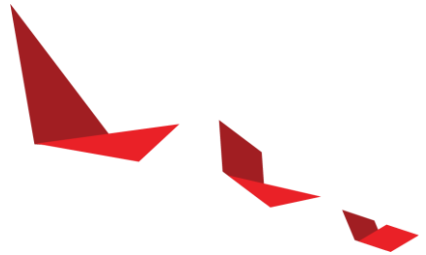
XILINX

深度学习算法在赛灵思 7nm Versal板卡上的优化与部署

Dr. Fan Zhang
Software & AI Tech Marketing

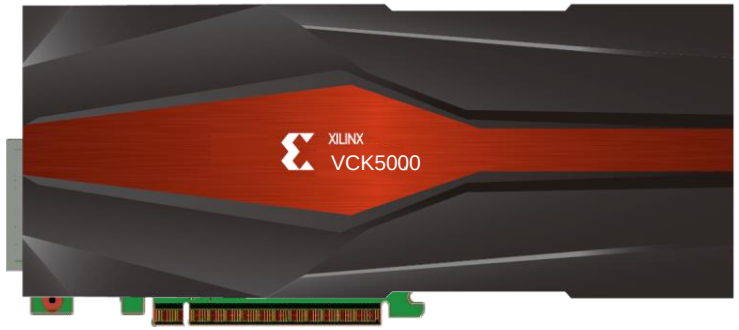
Agenda



- Overview of Xilinx 7nm Versal Card (VCK5000)
- IP Overlays of Deep learning Processing Unit (DPU)
  - for CNN
  - for MLPerf
  - for BERT
- Quick demo on VCK5000
- Summary

VCK5000: 数据中心加速板卡

- Vitis™ platform supported flow for data center, machine learning, and HPC application development
- Product support advisories
  - No Vivado® Design Suite based examples
  - No Petalinux BSP or collateral
  - Headstart program acceptance and FAE resource support required to purchase

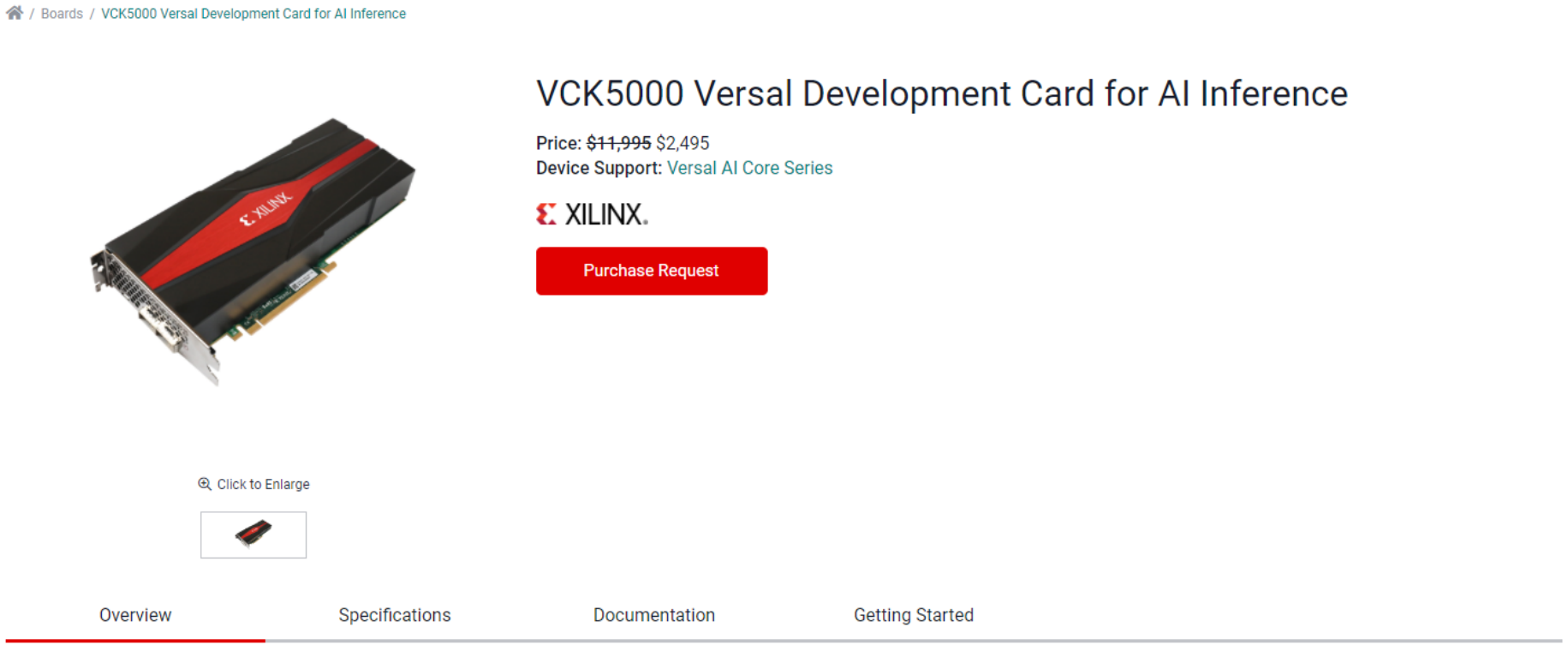


Features	
FPGA	Versal VC1902
Device Speed/Voltage	-2M
Width	Dual slot
Form Factor	FH3/4L Passive FHFL Active
Memory	16G DDR4-3200
PCIe	2x Gen4x8, 1x Gen4x8, 1x Gen3x16
Network I/F	2x QSFP28
Thermal	Passive or Active options
Power (Max TDP)	300W
Compute	400 AI Engines for optimized CNN/DNN performance
Shell	Hardened QDMA, DDR4 controller, and NoC

Schedule	
ES1 Early Access Shipping:	Now
Production Silicon Card Shipping:	Q3 ’21

VCK5000: 数据中心加速板卡

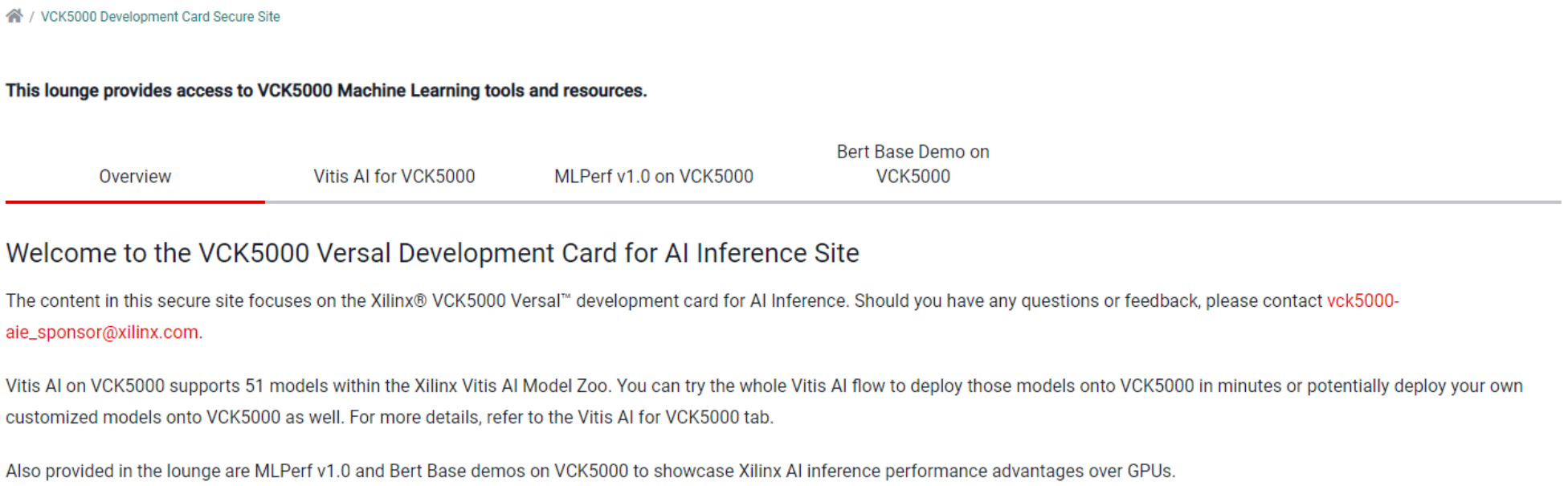
- VCK5000 Lounge
  - https://www.xilinx.com/products/boards-and-kits/vck5000.html



XILINX.

