异构融合计算：多核协同驱动的创新革命

朱首赫, 朱希研, 李皓

2025 年 6 月 20 日

目录

1 研究背景

近年来，自动驾驶、元宇宙、人工智能等应用不断创新发展，数据规模、算法复杂度以及算力需求爆发式增长。而**登纳德缩放定律（Dennard Scaling）**终结导致单核频率停滞于 3-5GHz；**摩尔定律**晶体管密度倍增周期延至 3-5 年，CPU 性能年增长率降至 3%；**阿姆达尔定律**揭示并行加速受限于任务串行部分，千核级 CPU 扩展收益递减 7，要提高性能亟需探索异构融合技术。同时，各类加速处理器已成为算力基础设施的重要组件，基于 CPU+xPU 的异构计算系统逐渐成为各算力场景的主流架构。然而，随着异构计算系统的种类和数量越来越多，xPU 性能与灵活性难以兼顾、各 xPU 间计算孤岛问题难以协同、调试和维护成本增高等问题愈发凸显，亟需从异构融合计算方向加强理论研究和实践探索。从 CPU 多核并行（如 AMD 96 核 EPYC）到“CPU+XPU”异构架构，再至超异构（Hyper-Heterogeneous）融合，旨在突破单一处理器性能-灵活性权衡困境，也是技术演进的必然。

以人工智能发展为例，《Nature Electronics》期刊在 2022 年 4 月的一篇文章显示：从 2018 年开始，随着 AI 大模型应用的涌现，算力需求平均每 2 个月翻一倍；摩根士丹利估计 “2022 年谷歌的 3.3 万亿次搜索，平均成本约为每个 0.2 美分”，John Hennessy 表示 “基于大模型搜索的成本是标准关键词搜索的 10 倍”。需求的变化和成本的约束，再加上 NoC(Network-on-Chip) 和 SiP(System in Package) 等新芯片技术的赋能，必将推动算力基础架构的变革。计算架构已逐渐从目前各自为政、孤岛式的异构计算，走向异构融合计算。同时，以系统设计为中心，按照应用需求来设计、定义和规划计算架构，推动多层级技术的融合已成为当前的最佳可行方案。

2 定义

异构计算（Heterogeneous Computing）是指不同类型指令集和体系结构的处理器组成的系统计算方式。CPU 和其他处理引擎最大不同在于：CPU 是 Self-Control（图灵完备的），可以独立运行，其他加速处理器需要在 CPU 的协助下运行。因此，异构计算通常是指 CPU+xPU 的异构计算架构（xPU 泛指其他各类非 CPU 的加速处理器）。

Intel 于 2019 年提出” **超异构计算**” 的概念，强调了超异构计算涉及的三个方面：系统架构、工艺和封装，以及统一的异构计算软件。但在最核心的系统架构层次，Intel 仅仅只强调了” 多”，并没有进一步对超异构计算进行阐述，以及设计实现的进一步细节说明。

” **异构融合计算**” 是一个全新的概念，目前行业还没有形成统一的定义。从概念上讲，” 异构融合计算” 属于异构计算的范畴，可以定义为异构计算的一种高阶形态。

狭义的” 异构融合计算”，是一种新的计算架构和方法，通过融合计算。而**广义的” 异构融合计算”**，则通过不同层次、不同类型的技术整合，来实现异构 CPU 和多种不同类型、不同架构的加速处理器，以实现更大规模、更高性能、更加高效的融合计算资源的高效利用。

广义的异构融合计算，主要包含以下几方面内容：

1. **超异构**：系统中异构处理器的数量为三个或三个以上。” 一个称为同构，两个称为异构，三个或三个以上称为超异构”。超异构是异构融合计算的前提。

2. **硬件融合**：强调不同处理器之间的深度协同（指单个工作任务由两个或两个以上处理器协作处理）和深度融合（指某个具体工作任务可以跨 CPU、GPU 和 DSA 等不同类型处理器运行，也可以

