异构融合计算：
多核协同驱动的创新革命

汇报：朱希研

PPT ： 朱首赫、李皓

Table of contents

01

异构融合计算技术的研究背景

Research Background of Heterogeneous Fusion Computing Technology

登纳德缩放定律终结

单核频率停滞于 **3-5GHz**

摩尔定律放缓

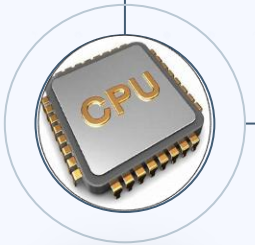
晶体管密度倍增周期延至 **3-5年**

CPU性能年增长率降至 **3%**

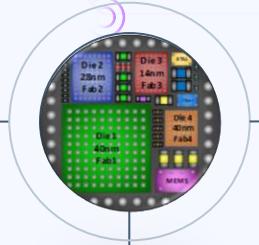
阿姆达尔定律揭示串行瓶颈

千核级CPU扩展收益 **递减**

同构计算方式



转向



基于CPU+xPU
的异构计算

随着种类和
数量越来越多

- 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

?

亟需探索异构融合计算技术

02

异构融合计算的定义

Definition of Heterogeneous Fusion Computing

广义的异构融合计算

超异构

系统中异构处理器的数量为三个或三个以上

硬件融合

强调不同处理器之间的深度协同和深度融合

软件融合

面向异构硬件计算环境，将OS、编程模型、通信协议等进行融合和优化，提供统一的软件运行环境和编译开发工具

系统融合

合理分配任务和调度资源，实现更高的计算性和效率。

03

异构融合计算技术的原理

——六大核心层面

Principles of Heterogeneous Fusion Computing Technology

(1) 异构处理器协同

- 内容：在抽象层面上对整个异构计算系统架构的设计
- 目的：通过协同机制发挥不同处理器的独特优势实现整体性能提升

处理器类型	主要功能优势
CPU	通用计算、逻辑控制、资源管理
GPU	大规模并行计算、图形渲染
FPGA	可重构硬件加速、定制逻辑
ASIC	特定功能硬件加速

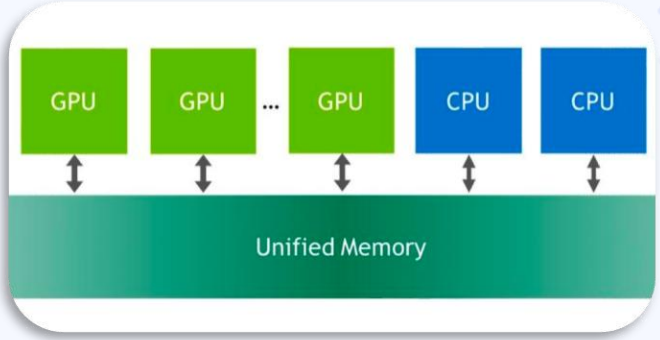
(2) 任务调度与划分

- 旨在最小化执行时间、平衡负载并减少通信开销
- 建模为有向无环图（AOV网的拓扑排序问题）

（3）统一内存架构

- 提供单一、连贯的内存地址空间，可由任何CPU或GPU访问
- 消除了程序员手动管理数据传输的需要，
  解决异构计算中数据传输瓶颈和编程复杂性
- 具体实现的例子

  按需分页迁移：通过共享指针而非数据拷贝来实现了交换数据



（4）高效互联通信

- 确保处理器间数据快速交换，是异构系统发挥潜力的关键，
- 现有技术：PCIe：　通用高速总线、
  CXL：　基于Pcle的高速串行协议
  NVLink：提供NVIDIA GPU间高速互联
  InfiniBand：提供服务器间高性能网络通信

（5）多架构编程模型

- 统一不同架构、指令集和编程模型的编程抽象
- 降低软件开发门槛，提高程序员生产力
- 例如：CUDA通过对流行编程语言的扩展，便于开发者利用GPU的强大并行处理能力加速应用。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。

