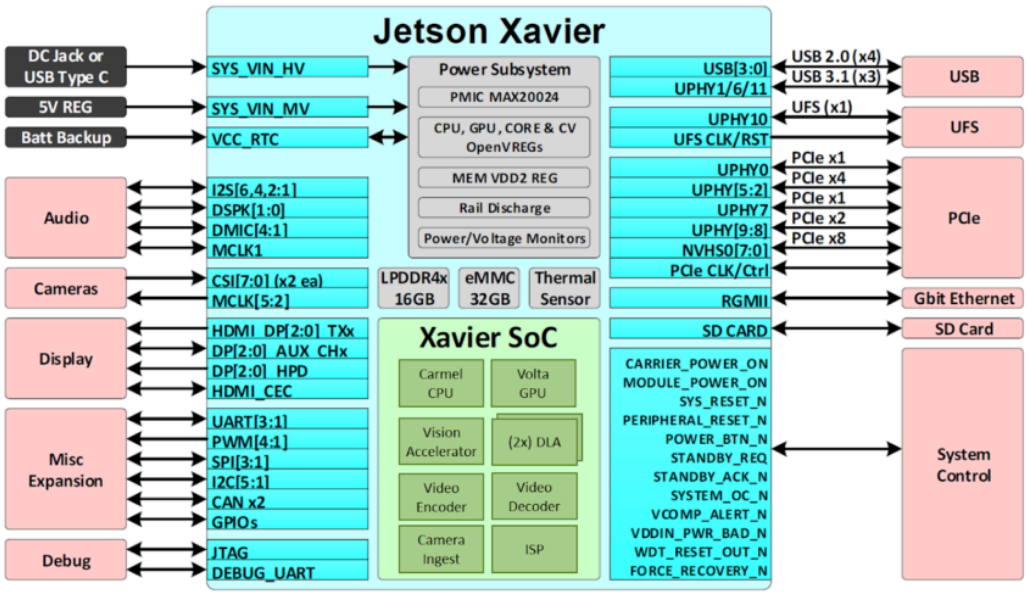
KI aus der Kiste

Das Advantech MIC 730AI ist ein robustes KI-Inferenzsystem, das auf dem NVIDIA Jetson® Xavier basiert. Mit 19x23x9cm hat das System die Abmessungen eines Mini-PCs. Das Hauptunterscheidungsmerkmal zu einem herkömmlichen Industrie-PCs ist die Tatsache, dass Xavier-Boards über eine integrierte GPU verfügen. Damit verfügt das System über ausreichend Leistung für die Verarbeitung anspruchsvoller KI-Anwendungen. Das Board ist unter anderem mit Video De- und Encoder sowie Deep Learning Beschleunigern (DLA) ausgestattet. Mit einer Betriebstemperatur von -10 bis 50°C, passiver Kühlung und geringem Stromverbrauch ist das System für den „Edge-Betrieb" ausgelegt. Edge-Systeme arbeiten in der Nähe der Sensoren, um die Übertragung großer Datenmengen über das Netzwerk zu vermeiden.



Hardware

Der Xavier im MIC-730AI ist in zwei Varianten erhältlich. Die kleinere Version verfügt über 8 GB RAM bei einem Speichertakt von 1333 MHz, 6 ARM sowie 364 Volta GPU / 48 Tensor-Kernen, während der große Bruder über 16 GB bei 2133 MHz Speichertakt und 8 ARM sowie 512 GPU / 64 Tensor-Kerne verfügt. Im Hinblick auf die effektive Leistung für KI-Anwendungen bietet die 16-GB-Version ungefähr doppelt so viel Rechenleistung wie die 8-GB-Version. Während die größere Version mehr Flexibilität in der Forschungs- und Entwicklungsphase bietet, sollte die kleinere Version bereits für die meisten industriellen Produktionsaufgaben ausreichen. Beide Versionen verfügen über 32 GB internen Speicher, der neben dem Betriebssystem ca. 16 GB freien Speicherplatz für zusätzliche Anwendungen bereit hält.

Anschrift:
EvoTegra GmbH
Kreuzwingert 11
D – 55296
Gau-Bischofsheim

Kontakt:
Tel.: +49 171 / 402 4242
E-Mail: info@evotegra.de
Homepage: www.evotegra.de

Gerichtsstand: Mainz
HR-Nr.: B 43519
USt.-ID Nr.: DE 217667197
Steuer-Nr.: 44/654/70280

Geschäftsführer:
Tobias Manthey

Bankverbindung:
Deutsche Bank 24 AG
IBAN DE46 5507 0024 0024 0820 00
BIC DEUTDEDBMAI

EvoTegra GmbH
Kreuzwingert 11 • D – 55296 Gau-Bischofsheim

EVOTEGRA
CONCEPTS & INTEGRATION

Konnektivität

Beide Versionen sind mit einem HDMI, 2x RJ45 GbE-Anschlüssen, 2x USB 2.0 sowie 2x USB 3.0-Anschlüssen an der Vorderseite ausgestattet. Für industrielle Anwendungen stehen 2 COM-Ports sowie 16 DI/DO-Anschlüsse zur Verfügung. Clever: Unter der Rückseite befinden sich weitere 3x USB 2.0-Anschlüsse auf der Platine sowie ein zusätzlicher COM-Port, eine 5V Stromversorgung sowie ein Nanosim-Slot.

