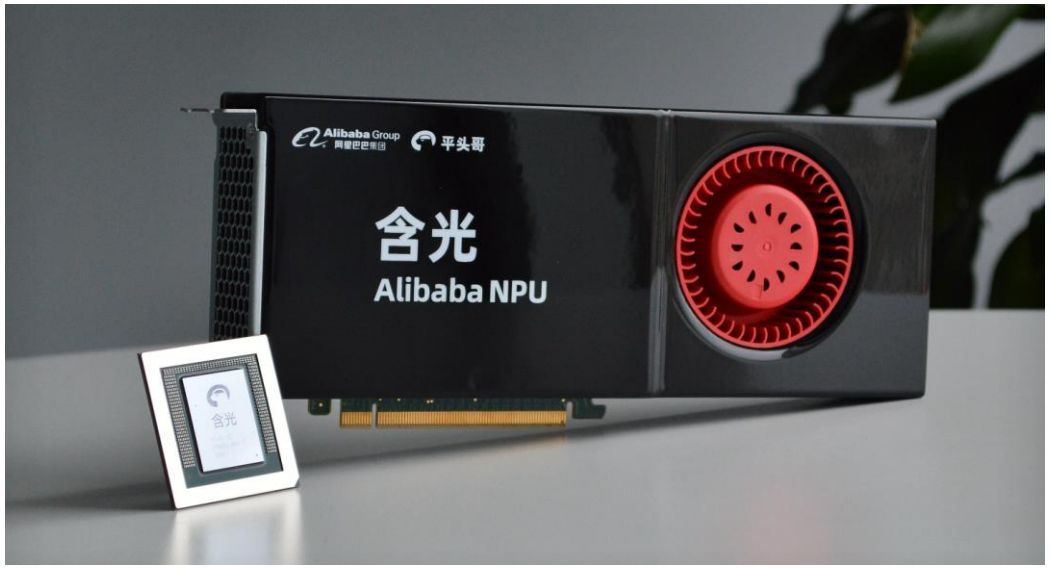
Hanguang 800 NPU
– The Ultimate AI Inference Solution for Data Centers



Yang Jiao, Liang Han, Xin Long, & Team

Alibaba Group

A Business-Driven Design

TCO Efficient

- For data center inference acceleration
- High throughput, low latency, power efficient
- High effective TOPs

CNN Optimized

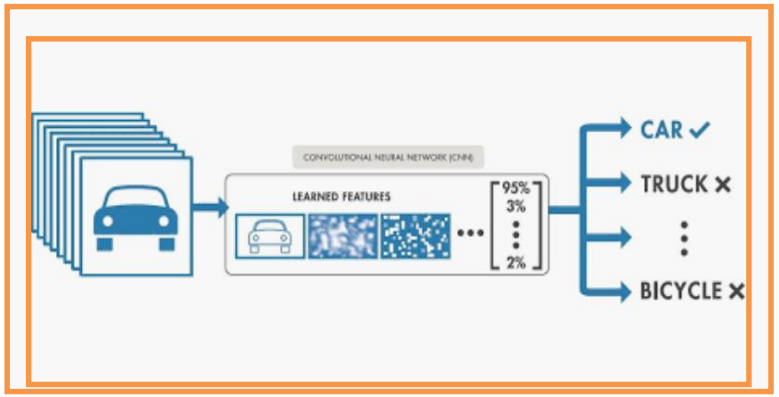
- Convolution-efficient architecture
- As well as GEMM acceleration

Domain Programmable

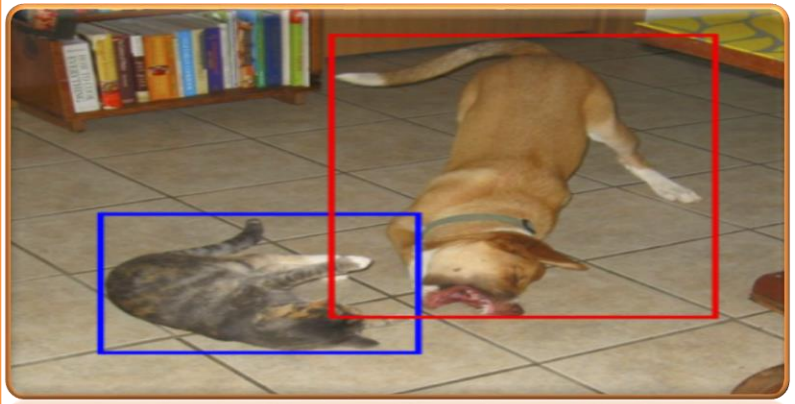
- Native OPs for computer vision deep learning tasks
- Expand to support future activation functions

Optimized for CV Tasks

Classification

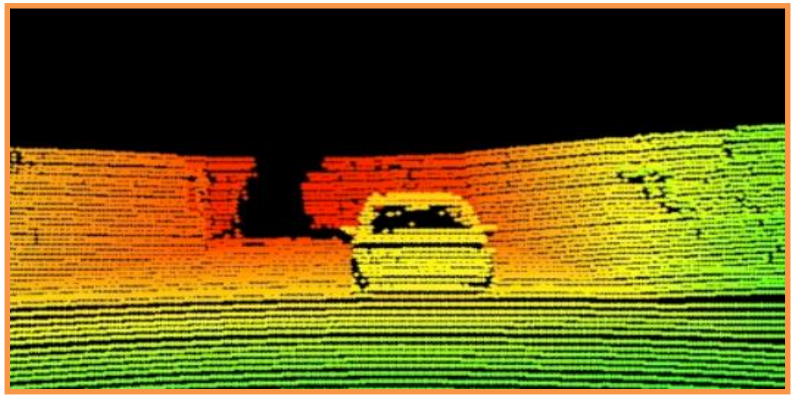


Object Detection



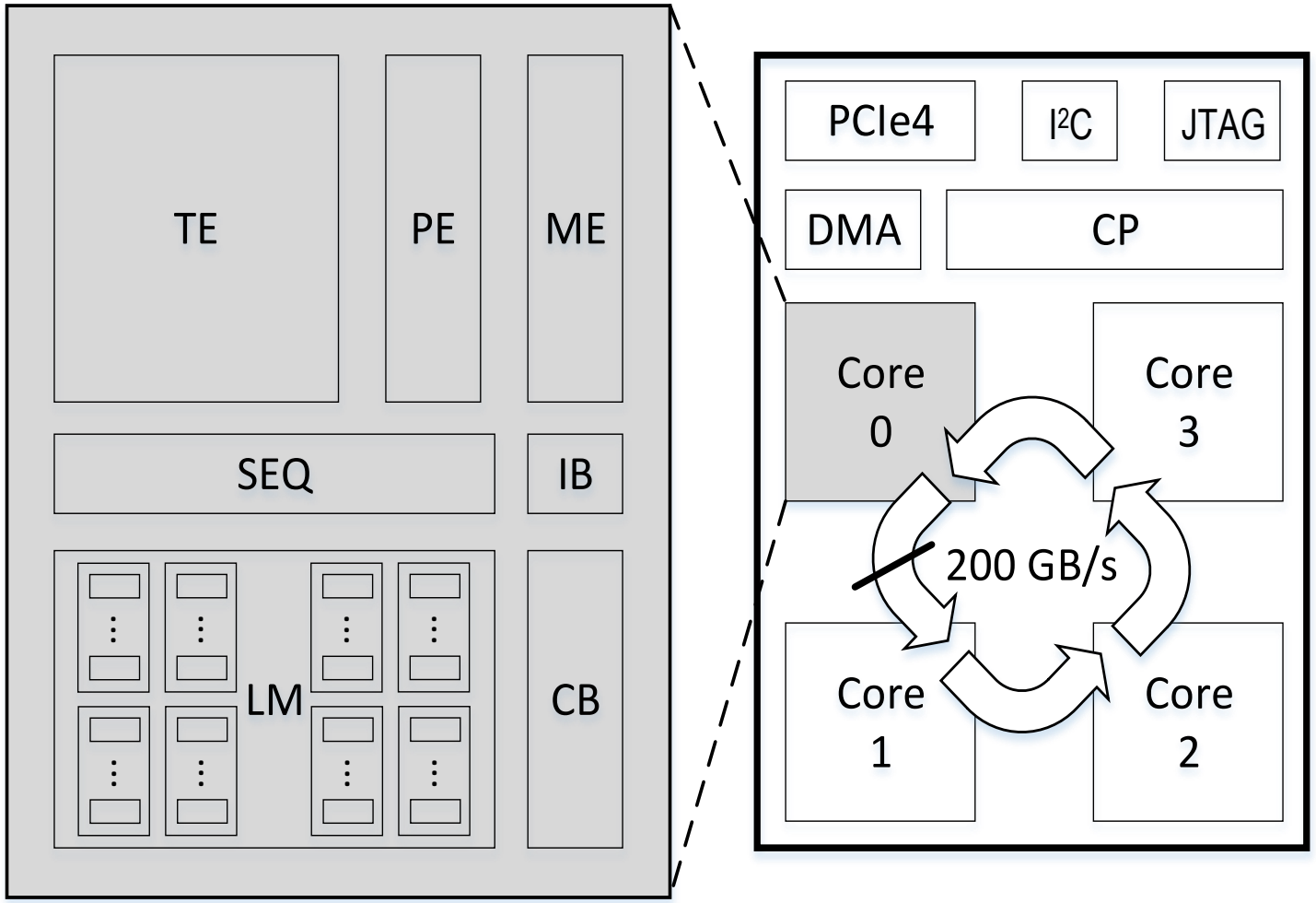
Segmentation

Point Cloud



Top Level Diagram

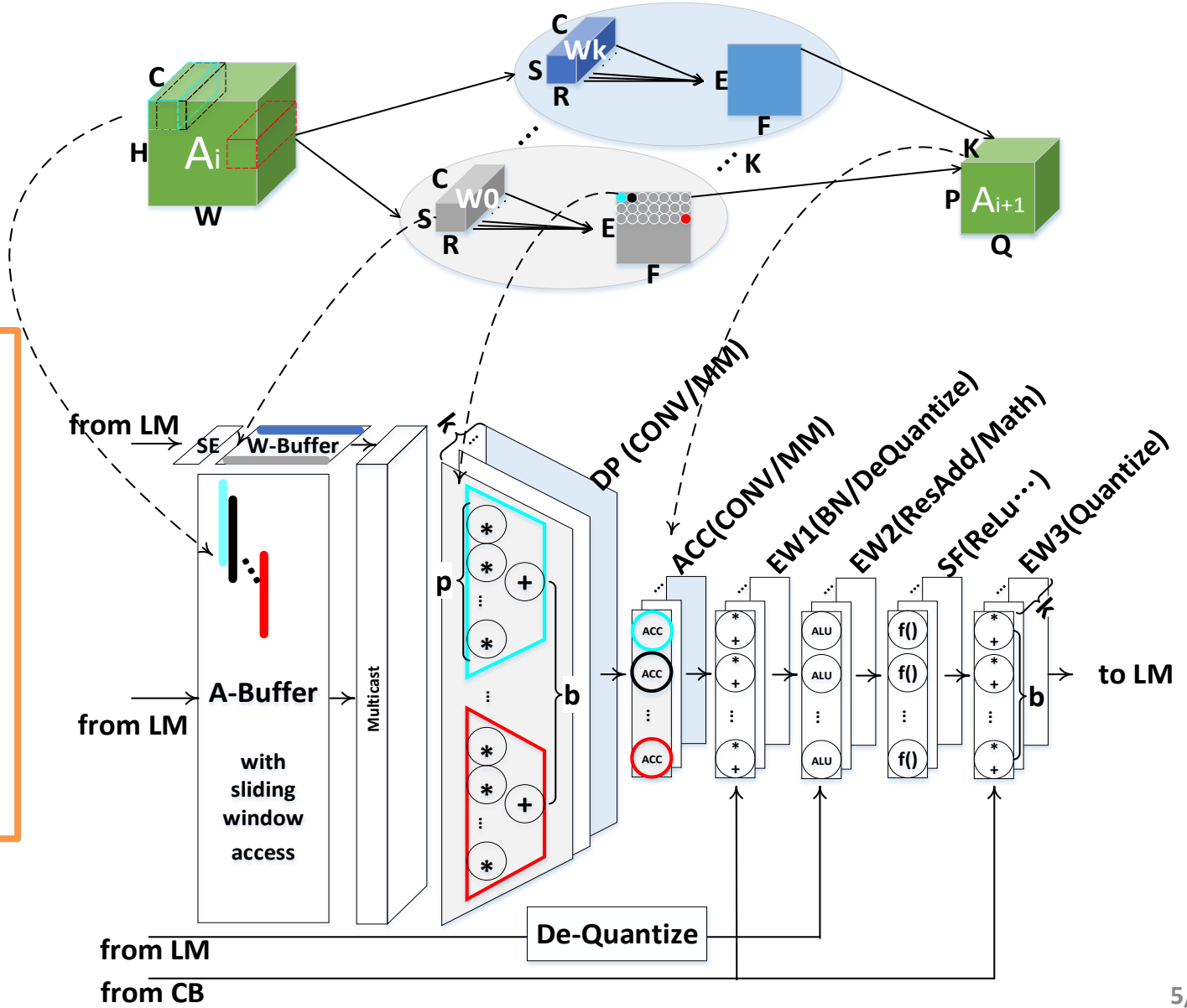
- Top Level:
  - 4 Cores w/ a ring bus
  - Command Processor (CP)
  - PCIE gen4 X16
- Each Core:
  - Tensor Engine (TE)
  - Pooling Engine (PE)
  - Memory Engine (ME)
- SRAM-Only:
  - 192MB Local Memory (LM)
  - Distributed shared
  - No DDR