跨同类型中的不同架构处理器运行）。各处理器之间可以通过高速总线或高性能网络进行通信和数据传输，通过更高层次的系统划分和任务调度实现协同计算。

3. **软件融合**：面向异构（硬件）计算环境，将操作系统、应用软件、编程模型、编程语言、通信协议、数据等技术资源进行融合和优化，提供统一的软件运行环境和编译开发工具，旨在降低异构融合计算系统的复杂度，实现计算任务的跨平台运行。

4. **系统融合**：通过合理地任务分配和资源调度，异构融合计算系统可以实现更高的计算性能和更好的计算效率。

传统异构计算，特指 CPU+xPU 的计算架构。异构融合计算与传统异构计算的差异点在于：传统异构计算仅有一种加速处理器类型，并且仅关注 CPU 和加速处理器的协同；而异构融合计算则具有两种或两种以上的加速处理器类型，并且需要重点关注所有处理器之间的协同和融合，以及硬件和软件之间的融合，系统内部及系统之间的融合问题。

3 异构融合计算技术的原理

异构融合计算技术的核心原理在于打破传统同构计算或松散耦合异构系统的界限，通过高效的管理和调度机制，将不同类型、不同架构、各具优势的计算单元紧密集成并高效协同工作，形成单一、强大的计算系统，从而实现计算资源的最优利用。这一技术更具体的工作原理可以概括为 6 个层面：**异构处理器协同**、**任务调度与划分**、**统一内存架构**、**高效互联通信**、**多架构编程模型**、**软硬件协同优化**。下面将围绕这六个核心层面详细展开介绍。

图 1: 不同处理器的优势

异构处理器协同是异构融合计算的基石，通过发挥不同处理器的独特优势实现整体性能提升。表 1 整理了不同处理器的特性对比。以一个典型的 CPU+GPU+FPGA 异构融合计算系统为例，其协同机制大致如下：CPU 预处理任务并分配给 GPU 和 FPGA，各单元通过 OpenCL 编程，结果经 PCIe 传回 CPU 整合。同时，**异构系统架构（HSA）**还通过统一虚拟地址空间，使 CPU 和 GPU 共享内存和任务，降低通信延迟。这种机制充分发挥了 CPU 的管理优势、GPU 的并行处理优势和 FPGA 的能效比与灵活性。

表 1: 异构处理器特性对比

处理器类型	**主要功能**	**核心优势**	**典型工作负载/应用场景**	**主要局限性**
CPU	通用计算、逻辑控制、资源管理	灵活、擅长串行处理、分支逻辑	操作系统、通用应用、复杂控制	并行计算能力有限、能效相对较低
GPU	大规模并行计算、图形渲染	高吞吐量、擅长数据并行(SIMD/SIMT)	游戏、AI 训练/推理、科学模拟、3D 建模	串行处理能力有限、通用性不如 CPU
FPGA	可重构硬件加速、定制逻辑	极高灵活性、能效比高、可现场重构	实时信号处理、硬件加速、网络处理、原型验证	设计复杂、开发周期长、频率相对较低
ASIC	特定功能硬件加速	极致性能、极低功耗、体积小	加密货币挖矿、AI 推理专用芯片、数据中心特定加速	缺乏灵活性、无法重构、开发成本高、周期长

异构融合计算中**高效任务调度与划分**至关重要，旨在最小化执行时间、平衡负载并减少通信开销。任务通常被建模为有向无环图，其调度问题在数据结构课程中已有讲解，这里不再赘述。

