aiWare™
Hardware innovations
for automotive AI

Case Study: silicon excellence + system & algorithm excellence = solution excellence

written by: Tony King-Smith

Abstract: When Nextchip set out to find the best partner to maximize the AI capabilities of the next generation of their award-winning Apache4 vision-based edge processor, they not only looked to established hardware IP leaders, but also they asked their most trusted customers. The result was a unique collaboration with an unlikely source of hardware IP: an innovative young European software and systems supplier with a passion for making AI hardware platforms for autonomous driving much better.

When disrupting a global market, together is better

To paraphrase the famous poet John Donne: no man – or company – is an island. The sheer complexity of the task to create safe and reliable yet cost-effective autonomous vehicles has led everyone in the automotive industry – from the largest OEMs and leading Tier1s, through to industry disrupters such as Tesla – to acknowledge that collaboration is essential. No one company, no matter how large or whatever the budget, has all the skills and know-how to create autonomous vehicles by themselves, nor the components and software needed to assemble and control them. Autonomous vehicles will be built through many collaborations and partnerships between companies whose skills complement one another, resulting in a more compelling product for their mutual customers.

In this white paper, we will explore the path taken by an innovative, well-established Korea-based specialist silicon and algorithm supplier to create something truly unique. We will see how by looking beyond the requirements of the hardware itself to the solutions their customers needed, Nextchip found in AImotive™ a unique and unexpected supplier of industry-leading hardware IP.

Building on strong foundations

Over the past 25 years, Nextchip has built an exceptional reputation for excellent hardware, algorithms and chip supply for embedded system. Their deep knowledge of image signal processing, computer vision algorithms, codecs and low-power, high-performance engineering know-how enabled them to create a series of successful chips targeting the fast-growing video and camera markets.

By focusing on imaging, Nextchip have created a series of highly specialized digital and analogue technologies, combined with well-engineered processing engines. These have been combined with advanced hardware IP from industry leaders such as Arm and Synopsys to create a wide range of highly optimized chips, ISP (Image Signal Processor) to mixed signal ASIC, transmission device and SoCs.

Several technologies stand out as key differentiators in Nextchip's capabilities:

ISP: the sophisticated capabilities of the Nextchip ISP (Image Signal Processor) have won them wide acclaim. Their ISPs have appeared both as dedicated ISP chips, and integrated into advanced SoCs

AHD: their unique analogue high-speed data transmission technology builds on Nextchip's expertise in vision systems, combined with advanced analogue and digital engineering capabilities

CV Algorithms: through their in-house development of advanced CV (Computer Vision) algorithms for challenging tasks such as ISP related algorithms like AE, AWB, HDR, LFM and ADAS related algorithms like pedestrian detection, vehicle detection, lane detection and moving object demonstrates their understanding of how to bring together advanced systems, hardware and software engineering

The culmination of their developments is Apache4: a complete IEP (Imaging Edge Processor) targeting primarily automotive ADAS applications. This award-winning design combines their advanced ISP with high performance Arm-based CPUs and powerful DSPs to deliver a compelling application platform.

The next step: delivering AI capabilities to cost-sensitive markets

When Nextchip saw the success and enthusiastic industry reception for Apache4, they knew they were on to a winner. They focused more of their broad engineering expertise to figuring out how to make the next generation better. How much more processing

power was needed? What sort of processors? And what was the application that was going to drive their market?

Almotive drives its cars daily on 3 continents

The answer to the last question was clear: AI. Almost every customer they visited wanted to know their plans to introduce powerful AI capabilities into their future products. But there were challenges: how much AI performance? What should the capabilities be? How do you balance high performance with flexibility and power consumption? Nextchip quickly realized they needed help to identify the right balance of performance and capabilities for their markets. They needed a partner who understood not only silicon, but could talk about the actual AI application itself to their customers. That was well beyond their capabilities.

The search begins

In order to answer some of these questions, Nextchip started by talking to all of their existing hardware IP suppliers. They all had IP products – how should they choose the best one when they didn't fully understand the application, nor AI?

When Nextchip spoke to their customers, they found that they had already spoken to many of the well-known hardware IP providers – and they all had the same concerns: Do these vendors really understand the specific challenges of automotive AI inference? Do they understand the challenges of executing demanding AI for HD camera-based computer vision? Are they focused on the inference market, or are they trying to address training as well for NNs? Indeed aren't they claiming their solution will do any form of AI, without really understanding what that means? Are they able to produce a solution fully optimized for automotive, vision-first safety-critical AI inference? And can I believe any of the numbers being claimed for realistic automotive applications?

