AI SoC Performance Profiling in RTL Design

by BC Lai, CTO and Roger Wang, VP of Sales and Marking

Introduction

AI SoCs (System-on-Chips) are becoming increasingly prevalent in many industries due to the high demand for efficient and accurate computation for AI applications. AI applications would usually require intensive computation as well as frequent data accesses. In a computation intensive system, performance is usually bottlenecked by the incapable processing engines or insufficient computation units. Deploying more processing engines and adding customized processing units could effectively boost the performance. In a memory intensive system, performance would be constrained by the long latency memory accesses of large volume of data. In this system, adding more processing units does not help. Designers should focus on the refinement of the data management policy, memory accesses, data layout, or system interconnect. There is also a chance that the characteristics of an AI application could switch between these two types. In this case, it is more difficult to optimize the system performance. Designers need to implement a more flexible/reconfigurable architecture to adapt to the run-time characteristics. To attain a high performance and efficient design, it requires in-depth understanding of the system performance and smart data/computation management.

These AI SoCs usually incorporate various components such as CPU, GPU, and NPU (Neural-network Processing Unit) to perform AI tasks at faster speed than traditional CPUs. However, the performance of AI SoCs can vary depending on the application, and thus, it is important to measure their performance accurately. In this whitepaper, we will introduce iPROfiler, an AI SoC Performance Profiler, to characterize the performance of AI SoCs at front-end design stage. It helps to identify the bottleneck, pinpoint the design issues, suggest proper architecture refinement, and verify the final system performance.

What is iPROfiler?

iPROfiler is a tool that allows engineers to measure the performance of AI SoCs for various AI applications at SoC front-end design stage. It provides a comprehensive analysis of the SoC's performance, including throughput, latency, transaction monitor, memory utilization and data reuse. All these are key factors in determining the efficiency of the SoC for a given task. Additionally, iPROfiler provides a detail breakdown of the SoC's performance on individual system components, such as CPU, GPU, NPU, Memory, System Interconnect, and other IP components.

How Does iPROfiler Work?

iPROfiler provides a clean RTL instance called Performance Measurement Tap (PMT) module which can be easily plugged-in or removed from an RTL design. PMT module collects SoC transaction data and sends them to the iPROfiler engine on an on-promise server or a cloud server. iPROfiler enables designers to set up analysis criteria that can generate benchmarks. These benchmarks are designed to evaluate the SoC's performance for specific AI tasks, such as object detection or speech recognition. iPROfiler measures not only the processing throughput and cycle (latency) breakdowns to perform these tasks, it also quantizes the memory and data reuse features of the target SoC system . The comprehensive analysis on individual system components further enables engineers to identify bottlenecks and optimize the system for the specific operations of applications.

Features of iPROfiler

Architecture Verification

Architecture Design Exploration: By benchmarking the performance of each IP components, iPROfiler helps designers select appropriate components, interfaces, and interconnects to meet the specifications. Some widely adopted examples are that iPROfiler provides the performance analysis information for designers to make proper decision on system interconnect bandwidth, cross protocol interconnect routing performance impact, and on-chip memory vs. off-chip memory size allocation.

IP Performance Measurement: Many IPs may not meet the proclaimed performance. It is hard to believe the specification on a datasheet without a test chip evaluation. It would be too late to find out when the SoC tape-out. iPROfiler helps evaluate the IP components in RTL design and could save design and turn-around time as well as the associated design/verification cost.

Performance Analysis

Port-to-Port Transaction Analysis (P2P-TA): It is critical to analyze the communication between different blocks or modules in a design. P2P-TA enables an approach to capture the flow of data or control signals between ports of a group of modules and to analyze the interaction (read/write) between them.

Cross-Protocol Analysis (CPA): CPA enables analysis of the interactions between different protocols used in an SoC system. iPROfiler supports AMBA and its subset protocols. However, if a system uses a proprietary interconnection protocol, PMT module can be easily reconfigured to support the proprietary protocol based on customer requests.

Memory Bandwidth Utilization: memory bandwidth utilization refers to the efficiency with which a system accesses and utilizes memory resources. Memory bandwidth utilization is an important consideration in RTL (Register Transfer Level) design because it directly affects the performance of the system. If memory bandwidth utilization is high, data can be transferred quickly and efficiently, resulting in faster processing and better system performance. On the other hand, if memory bandwidth utilization is low, data transfer rates may be slow, causing the system to operate at a lower speed and potentially leading to performance issues. iPROfiler helps SoC designers optimizing memory bandwidth utilization to ensure their design operate efficiently and effectively.

