RISC-V® Summit

Scalable, Configurable Neural Network Accelerator based on RISC-V core

Karthik Wali
Staff Design Engineer
LG Electronics

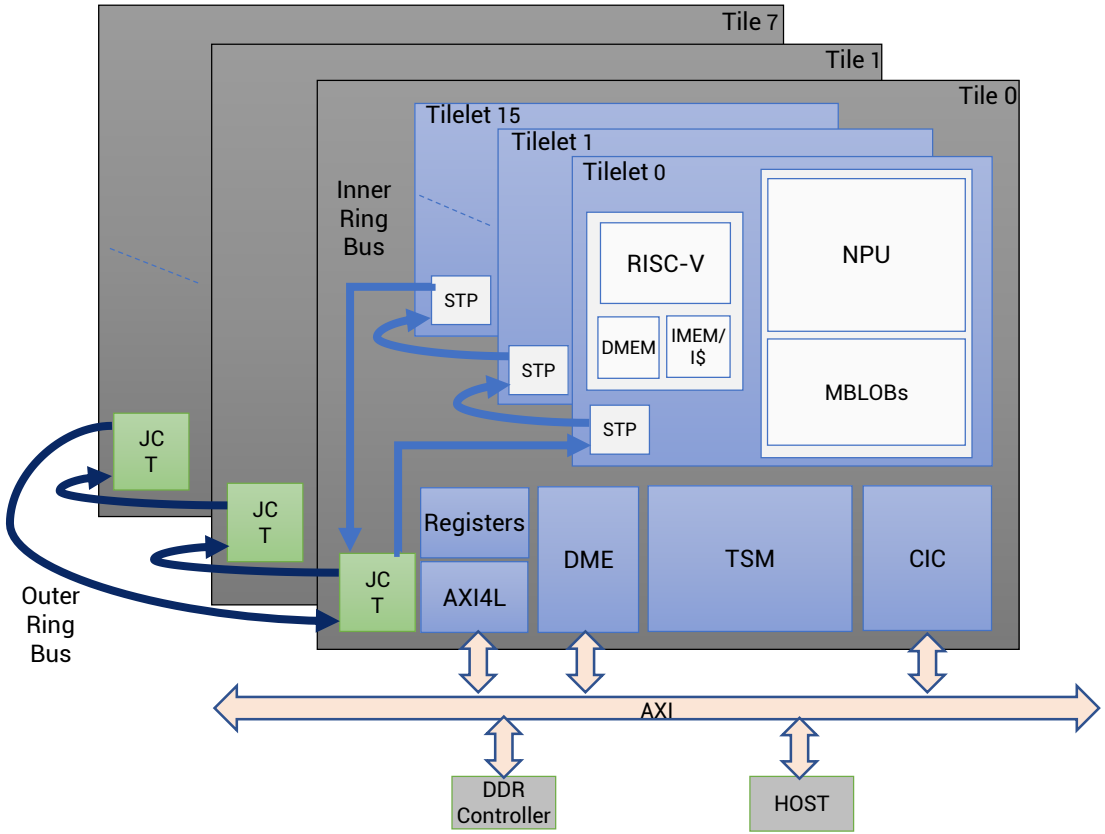
Introduction

- Neural Network algorithms are increasingly used in many machine learning tasks such as image classification and speech recognition

- Neuromorphic Accelerator is a class of accelerators which are highly optimized for executing Neural Networks in general and Convolution Neural Networks (CNN) in particular

