NVIDIA Tesla P4 GPU - Test Report

Date: February 20, 2026 **Prepared for:** Carrie (Vendor) **Prepared by:** Alphonzo Jones Jr.

Device Information

Field	Value
GPU Model	NVIDIA Tesla P4
VRAM	8 GB GDDR5 (7680 MiB)
Host System	HP Z240 Workstation
PCIe Bus ID	00000000:02:00.0
Driver Version	570.211.01
CUDA Version	12.8
OS	Ubuntu 24.04.3 LTS (Kernel 6.8.0-100-generic)

Summary

The NVIDIA Tesla P4 GPU fails under compute load with **Xid 79 errors** ("GPU has fallen off the bus") and **PCIe RxErr** (receive errors on the PCIe bus). The GPU responds to basic queries at idle but cannot sustain any workload. The card was confirmed defective via gpu-burn stress testing on February 19, 2026. All GPU workloads have been migrated to CPU-only operation.

Test Results

Test 1: nvidia-smi at Idle - PASS

The GPU responds to `nvidia-smi` when no compute workload is running.

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 570.211.01          Driver Version: 570.211.01    CUDA Version: 12.8    |
|-------------------------------+----------------------+----------------------+
| GPU  Name                     Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf              Pwr:Usage/Cap |          Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla P4                           Off | 00000000:02:00.0 Off |                    0 |
| N/A  57C    P8               8W /  75W |      0MiB /  7680MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
```

Result: GPU is detected and reports healthy status at idle. Temperature 57C, power draw 8W (idle).

Test 2: LLM Inference Workload (Ollama) - FAIL

Deployed Ollama (LLM inference server) with `nvidia.com/gpu: 1` resource limit, loading a 3B-parameter model (llama3.2).

Observed behavior: - Model loads into GPU memory initially (~4.9 GB VRAM consumed) - After variable runtime under inference load, the GPU becomes unresponsive - Kernel logs report **NVIDIA Xid 79** error - Pod events show: `failed to get device handle from UUID: Unknown Error` - Requires full node reboot to restore GPU to detectable state

Test 3: GPU Persistence Mode - FAIL

Attempted mitigation by enabling GPU persistence mode (`nvidia-smi -pm 1`) via a privileged init container before starting compute workloads. This is a standard workaround for intermittent Xid 79 errors.

Result: Persistence mode did not prevent the failure. GPU still fell off the bus under load.

Test 4: gpu-burn Stress Test - FAIL

Ran `gpu-burn` (CUDA stress test) to apply sustained compute load directly to the GPU.

Result: GPU failed with **Xid 79** and **PCIe RxErr** (PCIe receive errors), confirming the failure is hardware-related and occurs under any compute load, not specific to the Ollama application.

Error Details

NVIDIA Xid 79 - "GPU has fallen off the bus"

This is a critical hardware error indicating the GPU has lost communication with the host system over the PCIe bus. Common causes include:

- Defective GPU card
- Failing PCIe slot or riser
- Power delivery issues
- Thermal throttling leading to shutdown

PCIe RxErr (Receive Errors)

PCIe receive errors were logged alongside the Xid 79 events, indicating data corruption or signal integrity problems on the PCIe bus. This further supports a hardware-level failure.

Troubleshooting Steps Taken

Step	Action	Result
1	Verified NVIDIA driver installation (570.211.01)	Driver loads correctly
2	Verified NVIDIA container toolkit and device plugin	GPU detected, allocatable in Kubernetes
3	Tested nvidia-smi at idle	GPU responds normally
4	Ran LLM inference workload	GPU crashes with Xid 79
5	Enabled GPU persistence mode	Did not resolve issue
6	Rebooted host node	GPU recovers temporarily, fails again under load
7	Ran gpu-burn stress test	Confirmed failure: Xid 79 + PCIe RxErr
8	Switched all workloads to CPU-only mode	Workaround successful, services restored

---

Timeline

Date	Event
Feb 12, 2026	GPU deployed with Ollama LLM server, initial operation successful
Feb 15, 2026	GPU monitoring dashboards and temperature alerts configured
Feb 19, 2026	GPU failures began; Xid 79 errors observed under load
Feb 19, 2026	GPU persistence mode attempted as fix - unsuccessful
Feb 19, 2026	gpu-burn stress test confirmed GPU hardware failure
Feb 19, 2026	All workloads switched to CPU-only mode
Feb 20, 2026	nvidia-smi confirms GPU still responds at idle but cannot sustain load

Conclusion

The NVIDIA Tesla P4 GPU is defective. It passes basic detection and idle queries but **fails consistently under any compute workload** with Xid 79 (GPU fallen off the bus) and PCIe receive errors. Multiple mitigation attempts (persistence mode, node reboots) were unsuccessful. The failure pattern is consistent with a hardware defect in the GPU card itself. The GPU should be replaced under warranty/return.