Qualcomm Technologies AI Software

AI DEVELOPER TOOLS FOR QTI MOBILE AND DATA CENTER CHIPS

INTRODUCTION

Qualcomm has been developing AI hardware and software for nearly a decade and has recently expanded from the company's mobile chip space Snapdragon processors to enter the data center market with the Cloud AI100 platform. Starting from a collection of tools specific to various logic blocks such as CPU, GPU, and Hexagon processor, the company's AI software has now matured into an integrated stack for a fused AI hardware platform. This brief explores the latest enhancements of this AI software stack designed for the most widely deployed AI platform in the industry.

EFFICIENCY FOR DEVELOPERS OF EDGE AND DATACENTER AI

The newest Snapdragon 888 5G Mobile Platform features the company's 6th generation AI engine, a completely re-engineered Qualcomm® Hexagon™ 780 Processor that features a fused AI-accelerator architecture which brings the total Qualcomm AI Engine performance up to 26 TOPS, at 3X performance-per-watt improvement and 16X larger shared AI memory. The QTI Cloud AI100 took that power efficiency to the data center, as evidenced by the most recent MLPerf inference benchmark results, wherein QTI bested all submission by over 3X in performance/watt. Clearly, the QTI development team takes a comprehensive system-level view of performance and power efficiency.

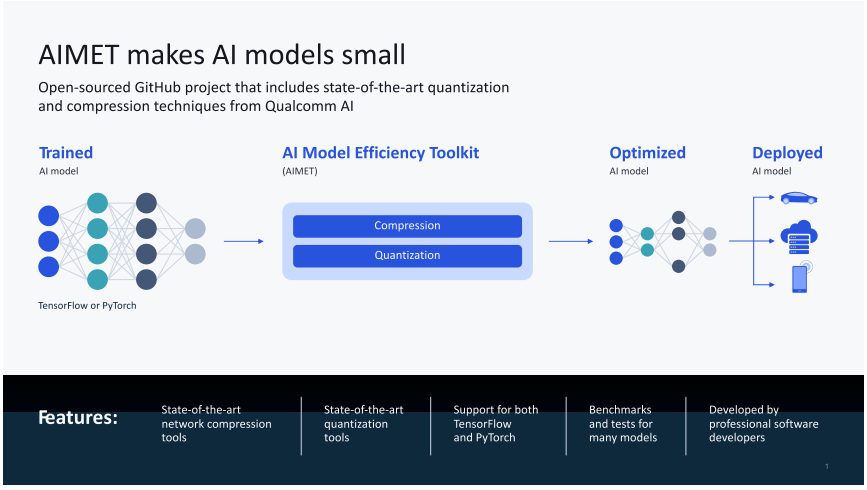
At the core of the 6th gen Qualcomm AI Engine is the Qualcomm Hexagon processor. This year, QTI introduced the Hexagon 780 processor. It's completely redesigned and features the biggest leap in architecture and performance in years, and named the *fused AI accelerator architecture*. In past generations, Snapdragon used separate scalar, vector, and tensor accelerators. For this new generation, QTI removed the physical distances between the accelerators and fused them together, so it's now on one big AI accelerator. QTI also added a dedicated large shared memory across the three different accelerators, so they can share and move data efficiently.

ENABLING A FULL SYSTEM APPROACH TO PERFORMANCE AND EFFICIENCY

Let's begin with the Qualcomm Neural Processing SDK, designed to help developers run one or more neural network models on Snapdragon mobile platforms, whether it is to run on the CPU, GPU or Hexagon processor. The SDK provides a high-level pipeline for machine learning, including everything a developer needs to go from a trained DNN model to an optimized network for inference processing, including:

- Android and Linux runtimes for neural network model execution
- Acceleration support for Qualcomm Hexagon processor, Adreno GPUs and Kryo CPUs
- Support for models in Caffe, Caffe2, ONNX, and TensorFlow formats
- APIs for controlling loading, execution and scheduling the runtimes
- Desktop tools for model conversion
- Performance benchmark for bottleneck identification
- Sample code and tutorials
- HTML Documentation

In addition to Qualcomm Neural Processing SDK, Qualcomm AI Engine direct, announced together with the new fused AI accelerator architecture on the Hexagon 780 processor, provides developers with access directly to the hardware, and not only for the Hexagon 780 processor, but also for the Adreno GPU and Kryo CPU. The Qualcomm AI Engine direct now has a unified AI API across the whole Snapdragon platform. In addition, this API is backward compatible and available on the previous 5th generation Qualcomm AI Engine. A developer or OEM can take advantage of this solution across Snapdragon platforms and leverage both the 5th and 6th generation AI Engine. QTI is focused on modularity and extensibility – expanding on user-defined operator concept to bring new capabilities for developers to create their own AI solutions, accelerated on Snapdragon.



In addition, developers can also use the open-source AI Model Efficiency Toolkit (AIMET), which provides advanced model quantization and compression techniques for trained neural network models. Obviously, reducing the size and precision of a neural network readies the application to run efficiently on a low-power, smaller mobile device. But the AIMET can also improve performance and power efficiency of running them on the larger CloudAI100 platform. AIMET also includes a model zoo, supporting image classification, semantic segmentation, object detection, super resolution, speech recognition and pose estimation. Having ready-made models speeds time to solution with tested, optimized models for common functions.

AIMET output is readied for deployment through the open-source TVM platform or the Qualcomm AI Engine direct. AI Engine Direct provides developers with direct access to the hardware, and not only for the Hexagon 780 processor but also for the Adreno GPU and Qualcomm Kryo CPU.

For autonomous vehicle functionality, the QTI solution is the Qualcomm® Snapdragon Ride™ Platform. Licensed developers can use it to develop solutions for the Qualcomm® Autonomy and Advanced Driver Assistance (ADAS) SDK, a C++ API containing numerous classes and methods for performing image processing functions specific to automotive.

And for robotic developers, the Qualcomm Robotics RB5 Platform offers a next-generation, solution that can be used to develop high-compute, artificial intelligence (AI)-enabled, low-power robots and drones for consumer, enterprise, defense, industrial, and professional service applications. Built around the Qualcomm QRB5165 processor, the Qualcomm Robotics RB5 platform includes many of the features found in the Snapdragon 865 Mobile Platform including its Hexagon Processor with specialized compute capabilities for AI, and 5G connectivity. This makes the platform well suited for adding AI operations such as on-device intelligence and computer vision algorithms to robotic and drone projects.

CONCLUSIONS

The Qualcomm Snapdragon and now the Cloud AI100 form one of the most widely deployed AI acceleration platforms in the industry. Consequently, developers need access to the tool chains that make it easy to develop, optimize, and deploy on these chipsets. The portfolio of tools we have investigated make it easy to port existing AI models, and to optimize those models to run as fast and efficiently as possible. Providing a unified development stack for both Mobile, Edge, and Cloud should also enable Qualcomm ecosystem partners to extend their applications from any deployment domain to another.

Combining the AI software stack with the portfolio of power-efficient Snapdragon and Cloud processors should help Qualcomm continue to lead the AI market in mobile, edge, and now establish a position in the cloud.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR: Karl Freund, Founder and Principal Analyst at Cambrian-AI Research

INQUIRIES:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This document was developed with QTI funding and support. Although the document may utilize publicly available material from various vendors, including QTI, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.