ML Inference Test

(Processing Unit Test Comparison for ML Operations)

RK3588, CM3588, Nvidia RTX-2060, 3050

Table of Contents:

Introduction

This document evaluates the efficiency of Swiss Vault's Edge-Computing hardware solution, leveraging task-oriented chips like **RK3588 and CM3588** with ARM-based NPUs for ML-AI model inference tasks configured to run on VaultFS. Swiss Vault's solution provides energy-efficient compute support for inferencing pre-trained ML models, achieving performance levels competitive with **Nvidia's RTX 2060 and RTX 3050 GPUs**. The main goal of this comparison is to assess Total Cost of Ownership (TCO), energy consumption, and performance overhead. Key findings are summarized in Table 1.

Hardware Details

Radxa Rock 5B powered by RK3588

High-performing multi-functional ARM based SBC (Single Board Computer) powered by the Rockchip RK3588 SoC. It uses a Quad Arm Cortex A76 CPU, Quad Arm DynamIQ Cortex A55, Arm Mali G610MC4 GPU. It also encloses an NPU with 6 TOPS, which supports TensorFlow, Pytorch, Caffe, ONNX, capable to perform inference tasks by using RKNN-toolkit-lite-2 and RKNN SDK.

The net energy consumption for the RK3588 while running the below tests stayed between **3-5W**. For more information, refer to: **RKNN-DataSheet**.[1]

FriendlyElec CM3588

High-performing computing module designed by FriendlyElec team, which uses Rockchip RK3588 as the main control processor equipped with 4/8/16 GB LPDDR4x memory and 0/64 GB eMMC flash memory. Follows the same RK3588 support for SDK and NPU. Additionally, this supports 4 100-pin-board-to-board connectors (4x M.2 Key-M, i.e. PCIe 3.0x1, NVMe SSD). For more information, refer to: **CM3588-Product-Sheet**.

[1] The tests have been performed with 8 GB and 32 GB versions of the RK3588s (See table-1) .

Nvidia RTX 2060 & 3050

The NVIDIA GeForce RTX 2060 and RTX 3050 are mid-range graphics cards from NVIDIA. The 20-series is based on the Turing architecture while the 30-series is based on Ampere Architecture. The RTX 2060 features 1920 CUDA cores and RTX 3050 features 2304 CUDA cores, both capable of delivering robust performance for gaming and content creation activities. Both are equipped with 6 GB of GDDR6 memory in this use-case, the RTX 2060 supports a 192-bit memory interface width, providing ample bandwidth for smooth gaming experiences at high resolutions while RTX 3050 supports a 96-bit memory interface width. The RTX 2060 has an average power consumption of **160-180W,** while the RTX 3050 consumes **70W**(Graphics Card Power Consumption). For more information refer to : **Nvidia Spec-Sheet**.

AMD-V2000 (extra-just for reference)

The AMD V2718 is a SoC, supporting integration in AI engines, programmable FPGA fabric, and ARM Cortex-A72 and Cortex-R5 cores for high-performance computing and real-time control. Enhanced with the GFX90c integrated GPU, it offers graphical processing capabilities which support AI and edge-computing tasks. The GFX90c iGPU supports up to 8 compute units, providing robust performance for graphics-intensive applications. The iGPU requires **25W** to run, equipped along with DDR4/LPDDR4 memory support and PCIe Gen4 for high-speed data transfer in V2718. For more information, refer to the **AMD V2000 Product Sheet**.

Hardware Comparison

RK3588s use NPU (Neural Processing Unit) while Nvidia RTX 2060, RTX 3050 and AMD GFX90c are GPUs. The Rockchip RK3588 demonstrates a significant advantage over the Nvidia RTX 2060, RTX 3050, and AMD GFX90C-V2000 series GPUs, particularly for AI model inference tasks in a distributed environment combined with its superior energy efficiency and cost-effectiveness. This makes RK3588 an appealing choice for AI inference applications where value and performance are critical considerations.

In terms of power consumption, the RK3588 operates at approximately 3-5 watts per chip, which translates to low operating expenses (OPEX). Additionally, there is no capital expenditure (CAPEX) associated with the RK3588 as it is included with Swiss Vault's Edge Storage devices. This makes the RK3588 a cost-efficient solution compared to Nvidia and AMD GPUs, which typically incur both CAPEX and higher OPEX due to greater energy demands.

