LUMISSIL Microsystems

A Division of ISSI

Application Note
X2000 for Edge AI

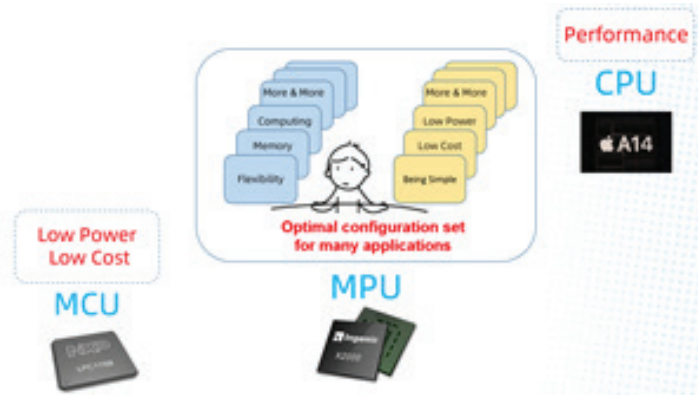
1. Overview

X2000, a SoC released by Ingenic, utilizes the Ingenic CPU core XBurst2. A rich feature set enable wide range of innovative solutions. Among its features are heterogeneous tri-CPU architecture, a built-in Gigabit Ethernet interface which is compliant with IEEE1588-2002, ability to connect 3 cameras simultaneously, low power consumption etc. This feature set is an excellent baseline for innovative solutions.

X2000's potential for AI solutions falls into two categories: First, there are resources built in the SoC such as computing capacity, memory on chip and low power technology base; Second, Ingenic offers the MAGIK deep learning neural network development kit, which can be used to train, optimize and deploy AI systems. These resources help boost Ingenic's SoC value for AIoT applications.

The AI capability of X2000 is the focus of thisapplication note.

2. Edge AI as a Major Strength of X2000



I X2000's feature set define it as a MPU. Figure 1 outlines the differentiation of a MPU, when compared to a CPU or a MCU.

For a lot of Edge-AI solutions , the MPU positioning as described in Figure 1 will be optimal - More powerful then a MCU and significantly less costly then a CPU.

Edge AI systems have several advantages, such as:

- Faster and closer interaction with users since the algorithms run locally with smaller latency.
- Autonomous service since the AI services can still be sustained without network connection.
- Better protected privacy since the related data is collected and stored at one location.

But edge AI also brings about many challenges, such as:

- The algorithms should not be too complex as the local computing capacity is limited and specific kits are needed to train and optimize models for the edge devices.
- The size of models is limited by memory size at the edge device, and therefore there is a tradeoff between performance and memory when building the models.
- The modules must be power efficient since the power supply at the edge device is typically limited.

X2000 basics

As a MPU, X2000 offers many advantages such as: Ability to run an OS like Linux or RTOS.

- Multi-media processing with H.264 coding and decoding.
- Connectivity with Gigabit Ethernet interfaces.

The most important and valuable strength of X2000 is that it can be used in an edge AI (or embedded AI) solution, utilizing its features such as:

- Tri-CPU compute capacity architecture and the tailored training and optimizing tools provided by Ingenic MAGIK.
- Memory on chip which provides a sound base for MAGIK to run itsmodels.
- Low power consumption technology.

Table 1 describes more comprehensively X2000's potential to enable edge-AI solutions:

Basic Functionality	Extended Functionality
X2000 can code and de-code video when used in a video camera.	Beyond video coding and decoding, X2000 can be used for human shape recognition, facial recognition and other objects recognition.
X2000 can process image when used for image capture.	Beyond image processing, X2000 can be used in object detections and image recognitions.
X2000 can process audio when used for speech compression and voice triggering.	Beyond audio processing, X2000 can be used in speech recognition and text-to-speech applications.
X2000 provides connectivity for terminals to communicate with other equipment.	X2000 can provide high performance connectivity for smart devices and build a strong network with cloud-edge communication.

TABLE 1 X2000's extended function

3. X2000's Computing Capacity

Armed with both an XBurst2 (which can be configured as 2 logical CPUs), and an XBurst0, a MIPS instruction set and SIMD extension instruction set, X2000 boasts a rather strong level of computing capacity.

The main CPU core —XBurst2

The XBurst2 is a new member of Ingenic's CPU core family. As the main CPU core of X2000, XBurst2's key features are :

- 1.2GHz working frequency.
- With Simultaneous Multi-Threading technology, the XBurst2 can be configured as 2 logical CPUs. In comparison to two physical CPUs, the two logical CPUs cooperate better and are more power efficient.
- Dual-execution per cycle per logical CPU.
- 32KB L1 Instruction Cache and 32KB Data Cache.
- Floating Point Unit and Programmable Memory Management Unit.
- 512KB L2 Cache.
- Advanced power management scheme including clocks turn-off to idle modules

The secondary CPU core —XBurst0

X2000 has a secondary XBurst0 CPU core, based on MIPS ISA. It's key features are:

- 240MHz working frequency..
- 32KB Tightly Coupled Sharing Memorywhich is accessible by the main CPU core and the DMA scheme.