Tensor Engine

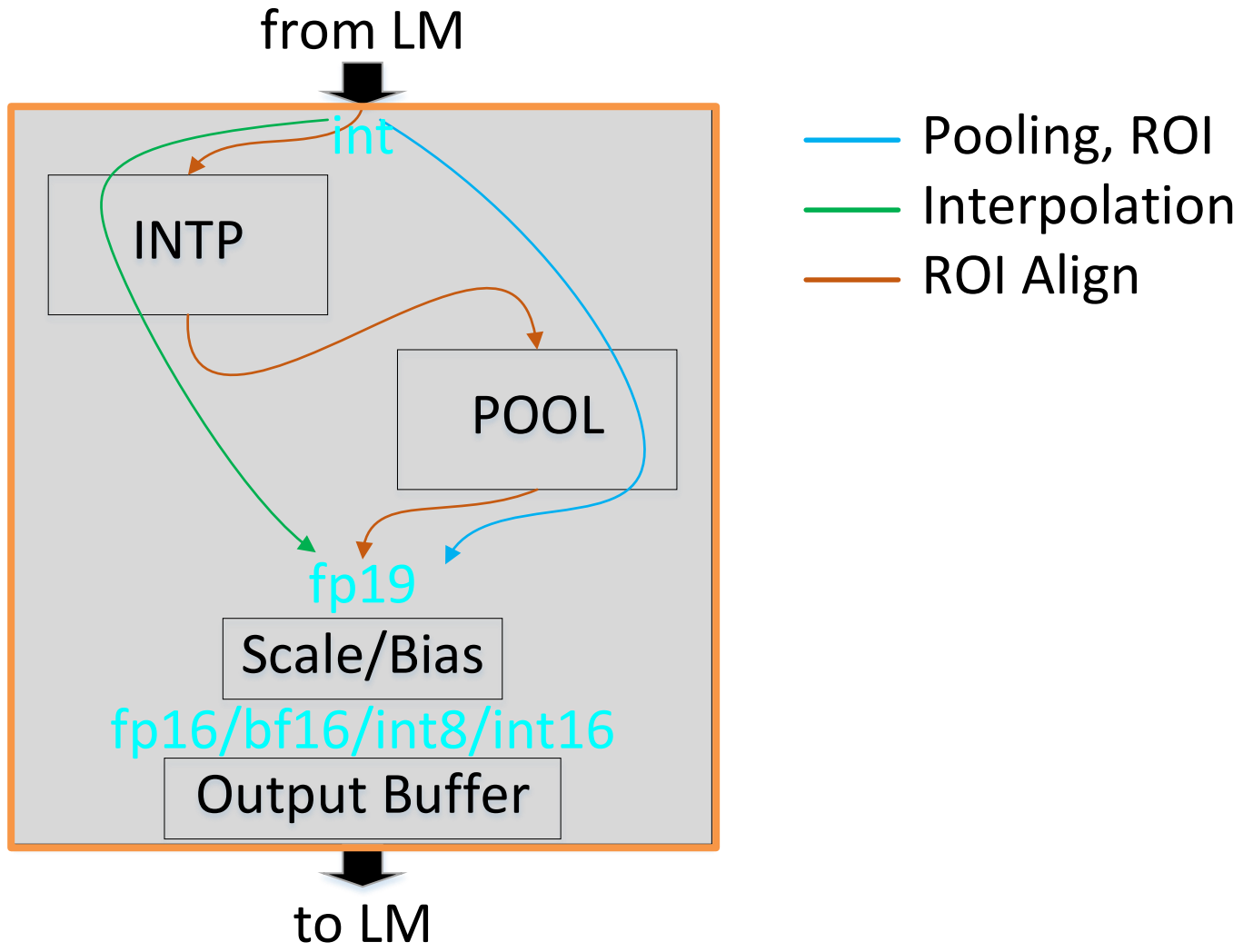
- Weight+Activation Stationary
  - Data reuse
    - via W-buffer, A-buffer, Multicasting
  - Avoid img2col()
    - via sliding-window access
- Fused OPs
  - e.g. CONV-BN-ResidualADD-ReLU