（6）软硬件协同优化

- 要求硬件设计与软件需求紧密结合
- 既需要从底层为异构计算设计的硬件架构，也需要能够促进异构计算的软件栈。
- 二者“双向奔赴”，才能充分释放异构计算的潜力

04

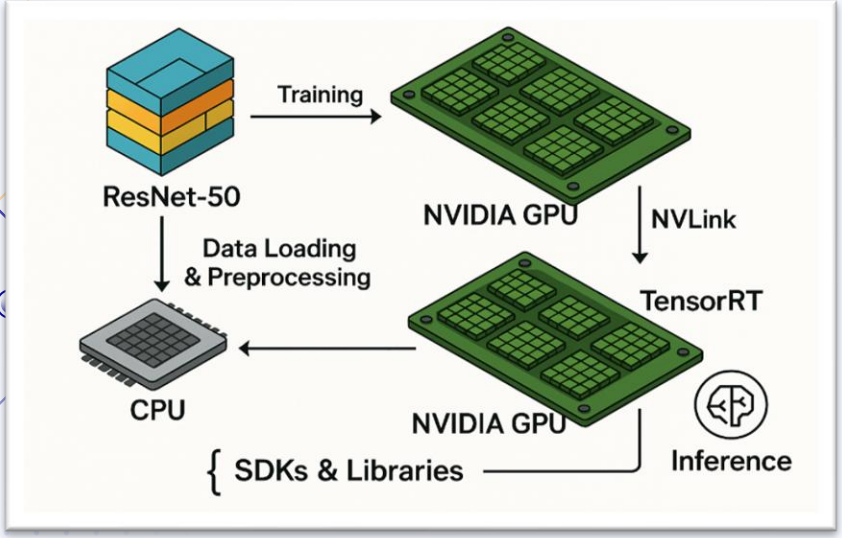
异构融合计算技术的发展应用

Applications of Heterogeneous Fusion Computing Technology

在人工智能与机器学习（AI/ML）领域的应用

满足AI/ML任务对算力的巨大需求

以大规模图像识别模型ResNet-50的训练为例：



GPU+配合cuDNN等：并行加速，大大缩短训练时间

CPU：数据加载、预处理和任务调度

NVLink：高速互联，提升数据交换效率

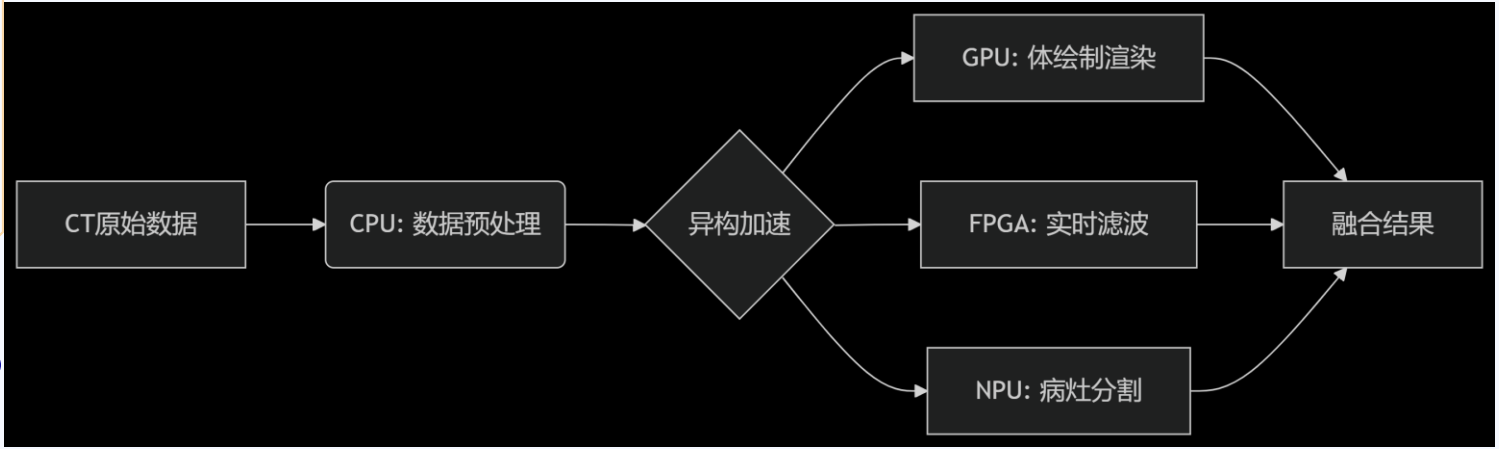
TensorRT：优化和量化

特定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

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。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。

在通信领域的应用

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

CPU	• **控制和管理任务** • Control and manage tasks
GPA	• **并行处理大规模矩阵运算** • Parallel processing of large-scale matrix operations
FPGA	• **实现特定的信号处理算法** • Implement a specific signal processing algorithm

