Accelerating Performance, Power and Functional Validation of Computer Vision Use cases on next generation Edge Inferencing Products

Yoga Priya Vadivelu, Arpan Shah, Deepinder Singh Mohoora, Ullas, Praveen Buddireddy
yoga.priya.vadivelu@intel.com, arpan.shah@intel.com, deepinder.singh.mohoora@intel.com, ullas@intel.com
Intel Technology India Pvt Ltd
#23-56P, Outer Ring Road Devarabeesanahalli, Varthur Hobli Bengaluru, Karnataka 560103

2022
DESIGN AND VERIFICATION™
DVCON
CONFERENCE AND EXHIBITION
UNITED STATES
VIRTUAL
FEBRUARY 28 - MARCH 3, 2022

Introduction

AI Edge inferencing products are powering a lot of visual and audio intelligence which includes smart cameras, virtual assistant, intelligent drones and autonomous driving. Vision Subsystem is one of the key subsystems for Edge Inferencing AI products to run complex use cases for Neural Networks and Compute Algorithms. Along with multiple camera inputs, multi-stream Media encode/decode capability, graphics processing, dual display support, the bring up of end-to-end functional use cases in a pre-silicon environment poses several challenges. Performance and Power analysis of these subsystems is a key aspect to ensure right architectural and design tweaks. Pre-silicon Emulation is a feasible platform to validate complex end to end firmware based neural network use cases such as RESNET, Mobile-NET, Yolo for object detection, image classification and tracking. The proposed methodology leverages IP/Subsystem environment at SOC level to generate not only the testbench but test vectors as well. Performance validation framework utilizes the output waveforms and run logs of the performance tests for assessing the system level performance metrics closer to silicon accuracy and this helps for left shifting by finding critical bugs at Pre-silicon validation. The environment is leveraged for the third validation vector – estimation of average and peak power consumption as well.

Problem Statement

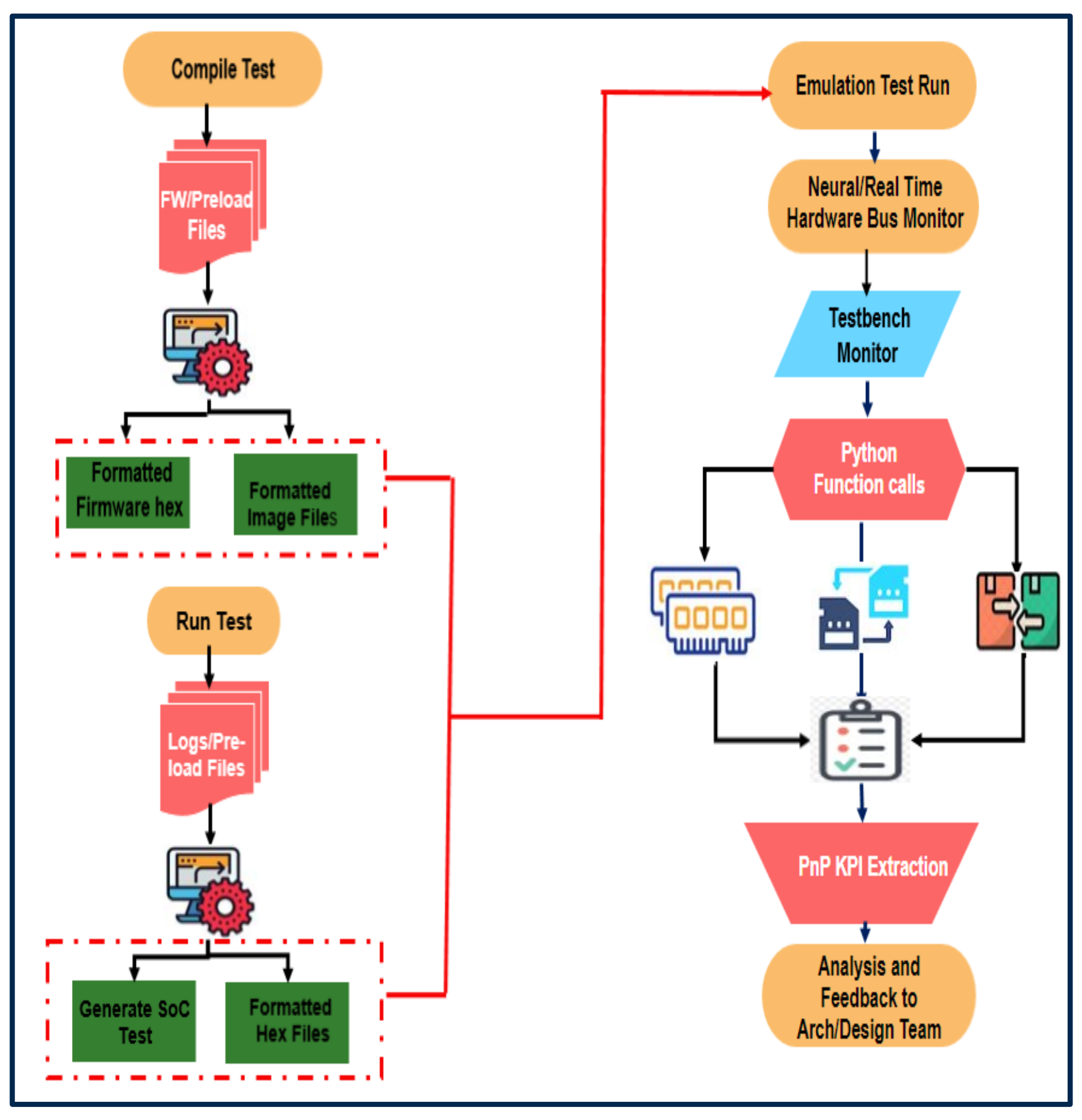
List of key problems and challenges involved in validation of Edge Inferencing SOC

