神经网络的异构推理

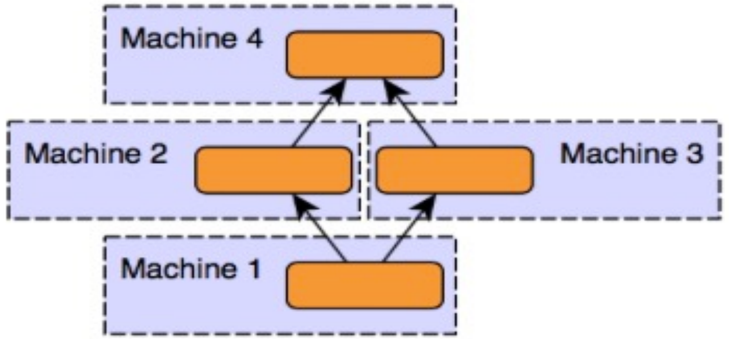
万物互联下的端侧智能模式

Chao Xing/2022.03.06

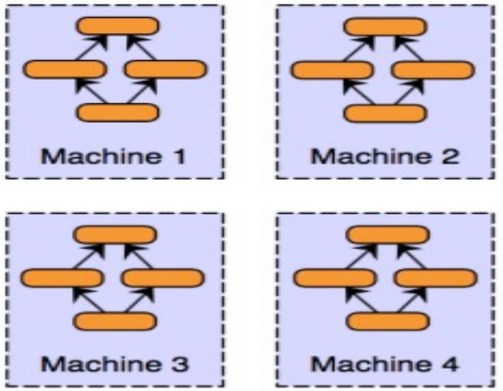
分布式神经网络

1. 分布式神经网络训练

1.1 模型并行

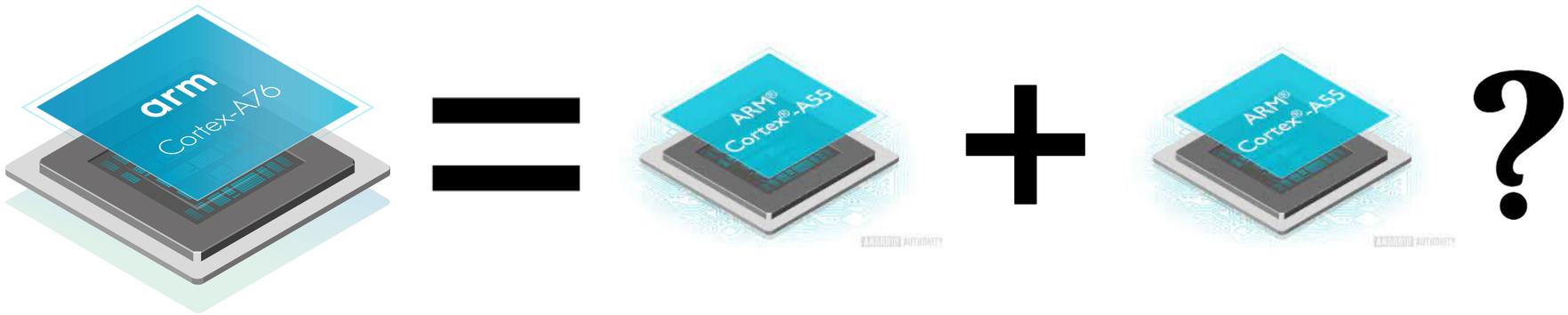


