IPADS
INSTITUTE OF PARALLEL AND DISTRIBUTED SYSTEMS

「网络天下·处理器技术论坛」

面向新算力硬件体系的调度技术挑战
“操作系统视角”

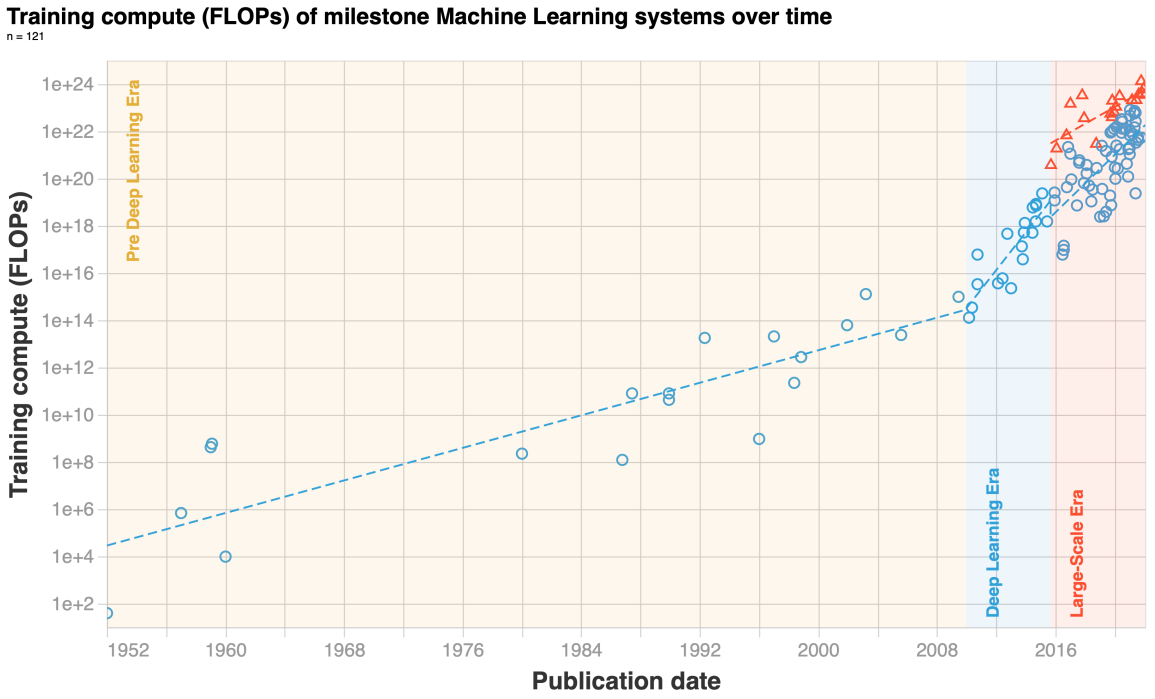
陈 榕

IPADS@上海交通大学

华为·松山湖 2023

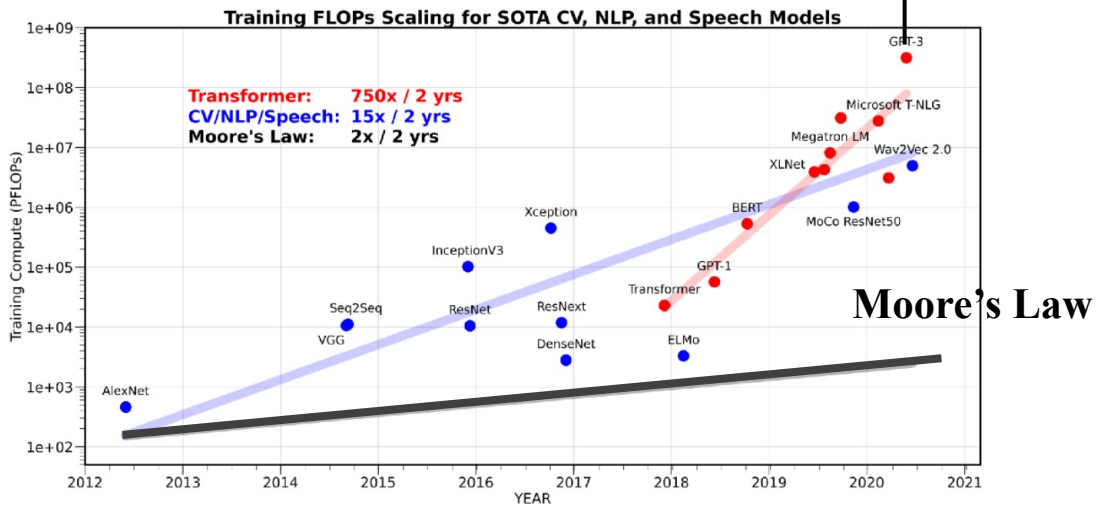
算力需求爆发增长

人工智能应用（尤其是深度学习）带来了爆发式的算力需求增长



Source: “Compute Trends Across Three Eras Of Machine Learning”, 2022
https://github.com/epoch-research/Compute-Trends

Inference : 740,000 GFLOPs
Training : 310,000,000 PFLOPs
GPT-3



Moore’s Law

Source: “AI and Memory Wall”, 2021
https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

算力硬件增长放缓

Moore's (CPU) Law: 2X / 2 Yrs vs. Huang's (GPU) Law: 25X / 5 Yrs ?

