DeepCPU:Serving RNN-based Deep Learning Models 10x Faster

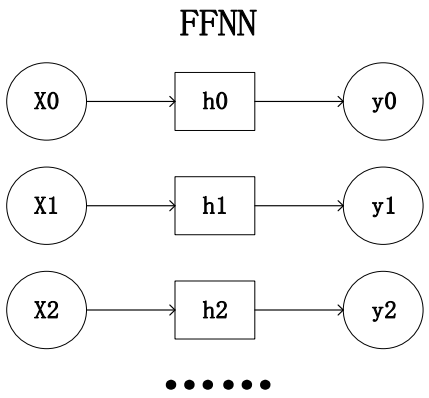
DeepCPU：使基于RNN的深度学习模型的服务速度提高10多倍

国光硕1807班
M201872992
张 超

FFNN and RNN

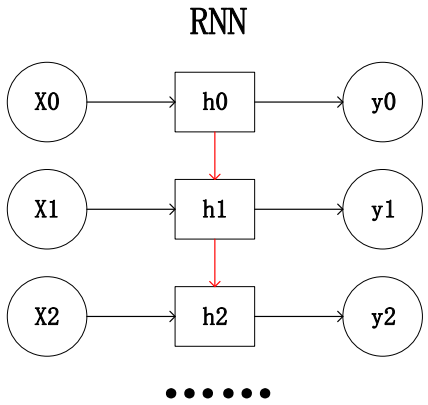
- 什么是神经网络?
- 前馈神经网络FFNN
- 循环神经网络RNN

序号	图片 x	类别 y
第 1 张图片		惊讶
第 2 张图片		晕菜
第 3 张图片		尴尬
:	:	:



- 一劳永逸? 

- 自然语言处理
- 机器翻译
- 机器阅读理解



本文的工作

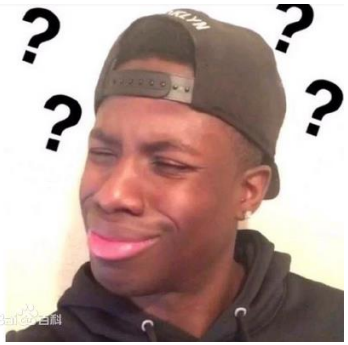
- 发现问题，传统的RNN服务的弊端 ✓
- 解决问题，设计并实现DeepCPU（一种更快的、运行在CPU上深度学习服务库）✓
- 实验测试，DeepCPU效果更好 ✓
- 结论&总结 ✓

目录

- 问题&分析
- DeepCPU，如何解决问题
- 实验，DeepCPU pk CPU&GPU
- 总结

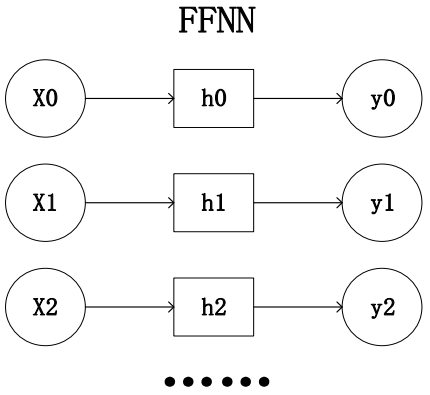
问题

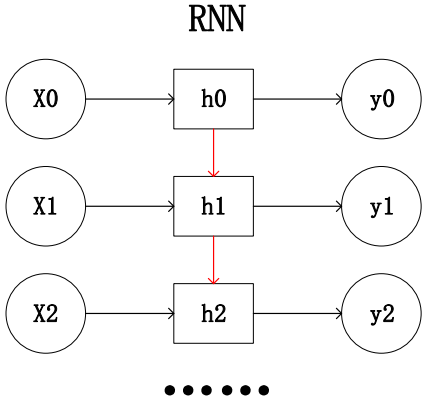
- 传统RNN服务的表现（SLA服务等级协议）
- 普通CPU计算RNN时性能<峰值的2%（以Xeon E5-2650为例讨论flops）