> **AS THE NUMBER OF QUESTIONS GREW, THE TRUST IN THE ANSWERS DROPPED: TOO MANY ANSWERS WERE GENERIC, NOT UNDERSTANDING THE UNIQUE NEEDS OF AUTOMOTIVE.**

And too many numbers were being produced using benchmarks that were simply not relevant for automotive grade, low-latency camera-based systems. Nextchip and its customers began to realize that automotive vision-based AI using high-resolution image sensors is a highly specialized area that very few people understood. They needed more than hardware IP – they needed answers they could believe in.

After an extensive search, they found a unique company based in Europe that combined extensive know-how in AI and embedded software for automotive applications with a unique approach to hardware IP for tackling one of the most compute-intensive tasks for autonomous driving: DNNs (Deep Neural Networks) for real-time inference.

This company's answers to their questions combined, for the first time, an in-depth understanding of the challenges of delivering automotive grade camera-based AI systems for volume production, with industry-leading NN design capabilities honed solely for the automotive ADAS market. Furthermore, they had also designed hardware IP intended to solve what they also saw was a significant hole in the market for silicon platforms capable

of executing these demanding, high performance AI applications hour after hour, day after day, in extreme operating conditions meeting the most demanding automotive safety and reliability standards such as ISO26262.

That company was AImotive Kft.

When Nextchip started asking questions about AImotive to their industry colleagues and to their customers, they were surprised to hear how well known AImotive was, and how respected their automotive AI technologies had become.

They said: *"if you can deliver their capabilities in your chips, we will trust your chips because we know that this company understands the real complexities and challenges delivering camera-based automotive AI".*

The search was over – Nextchip had found its new strategic partner.

An unusual supplier of hardware IP: a software company!

AImotive Kft. – a fast-growing, highly successful autonomous driving software and systems company based in Budapest, Hungary – is a unique technology powerhouse. It brings together under one roof AI, NN, embedded hardware and software and simulation expertise to create a broad, highly integrated yet modular technology portfolio for autonomous vehicles. Their fleet of test vehicles is a regular sight on roads in the US, Europe and Japan, and their expertise is often sought after in conferences around the world dedicated to the latest in autonomous vehicles technologies. It is AImotive's expertise in all these areas, plus their strong focus on camera-based solutions, that made AImotive's aiWare™ NN acceleration hardware IP of great interest to Nextchip.

Executives explain details of AImotive's test fleet to Nextchip engineers

When Nextchip studied AImotive in detail, it found a number of key attributes that made it a highly desirable partner:

- Respected in the automotive AI industry in every major region
- Well-funded with world-class strategic and financial investors and strong leadership
- a strong, stable and highly qualified engineering team including a significant number of PhD-qualified AI algorithm experts

- In-depth knowledge of applying the latest AI and CV techniques to cameras for automotive applications in autonomous driving, with extensive know-how and patents in algorithms as well as NN design
- Significant experience building and testing a range of cars running their complete AI software stacks in four locations (US, Hungary, Tokyo and France)
- Hands-on experience porting NN and other AI algorithms to a range of silicon platforms including Intel, Nvidia, Renesas and Mediatek SoCs
- Hardware IP designed with the sole purpose of accelerating NNs for real-time low latency inference for automotive camera (and other high-resolution) sensors in an ASIL B and higher certifiable environment

This combination of skills was unique. Nextchip quickly realized that by being able to deliver AImotive's unique portfolio of expertise and technologies to its customers as part of its new Apache5 SoC, it had a unique and compelling product offering that automotive OEMs and Tier1s would find both innovative and refreshing. Their customers could engage with Nextchip together with their new partner AImotive in ways few other vendors could offer. AImotive spoke their customers' language!

Creating the best product

When Nextchip started specifying their next generation Apache5 IEP for automotive camera applications, they recognized that already car makers were moving away from the high end "revolutionary" L4/L5 fully autonomous vehicles to more "evolutionary" L2/L2+/L3 solutions. As a volume chip supplier to a wide range of automotive OEMs and Tier1s, Nextchip saw that high efficiency was essential when executing real-time algorithms for applications such as smart rear vision, valet parking or highway driving. Only by understanding the actual NNs used in automotive environments – not just public NN benchmarks developed many years ago – could Nextchip demonstrate that its highly optimized chip could do the job. AImotive answered that requirement by providing that expertise

alongside a recognized award-winning* industry-leading NN hardware acceleration IP core.

AImotive was able to engage with Nextchip's customers to help explain why the unique combination of features in the NN accelerator on Apache5 would mean their application could achieve high performance at low power consumption. Being able to explain to Nextchip's customers how vision-based NN algorithms can be optimized for Apache5 was key to winning new business. And by demonstrating that Nextchip's customers would have access to AImotive's special skills needed to help them port and optimize new custom NNs to Apache5, AImotive enabled Nextchip to offer a level of technical support not available from other chip vendors.