1.2 数据并行



2. 分布式神经网络推理

2.1 为什么需要研究分布式神经网络推理?
2.2 如何进行分布式神经网络推理?
2.3 分布式神经网络推理的示例



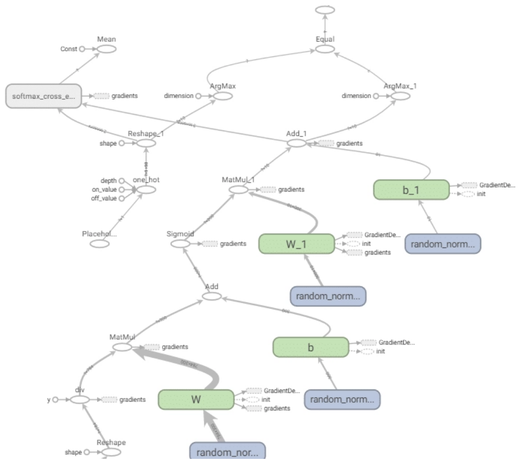
分布式神经网络推理

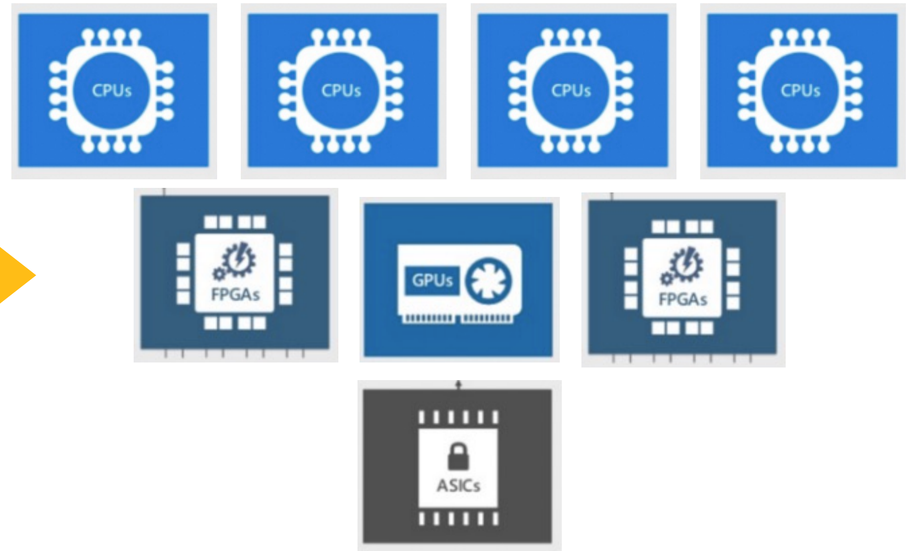
2.1 为什么需要研究分布式神经网络推理?

