Security PoC Report for Green Computing with Remote GPU over APN

August 2026

Table of Contents

1. Introduction

This document is a PoC report for the security aspects of "Green Computing with Remote GPU over APN for GenAI / LLM with IOWN GF technologies project".

Background

"Green Computing with Remote GPU over APN for GenAI / LLM with IOWN GF technologies project" assumes a use case in which users generate an LLM by training on GPU computing located on an Open APN-connected green data center using the large amounts of training data that exists at their location.

This use case enables lower energy costs per training time for LLM generation by leveraging energy-efficient green data centers, even in a remote access environment.

When handling sensitive data, such as medical or healthcare records or data used for drug discovery, in this use case, it is mandatory to implement data protection mechanisms to mitigate the risk of data leakage and to ensure that data owners maintain data sovereignty.

Purpose

The purpose of this PoC is to verify the feasibility of end-to-end (E2E) data protection architectures proposed in "Security RIM for Green Computing with Remote GPU over APN [IOWN PETs FA]" which fulfill the security requirements of the green computing in large-scale AI training environments across remote locations using next-generation All-Photonic Network (APN) technology advocated by the IOWN Global Forum (GF).

Structure

This report systematically summarizes the PoC background, purpose, validation scenarios, system structure, technical evaluation results, contributions to IOWN GF, and future outlook. Through this, we aim to provide technical and operational guidelines toward the societal implementation of green data centers and distributed AI infrastructures with data security.

2. Overview of the PoC Project

2.1 Scope and Achievement of the PoC

The scope of this PoC is to comprehensively verify the performance, efficiency, and sustainability of the AI development infrastructure in a remote GPU environment utilizing APN technology advocated by the IOWN GF and the domestically developed large language model "tsuzumi" with data security.

To evaluate its practical feasibility, this PoC includes measuring the performance impact of the remote GPU service configuration enhanced with data security. A large language model (LLM), "tsuzumi"[TSUZUMI], was used for AI analysis.
Performance measurements will be conducted under the following conditions.

- GPU and storage node placement: local connection or remote (APN) connection
- Data security: Yes or No

2.1.1 Hypotheses and Demonstrated Results

We hypothesized that the impact of distance on training time would remain within an acceptable range, as defined in Section 2.1.2, when the security requirements for direct access via APN are fulfilled.
The evaluation results demonstrated that the impact was within the defined performance target.

2.1.2 Main Goals of the PoC

The primary goal of this PoC is to evaluate the feasibility of remote GPU training across distance, utilizing the data security architecture defined in "Security RIM for Green Computing with Remote GPU over APN," while achieving an increase in training time of less than 10% compared to operations within the same data center.

In addition, this PoC aims to clarify the impact on the performance of large language model (LLM) training introduced by the data security architecture defined in the same document.

2.1.3 Summary of PoC Reference Implementation and Evaluation

As shown below. See Section 4 for details.

Table 2-1: Summary of PoC results

Use Case	Key Requirement (KR)	Validation Point	Result	Relevant Section
Direct access (via APN)	KR1.1	Increase of training time < 10%	Pass	4.1

2.2 PoC completion status

Overall PoC Project Completion Status: Complete.

Table 2-2: Summary of Benchmark Tests conducted

Total Training Time	Storage Access Speed (Throughput and/or Latency)	Power Consumption	Cost Evaluation
Complete	Complete	Complete	Complete

2.3 PoC Project Participants

PoC Project Name: Security PoC Report for Green Computing with Remote GPU over APN

Table 2-3: PoC project participants

Company	Name(s)
NTT, Inc.	fumiaki.kudoh@ntt.com

3. Confirmation of PoC Demonstration

3.1 Background of AI development environment in remote locations

Recently, with the rapid proliferation of generative AI, there is an increasing need in AI development environments to efficiently coordinate and transfer large amounts of data between remote locations. Specifically, for training and inference using large language models (LLM), high-speed and stable data access in a distributed environment is required. Additionally, while the use of green data centers leveraging renewable energy in remote areas is desirable from a cost perspective, there are concerns about processing performance due to the distance from the data origin.

At the same time, for services such as those listed below that involve the use of confidential data, concerns regarding the risk of data leakage become a barrier to AI operations and the utilization of remote computing resources. Examples include:

- Medical record information
- Drug discovery

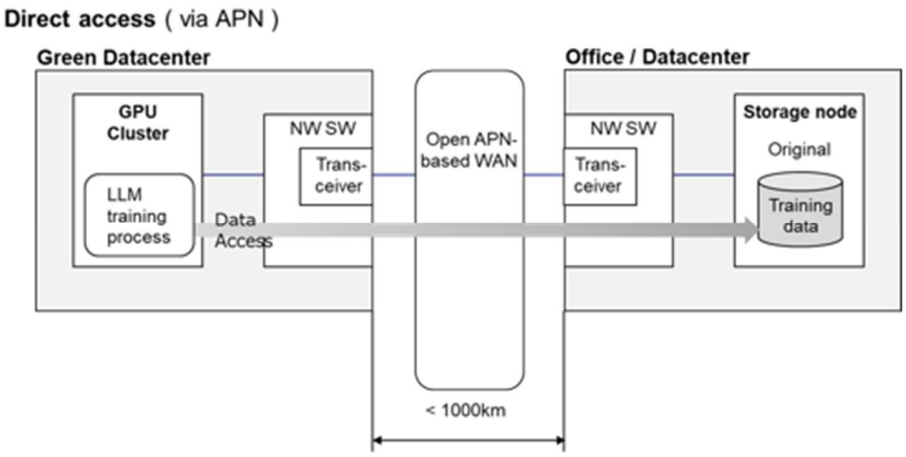
Against this background, we performed a comparative evaluation of training completion times in pseudo-IOWN APN (All-Photonic Network) and local environments to assess the effectiveness of IOWN (Innovative Optical and Wireless Network) in long-distance network environments while considering E2E data confidentiality.

For the training target, we decided to use large language models (LLMs), which have rapidly gained popularity in recent years, and to perform the training using GPU resources located remotely.

3.2 Supported Use Case

For verification, we used the domestically produced LLM 'tsuzumi' and conducted data transfer simulating an actual AI development environment.

Among the three system architectures defined in "the Reference Implementation Model and Proof-of-Concept Reference of Green Computing with Remote GPU", this PoC focuses on "Direct access". In this method, the GPU accesses training data on the Office side via Open APN.



The learning process of a typical LLM is shown in the figure below.

