NeuralScale: A RISC-V Based Neural Processor Boosting AI Inference in Clouds

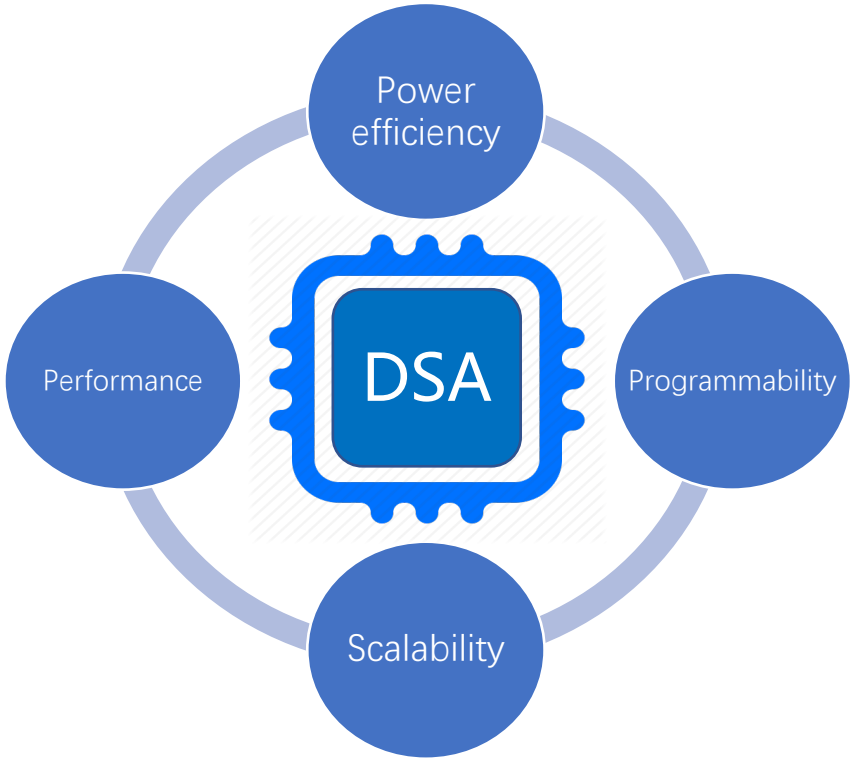
希姆计算, Stream Computing Inc.
詹荣开, Mark Zhan
2020-6-7

The Future of DSA AI Accelerator: General Purpose NPU

A New Golden Age for Computer Architecture: Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development

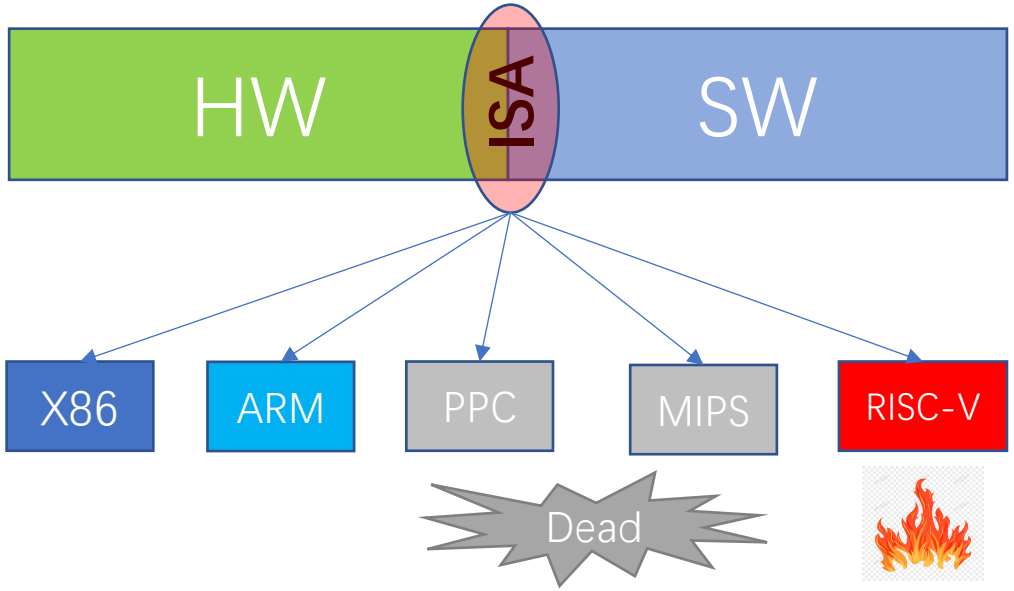
—— By John Hennessy & David Patterson

- The Computation of AI Workload: Massive Parallelism and Huge Memory Access Demands
- Deep Learning Theory Evolvement: New Operators or Activation Functions
- Operators Customization Needs in real AI Applications.
- Not all operators are computing intensive



