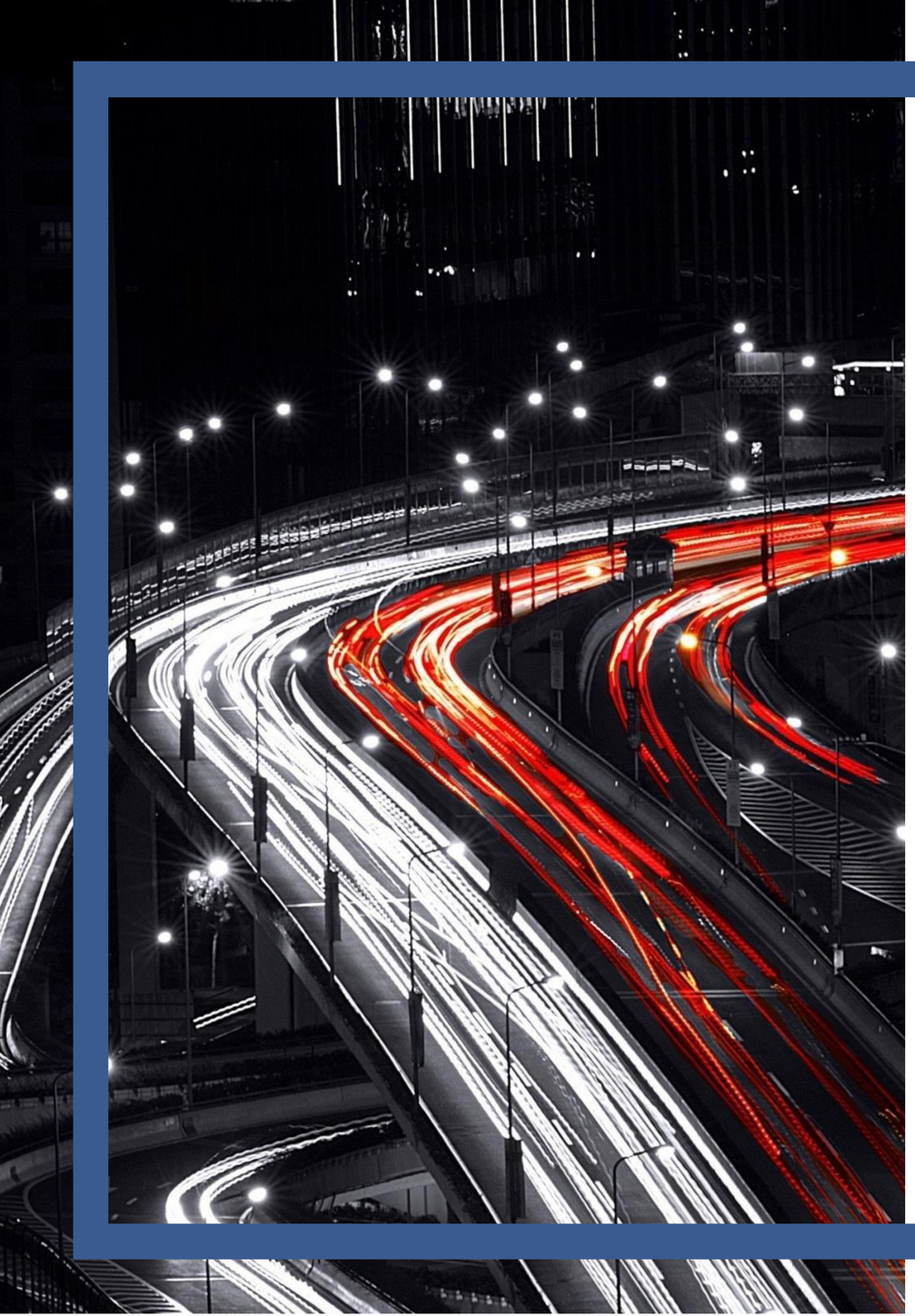
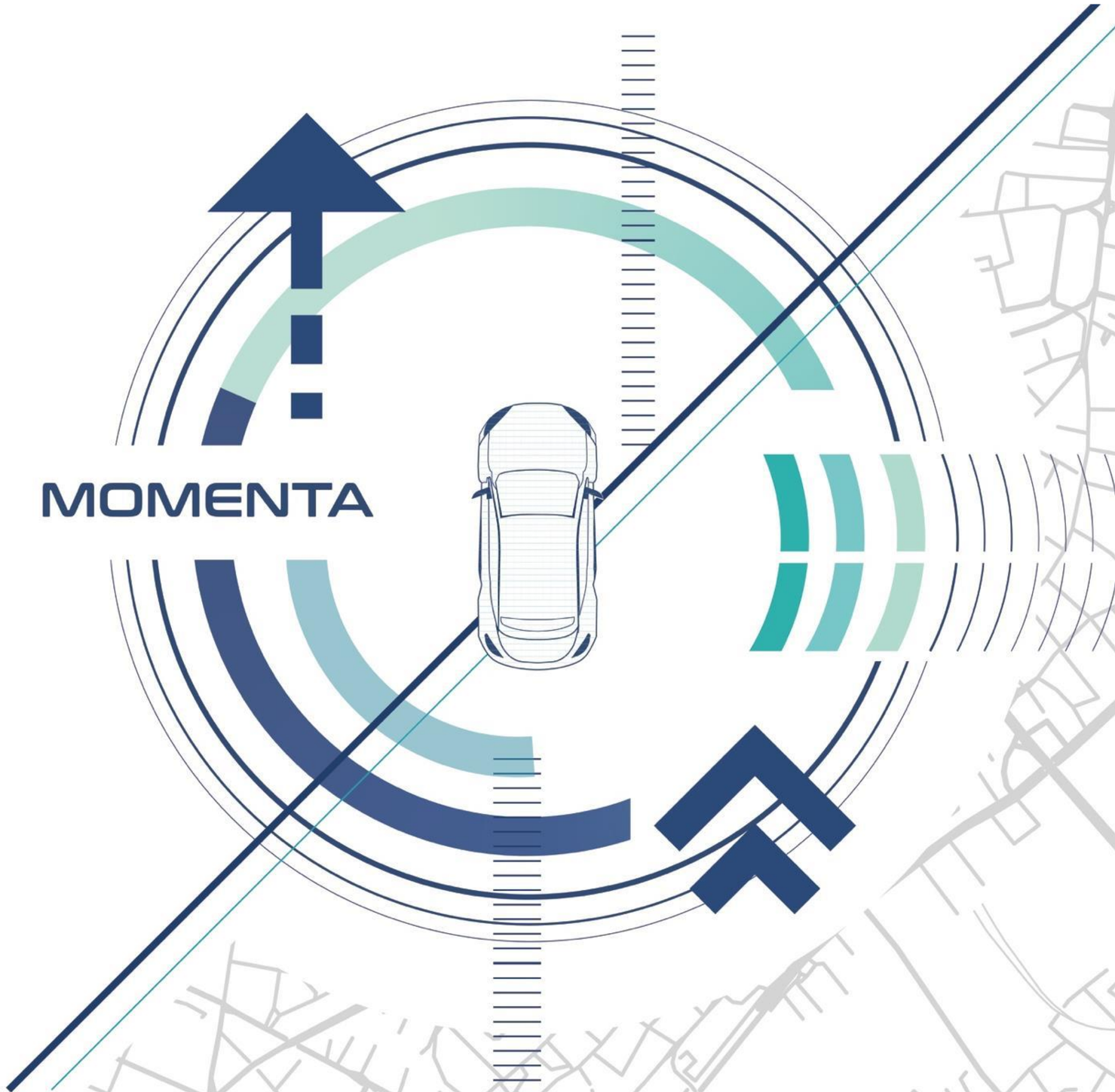


