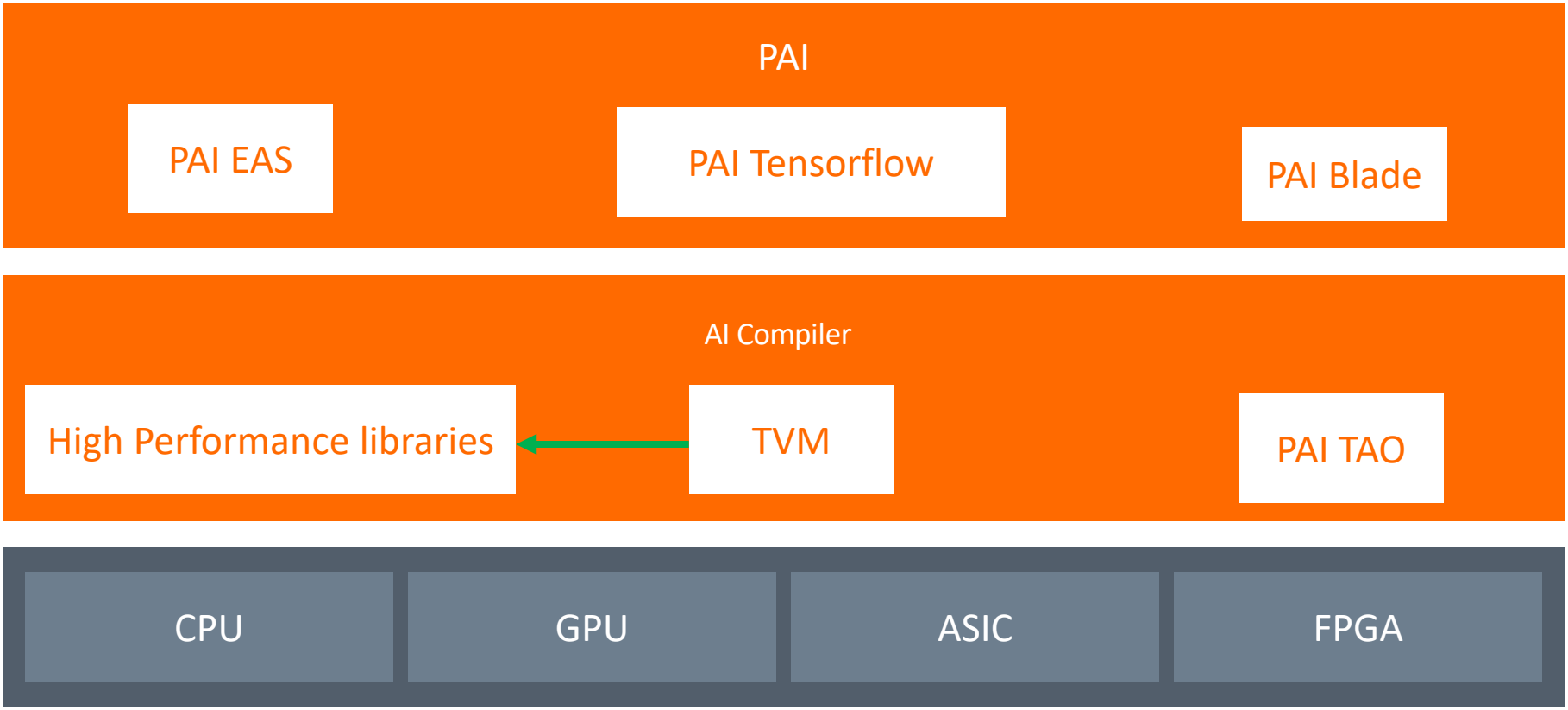
AI Compiler @ Alibaba

Xiaoyong Liu

PAI (Platform of AI)
Alibaba Cloud Intelligence

Presenting the work of many people !