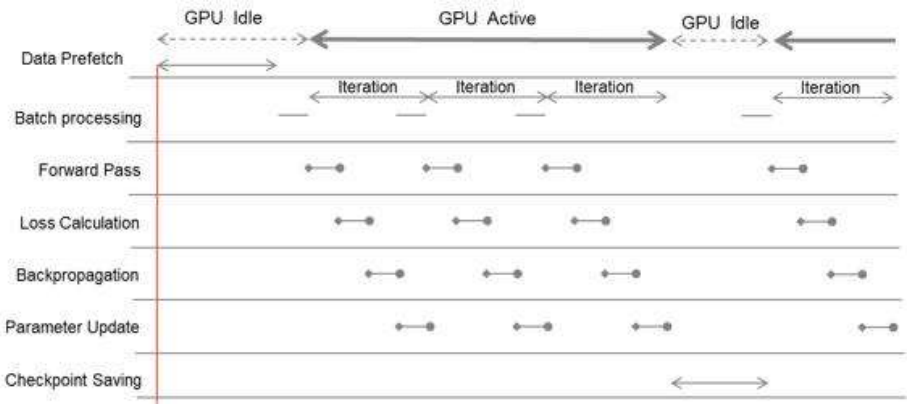


Figure 3-2: LLM Training Process and GPU Time

In this figure, Data Prefetch corresponds to the Read Time or data transfer waiting time, while Checkpoint Saving corresponds to the Save Time. It is important to minimize the proportion of Save Time, during which the GPU remains idle, by repeating iterations as many times as possible. In this PoC, we assumed that data security is maintained while using the remote environment, and checkpoints are saved on the remote environment where the GPU resides.

The data security configuration to be evaluated is as follows:

- The configuration ensures data confidentiality between on-premises storage and remote GPUs without relying on the infrastructure provider as a trusted party, while providing users with verifiable assurance of that confidentiality. It comprehensively protects data in all three states—**data at rest**, **data in motion**, and **data in use**—and provides seamless protection across transitions between these states. This security architecture assumes a trust model in which the chip

vendors providing the TEE and GPU TEE technologies, as well as the service provider that delivers distributed confidential computing based on these technologies, are trusted.

- Specifically, isolated computing environments based on **Trusted Execution Environments (TEEs)** are deployed at both the on-premises and remote GPU sites. After establishing mutual trust through **remote attestation**, which enables mutual verification of both TEE environments, the two sites are seamlessly connected via Layer 2 quantum safe encrypted communication which is ensured through MFS FA–compliant communications [MFS FA]. It uses two post quantum cryptographies for key exchange and AES-256 for MACsec. This foundation is further complemented by encrypted communication between the storage system and the TEE, as well as encrypted CPU–GPU communication leveraging a **GPU TEE**, thereby providing **end-to-end data protection from storage to GPU**.

3.3 Validation Contents

Experimental Configurations and Metrics

Conduct measurements of AI training time using GPUs across four configurations, considering node placement and with or without data security.

The primary metric is the increase in tsuzumi's training completion time (Read, GPU, and Save Time), measured by comparing the local and remote (APN) configurations, and an additional metric is also examined, which is the increase observed when comparing configurations with and without the data security feature.

Table 3-1: Experimental Configurations

		Node placement	
		Local	**Remote (APN)**
Data security	**without**	Configuration 1	Configuration 3
	with	Configuration 2	Configuration 4

For details on the implemented system configurations and measurement methodology, refer to Sections 5.1 and 5.2, respectively.

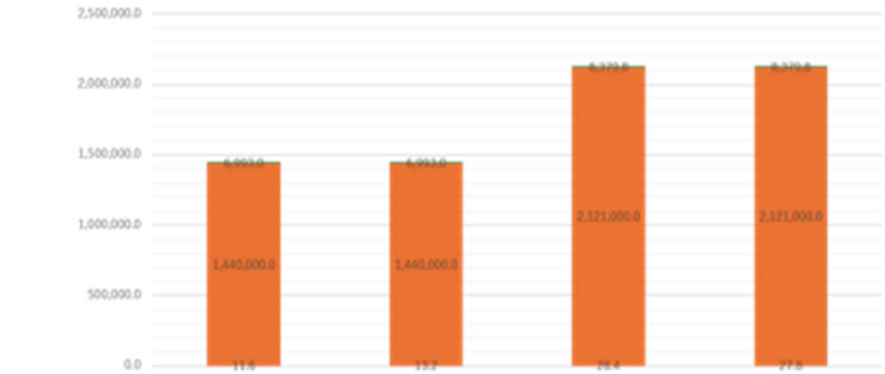
4. PoC Goals Status Report

4.1 Key Result

In this direct access use case, we evaluated the increase in training time and confirmed that it is less than 10% both with and without security, which is the key requirement (KR) described in Chapter 2. We also examined the impact on power consumption and cost.
Below are the evaluation results for training time and for power consumption and cost.

4.1.1 Training Time

The graph below shows the tsuzumi training time (Read Time, GPU Time and Save Time), measured for each of configuration 1 to 4 described in Chapter 3.3 and 5.1. For APN connections, the distance is calculated from the network latency.



	Configulation 1 MACsec/TEE OFF 0km	Configulation 3 MACsec/TEE OFF APN 330km (3.7ms)	Configulation 2 MACsec/TEE ON 0km	Configulation 4 MACsec/TEE ON APN 330km (3.7ms)
SAVE TIME (s)	6,993.0 (±906.1)	6,993.0 (±906.1)	8,379.8 (±255.9)	8,379.8 (±255.9)
GPU TIME (s)	1,440,000.0 (±3442.8)	1,440,000.0 (±3,442.8)	2,121,000.0 (±2,377.5)	2,121,000.0 (±2,377.5)
READ TIME (s)	11.6 (±4.4)	13.2 (±4.9)	28.4 (±10.4)	27.8 (±10.4)

mean (± standard deviation)

Figure4-1: Comparison of tsuzumi Training Time

The training time was measured for AI training using the tsuzumi large language model by measuring the training data loading time (Read Time), the GPU computation time for parameter calculation (GPU Time), and the checkpoint saving time (Save Time). Data is read once per measurement, with the cache cleared and newly loaded each time. The reported values represent the averages of 10 measurement runs for each configuration.

This training was conducted using a single H100 GPU and a 10 GiB training dataset, and GPU Time was calculated by multiplying the measured GPU Time by the total number of training steps, which was set to 150,000 in this evaluation. Save Time was calculated by multiplying the measured Save Time by the number of save operations. Checkpoint saves were performed every 2,000 steps, resulting in a total of 75 save operations (for details, see Section 5.2). Results may differ in multi-GPU or large-scale training environments.