让深度学习更高效运行的两个视角

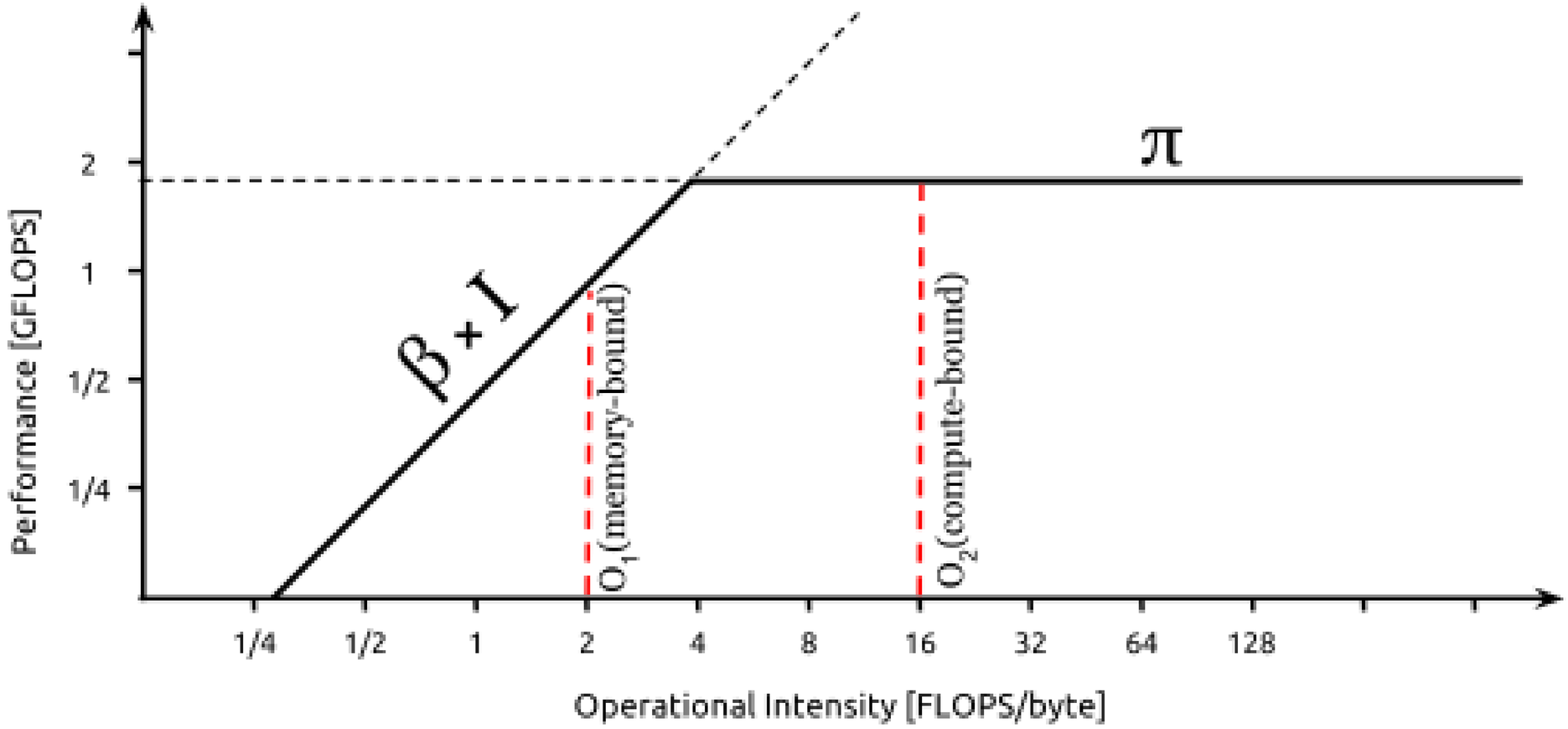
Momenta

关于
Momenta

- 打造自动驾驶大脑。
- 核心技术：基于深度学习的环境感知、高精度地图、驾驶决策算法



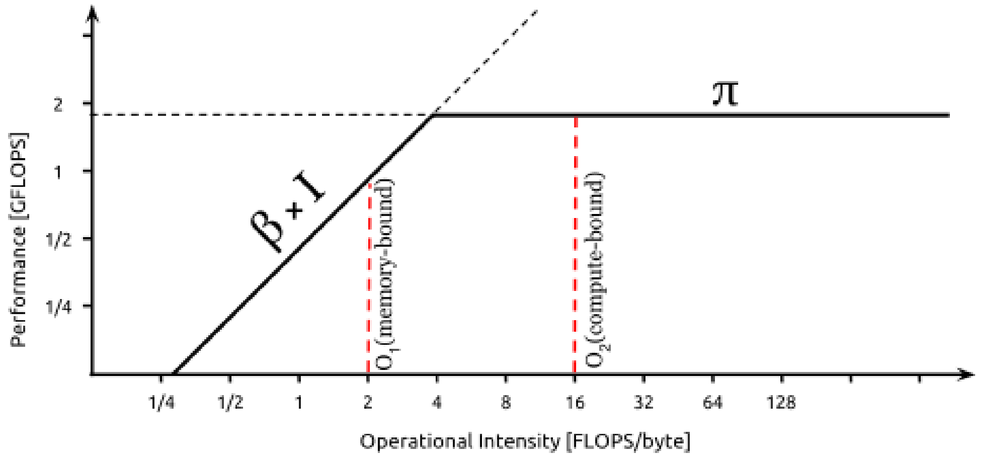
背景

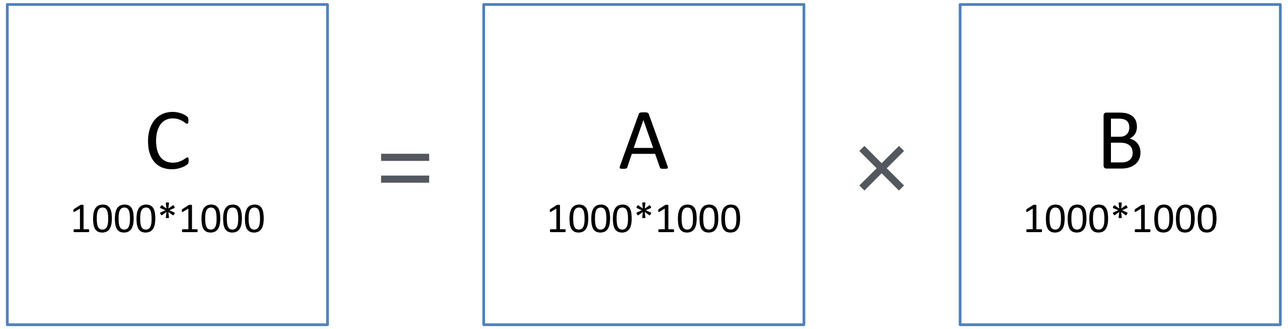


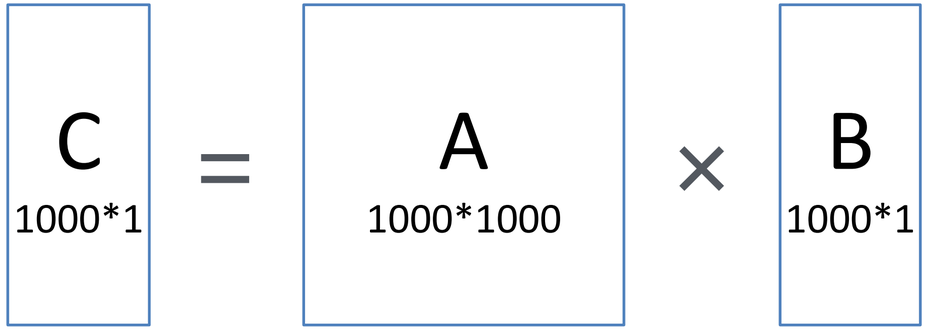
* https://en.wikipedia.org/wiki/Roofline_model

