中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

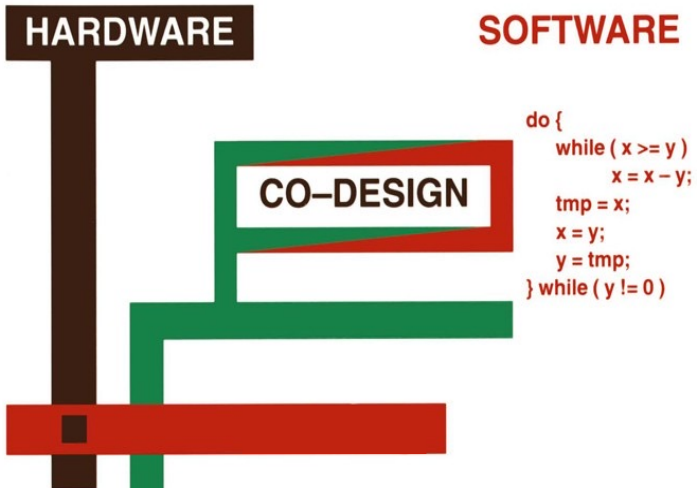
SKLP
处理器芯片全国重点实验室
State Key Lab of Processors, ICT, CAS

面向新型应用范式与新型体系结构的编译技术探索

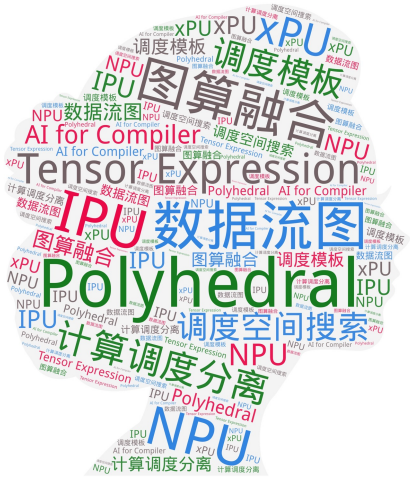
中科院计算所 处理器芯片全国重点实验室
崔慧敏

编译作为桥梁的角色


Fast
Slow


HARDWARE
SOFTWARE
CO–DESIGN
do {
while (x >= y)
x = x – y;
tmp = x;
x = y;
y = tmp;
} while (y != 0)