Furthermore, the RK3588 can be integrated into SwissVault's VaultFS architecture, enabling multiple units to be pooled together within a network. This setup allows for parallel utilization of their

computational capabilities, creating a highly scalable and compute-intensive environment. This flexibility in configuration provides notable improvements in inference times, as demonstrated by the data shown in the table.

Experiment

To assess the inferencing capabilities of various hardware configurations, including the Rockchip RK3588, CM3588, Nvidia RTX 2060, Nvidia RTX 3050 and AMD V2000, we conducted a series of tests focused on efficiency, energy consumption, and cost-effectiveness. The primary goal was to evaluate how these configurations perform when deployed with SwissVault's VaultFS for inferencing a pre-trained object detection model. Below is an overview of the environment setup, data details, and the process used to generate the results.

SETUP

Software and SDK Configurations

Each device was configured with the appropriate SDK to leverage its hardware capabilities for ML inference:

- Rockchip's RK3588 and CM3588 were set up using `RKNN-toolkit2` and `RKNN-toolkit-lite2`, optimized for their ARM-based NPUs.
- The Nvidia GPUs were configured with CUDA toolkit, while the AMD V2000 with ROCm.
- **VaultFS** was mounted across all systems, providing seamless data access within Swiss Vault's storage solution. This ensured uniform data retrieval and facilitated efficient data processing across different hardware configurations.

Dataset and Model Selection

- The inference model was run on a file located in VaultFS at `/mnt/vaultFS/test_3_1/Cosmos/Cosmos.A.Space.Time.Odyssey.S01E08.HDTV.x264-LOL.mp4`. This file is a 40-minute video comprising roughly 57,000 frames, offering a substantial data load for inferencing tasks.
- The ResNet18 model was utilized for this test given its well-established CNN pre-trained on ImageNet. ResNet18 was chosen for its efficiency in image classification tasks, making it suitable

for this experiment. The inference task involved running frame-by-frame classification on the same video file using ResNet18 on each hardware setup/configuration.

Inference Process

- A worker-coordinator script structure was employed to handle the inference task, processing the video frame-by-frame, when the task was initiated by a coordinator with the option of upscaling worker nodes based on preference and availability. Each frame underwent transformation before being passed to the model for classification. The script extracted top predictions and probabilities for every frame, providing detailed classification insights.
- For configurations involving multiple units (e.g., two RK3588s or configurations with CM3588 or Nvidia GPUs) the inference process was distributed. A coordinating script splits the video into chunks, enabling parallel processing across devices to enhance efficiency and test the capability to support nodal upscale. This setup allowed concurrent frame processing, significantly reducing overall inference time. [2]

Results Collection and Analysis

- Post-processing, the classification results, inference times, and energy consumption data were collected or prepared, as shown in the table below. The RK3588-based configurations demonstrated particularly high energy efficiency, with power consumption levels between 3-5W per chip during the inference tasks.
- Each hardware setup has been evaluated based on its inference speed, energy usage, and operational/purchase costs in the table, where inference speed for multiple prediction lengths were generated. The RK3588 configurations stood out for their low OPEX and competitive inference times, particularly when compared to traditional GPU setups like Nvidia's RTX series (in fact combining to beat Nvidia setup, see table).

This experiment underscores the RK3588's advantages in delivering top-tier performance with remarkable energy efficiency and cost savings when utilized with SwissVault's VaultFS, making it a reliable choice for tasks involving inferencing pre-trained ML-AI-LLM models.

---

[2] The Embarrassingly Parallel Task structure was followed for the distribution task.