Performance : GPU = CPU

Double (2X) FLOP/s

CPU: 2.32 Yrs (Historical)

GPU: 2.31 Yrs (Avg), **2.69 Yrs (Top)**

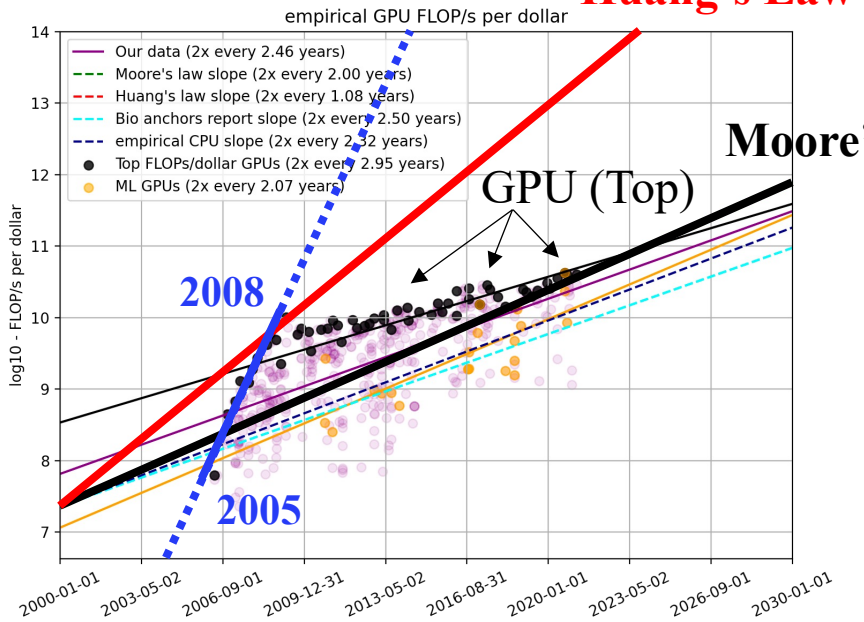
Price-Performance : GPU < CPU

Double (2X) FLOP/s per Dollar

CPU: 2.32 Yrs (Historical)

GPU: 2.46 Yrs (Avg), **2.95 Yrs (Top)**

Huang's Law @2018



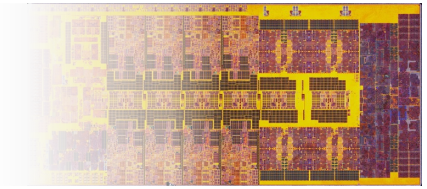
需要更精打细算的使用 GPU

Source: "Trends in GPU price-performance", 2022, https://epochai.org/blog/trends-in-gpu-price-performance

算力硬件演进方向

规模化

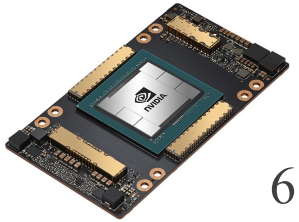
CPU：多核 / 众核 / NUMA



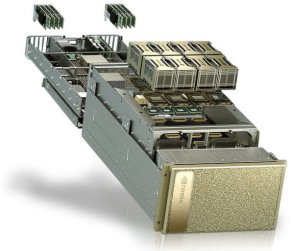
Intel CPU
Raptor Lake
24 cores

Huawei ARM
Kunpeng 920
64 cores

GPU：Cores / SMs / 多卡



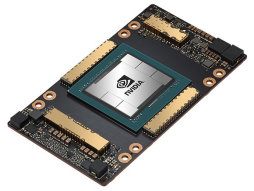
Nv A100
108 SMs
6,912 cores



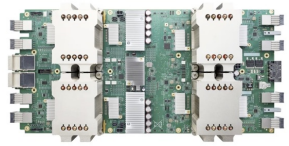
Nvidia DGX-2
A100 x8
AMD/64 x2

领域化

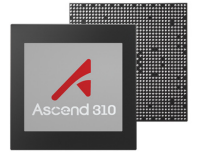
GPU



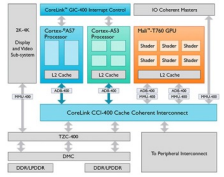
TPU



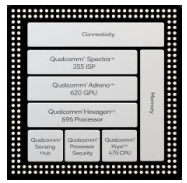
NPU



异构化



ARM
big.LITTLE



Qualcomm
AIE

Nvidia
Jetson Orin

应用的算力外需求

实时性

智能车

1. 障碍物检测
2. 红绿灯识别
3. 语音助理
4. 疲劳监测
5. …

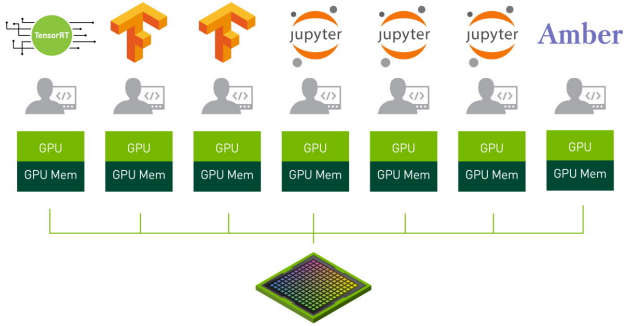
典型DNN推理任务：3.5～13.6 ms

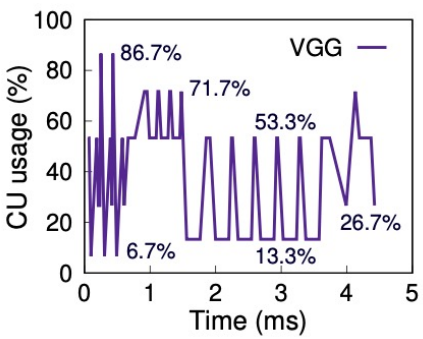
Model	ResNet	DenseNet	VGG	Inception	Bert
#Kernels	307	207	55	146	205
Exec. Time	13.6	3.5	4.4	8.3	5.4