MOMENTA

背景



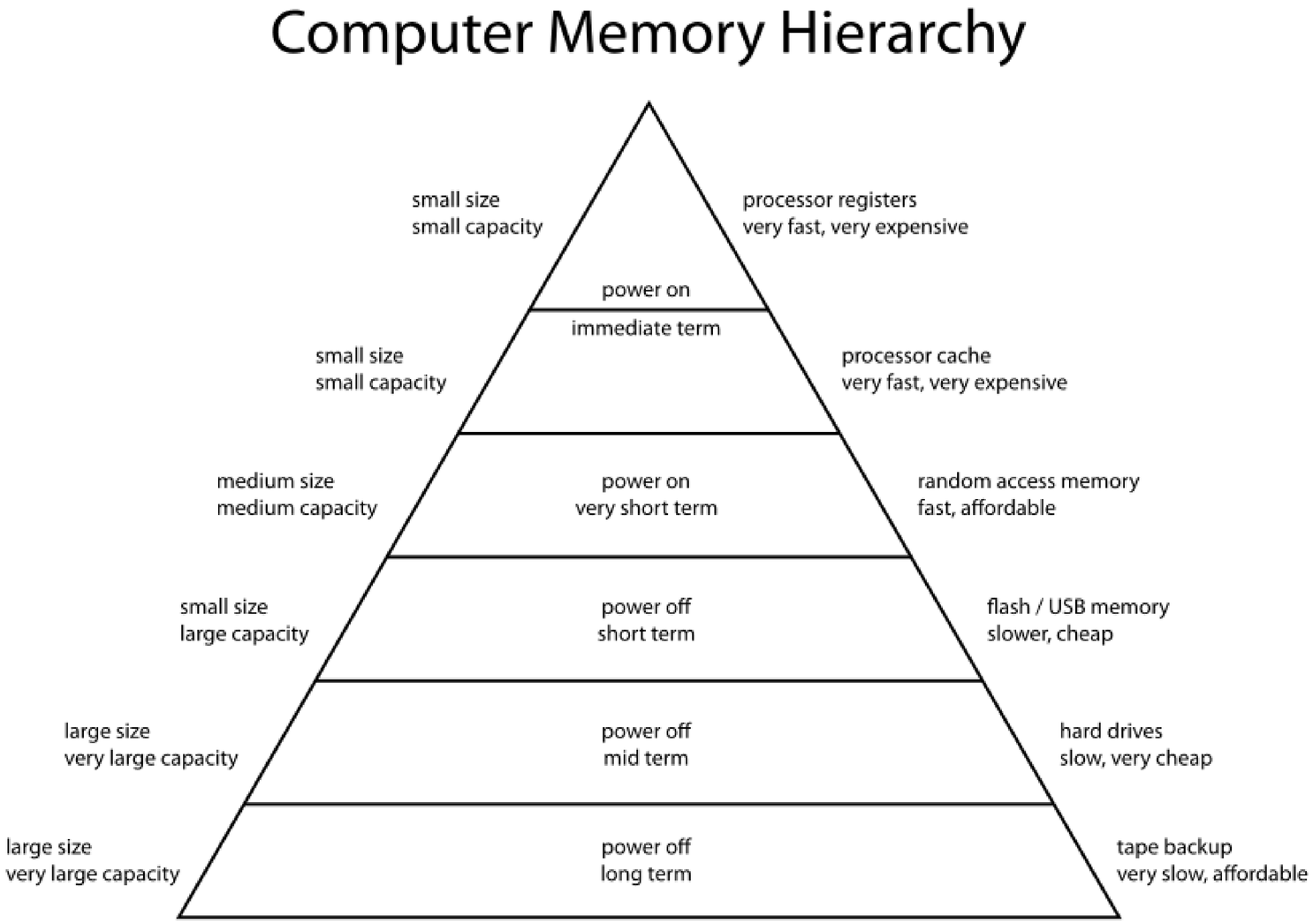




* https://en.wikipedia.org/wiki/Roofline_model

MOMENTA

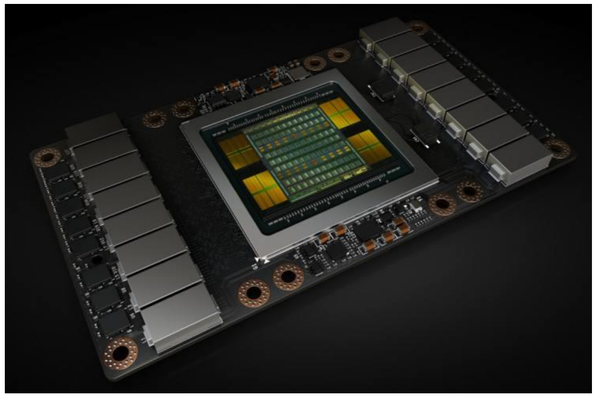
背景



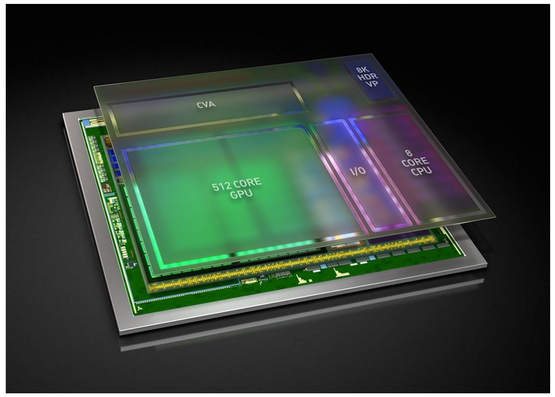
* https://en.wikipedia.org/wiki/Memory_hierarchy

MOMENTA

背景



NVIDIA Tesla V100
120T TensorCore FLOPS
HBM2 900 GB/s
20MB SM+16MB 缓存



NVIDIA Xavier
20T TensorCore FLOPS
10T DLA OPS
LPDDR4 137 GB/s
片上存储未知

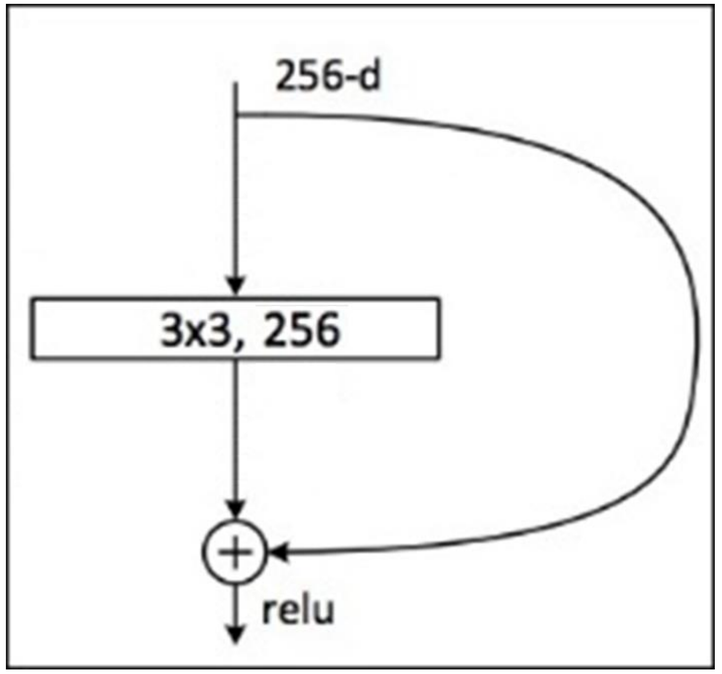
Raspberry Pi 3
38.4 GFLOPS (4x A53 1.2GHz)
LPDDR2 约 3.6 GB/s
512KB L2缓存

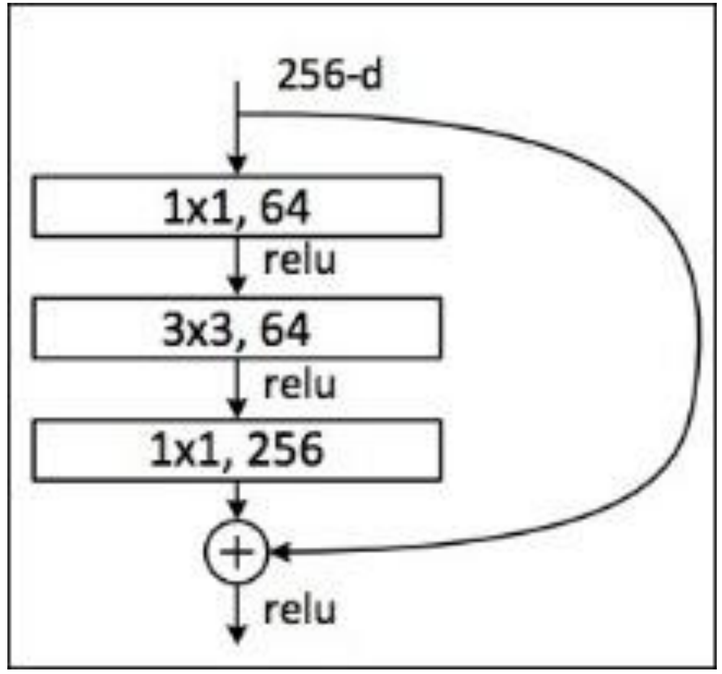
MOMENTA

两个视角：计算量和访存量

模型	计算量 / FLOPS	访存量 / byte	计算密度 / FLOPS/byte
VGG 16	31.0 G	675 M	45.9
ResNet 152	22.6 G	472 M	47.9
ResNet 50	7.72 G	211 M	36.6
ResNet 18	3.63 G	72.5 M	50.1
Inception V2	4.07 G	100 M	40.7
MobileNet	1.15 G	57.8 M	19.9
ShuffleNet-0.5x-g3	68.5 M	19.0 M	3.61
ShuffleNet-0.5x-g8	76.9 M	29.4 M	2.91