Key Findings

Table 1-Inference Test Results

Running Chip	GPU/Chip/ Config-X	Prediction Size	Inference Period	Cost		Energy Consumption
				CAPEX	OPEX	
RK3588	RK3588*1 (8GB) CONFIG-1	3	10m26.874s	0	$	3-5W
		5	10m21.186s	0	$	
		10	10m36.803s	0	$	
		20	10m28.456s	0	$	
		30	10m32.330s	0	$	
		100	10m28.808s	0	$	
	RK3588*1 (32 GB) CONFIG-2	3	10m18.147s	0	$	3-5W
		5	10m8.767s	0	$	
		10	10m17.597s	0	$	
		20	10m26.494s	0	$	
		30	10m31.929s	0	$	
		100	10m29.996s	0	$	
	RK3588*2 (32 GB) CONFIG-3	3	6m18.119s	0	$	5-8W
		5	6m15.757s	0	$	
		10	6m15.518s	0	$	
		20	6m17.780s	0	$	
		30	6m18.970s	0	$	
		100	6m30.080s	0	$	
	RK3588*2 + CM3588 (Technically 3* RK3588) CONFIG-4	3	4m12.239s	0	$	8-10W
		5	4m9.154s	0	$	
		10	4m10.818s	0	$	
		20	4m10.072s	0	$	
		30	4m11.745s	0	$	
		100	4m21.973s	0	$	
	RK3588*2 + CM3588 + RK3588 (8 GB) (Technically 4* RK3588) CONFIG-5	3	3m7.173s	0	$	Peaked at 15W
		5	3m9.433s	0	$	
		10	3m8.809s	0	$	
		20	3m8.435s	0	$	
		30	3m10.801s	0	$	
		100	3m13.963s	0	$	
Nvidia	RTX 2060*1 CONFIG-6	3	7m56.880s	$$	$$	160-180W
		5	7m56.666s	$$	$$	
		10	8m3.965s	$$	$$	
		20	8m13.495s	$$	$$	
		30	8m22.308s	$$	$$	
		100	9m15.437s	$$	$$	
	RTX 3050*1 CONFIG-7	3	5m31.791s	$$	$$	70W
		5	5m32.974s	$$	$$	
		10	5m59.728s	$$	$$	
		20	6m11.796s	$$	$$	

		30	6m15.377s	$$	$$	
		100	7m24.175s	$$	$$	
	RTX 2060 + RTX 3050 CONFIG-8	3	2m45.122s	$$$	$$$	230-250W
		5	2m46.461s	$$$	$$$	
		10	2m57.625s	$$$	$$$	
		20	3m8.920s	$$$	$$$	
		30	3m0.255s	$$$	$$$	
		100	3m51.156s	$$$	$$$	
AMD V2718	GFX90C*1 CONFIG-9	3	27m57.029s	$$	$	25W
		5	27m57.151s	$$	$	
		10	28m20.999s	$$	$	
		20	28m13.586s	$$	$	
		30	28m18.387s	$$	$	
		100	28m41.203s	$$	$	

Performance Highlights

Let's compare the performances as shown in the above table for RK3588's configurations (Config-1 to Config-5), where the inference duration reduces as compute nodes are added and the load is distributed.