Erweiterbarkeit

Das Gehäuse verfügt über Platz für ein 2,5-Zoll Laufwerk, während das Board über einen SATA-Anschluss sowie über einen M2 und einen Mini-PCI-Express-Steckplatz verfügt. Für die Unterstützung von 2 PCI-Express-Karten in voller Größe kann das Gehäuse mit dem Erweiterungsmodul MIC-75M20 erweitert werden. Clever: Über die „iDoor" Blende kann das MIC-730AI mit einer Vielzahl von Mini-PCIe-basierten Erweiterungen wie Wifi, LTE oder industriellen Feldbusadaptern erweitert werden.

Integration und Sicherheit

Das System ist wahlweise mit Montagewinkeln / Wandhalterung oder als Desktop-Version erhältlich. Um auf den Flash-Port zugreifen zu können, muss das Gehäuse geöffnet werden. Für sicherheitsrelevante Anwendungen kann das Gehäuse daher mit einem Siegel geschützt werden.

Test 1 Massenverarbeitung

Das erste Testszenario ist ein anspruchsvoller Anwendungsfall für die Massenverarbeitung mit geringer Latenzzeit. Typischerweise findet man diese Art der Verarbeitung z.B. in Fahrzeugen oder in einigen industriellen Anwendungsfällen.

Wir haben das folgende Setup gewählt:

- MIC-730AI
- 2 Basler ACE Industrie-Kameras mit einer Auflösung von 1920 x 1200 Pixel
- GPU-basiertes Stereokamerasystem für die 3D-Rekonstruktion

- Middleware-Software (Publish / Subscriber Message Bus)
- Visualisierung mit 5-10 Millionen 3D-Punkten / Sekunde
- Deep Learning-basierte Objekterkennung mit einer Auflösung von 960 x 600 Pixel
- Visualisierung der Objekterkennung

In dieser Konfiguration verarbeitet das System 625 Megabyte Daten pro Sekunde, wobei die verfügbare Speicherbandbreite ungefähr zu 40% ausgelastet wird. KI Algorithmen werden hauptsächlich auf der GPU verarbeitet. Daher beträgt die Auslastung aller CPU-Kerne nur 50%, während die GPU im Durchschnitt mit mehr als 80% ausgelastet ist.

Konfigurationen

Wir haben das MIC-730AI in 2 verschiedenen Leistungsmodi getestet. Durch die Anpassung der verwendeten CPU-Kerne sowie die Taktfrequenzen von CPU und GPU kann mit den Leistungsmodi der Stromverbrauch auf 10, 15 oder 30 Watt begrenzt werden.

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

Im MAXN-Power-Modus arbeiten sowohl CPU als auch GPU mit maximaler Geschwindigkeit. Dabei wird allerdings keine Obergrenze für den Stromverbrauch garantiert. Um das System auszulasten und die Wärmeableitung des passiven Kühlsystems zu validieren, haben wir in diesem Modus einen 24-Stunden-Belastungstest durchgeführt.
Der 30W ALL-Modus ist ein Modus, der den Stromverbrauch unter Verwendung aller 8 ARM-Kerne auf 30 Watt begrenzt.

Powermodus: MAXN
KI Ausführung: ~18ms
Stromverbrauch: ~50W
CPU-Last: ~50%

CPU Temperatur: 66,5°C
GPU-Last: ~80%
GPU Temperatur: 65,5°C
Platinentemperatur: 57,5°C
Kühler max. Temperatur: 49,2°C

Powermodus: 30WALL
KI Ausführung: ~35ms
Stromverbrauch: ~35W
CPU-Last: ~65%
CPU Temperatur: 53°C
GPU-Last: ~85%
GPU Temperatur: 53,5°C
Platinentemperatur: 48°C
Kühler max. Temperatur: 43,3°C

Testergebnisse
Während der Tests in beiden Modi verarbeitete das System insgesamt mehr als 100 Terabyte Daten. Bei Raumtemperatur erreichten sowohl die CPU als auch die GPU bis zu 66°C und erwärmten den passiven Kühler auf maximal 50°C. Dies lag jedoch weit unter der empfohlenen maximalen GPU-Temperatur von 88°C. Das System arbeitete zu jeder Zeit stabil.

