Tensor Processing Unit

——国防科大2020年高性能评测与优化课程小组讨论

报告人：龙　健

组　员：张树晓　龙　健

指导：龚春叶、甘新标、杨博

主要内容

1、需求分析
2、技术方案
3、效果（与CPU、GPU性能比较）
4、TPU可能的设计及对比分析
5、结论与思考

需求分析

- 神经网络的兴起对高性能计算的需求
- 必须通过特定领域的硬件来提高性价比

需求分析--神经网络（NN）介绍

- 神经网络的兴起
- 神经网络的目标
- 神经网络的两个阶段
- 三种神经网络：
  - 1. 多感知器(MLP)
  - 2. 卷积神经网络(CNN)
  - 3. 递归神经网络(RNN)

Name	LOC	Layers					Nonlinear function	Weights	TPU Ops / Weight Byte	TPU Batch Size	% of Deployed TPUs in July 2016
		FC	Conv	Vector	Pool	Total					
MLP0	100	5				5	ReLU	20M	200	200	61%
MLP1	1000	4				4	ReLU	5M	168	168	
LSTM0	1000	24		34		58	sigmoid, tanh	52M	64	64	29%
LSTM1	1500	37		19		56	sigmoid, tanh	34M	96	96	
CNN0	1000		16			16	ReLU	8M	2888	8	5%
CNN1	1000	4	72		13	89	ReLU	100M	1750	32	