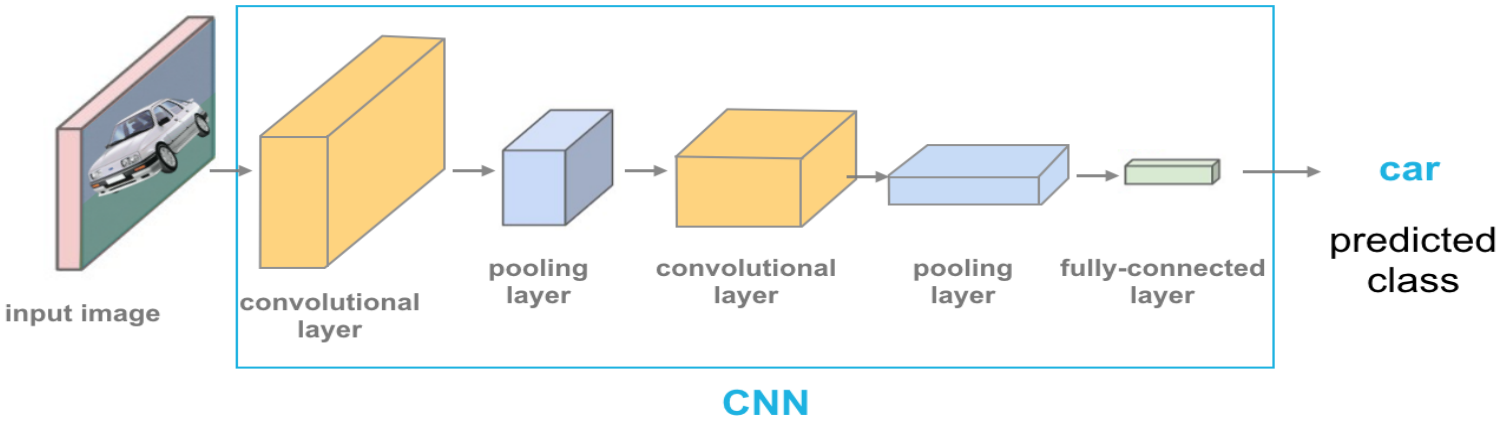
VCK5000: 数据中心加速板卡

- AI Inference Lounge on VCK5000
  - https://www.xilinx.com/member/vck5000-aie.html#overview

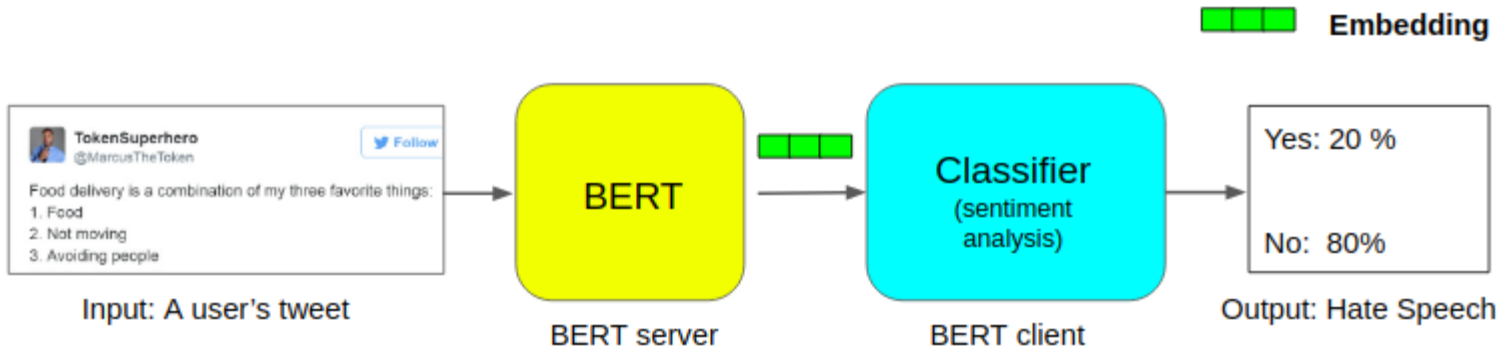


为什么需要多套硬件架构?

- CNN



- BERT

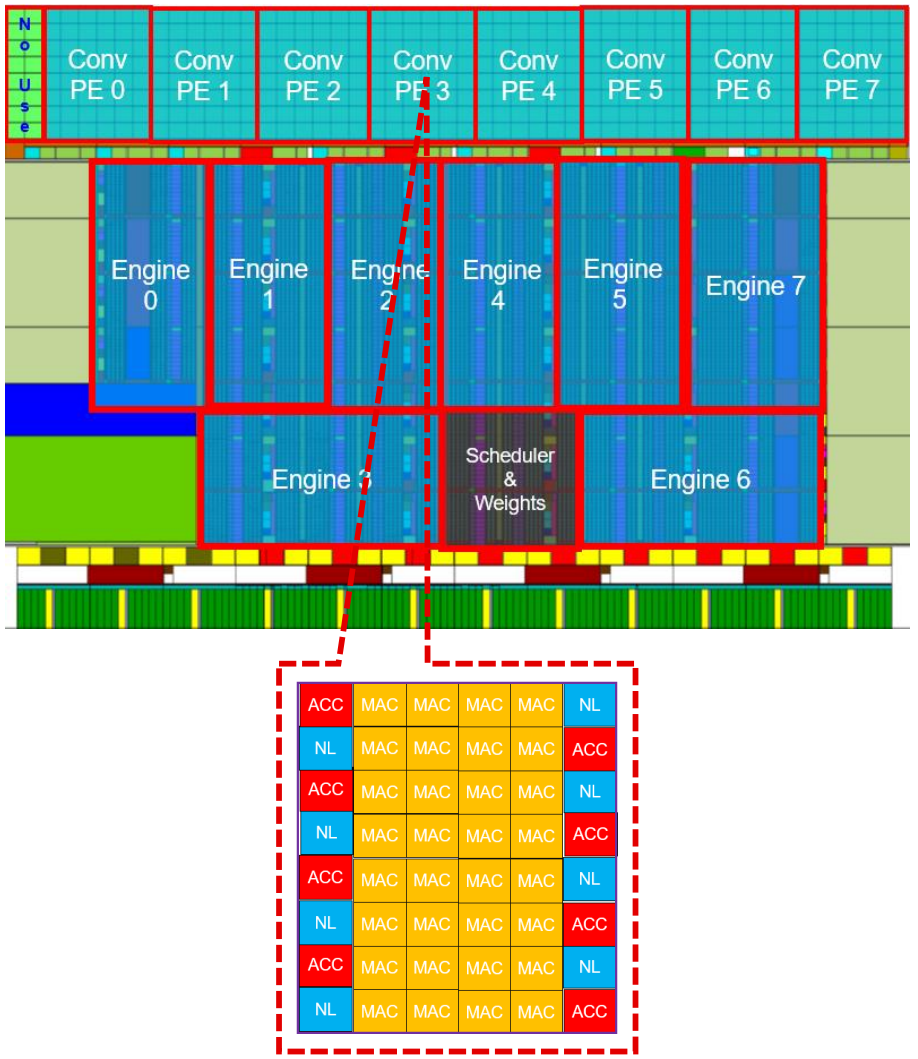


XILINX.