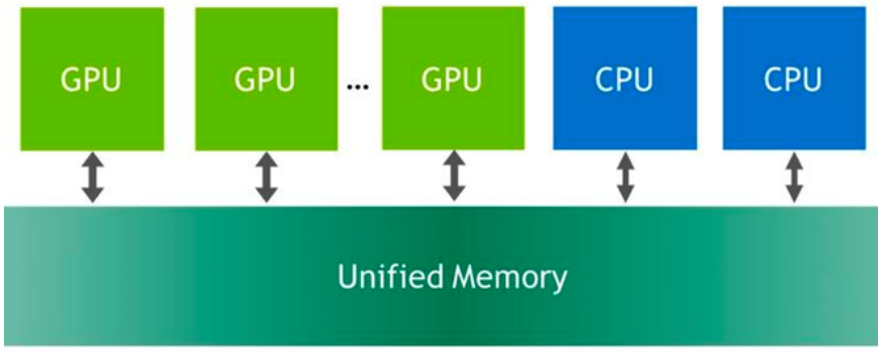

图 2: CUDA 的统一内存架构

统一内存架构解决异构融合计算中数据传输瓶颈和编程复杂性。传统 CPU 和 GPU 内存独立，需显式拷贝，往往会引入延迟和带宽瓶颈。而**统一内存（UM）**提供单一、连贯的内存地址空间，可由任何 CPU 或 GPU 访问，它消除了程序员手动管理数据传输的需要，从而显著简化了编程模型，提高了内存使用效率。以**按需分页迁移（On-demand Paging Migration）**为例：当某个处理器（如 GPU）尝试访问不属于其本地内存但属于统一内存空间的数据时，系统会自动触发页面错误（page fault），并将所需数据页从其当前位置（如 CPU 内存）迁移到访问该数据的处理器本地，即通过共享指针而非数据拷贝来实现了交换数据。

高效互联通信是异构系统发挥潜力的关键，确保处理器间数据快速交换。现有技术包括：**PCIe（Peripheral Component Interconnect Express）**通用高速总线用于连接 CPU、GPU、SSD 等高

带宽组件；**CXL（Compute Express Link）**作为一种基于 PCIe 物理层的高速串行协议，允许在计算机系统内部的不同组件之间进行快速、可靠的数据传输；**NVLink** 提供 GPU 之间的高带宽、低延迟通信；**InfiniBand**（无限带宽技术）提供服务器之间的低延迟和高吞吐量通信。

异构融合计算产品往往涉及不同架构、指令集和编程模型，给软件开发者带来了巨大挑战。为了降低开发门槛，急需要统一的编程抽象。**多架构编程模型**便是能连接异构硬件与软件应用之间的桥梁。它的存在能大大降低软件开发的复杂性，提高程序员生产力。例如我们所熟知的 **CUDA** 就是一种，它通过对流行编程语言的扩展，使开发者能够顺利利用 GPU 的强大并行处理能力。同时，它还提供了包括 GPU 加速库、编译器、开发工具和运行时库等工具，形成了一套强大的生态系统。

软硬件协同优化是异构融合计算实现极致性能和能效的关键，它要求硬件设计与软件需求紧密结合，“软硬兼施”充分释放计算潜力。要实现更大的效益，既需要从底层为异构融合计算设计的硬件架构，也需要能够促进异构融合计算的软件栈。正是目的性构建的硬件与在更大系统抽象框架内提供细粒度控制的软件栈的结合，才能够充分实现异构融合计算所能提供的深度优化。

4 异构融合计算技术的发展应用

4.1 在人工智能与机器学习（AI/ML）领域的应用

随着深度学习模型规模的迅速扩大，AI 计算对算力的需求激增，异构融合计算成为加速训练和推理的核心方案。其核心思想同样是根据不同计算单元的特性和 AI/ML 任务的计算需求进行合理的任务分配与协同处理。