Mapping a CONV2D to Tensor Engine

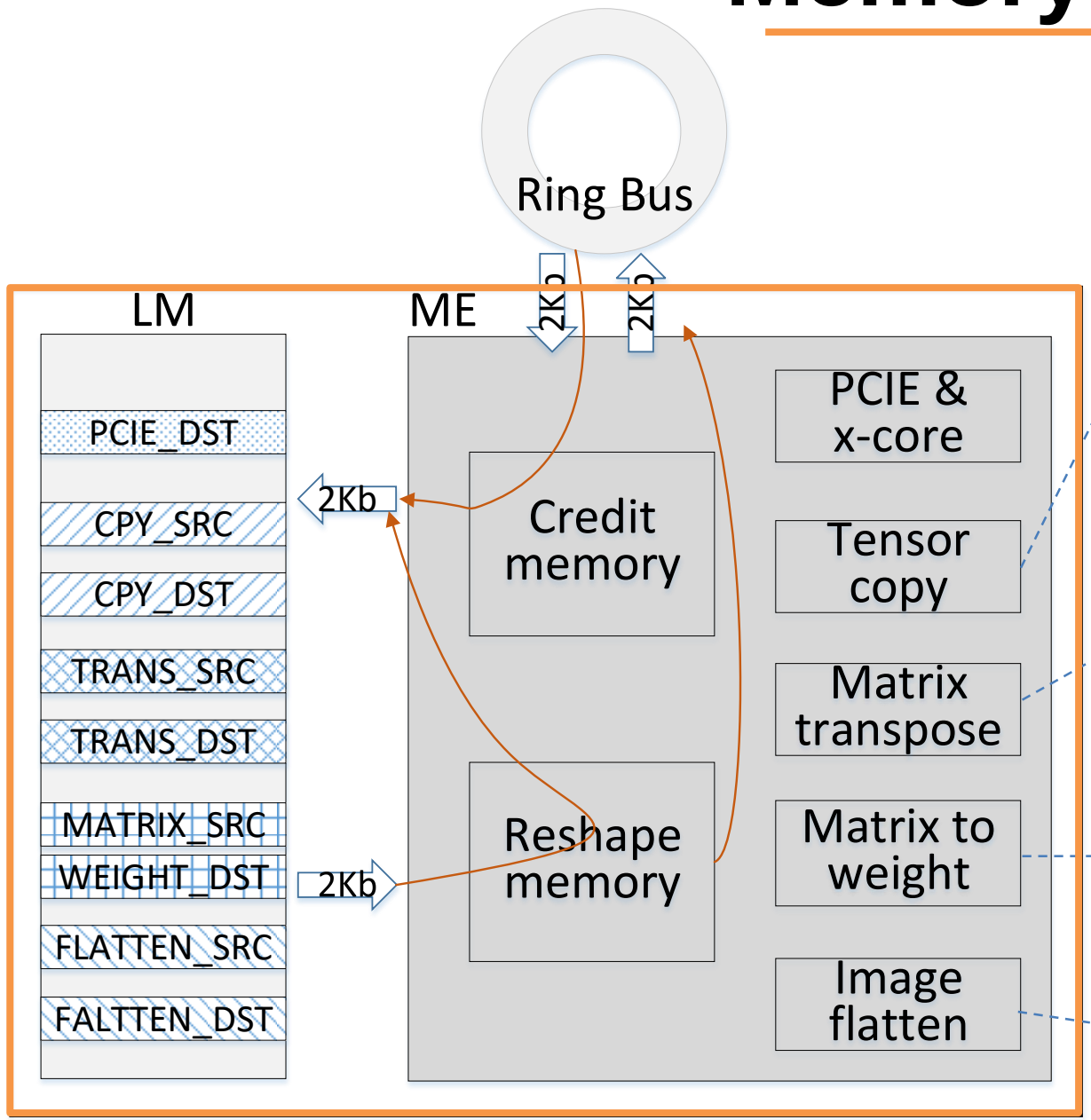


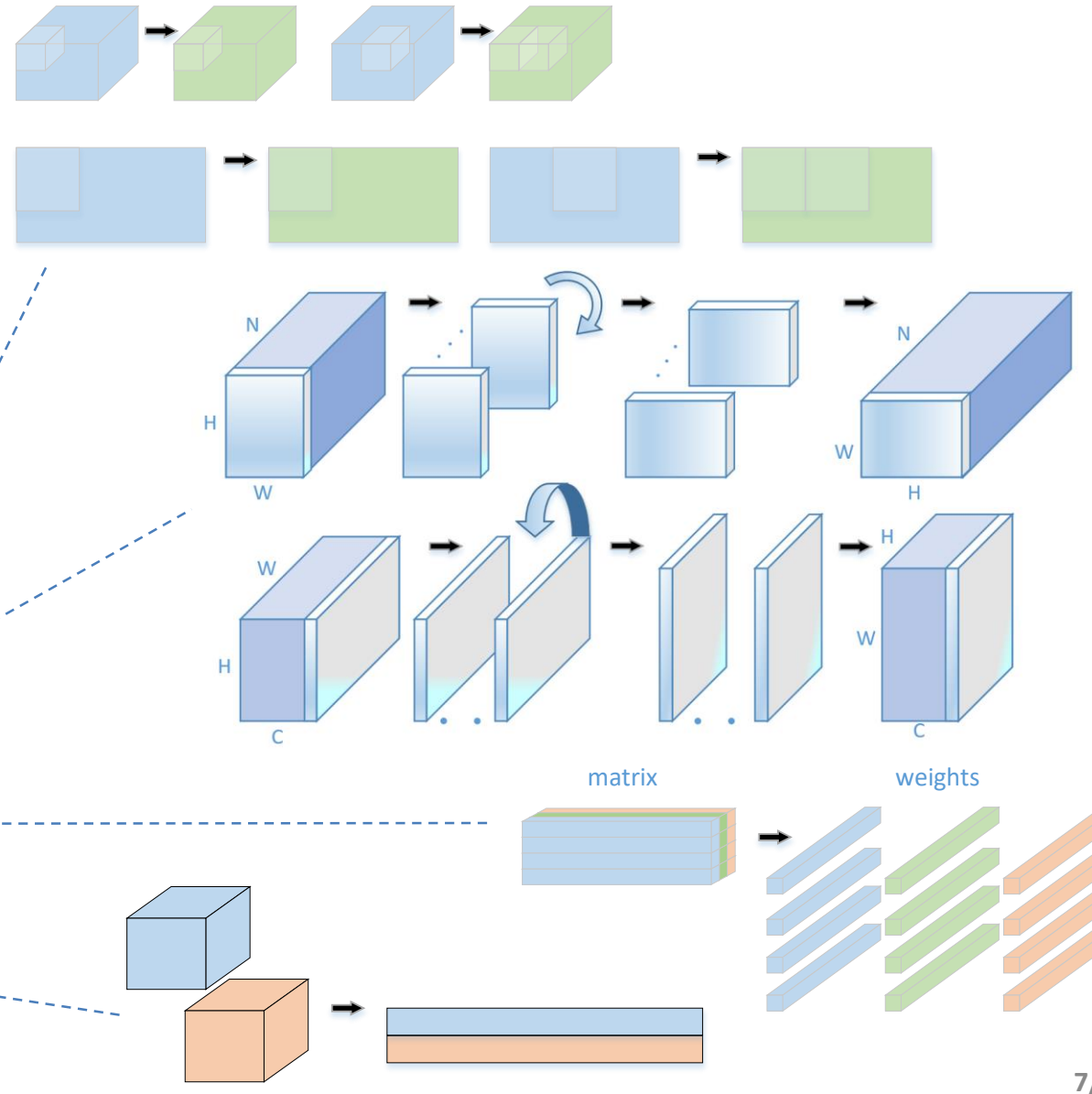
Pooling Engine

- POOL Unit
  - POOLs, ROIs
  - ROI_align_2
- INTP Unit
  - Interpolations
  - ROI_align_1
- Scale/Bias Unit
  - Scaling/Bias operations
  - Data formatting



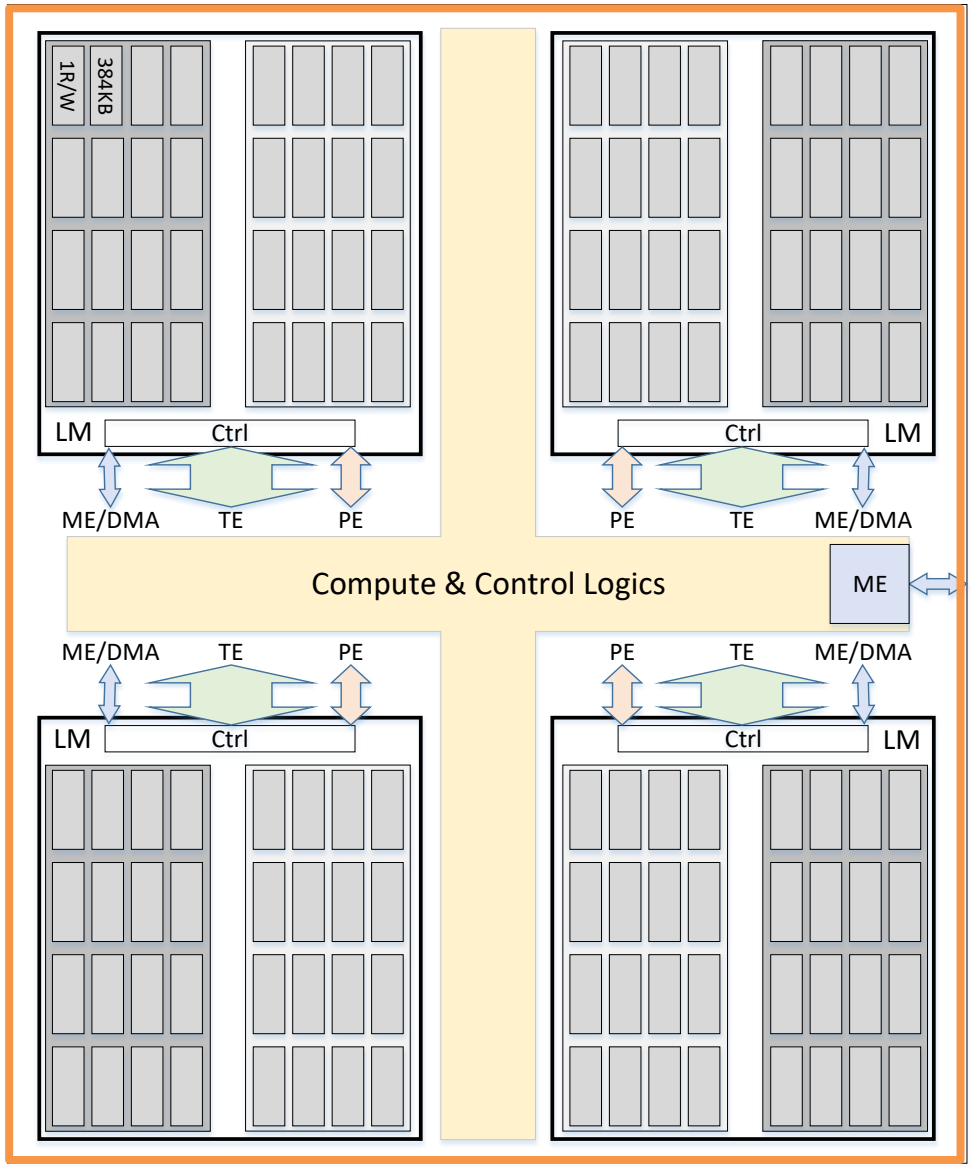
Memory Engine





192MB On-Chip SRAM

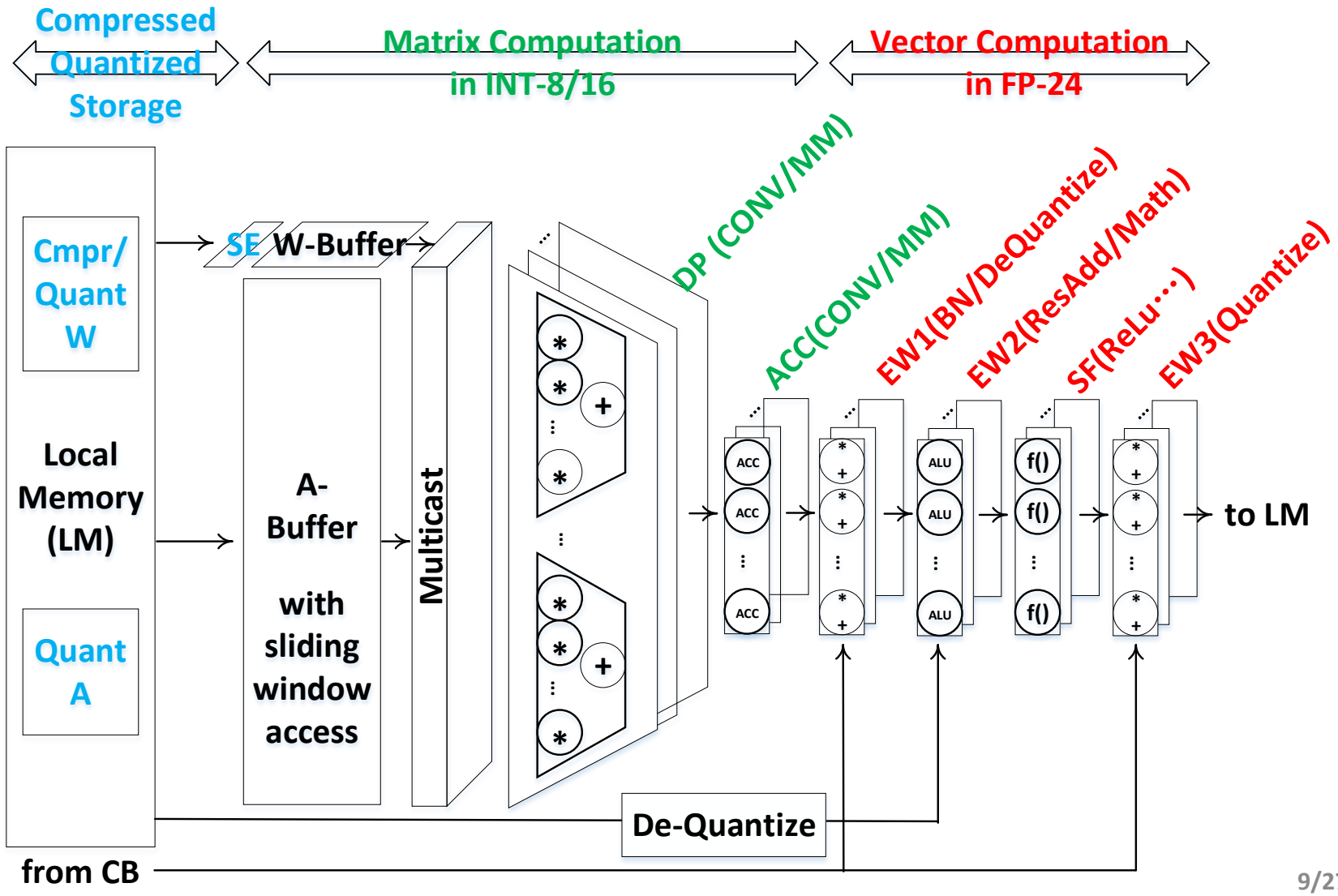
- 1-R/W SRAM Modules
  - High density, low power
- 2 blocks for W & A
  - 16 modules in each block
- 4 clusters
  - Distributed shared
- 48MB in each of the 4 cores
  - Memory copy across cores
    via ME & Ring Bus