- One such CNN accelerator is **LG Neural Engine (LNE)**

LG Neural Engine



LNE's Features

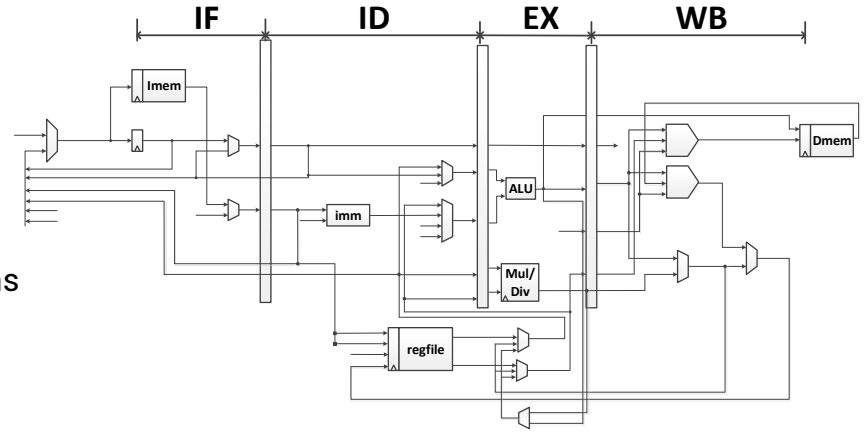
- Scalable and flexible implementation to meet requirements from IoT to high-end applications
- Supporting both inference and on-device learning for edge devices
- Customized ISA extension to RISC-V to support neural network functions
- Supported and verified over different neural networks such as
  - Image classification : GoogLeNet, MobileNet, ResNet-50, VGG-16, SqueezeNet, etc
  - Object detection : Tiny-Yolo, SacNet, SacNet-yolo-tiny, etc
  - Image segmentation : SegNet, ErfNet, etc
- Developed with area and energy-efficient implementation
- High speed design with target frequency of 1.0+ GHz
- Industry standard AXI bus interfaces for easy SoC integration
- Internal computations use either 8-bit or 16-bit fixed-point data format
- Support binarized weight for higher FPS and lower memory BW
- Customized, instruction-accurate SPIKE simulator
- Silicon proven architecture by TSMC 28nm HPC+

LNE's Architecture

- LNE can be scaled up to 128 cores (theoretically no limitation but physically)
- Each Tilelet contains one RISC-V and NPU w/ Memory BLOBs (MBLOBs)
- Each Tilelet can operate independently, simultaneously and synchronized if needed
- Tilelets share a common, asynchronous Data Movement Engine (DME)
- Tilelets share a Common Instruction Cache (CIC)
- Tilelets transfer data through Inner and Outer Ring Bus
- Tile Shared Memory (TSM) can serve as on-chip memory to lower DDR accesses, so power consumption
- Access to internal register space through AXI4L Slave
- Access to external (on-chip shared or DDR) memory is through the AXI4 Master

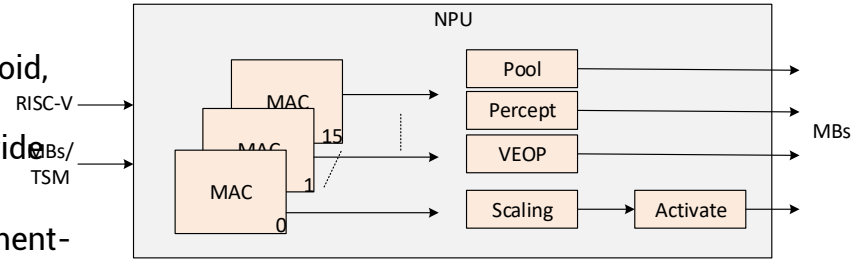
RISC-V

- Why RISC-V?
  - Free open source architecture
  - Ability to add custom instruction set
  - Easy migration to ASIC
  - SPIKE & RISC-V Toolchain
  - Parameter Computations
  - Support functionality not in NPU
- RISC-V Features
  - RV32IMC optional M and C extensions
  - 4-stage pipeline
  - High Speed Design
  - Configurable Multiplier
    - High Frequency (latency up to 32 cycles)
    - Low Frequency (latency 4 cycles)
  - Hardware Divider (latency 16 cycles)
  - Co Processor Interface
  - RISC-V cores support from IoT to high end devices