AI领域知识 + 传统编译模型 = 更多优化机会

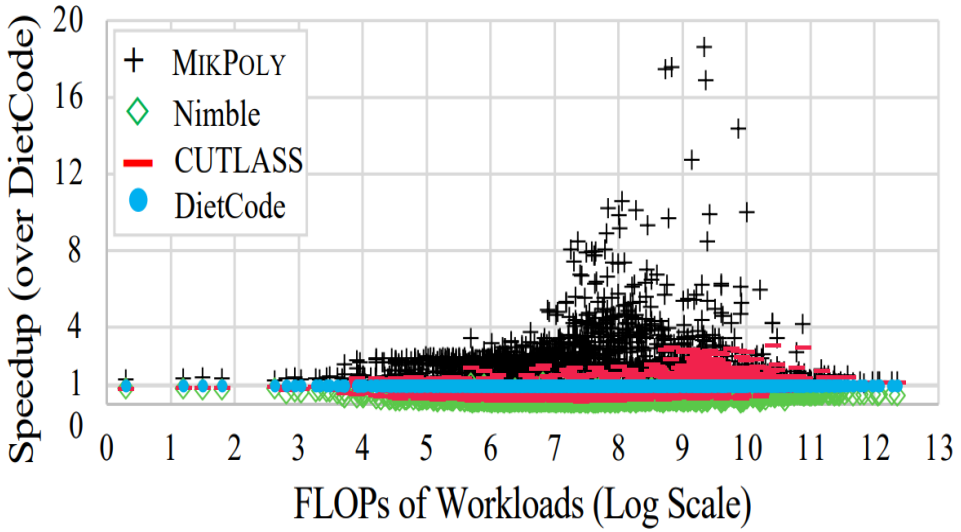


Sirius: 针对AI程序的全程序优化 – MLSYS2023

Souffle: 张量表达式全局分析优化AI应用– ASPLOS2024

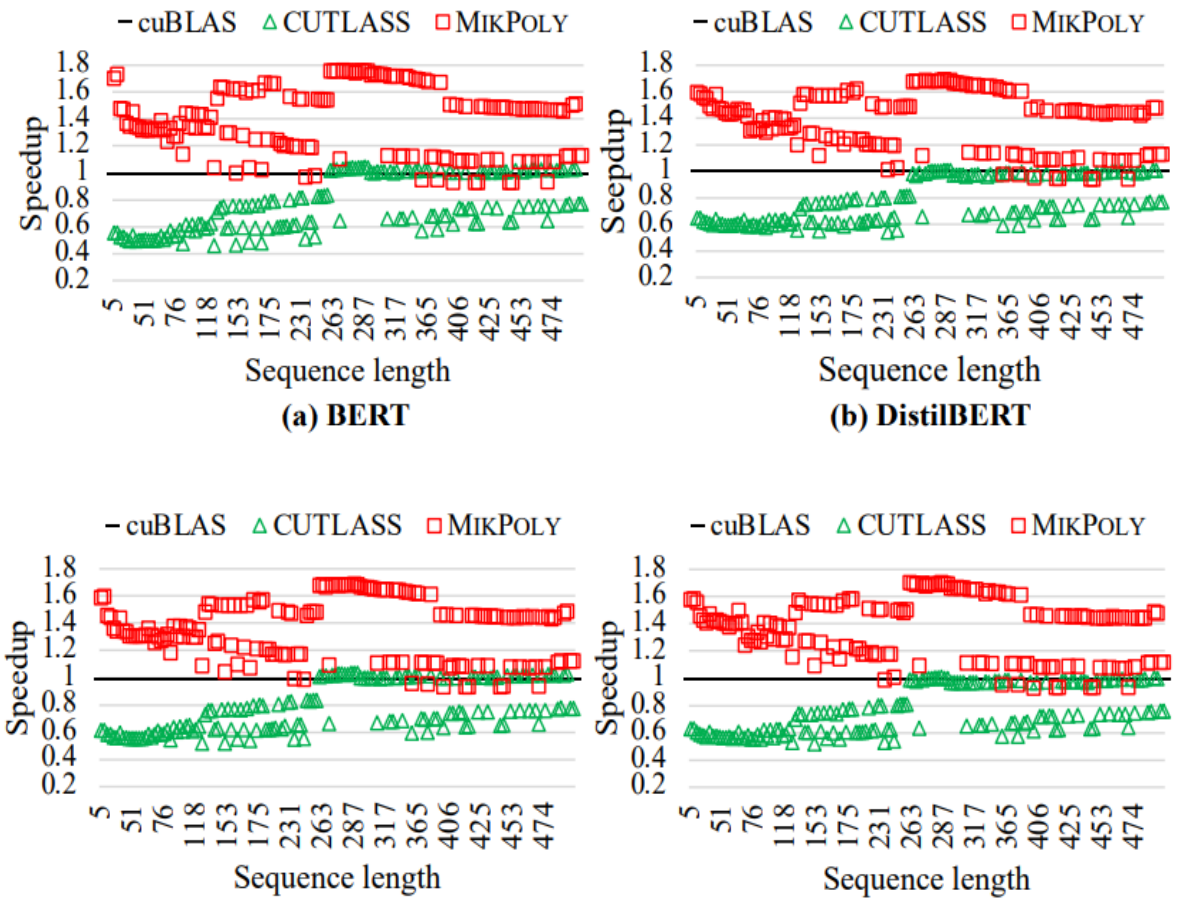