实现高效、低延迟的信号处理

05

异构融合计算技术的技术挑战与未来趋势

Technical challenges and future trends

技术挑战

（1）硬件层面

指令集与内存模型不统一：不同计算单元（如CPU的通用指令集与GPU的SIMD架构）的硬件设计差异导致任务调度和数据交换困难。

数据传输延迟与带宽限制：异构设备间通常依赖PCIe或NVLink互连，带宽有限（如PCIe 5.0仅128GB/s），成为性能瓶颈。

能效平衡问题：高能效单元（如ASIC）与灵活单元（如FPGA）的协同需动态功耗管理，但现有技术难以实时优化。异构融合计算系统中，硬件资源的管理与调度至关重要。

（2）软件层面

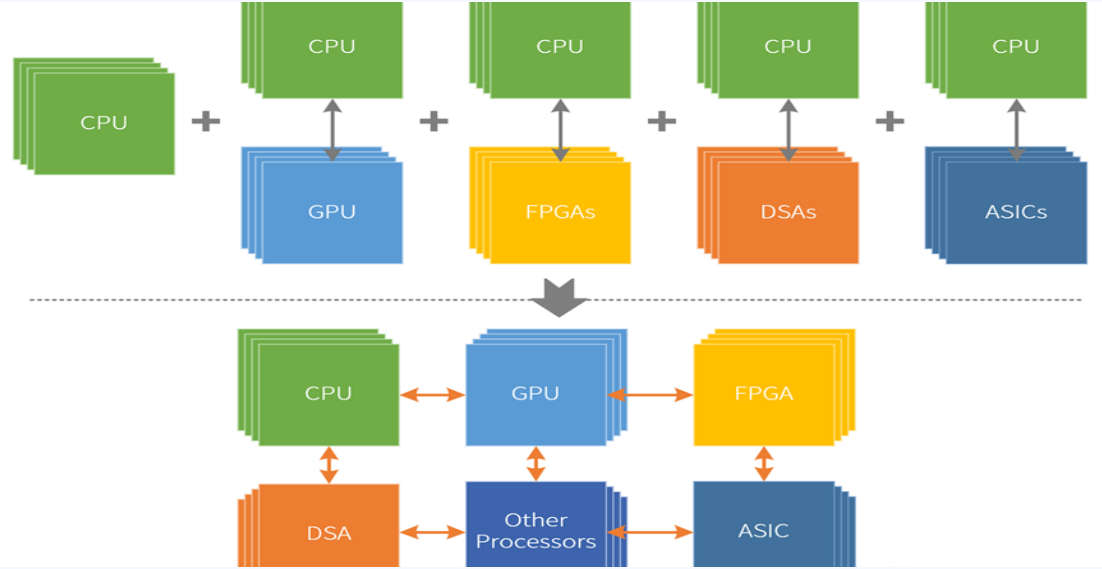
抽象层缺失：现有框架（如Kubernetes Device Plugin）对异构资源的抽象不彻底，无法完全屏蔽底层差异。

工具链碎片化：各厂商提供独立的Profiler（如NVIDIA Nsight、Intel VTune），缺乏跨平台调试方案

（3）新兴场景的适应性挑战

三类及以上芯片整合：超异构（如CPU+GPU+ASIC）需解决更复杂的任务划分与数据流优化，现有调度算法难以胜任。

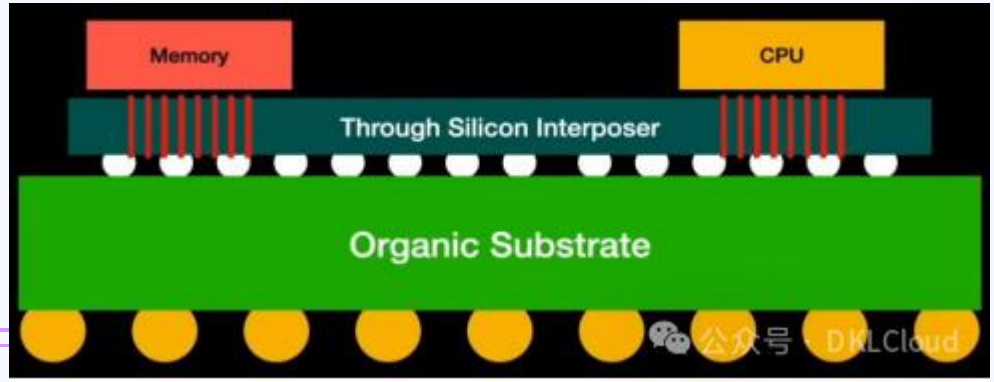
- **量子-经典混合计算**：量子处理器与经典异构系统的协同面临算法适配和错误校正等难题。