Programmable DSA AI Processor

RISC-V: DSA-Native ISA for General Purpose NPU

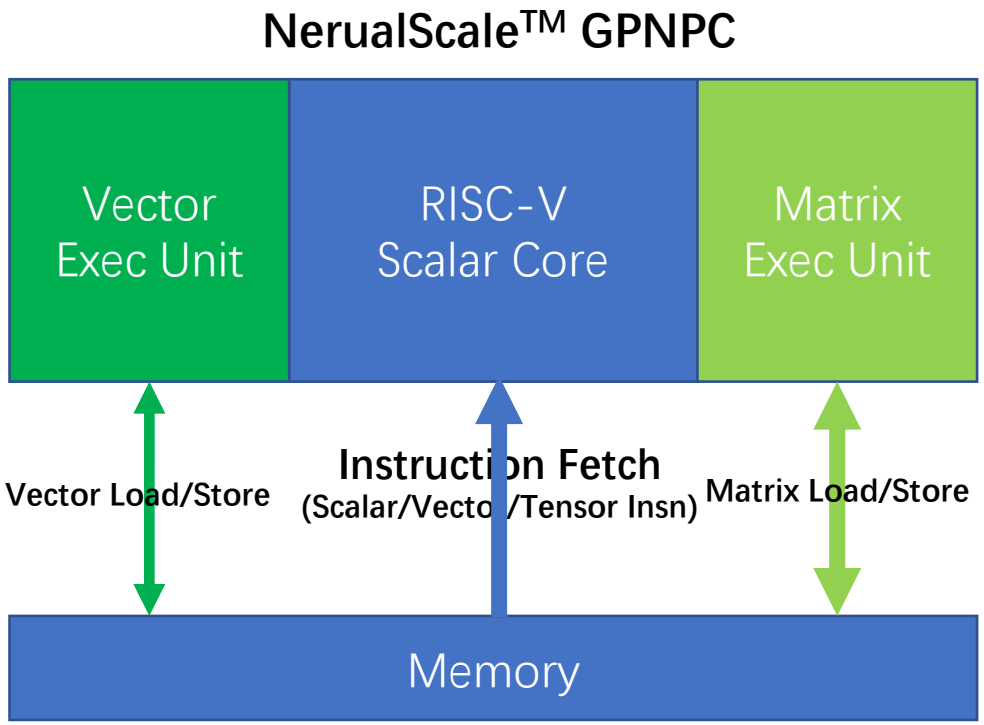


	RISC-V	intel / arm
Elegant ISA	Latest Modern ISA Design in industry: Clean、Efficient and Elegant Instructions	Due to backward compatibility, very complicated instructions
Modularity	Modularized Design: No need to implement all instructions	Not Supported
Extensibility	Silicon Vendor Can define his own private insn	Not Supported