Test 2 Anforderungsbasierte Verarbeitung

In diesem Szenario testen wir das System in einem typischen industriellen Anwendungsfall. Das System empfängt die Daten über die Netzwerkschnittstelle und das Ziel ist es, die Daten mit der geringstmöglichen Latenz zu verarbeiten.

- MIC-730AI
- Netzwerkschnittstelle (C++)
- Deep Learning-basierte Objekterkennung und -klassifizierung (1024x1024 Pixel)

Um Energie zu sparen, wird selbst im schnellsten Power-Modus (MAXN) der Takt für CPU und GPU bei geringer Last gedrosselt. Dadurch haben die ersten Bilder eine bis zu dreimal höhere Latenz, als die nachfolgenden Bilder. Während in den meisten Anwendungsfällen die Energiesparfunktionen unproblematisch sind, ist dies für die zeitkritische Verarbeitung kein erwünschtes Verhalten.
Die Lösung ist einen benutzerdefinierten Energiesparmodus zu erstellen, der den GPU- und CPU-Takt auf ihre maximale Frequenz begrenzt.

Powermodus: Benutzerdefiniert
KI Ausführungszeit: ~35ms
Stromverbrauch: ~30W
CPU-Auslastung: ~5%
CPU Temperatur: 40°C
GPU-Auslastung: ~10%
GPU Temperatur: 40.5°C
Platinentemperatur: 37°C
Kühler max. Temperatur: 34.4°C

Testergebnisse

Da das System zwischen der Verarbeitung der einzelnen Bilder Ruhephasen hat, wirkt sich die dauerhaft hohe Taktfrequenz von CPU und GPU nicht merklich auf die Temperatur aus.
Die Ausführungszeit beträgt etwa das 1,75-fache der Zeit, die ein auf Intel basierender Industrie-PC in Kombination mit einer NVIDIA RTX 2080Ti-GPU erreichen kann. Vergleicht man die technischen Daten beider Systeme, so verfügt der Industrie-PC im Vergleich zum MIC-730AI über etwa das 8-fache an Ressourcen. Die Tatsache, dass der 8-fache Vorteil bei den verfügbaren Ressourcen nur den 1,75-fachen Vorteil bei der tatsächlichen Verarbeitung ergibt deutet darauf hin, dass die Latenz hauptsächlich mit dem Speicherdurchsatz zusammenhängt.
Im Vergleich zum Industrie-PC hat der MIC-730AI jedoch nur 5-10% des Stromverbrauchs, keine beweglichen Teile und im Allgemeinen weit weniger Komponenten. Daher sind für den MIC-730AI geringere Wartungskosten zu erwarten.

Betriebssystem

Das System wird mit Linux Ubuntu 18.04 und dem „Jetpack" ausgeliefert. Das Jetpack enthält die NVIDIA-Bibliotheken und -Tools zur Beschleunigung von Deep Learning Anwendungen. Technisch bietet die Umgebung die gleichen Annehmlichkeiten wie ein GPU-basiertes PC-System. Die Bereitstellung von neuronalen Netzwerken, die auf einem PC oder in der Cloud trainiert wurden, ist daher unkompliziert. Obwohl es technisch möglich ist sogar neuronale Netze auf dem MIC-730AI zu trainieren, ist das System für grundsätzlich die Ausführung ausgelegt. Netzwerke können prinzipiell nativ mit Python ausgeführt werden. Um jedoch die beste Leistung zu erzielen, empfehlen wir C++ in Kombination mit Netzwerkoptimierung zu verwenden. Die NVIDIA-Netzwerkoptimierung TensorRT ist im „Jetpack" bereits vorinstalliert und kann Netzwerke aus verschiedenen Deep-Learning-Frameworks optimieren.

Zusammenfassung

Der MIC-730AI bietet eine hohe Leistung für eine Vielzahl von KI- und klassischen Anwendungen. Die verfügbaren Schnittstellen, das PCI-Erweiterungsmodul und der Erweiterungssteckplatz „iDoor" ermöglichen eine einfache Anpassung und Integration in eine Vielzahl von Anwendungsfällen. Geringer Stromverbrauch in Kombination mit passiver Kühlung ermöglichen den Einsatz in industriellen Umgebungen in der unmittelbarer Nähe von Maschinen oder in Fahrzeugen.
Das System kann ohne große Anpassungen verwendet werden und ermöglicht kurze Bereitstellungszyklen von KI-Anwendungen.

Evotegra GmbH

Die Evotegra GmbH ist Anbieter von kundenspezifischen AI-Lösungen zur Bilderkennung. Als Lösungsanbieter bieten wir Beratung, Datenmanagement sowie Integration in kundenspezifische Umgebungen. Unsere Expertise ist die Optimierung von KI-Lösungen auf Embedded-Plattformen.

Die Hardware wurde von Advantech zur Verfügung gestellt. Die Evotegra GmbH wurde für den Test nicht bezahlt.