未来趋势

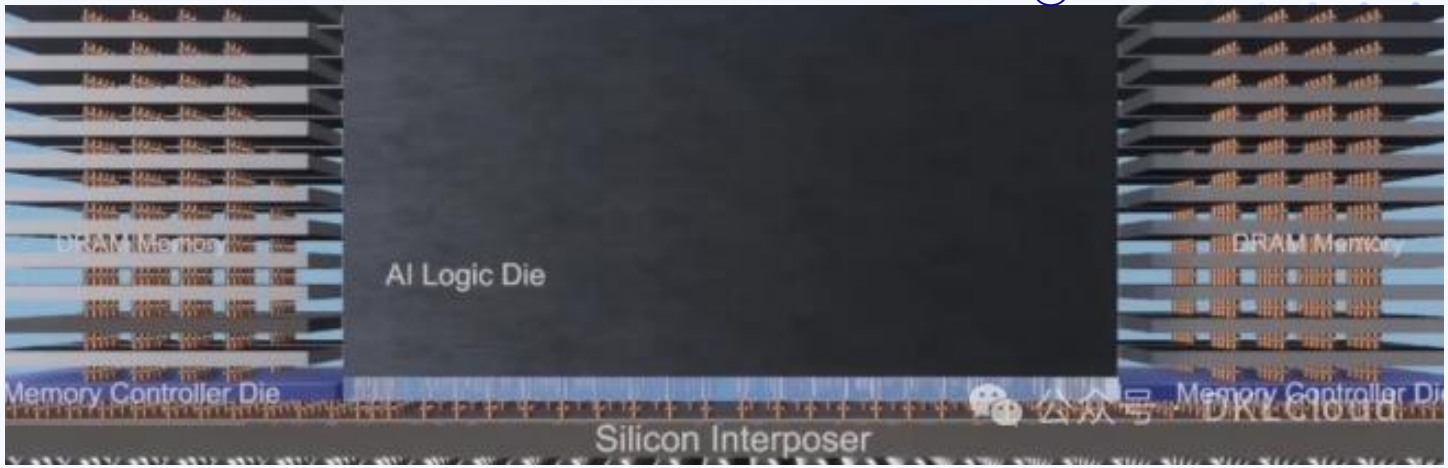
（1）硬件架构创新

更高集成度的异构芯片：未来，随着芯片制造工艺的不断进步，异构融合计算芯片的集成度将进一步提高。例如，通过采用Chiplet技术，将不同架构的处理器、内存、I/O等组件集成在一个芯片封装内，实现更紧密的协同和更高的性能。同时，3D堆叠技术也将得到更广泛的应用，通过在垂直方向上堆叠多个芯片，进一步提高芯片的性能和密度。



2.5D Chiplet封装示意图

未来趋势



3D Chiplet封装示意图

新型处理器架构的涌现：除了传统的CPU、GPU、FPGA等处理器，未来还将出现更多新型的处理器架构，如类脑计算芯片、量子计算芯片等。这些新型处理器具有独特的计算能力和优势，将在特定领域发挥重要作用。异构融合计算将不断融合这些新型处理器架构，以满足未来多样化、复杂化的计算需求。

未来趋势

(2) 软件技术发展

统一编程框架：如NVIDIA CUDA、Intel OneAPI、华为昇腾CANN等，提供跨硬件抽象层，降低开发门槛。

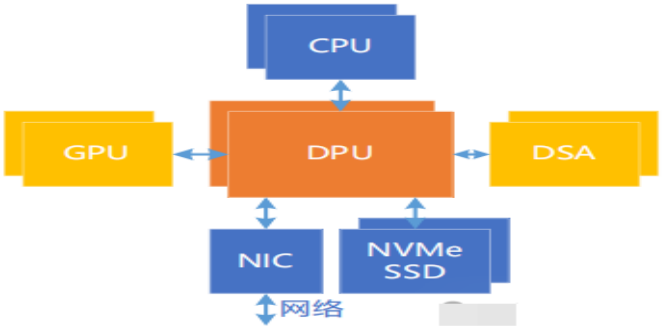
AI驱动的编译优化：自动任务分割与代码生成（如TVM、Halide），提升异构资源利用率。

云原生与Serverless调度：Kubernetes扩展支持GPU/NPU/FPGA动态调度，实现弹性资源池化

未来趋势

（3）垂直场景深度优化

AI与大模型：GPU+NPU协同加速训练与推理，如NVIDIA DGX H100与华为昇腾910B的组合。
自动驾驶：车载超异构芯片（如NVIDIA Thor、特斯拉HW5.0）实现2000+ TOPS算力，支持L4/L5自动驾驶。
边缘计算：轻量级异构架构（如RISC-V+NPU）满足低功耗、低延迟需求。
医学影像与科学计算：GPU+FPGA加速实时3D重建与流体仿真。



设备级异构融合案例：以 DPU 为中心的计算架构

Thanks !