The results confirm that, even when accessed remotely via the IOWN APN, both configurations—with and without data security mechanisms enabled—satisfy the key requirement (KR) of less than a 10% increase in training time when comparing local and remote environments. These results indicate that the remote GPU architecture over the IOWN APN remains feasible from a performance perspective, even with data security enabled.

A detailed breakdown of the training time shows that GPU Time is the dominant factor, accounting for approximately 99.6% of the total training time, while Read Time and Save Time contribute only marginally.

Similarly, the *"PoC Report – Green Computing with Remote GPU over APN (using tsuzumi-7B)"* presents measurement results conducted with eight H100 GPUs. Although the GPU Time is reduced to roughly one-third, it still remains the primary contributing factor (see the report for further details).

On the other hand, the impact of enabling data security processing itself is observed: when data security mechanisms (software MACsec and TEE) are enabled, the total training time increases by approximately 47% compared to the non-secure configuration under the same local or remote environment, due to an increase in GPU Time caused by the overhead of security processing, where GPU Time is the dominant component of the overall training time. This overhead is primarily attributed to encrypted communication between CPU and GPU (see Section 5.2 for the definition of GPU Time).

Detailed experimental results regarding Read Time and storage access speed are provided in the Appendix.

4.1.2. Power consumption and cost

The power consumption in this PoC is calculated based on the actual expected power consumption from catalog specifications. The power consumption is assumed to be 70% of the rated value specified in the catalog. Although this assumption is used in the calculations, it does not affect the analysis results because it cancels out when calculating percentage differences between configurations.

Table 4-1: Power Consumption of Equipment

Equipment Model	Power Consumption (kW)	70% of Power Consumption (kW)
PowerEdge R7525	1.80	1.26
NVIDIA H100 PCIe 80GB*	0.35	0.25
NetApp AFF C400	1.24	0.87
Galileo-1 (G1)	0.80	0.56
Alaxala AX2630S-24T4XW	0.05	0.03
Juniper EX4600-40F-AFO	0.37	0.26

*SEV-compatible, VBIOS applied

In this PoC, we trained the LLM in two types of environments:
Configurations 1 and 2, which operate entirely within a local data center, and
Configurations 3 and 4, which utilize a hybrid setup combining local and remote sites.

Note: Labels marked ‘Local’ in the column headers of the tables below indicate that the data in those columns relates to equipment located at the local sites. Labels marked ‘Remote (APN)’ indicate that the data corresponds to equipment at remote sites.

The total power consumption of all equipment used in each Configurations are listed in the following table.

Total Power Consumption of each equipment (kW)
= power consumption of equipment (kW) × quantity (Qty) of equipment used in the Configuration

Table 4-2: Total Power Consumption of equipment for each Configurations

Equipment Model	Configuration 1,2		Configuration 3,4				
	Local		Local		Remote (APN)		Total (Local + Remote (APN))
	Qty	Power Consumption (kW)	Qty	Power Consumption (kW)	Qty	Power Consumption (kW)	Total Power Consumption (kW)
PowerEdge R7525	2	2.52	1	1.26	1	1.26	2.52
NVIDIA H100 PCIe 80GB	1	0.25	0	0.00	1	0.25	0.25
NetApp AFF C400	1	0.87	1	0.87	0	0.00	0.87
Galileo-1 (G1)	2	1.12	1	0.56	1	0.56	1.12
Alaxala AX2630S-24T4XW	1	0.03	1	0.03	0	0.00	0.03
Juniper EX4600-40F-AFO	0	0.00	1	0.26	1	0.26	0.51
Total	-	4.78	-	2.98	-	2.32	5.30

The measured training time for each of the configurations 1 to 4 described in Chapter 5.1. and Total Energy Consumption of each Configuration is listed in the following table.

Total Energy Consumption for Training (kWh)
= Total Power Consumption (kW) × Total Training Time (h)

Table 4-3: Total Energy Consumption for training in each Configurations

	Total Training Time (s)	Total Training Time (h)	Total Power Consumption (kW)			Total Energy Consumption for Training (kWh)		
			Local	Remote (APN)	Total	Local	Remote (APN)	Total
Configuration 1	1,447,004.63	401.95	4.78	0.00	4.78	1,923.11	0.00	1,923.11
Configuration 2	2,129,408.15	591.50	4.78	0.00	4.78	2,830.04	0.00	2,830.04
Configuration 3	1,447,006.16	401.95	2.98	2.32	5.30	1,195.79	932.72	2,128.51
Configuration 4	2,129,407.52	591.50	2.98	2.32	5.30	1,759.72	1,372.58	3,132.30

To calculate the Total cost of training, PUE (Power Usage Effectiveness) and Power Unit Price of each data center should be considered. PUE is a parameter indicating the efficiency of a data center's power utilization. Lower PUE value is more efficient.

The PUE values and electricity unit prices for standard and green data centers are assumed as in the table below:

Table 4-4: PUE and Electricity Costs: Standard vs. Green Data Centers

Data Center Type	PUE Value	Power unit price ($/kWh)
Local: Standard Data Center	1.8	0.20
Remote: Green Data Center	1.2	0.10

Total Energy Cost for Training ($)
= Total Energy Consumption for Training(KWh) × PUE × Power Unit Price ($/kWh)

Total cost of training is listed in the table below.

Table 4-5: Total Energy Cost for Training in Each Configuration

	Total Energy Consumption for Training (KWh)		**PUE**		**Power Unit Price ($/kWh)**		**Total Energy Cost for Training ($)**		
	Local	**Remote (APN)**	**Local**	**Remote (APN)**	**Local**	**Remote (APN)**	**Local**	**Remote (APN)**	**Total**
Configuration 1	1,923.11	0.00	1.8	1.2	0.2	0.10	692.32	0.00	692.32
Configuration 2	2,830.04	0.00	1.80	1.20	0.20	0.10	1,018.82	0.00	1,018.82
Configuration 3	1,195.79	932.72	1.80	1.20	0.20	0.10	430.48	111.93	542.41
Configuration 4	1,759.72	1,372.58	1.80	1.20	0.20	0.10	633.50	164.71	798.21

Equipment Cost

Total equipment cost is listed in the table below.