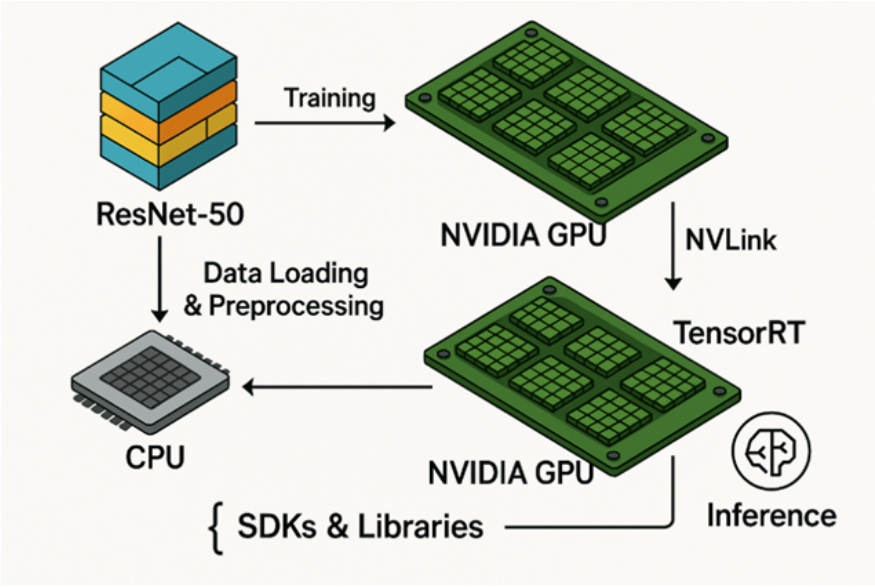

图 3: 基于 NVIDIA 异构融合方案的 ResNet-50 训练流程

NVIDIA 在 AI/ML 领域的核心策略是构建以其强大的 GPU 为中心的异构计算平台。以训练大规模图像识别模型（如 ResNet-50，图 3）为例，NVIDIA 的异构计算融合技术在深度学习模型训练中发挥关键作用。**NVIDIA GPU** 凭借其数千 CUDA 核心并行加速核心的矩阵乘法和卷积运算，配合 **cuDNN** 等库，将训练时间从数周缩短至数小时；CPU 则负责数据加载、预处理和任务调度，有效协

调整个训练流程。对于多 GPU 系统，**NVLink** 高速互联技术提升了数据交换效率，进一步加速了模型并行训练速度。训练后的模型通过 **TensorRT**（一个高性能的深度学习推理优化器和运行时）进行优化和量化，在 NVIDIA GPU 上实现高速低延迟推理，满足实时应用需求。此外，NVIDIA 还提供针对特定 AI/ML 任务的 SDK 和库，简化了开发和部署流程，共同构建了以 GPU 为核心的高效异构计算平台。这种软硬件协同设计的策略，使得 NVIDIA 的异构融合方案在 AI/ML 领域展现出强大的竞争力。

4.2 图像处理与计算机视觉领域的应用

图像处理任务通常涉及大量像素数据，GPU 的并行计算能力使其能够高效执行去噪、物体识别、边缘检测等任务。通过异构计算，GPU 和 CPU 的紧密协作提升了图像处理的效率、实时性和精度。

在医疗影像分析中，**量子算法与边缘计算的协同架构设计**为解决传统算力瓶颈与实时性需求提供了新思路。通过将量子计算的高效并行处理能力嵌入边缘节点，可在本地设备层面对海量医疗影像数据进行快速特征提取与模式识别，同时降低云端传输带来的延迟风险。例如，在医疗场景中，量子退火算法可加速对高分辨率 CT 图像的异常区域定位，而边缘计算节点通过分布式缓存机制实现影像数据的实时预处理，显著缩短诊断响应时间。

为保障跨领域应用的安全性，架构设计中整合了轻量级加密协议与量子密钥分发技术，确保医疗数据的隐私性。在计算流程层面，**量子变分算法**被用于优化边缘节点的特征选择过程，通过构建量子态编码的权重矩阵，有效提升影像分类任务中对抗噪声干扰的鲁棒性。此外，针对边缘设备的异构特性，架构采用**混合量子经典计算模式**，将卷积神经网络中的参数优化任务分解至量子处理器执行，而经典计算单元负责执行梯度更新与模型微调，从而在有限硬件资源下实现计算效率的指数级提升。这一融合架构不仅突破了传统算法在复杂场景下的性能瓶颈，更为后续的集成奠定了可扩展的技术基础。