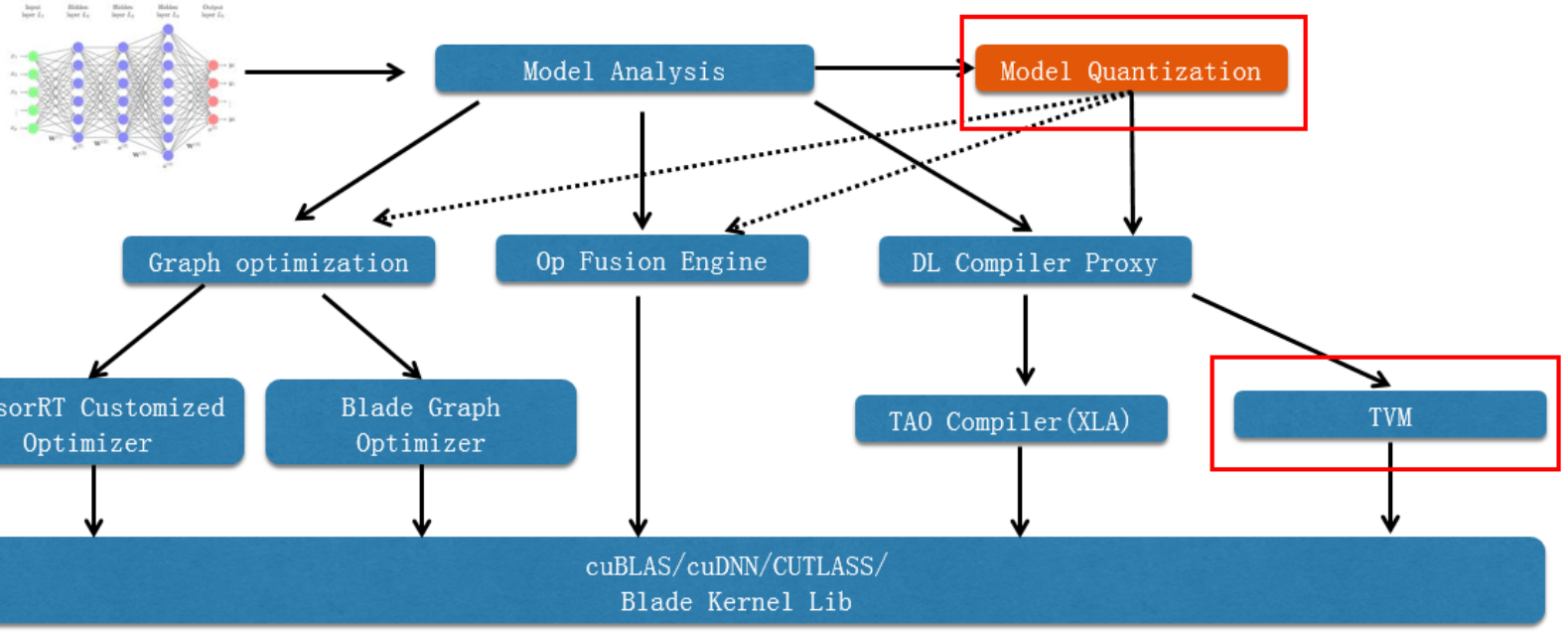
AI Compiler Stack



How TVM is used @ Alibaba

- **An End-to-End Deep Learning Compiler**
  - **Empower AI service**
  - **Generate high performance operators**
    - subgraph & kernel
    - heterogenous computing
- **An Optimizer & Compiler**
  - **Enable chips such as CPU, GPU, DSP & etc. ， potentially FPGA, AI Chips.**
  - **Deploy algorithm automatically**
- **All Scenarios**
  - **Cloud, Edge & IoT**
  - **Training & Inference**

TVM + AIService：PAI-Blade

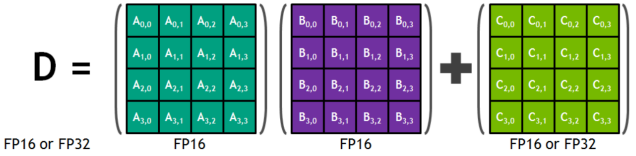


Things We Experienced

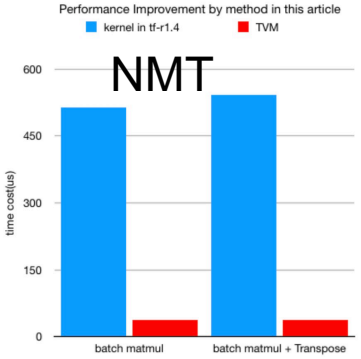
- **Current approach is too much engineering effort , difficult for platform service**
- **TVM is good at**
  - **To generate high-performance computing intensive kernels**
    - Automatic is the key
  - **Heterogenous hardware friendly, if ISA is provided**
    - Performance portability
  - **Software Architect friendly to Auto TVM / Schedule…**
  - **Whole-Graph Optimization**
- **Challenges**
  - **Easy of deployment, including coverage, quality & compatibility**
    - Correctness, Performance & easy of new device enabling
  - **Systems don’t interop**
  - **Maturity/Standardization …**