分析为什么这么慢

- CPU的计算≈矩阵相乘
- 数据重用性差
- 优化矩阵相乘的划分
- 优化并行计算



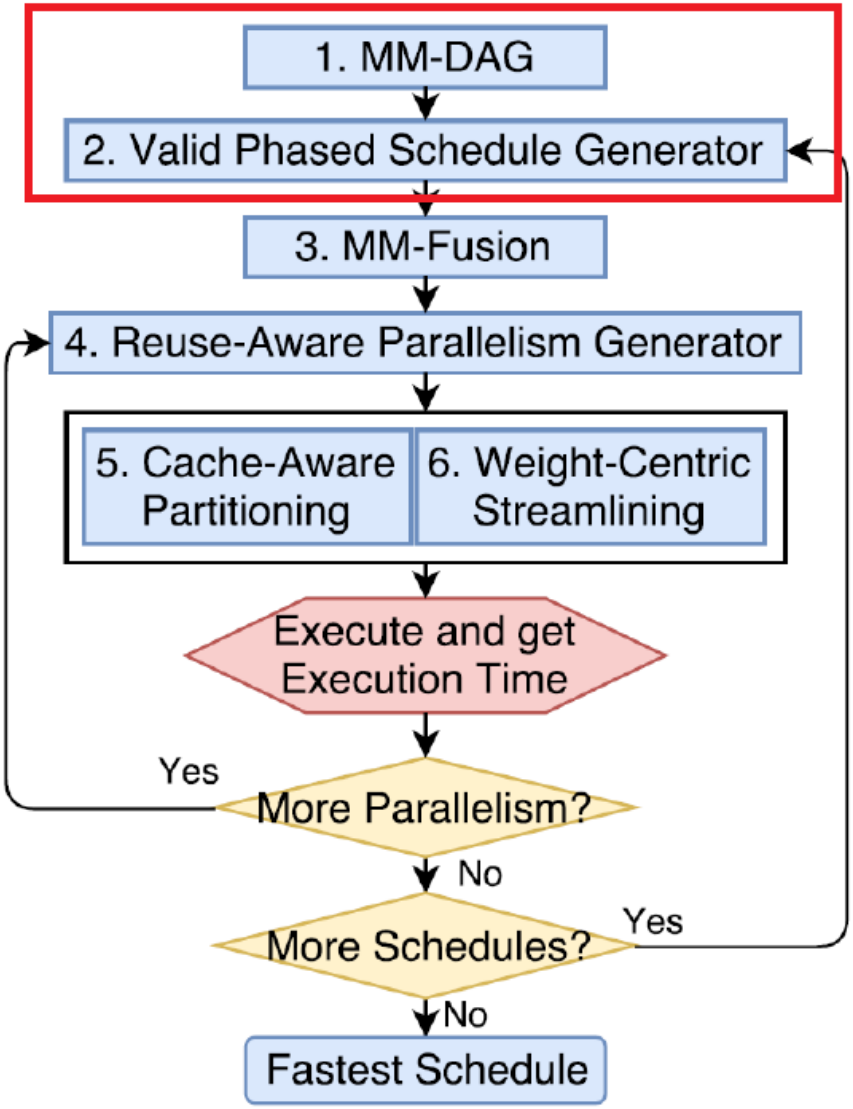


挑战&策略

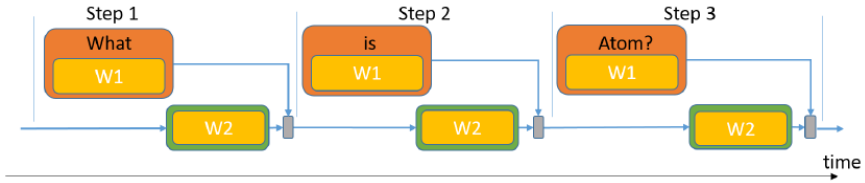
- 计算空间大
- 寻求最优RNN执行
- 硬件要求不能太高
- 参数的影响

- 搜索空间
- MM-DAG（矩阵相乘有向无环图）
- 整体&局部优化

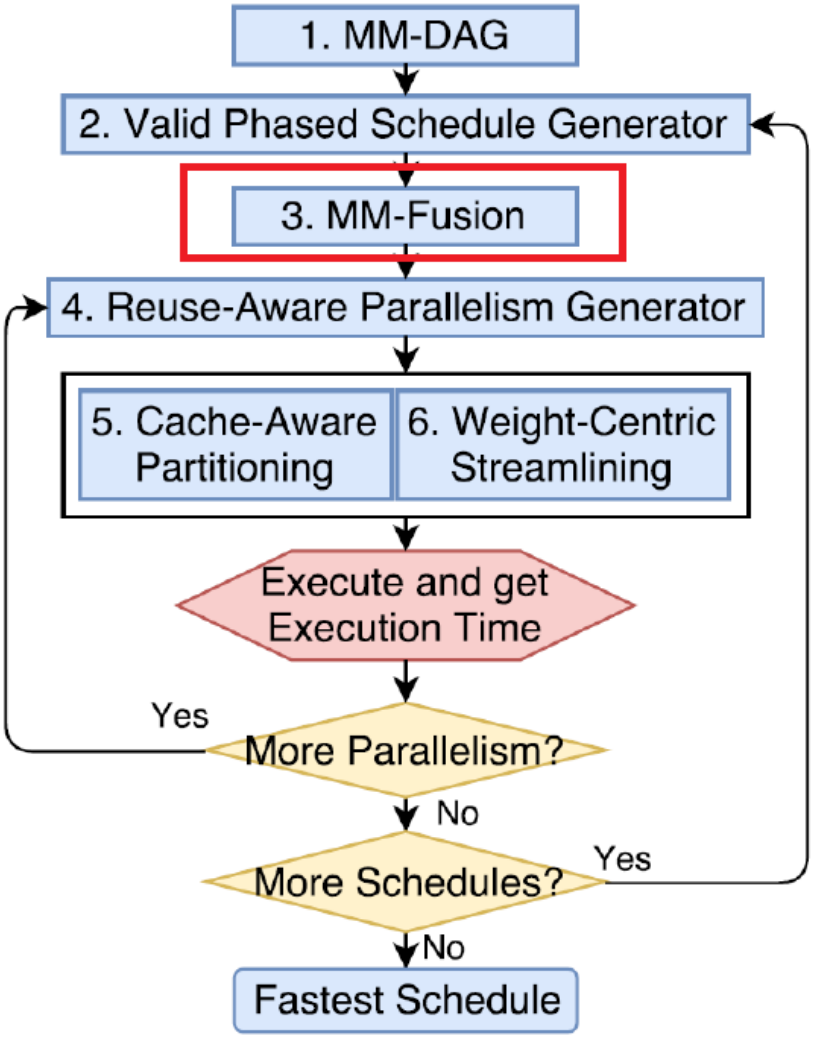