优化1：bottleneck

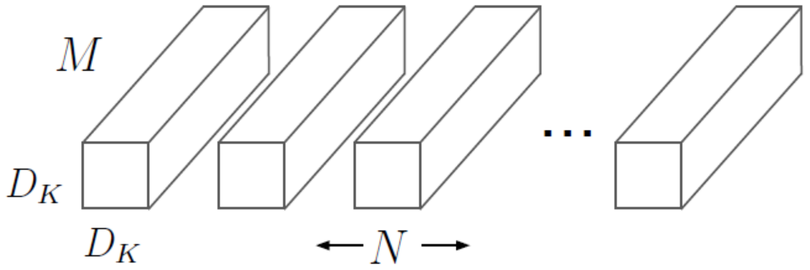




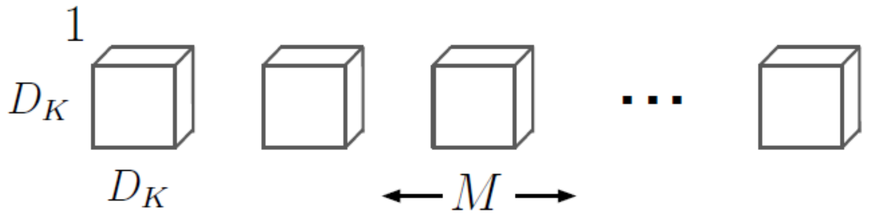
* https://arxiv.org/abs/1512.03385

MOMENTA

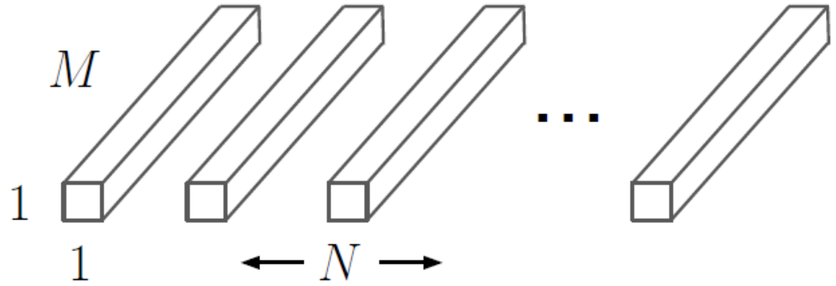
优化2：depthwise卷积



(a) Standard Convolution Filters



(b) Depthwise Convolutional Filters



(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

* https://arxiv.org/abs/1704.04861

MOMENTA

两个视角：计算量和访存量

模型	计算量 / FLOPS	访存量 / byte	计算密度 / FLOPS/byte
VGG 16	31.0 G	675 M	45.9
ResNet 152	22.6 G	472 M	47.9
ResNet 50	7.72 G	211 M	36.6
ResNet 18	3.63 G	72.5 M	50.1
Inception V2	4.07 G	100 M	40.7
MobileNet	1.15 G	57.8 M	19.9
ShuffleNet-0.5x-g3	68.5 M	19.0 M	3.61
ShuffleNet-0.5x-g8	76.9 M	29.4 M	2.91

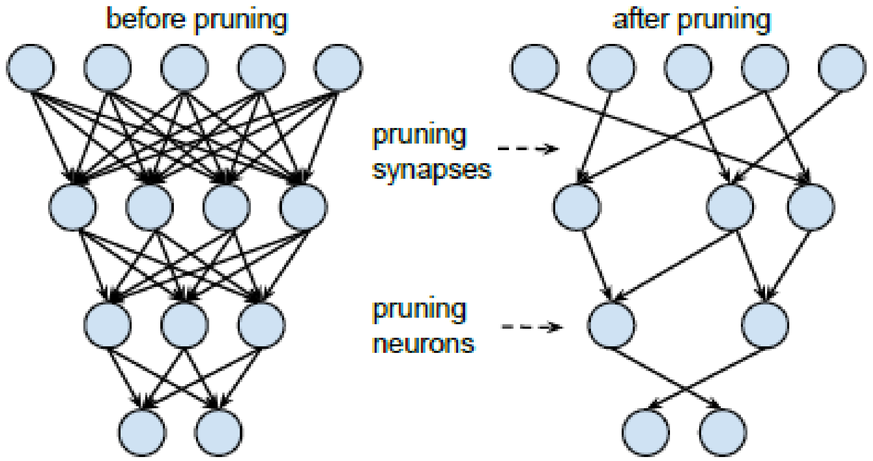
优化3：FFT / Winograd卷积算法

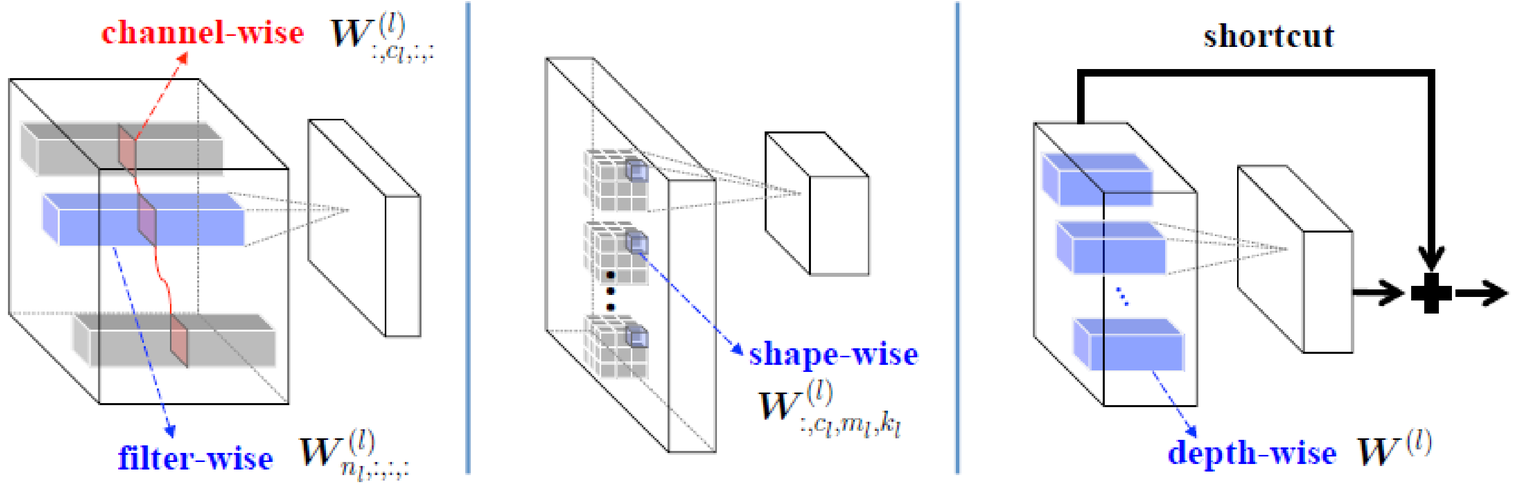
$$Y = A^T \left[[GgG^T] \odot [B^T dB] \right] A$$