Contributed to TVM Community

- **Automatic Tensor Core Scheduling**
  - Nvidia tensor core in V100/T4



- **Schedule Algorithm enablement, as Batch Matmul and etc.**
  - Know what/how



- **Support TFLite models**
  - Automatically

- **C++ RPC Server**
  - Tuning your program in embedded environment without python

Ongoing Effort to Community

- **Automatic Tensor Core Scheduling Enhancement**
  - vthread supporting
- **Operators: New Ops in Relay / TVM**
  - HashTable， Embedding…
- **Standardize GraphRuntime Exports Into a Single DLL**
  - A way to unify runtime models exports

Product-driven TVM enhancement

- **Brings online inference service**
- **Enhances Infrastructure**
- **Compiles heterogenous hardware at cloud & edge**
  - **Nvidia server GPU**
    - V100/T4 on FP16/INT8/INT4/INT1
  - **Intel X86 Server CPU**
    - on INT8/FP32/BF16
  - **ARM64 CPU**
    - on INT8 / FP32
  - **ARM32 CPU**
  - **HIFI4 DSP**
  - **Hexagon DSP**
  - **PowerVR GPU**
  - **Intel GPU**

Any general solution is planed to contribute back to TVM Community!

TVM produced more performance

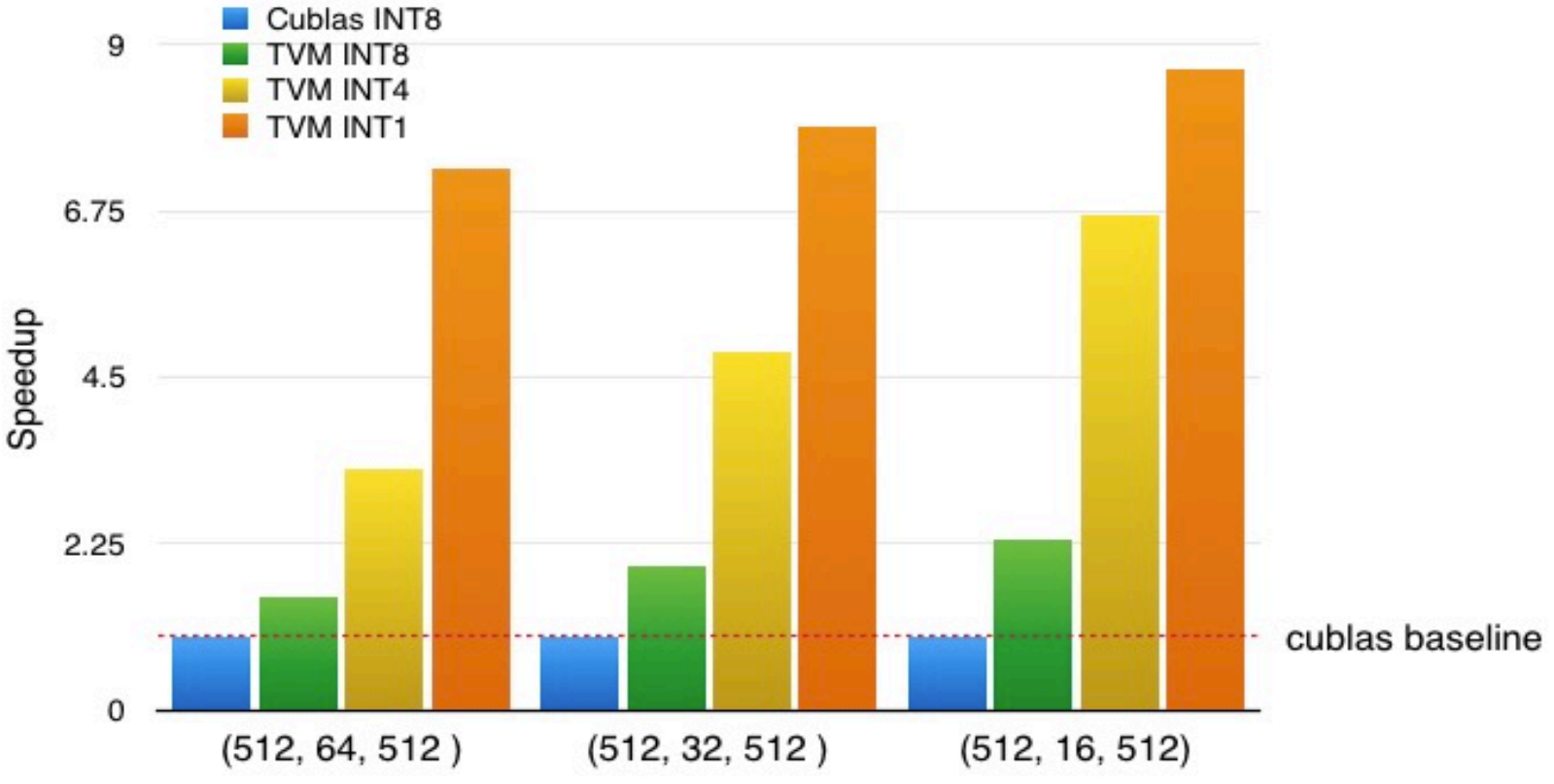
- **VS**
  - **Chip supplier's latest manually optimized high performance library**
  - **Assembly-level optimized edge machine learning framework**
- **Optimized to gain decent performance on various products**
  - **Server Nvidia GPU**