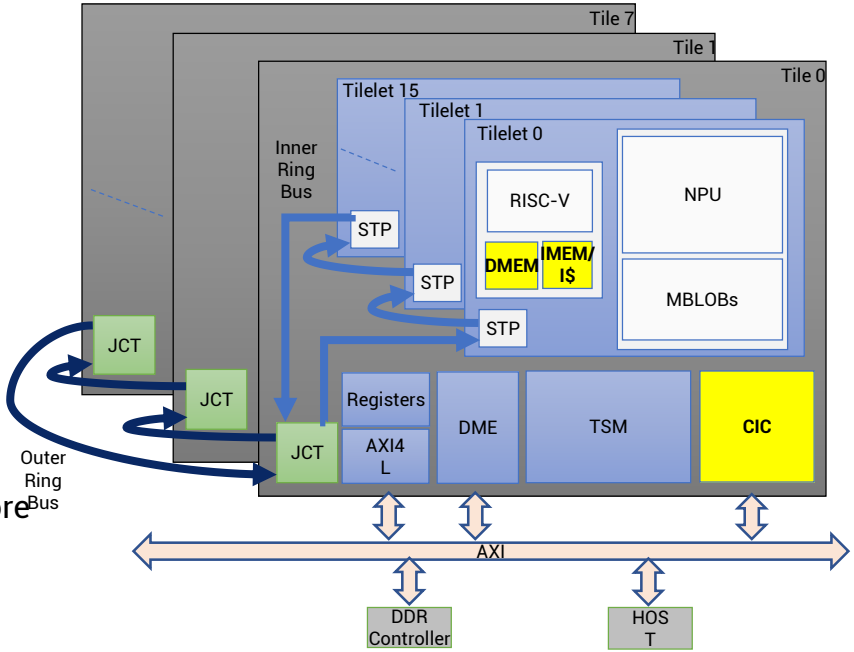
Neural Processing Unit (NPU)

- NPU accelerates all the neuro-morphic related operations with RISC-V extended ISA
- NPU supported functions
  - Activation - ReLU, pReLU, ReLU6, Sigmoid, TANH
  - Pool - Average, Max, Index, Variable stride
  - Percept
  - Vector operations - normalization, element-wise addition and multiplication
  - Convolution - 2D/3D with variable stride, dilation, bias per kernel and activate and binarized weights
- NPU interfaces with MBLOBs and TSM
- Configurable Pipeline Stages
- NPU has
  - 8 lanes of 16x16 MACs, or
  - 16 lanes of 8x8 MACs
- Total MACs in LNE
  - 8 Tiles x 16 Tilelets/Tile x 16 MACs/Tilelet = 2,048 MACs



