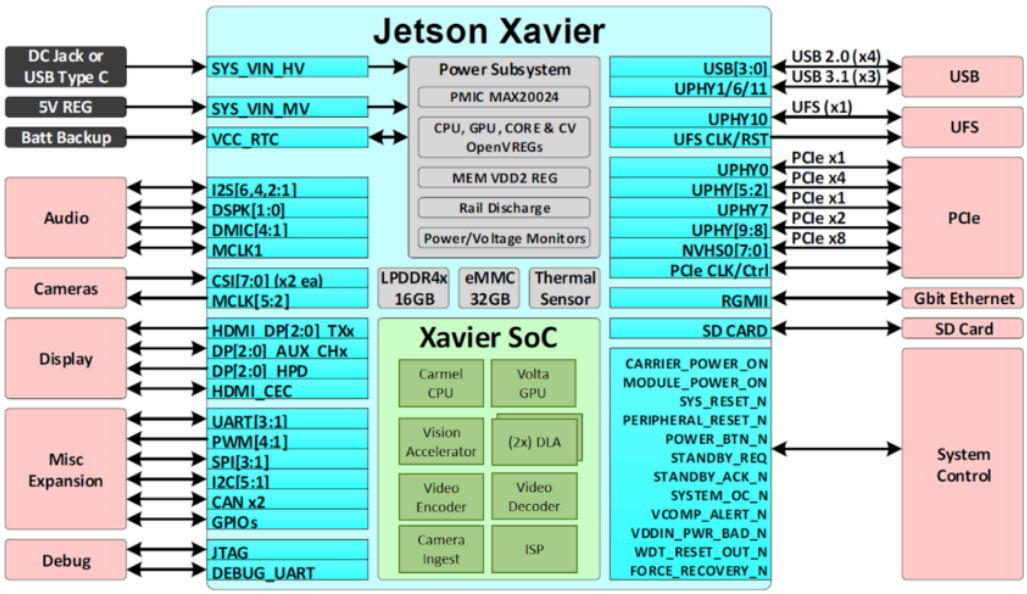
EvoTegra GmbH
Kreuzwingert 11 • D – 55296 Gau-Bischofsheim

EVOTEGRA
CONCEPTS & INTEGRATION

AI in the box

The Advantech MIC 730AI is a rugged AI inference system based on the NVIDIA Jetson® Xavier. With 19x23x9cm the system comes in the dimensions of a Mini-PC. The key differentiator to a common Industrial PC is the fact that Xavier boards have an integrated GPU, which has enough computational resources to process demanding AI loads. Among others the board is equipped with video de- and encoder as well as Deep Learning Accelerators (DLA). With an operating temperature of -10 to 50°C, passive cooling and a low power consumption the system is designed to operate on "the edge". Edge systems operate close to the sensor to avoid transfer of large amounts of data over the network.



Hardware

The Xavier in the MIC-730AI comes in 2 flavors. The smaller version comes with 8GB of RAM at a clock of 1333Mhz and 6 ARM as well as 364 Volta GPU / 48 Tensor cores while its big brother comes with 16 GB at 2133 Mhz and 8 ARM as well as 512 GPU / 64 Tensor cores. Looking at the effective performance for AI applications the 16GB version offers twice as much computational power as its 8GB counterpart. But while the bigger version allows more flexibility during a research and development stage, the smaller version should already be fine for most industrial production workloads. Both versions come with 32 GB internal storage that contains the operating system and about 16 GB free space for additional applications.

EvoTegra GmbH
Kreuzwingert 11 • D – 55296 Gau-Bischofsheim

EVOTEGRA
CONCEPTS & INTEGRATION

Connectivity

Both versions come with HDMI, 2 RJ45 GbE connectors, 2x USB 2.0 as well as 2x USB 3.0 connectors on the front panel. For industrial applications there are 2 COM ports as well as 16 DI/DO connectors. *Clever:* Under the back cover there are another 3x USB 2.0 connectors on the board along with a supplemental COM port, a 5V power supply as well as a Nanosim-Slot.

Extensibility

Talking about extensibility the case has a bay for a 2.5'' drive while the board offers a SATA connector and M2 as well as a Mini PCI-Express slot. Those who need support for full-size PCI-Express cards can extend the case with the MIC-75M20 expansion module that offers 2 slots for PCI-Express cards with 16/4 lanes. Clever: The "iDoor" on the front panel. This slot can be used to extend the MIC-730AI with a wide range of Mini-PCIe based extensions like Wifi, LTE or industrial field bus adapters.

Integration and safety

The system is available either with mounting brackets / wall mount or as desktop version. To access the flash port the case must be opened. For security sensitive applications it is therefore possible to seal the casing.

Test 1 Bulk-Processing

The first test scenario is a bulk processing use case optimized towards low latency which is typically found e.g. in automotive or some industrial use-cases. We chose the following setup:

- MIC-730AI
- 2 Basler ACE Industry cameras with 1920x1200 pixel resolution
- GPU-based stereo camera system for 3D reconstruction
- Middleware software (Publish/Subscriber message bus)
- Visualization with 5-10 million 3D points/sec
- Deep Learning based object detection with an input size of 960x600 pixels
- Visualization of object detection

In this configuration the system is processing 625 Megabytes of data per second which is using the available memory bandwidth for about 40%. AI loads are mainly processed on the GPU. Therefore the load on all the CPU cores is just 50% while GPU is loaded with more than 80% on average.