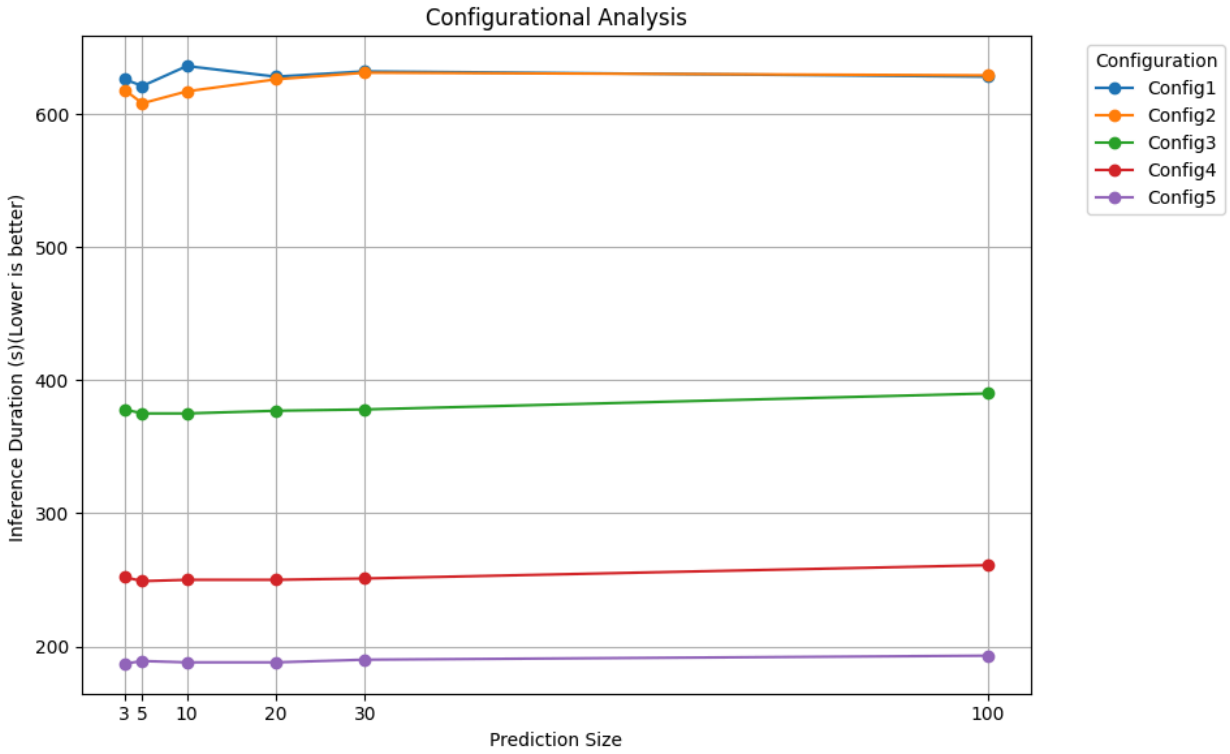

Fig-1 (Config-1 to Config-5, Inference Period)

As illustrated in the figure above (Fig-1), the inference time decreases consistently across all prediction sizes (3, 5, 10, 20, 30, and 100) as additional RK3588 compute nodes are added. During testing, the energy

consumption was observed to fluctuate between *3W and 15W,* underscoring the system's remarkable energy efficiency. This also defines the processing efficiency of RK3588, being able to handle the inference of the 57K frames, improving by the addition of every node.

The Config-1 and Config-2 serve as baselines for RK3588 Performance for the task, which when complimented with addition of multiple nodes for load balancing, displays improvement in speed. The addition of 1 node (Config-3, i.e. 2* RK3588) we see a **40%** faster inference period, and this pattern is continued when 2 nodes (Config-4, i.e. 3*RK3588) displays **55%** and 3 nodes (Config-5, i.e. 4*RK3588) displays **65%** faster average inference speed.