优化多设备连接下的AI模型推理性能

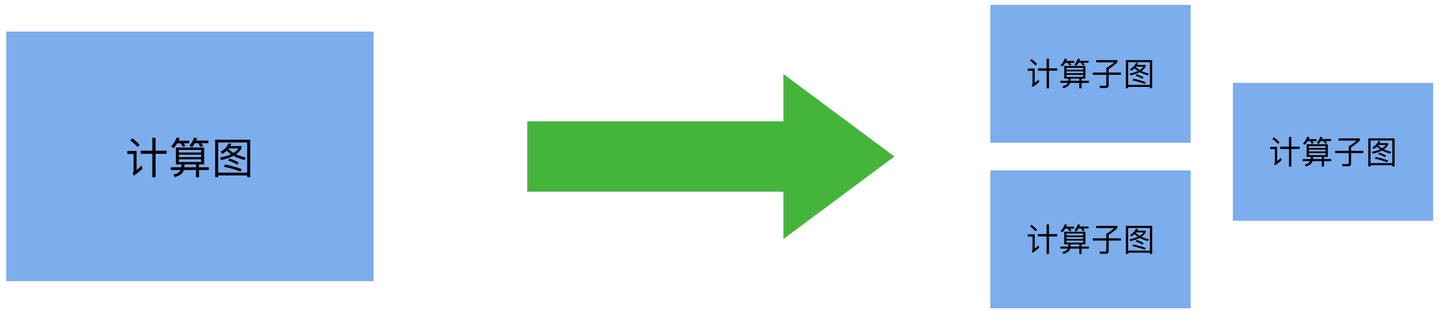
优化单设备多核CPU下的AI模型推理性能

优化单设备多种计算资源下的AI模型推理性能



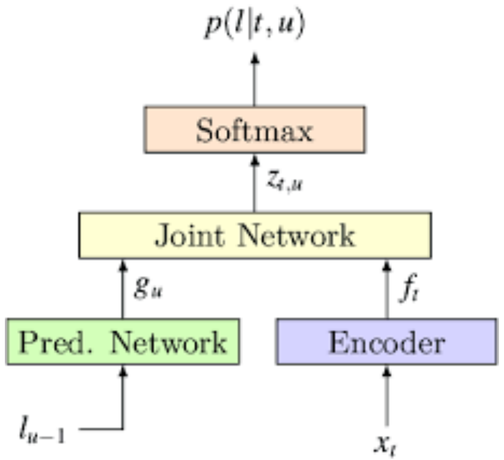


2.2 如何进行分布式神经网络推理?