Table 4-6: Total Equipment Cost

Equipment Model	Price ($)	Configuration 1,2		Configuration 3,4				
		Local		Local		Remote (APN)		Total
		Qty	Equipment Price ($)	Qty	Equipment Price ($)	Qty	Equipment Price ($)	Total Equipment Price ($)
PowerEdge R7525	21,825	2	43,650	1	21,825	1	21,825	43,650
NVIDIA H100 PCIe 80GB	31,500	1	31,500	0	0	1	31,500	31,500
NetApp AFF C400 *	144,000	1	144,000	1	144,000	0	0	144,000
Galileo-1 (G1)	13,000	2	26,000	1	13,000	1	13,000	26,000
Alaxala AX2630S-24T4XW	3,700	1	3,700	1	3,700	0	0	3,700
Juniper EX4600-40F-AFO	4,985	0	0	1	4,985	1	4,985	9,970
Total			248,850		187,510		71,310	258,820

* The cost of the NetApp AFF C400 is calculated as follows:

- The unit price of the AFF C400 is approximately $1.2K per TiB of effective capacity, including a 3-year maintenance fee.
- The configuration used in our testing was 120 TiB.
- This results in a total of roughly $144K (inclusive of the 3-year maintenance fee).

Total Cost

Total Cost is a sum of energy cost for training and equipment cost.

Total Cost ($) = Total Energy Cost for Training ($) + Total Equipment Price ($)

Table 4-7: Total Cost

	Total Energy Cost for Training ($)	Total Equipment Price ($)	Total Cost (Energy + Equipment) ($)
Configuration1	692.32	248,850.00	249,542.32
Configuration2	1,018.82	248,850.00	249,868.82
Configuration3	542.41	258,820.00	259,362.41
Configuration4	798.21	258,820.00	259,618.21

Analysis

The result datasets for each configuration are compared with those of Configuration 1, and the percentage differences obtained from these comparisons are shown in the Table 4-8.

Table 4-8: Comparison to Value of Configuration 1

	Total Training Time	Total Power Consumption	Total Energy Consumption for Training	Total Energy Cost
Configuration1 (Reference value) Local, Security: OFF	0.00%	0.00%	0.00%	0.00%
Configuration2 Local, Security: ON	47.16%	0.00%	47.16%	47.16%
Configuration3 Remote, Security: OFF	0.00%	10.68%	10.68%	-21.65%
Configuration4 Remote, Security: ON	47.16%	10.68%	62.88%	15.29%

As explained in Section 4.1.1, the total training time increases by 47.16% when data security mechanisms are enabled. Although energy consumption typically scales proportionally with processing time, a larger increase of 62.88% was observed at the remote site (Configuration 4).

The higher power consumption rate of 10.68% associated with the configurations at the remote site may be one of the contributing factors of this result.

Energy consumption is determined by processing time × power consumption, so increases in both factors compound with each other.
$(1+0.4716)\times(1+0.1068)-1=62.88\%$

In other words, the 47.16% increase in processing time and the 10.68% increase in power consumption do not simply add together. Instead, their combined effect results in a 62.88% increase in energy consumption.

Despite the 62.88% increase in energy consumption, the use of the green data center reduced the impact on electricity cost dramatically, limiting the cost increase to only 15.29%.

In the green data center, the combination of a low PUE and a lower electricity price reduces the energy cost per kWh to approximately one-third of that in the local environment. Configuration 4 processes 44% of its total energy consumption on this lower-cost remote infrastructure. As a result, although total energy consumption increases significantly, the impact on the overall energy cost is substantially mitigated.

This demonstrates that the green data center effectively absorbed the additional energy-related cost through its efficient PUE and lower electricity unit price.

Based on these findings, it is anticipated that, even in scenarios where the security feature is enabled, relocating a subset of devices to remote facilities characterized by lower electricity costs and higher energy efficiency will allow processing to be executed while reducing the electricity expenses associated with the relocated equipment.

CPUs and GPUs with TEE capabilities are required for this data security architecture. High-end CPUs and GPUs include TEE capabilities by default. When using such processors, there is no cost difference between systems with and without this security feature. Cost differences arise only when low-end or mid-range CPUs or GPUs without TEE support are used.

5. PoC Technical Report

5.1 Implemented System

Experimental Configurations and Metric

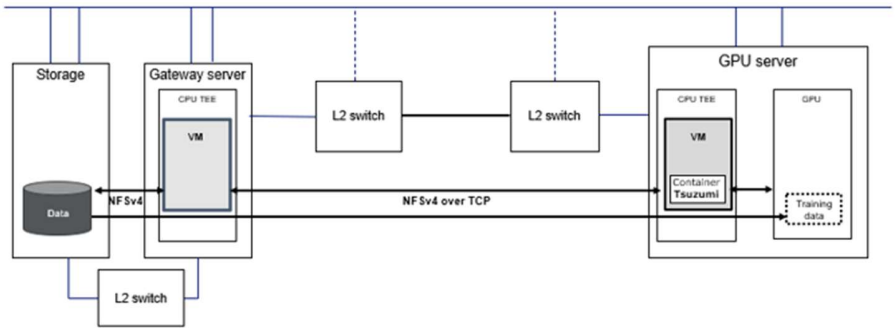
Conduct measurements of AI training time using GPUs across four configurations, considering node placement and with or without data security.

The diagrams for each of the configurations described above are shown below.

5.1.1 System Configuration (Local Connection)

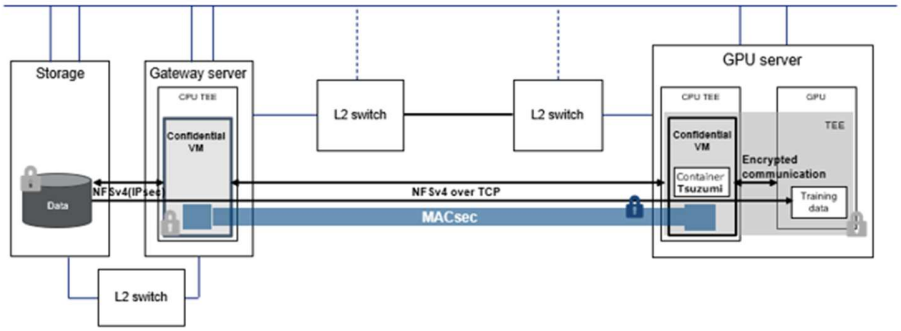
- **Configuration 1 (without data security)**

This configuration represents the pattern with a local connection and no security, where tsuzumi runs in a container inside the VM and accesses the storage data via NFSv4 in two hops.



- **Configuration 2 (with data security)**

This configuration represents the pattern with a local connection and security enabled, where the VM operates as a Confidential VM using TEE, the GPU is TEE-enabled, software MACsec is applied for communication between VMs, and IPsec is used between the Confidential VM and the storage.