Testbed：AMD Radeon Instinct MI50 GPU

性价比

GPU多任务共享



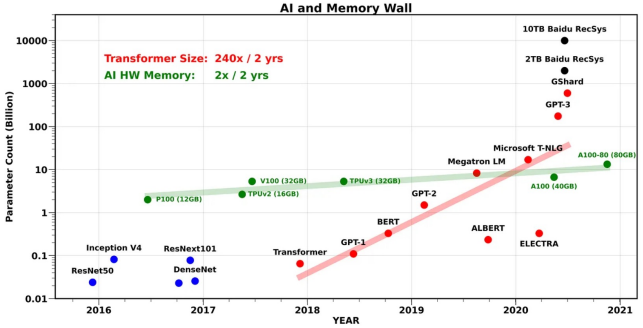


独占使用利用率低

大存储

#Param：Baidu RecSys = **10 TB**

GPT-3 = **175 GB**



需求：240X / 2 years

供给：2X / 2 years

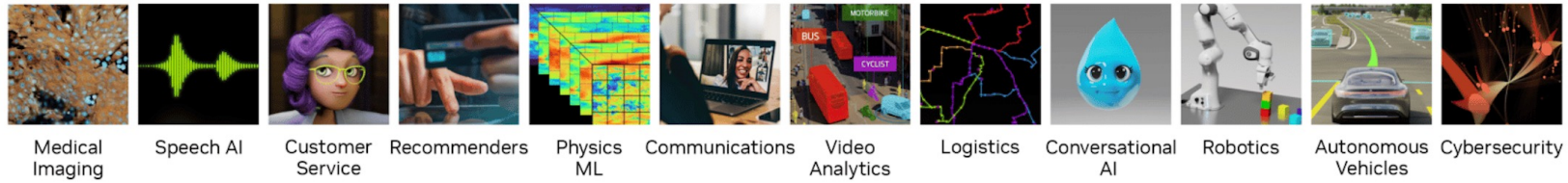
V100

16-32 GB

A100

40-80 GB

操作系统：调度机制挑战



应用

需求

大算力

实时性 / 性价比 / 大存储 / …

系统基础软件

调度策略

系统框架/服务

调度策略

去调度

操作系统

调度机制

资源调度

算力硬件

CPU

GPU

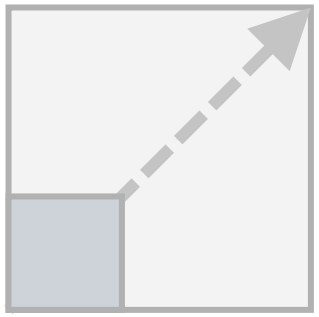
FPGA

ASIC

面向新算力硬件体系的调度机制挑战

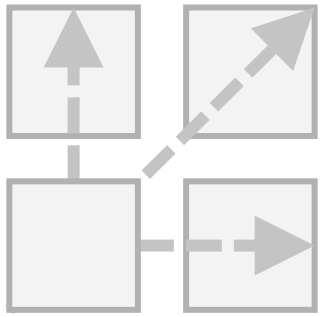
1.

领域化算力
资源的调度



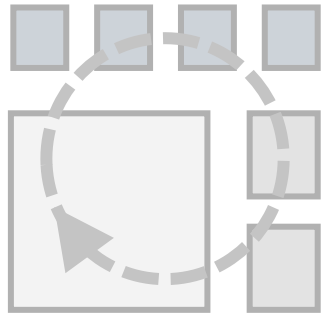
2.

规模化算力
资源的调度



3.

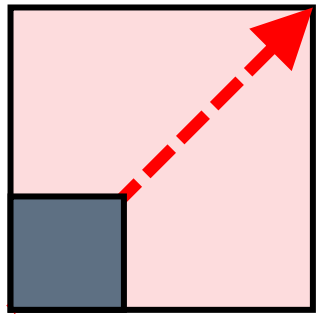
异构化算力
资源的调度



面向新算力硬件体系的调度机制挑战

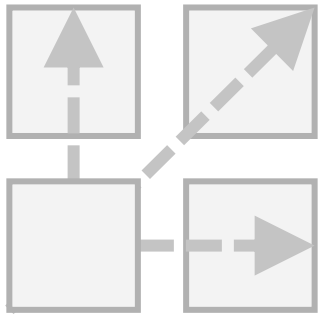
1.

领域化算力
资源的调度



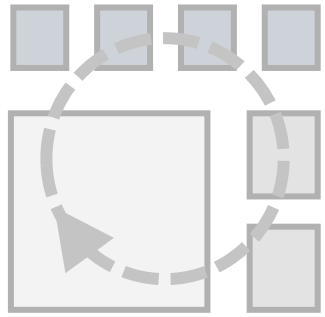
2.

规模化算力
资源的调度



3.

异构化算力
资源的调度



领域化算力资源 —— 高并发

GPU

高并发算力

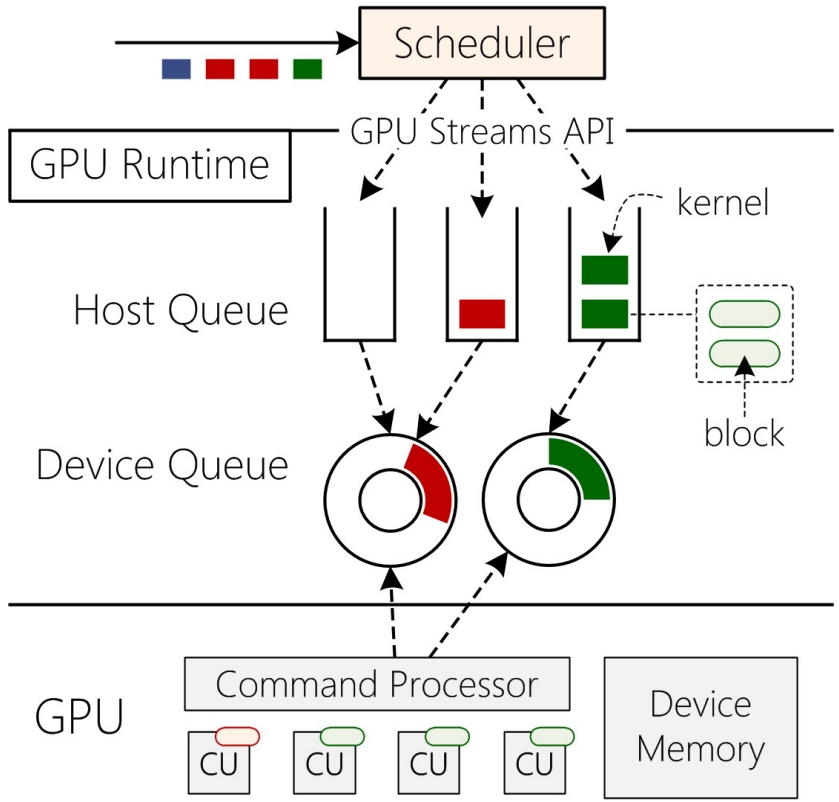
GPU Architecture
(Nvidia Ampere)