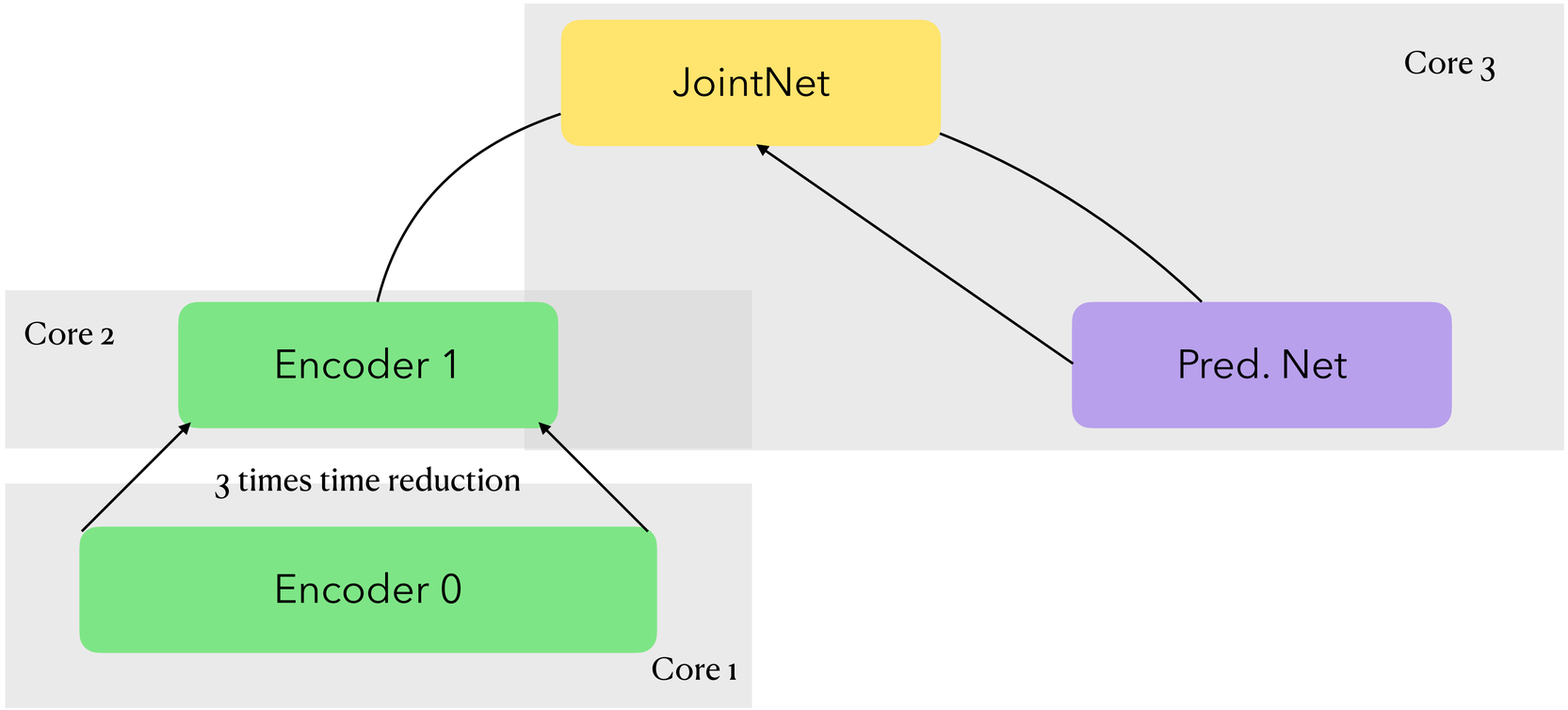


分布式神经网络推理

2.3 分布式神经网络推理的示例

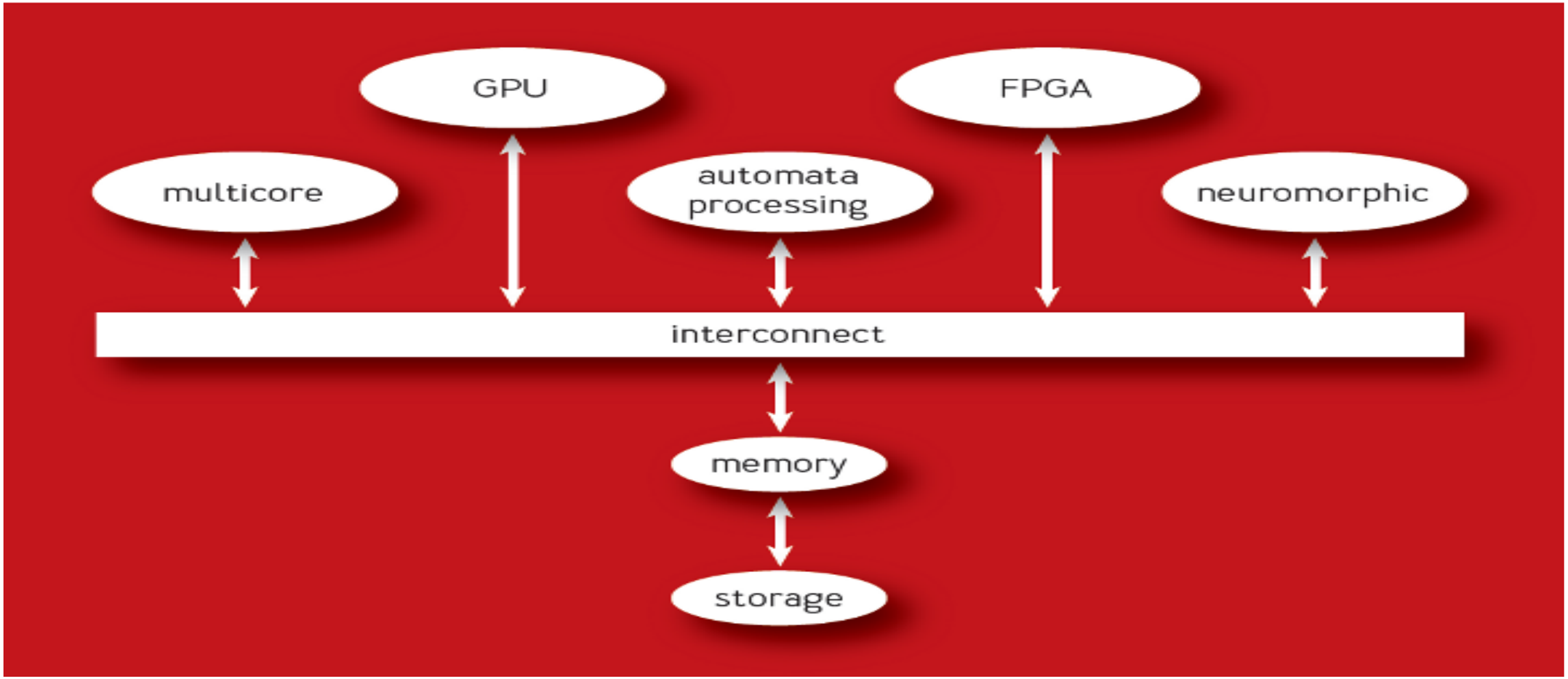


Google RNN-T