Data Reuse: Data reuse is essential for optimizing the performance and power efficiency of AI SoC designs. iPROfiler helps designer to design a good data reuse mechanism which can reduce the number of off-chip memory accesses that can improve performance, reduce power consumption, improve scalability and optimized resource utilization.

Data Reuse Distance: Data reuse distance is a metric used in AI SoC design to measure the number of memory accesses between two consecutive uses of the same data. It represents the distance in memory between two accesses of the same data element. iPROfiler helps designer to identify which data elements are accessed frequently and which ones are rarely utilized. This information can be used to optimize the layout of data in memory, such as grouping frequently accessed data together in cache or on-chip memory to reduce the number of off-chip memory accesses required. iPROfiler also helps identify opportunities for optimizing data access patterns, such as by reordering operations to minimize the distance between two consecutive uses the number of off-chip memory accesses and improve performance and power efficiency.

Verification and Debug

Transaction Tracking: iPROfiler monitors and analyzes the transactions that occur within the system. iPROfiler transaction tracking involves recording information

about each transaction, such as the source and destination addresses, the type of transaction (e.g., read or write), and the timing information (e.g., start and end times). It can then be used to analyze the performance of the system, identify bottlenecks, and optimize the design for improved performance and power efficiency. iPROfiler can also identify the sources of errors or unexpected behavior within the system which help designers to identify the transactions that are causing issues and take corrective action.

Anomaly Detection: The goal of anomaly detection of iPROfiler is to identify and resolve issues that may impact the performance or reliability of the system. iPROfiler identify unusual or unexpected behavior in the design during RTL phase. By detecting and resolving issues early in the design process, designers can reduce the risk or costly errors and improve the quality of the final product.

Data Access Report: Data movement is critical for achieving high performance and efficient use of hardware resources. iPROfiler provides a Data Access Report (DAR) that helps designers analyze how data is accessed and moved within the design, which can help identify potential bottlenecks and optimize the design for better performance.

Benefits of iPROfiler

iPROfiler provides numerous benefits to design and verification engineers who are developing AI applications. It enables accurate measurement of the performance of the AI SoC and identifies factors for optimization. Additionally, it can facilitate the close collaboration from AI model engineers to software and SoC hardware designers, by quantitively comparing the performance of different on-chip and off-chip memory features and the impacts of data allocation strategies. iPROfiler also helps to reduce the time and cost of development by identifying performance issues early in the RTL design stage.

Case Study: Using iPROfiler to identify the bottleneck of an AI SoC

To demonstrate the effectiveness of iPROfiler, we conducted a case study on a computation intensive AI SoC. This System on Chip has 10 IPs which include CPU, NPU, SRAM, DRAM controller, AXI/AHB, I/Os, and other IP components. The simulation of the key application kernel takes 100K cycles to complete. Thousands of waveforms, and signals that hard to identify which modules are performance factor and need optimization. iPROfiler do the following performance profiling.

1. Identify the bottlenecks

iPROfiler clearly shows that out of 10 IPs, two IPs (CPU and NPU) takes up 10% (10K cycles) of the execution cycles. 80% (80K cycles) of cycles are on memory (SRAM and DRAM) accesses. This conclude the current design is a memory intensive system. Optimizing computation is not the focus at this point.

2. Enhance Data Management (capture data reuse)
   Increase data reuse is the key to system performance. iPROfiler identifies that out of 90% of the data accesses are on only 5% of the data, which means these 5% of the data will be reused frequently. However, the current utilization of SRAM does not capture these reuses. iPROfiler feedbacks the key data addresses to the software programmers, and programmers can adjust the program to allocate the most valuable (most frequently reused) data on SRAM, and significantly reduce the unnecessary DRAM accesses. The overall system performance is also improved.

Conclusion

In conclusion, iPROfiler is an essential tool for engineers who are developing AI SoCs, or a data driven SoCs. It provides comprehensive analysis of the SoC's performance and enables engineers to optimize the system for their specific applications. With the increasing demand for efficient and accurate AI computation, iPROfiler is becoming an essential tool for SOC design engineers in many AI applications.