1 GPU
x 8 GPCs / GPU
x 8 TPCs / GPC
x 2 SMs / TPC

= 128 SMs*

SM Architecture

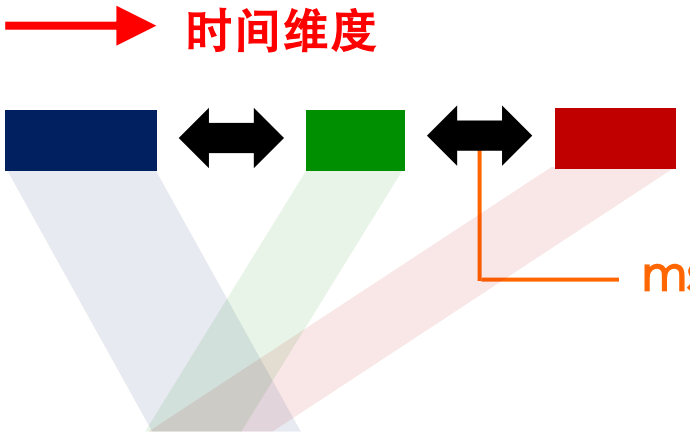
GPU任务调度



* GPU调度单元 NVIDIA使用SM、ARM使用CU

调度挑战 #1：时间维度

GPU 



vs. CPU 

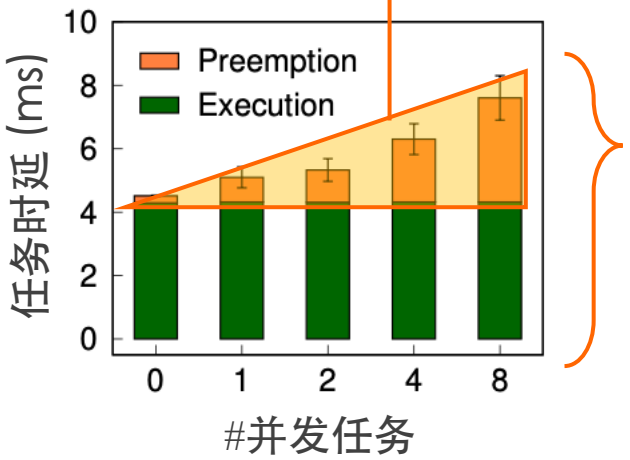
- 执行时间更长（5-100ms）
- 调度延迟更低（~5us）

ms级任务调度延迟

典型DNN推理任务：3.5～13.6 ms

Model	ResNet	DenseNet	VGG	Inception	Bert
#Kernels	307	207	55	146	205
Exec. Time	13.6	3.5	4.4	8.3	5.4

Testbed：AMD Radeon Instinct MI50 GPU



调度开销
接近一半*

调度机制挑战 #1：实时任务抢占

* 未考虑GPU内存切换、任务加载等开销

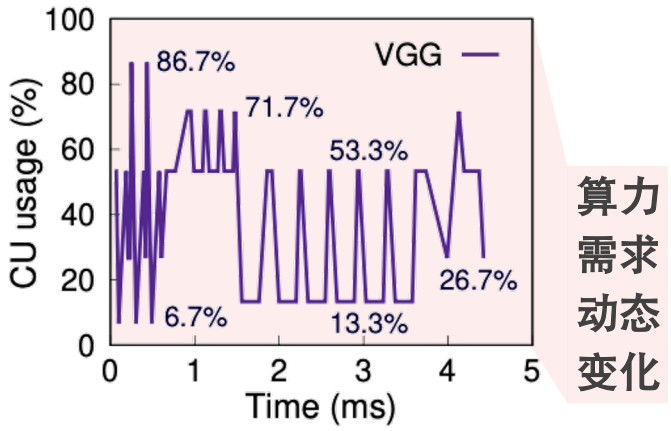
调度挑战 #2：空间维度

GPU

时间维度

空间维度

调度单位
SM / CU *



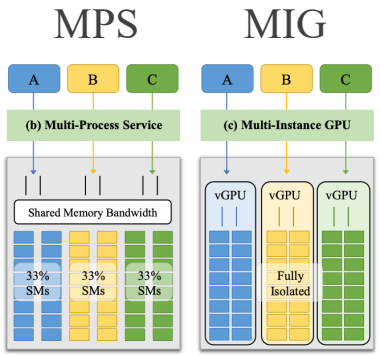
算力需求动态变化

vs. CPU

- CPU core 独立调度
- 无空间维度调度问题

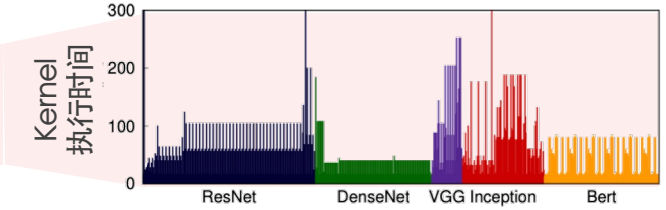
GPU共享机制

- MPS：共享调度
- MIG：静态划分



问题：静态配置

执行时间动态变化



调度机制挑战 #2：动态算力共享

* GPU调度单元 NVIDIA使用SM、ARM使用CU

调度挑战 #3：内存维度

GPU 

空间维度

内存维度

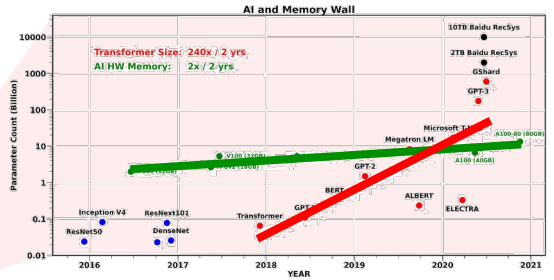
内存竞争 ⟷ GPU内存

- 容量受限
- 带宽受限

GPU内存数据切换
成为调度主要瓶颈

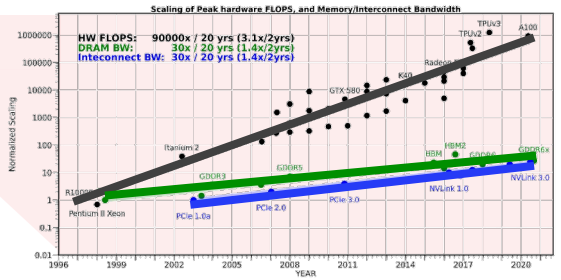
vs. CPU 

- 内存算力比远大于GPU



内存需求：240X / 2 Yrs

内存容量：2X / 2 Yrs



算力增长：3.1X / 2 Yrs

内存带宽：1.4X / 2 Yrs

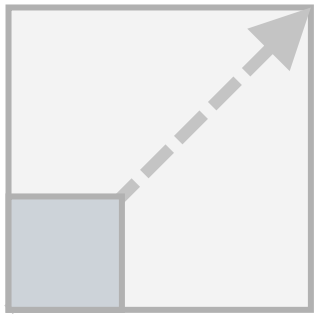
互联带宽：1.4X / 2 Yrs

调度机制挑战 #3：算力和内存的协同调度

面向新算力硬件体系的调度机制挑战

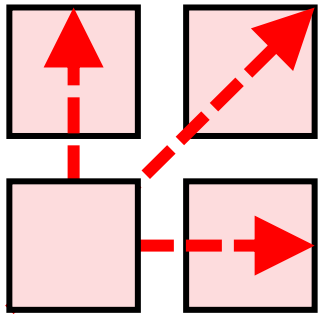
1.