Efficiency: the key to success

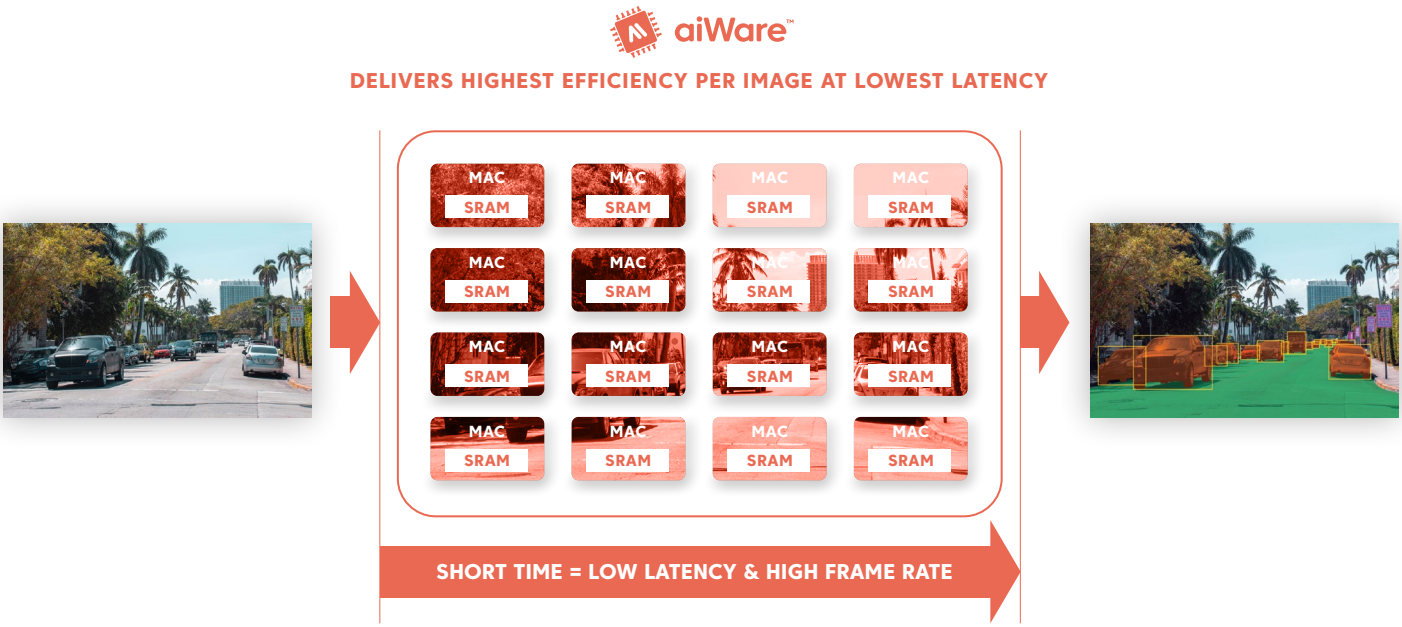
With any chip, the hardware resources are fixed. Unlike software, once the chip has been manufactured the functionality cannot be changed except by the software executing on it. Therefore, it is crucial that the hardware is sufficiently flexible to accommodate a wide range of applications, and to execute these at the highest possible efficiency.

Efficiency is all about what percentage of the hardware resources are being used to do useful work. This is vital for system integrators, since for more advanced silicon manufacturing processes such as that used by Nextchip for Apache5 suffer from a phenomenon known as "leakage". This means that even if a section of logic is doing nothing, it still consumes a significant amount of power as long as it is switched on. So if the hardware is consuming power, we need to make sure it is doing something useful for as much of the time as possible.

However getting answers about how efficient any hardware will be when executing your application is much harder than you think. System integrators usually select chips based on their claimed performance metrics, using well-known industry benchmarks. These are often designed to measure performance of specific features, which might not be relevant to your application. Not delivering the sustained performance claimed by these metrics

can waste valuable engineering effort, either during evaluation (for example analyzing results from the wrong benchmarks), or during design of the product itself (for example only discovering shortfalls in performance too late in the design process).