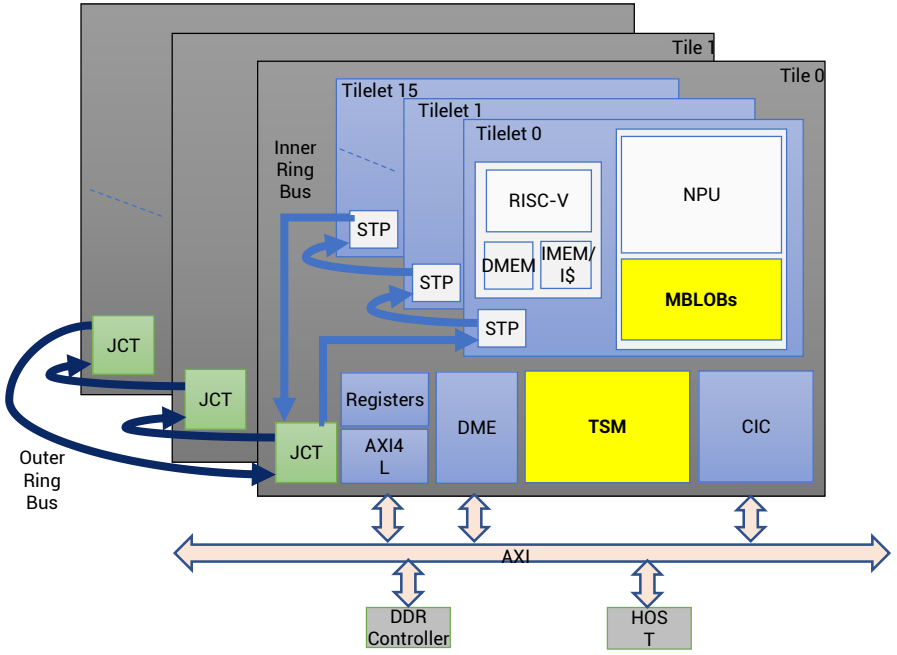
Common iCache/Data Memory

- Common iCache
  - 4KB/8KB options with 4-way set associativity
  - 256B cacheline
  - 4 Banks, address partitioned
  - Support for 16 requestors
- Per core iCache and Instruction Memory available as options to replace Common iCache.
- Data Memory
  - Localized memory to each RISC-V core
  - 2KB/4KB/8KB options
  - DDR accessible



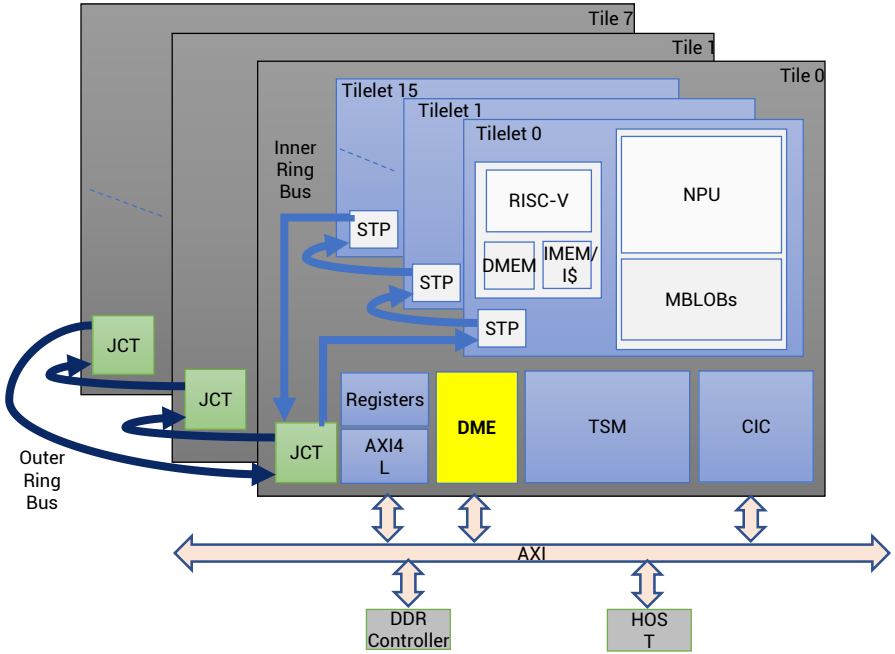
MBLOBs and TSM

- Memory BLOBs (MBLOBs)
  - MBLOBs – each as source for data, weights or destination for results
  - Extra MBLOB option is available for extended usages
  - 2KB/4KB/8KB options are available per MBLOB.
- Tile Shared Memory (TSM)
  - On chip memory for lower memory BW
  - Data movement between TSM and DDR
  - Data movement between TSM and MBLOBs
  - Data movement from TSM to NPU the data is directly from TSM, not from MBLOB
  - 64KB/128KB/256KB/512KB options