Note: A separate Application note,titled "X2000's Three CPUs" can provide more details.

The SIMD instruction set extension

XBurst2 implements 32-bit MIPS32 ISA R5 and SIMD instruction set extension. The extension consists of 2 parts:

- MIPS SIMD ISA: MSA128.
- Ingenic 128bit SIMD ISA: MXA128.

As a result, XBurst2 is capable of:

- Vector compute of both integer and floating point data.
- Audio/video processing acceleration and smart applications in which speech recognition, facial recognition and human/objects detections are involved

4. Memory in X2000

128MBytes LPDDR3.

5. Low power of X2000

The typical power consumption of an X2000 is less than 380mW.

Note: A separate Application note, ntitled "X2000's Low Power Potentials" can provide more details.

6. Operating Systems running on an X2000

Linux 4.4 is well verified to run on X2000. Other OS's are possible as well.

7. MAGIK - Deep learning neural network applications development kit

MAGIK, Ingenic's development kit for deep learning neural network applications, is a platform for AI model quantization, model transformation and model deployment. It can greatly accelerate AI application development for X2000. MAGIK's architecture is described in Figure 2.

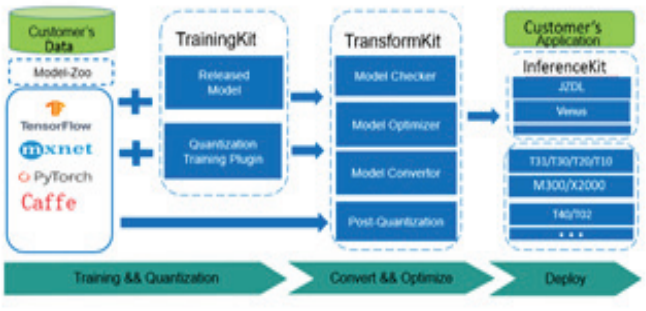


Figure 2: Ingenic Deep Learning Platform MAGIK

3 steps.

I. Model quantization and training

Several possible approaches are:

- MAGIK's Release Model can provide models developed by Ingenic. These models can be trained with customers' data to get 4~8-bit quantized models.
- MAGIK includes a Model Zoo, which consists of popular models which can be trained with customers' data to get 4~8-bit quantized models.
- TensorFlow, mxnet, PyTorch or Caffe can be used in training to get float or 4~8-bit quantized models.
- Well verified models can be transformed directly without quantization and training if performance of the AI system is not a major concern.

II. Model transforming

Models can be transformed in three steps:

- Model checking: The Model Checker figure out if the models are suitable for the hardware platform and whether the operators needed are in the library provided by MAGIK.
- Model Converting: The Model Converter converts models trained with different platforms into MAGIK models.
- Model Optimizing: The Model Optimizer optimizes models to best utilize the features of the hardware platform.

III. System deploying

MAGIK includes an inference firmware package named JZDL for X2000.

8. deploying an AI application on X2000

After being trained, models are deployed on hardware platforms. Inference is the process during which a trained model getting prediction (classification or regression) on the basis of available data.

JZDL is a module in MAGIK and it is the AI inference firmware package for X2000 with the following features:

Tailor-designed for X2000

- JZDL has been optimized based on XBurst2's SIMD and part of the coding is done in assembly language to ensure best efficiency.

Popular supported technological modules

- Supporting Operators such as common Convolution, Depth-wise Convolution, Pooling, Activation, Full-connection, Squeeze and Excitation, Concat, and Split among others.
- Supporting Multi-input or Branchy Neural Networks.
- Supporting forward inference with float precision or 8bit/4bit quantization.

Operation under Limited resources

- Well-designed memory management schemes and data structures and only 393KB memory is needed to run the inference package.

Applications

- Detections: human shape detection, human face detection, pet detection, biometric liveness detection (face anti-spoofing), cry detection, vehicle detection among others
- Recognitions: facial recognition, OCR, plate recognition among others

9. Conclusion

As a device for edge AI applications, X2000 has many

capabilities. First, available computing resources in the IC such as XBurst2, with the MIPS ISA and its extensions and advanced memory management. Second are well-matched operating systems and well-verified development kit packages. And last but not least, is the MAGIK AI development platform.

As it is, X2000's AI capacity is consisted of the combination of image processing, video coding/ decoding, and connectivity. Such capacity makes it possible for an X2000 to be deployed as a standalone edge AI agent. Moreover, it is possible for several X2000s to be implemented together in a system and provide much stronger AI capacity as a larger scale AI agent or distributed intelligent system.