VCK5000上的CNN加速引擎

	8 PEs
AIE cores number(%)	48*8
PL LUTs(%)	588941 (65.46%)
PL FFs(%)	647521 (35.99%)
PL DSPs(%)	528 (17.62%)
PL BRAMs(%)	912 (94.31%)
PL URAMs(%)	424 (99.53%)

** All resource utilization is calculated with VCK5000 info*

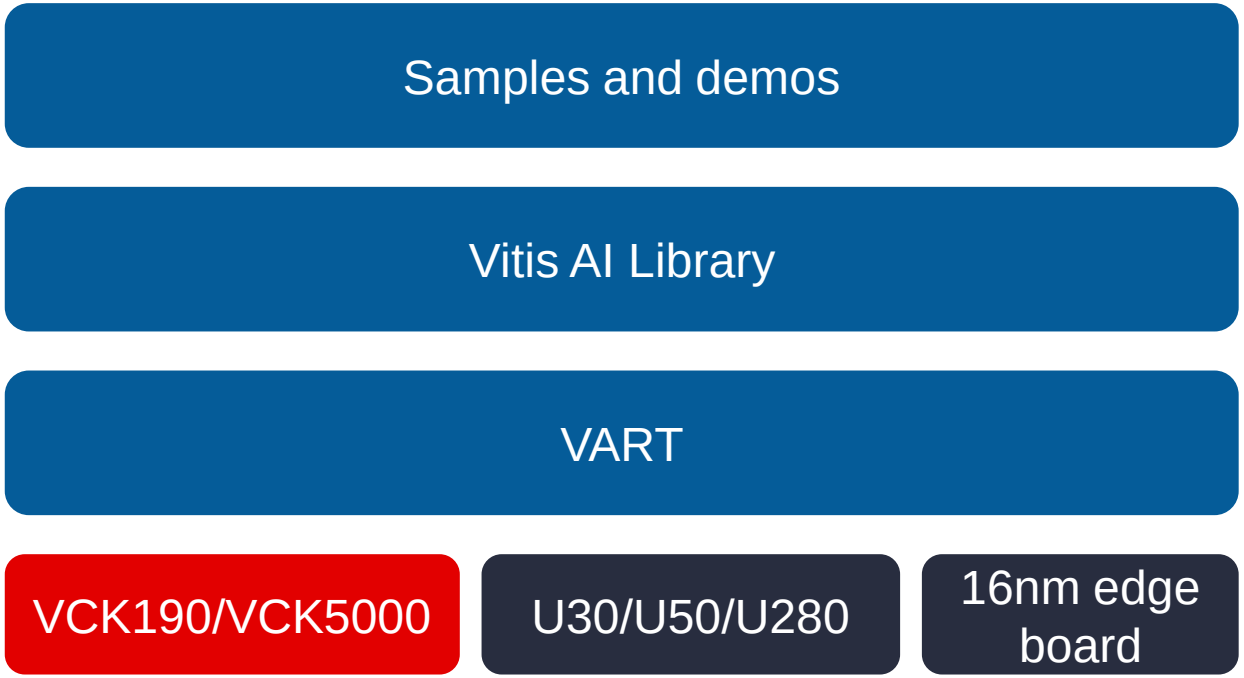


XILINX.

统一的软件开发套件

- Software stack
  - Vitis AI release
  - Unified software stack for cloud and edge
  - Unified samples and demos for 16nm and 7nm platforms
- Runtime (VART)
  - Support Versal device (VCK190 & VCK5000)
  - 70+ models support on Versal

Unified software platform for cloud and edge for 16nm solution and 7nm solution

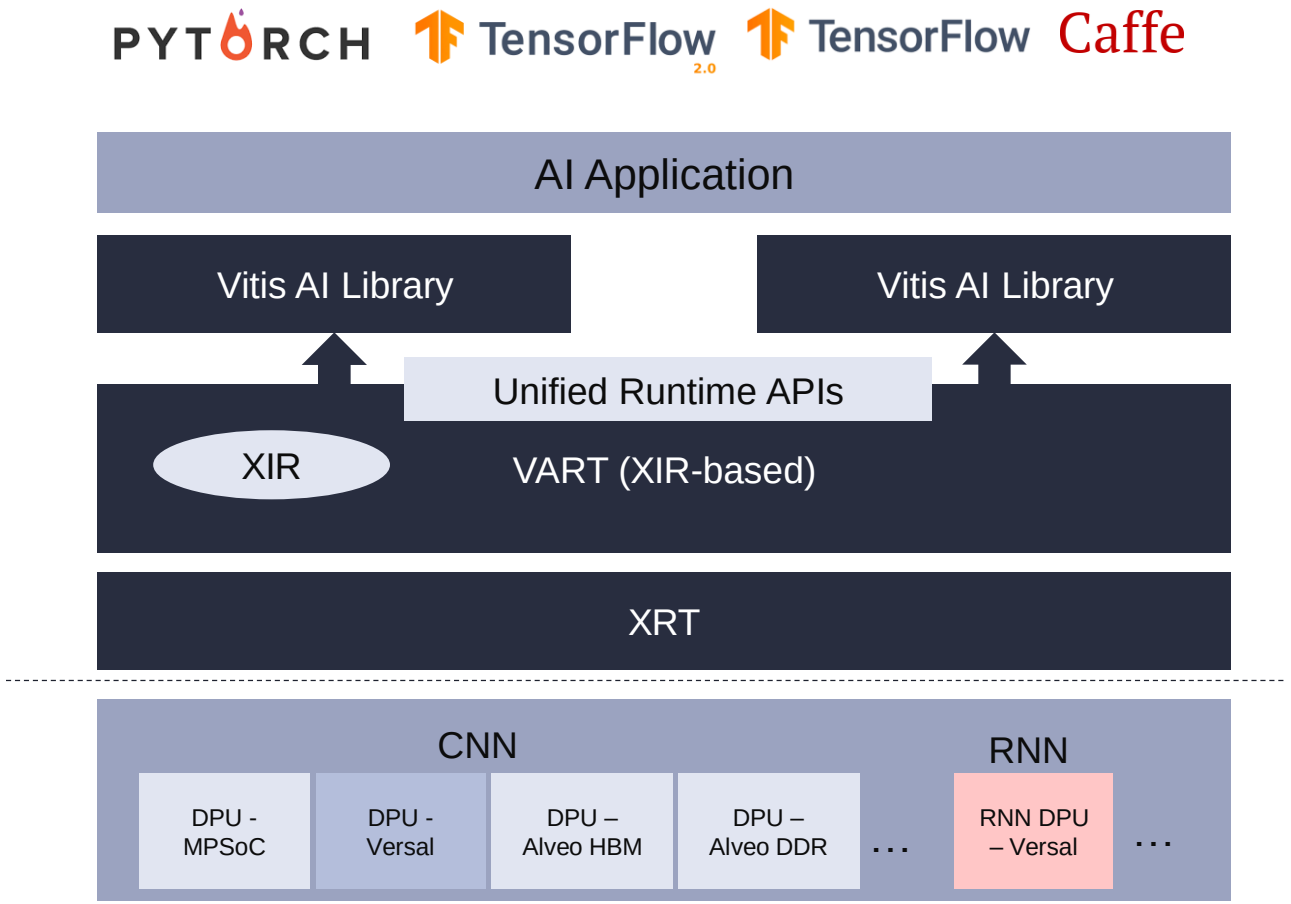


XILINX.