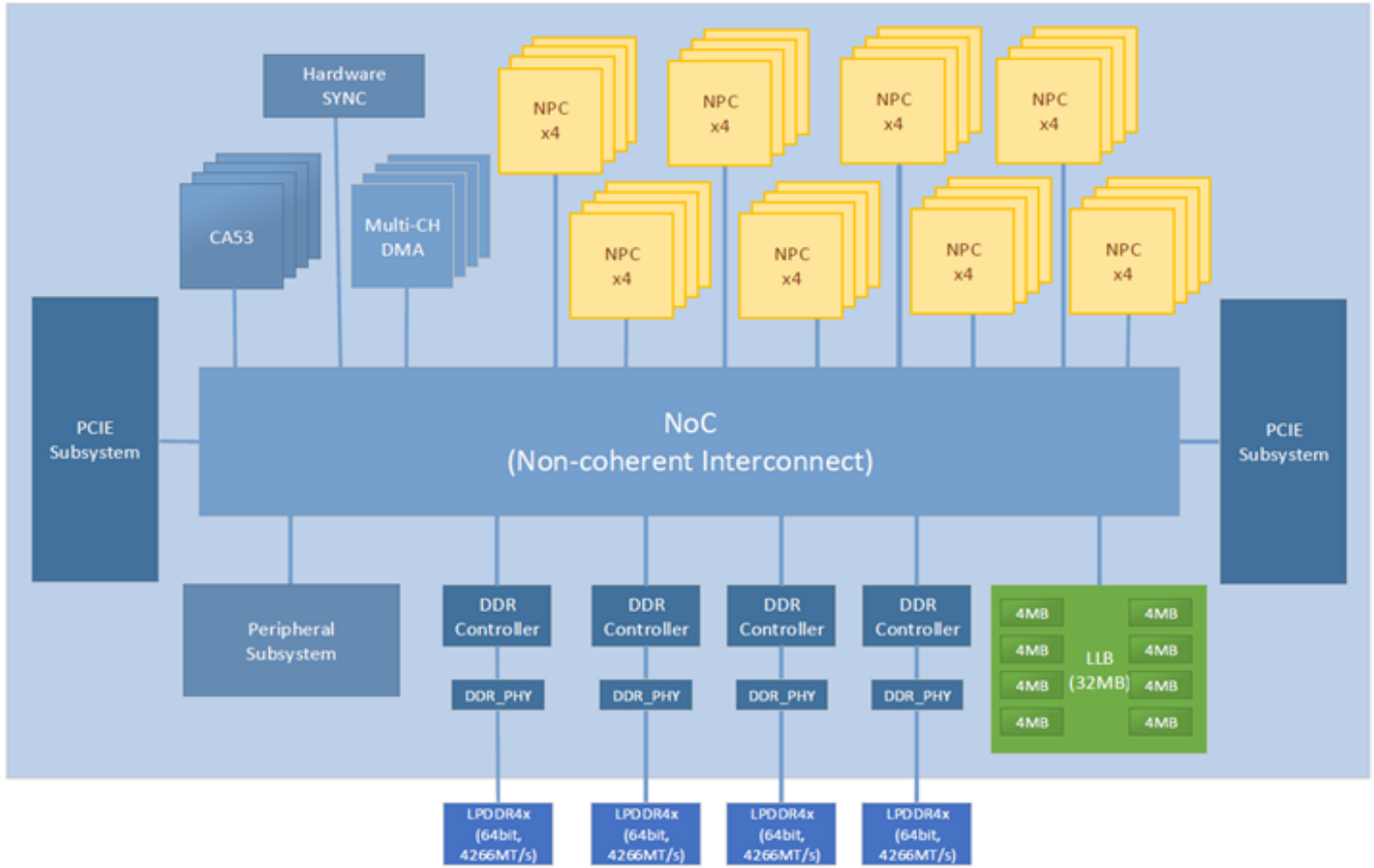
- **As a DSA-Native ISA Design, RISC-V is the Best Choice for General Purpose NPU, to achieve a balance both Programmability and Performance:**
  - **Turing-Complete base ISA: Enable C-Language Programming for NPU**
  - **V-Extension Vector Instructions: Base of AI Operators Implementation**
  - **Self-defined Instructions Extension: Matrix/Conv acceleration.**

NeuralScale: Stream Computing General Purpose NPC (GPNPC) Architecture

- Scalar Processor Core:
  - RV32 GC(IMAFDC) Instruction Set
  - IEEE-754 Compatible Single-Precision FPU
- Vector Exec Unit
  - RISC-V V-Spec Vector ISA w/ FP16 & INT8 data types
  - Extended V-Spec Instructions for sqrt/exp/log、post-increment etc.
- Matrix Exec Unit
  - Extended GEMM Instructions



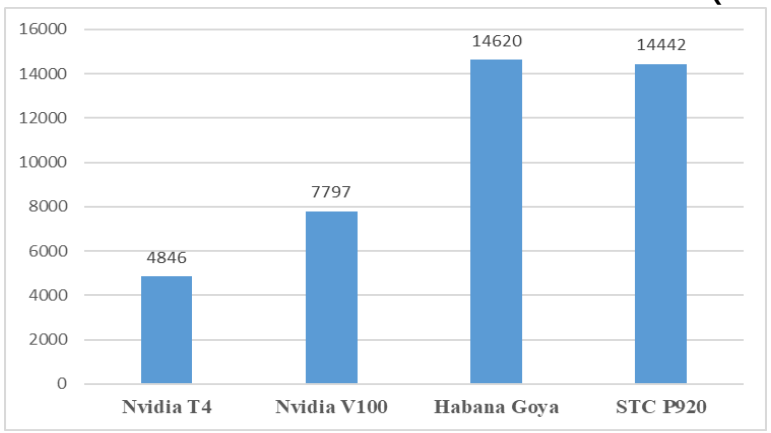
P920 NPU: Stream Computing 1st Gen NPU Product for AI Inference