The aiWare3P NN accelerator used in Nextchip's Apache5 has been designed from the ground up for maximum efficiency for vision-based NN applications. It is particularly efficient for large input sizes needed for the latest generation of 1Mpixel–2Mpixel and higher resolution cameras being used for automotive applications. Indeed, in benchmarks published by AImotive, aiWare delivers up to 96% efficiency in sustained execution of some vision applications. This means no power is wasted, resulting in lower overall power consumption and higher sustained performance for day in, day out operation.



aiWare focuses on processing each image separately as quickly as possible (Batch=1)

The unique, highly specialized low-level hardware architecture of aiWare enables it to use less power than most other NN hardware accelerators. That is because it is designed only for vision inference applications in embedded environments, so no power or silicon area is wasted on expensive and complex CPUs, GPUs or DSPs that offer flexibility that will not be used in an end application.

Specialized vs general purpose processing

When designing an SoC for embedded imaging applications, Nextchip knew that power and cost were key factors. Since Apache5 would be used for very specific production embedded applications, there was no need to provide excessive programmability if that cost them performance.

Nextchip were delighted to see the flexibility offered by the aiWare solution. AImotive explained that almost any embedded deep CNN could be executed totally within the aiWare engine. In effect, it was every bit as flexible as other NN accelerators designed around more programmable engines such as DSPs or GPUs – but without the complexities of low level code optimization necessary to get any sort of reasonable efficiency from such engines.

The benefits of aiWare became even more clear when AImotive showed them comparative benchmarks of aiWare vs well-known GPUs. While aiWare scaled with high efficiency as the size of input sensor grew from 1M to 8M pixels, the GPU failed to execute well-known benchmarks due to memory constraints. This highlighted the value of a more specialized architecture designed to handle the high bandwidth data challenges of CNN execution for camera-based systems.

> **AIMOTIVE ALSO EXPLAINED THE FLEXIBILITY OF THE HARDWARE RESOURCES IN AIWARE TO HANDLE ALL SORTS OF ACTIVATION AND POOLING FUNCTIONS.**

Customers can even ask AImotive to support any unusual layer functionality requests, which could include extending the SDK to deliver additional layer functions, or to suggestions for restructuring the NN itself to deliver very similar or identical results using existing capabilities.

A specialized architecture does not mean fixed functionality – indeed for applications like DNN execution the unique architecture enables more flexibility than other NN accelerators, even those consuming 10x the power!

AI is about more than NNs

Nextchip's Apache5 SoC enables their customers to implement complete AI-based vision subsystems. Not all algorithms will be NNs – that's why the powerful CPU cluster and ISP on-chip enable a wide range of AI and CV algorithms to be implemented. Thanks to AImotive's broad knowledge of all types of AI algorithms, they were able to give Nextchip detailed advice and assistance to ensure that all parts of the Apache5 SoC would deliver the performance and capabilities needed for advanced imaging AI applications. This combination of skills all from one partner is rare in the hardware IP industry.

Understanding cameras

A number of key partners in Nextchip's ecosystem are imaging sensor providers. When working with these sensor vendors, a great deal of low level knowledge is required to ensure that every part of the system, for sensor to output, works under all operating conditions.

This automotive reversing camera from Kyocera features aiDrive™ NN algorithms for highest performance

Thanks to AImotive's extensive test vehicle program, they have gained significant knowledge of how sensors actually behave in real-world conditions. This expertise is fed into the development program for their aiDrive modular suite of software for autonomous driving, enabling them to adapt the software to almost any sensor and vehicle setup.

Beyond this, AImotive has also used its aiSim simulation and end-to-end design & validation suite of tools to enable them to accurately model sensor behavior, complementing real-life test data. This AI-based expertise enables AImotive to complement perfectly Nextchip's deep knowledge in imaging sensors and CV algorithms. The result is an unrivalled support capability for Nextchip's customers, enabling them to get their AI-based imaging products to market sooner, with higher performance and capabilities.