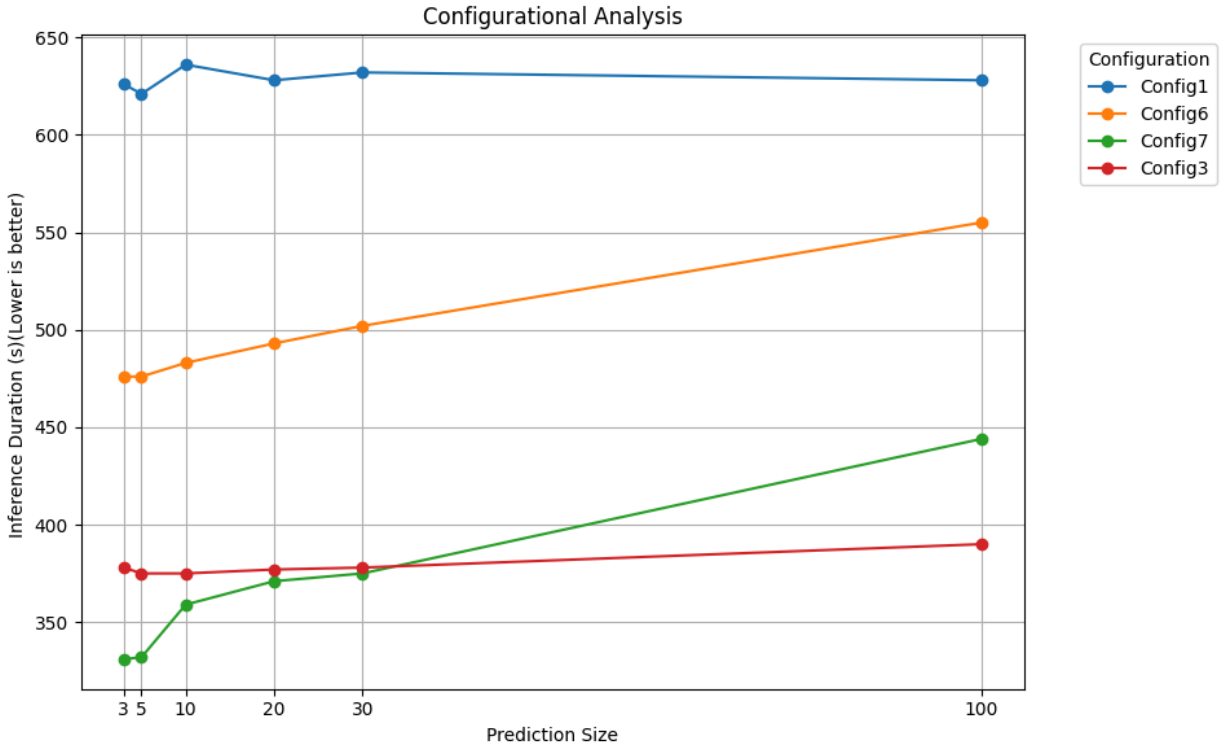

Fig-2 (Config-1,3 and Config-6,7 ; Inference Period)

Similarly, when we look at the results displayed in Fig-2 to compare Config-1 and Config-3 with Config-6 and Config-7, we see that dual node processing of RK3588 (Config-3, i.e. 2*RK3588) displays an average inference period nearly the same as Config-7 [**6.19 minutes**(Config-3) and **6.21 minutes**(Config-7)], which however causes this negligible difference to become more significant when considering the **CAPEX and OPEX** as presented in the table and Fig-4, where Config-3's efficiency in terms of operational and capital expenditures makes it a more cost-effective solution. Not to forget Config-3 turns out to be roughly **30%** faster than Config-6 (i.e. 1*RTX-2060).

An important observation is the behavior of inference periods as the prediction size increases across configurations (3, 5, 10, 20, 30, 100). While Config-6 and Config-7 display larger gaps in inference periods with increasing prediction size, Config-1 and Config-3 exhibit stable increments with minimal overhead. This trend results in *Config-3 outperforming Config-7* (see Fig-2) when the ***prediction size = 100.***