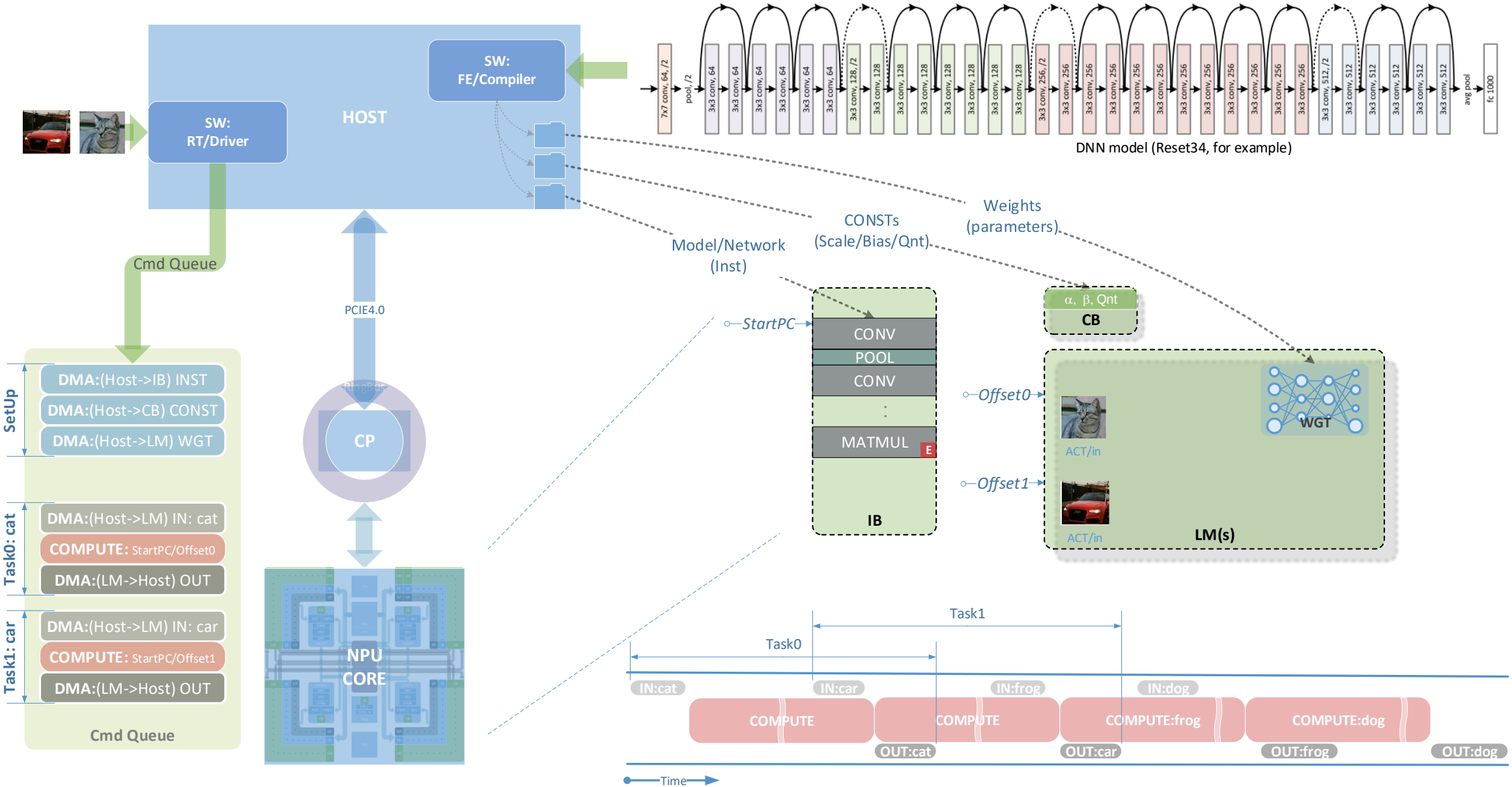
LM Organization in Each Core

Compressed and Quantized Storage/Processing

- **Compressed** Model
  - Sparsity Engine to unpack with bit-masks
  - Pruning is optional though
- **Quantized** computation and storage
- Vector Unit w/ **FP-24**
  - 1sign.8exp.15man

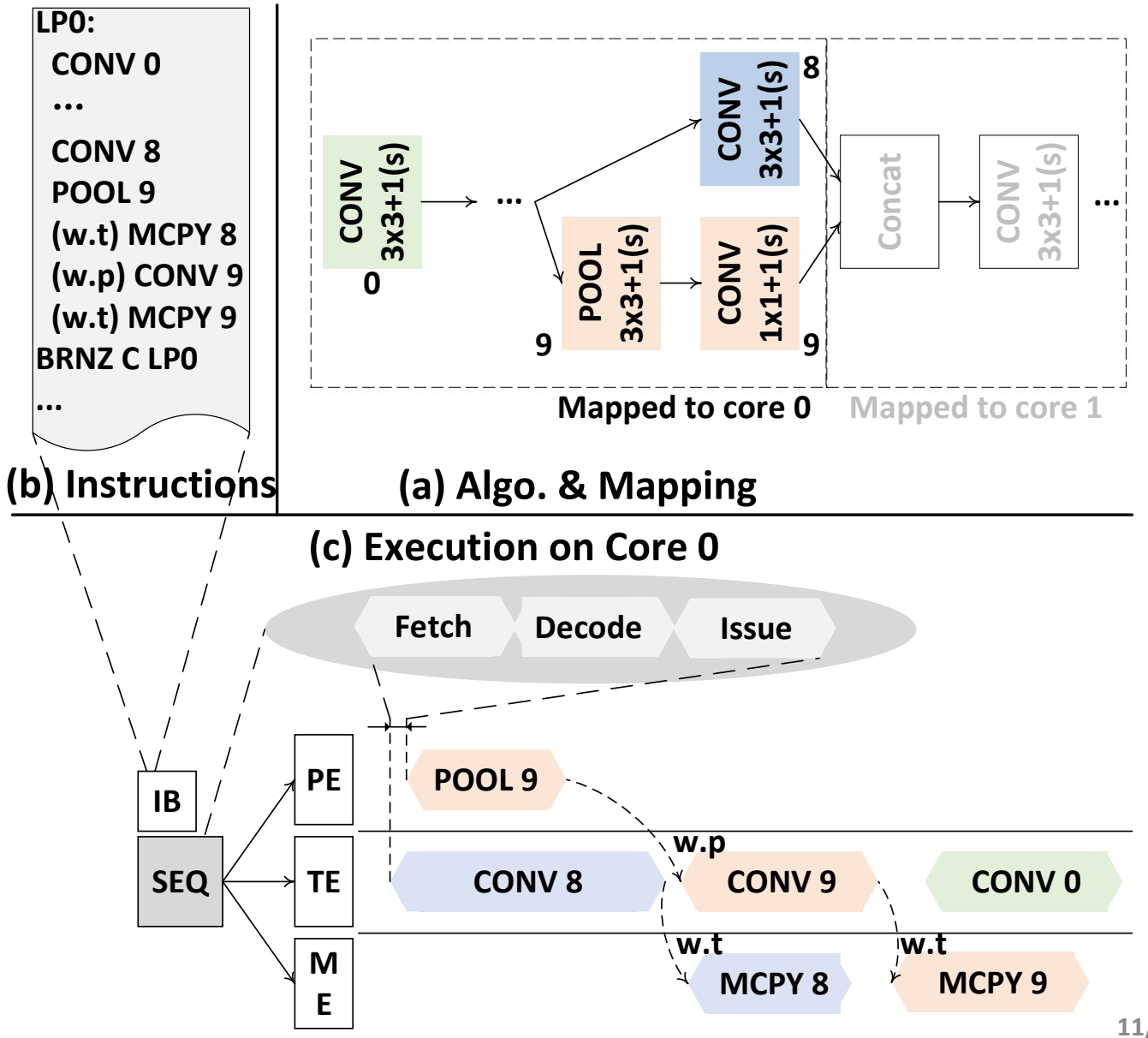


Workflow at Command Level

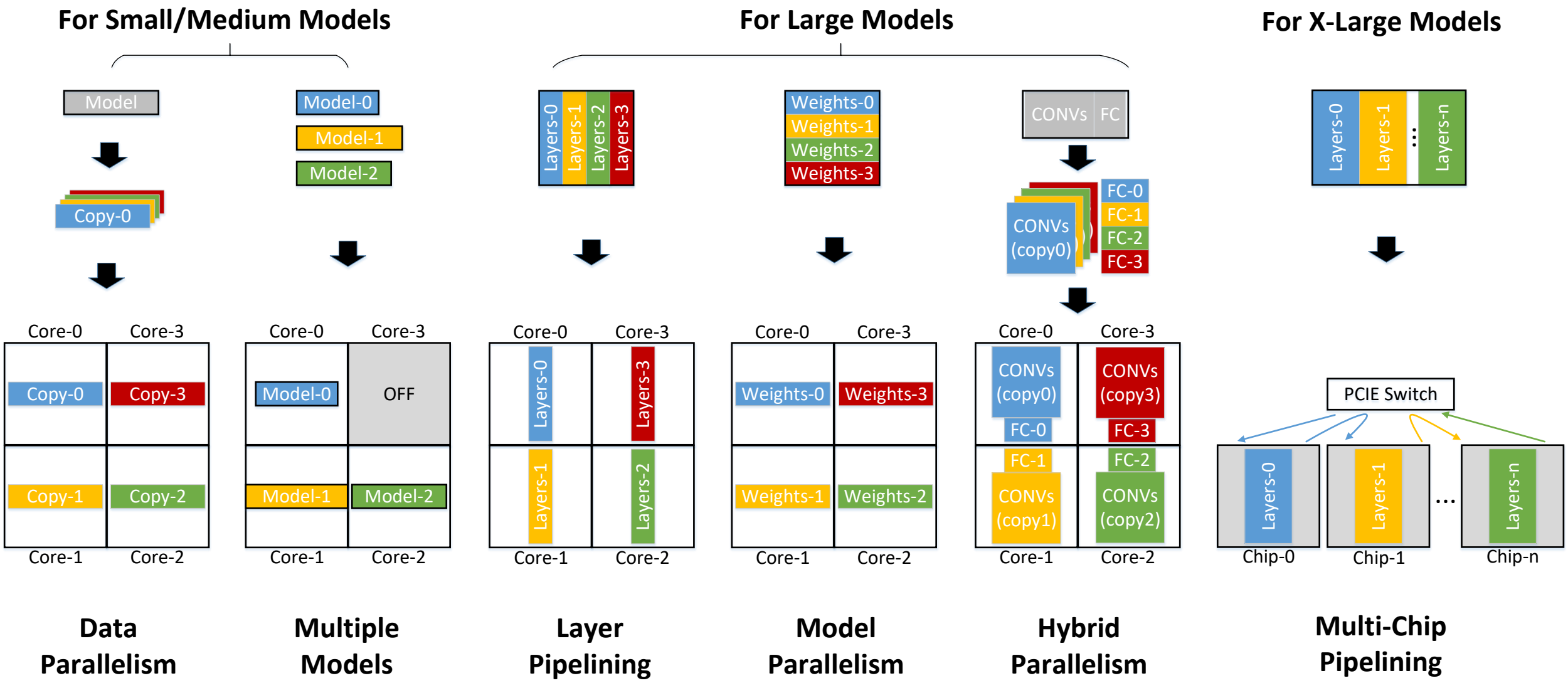


Workflow at Instruction Level

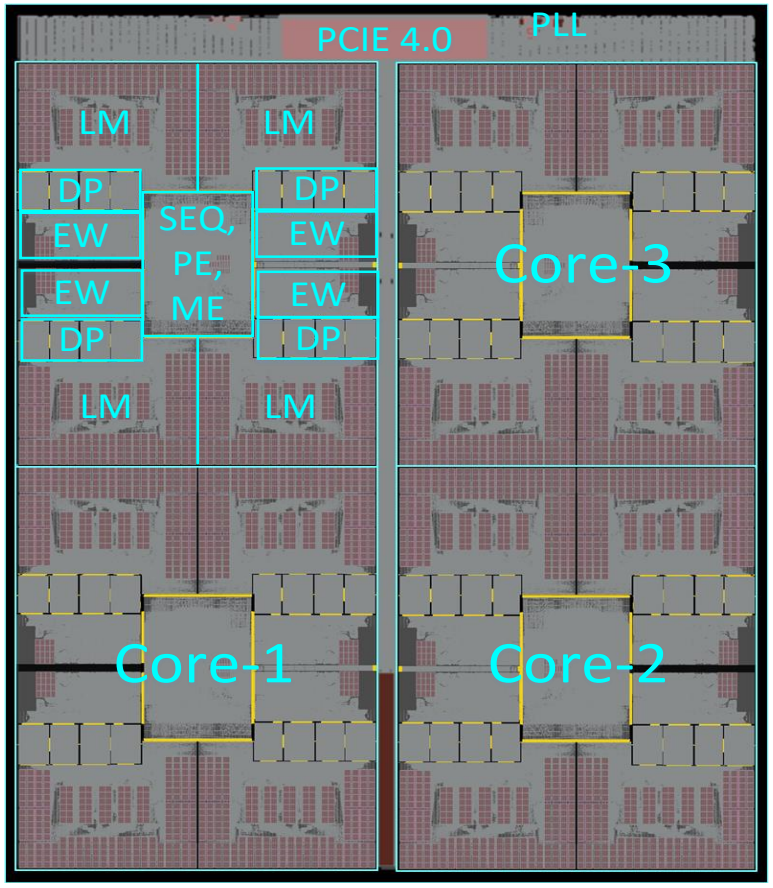
- Domain-specific instruction set
  - CISC-like, operation-fusion
  - Coarse-grain data at tensor level (HW decomposes & ctr)
- Synchronization among 3 engines
  - Embedded bits in instructions
  - Hardware dependency checker