i. Need for a methodology and validation platform which can capture activity profiles close to real world product scenarios involving several high bandwidth multimedia traffic initiators over multiple frames.
ii. Lack of a common platform to cover all 3 vectors – Functional, Performance and Power. Different validation platforms, test vectors and methodologies make it hard to infer consistent information which can lead to pin-pointed architectural decisions.
iii. Accurate performance and power modelling needs a concurrent system level platform to compliment IP and Sub-system level System-C and Simulation based platforms.
iv. Faster turn-around time for architecture & design feedback: Need to have early architecture and design feedback to avoid costly design re-spins.
v. Lack of SOC testbench infrastructure to run both Network path trace and Firmware based testcases.

Methodology

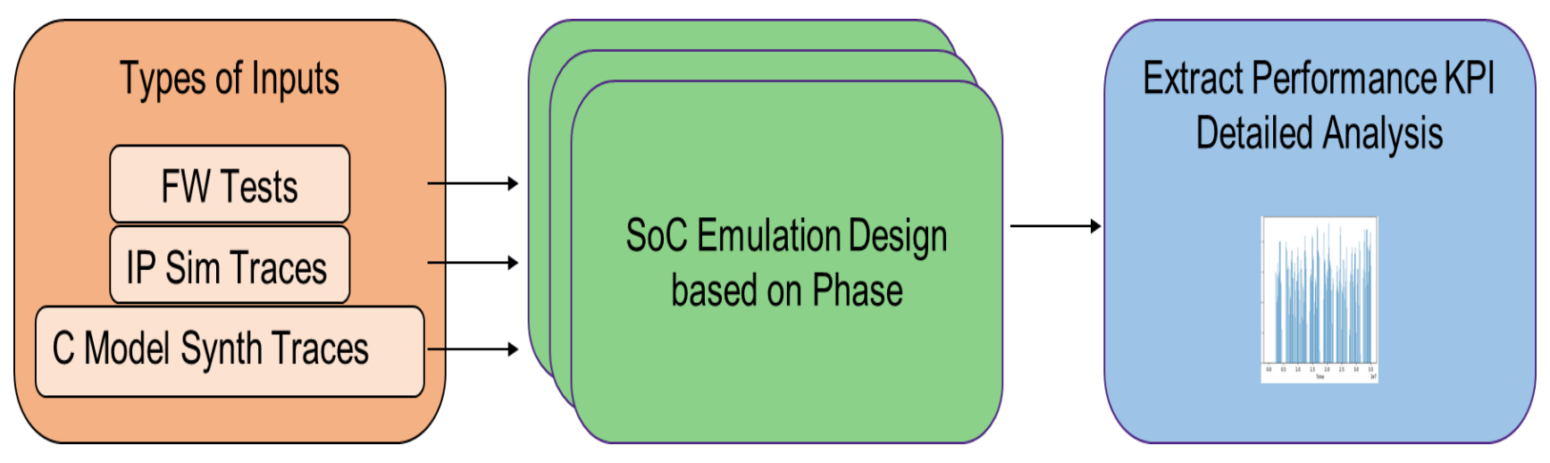
A. Use case Validation

i. In IP/Subsystem emulation environment, firmware images and image file inputs for each scenario is generated.
ii. Test sequences and testbench collaterals are extracted by post processing the logs of Subsystem and IP emulation setup and directed core level test sequences are created.
iii. The proposed methodology overcomes challenges faced in creating C test cases at SOC level by automating subsystem level setup and making use of firmware files, image files and the required test collaterals that are needed for SOC emulation testbench, thus saving significant effort by avoiding re-development of SOC level test scenarios.
iv. Example for Resnet network testcase, we generate firmware files, raw input image files and SOC test case configuration from IP/Subsystem testbench and these inputs are provided to SOC emulation testbench where the hardware bus monitors, and Python functions are integrated to monitor complex transactions and to do data integrity checks.



B. Power and Performance Validation

•Performance Analysis:
We have formulated a three-phase methodology for performance analysis