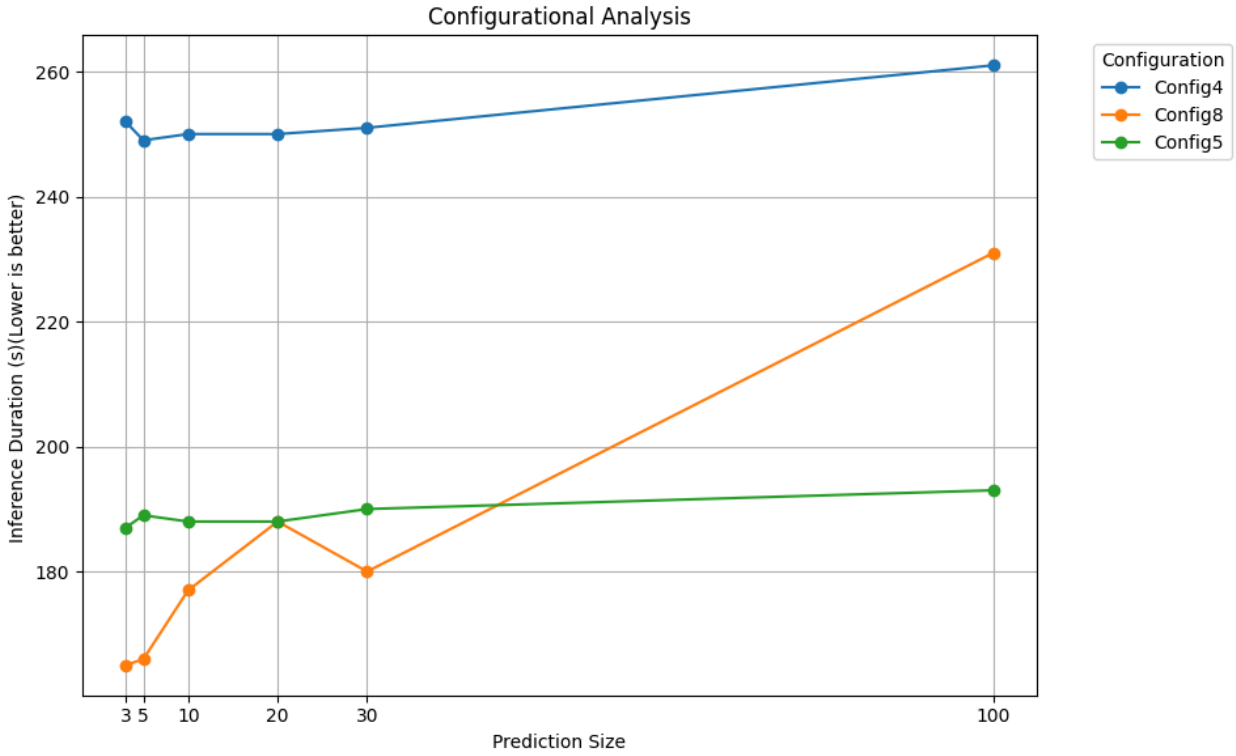

Fig-3 (Config-4,5 and Config-8 ; Inference Period)

In Fig-3, we see that when Config-4 (3*RK3588) and Config-5 (4*RK3588) are compared with Config-8 (RTX-2060+RTX-3050), a similar trend as seen previously is followed, i.e. 2 RK3588s can offer performance comparable to 1-mid range Nvidia GPU, at better efficiency with lower **CAPEX+OPEX** and hence Config-5 (Inference Period) ~ Config-8 (Inference Period).

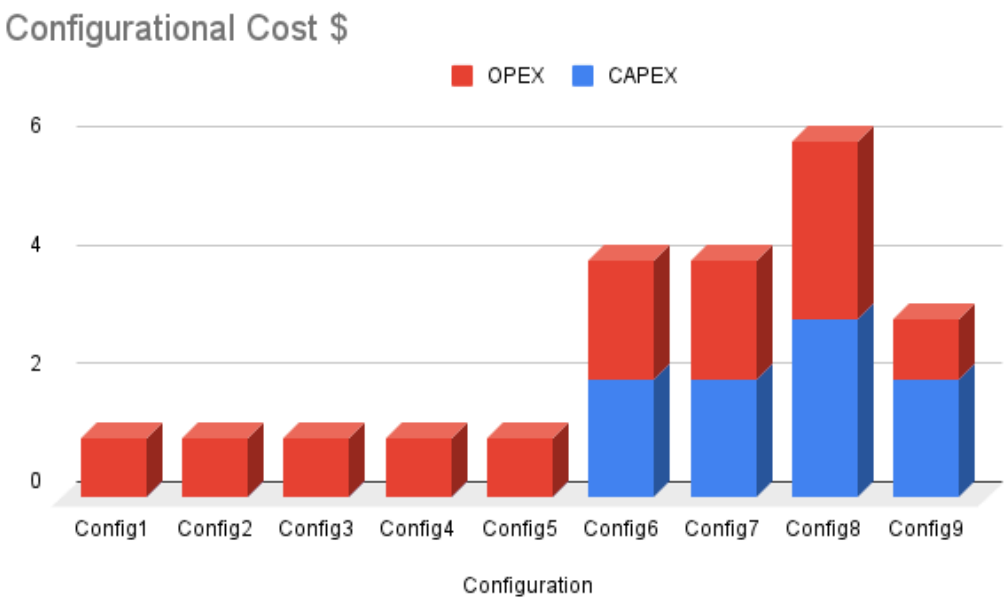

Fig-4 (Capex & Opex comparison for configurations)

Note that we have considered **CAPEX** for using RK3588 as $0, given it will be offered along with Swiss Vault's Storage Solutions, at no added additional cost. As shown in **Fig-4,** the OPEX is primarily driven by energy consumption costs, which are further illustrated in **Fig-5**.