Vitis AI 运行时 (VART)

- **Open source in Vitis AI 1.3**
- **Added new Python APIs**
  - APIs for TensorBuffer Operation
  - APIs of RunnerExt
- **Software tools**
  - Supports Xmodel compiled with XIR flow
- **DPU/Platform**
  - Supports edge to cloud platforms
    - **Versal VCK190 board**
    - **Versal VCK5000 board**

VART Stack



Coming soon

XILINX

AI Library：高级AI应用库

- **High-level API-based libraries**
  - Supports new models in AI Model Zoo
  - Supports Xmodel compiled with XIR flow
  - For multiple tasks: image classification, detection, segmentation…
- **New DPU/Platforms**
  - Supports DPUCVDX8G (Versal DPU) on VCK190

User Applications

Demo and Reference applications

Framework

Vitis AI Library: Classification, Detection, Segmentation, …

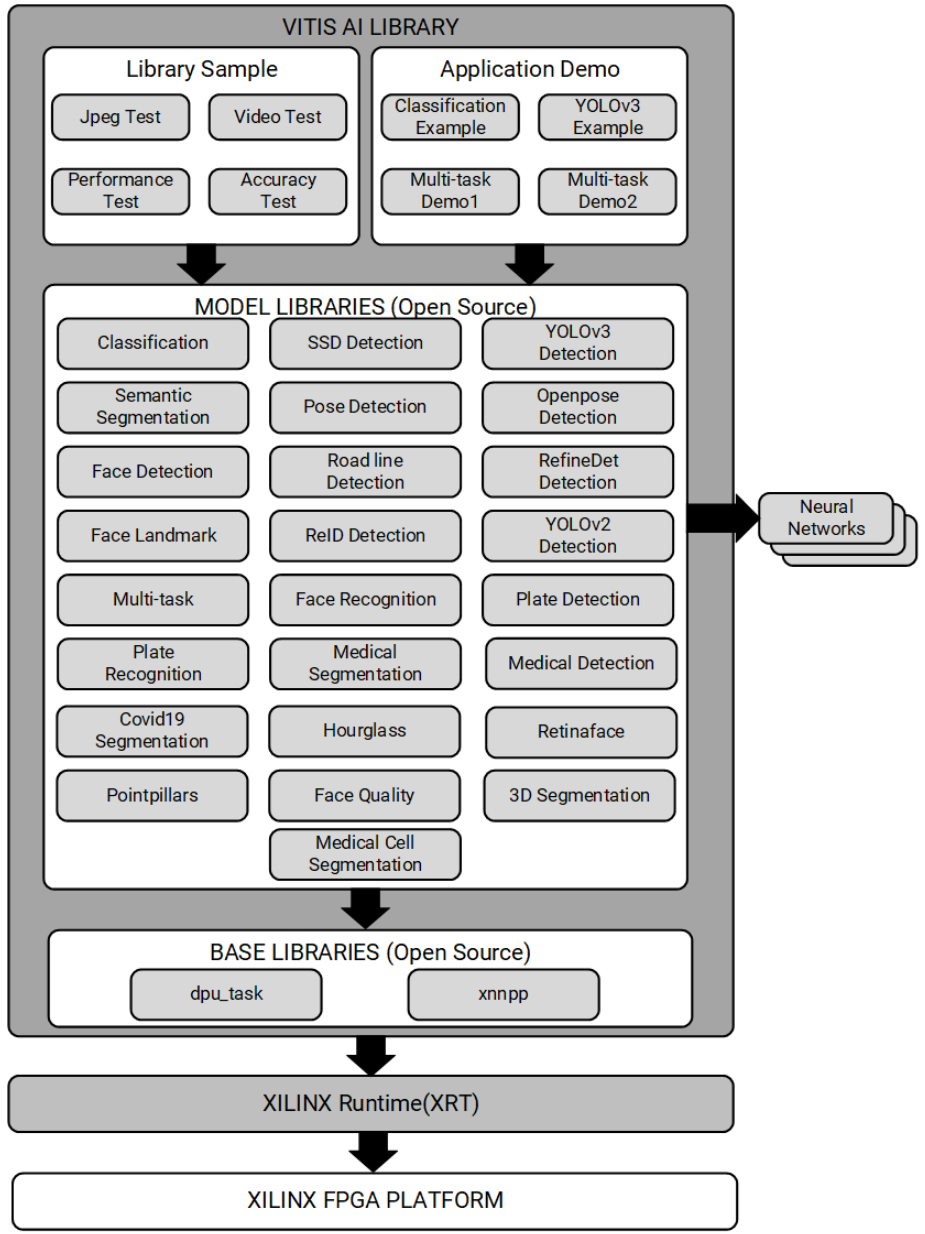
OS level packages

Ease-of-Use

Optimized

Open

AI Library Block Diagram



CNN架构的端到端性能

- New hardware aware pruning
  - Prune network according to hardware channel parallelism
  - MLPerf Resnet50_v1.5 remaining 74% pruning version meets Mlperf accuracy requirement
  - MLPerf Resnet50_v1.5 remaining 74% pruning version shows 23% hardware performance improvement
- MLPerf testing result
  - Kernel performance: 6433fps (measured by IP internal performance counter)
  - Sever mode performance: 5921fps (-5.6%)
  - Offline mode performance: 6257fps (-0.5%)
- Device power
  - 8PE PL300M/AIE1.25GHz device power is about 70W

CNN架构的网络支持规划

- All model zoo non-depthwise models have passed architecture simulation
- ~52 models can be supported in VAI1.3.1.