    Automatic TC Scheduling + tensorization + tensorcore
  - **Edge Arm64**
  - **IoT Arm32**

Performance on V100 (FP16)

M, N, K	cuBLAS TensorCore	TVM TensorCore	speedup
512, 16, 512	7.7470us	5.2570us	1.47X
512, 32, 512	8.0140us	6.0220us	1.33X
512, 64, 512	8.7530us	6.2390us	1.40X
512, 128, 512	9.0290us	7.1610us	1.26X
256, 256, 256	6.9380us	4.5930us	1.51X
1024, 32, 512	8.3320us	6.3770us	1.30X
2048, 32, 512	9.0640us	7.5070us	1.21X

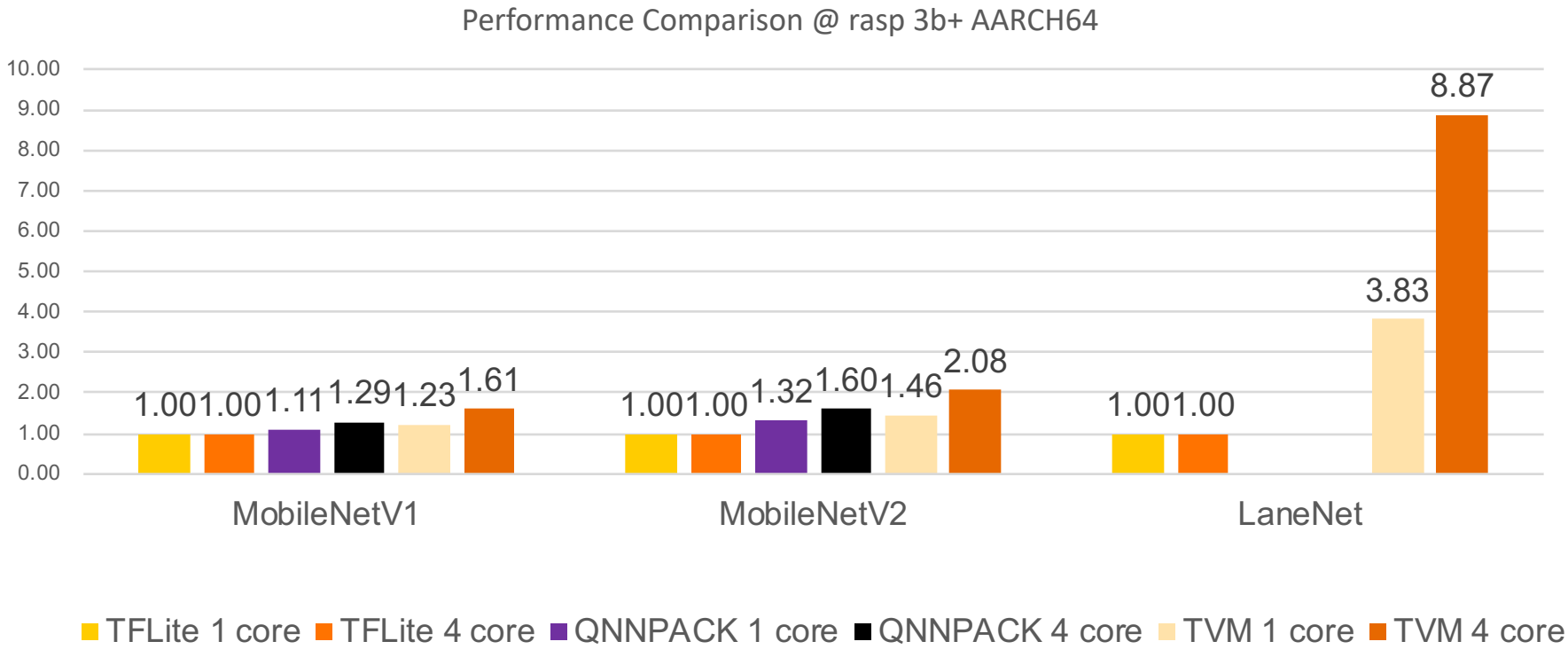
Performance on T4



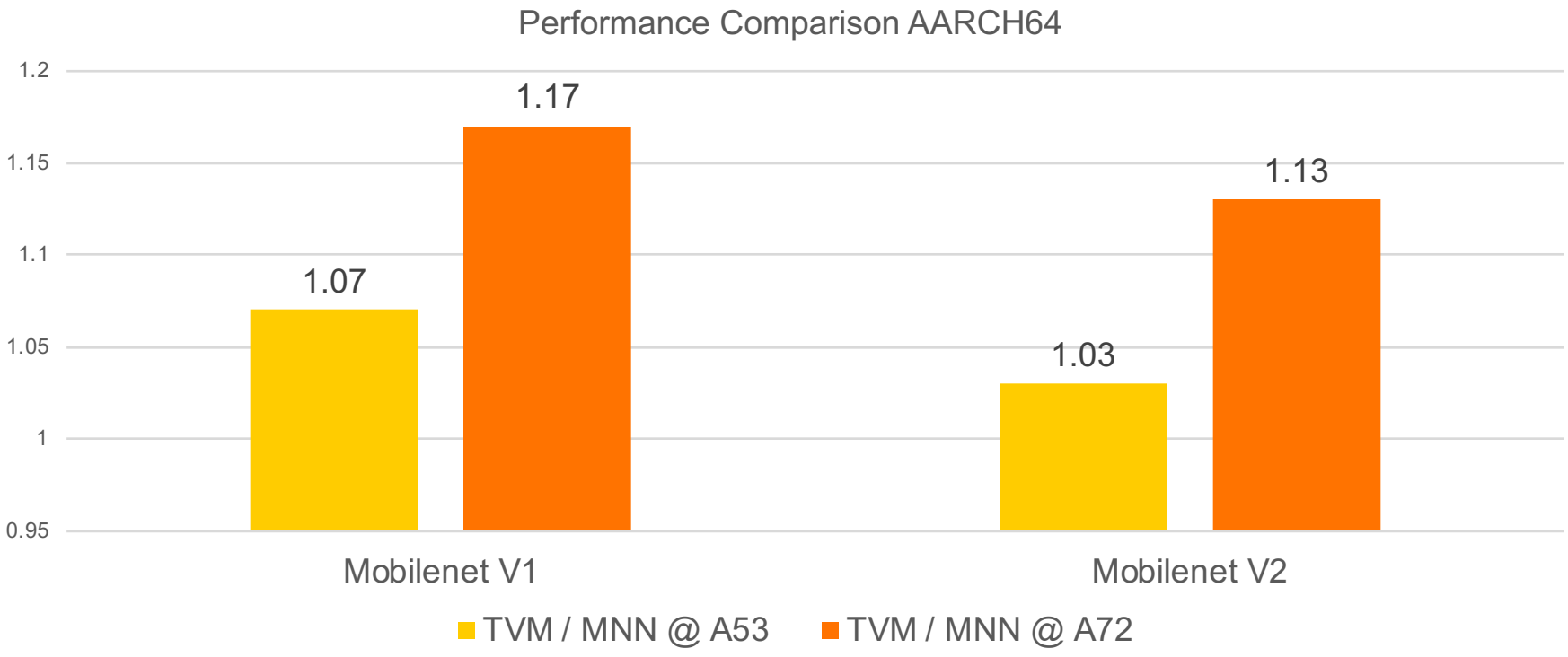
AliOS enhances TVM on vehicles

- **To accelerate NLU and AR-Nav models**
- **ARM64 CPU performance on INT8 / FP32**
  - NHWC/img2col+pack/no tensorize&co-optimized with llvm
  - Planning to contribute back to community
- **Hexagon DSP**
  - vrmpy tensorize/llvm-codegen
  - Could run end-to-end Mobilenet V2 INT8 model