MikPoly: 动态形状矩阵乘法优化 – ASPLOS2024

AI领域知识 + 传统编译模型 = 更多的优化机会



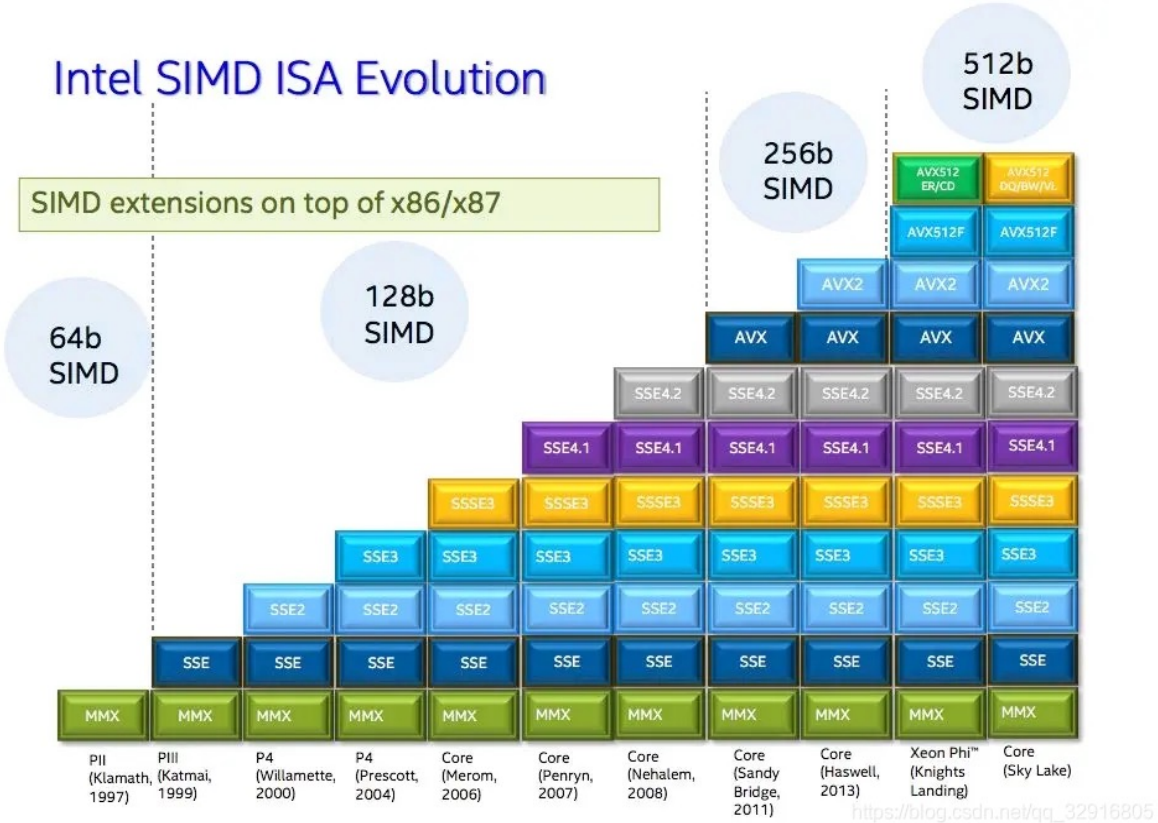
A100平台，平均1.47x over cuBLAS
昇腾平台，平均1.10x over CANN