领域化算力
资源的调度



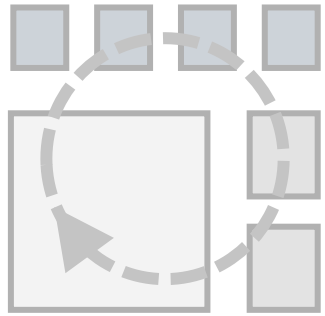
2.

规模化算力
资源的调度



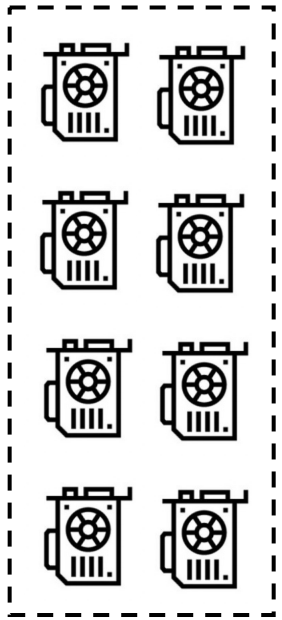
3.

异构化算力
资源的调度



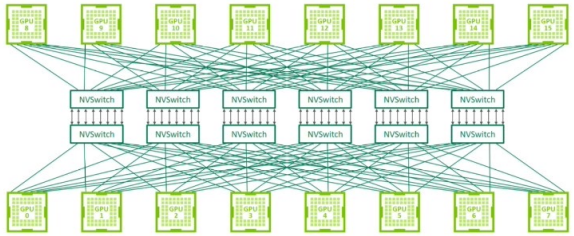
规模化算力资源 —— 单机多卡

GPU X 8

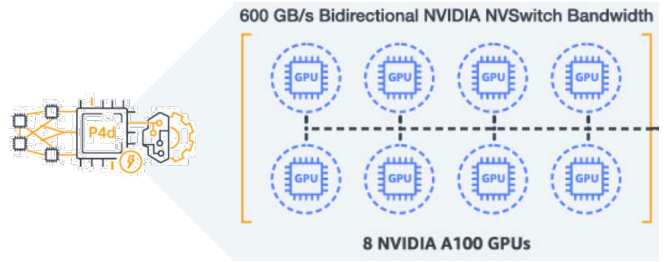


单机多卡 + 高速链接 = 主流配置

- **Nvidia DGX server**
  - 8-16 GPUs (V/A/H)
  - NVLink / NVSwitch



- **Amazon EC2 P4d**
  - 8 A100 GPUs
  - 600 GB/s NVSwitch

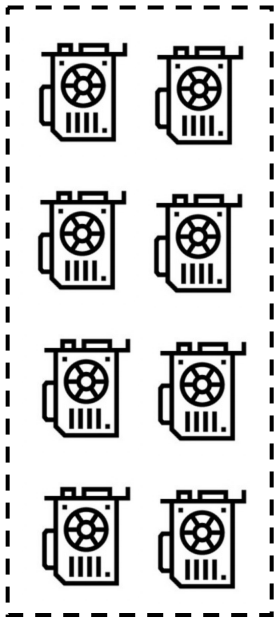


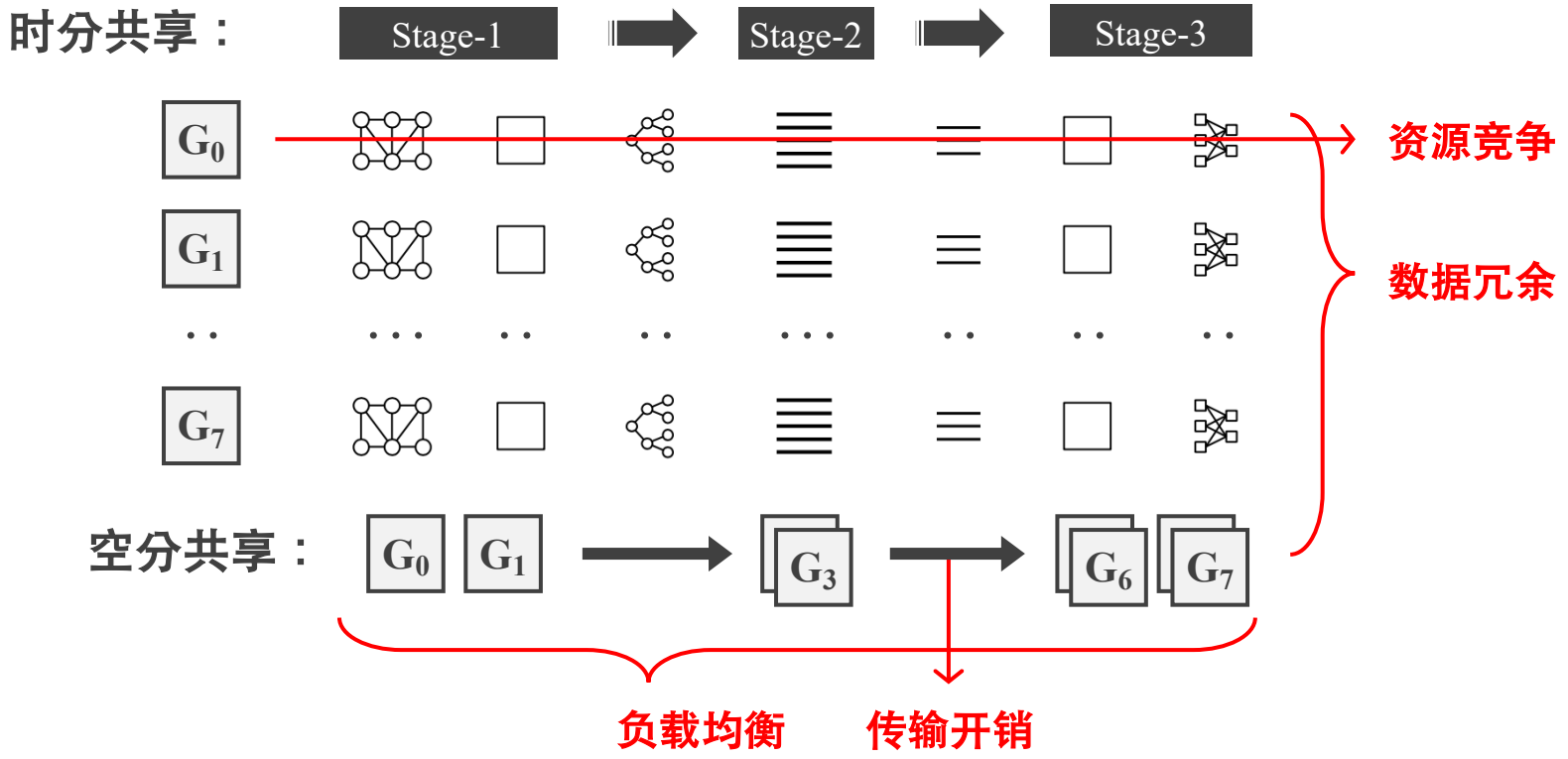
...