4.3 通信领域的应用

随着通信技术的飞速发展，尤其是 5G 和未来 6G 网络的部署，通信系统面临着巨大的数据处理和传输压力。异构融合计算在通信领域的应用，能够有效提升通信系统的性能和效率，满足日益增长的通信需求。

高效信号处理：在 5G/6G 基站中，需要处理大量的无线信号，包括信号的调制解调、编码解码、波束成形等。这些任务对计算能力要求极高，尤其是**大规模天线阵列（Massive MIMO）**技术的应用，需要实时处理海量的信号数据。异构融合计算通过将 CPU、GPU 和 FPGA 等处理器结合使用，可以显著提高信号处理的效率。例如，CPU 负责控制和管理任务，GPU 用于并行处理大规模矩阵运算，FPGA 则用于实现特定的信号处理算法，从而实现高效、低延迟的信号处理。

资源管理与调度：基站需要根据不同的用户需求和网络状态动态分配计算资源。异构融合计算系统可以根据实时的流量和信号质量，动态调整 CPU、GPU 和 FPGA 等处理器的负载，确保资源的高效利用。例如，当基站检测到某个区域的用户流量突然增加时，可以将更多的 GPU 核心分配给该区域的信号处理任务，同时利用 FPGA 的灵活性快速调整信号处理算法.

5 异构融合计算的技术挑战与未来趋势

5.1 技术挑战

异构融合计算虽然带来了巨大的性能潜力，但在实际落地和广泛应用中仍面临诸多技术挑战，这些挑战与前文所述的系统内部协同问题有所区分，更侧重于**跨生态、跨平台以及设计-部署全生命周期**的复杂性：

- **异构软硬件生态碎片化与互操作性难题**：当前，异构计算领域呈现出百家争鸣的态势，各类处理器（CPU、GPU、FPGA、ASIC 等）拥有各自独立的指令集、内存模型和编程范式。随之而来的是各种专属的编程语言（如 CUDA for NVIDIA GPU）、库（如 cuDNN）和开发工具。这种高度碎片化的生态环境导致开发者需要投入大量精力学习和适应不同的编程模型，使得应用程序难以在不同的异构平台上实现无缝迁移和高效运行，极大地增加了开发、调试和维护的复杂度。如何构建一个统一、开放、兼容性强的软件栈和开发工具链，实现真正的**跨平台互操作性**，是异构融合计算面临的首要挑战。

- **全栈优化与调试的复杂性**：异构融合计算系统是一个复杂的协同整体，性能优化不再局限于单一处理器层面，而是需要从应用层、编程模型、运行时、操作系统、互联总线直至底层硬件架构进行**全栈式的协同优化**。例如，内存访问模式、数据传输延迟、任务粒度划分、处理器负载均衡等任何一个环节的瓶颈都可能严重影响整体性能。同时，跨越多类型处理器和抽象层次的调试过程异常复杂，难以有效追踪和定位性能瓶颈或错误。传统单一处理器环境下的性能分析和调试工具往往无法胜任异构融合场景，亟需开发新的、能够提供**全局视图和细粒度控制**的诊断工具。

- **功耗与散热的严峻挑战**：随着异构处理器数量的增加和集成度的提高，系统的整体功耗急剧上升，尤其是在数据中心和高性能计算场景下，散热问题变得尤为突出。高功耗不仅增加了运营成本，也对硬件设计和封装技术提出了更高要求。如何**在追求极致性能的同时，有效地控制系统功耗并解决散热难题**，是异构融合计算走向大规模部署的关键制约因素。这需要从芯片架构、封装技术、冷却方案以及智能电源管理等多个维度进行创新。