需求分析--特定领域的硬件来提高性价比

- TPU的来源

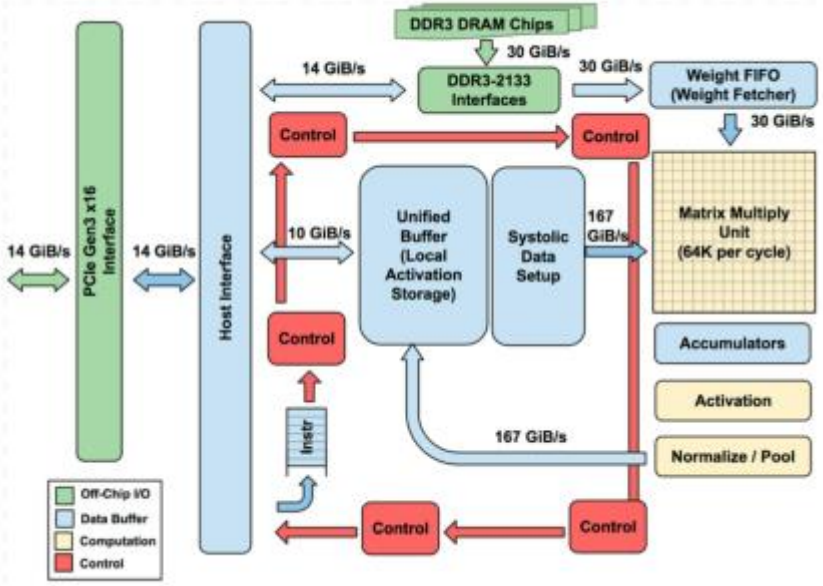
  定制ASIC芯片

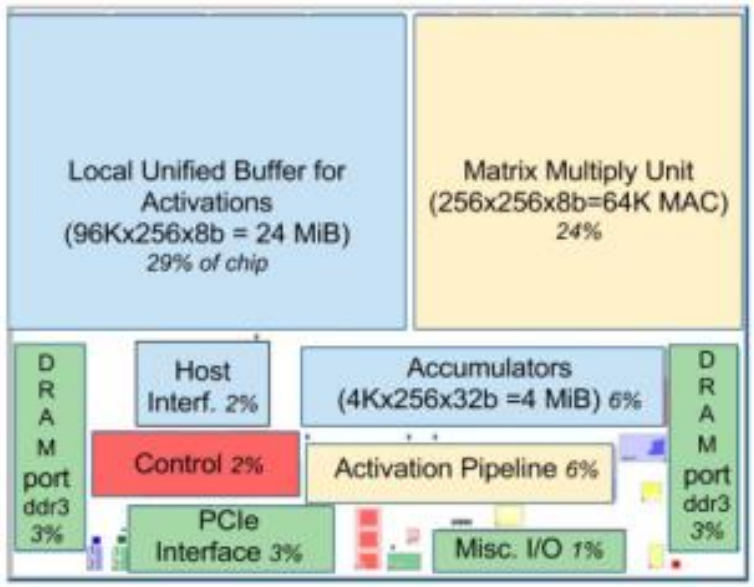
  性能提高10倍

  协处理器

TPU设计的技术方案

- TPU的架构



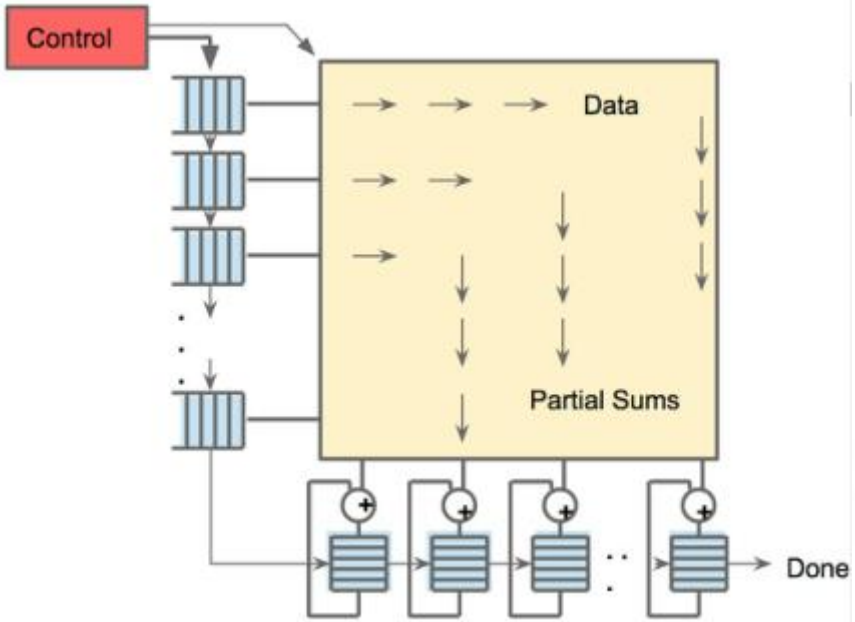


TPU设计的技术方案

- TPU的指令
  - Read_Host_Memory
  - Read_Weights
  - MatrixMultiply/Convolve
  - Activate
  - Write_Host_Memory

TPU设计的技术方案

- TPU的数据流



效果（与CPU、GPU性能比较）

- CPU、GPU和TPU平台
  - TPU：基准平台是2015年部署TPUs时可用的服务器级计算机
  - CPU：CPU服务器由Intel的18核双插槽Haswell处理器
  - GPU：GPU加速器是Nvidia K80