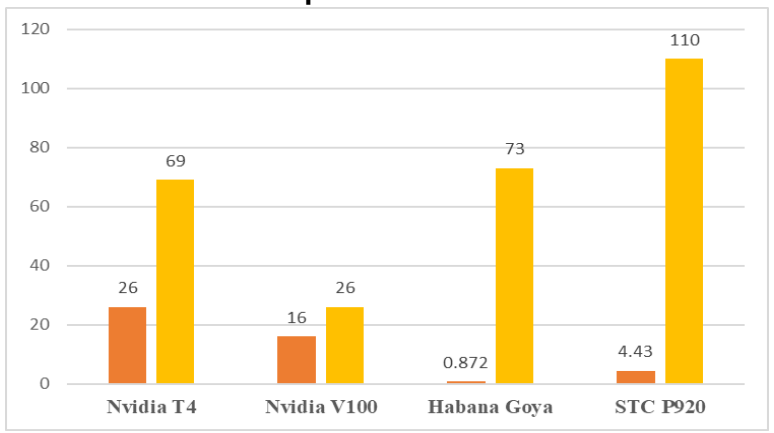
P920 NPU Features	
Chip Area	~400mm^2
Process	TSMC 12nm FinFET
GPNPC Cores	x32
Peak Performance (FP16)	128 TFLOPS
Peak Performance (INT8)	256 TOPS
TDP	130W
Memory	16GB LPDDR4
Host IF	PCI Gen4 x16

P920 NPU: Stream Computing 1st Gen NPU Product for AI Inference

ResNet-50 V1.5 (INT8) Performance Comparison



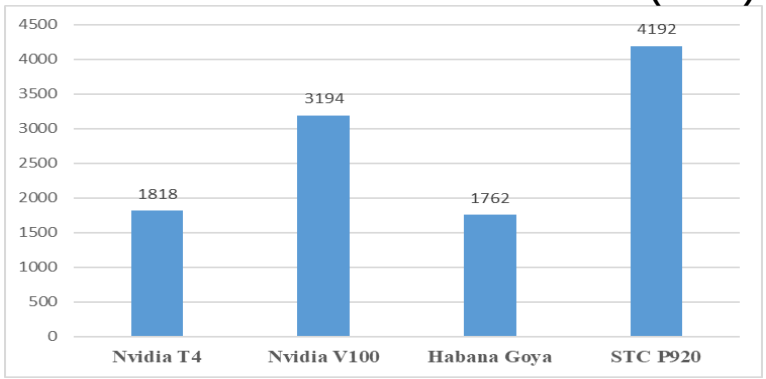
Performance (images/second)



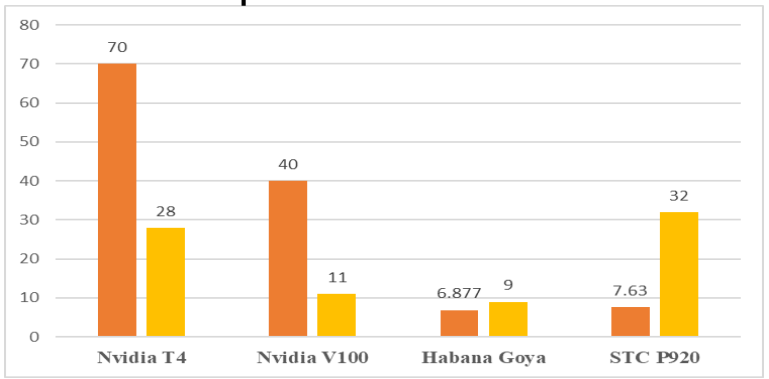
Latency (ms)

Power Efficiency (images/second/watt)

BERT (FP16) Performance Comparison



Performance (sentences/second)



Latency (ms)

Power Efficiency (sentences/second/watt)