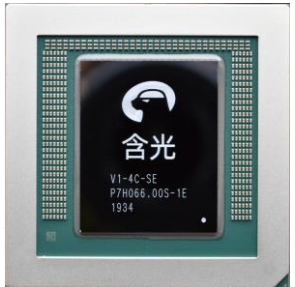
Scalable Task Mapping



Implementation



	NVIDIA T4	NVIDIA V100	Huawei Ascent910	NVIDIA A100	Alibaba Hanguang 800
Peak Performance (INT8 TOPS)	130	125 (FP16)	512	624/1248	825
Frequency (MHz)	1590	1530	1000	1410	700
TDP (Watt)	70	300	350	400	280
Area (mm^2)	545	815	1228	826	709
Process (nm)	TSMC 12nm	TSMC 12nm	TSMC 7+nm	TSMC 7nm	TSMC 12nm

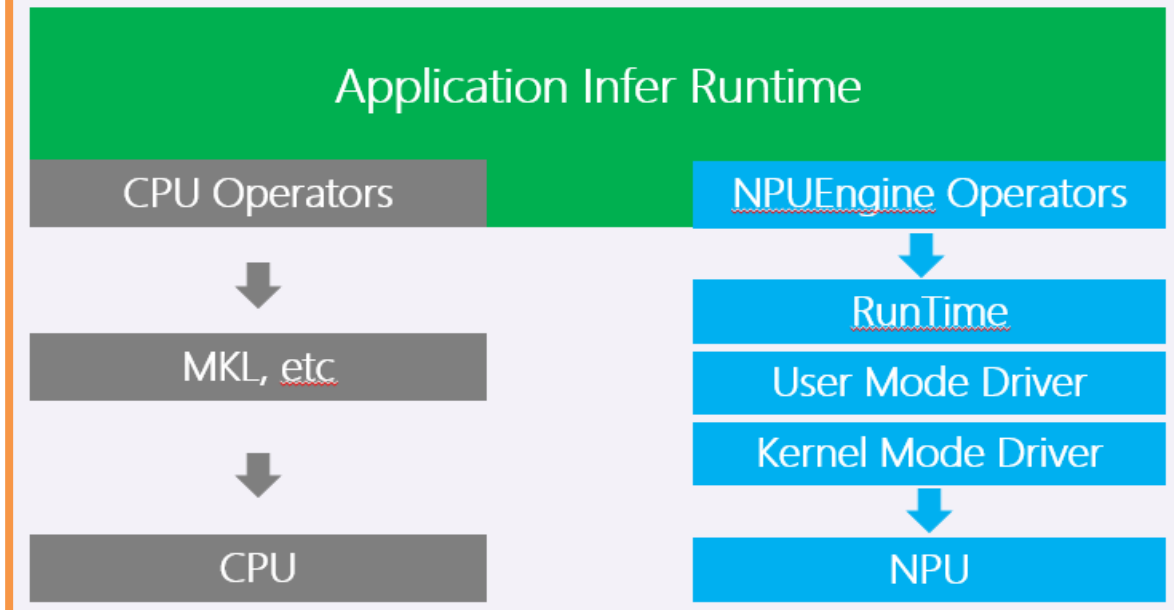


Software Stack

Caffe mxnet TensorFlow ONNX ...

Frameworks

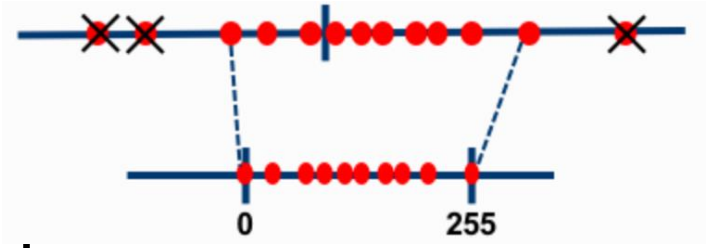
Compiler



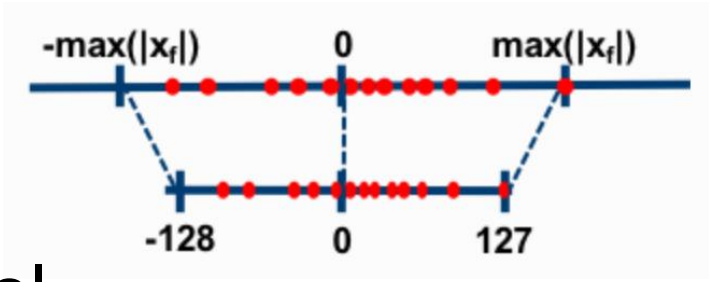
Run-time

Quantization Method

- Post-training quantization
  - Pruning is optional
- Static quantization at compile time
- Clip out-of-range post-rounding values



- Symmetric quantization

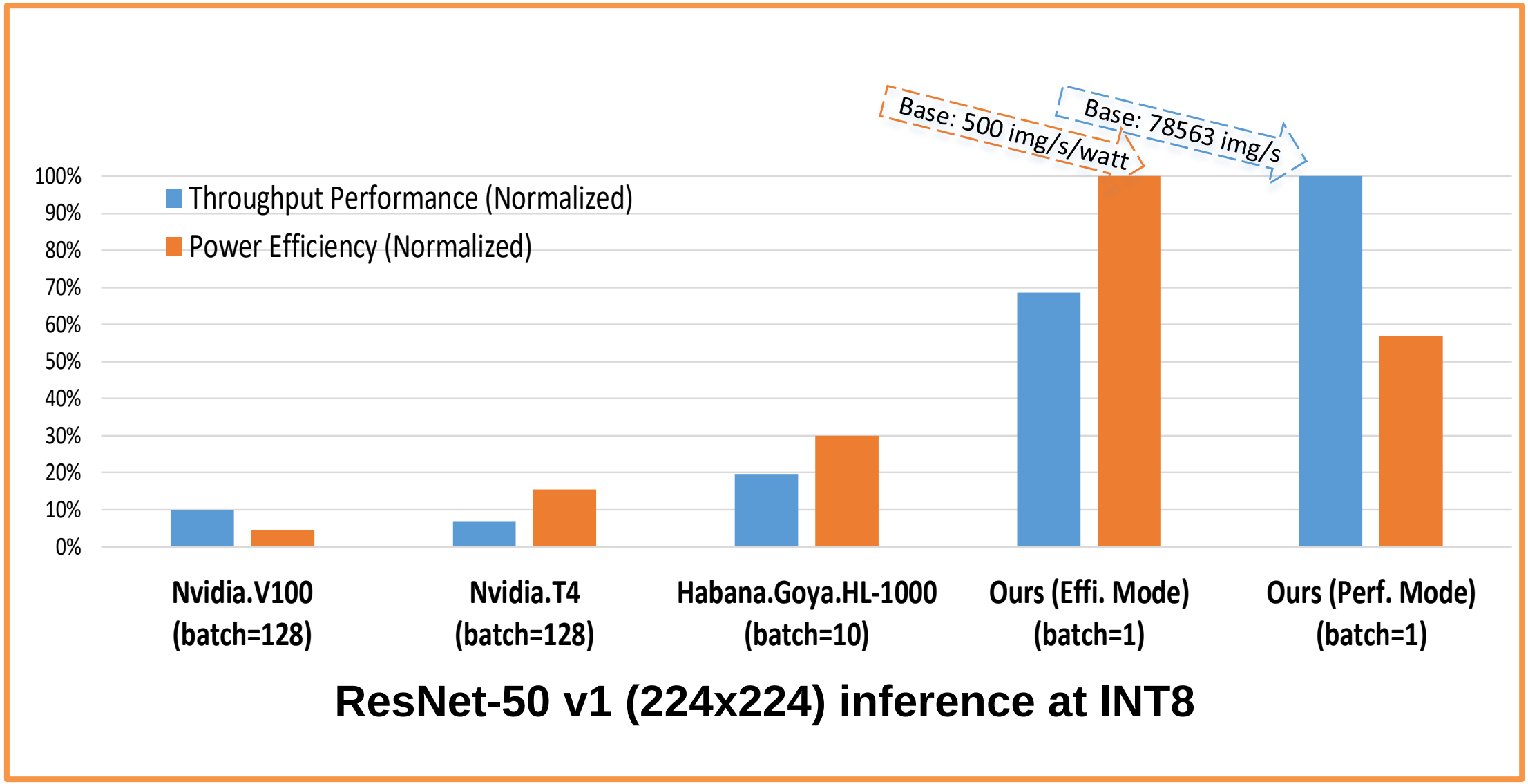


- Weights: per-channel
- Activations: per-tensor

Quantization Accuracy

Model	GPU FP32	NPU INT8
resnet_v1_50 (Top1/Top5)	75.18 / 92.18	74.93 / 92.47
resnet_v1_101 (Top1/Top5)	76.40 / 92.90	76.02 / 93.04
resnet_v1_152 (Top1/Top5)	76.80 / 93.20	76.45 / 93.41
resnet_v2_50 (Top1/Top5)	75.57 / 92.82	75.11 / 92.92
resnet_v2_101 (Top1/Top5)	77.00 / 93.70	76.58 / 93.81
inception_resnet_v2 (Top1/Top5)	80.40 / 95.30	80.19 / 95.26
inception_v3 (Top1/Top5)	77.99 / 93.94	77.86 / 94.07
inception_v4 (Top1/Top5)	80.2 / 95.2	80.11 / 95.20
ssd_resnet_50_fpn (mAP)	37.12	36.96
ssd_inception_v2_coco (mAP)	24	26.27
faster_rcnn_resnet50_coco (mAP)	25.89	25.96
faster_rcnn_inception_v2_coco (mAP)	24.8	24.66
mask_rcnn_resnet50_atrous_512 (mAP)	20.77	20.8