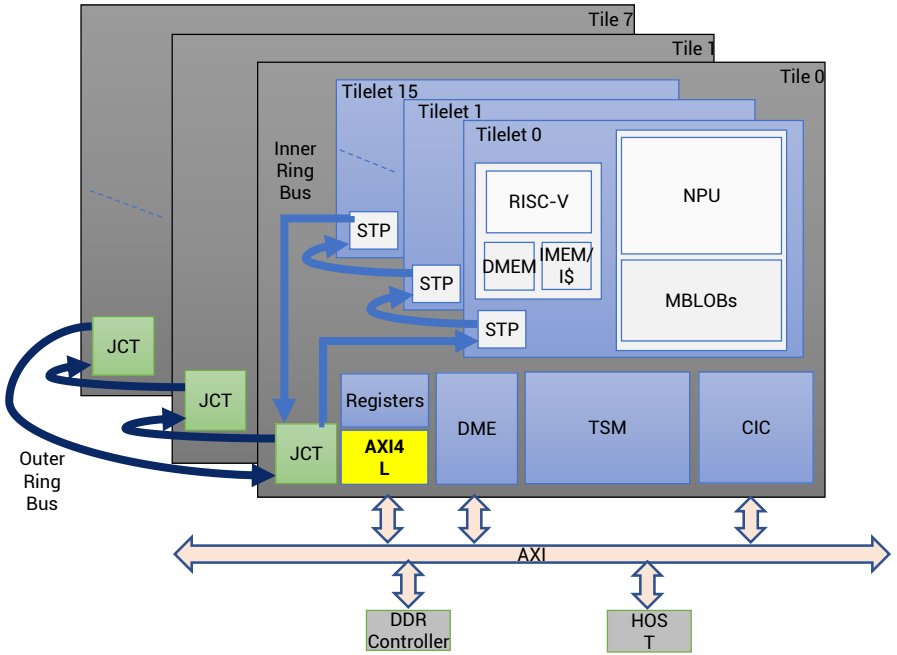
Data Movement Engine (DME)

- Operations are asynchronous to RISC-V and are enabled with DATAMV instructions from RISC-V
- Data packets can be stored or loaded with 2D/3D memory layout
- Support for data transfer from/to Tilelets with broadcast/multicast feature
- Support for data transfer from/to TSM
- Support for data transfer from/to DMEM



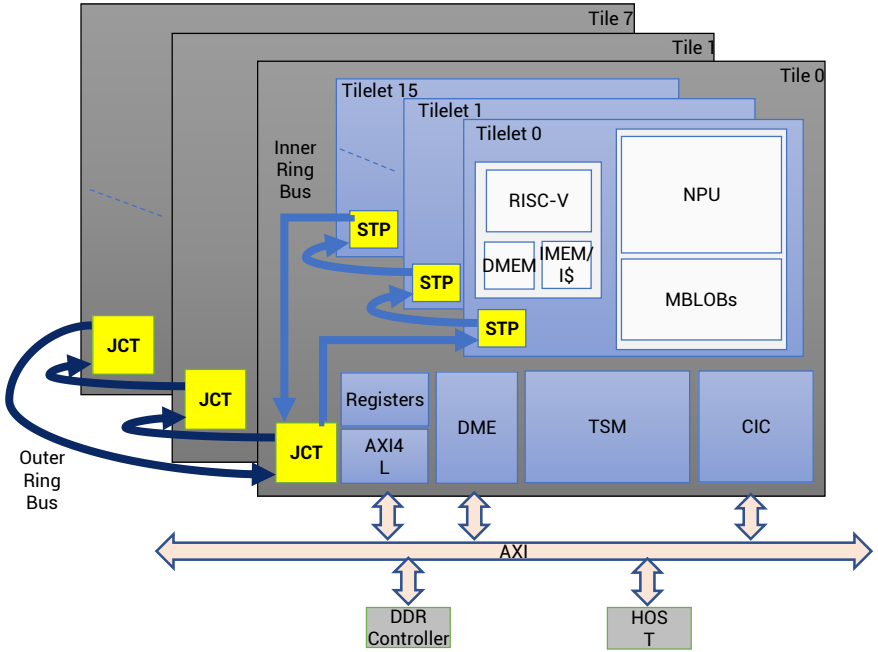
Host Access

- HOST can communicate with each Tile (and Tilelet) through AXI
- HOST can access Registers in Tile or the custom registers in RISC-V (not GPR of RISC-V)
- HOST can access DMEM and MBLOBs
- HOST can activate each Tilelet individually or altogether