规模化算力资源：Multi-GPU（vs. CPU Multi-socket）

调度挑战 #4：共享方式

GPU
X 8

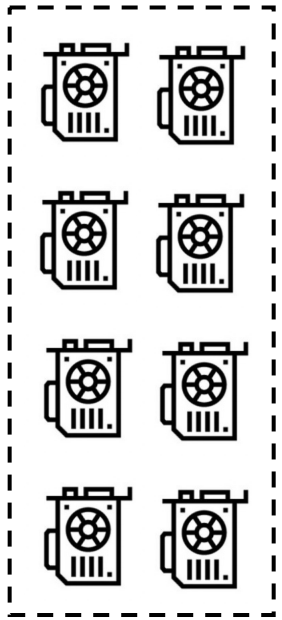


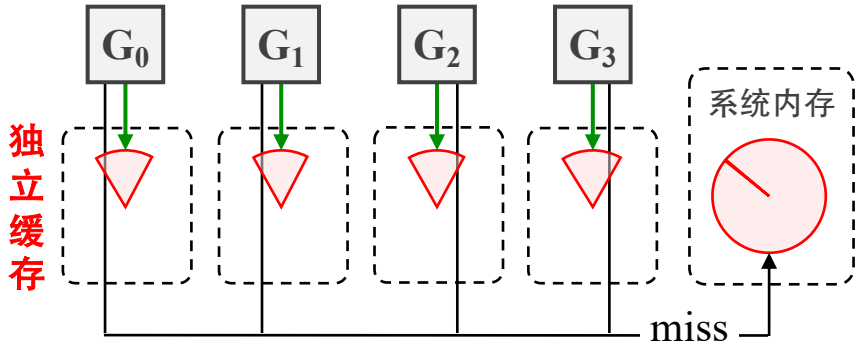


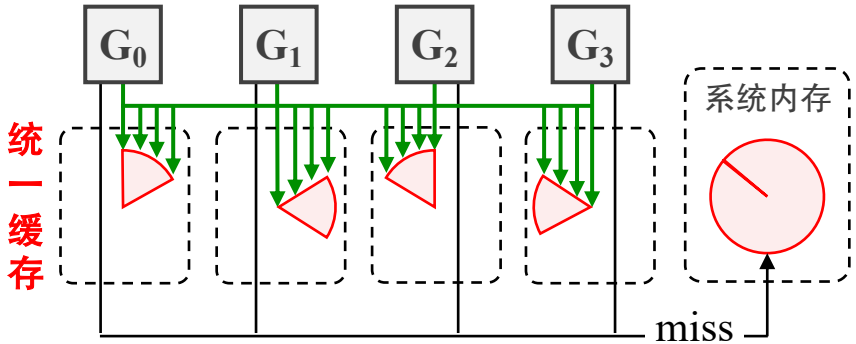
调度机制挑战 #4：算力资源的共享方式

调度挑战 #5：通信带宽

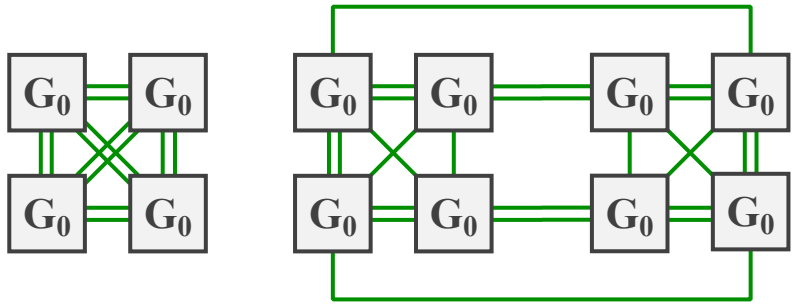
GPU
X 8

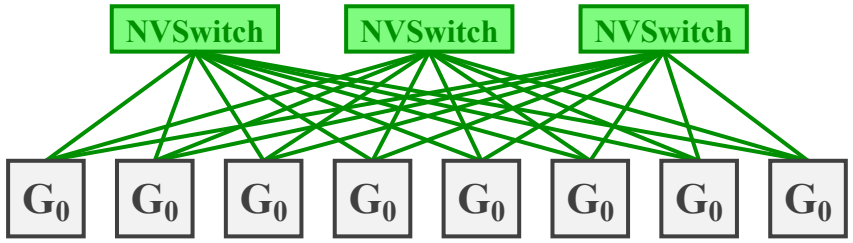






异构带宽拓扑



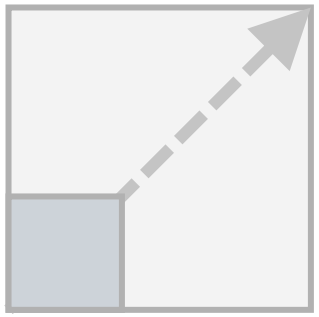


调度机制挑战 #5：带宽感知的调度优化

面向新算力硬件体系的调度机制挑战

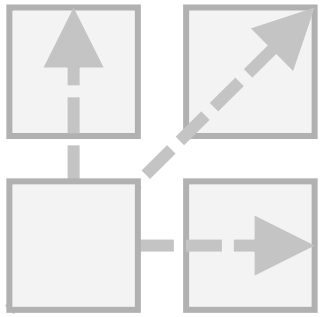
1.