Performance & Efficiency



ResNet-50 v1 (224x224) inference at INT8

Scalable Performance and Power

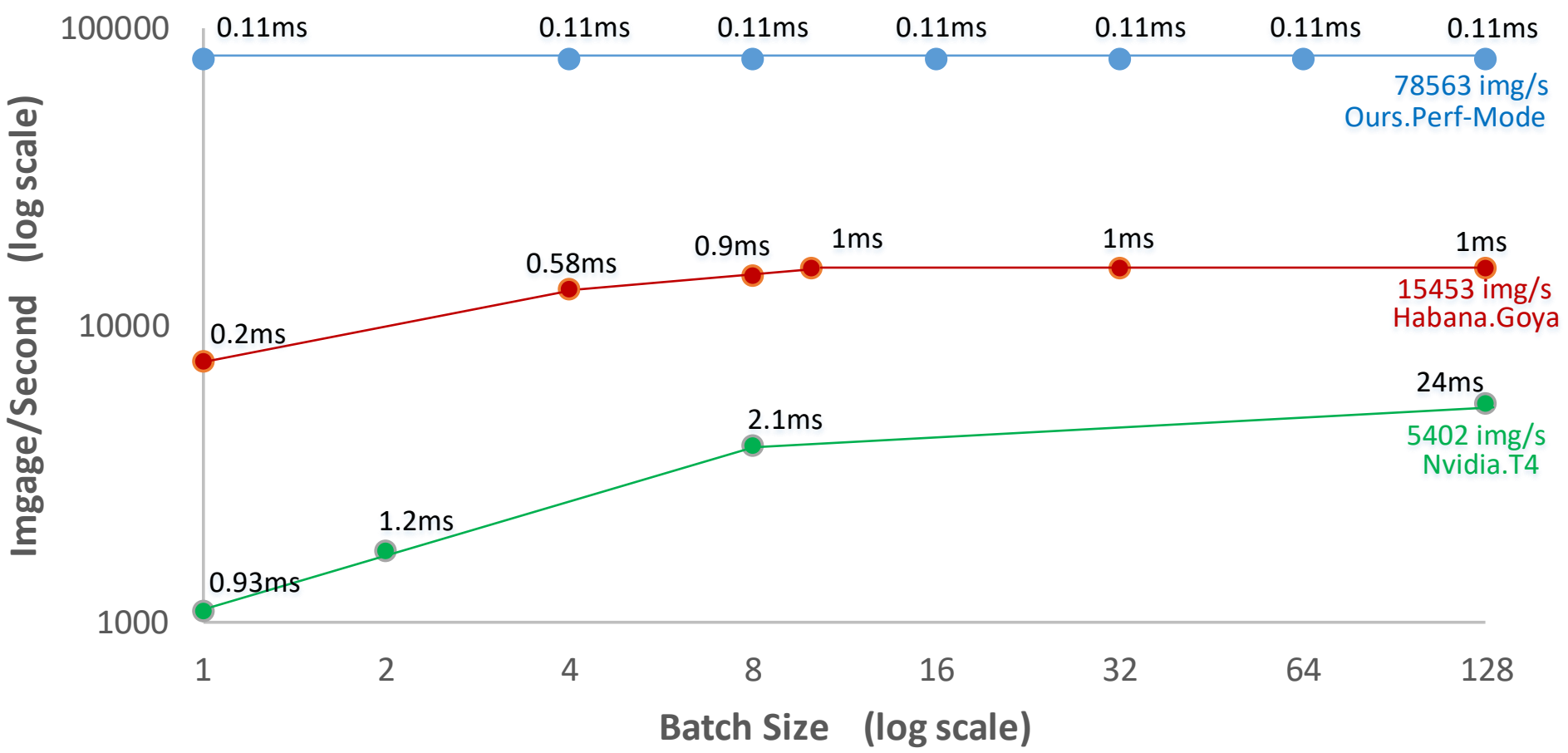
- Configurable frequency and voltage
- Can be deployed from data center to large edges

ResNet-50 v1

Configuration	Performance Mode	Efficiency Mode	Low Power Mode 1	Low Power Mode 2	Low Power Mode 3
Frequency (MHz)	700	475	475	475	475
Perf (img/sec)	78,563	53,983	37,500	25,000	12,500
Power (watt)	276	108	75	50	25

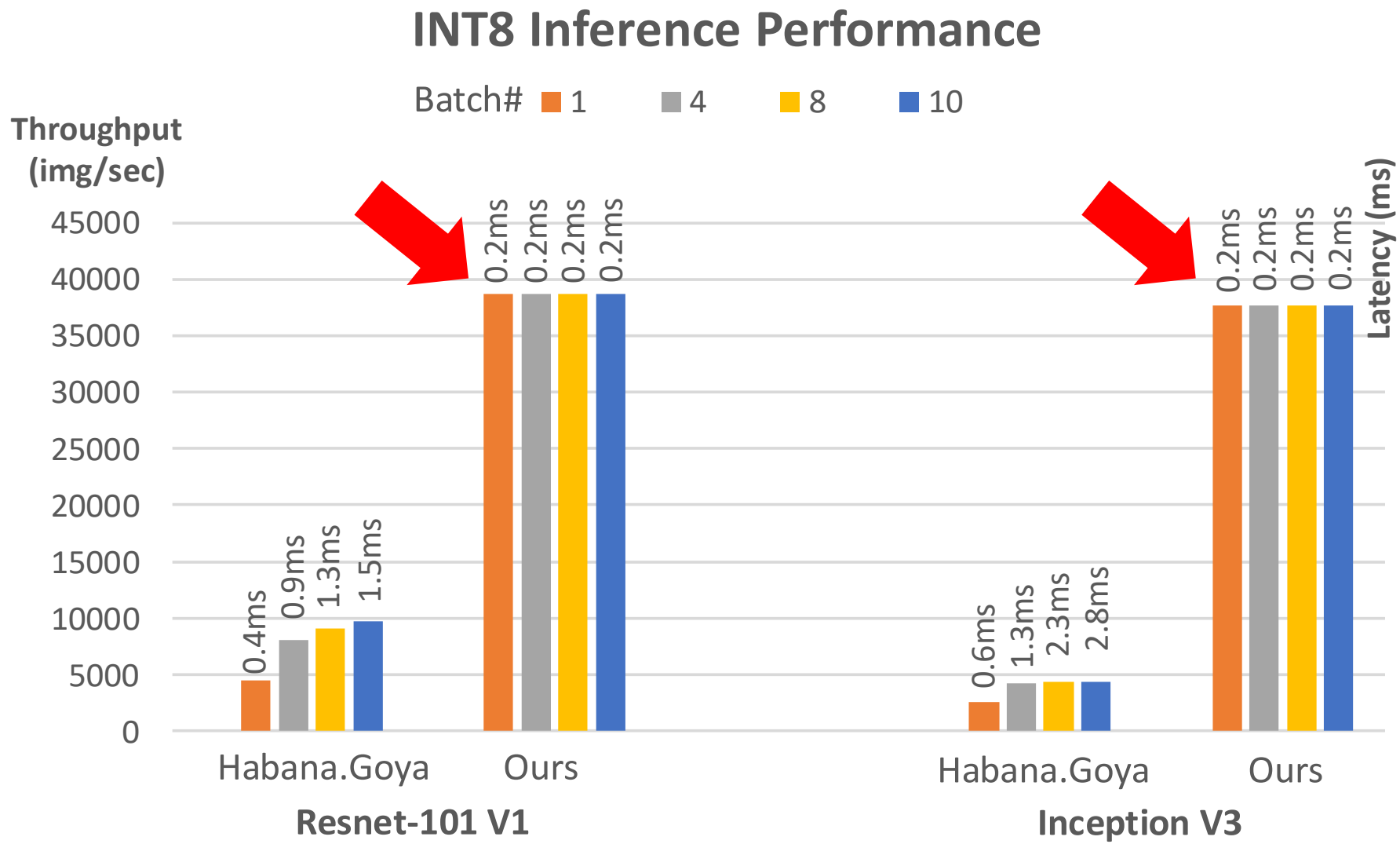
Batching-Independent Performance

Best in class latency, no performance drop



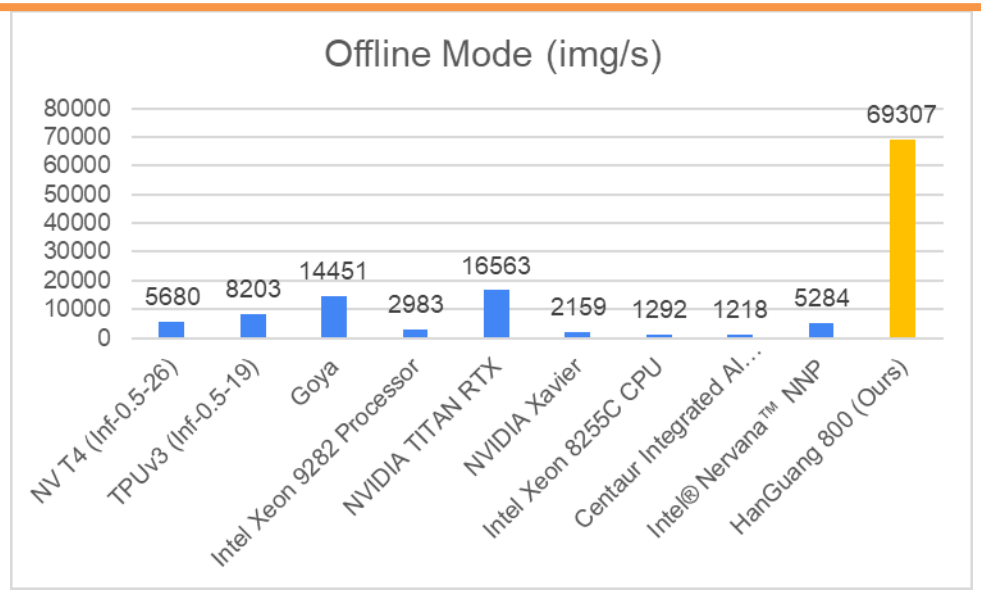
ResNet-50 v1 (224x224) inference at INT8