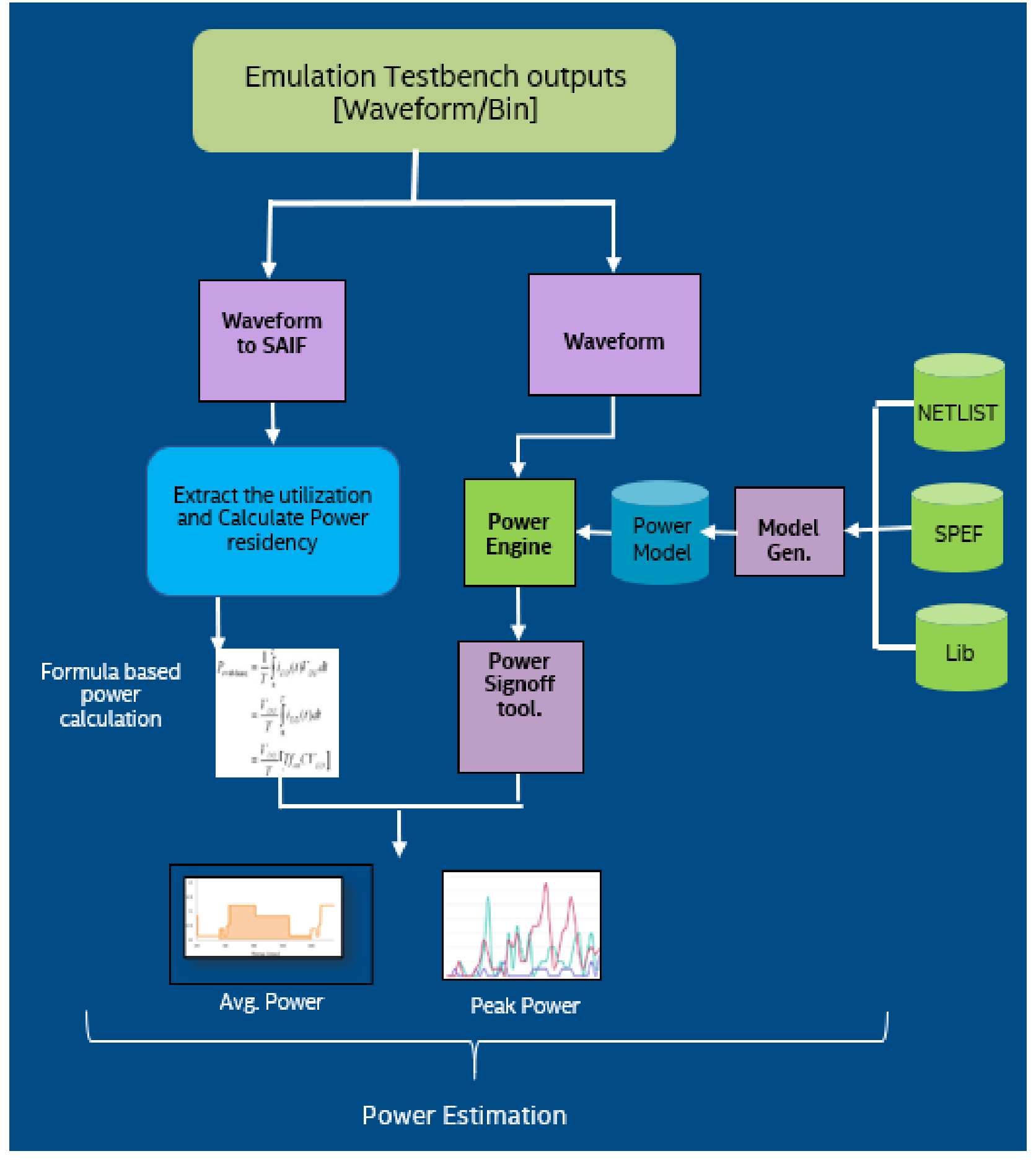

Types of Inputs
FW Tests
IP Sim Traces
C Model Synth Traces
SoC Emulation Design based on Phase
Extract Performance KPI Detailed Analysis

i. First phase uses transaction traces from an IP/Subsystem and plugs it into SOC infrastructure.
ii. Second phase uses Firmware based scenarios from IP/Subsystem and ports it to SOC level.
iii. Third phase is a unique hybrid methodology which can use Network traces and firmware tests together to generate a multi-frame activity profile comprising complex test scenario using firmware for all multimedia initiators concurrently.
iv. Key Performance Indicators (KPI) extraction and graphical representation of performance numbers is done using python scripts

Methodology

- **Power Estimation:**