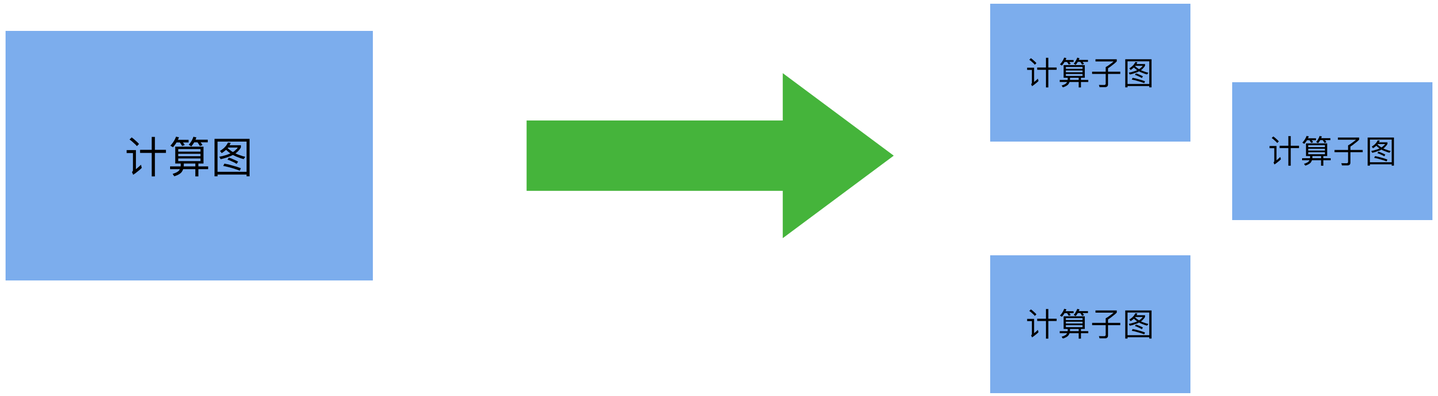


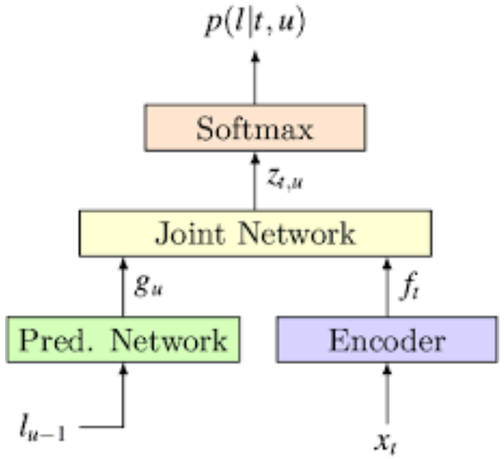
异构计算设备

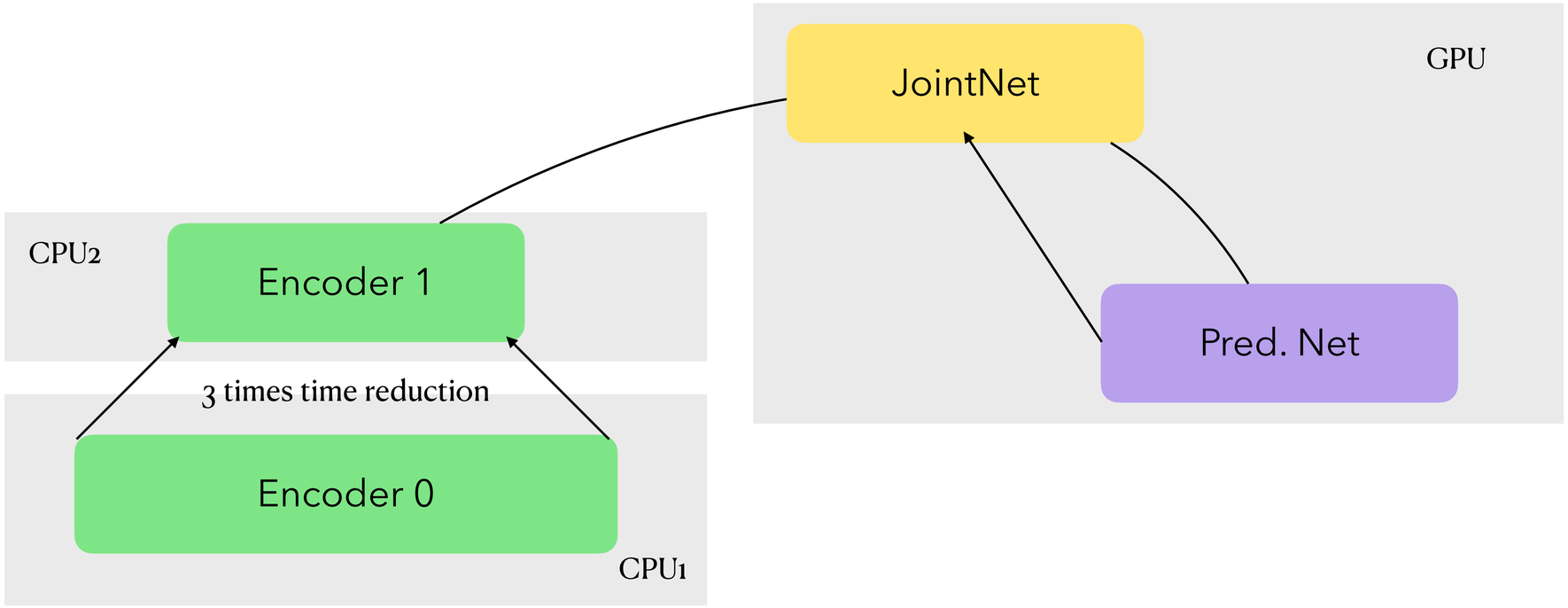
FIGURE 1: **GENERIC HETEROGENEOUS SYSTEM**



神经网络异构推理


计算图
计算子图
计算子图
计算子图


$p(l|t,u)$
Softmax
$z_{t,u}$
Joint Network
g_u
f_t
Pred. Network
Encoder
l_{u-1}
x_t