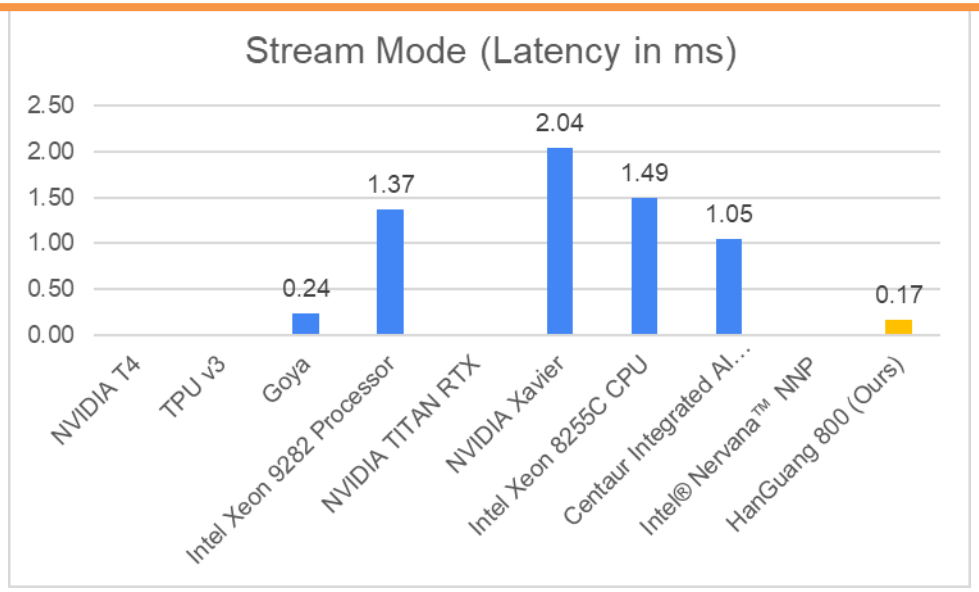
More Results

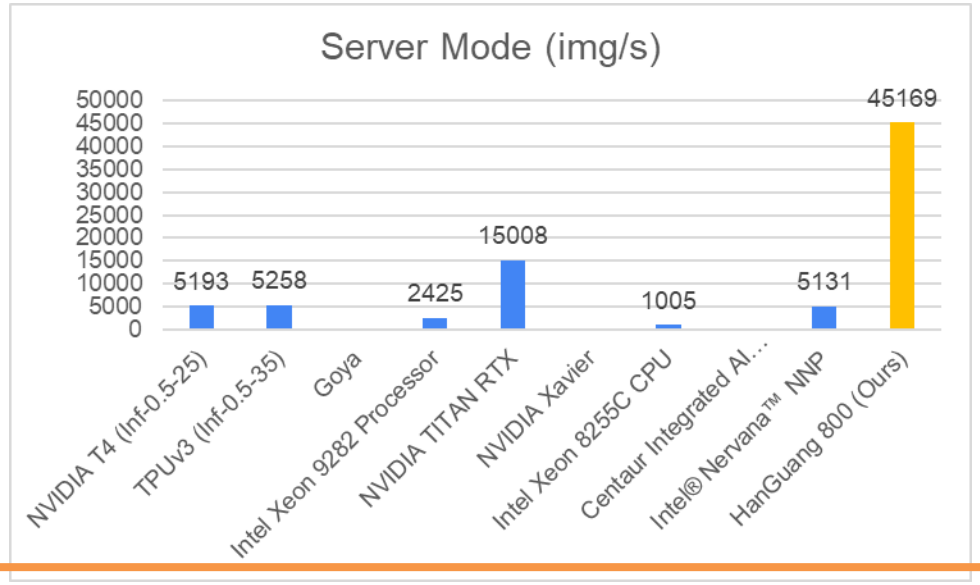


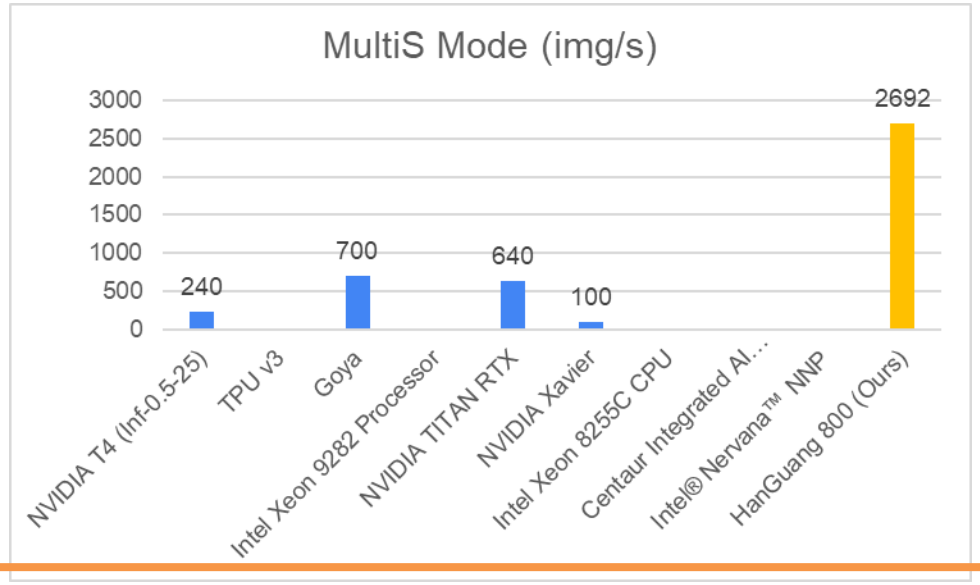
MLPerf Inference V0.5 (Single-Chip Performance*)

ResNet-50 v1.5



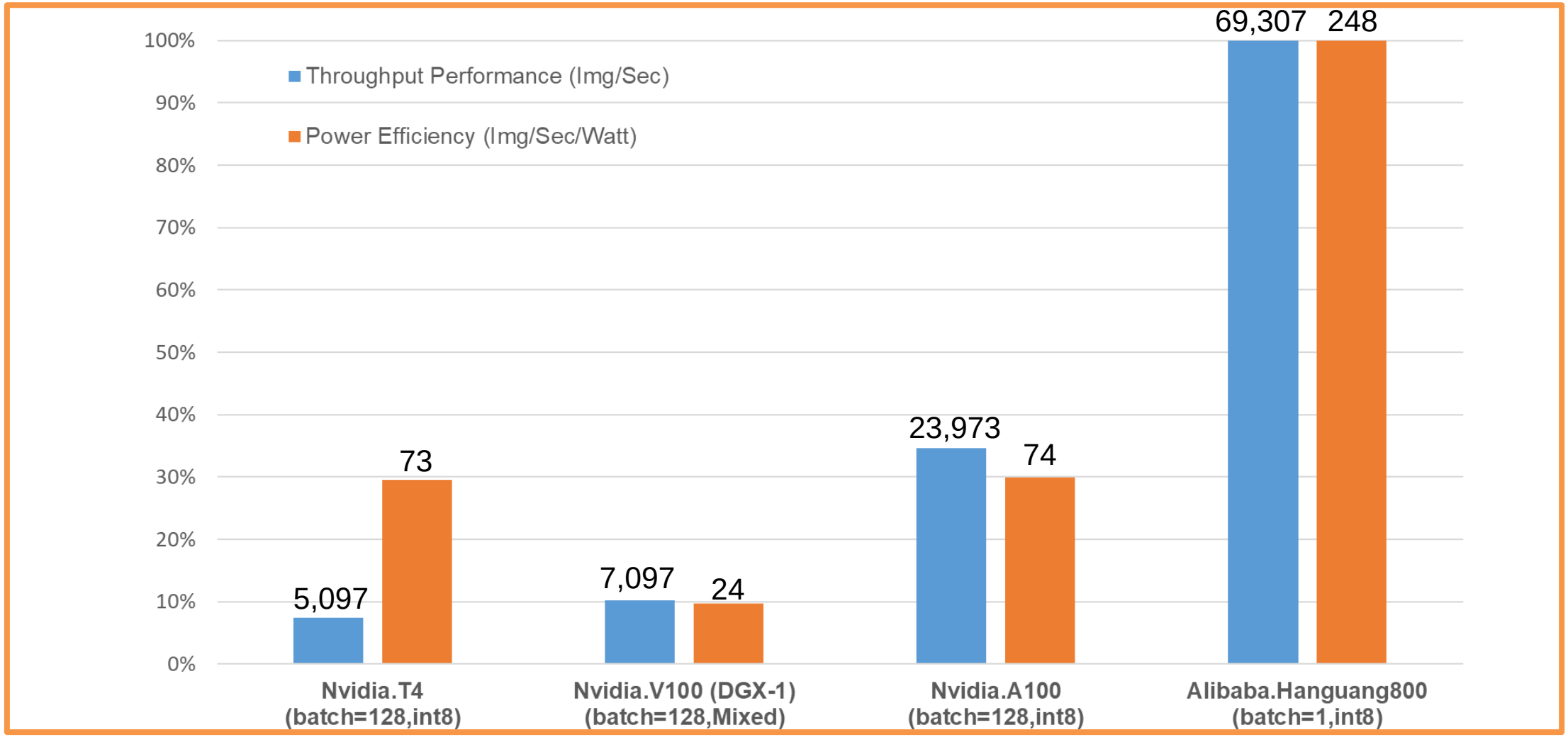






** Single-Chip Performance is not the primary metric of MLPerf. MLPerf name and logo are registered trademarks. See www.mlperf.org for more information.*

Comparing with Latest GPUs



ResNet-50 v1.5 Inference

Nvidia GPU Data Source: NVIDIA Data Center Deep Learning Product Performance

Deployment

Smart City

Image Search

Smart Fashion

5G Edge Cloud

Smart Medical

Video Cloud

Recommender Sys

Smart Manufacture

Smart Retail

In Data Centers

In Edge Servers

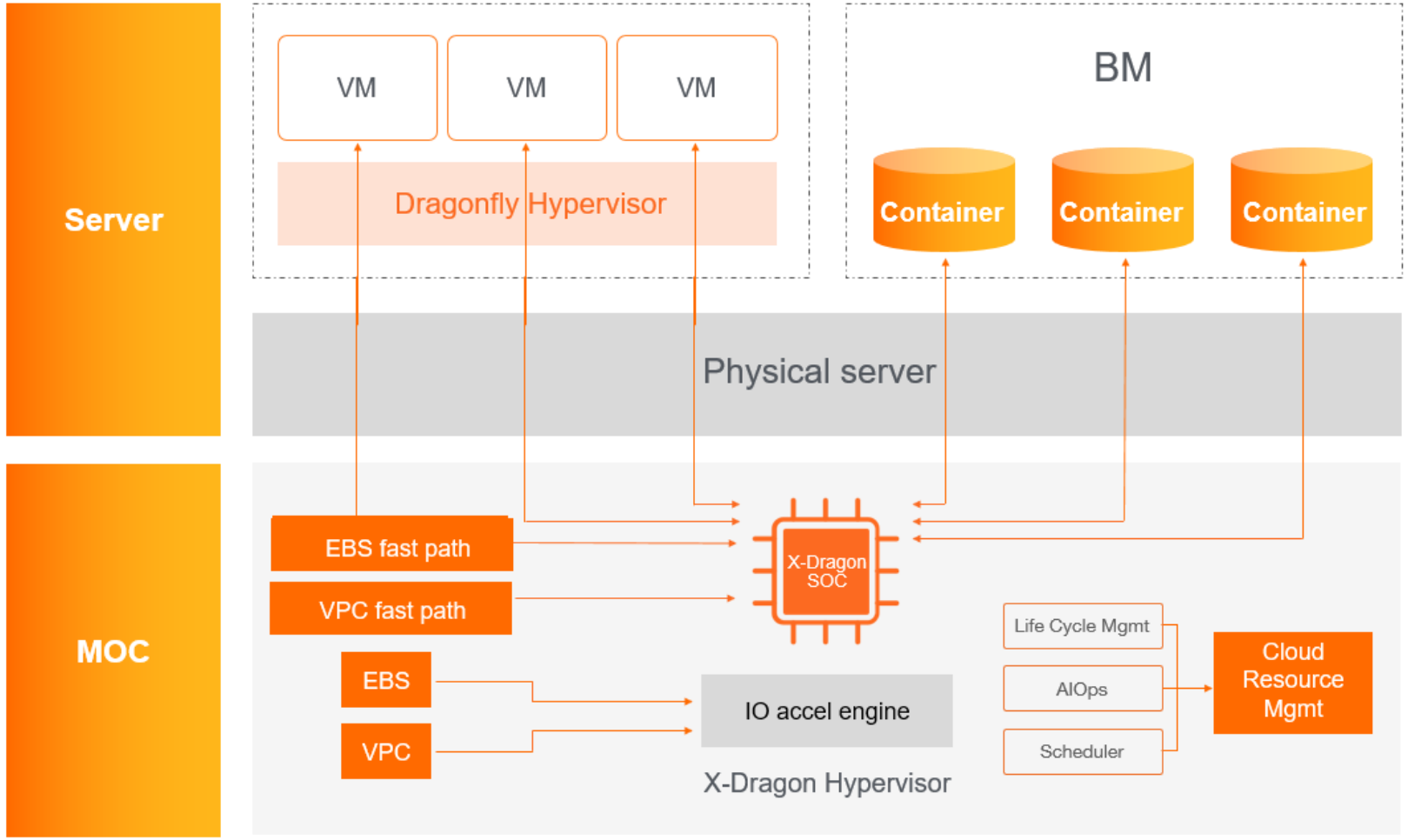
In Large End-Devices

Alibaba AI Cloud

Hanguang NPU Inside

Elastic Compute Service in Alibaba Cloud

X-Dragon architecture provides unified resource pool for Virtual Machines, Bare-Metals, and Containers



X-Dragon SoC

Higher Performance, Lower Latency, and Lower Jitter

H/W Based QoS

Multi-tier VPC and EBS QoS
Multidimensional Accurate Scheduling

Enhanced Network

2 x 50 Gb, Dual Uplink
BM Level Latency

Enhanced H/W Queues

1024 EBS + 1024 VPC Queues
Enhanced Cloud Service Isolation

Further Reading: "Introducing the Sixth Generation of Alibaba Cloud's Elastic Compute Service"