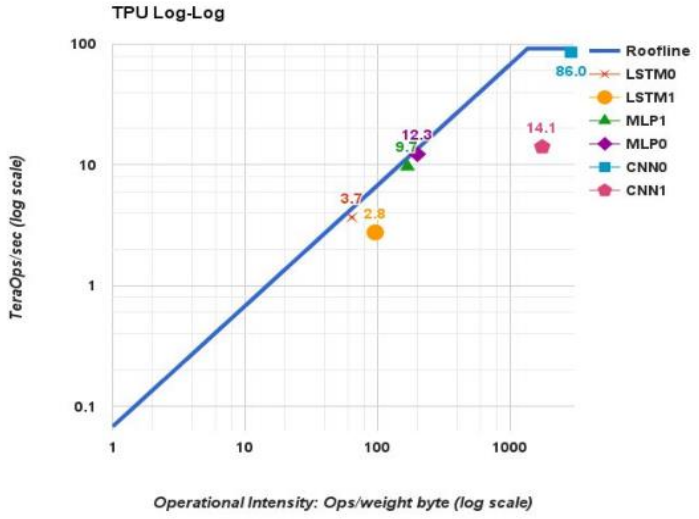
效果（与CPU、GPU性能比较）

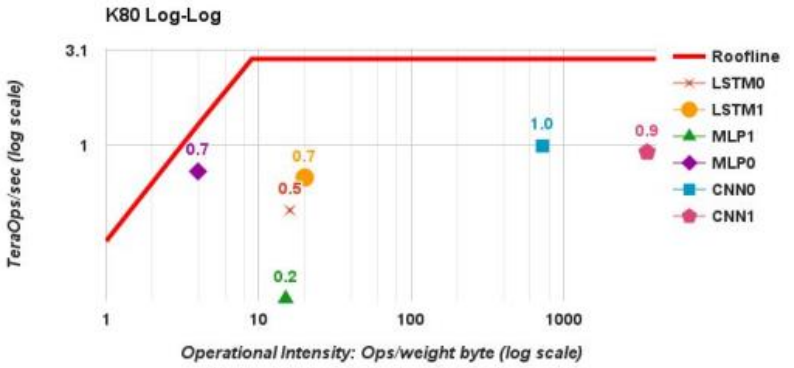
- 测试平台

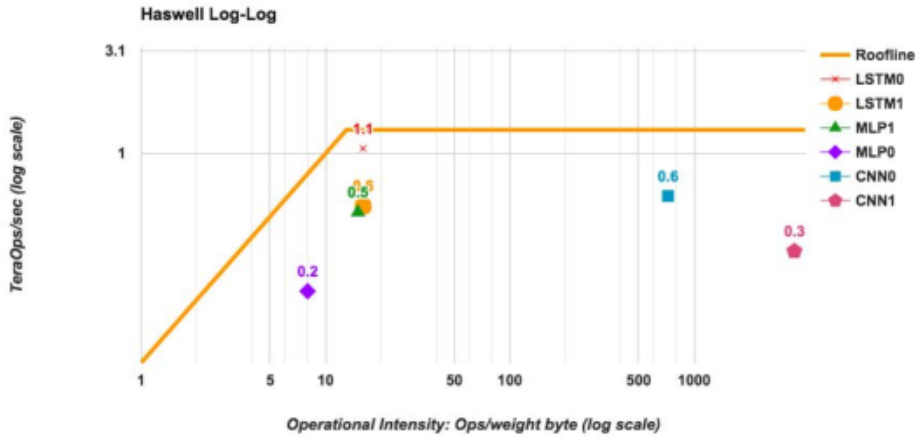
Model	*Die*										*Benchmarked Servers*				
	mm²	*nm*	*MHz*	*TDP*	*Measured*		*TOPS/s*		*GB/s*	*On-Chip Memory*	*Dies*	*DRAM Size*	*TDP*	*Measured*	
					Idle	*Busy*	8b	FP						*Idle*	*Busy*
Haswell E5-2699 v3	662	22	2300	145W	41W	145W	2.6	1.3	51	51 MiB	2	256 GiB	504W	159W	455W
NVIDIA K80 (2 dies/card)	561	28	560	150W	25W	98W	--	2.8	160	8 MiB	8	256 GiB (host) + 12 GiB x 8	1838W	357W	991W
TPU	NA*	28	700	75W	28W	40W	92	--	34	28 MiB	4	256 GiB (host) + 8 GiB x 4	861W	290W	384W

