽的责任和义务。



2022 年，数字中国万里行团队特地参观考察中国华电杭州华电江东热电有限公司，该公司通过对燃气发电性能深入分析有效指导开展节能降耗工作，通过燃机智慧运维云平台的辅助运行决策，可优化机组运行方式，从“基于直觉的低效率决策”向“基于数据的科学决策”转变；基于云边协同架构的平台，助力专家远程监控和辅助操作，实现从“信息孤岛、层级冗余”向“集成共享、扁平协作”转变；通过采集、生产等各个环节数据实时感知和共享；建立 AI 模型根据历史数据预测出未来情况，实现从“被动的事后反应”向“主动的预知反应”方式转变，成为燃气发电行业的数字化、智能化发展的典型实践案例。

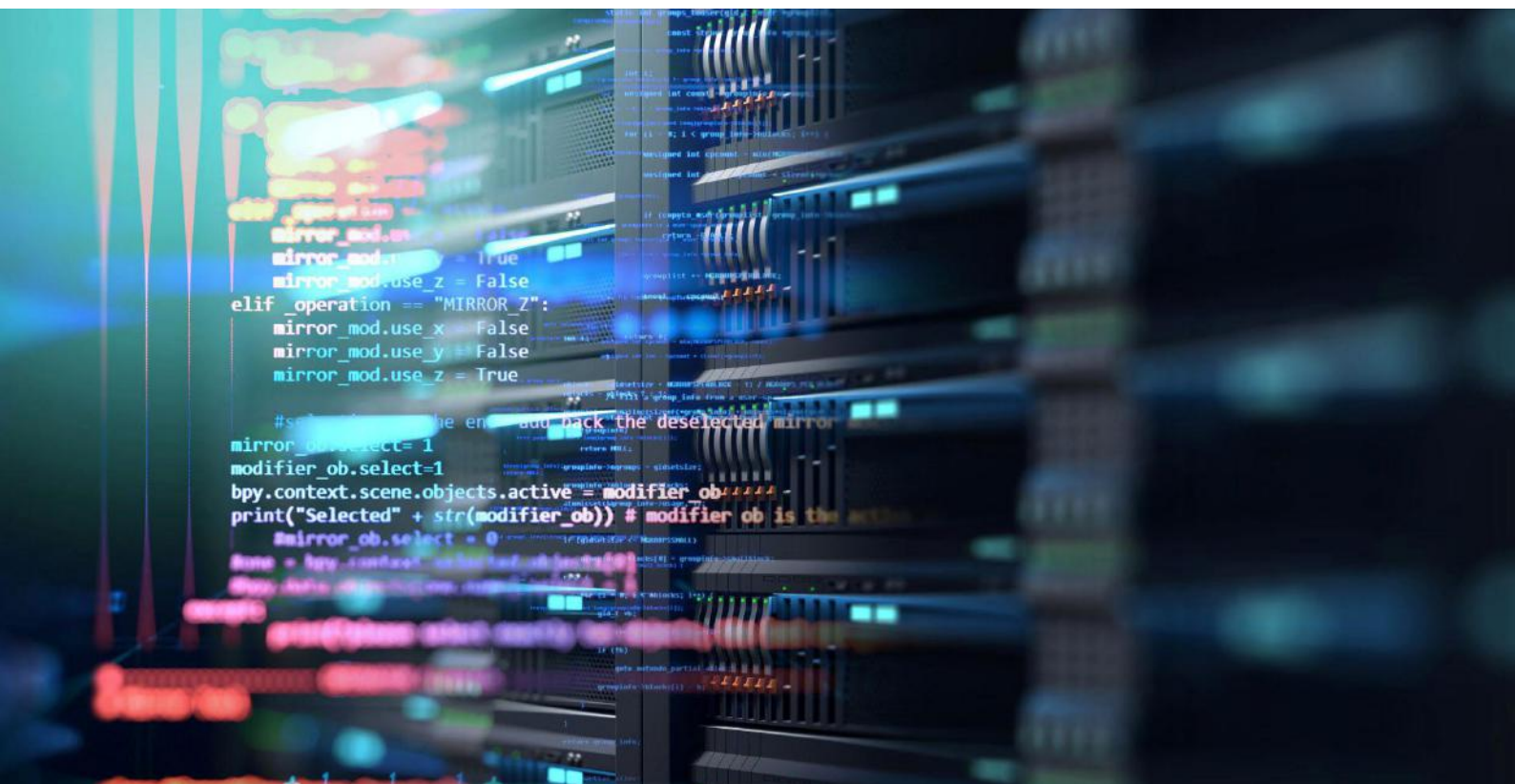
在新的能源产业变革，无论是数据中心的建设者还是数据中心的使用者，在实现绿色低碳转型中积极探索，将数字技术与电力电子技术、发展清洁能源与能源数字化相融合，从源头开始，多措并举，共建可持续发展的未来。





版权声明

《算力经济时代·数字中国万里行-2023 新型算力中心调研报告》版权属于中研益企（北京）信息技术研究院有限公司，并受法律保护；转载、摘编或利用其他方式使用本考察报告文字、图片或者观点的，应注明“来源：益企研究院”；违反上述声明者，本公司保留追究其相关法律责任的权利。



益企研究院
E7 · RESEARCH



E-mail: contact@e7acad.com