NVMe In-Line Accelerator Enhancing AI Ecosystems in the Data Center

Introduction

In the era of artificial intelligence (AI) dominance, data centers are undergoing a profound transformation to meet the escalating demands for computational power, storage efficiency, and accelerated processing. In this digital renaissance, where every bit and byte carry the promise of transformative insights, data centers stand as the champions of progress, adapting and evolving to meet the relentless demands of AI applications.

One significant advancement in this evolution is the integration of the Unigen Mercier NVMe accelerator, purpose-built as a hardware offload solution to augment AI workload processing. By seamlessly incorporating these accelerators into their infrastructure, data centers are not only meeting but also exceeding the demands of AI applications.

This paper explores how data centers are catering to the demands of AI and how integrating the Unigen Mercier NVMe accelerator optimizes resource allocation, security offload, and hardware virtualization, thereby revolutionizing the landscape of AI computing.

Security Offload

The Unigen Mercier NVMe accelerator addresses the demand for centralizing the security management feature in enterprise storage applications.

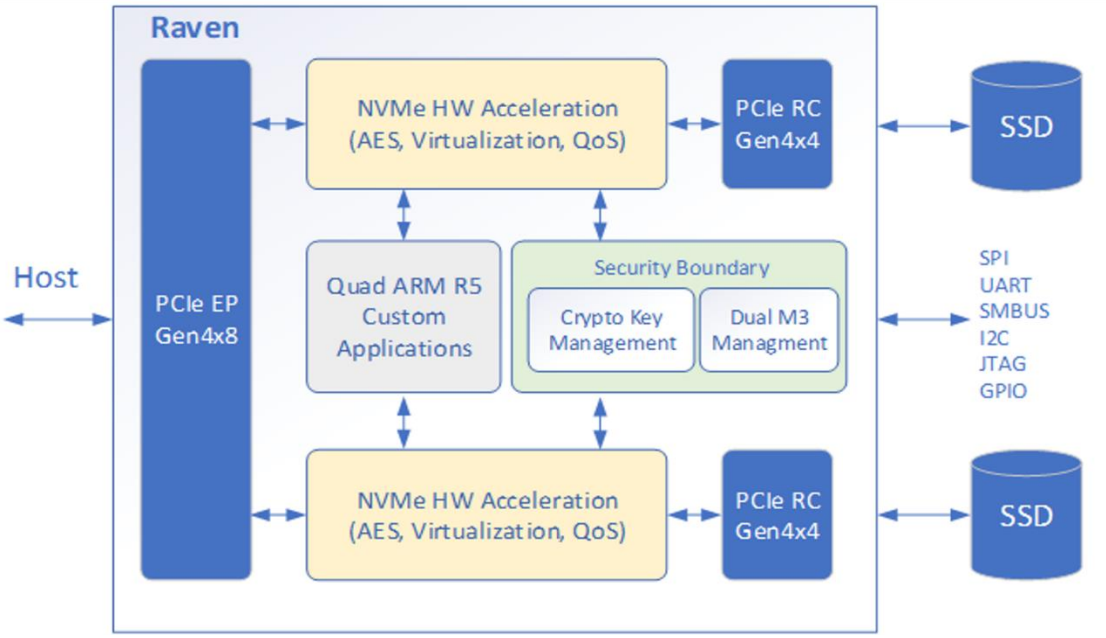
The offload scheme has three goals:

1. Encrypting the data in-flight to the SSD; adheres to the industry demand for data encryption in transit and at rest
2. Managing the encryption keys; alleviates administration of policies and procedures for protecting, storing, organizing, and distributing encryption keys away from the host and the SSD device

3. Providing FIPS 140-3 Level 2 certification

Integrating a solution that provides certification from the Federal Information Processing Standards (FIPS) paves the way for data centers seeking to be Federal Risk and Authorization Management Program (FedRAMP) compliant. For example, Google Cloud Services and Microsoft AZURE both claim various levels of FedRAMP compliance. Figure 1 provides a block- level overview of the Unigen Mercier NVMe accelerator.

Figure 1. Mercier Block Diagram



Hardware Virtualization

Efficient resource allocation is paramount for maximizing performance and optimizing infrastructure utilization. One groundbreaking solution that has emerged to address this challenge is the integration of single root I/O virtualization (SR-IOV) technology. SR-IOV- enabled devices emerge as a potent solution, leveraging virtualization technology to streamline resource allocation. By offloading I/O operations typically managed by a hypervisor and enabling direct access to SSD resources for virtual machines, SR-IOV enhances throughput and latency, crucial for AI workloads' demanding nature.