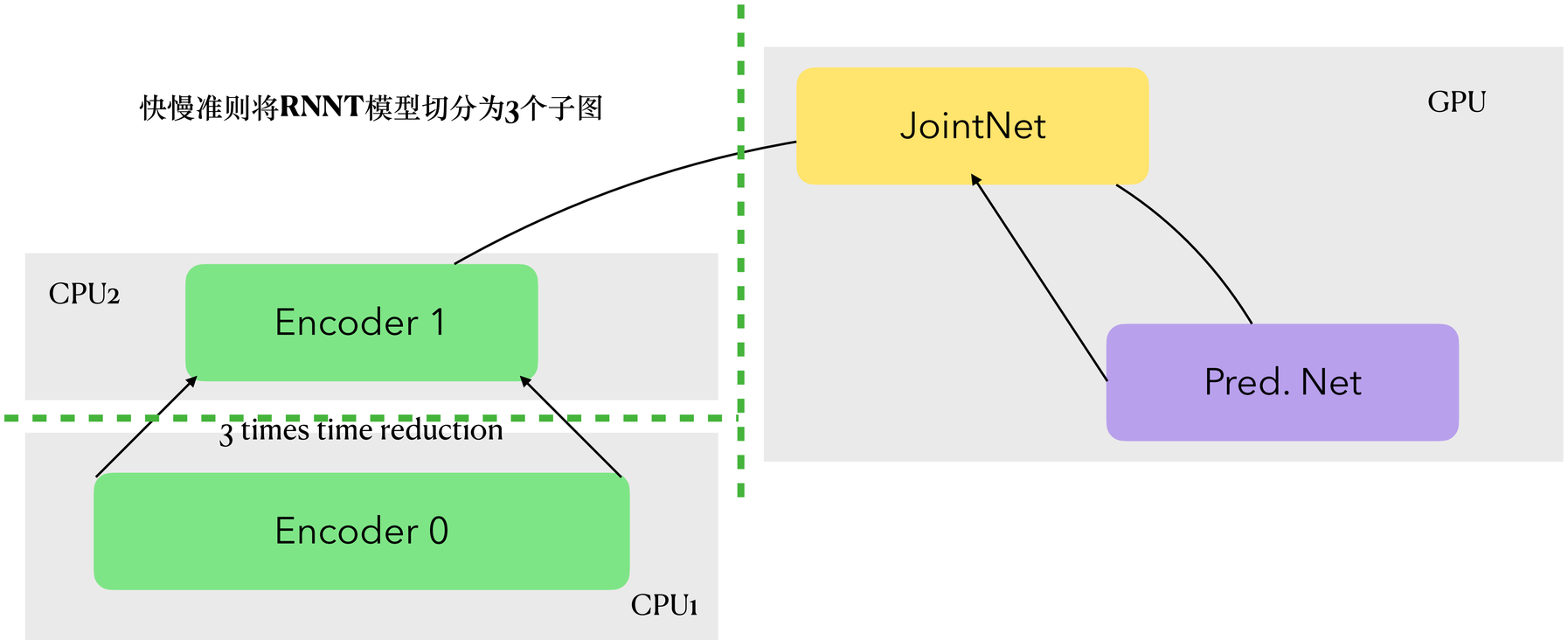

GPU
JointNet
Pred. Net
CPU2
Encoder 1
3 times time reduction
Encoder 0
CPU1

神经网络异构推理

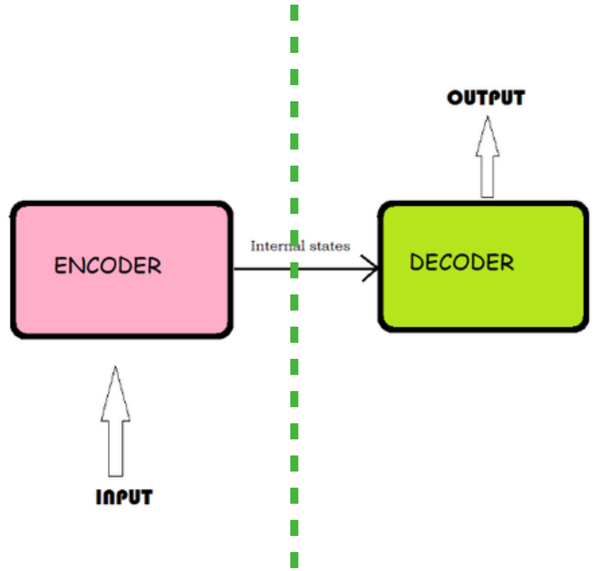
软硬一体化设计的可切分神经计算图切分准则：

准则1. 快慢准则：快慢准则是指切分计算图时，需要依据不同模块中计算频率的快慢来将计算图的不同模块切分成不同的子图

快慢准则将RNNT模型切分为3个子图



快慢准则将seq2seq模型切分为2个子图，其中encoder子图驱动一次，而decoder子图驱动多次



神经网络异构推理

软硬一体化设计的可切分神经计算图切分准则：

准则2. 运转效率优先准则：运转效率优先准则是指将不同计算子图驱动在不同的计算资源上时，由于数据传输带来的额外开销不应该高于单设备独立驱动完整计算图时的开销

准则3？？？

未来&讨论

- 未来的端侧计算节点一定含有多组异构计算资源，在这种异构计算资源的约束下，我们应该如何设计和实现新型神经网络结构来适配相关异构计算资源?
- 多模态神经模型或双塔结构的神经网络模型是天然适合使用异构计算资源驱动的，这种双塔式的神经网络应该如何设计与适配连接的多设备以及不同设备内的异构计算资源?
- ....