Elastic Bare-Metal AI Inference Instance

Public Cloud: ecs.ebman1.24xlarge

- CPU: Cascade Xeon 104 cores
- Memory: 384GB
- NPU: 4x 2-core Hanguang 800

Private Cloud

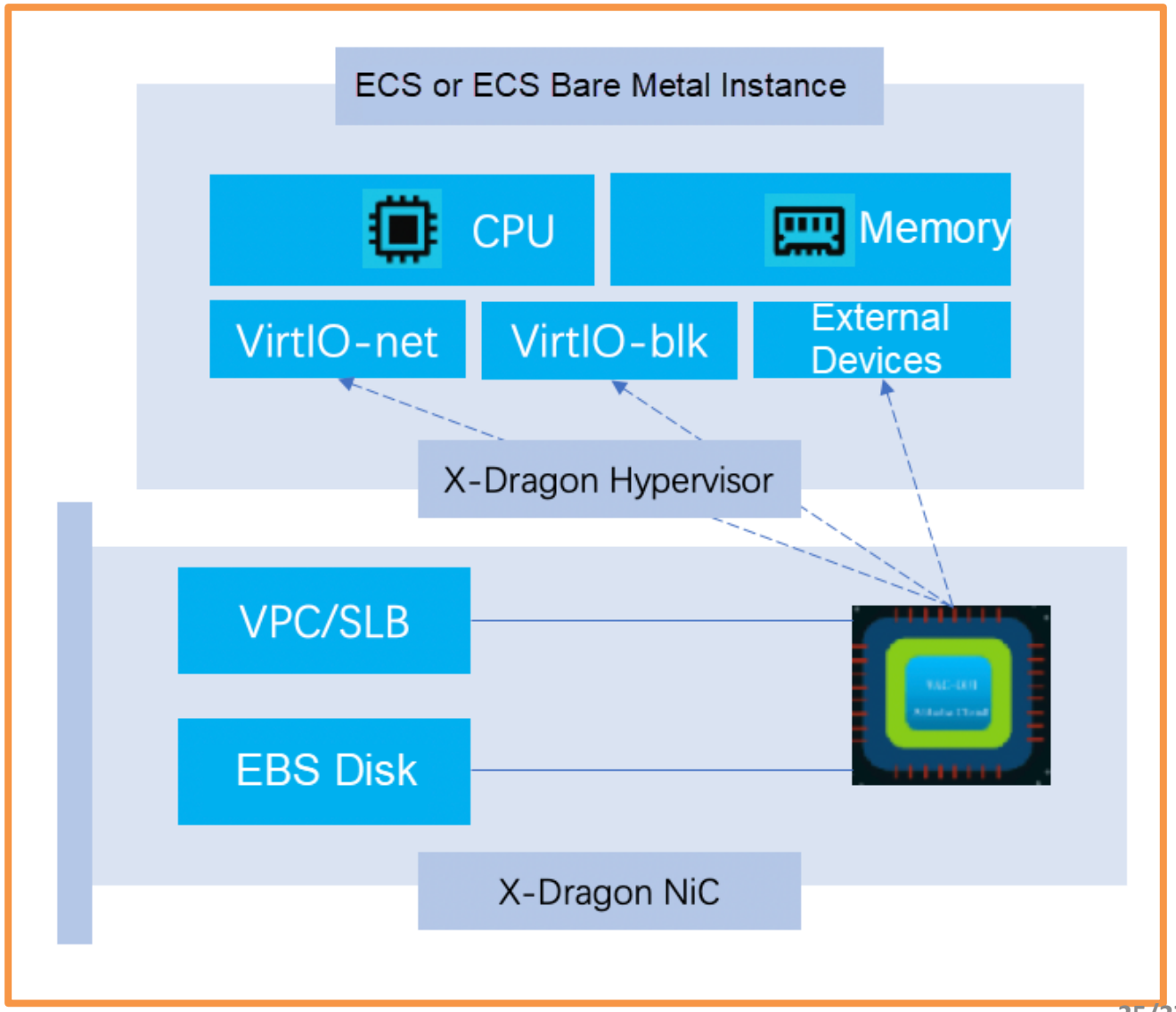
- Customized Configurations
- 1/2/4x 2-core/3-core/4-core Hanguang 800

X-Dragon Bare-metal Server

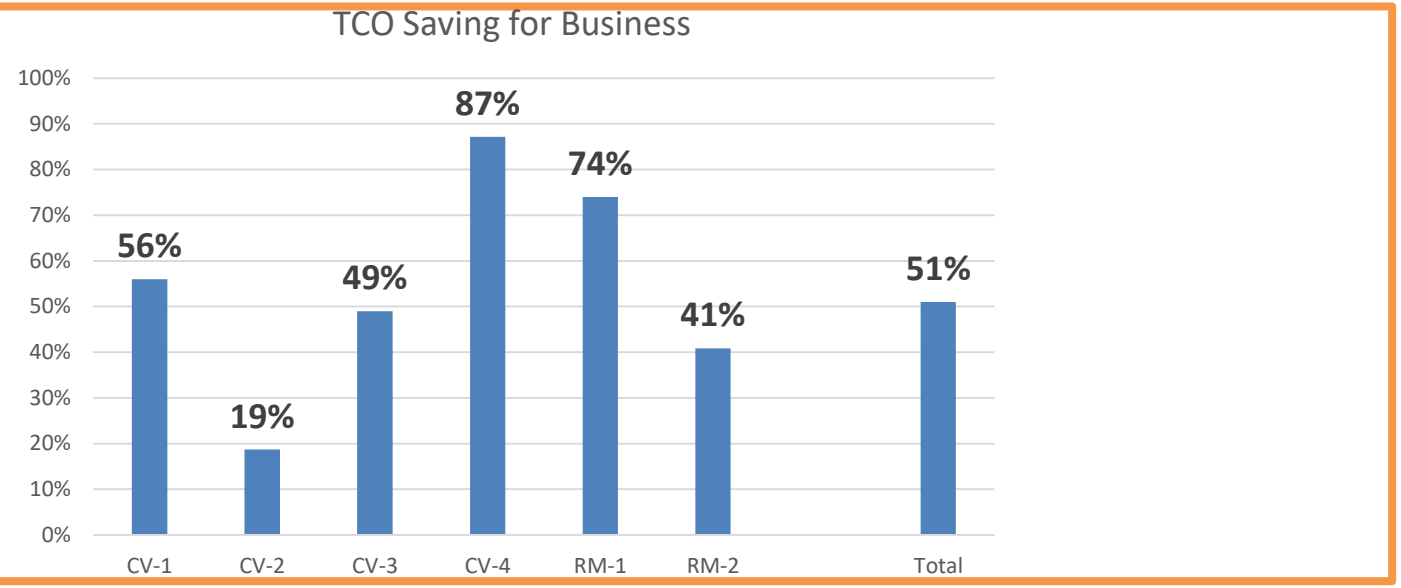
- Bare-metal performance
- Elasticity of Cloud Infrastructure

Support Common Framework

- Tensorflow
- MXNet
- PyTorch



Deployed Applications via Alibaba Cloud



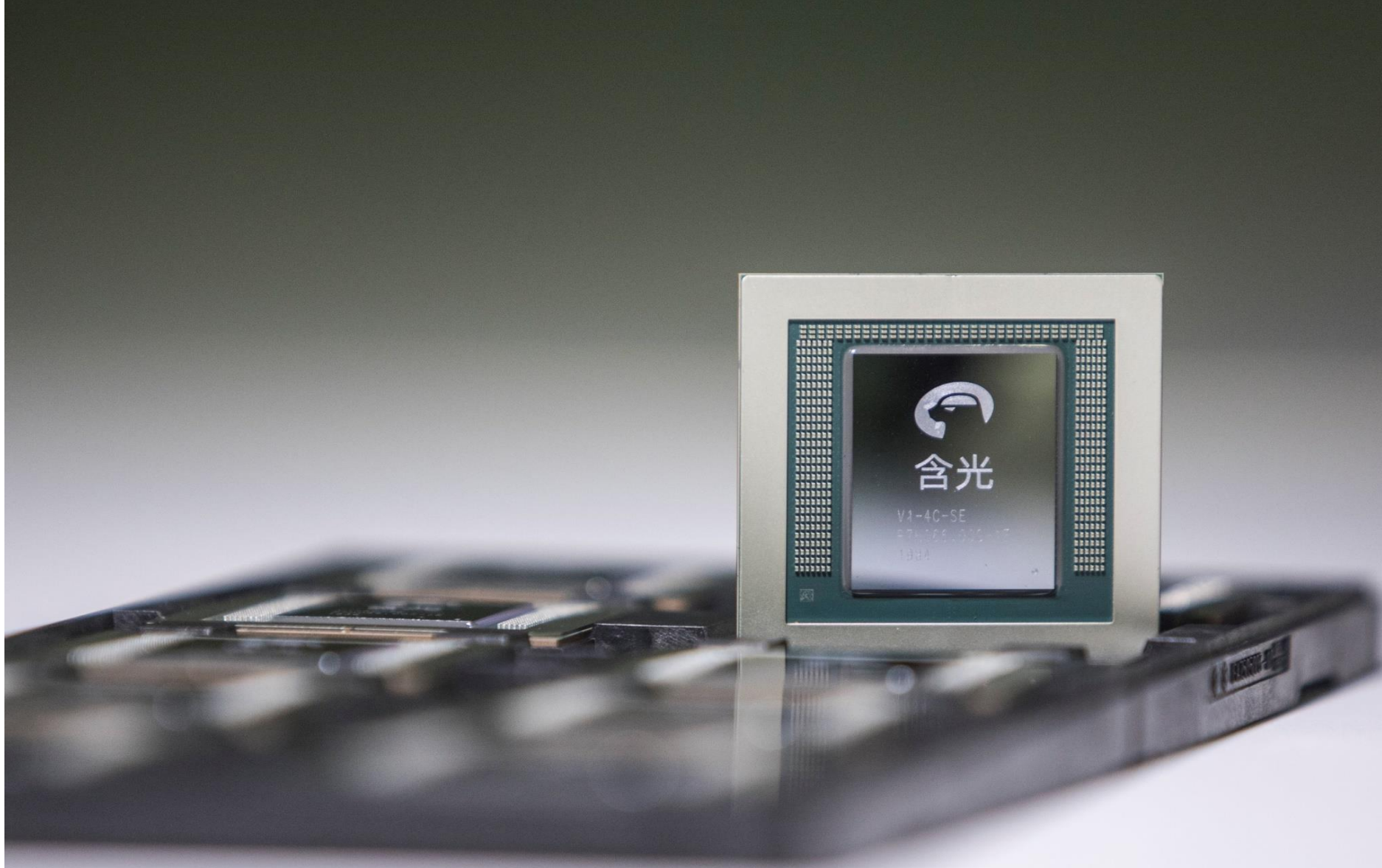
Top 6 businesses representing 99.1% deployment of Hanguang 800 NPUs.

Business	Major App. Type	GPU Server	NPU Server	Deployment %
CV-1	Image/Video	2x T4	2x 3-core	8.2%
CV-2	Image/Video	8x T4	2x 3-core	28.5%
CV-3	Image/Video	8x T4	2x 3-core	12.5%
CV-4	Image/Video	4x T4	4x 2-core	12.5%
RM-1	Recommender	1x T4	1x 4-core	12.5%
RM-2	Recommender	1x T4	1x 4-core	25.1%

Summary

- Hanguang 800 NPU takes lead in AI inference performance.
- Fully integrated with X-Dragon compute platform, Hanguang instances provide both performance of bare-metal and elasticity of cloud infrastructure.
- Being deployed to cloud customers, Hanguang estimates to lower system TCO by ~50% for various business applications.
- Stay tuned for more Hanguang powered Alibaba cloud services.

Q&A



上古三创 一曰含光

视不可见 运之不知其所触

泯然无际 经物而物不觉

Thank You!