Scalability – making products future-proof

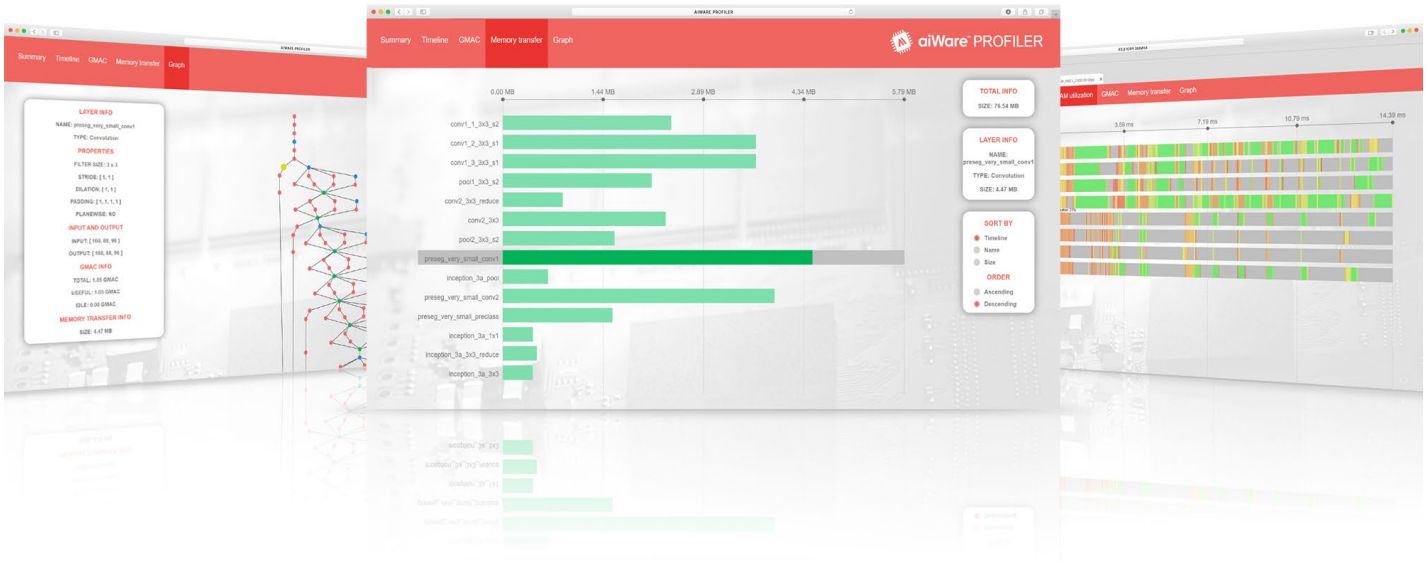
From the very beginning, every technology developed by AImotive is designed for scalability. This ability to scale from low-end L2 to high end L4/L5 using the same set of software and hardware modules is another unique asset that makes AImotive's technology portfolio so compelling for many in the autonomous driving ecosystem.

For Nextchip, this means that by investing in designs incorporating the aiWare scalable hardware IP, they can be confident that future generations of Apache SoCs will be able to deliver substantial increases in performance and capabilities for their NN acceleration.

Tools for optimizing NNs, not processors

The field of AI is changing constantly, with new development tools, network topologies, algorithmic techniques and more being discovered every week. It was therefore vital that Nextchip could offer its Apache5 customers the most flexible software environment to move their NN designs from the lab to the production hardware platform.

When providing tools for developers, many providers of NN acceleration IP deliver complex tools primarily to help developers optimize low level software executing on their processors. These tools, while sophisticated and impressive in their capabilities, often fall short in the key task: helping engineers move a NN from a server-based training environment to a tighly-constrained embedded hardware platform.



aiWare's advanced Performance Analysis tools help AI engineers optimize production NNs

Because AImotive has undertaken a series of projects to do just this task, they know that it will never become a simple "push button" operation. Indeed, this task requires extensive knowledge of NNs and how they behave. aiWare's Performance Analysis tools set a new benchmark for providing the information developers need to tackle this task. By providing an easy to use tab-based user interface, users can quickly take their compiled network and see immediately the performance on their target aiWare hardware. Indeed they don't even need the chip to be available! Thanks to the highly deterministic architecture of aiWare, the offline estimator can accurately estimate the GMACS (or TOPS) and external memory bandwidth per layer to within 10% of the final chip. And when the chip itself is available, the online tools read data from an extensive set of hardware registers to confirm exact NN execution metrics.

This offline capability is extremely powerful for OEMs and Tier1s. Now they can start to develop and refine their production NNs many months before actual hardware is available, saving crucial time in demanding production schedules.

For chip developers like Nextchip, these tools enable them to explore hardware configuration options such as aiWare core size, on-chip memory and internal configurations to find the best

balance of performance vs silicon area. This powerful capability helps Nextchip to ensure they are producing SoCs that can deliver the performance their customers expect. That means the commitment by their customers to the new chip is significantly de-risked, resulting in a closer relationship with their customers.

Unique capabilities, unique partnerships.

Apache5 will set a new standard for intelligence in imaging sensor pre-processors for automotive, and other markets. However, it is not just the chip itself that will attract attention. It is the unique partnership between two highly complementary automotive technology leaders, and the ability for Nextchip's customers to benefit from that combined expertise, that will set Apache5 apart as the leading IEP (Image Edge Processor) for future generations of intelligent vehicles.

Notes:
*aiWare won best vision processor award
at the Embedded Vision Summit 2018
www.embedded-vision.com/product-awards-2018-winners