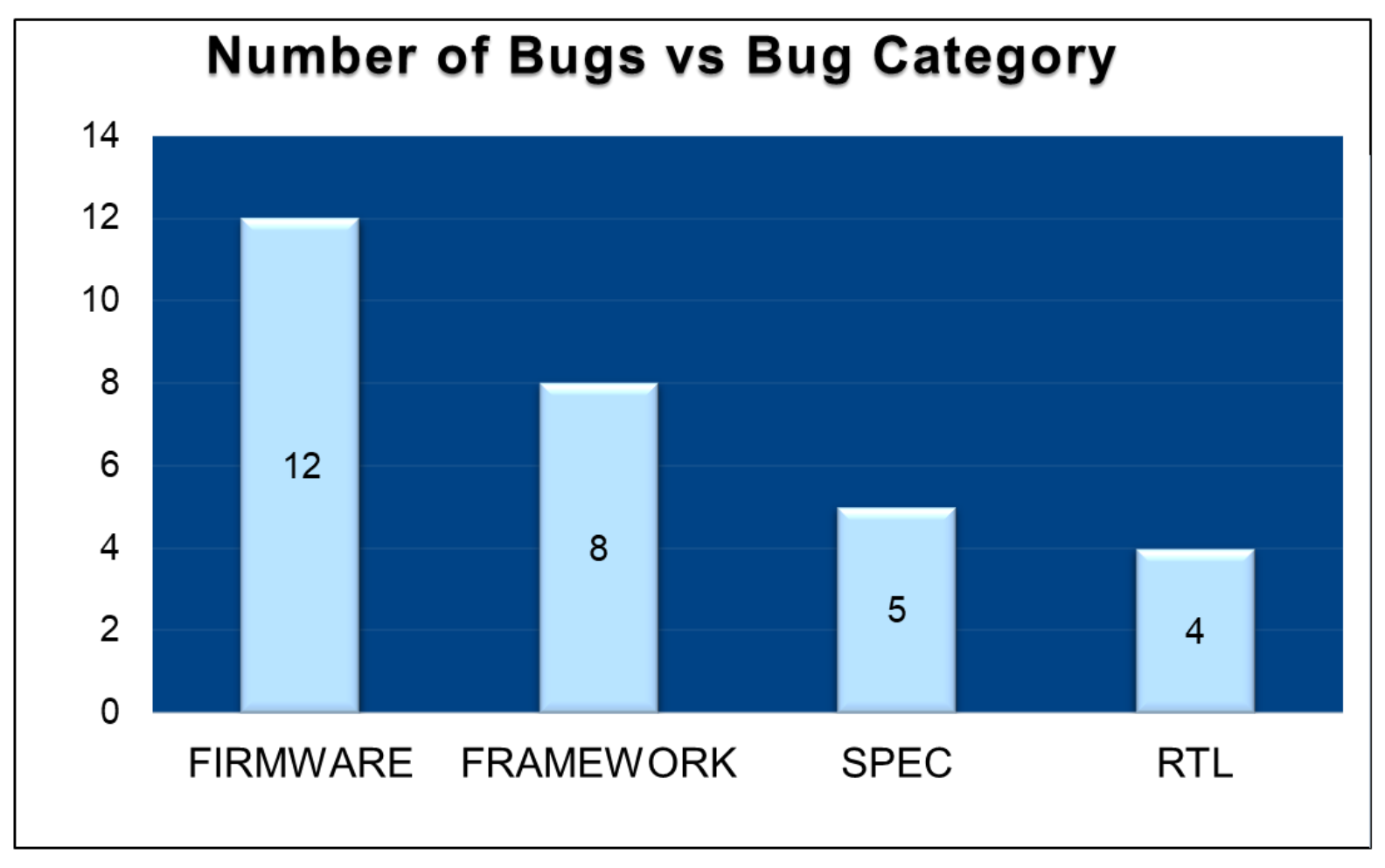
i. One of the key differentiating features of a product is to not only meet the performance targets but also to achieve those targets within power budget. Estimating power at early stage for concurrent scenarios is the key ability to estimate power at faster pace and increasing the scope of scenarios that can be analyzed early in the product design phase.
ii. The switching activity of the testcase is extracted out from the waveform database by executing end-to-end neural network algorithm of interest.
iii. Based on the activity, RTL SAIF files are generated. For more accurate power estimation, we have a methodology to generate Gate level SAIF. Average and peak power are estimated using SAIF and power model constructed by power engine tool.