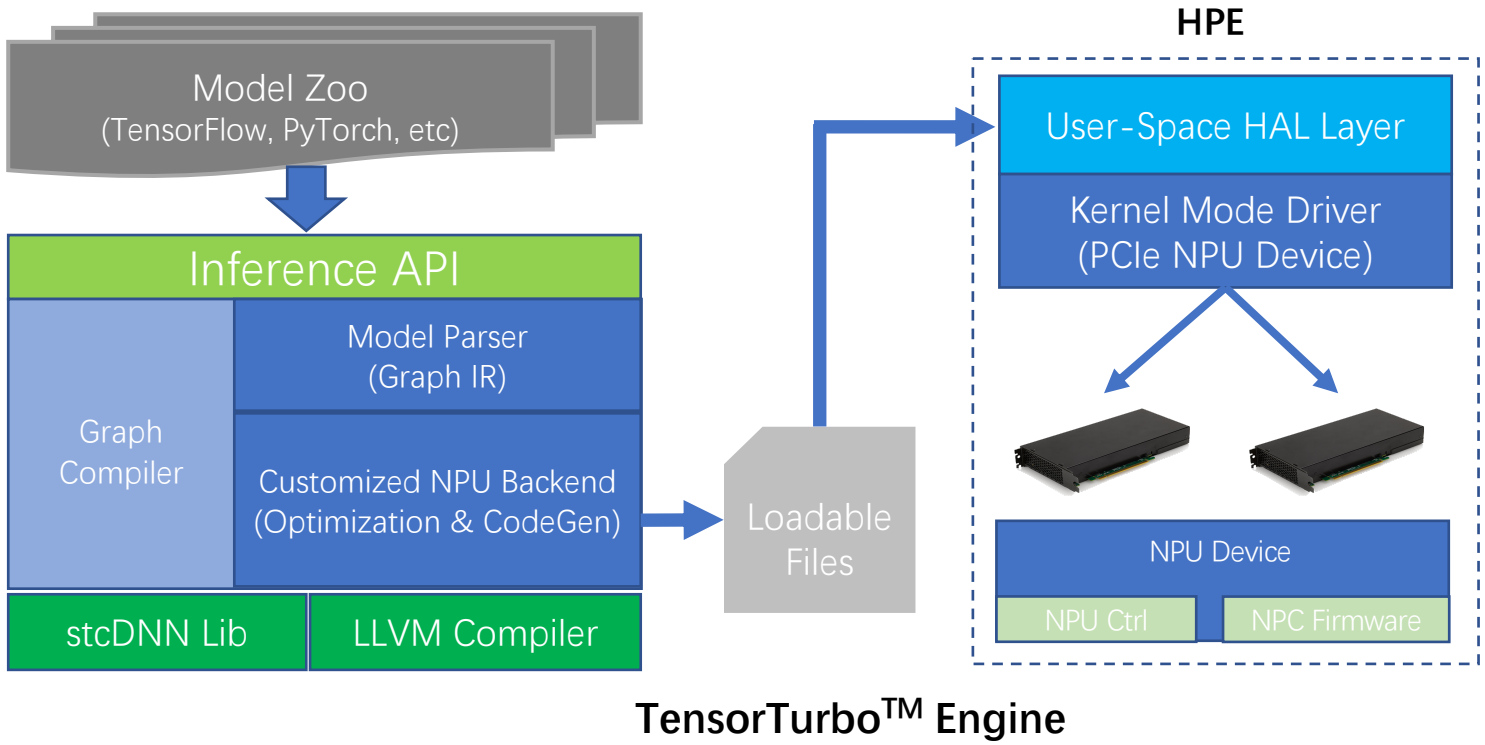
High Throughput for both CNN & NLP Models

Critical Latency Performance for Real-Time AI Inference

Power Efficiency: Balanced Design for Various AI Domain NN Models

TensorTurbo： E2E Inference Stack for STC P920 NPU

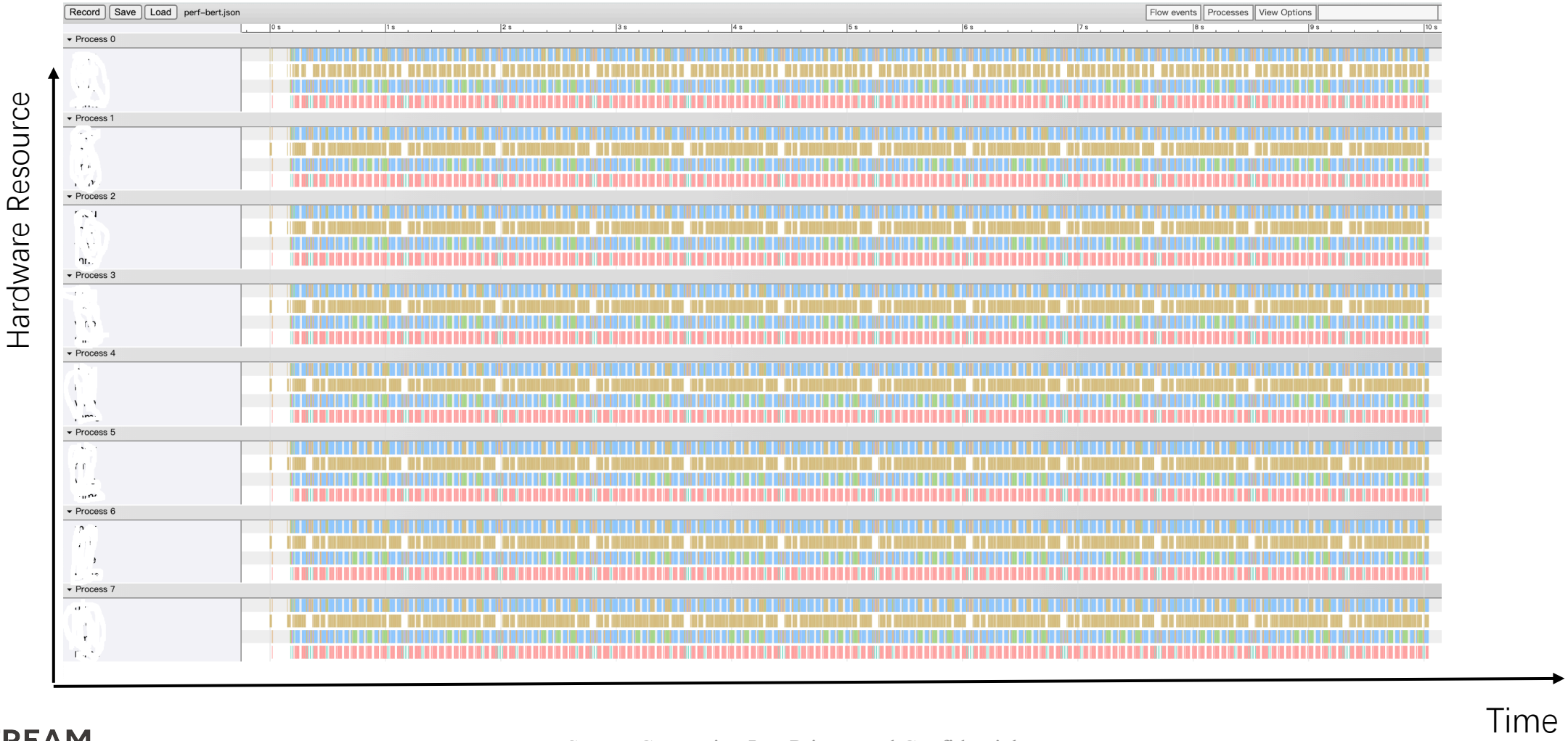
- **Use Case**: Customer has a pre-trained NN model, and wants to deploy it onto STC P920 NPU to execute AI Inference.