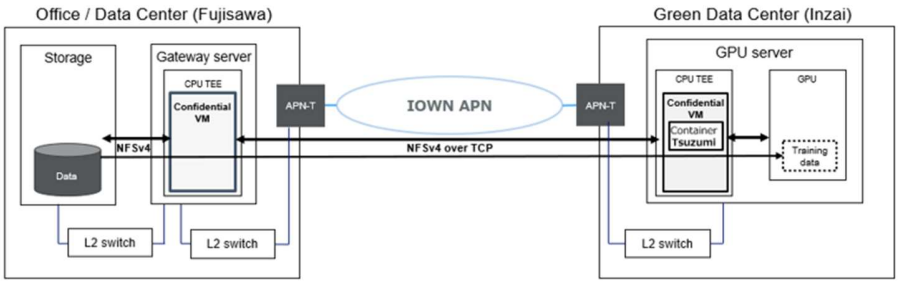


5.1.2 System Configuration (Remote (APN) Connection)

- **Configuration 3 (without data security)**

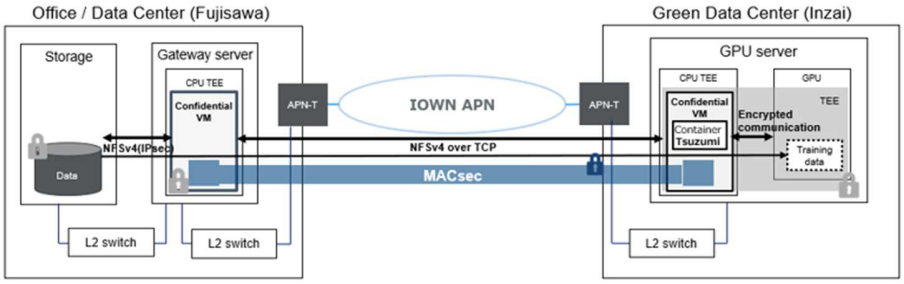
This configuration represents the pattern with a remote connection over the APN and no security. It is basically the same as Configuration 1 except for the remote connection via APN (the office is located in Fujisawa City and the green data center in Inzai City), which is approximately 330 km as estimated from the measured latency (3.7ms RTT*).

*Distance is estimated based on the following latency assumptions:
・0 km: 0.4 ms RTT
・80 km: 1.2 ms RTT



- **Configuration 4 (with security)**

This configuration represents the pattern with a remote connection over the APN and security enabled. It is basically the same as Configuration 2 except for the remote connection via APN (as in Configuration 3).



5.1.3 Software Specifications

Table 5-1: Software Specifications of the Implemented Systems

Equipment	Specifications
GPU Server	[Confidential VM] Distribution: Ubuntu 24.04 GPU Driver: 570.124.06 CUDA Toolkit: 12.8.1 NVIDIA Container Toolkit: 1.17.5 Kubernetes: v1.33.1 [Host] Distribution: Ubuntu 24.04 QEMU: 9.1.0 (AMD SEV compatible patch applied) [Benchmark software] tsuzumi-7b-v1_3-8k-instruct Checkpoint configuration: 1 batch Repeat batch number: 10
Gateway Server	[Confidential VM] Distribution: Ubuntu 24.04 NFS Server: v4 Kubernetes: v1.33.1 [Host] Distribution: Ubuntu 24.04 QEMU: 9.1.50 +IGVM compatible patch

5.1.4 Hardware specifications

Table 5-2: Hardware Specifications of the Implemented Systems

Equipment	Specifications/Model
GPU Server	PowerEdge R7525
GPU	NVIDIA H100 PCIe 80GB × 1
Storage	NetApp AFF C400
L2 switch/APN-T	Galileo-1
L2 switch	Alaxala AX2630S-24T4XW
L2 switch	Juniper EX4600-40F-AFO

5.2 Measurement Method

5.2.1 Measurement Metrics and Conditions

This section defines the performance metrics and describes the conditions under which the measurements were conducted during the full fine-tuning of the LLM tsuzumi.

Metrics

- **Read Time:** Time required for the application on the CPU to load training data from storage.
- **GPU Time:** Time required for training-related parameter computation on the GPU, including the time required for data transfer between the CPU and the GPU.
- **Save Time:** Time required to save training checkpoints.

Each metric was measured ten times, and the average value was used as the final evaluation result.

Measurement Conditions

The table below summarizes the conditions under which performance measurements for the LLM were conducted.

Table 5-3: Performance Measurement Conditions for LLM (tsuzumi)

Model	tsuzumi-7b-v1_3-8k-instruct
Training Method	Full fine-tuning **The measurement program is a modified version of the standard tsuzumi training program adapted for performance measurement.*
Training Data Size / Type	10 GiB training dataset (Created by duplicating the sample tsuzumi training dataset until the total size reached 10 GiB)
Training Termination Condition	Completion of training for 10 steps

In addition to the conditions listed in the table, the total number of training steps was determined based on the following values:

- **Examples per file:** 240,000
- **Number of files:** 20
- **Total batch size:** 32 samples per step

Using these numbers, the total number of training steps is calculated as follows:

(240,000×20)/32=150,000 steps

The total GPU Time is obtained by multiplying the measured GPU computation time per step by this total step count.

Similarly, the number of checkpoint saves is derived by dividing the total number of training steps by the following checkpoint save interval:

- **Save interval:** 2,000 steps per save

Therefore, the number of save operations is:

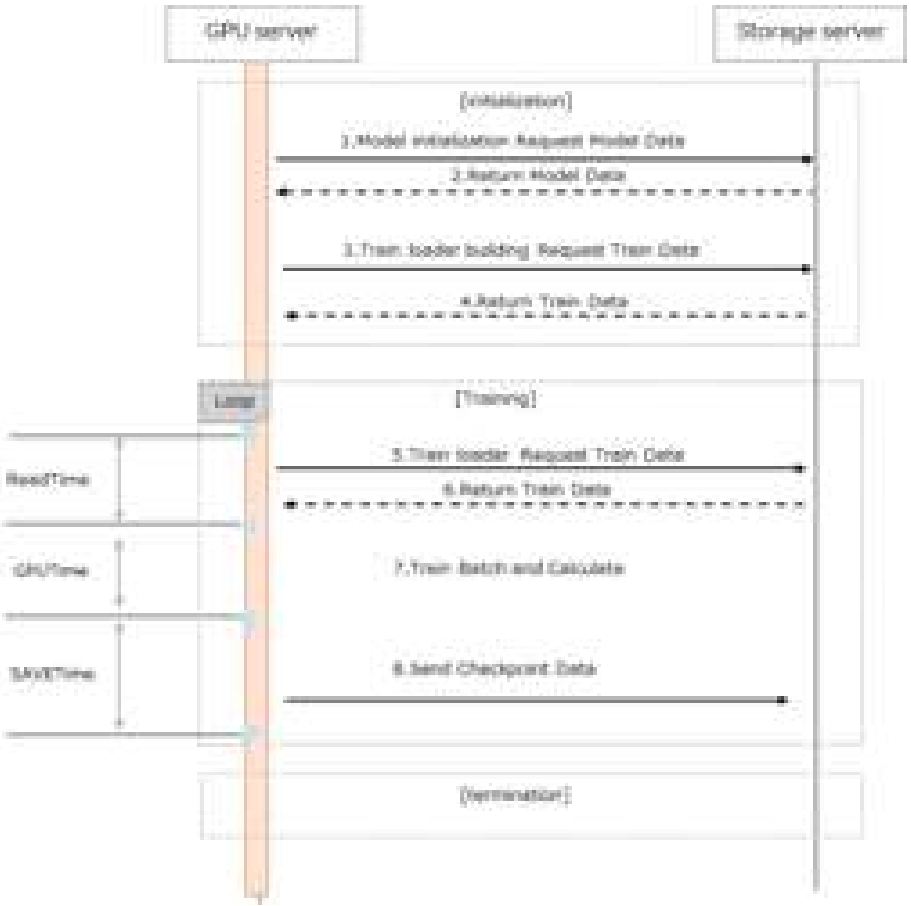
150,000/2,000=75 saves

The total Save Time is computed by multiplying the measured time per save operation by the 75 checkpoint saves.

Measurement Reference (System Sequence Diagram)

The measurement points for Read, GPU, and Save Times correspond to the timing events shown in the general system sequence diagram for AI training (Figure 5-5), quoted from Reference Implementation Model and Proof-of-Concept Reference of Green Computing with Remote GPU - IOWN Global Forum.

This diagram clarifies the measurement points in the AI training flow (data loading → GPU computation → checkpoint saving).



5.3 PoC Technical Finding

Through this proof of concept (PoC), the following technical insights were obtained regarding the impact of distance and data security configurations on remote storage access performance for training LLM (tsuzumi) across geographically separated data centers using APN.

Impact of distance on remote storage access

When considering the proportion of GPU computation time in the training time of the LLM (tsuzumi), the impact of physical distance between data centers on remote storage access performance is shown to be negligible.

Impact of data security configuration

The impact of the data security configuration on the training time of the LLM (tsuzumi) leads to an increase in GPU processing time due to the cryptographic processing overhead of TEE and GPU TEE. As a result, the utilization time of remote GPU resources increases by approximately 47%. However, these results were obtained using the NVIDIA H100. It is known that significant performance improvements can be expected by using Blackwell, the latest TEE-enable GPU model available at the time of writing this report. In addition, since performance in this PoC was measured using a single GPU, future work is expected to include performance evaluation of a GPU TEE-based 8-GPU cluster leveraging Blackwell.

5.4 From the marketing (business) perspective

Whether the 47% increase in remote GPU resource utilization time caused by the data security configuration is acceptable depends on a cost-benefit comparison that takes into account the following factors:

- the additional costs associated with selecting the security options and the increased service usage time
- the financial costs of purchasing and maintaining GPU resources in-house, along with the benefit of reduced security assessment effort when using external resources under the data security configuration

6. PoC's Contribution to IOWN GF

- RIM-TF: This PoC supports a reference implementation model of a distributed data center system for training generative AI models with data security.
- DSDT-TF: This report can also be regarded as a performance evaluation of a reference implementation of a data space with end-to-end confidentiality and can serve as a foundation for broader discussions on future data space implementations.

7. Conclusion

In this PoC, it was demonstrated that even with a data security configuration ensuring end-to-end confidentiality, the LLM training completion time in a remote access

environment equivalent to the IOWN APN 330 km distance remains within 10% of that in a local connection environment. And it was also clarified how the data security configuration affects the overall training completion time.

8. Acknowledgment

We sincerely thank NetApp for kindly loaning the storage system used in this PoC. Their support significantly contributed to the successful completion of the evaluation.

9. Document History

Version	Date	By	Description of Change
1.0			Initial draft

Reference

[MFS FA] Functional Architecture for Protection of Data in Motion: Multi Factor Security Key Exchange and Management, https://iowngf.org/wp-content/uploads/2025/02/IOWN-GF-RD-MFS_Functional_Architecture-1.0.pdf

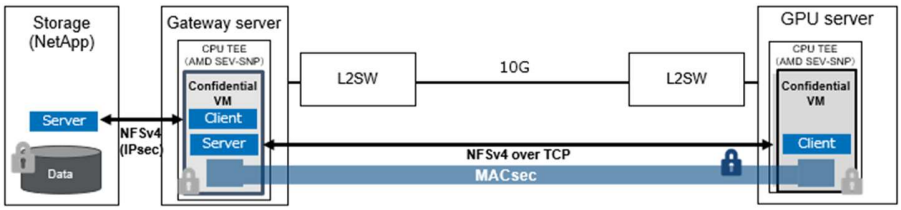
[Security RIM] Security RIM for Green Computing with Remote GPU over All-Photonic Networks, Security RIM for Green Computing with Remote GPU over All-Photonic Networks - IOWN Global Forum

[TSUZUMI] NTT's LLM "tsuzumi", https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202408fa1_s.html

Appendix

This section presents measurement results obtained in both single-hop and two-hop NFS environments. The results examine Read Time and data transfer performance across different network distances (0 km and various APN distances), as well as the impact of security configurations on performance characteristics under different NFS configurations.

First, the Read Time of tsuzumi training data was compared in a two-hop NFS environment under four combinations of encrypted communication (ON/OFF) and TEE (ON/OFF) (see Figure A-1).



The measurement results are shown in Figure A-2. Measurements were conducted under multiple network conditions, including a local (0 km) environment and APN distances of 330 km*, and 960 km*. For reference, additional data is shown on the right measured over the Reference IP Network, configured with 1 Gbps bandwidth and 10 ms added latency, referring to typical specifications of current Internet VPN connections.

*The APN distances were estimated based on the observed network latency.

Comparing the tsuzumi Read Time with the referenced cloud network, the remote connection over the APN shows a significant improvement, even with a data-security configuration.