* https://arxiv.org/abs/1509.09308

MOMENTA

优化4：稀疏化

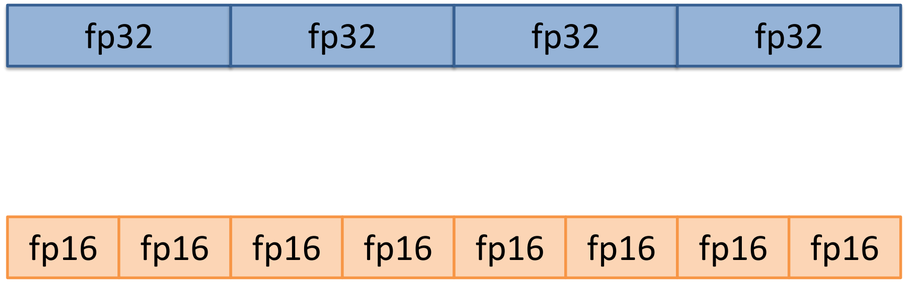




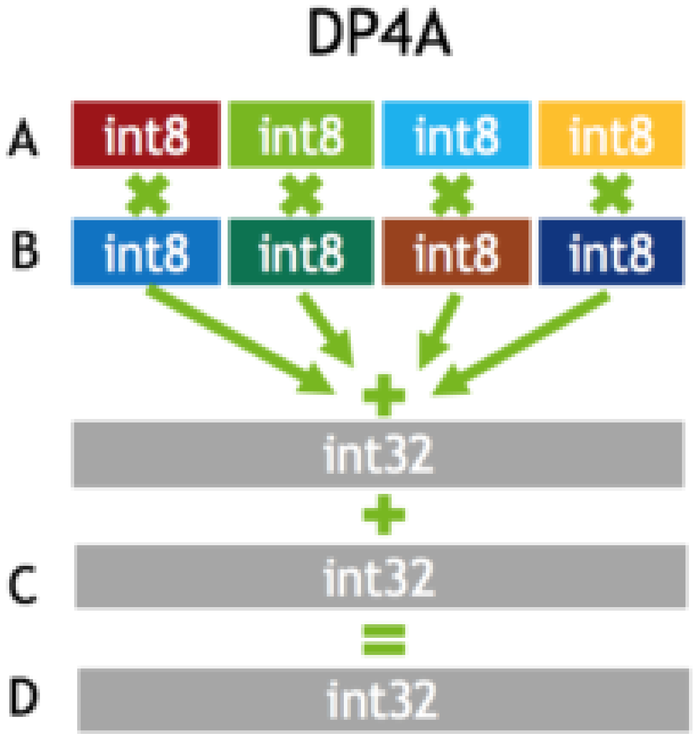
* https://arxiv.org/abs/1506.02626
* https://arxiv.org/abs/1608.03665

MOMENTA

优化5：低精度运算



NVIDIA 部分GPU
大多数移动GPU
新一代ARM



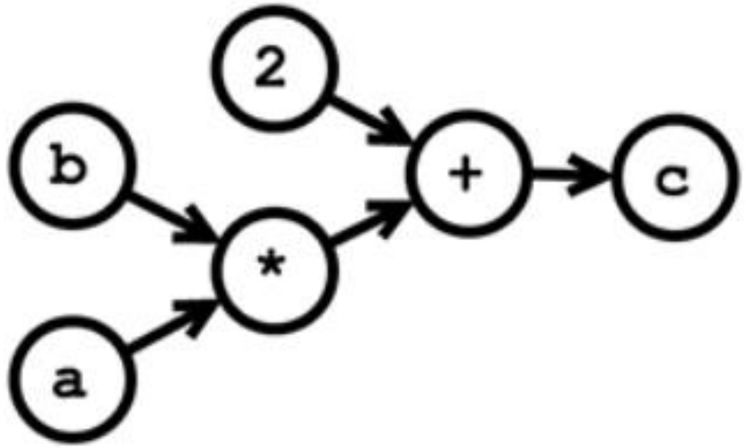
NVIDIA 部分GPU
新一代ARM

MOMENTA

从operator层面向上

Frontend `c = a * b + 2`

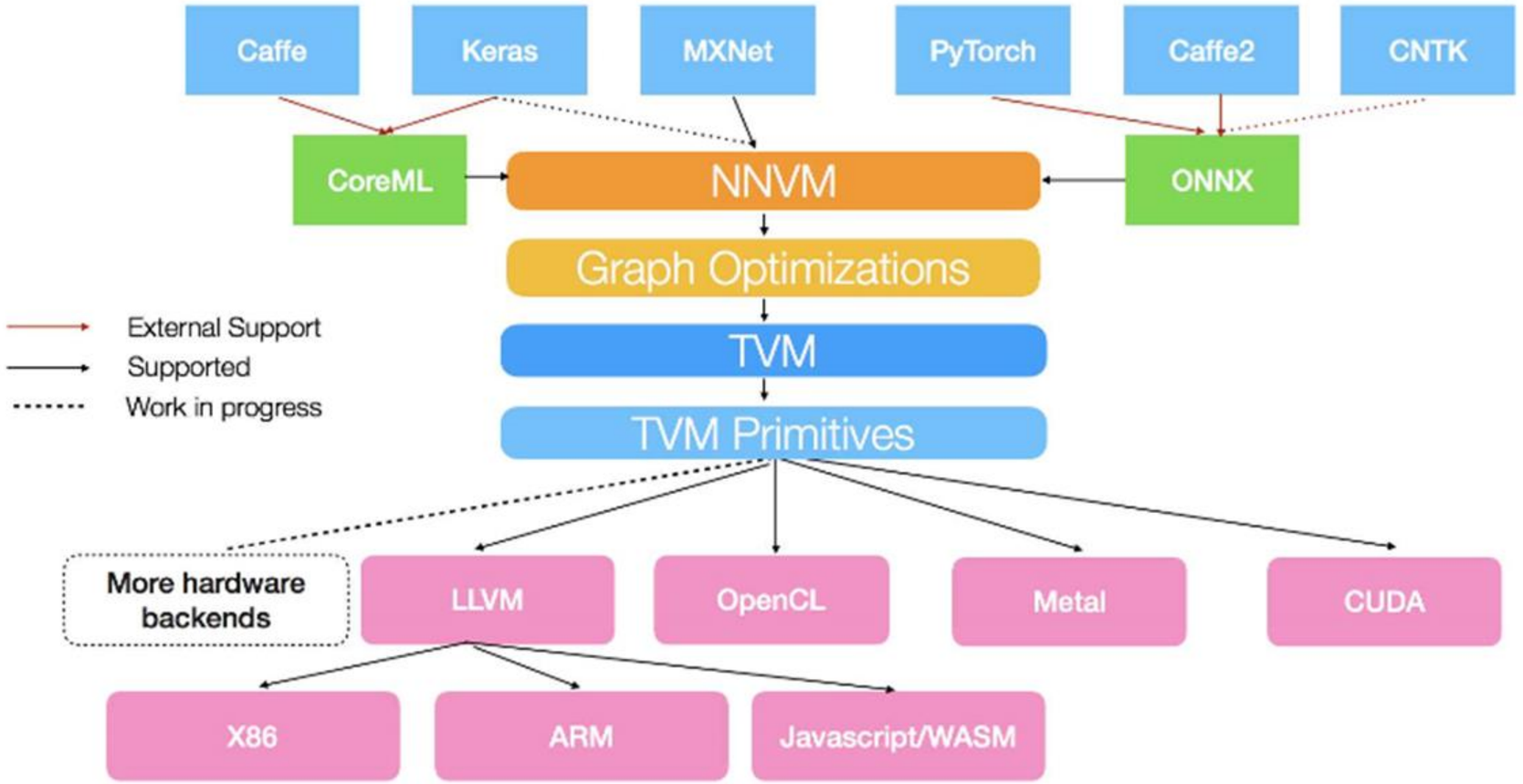
Computation Graph



* https://aws.amazon.com/cn/blogs/machine-learning/introducing-nnvm-compiler-a-new-open-end-to-end-compiler-for-ai-frameworks/