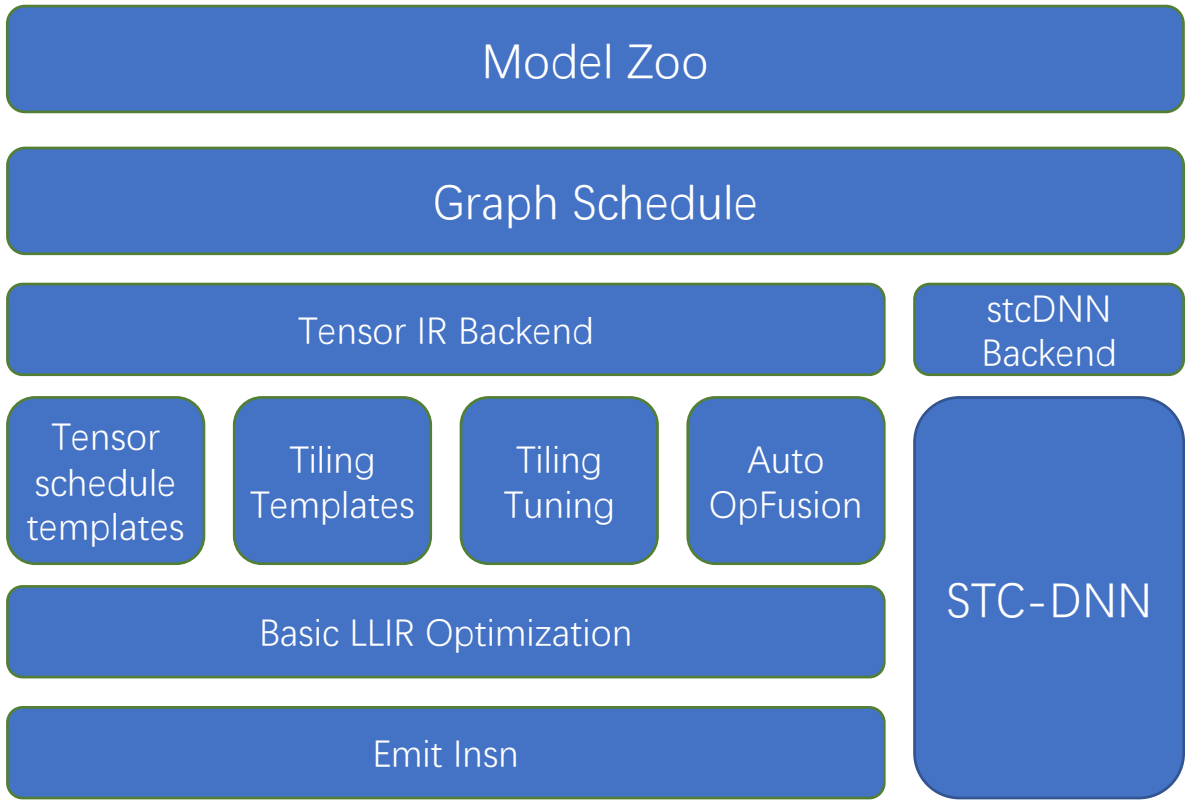
- **Model Zoo**: Pre-Optimized Models for STC P920 NPU
- **Graph Compiler**: TVM-based, deeply customized for NeuralScale architecture.
- **Heterogenous Program Engine (HPE):**
  - C/C++ Level Heterogenous Computing Kernel Program