DeepCPU ， 1~2



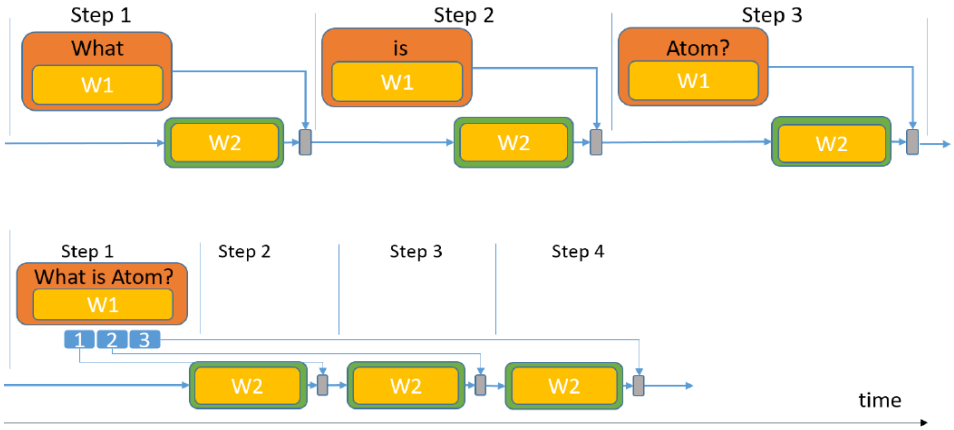
- DeepCPU
- MM-DAG
- 有效阶段生成器



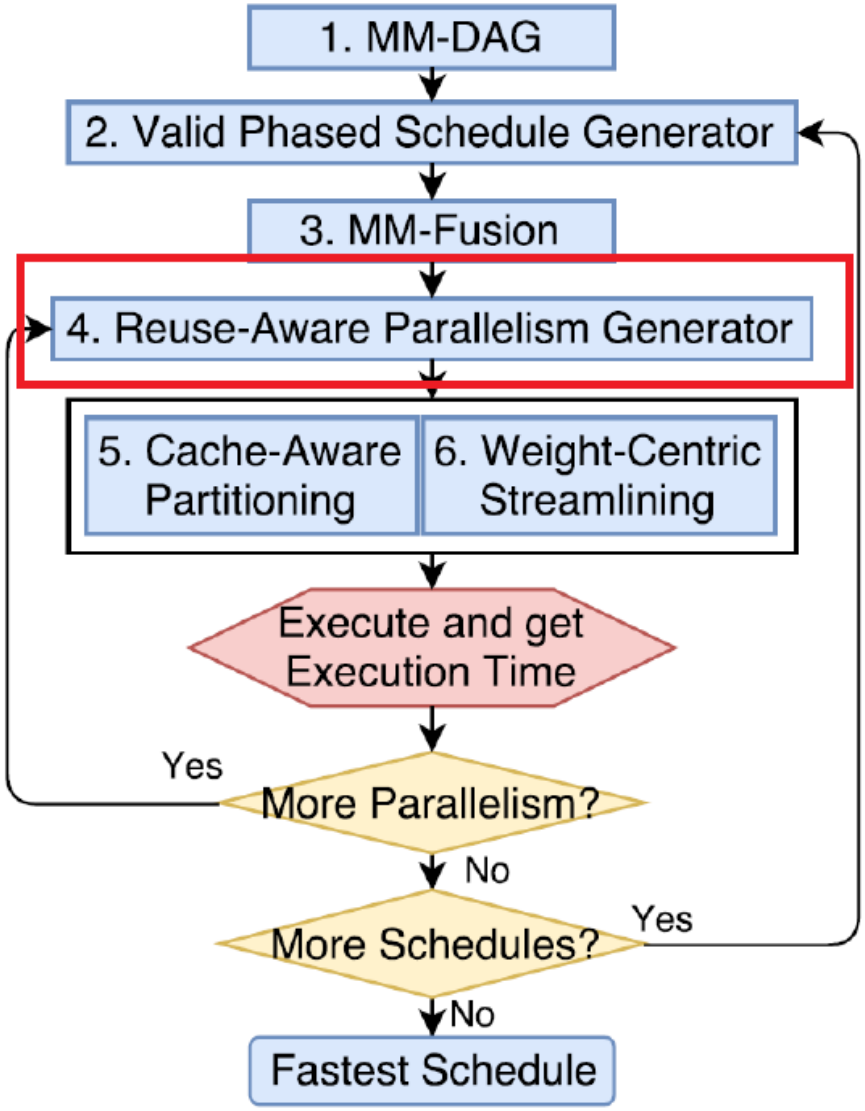
DeepCPU ， 3



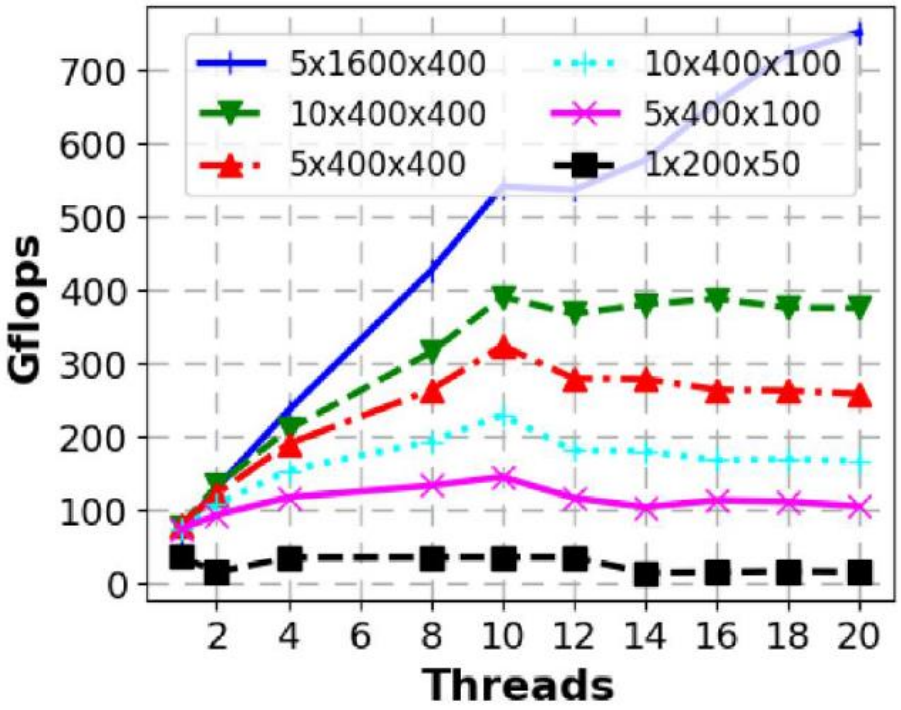
- DeepCPU
- 矩阵相乘的合并



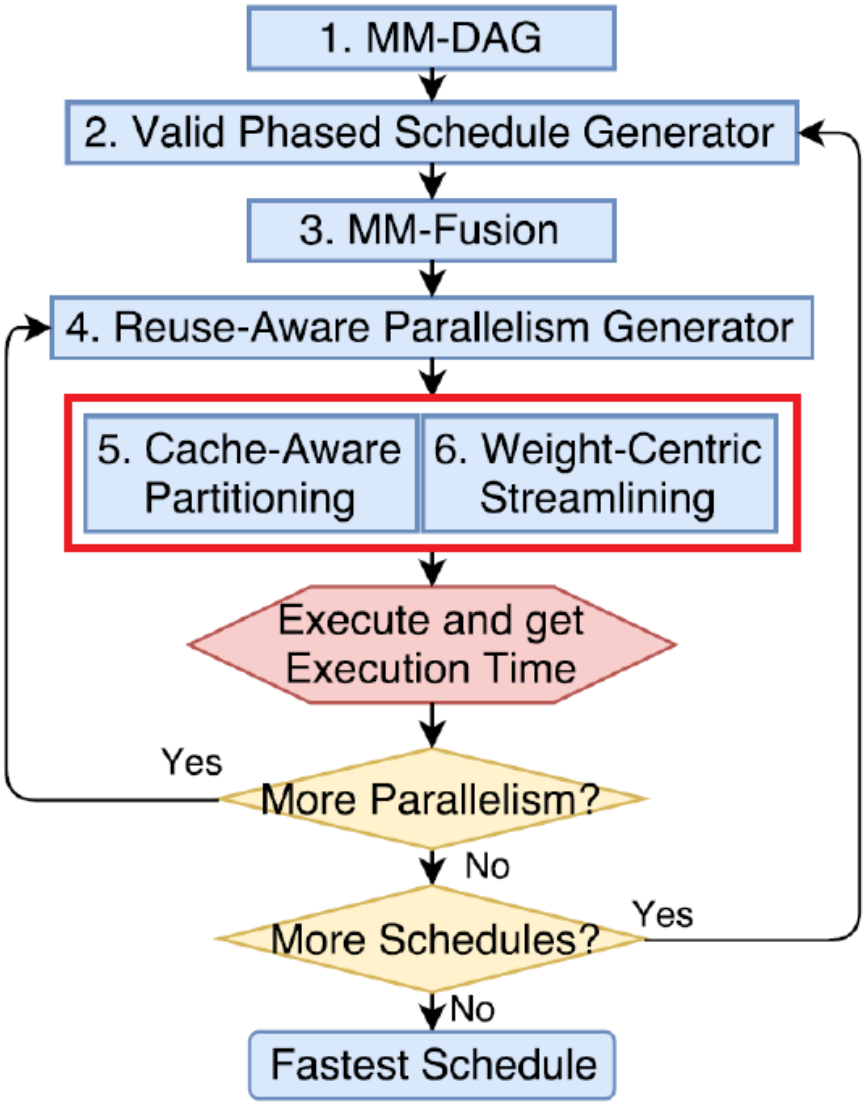