MOMENTA

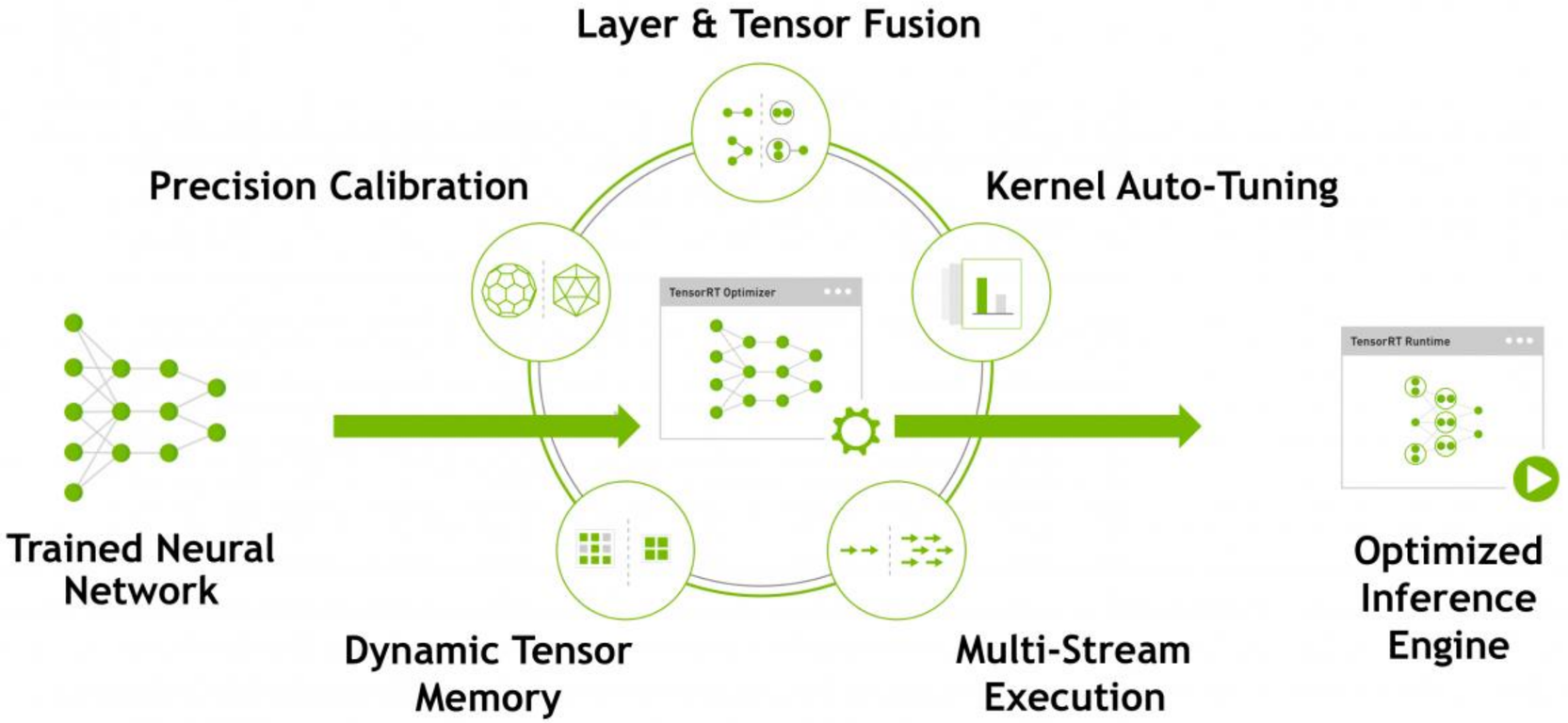
从operator层面向上



* http://www.tvmlang.org/2017/10/06/nnvm-compiler-announcement.html

MOMENTA

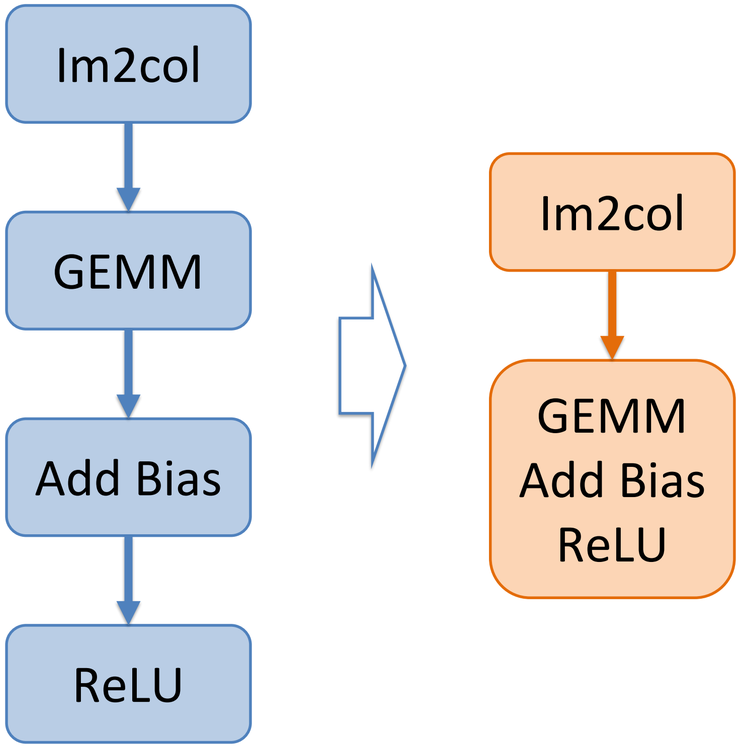
从operator层面向上

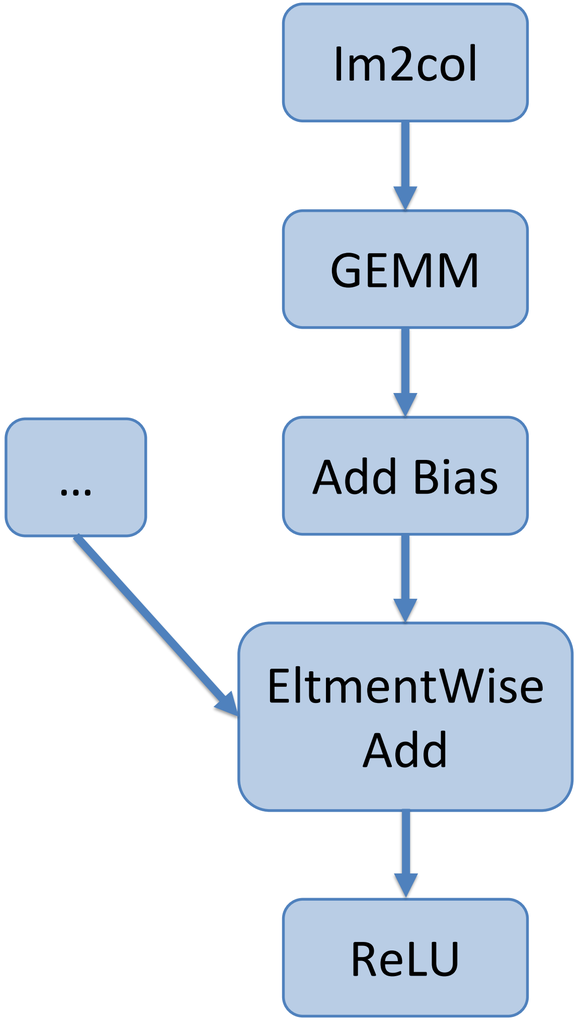


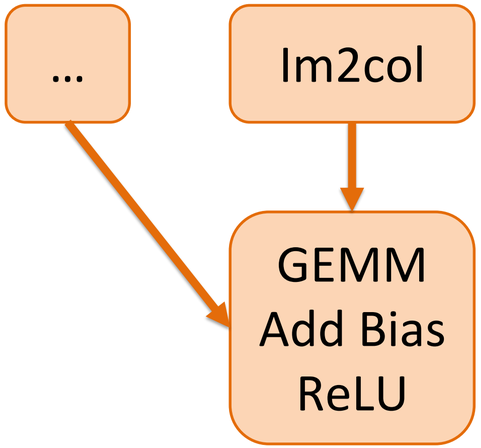
* https://developer.nvidia.com/tensorrt

MOMENTA

Operator的融合

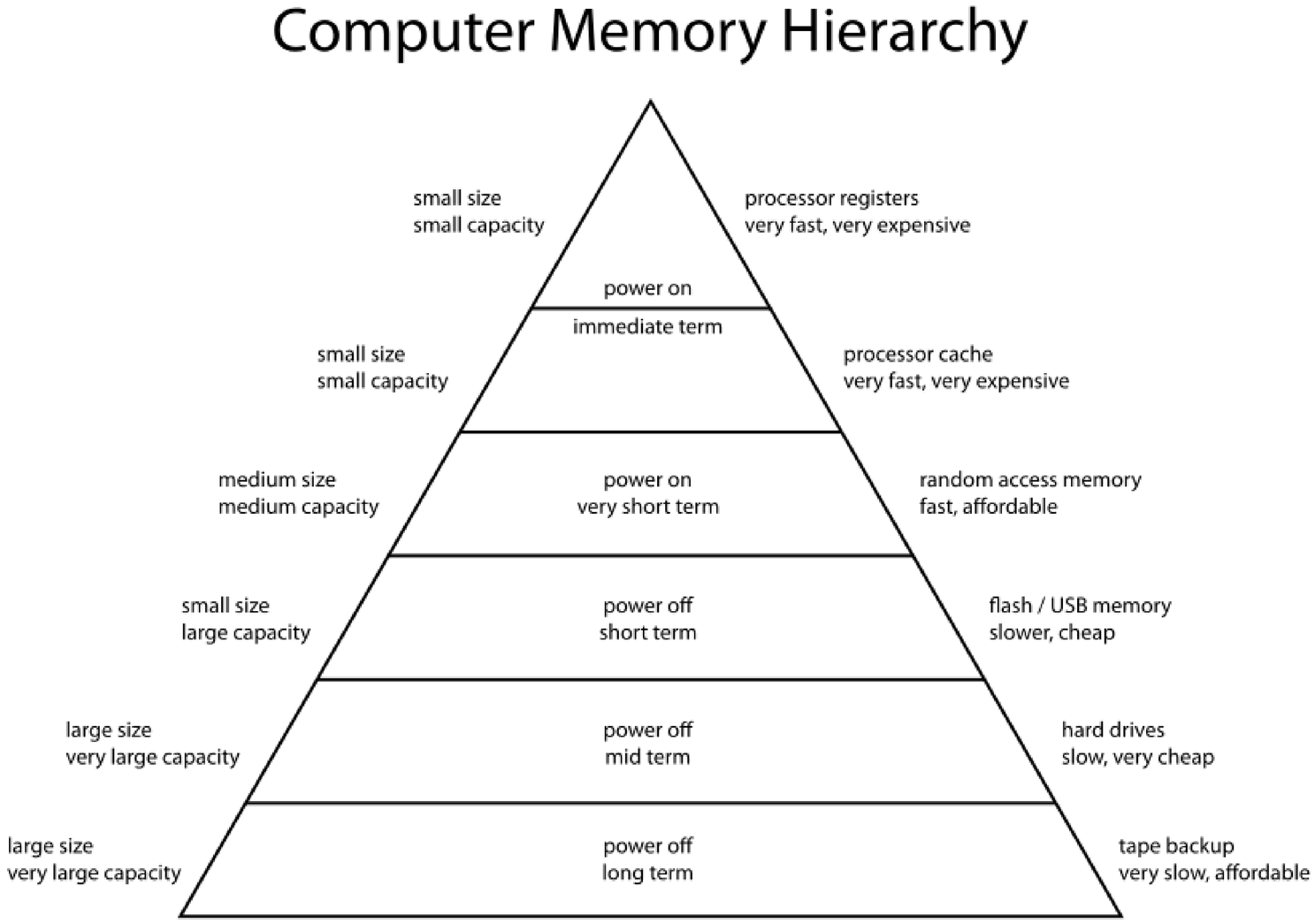