TensorTurbo： Neural Network Compilation

- **Fundamental Challenges for AI Compilation: It is all about how to schedule/optimize your AI program, so then you can maximize the hardware resource utilization to gain minimum AI program execution time.**



TensorTurbo： Neural Network Compilation

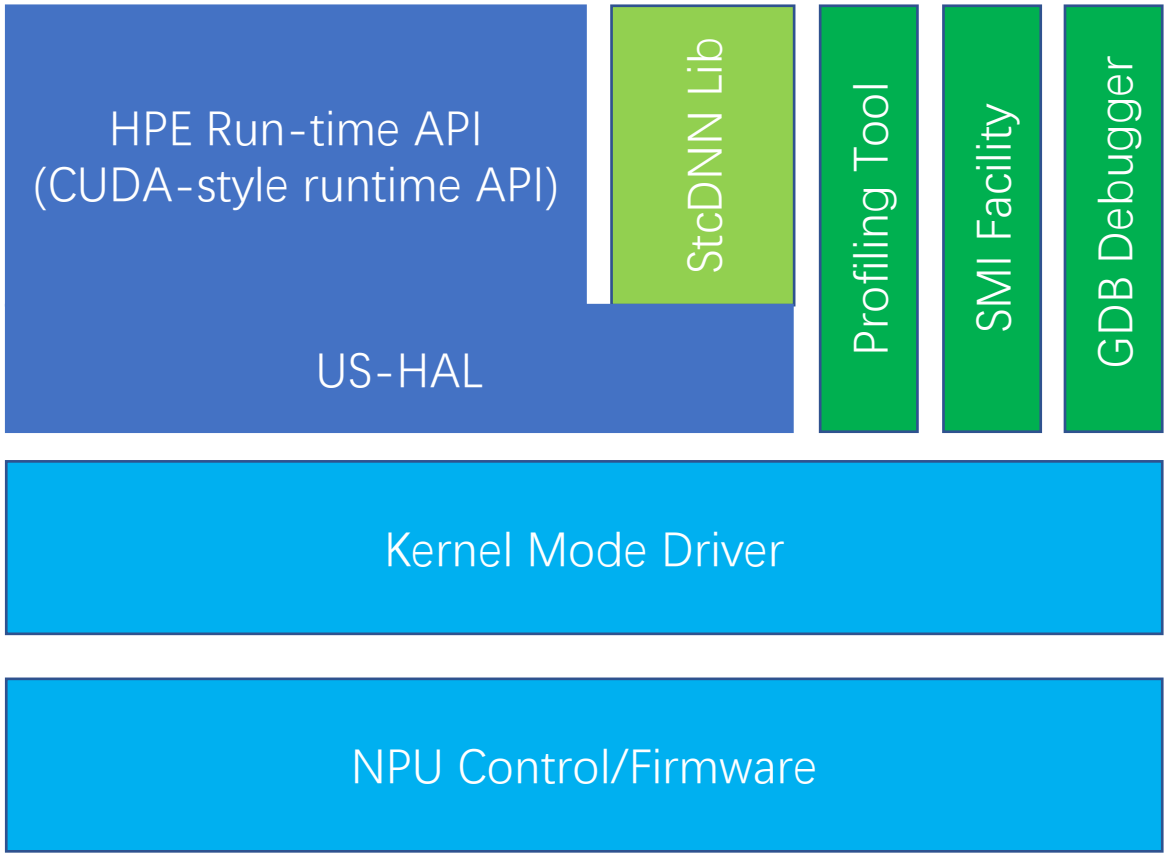


TensorTurbo TVM-based Graph Compiler

- Graph Schedule: Split Batch Dim Input Feature Map to make intermediate data resident on L1 Buffer
  - Heuristics Auto Graph Schedule!
  - A set of graph schedule APIs support users manually split batch-dimension feature map data.
- Tiling Strategies within an Operator:
  - Heuristics Templates
  - Auto Tiling
- Operators Schedule:
  - OpSchedule Template: Reduce the effort of TVM IR-based operator development
- Backend Optimization PASS:
  - VME/MME Insn Schedule, DMA Schedule： Exploit ILP
  - Auto-Sync Insert
  - Double Buffer
  - On-Chip Buffer Allocation/Bank Conflict.