Emulation Testbench outputs [Waveform/Bin]
Waveform to SAIF
Waveform
Extract the utilization and Calculate Power residency
Power Engine
Power Model
Model Gen.
NETLIST
SPEF
Lib
Formula based power calculation
Power Signoff tool.
Avg. Power
Peak Power
Power Estimation

Results

A. Use case Validation

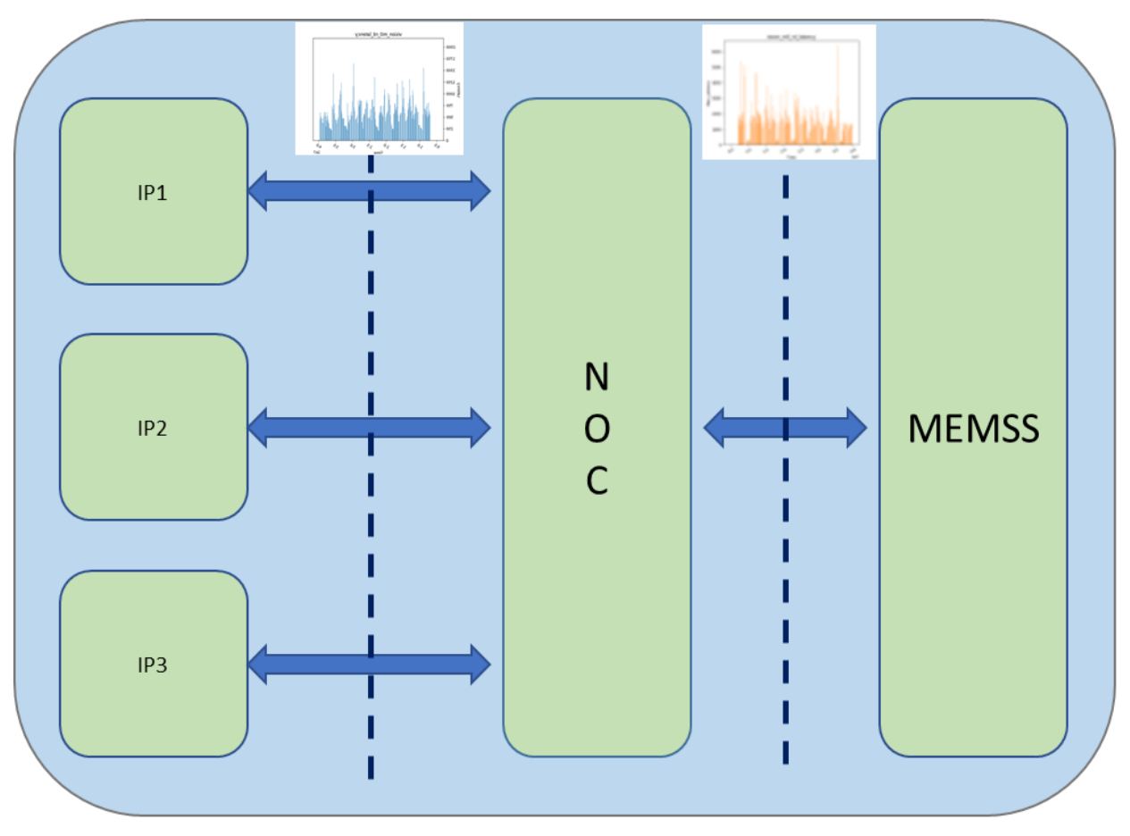
i. Faster bring-up of VPU (Vision processing Unit) core boot tests, DMA and Image Signal Processing test cases within 4 weeks of delivery time using the methodology which does automation and test collateral porting from Subsystem to SOC.
ii. Significant Left Shift by executing Neural network use cases before Tape-out. Neural network firmware-based tests had execution time of around two days in SOC simulation whereas in emulation it takes approximately 90 minutes thus achieving faster execution time and it provided ample window to deliver performance feedback to architecture and design team.
iii. 60+ test cases were validated and 30+ critical functional and performance bugs are found in NOC, DDR and VPU RTL.