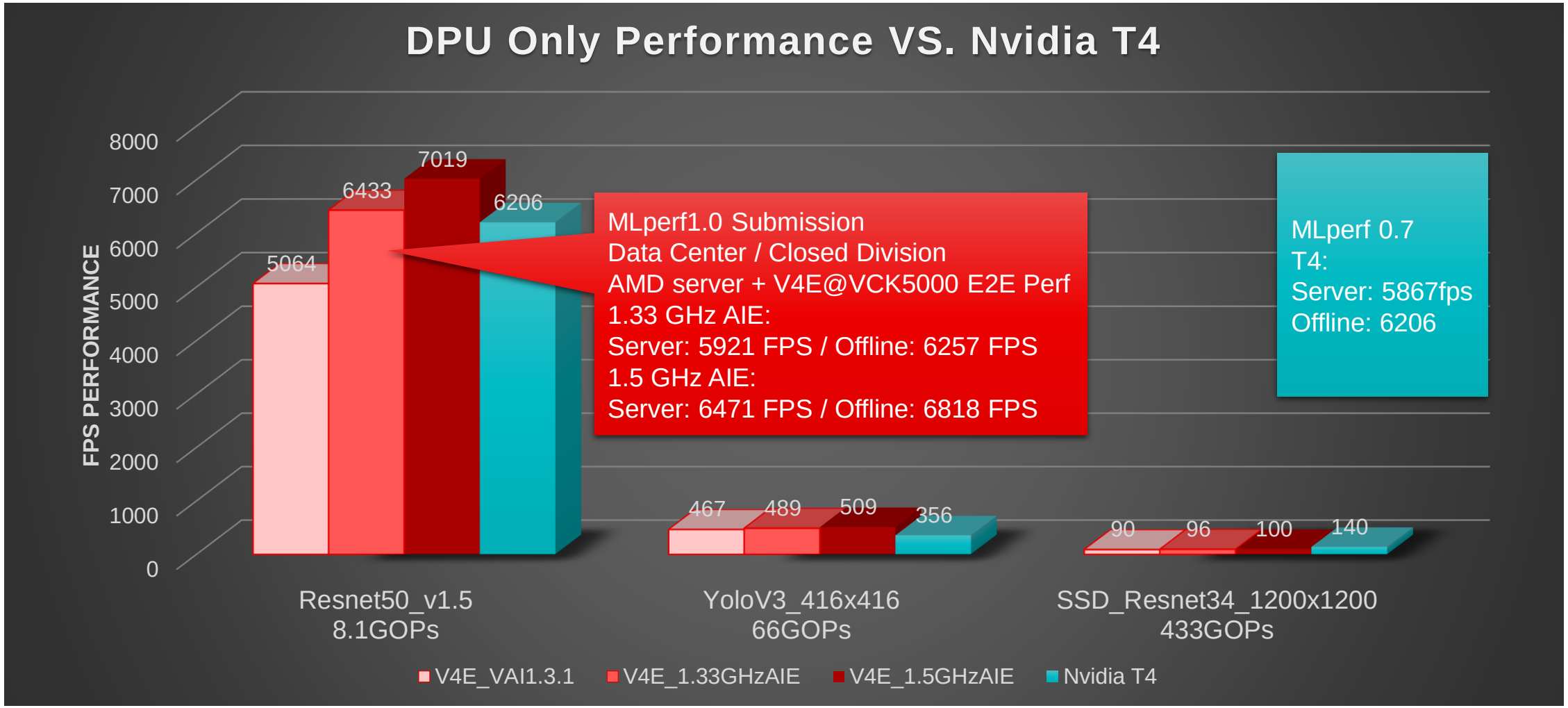
- 12 TF / TF2 CNN models
  - inception_resnet_v2_tf
  - inception_v1_tf
  - inception_v3_tf
  - inception_v4_2016_09_09_tf
  - resnet_v1_50_tf
  - resnet_v1_101_tf
  - resnet_v1_152_tf
  - ssd_resnet_50_fpn_coco_tf
  - yolov3_voc_tf
  - medical_seg_cell_tf2
  - resnet50_tf2
  - semantic_seg_citys_tf2
- 33 Caffe CNN models
  - resnet50
  - inception_v1/v2/v3/v4
  - squeezenet
  - resnet18
  - ssd_pedestrain_pruned_0_97
  - refinedet_pruned_0_8
  - ssd_adas_pruned_0_95
  - yolov3_voc / yolov2_voc
  - vpgnet_pruned_0_99
  - fpn
  - openpose_pruned_0_3
  - densebox_320_320 / 640_360
- 7 Pytorch CNN models
  - ENet_cityscapes_pt
  - face-quality_pt
  - FPN-resnet18_covid19-seg_pt
  - resnet50_pt
  - salsanext_pt
  - squeezenet_pt
  - unet_chaos-CT_pt

CNN架构下的性能分析

Neural Network	Input Size	GOPS	Performance (fps) (Single thread)	Performance (fps) (Multiple thread)
densebox_320_320	320x320	0.49	724.91	6120.58
yolov2_voc	448x448	34	248.07	939.84
ENet_cityscapes_pt	512x1024	8.6	26.55	137.75
face_landmark	96x72	0.14	3902.86	13348.30
tiny_yolov3_vmss	416x416	5.46	357.39	2566.91
face-quality_pt	80x60	0.06	4309.15	36995.40
fpn	256x512	8.9	116.16	945.26
FPN_Res18_Medical_segmentation	320x320	45.3	97.95	179.36
FPN-resnet18_covid19-seg_pt	352x352	22.7	310.18	704.55
FPN-resnet18_Endov	240x320	13.75	661.74	1545.64
inception_v1	224x224	3.2	1080.75	3572.83
inception_v1_tf	224x224	3	929.68	3767.30
medical_seg_cell_tf2	128x128	5.3	943.07	1928.51
MLPerf_resnet50_v1.5_tf	224x224	8.19	889.88	3754.93
mlperf_ssd_resnet34_tf	1200x1200	433	17.82	86.02
multi_task	288x512	14.8	105.21	666.11
openpose_pruned_0_3	368x368	49.9	29.03	166.09

https://www.xilinx.com/member/forms/download/eula-xef.html?filename=VCK5000_Network_Support_List-8PE.xlsx

MLPerf v1.0



- https://mlcommons.org/en/inference-datacenter-10/

XILINX.

BERT Project

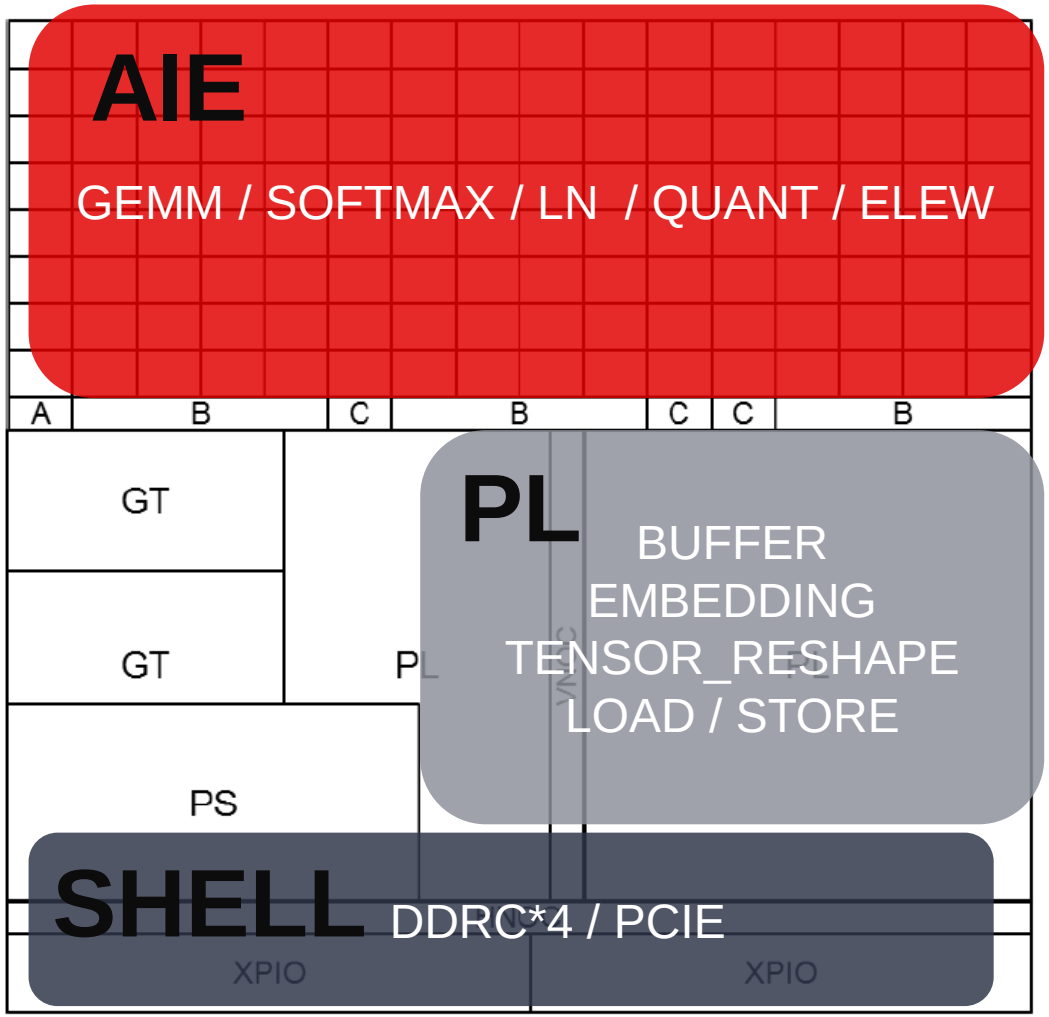
Backgroud:

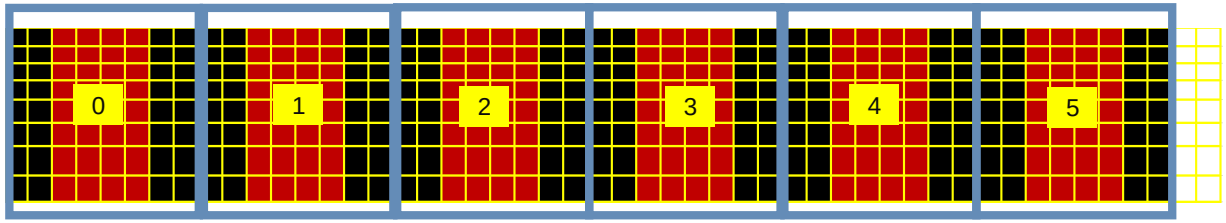
- **BERT**: Pre-trained deep bidirectional transformer. The hottest and state-of-the-art language model for NLP.
- **Transformer**: It is powerful not only in NLP, but also in image processing tasks such as classification, detection, and low-level tasks(denoise, super resolution, etc.), even in point-cloud.

Target:

- The first Xilinx BERT solution.
- Fully utilize the power of Xilinx Versal devices.
- Extend to broader transformer-based solutions.

BERT架构概述

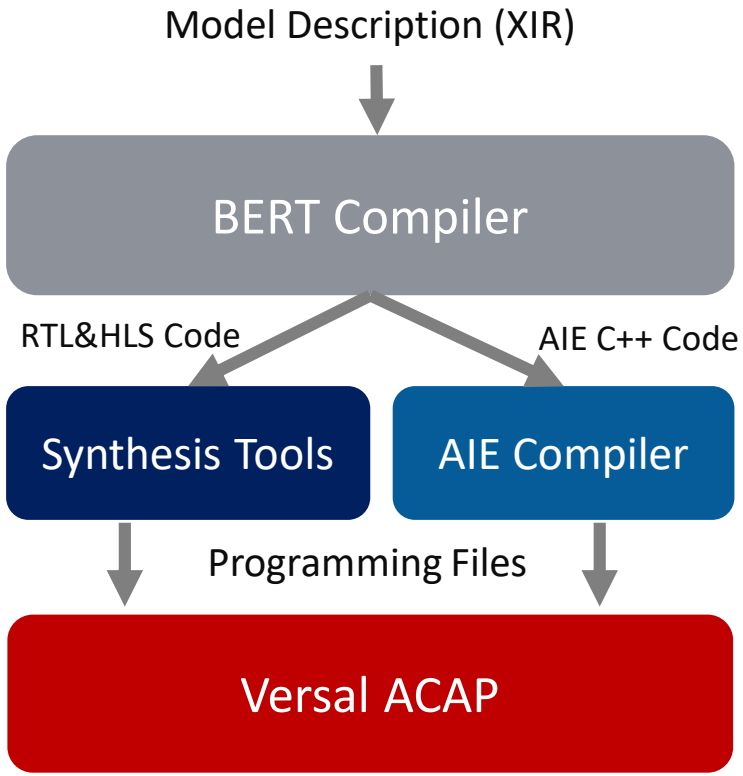




- **AIE:** The main computing engine. 384 of 400 AIE cores are used.
- **PL:** Responsible for tensor reshape and data path.
- **SHELL:** DDR & HOST interfaces
- **Supported Operators**:
  - *Embedding*
  - *GEMM*
  - *Softmax*
  - *Layer Normalization*
  - *Gelu*
  - *Positional Encoding*
  - *Elementwise Add*
  - *Concat*
  - *Taking Activation as Weight*

XILINX.

Templates-Driven 编译器



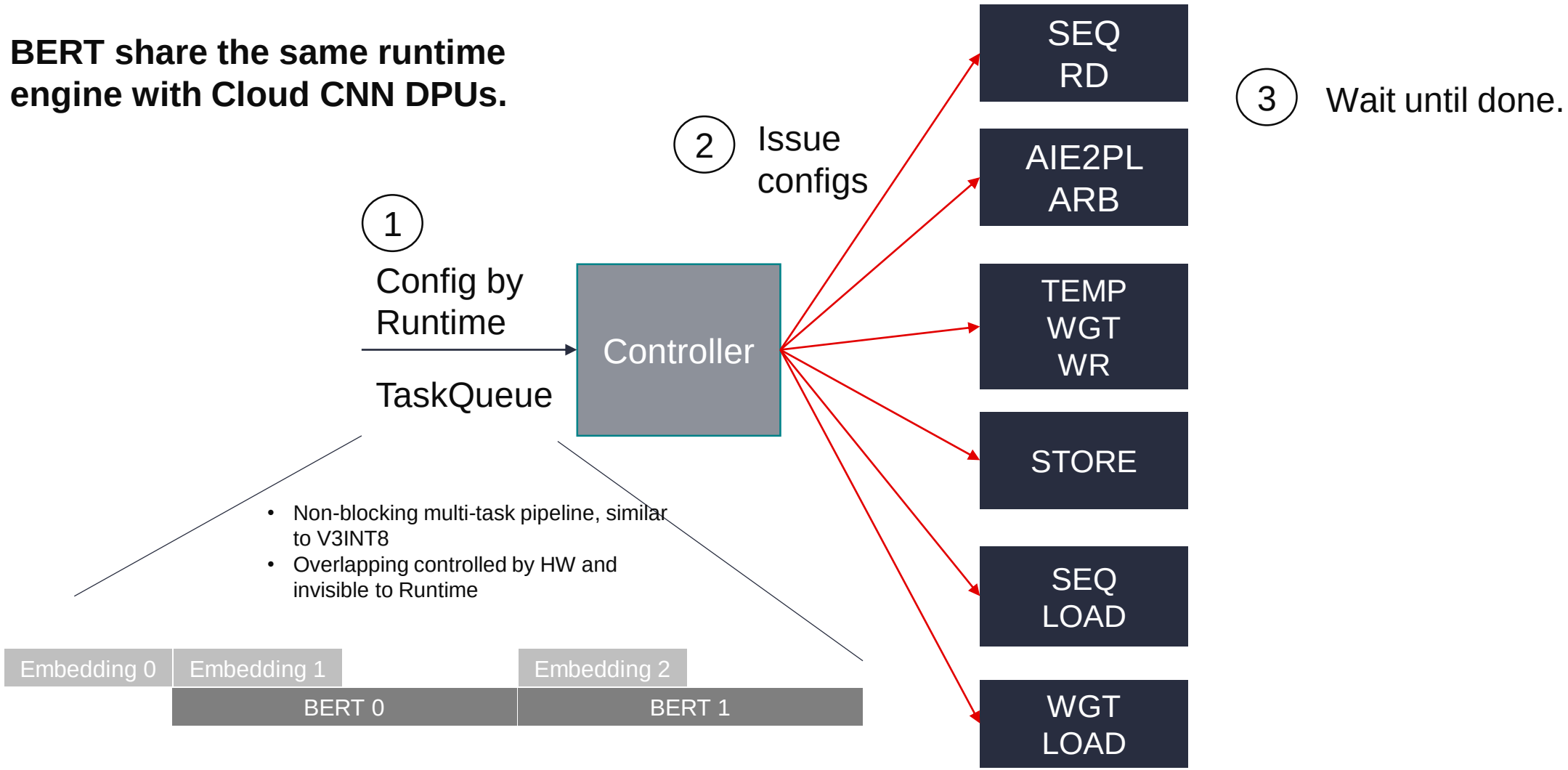
- Redefine the role/flow of **BERT Compiler**
  - Parse the model and check whether the BERT model is in the support range;
  - Rearrange the weights in the model based on the requirement of IP and forward them to Runtime usage;
  - Generate the configuration parameters for the PL and AIE code templates and output real PL and AIE code：

    `BertModel` ➔ `(BertPLImpl Code + BertAIEImpl Code)`

  - (Optional)Drive the RTL and AIE compiler to compile the generated PL & AIE code;
  - (Optional)Drive other necessary flows to generate the xclbin;

BERT 运行时

BERT share the same runtime engine with Cloud CNN DPUs.