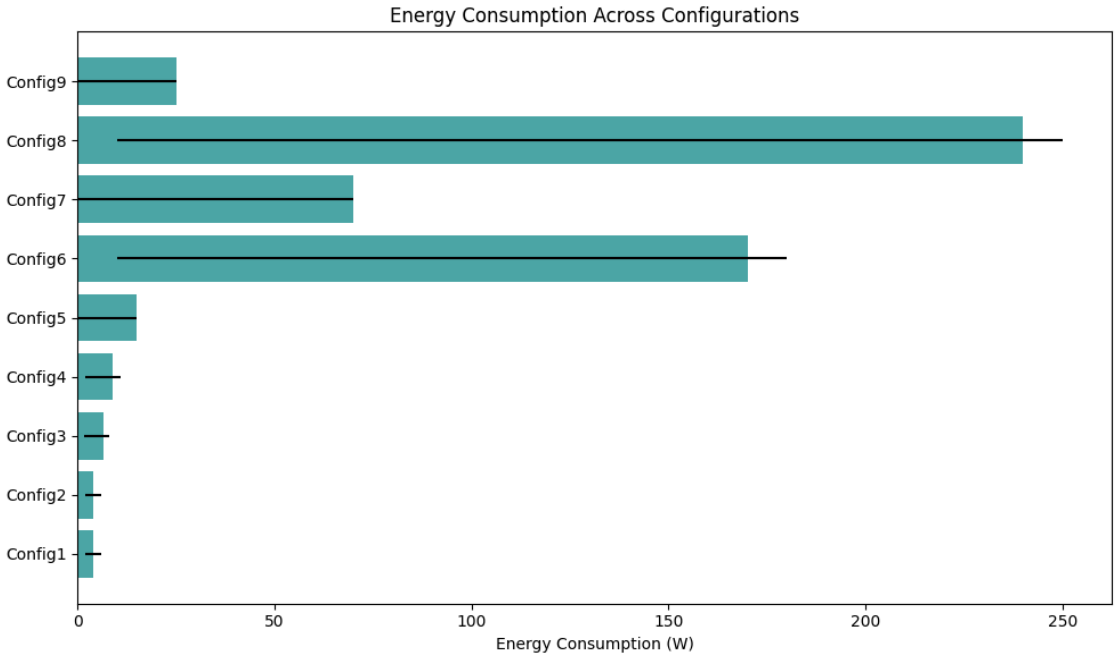

Fig-5 (Ranged Energy Consumption)

Fig-5 highlights variable configurational energy consumption (EC) with the black line in the middle indicating the range. **Config 1–5** exhibit significantly lower EC ranges, whereas **Config 6–8**, featuring power-hungry Nvidia GPUs, demonstrate substantially higher EC and consequently higher OPEX.

For reference, we also included **Config-9** (AMD iGPU setup for V2K), which ranked lowest in our tests for both inference duration and energy efficiency.

Our tests quantified **key performance indicators such as Energy Consumption, Performance, and OPEX/CAPEX**, providing actionable insights for comparison. Along with that the tests have clearly shown the use of RK-3588 in multiple setups in conjunction with the Vault File System (or VaultFS), to be an extremely efficient and confident choice for **EDGE AI** use cases, when inferring complex models with large sets of data and compute, keeping in mind the ease of access it provides.

Edge AI focuses on performing inference tasks on local edge devices, rather than relying on cloud servers, enabling real-time decision-making, low latency, and reduced bandwidth consumption. Energy-efficient processors like RK3588 excel in this context, providing high compute power while maintaining minimal energy consumption. Supported by a robust, scalable, and high-performing file system like VaultFS, RK3588-based setups offer a compelling solution for data-intensive and compute-heavy AI tasks. VaultFS enables efficient data access, storage, and retrieval, which complements RK3588's processing efficiency. This synergy allows organizations to deploy Edge AI models seamlessly, even for complex inference workloads involving massive datasets using our **SuperNAS.**

The future of Edge AI is driven by efficiency, scalability, and accessibility. By utilizing energy-efficient chips like RK3588 in conjunction with VaultFS, industries can unlock significant benefits:

1. **Energy Conservation:** Reduces power consumption while maintaining performance, crucial for sustainable AI deployments.
2. **Cost Effectiveness:** Minimizes both CAPEX and OPEX, making Edge AI viable at scale.
3. **Real-Time Performance:** Ensures low-latency inference, vital for autonomous systems, IoT, and real-time analytics.

The integration of RK3588 with scalable systems like VaultFS highlights a confident step forward toward sustainable Edge AI, enabling businesses to process complex models efficiently while contributing to a more energy-conscious future. This approach empowers industries across healthcare, manufacturing, smart cities, and beyond to **leverage AI where it matters most: at the EDGE.**

Get in touch with us:

VaultFS values its customers and has dedicated resources to ensure the best quality experience when demoing the product. Businesses can discover how VaultFS can serve their needs, by visiting ***Vaultfs*** .

For more information, you can also visit us at https://www.swissvault.global/