Integrating the Unigen Mercier NVMe accelerator with SR-IOV in today's data center can further enhance resource allocation and performance optimization, particularly for storage-intensive workloads. Here is how:

1. Improved storage performance: By integrating SR-IOV, virtual functions (VFs) can be directly assignable to virtual machines, bypassing the hypervisor, and

reducing overhead. This direct access enables faster data transfers and more efficient storage operations, enhancing workload performance.

2. Enhanced storage virtualization and security: SR-IOV allows for the creation of VFs that are dedicated to specific virtual machines (VMs) which provides efficient storage virtualization, with each VF having direct hardware access to SSD storage resources. SR-IOV provides security by isolating VMs, reducing the risk of one VM interfering with the operations of an adjacent VM. The MERCIER NVMe accelerator also incorporates firmware with the ability to detect malicious activity, such as bad physical region page access and admin command flooding, and to take action by rate limiting or potentially shutting down the VM altogether.
3. Reduced CPU overhead: CPU overhead related to storage I/O operations managed by the hypervisor is minimized. This frees CPU resources for other compute-intensive tasks, leading to better overall resource utilization and improved application performance.

Below are some examples of CPU utilization, performance, and latency improvements observed during our comparison testing in Figures 2, 3, and 4.

Figure 2. CPU Utilization Comparison

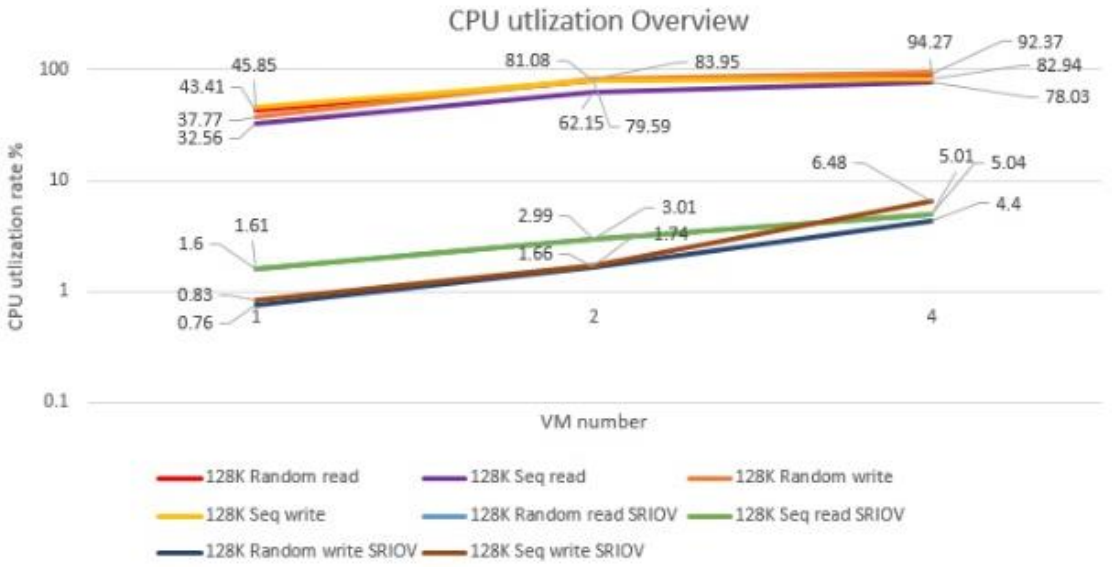


Figure 3. Performance Comparison

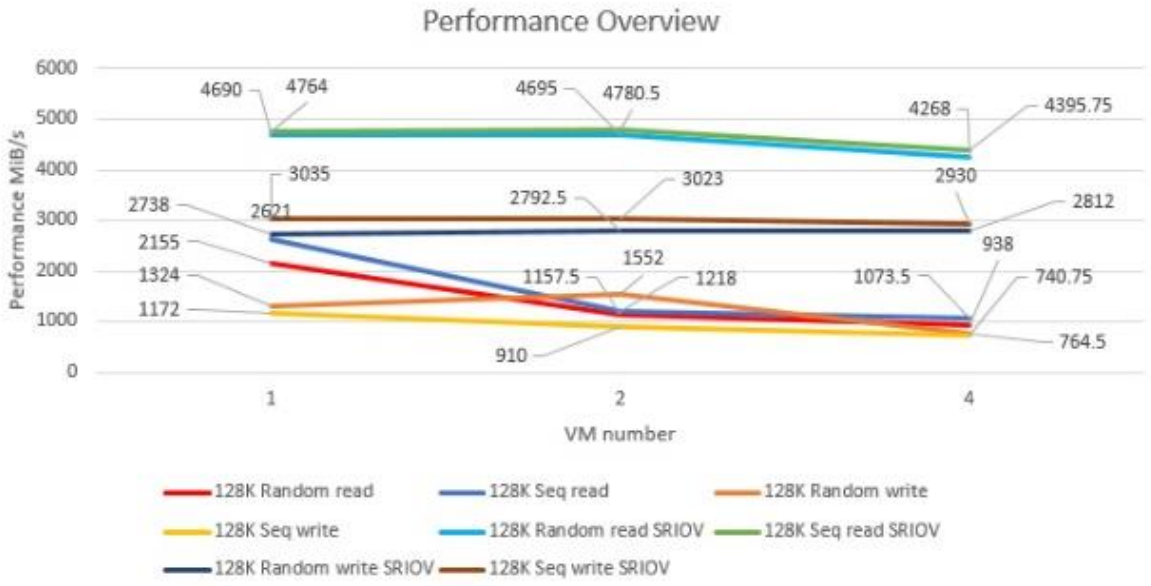
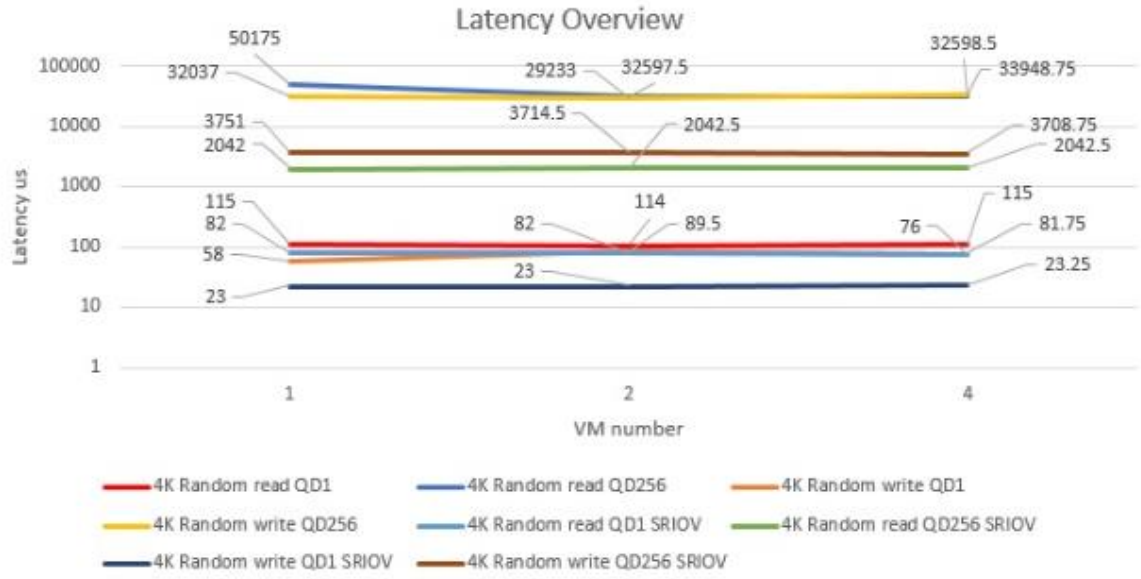