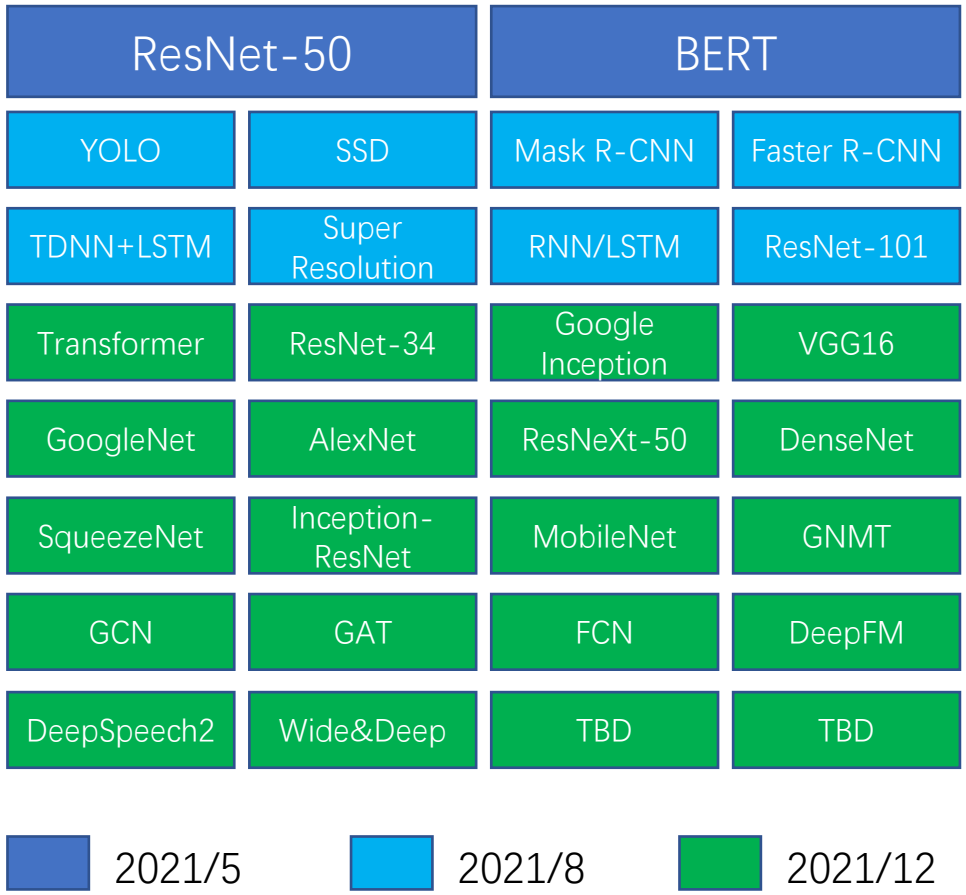
TensorTurbo： Heterogenous Program Engine (HPE)

- User-Space HAL provides CUDA-style Runtime APIs:
  - Device Management API
  - Kernel Launch & Management
  - Device Memory Management
  - Host/Device Memory Movement
  - Stream Programming APIs
  - Event Management cross multiple streams
- Utilities:
  - stcGDB: Use GDB to debug a heterogenous program
  - stcProf: Program Performance Profiling Tool
  - stcSMI: System Monitor Interface Tool
- NPU Firmware:
  - Manage Computing Kernels to be launched in NPU device



TensorTurbo™: Model Zoo

- Plan to support total ~30 Optimized NN models in CY2021 (2 + 8 + 20):
  - Image Classification
  - Object Detection & Segmentation
  - Video Enhancement
  - Speech Recognition
  - NLP



Stream Computing P920 NPU Competition Advantages Summary

Turing-Complete General NPC

NeualScale™: Advanced RISC-V based NPC Architect

- Good Flexibility
- Good Scalability: one NPC architecture fits for both inference and training from cloud-side to edge
- Good Programmability: Traditional C programming paradigm

Extreme Cost-effectiveness for AI Computation

Reduce TCO of AI Inference Server significantly

- High Throughput performance: ResNet-50 & BERT
- Low Latency: Satisfy latency-sensitive real-time AI application
- Low Power: 130W TDP
- Price：Market Competitive Price.

Advanced E2E Neural Network Toolchain

TensorTurbo:Industry Leading E2E NN Toolchain

- TVM-Based Graph Compiler: Deep NN Compilation Optimization maximize the performance
- Model Zoo：Pre-optimized models allows you getting started quickly.
- Operator Programming IF: customize layers/operators.

Thanks

Q&A