- **安全与可靠性保障**：异构融合计算系统因其复杂性和多组件特性，引入了新的安全漏洞和可靠性风险。例如，不同处理器之间的信任边界如何界定？共享内存区域的数据完整性和保密性如何保障？单个处理器故障如何影响整个系统的稳定性？在关键任务场景下（如自动驾驶），任何计算错误或系统失效都可能导致严重后果。因此，**如何在异构融合架构中构建端到端的安全防护机制，并确保系统在各种工况下的高可靠性运行**，是亟待解决的挑战。

5.2 发展趋势

异构融合计算正沿着以下几个方向演进，以克服现有挑战并充分释放其潜力：

- **硬件层面：走向更高集成度的 “芯粒（Chiplet）” 与 “SoC”**：未来异构融合将加速向更紧密的硬件集成迈进。**芯粒（Chiplet）技术**将不同功能、不同工艺的计算单元（如 CPU 核、AI 加速器、内存控制器、I/O 接口等）以小芯片的形式集成在一个封装内，通过高速互联（如 **UCIe**）实现片上级别的协同。这不仅能突破传统单片芯片的面积限制和良率问题，还能根据特定应用需求灵

活定制和组合，实现性能、功耗和成本的最优化。同时，**系统级芯片（SoC）**将进一步深化集成，把更多异构处理单元、存储和通信模块整合到单一芯片上，为异构融合提供更高效的物理基础。