- **Intel GPU**
  - Schedule algorithm
  - Boost 1.6X performance of Lanenet model

Performance on ARM64 INT8



Performance on ARM64 FP32

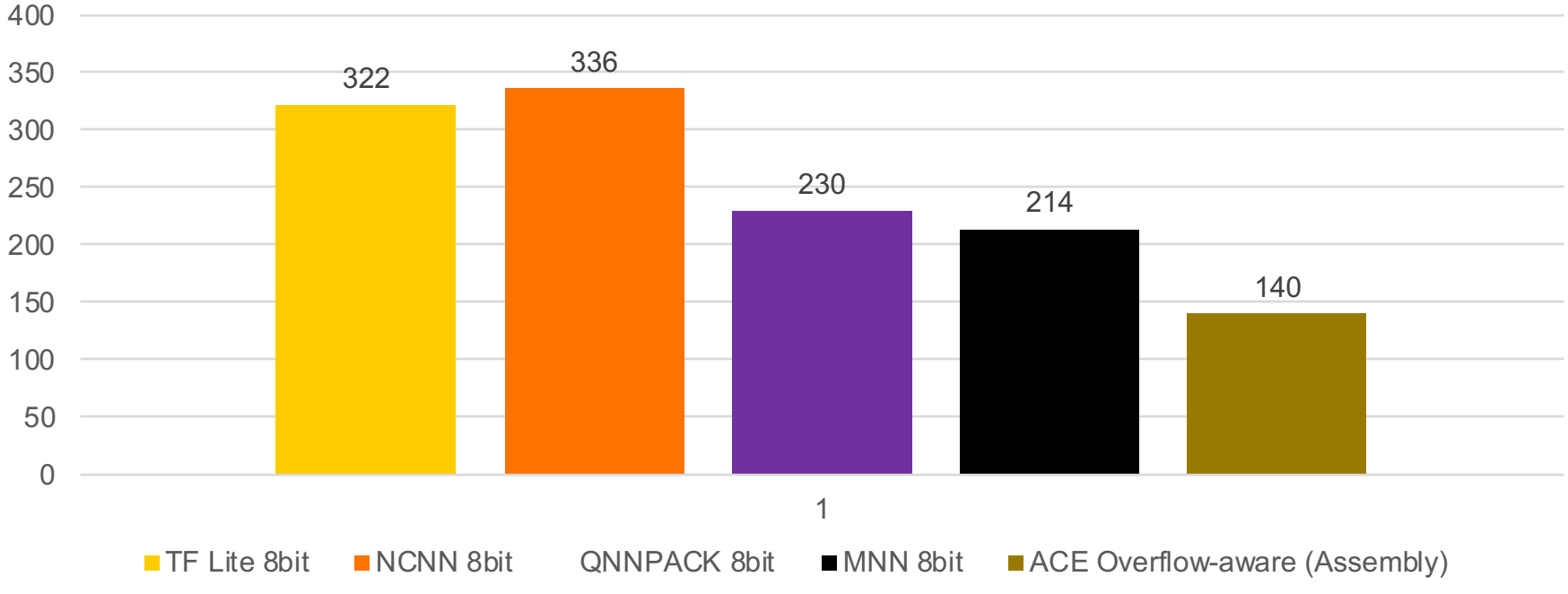


AI Labs Compiles TMallGenie Models

- **ARM32 CPU**
  - Overflow-Aware Quantization (INT16 = INT8 * INT8)
  - GEMM Tensorize
- **HIFI4 DSP**
  - GEMM Tensorize, 10X speed up
- **PowerVR GPU**
  - Schedule Algorithm