Figure 4. Latency Comparison



In Figure 2, with SR-IOV enabled you can achieve significant CPU utilization savings with over 80% better efficiency. Figures 3 and 4 delineate the clear advantage of hardware virtualization maximizing the performance of the attached SSD device. These data points can be viewed as a raw assessment on the potential savings due to variations of platforms, workloads, and implementation choice. Flexible IO tester was

used to generate specific workloads along with CPU utilization monitoring to garner resource allocation metrics. This provides end customers with additional tools to address TOC savings based on their targeted needs.

Scaling Virtual Machines per Server

Another benefit of utilizing SR-IOV is scaling the number of virtual machines (VMs) per server, which can yield substantial cost savings for data centers.

With more VMs running on a single server, data centers can achieve higher levels of server utilization, thereby maximizing the return on investment for hardware resources. Additionally, by reducing the number of physical servers needed to support a given workload, data centers can save on hardware costs, power consumption, cooling expenses, and physical space requirements.

Conclusion

In conclusion, the integration of security offload and SR-IOV-enabled devices heralds a paradigm shift in data center resource allocation at the storage level. This symbiotic relationship not only enhances efficiency but also fortifies security measures, ensuring a seamless and protected computing environment. Together, they pave the way for a future where resource allocation in today's data centers are both optimized and safeguarded, laying the foundation for unprecedented levels of performance and resilience.