效果（与CPU、GPU性能比较）

- 与CPU、GPU性能比较







效果（与CPU、GPU性能比较）

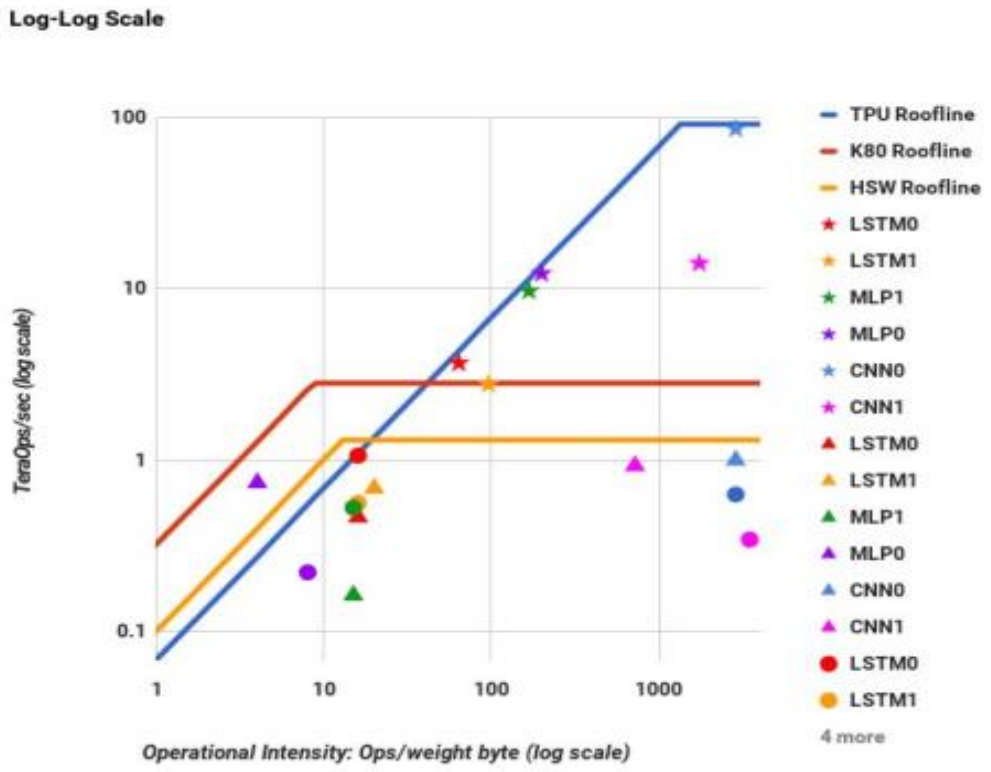
- 与CPU、GPU性能比较

Type	*Batch*	*99th% Response*	*Inf/s (IPS)*	*% Max IPS*
CPU	16	7.2 ms	5,482	42%
CPU	64	21.3 ms	13,194	100%
GPU	16	6.7 ms	13,461	37%
GPU	64	8.3 ms	36,465	100%
TPU	200	7.0 ms	225,000	80%
TPU	250	10.0 ms	280,000	100%

Type	*MLP0*	*MLP1*	*LSTM0*	*LSTM1*	*CNN0*	*CNN1*	*GM*	*WM*
GPU	2.5	0.3	0.4	1.2	1.6	2.7	1.1	1.9
TPU	41.0	18.5	3.5	1.2	40.3	71.0	14.5	29.2
Ratio	16.7	60.0	8.0	1.0	25.4	26.3	13.2	15.3