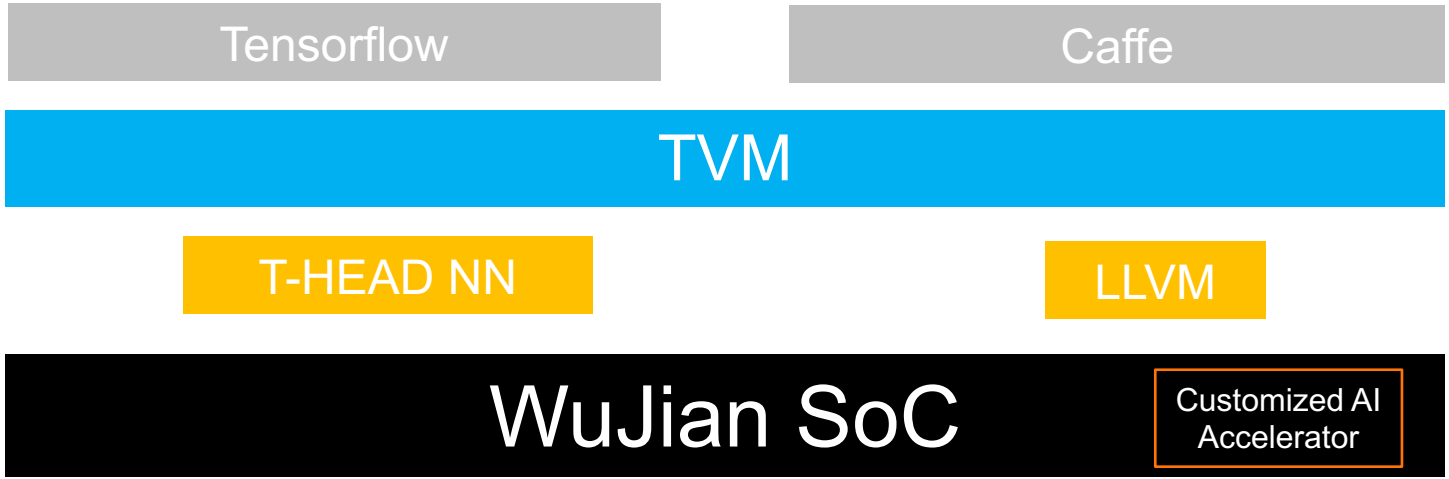
Performance

CPU：MTK8167S（ARM32 A35 1.5GHz）
Model：MobileNetV2_1.0_224



A DL Compiler in T-HEAD SoC

- **TVM has been integrated into WuJian(无剑) SoC toolchain**
- **Support Caffe Frontend**
  - Tested pass alexnet / resnet 50 / mobilenet v1 / mobilenet v2 / …



TVM Roadmap @ Alibaba

- **Keep contributing general effort back to community**
- **Auto Schedule (“with Berkeley team”)**
  - Auto* is the key to build machine-learning-powered system
- **Interpolate with top frameworks**
- **Auto heterogenous hardware placement in system level**
- **Infra Maturity**
  - Completeness & Seamless Deployment, as quantization, model compatibility
- **Workload Characterization**
  - To improve the key workloads within community
- **AI Service & Operators**
  - more chips, more models

Alibaba Group 阿里巴巴集团

Alibaba & OpenSource

Embrace OpenSource

Contribute OpenSource

Win-Win OpenSource

Takeaways

- **A golden age of deep learning compiler**
- **Industry-grade deep-learning compilation solution is still in evolution**
- **We are working to contribute to TVM**
  - Development & Research
- **Welcome to join us to contribute to TVM together**
  - Xiaoyong Liu (xiaoyong.liu@Alibaba-inc.com)

Alibaba Group 阿里巴巴集团

Thank you!

Q&A

Alibaba Group
阿里巴巴集团