XILINX.

BERT 量化

Mixed precision (Int8 & FP32) quantization works for BERT.

https://confluence.xilinx.com/pages/viewpage.action?pageId=131581806

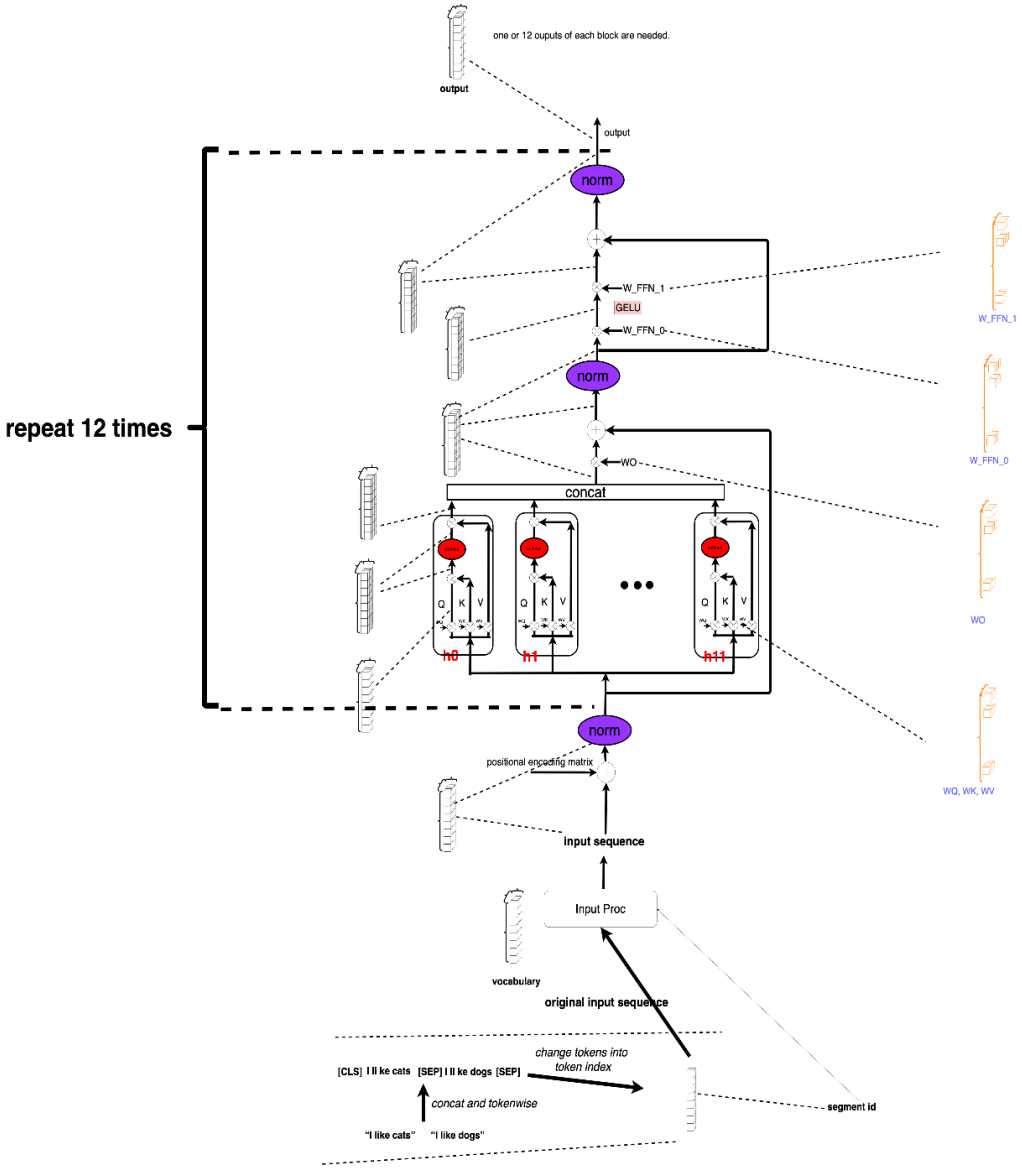
	Model	Dataset	Max seq length	Precision	Quantization Position	F1 on SQuAD1.1
MLPerf from zenodo	BERT Large	SQuAD1.1	384	fp32, int8	Except LN	fp32: 90.874% int8: **90.482%--> 90.12%**
Habana	BERT Base	SQuAD1.1	128	int16	GEMM in int16, some operators like LN in FP32	**fp32: 86.414%** int16: ~86.319% (At most 0.11% loss vs. fp32)
						128: **fp32: 86.97%** int16: 86.92% int8: 85.10% **int8 2^ * f32-LN: 85.76%** **int8 2^ * f32-LN (progressively):->86.28%**
Ours	BERT Base	SQuAD1.1	128/384	Up to int8	Weights and activations, including LN	384: fp32: 88.45% int16: 88.44% int8: 86.90% int8 2^ 16LN: 87.48%
Ours	BERT Large	SQuAD1.1	384	Up to int8	Weights and activations, including LN	90.16% (int8 2^ with rules) 90.19% (int8 2^ w/o LN) **90.62% (same quant as MLPerf)**