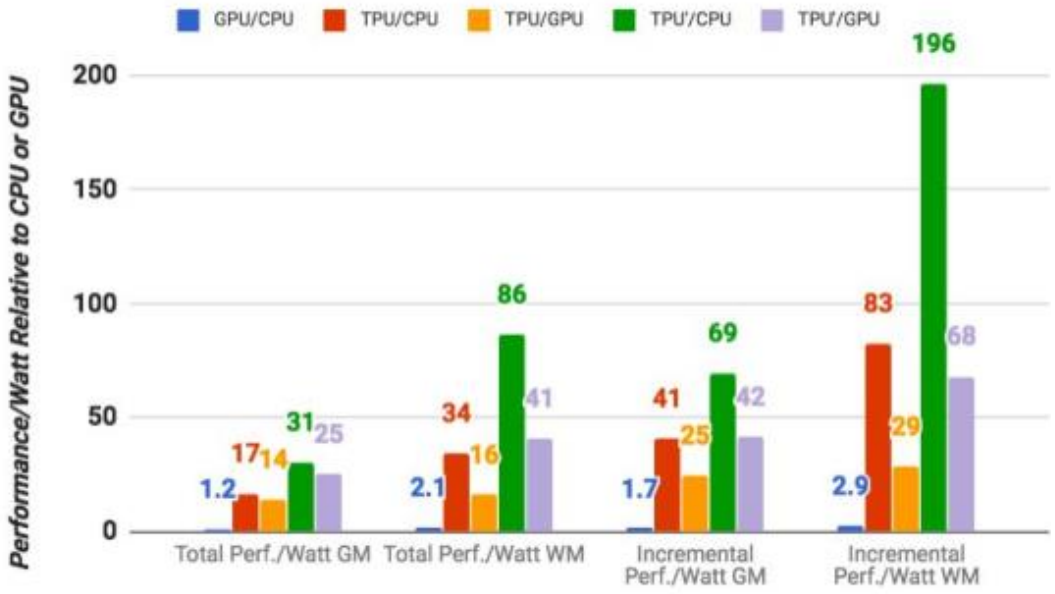
效果（与CPU、GPU性能比较）

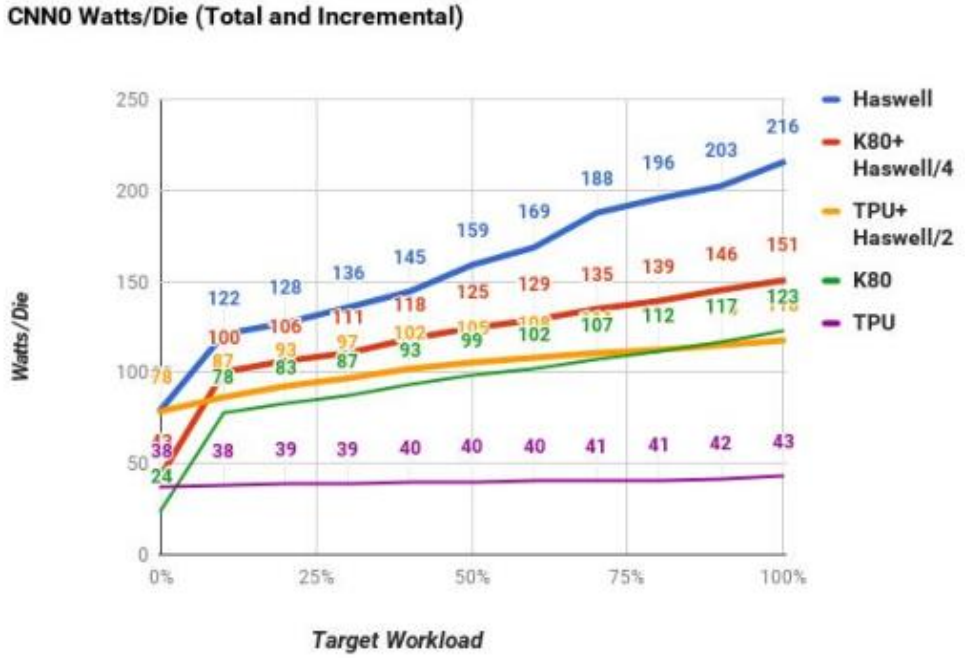
- 与CPU、GPU性能比较



效果（与CPU、GPU性能比较）

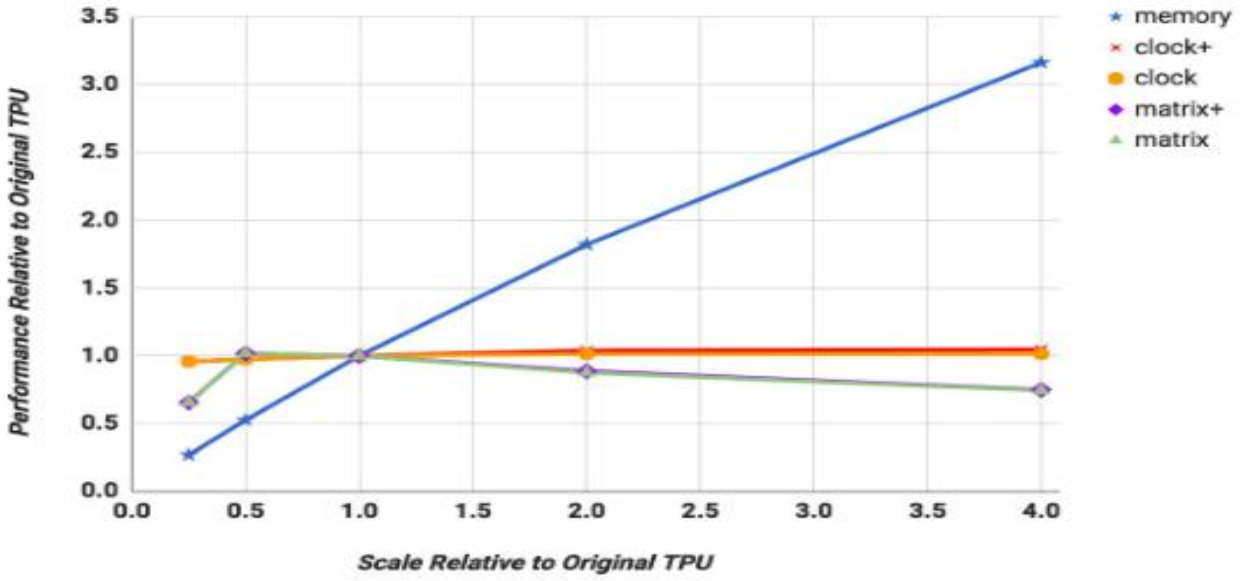
- 性价比、TCO和性能/瓦特
- 能量比例





TPU可能的设计及对比分析

- 评估可供选择的TPU设计



TPU可能的设计及对比分析

Chip	TPUv1	TPUv2	TPUv3
Announced	2016	May-17	May-18
Access	Internal-Only	Service Beta	*Undisclosed*
Introduction	2015	Feb 2018	*Undisclosed*
Process	28nm	*20nm est.*	*16/12nm est.*
Die Size	~300mm2	*Undisclosed*	*Undisclosed*
TOPS	92 / 23	45	90
Matrix Input	INT8 / INT16	bfloat16	bfloat16
Memory	8GB DDR3	16GB HBM	32GB HBM