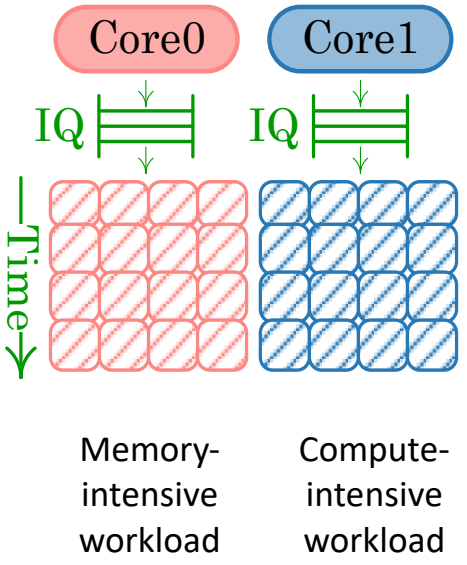
Transformer语言模型-端到端性能

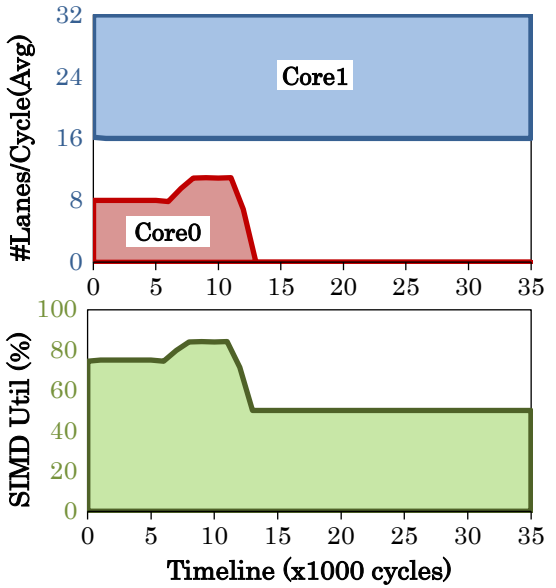


在A100 GPU平台上，端到端NLP模型**平均加速比1.37x-1.39x**

以弹性共享向量单元为例，探索软硬件协同



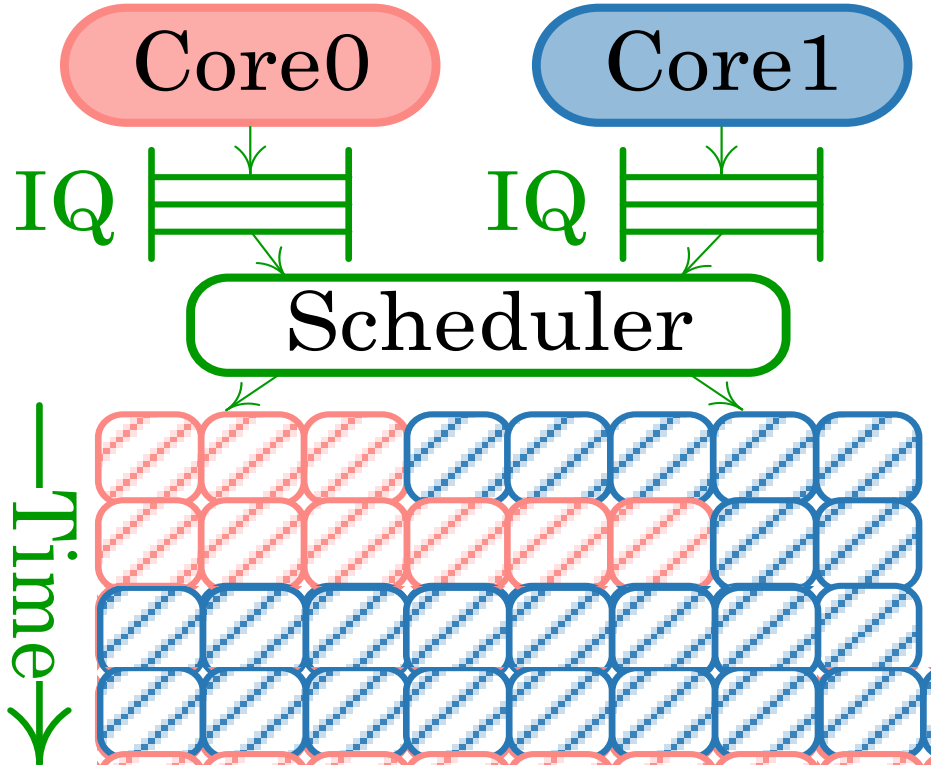


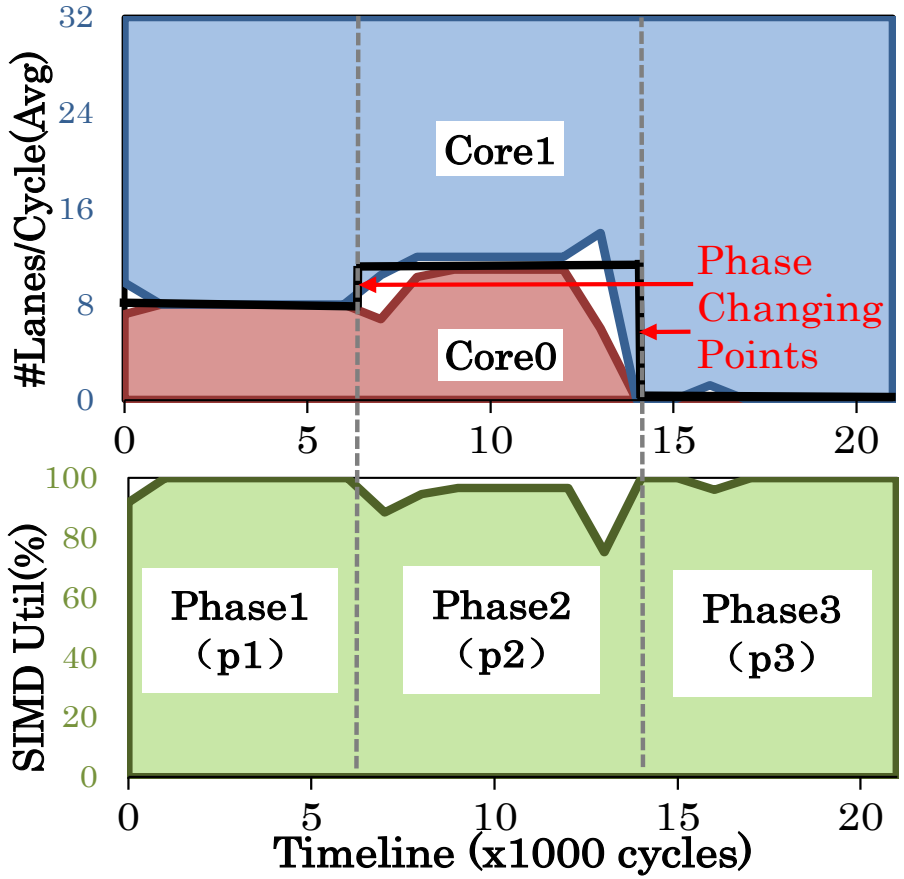


向量单元是CPU设计主要创新方向

传统向量单元为每个核私有，利用率低