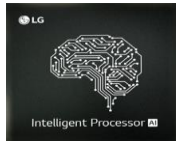


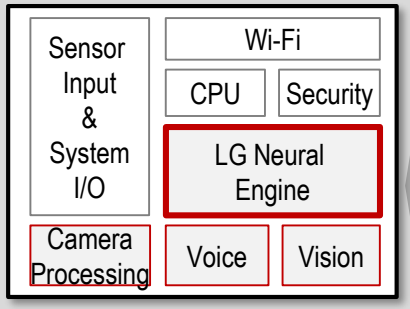
Ring Bus

- Data can transfer across the Tiles and Tilelets
- Instruction enabled asynchronous packet based multi-hop ring bus data transfers
- Stream Transfer Port (STP) transfers the data packets between Tilelets through Inner Ring Bus
- Junction (JCT) routes the data packets between Tiles through Outer Ring Bus



LG's AI SoC



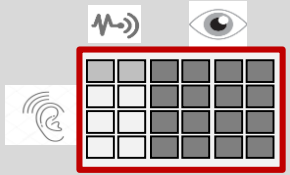


Brain

Sense

Efficiency & Privacy

On-Device neural network acceleration and learning



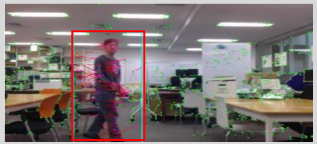
LG Neural engine

Scalability, Flexibility
On-Device Learning

Sensor processing of Vision, Hearing and Touch

Low light enhancement
with camera pre processing

Low Power Vision Intelligence and Secure A.I.

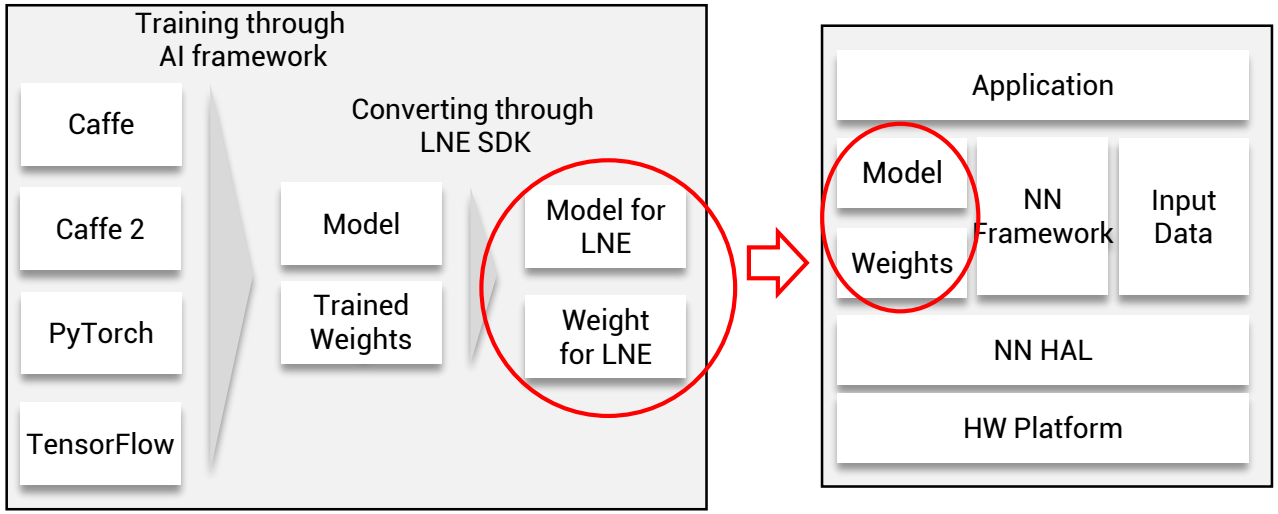


Human recognition and tracking

Meta Data

LNE's Software Flow

- LNE software framework supports trained models in Caffe, TensorFlow and PyTorch
- LNE SDK converts and maps the model and trained weights on Tiles and Tilelets
- Trained model is compiled with RISC-V GCC and ported on LNE



Demo

Finding Target

THANK YOU