Test configurations

We tested the MIC-730AI in 2 different power modes. The power modes can limit the power consumption to 10, 15 or 30W by adapting the number of CPU cores used as well as the clock frequencies of both CPU and GPU.

Mode Name	EDP	10W	15W	30W	30W2	30W3	30W4
	MAXN	MODE_10W	MODE_15W	MODE_30W_ALL	MODE_30W_6CORE	MODE_30W_4CORE	MODE_30W_2CORE
Power Budget	n/a	10W	15W	30W	30W	30W	30W
Mode ID	0	1	2	3	4	5	6
Number of Online CPUs	8	2	4	8	6	4	2
CPU Maximal Frequency (MHz)	2265,6	1200	1200	1200	1450	1780	2100
GPU TPC	4	2	4	4	4	4	4
CPU Maximal Frequency (MHz)	1377	520	670	900	900	900	900
DLA Cores	2	2	2	2	2	2	2
DLA Maximal Frequency (MHz)	1395,2	550	750	1050	1050	1050	1050
Vision Accelerator (VA) cores	2	0	1	1	1	1	1
VA Maximal Frequency (MHz)	1088	0	550	760	760	760	760
Memory Maximal Frequency (MHz)	2133	1066	1333	1600	1600	1600	1600

Xavier power modes

In the MAXN power mode both CPU and GPU operate at their maximum speed while there is no guaranteed power budget. In order to stress the system and to validate the heat dispersion of the passive cooling system, we conducted a 24h stress test in this mode.
The 30W ALL mode is a mode which limits the power consumption to 30 watts while using all 8 ARM cores.

Power-Mode: MAXN
CNN Inference time: ~18ms
Power consumption: ~50W
CPU load: ~50%
CPU temperature: 66,5°C
GPU load: ~80%
GPU temperature: 65.5°C
Board temperature: 57.5°C
Cooler Max temperature: 49.2°C

Power-Mode: 30WALL
CNN Inference time: ~35ms
Power consumption: ~35W (including peripherals)
CPU load: ~65%
CPU temperature: 53°C
GPU load: ~85%
GPU temperature: 53.5°C
Board temperature: 48°C
Cooler temperature: 43.3°C

Test results
During the tests in both modes the system processed more than 100 Terabytes in total. At room temperature both the CPU and GPU reached up to 66°C heating up the passive cooler to a maximum of 50°C. Nevertheless this was way below the maximum recommended GPU temperature of 88°C. The system operated stable at all time.

Test 2 On-Demand Processing
In this scenario we test the system in a typical industrial use-case. The system receives the data via the network interface and the goal is to process the data with the lowest latency possible.

- MIC-730AI
- C++ Network Client Software
- Deep Learning based object detection and classification with an input size of 1024x1024 pixels

To save energy even in the fastest power mode (MAXN) the clock for CPU and GPU is throttled under low load. The effect of the initially throttled GPU is that the first images have up to a 3 times higher latency than the subsequent images. While in most use-cases the energy saving features are unproblematic, for time critical processing this is an undesirable behavior.
The solution therefore is to define a custom power mode that keeps the GPU clock at its maximum clock at all times.

Power-Mode: CUSTOM
CNN Inference time: ~40ms
Power consumption: ~30W
CPU load: ~5%
CPU temperature: 40°C
GPU load: ~10%
GPU temperature: 40.5°C
Board temperature: 37°C
Cooler Max temperature: 34.4°C

Test results
As the system has idle time between the processing of the individual images keeping the GPU at maximum clock speed did not have a noticeable effect on the temperature.
The inference time is about 1,75 times as long as what a full size industrial PC based on the Intel Xeon in combination with a NVIDIA RTX 2080Ti GPU can achieve. Yet if we compare the technical specifications of both systems the industrial PC has about 8x as many computational resources compared to the MIC-

730AI. The fact that the 8x advantage in processing power just yields a 1.75x advantage in processing time indicates that the latency is mostly related to memory throughput.
But compared to the industrial PC the MIC-730AI has just 5-10% of the power consumption, no moving parts and far less components in general. Therefore we expect the MIC-730AI to have lower maintenance costs.

Operating System

The system comes with Ubuntu 18.04 and the "Jetpack". The Jetpack contains the NVIDIA libraries and tools to accelerate Deep Learning loads on the system. Technically the environment provides the same amenities as a GPU based PC-system. Therefore the deployment of a network trained on a PC or in the cloud is straightforward. While it is technically possible to train neural networks, the system is optimized for inference loads. Networks can be executed natively using Python scripts but to get the best value for money of the MIC-730AI we recommend to use C++ along with network optimizers. The NVIDIA optimizer TensorRT is preinstalled with the Jetpack and can optimize networks from various Deep Learning frameworks.

Summary

The MIC-730AI offers high performance for a wide range of AI and classic applications. Onboard interfaces, the PCI-expansion module and the "iDoor" extension slot allow easy customization and integration into a wide range of use-cases. Low power consumption in combination with passive cooling enable the use in industrial environments close to machines or inside vehicles.
The system can be used almost out of the box and supports short deployment cycles of AI applications.

Evotegra GmbH

Evotegra GmbH is a provider of customer specific AI solutions for image recognition. As a full stack solution provider we provide consulting, data, AI training and deep integration into custom environments. Our expertise is to optimize AI solutions on embedded platforms.

The hardware was provided by Advantech. Evotegra GmbH was not paid for the test.