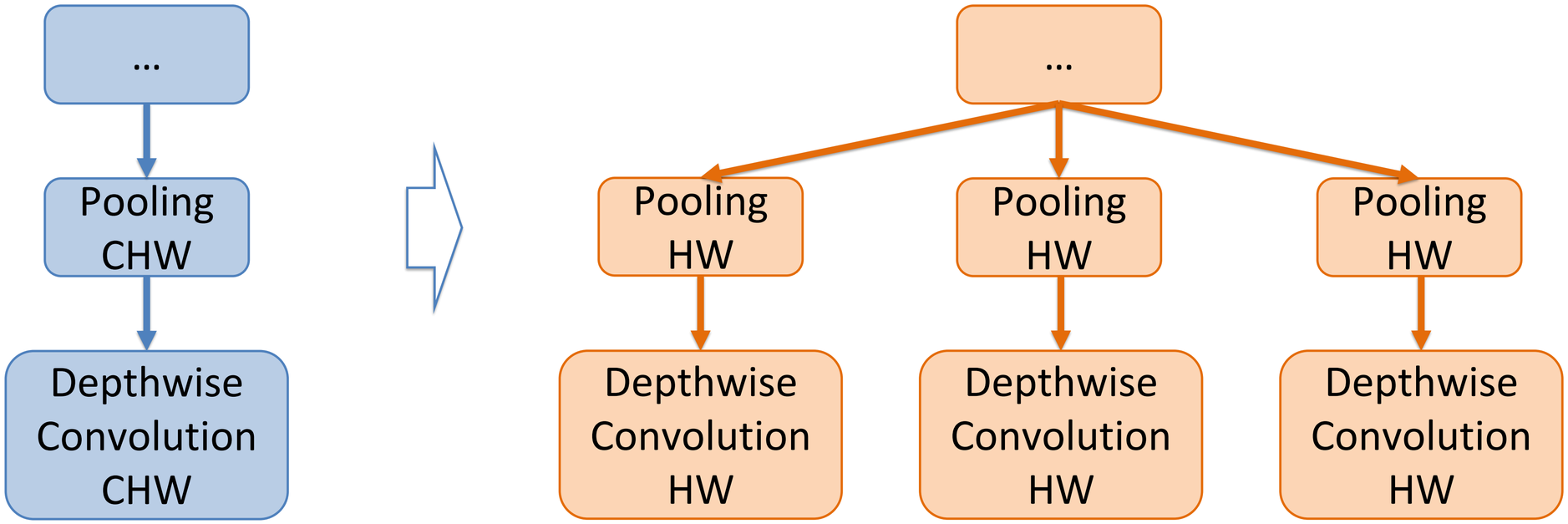
MOMENTA

Operator的融合

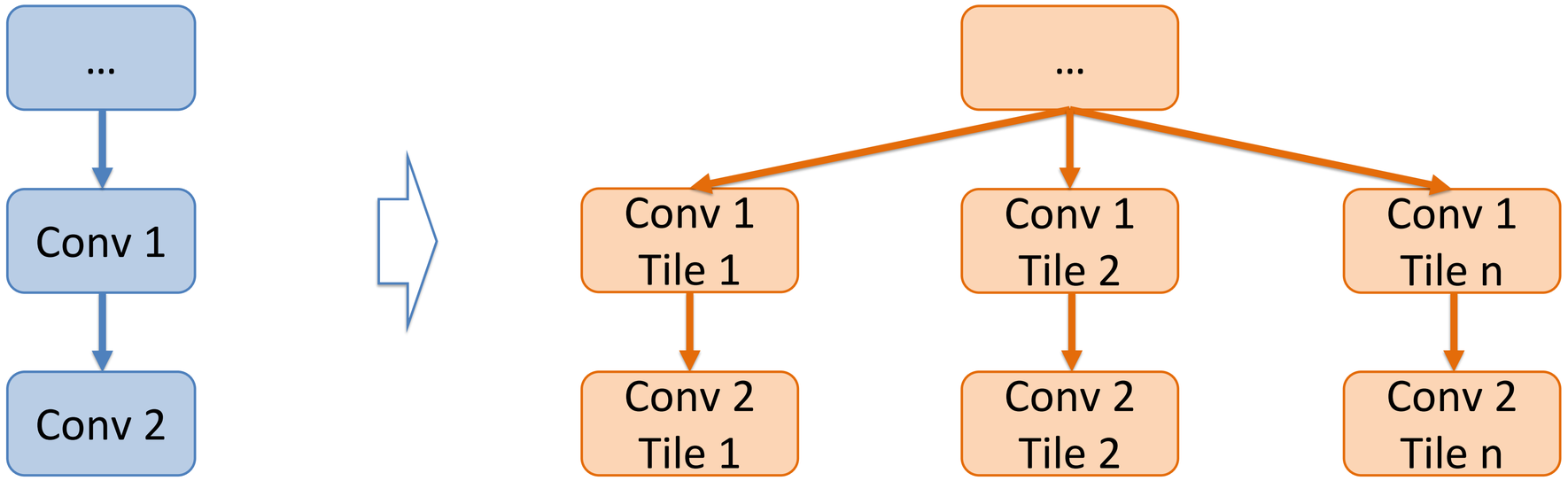


MOMENTA

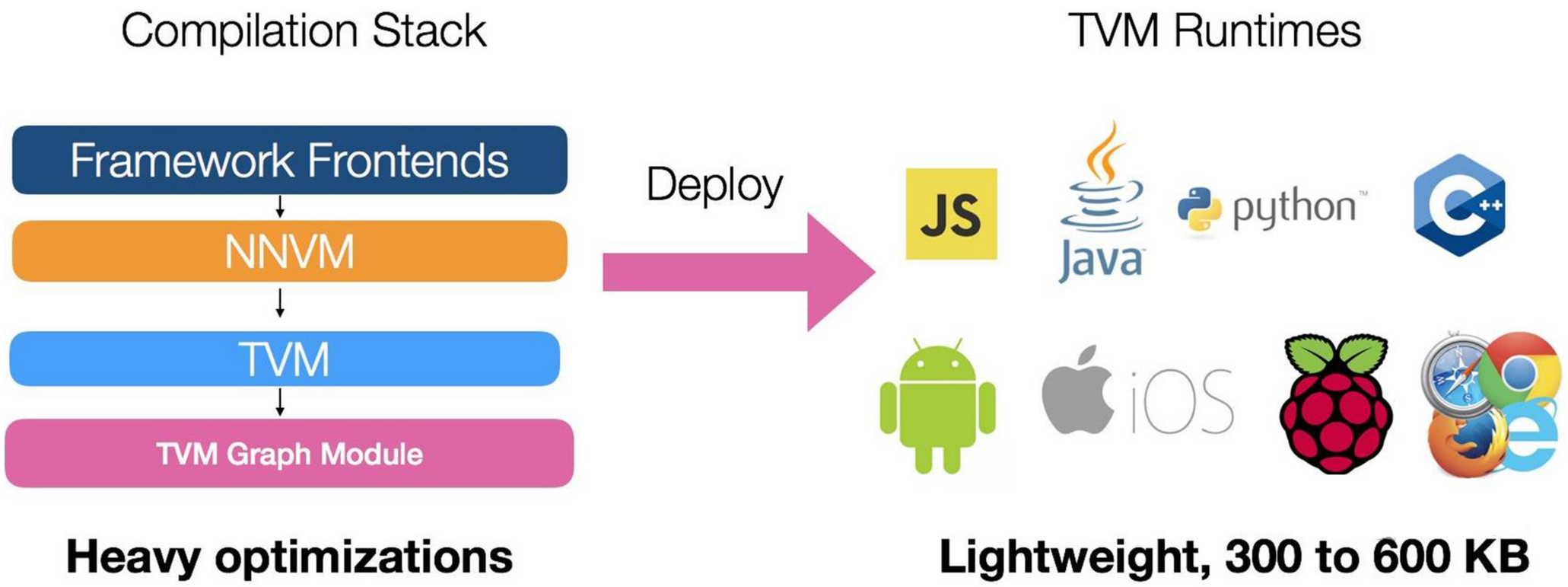
Operator的粒度



Operator的粒度



轻量化的运行时



* http://www.tvmlang.org/2017/10/06/nnvm-compiler-announcement.html

加速效果

	Caffe with Eigen	Momenta	加速比
VGG 16	11.6 s	2.75 s	4.22
ResNet 50	3.25 s	1.58 s	2.06
