← Back to Dashboard

vLLM Allows Remote Code Execution via PyNcclPipe Communication Service

Basic Information

CVE ID: CVE-2025-47277
DB ID: 3
Severity: CRITICAL
Date Reported: 10/06/2025, 03:39:03

Description

vLLM, an inference and serving engine for large language models (LLMs), has an issue in versions 0.6.5 through 0.8.4 that ONLY impacts environments using the `PyNcclPipe` KV cache transfer integration with the V0 engine.

AI Assessment

Re-Run AI Assessment

Vulnerability Assessment Report (OpenAI)

SCORE: 9.8

Summary

A critical remote code execution vulnerability exists in the vLLM tool when using the `PyNcclPipe` KV cache transfer. The vulnerability results from the TCPStore interface being incorrectly exposed on all network interfaces, enabling untrusted data deserialization, leading to potential remote code execution.

Technical Explanation

The vulnerability stems from the erroneous behavior of the TCPStore interface in the PyNcclPipe KV cache transfer integration. The intended functionality was to bind this interface specifically to a secure, private network as specified by the `--kv-ip` CLI parameter. Instead, the TCPStore interface listens on all available network interfaces, effectively bypassing network restrictions. This deprecated configuration allows for the deserialization of untrusted data, which can be exploited for executing arbitrary code.

Vulnerability Details

Field/Context	Type	Severity	Justification
PyNcclPipe KV cache transfer	Deserialization of Untrusted Data	CRITICAL	The TCPStore component listens on all interfaces, leading to potential RCE through deserialization.

Weakness

Deserialization of Untrusted Data (CWE-502)

Affected Software

vLLM versions >= 0.6.5, < 0.8.5

Potential Mitigations

- Upgrade to vLLM 0.8.5 or newer.
- Ensure that network interfaces are correctly restricted and monitored for all TCPStore uses.
- Implement network segmentation and isolation for critical services.

Related Attack Patterns

- CAPEC-152: Input Data Manipulation
- CAPEC-137: Parameter Injection

References

- https://github.com/vllm-project/vllm/security/advisories/GHSA-hjq4-87xh-g4fv
- https://docs.vllm.ai/en/latest/deployment/security.html
- https://github.com/vllm-project/vllm/pull/15988
- https://github.com/vllm-project/vllm/commit/0d6e187e88874c39cda7409cf673f9e6546893e7

Report Generated By

AI-Powered Vulnerability Assessment System

Confidential - For Internal Use Only

Vulnerability Assessment Report (Ollama Llama3)

SCORE: 9.8

Summary

Critical vulnerability in vLLM versions 0.6.5 through 0.8.4 that allows remote code execution via PyNcclPipe communication service.

Technical Explanation

The issue arises from the use of the `PyNcclCommunicator` class, which allows for peer-to-peer communication between distributed nodes without proper validation. This can lead to remote code execution if an attacker can manipulate the data being transmitted.

Vulnerability Details

Field/Context	Type	Severity	Justification
CWE-502	Deserialization of Untrusted Data	CRITICAL	The `PyNcclCommunicator` class does not properly validate the data being transmitted, allowing an attacker to inject malicious code.

Weakness

Deserialization of Untrusted Data

Affected Software

vllm-project/vllm

Potential Mitigations

- Limit the `TCPStore` socket to a private interface

Related Attack Patterns

- PyNcclPipe communication service

References

-
-
-
-

Report Generated By

AI-Powered Vulnerability Assessment System
Confidential - For Internal Use Only