DeepCPU ， 4



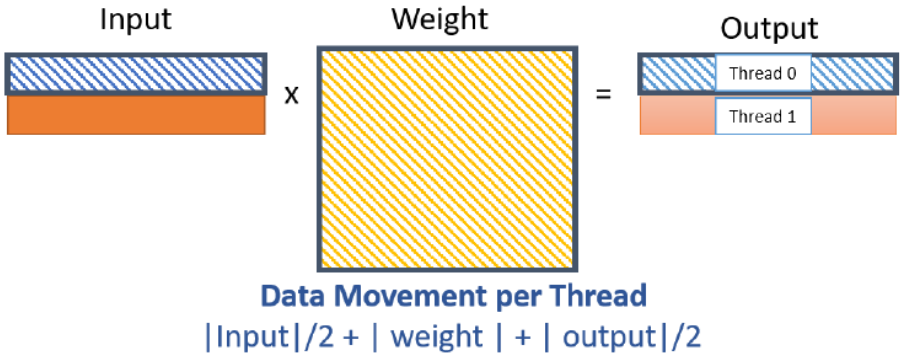
- DeepCPU
- 并行计算生成器

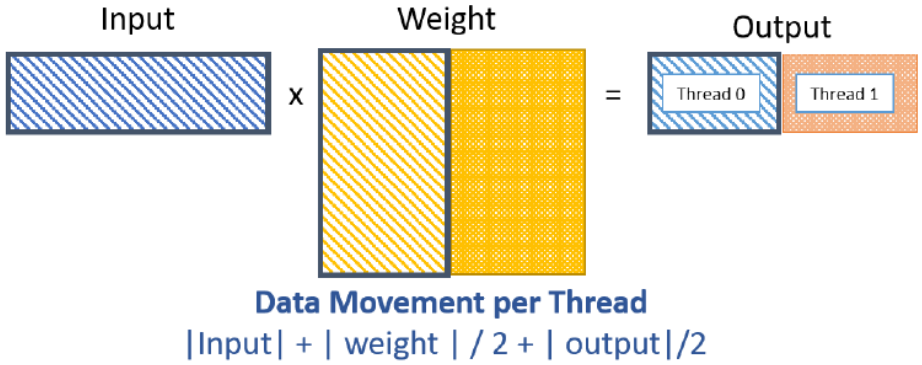


DeepCPU ， 5

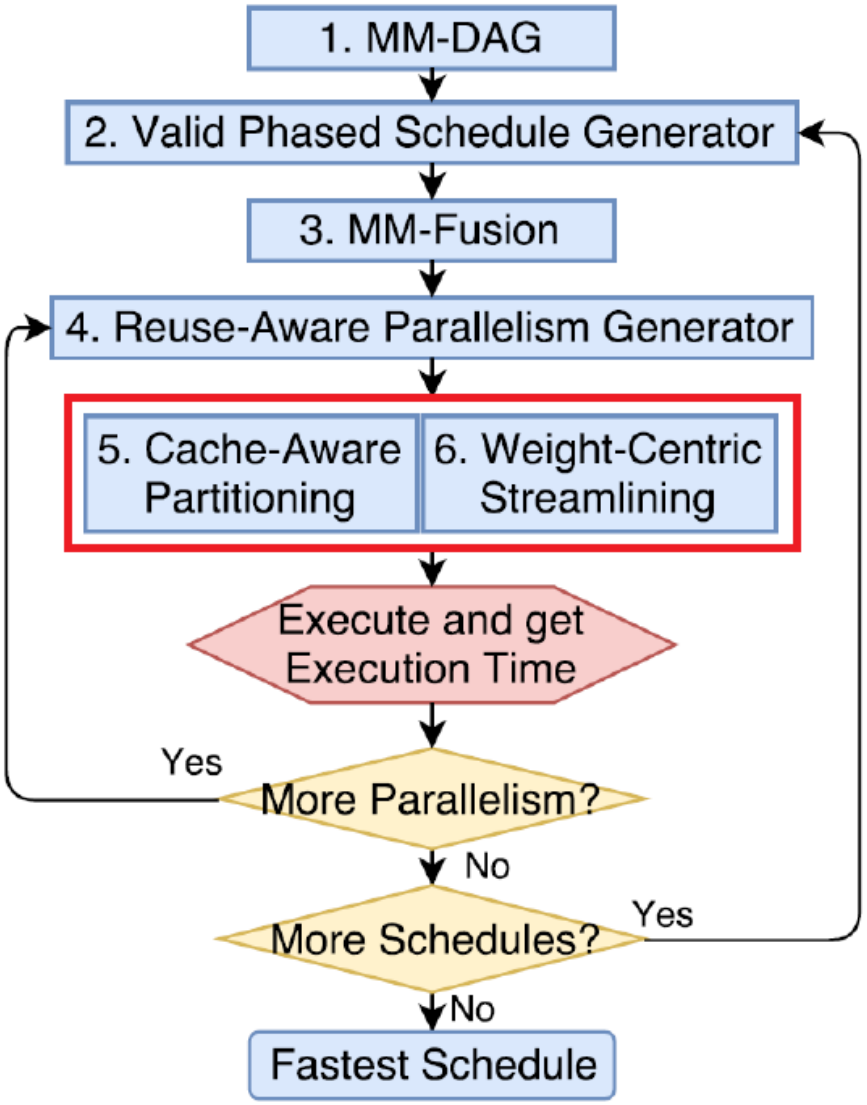


- DeepCPU
- 私有缓存分区

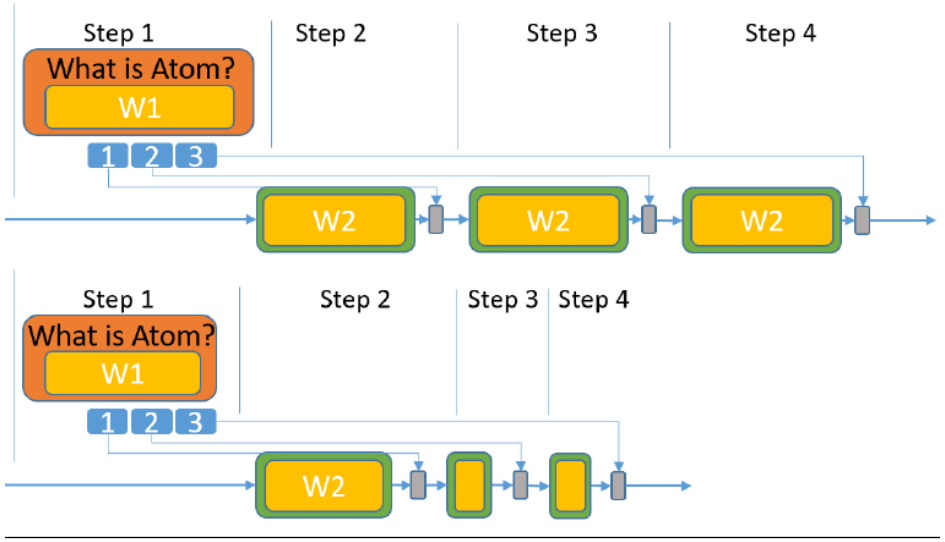