领域化算力
资源的调度



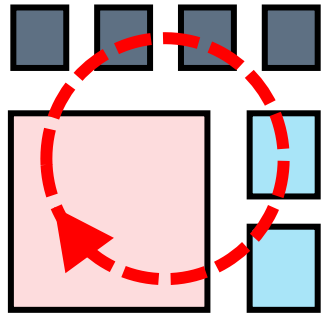
2.

规模化算力
资源的调度



3.

异构化算力
资源的调度



异构化算力资源 —— C+G+NPU

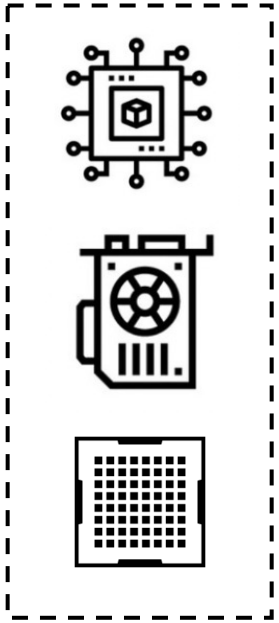
CPU

+

GPU

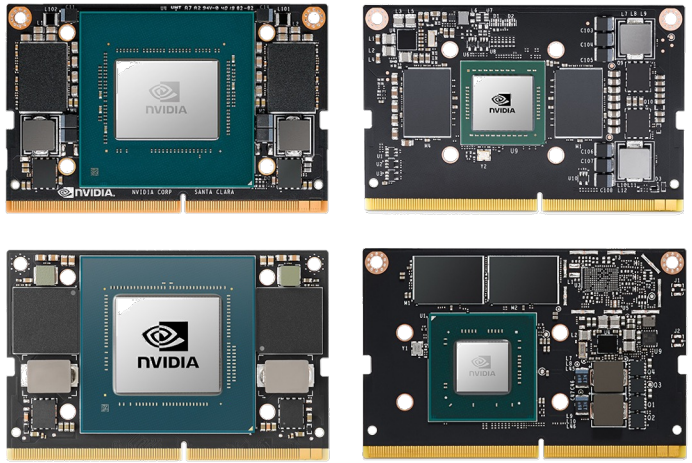
+

NPU



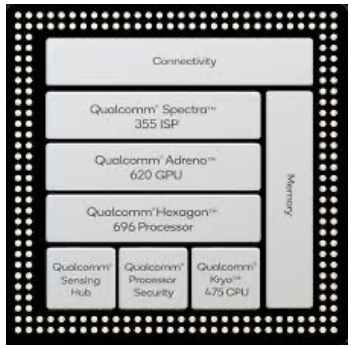
XPU：算力灵活性

- CPU 通用任务
- GPU 并行任务
- TPU 训练任务
- NPU 推理任务
- . . .