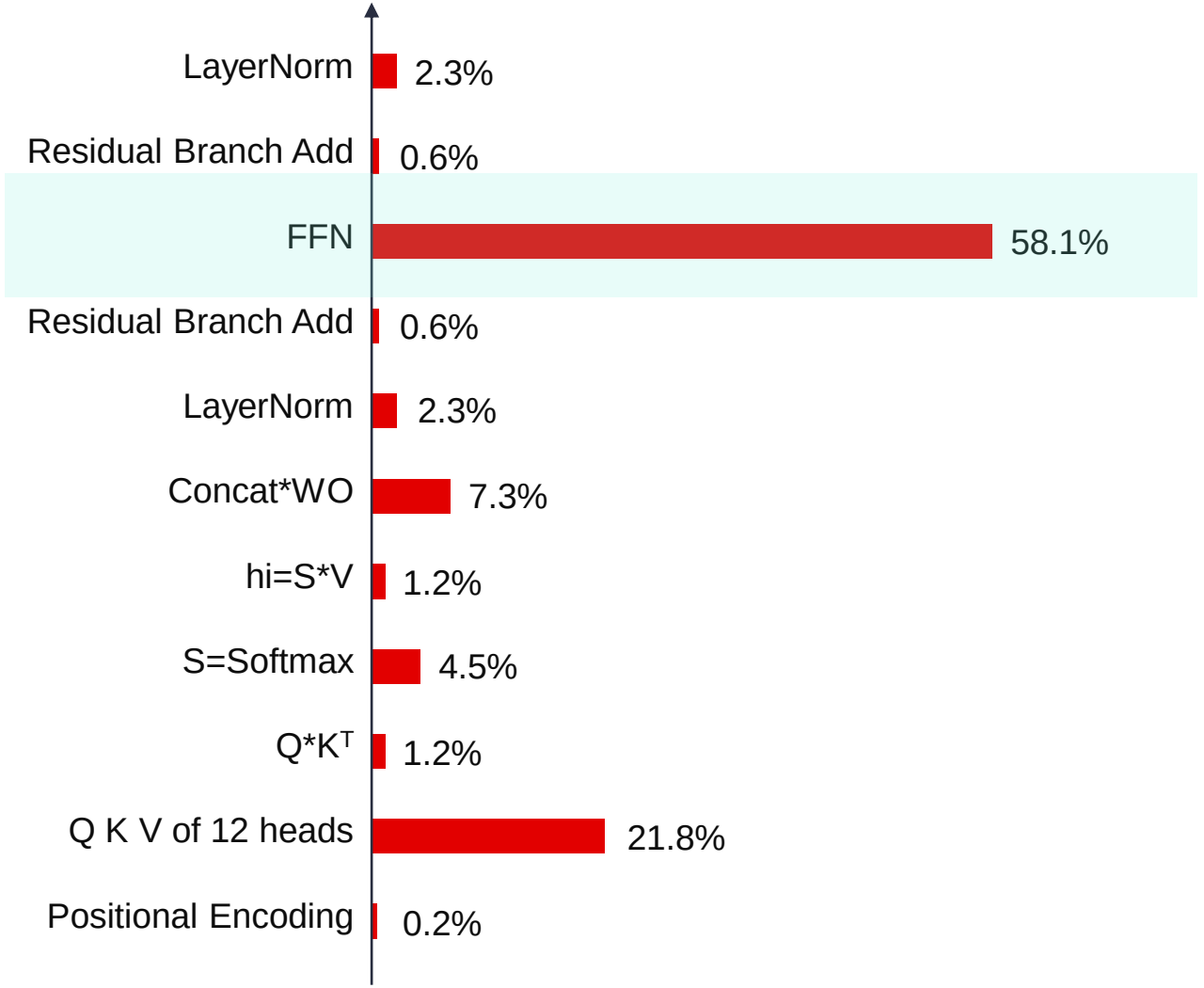
Latest(Dec.14): 86.5%

	Model	Method	Compression phase	Extra fine-tuning workload	Cost saving Params/FLOPS	vs. BERT base (Acc)
Benchmark	BERT Base		-	-	109M/2.7G	100%
Compressed BERT	DistilBERT	Distillation	Pre-train	None	6-layer: 2x/2x	97%
	TinyBERT	Distillation	Pre-train&Fine-tune	Data augmentation Fine-tuning distillation	4-layer: 7.5x/9.4x 6-layer: 2x/2x	96% 100%
	MiniLM	Distillation	Pre-train	None	12-layer: 3.3x/4x	100%
	BERT-OF-THESEUS	Pruning	Fine-tune	Module replacing	6-layer: 2x/2x	98%
	LayerDrop	Pruning	Pre-train	Selecting layers to keep/prune	6-layer: 1.6x/2x	100%
Architecture Modified BERT	ALBERT	-	Pre-train	None	12-layer: 12M/2.7G 24-layer: 18M/9.7G	97% 104%
	MobileBERT	-	Pre-train	None	24-layer: 4x/5x	99%

XILINX.

BERT 的逐层性能分析



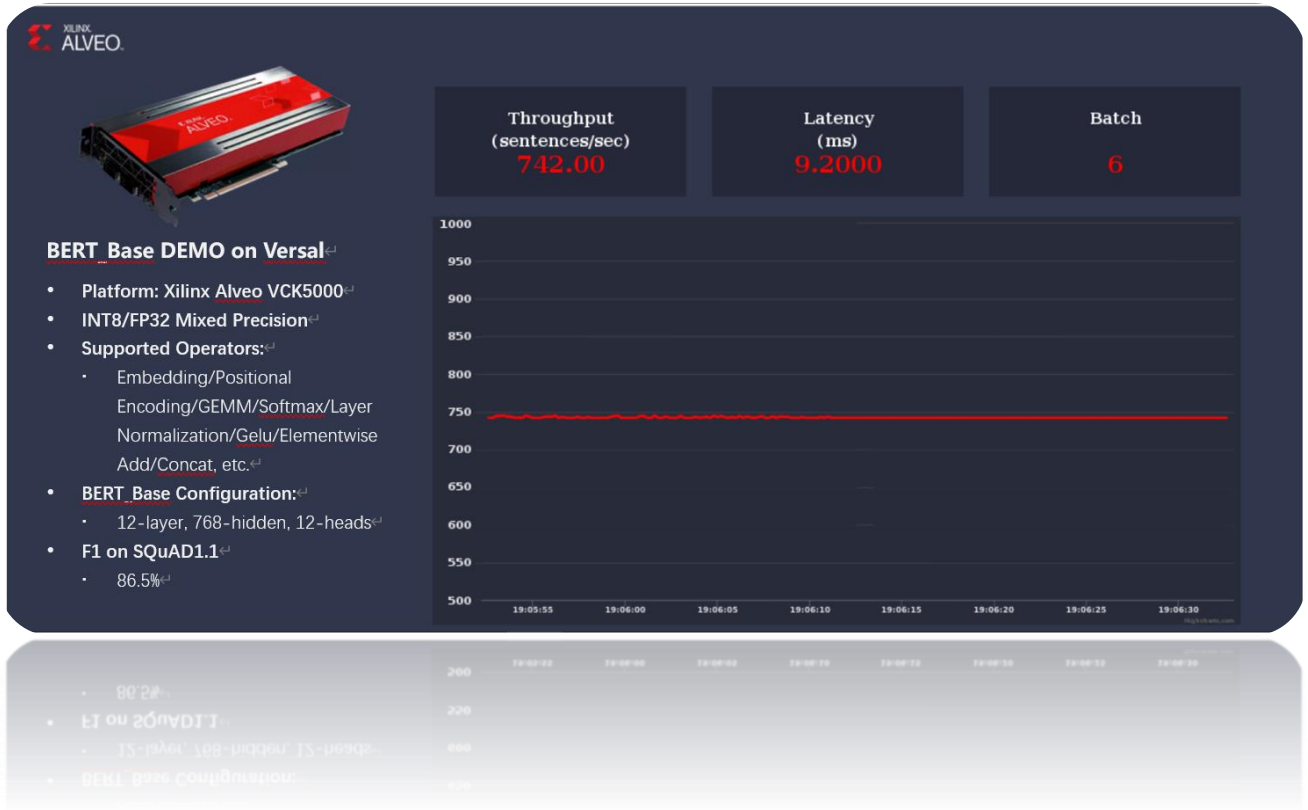


XILINX.

BERT DEMO on Versal

(Released in vai-1.3.1)

- **Platform: Xilinx Alveo VCK5000**
- **INT8/FP32 Mixed Precision**
- **BERT_Base Configuration:**
  - 12-layer, 768-hidden, 12-heads
- **F1 on SQuAD1.1**
  - 86.5%
- **Throughput**
  - 742 sentences/sec *(still optimizing)*
- **Latency**
  - 9.2 ms *(still optimizing)*



XILINX

总结

- Please visit our VCK5000 Lounge
  - https://www.xilinx.com/products/boards-and-kits/vck5000.html
- Solutions for CNN
  - DPUv4e on VCK5000 targeting on cloud application
- Demo
  - BERT demo on VCK5000
  - MLPerf v1.0 benchmark

Thank You