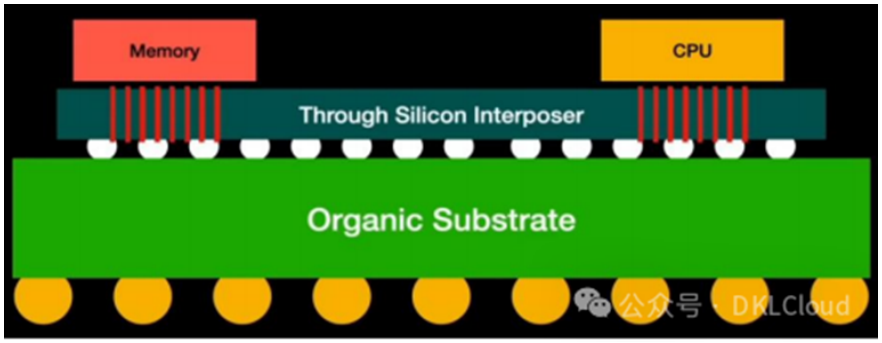

图 4: Chiplet 封装示意图

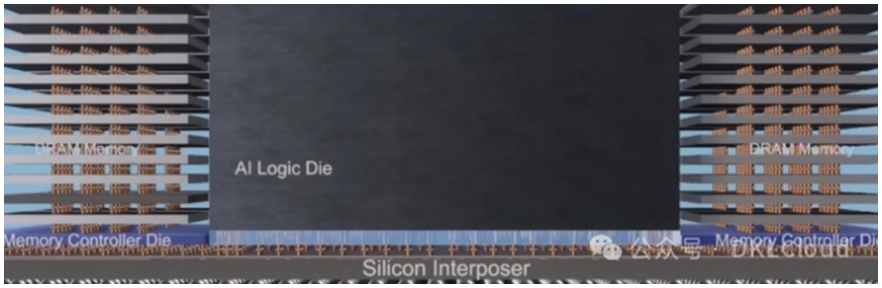

图 5: Chiplet 封装实物图

- **软件层面：构建统一、智能的运行时与编程模型**：为应对生态碎片化问题，未来将出现更加**统一且智能的运行时系统**。这些系统能够自动感知底层异构硬件资源，并根据任务特性进行智能调度和优化，实现代码的自动并行化、任务的动态迁移和资源的弹性分配。**声明式编程模型**（如领域特定语言 DSL）将进一步普及，使开发者能够专注于描述计算逻辑而非底层硬件细节，由编译器和运行时负责高效映射到异构硬件。此外，**AI 辅助的编译优化和性能调优工具**也将成为常态，利用机器学习技术预测最佳任务分配和资源配置方案。
- **架构层面：内存中心与数据流驱动**：传统冯 · 诺依曼架构中计算与存储分离导致的数据搬运瓶颈在异构系统中尤为突出。未来架构将更加强调**内存中心（Memory-Centric）设计**，例如采用**计算内存（Processing-in-Memory, PIM）**技术，将部分计算逻辑下沉到存储单元内部，减少数据在处理器和内存之间往返的开销。同时，**数据流驱动（Dataflow-Driven）架构**将成为趋势，计算任务的执行顺序不再严格依赖于指令序列，而是由数据的可用性驱动，从而更好地适应并行异构环境。
- **应用层面：深度融合与边缘到云的协同**：异构融合计算将不再局限于单一设备或数据中心，而是形成**从边缘设备、终端到云端的全链路协同计算模式**。例如，在自动驾驶中，部分实时性要求高的感知和决策在边缘 AI 芯片上完成，而复杂的路径规划和模型训练则在云端高性能异构集群上进行，两者通过高效通信协议协同工作。这种深度融合将催生更多创新应用，例如 AR/VR 中的实时渲染和交互、工业物联网中的智能制造与预测性维护、医疗健康领域的精准诊断等。

6 总结

异构融合计算通过硬件架构革新（Chiplet/3D 堆叠）、软件抽象突破（统一内存/跨平台编程）及系统级优化（算力网络/FaaS），实现了“灵活性与效率的协同进化”。它不仅是延续摩尔定律的工程技术路径，更是重构计算范式的基础设施革命。随着标准化推进（UCIe/CXL）及 AI/科学计算的需求爆发，异构融合架构将为自动驾驶、元宇宙、量子模拟等场景提供核心算力支撑。

尽管异构融合计算前景广阔，但其发展仍面临软硬件生态碎片化、全栈优化与调试复杂、功耗散热挑战以及安全可靠性保障等一系列严峻挑战。展望未来，随着芯粒技术、统一运行时系统、内存中心架构以及边缘到云深度协同等趋势的不断演进，异构融合计算将逐步克服这些难题，并有望彻底革新我们处理复杂计算任务的方式，为各行各业带来前所未有的创新机遇。

7 分工

朱首赫：贡献选题思路; 明确小组分工; 协调与整合组内工作; PPT1-4 节制作; 报告文档第 3-4 节撰写; 报告文档润色排版与翻译;

朱希研：贡献选题思路; 报告文档第 1、4、6 节撰写; PPT 引入和总结设计; 课堂汇报;

李皓：贡献选题思路; PPT4-5 节制作; 报告文档第 4-5 节撰写;

参考文献

[1] Muthukumaran Vaithianathan. The Future of Heterogeneous Computing: Integrating CPUs, GPUs, and FPGAs for High-Performance Applications. *International Journal of Emerging Trends in Computer Science and Information Technology*, 6(1), 2025.

[2] George Kyriazis. *Heterogeneous System Architecture: A Technical Review*. AMD, 2012.

[3] 工业和信息化部电子第五研究所. 异构算力统一标识与服务白皮书. 2023.

[4] Jonathon Evans, Michael Andersch, Vikram Sethi, Gonzalo Brito, and Vishal Mehta. *NVIDIA Grace Hopper Superchip Architecture Whitepaper*. Nov 10, 2022.

[5] Jing Pei, Lei Deng, Sen Song. Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature*.

[6] Muthukumaran Vaithianathan. The Future of Heterogeneous Computing: Integrating CPUs, GPUs, and FPGAs for High-Performance Applications. *International Journal of Emerging Trends in Computer Science and Information Technology*, 6(1), 2025.

[7] 量子算法融合边缘计算优化医疗影像与自动驾驶安全实践.