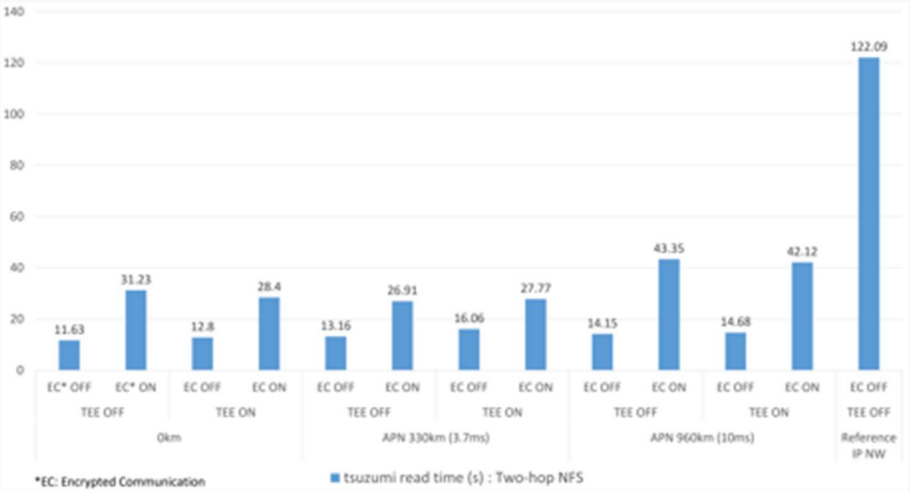
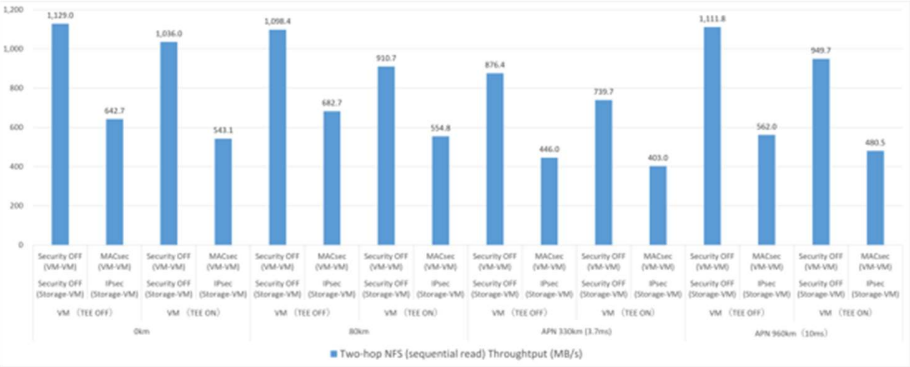
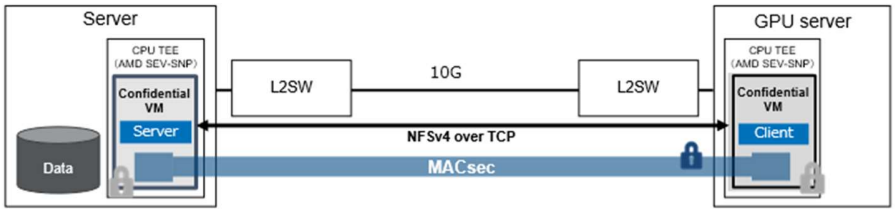


Figure A-3 shows measured results comparing sequential read throughput in a two-hop NFS environment (see Figure A-1 for the configuration). The vertical axis represents throughput in MB/s, while the horizontal axis indicates various VM configuration conditions, including security and IPsec on/off settings.

Measurements were conducted under multiple network conditions, including a local (0 km) environment and APN distances of 80 km, 330 km, and 960 km.

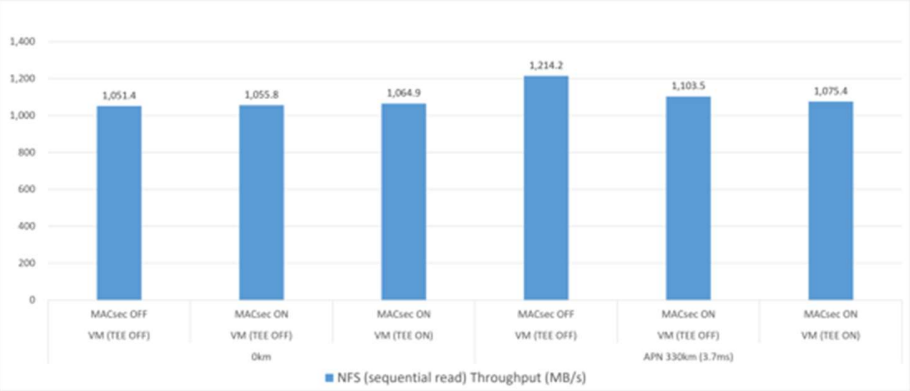


Next, NFS sequential read transfer performance was compared using a 10 GB workload in an NFS over TCP environment (see Figure A-3 for the configuration). The measurements were conducted with encrypted communication enabled, using FIO (Flexible I/O Tester) with a 256-bit encryption key.



The results are shown in Figure A-5. The vertical axis represents the NFS sequential read throughput in MB/s, while the horizontal axis indicates the respective measurement conditions, including security features (MACsec and TEE) with on/off configurations. Measurements were performed under two network conditions: local (0 km) and remote (APN 330 km) environment.

These results indicate that the NFS throughput itself does not seem to significantly degrade under the data-security configuration.



This section presents measurement results obtained to evaluate performance characteristics of a NetApp NFS server (see Figure A-6 for the configuration). Sequential and random read throughput, as well as application-level Read Time of tsuzumi data, were measured to assess the impact of security features (IPsec and TEE) on server-side performance.

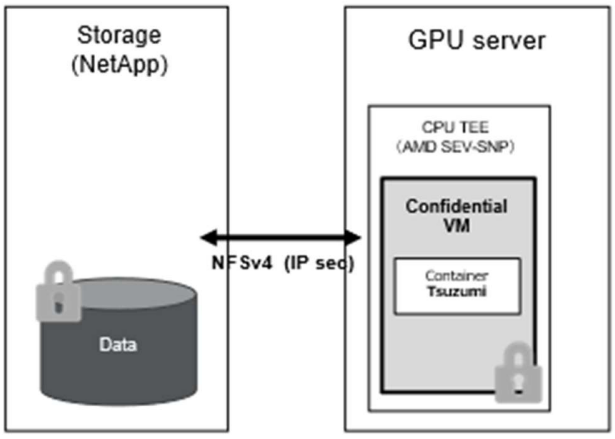


Figure A-7 shows measured results comparing sequential read throughput on a NetApp NFS server under different security configurations (IPsec on/off and TEE on/off).