CPU Interface	PCIe 3.0 x16	PCIe 3.0 x8	*PCIe 3.0 x8 est.*
Power Consumption	40W	*200-250W est.*	*200W est.*

结论与思考

结论

有五个架构因素可以解释TPU与CPU、GPU性能差距：

处理器：TPU只有一个处理器，而K80有13个，CPU有18个；单线程使系统更容易保持在固定的延迟限制内。

大型二维乘法单元：TPU有一个非常大的二维乘法单元，而CPU和GPU分别只有18个和13个较小的一维乘法单元；二维硬件在矩阵乘法中有很好的性能。

脉动阵列：二维组织支持脉动阵列，减少寄存器访问和能量消耗。

8位整型：TPU的应用使用 8 位整型而不是 32 位浮点运算来提高计算和内存效率。

弃掉的特征：TPU放弃了 CPU 和 GPU 需要但是 DNN 用不到的功能，这使得 TPU 更便宜，同时可以节约资源，并允许晶体管被重新用于特定领域的板载内存。

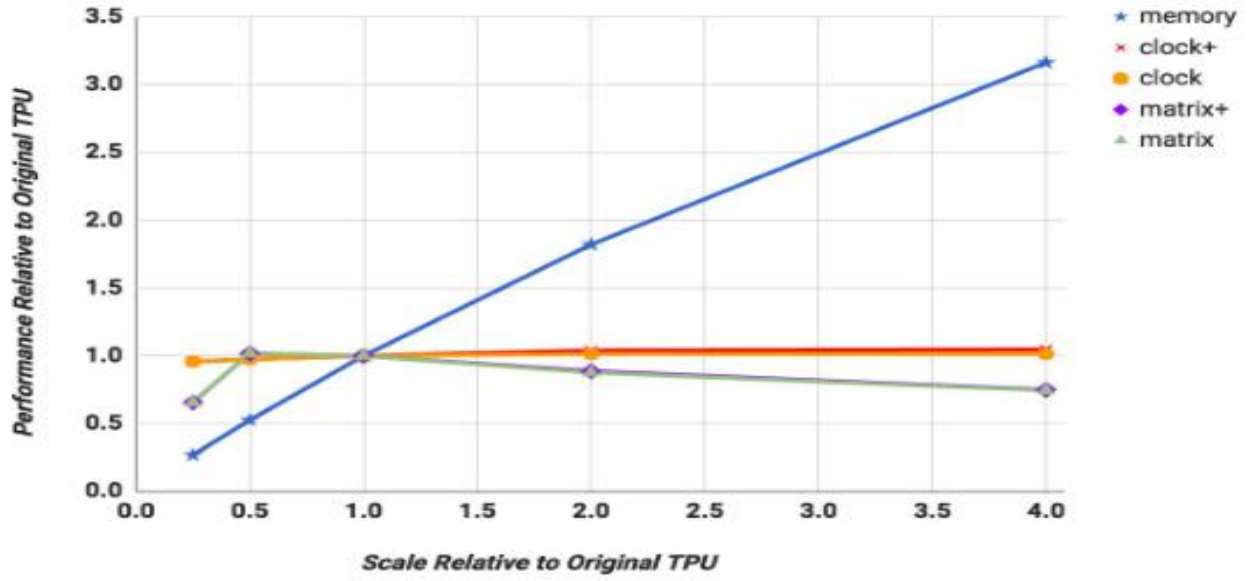
结论与思考

思考

关于TPU的发展以及性能提升等方面的一些思考

问题

TPU可选设计中，matrix增大为何性能反而略有下降？



谢 谢