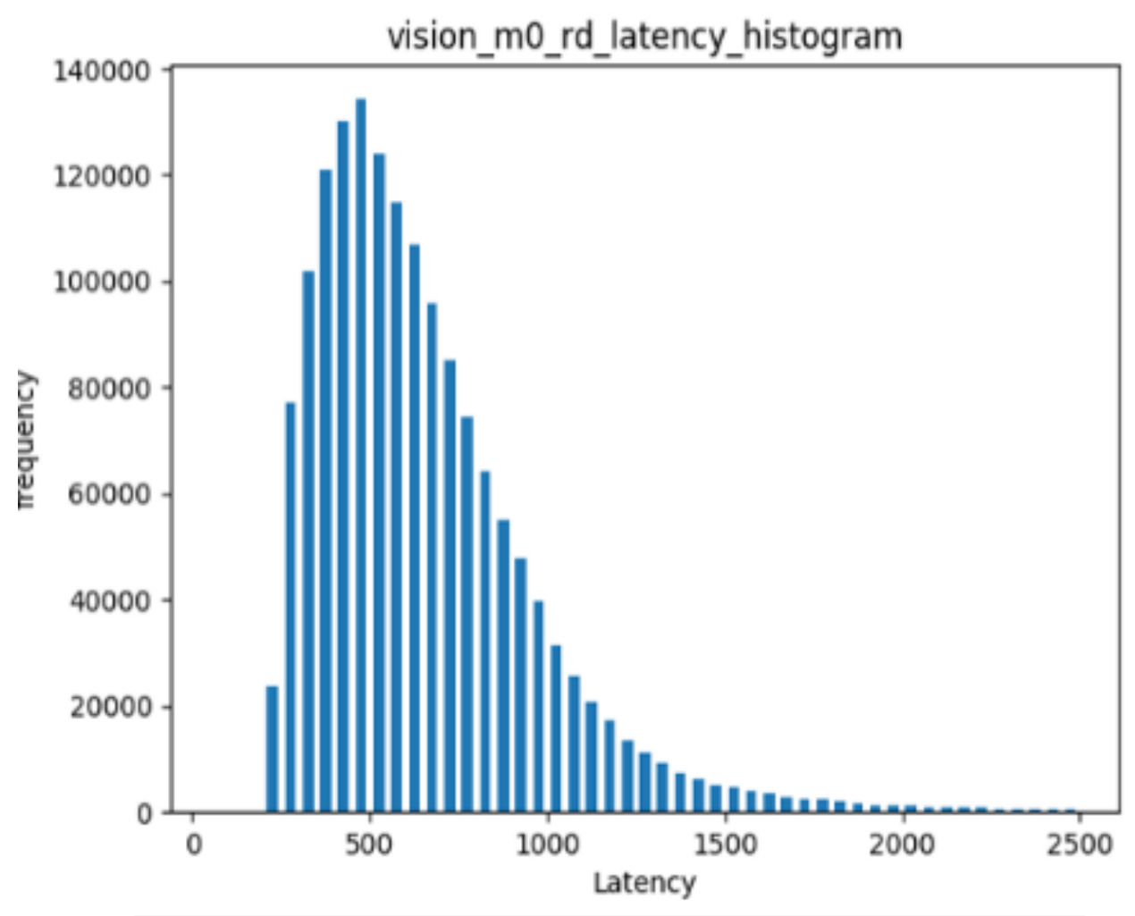


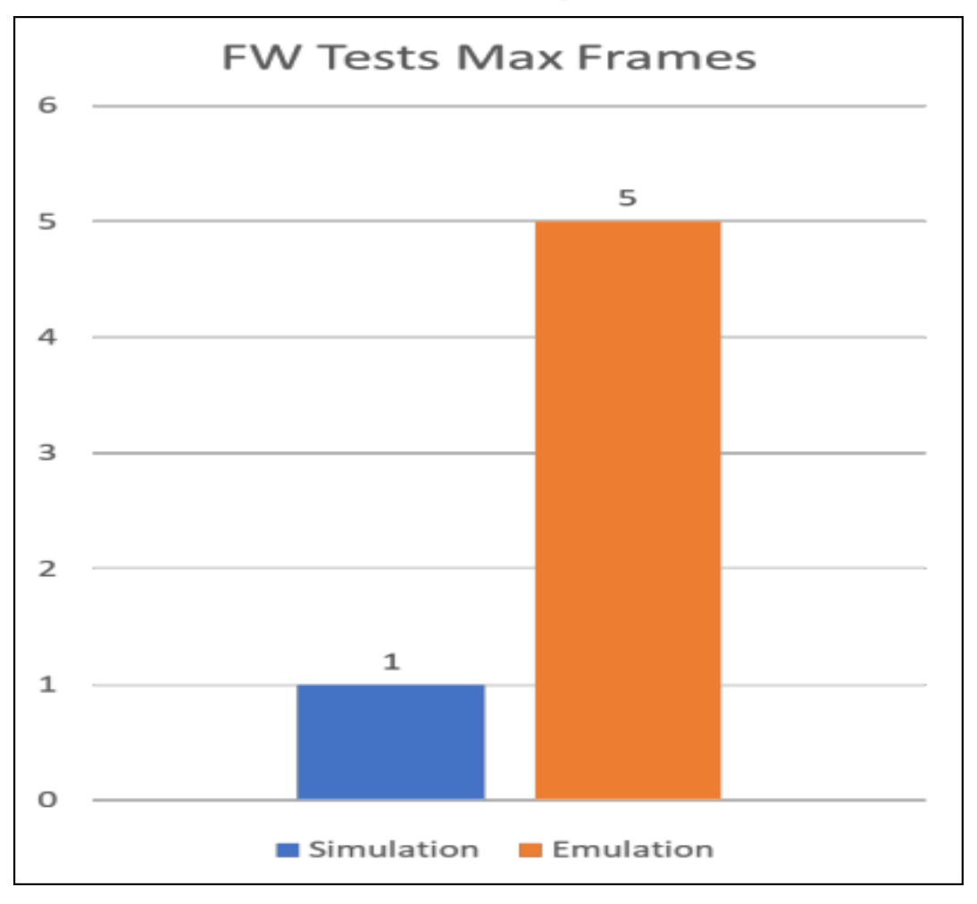
Results

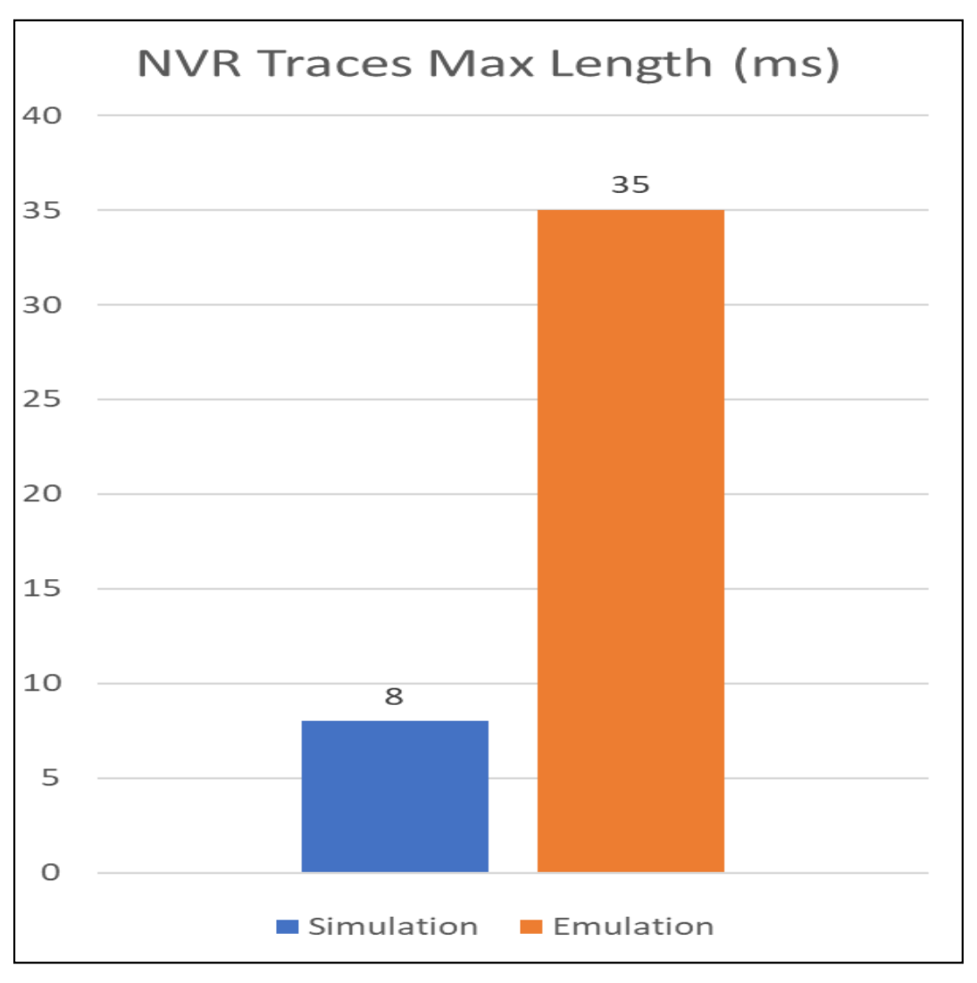
B. Power and Performance Validation

i. Performance KPIs are extracted and automated graphs for Bandwidth, Transaction Latencies and Outstanding transactions in NOC are created to infer key reconfigurations in the Architecture.









ii. Average and peak power estimates for critical IPs were estimated to meet KPI and thermal goals of the product. Table 1 shows the power estimation done using the methodology described in earlier sections. PoE is about 125x faster than the traditional Power analysis tool flow to estimate the power.

	Avg power		Peak power	
	PoE (power estimation using Emulation)	Power analysis Tool	PoE	Power Analysis Tool
IP-A	17.3mw	17.3mw	53mW	51.2mW
IP-B	272uW	297uW	734uW	705uW
IP-C	897uW	878uW	1.24mW	1.22mW
IP-D	12.3mW	12.3mw	734uW	705uW
	3.5hrs	~2days	1 hour	~2days

Conclusion

This methodology enabled us to bring-up and analyze all critical Neural network use cases along with other multimedia traffic initiators like Camera, Display, Graphics and Encoder/Decoder within the design freeze timelines. By estimating Power and Performance KPIs in pre-silicon stage, we left shifted all three validation vectors and influenced key architectural decisions.