ResNet 18	1.54 s	594 ms	2.59
MobileNet	1.87 s	670 ms	2.79
ShuffleNet-0.5x-g3	115 ms	38.8 ms	2.96

* 使用Raspberry Pi 3，Cortex A53 1.2GHz，单核测试

MOMENTA

总结

两个视角：计算量和访存量

通过网络结构设计优化计算量

通过Winograd、稀疏化、低精度运算等优化计算量

通过低精度运算、计算图优化等优化访存量

通过轻量化运行时避免额外开销

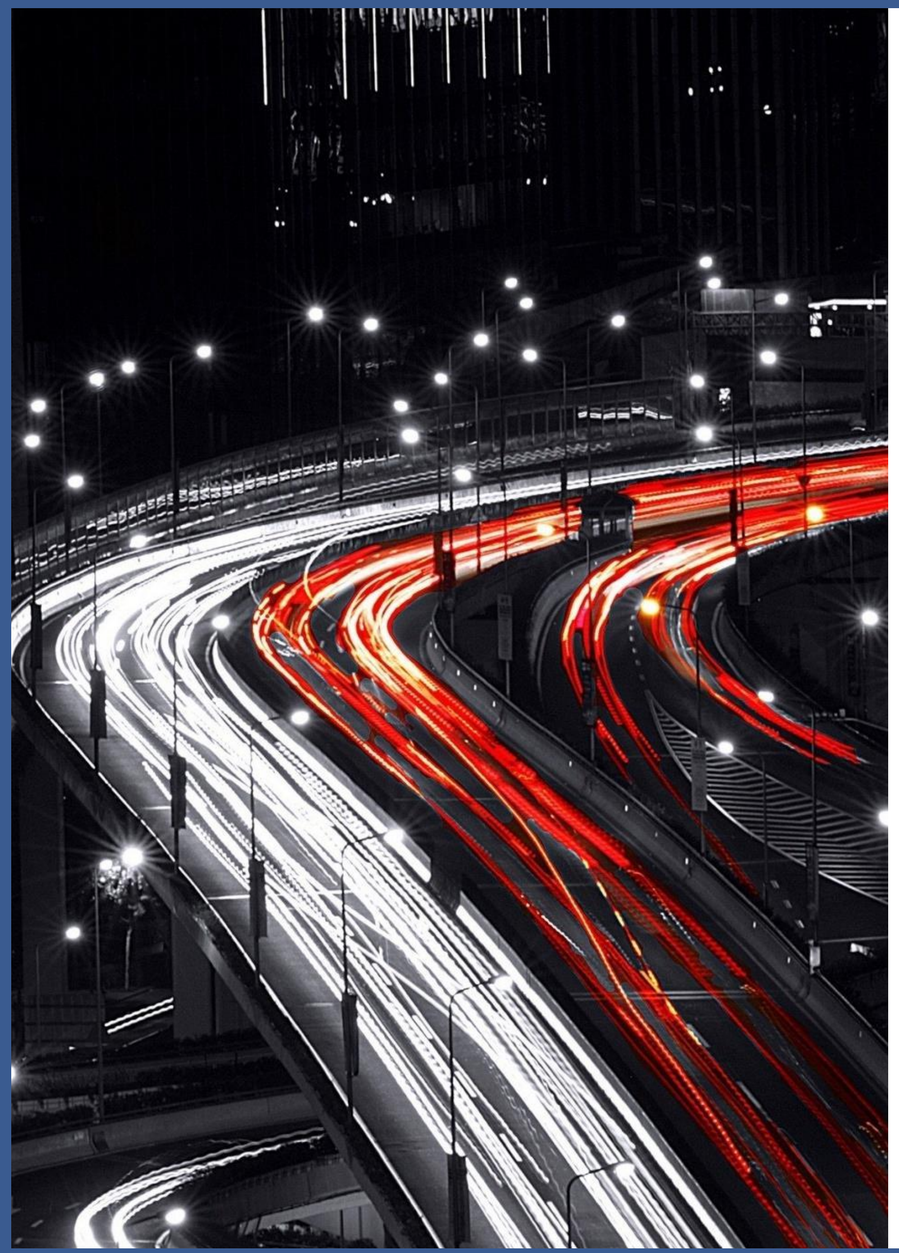
MOMENTA

THANK YOU

简历投递 jobs@momenta.ai

201802

MOMENTA



让深度学习更高效运行的两个视角

简历投递 jobs@momenta.ai

Momenta