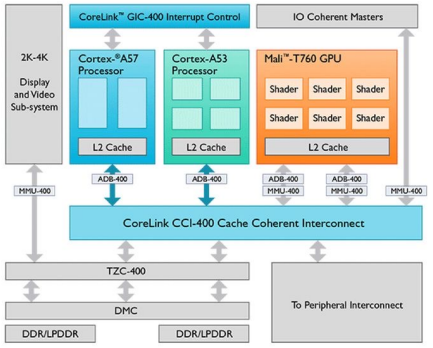
Nvidia Jetson Orin

- CPU : Arm
- GPU : Ampere
- NPU : DLA



Qualcomm AI Engine

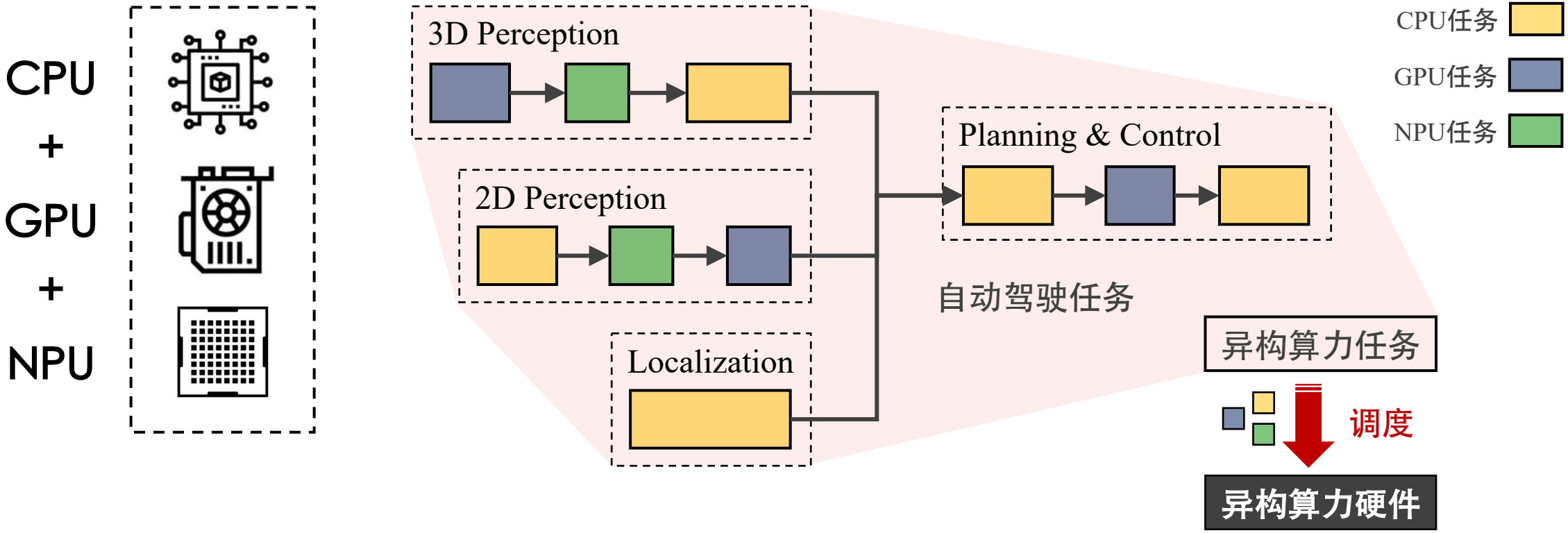
- Hexagon DSP
- Adreno GPU
- Kryo CPU



ARM big.LITTLE

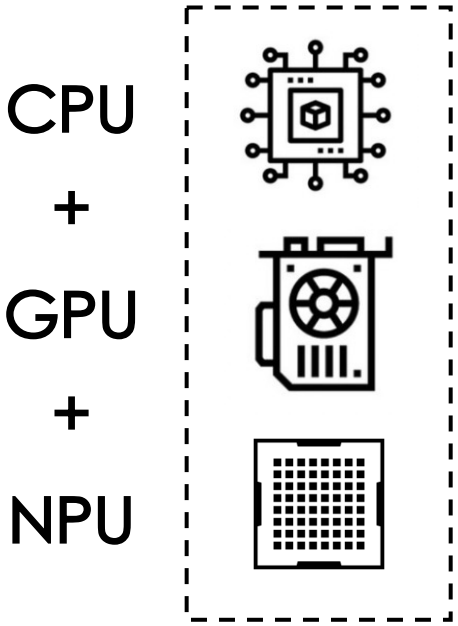
- CPU : Cortex-A57
- CPU : Cortex-A53
- GPU : Mali

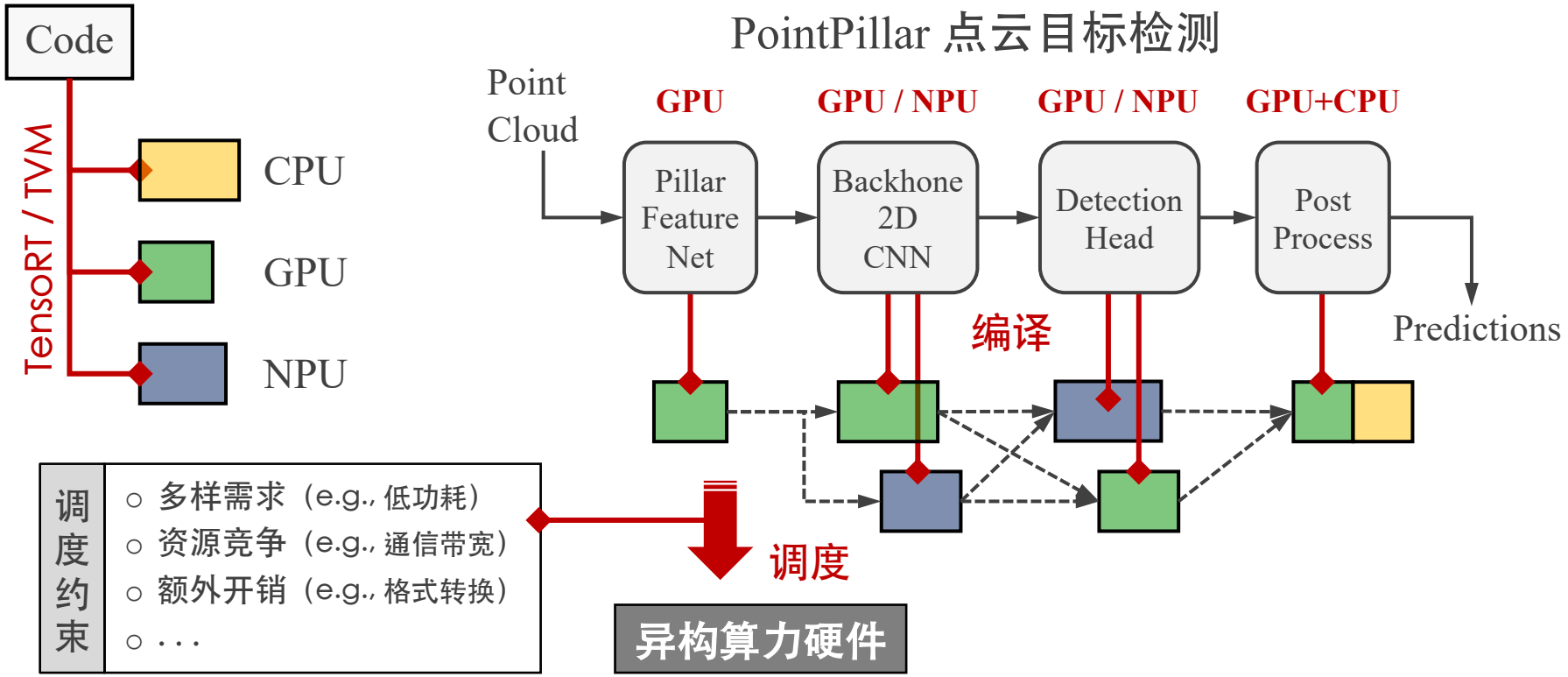
调度挑战 #6：异构调度



调度机制挑战 #6：异构算力任务的协同调度

调度挑战 #7：异构编译





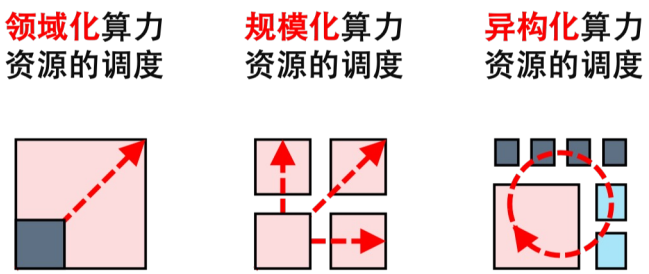
调度机制挑战 #7：面向异构算力硬件的编译+调度

小结 & 感谢

算力需求**爆发式增长** vs. 算力硬件演进呈现**领域化、规模化、异构化**特征

应用的**算力外需求**驱动操作系统在**调度机制**上突破

新算力硬件体系对调度机制造成了多方面的挑战



系统软件研究需要探索“经典”新路径 —— **“上下求索”**

例如：“应用特征反哺” + “硬件特性挖掘”

更多请见：https://ipads.se.sjtu.edu.cn/rong_chen

感谢！