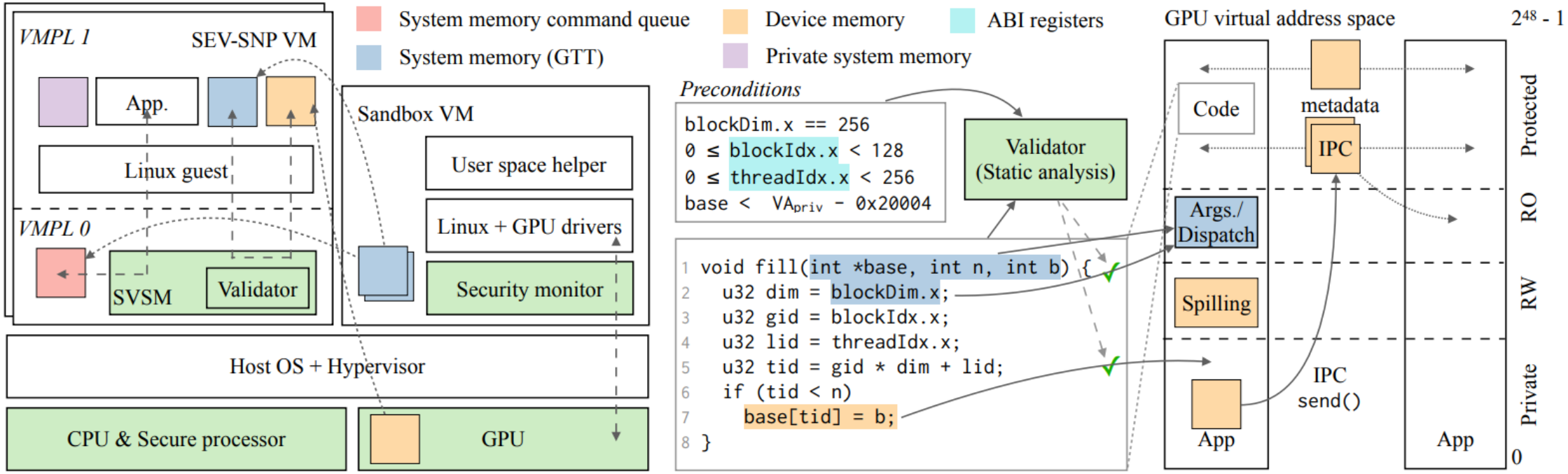
软硬协同的解决方案





能够及时感知软件行为变化，并按需调整资源使用量

异构加速器的安全性问题



- **基于CPU TEE构建GPU TEE**
  - CPU TEE (AMD SEV/SNP)
  - GPU被隔离在单独的VM中
  - Security Monitor检查所有CPU-GPU交互

- **编译验证GPU Kernel代码，保证行为边界**
  - 无悬空访问
  - 控制流完整
  - 应用程序访存安全属性：

谢谢！ Q&A

谢谢！ Q&A

谢谢！ Q&A