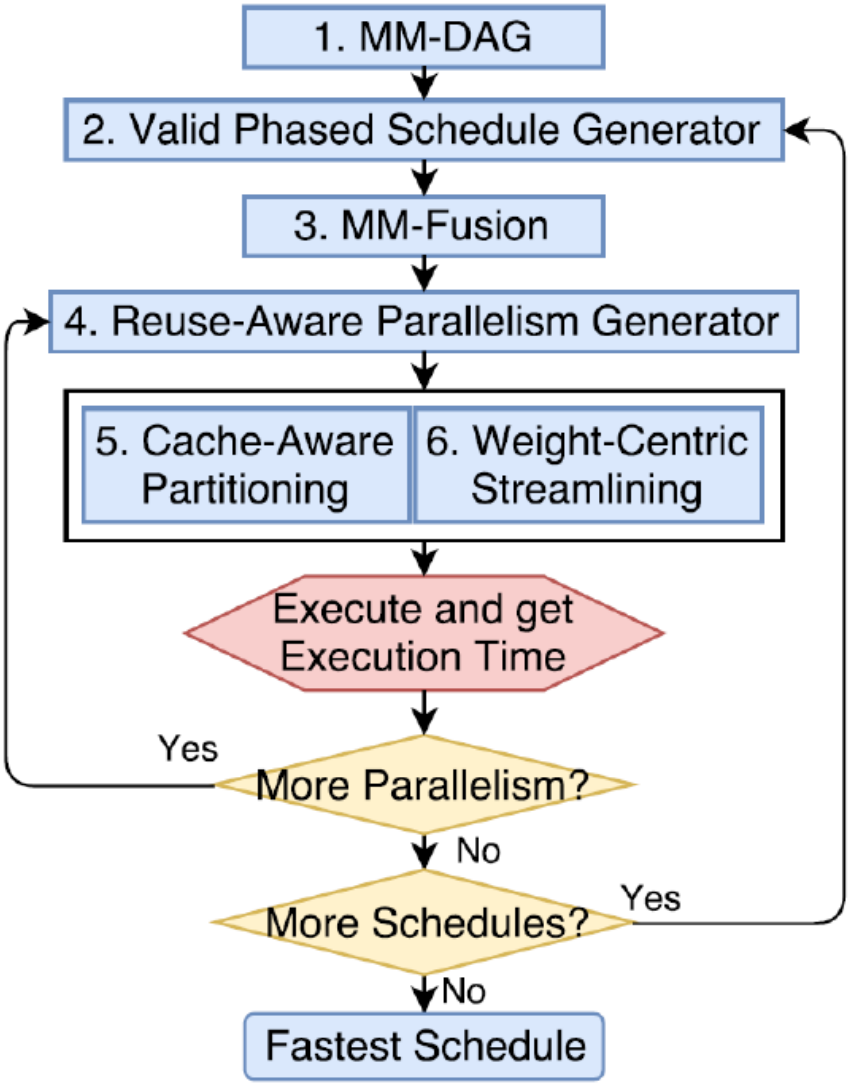
DeepCPU ， 6



- DeepCPU
- 权重精简



DeepCPU，总结



- 循环校验
- 一次构建，重复使用

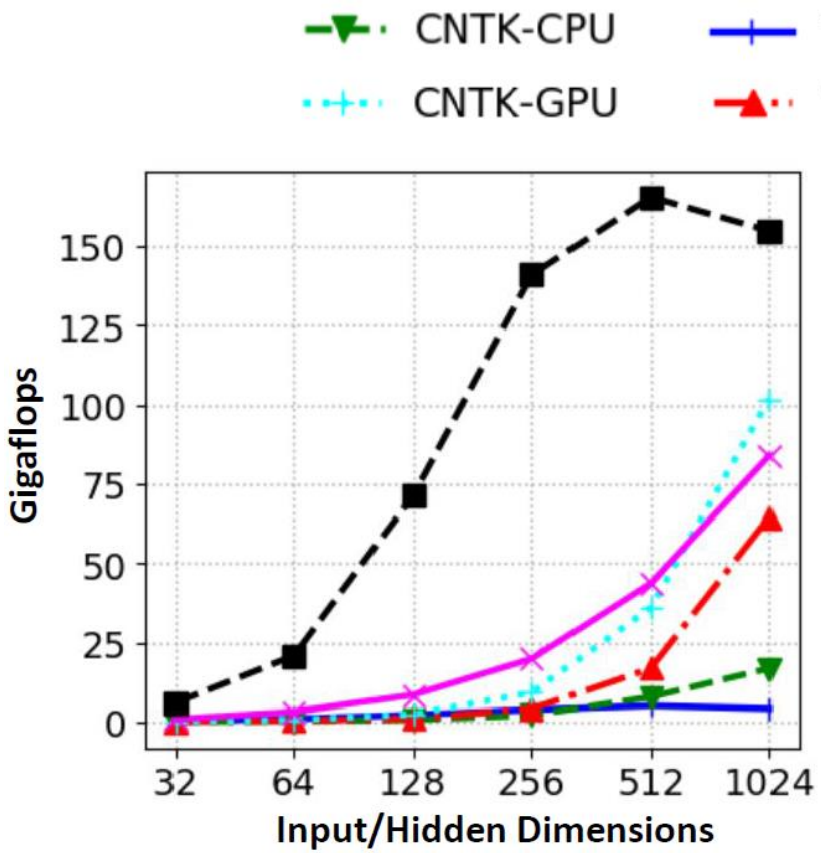
DeepCPU pk CPU

Performance : DeepCPU vs TF vs CNTK

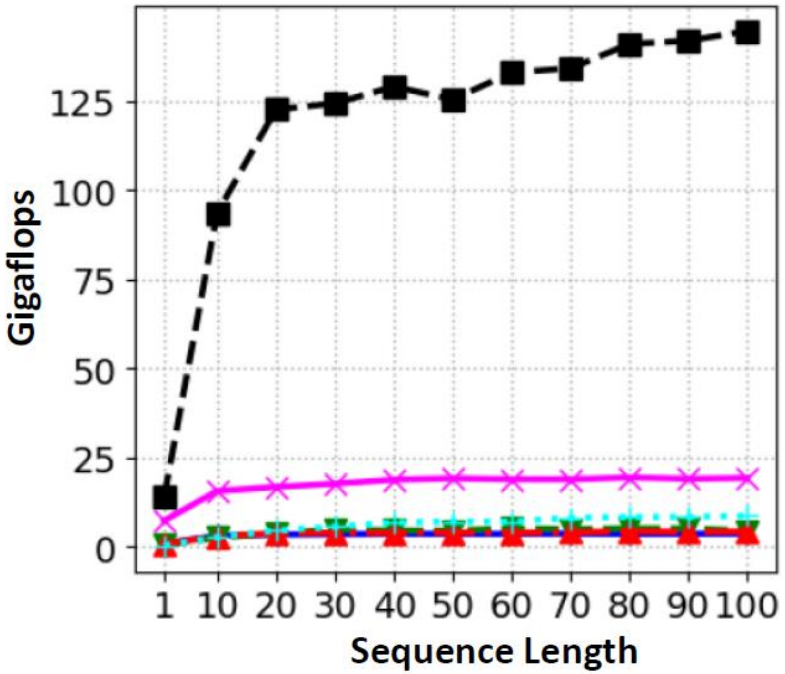
- Average LSTM speedup
  - **DeepCPU is 23x faster** than Tensorflow
  - **DeepCPU is 31x faster** than CNTK
- Average GRU speedup
  - **DeepCPU is 16x faster** than Tensorflow
  - **DeepCPU is 25x faster** than CNTK

Model parameters			LSTM speedup		GRU speedup	
input	hidden	seq. len.	TF	CNTK	TF	CNTK
64	64	100	26	81	11	36
256	64	100	34	93	17	45
1024	64	100	45	60	23	39
64	256	100	34	37	22	38
64	1024	100	28	4.6	17	5.8
1024	1024	100	42	3	23	4.8
256	256	1	14	16	17	19
256	256	10	21	18	21	24
256	256	100	38	28	24	28

DeepCPU pk GPU



Batch Size = 1, Sequence Length = 100



Batch Size = 1, Input/Hidden Dimension = 256

总结

- DeepCPU，一种针对RNN的深度学习库
- 使RNN在CPU上的服务速度提高了10x，进而使得在线提供某些RNN服务由不可能变成可能
- 使CPU在RNN服务上打败GPU
- 节约了机器成本，相同的服务，最初要使用10k台机器，现在减少到几百个，节约了数百万美金的成本

Thank You!