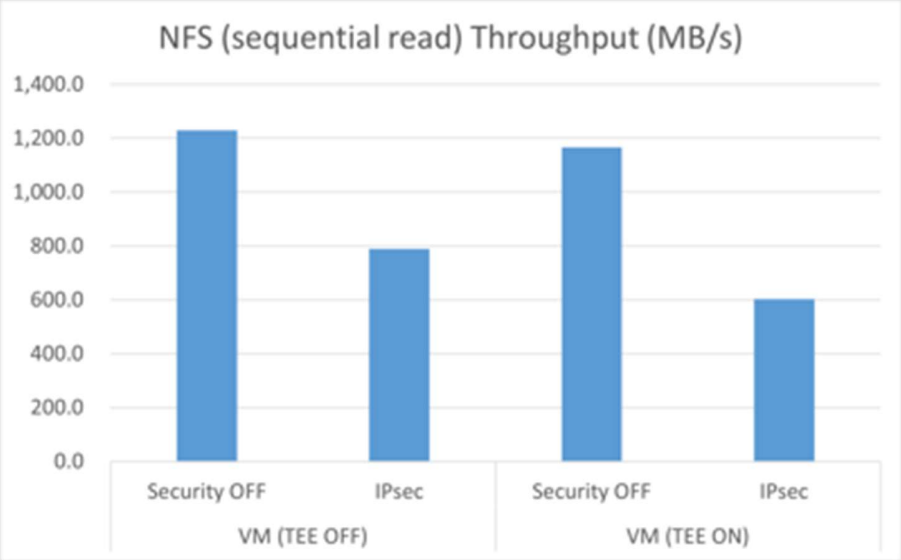


Figure A-8 shows measured results comparing random read throughput on a NetApp NFS server under different security configurations (IPsec on/off and TEE on/off).

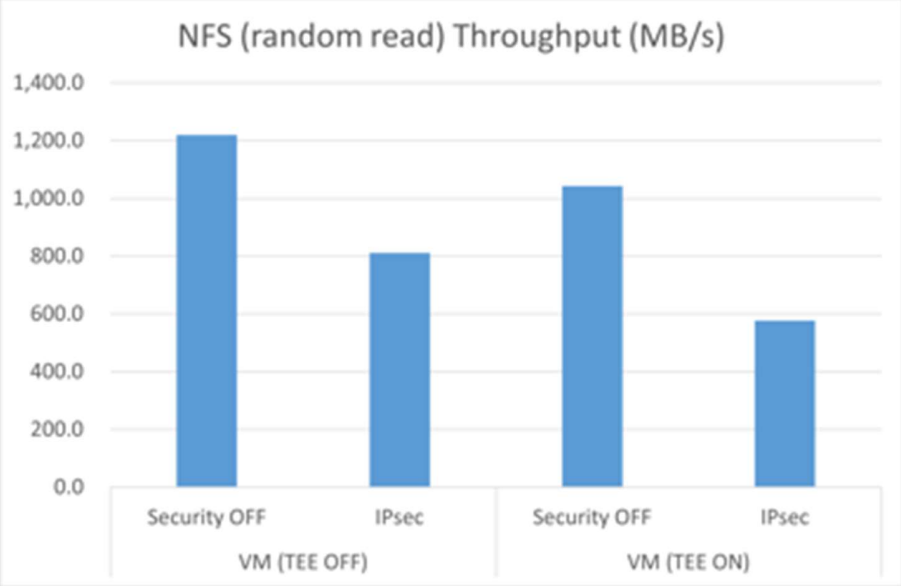
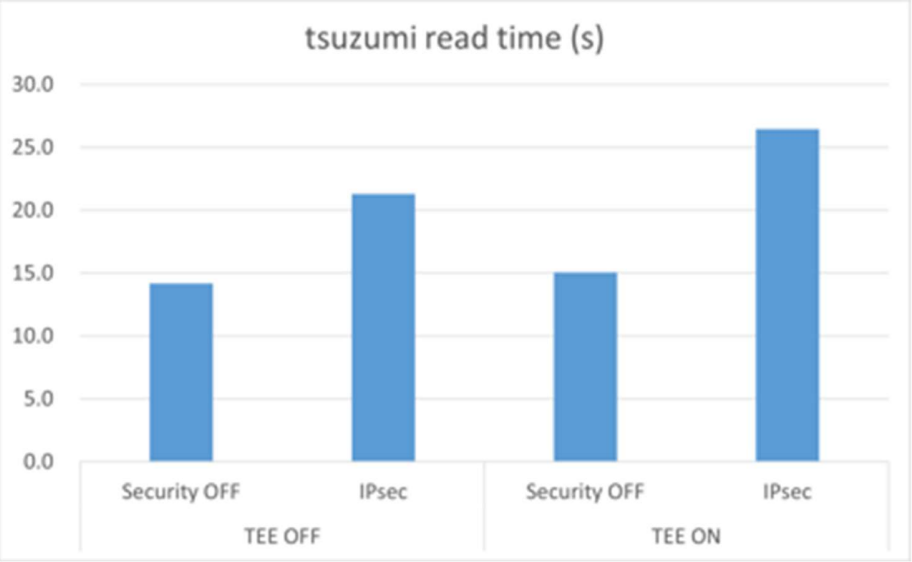


Figure A-9 shows measured results comparing tsuzumi data Read Time on a NetApp NFS server under different security configurations (IPsec on/off and TEE on/off).



The results shown in Figures A-7 through A-9 provide a quantitative view of the performance impact caused by enabling security features on a NetApp NFS server.

For sequential read workloads, enabling IPsec results in a throughput reduction of approximately 35–48%, while enabling both IPsec and TEE leads to a reduction of about 50%. Similar behavior is observed for random read workloads, where throughput decreases by approximately 34–45% with IPsec enabled and by about 53% when both IPsec and TEE are enabled.

At the application level, the tsuzumi data Read Time becomes approximately 1.5 to 1.8 times longer with IPsec enabled, and about 1.9 times longer when both IPsec and TEE are enabled.

Based on these measurements, it can be inferred that enabling security features on a NetApp NFS server typically results in a throughput reduction of roughly 35–50% for NFS data-transfer workloads and increases the application-level Read Time by approximately 1.5 to 1.9 times.